Keyword-Based Retrieval of Frequent Location Sets in Geotagged Photo Trails

Paras Mehta
Freie Universität Berlin
Germany
paras.mehta@fu-berlin.de

Dimitris Sacharidis
Technische Universität Wien
Austria
dimitris@ec.tuwien.ac.at

Dimitrios Skoutas
IMIS, Athena R.C.
Greece
dskoutas@imis.athena-innovation.gr

Agnès Voisard
Freie Universität Berlin
Germany
agnes.voisard@fu-berlin.de

ABSTRACT

We propose and study a novel type of keyword search for locations. Sets of locations are selected and ranked based on their co-occurrence in user trails in addition to satisfying a set of query keywords. We formally define the problem, outline our approach, and present experimental results.

CCS Concepts

•**Information systems** → **Association rules;** *Location based services;*

Keywords

spatial keyword search; frequent locations; geolocated photos

1. INTRODUCTION

By posting geotagged photos or tweets and checking-in at various locations, users generate spatio-textual "trails" that establish implicit relations among locations, keywords and users. These can be exploited to extract movement patterns and identify users with similar interests and behaviors. For locations, they can help extract semantics, measure popularity and find visiting patterns or other associations. In this work, we introduce the problem of finding sets of locations in geotagged photo trails that are (a) *relevant* to the user's information need, expressed by a set of query keywords, and (b) *popular* in terms of the number of trails supporting them. We utilize a user's posts to assess both relevance and popularity. Specifically, a user *supports* a location set, if she has made a relevant post near each location in the set, and additionally these relevant posts collectively cover (i.e., contain) all query keywords.

The problem is related to collective spatial keyword queries [4, 2], which retrieve a set of results that are located as close as possible to each other and/or to the user's location, while collectively satisfying a set of keywords. In our case, we measure the relatedness among locations not based on their spatial proximity but based on their co-occurrence in user trails. Thus, the problem is also related to mining frequent itemsets [1]. User trails are viewed as transactions, and co-occurrence of locations in an itemset is determined by the existence of a user that has posts nearby all these locations. However, instead of a two-way association between locations and users, there is a three-way association among locations, users, and keywords.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WebSci '16 May 22-25, 2016, Hannover, Germany

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4208-7/16/05.

DOI: http://dx.doi.org/10.1145/2908131.2908204

2. PROBLEM AND APPROACH

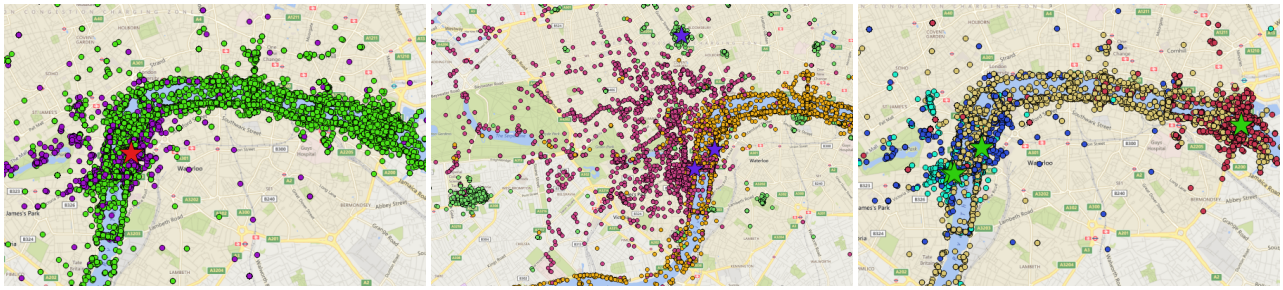
Assume a set of locations **L**, a set of users **U**, and a set of user posts **P**. Each post is a tuple $p = \langle u, \ell, \Psi \rangle$, where u denotes the user, $\ell = (lon, lat)$ the location, and Ψ a set of keywords characterizing the post. We say that a post p is *local* to location ℓ if the post's location is within distance ε to ℓ, i.e., if $d(p.\ell, \ell) \leq \varepsilon$, where d is a distance metric (e.g., Euclidean). Moreover, a post p is *relevant* to keyword ψ if the post's keyword set contains ψ, i.e., $\psi \in p.\Psi$.

Based on these, we define the notion of *support*, on which our approach is based. A user u *supports* a given location set L and keyword set Ψ, denoted as $u \in U_{L\Psi}$, if (a) for each keyword $\psi \in \Psi$, the user has made a post relevant to ψ and local to a location in L, and (b) for each location $\ell \in L$, the user has made a post local to ℓ and relevant to a keyword in Ψ. In other words, a user supports L, Ψ if she has posts that *simultaneously* cover the locality part L and the relevance part Ψ. This means there must exist a strong connection between a subset of the user's posts and given sets L, Ψ: each post in P should be both local to a location in L and relevant to a keyword in Ψ, and additionally each location in L should have a local post in P, and each keyword in Ψ should have a relevant post in P.

Accordingly, the *support* of a given location set L and keyword set Ψ is the number of users supporting L, Ψ, i.e., $sup(L, \Psi) = |U_{L\Psi}|$. Intuitively, a location set L is highly important with respect to a keyword set Ψ, if there exists a large number of users *supporting* this combination. Thus, given the definitions and concepts introduced above, our approach addresses the following problem: *Given a query keyword set* Ψ*, identify all the location sets, up to cardinality m, that are* frequent*, i.e., have support above a given threshold* σ.

3. EXPERIMENTAL RESULTS

We conducted a set of experiments using real-world data from the area of London. The dataset comprises 1,129,927 geolocated photos from Flickr [3], belonging to 16,171 different users and having in total 266,495 different tags. Moreover, we used as input 48,547 distinct locations, corresponding to Foursquare venues in the same area. To construct a query set, we extracted frequent tag combinations from the photos. We retrieved the top 100 most frequent tags, where the frequency of a tag was measured by the number of users having photos with that tag. From those, we manually picked a set of 30 tags, removing more generic ones, such as *"london"*, *"england"*, *"uk"*, *"iphone"*, *"canon"*, etc. Then, we combined these frequent tags to create tag sets of cardinality 2, 3 and 4. In each case, we selected the top 20 combinations according



(a) $\Psi = \{$*"london eye"*, *"thames"*$\}$ (b) $\Psi = \{$*"museum"*, *"thames"*, *"westminster"*$\}$ (c) $\Psi = \{$*"big ben"*, *"london eye"*, *"thames"*, *"tower bridge"*$\}$

Figure 1: Sample of results for London.

Table 1: Top 5 tag sets used as queries.

$\|\Psi\|$	Tag sets
2	london+eye, thames (922); big+ben, london+eye (908); thames, westminster (898); park, thames (880); big+ben, thames (846)
3	big+ben, london+eye, thames (557); big+ben, thames, westminster (497); big+ben, london+eye, westminster (472); london+eye, thames, westminster (464); park, thames, westminster (440)
4	big+ben, london+eye, thames, westminster (358); big+ben, london+eye, thames, tower+bridge (293); art, green, park, thames (258); green, park, thames, trees (257); park, statue, thames, westminster (257)

to the number of users having photos with those tags. The resulting tag sets were used as the query keyword sets in the experiments. Table 1 lists the top 5 combinations for different cardinalities.

Figure 1 presents the top result for three of the selected queries with different cardinalities. For each keyword in the corresponding query, we retrieve the list of users having photos with that keyword, and we intersect these lists to obtain a list of users having photos with all the query keywords. We display the locations of those photos on the map, using different colors for each keyword. Then, the location(s) contained in the top location set returned by our method are displayed with a star. Figure 1(a) illustrates the results for the query with keyword set $\Psi = \{$*"london eye"*, *"thames"*$\}$. The green (resp., purple) points denote the locations of photos that contain the tag *"thames"* (resp., *"london eye"*) and belong to a user that has also posted photos containing the tag *"london eye"* (resp., *"thames"*). Photos about *"thames"* are spread across the whole length of the river. On the other hand, London Eye is a landmark having a specific location; nevertheless, due to its high visibility, relevant photos can be found at various other locations, especially in and around St. James Park, for example. In this case, since London Eye is located at the banks of river Thames, the regions covered by the respective sets of relevant photos have a high overlap. In fact, the location set found to have the highest support for this query comprises a single location, which, as depicted in the figure, is situated in an area where a large number of photos containing both tags apparently exists.

The results for the query $\{$*"museum"*, *"thames"*, *"westminster"*$\}$ are illustrated in Figure 1(b). Two nearby but distinct locations are included in the top result corresponding to the river Thames and the Westminster Abbey. With respect to the keyword *"museum"*, we can observe in the figure that there exist (at least) two prominent regions with high density of relevant photos, namely one around the British Museum and one around the Natural History Museum and the Victoria and Albert museum. The former has been selected in the top result, indicating that this combination occurs more frequently. Finally, Figure 1(c) shows the top location set for the query

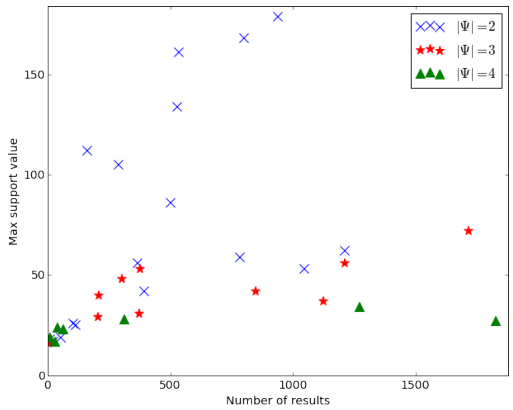


Figure 2: Query distribution in terms of number of results and maximum support value.

$\{$*"big ben"*, *"london eye"*, *"thames"*, *"tower bridge"*$\}$.

Figure 2 shows how the number of results and the support scores vary across queries of different keyword set cardinality. We computed the results for all queries with keyword set cardinality $|\Psi| \in [2,4]$, i.e., a total of 60 queries. For this experiment, we set the support threshold parameter to 16, i.e., 0.1% of the total number of users. For each query, we measured the number of results and the support of the top result. Queries having only two keywords tend to produce results with high support (e.g., up to around 3% of the total number of users). As the number of keywords in the query increases to 3 or 4, the maximum support among the returned results reduces significantly, dropping close to the support threshold; however, the number of returned results becomes much higher.

Acknowledgements

This work was partially supported by the EU Project City.Risks (H2020-FCT-2014-653747).

4. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[2] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi. Collective spatial keyword querying. In *SIGMOD*, pages 373–384, 2011.

[3] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[4] D. Zhang, Y. M. Chee, A. Mondal, A. K. H. Tung, and M. Kitsuregawa. Keyword search in spatial databases: Towards searching by document. In *ICDE*, pages 688–699, 2009.