# Recommendation Algorithm for Digital Libraries

Oscar Cruz-García, Jesús-M Olivares-Ceja

Centro de Investigación en Computación del IPN,
Av. Juan de Dios Bátiz esq. M. Othon de Mendizabal S/N
Unidad Profesional "Adolfo López Mateos"
07738 Ciudad de México, MÉXICO
jesuso@acm.org, oscarcruzgarcia@hotmail.com
http://www.jesusolivares.com

**Abstract.** Digital Libraries provide users with information access in a variety of geographical places using different content formats. Search engines have been developed to help users in finding information tasks. Despite of the speed and advances in search techniques, users still face problems while trying to locate information which meet their requirements. Researchers have developed recommendation tools that attempt to provide users with information that narrow the huge amount of available sources. On the other hand, Software Agents accomplish proactive searches on user's behalf. Agents employ a user model for information findings. Currently many recommendation techniques and methods have been developed. In this paper we describe an algorithm that provides both objective and subjective points of view for recommendations. The algorithm employs an ontology that contains digital library concepts. User and document model points to nodes in the ontology. The algorithm calculates the closest documents that fulfill user requests using a metric called confusion theory.

**Keywords:** Knowledge representation, Digital Library, Ontology, Objective and Subjective recommendations.

## 1. Introduction

Nowadays, the Web is populated with a huge amount of information and knowledge sources that are used to satisfy user requirements. Digital Libraries are part of this repository. Many traditional libraries are introducing networks to support user services. Despite of the fact that many different search engines that have been developed, many users still face problems trying to find knowledge sources that closely approximate their requirements. As a consequence, recommendation systems have been proposed to assist users in the process of locating information.

*Digital Library*[1] is a term that refers to "a library [7] in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers. The digital content may be stored locally, or accessed remotely via

[1] http://en.wikipedia.org/wiki/Digital_library


*Advances in Computer Science and Engineering.*
*Research in Computing Science 34, 2008, pp. 289-295*

*Received 02/04/08*
*Accepted 26/04/08*
*Final version 05/05/08*

computer networks. A digital library is a type of information retrieval system. The first use of the term digital library in print may have been in a 1988 report to the Corporation for National Research Initiatives. The term digital libraries was first popularized by the NSF/DARPA/NASA Digital Libraries Initiative in 1994."

A *Recommender System*[2] "is a type of information filtering (IF) technique that attempts to present information items (movies, music, books, news, images, web pages) that are likely of interest to the user. Typically, a recommender system compares the user's profile to some reference characteristics. These characteristics may be from the information item (the content-based approach) or the user's social environment (the collaborative filtering approach).

*Ontology* is an explicit specification of a conceptualization. In this paper, the ontology is a set of hierarchical nodes that represent digital library concepts.

A *User Profile* describes user's interests alongside with identification and personal information, and the main interest is related with the set of concepts of the ontology that representing user's interests. When a user explicitly specifies the nodes that is interest on, we obtain objective recommendations. When a user issues queries directly to the digital library, the supplied keywords are matched against the ontology producing nodes that we call the subjective interest, therefore, subjective recommendations result from such information usage.

A *Source Profile* (we use the acronym PF for the Spanish equivalent 'Perfil de la Fuente') is related with the set of keywords of a document (paper, video, book) in the Digital Library that maps the document with nodes in the ontology.

Since 80's we find works related with information user recommendations. In [9] user queries were modeled with point in Euclidean space, documents that satisfy user requirements are also points in that space, therefore a linear distance is used to find the documents that are closer to user queries.

Since 80's, many researches have been working in user profiling and information modeling. Particularly, digital libraries have been focused as a field of study. In [10] we found a similar work that employs user actions for building user profile that is similar as our subjective recommendations. In our paper we let user to select keywords to produce objective recommendations.

The rest of the paper is organized as follows. In the section 2 we explain the algorithm. Section 3 explains the system that is under development that uses the proposed algorithm. Conclusions and references are given at the end of this paper.

## 2. Recommendation Algorithm

As part of the tools that are being incorporated within a digital library, Recommender Systems provide advice and suggestions to users. Some recommendations are based on groups and others based on individual profiles.

Our proposal is an approach that focuses individual users' interests stored in user profile. As mentioned above user profiles (PU) are mapped against the ontology and -

[2] http://en.wikipedia.org/wiki/Recommender_system

also document profiles (PF) are also mapped against the ontology.

Generally, nodes from user profile does not coincide with the nodes from documents, therefore it is mandatory to use a metric that evaluates how closely a document matches user's information requirements (figure 1).

Our algorithm is based on the confusion theory proposed by Levachkine and Guzman [1]. This theory provides a measure of the confusion on the utilization of a node belonging to the ontology instead of other node. For example, somebody requires a feline and he is provided with a cat. Here the confusion is zero. But if instead of a cat he is provided with a rose, then there is a difference. This difference is calculated using the relative distance among nodes in the ontology.

Consider the following nodes as an example (indentation represents relatives, for example FELINE is son of ANIMAL):

```
THING
  ANIMAL
    FELINE
      Cat
      Tiger
    CANINE
      Dog
  PLANT
    FLOWER
      Rose
      Carnation
    FRUIT
      Strawberry
```

The Confusion Algorithm states that if I am requested with a node that is below the provided node, for example, I was requested with a Feline but a Cat is received instead, the confusion is zero. But if I was asked to provide a Tiger but a Feline is given, then confusion is one in this case, because I am not sure that the specific node concept is given.

In our work we take these ideas and divide the confusion by the longest path between nodes involved to provide in a number within 0 and 100 % in this form.

Recommendation algorithm has the following steps:

1. User keywords ($k_1$, $k_2$, ..., $k_n$) are mapped with ontology nodes. This mapping is done using ontology nodes keywords ($c_1$, $c_2$, ..., $c_l$). We obtain a *percentage of satisfaction* (S), this a value in the interval [0, 1], calculated as follows:

$$S = \frac{1}{n} \sum_{i=1}^{n} \text{equal}(k_i, c_j)$$

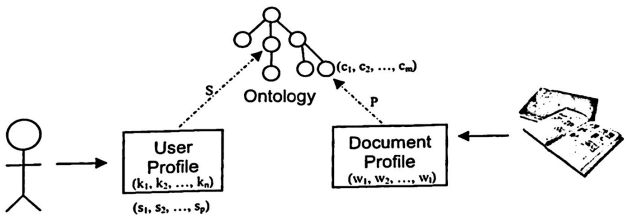


**Fig. 1.** Context of our recommendation algorithm

where:

$$\text{equal}(k_i, c_j) = \begin{cases} 1 & \text{if } w_i = c_j,\ j=1,..,l \\ 0 & \text{in other case} \end{cases}$$

2. Mapping document keywords ($w_1$, $w_2$, ..., $w_m$) against ontology nodes, as stated above by keywords ($c_1$, $c_2$, ..., $c_l$). In this case, we obtain a *percentage of belonging* (P), a value in the interval [0, 1], calculated with the formula:

$$P = \frac{1}{m} \sum_{i=1}^{m} \text{equal}(w_i, c_j)$$

where:

$$\text{equal}(w_i, c_j) = \begin{cases} 1 & \text{if } w_i = c_j,\ j=1,..,l \\ 0 & \text{in other case} \end{cases}$$

3. Check up the new acquisitions table and calculate R that is the recommendation percentage for each document in the table.

$$R = S\ P\ (1 - \text{conf}(K, W))$$

where:

K is the set of user keywords.
W is the set of document keywords.

4. User recommendations are those documents for which R is maximal.

### 2.1. User Modeling

In our algorithm users are modeled using keywords that reflect user interests. These keywords are used during recommendation process to find ontology nodes that maximize matching among user and node keywords. For each user keyword a number S related with each ontology node is calculated. The five top nodes are selected as the best nodes that define user interests, automatically. There is another way to determine ontology nodes that define user interests. That is the set of nodes which S is below a threshold established by each user.

After calculating S, user interests are a set of five or less nodes with a percentage of user satisfaction. We have selected five nodes, but another number of nodes could be used according with user's requirements. The third option to obtain user interests is to select directly from ontology those nodes that fulfil user's requirements; in this case S should be 100 % or users themselves may choose this percentage.

### 2.2. Document Modeling

Documents are modeled using their keywords that are mapped in a similar fashion as user keywords. The same percentage is provided for each node that contains at least one document keyword. For example if the document contains the keywords: database, relational-model and data-design; and upon calling the database a node appears that contains two keywords: database and information; then the node is assigned a 33 % because one of three keywords of the document matches with concept keywords.

### 2.3. Objective Recommendation

Once we have user profile (PU) and a set of document profiles (PF), the recommendation algorithms works attempting to find the documents that should be provided to the user as recommendations.

The algorithm works with a database table that contains new a acquisition to narrow the universe of calculations. For each document that appears as a new acquisition, the following formula is calculated to determine the rank of recommendation:

$$R = S\ P\ (1 - \text{conf}(K, W))$$

where:

K is the set of user keywords.

W is the set of document keywords.

with the above we expect to obtain a number within [0, 1] that represents a percentage of recommendation. The recommendations above a threshold contained in the user profile are showed to the user. These are objective recommendations because they are based on the nodes that have been selected by the user.

### 2.4. Subjective Recommendation

We found an alternative approach to calculate recommendations. This approach takes into account the keywords issued by the user when he or she searches the digital library. The keywords are stored in the user profile in a place different from keywords issued directly by the user.

The term subjective recommendation is used due to the fact that the user did not request those keywords explicitly but instead he or she uses these keywords during search tasks.

Recommendation calculus is done in the similar way as for objective recommendations. Subjective recommendations appear below objective recommendations unless user states differently in the user profile options.

## 3. Recommender Agent System Development

Currently an Agent Based System for a Digital Library is under development, and the proposed algorithm explained in this paper is part of it to determine what the recommendations to users are.

The system is developed using MySQL database and Java Servlets in a Web environment. User profile is generated using an HTML screen and a set of servlets are employed to expand the nodes of the ontology to enable user the selection. Nodes are stored in his or hers user profile in the database. Ulterior update is possible by erasing undesired keywords and inserting new keywords. The system automatically searches the ontology to determine which nodes represent user interest and also the percentage of accuracy.

A similar treatment is done with document keywords to determine the nodes and accuracy of the recommendation.

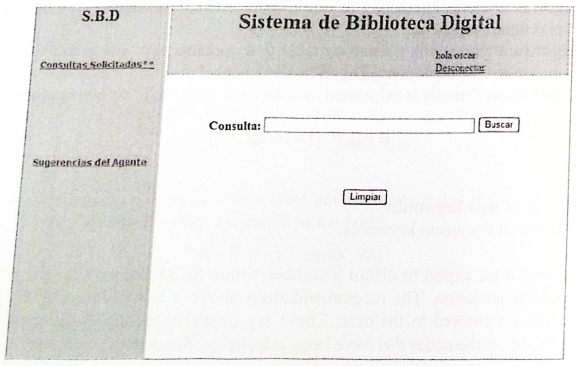


*Fig. 2.* Screenshot of the main user interaction screen

As the system is agent based, there is an agent server that at schedules the time to activate the agents that issues user recommendations.

Currently recommendations are stored in the database and they are optionally browsed by the user. They appear as a link in the main user screen called "Sugerencias del Agente" (figure 2). We are developing a tool to send proactively the suggestions generated to the user, according with the relative importance given by user. The importance is also stored in the user profile.

## 4. Conclusions

We have presented an algorithm to calculate user recommendations of documents out of a digital library. User and documents are modeled as a set of keywords that points to nodes of an ontology. For each node the percentage of reference is calculated and used further on to calculate of objective and subjective recommendations to the user.

A set of computer programs are under development to provide a tool for improving user's employment of a digital library.

## References

1. Serguei Levachkine, Adolfo Guzman, Hierarchy as a new data type for qualitative variables in *Expert Systems with Applications*, volume 32, issue 3, ISSN 0957-4174, april 2007
2. Jesús Manuel Olivares C., Sistema Evolutivo para Representación del Conocimiento (bachelor degree theses), IPN-UPIICSA, clasif. 7.152, Mexico City, abril 1991
3. Thomas R. Gruber, Toward Principles for the Design of Ontologies Used for Knowledge Sharing en Formal Ontology in *Conceptual Analysis and Knowledge Representation*, edited by Nicola Guarino and Roberto Poli, Kluwer Academic Publishers, Italy 1993
4. Natascha Hoebel, Sascha Kaufmann, Karsten Tolle, Roberto V. Zicari, The Design of Gugubarra 2.0: A Tool for Building and Managing Profiles of Web Users in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, December 2006
5. Stuart E. Middleton, Nigel R. Shadbolt, David C. De Roure, Ontological user profiling in recommender systems in *ACM Transactions on Information Systems (TOIS)*, Volume 22 Issue 1, January 2004
6. Adolfo Guzmán A., Finding the Main Themes in a Spanish Document *en Journal Expert Systems with Applications*, Vol. 14, No. 1/2, 139-148, Jan./Feb. 1998
7. Greenstein, Daniel I., Thorin, Suzanne Elizabeth. "The Digital Library: A Biography" Digital Library Federation (2002) ISBN 1933645180
8. Tom Gruber, A translation approach to portable ontology specifications in *Knowledge Acquisition* 5, (1993) pp. 199-220
9. Robert Korfhage, Query Enhancement by user profiles in *Proceedings of the 7th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, (July 1984), British Computer Society
10. Stuart E. Middleton, Nigel R. Shadbolr, David C. De Roure, Ontological User Profiling in Recommender Systems in *ACM Transactions on Information Systems (TOIS)*, Volume 22, Issue 1, January 2004