Computing Similarity between Users on Location-Based Social Networks

Soha Mohamed*†, Alia I. Abdelmoty*
*School of Computer Science & Informatics, Cardiff University, Wales, UK
Email: {AlySA, AbdelmotyAI}@cardiff.ac.uk
† Faculty of Computers and Information, Helwan University, Cairo, Egypt

***Abstract*—With the current trend of embedding location services within social networks, an ever growing amount of users' spatiotemporal tracks are being collected. These tracks can be used to generate user profiles to reflect users' interests in places. User-contributed annotations of places, as well as other place properties, add a layer of important semantics that if considered, can result in more refined representations of the users? profiles. In this paper, we study how such place-oriented profiles can be used to represent similarity between users of Location-Based Social Networks (LBSN). Spatial as well as semantic dimensions of the user-provided information are used within a folksonomy data model to represent relationships between users, places and tags. The model allows simple co-occurrence methods and similarity measures to be applied to build different views of personalized user profiles. Basic profiles capture direct user interactions, while enriched profiles offer an extended view of user's association with places and tags that takes into account relationships in the folksonomy. The main contribution of this work is the demonstration of how the different data dimensions captured on location-based social networks can be combined to represent useful views of user profiles and to compute similarity between users.**

***Index Terms*—GeoFolksonomy; User Profiles; Location-based Social Networks.**

I. INTRODUCTION

This work focuses on Location-Based Social Networks (LBSN) that collect information on users' interests in physical places in the real world. By "switching on" location on devices, we are giving away information on our whereabouts, our daily routines, activities, experiences, and interests. Thus, in comparison to other personal information, location data are possibly the most crucial type of data of relevance to privacy, as it pulls together our virtual and physical existences and thus raises critical questions about privacy in both worlds. This work proposes methods for constructing user profiles using different dimensions of the data captured from users on LBSN [1], and demonstrates how these profiles can be used to measure different aspects of similarity between users.

Studying user similarity from LBSN data is useful, as information available about users, their locations and activities are normally sparse. User similarities can be exploited to predict types of activities and places preferred by a user based on those of users with similar preferences. So far, previous works have studied data produced from LBSN from the point of view of enhancing the services provided by these networks, namely, for point of interest (POI) recommendations. There, the question of concern is to find places of interest to a user based on their history of visits to other places and their general interaction with the social network. Most works relied mainly on the spatial dimension of user data [2], with some works more recently exploring the relevance of the social and content data dimensions on these networks [3]. However, data dimensions are normally treated separately, or their outputs are combined in fused models.

In this paper, both semantic and spatial interactions of users are used to project distinct and complementary views of personalised user profiles. Thus, user's annotations on places they visit are compiled in semantic profiles, while collective user annotations on places are used to create specific profiles for places that encapsulate user's experiences in the place. Place profiles, in turn, are used to construct personalised user profiles. In comparison to previous works in the area of recommendations, LBSN data are treated as folksonomies of users, places and tags. User annotations in the form of tips, their interaction with places, in the form of check-ins, as well as general place properties, namely, place categories and tags, are analysed concurrently to extract relations between the three elements of the folksonomy. Simple co-occurrence methods and similarity measures are used to compute direct and enriched user profiles.

Similarity between users can then be computed using the different views of user profiles; using their direct interactions with the social network or extended with a holistic view of other users' interaction with the network in different regions of geographic space. Previous works attempting a similar approach used matrix factorization techniques to handle the multiple data dimensions, but did not consider the use of the range of content data as used in this paper. Sample realistic data from Foursquare are used to demonstrate the approach and evaluation results show its potential value.

The main contributions of this work can be summarized as follows: 1) Collecting users' direct feedback on venues from LBSNs. Users' interaction on LBSNs can be regarded as user feedback on geographic places they visited and interacted with. User's visits to places are recorded along with their comments and tags. 2) Modelling different levels of user profiles extracted from the heterogeneous user feedback in LBSNs. User-generated traces at venues in LBSNs include both spatial and implicit semantic content. The location traces are treated equally to the semantic traces inferred from their interaction with the place through tagging and tipping. Collective behaviour of users on the network are also used to understand the place characteristics and these in turn are further used in the modelling of user profiles. 3) Similarity between users on LBSN is approached in a uniform manner within the proposed framework, thus providing means of computing spatial, semantic or a combined view of user similarity on these networks. 4) Evaluation experiments are carried out using samples of realistic data sets for a representative number of users with different levels of usage of the LBSN.

The rest of the paper is organised as follows. Section II provides an overview of related work and approaches. In Section III, a geo-folksonomy data model for LBSN is introduced and in Section IV different types of user profiles are defined using this model. Section V describes the approaches to computing similarity between users and detailed evaluation experiments are reported in Section VI. The paper concludes in Section VII with an overview of future work.

II. Related Work

Works on modelling user data in LBSN mainly consider two problems; a) place (or point of interest) recommendation, and b) user similarity calculation. Different types of data are used by different approaches, namely, geographic content, social content as well as textual annotations made by users. Also, different methods are used in analysing the data, for example, distance estimations for geographic data modelling and topic modelling for annotation data analysis. In the area of POI recommendation, works range from generic approaches that uses the popularity of places [4] to recommendation methods that are based on user's individual preferences [5]. A useful survey of these approaches can be found in [6].

Based on check-in data gathered through Foursquare, Noulas and Mascolo [7] exploit factors such as the transition between types of places, mobility between venues and spatiotemporal characteristics of user check-in patterns to build a supervised model for predicting a user's next check-in. Ye, Lui and Lee [5] investigated the geographical influence with a power-law distribution. The hypothesis is that users tend to visit places within short distances of one another. Other works considered other distance distribution models [8]. Gao, Tang and Liu [9] considered a joint model of geo-social correlations for personalized POI recommendation, where the probability of a user checking in to a new POI is described as a function of correlations between user's friends and non-friends close to, and distant from a region of interest. Liu, Xiong and Papadimitriou [10] approached the problem of POI recommendations by proposing a geographical probabilistic factor model that combines the modelling of geographical preference and user mobility. Geographical influence is captured through the identification of latent regions of activity for all users of the LBSN reflecting activity areas for the entire population and mapping the individual user mobility over those regions. Their model is enhanced by assuming a Poisson distribution for the check-in count which better represents the skewed data (users visiting some places one time, while other places 100s of times). Whilst providing some useful insights for modelling the spatial dimension of the data, the above works do not consider the semantic dimension of the data.

Correlations between geographical distance and social connections were noted in [11] [3]. Techniques of personalized POI recommendation with geographical influence and social connections mainly study these two elements separately, and then combine their output together within a fused model. Social influence is usually modeled through friend-based collaborative filtering [12] [5] [13] with the assumption that a user tends to be friends with other users who are geographically close to him, or would want to visit similar places to those visited by his friends. Ying, Lu, Kuo, and Tseng [14] proposed to combine the social factor with individual preferences and location popularity within a regression-tree model to recommend POIs. The social factor corresponds to similar users; users with common check-ins to the user in question. In this paper, we also use this factor when extending user profiles to represent places of interest within the region of user activity.

More recently, the importance of content information for POI recommendation was recognised. Two types of content can be considered, attributes of places and user-contributed annotations. Place categories are normally used as an indication of user activity, thus a user visiting a French restaurant would be considered as interested in French food, etc. User annotations in the form of tips and comments are analysed collectively to extract general topics to characterise places or to extract collective sentiment indications about the place. Examples of works that considered place categories are [15] [16] [17] [18]. In [15] [16], Latent Dirichlet Allocation (LDA) model was used to represent places as a probability distribution over topics collected from tags and categories or comments made in a place and similarly aggregate all tips from places a user has visited to model a user's interest. Aggregation was necessary as terms associated with a single POI are usually short, incomplete and ambiguous. [17] on the other hand modelled topics from tweets and reviews from Twitter and Yelp, and assumed that the relations between user interests and location are derived from the topic distributions for both users and locations. In [18], a probabilistic approach is proposed that utilize geographical, social and categorical correlations among users and places to recommend new POIs from historical check-in data of all users. In this paper, we also model user's association to place through the place's relation to tags, but add the influence of other users relations in the place to the equation. Aiming at improving the effectiveness of location recommendation, Yang, et al [19] proposed a hybrid user POI preference model by combining the preference extracted from check-ins and text-based tips which were processed using sentiment analysis techniques. Sentiment analysis is an interesting type of semantics which we do not consider in this work, but can be incorporated in future work.

So far, most works on user similarity mainly focused on structured, e.g., geographic coordinates, or semi-structured, e.g., tags and place categories, data. Recently, Lee and Chung [20] presented a method for determining user similarity based on LBSN data. While the authors made use of check-in information, they concentrated on the hierarchy of location categories supplied by Foursquare in conjunction with the frequency of check-ins to determine a measure of similarity. Mckenzie, Adams, and Janowicz [16] suggest exploring unstructured user-contributed data, namely tips provided by users. A topic-modelling approach is used to represent users' interests in places. Venues (places in Foursquare) are described as a mixture of a given number of topics and topic signatures are computed as a distribution across venues. User similarity can then be measured by computing a dissimilarity metric between users' topic distribution. Their method of modelling venues is interesting, but it limits the representation of user profiles, where profiles are based on generated topics derived from collective user annotation on places. Thus, individualised association of users with the place is somewhat ignored. In contrast to the above approach, our model does not assume constraints on the number of topics represented by the tags, but combines the individual's association with both tags and place in the creation of user profiles.

Social links between users have also been widely utilized to improve the quality of location-based recommender systems, since the social friends are more likely to share common interests on POIs than strangers. Most current works derive the similarities between users from social links and put them into the traditional memory-based or model-based collaborative filtering techniques. For example, some literature [8], [13], [21]–[23] seamlessly integrated the similarities of users into the user-based collaborative filtering techniques, while others [19], [24], [25] employed the user similarities as the regularization terms or weights of latent factor models.

III. The Geo-Folksonomy Model

The location-based social networking platform Foursquare was used as our source of data. It holds a large number of crowdsourced venues ($>$ 65 million places) from a user population estimated recently to around 55 million users. As the application defines it, a venue is a user-contributed "physical location, such as a place of business or personal residence." Foursquare allows users to check in to a specific venue, sharing their location with friends, as well as other online social networks, such as Facebook or Twitter. Built with a gamification strategy, users are rewarded for checking in to locations with badges, in-game points, and discounts from advertisers. This game-play encourages users to revisit the application, compete against their friends and contribute check-ins, photos and tips. Tips consist of user input on a specific venue, normally describing a recommendation, experience or activity performed in the place.

In this work, we use a folksonomy data model to represent user-place relationships and derive tag assignments from users' actions of check-ins and annotation of venues. In particular, tags are assigned to venues in our data model in two scenarios as follows.

1) A user's check-in results in the assignment of place categories associated with the place as tags annotated by this user. Thus, a check-in by user u in place r with the categories (represented as keywords) x, y and z, will be considered as an assertion of the form $(u, r, (x, y, z))$. This in turn will be transformed to a set of triples $\{(u, r, x), (u, r, y), (u, r, z)\}$ in the folksonomy.
2) A user's tip in the place also results in the assignment of place categories as tags, in addition to the set of keywords extracted from the tip. Thus, in the above example, a tip by u in r with the keywords $(t_1, \cdots, t_n)$, will be considered as an assertion of the form $(u, r, (x, y, z, t_1, \cdots, t_n))$, and is in turn transformed to individual triples between the user, place and tags in the folksonomy.

The process of extracting keywords from tips is done by tokenizing the tip into a set of words (terms) on white space and punctuation. Then we remove all words with non-latin characters and stop words. The output is a set of single words (term vector). Furthermore, we use Wordnet syntactic category and logical groupings for classifying the extracted terms. For example, Wordnet 'noun.act' category is used to filter action verbs and nouns to describe a user- or place- associated activity (ex. swimming, buying or eating).

The data capturing process results in the creation of a *geo-folksonomy*, which can be defined as a quadruple $\mathbb{F} := (U, T, R, Y)$, where U, T, R are finite sets of instances of users, tags and places respectively, and Y defines a relation, the tag assignment, between these sets, that is, $Y \subseteq U \times T \times R$, [26] [27].

A geo-folksonomy can be transformed into a tripartite undirected graph, which is denoted as folksonomy graph $\mathbb{G}_{\mathbb{F}}$. A geo-Folksonomy Graph $\mathbb{G}_{\mathbb{F}} = (V_{\mathbb{F}}, E_{\mathbb{F}})$ is an undirected weighted tripartite graph that models a given folksonomy $\mathbb{F}$, where: $V_{\mathbb{F}} = U \cup T \cup R$ is the set of nodes, $E_{\mathbb{F}} = \{\{u, t\}, \{t, r\}, \{u, r\} | (u, t, r) \in Y\}\}$ is the set of edges, and a weight w is associated with each edge $e \in E_{\mathbb{F}}$.

The weight associated with an edge $\{u, t\}, \{t, r\}$ and $\{u, r\}$ corresponds to the co-occurrence frequency of the corresponding nodes within the set of tag assignments Y. For example, $w(t, r) = |\{u \in U : (u, t, r) \in Y\}|$ corresponds to the number of users that assigned tag t to place r.

Figure 1 depicts the overall process of user profile creation. The process starts with data collection of check-ins and tip data from Foursquare, that are then processed to extract users, places and tags and their associated properties. The modelling stage includes the definition of relationships between the three entities and the application of folksonomy co-occurence methods to extract the different types of profiles. Place and tag similarity calculations are used to further extend the basic profiles to build different views of enriched user profiles.

IV. User Modelling Strategies

We propose an approach to modelling users in LBSN that represents a user's spatial, semantic and combined spatio-semantic association with place. A spatial user profile represents the user's interest in places, while a tag-based profile describes his association with concepts associated with places in the folksonomy model. A spatio-semantic profile describes the user specific interest in certain concepts associated with places in his profile. A user profile is built in stages. Starting with a basic profile that utilises direct check-in and annotation histories, a user profile is then extended by computing the

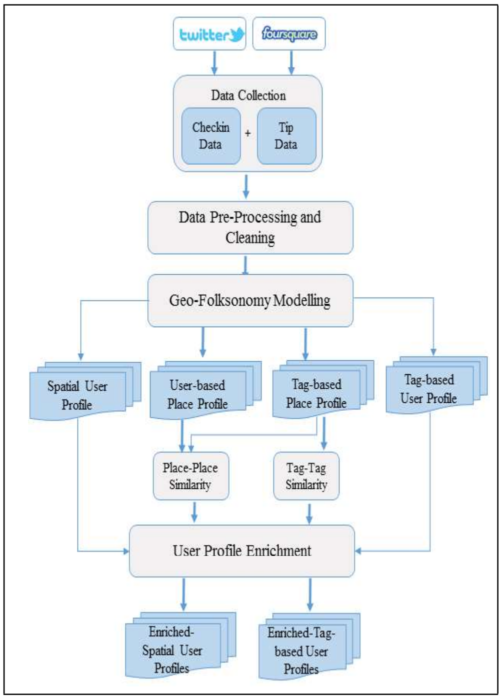


Fig. 1. The framework of the proposed system.

relationship between places and concepts derived from the collective behaviour of other users in the dataset. A basic profile represents actual interactions with places, while the extended profile describes “recommended” associations given overall interactions between users, places and concepts in the dataset. We are able to model such interactions separately in the extended profile by controlling the similarity function used to create the profile. For example, we can focus on modelling the types of places visited by the user or take into account visit behaviour of other users whose profiles overlap with the user, as discussed below.

A. Basic User Profiles

Definition 1: **Spatial User Profile** A spatial user profile $P_R(u)$ of a user u is deduced from the set of places that u visited or annotated directly.

$$\begin{aligned} P_R(u) =& \{(r, w(u,r)) | (u,t,r) \in Y, \\ & w(u,r) = |\{t \in T : (u,t,r) \in Y\}|\} \end{aligned} \qquad (1)$$

$w(u,r)$ is the number of tag assignments, where user u assigned some tag t to place r through the action of checking-in or annotation. Hence, the weight assigned to a place simply corresponds to the frequency of the user reference to the place either by checking in or by leaving a tip.

We further normalise the weights so that the sum of the weights assigned to the places in the spatial profile is equal to 1. We use $\overline{P_R}$ to explicitly refer to the spatial profile where the sum of all weights is equal to 1, with $\overline{w}(u,r) = \frac{|\{t \in T : (u,t,r) \in Y\}|}{\sum_{i=1}^{n} \sum_{j=1}^{m} |\{t_i \in T : (u,t_i,r_j) \in Y\}|}$, where n and m are the total number of tags and resources, respectively. More simply, $\overline{w}(u,r) = \frac{N(u,r)}{N_T(u)}$, where $N(u,r)$ is the number of tags used by u for resource r, while $N_T(u)$ is the total number of tags used by u for all places.

Correspondingly, we define the tag-based profile of a user; $P_T(u)$ as follows.

Definition 2: **Semantic User Profile** A semantic user profile $P_T(u)$ of a user u is deduced from the set of tag assignments linked with u.

$$\begin{aligned} P_T(u) =& \{(t, w(u,t)) | (u,t,r) \in Y, \\ & w(u,t) = |\{r \in R : (u,t,r) \in Y\}|\} \end{aligned} \qquad (2)$$

$w(u,t)$ is the number of tag assignments where user u assigned tag t to some place through the action of checking-in or annotation.

$\overline{P_T}$ refers to the semantic profile where the sum of all weights is equal to 1, with $\overline{w}(u,t) = \frac{N(u,t)}{N_R(u)}$, where $N(u,t)$ is the number of resources annotated by u with t and $N_R(u)$ is the total number of resources annotated by u.

Furthermore, we define a spatio-semantic profile of a user $P_{RT}(u)$, that is a personalised association between user, place and tag.

Definition 3: **Spatio-Semantic User Profile** Let $\mathbb{F}_u = (T_u, R_u, I_u)$ of a given user $u \in U$ be the restriction of $\mathbb{F}$ to u, such that, T_u and R_u are finite sets of tags and places respectively, that are referenced from tag assignments performed by u, and I_u defines a relation between these sets: $I_u := \{(t,r) \in T_u \times R_u | (u,t,r) \in Y\}$.

A spatio-semantic user profile $P_{RT}(u)$ of a user u is deduced from the set of tag assignments made for place r by u.

$$\begin{aligned} P_{RT}(u) =& \{([r,t], w_u([r,t])) | (t,r) \in I_u, \\ & w_u([r,t]) = |\{t \in T_u : (t,r) \in I_u\}|\} \end{aligned} \qquad (3)$$

where $w([r,t])$ is how often user u assigned tag t to place r.

$\overline{P_{RT}}$ is the spatio-semantic profile where the sum of all weights is equal to 1, with $w_u([r,t]) = \frac{N(u,[r,t])}{N_{RT}(u)}$, where $N(u,[r,t])$ is the number of times u annotate r with t, and $N_{RT}(u)$ is the total number of tags assigned by u for r. (Note that tag assignment by users for a place comes from both the explicit action of annotation as well as implicit action of checking-in as represented in the geo-folksonomy model).

B. Basic Place and Tag Profiles

So far, the basic user profile provides only a limited view of the user association with places and concepts derived directly from captured data. Basic profiles reduce the dimensionality of the folksonomy space by considering only 2 dimensions at a time; user-place and user-tag, leading to a loss of correlation information between all three elements.

Users profiles can be extended to represent possible latent relationships in the data. Thus a user profile can be used to present places (respectively tags) similar to those in the basic profile, where similarity between places (respectively tags) is measured through the collective actions of other users of check-ins and annotations.

To compute tag-tag similarity, profiles for tags are first defined through the places they are used to annotate. Thus, a *place-based tag profile* ($P_R(t)$) of a tag t is a weighted list of places r that are annotated by t. That is, $w(r,t)$ is determined by the number of users' check-ins and tips that resulted in assigning t to r in the geo-folksonomy. Similarity between tags is defined as the cosine similarity between their place-based tag profiles as follows.

$$CSim(t_1,t_2) = \frac{|P_R(t_1) \cap P_R(t_2)|}{\sqrt{|P_R(t_1)|.|P_R(t_2)|}} \qquad (4)$$

On the other hand, similarity between places is defined by measuring the similarity of their tag-based and user-based profiles. Let $P_T(r)$ and $P_U(r)$ be the tag-based place profile and user-based place profile for place r (defined in a similar manner to user profiles above). Conceptually, a tag-based place profile is a description of the place by the tags assigned to it and a user-based place profile is an account of users' visits to the place.

Cosine similarity between tag-based place profiles ($CSim_{tag}(r_1,r_2)$) and between user-based place profiles ($CSim_{user}(r_1,r_2)$) define a tag-oriented ranking and user-oriented ranking, respectively. These similarity rankings can be aggregated using the so-called Borda method [28] to compute a generalised similarity score between two places.

$$PSim(r_1,r_2) = \gamma * CSim_{tag}(r_1,r_2) + (1-\gamma) * CSim_{user}(r_1,r_2) \qquad (5)$$

where $0 \leq \gamma \leq 1$ is a parameter that determines the balance of importance given to similarity scores from $P_T(r)$ and $P_U(r)$. Conceptually, similarity between two places is a function of the overlap between their tag assignments only (for $\gamma = 0$), a measure of their common visitors only (for $\gamma = 1$), or both (for γ between 0 and 1).

C. Enriched User Profiles

We extend the basic user profiles by the information extracted from the computation of tag and place similarity above. The enriched user profiles will therefore present a modified view of how users are associated with places that reflect collective user behaviour on the LBSN.

Definition 4: **Enriched Spatial User Profile** An enriched spatial user profile $\acute{P}_R(u)$ of a user u is an extension of the basic profile by places with the highest degree of similarity to places in $\overline{P_R(u)}$. Let R_u be the set of all places in $\overline{P_R(u)}$ and w_i is the weight associated with place i in the profile.

$$\acute{P}_R(u) = \{< r_i, w_i > | \\ w_i = \left\{ \begin{array}{ll} w_i & ,\text{if } r_i \in R_u \\ w_i * Max(PSim(r_i,r_j)) & ,\forall(r_i \in \{R-R_u\} \wedge r_j \in R_u) \end{array} \right\} \qquad (6)$$

We extend the profile by the 10 most similar places to every place in the user profile. The process of building the enriched spatial profile from place similarity with γ as an input is shown in Figure 2. The complexity of the enrichment algorithm is $O(N * M)$, where N is the number of users and M is the number of places in the user profile.

```
procedure SpatialEnrichment($P_R(u)$,$\gamma$)
    for all place $r_i$ in Spatial-Profile $P_R(u)$ do
        Compute $PSim(r_i, r \in P_R(u))$ from Eq. (5).
        Find top-10 similar places($(r_j, sim_j)$)
        for each $< r_j, sim_j >$ in top similar places do
            $w_j = w_i * sim_j$
            add $< r_j, w_j >$ to $P_R(u)$
        end for
    end for
    return $\acute{P}_R(u)$
end procedure
```

Fig. 2. Algorithm for building the enriched user profile.

Definition 5: **Enriched Semantic User Profile** An enriched semantic user profile $\acute{P}_T(u)$ of a user u is an extension of the basic profile by tags with the highest degree of similarity to tags in $\overline{P_T(u)}$. Let T_u be the set of all tags in $\overline{P_T(u)}$ and w_i is the weight associated with tag i in the profile.

$$\acute{P}_T(u) = \{< t_i, w_i > | \\ w_i = \left\{ \begin{array}{ll} w_i & ,if t_i \in T_u \\ w_i * Max(Sim(t_i,t_j)) & ,\forall(t_i \in \{T - T_u\} \wedge T_j \in T_u) \end{array} \right\} \qquad (7)$$

A similar algorithm to that of enriching place profiles is used for choosing the tags and weights.

Definition 6: **Enriched Spatio-Semantic User Profile**

An enriched spatio-semantic user profile $\acute{P}_{RT}(u)$ of a user u is an extension of the basic profile by tags and places with the highest degree of similarity to tags in $P_{RT}(u)$. Let T_u be the set of all tags in $\overline{P_T(u)}$, R_u be the set of all places in $\overline{P_R(u)}$ and w_{ij} is the weight associated with tag i and place j in the profile.

$$\acute{P}_{RT}(u) =< [r_i,t_j], w_u(r_i,t_j) > |w_u(r_i,t_j) = \\ \left\{ \begin{array}{ll} w_u(r_i,t_j) & ,\text{if } r_i \in R_u and t_j \in T_u \\ w_u(r_i,t_j)*Max(PSim(r_i,r_k)) & ,t_j \in P_T(r_k) \wedge r_k \in \{R - R_u\} \\ 0 & otherwise \end{array} \right\} \qquad (8)$$

The spatio-semantic profile is extended with the most similar places to the user profile and these are assigned a weight computed using the place similarity value for all tags in their place-tag profiles and 0 for tags that are not in their profile. Thus the user simply inherits relationships with all the tags and their associated weights from basic places that are deemed similar to those in his profile.

1) User Profile Example: Here an example is given of a sample user profile created from the dataset used in this work. 'user164' checked in 600 different venues, with associated 400 venue categories. Note that one venue can have more than one venue category. Figure 3 shows the top 20 tags in his semantic user profile. Figure 4 shows the filtered tags from his profile that represent human activity (approximately 5% of all tags), as derived by mapping to Wordnet *noun.act* category.

Figures 5 and 6 show the spatial profile and the enriched spatial profile for user 'user164', respectively. $\gamma = 0.5$ was used in the place similarity equation of the enriched profile.

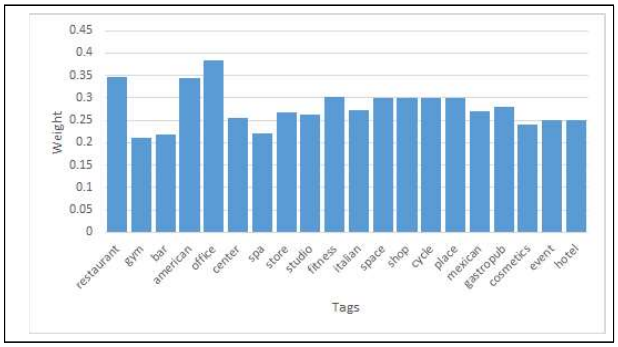


Fig. 3. Example Semantic user profile for user 'user164'.

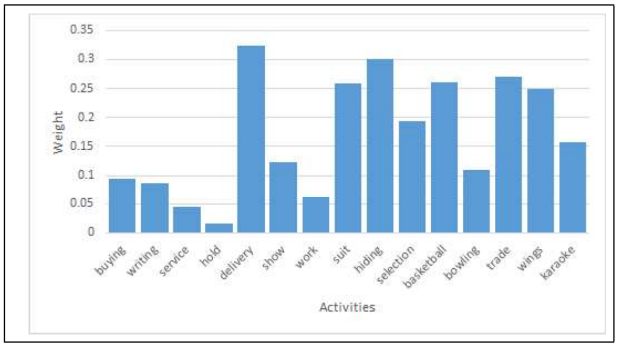


Fig. 4. Tags representing activities in the Semantic user profile of user 'user164'.

The size of the dots in the figures represents the weight (representing the degree of association) of the place in the profile. As shown in the figure, the level of association is more prominent for many more places in the enriched spatial profile.

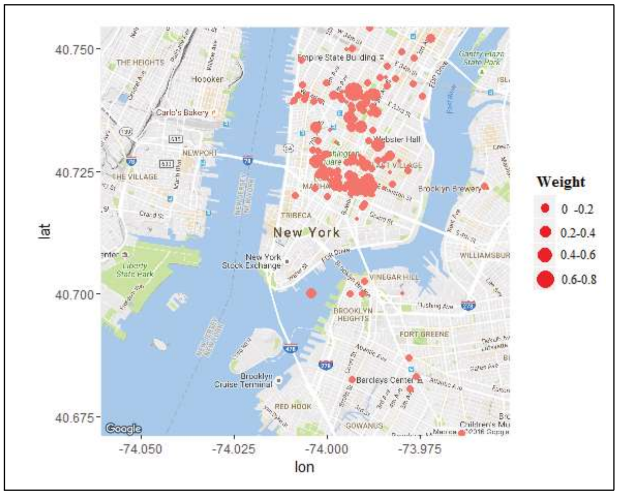


Fig. 5. Spatial user profile for user 'user164'.

V. Measuring User Similarity

Similarity between users can be measured on the basis of their spatial, semantic or spatio-semantic profiles. Spatial profiles gives a measure of user preferences in places. While the basic spatial profile will discover a map of common places that the users visited or annotated, the enriched spatial profile will produce an extended map of places that are likely to be of interest to both users. Similarity of spatial profiles can answer the question of which other users visiting habits to places are similar to mine?

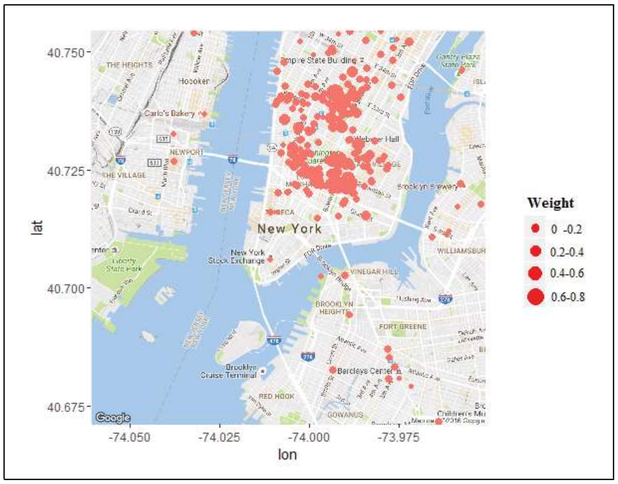


Fig. 6. Enriched spatial user profile for user 'user164' with $\gamma = 0.5$.

Semantic profiles, on the other hand, is a conceptual measure of user interests. Semantic filters on the types of concepts, e.g., themes of user activity or place type, can be applied to the folksonomy to give a more focussed view of user interests. Similarity of semantic profiles can answer questions such as, which other users share the same sort of activities as I do?

Spatio-semantic user similarity is a measure of personalised interests in places, as well as their associated concepts. It gives a holistic view of user preferences in place and will answer questions of which other users are interested in this (specific) place and share the same experiences or interests in this place.

Cosine similarity between any of the above types of profiles can be used to compute the similarity between users. The application of this process can be constrained by region of interest, by considering only users who have a high degree of similarity between their basic spatial profiles.

Figure 7 shows a bar chart of similarity values between 'user164' and other users, using their basic spatial and enriched spatial profiles. The figure demonstrates the impact of enrichment on user similarity, where this user appears to become more similar to other users in his profile (e.g., with 'user134'), given an extended view of their interests in places and their associated concepts.

VI. Experiments and Results

A. Datasets

Data about venues, tips and users who left the tips can be collected directly from Foursquare. However, users' check-in data are normally private. Many Foursquare users tend to push their check-in activity through Twitter; thus allowing another means of tracing the check-in information.

Approximately (10 months) of check-in data in New York city were collected from Foursquare between April 2012 and February 2013. This data consists of 227,428 anonymized user check-ins, with venue ids, venue category, longitude and

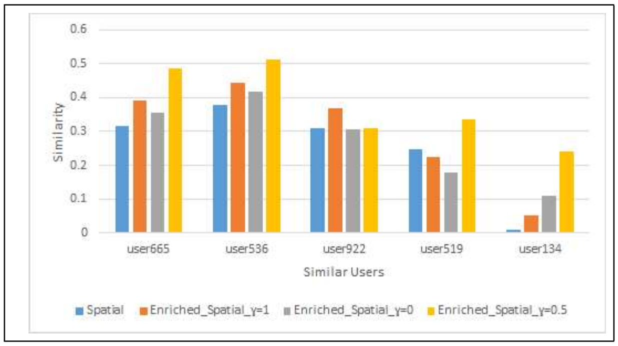


Fig. 7. Similarity between 'user164' and other users using their spatial and enriched spatial profiles.

latitude of venues and time stamps of check-ins. The data was then used to recursively extract venue-related tips (tip id, text and time stamp), and subsequently all venues for users related to the tips collected. 604,924 tips were collected for 167,786 users in 36,940 venues. Time stamps of the tip data range from January 2009 to June 2015. Figure 8 shows the number of places versus the number of users in the collected dataset. As the figure shows, about 94% of the users visited less than 10 places and about 3% of users visited 11 to 20 places and the remaining 3% visited 21 to 400 places.

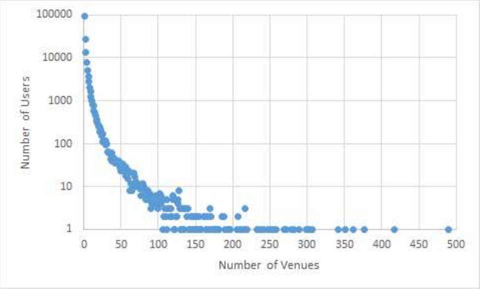


Fig. 8. Number of users versus the number of venues visited in the dataset.

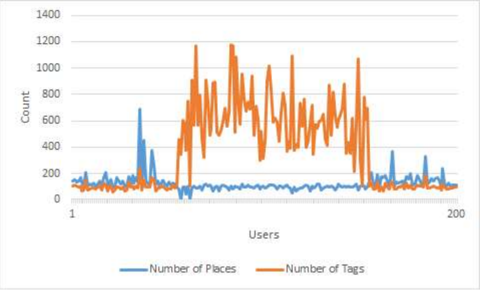


Fig. 9. Number of distinct places and tags for each user in the dataset.

Experiments were carried out using a sample of 400 users (200 users with a high frequency of check-ins and co-location rate and 200 users with a low frequency usage). Evaluation results for the profiles are presented for the high frequency users, but later a comparison between the two data sets is also given. Tables I, II shows summary statistics of the used sample datasets for both groups of users.

TABLE I. High Frequency Users' Dataset

Number of Venues	10,988
Total number of Checkins	50,584
Total Number of Tips	10,469
Total Number of Disitnct Tags	13,396
Number of users	200
Total Number categories	495
Total Number of Relationships	165,453

TABLE II. Low Frequency Users' Dataset

Number of Distinct Venues	4,411
Total number of Checkins	4,212
Total Number of Tips	2,900
Total Number of Tags	5,949
Number of users	200
Total Number categories	374
Total Number of Relationships	57,786

B. Evaluation of User Profiles

The evaluation experiment aims to measure the impact of using the full range of content captured on LBSN when building user profiles in comparison to using only partial views based on the check-in information. The experiment takes the form of place (and tag) top-N recommendation problem using the different constructed user profiles and seeks to establish how well the profiles reflect the user spatial and semantic characteristics when using the LBSN. The algorithm used for computing the top-N recommendations using spatial profiles is shown in Figure 10.

We use recall@N, precision@N and F1@N as our success measures, where N is a predefined number of places (or tags) to be recommended. Recall measures the ratio of correct recommendations to the number of true places (or tags) of a test check-in or tip record, whereas precision measures the ratio of correct to false recommendations made. Recall and precision are given by the following equations.

$$\begin{aligned} recall &= \frac{TP}{TP+FN} \\ precision &= \frac{TP}{TP+FP} \end{aligned}$$

True positives (TP) is the number of correct place (or tags) recommended; false positives (FP) is the number of wrong recommendations and false negatives (FN) is the number of true place (or tags) which were not recommended. F1 is a

TABLE III. Descriptive statistics of Check-ins for User Categories

Check-ins	**Low Frequency Users**	**High Frequency Users**
Mean	26.685	123.455
Median	28	105.5
Mode	29	104
Standard Deviation	6.221682	65.91199
Sample Variance	38.70932	4344.39
Range	29	648
Minimum	9	42
Maximum	38	690
User Count	200	200

```
procedure SPATIO-SEMANTIC TOP-K RECOMMENDER($\gamma$,TopK)
    for each $u_i$ do
        SpatialEnrichment($P_R(u_i), \gamma$)
    end for
    for all $u_i, u_j$ do
        Fetch profiles $P_R(u_i), P_R(u_j)$
        Compute CSim($u_i, u_j$) .
    end for
    for each $u_i$ do
        Fetch most similar user $u_j$
        Sort $< r_j, w_j >$ of $P_R(u_j)$
        Recommend TopK $r_j$ that are not in $P_R(u_i)$
    end for
    return TopK $< r_j, w_j >$
end procedure
```

Fig. 10. Spatio-semantic Top-K recommendation algorithm.

combined measure of recall and precision and is given by

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

The values of TP, FP and FN are determined by randomly splitting the users into two sets; the training set and the testing set. Multi-fold cross-validation was used to ensure a fair partitioning between test data and training data. Data were split 90% for training and 10% for testing, and the process was repeated 5 times to create 5 folds and the mean performance was reported.

1) Evaluation of Spatial Profiles: Results for the enriched user profiles using the proposed top-N recommendation method are presented. Different versions of the enriched spatial profiles, using different place similarity measures were created, a) using $\gamma = 0$ (to represent enrichment with place-tag similarity only), b) using $\gamma = 1$, (to represent enrichment with place-user similarity only), and c) using $\gamma = 0.5$ for an aggregated view of both effects. Hence, result sets are shown for the following user profiles. 1. Enriched-(Spatial + Tag) 2. Enriched-(Spatial+ User) 3. Enriched-(Spatial + All).

We compare the results of the top-N recommendation using the three different profiles with traditional Item-based Collaborative Filtering (IBCF) [29] and User-based collaborative Filtering (UCBF) [30] approaches, applied against the basic spatial user profile for recommending top-1, 2, 3, 4, 5, 10, 20, 30, 40, 50. Figures 11, 12 and 13 show the precision, recall and F1-measure for all approaches. As is shown in the figures, enriched user profiles demonstrate significantly better performance in comparison to the traditional approaches. In particular, the F1 measure for the combined profile (Spatial + All) outperforms the UBCF approach by 10% on average and the IBCF approach by 12% on average.

2) Evaluation of Semantic profiles: A similar experiment was carried out to evaluate the semantic user profiles. Again, the results were compared to the UBCF and IBCF approaches. Figures 14, 15 and 16 show the results of the top-10, 20, 30, 40, and 50 tag recommendations using the different methods. As shown in Figure 14, the enriched semantic profile demonstrates significant improvements with respect to both the UBCF and IBCF approaches. Results demonstrates the quality of the enriched semantic user profiles, and thus confirm their utility for more accurate representations of user profiles.

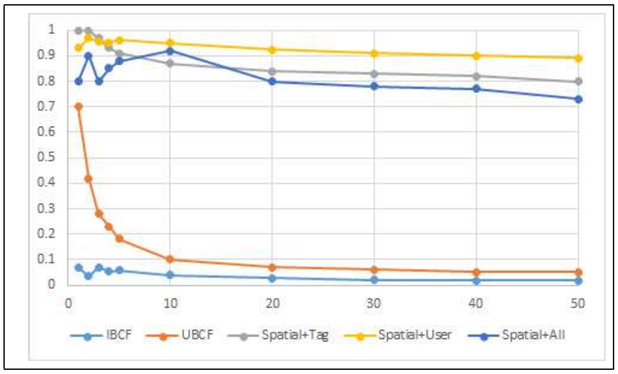


Fig. 11. Precision values for the top-N place recommendations.

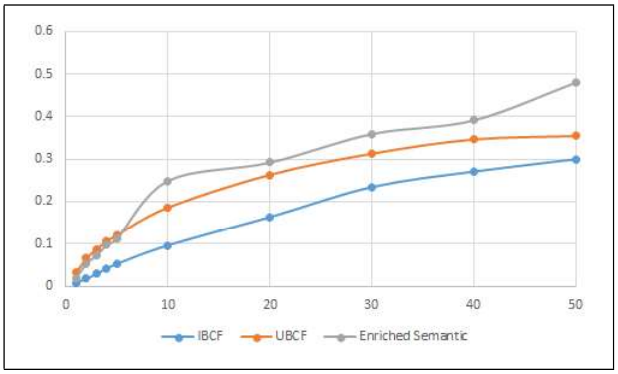


Fig. 12. Recall values for the top-N place recommendations

C. Evaluation of User Similarity Approaches

1) Methodology: The measure of user similarity is evaluated as an Information Retrieval problem where we search for the most similar user to a particular user in question. Place categories are used as a basis of ground truth comparison and evaluation.

Table IV shows an example; where distinct categories for the top-10 most visited places are shown for two sample users (with similarity value of 0.65). Foursquare attaches more than

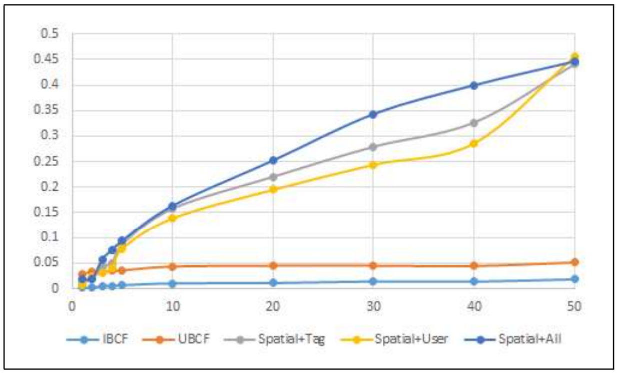


Fig. 13. F1 measure values for the top-N place recommendations

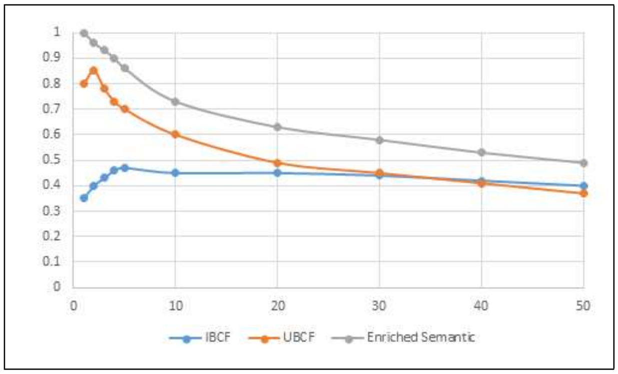


Fig. 14. Precision values for top-N tag recommendations.

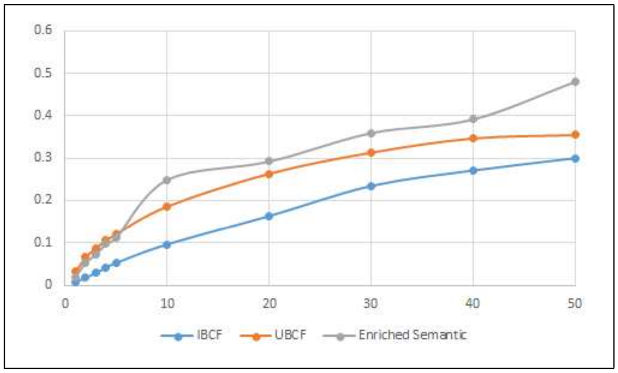


Fig. 15. Recall values for top-N tag recommendations.

one category to a place, and thus, there may be more than 10 categories for the top-10 places . The highlighted cells show the common categories between the two users.

Precision, recall and F-Measure are used as evaluation metrics in the same way they are used in the IR literature. These are defined below.

$$Precision = \frac{|f(u) \bigcap f(u_{sim})|}{f(u_{sim})} \tag{9}$$

$$Recall = \frac{|f(u) \bigcap f(u_{sim})|}{f(u)} \tag{10}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{11}$$

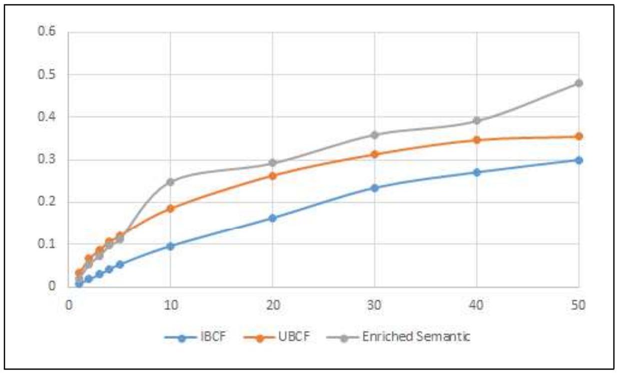
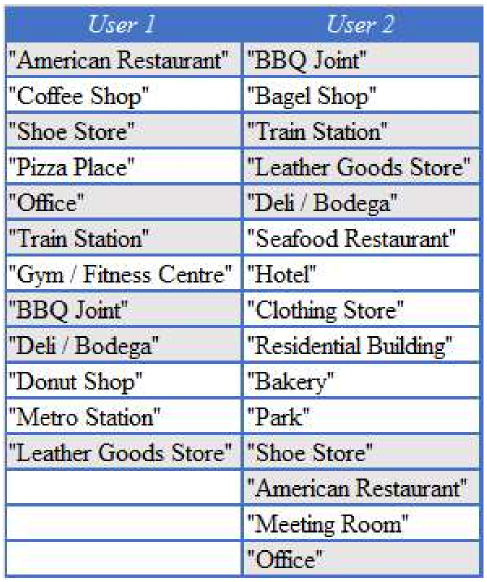


Fig. 16. F1 measure values for top-N tag recommendations.

TABLE IV. Distinct categories for top-10 most visited place for two users with similarity value of 0.65



User 1	User 2
"American Restaurant"	"BBQ Joint"
"Coffee Shop"	"Bagel Shop"
"Shoe Store"	"Train Station"
"Pizza Place"	"Leather Goods Store"
"Office"	"Deli / Bodega"
"Train Station"	"Seafood Restaurant"
"Gym / Fitness Centre"	"Hotel"
"BBQ Joint"	"Clothing Store"
"Deli / Bodega"	"Residential Building"
"Donut Shop"	"Bakery"
"Metro Station"	"Park"
"Leather Goods Store"	"Shoe Store"
	"American Restaurant"
	"Meeting Room"
	"Office"

where f(u) is the set of the distinct categories of the top-k places of user u, u_{sim} is the most similar user, and f(u_{sim}) is the set of distinct categories of the top-k places of the most similar user u_{sim}. Hence, precision represents the ratio of common categories between the two users in reference to those of the first user, while recall presents the same ration with respect to the second user. The F1 measure is the harmonic mean of precision and recall.

2) *Similarity of Spatial Profiles:* The evaluation experiment aims to measure the impact of using the full range of content captured on LBSN when building user profiles in comparison to using only partial views based on check-in information. We calculate the user similarity between the following user profiles:

1) Spatial User profile; ($user\_sim$)
2) Enriched Spatial with $CSim_{tag}$, $\gamma = 1$; ($user\_sim_{tag}$)
3) Enriched Spatial with $CSim_{user}$, $\gamma = 0$; ($user\_sim_{user}$)
4) Enriched Spatial with combined similarity, $\gamma = 0.5$); ($user\_sim_{combined}$)

Table V is a compilation of the precision, recall and F1-measure values for the various user similarities. For each profile, we fetch the frequent top-5, 10, 20, 30, 40, 50 venues and then evaluate their categories using the equations.

As can be shown in the table, similarity computation with the enriched spatial profiles produce a higher degree of precision, recall and F-measures in general, whilst the best results are for the enriched profiles with the combined place similarity. Results indicate that location tracks may not be the best basis for finding similar users and that a combined treatment of both the spatial and semantic dimensions can produce more accurate views of user profiles.

To further clarify the improvements in the evaluation metrics figures 17, 18 and 19 present the improvement in percentage of the metrics of the enriched profile over the basic spatial profile.

TABLE V. User Similarity Evaluation: Precision, Recall, F1-measure.

Precision

Top-K Places	$user\_sim$	$user\_sim_{tag}$	$user\_sim_{user}$	$user\_sim_{combined}$
Top-5	0.29016885	0.2836818	**0.2936379**	0.27810863
Top-10	0.32131577	0.28528178	0.28525722	**0.35280818**
Top-20	0.3590904	0.35163218	0.35682544	**0.3996159**
Top-30	0.38940138	0.39706513	0.40721306	**0.42995644**
Top-40	0.41870615	0.43158174	0.45326504	**0.4587258**
Top-50	0.43747735	0.48375404	0.49606603	0.46999252

Recall

Top-K Places	$user\_sim$	$user\_sim_{tag}$	$user\_sim_{user}$	$user\_sim_{combined}$
Top-5	0.2910496	0.26843706	0.28794414	**0.2881556**
Top-10	0.31549093	0.28207284	0.28236988	**0.36440614**
Top-20	0.35180694	0.16477493	0.35360697	**0.4058911**
Top-30	0.37913677	0.38837454	0.4045744	**0.44445464**
Top-40	0.39773872	0.41915244	0.44400847	**0.4669912**
Top-50	0.40748206	0.45438704	**0.48419788**	0.47782102

F1-measure

Top-K Places	$user\_sim$	$user\_sim_{tag}$	$user\_sim_{user}$	$user\_sim_{combined}$
Top-5	0.29060855	0.27584896	**0.29076314**	0.28304298
Top-10	0.3183767	0.28366823	0.28380620	**0.35851338**
Top-20	0.35541135	0.35070109	0.35520891	**0.40272906**
Top-30	0.38420052	0.39267175	0.40588944	**0.43708534**
Top-40	0.40795319	0.42527629	0.44858900	**0.46282160**
Top-50	0.42194730	0.46861089	0.49006011	**0.4738744**

A positive value means an improvement in performance. As can be observed, the average gain in precision and recall is best demonstrated in the case of enriched profile with the combined spatial and semantic measures.

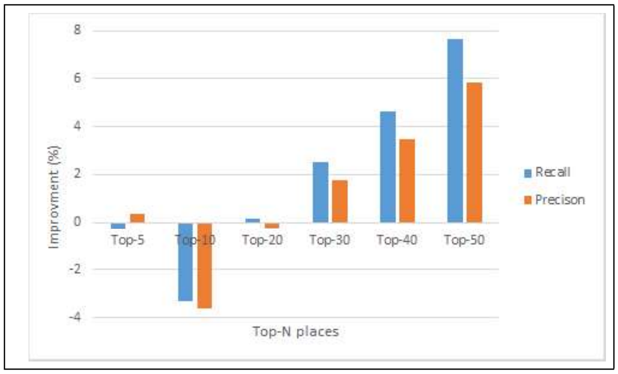


Fig. 17. $user\_sim_{user}$ versus user_sim

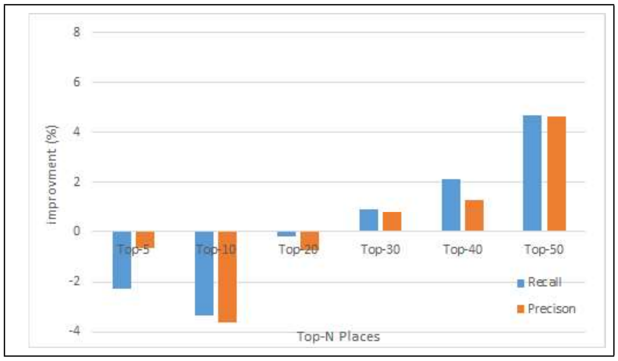


Fig. 18. $user\_sim_{tag}$ versus user_sim

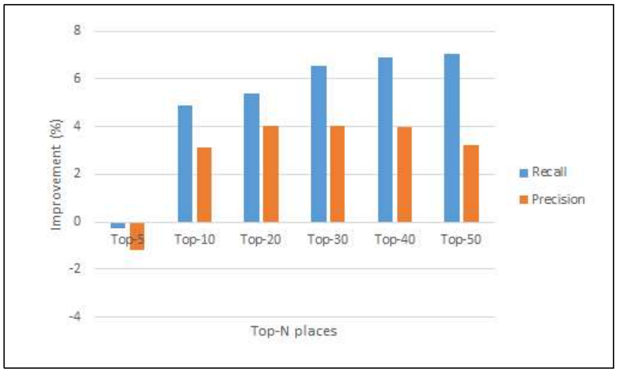


Fig. 19. $user\_sim_{combined}$ versus user_sim

To evaluate the overall performance of similarity methods, *the Mean Average Precision* (MAP) measure is employed. MAP is a commonly used summary measure of a ranked retrieval run. In our experiment, it stands for the mean of the precision score after each relevant user is retrieved for different top-N values, as in Equation (12).

$$MAP = \frac{\sum_1^N p@n}{N} \tag{12}$$

Figure 20 shows a comparative study of MAP between the different user similarities from different profiles baselines, and confirms the improved results for the enriched combined user similarity.

3) Similarity of Semantic Profiles: A similar experiment to the above is carried out for evaluating both the basic and enriched semantic profiles. Table VI shows the precision, recall and F1 measure values. As can be seen in the table, the enriched semantic similarity method performed better than the basic one. A compilation of an overall picture of the spatial

TABLE VI. Semantic Similarity Evaluation

	Precision		**Recall**		**F1 measure**	
	Semantic	*Enriched Semantic*	*Semantic*	*Enriched Semantic*	*Semantic*	*Enriched Semantic*
Top-5	0.304228	0.284233	0.302118	0.275948	**0.296848**	0.277169
Top-10	0.337925	0.283362	0.326494	0.279485	**0.319749**	0.279089
Top-20	0.389438	0.365204	0.344888	0.360197	0.348238	**0.35966**
Top-30	0.423739	0.410384	0.351962	0.405662	0.361898	**0.402931**
Top-40	0.456672	0.452301	0.369331	0.443004	0.382173	**0.441523**
Top-50	0.468839	0.49545	0.372371	0.474493	0.386104	**0.475632**

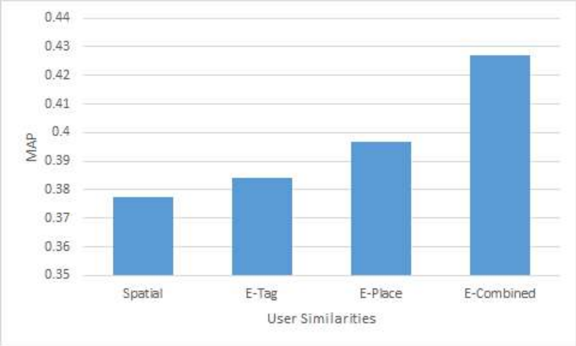


Fig. 20. Mean Average Precision (MAP) values for the different user similarities

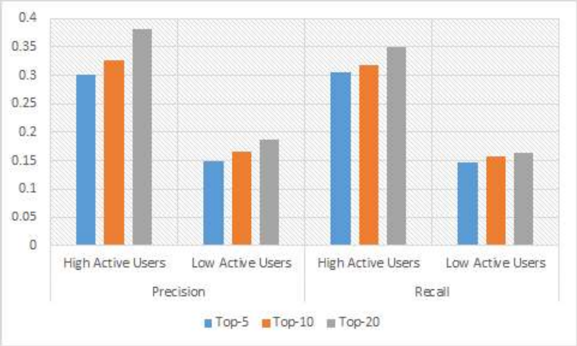


Fig. 22. Activity effect on combined user similarity

similarity against the semantic similarity methods is shown in Figure 21.

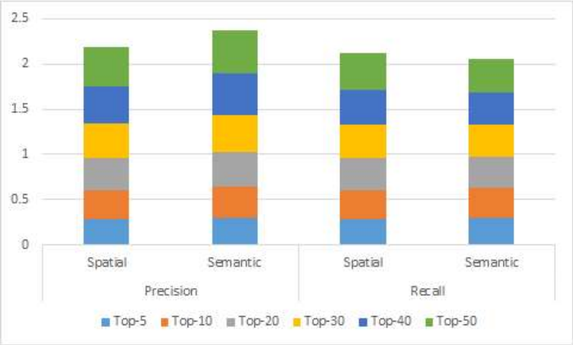


Fig. 21. Precision and Recall values for Spatial Versus Semantic User Similarity Evaluation

4) Influence of User Activity: Here a comparison is made between the similarity methods, given different levels of user activity on the LBSN. 200 users were chosen with high frequency of usage of the network and 200 others with a much reduced frequency as described in Table III.

Figure 22 summarises the results of the evaluation using the similarity method with $\gamma = 0.5$. As can be expected, the figure demonstrates how both the precision and recall values are higher in the case of frequent user.

VII. Conclusions

This paper considers the problem of user profiling on location-based social networks. Both the spatial (where) and the semantic (what) dimensions of user and place data are used to construct different views of a user's profile. A place is considered to be associated with a set of tags or labels that describe its associated place types, as well as summarise the users' annotations in the place. A folksonomy data model and analysis methods are used to represent and manipulate the data to construct user profiles and place profiles. It is shown how user profiles can be extended from a basic model that describes user's direct links with a place, to enriched profiles describing richer views of place data on the social network. The model is flexible and can be adjusted to focus on the spatial and semantic dimensions separately or in combination. Results demonstrate that the proposed methods produce user profiles that are more representative of user's spatial and semantic preferences. The framework is used a a basis for computing different methods of similarity between users. Experimental results were carried out on a representative set of users of a LBSN and demonstrate the efficacy of basing the similarity on profiles that combine both the semantic and spatial information in the data. To our knowledge, no other works have proposed similar treatments of the problem before. Future work will consider the temporal dimension of the data, which adds another layer of complexity as well as explore further inference of useful semantics from the data, e.g., representation of activities or experiences carried out in geographic places.

VIII. Acknowledgment

This work was performed using the computing facilities of the Advanced Research Computing @ Cardiff (ARCCA) Division, Cardiff University.

References

[1] S. Mohamed and A. Abdelmoty, "Uncovering user profiles in location-based social networks," in *GEOProcessing 2016: The Eighth International Conference on Advanced Geographic Information Systems, Applications, and Services*, 2016, pp. 14–21.

[2] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma, "Mining user similarity based on location history," in *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. ACM, 2008, pp. 34–42.

[3] H. Gao, J. Tang, and H. Liu, "Exploring social-historical ties on location-based social networks." in *ICWSM*, 2012, pp. 114–121.

[4] X. Cao, G. Cong, and C. S. Jensen, "Mining significant semantic locations from gps data," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 1009–1020, 2010.

[5] M. Ye, X. Liu, and W.-C. Lee, "Exploring social influence for recommendation-a probabilistic generative model approach," *arXiv preprint arXiv:1109.0758*, 2011.

[6] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel, "Recommendations in location-based social networks: a survey," *GeoInformatica*, vol. 19, no. 3, pp. 525–565, 2015.

[7] A. Noulas and C. Mascolo, "Exploiting foursquare and cellular data to infer user activity in urban environments," in *14th International Conference on Mobile Data Management*, vol. 1. IEEE, 2013, pp. 167–176.

[8] J.-D. Zhang and C.-Y. Chow, "igslr: personalized geo-social location recommendation: a kernel density estimation approach," in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2013, pp. 334–343.

[9] H. Gao, J. Tang, and H. Liu, "gscorr: modeling geo-social correlations for new check-ins on location-based social networks," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 1582–1586.

[10] B. Liu, H. Xiong, S. Papadimitriou, Y. Fu, and Z. Yao, "A general geographical probabilistic factor model for point of interest recommendation," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 27, no. 5, pp. 1167–1179, 2015.

[11] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1082–1090.

[12] D. Zhou, B. Wang, S. M. Rahimi, and X. Wang, "A study of recommending locations on location-based social network by collaborative filtering," in *Advances in Artificial Intelligence*. Springer, 2012, pp. 255–266.

[13] H. Wang, M. Terrovitis, and N. Mamoulis, "Location recommendation in location-based social networks using user check-in data," in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2013, pp. 374–383.

[14] J. J.-C. Ying, E. H.-C. Lu, W.-N. Kuo, and V. S. Tseng, "Urban point-of-interest recommendation by mining user check-in behaviors," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. ACM, 2012, pp. 63–70.

[15] B. Liu, Y. Fu, Z. Yao, and H. Xiong, "Learning geographical preferences for point-of-interest recommendation," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1043–1051.

[16] G. McKenzie, B. Adams, and K. Janowicz, "A thematic approach to user similarity built on geosocial check-ins," in *Geographic Information Science at the Heart of Europe*. Springer, 2013, pp. 39–53.

[17] B. Hu and M. Ester, "Spatial topic modeling in online social media for location recommendation," in *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 2013, pp. 25–32.

[18] J.-D. Zhang and C.-Y. Chow, "Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 443–452.

[19] D. Yang, D. Zhang, Z. Yu, and Z. Wang, "A sentiment-enhanced personalized location recommendation system," in *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, 2013, pp. 119–128.

[20] M.-J. Lee and C.-W. Chung, "A user similarity calculation based on the location for social network services," in *Database Systems for Advanced Applications*. Springer, 2011, pp. 38–52.

[21] G. Ference, M. Ye, and W.-C. Lee, "Location recommendation for out-of-town users in location-based social networks," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 2013, pp. 721–726.

[22] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee, "Exploiting geographical influence for collaborative point-of-interest recommendation," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 325–334.

[23] J. J.-C. Ying, W.-N. Kuo, V. S. Tseng, and E. H.-C. Lu, "Mining user check-in behavior with a random walk for urban point-of-interest recommendations," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 40, 2014.

[24] C. Cheng, H. Yang, I. King, and M. R. Lyu, "Fused matrix factorization with geographical and social influence in location-based social networks." in *Aaai*, vol. 12, 2012, pp. 17–23.

[25] Y.-L. Zhao, L. Nie, X. Wang, and T.-S. Chua, "Personalized recommendations of locally interesting venues to tourists via cross-region community matching," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 50, 2014.

[26] F. Abel, "Contextualization, User Modeling and Personalization in the Social Web," PhD Thesis, Gottfried Wilhelm Leibniz University Hannover, April 2011.

[27] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "Information retrieval in folksonomies: Search and ranking," in *Semantic web: research and applications, proceedings*. Springer, 2006, pp. 411–426.

[28] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 613–622.

[29] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 285–295.

[30] M. J. Pazzani, "A framework for collaborative, content-based and demographic filtering," *Artificial Intelligence Review*, vol. 13, no. 5-6, pp. 393–408, 1999.