Effective Community Search Over Location-Based Social Networks: Conceptual Framework with Preliminary Result

Ismail Alaqta[1,2(✉)], Junho Wang[1], and Mohammad Awrangjeb[1]

[1] School of Information and Communication Technology, Griffith University, Brisbane, Australia
ismail.alaqta@griffithuni.edu.au,
{j.wang,m.awrangjeb}@griffith.edu.au
[2] Department of Computer Science, Jazan University, Gizan, Saudi Arabia
ialaqta@jazanu.edu.sa

Abstract. Over the past decade, the volume of data has grown exponentially due to global internet service propagation. The number of individuals using the internet has expanded, especially with the use of social networks. Utilising GPS-enabled mobile devices, social networks have been labelled Location-based Social Networks (LBSN). This service enables users to share their current spatial information by 'checking-in' with their friends at different locations. This article proposes a conceptual framework to enhance the effectiveness of community search over LBSN. As users are more likely to look for people whom they share similar personalities and interests, these keywords plus the spatial information could help a lot in finding the most appropriate query-based social community. As a result, this paper aims to contribute to the existing body of knowledge as well as the industry in the field of community search (CS). In particular, this work is focusing on CS in the environment of LBSN to benefit from factors of spatial, keywords and time in order to enhance community search models by these factors. Therefore, in this study, we focus on the current state-of-the art of CS and the limitations of integrated models. The preliminary results confirm that user's checkins can present an alternative approach to produce and update the users' interests with which we use to boast effectiveness of attributed community search along with spatial information.

Keywords: Community search · User interests · Spatial graph

1 Introduction

Over the past decade, the volume of data has grown exponentially due to global internet service propagation. According to the latest report issued by the UN's international telecommunications union (ITU), the number of individuals using

L. Chang et al. (Eds.): ADC 2019, LNCS 11393, pp. 119–131, 2019.
https://doi.org/10.1007/978-3-030-12079-5_9

the internet exceeded 3.5 billion by 2017[1]. Social network applications, such as Facebook[2], Twitter[3] and Foursquare[4] are the most common internet applications. These applications have consequently attracted millions of users. For example, over 1.75 billion of Facebook users are active monthly[5]. Because they use GPS-enabled mobile devices, social networks have been called Location-based Social Networks (LBSNs). This service enables users to share their current spatial information by 'checking-in' with their friends at different locations. Foursquare, on which more than 30 million users are accommodated, receives millions of check-ins daily[6]. Other traditional social networks such as Facebook and Twitter[7] also provide users with the facility of check-ins, which can be utilised for many business purposes. In most cases, a check-in generates a triplet $\langle u, l, t \rangle$ indicating that user u checked-in at location l associated with spatial information $\langle x, y \rangle$ at a specific *time* t, which also shows that the user is temporally online. Consequently, this leads both industry and academia to consider the time dimension. People on social networks communicate with each other and this interaction is recorded with time. For example, consider social network users on the Gold Coast who are interested in a coffee shop at which their friends have already checked-in. This group of people have planned to meet up at a certain place and time. The coffee shop (e.g. Merlo)can also utilise its customers' profiles on Facebook to provide location-specific advertisements to potential customers, who might also be interested in other items offered by the coffee shop. However, this increases the complexity of the social network. Moreover, due to the vast development of online social networks, people can create and update their profiles. A huge amount of textual information is associated with users because they can express themselves easily through blogging. If a Flickr user utilises many keywords related to travelling (e.g. posts many photos about trips with keywords), these keywords help interested users to find people with similar interests. Basically, users are more likely to search for people with whom they share similar personalities and interests or those who share similar work and research areas. Users are progressively geo-coded and geo-positioned on social networks and there is increased availability of textual descriptions regarding interests, such as tourist attractions and cafes.

This research contributes to the existing body of knowledge as well as the industry in the field of community search. This work focuses on the social community in the environment of LBSN. Due to the variant data type of LBSN, the significance of this research can be classified into three dimensions: social relationships, attributes, and spatio-temporal. In terms of social matter, friend recommendations, in which the system searches for similar users to recommend

[1] http://www.itu.int/en/ITU-D/statistics.
[2] http://www.facebook.com.
[3] http://twitter.com.
[4] https://foursquare.com.
[5] http://www.statisticbrain.com/facebook-statistics.
[6] Foursquare statistics. https://foursquare.com/about/.
[7] www.twitter.com.

them to each other, is one of the most important outputs of community search. Moreover, as the users of LBSN can have keywords or tags to describe themselves or their businesses, a self-drive tour of a set of POIs or a minimum group of people, who share similar interests, could be achieved using an attributed community.

To model and search complex social graphs meaningfully, the simple graph model is often not adequate to capture many real-world social network datasets. As previously noticed from the examples, for most social networks, information is not only available about social connections but also about user demographics, preferences, actions performed, and so on. Combining both the explicit spatial association of a place and the implicit semantics of interaction with s place provides a unique opportunity for in-depth understanding of both places and users. Hence, in this research we investigate the possibilities of *spatio-attributed community search* to enrich the simple graph model.

2 Related Work

Community search is a community retrieval approach that aims to find a densely populated query-based on-line connected community (Fang et al. 2017; Li et al. 2015). For example, *k-core* (Seidman 1983) was utilised in (Li et al. 2015; Sozio and Gionis 2010). (Sozio and Gionis 2010) designed the first algorithm Global to retrieve the connected *k-core* that includes the vertex q. In detail, the problem was formulated as Q, a set of query nodes or seeds against a graph $G = (V, E)$ to retrieve a connected subgraph including Q. Thus, the authors suggested a function called the 'goodness function' f to measure the goodness of the subgraph. Moreover, this work (Sozio and Gionis 2010) considered subgraph density by using two other functions: the average and minimum degree of the subgraph nodes f_a and f_m, respectively.

2.1 Attributed Community Search

An attributed community is represented by vertices associated with text or keyword-named attributes. These attributes can effectively provide more features such as ease of interpretation and personalization (Fang et al. 2017). Recently, (Shang et al. 2017) proposed an attributed community search method, which was enhanced by (Huang et al. 2014), with a refining technique. The main idea was to reconstruct the graph based on topology-based and attribute-based similarities. The new reconstructed graph was called the TA-graph. Based on the TA-graph structure, an index named AttrTCP-index based on TCP-index (Huang et al. 2014) was created. Thus, queries that are on the new index AttrTCP-index return to communities that satisfy the queries. Moreover, (Fang et al. 2017) investigated the attributed community search by combining a cohesive structure and keyword. The data model in this study was similar to the previous one (Shang et al. 2017), specifically in keywords for which each vertex v is associated with a set of keywords. However, this work utilised the $k - core$ technique

(Seidman 1983) and the decomposition algorithm proposed in (Batagelj 2003) to find a cohesive structure called a connected $k-core$ denoted by $\widehat{k-core}$. More significantly, the study designed an index called the Core Label tree (CL-tree), which puts the $\widehat{k-core}$ and keywords in a tree structure. Based on the $k-core$ definition, the authors identify the research problem as given $G = (V, E)$, a positive integer k, a vertex $q \in V$ and a set of keywords $S \subseteq W$. In community search, index construction plays a key role due to the effective and efficient impact on results. Since cores can be nested (Batagelj 2003), the CL-tree index (Fang et al. 2017) was constructed. Obviously, a $\widehat{k-core}$ must contain $\widehat{(k+1)-core}$. Thus, a tree structure is the most suitable data structure for such $k-cores$.

2.2 Spatial Community Search

Spatial graphs are on-line social networks on which users can share their location information, e.g. their position during check-ins. Spatial community search can perform community retrieval techniques, e.g. *k-core or k-truss* on a spatial social network. For example, given a Geo-Social Graph G, and a query vertex q, the task of spatial community search is to find a subgraph of G. This subsection reviews the most considerable works in terms of a spatio-social community search, as previously reviewed works assume non-spatial graphs (Cui et al. 2013; 2014; Huang et al. 2014; Li et al. 2015; Sozio and Gionis 2010). It can be said that a recent work named *spatial-aware community* (SAC) (Fang et al. 2016) has adopted the concept of minimum degree, which basically depends on the $k-core$ technique. SAC is a subgraph denoted by $H = (V_H, E_H)$, which needs to satisfy the following:

- Connectivity, $G_q \in G$ is connected and q exists.
- Structure cohesiveness $\Rightarrow$ all vertices are intensively linked in H.
- Spatial cohesiveness $\Rightarrow$ all vertices are almost at the same spatial location.

Compared to traditional CS works, condition three is intuitively what distinguishes SAC. So, spatial cohesiveness in SAC is defined to achieve a minimum covering circle (MCC) with the smallest radius. The formal definition is that given a set of vertices **S**, the MCC of **S** is the spatial circle that contains all vertices in S with the smallest radius. SAC follows the two-step framework: (1) find a community S of vertices, based on some CS algorithm (Sozio and Gionis 2010); and (2) find a subset of S that satisfies both structure and spatial cohesiveness.

All reviewed methods consider social constraints. Some reinforce social queries with extra constraints, e.g. keywords, location, and time. However, there is a lack of integration of all constraints into one CS framework. Therefore, this article proposes a conceptual framework to enhance the effectiveness of community search models over LBSN. The enhancement has been enforced by integrating compatibly text-mining techniques with a community search model as demonstrated in Fig. 1.

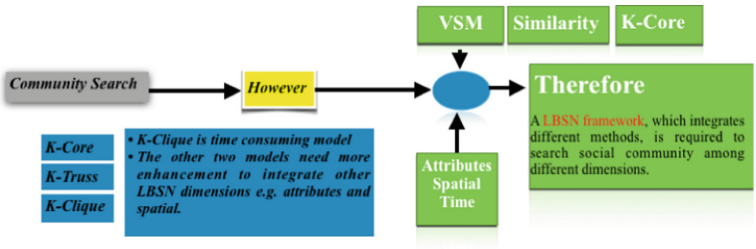

Fig. 1. Research gap

3 Methodology

This research has developed a hybrid approach, which aims to search for desired query-based social communities over LBSN. Our hybrid approach considers three different dimensions, *including keywords, location, and time*, which significantly enhance the effectiveness of social community search outputs. Therefore, the approach has combined different methods in which each method is required to achieve its research objectives under one framework.

3.1 Problem Formulation

In this section, we provide definitions that will be used throughout the paper. Also, this section provides the problem statement followed by an example to elaborate our research problem.

Data Model: We consider the location-based social network $G = (V, E, X)$ as an attributed graph, where V is a set of all users. Each edge $e\,(u, v) \in E$ indicates that a friendship exists between two users. X denotes a matrix $[X]\,n \times l$ where l is the number of all possible distinct keywords W, which are associated with places P that have been visited by users in form of 4-tuple check-in point CK. So, $CK = \{\langle u_i, p_k, t, W_{p_k}\rangle | u_i \in V, p_k \in P\}$ where p_k is identified by a unique GPS coordinate and t is a time-stamp when a user u_i checked-in p_k. For example, in Fig. 2 there are nine users, i.e. $u_1, ..., u_9$. Some conform with the conditions of inducing dense subgraphs. For instance, $\langle u_1, u_2, u_3, u_4\rangle, \langle u_7, u_8, u_9\rangle$ are two subsets, which form socially dense subgraphs. Moreover, our example shows that users could visit places either as a group or individually, e.g. $\langle u_1, u_2, u_3, u_4\rangle$ checked-in at the same time t_1 and the same place as well, which results in keeping a dense spatio-temporal relationship. Later, we will learn how to define our query model to retrieve communities. Based on our data model, we give the following definitions followed by the query model.

Query Model: The main goal of our framework is to search the community of a location-based social graph. Our query model is maintained by several constraints that need to be satisfied to return an Attri-Spatial Social Community.

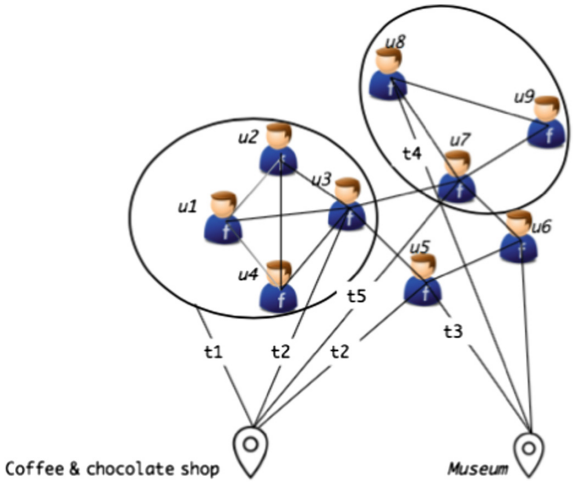

Fig. 2. Motivation example of location-based social network

Let us consider an example of a LBSN user u, who enquires where her friend *Majid* and his friends went for *coffee and chocolate* last year. People acquire each other's choices and interests e.g. user u likes the choice of her friend, *Majid*, in coffee and chocolate. To model this query, let $\boldsymbol{q}$ be a query that needs to retrieve all u's friends who visited a place p_u in *time t*. In addition, the query $\boldsymbol{q}$ has a keyword constraint that has the keywords of *coffee and chocolate*. Similarly, $\boldsymbol{q}$ can be asked to retrieve all places P_u visited by u's friends in a certain *time* t and keywords. The result set of $\boldsymbol{q}$ is $V' \subset P_u \subset V$.

3.2 Definitions

Based on the data and query model in the previous section, the following are definitions with which to draw up our framework.

Definition 1 (*User's Interests*): Let $U \subseteq V$ be a set of Users U who have CKs. Each user $u \in U$ is associated with a vector of keywords W_u. These keywords are extracted from places P, that have been visited by the user u, to represent users' interests as vectors in the space model.

Definition 2 (*Interest Weight*): Let each interest be $w \in W$ where W is a keyword set. Each $w \in u$ is associated with a weight named Relevance Score RS to indicate the interests' weights $\forall u \in U$.

Definition 3 (*Similarity-Based Graph SBG*): Given an attributed graph $G = (V, E, W)$, the SBG is a refined social graph constructed by computing interest-based similarities, which can be measured using a similarity function. The SBG can enhance the relationships between users regarding the user's interests. In addition, the SBG helps in returning accurate, query-based communities.

Definition 4 (*k-core* (Seidman 1983)): Given an integer $k \geq 0$, an existing connected subgraph $G(V')$ is called $\widehat{k-core}$ iff $\forall v \in G(V')$ has $deg(v) \geq k$, and $G(V')$ is connected.

Definition 5 (*Core Relevance Score CRS*): Given a query-based attributed subgraph $H \subseteq G$, and an interest w, the weight of interest w is RS as in Definition 2. Thus, CRS computes the relation between each subgraph H_i and their interest weight $\forall w \in H$

$$CRS_{H_i} = \frac{\sum_{w \in H_i} RS}{|H_i|} \tag{1}$$

where $|H_i|$ is the number of users $u \in H_i$.

Problem Definition: Given an undirected LBSN $G = (V, E)$, an integer k, $q \in V$, $w \in W_q$ and r, returns a subgraph $G_q \subset G$, which satisfies the following properties (Table 1):

Table 1. Query property

$Q = (q, k, w, r)$	
Property	Meaning
Connectivity	$G_q \subseteq G$ is connected and $q \in G_q$
Structure cohesiveness	$\forall v \in G_q, deg(v) \geq k$
Interest cohesiveness	Ensures that G_q has maximum CRS
Spatial range	A given radius r that ensures $\widehat{k-cores}$ are located within the range

3.3 Framework

In this section, we explain the framework and demonstrate how the three phases can interact with each other as one architecture. As shown in Fig. 3, our proposed architecture initially processes tags associated with places visited by users, followed by processing the social graph by linking each user with a vector of interests; each interest also carries weight. Once the attributed graph is created with associated vectors, we select all the pairs, which are unacquainted, but have at least one common neighbour, to measure the similarity between each pair. Based on comparing the similarity with a given threshold θ, we add any pairs that satisfy the minimum θ. Next, we compute the core decomposition. Finally, an index named AttriSpatial is created. The index is composed of two components - keywords and spatial -to handle the query $Q = (q, k, w, r)$.

User-Keyword TF-IDF Matrix. In this model, the well-known information retrieval model *Term Frequency Inverse Document Frequency TF-IDF* has been adapted to calculate the weight of users' interests. Accordingly, keywords W_p, extracted from places P_u that were checked into by a user u, are regarded as terms

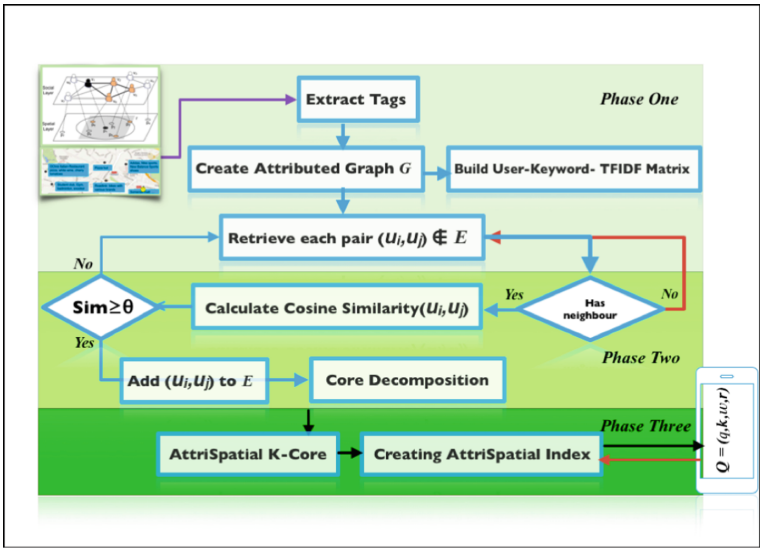

Fig. 3. Conceptual framework

and each user u is regarded as a document. Thus, each keyword of interest $w \in$ *user* u is represented as a dimension in the vector space model and, consequently, each user u_i is represented as a vector. To compute the weight of each keyword RS defined by Definition 3, the following is the $TFIDF$ model.

$$RS_{i,j} = \frac{f_{w_i,u_j}}{\sum_{w_i \in u_j} f_{w_i,u_j}} \cdot log \frac{\mid U \mid}{\mid u \in U : w_i \in u \mid} \tag{2}$$

where $f_{w,u}$ is the keyword frequency for each *user* u, and $\sum_{w \in u} f_{w,u}$ is the total number of a user's keywords W_u, which were acquired via place check-ins P_u and $log \frac{|U|}{|u \in U : w \in u|}$ calculates the inverse user frequency of keywords w.

Intra-similarity of Users. After phase one is complete, phase one output, which is the result of Algorithm 1, is used to input phase two. In this phase, the intra-similarity of users depends on each keyword relevance score associated with them. In such a case, we adopt the cosine similarity to calculate the similarity between a pair of users. Each user is represented by an interest weight vector x_i.

$$attr - sim(v_i, v_j) = \frac{x_i \cdot x_j}{\|x_i\|_2 \cdot \|x_j\|_2} \tag{3}$$

To guarantee that there is a minimum familiarity, each pair must have at least one common neighbour friend before adding them as friends.

On-line Query Processing. The aim of phase three is to search for the best cohesive query-dependent communities based on the attributed-spatial constraints. The cohesiveness of communities, returned by queries, attempts to align all constraints: keywords, social, and spatial. As a result of the output of the previous phases, during phase three we have produced a technique named *AttriSpatial K-Core*. In addition, a ranking function, called *Core Relevance Score CRS*, can be derived from the technique. The task of the derived function is to rank the retrieved communities based on the interest weight from phase one as well as the community structure from phase two. More significantly, this technique is employed to construct an efficient hybrid index to improve our dimensional query processing.

Data: $G = (V, E)$, *Users Check-ins*$CK = \{\langle u_i, p_k, t, W_{p_k}\rangle\}$
Result: *Attributed Graph with weighted attributes* $G = (V, E, X)$
begin
  initialisation
  $V \longleftarrow U$
  $W \longleftarrow \emptyset$
  $X[V, W]$
  for $v \in V$ **do**
    $X[v, RetKeyword(v)]$
  end
  for $v, w \in X[]$ **do**
    Compute RS
  end
end

Algorithm 1. Attributed graph extracting

4 Case Study of a Location-Based Social Network

For the sake of producing preliminary results, a conceptual framework (Armenatzoglou et al. 2013) has been reimplemented with modifications. Also, a dataset called *Weeplaces* has been used to evaluate our proposed hybrid approach.

4.1 Dataset

Weeplaces is a dataset (Liu et al. 2014) that has been collected from a website named Weeplaces, in which users' check-in activities can be visualised in LBSN. It has been integrated using the API of other well-known LBSNs, e.g. Facebook Places, Foursquare, and Gowalla. The dataset contains more than 7.5 million check-ins by 15,799 users across 971,309 geolocations. Most importantly, we have revealed that Weeplaces has a reasonable number of communities. As demonstrated in Fig. 4, various values of k affect the order of cores as defined in Definition 4.

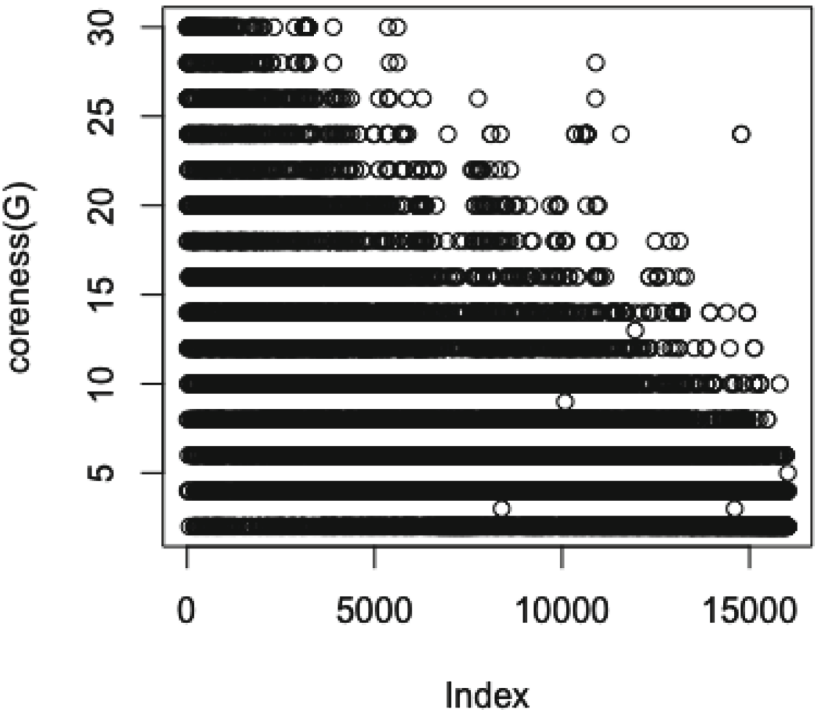

Fig. 4. Coreness distribution

4.2 Preliminary Results and Discussion

Setup. Two different data storage approaches have been employed for social and spatial layers. The two storage schemes have been implemented on MongoDB, a document-oriented database with the Python programming language. At the social layer, the social graph has been stored as a set of documents regarding the adjacency list representation. Moreover, each user has a list of keywords representing weighted interests. The spatial layer, which shows each place visited by a user, creates a document to represent that place. Thus, this document has a place ID, the visiting users' IDs, the tags that describe the place, and location coordinates.

Results. It is worth noting that our dataset has keywords to describe places that have been checked into by users. Keywords help us to understand users' interests by investigating the keyword frequency of each user compared to other users. Initially, we must represent the keywords distribution for the entire dataset over 100 users, in which each user is represented as a document. As shown in Fig. 5, on axis x, we plotted the graph attributes extracted by Algorithm 1. These attributes were associated with the weights calculated by the TFIDF schema in

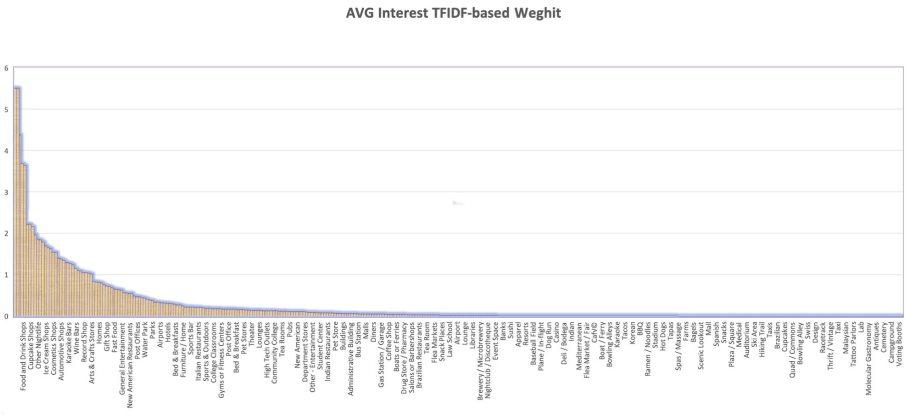

Fig. 5. Interest weights

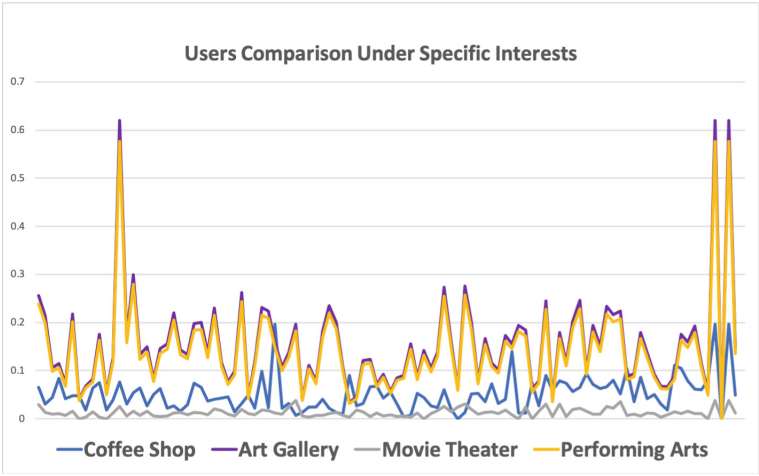

Fig. 6. Interest user comparison

Sect. 3.3. The goal is to differentiate between users using their interests. TFIDF is usually utilised as features to represent users as documents.

In Fig. 5, we grouped interests based on the global average TFIDF. The average will help us to select the appropriate threshold to retrieve query-based communities from which members contain attributes that are associated with greater weights than the threshold. The distribution in Fig. 5 demonstrates that a certain number of interests (graph attributes) are associated with weight score. This places these attributes in the area under the curve at which users share significant and representative interests, according to the feature selection technique, TFIDF. This leads us to the main contribution of our research; there may be other similar users who could increase the possibility of retrieving cohesive communities, in terms of interests.

We looked closely at Fig. 6, in which users are compared to each other under specific graph attributes (**Coffee Shop, Art Gallery, Movie Theatre and Performing Arts**). We found that users were had very similar interests (**Art Gallery and Performing Art**) due to their visits to similar places, although (**Coffee Shop and Movie Theatre**) were lower. Consequently, this outcome encouraged us to continue semantically investigating keywords associated with places that users prefer to visit.

Discussion. Previous related studies emphasise the significance of attributed graphs in community search. However, most of these studies investigated attributes, such as users' interests, as static keywords in community search. This study aims to integrate text analytic techniques into a framework of community search to keep users' profiles updated with their interests. This guarantees that communities retrieved by the framework are larger, more familiar, and more accurate than communities returned by community search models only.

5 Conclusion and Future Work

In this article, a recent body of knowledge regarding community search over a social graph has been reviewed. Specifically, two dimensions, including attributes and geolocations, have been investigated in more detail. This paper proposes a conceptual framework with preliminary results. Technically, this paper has shown that we can enrich users' profile interests by extracting keywords associated with places they visit. These interests have been analysed to produce an attributed social graph. As this work is part of ongoing research, in the future, we will conduct extensive experiments on various datasets of LBSN. Experimental work will include updating the social graph based on interest-based similarity among users as indicated in phase two. We will then create an efficient hybrid index that can handle different types of data, as explained in phase three. Furthermore, one of our future goals is to perform several comparisons, either using a baseline system or state-of-the-art work. Our future work will include the validation and feasibility of the proposed framework, in terms of effectiveness and efficiency.

References

Armenatzoglou, N., Papadopoulos, S., Papadias, D.: A general framework for geo-social query processing. Proc. VLDB Endowment **6**(10), 913–924 (2013). ISSN 21508097

Batagelj, V.: Efficient Algorithms for Citation Network Analysis. Networks, pp. 1–27 (2003)

Cui, W., Xiao, Y., Wang, H., Lu, Y., Wang, W.: Online search of overlapping communities. In: Proceedings of the 2013 International Conference on Management of Data - SIGMOD 2013, p. 277 (2013). ISSN 07308078

Cui, W., Xiao, Y., Wang, H., Wang, W.: Local search of communities in large graphs. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data - SIGMOD 2014, vol. 1, pp. 991–1002 (2014). ISSN 07308078

Fang, Y., Cheng, R., Li, X., Luo, S., Hu, J.: Effective community search over large spatial graphs. Proc. VLDB Endowment **9**(12), 1233–1244 (2016). ISSN 21508097

Fang, Y., Cheng, R., Chen, Y., Luo, S., Hu, J.: Effective and efficient attributed community search. VLDB J. **26**(6), 803–828 (2017). ISSN 0949877X

Huang, X., Cheng, H., Qin, L., Tian, W., Yu, J.X. Querying k-truss community in large and dynamic graphs. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data - SIGMOD 2014, vol. 2, pp. 1311–1322 (2014). ISSN 07308078

Li, R.-H., Qin, L., Yu, J.X., Mao, R.: Influential community search in large networks. Proc. VLDB Endowment **8**(5), 509–520 (2015). ISSN 2150-8097

Liu, Y., Wei, W., Sun, A., Miao, C.: Exploiting geographical neighborhood characteristics for location recommendation. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, CIKM 2014, pp. 739–748. ACM, New York (2014). ISBN 978-1-4503-2598-1. https://doi.org/10.1145/2661829.2662002

Seidman, S.B.: Network structure and minimum degree. Soc. Netw. **5**(3), 269–287 (1983). ISSN 03788733

Shang, J., Wang, C., Wang, C., Guo, G., Qian, J.: An attribute-based community search method with graph refining. J. Supercomput. **1**(1), 1–28 (2017). ISSN 1573-0484

Sozio, M., Gionis, A.: The community-search problem and how to plan a successful cocktail party. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 939–948 (2010)