Research Paper Recommender System with Serendipity using Tweets vs. Diversification

Chifumi Nishioka[1], Jörn Hauk[2], and Ansgar Scherp[3]

[1] Kyoto University Library, Kyoto, Japan
nishioka.chifumi.2c@kyoto-u.ac.jp
[2] Kiel University, Kiel, Germany
[3] University of Essex, Colchester, UK
ansgar.scherp@essex.ac.uk

Abstract. So far, a lot of works have studied research paper recommender systems. However, most of them have focused only on the accuracy and ignored the serendipity, which is an important aspect for user satisfaction. The serendipity is concerned with the novelty of recommendations and to which extent recommendations positively surprise users. In this paper, we investigate a research paper recommender system focusing on serendipity. In particular, we examine (1) whether a user's tweets lead to a generation of serendipitous recommendations and (2) whether the use of diversification on a recommendation list improves serendipity. We have conducted an online experiment with 22 subjects in the domain of computer science. The result of our experiment shows that tweets do not improve the serendipity, despite their heterogeneous nature. However, diversification delivers serendipitous research papers that cannot be generated by a traditional strategy.

Keywords: Recommender system · Scientific publication · Experiment.

1 Introduction

Various works have developed recommender systems for research papers to help researchers overcome the information overload problem [3]. Recommendations are generated based on, e. g., a user's own papers [21] or the papers a user has accessed in the past [16]. Most of the previous works have focused only on the accuracy of recommendations, e .g., on nDCG. However, several works in recommender systems in other domains (e. g., movies) argue that there are important aspects other than the accuracy [14, 10]. One of these aspects is the *serendipity*, which is concerned with the novelty of recommendations and in how far recommendations may positively surprise users [6].

In this paper, we study a research paper recommender system focusing on the serendipity. Sugiyama and Kan [22] have investigated serendipitous research paper recommendations focusing on the influence from dissimilar users and co-author network to the recommendation performance. In contrast, our study investigates the following research questions: (RQ1) Do a user's tweets enable

serendipitous recommendations? (RQ2) Is it possible to improve the serendipity in a recommendation list by diversification? We run an experiment to empirically investigate the two research questions using three factors. For RQ1, we employ the factor *User Profile Source*, where two different user profile sources are compared: a user's own papers and the user's tweets. User's own papers are used in existing recommender systems such as Google Scholar[4]. We assume that user's tweets produce recommendations that cannot be generated based on own papers, since researchers tweet about very recent developments and interests that are yet not reflected in their papers (e. g., what they found interesting at a conference [12] or in the social network). In the domain of economics, recommendations based on a user's tweets received a precision of 60%, which is fairly high [18]. In addition, we analyze the factor *Text Mining Method*, which applies different methods for computing user profiles from different content, e. .g., tweets, research papers, etc. As text mining methods, we compare the classicial TF-IDF [19] with two of its recent extensions that are known to perform well in recommendation tasks, namely CF-IDF [7] and HCF-IDF [17]. For RQ2, we introduce the factor *Ranking Method*, where we compare two ranking methods: classical cosine similarity and the established diversification algorithm IA-Select [2]. IA-Select ranks candidate items with the objective to diversify recommendations in a list. Since it broadens the coverage of topics in a list, we assume that IA-Select delivers serendipitous recommendations compared to cosine similarity.

Along with the three factors, we conduct an experiment with 22 subjects. The result of the experiment reveals that using the user's tweets for making scientific paper recommendations did not improve the serendipity. On the other hand, we confirm that the diversification of a recommendation list by IA-Select delivers serendipitous recommendations.

The paper is organized as follows: Section 2 describes the recommender system and the experimental factors. The evaluation setup is described in Section 3. Section 4 reports and discusses the results before concluding the paper.

2 Recommender System and Experimental Factors

In this paper, we build a content-based recommender system along with the three factors *User Profile Source*, *Text Mining Method*, and *Ranking Method*. It works as follows: (a) Candidate items of the recommender system (i .e., research papers) are processed by one of text mining methods and paper profiles are generated. (b) A user profile is generated based on his/her user profile source (i .e., own papers or tweets) by a text mining method, which is applied to generate paper profiles. (c) One of ranking methods orders recommendations to determine which papers are suggested. The design of the experiment is illustrated in Table 1, where each cell is a possible design choice in each factor. In the following paragraphs, we describe the details of each factor.

[4] `https://scholar.google.co.jp/`

Table 1. Factors and their choices panning in total $3 \times 2 \times 2 = 12$ strategies.

Factor	**Possible Design Choices**		
User Profile Source	Twitter		Own Papers
Text Mining Method	TF-IDF	CF-IDF	HCF-IDF
Ranking Method	Cosine Similarity		IA-Select

User Profile Source In this factor, we compare the data sources that are used to build a user profile: own papers and tweets. As baseline, we use the own papers of the users. This approach is motivated from existing research paper recommender systems such as Sugiyama and Kan [21] and Google Scholar. In contrast, we assume that using tweets provide more serendipitous recommendations. It is common among researchers to tweet about their professional interests on Twitter [12, 9]. Thus, tweets can be used for building a user profile in the context of research paper recommender system. We hypothesize that a user's tweets produce serendipitous recommendations, since researchers tweet about their most recent interests and information that are (yet) not reflected in their papers.

Text Mining Method For each of the two sources on content, i. e., the user's own papers or his/her tweets, we apply a profiling method using one of three text mining methods. This constitutes the second factor of the study. We compare three methods TF-IDF [19], CF-IDF [7], and HCF-IDF [17] to build paper profiles and a user profile. Since text mining method that works well differs depending on the type of contents to be analyzed (e .g., tweets, research papers), we introduce this factor. Thus, this factor is integrated into RQ1. First, we use TF-IDF [19], which is often used in recommender systems as baseline [13, 7]. Second, Concept Frequency Inverse Document Frequency (CF-IDF) [7] is an extension of TF-IDF, which replaces terms with semantic concepts from a knowledge base. The use of a knowledge base decreases noise in profiles [1, 8]. In addition, since a knowledge base can store multiple labels for a concept, synonyms are integrated into one feature. Finally, we apply Hierarchical Concept Frequency Inverse Document Frequency (HCF-IDF) [17], which is an extension of CF-IDF. It applies a propagation function [11] over a hierarchical structure of a knowledge base to give a weight to concepts in higher levels. Thus, it identifies concepts that are not mentioned in a text but highly relevant. HCF-IDF calculates a weight of a concept a in a text t as: $w(a,t) = BL(a,t) \cdot \log \frac{|D|}{|\{d \in D : a \in d\}|}$. $BL(a,t)$ is a propagation function Bell-Log [11], which is defined as: $BL(a,t) = cf(a,t) + FL(a) \cdot \sum_{a_j \in pc(a)} BL(a_j,t)$, where $FL(a) = \frac{1}{\log_{10}(nodes(h(a)+1))}$. The function $h(a)$ returns the level, where a concept a is located in the knowledge base. *nodes* provides the number of concepts at a given level in a knowledge base. $pc(a)$ returns all parent concepts of a concept a. We employ HCF-IDF, since it showed to work well for short texts such as tweets [18].

Ranking Method Finally, we rank all candidate items to determine which items are recommended to a user. In this factor, we compare two ranking methods:

cosine similarity and diversification with IA-Select [2]. As baseline, we employ a cosine similarity, which has been widely used in content-based recommender systems [13]. Top-k items with largest cosine similarities are recommended. Second, we employ IA-Select [2], in order to deliver serendipitous recommendations. IA-Select diversifies recommendations in a list to avoid suggesting similar items together. Although it has been originally introduced in information retrieval [2], it is also used in recommender systems to improve the serendipity [24]. The basic idea of IA-Select is that it lowers iteratively the weights of features in the user profile, which are already covered by selected recommended items. First, IA-Select computes cosine similarities between a user and each candidate item. Subsequently, it adds an item with the largest cosine similarity to the recommendation list. After picking the item, IA-Select decreases weights of features covered by the selected item in the user profile. These steps are repeated until k recommendations are determined.

3 Evaluation

We conduct an online experiment with $n = 22$ subjects to answer the two research questions. The setup of the experiment follows the previous works [18, 5]. In the subsequent paragraphs, we describe the procedure of the experiment, subjects, dataset, and metric, respectively.

Procedure We have implemented a web application where human subjects evaluate the twelve recommendation strategies. First, subjects start on the welcome page, which asks for consent to the data collection. Thereafter, subjects are asked to input their Twitter handle and their name as recorded DBLP Persons[5]. Based on their name, we retrieve a list of a user's papers and obtain the content of the papers by mapping them to the ACM-Citation-Network V8 dataset, which is described later. The top-5 recommendations are computed for each strategy. Thus, subjects have to evaluate $5 \cdot 12 = 60$ items as "interesting" or "not interesting" based on relevance to their research interests. Items are displayed with the bibliographic information including authors, title, year, and venue. Furthermore, subjects can directly access and read the research paper by clicking on the link of an item. In order to avoid a bias, the sequence of the twelve strategies is randomized for each subject. The list of top-5 items of each strategy is randomized as well, to avoid the well-known ranking bias [4, 5]. After evaluating all strategies, subjects are asked to fill out a form about demographic information such as age and profession. Finally subjects can state qualitative feedback about the experiment.

Subjects $n = 22$ subjects were recruited through Twitter and mailing lists. Subjects are on average 36.45 years old (SD: 5.55). Regarding the academic degree, two subjects have a Master, thirteen a Ph.D., and seven are lecturer/professors. Subjects published on average 1256.97 tweets (SD: 1155.8). Regarding research papers for user profiling, on average a subject has 11.41 own papers (SD: 13.53).

[5] https://dblp.uni-trier.de/pers/

Datasets As a corpus of research papers, we use the ACM-Citation-Network V8 dataset provided by the ArnetMiner [23]. From the dataset, we use 1,669,237 of 2,381,688 research papers that have title, author, year of publications, venue, and abstract. We use title and abstract to generate paper profiles. As a knowledge base for CF-IDF and HCF-IDF, we use the ACM Computing Classification System (CCS) [6]. It focuses on computer science and is organized in a hierarchical structure. It consists of 2,299 concepts and 9,054 labels.

Metric To evaluate the serendipity of recommendations, we use the Serendipity Score (SRDP) [6]. It takes into account both of unexpectedness and usefulness of candidate items, which is defined as: $SRDP = \sum_{d \in UE} \frac{rate(d)}{|UE|}$. UE denotes a set of unexpected items that are recommended to a user. An item is considered as unexpected, if it is not included in a recommendation list computed by the primitive strategy. We use the strategy Own Papers × TF-IDF × Cosine Similarity as a primitive strategy, since it is a combination of baselines. The function $rate(d)$ returns an evaluation rate of an item d given by a subject. If a subject evaluates an item as "interesting", it returns 1. Otherwise, it returns 0.

4 Result and Discussion

Table 2 shows the results of twelve strategies in terms of SRDP. We use the strategy Own Papers × TF-IDF × Cosine Similarity as a primitive strategy. Thus, mean and standard deviation of SRDP are 0.00 for that strategy as shown at the bottom in Table 2. An ANOVA is conducted to verify significant differences between strategies. The significance level for statistical tests is set to $\alpha = .05$. The Muchly's test ($\chi^2(54) = 80.912$, $p = .01$) detects a violation of sphericity. Thus, a Greenhouse-Geisser correction with $\epsilon = 0.58$ is applied. The ANOVA reveals significant differences between the strategies ($F(5.85, 122.75) = 3.51$, $p = .00$). Furthermore, a Shaffer's MSRB procedure [20] is conducted to find the pairwise differences. We observe several differences, but all of them are differences between the primitive strategy and one of the other strategies.

In order to analyze the impact of each experimental factor, a three-way repeated measures ANOVA is conducted. The Mendoza Test identifies violation of sphericity [15] for the global ($\chi^2(65) = 101.83$, $p = .0039$) and the factor *Text Mining Method* × *Ranking Method* ($\chi^2(2) = 12.01$, $p = .0025$). Thus, the three-way repeated-measure is applied with a Greenhouse-Geiser correction of $\epsilon = .54$ for the global and $\epsilon = .69$ for the factor *Text Mining Method* × *Ranking Method*. Table 3 shows the result with F-Ratio, effect size η^2, and p-value. Regarding the single factors, *Ranking Method* has the largest impact on SRDP, as an effect size η^2 presents. A post-hoc analysis reveals that the strategies using IA-Select make higher SRDP than those with cosine similarity. In addition, we observe a significant difference in the factors *User Profile Source* × *Ranking Method* and *Text Mining Method* × *Ranking Method*. In both factors, post-hoc

[6] https://www.acm.org/publications/class-2012

Table 2. SRDP and the number of unexpected items of the twelve strategies. The values are ordered by SRDP. M and SD denote mean and standard deviation.

	Strategy			SRDP	|UE|
	Text Mining Method	**Profiling Source**	**Ranking Method**	**M (SD)**	**M (SD)**
1.	TF-IDF	Own Papers	IA-Select	.45 (.38)	2.95 (1.05)
2.	CF-IDF	Twitter	CosSim	.39 (.31)	4.91 (0.29)
3.	TF-IDF	Twitter	IA-Select	.36 (.29)	4.91 (0.43)
4.	CF-IDF	Twitter	IA-Select	.31 (.22)	4.95 (0.21)
5.	CF-IDF	Own Papers	CosSim	.26 (.28)	4.91 (0.29)
6.	CF-IDF	Own Papers	IA-Select	.25 (.28)	4.91 (0.29)
7.	HCF-IDF	Own Papers	IA-Select	.24 (.22)	4.95 (0.21)
8.	HCF-IDF	Twitter	CosSim	.22 (.28)	5.00 (0.00)
9.	TF-IDF	Twitter	CosSim	.20 (.24)	4.95 (0.21)
10.	HCF-IDF	Twitter	IA-Select	.18 (.21)	5.00 (0.00)
11.	HCF-IDF	Own Papers	CosSim	.16 (.18)	5.00 (0.00)
12.	TF-IDF	Own Papers	CosSim	.00 (.00)	0.00 (0.00)

analyses reveal significant differences, when a baseline is used in either of the two factors. When a baseline is used in one factor, $|UE|$ becomes small unless a method other than baseline is used in the other factor.

Table 3. Three-way repeated-measure ANOVA for SRDP with Greenhouse-Geisser correction with F-ratio, effect size η^2, and p-value.

Factor	**F**	η^2	**p**
User Profile Source	2.21	.11	.15
Text Mining Method	3.02	.14	.06
Ranking Method	14.06	.67	**.00**
User Profile Source × *Text Mining Method*	0.98	.05	.38
User Profile Source × *Ranking Method*	18.20	.87	**.00**
Text Mining Method × *Ranking Method*	17.80	.85	**.00**
User Profile Source × *Text Mining Method* × *Ranking Method*	2.39	.11	.11

The results of our experiment reveal that tweets do not improve the serendipity of recommendations. As shown at the rightmost column in Table 2, tweets deliver unexpected recommendations to users. However, only a small fraction of these serendipitous recommendations were interesting to the users. The results show further that the IA-Select algorithm delivers serendipitous research paper recommendations. Thus, we extend on the related works of using IA-Select to improve serendipity to the context of a research paper recommender.

5 Conclusion

In this paper, we investigate whether tweets and IA-Select deliver serendipitous recommendations. We conduct an experiment following the three factors. The result of the experiment reveals that tweets do not improve the serendipity of recommendations, but IA-Select does. This insight contributes to future recommender systems in such a sense that a provider can make informed design choices for the systems and services developed.

References

1. Abel, F., Herder, E., Krause, D.: Extraction of professional interests from social web profiles. In: Proceedings of ACM Conference on User Modeling, Adaptation and Personalization (2011)
2. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proceedings of ACM International Conference on Web Search and Data Mining (WSDM). pp. 5–14. ACM (2009)
3. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: A literature survey. International Journal on Digital Libraries **17**(4), 305–338 (2016)
4. Bostandjiev, S., O'Donovan, J., Höllerer, T.: Taste- Weights: a visual interactive hybrid recommender system. In: Proceedings of ACM Conference on Recommender Systems (RecSys). ACM (2012)
5. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: Proceedings of SIGCHI Conference on Human Factors in Computing Systems (CHI). ACM (2010)
6. Ge, M., Delgado-Battenfeld, C., Jannach, D.: Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In: Proceedings of ACM Conference on Recommender Systems (RecSys). pp. 257–260. ACM (2010)
7. Goossen, F., IJntema, W., Frasincar, F., Hogenboom, F., Kaymak, U.: News personalization using the CF-IDF semantic recommender. In: Proceedings of International Conference on Web Intelligence, Mining and Semantics (WIMS). ACM (2011)
8. Große-Bölting, G., Nishioka, C., Scherp, A.: A comparison of different strategies for automated semantic document annotation. In: Proceedings of the 8th International Conference on Knowledge Capture (K-CAP). pp. 8:1–8:8. ACM (2015)
9. Große-Bölting, G., Nishioka, C., Scherp, A.: Generic process for extracting user profiles from social media using hierarchical knowledge bases. In: Proceedings of International Conference on Semantic Computing (IEEE ICSC). pp. 197–200. IEEE (2015)
10. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS) **22**(1), 5–53 (2004)
11. Kapanipathi, P., Jain, P., Venkataramani, C., Sheth, A.: User interests identification on Twitter using a hierarchical knowledge base. In: Proceedings of European Semantic Web Conference (ESWC). Springer (2014)
12. Letierce, J., Passant, A., Breslin, J.G., Decker, S.: Understanding how Twitter is used to spread scientific messages. In: Proceedings of Web Science Conference. Web Science Trust (2010)
13. Lops, P., De Gemmis, M., Semeraro, G.: Content-based recommender systems: State of the art and trends. In: Recommender systems handbook, pp. 73–105. Springer (2011)
14. McNee, S.M., Riedl, J., Konstan, J.A.: Being accurate is not enough: How accuracy metrics have hurt recommender systems. In: CHI Extended Abstracts on Human Factors in Computing Systems. pp. 1097–1101. ACM (2006)
15. Mendoza, J.L.: A significance test for multisample sphericity. Psychometrika **45**(4) (1980)

16. Nascimento, C., Laender, A.H., da Silva, A.S., Gonçalves, M.A.: A source independent framework for research paper recommendation. In: Proceedings of International ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 297–306. ACM (2011)
17. Nishioka, C., Große-Bölting, G., Scherp, A.: Influence of time on user profiling and recommending researchers in social media. In: Proceedings of International Conference on Knowledge Technologies and Data-driven Business (i-KNOW). ACM (2015)
18. Nishioka, C., Scherp, A.: Profiling vs. time vs. content: What does matter for top-k publication recommendation based on Twitter profiles? In: Proceedings of IEEE/ACM Joint Conference on Digital Libraries (JCDL). pp. 171–180. ACM (2016)
19. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management **24**(5) (1988)
20. Shaffer, J.P.: Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association **81**(395) (1986)
21. Sugiyama, K., Kan, M.Y.: Scholarly paper recommendation via user's recent research interests. In: Proceedings of ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL). ACM (2010)
22. Sugiyama, K., Kan, M.Y.: Towards higher relevance and serendipity in scholarly paper recommendation. ACM SIGWEB Newsletter (Winter), 4 (2015)
23. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: Extraction and mining of academic social networks. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). pp. 990–998. ACM (2008)
24. Vargas, S., Castells, P., Vallet, D.: Explicit relevance models in intent-oriented information retrieval diversification. In: Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). pp. 75–84. ACM (2012)