Acquiring semantic context for events from online resources

João Oliveirinha
Centro de Informática e Sistemas
Universidade de Coimbra
Coimbra, Portugal
jmforte@dei.uc.pt

Francisco Pereira
Centro de Informática e Sistemas
Universidade de Coimbra
Coimbra, Portugal
camara@dei.uc.pt

Ana Alves
Centro de Informática e Sistemas
Universidade de Coimbra
Coimbra, Portugal
ana@dei.uc.pt

ABSTRACT

During the last few years, the amount of online descriptive information about places and their dynamics has reached reasonable dimension for many cities in the world. Such enriched information can now support semantic analysis of space, particularly in which respects to what *exists* there and what *happens* there.

We present a methodology to automatically label places according to events that happen there. To achieve this we use Information Extraction techniques applied to online Web 2.0 resources such as Zvents and Boston Calendar. Wikipedia is also used as a resource to semantically enrich the tag vectors initially extracted.

We describe the process by which these semantic vectors are obtained, present results of experimental analysis, and validated these with Amazon Mechanical Turk and a set of algorithms. To conclude, we discuss the strengths and weaknesses of the methodology.

Keywords

information extraction, meaning of places, events, context-aware

1. INTRODUCTION

The life of any city is full of daily rhythms, the *business as usual* behaviours of citizens, as well as unique, sometimes cyclic moments, where something different happens for some reason that changes that routine. In most of the cases, these moments are *public special events* (PSE) that gather crowds during a period of time. Typically, PSEs include music concerts, theatre shows, festivals, parades, etc. Understanding space with respect to events, or to what *happens* there, becomes a powerful context-awareness tool, useful in a wide range of situations, from location based services to sociological analysis of space. Beyond simply identifying that an event is happening with a specific title in a specific venue, it is valuable to know the type of event, its performers and their popularity, for example.

Events, particularly the largest ones, are now available online in many modern cities. Worldwide event websites such as upcoming.org or zvents, or local event pages, provide such information in a timely fashion, often plenty in advance. In this context, the ability to process or search all this data and enrich it using other online resources in acceptable time becomes a powerful asset. Indexing tools that facilitate search and processing operations on large-scale information retrieval systems are now a growing trend, particularly focusing on extraction of information from natural language texts.[18]

One of the resources available for indexing information are the *tags*, which consist of arbitrary annotations that do not follow any taxonomy or ontology, since they are assigned by Web users. Tagging - the act of adding tags to a resource - is a method that social networks apply to explain content in a more flexible and variable way, resulting in very rich but unstructured knowledge. This knowledge becomes a list of concepts, where each concept is nothing more than a cognitive unit of meaning, an abstract idea or mental symbol that can be represented as one or more words[10].

Furthermore, one of the characteristics of information provided by web resources is geo-referenced information, meaning that we have an absolute position, such as the pair longitude-latitude, introducing location-aware concepts to information retrieval systems. Taking as an example, there are events or even tweets, that provide geo-referenced information. This type of geo-referenced information introduces a new opportunity to improve location-aware services, enhancing the definition of place[12].

If we now consider that the information has a time stamp associated, a place becomes represented by a list of concepts that changes in time. We call such a list a semantic index. In this work, we determine the description of a places in a time window by exploiting the events it holds, and as a consequence describe the place itself.

We present a methodology for extracting semantic information about events, and consequently places (or venues) hosting them, from online resources like Yahoo Upcoming, Boston Calendar and Wikipedia. By extracting semantic knowledge from events, it becomes possible to have a view on the dynamic life of places through the flow of events that happen in the city. Our basic approach is to apply Information Extraction techniques to web resources in order to generate a ranked semantic index about a given event, that

LocWeb 2010, November 29, 2010 Tokyo, Japan

can be visualized as a tag cloud and can be used by a wide number of applications (from search indexing to semantic user profiling or navigation).

We make a set of experiments that are bounded in terms of space and time (Boston, from August 2009 to September 2010). The next section will be dedicated to a state of the art overview, and then we will describe the methodology that we follow. The experiments will be presented afterwards, followed by a validation model and the paper will end with a discussion and conclusion about the overall system.

2. RELATED WORK

Lemmens and Deng [4] argue that Web 2.0 and Semantic Web have complementary characteristics, and so they suggested an iterative approach of integrating Web 2.0 tags with Ontologies.

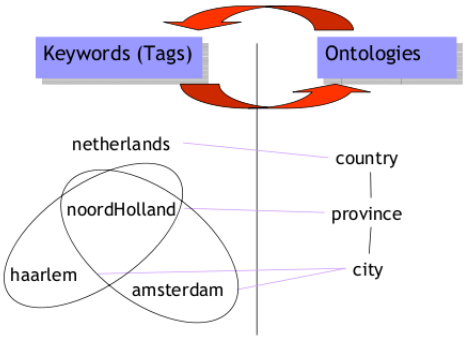

Figure 1: Web 2.0 and Ontology

This approach could be used as a semi-automatic tagging process and in fact they try to share the formal soundness of ontologies with the informal perspective of social networks which does not follow any hierarchical structure. However it is almost impossible to implement this system as the main choice points have to be made manually, and for each new POI/category. If we now consider the fact that this type of information is very dynamic, particularly when depending on Web 2.0 social networks, it would demand a set of constant up-to-date resources. In addition to this limitation, they also assume that users have the basic knowledge of semantic standards to make the corresponding match between ontology concepts and tags, which seems a little away from reality for the current days.

In 2007, Rattenbury et al [15] developed a way to detect events from the Flickr[1] photo Web Service. The idea behind it was to exploit the regularities on the tags assigned to the photos in which regards to time and space of several scales, so when several tags are found within the same small region/place, they become an indicator of event of a meaningful place (See figure 2). Then, the reverse process is possible, that of searching 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 and semantic information to their results. In second place, their approach is limited to the specific scenarios of Web 2.0 platforms that carry significant geographical reference information.

[1] http://www.flickr.com

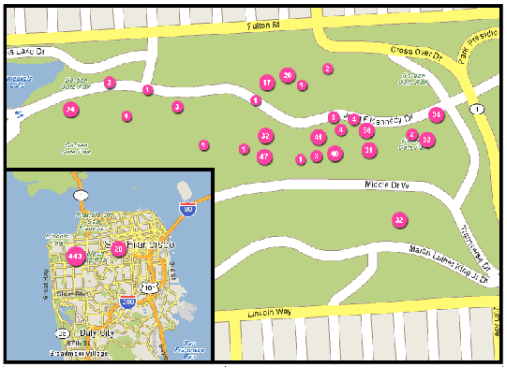

Figure 2: Flickr Tags

Similar approaches were also made towards analysing Flickr tags by applying ad-hoc approaches to determine "important" tags within a given region of time [5] or space [8] by exploiting the inter-tag frequencies. However, no determination of the properties or semantics of specific tags was provided [15].

In the Web-a-Where project, Amitay et al [3] tries to link web pages to geographical locations to which they are related. In addition they also assign to each page a geographic focus that is retrieved by the content the page discusses as a whole. Furthermore, their "tag enrichment" process consists of finding place 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 seem to be encouraging, however the authors do not explore the idea other than using explicit geographical references. An extension to this project could be added so it was capable of detecting places using other patterns like Rattenbury et al exploited, and thus without introducing the limitation of explicit geographic content. In fact, Serdyukov et al [17] recently exploited this behaviour by developing a system where pictures are placed in the map given a vector of tags associated to the image.

3. METHOLODOGY

The inference of event semantics is focused on the individual entity of a POI (Point Of Interest) and a specific time. In the next paragraphs, we summarize the process of building the semantic index of a place (i.e. a list of concepts associated to the event). It works in two major steps: event retrieving; meaning extraction.

3.1 Event selection and retrieval

The first step is responsible for finding information about events, retrieving that information from various sources using screen-scraping if no API is provided, and storing it in a database. The essential information that is extracted from the event is: event name, place name, event description, geographical address, start and end time, and if possible we also retrieve the official website of the event. Since we have the start and end time of each event we can then compute the semantic vector that classifies a place or area as a function where time is a variable.

As an additional information we also retrieve the categories of each event and venue so we can do experiments

with the data and develop validation models as explained in the next sections.

3.2 Building the semantic index

The next step, of meaning extraction, starts with keyword extraction. Our system, Kusco[1] mines these event descriptions and extracts relevant terms related to those events. This is achieved in a pipeline fashion with Part-of-Speech (POS) tagging [19], Noun Phrase chunking [14] and Named Entity Recognition (NER) [9]. POS taggers label each word as a noun, verb, adjective, etc. Then, individual noun phrases are inferred with *Noun Phrase chunking*, which concentrates on identifying *base* noun phrases, which consist of a *head* noun and its *left modifiers* (e.g. Mexican food). Finally, 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 cues, 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 event description, KUSCO ranks the concept with Term-Frequency [16] that will represent a given event. These nouns are contextualized on WordNet and thus can be seen not only as words but more cognitively as a concept (specifically a synset - family of words having the same meaning, i.e., synonyms [6]). 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 statistics from WordNet annotated corpus (Semcor[11]). It is important to notice that presently the system only deals with English descriptions, as all NLP resources used by this module are prepared to process this language.

The list obtained at this point carries however large quantities of *noise*, which corresponds to words that do not add new information 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 algorithm [13] to deduce the root of each word and define IDF (Inverse Document Frequency) value for each stem. We then select all words relatively common 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 is not eligible 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 [20] which states that frequency decreases very rapidly with rank. In the end, each event is represented by a list of the more relevant WordNet concepts and NE terms, or, in other words, by its *Semantic Index*. The final TF-IDF of each concept in each semantic index can then be calculated using as a corpus for the IDF based on the entire event database.

3.3 Enriching the index with Wikipedia

Being one of the paradigmatic examples of Web2.0 in practice, the Wikipedia relies on individual contributions from users of the entire world to build an "open source" encyclopedia. From the perspective of the information on place, and specifically for the application of Kusco Information Extraction, Wikipedia pages provide a fix structure with an initial abstract followed by a table of contents, the detailed content (which can vary considerably among pages), and then a set of references and external links. The abstract, for it is a summarization of the concept, catches the main highlights of each concept and it is therefore the perfect candidate for mining.

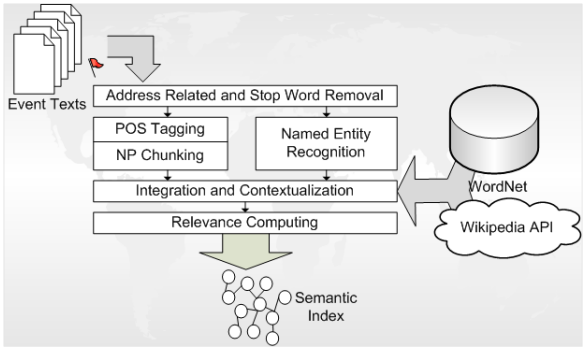

Figure 3: Kusco Architecture

After the extraction of initial concepts (either from open web or upcoming.org) using Kusco, we *enrich* each one of them by searching for the relevant page in Wikipedia and applying again Kusco to the union of all found abstracts using this process. It then extracts the concepts and ranks them according to Term Frequency.

The system could retrieve a number of more distant, yet potentially relevant, associations. For each of the events (available in our database), we pick the top 5 words from the description and then perform the aggregation of wikipedia abstracts for those words. Then, we re-rank again, thus obtaining the top-5 words listed (in order of relevance). Figure 3 shows the final architecture of the semantic vectors extraction module.

3.4 Venue semantic index

Since we have a list of all events in a specific place, the *venue*, we build a semantic index that aggregates all events in a single list. We process the events in three steps: first, we build the individual event indexes and compute the new TF-IDF value for each concept based on this list of event as the corpus; then, we select the top five concepts and retrieve the wikipedia abstracts of each concept; as a final step, we feed Kusko with the abstracts of the articles and retrieve the final list of concepts ranked in TF-IDF.

4. EXPERIMENTS

In this section we present some examples of the results obtained using our methodology applied to events from Boston. The events here presented are a small subset of our database that currently has 148303 events hosted in 11197 different venues in Boston and were extracted from the Boston Calendar and Zvents sources in a time window starting in August 25th 2009 to September 20th 2010. The average and standard deviation of events per venue is 11.64 and 58.04, respectively. Our database presently contains 189290 distinct concepts, of which 18851 are linked to Wordnet synsets and 24046 are linked to Wikipedia articles.

The table 2 contains a sample of the results obtained for each event and table 1 contains the descriptions that were used to extract the semantic vectors for those events. The column Concepts has the semantic index after the first iteration of Kusco, and the next column has the new semantic index after the enrichment stage is applied.

By looking at the tables, it is possible to recognize that the system was capable of extracting new concepts that were not available in the original description. This happens because we added semantic knowledge by processing the Wikipedia summaries for each concept of the first semantic index.

On the other hand, the last four events have very poor results which are justified by the lack of information related to the event in their descriptions. It is also possible to notice that some words are noise introduced by the user description or event of a faulty screen scraping. In other times, this happens because the description is so small that the term frequency of all terms is nearly constant, thus the ranking is not efficient.

Another important aspect that is extracted from table 2 is the small subset of concepts that are a compound of multiple words. This is a consequence of the weight system (TF-IDF) because these type of concepts tend to have a higher *inverse document frequency* (IDF) value but in contrast they have a very small *term frequency* (TF), as such, the final TF-IDF value is not high enougth to appear in the *top 5* list. In fact, this is the best approach because it makes possible to filter out most of the concepts that are very specific (eg. persons) and that do not add rich semantic information to the final vector.

Table 3 introduces the semantic indexes of four places that were computed by merging the semantic indexes of events held in the same place. As it is possible to notice, the more documents/events used to calculate the result, the better the definition will be. The first place in the table is a point of interest extracted from Boston Calendar that is used to represent the city and some events without a specific hosting place. This can explain the generality of the resulting tags.

ID	Description
A	This half-mile trail is part of a 32-acre cranberry wetland system and wooded area with a bridge and observation platform stretching across a six-acre pond. - June Wulff, Globe Staff
B	The Salem Farmers Market is a tradition that dates back to 1634. With it's peak around 1930, The City of Salem is now renewing its tradition of the Salem Farmers Market in downtown Salem, MA. Opening day: June 25, 2009, the 375th anniversary of the birth of the Salem Farmers Market and the rebirth of a Salem tradition. The Salem Market works to Provide a convenient and congenial means of purchasing locally grown or prepared food products and Support local agriculture and producers.
C	The Cadillac La Salle Club is going to be making their 8th annual appearance at Ray Ciccolos Cadillac Village of Norwood. The club is expected to bring over seventy antique Cadillacs from the 1920's up through the 70's. This is event is FREE and open to the public, and will feature a DJ and free refreshments
D	The paintings, sculptures, drawings, photographs, and manuscripts in this summer installation draw from the collections of the Boston Athenium and add to the wealth of objects always on public view on the Athenium first floor. Over 40 artists are represented, ranging from Italian and Scottish to American and from the 16th to the 21st century. The objects on view are as varied in style as in subject matter, and include: a portrait by the 16th-century
E	EPOCH Senior Healthcare of Chestnut Hill and EPOCH Assisted Living at Boylston Place, will be collecting non-perishable food items in their lobbies throughout the month of August. EPOCH will donate the food collected to the Brookline Food Pantry. Contact Mary Rivera at 617-243-9990 for more information.
F	10:00am - Noon: Volunteer 1:00 - 4:00pm: Games, tree climbing, and family nature walks in the woods 12:00 - 2:00 pm: Landscape watercolor painting workshop - all materials provided / on Schoolmaster Hill ***Meet at the Resting Place / Shattuck Picnic Grove Bus Route #16 from Forest Hills (Behind Shattuck Hospital across from Forest Hills Cemetery)
H	Come early and see Army Blackhawk Helicopter, Humvees, Playstation on Jumbotron and more. A benefit by cops for kids with cancer.
I	Lexington Farmers Market, corner of Massachusetts Ave, Woburn St., and Fletcher Ave. in Lexington Center. Tuesdays, June 9 through October 27, 2009, 2-6:30 p.m., rain or shine. Features locally grown produce, a variety of meats, fish, baked goods and other prepared foods, and artisans tent. Admission free. For more information, and to subscribe to the weekly newsletter, visit www.lexingtonfarmersmarket.org

Table 1: Examples of events

5. VALIDATION

5.1 Category persistence

This research project faces an important challenge of understanding the actual quality of the results in terms of the correctness of the words assigned to places. The list of words that best describes a place is by nature subjective, because a place can be defined according to different perspectives, and each perspective can vary with subject. In terms of validation, this raises difficult questions even for the typical user survey. The only way to guarantee a good validation using

ID	Concepts	Wiki Concepts
A	Globe, system, plataform, trail, pond	Pond, Falls, streams, currents, winter
B	Farmers, birth, rebirth, anniversary, agriculture	cultures,consumption, carbohydrate, Food safety, gastronomy
C	Cadillac, Norwood, apperance, refreshments, Cadillac La Salle Club	Cadillac, Michigan, Automobile, General Motors Company, vehicles
D	objects, Boston Atheneum, portrait, sculptures, collections	Paintings, Eastern, scenes, Sistine Chapel, Mona Lisa
E	Boylston, Chestnut, Healthcare, items, lobbies	Monoclonal, Surgery, Medicine,Dentistry, health systems
F	Forest Hils Cemetery, Volunteer, Games, Noon, materials	Trees, Collins, plants, Macmillan, Sequoia sempervirens
H	Jumbotron, kids, Playstation, benefit, cancer	Cancer, cells, abnormalities, neoplasm, treatment
I	meats, Massachusetts, Fletcher, tent, Lexington	tent, camping, shelter, rope, poles

Table 2: Boston Events Top Five Concepts

Name	Concepts	Num Docs
City of Boston	traffic, boston, competition, intersection, lanes, vehicle, rivals, frredom trail	11
Tremont Temple Baptist Church	bible, category judaism, tanahk[2], prayer, language, christians, meditation	21
New England Aquarium	aquaria, presentation, animals, fur seals, turtle	135
MIT[3]	community, dance, massachusetts, questions, students, seminar, lecture, university, skills	270

Table 3: Boston Places

ID	Cat[1]	Cat[2]	k
730VB	University	Theater	1
766VB	School	University	4
768VB	Nightclub	Club	1
813VB	University	University	1
813VB	Non-profit	University	7
820VA	Library	BookStore	1
817VA	Dance Hall	Community Center	1
965VA	Theater	Cultural center	1
809EA	Music	Rap/Hip-Hop	1
1225EA	Jazz	Jazz	1

Table 4: kNN Category classification results.

human resources is by using a large sample of people that they know all the places, which then becomes unpractical.

Thus, we first decided to analyze our results according to category consistency. Each POI has one or more category, so, the task is to verify the persistence of the word patterns according to those categories. To achieve these results we have implemented the following approach.

As could be seen from the resulting semantic vectors, an accurate comparison of indexes needs to take into account semantic distance as opposed to simple string matching. In other words, for example the concept "jazz music" is closer to "classical music" than to "football". We developed a method that takes into account the semantic indexes and the weight of each concept in order to be able to do reasonable evaluation of similar documents and classify with a closer category or topic. One of the best algorithms for this type of classification is the K Nearest Neighbor if it is well adapted with a weight system.

In fact, Eui-Hong Han et al argued that a well adapted Weight Adjusted Nearest Neighbor Classification algorithm can outperform other algorithms, such as C4.5, RIPPER, Naive-Bayesian, PEBLS and VSM[7] in tasks that depend on semantic similarity. The reason for these results is because this algorithm finds the k documents that are closer to the document to classify and those k categories of the documents "vote" for the category of the new document.

Our method for category persistance analysis consists in applying the *kNN* algorithm with *cosine distance*[16] as a document similarity measure to determine the "distance" between semantic vectors. Table 4 shows some examples of the results obtained with this method for the different number of k values, which represents the number of neighbours allowed to vote in the kNN algorthim.

Since we use many variables (eg. number of neighbours, similarity functions, distance threshold), we can not make a real estimation of the positive and negatives match in the classification process without defining a fixed value to the majority of each variable.

And if we do this we have no guarantee of correctness because we can not assume that, for example, two documents/events are similar if the *cosine distance* is lower than 0.2. If we assume a fixed threshold, then we needed to validate those results with the help of volunteers. In addition, we also do not know what is the best number of neighbours to be used in the algorithm. In conclusion, we can say that this methodology of classification has proven to be correct by the use of these examples, where we can analyse that the real category and the calculated one are semantically close. In fact, in some cases we can even get a more precise category match than the manually assigned one. For instance, the example with id 730VB, which was manually assigned with the Theater category, is in fact an University (Brandeis University). But on the other hand, it can be noticeable some problems when the number of *neighbours* allowed to vote is too small or too large. For instance, 813VB and 809EA are good examples of poor categories matches caused by a

[1]Category obtained after running classified via kNN.
[2]Real category extracted from the source
[3]Document similarity function used

Options	Before	After
Very Relevant	43%	8%
Relevant	46%	40.9%
Not Relevant	10%	50.5%

Table 5: Statistical Results - Events Batch 1

Options	Before	After
Very Relevant	38%	33%
Relevant	46%	53%
Not Relevant	11.5%	14%

Table 6: Statistical Results - Venues Batch 1

large and small value of *kNN* neighbours used, respectively. This happens because as the value of k increase, more distant neighbours, which are semantic vectors that represents events, will be able to vote for the result category. As such, this tends to lead to a result category that is more distant in a semantic way from the real category.

5.2 MTurk

As in many tasks that involve Natural Language Processing, it becomes practically impossible to validate results for their *universal and absolute correctness* due to ambiguity of language and subjectivity of the task itself. This makes it even more mandatory to ask a large number of subjects for their opinions on the task at hand: how well does a semantic index describe an event. We use the Amazon Mechanical Turk (MTurk) which makes possible to publish our output (semantic indexes) in their servers and have multiple persons from multiple places in the world validate our system in turn of a small cost. Extreme care must be taken to identify users that randomly fill surveys ("spammers") or that simply are inconsistent in their responses. Amazon does a good work by providing tools that help in filtering these cases. Some of the properties about the user that we can choose are the positive feedback required to participate in the survey and the country. This positive feedback value is voted by other survey owners and is used to determine the quality of the user answers, thus enabling the spam to be filtered out. We also make the same question to different users, to achieve a *quorum*. We also only let users with a 95% positive feedback participate and we choose to make the same questions to 3 distinct users, so in the final analysis we can select the best of 3 responses for each event by choosing the one that both users agreed.

We ran two types of batches in MTurk, and for each type we execute one batch for events and another for venues, making a total of 4 batches. In the first 2 batches we provided the user with the following data: an event/place name, an event/place description, the official website, if provided, and 2 lists of concepts (semantic indexes), one before wikipedia enrichment and another after it. Each semantic index is composed of the top five concepts ranked by the TF-IDF weighting system. With this information, the user was asked to classify the relevance of each semantic vector with 3 levels: *Not Relevant*, *Relevant* and *Very Relevant*. In both batches we provided 960 events and 200 venues to be validated by 69 and 19 distinct users, respectively.

Table 5 shows the results that we obtained from MTurk. As we can see, the results obtained before the enrichment with wikipedia are relatively good considering that only 10% are classified as not relevant to the event. However, after the enrichment the values dropped significantly. This testifies the risky game of semantic enrichment: adding external concepts can also introduce noise. In Table 7, we can see that when we applied the enrichment process, 10.7% of those semantic indexes have improved, 64.6% degraded and 24.5% maintained the relevance score. These results can be explained by the fact that the events were ranked by TF-IDF using as corpus the whole database of events. What we can learn from this is that this approach of using all the events as the corpus introduces noise to the semantic vectors because events seem to have a higher correlation with other closer events in terms of space. This corroborates the results obtained for the venues because the corpus used was only a subset of events (events held in the venue).

	Events	Venues
Improved	10.7%	26.5%
Degraded	64.6%	29.5%
Mantain	24.5%	44.5%

Table 7: Statistical Results - Batch 1

Relevance	VR	R	NR
Concept 1 BE	66.5%	28.6%	4.7%
Concept 2 BE	65.7%	26.3%	7.9%
Concept 3 BE	61.2%	26.4%	12.2%
Concept 1 AE	35.9%	26.1%	37.9%
Concept 2 AE	34.1%	24.1%	41.6%
Concept 3 AE	22.5%	33.8%	43.6%

Table 8: Statistical Results - Events Batch 2

The second batch that we ran in MTurk was with the same data, but instead of asking the user to classify the relevance of each semantic index, we asked them to classify the relevance of each concept in the semantic index. Validating the entire index at once proved very difficult for users since often only a subset of the concepts was relevant and they had to decide the overall veredict.

In this batch the number of concepts to be validated from each semantic index decreased from 5 to 3 mainly because of the combinatoric explosion of different questions. The number of users who validated this batch was 103 for the events batch and 24 for the venues batch.

In the tables 8 and 9, we can see the results obtained for the second batch for events and venues respectively. Each table is composed by a first row that contains the levels of relevance: *Very Relevant* (VR), *Relevant* (R) and *Not Relevant* (NR); and the columns refer to the first 3 concepts of the semantic index before enrichment (BE), followed by the list of three concepts after enrichment (AE).

These results confirm our suspicions that volunteers were

Relevance	VR	R	NR
Concept 1 BE	36%	58.5%	5.5%
Concept 2 BE	59%	37%	4%
Concept 3 BE	66%	30%	4%
Concept 1 AE	67.5%	28%	4.5%
Concept 2 AE	61.5%	32%	6.5%
Concept 3 AE	36%	54.5%	9.5%

Table 9: Statistical Results - Venues Batch 2

having doubts on how to classify the semantic index in the first batch if one bad (poor semantic information or noise) concept appeared. We can see that on the events and venues batch only a small subset of concepts were classified as *Not Relevant*. However, we also have to take into account that we dropped the last two concepts from the semantic index and that may have some influence in results. Another important aspect is that, on the venues batch, the process of enrichment has very good results: the concept that take the first position makes an improvement from almost 30% to 70%, but this tend to decrease for the following concepts which makes sense because of the TF-IDF ranking system. On the other hand, the events batch still present us with low results, mainly because of the same reason explained in the first batch, which is the use of all events for the corpus that is used to compute the IDF value.

6. DISCUSSION

Although the results show that the system can extract relevant concepts, it is also noticeable that some amount of noise will always be present. The best way to prevent this is to implement a stopword filter that reduces considerably this noise, but the ideal threshold has to be carefully negotiated in order to minimize false negatives/positives. We also want to remove the concepts that do not bring much semantic information to the semantic index. For instance, the concept Boston is not very important to appear in our semantic index if we know we are exploring events from Boston.

From the empirical analysis of the experiments, the results obtained from calculating the semantic indexes for the venues have the most quality. This happens because the corpus used to calculate the TF-IDF value for each concept is the number of documents/events that are held in that venue, instead of the whole corpus of the systems being used as in the case of calculating the semantic index of events.

Other than what was already applied in the experiments, the validation of these results is extremely difficult. Knowing the "correct" set of words for each POI/event is *per se* an ambiguous task. Furthermore, even for making a voluntary survey, the range of possible choices is enormous (which POIs to choose, which filter threshold, which perspectives, with and without wikipedia) becoming a potential demanding effort to reply. The other option is to make small sets of questions but aiming for a larger sample of respondents.

Another limitation is the performance of the system. The major bottleneck is on the side of Kusko because some parts are extremely slow such as the NER algorithm. We improved this from earlier versions by exporting Kusko as a Web Service so we do not need to load the system every time we extract a semantic index from a document. The other drawback in the performance is the web searching and screen scrapping in Wikipedia, but we implemented our own Wikipedia cache, improving the speed considerably. Using these approaches we managed to improve the time of processing a single event from a maximum of 55 seconds to 5 seconds. This still is a large amount of time, not surprising when dealing with unstructured Natural Language texts, but it raises the obvious question of scalability. Besides the code optimization, for large scale application, a careful choice on the *perspective*, coverage, constraints on the input has to be balanced against precision.

In previous versions of Kusco, which stands for *Knowledge discovery via Unsupervised Search from web to instantiate Common sense Ontologies*, we used Semantic Web ontologies in the process, namely Restaurant, Hotel and Museum related. However, as explained in [2], these ontologies were extremely poor in terms of domain knowledge and the results were weak.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a methodology for the extraction of semantics of places and events from online resources. We enrich extracted tags using the Wikipedia that results in a richer and diverse tag vectors of the first resources. In practical terms, we retrieve the semantic index that better describes an event or place. These tags or semantic index can be useful for various applications, namely, POI search, context-aware applications on ubiquitous systems, automatic advertising.

Results also show that despite the problems and difficulties inherent to systems using Natural Language Processing for unstructured texts, it is possible to obtain meaningful descriptions of place from dynamic web sources.

A challenging aspect is the validation of the methodology. We used automatic/objective methodologies based on machine learning classification algorithms to understand the persistence of words patterns with respect to category. We modified the similarity measure of those algorithms to consider semantic distance between different concepts and the results show that the semantic indexes extracted have a relevant degree of persistence among categories that is not due to chance.

More importantly, we ran online surveys with Amazon Mechanical Turk that showed encouraging results, particularly with respect to pre-enrichment results and enriched information about the venues.

As for future work, there are some ideas that can be explored in order to improve our methodology and system that are related to a better way to integrate perspectives/sources. It may be possible to set a weight to each perspective so we can define which perspective or source is more important. After this feature is implemented we can try to improve it to adapt the weights dynamically.

Other idea to explore is the semantic information that can be retrieved from the links and references of each Wikipedia article. The main idea is to extract a graph where articles are connected by shared concepts. With this information we can know how each concept relates to each other and possibly find new patterns between documents/events, or even improve the semantic index by inferring new concepts. Finally, another idea is to classify a place based on its events dimensionality. That is, trying to infer other concepts from bursts/patterns of regular events. For instance, a stadium that hosts different types of sports depending on the season.

8. REFERENCES

[1] A. Alves, F. C. Pereira, A. Biderman, and C. Ratti. Place enrichment by mining the web. In *Proc. of the Third European Conference on Ambient Intelligence*, 2009.

[2] A. O. Alves, B. Antunes, F. C. Pereira, and C. Bento. Semantic enrichment of places: Ontology learning from web. *International Journal of Knowledge-Based and Intelligent Engineering Systems (IOS Press)*, 2009.

[3] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280, New York, NY, USA, 2004. ACM.

[4] D. Deng and R. Lemmens. Web 2.0 and semantic web: Clarifying the meaning of spatial features, 2008.

[5] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 193–202, New York, NY, USA, 2006. ACM.

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

[7] E.-H. S. Han, G. Karypis, and V. Kumar. Text categorization using weight adjustedk-nearest neighbor classification. In *Advances in Knowledge Discovery and Data Mining*, volume 2035 of *Lecture Notes in Computer Science*, pages 53–65. Springer Berlin / Heidelberg, 2001.

[8] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 89–98, New York, NY, USA, 2006. ACM.

[9] V. Krishnan and C. D. Manning. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 1121–1128, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[10] E. Margolis and S. L. A. O. or Mental Representations? The ontology of concepts.

[11] R. Mihalcea. Semcor semantically tagged corpus. Technical report, CiteSeerX - Scientific Literature Digital Library and Search Engine [http://citeseerx.ist.psu.edu/oai2] (United States), 1998.

[12] F. C. Pereira, A. Alves, J. Oliveirinha, and A. Biderman. Perspectives on semantics of the place from online resources. *International Conference on Semantic Computing*, 0:215–220, 2009.

[13] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[14] L. Ramshaw and M. Marcus. Text Chunking using Transformation-Based Learning. In *Proceedings of the 3rd Workshop on Very Large Corpora: WVLC-1995*, Cambridge, USA, 1995.

[15] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110, New York, NY, USA, 2007. ACM.

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

[17] P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491, New York, NY, USA, 2009. ACM.

[18] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 327–336, New York, NY, USA, 2008. ACM.

[19] K. Toutanova, D. Klein, and C. Manning. Feature-rich part-of-speech tagging with a cyclic dependency network.

[20] G. K. Zipf. *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology*. Addison-Wesley, 1949.