Semantic Enrichment of Places

Understanding the Meaning of Public Places from Natural Language Texts

Ana Cristina da Costa Oliveira Alves

Department of Informatics Engineering

Faculty of Sciences and Technology of the University of Coimbra

A thesis submitted for the degree of

PhilosophiæDoctor (PhD)

2011 December

Abstract

In this thesis, we present our approach to the challenge of assigning semantic annotations to places, what we call *Semantic Enrichment of Places*. These annotations are automatically extracted by applying natural language processing and information extraction techniques that have been thoroughly applied and tested using the World Wide Web as the primary source. Here, we are particularly focused on extracting information that allows an external system to distinguish one place from other places that are spatially or conceptually close. This is because the *meaning of a place* is a function of its most salient features, present in the textual descriptions found in online resources about that place. In the situation under investigation, places correspond to *Points Of Interest* (POIs), as these are abundant on the Web. By definition, a POI is a place with meaning to someone and, if it is available online, it is likely that that person's interest is shared by many people. In this approach, the Web is first crawled to obtain a large number of POIs and then each of them is analyzed in order to obtain their individual *Semantic Index*: the set of words that best define them. Besides analyzing POIs, we also propose the application of such an approach in several different contexts and we integrate these contexts in a multi-faceted view of place.

Resumo Alargado / Extended Abstract

O conceito de lugar tem sido (re)definido de diferentes modos ao longo da história, independentemente dos aspetos culturais e desenvolvimentos tecnológicos. Do ponto de vista do utilizador, um lugar está sempre associado a um significado, que pode ser pessoal, consequência das experiências naquele lugar, ou público, ou seja, como o lugar é reconhecido de uma forma geral. Um lugar pode ser descrito ou referenciado sob diferentes perspetivas, dependendo do que pretende ser comunicado (e.g. a funcionalidade de um lugar, as suas características físicas, os serviços oferecidos, a relação do utilizador com o lugar).

Com o aumento da disponibilidade de dados georreferenciados na Web, o surgimento de novos serviços baseados em localização[0.1] tem focado a atenção da comunidade científica e da indústria em torno da definição de lugar. O potencial de sucesso de tais serviços é altamente dependente do modo como estes percebem o contexto.

Nos últimos anos, a quantidade de informação descritiva sobre lugares tem aumentado consideravelmente. Como grande parte desta informação é textual, é necessário aplicar técnicas de Processamento da Língua Natural (NLP)[0.2] e Extração de Informação (EI) [0.3] para obter o *significado dos lugares* intrínseco nesta quantidade massiva de informação. Esta informação é na sua maioria produzida por utilizadores que são baseados no senso comum para descrever a funcionalidade, serviços e propriedades dos lugares em questão.

Nesta tese, é apresentada uma abordagem ao desafio de assinalar anotações semânticas a lugares, definida formalmente como *Enriquecimento Semântico*

[0.1] Em inglês *Location-Based Services (LBSs)*
[0.2] Em inglês *Natural Language Processing*
[0.3] Em inglês *Information Extraction*

de Lugares. Estas anotações são automaticamente extraídas da Web através de técnicas de NLP e EI testadas e aplicadas no mesmo âmbito. Esta tese foca-se na extração de informação que permita a outros sistemas/utilizadores distinguir um lugar de outro que estão geograficamente próximos ou que pertençam a uma mesma categoria. Isto porque o *significado de um lugar* é uma função das suas mais salientes características presentes em descrições textuais extraídas de recursos *online*.

No âmbito desta tese, lugares correspondem a Pontos de Interesse (POIs)[0.4], uma vez que esta representação é atualmente abundante na Web. Pressupõe-se a hipótese de que se um POI é um lugar com significado para alguém e, se estiver presente *online*, é provável que seja partilhado por outras pessoas. Na abordagem proposta nesta tese implementada pelo sistema KUSCO, é feita inicialmente uma pesquisa tanto para obter POIs em grande quantidade, como para obter descrições textuais sobre estes. Cada uma desta descrições é processada e analisada em conjunto de modo a obter para cada POI o seu *índice semântico*: o conjunto de palavras que melhor o definem. Além de analisar POIs, propomos também a aplicação desta abordagem num contexto mais genérico e sob diferente perspetivas.

Um sistema que seja capaz, de extrair a semântica relevante de lugares pode ser útil para qualquer sistema baseado em contexto que tenha comportamentos diferentes de acordo com as características do lugar onde é utilizado. O nível de detalhe da informação considerada nesta tese representa um camada adicional a outros sensores (e.g. GPS, acelerómetros, bússolas, sensores luminosidade e proximidade), potenciando, eventualmente, o comportamento mais inteligente dos dispositivos mveis. Por exemplo, um algoritmo de *machine-learning* num *smartphone* poderia ser treinado para apresentar uma interface diferente consoante o tipo de atividade realizada pelo seu utilizador (e.g. lazer, trabalho, compras). Outras utilizações podem ser ilustradas, desde aplicações para navegação (e.g. navegação por conceitos, procura de um lugar dado um conjunto de palavras relacionadas)

[0.4] Em inglês *Points-of-Interest*.

até à análise de interações sociais com os lugares e utilização do espaço (e.g. procurar correlações entre eventos e a presença de pessoas).

To my family, Rui, my husband, and our beloved sons, Eduardo and João Gabriel.

Acknowledgements

I would like to acknowledge some people who have been fundamental during the process of completing this work. With their support it was possible to carry on even at those times when there was no light at the end of tunnel. I hope not to forget anyone's name and would like to acknowledge the help of everyone involved, but, as I am only human I may miss someone out. If so, my apologies to anyone not mentioned below.

I start with the main person responsible for bringing me to this academic research, my supervisor Francisco. As well as being an advisor and interrogator about the subject under discussion, he is a friend. For some, where to separate the two could be a problem, for me both of these roles have been a great help. He is an exemplar of persistent research and long-term studying, who has taught me that there are no fundamental answers, but always challenging questions to tackle.

A great part of this work is also devoted to my family, my husband Rui and our children Eduardo and João. They are incredible guys who have been extremely patient with the limited time I have had available until now. I thank them for all those times that I could not be with them, especially Eduardo; I hope now to be able to watch all the matches you play. I am sure that my family have been looking forward to the conclusion of this work even more than I have.

For my mother and father, I would like to acknowledge all their advice and the opportunity they gave me to carry on my studies. Without their encouragement I could never reach this milestone. For my brother, I hope I have given him an example of perseverance.

I would like to thank Filipe and João for their research and for their work in using the data produced by the system developed during this thesis.

For the AmILab staff, Professors Ana Almeida and Carlos Bento, and for researchers Marisa, Merkebe, Marco, Nuno, Norberto, António, Fábio, Jorge, Bruno and José, I am grateful to have worked in such a great environment and atmosphere of friendship.

List of Figures

List of Tables

List of Algorithms

Contents

1

Introduction

This chapter begins with the motivation for this thesis, introducing most of the terminology and concepts used in the remaining chapters. The problems and goals addressed are then outlined, defining the scope of the developed work. Finally, the organization of the thesis is described. The main ideas have been published elsewhere in several papers, on which important parts of this document are based. However, this thesis describes the issues addressed in these papers in greater detail and with examples, along with other issues not covered in these documents. The publications which have been used are:

- Ana O. Alves, Filipe Rodrigues and Francisco C. Pereira, 2011. **"Tagging Space from Information Extraction and Popularity of Points of Interest"**. In Proceedings of the Second International Joint Conference on Ambient Intelligence (*AmI'11*).

- Ana O. Alves, Francisco C. Pereira, Filipe Rodrigues and João Oliveirinha, 2010. **"Place in perspective: Extracting on-line information about Points of Interest"**. In Proceedings of the First International Joint Conference on Ambient Intelligence (*AmI'10*).

- João Oliveirinha, , Francisco C. Pereira and Ana O. Alves, 2010. **"Acquiring semantic context for events from on-line resources"**. In Proceedings of the 3rd International Workshop on Location and the Web (*LocWeb'10*).

- Ana O. Alves, Francisco C. Pereira, Assaf Biderman and Carlo Ratti, 2009. **“Place Enrichment by Mining the Web”**. In Proceedings of the 3rd European Conference on Ambient Intelligence (*AmI’09*).
- Francisco C. Pereira, Ana O. Alves and Assaf Biderman, 2009. **“Fusion of Semantics with Mobility Information: Prospects and Opportunities”**. In Proceedings of the International Conference on Pervasive Computing *Pervasive’09 Workshop: InMotion’09 - Pervasive Technologies for Improved Mobility and Transportation.*
- Ana O. Alves, Bruno Antunes, Francisco C. Pereira and Carlos Bento, 2009. **“Semantic Enrichment of Places: Ontology Learning from the Web”**. *International Journal of Knowledge-Based & Intelligent Engineering Systems - Intelligent agents and services for smart environments*, Volume 13 Issue 1, IOS Press Amsterdam, The Netherlands.
- Bruno Antunes, Ana O. Alves, Francisco C. Pereira, 2008. **“Semantics of Place: Ontology Enrichment”**. In Proceedings of the 11th Ibero-American Conference on Artificial Intelligence (*IBERAMIA 2008*).
- Ana O. Alves, Alves, Raquel Hervás, Francisco C. Pereira, Pablo Gervás and Carlos Bento, 2007. **“Conceptual Enrichment of Locations Pointed Out by the User”**. In Proceedings of the 11th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (*KES 2007*).
- Ana O. Alves, 2007. **“Semantically enriched places: An approach to deal with the position to place problem”**. In the Adjunct Proceedings of the International Conference on Ubiquitous Computing *Ubicomp’07 Doctoral Colloquium.*

1.1 Motivation

The concept of place has been recurrently inconsistent throughout history, regardless of culture and developments in communication technology. From the perspective of a user, places are often associated with meaning, and different people relate to places in different ways. It is noticeable that a place can be described or referenced according to a range of perspectives, depending on what is intended to be communicated (e.g. a place's function or its physical properties or its content or the user's relationship with the place).

In his thorough overview of the philosophical concept of place, Edward Casey [Casey, 1998] gathers perspectives from Pre-Socratic to 20th-century philosophers. Plato suggests that "all existence must of necessity be in some place and occupy a space". For Aristotle, "place is a container; it is not matter or form but the limit or vessel that contains a thing". For many cultures, the notion of place has connotations of a substantial conceptual framework that contributes to the definition of things and people. The Pre-Socratics were known by the places they lived in, and this tradition persists in many cultures nowadays. After Aristotle, the philosophy of place became marginalized for many centuries, but was strongly revitalized during the Enlightenment era. In *An Essay on Human Understanding*, John Locke suggested that "Our idea of place is nothing else but such a relative position of anything". Later on, Descartes and Newton established much of the ground on which we now base our concepts of place and space. More recently, new philosophical trends have revitalized the discussion with new ingredients (e.g. relativism, post-modernism, place as a social function, and virtual reality).

Until recently, the importance of this discussion was restricted to the realm of philosophy, with only some minor implications for practical fields such as geography or physics. However, it has become hugely relevant, with the emergence of Geographical Information Systems (GISs) over the last two decades. The vast number of GIS databases that are mutually incompatible is remarkable, and one of the reasons for this is the ambiguous meaning of the concept of place. More recently, the new trends in Location Based Services (LBSs) emanating from the Mobility and Ubiquitous Computing communities have made the meaning of place prominent issue to tackle. This meaning of place has to be shared with users in unambiguous and, more importantly, practical ways. Flexible representations that allow different perspectives are becoming ever more important.

1. INTRODUCTION

The vision of Ubiquitous Computing is rapidly becoming a reality as our environment grows increasingly replete full of sensors, widespread pervasive information and distributed computer interfaces. New challenges are emerging in creating coherent representations of information about places using this multitude of data. These challenges are being addressed in the various areas that involve *Data Fusion*, as significant progress has been made at specific levels of representation. Many off-the-shelf products integrate GPS, Wi-Fi, GSM, accelerometer, and light sensor data and furthermore employ elaborate software that is capable of integrated contextual processing. Many have noted, however, that a piece in this puzzle is missing, without which it is difficult to enable context-aware scenarios: semantic information. While semantic information has been available for centuries, the Internet has dramatically increased its abundance and availability. In each of the four dimensions of context awareness (*who*, *what*, *where*, *when*), semantics is present to varying degrees. This thesis focuses on the “what” dimension having the location (i.e. “where”) already defined: the semantics of place.

With the growing amount of geo-referenced data available on the Web, the increasing number of LBSs has intensified the spotlight that focuses on the definition of place. The potential for the success of such services is highly dependent on how they perceive the context. For example, a context-aware system should be able to adapt according to the place in which the user is (e.g. work, home, cinema, shopping center, etc.), and a simple service for a smartphone would be to detect the availability of the user according to place, or maybe change the interface (e.g. a “different skin”, a different set of applications, or ringtone volume), but this can only be achieved via the perception of place. Such perception is not trivial for the reasons just explained: place has many dimensions inherently associated with it. For example, a place can be described with geographical, demographic, environmental, historical, and perhaps also commercial attributes. The meaning of place is derived from social conventions, its private or public nature, the possibilities for communication, and many other factors.

The difficulty in the unambiguous conceptualization of place stems from its association with *space* and from the number of different perspectives that may arise. Considering the simple question “Where am I?”, there is a range of possible answers: relative to function (“I’m at work”); relative to someone (“I’m at my friend’s place”, "I'm with John"); relative to scale ("I'm in the US", "I'm in New York"; "I'm on 14th Street"); relative to objects ("I'm in my car", "I'm outside the stadium"). To this list of physical references, we can add the wealth of metaphorical creations of place (e.g. "I'm in second life", "My mind is somewhere else").

Over the last few years, the amount of online descriptive information about places has reached considerable proportions for many cities in the world. As such information is mostly in Natural Language text, Information Extraction techniques are needed to obtain the *meaning of places* that underlies these massive amounts of commonsense and user-made sources.

A system that is able to extract relevant semantics from places can be useful for any context-aware system that behaves according to position. The level of information considered in this thesis adds another layer to other sensors (GPS, accelerometer, compass, communications, etc.), eventually pushing forward the potential for intelligent behaviour. For example, a machine-learning algorithm in a smartphone could be trained to present a different interface according to the type of activity (e.g. leisure, work, shopping) inferred from place tags. Other uses can be imagined, from navigation applications (e.g. navigating by concepts, searching for a place given related words) to the analysis of social interactions and space use (e.g. finding correlations between events and the presence of people).

1.2 Research Question

The problem with the semantics of place has already been noted by many in the field of Ubiquitous Computing [Aipperspach et al., 2006; Harrison and Dourish, 1996; Hightower, 2003] as a valid research challenge. This thesis agrees with their perspective and aims to further explore this topic with particular emphasis on methodology and short-term real-world applications. The focus is on the most elementary and unambiguous information about a place: its latitude/longitude. Our question is: what does a specific position *mean* from a common sense perspective? The answer we propose involves the representation of concepts through a *tag cloud*, where a concept here is a noun in a given context and this context is given by its related concepts. Our task is thus to define the most accurate method of finding that set of concepts from a given source, i.e. the Internet. Our method is based on the hypothesis that the tag cloud of any given point in space will be a function of the semantics of its surrounding points of interest (POIs). A POI is a tuple with a latitude/longitude pair, a name and, optionally, a category such as *restaurant*, *hotel*, *pub*, *museum*, etc. It represents a place with meaning to people. The work presented here focuses on the semantic enrichment of POI data.

Bearing in mind not only the importance of the automatic extraction of information but also the ability to retrieve the right information for a given *place*, the research goal of this thesis is the development of a system capable of locating related information about places on the Web, extracting relevant terms associated with them, and presenting this information on a semantic level using lightweight ontologies.

1.3 Main Contributions

This thesis presents an approach to such a system and its implementation, resulting in an architecture called KUSCO: Knowledge Unsupervised Search for instantiating Concepts on lightweight Ontologies. KUSCO has an architecture based on Web Services, which fits well into a distributed environment well and promotes sharing in a Web 3.0 philosophy.

During the development of this system, some research issues were addressed. Some of these contributions are:

- a modular methodology for the assignment of semantics to a place, which is divided into three main steps: the retrieval of related information on the Web, the extraction of terms from this information (mostly textual descriptions), and the contextualization and calculation of the relevance of these terms;
- a new perspective of semantic enrichment applied to space analysis, and its implementation;
- a proposed representation of place semantics;
- an implemented system, KUSCO, easily deployable online, which has already been applied in other projects;

- an independent validation method that compares KUSCO's performance to other Information Extraction services, currently being used in many projects.

1.4 Methodology

The strategy adopted to approach the research question outlined above lies in *empirical generalization*[Cohen, 1995] and this can be classified as an exploratory and an assessment study since, in the first place, it collects large quantities of data and analyzes them in many ways to find regularities in order to detect patterns; and, in the second place, it establishes baselines and benchmarks for use when making comparisons with other equivalent systems.

Specifically in its first phase, KUSCO used third-party ontologies about the generic types of places (e.g. restaurants, museums, cinemas, etc.) to select the most relevant concepts in each POI. After some experiments, we observed that the quality of data extracted was not of a high level and this is the most important factor behind later instantiating place ontologies. Also, we concluded that it would be hard to justify the choice of a specific ontology instead of another from those third-party ontologies. As an alternative to this first attempt, the final methodology which will be detailed in this thesis is the exploration and extraction of information from other sources like Wikipedia and the reuse of Upper Level Model Ontologies that are connected and mapped to a range of Common Sense Resources like WordNet, Yago, and so on. Since data extracted were directly related to the sources they come from, we opted to establish different "perspectives" dependent on the different sources and different ways of extracting this data. These perspectives were later compared and the similarity and the complementarity between them was analyzed.

With regard to the integration of the data produced from other public taxonomies, the ontologies first considered were then replaced by Upper Level Ontologies that were further instantiated with the data produced. Figure 1.1 presents the adapted architecture with focus on the extraction phase (emphasized module). The concept definitions for inferring the overall *meaning of place* were retrieved from generic taxonomies that were also considered by some as Upper Level Ontologies. Hence we tried to reuse available resources from the scientific community to integrate them in a modular way to produce structured representations about places.

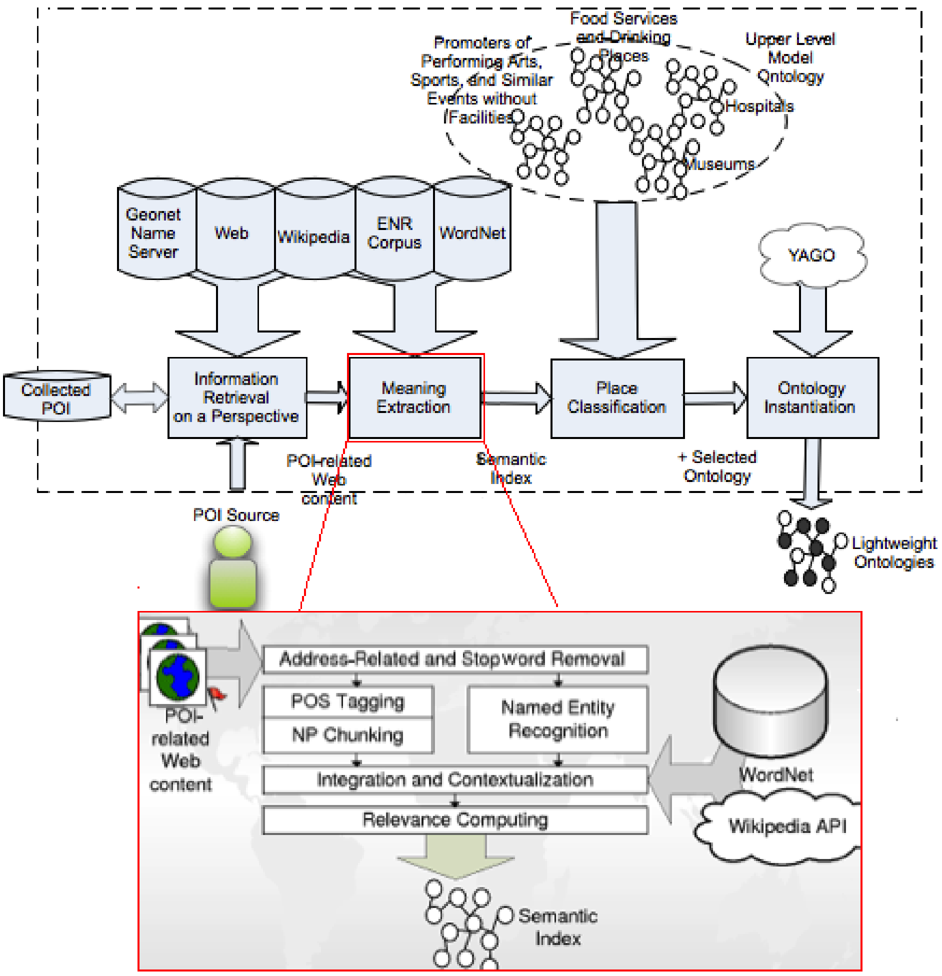


Figure 1.1: Present KUSCO system architecture.

The main dimension for measuring the performance of the KUSCO system is the accuracy of the data produced. There are some modules inside the system that we chose to validate independently beyond the overall accuracy of the whole system. This is because we were also interested in evaluating how good the behavior of each isolated module is, depending on the level of accuracy of the data used as input. The main source used to validate each module was based mainly on manual judgment through online surveys. Other exploratory studies were conducted, such as clustering of the data.

1.5 Thesis Outline

This thesis comprises nine chapters. Chapter 1 presents its motivations, goals, contributions and structure. Chapter 2 introduces the main concepts used in this thesis and presents related work. It starts by describing place representations in Location Context, Natural Language Processing, and Information Retrieval and Extraction, and finishes with an overview of the representation of Semantics with a particular focus on Knowledge Resources. Chapter 3 presents a generic model for Semantic Enrichment and its application to places. The KUSCO system is presented across four chapters: chapter 4 is dedicated to Information Retrieval about places; chapter 5 uses the Natural Language Processing concepts previously introduced to detail the information extraction process; chapter 6 shows that places can be described from different perspectives, each one using either a different source of information or the same source in a different manner; and the final chapter about the KUSCO system, chapter 7, is devoted to the semantic aspect of Place representation. Chapter 8 presents several experiments that were carried out to assess the performance of the system and the quality of the data produced. Chapter 9 concludes this thesis by presenting the main contributions of this work. This chapter also outlines new research and application directions in areas such as information extraction, term-weighting techniques, modeling and common-sense reasoning about places.

2

Literature Review

The goal of this chapter is to introduce the reader to the main subjects present in this thesis and to the related state of the art, in order that the reader may become familiar with the concepts and algorithms used. The first section starts by describing Location in Context Awareness, which is a topic of Ambient Intelligence (itself a subfield of Artificial Intelligence) which deals with implicit situational information related to *where* users or objects are located. Focusing on the symbolic aspect of Location Context, this section introduces related research to represent Places. In a different subfield of Artificial Intelligence, the second section provides useful approaches from Natural Language Processing that will later be integrated with the system here proposed. As online information is so important in this thesis, the state of the art on how to retrieve and extract relevant pieces of data is presented in the following section. Finally, the fourth section introduces Semantics, which is an important topic for an understanding of the context given to the data extracted.

2.1 Location in Context Awareness

Context is any information that can be used to characterize the situation of an entity at a given time and/or location. An entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and application themselves. Context generally refers to all types of information pertaining to a service and/or the user of the service [Abowd et al., 1999]. A system is context-aware if it observes, reacts and changes according to the context. Context information can be gathered from several sources including sensors, devices, data repositories and information services. Context data can be used to make inferences.

The key aspects of context are: location, agent or person, time and activity. These elements are used to answer basic questions related to the user, place or object which is the target of context representation: *who*, *what*, *where*, *when*. The main focus of this research is to represent *What* assuming that location is defined already. The representation used is based on labels that can be organized through common-sense lightweight ontologies. The next subsection deals with how location is presently represented in Pervasive Computing. The following subsection extends this information in order to attach more symbolic and user-readable labels.

2.1.1 Location Modeling

As a formal definition, location models can be classified into four main types [Ye et al., 2007]: Geometric, Symbolic, Hybrid or Semantic. *Geometric location models* are sometimes also referred as metric or coordinate, and are based on geometric coordinates, as used by GPSs, referring to a point or geometric figure in a multi-dimensional space, typically, a plane or a three-dimensional space. The topological properties of such a space allow the calculation of distances between locations and their inclusion in other locations. *Symbolic location models* are also called hierarchical, topological and *descriptive*, and are based on symbolic coordinates which define positions in the form of abstract symbols, e.g. room and street names, etc. In contrast to geometric coordinates, the distance between two symbolic coordinates is not formally defined. Also, topological relations like spatial containment cannot be determined without further information about the relationship between symbolic coordinates. Symbolic location models provide this additional information on symbolic coordinates. The *hybrid location model* considers both geometric and symbolic coordinates. While the first three model types (geometric, symbolic, hybrid) are mainly devoted to the spatial relationship between locations, the last one, the *semantic location model*, which is the focus of this thesis, is orthogonal to symbolic and geometric representations.

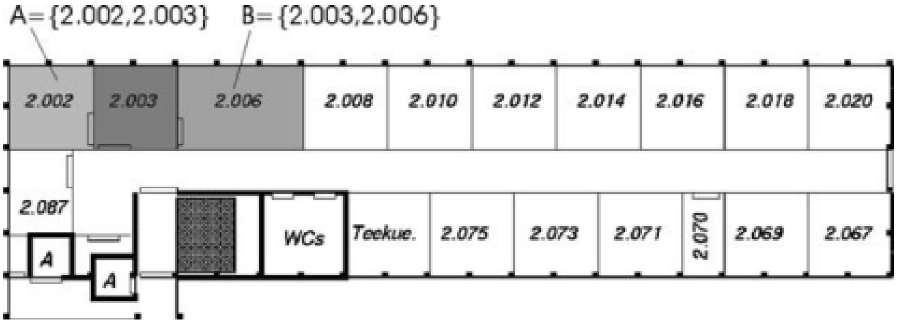


Figure 2.1: Symbolic location model based on sets [Coschurba et al., 2002].

As an example of a geometric location model, Coschurba's model [Coschurba et al., 2002] proposes a 2.5-dimensional approach to describe a three-dimensional (3D) shape by specifying its base as a two-dimensional (2D) and its height as a numeric value (0.5D). Only the coordinates for the space's base shape are recorded, which can reduce the amount of coordinate data and will allow the application of geometric computations on each shape.

A simple approach to the representation of symbolic coordinates is partially-ordered sets[Becker and Dürr, 2005]. A set L of symbolic coordinates forms the basis for the approach. Locations comprising several symbolic coordinates are defined by subsets of the set L. As a simple example, consider a building with several floors. The set L consists of all the room numbers of this building. The second floor as shown in Figure 2.1 can be modeled by the set $L_{floor2} = 2.002, 2.003, ..., 2.067$. Further arbitrary locations may be defined, e.g. the locations $A = 2.002, 2.003$ and $B = 2.003, 2.005$. This model can be used to determine overlapping locations and, as a special case of overlapping locations, the containment relation, by calculating the intersection of two sets $L1$ and $L2$. If $L1 \cap L2 \neq \emptyset$, then $L1$ and $L2$ overlap. If $L1 \cap L2 = L1$, then L2 contains L1. Thus, this model can be used in a range of queries where the range is defined by one set R of symbolic coordinates, and all subsets of R define locations within R.

A more representative example of symbolic model is presented by Brumitt and Shafer [Brumitt and Shafer, 2001]. The model is not a geometric model since it can represent containment and connectedness relationships within a space and not with any specific geometric position of the object in that space. Moreover, it has a lattice structure and a friendly naming system, both of which allow a person to perform queries according to information of symbols. The prime advantage of symbolic models is that they have clear representations of spatial relationships which are easily understandable by humans. However, it is extremely costly to construct and maintain symbolic models manually.

A hybrid location model was proposed by Jiang and Steenkiste [Jiang and Steenkiste, 2002], where they decompose the physical environment into different levels of precision and feature a self-descriptive location representation of each level. At a lower level of decomposition, they use a local and 3D-coordinate system (latitude, longitude, altitude) to define points or areas for which there is no name in the hierarchical tree. As an example, they can identify a specific printer in the Carnegie Mellon University campus through the identifier *cmu/wean-hall/floor3/3100-corridor#(10,10,0).*

A semantic representation provides other information around a place, such as a bus route or a snapshot of interest. Thus, semantic location models introduce semantics into location information and generate human-readable semantic locations. They can structure space according to the special requirements of the application or service. The HP Cooltown [Kindberg. et al., 2000] was the first to introduce a semantic representation of locations. Its main goal was to support a web presence for people, places and things. They used Universal Resource Identifiers (URIs) for addressing, physical URI beaconing and sensing of URIs for discovery, and localized web servers for directories in order to create a location-aware ubiquitous system to support nomadic users. They also distinguished in addition to physical and semantic locations a third type of location, the geographical location. Geographical locations are city names, zip codes and postal addresses.

The semantic location model Ubiquitous Web [Vazquez et al., 2006] was planned as a pervasive web infrastructure in which all physical objects are socially tagged and accessible by URIs, providing information and services that enrich user's experiences in their physical context, as the web does in cyberspace.

Another semantic location model was proposed in the Nimbus framework [Roth, 2004], which encapsulates all functions related to positioning and mapping to semantic locations. A semantic location is more than a position since it generally refers to areas and represents spatial entities. For example, for the physical place *Campus, University of Hagen*, the framework created a unique mapping: *campus.university_hagen.de*, thus forming a hierarchical namespace. A hierarchy contains domains with a similar meaning, e.g. domains of cities or geographical domains. Each hierarchy has a root domain and a number of subdomains; each of these can in turn be divided into other subdomains. The system processes information about location at a symbolic level which can be of different types: locations with political meaning (countries, states or cities), geographical locations (mountain, rivers or forests), temporary locations (construction zones or fairs) or other locations (campuses, malls, city centers). In our system, while the first two types are more static entities, not changing their boundaries frequently and being easily found in gazetteers[2.1], the last two location types are dynamic and it is a challenge to discover information about them. Furthermore, we are mainly interested in non-geographical features of a given location, including its functions, reception, activities, or services.

2.1.2 From Location to Place

As argued in [Hightower, 2003], absolute position such as the latitude/longitude pair is a poor representation of place. From the human perspective, places are often associated with meaning, and different people relate to places in different ways. The meaning of place is derived from social conventions, its private or public nature, possibilities for communication, etc. [Genereux et al., 1983; Kramer, 1995]. As argued by [Harrison and Dourish, 1996] with regard to the distinction between the concept of place from space, a place is generally a space with something added - social meaning, conventions, cultural understandings about role, function and nature. Thus, a place exists once it has meaning for someone and the perception of this meaning is the main objective of our research.

[2.1] Geographical dictionaries.

Beyond position, the name and labels associated with location are also of great importance in the characterization of places [Zhou et al., 2007]. While this is obvious to *personal* places, we think this is also true for public places. According to [Relph, 1976], place consists of three components: physical setting, i.e. the locale of a place; activities performed at a place; and the meanings of a place to the *public* and the individual. In an urban view of a city, people generally create more POIs referring to *buildings* than to other categories of places like parts inside buildings, regions, junctions, and others [Falko Schmid, 2009].

We have also investigated the possibility of automatically associating labels in the relevant literature. These approaches either use additional information, such as time of day and point-of-interest databases to determine the type of building, or attempt to assign labels by comparing places across users. Some works use machine-learning algorithms to infer these labels based also on other variables (time of day, weekday, etc.). These labels are limited to generic and personal ones like work, home, friend, etc. However, our approach is not centered on the user. Instead, it focuses on the *place* itself and this representation should include public aspects and the functionality of places, since the relation between a specific individual and the place itself is not of great importance. We think that a richer representation of *place* with more meaningful common-sense associated concepts associated will complement works such as those previously described.

In [Lemmens and Deng, 2008], the authors propose a semi-automatic process of tag assignment, which integrates knowledge from Semantic Web ontologies and the collection of Web 2.0 tags. This approach should be correct in theory: it shares the formal soundness of Ontologies with the informal perspective of social networks. However, it is in essence impracticable: for each new POI/category the main points have to be chosen manually. The dynamics of this kind of information, particularly when depending on Web 2.0 social networks, would demand enormous resources to keep the information up to date, and the required compliance with semantic standards by individual users already seems a lost battle.

Working in a different direction, Rattenbury et al. [Rattenbury et al., 2007] identify place and event from tags that are assigned to photos on Flickr. They exploit the regularities of tags which regard to time and space at several levels, so when “bursts” (sudden high quantities of a given tag in space or time) are found, they become an indicator of an event of meaningful place. Accordingly, the reverse process is possible, the search for the tag clouds that correlate with that specific time and space. They do not, however, make use of any enrichment from external sources, which could add more objective information, and their approach is limited to the specific scenarios of Web 2.0 platforms that carry significant geographical reference information.

Other attempts have also been made towards analyzing Flickr tags [Dubinko et al., 2006; Jaffe et al., 2006] by applying ad-hoc approaches to determine “important” tags within a given region of time [Dubinko et al., 2006] or space [Jaffe et al., 2006] based on inter-tag frequencies, or visualizing them over areas of the World [Naaman et al., 2007]. However, no determination of the properties or semantics of specific tags has been provided [Rattenbury et al., 2007].

Regarding the use of other social networks, the potential of other similar location-based services (like Gowalla[2.2], Foursquare[2.3] and Facebook Places[2.4]) has already been demonstrated in recent work and it is being increasingly exploited as the dimensions of such services grow. Cheng et al. [Cheng et al., 2011] provide an assessment of human mobility patterns by analyzing the spatial, temporal, social, and textual aspects associated with the hundreds of millions of user-driven footprints (i.e. “check-ins”) that people leave with these services. Anastasios et al. [Noulas et al., 2011] provide a similar study but they also analyze activity and place transitions. Both of these studies are very interesting and motivating for a further exploitation of this kind of services. For example, in [Berjani and Strufe, 2011] the authors exploit the use of Gowalla to develop a Recommender System for places in location-based Online Social Network services (OSN) based on the check-ins of the entire user base.

In the Web-a-Where project, Amitay et al. [Amitay et al., 2004] associate web pages with geographical locations to which they are related, also identifying the main

[2.2] http://www.gowalla.com

[2.3] https://foursquare.com/

[2.4] http://www.facebook.com/places/

"geographical focus". The "tag enrichment" process consists of finding words (normally Named Entities) that show potential for geo-referencing, and then applying a disambiguation taxonomy (e.g. "MA" with "Massachusetts" or "Haifa" with "Haifa/Israel/Asia"). The results are very convincing, but the authors do not explore the other idea beyond using explicit geographical references. An extension could be to detect and associate patterns such as those referred to above in [Rattenbury et al., 2007] without the need for explicit location referencing.

Our work focuses on the semantic aspect of location representation. Furthermore, we also take advantage of information available on the Web about public places. With the rapid growth of the World Wide Web, a continuously increasing number of commercial and non-commercial entities are acquiring a presence online, whether through the deployment of proper web sites or by referral by related institutions. This presents an opportunity for identifying the information which describes how different people and communities relate to places, and thereby enrich the representation of POIs. Nowadays, this information found on the Web is rarely structured or tagged with semantic meaning. Indeed, it is widely known that the majority of online information contains unrestricted user-written text. Hence, we become dependent primarily on Information Extraction (IE) techniques for collecting and composing information from textual descriptions, as described in section 2.4.

2.2 Natural Language Processing

Natural Language Processing (NLP) is an AI topic which has gained plenty of attention from other research communities since, basically, anything can be described by free text. Processing these textual descriptions means not only editing and presenting them to humans, but also selecting relevant pieces and representing them in a more structured way so as to be understandable by machines.

One approach to NLP, sometimes referred to as *'symbolic'*, consists of rules for the manipulation of symbols. It usually works top-down by imposing known grammatical patterns and meaning associations upon texts. A second approach is rooted in the statistical analysis of language and is characterized as *'empirical'*. Differently from the former, it works bottom-up from the texts themselves, looking for patterns and associations, normally assigning probabilities when dealing with ambiguity [Jackson and Moulinier, 2007]. With great emphasis on the second approach, some techniques for extracting relevant information are introduced below.

There are NLP frameworks available to pre-process texts in a pipeline fashion. GATE [Cunningham et al., 2002] and UIMA [Ferrucci and Lally, 2004] are examples of such frameworks. Many NLP libraries are also freely available for download, under the OpenNLP project [The Apache Software Foundation, 2010] or from the creators themselves. In the following subsections, some of these frameworks and tools will be introduced. In order to illustrate the results obtained by each step, let us trace the natural language processing of a given text.

2.2.1 An Illustrative Example

Before presenting some resources that can be used for information extraction, the following text describing a given public place will serve as an illustrative example of raw material over which NLP tools can be applied throughout this section [2.5]:

> *The White House is the official residence and principal workplace of the President of the United States. Located at 1600 Pennsylvania Avenue NW in Washington, D.C., the house was designed by Irish-born James Hoban, and built between 1792 and 1800 of white-painted Aquia sandstone in the Neoclassical style. It has been the residence of every U.S. President since John Adams. When Thomas Jefferson moved into the house in 1801, he (with architect Benjamin Henry Latrobe) expanded the building outward, creating two colonnades that were meant to conceal stables and storage.*
>
> *In 1814, during the War of 1812, the mansion was set ablaze by the British Army in the Burning of Washington, destroying the interior and charring much of the exterior. Reconstruction began almost immediately,*

[2.5] Summary of White House Wikipedia article available at http://en.wikipedia.org/wiki/White_House *and President James Monroe moved into the partially reconstructed house in October 1817. Construction continued with the addition of the South Portico in 1824 and the North in 1829. Because of crowding within the executive mansion itself, President Theodore Roosevelt had all work offices relocated to the newly constructed West Wing in 1901. Eight years later, President William Howard Taft expanded the West Wing and created the first Oval Office which was eventually moved as the section was expanded. The third-floor attic was converted to living quarters in 1927 by augmenting the existing hip roof with long shed dormers. A newly constructed East Wing was used as a reception area for social events; Jefferson's colonnades connected the new wings. East Wing alterations were completed in 1946, creating additional office space. By 1948, the house's load-bearing exterior walls and internal wood beams were found to be close to failure. Under Harry S. Truman, the interior rooms were completely dismantled and a new internal load-bearing steel frame constructed inside the walls. Once this work was completed, the interior rooms were rebuilt.*

Today, the White House Complex includes the Executive Residence, West Wing, Cabinet Room, Roosevelt Room, East Wing, and the Old Executive Office Building, which houses the executive offices of the President and Vice President.

The White House is made up of six stories - the Ground Floor, State Floor, Second Floor, and Third Floor, as well as a two-story basement. The term White House is regularly used as a metonymy for the Executive Office of the President of the United States and for the president's administration and advisers in general. The property is owned by the National Park Service and is part of the President's Park. In 2007, it was ranked second on the American Institute of Architects list of "America's Favorite Architecture".

2.2.2 Part-of-Speech Tagging

Words in a sentence are tagged by *Part-of-Speech* (PoS) *taggers* which label each word with a grammatical category coming from a fixed set. The set of tags includes conventional parts of speech such as noun, verb, adjective, adverb, article, conjunction, and pronoun, and their subtypes. Examples of well-known tag sets are the Brown tag set which has 179 total tags, and the Penn Treebank tag set that has 45 tags [Manning and Schütze, 1999]. Table 2.1 presents some morphological classes and respective labels/tags [Francis and Kucera, 1983].

There are two main approaches to PoS tagging: *rule-based* and *stochastic*. A rule-based tagger tries to apply some linguistic knowledge to rule out sequences of tags that are syntactically incorrect. This can be in the form of contextual rules such as: *If an unknown term is preceded by a determiner and followed by a noun, then label it as an adjective.*

On the other hand, a stochastic tagger always relies on training data. The simplest implementation disambiguates word tags based solely on the probability that that word occurs with a particular tag. This probability is typically computed from a training set in which words and tags have already been matched by hand.

There are also PoS taggers that combine the main advantages of both types, like Brill's trainable rule-based part-of-speech tagger [Brill, 1994], which benefits from training procedures on tagged corpora and captures the learned knowledge in a set of simple deterministic rules. Brill reports an overall precision of 96.5% with this algorithm.

For the purpose of seeing PoS tagging effectiveness, tags can be attached to each term in our previous example, using Brill's tagger in the following manner. (Here only a few sentences are shown, for practical reasons.)

> The/DT White/NNP House/NNP is/VBZ the/DT official/JJ residence /NN and/CC principal/JJ workplace/NN of/IN the/DT President/NNP of/IN the/DT United/NNP States/ NNPS ./. Located/VBN at/IN 1600/ CD Pennsylvania/NNP Avenue/NNP NW/NNP in/IN Washington/NNP

Table 2.1: Most common tags used in PoS tagging [Francis and Kucera, 1983].

Tag	Description
.	sentence closer (. ; ? *)
(	left paren
)	right paren
,	comma
:	colon
CC	coordinating conjunction (and, or)
CD	cardinal numeral (one, two, 2, etc.)
DT	singular determiner/quantifier (this, that)
FW	foreign word (hyphenated before regular tag)
IN	preposition
JJ	adjective
NN	singular or mass noun
NNP	singular proper noun
NNS	plural noun
PRP	personal pronoun
PRP$	possessive pronoun
RB	adverb
TO	infinitive marker to
VBD	verb, past tense
VBG	verb, present participle/gerund
VBN	verb, past participle
VBP	verb, non-3$^{\text{rd}}$ singular present
VBZ	verb, 3$^{\text{rd}}$ singular present
WRB	Wh adverb (how, where)

,/, D./NNP C./NNP ,/, the/DT house/NN was/VBD designed/VBN by/IN I rish-born/NNP James/NNP Hoban/NNP ,/, and/CC built/VBD between/ IN 1792/CD and/CC 1800/CD of/IN white-painted/NNP Aquia/NNP sandstone/NN in/IN the/DT Neoclassical/NNP style/NN ./.

2.2.3 Noun Phrase Chunkers

Noun Phrase (NP) Chunkers are typically partial (sometimes called *shallow*[Abney and Abney, 1991]) parsers and take us beyond part-of-speech tagging to the extraction of clusters of words that represent people or objects. In English, they tend to concentrate on identifying *base* noun phrases, which consist of a *head* noun, i.e. the main noun in the phrase, and its *left modifiers*, i.e, determiners and adjectives occurring just to the left of it. They are less likely to identify prepositional phrases and resolve their attachments, as would be required by “the man in the park with the telescope”, where NP chunkers are usually not able to detect the prepositional phrase “with the telescope”.

Inspired by Brill’s tagger presented earlier, Marcus and Ramshaw presented an algorithm for noun phrase chunking using transformation-based learning [Ramshaw and Marcus, 1995]. Results can be scored based on the correct assignment of tags, or on recall and precision of complete base NPs. The latter is normally used as the metric, since it corresponds to the actual objective. Different tag sets can be used as an intermediate representation. Marcus and Ramshaw obtained about 92% recall and precision with their system for base NPs. They mention two major sources of error: participles and conjunctions.

Next, we present the Marcus and Ramshaw NPC’s result for the first paragraph of our example, where bracket-enclosed parts are those considered as NPs:

> [The White House] is [the official residence and principal workplace] of [the President] of [the United States]. Located at [1600 Pennsylvania Avenue NW] in [Washington], [D.C.], [the house] was designed by [Irish-born James Hoban], and built between [1792 and 1800] of [white-painted Aquia sandstone] in [the Neoclassical style]. [It] has been [the residence] of [every U.S. President] since [John Adams]. When [Thomas Jefferson] moved into [the house] in [1801], [he] (with [architect Benjamin Henry Latrobe]) expanded [the building] outward, creating [two colonnades] [that] were meant to conceal [stables and storage].

2.2.4 Named Entity Recognizers

Named Entity Recognition (NER) tries to identify proper names in documents and may also classify these proper names as to whether they designate people, places, companies, organizations, and so on. Unlike noun phrase extractors, many NER systems choose to disregard part-of-speech information and work directly with raw tokens and their properties (e.g. capitalization clues, adjacent words such as "Mr." or "Inc."). The ability to recognize previously unknown entities is an essential part of NER systems. Such ability hinges upon recognition and classification rules triggered by distinctive features associated with positive and negative examples.

The state-of-the-art method for NER is Conditional Random Fields (CRFs) [Lafferty et al., 2001]. CRFs provide a powerful and flexible mechanism for exploiting arbitrary feature sets along with dependency in the labels of neighboring words. The underlying idea is that of defining a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences. Another attractive aspect of CRFs is that one can implement efficient feature selection and feature induction algorithms for them. That is, rather than specifying in advance which features of (X, Y) to use, we could start from feature-generating rules and evaluate the benefit of generated features automatically on data. Rather than classifying each word independently as one of either Person, Location, Organization or Other, CRFs assume the named-entity labels of neighboring words are dependent; for example, while *New York* is a location, *New York Times* is an organization. Empirically, they have been found to be superior to all the earlier proposed methods for sequence labeling [Finkel et al., 2005].

It might be supposed that this task could be simplified by using lists of people, places and companies, but this is not the case [Jackson and Moulinier, 2007]. New companies, products, etc. come into being on a daily basis, and just using a directory or gazetteer does not necessarily help you decide whether "Philip Morris" refers to a person or a company. To structure this difficulty, Manning presents five types of typical errors produced by NER systems [2.6]:

[2.6] The type of errors are inspired by an informal publication by Christopher Manning - http://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html

- The system hypothesized an entity where there is none.
- An entity was completely missed by the system.
- The system noticed an entity but gave it the wrong label.
- The system noticed there is an entity but got its boundaries wrong.
- The system gave the wrong label to the entity and got its boundary wrong.

Using the first paragraph of our illustrative example, the following presents the named entities that could be found by the NER tool developed and made available by Stanford NLP Group [Finkel et al., 2005]:

> The White/ORGANIZATION House/ORGANIZATION is the official residence and principal workplace of the President of the United/LOCATION States/LOCATION. Located at 1600/LOCATION Pennsylvania/LOCATION Avenue/LOCATION NW/LOCATION in Washington/LOCATION, D.C./LOCATION, the house was designed by Irish-born James/PERSON Hoban/PERSON, and built between 1792 and 1800 of white-painted Aquia/LOCATION sandstone in the Neoclassical style. It has been the residence of every U.S./LOCATION President since John/PERSON Adams/ PERSON. When Thomas/PERSON Jefferson/PERSON moved into the house in 1801, he (with architect Benjamin/PERSON Henry/PERSON Latrobe/PERSON) expanded the building outward, creating two colonnades that were meant to conceal stables and storage.

[Finkel et al., 2005] show that the NER system outperforms the baseline with greater than 95% confidence, using the standard *t-test* for the CoNLL'03 and CMU Seminar Announcements respectively, thereby demonstrating the stability of their method.

2.2.5 String Metrics

String Metrics (also known as similarity metrics) are a class of metrics that measure similarity or dissimilarity (distance) between two text strings. The task of matching entity names has been explored by a number of communities, including those researching statistics, databases, and artificial intelligence. Each community has formulated the problem differently, and different techniques have been proposed. Interested readers may explore different works covered in some of the available surveys [Christen, 2006; Cohen et al., 2003; Winkler et al., 2006; Yancey and Yancey, 2005]. From these surveys, we can conclude the most widely known string metric is a simple one called the Levenshtein Distance (also known as Edit Distance) [Levenshtein, 1966]. String Edit Distance is a metric that can be used to determine how close two strings are to each other. This value is obtained in terms of the number of character insertions and deletions needed to convert one into the other. For example the string edit distance between “sitten” and “sitting” is 2 because:

1. sitten → sittin (substitution of ’i’ for ’e’)

2. sittin → sitting (insertion of ’g’ at the end).

The distance can be converted into a similarity measure (between 0.0 and 1.0) using the equation 2.1.

$$sim_{ed}(s,t) = 1.0 - \frac{dist_{ed}(s,t)}{max(|s|,|t|)} \tag{2.1}$$

with $dist_{ed}(s,t)$ being the actual edit distance function which returns a value of 0 if the strings are the same or a positive number of edits if they are different. The edit distance is symmetrical and it always holds that $0 \leq dist_{ed}(s,t) \leq max(|s|,|t|)$, and $abs(|s| - |t|) \leq dist_{ed}(s,t)$.

A broadly similar metric, which is not based on an edit-distance model, is the *Jaro* metric [Jaro, 1989] which is based on the number and order of common characters between two strings. Given strings $s = a_1 \cdots a_K$ and $t = b_1 \cdots b_L$, one can define a character a_i in s to be in common with t if and only if there is a $b_j = a_i$ in t such that $i - H \leq j \leq i + H$, where $H = min(|s|,|t|)/2$. Let $s' = a'_i \cdots a'_K$ be the characters in s that are common with t (in the same order they appear in s), and let $t' = b'_1 \cdots b_L$, be analogous. Then define a transposition for s', t' to be a position i such that $a_i = b_i$. Let $T_{s',t'}$ be one-half the number of transpositions for s' and t'. The Jaro metric for s and t is defined in equation 2.2.

$$sim_{jaro}(s,t) = \frac{1}{3}\left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{s'}\right) \tag{2.2}$$

	W	I	L	L	I	A	M
W	$\underline{\mathbf{1}}$	**0**	**0**	0	0	0	0
I	**0**	$\underline{\mathbf{1}}$	**0**	**0**	1	0	0
L	**0**	**0**	$\underline{\mathbf{1}}$	**1**	**0**	0	0
L	0	**0**	**1**	$\underline{\mathbf{1}}$	**0**	**0**	0
L	0	0	**1**	**1**	$\underline{\mathbf{0}}$	**0**	**0**
A	0	0	0	**0**	**0**	$\underline{\mathbf{1}}$	**0**
I	0	1	0	0	**1**	**0**	$\underline{\mathbf{0}}$
M	0	0	0	0	0	**0**	**1**

Table 2.2: The Jaro metric. The underlined entries are the main diagonal, and each bold character is in common with the string "WILLIAM" ("WILLLAIM")[Bilenko et al., 2003].

In order to better understand the intuition behind this metric, consider the matrix M in table 2.2, which compares the strings $s =$"WILLLAIM" and $t =$"WILLIAM". The boxed entries are the main diagonal, and $M(i, j) = 1$ if and only if the ith character of s equals the jth character of t. The Jaro metric is based on the number of characters in s that are in common with t. In terms of the matrix M in the table 2.2, the ith character of s is in common with t if $M_{i,j} = 1$ for some entry (i, j) that is "sufficiently close" to M's main diagonal, where sufficiently close means that $|i-j| < min(|s|, |t|)/2$ (shown in the matrix in bold).

Good results for name-matching tasks [Cohen et al., 2003] have been reported using a variant of the Jaro metric proposed by [Winkler, 1990]. The Winkler algorithm improves upon the Jaro algorithm by applying ideas based on empirical studies which found that fewer errors typically occur at the beginning of names. The Winkler algorithm therefore increases the Jaro similarity measure for agreeing initial characters (up to four), using a prefix scale p which gives more favorable ratings to strings that match from the beginning for a set prefix length ℓ . Given two strings s and t, their Winkler similarity $sim_{winkler}$ is stated in equation 2.3.

$$sim_{winkler}(s,t) = sim_{jaro}(s,t) + \ell \times p(1 - sim_{jaro}(s,t)) \tag{2.3}$$

where:

String Metric	Computed Similarity	Assumptions
Edit	$sim_{ed}(s,t) = 1.0 - \frac{2}{7} = 0.71$	$\lvert s\rvert = 6$, $\lvert t\rvert = 7$
Jaro	$sim_{jaro}(s,t) = \frac{1}{3}\left(\frac{5}{6} + \frac{5}{7} + \frac{5-0}{5}\right) = 0.85$	$\lvert s'\rvert = 5$, $\lvert t'\rvert = 5$ no transposition needed
Winkler	$sim_{winkler}(s,t) = 0.85 + 4 \times 0.1 \times (1 - 0.85) = 0.91$	$\ell = 4$, $p = 0.1$

Table 2.3: Different string metrics applied to the same example: the similarity between the strings “sitten” and “sitting”. These metrics are normalized to result in a similarity between 0 (very dissimilar) to 1 (equal).

- sim_{jaro} is the Jaro similarity for strings s and t;
- ℓ is the length of common prefix at the start of the string up to a maximum of 4 characters;
- p is a constant scaling factor for how much the score is adjusted upwards for having common prefixes. The standard value for this constant in Winkler's work is $p = 0.1$.

Regarding the same example used above, this time the similarity between s “sitten” and t “sitting” across the three metrics here presented can be observed in table 2.3.

An implemented open-source, Java toolkit of string metrics that includes a variety of different techniques named SecondString is available online [Cohen and Ravikumar, 2003]. Later, this library will be used on two distinct modules of our system, firstly to match the similarity between POIs (see section 4.1.4) and then later in another task of comparing entity names in the Meaning Extraction module (see section 5.3).

2.3 Information Retrieval

As Information Retrieval (IR) is a field of study that helps users find needed information from a large collection of text documents or web pages, it has became so popular that people make fewer trips to libraries, but more searches on the Web [Liu, 2009].

Formally speaking, IR is defined as finding materials (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [Manning et al., 2008].

In IR systems, vast collections of documents are pre-processed and *indexed* in order that a given query can be answered efficiently. This query is matched against indexed documents using a similarity metric. Documents are generally Web Pages and the act of querying an IR system is commonly known as Web Search. In the following sections, the phases: Indexing, Search and Matching Similarity, are analyzed in detail, and specific terminology in the field is introduced.

2.3.1 Indexing Documents

Traditional IR assumes that the basic information unit is a *document* and that a large collection of documents is available to form the text database. On the Web the documents are *web pages*. A document is split into term units. This definition will be further extended below (see section 2.4.3.2). Besides the fact that some interesting techniques have been demonstrated in multi-word indexing [Evans and Zhai, 1996], the great majority of search engines use single words to structure both documents and queries.

Before the documents in a collection are used for retrieval, some pre-processing tasks are usually performed [Liu, 2009]. For traditional text documents (which have no HTML tags), the tasks are mainly stopword removal and stemming, while for web pages, additional tasks such as identification of main content blocks also require careful considerations:

- *Stopwords* are frequent and meaningless words in a language. They belong to closed morphological classes, such as articles, prepositions and conjunctions. Stopword removal is applied both in documents before indexing and storing, and in the query.
- *Stemming* refers to the process of reducing words to their roots. A **stem** is the portion of a word that is left after removing its prefixes and suffixes (e.g. “study”,

"students", and "studying" are reduced to "stud"). In English, the most popular stemming algorithm was introduced by Martin Porter [Porter, 1997].

- *Identification of main content blocks* in a web page requires a careful analysis of web page structure. Especially in focused directory sites (e.g. http://www.tripadvisor.com or http://www.urbanspoon.com), the main block content is the only distinct information seen when we visit different web pages within the same site. Banner ads, lateral information and menus are obviously noisy information to avoid when trying to extract relevant terms to index. Considering the fact that web pages of a given Web site use the same template, research has been done to identify this underlying structure in order to extract what differs from page to page (perhaps its main content). This problem has gained the attention of the IR and the Database community and is explored in detail in section 2.4.

Among many index schemes for text [Liu, 2009], the **inverted index** is a popular one which has been shown to be superior to most other forms of indexing. In its simplest form, the inverted index of a document collection is basically a data structure that attaches to each distinctive term a list of all documents that contains that term. Thus, in retrieval, it always takes the same time to find the documents that contains a query term.

2.3.2 Search

Given a user query, searching involves the following steps [Liu, 2009]: 1) pre-processing the query using stopword removal and stemming; 2) finding pages that contain all (or most of) the query terms in the inverted index; 3) ranking the pages and returning them to the user.

The ranking algorithm is the heart of a search engine. Beyond considering document content as a traditional IR system does, on the Web another important factor is taken into account: hyperlink structure. PageRank [Brin and Page, 1998] is the most well-known algorithm making use of the link structure of web pages to compute a quality or reputation score for each page. It is based on the idea that "a link from page x to page y means y is reputed by x" since the author of page x believes that page y contains quality. Tomlin [Tomlin, 2003] proposes a generalization of the PageRank algorithm that computes flow values for the edges of the Web graph, and a TrafficRank value for each page. This last ranking algorithm is the one used in the Yahoo! search engine.

Location-based web search (or *Local Search*) is one of the popular tasks expected from search engines. A location-based query consists of a topic and a reference location. Unlike general web search, in location-based search, a search engine is expected to find and rank documents which are not only related to the query topic but also geographically related to the location with which the query is associated. Besides the lack of geographical information associated with the Web resources, another issue is that, in general search engines, the rank score for each page is calculated globally while in location-based search, the web pages must be analyzed and evaluated locally. There are several issues concerning developing effective geographical search engines and, as yet, no global location-based search engine is reported to have achieved them [Asadi et al., 2009]. Amongst the most notable difficulties are location ambiguity, lack of geographical information on web pages, language-based and country-dependent variation in address styles, and multiple locations related to a single web resource.

Search engine companies have started to develop and offer location-based services. However, they are still geographically limited, mostly to the United States, such as Yahoo! Local, Google Maps and MSN Live Local, and have not become as successful and popular as general search engines. Also, generally speaking, the results presented are related to their business directories and not to Web documents. Despite this, a lot of work has been done in improving the capabilities of location-based search engines [Ahlers and Boll, 2007; Amitay et al., 2004], but this is beyond the scope of this thesis. Instead, this thesis makes use of generally available search engines and formulate queries using the geographical reference to retrieve information about places (section 4.2.1). The work in this context is more focused on contributing to the indexing capabilities of such engines in terms of local search (finding an inspiration in [Tanasescu and Domingue, 2007]) than on becoming any alternative form of search.

2.3.3 Matching Similarity

The vector derived from document d is denoted by $\overrightarrow{V}(d)$ with one component in the vector for each dictionary term. The set of documents in a collection may then be viewed as a set of vectors in a vector space, in which there is one axis for each term. This representation loses the relative ordering of the terms in each document but is the simplest one.

Given a query and documents represented by vectors of weighted terms[2.7] denoted by $\overrightarrow{V}(q)$ and $\overrightarrow{V}(d)$ respectively, documents are ranked according to a given similarity metric. Since query and document vectors are of different size (in general, queries are much smaller than documents), *Cosine* similarity [Salton and Buckley, 1988] (equation 2.4) is most often used:

$$score(q,d) = \frac{\overrightarrow{V}(q) \cdot \overrightarrow{V}(d)}{|\overrightarrow{V}(q)||\overrightarrow{V}(d)|} \tag{2.4}$$

where the numerator represents the *dot product* (also known as the *inner product*) of the vectors $\overrightarrow{V}(q)$ and $\overrightarrow{V}(d)$, and the denominator is the product of their *Euclidean lengths*. The dot product $\overrightarrow{x} \cdot \overrightarrow{y}$ of two vectors is defined as $\sum_{i=1}^{M} = x_i y_i$. Let $\overrightarrow{V}(d)$ denote the document vector for d, with M components $\overrightarrow{V}_1(d) \cdots \overrightarrow{V}_M(d)$. The Euclidean length of d is defined as $\sqrt{\sum_{i=1}^{M} \overrightarrow{V}_i^2(d)}$. The ratio is calculated to normalize the length of documents since long documents tend to have high term frequencies. Thus this metric compensates for the effect of difference in length of query and document, since the document may have a high cosine score for a query, even if it does not contain all the query terms [Manning et al., 2008].

In order to measure the performance of IR systems, *Precision* and *Recall* metrics are generally used. While Precision (P) is the fraction of retrieved documents that are relevant, Recall (R) is the fraction of relevant documents that are retrieved [Manning et al., 2008]. Precision can be seen as a measure of exactness or fidelity, whereas recall is a measure of completeness. With reference to the contingency table 2.4 that classifies each type of result, P and R are formally defined by equations 2.5 and 2.6 respectively.

[2.7] Term Weighting methods are discussed in section 2.4.3.3

	relevant	non-relevant
retrieved	true positives(tp)	false positives(fp)
not retrieved	false negatives(fn)	true negatives(tn)

Table 2.4: Classification of each type of result from IR systems

Precision is defined as the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class, while recall is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been). An alternative metric is *Accuracy*, which is the fraction of IR system classifications that are correct. In terms of the contingency table (table 2.4), Accuracy can be defined by equation 2.7, which is the proportion of true results (both true positives and true negatives) in the population. F-measure (also known as F_1 score) is the harmonic mean between precision and recall defined by equation 2.8. In most experiments, there is no particular reason to favor precision or recall, so most researchers use equal weight of precision and recall to compute F-measure.

$$Precision = \frac{tp}{(tp + fp)} \tag{2.5}$$

$$Recall = \frac{tp}{(tp + fn)} \tag{2.6}$$

$$Accuracy = \frac{tp + tn}{(tp + fp + fn + tn)} \tag{2.7}$$

$$F - measure = 2 \times \frac{recall \times precision}{recall + precision} \tag{2.8}$$

These metrics are first defined for the simple case where an IR system returns an unranked set of documents for a query. However, in a ranked retrieval context case what matters is rather how many good results there are on the first k retrieved documents. This is referred to as *precision at k*, for example "precision at 10" [Manning et al., 2008]. It has the advantage of not requiring any estimate of the size of the set of relevant documents; the disadvantage is that it is the least stable of the commonly used evaluation measures and that it does not average well because the total number of relevant documents for a query has a strong influence on its “precision at k”.

2.4 Information Extraction

Information Extraction (IE) refers to the automatic extraction of structured information such as entities, relationships between them, and attributes describing entities from unstructured sources [Sarawagi, 2008]. Traditional IE, the main concern of this thesis, aims to extract data from totally unstructured free texts that are written in natural language. In contrast, the task of Web IE is very different from traditional IE tasks, in that it processes online documents that are semi-structured and usually generated automatically by a server-side application program. As a result, traditional IE usually takes advantage of NLP techniques such as lexicons and grammars, whereas Web IE usually applies machine-learning and pattern-mining techniques to exploit the syntactical patterns or layout structures of the template-based documents [Kayed and Shaalan, 2006].

With regard to textual information, IE is a task much linked to Text Mining, both being sub-topics of the wider area of IR. IE applies classic NLP techniques and resources over unstructured pages written in natural language. In contrast, Text Mining usually applies machine learning and pattern mining to exploit the syntactical patterns or layout structures of the template-based documents. Since it is impossible to guarantee that public places are only represented in structured pages from directory sites (e.g. from directory sites such as http://www.tripadvisor.com or http://www.urbanspoon.com), we will eventually need to apply NLP techniques to extract relevant information. In fact, some important places may have their own pages, which impedes the template-based learning.

Broadly speaking, the role of IE is to obtain meaningful knowledge out of large quantities of unstructured data. IE systems can be classified according to the following features [Sarawagi, 2008]:

1. the type of structure extracted (entities, relationships, lists, tables, attributes, etc.).
2. the type of unstructured source (short strings or documents, templatized or open-ended).
3. the type of input resources available for extraction (structured databases, labeled unstructured data, linguistic tags, etc.).
4. the method used for extraction (rule-based or statistical, manually coded or trained from examples).
5. the output of the extraction (an annotated unstructured text or a database).

Besides the fact that a lot of work has been done in extracting different types of structures from information sources, such as relationships [Etzioni et al., 2008; Shinyama and Sekine, 2006; Zelenko et al., 2002] or opinion/attributes [Hu and Liu, 2004], the projects studied here extract the most basic structure: terms. This section focuses on introducing and presenting the main features of IE systems. For a thorough comparison of works not covered here, the reader is redirected to a large survey described in [Sarawagi, 2008]. In the following subsections, sets of IE systems are grouped according to the features described above.

2.4.1 Source of Information

Sources of information can be classified by their granularity (1) and heterogeneity(2) across different documents. *Granularity* is associated with the size of analyzed text. Most popular sources of information are small text snippets [Sarawagi, 2008] such as records or sentences, but only those systems that deal with entire paragraphs and documents are described in detail, since it is necessary to consider the context of multiple sentences or an entire document for meaningful extractions.

Typically, for extracting information from longer units the main challenge is to design efficient techniques for filtering only the relevant portion of a long document (in texts) or the main content (in web pages).

In the case of a web page, its structure and layout varies depending on the different content type it will represent or the tastes of the designer styling its content. The position of the main content or the tag inside the HTML structure which refers to the main content differs in a variety of websites [Rahmani et al., 2010]. Modern web documents contain far more data than their main content. Navigation menus, advertisements, functional or design elements are typical examples of additional contents which extend, enrich or simply accompany the main content [Gottron, 2008].

The process of determining the main content of an HTML document is commonly referred to as Content Extraction (CE) and was first introduced by Rahman et al. in [Rahman et al., 2001]. The algorithm developed by them considers all detailed pages of a website as pages of the same class. It runs a learning phase with two or more pages as its input, finds the blocks that their pattern repeats between input pages and marks them as non-informative blocks, then saving them in storage. These non-informative blocks are mostly copyright information, the header, the footer, sidebars and navigation links. Then, when the CE algorithm is used in the real world, it first eliminates non-informative patterns from the structure of its input pages based on the stored patterns in its storage for a specific class of input pages. Finally, from the remaining blocks in the page, it will return the block of text containing the largest text length. CE needs a learning phase because it cannot extract the main content randomly from just one input web page.

The *heterogeneity* dimension has an impact on the complexity and accuracy of an extractor. The more homogeneous the source is, the more precise and simple the extraction process becomes. In decreasing order of homogeneity, information can arise from *machine-generated pages*, *partially structured domain specific sources* or *open-ended sources* [Sarawagi, 2008].

The extractors for highly templatized *machine-generated pages* are popularly known as wrappers [Crescenzi et al., 2001; Yi et al., 2003]. Wrapper Induction (WI) is a kind of software tool that is designed to generate wrappers. A wrapper usually performs a pattern-matching procedure which relies on a set of extraction rules. Tailoring a WI system to a new requirement is a task that varies in scale depending on the text type, domain, and scenario. To maximize reusability and minimize maintenance cost, designing a trainable WI system has been an important topic in the research fields of message understanding, machine learning, data mining, etc. These tools are a great help in populating databases from web pages, but they also require a period of time (which increases dramatically as the number of pages grows) for learning each new template from a different domain.

Partially structured domain-specific sources are the most studied setting for IE. Hence, the input source is from within a well-defined range and there is at least an informal style that is roughly followed, for example news articles [Doddington et al., 2004; Grishman and Sundheim, 1996; Marsh and Perzanowski, 1998; Tjong Kim Sang and De Meulder, 2003], or classified ads [Soderland, 1999], or citations [Borkar et al., 2001], or resumés. Thus it is possible to develop an adequate extraction model given enough labeled data, but there is a much greater variety between one input and another than in machine-generated pages.

Over recent years, there has been increased interest in extracting entities and relationships between entities from *open-ended sources*, such as the web, where there is little that can be expected in terms of homogeneity or consistency [Etzioni et al., 2008; Shinyama and Sekine, 2006; Urbansky et al., 2010; Wang and Cohen, 2007]. In such situations, one important task is to exploit the redundancy of the extracted information across many different sources. For instance, previous approaches have assumed that they are dealing with small, domain-specific corpora and limited sets of relations. The use of NERs as well as syntactic or dependency parsers is a common thread that unifies most previous work, but this rather heavy linguistic technology runs into problems when applied to the heterogeneous text found on the Web. While the parsers work well when trained and applied to a particular genre of text, such as financial news data in the Penn Treebank, they make many more parsing errors when confronted with the diversity of Web text [Etzioni et al., 2008].

2.4.2 Method

There is a range of approaches to building IE systems. One approach is to manually develop IE rules by encoding patterns (e.g. regular expressions) that reliably identify the desired entities or relations. However, due to the variety of forms and contexts in which the desired information can appear, manually developing patterns is very difficult and tedious and rarely results in robust systems. Consequently, machine-learning has become the most successful approach to developing robust IE systems [Cardie, 1997]. Among machine-learning methods used in IE, there is a clear distinction between rule-based and statistical methods. Rule-based extraction methods are driven by hard predicates, whereas statistical methods learn from labeled examples. Rule-based methods are easier to interpret and develop, whereas statistical methods are more robust in relation to noise in the unstructured data. Therefore, rule-based systems are more useful in closed domains where human involvement is both essential and available. In open-ended domains the soft logic of statistical methods is more appropriate [Sarawagi, 2008].

Another distinction in machine-learning methods is whether they are supervised or unsupervised. For both types, it is necessary to possess an understanding of machine learning to be able to choose between various model alternatives and to define features that will be robust in relation to unseen data [Sarawagi, 2008].

The supervised approach [Turney, 2000] regards information extraction as a classification task. In this approach, a model is trained to determine whether a candidate term of the document is a keyphrase, based on statistical and linguistic features. A document set with human-assigned keyphrases is required as a training set. Thus, even in the learning-based systems, domain expertise in identifying and labeling examples that will be representative of the actual deployment setting is needed.

As an example of an unsupervised approach, the graph-based ranking system TextRank [Mihalcea and Tarau, 2004] regards information extraction as a ranking task, where a document is represented by a term graph, based on term relatedness, and then a graph-based ranking algorithm (e.g. PageRank [Brin and Page, 1998]) is used to assign importance scores to each term. Each vertex on a term graph is a (single or multi-word) term on the text and existing methods usually use term co-occurrences within a specified window size in the given document as an approximation of term relatedness [Mihalcea and Tarau, 2004].

Due to the specificity of some IE tasks, supervised machine-learning methods trained on human-annotated corpora are becoming the most successful approach to developing robust and domain-specific IE systems [Cardie, 1997]. The state of the art in this area is currently represented, on the one side, by supervised learning methods, where a system is trained to recognize keywords in a text, based on lexical and syntactic features; and on the other side, by unsupervised learning methods, where graph-based ranking methods are becoming the most widely used unsupervised approach for keyphrase extraction.

In the first class of systems, the most prominent is KEA [Wu et al., 2005]. In KEA, the candidate terms are represented using three features: TF-IDF[2.8], distance (the number of words that precede the first occurrence of the term, divided by the number of words in the document) and keyphrase frequency (the number of times a candidate term occurs as a key term in the training documents). The classifier is trained using the naive Bayes learning algorithm. Thus, KEA depends on the training set and may provide poor results when the training set does not fit well with the processed documents. This approach is domain-specific, and therefore it largely fails in open domain extraction, not being able to deal with documents of different categories.

In unsupervised learning, Litvak and and Last [Litvak and Last, 2008] applied to TextRank system [Mihalcea and Tarau, 2004] the HITS ranking algorithm instead of PageRank since the former is more appropriated for directed graphs. Other methods improve upon the graph-based method presented in the TextRank by using clustering techniques on term graphs for keyphrase extraction [Grineva et al., 2009; Liu et al., 2009].

The Yahoo! Term Extraction API [2.9] looks for the appearance of popular search terms in a webpage when extracting keywords. This tool is integrated in a broader IE/IR system named Y!Q (Yahoo Contextual Search). Yahoo! Term Extraction is a

[2.8] TF-IDF means *term frequency - inverse document frequency*; see section 2.4.3.3

[2.9] http://developer.yahoo.com/search/content/V1/termExtraction.html context-based IE tool as described in [Kraft et al., 2005], which means it can generate results based on the passages of text that constitute a document. Each passage of text (or context) is represented by a *Context Term Vector*, a dense representation of a context that can be obtained using various text- or entity-recognition algorithms represented in the vector space model [Yu et al., 1982]. In this model, extracted terms are typically associated with weights, which represent the importance of a term within the context, and/or additional meta-data. The context term vector and its associated meta-data information (e.g. term weights, entities) are passed to other components in the Y!Q Contextual Search Engine for further processing, such as the Query Planning and Rewriting Framework (QPW) and Contextual Ranking (CR). However, for the focus of this thesis, we are interested in the first component: Content Analysis (CA).

The terms of a context term vector may represent (but are not limited to) a subset of the words/phrases/entities in the content of the context. Y!Q makes this context available in the form of a semantic network (derived from a large document corpus and query logs) provided by a service. This service, named Content Analysis (also known as Yahoo! Term Extraction), comprises three major components related to Information Extraction (see Figure 2.2):

1. The Vectorizer performs term extraction (e.g. using statistical and/or linguistic analysis methodologies [Maynard and Ananiadou, 1999]) given a context (plus optionally a query) to return its key concepts. As a result of the term extraction step, Y!Q generates a context term vector representation of the input. This digest is then be used to determine subsequent actions (e.g. what information sources to use, how to articulate contextual search queries, etc.).

2. The Entity Identifier is used for cross-referencing terms previously extracted. Then entity information is associated with the extracted terms as meta-data. The Yahoo entity dictionary is periodically reviewed by an editorial team to keep its content fresh and of high quality.

3. Topic Disambiguation is applied to terms since they can have multiple meanings (e.g. the term “jaguar” could represent a car, an animal, or an operating system). All terms in the term context vector are disambiguated using various term disambiguation techniques [Maynard and Ananiadou, 1998].

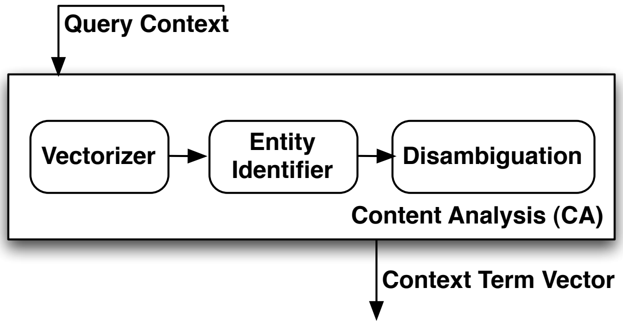


Figure 2.2: Yahoo Term Extraction overview [Kraft et al., 2005]

The Y!Q authors tested the performance of this tool compared with the standard Yahoo! Web Search (which has no context in query formulation and matching). They noted that users prefer the results given by search using context, since in some cases there were no answers in the traditional search. One example was a context about *apache helicopter* and the query was *apache cost*. Yahoo! Web Search returned irrelevant results which had nothing to do with helicopters. Again, apache is an ambiguous query (other meanings include *Apache Indians* and *Apache web server*, that in Y!Q the right one must be chosen by users). But it is assumed that context must be explicitly provided by users, being when additional text (context) is provided in a contextual search, or when a provider marks by known HTML tags the context of a given web page.

In order to discover which systems perform better, it is important to test them on the same data set. For this purpose, evaluation contests have been organized by the scientific community, as discussed later in section 2.4.4.

2.4.3 Output of IE

Another way to distinguish between different IE systems is which form of output is intended: a common web page template induced from web pages of the same Web site in order to retrieve k-tuples resembling database records about entities properties; or the main content of a web page, ignoring advertisements, menus and other navigational lateral information; or, finally, a list of relevant terms representing the content of a given document (web page or text), also known as Term Extraction (TE).

The main aim in TE is to determine whether a word or phrase is a term which characterizes the target domain. A general TE consists of two steps:

1. It makes use of various degrees of linguistic filtering (e.g. part-of-speech tagging, noun phrase chunking, etc.), through which candidates of various linguistic patterns are identified (e.g. noun-noun, adjective-noun-noun combinations, etc.).
2. The use of frequency- or statistical- based evidence measures to compute weights, indicating to what degree a candidate qualifies as a terminological unit.

Thus, in the following subsections we will cover in depth different approaches that are dedicated only to term extraction and subsequent statistical computation of their relevance.

2.4.3.1 Definition of Term

Before analyzing approaches to dealing with Term Extraction, it is important to define in an unambiguous way the terminology used in the field. A term is a linguistic representation of a concept. A concept is a mental representation of an object in a given context. The same concept can be identified by different terms. While some work differentiates a simple term (one word) from a complex term (two or more words) [Kageura and Umino, 1996; Wong et al., 2008], in this thesis a term means an expression (usually a noun phrase) composed of one or more words. A keyphrase (or keyword or keyterm) is defined as a meaningful and significant term that in a given set can serve as a highly condensed summary for a document. Keyphrases can also be used as a label for the document, to supplement or replace the title or summary, or they can be highlighted within the body of the document to facilitate the user's fast browsing and reading [Wan and Xiao, 2008].

2.4.3.2 Automatic Term Extraction

In order to better understand and organize the work produced in the field of Automatic Term Extraction (or Automatic Term Recognition - ATR), it can be useful to identify two mainstream approaches to the problem. On the one hand, statistical measures have

Linguistic Filter	Author(s)
$(N)^+N$	[Dagan and Church, 1994]
$(N\|A)^+N$	[Frantzi et al., 2000]
$((A\|N)^+\|(A\|N)(NP)?(A\|N))N$	[Justeson and Katz, 1995]
$((A\|N\|\#)^+\|(A\|N\|\#)^*(NP)?(A\|N\|\#)^*)N$	[Korkontzelos et al., 2008]

Table 2.5: Different linguistic filters proposed in literature. A means *Adjective*, N *Noun*, P *Preposition* and # *Numerals*. The symbols used in the above expression mean respectively: + one or more, $*$ zero or more, ? optional, | mathematical or

been proposed to define the degree of relevance of candidate terms, i.e. to find appropriate measures that can help in selecting good terms from a list of candidates. On the other hand, computational terminologists have tried to define, identify and recognize terms by looking at pure linguistic properties, using linguistic filtering techniques which aim to identify specific syntactic patterns of terms. Finally, hybrid approaches try to use these two views together, taking into account both linguistic and statistical hints to recognize terms [Pazienza et al., 2005].

Linguistic filters are generally made up of PoS taggers, stopword lists and parsing rules to select term candidates. These rules are formed relying on the analysis produced by a shallow syntactic parser, and are chosen in an empirical way by looking at experimental data. Different linguistic filters for English have been proposed by different authors, as can be seen in table 2.5. In increasing order of complexity as we go down the list, filters can range from *open* (which let a lot of candidate terms through) to *closed* filters (which are more restrictive). Accordingly, it is important to note that the more the filter is open in length, the greater its coverage.

To justify which direction to follow, open or closed filters, Daille et al. [Daille et al., 1996] confirmed, through the analysis of manually produced terminological data banks, that terms generally appear in the form of short noun phrases, mostly composed of only two main items such as nouns, adjectives and adverbs. These core terms, consisting of one or two main items, are called base-terms. However, the more specific and complete a term is, the more information and meaning it can represent. Hence, a compromise needs to be found in order either to present a great coverage of short and simple terms or to extract specific and long terms.

2.4.3.3 Term Weighting

In order to distinguish the relevance of a term, Kageura and Umino [Kageura and Umino, 1996] proposed an important feature of domain-specific terms called *Termhood*, which refers to the degree to which a term is related to a domain-specific concept. In contrast to this, they also defined the concept of *Unithood*, i.e. the degree of strength or stability of syntagmatic combinations or collocations. In other words, Unithood is concerned with whether or not sequences of words should be combined to form more stable lexical units, while Termhood is the extent to which these stable lexical units are relevant to some domains. While the former is only relevant to complex terms, the latter is concerned with both simple terms and complex terms.

For example, in a music events corpus, "distorted electric guitar" is a valid term, which has both high termhood and unithood. However, its frequently occurring substring "distorted electric", has high unithood and low termhood, since it does not refer to a key domain concept. In contrast, in "hard rock music", both substrings "hard rock" and "rock music" have high unithood and termhood. Thus, it is also important to classify ATR systems as domain-specific. It is also important to stipulate whether the intention is to compute: *termhood* or *unithood*, since the first is a metric that is more appropriate to domains, while the second is related to the strength of a complex term independently of the domain.

Ranking candidate terms based on a particular criterion is the purpose of *term weighting* methods. Candidate terms above a given threshold are selected for further processing, for example, to be used to compute similarity between documents. Term weighting is mainly applied in hybrid ATR systems or to provide feature values for machine-learning methods. Although numerous term-weighting methods have been proposed in the literature, only the most popular and standard are presented here. First, three *termhood* measures are presented, before we come to a recent trend in term weighting that also incorporates *unithood* in its computation at the end of this section.

The simplest approach, *Term Frequency*, consists of computing the number of occurrences of a term t in a document d and is denoted $tf_{t,d}$. This value can be normalized if the total number of candidate terms $|T|$ is considered. In short, the equation of term frequency is defined in 2.9.

$$tf_{t,d} = \frac{n_{t,d}}{|T_d|} \tag{2.9}$$

However, as terms are not equally important in relevance, a metric that can compute the discriminative power of a term in a given document in the collection is referred to as *Inverse Document Frequency* (equation 2.10). It considers the numbers of document where a given term appears (the *document frequency* - df_t) in relation to the total number of documents in the collection $|N|$. The inverse document frequency (idf) of a rare term is high, while that of a frequent term is likely to be low.

$$idf_t = log\frac{N}{|df_t|} \tag{2.10}$$

TF-IDF (equation 2.11) is a standard statistical method that combines the frequency of a term in a particular document with its inverse document frequency in general use [Salton and Buckley, 1988]. This score is high for rare terms that appear frequently in a document and are therefore more likely to be significant. In a pragmatic view, $tf\text{-}idf_{t,d}$ assigns to term t a weight in document d that is: highest when t occurs many times within a small number of documents; lower when the term occurs fewer times in a document, or occurs in many documents; lowest when the term occurs in virtually all documents [Manning et al., 2008]. While in IR, the TF-IDF is primarily used to rank documents, it can also be used to rank words and word sequences of a document as term candidates for (the domain of) the document. Arguably, a high frequency and a high degree of concentration of a term in a given document speaks in favor of its being document-specific.

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t \tag{2.11}$$

Frantzi [Frantzi, 1997] proposed a measure known as *C-Value* for extracting complex terms. Considered as a unithood measure to identify recurrent multi-word terms, the measure is based upon the claim that a substring of a term candidate is a candidate itself, given that it demonstrates adequate independence from the longer version it appears in. For example, in “real time image generation”, “real time” and “image generation” are acceptable as valid complex candidate terms. However, “time image generation” is not. The measure is built using statistical characteristics of the candidate term. These are:

1. the total frequency of occurrence of the candidate term in the corpus ($f(t)$)
2. the frequency of the candidate term as part of other longer candidate terms ($f_{Nested}(z)$)
3. the number of these longer candidate terms ($|Nested|$)
4. the length of the candidate term (in number of words) ($|t|$)

The final formula is presented as 2.12. The idea is to subtract from the candidate-specific score (based on frequency and unit length) the average frequency of longer candidates of which the given candidate is part. The C-value is only applied to multi-word strings that have passed a linguistic filter. As described earlier, a linguistic filter is based on part-of-speech tags and a stop-list; it can be varied so as to balance precision against recall. Frantzi observed that the C-value would be high for long non-nested strings that have a high absolute frequency in the studied corpus. On the other hand, non-maximal candidates that are part of longer candidates with high frequencies would get a low C-value.

$$Cvalue_t = \begin{cases} log_2|t| \times f(t) & \text{if } t \text{ is not nested;} \\ log_2|t| \times (f(t) - \frac{1}{|Nested|} \times \sum_{z\epsilon Nested} f(z)) & \text{otherwise} \end{cases} \tag{2.12}$$

2.4.4 Evaluation Contests

Evaluations of IE systems are critical to the scientific progress of this field, and they have been performed at conferences or contests set up by government agencies, sometimes acting in coordination with contractors or academics. As a valuable outcome of these events, tagged data sets are made public in order to automatically compute the performance of each participating system. A large proportion of these data is still used after the contest period, known as *Golden Sets*, and these are a great help in recreating past scenarios and improving current algorithms. Table 2.6 gives an overview of different contests that have taken place in the past. In some cases, they occur in cycles over

Event Title	Acronym	Language	Year	Tracks of Interest
Evaluation Exercises on Semantic Evaluation [SemEval Portal, 2011]	SemEval	English	1998-present (3-year cycle)	Automated Keyphrase Extraction, Word Sense Disambiguation
TAC Knowledge Base Population Evaluation [tac, 2010]	TAC/ KBP	English	2009-present	NER, Wikipedia Info-box Population
Automatic Content Extraction Program [Doddington et al., 2004]	ACE	English	2000-2008	NER
Evaluation contest for NER in Portuguese [Santos et al., 2008]	HAREM	Portuguese	2004-2008	NER
Conference on Computational Natural Language Learning [Tjong Kim Sang and De Meulder, 2003]	CoNLL	Language-Independent	2002-2003	NER
Message Understanding Conference [Marsh and Perzanowski, 1998]	MUC	English	1987-1999	NER

Table 2.6: Some Evaluation Contests on NLP tasks that make available tagged data sets.

a fixed interval of years (e.g. the Semantic Evaluation contest).

When no appropriate Golden Set is available to properly evaluate a system, mainly for those processing large modern collections, it is usual for relevance to be assessed for only a subset of the documents involved. The most standard approach is *pooling*, where a subset of the collection is created by a number of different IR systems (usually the ones to be evaluated) from the top k documents. The relevance of the subset is then evaluated by humans. However, a human is not a device that reliably reports a gold standard judgment of relevance of documents to a query [Manning et al., 2008]. Thus it is interesting to consider and measure how much agreement between judges there is in judgments of relevance. In the social sciences, a common measure for the agreement between two human judges (raters) is the *kappa statistics* [Cohen, 1960], defined by the equation 2.13.

$$kappa = \frac{P_a - P_e}{1 - Pe} \tag{2.13}$$

where P_a is the proportion of the time the judges agreed, and P_e is the proportion of the times they would be expected to agree by chance. Fleiss [Fleiss et al., 1971] extended the kappa coefficient for any number of raters, also known as *Fleiss generalized kappa* and defined by the equation 2.14.

$$Fleiss\_kappa = \frac{P_a - P_{e.k}}{P_{max} - P_{e.k}} \tag{2.14}$$

where P_{max} is the maximum value that rater agreement can take in the case of all raters agreeing on all cases, which is 1. P_a is the proportion of observed rater agreement and is defined by the equation 2.15.

$$P_a = \frac{1}{nr(r-1)}(\sum n_{ij}^2 - nr) \tag{2.15}$$

where n is the total number of subjects, or items; r is the number of raters; and n_{ij} is the number of ratings in each cell (for item i and category j). Finally, the formulation of the chance agreement proportion, $P_{e.k}$ is given by the equation 2.16.

$$P_{e.k} = \sum_{j=1}^{q}(\sum_{i=1}^{n} r_{ij}/r)^2 \tag{2.16}$$

where j is the category number; q is the total number of categories; and r_{ij} is the number of raters selecting category j for subject/item i. $P_{e.k}$, then, is the sum across all categories of the square of the proportion of rater use in each category. The *kappa* value is 1 if judges always agree, 0 if they agree only at the rate given by chance, and negative if the rate of agreement is worse than random. As a rule of thumb, a kappa value above 0.8 is taken as a very good agreement, a kappa value between 0.6 and 0.8 is considered a good agreement, between 0.4 and 0.6 is considered a moderate agreement, between 0.2 and 0.4 a fair agreement and an agreement below 0.2 is a poor agreement and it is seen as data providing a dubious basis for an evaluation. But the acceptance of this measure is still far from consensus in the NLP scientific community [Koller et al., 2007] with respect to the fact that low kappa (< 0.4) values must be rejected. In fact, from those evaluation contests presented on table 2.6, only some systems presented kappa statistics besides F-measure on shared tasks [Hendrickx et al.; Mukund et al., 2011].

To summarize, this section has introduced the features which distinguish IE systems and can be reused or improved upon by the work described in this thesis. The new system proposed deals with entire texts and web pages content in order to extract all relevant terms (TE) from these sources of information. As a next step, its goals are to contextualize those terms in order for them to become concepts (terms with meaning), and to find their attributes and relationships, using the background presented in the next section.

2.5 Semantics

The growing amount of information available on the web demands the development of efficient and practical IE approaches, in order to avoid the overloading of information for users. This need for new ways of extracting information from the web has stimulated a new vision, the Semantic Web [Berners-Lee et al., 2001], where available resources have associated machine-readable semantic information. For this to come about, a knowledge representation structure for representing the semantics associated with resources would be necessary, and this is where ontologies [Zuniga, 2001] have assumed a central role in the movement of the Semantic Web.

Because it would be an over-ambitious task to design an ontology of the world, research has focused on the development of domain-specific ontologies, in which construction and maintenance are time-consuming and error-prone when done manually. In order to automate this process, research into Ontology Learning has emerged, combining IE and learning methods to automatically, or semi-automatically, build ontologies. In the same way, a great effort has been made in making open and free Common Sense knowledge resources available (such as WordNet [Miller et al., 1990], OpenCyc [OpenCyc, 2011], ConceptNet [Liu and Singh, 2004] and others). To represent this knowledge in a more formal way, it is also valuable to structure knowledge representation and show links between concepts.

2.5.1 Knowledge Resources

When using data from different sources, integration of information is imperative in order to avoid redundancy. In textual descriptions of places we have a two-dimensional space: location × attributes. While it is very straightforward to match locations, terms describing attributes and related concepts are generally expressed by compound words that are sometimes misspelled or written in different ways (e.g. using synonyms). In addition to this, a term may have different senses, so it would be useful to find the most appropriate in order to represent it semantically. To discover the meaning of each term, the use of common-sense and domain-knowledge sources is becoming popular, the first of which is the main focus of this thesis.

2.5.1.1 WordNet

WordNet [Miller et al., 1990] is a computational lexicon[2.10] of English based on psycholinguistic principles, created and maintained at Princeton University. It encodes concepts in terms of sets of synonyms (called synsets). Its latest version, WordNet 3.0, contains about 155,000 words organized in over 117,000 synsets. For example, the concept of car is expressed with the following synset (assuming the notation followed by WordNet and subscript $word\#p\#n$ where p denotes the part-of-speech tag and n the word's sense identifier, respectively):

$$car\#n\#1, auto\#n\#1, automobile\#n\#1, machine\#n\#6, motorcar\#n\#1$$

where the words are arranged by frequency order: the most frequent word used to refer that meaning comes first, and so on. The reader may notice that the word "machine" is not so common to refer to this meaning, being most often used to describe a more general concept like "any mechanical or electrical device that transmits or modifies energy to perform or assist in the performance of human tasks". The complete list of senses related to the noun "car" and their respective definitions can be seen in the following list:

[2.10] in other words, a set of words encoded in the dictionary

- $car\#n\#1, auto\#n\#1, automobile\#n\#1, machine\#n\#6, motorcar\#n\#1$: a motor vehicle with four wheels; usually propelled by an internal combustion engine;
- $car\#n\#2, railcar\#n\#1, railwaycar\#n\#1, railroadcar\#n\#1$: a wheeled vehicle adapted to railroad rails;
- $car\#n\#3, gondola\#n\#3$: the compartment that is suspended from an airship and that carries personnel, the cargo and the power plant;
- $car\#n\#4, elevatorcar\#n\#1$: where passengers ride up and down;
- $cablecar\#1, car\#5$: a conveyance for passengers or freight on a cable railway;

A synset can be seen as a set of word senses all expressing the same meaning. Each word sense uniquely identifies a single synset. For instance, given *car#n#1* the corresponding synset *car#n#1*, *auto#n#1*, *automobile#n#1*, *machine#n#6*, *motorcar#n#1* is uniquely determined. Figure 2.3 shows an excerpt of the WordNet semantic network containing the *car#n#1* synset. As words are not always so ambiguous, a word *w#p* is said to be *monosemous* when it can convey only one meaning. For instance, *restaurant* is a monosemous word, as it denotes a single sense, that of a building where people go to eat. Alternatively, *w#p* is *polysemous* if it can convey more meanings (e.g. *bank#n* as a piece of sloping land, a long ridge or pile, a slope in the turn of a road or track, etc.), which are quite often related in an etymological way. Senses of a word *w#p* which can convey unrelated meanings are *homonymous* (e.g. *bank#n* as a slope vs. *bank#n* as a financial institution).

For each synset, WordNet provides the following information:

- A gloss, that is, a textual definition of the synset possibly with a set of usage examples (e.g. the gloss of car#n#1 is "a motor vehicle with four wheels; usually propelled by an internal combustion engine; 'he needs a car to get to work"');
- Lexical relations, which connect pairs of word senses. Concerning only relations between nouns, Antonym is a lexical relation available (e.g. girl#n#1 the opposite noun of boy#n#1);

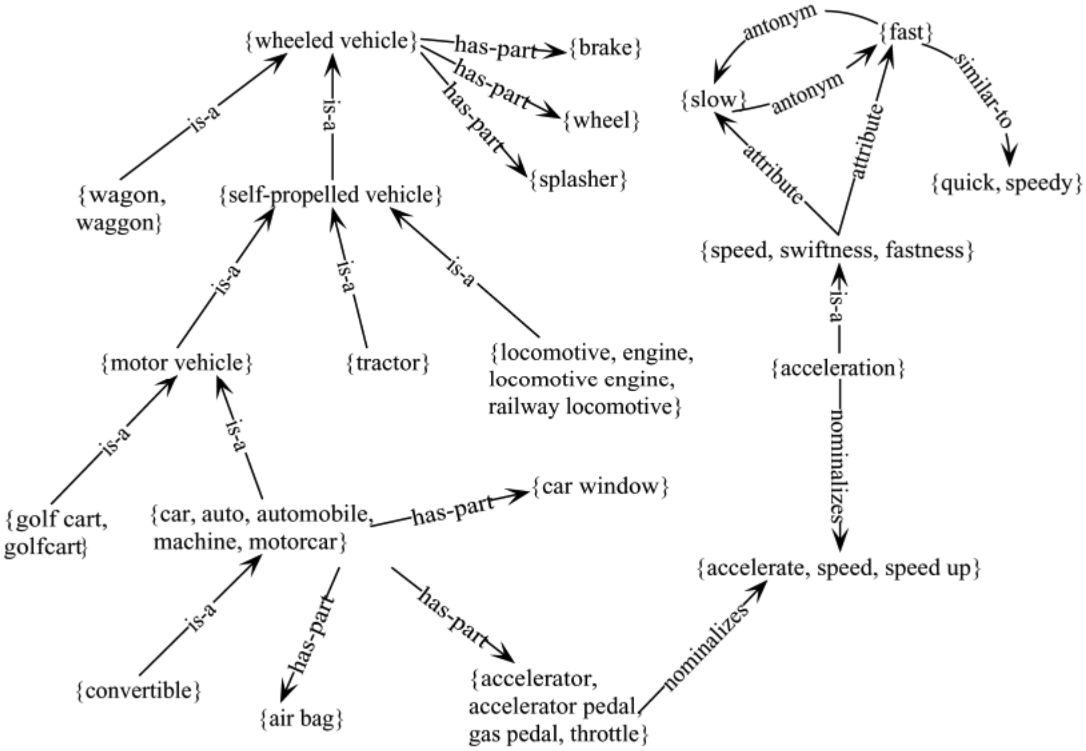


Figure 2.3: WordNet semantic network around car#n#1 meaning [Navigli, 2009]

- Semantic relations, which connect pairs of synsets. There are two semantic relations of interest to explore: Hypernym/Hyponym and Meronym/Holonym. The first pairing refers to inheritance between nouns, also known as an *is-a*, or *kind-of* relation and their respective inverses. Y is a hypernym of X if every X is a (kind of) Y (motor_vehicle#n#1 is a hypernym of car#n#1 and, conversely, car#n#1 is a hyponym of vehicle#n#1). The second pairing encloses the membership idea, and is also known as a *part-of*, or *member-of* relation and their respective inverses. Y is a meronym of X if Y is a part of X (accelerator#n#1 is a meronym of car#n#1, while car#n#1 is a holonym of accelerator#n#1).

SemCor [Mihalcea, 1998] is a textual corpus in which words are syntactically and semantically tagged in WordNet. It is the largest and most used sense-tagged corpus, which includes 352 texts extracted from the Brown corpus [Francis and Kucera, 1983] tagged with around 234,000 sense annotations. All the words in the corpus have been syntactically tagged using Brill's part-of-speech tagger [Brill, 1994]; the semantic tagging was done manually for all the nouns, verbs, adjectives and adverbs, each of these words being associated with its correspondent WordNet sense.

In summary, WordNet seems to be a powerful tool for inferring concepts related to a given place. Normally, these concepts and their semantic relationships are attached to a generic perspective, thus not representing any instance in themselves. For example, the concept of *library* can be generically described as *a building that houses a collection of books and other materials* (in WordNet), but if we talk about a specific library (e.g. the U.S. National Library of Medicine), further exploration is needed to achieve a more precise meaning of that place (which is generally not possible using common-sense Ontologies).

2.5.1.2 Wikipedia

Wikipedia is a multilingual web-based encyclopedia [Wikipedia, 2004]. Being a collaborative open-source medium, it is edited by volunteers. Wikipedia provides a very large domain-independent and interlinked encyclopedic repository. Its extensive network of links, categories and info-boxes[2.11] provide a variety of explicitly defined semantics that other corpora lack. However, it does not always engender the same level of trust or expectation of quality as traditional resources, because its contributors are largely unknown and some may be considered non-experts. It is also far smaller and less representative of all human language use than the web as a whole [Medelyan et al., 2009].

Since authors are free to create any content titled with different denominations, it is common that the same page can be found with different titles. This redirection availability, expressly edited by authors, also makes Wikipedia a good place to see entities' synonyms (e.g. the United States may be referenced by its acronym U.S. or by its complete name the United States of America). Presently, Wikipedia contains more than 4 million redirect pages.

Authors are also encouraged to categorize their articles by choosing the most appropriate categories (already existing or new ones). For example, the article "Museum"

[2.11] A special type of template that displays factual information in a structured, template-based format.

falls into these categories: Museums, Museology, Tourist activities, and Greek loanwords. Assigning the categories themselves to other more general categories is also promoted; Museums belongs to Educational buildings, which in turn belongs to Educational environment. These categorizations, like the articles themselves, can be modified by anyone. There are almost 650,000 categories in the English Wikipedia, each one not being considered an article in itself but merely a node for organizing the articles that it contains, with a minimum of explanatory text. The goal of the category structure is to represent an information hierarchy. It is not a simple tree-structured taxonomy, but a graph in which multiple organization schemes coexist. Both articles and categories can belong to more than one category [Medelyan et al., 2009].

Instead of taking readers to an article named by the term, as “Museum” does, the Wikipedia search engine sometimes takes them directly to a disambiguation page where they can click on the meaning they want (as in the case of “POI”, which refers to different pages). Currently, the English version contains 320,000 disambiguation pages, all with the word “disambiguation” in the title or belonging to the disambiguation pages category. Geo-referentiation is another attribute attached by authors to some pages related to geographical entities. Presently there are more than 171,000 geo-referenced pages in the database [Creative Commons, 2010].

2.5.1.3 Wikipedia as an Ontology

The organization of objects into categories is a vital part of knowledge representation. Documents in the Wikipedia collections are organized in a hierarchy of categories defined by the authors of the articles (section 2.5.1.2). The Wikipedia category system is a taxonomy for arranging articles into categories and sub-categories. Great effort has been made to organize this encyclopedic knowledge in an Ontology [Auer et al., 2007; Ponzetto and Strube, 2007; Siorpaes and Bachlechner, 2006; Suchanek et al., 2008; Syed et al., 2008]. The last three authors made their compiled data available for use in academic research and their work is explained in detail in the following paragraphs.

YAGO [Suchanek et al., 2008] aims to be a giant taxonomy by mapping Wikipedia’s leaf categories onto the WordNet taxonomy of synsets and by adding the articles belonging to those categories as new elements. YAGO extracts 14 relationship types, such as *subClassOf*, *type*, *familyNameOf*, *locatedIn*, etc. from different sources of information in Wikipedia. One source is the Wikipedia category system itself and another one is the system of Wikipedia redirects. The Wikipedia categories are organized in a directed acyclic graph, which yields a hierarchy of categories. The YAGO's authors define a mixed suite of heuristics for extracting further relations to augment the taxonomy. They also make use of the info-box structure to parse attributes and values (with ranges and domains) about the article entity. The YAGO classification schema consists of 286,000 classes which form a deep subsumption hierarchy. Details of the mapping algorithm are described in [Suchanek et al., 2008]. Characteristics of the YAGO hierarchy are its deepness and the encoding of much information in one class (e.g. the class *MultinationalCompaniesHeadquarteredInTheNetherlands*). Manual evaluation of sample facts shows 91-99% accuracy, depending on the relation.

On a larger scale, but less formally structured, the DBpedia project [Auer et al., 2007] was initially started by extracting facts from the info-boxes of particular types of Wikipedia articles (e.g. on people, cities, companies, music bands, etc.). This transforms Wikipedia's structured and semi-structured information (most notably info-boxes) into a vast set of RDF triples. Roughly, RDF (Resource Description Framework) is a language where Web information can be structured through relations (or triples) between a subject and an object.

DBpedia uses the words from the info-boxes as relation names. Recursive regular expressions are used to parse relational triples from all commonly used Wikipedia info-box templates containing several predicates. The templates are taken at face value; no heuristics are applied to verify their accuracy. This way, DBpedia can extract a wealth of facts from the info-boxes, but there is a drawback, in that the same relationship may appear with different names (e.g. length, length-in-km, length-km). The developers have manually mapped the most commonly used Wikipedia templates into a subsumption hierarchy consisting of 170 classes and then mapped 2,350 attributes from within these templates to 720 ontology properties [Bizer et al., 2009b]. However, links between categories are still merely extracted and labeled with the relation *isRelatedTo*. As with YAGO, Wikipedia categories are treated as classes and articles as individuals. The unsurpassed quantity of information in DBpedia is a wonderful resource for the research community, particularly given its multilingual character, and it is becoming a kind of a hub for free large-scale data repositories, but at the moment no formal evaluation of this resource has been reported [Bizer et al., 2009b].

WikiNet is a taxonomy derived from Wikipedia categories and the links between them. Although this approach pursues similar goals to YAGO's, it leads to lower quality and a more restricted scope [Ponzetto and Strube, 2007]. Ponzetto and Strube begin by identifying and isolating isA relations from the rest of the category links (which they call notIsA). Here isA is thought of as subsuming relations between two classes $isSubclassOf(\textit{Apples, Fruit})$ and between an instance and its class $isInstanceOf(NewZealand, Country)$. Several steps are applied. One of the most accurate steps matches the lexical head and modifier of two category names. Sharing the same head indicates isA, e.g. $isA(\textit{British computer scientist}, \textit{Computer scientist})$. Modifier matching indicates notIsA, e.g. $notIsA(\textit{Islamic mysticism, Islam})$. Another method uses co-occurrence statistics of categories within patterns to indicate hierarchical and non-hierarchical relations. For noun phrases "A and B", "A such as B" (e.g. fruit such as apples) indicates isA, and the intervening text can be generalized to a textual construction like, ", especially", and so on. Similarly, "A is used in B" (fruit is used in cooking) indicates notIsA. Comparison with relations manually assigned to concepts with the same lexical heads in ResearchCyc (a version of the Cyc ontology like OpenCyc [OpenCyc, 2011] but devoted to research purposes) shows that the labeling is highly accurate, depending on the method used, and yields an overall F-measure of 88%.

2.5.2 Word Sense Disambiguation

Word Sense Disambiguation (WSD) aims to choose the right sense given the context of a word. WSD has been described as an AI-complete problem, that is, by analogy to NP-completeness in complexity theory, a problem whose difficulty is equivalent to solving central problems of AI. Several algorithms have been developed to deal with this issue [Navigli, 2009], but the difficulty of WSD is also confirmed by its lack of applications to real-world tasks. On the other hand, WSD heavily relies on knowledge. In fact, the skeletal procedure of any WSD system can be summarized as follows: given a set of words (e.g. a sentence or a *bag of words*), a technique is applied which makes use of one or more sources of knowledge to associate the most appropriate senses with words in context. Knowledge sources can vary considerably from corpora (i.e. collections) of texts, either unlabeled or annotated with word senses, to more structured resources, such as machine-readable dictionaries, semantic networks, etc. Without knowledge, it would be impossible for both humans and machines to identify the meaning of sentences [Navigli, 2009].

A baseline is a standard method to which the performance of different approaches is compared. To compare WSD systems' performance, two basic baselines are commonly used, the random baseline, where a sense from those available in the knowledge source is selected by chance, and the first sense baseline. The first sense baseline is a real challenge for all-words WSD systems [Navigli, 2009] in that only a few systems are able to exceed it. This is not the case in lexical sample WSD, as more training data is usually available and the task is less likely to reflect the real distribution of word meanings within texts (e.g. if we consider the extreme case of an equally balanced frequency of the meanings of a word: the first sense baseline would perform with a level of accuracy equal to the random baseline). The fact that most methods find it difficult to overcome the first sense baseline is an indicator that most of the attempts seem to have been of little use.

Despite the popularity of WordNet in WSD tasks, Wikipedia has gained more attention as it offers less granularity in sense definition (making WSD easier) and more context (complete Wikipedia articles) to each sense than a restricted number of sentences (glosses). As a collaborative platform, Wikipedia defines only those senses on which its contributors reach consensus [Medelyan et al., 2009]. Despite the fact that there is still far less research on word sense disambiguation using Wikipedia than using WordNet, significant advances have been made. Over the last three years the accuracy of mapping documents to relevant Wikipedia articles has improved by one third [Milne and Witten, 2008]. However, for fair comparison, the same version of Wikipedia and the same training and test sets should be used, as has been done for WordNet[2.12]. This is not currently the case. Similar to named entity extraction research, where each research group concentrates on different types of entities, e.g. persons or places, the use of Wikipedia to disambiguate terms must be evaluated by extrinsic evaluations of

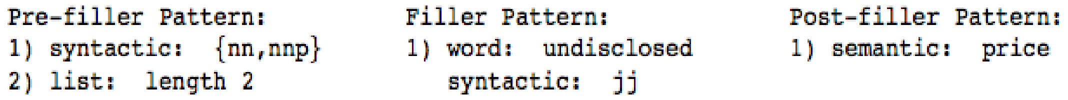
[2.12] The "Senseval" contest now named "SemEval" [SemEval Portal, 2011] what improvement was gained in the performance of a given system.

There is a clear distinction between approaches that are concerned with finding *named entity* mappings, and others that are concerned with mapping all kinds of concepts to Wikipedia [Medelyan et al., 2009]. For the first group the main motivation is that Wikipedia is recognized as the largest available collection of such entities (proper nouns such as geographical and personal names, and titles of books, songs and movies). It has become a platform for discussing news, and contributors put issues into encyclopedic context by relating them to historical events, geographical locations and significant personages, thereby increasing the coverage of named entities. For the second type of research, every term or phrase is an eligible entry for searching in Wikipedia articles.

2.5.3 Knowledge Resources and Information Extraction

Apart from contextualization, knowledge resources can also directly act in the IE process, such as in the generalization of extraction rules or selecting candidate terms. In [Milne and Witten, 2008], it is argued that they have enormous potential, since these algorithms cross-reference documents with large knowledge bases, and can provide structured knowledge about any unstructured document. Thus any task that is currently addressed using the bag-of-words model could probably benefit from these techniques.

With regard to external knowledge bases, some IE systems use lexical semantic databases, such as WordNet (section 2.5.1.1), which provide word classes that can be used to define more general extraction patterns, such as the RAPIER system [Califf and Mooney, 2003]. This IE system implements a popular rule-learning algorithm, the Bottom-up rule learning, in which the starting rule is a very specific rule covering just a pair of specific instances. This rule is gradually made more general so that the coverage increases with a possible loss of precision. For instance, figure 2.4 shows a rule learned by RAPIER for extracting the transaction amount from a newswire concerning a corporate acquisition. This rule extracts the word “undisclosed” from phrases such as “sold to the bank for an undisclosed amount” or “paid Honeywell an undisclosed



```
Pre-filler Pattern:         Filler Pattern:           Post-filler Pattern:
1) syntactic:  {nn,nnp}     1) word:  undisclosed     1) semantic:  price
2) list:  length 2             syntactic:  jj
```

Figure 2.4: A RAPIER's rule for extracting the transaction amount from a newswire concerning a corporate acquisition. "nn" and "nnp" are the part-of- speech tags for noun and proper noun, respectively; "jj" is the part-of-speech tag for an adjective [Califf and Mooney, 2003].

price". The pre-filler pattern[2.13] matches a noun or proper noun (indicated by the PoS tags 'nn' and 'nnp', respectively) followed by at most *two* other unconstrained words. The filler pattern[2.14] matches the word "undisclosed" only when its PoS tag is "adjective". The post-filler pattern[2.15] matches any word in WordNet's semantic class named "price".

Some authors go further on name contextualization and propose to completely automate the link insertion which is done manually at present. This process is named "Wikification" [Mihalcea and Csomai, 2007]. They disambiguate terms (words or phrases) that appear in plain text to Wikipedia articles, concentrating exclusively on *important* concepts. Thus, it can be considered a term extraction task with Wikipedia as the controlled vocabulary. There are two stages: extraction and disambiguation. In the first, terms that are judged important enough (by statistical measure) to be highlighted as links are identified in the text. The *keyphraseness* feature of a candidate term is defined as the number of Wikipedia articles in which the term appears and is marked up as a link divided by the total number of Wikipedia articles where the term appears. This feature can be interpreted as the probability that the candidate term is selected as a key term in a Wikipedia article as, according to the Wikipedia editorial rules, only key terms should be used as links to other articles. Wikify! uses keyphraseness as the only feature to select key terms. The number of key terms selected by the method is specified by the special external parameter density as a fraction of the overall number of words in a document being processed. In the second stage, these terms are disambiguated to Wikipedia articles that capture the intended sense. The creators have

[2.13] the text immediately preceding the phrase to be extracted
[2.14] the string to be extracted
[2.15] the text immediately following the phrase to be extracted

implemented the *Wikify!* system available online as a demo[2.16].

Finally, [Grineva et al., 2009] present an approach that uses Wikipedia as a lexicon for topic descriptors. The authors exploit the concept of *community* which is formed by applying the Girvan-Newman network analysis [Newman and Girvan, 2004] on extracted terms from each Wikipedia article. The method consists of the following five steps:

1. The extraction of candidate terms considers all n-grams which refer to Wikipedia article titles.

2. Word sense disambiguation is applied when more then one article (ambiguous) can be matched against one candidate term.

3. A semantic graph is built which is a weighted graph, i.e. a semantic graph, where each vertex is a term; an edge between a pair of vertices means that the two terms corresponding to these vertices are semantically related; and the weight of the edge is the semantic relatedness measure of the two terms.

4. A community structure of the semantic graph is discovered through a community-detection algorithm to group terms semantically.

5. Valuable communities are selected, where the groups are then ranked based on the semantic similarity and keyphraseness of its members and the most significant community members are chosen as topics. This approach uses the link structure of candidate Wikipedia entities to first group them into clusters and then rank these clusters by the overall importance of their entities, measured using their in-links, in the way also used by the Wikify! system [Mihalcea and Csomai, 2007].

For testing the approach, 252 posts from technical blogs were manually annotated by 22 volunteers and the results obtained on this data set was 67.7% recall and 46.1% precision, using Wikipedia as the controlled vocabulary.

[2.16] http://wikifyer.com/

2.5.4 Ontology Learning

According to [Petasis et al., 2007], ontology learning can be described as *"the process of automatic or semi-automatic construction, enrichment and adaptation of ontologies"*. It relies on a set of algorithms, methods, techniques and tools to automatically, or semi-automatically, extract information about a specific domain to construct or adapt ontologies. The process of ontology learning comprises four different tasks: ontology population, ontology enrichment, inconsistency resolution and ontology evaluation. Ontology population is the task that deals with the instantiation of concepts and relations in an ontology, without changing its structure. In contrast, ontology enrichment is the task of extending an ontology by adding new concepts, relations and rules, which results in changes in its structure. Also in contrast to ontology learning, ontology enrichment by population and instantiation can be considered a less ambitious task to automatize. Indeed, Stumme et. al [Stumme et al., 2006] state that the purpose of Ontology Learning is not to replace the human, but rather to provide her with more and more support.

As human language is a primary mode of knowledge transfer, ontology learning from relevant text collections seems indeed a viable option as illustrated by a number of generic systems that are based on this principle, e.g. ASIUM [Faure and Ndellec, 1998], TextToOnto [Maedche and Staab, 2001], OntoMiner [Davulcu et al., 2003] and Ontolearn [Navigli and Velardi, 2004]. Particularly, each system presents some specificities as detailed in the following list:

- ASIUM [Faure and Ndellec, 1998] learns ontologies by bottom-up conceptual clustering from parsed corpora in specific domains. The clustering method also generalizes the grammatical relations between verbs and complement heads as they are observed in the corpora. The set of grammatical relations learned for a given verb forms the verb subcategorization frame.

- TextToOnto [Maedche and Staab, 2001] is an ontology-learning environment which has a bookshop of learning methods such as formal concept analysis, association rules which permit the extraction of concepts, taxonomic relations and non-taxonomic relations, a linguistic tool (an analyzer) and heuristics.

- OntoMiner [Davulcu et al., 2003] analyzes sets of domain-specific web sites and generates a taxonomy of particular concepts and their instances. This tool uses HTML regularities within web documents in order to generate a hierarchical semantic structure encoded in XML. It explores directories of home pages in order to detect key domain concepts and relations between them.

- OntoLearn [Navigli and Velardi, 2004] is a tool based on linguistic and semantic techniques. It extracts domain terminology from Web documents by using a linguistic processor and a syntactic parser. The semantic interpretation task consists of finding the appropriate WordNet concept for each term.

However, all of these combine a certain level of linguistic analysis with machine-learning algorithms to find potentially interesting concepts and relations between them applied to any domain. These systems implement a typical approach in ontology learning from texts which involves term extraction from a domain-specific corpus through a statistical or rule-based process that determines their relevance for the domain corpus at hand. These are then clustered into groups with the purpose of identifying a taxonomy of potential classes. Subsequently, relations can be also identified by computing a statistical measure of connectedness between identified clusters. The existing approaches rely on the assumption that documents have either a similar structure or a similar content, an assumption which seems unrealistic due to the heterogeneity of the Web.

With regard to specific domains, the Artequakt [Alani et al., 2003] system uses natural language tools to automatically complete knowledge about artists from multiple Web Pages, based on a pre-defined and hand-crafted ontology to generate artist biographies. The system uses a biography ontology and a database of artists with missing attributes, especially built for this purpose, which defines the data for an artist biography. In this sense, this system does not learn a domain ontology, but instead populates concepts and attributes on a manually built ontology for a given instance (an artist). The Information for ontology instantiation is collected by parsing text found on the Web and is subsequently presented using templates. The NLP environment GATE [Cunningham et al., 2002] is used to extract named entities and to resolve co-reference from Web documents retrieved from a search engine with the artist's name as query.

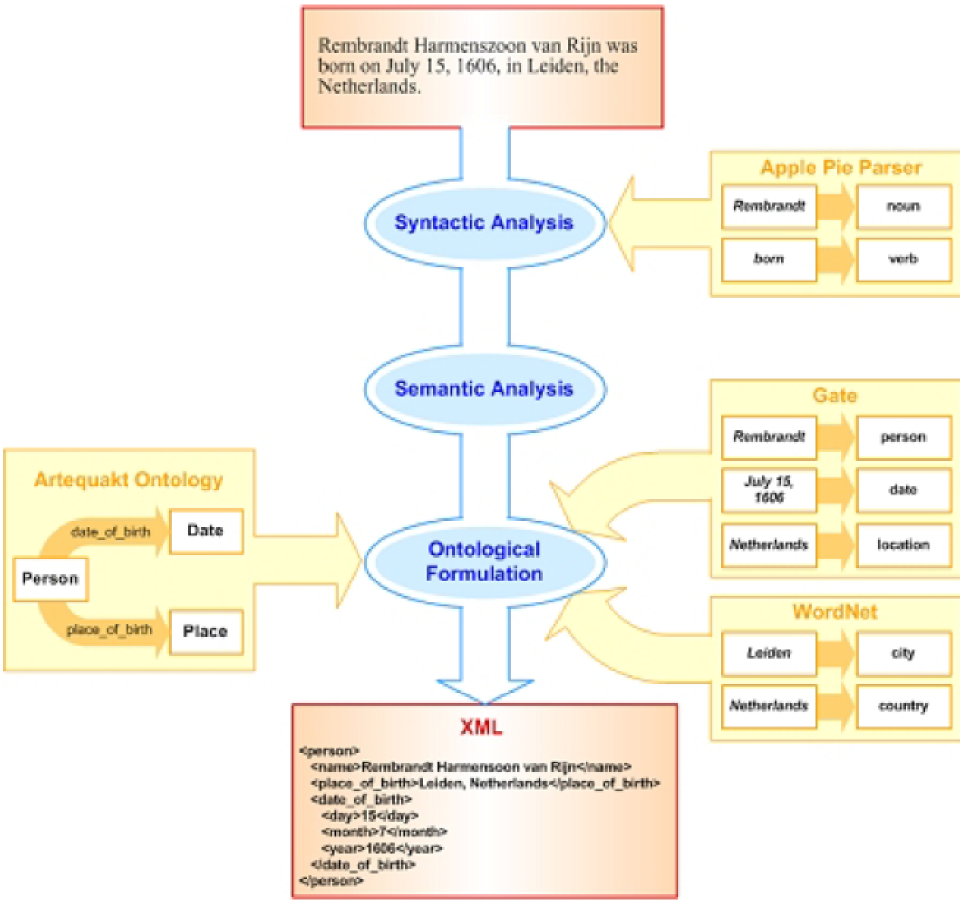


Figure 2.5: The Artequakt extraction process [Alani et al., 2003].

It assumes that web pages are mostly composed of syntactically well-constructed texts in order to extract knowledge triples (concept - relation - concept). web pages are divided into paragraphs, and then into sentences. Each sentence, which heuristically corresponds to a grammatical construction of the form Subject-Verb-Concept, is then used to complete a triple. These triples are only extracted to complete the missing information from a knowledge base of artists. For instance, if for a given artist it is missing information about their birth location and date, the system harvests the Web looking for this information, as shown in figure 2.5.

Semantic Web Mining [Stumme et al., 2006] is the conjunction of Web Mining and the Semantic Web. From Web Mining, Information Extraction (section 2.4) is the perfect support for knowledge identification and extraction from Web documents, as it can, for example, provide support in the analysis of documents either in an automatic way (unsupervised extraction of information) or in a semi-automatic way (e.g. as support for human annotators in locating relevant facts in documents, via information highlighting). A materialization of this idea is proposed on the Web → KB system [Craven et al., 2000], where a knowledge base is built from Web Pages given an ontology and training examples from the Web that describe instances of these classes and relations. Then the system learns general procedures for extracting new instances of these classes and relations from the Web.

Figure 2.6 presents an overview of the system to represent the knowledge base of a university. The top part of the figure shows an ontology that defines the classes and relations of interest. Each partial box represents a class, and the arrows indicate specialization relationships. The other defined relations for each class are listed inside of its corresponding box. The bottom part shows two web pages identified as training examples of the classes "course" and "faculty'. Given the ontology and a set of training data, Web → KB learns to interpret additional web pages and hyperlinks to add new instances to the knowledge base, such as those shown in the middle of the figure. These instances are represented by dashed partial boxes. The dashed lines show the relationships between the instances and their Web sources. Despite the fact that this system uses as input a complete ontology and an example for each class and relation for further instantiation, Craven et al. [Craven et al., 2000] assume some hypothesis, for example, that each class instance is represented by a single page; or that there is a link (or a sequence of links) between two instances.

2.6 Semantic Annotation

Document content is readable by humans, but, unless it is semantically annotated, it is not machine readable, in the sense that it cannot be automatically interpreted in any reasonable manner. Instead of tagging entities in texts with limited classes derived directly from named entity recognizers (e.g location, person), Semantic Annotation gives meaning to each label it refers to. This meaning is generally derived from ontologies and lexical resources. For instance, rather than just annotating the word Cambridge

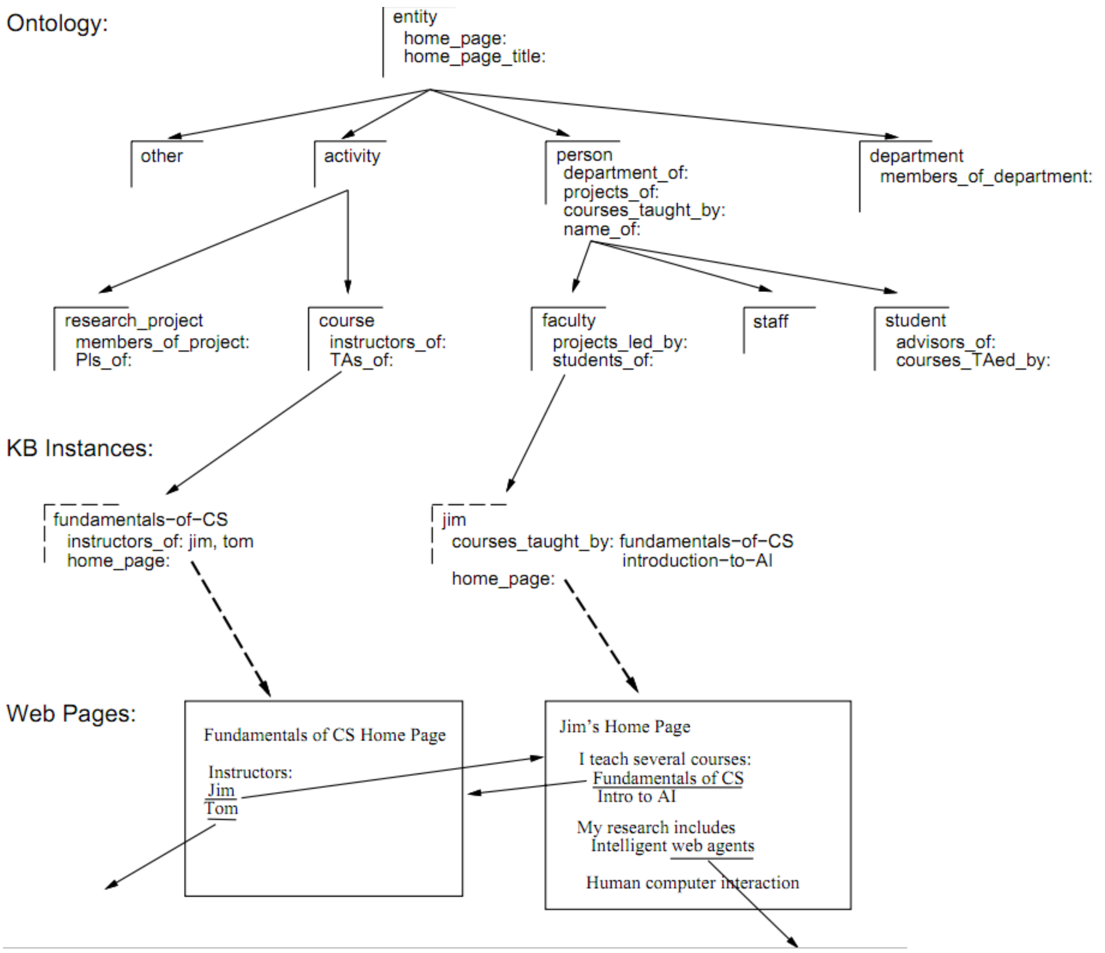


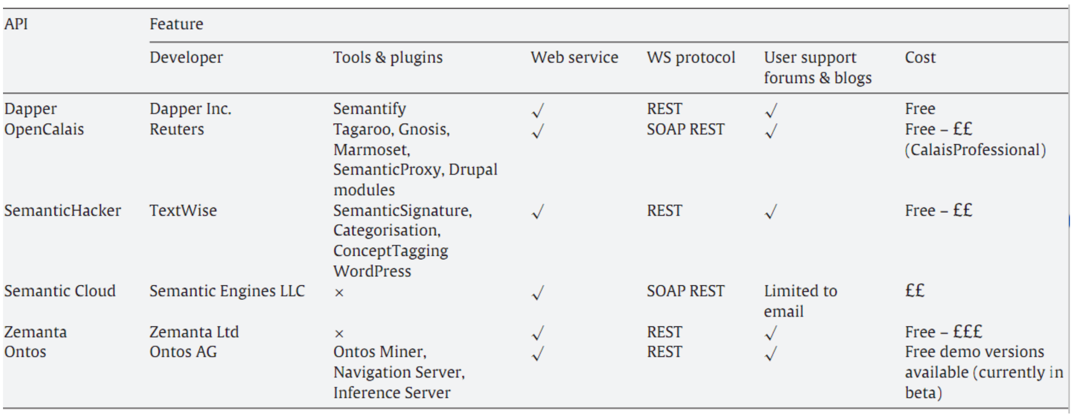
Figure 2.6: Web → KB system overview for the University domain [Craven et al., 2000].

as a location, linking it to an ontology instance with a spatial attribute would allow a system to differentiate between Cambridge in the United Kingdom and Cambridge, in Massachusetts, USA. Thus, in this case it would also be possible to use the knowledge behind this concept that the document mentions a city in Europe. This inference is only possible because ontologies tell us that this particular Cambridge is part of the country called the United Kingdom, which is part of the continent Europe. Semantic Annotation is sometimes crucial for the Ontology Learning process since after the text has been annotated (manually, semi-automatically or automatically), it is used for populating and enhancing the ontology. Actually, these two tasks are combined in a cyclical process: ontologies are used for interpreting the text at the right level for Semantic Annotation and Semantic Annotation highlights new knowledge from the text, to be integrated into the ontology.

MnM [Vargas-Vera et al., 2002] is a semantic annotation tool that, instead of building an ontology, tags texts with ontological semantic annotations. When integrated with the Amilcare IE system [Ciravegna, 2001], it automatically extracts text phrases from documents, which facilitates the semantic annotation phase conducted by an ontologist. After being annotated, a training corpus is then fed into a wrapper induction system in order to induce rules that can be used to extract information from corpus texts.

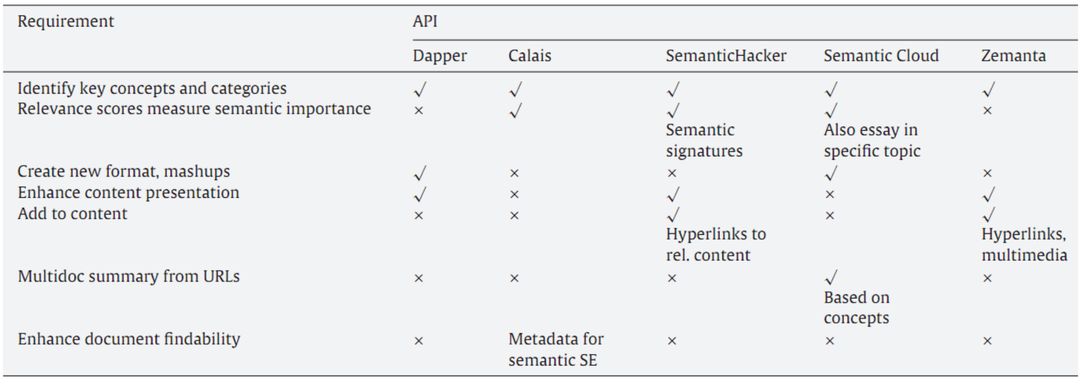
Armadillo [Dingli et al., 2003] is also a semantic annotation tool which utilizes the Amilcare IE system [Ciravegna, 2001] to perform wrapper induction, but this time on web pages. This system uses a pattern-based approach to find entities, finding its own initial set of seed-patterns rather than requiring an initial set of seeds. Manual patterns are used for the named entity recognizer, but no manual annotation of corpus documents is required. Once the seeds are found, pattern expansion is then used to discover additional entities. Information redundancy, via queries to Web services such as Google and CiteSeer, is used to verify discovered entities by analyzing query results to confirm or deny the existence of an entity.

Nowadays, some state-of-the-art semantic annotation tools (e.g. [Calais, 2008; Hacker, 2008]) have been adopted in the industry because of their availability through



API	Feature					
	Developer	Tools & plugins	Web service	WS protocol	User support forums & blogs	Cost
Dapper	Dapper Inc.	Semantify	√	REST	√	Free
OpenCalais	Reuters	Tagaroo, Gnosis, Marmoset, SemanticProxy, Drupal modules	√	SOAP REST	√	Free – ££ (CalaisProfessional)
SemanticHacker	TextWise	SemanticSignature, Categorisation, ConceptTagging WordPress	√	REST	√	Free – ££
Semantic Cloud	Semantic Engines LLC	×	√	SOAP REST	Limited to email	££
Zemanta	Zemanta Ltd	×	√	REST	√	Free – £££
Ontos	Ontos AG	Ontos Miner, Navigation Server, Inference Server	√	REST	√	Free demo versions available (currently in beta)

Figure 2.7: Product information of each Semantic Annotation APIs [Dotsika, 2010].



Requirement	API				
	Dapper	Calais	SemanticHacker	Semantic Cloud	Zemanta
Identify key concepts and categories	√	√	√	√	√
Relevance scores measure semantic importance	×	√	√ Semantic signatures	√ Also essay in specific topic	×
Create new format, mashups	√	×	×	√	×
Enhance content presentation	√	×	√	×	√
Add to content	×	×	√ Hyperlinks to rel. content	×	√ Hyperlinks, multimedia
Multidoc summary from URLs	×	×	×	√ Based on concepts	×
Enhance document findability	×	Metadata for semantic SE	×	×	×

Figure 2.8: Requirements-based decision making in comparing Semantic Annotation APIs [Dotsika, 2010].

APIs. A recent survey [Dotsika, 2010] compares such tools across different functionalities. Table 2.7 briefly presents each tool. Observing these characteristics, we will focus on those that are most popular, able to be used (free of charge) and rely on the following aspects: identifying key concepts and categories; measuring semantic importance through relevance scores; creating new formats, (e.g. mash-ups); and adding to content. These aspects are summarized in table 2.8. Next, we will look at in detail those tools which are free. However, as they are commercial products, the information provided is from consultation of the relevant website, so it is not possible to deeply analyze which specific algorithms they employ.

OpenCalais [Calais, 2008; Reuters, 2008] is an automatic generator of semantic

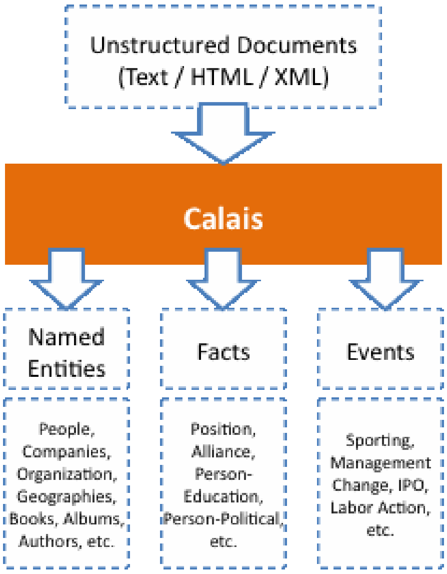

Unstructured Documents
(Text / HTML / XML)
Calais
Named Entities
Facts
Events
People, Companies, Organization, Geographies, Books, Albums, Authors, etc.
Position, Alliance, Person-Education, Person-Political, etc.
Sporting, Management Change, IPO, Labor Action, etc.

Figure 2.9: OpenCalais architecture

metadata in RDF format from unstructured text. It works on text only (no other media files are supported) and its architecture is shown in figure 2.9. The API reads unstructured documents (plain text, HTML, XML), recognizes a number of different entities and annotates them semantically in RDF. Existing entities include person, company, place and event. Apart from the list of entities, the service returns a number of occurrences and relevancy scores measuring the semantic importance of the various entities. The latest version, Calais 4.0, can assign content to ten different categories: health, politics, sports, technology, law, business & finance, entertainment & culture, travel, weather and environment.

OpenCalais has mainly been used for the automatic tagging of web blog posts. Table 2.7 shows an example of the results that OpenCalais returns. OpenCalais is free of charge. However, for a service where the amount of daily submissions is expected to exceed 40,000, there is CalaisProfessional, a paid equivalent to OpenCalais. CalaisProfessional offers a higher-class service, a service level agreement and a five times higher submission rate capability.

SemanticHacker [Hacker, 2008] is a semantic annotation API developed by TextWise that takes text as input, highlights a list of concepts and categorizes the document content creating what TextWise calls its *"semantic signature"*. To calculate all of this, TextWise uses a variation of the Vector-Space Model that they have created called Trainable Semantic Vectors (TSVs). A TSV calculates a semantic index, i.e. a semantic signature, which consists of a weighted vector of typically thousands of *semantic dimensions*. These semantic dimensions are a result of an one-time supervised training process from an appropriate classification schema for the domain and can be labels that represent categories extracted from the Open Directory Project [The Open Directory Project, 2002].

In addition, they do not require the manual construction or maintenance of ontologies. Instead, a TSV automatically generates its own semantic dictionary during training that contains the vocabulary known to be relevant to the application domain.

The matching service provides a similarity search that matches the text's semantic signature to a number of context indexes such as Wikipedia, YouTube videos, etc., and includes a number of links that may be relevant for further reading. The listed items are ordered based on their match score. The index call is available only after licensing a Custom Content index for a fee. The API then allows users to perform similarity searches against their own custom content. SemanticHacker works under a license agreement. Users are sent a token upon registration which enables access to the API and allows a limited number of queries. Additional queries, along with custom dictionary development, can be purchased after contacting TextWise.

The Zemanta API [Zemanta, 2009] takes in unstructured text and returns tags, categories, links, photos, and related articles. The service acts as a single-point entry to various pre-indexed content databases. Originally conceived to facilitate the editing of blogs, Zemanta analyzes the postings, discovers relevant content and adds it to the page or document. The system is powered by NLP and semantic algorithms. It categorizes content by comparing it to their pre-indexed database. The categorization process is constantly enhanced by end-user input and machine-learning methods.

Text	Entities	Topics	Relations
George Bush was the President of the United States of America until 2009. Barack Obama is the new President of the United States now.	Country: United States of America(0.43) Person: Barack Obama(0.29) George Bush(0.29)	Politics	*PersonPoliticalPast:* Person: George Bush Position: President *PersonPolitical:* Person: Barack Obama Position: President of the United States

Table 2.7: An example of output from OpenCalais for a given input text.

3

A Model for the Semantic Enrichment of Places

In this chapter, we will take a top-down approach, presenting the requirements for an *abstract semantic enrichment model* and then progressing towards a formal model applied for the *semantic enrichment of places*, which will be the subject of the implementation presented in the the next four chapters. In the actual implementation of the proposed system many choices were made, both in terms of what aspects to focus on (the sources from which to retrieve information, what to extract) and in terms of practical decisions (e.g. implementing algorithms, or choosing representations). Whenever necessary, we will justify each decision taken throughout the conceptualization and implementation of the *Semantic Enrichment of Places.*

3.1 A Generic Semantic Enrichment Model

From our perspective, the *semantic enrichment* process may be able to attach information to any given entity[3.1] (e.g. a place, a event, a person). In chapter 2, questions regarding semantic enrichment were raised. Accordingly, a set of principles followed in the synthesis of the model for semantic enrichment proposed is recalled here:

- Information Retrieval: is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large

[3.1] We will apply the term *entity* to any generic object, physical or abstract, on which the semantic enrichment is carried out.

collections (usually stored on computers)[Manning et al., 2008]. A model for semantic enrichment should find the right information about a given entity in a large collection of documents.

- Information Extraction: refers to the automatic extraction of structured information such as entities, the relationships between them, and attributes describing entities from unstructured sources [Sarawagi, 2008]. There are two general problems: extracting information from natural language texts and extracting structured data from Web pages. A model for semantic enrichment should also be able to perform these two tasks: in a first phase, it should be able to find entities and directly related information from Web Pages and, in a second stage, it should be able to extract relevant information about each entity in the documents previously found by Information Retrieval.

- Knowledge Representation: can be defined as *the study of how to put knowledge into a form that a computer can reason with*[Russell and Norvig, 2003]. A model for semantic enrichment must be able to build a structured representation of the information extracted about the entity studied. Each term previously extracted must be contextualized in a *Common Sense Ontology* and given a context where it appears. After being *disambiguated*, these terms are better described as concepts, since they have a meaning and are related to other concepts by semantic relations. This representation must allow us to infer more knowledge about each concept and also, in a wider view, about the entity itself, as these concepts are present and interlinked with other domain ontologies.

From these three main parts and intermediate aspects, we propose a Generic Semantic Enrichment model. Inspired by the fact that most Ontology Learning systems [Buitelaar et al., 2005] use texts entered by humans as their primary source, our model contributes a novel way to discover this information from pre-selected on-line sources. Figure 3.1 shows a model that considers the many aspects referred to above. Given an on-line source of entities of the same kind chosen by the user (generically any kind, e.g. POIs, events, blogs, artists), the Information Retrieval module should be responsible for two of the main steps in the Semantic Enrichment process: mining of entities in the source; and later finding related documents associated with these entities. The first step is achieved using scraping techniques, the source being a directory Web site or

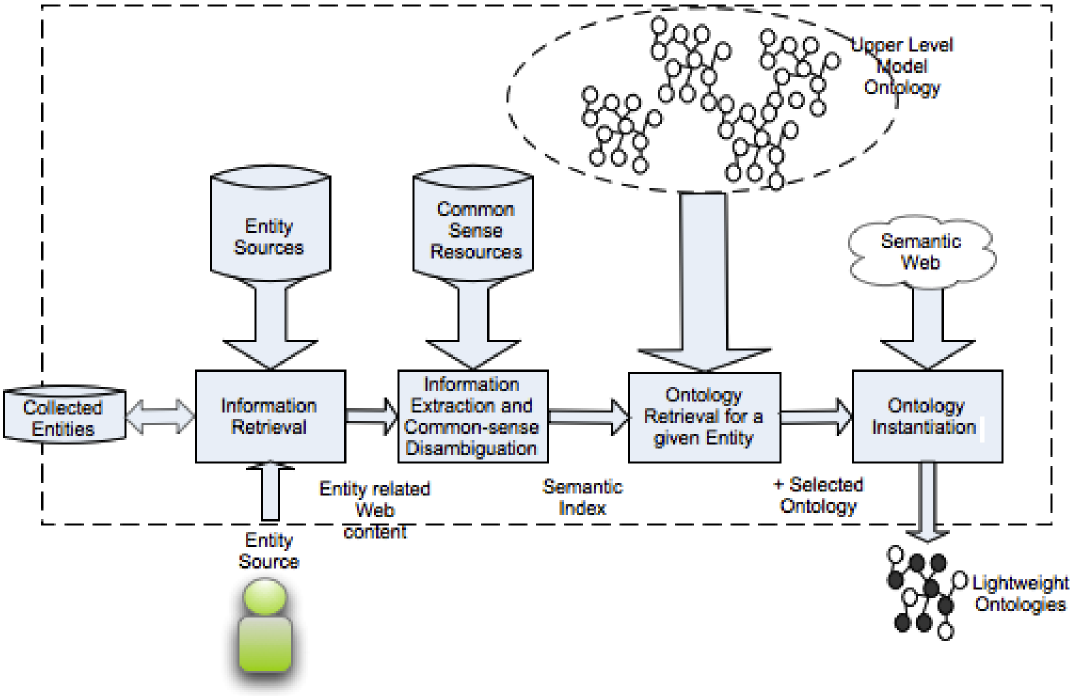


Figure 3.1: Generic model of Semantic Enrichment.

a database of entities. The next module, Information Extraction and Common-sense Disambiguation, is responsible for extracting relevant terms from those documents and contextualizing them in common-sense resources. The classification of a given entity in an Upper Level Ontology allows the generic model to infer abstract concepts about the entity. Furthermore, this Upper Level Ontology facilitates the materialization of abstract concepts connecting them to instances (e.g. Named Entities). This instantiation is possible using the Semantic Web, where each concept is not only contextualized by its meaning but also interlinked with other concepts and resources in a structured way.

The model just presented is an abstraction of the process conceived for the Semantic Enrichment of an entity. Each module can be seen as a road map to follow while making other choices or using different techniques to solve each intermediate task. For example, in the selection of sources for Information Retrieval, a range of alternatives could be explored in order to retrieve the information associated with each entity beyond the ones we have studied[3.2]. Also, the Automatic Term Recognition phase in the Information Extraction module could be implemented using different techniques available in the literature. The implementation of this generic model to enrich a specific kind of entity is covered in the next section and detailed later in the following chapters, leaving some of the last modules to be explored and accomplished by others in future work.

3.2 Description of the Implemented Model

The implementation of the generic model suggested earlier is outlined in figure 3.2. Each module in this architecture is explained and discussed in detail in the respective chapters as shown in the figure. The expected processing and data flow in this system is as follows:

- *A POI* Source is specified as input for the system. In the present architecture, the structure of this source (e.g. a collection of documents, a directory Web site, a POI-finding API) is mapped to the POI entity on the conceptual data model in order to automatically populate a database of POIs. This intensive extraction, or POI Mining, may be accomplished by simply invoking API (when available) or by Web scraping.

- *Information Retrieval on a Perspective* consists of finding documents about each POI from a given background collection. The World Wide Web and Wikipedia were explored to retrieve such information applying two different approaches for each collection. The system explores the Web in two different search methods: open and focused. The Wikipedia knowledge source is used through two distinct ways to locate generic and specific information about places. For these four combinations of selecting the subset of documents in the source collection, we name each one as a different *Perspective* of the semantic enrichment.

- *Meaning Extraction* finds the bag of relevant concepts in a document retrieved from a given source (Web or Wikipedia). Instead of a common bag of words, this set contains terms with meaning, since the disambiguation of each term is performed in this module. Therefore, the intermediate output of the Meaning

[3.2] detailed in chapter 6, which we call *Perspectives* of an entity.

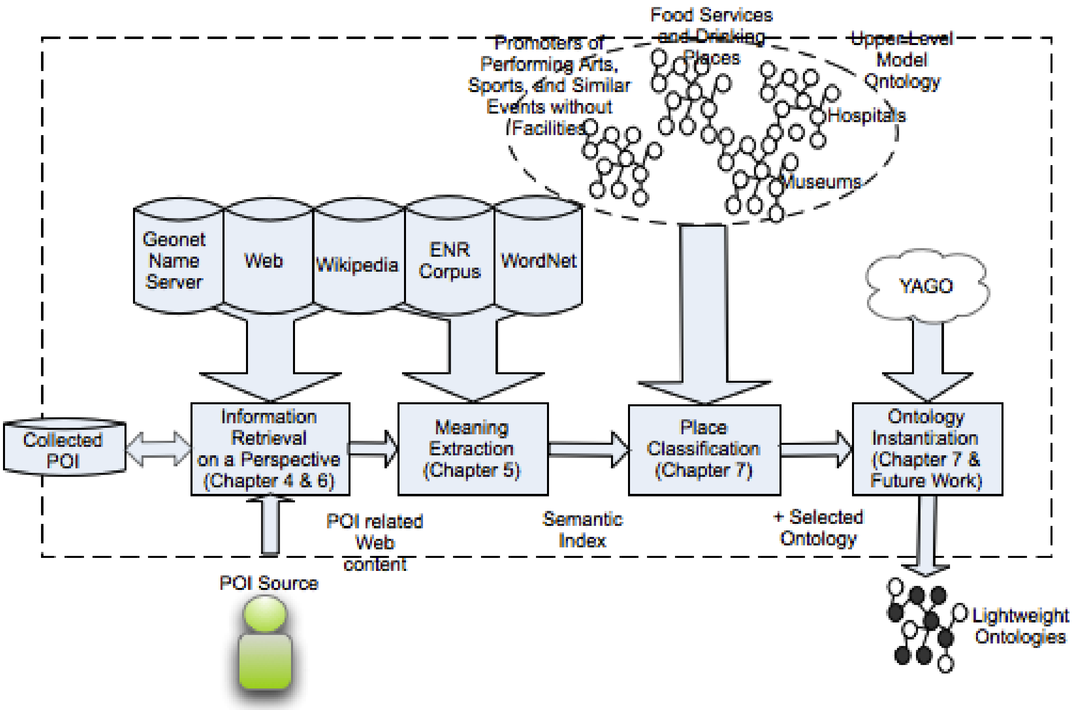

Food Services and Drinking Places
Promoters of Performing Arts, Sports, and Similar Events without Facilities
Upper-Level Model Ontology
Hospitals
Museums
Geonet Name Server
Web
Wikipedia
ENR Corpus
WordNet
YAGO
Collected POI
Information Retrieval on a Perspective (Chapter 4 & 6)
Meaning Extraction (Chapter 5)
Place Classification (Chapter 7)
Ontology Instantiation (Chapter 7 & Future Work)
POI related Web content
Semantic Index
+ Selected Ontology
POI Source
Lightweight Ontologies

Figure 3.2: Implemented Model for Semantic Enrichment of Places.

Extraction module is called a *Semantic Index*, where each concept is weighted using statistical relevance metrics.

- *Place Classification* performs the task of classifying a given POI within an Upper Level Ontology, allowing the next module to specify the concepts, relations and attributes in this new instance of the suggested ontology.

- *Ontology Instantiation* is the final and proposed module for reusing shared knowledge on Web 3.0 to find and induce already known and new concepts, relations and attributes about a given POI.

- *Lightweight Ontologies* are the output of the system as the final representation of a place. We chose this less formal knowledge representation since the scope of this work is not to establish rules or axioms about the entity represented but, instead, to reuse and connect shared information that is dispersed on-line.

4

Retrieving Information about Places from the Web

Before extracting and enriching information about Public Places, it is crucial to locate and collect potential sources of such data. We search directly in the web, to collect either Places (POIs) themselves or related textual descriptions. The two main phases in this process are described in the following sections. Section 4.1 defines a POI and describes its extraction from Yellow Pages services or directories and explains how to recognize POI duplicates from different sources and organize them. Section 4.2 presents the KUSCO module of retrieving documents associated with POIs and with their categories from the web.

4.1 Points of Interest (POIs)

The work here presented was developed in collaboration with Filipe Rodrigues [Rodrigues, 2010], under the joint supervision of my supervisor, Francisco Pereira, and myself. It received the national AI dissertation award (TleIA2010) and was published in 2012 GEOProcessing conference proceedings [Rodrigues et al., 2012]. The work developed primarily focused on the extraction, analysis, manipulation and classification of POIs. A POI is a specific point location that someone may find useful or interesting. POIs can be used in navigation, place descriptions, sociological studies, the analysis of city dynamics, the geo-reference of texts, etc. This kind of simple information structure can be used and enriched in such a way that context-aware systems behave more intelligently.

4.1.1 POI Sources

In spite of their importance, the production of POIs is scattered across a myriad of different websites, systems and devices, thus making it extremely difficult to obtain an exhaustive database of such wealthy information. There are hundreds, if not thousands, of POI directories in the Web, with POIs for places all over the World. The information provided is obviously not homogeneous, but all sources are templatized having the following fields generally available for each POI: Name, Address (and/or GPS Position), Category(ies), Official Web Site (optional). The sources studied are presented and characterized in table 4.1.

A clear distinction is made between local business directories and platforms that are based on social networks. In the first group the company or service referred to by the POI is usually introduced by the owner itself. Here, the information provided is usually more accurate since the company wants to be easily found by potential customers, and so the chosen categories are more precise. The information about POIs provided in the second group is collaboratively created by a social network. Each individual is free to create a third-party reference to a geo-referenced company or service. The information entered is not validated by any authority, and the categories used are generic ones, such as restaurant, theater or hospital. In short, business directories are more accurate and have more specific and structured category titles, as will be discussed in section 4.1.2. As a result, local business directories were used as a source to create a massive POI database, as described in the next section.

Special attention should be paid to Yahoo data; to the Dun & Bradstreet (D&B) commercial database [Dun & Bradstreet, 2011], created by a consultancy company that specializes in commercial information and insights for businesses; and to InfoUSA[4.1], provided by the Harvard Center for Geographic Analysis (ESRI Business Analyst Data). In the first source (Yahoo), the database was essentially built from user (owner) contributions. In the other two, the data acquisition process was semi-automatic and involved

[4.1] http://www.infousa.com

Name	Website	API	Coverage	Language
Yahoo	http://local.yahoo.com	Yes	US cities and cities from main EU countries	English
Manta	http://www.manta.com	No	Worldwide	English
City Search	http://www.citysearch.com	No	US cities	English
Yellow Pages	http://www.yellowpages .com	No	US cities	English
Boston Globe	http://www.boston.com/ bostonglobe/	No	Boston, MA	English
Upcoming	http://upcoming.yahoo .com	Yes	Worldwide	English
Yelp	http://www.yelp.com	No	Canada & US cities and cities from main EU countries	Location dependent
Sapo	http://mapas.sapo.pt	Yes	Portuguese cities	Portuguese
Páginas Amarelas	http://www.pai.pt	No	Portuguese cities	Portuguese
Dun & Bradstreet	http://www.dnb.com	No	Worldwide	English
InfoUsa	http://www.infousa.com	No	US cities	English

Table 4.1: POI sources studied: the first nine are public, while the last two are private and require payment for access.

the integration of official and corporate databases, statistical analysis and manual evaluation [Dun & Bradstreet, 2011]. These sources were later used in experiments in the automatic classification of POIs (section 4.1.5).

4.1.2 POI Taxonomies

There are distinct ways to organize categories of POIs provided from different sources. While some provide a flat list, or even a mix of labels arbitrarily chosen by users, others offer a more structured organization in the form of a taxonomy. Few of them follow or provide an external official classification system associated with POIs. At the same time, associations between POIs and these categories are not always unique, as one place can be classified within different categories, which may or may not be related. One example is the POI named “Extreme Rims & Sounds Incorporated”, from the Yahoo! Local web service. This POI is categorized under: Auto Conversions, Auto Detailing, Auto Body Shops, Race Car Parts, Vehicle Painting & Lettering; all descending from the common super category: Automotive. To deal with these issues, the following two subsections are presented. The first introduces official business classification systems adopted in different regions around the world. The second, as well as laying out the organization of categories of POI sources, also proposes an approach to specify only one top category that subsumes the others associated with a given POI in a given taxonomy.

Official Taxonomies

In business, classification systems serve to communicate important facts about a company. These codes are generally controlled by a governmental, a professional, a trade or an international standards organization. They often serve as shorthand for users interested in material in a particular area of industry or a specific business sector [Hodge, 2000]. The North American Industry Classification System (NAICS) [NAICS, 2011], the International Standard Industrial Classification (ISIC) [United Nations, 2011], and the Classificação de Actividades Económicas (the *Economic Activities Classification* - CAE)[CAE, 2011] are examples of official and standard POI classification systems. Table 4.2 presents a summary of these classification systems. All the responsible entities for these classification systems provide a complete listing of the categories online, and

Classification System	Countries Covered	Access	Taxonomy Depth	Number of Top Major Industry Sectors
NAICS	U.S., Canada and Mexico	Paid	5 levels	20
SIC	U.S. (replaced by NAICS in 2000)	Public	3 levels	10
ISIC	Worldwide	Paid	3 levels	21
CAE	Portugal	Public	6 levels	17

Table 4.2: Some of the business classification systems studied.

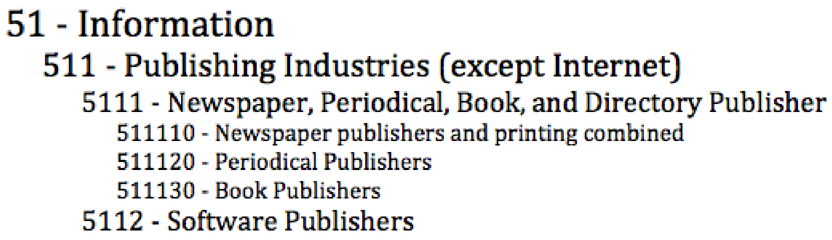
the mapping between them is also available[4.2].

Coding systems usually group industries in a hierarchy. At the top of the hierarchy are "major industry sectors". Although some coding systems have a different level of detail in taxonomy, all of them classify a company by its most profitable activity when this company is in different industry sectors. In our case, we will adopt the NAICS since the majority of the POIs in the database are from U.S. cities, mainly from Boston and New York.

The North American Industry Classification System (NAICS) is the standard system used by Federal statistical agencies in classifying business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy [Bureau, 2010]. The NAICS was developed under the auspices of the Office of Management and Budget (OMB), and was adopted in 1997 to replace the old Standard Industrial Classification (SIC) system.

The NAICS is a two through six-digit hierarchical classification code system, offering five levels of detail. Each digit in the code is part of a series of progressively narrower categories, and the more digits in the code, the greater the classification detail. The first two digits designate the economic sector, the third digit designates the sub-sector, the fourth digit designates the industry group, the fifth digit designates the NAICS

[4.2] The United Nations maintains a website containing international and national classification systems and their interconnection at http://unstats.un.org/unsd/cr/ctryreg/default.asp?Lg=1 industry, and the sixth digit designates the national industry. A complete and valid NAICS code contains six digits [Association, 2010]. Figure 4.1 shows part of the NAICS hierarchy.



51 - Information
511 - Publishing Industries (except Internet)
5111 - Newspaper, Periodical, Book, and Directory Publisher
511110 - Newspaper publishers and printing combined
511120 - Periodical Publishers
511130 - Book Publishers
5112 - Software Publishers

Figure 4.1: Example of the NAICS hierarchy

Categories Provided by Public POI Sources

Different classifications are used by providers to organize POIs in Public sources, ranging from simple labels freely associated with POIs to taxonomies with reasonable detail in grouping categories. A common particularity is shared between them, in that all category information is free to obtain, either by an available API or by screen scraping (such as the case with POI extraction 4.1.3)[4.3]. Table 4.3 shows a comparison between those POI sources first introduced in section 4.1.1.

From the category organizations studied, those that are structured in taxonomies are richer with POIs classified through different levels of specificity of economic activities. These taxonomies may be created by the company itself or may be collaboratively built by the suggestions of users (who may or may not be owners) who feed the system with new POIs. The Yahoo taxonomy is one of the few that allows multiple categorizations of a POI. This diversity is rich but sometimes also introduces very low granularity, high heterogeneity and high ambiguity. In these cases it is interesting to induce the *Least Common Subsumer*[Cohen et al., 1992] (or *lcs*) of these

[4.3] By screen Scraping (or Web Scraping) we mean to extract valuable information from web pages by using regular patterns. We only used this technique when it was allowed by the Terms of Use in question or when the given source was previously contacted. Furthermore, the use of Web Scraping was limited to search results in specific geographical areas and specific categories. Thus, we discourage the broad employment of such technique without the application of these rules.

POI source	Category organization	Multiple categories allowed	Provides official classification system	Classification Extraction
Yahoo	Taxonomy (depth 4 depending on the city)	Yes	No	Web Scraping
Manta	Taxonomy (depth 3 with greater compound names)	No	NAICS and SIC code	Web Scraping
City Search	Taxonomy (depth 3)	Yes	No	Web Scraping
Yellow Pages	Taxonomy (depth 2 with '-' meaning sub-specializations in some cases)	Yes	No	Web Scraping
Boston Globe	Taxonomy (depth 2)	No	No	Web Scraping
Upcoming	It uses Yahoo! Local to contextualize POIs.			
Yelp	Taxonomy (depth 3)	Yes	No	API
Sapo	Free labels chosen by users	No	No	API
Páginas Amarelas	Flat list (with '-' meaning specializations)	No	No	Web Scraping

Table 4.3: Category organizations provided across different Public POI sources.

categories, which corresponds to the most specific category in the taxonomy that subsumes the other categories associated with a given POI. For example, assuming the case of the POI "Pomodoro Rosso" on Columbus Avenue, New York, its categories are: Italian Restaurants; Carry Out & Take Out; Restaurants. As the first and the second categories are specializations of the last category, it is possible to conclude that $lcs(ItalianRestaurants, CarryOut\&TakeOut, Restaurants) = Restaurants$. However, in businesses with a greater diversity in products and services offered (like "Sears" located on Beverley Road, Brooklyn, with the categories: Opticians; Eyewear; Clothing Stores), it is sometimes not possible to select just one category as the Least Common Subsumer. In these cases, the only possibility is to verify if there are some categories that are specializations of others, merging them as the most general, which for the example just described would be: Eye Care and Clothing Stores.

Let $\mathcal{L}$ be a description logic and let $\Longrightarrow$ be its subsumption relationship, i.e. if description A subsumes B then we will write B$\Longrightarrow$A. By definition, $\Longrightarrow$ must be a partial order on (denotations of) descriptions. It is possible that A$\Longrightarrow$B and B$\Longrightarrow$A without A and B being syntactically or lexically identical; in this case, we will say that A and B are *semantically equivalent*, and write A≡B. C $\in \mathcal{L}$ is a *least common subsumer (LCS)* of A and B iff a) C subsumes both A and B, and b) no other common subsumer of A and B is strictly subsumed by C.

Definition 1. *If A and B are concepts in a description language $\mathcal{L}$ then lcs(A,B) is a set of concepts $\{C_1,\ldots,C_i,\ldots\}$ such that a) each C_i is a least common subsumer of A and B, b) for every C that is a least common subsumer of A and B, there is some C_i in lcs(A,B) such that $C_i \equiv C$, and c) for $i \neq j$, $C_i \not\equiv C_j$.*

The taxonomy of Yahoo's categories comprises approximately more than 1,300 distinct categories distributed across a hierarchy which can go 4 levels deep. This depends on the city studied and its respective coverage. For example, in New York there are categories (e.g. Indoor Air Quality, Miniature Golf) that are not seen in the Boston local directory and vice-versa (e.g. Baby Sitters, Boat Repair). Appendix D presents the general view of the Yahoo taxonomy integrating categories from New York and Boston.

Since in most cases no API is currently available from those local directories studied, a wrapper was created based on regular expressions, in order to automatically extract the category taxonomy from each local directory. Only the Yelp web site provides the complete list of categories[4.4], while Yahoo! Local only presents it through menu navigation across its web site. Curiously, this dynamic is also observed in the fact that this taxonomy is different depending on which city we are virtually visiting. Namely, Yahoo! Local builds its menus dynamically, thus presenting proper taxonomies for distinct cities as seen earlier. Through time, this taxonomy grows with new types of services and places. In this way, by using specific wrappers for each POI provider, it is possible to run them periodically in order to integrate new categories in the respective stored taxonomy.

4.1.3 POI Extraction

At this point, it is possible to conclude that there are in fact diverse templatized sources of information about POIs. However, each one uses its own format to represent the POIs and its own taxonomy to classify them. Also, the Web servers that provide POI information (e.g. Yahoo, Manta, Yellow Pages, CitySearch, Upcoming) are mere repositories and consequently do not take advantage of the full potential of such information.

Currently, there are no specialized tools for POI extraction that we know of. However, there are commercial web-scraping tools that allow users to visually select parts of web pages they want to extract, define repetition patterns and other options, and extract the selected information directly into a user-defined file or database. Besides the high price, these kinds of tools are not very flexible and so they cannot be applied in all situations. They do not deal with problems like hidden information in the HTML or in the JavaScript code, request limitations imposed by the server, information spread across multiple pages, poorly formatted web pages, authentication requirements, URL manipulation, etc. In summary, these tools are appropriate for small simple scraping jobs. For larger, highly customized tasks it was better and cheaper to develop our own scraper tool. Furthermore, the development time was not that long if a framework was used that automatically performed the connection handling, database management,

[4.4] http://www.yelp.com/developers/documentation/category_list and other required tasks.

Extracting POIs from Web resources is essentially a Web scraping task. There are multiple ways to scrape a web page, the most commonly used being XPath[4.5], based on string manipulation and regular expressions. The approach implemented used regular expressions, which makes it more robust and flexible. Other approaches were more error-prone and required more frequent patching and updating. The use of regular expressions also made the source code a lot clearer.

Since POI data is scattered across various sources in different formats 4.1.1, it was necessary to create a relational database schema that could accommodate POIs with different levels of information. Figure 4.2 presents an excerpt of an E-R model of the POI database (those tables which are particularly affected by the POI extraction process). This database was modeled to integrate all the information retrieved, extracted and enriched by the KUSCO system.

In many cases, POI websites provide a developer API, using REST or SOAP Web services, which makes the work of programmers easier. However, most of the websites still do not have such an API, and, what is worse, some of these websites have poorly formatted HTML, which makes the scraping task much harder.

Considering these specificities, we developed a multi-thread POI Extractor that is currently able to extract POIs from Yahoo, Manta, City Search, Yellow Pages, Boston Globe, Upcoming, Yelp, Sapo and Páginas Amarelas. The POI Extractor uses an API when it is available from the POI source, or otherwise, web scraping. This last step is done following the conditions of the terms of service of each POI source and taking the time/limitation of requests into consideration. An important feature worth mentioning is the ability of the framework to identify similar POIs, i.e. POIs from multiple Web sources that refer to the same place. The next section explains the POI-matching algorithm in detail.

[4.5] A technology used to navigate through elements and attributes in an XML document. XPath is a major element in W3C's XSLT standard - and XQuery and XPointer are both built on XPath expressions [Berglund et al., 2007]

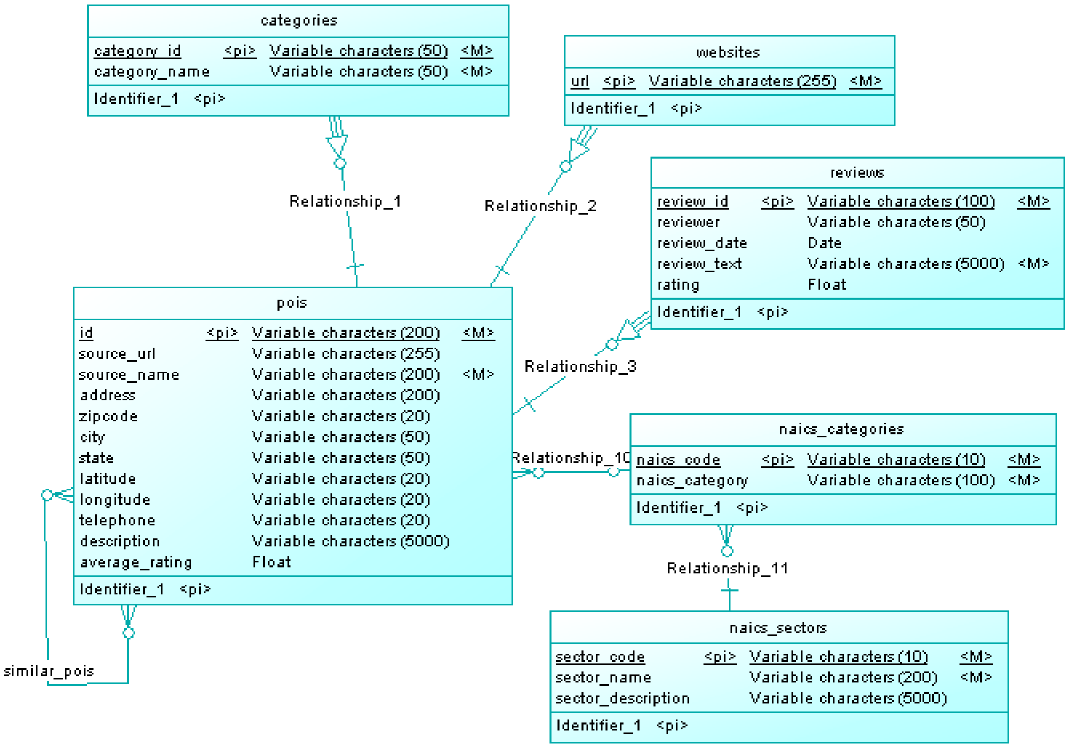

categories
category_id <pi> Variable characters (50) <M>
category_name Variable characters (50) <M>
Identifier_1 <pi>
websites
url <pi> Variable characters (255) <M>
Identifier_1 <pi>
reviews
review_id <pi> Variable characters (100) <M>
reviewer Variable characters (50)
review_date Date
review_text Variable characters (5000) <M>
rating Float
Identifier_1 <pi>
Relationship_1
Relationship_2
Relationship_3
Relationship_10
Relationship_11
pois
id <pi> Variable characters (200) <M>
source_url Variable characters (255)
source_name Variable characters (200) <M>
address Variable characters (200)
zipcode Variable characters (20)
city Variable characters (50)
state Variable characters (50)
latitude Variable characters (20)
longitude Variable characters (20)
telephone Variable characters (20)
description Variable characters (5000)
average_rating Float
Identifier_1 <pi>
naics_categories
naics_code <pi> Variable characters (10) <M>
naics_category Variable characters (100) <M>
Identifier_1 <pi>
naics_sectors
sector_code <pi> Variable characters (10) <M>
sector_name Variable characters (200) <M>
sector_description Variable characters (5000)
Identifier_1 <pi>
similar_pois

Figure 4.2: POI database fed by POI Extraction

4.1.4 POI Matching

Here, we propose an algorithm for POI matching that makes use of a string comparison library to identify similarities between POIs from different sources. When we were extracting POIs from multiple sources, it was important to have a way to identify similar POIs, so that we did not end up with redundant information and also to collect as much information as possible about a given place. This required a way to identify similarities based not only on proximity, but also on name likeness. Since most of the POI sources are dependent on the contribution of the user (i.e. the POI's owner, submitting and updating information about places), it was not feasible to only rely on a naïve string comparison between POI names. Also, in POI matching, it is crucial to have a high level of precision and good recall results. The ultimate objective is to get a good recall while keeping the precision close to 100%.

The approach here proposed for POI matching made use of the Jaro-Winkler string metric (equation 2.3) to identify close names, while ignoring spelling errors and some abbreviations. The URL of the POI's official website, when available, was also used to identify not only matches that were not recognized by using only the proximity and the name similarity but also some mismatches that might otherwise have been considered matches.

It is important to note that the similarity between POIs was heuristically computed. Some thresholds were then defined that enabled a certain level of confidence when deciding whether they referred to the same place or not. Taking this into account, two POIs were considered similar by the algorithm if they fitted into one of the following groups:

- The distance between the two POIs is less than 80 meters, the name similarity is above 0.70 and one or both POIs do not have website information.
- The distance between the two POIs is less than 80 meters, the name similarity is above 0.70 and the website similarity is higher than 0.60.
- The distance between the two POIs is less than 80 meters, the name similarity is above 0.60 and the website similarity is higher than 0.95.

These various thresholds were obtained through an analysis of results from initial experiments that were done using low thresholds and which were followed by an iterative process of threshold tuning, experimentation and evaluation of the results obtained. The tuning process was not completely blind since, besides comparing the changes made by the previous tuning, the similarity values were obtained for the different metrics (i.e. name and website likeness and Euclidean distance of the location). Therefore, for each iteration there was a rough idea where the threshold should be.

At first, it was also considered that the POI categories might be a possible improvement of the algorithm. However, it was soon realized that in doing so, although this might improve recall, it would certainly have reduced precision due to the lack of coherence between the taxonomies of the different POI sources. The only possibility would have been to use a common taxonomy to further refine the matches. Unfortunately, that would have required both POIs to be classified in the same taxonomy. In the next section, classification of POIs into taxonomies is explained in detail.

4.1.5 Automatic POI Classification

In this section, different approaches were implemented and tested to classify POIs into a common taxonomy. Since a given source of POIs generally has a proper taxonomy of POI categories, each one classifying them by related terms (e.g. Entertainment Venues or Live Theaters), it was necessary to select a common classification to use. The aim of this focused research was to be able to classify POIs from different sources in accordance with a common and more widespread taxonomy like NAICS and ISIC. This was essential in order to perform a proper analysis of the extracted POIs. If the POIs were not mapped to a common taxonomy, we would not be able to determine, for instance, how many POIs of museums exist in a given area because one POI source may classify them only as "Museums" and another, with more information, as "Art Museums & Galleries".

The approaches proposed here, to automatically classify POIs in a given taxonomy, relied mainly on category information from POI sources which were also organized in taxonomies (table 4.3). Because the same POI was sometimes assigned to more than one category in some POI sources, the number of possible combinations could have been huge. As a consequence, finding mappings between the source taxonomy and the target taxonomy was not always a simple task. Consider the following mappings:
"Newspaper Publishers" -> "Newspaper Publishers"
"Newspapers Printing" -> "Newspaper Publishers"
"Laboratories" -> "Research & Development in Biotechnology"

Even though the first mapping is obvious, the other two are not, especially the last one. This is mainly due to the different levels of granularity in category names. The fact that the POIs can belong to multiple categories can help to differentiate target categories, thus fixing some granularity issues. However, sometimes this is not enough, and, in order to determine such mappings, additional information about the POIs would be necessary in order to find the correct mapping.

The aim was to classify POIs in accordance with more widespread taxonomies like NAICS, ISIC or CAE. Among these, we chose to classify POIs in accordance with the NAICS code mainly because most of the data available was from North America, particularly from Boston and New York. Notwithstanding this, it would be possible to apply the same approaches proposed here to ISIC or CAE, since the methodology would be analogous.

In some experiments (appendix C), POIs were classified into different NAICS levels (i.e. NAICS categories with different granularities), particularly two, four and six-digit NAICS codes. This choice is typical in Urban Planning, depending on the type of study in question (e.g. level 2 allows an analysis of economic sectors, while level 6 goes to the level of the establishment specificities).

Different approaches were studied, including Ontology Matching and Lexical/Semantic Similarity, but the most profitable was the application of machine-learning (ML) algorithms to automatically classify the NAICS code of a given POI. This approach was subdivided into the following two different techniques which varied whether they extended the data used by ML algorithms:

- *Flat Classification:* Weka [Hall et al., 2009] is a large collection of machine-learning algorithms for data-mining tasks. In this approach, ML was applied using different models such as Bayesian networks, tree-based learners, instance-based learners (lazy learners), rule-based learners and neural networks. Table 4.4 provides a brief description of the Weka algorithms tested. It is beyond the scope of this thesis to describe any of the algorithms in detail. The interested reader is redirected to the dedicated literature (e.g. [Mitchell, 1997; Witten and Frank, 2005]).
- *Extension with Semantic Annotations:* Describing POIs only by one or two concepts (their categories) may not be sufficiently diverse in order to help data-mining algorithms to classify more precisely. The use of semantically enriched information about places incorporating more attributes in order to refine POI details was

also proposed. This enrichment was made by the KUSCO system and it is presented in detail in section 6.3.1 of chapter 6, which basically consists of gathering textual descriptions available on Wikipedia and applying Information Extraction and NLP techniques on them (for detail, see chapter 5). As one POI may belong to more than one category, all its category descriptions were processed. These tags were also semantically contextualized in WordNet [Miller et al., 1990] in order to aggregate synonym tags in just one entry. Table 4.5 shows some examples of the POI indexes produced.

As one would expect, the computational complexity of the classifiers increased many times, depending on the base machine-learning algorithm used. As a result, it was not possible to test some of the more computationally intensive algorithms. However, as was done for the flat classification, different types of machine-learning algorithms were tested, such as: Bayesian networks, tree-based learners, instance-based learners and rule-based learners. It was not possible to test neural networks due to their computational demands for this specific classification problem, both in processing power and memory. Experiments done using these algorithms are presented in appendix C.

4.2 Retrieving POI Information

It became necessary to enrich the data available from public or commercial POI sources with information about POIs from the Web. Initially, the entire Web was used to retrieve such information but the high level of noise obtained from the open-ended web pages led to the approach of selecting specific and constantly updated repositories. We applied several approaches: location-based Web search; locating the About page of a POI; and retrieval of Wikipedia related articles. Since the information retrieved about a place is in the form of open-ended, semi- or fully-templatized documents (free texts, structured Wikipedia articles or web pages respectively), it is the raw material for the next phase in the Semantic Enrichment process, which will be studied in the following chapter.

4.2.1 Location-based Web Search

This approach, named Location-based Web Search, is responsible for finding web pages related to POIs. It is important to differentiate this aim from the aim of searches made

Implementation	Description
FT	Classifier for building 'Functional trees', which are classification trees that could have logistic regression functions at the inner nodes and/or leaves.
ID3	Unpruned decision tree based on the ID3 algorithm.
J48	Pruned or unpruned C4.5 decision tree.
J48graft	Grafted (pruned or unpruned) C4.5 decision tree.
RandomForest	Forest of random trees.
RandomTree	Tree that considers K randomly chosen attributes at each node. Performs no pruning. Also has an option to allow estimation of class probabilities based on a hold-out set (backfitting).
DecisionTable	Simple decision-table majority classifier.
JRip	Propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP.
IBk	K-nearest neighbors classifier. Can select appropriate value of K based on cross-validation. Can also do distance weighting.
IB1	1 - nearest-neighbor classifier. Simplification of IBk.
K*	K* is an instance-based classifier, that is, the class of a test instance is based upon the class of those training instances similar to it. It differs from other instance-based learners in that it uses an entropy-based distance function.
BayesNet	Bayesian Network
NaiveBayes	Naive Bayes model
Multilayer Perceptron	A classifier that uses back-propagation to classify instances.
ZeroR	Predicts the mean (for a numeric class) or the mode (for a nominal class).
OneR	Uses the minimum-error attribute for prediction, making numerical attributes discrete.

Table 4.4: Brief description of each Weka algorithm tested

POI Name	Yahoo Categories	Tags
Willett Institute of Finance	Financial Planning, Investment Services	income, shares, Security Analysis, plan, Cash
W3 Edge LLC	Computer Communications, Web Services	network, software system, devices, servers, wireless technologies
Curis Incorporated	Doctors & Clinics, Laboratories, Medical Laboratories	hospital, General practice, chemistry, clinic, Clinical laboratory

Table 4.5: Some tags produced by KUSCO.

by local search engines that are generally POI-oriented. In others words, here the POI already exists and the aim is not to find POIs in a given area, but instead to gather more information about them available on the Web. Since local search engines like Yahoo! Local, GoogleMaps and so on, do not retrieve web pages related to a given location, the approach proposed here took advantage of a general Web search engine using only POI data as keywords: place name and geographical address. This last element is composed of the City name (where the POI is located) and it was initially obtained from Gazetteers [4.6] given a latitude/longitude pair. But recently, this information has already been available from the directory where the POIs were extracted. This search was made by the freely available Yahoo! Search API. KUSCO applies a heuristic that uses the geographical reference as another keyword in the search. Thus, assuming a POI is a tuple (Latitude, Longitude, Category, Name), the final query to the search is: <City Name> <Name>. Other attempts using other parts of the POI address were employed in the query, i.e. with a deeper level of granularity (e.g. street name and zipcode) but, as expected, in most cases pages related to the contacts in the company were retrieved, even when the official web site was found. Instead of contacts web pages, other pages in the same website would have given more relevant information about the POI (as explained later in section 4.2.2).

[4.6] A geographical dictionary (as at the back of an atlas) generally including position and geographical names like Geonet Names Server and Geographic Names Information System [GNS, 2009].

To automatically select only specific pages centered on a given Place, KUSCO also filters out generic Web Pages applying the following heuristics:

1. The title must contain the POI name;
2. The page body must contain an explicit reference to the POI geographical area (i.e. the city name);

The first heuristic follows the premise stated by Manning [Manning et al., 2008] that, generally, document zones may suggest different importance levels in a text, such as title (having more weight in the extraction task). Thus, the hypothesis is that, if the title refers to the POI name, it is likely that the page is focused on the given POI. The second heuristic relies on KUSCO's sensitiveness to geographical location of Public Places. For example, after looking for specific Web information for a given POI named "Carnegie Hall" in New York, the system found many relevant results all referring to the same place: a concert venue. In another example, given a POI in the same city about "Mount Sinai" (a hospital), a geographic-based local search resulted in other definitions different from a hospital, such as a metropolitan neighborhood. This demonstrates that this approach can become very dependent on search algorithms and on the Web's representativeness of places.

Once a threshold (N) has been established, the top N most relevant pages are selected (as suggested by the search engine) at the end of this process. A pitfall to avoid occurs when the search returns duplicate content referring to the same web page, copies which were stored in different Web Servers. A more difficult problem is to find *Near-Duplicates* that use different character sets, formats, or inclusions of advertisement or current date. By some estimates, as many as 40% of the pages on the Web are duplicates of other pages [Manning et al., 2008].

Search engines try to avoid indexing multiple copies of the same content to keep storage and processing overheads down. Both Altavista (now owned by Yahoo) and Google have been awarded US patents [4.7]that improve on existing methods for classifying duplicate content. The secret is to make comparisons quickly without doing

[4.7] respectively for Altavista: 5,970,497 and 6,138,113 , and for Google: 6,615,209 and 6,658,423

Algorithm 4.1 Location-based Web Search: Find POI related web pages

```
Input: POI instance with name composed of ${token_1 \ldots token_n}\vee$
  $address \Leftarrow$ {Number, Street, Zipcode, City, State, Country}
Output: N Web Pages

      $queryList \Leftarrow$ {``City'' + ``$token_1 + ... + token_n$'';
                        ``City'' + ``$token_1$'' + ... + ``$token_n$'';
                        ``$token_1 + ... + token_n$''}

  $pagesFound \Leftarrow \emptyset$
  $filteredPages \Leftarrow \emptyset$
  for all $query_i \in queryList$ do
    if $(|filteredPages| \leq 2)$ then
      $pagesFound \Leftarrow$ top-40 of Yahoo!  Search($query_i$)
      $filteredPages \Leftarrow$ Validate($pagesFound, poi$)
    else
      break
    end if
  end for
  return $filteredPages$

  function Validate($pagesFound, poi$): $filteredPages$
  $filteredPages \Leftarrow \emptyset$
  for all $page \in pagesFound$ do
    if ($page.title \supseteq poi.name$ and $page.body \supseteq poi.city$) then
      insert $page$ into $filteredPages$
    end if
    if $(|filteredPages| = 10)$ then
      break
    end if
  end for
  return $filteredPages$
  end function
```

any kind of word-by-word matching. One of Altavista's patents looks for similarities in the outbound links on a page. Two pages are selected, a first page and a second page. For each selected page, the number of outgoing links is determined. The two pages are marked as near duplicates based on the number of common outgoing links for the two pages. Besides this, a method that can eliminate near-duplicate documents from a collection of hundreds of millions of documents, by computing independently a vector of features less than 50 bytes long for each document and comparing only the vectors rather than entire documents, has been presented by Andrei Z. Broder [Broder, 2000]. Provided that m is the size of the collection, the entire processing takes time $O(m \log m)$. This algorithm has been successfully implemented and applied in the context of the Altavista search engine since then and is currently applied in Yahoo! Search [Kumar and Govindarajulu, 2009]. Considering that the problem of near duplicates is dealt with by the search engine used (in our case the Yahoo! Search API), we present here the Algorithm 4.1 proposed and used by KUSCO to retrieve web pages related to a given POI.

4.2.2 The *About Page* on a POI Official Website

Instead of considering all web pages from different domains, this approach restricts this universe to the Official Website of a POI when it is available from POI sources. We try to find in the POI Official Website the information which is focused on the purpose, services offered or mission, or in other words, what the POI itself is. An attempt to solve this problem was made by searching in the given website for a specific statement including that information. After some observations, we concluded that all the information the system needs is quite often found on the "About Web Page" or "Info Page" which concisely presents the company/POI to visitors.

With the availability of general search engines like Google[4.8] to restrict a web search to a given website, the approach to retrieve the About Page of a given POI basically consisted of using this restricted search on the POI Official Website with different combinations of queries around "About us", "About this company", etc. As there are no strict rules followed by web designers about the name and location of the About Page in any given Web site, two heuristics were applied in order to find and filter

[4.8] An example is given at Google Domain Search which is also available as an API.

possible candidates related to these characteristics respectively: the keywords used to find the About Page inside the web site were, in order of preference, the words “About us”, “About” +“us” and “about”. The About Page was assumed not to be at a deep level within the web site.

4.2.3 POI Information on Wikipedia

An alternative to the World Wide Web and its plentiful heterogeneity in web page structures is Wikipedia. Wikipedia has a known document template which is used by editors and one is advised to follow the good practices [4.9] suggested by the Wikimedia Foundation (responsible for the Wikipedia project).

Wikipedia provides an API to access its content [4.10], with specific methods to search for articles. They actually make the entire Wikipedia database public, free to be downloaded and used. In this way, beyond its size and continuous evolution, Wikipedia is completely available to be explored for academic purposes. The following subsections present two different approaches proposed to retrieve POI-related articles from Wikipedia.

4.2.3.1 Category Articles

In this approach, it is proposed to retrieve information on Wikipedia about the category(ies) of a POI. Local POI directories are normally structured in a hierarchical tree/taxonomy of categories (section 4.1.2) instead of in a flat way. How they are defined is dependent on each provider. Their taxonomy may be created by the company itself or be collaboratively built by the suggestions of users who feed the system with new POIs. Neither a rigid organization nor a consistent validation of such a taxonomy is assumed here. Accordingly, node duplication and multiple inheritance may be a reality that a generic methodology must face. In fact, in our database it is normal for each POI to have multiple categories. In our approach, it is not mandatory to have a root node linking all main categories at the base of the taxonomy, and hence it is possible to process different types of taxonomies. A main category is that which does not have

[4.9] Available at http://en.wikipedia.org/wiki/Wikipedia:Writing_better_articles

[4.10] http://www.mediawiki.org/wiki/API any more generic category subsuming it.

Since some POI sources are well structured, that is, they have a hierarchical taxonomy of disjunct categories, we believe that the automatic tagging of places, initially considering the category they belong to, is a first step in labeling that place using encyclopedic knowledge (Wikipedia). As this is a rich source of common-sense knowledge, contextualizing category names in their corresponding Wikipedia articles is based mainly on string similarity between them. We opted for a top-down approach, beginning with the main categories and continuing until the taxonomy leaves were reached. To increase the confidence in this process, we chose to disambiguate the main categories manually at first and to make sure that at least a more specific category would be connected to the wikipages of its hypernym. When a POI has many categories, the system obtains the articles for each one and considers the union of all the resulting articles as the source of analysis. Since there are many different combinations of categories, it is possible to guarantee that each POI is subjected to its own specific flavor of category analysis. Algorithm 4.2 presents the process for disambiguating the complete category taxonomy.

Every search made using the Wikipedia API tries the original name of a category first. However, as has been observed many times, categories are generally compound names or variations of a lemma[4.11]. Some linguistic patterns are applied (line 10) to break down compound names that otherwise could not be mapped to any Wikipedia articles. For example, the category name "Computer & Electronics" derives two category names, "Computer" and "Electronics", which are stored in an ordered list just after the category's original name. More complex interpretations are also made, as in "Lesbian, Gay, & Bisexual Bars" which produces: "Lesbian Bars", "Gay Bars" and "Bisexual Bars". These patterns also include the lemmatization of category names (whether compound or not), by looking at WordNet [Miller et al., 1990] dictionary entries. As stated above, since WordNet is not a good resource for finding compound names, this inflection is normally applied to the head noun, so in the case of "Sports Cars" it would be "Sports Car".

[4.11] In linguistics, a lemma is the basic form of a word, for example the singular form of a noun or the infinitive form of a verb, as is shown at the beginning of a dictionary entry [Crowther, 1998].

Algorithm 4.2 Category Disambiguation: Find related Wiki pages to each category

```
Input: taxonomy$rootNodes \vee$manually disambiguated main categories $rootTitlesList$
Output: fully disambiguated categories $resultTitlesList$
  for all $node_i in rootNodes$ do
    $childrenList \Leftarrow$ GetDirectChildren($node_i$)
    for all $child in childrenList$ do
      $resultTitles \Leftarrow$ DisambiguateDescendentTree($child, rootTitlesList_i$)
      insert $resultTitles$ into $resultTitlesList$
    end for
  end for
  return $resultTitlesList$

  function DisambiguateDescendentTree($node, ascendantTitles$): $resultTitles$
  $uniqueNames \Leftarrow$ LinguisticPatterns($nodeName$)
  $wikiTitles \Leftarrow \emptyset$
  for all $uniqueName_i in uniqueNames$ do
    $wikiTitles \Leftarrow wikiTitles \bigcup$ QueryWikiAPI($uniqueName_i$)
  end for
  if ($wikiTitles = \emptyset$) then
    $wikiTitles \Leftarrow \{ascendantTitles\}$
  end if
  $resultTitles \Leftarrow \{(node, wikiTitles)\}$
  if (isLeafNode($node$) is false) then
    $childrenList \Leftarrow$ GetDirectChildren($node$)
    for all $child in childrenList$ do
      $resultTitles \Leftarrow resultTitles \bigcup$ DisambiguateDescendentTree($child, wikiTitles$)
    end for
  end if
  return $filteredPages$
  end function
```

4.2.3.2 POI Article

While the above approach is centered on place category, in this approach the focus is on place name. Once more, string similarity is used to match place name to a Wikipage title in order to find the Wikipedia description for a given place. At first glance, this method is efficient in mapping compound and rare place names such as “Beth Israel Deaconess Medical Center” or “Institute of Real Estate Management’. However, it could naively induce some wrong mappings for those places with very common names (e.g. Highway - a clothing accessories store in New York, Registry - a recruitment company in Boston, Energy Source - a battery store in New York). This problem was tackled by determining the specificity of place names and considering only those with high Information Content (IC)[Resnik, 1995]. The Information Content of a concept is defined as the negative log likelihood, $-logp(c)$, where $p(c)$ is the probability of encountering such a concept. For example, ‘money’ has less information content than ‘nickel’ as the probability of encountering the concept of ‘money’, $p(Money)$, is larger than the probability of encountering the concept of ‘nickel’, $p(Nickel)$, in a given corpus. For those names present in WordNet (e.g. Highway, Registry), the IC is already calculated [Mihalcea, 1998], whereas for those not present, it is heuristically assumed that they would only be considered by the system if they were not a node in the Wikipedia taxonomy, i.e. a Wikipage not representing a Wikipedia category (as in the case of Energy Source), but only being a Wikipedia article.

5

Automatic Tagging of Texts Relying on Geography

The information related to POIs can be obtained from different sources on the Web as explained earlier (chapter 4). This chapter presents the process of extracting relevant terms from this information. To support the chapter, in the first section we give an example which will be used in the remaining sections. Then, in the second section, Natural Language Processing techniques are applied to textual descriptions in order to extract key-phrases related to a given POI. In the third section, contextualization and integration of information is done because words are naturally ambiguous and it is essential to detect duplication of concepts. Finally, in the last section, the most relevant concepts are selected to build the index of the original POI under investigation.

5.1 An Example

To follow the whole process of semantic enrichment of POIs described here, we propose, as a first example, a university with a bar/restaurant, using location-based web search to find related web pages (section 4.2.1). More examples are contextualized and presented in appendix A. At first, the POI is only a point in the map having a name associated with it, for example, $(42.339078, -71.099239, \texttt{SimmonsCollege})$. The reverse geo-coding gives us the city to which it belongs, so the next step is to browse the Web using the Yahoo! Search API with the following queries in the format [City] + POI Name: *"Boston"+"Simmons College"*. From this query, a set of relevant pages is retrieved and downloaded for the next phase.

From 40 web pages only 10 at most are selected following the criteria described in section 4.2.1. Table 5.2 presents the web pages selected for this POI with the relevant summary presented in Yahoo! Search results. Due to the limitations of page size, only the domain for each page is provided. (For interested reader, a simple search query will yield the exact page). It is also important to note that, as the Web grows and changes every day, this result set will be different in future searches.

5.2 Term Extraction

Having a set of pages as input, the Meaning Extraction module extracts a ranked list of terms. Our hypothesis is that these terms are the main concepts of the place represented by the input POI. Figure 5.1 presents its overall architecture. This process starts with Noun Phrase chunking and Named Entity Recognition (NER) using available Natural Language Processing (NLP) tools (section 2.2). Linguistic analysis of text typically proceeds in a layered fashion. Texts are broken up into paragraphs, paragraphs into sentences, and sentences into words. A sentence analyzer identifies the boundaries of sentences in a document and a tokenizer decomposes each sentence into tokens. Tokens are obtained by splitting a sentence along a predefined set of delimiters like spaces, commas and periods. A token is typically a word or a digit or a punctuation mark.

Words in a sentence are then tagged by the Brill's *Part-of-Speech* (POS) *tagger* [Brill, 1994] which labels each word as a noun, verb, adjective, etc. For this, an implementation of this tagger in Java [Lin, 2004] is used. A *Noun Phrase chunker* [Ramshaw and Marcus, 1995] is then applied in order to identify every group of words with a head noun which functions together just as a single term, with a Java implementation of the Mark & Ramshaw's algorithm [Greenwood, 2005] being used.

At the same time, the original text is also processed by a Named Entity recognizer [Finkel et al., 2005] to identify proper names in the text. We use the Stanford NLP group's implementation[5.1]. This NER is a three-class CRF-based implementation that

[5.1] Available at http://nlp.stanford.edu/software/index.shtml

ID	Title	Domain Url
A	*Simmons College - Boston, Massachusetts*	www.simmons.edu
Simmons College is a nationally distinguished, small university in the heart of Boston providing exceptional graduate and undergraduate study integrated with career preparation...		
B	*Simmons College (Massachusetts) - Wikipedia, the free ...*	en.wikipedia.org
Simmons was founded in 1899 with a bequest by John Simmons to educate women, so they could have an independent livelihood. Simmons is a member of the Colleges of the Fenway consortium...		
C	*Simmons College - Boston, MA, 02115 - Citysearch*	boston.citysearch.com
(617) 521-2000, 300 Fenway, Boston, MA 02115 Years in business Established in 1899 Last updated 8.25.10 Category: Colleges , K-12 Schools , Private & Parochial ...		
D	*Simmons College near Boston, MA*	local.yahoo.com
simmons college in Boston, MA on Yahoo! Local Get Ratings & Reviews on simmons college with Photos, Maps, Driving Directions and more.		
E	*Simmons College in Boston, Massachusetts ...*	www.earnmydegree.com
Find detailed information about Simmons College in Boston, MA.		
F	*Simmons College in Boston, Massachusetts*	www.cappex.com
Simmons College in Boston, MA. See Simmons College facts, admissions, cost, financial aid, programs, majors, sports, campus life info on Cappex.com.		
G	*Simmons College*	www.insiderpages.com
For the person who said they didn't like the college because the students made comments about other students in class is...		
H	*Simmons College - Boston, MA*	www.yelp.com
7 Reviews of Simmons College "I'm about to start my second year in the GSLIS (library school) program. After a terrible undergrad experience at the corporation calling itself Boston University, I really do enjoy attending Simmons. The classes are"...		
I	*Simmons College Food Delivery ...*	www.grubhub.com
GrubHub.com: Find restaurants that deliver to Simmons College in Boston. Browse delivery menus, reviews, and coupons. Order online or by phone.		
J	*Simmons College — Facebook*	www.facebook.com
Welcome to the official Facebook Page about Simmons College. Join Facebook to start connecting with Simmons College.		

Table 5.1: The most relevant pages obtained by Yahoo.

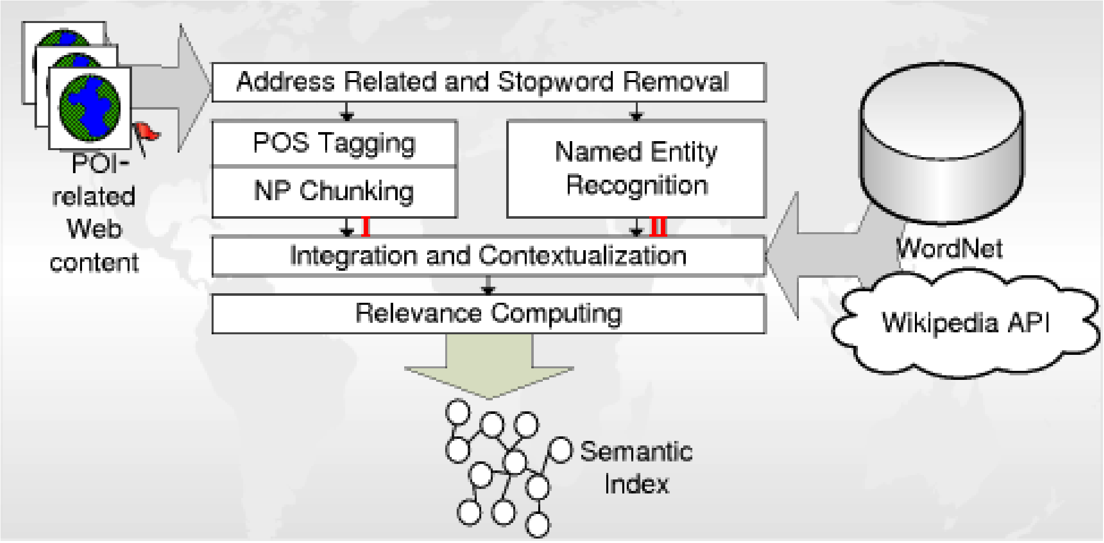


Figure 5.1: Meaning Extraction in KUSCO architecture

can label entities as PERSON, ORGANIZATION, and LOCATION. This serialized classifier was trained on data from CoNLL, MUC6, MUC7, and ACE (section 2.4.4). This model is considered robust for the two national varieties of English (British and American), because it was trained on both US and UK newswires. Despite knowing that performance depends on many factors (e.g. training and test data, initial parameters, or the feature set of the CRF), Finkel et al. present the statistics for the version of classifier used by KUSCO's Meaning Extraction module from one machine on one test set:

ner-eng-ie.crf-3-all2006-distsim.ser.gz			
Memory:	320MB		
Class:	PERSON	ORGANIZATION	LOCATION
F-measure:	91.88	82.91	88.21

In the Meaning Extraction module (see figure 5.1), noun phrases (flow I, which applies POS tagging and NP chunking) are represented by common nouns while the entities (flow II, which applies NER) are represented by proper nouns. Each instance in both groups is conceptualized using single or compound nouns. Instead of being limited by only the solutions provided in a given NLP framework such as those presented in section 2.2 in chapter 4, it was decided to apply the most successful and adopted algorithms to each NLP task. Furthermore, a widely used stopword list[5.2] is consulted

[5.2] Available at http://snowball.tartarus.org/algorithms/english/stop.txt to filter future concept candidates. Besides this, some heuristics are used to determine the validity of each term before the next phase, that of information integration, begins:

- Terms must be no more than 3 words long (only for NPs) and have at least two letters.
- The head noun of a valid noun phrase cannot be a stopword, must start with a letter and must contain at least two consecutive letters.
- No words in noun phrases can have a change of a case inside (e.g. “lOve” is not valid, while “Love” is).
- The head noun of valid noun phrases cannot contain numbers or unusual characters like {= _;!. :?,)(”/||& >< []*@%$#}.

For the example above, the system extracted the following entities (flow II) from web page A in the table : *Simmons Black Student Organization, BSO.*

Using the same text excerpt, the system selected the following list of NPs (flow I): *the beginning, Black History Month, style, the cake, ceremony, those lucky, a free piece, the BSO, several events.*

As a result, the term index produced comprises the union of Named Entity terms and noun phrase terms. This index is a weighted term vector that is semantically contextualized in the next phases. From our point of view this is simpler than inverted indexes since our final objective is the representation of place by its relevant concepts and not the retrieval of places.

As the same text excerpt may be at the same time both a noun phrase and a named entity, integration of this information has to deal with and avoid such duplication. Algorithm 5.1 presents the order in which NPs and NEs found in the text are selected. Since not all terms identified are relevant to this POI, some weighting scheme must be applied to select the most prominent characteristics of places. The next sections present the contextualization, integration and the computing of relevance of terms found in this phase.

Algorithm 5.1 Term Extraction: extract candidate terms (NPs or NEs)

```
Input: $T$ : text to analyze
Output: $NPs$ : set of noun phrases found $\vee$ $NEs$ : set of NEs found
  {concurrent execution}
  $POStaggedText \Leftarrow$ BrillTagger($T$)
  $NEs \Leftarrow$ StanfordNER($T$)
  $NPs \Leftarrow$ Marcus&RamshawNounPhraseChunker($POStaggedText$)
  for all $ne_i \in NEs$ do
    for all $np_j \in NPs$ do
      if ($np_j \supseteq ne_i$) then
        remove $np_j$ from $NPs$
      end if
    end for
  end for
  return $NPs, NEs$
```

5.3 Word Sense Disambiguation

In this phase we have to consider terms as *concepts* instead of as simple lexical terms. Concepts have meaning and are contextualized by disambiguating the terms found. To solve this problem we treat differently common nouns (generally denoting concepts) from proper nouns (generally Named Entities found), as can be seen in the following subsections. The integration of concepts and entities contextualized in different lexical resources is also proposed in order to have a common criterion of comparison when computing the semantics of entities and common concepts.

The sense of each term is acquired by using external semantic resources in order to contextualize each term previously found. As noted above, WordNet [Miller et al., 1990] is a widespread lexicon and common-sense ontology since, as well as behaving as a dictionary, it also provides semantic relations between concepts (which are called a *synset*, or in other words, a family of synonymous words)[5.3]. These nouns are contextualized on WordNet and thus can be considered not only as a word but more cognitively as a concept (specifically a *synset*). Given that each word present in WordNet may

[5.3] WordNet was presented in detail in section 2.5.1.1.

have different associated meanings, its most frequent sense is selected to contextualize a given term. All words (compound or not) are first searched in the original form, and if no entry in WordNet is found, its lemma is looked up next. For instance, when the system looks up the concept associated with the word “financial aids”, there is no exact correspondence to this form, but there is an entry for the singular “financial aid”.

Although many approaches have been studied in Word Sense Disambiguation (see section 2.5.2), none of them has achieved a satisfactory level of accuracy. Moreover, the heuristics of taking the most common sense of a given word shows better results that computationally intensive and complex techniques [Navigli, 2009]. Bearing this in mind, WordNet is used here to map words to concepts choosing in each case the most probable sense through corpus occurrence. For example, the term “library” has (at least) two meanings in WordNet: “a room where books are kept” or “a collection of standard programs and subroutines that are stored and available for immediate use”; the first meaning is the most frequently used if we consider statistics from a corpus. The corpus used is the SemCor corpus which was especially developed for WordNet and which is manually annotated with *synset* occurrences [Mihalcea, 1998].

WordNet is not intended to be a complete resource. We can still check if a given noun phrase corresponds to an instance in WordNet, but even for compound nouns the lack of information is broadly recognized. To solve this problem, we use Wikipedia. This is, in our view, the best open common-sense resource suited for compound noun disambiguation, particularly for named entities. If, on the one hand, it is continuously growing by integrating contributions from authors, on the other hand it has a very complete list of variations of the name of a given entity or compound noun. To disambiguate proper or compound nouns not found in WordNet, the same approach is used here. The most common use given to a word is queried in Wikipedia. If it returns a *disambiguation page*, then the system simply continues with the term decontextualized.

In Table 5.2, we show the application of this process to our Simmons College example. There are some points that must be explained: in the case where words can be both recognized as Named Entities and Noun Phrases, which prevails? In which order should the disambiguation begin, in WordNet or in Wikipedia? If only a part of an NP

WordNet concepts	
cake	baked goods made from or based on a mixture of flour, sugar, eggs, and fat
ceremony	the proper or conventional behavior on some solemn occasion
piece	a portion of a natural object
style	a way of expressing something (in language or art or music etc.) that is characteristic of a particular person or group of people or period
women's	an adult female person (as opposed to a man)
Wikipedia terms	
Black History Month	http://en.wikipedia.org/wiki/Black_History_Month
Boston	http://en.wikipedia.org/wiki/Boston
graduate_school	http://en.wikipedia.org/wiki/graduate_school
Massachusetts	http://en.wikipedia.org/wiki/Massachusetts
Ambiguous or, not found, terms (even in Wikipedia)	
BSO, Simmons Black Student Organization	

Table 5.2: WordNet meaning or Wikipedia mapping for some concepts or some proper nouns associated with "Simmons College"

is considered an NE, is the whole NP discarded? When duplicates are detected, which one remains? To answer theses questions, Algorithm 5.2 explains term disambiguation in detail and Algorithm 5.3 presents how duplicates are compared and counted. Both algorithms are applied in the Meaning Extraction module.

For each noun phrase found in a text, the corresponding mapping is retrieved using WordNet (see function in line 9 of the algorithm 5.2) and Wikipedia (see function in line 15 of the same algorithm). While in Wikipedia the only restriction is to guarantee that a given term is not ambiguous (i.e. the obtained sense is not a disambiguation page), for WordNet, the sense chosen has to be the most frequently used in the SemCor corpus and it is only considered if it does not refer to an instance of something. As an example, consider the terms "Secession", "Human Genome Project", "Leviticus",

Algorithm 5.2 Noun Phrase Disambiguation: map a Noun Phrase to a lexical resource entry

```
Input: $np$ : noun phrase
Output: $map$ : mapped entry in lexical resource with a term count associated $\vee$ $SemCor$ : Anotated WordNet Corpus
$WNmap \Leftarrow$ WordNetDisambiguation($np, SemCor$)
$WKmap \Leftarrow$ WikipediaDisambiguation($np$)
if (NumberOfWords($WKmap$) $>$ NumberOfWords($WNmap$)) then
  $map \Leftarrow WKmap$
else
  $map \Leftarrow WNmap$
end if
return $map$

function WordNetDisambiguation($np, SemCor$): $WNmap$
$candidateSynsets \Leftarrow$ FindSynsetsInWordNet($np$)
$selectedSynset \Leftarrow MostFrequentNoInstanceSynset(candidateSynsets, SemCor)$
{each mapping contains also the hit count, every count is initialized with 1}
$WNmap \Leftarrow (selectedSynset, 1)$
return $WNmap$
end function

function WikipediaDisambiguation($np$): $WKmap$
$candidatePage \Leftarrow$ SearchWikipediaAPI($np$)
if (IsDisambiguationPage(candidatePage) is False ) then
  $selectedPage \Leftarrow candidatePage$
  $WKmap \Leftarrow (selectedPage, 1)$
end if
return $WKmap$
end function
```

"law school" and "Bakke decision". The first three terms are instances of WordNet concepts. In these cases, their mappings to Wikipedia page titles are selected since there is an article for each term (main subject) and this article is not a disambiguation page. In the case of "law school", there is also an entry in each semantic resource, but the WordNet meaning is selected since this is not an instance term. Finally, besides the fact that "Bakke decision" is an instance term in WordNet, there is no reference to such a term in Wikipedia. For these reasons, this last term remains decontextualized.

5.4 Information Integration

When using data from different sources, the integration of information is imperative in order to avoid duplicates. Beyond detecting lexical duplication when comparing terms, another approach has to be taken in comparing semantically close terms. As well as the meaning of each term found in the last section, we also use the lexical resources WordNet and Wikipedia to infer the level of relatedness among terms.

At this stage, there are three types of terms in the system: simple term, WordNet term, and Wikipedia term. The last two are specializations of the first. WordNet is used to detect if two WordNet terms are synonyms or *duplicates* (i.e. belonging to the same *synset*). Wikipedia is used to determine if two Wikipedia terms (page titles) are duplicates (i.e. related or equivalent by redirections). However, for comparing simple terms (simple lexical terms with no associated semantics) we compute their (lexical) similarity using the Jaro-Winkler string metric (equation 2.3). Algorithm 5.3 presents the whole process of information integration in the Meaning Extraction module.

Each mapping to a lexical entry is represented by a pair $(link, count)$ where *link* is the entry in the lexical resource (the *synset* identifier in WordNet, or otherwise the Wikipedia link) and *count* is the number of occurrences in the text. Count is initialized as 1 but, as later duplicates are detected, it is increased according to the number of duplicate occurrences. The detection of intraclass similarity (in the same lexical resource) is straightforward: between WordNet terms it is only detected if they are in the same *synset* (synonyms); between two Wikipedia terms (page titles) it is detected if they

Algorithm 5.3 Term Integration: Merge three sets of terms - simple, WordNet and Wikipedia terms

```
Input: $WNmap$ : Set of WordNet mappings found$\vee$
  $WKmap$ : Set of WordNet mappings found$\vee$
  $simpleTerms$ : Set of remaining terms
Output: $resultMap$ : union of previous sets without duplicates
  $uniqueWordNetMap \Leftarrow$ GroupDuplicatesInWordNet($WNmap$)
  $uniqueWikipediaMap \Leftarrow$ GroupDuplicatesInWikipedia($WKmap$)
  $uniqueSimpleTerms \Leftarrow$ GroupDuplicates($simpleTerms$)
  $resultMap \Leftarrow \emptyset$
  for all $map_i \in uniqueWordNetMap$ do
    $synsetTerms \Leftarrow$ all terms from synset $map_i$
    for all $map_j \in uniqueWikipediaMap$ do
      $titleTerms \Leftarrow$ redirections to the same page $map_j$
      if (Similar(titleTerms,synsetTerms) is True ) then
        $count_{map_i} \Leftarrow count_{map_i} + count_{map_j}$
        remove $map_j$ from $uniqueWikipediaMap$
      end if
    end for
    for all $term_k \in uniqueSimpleTerms$ do
      if (Similar($\{term_k\}$,synsetTerms) is True) then
        $count_{map_i} \Leftarrow count_{map_i} + count_{term_k}$
        remove $term_k$ from $uniqueSimpleTerms$
      end if
    end for
    insert $map_i$ into $resultMap$
  end for
  {inspect remaining Wikipedia mappings and simple terms}
  for all $map_j \in uniqueWikipediaMap$ do
    $titleTerms \Leftarrow$ redirections to the same page $map_j$
    for all $term_k \in uniqueSimpleTerms$ do
      if (Similar($\{term_k\}$,titleTerms) is True) then
        $count_{map_j} \Leftarrow count_{map_j} + count_{term_k}$
        remove $term_k$ from $uniqueSimpleTerms$
      end if
      insert $map_j$ into $resultMap$
    end for
  end for
  $resultMap \Leftarrow resultMap \bigcup uniqueSimpleTerms$
  return $resultMap$
```

redirect to the same page or at least they are lexically similar above a given threshold. However, for interclass comparisons (from different lexical resources) a specific order is followed: WordNet is first inspected, then Wikipedia, and, finally, terms that are not found in either resource or are ambiguous are just lexically compared.[5.4].

In order to summarize this approach of integrating different terms (simple, Wikipedia and WordNet) KUSCO considers two terms a and b equivalent iff:

- a and b are WordNet terms and belong to the same *synset*;
- a and b are Wikipedia terms which either are the same article or redirect to the same article;
- a and b are simple terms and their string similarity is ≥ 0.95;
- a and b are WordNet and Wikipedia terms respectively (or vice-versa) and A is the set of WordNet terms belonging to the synset of a, and B is the set of Wikipedia redirections to/from the article of b, and $\exists t_a \in A$ and $t_b \in B : sim_{winkler}(t_a, t_b) \geq 0.95$;
- a and b are WordNet and simple terms respectively (or vice-versa) and A is the set of WordNet terms belonging to the synset of a and $\exists t_a \in A : sim_{winkler}(t_a, b) \geq 0.95$
- a and b are Wikipedia and simple terms respectively (or vice-versa) and A is the set of Wikipedia redirections to/from the article of a, and $\exists t_a \in A : sim_{winkler}(t_a, b)$ ≥ 0.95

For our Simmons College example used above, some terms found are shown in table 5.3. Each term was disambiguated using the algorithm 5.2 and later merged following the algorithm 5.3. Besides the fact that disambiguation is done in a simple and computationally efficient way, the system is modular, so it is relatively easy to implement other disambiguation approaches. Some interesting examples of contextualized terms were detected and are exemplified in the same table 5.3. While for some common concepts, their synonyms may have been frequent in the text (e.g. student, scholar), in

[5.4] When a term which does not exist in WordNet or is an instance and its Wikipedia page is a disambiguation page or even does not exist at all.

the case of compound nouns and named entities, finding equivalent terms in the text was more difficult as the represented idea was more specific (e.g. public administration, public office, federal administration, government management).

Table 5.3: Disambiguated terms and some excerpts of their context.

Term	**TF**	**Mapping**	**Context(few examples)**
Public Administration	0.0129	[WK]Public_ Administration	**Public Administration** and Social Service Professions (site F)
delivery	0.0129	[WN]the act of delivering or distributing something (as goods or mail)	Simmons College Food **Delivery** — Simmons College Restaurant... See all 48 Simmons College Pizza **delivery** restaurants ...they get 4 stars for the excellent pasta and the relatively quick **delivery** (site I)
career	0.0079	[WN]the particular occupation for which you are trained	**Career** Education Center **Career** Services Office...**Career** Resource Center (site A) ... professionally skilled and technologically proficient for cutting-edge **careers** (site F)
business	0.0070	[WN]a commercial or industrial enterprise and the people who constitute it	Dean Deborah Merrill-Sands Writes on "Principled Leadership" in Women's **Business** Boston (site A)... Simmons College **Business** Overview (site D)... **Business** with a Legal Studies Minor (Bachelor's) (site E)
spanish	0.0060	[WN]the Romance language spoken in most of Spain and the countries colonized by Spain	**Spanish** Language and Literature (site F)
student	0.0050	[WN]a learner who is enrolled in an educational institution	Dix **Scholars** Information Session (site A)... The undergraduate program is single-sex, with 2,060 **students** (site B)
school	0.0040	[WN]an educational institution	School of Health Sciences...School of Management...School of Social Work (site B)

Continued on next page...

Table 5.3 – Continued

Term	TF	Mapping	Context(few examples)
Medical Transcriptionist	0.0029	[WK]Medical_Transcriptionist	*...certified medical transcriptionist* (CMT) certifications... (site J)
heart	0.0020	[WN]the locus of feelings and intuitions	Located in the **heart** of Boston (site A)... historic New England campus in the **heart** of Boston (site F)
degree program	0.0020	[WN]a course of study leading to an academic degree	Popular Online **Degree Programs** (site E)
Information Systems Schools	0.0020	Not found in WordNet nor in Wikipedia	See more **Information Systems Schools** or Online Information Systems School Degree Programs (site F)

To sum up, we use WordNet and Wikipedia as resources in the extraction process. We also accept terms that do not exist in these lexical resources, but in this case, the terms are not able to be integrated with other related terms. In the next section, term weighting techniques are employed in order to compare and deduce term relevance in a broader set of POIs.

5.5 Term Weighting

At this point, on completion of the previous subtasks for each POI considered as a unique document, KUSCO ranks extracted terms with a TF value in order to select the most relevant terms (only common or proper nouns) that represent a given place. The importance of each term is computed considering all the pages related to a POI. However, as new POIs become enriched, more global relevance computing has to be applied in order to select what is really distinctive in each POI. Diverse measures were studied above in section 2.4.3.3. The most accepted and widely used is TF-IDF frequency (equation 2.11). Another point to make clear is that, in the literature, only an exact match needs to be obtained between terms for them to be considered as equivalent and, as a minimum, the stems of the two terms need to be deduced before comparison to avoid false negatives. In KUSCO, equality is not only considered from a lexical but also from a semantic perspective. For example, the concept "teacher" and "instructor" are considered to be equivalent since they are synonyms considering their most frequent use by WordNet. In the same way, "American Federation of Teachers" and "American Educator" are equivalent according to Wikipedia[5.5].

As we were using POI directories as resources for the Information Retrieval process (chapter 4), we assumed that the information usually found about a POI was already known, such as its address, city, state or province, zipcode, etc. Thus, the terms related to a place's geographical location were not so relevant to its meaning with regard to the place's function: what we name *geographical redundancy*. Further clarification of this subject will be provided.

Another distinctive point that should be established from the traditional TF-IDF, is that usually there were many POIs related to the same company, known as "chain stores". Although they were located in different zones of a given city, they very often offered the same services and product. Table 5.4 presents the 5 most frequent chain stores in the POI database containing POIs from three states in the US: California, Massachusetts, and New York. In these cases, if each POI was considered a different document independently, there would be a lot of documents of the same nature. The terms associated with these documents would have a low IDF as they would become very "common". To balance these results, a "Chain Based TF-IDF" it was computed, that considered not each POI as a document, but each type of POI. Considering POI attributes, two POIs from the same source are considered to be of same type (as part of the same *chain*) if they have exact names and belong to the same categories from their common source taxonomy. In this sense, the influence of geographical terms receives lower relevance as different cities are studied and their respective POIs are stored in the POI database. For example, if a given POI in New York has some associated terms referring to a neighborhood, such as Manhattan, Brooklyn and so on, these terms would have lost importance since other cities, e.g. Boston, were being inspected by searching

[5.5] *American Educator* is a quarterly journal published by the *American Federation of Teachers* focusing on various issues about children and education.

POI Name	Categories	Occurrences
Verizon Wireless	Cellular Providers	557
US Post Office	Post Offices	395
Rite Aid	Greeting Cards Printing;Drug Stores;First Aid;Pharmacies;Medical Supplies & Equipment; Cosmetics Retailers;Photography Labs;Beauty Supplies;Bath & Body Products;Greeting Cards	359
Dunkin Donuts	Bagel & Donut Shops;Driveway & Sidewalk;Grading & Paving;Restaurants;Coffee Houses	343
Starbucks	Cafes,Coffee Houses,Restaurants	202

Table 5.4: The top 5 most popular chain stores in the POI database.

for POIs of the same chain, but in this case with other geographical associated terms, like South Boston, Brookline, etc. Ultimately, it is anticipated that the most relevant concepts with no geographical redundancy will enhance the meaning of the POI in question.

As a result of this system, each POI is represented by the list of its more relevant WordNet, Wikipedia and simple terms, that is by its *Semantic Index*. In the Simmons College example, its semantic index is composed of the following concepts: *Criminal Justice, MBA Schools, Student Loan Center, Post-Master, Postbachelor, Chemistry Program, Political Science Program, Economics Program, Africana Studies Program, Communications Program*, bearing in mind the decreasing TF-IDF order.

5.6 Noise Removal

However, the list obtained at this point carried large quantities of *noise*, which corresponds to any word that does not contribute in any way to the meaning of the place. This includes technical keywords (e.g. http, php), common words in web pages (e.g. internet, contact, email, etc.) as well as geographically related nouns that become redundant when describing the place (e.g. for a POI in Brooklyn Bridge, NY, nouns like ‘New York” or “Brooklyn” are unnecessary). This is what we term its *geographical redundancy*. We apply a filter that gathers a set of fixed common words (a “stopword list") as well as a variable set of "redundant words". The latter set was obtained from an analysis of a large set of texts: we grouped all original texts retrieved, tokenized them to isolate words, applied a stemmer algorithm [Porter, 1997] to deduce the root of each word and defined an IDF (Inverse Document Frequency) value for each stem. We then selected all words relatively commonly occurring in at least 30% or more of our corpus to become also "special stopwords", in the sense that if the stem of some candidate word was present in this last list, it was considered a common word and not eligible to be a descriptive concept. These "special stopwords", in our case, only represented 3% of our stem list of all words processed. This is supported by Zipf's Law [Zipf, 1932], which states that frequency decreases very rapidly with rank.

The identification of all valuable concepts regarding a POI, given a set of web pages, seems to be an achieved goal within KUSCO. However, the emergence of large quantities of redundant, lateral or simply page-format data hinders the determination of accurate semantics. In other words, while recall may be very high, precision is instead very low. We rely on statistical evidence to find very frequent words that bring little information content, together with some heuristics to filter out insignificant words (e.g. a geographical description of the place, such as the name of the city, which becomes redundant).

The *geographical redundancy* of a POI Semantic Index is represented by terms related to countries, cities and other address-related information, which can be found in gazetteers. Most of the time, these gazetteers are available online[5.6] and contain detailed geographical information which can be at the village or street level. Beyond filtering terms related to the POI address out from its Semantic Index, those geographical terms can be also considered as irrelevant to the meaning of a place. Thus, an offline list of gazetteer terms might be built in order to set apart these geographical terms.

Another possible source of noise is lateral information frequently found on directory web sites. In these web sites, only a part of a web page is really related to the POI

[5.6] For example, the Geographic Names Information System maintained by an agency of the US Government [United States Geological Survey, 2007].

under investigation and other components of the web page (e.g. external links, related POIs, directory web site information and advertising) are presented at the same time. Instead of training a wrapper to induce the template used by each directory site, that would demand manual effort in annotating a considerable set of examples, an offline method that automatically detected directory web pages from those retrieved was employed. Thus, if the number of web pages of a given directory web site was above a predefined threshold, these web pages would be the documents used to create the *Web Domain Semantic Index*. Therefore, the expectation was that there would be natural evidence of common properties between web pages for different POIs but within the same website, since it is possible to filter these terms later. To summarise, we aimed to detect the main content among different web pages from the same directory web site using only the content while avoiding those terms that were common in a large number of documents from this collection.

6

Perspectives on Semantic Enrichment of Places

In previous chapters information about Places was retrieved from the Web. This information was obtained from different sources, each one corresponding to a "perspective" of that place. As we will see, perspectives can be correlated and compared to provide different views according to what needs to be known about a place.

6.1 The Open Web Perspective

The *Open Web* perspective consists of crawling the web using a search engine (e.g. Yahoo) given a POI name and address. Location-based Web Search (section 4.2.1) is used to retrieve web pages related to this POI and then these documents serve as input for the Meaning Extraction module of the KUSCO system (chapter 5). The term "open" means that the search is not limited to any particular web domain. Beyond a fixed list of common stopwords, address tokens are filtered out from candidate terms for the Semantic Index. For instance, if a given POI is at this address "2666 Broadway, New York, NY 10025", then terms containing place-related information like "Broadway","New York" or "NY" are filtered out.

The TF-IDF was computed against the entire collection of web pages retrieved for different POIs. As described earlier, POIs of the same type (equal name and belonging to the same categories), are grouped in order to avoid the overweighting of terms from chain stores.

As expected, a lot of directory pages were found using the Open Web perspective. After a first run, a relatively large set of pages from the most recurrent directory web sites were used to build the respective *Web Domain index*. Table 6.1 gives the top 3 most popular directory web sites and the respective percentages of POIs having a web page from each source. Also, for each directory web site, which we call *Web Domain*, the set of the most relevant terms produced by KUSCO given all the web pages of that domain is presented.

Web domain	Representativeness on retrieved Web Pages	Common terms
local.yahoo.com	97%	Stumpfl, Employment, placement, intersection, page
citysearch.com	53%	Neighborhood, Battery, Alphabet, Metro, Twitter
www.manta.com	35%	Manta, input stream, response, Xbox, contact

Table 6.1: The most popular directory sites found from web pages retrieved in Location-based Web search. The percentages refer to the number of POIs which have a page in a given Web Domain.

In table 6.1 the high level of representativeness of the coverage of the Yahoo! Local web domain was because each POI was originally retrieved from this source, with the result that it was very likely that a Web search would find this web page. Each Web domain index contains the top 5 most relevant terms in all web pages in the database for a given domain.

6.2 The About Perspective

Taking advantage of information previously extracted from POI directory sites (see section 4.1.3), a more focused Web search is proposed at the POI official website if it is available. This perspective tries to capture the mission, objectives, services and products offered by POIs and to represent them through the most relevant terms. These days, almost all Web sites have at least one page about the company. Even e-commerce domains contain a small "About Us" link with some basic information about the vendor. Some businesses that cannot feasibly sell their products online and require human interaction to conduct a transaction (like a law firm or company-level software provider) rely heavily on the About section to help build their image as a reputable company.

The title of the *About* page most of time follows one of three options [Potts, 2007]: "About Us", "About [company name]", "About", only differing in the style employed: while the first option is written in the first person, the second option is in the third person, and the last option is often used when there are subsections detailing the information presented: services, products, contacts, etc. However, all these variants generally present an overview of the company on the first page with a fact-based statement summarizing the business, which is sometimes all that a visitor is looking for. The language used is direct, never assuming any prior knowledge from the reader. Every visitor can be considered a complete stranger and has no clue what the company does. This might amount to only a page of content. The content should actively link to the company's primary products and services pages. As key terms and concepts surface, it is important for users to be able to jump to more detailed passages.

In short, good practices in Web design indicate that a good point to locate what the company is, its products and services is on the "About Page". In order to build the *About* perspective, a restricted search for this page (section 4.2.2) was made and the meaning extraction module was applied. As in the case of the Open Web, in the About perspective TF-IDF computation was done against all other web pages previously retrieved (even from the Open Web perspective). This was done because we saw both perspectives as having the same universe of raw material in common: web pages. Thus, even if there were different types of web pages stored (structured, semi-structured or free-text), all were considered to be from the same source: the World Wide Web.

Another characteristic common to the Open Web and About perspectives is that as both were concerned with a given company/POI, its address was removed from the candidate term list when building the final Semantic index. However, in contrast to the previous perspective, no Web Domain index was built since each web page was specific to each official web site. After a considerable number of POIs were enriched with the Web as the source, the final TF-IDF value for this perspective was computed.

6.3 The Wikipedia Perspective

Wikipedia provides us with a massive database of partially structured textual information, currently on over 3 million topics. Plenty of relevant information about places is obtainable, both directly by searching for the actual Wikipedia page of a POI (e.g. Starbucks) and indirectly by finding information related to its category (e.g. Restaurant). The two variations implemented, the Red Wiki (focused on POI categories) and the Yellow Wiki (focused on the POI Wikipage when available) are presented below.

6.3.1 Red Wiki - Low-detail labeling

In the Red Wiki perspective, the Wikipedia page corresponding to the identified set of categories of a POI is processed. As mentioned above, each POI has one or more categories, and except in rare situations, each such category is described in a Wikipedia page. Thus, the *red wiki perspective* of a POI extracts the semantic index from the union of its category Wikipedia articles.

Only the summary section of each category's Wikipedia articles is processed by the Meaning Extraction module of KUSCO. Since this is a Wikipedia perspective, there is no sense in computing the TF-IDF with web pages as the document collection, as the universe for the calculation of term relevance consists solely of Wikipedia articles (associated with either the Yellow or Red Wiki perspectives). Also, particularly in this perspective, no filtering is applied at all, since a POI address is unlikely to appear in a Wikipedia article describing a category. Obviously, the names of those categories used as seed for this enrichment are not intended to form part of the semantic index of a given POI, since they are already known.

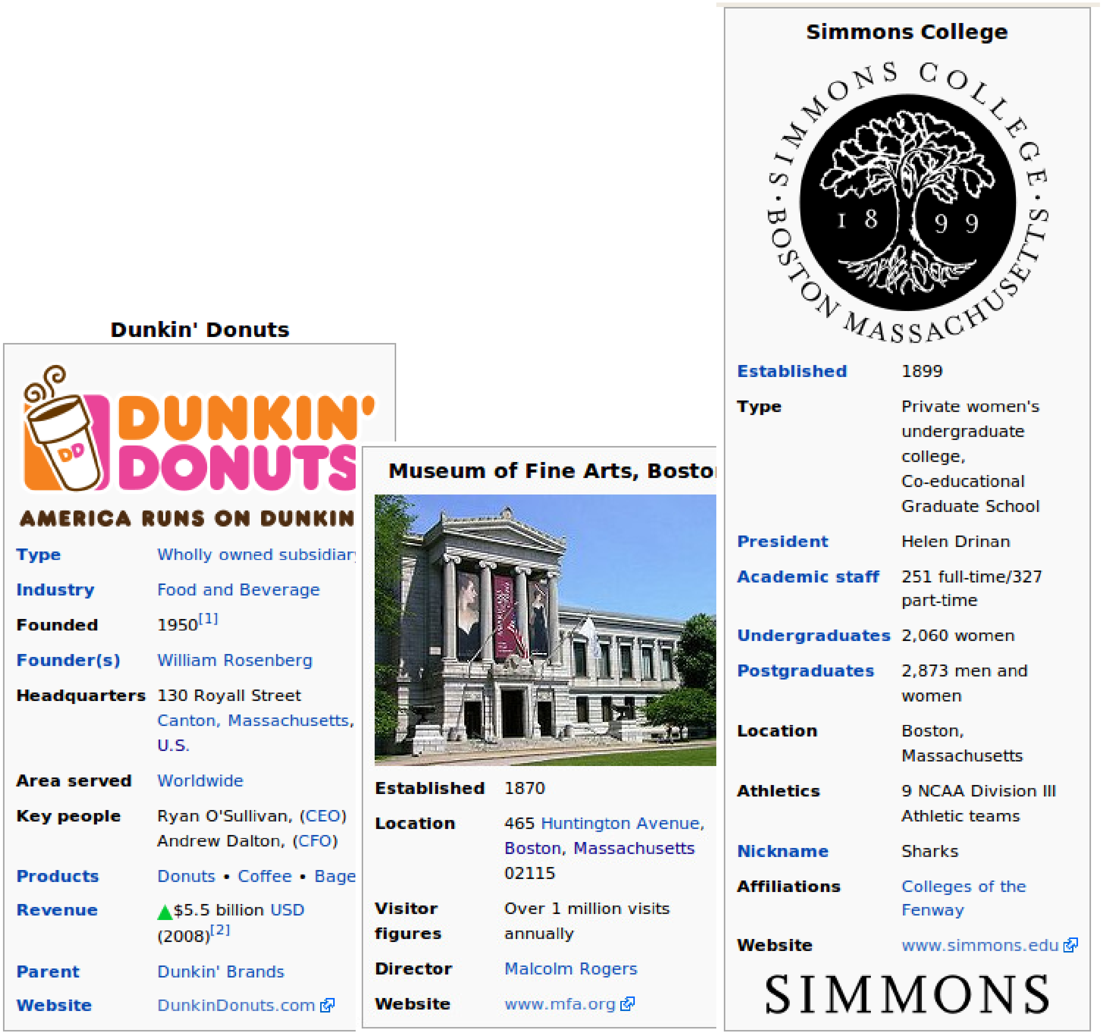
The information provided by Wikipedia about POI categories can be viewed as an extension of WordNet glosses, since the summary of the Wikipedia article sums up the most common sense given to the category name. This information is sometimes very generic considering the POI itself (low specificity). On the other hand, the larger the set of categories to which the POI belongs, the less precise is the final information extracted, since there is no indication from the POI sources which category is more relevant to the POI. In this way, for instance, the categories "Photography Labs" and "Pharmacies" which are associated with the POIs from the *CVS* store chain are considered of equal importance in the POI activity.

6.3.2 Yellow Wiki - Medium-detail labeling

Representing a place by generic terms does not always give a complete vision of it. For comparison between POIs belonging to the same categories but offering concurrent services and/or products, a more specific approach is needed. Accordingly, the Yellow Wiki perspective searches Wikipedia for the specific article about a given POI (section 4.2.3.2). Meaning Extraction is then applied over the article summary and the Semantic Index is built. As info-boxes are a valuable resource used by authors to outline article information, particularly in the case of those articles about companies, artists, geographic entities, etc., some properties about company/POI profiles are judged to be sufficiently discriminating for them to form part of the final set of the most relevant terms. Figure 6.1 presents the most commonly used info-box templates associated with the kind of place studied in this thesis (not political or geographical entities but company/trade businesses or Points of Interest).

From the majority of Wikipedia articles retrieved with info-boxes, some properties were more informative than others. Knowledge of the type of place and the products and services offered is more valuable than knowing who the company's current CEO is. There was no noise in extracting these attribute-value pairs from the Wikipedia database because they were entered manually by authors. In this way, these terms, when they were available, were merged into the final Semantic Index built for a given POI.

In the same way as in the previous perspective, the TF-IDF value was computed over every Wikipedia article retrieved from both perspectives. In addition, in contrast to the



Dunkin' Donuts

Type	Wholly owned subsidiar
Industry	Food and Beverage
Founded	1950[1]
Founder(s)	William Rosenberg
Headquarters	130 Royall Street Canton, Massachusetts, U.S.
Area served	Worldwide
Key people	Ryan O'Sullivan, (CEO) Andrew Dalton, (CFO)
Products	Donuts • Coffee • Bage
Revenue	▲$5.5 billion USD (2008)[2]
Parent	Dunkin' Brands
Website	DunkinDonuts.com

Museum of Fine Arts, Bosto

Established	1870
Location	465 Huntington Avenue, Boston, Massachusetts 02115
Visitor figures	Over 1 million visits annually
Director	Malcolm Rogers
Website	www.mfa.org

Simmons College

Established	1899
Type	Private women's undergraduate college, Co-educational Graduate School
President	Helen Drinan
Academic staff	251 full-time/327 part-time
Undergraduates	2,060 women
Postgraduates	2,873 men and women
Location	Boston, Massachusetts
Athletics	9 NCAA Division III Athletic teams
Nickname	Sharks
Affiliations	Colleges of the Fenway
Website	www.simmons.edu

Figure 6.1: Most common info-box templates associated with POIs.

Red Wiki perspective, the address-related terms were filtered out from the Semantic Index, as were the category terms.

6.4 The Event Perspective

The acquisition of semantics related to events in places works as an extension of the previous section but the focus here is on dynamic online resources that provide information associated with time. The word *dynamic* refers to websites that change content at least on a daily basis. In other words, the main difference lies in the selection of web resources and in the specific attention given to time (having exact information about dates and times is important).

Thus far, the information extracted for places has been relatively static, not changing in time. However, there are places that are not so strongly defined in themselves, but instead are better described by what happens there or, in other words, by the events hosted by a given venue. Examples of these kinds of places are concert halls, arenas, theaters, and playing fields that are used for different sports depending on the time of year. In these cases, a more realistic approach is to collect information not only from POI directories but also from events sources in order to add a different axis in the meaning of a place: time.

The work here presented was developed in collaboration with João Oliveirinha [Oliveirinha, 2010], under the joint supervision of my supervisor, Francisco Pereira, and myself. It is primarily focused on the gathering, extraction and enrichment of information of events in a given city.

Within the Internet, the range of dynamic resources about events is rapidly growing throughout the world, becoming a challenge in itself simply to enumerate the existing variety. Accordingly, clear selection criteria are required: event coverage; geographical coverage; richness of content; availability of historical data and reliability of sources. Event coverage means the ratio of events in the database to those that actually take place in a geographical area. Of course, the ideal value is 1 (every event is reported in the database). Geographical coverage corresponds to the area associated with the database. For experimental purposes, event coverage was favored over geographical coverage, but the selected area needed to have significant event life. Richness of content signifies how detailed the available information is apart from the mandatory data (name, date, time, place) namely the availability of some text description. The availability of historical data is important for practical purposes: storing all event information for research analysis can demand large computational resources and take a long time. The Yahoo! Upcoming [6.1] and Zvents[6.2] web sites achieve a good compromise

[6.1] http://upcoming.yahoo.com/

[6.2] http://www.zvents.com

in those respects, and, although their coverage is not perfect, they are well organized with historical data. For each event, they provide its category (e.g. Music, Social, Commercial, etc.), name, date, time and textual description. Also, both provide an API and are worldwide.

A multi-thread script was implemented to accomplish the extraction and indexing of events and venues in three stages:

- The first stage is where the information from the two sources is retrieved using several parallel threads depending on the source, the method used for retrieving information (API or screen scraping) and the limitations of the service. All the information retrieved is stored in a database in a concurrent way.
- The second stage is responsible for feeding the Meaning Extraction of the KUSCO system with event description, so a list of ranked concepts is retrieved. After this stage is completed, the database is updated with the top N words that best describe the document/event and the corresponding Term Frequency of each concept.
- The last stage is where we compute and update the value of the TF-IDF for all the concepts extracted in the previous stage in the database. This phase is only started after all the events have been retrieved for a given period of time.

http://www.16beavergroup.org/events/archives/000830.php
http://www.giganticartspace.com/
http://upcoming.yahoo.com/event/3909/
http://calendar.artcat.com/event/view/8/3965
http://rhizome.org/discuss/view/12756
http://www.photography-now.com/institutions/I7335022.html

Table 6.2: List of retrieved websites for “Gigantic Artspace”.

To have an idea of the appropriateness of this perspective for those kinds of POIs that are mainly event venues (e.g. exhibition hall, conference center), where their functionality depends heavily on the events taking place within them, we provide an

Concept	Score	WordNet gloss
e-archive	2.559	
gas	1.704	a fluid in the gaseous state having neither independent shape nor volume and being able to expand indefinitely
inquiries	1.033	a search for knowledge
Foreign Legion	0.698	
Bil Bowen	0.591	
Mari Kimura	0.537	
Eric Singer	0.537	
galleryartist	0.512	
Lee Ranaldo	0.442	
pursuit	0.404	a diversion that occupies one's time and thoughts (usually pleasantly)
Joshua Fried	0.394	
Tongue Press	0.349	
Nassau Street	0.349	
Lower Manhattan Cultural Council	0.349	
interstices	0.349	
Franklin Street	0.315	

Table 6.3: Concepts for "Gigantic Artspace"

illustrative example of two different perspectives of the same place. For example, if we consider a POI in New York (-74.0028, 40.7171, "Gigantic Artspace"), which has an official website [6.3], following the Open Web perspective (section 6.1), KUSCO retrieves from Location-based Web search the web sites listed in table 6.2. Then it extracts the concepts and applies filtering. Table 6.3 shows the obtained list. The concept, "gas", is the acronym of "Gigantic ArtSpace", while the several names ("Mari Kimura", "Eric Singe", etc.) correspond to artist names. Open web extraction is thus extremely dependent on external factors (correct ranking of the appropriate pages, quantity of noise in those pages, depth of description of the place) and represents what we call the "hardest scenario" for the extraction of semantics. Here we regard it as a *static* perspective, but

[6.3] http://www.giganticartspace.com/

in reality this is dependent on the current content of the pages (in the example above, some information is actually dynamic, such as the performer names).

In the *Event* perspective, the events are more important that the venue itself. For instance, for the same POI an excerpt of the description text for an event happening there is shown below:

> Event id: 353171
> Tile:"The String Orchestra of Brooklyn: Winter Concert"
> Date: 2007-12-15, 20:00:00
> Description:
> Bach: Erbarme dich, mein Gott from the St. Matthew's Passion Torelli: Christmas Concerto Vaughn Williams: Fantasia on a Theme by Thomas Tallis Mozart: Serenata Notturna, k239 Eli Spindel, conductor Kimberly Sogioka, alto Suggested donation: $10

From this text, the Meaning Extraction module from KUSCO was able directly extract the concepts found in table 6.4 (with their WordNet glosses, when available). Notice that, in this case, the system was not required to perform any web search so this text was the only resource used.

The semantic indexes extracted from the event description using only the Meaning Extraction module at KUSCO as a resource are very poor and provide little more information about an event than the one provided by the raw description. Therefore, after the last stage mentioned above, the system tries to retrieve the article summaries of the pages available in the Wikipedia for each concept in the semantic index. Using this approach, all Wikipedia page summaries related to a single event are gathered into one single file and fed into the Meaning Extraction module again, which results in a new list of concepts ranked by Term Frequency. The last step is to calculate again the TF-IDF for each concept, selecting the best ranked concepts to be used as labels for each event.

After the initial concepts are extracted using KUSCO (either from the Open Web perspective for venues or from the Event perspective for events), they are *enriched* by

Concept	Score	WordNet gloss
Bach	0.637	
dich	0.637	
Eli Spindel	0.637	
Gott	0.637	
Kimberly Sogioka	0.637	
Mozart	0.637	the music of Mozart
Serenata Notturna	0.637	
Thomas Tallis	0.637	
Torelli	0.637	
Vaughn Williams	0.637	
Fantasia	0.546	a musical composition of a free form usually incorporating several familiar themes
alto	0.546	a singer whose voice lies in the alto clef
donation	0.431	a voluntary gift (as of money or service or ideas) made to some worthwhile cause
Theme	0.0.407	a unifying idea that is a recurrent element in literary or artistic work

Table 6.4: Concepts from event id 353171 of Yahoo! Upcoming

searching for the relevant page in Wikipedia and applying again the Meaning Extraction module of KUSCO to the union of all abstracts found using this process. The concepts are then extracted and ranked. Table 6.5 shows the sequence of word lists obtained for the example in table 6.3 ("Gigantic Artspace") before filtering, after filtering, and after Wikipedia enrichment (applied to the filtered list). The system is able to retrieve a number of more distant, yet potentially relevant, associations.

6.5 Similarity between Perspectives

We call perspectives on semantic enrichment of places to the way that distinct sources and modes are used to retrieve information about places. At this point, after all pieces of information processed by KUSCO have been presented, figure 6.2 depicts the KUSCO's conceptual model. The main entities in the system are: POIs; related Websites; categories that POIs belong to; and semantic indexes which are built using some source in

Before filter	**Filter**	**Wikipedia**
gas, e-archive, inquiries, Galleries, Eric Singer, Mari Kimura, Bil Bowen, Foreign Legion, Lee Ranaldo, art, Suggestion Board, Joshua Fried, News Blog, pursuit, performance, community, means, traditions, Tongue Press, interstices, beliefs, soldiers, Museums, Nassau Street, friend, premiere, Lower Manhattan Cultural Council, climates, opera, works, Australia, silence, rhizome, ...	gas, e-archive, inquiries, Foreign Legion, Bil Bowen, Eric Singer, Mari Kimura, performance, gallery artist, pursuit, Lee Ranaldo, Discussion, Joshua Fried, News Blog, Suggestion Board, beliefs, climates, interstices, Lower Manhattan Cultural Council, means, Nassau Street, opera, soldiers, Tongue Press, Franklin Street, rhizome, Appointment, awareness, concepts, critique	gas, inquiry, **Violin**, **audience**, Mari Kimura, **band**, **Music**, performance, **performers**, **matter**, **aim**, **artist**, **Landscape**, **Ohio**, **Drums**, **metal**, **drummer**, **instruments**, **subharmonics**, **Rothstein**, **Kaiser**, **Eric Doyle Mensinger**, **Alice Cooper**

Table 6.5: After filtering and after Wikipedia enrichment

the Web. These indexes are composed mainly of concepts that can be contextualized in WordNet or Wikipedia.

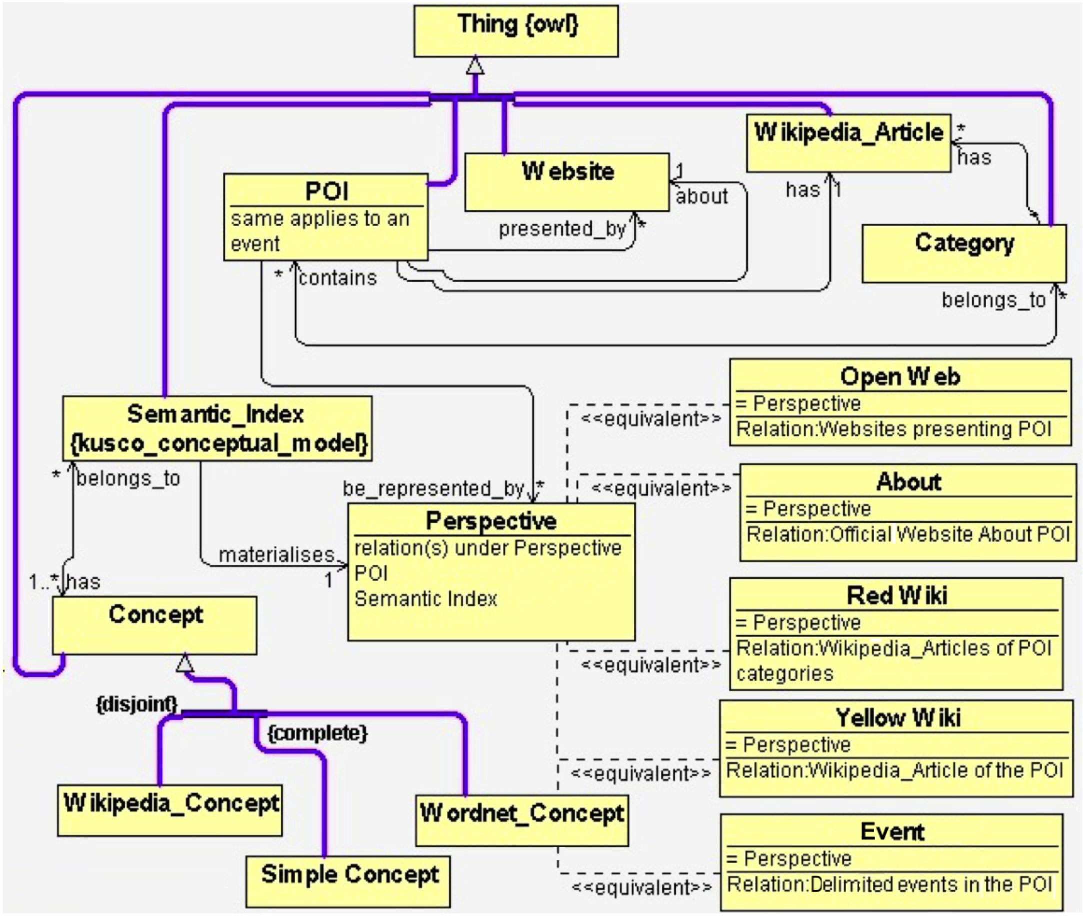


Figure 6.2: Conceptual model of the KUSCO system.

An analysis of semantic similarity between perspectives allows us to verify the recurrence of word patterns from the different sources. The assumption is that for the same POI the words from different perspectives should be related and/or similar. This relatedness is computed by Cosine Similarity (equation 2.4) between indexes from different perspectives for a given POI ranging from 0 (most dissimilar) to 1 (most similar).

Firstly, a sparse matrix of the occurrences of Terms/Indexes is created by weighting each occurrence using a TF-IDF weighting for each term. The contextualization of each term (in WordNet or Wikipedia) is used in order to identify synonyms from different indexes (section 5.3). For example, if the term “nightclub” appears in a given perspective mapped to WordNet as the concept: “a spot that is open late at night and that provides entertainment (such as singers or dancers) as well as dancing and food and drink”, any word representing this concept is considered as a match (e.g. cabaret, night club, club, nightspot). Remaining terms are compared by String Similarity in order to find small variations of names like “market intelligence” and “marketing intelligence”.

Equation 6.1 shows how similarity is computed based on cosine similarity (equation 2.4). This similarity metric is sufficiently generic and can be used to compute the semantic similarity between two semantic indexes in the same or in different perspectives. If X and Y are two semantic indexes, the similarity is computed by:

$$similarity(X,Y) = cos\theta = \frac{\overrightarrow{V}(X) \cdot \overrightarrow{V}(Y)}{|\overrightarrow{V}(X)||\overrightarrow{V}(Y)|} \tag{6.1}$$

where, θ is the angle between the gradients of the semantic index X and Y in the n-dimensional term space. X and Y are very similar as θ approaches 0 ($cos(0) = 1$), and very dissimilar as θ approaches 90 ($cos(90) = 0$). Thus:

- $\overrightarrow{V}(X) \cdot \overrightarrow{V}(Y) = \sum X_i * Y_i$
- $|\overrightarrow{V}(X)| = \sqrt{\sum X_i 2}$
- $|\overrightarrow{V}(Y)| = \sqrt{\sum Y_i 2}$

$\forall i \in 0,..,n-1$, where n is the number of terms in each semantic index. Two terms are equivalent if they are considered to be similar, as explained in the information integration approach to KUSCO’s Meaning Extraction module (section 5.3).

7

Semantically Enriched Places by Lightweight Ontologies

Based on the knowledge that the final result of the process of Semantic Enrichment is a semantic index which contextualizes a POI by a set of concepts, this chapter presents the process of linking these concepts to the Semantic Web [Berners-Lee et al., 2001]. The Semantic Web in our system is represented by the Linked Web Data project [Bizer, 2009], which connects different Ontologies and provides Web APIs in order to consult interlinked data. Thus, the POI can be represented not by a bag of concepts but instead by an interlinked cloud of concepts that enable us to infer more knowledge about a place.

7.1 Semantic Web

Although a decade has passed since its definition by Tim Berners-Lee [Berners-Lee et al., 2001], the *Semantic Web* is a visionary architecture of the Web where all on-line information can be processed by machines. This could only happen if data were structured probably as in Ontologies, by axioms defining entities, their properties and the relations between them.

Since it is an ambitious task to represent all knowledge about the World, even for a restricted subset such as Places, the focus of this research is not to create a complete general Ontology about places from scratch or even Domain Ontologies about different types of places. Since one of the first motivations to build ontologies is for knowledge sharing, here the intention is to reuse structured common-sense knowledge and instantiate this knowledge, mainly concepts, related to Public Places. If we want to use ontology technology for increasing interoperability between multiple representations or increased access to existing data, we need to build ontologies that are linked to existing knowledge organization systems [Hepp, 2008]. In this context, WordNet (presented in section 2.5.1.1) is considered, in this work, as the generic Ontology to represent concepts and the semantic relations between these concepts.

Technically, *Linked Data* refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and it can in turn be linked to and from external data sets [Bizer et al., 2009a]. While the primary units of the hypertext Web are HTML (HyperText Markup Language) documents connected by untyped hyperlinks, Linked Data relies on documents containing data in RDF (Resource Description Framework) format [Klyne and Carroll, 2004].

Another difference is that Linked Data, rather than simply connecting these documents, uses RDF to make typed statements that link arbitrary things in the world. The result, which we will refer to as the Web of Data, may more accurately be described as a web of things in the world, described by data on the Web.

The Open Linked Data project[7.1] includes DBpedia [Bizer et al., 2009b], Yago [Suchanek et al., 2008] and WordNet [Miller et al., 1990], presented in section 2.5.1.3, which cover multi-domain knowledge. Through synthesis, this knowledge is used in our system to contextualize terms in the following way: WordNet, with its clean and carefully manually assembled hierarchy of thousands of concepts, serves as a background taxonomy for common terms (or common nouns); Yago, with its more than 1.7 million of precisely specified Wikipedia entities, is used to disambiguate specific terms (or named entities); and finally with DBpedia, we cover all those other articles in Wikipedia referring not only to leaf articles (entities) but more generic articles that

[7.1] Available at http://linkeddata.org/

in our case we use to infer related concepts about a category name.

Two other resources interlinked in the Open Linked Data project are also presented in a new perspective: NAICS and SUMO. The first is the acronym for the **N**orth **A**merican **I**ndustry **C**lassification **S**ystem and was introduced in section 4.1.2. The second resource, an acronym for **S**uggested **U**pper **M**erged **O**ntology, is a generic ontology where most abstract classes in the world are classified. Our approach uses the connection between NAICS and SUMO to find the right abstract class to classify a given POI according to its main activity (e.g. a Restaurant, a Museum, or a School). This chosen class will facilitate the future instantiation of this generic model by incorporating new concepts related to this new entity (the POI properties and attributes).

The Suggested Upper Merged Ontology (SUMO) is an upper level ontology developed by IEEE's Standard Upper Ontology (SUO) Working Group as an open source initiative with the purpose of creating a public standard, accessible through the Web. The idea is to use SUMO as a source for general-purpose definitions and as the foundation for the construction of middle-level and domain-specific ontologies. SUMO will promote data interoperability, information retrieval, automated inference, and natural language processing [Breitman et al., 2007].

SUMO was developed by the Teknowledge Corporation using input received from the Standard Upper Ontology (SUO) Working Group. At that time, SUMO was the first attempt to synthesize content from several available formal ontologies, including John Sowa's upper level ontology [Sowa, 2000]. Some of the general topics covered in SUMO include [Niles and Pease, 2001]: structural concepts such as instance and subclass; general types of objects and processes; abstractions including set theory, attributes, and relations; numbers and measures; temporal concepts, such as duration; parts and wholes; basic semiotic relations; agency; and intentionality. Figure 7.1 illustrates SUMO's top classes.

The ontology is being progressively created through the integration of public content. For instance, the YAGO-SUMO integration incorporates millions of entities from

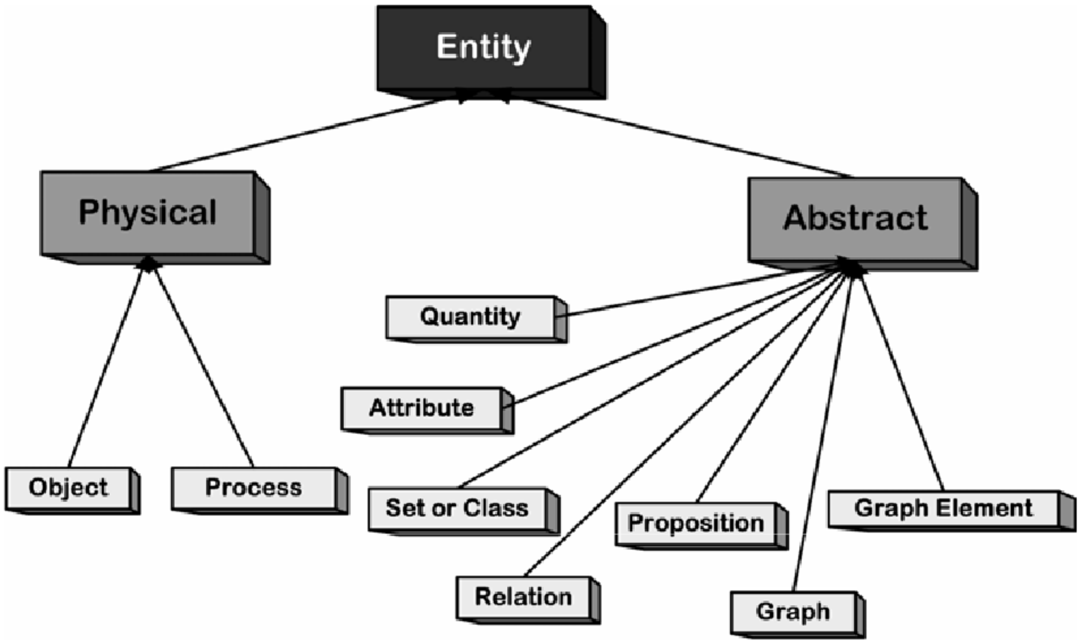
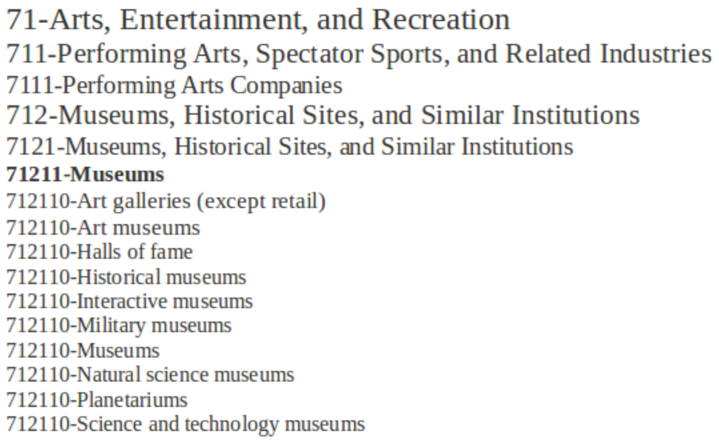


Figure 7.1: SUMO's top classes [Breitman et al., 2007].

YAGO into SUMO. With the combined force of the two ontologies, an enormous, unprecedented corpus of formalized world knowledge is available for automated processing and reasoning, providing information about millions of entities such as people, cities, organizations, and companies [de Melo et al., 2008]. Compared to the original YAGO, more advanced reasoning is possible due to the axiomatic knowledge delivered by SUMO. For example, a reasoner can conclude that a child of a human must also be a human and cannot be born before its parents, or that two people sharing the same parents must be siblings. An example specifically related to place properties could be a reasoner inferring that a service elevator is a transportation device consisting of a car that moves up and down in a vertical shaft for carrying freight so that objects can move from one floor to another in a building. Another instance of such interlinked data is the integration of NAICS-SUMO achieved by SUMO's authors [7.2]. Despite the fact that the NAICS taxonomy is barely mapped to SUMO, the main economy sectors are uniquely indicated. This is illustrated by comparing the excerpt from the NAICS

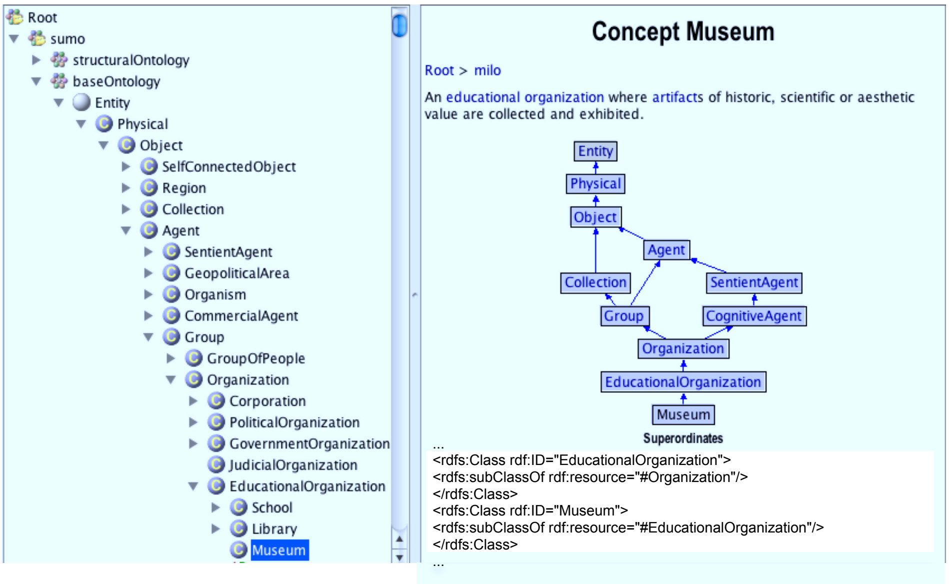
[7.2] Available through the SUMO ontology portal: http://sigmakee.cvs.sourceforge.net/viewvc/ sigmakee/KBs/naics.kif



71-Arts, Entertainment, and Recreation
711-Performing Arts, Spectator Sports, and Related Industries
7111-Performing Arts Companies
712-Museums, Historical Sites, and Similar Institutions
7121-Museums, Historical Sites, and Similar Institutions
71211-Museums
712110-Art galleries (except retail)
712110-Art museums
712110-Halls of fame
712110-Historical museums
712110-Interactive museums
712110-Military museums
712110-Museums
712110-Natural science museums
712110-Planetariums
712110-Science and technology museums

Figure 7.2: Example of the NAICS hierarchy

hierarchy in figure 7.2 with the same information inferred as RDF triples in SUMO, shown in figure 7.3.



Concept Museum

Root > milo

An educational organization where artifacts of historic, scientific or aesthetic value are collected and exhibited.

```
...
<rdfs:Class rdf:ID="EducationalOrganization">
<rdfs:subClassOf rdf:resource="#Organization"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Museum">
<rdfs:subClassOf rdf:resource="#EducationalOrganization"/>
</rdfs:Class>
...
```

Figure 7.3: Excerpt of the NAICS hierarchy mapped to SUMO.

7.2 Linking Semantically Enriched Places to Ontologies

Two clearly distinct approaches were followed in representing a semantic index about a place as a lightweight ontology with concepts and relations about a given POI. The first approach was based on the hypothesis that third-party ontologies representing the most popular place categories were available and validated so they could be instantiated by our system. The second and more realistic approach assumes that these ontologies do not exist (or at least are not complete) and we started by finding an upper level ontology with generic classes about the most prominent category of a POI and after that instantiate this generic model using the concepts and attributes found about this new instance. Both approaches are described in greater detail in the following subsections.

7.2.1 First Approach: Populating a Generic Place Ontology

We define Generic Place Ontologies as a set of knowledge structures representing a collection of common sense and generic information about well-known place categories, like restaurants, cinemas, museums, hotels, hospitals, etc. As a first stage, this information was manually collected from well-known and shared third-party Ontologies but as the system was used, it would have been dynamically fed by new examples, and thus instantiated and populated by specific facts about these instances that represent real-world places. In order to infer a place's meaning, ontologies were contextualized in WordNet. For each term in an ontology, a definition was looked for in WordNet. At first, only concepts describing places were extracted from the Web. After this, relations between those concepts were instantiated by harvesting a huge Web database of common-sense facts, KnowItAll [Etzioni et al., 2008]. The most common relation between two concepts was chosen for each pair, but in the near future, we aim to instantiate relations between these concepts using the original context where they appear (web pages related to a given place) instead. For instance, given the concept "collection of artifacts", from the original source (web pages) we know that "a museum houses a *collection of artifacts*" while from KnowItAll "a museum is home to a *collection of artifacts*".

Common Sense Ontologies are formally collections of simple and semantic knowledge that allow the extension of the computational reasoning process. We were able to focus on generic concepts and relationships about a known category of places (restaurant, museum, hospital, cinema, pharmacy, etc.) in order to build a *Common Sense Place Ontology* comprising not only semantically related concepts for a given category but all concepts referred to by descriptive definitions (or glossaries).

7.2.1.1 Place Classification

In order to evaluate the system's capacity to categorize POIs (i.e. to identify if they represent restaurants, museums, bars, etc.), we selected a set of ontologies using popularity-based criteria. The result of this ontology selection process was a set of four ontologies about different domains: restaurants[7.3], museums[7.4], travel[7.5] and shows[7.6].

To start, as described above, POIs were associated with a set of WordNet [Miller et al., 1990] concepts. To facilitate the categorization of the POIs against this set of ontologies, we also mapped the concepts of the selected ontologies in WordNet. The mapping comprised three phases:

1. Term Extraction. The terms were extracted from the names of the concepts contained in the ontology. Because these names usually comprise one or more terms, they were divided up according to use of upper case letters and special characters such as '-' and '_'.

2. Term Composition. In a preliminary analysis of the results obtained in the previous phase, we found that some of the terms represented compound entities such as "fast food" or "self-service". To avoid losing these compound entities, the different terms extracted from each concept were combined and the resulting combinations were included as terms associated with the concept.

3. Concept Identification. The terms and combinations of terms, extracted in the previous phases, were then searched for in WordNet. When more than one sense was found for each term, these were disambiguated by selecting the sense with the greatest tag count value.

[7.3] http://gaia.fdi.ucm.es/ontologies/restaurant.owl

[7.4] http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs

[7.5] http://protege.cim3.net/file/pub/ontologies/travel/travel.owl

[7.6] http://www-agentcities.doc.ic.ac.uk/ontology/shows.daml

The result of the mapping process was that all the concepts of each ontology became associated with one or more WordNet concepts. With the ontologies already mapped in WordNet, the categorization process proceeded. The categorization was conducted with three different alternative mappings, which we called *simple mapping*, *weighted mapping* and *expanded mapping*, as defined below:

- *Simple mapping*, as its name suggests, is the simplest approach and represents the direct mapping between the concepts associated with POIs and the concepts associated with the ontologies. The mappings between concepts of the two structures are counted and the POI is categorized in the ontology with the greatest number of mappings.
- *Weighted mapping* takes advantage of the TF-IDF [Salton et al., 1975] value of each one of the concepts that are associated with a POI. The TF-IDF value represents the weight of the concept in relation to the POI it is associated with. In this way, each mapping has a weight equal to the weight of the concept that originated the mapping. The POI is then categorized in the ontology with the greatest sum of mapping weights.
- *Expanded mapping* is based on the idea that the expansion of the concepts to their hyponyms makes the mapping more tolerant and extensive. One may argue that when searching for *mammals*, we are implicitly searching for all kinds of mammals, such as a *dog* or a *cat*. Following this idea, the concepts associated are expanded to their hyponyms and the concepts that result from this expansion are associated with a POI. Then, the mapping between the POI and ontologies is performed as in the simple approach.

An experiment [Antunes et al., 2008] done using this approach is presented in appendix B.

7.2.2 Second Approach: Instantiating an Upper Level Ontology

In terms of concept instances, a Semantic Index of a place in a given perspective is generally made up of both types of elements: generic concepts and entities. The first group is often contextualized in WordNet, while the second group representing instances of generic concepts is quite often contextualized in Wikipedia when there is unambiguously a related article about each instance. However, as we are interested in the *meaning of a place*, this knowledge sometimes has to be generalized in order to infer which generic concepts are most relevant. If we consider the example of the POI “White House” located in Washington, 1600 Pennsylvania Avenue NW, the *Yellow Wiki* perspective retrieves the Wikipedia article about the White House[7.7]. The appearance of extracted entities referring to former presidents (e.g “John Adams”, “Thomas Jefferson”, “James Monroe”, “William Howard Taft”, “Harry S. Truman”, “Theodore Rosevelt”) or even the present incumbent are not so important, because they are all instances of the concept “President of the United States”.

The generalization of instances consists basically of exploring the *Hypernym* semantic relation which allows us to infer more abstract knowledge about instances. In WordNet, this is relatively direct because it has had a built-in relationship since its first release, but for instances in Wikipedia this property is not self-contained in the encyclopedic database. In this case, we can take advantage of the DBpedia project [Auer et al., 2007] and its continuously updated ontology of Wikipedia pages organized into classes whose instances are the articles themselves. These classes are sometimes directly connected to WordNet synsets using Yago [Suchanek et al., 2008] mapping. As an example, the Wikipedia article entitled “Theodore Rosevelt” corresponds to the node in the DBpedia ontology http://dbpedia.org/page/Theodore_Roosevelt and “John Adams” to http://dbpedia.org/page/John_Adams; both of these have the property “rdf:type” corresponding to the value “yago:PresidentsOfTheUnitedStates”, which in turn is mapped to the “president#n#3”[7.8] synset in WordNet. The Semantic Index that undergoes this generalization process is then called a *Generalized Semantic Index*.

As each instance is generalized to the most specific concept which subsumes it, the relevance of each concept in the Generalized Semantic Index needs to be recomputed. Thus, a generic concept which is directly expressed in the source document(s) or its instances are considered all together just as a single reference to the concept and the sum of their TF-IDF values produces the new weight of the resulting unified concept. This recalculation is done for each perspective and consequently the semantic indexes

[7.7] http://en.wikipedia.org/wiki/White_House

[7.8] This sense of the word “president” means “the chief executive of a republic”.

can become shorter. This occurs mainly when sibling instances or instances of concepts are present in the same semantic index.

7.2.2.1 Place Classification

Different approaches were implemented and tested to classify POIs to a common taxonomy, which would allow, for instance, a place to be directly connected to the SUMO Upper Level Ontology. Since a given source of POIs generally provides a proper taxonomy of POI categories, classifying them by related terms (e.g. Entertainment Venues or Live Theaters), it was necessary to select a common classification to use. The aim was to be able to classify POIs from different sources to a common and more widespread taxonomy like NAICS. The proposed approaches for automatically classifying POIs in a given taxonomy were introduced in section 4.1.5 and they rely mainly on category information from POI sources which are also organized in taxonomies. Among the different approaches studied, from Ontology Matching to Lexical/Semantic Similarity, the most profitable was the application of Machine-Learning (ML) algorithms to automatically classify the NAICS code of a given POI [Rodrigues et al., 2012]. More precisely, the most successful algorithm was a K-nearest neighbor classifier, named IBk Aha et al. [1991] (with $k = 1$), which essentially finds the similar test case and assigns the same NAICS code. This algorithm obtained approximately an accuracy of 85%, 76.8% and 73.1% for the automatic classification of POIs in the two, four and six-digit NAICS code respectively (explained in more detail in appendix C).

7.2.2.2 Lightweight Ontology Instantiation

In the Meaning Extraction Module, only concepts are extracted from web pages describing places. However, here, once the right NAICS category is found for a given POI, those concepts would be used to instantiate its corresponding mapping in the SUMO ontology. In our system, we aim to derive informal descriptive lightweight ontologies, since they are made up of atomic labels and *is-a* edges. As the Ontology is composed not only of concepts and *is-a* relations but also of other non-taxonomic relations and attributes, we also propose using the original context where concepts appeared inside the web page to be used to instantiate relations between them. For instance, considering a POI semantic index which contained concepts like: “facility”, “vegetarian”, “Portuguese”, “terrace”, and “bar”, this POI was correctly categorized as a Restaurant by matching the concepts against the corresponding classes in the SUMO ontology: Facility, VegetarianCuisine, PortugueseCuisine, OutdoorSeatingOnTerrace and WineBarCuisine, respectively.

An ontology typically does not contain all the possible instantiations of generic concepts in a given domain. This instantiation process is made possible by WordNet semantic relations, which are implicit through contextualizing concepts from web pages, and non-taxonomic relations, which connect semantically related concepts. For example, therefore, although the concept "Dim Sum" from a POI semantic index does not appear explicitly in the Ontology, the WordNet semantic relation in "Dim Sum" *is a type of* "Cuisine" is useful for capturing which classes and their types occur in a given POI. Another example comes from the concept "private dining room facility": although it does not appear in WordNet, we are able to identify its connection with the "Facility" class by lexico-syntactic patterns [Schutz and Buitelaar, 2005].

8

Experiments and Evaluation

In the development process, we placed great emphasis on ensuring the short-term applicability of this project. It is an online resource for extracting semantic information about places, intended for use by other projects and applications. The choices made regarding both POIs, events and dataset samples were motivated by our desire to make the system relevant to a wide user community, and to ensure it reflects the unstructured nature of the internet. No priority or special emphasis was given to POIs that provide more information than the average. Accordingly, in the first section of this chapter, we present different scenarios, each one corresponding to an experiment that demonstrates the behavior of the system in an uncontrolled environment. In the second section, a more methodological evaluation of the proposed system is conducted.

8.1 Experiments

Alongside the development of the KUSCO system and the writing of related papers, several experiments were done with different samples of POI and Event data sets. The geographical region studied was mainly metropolitan areas in the USA, apart from initial experiments which considered also POIs from UK and Australia. Therefore, the data produced by KUSCO could be used by other projects which focus on these areas. In the following subsections, each experiment performed alongside the development of this thesis is described, together with the results obtained. The experiments are presented in chronological order so the reader will also perceive the evolution of the system in terms of the modules used. We will not hide the difficulties in analyzing

KUSCO, namely in with regard to the value (or quality or usefulness) of the results. Nevertheless, we hope to provide a set of objective conclusions and benchmarks that may be useful for future comparisons.

8.1.1 KUSCO and the Open Web perspective

In the first experiment [Alves et al., 2009], the POI categories we chose to analyze were restaurants and hotels. These are described mostly in dedicated listing website pages such as tripadvisor.com, hotels.com, lastminute.com, and others. While these websites can provide rich content for each POI, in the majority of cases they provide only limited detail as well as plenty of noise. There is a very large number of hotels and restaurants described online and these categories do not represent a set of hand-picked points. These conditions made hotels and restaurants a good basis for our analysis. Also, the Internet is widely used by the public to explore these POIs, which increases the relevance of the metadata that our system creates.

A set of experimental results was obtained for over 215 POIs which were randomly selected from 4,989 POIs of hotels and restaurants in the U.K., Australia and New York city. They were collected from different POI sharing websites [8.1] and also from the Yahoo! Local search directory. We also addressed a few questions about the effectiveness of the KUSCO System. Here, we focus our experimental evaluation of two distinct modules of the system: Location-based Web Search and Meaning Extraction.

For the 215 POIs, 1,091 web pages were processed by KUSCO. With a great diversity of web page sources, 477 different domains were retrieved, most of which were directory Web sites. Following our initial queries with the Yahoo! Search engine, we repeated an identical process using Google search. For our POI set, we automatically selected 864 pages from the total retrieved, using the same heuristic described above. Table 8.1 presents statistics from different sets of pages, one group retrieved via Yahoo and the other retrieved via Google. The Yahoo results exhibited a greater diversity than Google, which could be explained by the fact that more than 50% of web pages retrieved by Yahoo were not considered relevant by Google.

[8.1] Such as POIfriend.com, Pocket GPS World, GPS Data Team, POI Download UK.

	Yahoo	Google
web pages per POI (with a 'top N' threshold of 10)	5.07	4.02
Total of distinct domains	477	300
Common web pages (from the total retrieved by each search engine)	49	73
POIs with common web pages retrieved by both search engines	215 (All)	

Table 8.1: Location-based web search results for 215 POIs from two Search APIs: Yahoo and Google.

Related to the Meaning Extraction module (ME), we chose to qualitatively analyze the results that we obtained. We needed to understand *the contribution of the Meaning Extraction module*. For each text, the ME in KUSCO extracts the most relevant terms, which are then contextualized in WordNet. Semantic Indexes produced by KUSCO have an average size of 35 terms (both concepts and named entities). We applied the Information Content (IC) measure from WordNet concepts (a combination of specificity and term frequency of each term in relation to large corpora [Resnik, 1995]) to the KUSCO semantic indexes. In this respect, we obtained 71 ± 6.0% of IC. Looking for Named Entities, we verified that the average TF measure for these concepts was 59%. These measures, however, mean that KUSCO demonstrated a considerable level of efficiency in retrieving valid concepts with regard to their relevance to the place. For instance, concepts like "New York" or "address", present in the Web Pages processed, received a high IC value or high frequency, but they do not add novel information about the place, so they were not selected for a semantic index.

8.1.2 Comparison of Perspectives

In the second experiment [Alves et al., 2010], we collected a large set of POIs from Boston, New York and San Francisco. The selection of these places was related to the interaction with other projects that we were running in each of these cities.

Applying the Event perspective, we explored the Boston Globe Calendar[8.2] to retrieve information from events and the venues (POIs) hosting them. The Semantic Enrichment for those POIs and events needed some processing time. The average for

[8.2] http://calendar.boston.com/

	New York	Boston	San Francisco	Overall
Yahoo	183144	64133	94466	341743
YellowPages	7694	12878	-	20572
Boston Calendar	13999	2867	9497	26364
OpenWeb	757	2020	-	2777
Red Wiki	69011	20309	-	89320
Yellow Wiki	4400	1928	-	6328
Events	-	7591	-	-
Enriched Events	-	3827	-	-

Table 8.2: Above: number of POIs per perspective/city; Below: number of enriched POIs per perspective/city.

a POI analysis from the Open Web perspective was approximately 108 seconds, while for the Red and Yellow Wiki perspectives it was 57 and 31 seconds, respectively. The events analysis took 30 seconds on average. The Open Web perspective is naturally more time-consuming since it searches the entire web (using the Yahoo! Search engine), while any of the other perspectives uses a more bounded search. In Table 8.2, we present the overall statistics.

Regarding the words obtained, we had a total of 77,558 different words, of which 9,746 (12.6%) were also identified in WordNet. These concepts were analyzed with regard to the average information content (IC) obtained. The IC [Resnik, 1995] reflects the balance of specificity of the concept on a scale of 0 to 17. This average was 16.31 ± 1.73, meaning that the concepts were in general very specific, thus carrying a rich content for the definition of POIs. However, this a risky game: if concepts are generic, the probability of being correct with respect to the place is much higher than when they are very specific. Since these words came from the actual text, in general they should be correct.

We show in Table 8.3 an excerpt of the "good" and "bad" examples found. This choice intends to reveal the benefits and drawbacks of the approach. A less subjective perspective of the results is presented in the next section. Except for the Red Wiki, we only put one category for each POI (many of which have more than one) to make the table more legible and enable the reader to understand the type of place.

Name	**Categories**	**Terms**
Open Web		
Envirotech Incorporated	Waste and Environmental Consulting	Industrial Services, Asbestos Management, Mildew Removal, Asbestos Removal, Residential Services
Grasshopper	Telecommunications	Boston Telecommunications, Gary, Communication Services, Boston Business Directory, Telephone Communications
Cambridge Savings Bank	Banks	Houston, Reading Room, Allston, Senior Commercial Loan and Business Development, Federally Chartered
I Have A Dream Foundation	Educational Consulting	Boulder County, Dany Garcia, Arne Duncan, Jeffrey Gural, National Partners
Universal Gear Incorporated	Clothing Retailers	Banana, Lafayette St, Yahoo Services, Terms of Service, Category Sponsors
Monroe Paint Distributors Incorporated	B2B Paint & Wall Coverings	movie theater, latitude, beauty salon, Delicious, Construction
Red Wiki		
Bowdoin Square Exxon	Gas Stations	pumps, gasoline, fuel dispenser, filling station, gasoline stand
Harvard Market	Grocery Stores	groceries, retailing, food, vegetables, products
Kim Depole Design Incorporated	Interior Design	office space, architects, private residence, code, decoration
Little Basil	Restaurants, Thai Restaurants	restaurant, restaurateur, cuisines, delivery service, meals
Cambridge Library	Libraries	collection, library, information needs, public body, access points
Harvard Magazine	Marketing Agencies, News Services	pool, product, industry trade group, farmers, consumers
Yellow Wiki		
Victoria's Secret	Clothing Women's, Clothing	Victoria, wear, Limited Brands, top Model, fashion models
Boston Police Department	Law Enforcement	Massachusetts, law enforcement agency, correction, investigation, responsibility
TD Garden	Entertainment Venues	Boston Celtics, arena, Boston Blazers, naming rights, National Lacrosse League,

Continued on next page

Table 8.3 – continued from previous page

Name	Categories	Terms
Starbucks Coffee	Coffee Houses	stores, Seattle, Washington, drip, Israeli
Babbo	Restaurants	Marchen Awakens Romance, Ginta Toramizu, team, Nobuyuki Anzai, manga series
Blue Smoke	Steak Houses	Nora Roberts, Blue Smoke, Television film, novel
Boston Calendar Events		
Nature Trail and Cranberry Bog	Nature	Pond, Falls, streams, currents, winter
The Haunted House	Theatre	Orpheum, journey, spirit, tale, surprises
Salem Farmers' Market	Farmers' Markets	cultures, consumption, carbohydrate, Food safety, gastronomy
8th Village Cadillac Day	other	Cadillac, Michigan, Automobile, General Motors Company, vehicles
Fall Forest Festival	Community	Trees, Collins, plants, Macmillan, Sequoia sempervirens
5K Road Race Run	Running	Cancer, cells, abnormalities, neoplasm, treatment
Lexington Farmers' Market	Farmers markets	tent, camping, shelter, rope, poles

Table 8.3: Some examples from experiments (in each perspective, first block is "good", second block is "bad").

The results from the Open Web perspective were extremely dependent on the precision of the initial search. In other words, if the correct webpage about the POI was found, then generally the results were acceptable, but this was not always easy to guarantee, depending on the nature of the POI. For example, if its name was a common noun (e.g. "Gap"), there were too many unrelated pages. In contrast, if it did not have a webpage (e.g. it only existed in directory listings), there was no related page. The Red Wiki perspective easily obtained meaningful words, although these were hardly specific to the POI, which was to be expected since it works within its own category. It is therefore a very "safe" perspective in terms of guaranteeing the correctness of the obtained semantics. The Yellow Wiki was able to obtain much more refined results (e.g. the TD Garden is in fact the arena where Boston Celtics play NBA games) but it was more fragile when the wrong Wikipedia page was found (e.g. the Blue Smoke Steak House was taken as a film with the same name) and is easily fooled by lateral information (e.g. the fact that Starbucks originated in Seattle should be less relevant than, for example, the fact that it serves coffee or cappuccino).

Finally, the events database brought a different kind of results. We were no longer defining the place as it was interpreted by others; instead we were defining events (which in turn provided a different perspective on place) as they were described in the database and enriching that analysis with Wikipedia. This again allowed for elaborate results (e.g. the description of the environment of a nature trail) but it could focus exaggeratedly on secondary concepts (e.g. describing the concept of tent in the Lexington Farmers' Market).

8.1.2.1 Focusing on the Event Perspective

Apart from the comparison of perspectives seen above, a third experiment [Pereira et al., 2009] was related to a specific project in which a correlation was sought between semantics of place and mobility data. Therefore, the study area and time window were constrained by the available data for the project. More specifically, we worked with the Lower Manhattan area around the New York City Waterfalls exhibit by the artist Olafur Eliasson. This corresponded to a polygon that fell within the bounding box from 40.698191", -73.991739" to 40.715693", -74.021560" (approximately 3 × 2 km^2). We called this the "waterfalls area". For some experiments, we also considered an "extended area" (which covered a larger portion of Manhattan, up to 7th street, 40.68366", -73.960933" to 40.727915", -74.0401914", approx. 7 × 5 km^2) to allow us to obtain a larger number of points. The time window chosen went from August 2007 to August 2008, covering most of the Waterfalls exhibition period. As with the area, this choice was made in synchrony with the mobility analysis project. In both cases, we believe the choices to have been valid. For this experiment, the source of events selected was Yahoo! Upcoming [8.3].

In order to allow the comparison of results from different *perspectives*, we used the same set of POIs, essentially corresponding to 107 venues in the "waterfalls area" and 716 in the "extended area". For the Open Web perspective on the initial set of POIs, we took a sample of 292 different venues that included all of those from the "waterfalls area" and some from the "extended area". 2,150 pages were retrieved (an average of 7.36 ± 3.07 pages per POI). Such a variety of information gave rise to a Semantic Index

[8.3] http://upcoming.yahoo.com/

Percentile	average size	std. deviation
5%	56	29.49
10%	43	22.22
20%	33	17.73
50%	25	14.30
90%	23	13.18

Table 8.4: Filter progress with different configurations

of 56 ± 28.49 concepts on average.

The filter of stopwords proposed in section 5.6 was not just a fixed set of common words but also included a dynamic set of special stopwords (which were very common in many of the documents). From all the semantic indexes, the top 5 most popular terms retrieved were "Terms of Service" (131 times), "Neighborhood" (123), "Zip" (120), "Suggestion Board" (79) and "Search Local" (70). These had little or no valid semantic information and indicated that the filter needed to be improved. We observed that the definition of *common* needed to be fine-tuned. In this regard, we ran the algorithm with several percentiles of values where a word would be considered "common" (occurring in 5%, 10%, 20%, 50% or 90% of the documents). From a subjective analysis of the end results (see table 8.4), we can see that the filter was effective in removing noise, although it still failed with those concepts, meaning that work had to be done to improve it.

With regard to the source Yahoo! Upcoming, for each event in the limited area and during the study period, we extracted the top 5 words from the Event perspective. The total number of words obtained was 724, of which 306 were duplicates. The word "music" is the one that appeared most times (23), followed by the words "internet" and "artists". Overall, the full word spectrum had the following pattern: 41.3% of the words showed up only once, 12.9% twice, 14.5% 3 times, 6% 4 times, which was to be expected given the small number of events covered. Regarding the TF-IDF statistics obtained, the average value was 0.0357 ± 0.0675. From the empirical observation, the words obtained were in general relevant to the topic, which means that both the filtering and the enrichment phase worked well.

Table 8.5 shows the top 5 words for 25 events from Yahoo! Upcoming. For each of these events which were available on the Upcoming website, with their corresponding event id, we picked the top 5 words from the description and then performed the aggregation of Wikipedia abstracts for those words. Then, we ranked them again, thereby obtaining the top 5 words listed (in order of relevance).

Id	**Category**	**Name**	**Top 5 words**
353171	Music	The String Orchestra of Brooklyn: Winter Concert	music, suites, johann sebastian bach, style, forms
462350	Media	Aleksey Budovskiy: Russian cartoons recent and classic	country, animators, soviet, language, arms
449040	Family	Easter is 'Egg'cellent in Lower Manhattan	families, children, symbol, nest, units
250921	Education	The Apartment (1960): Movie Nights on the Elevated Acre	star, comedy, part, core, title
396331	Performing/ Visual Arts	Renaissance: International New Media Exhibition	premiere, artists, performances, exhibition, term
447037	Other	NY Giants' Justin Tuck Autograph Signing at J&R Music World	team, bowl, world, game, eastern
282198	Festivals	Stone Street Oysterfest	oysters, pub, term, shell, fogelson
350299	Other	Trinity Church Choir Live at J&R on 12/6	spirit, performance, people, performers, example
692856	Other	Regina Belle Performance & Autograph Signing	cd, autographs, disc, media, minutes
323193	Commercial	IBM and ACORD eForms+ Development Tour: Extend electronic forms capabilities	forms, industry, acord, workshop, area
386182	Other	New York Social Network Group Dinner at the Seaport	food, cultures, ships, carmine, methods
765065	Other	Thought @ Rebar	dance, night, physics, body, form
495775	Other	JD Allen Free Live Performance	detroit, population, saxophone, census, michigan
341172	Music	From the Ocean to the Gulf	musicians, students, music, tarab, repertoire
221078	Music	Lullatone USA tour	myspace, lullatone, company, hills, interactive
319392	Sports	Friday Night Fights	school, home, tradition, club, example

Continued on next page

Table 8.5 – continued from previous page

Id	Category	Name	Top 5 words
833448	Social	Unfancy Food Show	honeys, bees, honey, sweetness, mcclure's
454079	Social	Swinger's Auction	bedford, auction, bidders, prices, borough
381393	Media	Book Signing with Garry Kasparov	life, view, chess, forms, substances
491481	Social	Washington's Inaugural Address 219 Years Later	washington's, president, bill, continental congress, government
860128	Music	Pedals & Pumps: Ludmila Golub	festival, feast, hemisphere, instrument, consoles
813040	Festivals	Egg Roll and Egg Cream Festival	bus, train, canal, power, eldridge street
917489	Media	An Evening of Listening Pleasure	poetry, book, poet, night, form
227838	Other	Chance Encounter: Free, Live & Experiment Urban Musical Concert	music, experience, concept, susan narucki, life

Table 8.5: Wikipedia enrichment of 25 Yahoo! Upcoming events (top 5 words)

8.1.3 Clustering

In the fourth experiment [Alves et al., 2011], we proposed an approach to visualize the urban space through tags taking into account available online information (static knowledge) about, and popularity (dynamic knowledge) of, places on the social Web. To accomplish this, on top of extracting and enriching a massive quantity of POIs for a given city via three perspectives (About, Red and Yellow Wiki), a social network (Gowalla[8.4]) was used to infer the popularity of places among this community, in order to compute the social significance of a given area and to select a tag that best represented it.

Clustering allows the identification of groups of data instances that are similar in some sense. In this context, clustering allowed us to identify groups of nearby POIs in the city according to the geographical distance between their coordinates.

[8.4] http://www.gowalla.com

A subgroup of density-based clustering algorithms was devised to discover clusters of arbitrary shapes where each was regarded as a region in which the density of data instances exceeded a threshold, making them perfect for the identification of "hotspots" of POIs in the city (i.e. places with high concentrations of POIs). In this experiment, we used DBSCAN [Ester et al., 1996] to identify such "hotspots" that would be the basis of the Semantic Enrichment and Visualization processes.

In addition to the traditional TF-IDF values computed for individual POIs against other indexes in the POI database, we also used Gowalla to infer a popularity-based TF-IDF value for the terms of a POI p in a given cluster c using the POI check-ins. The idea was that concepts associated with POIs that were very popular should be weighted favorably. Equation 8.1 shows how the popularity-based TF-IDF was calculated for each concept i in a given cluster c based on the POIs p that belong to that cluster.

$$Popularity\text{-}based\ TF\text{-}IDF_{i,c} = \frac{1}{|c|}\sum_{p \in c} TF\text{-}IDF_{i,p} \times check\text{-}ins_p \tag{8.1}$$

Using the greater metropolitan area of Boston as a test scenario, we extracted 156,364 POIs from the Yahoo public API. Each POI had an average of 2 categories and the Yahoo taxonomy was spread across three different levels of specificity, where the top level had 15 distinct categories and the lower level had a total of 1,003 categories[8.5]. Table 8.2 shows the distribution of the extracted POIs across the top categories.

In order to cater for the categories in the clustering process, we adopted a two-level clustering approach. In the first level, we grouped together POIs that were closer to each other according to their proximity in the Yahoo taxonomy, and in the second we applied DBSCAN to these groupings, thus producing clusters of geographically close POIs that also had a similar set of categories. This approach gave us a different perspective of the POI data.

In the clustering phase, we grouped together POIs that shared the same top-level category, and then for each top-level category we applied DBSCAN using the POI coordinates. The parameters of DBSCAN were manually tuned. The goal was to choose a set of parameters that produced a balanced number of clusters that covered most of the different areas of the city. Figure 8.1 depicts the centroids of a possible clustering solution using the POI data enriched via the RedWiki perspective applied to the Yahoo

[8.5] These numbers refer only to the data we collected.

category of the POIs inside them.

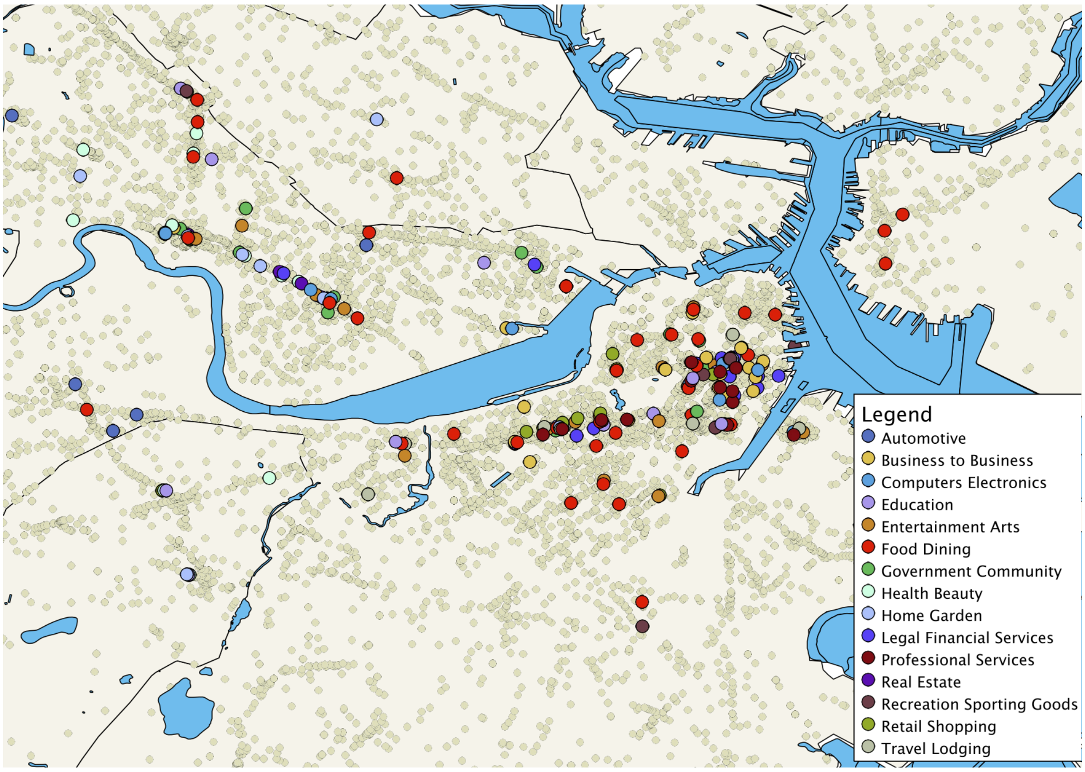


Figure 8.1: Centroids of the clusters identified using the POI data enriched via the RedWiki perspective and the corresponding Yahoo categories.

Figure 8.2 shows, for each Yahoo top category, the percentage of enriched POIs according to each perspective. We can observe the greater coverage of the Red Wiki perspective in contrast to the Yellow Wiki perspective. This can be explained by the fact that almost every POI is categorized under at least one category in Yahoo, and each category was mapped to at least one Wikipedia article (except in the case where there was more than one mapping to Wikipedia, e.g. Computers & Electronics) while in the Yellow Perspective, KUSCO searched for more specific information in Wikipedia, namely the POI article, when it existed. The enrichment process was validated (see section 8.2.2 for details) and we obtained a precision of 62 ± 22%, using a survey covering a population of 30 individuals (visitors or inhabitants of Boston).

In order to visualize and understand the whole process, we considered the top 5

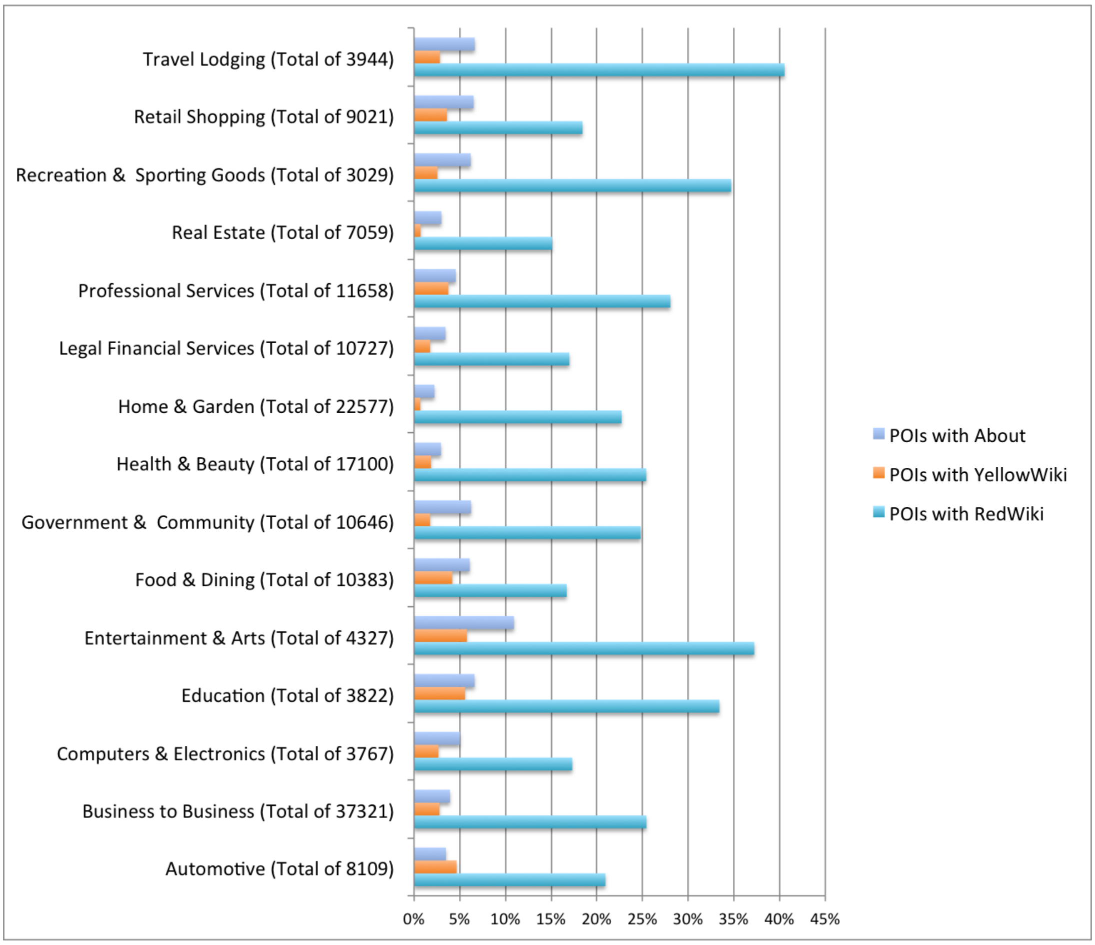


Figure 8.2: POI distribution over the different Yahoo categories for the different perspectives.

most popular categories in Gowalla (and the respective Yahoo category) [8.6] for the greater metropolitan area in Boston. They were: Food (Food & Dining), Shopping & Services (Retail Shopping), Architecture & Buildings (Real Estate), Nightlife (Entertainment & Arts), and College & Education (Education). From the first view of the city (Figure 8.3), we observed a great predominance of common concepts because the system was dealing with generic information associated with the POI categories (the Red Wiki perspective). The more relevant a tag, the greater its font size. For instance, the term *health services* came from POIs which belonged to a category that Gowalla regarded as not so popular: Health & Beauty. However, the concentration of very similar POIs related to the health services was so high [8.7] that this specialized zone was identified and the most relevant tag in all these related subcategories of the top category of *Health* was chosen (regardless of the popularity of its POIs).

Another interesting example that helped us understand how popularity is crucial in determining the most relevant tag was the cluster identified by *secondary education*. In this cluster, different types of POIs were grouped, in this case in a *zone* not as specialized as in the previous example[8.8], but the fact that the most popular POI among them was a High School with a lot of associated check-ins biased the weight of the tag that was ultimately chosen.

Considering a different region of the city (Figure 8.4), by using the Yellow Wiki perspective we found tags that were more specific as we were dealing with the proper Wikipedia article for each POI (when it existed). In the Yellow Wiki perspective, as we only had extracted information about each POI itself, we opted for displaying only the most popular POIs. In this figure, we can see interesting POI-tag relationships: the Cambridge Innovation Center - *business incubator*; Boston Common - *Central Burying Ground*; Massachusetts General Hospital - *Harvard Cancer Center*; Boston University - *Colleges*; Louisburg Square - *Beacon Hill*; Best Buy - *Forbes*; California Pizza Kitchen - *Richard.* The last two examples reflect some of the difficulties that we faced in the present methodology. Firstly, the company "Best Buy" being related to the concept "Forbes" (a magazine) is not very relevant to understand the POI, since

[8.6] Data extracted from Gowalla in May, 2011.

[8.7] e.g. MT Auburn Pulmonary Service, Cambridge Urological Association, Associated Surgeons, Cambridge Gastroenterology

[8.8] e.g. Somerville High School, GC Vocal Studio, Dexter Painting and Carpentry, FISH Magical Enterprises this fact [8.9] is only referred to in the summary of the Wikipedia article in the second paragraph. Secondly, with the last POI-tag pair showed before, it is relatively straightforward to verify that Richard Rosenfield is a co-founder of California Pizza Kitchen. In this sense, if we knew more about this POI via DBpedia, particularly the semantics behind it[8.10], it would be possible to infer more related knowledge (e.g. *founder(California_Pizza_Kitchen, Richard Rosenfield)*).

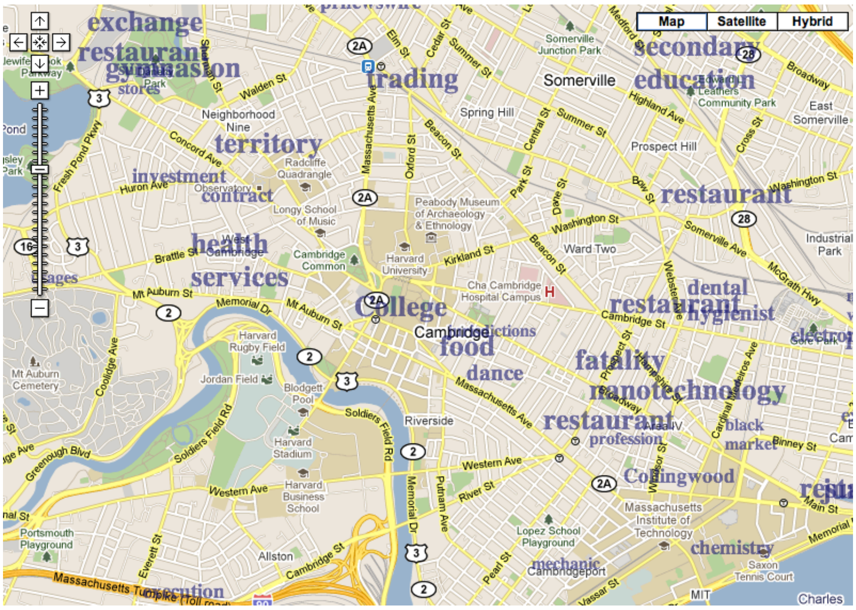


Figure 8.3: The most relevant tags from the Red Wiki perspective using DBSCAN (epsilon=0.0005, minPoints=15) to cluster POIs.

8.2 Validation

Before analyzing the Semantic Indexes produced by KUSCO, it is important to evaluate the quality of the material retrieved for each perspective. In contrast to classical ATR tasks where the source is from a controlled domain and based on plain texts, we deal in this thesis with web pages related to a given place, and this place is represented by a POI. In the following subsections, the two implemented modules of KUSCO, in the scope of this thesis, are evaluated and also the distribution of terms is analyzed.

[8.9] Best Buy won Forbes prizes in two consecutive years.

[8.10] http://dbpedia.org/page/California_Pizza_Kitchen

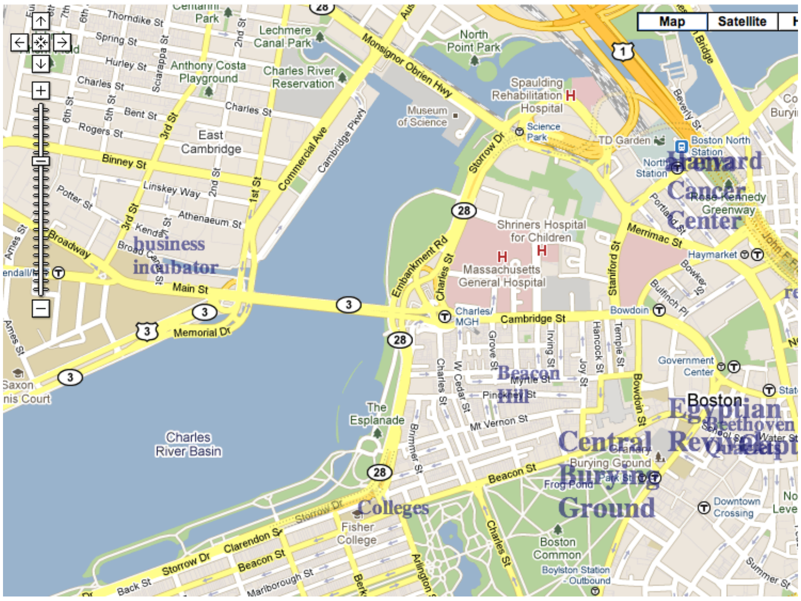


Figure 8.4: The most relevant tags from the Yellow Wiki perspective using the most popular POIs according to Gowalla (no clustering).

8.2.1 Information Retrieval

Since this was a subjective and hard task to perform manually, we asked volunteers to evaluate how a given web page or Wikipedia article was related to the POI indicated. In the Open Web and About perspectives, the evaluation of the retrieved web pages was more subjective than in the Red and Yellow Wiki perspectives, since it was not possible to say that there was only one page that described a place in the Word Wide Web, even in the About section of a given POI website, since it might be structured in multiple web pages, each one depicting one aspect or service offered (e.g. the Massachusetts Institute of Technology - the MIT *about* page[8.11]). In the Red and Yellow Wiki perspectives, the question was more directed because we were evaluating if a given article was "the" category or POI Wikipedia article.

With the Open Web and About perspectives we took advantage of a crowd-sourcing network, such as Amazon's Mechanical Turk (AMT)[8.12], to obtain a large quantity of results in a randomly selected sample of the web pages retrieved. For the Open Web perspective, the task was: "*Given a list of urls choose the most appropriate rating for*

[8.11] http://web.mit.edu/aboutmit/

[8.12] https://www.mturk.com

Perspective	# of POIs	Web Pages per POI	Accuracy		
			0	1	2
1st web page	58	8.5 ± 1.3	23%	36%	40%
2nd web page			27%	34%	39%
3rd web page			23%	32%	45%
4th web page			23%	31%	45%
5th web page			26%	32%	42%
Open Web			Average of top 5 web pages		
			24.7 ± 1.8%	33.1 ± 2.2%	42.2 ± 2.9%
About	50	1	14%	14%	72%

Table 8.6: Evaluation of POI and respective web pages retrieved for Open Web and About perspectives. In the Open Web perspective, each top-k web page was analyzed separately.

each url ranging from 0 (non-related link), 1 (containing several places including this), or 2 (centered only on this place) in relation to the Place indicated". Then, for each POI contextualized by its name, by the Yahoo! Local categories and by the Yahoo! Local url, an average of 10 url links were ranked by volunteers. The About perspective task was slightly different, as the set of possible web pages considered as *about* pages was smaller. In this case, the task was: "*Given a web page verify if it is the right one describing the Place (the Info/About Page). Please choose 0 (it is not the Info Page), 1 (ok, but there is better), or 2 (it is the Info Page)*'. Table 8.6 shows the results obtained for each perspective. While for the Open Web perspective the top 5 most relevant web pages were presented, for the About perspective only the most probable About page was presented and the volunteer evaluated the quality of information that it contained. For all the web pages presented, the same range of appropriateness was presented, from 0 (poor) to 2 (good). From the answers collected, the About perspective had the higher level of precision with regard to the right information, as 72% of the web pages presented were considered as the *About* page for each POI.

The level of agreement between participants in the AMT evaluation task was obtained by the *kappa statistics* introduced in section 2.4.4 in chapter 2. For each item being classified, three participants were invited to rate it in one of three distinct categories (0, 1 or 2). Since more than two participants were considered in rating a item,

Perspective	**Open Web**	**About**
# of items	58 POIs × 5 web pages	50 POIs
Total of distinct raters in the task	15	19
# of ratings per item	3	3
# of categories	3	3
kappa statistics		
Proportion of rater agreement - P_a	0.42	0.73
kappa value	0.11 ± 0.4	$0.40 \pm .06$
Confidence interval		
Lower 95% limit **Upper 95% limit**	$[0.035, 0.186]$	$[0.277, 0.518]$

Table 8.7: Inter-volunteer agreement for Open Web and About perspectives. Each POI was repeated 3 times and evaluated by different volunteers.

the *Fleiss kappa* value[8.13] 2.14 was computed using a SPSS[8.14] macro[8.15] [King, 2004] in both perspectives and is presented in table 8.7. As it would have been impracticable to ask the same participants to complete all the surveys for all the POIs under evaluation, 15 and 19 participants answered a total of 58 × 3 = 174 and 50 × 3 = 150 surveys for the Open Web and About perspectives respectively. In the case specific of the Open Web Perspective, each web page from the top-5 was considered a different item to be ranked by raters. From the kappa values obtained, we were able to conclude that the validation of the About perspective was more consistent than that for the Open Web, and this was considered a moderate level of agreement. This might have been because of the subjective nature of the task, since it was easier to judge if a given web page was or was not the *About* page than to evaluate the quality of information of a certain web page related to a given POI.

In the Red and Yellow Wiki perspectives, the retrieval of Wikipedia articles was, in our opinion, less ambiguous because the universe of search was only the set of ency-

[8.13] kappa value for short.

[8.14] *Statistical Package for the Social Sciences*, a software package developed by IBM.

[8.15] Available at http://www.ccitonline.org/jking/homepage/interrater.html clopedia pages. For this reason, we opted for a functional test instead of a evaluation survey about the appropriateness of each mapping, as made earlier for the Open Web and About perspectives. In this case, for both perspectives, we took a sample of random POIs and asked some volunteers from the AmILab (6 for the Red Wiki perspective and 5 for the Yellow Wiki) to classify each article with regard to the perspective in question.

For the Red Wiki perspective, volunteers were asked to rank the mapping between category names and 1,018 Wikipedia articles. Only the non-root categories (automatically mapped to Wikipedia) were evaluated. As in the previous process, different answers were accepted as an evaluation of the Wikipedia mapping for each category: 0 - completely unrelated, 1 - ambiguous, 2 - right. However, in contrast to the previous process, since every Wikipedia article in the random sample was evaluated just once by one volunteer, no deviation is presented in the final overall precision for each perspective. In this perspective, different levels of abstraction appeared when dealing with a POI and its related category articles. Sometimes it was not possible to find the specific category, in which case the upper level category article was used. For example, the specific category of POIs named *Bathroom Remodeling* was not at that time referred to by any article describing its meaning. In this case, the proposed article mapping was related to the upper POI category *Construction, Repair, & Improvement.* Another situation of a lack of specificity occurred when the category name of a POI did not correspond exactly to the meaning presented in the article, but was somehow related, and therefore was not considered completely wrong. As a specific example, we were able to observe this phenomenon in the category *Business Associations.* Although the name was correctly found in Wikipedia[8.16], it is redirected to an article titled "Companies law". Besides the fact this article seems at first sight only about law applied to companies, it explains the concept and types of business organizations. A concrete example of ambiguity, due to the redirection of related Wikipedia titles but not exactly meaning the same concept, is the the category *Drug Stores*[8.17]. Despite the name is correctly found on Wikipedia it is redirected to *Pharmacy*[8.18] and this article is more related to the profession and not the place where medicaments are dispensed, being the article titled *Community Pharmacy*[8.19] the more precise choice.

For the Yellow Wiki perspective, two test sets were prepared with a random sample

[8.16] http://en.wikipedia.org/wiki/Business_organizations
[8.17] http://en.wikipedia.org/wiki/Drug_stores
[8.18] http://en.wikipedia.org/wiki/Pharmacy
[8.19] http://en.wikipedia.org/wiki/Community_pharmacy

Perspective	**# of Wikipedia Articles**	**Correspondence**	**# of Volunteers**	**Avg. of Articles per Volunteer**	**Accuracy**		
					0	1	2
Red Wiki	1018	1 Wikipedia article for each category Each POI may have more than 1 category	6	124.8 ± 4.9	5%	17%	**78%**
Yellow Wiki	100	1 Wikipedia article for each POI	5	20 ± 7	71%		
Yellow Wiki (only with Infobox)	100	1 Wikipedia article for each POI	5	20 ± 5	86%		

Table 8.8: Validation of Wikipedia articles retrieved for the mapping of categories and POIs in Red and Yellow Wiki perspectives respectively

of 100 POIs in each one. In this perspective, the question posed to volunteers was simply *"Is this* ***the*** *specific Wikipedia page for this POI?"*. The first test set of Wikipedia articles was randomly chosen from the complete set obtained from mapping from POIs to Wikipedia articles. The second, in its turn, was randomly chosen only from those Wikipedia articles which had an info-box. The overall precision of the Red and Yellow Wikipedia mapping is presented in table 8.8.

8.2.2 Automatic Term Extraction

An important characteristic of a given Automatic Term Extraction system is to compute how efficient it is by a methodological evaluation process. As a possible solution for comparing the terms extracted, manual term extraction by humans is not widely accepted as a feasible option, not only because of its cost, but also because of its lack of agreement. In fact, manual assessment of users' tags by human evaluators has shown a level of precision of 59% [Mishne, 2006] and 49% [Sood and Hammond, 2007].

Looking at the results obtained in the recent Keyphrase Extraction Contest [SemEval Portal, 2011] which was devoted specifically to the extraction of keyterms from scientific articles, we should note that, although the precision and recall of most current keyphrase extractors is still much lower compared with other NLP tasks (in the range $\in [0.05, 0.31]$[8.20]), this does not indicate poor performance because different annota-

[8.20] http://semeval2.fbk.eu/semeval2.php?location=Rankings/ranking5.html tors may assign different keyphrases to the same document. As described in [Wan and Xiao, 2008], when two annotators were asked to label keyphrases on 308 documents, the kappa value for measuring inter-agreement among them was only 0.70.

Generally speaking, the Meaning Extraction module was expected to bring out the relevant concepts and entities mentioned in textual descriptions about entities. For a comparison with another Automatic Term Extraction system, we chose to use the Yahoo! Term Extraction API (Yahoo!TE) [Yahoo!, 2009] in order to compare their precision and recall on the same task. The Yahoo!TE API provides a list of significant words or phrases extracted from a larger content and is currently used to create indexes for web pages for Information Retrieval purposes.

Concerning to a generic type of entity, we randomly selected a set of Wikipedia article abstracts from the category “artists”. Then, we compared the performance of the Meaning Extraction module in KUSCO against Yahoo!TE over a set of 1,527 Wikipedia article abstracts manually annotated by 20 volunteers. These consisted of volunteers with either B.Sc. or M.Sc. degree in computer science and were recruited in the Center of Informatics and Systems of the University of Coimbra. The specific task presented to them was to highlight the most relevant terms to characterize a given artist.

Table 8.9 show the overall precision and recall of KUSCO using simple TF-IDF statistics and Yahoo!TE considering the top 5, top 10 and all relevant tags. An example of a given Wikipedia abstract is also showed in table 8.10, where the tags from each participant in the test (human, KUSCO, Yahoo!TE) are highlighted. Even in a general task of Automatic Term Extraction, the system showed interesting results outperforming the precision presented by Yahoo!TE, but presented lower values for recall. This can be due the fact that Yahoo!TE was more restrictive than KUSCO, building shorter indexes than our system, as so as the number of tags highlighted by annotators that were in average not longer than 7 tags $\approx 6.1 \pm 4.6$. Also term weighting plays an important role, since if the whole set of keywords tagged by humans is taken into consideration, the precision of the system increases.

In respect to geo-referenced entities, in our case POIs, we compared the Meaning Extraction module against Yahoo!TE using POI web pages. As there was no location-based semantics benchmark dataset to test and validate our results, we compared both systems in terms of qualitative information extracted in order to examine the diversity

Term Extraction System	Index Length	Precision at k	Recall	F-measure
KUSCO	Top 5 ($k = 5$)	**0.31**	0.24	**0.27**
	Top 10 ($k = 10$)	**0.42**	0.32	**0.36**
	All avg. 14.3 ± 13.6	**0.55**	0.36	**0.41**
Yahoo!TE	Top 5 ($k = 5$)	0.30	**0.26**	**0.27**
	Top 10 ($k = 10$)	0.39	**0.35**	0.35
	All avg. 8.2 ± 6.2	0.48	**0.39**	0.38

Table 8.9: Precision levels at 5, 10, and all tags from KUSCO and Yahoo!TE compared with human-annotated tags on Wikipedia abstracts.

Wikipedia article abstract	{[<Martin Mull>]} (born August 18, 1943) is an {[<American actor>]} who has starred in his {[own <television sitcom>]} and acted in {[prominent <films>]}. He is also a {[<comedian>]}, {[painter]}, and {[recording artist]}. He is a {[<satirist>]} and incorporates his {[comedic sense]} into all of his {[work]}.

Table 8.10: Example of a Wikipedia article abstract tagged by 3 different competing systems: human (enclosed in "[]"), KUSCO (in "{}") and Yahoo!TE (in "<>").

and richness of our module.

Using the same web pages for both systems, we needed to understand *the contribution of the Meaning Extraction module compared to the Yahoo!TE API*. Since neither the Yahoo!TE nor a search over the Web accepted a location as a parameter, we applied the same selected web pages (downloaded for each POI) as input for this API output. For each text, Yahoo!TE extracted the most relevant terms, which were then contextualized in WordNet in the same way as KUSCO does. Considering a threshold baseline of equal value for both Extraction Meaning systems, the Semantic Indexes produced by KUSCO had an average size of 35 terms (both concepts and named entities), while those built using the Yahoo!TE API had 44 terms on average. For further comparison with the Yahoo!TE, we applied the Information Content (IC) measure from the WordNet concepts (a combination of specificity and term frequency of each term in relation to large corpora [Resnik, 1995]) to both semantic index lists. In this respect, KUSCO and Yahoo had very similar results with similar standard deviations (71 ± 6.0% and 70 ± 6.0% respectively). Looking from the same perspective for Named Entities, we sought the average TF measure for these concepts in both approaches. Here, KUSCO slightly outperformed Yahoo (59% and 50% respectively). These measures, however, meant that both systems had the same level of efficiency (with a slight advantage for KUSCO) in obtaining valid concepts regardless of whether or not they were significant for the meaning of the place (e.g. concepts like "New York" or "address", present in the Yahoo!TE index, received a high IC value or high frequency, without adding new novel information about a place).

From the 2 Wikipedia perspectives (Red and Yellow Wiki), we randomly selected a sample of 420 Semantic Indexes about Boston POIs which were then manually validated, this time by 30 volunteers who knew the city in question and who, for each term, judged whether it was related to the POI or not (just two possible answers). In some cases the volunteers were not sure about the appropriateness of some terms instead of others which was advised to leave the evaluation of those term in blank. We obtained an accuracy level of 60 ± 14% and 62 ± 22% for the Yellow and Red Wiki perspectives respectively, considering all unanswered terms as invalid.

In order to generalize the validity of Semantic Indexes to include other perspectives, we randomly chose a large set of indexes to evaluate with AMT volunteers covering all static perspectives. The selected indexes from Boston were weighted with Semantic

TF-IDF (presented in section 5.5). The POIs whose Semantic Indexes were being analyzed were contextualized by their name, by Yahoo! Local categories and by Yahoo! Local urls, and a set of top 5 keywords was presented to be evaluated according to their level of relatedness to the given POI. The same question *"Given a list of tags choose the most appropriate rating for each tag ranging from 0(non-related), 1(non-relevant), 2(somewhat relevant) to 3(highly relevant) in relation to the Place indicated"* was presented three times for each POI through the 4 distinct perspectives: the Open Web, About, Yellow Wiki and Red Wiki perspectives; and two baselines perspectives: Naive Random and Random. The first baseline, Naive Random, considered for each POI a random set of concepts from the universe of concepts in the database. The second baseline, Random, limited this universe to the union of concepts from different perspectives of a given POI. Table 8.11 summarizes the results obtained.

Each volunteer answered for one POI at a time considering the top-5 tags selected by KUSCO, with each volunteer being allowed to answer only once for each POI. As the number of participants involved was considerable, table 8.12 presents the kappa statistics in order to verify the level of agreement among volunteers on this task. In the same line with previous AMT evaluation tasks, it would have been impracticable to ask the same participants to complete all the surveys for all the POIs under evaluation. A total of 49 distinct participants answered a total of $(96+50+50+50+50+50)\times 3 = 1038$ surveys. Since more than two participants were considered to rate each item (in this case, a tag), the *Fleiss kappa* value[8.21] 2.14 was computed for the Naive Random and the other perspectives. This distinction was made because we believed the Naive Random was clearly poor in quality and it might had influenced the final evaluation. The categories chosen by raters were also grouped and the respective kappa value was computed once again. The kappa value considered their agreement in four and two categories of answers: relevant (score of 3 and 2) and non-relevant (score of 1 and 0)[8.22]. From the kappa values obtained, we were able to conclude that, as expected, there was more agreement when the number of categories was lower, and also that the validation of the Naive Random perspective was less consistent than the others. These others perspectives had a fair level of agreement, considering again 2 categories of answers (relevant and non-relevant). This might had been because the tags presented in the Naive Random perspective were so random that would be impossible to predict the answers from raters. In respect to the other perspectives, the agreement was indeed

[8.21] kappa value for short.

[8.22] As showed in table 8.11.

Perspective	# of Semantic Indexes	Accuracy			
		0	1	2	3
		non-relevant		relevant	
Naive Random baseline	96	95 ± 2%	3 ± 1%	1 ± 1%	1 ± 0%
		98%		2%	
Random baseline	50	51 ± 6%	23 ± 4%	13 ± 3%	13 ± 2%
		74%		26%	
Open Web	50	49 ± 2%	22 ± 1%	20 ± 1%	9 ± 1%
		71%		29%	
About	50	38 ± 2%	22 ± 3%	23 ± 3%	17 ± 5%
		60%		40%	
Yellow Wiki	50	37 ± 2%	22 ± 4%	23 ± 4%	18 ± 4%
		59%		41%	
Red Wiki	50	32 ± 3%	21 ± 3%	24 ± 2%	23 ± 2%
		53%		47%	

Table 8.11: Evaluation of Semantic Index relevance from different perspectives.

below to the moderate level of agreement (< 0.4) indicating that tagging depends on individual opinion. Another possibility is that this specific task was more subjective than the evaluation about the relatedness of the About web page of a given POI 8.7. This subjectivity suggests that the evaluation of KUSCO must also include a broader statistical evaluation independent of human evaluation as presented in the next section.

8.2.3 Term Coherence

We faced an important challenge in understanding the actual quality of the results in terms of the *correctness* of the words assigned to places. Any *ideal* list of words is by nature subjective. As mentioned above, a place can be defined according to different perspectives, and each perspective can vary according to the subject. In terms of validation, this raises difficult questions even for the typical user survey. A very large sample of people who *know* the specific places is necessary in order to achieve believable results, but this then becomes impractical and costly. We decided, firstly, to analyze our results according to two dimensions: category consistency and coherence among perspectives. We also analyzed the distribution of the words in each perspective.

Each POI has ultimately one category[8.23], so, in the *category consistency* validation,

[8.23] In reality, only a portion of the POIs have a single category, but we determined the *Least Common*

Perspective		**Naive Random Baseline**	**Random Baseline, Open Web, About, Yellow and Red Wiki**
# of items = 5 tags × # POIs		96 POIs	50 POIs × 5 perspectives
Distinct raters in the task		16	33
# of ratings per item		3	
kappa statistics for 4 categories			
Proportion of rater agreement P_a		0.92	0.39
kappa value		0.16 ± 0.02	0.15 ± 0.01
Confidence interval			
Lower 95% limit	**Upper 95% limit**	$[0.119, 0.200]$	$[0.127, 0.165]$
kappa statistics for 2 categories Non-Relevant/Relevant			
Proportion of rater agreement P_a		0.97	0.71
kappa value		0.10 ± 0.03	0.37 ± 0.02
Confidence interval			
Lower 95% limit	**Upper 95% limit**	$[0.053, 0.156]$	$[0.339, 0.403]$

Table 8.12: Inter-volunteer agreement for rating indexes from different perspectives. Each index was evaluated 3 times by distinct volunteers.

the task was to verify the persistence of the word patterns according to those categories (15 for POIs and 10 for events). The first approach was to apply a clustering algorithm such as K Means, where K corresponds to the number of different categories. After clustering with a training set, we applied a classification task: given the top 5 words of a POI from the test set, we classified the POI in one of the categories. We applied 10-fold cross validation[8.24]. In order to get basic benchmarks to analyze the results, we set up two baselines: the *random baseline* consisted of the accuracy of a random classifier (applied to all cases of the data set) and the *fixed baseline* classifier selected the most popular class.

The resulting accuracy of clustering was in fact extremely poor, even when compared with the baselines. The highest value obtained (37.66%) was for the Open Web perspective which was actually lower than the *fixed baseline* of 47.49% (the accuracy obtained by a dumb model which basically always assigned the most popular category to any POI). This implies that either the word patterns were not consistent with respect to category or they were too elaborate for the clustering algorithms to deal with. We tried Bayesian networks, which are actually more common in text categorization, and the results improved considerably (the level of accuracy ranged from 57.12% to 97.3%). In Table 8.13, we summarize the results.

The high value for the Red Wiki perspective reflects the fact that our algorithm was able to extract sufficiently specific words from Wikipedia category definitions such that they became easily distinguishable from each other. This was interesting because many POIs (those that had multiple categories) were assigned a more generic category for classification, and thus POIs from different original areas of the category hierarchy and with many different words were gathered in the same *class*. Taking this into account, we were able to conclude that the original assignment of categories to the POIs was itself very consistent (e.g. Food & Dining subcategories were rarely mixed up with Home & Garden ones).

For the Open Web perspective, a careful analysis revealed that there was still a reasonable quantity of noise in the indexes, while in the Event perspective, we were categorizing *events* as opposed to places, producing less stable patterns. For example,

Subsumer for POIs with multiple categories in the hierarchy, which consists of the most specific upper category that contains the categories of the POI in its descendants.

[8.24] Divide data set into 10 folds; each fold will become a test set for a model built with the remaining 9 folds [Witten and Frank, 2005].

	Red Wiki	**Yellow Wiki**	**OpenWeb**	**Event**
Random baseline	9.199%	13.483%	23.781%	17.3%
Fixed baseline	16.54%	23.25%	42.49%	13.86%
K Means	24.40%	28.26%	18.71%	24.93%
Bayesian Network	97.3%	66.56%	57.12%	51.08%

Table 8.13: Category consistency results.

an exhibition about food pictures would produce words that matched with a gastronomy festival (thus, a different kind of event), while in the other perspectives a POI from the category of Restaurants ended up becoming recurrent (as gastronomy-related words should not appear in the other types of places). This might be due to the fact that we were collecting information from different Web sites with distinct templates and lateral information (e.g. advertisements, ads by Google, news headlines from RSS feeds, etc.) while in Wikipedia we had an available API to extract only useful and structured information. We could thus conclude that, for all perspectives, our system brought a degree of consistency that is relevant, particularly considering the two baselines.

With regard to event categories, since Yahoo! Upcoming provides category information for each event extracted, we were also able to perform classification analysis considering this type of information (clustering with K Means and “Farthest First” [Hochbaum and Shmoys, 1985] and generating a priori association rules [Scheffer, 2001]). The results show some coherence between the Event perspective and the event categories, which is remarkable for such a small set. In table 8.14, we can see that K Means organized data into 4 categories(“Performing/Visual Arts”, “Music”, “Education” and “Comedy”), and some of the words were not directly related to the categories (e.g. “dr” or “countries”). With the “Farthest First” algorithm, a centroid for each category was inferred from the examples. We also applied the a priori algorithm (table 8.15), confirming the same general conclusion: each event category attracted its own set of words.

The analysis of relatedness among the perspectives allowed us to see the recurrence of word patterns from the different sources. The assumption was that for the same POI the words from different perspectives should be related and/or similar. This analysis was limited, however, to the POIs that had already been analyzed for more than one

Cluster Analysis		
KMeans	Farthest First	Category
artists, music, seminar, countries	play	Performing/ Visual Arts
dr	bag	Education
myspace, music	music	Music
	market	Social
	art	Festivals
	behaviour	Media
	forms	Commercial
	school	Sports
	wars	Politics
film	farce	Comedy

Table 8.14: Word cluster centroids according to event category

perspective (which corresponds to more than 8,500 over a total of 120,000 semantic indexes). The sample sets was randomly chosen among POIs which had both perspectives and their semantic index in each perspective had 5 concepts at least. Firstly, for each comparison of perspective pair, we created a sparse matrix of Top-15 Terms/Indexes occurrences by weighting each occurrence using the semantic TF-IDF weight of each term. Then, the Cosine Similarity (equation 2.4) was computed between indexes from the two different perspectives for a given POI ranging from 0 (most dissimilar) to 1 (most similar). Table 8.5 presents box-plot with the distribution of similarity values computed between semantic indexes from different perspective for the same POI in the Boston Area. For a total of 4,076 semantic indexes in different perspectives whose POIs were randomly selected, only non-zero/non-missing similarity values were considered. Table 8.16 shows the representativeness of which pair of perspectives.

Terms could be simple, WordNet or Wikipedia terms, and two terms were considered similar according to the algorithm 5.3 presented in section 5.4. In respect to integration of terms, those terms present in WordNet (i.e. WordNet terms) were contextualized as concepts (synsets) in order to identify synonyms from different indexes. For example, if we had the term “nightclub” in a given perspective mapped to WordNet as the concept “a spot that is open late at night and that provides entertainment

word	→	category	accuracy
century		Performing/Visual Arts	0.99318
religion		Performing/Visual Arts	0.99318
practices		Performing/Visual Arts	0.99318
choir		Performing/Visual Arts	0.99318
myspace		Music	0.98849
hills		Music	0.96257
legislature		Education	0.96257
friends		Music	0.96257
david crane		Music	0.96257
musicians		Music	0.96257
bar		Music	0.96257
images		Performing/Visual Arts	0.96257
children		Family	0.96257
dj		Music	0.96257
jockey		Music	0.96257
instrument		Music	0.96257
artists		Performing/Visual Arts	0.95582

Table 8.15: An excerpt from the association rules with highest accuracy found (the numbers on the right indicate level of accuracy).

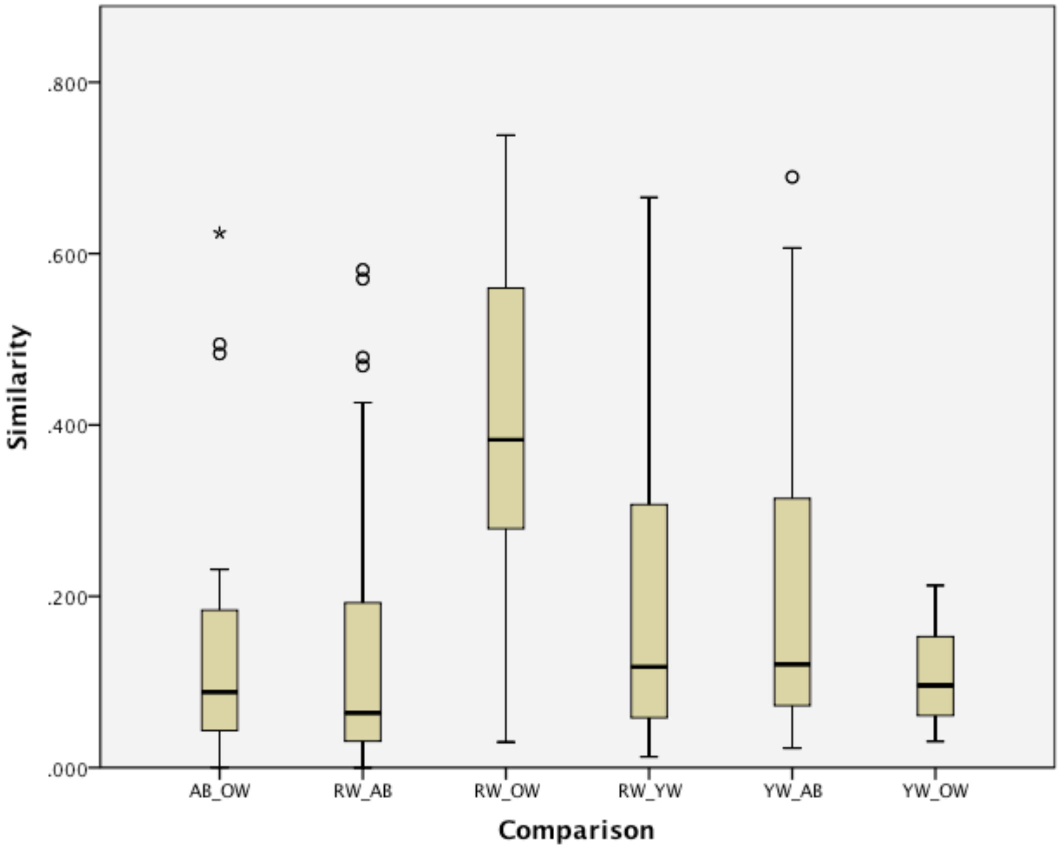


Figure 8.5: Semantic similarity among perspectives where RW means Red Wiki, YW Yellow Wiki, AB About and OW Open Web pespective.

(such as singers or dancers) as well as dancing and food and drink"; then any word representing this concept will be considered as a match (e.g. cabaret, night club, club, nightspot). Wikipedia terms were considered equivalent if they were redirected to the same Wikipedia article (e.g. "Western countries" is equivalent to "Western world"). The remaining terms were compared by String Similarity in order to find little small variations in names, like "operating system market and "operating system marketing.

Analyzing the numbers in chart 8.5 and table 8.16, we can see that the effect of filtering already known information about POIs, like category names in the Red Wiki perspective, and the POI address related terms in the other perspectives, made the remaining terms more likely to be unique and diverse in each perspective. The low representativeness of the Open Web perspective, we believe, was due to its quantity of

		Semantic Indexes					
		Valid		Missing		Total	
	Comparison	N	percent	N	percent	N	percent
Similarity	AB_OW	50	15.4%	274	84.6%	324	100.0%
	RW_AB	144	12.5%	1004	87.4%	1148	100.0%
	RW_OW	46	4%	1116	96.0%	1162	100.0%
	RW_YW	168	16.6%	846	83.4%	1014	100.0%
	YW_AB	102	23.8%	326	76.2%	428	100.0%
	YW_OW	24	7.0%	320	93.0%	344	100.0%

Table 8.16: Number of semantic indexes considered in the initial set randomly selected.

noise, making this perspective semantically distant to the majority of concepts in the other perspectives. Considering the 3 most representative comparison pairs in table 8.16, the assessments of the Red Wiki perspective were the most related to the others. Actually this perspective shows the overall highest similarity in average with the Yellow Wiki perspective. This was explained by the common domain used by both perspectives, the Wikipedia. As the About perspective was not so easily obtained for several reasons, such as the lack of official website reference in the POI directory information that was retrieved, the sample sets used in the comparison of this perspective were shorter than those retrieved in the other perspectives. Even so, for those POIs when it was possible, this perspective showed the highest similarity to each of the other perspectives.

The best case of similarity was the closest pair for a given POI, and the worst case was the one that was farthest apart. The overall best and worst cases are detailed in Table 8.17 and 8.18 respectively. Terms mapped into WordNet and Wikipedia are complemented with synonyms and related terms enclosed by parentheses, "()". Looking at the examples, the disparity between perspectives was not a disadvantage but a richness acquired by the contribution of distinct terms from different perspectives. The common concepts between perspectives seem to be relevant to the POI in question. For instance, looking at the Wikipedia article about the POI *Harpoon Brewery*[8.25], the concept *Windsor* is related to the location of the second oldest Harpoon Brewery in U.S. (apart from the original in Boston). In addition to this, perhaps, for those

[8.25] http://en.wikipedia.org/wiki/Harpoon_Brewery concepts that are not exactly synonymous but are specializations (e.g. brewery and microbrewery) or are related (e.g. tomatoes, onions, lamb and potatoes are food) we could apply other semantic similarity measures.

Finally, we also checked the shape of the word frequency histogram. As we can see in Figure 8.6, in every perspective, the distribution of words followed the typical long-tail distribution that matches Zipf's law for word frequency [Zipf, 1932], as expected.

POI	**Location**	**Categories**
Marketplace Technologies	342 Broadway Cambridge, MA	**Investment Services**, Financial Planning, **Marketing Agencies**
Perspective	**Similarity**	**Terms**
Red Wiki	0.738	Good, income, Industry trade group (Industry association, Trade Association, Trade Organization, Trade bodies), shares, product line, ...
Open Web		**Marketing Agencies**, **Investment Services**, Regional Account Reps, Morgan Stanley Smith Barney, Brattle Sq, ...
Thrift Shop	1194 Washington St, Norwood, MA	**Thrift Stores**
Perspective	**Similarity**	**Terms**
Red Wiki	0.666	charitable organization (Charitable cause, Charity Law, Voluntary Welfare Organisation), **costs**, **op shop (Charity shops, Opportunity shop, Resale shops, Second hand store, Thrift Store)**, second-hand shop, second-hand good (Pre-owned, Used Goods), ...
Yellow Wiki		**costs**, hospice, mortgage, Charity, ..., **Charity shops (Op Shop, Opportunity shop, Resale shops, Second hand store, Thrift Store)**, ...
Harpoon Brewery	306 Northern Ave, Boston, MA	Tourist Attractions, Bars & Pubs, Breweries, All Bars, Pubs, & Clubs, Restaurants, American Restaurants, Wineries
Perspective	**Similarity**	**Terms**
About	0.607	brewery, **Beer**, **Windsor**, taste (taste sensation, gustatory sensation, perception, gustatory perception), neighbor, ...
Yellow Wiki		**Windsor**, **Beer**, Vermont, St. Patrick's Day (Green beer, St. Patrick's Day, Parade, St patty day), microbrewery, ...

Table 8.17: Examples of High similarity between perspectives for a given POI.

POI	**Location**	**Categories**
Sports Club Management	400 Washington St, Braintree, MA	Human Resources, Other Business Services, Management & Consulting
About	≈ 0	domain (knowledge domain, knowledge base), Tennis (lawn tennis), Verizon, safety (base hit), SCM, ...
Open Web		beauty salon (salon, beauty parlor, beauty parlour, beauty shop), document (written document, papers), Emergency (exigency, pinch), Bonus (incentive), Rowan (rowan tree, European mountain ash, *Sorbus aucuparia*), ...
Enterprise Rent-A-Car	522 Washington St, Stoughton MA	Car Rentals, Truck Rentals
Open Web	0.031	Rutland, Passenger Car Rental, Free Pick-Up, USA Cont, Careers, ...
Yellow Wiki		Missouri, Questions, lists, air bags (Sensing and Diagnostic Module, Supplementary Restraint System), Clayton, ...
Passage to India	1900 Massachusetts Ave, Cambridge, MA	Catering Services, Indian Restaurants, Restaurants, Southeast Asian Restaurants
About	≈ 0	spices, Lamb, potatoes, tomatoes, onions, ...
Red Wiki		chefs, cuisine (culinary art), India, evolution (development), Voyages, ...

Table 8.18: Examples of Low similarity between perspectives for a given POI.

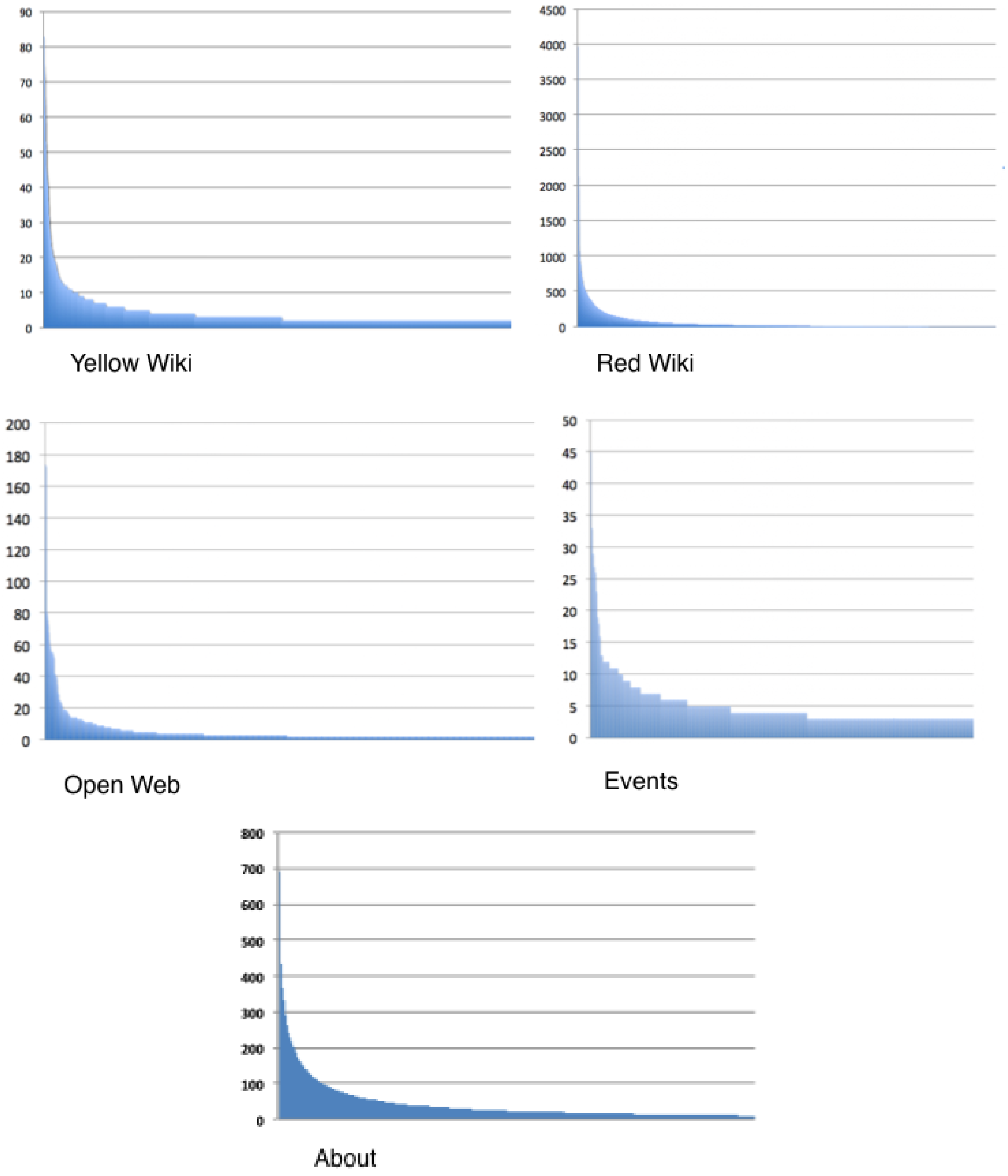


Figure 8.6: Word distribution according to perspective.

9

Conclusions and Future Work

This chapter presents the main contributions of this thesis. The main areas of contribution are: Context Awareness, Information Retrieval, Automatic Tagging and Term Weighting. Alongside the development of KUSCO, new research directions were also identified, but due to the research required for these, they have been postponed for future work. Here, we present new research challenges and improvements that can be performed on Semantic Enrichment.

9.1 Thesis Contributions

In this thesis, we have defined the general process of *"Semantic Enrichment"* of entities and the approach proposed to deal with the specific scenario of *Semantic Enrichment of Places*, thereby developing a system, KUSCO, which builds a semantic index associated with a given Point of Interest (a latitude/longitude pair and a name). For each POI, KUSCO finds related information on the Web and executes a sequence of Information Extraction and Natural Language Processing steps to automatically extract the relevant related terms. Each term is contextualized in lexical resources (WordNet and Wikipedia) which guide the extraction process by validating common-sense terms and which are also used to infer the meaning of each term. Once these terms are contextualized, they are called concepts. Their relevance is computed through an extended version of TF-IDF, which considers the semantics of each term. The system has also been subject to a series of tests. In comparison with related work, specifically, the generic term-extraction tool from Yahoo (Yahoo! Term Extraction), KUSCO has shown better results.

The main contribution in this work includes a clear and well defined methodology for creating semantic information about places from web pages and a new set of benchmarking datasets (semantic indexes) that allow for future comparisons and gradual improvements on the current results and methods. The implemented system is able to gather a massive amount of POIs and analyze a considerable proportion of these, clearly enough for a valid analysis. The process is fully unsupervised since no labeled examples are needed. The experiments show that the semantic indexes obtained have a generally good quality, and we have presented several different "perspectives" that can be used according to the context.

We have also presented a methodology for extracting semantic information about arbitrarily sized areas, depending on the availability of Points of Interest. The nature of this process is ultimately subjective since all the information is extracted automatically from crowd-sourced resources. However, we rely ourselves on techniques that favor statistical relevance and specificity to select the words (or tags) that should better represent the context according to "how people understand that space". Furthermore, the concept of "perspectives" explicitly models the unavoidable ambiguity in this problem. We have taken two approaches to Wikipedia and two approaches to the World Wide Web as four ways of understanding the same space.

KUSCO is currently being applied in the context of three different research projects as a semantic enrichment source: an intelligent route planner; a project for analysis of correlations of cell-phone activity and events in the city; and a platform for fusion of traffic and land use data.

In conclusion, for each contribution expected in section1.3 we explain how it was achieved during the development of this thesis:

- *a generic model of semantic enrichment* which was presented in Chapter 3. This model is mainly based in searching and extracting online information. This information is mapped to knowledge sources in order to let external systems infer more information. It was conceived to be ideally extended to any named and georeferenced entity that is present in the Web and its related information is indexed by a search engine. The geographical location of these entities is used mainly to disambiguate the information retrieved. Additionally, this geo reference is also important to adapt NLP algorithms used in the process of information extraction and to select alternative knowledge sources in the native language spoken in that location;

- *a modular methodology for the assignment of semantics to a place, which is divided into three main steps*: *the retrieval* of related information on the Web which was mainly discussed in Chapter 4, *the extraction of terms* from this information (mostly textual descriptions), and *the contextualization and computation* of the relevance of these terms which were mainly discussed in Chapter 5. In these steps some state-of-the-art algorithms and systems were reused, namely a search engine in the retrieval step and NLP algorithms in the extraction module. Therefore, it was possible to focus the work in the whole semantic enrichment system and to leave specific sub-tasks able to be implemented by similar algorithms independently of its nature (e.g. a search engine based on the Page Rank vs. HITS ranking algorithm, or rule-based vs. stochastic POs tagging). Furthermore, no training phase is required and consequently no previous manually enriched places are required to start KUSCO. Time to time, it is necessary to recompute the relevance of terms recomputing the Semantic based TF-IDF. This extension of traditional TF-IDF was also proposed in order to correlate semantic related terms that otherwise would be considered unrelated;

- *new perspectives of semantic enrichment applied to space analysis, their implementation and a proposed way to correlate these perspectives to a given POI* which were mainly discussed in Chapter 6. By perspective, we mean the source of information and the way the information is retrieved from this source. Specifically speaking, it was proposed two perspectives that take advantage of information available on Wikipedia. On one hand, Red Wiki perspective retrieves generic Wikipedia articles related to the categories of a place. On the other hand, Yellow perspective searches for *the* specific Wikipedia article about a place when it is available. For a more broad coverage in the Web domain, it was also proposed two other perspectives, the Open Web and About perspectives, which respectively searches for web pages related to a place. The latter perspective centers the search for the *About* page of a POI;

- *a proposed representation of place semantics* which was mainly discussed in Chapter 7. This intended representation shares the formalism of the Semantic Web in what concerns to the availability of semantic indexes as a lightweight ontology. In this proposal the concepts extracted by the system are connected into the Linked Open Data initiative;

As another contribution and not less important is the fact that this thesis has been used to show new directions on the research of Information Extraction applied to Ambient Intelligence. These directions are now being followed internally in the AmI laboratory, part of Cognitive Media Systems group of the Center of Informatics and Systems of the University of Coimbra:

- by the CROWDS project which consists on understanding urban land use from digital footprints of crowds. A characterization of population dynamics by type, neighborhood, or region would enable customized services (and advertising), as well as the accurate timing of urban service provisions, such as scheduling transit service based on daily, weekly, or monthly mobility demand. In general, more synchronous management of service infrastructures clearly could play an important role in urban mobility management. One of its main objective shapes on the intersection of information extracted by KUSCO about space with other digital footprints, such as cell phone usage or taxi demand.;

- by the TICE.Mobility project exploring new, more efficient and comprehensive solutions for urban transportation, through the use of communication and information technologies (CIT) to make it possible to integrate the various available solutions, in an ecological, energy-efficient way with better quality for users. KUSCO is being adapted to enrich Portuguese places with automatic detection of retrieved information language. This automatic recognition of language will enable KUSCO to search and handle distinct sources of information depending on the user (or external system) native language;

- by the PhD thesis of Filipe Rodrigues which title is “The internet as a sensor: machine learning approaches to extract relevant information for urban mobility”. KUSCO challenges and results inspired the objectives of this new PhD thesis in order to learn how people tag space.

9.2 Future Work

The future steps for this system include the exploitation of the structured knowledge resources DBPedia and Yago as proposed in chapter 7 that are directly connected to Wikipedia and WordNet, which can provide broader and integrated common-sense semantics as well as specific information on the idiosyncrasies of each POI (e.g. Restaurants have a menu, Museums have a theme). This, we expect, should show better results than using OWL ontologies, a process we tried initially with the same categories of POIs studied (for more details, please see appendix B). In short, with regard to the use of external ontologies, an important point to be noted is that the quality of the ontologies used in the process is crucial to the quality of results obtained, which means that for a more careful selection and evaluation of such ontologies is required. Concretely, an approach which uses semantic enriched places and the Linked Open Data project to instantiate an Upper Ontology was proposed in section 7.2.2. The model which classifies places in a standard taxonomy (NAICS) that is mapped to an Upper Ontology (SUMO) is already built and tested (appendix C), remaining only the instantiation phase to be implemented.

Automatic Tagging revealed a ambitious task to achieve. It was presented in the literature review and verified by the Meaning Extraction module in KUSCO. It could benefit from similar ATR systems using a voting scheme. Also the Named Entity recognizer algorithm could be trained and tested on specific examples of related Web Pages. Also in the relevance computing of terms, the word sense disambiguation proposed in section 5.3 can be extended to consider the given context of a place when choosing the most appropriate meaning for a given term instead of just picking up the most common used for such a term represented in the Knowledge Source.

Choosing the most used sense of a given term in knowledge resources was the disambiguation method used by KUSCO to infer its meaning. With different perspectives of a given place being available, we think the use of this context can help the disambiguation task. This is true mainly for generic and ambiguous concepts that are related in similar perspectives for a given place (e.g. *bank* and *account*, each one present in different perspectives), which can benefit of the domain in question (e.g. Finance Services) and other unambiguous concepts found for the same place (e.g. tax, monetary fund).

The use of popularity, from Gowalla, helps to understand the social dimension of space. From a purely democratic point of view, the more people that enter a place (and report on it positively), the more relevant it is for the community. Of course, this raises questions itself on how representative the population that uses Gowalla is, and how they actually report on the places they visit (e.g. do they check in more often when waiting in a fast-food queue than when having fun?). The work presented will become more representative as such communities grow. As a further improvement of our approach, we plan also to use the POI radius available from Gowalla as a feature to be considered in the cluster algorithm, since currently a very wide POI (e.g. MIT) has the same weight as a small POI (e.g. Starbucks). Other sources could be included (e.g. using Twitter, Facebook, Eventful, etc.) to infer the popularity of POIs.

References

Text Analysis Conference - Knowledge Base Population Track, 2010. URL `http://nlp.cs.qc.cuny.edu/kbp/2010/index.html`. Last visited: December, 2011. 47

S. Abney and S. P. Abney. Parsing by chunks. In *Principle-Based Parsing*, pages 257–278. Kluwer Academic Publishers, 1991. 23

G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles. Towards a better understanding of context and context-awareness. In *Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing*, HUC '99, pages 304–307, London, UK, 1999. Springer-Verlag. ISBN 3-540-66550-1. 11

D. W. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. *Mach. Learn.*, 6:37–66, Jan. 1991. ISSN 0885-6125. 142

D. Ahlers and S. Boll. Location-based web search. In A. Scharl and K. Tochtermann, editors, *The Geospatial Web*. Springer, London, 2007. ISBN 1-84628-826-6. URL `http://www.geospatialweb.com/`. Last visited: December, 2011. 31

R. Aipperspach, T. Rattenbury, A. Woodruff, and J. F. Canny. A quantitative method for revealing and comparing places in the home. In *UbiComp'06*, pages 1–18, 2006. 5

H. Alani, S. Kim, D. Millard, M. Weal, W. Hall, P. Lewis, and N. Shadbolt. Automatic extraction of knowledge from web documents. In *Workshop on Human Language Technology for the Semantic Web and Web Services, 2 Int. Semantic Web Conf.*, 2003. ix, 62, 63

A. Alves, F. Pereira, F. Rodrigues, and J. Oliveirinha. Place in perspective: Extracting online information about points of interest. In B. de Ruyter, R. Wichert, D. Keyson, P. Markopoulos, N. Streitz, M. Divitini, N. Georgantas, and A. Mana Gomez, editors, *Ambient Intelligence*, volume 6439 of *Lecture Notes in Computer Science*, pages 61–72. Springer Berlin / Heidelberg, 2010. ISBN 978-3-642-16916-8. 147

A. Alves, F. Rodrigues, and F. Pereira. Tagging space from information extraction and popularity of points of interest. In D. Keyson, M. Maher, N. Streitz, A. Cheok, J. Augusto, R. Wichert, G. Englebienne, H. Aghajan, and B. Krse, editors, *Ambient Intelligence*, volume 7040 of *Lecture Notes in Computer Science*, pages 115–125. Springer Berlin / Heidelberg, 2011. ISBN 978-3-642-25166-5. 154

A. O. Alves, F. C. Pereira, A. Biderman, and C. Ratti. Place enrichment by mining the web. In *Proceedings of the European Conference on Ambient Intelligence*, AmI '09, pages 66–77, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-05407-5. 146

E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '04, pages 273–280, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4. 17, 31

B. Antunes, A. Alves, and F. C. Pereira. Semantics of place: Ontology enrichment. In *IBERAMIA*, pages 342–351, 2008. 140, 219

S. Asadi, X. Zhou, and G. Yang. Using local popularity of web resources for geo-ranking of search engine results. *World Wide Web*, 12:149–170, 2009. ISSN 1386-145X. 31

N. Association. NAICS Association: Frequently Asked Questions, February 2010. URL `http://www.naics.com/faq.htm`. Last visited: December, 2011. 82

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives. DBpedia: A nucleus for a web of open data. In *In 6th International Semantic Web Conference, Busan, Korea*, pages 11–15. Springer, 2007. 54, 55, 141

C. Becker and F. Dürr. On location models for ubiquitous computing. *Personal Ubiquitous Comput.*, 9:20–31, January 2005. ISSN 1617-4909. 13

REFERENCES

A. Berglund, S. Boag, D. Chamberlin, M. F. Fernández, M. Kay, J. Robie, and J. Siméon. XML Path Language (XPath) 2.0 (W3C Recommendation), January 2007. URL `http://www.w3.org/TR/xpath20/`. Last visited: December, 2011. 86

B. Berjani and T. Strufe. A recommendation system for spots in location-based online social networks. In *Proceedings of the 4th Workshop on Social Network Systems*, SNS '11, pages 4:1–4:6, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0728-4. 17

T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284:34–43, 2001. 49, 133

M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18:16–23, September 2003. ISSN 1541-1672. xi, 27

C. Bizer. The emerging web of linked data. *IEEE Intelligent Systems*, 24:87–92, 2009. ISSN 1541-1672. 133

C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst*, 5(3):1–22, 2009a. 134

C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165, 2009b. ISSN 1570-8268. The Web of Data. 55, 56, 134

V. Borkar, K. Deshmukh, and S. Sarawagi. Automatic segmentation of text into structured records. *SIGMOD Rec.*, 30:175–186, May 2001. ISSN 0163-5808. 37

K. K. Breitman, M. A. Casanova, and W. Truszkowski. Ontology sources. In *Semantic Web: Concepts, Technologies and Applications*, NASA Monographs in Systems and Software Engineering, pages 175–199. Springer London, 2007. ISBN 978-1-84628-710-7. x, 135, 136

E. Brill. Some advances in transformation-based part of speech tagging. In *National Conference on Artificial Intelligence*, pages 722–727, 1994. 21, 52, 102

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V. 30, 38

A. Z. Broder. Identifying and filtering near-duplicate documents. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, COM '00, pages 1–10, London, UK, 2000. Springer-Verlag. ISBN 3-540-67633-3. 96

B. Brumitt and S. Shafer. Topological world modeling using semantic spaces. In *In UbiComp 2001 Workshop on Location Modeling for Ubiquitous Computing*, 2001. 13

P. Buitelaar, P. Cimiano, and B. Magnini. *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, 2005. ISBN 978-1-58603-523-5. 72

U. C. Bureau. North American Industry Classification System (NAICS): Introduction, February 2010. URL `http://www.census.gov/eos/www/naics/`. Last visited: December, 2011. 81

CAE. Código de actividades económicas. instituto nacional de estatística., 2011. Last visited: December, 2011. 80

Calais. Opencalais whitepaper. Technical report, Thomson Reuters, 2008. 66, 67

M. E. Califf and R. J. Mooney. Bottom-up relational learning of Pattern Matching rules for Information Extraction. *J. Mach. Learn. Res.*, 4:177–210, December 2003. ISSN 1532-4435. ix, 58, 59

C. Cardie. Empirical methods in Information Extraction. *AI Magazine*, 18(4):65–80, 1997. 38, 39

E. Casey. *The Fate of Place: A Philosophical History (Centennial Books)*. University of California Press, Nov. 1998. ISBN 0520216490. 3

Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring Millions of Footprints in Location Sharing Services. In *In ICWSM '11*, 2011. 17

P. Christen. A comparison of personal name matching: Techniques and practical issues. In *Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*, pages 290–294, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2702-7. 26

F. Ciravegna. Adaptive Information Extraction from text by rule induction and generalization. In *Proceedings of the 17th international joint conference on Artificial intelligence - Volume 2*, pages 1251–1256, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-812-5, 978-1-558-60812-2. 66

J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, Apr. 1960. 48

P. R. Cohen. *Empirical methods for artificial intelligence*. MIT Press, Cambridge, MA, USA, 1995. ISBN 0-262-03225-2. 7

W. W. Cohen and P. Ravikumar. SecondString: An open-source java toolkit of approximate string-matching techniques. In *Neural Information Processing Systems*, 2003. 28

W. W. Cohen, A. Borgida, and H. Hirsh. Computing Least Common Subsumers in Description Logics. In *Proceedings of the 10th National Conference on Artificial Intelligence*, pages 754–760. MIT Press, 1992. 82

W. W. Cohen, P. Ravikumar, and S. E. Fienberg. *A comparison of string distance metrics for name-matching tasks*, pages 73–78. Number C. 2003. 26, 27

P. Coschurba, K. Rothermel, and F. Drr. A fine-grained addressing concept for geocast. In H. Schmeck, T. Ungerer, and L. Wolf, editors, *Trends in Network and Pervasive Computing - ARCS 2002*, volume 2299 of *Lecture Notes in Computer Science*, pages 1–9. Springer Berlin / Heidelberg, 2002. ISBN 978-3-540-43409-2. 10.1007/3-540-45997-9_9. ix, 13

M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the World Wide Web. *Artif. Intell.*, 118:69–113, April 2000. ISSN 0004-3702. doi: 10.1016/S0004-3702(00)00004-7. ix, 64, 65

Creative Commons. The GeoNames geographical database, 2010. URL `http://www.geonames.org/`. Last visited: December, 2011. 54

V. Crescenzi, G. Mecca, and P. Merialdo. RoadRunner: Towards Automatic Data Extraction from Large Web Sites. In *Proceedings of the 27th International Conference on Very Large Data Bases*, VLDB '01, pages 109–118, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-804-4. 36

J. Crowther, editor. *Oxford Advanced Learner's Dictionary*. Cornelsen & Oxford, 5th edition, 1998. 98

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002. 19, 62

I. Dagan and K. Church. Termight: identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing*, pages 34–40, Morristown, NJ, USA, 1994. Association for Computational Linguistics. 43

B. Daille, B. Habert, C. Jacquemin, and J. Royauté. Empirical observation of term variations and principles for their description. *Terminology: International Journal of Theoretical and Applied issues in Specialized Communication*, 1996. ISSN 0929-9971. 43

H. Davulcu, S. Vadrevu, S. Nagarajan, and I. V. Ramakrishnan. OntoMiner: Bootstrapping and Populating Ontologies from Domain-Specific Web Sites. *IEEE Intelligent Systems*, 18:24–33, September 2003. ISSN 1541-1672. 61, 62

G. de Melo, F. Suchanek, and A. Pease. Integrating YAGO into the Suggested Upper Merged Ontology. In *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008)*. IEEE Computer Society, Los Alamitos, CA, USA, 2008. 136

A. Dingli, F. Ciravegna, and Y. Wilks. Automatic Semantic Annotation using Unsupervised Information Extraction and Integration. In *Workshop on Knowledge Markup and Semantic Annotation*, 2003. 66

REFERENCES

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. *Proceedings of LREC 2004*, pages 837–840, 2004. 37, 47

F. Dotsika. Semantic APIs: Scaling up towards the semantic web. *International Journal of Information Management*, 30(4):335 – 342, 2010. ISSN 0268-4012. doi: DOI:10.1016/j.ijinfomgt.2009.12.003. ix, 67

M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *WWW '06*, pages 193–202, New York, USA, 2006. ACM. ISBN 1-59593-323-9. 17

Dun & Bradstreet. D & B website, February 2011. URL `http://www.dnb.com/`. Last visited: December, 2011. 78, 80

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996. 155

O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open Information Extraction from the Web. *Commun. ACM*, 51:68–74, December 2008. ISSN 0001-0782. 35, 37, 138

D. A. Evans and C. Zhai. Noun-phrase analysis in unrestricted text for Information Retrieval. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 17–24, Morristown, NJ, USA, 1996. Association for Computational Linguistics. 29

C. K. Falko Schmid. In-situ communication and labeling of places. In *6th International Symposium on LBS & TeleCartography*. Springer, 9 2009. 16

D. Faure and C. Ndellec. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *In LREC workshop on*, pages 5–12, 1998. 61

D. Ferrucci and A. Lally. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10:327–348, September 2004. ISSN 1351-3249. doi: 10.1017/S1351324904003523. 19

J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL '05*, pages 363–370, 2005. doi: http://dx.doi.org/10.3115/1219840.1219885. 24, 25, 102

J. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382, 1971. 48

W. N. Francis and H. Kucera. Frequency analysis of English usage: Lexicon and grammar. 1983. xi, 21, 22, 52

K. T. Frantzi. Incorporating context information for the extraction of terms. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL-35, pages 501–503, Morristown, NJ, USA, 1997. Association for Computational Linguistics. 45

K. T. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: the C-value/NC-value method. *Int. J. on Digital Libraries*, 3(2):115–130, 2000. 43

R. Genereux, L. Ward, and J. Russell. The behavioral component in the meaning of places. *Journal of Environmental Psychology*, 3:43–55, 1983. 15

GNS. Geonet Names Server. National Imagery and Mapping Agency, 2009. URL `http://earth-info.nga.mil/gns/html/index.html`. Last visited: December, 2011. 93

T. Gottron. Combining Content Extraction Heuristics: The CombinE System, 2008. 36

M. Greenwood. Java implementation of the Ramshaw and Marcaus BaseNP chunker, 2005. Last visited: December, 2011. 102

M. Grineva, M. Grinev, and D. Lizorkin. Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 661–670, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4. 39, 60

R. Grishman and B. Sundheim. Message Understanding Conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics -*

Volume 1, pages 466–471, Morristown, NJ, USA, 1996. Association for Computational Linguistics. 37

J. Gurland and R. Tripathi. A simple approximation for unbiased estimation of the standard deviation. In *American Statistician*, volume 25, pages 30–32, 1971. 222

S. Hacker. White Paper: Trainable Semantic Vectors & Semantic Signatures. Technical report, TextWise, 2008. 66, 69

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11, 2009. 90

S. Harrison and P. Dourish. Re-place-ing space: the roles of place and space in collaborative systems. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*, CSCW '96, pages 67–76, New York, NY, USA, 1996. ACM. ISBN 0-89791-765-0. 5, 15

I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. . Saghdha, L. Romano, and S. Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. 49

M. Hepp. Ontologies: State of the art, business potential, and grand challenges. In M. Hepp, P. Leenheer, A. Moor, and Y. Sure, editors, *Ontology Management*, volume 7 of *Semantic Web and Beyond*, pages 3–22. Springer US, 2008. ISBN 978-0-387-69900-4. 10.1007/978-0-387-69900-4_1. 134

J. Hightower. From position to place. In *Proc. of LOCA*, pages 10–12, 2003. Ubicomp. 5, 15

D. Hochbaum and D. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10(2):180–184, 1985. 172

G. Hodge. Systems for Knowledge Organization for Digital Libraries: Beyond traditional authority files. Technical report, Digital Library Federation, April 2000. URL `http://www.clir.org/pubs/reports/pub91/contents.html`. Last visited: December, 2011. 80

M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artificial intelligence*, AAAI'04, pages 755–760. AAAI Press, 2004. ISBN 0-262-51183-5. 35

P. Jackson and I. Moulinier. *Natural Language Processing for Online Applications*. Natural Language Processing 5. John Benjamins, Philadelphia, 2007. 19, 24

A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *MIR '06*, pages 89–98, 2006. ISBN 1-59593-495-2. 17

M. A. Jaro. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989. doi: 10.2307/2289924. 26

C. Jiang and P. Steenkiste. A hybrid location model with a computable location identifier for ubiquitous computing. In *Proceedings of the 4th international conference on Ubiquitous Computing*, UbiComp '02, pages 246–263, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-44267-7. 14

J. Justeson and S. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, pages 9–27, 1995. 43

K. Kageura and B. Umino. Methods of automatic term recognition: A review. *Terminology*, 3(2): 259–289, 1996. 42, 44

M. Kayed and K. F. Shaalan. A survey of web information extraction systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411–1428, 2006. ISSN 1041-4347. Member-Chia-Hui Chang and Member-Moheb Ramzy Girgis. 34

T. Kindberg., J. Barton, and J. et al. Morgan. People, places, things: Web presence for the real world. In *Proc. of WMCSA2000*, 2000. 14

J. E. King. *Software Solutions for Obtaining a Kappa-Type Statistic for Use with Multiple Raters*, volume Dallas, Te. 2004. 162

G. Klyne and J. J. Carroll, editors. *Resource Description Framework (RDF): Concepts and Abstract Syntax*. W3C Recommendation. World Wide Web Consortium, Feb. 2004. URL `http://www.w3.org/TR/rdf-concepts/`. Last visited: December, 2011. 134

REFERENCES

A. Koller, J. Moore, B. di Eugenio, J. Lester, L. Stoia, D. Byron, J. Oberlander, and K. Striegnitz. Shared task proposal: Instruction giving in virtual worlds. Working group report, Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, 2007. 48

I. Korkontzelos, I. Klapaftis, and S. Manandhar. Reviewing and evaluating automatic term recognition techniques. In *Proceedings of the 6th International Conference on Natural Language Processing, GoTAL 2008*, Gothenburg, Sweden, August 2008. 43

R. Kraft, F. Maghoul, and C. C. Chang. Y!Q: contextual search at the point of inspiration. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 816–823, New York, NY, USA, 2005. ACM. ISBN 1-59593-140-6. ix, 40, 41

B. Kramer. Classification of generic places: Explorations with implications for evaluation. *Journal of Environmental Psychology*, 15:3–22, 1995. 15

J. P. Kumar and P. Govindarajulu. Duplicate and near duplicate documents detection: A review. *European Journal of Scientific Research*, 32(4):514–527, 2009. 96

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. 24

R. Lemmens and D. Deng. Web 2.0 and semantic web: Clarifying the meaning of spatial features. In *11th International Conference on Geographic Information Science. Workshop: Semantic Web meets Geopatial Applications.*, AGILE'08, 2008. 16

V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–+, Feb. 1966. 26

J. Lin. Java version of Brill's Part-of-Speech Tagger, 2004. Last visited: December, 2011. 102

M. Litvak and M. Last. Graph-based keyword extraction for single-document summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, MMIES '08, pages 17–24, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. ISBN 978-1-905593-51-4. 39

B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer, 1st ed. 2007. corr. 2nd printing edition, January 2009. ISBN 3540378812. 28, 29, 30

H. Liu and P. Singh. Conceptnet: A practical commonsense reasoning toolkit, 2004. 49

Z. Liu, P. Li, Y. Zheng, and M. Sun. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 257–266, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. 39

A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16:72–79, March 2001. ISSN 1541-1672. 61

C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-13360-1. 21

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008. ISBN 0521865719. 29, 32, 33, 45, 47, 72, 94

E. Marsh and D. Perzanowski. MUC-7 Evaluation of IE Technology: Overview of Results, 1998. URL `http://www.aclweb.org/anthology-new/M/M98/M98-1002.pdf`. Last visited: December, 2011. 37, 47

D. Maynard and S. Ananiadou. Term sense disambiguation using a domain-specific thesaurus. In *In Proc. of 1st International Conference on Language Resources and Evaluation (LREC*, pages 681–687, 1998. 40

D. Maynard and S. Ananiadou. Identifying contextual information for multi-word term extraction. In *In (Sandrini*, pages 212–221, 1999. 40

O. Medelyan, D. Milne, C. Legg, and I. H. Witten. Mining meaning from Wikipedia. *Int. J. Hum.-Comput. Stud.*, 67:716–754, September 2009. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2009.05.004. 53, 54, 57, 58

R. Mihalcea. SemCor semantically tagged corpus, 1998. URL `http://lit.csci.unt.edu/~rada/downloads/semcor/semcor3.0.tar.gz`. Last visited: December 2011. 52, 100, 107

R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9. 59, 60

R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004. 38, 39

G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to WordNet: An on-line lexical database. *Int J Lexicography*, 3(4):235–244, 1990. doi: 10.1093/ijl/3.4.235. 49, 50, 91, 98, 106, 134, 139

D. Milne and I. H. Witten. Learning to link with Wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. 57, 58

G. Mishne. AutoTag: a collaborative approach to automated tag assignment for weblog posts. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 953–954, New York, NY, USA, 2006. ACM. ISBN 1-59593-323-9. 164

T. M. Mitchell. *Machine Learning.* McGraw-Hill, New York, 1997. 90

S. Mukund, D. Ghosh, and R. Srihari. Using sequence kernels to identify opinion entities in urdu. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 58–67, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. 49

S. Naaman, M. Nair, R. Yang, and J. Ahern. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. *International Conference on Digital Libraries, Vancouver, BC, Canada*, 2007. 17

NAICS. North American Industry Classification System, Mexico's Instituto Nacional de Estadística e Geografía Informática (INEG) and Statistics Canada and the United States Office of Management *and* Budget (OMB), 2011. URL `http://www.naics.com/index.html`. Last visited: December, 2011. 80

R. Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):1–69, 2009. ISSN 0360-0300. doi: http://doi.acm.org/10.1145/1459352.1459355. ix, 52, 56, 57, 107

R. Navigli and P. Velardi. Learning domain ontologies from document warehouses and dedicated web sites. *Comput. Linguist.*, 30:151–179, June 2004. ISSN 0891-2017. 61, 62

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004. doi: 10.1103/PhysRevE.69.026113. 60

I. Niles and A. Pease. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, New York, NY, USA, 2001. ACM. ISBN 1-58113-377-4. 135

A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *In ICWSM '11*, 2011. 17

J. Oliveirinha. Semantics in Place and Time. Master's thesis, Faculty of Sciences and Technology, University of Coimbra, Portugal, 2010. 125

OpenCyc. http://www.opencyc.org, 2011. Last visited: December, 2011. 49, 56

M. Pazienza, M. Pennacchiotti, and F. Zanzotto. Terminology Extraction: An analysis of linguistic and statistical approaches. In S. Sirmakessis, editor, *Knowledge Mining Series: Studies in Fuzziness and Soft Computing.* Springer Verlag, 2005. 43

F. C. Pereira, A. Alves, J. Oliveirinha, and A. Biderman. Perspectives on semantics of the place from online resources. In *Proceedings of the 2009 IEEE International Conference on Semantic Computing*, ICSC '09, pages 215–220, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3800-6. 151

G. Petasis, V. Karkaletsis, and G. Paliouras. Ontology population and enrichment: State of the art. Public deliverable d4.3, BOEMIE Project, 2007. 61

S. P. Ponzetto and M. Strube. Deriving a large-scale taxonomy from Wikipedia. In *AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence*, pages 1440–1445. AAAI Press, 2007. ISBN 978-1-57735-323-2. 54, 56

M. F. Porter. Readings in information retrieval. chapter An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. ISBN 1-55860-454-5. 30, 117

K. Potts. *Web Design and Marketing Solutions for Business Websites.* friends of ED, 2007. ISBN 1590598393. 121

A. F. R. Rahman, H. Alam, and R. Hartono. Content extraction from html documents. In *In 1st Int. Workshop on Web Document Analysis (WDA2001)*, 2001. 36

A. M. Rahmani, M. M. Pedram, and M. Asfia. Main Content Extraction from Detailed Web Pages. *International Journal of Computer Applications*, 4 (11):18–21, August 2010. Published By Foundation of Computer Science. 36

L. Ramshaw and M. Marcus. Text Chunking using Transformation-Based Learning. In *Proc. of WVLC-1995*, Cambridge, USA, 1995. 23, 102

T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *SIGIR '07*, pages 103–110, New York, USA, 2007. ACM. ISBN 978-1-59593-597-7. 16, 17, 18

E. Relph. *Place and placelessness*, volume 67. Pion, 1976. 16

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453, 1995. 100, 147, 148, 167

Reuters. OpenCalais API. `http://www.opencalais.com/`, 2008. Last visited: December, 2011. 67

F. Rodrigues. POI Mining and Generation. Master's thesis, Faculty of Sciences and Technology, University of Coimbra, Portugal, 2010. 77

F. Rodrigues, A. O. Alves, F. C. Pereira, S. Jiang, and J. Ferreira. Automatic classification of Points-of-Interest for land-use analysis. In *Proceedings of Geoprocessing'2012.*, 2012. ISBN 978-1-61208-178-6. 77, 142

J. Roth. Accessing location data in mobile environments - the Nimbus location model. In *Mobile HCI 03 Workshop on Mobile and Ubiquitous Information Access*, pages 256–270. Springer-Verlag, 2004. 15

S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice-Hall, Englewood Cliffs, NJ, 2nd edition edition, 2003. 72

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988. ISSN 0306-4573. doi: 10.1016/0306-4573(88)90021-0. 32, 45

G. Salton, A. Wong, and A. C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:229–237, 1975. 140

D. Santos, C. Freitas, H. Oliveira, and P. Carvalho. Second HAREM: New challenges and old wisdom. In A. Teixeira, V. de Lima, L. de Oliveira, and P. Quaresma, editors, *Computational Processing of the Portuguese Language*, volume 5190 of *Lecture Notes in Computer Science*, pages 212–215. Springer Berlin / Heidelberg, 2008. 47

S. Sarawagi. Information Extraction. *Found. trends databases*, 1:261–377, March 2008. ISSN 1931-7883. doi: 10.1561/1900000003. 34, 35, 36, 38, 72

T. Scheffer. Finding association rules that trade support optimally against confidence. In *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 424–435, 2001. 172

A. Schutz and P. Buitelaar. Relext: A tool for relation extraction from text in ontology extension. In *The Semantic Web - ISWC 2005, 4th International Semantic Web Conference, ISWC 2005, Proceedings*, volume 3729 of *Lecture Notes in Computer Science*, pages 593–606, Galway, Ireland, 2005. Springer. ISBN 3-540-29754-5. 143

SemEval Portal. In ACLwiki, 2011. URL `http://aclweb.org/aclwiki/index.php?title=SemEval_Portal`. Last visited: December, 2011. 47, 57, 164

Y. Shinyama and S. Sekine. Preemptive Information Extraction using unrestricted relation discovery. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 304–311, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. 35, 37

K. Siorpaes and D. Bachlechner. Harvesting Wiki Consensus - Using Wikipedia entries as ontology elements. pages 54–65, 2006. 54

S. Soderland. Learning information extraction rules for semi-structured and free text. *Mach. Learn.*, 34:233–272, February 1999. ISSN 0885-6125. doi: 10.1023/A:1007562322031. 37

S. C. Sood and K. J. Hammond. Tagassist: Automatic tag suggestion for blog posts. In *In International Conference on Weblogs and Social*, 2007. 164

J. F. Sowa. *Knowledge representation: logical, philosophical and computational foundations.* Brooks/Cole Publishing Co., Pacific Grove, CA, USA, 2000. ISBN 0-534-94965-7. 135

G. Stumme, A. Hotho, and B. Berendt. Semantic Web Mining: State of the art and future directions. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(2):124 – 143, 2006. ISSN 1570-8268. doi: DOI:10.1016/j.websem.2006.02.001. Semantic Grid –The Convergence of Technologies. 61, 63

F. Suchanek, G. Kasneci, and G. Weikum. YAGO - a large ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*, 6(3):203–217, September 2008. 54, 55, 134, 141

Z. Syed, T. Finin, and A. Joshi. Wikipedia as an Ontology for Describing Documents. In *Proceedings of the Second International Conference on Weblogs and Social Media.* AAAI Press, March 2008. 54

V. Tanasescu and J. Domingue. A differential notion of place for local search. In *LOCWEB '08*, pages 9–16, New York, USA, 2007. ACM. ISBN 978-1-60558-160-6. 31

The Apache Software Foundation. OpenNLP, 2010. URL `http://incubator.apache.org/opennlp/`. Last visited: December, 2011. 19

The Open Directory Project. The open directory project website. `http://dmoz.org`, 2002. Last visited: December, 2011. 69

E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In W. Daelemans and M. Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada, 2003. 37, 47

J. A. Tomlin. A new paradigm for ranking pages on the world wide web. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 350–355, New York, NY, USA, 2003. ACM. ISBN 1-58113-680-3. 30

P. D. Turney. Learning algorithms for keyphrase extraction. *Inf. Retr.*, 2:303–336, May 2000. ISSN 1386-4564. doi: 10.1023/A:1009976227802. 38

United Nations. International Standard Industrial Classification of all economic activities, 2011. Last visited: December, 2011. 80

United States Geological Survey. Geographic Names Information System (GNIS), 2007. URL `http://geonames.usgs.gov/`. Last visited: December, 2011. 117

D. Urbansky, M. Feldmann, J. Thom, and A. Schill. Entity Extraction from the Web with WebKnox. In V. Snášel, P. Szczepaniak, A. Abraham, and J. Kacprzyk, editors, *Advances in Intelligent Web Mastering - 2*, volume 67 of *Advances in Soft Computing*, pages 209–218. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-10687-3_20. 37

M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna. MnM: Ontology driven semi-automatic and automatic support for semantic markup. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, EKAW '02, pages 379–391, London, UK, 2002. Springer-Verlag. ISBN 3-540-44268-5. 66

J. Vazquez, J. Abaitua, and D. D. I. na. The ubiquitous web as a model to lead our environments to their full potential. In *W3C Workshop on the Ubiquitous Web. Position paper*, 2006. 14

X. Wan and J. Xiao. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*, pages 855–860. AAAI Press, 2008. ISBN 978-1-57735-368-3. 42, 165

R. C. Wang and W. W. Cohen. Language-independent set expansion of named entities using the web. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 342–350, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-3018-4. doi: 10.1109/ICDM.2007.104. 37

Wikipedia. http://en.wikipedia.org, 2004. Last visited: December, 2011. 53

W. E. Winkler. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research*, pages 354–359, 1990. 27

W. E. Winkler, W. E. Winkler, and N. P. Overview of record linkage and current research directions. Technical report, Bureau of the Census, 2006. 26

I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques, 2nd Edition.* Morgan Kaufmann, 2005. 90, 171

W. Wong, W. Liu, and M. Bennamoun. Determination of unithood and termhood for term recognition. In M. Song and Y. Wu, editors, *Handbook of Research on Text and Web Mining Technologies.* IGI Global, 2008. 42

Y.-f. B. Wu, Q. Li, R. S. Bot, and X. Chen. Domain-specific keyphrase extraction. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 283–284, New York, NY, USA, 2005. ACM. ISBN 1-59593-140-6. 39

Yahoo! Term extraction documentation for search web, 2009. URL `http://developer.yahoo.com/search/content/V1/termExtraction.html`. Last visited: December, 2011. 165

W. E. Yancey and W. E. Yancey. Evaluating string comparator performance for record linkage. Technical report, Bureau of the Census, 2005. 26

J. Ye, L. Coyle, S. Dobson, and P. Nixon. A unified semantics space model. In J. Hightower, B. Schiele, and T. Strang, editors, *LoCA*, volume 4718 of *Lecture Notes in Computer Science*, pages 103–120. Springer, 2007. ISBN 978-3-540-75159-5. 12

L. Yi, B. Liu, and X. Li. Eliminating noisy information in Web pages for Data Mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 296–305, New York, NY, USA, 2003. ACM. ISBN 1-58113-737-0. 36

C. T. Yu, K. Lam, and G. Salton. Term Weighting in Information Retrieval Using the Term Precision Model. *J. ACM*, 29:152–170, January 1982. ISSN 0004-5411. 40

D. Zelenko, C. Aone, and A. Richardella. Kernel methods for relation extraction. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 71–78, Morristown, NJ, USA, 2002. Association for Computational Linguistics. 35

Zemanta. Zemanta API. `http://www.zemanta.com/api/`, 2009. Last visited: December, 2011. 69

C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen. Discovering personally meaningful places: An interactive clustering approach. *ACM Trans. Inf. Syst.*, 25, July 2007. ISSN 1046-8188. 16

G. K. Zipf. *Selective Studies and the Principle of Relative Frequency in Language.* Harvard University Press, 1932. 117, 177

G. L. Zuniga. Ontology: Its transformation from philosophy to information systems. In *Proceedings of the International Conference on Formal Ontology in Information Systems*, pages 187–197. ACM Press, 2001. 49

Appendices

Appendix A

Semantic Enrichment of Places by Examples

A list of some semantic indexes is presented in the following subsections as the retrieval of information made previously by KUSCO for each perspective. Thus, for the same set of POIs, different views are presented using the Web or Wikipedia as source for the semantic enrichment process. Considering the states of Massachusetts and New York in the U.S., we chose the Yahoo! Local API as the POI source for the examples used here (see table A.1). In these areas, the taxonomy of POI categories used by Yahoo! Local is detailed in appendix D.

Following a generic to specific view of places, we start with a more controlled universe of search, the Wikipedia. In the first two sections the Red and Yellow Wiki perspectives are demonstrated for each POI in table A.1. Afterwards, extending the search for the semantic enrichment of places to the World Wide Web, the subsequent sections respectively present the About and Open Web perspectives for the same POIs.

The Red Wiki Perspective

In the Red Wiki perspective, Wikipedia is used to retrieve the articles about POI categories. These categories are inferred from the source where the POI is extracted, in our case the Yahoo!Local API, and are organized in a hierarchical taxonomy of categories (detailed in appendix D). Table A.2 presents the mapping of POI categories to Wikipedia articles made by KUSCO. Only the Wikipedia article abstracts are used by the meaning extraction module. In order to let the reader contextualize the content of a given article, the initial part of each abstract is also showed.

POI name	Location	Categories	POI source url
Au Bon Pain	1 State St Plz New York, NY	Carry Out & Take Out, Bakeries, Restaurants	http://local.yahoo.com /info-11039898-au-bon-pain-new-york
Boston Athenaeum	10 Beacon St, Boston, MA	Tourist Attractions, All Entertainers, Art Museums & Galleries, Libraries	http://local.yahoo.com /info-10150557-boston-athenaeum-boston
Duane Reade	460 8th Ave, New York, NY	Drug Stores, Photography Labs, First Aid, Pharmacies	http://local.yahoo.com /info-11017289-duane-reade-new-york
Erbaluce	69 Church St Boston, MA	Italian Restaurants, Restaurants	http://local.yahoo.com /info-46631862-erbaluce-boston
Harvard Medical School	250 Longwood Ave, Boston, MA	General Practice Medicine, Colleges & Universities, Doctors & Clinics, Neurology	http://local.yahoo.com /info-10168845-harvard-medical-school-boston
Mail Boxes Etc	258 Harvard St, Brookline, MA	B2B Courier Services, Cargo Services, Photocopying, Direct Mail, Mail Services, Packaging, Fax Services	http://local.yahoo.com /info-10233127-mail-boxes-etc-brookline
Petco	1210 Providence Hwy, Norwood, MA	Pet Supplies, All Animal Services	http://local.yahoo.com /info-10144601-petco-norwood
Radio Shack	925 Lexington Ave, New York, NY	Computer Software, Electronics Retailers, Cellular Phones	http://local.yahoo.com /info-11112320-radio-shack-new-york
Sbarro	350 Longwood Ave, Boston, MA	Carry Out & Take Out, Pizza, Restaurants	http://local.yahoo.com /info-10168837-sbarro-boston

Table A.1: Some example of POIs from the greater metropolitan area of Boston and New York considering the Yahoo! Local API.

As explained in section 4.2.3.1, when there is no direct mapping in Wikipedia for a given category name or when this name is ambiguous it is used the immediately upper category mapping in the POI taxonomy. In table A.2 these categories are highlighted in **bold** (e.g. "B2B Courier Services" is mapped to Wikipedia though its subsumer category “Shipping”). When a category is a compound name and its not exactly represented by a Wikipedia article, this compound name is broken in shorter names and retrieved independently in Wikipedia (e.g. Colleges & Universities). For those category names highlighted in *italics* in table A.2, there were not a unambiguous correspondence in Wikipedia (e.g. the article titled “Doctors” is redirected to a disambiguation page).

Table A.2: The *Wikipedia* article abstracts retrieved by KUSCO for the categories of those POIs in study.

POI category	**Wikipedia article**	**An excerpt of the Wikipedia article abstract**
All Animal Services	Animals	Animals are a major group of multicellular, eukaryotic organisms of the kingdom Animalia or Metazoa. ...
All Entertainers	Entertainers	Entertainment consists of any activity which provides a diversion or permits people to amuse themselves in their leisure time. ...
Art Museums & *Galleries*	Museums	A museum is an institution that cares for a collection of artifacts and other objects of scientific, artistic, cultural, or historical importance and makes them available for public viewing through exhibits that may be permanent or temporary. ...
B2B Courier Services	Shipping	Shipping has multiple meanings. It can be a physical process of transporting commodities and merchandise goods and cargo, by land, air, and sea. It also can describe the movement of objects by ship. ...
Bakeries	Bakeries	A bakery (or baker’s shop) is an establishment which produces and sells flour-based food baked in an oven such as bread, cakes, pastries and pies. ...
Cargo Services	Cargo	Cargo (or freight) is goods or produce transported, generally for commercial gain, by ship, aircraft, train, van or truck. ...
Carry Out & Take Out	Carry-out	Take-out or takeout, carry-out, take-away ... is food purchased at a restaurant for the purpose of being eaten elsewhere. ...

Continued on next page. . .

Table A.2 – Continued

POI category	**Wikipedia article**	**An excerpt of the Wikipedia article abstract**
Cellular Phones	Cellular_phones	A mobile phone (also known as a cellular phone, cell phone and a hand phone) is a device that can make and receive telephone calls over a radio link whilst moving around a wide geographic area. ...
Colleges & Universities	Colleges	A college (Latin: collegium) is an educational institution or a constituent part of an educational institution. ...
	Universities	A university is an institution of higher education and research, which grants academic degrees in a variety of subjects. ...
Computer Software	Computer_software	Computer software, or just software, is a collection of computer programs and related data that provides the instructions for telling a computer what to do and how to do it. ...
Direct Mail	Advertising	Advertising is a form of communication used to encourage or persuade an audience (viewers, readers or listeners) to continue or take some new action. ...
Doctors & Clinics	Clinics	A clinic (or outpatient clinic or ambulatory care clinic) is a health care facility that is primarily devoted to the care of outpatients. ...
Drug Stores	Drug_stores	Pharmacy is the health profession that links the health sciences with the chemical sciences and it is charged with ensuring the safe and effective use of pharmaceutical drugs. ...
Electronics Retailers	Home_electronics	Consumer electronics are electronic equipment intended for everyday use, most often in entertainment, communications and office productivity. ...
Fax Services	Fax	Fax (short for facsimile), sometimes called telecopying, is the telephonic transmission of scanned printed material (both text and images), normally to a telephone number connected to a printer or other output device. ...

Continued on next page...

Table A.2 – Continued

POI category	**Wikipedia article**	**An excerpt of the Wikipedia article abstract**
First Aid	First_aid	First aid is the provision of initial care for an illness or injury. ...
General Practice Medicine	Clinics	A clinic (or outpatient clinic or ambulatory care clinic) is a health care facility that is primarily devoted to the care of outpatients. ...
Italian Restaurants	Italian_restaurant	Italian cuisine ... has developed through centuries of social and political changes, with roots as far back as the 4th century BCE. ...
Libraries	Libraries	In a traditional sense, a library is a large collection of books, and can refer to the place in which the collection is housed. ...
Mail Services	Mail	Mail, or post, is a system for transporting letters and other tangible objects: written documents, typically enclosed in envelopes, and also small packages are delivered to destinations around the world. ...
Neurology	Neurology	Neurology ... is a medical specialty dealing with disorders of the nervous system. ...
Packaging	Shipping	Shipping has multiple meanings. It can be a physical process of transporting commodities and merchandise goods and cargo, by land, air, and sea. It also can describe the movement of objects by ship. ...
Pet Supplies	Animals	Animals are a major group of multicellular, eukaryotic organisms of the kingdom Animalia or Metazoa. ...
	Pets	A pet is a household animal kept for companionship and a person's enjoyment, as opposed to wild animals or to livestock, laboratory animals, working animals or sport animals, which are kept for economic or productive reasons. ...
Pharmacies	Pharmacies	Pharmacy is the health profession that links the health sciences with the chemical sciences and it is charged with ensuring the safe and effective use of pharmaceutical drugs. ...

Continued on next page...

Table A.2 – Continued

POI category	**Wikipedia article**	**An excerpt of the Wikipedia article abstract**
Photocopying	Professional_services	Professional services is an industry of infrequent, technical, or unique functions performed by independent contractors or by consultants whose occupation is the rendering of such services. ...
Photography Labs	Photography	Photography is the art, science and practice of creating durable images by recording light or other electromagnetic radiation, either electronically by means of an image sensor or chemically by means of a light-sensitive material such as photographic film. ...
Pizza	Pizza	Pizza ... is an oven-baked, flat, round bread typically topped with a tomato sauce, cheese and various toppings. ...
Restaurants	Restaurants	A restaurant ... is an establishment which prepares and serves food and drink to customers in return for money, either paid before the meal, after the meal, or with a running tab. ...
Tourist Attractions	Tourist_attractions	A tourist attraction is a place of interest where tourists visit, typically for its inherent or exhibited cultural value, historical significance, natural or built beauty, or amusement opportunities. ...

Following the pipeline of the KUSCO system, the Meaning Extraction module select terms in abstracts extracted from the Wikipedia articles in table A.2. Each term is contextualized, when it is possible, in Wordnet or Wikipedia and becoming a *concept*. Thus, it is possible to integrate concepts which are semantically related (i.e., synonyms according to Wordnet or Wikipedia). For those concepts not contextualized in any knowledge resource, they are compared in lexical terms as presented in algorithm 5.3. After all concepts have been extracted, the semantic TF-IDF is computed for each them having in mind that a document in the *Inverse Document Frequency* component of this metric is a POI. This weighted vector of concepts we call the *Semantic Index*

POI name	Semantic Index
Au Bon Pain	Pizza delivery, food, Döner kebab, fast food, table service, franchise, Examples, takeout, Take-out food, Asian countries
Boston Athenaeum	opportunities, dance, Frankish, Museum, collection, Le Louvre, Madrid, poetry readings, music concert, ballet
Duane Reade	medication, pharmaceutical company, Greek, digital camera, chemical sciences, patient care, health sciences, photographic lens, health professional, ingredients
Erbaluce	restaurateur, cuisines, delivery service, Meals, chefs, service model, artisans, line cooks, Seinfeld, drink
Harvard Medical School	health services, hospital, primary health care, group, physiotherapists, medical school, outpatient clinic, full dress, institutions, medical specialty
Mail Boxes Etc	Shipping, ship, Postal system, truck, document, technologies, contracts, destination, international trade, advantages, formats
Petco	animals, pets, dog, group, owners, humans, Nematomorpha, fossil record, social interaction, organisms
Radioshack	place, home, data, Mobile, physical device, mobile phone, subscriptions, barrier, hardware, adornment
Sbarro	food, fast food, delis, takeout, kebabs, cue, Windows, Examples, franchise, Take-out food

Table A.3: The semantic indexes in the Red Wiki perspective for the POIs in study.

of a POI in a given perspective. Table A.3 presents the Semantic Index built for each POI in study considering its top-10 concepts ordered by semantic TF-IDF relevance.

It is important to remember that, in this perspective, all terms that are already known as category names of a given POI are filtered out in its semantic index in order to avoid duplicate information (as explained in section 6.3.1). For instance, in the case of the POIs *Au Bon Pain* and *Boston Athenaeum*, the terms *restaurant* and *Library* were removed from their respective semantic indexes. Furthermore, in this last POI the term *Museums* was maintained by KUSCO as it is a generalization of *Art Museum* and not meaning the same of one of the categories already known from the POI *Boston Athenaeum.*

The Yellow Perspective

In the Yellow Wiki perspective (section 4.2.3.2), Wikipedia is used to retrieve the specific article about a given POI. Table A.4 presents the mapping of POIs to Wikipedia articles made by KUSCO. In order to let the reader contextualize the content of a given article, the initial part of each abstract is also showed.

In the same line as the Wiki perspective, following the pipeline of the KUSCO system, the Meaning Extraction module select terms in articles extracted from the Wikipedia articles in table A.4. Table A.5 presents the Semantic Index built for each POI in study considering its top-10 concepts ordered by semantic TF-IDF relevance.

The About Perspective

In the About perspective, the POI source url is used to retrieve the *About* page of a given POI using some heuristics described in section 4.2.2. Table A.6 presents the respective About page found by KUSCO for some POIs in study. For some POIs, it was not possible to find a candidate for the *About* page due or to the completely nonexistence of such page in the POI website (e.g. Duane Reade's website) or due to the organization of the About section with other submenus making it hard to find the best one to elect as the About page (e.g. Erbaluce's website).

In the same line as the other perspectives, following the pipeline of the KUSCO system, the Meaning Extraction module select terms in texts extracted from the Web pages in table A.6. Table A.7 presents the Semantic Index built for each POI in study considering its top-10 concepts ordered by semantic TF-IDF relevance.

The Open Web Perspective

Table A.8 presents the Web pages retrieved by Information Retrieval module using Location-based Web Search (section 4.2.1) for each POI in table A.1 with the relevant summary presented in Yahoo! Search results. Due to the limitations of page size, only the domain for each page is provided. (For interested reader, a simple search query will yield the exact page). It is also important to note that, as the Web grows and changes every day, this result set will be different in future searches.

POI name	**Wikipedia article**	**First sentence in the Wikipedia article**
Au Bon Pain	Au_Bon_Pain	Au Bon Pain ... is a fast-casual bakery and cafe chain headquartered in Boston, Massachusetts.
Boston Athenaeum	Boston_Athenaeum	Boston Athenum is one of the oldest independent libraries in the United States.
Duane Reade	Duane_Reade	Duane Reade Inc., a subsidiary of the Walgreen Company, is a chain of pharmacy and convenience stores, primarily located in New York City, known for its high volume small store layouts in densely populated Manhattan locations.
Erbaluce	Erbaluce	Erbaluce or Erbaluce Bianca is a white Italian wine grape grown primarily in the Piedmont region around Caluso. In addition to dry table wines, it is used to make sweet wines passito with deep golden coloring.
Harvard Medical School	Harvard_Medical_School	Harvard Medical School (HMS) is the graduate medical school of Harvard University.
Mail Boxes Etc	Mail_Boxes_Etc.	Mail Boxes Etc. (MBE) is a global retail chain of business service centers.
Petco	Petco	PETCO is a chain of retail stores that offers pet supplies and services such as grooming and dog training.
Radioshack	Radio_Shack	RadioShack Corporation (formerly Tandy Corporation) (NYSE: RSH) is an American franchise of electronics retail stores in the United States, as well as parts of Europe, South America and Africa.
Sbarro	Sbarro	Sbarro is a chain of pizza restaurants that specializes in traditional Italian cuisine,[2] including its most popular menu item ‘pizza by the slice.”

Table A.4: The *Wikipedia* article abstracts retrieved by KUSCO for the POIs in study.

POI name	Semantic Index
Asbury Automotive Group Incorporated	asburyauto, franchises, revenues, snapshots, retailers, auto dealership, brands, parts
Au Bon Pain	Good, restaurant, locations, cafes, Thomas John, Compass Group, Siam Square, panerabread, continuation, airports
Boston Athenaeum	Library, collections, landmark building, galleries, Edward Clarke Cabot, Phineas Adams, winter, Anthology Club, Researchers, classification system
Duane Reade	layouts, convenience stores, locations
Erbaluce	Clarke Encyclopedia, table wines, Caluso, Erbaluce, wine grape, sweet wines, Harcourt Books, grape, hills, region
Harvard Medical School	departments, Chief Academic Officer, students, program, Neighborhood, problem-based learning, full professor, Harvard-MIT Division of Health Sciences and Technology, Massachusetts Mission Hill, Massachusetts Institute of Technology
Mail Boxes Etc	Locations, mbe, Fineffe Group, Milan, brands, couriers, North American, franchisees, Store brand, Italy
Petco	pet foods, Pets, Cesar Millan, San Diego, Halo Brand, food, section, cat food, dog, natural products
Radioshack	Sponsor, brands, Electronics, telephones, Contacts, exterior, Equipment, video cables, Accurian, digital picture frame
Sbarro	Pizza, Anthony Missano, John Brisco, Peter Beaudrault, Melville, pasta, Brooklyn, slice, dishes, restaurants

Table A.5: The semantic indexes in the Yellow Wiki perspective for the POIs in study.

POI name	POI Web site second Yahoo!Local	POI *About* page
Au Bon Pain	http://aubonpain.com/	http://www.aubonpain.com/aboutus/
Boston Athenaeum	http://bostonathenaeum.org/	http://bostonathenaeum.org/
Duane Reade	http://duanereade.com	*not found*
Erbaluce	http://www.erbaluce-boston.com	*not found*
Petco	http://petco.com/	http://about.petco.com/
Radioshack	http://radioshack.com/	http://radioshack.com/

Table A.6: The *About* pages retrieved by KUSCO for some of the POIs in study given the POI Web site by the Yahoo!Local API.

POI name	Semantic Index
Au Bon Pain	Soups, breads, Pain, plate, kiosks, Au Bon Pain, Louis Kane, Thailand, Taiwan, Youll
Boston Athenaeum	Athen, associate, Meet-ups, Membership, gathering, fund, Kicker, talk, announcement query
Petco	sheet, Release, discus, sponsorship, Scoop, Were, companion, sustenance, priority, retailer, seller
Radioshack	Tandy, Realist, Album, Catalog, computer store, format, Milton Deutschmann, electronics industry, brother, Equipment

Table A.7: The semantic indexes in the About perspective for some of the POIs in study.

Table A.8: The most relevant pages obtained by Yahoo for each POI in table A.1

<table>
<tr><th>POI name</th><th>Web page title</th><th>Web page url</th></tr>
<tr><td><b>Au Bon Pain</b></td><td>Au Bon Pain - Welcome to Au Bon Pain</td><td>http://www.aubonpain.com</td></tr>
<tr><td></td><td>Au Bon Pain - Wikipedia, the free encyclopedia</td><td>http://en.wikipedia.org/wiki/Au_Bon_Pain</td></tr>
<tr><td></td><td>Au Bon Pain - New York, NY, 10005 - Citysearch</td><td>http://newyork.citysearch.com/profile/37744817</td></tr>
<tr><td></td><td>Au Bon Pain - Financial District - New York, NY</td><td>http://www.yelp.com/biz/au-bon-pain-new-york-11</td></tr>
<tr><td></td><td>Au Bon Pain - Restaurant - New York - HopStop.com - Transit ...</td><td>http://www.hopstop.com/Au_Bon_Pain-Restaurants-New_York</td></tr>
<tr><td></td><td>Au Bon Pain Restaurant New York NYC NY Reviews - Gayot</td><td>http://www.gayot.com/restaurants/au-bon-pain-new-york-ny-10003_1ny99894-27.html</td></tr>
<tr><td colspan="3">At Au Bon Pain, we take our service - and menu - beyond the expected. From our authentic artisan breads and scrumptious pastries to inspired menus filled with savory ... Au Bon Pain has three locations in New York City's Port Authority Bus Terminal. The chain is also very successful on college campuses: ... (212) 952-9007 - , New York, NY 10005 (212) 952-9007 - In Short - The marketplace-like atmosphere of this bustling eatery showcases sandwiches and salads in cold cases and various breads on ...
(212) 962-8453 - 222 Broadway (between Ann St & Fulton St) "I guess this NY's version of Panera Bread. They sell sandwiches, salads, baked goods, coffee, etc. I ... Read a professional restaurant review of Au Bon Pain at GAYOT.com, where we have reviews for many Sandwiches restaurants in New York.</td></tr>
<tr><td><b>Boston Athenaeum</b></td><td>Boston Athenaeum</td><td>http://bostonathenaeum.org</td></tr>
<tr><td></td><td>Boston Athenaeum - Downtown - Boston, MA</td><td>http://www.yelp.com/biz/boston-athenaeum-boston</td></tr>
<tr><td></td><td>Cleanup underway at Boston Athenaeum after water leak ...</td><td>http://www.boston.com/yourtown/news/beacon_hill/2011/01/cleanup_underway_at_boston_ath.html</td></tr>
<tr><td></td><td>Boston Athenaeum - Partner - Forum Network - Free Online ...</td><td>http://forum-network.org/partner/boston-athenaeum</td></tr>
<tr><td></td><td>Boston Athenaeum - New England Travel</td><td>http://www.newenglandtravelplanner.com/go/ma/boston/sights/athenaeum.html</td></tr>
<tr><td></td><td>Boston Athenaeum - Boston Behind the Scenes Podcast</td><td>http://www.bostonbehindthescenes.com/boston-athenaeum</td></tr>
</table>

Continued on next page. . .

Table A.8 – Continued

<table>
<tr><th>POI name</th><th>Web page title</th><th>Web page url</th></tr>
<tr><td></td><td>BOSTON ATHENAEUM — BOSTON MA — ORGANIZATION DIRECTORY ...</td><td>http://www.artsboston.org/org/detail/7073</td></tr>
<tr><td colspan="3">(617) 227-0270 - 10 1/2 Beacon St. “If I lived closer to Boston I’d consider joining.” ... “There were magnificent statues and books along the walls.” The Boston Athenaeum, the landmark membership library on Beacon Hill that is more than 200 years old, has sent thousands of books to a specialist for freeze-drying ... The Boston Athenaeum, one of the oldest and most distinguished independent libraries in the United States, was founded in 1807 by members of the Anthology Society, a ... All about the Boston Athenaeum, a historic private research library, museum and art gallery on Beacon Hill off Boston Common near the State House in Boston ... This time well visit the largest membership library in the country, the Boston Athenaeum. The Athenaeum is in the middle of celebrating its 200th year with an ... Boston Athenaeum Comment on Facebook. One of Boston’s cultural treasures since 1807, the Boston Athenum maintains the largest membership of any independent ...</td></tr>
<tr><td><b>Duane Reade</b></td><td>DUANEreade Your City. Your Drugstore.</td><td>http://duanereade.com</td></tr>
<tr><td></td><td>Duane Reade - Wikipedia, the free encyclopedia</td><td>http://en.wikipedia.org/wiki/Duane_Reade</td></tr>
<tr><td></td><td>DUANEreade Your City. Your Drugstore. - HopStop.com - Transit ...</td><td>http://www.hopstop.com/dr</td></tr>
<tr><td></td><td>Duane Reade - Hell’s Kitchen - New York, NY</td><td>http://www.yelp.com/biz/duane-reade-new-york-34</td></tr>
<tr><td></td><td>Duane Reade - New York, NY, 10128 - Citysearch</td><td>http://newyork.citysearch.com/profile/42278726/new_york_ny/duane_reade.html</td></tr>
<tr><td></td><td>Duane Reade - New York Store & Shopping Guide</td><td>http://nymag.com/listings/stores/duane-reade</td></tr>
<tr><td></td><td>Home - DuaneReade Walkin</td><td>http://www.drwalkin.com</td></tr>
<tr><td></td><td>Duane Reade - New York</td><td>http://www.insiderpages.com/doctors/duane-reade-new-york-19</td></tr>
<tr><td colspan="3">Duane Reade is New York’s pharmacy, with over 250 convenient locations to fill your prescription, photo, and day-to-day health, wellness, and beauty needs ... Duane Reade Inc., a subsidiary of the Walgreen Company, is a chain of pharmacy and convenience stores, primarily located in New York City, known for its high volume ... New York Living; New Stores; Special Offers; Drugs; Diseases/Conditions; Community Support; Natural Health Products; Lab Tests and Procedures; Health Tools; Find a Duane Reade near:</td></tr>
</table>

Continued on next page...

Table A.8 – Continued

<table>
<tr><th>POI name</th><th>Web page title</th><th>Web page url</th></tr>
<tr><td colspan="3">(212) 246-0168 - 721 9th Ave (between 49th St & 50th St) - “The people working here on my most recent visit were actually friendly and having a good time. In fact ... (646) 672-1760 New York’s ubiquitous drugstore features a pharmacy and offers cosmetics, nutritional and hygienic products, vitamins, greeting cards, snacks and more. This is the closest thing New Yorkers have to a five-and-dime and pharmacy all in one. See the profile of this NYC store. ... in a New York minute! Access to urgent affordable health care in New York is now as easy as walking into a Duane Reade store. Our professional medical staff will ... Get reviews for the things that matter most in New York, NY including Health Food Stores on Insider Pages...</td></tr>
<tr><td><b>Erbaluce</b></td><td>Erbaluce</td><td>http://erbaluce-boston.com</td></tr>
<tr><td></td><td>Erbaluce - Boston, MA</td><td>http://www.yelp.com/biz/erbaluce-boston</td></tr>
<tr><td></td><td>Erbaluce in Boston - Find Restaurant Info - Boston.com</td><td>http://calendar.boston.com/boston-ma/venues/show/1047425</td></tr>
<tr><td></td><td>Erbaluce, Boston - Restaurant Reviews - TripAdvisor</td><td>http://www.tripadvisor.com/Restaurant_Review-g60745-d1233630-Reviews-Erbaluce-Boston_Massachusetts.html</td></tr>
<tr><td></td><td>Dining Out: Erbaluce - Boston Magazine</td><td>http://www.bostonmagazine.com/dining_food_wine/articles/dining_out_erbaluce</td></tr>
<tr><td></td><td>Erbaluce in Boston, MA 69 Church St, Boston, MA</td><td>http://www.superpages.com/bp/Boston-MA/erbaluce-L2096101643.htm</td></tr>
<tr><td colspan="3">(617) 426-6969 - 69 Church St (between Piedmont St & Shawmut St) - “And did I mention the panna cotta .” ... “He had the wild boar and really enjoyed it.” ... Come to Boston.com to get information, reviews, photos, and directions on Erbaluce and other restaurants in Boston, MA. ... “My husband and I ate here for the first time following a show at the Wang. It was a ...” “Really difficult location for taxis to find. We drove around ... Dining Out: Erbaluce Long-lost culinary vet Charles Draghi returns to the scene with a poetic homage to high-art northern Italian. But it’s not for the timid of palate. ... Erbaluce in Boston, MA - Map, Phone Number, Reviews, Photos and Video Profile for Boston Erbaluce. Erbaluce appears in: Restaurants ...</td></tr>
<tr><td><b>Harvard Medical School</b></td><td>Harvard Medical School</td><td>http://hms.harvard.edu/hms/home.asp</td></tr>
<tr><td></td><td>Harvard Medical School - Wikipedia, the free encyclopedia</td><td>http://en.wikipedia.org/wiki/Harvard_Medical_School</td></tr>
<tr><td></td><td>Harvard Medical School - Boston, MA, 02115 - Citysearch</td><td>http://boston.citysearch.com/profile/4728467/boston_ma/harvard_medical_school.html</td></tr>
<tr><td></td><td>Department of Cell Biology - Harvard Medical School</td><td>http://cellbio.med.harvard.edu</td></tr>
</table>

Continued on next page...

Table A.8 – Continued

<table>
<tr><th>POI name</th><th>Web page title</th><th>Web page url</th></tr>
<tr><td></td><td>Systems Biology Harvard Medical School - Harvard University</td><td>http://sysbio.med.harvard.edu</td></tr>
<tr><td></td><td>Harvard Medical School, Boston - Who's what, where?</td><td>http://www.boston.com/business/whoswhat/2011/04/harvard_medical_3.html</td></tr>
<tr><td></td><td>Department Of Psychiatry - Harvard Medical School</td><td>http://medapps.med.harvard.edu/psych</td></tr>
<tr><td></td><td>Harvard Medical School — Admission and Application Information</td><td>http://www.medicalschooladmission.com/harvard</td></tr>
<tr><td></td><td>Harvard Medical School Food Delivery ...</td><td>http://www.grubhub.com/boston/harvard-medical-school</td></tr>
<tr><td></td><td>Boston School Of Electrolysis At Harvard Medical School</td><td>http://www.bostonschoolofelectrolysis.com/harvard.php</td></tr>
<tr><td colspan="3">Children's Hospital Boston; Dana-Farber Cancer Institute; Forsyth Institute; Harvard Pilgrim Health Care; Hebrew SeniorLife; Joslin Diabetes Center; Judge Baker Children's Center; ... Harvard Medical School (HMS) is the graduate medical school of Harvard University. ... The architect for the campus was the Boston firm of Shepley, Rutan and Coolidge. ... (617) 495-1000 - 250 Longwood Ave Ste 304, Boston, MA 02115, Category: Medical Schools , Internal Medicine Doctors , Radiologists ... The Department of Psychiatry of Harvard Medical School coordinates the psychiatric resources of nine major teaching institutions in the Greater Boston area into a ... admission and other information for harvard medical school. Home: Admission Counselors: Book Reviews: Discussion Board: Feature Content: ... Boston, MA 02115. ... GrubHub.com: Find restaurants that deliver to Harvard Medical School in Boston. Browse delivery menus, reviews, and coupons. Order online or by phone. ... Boston School Of Electrolysis, for permanent hair removal, is an electrolysis clinic operated by Kimberly Williams, a licensed and registered electrologist, in...</td></tr>
<tr><td><b>Mail Boxes Etc</b></td><td>Mail Boxes Etc. - BROOKLINE, MA - Home</td><td>http://www.mailboxesetclocal.com/0759</td></tr>
<tr><td></td><td>Mail Boxes Etc. - Coolidge Corner - Brookline, MA</td><td>http://www.yelp.com/biz/mail-boxes-etc-brookline</td></tr>
<tr><td></td><td>Mail Boxes Etc - Brookline, MA, 02446 - Citysearch</td><td>http://boston.citysearch.com/profile/4790529/brookline_ma/mail_boxes_etc.html</td></tr>
<tr><td></td><td>Mail Boxes Etc, Brookline, MA - Company Profiles & Company ...</td><td>http://www.manta.com/c/mtx00lv/mail-boxes-etc</td></tr>
<tr><td colspan="3">Welcome to the Mail Boxes Etc in Brookline, MA. Also proudly serving customers for the</td></tr>
</table>

Continued on next page...

Table A.8 – Continued

POI name	Web page title	Web page url
Boston area. Mailboxes, Printing, Faxing, Business Cards, Shredding, Moving ... 258 Harvard Street - "I went here to buy a T pass and the salesperson was incredibly friendly and helpful. It was a drawback that they only took cash, lucky my bank ... (617) 903-7774 - 258 Harvard St, Brookline, MA 02446 - Last updated 3.15.11 Category: Direct Mail Advertising , Advertising , Business Services ... Mail Boxes Etc in Brookline, MA is a private company categorized under Mailing and Shipping Services. Current estimates show this company has an annual revenue of $1 ...		
Petco	Petco Norwood Store Location - Pet Supplies - Pet Products ...	http://www.petco.com/Content/Locator/Details. aspx?storeid=745
	Norwood PETCO Events, Information - Norwood, MA - Boston.com	http://calendar.boston.com/norwood-ma/venues/show/38997-norwood-petco
	Petco in Norwood, MA by Yellowbook	http://www.yellowbook.com/profile/petco_1536960510.html
	Petco in Norwood, MA - YellowBot	http://www.yellowbot.com/petco-norwood-ma.html
	Petco - Learn More About Our Company, Vision & Job Opportunities	http://careers.petco.com/info.asp
	Adoption Day Norwood - Adoption Day at Petco - NORWOOD ...	http://eventful.com/norwood/events/adoption-day-/E0-001-032813308-3
Pet Supplies - Pet Products - Pet Food — Petco.com. Petco's commitment to Natural, Holistic, and Organic pet food is unparalleled in the pet industry. Petco offers a ... Norwood PETCO, Norwood, MA: Get Reviews, Ratings, Photos, Directions and more with Boston.com. Petco at 1210 Providence Hwy Ste 1, Norwood, MA 02062 ... Status: Open. Sunday 10:00am-6:00pm Monday 9:00am-9:00pm Tuesday 9:00am-9:00pm Wednesday Business Listing Information for Petco in Norwood, MA by Yellowbook. ... Interested in pursuing a career at Petco? We invite you to learn more about our company, vision, and job opportunities. ... Adoption Day in Boston, MA at Petco - NORWOOD. Stop by and visit our table. Donna will be showing photos and taking applications. Also, we will be...		
Radio Shack	RadioShack - New York, NY, 10036 - Citysearch	http://newyork.citysearch.com/profile/7183431/new_york_ny/radioshack.html
	RadioShack - mobile phones, MP3 players, laptops, and more	http://www.radioshack.com
	Radio Shack in New York, NY - 50 E 42nd St, New York, NY	http://www.superpages.com/bp/New-York-NY/Radio-Shack-L2050758302.htm

Continued on next page...

Table A.8 – Continued

POI name	Web page title	Web page url
	Radioshack in New York, NY - New York Radioshack - YP.com	http://www.yellowpages.com/union-square-new-york-ny/radioshack
	Radioshack - Yorkville - New York, NY	http://www.yelp.com/biz/radioshack-new-york-16
	RadioShack - Wikipedia, the free encyclopedia	http://en.wikipedia.org/wiki/RadioShack
	Radio Shack locations in New York, NY - Mojopages	http://www.mojopages.com/brands/radio-shack/new-york/ny
(212) 944-2540 - In Short, while the name dates back to the invention of the radio at the turn of the 20th century, today the electronic store offers a ... Thank you for visiting Radio Shack. If you need assistance with shopping on our site, please call us at 800-843-7422 and a customer care representative will be happy ... Radio Shack in New York, NY- Map, Phone Number, Reviews, Photos and Video Profile for New York Radio Shack. Radio Shack appears in: Electronic Equipment & Supplies ... (212) 426-2160 - 1668 1st Ave - “I stopped in this Radio Shack today since I was in the neighborhood. I had been planning on buying a netbook for a while and I knew ... The company was started as Radio Shack in 1921 by two brothers, Theodore and Milton Deutschmann, ... “Fix-It Service Remodels Radio Shack”. The New York Times. ... Find Radio Shack locations in the New York, NY area with the Mojopages location finder. Maps, reviews and phone numbers for New York Radio Shack locations. ...		
Sbarro	Sbarro in Boston - Find Restaurant Info - Boston.com	http://calendar.boston.com/boston-ma/venues/show/762771-sbarro
	Sbarro - Boston, MA, 02108 - Citysearch	http://boston.citysearch.com/profile/4715672/boston_ma/sbarro.html
	Sbarro : DiningGuide Restaurant Profile of Sbarro in Boston ...	http://boston.diningguide.com/data/d103775.htm
	Sbarro	http://sbarro.com
	Sbarro - East Boston - Urbanspoon	http://www.urbanspoon.com/r/4/54076/restaurant/Boston/Sbarro-East-Boston
	Sbarro - Back Bay - Boston 02199	http://www.menuism.com/restaurants/sbarro-boston-617825
	Sbarro - Wikipedia, the free encyclopedia	http://en.wikipedia.org/wiki/Sbarro
Come to Boston.com to get information, reviews, photos, and directions on Sbarro and other Pizza restaurants in Boston, MA. ... (617) 423-2083 - In Short - Founded as an Italian grocery in 1959 Brooklyn, this family-owned chain has since morphed into the		

Continued on next page. . .

Table A.8 – Continued

POI name	Web page title	Web page url
world's largest shopping mall-based ... Sbarro: Boston DiningGuide Restaurant Profile Page ... The Profile Page for this restaurant is brought to you by the DiningGuide.com service. ... Sbarro Mama Sbarros Carmela's of Brooklyn. Search Jobs. SEC Filings Investor Presentations Press Releases. History Management Team. Atlantic City Las Vegas New York Philadelphia ... Sbarro, Pizza Place in East Boston. See the menu. Reviews from critics, food blogs and fellow diners. ... Read reviews of Sbarro in Back Bay Boston from trusted Boston restaurant reviewers. Includes the menu, user reviews, photos, and 48 dishes from Sbarro. ... Sbarro restaurants are located in department stores, shopping malls, airports, service areas, cinemas and college campuses. ... Boston Pizza; California Pizza Kitchen; ...		

In the same line as the other perspectives, following the pipeline of the KUSCO system, the Meaning Extraction module select terms in texts extracted from the Web pages in table A.8. Table A.9 presents the Semantic Index built for each POI in study considering its top-10 concepts ordered by semantic TF-IDF relevance.

POI name	Semantic Index
Au Bon Pain	Poopin, Riverside Menuism, Real Estate, Coast Cafe, Sweet Touch Cafe, Designers and Engineers, Concord Ln, Urbanspoon Boston, Angelos Pizza
Boston Athenaeum	reading, feed, non-member, tick, issue, computers, fun, sore throat, music, lens
Duane Reade	placement, Neighborhood, Krey, Swine Influenza, Italiano, Delancey, Brooklyn, American Express, Las Vegas, San Diego
Erbaluce	image, favorite, Motel, small town, tender, Central, fleming, tarragon, enjoyment, bottle
Harvard Medical School	Department Of Psychiatry, Kimberly Williams, Systems Biology, Kaplan University, Feature Content, Post Bacc, Employee Guide, Shattuck Street, Jules Dienstag, Drosophila
Mail Boxes Etc	Globe, US Post Office, United Parcel Svc Inc, Local Search, Private Mailbox With Street Address, Ups Authorized Outlet, Coolidge Corner South Side, Lifestyle, MA-NH Metro Area, Contact Manta
Petco	ratification, Neighborhood, fundraiser, Hope, return, turn, Itinerary, Tue, push, globe
Radioshack	Tandy, Arnold, Arnold Worldwide Partners, Warlox Wireless, Olufsen, Sony Style, Gale Group, Sprint Store, Electronics Retailers, Apple Store
Sbarro	Povo, Fenway-Kenmore, Danvers Massachusetts, Gina Goff, Beverly Hills, Anthony Missano, Peabody Pizza, Style Pizza, Stuffed Pizza, International Food Pavilion

Table A.9: The semantic indexes in the Open Web perspective for the POIs in study.

Appendix B

Ontology-based Place Classification

As introduced in section7.2.1, we defined *Generic Place Ontologies* a collection of commonsense and generic information about well-known place categories, like restaurants, cinemas, museums, hotels, hospitals, etc. In order to evaluate the three categorization approaches using such external ontologies, we conducted some preliminary experiments[Antunes et al., 2008] with four sets of POIs, already categorized as restaurants and museums from POI directory sites. We then used those three categorization approaches to categorize the POIs according to two ontologies from the four previously selected and mapped in WordNet (section 7.2.1.1). The percentages of correctly categorized POI's for each set are presented in Table B.1.

Although this is a preliminary experimentation, using a total of 116 POI's, the results obtained reveal interesting hints. As expected, the quality of the ontologies

Table B.1: Percentages of correctly categorized POI's.

	Simple	**Weighted**	**Expanded**
Restaurants (I)	71%	29%	59%
Restaurants (II)	70%	41%	69%
Museums (I)	0%	15%	15%
Museums (II)	14%	14%	14%

is crucial to the results of the categorization process. In our experimentation scenario, the ontology representing the restaurants domain was clearly more detailed than that representing the museums domain. Furthermore, the museums ontology was very abstract, which decreases the probability of matching with the specific concepts associated to POIs. In part, this explains the bad results of the POIs representing museums.

Another interesting result is that the simple approach performs better than the weighted approach in most cases. This reveals that somehow the TF/IDF value used for weighting the concepts associated to the POIs is not reflecting the real weight of the concept, which should be improved in a near future. Also, we can conclude that the expanded approach stays very close to the simple approach. In this situation, there is not an evident gain on expanding the concepts to their hyponyms. Again, the quality and detail of the ontologies used may have a strong impact in the results obtained with this approach in the way that when ontologies are not specific enough there is no point on specifying the concepts associated to the POIs.

Also, we can conclude that the expanded approach stays very close to the simple approach. In this situation, there is not an evident gain on expanding the concepts to their hyponyms. Again, the quality and detail of the ontologies used may have a strong impact in the results obtained with this approach in the way that when ontologies are not specific enough there is no point on specializing the concepts associated to the POIs.

Appendix C

A Learning Model for Place Classification

In an experiment of POI Classification, the data used consisted, firstly, of a large set of POIs extracted from Yahoo through their public API, secondly a set provided by Dun & Bradstreet (D&B) and finally a third set from InfoUSA (see table 4.1 for a description of the sources). The POIs from D&B and InfoUSA had a NAICS code assigned (2007 version), but the ones from Yahoo did not. However, each POI from Yahoo was assigned, on average, to roughly two categories from the Yahoo category taxonomy. In summary, the total numbers of POIs used were respectively: 156,364 POIs from Yahoo, 29,402 from D&B and 196,612 from InfoUSA for the area of Boston, Massachusetts. A set of 331,118 POIs from Yahoo and 16,852 from D&B for the New York city area was also used to see how the previously trained model would perform in a different city.

Given its nature, the growth of the Yahoo database (or any other user content platform) is considerably faster than D&B and InfoUSA, and the POI categorization follows less strict guidelines, which in some cases may become subjective. The hypothesis is that there is considerable coherence between Yahoo categories and NAICS codes, such that a model can be learned that automatically classifies incoming Yahoo POIs.

There was, however, a major hurdle that needed to be overcome: the same POI (name, address, latitude, longitude) did not often have the same representation in both databases. This called for a careful *POI Matching* operation (section 4.1.4). Using the thresholds defined in section 4.1.4, by manually validating a random subset of the POI matches identified (6 sets of 50 random POIs assigned to 6 volunteers), we concluded that the percentage of correct similarities identified was above 98% ($\sigma = 1.79$). Differently from validations mentioned in this document, this is an extremely objective one, not demanding external participants or a very large sample[C.1].

After matching Yahoo POIs to D&B and InfoUSA, two different POI databases were built, where each POI contained a set of categories from Yahoo and a NAICS classification provided by D&B and InfoUSA respectively. From this point on, we shall refer to the initial dataset, with results from POI matches between Yahoo and D&B, as dataset A, and to the dataset resulting from the POI matching between Yahoo and InfoUSA as dataset B.

	Dataset A	**Dataset B**
NAICS source	D&B	InfoUSA
Total POIs	7289	44634
Distinct NAICS	504	689
Distinct categories	802	1109
Distinct category combinations	569	1002
Category combinations that appear only once	136	92
Categories that appear only once	181	107
NAICS that appear only once	115	96

Table C.1: Some statistics of datasets A and B for Boston

Table C.1 shows some statistical details of both datasets used. Dataset A contained 7,289 POIs for Boston and Cambridge and 2,415 for New York. In comparison with the original databases, these were much smaller sets due to a very conservative matching approach (string similarity of at least 80%, max distance of 80 meters). However the POI quantities were high enough to build statistically valid models. A detailed analysis of this data was performed and it identified 569 different category combinations, which included only 802 distinct categories from the full set (of over 1,300). From D&B, the data covers 504 distinct six-digit NAICS codes. However, the 2007 NAICS taxonomy has a total of 1,175 six-level categories, meaning that this sample data only covered some of the most common NAICS codes, which only represented about 43% of the total

[C.1] Using the central limit theorem, the standard error of the mean should be near 0.73. Assuming an underestimation bias for n=6 of 5% (according to [Gurland and Tripathi, 1971]), accuracy keeps very high, being the 95% confidence interval [96.5%, 98.7%]

number of NAICS categories.

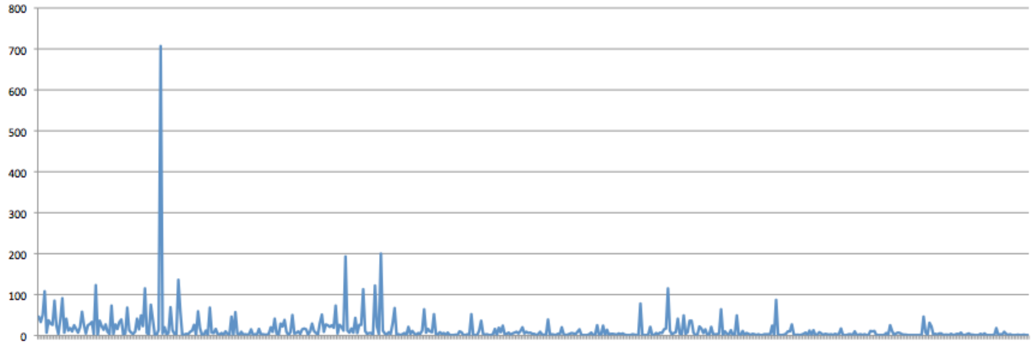


Figure C.1: Distribution of the POIs in dataset A along the different NAICS code

Figure C.1 shows the distribution of POIs among the different NAICS codes for dataset A. As can be seen in the chart, the distribution was far from being uniform. Further analysis of the coherence between NAICS and Yahoo shows that only in 80.2% of the POIs in dataset A was the correspondent NAICS code consistent with the most common one for that given set of categories, which means that about one fifth of the POIs were no coherent with the rest of the sample. For different NAICS levels, particularly for two-digit and four-digit NAICS codes, the same analysis showed, as expected, a higher level of coherence, for the two and four-digit NAICS, 87.1% and 83.4% of the POIs, respectively. Therefore, by having the same set of Yahoo categories mapping to different NAICS codes on different occasions, it is not possible to expect to obtain a perfect model that classifies correctly all test cases. In order to understand the impact of these inconsistencies in the results, the POI dataset was also modified so that the NAICS code of a given POI would match the NAICS codes of the other POIs with the same category set, assigning to each POI the most common NAICS code for that given category set in the dataset.

Tables C.2 and C.3 show, respectively, the five most common NAICS and Yahoo categories that were identified in dataset A. Regarding dataset B, 689 distinct NAICS codes were identified, as were 1,109 distinct categories of the more than 1,300 that Yahoo has. This number of distinct categories was almost double that of dataset A (only 802) and therefore provided a better coverage of the source taxonomy. The number of distinct category combinations almost doubled when compared to dataset A, which leads to greater diversity in the training data and hopefully more accurate classifiers.

NAICS code	**Description**	**Occurrences**
423730	Warm-Air Heating and Air-Conditioning Equipment and Supplies Merchant Wholesalers	707
446130	Optical Goods Stores	200
314999	All Other Miscellaneous Textile Product Mills	193
493120	Refrigerated Warehousing and Storage	136
332997	Industrial Pattern Manufacturing	123

Table C.2: Most common NAICS in the dataset A

Yahoo category	**Occurrences**
Salons	157
All Law Firms	129
Government	116
Trade Organizations	115
Architecture	86

Table C.3: Most common Yahoo categories in the dataset A

Figure C.2 shows the distribution of POIs among the different NAICS codes for dataset B. Like that for dataset A, shows an irregular distribution.

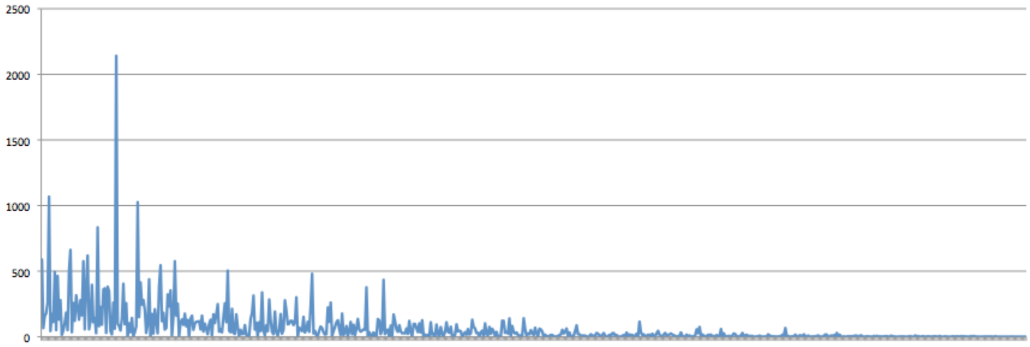


Figure C.2: Distribution of the POIs in dataset B along the different NAICS code

Another possibility for generating a training set would be to manually classify a small set of Yahoo POIs to the NAICS codes. Even though this would be a terribly laborious and time-consuming task, it might generate a more consistent training set, since the NAICS classifications provided by D&B and InfoUSA result from the contribution of multiple users/sources, which makes them somehow subjective (bearing in mind that the NAICS codes of businesses are not always simple to identify). However, here the objective is to automate the classification process as much as possible, therefore the previously described approach through POI Matching (section 4.1.4) was chosen. Also, manually producing training sets with the dimensions of those used in practice would not be feasible.

Table C.4 shows the accuracies obtained using different machine learning algorithms (presented in Chapter 4 on section 4.1.5) for different NAICS levels (two, four and six-digit codes) using dataset A. There are some missing results in the table because the algorithm took over 72 hours to run. Before we start analyzing the machine learning results, it is important to mention that the ZeroR and OneR algorithms, because of the way they work, were not applied to "compete" for the best results against the other algorithms. Instead, they merely serve as baselines for the other algorithms.

As expected, we got better results classifying POIs to the two-level NAICS than for the six-level NAICS, since the eventual noise due to ambiguous classifications in

Algorithm	**NAICS2**	**NAICS4**	**NAICS6**
FT	85.759	-	-
ID3	84.248	75.837	72.119
J48	83.397	75.755	71.282
J48graft	83.823	76.358	71.776
RandomForest	84.879	77.099	72.983
RandomTree	84.207	75.906	72.379
DecisionTable	77.840	71.256	-
JRip	79.624	72.187	67.838
IB1	80.736	70.952	65.299
IBk	84.989	76.811	73.052
K*	84.893	77.566	73.408
BayesNet	80.681	56.394	42.440
NaiveBayes	74.547	40.354	28.444
MultilayerPerceptron	5.762	-	-
ZeroR	14.586	9.701	9.701
OneR	21.858	12.349	12.088

Table C.4: Results obtained for the different machine learning algorithms with POIs from dataset A for the Boston area

Algorithm	**NAICS2**	**NAICS4**	**NAICS6**
ID3	92.975	89.728	88.680
RandomForest	93.609	90.805	89.846
IBk	94.170	91.189	89.979

Table C.5: Results obtained for the different machine learning algorithms using a re-classified dataset

the POI datasets is smaller.We can see that the tree-based (e.g. ID3, RandomForest) and instance-based learning approaches (e.g. IBk, K*) are the ones that perform better in this classification task, especially the latter. Notice that only 80.2% of data is classified in a totally non-ambiguous way. The most successful algorithm is IBk (with $k = 1$), which essentially finds the similar test case and assigns the same NAICS code. The difference in accuracy between tree-based and instance based approaches is very small to make strong conclusions, however we could expect that instance based models bring better results since the distribution of the different Yahoo! categories is relatively even among examples of the same NAICS code (implying no clear "dominance" of some categories over others). Understandably, the Naive Bayes algorithm performs badly because the assumption that different Yahoo! categories for the same NAICS classification are independently distributed is obviously false (e.g. "Doctors & Clinics, Laboratories, Medical Laboratories" are correlated). Such assumption is not fully necessary in Bayesian Networks, which actually brings better results. Unfortunately, we could not find a model search algorithm that performs in acceptable time (less than 72 hours) and produces a more accurate model. We used Simulated Annealing and Hill Climbing.

In table C.5 we can see the results obtained by modifying the POI dataset, so that the NAICS codes of POIs where ambiguities arise are grouped together in the same "super-category", eliminating the inconsistencies. By comparing the results in table C.5 with the results in table C.4, we realize that the NAICS labeling inconsistencies in the POI data have a major negative effect in the performance of the machine learning algorithms, reducing the accuracy in more than 16% in some cases for the six-level NAICS codes.

It would be expectable to obtain accuracies more close to 100% for the results in table C.5. However, that does not happen due to the fact that 115 of the 514 NAICS

Algorithm	NAICS2	NAICS4	NAICS6
ID3	83.967	75.986	72.087
J48	82.935	75.504	71.559
RandomForest	86.467	78.876	75.848
RandomTree	81.467	73.417	69.289
JRip	81.307	73.509	69.289
IB1	85.435	76.582	72.706
IBk	86.903	79.059	75.619
K*	86.834	79.541	76.261
BayesNet	80.183	56.467	40.137
NaiveBayes	74.541	30.688	20.091

Table C.6: Results obtained for the different machine learning algorithms with POI data from the Boston area using semantic annotations

codes covered by our dataset A only occur once. Therefore, when we split the dataset to perform the ten-fold cross-validation, a significative number of the test cases will have NAICS codes that the algorithm was not trained for, causing it to incorrectly classify them.

Table C.6 shows the results obtained using both the categories from Yahoo! and the semantic annotations. Please remember that the presented results are based on semantic enriched POIs from dataset A. By comparing the results in table C.6 (with semantics) with the ones presented in table C.4 (without semantics), we can see that there was some improvement in some algorithms (like IBk). This is somehow understandable if we consider the way IBk (k-nearest neighbor) works. Since it measures the euclidean distances from the test case to all training examples, having more information about the POI (in this case semantic annotations) would supposably help, thus increasing the accuracy. On the other hand the performance of other algorithms such as ID3 decreased when compared to the results from table C.4. This fact suggests that having semantic information about the POIs might be difficulting the choice of the next feature to use in the decision tree by messing with the entropy and gain computation.

After comparing several classification approaches with and without enriched information obtained from semantic annotations, the results were applied to the urban modeling task of estimating employment size at a disaggregated level. This task is traditionally made at a higher, less focused level (Traffic Analysis Zone, Census Tract or Block Group level) than what might now be possible. This part of the work was done in collaboration with Shan Jiang (shanjang@mit.edu) and Professor Joseph Ferreira (jf@mit.edu) at MIT.

Appendix D

Yahoo Taxonomy

Figure D.1 presents the root categories of Yahoo! Local. This taxonomy was automatically extracted from http://local.yahoo.com. For each root category, the direct descendants are presented in the table D.1. We opted for only showing until the 2nd level in the taxonomy hierarchy due to page-size limitation. The complete taxonomy can be seen in detail at http://eden.dei.uc.pt/ ana/phd/yahootaxonomy. This taxonomy comprises the POI categories from Boston and New York as they were presented in May 2011.

	Taxonomy depth
1st level	**2nd level**
Automotive	All Terrain Vehicles
	Antique Cars
	Auto Appraisers
	Auto Conversions
	Auto Detailing
	Auto Wreckers
	Automotive Buyers Services
	Automotive Supplies
	Car Accessories
	Car Security
	Car Transport
	Car Washes
	Customs & Tuning
	Dealers
	Driving Schools
	Gas Stations
	Motor Scooters

Continued on next page

Table D.1 – continued from previous page

	Taxonomy depth
1st level	**2nd level**
	Motorcycles
	Parts
	Racing
	Recreational Vehicles
	Rental
	Repair & Service
	Tires
	Towing
	Trailers
	Trucks
	Vans
	Vehicle Painting & Lettering
Business to Business	
	Advertising
	Business Services
	Communications & Media
	Computers & Electronics
	Construction & Real Estate
	Conventions & Trade Shows
	Education & Training
	Energy & Mining
	Entertainment & Recreation
	Environmental & Ecological Services
	Finance & Investment
	Food & Agriculture
	Government & Law
	Health & Medicine
	Manufacturing & Industrial Supplies
	Office Supplies & Equipment
	Printing & Publishing
	Science & Technology
	Security
	Shipping
	Transportation
Computer & Electronics	
	Communications & Networking
	Computer Furniture
	Computer Graphics
	Computer Multimedia
	Computer Rental
	Computer Repair

Continued on next page

Table D.1 – continued from previous page

Taxonomy depth	
1st level	**2nd level**
	Computer Software
	Computer Stores
	Computer Training
	Consulting
	Desktop Publishing
	Hardware
	Internet Services
	Organizations
Education	
	Academic Specialty Schools
	Adult
	Colleges & Universities
	K-12
	Language Schools
	Learning Disabilities Schools
	Preschools
	Private & Parochial Schools
	School Supplies
	Schools for the Handicapped
	Sports Training
	Tutoring
Entertainment & Arts	
	Artists
	Bars
	Dance
	Entertainers
	Entertainment Production
	Entertainment Venues
	Event Planners
	Movies & Film
	Museums & Galleries
	Music
	Party Rentals
	Talent Agencies
	Theatre
	Tickets
	Video
	Writers
Food & Dining	
	Banquet Rooms
	Beverages

Continued on next page

Table D.1 – continued from previous page

Taxonomy depth	
1st level	**2nd level**
	Candy & Sweets
	Catering Services
	Culinary Schools
	Grocery Stores
	Meats
	Natural & Organic Foods
	Poultry Retailers
	Produce Retailers
	Restaurants
	Seafood
	Specialty Food Stores
Government & Community	
	Animal & Humane Societies
	City Hall
	Community Centers
	Correctional Institutions
	Courts of Law
	Crime Prevention
	Disabled Services
	Embassies & Consulates
	Emergency Services
	Family Services
	Finance & Taxation
	Fire Protection
	Food Hunger
	Funerals & Memorials
	Government
	Housing
	Housing Authorities
	Immigration Services
	Law Enforcement
	Legal Services
	Lesbian
	Libraries
	Military
	Other Organizations
	Philanthropy
	Post Offices
	Recycling
	Religion & Spirituality
	Senior Services

Continued on next page

Table D.1 – continued from previous page

Taxonomy depth	
1st level	**2nd level**
	Social Services
	Vehicle Registration
	Veteran Organizations
Health & Beauty	
	Abuse Treatment Centers
	Alternative Medicine
	Barbers
	Bath & Body Products
	Beauty Salons
	Beauty Supplies
	Care Providers
	Cosmetics
	Cosmetology
	Day Spas
	Dental
	Diet Centers
	Doctors & Clinics
	Drug Stores
	Eye Care
	First Aid
	Fitness
	Fragrances
	Hair
	Hospitals & Medical Centers
	Laboratories
	Medical Education
	Medical Supplies & Equipment
	Mental Health
	Nail Care
	Pharmacies
	Skin Care
	Tanning Salons
	Tattoos & Piercings
	Women's Health & Reproduction
Home & Garden	
	Appliances
	Bed & Bath
	Carpets & Rugs
	Cleaning Services
	Construction
	Furniture

Continued on next page

Table D.1 – continued from previous page

	Taxonomy depth
1st level	**2nd level**
	Home & Garden Lighting
	Home & Garden Retailers
	Home Decor
	Housewares
	Indoor Air Quality
	Lawn & Garden
	Locksmiths
	Pest Control
	Plumbing
	Pools
	Safety
	Snow Removal
	Telephone
	Tree Service
	Utilities
	Waste Management
	Welding Services
	Window Treatments
Landmark	
	Basin
	Bay
	Beaches
	Bend
	Border Crossings
	Bridges
	Buildings
	Canals
	Cape
	Channel
	Cliffs
	Dams
	Falls
	Flats
	Forests
	Gaps
	Harbors
	Historical Military Facilities
	Historical Monuments
	Islands
	Isthmus
	Lakes

Continued on next page

Table D.1 – continued from previous page

Taxonomy depth	
1st level	**2nd level**
	Levees
	Mines
	Natural Arch
	Natural Springs
	Oil Fields
	Parks
	Pillars
	Ranges
	Rapids
	Reserve
	Reservoirs
	Ridges
	Seas
	Streams
	Summit
	Swamps
	Towers
	Trails
	Tunnels
	Valleys
	Wells
	Woods
Legal & Financial Services	
	Accounting & Bookkeeping
	Arbitration & Mediation
	Bail Bonds
	Bankruptcy Services
	Banks
	Cash & Check Advances
	Court Reporting Services
	Credit & Debt Services
	Credit Reporting Agencies
	Credit Unions
	Currency Exchanges
	Financing
	Fingerprinting
	Fund Transfer Services
	Insurance
	Investment Services
	Law Firms
	Notaries

Continued on next page

Table D.1 – continued from previous page

Taxonomy depth	
1st level	**2nd level**
	Taxes
Professional Services	
	Animals & Pets
	Auctioneers
	Business Organizations
	Dating Services
	Design
	Employment
	Genealogy
	News & Media
	Office Supplies & Services
	Paranormal Phenomena
	Photocopying
	Photography
	Printing
	Private Investigation
	Publishing
	Sex
	Shipping
	Storage
	Telecommunications
Real Estate	
	Apartments
	Corporate Housing
	Escrow Services
	Home Builders
	Mobile Home Parks
	Other Real Estate
	Planned Communities
	Property Management
	Real Estate Agents
	Real Estate Appraisal
	Real Estate Inspection Services
	Real Estate Title Companies
	Relocation Services
	Rental Agencies
	Timeshares
	Vacation Rentals
Recreation & Sporting Goods	
	Amusement & Theme Parks
	Archery

Continued on next page

Table D.1 – continued from previous page

Taxonomy depth	
1st level	**2nd level**
	Arenas & Stadiums
	Aviation
	Baseball
	Basketball
	Billiards
	Boating
	Bowling
	Camping
	Climbing
	Cycling
	Equestrian
	Fencing
	Fishing
	Football
	Gambling
	Golf
	Gymnastics
	Hobbies
	Hunting
	Marinas
	Martial Arts
	Other Sports
	Outdoor Parks
	Paddling
	Paint Ball
	Playgrounds
	Skating
	Skiing
	Skydiving
	Soccer
	Sporting Goods
	Surfing
	Swimming & Diving
	Tennis
Retail Shopping	
	Antiques & Collectibles
	Arts & Crafts
	Baby Accessories & Services
	Bookstores
	Clothing
	Convenience Stores

Continued on next page

Table D.1 – continued from previous page

	Taxonomy depth
1st level	**2nd level**
	Events & Occasions
	Flowers
	Footwear
	Gifts
	Holiday
	Home Electronics
	Jewelry & Watches
	Luggage & Accessories
	Shopping Venues
	Toys & Games
	Trophies
	Weapons
	Weddings
Travel & Lodging	
	Airlines
	Airports
	Charter Buses
	Charter Flights
	Concierge Services
	Cruises
	Hotels & Lodging
	Limos & Shuttles
	Mass Transit
	Parking Services
	Passport & Visa Services
	Taxi Services
	Tour Operators
	Tourist Attractions
	Travel Agents

Table D.1: The first two levels in Yahoo taxonomy of POI categories. The Landmark category is not included since it is composed of geographical land marks, and not by public places.

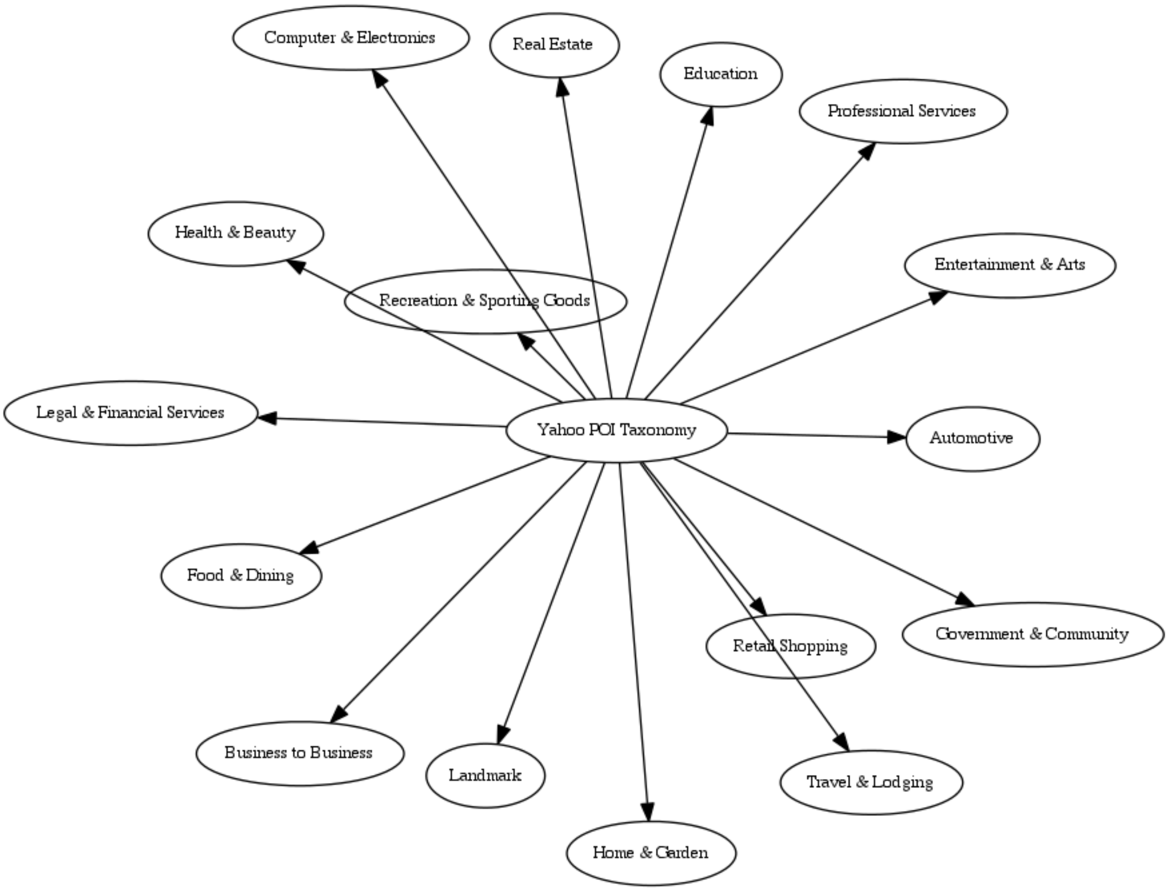


Figure D.1: Root categories in Yahoo taxonomy.