Tagging Space from Information Extraction and Popularity of Points of Interest

Ana O. Alves[1,2], Filipe Rodrigues[1], and Francisco C. Pereira[1]

[1] CISUC, University of Coimbra, Portugal
{ana,camara}@dei.uc.pt, fmpr@student.dei.uc.pt
[2] ISEC, Coimbra Institute of Engineering, Portugal
aalves@isec.pt

Abstract. This paper is about automatic tagging of urban areas considering its constituent Points of Interest. First, our approach geographically clusters places that offer similar services in the same generic category (e.g. Food & Dining; Entertainment & Arts) in order to identify specialized *zones* in the urban context. Then, these places are analysed and tagged from available information sources on the Web using KUSCO [2,3] and finally the most relevant tags are chosen considering not only the place itself but also its popularity in social networks. We present some experiments in the greater metropolitan area of Boston.

Keywords: Context-Awareness, Semantic Enrichment, Web Mining.

1 Introduction

Understanding local context has been a recurrent challenge in Pervasive Computing. Besides information from sensors (e.g. latitude/longitude, wifi, cell-id, bluetooth, etc.), some functional or semantic properties can be collected that inform the overall context. Progress has been done in identifying such information for individual POIs [3]. However, besides just knowing about the exact place where we are (e.g. a specific Point of Interest), which per se can be more than enough to support many location based services, identifying the "big picture" of the area can bring this context to another level. Aided by social network popularity indicators (e.g. Gowalla[1] check-ins), we can even ground ourselves on what are, according to these communities, the most relevant spots to consider.

Ideally, the spatial resolution of this local context can vary from the point of interest to the entire city, depending on the application. A user might want enriched information about a specific place or about an entire region. In any case, for our research, the smallest entity is indeed the POI (Point-of-Interest). We consider a POI to be a touristic place (e.g. Boston Common) or a space with a given functionality (e.g. a post office agency).

These POIs are nowadays available from commercial or public POI sources (e.g. Manta[2] and Yahoo!Local[3]) and they generally refer to *buildings* rather

[1] http://www.gowalla.com
[2] http://www.manta.com
[3] http://local.yahoo.com

D. Keyson et al. (Eds.): AmI 2011, LNCS 7040, pp. 115–125, 2011.

than other kinds of places, like: parts inside buildings, regions, junctions, and others[9].

We propose an approach to visualize the urban space through tags taking in account available online information (more static knowledge) and popularity (more dynamic) about places in the social Web. This aim is accomplished by following some steps for a given city: First, POIs are massively extracted from public POI sources. These POIs are then grouped by Machine Learning techniques as clusters of related services considering generic categories (e.g. Recreation & Sporting Goods; Government & Community) that are geographically close. In parallel, tags are associated to each POI using Web Mining and Natural Language processing to extract relevant concepts which best describe the given place. And finally, a social network (Gowalla) is used to infer the popularity of places in order to compute the social significance of a given area considering this community and to select a tag that best represents it. Hence, the main contribution described in this paper is to extend KUSCO's tag ranking. KUSCO previously has considered all POIs in the city of equal importance, but the approach here proposed adds another dimension to the selection of most representative tags, the popularity of POIs.

The remaining of this paper is organized as the following: in the next section we present the related work. In section 3, the methods to obtain the POI data and group it in clusters are described. The process of information retrieving, extracting and computing the relevance of tags is detailed in section 4. Some experiments and validation are summed up in section 6. And finally, in section 7 we presents the conclusions and discuss future work.

2 Related Work

Context is any information that can be used to characterize the situation of an entity at a given time. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves. Context generally refers to all types of information pertaining to a service and/or the user of the service[1]. Knowledge typically refers to more general information, of which context is a specific type. Knowledge would typically include information about users and their preferences, and also information that can be inferred from other sources. A system is context aware if it observes, reacts and changes accordingly to the context. Context information can be gathered from several sources including sensors, devices, data repositories and information services. Context data can be used to make inferences.

The key aspects of context are: location, agent or person, time and activity. These elements are used to answer basic questions related to a user, place or object which is target of context representation: *Where*, *Who*, *When*, *Why*. The main focus of this research is to represent *Where* in a more meaningful way with semantic tags.

The possibility to automatically associate labels has been investigated in the literature [4,7,10,14,11]. These approaches either use additional information, such as time of day and point-of-interest databases, to determine the type of building, or attempt to assign labels by comparing places across users. Some works use machine learning algorithms to induce these labels based also in other variables (time of day, weekday, etc.). These labels are limited to generic and personal ones like: work, home, friend and other. But our approach is not centered in the user, instead in the *Place* itself. And this representation should include public aspects and the functionality of places, not being of great importance the relation between some individual and the place itself. We think that a richer representation of Place with more meaningful common-sense concepts associated will help works like these described above.

From these works the most related is that proposed by Rattenbury et al. [14]. They 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. And the main difference to our approach is that we automatically *generate* tags not depending on contribution given by users.

Regarding the use of Gowalla, its potential and the potential of other similar location-based services like Foursquare[4] and Facebook Places[5] has already been demonstrated in recent work and it is being increasingly exploited as the dimensions of such services grow. Cheng et al. [6] provide an assessment of human mobility patterns by analyzing the spatial, temporal, social, and textual aspects associated with the hundreds of millions of user-driven footprints (i.e., "check-ins") that people leave with these services. Anastasios et al. [13] provide a similar study but they also analyze activity and place transitions. Both of these studies are very interesting and motivating for a further exploitation of this kind of services. For example, in [5] the authors exploit the use of Gowalla to develop a Recommender System for places in location-based Online Social Network services (OSN) based on the check-ins of the entire user base.

3 POI Mining

POI Mining refers to the processes of extraction, pre-processing and pattern recognition (namely clustering) in POI data that are the basis of the approaches that will be later presented in this paper.

[4] https://foursquare.com/

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

3.1 POI Extraction

The growing number of smartphones and social networks during the latest years has been producing a vast amount of geo-referenced information on the Web. Capture devices such as camera-phones and GPS-enabled cameras can automatically associate geographic data with images, which is significantly increasing the number of geo-referenced photos available online. Social networks also have an important role. They are a great medium where users can share information they collect with their mobile devices. As a consequence, the amount of online descriptive information about places has reached reasonable dimensions for many cities in the world.

In spite of their importance, the production of POIs is scattered across a myriad of different websites, systems and devices, thus making it extremely difficult to obtain an exhaustive database of such a wealthy information. There are thousands of POI directories in the Web, with POIs for places all over the World. These are in fact great sources of information. However, each one uses its own format to represent the POIs and its own taxonomy to classify them. Also, the Web servers that provide POI information (e.g., Yahoo!, Manta, Yellow Pages, CitySearch, Upcoming) are mere repositories, and therefore, they don't take advantage of the full potential of such information.

Despite the fact that our approach could be applied to any POI source and that for the sake of completeness multiple POI sources should be used, doing so would require a careful process of POI matching and integration, therefore, for just a matter of simplicity, in this paper we focus only on POI data extensively extracted from Yahoo! through its public API for the Boston Metropolitan Area.

3.2 Clustering

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

The subgroup of density-based clustering algorithms are devised to discover clusters of arbitrary shapes where each is regarded as a region in which the density of data instances exceeds a threshold, making them perfect for the identification of "hotspots" of POIs in the city (i.e. places with high concentrations of POIs). In this paper we use DBSCAN [8] to identify such "hotspots" that would be the basis of the Semantic Enrichment (Section 4) and Visualization (Section 6) processes.

In order to take the categories in consideration in the clustering process, we adopted a two-level clustering approach. In the first level, we group together POIs that are closer to each other according to their proximity in the Yahoo! taxonomy, and in the second we apply DBSCAN over these groupings, thus producing clusters of geographically nearby POIs that also have a similar set of categories. This approach will give us a different perspective of the POI data.

4 Semantic Enrichment of POIs

An approach that is able to extract relevant semantics from places can be useful for any context-aware system that behaves according to position. The level of information considered in this work brings another layer to add to other sensors (GPS, accelerometer, compass, communications, etc.), eventually pushing forward the potential for intelligent behavior. This section presents an approach to such a system and its implementation, resulting in an architecture called KUSCO a system which tags POIs using available sources of information (perspectives) on the Web. We briefly summarize this system for completeness of this paper, but redirect the reader to any of earlier publications for further details and evaluation of this approach[2,3].

Beyond the data available from commercial and community based POI sources, enrichment with public information is desired. Initially, the entire Web was used to retrieve such information but the great level of noise obtained led us to constrain our enrichment source to Wikipedia[6]. It is on this database that we apply two different methods, the “red and yellow perspectives” of place.

Low-detail labeling: the red wiki perspective

In the red wiki perspective, we extract the Wikipedia page corresponding to the identified category of a POI. Local POI directories are normally structured in a hierarchical tree of categories. This taxonomy may be created by the company itself or be collaboratively built by suggestion of users who feed the system with new POIs.

Since no API is currently available from Yahoo! to extract the entire taxonomy of POI categories, we have created a wrapper based on regular expressions in order to automatically extract it. Yahoo! only presents categories through menu navigation along its web site.

Each POI is only associated to leaf categories (more specific ones) instead of generic categories. These categories, in the middle and top of the taxonomy, are completely hidden from API when retrieving POIs. Curiously, a dynamic property of this POI source is also observed in the fact that this taxonomy is different depending on which city we are virtually visiting. Namely, Yahoo builds dynamically their menus, thus presenting proper taxonomies to distinct cities. Through time, this taxonomy grows with new types of services and places.

To contextualize each category in the corresponding Wikipedia article we base ourselves on string similarity between the category name and article title. We have opted for a top-down approach, from main categories to taxonomy leaves. To increase the confidence of this process, we disambiguate manually the main categories to start with and make sure that at least a more generic category will be connected to the Wikipages of its hypernym. When a POI has many categories, we obtain the articles for each one and consider the union of all the resulting articles as the source of analysis. Since there are many different combinations of categories, we can guarantee that each POI gets its own specific flavor of category analysis. For instance, consider the POI *Boston University* which is

[6] http://www.wikipedia.org/

classified under the Yahoo! categories: (1) Colleges & Universities; (2) High Schools; (3) School Districts. These categories are automatically mapped by KUSCO to the respective Wikipedia articles: (1) http://en.wikipedia.org/wiki/Universities & http://en.wikipedia.org/wiki/Colleges; (2) http://en.wikipedia.org/wiki/High_schools; (3) http://en.wikipedia.org/wiki/School_districts.

Medium-detail labeling: the yellow wiki perspective

While the previous approach is centered on place category, here we focus our attention on Place name. We use string similarity to match Place name to Wikipage title in order to find the Wikipedia description for a given place. On a first glance, this method is efficient in mapping compound and rare place names such as 'Beth Israel Deaconess Medical Center' or 'Institute of Real State Management', however it can naively induce some wrong mappings for those places with very common names (e.g., Highway - a clothing accessories store in New York, Registry - a recruitment company in Boston, Energy Source - a batteries store in New York). We approach this problem by determining the specificity of place names, and only considering those with high Information Content (IC)[15]. The Information Content of a concept is defined as the negative log likelihood, -logp(c), where p(c) is the probability of encountering such concept. For example, 'money' has less information content than 'nickel' as the probability of encountering the concept, p(Money), is larger than encountering the probability of p(Nickel) in a given corpus. For those names present in Wordnet (e.g. Highway, Registry), IC is already calculated [12], while for those not present in Wordnet, we heuristically assume that they are only considered by our approach if they are not a node in Wikipedia taxonomy, i.e., a Wikipage representing a Wikipedia category (case of Energy Source), but being only a Wikipedia article.

Having a set of textual descriptions as input, KUSCO extracts a ranked list of concepts. This ranking is based solely on TF-IDF [16] (Term Frequency × Inverse Document Frequency) value in order to extract the most relevant terms that will represent a given place.

5 Tag Relevance Computing Based on Popularity

Beyond the traditional TF-IDF computed for individual POIs against other indexes in the POI database, we also use Gowalla to infer a popularity-based TF-IDF for the terms of a POI p in a given cluster c using the POI check-ins. The idea is that concepts associated with POIs that are very popular should be weighted favorably. Equation 1 shows how the popularity-based TF-IDF is calculated for each concept i in a given cluster c based on the POIs p that belong to that cluster.

$$Popularity\text{-}based\ TF\text{-}IDF_{i,c} = \frac{1}{|c|} \sum_{p \in c} TF\text{-}IDF_{i,p} * check\text{-}ins_p \qquad (1)$$

6 Experiments

Using as a test scenario the greater metropolitan area of Boston, we extracted 156364 POIs from the Yahoo! public API. Each POI has an average of 2 categories and the Yahoo! taxonomy is spread across three different levels of specificity, where the top level has 15 distinct categories and the lower level has a total of 1003 categories[7]. Table 1 shows the distribution of the extracted POIs over the top categories.

Table 1. POI distribution over the different Yahoo! categories for the different perspectives

Yahoo! Category	Total number of POIs	# POIs with RedWiki	# POIs withi YellowWiki
Automotive	8109	1698 (20.9%)	377 (4.6%)
Business to Business	37321	9488 (25.4%)	1034 (2.8%)
Computers & Electronics	3767	652 (17.3%)	100 (2.7%)
Education	3822	1277 (33.4%)	213 (5.6%)
Entertainment &Arts	4327	1611 (37.2%)	250 (5.8%)
Food & Dining	10383	1734 (16.7%)	433 (4.2%)
Government & Community	10646	2640 (24.8%)	186 (1.7%)
Health & Beauty	17100	4344 (25.4%)	317 (1.9%)
Home & Garden	22577	5127 (22.7%)	153 (0.7%)
Legal Financial Services	10727	1823 (17.0%)	188 (1.8%)
Professional Services	11658	3270 (28.0%)	436 (3.7%)
Real Estate	7059	1066 (15.1%)	50 (0.7%)
Recreation & Sporting Goods	3029	1051 (34.7%)	77 (2.5%)
Retail Shopping	9021	1663 (18.4%)	324 (3.6%)
Travel Lodging	3944	1599 (40.5%)	111 (2.8%)

In the clustering phase we grouped together POIs that shared the same top-level category, and then for each top-level category we applied DBSCAN using the POI coordinates. The parameters of the DBSCAN we manually tuned by running the clustering algorithm many times with different parameter setting and visually validating the results in a map. The goal was to choose a set of parameters that produced a balanced number of clusters that covered most of the different areas of the city. Figure 1 depicts the centroids of a possible clustering solution. We can see that the dominance of some categories over the others is also reflected in the clustering.

Table 1 shows for each Yahoo! top category the number of POIs extracted as the number of POIs enriched for both perspectives by KUSCO. We can observe the greater coverage of Red Wiki perspective as opposite to Yellow Perspective. This can be explained by the fact that almost every POI is categorized under at least one category in Yahoo!, and each category is mapped to at least one Wikipedia article (except in the case where there are more than one mapping

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

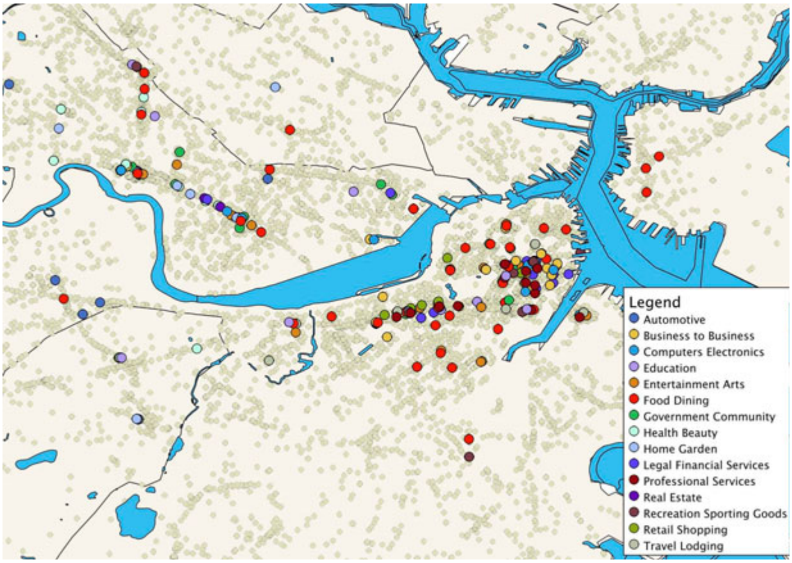

Fig. 1. Centroids of the clusters identified using the POI data enriched with the Red-Wiki perspective and the correspondent Yahoo! categories

to Wikipedia, e.g. Computers & Electronics). While in the Yellow Perspective, KUSCO searches for more specific information on Wikipedia: the POI article, when it exists. The enrichment process was validated earlier in previous studies [3] and we obtained a precision over than 60% ($\sigma = 20\%$) using a survey to 30 visitors or inhabitants of Boston.

In order to visualize and understand the whole process, consider the top 5 most popular categories in Gowalla (and the respective Yahoo! category) [8] for the greater metropolitan area in Boston. They are: Food (Food & Dining), Shopping & Services (Retail Shopping), Architecture & Buildings (Real State), Nightlife (Entertainment & Arts), College & Education (Education). From the first view of the city (Figure 2), we can observe a great predominance of common concepts as the system is dealing with generic information associated to the POI categories (perspective Red Wiki). The more relevant is a tag, the greater its font size. For instance, the term *health services* comes from POIs belonging to a not so popular category regarding Gowalla: Health & Beauty. But the concentration of very similar POIs related to the health services is so high [9], that this specialized zone is identified and the most relevant tag in all these related subcategories of Health top category is chosen (no matter the popularity of its POIs).

Another interesting example that helps us understand how the popularity is crucial to determine the most relevant tag, is the cluster identified by *secondary*

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

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

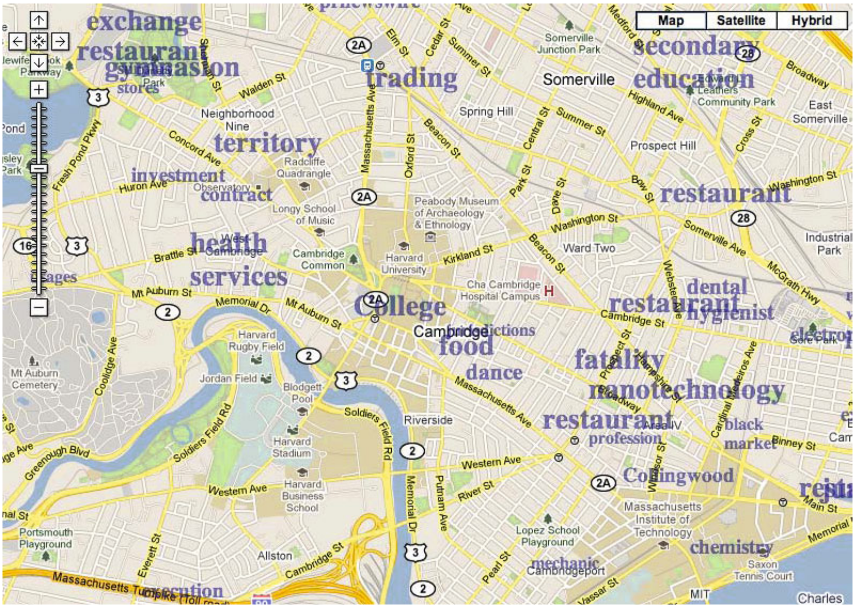

Fig. 2. Most relevant tags from the RedWiki perspective using DBSCAN (epsilon=0.0005, minPoints=15) to cluster POIs

education. In this cluster, different types of POIs we grouped, in this case in a not a so specialized *zone* [10], but as the most popular POI among them is a High School with a lot of checkins associated, this fact biased the weight of chosen tag.

Considering a different region of the city (Figure 3), by the Yellow Wiki perspective we find more specific tags as we are dealing with the proper Wikipedia article to each POI (when it exists). In the Yellow perspective, as we only have extracted information about each POI itself, we opted for displaying only the most popular POIs. In this figure, we can see interesting POI-tag relationships: Cambridge Innovation Center - *business incubator*; Boston Common - *Central Burying Ground*; Massachusetts General Hospital- *Harvard Cancer Center*; Boston University - *Colleges*; Louisburg Square-*Beacon Hill*; Best Buy-*forbes*; California Pizza Kitchen-*Richard*. The last two examples reflect some difficulties that we have faced in the present methodology: the company 'Best Buy' related to the concept Forbes (a magazine). This is not so relevant to understand the POI since this fact [11] is only referred in the summary of Wikipedia article since the second paragraph. This could be used to decrease the weight of extracted concepts: how long they are from the first paragraph.

[10] e.g. Somerville High School, GC Vocal Studio, Dexter Painting and Carpentry, FISH Magical Enterprises.

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

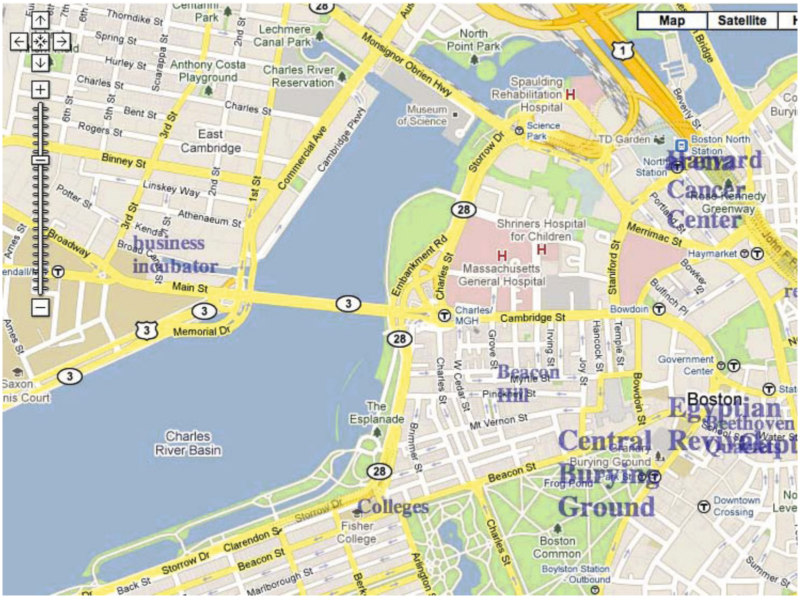

Fig. 3. Most relevant tags from the YellowWiki perspective using most popular POIs according to Gowalla (no clustering)

Considering the last pair POI-tag showed before, it is relatively straightforward to verify that Richard Rosenfield is a co-founder of California Pizza Kitchen. In this sense if we knew more about this POI, namely the semantic behind it, regarding DBpedia*http://dbpedia.org/page/California_Pizza_Kitchen* it would be possible to infer their relationship: founder(California_Pizza_Kitchen, Richard Rosenfield).

7 Conclusions and Further Work

We presented a methodology for extracting semantic information about arbitrary sized areas, depending on the availability of Points of Interest. The nature of this process is ultimately subjective since all information is extracted automatically from crowd sourced resources. However we rely ourselves in techniques that favor statistical relevance, specificity and popularity to select the words (or tags) that should represent better the context according to "how people understand that space". Furthermore, the concept of "perspectives" explicitly models the unavoidable ambiguity in this problem. We took two approaches to Wikipedia as two ways to understand the same space. Others could be included (e.g. using twitter, facebook, eventful, etc.).

The use of popularity, from Gowalla, helps understanding the social dimension of space. From a purely democratic point of view, the more people that enters a place (and happily reports it), the more it is relevant for the community.

Of course, this raises questions itself on how representative is the population that uses Gowalla, and how they actually report the places they visit (e.g. don't they check in more often when waiting in a fast food queue than when having fun?). The work here presented will become more representative as such communities grow. As a further improvement of our approach, we plan to also use the POI radius available from Gowalla as a feature to consider in the cluster algorithm, since very wide POIs (e.g. MIT), at this time, have the same weight of the small POIs (Starbucks).

References

1. Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P.: Towards a Better Understanding of Context and Context-Awareness. In: Gellersen, H.-W. (ed.) HUC 1999. LNCS, vol. 1707, pp. 304–307. Springer, Heidelberg (1999)
2. Alves, A.O., Pereira, F.C., Biderman, A., Ratti, C.: Place Enrichment by Mining the Web. In: Tscheligi, M., de Ruyter, B., Markopoulus, P., Wichert, R., Mirlacher, T., Meschterjakov, A., Reitberger, W. (eds.) AmI 2009. LNCS, vol. 5859, pp. 66–77. Springer, Heidelberg (2009)
3. Alves, A.O., Pereira, F.C., Rodrigues, F., Oliveirinha, J.a.: Place in perspective: extracting online information about points of interest. In: Proc. of AmI 2011 (2011)
4. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: geotagging web content. In: SIGIR 2004, pp. 273–280. ACM, New York (2004)
5. Berjani, B., Strufe, T.: A Recommendation System for Spots in Location-Based Online Social Networks. In: Proc. of Eurosys Works on Social Network Systems (2011)
6. Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z.: Exploring Millions of Footprints in Location Sharing Services. In: ICWSM 2011, Barcelona, Spain (2011)
7. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., Tomkins, A.: Visualizing tags over time. In: WWW 2006, pp. 193–202. ACM, New York (2006)
8. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, pp. 226–231 (1996)
9. Falko Schmid, C.K.: In-situ communication and labeling of places. In: 6th International Symposium on LBS & TeleCartography. Springer, Heidelberg (September 2009)
10. Jaffe, A., Naaman, M., Tassa, T., Davis, M.: Generating summaries and visualization for large collections of geo-referenced photographs. In: MIR 2006, pp. 89–98 (2006)
11. Lemmens, R., Deng, D.: Web 2.0 and semantic web: Clarifying the meaning of spatial features. In: Semantic Web meets Geopatial Applications, AGILE 2008 (2008)
12. Mihalcea, R.: Semcor semantically tagged corpus. Tech. rep., University of North Texas (1998), `http://citeseer.ist.psu.edu/250575.html`
13. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: An Empirical Study of Geographic User Activity Patterns in Foursquare. In: ICWSM 2011, Barcelona, Spain (2011)
14. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: SIGIR 2007, pp. 103–110 (2007)
15. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: IJCAI, pp. 448–453 (1995)
16. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management 24(5), 513–523 (1988)