The Effect of Diversity Implementation on Precision in Multicriteria Collaborative Filtering

Wiranto
Informatics Department
Sebelas Maret University
Surakarta, Indonesia

Edi Winarko
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia

Sri Hartati
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia

Retantyo Wardoyo
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia

Abstract —**This research was triggered by the criticism on the emergence of homogeneity in recommendation within the collaborative filtering based recommender systems that put similarity as the main principle in the algorithm. To overcome the problem of homogeneity, this study proposes a novelty, i.e. the diversity of recommendations applied to the multicriteria collaborative filtering-based document recommender systems. Development of the diversity recommendation was made by the two techniques, the first is to compare the similarity of content and the second is to use a variation of the criteria. The application of diversity, both content and criteria-based, was proven to provide a sufficiently significant influence on the increase of recommendation precision.**

Keywords—Algorithms; multicriteria; content; collaborative; filtering; systems; similarity; diversity; precision

I. INTRODUCTION

The development of collaborative filtering based recommender systems always puts the aspect of similarity as the main reference in the algorithm, and the main parameter used to assess the performance is the accuracy of prediction. Therefore, most studies on the recommender systems are focused on improving the accuracy of predictions, including when developing a multicriteria collaborative filtering model [1] [2] [3]. The implication of similarity implementation is the resultant recommendation that is homogeneous in nature. It is the advantage of collaborative filtering approach, but on the other hand it can also be disadvantage. The homogeneity of recommendation is due to process in collaborative filtering algorithm that does not involve the description or the content of recommendation object, so that the system does not accommodate the existence of new items [4]. The adverse effect is the case where many objects whose content is very interesting for the users, but it were never promoted to be a part of a list of recommendations.

Based on this fact, it is necessary to conduct a study with a focus on the development of recommendation diversity, but it remains in the corridor of multicriteria collaborative filtering.

Recommendation diversity is very important to be taken into account because it is closely related to the level of user satisfaction. In fact, it is not only important in recommender systems, but is also very important in developing a model of information retrieval systems and social media with very rapid development. The ideas of diversity developed in this study are of two kinds, content-based diversity and criteria-based diversity, which were done on multicriteria collaborative filtering model.

Besides prediction accuracy, there is other parameter used to measure the performance of recommender systems. The parameter is recommendation precision, which is defined as a number that indicates the percentage of items that were given a high predictive value by recommender systems, as well as by users. In this study, how much the influence of the implementation of recommendation diversity on the increase of precision in the multicriteria collaborative filtering applied to construct a scientific document recommender systems will be measured.

The writing of paper is organized as follows. Section 2 describes the content-based diversity. Section 3 describes the criteria-based diversity. The testing of recommendation precision is presented in section 4, while the discussion of the results of test is written in section 5. The writing of the paper is concluded by section 6.

II. CONTENT BASED DIVERSITY

The objects of recommendation in this study are scientific documents with text format, making it possible to do an analysis of its content. The results of the analysis of a document can be compared with other document contents. The results of the comparison of the document contents generate similarity values that are then used as the basis of determining which documents that need to be recommended to the users, with the guideline that the higher the content similarity, the lower the document diversity [6][7]. The scenario for determining the diversity based on the document content can be explained as follows:

a) Documents whose contents are analyzed are those already included into the list of Top-N produced by a multicriteria collaborative filtering engine.

b) Content analysis is sufficiently done on document abstract.

c) Content analytic process is meant to find out or measure the similarity.

d) One of the documents with high enough similarity values are chosen to be included into the list of the recommendation.

The analytic process of document content is done in two steps, i.e. indexing process and similarity measurement.

B. Document Indexing

Document index is a set of terms representing the content. Each document is represented with bag-of-words. The process is started by transforming the document into a bag containing independent words. Each word is stored into a database that is arranged as an inverted index. The arrangement of inverted index required the involvement of linguistic processing with aim of extracting important terms by deleting stop-words and stemming. The definition of stop-words is 'words that have no relevance with main subject, although the words often appear in many documents'. The example of stop-words include a, an, all, also, after, although, because, beside, every, the, this, it, these, those, his, her, my, our, their, your, few, many, several, some, for, and, nor, bit, or, yet, so, if, unless, on, off, over, of, during and etc. Meanwhile, stemming is an operation to gain a form of the roots of word by deleting the prefix or suffix. By the technique, a group of suitable words, where words in the group are variants, will be gained. As an example, the words *write, written, writer, writing* are interchangeably used in term with the general stem of *write*.

Forming the inverted index requires five steps, i.e.:

a) The deletion of format and document markup with many tags and formats such as HTML document.

b) Tokenization. The words in sentence, paragraph or pages are separated into token or pieces of a single word or stemmed word. Being included into the step is to delete certain characters such as punctuation mark and to change all the tokens into lower case.

c) Filtering, i.e. to determine which terms will be used to represent document in order to describe the document content and distinguish them with other documents. The terms with the very high level of frequency in appearance cannot be used for the purpose because they are unable to be discriminator inter-documents or often called by a term of poor discriminator. Moreover, terms often appearing in many documents do not also reflect the definition of the topic of sub-topic of documents. Therefore, the terms often used can be considered as stop-word and must be deleted. In order that the process of stop-word deletion goes fast, a book of stop-word or the stop-list of term that will be deleted must be arranged.

d) The retrieval of term into a form of root. Document can be expanded by searching synonymous for certain term within it. Synonym is words that have similar meaning but

morphologically seem different. The step is similar with stemming process, but what want to find is a group of relevant words. The main difference is that synonym does not share in use of term, but found based on thesaurus.

e) The weighting of term. To do the weighting, it can be selected local or global weighting model or the combination of both. The model often used in several applications is a combined weighting by the multiplication of the local weight of term frequency and the global inverse document frequency, written by tf.idf. [8]

C. The Measurement of Content Similarity

To measure the similarity of text formatted document, the bag-of-words need to be converted first into the vector space model with each document represented as a multidimensional vector with dimensions in accordance with the chopped term in the database. Figure 1 shows an example of visualization of three-dimensional vector space models with the terms of T_1, T_2, and T_3 as well as two documents of D_1 and D_2.
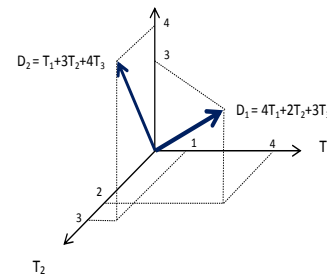

Fig. 1. An Example of Three-Dimensional Vector Space Model

The database of documents is represented by the matrix of term-document or term-frequency where each cell match with the weight given. The value of zero shows that the term does not appear in the document. Figure 2 is an example of term-document matrix for the database containing n document and t term.

Fig. 2. The Example of Term-Document Matrix

Based on the term-document matrix formed and the weighting of tf-idf, the numeric value of the document can be known, thus the inter-document nearness can be calculated. The nearer the two vectors are, the more similar the two dua documents. The similarity of text content document can be calculated by using a cosine similarity formula. For example, two vectors representing documents d_j and d_k were given, so

the content similarity between both documents was defined as [9].

$$sim(d_j, d_k) = \frac{\sum_{i=1}^{t}(w_{ij}.w_{ik})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 . \sum_{i=1}^{t} w_{ik}^2}}$$

To make understanding easier, the example of 3 documents was given by vector representation as follows :

$$D_1 = 2T_1 + 5T_2 + 6T_3$$
$$D_2 = 5T_1 + 3T_2 + 4T_3$$
$$D_3 = 4T_1 + 5T_2 + 5T_3$$

Thus, the inter document similarity value can be calculated as follows:

(1) similarity D_1 and D_2 :

$$sim(D_1, D_2) = \frac{(2*5 + 5*3 + 6*4)}{\sqrt{(4 + 25 + 36).(25 + 9 + 16)}}$$
$$= \frac{49}{\sqrt{(65)*(50)}} = 0.86$$

(2) similarity D_1 and D_3 :

$$sim(D_1, D_3) = \frac{(2*4 + 5*5 + 6*5)}{\sqrt{(4 + 25 + 36)*(16 + 25 + 25)}}$$

$$= \frac{63}{\sqrt{(65)*(66)}} = 0.96$$

(3) similarity D_2 and D_3 :

$$sim(D_2, D_3) = \frac{(5*4 + 3*5 + 4*5)}{\sqrt{(25 + 9 + 16)*(16 + 25 + 25)}}$$
$$= \frac{55}{\sqrt{(50).(66)}} = 0.95$$

From the three values of document similarity above, it can be known that document D_3 had similarity with the other two documents. The smallest similarity value was gained between document D_1 and document D_2, so the document prioritized to recommend was D_1.

III.   CRITERIA BASED DIVERSITY

Referring to a construction of document recommendation system by using multicriteria collaborative filtering, actually a space is available to engineer at the step of recommendation generation to make sure the presence of diversity [2] [10]. The new concept of diversity sufficiently bases at four individual criteria that was determined and used since earlier, different from the concept of document content-based diversity whose process was long enough and need the step of indexing. The construction of document recommendation system by using multicriteria collaborative filtering whose recommendation generation takes criteria-based diversity into account is shown in Figure 3.
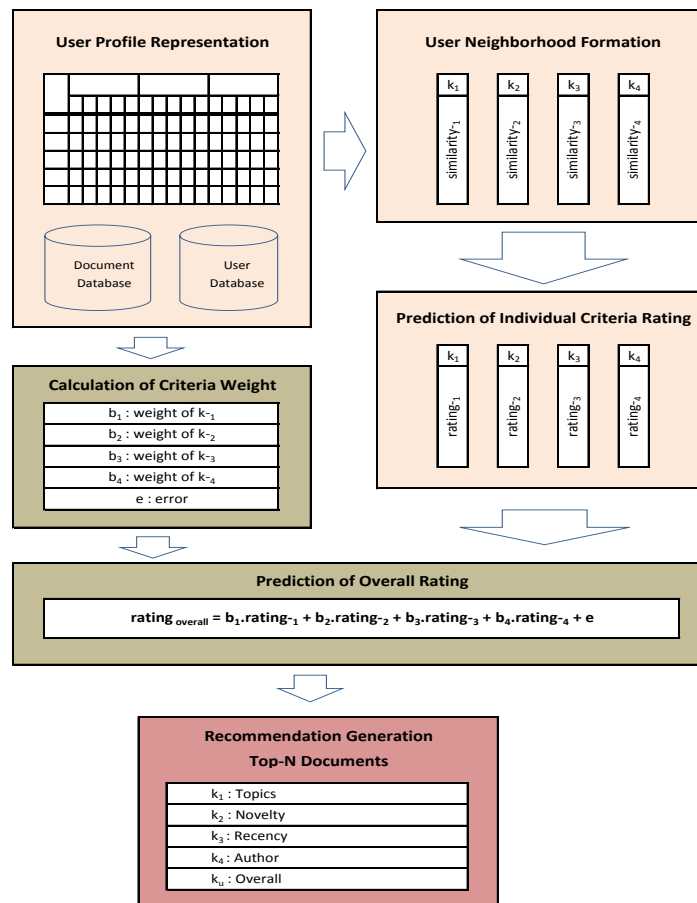

Fig. 3.   The Construction of Multicriteria Collaborative Filtering (MCF) Recommender Systems Model Applying the Criteria-Based Diversity.

Figure 3 shows that the scenario run was still in collaborative filtering paradigm with four individual document criteria and one overall criterion. Only the modification of document selection procedures and its generation process are required, so that the recommended documents are more various.

In the Multicriteria Collaborative Filtering, prediction process of ratings is done for each criterion [11]. So, by using the four individual criteria and one overall criterion, actually five values of prediction results were generated and each can be used to generate recommendation. For each criteria (topic, novelty, recency, author and overall), a number of document Top-N with the highest predictive value was generated. After the step, one document was taken respectively and put into a list of document recommended to the users. Thus, there will be five document variations recommended based on different criteria, although it is still possible for the emergence of the same document.

IV. EXPERIMENT

Initially the precision is only used in information retrieval systems and has always been associated with another metric called as recall that is defined as follows: [5]

$$Precision = \frac{|\{relevant\ documents\} \cap \{documents\ retrieved\}|}{|\{documents\ retrieved\}|}$$

$$Recall = \frac{|\{relevant\ documents\} \cap \{documents\ retrieved\}|}{|\{relevant\ documents\}|}$$

For measuring the precision in this study, the term is modified to be "Precision in Top-N" that is defined as a percentage of documents with the high production value, becoming the most relevant N document for the users.

In the testing, rating value will be categorized as high if the value was larger or equal to 4.0. The measuring of precision was done when the number of the users and documents reached 200x400, while the rating value used was the value for the overall criteria. The variation of the testing was based on the neighborhood size by determining a number of the users with the highest similarity value. In the testing, three neighborhood sizes were selected, including 5 users, 10 users, and 50 users. At the last option, the large neighborhood size led the meaning of nearest-neighborhood was bias and the load of computation become large also, but the measuring under the condition still need to do for the performance of system. Meanwhile, the Top-N values used were 5, 10 and 15.

The testing was done under two different conditions. The first was when the system did not apply the recommendation diversity yet, and the second was after the system applied the diversity. There were three variations in recommender systems run in the testing, namely : classic collaborative filtering (CF), multicriteria collaborative filtering using cosinus-based similarity (MCF Cosinus) and multicriteria collaborative filtering using multidimensional distance-based similarity (MCF MD Distance). Result of the testing of precision before the implementation of diversity was presented in Table 1.

From Table 1, it can be indicated that the larger the neighborhood size used, the higher the precision given by all of the recommendation systems models.

TABLE I.   PRECISION WITHOUT DIVERSITY IMPLEMENTATION

NO	NEIGHBORHOOD SIZE	MODEL	PRECISION (%)		
			Top-5	Top-10	Top-15
1	5 Users	Collaborative Filtering	65.0	65.0	67.4
		MCF Cosinus	68.2	68.4	68.9
		MCF MD Distance	70.1	71.8	73.3
2	10 Users	Collaborative Filtering	66.9	66.7	68.5
		MCF Cosinus	68.8	68.9	69.3
		MCF MD Distance	70.8	71.6	74.4
3	50 Users	Collaborative Filtering	70.8	69.8	70.1
		MCF Cosinus	71.8	69.9	70.2
		MCF MD Distance	71.9	75.1	76.4

In computation perspective, it can be concluded that the more the members accommodated in collaborative process, the more relevant and appropriate the resultant recommendations for meeting the need of the users. Observation on the size of Top-N document determined also resulted in the same information, where the larger the size of Top-N, the larger the precision value had by all the models. The highest value of recommendation precision was reached by MCF MD Distance at the neighborhood size of 50 users and with Top-N of 15 documents, i.e. 76.4%.

Results of the measuring of recommendation precision after the content-based diversity applied were presented in Table 2. Actually, the precision value increased significantly. It can be clearly seen from comparison between them with results of the measurement of recommendation precision when the diversity was not applied as presented in Table 1.

TABLE II.   PRECISION WITH CONTENT BASED DIVERSITY

NO	NEIGHBORHOOD SIZE	MODEL	PRECISION (%)		
			Top-5	Top-10	Top-15
1	5 Users	Collaborative Filtering	67.0	67.8	69.5
		MCF Cosinus	69.2	69.9	70.4
		MCF MD Distance	72.1	74.8	75.3
2	10 Users	Collaborative Filtering	68.0	68.8	70.5
		MCF Cosinus	70.1	71.9	72.7
		MCF MD Distance	72.1	74.8	75.3
3	50 Users	Collaborative Filtering	71.8	72.8	73.1
		MCF Cosinus	72.2	73.9	74.2
		MCF MD Distance	76.9	77.1	77.8

The concept of content-based diversity can be applied in all of the recommendation system models, while the concept of criteria-based diversity can only be applied in multicriteria collaborative filtering model with a process scheme simply illustrated by Figure 3. The results of the testing on the effect of criteria-based diversity concept application on the increase

of recommendation precision also give sufficiently positive information. It can be seen at the results of recommendation precision measurement as presented in Table 3.

The results of the testing increasingly affirm that the larger the neighborhood size used, the higher the precision given by all recommendation system models. Moreover, the larger the Top-N values selected, the larger the precision value had by all models. The highest value of recommendation precision was reached by MCF MD Distance at the neighborhood size of 50 users and the Top-N of 15 documents, i.e. 77.5 %.

TABLE III.　　　PRECISION WITH CRITERIA BASE DIVERSITY

NO	NEIGHBORHOOD SIZE	MODEL	PRECISION (%)		
			Top-5	Top-10	Top-15
1	5 Users	MCF Cosinus	68.2	70.9	70.8
		MCF MD Distance	72.0	73.8	75.4
2	10 Users	MCF Cosinus	70.2	72.5	72.6
		MCF MD Distance	72.6	74.6	76.2
3	50 Users	MCF Cosinus	72.4	74.1	74.4
		MCF MD Distance	76.8	77.3	77.5

V. DISCUSSIONS

The idea of recommendation diversity was generated with the aim to provide the added value, making it possible for the users to get documents that are more relevant to their needs. It can be expected that after getting the relevant documents, the users will be satisfied and give the high value of rating on the documents. This was consistent with the theory of consumer behavior, explaining that when a person feels satisfied and so happy with the service, it will provide a high and sustained appreciation. The more the documents are given the high value of rating by the users, the more the increase of recommendation precision. For the reason, in this testing, the measurement of recommendation precision as recommendations are generated involved content- and criteria-based document diversity.

The two concepts of diversity give a special feature in the process of generating recommendations, so that there is diversity within uniformity. The higher the level of the content similarity, the lower the level of document diversity. The main implication of the application of content diversity was that among the documents with high rating, some documents with relatively different contents are selected. For the criteria-based diversity, it is sufficiently determined based criteria variation. It means that among the documents with high rating, several documents with different criteria are selected.

VI. CONCLUSIONS

Based on the results of the measuring of recommendation precision, it can be concluded that the application of diversity in multicriteria collaborative filtering-based recommendation document system had a positive effect, namely, to increase the recommendation precision. It can be interpreted that basically the users want various recommendations, although generated by a system built on the collaborative filtering concept based on the principle of similarity. The results of the study indicate that each effort to develop the recommender systems should accommodate the idea of diversity in order to produce a kind of recommendation that is more relevant and able to meet the subjective needs of the users. Thus, the principle of similarity in the collaborative filtering can be enriched by the feature of diversity.

REFERENCES

[1] A. Umyarov and A. Tuzhilin, "Improving Rating Estimation in Recommender Systems Using Aggregation- and Variance-based Hierarchical Models," in *Proceedings of the Third ACM Conference on Recommender Systems*, New York, NY, USA, 2009, pp. 37–44.

[2] Wiranto, E. Winarko, S. Hartati, and R. Wardoyo, "Improving The Prediction Accuracy of Multicriteria Collaborative Filtering by Combination Algorithms," *International Journal of Advanced. Computer Science and Application.*, vol. 5, no. 4, May 2014.

[3] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," in *Proceedings of the 10th International Conference on World Wide Web*, New York, NY, USA, 2001, pp. 285–295.

[4] M. Zhang and N. Hurley, "Avoiding Monotony: Improving the Diversity of Recommendation Lists," in *Proceedings of the 2008 ACM Conference on Recommender Systems*, New York, NY, USA, 2008, pp. 123–130.

[5] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, "Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach," *ACM Trans Inf Syst*, vol. 23, no. 1, pp. 103–145, Jan. 2005.

[6] H. Shimodaira, "Similarity and Recommender Systems." http://www.inf.ed.ac.uk/teaching/courses/inf2b-learn-note02-2up.pdf, 21-Jan-2014.

[7] M. Ge, 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*, New York, NY, USA, 2010, pp. 257–260.

[8] N. Polettini, *The Vector Space Model in Information Retrieval- Term Weighting Problem*, Department of Information and Communication Technology University of Trento, Italy, 2004.

[9] B. M. Kim, Q. Li, C. S. Park, S. G. Kim, and J. Y. Kim, "A New Approach for Combining Content-based and Collaborative Filters," *J Intell Inf Syst*, vol. 27, no. 1, pp. 79–91, Jul. 2006.

[10] K. Chapphannarungsri and S. Maneeroj, "Combining Multiple Criteria and Multidimension for Movie Recommender System," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong, 2009, vol. I.

[11] L. Liu, N. Mehandjiev, and D.-L. Xu, "Multi-criteria Service Recommendation Based on User Criteria Preferences," in *Proceedings of the Fifth ACM Conference on Recommender Systems*, New York, NY, USA, 2011, pp. 77–84.

AUTHORS PROFILE

Wiranto is a lecturer at the Bachelor Informatics Program, Faculty of Mathematics and Natural Sciences, Sebelas Maret University in Surakarta. He received both Bachelor and Master of Computer Science from Gadjah Mada University in Yogyakarta, Indonesia. He is currently taking his Doctoral Program at Department of Computer Science and Electronics, Gadjah Mada University.

Edi Winarko is an Associate Profesor and Head of Department of Computer Science and Electronics, Gadjah Mada University in Yogyakarta, Indonesia. He received his bachelor degree in Statistics from Gadjah Mada University, M.Sc in Computer Sciences from Queen's University, Canada and Ph.D in Computer Sciences from Flinders University, Australia. His research interests are data mining, data warehousing and information retrieval.

Sri Hartati is an Associate Profesor and Chair of Computer Science Graduate Program, Gadjah Mada University in Yogyakarta, Indonesia. She received her bachelor degree in Electronics and Instrumentation from Gadjah Mada University, both M.Sc and Ph.D in Computer Sciences from New Brunswick, Canada. Her research interests are intelligent systems, decision support systems, medical computing and computational intelligence.

Retantyo Wardoyo is an Associate Profesor and Former Head of Department of Computer Science and Electronics, Gadjah Mada University in Yogyakarta, Indonesia. He received his bachelor degree in Mathematics from Gadjah Mada University, M.Sc and Ph.D in Computer Sciences from University of Manchester, United Kingdom. His research interests are fuzzy systems and expert systems.