Semantic Enrichment of Places: Ontology Learning from Web

Ana Alves, Bruno Antunes, Francisco C. Pereira, Carlos Bento
CISUC, Department of Informatics Engineering
University of Coimbra
Polo II, 3030-290 Coimbra, Portugal
http://cisuc.dei.uc.pt
{ana, bema, camara, bento}@dei.uc.pt

November 15, 2008

Submitted for a Special Edition: Intelligent Agents and Services for Smart Environments
Special Edition Guest Editors: F. S. Corra da Silva and S. Bandini

Abstract

In this paper, we present an approach to a challenge well known from the area of Ubiquitous Computing: extracting meaning out of geo-referenced information. We believe that Public Places can also be an environment where users can be confronted with intelligent services depending on their needs. To make this a reality we need firstly an intelligent way to represent places. The importance of this "semantics of place" problem is proportional to the number of available services and data that are common nowadays. Having rich knowledge about a place, we open up a new realm of "Location Based Services" that can behave more intelligently. Our approach builds on Ontology Engineering techniques in order to build a network of semantic associations between a place and related concepts. We briefly describe the KUSCO system and present some preliminary results.

Keywords: Point of Interest, Place, Semantic of Place, Ontologies, Ontology Enrichment, Ontology Evaluation.

1 Introduction

The current ubiquitous availability of localization technologies (particularly GPS) has driven to the emergence many new applications (the "Location Based Services" or LBS) and enormous amounts of geo-referenced data. However, although we have already available rich and sophisticated knowledge representation and techniques (e.g. Semantic Web and Ontology Engineering) that allows for elaborate uses, information on location or *place* tends to be poor, with little or no directly associated semantics (e.g. the typical Point Of Interest or POI simply has a description and a generic type; in other cases we only have the latitude/longitude pair; and sometimes an LBS has its own purpose driven semantics, unusable to others). The association of a set of semantic meanings to a place should allow the application of those sophisticated techniques and foster the quality of current and future LBS (e.g. with better indexing). So, one asks: How can we extract semantics from a place? What is the meaning of place? KUSCO is a

system that intends to find and associate a set of semantic tags to Points Of Interest (or POI's). A POI is a geo-referenced tag that contains a latitude/longitude pair, a name and a type (e.g. restaurant, museum, gas station). Such information is used by KUSCO, which applies a number of techniques to automatically extract information from the Web about that POI. The system starts by doing a web search (e.g. using Google), then it extracts and calculates statistics about the main words used. Afterwards, it associates these words to concepts (of WordNet [1]), and finally it determines the relationships of these concepts with the POI by using Place Ontologies.

In this paper, we describe the Ontology Enrichment process of the KUSCO system. This process is necessary to establish the association between the Generic Place Ontologies and specific POI instances, which is also normally called Ontology Instantiation. Given a POI and an Ontology, KUSCO seeks for the associations between the Ontology *Terms* and the more relevant words found in the Web pages most related to the POI. In the next section, we present "Semantics of Place" problem, well known in the area of Ubiquitous Computing. Then, we present some Ontology Engineering concepts that are essential to this work: Learning and Evaluation. In section 3, we present KUSCO and show some preliminary results. We end the paper by discussing next steps (section 4) and final remarks (section 5).

2 State of the Art

Being an interdisciplinary work, to associate meaning to Public Places involves not only Semantic Reasoning, the final step, but also a pre-processing phase over our initial representation about places: a POI. With the aim of knowing the context about a place, first we use algorithms and techniques from Information Extraction over Web, a subfield in Natural Language Processing. To efficiently represent this extracted knowledge we represent this information using Ontologies and we take advantage of them formal rules and power to generalize to structure and deduce other implicit relationships between concepts. In this section, we present the three main areas of knowledge where our work is based: Ambient Intelligence, Natural Language Processing and Ontologies, focused in the topic which we apply in KUSCO system: Semantics of Place, Web Information Extraction and Ontology Engineering.

2.1 Semantics of Place

First introduced in [2], Jeffrey Hightower argues that location must have more associated information than simply the absolute position in a global coordinate system. Location representation needs more human-readable information including geographic, demographic, environmental, historical and, perhaps, commercial attributes. The meaning of place derives from social conventions, their private or public nature, possibilities for communication, etc. [3, 4]. As argued by [5] on distinguishing 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 - having also temporal properties, once the same space can be different places at different times. Thus, a place only exists if it has some meaning for someone and the construction of this meaning is the main objective of our research.

As a formal definition, location models can be classified into four main types [6]: Geometric, Symbolic, Hybrid or Semantic. While the first three models (the third considers both geometric and symbolic) are mainly devoted to spatial relationship between locations, the last one, the Semantic Location Model, is orthogonal to symbolic and geometric representations. The semantic representation provides other information around its place, such as a bus route or a snapshot of interest. As an example of semantic representation, the HP Cooltown [7] introduces a semantic representation of locations. Its main goal is

to support web presence for people, places and things. They use Universal Resource Indentifies (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. In the same line, Ubiquitous Web [8] was envisioned as a pervasive web infrastructure in which all physical objects are socially tagged and accessible by URIs, providing information and services that enrich users experiences in their physical context as the web does in the cyberspace.

Automatic association between locations and places and category names is made in [9], where a diary containing visited places is incrementally built as a user is walking and staying in some locations in the city. By using reverse geocoding and White-pages Web services, the system is able to collect locations and associate them which public places they refer, being possible to retrieve information from this closed structured database (white-pages). A quality step forward would be reasoning about the common-sense meaning of those places (perhaps using WordNet [1]).

While the focus of our work is the Semantic aspect of Location Representation, we also take advantage of information available on the Web about public places. With the growth of the World Wide Web, we think that almost every commercial and non-commercial entities of public interest are or tend to become present on-line by proper web sites or referred by other related institutions. This should become even more relevant for places considered interesting for a group of people (i.e. they are in a sense a *Point Of Interest*). But differently from the two previous semantic models, we don't assume that Semantic Web is already a reality, with all information semantically structured and tagged. Actually, it is widely accepted that the majority of on-line information is composed of unrestricted user-written texts, so we get mainly dependent on the Information Extraction (IE) capabilities (we will discuss this later on section 2.2).

2.2 Web Information Extraction

Generating or populating ontologies from the Web is not a new topic of research, in [10] the authors take advantage of the natural interlinked organization of the Web to generate a taxonomy of keywords. They propose an approach based on extracting information from menus and navigation indicators to automatically generate a first raw ontology about a Web site. However, once it is widely accepted that the majority of on-line information is composed of unrestricted user-written texts, we are mainly concerned about the Information Extraction (IE) field, which is a research subtopic in IR devoted to extract useful information from a body of text, including techniques like Term Extraction and Name Entity Recognition. Being Web Information Extraction (WIE) subtopics within Information Retrieval (IR) devoted to extract useful information from written data, IE applies classic Natural Language Processing (NLP) techniques and resources over unstructured pages written in natural language where no structure can be found. Differently, WIE usually applies machine learning and pattern mining to exploit the syntactical patterns or layout structures of the template-based documents. In our case, since it is impossible to guarantee that public places have structured or semi-structured pages presenting their services, it is impossible to learn the layout for every new page. In the Natural Language Processing field, there are other techniques which will be further used, including part-of-speech tagging and word sense disambiguation to discover meaningful key concepts from the Web and contextualize it in a Common Sense Ontology. In [11], the Artequakt system uses natural language tools to automatically extract knowledge about artists from multiple documents based on a predefined ontology to generate artist biographies. The system uses a biography ontology, which defines the data for an artist biography. Information is collected by parsing text found on the Web and is subsequently presented using templates. It assumes that Web pages are syntactically well-constructed in order to extract knowledge triples (concept - relation - concept). Web pages are divided into paragraphs, and consequently in sentences. Each sentence, which heuristically

corresponds to a grammatical construction of the form Subject-Verb-Concept, is then used to fulfill a triple. In our case, we assume that the Web is a huge repository of knowledge but with no structure. Unlike texts, there are no guarantees that we can find such syntactic properties in facts about places.

2.3 Meaning and Ontologies

Common Sense Ontologies (such as WordNet, OpenCyc [12], ConceptNet [13] or others) are collections of trivial and semantic knowledge that allow the extension of the computational reasoning process. We can focus on generic concepts and relationships about a known category of places (restaurant, museum, hospital, cinema, pharmacy, etc.) in order to be build a *Common Sense Place Ontology* comprising not only semantic related concepts to a given category but all concepts referred by descriptive definitions (or glossaries).

The growing amount of information available on the web demands for the development of efficient and practical information extraction approaches, in order to avoid the actual user's overloading of information. This need for new ways of extracting information from the web stimulated a new vision, the Semantic Web [14], where resources available have associated machine-readable semantic information. For this to come true, a knowledge representation structure for representing the semantics associated to resources would be necessary, and that was where ontologies [15] assumed a central role in the movement of the Semantic Web. Because it is nearly impossible to design an ontology of the world, research focused on the development of domain-specific ontologies, in which construction and maintenance are time-consuming and error-prone when manually done. In order to automate this process, research on ontology learning has emerged, combining information extraction and learning methods to automatically, or semi-automatically, build ontologies.

2.3.1 Ontology Learning

According to [16], 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. On the other hand, ontology enrichment is the task of extending an ontology by adding new concepts, relations and rules, which results in changes on its structure. Because errors and inconsistencies can be introduced during ontology population and enrichment, inconsistency resolution aims to detect these inconsistencies and generate appropriate resolutions. Finally, the ontology evaluation task assesses the ontology by measuring its quality with respect to some particular criteria (see section 2.3.2).

The ontology learning process can be performed through three different major approaches [16]:

- The integration of ontologies by capturing the features that are shared between them. The integration process can assume different forms: the creation of a single ontology from the merge of the others; the alignment between ontologies, using links between them that allow their reuse from one another; and the mapping of ontologies through corresponding elements between them.

- The construction of a new ontology from scratch, based on the information extracted from data about a specific domain.

- The specialization of a generic ontology by adapting it to a specific domain.

Following the work of Buitlaar et al. [17], the ontology learning process deals with six different aspects related with the structure of an ontology: terms, synonyms, concepts, concept hierarchies, relations and rules.

2.3.1.1 Term Identification

The term extraction phase is in the basis of every ontology learning process. A term is *"an instance of a recognisable entity in a corpus that conveys a single meaning within a domain (concept)"* [16]. The main objective of term identification is to find terms in the corpus, that possibly represent a concept and can be used to enrich an ontology. The most successful approaches for term recognition are those based on statistical methods, which usually support on occurrence frequencies to find the importance of each term in relation to others. The TF/IDF [18] metric is commonly applied [19] and complemented with other methods, such as latent semantic indexing [20], or co-occurrence information [21]. Recurring to clustering techniques and other resources, such as WordNet [1], groups of similar terms are created, which possibly represent the same concept [22]. Natural Language Processing [23] techniques are commonly applied, such as morphological analysis, part-of-speech tagging and syntactic analysis, to enhance frequency and clustering approaches [24].

2.3.1.2 Synonym Identification

When a set of terms refers to the same concept or relation, they are said to be synonyms. A lot of work has been done concerning the identification of synonyms, especially using resources such as WordNet [1] and applying word sense disambiguation techniques to find the sense of each term. Other techniques include clustering approaches [25] and information retrieval algorithms, such as Latent Semantic Indexing algorithms (LSI, LSA, PLSI, etc.) [26]. Since terms are domain-specific and the majority of terms consist of more than one word, Term Sense Disambiguation has been proposed [27] making no use of any general language resources but taking the Web to retrieve contextual information.

2.3.1.3 Concept Identification

Concepts are an important aspect of any ontology, but different views exist on what constitute a concept [17]:

- An intentional definition of concept. This definition can be of two types: informal or formal. The informal definition defines a concept in a descriptive way, while the formal definition defines a concept in terms of properties and relations between them.
- A set of concept instances. This is achieved through a process known as ontology population or ontology tagging.
- A set of realizations (i.e. terms). This is based on clusters of terms that form the realizations of the concept.

2.3.1.4 Taxonomy Construction

A hierarchy of concepts, or taxonomy, is constructed with inclusion relations (usually known as *"is-a"* relations) and represent an important aspect of an ontology. These relations are typically identified using lexico-syntactic patterns [28]. Recent systems apply machine learning and pattern learning algorithms to automate the process of defining such patterns [29].

2.3.1.5 Semantic Relations Extraction

Besides the inclusion relations, an ontology typically contains non-taxonomic relations that connect semantically related concepts. Again, lexico-syntactic patterns are commonly used to identify this kind of relations. Some approaches exploit the fact that verbs represent an action or relation between concepts in sentences [30].

2.3.1.6 Rule Acquisition

Rule acquisition is the least explored aspect of ontology learning. Some work have been developed for extracting rules from text [31] and advances have been made in inductive logic programming to address reasoning in the Semantic Web [32].

2.3.2 Ontology Evaluation

In this subsection, we will discuss some of the techniques and metrics that can be used for ontology evaluation. The need for well defined techniques of ontology evaluation arises from the fact that different ontology conceptualizations can be constructed from the same body of knowledge. Much of the work developed in this field came from the context of ontology learning and enrichment, where different evaluation approaches were explored to evaluate the resulting ontologies. Also, the increasing development of semantic-aware applications, that make use of ontologies, uncovered the need to evaluate the available ontologies and choose the one that best fits the specific needs of the application.

According to [33], there are four different categories of techniques used for ontology evaluation: those based on a "golden standard", those based on the results of an application that makes use of the ontology, those based on the use of a corpus about the domain to be covered by the ontology, and those where evaluation is done by humans.

When the semantic characterization of place involves the construction and enrichment of place ontologies, it becomes necessary to apply some of the techniques developed for ontology evaluation, so that we can assess the quality of the ontology produced and validate the proposed ontology enrichment approach.

2.3.3 Ontology Selection

The ontology selection topic, also known as ontology ranking, has been gaining importance within the Semantic Web community, to a great extent due to the increasing number of ontology repositories available in the web [34]. Ontology selection is defined *"as the process that allows identifying one or more ontologies or ontology modules that satisfy certain criteria"* [34] and has ontology evaluation in its basis, since most of the approaches for ontology selection rely on an ontology evaluation criteria to achieve their objective.

According to [34], the different ontology selection approaches can be distinguished by the selection criterion that is applied: popularity, i.e., when the selected ontology must be the most referenced among

all considered ontologies; richness of knowledge relying on the structure of each ontology; and finally, topic coverage, or by other words, the extent to which an ontology cover a certain knowledge domain.

In relation to the semantic of place topic, ontology selection may have an important role when it comes to choose the right ontology to represent places and their semantic. Also, in a process of ontology learning or enrichment, when we want to have an ontology as a start point for the whole process, it is important to have a well defined criteria for choosing the ontology that better adjusts to the semantic of the domain we pretend to represent.

3 KUSCO

The problem of "position to place" is a well known challenge within the area of Ubiquitous Computing and relates deeply with the connection humans have with places, their functionality and meaning. Attached to a tag name, even when a category is included, a place needs a richer semantic representation in our perspective in order to be understood. This knowledge can be used for whatever processes that demand semantics of place (e.g. understanding POIs while in navigation; searching for a place that has specific characteristics; route planning using locations with specific functionalities; inferring users activity, etc.). We formally name this process as Semantic Enrichment of Place and it consists of using available Common Sense Ontologies and Web information to build a collection of generic and instance facts about these places. We present here a PhD project named KUSCO (Knowledge discovering by Unsupervised Search from web to instantiate Community's Ontologies) where from a Point of Interest (POI) we tend to instantiate the specific Ontology which this place belongs to. Using the figure of KUSCO Architecture (see 5) as guide to explain the component modules, our system consists of:

- *Generic Place Ontologies*: For each place category we plan to work with, we try to find the most popular, already built and on-line ontology representing the given domain.
- *Geo Web Search*: Once ontologies have been chose, we collect POIs from different sources and in different formats to feed our system. For each POI, we apply reverse geocoding in order to find geographic information like City and Village from where the POI is related. This information added to Place Name is used to retrieve its most relevant Web pages.
- *Meaning Extraction*: NLP techniques are applied to the set of most relevant pages to retrieve key concepts to the given place. These concepts are also contextualized in WordNet to take advantage of meaning of concepts and semantic relations between them.
- *Place Categorization*: POI often is composed of a triple (Lat, Long, Place Name). The categories, or ontology, of these places are found in this module using some semantic similarity measures against relevant concepts found previously.
- *Ontology Population*: Once a POI has been classified and its specific ontology retrieved, the meaning of this place (its relevant concepts) are used to populate the given ontology.

After this brief description, the following subsections will explain in detail each component of this system.

3.1 Generic Place Ontologies

The module of Generic Place Ontologies represents a collection of commonsense and generic information about well-known place categories, like restaurants, cinemas, museums, hotels, hospitals, etc. At a first stage, this information is manually collected from well-known and shared Ontologies (retrieving and selecting the most popular using ontology search engines like [35]). But as the system is used, it is 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 place meaning, ontologies are contextualized on WordNet [1]. For each term in an ontology, a WordNet's definition will be looked for.

3.2 Geo Web Search

This module is responsible for finding Web pages using only POI data as keywords: place name and geographical address. This last element is composed of the City name (where POI is located) and is obtained from Gazetteers [1] available on Web). This search is presently made by the freely available Yahoo API.We are applying a simple heuristic that use the geographical reference as another keyword in the search. Thus, assuming a POI is a quadruple (Latitude, Longitude, [Category,] Name)[2], the final query to search will be: "City Name" + ["Category" +] "Name". At this moment our system is very sensitive to geographical location of Place Name. For example, after looking for specific Web information for a given POI named "Carnegie Hall" in New York, we find 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 geographical search gives us other definitions different from a hospital, such as a metropolitan neighbourhood. This shows us that this approach can become very dependent of search algorithms and of the Web's representativeness of places. At the end of this process, the N more relevant pages are selected (as suggested by the search engine).

3.3 Meaning Extraction

Having the set of Web pages found earlier, keyword extraction and contextualization on Wordnet is made at this point. This processing includes POS tagging and Word Sense Disambiguation using available NLP tools [37, 38]. On completion of these sub tasks for each web page, we are able to extract the most relevant terms (only common or proper nouns) that will be used in the categorization task (next module). These nouns are contextualized on WordNet and thus can be thought not only as a word but more cognitively as a concept (specifically a synset - family of words having the same meaning, i.e., synonyms [1]). Each concept's importance is computed by tf-idf weighting [39] by two ways: considering only the most relevant WebPages retrieved for that POI (local TF/IDF) and considering the most relevant WebPages for all POI's on that category (global TF/IDF). At this stage each POI is represented by a list of more relevant WordNet concepts and NE terms, or in other words by its *semantic index*.

3.4 Place Categorization

In order to evaluate the capacity of categorizing POI's (i.e. if they represent restaurants, museums, bars, etc.) we selected a set of ontologies using a popularity based criteria (see section 2.3.3). The result of this

[1] 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 [36].

[2] This category refers to the type of POI in question, a museum, a restaurant, a pub, etc. This information is optional, once sometimes it may not be present in the POI.

ontology selection process was a set of four ontologies about different domains: restaurants[3], museums[4], pubs[5], travel[6] and shows[7].

In an initial phase, already described, POI's were associated to a set of WordNet concepts resulting in their semantic indexes. To facilitate the categorization of a POI semantic index against this set of ontologies, we also map the concepts of the selected ontologies in WordNet. The mapping comprises three phases:

1. *Term Identification.* The terms are extracted from the name of the concepts contained in the ontology. Because these names are usually comprised of one or more terms, they are split by upper case letters and special characters such as '-' and '_'. For instance, a class named 'PortugueseCuisine' will be split in two terms: 'Portuguese' and 'Cuisine'.

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

3. *Concept Identification.* The terms and combinations of terms, extracted in the previous phases, are then searched in WordNet. When more than one sense is found for each term, these are disambiguated by selecting the sense with the greatest tag count value. The tag count is a value given by WordNet for each word sense and represents the frequency of that word sense in a textual corpus.

The result of the mapping process is that all the concepts of each ontology became associated to one or more concepts of WordNet. With the ontologies already mapped in WordNet, the categorization process proceeds with three different approaches, which we called of simple approach, weighted approach and expanded approach.

3.4.1 Simple Approach

The simple approach, as its name tells, is the most simple approach and represents the direct mapping between the concepts contained in a POI semantic index and the concepts associated to the ontologies. For instance, if a POI semantic index contain the concept 'Buffet' and there is a class in the ontology that is also associated to this concept, the mapping between this two concepts is taken into account. The mappings between concepts of the two structures are counted and the POI is categorized in the ontology with the greatest number of mappings.

3.4.2 Weighted Approach

The weighted approach takes advantage of the TF/IDF [18] value of each one of the concepts that are associated to POI's. The TF/IDF value represents the weight of the concept in relation to the POI it is associated to. This way, each mapping has a weight equal to the weight of the concept that originated

[3]http://gaia.fdi.ucm.es/ontologies/restaurant.owl
[4]http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs
[5]http://www.csd.abdn.ac.uk/research/AgentCities/ontologies/pubs
[6]http://protege.cim3.net/file/pub/ontologies/travel/travel.owl
[7]http://www-agentcities.doc.ic.ac.uk/ontology/shows.daml

the mapping. For instance, if the weight of the 'Buffet' concept, used in the previous example, is '0.7', the mapping between this concept and the one found in the ontology will count with a value of '0.7'.The POI is then categorized in the ontology with the greatest sum of mapping weights.

3.4.3 Expanded Approach

The expanded approach is based on the idea that the expansion of the concepts to their hyponyms make the mapping more tolerant and extensive. One may argue that when searching for *restaurants*, we are implicitly searching for every kind of restaurant, such as an *italian restaurant* or a *self-service restaurant*. Following this idea, we have applied three types of expansion:

- *Expanded POIs*: The concepts associated to POI's semantic indexes are expanded to their hyponyms and the concepts that result from this expansion are attached to the POI semantic index. For instance, a POI associated to the concept 'Restaurant', become associated to all the hyponyms of this concept, namely *italian restaurant* or *self-service restaurant*. Then, the mapping between POI's and ontologies is performed as in the simple approach.

- *Expanded Ontology*: The concepts associated to a given ontology are expanded to their hyponyms as in the previous approach an the resulting concepts are attached to the WordNet mapping of this ontolgy. As an example of this process, supposing the class *beer* will be expande to comprise several kinds of beer such as *draft beer*, *suds*, *larger* or *ale*. After this expansion, the mapping between POI's and ontologies is also performed as in the simple approach.

- *Expande Double*: The two previous approach are combined here to promote the specialization of both ontologies and POI semantic indexes.

3.5 Ontology Instantiation

At the Meaning Extraction Module, only concepts are extracted from Web pages describing places. But here, once the right Ontology has been found for a given POI, those concepts will be used to instantiate this Ontology. As the Ontology is composed not only by concepts but also of relations, the original context where concepts appear inside Web page will be used to instantiate relations between concepts. For instance, suposing a POI semantic index which contains concepts like: 'facility', 'vegetarian', 'Portuguese', 'terrace', and 'bar'. This POI is correctly categorized as a Restaurant by respectively matching of some classes in the corresponding ontology: Facility, VegetarianCuisine, PortugueseCousine, OutdoorSeatingOnTerrace and WineBarCuisine.

Once that an ontology typically doesn't have all possible instantiations of generic concepts in a given domain. This instantiation process is possible by WordNet semantic relations, which are implicit by contextualizing concepts from Web pages, and non-taxonomic relations that connect semantically related concepts. So, although the concept 'Dim Sum' from a POI semantic index doesn't appear explicitily in the Ontology, the WordNet semantic relation in 'Dim Sum' *is a type of* 'Cousine' is useful to capture which classes and type of them occur in a given POI. Another example, come from the concept 'private dinning room facility' that, besides the fact that it doesn't appear in WordNet, by Lexico-syntactic patterns [30] we are able to identify its connection with 'Facility' class.

3.6 Preliminary Results

In order to evaluate the five categorization approaches described before, we conducted some preliminary experiments with three sets of POI's, manually categorized as pubs (27 POI's), museums (15 POI's) or restaurants (17 POI's) geographically distributed as we can see in Figure 5. As it can be observed, all POI's are located on English-Native Countries, once, at this moment, we are only using lexical and semantic resources on this language. Each POI was fed to KUSKO using 2 different TF/IDF calculations on Meaning Extraction module: one considering the total of documents as only the number of related Web pages to each POI (relatively small); and other taking in account all the pages of a given category (relatively bigger). As result of Meaning Extraction phase we obtained 118 POI's semantic indexes.

We then used the five categorization approaches to categorize the POI's according to the five ontologies previously selected and mapped in WordNet. The percentages of correctly categorized POI's for each set are presented in Table 1.

Although this is a preliminary experimentation, using a total of 118 POI's semantic indexes, the results obtained reveal interesting hints. As expected, the quality of the ontologies is crucial to the results of the categorization process. In our experimentation scenario, the ontology representing the restaurant domain was clearly more detailed than that representing the pubs and museums domains. Furthermore, the museums ontology was very abstract, which decreases the probability of matching with the specific concepts associated to POI's. In part, this explains the bad results of the POI's representing museums.

Another interesting result is that the simple approach performs better than the weighted, as the POI TF/IDF (considering only the pages related to a POI) overcomes the Category TF/IDF (considering all pages for a given category) approach in most cases. This reveals that somehow the TF/IDF value used for weighting the concepts associated to the POI's is not reflecting the real weight of the concept. This can be due to the fact that of heterogeneity of our data. For instance, in our set of POI's about museums, we have different kinds of museums, with inumerous attractions and expositions, ranging from the Wolfsonian Museum in Miami to the Museum of the City of the New York. This fact can be observed in detail with we confront the previous results with other experiments considering only a extract of POI semantic index (10% of the original length) - in Table2. Another explanation to this could be the fact that we are actually computing concept frequency rather than term frequency, once this TF/IDF is calculated after WordNet disambiguation in the Meaning Extraction module. We are planning to anticipate this in a new battery of tests and then, considering only terms and not concepts yet, compute TF/IDF over raw terms as it has been done in traditional Information Extraction approaches.

To see if our categorization module performs better with other data extracted from different ways, we used the Term Extraction Yahoo API [40] to collect new POIs semantic indexes. For each Web page, we obtained from Yahoo API an index of most relevant terms, and mapped them to WordNet selecting the most frequent meaning to each term. These new semantic POI indexes were categorized using the same 5 approaches as before and the results are presented in Table 3. As we can see, the results are generally better, what we can conclude that our concept selection may be improved by using some traditional Information Extraction techniques (as TF/IDF over terms, as said before) like stemming. This reorganization of the Meaning Extraction module may produce new semantic POI indexes with more relevant terms.

Also, we can conclude that the expanded approach only performs better where the original ontology is not enough detailed. In this situation, there is an evident gain on expanding the concepts to their hyponyms, as in the case of Museums Ontology. If we examine in detail the bottom part of the hierarchy of classes in Museums Ontology, the most specific concepts in this domain are very abstract, as we can observe in Figure5, while Restaurants and BarsPubs Ontologies contain more specific concepts like:

SmookingRoom, BrailleMenu, Beer, Wine, Ale, ScothAle, Pizza, VegetarianCuisine, etc. Again, the quality and detail of the ontologies used may have a strong impact in the results obtained with this approach.

4 Future Work

This is an ongoing work and a lot of ideas are planned to be tested in the near future, some of them were extracted from the results obtained so far.

We believe that other kind of information can be associated to POIs. For instance, Named Entity Recognition (NER) techniques can be used to extract entities from the web pages that describe the POIs. These entities are not detected in the term extraction phase and associate more specific information to the POI. Also Word Sense Disambiguation on WordNet seems not to be enough specialized to find the actual meaning for each term present both in ontology and POI Web Pages. In these case we are studying to apply also other more independent techniques as Term Sense Disambiguation [27] to complement the knowledge extracted from Web Pages in order to recognize compound nouns as one entity and not only property of one base concept.

Another fact already referred is that the quality of the ontologies used in the process is crucial to the results obtained, which demands for a more carefully selection and evaluation of such ontologies. Some of the approaches developed in areas such as ontology evaluation (see section 2.3.2) and ontology selection (see section 2.3.3) may be applied, in order to guarantee the quality of the ontologies used in the system.

The techniques of ontology evaluation, discussed before (see section 2.3.2), should also be applied in the final step of the KUSCO architecture, in order to evaluate the results of the ontology population process. The evaluation process should be especially careful at the lexical, contextual and relational levels, because it is at these levels that the enrichment process will be focused on.

5 Conclusions

It is clear that, in order to improve current and future location based services, more information must be associated to common POI's. Location representation needs more human-readable information including geographic, demographic, environmental, historical and, perhaps, commercial attributes. KUSCO, the system we are developing, implements a process that we call as Semantic Enrichment of Place, which consists of using available Common Sense Ontologies and Web information to build a collection of generic and instance facts about these places. We have described the system architecture and foccused in the process of association between the Generic Place Ontologies and specific POI instances. Interesting results were obtained in a preliminary experimentation, which revealed important hints that will be used for future improvements.

References

[1] Fellbaum: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)

[2] Hightower, J.: From position to place. In: Proc. of the Workshop on Location-Aware Computing. (2003) 10–12 part of the Ubiquitous Computing Conference.

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

[4] Kramer, B.: Classification of generic places: Explorations with implications for evaluation. Journal of Environmental Psychology **15** (1995) 3–22

[5] Harrison, S., Dourish, P.: Re-place-ing space: the roles of place and space in collaborative systems. In: CSCW '96, New York, NY, USA, ACM Press (1996) 67–76

[6] Ye, J., Coyle, L., Dobson, S., Nixon, P.: A unified semantics space model. In Hightower, J., Schiele, B., Strang, T., eds.: LoCA. Volume 4718 of LNCS., Springer (2007) 103–120

[7] Kindberg., T., Barton, J., Morgan, J.e.a.: People, places, things: Web presence for the real world. In: Proc. of WMCSA2000. (2000)

[8] Vazquez, J., J., A., D.L., I.: The ubiquitous web as a model to lead our environments to their full potential. In: W3C Workshop on the Ubiquitous Web. Position paper. (2006)

[9] Bicocchi, N., Castelli, G., Mamei, M., Rosi, A., Zambonelli, F.: Supporting location-aware services for mobile users with the whereabouts diary. In: MOBILWARE '08, Brussels, Belgium, Belgium, ICST (2007) 1–6

[10] Wang, C., Lu, J., Zhang, G.: Mining key information of web pages: A method and its application. Expert Syst. Appl. **33**(2) (2007) 425–433

[11] Alani, H., Kim, S., Millard, D., Weal, M., Hall, W., Lewis, P., Shadbolt, N.: Automatic extraction of knowledge from web documents (2003)

[12] Opencyc: http://www.opencyc.org (2008)

[13] Liu, H., Singh, P.: Conceptnet: A practical commonsense reasoning toolkit (2008) Avaiable at http://citeseer.ist.psu.edu/liu04conceptnet.html.

[14] Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284** (2001) 34–43

[15] Zuniga, G.L.: Ontology: Its transformation from philosophy to information systems. In: Proceedings of the International Conference on Formal Ontology in Information Systems, ACM Press (2001) 187–197

[16] Petasis, G., Karkaletsis, V., Paliouras, G.: Ontology population and enrichment: State of the art. Public deliverable d4.3, BOEMIE Project (2007)

[17] Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press (2005)

[18] Salton, G., Wong, A., Yang, A.C.S.: A vector space model for automatic indexing. Communications of the ACM **18** (1975) 229–237

[19] Ahmad, K., Davies, A., Fulford, H., Rogers, M.: What is a term? The Semi-Automatic Extraction of Terms from Text. John Benjamins Publishing Company, Amsterdam (1994)

[20] Fortuna, B., Grobelnik, M., Mladenic, D.: Visualization of text document corpus. Informatica (Slovenia) **29**(4) (2005) 497–504

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

[22] Agirre, E., Ansa, O., Hovy, E.H., Martínez, D.: Enriching very large ontologies using the WWW. In: Proc. of OL'2000 in conjunction ECAI'2000. Volume 31., Berlin, Germany, CEUR-WS.org (2000)

[23] Jurafsky, D., Martin, J.H.: Speech and Language Processing. Prentice Hall (2000)

[24] Haase, P., Stojanovic, L.: Consistent evolution of OWL ontologies. In: Proc. of ESWC 2005. Volume 3532 of LNCS., Heraklion, Crete, Greece, Springer (2005) 182–197

[25] Lin, D., Pantel, P.: Concept discovery from text. In: Proceedings of the International Conference on Computational Linguistics (COLING). (2002) 577–583

[26] Landauer, T., Dumais, S.: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. In: Psychological Review. Volume 104/2. (1997) 211–240

[27] Klapaftis, I.P., Manandhar, S.: Term sense disambiguation for ontology learning. isda **2** (2006) 844–849

[28] Kietz, J.U., Maedche, A., Volz, R.: A method for semi-automatic ontology acquisition from a corporate intranet. In: Proc. of Workshop Ontologies and Text, co-located with EKAW'2000, Juan-Les-Pins, France (2000)

[29] Snow, R., Jurafsky, D., Ng, A.Y.: Learning syntactic patterns for automatic hypernym discovery. In: Proceedings of Advances in Neural Information Processing Systems (NIPS). (2004)

[30] Schutz, A., Buitelaar, P.: Relext: A tool for relation extraction from text in ontology extension. In: Proc. of ISWC 2005. Volume 3729 of LNCS., Galway, Ireland, Springer (2005) 593–606

[31] Lin, D., Pantel, P.: DIRT-discovery of inference rules from text. In Provost, F., Srikant, R., eds.: Proceedings of the Seventh ACM SIGKDD, New York, ACM Press (2001) 323–328

[32] Lisi, F.A.: Principles of inductive reasoning on the semantic web: A framework for learning in AL-log. In: Proc. of PPSWR 2005, Proceedings. Volume 3703 of LNCS., Dagstuhl Castle, Germany, Springer (2005) 118–132

[33] Brank, J., Grobelnik, M., Mladenic, D.: A survey of ontology evaluation techniques. Technical report, Josef Stefan Institute (2005)

[34] Sabou, M., Lopez, V., Motta, E., Uren, V.: Ontology selection: Ontology evaluation on the real semantic web. Proceedings of the Evaluation of Ontologies on the Web Workshop, held in conjunction with WWW (2006)

[35] Swoogle: Semantic web search engine in http://swoogle.umbc.edu (2008)

[36] Imagery, N., Agency, M.: Geonet names server (gns) in http://earth-info.nga.mil/gns/html/index.html (2007)

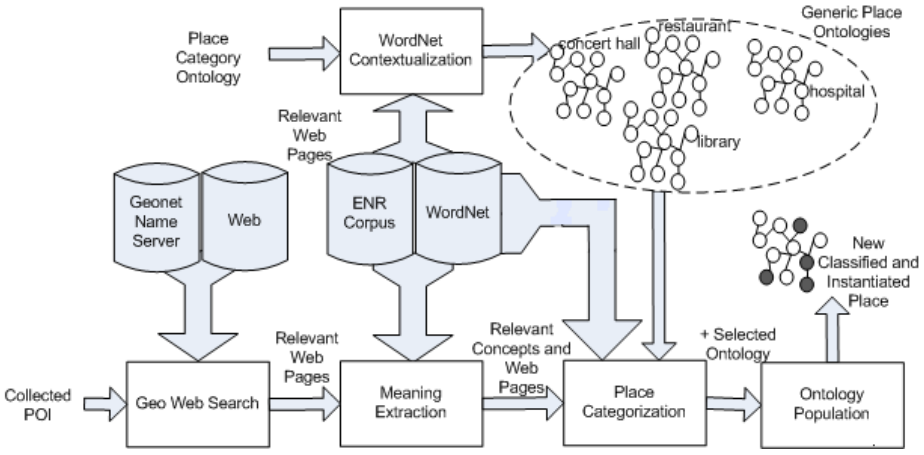

Figure 1: The architecture of KUSCO.

Table 1: Percentages of correctly categorized POI's.

	Simple	**Weighted**	**Expanded POIs**	**Expanded Ontology**	**Expanded Double**
Restaurants (POI TF/IDF)	82%	41%	88%	65%	82%
Restaurants (Category TF/IDF)	82%	12%	12%	59%	71%
Pubs (POI TF/IDF)	15%	33%	41%	70%	48%
Pubs (Category TF/IDF)	15%	22%	37%	70%	44%
Museums (POI TF/IDF)	0%	0%	13%	0%	0%
Museums (Category TF/IDF)	0%	13%	13%	0%	0%

[37] Toutanova, K., Klein, D., Manning, C.: Feature-rich part-of-speech tagging with a cyclic dependency network (2003)

[38] Patwardhan, S., Banerjee, S., Pedersen, T.: Senserelate: : Targetword-a generalized framework for word sense disambiguation. In: AAAI. (2005) 1692–1693

[39] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management **24**(5) (1988) 513–523

[40] Yahoo!: Term extraction documentation for yahoo search web services. http://developer.yahoo.com/search/content/v1/termextraction.html (2008)

Figure 2: Geographic distribution of selected POI's.

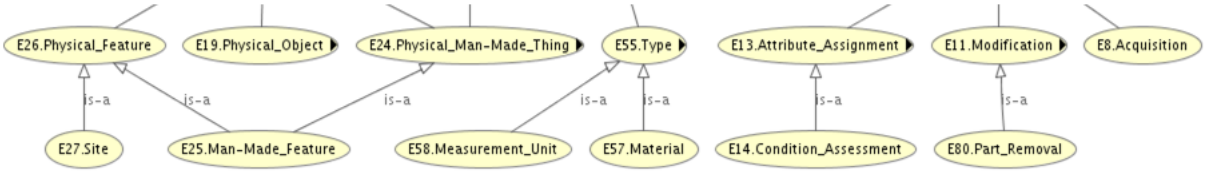

Figure 3: Most specific concepts of Museums Ontology.

Table 2: Percentages of correctly categorized museums POI's considering only a subset of most revelant concepts.

	Simple	**Weighted**	**Expanded POIs**	**Expanded Ontology**	**Expanded Double**
Museums (POI TF/IDF)	0%	0%	0%	0%	6%
Museums (Category TF/IDF)	7%	0%	13%	27%	40%

Table 3: Percentages of correctly categorized POI's considering also semantic indexes from Yahoo Term Extraction API.

	Simple	**Weighted**	**Expanded POIs**	**Expanded Ontology**	**Expanded Double**
Restaurants (POI TF/IDF)	82%	41%	88%	65%	82%
Restaurants (from Yahoo TE API)	82%	59%	94%	24%	59%
Pubs (POI TF/IDF)	15%	33%	41%	70%	48%
Pubs (from Yahoo TE API)	41%	37%	52%	59%	30%
Museums (POI TF/IDF)	0%	0%	13%	0%	0%
Museums (from Yahoo TE API)	7%	13%	7%	7%	20%