Place Enrichment by Mining the Web

Ana O. Alves[1,2], Francisco C. Pereira[2], Assaf Biderman[3] and Carlo Ratti[3]

[1] ISEC, Coimbra Institute of Engineering,Portugal
aalves@isec.pt,
[2] CISUC, University of Coimbra, Portugal
ana | camara@dei.uc.pt
[3] SENSEable City Lab, MIT, USA
abider|ratti@mit.edu

Abstract. In this paper, we address the assignment of semantics to places. The approach followed consists on leveraging from web online resources that are directly or indirectly related to places as well as from the integration with lexical and semantical frameworks such as Wordnet or Semantic Web ontologies. We argue for the wide applicability and validity of this approach to the area of Ubiquitous Computing, particularly for Context Awareness. We present our system, KUSCO, which searches for semantics associations to a given Point Of Interest (POI). Particular focus is provided to the experimentation and validation aspects.

1 Introduction

The vision of Ubiquitous Computing is rapidly becoming a reality as our environment grows increasingly replete with different sensors, widespread pervasive information, and distributed computer interfaces. New challenges still emerge, however, 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*. Significant progress has been made in Data Fusion 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 (*where*, *what*, *who*, *when*), semantics are present to various degrees. In this paper, we focus on the "where" dimension: the semantics of place.

The problem with the semantics of place has already been noted by many in the field of Ubiquitous Computing [1–3] as a valid research challenge. We stand in agreement with their perspective and aim in our work to further explore this topic with particular emphasis on methodology and short-term real-world applications. Our 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 near concepts in the tag cloud. Our task is thus to define the most accurate method to find 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 (POI). 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.

We believe that such information is highly relevant to the majority of context-aware systems. For example, navigation systems: searching for a place that has specific characteristics; route planning using locations with specific functionalities; inferring users activities, etc. We refer to this process as "Semantic Enrichment of Place". It involves using available common sense ontologies and Web information to build a collection of generic and instant facts about places.

The main contributions described in this paper are: (1) A modular methodology for the assignment of semantics to a place; (2) An implemented system, KUSCO, available online; (3) An independent validation method that compares KUSCO's performance with the Yahoo! terms Information Extraction service, currently being used in numerous projects.

This paper starts with an overview of the approaches that involve the semantics of place within several areas: ubiquitous computing, location-based Web search, and Information Extraction. Our system, KUSCO, is presented in the following section, which includes a detailed description of its processes in addition to several examples. Before finishing with a conclusion, we present our project results and validation process and raise a number of important questions.

2 Literature Review

2.1 Semantics of Place

As argued in [1], absolute position such as the pair latitude/longitude 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 derives from social conventions, their private or public nature, possibilities for communication, and many more. As argued by [2] on 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. In [4], the authors propose a semi-automatic process of tag assignment which integrates knowledge from Semantic Web ontologies and the collection of Web2.0 tags. This approach should be the theoretically correct one: it shares the formal soundness of Ontologies with the informal perspective of social networks. However it is essentially impracticable: the main choice points have to be made manually, and for

each new POI/category. The dynamics of this kind of information, particularly when depending on Web 2.0 social networks, would demand drastic resources to keep the information up to date, and the compliance to semantic standards by individual users also seems a lost battle, of which the lack of success of the Semantic Web as a global vision is the most salient expression.

On a different direction, Rattenbury et al [5] identify place and event from tags that are assigned to photos on Flickr. They exploit the regularities on tags in which regards to time and space at several scales, so when "bursts" (sudden high intensities of a given tag in space or time) are found, they become an indicator of event of meaningful place. Then, the reverse process is possible, that of 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 were also made towards analysing Flickr tags [6, 7] by applying ad-hoc approaches to determine "important" tags within a given region of time [6] or space [7] based on inter-tag frequencies. However, no determination of the properties or semantics of specific tags was provided [5].

In the Web-a-Where project, Amitay et al [8] associate web pages to geographical locations to which they are related, also identifying the main "geographical focus". The "tag enrichment" process thus 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, however the authors do not explore the idea other than using explicit geographical references. An extension could be to detect and associate patterns such as those referred above in [5] without the need for explicit location referencing.

While our work focuses on the semantic aspect of location representation, 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 acquire presence on-line, whether through the deployment of proper web sites or by referral of related institutions. This presents an opportunity for identifying the information which describes how different people and communities relate to places, and by that enrich the representation of Point Of Interest. Notwithstanding the effort of many, the Semantic Web is hardly becoming a reality, and, therefore, information is rarely structured or tagged with semantic meaning. Currently, it is widely accepted that the majority of on-line information contains unrestricted user-written text. Hence, we become dependent primarily on Information Extraction (IE) techniques for collecting and composing information on the Web, as described below.

2.2 Location-Based Web Search

Location-based web search (or *Local Search*) is one of the popular tasks expected from the 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 which the query is associated with. There are several issues for developing effective geographic search engines and, as yet, no global location-based search engine has been reported to achieve them[9]. Location ambiguity, lack of geographic information on web pages, language-based and country-dependent addressing styles, and multiple locations related to a single web resource are notable difficulties.

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. Despite this, lot of work has been done in improving the capabilities of location-based search engines [10], but it is beyond of the scope of this paper to develop them. Instead of this, we make use of generally available search engines and formulate queries using the geographical reference to retrieve information about places (section 3.1). The role of our work in this context is more on the side of contributing to the indexing capabilities of such engines in terms of local search (finding an inspiration in [11]) than on becoming any alternative form of search.

2.3 Information Extraction

The role of the Information Extraction (IE) task is to obtain meaningful knowledge out of large quantities of unstructured data. In which regards to textual information, IE is a task much linked to Text Mining, being both subtopics of the wider area of Information Retrieval (IR). IE applies classic Natural Language Processing (NLP) techniques and resources over unstructured pages written in natural language. Differently, Text Mining usually applies machine learning and pattern mining to exploit the syntactical patterns or layout structures of the template-based documents. However, 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). In fact, some important places may have their proper pages.

Nowadays, freely available common sense lexicon resources (such as WordNet, OpenCyc or others) are helpful tools to deduce semantic meaning of several concepts, including places, in that they carry relational networks that allow for new associations. Normally, these concepts and their semantic relationships are built with a generic perspective, thus rarely 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. U.S. National Library of Medicine), further exploration beyond those resources is needed to grab a more precise meaning of that place. It is at this point that Information Extraction is mandatory in the inference of the semantics of a place: the concepts that are specific to that place are the ones that best distinguish it from others.

Within IE, and excluding some of the projects referred to above, the Artequakt [12] system is the one that is closest to KUSCO. It uses natural language tools to automatically extract knowledge about artists from multiple Web Pages based on a predefined and hand-crafted ontology to generate artist biographies. The system uses a biography ontology, especially built for this purpose, which defines the data for an artist biography. Information is collected by parsing text found on the Web and is subsequently presented using templates. Diferently from our approach, 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.

3 KUSCO

When given a Point Of Interest, KUSCO [13] searches the Web for related pages by using reverse geocoding to formulate the query. Afterwards, Information Extraction is applied to those Web pages in order to extract meaningful concepts related to the intendend Place referred by the POI. We name this output as a Semantic Index since it contains concepts contextualized in two distinct types: *common concepts* for that place (e.g. smooking room, wheelchair access) which can be found on Common Sense Ontologies (e.g. WordNet) and *specific concepts* or related Named Entities (e.g. Carlsberg, Rissoto) generally as proper nouns. For example, the POI (52.212944, 0.119241, Arundel House Hotel) will trigger KUSCO to build a Semantic Index for that Place composed by Named Entities like Cambrige Guide, Conservatory Brasserie, Cambridgeshire; and some WordNet concepts like airport: "an airfield equipped with control tower and hangars as well as accommodations for passengers and cargo", comfort: "a state of being relaxed and feeling no pain", cleanliness: "diligence in keeping clean". In this paper, we describe in detail the two main processes responsible for mining the *meaning of the place*: Geo Web Search and Meaning Extraction.

3.1 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 by the City name (where the POI is located) and is obtained from Gazetteers [4] available on Web). The search is 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, Name), the final query to each search will be: <City Name> <Name>. To automatically select only pages centered on a given Place, we apply also the

[4] A geographical dictionary generally including position and geographical names like Geonet Names Server and Geographic Names Information System [14].

following heuristics to filter out unuseful Web Pages:(1) The title must contain the POI name; (2) The page body must contain an explicit reference to the POI geographical area; (3) Out of date pages will not be considered.

3.2 Meaning Extraction

Having the set of Web pages found earlier, keyword extraction and contextualization on Wordnet is made at this point. This process includes Part-of-Speech tagging, Noun Phrase chunking and Named Entity Recognition (NER) using available NLP tools [15–17]. Linguistic analisys of text typically proceeds in a layered fashion. Texts are broken up into paragraphs, paragraphs into sentences, and sentences into words. Words in a sentence are then tagged by *Part-of-Speech* (POS) *taggers* which label each word as a noun, verb, adjective, etc. *Noun Phrase chunking* is made typically by partial (sometimes called 'shallow') parsers and go beyond part-of-speech tagging to extract clusters of words that represent people or objects. 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.

Named Entity Recognition tries to identify proper names in documents and may also classify these proper names as to whether they designate people, places, companies, organizations, and the like. Unlike noun phrase extractors, many NER algorithms 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.

On completion of these subtasks for each web page, KUSCO ranks the concept either with TF-IDF [18] (Term Frequency $\times$ Inverse Document Frequency) value in order to extract the most relevant terms (only common or proper nouns) that will represent a given place. 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 [19]). Given that each word present in WordNet may have different meanings associated, its most frequent sense is selected to contextualize a given term. For example, the term "wine" has two meanings in WordNet: "fermented juice (of grapes especially)" or "a red as dark as red wine"; being the first meaning the most frequent used considering statiscs from WordNet annotated corpus (Semcor[20]).

When using data from different sources, integration of information is imperative to avoid duplicates. To solve this problem we treat differently common nouns (generally denoting concepts) from proper nouns (generally Named Entities found). Although we use WordNet to find synonyms in the first group, we don't have a list of all possible entities in the world to match words from the second group. So, we take advantage of the relatively mature field of *String metrics* to find the distance between strings using an open-source available library

ID	Title - Url		Rank
A	*Whitneys' J Byrne Bar - Restaurant New York, NY : Reviews and maps...*	http://local.yahoo.com	1
B	*Whitneys' J Byrne Bar - Restaurant, New York City, NY : Reviews of...*	http://travel.yahoo.com	3
C	*Whitneys' J Byrne Bar & Restaurant New York, NY on Yahoo! Local*	http://local.yahoo.com	4

Table 1. Most relevant pages obtained by Yahoo!.

with different algorithms implementations [21]. The importance of each concept is computed by tf weighting considering all pages related to a POI. As result of the system, each POI is represented by a list of more relevant WordNet concepts and NE terms, or, in other words, by its *Semantic Index*.

3.3 Illustrative Example

To follow the whole process of Semantic Enrichment of Places of Interest here described, we propose an illustrative example with a bar restaurant. In the beginning of the process, it is only a point in the map having a name associated:*(40.708925,-74.005111, Whitneys J Byrne Bar & Restaurant)*. The reverse geocoding gives us the city where it belongs to. So the next phase is to browse the Web using the Yahoo!Search API with the following queries in the format [City] + POI Name: *"New York"+Whitneys+J+ Byrne+Bar+Restaurant*.

For each query a set of relevant pages is retrieved and downloaded to the next phase. From 20 Web pages only 10 at most are selected following the criteria described above. Table 1 presents the Web pages selected for this POI, with the corresponding Yahoo! search ranking.

As we can see from table 1, there are sites pointing out to the same information, only differing the server where the page is hosted. In these cases, we see a possible solution by using *string metrics* between urls found. Analysing more deeply, the content is the same, only differing, in some cases, for a few characters - say, a notation showing the date and time at which the page was last modified. This widespread phenomenon on the Web, named *near duplication*, can be detected by a technique, *shingling*, that creates contiguous subsequences of tokens in a document, that can be used to gauge the similarity of two documents [22].

The Meaning Extraction module receives as input for each POI a set of N most relevant pages and extract in the first place the Named Entities as it works about raw text. For the example above, we can find the following relevant Named Entities (enclosed by brackets with their respective context excerpt in raw text):

Categories: [Steak Houses], Restaurants, American Restaurants, Irish Restaurants Reviews of New York Restaurants on Yelp.com New York City Guide > Food & Dining > Restaurants > [Steak Houses] > Whitneys' J Byrne Bar & Restaurant (site A) You Might Also Like [Nebraska Beef] 15 Stone St, New York, NY (site A)

To find common concepts, POS tagging is applied to recognize nouns and NP chunking to isolate Noun Phrases on text. For our POI in the example above, relevant concepts were found and contextualized in WordNet (table 2).

Hotels	a building where travelers can pay for lodging and meals and other services
Irish	the Celtic language of Ireland
American	a native or inhabitant of the United States
Attractions	he force by which one object attracts another
Travel	the act of going from one place to another
Deals	(often followed by 'of') a large number or amount or extent
Services	(sports) a stroke that puts the ball in play; "his powerful serves won the game"
Steaks	a slice of meat cut from the fleshy part of an animal or large fish
Reservations	a district that is reserved for particular purpose
Plan	a series of steps to be carried out or goals to be accomplished

Table 2. Wordnet meaning for each concept associated to "Whitneys J Byrne Bar Restaurant"

As a result, the semantic index produced comprises the union of Named Entity concepts and Wordnet concepts. The ordering follows Term Frequency (TF) ranking and, in this POI, it is composed by the follwoing concepts: *Hotels, Steak Houses, Irish, American, Attractions, Travel, Deals, Nebraska Beef, Services, Steaks.*

4 Avoiding Noise

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 is very high, precision is very low. We rely on statistical evidence to find very frequent words that bring little information content, as well as some heuristics to filter out insignificant words (e.g. geographical description of the place, such as the name of the city, which becomes redundant).

The list obtained at this point carries 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). 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 is obtained from an analysis of a large set of texts: we group all original texts retrieved, tokenize them to isolate words, apply a stemmer algotithm [23] to deduce the root of each word and define IDF (Inverse Document Frequency) value for each stem. We then select all words relatively commom 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 is present in this last list, it is considered a common word and not elegible to be a descriptive concept. These "special stopwords", in our case, only represent 3% of our stem list of all words processed. This can be supported by Zipf's Law [24] which states that frequency decreases very rapidly with rank.

5 Results and Experimental Evaluation

In the development process, we put an emphasis on assuring 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. Both the choice of POI types and dataset samples was 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. In other words, we present a basic scenario that demonstrates the behavior of the system in an uncontrolled environment.

The POI categories we chose to analyze were restaurants and hotels. They are described mostly in dedicated listing websites pages such as Tripadvisor.com, hotels.com, lastminute.com, and more. 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 make 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 our system creates.

A set of experimental results was obtained for over 215 POIs which were randomly selected from 4989 POIs of hotels and restaurants in the U.K., Australia and New York city. They were collected from different POI sharing websites [5] and also from Yahoo!Local search directory. We also address a few questions about the effectiveness of the Kusco System. The following sections describe our experimental evaluation of two distinct modules of the system: Geo Web Search and Meaning Extraction.

5.1 GeoWeb Search Results

For the 215 POIs, 1091 Web pages were processed by Kusco. With a great diversity of Web pages sources, 477 different domains were retrieved, most of which were directory Web sites. Following our initial queries with 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 5.1 presents statistics from different sets of pages, one group retrieved by Yahoo and the other retrieved by Google. The Yahoo results exhibit 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.

[5] Such as POIfriend.com, Pocket GPS World, GPS Data Team, POI Donwload UK.

	Yahoo	Google
Web Pages per POI (in 10 possible)	5.07	4.02
Distinct Domains	477	300
Common Web pages (from the total of each side)	48.96	73.23
POIs with common Web pages from two sides	215 (All)	

Table 3. Geo web search results from two Search APIs: Yahoo and Google.

5.2 Meaning Extraction Results

The Meaning Extraction module is expected to bring out relevant concepts and entities mentioned in the POI Web pages. As there is location-based semantics benchmark dataset to test and validate our results, we chose to use Yahoo!Term Extraction API (Yahoo!TE) [25] to examine the diversity and richness of our module. Yahoo!TE API provides a list of significant words or phrases extracted from a larger content and is currently used to create indexes over Web pages for Information Retrieval purposes.

Using the same Web pages for both systems, we need to understand *the contribution of Meaning Extraction module compared to Yahoo!TE API.* Once Yahoo!TE doesn't receive location as a parameter neither does search over the Web, we will apply the same selected Web pages (downloaded for each POI) as input to this API output. To each text, Yahoo!TE extracts most relevant terms, which are then contextualized in WordNet as KUSCO does. Considering a threshold baseline of equal value for both Extraction Meaning systems, Semantic Indexes produced by KUSCO have an average size of 35 terms (both concepts and named entities), while those built using Yahoo!TE API have 44 terms on average. For further comparison with Yahoo!TE, 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 [26]) to both semantic index lists. In this respect, KUSCO and Yahoo! have very similar results (71% and 70% respectively), with similar standard deviation (aprox. 6.0). Looking for the same perspective for Named Entities, we sought for the average TF measure for these concepts in both approaches. Here, KUSCO slightly outperforms Yahoo (59% and 50% respectively). These measures, however, mean that both systems have the same level of efficiency (with slight advantage for KUSCO) in getting valid concepts regardless of being or not significant for the meaning of the place (e.g. concepts like "New York" or "address", present in the Yahoo!TE index, get high IC value or high frequency, but they don't add novel information about the place).

6 Conclusions

In this paper, we presented an approach to the problem of "Semantics of Place". We developed a system, Kusco, which builds a semantic index associated to a given Point Of Interest (a latitude/longitude pair and a name). For each POI,

Kusco executes a sequence of Information Extraction and Natural Language Processing steps based on algorithms and tools that have been thoroughly tested. 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!Terms Extraction), Kusco has shown better results. Kusco is expected be launched as an open platform on-line in the near future.

The main contribution in this work includes a clear and well defined methodology for creating semantic information about place from web pages, a stable tool that will be available online, and a new set of benchmarking datasets (semantic indexes) that allow for future comparisons and gradual improvements on the current results and methods. Kusco is currently being applied in the context 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; a platform for data fusion with traffic and land use data.

The future steps for this system include the exploitation of structured knowledge resources (e.g. openmind, framenet, wikipedia) that can provide broader common sense semantics as well as specific information on the idiosyncrasies of each POI (e.g. Restaurants have a menu; Museums have a topic; etc.). This, we expect, should show better results than using OWL ontologies, a process we tried before with the several categories of POIs studied previously [13, 27, 28].

References

1. Hightower, J.: From position to place. In: Proc. of LOCA. (2003) 10–12 Ubicomp.
2. Harrison, S., Dourish, P.: Re-place-ing space: the roles of place and space in collaborative systems. In: Proc. of CSCW '96, New York, USA, ACM Press 67–76
3. Aipperspach, R., Rattenbury, T., Woodruff, A., Canny, J.F.: A quantitative method for revealing and comparing places in the home. In: UbiComp'06. 1–18
4. Lemmens, R., Deng, D.: Web 2.0 and semantic web: Clarifying the meaning of spatial features. In: Semantic Web meets Geopatial Applications. AGILE'08
5. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: SIGIR '07, New York, USA, ACM 103–110
6. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., Tomkins, A.: Visualizing tags over time. In: WWW '06, New York, USA, ACM 193–202
7. Jaffe, A., Naaman, M., Tassa, T., Davis, M.: Generating summaries and visualization for large collections of geo-referenced photographs. In: MIR '06. 89–98
8. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: geotagging web content. In: SIGIR '04, New York, USA, ACM 273–280
9. Asadi, S., Zhou, X., Jamali, H., Mofrad, H.: Location-based search engines tasks and capabilities: A comparative study. Webology **4** (December 2007)
10. Ahlers, D., Boll, S.: Location-based web search. In Scharl, A., Tochtermann, K., eds.: The Geospatial Web. Springer, London (2007)
11. Tanasescu, V., Domingue, J.: A differential notion of place for local search. In: LOCWEB '08, New York, USA, ACM 9–16
12. Alani, H., Kim, S., Millard, D., Weal, M., Hall, W., Lewis, P., Shadbolt, N.: Automatic extraction of knowledge from web documents (2003)

13. Alves, A., Antunes, B., Pereira, F.C., Bento, C.: Semantic enrichment of places: Ontology learning from web. Int. J. Know.-Based Intell. Eng. Syst. **13**(1) (2009) 19–30
14. GNS: Geonet names server. national imagery and mapping agency (2009)
15. Toutanova, K., Klein, D., Manning, C.: Feature-rich part-of-speech tagging with a cyclic dependency network
16. Ramshaw, L., Marcus, M.: Text Chunking using Transformation-Based Learning. In: Proc. of WVLC-1995, Cambridge, USA
17. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: ACL '05. 363–370
18. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management **24**(5) (1988) 513–523
19. Fellbaum: WordNet: An Electronic Lexical Database. MIT Press (May 1998)
20. Mihalcea, R.: Semcor semantically tagged corpus. Technical report (1998)
21. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string distance metrics for name-matching tasks. In: IJCAI-03 Works. on Information Integration. 73–78
22. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (July 2008)
23. Porter, M.: An algorithm for suffix stripping. (1997) 313–316
24. Zipf, G.K.: Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology. Addison-Wesley (1949)
25. Yahoo!: Term extraction documentation for search web: http://developer.yahoo.com/search/content/v1/termextraction.html (2009)
26. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: IJCAI. (1995) 448–453
27. Antunes, B., Alves, A., Pereira, F.C.: Semantics of place: Ontology enrichment. In: IBERAMIA. (2008) 342–351
28. Alves, A.O.: Semantically enriched places: An approach to deal with the position to place problem. In: Doctoral Colloquium of Ubicomp - Adjunct Proceedings. (2007)