Early Findings from a Large-scale User Study of CHESTNUT: Validations and Implications

Xiangjun Peng[1], Zhentao Huang[1], Chen Yang[2], Zilin Song[1] and Xu Sun[1]

[1] User-Centric Computing Group, University of Nottingham Ningbo China
[2] Department of Computer Science, Syracuse University, Syracuse, United States
xu.sun@nottingham.edu.cn

Abstract. Towards a serendipitous recommender system with user-centred understanding, we have built ***CHESTNUT***, an Information Theory-based Movie Recommender System, which introduced a more comprehensive understanding of the concept. Although off-line evaluations have already demonstrated that ***CHESTNUT*** has greatly improved serendipity performance, feedback on ***CHESTNUT*** from real-world users through online services are still unclear now. In order to evaluate how serendipitous results could be delivered by ***CHESTNUT***, we consequently designed, organized and conducted large-scale user study, which involved 104 participants from 10 campuses in 3 countries. Our preliminary feedback has shown that, compared with mainstream collaborative filtering techniques, though ***CHESTNUT*** limited users' feelings of **unexpectedness** to some extent, it showed significant improvement in their feelings about certain metrics being both **beneficial** and **interesting**, which substantially increased users' experience of serendipity. Based on them, we have summarized three key takeaways, which would be beneficial for further designs and engineering of serendipitous recommender systems, from our perspective. All details of our large-scale user study could be found at https://github.com/unnc-idl-ucc/Early-Lessons-From-CHESTNUT

Keywords: Serendipity, Recommeder Systems, User Study

1 Introduction

Towards a more comprehensive understanding of serendipity, we have built ***CHESTNUT***, the first serendipitous movie recommender system with an Information Theory-based algorithm, to embed a more comprehensive understanding of serendipity in a practical recommender system [16, 24]. Although experimental studies on static data sets have shown that ***CHESTNUT*** could achieve significant improvements (i.e. around 2.5x), compared with other mainstream collaborative filtering approaches, in the incidence of serendipity, it remains necessary for a user study to be conducted to allow validations of ***CHESTNUT*** and impose further investigations into the concept of serendipity and the engineering of serendipitous recommender systems.

Therefore, we carried out a large-scale user study around ***CHESTNUT***, along with its experimental benchmark systems. To enable a detailed study, we first designed a plan to ensure all participants were capable of experiencing serendipity, by excluding the effects of any environmental factors as much as possible. We then launched the study, and invited 104 participants to contribute, from whom we collected extensive data and qualitative records from real-world users over a ten-month period.

Our initial results indicate that, although ***CHESTNUT*** limited users' feelings of "unexpectedness", when compared with item-based and user-based collaborative filtering approaches, it did show significant improvement in users' feelings about certain metrics being both "beneficial" and "interesting", which substantially increased their experience on serendipity. The low quantity of "unexpectedness", through our interviews, have been addressed due to relatively old movies from ***CHESTNUT***.

Based on these preliminary statistics and context-based investigations, we summarized three key takeaways for future work, which lied on **the Design Principles of User Interfaces**, **Novel Integration of More Content-based Approaches** and **Introspection of Serendipity Metrics**. We believe they are extremely useful for further designs and engineering of serendipitous recommender systems.

More specifically, we have made three main contributions here:

(1) A Large-scale User Study among *CHESTNUT* and two mainstream Collaborative Filtering Systems. We have performed a large-scale user study among ***CHESTNUT***, Item-based and User-based Collaborative Filtering approaches. Our study has lasted for around 10 months, which involved 104 participants across 3 countries. All details of our large-scale user study could be found at https://github.com/unnc-idl-ucc/Early-Lessons-From-CHESTNUT

(2) Validations and Implications of the Improvements from *CHESTNUT* in Serendipity. Through this study, we have validated the effectiveness of ***CHESTNUT*** in terms of serendipitous recommendations, compared with widely commercialized algorithms. Our initial results also indicates some limitations of our current end-to-end prototype, which has limited the performance of ***CHESTNUT***.

(3) Takeaways for Principles of Designs, Developments and Evaluations in Engineering of Serendipitous Recommender Systems. Based on several implications from this study, we have summarized three key takeaways, as potential future work directions, to discuss about future principles of designs, developments and evaluations for serendipitous Recommender Systems.

This paper would be organized as follow. Section 2 would provide necessary background information and illustrate our motivation of this study. Section 3 would introduce details around this study, spanning from methodology to technical adjustments. Section 4 would report our initial results and relevant analysis from this study. Section 5 would present our discussion and introspection to motivate and stimulate potential principles and follow-up work in the future.

2 Background and Motivation

For a decade, serendipity has been understood narrowly within the Recommender System field, and it has been defined in previous research as receiving an unexpected and fortuitous item recommendation [13]. Such mindset have led to many efforts in the development and investigation of serendipitous recommender systems through modelling and algorithmic designs and optimizations, instead of rethinking the natural understanding of the concept. [1, 2, 5, 6, 3, 8–10, 12, 11, 14, 15, 17, 18, 20–22].

CHESTNUT was built to validate a novel insight around serendipitous recommender system, by merging **insight**, **unexpectedness** and **usefulness** to provide a more comprehensive understanding of serendipity, as the first user-centered serendipitous recommender system. In the context of movie recommendation, ***CHESTNUT*** enables connection-making between users through their directors' information (**cInsigt**), filter out non-popular and non-familiar movies (**cUnexpectedness**) and then generate recommendations through rating prediction (**cUsefulness**). The above three steps ensured relevance, unexpectedness and values respectively.

Although the theoretical support of ***CHESTNUT*** [24], its effectiveness [23] and practical system performance [16] has been examined earlier, **the missing validation from real-world users is still missing**. Also, a large-scale user study would also help to uncover several issues, which are not capable to be found through off-line evaluations, and enhance its practicality. Therefore, we have performed a large-scale user study since we believe such a study is essential, important and meaningful for both ***CHESTNUT***-related work and the communities of Recommender Systems and Information Management.

3 Methodology

In this section, we introduce the research methods used in the ***CHESTNUT*** user study. Other than environmental factors, previous user studies of serendipity has pointed out that users' willingness to participate would undoubtedly affect their serendipitous experiences [23]. To allow us to collect satisfactory feedback, we scheduled face-to-face interviews as suggested by participants. However, we were unable to manage all interviews in this way, owing to geographical limitations. For those who couldn't attend in person, we applied a mobile diary method to record relevant details, which was a systematic method used in previous user studies on serendipity [19, 23].

3.1 Participants

In total, 104 undergraduate students were invited to take part in this user study, with each participant having made at least 30 movies' ratings. Although a previous study invited professional scholars to take part in serendipity interviews

Table 1. Geographical Distribution of Participants

Affiliation	Number of Participants	Country
Campus 1	79	China/United Kingdom
Campus 2	13	United States
Campus 3	4	China
Campus 4	2	China
Other Campuses	6	China

Table 2. Personal Information Collection from Participants

Affiliation	Average Age	Male-to-Female Ratio
Campus 1	19.468	33:46
Campus 2	20.769	5:8
Campus 3	18.750	3:1
Campus 4	20.000	1:1
Other Campuses	19.000	5:1
All Campuses	19.586	47:57

Table 3. Levels of Involvements from Participants

Affiliation	Average Number of Co-rated Items
Campus 1	33.363
Campus 2	22.411
Campus 3	24.250
Campus 4	41.500
Other Campuses	29.333
All Campuses	32.362

(i.e. because their speciality made experiencing serendipity easier), this study aimed to investigate serendipity within a more generalized group [23].

Details about all participants' geographical distribution are reported in Table 1. Details about all participants' personal information are reported in Table 2. Details about the levels of involvements from participants are reported in Table 2. All participants' names reported in this study are aliases.

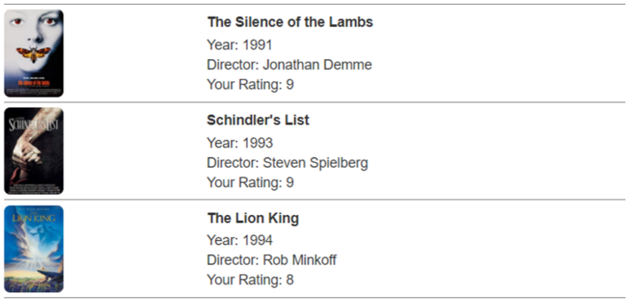
There are two things to be illustrated in Table 1: 1) The term **Countries** refers to those countries which the corresponding campus bases on; 2) The term **Other Campuses** refers to those campuses, which only has one participant.

3.2 Procedure

Before the bulk of this study began, a pilot study was performed with two male participants on campus for a period of four days. The detailed experiment issues such as time arrangement, system functionality and interaction preparation, were all decided based on this pilot study.

The bulk of this study was then conducted. For each participant, there were two parts to the whole study - a pre-interviews and an empirical interviews. First, for the pre-interviews, each participant was invited individually to a short meeting (around 30 minutes), to introduce the purpose of study and to collect their own movie rating records. Since the majority of users do not use IMDb as their channel with which to manage their favorite movies, we had to perform this collection procedure within the interview, which also meant that participants could prepare in advance.

Second, for the empirical interviews, each participant made their own schedule in advance. During this stage, users were able to view their recommendation results from the website. Within the time period between the two interviews, we processed collected user profiles in ***CHESTNUT***, and placed their recommendation results online. During the interviews, participants were guided, both by the researchers and the system, to review their profiles, and to check and comment on their recommendation results step by step. We have sketched the interfaces in Figures 1 and 2.

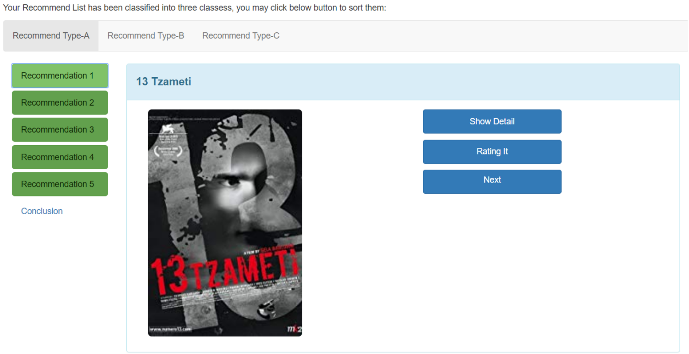


The Silence of the Lambs
Year: 1991
Director: Jonathan Demme
Your Rating: 9

Schindler's List
Year: 1993
Director: Steven Spielberg
Your Rating: 9

The Lion King
Year: 1994
Director: Rob Minkoff
Your Rating: 8

Fig. 1. User Profile Review Page



Your Recommend List has been classified into three classess, you may click below button to sort them:

Fig. 2. User Ratings and Comment Page

This user study has lasted for ten months because we followed the "user-centred" arrangements, and eac schedule was determined by the participants involved. In addition to ***CHESTNUT***, we also set up two benchmark systems (i.e. an item-based and a user-based approach) and included their results together within this study. Each system produced five recommendation results, according to their submitted profiles. Since not every user could attend a face-to-face interview, we had to perform interviews via online applications (e.g. Wechat), where necessary.

3.3 Data Collection

Two types of data were collected: 1) the user experiences from all recommended movies, generated from all three systems. For each movie, the participants were able to rate feelings on whether the information they were presented with was "unexpected", "Interesting" and "Beneficial", according to the scale shown in Table 2. 2) under each page, the users also had the option to leave their comments on any of the movies, if they felt it was necessary.

Table 4. Geographical Distribution of Participants

Rating	Context
5	Extremely
4	Quite A Bit
3	Moderately
2	A Little
1	Not At All

3.4 System Modifications

Previous studies on ***CHESTNUT*** applied the relatively small HeteRec 2011 data set from Movielens (i.e. 2113 users with 800,000 ratings) [4]. Since ***CHESTNUT*** is a memory-based collaborative filtering system, its baseline data set had to be expanded to adapt the system into a real-world scenario. We chose the most recent 30 years' movies and related ratings from ml-20m [7], which is one of the latest and largest data set from Movielens (i.e. 138471 users with 150,000,000 ratings).

4 Results

In this section, we will introduce the preliminary results and analysis, drawn from our user study. After collecting all feedback, we performed a series of preliminary analysises to examine the effectiveness of ***CHESTNUT***. First, we will give a performance overview of ***CHESTNUT*** and its benchmark systems, which relied on data collected from online questionnaires (i.e. to sketch the level of users' feelings under the different metrics). Next, we will sketch out our perspectives as preliminary hypothesis, which will entail additional investigations and discussions around both serendipity and ***CHESTNUT***.

4.1 Performance Overview

Our performance overview is divided into three parts, as arranged in the online questionnaires. We will examine the rating levels for whether the participants thought the information provided was "unexpected", "interesting" and "beneficial" separately. This was done for ***CHESTNUT***, and for the item-based and user-based systems respectively.

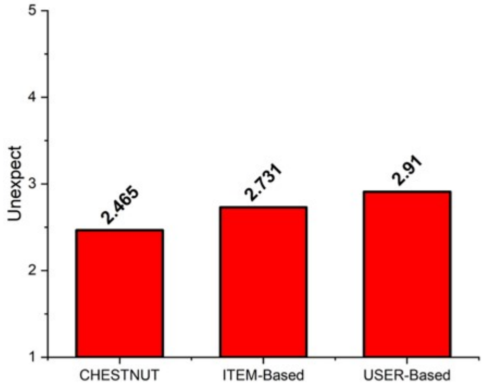


Fig. 3. Average Rating of "unexpectedness"

First, we examined all participants' levels of "unexpected" feelings towards their results, as shown in Figure 3. On the one hand, ***CHESTNUT*** performed the worst with regard to users' feelings of unexpectedness, with the level only reaching 2.465 on average. On the other hand, the user-based and item-based approaches achieved better ratings for this factor, with levels of 2.731 and 2.91 on average respectively.

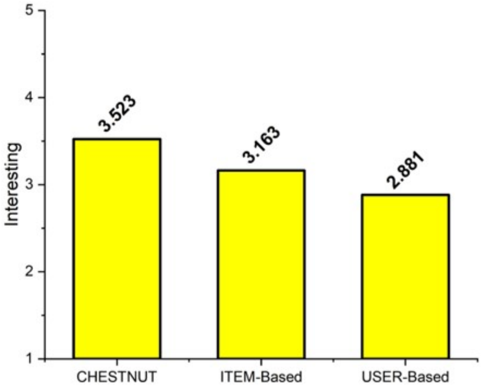


Fig. 4. Average Rating of "interesting"

We then explored the levels of feelings on relating to whether they found the information to be "interesting", and the results are shown in Figure 4. In terms of providing interesting recommendation results, ***CHESTNUT*** achieved 3.523 on average. As for the user-based and item-based approaches, they only achieved average ratings of 3.163 and 2.881 respectively. The results support the fact that ***CHESTNUT*** is able to provide more interesting recommendation results.

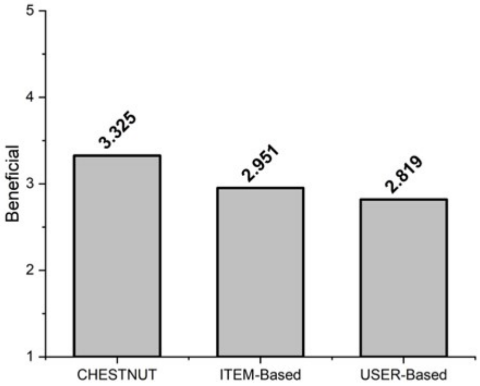


Fig. 5. Average Rating of "beneficial"

Finally, we checked quantitatively the levels of feelings relating to whether they found the information to be ”beneficial”, as illustrated in Figure 5. In this case, ***CHESTNUT*** achieved a rating of 3.325 on average, but the user-based and item-based approaches only reached 2.951 and 2.819 on average respectively. Similar to the results for how ”interesting” the participants found the information, the results support the fact that ***CHESTNUT*** is able to provide much more beneficial recommendation results than the two conventional approaches.

4.2 Why Were the Results Not So ”Unexpected”?

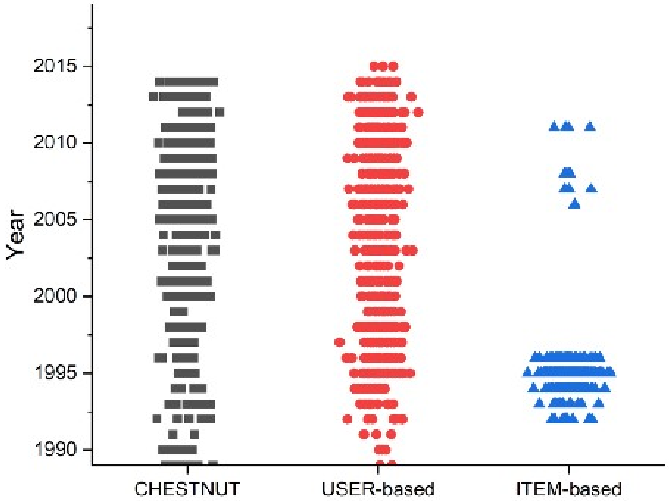


Fig. 6. Year Distribution of Recommendation Results

The major concern of our user study is that results from ***CHESTNUT*** were not as highly rated in terms of their unexpectedness as we predicted. Normally, ***CHESTNUT*** has a particular functional unit (i.e. ”cUnexpectedness”) to ensure that all results are unexpected. However, during this study, ***CHESTNUT*** seemed to fail to make users feel that the results were unexpected.

Having looked further into this phenomenon, we found that, as participants claimed, they felt the results were unexpected when they encountered relatively old movies (i.e. those made approximately 20 years ago), because most only kept track of more recent productions. We further confirmed this by drawing the year distribution of recommendation results from different systems, as shown in Figure 6. We believed this is the main reason why they outperformed than ***CHESTNUT*** in their ”unexpected” ratings.

5 Discussions and Takeaways

Based on preliminary results and analysis from our study, we hereby discuss revealed issues and relevant takeaways, to stimulate novel insights and follow-up investigations. In general, there are three aspects, which we want to highlight:

- **Design Principles of User Interfaces.**
  The current User Interface design of ***CHESTNUT*** directly reflects high-level overview information of different items, which has indicated that general design choices could potentially hurt users' capability to encounter serendipitous information. Particularly, in our case, the tag "year" has led to a lot of negative effects in the domain of "unexpectedness".

  We believe future serendipity-oriented interfaces demands content-based interaction and personalized mechanisms. For instance, in the context of movies, we could use Movie Trailer as preview resources, and let users customize their interfaces by re-ordering the priority of displayed information as they prefer.

- **Novel Integration of More Content-based Approaches.**
  The current study of ***CHESTNUT*** has only been exploited with the setting of "Director", as the connection-making resources. Given the fact that Directors are dependent to active periods, levels of productivity and genres, it's reasonable to lead to "old movies appear frequently".

  We believe future studies among ***CHESTNUT*** and other serendipitous recommender systems would take the categories of information, as the guiding resources, into account. In ***CHESTNUT***, we have already included support for other information categories in **cInsight**, such as Years and Genres. This would also impose novel integration of different content-based approaches together, which aims to provide personalized recommendation results.

- **Introspection of Serendipity Metrics.**
  The comparison between our real-world study and off-line evaluations have indicated that, there is a huge gap of current serendipity measures in the context of Recommender Systems.

  We believe this also imposes a lot of opportunities to develop novel schemes and frameworks for the validations of serendipitous designs and implementations. More specifically, the results from our experimental study and real-world feedback around ***CHESTNUT*** are extremely valuable, especially when combining both of them together.

Although we have addressed three aspects of our takeaways, we still believe there are a lot of challenges and opportunities beyond ***CHESTNUT***. Hereby, we only provide reflections from our own experiences and we hope they would stimulate more interesting and novel ideas for serendipitous designs and engineering, in both Recommender Systems and other relevant communities.

6 Conclusion and Future Work

In this paper, we have presented our early findings from a large-scale user study of ***CHESTNUT***, which involved 104 participants over a ten months period. According to our initial analysis, the results have shown that, compared with mainstream collaborative filtering techniques, though ***CHESTNUT*** limited users' feelings of **unexpectedness** to some extent, it showed significant improvement in their feelings about certain metrics being both **beneficial** and **interesting**, which substantially increased users' experience of serendipity. Based on them, we have summarized three key takeaways, which would be beneficial for further designs and engineering of serendipitous recommender systems.

Our future work will make variants of further in-depth studies, based on our summarized takeaways, to investigate both the concept of serendipity and the optimizations of ***CHESTNUT*** through these empirical data. Beyond ***CHESTNUT***, we are particularly interested in bridging real-world feedback and off-line results for more sophisticated frameworks, to further validate many variants of other serendipitous recommender system designs and implementations.

7 Acknowledgement

We thank for valuable feedback and suggestions from our group members and anonymous reviewers, which have substantially improved the overall quality of this paper. In particular, we thank for our ex-member, Weikai Kong, for his great efforts in setting up and assisting to perform the bulk of this study. This research is generously supported by National Natural Science Foundation of China Grant No. 71301085 and Hefeng Creative Industrial Park in Ningbo, China.

References

1. Zeinab Abbassi, Sihem Amer-Yahia, Laks V. S. Lakshmanan, Sergei Vassilvitskii, and Cong Yu. Getting recommender systems to think outside the box. In *Proceedings of the 2009 ACM Conference on Recommender Systems, RecSys 2009, New York, NY, USA, October 23-25, 2009*, pages 285–288, 2009.
2. Panagiotis Adamopoulos and Alexander Tuzhilin. On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM TIST*, 5(4):54:1–54:32, 2014.
3. Upasna Bhandari, Kazunari Sugiyama, Anindya Datta, and Rajni Jindal. Serendipitous recommendation for mobile apps using item-item similarity graph. In *Information Retrieval Technology - 9th Asia Information Retrieval Societies Conference, AIRS 2013, Singapore, December 9-11, 2013. Proceedings*, pages 440–451, 2013.
4. Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011, New York, NY, USA, 2011. ACM.

5. Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Cataldo Musto. An investigation on the serendipity problem in recommender systems. *Inf. Process. Manage.*, 51(5):695–717, 2015.
6. Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010*, pages 257–260, 2010.
7. F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *TiiS*, 5(4):19:1–19:19, 2016.
8. Hiroaki Ito, Tomohiro Yoshikawa, and Takeshi Furuhashi. A study on improvement of serendipity in item-based collaborative filtering using association rule. In *IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2014, Beijing, China, July 6-11, 2014*, pages 977–981, 2014.
9. Junzo Kamahara, Tomofumi Asakawa, Shinji Shimojo, and Hideo Miyahara. A community-based recommendation system to reveal unexpected interests. In *11th International Conference on Multi Media Modeling (MMM 2005), 12-14 January 2005, Melbourne, Australia*, pages 433–438, 2005.
10. Noriaki Kawamae. Serendipitous recommendations via innovators. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*, pages 218–225, 2010.
11. Kibeom Lee and Kyogu Lee. Using experts among users for novel movie recommendations. *JCSE*, 7(1):21–29, 2013.
12. Kibeom Lee and Kyogu Lee. Escaping your comfort zone: A graph-based recommender system for finding novel recommendations among relevant items. *Expert Syst. Appl.*, 42(10):4851–4858, 2015.
13. Sean M. McNee, John Riedl, and Joseph A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *Extended Abstracts Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006, Montréal, Québec, Canada, April 22-27, 2006*, pages 1097–1101, 2006.
14. Kenta Oku and Fumio Hattori. Fusion-based recommender system for improving serendipity. In *Proceedings of the Workshop on Novelty and Diversity in Recommender Systems, DiveRS 2011, at the 5^{th} ACM International Conference on Recommender Systems, RecSys 2011, Chicago, Illinois, USA, 23 October 2011*, pages 19–26, 2011.
15. Kensuke Onuma, Hanghang Tong, and Christos Faloutsos. TANGENT: a novel, 'surprise me', recommendation algorithm. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 657–666, 2009.
16. Xiangjun Peng, Hongzhi Zhang, Xiaosong Zhou, Shuolei Wang, Xu Sun, and Qingfeng Wang. CHESTNUT: Improve Serendipity Performance in Movie Recommendation by an Information Theory-based Collaborative Filtering Approach. In *Proceedings of the 22th Springer International Conference on Human-Computer Interaction (HCI'20)*, 2020.
17. Markus Schedl, David Hauger, and Dominik Schnitzer. A model for serendipitous music retrieval. In *Proceedings of the 2nd Workshop on Context-awareness in Retrieval and Recommendation, CaRR 12, 2012, Lisbon, Portugal, February 14-17, 2012*, pages 10–13, 2012.
18. Giovanni Semeraro, Pasquale Lops, Marco de Gemmis, Cataldo Musto, and Fedelucio Narducci. A folksonomy-based recommender system for personalized access to digital artworks. *JOCCH*, 5(3):11:1–11:22, 2012.

19. Xu Sun, Sarah Sharples, and Stephann Makri. A user-centred mobile diary study approach to understanding serendipity in information research. *Inf. Res.*, 16(3), 2011.
20. Maria Taramigkou, Efthimios Bothos, Konstantinos Christidis, Dimitris Apostolou, and Gregoris Mentzas. Escape the bubble: guided exploration of music preferences for serendipity and novelty. In *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*, pages 335–338, 2013.
21. Hisaaki Yamaba, Michihito Tanoue, Kayoko Takatsuka, Naonobu Okazaki, and Shigeyuki Tomita. On a serendipity-oriented recommender system based on folksonomy and its evaluation. In *17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, KES 2013, Kitakyushu, Japan, 9-11 September 2013*, pages 276–284, 2013.
22. Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. Auralist: introducing serendipity into music recommendation. In *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012*, pages 13–22, 2012.
23. Xiaosong Zhou, Xu Sun, Qingfeng Wang, and Sarah Sharple. A context-based study of serendipity in information research among chinese scholars. *Journal of Documentation*, 74(3):526–551, 2018.
24. Xiaosong Zhou, Zhan Xu, Xu Sun, and Qingfeng Wang. A new information theory-based serendipitous algorithm design. In *Proceedings of the 19th Springer International Conference on Human-Computer Interaction (HCI'17)*, pages 314–327, 2017.