Geo-spatial Domain Expertise in Microblogs

Wen Li[1], Carsten Eickhoff[2], and Arjen P. de Vries[3]

[1] Delft University of Technology, 2628 CD Delft, The Netherlands
wen.li@tudelft.nl
[2] ETH Zurich, 8092 Zurich, Switzerland
c.eickhoff@acm.org
[3] Centrum Wiskunde & Informatica, 1098 XG Amsterdam, The Netherlands
arjen@acm.org

Abstract. In this paper, we present a framework for describing a user's geo-spatial domain expertise in microblog settings. We investigate a novel way of casting the expertise problem by using *points of interest* (POI) as a possible categorization of expertise. To this end, we study a large-scale sample of geo-tagged tweets and model users' location tracks in order to gain insights into their daily activities and competencies. Based on a qualitative user study among active Twitter users, we present an initial exploration of domain expertise indicators on microblogging portals and design a classification scheme that is able to reliably identify domain experts.

Keywords: Domain expertise, Geo-tagging, Twitter.

1 Introduction

Empowered by affordable Internet-enabled mobile devices, many online services such as social networks or microblogging portals allow users to share their current geo-spatial context. The resulting data traces are a unique combination of digital and real-world activity that allow for a range of interesting academic and industrial applications including location prediction [6], localized search personalization [4], or contextual advertisement [2].

Expert finding is concerned with identifying those individuals that are most knowledgeable about a given topic. This task was originally applied for locating expertise holders in corporate settings [1] and has since then been extended to a wide range of scenarios. Expertise is typically modelled along a number of pre-defined topics and is estimated based on the users' historic activity in the form of document authorship or project participation.

In this paper, we investigate expertise in terms of a user's knowledge about a place or a class of places. Previous work has found a strong connection between places and their function [7], making a *point of interest* (POI) a proxy for the typical range of activities that are carried out there. Based on this finding, we hypothesise that a user's location tracks constitute evidence for expertise towards the place's function. In this way, we model which of your friends to ask

M. de Rijke et al. (Eds.): ECIR 2014, LNCS 8416, pp. 487–492, 2014.

for advice on "historic museums in Shanghai" or who to best turn to for a menu recommendation at that "new tapas place that opened down town". We rely on Twitter and Foursquare as data sources for investigation and experimentation.

Our work makes 3 novel contributions beyond the state of the art in domain expertise modelling and expert finding. **(1)** We propose a novel domain expertise framework based on the topology of points of interest. **(2)** We conduct a survey among active Twitter users in order to better understand their usage of geo-tagged tweets. **(3)** Based on the insights gained from an initial user study, we design and evaluate an automatic method that is able to reliably identify domain experts.

2 Related Work

The task of expertise retrieval originated in the domain of enterprise search, in which authorship of documents or group affiliations are used as evidence to determine an individual's topical expertise [1,5]. For social networks like Twitter, Wagner *et al.* [8] suggested using external resources other than tweet content for identifying users' expertise in order to overcome the high amounts of noise pertaining to the domain. Bar-Haim *et al.* [3] tried to identify stock experts on Twitter by evaluating their expertise according to stock market events and their tweeted buys and sells. Weng *et al.* [9] combined knowledge from topics (distilled by LDA) and social networks to produce a topic-specific network and used random walk methods to find topic-specific influential users (experts on the topic).

In this paper, we cast the problem in analogy to the structure of the real world rather than document-derived structures and topic-specific influential users by finding experts knowledgeable about a location or a class of locations. In previous works, geo-location-related information is generally under-represented. To capture such knowledge, we rely on the POI-tags on tweets rather than the tweets themselves to profile our candidates. To the best of our knowledge, this is the first attempt to the problem.

3 Methodology

Many social media applications represent location information in the form of so-called geo-tags added to the original content. For example, one may post a tweet, "*I really love sandwiches here*", with a POI-tag *Blue Barn Gourmet* containing detailed information about the place. In the context of Foursquare, such geo-tagged messages are also referred to as check-ins. On Twitter, geo-tagged tweets contain a pair of coordinates, a place name, and an address. On Foursquare, a check-in may also include the category that the place belongs to. Categories on Foursquare are organized in a multi-level hierarchy, effectively forming an ontology of geographic entities. In this paper, we refer to the categories at the top level as top-level-categories, and other categories (at lower levels) just as

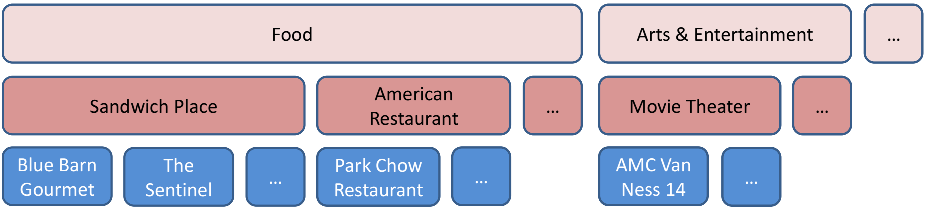

Fig. 1. POI and POI-Category Hierarchy

categories. For example, as shown in Figure 1, Blue Barn Gourmet is a sandwich place in San Francisco, CA, which is categorized as Food at the top level.

Intuitively, an expert knowledgeable in a given topic should have many contact points with it. In previous works, those contact points are modelled as authorships or topical friendships. In this paper, we focus on the check-in activity and postulate three properties a good pair of geo-topic (a location or a class of locations) and geo-expert should satisfy.

Within-Topic Activity. ($\mathcal{S}_n$) The first property is general activity at a given location or category. Intuitively, the more frequently the user interacts with the topic l, the more they will know about it. Since we can only measure check-ins ($C_e = \{(l_c, t_c)\}$) of an expert e at time t, we may not be able to capture all actual physical visits. The check-ins instead represent a lower bound to the number of visits. In this way the expertise can be measured by $\mathcal{S}_n(e, l) = \sum_{(l_c,t_c)\in C_e} \mathbf{1}(l_c = l)$, where $\mathbf{1}(.)$ is an indicator function that equals 1 if and only if the condition in the parenthesis is satisfied, 0 otherwise.

Within-Topic Diversity. ($\mathcal{S}_d$) Secondly, we require an expert to know more than just a single instantiation of a category or top-level category. Accordingly, we consider check-ins to a large number of different POIs within a category a stronger indication of expertise than only check-ins at a single location. That is $\mathcal{S}_d(e, L) = \sum_{l\in L} \log \sum_{(l_c,t_c)\in C_e} \mathbf{1}(l_c = l)$.

Recency. ($\mathcal{S}_r$) Finally, we require the evidence of expertise to be as fresh as possible. Old check-ins may not represent the user's current range of interests and occupations accurately any more. That is $\mathcal{S}_r(e, l) = \sum_{(l_c,t_c)\in C_e} e^{t_c - t}\mathbf{1}(l_c = l)$.

4 Understanding Geo-spatial Expertise

In order to gain a better understanding of how people use geo tags in their tweets, we issued a survey among active users of the microblogging portal Twitter. The survey was distributed via Crowdflower (`http://crowdflower.com`) and required workers to have a Twitter account to ensure the subjects' familiarity and personal experience with the domain and terminology. A total of 164 forms were received. In the following, we discuss the main findings and implications.

4.1 Geo-spatial Recommendations

Our first research question **(Q1)** is concerned with *how often, in which way, and to whom are people looking for, or giving poi-advice?* We asked our survey participants an initial set of three questions: A) "How often do you ask your friends, family, colleagues or any other people for advice about a place to go?" B) "Which of the following groups do you trust the most when they give you suggestions on places to go?" C) "If you need advice, but you do not know who can help, which of the following channels do you prefer?" For each question, a number of options as well as the possibility to give free-text answers were offered. Participants were invited to select more than one option if applicable. Figure 2 shows the answer distribution for the questions. Most participants stated they rely on location advice from time to time. Only 13% replied that they generally prefer to research places themselves. When it comes to accepting advice, trust in the advisor seems to be a key issue. We observe a clear preference order favouring family and friends over on-line contacts or even unknown review writers. In the case that no friend or family member knows advice, a broadcast to the personal social network and twitter followers is favoured over posting to forums or starting a blog post.

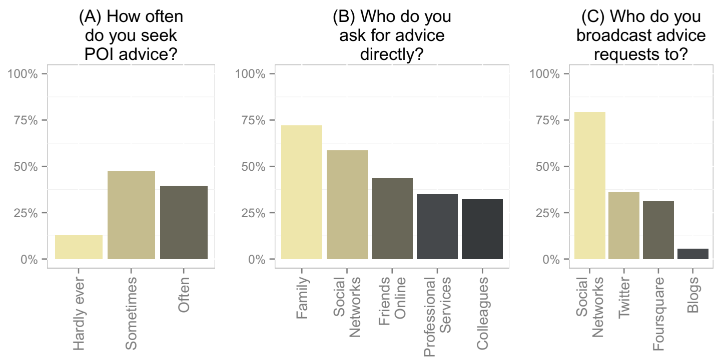

Fig. 2. Geo tag usage on Twitter and Foursquare

4.2 Measures of Expertise

Previously, we established that on-line communication channels are a realistic source of POI-related recommendations and advice. Trust in the advisor appears to be a key issue especially on the Web. Consequently, in our second research question **(Q2)**, we investigate *how to determine the geo-spatial expertise of a person.* In a short experiment, we showed our survey participants examples of anonymous Twitter profiles along with their geo-tagged check-in history. The judges were supposed to determine geo-spatial domain expertise for several profile/topic pairs. Afterwards, the participants were asked to explain their reasoning process and detail which, if any, of the rules presented in Section 3 they used to make their decision. Again, multiple answers were possible. This question resulted in a clear ordering of criteria according to the frequency at which they were chosen by participants: Within-topic activity (70.1%) > Within-topic diversity (48.8%) $\geq$ Recency (45.1%) > General activity (34.8%).

4.3 Predicting Expertise

To complete off our inspection of geo-spatial notions of expertise, we now aim to make the qualitative insights from our survey usable in an algorithmic manner. Our third and final research question **(Q3)** is therefore, *how well do automatic representations of the survey participants' criteria predict POI-related expertise?*

To achieve this, we algorithmically render different variants of our rules presented in Section 3. Each version aggregates a candidate's check-ins at a location or category. For the variant emphasizing recency, we discounted old check-ins, resulting in candidates with more recent check-ins being ranked at the top. For the variant emphasizing diversity, we discounted repetitive check-ins at the same place, *i.e.*, multiple check-ins at the same place will be given less importance than a check-in at a previously unseen place.

As a first qualitative evaluation of our methods, we approached active users of the Twitter geo-tagging functionality that were dominant in a collection of tweets we crawled between June and August, 2013. A group of 10 such users volunteered to work with us. We presented each of them with our model's predictions of their individual expertise and asked them to judge their actual knowledge about the topic on a 5-point scale from 1 ("I do not know about this") to 5 ("I am an expert"). 6 participants indicated high expertise (Grades 4–5) in the topics predicted by our method. Another 3 reported reasonable competency (Grade 3) towards the predicted topics. Only a single participant indicated mild expertise (Grade 2) towards the predictions. Given the high dimensionality of our problem space (400 categories and thousands of individual POIs), the results of our initial qualitative evaluation study look very promising and encourage future quantitative confirmation in a Web-scale setting.

5 Conclusion

In this paper, we presented a novel categorization of domain expertise along the hierarchy of *points of interest* (POI) as observed on Twitter and Foursquare. We conducted a qualitative user study and investigated the way in which Twitter and other communication channels are used for searching, receiving and giving location-related advice and recommendations. Doing so, we found on-line communication with close friends and family or even the wider social network to be among the major channels for obtaining advice. We also presented participants with examples of Twitter streams and asked them to judge the expertise of the showcased user towards a number of topics, as well as, to explain which criteria influenced their decision. Within-class coverage and diversity turned out to be the most frequently named features. On the basis of these qualitative insights, we designed an automatic classification method that was able to reliably predict domain expertise with high agreement to the profiled persons' self assessment.

This paper describes an ongoing piece of work in progress. There is a substantial amount of envisioned extensions that would have gone beyond the limits of this work. **(1)** In this paper, we reported the results of a user study at limited scale. In the future, we will take our qualitative insights to Web scale and quantitatively verify their performance on a realistic sample of the Twitter user base. **(2)** In the future, we will investigate using the textual tweet content as additional evidence for domain expertise. **(3)** Geo tags are a powerful type of semantic annotation that has been demonstrated to hold significant potential for a wide range of applications. Unfortunately, their coverage amounts to less than 1% of the overall Tweet volume. It is therefore crucial to investigate bootstrapping methods that can help annotate untagged tweets with latent geo tags based on their content or temporal dynamics.

References

1. Balog, K., Azzopardi, L., De Rijke, M.: Formal models for expert finding in enterprise corpora. In: SIGIR 2006, pp. 43–50. ACM (2006)
2. Banerjee, S., Dholakia, R.: Mobile advertising: does location based advertising work? International Journal of Mobile Marketing (2008)
3. Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., Goldstein, G.: Identifying and following expert investors in stock microblogs. In: EMNLP 2011, pp. 1310–1319 (2011)
4. Bennett, P.N., Radlinski, F., White, R.W., Yilmaz, E.: Inferring and using location metadata to personalize web search. In: SIGIR 2011, pp. 135–144. ACM (2011)
5. Campbell, C.S., Maglio, P.P., Cozzi, A., Dom, B.: Expertise identification using email communications. In: CIKM 2003, pp. 528–531 (2003)
6. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: CIKM 2010, pp. 759–768. ACM (2010)
7. Li, W., Serdyukov, P., de Vries, A.P., Eickhoff, C., Larson, M.: The where in the tweet. In: CIKM 2011, pp. 2473–2476. ACM (2011)
8. Wagner, C., Liao, V., Pirolli, P., Nelson, L., Strohmaier, M.: It's Not in Their Tweets: Modeling Topical Expertise of Twitter Users. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, pp. 91–100 (2012)
9. Weng, J., Lim, E.-P., Jiang, J., He, Q.: TwitterRank: Finding Topic-sensitive Influential Twitterers. In: WSDM 20210, pp. 261–270 (2010)