Stories around You: Location-based Serendipitous Recommendation of News Articles

Yonata Andrelo Asikin[1] and Wolfgang Wörndl[2]

[1] BMW Group, Munich, Germany
yonata-andrelo.asikin@bmw.de
[2] Department of Informatics, Technische Universität München, Germany
woerndl@in.tum.de

Abstract. Existing studies in serendipitous recommendation mostly focus on extending the metrics of desired goals such as accuracy, novelty and serendipity with respect to the user preferences. This work aims at serendipity by exploiting the prevailing location (spatial) contexts of the recommendation. For this purpose, we propose a novel spatial context model and a number of recommendation techniques based on the model. A user study on a real news dataset shows that our approach outperforms the baseline distance-based approach and thereby improves the overall user satisfaction with the recommendation result in the absence of the user's personal information.

Key words: serendipity, location-based recommender systems

1 Introduction

Serendipity means a pleasant surprise or happy accident of discovering something good or useful while not specifically searching for it. In the research field of recommendation system, serendipity is regarded as an important objective for ensuring user satisfaction with the recommendation quality [10]. Existing approaches to recommending serendipitous contents mostly focus on extending item evaluation metrics beyond accuracy and analysing existing structure of user variables such as preferences or relations to items and other users. However, this information is not always available such as in a new system or for new user (called cold start problem) or due to the privacy concerns and the willingness of the user to provide information.

In fact, serendipity can potentially happen to a lot of people due to a certain circumstance. For instance, let *Alice* be a person who does not like country music. While walking in a village near a line of mountains with a beautiful country-side scenery, she listens to radio from her mobile device. Suddenly, the radio plays a country song and she gets really interested in the song. This can be regarded as a serendipitous experience regarding her music taste. Starting from this motivation and in order to address the above mentioned problem, we propose approaches that exploit the current context variables of the recommendation which are less sensitive regarding privacy compared to the user's personal information.

Specifically, this work focuses on location or spatial variables as context. Existing works in location-based recommendation mostly emphasize the distance between the current user location with the items' coordinates as well as the user preferences. Furthermore, the works do not consider different possible associations between an item and the tagged locations that can potentially affect or enrich the recommendation result. For instance, a news item can be associated with a city because it tells a story about a person who was born there. Therefore, we model the spatial information beyond the geographic coordinates and study the associations of the location with the news articles as a part of the prior processes. Using this spatial model and considering the prior processes enable us to build various approaches to finding serendipitous items despite the absence of user preferences. To the best of our knowledge, no previous work has studied context-based serendipitous recommendation (and in particular, location-based). In brief, the contributions of this work can be listed as follows: (1) This study presents a comprehensive spatial model for recommending news articles that goes beyond the standard geographical information; (2) We introduced location-based recommendation approaches aiming at serendipity by exploiting the spatial context; (3) We conducted a user study on a real news dataset for evaluating the approaches, in which our approach outperformed the baseline algorithm in terms of surprising and serendipity of the results.

In the remainder of this paper, Section 2 presents related studies in location-aware and serendipitous recommendation. Section 3 briefly describes our spatial model as basis for the recommendation approaches in Section 4. Section 5 discusses the evaluation of the approaches that is concluded in Section 6.

2 Related Work

This work closely relates to the research on recommendation approaches focusing on serendipity and location-aware venues and news recommendation.

Serendipitous Recommendation: The traditional collaborative filtering algorithm (like-minded-people concept) can be extended by modifying the recommendation objective or similarity metrics to introduce serendipity into the recommendation result [7]. Often with this approach, accuracy is sacrificed (significantly) for the sake of other metrics. The study conducted in [10] focuses on balancing the accuracy with other factors (novelty, diversity, and serendipity) simultaneously. Social-related variables of a user can be employed to discover surprising and useful items for the user, e.g. the interaction history [3] or social relationships and trust [5]. Other researchers also modelled and analysed the user-item relations: graph-based [9] and semantic-based [1]. None of these approaches could work without sufficient user information. Our approaches, in contrast, count on contexts to deliver serendipitous items generally for all users.

Location-aware Recommendation: Location-aware recommender systems (LARS) can be classified based on a taxonomy introduced in [4]: (non-)spatial ratings for (non-)spatial items. Following this taxonomy, a location-based news recommendation uses the schema of spatial ratings for non-spatial items or for spatial items if the news is geo-tagged. This work and other studies

generally assume that the items are already tagged with geographical coordinates, and emphasize the distance between the current user location with the items' coordinates as well as the user preferences. This is shown for both venue recommendation [6] and location-aware news recommendation [2][8].

3 Spatial Model for News Recommendation

Our *spatial model* represents the broad scope of spatial information of a location in three classes: *geographical information*, *physical character*, and *place identity*. The geographical information includes the geographic coordinate (latitude and longitude) as well as the location names. The physical character of a location or *landform* generally defines the character of scenery seen by human nature. Finally, the place identity concerns the meaning and significance of places for their inhabitants and users. A news article may contain geographical information, e.g. *location name* where the news was released and *geographic coordinates* (through *geotagging* which recognizes and resolves references to geographic locations in text documents). In our approach, physical character and place identity features will be mined from a news article. We call this feature extraction process *location inference* and the further associating process *location association*.

Let $\mathcal{C} = \{c^{(1)}, ..., c^{(m_c)}\}$ as the set of m_c global available news articles, where $c^{(i)} = (u, \mathcal{D})$ is a tuple containing creator $c^{(i)}.u$ and text features vector $c^{(i)}.\mathcal{D}$. All physical locations on the earth can be represented as a set of all point locations denoted by $\mathcal{L}_G \subset \mathbb{R}^2$, where a point location $l \in \mathcal{L}_G$ is a tuple of latitude and longitude. Alternatively, $\mathcal{L}_N = \{L^{(1)}, ..., L^{(m_l)}\}$ denote the set of m_l physical places where $L^{(j)} \subseteq \mathcal{L}_G$ (allowing a place to be either a point or a region) and consequently $\mathcal{L}_N \subseteq \mathcal{P}(\mathcal{L}_G)$. Since a location can physically belong to another location (e.g. a city belongs to a country), we define a containment relation $\text{cont}_D : \mathcal{L}_N \times D \rightarrow \{0, 1\}$ where $D \in \{\mathcal{L}_G, \mathcal{L}_N\}$. Based on this representation, the different spatial information classes can be developed by introducing a set of n_l global location features $\mathcal{F}_L = \{f^{(1)}, ..., f^{(n_l)}\}$. A location feature $f^{(k)}$ can be a place name (LN) (e.g. *Munich*, *Eiffel Tower*) or a low level feature that solely or together with other features defines the physical character (LPC) (e.g. *mountain*, *beach*), or the place identity (LPI) (e.g. *industrial*, *cultural*). Let $\mathcal{F}_{LN}, \mathcal{F}_{LPC}, \mathcal{F}_{LPI} \subset \mathcal{F}_L$ be the sets of features for the particular representation of LN, LPC, and LPI, respectively. The location features are gained through the geographical mapping functions $\psi_D : \mathcal{L}_N \rightarrow \mathcal{P}(D)$, where $D \in \{\mathcal{F}_{LN}, \mathcal{F}_{LPC}, \mathcal{F}_{LPI}\}$.

Through *location inference*, spatial information is extracted from a news article. During *geotagging*, words or phrases that can be place names (called *toponym*) are firstly found in the article (this searching step is called *toponym recognition*). Afterwards, each toponym will be assigned to the right geographic coordinate (called *toponym resolution*). Formally, location inference is used to extract a set of features in $\mathcal{F}_L$ from $\mathcal{C}$. For LPC and LPI, the inference functions are denoted as $\text{inf}_{LPC} : \mathcal{C} \rightarrow \mathcal{P}(\mathcal{F}_{LPC})$ and $\text{inf}_{LPI} : \mathcal{C} \rightarrow \mathcal{P}(\mathcal{F}_{LPI})$, respectively. Since each feature $f^{(k)} \in \mathcal{F}_{LN}$ (a toponym) still has to be disambiguated to an exact $L \in \mathcal{L}_N$, the location inference is defined differently for

LN. The location inference function for LN is defined as the composition of the toponym recognition and toponym resolution functions: $\inf_{LN} = \inf_{rec} \circ \inf_{res}$ where $\inf_{rec} : \mathcal{C} \rightarrow \mathcal{P}(\mathcal{F}_{LN})$ and $\inf_{res} : \mathcal{P}(\mathcal{F}_{LN}) \rightarrow \mathcal{P}(\mathcal{L}_N)$.

People can draw a myriad of associations between news and locations. For instance, a news article can tell the history of a place and therefore, an association called *telling history* is built between the article and the place. The news articles combined with the respectively inferred locations form a set of *localized* recommendable items $\mathcal{X} = \{X^{(1)}, ..., X^{(m_x)}\}$ where $m_x \leq m_c$ is the total number of items. The tuple in $c^{(i)}$ is extended for $X^{(i)}$ resulting in $X^{(i)} = (u, \mathcal{D}, F_L, L_N)$ where $F_L \subset \mathcal{F}_L$ and $L_N \subset \mathcal{L}_N$ are the inferred location features and geographic coordinates, respectively. The associations between an item and the inferred locations can be built by means of a function $\text{association}_I : \mathcal{X} \times \mathcal{P}(\mathcal{L}_N) \rightarrow \mathcal{P}(\mathcal{A}_i)$ where $\mathcal{A}_I$ is the global set of possible associations between X and L.

4 Algorithms for Serendipitous Recommendation

For the sake of completeness, we defined a set of m_u users (either the consumer or creator of an item) as $\mathcal{U} = \{U^{(1)}, ..., U^{(m_u)}\}$. The items with inferred and associated locations together with user and location information provide building blocks for the context-aware news recommendation schema: $R : \mathcal{U} \times \mathcal{X} \times \mathcal{L}_N \rightarrow \mathbb{R}$. Given a current location L of a user u, a recommender approach suggests an item X based on L by exploiting the spatial information contained in both X and L. A baseline approach can simply be based on the distance between both of them, e.g. news near you (analogously to places near you). This method, called **Nearest Distance (ND)**, suggests a single item $X^{(i)}$ that contains $L^{(j)} \in X^{(i)}.L_N$ with smallest distance to L. To show how different utilizations of spatial model can affect the recommendation quality and in particular achieve serendipity, we propose a number of approaches below.

Geographical Hierarchy (GH) uses geographical hierarchy information of a location L and considers its parent-locations. Formally, GH looks for items with an inferred location $L^{(i)}$ where $\text{cont}_{L_N}(L^{(i)}, L) = 1$ and picks one of them randomly. Low serendipity is expected to be seen from the recommended items, since the news articles picked by this approach can be very general and well-known in a larger area of the location.

Event Association (EA) suggests the next located item from L with the association *describing event at location* with L. In this study, we define a set of associations $\mathcal{A}_I = \{\text{describing location, describing event at location}\}$. The associations are defined in this work simply by classifying based on the existence of certain keywords. Formally, we assume that if an item $X^{(i)}$ with inferred location L belongs to the class *describing location*, then the $\text{association}_i(X^{(i)}, L) = \{\text{describing location}\}$. By picking a news with a less-typical association, this approach may retrieve a more serendipitous item.

Place Identity (PI) and Combination (ND+PI): this method suggests an item with a topic that is not usual at that particular location (based on the place identity). Given current location L, the place identity is defined as $\psi_{\mathcal{F}_{LPI}}(L)$. Here, the place identity is defined as a set of topics that are often discussed at

L. Therefore, the approach will retrieve items whose topics have low similarity to the place identity, i.e. news that are not usual at L. By introducing this diversity, the serendipity is expected to be induced by this approach. Since there can also be multiple retrieved items, we can pick one item randomly (PI) or pick the nearest one (ND+PI).

5 Evaluation: Stories around You

To show how the approaches recommend serendipitous items, an online user study based on real crowd-sourced news dataset was performed. The dataset originates from an online crowd-sourced idea finding portal Jaring-Ide[1]). Specifically, it consists in a set of text articles which are ideas generated for an idea contest called *My Indonesian Moment* which is a contest about a (tourism) moment that someone experienced in a location in Indonesia. After filtering out inappropriate ideas (e.g. no text content), we get $m_c = 1869$ from 1914 ideas.

The dataset is not tagged with any spatial information and therefore, location inference (and association) are necessary. However due to the nature of the data (mixed languages, informal writing, etc.), automatic toponym recognition technique did not perform well. Therefore, we compiled a set of sub-strings of the texts that represent the correct location context of the articles. This resulted in 5293 toponyms that still have to be resolved. For the toponym resolution on $c^{(i)}$, we use gazetteer from GeoNames[2]. The inference $\inf_{LN}$ resolved the total of 4297 toponyms that corresponds to 1818 resolved items (97.27% of all available items). This forms a set of recommendable items $\mathcal{X}$ with $m_x = 1818$. Since no ground truth for disambiguated (resolved) locations (with latitude and longitude coordinates) is available, we have to relate the performance of this technique with the appropriateness evaluation of the recommendations.

5.1 Model of Place Identity

For modelling the place identity used by PI and ND+PI, we first performed items clustering based on the inferred locations of the items with *leader-follower* method (distance threshold = 200 km). This results in 57 clusters over all items with maximal distance of a cluster member to centroid is about 230 km. Next, for each cluster, we want to define the common topics in that cluster. For this purpose, we created a vector over all terms in the whole dataset for each item using TF-IDF (*term frequency-inverse document frequency*). We defined the central topics in a cluster by computing the mean centroid of the term vectors in each cluster. Next, the similarity of each cluster item with the centroid is computed with the *cosine similarity*. Table 1 shows an example of cluster resulting from the approach described above. The cluster consists of 53 items and the average of similarity computations to the centroid is 0.341. To recommend an item in a given location L, PI first looks for the nearest cluster with the smallest distance between its centroid and L. Next, the average of item similarity with the centroid (the place identity) is computed and an item with a lower similarity than

[1] http://www.jaring-ide.com/
[2] http://www.geonames.org/

Table 1: The topic extraction of a cluster showing 4 (of 53) example members.

Centroid topics ($avg = 0.341$)**:** Aceh, tsunami, fish*, fisherman*, beach*		
Items	**Sim**	**Topics**
Above *avg*	0.665	Aceh, tsunami, Province*, island*, hit*
Above *avg*	0.615	Aceh, fishing*, fish*, sun*, region*
Below *avg*	0.329	dance performance*, colonialism*, Dance*, Aceh, allowed*
Below *avg*	0.089	art*, element*, festive*, epoch*, Dance*

*word translated from Bahasa to ease the observation

the average similarity (hence, not similar to the usual topics) is picked (items labelled as *Below avg* in Table 1).

5.2 User Study

We performed a user study using a web application that shows suggested stories (news articles) based on a current location. The assumed current location is generated randomly from a set of about 300 regencies and cities in Indonesia. In every recommendation session, four stories are suggested by four approaches: ND (as a baseline algorithm), GH, EA, and PI. In every recommendation session (on a web page), the order of these stories is shuffled and hence the user can not find out which item is recommended by which technique. For each suggested story the user is asked to submit evaluations in three categories: ***appropriate, like, surprising***. The category *appropriate* is the measure of how suitable the story is with the given location (since the toponym resolution was performed without ground truth as in a real-life application). Next, user can assess the quality of the story in the category *like*. Finally, the category *surprising* defines the metric of how unusual the topic of the story in the area of the given location is. The evaluation is submitted in form of a 5-scale rating (from disagree to agree).

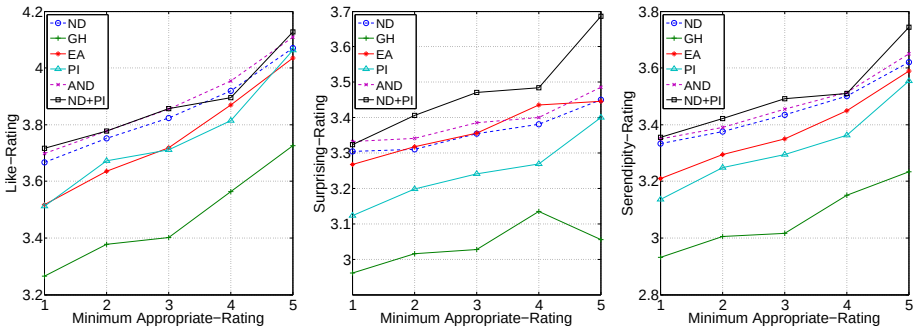
In this user study, 44 users with general knowledge about locations in Indonesia were asked to assess the recommended articles in each given location (165 locations were randomly given across the experiment). The result comprises 827 ratings distributed over stories that were recommended by the approaches (ND: 207, GH: 204, EA: 205, PI: 211) on 232 recommendation pages (which means that some pages did not receive complete ratings for all 4 stories). In addition to the online elicited ratings, we defined *serendipity-rating* as **serendipity = (like + surprising) / 2** (since serendipity involves unexpected (surprising) but pleasant (liked) aspects). We also run offline recommendation on the already rated stories with **ND+PI** and **AND** (Absolute Nearest Distance) as another baseline. This is done because: (1) GH, EA, and PI do not have real objective functions (partially random); (2) not all stories were rated completely on every recommendation page. For every page with missing ratings for ND, AND recommends other rated items with the nearest distance.

The summary of the evaluation results is partly presented in Table 2. The table presents the average of ratings for each technique and each rating category with appropriate-rating = 5 (assumed to be recommended appropriately). The result from ND can for instance be regarded as the parameter for the overall

Table 2: Results of ratings in the experiment *Stories around You*.

	ND	**GH**	**EA**	**PI**	**AND**	**ND+PI**
#appropriate $\geq$ 5	100	73	83	65	111	86
like-rating	4.070	3.726	4.036	4.062	4.108	**4.128**
new-rating	3.450	3.055	3.446	3.400	3.486	**3.686**
serendipity-rating	3.620	3.233	3.590	3.554	3.649	**3.744**

appropriateness of the recommendation: 160 out of 232 items (about 68.9%) were evaluated with rating ≥ 4. Aside from the fact that the inference may have been wrong at the first place, there may be 3 other causes for an inappropriate recommendation: (1) not enough news articles to recommend at the location; (2) a nearest item is from another adjacent regency or even another adjacent province (since no shared-parent check); (3) the participants think the location is not central to the story even though it is inferred correctly.



(a) Like (b) Surprising (c) Serendipity

Fig. 1: Ratings based on the appropriateness range

The overall comparison and the development of the like-, surprising- and serendipity-ratings of the approaches along the ranges of appropriate-rating is illustrated in Figure 1. As can be seen in this figure, our approach ND+PI can perform as well as both of the baseline approaches ND and AND in term of the like-rating (Figure 1a). In terms of both surprising- and serendipity-rating (Figure 1b and 1c), the approach outperforms the baseline approaches in almost all value ranges of appropriate-ratings. PI and EA, in contrast, did not perform well in both surprising- and serendipity-rating as expected originally. We argue that this is caused by the random nature of these approaches as well as the availability of the data (e.g. not enough data with the desired association near the location). Another important insight is to see how the items recommended by GH were seen as less-favoured (even with appropriate-rating = 5), and expectedly less-surprising for the users since the recommended news articles would be more general. This shows the effectiveness of our location inference approach to assign the locations to the correct geographical hierarchy level.

6 Conclusion

We presented approaches for recommending news article by using spatial variables as the main factor of relevance. The aim of these approaches is to deliver serendipitous recommendation and improve the user satisfaction in absence of user preferences. A user study showed that the approaches can find items that are in general more serendipitous (surprising but still favoured) than the ones retrieved by the baseline (distance-based) algorithm. This study can motivate further investigations of context-based serendipitous recommendation by using more complex spatial model (e.g. based on LDA instead of TF-IDF) and location associations, as well as the integration of user preferences where applicable.

References

1. N. Auray. Folksonomy: The new way to serendipity. *Communications & Strategies, No. 65, 2007*, 2007.
2. J. Bao, M. Mokbel, and C.-Y. Chow. Geofeed: A location aware news feed system. In *IEEE 28th International Conference on Data Engineering (ICDE)*, pages 54–65, 2012.
3. Y.-S. Chiu, K.-H. Lin, and J.-S. Chen. A social network-based serendipity recommender system. In *Intelligent Signal Processing and Communications Systems (ISPACS), 2011 International Symposium on*, pages 1 –5, dec. 2011.
4. J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. Lars: A location-aware recommender system. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering*, ICDE '12, pages 450–461, Washington, DC, USA, 2012. IEEE Computer Society.
5. D. Mican, L. Mocean, and N. Tomai. Building a social recommender system by harvesting social relationships and trust scores between users. In W. Abramowicz, J. Domingue, and K. Wecel, editors, *Business Information Systems Workshops*, Lecture Notes in Business Information Processing, pages 1–12. Springer Berlin Heidelberg, 2012.
6. A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 144–153, 2012.
7. M. Schedl, D. Hauger, and D. Schnitzer. A model for serendipitous music retrieval. In *Proceedings of the 2nd Workshop on Context-awareness in Retrieval and Recommendation*, CaRR '12, pages 10–13, New York, NY, USA, 2012. ACM.
8. W. Xu, C.-Y. Chow, M. L. Yiu, Q. Li, and C. K. Poon. Mobifeed: a location-aware news feed system for mobile users. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '12, pages 538–541, New York, NY, USA, 2012. ACM.
9. H. Yin, B. Cui, J. Li, J. Yao, and C. Chen. Challenging the long tail recommendation. *PVLDB*, 5(9):896–907, 2012.
10. Y. C. Zhang, D. O. Séaghdha, D. Quercia, and T. Jambor. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 13–22, New York, NY, USA, 2012. ACM.