Semantic Tagging of Places based on User Interest Profiles from Online Social Networks

Vinod Hegde, Josiane Xavier Parreira, and Manfred Hauswirth

Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland
{vinod.hegde,josiane.parreira,manfred.hauswirth}@deri.org

Abstract. In the recent years, location based services (LBS) on mobile devices have become very popular. With the growing number of smartphone users, the demand for services that can provide recommendation of places based on user location and interest has increased rapidly. The performance of such LBS depends on a number of factors, including how well the places are described. A number of location based services allow users to check-in at places, i.e. users can let others know of their whereabouts. Even though they also enable users to manually tag places they have visited, users rarely do so. Moreover, the available information attached to places (e.g. their names) is often ambiguous or insufficient for service providers to automatically generate tags. On the other hand, users often provide information about their interests in online profiles via online social networks. The common interests of a group of people that has visited a particular place can potentially provide further description for the place.
In this work we present an approach that automatically assigns semantic tags to places, based on interest profiles and check-in activities of users. The approach consists of: (i) an interest profile expansion algorithm to derive semantic concepts related to the user interests; (ii) a model to determine the probability that a particular semantic concept describes a place, based on the check-in activities of users; and (iii) a noise removal approach, using a hierarchical clustering technique, which is applied on the top probable semantic concepts to derive the final semantic tags for places. We have evaluated our approach with real world datasets from popular social network services, against a set of manually assigned tags. The experimental results show that not only we are able to automatically derive meaningful tags for different places, but also that the sets of tags assigned to places are expected to stabilise as more unique users check-in at places. This indicates that top probable tags derived can be consistently assigned to places irrespective of the number of people who have checked-in at those places.

Keywords: place tagging, recommendation systems, data mining, online social networks, location based services

1 Introduction

Mobile devices have never been so ubiquitous. Equipped with sophisticated sensors such as GPS sensors and cameras, they now enable a new range of location based services (LBS). These services determine the physical location of the user and provide a number of functionalities. For instance, users can check-in at places, i.e. users can let others know of their whereabouts. Check-in activities are already being explored to understand user behaviours for personalised advertising and promotion of

businesses [27,28,2,1]. Another functionality common in LBS is place recommendation: nearby locations are suggested to the user, by matching the description of the places with the user needs or interests. The performance of LBS recommendation depends on the richness of the geographic data used. This geographic data includes places or points of interests (POIs), comments, ratings about places and metadata about places such as tags. Some LBS allow users to manually assign any descriptive or categorical tags to places they have visited. By descriptive tags, we mean any short keywords which are semantically related to a place. For example, it would be appropriate to tag a *Computer Science Building* with tags such as *Software*, *Engineering*, and *Programming*. A categorical tag such as *Academic Building* for a place is much more abstract and less informative. Even though users often use LBS for check-in activities, they rarely tag a place. Currently, most of the places used by LBS are poorly tagged. A study on one such service showed that 30% of the places do not contain any tags [26]. Service providers can not automatically generate tags since the available information about the places is often ambiguous or insufficient. However, it has been noted that mobile phone users are the prominent consumers of LBS and their information requirements are highly dependent on their real time social and physical context [23]. In other words, users look for very specific information on the mobile phones based on their current physical location and social contexts. Therefore, there is an urgent need for techniques to assign semantically related tags to the places automatically so that search and recommendation can be more effective.

In a different context, many complex problems related to information generation on the Web have been solved utilising the wisdom of the crowd [10,4,16]. For example, in [6,21], various ways in which explicit or implicit information provided by the users can be utilised to enrich the information on the Web have been discussed. Web users leave their footprint on the Web using resources such as online social networks (OSNs) and microblogging systems, which can be used to derive the user's preferences and interests. Many of the users of OSNs also use location based services to check-in into places. Based on the above observations, the common interests of a group of people that has visited a particular place can potentially provide further description for the place.

In this work, we describe how the two sources of information combined – user interest profiles on OSNs and check-in logs – can be utilised to derive tags for a place. We present an approach that automatically assigns semantic tags to places, based on interest profiles and check-in activities of users. We first extract semantic concepts from the interest profiles of users available on OSNs. However, the interest profiles of users are often sparse and contain only a few keywords. We provide an interest expansion algorithm that discovers "hidden" interests by expanding the user interest profile in a controlled manner. The expansion algorithm is able to derive more concepts without deviating from the user interests. We provide a model to determine the probability that a particular semantic concept describes a place, based on the expanded interest profiles and check-in activities of users at places. We consider the top-k probable semantic concepts for any given place and perform a hierarchical clustering on those concepts to derive the final set of tags.

We have evaluated our algorithm with real world datasets from popular social networking services, against a set of manually assigned tags. We have also studied the nature of tag probability distributions against the check-in activities by users in order to understand the quality of the top probable tags and collective interests of people visiting places. The experimental results show that the automatically generated tags are

similar to the manually assigned tags, and also that the sets of tags assigned to places are expected to stabilise as more unique users check-in at places. This indicates that top probable tags derived can be consistently assigned to places irrespective of the number of people who have checked-in at those places.

The rest of the paper is organised as follows. We discuss the related work in Section 2. In Section 3, we present our probabilistic model to derive tags to places based on interest profiles and check-in activity of the users, as well as our interest profile expansion algorithm. In Section 4 we present an experimental evaluation and analysis of our approach. Section 5 concludes the paper and discusses future work.

2 Related Work

In the recent years, there has been an increased interest in the area of analysis and enrichment of geographic data. The amount of volunteered geographic information (VGI) is rising, as more users are equipped with sophisticated mobile devices which enable them to actively contribute with geographic data. [22,17] have studied various approaches to deriving and recommending tags to annotate images based on various types of user data. The work in [8] gives an example of how GPS traces and other geographic data provided by people can be used to create an accurate map of the world. In [24] various approaches that can be adopted for manually tagging places using mobile phones are discussed. Lin et al. [15] study the naming preferences of people regarding the places they visit and shows that such preferences depend on the context of the person naming a place. It also finds that on an average places have very few description names. All these works indicate that there is a need for obtaining and enriching geographic information and that the manual effort to generate such information is not enough. In [13], an automatic place naming technique based on user check-in activities is discussed. However, this deals with deriving only the names of the places while our approach provides descriptive tags for the places. Noulas et al. [18] provides a good example of the importance of semantic annotations, where they show that identification of user communities and comparison of urban neighbourhoods can be done using the annotations of places. In [26], the authors find that significant amount of places lack even the abstract textual descriptions and hence focus on deriving the categorical tags for place categories such as *restaurant* and *cinema*. Our work, on the other hand, focuses on deriving more descriptive tags. To the best of our knowledge, assigning places with automatically derived semantic tags has not been studied yet. Such methodology is much needed as users rarely assign specific tags to places and rich information is needed for search and recommendation of places.

3 Semantic Tagging of Places

Online social networks enable users to express their social interests and other personal information via their user profiles. In addition, location based social networks let users express their location information with check-in activities. In this section we describe how we use both the user interests listed in OSN profiles and the check-in activities of users to derive descriptive tags for places. We first present our probabilistic model for determining the probability that a given semantic tag describes a place, based on the interests of users that have visited the place. A hierarchical clustering technique is applied on the top probable semantic concepts to remove possible 'noise' tags and derive the final semantic tags for places.

It has been found that user profiles in OSNs have very few fields under various categories such as *work*, *interests*, and *education* and have considerable textual descriptions which are complex to analyse [11,29]. We present an interest expansion algorithm that removes ambiguous concepts and expands the initial set of users interests. The expansion is done in a way to derive hidden related concepts, without deviation from the initial interests.

3.1 Probabilistic Model for Deriving Tags for a Place

Our probabilistic model considers the check-in activities of users and their interests to derive the most probable tags for a place. Let U denote the set of all users who check-in at places and let P denote the set of all places (or POIs) the users can check-in. A user check-in is modelled as a tuple of the form $\langle u, t, p \rangle$, where $u \in U$, $p \in P$ and t is the timestamp of the check-in activity. The set of all user check-ins is denoted by CH. From CH we can extract CH_{ip}, which is total the number of check-ins of user i user at place p, and CHU_p which is the set of users who have checked in at least once at place p. The set of concepts in the interest profile of user i is given by K_i.

When the i^{th} user checks in at p, we consider each concept in K_i as candidate tag for p. We do so with the hypothesis that there is a possible semantic relationship between a place and any concept in the interest profile of the person checking in at that place. The check-in action by any users at p contributes to the expansion of the candidate tag set CT_p which is defined as $CT_p = \bigcup_i K_i$ where $i \in CHU_p$. Given the p^{th} POI, the probability that p is checked in by i^{th} user is given by

$$Pr(U_{ip}) = \frac{CH_{ip}}{\sum_j CH_{jp}}$$

where $i \in CHU_p$ and $\sum_j CH_{jp}$ is the total number of check-ins by all users at p.

The conditional probability that i^{th} user with n concepts in K_i attaches one of the concepts k_j as tag to a POI is given by

$$Pr(k_j|U_{ip}) = \frac{1}{n} \qquad \forall p \in P, k_j \in K_i.$$

This is with the assumption that all concepts in a user interest profile equally represent the interests of a user. The total probability that the p^{th} POI is attached with the concept k_j as a tag is given by

$$Pr(k_j) = \sum_i Pr(k_j|U_{ip}) Pr(U_{ip}) \qquad \forall k_j \in CT_p.$$

We call this the *Tag Probability* of the concept k_j. It is easy to see that $\sum_j P(T_p = k_j) = 1$ and $0 < P(T_p = k_j) \leq 1$ where $k_j \in CT_p$. This means that a categorical random variable T_p defines the probability distribution of the tags for the place p where the sample space $\Omega = CT_p$. We can see that a random variable T_p^n can be defined by considering the check-in activities of the first n unique users at place p (denoted by CT_p^n) with the sample space $\Omega = CT_p^n$. In our model to derive tag probabilities, concepts in the interest profile of a frequent visitor are considered as more probably related to the corresponding place.

3.2 Hierarchical Clustering of Top Probable Tags

The work in [19] successfully employs hierarchical clustering to obtain clusters of interests from interest profiles of users without considering the geographical aspects of users. In our approach to deriving semantic tags for places, though we derive the top

probable tags, not every tag derived needs to be semantically related to the corresponding place. This observation demands clustering of the tags so that we could obtain one or more "natural" clusters of tags to tag a place and discard unrelated tags which are noise. Hierarchical clustering is one of the widely used clustering method for efficient clustering and [9] lists various techniques and advantages of hierarchical clustering.

We compute the semantic similarity between the tags and use the agglomerative nesting algorithm with the group average method as it is one of the best methods for clustering documents [3]. Determining the number of clusters given a set of elements is a well known problem and various techniques for deriving the appropriate number of clusters have been proposed. In [12] a novel method for cutting the dendrogram obtained from hierarchical clustering to obtain clusters is discussed. We used this method to obtain the clusters of tags corresponding to each random variable for each place. Any tag that does not fall into the generated clusters is then discarded. As we will show in the next section we use Wikipedia [1] concepts as probable tags. Therefore we used the Wikipedia Link Vector Model (WLVM) [25] to obtain the semantic similarity between the tags.

3.3 Interest Profile Expansion Algorithm

Users describe themselves and their interests on online social networking profiles. Such profiles are a great source of information about the user, but they often contain only few short textual snippets or keywords. Many of the field values are textual descriptions which are inherently ambiguous and complex to analyse. Our expansion algorithm uses Wikipedia to disambiguate and expand the interest profile of a user. Wikipedia is a vast repository of knowledge constantly updated and refined by a large user community. It has the advantage that all the concepts defined are rich in their article content with numerous links to related concepts. The concepts and the links between them form the Wikipedia graph structure where concepts represent the nodes and links represent the edges. We use the term concept and node interchangeably in the work. In order to get a disambiguated user profile, we retain only those keywords which match to a single Wikipedia concept and discard remaining keywords so that a modified user profile contains unambiguous concepts.

Next we apply our user interest profile expansion algorithm to expand the disambiguated profile. The algorithm considers the fact that a Wikipedia concept can be associated with its *related* concepts based on the links in its article content on Wikipedia. The algorithm also takes into account the fact that concepts with a large number of inlinks from other concepts tend to be more general [5] and hence does not include such concepts in the expansion. This ensures that general concepts such as *Education* and *United States* which have high indegree are not present in the expanded profile and hence not used as tags for places.

Algorithm 1 describes how the expansion is done. It considers each concept in the user profile and attempts to expand it in a depth first manner. The parameters R and R_{glob} control the expansion of any node by limiting the number of nodes that can be expanded. The parameter $Indeg_{threshold}$ defines the maximum number of inlinks that a concept can have so that its not considered to be a general concept. The $distance$ function computes the shortest distance between any two concepts which is the minimum number of links to be traversed from one concept to the other in the Wikipedia graph

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

structure. The set of neighbour nodes which would be expanded from a given node is decided by the proximity of those nodes to the nodes in W. The measure of proximity of a node u is stored in $r[u]$ as seen in the algorithm. For a given node, the algorithm only expands those nodes that are closest to the set of nodes in W. This ensures that only those nodes more related to the original interests of a user are expanded further.

A node v_i is expanded only if $\prod_{k=i-1}^{0} \frac{1}{outdegree(v_k)} \geq R_{glob}$ where i is the height of the node v_i in the expansion tree and v_{i-1},v_{i-2},$\ldots v_0$ represent the ancestors of v_i in the expansion tree. During the expansion j^{th} node v_{ij} at height i, at most N_{ij} neighbours are added to the expansion list which are at unit distance from v_{ij} in the Wikipedia graph. At most k nodes are considered for expansion from any given node. So, the maximum number of nodes added due to the expansion of a node is $M_0 + M_1 + M2 \ldots + M_h$ or $O\big(\sum_{i=0}^{h} M_i\big)$, where $M_i = \sum_j N_{ij}$ and h is the maximum height possible for all the non-leaf nodes in the expansion tree. For any M_i, neighbours of at most k^i nodes are considered.

The result from the interest profile expansion algorithm for a user i corresponds to the set K_i in the probabilistic model. In the next section, we evaluate how both approaches combined can provide meaningful descriptive tags for places.

Algorithm 1 Interest Profile Expansion

```
function ExpandProfile(W)
    U ← φ
    for all c ∈ W do
        AddNode(c,1,W);
    end for
end function
function AddNode(v, R, W)
    if R ≥ R_glob then
        N ← {u | dist(u, v) = 1}
        for all u ∈ N do
            if indegree(u) < Indeg_threshold then
                for all c ∈ W do
                    r[u] ← r[u] + distance(c, u) + distance(u, c);
                end for
                add(u,U);
            end if
        end for
        for all t ∈ TopKNeighbor(r) do
            AddNode(t, R * 1/|outdegree(v)|, W);
        end for
    else
        return;
    end if
end function
```

4 Experimental Evaluation

We have performed an experimental evaluation in order to verify the effectiveness of our approach. We first describe the real world datasets used in the experiments and then present the results of our evaluation. The evaluation is divided into different parts. We report on the expansion algorithm, the parameters used and the distribution of the profile sizes. We show how the assigned tags evolve with the increasing number of user check-in activities and how they compare to a set of manually assigned tags. Finally, we analyse the nature of the tag probability distributions which indicates that the set of automatically generated tags is expected to stabilise with the increasing number of unique user check-ins.

4.1 Datasets Description

We collected data from Foursquare[2] for over one million random places in UK, USA and Ireland between June and July 2012, to check how well the places are described. Only 7% of the places had any descriptive tags and only 21% of the places had any tips/comments in the form of short text snippets, which again confirmed the lack of rich description of places.

We then collected the Facebook [3] and Foursquare user profiles of 104 volunteers residing in the city of Galway, Ireland. These were random users as we requested people to participate through various social media and announced prizes for their contribution. The social interests of the users were obtained from their Facebook profiles by extracting the text in the fields corresponding to hometown, interests, activities, education, work, and events. We have found that interest profiles were sparse in terms of the keywords and our observations are indeed similar to the figures stated in [11,29]. The size of the user profiles in terms of number of keywords can be fit with a Poisson distribution using a *Maximum Likelihood Estimation* (MLE) (n = 104, λ = 362.1, $S.E$ = 1.9) as shown in Figure 1(a). We have obtained check-in activities from both Foursquare and Facebook profiles of the volunteers. The check-in activity data contains 4476 records of check-ins of users which they had generated using their Facebook and Foursquare mobile applications. There are 1633 unique places where users had checked-in and 215 places where at least 2 users had checked-in.

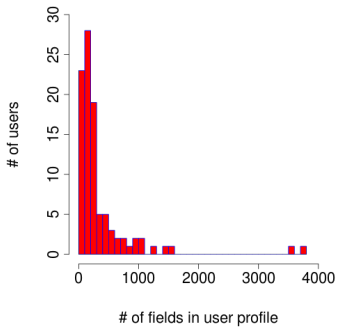
4.2 Evaluation

Interest Profile Expansion Algorithm We first disambiguated the interest profiles and found that 20% of the keywords in user profiles matched to an exact Wikipedia concept. For generating the values assigned to the different variables in the expansion algorithm we have proceed as follows: We have sorted the concepts by the number of inlinks to them and manually inspected many of the top concepts. This has shown that indeed such concepts were very general in nature. Since we have not found any formal approaches to decide the generality of Wikipedia concepts, we discarded top 1% of the concepts and obtained the statistics for the inlinks of the remaining concepts. All the remaining concepts had very few inlinks (n = 3537875, min = 0, max = 221, $mean$ = 9.274). Hence we set the value of $Indeg_{threshold}$ to 221 which ensured that nodes with more than 221 inlinks were not added during expansion. We set the expansion controller variable R_{glob} to $1/100$ which meant that a concept is expanded only if it has no more than 100 ancestors considered during the expansion. The expansion algorithm considerably enriched the user interest profiles with related concepts in Wikipedia. The expanded user interest profiles were significantly larger compared to their original size and we could fit the size with Poisson distribution using MLE (n = 104, λ = 3843.835, $S.E$ = 6.295285) as shown in Figure 1(b).

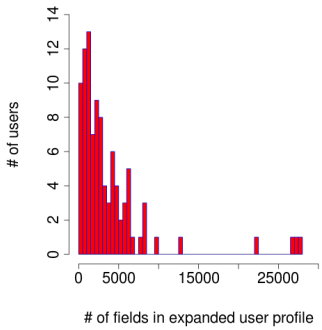
Automatic Semantic Tagging Results For the automatic semantic tagging we have considered only those places which were checked-in by at least 2 users. For each place p, we have computed the random variable CT_p^n by incrementally considering the unique users who had checked-in at p. This process defined CHU_p number of random variables corresponding to tag probabilities for p. We then applied the hierarchical clustering

[2] https://foursquare.com/

[3] http://www.facebook.com/

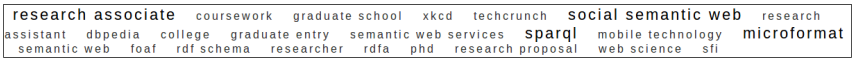


(a) Number of keywords before expansion.

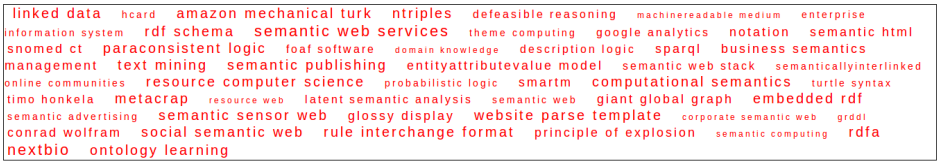


(b) Number of semantic concepts after expansion.

Fig. 1: Sizes of interest profiles before and after expansion.



(a) Manually assigned tags.



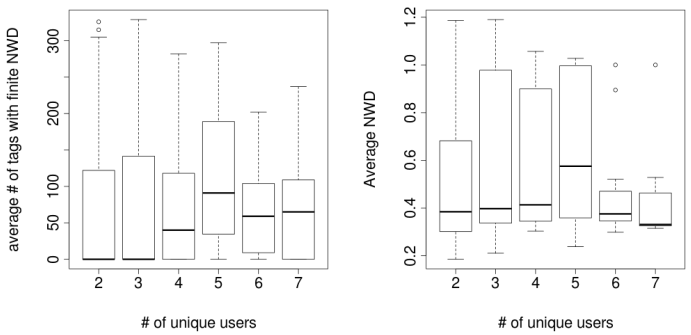
(b) A cluster of tags derived from top probable tags after check-ins by 10 users.

Fig. 2: Manual and derived tags assigned to Digital Enterprise Research Institute

method to obtain the clusters of tags corresponding to each random variable for each place.

In order to evaluate the quality of the derived tags, we have used a set of manual tags assigned by volunteers as ground truth. Seven volunteers manually tagged the places they knew among the places in the collected check-in records. They tagged a total of 25 unique places with multiple tags (mean number of tags per place = 22.96). Manual inspection of automatically derived tags and manually assigned tags revealed that most of the tags in such clusters were highly related to the places under consideration, though users had not tagged places with the derived tags. Figure 2 shows both the manual and automatically derived tags for Digital Enterprise Research Institute (DERI), a Semantic Web research institute. Frequent manual tags and the most probable tags are shown in larger fonts. We can see that though automatically derived tags are not exactly the same as the manual tags, they are good candidate tags for DERI.

Automatic Semantic Tagging Evaluation For a systematic evaluation of the generated tags we have measured the Normalised Web Distance [7] between tags and the place names. Normalised Web Distance (NWD) has been extensively used to obtain the



(a) Number of derived tags with finite NWD.

(b) Average NWD scores of tags.

Fig. 3: Variation in the Normalised Web Distance scores against the number of unique users.

semantic relatedness between any two strings, where the extensive data on the Web is used. Formally, the NWD between any two strings x and y is given as

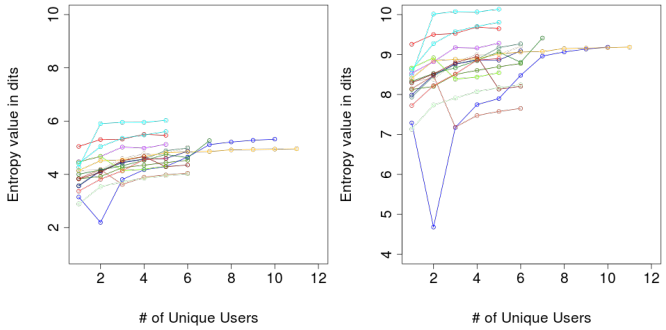
$$d_{nwd}(x,y) = \frac{max\{logf(x), logf(y)\} - logf(x,y)}{logN - min\{logf(x), logf(y)\}}$$

where $f(x)$ is the number of Web pages containing the string x, $f(y)$ is the number of Web pages containing the string y, $f(x,y)$ is the number of pages where both x and y appear, and N is the total number of pages indexed by a specific search engine.

We first analyse how different users visiting a place affect the set of generated tags. For each random variable T_p^n, we have computed the d_{nwd} between the top 350 automatically derived tags and place names using the index provided by *Yahoo*[1]. It is possible that some tags have an infinite NWD to a place, which were considered as invalid and discarded. Figure 3(a) shows the box plot of number of valid tags, i.e. tags with a finite NWD, over all places. Please note that, for instance, for the case of 6 users, only places which have at least 6 distinct users were considered. We can see that the more unique users check-in at places, the more valid tags are generated.

We then computed the values of d_{nwd} between place names and the manually assigned tags to compare the performance of our semantic tagging technique. The five-number summary of d_{nwd} between manual tags and place names is (min=0.0000, $Q1$=0.1216, $median$=0.3032, $Q3$=0.8732, max=1.9030) with $mean$=0.4730. The five-number summary of d_{nwd} between automatic tags and place names, considering all check-in activities, is (min=0.0000, $Q1$=0.2053, $median$=0.5340, $Q3$=1.0000, max= 3.4930) with $mean$=0.5719. This shows that automatic tags exhibited d_{nwd} values comparable to those of the manual tags. The Welch's t-test showed that mean value of d_{nwd} for automatic tags is greater than that of manual tags with 95% confidence interval of (0.053, 0.144) where H_A is that true difference in means is not equal to 0. This means that on an average, the d_{nwd} scores obtained by automatically derived tags are not much higher than the ones obtained by the manual tags.

[1] http://developer.yahoo.com/search/boss/



(a) Entropy values with unexpanded interest profiles.

(b) Entropy values with expanded interest profiles.

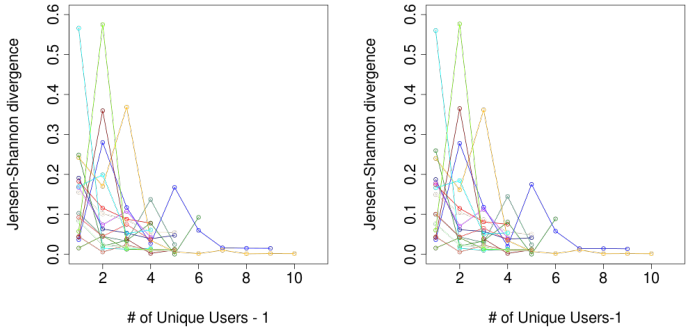
Fig. 4: Entropy values observed over the tag probability distributions w.r.t. the number of unique visitors.

Figure 3(b) shows the average values of d_{nwd} for the valid tags obtained against the number of unique users. We can see that in spite of more unique users visiting a place, the average scores of d_{nwd} obtained by the tags remain close to the ones achieved by manually assigned tags.

We noted that we could derive an average of 158 tags for places with expanded user profiles whereas we could derive 51 tags with unexpanded profiles. We also observed that only 9% tags obtained from expanded interest profiles had infinite values of d_{nwd} against places whereas this was 17% for the unexpanded user profiles. This clearly indicated the advantages of carefully expanding the concepts in user profiles and using them as probable tags. Clustering the top probable tags obtained from expanded user profiles showed that 30% of the tags belonged to some cluster and were related to each other and only 2% of the tags had infinite normalised web distance. 70% of the tags did not belong to any cluster and were not related to each other and 8% of such tags had infinite normalised web distance. This showed that clustering the tags fetched tags related to each other and to the place thereby removing any 'noise' tags among the top probable ones.

Nature of the tag probability distributions We have studied the nature of the tag probability distributions of a place over the number of unique visitors of that place. We considered only those places which had been checked-in by at least 5 distinct users to study the variation in the tag probability distributions. We have computed the entropy [20] to analyse the information content or *randomness* of tag probability distributions, and we have used Jensen-Shannon divergence [14] to analyse the variations among tag probability distributions.

We depict the variation in entropy of T_p^n when unexpanded user profiles are considered in Figure 4(a). Figure 4(b) shows the variation in entropy when expanded user profiles are considered. We see that the increase in the entropy values is lesser after more unique users check-in. This indicates that the information content of T_p does not increase in spite of increased sample space and stabilises with the number of unique users visiting place p. It also implies that some of the semantic tags become more probable and thereby reduce the entropy in spite of increased sample space.



(a) Jensen-Shannon divergence values with unexpanded interest profiles.

(b) Jensen-Shannon divergence values with expanded interest profiles.

Fig. 5: Jensen-Shannon divergence w.r.t. the number of unique visitors.

We computed the Jensen-Shannon divergence between T_p^n and T_p^{n+1}. We show how the divergence value diminishes based on the number of unique users in Figure 5(b) when expanded user profiles are considered. Interestingly, the divergence values obtained for the random variables when expanded profiles were used are very similar to the ones corresponding to the unexpanded profiles and are shown in Figure 5(a). This indicated that in spite of considering various interests of users to derive tag probability distributions of a place, such distributions showed high dependence as interests of more users were considered.

5 Conclusion and Future Work

In this work, we have presented an algorithm to automatically derive descriptive semantic tags for places, based on users' interests found in online profiles and their check-in activities. Specifically, we derived from each user a set of concepts based on the user interests, using our interest profile expansion algorithm. The sets are used in our probabilistic model together with the hierarchical clustering techniques to derive a set of tags for a place, based on the users that have visited the place. We performed an experimental evaluation that shows that not only we are able to automatically derive meaningful tags for different places, but also that the sets of tags assigned to places are expected to stabilise with the increasing number of user check-ins. In the future work, we plan on obtaining larger datasets to validate our findings rigorously, and we will also consider other online sources of user data, such as Twitter.

Acknowledgements: *This research has been supported by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Líon-2).* Also, we thank Dr.Milovan Krnjajić for providing useful feedback about the paper.

References

1. Berjani, B., Strufe, T.: A recommendation system for spots in location-based online social networks. In: Proceedings of the 4th Workshop on Social Network Systems. ACM (2011)
2. Cheng, Z., Caverlee, J., Lee, K., Sui, D.: Exploring millions of footprints in location sharing services. AAAI ICWSM (2011)
3. El-Hamdouchi, A., Willett, P.: Comparison of hierarchic agglomerative clustering methods for document retrieval. The Computer Journal 32(3) (1989)

4. Fuxman, A., Tsaparas, P., Achan, K., Agrawal, R.: Using the wisdom of the crowds for keyword generation. In: WWW. ACM (2008)
5. Gabrilovich, E., Markovitch, S.: Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research 34(1) (2009)
6. Goodchild, M.: Citizens as sensors: web 2.0 and the volunteering of geographic information. GeoFocus (Editorial) 7 (2007)
7. Gracia, J., Mena, E.: Web-based measure of semantic relatedness. WISE (2008)
8. Haklay, M., Weber, P.: Openstreetmap: User-generated street maps. Pervasive Computing, IEEE 7(4) (2008)
9. Johnson, S.: Hierarchical clustering schemes. Psychometrika 32(3) (1967)
10. Kittur, A., Chi, E., Pendleton, B., Suh, B., Mytkowicz, T.: Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. World Wide Web 1(2), 19 (2007)
11. Lampe, C., Ellison, N., Steinfield, C.: A familiar face (book): profile elements as signals in an online social network. In: SIGCHI. ACM (2007)
12. Langfelder, P., Zhang, B., Horvath, S.: Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. Bioinformatics 24(5) (2008)
13. Lian, D., Xie, X.: Learning location naming from user check-in histories. In: SIGSPATIAL. ACM (2011)
14. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory 37(1) (1991)
15. Lin, J., Xiang, G., Hong, J., Sadeh, N.: Modeling people's place naming preferences in location sharing. In: Proceedings of the 12th ACM international conference on Ubiquitous computing. ACM (2010)
16. Mendes, P., Passant, A., Kapanipathi, P.: Twarql: tapping into the wisdom of the crowd. In: Proceedings of the 6th International Conference on Semantic Systems. p. 45. ACM (2010)
17. Moxley, E., Kleban, J., Manjunath, B.: Spirittagger: a geo-aware tag suggestion tool mined from flickr. In: Proceeding of the 1st ACM international conference on Multimedia information retrieval. ACM (2008)
18. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. SMW (2011)
19. Paolillo, J., Wright, E.: Social network analysis on the semantic web: Techniques and challenges for visualizing foaf. Visualizing the semantic web 2 (2005)
20. Shannon, C., Weaver, W., Blahut, R., Hajek, B.: The mathematical theory of communication, vol. 117. University of Illinois press Urbana (1949)
21. Sheth, A.: Citizen sensing, social signals, and enriching human experience. IEEE Internet Computing 13(4) (2009)
22. Sigurbjörnsson, B., Van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: WWW. pp. 327–336. ACM (2008)
23. Sohn, T., Li, K., Griswold, W., Hollan, J.: A diary study of mobile information needs. In: SIGCHI. ACM (2008)
24. Wang, J., Canny, J.: End-user place annotation on mobile devices: a comparative study. In: CHI extended abstracts on Human factors in computing systems. ACM (2006)
25. Witten, I., Milne, D.: An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy (2008)
26. Ye, M., Shou, D., Lee, W., Yin, P., Janowicz, K.: On the semantic annotation of places in location-based social networks. In: SIGKDD. ACM (2011)
27. Ye, M., Yin, P., Lee, W., Lee, D.: Exploiting geographical influence for collaborative point-of-interest recommendation. In: SIGIR (2011)
28. Yu, C., Chang, H.: Personalized location-based recommendation services for tour planning in mobile tourism applications. E-Commerce and Web Technologies (2009)
29. Zhao, S., Grasmuck, S., Martin, J.: Identity construction on facebook: Digital empowerment in anchored relationships. Computers in Human Behavior 24(5) (2008)