Using Recommendation Systems for Explicit Knowledge Dissemination and Profiling Identification for Scientific and Engineering Contexts

Carlos Eduardo Barbosa², Jonice Oliveira¹, Luciano Maia², Jano M. de Souza^{1,2}

1 – COPPE-PESC/UFRJ - Graduate School and Research in Engineering (COPPE) - Department of Computer and Systems Engineering (PESC) / Federal University of Rio of Janeiro (UFRJ)

PO Box 68.511 – ZIP 21.941-972 - Rio de Janeiro – Brazil

Telephone: (+55 +21) 2562-8696 Fax: (+55 +21) 2562-8676

edu@ufrj.br, {jonice, jano}@cos.ufrj.br, lucianomaia@gmail.com 2 – IM-DCC/UFRJ - Institute of Mathematics (IM) - Computer Science Department (DCC) / Federal University of Rio of Janeiro (UFRJ)

ABSTRACT

In scientific and engineering scenarios we can notice the predominance of explicit knowledge being manipulated and distributed, which makes recommender systems very useful in this environment. But along with a knowledge management approach this kind of system can support the organization in better identifying competences, help engage users in a continuous and dynamic knowledge exchange, and customize knowledge dissemination as much as possible. In this work we detail a collaborative recommender system which is used in a Knowledge Management Environment for Scientific and Engineering contexts; we show how this approach can be aimed at a KM process and how this approach can deal with other kinds of knowledge used in research centers and universities.

Keywords

Recommendation system, knowledge management, explicit knowledge, profiling, expertise identification.

1. INTRODUCTION

Observing the considerable and increasing amount of digital information available it becomes more and more difficult to discern which the most relevant information is. Unfortunately we do not have enough time to read, interpret and assimilate all data about a subject of our interest. How can we go directly to the most appropriate contents? There is not a fully efficient answer but we can try to optimize the choice with the use of Recommendation Systems. We usually make choices without the complete knowledge of the possible alternatives. In everyday life we are based on recommendations from other people who have a similar profile to ours to choose movies, books, restaurants, etc. So Recommendation Systems aid and expand this social and natural process becoming useful in decision support, too.

People who dominate a subject can determine whether a document is interesting or not just by quickly browsing through it. However, they can not externalize the rules on which they base their evaluation of a material because this decision relies mainly on tacit knowledge and on the evaluator's personal competences. Thus, Recommendation Systems can be a powerful tool in the identification of such competences, aiming to identify them in an organization.

Web-based product and service recommendation systems have become ever popular business practice with increasing emphasis on modeling customer needs and providing them with targeted or personalized service solutions in real-time interaction. This kind of solution is especially useful to Knowledge Management whose main challenge is to identify and to enable the right knowledge (tacit and explicit) at the appropriate time.

The proposal of this work is the creation of a Recommendation System which was incorporated into a Knowledge Management environment, allowing engineers, researchers and students to receive, in an optimized way, explicit and useful knowledge for their research work. Apart from selectively disseminating explicit knowledge, this proposal also helps in identifying tacit knowledge and competences of the institution. To describe this work, a small revision of Recommendation Systems (section 2) will be made at first, showing different approaches, as well as advantages and disadvantages of each one. Next, the application scenario for this solution will be shown (section 3), as well as the approach itself (section 4). For a better understanding, a small case study will be provided (section 5). Finally we will mention future paths for works and the reach the conclusion of this work (section 6).

2. RECOMMENDATION SYSTEMS

Recommendation Systems provide access to relevant information, making suggestions based on previous examples of what an user likes or not. These systems are especially useful when the universe of choices is big and unknown, when the user does not have enough knowledge or expertise to decide, or when there is a record of similar behavior from other people.

There are two approaches to Recommendation Systems [1]: the content-based approach (section 2.1) and collaborative approach (section 2.2). It is important to remember that there are hybrid approaches to make a recommendation, which combine the two approaches.

2.1 Content-Based Approach

The premise, in which the content-based approach relies on, is that the user would like to see similar documents from others, which have already been seen and well evaluated previously. In this approach the similarity is calculated among evaluated documents and others which have not yet been evaluated by

the user. If the similarity among the evaluated and nonevaluated documents is big, and the evaluated document had obtained a good grade, the non-evaluated document is recommended.

One of the main problems related to this approach is to maintain the recommendation for the subjects that have been specifically evaluated by the user. A document on a new subject would never be recommended because it would not be similar to any of other documents which were positively evaluated by the user [1].

A document which was recommended by this approach may not necessarily have quality comparable to the document with which it was compared in the recommendation. This means that two documents can be similar, however with very different quality between them. One last factor to bear in mind is that there still are no efficient methods to compare objects such as video and audio [9] because content analysis is made especially by mining techniques.

2.2 Collaborative Approach

In the collaborative approach the system seeks similarities among the users in order to recommend a well-evaluated document for a similar user. The approach follows the principle that users are seeking a document which has already been evaluated positively by other users similar to him/her [6]. This kind of evaluation type is supposedly more efficient because people evaluate documents in a more proficient way than computational systems, due their previous experiences, expertise, and understanding of the area, quick association with correlated works and comparison with work of the same area.

In the collaborative approach it is possible to solve the problem found in the recommendation made via a content-based approach where the user only receives documents of similar contents making the access to new subjects and the possible analysis of existent subjects more difficult.

3. THE APPLICATION SCENARIO

With the intention of aiding Knowledge Management in research centers and universities the GCC was created [5], a proactive environment, capable of taking actions based on the researcher's profile and his/her domain, helping him/her in his/her tasks, promoting knowledge exchange and helping in the identification of the institution's competences.

There are several kinds of explicit scientific knowledge, such as models, formulas, description of scenarios and experiments, processes, mental maps, but mainly scientific knowledge is disseminated in the shape of documents, as publications, theses and technical reports. Faced with this scenario, and due to the growing number of publications on a research subject, a recommendation module was incorporated to GCC. This module was implemented to allow the evaluation of any kind of explicit scientific knowledge but initially it is being used for recommending documents.

In the initial page of the system the user is notified about the documents which have not yet been evaluated by him/her but that can be of his/her interest as shown in Figure 1. The user can upload and evaluate a document in his/her community but he/she can navigate to existing communities to access available documents and to evaluate them at any time, as

shown in Figure 2. We will explain the recommendation module, called MISIR, in detail below.



Figure 1. The user is notified about new interesting documents when logging into GCC [5].



Figure 2. The user can evaluate documents posted in a community.

4. THE MISIR APPROACH

We can divide the MISIR Approach in two stages: a prediction stage and recommendation stage. In the prediction stage, the algorithm seeks documents that the user has not evaluated and, based on each grade given by the other users and on profile similarities it calculates a grade for each non-evaluated document found. In the recommendation stage documents whose predicted grades were higher than a stipulated limit are selected.

In the first stage the similarity among users is calculated taking into account the "users x documents" matrix. In the following stage, all predicted grades are verified and, if the prediction grade is higher than the reference value, the document will be chosen for recommendation to the user.

By analyzing the comparisons for approaches, models and kinds of algorithms which can be developed, related to the reality in which GCC is found, we decided to invest in the collaborative approach, using the model in memory and the Pearson's algorithm's Correlation [2].

This choice proves ideal for our case because the GCC does not simply recommend documents, eliminating the content-based approach. The recommendation based on memory was chosen because it is more effective than the recommendation based on the model, as we can see in tests done by [2], and of easy implementation. In this section we will describe, in general terms, the Pearson's algorithm's Correlation. Initially, for the calculation of grade prediction [2], we used Equation

$$P_{a,j} = m_a + \frac{\sum_{i} w(a,i) \times (n_{i,j} - m_i)}{\sum_{i} |w(a,i)|}$$
 (1)

where

- Pa,j is the prediction grade of user a for item j;
- a is the user who wishes to make the prediction;
- j is the document upon which we wish to calculate the evaluation:
- w(a, i) is the correlation between user a and user i;
- mi is the arithmetical mean of user i's evaluations;
- ni,j is the grade which was given to document j by user i;

Calculation of the arithmetical mean of an user's grades is given by Equation 2.

$$m_i = \frac{\sum_{j} n_{i,j}}{N_i}$$
 (2)

where

- m_i is the arithmetical mean which we wish to calculate;
- j is a document which was already evaluated by the user:
- N_i is the number of documents evaluated by user i;
- n_{i,j} is the grade which was given to document j by user i:

For the Pearson Method, the value of w(a, i) is calculated based on the users' correlation. The correlation between user a and user i is calculated by Equation 3, where j is a document which was evaluated by both users.

$$w(a,i) = \frac{\sum_{j} (n_{a,j} - m_a) \times (n_{i,j} - m_i)}{\sqrt{\sum_{j} (n_{a,j} - m_a)^2} \times \sqrt{\sum_{j} (n_{i,j} - m_i)^2}}$$
 (3)

The Pearson coefficient w(a,i) obtains values in the [-1, +1] interval. The closer it is to -1, the less related the users are. On the other hand, the closer it is to +1, the more similar they are.

5. CASE STUDY

With the intention of demonstrating the efficiency of the recommendation, a small evaluation of the system will be described. The evaluation consists of comparing grades that the users had evaluated and their respective recommendations that would be made by the system. Comparing these two grades we can prove the efficiency of the recommendation. Thus, 10 books were chosen from the degree course in Computer Science. Five graduation students from the Computer Science course at UFRJ¹ took part in the evaluations. These students used the GCC to evaluate the books. Thus, 50 evaluations of the participant users were obtained.

The evaluation can be divided into three steps. The first step consists of randomly erasing part of the evaluations. In the second step the algorithm is executed and does the prediction for the erased grades. In the third and final step, the precision of the algorithm through metrics defined in [4] is calculated. The identification codes and the title of books are listed in Table 1

Table 1. List of different books used in different classes

Book	Title (in Portuguese)	Class
1	C Completo e Total	Computing (I and II)
2	Análise Combinatória e Probabilidade	Combinatory Analysis
3	Números Inteiros e Criptografia RSA	Integer Numbers and Cryptography
4	Estruturas de Dados e seus Algoritmos	Data Structure I
5	Introdução aos Sistemas Digitais	Logical Circuits
6	Física vol. I	Physics I
7	Redes de Computadores e a Internet	Network
8	Estatística Básica	Statistics and Probability
9	Engenharia de Software	Software Engineering
10	Compiladores: princípios, técnicas e ferramentas	Compilers I

The 50 evaluations are represented in Table 5. For the sake of simplification the document's code will be used as its identification in the following tables of this study.

In the first step of the study, we randomly erased 20% of the books evaluations. Notice that they could be different percentages; however, the bigger the number of erased grades, the greater will be the error in the prediction algorithm. The erased grades are marked in Table 2.

Table 2. Material Grades by Users

		Students (Users)					
		André	Daniel	Eduardo	João	Robson	
	1	3	3	5	5	5	
	2	3	2	3	3	4	
	3	4	5	5	5	5	
	4	3	5	5	5	5	
Books	5	4	3	4	5	4	
	6	4	3	3	1	3	
	7	4	4	5	3	2	
	8	1	4	3	1	1	
	9	2	4	4	5	2	
	10	3	2	4	4	5	

Then the erased grades were predicted. In this step, during the execution of the prediction algorithm, three matrices were generated: matrix of arithmetical mean for the user's

¹ UFRJ - Federal University of Rio de Janeiro, Brazil.

evaluations, matrix of correlation between the users and matrix of prediction.

The first matrix calculated in the algorithm is the arithmetical mean of the user's evaluations, generated for the program and illustrated in Table 3.

Table 3. Average Grades by Users

Student	Average
André	2,875
Daniel	3,500
Eduardo	4,125
João	3,625
Robson	3,750

The second generated matrix contains the correlations between the users, as shown in Table 4. The correlation between the users is calculated on the basis of the documents that both had evaluated.

Table 4. Symmetric matrix w obtained by Pearson's algorithm

Correlation Coefficient w(a,i)		User i					
		André	Daniel	Eduardo	João	Robson	
User a	André	1	0,017	0,659	0,218	0,479	
	Daniel	0,017	1	0,622	0,306	-0,203	
	Eduard	0,659	0,622	1	0,669	-0,133	
	João	0,218	0,306	0,669	1	0,271	
	Robson	0,479	-0,203	-0,133	0,271	1	

The matrix of correlation between the users is symmetrical. Correlations get values between [-1, +1]. The closer the correlation is to +1, the more similar the users are in the way of evaluating, i. e., they like or dislike the same documents. The closer the correlation is to -1, the more dissimilar the users are in the way of evaluating.

The third matrix calculated in the algorithm is equivalent to the prediction of the erased grades. The results are illustrated in Table 5.

Table 5. Matrix of grades, with 20% of prediction

		Students						
		André	Daniel	Eduardo	João	Robson		
	1	3	3	4,377	5	5		
	2	3	2	3	2,863	4		
	3	4	5	5	5	5		
	4	3	5	5	5	3,761		
oks	5	2,9	3	4	3,513	4		
Books	6	4	2,3	3,327	1	3		
	7	2,582	4,113	5	3	2		
	8	1	4	3	1	2,313		
	9	2	4	4	5	2		
	10	3	2	4	4	5		

The prediction having been concluded, we are going to measure the precision of the prediction algorithm. Initially, it is necessary to get the error for each prediction, comparing the grades calculated by the algorithm (prediction's grade) with the user's evaluation (real grade).

In [4] the author proposed to evaluate the collaborative filtering algorithms on the basis of Mean Absolute Error (MAE). The MAE formula shown in Equation 4, the closer the MAE is to 0, the more precise the algorithm is.

$$MAE = \frac{\sum_{i} |E_i|}{N} . \tag{4}$$

where:

- i corresponds to a prediction;
- E_i is the error of each prediction;
- N is the number of predictions carried out;

In the same way, we define the MPE, as in Equation 5. The closer the MPE is to 0, more precise the algorithm is.

$$MPE = \frac{\sum_{i} |pE_{i}|}{N} . \quad (5)$$

where:

- i corresponds to a prediction;
- pE_i is the percentage error of each prediction;
- N is the number of predictions carried out;

Indexes MAE and MPE were calculated using the data collected in this study. Predictions' values are listed in Table 6. And, as expected, it was found that the Mean Absolute Error value is close to zero, as demonstrated in [2].

Table 6. List with prediction grades

Student	Book ID	Pred. Grade	Real Grade	Diff.	% Error	Abs. Error
André	5	2,9	4	-1,1	22	1,1
Andre	7	2,582	4	-1,418	28,4	1,418
Daniel	6	2,3	3	-0,7	14	0,7
Daillei	7	4,113	4	0,113	2,3	0,113
Eduardo	1	4,377	5	-0,623	12,5	0,623
Eduardo	6	3,327	3	0,327	6,5	0,327
João	2	2,863	3	-0,137	2,7	0,137
3000	5	3,513	5	-1,487	29,7	1,487
Robson	4	3,761	5	-1,239	24,8	1,239
Kooson	8	2,313	1	1,313	26,3	1,313
				4 051	160.2	Q 157

 -4,951
 169,2
 8,457
 Total

 MPE
 16,9
 0,8457
 MAE

Finally, in possession of the predictions, the recommendation system rounds the predicted grade to the next integer and verifies if it reached the threshold that we stipulate as being grade 4 (in a 1 to 5 scale). Thus, the MISiR selects the documents to be recommended for each user, as demonstrated in Table 7 of our example, where 40% of the predictions were converted into recommendations.

We can see that the system makes many recommendations and these recommendations are precise (as it can be found by comparing Tables 2 and 5). It indicates that it is possible to have a good idea of the user's taste only by comparing his/her evaluations with other users' evaluations.

The pure MAE index value can lose some of its meaning since we are not making comparisons of efficiency between recommendation systems. However, its percentage error, Mean Percentage Error (MPE) is of 16.9%. This means that the predicted grades have, on average, 16.9% error.

Another thing that can be seen is that, in their majority, the biggest errors occur with grades 1 and 5. This happens because these upper/lower limits behave in an asymptotic way in relation to the recommendation. Then, even if these errors are physically bigger on average, they are less representative because they do not interfere in the recommendation of a document

Table 7. List of Recommendations

Student	Book	Predicted
Daniel	Redes de Computadores e a Internet	4,113 ≅ 4
Eduardo	C Completo e Total	4,377 ≅ 4
João	Introdução aos Sistemas Digitais	3,513 ≅ 4
Robson	Estruturas de Dados e seus	3,761 ≅ 4

According to [8], tacit knowledge is directly associated to our actions. So, when we evaluate a document, we are making use of our knowledge, degree of "expertise" and interests. MISIR helps in knowledge identification given that the approach is directly connected to the identification of explicit knowledge in an institution, and consequently in the knowledge dissemination, too. Moreover, it helps in identifying individual competences, since through the evaluation of documents we can track evaluators' competences and interests.

6. MISIR AND KNOWLEDGE MANAGEMENT

Several authors have tried to define and classify the different activities related to Knowledge Management. In [10], the author analyzed the main models of Knowledge Management and established a generic model for this activity. According to the author, a process of KM can be divided into: i) Identification, ii) Capture, iii) Selection and Validation, iv) Organization and Storage, v) Dissemination, vi) Application, and vii) Creation. Due to its being a generic model, we based the evaluation of our work and its applicability in KM on this model. MISIR helps in all of the stages, as shown in Figure 3, except in the Organization and Storage process, as we shall discuss below.

According to [10], tacit knowledge is directly associated to our actions. So, when we evaluate a document, we are

making use of our knowledge, degree of "expertise" and interests. MISIR helps in the process of **Identification** given that the approach is directly connected to the identification of explicit knowledge in an institution. Moreover, it helps in identifying individual competences, since through the evaluation of documents we can track evaluators' competences and interests.

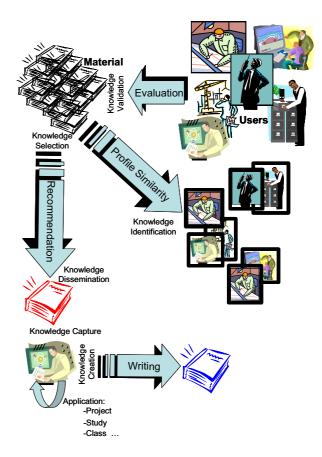


Figure 3- Knowledge Transformation with MISIR

The Capture process is responsible for the acquisition of knowledge necessary to create and maintain the essential competences in an organization. As this proposal seeks to emphasize explicit knowledge in document form which is more similar and useful to an user it helps in the development and enrichment of such competences. Due to its characteristic of trying to show more relevant results to the user the approach is also directly related to the process of Selection and Validation. A Recommendation System is pro-active because it recommends a document without the user asking explicitly for that. However, the information is always available to the user.

In relation to knowledge **Dissemination**, the implementation of some mechanism capable to disseminate the knowledge automatically for the several interested parties is necessary, so that new knowledge or information is quickly passed on to those who need it. This is the case when users are notified and recommended on the existence of new documents, according to one's profile.

In relation to the **Application** stage, although the knowledge, the experiences and the information are available and can be shared, it is fundamental that they are used and also applied to

real situations of the organization in a way to produce concrete benefits such as performance improvements, release of new products and the conquering of new markets. In our context, the approach can help in the application of new knowledge (what was recommended, worked on, and assimilated) in a research scenario, as a successful experiment, design project or practice, for example.

Taking into account that the solution always seeks to make new, useful knowledge available, which also fits the interests of the user, it also helps the process of **Creation** of new knowledge.

7. CONCLUSION AND FUTURE WORKS

This work presented an environment of recommendation, using the collaborative approach, for Scientific and Engineering Knowledge Management scenarios. As seen in the previous section this work assists in several stages of the Knowledge Management process also being a useful tool in the identification of each person's competences.

Although it has advantages, this approach does not deal with problems such as the insertion of new documents in the database, which will be recommended only later, after some users have evaluated it. A tougher issue to correct is the treatment given to users who have interests that not fit in groups. This particular kind of user won't have - or have only a few - similar users which the MISIR can rely on to make the predictions.

We can mention other problems which are not treated in this approach. One of them is how much we can trust in a recommendation. This is another task gleaned from discussions with users.

To solve problems related to sparse matrix (few matching), latter recommendations and a trusty recommendation, we intent to use social network information to improve the process, during two moments: i) the correlation coefficient calculus and ii) recommendation. We have another project which identify and map scientific and technical social networks, and information about a social network will help in search and dissemination straighter, more useful and important information.

The biggest advantage of the chosen approach is its extensibility. The described algorithm can be used to recommend any types of explicit knowledge and widely used in the scientific scenario, such as links, audio, video, processes, mental maps, models, primary data, description of experiments, practices, amongst others. As future work we can mention the enlargement of the proposal aimed at the work with these kinds of explicit knowledge. Moreover, we will extend this proposal to a hybrid proposal, taking into consideration the user's experience time and expertise in order to ponder and to differentiate the evaluation provided.

A hybrid proposal would improve the recommendation of some types of media. Mental maps, models, processes and documents would have an improved recommendation with this method [7].

8. REFERENCES

- Balabanovic, M., & Shoham, Y., 1997, "Fab: Content-Based, Collaborative Recommendation",
 Communications of the ACM, v. 40, n. 3, pp. 66-72.
- [2] Breese, J. S., Heckerman, D., Kadie, C., 1998, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering". In: Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 43-52, Madison, WI, USA.
- [3] Herlocker, J., Konstan, J., Borchers, J., et al., 1999, "An algorithmic framework for performing collaborative filtering". In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, p.230-237, August, Berkeley, California, USA.
- [4] O'Connor, M. & Herlocker, J., 1999, "Clustering Items for Collaborative Filtering". In: Workshop on Recommender Systems: Algorithms and Evaluation, 1999 Conference on Research and Development in Information Retrieval, California, USA.
- [5] Oliveira, J. et al., 2005, "GCC: An Environment for Knowledge Management in Scientific Research and Higher Education Centres" Proceedings of I-KNOW '05. Graz, Austria, June, 2005
- [6] Resnick, P., Iacovou, N., Sushak, M., et al, 1994, "GroupLens: An open architecture for collaborative filtering of netnews". In: Proceedings of the 1994 ACM conference on Computer supported cooperative work, pp. 175 – 186, Chapel Hill, North Carolina, USA.
- [7] James Salter, Nick Antonopoulos, "CinemaScreen Recommender Agent: Combining Collaborative and Content-Based Filtering," IEEE Intelligent Systems, vol. 21, no. 1, pp. 35-41, Jan/Feb, 2006.
- [8] Schuman, L., 1987, "Plans and Situated Actions: The Problem of Human-Machine Communication, Cambridge University Press, Cambridge, UK.
- [9] Shardanad, U. & Maes, P., 1995, "Social information filtering: Algorithms for automating "word of mouth"".
 In: Proceedings of the Conference on Human Factors in Computing Systems – CHI 95, Denver, pp. 210-12
- [10] Stollenwerk, M. F. L., 2001, "Gestão do Conhecimento: Conceitos e Modelos". In Tarapanoff, Kira, Inteligência Organizacional e Competitiva, Chapter 5, Brasília, Editora UnB (in Portuguese)