Automatic Generation of a Formal User Profile Depending on Several Factors

Labriji Amine
University Hassan II Casablanca, Morocco

Charkaoui Salma
University Hassan II Casablanca, Morocco

Abdelbaki Issam
University Hassan II Casablanca, Morocco

Namir Abdelwahed
University Hassan II Casablanca, Morocco

ABSTRACT
The search for information is not a recent activity. It's a newly rediscovered activity as his control seems to be more and more required. Finding information quickly and efficiently is, in fact, extremely important.

The information research systems tend to customize access to information. They have for objective to issue to the user information that is relevant and appropriate to its preferences, its centers of interest or more its overall profile. The information research systems tend mainly to model the user according to a profile and then to integrate it into the chain of access to information, in order to better respond to its specific need.

This paper presents a technique of implicit construction of the user profile that is inscribed in a formal approach using the behavior of the user as a source for predicting implicitly his need. This technique focuses particularly on the interaction of the user with the information search system as well as its geographical affiliation, its linguistic affiliation and the date of the interaction.

Keywords
User Profile, formal context, Customizing, information retrieval systems (IRS).

1. INTRODUCTION
The general models of information retrieval are based on the assumption that the user is represented by its query, and for a given query the information research systems return the same list of results, yet users have different information needs. The work is currently moving toward a larger definition of the user. It is a stream of research which aims at the implementation of user centered systems by representing him as a user profile.

The analysis of the user behavior reveals a particular importance. In fact, by knowing perfectly how the user will develop its information search strategies, it will be possible to propose significant information for his research. The profiles modeling and the way to adapt them to different users with no clear idea about the seeking information, allows us to provide a personalized access to the content of scientific papers based on the exploitation of the user profile.

We propose a formal profile construction method. We are going to build user profiles for a custom access in a meta-search engine. Aside from the user interactions, we formalize the user profile by 3 parts, namely the geographic, linguistic and the interaction date.

The first section defines the user profile concept, the second shows an art state overview, the third section presents our approach main axes, and finally in the last section, we give a conclusion and an overview of our prospects.

2. USER PROFILE
According to [Wah, 11], a user profile (or even user model) is a set of user related data in an IT service. It is a knowledge source that contains acquisitions on all user aspects which may be useful for the system behavior.

The user profiles are often used by the operating systems, database management systems, search engines and meta-search engines. For example, they allow enriching the users query. Other works are interested in the users feedback during the query launch including the information research systems distribute Peer-to-Peer where a node can be both a client and a server. Indeed, on one hand, the user retrieves a results list; on the other hand the IRS supplies their knowledge base by the information provided by the user, including the query logs and clicks traces, in order to improve the results relevance.

2.1 Profile content
The user profile can combine various information according to the needs. Among the latter are to be distinguished:

- Personal characteristics that may be useful (age, sex, etc.)
- History of interactions with the system, which can allow to model the behavioral habits
- A measure of the psychological state (stress, fatigue, etc.) which remains difficult to determine

2.2 Data recovery
The user profile data are represented according to the needs. In general, they are stored in the knowledge base of the system in the form of property-value pairs.

According to the adaptability of the system, the user profile data can be entered by the user himself (age, sex, etc.) or by selecting a predefined profile by experts or alimented during use.

3. STATE OF THE ART
The user center of interest is represented by its query submitted to the IRS. There are several representation techniques of centers of interests constituting the user profile. A naive representation of centers of interest is based on key words, such as the case of web portals like MyYahoo, InfoQuest, etc. There are other representations more developed to illustrate the centers of interest of the user. [Gow, 6] and [Sie, 7] represents the centers of interests according to vectors of weighted terms. On the other hand [Sie, 8] and [Cha, 9] represents them semantically according to concepts weighted by a general ontology, or according to the matrices of concepts by [Liu, 10].

[Gow, 6] and [Sie, 7] have proposed a user profile modeling according to a class of vectors each of which represents a user center of interest, thus, the classes centroids represents the centers of interest of the user. The semantic representation approaches exploits ontology of reference to represent the user centers of interest according to vectors of weighted concepts of the used ontology. We include the concepts hierarchy of "Yahoo" or ODP 3 as the more used sources of evidence in this type of approaches. [Cha, 9] constructs the user profile on a supervised classification technique of documents considered to be relevant according to a similarity measure vector with the concepts of the ODP ontology. This classification allows, on several research sessions, to associate each ontology concept with a weight calculated by aggregation of similarity scores of classified documents under this concept. The user profile will be constituted by the whole concepts having the highest weight representing the user centers of interest. On the other hand, [Sie, 8] uses simultaneously the user centers of interest represented by vectors of weighted terms and the concepts hierarchy of "Yahoo". The user profile will be constituted by contexts each of them formed by a pair of the hierarchy concepts: one represents the adequate concept to the research, and the other one represents the concept to exclude in the search.

A matrix representation of the user profile is adopted in [Liu, 10]. The matrix is constructed incrementally from the user search history in order to put in place the categories representing the user centers of interest and the associated weighted terms reflecting the degree of the user interest for each category.

4. USER PROFILE CONSTRUCTION

In this section we present our method used to construct the user profile so that the meta-search engine can use it during the classification phase. In our case there is a need to record two information, namely, the relationship between the request terms and the documents and the relationship between the query terms and the search engines. In fact, when a user visits the document, the meta-engine will save this action, completing the query terms with the visited document and the search engines that have returned this document.

4.1 User profile Content

Experience shows that it is difficult to anticipate all the user characteristics in a research session to help him in all possible contexts [Ber, 12]. Thus, it was decided to add other information likely to be interesting for the personalized access to information including geographical and linguistic belonging. In fact, users with the same geographical location often share the same culture and the same center of interest. On the other hand, despite the distance that can separate the internet surfers, we can estimate that their behavior may be similar if they share the same language. Finally, some documents are no more updated therefore are no longer relevant, leading us to add the interaction date so that the information research systems are able to learn whether the document is always visited or not. Finally, the personal history of a given user can be interesting if the user is authenticated in the information research system. Thus we can summarize the characteristics of our user template in the

following figure.

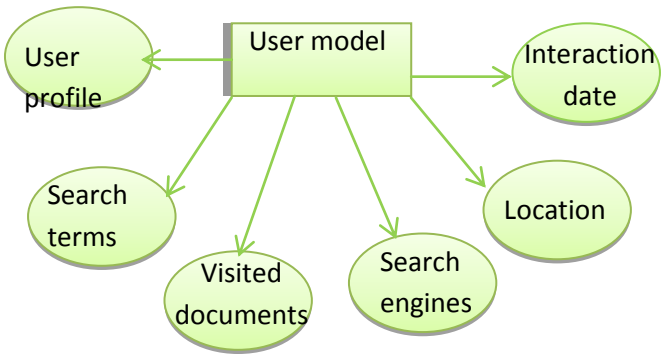

Figure 1: The characteristic of user

4.2 User profile process of construction

We apply a formal method that we can schematize by the following process.

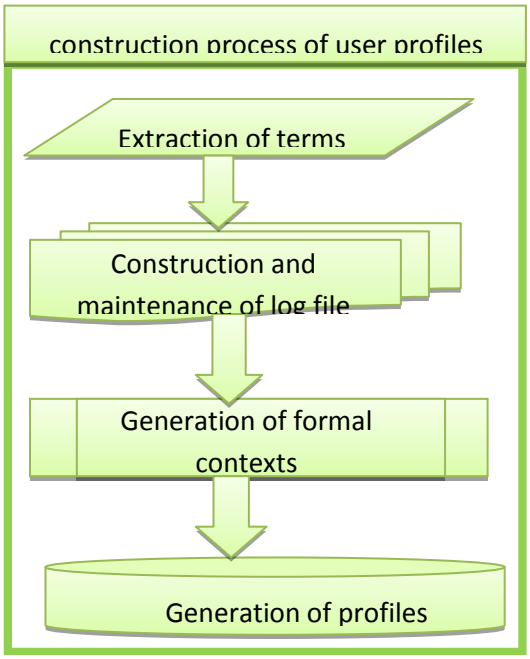

Figure 2: User profile process of construction

There are the 4 main stages, the first one is the query terms extraction phase, then we extract the information from the user navigation history in a XML log file. The third phase is the formal context construction from the log file generated in the previous step, and the last is the profile generation from the formal contexts previously generated.

4.2.1 Terms extraction

To extract the query terms, we chose to apply a form study using the tool TALN (Automatic natural language Processing) Treetagger as a morpho-syntactic analyzer. It is distributed freely for research purposes. It is a tool that allows you to annotate a text with the information deemed relevant. It has been developed by Helmut Schmid in the framework of the project "TC" in the ICLUS utility (Institute for Computational Linguistics of the University of Stuttgart). TreeTagger allows the labeling of German, English, French, Italian, Deutch, Spanish, Bulgarian, Russian, Greek, Portuguese, Chinese and the old French texts. It is adaptable to other languages if the lexicons and the corpus labeled manually are available. Finally, it is customizable according to our needs by expanding the desired specifications.

Following our needs, we proceeded as follows:

- Segmentation: Find basic units which would correspond to the words.

Example: Today (we must locate the separator, in this case " ' " is not a separator).

- Recomposition: Find the compound words .
- Lexical Analysis: Bring the words to a morphological basis form (conjugation, kind, number).
- Stemming: Consists in grouping words with same origin.

Thus, for each request R, we have a list of matching terms Ti.

4.2.2 Log file construction and update

Based on user interaction, we get information about the request: the application identifier (system generated), the user identifier (if authenticated), the location, the request execution date and the consulted terms and documents. Indeed, when the user enters a query, it consults some documents from wish we can deduce the search engines giving them as results. These search engines and documents are called assets in relation to this request.

Example

A query R contains terms (T1 -T2 -T3) with several results, an anonymous user has chosen a set of active documents (D1-D2) so we have a set of search engines associated (M1-M3-M4).

```
<Requests>
<Request login = "R1" >
        <Terms>
                <term>T1</term>
                <term>T2</term>
                <term>T4</term>
                <term>T6</term>
        </Terms>
        <Execution_date>18/06/2013</Execution_date>
        <Location>10.67.32.10</Location>
        <Engines>
                <engine>E1</engine>
                <engine>E2</engine>
        </Engines>
        <Documents>
                <document>D1</document>
                <document>D2</document>
                <document>D3</document>
        </Documents>|
</Requests>
<Requests>
<Request login = "R2" >
        <Terms>
                <term>T1</term>
                <term>T2</term>
                <term>T4</term>
        </Terms>
        <Execution_date>18/06/2013</Execution_date>
        <Location>10.67.32.10</Location>
        <Engines>
                <engine>E1</engine>
                <engine>E10</engine>
                <engine>E5</engine>
        </Engines>
        <Documents>
                <document>D1</document>
                <document>D2</document>
                <document>D7</document>
        </Documents>
</Requests>
<Request login = "R3" >
        <Terms>
                <term>C1</term>
                <term>C2</term>
                <term>C3</term>
        </Terms>
        <Execution_date>18/06/2013</Execution_date>
        <Location>10.67.32.10</Location>
        <Engines>
                <engine>E1</engine>
                <engine>E3</engine>
                <engine>E4</engine>
        </Engines>
        <Documents>
                <document>D1</document>
                <document>D2</document>
                <document>D3</document>
        </Documents>
</Requests>
```

Figure 3: Example of a generated log file

Each request has an identifier, a location, an execution date, and owns as sub-set a whole list of terms and a sub-set of active search engines as well as a subset of active documents in relation to the query.

4.2.3 Formal contexts generation

This is an intermediate step which consists of manipulating the user history in order to generate later knowledge. The latter will be stored in our system in order to provide him with the necessary elements to define the user profile. The formal concept analysis (FCA) is committed to explore the concepts when they are formally described in order to define them precisely.

The FCA allows you to classify subset of terms, their active documents and search engines within the formal concepts. Be O a set of object, P a set of property and R a binary relationship between P and O. a formal context is defined by the triplet (O, P, R). The elements of O are called the objects and the elements of P are called the properties of the context. To express that an object o of O is in relationship with a property p of P, we write oRp. This means that the object o has the property p.

In our case, the terms are the objects, the properties are either the active documents or the active search engines, thus we define two context types:

- Document Context Term (DCT): defines a relationship between a set of terms (objects) and a set of documents (property)
- Document Context Engine (DCE): defines a relationship between a set of terms (objects) and a set of engines (property)

In our case, we say that an object Oi have the property Pj when the latter is always present with the presence of the object Oi. It can be represented in a matrix where 1 means that the object Oi owns the property PJ and 0 otherwise.

Example

Table 1: Example of a matrix showing the relationship between Object and Property

	O1	O2	O3	O4	O5
P1	1	1	1	0	0
P2	1	1	0	1	1
P3	1	1	0	1	1
P4	1	1	0	1	0

4.2.4 Profiles generation

We have two types of profile from the DCT and DCE contexts: The first one represents the link between the passed queries and the active search engines called Profile Engine Term (PMT). The second one represents the link between the passed queries and the active documents called Profile Document Term (PDT), they are defined in the following form: ({m1, . . . , mi}; {t1, . . . , tj}), respectively, ({d1, . . . , dt}; {t1, . . . , tk}), such that {m1, . . . , mi} is a search engine set, which jointly owns all the terms {t1, . . . , tj} and {d1, . . . , dt} is a document set which have all the terms in common.

The profiles set represents a cover, in our case, there are two cover types, one for the PMTs noted C1 and the other for the PDTs noted C2, these two covers represents a knowledge basis generated during the learning phase noted B(C1,C2).

In the table 1, the objects {O1, O2, O4} have the properties {P2, P3, P4}, so we can define a profile P = ({O1, O2, O4}, {P2, P3, P4}).

5. EVALUATION

In order to validate our proposal, we have conducted experiments to assess the impact of the formal method use taking into account several factors during the user profile learning phase, the latter will be used during the classification phase of the results in our meta-search engine.

We used two measures as basic indicators to test the methods effectiveness, it is the "recall rate", that is to say the ratio between the number of relevant documents found during a search and the total number of relevant documents existing in the system. The other indicator is the "accuracy rate", that is the ratio between the number of relevant documents found during a search and the total number of documents found in response to the question.

5.1 TREC Collection

Given that there is currently no standard evaluation framework for a personalized access model to information, we propose an evaluation framework by collections TREC (Text Retrieval Conference), it is an American conference wish purpose is to enable performance comparison between information research systems on large volumes of data. It brings together the tool boxes designers and full-text information research software. It becomes a reference and an international standard in the information evaluation field.

We chose to evaluate our model using the NIST collection (Disk 4- 5) of the TREC evaluation having a size of 741670 documents.

5.2 Learning Phase

First, there is a need to enrich our new knowledge base using our user model construction method. To this end, we have launched the first 10,000 queries.

5.3 Experimental results

We measured our approach in order to build a new knowledge base, the latter will be compared with the old knowledge base (the user interactions history) using the same algorithm for classification. Figure 2 presents the results obtained for the two measures Precision and recall for the two knowledge bases.

Table 1. Table captions should be placed above the table

Number of request	Precision		Recall	
	Trace clic	Formel profil	Trace clic	Formel profil
100	0.840	0.879	2.156	2.182
200	0.851	0.876	2.153	2.179
300	0.849	0.879	2.159	2.179
400	0.859	0.876	2.159	2.179
500	0.858	0.881	2.158	2.181

The first tests presented in this figure are very encouraging. The comparison of our approach with the existing ones shows that our approach is competitive.

6. CONCLUSION AND PROSPECTS

We have presented throughout this paper an implicit construction method of a user profile. It is a formal method that constructs the user template with certain characteristics, including the user interactions, geographic affiliation, its linguistic affiliation and the interaction date. These user profiles are saved in our knowledge database. We intend to use our user profile construction method to classify the results in our meta-search engine. We are also considering using it to put in place a diagnostic system in order to evaluate our meta-search engine.

7. REFERENCES

[1] FU H., NGUIFO E. M., « Etude et conception d'algorithmes de génération de concepts formels », Revue Ingénierie des Systèmes d'Information, vol. 9, no 3-4, p. 109–132, 2004.

[2] FLOC'H A. L., FISETTE C., MISSAOUI R., VALTCHEV P., GODIN R., « JEN : un algorithme efficace de construction de générateurs pour l'identification des règles d'association », Numéro spécial de la revue des Nouvelles Technologies de l'Information, Vol. 1 No. 1, Editions Cépaduès, p. 135–146, 2003.

[3] Sanderson, M. 1994, "Word sense disambiguation and information retrieval », dans SIGIR, proceedings of the 17th Annual ACM SIGIR Conference on Research & Development in Information Retrieval,p.142_151, 1994.

[4] Jacques Savoy, Yves Rasolofo, Faïza Abbaci, "Fusion de collections dans les métamoteurs" dans.JADT 2002 : 6es Journées internationales d'Analyse statistique des Données Textuelles 2002.

[5] Glover, G.W. Flake, S. Lawrence, W.P. Birmingham, A. Kruger, C.L. Giles, et D.M. Pennock. « Improving category specific web search by learning query modifications ». Dans Proceedings of Symposium on Applications and the Internet, pages 23–31, January 2001.

[6] Gowan J., « A multiple model approach to personalised information access », Master thesis in computer science, Faculty of science, Université de College Dublin, February, 2003.

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

[8] Sieg A., Mobasher B., Burke R., Prabu G., Lytinen S., « Representing user information context with ontologies », uahci05, 2005.

[9] Challam V., Gauch S., Chandramouli A., « Contextual Search Using Ontology-Based User Profiles », Proceedings of RIAO 007, Pittsburgh USA, 30 may - 1 june, 2007.

[10] Liu F., Yu C., Meng W., « Personalized Web Search For Improving Retrieval Effectiveness », IEEE Transactions on Knowledge and Data Engineering, vol. 16, n° 1, p. 28-40, 2004.

[11] Wahlster W. et Kobsa A.,Dialogue-based user models. In Proceedings of IEEE, Vol. 74(7), pp. 948-960, 1986.

[12] Berisha-Bohé S., « Modélisation de l'utilisateur pour la recherche d'information dans des bibliothèques numériques », Master Recherche, INSA Lyon, 2005.