Diagnostic evaluation of a personalized filtering information retrieval system. Methodology and experimental results

Christine Michel

Diagnostic Evaluation of a personalized filtering information retrieval system. Methodology and experimental results.

Christine MICHEL
Laboratoire CEM-GRESIC
MSHA - Esplanade des Antilles, D.U
33607 PESSAC Cedex - FRANCE
Tél : +33 (0)5 56 84 68 13/ 68 14
Fax : +33 (0)5 56 84 68 10
Christine.Michel@montaigne.u-bordeaux.fr

Study made in the laboratory RECODOC (University Claude Bernard Lyon I – FRANCE)

Abstract

The study presented in this paper deals with the diagnostic evaluation of a system being implemented. The tested system's particularity is to provide a filtering process taken into user's account personal characteristics. The aim of diagnostic evaluation is to choose one filtering process between 8 proposed ones. 16300 interrogations are used as a representative sample. It combines characteristics relating to: the user's profile, the user's need of information and the filtering process. Answers are compared relating to: the number of common documents, the rank of common documents and the specificity degree of the query. These criteria give indication about the filtering impact.

Introduction

Hirschman et al (Hirshman 95) distinguish three evaluation types. The **adequacy evaluation** determines the fitness of a system for a purpose. The **diagnostic evaluation** is the production of a system performance profile with respect to some "taxonomisation" of the space of possible inputs[1]. The software engineering teams also uses it to compare two generations of the same system (regression testing). The **performance evaluation** is the measure of the system performance in one or more specific area. It's typically used to compare like with like between two alternative implementations of technology. In *"information retrieval itself, a classic criterion is precision …/…, a measure is the percentage of document retrieved with are in fact relevant…/…, and a method for computing is to simply average over some number of test queries the ratio achieved by the system under test."*

The TREC well-known experiments are performance evaluation. *"One of the goal of TREC is to provide task evaluation that allows cross systems comparison which has proven to be the key strength in TREC. ../… The addition of secondary tasks (called tracks) in TREC-4 combined these strengths by creating a common evaluation for retrieval sub problems" (Voorhees 98)*. The methodology presented here is half a diagnostic and performance evaluation. It allows quick auto evaluation of information retrieval systems during the conception step. The test aim is to quantify the stability or the reactivity of the system submitted to different personalized filtering criteria. The tested system is often just a prototype so real users with personal information need can't make direct interrogations. Those interrogations have to be simulated in laboratory. We consider the system as a black box submitted to different contexts of information. Protocols recommended in those cases are purely quantitative in order to have an exact control on the variables. Each particular component is isolated and observed on how it modifies the system's answers.

[1] It's typically used by system developers, but sometimes offered to end-users as well. It usually requires the construction of a large and hopefully representative test suite.

According the diagnostic evaluation, we have made a "taxonomy" of the possible input i.e. different types of users in search (users are represented by a specific profile and specific information need). Nevertheless it is also a performance evaluation because for the same systems we compare 8 different filtering algorithms by a performance criterion: the filtering impact i.e. the degree of similarity between the neutral answer without any filtering and the filtering answer.
The system, tested in our experiment, presents answers in a **ranked clustering way**. In the first part of this paper we present which alternatives are proposed to experimenters in this case. In the second part we present the tested system: Profil-Doc, in the third part the experimental protocol and in the last part the results.

1 The ranked clustering answer as a problem in evaluation.

Clustering process is used to improve the visibility of an information set. *"Document clustering has long been investigated as a post retrieval document visualization techniques. Document clustering algorithms attempt to group documents together based on their similarities …/… This can help users both in location interesting document more easily and in getting an overview of the retrieved document set". "Information Retrieval community has long explored a number of post-retrieval document visualization techniques as alternatives to the ranked list presentation …/… : document networks, spring embeddings, documents clustering, and self organizing map. Of the four major techniques, only document clustering appears to be both fast enough and intuitive enough to require little training or adjustment time from the user."* (Zamir, 99) In our case, Profil-Doc via SPIRIT[2] uses a clustering process and also ranks the different cluster by order of relevance. As Kantor said *"Clusters of documents as clusters of terms represents concepts. While each document no doubt contains many concepts, the cluster will rank some concepts more highly"* (Kantor, 94). The SPIRIT's ranking method is presented in (Fluhr, 84).
For example, the answer given by SPIRIT for the query "*large-scale system evaluation*" is composed of 104 documents into 12 clusters, ranked by order of relevance (cf table 1).

Cluster rank	Cluster name	Document reference	Cluster document number
1	system-evaluation-large-scale	docu462	1
2	system-evaluation, scale	docu104	1
3	system, evaluation, large, scale	docu264, docu457	2
4	evaluation, large, scale	docu262, docu263 ; docu265	3
5	evaluation, large	docu199	1
6	system, large, scale	docu259, docu456, docu459 ...	5
7	system, large	docu458	1
8	system, evaluation, scale	docu36, docu29, docu317	12
9	system, evaluation	docu49, docu288, docu318 ...	4
10	large, scale	docu261, docu463	2
11	evaluation, scale	docu213, docu245 ...	7
12	system, scale	docu230, docu211, docu196 ...	65
12	system, scale	docu230, docu211, docu196 ...	

Table 1 : Example of ranked clustering answer

According to Tague (Tague 95, Fricke 98) the documents' rank of presentation is one of the five

[2] SPIRIT (Syntactic and Probabilistic Indexing and Retrieval Information System) is a commercial product of T.GID. Searches about SPIRIT are made according to the CEA-DIST (Commissariat à l'Energie Atomique - Scientific and Technique Information Direction) – http://www.dist.cea.fr/

aspects to take into account to evaluate the quality of a SRI or of an information research center. Indeed *"algorithmically ranked retrieval results become interpreted and assessed by users during session time. The judgment is in accordance with the users' dynamic and situational perceptions of a real or simulated information retrieval task"* (Borlung 98). Tague quotes many studies that highlight the delay induced in the satisfaction of an informational need, induced by a possible modification of this order of presentation (Tague 96).

Measures of Recall, Precision, Jaccard, Cosine, … used in the case of collection test evaluations or comparative test protocols do not take the documents' order of presentation into account. They are the results of intersections or unions of the compared sets.

In the trec_eval package of TREC 7 (Voorhees 98) report, we can find several adaptations of Recall and Precision, used when systems return a ranked list of documents. There is 85 numbers per run in the trec_eval package. For example **P(10)** is the precision after the first 10 documents are retrieved, **P(30)** is the precision after the first 30 documents are retrieved, **R-Prec** is the precision after the fist R documents are retrieved, where R is the number of relevant documents for the current topic, **mean ave precision** is the mean of average precision, **R(1000)** is the recall after 1000 documents are retrieved, **rank first rel** is the rank of the first relevant document retrieved. We can see in this example that several measures are roughly the same, they are varying according to the cut off level. A statistical study on 8 measures (Voorhees 98) shows that several are correlated i.e. measure the same things. This example notices the importance of a **good and global evaluation measure**.

We propose (Michel 99) new mathematical methods and formalisms allowing us to build measures of proximity taking the documents' rank of presentation into account. We call them **OS measures**, i.e. measures of Ordered Similarity. The experiment presented in section 3 proposes an example of the OS measure.

2 The tested system

Profil-Doc is a full-text information retrieval system made specifically for searcher from scientific and technical information fields. *"Its aim is to carry out a pre orientation toward an information corpus restricted to user relevant information determined with the aid of utility criteria"* (Laine 96). The pre-orientation system includes three fundamental operations:

- **a characterization of the user's profile** relatively to four criteria : educational level (student, doctor, confirmed searcher) , disciplinary field (information science, mathematics, …), search stage (state of the art, definition of a subject, experimentation, discussion …) and type of search (specific or general)
- **a segmentation of texts to be processed into part of text** relatively to three criteria: the type of part (resume, introduction, experimentation, .), the discursive form of the part (argumentative, descriptive, …) and the format of presentation (text, equation, image, .). Characterized parts of texts are called "documentary units". The description format of the database is a part of the system ; it's so specific that it forbids us to use a classical large-scale test collection as in the TREC example.
- **a filtering process** that selects the useful parts for the identified user. 8 different filtering algorithms are under evaluation.

From a specific user's profile, the filtering process chooses the usefulness properties of the part of text. They made the extraction of a personalized corpus possible. *"Once the "personalized" corpus has been defined, documentary software can be used to implement a classical search procedure to process user queries"*(Laine, 96) The chosen documentary system is SPIRIT, a full-text and natural language querying system. As we already said, SPIRIT ranks answers texts in cluster in function of concepts of query it treats. The higher ranks are the more pertinent to the query clusters are. We work with the SPIRIT-W3 version, i.e. "*SPIRIT databases could be carried through a standard browser."* (Fluhr 97). So the system is presented as a Web server. The tested prototype is composed of a 505 documentary units database, and a non-ergonomic interrogation interface i.e. the user can't

directly give his profile for the filtering process. He has to give manually the type of parts of text he wants. So, Profil-doc interrogations are composed of **factual criteria** defining properties of the useful part of text for the user's profile and **textual query** defining user's information need.

3 Evaluation Methodology

As shown in the figure 1, the protocol is composed of 3 steps:
First, as Hirschman (Hirschmann, 95) said, there is a "taxonomy" of the system inputs, i.e. the interrogations. We have to compare 8 different filtering algorithms. We characterize a context of interrogation by a specific user's profile and a query.
The system querying is made automatically, and answers are compared in order to evaluate the filtering impact. It's the degree of similarity between answers obtained with a filtering process and without (i.e. neutral answers).

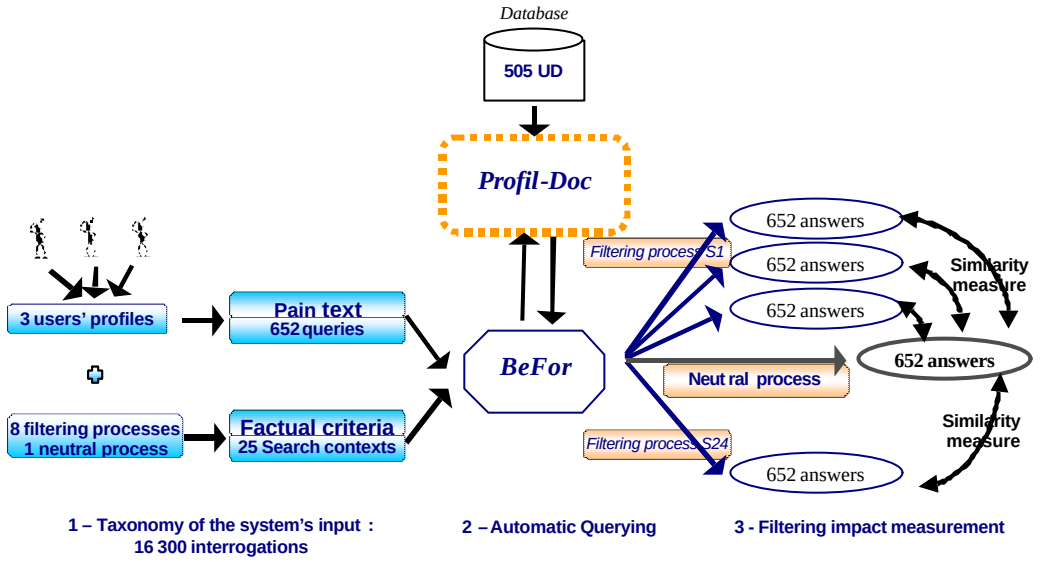

Figure 1 : Evaluation methodology

3.1 Taxonomy of the system's inputs

a) Construction of large-scale plain text query corpus

We chose to construct automatically a large-scale query corpus representing the most database concepts in order to limit as much as possible the influence of each query. The process is not really original i.e. a statistical random choice of sentences from original texts of database.
The database is composed of about 7000 sentences. We extract 10% of them chosen at random. Quick reading permits to suppress some words in order to make the sentence looks like query (else the sentence is suppressed). At the end we have **652 queries**. This random extraction produces artificial queries which are less or not general: for example the query *"University formation program"* is very general and have 216 answer documents, and query *"Polymer paint color journal"* is very specialized and have less than 5 answer documents (Michel 99). We choose to analyze our query corpus under this criterion in order to control a hypothetic bias. As we can see in this example, the number of query's words is not a good indication of the specificity degree. So we directly rank queries in a decreasing way in function of the number of documents answers have when the system is used in a neutral way.
In the figure 2 we can observe the corresponding distributions.
We can notice that they are following a Zipfian low. A Zone represents general queries where answer have more than 150 documents. In opposition, B Zone represents specialized queries with less than 20

documents in answer. Regarding the regularity of the curve we must suppose that this query corpus will not induce any bias in the experimentation.

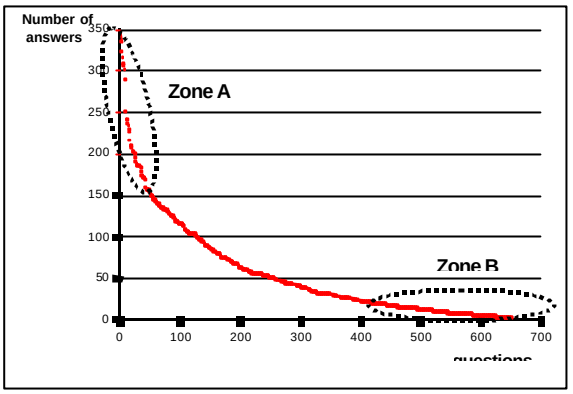

Figure 2 : The number of answered documents by query

b) Simulation of filtering process: construction of 24 contexts of search

We choose to present three different user's profiles. P1 is a student of the information science domain, who needs a very precise information to make a state of the art. P2 is a researcher in the biology domain who needs information to generalize his results to other domains. The last user P3 is a researcher in information science domain who needs a very precise information to finish the redaction of an article. These three situations are very different in terms of information utility.

Each user tests the 8 filter algorithms. Considering the three user's profiles, we have 24 different personalized filtering searches defined by various factual criteria corresponding to user's useful part of text properties.

We call S1, S2, S3, … and S8 the personalized search of profile P1,
S9, S10, … and S16 the personalized search of profile P2 and
S17, S18, … and S24 the personalized search of profile P3.

Filter	P1	P2	P3
1	S1	S9	S17
2	S2	S10	S18
3	S3	S11	S19
4	S4	S12	S20
5	S5	S13	S21
6	S6	S14	S22
7	S7	S15	S23
8	S8	S16	S24

Table 2 : The search strategies 1 to 24

We submit the 652 queries to the system in these 24 contexts, and compare the results with the neutral ones (I.E. results without any filtering). This means more than 16000 interrogations to do. It's quite unrealistic in a manual way[3].

[3] To make some comparison, in the last TREC 7 the participants have only 50 news natural language topic statement (and 350 derived from previous TREC experiments).

3.2 Automatic Querying of the system

The application "BeFor" developed by the CEA (Charron, 99) is used to query search engines on special topics in order to make appear the "invisible Web" and store not only HTML files but also dynamic web pages created from database. We use its initial function to make large-scale automatic querying of our system.

XML is used to formalize queries and answers. Two XML files are used to make a description of the structure and the content of queries submitted to the system. The XML file used to represent our query structure is:

```
<?XML VERSION=« 1.0 »?>
<!DOCTYPE SEARCH_PROFILES SYSTEM « search_profiles.dtd »>
<SEARCH_PROFILES>
<SEARCH_ENGINE>
<!--spirit/W3/totale-->
<ENGINE_NAME>Spirit-W3(totale)</ENGINE_NAME>
<URL METHOD=« post »>
    <PROTOCOL>http</PROTOCOL>
    <SERVER>www.recodoc.univ-lyon1.fr</SERVER>
    <PORT><PORT>
    <FILE>spirit/...../traitquest</FILE>
    <PARAMATER><NAME>T:TITRE,TITRE_REVUE,AUTEUR,COAUTEUR,AFFILIATION,TEXTE.</NAME><TYP
E> Query </TYPE></PARAMATER>
    <PARAMATER><NAME>F:PROFESSION,DISCIPLINE,COMMUNAUTE.</NAME>
       <TYPE>Production</TYPE></PARAMATER>
    <PARAMATER><NAME>F:ENV_EDITO.</NAME><TYPE>Diffusion</TYPE></PARAMATER>
    <PARAMATER><NAME>F:TYPE_UD,FORME_DISC,STYLEF.</NAME><TYPE>Unite</TYPE></PARAMATER>
    <PARAMATER><NAME>F:TYPE_UD.</NAME><TYPE> UD_Type </TYPE></PARAMATER>
    <PARAMATER><NAME>F:FORME_DISC.</NAME><TYPE>Disursive_Form</TYPE></PARAMATER>
    <PARAMATER><NAME>F:STYLEF.</NAME><TYPE>Presentation_Style </TYPE></PARAMATER>
    <PARAMATER><NAME>idbase</NAME><VALUE>totale</VALUE></PARAMATER>
</URL>
</SEARCH_ENGINE>
</SERCH_PROFILES>
```

<SEARCH ENGINE>, <URL>, <METHOD>, <PROTOCOL>, <SERVER>, <PORT> correspond to technical interrogation formats. <FILE> is the query file address and <PARAMETERS> is the query description with names of each queries fields (<TYPE>) and the corresponding querying SPIRIT fields (<NAME>). The field **query** corresponds to the textual query and the fields **production, diffusion, …presentation_Style**, to the factual criteria.

The corresponding effective interrogations are stored in an XML file as follows:

```
<?XML VERSION=« 1.0 »?>
<!DOCTYPE QUERIES SYSTEM « query.dtd »>
<QUERIES>
<QUERY>
    <NUMBER>1</NUMBER>
    <TYPE>query</TYPE><VALUE>large-scale system evaluation</VALUE>
    <TYPE> UD_Type </TYPE><VALUE> resume, introduction </VALUE>
    <TYPE> Disursive_Form </TYPE><VALUE>descriptive</VALUE>
    <TYPE>Presentation_Style</TYPE><VALUE>text</VALUE>
</QUERY>
<QUERY>
    <NUMBER>2</NUMBER>
    <TYPE>query</TYPE><VALUE> large-scale system evaluation </VALUE>
    <TYPE> UD_Type </TYPE><VALUE>development, experimentation</VALUE>
```

<TYPE> **Disursive_Form** </TYPE><VALUE>*argumentatives*</VALUE>
<TYPE>**Presentation_Style**</TYPE><VALUE>*text with numeric data*</VALUE>
</QUERY>
<QUERY>
<NUMBER>3</NUMBER>
<TYPE>**query**</TYPE><VALUE>*plain text information retrieval systems*</VALUE>
<TYPE> **UD_Type** </TYPE><VALUE> resume, introduction </VALUE>
<TYPE> **Disursive_Form** </TYPE><VALUE>descriptive</VALUE>
<TYPE>**Presentation_Style**</TYPE><VALUE>text</VALUE>
</QUERY>
</QUERIES>

Query 1 and 2 represent the same content but not the same purpose. Queries A and 3 have the same purpose but not the same content.
This succession of queries is given to "BeFor" which automatically queries Profil-Doc through its web interface and answers are given back in an XML format.

3.3 Calculus of the filtering impact

a) Comparison of answer sets: The P_d Ordered Similarity measure

Each answer is composed of a succession of cluster, each cluster has a rank of presentation and a documents list. Let's use the formalism presented in (Michel 99). If C and C' are the answers to be compared, m and m' the number of classes of C and C'. C_i and C'_j are the clusters (class) of C and C', i and j varying between 1 to m and 1 to m'.
Then *Pd* is defined by :

$$P_d(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} J(C_i, C'_j) \times \delta^{mo}\left(i\left(|i-j|+1\right)\right) \times \delta^{mo}\left(j\left(|i-j|+1\right)\right)$$

With $J(A,B) = \dfrac{|A \cap B|}{|A| + |B| - |A \cap B|}$.(J is the Jaccard's indices), and δ^{mo} a function from $[1, m_0^{\,2}]$ to IR^+ defined by :

$$\delta^{mo}(n) = \frac{\sqrt{6m_0^{\,3}}}{\sqrt{6m_0^{\,4} - 6m_0^{\,3} + 8m_0^{\,2} - 3m_0 + 1}}\left(1 - \frac{n-1}{m_0^{\,2}}\right).$$

The function δ^{mo} is called the "delay" function. It measures the delay of reading due to the difference in document's order of presentation between answers C and C'. δ^{mo} is a decreasing function so the smaller i and j are (i.e. classes of documents are in the first places), the more they will be taken into account in the similarity measure. And the more distant i and j are, the less the corresponding class will be take into account in the similarity measure.

b) Results interpretation

For each of the 652 queries, we calculate the similarity *Pd* between neutral answers and filtered answer for S1, S2, … and S24 search contexts. Policies in evaluation experiments consist in general indicators describing the system. In TREC for example, means of measure (recall, precision, …) given in the 50 search contexts results is generally calculated as the latest and final indicator. With this treatment we lost information relating to the system's comportment submitted to general or specialized queries. We choose to analyze curves showing the comparison of answers from less to most specialized queries. A bibliometric study (Michel 99) shows that it's useful to construct **44 groups**. G1 is composed of questions having from 1 to 5 documents per answer[4], G2 of questions having from 6 to 10 documents per answer, …., G43 of questions having from 211 to 215 documents per answer and

[4] The system is used in a neutral way, i.e. without filtering process.

G44 of questions having more than 215 documents per answer.
The curves look like figure 3. In X-Axis there is groups G1, G2, …, G44. In Y-Axis there is the value of *Pd* varying between 0 to 1.

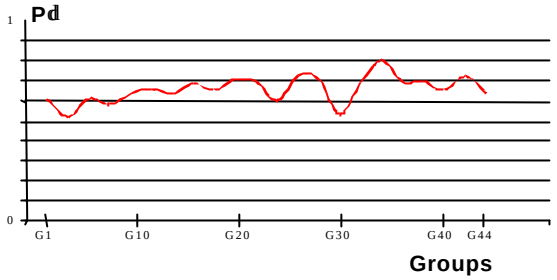

Figure 3 : Pδ results for 44 groups of queries in a search context Sn

It gives a quantitative estimation of the **filtering impact**. The higher *Pd* is, the nearer filtered answers to neutral answers are, so the less influent the filtering is. Conversely, the smaller *Pd* is, the more influent the filtering is.

4 Results

In the following figures we can see the difference of results obtained in the 24-search context. The 24 curves are presented in the 8 following figures : They present each of the 8 filtering types applied to users P1, P2 and P3.

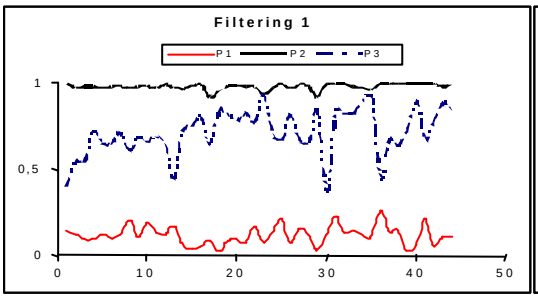

Figure 4 : Filtering 1

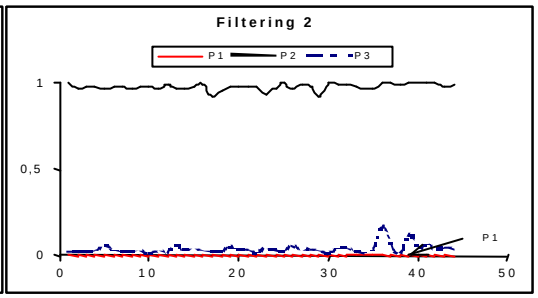

Figure 5 : Filtering 2

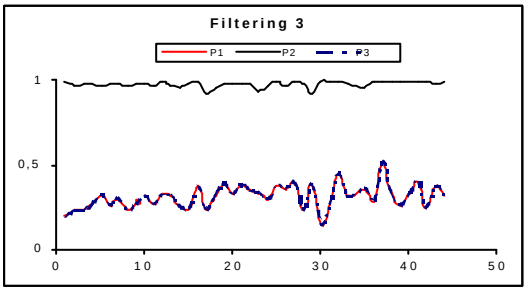

Figure 6 : Filtering 3

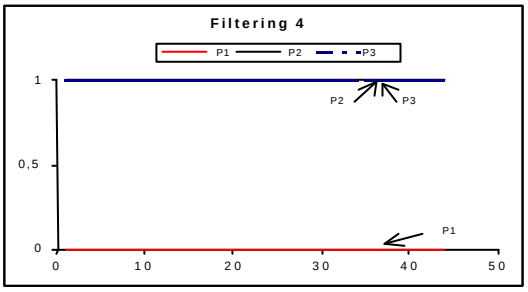

Figure 7 : Filtering 4

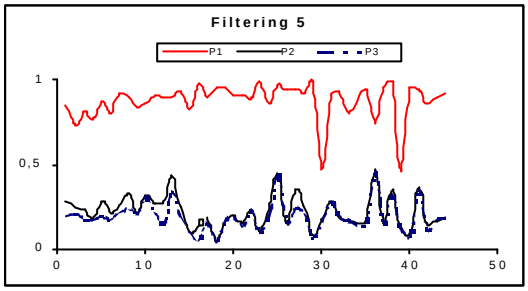

Figure 8 : Filtering 5

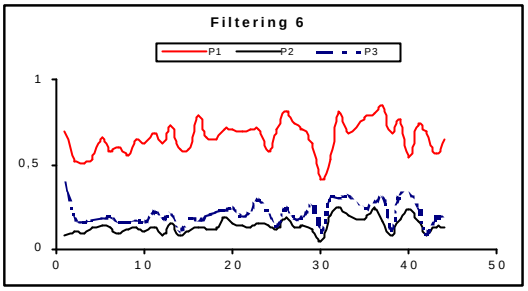

Figure 9 : Filtering 6

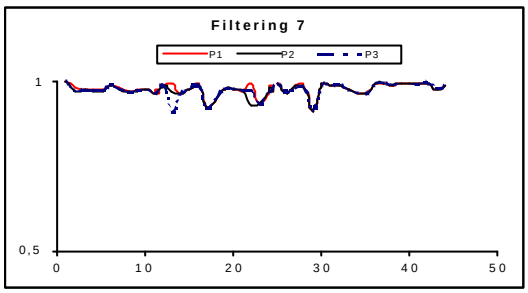

Figure 10 : Filtering 7

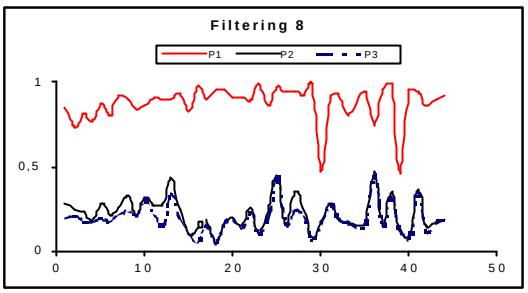

Figure 11 : Filtering 8

The first observation is that there is no specific profile, I mean a very large or very strict profile. For each of them the 8 filtering have a different action. For example, in the case of filtering 1, 2 ,3 ,4 the proximity values are low for profile 1 ,indeed, proximity near to 0 in the case of filtering 1 and 3, and equal to 0 in the case of filtering 2 and 4. In these cases, the filtering is very robust. Conversely, the filtering process for user P2 has no real effect (proximity is near to 1 in the case of 1, 2, 3 and equal to 1 in the case of filtering 4). If we observe now the results of filtering 5, 6, 8 we have the inverse trend : filtering is not influent for P1 and very influent for P2. **So the observed results reflect exactly the action of the different filtering and are not due to a specific user's profile**.

Except in the filtering 1, there is always superposition of two or more curves, meaning that the two or more concerned profiles are treated exactly on the same way by the filter. For example profile 2 and 3 are treated exactly in the same way by filtering 5, 6, and 8; and profile P2 and P3 are treated exactly on the same way by filtering 2 and 3. Filtering 7 is particular, it produces any real filtering for any profile, indeed we have a value near to 1 for each answer.

Curves of filtering 1 show that the segmentation between the different curves is effective and that users are treated individually. Indeed, for each of them the filtering have been more or less influent. Knowing that the observed system must provide a personalized answer regarding the user's profile we could advise the filtering 1.

Conclusion

The protocol presented here permits to highlight which filtering process is the most personalized according to the user's profile. Regarding the results, the filtering process 1 proposed in Doc-Doc appears to be useful.

We try to be as exhaustive as possible in the definition of the system's input. Indeed, the query corpus is representative of the database content. There are not many user's profiles but they are really different in terms of usefulness. The protocol is globally automatic and as less dependent of the domain as possible in order to repeat it as many time as it's necessary and in the various possible fields. The protocols is really open i.e. it can be used in several contexts. For example in the case of a

performance evaluation it's possible to use a large-scale test collection query corpus, compare answers with an OS measure and interpret results in curves.
The criterion of **query degree of specificity** is the first originality of this experiment. A bibliometric study may highlights if there is any bias in the query corpus (for example if the distribution show in figure 2 is not regular).
The second originality deals with the **similarity measurement** employed to compare answers**.** It's the first protocol where the ranking and clustering presentation of answer is one of the analyzed criteria ; in spite of the fact that many systems, even Web systems (as SPIRIT-W3), used this presentation way. Indeed, usually, Web engines rank totally document in order of relevance but Zamir (Zamir 99) presents some Web experiments in which the clustering process is used. For example Northernlight offers to users the possibility of viewing answers into *Custom Search Folders* labeled by a short phrase. Documents are grouping under criteria such as the Subject (e.g., hypertension, baseball, camping, expert systems, desserts) , the Type (e.g., press releases, product reviews, resumes, recipes) , the Source (e.g., commercial Web sites, personal pages, magazines, encyclopedias, databases) , and the Language (e.g., English, German, French, Spanish). We can noticed that these criteria looks like the one used in Profil-Doc.

Reference

(Borlung 98) : BORLUND P., INGWERSEN P. – Measures of relative relevance and Ranked Half-Life : Performance indicators for interactive IR. – In *Proceeding of the SIGIR 98, 24-28 august 1998*, Melbourne, Australia

(Boyce 94) : BOYCE B.R., MEADOW C.T., KRAFT D.H. - *Measurement in information science*. – Academic Press – 1994 – 283 p.

(Charron 99) CHARRON J, FLUHR C – BeFor (Beyond Forms). Un modèle de représentation du Web invisible – *Hypertext, Hypermedia et internet (H2PTM'99)* – Hermès - Paris

(Ellis 96) : ELLIS D. – The dilemma of measurement in information retrieval research. - *Journal of the American Society for Information Science* – 47(1) - 1996 - pp 23-36.

(Fluhr, 84) FLUHR C. – Le traitement et l'interrogation des bases de données textuelles – In *Informatique et droit en Europe* – Université libre de Bruxelles – Edition Bruxlent. 1984 – pp 97-114.

(Fluhr 97) : FLUHR C. – SPIRIT-W3 : A distributed Cross-Lingual Indexing and Search Engine - *Proceeding of the INET 97 « The Seventh Annual Conference of the Internet Society »* - june 24-27 111997 – Kuala Lumpur, Malaysia

(Frické 98) : FRICKE M – Jean Tague-Sutcliffe on measuring information – In *Journal of the American Society for Information Science* – 34(4) – 1998 – pp 385-394.

(Harter 96) : HARTER S.P. – Variation in Relevance assessment and measurement of retrieval effectiveness – In *Journal of the American Society for Information Science* – 47(1) - 1996 -pp 37-49.

(Harter 97) : HARTER S.P., HERT C.A. - Evaluation of information retrieval systems : Approches, Issues, and methods. In *Annual review of information science and technology* - 32 - 1997 -pp 1-93.

(Hirschman 95) : HIRSCHMAN L, THOMSON H.S., - Overview of evaluation in speech and Natural Langage Processing. In *Survey of the state of the art in human langage technology* collective direction COLE R./MARIANI J./USZKOREIT H./ZAENEN A./ZUE V. - Rapport NFS/CEE. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html. To be published Cambridge University Press et Giardini Publ.

(Kantor 94) : KANTOR P. B. – Information retrieval techniques. In *Annual review of information science and tecnologie (ARIST*) vol 29, 1994, Martha E Williams Editor – Published for the American

Society for Information Sciences (ASIS). pp53-91

(Laine 96) LAINE-CRUZEL S., LAFOUGE T., LARDY J.P., BEN ABDALLAH N. - Improving information retrieval by combining user profile and document segmentation.- In *Information Processing and management* -1996- vol 32 n 3 - pp 305-315.

(Losee 90) : LOSEE R. M. – *The science of information. Measurement and applications* – Academic Press, Inc. – 1990 – 293 p.

(Michel 99): MICHEL C - Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes. Réalisation et evaluation d'un prototype de système de recherche d'information avec filtre selon les profils des utilisateurs.- PhD Thesis - University Lyon II - 6 January 1999 -322 p.

(Mizzaro 97): MIZZARO S - Relevance : the whole history In *Journal of the American Society for Information Science* - 48(9) - 1997 - pp 810-832.

(Park 94) : PARK T.K. - Toward a theory of user-based relevance : A call for a new paradigm of inquiry. In *Journal of the American Society for Information Science* - 45 - 1994 - pp 135-141.

(Radasoa 88) : RADASOA H. Méthodes d'amélioration de la pertinence des réponses dans un système de bases de données textuelles. PhD Thesis. University Paris Sud-Orsay - 28 November 1988 - pp. 156.

(Robertson 92) : ROBERTSON E.S. – On the evaluation of IR systems – In *Information Processing and Management* – 28(4) - 1992 – 457-466 pp.

(Robertson 94) : ROBERTSON E.S. – Computer retrieval – In *Fifty years of information progress. A journal of documentation review* – Edited by Brian C Vickery – ASLIB – 1994 – 235 p.

(Salton 83) : SALTON G. McGILL M. J. – *Introduction to modern Information Retrieval*. New York : McGraw-Hill.

(Saracevick 88I) : SARACEVICK T. - KANTOR P. - CHAMIS A. Y. – TRIVISON D - A study of information seeking and retrieving I. Background and methodology. In *Journal of the American Society for Information Science* - 39 - 1988 - pp161-176.

(Saracevick 88II) : SARACEVICK T. - KANTOR P. - A study of information seeking and retrieving II. Users, questions, and effectiveness. In *Journal of the American Society for Information Science* - 39 - 1988 - pp177-196.

(Saracevick 88III) : SARACEVICK T. - KANTOR P. - A study of information seeking and retrieving III. Searchers, searches, and overlap. In *Journal of the American Society for Information Science* - 39 - 1988 - pp197-216.

(Spark Jones 76) : SPARCK JONES K., VAN RIJSBERGEN C.J. - Information retrieval test collection. In *Journal of documentation* - 32(1) - 1976 - pp 59-75.

(Su 94) : SU L. T. – The relevance of Recall and Precision in user evaluation . In *Journal of the American Society for Information Science* – 45(3) - 1994 – pp207-217.

(Tague 90) : TAGUE J. – Rank and sizes : some complementarities and contrasts – In *Journal of information science* - 1990, 16(1) - p 29-35.

(Tague 92) : TAGUE-SUTCLIFFE J. - The pragmatics of information retrieval experimentation, revisited. In *Information processing and management* - 1992 - 28 (4) - pp 467-490.

(Tague 95) : TAGUE-SUTCLIFFE J. - *Mesuring information. An information services perspectives*. - Academic Press. - 1995 - 206 p.

(Tague 96) : TAGUE-SUTCLIFFE J. – Some perspectives on the evaluation of information retrieval systems In *Journal of the American Society for information science*. - 47(1) - 1996 - pp 1-3.

(Van Rijsbergen 86) : VAN RIJSBERGEN C.J. - A new theoretical framework for information retrieval. In *ACM SIGIR International conference on reserche and development in information retrieval* - Pise, Italie - 1986 - p 200.

(Voorhees 98) : VOORHEES E.M. – HARMAN D. – Overview of the seventh Text Retrieval Conference TREC 7. In *Proceedings of the seventh Text Retrieval Conference TREC 7.* – Gaitherburg 9-11 nivember 1998 - p.

(Zamir 99) : ZAMIR O., ETZIONI O. : Grouper : A dynamic clustering interface to web search results – In *Proceeding of the Eighth International World Wide Web Conference* – May 11-14 1999 – Toronto, Canada (http://www8.org)