
A System based on metadata for creating a Community cache

**Aref BOUKHRIS*– Habib SMEI*–
Abdelmajid BEN HAMADOU*– Mesaac MAKPANGOU****

** Laboratoire de recherche LARIM, Institut supérieur d'informatique et de multimédia Sfax – Tunisie.*

arfb1@yahoo.fr.

habib.smei@isetsf.rnu.tn,

abdelmajid.benhamadou@isimsf.rnu.tn,

*** Laboratoire INRIA Paris*

mesaac.makpangou@inria.fr

ABSTRACT: *The Web is a gigantic database where it is difficult to find, in a precise and fast way, desired information. Indeed the existing tools of research and filtering are not enough powerful to fulfil the users requirements of documents. This is true for a community of users where the demands for documents are very pointed and where the members of a given community wish to have the tools their allowing having research targeted possibilities. In this paper we present a cache system containing pedagogical documents described by metadata. The objective of this system is bound to the members of the community in a collaborative spirit, to exchange resources and ideas, to facilitate the access to these resources by a search engine which uses an approach of filtering by coupling metadata and user profile to improve the relevance of the pedagogical documents. Thus, we present the total architecture of this system and its implementation.*

RÉSUMÉ : *Le Web est une base de données gigantesque où il est difficile de trouver, de façon précise et rapide, l'information souhaitée. En effet les outils de recherche et de filtrage existants ne sont pas assez performants pour répondre aux exigences des utilisateurs demandeurs de documents. Ceci est plus vrai pour une communauté d'utilisateurs où les demandes en documents sont très pointues. Dans cet article nous présentons un Système à base de métadonnées pour la création d'un cache communautaire appliqué au contexte pédagogique. L'objectif de ce système est de créer des liens entre les membres de la communauté dans un esprit de collaboration, d'échanger des ressources et des idées, de faciliter l'accès à ces ressources par un moteur de recherche qui s'articule autour d'une approche de filtrage par appariement entre métadonnées et profil utilisateur, d'améliorer la pertinence lors de la recherche des documents pédagogiques hébergés dans le système de cache que nous proposons. Ainsi nous présentons l'architecture globale de ce système et son implémentation.*

KEYWORDS: *user profile, metadata, Community cache, pedagogical document, Information retrieval System.*

MOTS-CLÉS : *profil utilisateur, métadonnées, cache communautaire, document pédagogique, Système de Recherche d'Information.*

1. Introduction

Currently, we attend an enormous quantity of documents which are created, published and diffused each day thanks to Internet network. Indeed, we count more than 4 billion Web pages [BRPL 04] [MEATM] indexed by Google [GOGL2] excluding the 84 billion pages estimated as "Deep Web" [BRPL 04]. This generates a continuous growing of Internet estimated by the Metamend site [METAM] by more than 10 million pages per day, which makes consequently difficult the localization of the "good" information sought by the users. Indeed, the least of the queries on a research tool returns frequently a significant number of responses, and it is very difficult to find relevant information in this huge mass of documents. Thus, the user is confronted to the noise problems (a lot of returned documents) generated by the research tools and passes a lot of times to find pertinent information.

Note that this work belongs to the project SYFAX [SFX1][SFX2] which is interested in the pedagogical context. SYFAX ensures the excavation of the Web, the extraction of the documents metadata which could be lodged in the Community cache for a possible fast access; it ensures also the assumption of responsibility of the user data like his centers of interest (user profile). This system proposes an approach of filtering by pairing metadata - user profile used in the process of research to eliminate the noninteresting documents. SYFAX offers to the user's mechanisms of co-operation to enable them to share their experiments as well as information or documents.

The second section will describe the existing solutions to improve research, the 3rd will present the metadata in the pedagogical context, section 4 will describe the user profile and will propose an approach of filtering by pairing metadata - user profile used in a system of pedagogical Community cache, and section 5 will describe the total architecture of the proposed cache system and finally section 6 will present its implementation.

2. The existing solutions to improve research

To seek information on the Web, automatic research tools are used (search engine, directories, meta-engines, etc). They are powerful software allowing traversing the whole Web to search new sites and to index them and integrate them in their databases. Consequently, when the Net surfer formulates his query, the research tool looks in its database, to found the documents in association with the query. These research systems, known as "traditional" use the vocabulary of the language like bond of correspondence between the user and the document (matching query/content).

Recently, we could see developed other ideas to improve documents research like: research multi criteria by adding the possibility of refining research by other elements than keywords, the integration of data of semantic type to describe the

documents, the assumption of responsibility of the user's centers of interests and his profile with the process of research for better reply to its needs and the restriction of the documents at a community through Community gateways.

We recommend in our cache system [CACH] of Community associating the addition of the metadata and the user profile by applying it at a community, which makes it possible to improve research considerably by adopting mechanisms which automatically feed the base of documents available from the Web. And system of recommendation also based on the system of pairing metadata - user profile.

3. Metadata and to the pedagogical context

The term metadata is used to define the whole technical and descriptive information added to the documents for their better qualification. To have been used by others users. The real interest of these metadata is the addition of semantic contents of nature to the documents published, which strongly increases the quality of information available for the research tools.

In the field of education and e-learning, working groups tried to define and specify adapted elements of description. We can quote mainly: Dublin Core Education [DCEd], Learning Object Model [LOM], Learning Technology Standards Comitee (LTSC), European Schoolnet [ESclNet] and The working group Metadata Education (GTME 1.1) [GTME 01]...

We adopted the diagram of the GTME 1.1, because of its simplicity, with which we can modify some elements, and we adapted it to our needs. The identifiers and the diagrams of encoding suggested are in French. Within the framework of our cache system, we retained 13 elements, and modified the properties of several elements and added 2 elements to obtain 15 elements of metadata to describe the pedagogical documents. These elements are: Title, Auteur, keywords, Description (summarized), Date (of the last modification), Type of document, Format, Langue, Status, level, Discipline, Under discipline, Nature, Media, Public.

4. Proposal for an approach of filtering by pairing: metadata – user profile

A user profile is a collection of information about the user [LEXMAIA]. This collection can be seen as a grow of characteristics with associated values containing for example what the user prefers, or what it is able to make, his profession... There are several types of profiles: static profiles, reflexive profiles, and dynamically corrected Profiles. In our system we use the reflexive profiles and we tend to adopt the profiles corrected dynamically to adapt to the new requirements of the users. Admittedly, it is completely commonplace to recognize that the best means to collect relevant documents is to take into account the profile of the user, in the process of information retrieval; In this paper we focus on the novice user by a

system we call pairing metadata - user profile. To achieve our goal, we adopt following methodology: analyzes the needs for information of the user in order to determine the characteristics of the user, then deduce the useful metadata to associate with the documents, finally establish a correspondence between the characteristics of the user and documents metadata. Our cache system is based on a pedagogical corpus of documents. We use three questions [LAI 99] like three successive stages to determine the characteristics of the user in the pedagogical context: Who is it? What does he want? What to do (which information)? Which is the final goal of the use of the selected documents? The first question corresponds to the profession, the second to the type of desired documents, and the third corresponds to the disciplines and the under-disciplines. The association between the user profile and the metadata can be represented in the shape of a binary table of association, or a variable, flexible. Nevertheless, the precision of the value by a progressive seizure at the time of the formulation of the queries makes it possible to associate a growing number of metadata. We can also associate ontology for each metadata for a more complete association the elements (characteristic) of the user profile.

5. The cache system structure

To feed the database of documents of the Community cache from the Web (see figure 1), the cache server calls upon a robot which will traverse the Web and will bring back documents, then by a stage of pre-filtrate. We discriminate the nonintersecting documents and which do not correspond to the profile of the community (definite in the form of characteristics through a meta-description of the community). These pre-selected documents will pass by a stage of analysis and extraction of metadata (automatic or semi-automatic). Thus, we obtain a database containing documents and their respective metadata; this base will be sorted manually to keep only documents that interest really the community.

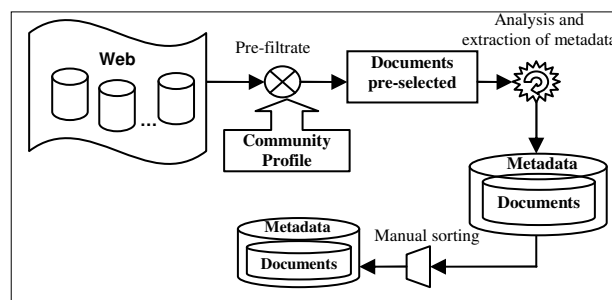


Figure 1. *structure of the extraction metadata and filtering systems.*

In the phase of inscription to the system, the user must seize his user profile and his centers of interest, then when this user connects himself to the cache system it has the possibility of lodging his own documents to feed the database while adding the metadata for each document (the author is the person more qualified to describe its document). This stage of description can be carried out in three manners: manual (by the author of the document), automatic (through an automatic system of extraction of metadata) or semi-automatic (if after the automatic extraction misses some elements, we will supply them manually). The database of documents can be alimented automatically or in a semi-automatic way starting from the Web (see① in figure 2); Research is multi criteria based on keywords and can be refined by the metadata, the user profile can be activated or deactivated to widen the fields of research. The user can also automatically receive the new documents lodged in the system and recommended by other expert users and who correspond to his his centers of interests through the system of pairing user profile - metadata. The community also has tools of communication and collaborations like the forums in order to allow the community members to exchange documents and ideas on specific topics to consolidate the spirit of the community.

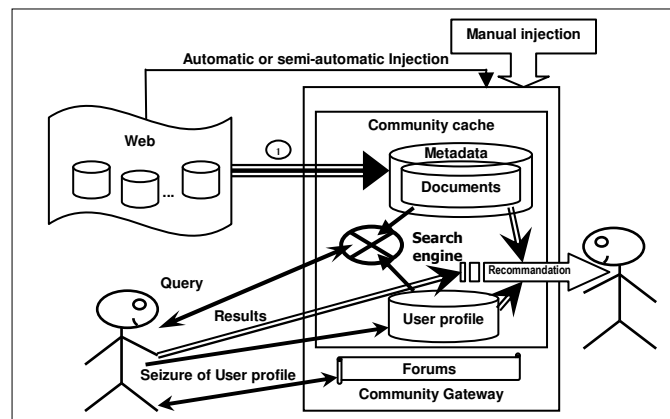


Figure 2. Total architecture of the system Community cache

The prototype was implemented in the form of Community gate, the extraction part of the metadata is developed for an automatic semi seizure of the metadata, the number of automatically extracted fields varies according to the format of the document. Now we are designing and producing a prototype of framework for the automatic extraction of the metadata and for various formats and types of documents used in the pedagogical context which one will integrate in our system of Community cache SYFAX. The prototype of the cache currently is under tests and validation in order to validate the results obtained and possibly to improve them.

6. Conclusion and prospects

In this paper we presented an information system dedicated to communities of users given. This is one of the major challenges in order to facilitate and to make more effective the search for information on the Web. That supports the sharing and collaboration between users who belong to the same community. The cache system proposes at the same time a methodology and tools to build and exploit information systems dedicated to communities of users. Recommended methodology consists in attaching metadata to the information shared by the community. The system of filtering integrated in the cache system depends of an approach of filtering by a pairing between metadata and user profile. The developed prototype enabled us to cure certain problems current search engines. It can also constitute a core for the lodging of the pedagogical documents for the service of the pedagogical community in order to exchange the experiments and to be used as a support for pedagogy. This work can be extended and improved in order to reinforce the concept of collaboration and communication between the members of a community.

References:

- [BRPL 04] http://www.brightplanet.com/deepcontent/deep_web_faq.asp#DeepWebSize
- [CACH] H. SMEI, A. BEN HAMADOU, Mr. MAKPANGOU, the caches Web on Internet, *GEI 2001 - Mars 2001 -Sousse - Tunisia*
- [DCED] Dublin Core Education <http://dublincore.org/groups/education/>
- [ESCLNet] European Schoolnet, <http://www.en.eun.org/>
- [GTME 01] Working group Metadata Education (GTME), CRDP of Montpellier June 2001, <http://www.ac-montpellier.fr/ressources/GTME1-1.doc>
- [GOGL] Google <http://www.google.com>
- [LAI 99] LAINE-CRUZEL Sylvie, "ProfilDoc, to filter exploitable information", *BBF*, T 44, n°5, 1999, p 60-64.
- [LEXMAIA] R.Charton, Lexicon of project MAIA: Autonomous Intelligent machine, 2001, http://www.loria.fr/equipements/maia/lexique/profil_utilisateur.html
- [LOM] LOM Learning Technology Comitee Standards, <http://ltsc.ieee.org/>
- [METAM] How Big Is The Internet? How Fast Is The Internet Growing? <http://www.metamend.com/internet-growth.html>
- [SFX1] H. SMEI, Mr. MAKPANGOU, A. BEN HAMADOU, SYFAX: A system of Semantic Web cache for Distributed communities, International Conference - *MediaNet 2002 - June 2002 - Sousse - Tunisia*
- [SFX2] H. SMEI, Mr. MAKPANGOU, A. BEN HAMADOU, Towards an Automatic System of Excavation and extraction of information: SYFAX, Case of pedagogy. *GEI 2002 - Mars 2002 - Hammamet - Tunisia*