A Location-Based News Article Recommendation with Explicit Localized Semantic Analysis

Jeong-Woo Son A-Yeong Kim Seong-Bae Park
School of Computer Science and Engineering
Kyungpook National University
Daegu, Korea
[jwson, aykim, sbpark]@sejong.knu.ac.kr

ABSTRACT

The interest of users in handheld devices is strongly related to their location. Therefore, the user location is important, as a user context, for news article recommendation in a mobile environment. This paper proposes a novel news article recommendation that reflects the geographical context of the user. For this purpose, we propose the Explicit Localized Semantic Analysis (ELSA), an ESA-based topical representation of documents. Every location has its own geographical topics, which can be captured from the geo-tagged documents related to the location. Thus, not only news articles but locations are also represented as topic vectors. The main advantage of ELSA is that it stresses only the topics that are relevant to a given location, whereas all topics are equally important in ESA. As a result, geographical topics have different importance according to the user location in ELSA, even if they come from the same article. Another advantage of ELSA is that it allows a simple comparison of the user location and news articles, because it projects both locations and articles onto an identical space composed of Wikipedia topics. In the evaluation of ELSA with the New York Times corpus, it outperformed two simple baselines of Bag-Of-Words and LDA as well as two ESA-based methods. Rt10 of ELSA was improved up to 46.25% over other methods, and its NDCG@k was always higher than those of the others regardless of k.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

General Terms

Algorithm, Experimentation, Performance

Keywords

Localized recommender system; Explicit localized semantic analysis; Geographical context; Local topic distribution

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

1. INTRODUCTION

Content-based news article recommendation aims to offer news articles to a reader based on his/her interests. To reflect the interests of a specific reader into article recommendation, the interests are predicted from the user profile that is a collection of articles he/she has read in the past or explicit information about the reader such as age and gender. This kind of user profile is static or at least almost static. The profile remains same regardless of the user location, even if a large number of users of handheld devices (i.e. smart phone or tablet PC) access online news providers anywhere, and their interests depend on their location.

The fact that the interests of a user are strongly related to their location, implies that his/her geographical context is important for localized news article recommendation. For instance, assume that a user is reading an article "After Delays, Wireless Web Comes to Parks" of The New York Times. This article delivers news about establishing wireless networks at the prominent parks in New York city. Therefore, it has two major topics, a new facility of public parks and a company that won a contract for the wireless network service. The people who are strolling at a public park will focus on the former, while those working at Wall Street would prefer the latter.

The key of news article recommendation is obviously a representation of the articles. Word frequencies or topics are the most widely used representation of articles [11, 18], but recently, topic representation is preferred because a topic is a good proxy for article contents [23]. For instance, Wang and Blei proposed a collaborative topic model using Latent Dirichlet Allocation (LDA) to recommend scientific articles [2]. Egozi et. al [5] projected articles onto a topic space constructed with Explicit Semantic Analysis (ESA) [7]. In ESA, each Wikipedia concept is regarded as a possible topic, and an article is represented as a topic vector with Wikipedia concepts. Because Wikipedia has a large volume of concepts, articles can be expressed efficiently and accurately. On the other hand, such topic representations of articles are geo-neutral. Therefore, a topic model that reflects the geographical context of a user is needed to localize an article recommendation.

When the location at which a user is standing is the only information available for the user profile, the geographical context of a user is equivalent to the geographical context of his/her location. The geographical context of a location can also be expressed as a topic vector, because most locations have their own geographical topics that are defined as spatially coherent meaningful themes [25]. Most topic models

proposed for extracting geographical topics from a location identify topics actually from a set of geo-tagged documents associated with the location [13]. Nevertheless, they have a problem in that the geo-tagged documents include not only the topics that are related directly to the location but also some other topics that have no relationship with the location.

This paper proposes a novel method for the topical representation of documents, Explicit Localized Semantic Analysis (ELSA). ELSA basically represents news articles as topic vectors considering the geographical context of a location. It inherits all the advantages of ESA as an extension of ESA, but overcomes the drawback in that ESA is geo-neural. That is, the topic vectors of the articles in ELSA are biased to the geographical context of the location. For this purpose, ELSA estimates the distribution of local topics for a given user location. All topics for a location are found from the geo-tagged documents of the location, but they are generally dependent on one another. For instance, the topics 'FC Barcelona' and 'Lionel Messi' are somehow dependent from each other, as they are related to *Camp Nou*, a football stadium in Barcelona. To express the topic dependency, the link information within Wikipedia concepts is adopted. All Wikipedia concepts, which are related to the location, are expressed as a graph with their link information. The hub nodes of this topic dependency graph are the actual *local* topics associated with the location.

Another advantage of ELSA is that it can be used to represent not only articles but also locations. This is because the user location is also expressed with the topics from a set of geo-tagged documents retrieved with that location. That is, the articles and locations are all represented as vectors in identical topic space. This leads to a simple comparison of articles and locations, and efficient recommendation of the articles that are relevant to the user location.

A series of experiments were carried out to verify the recommendation performance of ELSA. In the experiments, the articles were recommended for fifteen locations in three distinguished categories: Airport, Baseball stadium, and National park. The New York Times corpus was used as the recommendation candidates, and the performance of ELSA was compared with four other methods: Bag-of-Words, Latent Dirichlet Allocation (LDA) [2], Explicit Semantic Analysis (ESA) [7], and Probabilistic Explicit Semantic Analysis (PESA) [21]. According to the experimental results, ELSA outperformed the others in both Rt10 (Rating of Top 10) [15] and NDCG@k. These results highlight the effectiveness of ELSA in the localized article recommendation.

2. RELATED WORK

2.1 Localized Recommendation

The popularity of handheld devices with a GPS makes the user location available for various kinds of recommendations [12, 17, 19]. Most previous studies on localized recommendation have focused on the physical attributes of locations. Dao et al. proposed a context-aware collaborative filtering for location-based advertising [4]. In their study, the item scores for a specific user are determined by considering user's location. That is, the user-item matrix of this method is expanded with the user locations. Thus, the user similarity can be obtained using the items shared by other users at the same location.

Zheng et al. proposed the user-centered collaborative location and activity filtering (UCLAF) to recommend tourism spots and activities [27]. In UCLAF, the similar users of a specific user are obtained based on their GPS trajectories and activities at tourism spots. UCLAF recommends tourism spots for the user by analyzing similar users. CityVoyager system proposed by Takeuchi and Sugimoto [22] uses the shops previously visited by a user to recommend new shops for him/her. Ye et al. proposed Geo-Measured Friend-based Collaborative Filtering (GM-FCF) to recommend places such as stores, movie theaters, etc [24]. The recommendation by GM-FCF is made using the places visited by his/her friends in the social network. GM-FCF ranks the places by their physical distance from his/her location. Yu and Chang suggested a tour planning system [26]. This system also recommends sightseeing spots, hotels, and restaurants based on user's location.

As shown in these studies, the user's location helps recommendation systems improve their performance. This is because it reveals valuable information of the user. This was recently taken into account for news recommendation. GeoFeed is a location-aware news feed system [1]. This system provides its users with the news that are spatially related with the users. The spatial relationship between the user and a news in GeoFeed is determined using the distances from his/her to the locations tagged in the news. Mokbel et al. employed user locations to recommend *local* news [14]. In their study, they considered the distance between the user and location in which the news article was published. Although these studies proved the feasibility of their idea by implementing practical localized recommender systems, only the physical attributes of the user location was considered but its latent attributes are ignored. To the best of our knowledge, no study has used the latent attributes of the user locations in news article recommendation.

2.2 Geographical Topic Model

A location is described physically by its longitude and latitude. This is because it is difficult to transform a location into other forms but the coordinate of the longitude and latitude, even though most locations have their own geographical context and the context can be expressed with topics at these locations. For instance, the topics '*shopping*' at a department store and '*trip*' at a sightseeing spot describe their locations well. One of the efforts to extract the geographical topics of a location is the geographical topic model. Therefore, range of geographical topic models have been proposed [20, 28].

The main issue in geographical topics is how to associate topics with locations. Eisenstein et al. proposed a generative model to determine the topics of a set of geo-tagged documents and the regions corresponding to the topics simultaneously [6]. In order to determine them simultaneously, they assumed that a document is generated from two latent variables of geographical topics and regions. Yin et al. proposed Latent Geographical Topic Analysis (LGTA) [25] to solve the same problem. The contribution of LGTA is that it allows geographical topics to be associated with multiple regions. Hong et al. proposed a generative model to discover geographical topics in the tweet stream [9]. They made a similar assumption using the above-mentioned methods, but the idiosyncracy of their method comes from the fact that the method attempts to reflect the preferences of

twitter users and the dependency between regions and topics.

These geographical topic models were applied successfully to geographical topic discovery. On the other hand, it is difficult to apply them to news article recommendation. In order to recommend articles for a specific location, all the topics at the location should be identified first. However, the topic models for topic discovery try to find major topics for the location. As a result, some topics are ignored, even if they deliver some information for the location. Thus, a topic model that utilizes all topics of a given location is needed.

One candidate to represent a location with all its topics is Probabilistic Explicit Semantic Analysis (PESA) [21]. PESA is a probabilistic topic model based on Explicit Semantic Analysis (ESA) [7], and is designed to compare locations with their topics. For this, it maps locations onto a topic space spanned by Wikipedia concepts. Because all Wikipedia concepts are employed in representing locations, there are no missing topics associated with a location in PESA.

PESA aims to compare locations, and its performance in location comparison is high. However, it has a critical problem when being applied to localized article recommendation. PESA computes the topic prior distribution to reflect the dependency among topics. This prior distribution forces the locations in the same geographical category to be similar locations. It is global in that it comes from the dependency relations of all possible topics. Nevertheless, the topics relevant only to a specific location are more influential than those observed at almost all locations in the localized recommendation. For instance, assume two baseball stadiums, Rangers Ballpark and Angel Stadium. They share many topics regarding baseball games. Hence, PESA considers them to be similar locations. However, they should be treated differently in localized article recommendation because they are physically far and can have different geographically-affected topics. That is, Rangers Ballpark could have topics about Texas, whereas Angel Stadium could have topics about Los Angeles or California.

3. LOCALIZED NEWS ARTICLE RECOMMENDATION

Assume that a user is specified only by his/her location $l = \{longitude, latitude\} \in R^2$. The localized news article recommendation then ranks a set of articles, $\mathbf{A} = \{a_1, a_2, \ldots, a_n\}$ so that the order between ranks $a_1 \succ a_2 \succ \ldots \succ a_n$ satisfies the geographical preference of the user at l. The preference relation is determined by a score function $f(a, l)$ that always meets $f(a_i, l) > f(a_j, l)$ if $a_i \succ a_j$. That is, the top k ($k \leq n$) articles from $\mathbf{A}$ scored by this function are suggested as appropriate news articles according to the location l.

The score function $f(a, l)$ measures the appropriateness of each a to l. If both a and l are represented as vectors in the same topic space, it can be replaced by a similarity function [3]. That is, the score function $f(a, l)$ can be defined as

$$f(a, l) = \frac{\Phi(a) \cdot \Phi(l)}{|\Phi(a)| \cdot |\Phi(l)|}, \quad (1)$$

where Φ is a mapping function of news articles and locations into a topic space.

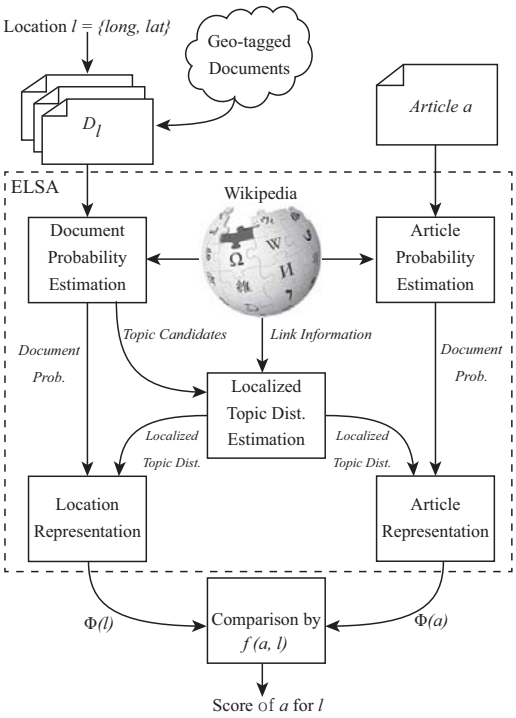


Figure 1: Overall process of the localized article recommendation.

Figure 1 depicts the overall process of measuring how appropriate a news article a is to a specific location l by Explicit Localized Semantic Analysis (ELSA). When l, the location of a user is given, D_l, a set of documents for l, is crawled from geo-tagged documents such as Twitter, Facebook, and Flickr tags. Each document d in D_l is represented as a topic vector of which dimension is the number of concepts in Wikipedia as done in ESA. Then, the location l is represented as a product of all d's in D_l. The concepts with non-zero values in this product vector are the possible topics for l, even though some non-relevant topic are also included at this step.

We assume that the topic would be dependent on one another if they are actually related to l. In order to reflect this dependency into the recommendation, a dependency graph of the topics is constructed using the link information in Wikipedia, and the local topic distribution is estimated from this graph using PageRank [16]. This local topic distribution explains which topics are actually important with respect to l when the topics are related to each other. Therefore, it is used as a weight vector for the product vector of D_l. Then, the final mapping $\Phi(l)$ of a location l into the Wikipedia topic space is the multiplication of the product vector and the local topic distribution.

News articles are also projected onto the Wikipedia topic space in the same way. A news article a is first expressed as a topic vector as ESA does. Then, the final mapping $\Phi(a)$ of a is a multiplication of the topic vector of a and the local topic distribution. Note that even the same article would be represented differently according to the location, because every different location has a different local topic distribu-

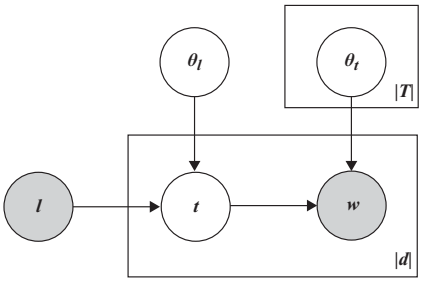


Figure 2: Graphical representation of ELSA.

tion. At last, the appropriateness of a to l is determined by $f(a, l)$.

4. EXPLICIT LOCALIZED SEMANTIC ANALYSIS

4.1 Location Representation

Explicit Semantic Analysis (ESA) [7] is a topical representation of unrestricted natural language documents. When a set of geo-tagged documents D_l for a location l is given, it maps, for the representation of l, D_l onto a topic space which is explicitly defined with Wikipedia concepts. The dimensionality of the topic space is equivalent to the number of Wikipedia concepts. Let T be a set of Wikipedia concepts. Then, D_l is expressed as a topic vector

$$D_l = \langle \phi_1(D_l), \phi_2(D_l), \ldots, \phi_{|T|}(D_l) \rangle.$$

Here, $\phi_i(D_l)$ is the weight of the i-th topic for D_l and is given as

$$\phi_i(D_l) = \sum_{d \in D_l} \sum_{w \in d} tfidf(w, d, D_l) \cdot tfidf(w, c_i, T)), \qquad (2)$$

where c_i is an article of a Wikipedia concept $t_i \in T$, and $tfidf(w, d, D_l) = tf(w, d) \cdot idf(w, D_l)$. Here, $tf(w, d)$ is the term frequency of w in d and $idf(w, D_l)$ is the inverse document frequency of w in D_l. That is, the location is represented as a topic vector through the geo-tagged documents related to the location.

Because news articles, themselves are documents, they can also be represented as topic vectors by ESA. Thus, ESA provides an identical semantic representation for both locations and articles. In addition, this is a fine-grained representation of locations and articles due to the rich world-knowledge of Wikipedia. However, some topics that are irrelevant to l can be included in D_l. Since ESA does not consider any geographical context of a location, it is affected by those irrelevant topics.

ELSA is different from ESA in that it considers the geographical context of a location in expressing the locations and articles. ELSA reinforces ESA with a local topic distribution. Figure 2 shows the graphical representation of ELSA. Here, a document $d \in D_l$ depends on topics $t \in T$, and the topics are dependent on a location l. Note that l and d are the observed variables, while t is the unobserved one. θ_l denotes the local topic distribution at l and θ_t is the word distribution of the topic t. Both are also unobserved variables.

ELSA represents a location l as

$$\Phi(l) = \left\langle p(D_l, t_1, l|\theta_l, \theta_{t_1}), \ldots, p(D_l, t_{|T|}, l|\theta_l, \theta_{t_{|T|}}) \right\rangle, \qquad (3)$$

where $p(D_l, t_i, l|\theta_l, \theta_{t_i})$ is the probability of the i-th topic. Since l is independent of θ_l and θ_t, $p(D_l, t_i, l|\theta_l, \theta_t)$ becomes $p(D_l, t_i|l, \theta_l, \theta_{t_i}) \cdot p(l)$. Note that only l is related directly to t in Figure 2. Therefore,

$$\begin{aligned} p(D_l, t_i, l|\theta_l, \theta_{t_i}) &= p(D_l, t_i|l, \theta_l, \theta_{t_i}) \cdot p(l) \\ &= p(D_l|t_i, \theta_{t_i}) \cdot p(t_i|l, \theta_l) \cdot p(l) \\ &= p(t_i|l, \theta_l) \cdot p(l) \cdot \prod_{d \in D_l} p(d|t_i, \theta_{t_i}). \end{aligned}$$

If $p(l)$ is assumed to be uniform for all locations,

$$p(D_l, t_i, l|\theta_l, \theta_{t_i}) \propto p(t_i|l, \theta_l) \cdot \prod_{d \in D_l} p(d|t_i, \theta_{t_i}).$$

Then, for the scalability,

$$\begin{aligned} p(D_l, t_i, l|\theta_l, \theta_{t_i}) &\propto \log(p(D_l, t_i, l|\theta_l, \theta_{t_i})) \\ &\propto \log p(t_i|l, \theta_l) + \sum_{d \in D_l} \log p(d|t_i, \theta_{t_i}). \end{aligned}$$

Therefore, $p(d|t_i, \theta_{t_i})$ and $p(t_i|l, \theta_l)$ should be estimated for each topic t_i.

When we assume all words are independent, $\log p(d|t_i, \theta_{t_i})$ can be expressed as

$$\log p(d|t_i, \theta_{t_i}) = \sum_{w \in d} \log p(w|t_i, \theta_{t_i}).$$

Because Wikipedia concepts are adopted as topics, each topic has its own Wikipedia article. $p(w|t_i, \theta_{t_i})$ then measures how w is possibly generated from c_i, which is the Wikipedia article for t_i. Thus, after add-one smoothing, $p(w|t_i, \theta_{t_i})$ is

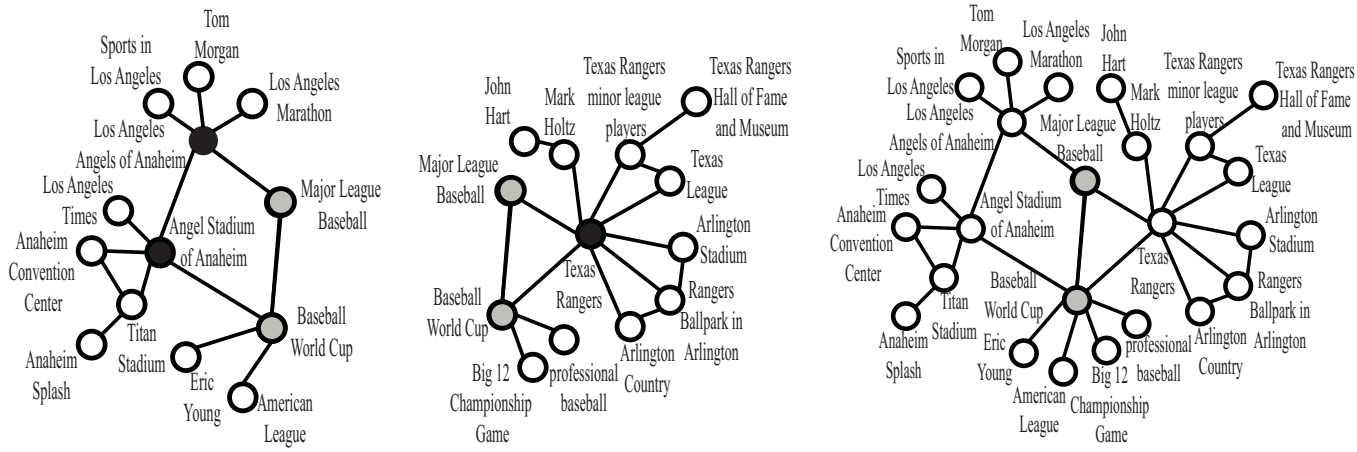
$$p(w|t_i, \theta_{t_i}) = \frac{1 + n_{c_i}(w)}{|W_{c_i}| + \sum_{w \in c_i} n_{c_i}(w)}, \qquad (4)$$

where W_{c_i} is a set of words in c_i and $n_{c_i}(w)$ is the frequency of w in c_i.

4.2 Estimating Local Topic Distribution

The local topic distribution $p(t_i|l, \theta_l)$ denotes the probability of a topic t_i at the location l. However, most topics at l are often coherent so that they are dependent on one another. This dependency among topics makes it difficult to estimate $p(t_i|l, \theta_l)$ directly. Thus, ELSA first constructs the dependency structure of local topics of l, and then approximates $p(t_i|l, \theta_l)$ with it.

As a first step to derive $p(t_i|l, \theta_l)$, all topics related to l should be identified. Because $p(D_l|t_i, \theta_{t_i})$ is the probability that topic t_i generates geo-tagged documents D_l of l, the actual topics of l are those with non-zero $p(D_l|t_i, \theta_{t_i})$. We denote the set of these actual topics as T_l. Then, the dependency structure of the topics of T_l is constructed using the link information within Wikipedia articles. Wikipedia articles often contain hyperlinks to other articles. These hyperlinks have semantics because all Wikipedia articles are published and managed carefully and manually by massive users, According to Kamps and Kollen [10], the hyperlinks in Wikipedia articles reflect semantic relationships among Wikipedia concepts. Because ELSA uses Wikipedia concepts as its geographical topics, it is natural to use these link



(a) Topic graph for *Angel Stadium* (b) Topic graph for *Rangers Ballpark* (c) Combined topic graph for both locations

Figure 3: Different major topics of ELSA and PESA at two baseball stadiums.

information among topics in T_l for the dependency structure of the topics. ELSA represents the dependency structure as a directed graph. For instance, when *Central park* has a hyperlink to *National historic landmark*, the corresponding dependency exists from *Central park* to *National historic landmark*.

When the dependency structure of T_l is given, ELSA approximates $p(t_i|l,\theta_l)$ from it according to the PageRank algorithm [16], because the values obtained from PageRank can be regarded as probabilities of the graph nodes in a complicated link structure [8]. That is, $p(t_i|l,\theta_l)$ is

$$\begin{aligned} p(t_i|l,\theta_l) &\approx PR(t_i) \\ &= \frac{1-\alpha}{|T_l|} + \alpha \cdot \sum_{t' \in in(t_i,T_l)} \frac{PR(t')}{L(t')}, \end{aligned} \quad (5)$$

where α is a damping factor, $in(t_i,T_l)$ is the set of topics in T_l with a link to t_i, and $L(t')$ is the number of outbound links from t'.

4.3 Article Representation

ELSA projects a location onto the topic space of Wikipedia concepts through a set of geo-tagged documents. Therefore, when a set of news articles, $\mathbf{A} = \{a_1, a_2, \ldots, a_n\}$ is given, all articles of $\mathbf{A}$ can be projected onto the topic space because news articles are also documents. Thus, the projection of an article $a \in \mathbf{A}$ can be written as

$$\Phi(a) = \left\langle p(a,t_1,l|\theta_l,\theta_{t_1}), \ldots, p(a,t_{|T|},l|\theta_l,\theta_{t_{|T|}}) \right\rangle.$$

Since we assume that the location probability $p(l)$ is uniform,

$$\begin{aligned} p(a,t_i,l|\theta_l,\theta_{t_i}) &= p(a,t_i|l,\theta_l,\theta_{t_i}) \cdot p(l) \\ &\propto p(a,t_i|l,\theta_l,\theta_{t_i}). \end{aligned}$$

According to Figure 2, $p(a,t_i|l,\theta_l,\theta_{t_i}) = p(a|t_i,\theta_{t_i}) \cdot p(t_i|l,\theta_l)$. Thus,

$$\begin{aligned} p(a,t_i,l|\theta_l,\theta_{t_i}) &\propto p(a|t_i,\theta_{t_i}) \cdot p(t_i|l,\theta_l) \\ &= p(t_i|l,\theta_l) \cdot \prod_{w \in a} p(w|t_i,\theta_{t_i}) \\ &\propto \log p(t_i|l,\theta_l) + \sum_{w \in a} \log p(w|t_i,\theta_{t_i}) \\ &= \log p(t_i|l,\theta_l) \cdot \sum_{w \in a} \log \left(\frac{1 + n_{c_i}(w)}{|W_{c_i}| + \sum_{w \in c_i} n_{c_i}(w)} \right), \end{aligned}$$

where $n_a(w)$ denotes the frequency of a word w in a. The term $p(t_i|l,\theta_l)$ is the local topic distribution given in Equation (5). Because this term depends on l, it becomes different for every different l. As a result, even the same news article is represented differently if the location changes.

4.4 Comparison with PESA

The probabilistic ESA (PESA) is another topical representation of documents based on ESA [21], and is designed to compare locations through geo-tagged documents. It assigns high similarity to the pairs of locations in the same geographical category, while assigning very low similarities to the pairs of locations from different categories. For this, it regards, as geographical topics, the Wikipedia concepts of which a Wikipedia article has geo-tags or hyperlinks to other geo-tagged Wikipedia articles. Then, PESA uses these geographical topics to estimate a topic distribution.

PESA also estimates a topic distribution to reflect dependencies among topics. Because it considers the dependency of almost all topics, the global topics shared by most locations in a geographical category have high probabilities in the topic distribution of PESA. These global topics are important in PESA, since PESA focuses on location comparison in which similar locations share geographical topics. On the other hand, the task of a localized news article recommendation aims to offer appropriate articles for each specific location. That is, the local characteristics are more important in this task than the global characteristics. Many

Table 1: Locations used in the experiments and their categories

Geographical category	Locations
Airport	George Bush Airport
	John F. Kennedy Airport
	La Guardia Airport
	Phoenix Sky Harbor Airport
	Washington Dulles Airport
Baseball stadium	Angels Stadium of Anaheim
	Busch stadium
	Citizens Bank Park
	Rangers Ballpark
	Turner Field
National park	Arches
	Grand Canyon
	Hawaii Volcanoes
	Rocky Mountain
	Yellowstone

topics that are important globally are often not important locally. Therefore, PESA is unsuitable for this task.

ELSA estimates the distribution of only local topics. In the topic distribution of ELSA, the topics directly associated with a given location have high probabilities. As a result, even the locations from the same category have different topic distributions. This is reasonable because the locations could have different geographical context even if they belong to the same category.

The advantage of ELSA can be addressed by a simple example. Figure 3 shows the topic structures of two locations: *Angel Stadium* at Anaheim, CA (Figure 3(a)) and *Rangers Ballpark* at Arlington, TX (Figure 3(b)). In these graphs, the nodes are topics and the edges are their dependencies. The top thirteen topics were chosen from each location according to the probabilities over its D_l.

ELSA assigns high probabilities to the black nodes in each graph, because the nodes are local hubs. When the location is *Angel Stadium*, "Los Angeles, Angels of Anaheim" and "Angel Stadium of Anaheim" have high probabilities. On the other hand, the highest probability is assigned to the topic of "Texas Rangers" in the topic distribution of *Rangers Ballpark*. As a result, even the same article would be represented differently due to these different local topics.

PESA uses the dependency structure of all topics. Thus, it determines the topic distribution from the combined graph in Figure 3(c). After the graphs of Figure 3(a) and 3(b) are combined, the topics such as "Major League Baseball" and "Baseball World Cup" (gray nodes in this graph) are considered as hubs. These topics shared by two locations force the topic vectors of the locations to be similar. Therefore, PESA is effective in location comparison. However, the emphasis on global topics is adverse to localized news article recommendation in which local topics are preferable.

5. EXPERIMENTS

5.1 Experimental Settings

For the evaluation of ELSA, fifteen locations in USA are selected manually. Table 1 lists the locations and their categories. All locations belong to one of three geographical categories: 'Airport', 'Baseball stadium', and 'National park'. The keywords of the Flickr images are used as a set of geo-tagged documents, D_l for location l. These are adopted as geo-tagged documents because the keywords are often a description of the location. For each location, l in Table 1, the images that are closely associated with l are crawled through Flickr API (`http://www.flickr.com/services/api`). The GPS information attached to the images is used to determine their association with l. We determined that the image would be associated with l if the distance from l to the image location is within a predefined threshold. The threshold used was 0.2 km for 'Baseball stadium', 1 km for 'Airport', and 10 km for 'National park'. The keywords tagged to an image are regarded as a geo-tagged document.

Table 2: Statistics of the Flickr data

Category	Avg. documents	Avg. keywords	Unique keywords
Airport	177.8	7.49	1,895
Baseball stadium	179.0	8.66	900
National park	201.0	6.74	1,010

Table 2 describes the Flickr data set used in the experiments. The locations in the category of 'Airport' have 177.8 documents on average, whereas those in 'Baseball stadium' and 'National park' have 179 and 201 documents, respectively. The average number of keywords in a document is 7.49 in 'Airport', 8.66 in 'Baseball stadium', and 6.74 in 'National park'. The last attribute is the number of unique keywords. 'Airport' has 1,895 unique keywords, and 'Baseball stadium' and 'National park' have 900 and 1,010 unique keywords respectively.

A Wikipedia snapshot as of May 2, 2012 was used to generate geographical topics. A preprocessing described in [21] was adapted to the snapshot to exclude irrelevant topics, such as digit, time, and so on. After preprocessing, 1,116,275 concepts remain, and they are used for a topic set T in all the experiments below. The number of unique terms in this Wikipedia snapshot was 1,498,045.

The New York Times (NYT) corpus was adopted as a pool of news articles. This corpus contained 1,841,402 articles published by the New York Times from 1987 to 2007. Among them, the articles from January 1, 2006 to June 19, 2007 were actually used as a set of news articles, $\mathbf{A}$. This set contained 127,005 articles composed of 356,595 distinct words.

ELSA was compared using four baseline methods: BOW, LDA, ESA, and PESA. BOW is the bag-of-words model that represents locations and articles with the frequency of words. Thus, the need for a topic model in a localized article recommendation can be shown by a comparison with BOW. Three topic-based models, LDA, ESA, and PESA were prepared to evaluate the effectiveness of ELSA. LDA is the Latent Dirichlet Allocation that is used widely in the various applications to extract topics from documents. ESA and PESA both use Wikipedia concepts as their topics, but PESA additionally models the prior distribution of geographical topics.

The performance of these methods were compared with two metrics: (i) Rt10 (Rating of Top 10), the average score of recommended articles, and (ii) Normalized Discounted Cumulative Gain at top k (NDCG@k) which is widely used

Table 3: Kappa values of the articles scores by the human evaluators and actual number of news articles per location.

Category	Kappa value	Avg. distinct articles per location
Airport	0.510	32.8 ± 2.2
Baseball stadium	0.677	28.0 ± 6.5
National park	0.680	31.0 ± 3.7
Overall	0.623	30.6 ± 4.6

Table 4: Rt10 of the experimental methods

Category	BOW	LDA	ESA	PESA	ELSA
Airport	1.490	1.620	1.710	1.700	**2.119**
Baseball stadium	2.110	1.940	1.290	1.800	**2.400**
National park	1.470	0.700	**1.620**	1.490	1.560
Overall	1.690	1.420	1.540	1.663	**2.026**

for measuring the rank accuracy. Two human evaluators are employed to label the scores of news articles manually. Since there are too many articles in the data set, the evaluation was done only for the recommended articles. That is, for a given location, all five methods including ELSA select their own top ten articles. Accordingly, the number of news articles per location is at most 50. Since two or more methods can recommend the same articles, the evaluators score normally less than 50 articles per each location. The guidelines of scoring news articles are as follows.

- 0 point: The article has nothing to with the geographical category.
- 1 point: The article is weakly related to the geographical category.
- 2 point: The article belongs to the geographical category, but is not related to the location.
- 3 point: The article is strongly related to the location.

Table 3 lists the Kappa values of the scores by the evaluators. As shown in this table, Kappa values for 'Baseball stadium' and 'National park' were approximately 0.68, and approximately 30 distinct articles are recommended for each location. These values corresponded to *substantial agreement*. For 'Airport', the Kappa value was 0.51, which falls in *moderate agreement*, and the average distinct articles per location was 32.8.

5.2 Experimental Results

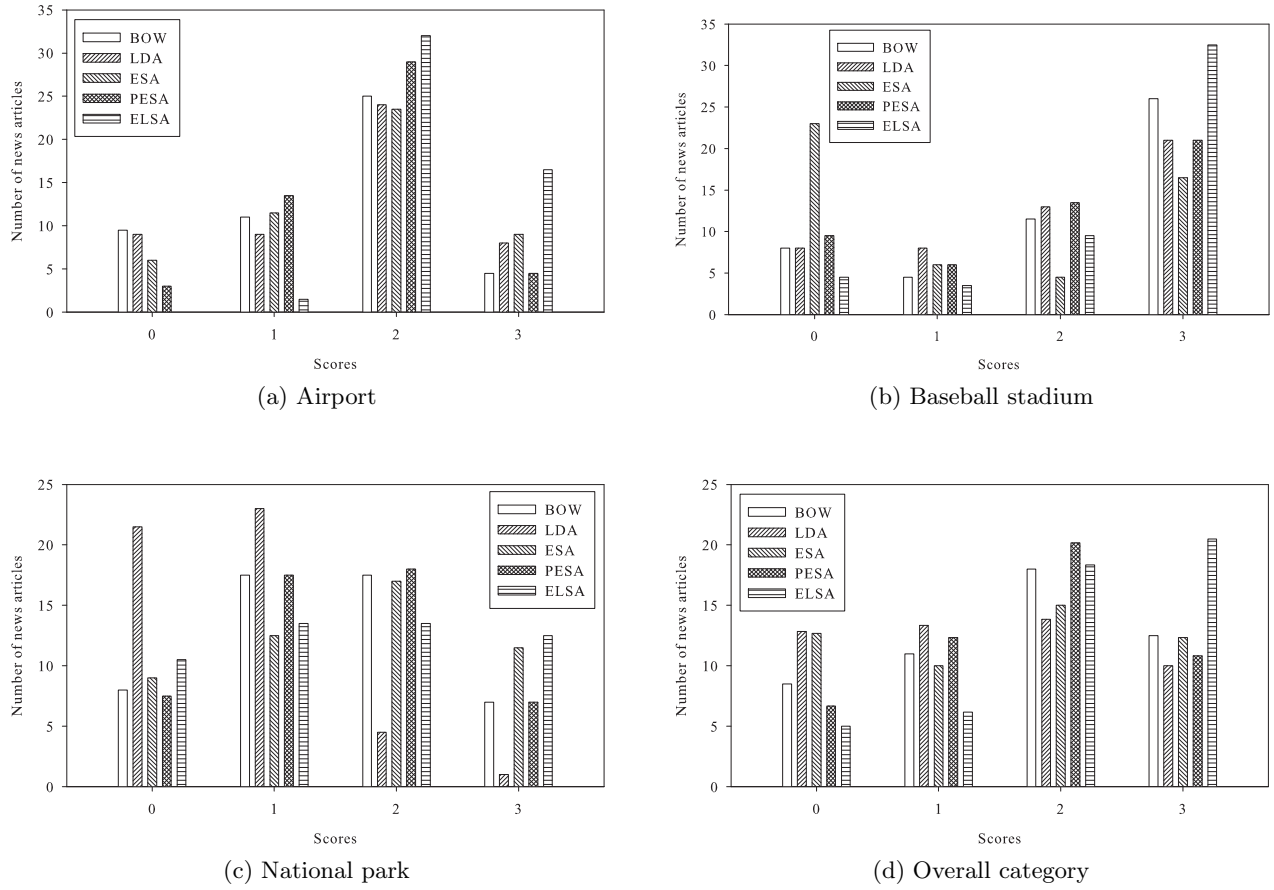
Table 4 shows the Rt10 values of the five methods. As shown in this table, LDA achieved 1.420, which was the worst performance. This Rt10 was obtained when LDA shows the best performance with 500 topics. The fact that LDA was worst among the five methods implies that it fails to capture diverse topics of news articles and locations. On the other hand, all methods based on Wikipedia concepts outperform LDA. This result highlights the effectiveness of the Wikipedia concepts as geographical topics.

BOW showed the second best performance with 1.69 of Rt10. This is an unexpected result, because BOW does not adopt topics to represent articles and locations. The good performance of BOW originates mostly from the locations of 'Baseball stadium'. Most geo-tagged documents for baseball stadiums contain the name of home teams, which allows a simple lexical matching to work well. As a result, the Rt10 score of BOW for 'Baseball stadium' was very high. However, its performance in other categories was lower than ESA-based topic models. Therefore, from the results of LDA and BOW the topical representation leads to high performance in localized article recommendation, if the topic space is rich enough to cover the entire topics.

ELSA was found to be the best method in this experiment. ELSA achieved significantly higher performance than the other methods. Its Rt10 in 'Airport' and 'Baseball stadium' was much higher than those of others. In 'National park', ESA showed the best Rt10 of 1.620, but the difference between ELSA and ESA was only 0.06. As a result, ELSA outperformed all other methods with an overall Rt10 of 2.026, and was the only method of which the overall Rt10 was more than 2.0. The outperformance of ELSA over other methods was statistically significant with a confidence interval of 95%. The p-values of the T-test against the null hypothesis that ELSA does not outperform other methods were all below 0.05. The essential factor of ELSA that distinguishes it from ESA and PESA is the use of the *local* topic distribution as the geographical context of a location, Therefore, the best performance of ELSA proves the importance of a geographical context of locations in location and article representation.

The overall quality of the recommended articles can be verified by Rt10. However, the quality of a method can be poor, even if the method achieves a high Rt10. This occurs when a large number of recommended news articles receive 1 or 2 point by human evaluators. Therefore, it is important to determine how many articles receive 2 or 3 point. Figure 4 shows the number of recommended articles for each score. As shown in this figure, the quality of ELSA is best in that most of the articles recommended by it received 2 or 3 point. In particular, an examination of the overall category (Figure 4-(d)) showed that, the number of articles with 3 point recommended by ELSA exceeded those by other methods. Note that most articles recommended by PESA received 2 point. This is because PESA uses the distribution of all topics, which makes it difficult for the locations within the same category to be distinguished. On the other hand, even though ELSA shares topics with PESA, it effectively characterizes the geographical context of each location by its local topic distribution. As a result, ELSA recommends the articles that match the given location better than the other methods.

The Rt10 scores and number of articles with high scores have a tendency to decrease, as the location boundary becomes larger. The Rt10 score for 'National park' that has a boundary of 10 km was the worst, whereas that for 'Baseball stadium' with a boundary of 0.2 km was the best. This is because a location has a range of geographical topics as its boundary becomes large, and this variety of topics decreases the locality effect of the topics. However, ELSA outperformed the other methods regardless of how large the location boundary was.



(a) Airport

(b) Baseball stadium

(c) National park

(d) Overall category

Figure 4: Number of news articles for each score.

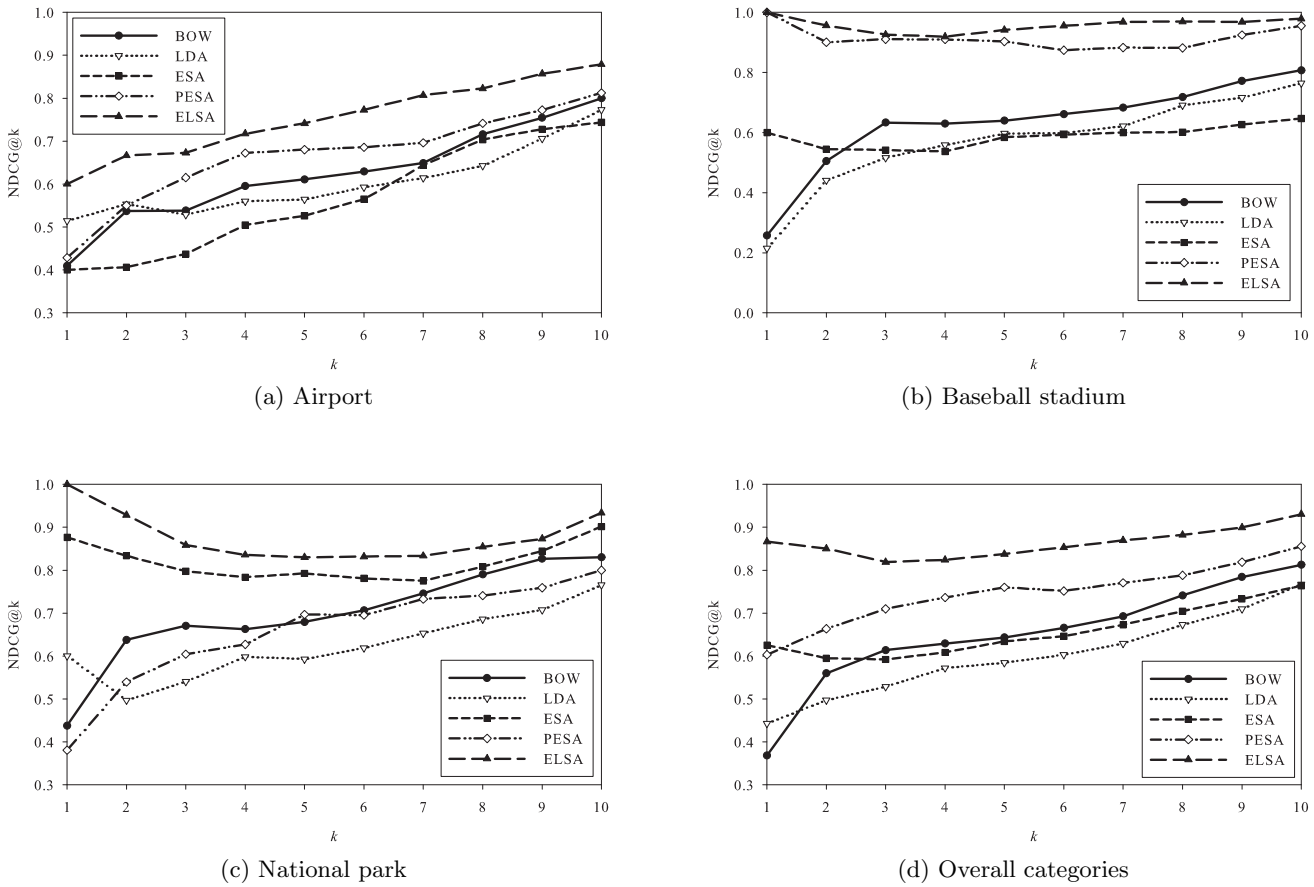
The ranking performance of the methods was also measured with NDCG@k. Figure 5 shows the NDCG@k values with k from 1 to 10. Even with this metric, ELSA outperformed the others for all geographical categories. Although PESA achieved a similar performance in 'Baseball stadium', its performance was lower than ESA and BOW in 'National park'. BOW showed the lowest NDCG@k when k was less than 3. Because it considers only the shared words by the news articles and geo-tagged documents, it recommends many irrelevant articles with some general words. For instance, many articles with a word, *angel* are irrelevant to Angel Stadium, and those with *bush* are irrelevant to George Bush Airport.

ELSA outperformed the others consistently for all k's. All improvements in NDCG@k of ELSA over the other methods are also statistically significant, because the p-values against the null hypothesis were less than 0.05. In particular, when $k \leq 3$, a large gap between ELSA and other methods was observed. As k was increased up to 10, this gap became smaller but was still higher than others. This result suggests that ELSA recommends news articles successfully that are appropriate to the geographical context.

PESA was the second best and also reflected the distribution of geographical topics in representing articles with topics. Thus, the superiority of ELSA and PESA over other methods proves the importance of geographical topics in the localized news article recommendation. Interestingly, ELSA and ESA showed a similar tendency in NDCG@k for all categories as k increased even though NDCG@k of ELSA was always higher than that of ESA. This fact implies that ELSA improves ESA successfully and substantially as its expansion for localized news article recommendation by adopting the *local* topic distribution.

6. CONCLUSION

This paper proposed a localized news article recommendation with the Explicit Localized Semantic Analysis (ELSA). The location of a user provides valuable information as user context, because each location has its own geographical topics, which are often strongly related to the user context at that location. Therefore, the proposed method recommends the news articles that are appropriate to the location by reflecting the geographical context of user. For this, ELSA represents news articles as the topic vectors of Wikipedia concepts. The weight of each topic for an article is adjusted to reflect the geographical context of the user. The user location is also expressed as a topic vector through the geo-tagged documents of the user location. Because the concerns of the user are affected by his/her location, even the same article should be represented differently according to



(a) Airport

(b) Baseball stadium

(c) National park

(d) Overall categories

Figure 5: NDCG@k scores with various k.

the user's location. To reflect it into article recommendation, ELSA models the dependencies of local topics. Then, the local topic distribution estimated from the topic dependencies is applied to represent both locations and news articles.

ELSA was evaluated with the New York Times corpus for fifteen famous locations. It was compared with four other methods: BOW, LDA, ESA, and PESA. All the methods were evaluated with two kinds of metrics: Rt10 and NDCG@k. According to the experimental results, ELSA outperformed all other methods substantially. The superiority of ELSA over BOW and LDA implies the effectiveness of the topic space by the Wikipedia concepts. PESA, another ESA-based topic model achieved higher performance than other methods including ESA, itself. Although both ELSA and PESA adopt the topic space by Wikipedia concepts, they are different in that ELSA uses a local topic distribution whereas PESA uses a global topic distribution. The performance of ELSA was superior to that of PESA in both metrics, because the local topic distribution is more appropriate than the global topic distribution in localized article recommendation.

The contributions of this paper are two-fold. First, this is the first effort to recommend articles according to user locations. Although every location has its own geographical topics, most legacy localized recommendation systems focused on their physical attributes. In this paper, by considering topics of a location as its geographical context, the articles that match the geographical context of a location could be recommended. Second, we manually constructed a data set for localized article recommendation. 459 news articles out of approximately 120,000 articles of the New York Times were tagged with scores from 0 to 3 with respect to their relevance to fifteen locations. Because this is the first data set for localized article recommendation, it can be used for further developments.

Acknowledgement

This work was supported by the Industrial Strategic Technology Development Program (10035348, Development of a Cognitive Planning and Learning Model for Mobile Platforms) funded by the Ministry of Knowledge Economy (MKE, Korea).

7. REFERENCES

[1] J. Bao, M. Mokbel, and C. Chow. GeoFeed: A location-aware news feed system. In *Proceedings of the 28th IEEE International Conference on Data Engineering*, pages 54–65, 2012.

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] T. Bogers and A. Bosch. Comparing and evaluating information retrieval algorithms for news recommendation. In *Proceedings of ACM conference on Recommender systems*, pages 141–144, 2007.

[4] T. Dao, S. Jeong, and H. Ahn. A novel recommendation model of location-based advertising: Context-aware collaborative filtering using ga approach. *Expert Systems with Applications*, 39(3):3731–3739, 2012.

[5] O. Egozi, S. Markovitch, and E. Gabrilovich. Concept-based information retrieval using Explicit Semantic Analysis. *ACM Transactions on Information Systems*, 29(2):8:1–8:34, 2011.

[6] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.

[7] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007.

[8] T. Griffiths, M. Steyvers, and A. Firl. Google and the mind: Predicting fluency with PageRank. *Psychological Science*, 18(12):1069–1076, 2007.

[9] L. Hong, A. Ahmed, S. Gurumurth, A. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*, pages 769–778, 2012.

[10] J. Kamps and M. Koolen. The importance of link evidence in Wikipedia. In *Proceedings of the 30th European Conference on IR Research*, pages 270–282, 2008.

[11] R. Krestel, P. Fankhauser, and W. Nejdl. Latent Dirichlet Allocation for tag recommendation. In *Proceedings of the 6th ACM Conference on Recommender Systems*, pages 61–68, 2009.

[12] Y. Li, A. Guo, S. Liu, Y. Gao, and Y. Zheng. A location based reminder system for advertisement. In *Proceedings of the 18th International Conference on Multimedia*, pages 1501–1502, 2010.

[13] Q. Mei, C. Liu, and H. Su. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th International Conference on World Wide Web*, pages 533–542, 2006.

[14] M. Mokbel, J. Bao, A. Eldawy, J. Levandoski, and M. Sarwat. Personalization, socialization, and recommendations in location-based services 2.0. In *Proceedings of the International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases 2011, collocated with VLDB*, pages 1–6, 2011.

[15] R. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *Proceedings of the SIGIR-99 Workshop on Recommender Systems: Algorithms and Evaluation*, 1999.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd. *The pagerank citation ranking: Bringing order to the web.* Technical report, Computer Systems Laboratory, Stanford University, 1998.

[17] M. Park, J. Hong, and S. Cho. Location-based recommendation system using bayesian user's preference model in mobile devices. In *Proceedings of the 4th International Conference on Ubiquitous Intelligence and Computing*, pages 1130–1139, 2007.

[18] D. Ramage, D. Hall, R. Nallapati, and C. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[19] N. Savage, M. Baranski, N. Chavez, and T. Höllerer. I'm feeling loco: A location based context aware recommendation system. In *Advances in Location-Based Services*, pages 37–54, 2012.

[20] S. Sizov. GeoFolk: Latent spatial semantics in web 2.0 social media. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 281–290, 2010.

[21] J. Son, Y. Noh, H. Song, S. Park, and S. Lee. Location comparison through geographical topics. In *Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 311–318, 2012.

[22] Y. Takeuchi and M. Sugimoto. CityVoyager: An outdoor recommendation system based on user location history. In *Proceedings of the 3rd International Conference on Ubiquitous Intelligence and Computing*, pages 625–636, 2006.

[23] X. Wei and B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185, 2006.

[24] M. Ye, P. Yin, and W. Lee. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 458–461, 2010.

[25] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web*, pages 247–256, 2011.

[26] C. Yu and H. Chang. Personalized location-based recommendation services for tour planning in mobile tourism applications. In *Proceedings of the 10th International Conference on E-Commerce and Web Technologies*, pages 38–49, 2009.

[27] V. Zheng, B. Cao, Y. Zheng, X. Xie, and Q. Yang. Collaborative filtering meets mobile recommendation: A user-centered approach. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 236–241, 2010.

[28] Y. Zhou and J. Luo. Geo-location inference on news articles via multimodal pLSA. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 741–744, 2012.