Models of Digital Educational Resources Indexing and Dynamic User Profile Evolution

http://dx.doi.org/10.3991/ijet.v11i01.4932

H. Slimani, N. El Faddouli, S. Bennani, N. Amrous
Mohammed Vth University Rabat, Rabat, Morocco

Abstract—**The modelling of the user profile and its integration into the search process is an effective way in personalized information search within a repository of educational digital resources. Therefore, it raises gradually the issue concerning the dynamic development of this profile so as the information requester sets up queries. In our approach presented in this paper, we propose two models for personalized search on digital educational resources. The first is to establish an index of repository resources while the second is to build the user profile and boost its evolution after each query submitted by the user based on a classical Bayesian network representing a search activity.**

***Index Terms*—User Profile, personalized search, classical Bayesian networks, LOM, Dewey Decimal Classification, Digital Educational Resources, indexing, Lucene engine, relevance document.**

I. INTRODUCTION

The rapid development of Internet and the increasing use of Information and Communication Technologies for Education (ICTE) led, among others, to the proliferation of Digital Educational Resources (DER) within the information retrieval systems in general and the educational content storage systems in particular. These resources are increasingly available in various forms: electronic courses, text files, recordings, animations, audiovisual contents, multimedia documents, etc.

Many works have treated approaches of the integration and evolution of user profiles in the search for information. This is expressed by adding new information to the query [1] [2], by modeling based DBN (Dynamic Bayesian Network) [3], by using a trace-based system [4], by exploiting search history [5] or by associating with resource description user profile elements [6]. The originality of our approach lies in the fact of finding a link or bridge between the user profile and digital resource descriptors by modeling the user through the indices of Dewey Decimal Classification (DDC). These indices are part of standard metadata, in this case the LOM (the ninth category "Classification" element 9.2.2.1).

In a previous work, we presented a personalized search approach according to the profile of information user [7]. This profile is evolving from a static way. The present work is a continuation in this axis of research. It aims at making dynamic the development of integrated user profile in the personalized search.

Indeed, in order to improve searching and therefore satisfy the information needs of the user, we propose a new indexing model inspired from a search engine, which is part of free and open-source software. This is the Lucene search engine (http ://lucene.apache.org/) [8]. The choice of this engine is mainly thanks to its easy use and its efficiency. For the modeling of the user, we use an approach based on classical Bayesian networks to integrate the resulting profile in the process of finding information. This approach also deals with the dynamic evolution of this profile.

We present initially the issue related to the integration and dynamic evolution of the user profile in the information search process. Then we present our own indexing model. The proposed integration and dynamic evolution model of the user profile is described in section4. We will subsequently summarize our research so far. Finally, we will propose research directions likely to interest future researchers in this area.

II. PROBLEM FORMULATION

The large growth in the production of digital educational resources in the recent past, combined with the varied typology, enormously heterogeneous nature, multiform and multilingual of information present obstacles that current information retrieval systems must resolve to fully enhance the level of relevance of returned resources. This relevance is reflected in the adaptation degree of the results returned to recurrent preferences, knowledge, user interests or just to user profiles. Obviously, the best way to achieve this goal is to develop a clear representation of the user and insert the resulting model in at least one of the phases of the search process. This is called personalized search, intelligent or adaptive according to the user profile [9] [10]. This raises the issue related to the dynamic evolution of this profile as the information user poses its queries. So the problem is: how to make dynamic the evolution of user profile integrated into an educational content storage system as the user makes each search query?

To address this issue, we propose a solution, which is based on a bridge between metadata descriptors of digital educational resources (element 9.2.2.1 Table II) and the user modeling (Table III). In this sense, our work has two aims, first propose an indexing model of digital educational resources inspired of the Lucene search engine and, second, describe our approach to the building and evolution of the user profile by exploiting each request operated by the user who requests the information.

III. INDEXING MODEL OF DIGITAL EDUCATIONAL RESOURCES

At large, indexing is to describe and represent the intellectual content of a document (whatever its medium) to facilitate the search. In other words, it is the content description, consisting in translating keywords, from the analysis, in a natural or documentary language (Dewey Decimal Classification, Universal Decimal Classification, Thesaurus, etc.)

Once a document is indexed, it will be easily manipulated by a machine in order to be exploited by the searcher of information (Figure 1) [11].

The most complete metadata schema and the most frequently used in the indexing of digital educational resources is the Learning Object Metadata (LOM), which we will define below.

A. Learning Object Metadata (LOM)

LOM is the acronym for "Learning Object Metadata" [12]. Technically, its name is IEEE 1484.12.1-2002 (LOM)[13]. LOM is a Descriptive Metadata Scheme (DMS) associated with digital educational resources. This standard was published in June 12th, 2002 by the International IEEE-LTSC-LOM committee (Institute of Electrical and Electronics Engineers, Inc. - Learning Technology Standards Committee - Learning Objects Metadata Working Group) based on the technical specifications of ARIADNE metadata (Alliance of Remote Instructional Authoring and Distribution Networks for Europe), IMS (Instructional Management System) and Dublin Core. It aims at providing a common working framework at the international level to foster the sharing and exchanging of digital educational resources.

The version 1.0 descriptors of the LOM are divided into 9 categories, grouping 68 metadata, 10 among them are composed. Table I provides a brief description of each category.

In the proposed model, we are interested particularly in the ninth category "Classification"; it describes the location where there is such a digital educational resource where the classification system used, and more precisely, in the element 9.2.2.1 (Table II). This is the identifier of the taxon that contains an alphabetical expression, digital or mixed, provided by the classification system selected in the element 9.2.1.

The element 9.2.2 represents the taxon, which is a particular term in a given taxonomy. This composed element can be repeated because it is possible that the same resource is referenced by more than one term in the same system. Thus, each digital educational resource may have one or more identifiers of the taxon.

For our study, the value of the taxon identifier consists only of numbers, given that the classification system used is the Dewey decimal classification (DDC) described below. It should be noted that our approach is more general; it can be used with any other classification. The choice of the DDC is because it is considered one of the most used classifications today.

B. The Dewey Decimal Classification (DDC)

In general, classification provides an orderly and hierarchical system of categorization of objects. It is used to arrange the knowledge represented in any form, such as, digital resources, living species, diseases, documents in a library, etc.

The Dewey Decimal Classification is a classification system developed by Melvil Dewey in *1876*. It organizes all human knowledge in a hierarchical structure, ranging from general concepts to specific objects. DDC is divided into *10* main classes, *100* divisions, *1000* sections and a multitude of subsections. The folowing figure (Figure 2) shows the example of *512.5* Index.

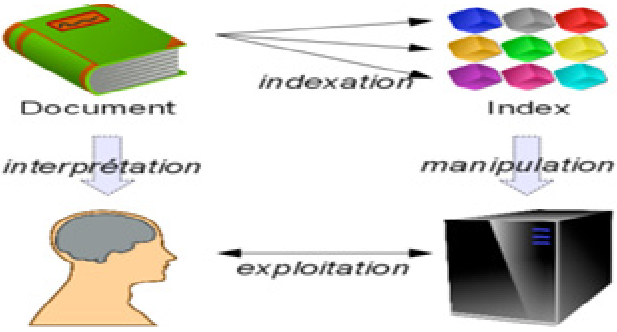

Figure 1. Principle of indexing

TABLE I.
DESCRIPTION OF THE LOM CATEGORIES

Categories	Description
1. General	Characteristics of the DER that are independent of the context of use
2. Life cycle	Describes the current state of the DER and the history of its various versions
3. Meta-Metadata	Metadata information rather than DER
4. Technical	Characteristics and technical requirements of DER
5. Educational	Educational characteristics of the DER
6. Rights	Intellectual property and conditions of use of the DER
7. Relation	Definition of the links between the DER and others
8. Annotation	Comments on the educational use of DER
9. Classification	Allows the classification of subjects treated with the resource

TABLE II.
STRUCTURE OF THE CATEGORY CLASSIFICATION

<table>
<tr><td rowspan="6">9. Classification</td><td colspan="3">9.1 Purpose [R]</td></tr>
<tr><td rowspan="3">9.2 Taxon Path [O]</td><td colspan="2">9.2.1 Source [O]</td></tr>
<tr><td rowspan="2">9.2.2 Taxon [O]</td><td>9.2.2.1 Id [O]</td></tr>
<tr><td>9.2.2.2 Enty [O]</td></tr>
<tr><td colspan="3">9.3 Description [R]</td></tr>
<tr><td colspan="3">9.4 Keyword [R]</td></tr>
</table>

[O]: Field must be filled (compulsory).
[R]: Field can be filled (recommended).

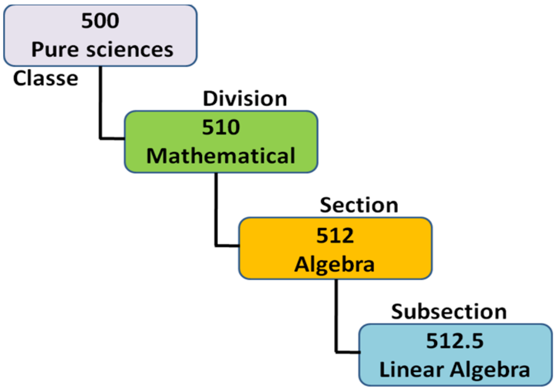

Figure 2. Tree of the Dewey Decimal Classification

The DDC index or notation is a numeric value. It consists of at least three digits. The fourth digit is separated from the third by a point, from sixth digit each three digits are separated by a space. More along notation is, the more the subject is precise, for example the following index follows the rules of the notation: 830.900 8.

The indices of the Dewey Decimal Classification are used among the research criteria on the Web by some search engines.

C. *The Lucene search engine*

Lucene is a text-oriented search engine. It is developed in Java by Doug Cutting in *1997*. In October *2001*, this engine was taken over by the Apache Foundation. It makes it possible both to index and retrieve resources. The indexing and search operations, as performed by Lucene, will be described in the following:

1) Indexing: The principle of the Lucene indexing is based on a mechanism that leads to automatically create a structure that provides a powerful and quick search of records in a file. This structure is commonly called index. This principle is similar to the operation of searching for a book in a library using a catalog.

A resource is represented in the form of a record through a set of fields or metadata deemed relevant such as (Title, Author, Keywords, Abstract, taxon identifier, etc.). All the terms that constitute these fields are then filtered by analyzers; they allow bringing back the terms to their radical, for example:

1. – Cut out the text term by term,
2. – Render all the terms in lowercase,
3. – Render all the terms in the singular,
4. – Remove all suffixes and prefixes,
5. – Dispose of stop words such as: "of", "the", "my", "a", "you", "and" etc.

Once the filtering-terms operation is completed, the next step is to establish the index (Figure 3), associating thus each term resulting from the previous phase with the number(s) of its record(s).

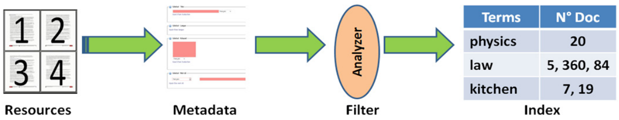

Figure 3. Steps in the indexing operation with Lucene

In our indexing approach, we will follow the same model; the proposed change will be in the constitution of the index which will be different.

2) Search: Search is the operation that consists to obtain references to resources, by comparing each term, object of search with those on the index, which is pre-established in the indexing phase. Two major factors intervene in the search quality are: (1) the relevance of returned resources (2) the execution time of the query.

When a user enters a term or a set of terms in the graphic interface offered by the search engine in question, these terms will be subsequently filtered by the analyzers in the same manner as that of the indexing operation. Then the system compares the resulting terms with those contained in the field (Terms) of the index (Figure 4). This way the system manages to extract the identifier of the resource stored in the database.

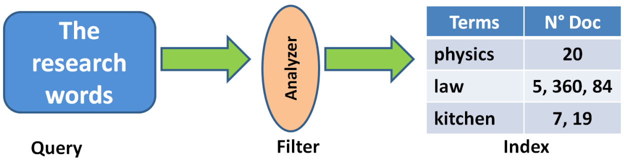

Figure 4. Steps in the search operation with Lucene

In the index, a term may be associated with multiple resources at once. Lucene attributes for each significant term, object of search, a degree of conformity with that of the index, for each resource; this value is expressed as a percentage.

D. *Proposed indexing model*

The indexing model that we propose is similar to that of the Lucene search engine. The difference lies in the index creation phase, to which we added a third column "DDC Indices», (Table III) containing the value (s) of taxon identifier element 9.2.2.1. In the case of multiple values, they must be separated by commas. This is for each record representing a resource of the second column.

These DDC indices represent the disciplinary fields or the themes which gather corresponding digital educational resources. A resource may have one or more indices. It depends on the context where such resource is used, taking the following examples: (1) a resource that deals with a theme related to the physical matter; it can have either the index *530* representing the generalities on physical matter, or the index *621* representing the applied physics or both (2) same thing for the woman disciplinary field which may have for example:

1. – *305.4*: interdisciplinary works on women, on females,
2. – *155.633*: psychology of women,

1 – *346.013 4*: legal status of women.

TABLE III.
THE PROPOSED INDEX

Terms	N Doc	Indices DDC
Physics	*20*	*530, 621*
Law	*5*	*340*
	360	*340*
	84	*340*
Kitchen	*7*	*641.5*
	19	*641.5*

Another point of difference with respect to the index of Lucene: when a term is cited in several resources at the same time, the structure of the numbers of the records (second column of the proposed index) became vertical instead of a horizontal representation. This is done to discriminate the corresponding DDC indices to each record.

The third column added will serve us for the dynamic evolution of the user profile using classical Bayesian networks that are the subject of the next section.

IV. EVOLUTION OF THE USER PROFILE

A. *Theoretical framework of classical Bayesian networks*

A classical or static Bayesian network is a combination of the theories of probabilities and of graphs; it allows to reason with missing information and represent probabilistic knowledge, this in the form of a DAG (Directed Acyclic Graph) $G = (X, E)$ whose set of nodes X are associated with random variables such as: $X = \{X_1, X_2, ..., X_{n-1}, X_n\}$ and the set of arcs E that represent probabilistic dependencies among those variables. These arcs determine the conditional probability distributions for each X_i knowing the discrete states of its parents [14]. Figure 5 shows an example.

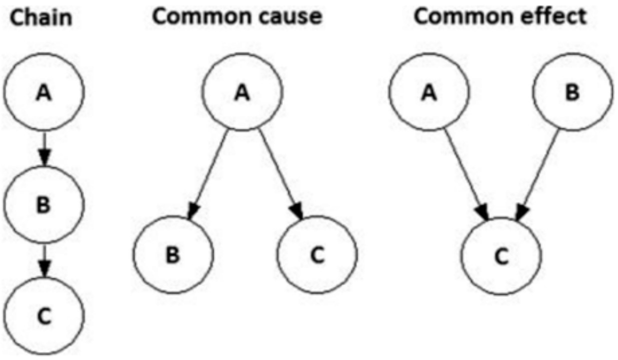

Figure 5. Bayesian networks example

The calculation of different probabilities from a Bayesian network is performed using the determination of conditional probability tables. These tables give the probability for each value of the node knowing the combinations of the values of the parents of the node. If X_i is a root node, its probability distribution is said a priori or unconditional; otherwise, it is said conditional.

For each variable X_i, the joint probability can be calculated by multiplying the conditional probabilities of each variable given its parents (1) [16]:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|parents(X_i)) \qquad (1)$$

B. *Calculating the relevance of a resource*

The relevance of a resource stored in a repository of digital educational resources is measured in relation to the importance of belonging of a term t_i to this resource, t_i contained in the request sent by the user. To evaluate this relevance in our approach we used the TF-IDF weighting method (Term Frequency-Inverse Document Frequency). This method is based on Zipf's law [16].

The calculation of the relevance of a resource or document d_j in relation to a term t_i is obtained by calculating the t_i weight in d_j. This weight $w(t_i, d_j)$ is between 0 and 1. It is defined by the product of the one part the Term Frequency which is the number of occurrences t_i in $d_j(tf_{i,j})$ and secondly the Inverse Document Frequency which is a measure of the importance t_i in the entire repository (idf_i) as follows (2) [17] :

$$w(t_i, d_j) = tf_{i,j} X idf_i \qquad (2)$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (3)$$

- $n_{i,j}$: Number of term t_i occurrences in the document d_j;
- $\sum_k n_{k,j}$: Total number of terms in the document.

$$idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \qquad (4)$$

- $|D|$: Total number of documents in the repository ;
- $|\{d_j : t_i \in d_j\}|$: Number of documents that contain the term t_i.

The calculation of the relevance of digital educational resources is an important factor. It will be employed in our approach to the dynamic evolution of the user profile, and this will be detailed in the next section.

C. *The dynamic evolution of the user profile approach*

1) User modeling: The integration of the user model in at least one phase of the process of finding information is very useful in searching relevant results in large databases of digital educational resources; otherwise, the same resources will be returned to the same queries even if they were submitted by different users who request the information; they have different contexts, different informational needs, different user interests, etc.

Research has been carried out on the representation of the user model; we cite three most used approaches: ensemblist or vector approaches like [18], conceptual or semantic approaches like [19] and multidimensional approaches like [20].

In our approach, the proposed user model is ensemblist, described more accurately, by a list of numeric values or specifically DDC indices, where each index corresponds to a specific interest.

[2] TABLE IV
AN EXAMPLE OF PROFILE REPRESENTED BY DDC INDICES

User profile
530
621
340
641.5

The same requester of information may have, at different times, different interests. This generates the question of the dynamic evolution of this profile as the user sends requests.

2) Modeling the dynamic evolution of the user profile: The proposed user profile based on an ensemblist representation; in this case its construction and dynamic evolution are expressed by the systematic addition of new DDC indices to the list that constitutes this profile. These indices are extracted and ordered from the latest resources collected which are deemed relevant, following the various search operations carried out by the requester of information.

The model is a classical Bayesian network; it represents a search activity, which associates the query node *q* (Figure 6), the term nodes *{t_1, t_2,...., t_i}*, the document nodes *{d_1, d_2,...., d_n}* and the user interest nodes *{c_1, c_2, ..., c_m}*.

All the nodes in our model represent binary random variables where the domain is *Dom(n) = {n,$\overline{n}$}*, where *n* represents the case where the node *n* is instantiated positively whereas *n* designates that this node is instantiated negatively or it is not instantiated. The Table IV summarizes the possible states of the nodes:

TABLE IV.
STATES NODES OF THE PROPOSED MODEL

Query	q	:	Query *q* instantiated positively.
	$\overline{q}$	:	Query *q* is not instantiated.
Terms	t_i	:	Term *ti* belongs to the query *q*.
	$\overline{t_i}$	:	Term *ti* does not belong to the query *q*.
Documents	d_j	:	Document d_j.
	$\overline{d_j}$	:	Document d_j.
User Interests	c_k	:	Index c_k belongs to at least one document d_j.
	$\overline{c_k}$	:	Index c_k does not belong to any document d_j.

In our approach, the outbreak of the search process activity is owed to the arrival of a query emitted by a user; that is why the case of q where the query is not instantiated does not interest our study.

The causal links between these nodes are represented by directed arcs respectively as follows: (1) the query node to the terms nodes (2) the term nodes to the document nodes (3) the document nodes to the user interests nodes.

When a user sends a query in order to search for specific information in digital educational resources, the evaluation of the latter is made according to the following sequence of events (Figure 7):

1. *Phase of search*: The query is cut up into terms {t_1,t_2,....,t_i} after having filtered them by the analyzers. These terms will be compared with those contained in the Terms field of the index (Table III) if none of them is found, a message is displayed for this purpose, otherwise we move to phase 2.
2. *Phase of calculation of the relevance*: On the basis on the mathematical formulas addressed in section 4.2 a corresponding numerical value to the weight of the term-document matching is assigned to each document. If this value exceeds the threshold, the DDC indices of the third column of our index (Table III) corresponding to the mentioned document will be recovered. Otherwise, a message is displayed to this effect.
3. *Phase of the profile update*: The dynamic evolution of the user profile is done by adding at the beginning of the list of new DDC indices (Table 4), recovered during phase 2 after every search activity carried out by an information requester, which model the user. Thus, the profile is updated as of the repository puisements.

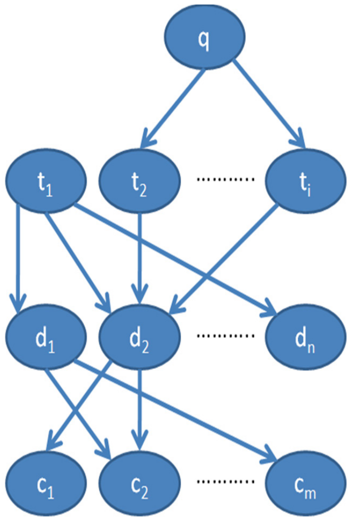

Figure 6. Modeling a search activity

4. *Phase of the display:* This phase corresponds to the step of displaying the follow-up to treatment of the query posed by the user. For this, the system be either :

- Prepares and displays the collected results.
- Or, displays a message informing the user that no relevant resource is found.

V. DISCUSSION

We divided our approach into two main phases: the first is to establish an index after the indexing operation of digital educational resources; the second phase is to build the user profile and make its evolution dynamic after each query submitted by the applicant information.

In an information retrieval system, the purpose of the integration of the user profile in the process of searching is the issuance of relevant results corresponding to the context of the requester of information. We treated this point in our work by doing the calculation of this value for each term relative to the document where it is mentioned. For the evaluation of this measure we have compared with a reference value or threshold to determine the degree of relevance of such a document. It remains to determine in the next step of our study the exact value of this threshold.

VI. CONCLUSION

In this article, we discussed the approach of the determination and dynamic evolution of the user profile in a personalized search. The proposed model is represented by a classical Bayesian network that represents a search activity, which combines the query node, term nodes, document nodes and user interest nodes. The causal links between these nodes are represented by directed arcs.

In the future, we will focus on determining the relevance threshold value of digital educational resources subject of the personalized search and implementing the proposed models, experimenting and evaluating them in relation to personalized information retrieval models already existing.

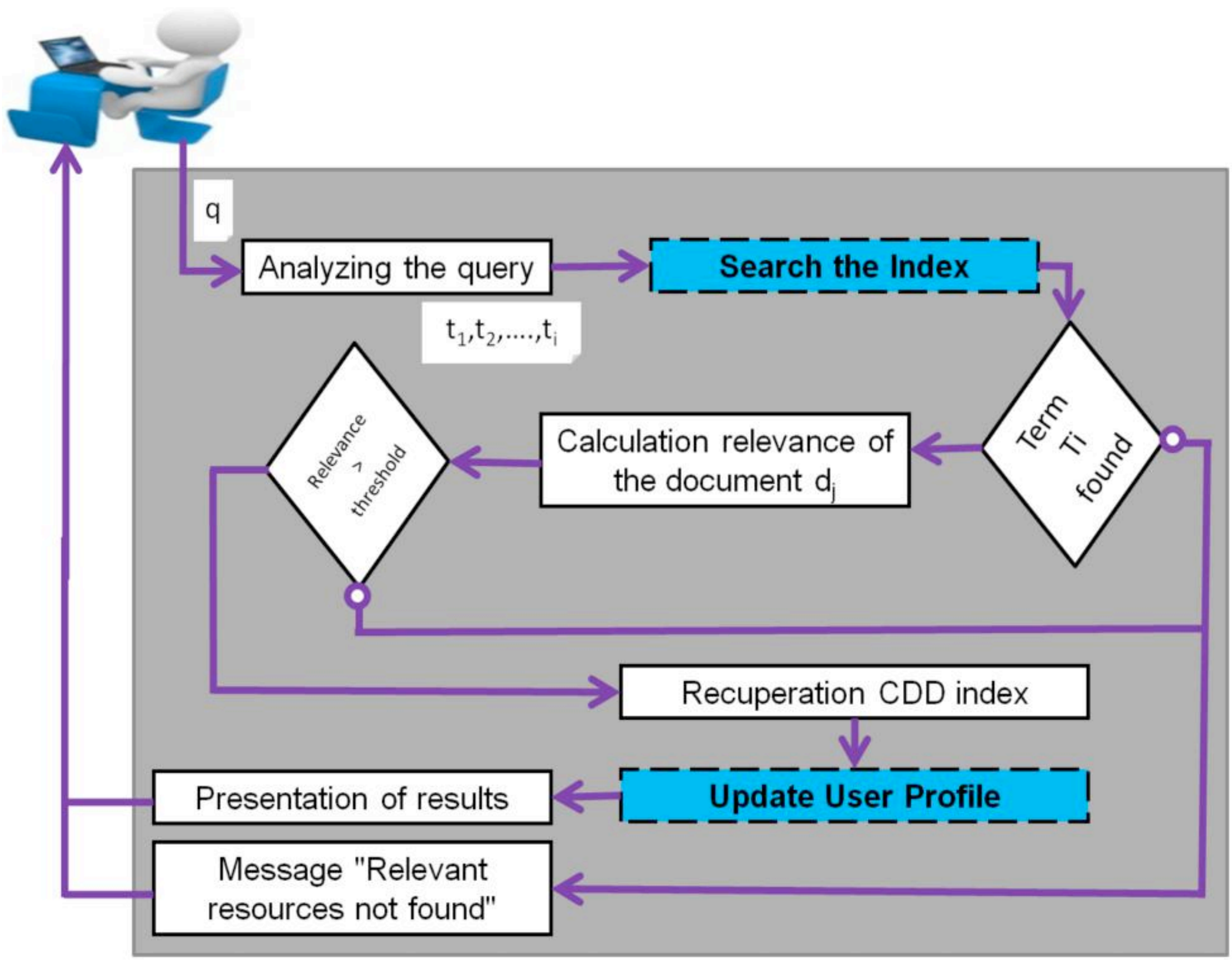

Figure 7. The sequence of events of a query

REFERENCES

[1] S. Berisha-Bohé and B. Rumpler, "Modèle évolutif d'un profil utilisateur" in *CORIA*, 2007, pp. 197–210.

[2] I. Ruthven and M. Lalmas, "A survey on the use of relevance feedback for information access systems," *The Knowledge Engineering Review*, vol. 18, no. 02, pp. 95–145, 2003. http://dx.doi.org/10.1017/S0269888903000638

[3] F. Achemoukh and R. Ahmed-Ouamer, "Modélisation d'évolution de profil utilisateur en recherche d'information personnalisée" in *CORIA*, 2012, pp. 83–97.

[4] V. Butoianu, O. Catteau, P. Vidal, and J. Broisin, "Un Système à Base de Traces pour la Recherche Personnalisée d'Objets Pédagogiques : le cas d'Ariadne Finder," *Atelier Personnalisation de l'apprentissage : quelles approches pour quels besoins ?"* , EIAH 2011, 2011.

[5] L. Tamine and W. Bahsoun, "Définition d'un profil multidimensionnel de l'utilisateur : vers une technique basée sur l'interaction entre dimensions," in *Actes de la Conférence francophone en Recherche d'Information et Applications (CORIA 2006)*, 2006, pp. 225–236.

[6] Y. Amerouali, "Metadata et profil utilisateur," in *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI*, 2013.

[7] H. Slimani, N. El Faddouli, M. K. Idrissi, and S. Bennani, "Sharing the digital pedagogical resources among institutions of higher education in Morocco," in *Information Science and Technology (CIST), 2014 Third IEEE International Colloquium in*, 2014, pp. 266–271.

[8] M. McCandless, E. Hatcher, and O. Gospodnetic, *Lucene in Action: Covers Apache Lucene 3.0.* Manning Publications Co., 2010.

[9] D. Kostadinov, "Personnalisation de l'information : une approche de gestion de profils et de reformulation de requêtes," Université de Versailles-Saint Quentin en Yvelines, 2007.

[10] M. Chevalier, C. Julien, and C. Soulé-Dupuy, *Collaborative and Social Information Retrieval and Access : Techniques for Improved User Modeling : Techniques for Improved User Modeling*. IGI Global, 2009. http://dx.doi.org/10.4018/978-1-60566-306-7

[11] *Intérêt et usage des métadonnées, indexation et recherche (n.d.)* [Online]. Available : http://scenari.utc.fr/stc/pro/pres/20100504vocabnomen/site/co/section1.html

[12] B. De La Passardière, M. Grandbastien, and others, "Présentation de LOM v1. 0, standard IEEE," *Revue Sciences et techniques éducatives*, 211–218, 2003.

[13] LOM specification, Learning Object Metadata, Retrieved Octobre 20, 2014, website of IEEE Standards Association.[Online]. Available : http ://ltsc.ieee.org/wg12/index.html

[14] H. Turtle and W. B. Croft, "Evaluation of an inference network-based retrieval model," *ACM Transactions on Information Systems (TOIS)*, vol. 9, no. 3, pp. 187–222, 1991. http://dx.doi.org/10.1145/125187.125188

[15] A. Hasman, “Probabilistic reasoning in intelligent systems : Networks of plausible inference : by Judea Pearl, Morgan Kaufmann Publishers Inc., San Mateo, California, 552 pp.,” *International Journal of Bio-Medical Computing*, vol. 28, no. 3, pp. 221–225, Jul. 1991. http://dx.doi.org/10.1016/0020-7101(91)90056-K

[16] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information processing & management, vol. 24, no. 5, pp. 513–523, 1988. http://dx.doi.org/10.1016/0306-4573(88)90021-0

[17] Term Frequency-Inverse Document Frequency, “TF-IDF” Wikipédia. 17-Jul-2014. http ://ltsc.ieee.org/wg12/index.htm

[18] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, “Stuff I’ve seen : a system for personal information retrieval and re-use,” in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 72–79. http://dx.doi.org/10.1145/860435.860451

[19] F. Liu, C. Yu, and W. Meng, “Personalized web search for improving retrieval effectiveness,” Knowledge and Data Engineering, IEEE transactions on, vol. 16, no. 1, pp. 28–40, 2004.

[20] M. Bouzeghoub and D. Kostadinov, “Personnalisation de l’information : aperçu de l’état de l’art et définition d’un modèle flexible de profils.” *CORIA*, vol. 5, pp. 201–218, 2005.

AUTHORS

H. Slimani (First author) is with RIME TEAM-Networking, Modeling and e-Learning- LRIE Laboratory Re-search in Computer Science and Education Laboratory Mo-hammadia School of Engineers (EMI) - Mohammed Vth University Agdal AV. Ibn Sina Agdal Rabat BP. 765 Morocco (hamidslimani@research.emi.ac.ma).

N. El Faddouli is a professor at Mohammadia School of Engineers (EMI) - Mohammed Vth University Agdal AV. Ibn Sina Agdal Rabat BP. 765 Morocco (faddouli@emi.ac.ma).

S. Bennani is also a professor at Mohammadia School of Engineers (EMI) - Mohammed Vth University Agdal AV. Ibn Sina Agdal Rabat BP. 765 Morocco (sbennani@emi.ac.ma).

N. Amrous is a professor of information science at the School of Information Science in Morocco (namrous@esi.ac.ma)

Submitted 05 August 2015. Published as resubmitted by the authors 26 December 2015.