Modeling and Construction of a user profile on the ontology's structure basis

Issam Abdelbaki[1], El habib Ben lahmar[2], Elhoussin Labriji[3]

[1] Department of Mathematics and Information Technology, faculty of sciences Ben M'SIK, University Hassan II – Mohammedia, Casablanca, Morocco

[2] Department of Mathematics and Information Technology, faculty of sciences Ben M'SIK, University Hassan II – Mohammedia, Casablanca, Morocco

[3] Department of Mathematics and Information Technology, faculty of sciences Ben M'SIK, University Hassan II – Mohammedia, Casablanca, Morocco

Abstract

Information research isn't a recent activity, it's a rediscovered one since this discipline seems to be more and more required. For instance, to know how to quickly and efficiently browse up information is extremely important. Information browsing systems tend to personalize the access to the information; their goal is to deliver to the user the appropriate information answering his needs, his interests and in general, his profile. Information browsing systems tend to model the user according to a profile then include it in the access channel to the information, in order to answer more efficiently to his needs.
This article describes a modeling and construction technic of a user profile based on a general ontology that uses its structure. It will also detail our method of concepts request extraction and similarities measurement within interests. In addition, we evaluated our approach by using user profiles on delivered results by a browsing website. Consequently, we will present some experimental results.

Keywords: *User profil, Ontology, Web semantics similarity measure, Information browsing system*.

1. Introduction

The general models of information research are built on the assumption that the user is presented by his request, therefore for a given request, browser systems shuffle the same results list even though the users have a different need. The recent works are oriented to a broad definition of the user, it's a research trend that aims to use the systems centered on users that are presented by a user profile.
The analysis of the user behavior is of a particular importance. For instance, by knowing how the user will develop strategies to search for information, it will be possible to propose meaningful information for research. Modeling profiles and how to adapt them to different users who do not have a clear idea about the information they are looking for will enable to offer personalized access to content of scientific papers based on the operating use profile.
There are several definitions of the user profile (Wahlster et al., 1986) defined it as follows: "A user profile (or user model) is a set of data about the user of a computer service. It is a source of knowledge acquisition that contains all aspects of the user that can be useful for the system's behavior. The user profile is extracted from the history of the user's requests, the goal is to find documents that were checked, and these documents are called "active" with respect to the query. The user profile is generally used by the information retrieval systems across the access to the information chain , further work is around the feedback from users when launching an application including systems Information Retrieval RID Distributed Peer-to-Peer , where a node can be both a client and a server. Indeed, on one hand, the user retrieves a list of results, on the other hand the search information feeds its knowledge base information provided by the user, including newspapers queries and traces of clicks to improve the relevance of results.
Our goal is to build a user profile based on ontology to semantically interpret the user query. In fact, we can find the same list of results for a single query submitted by users with different information needs. For example, for the "Apple" request, some users want to retrieve results treating the "Apple" brand, while others are interested in finding results treating the "Apple" fruit.
Our contribution focuses on the structural representation of the user profile based on the ODP ontology. In fact, ODP

is represented by a tree structure of concepts (referred to as categories), this structure can be regarded as a graph where each node is a concept. More specifically, our contribution is divided into three main areas: modeling and construction of user profile, the choice and definition of the focus of the user and the extraction of the concepts of user request. Finally, to test the effectiveness of our approach, we have applied a technique of reordering results in the Google search engine based on the similar user profiles. The first section gives an overview of the existing work, the second section presents our approach with different axes, the third section presents some experimental results evaluating the performance of our approach and finally the final section ends with a conclusion and gives an overview of our prospects.

2. Related works

The user interests center is represented by its request to the system of information retrieval. There are many representation techniques of the centers of interests that form the user profile. A naive representation of interests is based on key words, such as the case of web portals MyYahoo, InfoQuest, etc. There are other more elaborate representations to illustrate the interests of the user. (Sieg & al. , 2005 and Challam & al., 2007) illustrate it according to the present semantic concepts weighed general ontology, or as matrices by concepts (Liu and al . 2004).

(Gowan, 2003 & Sieg and al., 2004) proposed a model of the user profile based on a class of vectors each of which represents an area of interest of the user. Approaches to semantic representation operate as references ontology to represent user interests by the weighed vectors of concepts of the ontology used. We include the hierarchy of concepts of "Yahoo" or ODP as sources of evidence most often used in this type of approach. (Challam and al., 2007) builds the user profile on a technique of supervised classification of documents deemed relevant by a measure of similarity vector with ontology concepts of the ODP. This classification allows multiple search sessions, to associate with each concept of the ontology, a weight calculated by aggregating the similarity scores of documents classified under this concept. The user profile will consist of all concepts with the highest weight representing user interest centers. On the other hand (Sieg and al., 2005) simultaneously exploits the interests of the user represented by vectors of weighted terms and hierarchy “Yahoo” concepts. The user profile will consist of contexts each formed of the adequate concept representation for the research and the other is the concept representation to exclude from the search.

A matrix representation of the user profile is adopted in (Liu et al., 2004), the matrix is constructed from the search history of the user to incrementally establish categories representing the interests of the user and associated weighed terms reflecting the degree of interest of the user for every category.

3. Representation of the user profile based on ontology’s structure

In this section, we start with a representation of the reference ontology used to represent user profiles, then we present our approach to modeling and building user profile, we highlight the concept extraction methods of the user request and the method of calculating similarity of focus.

3.1 The ontology of reference

We chose to use the ontology domain ODP (Open Directory Project) as a reference, it is our source of semantic knowledge in the extraction of concepts of the application process. Semantic categories of an ontology are connected with the type relationships "is a", each category of the ODP represents a concept, it is the area of interest of a user. ODP editors are manually designed to match each concept to a set of web pages whose content matches the semantics related to the category. The ODP data is represented by two files of type "RDF", the first "Structure.rdf" contains the tree structure of the ontology and the second "Content.rdf" list web pages associated with each category. These pages are considered the most relevant for the query selected by domain experts. Every category is represented by a title and a description of its meaning, the description contains the titles and descriptions of its subcategories. Above is an excerpt from the architecture of the ODP.

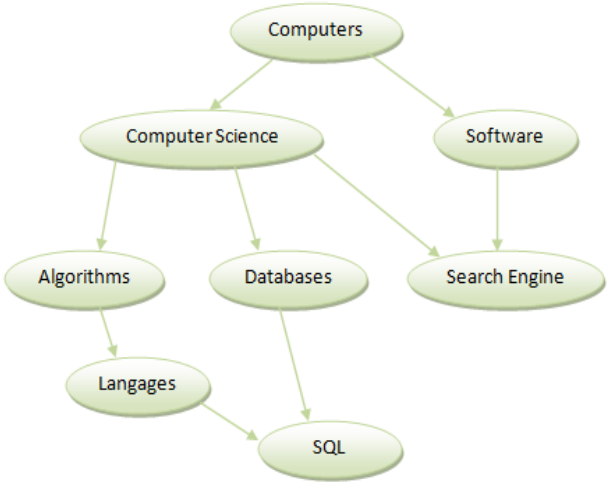

Figure 1: Extract of the ODP ontology architecture.

3.2 User profile presentation

We propose a semantic representation of the interests of the user based on the architecture of the ODP ontology. Indeed, as mentioned earlier the ODP data is represented by two files of type "RDF" (Structure.rdf and Content.rdf), the first file contains the tree structure of the ontology and the second file list pages associated with each category web. The aim is to represent the user profile using the same structure of the "Content.rdf" file to exploit the architecture of the ontology. We define the "Profil.rdf" file constituting the user center of interest, each category is referenced by the following information:

- Catid: Identifier of the concept,
- LastUpdate: The date of the last modification of the category
- Termes: Weighted list of the category's terms.
- Documents: List of active documents relative to the request
- Score: the score of the category.

Each concept (also called category) is represented by a vector of weighed terms selected from associated web pages (pages considered most relevant by the publisher of the ontology), this vector is stored in the property "Terms" in the file "Profil.rdf" to be used in the measurement of the extraction of the concepts of the user query. We detail in the following sections the process of representation of categories by a vector of weighed terms, as we give a detailed description of our concept extraction process.
After extracting the query concepts, the information retrieval system stores the user's interaction to determine the documents visited in relation to these concepts, these documents are called assets relatively to the query. Thus, for each concept, all these active documents are saved in the property "Documents" of the file "Profil.rdf". The score of a category is then the number of its uses by the user in order to give a weight to the user center of interest.
Finally, a similarity measure of the center of interest is applied between all concepts of the application and user profiles in stock so they are used in the access to information system.

3.3 Categories representation

In order to represent every category by a vector of weighted terms, we use the pages considered relevant defined by domain experts in the "Content.rdf" file. For this, we apply a study form using a tool for the automatic processing of natural language. We chose to use TreeTagger as a morph-syntactic analyzer. It is distributed free of charge for research purposes. It is a tool that allows annotating a text with information deemed relevant. It was developed by Helmut Schmid in the project "TC" in ICLUS (Institute for Computational Linguistics of the University of Stuttgart). TreeTagger allows labeling of German, English, French, Italian, Spanish, Bulgarian, Russian, Greek, Portuguese, Chinese and old French texts. It is adaptable to other languages if their lexicon and manually labeled corpus is available. Finally, it is possible according to our needs custom develop the desired specifications.
Following our needs we proceed as follows:

- Elimination: taking off insignificant words (in English: one, and, the, in, under …), these words are called "empty words".
- Segmentation: looking for elementary units that match the words.
- Recomposition: finding compound words.
- Lexical Analysis: Recovering words to a morphological basis form (gender, number).
- Lemmatization: consists on grouping words with the same origin.

Thus, each category noted Ci is represented by a vector Vi, the vector contains a list of weighed terms. Wij the weight of term Tj in the category Ci is calculated as follows:

$$wij = pij * \log(N/Ni)$$

- pij : degree of Tj representation capability in Di.
- N : number of subcategories.
- Ni : number of subcategories containing the Tj term.

3.4 Concepts Extraction

After representing every category of the ODP by the vector model, we proceed to the extraction of the concepts of the query by a vector similarity measure between vectors representing all categories of the ODP noted V(Ci) and the vector representing the query noted V(R). Indeed, the query is made by the vector of its significant terms, these terms are derived by applying the same survey form used in the previous phase. By applying the cosine similarity measure, the proximity of a request R to a class Ci is given by:

$$\cos(V(\mathrm{Ci}).V(R)) = \frac{card(E(V(\mathrm{Ci})) \cap E((V(R))}{\sqrt{card(V(\mathrm{Ci})) \cdot card(V(R))}}$$

- E(V) is the components of the V vector.

Mathematically it is considered that two vectors are similar when the cosine of the nail formed by these two has a

value greater than 0.8. Finally, the concepts with the most similar representative vectors to that of the query will be considered as concepts of the query.

Example:
Consider the query "course java". We Suppose that TreeTagger tool generated the vector representing the following query:

V(R) = {course, java}

We will calculate the similarity of the vector with that of the already given concept Ci generated.

V(Ci) = {development, informatique, software}

The similarity between the vectors is:

$$\cos(V(\mathrm{Ci}), V(R)) = \frac{1}{\sqrt{2 \cdot 4}} = 0.35$$

Since the cosine is less than 0.8, we know that the concept is not the query concept.

3.5 Measures of similarities of centers of interests

To use the profiles of users with the same interest center, we need to measure the similarity of their interest center. The interest center of the user is defined as a weighed set of concepts. In the literature, several studies have been developed, (Rada and al., 1989) have suggested that semantic similarity can be calculated based on taxonomic relationships " is-a ". More generally, the similarity computation may be based on the reporting relationships of specialization/generalization. One of the most obvious ways to rate the semantic similarity in taxonomy is to calculate the distance of the shortest path. The authors emphasize that this proposal is valid for all links hierarchical (is-a, kind -of, part-of, ...), but must be adapted to other types of links (cause, etc. .) . (Budanitisky and Hirst, 2001) compare five measures of semantic similarities or distances using WordNet (Fellbaum, 1998) (where the relation " is-a " is restricted to nouns and verbs). A complete state of the art is presented by (Patwardham, 2003) which compares these measures in relation to assessments made by human subjects. The vector space models are widely adopted (Bar, 99) and (Sal, 83). These approaches use a feature vector in a dimensional space, each representing the similarity concept and calculate based on the measure of the Euclidean distance or cosine.
We represented the center of interest of a user by a weighed set of concepts, the weight represents the rate of use by the user. Every concept is referenced by a vector of weighted terms. For a more accurate similarity at least, the interest center is represented by a vector of weighted terms of its concepts VC(Ti). On the other hand, the user expresses his need by a query containing a vector of weighed term VR(Ti). Thus, to infer the interests of the current user, we measure the similarity between these two vectors. We chose to use the cosine measure, applying the same principle to the concept of extraction.

In summary, for each user request, we deduce the concepts of the query that is the center of interest, then we select the profiles of users with the same interest center by measuring their similarity.

4. Experimental Evaluation

To evaluate our approach, we adopted the technique of reordering search results using user profile, our reordering function is based on the combination of the ranks of documents and score documents provided by all user profiles with the same interest center (we explained in the previous section, our method of calculating similarity of interests). Thus, examining these user profiles, the score of the document is its attendance rate compared to all concepts of the query.
The final score of the document is calculated by the following formula:

SF(Di) = SdR(Di) * (Nb – Rang(Di))

- SdR(Di) : Score of the Di document in respect to the query,
- Nb : Number of results provided by the browser
- Rang(Di) is the rank of the Di document.

We relied on two measures commonly used in classification, recall and precision. This is the "rate of return", ie the ratio between the number of relevant documents found during a search and the total number of existing relevant documents. The other indicator is the "accuracy rate" 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. These two concepts are often used because they reflect the point of view of the user: if precision is low, the user will be dissatisfied because it will waste time reading information that is not interested. If the recall is low, the user will not have access to information they wished to have. We measured our approach with 100 requests for multiple domains, the following figure shows the results for both precision and recall measures. The first tests presented in this figure are very encouraging. The comparison of our approach with existing ones shows that our approach is competitive.

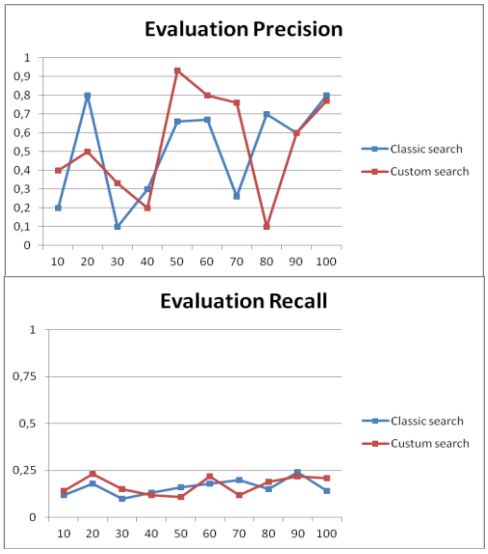

Figure 2: Evaluation precision / recall

5. Conclusion

We presented through this paper a method of modeling and construction of implicit user profile using the structure of the ODP ontology. We can use the wealth of theories graphs across the access to information chain. We intend to use our method to classify the results in our meta-search engine. We also plan to use it to develop a diagnostic system to assess our meta search engine.

References

[1] I.Abdelbaki, Z.Rachik, E.Ben lahmar, E.Labriji, Int.J.Computer Technology & Applications,Vol 4 (3), 2013,414-418.
[2] I.Abdelbaki, E.Ben lahmar, E.Labriji, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol 4 (2) , 2013, 194 – 198.
[3] Wahlster W. et Kobsa A.,Dialogue-based user models. In Proceedings of IEEE, Vol. 74(7), pp. 948-960, 1986.
[4] Gowan J., A multiple model approach to personalised information access, Master thesis in computer science, Faculty of science, Université de College Dublin, February, 2003.
[5] Challam V., Gauch S., Chandramouli A., « Contextual Search Using Ontology-Based User Profiles», Proceedings of RIAO 2007, Pittsburgh USA, 30 may - 1 june, 2007.
[6] R. Rada, H. Mili, E. Bichnell et M. Blettner, Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics. 1989. pp 17-30
[7] C. Leacock et M. Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification. In WordNet: An Electronic Lexical Database, C. Fellbaum, MIT Press, 1998.
[8] R.Baeza-Yates et B.Ribeiro-Neto. Modern Information Retrieval. ACM Press; Addison-Wesley: New York; Harlow, England; Reading, Mass., 1999.
[9] Auger, P., Drouin. Filtact© : un automate d'extraction des termes complexes. Terminologies nouvelles, (15), (1996). p. 48–49.
[10] G.Salton et M. J.McGill, Introduction to modern information retrieval. McGraw-Hill. New York, 1983.
[11] Lemay& Drouin. Two Methods for Extracting "Specific" Single-Word Terms from Specialized Corpora. International Journal of Corpus Linguistics, 10(2), (2005). p. 227–255.
[12] Tamine L., Zemirli W., Bahsoun. W., « Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information », Information - Interaction - Intelligence, Cépaduès Editions, 2007c.
[13] Daille, B., _E. Gaussier & J.-M. Langé. \Towards Automatic Extraction of Monolingual and Bilingual Terminology". Coling '94. Proceedings of the Fifteenth International Conference on Computational Linguistics. (1994). 515-521.
[14] H. Zargayouna et S. Salotti. Mesure de similarité sémantique pour l'indexation de documents semi-structurés. In 12ème Atelier de Raisonnement à Partir de Cas, 2004.
[15] BAZIZ. indexation conceptuelle/sémantique guidée par ontologie pour la recherche d'information, Thèse de Doctorat en informatique effectuée à l'Institut de Recherche en Informatique de Toulouse (IRIT). (2005).
[16] CLAVEAU. Acquisition automatique de lexiques sémantiques pour la recherche d'information, Thèse de doctorat, Université de Rennes 1. (2003).
[17] Ma Z., Pant G., Sheng, « Interest-based personalized search », ACM Transactions on Information Systems, 2007.
[18] Robertson S. E., Walker S., Hancock-Beaulieu M. M. Large Test Collection Experiments on an Operational, Interactive System: OKAPI at TREC. Information Processing & Management, vol. (31):345-360. (1995).
[19] Budanitsky, A. & Hirst, G. Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. In Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2001), Pittsburgh, PA.
[20] Patwardham S. Incorporating Dictionary and Corpus Information in a Measure of Semantic Relatedness, M.S. Thesis, August. (2003).
[21] Sieg A., Mobasher B., Burke R., Prabu G., Lytinen S., « Representing user information context with ontologies », uahci05, 2005.
[22] 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.