Research Article

Concept-based Recommendation System for Finding Serendipity

Kodai Tsukahara[1], Eiji Kamioka[2], Phan Xuan Tan[3]

[1,2,3,]Shibaura Institute of Technology
ma19055@shibaura-it.ac.jp[1], kamioka@shibaura-it.ac.jp2, tanpx@shibaura-it.ac.jp[3]

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Current information recommendation systems obtain users' preferences from Web browsing histories and activities such as purchase of products, and efficiently provide the users with their preferable information. In such a case, however, the same or similar information is always recommended, which is called filter bubble and it decreases the users' satisfaction to the systems. If information recommendation systems could provide users with something surprising and useful as output information, the user's satisfaction to the systems would drastically increase. Therefore, "serendipity" is paid attention to in this research. In this paper, a new information recommendation system using a concept-based information retrieval is proposed to provide the users with serendipitous information. In this system, concepts which describe features or roles of items are input instead of the items themselves, and information which can meet the concepts are output as candidates of serendipitous information. The serendipitous information is extracted from the output information using the criteria which are the indexes of serendipity defined in this research. Through the evaluation experiment, it is revealed that the proposed system achieves the accuracy of 70% for the serendipitous information determination and the accuracy of 100% for the information retrieval, which are satisfactory for this research purpose.

Keywords: serendipity, recommendation system, relevance, unexpectedness, usefulness

1. Introduction

In the current information society, recommendation systems are widely utilized to efficiently provide the users with preferable information extracted from a huge amount of resources. In a typical recommendation system, the users' preferences are obtained from their Web browsing histories and activities such as purchase of products. Based on the preferences, useful information for the users are recommended.

Pariser (2011) and Jauhar (2015) stated that although recommendation systems can recommend favorite information for the users, those information are unnecessary in most cases. This is because such recommended information is always similar to the ones recommended before. In addition, if only preferable information is recommended to the users, it causes a filter bubble problem in which information in other categories cannot be recommended.

In this paper, an approach to easily obtain serendipitous information is discussed. For this purpose, the definition of serendipity must be clarified. Serendipity has its origin in a fairy tale, the name of which is "The Three Princes of Serendip". A writer, Horace Walpole, created the name of "serendipity" as a concept from a story of the fairy tale, which is "The three princes discover what they were not originally looking for from events that happened by chance with their intelligence" (Foster and Ford (2003).

Oku and Hattori (2013) stated that users can easily obtain serendipitous information using an interactive approach.Moreover, Adamopoulos and Tuzhilin (2015) stated that the information located in the middle of the expected information and the irrelevant information is serendipitous information. Hence, in this paper, it is assumed that actively performing information retrieval on such "unexpected and irrelevant information" would lead to the serendipitous information. To support such an action, a new information recommendation system using a concept-based information retrieval is proposed. In this system, the users input concepts (features) of an item that they want to search for, and then serendipitous information will be output. In this paper, the approach to realize the proposed system is thoroughly discussed and the performance to output serendipitous information is also evaluated.

The rest of the paper is organized as follows: Section 2 describes related works on serendipity and section 3 describes the proposed system and how to acquire serendipitous information. Section 4 evaluates the proposed framework based on experiments. Finally, section 5 summarizes this paper.

2. Related Works

In this section, the definition of serendipity in this paper is determined with reference to the related works.

Herlocker et al. (2004) stated that even if a recommenation system recommends information in a genre that a user likes, this recommended information will be meaningless to the user if it has already been evaluated by the user. To solve this problem, Herlocker et al. stated that "novelty" and "serendipity" should be taken into consideration for recommendation systems. Novelty is an index for evaluating whether a user has known the recommended information before or not. On the other hand, serendipity is an index for measuring how unexpected the recommended information is as well as how attractive it is for a user. It is difficult to recommend useful information to users only with novelty. Furthermore, the definition of serendipity is complicated, thus needs to be redefined more simply.

Adamopoulos and Tuzhilin (2015) stated that even though a recommended information is completely different from the one that a user has, it does not always lead to the recommendation of serendipitous information regarding the novelty and surprise in existing systems. Furthermore, the further away from the expected level the obtained information is, the more unexpected such a recommendation becomes until it is recognized as irrelevant to the user. In other words, serendipitous information exists between unexpected information and irrevant information, and recommendation systems need to obtain this. However, the details on the user's prediction were not clear in their paper.

Oku and Hattori (2013) proposed a fusion-based recommendation system that generates intrinsic chances. In this system, the users can actively operate the system and find serendipitous books. However, this system cannot be utilized effectively unless the users have a sufficient motivation to find serendipitous books.However, an interactive approach like their system could stimulate users' inquisitiveness and make it easier to find serendipity.

Referring to abovementioned related works, serendipity in this paper is defined as "unexpected but useful information for users". However, "novelty" is not discussed in this study. This is because serendipitous information does not always exist in the information that the user does not have. On the other hand, irrelevant information to the information that the user expects can be regarded as unexpected information for the user, leading to serendipity. Therefore, "relevance" is introduced instead of "novelty" in this study. Hence, the indexes of serendipity in this paper are "relevance", "unexpectedness" and"usefulness".

3. Research Methodology

In this section, a system, which provides the users with serendipitous items from concepts of a word, is proposed. In addition, the index of "similarity" is introduced to define the relevance between two words.

The purpose to develop the proposed system is to extract serendipitous information when someone wants to perform information retrieval. Figure 1 shows how the proposed system works. For instance, suppose that a user wants to get information of an item. Then, the user gives the system some words which are relevant to the item as concepts. In this case, "X is used for study", "X is at library" and "X has knowledge" can be the concepts, and "study", 'library" and "knowledge" are the relevant words to the item in each concept, respectively. Note that as known well, the concepts used above can be described as "UsedFor", "AtLocation" and "HasA". The system searches for several words which correspond to X in each concept, and finally outputs the words which meet all the concepts among them. The output words could be "book", "newspaper" or "Internet" in this case. One of these can be the expected item. At the same time, there is a possibility that some unexpected words can also meet all the concepts, which may be serendipitous information. To realize this system, a database, which shows the relation between words and multiple concepts, is needed. In this system, ConceptNet (2004) is used as the database.

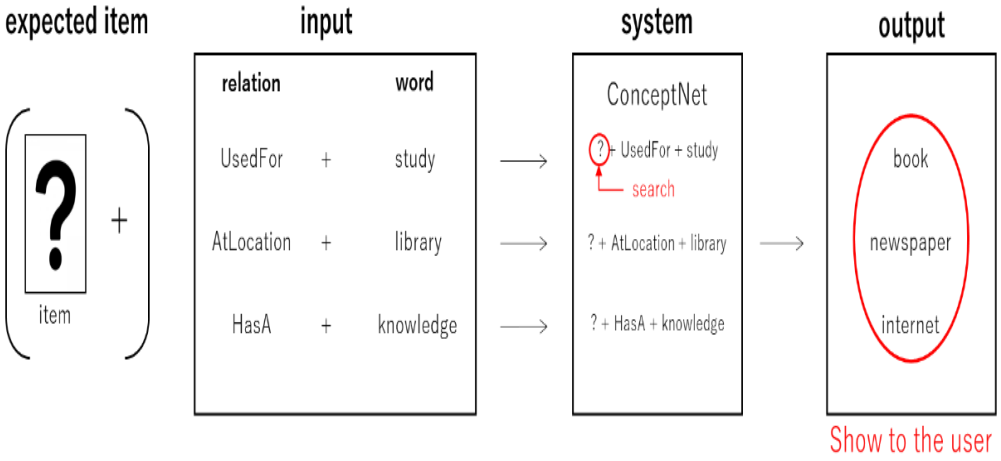

Figure 1. Overview of the proposed system

The most important thing here is to clarify if the output words which meet all the concepts are serendipitous information or not. To do that, the indexes of serendipity mentioned in section 1, that is to say, "relevance", "unexpectedness" and "usefulness" are utilized. Regarding "relevance" and "unexpectedness", "similarity" between the expected item and the output word can be evaluated. Note that "similarity" can be obtained using ConceptNet. It seems that there is a strong negative correlation between "relevance" and "unexpectedness". However, "relevance" mainly depends on the knowledge of words, and "unexpectedness" depends on the user's experience. In addition, "usefulness" varies from person to person. Therefore, subjective evaluations are also performed to these indexes in this paper.

In this study, Concept Net Numberbuch (CNN), which is an extended database of ConceptNet, is utilized to get "similarity". CNN consists of a set of semantic vectors that can be used directly, representing the meaning of words. It evaluates the strength of connection between words from the common-sense viewpoint based on the semi-structured common-sense knowledge of ConceptNet. It refers to not only the data of ConceptNet but also the data of other tools used for the same purpose. In this paper, the "similarity" between the expected item and the output word is calculated using CNN, and the correlation between the value of similarity and the subjective evaluation score of each index ("relevance", "unexpectedness" and "usefulness") is investigated.

4. Evaluation

In this section, the effectiveness and the accuracy of the proposed system is evaluated through an experiment based on "similarity" and subjective evaluation. The details of the evaluation method and the results are described in the following subsections.

4.1 Experimental methods

In the evaluation experiment, the proposed system was utilized by 10 subjects (male: 9, female: 1) who were students of the university to which the authors belong. The average of the subjects was in the middle of twenties. None of the subjects had any knowledge of the proposed system. The experimental procedure is as follows:

(1) Subject decides a word, which is searched for and regarded as the "expected item". Then, the subject describes it as its concept with a relevant word, as explained in section 3.1. One "expected item" is described by at least three types of concepts.
(2) The subject gives the proposed system these relevant words as elements of its concepts, and then, evaluates the words output from the proposed system by a 6-point MOS scale based on the indexes of serendipity in this study, which are "relevance", "unexpectedness" and "usefulness".
(3) The subject also evaluates if the output words can be regarded as serendipitous information or not (Yes/No evaluation).

Note that MOS used in the above (2) was the average score obtained from ten subjects: one subject who actually performed the experiment with the expected item, and other nine subjects.)

4.2 Experimental results

In this section, the results of evaluation experiment are discussed.

4.2.1 Correlation between similarity and indexes of serendipity

Figure 2 shows the correlation between "similarity" and "relevance". On the other hand, Fig. 3 shows the correlation between "similarity" and "unexpectedness". The total number of data output from ConceptNet to the given concepts by the 10 subjects is 369. The similarity between the expected item and the output word was obtained from CNN. The larger the similarity value is, the more similar the output word is to the expected item. The subjective scores of "relevance" and "unexpectedness" were obtained from the subjective evaluation. The larger the subjective score is, the stronger the degree is.

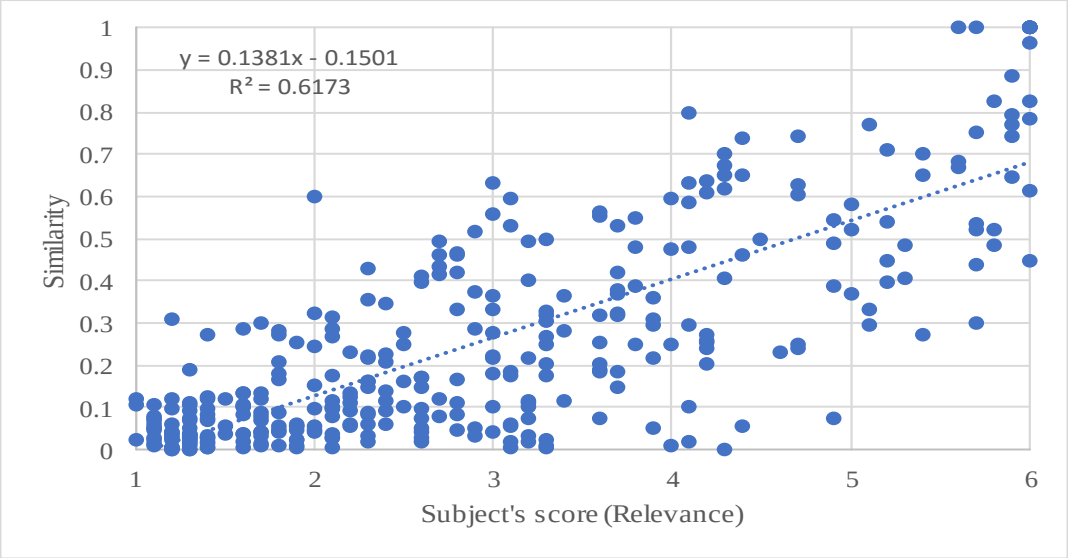

Figure 2. Correlation between Similarity and Subjective score of "Relevance"(left graph)

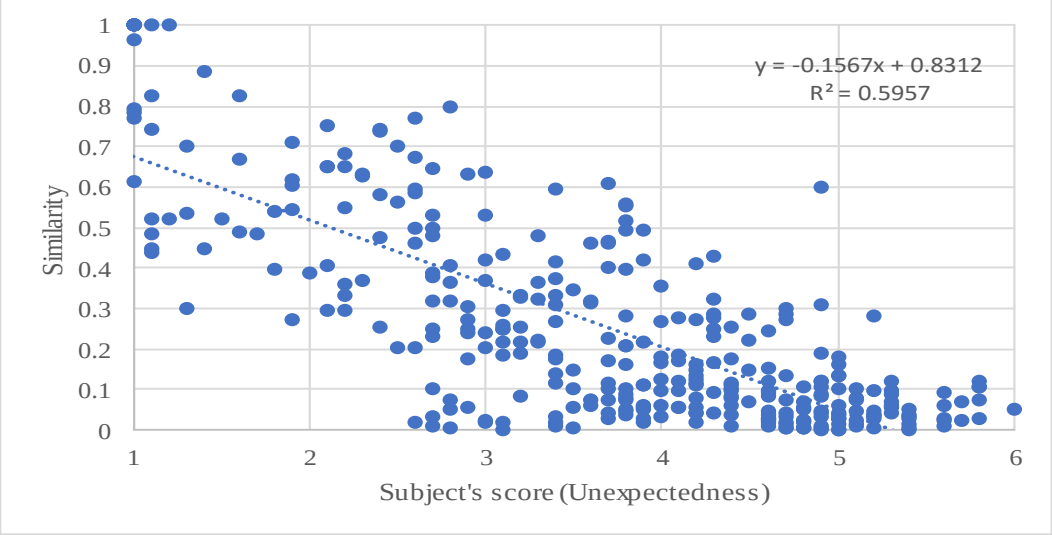

Figure 3. Correlation between Similarity and Subjective score of "Unexpectedness" (right graph)

As seen in Fig. 2, there is a high positive correlation between "similarity" and "relevance" with the correlation coefficient of 0.786. Figure 3, on the other hand, shows a high negative correlation between "similarity" and "unexpectedness" with the correlation coefficient of -0.772. The tendencies of the two are completely opposite. This can be predicted because if an expected item is strongly relevant to an output word, the unexpectedness of the output word to the expected item is low in general. As stated in section 3.1, however, "unexpectedness" depends on the user's experience. Therefore, the results of subjective evaluation will be carefully investigated in section 4.2.2.

4.2.2 Correlation among indexes of serendipity

In this section, the relation among the indexes of serendipity based on the results of subjective evaluation is investigated. For subjective evaluation, the evaluation of 10 subjects was used. Figure 4, Fig. 5 and Fig. 6 illustrate the correlations between "unexpectedness" and "relevance", "usefulness" and "unexpectedness", and "usefulness" and "relevance", respectively. The data number of data analyzed here is the same as the one described in section 4.3.2, which is 369. The error bars in the graphs show the standard deviation.

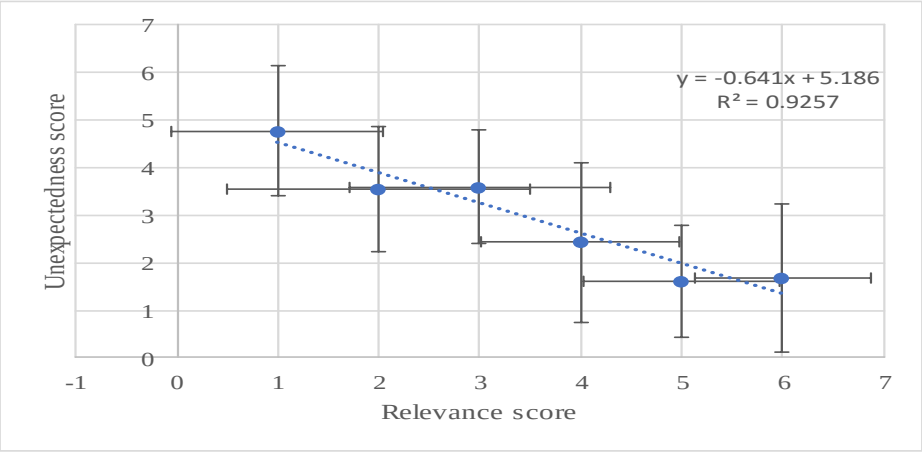

Figure 4. Correlation between "Unexpectedness" and "Relevance" (Left)

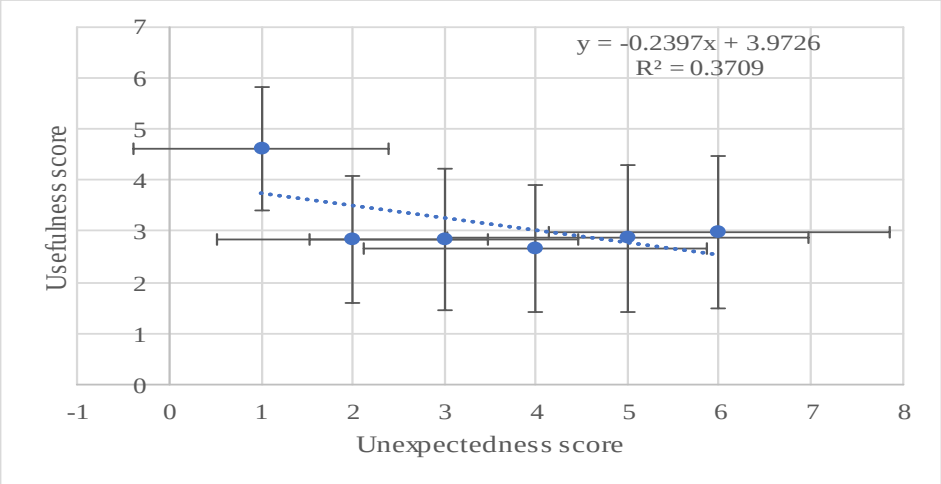

Figure 5. Relationship between "Usefulness" and "Unexpectedness" (Center)

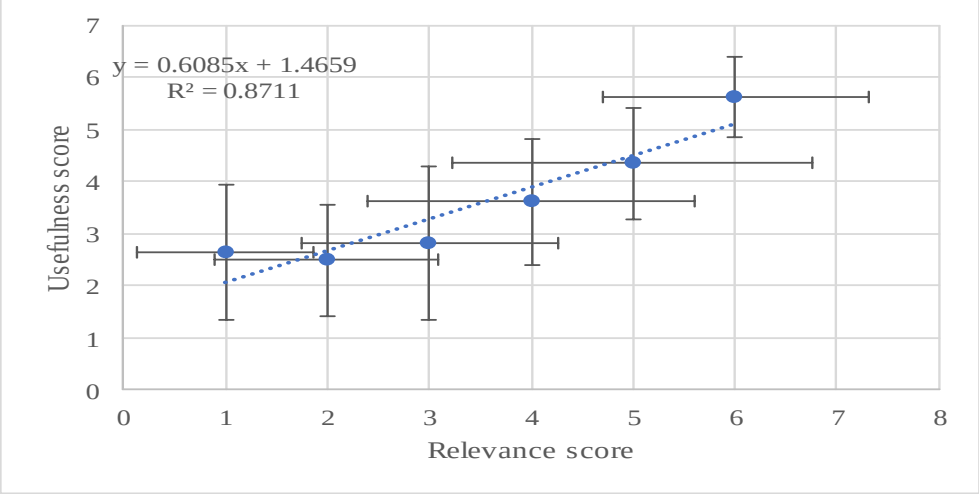

Figure 6. Relationship between "Usefulness" and "Relevance" (Right)

Figure 4 shows a high negative correlation between “unexpectedness” and “relevance”. This can be predicted from the results discussed in section 4.2.1. However, the width of the error bar is large. This is because the subjects did not simply evaluate that irrelevant words were unexpected, but used their subjective views as well for the evaluation of “unexpectedness”. Figure 5 does not indicate any significant correlation between “usefulness” and “unexpectedness”, meaning that it is difficult to obtain useful information from the viewpoint of “unexpectedness”. Figure 6 shows a high positive correlation between “usefulness” and “relevance”. This can also be predicted since when words are relevant to the expected item, they could be useful with a high probability. However, when the score of “relevance” is low, the width of the error bar is large. It infers that serendipity may be lurking here.

4.2.3 Accuracy of Finding Serendipitous information and expected items

In this section, the evaluation of the proposed system in terms of the accuracy of finding serendipitous information and expected items is discussed. In this study, the output word is estimated as serendipitous information when all the following three conditions are met: (1) Subjective score of “relevance” is less than 3 out of 6. (2) Subjective score of “unexpectedness” is more than 4 out of 6. (3) Subjective score of “usefulness” is more than 4 out of 6.

The value of similarity obtained by CNN is categorized in four types as shown in Table 1 and Table 2 shows the statistics of the results obtained by the evaluation experiment with the range of similarity value. The columns of “Word” and “Proportion” indicate the number of words output from the system, and the ratio of the number of output words in the range of similarity value to the total number of output words, respectively. The columns of “SRDP (subjective)” and “SRDP (estimated) indicate the number of output words determined as serendipitous information by the subjects, and the number of output words estimated as serendipitous information by the above three conditions, respectively. The column of “SRDP (determined)” indicates the number of output words estimated as serendipitous information included in the serendipitous information determined by the subjects. Therefore, the accuracy of finding serendipitous information can be calculated by the ratio of the total number of SRDP (determination) to the total number of SRDP (estimated). In addition, the lows of “Accuracy (exp)” and “Accuracy (SRDP)” show the accuracy that the expected item is included in the output words, and the accuracy of serendipity determination.

Table 1. Categorization of similarity value

Range of similarity value (sim)	Level of similarity
sim = 1	Outout word is identical with the expected item
$1 > sim \geq 0.5$	Output word is close to the expected item
$0.5 > sim > 0$	Output word is not close to the expected item
sim= 0	Output word is not supported by CNN

Table 2. Statistics of results obtained by evaluation experiment

Range of similarity value (sim)	Word	Propotion	SRDP (subjective)	SRDP (estimated)	SRDP (determined)
Sim = 1 (expected)	10	3%	0	0	0
$1 > sim \geq 0.5$	49	13%	6	1	1
$0.5 > sim > 0$	280	76%	35	46	32
sim = 0	30	8%	-	-	-
Accuracy (exp)					100%
Accuracy (SRDP)					70%

As seen in Table 2, Accuracy (exp) is 100%, meaning that the proposed system achieved to output all the expected items to all the subjects. Hence, it was revealed that the proposed system can be used to search for the expected items using their concepts.

As for the accuracy of serendipity determination, the serendipitous information tends to be found from the output words which are less relevant to the expected items ($0.5 > sim > 0$), where the number of output words is very large. Nevertheless, the accuracy of serendipity determination is 70%, which is quite high. This is because the indexes of serendipity defined in this study, which are “relevance”, “unexpectedness” and “usefulness”, work effectively to determine serendipitious information.

5. Conclusion

In this paper, an approach to easily obtain serendipitous information when users perform information retrieval was discussed based on the proposed indexes of the serendipity, which are "relevance", "unexpectedness" and "usefulness". In addition, the developed system to realize this approach using ConceptNet was demonstrated and evaluated. The system achieved the serendipitous information determination accuracy of 70%. In addition, it achieved the accuracy of information retrieval of 100% based on input of concepts. It was revealed that the proposed approach is effective to find serendipitous information from concept-based information retrieval. In this study, subjective evaluation results were used to estimate serendipitous information. As future study, how to use similarity values obtained from CNN with the subjective scores will be investigated.

References

1. Eli Pariser. 2011. The Filter Bubble: What The Internet Is Hiding From You. Penguin Press.
2. Foster. A, Ford. N. 2003. Serendipity and information seeking: an empirical study. Journal of Documentation. Vol 59. No 3. 321-340.
3. H Liu, P Singh. 2004. ConceptNet — a practical commonsense reasoning tool-kit. BT Technology Journal. Vol 22. No 4. 211-226.
4. Jauhar, J., Ghani, A.B.A., Joarder, M.H.R., Subhan, M., Islam, R. (2015). Brain drain to Singapore: A conceptual framework of Malaysians' diaspora. Social Sciences (Pakistan), 10 (6), pp. 702-711.
5. Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, John T. Riedl. 2004. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems. Vol 22. No 1. 5-53.
6. Kenta Oku, Fumio Hattori. 2013. Fusion-based Recommender System for Serendipity-Oriented Recommendations. Japan Society for Fuzzy Theory and Intelligent Informatics. Vol. 25. Issue 1. 524-539.
7. Panagiotis Adamopoulos, Alexander Tuzhilin. 2015. On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected. ACM Transactions on Intelligent Systems and Technology. Vol 5. Issue 4. No 54. 11-18.