PIRSIDIL: A personalized information retrieval system in a digital library

Thameur Dhieb, Mahmoud Neji
MIRACL, Faculty of Economics and Management of Sfax
University of Sfax
FSEG PB:1088, 3018 Sfax Tunisia
{thameur.dhieb, mahmoud.neji}@issatgf.rnu.tn

ABSTRACT: *The personalization in information retrieval has for goal to provide applicable information that corresponds precisely to the user's needs. In this paper, we are interested in creating a "Personalized Information Retrieval System In a Digital Library" called PIRSIDIL in order to present to the current user relevant retrieved documents adapted to their centers of interests and their needs. To achieve this goal, we are interested in modeling the user profile and in describing the metadata of documents and use them in the matching. The first experimental evaluations show a significant improvement of results offered by our system compared to Google.*

Key words: Information Retrieval System, Personalization, User Profile, Metadata, Matching, Digital Library

Received: 14 June 2011, Revised 31 July 2011, Accepted 6 August 2011

1. Introduction

The considerable development of the Internet in these last years has led to a large increase in the number of users and also the number of heterogeneous resources available (structured data, textual documents, software components, pictures, etc.): According to a statistical study published in the website "http://www.worldwidewebsize.com" covering several search engines, the size of the Web is estimated between 20 and 40 billions of web pages in the year 2010. Thus, when a user issues a query, the systems of information retrieval deliver massive results satisfying only the query and suppose that the user is completely represented by his query without taking into account of his center of interest, his preference and his context of research [1] [2].

Overcome this limit and allow the user to easily find information that interests him are the subject of several recent research projects [3] [4] [5]. This is the case of information retrieval known as personalized information retrieval that consists to integrate the user in the process of finding information in order to satisfy the needs in information of every user.

In this context, our work focuses on the creation of a personalized information retrieval system in a digital library that we named it PIRSIDIL in the goal to assist the user in the process of information retrieval. We are interested in this paper to describe the metadata of documents and to build the user profile which will be used in a process of personalized information retrieval by matching "user profile" with "metadata" and ordering search results according to the centers of interests of the user.

We begin to explore the dimensions of our personalization, then we illustrate the phase of exploitation of personalization in a process of information retrieval based on our method of matching user profile with metadata, afterwards, we present our implementation and the results of our first tests. We end with a conclusion summarizing our contribution.

2. Dimension of personalization

Under our system, we propose to define two main dimensions of search personalization:

1- Dimension describes the user across their centers of interests and their preferences of research while building the user profile.

2- Dimension related to documents by presenting their metadata.

2.1 User profile

We define the user profile as a set of data (personal, professional and search preferences) that describe the user in order to improve the result of information retrieval. The most important step in building the user profile is the acquisition that can be: [6] [7].

i) Explicit: based on the collection of user's data directly entered by him through an interface of the system.

ii) Implicit: based on the collection of user's data by observing their interactions with the system during their activities of research.

iii) Mixed: based on a combination of explicit initialization phase, that constructs the user profile and implicit phase for completing the profile or infer new centers of interests not set by the user.

We chose to use the mixed acquisition because it combines the advantages of the two previous acquisitions methods. After the acquisition phase, we attain the representation phase of user profile. It exists in the literature three main types of representation: vector, semantic and multidimensional representation. We present the description, the strength and the weakness of the three user profile representations in Table 1.

	Representation		
	Vector	***Semantic***	***Multidimensional***
Description	Based on the vector model proposed by Salton in 1971 [8]. The profile is represented by one or several vectors defined in a space terms. The coordinates of the vectors correspond to the weights of terms in the profile [9].	Based on the construction of a hierarchy of concepts or ontology instead of a personal set of independent terms. Each category of the hierarchy represents the knowledge of a domain of interest of the user [10] [11].	Captures and categorize all information characterizing the user profile. The last may contain several types of information such as demographics data, centers of interests, historic information. Each type of information represents a dimension in the model [4].
Strength	- Simplicity of implementation	- Avoid semantic ambiguity	- Better interpretation of user profile.
Weakness	- Neither makes interpretation nor considering levels of characterizing general user.	- Do not use the hierarchic structure to capture dynamics profile change.	- Research strongly linked to the databases domain.

Table 1. User profile representations

We opted for a multidimensional representation; our choice justified by the fact that the profile is characterized by several categories of information that is to say several dimensions that can be decomposed into sub-dimensions. Some dimensions will be given explicitly by the user and others will be implicitly deducted by the system.

The profile is structured in our system along three dimensions: The first is called “personal data” which contains all data from user's identity (first name, last name, civil status, date of birth, country, city, and email). The second is a dimension that represents the “professional data”, it contains the level of study, the specialty and the centers of interests of the user and finally a third dimension designed by “search preferences” that defines the year of publication, the type, the format and the language of resources that the user wants to look for in the digital library.

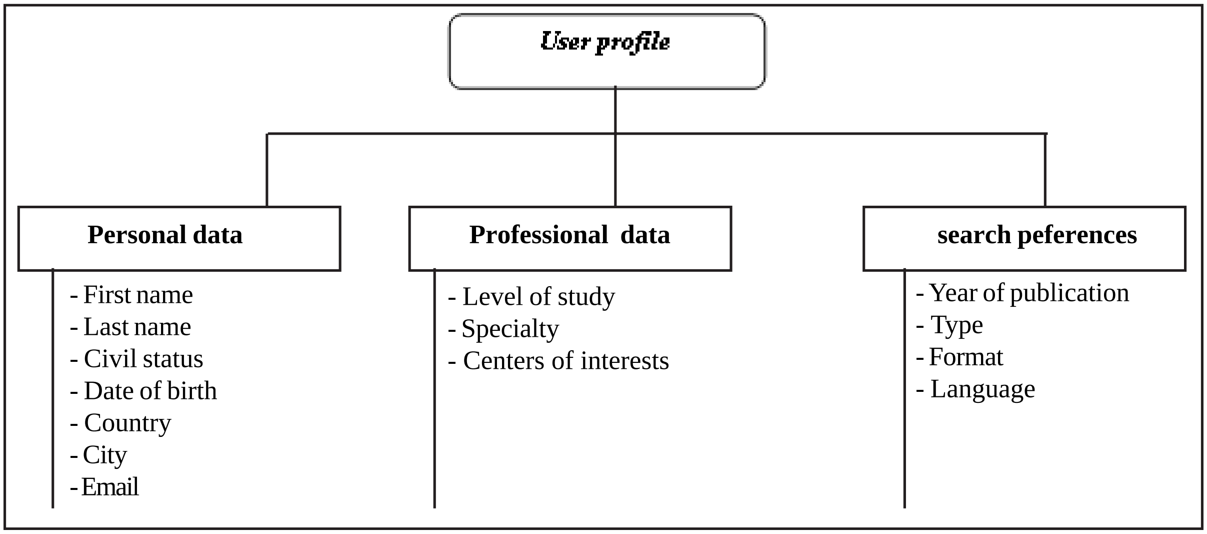


Figure 1. User profile model in PIRSIDIL

2.2 Metadata

Metadata is by definition "a data about data, specifically, it is a structured set of information describing any resource" [12]. To make the data usable by others, it must register in models widely recognized by actors in the Web. The metadata concept existed before the advent of the Internet but its interest grew with the number of electronic publications and digital libraries. The solution proposed by the World Wide Web Consortium (W3C) is to use the metadata to describe data in order to facilitate information retrieval. They guarantee interoperability by sharing and exchange of information that make the content readable and understandable by machines.

Several organisms of standardization proposed and published metadata diagrams susceptible to be used by the largest number of users. The most used metadata diagram is proposed by the organization Dublin Core Metadata Initiative (DCMI) [13]; it is often called the Dublin Core. The Dublin Core aims since its creation to solve the problem of the unified description to the electronic resources of information. It becomes a standard ISO 15836 since February 2003.

The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description that are: title, creator, subject and keywords, description, publisher, date, type, format, language, identifier, contributor, source, relation, coverage, rights. The real interest of these metadata is the addition of semantic content of nature to the documents published, which greatly increases the quality of information available and improves the relevance of results for the research tools.

To build our model of metadata that will be used in PIRSIDIL, we adopted the model of Dublin Core because of its simplicity, with which we can modify some elements and we adapted it to our needs. We retained ten of properties from Dublin and we modified the properties of some elements that are: title, Author, keywords, description, publisher, date (date of publication), type, format, language, identifier. In addition; we added 4 other properties that are: size, level, discipline and under discipline in order to adequately characterize the database resources of our digital library.

3. Proposal of a method of matching

To search for documents, users filled up a form of multi-criteria query. At least one of fields "keywords", "title" or "author" is filled up by the user, the other optional fields are: Year of publication, type, format and language, the system compares it's with those metadata of documents available in the database of digital library . The objective in this first level matching is to eliminate documents that do not correspond to the preferences of the user in his research. Thus we define nine criteria for comparison to establish the matching query metadata E(Q) in Table 2.

For criteria C_1, the system verifies the existence of query keywords in the metadata "keywords", "description" and "title" of the documents, so it gives the note 1 in the case of existence of keywords in any of these metadata; otherwise it gives the note 0. For the remaining criteria, the system makes a Boolean comparison providing a note 0 or 1.

C_i	Query	Metadata
C_1	Keywords	Keywords, Title, Description
C_2	Title	Title
C_3	Author	Author
C_4	Year of publication	Date (Year)
C_5	Type	Type
C_6	Format	Format
C_7	Language	Language

Table 2. Element of comparison for E(Q)

The matching query metadata E(Q) will be calculated in the formula (1) by the product of the note affected to each evaluation criteria (C_i) specified in Table 2.

$$E(Q) = \prod_{i=1}^{7} E(Ci) \qquad (1)$$

The user can activate or deactivate his profile. The profile can be activated for a research centered on the domain and the centers of interests of the user or can be deactivated to widen the field of research.

- If the query is formulated without taking into account of the user profile, the documents selected through the matching query metadata E(Q) represent the final result.

- If the query is formulated by taking into account of the user profile, we integrate the matching user profile metadata E(P) in the calculations.

For establishing the matching E(P), we will associate the user profile elements that we identified through a questionnaire that we have made references to the services of the library of the faculty of economics and management of Sfax, the results of this questionnaire allowed us to classify the user profile elements in order of importance:

Specialty or the domain: (40% positive responses)
Centers of interests: (35% positive responses)
Level of study: (25% positive responses)

So, we define in Table 3 three criteria of comparison for E(P).

C_i	User profile	Metadata	Weight(Wi)
C_1	Specialty	Discipline	40%
C_2	Centers of interests	Under discipline	35%
C_3	Level of study	Level	25%

Table 3. Element of comparison for E(P)

The matching user profile metadata E(P) in the formula (2) is evaluated by the sum of the scores affected to each criteria weighted by the weights clarified in Table 3.

$$E(P) = \sum_{i=1}^{3} Wi * E(Ci) \qquad (2)$$

The results of search will be ordered according to the ordering of the centers of interests of the user.

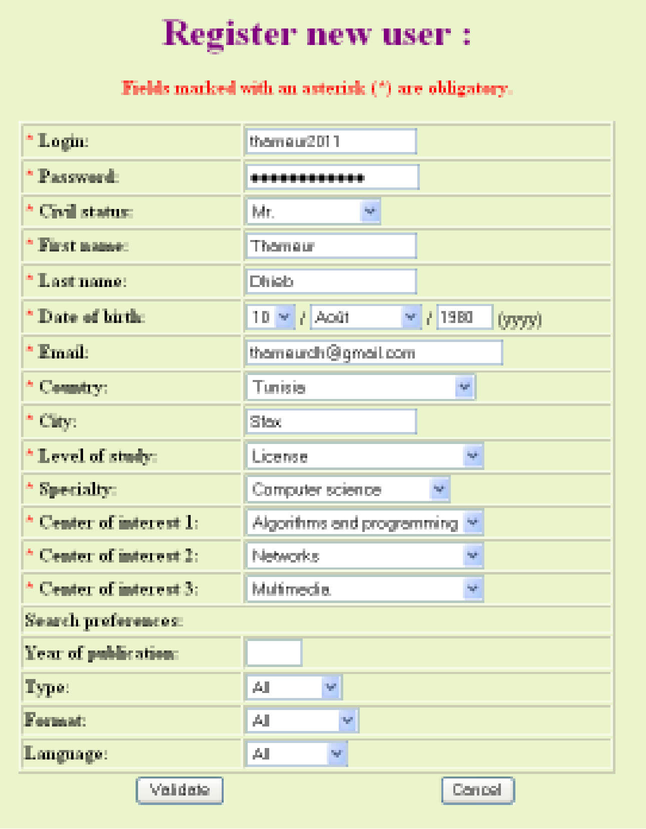
4. Implementation

Our system was created by EasyPHP v5.3.1 (Apache Server, PHP language and MySQL data base) which allows hosting it on several platforms and multiple operating systems. The choice of this technology is justified by the availability. It has proven itself in terms of robustness and compatibility.

During the enrollment phase system, the user must fill up a form by entering his personal data, his professional data and his search preferences.

When the user connects to the system, he has the possibility to update his user profile, to search documents or to host resources to feed the database of our digital library while adding the metadata for each document.

In PIRSIDIL, research is multi criteria and takes place by the seizure of the keywords. Thus, the user can define many filters of research as the type, the format, the language, etc. In the filling up of research form, the user can activate or deactivate his profile.

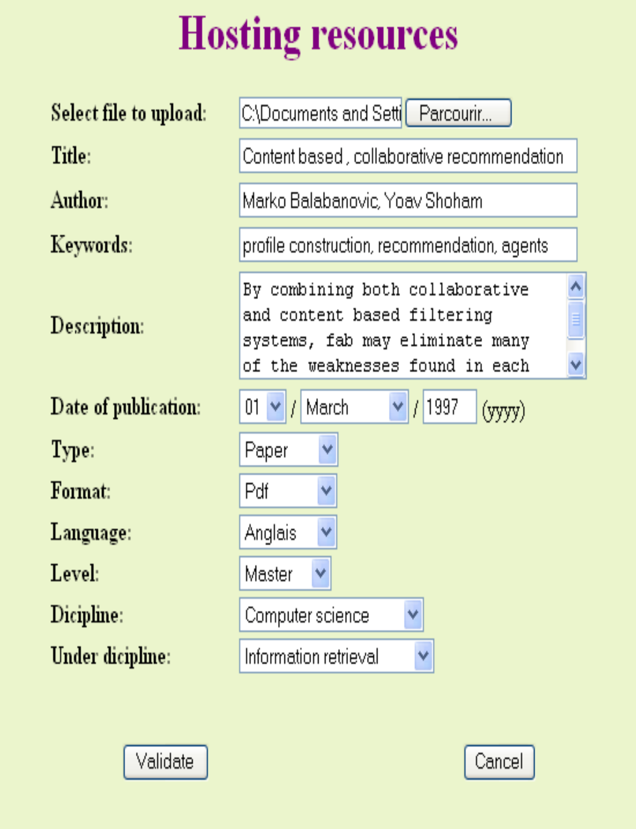


Register new user :

Fields marked with an asterisk (*) are obligatory.

* Login:	thameur2011
* Password:	*************
* Civil status:	Mr.
* First name:	Thameur
* Last name:	Dhieb
* Date of birth:	10 / Août / 1980 (yyyy)
* Email:	thameurdh@gmail.com
* Country:	Tunisia
* City:	Sfax
* Level of study:	License
* Specialty:	Computer science
* Center of interest 1:	Algorithms and programming
* Center of interest 2:	Networks
* Center of interest 3:	Multimedia
Search preferences:	
Year of publication:	
Type:	All
Format:	All
Language:	All

Figure 2. Register new user



Hosting resources

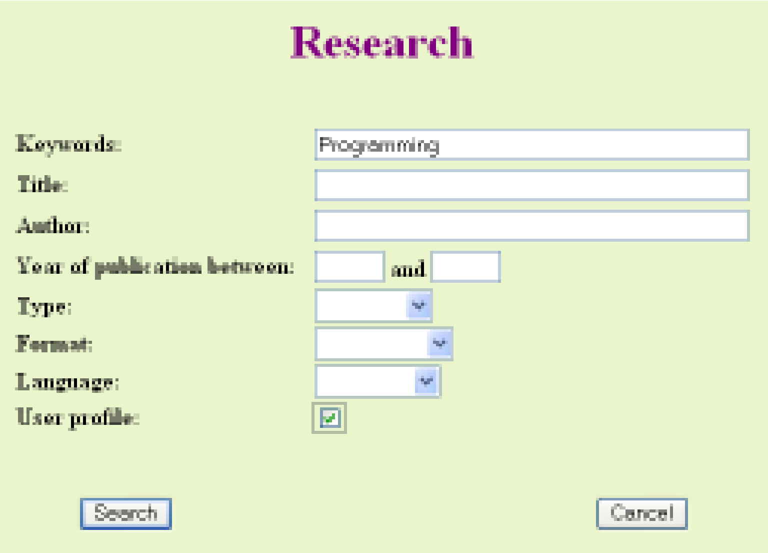
Select file to upload:	C:\Documents and Setti
Title:	Content based , collaborative recommendation
Author:	Marko Balabanovic, Yoav Shoham
Keywords:	profile construction, recommendation, agents
Description:	By combining both collaborative and content based filtering systems, fab may eliminate many of the weaknesses found in each
Date of publication:	01 / March / 1997 (yyyy)
Type:	Paper
Format:	Pdf
Language:	Anglais
Level:	Master
Dicipline:	Computer science
Under dicipline:	Information retrieval

Figure 3. Hosting resources

PIRSIDIL has been tested by users with two modes of research (with and without activation of user profile) We will take an example of research with the registered user shown in Figure 2. Their professional data are:
specialty: computer science, center of interest 1: algorithms and programming, center of interest 2: networks, center of interest 3: multimedia, level of study: license.

We will focus more specifically for research with keywords, as this option allows us to better highlight the advantages of our system. If the user searches for the keyword “programming”, it will have the following results:

- First case: The user does not activate his user profile in his query of research : The first ten documents returned by PIRSIDIL shown in Figure 5.



Research

Keywords: Programming

Title:

Author:

Year of publication between: and

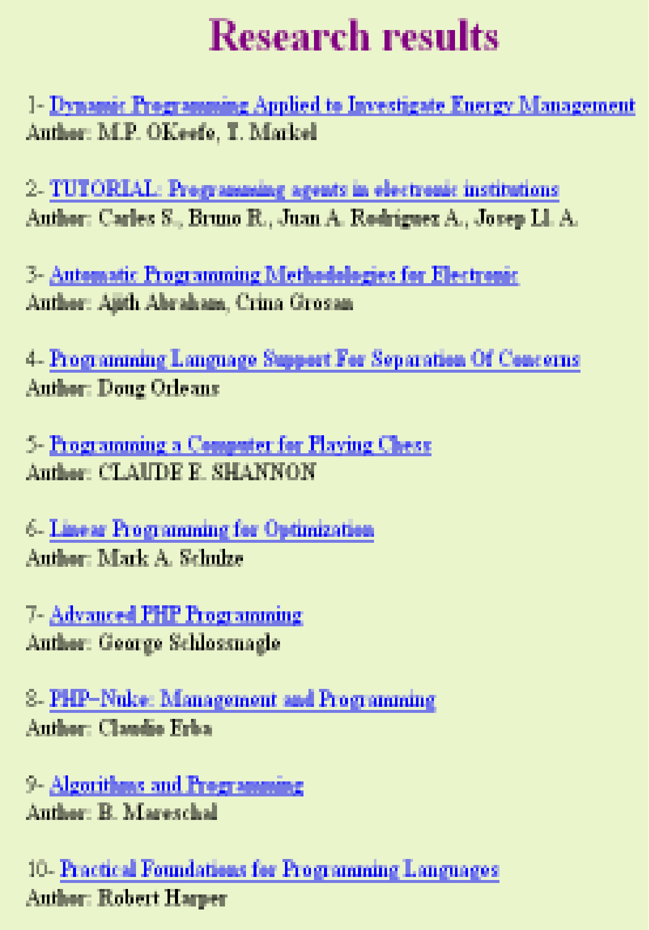
Type:

Format:

Language:

User profile: ☑

Figure 4. Research form



Research results

1- Dynamic Programming Applied to Investigate Energy Management
Author: M.P. OKeefe, T. Markel

2- TUTORIAL: Programming agents in electronic institutions
Author: Carles S., Bruno R., Juan A. Rodriguez A., Josep Ll. A.

3- Automatic Programming Methodologies for Electronic
Author: Ajith Abraham, Crina Grosan

4- Programming Language Support For Separation Of Concerns
Author: Doug Orleans

5- Programming a Computer for Playing Chess
Author: CLAUDE E. SHANNON

6- Linear Programming for Optimization
Author: Mark A. Schulze

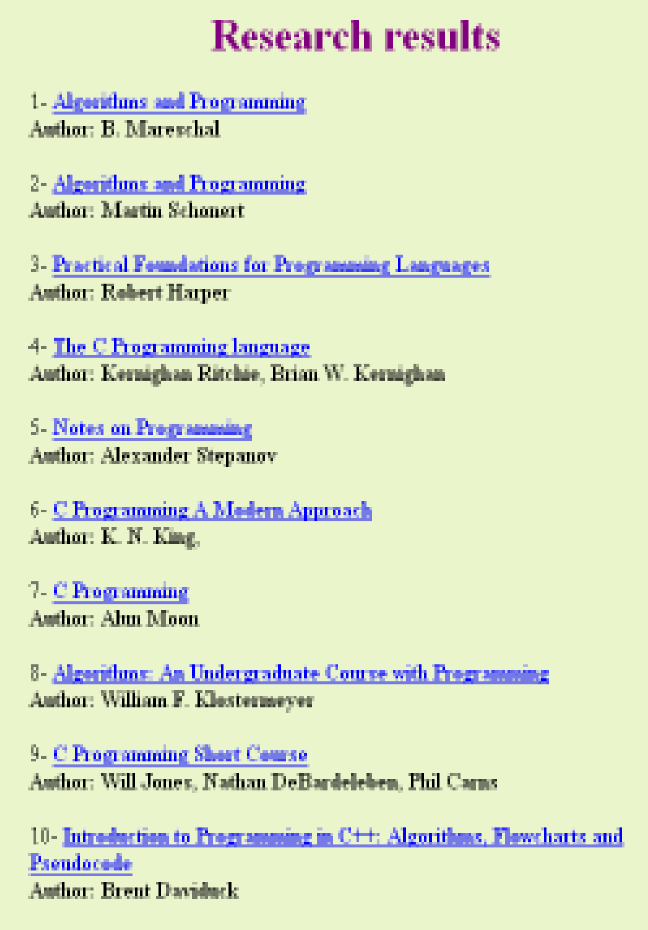
7- Advanced PHP Programming
Author: George Schlossnagle

8- PHP-Nuke: Management and Programming
Author: Claudio Erba

9- Algorithms and Programming
Author: B. Mareschal

10- Practical Foundations for Programming Languages
Author: Robert Harper

Figure 5. Research results



Research results

1- Algorithms and Programming
Author: B. Mareschal

2- Algorithms and Programming
Author: Martin Schonert

3- Practical Foundations for Programming Languages
Author: Robert Harper

4- The C Programming language
Author: Kernighan Ritchie, Brian W. Kernighan

5- Notes on Programming
Author: Alexander Stepanov

6- C Programming A Modern Approach
Author: K. N. King,

7- C Programming
Author: Alun Moon

8- Algorithms: An Undergraduate Course with Programming
Author: William F. Klostermeyer

9- C Programming Short Course
Author: Will Jones, Nathan DeBardeleben, Phil Carns

10- Introduction to Programming in C++: Algorithms, Flowcharts and Pseudocode
Author: Brent Daviduck

Figure 6. Research results

In this case, the system returned a list of all documents containing the word « programming » in the title, in the description or in the keywords metadata of documents. If we look more precisely to the first ten documents returned, we find that the first, the second and the third documents dealing with the electronic domain, the fourth, the fifth and the sixth documents dealing with the computer science domain, but on the doctoral level. The seventh and the eighth dealing with the computer science domain, on the license level but touching the under discipline multimedia, the ninth and the tenth dealing with the computer science domain, on the license level with the under discipline algorithms and programming.

All this means that the user who wants to search documents related to his professional data and he deactivate his profile in his query of research will get irrelevant documents, in this example, the first document that is relevant for the user located in the ninth result.

- Second case: The user activate his user profile in his query of research

The first ten documents returned by PIRSIDIL shown in Figure 6.

The first ten documents returned dealing with the computer science domain, on the license level and with the under discipline algorithms and programming. So, in this case, the user finds all these documents relevant for him.

To validate our system, we created hundreds of annotated resources. The evaluation process is to compare the search results provided by Google and our system PIRSIDIL with activation of user profile for the same information need. Then, we calculated separately for the 50 first responses provided by Google and PIRSIDIL the number of relevant documents and the number of irrelevant documents. This is interesting since we know that users of information systems generally only explore the first responses which have been provided [14]. Our goal is to determine and compare between Google and PIRSIDIL the recall and precision rates.

We have found in the tests that the results provided by our system are generally better quality than those provided by Google. The shape of the precision-recall curves obtained from the basis of 5 search sessions shown in Figure 7.

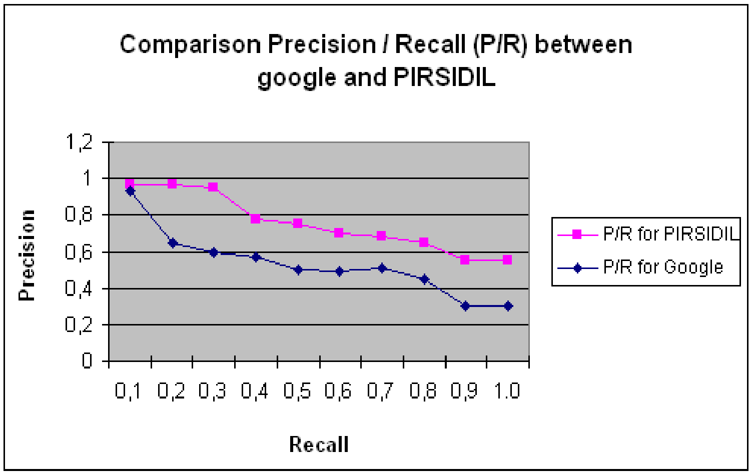


Figure 7. Comparison Precision/Recall between Google and PIRSIDIL

We note that for the first documents in the list returned by our system, recall/precision rates are close to 1.
Thus, these experimental studies have allowed us to demonstrate the impact of the personalization on the performance of results.

5. Conclusion

We presented in this paper our personalized information retrieval system in a digital library which we constructed the user profile model and we represented the metadata of documents and then we incorporated a method of matching user profile with metadata. Our system presents of the user aware the answer documents relevant and responsive to their centers of interests and their preferences of research.

The first results to validate our system show an encouraging improvement in the quality of responses to the user. Moreover, as our sample of users is not sufficiently representative, our current work will extend the existing tests on a larger number of users, clearly differentiated profiles and conduct a comparative evaluation with other personalized information retrieval systems given by other authors.

The perspective of improvement of our system is located in the integration of ontology in our method of matching and the sharing of profiles between users within a collaborative work.

References

[1] Mianowska, B., T Nguyen, N. (2011). A Method for User Profile Adaptation in Document Retrieval. *In*: Proc of the Third Asian Conference on Intelligent Information and Database Systems. (ACIIDS 2011), p. 181-192. Lecture Notes in Computer Science Volume 6592/2011.

[2] Budzik, J., Hammond, K. (2000). Users interactions with everyday applications as context for just in- time information access, *In*: Proc of the 5th international conference on intelligent user interfaces, (IUI 2000), p. 41-51.

[3] Benammar, A., Hubert G., Mothe, J. (2003). Proposition à l'intégration des profils dans le processus de recherche d'information, *International Journal of Info & Com Sciences for Decision Making.*

[4] Zemirli, N. (2008). Modèle d'accès personnalisé à l'information basé sur les diagrammes d'infuence intégrant un profil utilisateur évolutif, in the thesis submitted for obtaining the degree of Doctor of Computing, University Paul Sabatier.

[5] Daoud, M. (2009). Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche, in the thesis submitted for obtaining the degree of Doctor of Computing, University Paul Sabatier.

[6] Girardi, R., Leite, A. (2010). The Specification of Requirements in the MADAE-Pro Software Process, *Journal of Information Systems*, (ISYS 2010) 3.

[7] Lu, Y., Sebe, N., Hytnen, R., Tian, Q. (2011). Personalization in multimedia retrieval: A survey, *Multimed Tools Appl*. 51. 247–277

[8] Salton, G. (1971). The SMART Retrieval System, Experiments in Automatic Document Processing, in Prentice-Hall Inc, NJ.

[9] Gowan, J. (2003). A multiple model approach to personalised information access, in the master thesis in computer science, Faculty of science, University of College Dublin.

[10] Sieg, A., Mobasher, B., Lytinen S., Burke, R. (2004). Using Concept Hierarchies to Enhance User Queries in Web-based Information Retrieval , *In*: the Artificial Intelligence and Applications(AIA 2004).

[11] Challam, V., Gauch, S., Chandramouli, A. (2007). Contextual Search Using Ontology-Based User Profiles, *In*: Proc of the RIAO 2007, Pittsburgh USA.

[12] Peccatte, P. (2007). Métadonnées: une initiation. Dublin Core, IPTC, Exif, RDF, XMP, etc., http://peccatte.karefil.com/software/Metadata.htm

[13] Dublin Core Metadata Initiative website. (2010), http://www.dublincore.org/documents/2010/10/11/dces/

[14] Jansen, B.J., Spink, A., Bateman, J.Q., Saracevik, T. (1998). Real Life Information Retrieval: A study of User Queries On the Web, SIGIR Forum, 32 (1) 5-17.