Towards Personalized Maps: Mining User Preferences from Geo-textual Data

Kaiqi Zhao[1] Yiding Liu[2] Quan Yuan[3] Lisi Chen[4] Zhida Chen[5] Gao Cong[6]

Nanyang Technological University

{[1]kzhao002@e.,[2]ydliu@,[3]qyuan1@e.,[5]chen0936@e.,[6]gaocong@}ntu.edu.sg

Hong Kong Baptist University

[4]chenlisi@comp.hkbu.edu.hk

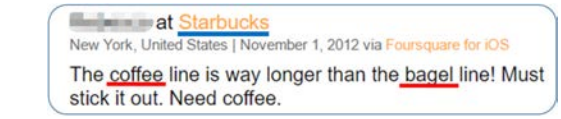
ABSTRACT

Rich geo-textual data is available online and the data keeps increasing at a high speed. We propose two user behavior models to learn several types of user preferences from geo-textual data, and a prototype system on top of the user **pre**ference models for **min**ing and s**ear**ch geo-textual data (called PreMiner) to support personalized maps. Different from existing recommender systems and data analysis systems, PreMiner highly personalizes user experience on maps and supports several applications, including user mobility & interests mining, opinion mining in regions, user recommendation, point-of-interest recommendation, and querying and subscribing on geo-textual data.

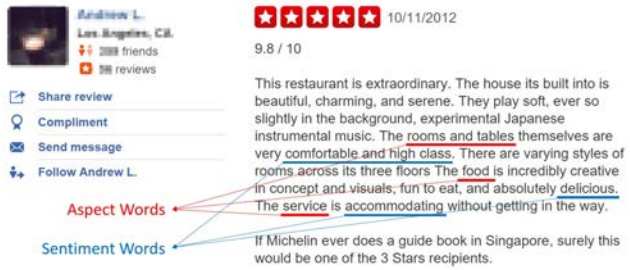
1. INTRODUCTION

People post a variety of content to the internet everyday through GPS-equiped mobile devices. Such posts are associated with geographical coordinates (latitude and longitude), and some of them are associated with semantic places, i.e., points-of-interest (POIs). They also contain words that imply semantic topics (see Figure 1(a)), or words that imply user's opinions on different aspects of a POI (see Figure 1(b)). With multiple types of information available from geo-textual posts, we face a great opportunity to mine different kinds of user preferences, including preferences on topic, region, POI aspect, and category. For example, a user who prefers topic "sports" may often mention words like "shoot" and "goal" in their posts. As another example, a user may frequently visit shops in a shopping area she likes (i.e., preferences on region).

However, building a unified model that captures different types of user preferences poses three main challenges. First, the interactions among different types of latent variables (e.g., aspect, sentiment, region, topic) and observable variables (e.g., text, time, category, POI) are unclear. Second, the data could be in different formats (continuous and discrete) from different data sources (e.g., Yelp and Foursquare). The variety of data makes the modeling and parameter learning complicated. Third, the latent variables in different scopes further complicate the model learning. For example, each sentence in a review is often related to one aspect and the whole review should be posted in some latent region. This implies that aspect and sentiment are often modeled in the scope of sentence, while region is modeled in the scope of document.



(a) Words about topic "eat & drink" in a check-in post



(b) Words about different aspects (environment, food and service) and their corresponding sentiments in a review

Figure 1: Screenshot of short text (e.g., Foursquare check-ins) and long text (e.g., Yelp reviews) data

To tackle these challenges, we design two probabilistic models, namely **W**ho, **W**here, **W**hen, **W**hat (W4) model [7, 8] and **S**entiment, **A**spect, **R**egion (SAR) model [9] for short text and long text data, respectively. W4 mines user preferences on topics and regions from short geo-textual documents with temporal information (e.g., check-ins), while SAR mines user preferences on aspects, categories and regions from geo-textual documents in which temporal information is not available but the text is long enough for sentiment analysis (e.g., geo-tagged reviews). Both user behavior models support mining several types of user preferences and hence cater the needs for various applications. The proposed models achieve better performances than other models in many applications. For example, SAR achieves at least 60% higher accuracy than other models in POI recommendation, and W4 performs at least 80% more accurate than other models in location prediction.

In this demonstration, we propose a prototype system[1], namely PreMiner, which is built on top of the two user behavior models. Our system supports querying and mining geo-textual data for personalized map services, based on the two models and techniques proposed in our previous work [1, 3]. It supports, but not limited to the following applications:

[1]The system is available at `http://spatialkeyword.sce.ntu.edu.sg/PreMiner`.

Proceedings of the VLDB Endowment, Vol. 9, No. 13

User mobility & interests mining: We support mining user mobility patterns and their interests in *regions* (e.g., shopping centers), *categories* (e.g., restaurants), *aspects* (e.g., price of a restaurant), and *topics* (e.g., sports).

Aspect analysis in a region: PreMiner returns the most positive and negative aspects in the user specified region and category. Market investigators can use the results to improve their businesses.

User recommendation to business: The business (POI) owner may want to promote some of its aspects (e.g., discounts), we recommend the users who may be interested in it.

Personalized POI recommendations: User specify different criterions (e.g., aspect, category) to receive POI recommendations.

Personalized search and subscription: We also support different types of personalized spatial-keyword queries and subscribing streaming data (e.g., tweets) on POIs or regions.

2. RELATED WORK AND NOVELTY

Studies on geo-textual data gain much attention in recent years [1, 6, 5]. Most of the existing user behavior models on geo-textual data consider user interests in either regions or topics, but ignore other features such as aspect and category. The lack of support for different types of user interests makes them only applicable to few applications such as context-free POI recommendation. Moreover, existing search / recommender systems built on geo-texual data [1, 6, 5] also target at specific applications such as spatial object querying or POI recommendation. However, personalized map users may have different intended needs (e.g., business needs analysis, personalized recommendations and search). Existing models and systems cannot fulfill different application needs.

Our prototype system is built on top of two novel user behavior models, namely SAR and W4 [7, 8, 9], to satisfy different application requirements. Compared to the previous user behavior models, SAR is the first unified model that explores user's preferences on aspect, category, and region. It automatically and simultaneously learns latent aspects (each represented by a distribution over words) and regions (each represented by a Gaussian distribution and a word distribution), and conducts sentiment analysis on each aspect. None of existing methods manage to model all the factors in a unified model. W4 model is novel in that (1) it takes four factors of a geo-textual document, i.e., user, POI, time and text, into account, while other models consider only some of them; and (2) it models user behaviors with personalized regions. Compared to other models who mine regions from all users, the personalized regions in W4 are more appropriate for characterising each user's mobility patterns. Both models support mining more types of user preferences than existing models because they capture the interdependencies among different factors. In the view of application, PreMiner also supports new application scenarios such as aspect analysis and user recommendation.

3. SYSTEM OVERVIEW

We first present the architecture of PreMiner, and then outline the underlying techniques.

3.1 System Architecture

We show the structure of PreMiner in Figure 2. PreMiner contains a client side and a server side. The client side is a user interface for users to interact with the functionalities of the system. The server side consists of offline and online components. The offline components are used to collect and process geo-textual data and provide suitable application program interfaces (APIs) for the online components. The online components response user's requests for different applications.

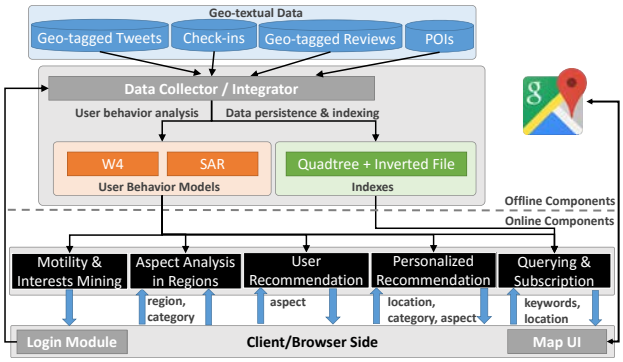


Figure 2: System Overview of PreMiner

Offline components: PreMiner contains three offline components, namely *Data Collector*, *User Behavior Models*, and *Indexes*.

1) *Data Collector*: The data collector grabs and integrates the historical data from difference sources (e.g., Twitter, Foursquare and Yelp). It also keeps monitoring the streaming data to support pushing news for the online subscription component.

2) *User Behavior Models*: The user behavior models analyze user's historical geo-textual posts to reveal the user's interests in regions, topics, categories, and aspects. The user preferences are then used by all the online applications.

3) *Indexes*: To support efficiently updating and querying of geo-textual data (Foursquare check-ins and Yelp reviews), we use a simple hybrid data structure that combines a shallow Quad-tree and inverted file to index geo-textual data, which is called *IQ-tree* [2]. The IQ-tree is also used to index subscription queries if there are a large number of such queries.

Online components: The online components include solutions for different applications using the user preferences learnt by the user behavior models in the offline part. The applications include *mobility & Interests Mining*, *Aspect analysis in regions*, *User recommendation to business*, *Personalized POI recommendation* and *Personalized querying & subscription*. The details of each application will be presented in Section 3.3.

3.2 Probabilistic User Behavior Models

For different types of data, we adopt different strategies on analyzing user's preferences. For long texts (e.g., geo-tagged reviews), we design a model (SAR) [9] that considers user preferences on regions, categories and aspects. For short texts (e.g., check-ins), since there are fewer words about aspect, we cannot analyze the aspects and sentiments precisely. Nevertheless, the temporal information is available because the check-in time is the right time when the user was in the location, while geo-tagged reviews may be posted several hours or days after the user visited a place. Thus, we design another time-aware model (W4) [7] that considers user's preferences on regions and topics.

SAR model: For long text data, we develop a Bayesian model called **S**entiment **A**spect **R**egion (SAR) model [9]. In this model, we mine user preferences on latent regions, aspects, and categories. For example, a user may have some preferred regions for shopping, and she may prefer to visit bargain shops rather than expensive ones. Different from the previous user behavior models, our model considers user mobility mining and sentiment analysis simultaneously. Our model has at least two benefits compared with other

models. First, it can discover more accurate user preferences on regions because it considers the user's sentiments. A user keeps posting negative reviews in a region may not like the region. Second, our model can discover a user's preferences on several attributes, including regions, categories, and aspects. These preferences could be used for various applications such as aspect analysis in regions and personalized recommendations.

SAR models user preferences on aspects, categories, and latent regions as conditional probabilities, i.e., $p(a|c,u)$, $p(c|u)$, and $p(r|u)$. The variables u, a, c, r are user, aspect, category, and region, respectively. User may select a category of POIs or a region to visit according to these preferences. Along with the user preferences, we build SAR with the following observations: (1) A user may choose to visit a POI l by considering its category and location, we take the user's selected region and category into account to model the probability of visiting a POI, i.e., $p(l|r,c)$; (2) The user may write words **w** on the POI by considering some aspect and the location information, and thus we model the probability of writing each word w conditioning on aspect, sentiment, and region, i.e., $p(w|a,s,r)$; (3) Each POI has overall probabilities (profiles) of different sentiments (positive, neutral, and negative) for each of its aspects $p(s|a,l)$. To this end, we compute the likelihood of the training data as:

$$p(D) = \prod_{d}^{D} p(u_d) \sum_{r} p(r|u_d) p(l_d, \mathbf{w}_d|r, u_d), \quad (1)$$

where D is the set of documents and u_d is the author of document d. The probability $p(l_d, \mathbf{w}_d|r, u_d)$ is computed by applying Bayes' rule to the aforementioned conditional probabilities.

Because SAR has latent variables (regions and aspects) in different scopes, it is hard to estimate the parameters by maximum likelihood methods. To address the technical challenge, we propose a two-level expectation-maximization (EM) algorithm. EM algorithm iteratively estimates the posterior distribution of latent variables (E-step) and updates the model parameters (e.g., the user preferences) by maximizing the likelihood computed by Equation 1 (M-step). In each EM iteration, our two-level EM algorithm estimates the latent regions in the first level and the aspects with the regions fixed in the second level. As a result of model learning, we output the user preferences and POI profiles.

W4 model: For short text data, we build a Bayesian model called **W**ho, **W**here, **W**hen, **W**hat (W4) model [7, 8]. This model considers four factors of a post (i.e., user, location, time and words) to model user's interests in regions and topics in each region. For instance, a user may visit office in weekdays while visit a shopping center in weekends. The user may prefer to talk about football in a bar but foods in a restaurant. Compared to the previous models, our model captures the relationships among the four factors, while previous models only consider some of them. In addition, W4 personalizes regions to better fit users' mobility patterns, while other models fit regions base on the visiting history of all users.

Similar to SAR, W4 models user preferences on topics, regions and time as conditional probabilities, i.e., $p(z|u)$, $p(r|u,t)$, $p(t|u)$. The variables u, z, t, r are user, topic, time, and region, respectively. The time dimension is divided into 1-hour slots. We build W4 with the following observations: (1) The user visits a POI l by considering the region and topic, i.e., $p(l|z,r)$; and (2) The user writes each word on the POI according to the region and topic, i.e., $p(w|z,r)$. The likelihood of the training data is then computed as:

$$p(D) = \prod_{d}^{D} p(u_d) \sum_{r} p(r|u_d, t_d) \sum_{z} p(l_d|r,z) p(\mathbf{w}_d|r,z), \quad (2)$$

where D is the set of posts and u_d is the author of post d.

The technical challenge of training W4 lies in formulating $p(l|z,r)$. Note that the coordinates in regions are continuous, while popularity of a POI in topics is discrete. To overcome this problem, we propose to standardize the continuous distributions (e.g., Gaussian) to categorical distributions to allow both types of distributions in a unified model. Based on the standardizing method, we use EM algorithm to infer the model parameters.

Online applications of SAR and W4: Since SAR and W4 model different types of user preferences for different types of data, they can be applied to different applications. SAR is applicable to applications that require aspect analysis, i.e., user interests mining on aspects, aspect analysis in regions. W4 is applicable to mining user interests in topics and temporal mobility patterns in personalized regions, and personalized search and subscription. Both models are applicable to POI recommendation and user recommendation.

3.3 Online Mining & Querying

We next proceed to explain how to make use of the learnt user preferences to support the online applications.

Mobility & interests mining for users: For both SAR and W4, we use the probabilities (e.g., $p(a|c,u)$, $p(c|u)$) learnt by the models to rank regions, categories, aspects and topics, and then pick top ones of each type to represent the interests of a user. For each region in W4, we further use the probability $p(r|u,t)$ to capture the user's temporal mobility patterns in the region.

Aspect analysis in regions: This is a new scenario for business analysis supported by SAR. Given a user specified region and category, we return the top positive (and negative) aspects in the region and category. We apply Bayes' rule to compute the probability of being positive / negative given the specified category and region for each aspect. The aspects that with highest probability to be positive / negative are returned to the user.

User recommendation to business: This function allows business owners to find customers that may be interested in their advertisement on some aspects (in SAR) or topics (in W4). The ranking score of a user is computed by multiplying the user preferences with the profile of the business in both models. We return to the business owner a list of top ranked users.

Personalized POI recommendation: We compute the ranking score of a POI by comparing the user preferences learnt from SAR/W4 to the POI's profiles as proposed in our previous work [7, 9]. In addition to context-free POI recommendation, PreMiner supports user specified contexts (e.g., aspect, category and region) by computing the ranking score only with respect to the specified contexts.

Personalized search & subscription: PreMiner supports personalized search and subscription by adding the user's topic interests learnt by W4 into the ranking function of our techniques [4, 2], i.e., topic similarity is considered together with spatial proximity.

4. DEMONSTRATION

Dataset: PreMiner is built on three datasets: Foursquare, Yelp and Twitter. Tweets are only used for personalized search and subscription. The Foursquare and Yelp data are used to train W4 and SAR, respectively. We collect 4 million tweets from 144K users in Singapore and keep tracking new tweets using the Streaming API[2]. For Foursquare, we crawl 18K check-ins from 3,691 users and all 323K POIs from Singapore. For Yelp, we adopt the data from the 10 cities published by Yelp[3] and collect the data in Singapore, which results in 368K users, 70K POIs and 1.62M reviews.

[2]https://dev.twitter.com/streaming/

[3]http://www.yelp.com/dataset challenge

Mobility & interests mining: When a user logs in, PreMiner will automatically retrieve her interests learnt by SAR/W4, For a new user without any historical data, PreMiner asks the user to sign up with Foursquare or Yelp account, or to specify some preferred POIs and keywords based on which it estimates the initial user interests. User's interests in topics (for Foursquare user), categories and aspects (for Yelp user) will be shown in the user's interests panel while personalized regions will be drawn on the map. Topics in W4 are expressed as a set of keywords while aspects in SAR are manually named according to the word distributions of the aspects. Foursquare user can explore the temporal mobility patterns and the keywords in the personalized regions (see Figure 3). When the user clicks on one of her mobility region, e.g., region B, the keywords shown in the right side of Figure 3 represent the user's activities and interests in the region, and the time patterns indicate the probability of the user staying in region B at different time.

Figure 3: Temporal mobility patterns in personalized regions

Aspect analysis in a region: As shown in Figure 4, user can arbitrarily draw a region on the map and specify a category. PreMiner will return the most positive and negative aspects to the user.

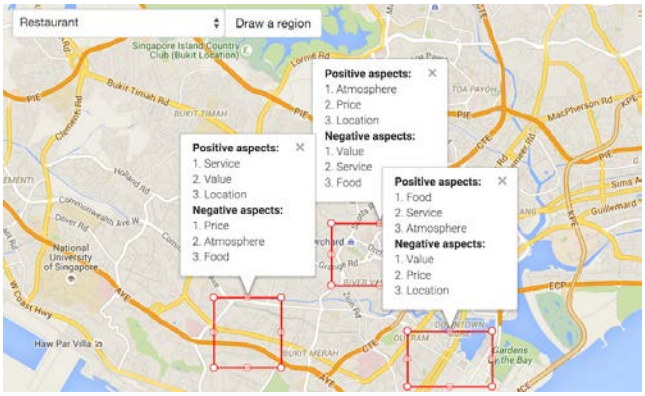


Figure 4: Aspect analysis in regions

User recommendation to business: Business owner can select one POI and submit an aspect or topic (learnt offline) according to their advertisement theme (e.g., select the aspect "price" for an discount event). PreMiner finds the users who have high preference on the aspect and may rate positive for the POI. We also report the users' interests in aspects and topics, and their approximate distances to the POI by averaging on their activity regions in the result panel. Figure 5 shows an example (aspect = "service").

Personalized POI recommendation: In PreMiner, context-free recommendation is made automatically based on user interests when

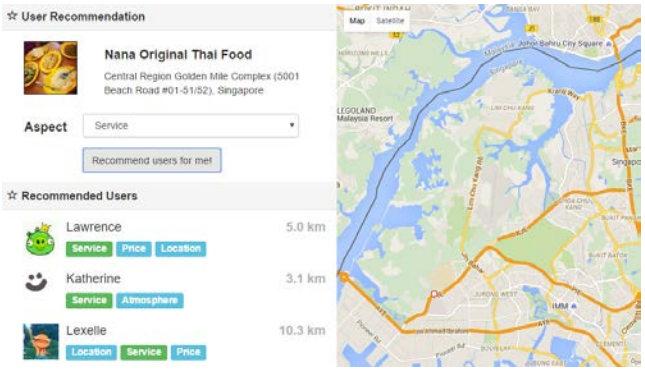


Figure 5: User recommendation to business

the user logs in. In addition, if the user has further requirements to be satisfied, she can specify a location and choose preferred category (e.g., restaurant) and aspect (e.g., price) to initiate context-aware POI recommendation. The recommended POIs are shown together with some aspects that the user may like.

Personalized search & subscription: Users can query for tweets and POIs that satisfy her interests and the query (keywords and location). On each retrieved POI or any arbitrary region drawn by the user, the user can subscribe streaming geo-textual data (e.g., tweets) that are spatially and textually relevant to the POI / region. New messages that match the subscription will be fed to the user and displayed below the user profiles.

Acknowledgment This demo was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Prime Minister's Office, Singapore, under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office. This work is also supported in part by a Tier-1 grant awarded by Ministry of Education Singapore (RG22/15). In addition, this work was done when Lisi Chen was in Nanyang Technological University.

5. REFERENCES

[1] X. Cao, G. Cong, C. S. Jensen, J. J. Ng, B. C. Ooi, N.-T. Phan, and D. Wu. Swors: A system for the efficient retrieval of relevant spatial web objects. *PVLDB*, 5(12):1914–1917, 2012.

[2] L. Chen, G. Cong, and X. Cao. An efficient query indexing mechanism for filtering geo-textual data. In *SIGMOD*, pages 749–760, 2013.

[3] L. Chen, Y. Cui, G. Cong, and X. Cao. SOPS: A system for efficient processing of spatial-keyword publish/subscribe. *PVLDB*, 7:1601–1604, 2014.

[4] G. Cong, C. S. Jensen, and D. Wu. Efficient retrieval of the top-k most relevant spatial web objects. *PVLDB*, 2(1):337–348, 2009.

[5] R. Deveaud, M. Albakour, J. Manotumruksa, C. Macdonald, I. Ounis, et al. Smartvenues: Recommending popular and personalised venues in a city. In *CIKM*, pages 2078–2080, 2014.

[6] A. Magdy, L. Alarabi, S. Al-Harthi, M. Musleh, T. M. Ghanem, S. Ghani, S. Basalamah, and M. F. Mokbel. Demonstration of taghreed: A system for querying, analyzing, and visualizing geotagged microblogs. In *ICDE*, pages 1416–1419, 2015.

[7] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Who, where, when and what: Discover spatio-temporal topics for twitter users. In *KDD*, pages 605–613, 2013.

[8] Q. Yuan, G. Cong, K. Zhao, Z. Ma, and A. Sun. Who, where, when, and what: A nonparametric bayesian approach to context-aware recommendation and search for twitter users. *ACM Trans. Inf. Syst.*, 33(1):2:1–2:33, 2015.

[9] K. Zhao, G. Cong, Q. Yuan, and K. Zhu. Sar: A sentiment-aspect-region model for user preference analysis in geo-tagged reviews. In *ICDE*, pages 675–686, 2015.