Perspectives on semantics of the place from online resources

Francisco Pereira, Ana Alves, João Oliveirinha
Centro de Informatica e Sistemas da Universidade de Coimbra
CISUC
Coimbra, Portugal
{camara,ana,jmforte}@dei.uc.pt

Assaf Biderman
SENSEable City Lab
Massachusetts Institute of Technology (MIT)
Cambridge, MA, USA
abider@mit.edu

***Abstract*—We present a methodology for extraction of semantic indexes related to a given geo-referenced place. These lists of words correspond to the concepts that should be semantically related to that place, according to a number of perspectives. Each *perspective* is provided by a different online resource, namely upcoming.org, Flickr, Wikipedia or open web search (using Yahoo! search engine). We describe the process by which those lists are obtained, present experimental results and discuss the strengths and weaknesses of the methodology and of each perspective.**

***Index Terms*—Information extraction; Semantics of place; Automatic tagging.**

I. Introduction

The concept of place is one with recurrent inconstancy throughout history, culture and communication. It is noticeable that a place can be described or referenced according to many perspectives, dependent on what is intended to be communicated (e.g. its function, its physical properties, its content, its relationship with the subject). With the growing number of geo-referenced data available on the Web, an increasing number of Location Based Services (LBS) intensifies the spotlight that focuses on the definition of place. The potential of success of such services is highly dependent on how they perceive the context, particularly the place. For example, a context-aware system should be able to adapt according to the place in which the user is (e.g. work, home, cinema, shopping), a simple service for a smartphone would be to detect the availability of the user according to place, or maybe change the interface (e.g. a "a different skin", a different set of applications, ring volume), but this can only be achieved via the perception of place. Such perception is not trivial for the reasons just explained: place has inherently many dimensions associated.

In this paper, we present a methodology for extracting semantic information about place from online resources. Also from the point of view of each resource, a different perspective on place can be found. In Wikipedia, places are described with historical and geographical viewpoints; in Flickr, places become aggregations of pictures and tags; in Upcoming.org, the dynamic life of places become apparent through the flow of events that happen in the city; using regular web search, the possibilities are immense, as well as the resultant ambiguity, but often places are described thoroughly in their specific homepages.

Our basic approach is to apply Information Extraction techniques to web resources in order to generate a ranked semantic index associated to a place, this can be visualized as a tag cloud and can be used for a wide number of applications (from search indexing to semantic user profiling or navigation) [1]. In this paper, we propose the application of such approach to several different resources and integrate them in a multi-faceted view of place. We make a set of experiments that are bounded in terms of space and time (Lower Manhattan, New York; August 2007 to August 2008).

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 and the paper will end with a discussion and conclusion.

II. Related Work

Rattenbury et al [2] 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.

The World Explorer project [3] dedicates to organize the most important tags from Flickr around an area (with different scales, from street to world level). The choice of "representative tags" is made in a sequential manner that starts with clustering images (with K-means), then it makes a tf-idf analysis over the set of tags for each cluster, which then ranks those words in that cluster. The authors provide the use of their database via the *tagmap* API, and a detailed look on that that shows strong prominence for geo-graphically related information (e.g. name of the place, of the tourist attraction, etc.), which although correct becomes redundant in terms of the semantics of that place. A richer representation will include not only such information but also other associative knowledge about that place (e.g. if it is a theatre, what kinds of plays or actors are usually there?). Such information is partially present in Flickr, but it becomes hidden by the author's ranking method.

Also based on Flickr images, [4] integrate analysis on image, text and time. They apply a *mean shift* analysis to determine relevant landmarks and images (i.e. a kind of cluster centroid) and then associate these to the most salient words in statistical terms. They also show a classification experiment in which they estimate image geographical location based on visual cues, text tags and both. The combined approach outperforms considerably the isolated ones. However, in this project, there is no semantic analysis (e.g. semantic distance) between words, stemming, synonymy check, enrichment with external resources, or even basic filtering. In other words, the semantic tags are taken as atomic strings.

Other attempts were also made towards analysing Flickr tags [5], [6], which applied ad-hoc approaches to determine "important" tags within a given region of time [5] or space [6] based on inter-tag frequencies. However, no determination of the properties or semantics of specific tags was provided [2].

In the Web-a-Where project, Amitay et al [7] 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 further than strictly geographical meaning. An extension could be to detect and associate patterns such as those referred above in [2] without the need for explicit location referencing.

III. Methodology

A. Kusco basic approach - the open web perspective

The inference of place semantics is focused on the individual entity of a POI (Point Of Interest). Our system, Kusco mines the web for relevant terms related to that POI. In the next paragraphs, we summarize the process of building of the semantic index of a place (i.e. a list of words associated to the place). It works in two major steps: GeoWeb search; Meaning extraction.

GeoWeb search is responsible for finding Web pages using only POI data as keywords: place name and geographical address. This last element is composed of the address (where the POI is located) and is obtained from Gazetteers available on Web [8]. The search is made by the freely available Yahoo!Search API. The system applies a heuristic that uses the geographical reference as another keyword in the search. Thus, assuming a POI is a quadruple (Latitude, Longitude, [Category,] Name)[1], the final query to each search will be: <City Name> <Name> + [<Category>] . To automatically select only pages centered on a given Place, we apply also the following heuristic to filter out unuseful Web Pages:

1) *The title must contain the POI name.* For example, looking for Web pages about the POI named “Almonds Bistro” in Carlisle, UK using the Yahoo!Search API, we have found the url http://www.idorla.com/carlisle-accommodation-cumbria-in-england.html as a relevant page. But if we examine the title of the page “IDORLA - Carlisle bed & breakfast accommodation directory” we can conclude that it is not specific about that place, but instead a general directory Web site of Services in the Carlisle area.
2) *The page body must contain an explicit reference to the POI geographical area.* We think geo reference in the query is crucial to retrieve proper Web pages about a given POI. For example, if we search the Web only using the name of a Hospital in NY City named “Mount Sinai”, 7 of the Top-10 pages retrieved by Yahoo are related to the Biblic place in Egypt.
3) *Out of date pages will not be considered.* Some Web pages may be outdated due to varied factors, including changes in the characteristics of the place. In those cases, despite the funcionality that Yahoo API can retrieve cached Web pages, these pages are not processed by KUSCO.

The next step, of meaning extraction, starts with keyword extraction which is achieved in a pipeline fashion with Part-of-Speech (POS) tagging [9], Noun Phrase chunking [10] and Named Entity Recognition (NER) [11]. 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 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 with TF-IDF [12] (Term Frequency $\times$ Inverse Document Frequency)[2] that will represent a given place. These nouns are contextualized on WordNet and thus can be seen not only as a word but more cognitively as a concept (specifically a synset - family of words having the same meaning, i.e., synonyms [13]). 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[14]). 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 algotithm [15] 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 [16] which states that frequency decreases very rapidly with rank. In the end, each POI is represented by a list of the more relevant WordNet concepts and NE terms, or, in other words, by its *Semantic Index*.

We show now an example of a POI in New York (-74.0028, 40.7171, “Gigantic Artspace”), which as an official website (http://www.giganticartspace.com/). Kusco retrieved from Open Web search the web sites listed in table I. Then it extracts the concepts and applies the filter. Table refgiganticConcepts shows the obtained list.

The first concept, “gas”, is the acronym of “Gigantic ArtSpace”, while the several names (“Mari Kimura”, “Eric Singer”, etc.) correspond to artist names. Open web extraction is thus extremely dependent on external factors (correct ranking of the appropriate pages, quantity of noise in those pages, depth of description of the

[1]Category refers to the type of POI in question, a museum, a restaurant, a pub, etc. This information is optional, and often not present in the POI information.

[2]TF measures the frequency of the word in the observed document, IDF identifies the frequency of the word among the universe of documents - common words get a low IDF.

http://www.16beavergroup.org/events/archives/000830.php
http://www.giganticartspace.com/
http://upcoming.yahoo.com/event/3909/
http://calendar.artcat.com/event/view/8/3965
http://music.columbia.edu/pipermail/art+tech/2004-September/000107.htm
http://rhizome.org/discuss/view/12756
http://www.galleryartist.com/giganticartspace/
http://05.performa-arts.org/venues/gigantic-artspace
http://www.photography-now.com/institutions/I7335022.html

TABLE I
LIST OF RETRIEVED WEBSITES FOR "GIGANTIC ARTSPACE".

Concept	Score	Wordnet gloss
e-archive	2.559	
gas	1.704	a fluid in the gaseous state having neither independent shape nor volume and being able to expand indefinitely
inquiries	1.033	a search for knowledge
Foreign Legion	0.698	
Bil Bowen	0.591	
Mari Kimura	0.537	
Eric Singer	0.537	
galleryartist	0.512	

TABLE II
CONCEPTS FOR "GIGANTIC ARTSPACE" (WITH SCORE ABOVE 0.5)

place) and represents what we call the "hardest scenario" for the extraction of semantics. Here we assume it as a *static* perspective, but in reality that is dependent on the actual pages (in the example above, some information is actually dynamic, such as the performer names).

B. Event semantics - upcoming.org perspective

The inference of semantics related to events in places works as an extension of the previous section: we now focus on dynamic online resources that provide information associated with time. By *dynamic* we refer to websites that change content at least on a daily basis. In other words, the main difference lies in the selection of web resources and in the specific attention given to time (having an exact information of date and time is important).

Within the internet, the range of dynamic resources about events is rapidly growing throughout the world, becoming a challenge by itself to simply enumerate the existing variety. It thus demands clear criteria: event coverage; geographical coverage; richness of content; availability of historical data and reliability of sources. We consider event coverage being the ratio of events in the database with the ones that actually happen in a geographical area. Of course, the ideal value is 1 (every event is reported in the database). Geographical coverage corresponds to the area associated to the database. For experimental purposes, we favor event coverage over geographical coverage, but the selected area should have significant event life. By richness of content, we mean how detailed the available information is aside from the mandatory data (name, date, time, place) namely the availability of some text description. The availability of historical data is important for practical purposes: storing all event information for research analysis can demand large computational resources and take long time. The upcoming.org website brings a good compromise in those aspects, although its coverage is not perfect, it is a well organized database with historical data. For each event, it provides its category (e.g. Music, Social, Commercial, etc.), name, date, time and textual description.

Below, we show an excerpt of the description text for event 353171

Concept	Score	Wordnet gloss
Bach	0.637	
dich	0.637	
Eli Spindel	0.637	
Gott	0.637	
Kimberly Sogioka	0.637	
Mozart	0.637	
Serenata Notturna	0.637	
Thomas Tallis	0.637	
Torelli	0.637	
Vaughn Williams	0.637	
Fantasia	0.546	a musical composition of a free form usually incorporating several familiar themes
donation	0.431	act of giving in common with others for a common purpose especially to a charity
Theme	0.0.407	a unifying idea that is a recurrent element in literary or artistic work; "it was the usual 'boy gets girl' theme"

TABLE III
CONCEPTS FROM EVENT ID 353171 OF UPCOMING.ORG.

("The String Orchestra of Brooklyn: Winter Concert", 2007-12-15, 20:00:00, at "St. Ann and the Holy Trinity Church"):

> Bach: Erbarme dich, mein Gott from the St. Matthew's Passion Torelli: Christmas Concerto Vaughn Williams: Fantasia on a Theme by Thomas Tallis Mozart: Serenata Notturna, k239 Eli Spindel, conductor Kimberly Sogioka, alto Suggested donation: $10

From this text, using the Meaning Extraction module from Kusco, we can directly extract the concepts found in table III (with their Wordnet glosses, when available). Notice that, in this case, we're not asking the system to perform any web search so this text is the only resource used.

This index can be enriched with the Wikipedia, as shown in the next section.

C. Encyclopedic descriptions - Wikipedia perspective

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 references and external links. The abstract, for it is a summarization of the concept, catches the main highlights of each concept and is therefore the perfect candidate for mining.

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. In table IV, we show the sequence of word lists obtained for the example of section III-A ("Gigantic Artspace"), before filtering, after filtering, and after Wikipedia enrichment (applied on the filtered list).

The system could retrieve a number of more distant, yet potentially relevant, associations. For a more constrained example, with upcoming.org, table V shows the top 5 words obtained with the same process. For each of the events (available in the upcoming.org website, with the corresponding event id), we pick the top 5 words

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

TABLE IV
AFTER FILTERING AND AFTER WIKIPEDIA ENRICHMENT

Tagmap list	After filter	Wikipedia
KnittingFactory, tribeca, ghostbusters, Manhattan, bull, BoweryBallroom, LowerEastSide, WallStreet, ManhattanBridge, CafeHabana, LittleItaly, cityhall, MunicipalBuilding, Spring, eastriver, Brooklyn, BrooklynBridge, TrinityChurch, CityHallPark, courthouse, SouthStreetSeaport, subway, wtc, WorldTradeCenter, AppleStore, chinatown, new, PrinceStreet, uniqlo, soho	Knitting Factory, tribeca, ghostbusters, Manhattan, bull	bull, Factory, festival film, Ghostbusters, Poster, tribeca, neighborhood, Hudson, magazine, county, TribecaTribeca, HaroldRamis, boroughs, footnotes, Elevation, IvanReitman, warehouses, community, series, size, seat, title, power, boxofficemojo, ghost, adjustments, comedies, Funniest

TABLE VI
FLICKR TAG LIST BEFORE, AFTER FILTERING AND AFTER WIKIPEDIA ENRICHMENT

Id	Category	Name	Top-5 words
353171	Music	The String Orchestra of Brooklyn: Winter Concert	music, suites, johann sebastian bach, style, forms
462350	Media	Aleksey Budovskiy: Russian cartoons recent and classic	country, animators, soviet, language, arms
449040	Family	Easter is 'Egg'cellent in Lower Manhattan	families, children, symbol, nist, units
250921	Education	The Apartment (1960): Movie Nights on the Elevated Acre	star, comedy, part, core, title
396331	Performing/ Visual Arts	Renascence: International New Media Exhibition	premiere, artists, performances, exhibition, term
447037	Other	NY Giants' Justin Tuck Autograph Signing at J&R Music World	team, bowl, world, game, eastern
282198	Festivals	Stone Street Oysterfest	oysters, pub, term, shell, fogelson
350299	Other	Trinity Church Choir Live at J&R on 12/6	spirit, performance, people, performers, example
692856	Other	Regina Belle Performance & Autograph Signing	cd, autographs, disc, media, minutes
323193	Commercial	IBM and ACORD eForms+ Development Tour: Extend electronic forms capabilities	forms, industry, acord, workshop, area

TABLE V
WIKIPEDIA ENRICHMENT OF 10 UPCOMING.ORG EVENTS (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).

D. Free association - Flickr perspective

In Flickr, users can add arbitrary tags to pictures. Given its current growth rate (of aprox. 3 million new pictures every month), the potential for statistical strength on data analysis is enormous. In section II, we already enumerated a number of related projects. One of those works [3] provides an API for accessing the *tagmaps* associated to a given point or bounding box. In spite of being biased towards geographical descriptors (thus becoming redundant for the meaning of the place), these tag lists can become useful after applying our stopword filter. The question can be raised, however, whether this "Flickr perspective" can add more valuable information on place than what is achievable with the other resources. An interesting experiment is to enrich Flickr data itself with Wikipedia knowledge. In this way, we can look for "explanations" of the tags used. As an example, we choose a tagmap spot that contains two POIs ("Knitting Factory" and "Ghostbusters HQ") and that allows to see the strengths and weaknesses of the method. In table VI, we can see the results before, after filtering, and with filtering and Wikipedia enrichment.

The filter removed words that are not discriminant of the place within the area. Surprisingly, some of the tags (e.g. "wtc", "AppleStore") are so common in the surroundings that they end up not revealing the place. Also notice that almost all geographically related references were removed, ending up with only 5 concepts. These words are in turn the *seeds* for the Wikipedia enrichment. The resulting list is the selection of the words that were frequent (*tf-wise*) among all the retrieved wikipedia abstracts. Understandably, the word "bull" is not present in Wikipedia. The entry for "Knitting Factory" has an extremely poor abstract (one sentence). There are still two geographical references ("Manhattan" and "tribeca") which end up contributing with some words related to the area (e.g. "Hudson", "neighborhood", "burroughs", "county"). The remaining words show a mix of informative words on Ghostbusters (e.g. "Ivan Reitman" was the director of "Ghostbusters"; the style is in "comedies", it is about "ghosts" and became a TV "series") and generic words on Factory related information.

IV. EXPERIMENTS

The experiments made are related to a specific project in which correlation is sought between semantics of place and mobility data, therefore the study area and time window are constrained by the available data for the project. More specifically, we work with the Lower Manhattan area around the New York City Waterfalls exhibit from the artist Olafur Eliasson. This corresponds to a polygon that falls within the bounding box from 40.698191", -73.991739" to 40.715693", -74.021560" (aproximately 3×2 km^2). To this we call the "waterfalls area". For some experiments, we also consider an "extended area" (which covers a larger portion of Manhattan, until 7th street, 40.68366", -73.960933" to 40.727915", -74.0401914", aprox. 7×5 km^2) to let us obtain a larger number of points. The time window chosen goes from August 2007 to August 2008, covering the largest part of the Waterfalls exhibition period. As with the area, this choice was made in synchrony with the mobility analysis project. In either case, we believe the choices are valid.

In order to allow the comparison of results among the different *perspectives*, we use the same set of POIs, essentially corresponding to 107 venues in the "waterfalls area" and 716 in the "extended area".

For the open web perspective we initial set of POIs, we took a sample of 292 different venues that include all from the "waterfalls area" and some from the "entended area". 2150 pages were retrieved (average of 7.36 pages per POI with std. deviation of 3.07). Such variety of information raised in a Semantic Index of 56 concepts in average (std. deviation=28.49). From all semantic indexes, the 5 top most popular terms retrieved were "Terms of Service" (131 times), "Neighborhood" (123), "Zip" (120), "Suggestion Board" (79) and "Search Local" (70). These have no valid semantic information and raise the issue that the filter needs to be improved. In this regard, we ran the algorithm with several percentile values (5%, 10%, 20%, 50% and 90%) as can be seen in table VII. From a subjective analysis on the end results, we see that the filter is effective in removing noise, however it still fails with those concepts, meaning that work has to be done to improve it.

For the experiments with Flickr data, for each of the 716 venues found for the extended study area, we defined a bounding box with aproximately 250 meters in each direction. We then retrieve the tagmap list for each of the points. The total number of tags is 17961, with an average of 25 tags per point. However, the diversity is extremely small. There are only 125 different tags. Table VIII shows the top-15 list of tags in terms of appearances in the tagmap lists.

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

TABLE VII
FILTER PROGRESS WITH DIFFERENT CONFIGURATIONS

Flickr tag	Count	Flickr tag	Count
New York	716	Tompkins Square Park	453
NYC	716	pickles	444
chinatown	672	d b a	444
subway	654	Mercury Lounge	444
East Village	473	WD 50	444
cbgb	464	soho	435
Katz s	462		
sin e	462		
delancey	462		

TABLE VIII
TOP 15 TAGS FROM TAGMAPS OF THE EXTENDED STUDY AREA.

This lack of diversity may demonstrate that the tagmap lists are poor in differentiation among neighbor areas, and reiterates the fact that they are essentially geographical reference oriented as opposed to a perspective on content or function of the place. In reality, the algorithm of [3] may be hiding rich detail underneath their clustering method and to overcome this, the obvious solution seems to be to implement a clustering algorithm ourselves (or ideally adding our filtering philosophy to tagmap).

For each of the POIs, the process of filtering and enrichment was applied (as in section III-D). There was an average of 20 filtered tags per point. After enrichment from Wikipedia, we get an average of 137.45 new words, the TFIDF mean being 0.0540 (the std. deviation is a very high 0.1678, since the top words have high values).

In the case of upcoming.org, for each event on the waterfalls area and during the study period, we allowed Kusco to extract the top-5 words from the text description and then enriched those words with Wikipedia. The total number of words obtained is 724, of which 418 are different from each other. The word "music" is the one that appeared more times (23), followed by the words "internet" and "artists". Overall, the full word spectrum has the following pattern: 41.3% of the words show up only once, 12.9% twice, 14.5% three times, 6% 4 times, then there is a slow decay, i.e. the majority of the words appear only once or twice, which would be expectable given the small number of events covered. Since upcoming.org provides category information, we can perform classification analysis (clustering with K means and "Farthest First" [17]; generating association rules with apriori [18]). The results show some coherence with the event categories, which is notable for such a poorly sized set. In table IX, we can see that K means organized data into 4 categories("Performing/Visual Arts", "Music", "Education" and "Comedy"), and some of the words are not directly related to the categories (e.g. "dr" or "countries"). With the "Farthest First" algorithm, a centroid for each category was inferred from the examples. These results indicate that there still plenty to be improved in many respects but we area already obtaining useful results. A confirmation can be given in table X, the result of the apriori algorithm.

Regarding the tf-idf statistics obtained, the mean value is 0.1734 (std. deviation=0.333) for the 5 world blocks obtained (out of a list with 143.7 concepts in average). From the empirical observation, the words obtained are in general relevant to the topic.

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

TABLE IX
WORD CLUSTER CENTROIDS ACCORDING TO CATEGORY

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

TABLE X
AN EXCERPT FROM THE ASSOCIATION RULES WITH HIGHEST ACCURACY FOUND (NUMBERS ON THE RIGHT INDICATE ACCURACY).

V. 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 use of the stopword filter reduces considerably this noise, but the ideal threshold has to be carefully negotiated in order to minimize false negatives/positives. The value chosen should also depend on the perspective: for open web, a conservative choice may become preferred (prefer to eliminate good words than to get too much noise); for Flickr, it depends on whether geographically related descriptions are wanted; in the case of upcoming, the results indicate that a non-conservative value (i.e. low value) can be used, as the descriptions are very concise in general.

From the empirical analysis of the experiments, the results from upcoming.org are the ones that demonstrate more quality. This happens due to a number of factors: their text descriptions are objective and normally concise; they relate to entities that often are present in Wikipedia (e.g. artists, films, workshop themes); they are free of any lateral information such as publicity.

On the other side of the spectrum, the "open web" mode is the most fragile in that there is no control on the resources. It is however very useful in two important ways: to test the system in the most demanding scenario; to allow the search for semantics even for places that are badly represented.

Flickr tagmaps showed surprising limitations and to unravel those, it seems to be mandatory to work directly with its original data, raising issues that were assumed to be solved by that project, namely regarding the clustering of words using tf-idf.

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 is *per se* an ambiguous task, as referred in the introduction. 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. Given the variability of words and places, to get a statistically valid sample we would need thousands of replies. While it is not out of the question to apply this technique, it is clear that the "ground truth" validation is a strong limitation of this methodology and it will demand a very careful consideration.

Another limitation is on the side of performance. Some parts of Kusco (which is implemented in Java) are extremely slow such the NER algorithm. The application of the filter becomes often slow because of the need of stemming the original words. It also makes web searching and screen scrapping (in Wikipedia). Depending on the configuration (from just retrieving tags and filtering to perform the complete processing) , we need from 5 to 55 seconds to process each POI. Since that, except for the Flickr tags, we are dealing with unstructured Natural Language texts, these values are not surprising, but it raises the obvious question of scalability. Besides the code optimization, for a 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 actually 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 [1], these ontologies were extremely poor in terms of domain knowledge and the results were weak. In future iterations, we plan to revisit this approach.

VI. Conclusions and Further Work

In this paper, we presented an approach to the extraction of semantics for places from online resources. We consider the different perspectives that underlie each of those resources and explore each other in order to obtain the richest knowledge possible. In practical terms, what we intend is to generate a semantic index that describes each perspective. This index can be useful for a wide number o applications, namely POI search, context-awareness or the study of the city as an aggregation of semantically rich space.

Results show that, in spite of the inherent complexity of the problem emanating from the Natural Language nature of texts, we can obtain meaningful descriptions of place both from the static perspective (what the place *is* throughout time) as well as from the dynamic perspective (what *happens* in the place).

We described a number of perspective integrations, namely the use of Wikipedia for enrichment of the other resources. However, a fully "integrated perspective" can be proposed that consists on facing the meaning of place itself as a function. Assuming that OW, UP and $Flickr$ correspond to the TF-IDF value of a concept in the perspectives of Open Web, upcoming.org and Flickr, respectively, the rank of each concept should be given by the following weighted sum:

$$OW \times w_1 + UP \times w_2 + Flickr \times w_3$$

The next steps for this project will be to integrate these perspectives as shown and study its effects. Finally, we intend to add other online resources (e.g. RSS feeds, Twitter, delicious).

References

[1] A. C. 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.

[2] 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*. New York, NY, USA: ACM, 2007, pp. 103–110.

[3] S. Ahern, M. Naaman, R. Nair, and J. H. Yang, "World explorer: visualizing aggregate data from unstructured text in geo-referenced collections," in *JCDL '07: Proceedings of the 2007 conference on Digital libraries*. New York, NY, USA: ACM Press, 2007, pp. 1–10. [Online]. Available: http://dx.doi.org/10.1145/1255175.1255177

[4] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *WWW '09: Proceeding of the 18th international conference on World Wide Web*, 2009. [Online]. Available: http://www2009.eprints.org/77/

[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*. New York, NY, USA: ACM, 2006, pp. 193–202.

[6] 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*. New York, NY, USA: ACM, 2006, pp. 89–98.

[7] 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*. New York, NY, USA: ACM, 2004, pp. 273–280.

[8] N. Imagery and M. Agency, "Geonet names server (gns)," 2007. [Online]. Available: http://earth-info.nga.mil/gns/html/index.html

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

[10] 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.

[11] 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*. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 1121–1128.

[12] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988. [Online]. Available: http://portal.acm.org/citation.cfm?id=54260

[13] Fellbaum, *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998. [Online]. Available: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/026206197X

[14] R. Mihalcea, "Semcor semantically tagged corpus," CiteSeerX - Scientific Literature Digital Library and Search Engine [http://citeseerx.ist.psu.edu/oai2] (United States), Tech. Rep., 1998. [Online]. Available: http://citeseer.ist.psu.edu/250575.html

[15] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980. [Online]. Available: http://portal.acm.org/citation.cfm?id=275705

[16] G. K. Zipf, *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology*. Addison-Wesley, 1949. [Online]. Available: http://dblp.uni-trier.de/rec/bibtex/books/aw/Zipf49

[17] D. Hochbaum and D. Shmoys, "A best possible heuristic for the k-center problem," *Mathematics of Operations Research*, vol. 10, no. 2, pp. 180–184, 1985.

[18] T. Scheffer, "Finding association rules that trade support optimally against confidence," in *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, 2001, pp. 424–435.