Serendipitous Recommendation for Scholarly Papers Considering Relations Among Researchers

Kazunari Sugiyama*
National University of Singapore
Computing 1, 13 Computing Drive,
Singapore 117417
sugiyama@comp.nus.edu.sg

Min-Yen Kan
National University of Singapore
Computing 1, 13 Computing Drive,
Singapore 117417
kanmy@comp.nus.edu.sg

ABSTRACT

Serendipity occurs when one finds an interesting discovery while searching for something else. In digital libraries, recommendation engines are particularly well-suited for serendipitous recommendations as such processes work without needing queries. Junior researchers can use such scholarly recommendation systems to broaden their horizon and learn new areas, while senior researchers can discover interdisciplinary frontiers to apply integrative research. We adapt a state-of-the-art scholarly paper recommendation system's user profile construction to make use of information drawn from 1) dissimilar users and 2) co-authors to specifically target serendipitous recommendation.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering, Search process; H.3.7 [**Digital Libraries**]: Systems issues

General Terms

Algorithms, Experimentation, Human factors, Performance

Keywords

Recommendation, Serendipity, User modeling

1. INTRODUCTION

Scholars will often seek out colleagues for informal advice and attend seminars and conferences outside of their areas to broaden their horizons. Such interactions can lead to serendipitous discovery of new ideas, approaches or ways of thinking, and applies to researchers of all levels of experience. For example, junior researchers need to broaden their range of research interests to acquire knowledge. Senior researchers have expertise in their own fields, but may seek to apply their knowledge towards other areas or import ideas familiar to other areas to their own. Such cross-pollination work is a hallmark of productive interdisciplinary exchange. We see that both types of researchers would thus benefit from serendipitous recommendation.

The digital library recommender systems [8, 9], in contrast, generate recommendations of related scholarly work relevant to a user, often without the need for an explicit query. In such cases, contextual information about the user can provide evidence for recommendation. Because of this crucial distinction, users perceive recommendation differently from search and may be more willing to accept or explore serendipitous recommendations. It may even be argued that serendipitous recommendations are more important than strongly similar relevant recommendations, since directed searching can uncover the latter.

To achieve this, we utilize information inherent in other researchers to improve recommendations for a target researcher through two approaches. Our approaches are analogous to asking colleagues for advice or recommendation on what they find interesting. The preferences gathered from other users are used in the construction of the target researcher's user profile, used in matching candidate documents in the recommendation processes.

From one perspective, a scholar may find the most surprising ideas from fields furthest away from his own study. In our first approach, we model this by gathering recommendation evidence from the *dissimilar users* to the target researcher.

From another perspective, a scholar may find the most useful recommendations from his scholarly social network. Researchers who have collaborated before may communicate to each other interesting trends and help bridge gaps in each other's knowledge on related fields. In our second approach, we model this by collecting evidence from *co-authors* of the target researcher.

2. RELATED WORK

Ziegler *et al.* [11] proposed a similarity using a taxonomy-based classification and use it to compute an intra-list similarity to determine the overall diversity of the recommended list. They provide a heuristic algorithm to increase the diversity of the recommendation list. Nakatsuji *et al.* [7] improved Ziegler *et al.*'s approach by equating highly novel items as those which belong to the class with the smallest distance from the class the user accessed before. Zhang and Hurley [10] focused on intra-list diversity and optimized the tradeoff between users' preferences and the diversity of the top-N results. They modeled the competing goals of maximizing the diversity of the searched list while maintaining adequate similarity to the user query as a binary optimization problem. Andre *et al.* [2] proposed a method for performing serendipitous searches for Web information retrieval. They first defined the potential for serendipity as search results that are interesting but not highly relevant. In another publication, they discussed serendipity from human cognitive point of view [1]. They hypothesized that a reconsideration of serendipity from other angles may help designing systems to support the desired effects of serendipitous revelation. Lathia *et*

*Dr Sugiyama was supported in part by the Global Asia Institute under grant no. GAI-CP/20091116. Both authors would like to acknowledge support from the National Research Foundation's grant no. R-252-000-325-279.

JCDL'11 June 13–17, 2011, Ottawa, Ontario, Canada.

al. [6] found that temporal diversity is an important facet of recommender systems by showing how data of collaborative filtering changes over time. Kawamae [5] emphasized the surprise of each user in the recommendation focusing on the estimated search time that the users would take to find the item by themselves. Their recommender system assumed that items recently purchased by an "innovator", who has well-proven unpredictable trait, will surprise other users more than other items.

3. PROPOSED METHOD

Our work in this paper extends our prior work in [8]. At one level, our work fits a standard recommender system architecture consisting of the following three steps:

1. We first construct a basic user profile $\boldsymbol{P}_u$ for each researcher u that we need to generate recommendations for. Then, by using this basic user profile, we construct a user profile $\boldsymbol{P}_u^{srdp}$ for serendipitous recommendation;
2. We then construct feature vectors $\boldsymbol{F}^{p_{rec_j}}$ $(j = 1, \cdots, t)$ for candidate papers to recommend by using citation papers and reference papers of the target paper;
3. We compute cosine similarity $Sim(\boldsymbol{P}_u^{srdp}, \boldsymbol{F}^{p_{rec_j}})$ between $\boldsymbol{P}_u^{srdp}$ and $\boldsymbol{F}^{p_{rec_j}}$ $(j = 1, \cdots, t)$ to generate recommendations, returned in the order of high similarity.

3.1 Basic User Profile Construction

In the scheme, we built a feature vector $\boldsymbol{f}^p$ representing his paper p using the TF of terms that appear in the publication[1]. In the novel step, we modified the assigned weights for terms to account for the influence of the papers in the citation network neighborhood. Papers that cite the target paper (termed *citation papers*) as well as those that the target paper references (termed *reference papers*) influenced the original $\boldsymbol{f}^p$ weighting.

For *junior researchers*, defined as having only one recently published paper (p_1) which has not been cited, the user profile is simply the vector $\boldsymbol{f}^{p_1}$ for single paper p_1, with the weighting modified by its context of reference papers:

$$\boldsymbol{P}_u = \boldsymbol{f}^{p_1} + \sum_{y=1}^{l} W^{p_1 \to ref_y} \boldsymbol{f}^{p_1 \to ref_y}, \tag{1}$$

where $W^{p_1 \to ref_y}$ $(y = 1, \cdots, l)$ denotes the cosine similarity assigned to paper $p_1 \to ref_y$, computed on the basis of paper p_1.

For *senior researchers*, characterized as having n past papers p_i $(i = 1, \cdots, n)$, the individual feature vectors for each paper have an enlarged context accounting for possible citation papers (corresponding to the additional third term below):

$$\boldsymbol{F}^{p_i} = \boldsymbol{f}^{p_i} + \sum_{y=1}^{l} W^{p_i \to ref_y} \boldsymbol{f}^{p_i \to ref_y} + \sum_{x=1}^{k} W^{p_{cx} \to p_i} \boldsymbol{f}^{p_{cx} \to p_i}. \tag{2}$$

Secondly, as research interests of a (senior) researcher change over time, the user profile construction process models this by using a tunable *forgetting factor* that assigns less weight to papers published further in the past. The user profile for the general case is thus defined as:

$$\boldsymbol{P}_u = \sum_{z=1}^{n-1} e^{-\gamma \cdot d_{n \to z}} \boldsymbol{F}^{p_z} + \boldsymbol{F}^{p_n}, \tag{3}$$

where $e^{-\gamma \cdot d_{n \to z}}$ denotes the weight between $[0, 1]$ assigned to paper $p_{n \to z}$ computed on the basis of the most recent paper p_n. Here,

[1]Note that we prefer TF over standard TF-IDF in the construction process, as the limited size of a researcher's publication list often does not allow reliable estimates of IDF.

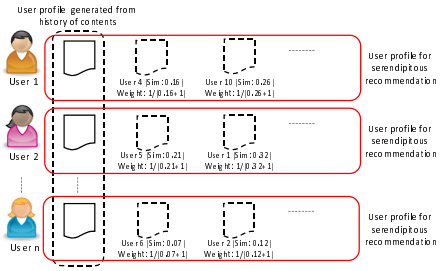


Figure 1: User profile construction with dissimilar users for serendipitous recommendation.

γ is the forgetting coefficient $(0 \leq \gamma \leq 1)$ and $d_{n \to z}$ is the difference between the published year of the most recent paper n and the previously published work z.

In this paper, we investigate how serendipitous recommendation can be influenced through the modification of the following user profile construction process.

3.1 Profile Construction via Dissimilar Users (DU)

Researchers whose interests differ from a target user may be able to generate interesting and surprising recommendations. For this reason, our dissimilar users (DU) approach utilizes profiles from users that are maximally different from our target user. We use the reciprocal of similarity between the target user and a candidate user to rank candidate users with respect to their dissimilarity.

Figure 1 shows user profile construction with dissimilar users for serendipitous recommendation.

Suppose, for example, that User 1 is the target user and that the similarity between Users 1 and 4 is 0.16. In this case, the weight assigned to the profile of User 4 for User 1 is computed by taking the reciprocal of (the cosine similarity + k) : 1/(0.16+k). k is used to place a bounded limit on the dissimilarity value; in our work, we set $k = 1$, such that dissimilarity values range between $[1, \frac{1}{2}]$.

Weights assigned to the other users' profiles for the target user are computed in the same manner and combined together with the original target user profile from the baseline. Let $\boldsymbol{P}_u^{srdp}$ be the modified user profile for user u for serendipitous recommendation. This scheme is formalized as follows:

$$\boldsymbol{P}_u^{srdp} = \boldsymbol{P}_u + \sum_{v=1}^{N_{du}} \left(\frac{1}{sim(\boldsymbol{P}_u, \boldsymbol{P}_v) + k} \times \boldsymbol{P}_v \right), \tag{4}$$

where $\boldsymbol{P}_u$ and $\boldsymbol{P}_v$ are the basic user profiles of user u and users $v = (1, \cdots, N_{du})$, who are the dissimilar users for user u. The fractional term is the weighting factor for the dissimilar user; as described above, it is essentially the reciprocal of the cosine similarity between u and v's profile. We experimented with three methods for selecting dissimilar users:

1. **Selection by common title words (DU-title):** The words in the title are good cues to find serendipitous papers because we often find papers serendipitous from titles in the proceedings or conference program. Thus, we employ titles of the paper to select dissimilar users.

2. **Selection by common references (DU-refs):** We expect that references are also good cues to find serendipitous papers. If different users refer the same paper, user profile with different topics can be easily constructed. We expect such user profiles contribute to serendipitous recommendation.

3. **Selection by thresholding the cosine similarity (SIM-x):** In this approach, we first set the threshold of similarity between users to select dissimilar users, and then construct user

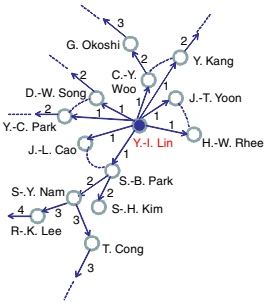


Figure 2: User profile construction with co-authors' network for serendipitous recommendation.

profile for serendipitous recommendation. Let x be the value of threshold. We select users whose similarity between a target user is less than x. The reason why we focus on 'less than' is that our aim is to construct user profile with dissimilar users to recommend serendipitous papers.

3.2 Profile Construction via the Co-author Network (CAN)

Researchers are often collaborative; teaming up with others to do research and capitalize on each other's expertise. A trusted co-author often serves as a sounding board for ideas, lends inspiration and motivation, and importantly for us, can give a different perspective on a research area.

In much the same way, our second approach modifies the construction of the user profile by utilizing a target user's co-author network. We modify the co-author network using the following two constraints, which give rise to networks similar to Figure 2.

(c1) We place co-authors in the network at a "collaborative distance" with respect to their minimal transitive co-authorship distance to the target user u.

(c2) We ignore authoring relationships between other co-authors; in other words, we consider only the radial network centered on the target user.

In the (CAN) scheme, we define $\boldsymbol{P}_u^{srdp}$ as the user profile for serendipitous recommendation of user u as follows:

$$\boldsymbol{P}_u^{srdp} = \boldsymbol{P}_u + \sum_{pl=1} \sum_{ca_{(pl)}=1}^{N_{(pl)}} w_{ca_{(pl)}} \boldsymbol{P}_{ca_{(pl)}} \quad , \quad (5)$$

where $\boldsymbol{P}_u$ and $\boldsymbol{P}_{ca_{(pl)}}$ are basic user profiles of user u and co-author $ca_{(pl)}$ that is separated from u with a path length of pl from the user u, respectively. Here, $w_{ca_{(pl)}}$ is the multiplicative coefficient used to integrate $\boldsymbol{P}_{ca_{(pl)}}$ with $\boldsymbol{P}_u$. In constructing the modified user profile for recommendation, we explore four methods (W1 to W4 below) to set $w_{ca_{(pl)}}$:

W1. **Linear Combination (LC):** This weighting scheme simply combines user profile $\boldsymbol{P}_u$ of user u and user profile $\boldsymbol{P}_{ca_{(pl)}}$ of co-author $ca_{(pl)}$ as follows: $w_{ca_{(pl)}} = 1$.

W2. **Reciprocal of Path Length (RCP-PL):** This weighting scheme assigns larger weights to closer co-authors and smaller weights to distant co-authors from the target user, using reciprocal weighting: $w_{ca_{(pl)}} = 1/(pl + k)$, where k is a constant. We set $k = 1$ for simplicity, similar to what was done in the DU method.

W3. **Reciprocal of the Similarity (RCP-SIM):** The purpose of this work is to provide serendipitous recommendation. Therefore, we assign larger weight to dissimilar users. As in Equation (4), the reciprocal of similarity between user u and co-author $ca_{(pl)}$ whose pl is distant from user u, is defined as: $w_{ca_{(pl)}} = 1/(sim(\boldsymbol{P}_u, \boldsymbol{P}_{ca_{(pl)}}) + k)$, where k is again set to 1, for identical reasons.

W4. **Product of W2 and W3 (RCP-PLSIM):** This weighting scheme $w_{ca_{(pl)}}$ combines path length and cosine similarity by taking the product of the equations in W2 and W3: $w_{ca_{(pl)}} = (1/(pl + k)) \times (1/(sim(\boldsymbol{P}_u, \boldsymbol{P}_{ca_{(pl)}}) + k))$.

Subsequent steps in the construction of feature vectors for candidate papers and in matching feature vectors for (serendipitous) recommendation are identical to our prior work in [8].

4. EXPERIMENTS

We use the same dataset[2] constructed in our previous work [8]. This is a publicly available scholarly recommendation dataset, comprising of the following components: feature vectors of candidate papers to recommend constructed from the ACL Anthology Reference Corpus[3] [3]; canonical citation and reference links for each potential paper to be recommended; as well as feature vectors representing users' research interests, derived from DBLP[4] list published by the 28 users surveyed in the dataset.

In our current work, we also asked each of the researchers to mark papers that are serendipitous papers as well as relevant papers to their research interest.

We assess our two user profile construction methods using four metrics: normalized discounted cumulative gain (nDCG), mean reciprocal rank (MRR), recall, and a normalized version of item novelty metric (nITN), employed in [10]. Due to space limits, we only show experimental results evaluated with nITN, and specifically examine nITN at $n = 10$ (nITN@10). Evaluation using other metrics were similar.

4.1 Results with Dissimilar Users (DU)

As shown in Figure 3 (a,b), when the number of dissimilar users N_{du} is small ($N_{du} \leq 4$ and $N_{du} \leq 5$ for junior and senior researchers, respectively), recommendation accuracy obtained by (DU), (DU-title), and (DU-refs) is low.

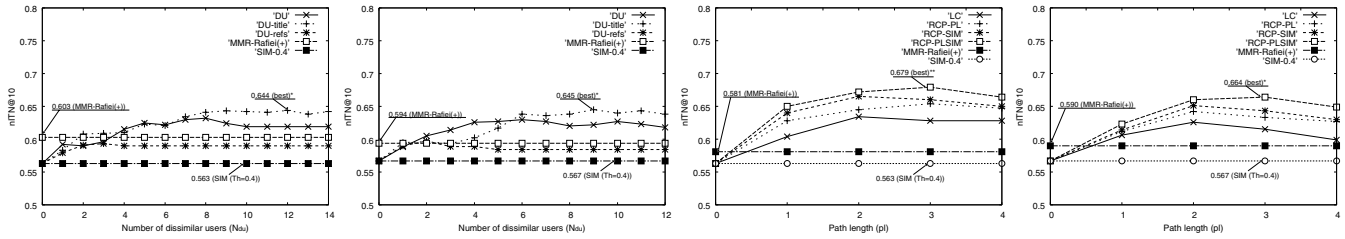
However, when N_{du} is tuned larger ($N_{du} \geq 8$ and $N_{du} \geq 6$ for junior and senior researchers, respectively), the user profiles start being effective. We conjecture that idiosyncrasies from individual DUs are smoothed over to give better recommendations in these cases. We obtained the best results in using (DU) for (DU-title) with $N_{du} = 12$ and $N_{du} = 9$ for junior and senior researchers, respectively. This indicates that user profiles that contain a variety of topics can be constructed by using more dissimilar users and it leads to recommendation of more serendipitous papers. Importantly, diversification beyond this level did not improve results, perhaps due to the gradual dilution of the signal that the approaches generate.

Among the three DU methods explored, (DU-title) proved the most effective. For junior and senior researchers, the best recommendation accuracies in nITN@10 are 0.644 and 0.645, respectively. We also find that in nITN@10 for junior researchers, we can observe convergence as the value of N_{du} becomes larger in (DU), (DU-title), and (DU-refs).

4.2 Results with the Co-Author Network (CAN)

For both junior and senior researchers, we observe that all co-authors often appear up to four paths from the target researcher. In addition, as shown in Figure 3 (c,d), we observe convergence of recommendation accuracy at a path length of three from the target user ($pl = 3$) for both junior and senior researchers.

[2]http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html
[3]Version 20080325, http://acl-arc.comp.nus.edu.sg/
[4]http://www.informatik.uni-trier.de/~ley/db/index.html



(a) nITN@10 via DU (for juniors) (b) nITN@10 via DU (for seniors) (c) nITN@10 via CAN (for juniors) (d) nITN@10 via CAN (for seniors)

Figure 3: Recommendation accuracy measued by nITN@10 for junior and senior researchers obtained by user profile constructed using dissimilar users (DU) in subfigures (a) and (b), respectively; and using the co-author network (CAN) in subfigures (c) and (d), respectively. "" and "*" denote the difference between MMR-Rafiei(+) and the best value of 'DU (title)' in (a,b) and 'RCP-PLSIM' in (c,d) is significant for $p < 0.01$ and $p < 0.05$, respectively.**

We compare across (CAN) and (DU) methodologies. In junior researchers, the best recommendation accuracy of nITN@10 (0.679) in (CAN) outperform that of ITN@10 (0.644) in (DU). This shows that user profile construction using co-author network is effective approach to recommend serendipitous papers compared with that using dissimilar users. In contrast, for senior researchers, the best value of nITN@10 (0.664) in (CAN) outperform that of nITN@10 (0.645) in (DU). Consistent with our observations with junior researchers, it is more effective approach to construct user profile using co-author network (CAN) rather than dissimilar users (DU).

4.3 Comparing MMR with Our Approaches

We compared the effectiveness of our approaches with the well-known diversifications strategy of Maximal Marginal Relevance (MMR) [4], which combines query relevance with result novelty.

According to Figure 3 (a,b), in (DU), the recommendation accuracy obtained by MMR-Rafiei(+) always outperforms our baseline system (SIM-0.4), which indicates that diversification in general boosts results.

Serendipitous recommendation approaches leveraging the co-author network (CAN) almost always outperforms results obtained by MMR-Rafiei(+). Our results indicate that this finding is independent of the choice of weighting scheme (LC), (RCP-PL), (RCP-SIM), and (RCP-PLSIM), and independent of path length (pl). In summary, the recommendation accuracy obtained by both of our proposed approaches, (DU) and (CAN), effectively improve recommendation accuracy over MMR-Rafiei(+).

According to our experiments described above, our two approaches can enrich the baseline user profile to recommend serendipitous papers to both junior and senior researchers by considering relations with neighboring researchers. In user profile construction with dissimilar users (DU), we achieve the best recommendation accuracy by selecting 12 and 9 dissimilar users with common titles, for junior and senior researchers, respectively. In profile construction via the co-author network (CAN), we can achieve the best recommendation accuracy when the path length from the target user is three for both junior and senior researchers.

5. CONCLUSION

We have implemented and evaluated two approaches user profile construction which considers relations between a target researcher and "neighboring" researchers. The two techniques utilize dissimilar users' or co-authors' knowledge to generate serendipitous recommendations. Experimental results along the standard evaluation metrics show that our proposed approaches work to generate statistically better recommendations over a suitable MMR-based diversity algorithm. Further microanalysis confirms the serendipitous nature of the recommendations generated by our systems.

In future work, we plan to further develop a method for constructing user profile that provides highly accurate recommendation of serendipitous papers, especially, focusing on how to select co-authors and appropriate weighting scheme to them.

Acknowledgments: We wish to thank the researchers of the Innovation Policy Research Center, at the University of Tokyo, whom have given us valuable comments on this research.

6. REFERENCES

[1] P. Andre, m.c. schraefel, J. Teevan, and S. T. Dumais. Discovery is Never by Chance: Designing for (Un)Serendipity. In *Proc. of the 7th SIGCHI Conference on Creativity and Cognition (C&C'09)*, pages 305–314, 2009.

[2] P. Andre, J. Teevan, and S. T. Dumais. From X-Rays to Silly Putty via Uranus: Serendipity and its Role in Web Search. In *Proc. of the 27th International Conference on Human Factors in Computing Systems (CHI 2009)*, pages 2033–2036, 2009.

[3] S. Bird, R. Dale, B. J. Dorr, B. Gibson, M. T. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. R. Radev, and Y. F. Tan. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proc. of the 6th International Conference on Language Resources and Evaluation Conference (LREC'08)*, pages 1755–1759, 2008.

[4] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 335–336, 1998.

[5] N. Kawamae. Serendipitous Recommendations via Innovators. In *Proc. of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*, pages 218–225, 2010.

[6] N. Lathia, S. Hailes, L. Capra, and X. Amatriain. Temporal Diversity in Recommender Systems. In *Proc. of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*, pages 210–217, 2010.

[7] M. Nakatsuji, Y. Fujiwara, A. Tanaka, T. Uchiyama, K. Fujimura, and T. Ishida. Classical Music for Rock Fans?: Novel Recommendations for Expanding User Interests. In *Proc. of the 19th International Conference on Information and Knowledge Management (CIKM'10)*, pages 949–958, 2010.

[8] K. Sugiyama and M.-Y. Kan. Scholarly Paper Recommendation via User's Recent Research Interests. In *Proc. of the 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL '10)*, pages 29–38, 2010.

[9] D. Yang, B. Wei, J. Wu, Y. Zhang, and L. Zhang. CARES: A Ranking-Oriented CADAL Recommender System. In *Proc. of the 9th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2009)*, pages 203–211, 2009.

[10] M. Zhang and N. Hurley. Avoiding Monotony: Improving the Diversity of Recommendations. In *Proc. of the 2008 ACM Conference on Recommender Systems (RecSys'08)*, pages 123–130, 2008.

[11] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving Recommendation Lists Through Topic Diversification. In *Proc. of the 14th International World Wide Web Conference (WWW2005)*, pages 22–32, 2005.