A New Information Theory-Based Serendipitous Algorithm Design

Xiaosong Zhou
University of Nottingham
Ningbo China
Ningbo, 315100, China
Xiaosong.zhou@nottingham.edu.cn

Qingfeng Wang
University of Nottingham
Ningbo China
Ningbo, 315100, China
Qingfeng.wang@nottingham.edu.cn

Zhan Xu
University of Nottingham
Ningbo China
Ningbo, 315100, China
Zhan.xu@nottingham.edu.cn

Xu Sun
University of Nottingham
Ningbo China
Ningbo, 315100, China
Xu.sun@nottingham.edu.cn

Abstract

The development of information technology has stimulated an increasing number of researchers to investigate how to provide serendipitous experiences to users in the digital environment, especially in those fields such as information research and recommender systems. Although a number of advances have been made in understanding the nature of serendipity in the area of information research, few of these have been employed in the design of recommender systems. This paper proposes a new serendipitous recommendation algorithm, based on the understanding of serendipity from the area of information research, by taking into consideration the three most important elements of serendipity, namely “unexpectedness”, “insight” and “value”. By applying this algorithm to an empirical study, we found that it successfully provided the participants with more possibilities to experience serendipitous encounters.

Author Keywords

Serendipity; recommender system; information theory.

ACM Classification Keywords

H.3.3 [Information Search and Retrieval]: Information filtering

Introduction

Serendipity is experienced widely, both in the humanities and the sciences. It is defined as "an unexpected experience prompted by an individual's valuable interaction with ideas, information, objects, or phenomena" [6]. So far, studies relating to serendipity focus mainly on the following two directions: the ongoing theoretical studies in the area of information research which aim to further investigate the nature of serendipity, and the empirical studies with the purpose of developing applications or algorithms that provide users with serendipitous encounters, especially in the digital environment.

One of the areas which is attempting to employ serendipity applications is the design of recommender systems. The overload of information in cyberspace has left current users no longer satisfied by simply recommending "accurate" information. Rather, users aim to receive recommendations about information which is more serendipitous and interesting. However, a rising concern which we identified based on our review of relevant studies, is that those discoveries from information research regarding the nature of serendipity have not received sufficient attention in recommender systems designs. This paper proposes a new algorithm to support serendipitous recommendations by combing the fruits of recent research into serendipity in the area of information research.

The Problem and the Research Question

Recommender systems researchers often consider serendipity as "unexpected" and "useful" [2], and they have designed recommendation algorithms through either content-based filtering [3] or collaborative filtering [7]. However, most of the recommendation algorithms focus mainly on providing "unexpectedness" to users, and have treated "usefulness" as purely a metric value with which to measure the effectiveness of their algorithms, rather than considering it as a design clue [1].

A possible reason for such a phenomenon is that "usefulness" is such a subjective conception which is largely dependent on personal requests. A potential solution to address the problem may arise from the study of serendipity in information research, where serendipity is considered to have three main characteristics: unexpectedness, insight and value [4]. The "unexpectedness" and "value" of serendipity is consistent with the current view of serendipity in recommender systems, however, the "insight" aspect tends to be neglected.

"Insight" is considered as the ability to find some clues in the current environment, then "making connections" between such clues and one's previous knowledge or experience, and finally shifting the attention to the newly discovered clue [10]. Supporting the process of "making connections" has always been considered to be an important element of any serendipitous design strategy [5].

Based on the above, we then raised our research question. Namely, is it possible to combine the theoretical studies of serendipity in information research, especially the aspects of "insights" or "making connections", into a recommender system design?

We proposed a collaborative-filtering based algorithm, taking into consideration the theoretical discoveries concerning serendipity in information research. Based on our discovery that serendipity is often encountered in a relaxed and leisure atmosphere [11], we applied the algorithm to a game-based application and conducted an empirical study. The results show that our algorithm is able to provide participants with serendipitous encounters.

Proposed Algorithm

There are two major concerns to consider when providing serendipitous encounters in a recommendation system design: the first concern is how to balance "unexpectedness" and "usefulness". As pointed out by [1], there should be "a most preferred distance" between the two values, as a high level of unexpectedness may cause a user's dissatisfaction with the recommended information, while users may also lose interest in such information with a low level of unexpectedness. The second concern is how to combine "insight" into a system design to stimulate the process of "making connections".

We have addressed the two concerns from the perspective of "relevance" with two hypotheses:

- Hypothesis 1: Given the information that is highly relevant to a user's personal profile, the information would also have a high potential value for the user;
- Hypothesis 2: Given the information is relevant to the user while it is not known by the user, user may feel unexpected of receiving the information.

Consider a target user A, who is the user who will be provided with the recommended information, a user B, who is highly relevant to user A, and a user C, who is highly relevant to user B but is not known by user A. User A may experience serendipity by providing information about user C, which is unexpected to him/her, and by introducing the relationship between user B and user C, which may be of further interest or usefulness to user A (Figure 1). The following part of this section illustrates a detailed implementation of the algorithm.

1. Prior given target user profile

Consider a table of a target user profile U1 whose category set C= $\{C_1, C_2, C_3 \ldots C_i \ldots C_n\}$, where C_i represents the i^{th} category of the user profile. All the categories are arranged through the value of their weights in the user profile. The weight can either be a given weight by the dataset or calculated through clustering analysis. To simplify the introduction of our proposed algorithm here, it is more convenient to consider that the weight for each Ci is given by the dataset at the very beginning. For each category set C_i, consider $C_i = \{a_1, a_2, a_3 \ldots a_i \ldots a_n\}$, where a_i is the corresponding attribute to each vector C_i. In particular, each a_i represents the dimension with the specific socialization characteristics according to which a new user profile may be produced (e.g. authors of literature; musicians). The values for each a_i are also arranged by their weight in each vector C_i and can be calculated through semantic analysis, such as the tf*idf weight calculation [8].

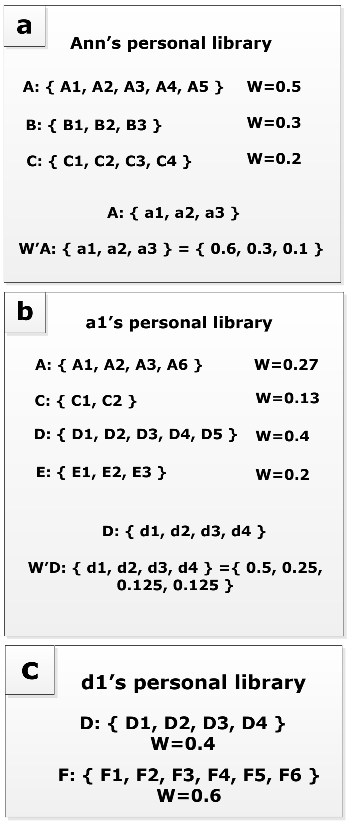

Figure 1: An example of the proposed algorithm: (a) target user Ann's personal library; (b) user a1's personal library generated by Ann; (c) user d1's personal library generated by a1.

2. Screen the weight

C_1 is the largest weight in the C set and a_1 is the largest weight in the C_i set. Set a threshold τ to eliminate the low weight value in the C_i set:

$$w_{U1} = \{w_{C_1}, w_{C_2}, w_{C_3}, \ldots, w_{C_i} \,|\, w_{C_i} \geq \tau\}$$

Similarly, set a threshold θ to eliminate the low weight value in the C_i item:

$$w_{C_i} = \{w_{C_i,a_1}, w_{C_i,a_2}, w_{C_i,a_3}, \ldots, w_{C_i,a_i} \,|\, w_{Ci,a_i} \geq \theta\}$$

3. Generate a new user profile

As each a_i represents a specific socialization characteristic in the C_i set, a new user profile can thus be produced. Here, the generation of the user profile arranges from the largest weight of w_{C_i,a_1} to the smallest weight of w_{C_i,a_i}.

4. Iteration and End condition

Based on the weight arrangement in a user profile, we intuitionally consider that for an attribute a_i with a large weight, it is also more possible for the current user to have acknowledged the information of a_i. In other words, the probability for a current user U_i to find the next user profile U_{i+1} is proportional to the weight of the attribute in the current user profile:

$$P(U_{i+1}|U_i) = \lambda w_{c_i} * w_{c_i,a_i}$$

where λ is the proportionality coefficient of the probability to the relevant weight.

The probability can be further extended to the i^{th} user to the target user U_1 if only the generated user is always new to the prior generated ones:

$$P(U_i|U_1) = P(U_2|U_1) * P(U_3|U_2) * \ldots * P(U_i|U_{i-1}) \quad (2)$$

The iteration to find the next user would not continue until it meets the following two end conditions:

- the generated user is no longer new to all the previous generated users;
- $P(U_i|U_1)$ comes to a threshold δ, where δ represents an appropriate threshold of the probability.

5. Recommendation

When the iteration is finished, the content with the largest weighted category in the current candidate will be provided to the target user, in addition to the relevant information of the previous searched users that resulted in the current user.

6. An example of the proposed algorithm

An example of the proposed algorithm is provided in Figure 1. Consider Ann as the target user (U1) with different literature categories of {A, B, C} in her personal library, whose weight is {0.5, 0.3, 0.2} (Figure 1-a). The authors' names of the literature are set as the attributes for each category and, according to the tf*idf weight calculation, there are three values {a1, a2, a3} in category A, where the weight W'A= {0.6, 0.3, 0.1}. Set λ=1 for each probability of the current user to find the next user profile, and the probability for Ann to find a1's profile (U2) can be calculated according to equation (3):

$$P(U_2|U_1) = w_A * w_{A,a_1} = 0.5 * 0.6 = 0.3$$

The profile of a1 is then produced as Figure 1-b. Likewise, among the four authors in the D category,

Figure 2: Given picture

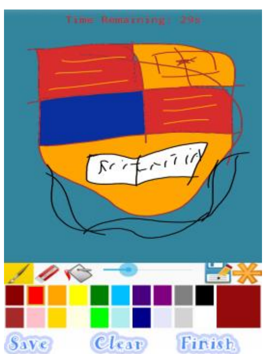

Figure 3: participant's drawing

author d1 (U3) weights largest and then produces d1's profile (Figure 1-c):

$$P(U_3|U_2) = w_D * w_{D,d_1} = 0.4 * 0.5 = 0.2$$

According to equation (3), the probability for Ann (U1) to find d1's profile (U3) is:

$$P(U_3|U_1) = P(U_2|U_1) * P(U_3|U_2) = 0.3 * 0.2 = 0.06$$

Set the threshold δ as 0.06, then the iteration of the algorithm stops and recommends literature of category F in d1's profile to Ann, in addition to the relevant information of d1 and a1. For example, the recommended information can be "these papers (category F) are most stored by d1, who have published papers (d1, d2, d3, d4) with a1".

Description of the Proposed Algorithm

We consider the proposed algorithm to be collaborative filtering based, and hence it is more appropriate to those datasets whose content is generated by different users, according to which it will be easier for a current user to produce the next user's profile.

The proposed algorithm relates with serendipity from the following three aspects:

Unexpectedness: by setting the value of probability $P(U_i|U_1)$. We consider the smaller the probability of the target user finding another user, the more unexpected the provided information of the current candidate would be to the target user.

Insight: As already mentioned, the ability to connect a new clue with previous knowledge/experience is also a key element in the occurrence of serendipity. Hence, we provided the related information of the current user (recommendation source) to the target users by demonstrating the searched clues between them.

Value: the value of the provided information is related to the potential needs/concerns of the user. Hence, we begin our algorithm with the element with the largest weight value. We consider the greater the weight the element possesses, the more potential value it may bring to the target user.

Empirical Study

According to the discoveries from our information research, serendipity is often experienced in a relaxed and leisure state [11], so we applied the algorithm to a game-based application and conducted an empirical study to investigate whether our proposed algorithm could provide serendipitous encounters to researchers. The study is described in detail below:

Participants

28 PhD students (14 males and 14 females) from different disciplines were invited to participate in the study. They were asked to conduct a drawing game on a mobile application which was developed by our research group.

Game-based Application Design

As all the participants were PhD researchers, the algorithm was designed based on three initial hypotheses:

- Hypothesis 1: For each PhD student, we considered their supervisor to have a large weighting attribute in their personal profile.

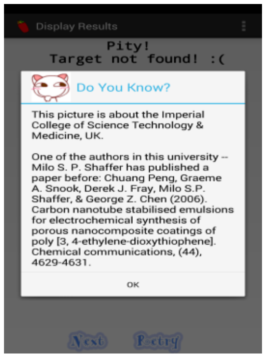

Figure 4: Provided related information

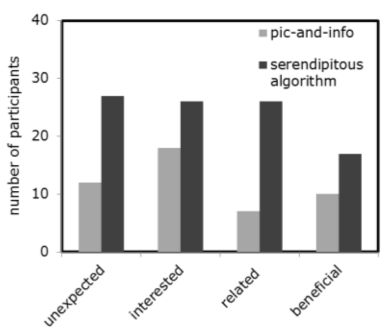

Figure 5: Questionnaire Result

- Hypothesis 2: For each PhD's supervisor, any co-authors for their publications were large weighted attributes in the supervisors' profile.
- Hypothesis 3: For each co-authors' personal profile, their working institution was a large weighted attribute.

Therefore, the aim of our study was to provide each PhD student with information about their supervisors' co-authors' institutions. We began our study by providing participants with pictures showing the logos of the relevant institutions (Figure 2). Participants were then asked to draw the picture within three minutes (Figure 3). At the end of the task, the related text information for the pictures was provided to the participants (Figure 4). The information given relating to the pictures included two levels: (1) the introduction of the institution; (2) the publications of both the participants' supervisors and the co-authors.

As a comparison, a traditional model of "picture with related information" (pic-and-info) was also included in the application. Two pictures from the website of Nature were chosen and provided to the participants, together with the related text information from the website. Each participant was required to finish drawing at least two pictures with our proposed algorithm and two with the traditional "pic-and-info" model.

Evaluation
Each participant was given a questionnaire with the four dimensions: "unexpected", "interested", "related" and "beneficial" Each dimension included a Likert scale ranging from 1 representing "not at all" to 5 representing "extremely". The design of the questionnaire was also based on the evaluation metrics from the information research [9].

In addition, a 15 minute post-interview was conducted immediately after each participant finished the experiment. The participants' subjective experience and further reasons for their ratings of the four dimensions were collected during the interview process.

Results

Questionnaire
In total, 20 effective questionnaires were selected from the 28 participants, because the remaining eight participants were too focused on the gameplay and failed to read the related information for the pictures. The questionnaires collected feedback for 40 pictures of the common "pic-and-info" model and a further 40 pictures based on our algorithm.

Only the marks of four or five were considered to be effective values on the corresponding dimension, which is shown in Figure 5. It is obvious that compared with the conventional "pic-and-info" model, more participants felt the information in the pictures based on our algorithm was more unexpected, interesting, related and beneficial, which indicates the greater potential possibility for the participants to encounter serendipity.

Interview
During the interview process, most participants reported a sense of serendipity from the following two perspectives:

(1) The unexpected relationship between the picture and the information provided.

I've never thought the picture is related to my supervisor. The information really surprised me and I really think this is a very good design to provide me with the information in such a context!

(2) The benefit gained from the information provided.

I never knew that my supervisor had published such a paper with him (the co-author) before…… I'll check the details of the paper later.

The feedback from the participants demonstrates that our proposed algorithm effectively supports the design strategies for experiencing serendipity.

Conclusion and Future Work

In this paper, we have presented a new serendipitous recommendation algorithm based on our findings from information research. In particular, different from current designs which focus on the property of "unexpectedness", our proposal extends serendipitous recommendation algorithm by including two other vital aspects of serendipity - i.e., "insight" and "value".

We also carried out an empirical study among researchers based on the proposed algorithm, and the results demonstrate that compared to the conventional "pic-and-info" design, the application of our algorithm has effectively resulted in serendipitous encounters among our participants.

However, the study is limited by the small sample number of participants, so our future work will aim to explore the algorithm through more datasets. We will also compare our proposed algorithm with other existing algorithms so as to better evaluate and optimize the algorithm in different situations.

References

1. Adamopoulos, P. and A. Tuzhilin, On unexpectedness in recommender systems: Or how to better expect the unexpected. ACM Transactions on Intelligent Systems and Technology, 2014. 5(4).
2. Ge, M., C. Delgado-Battenfeld, and D. Jannach. *Beyond accuracy: evaluating recommender systems by coverage and serendipity*. in *Proceedings of the fourth ACM conference on Recommender systems*. 2010. ACM.
3. Jenders, M., et al., *A Serendipity Model for News Recommendation*, in *KI 2015: Advances in Artificial Intelligence*. 2015, Springer. p. 111-123.
4. Makri, S. and A. Blandford, *Coming across information serendipitously – Part 1.* Journal of Documentation, 2012. **68**(5): p. 684-705.
5. Makri, S., et al., *"Making my own luck": Serendipity strategies and how to support them in digital information environments.* Journal of the Association for Information Science and Technology, 2014. **65**(11): p. 2179-2194.
6. McCay-Peet, L. and E.G. Toms, *Investigating serendipity: How it unfolds and what may influence it.* Journal of the Association for Information Science and Technology, 2015. **66**(7): p. 1463-1476.
7. Oku, K., & Hattori, F. (2011, October). Fusion-based recommender system for improving serendipity. In Proceedings of the Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011), at the 5th ACM International Conference on Recommender Systems (RecSys 2011) (p. 19).

8. Pazzani M J, Billsus D. Content-based recommendation systems[M]//The adaptive web. Springer Berlin Heidelberg, 2007: 325-341.

9. Pontis, S., Kefalidou, G., Blandford, A., Forth, J., Makri, S., Sharples, S., ... & Woods, M. (2015). Academics' responses to encountered information: Context matters. Journal of the Association for Information Science and Technology.

10. Rubin, V.L., J. Burkell, and A. Quan-Haase, *Facets of serendipity in everyday chance encounters: A grounded theory approach to blog analysis.* Information Research, 2011. **16**(3).

11. Sun, X., S. Sharples, and S. Makri, *A user-centred mobile diary study approach to understanding serendipity in information research.* Information Research, 2011. **16**(3).