Web Information Retrieval Based on User Profile

Rachid Arezki[1], Pascal Poncelet[1], Gérard Dray[1], David W. Pearson[2]

[1] Centre LGI2P EMA, Site EERIE Parc Scientifique Georges Besse
30035 Nimes Cedex 1, France
{rachid.arezki, pascal.poncelet, gerard.dray }@ema.fr
[2] EURISE, Jean Monnet University of Saint-Etienne 23, rue du Docteur Michelon
42023 Saint-Etienne, France.
david.pearson@univ-st-etienne.fr

Abstract. With the growing popularity of the World Wide Web, the amount of available information is so great that finding the right and useful information becomes a very hard task for an end user. In this paper, we propose a new approach for personal Web information retrieval. The originality of our approach is a choice of indexing terms depending on the user request but also on his profile. The general idea is to consider that the need of a user depends on his request but also on his knowledge acquired through time on the thematic of his request.

1 Introduction

With the growing popularity of the World Wide Web, the amount of available information is so great that finding the right and useful information becomes a very difficult task. The end user, generally overloaded by information, can't efficiently perceive such information. In order to help the user in his task, the search engines available on the Web, propose through requests expressed by user in form of key words, a set of documents. Unfortunately, the quantity of returned results is also very large. Moreover, some relevant documents are often badly ranked and thus rarely consulted. *Blair&Maron* showed that the poor performance of IR systems is mainly due to the incapacity of users to formulate adequate requests [1]. Indeed, requests only formulated by key words express badly user information needs. In fact, these needs depend of course on the formulated request but also on the knowledge acquired by the user in his search domain: two users can formulate the same requests for different needs, and the same user for the same request may expect different answers in different periods of time [2]. For example, the results expected by an expert in java language who formulates the request "java course" are different from the results expected by a non expert which formulates the same request. A possible solution of this problem is to take into account the user profile in order to refine the ranks of the results returned by the Web search engines. In other words, the personalized Web information retrieval consists in finding a model able to consider efficiently user interests. In this article we present a new approach for information retrieval based on user

profile. The originality of our approach is a choice of indexing terms depending on the user request but also on his profile. The general idea is to consider that the need of a user depends on his request but also on his knowledge acquired through time on the thematic of his request.

We have developed *PAWebSearch*, a personal agent for Web information retrieval which supervises the user's actions and learns dynamically the user profile through the consulted documents (Web pages). For each information retrieval request carried out via a Web search engine (*Google, Yahoo,..*), *PAWebSearh* considers user request and results provided by the Web search engine for ranking these results according to the user profile.

The general principle of our approach is as follows. From a request q carried out by a user on a Web search engine, we recover all the results. Then, an analysis of user profile (user knowledge) allows us to obtain a set T of indexing terms. The construction of the indexing terms set T depends both on the user profile and on the user request q. We thus index all documents returned by search engine and request q according to the indexing term set T [1]. Then, to better adapt to the user's needs, the initial request vector q is transformed into q'. Proposing documents to the user is done by the calculation of similarities between the documents returned by the Web search engine and the request q'.

2 User profile representation

A user is defined by a tuple $p =< id, G >$, where id stands for a unique user identifier and G is a graph representing documents consulted by the user. The general idea is to analyze the content of the different documents and to store in the graph G co-occurrence frequency between various terms (words) of a document, as well as occurrence frequency of these terms. More precisely, $G= < V, E >$ is a labelled graph such as: (i) $V=\{(t_1, f_{t_1})..(t_n, f_{t_n})\}$ is a set of vertices of G, where each vertex (t_i, f_{t_i}) is represented by a term t_i and its frequency f_{t_i}. (ii) $E=\{(t_i, t_j, fco(t_i, t_j))/t_i, t_j \in V\}$ is a set of edges of G, where $fco(t_i, t_j)$ represents co-occurrence frequency between the terms t_i and t_j.

Algorithm 1: User Profile Learning
Input: consulted document d,
the user profile $p =< id, G >$
Output: updated user profile $p =< id, G >$
begin
1. construction of the co-occurrence graph G_d
2. **for** *each term* t_i *of* G_d **do**
  if $t_i \in G$ **then** $f^G_{t_i} = f^G_{t_i} + f^{G_d}_{t_i}$
  else
    create a new vertex (t_i, f_{t_i}) in the graph G such as
    $f^G_{t_i} = f^{G_d}_{t_i}$
3. **for** *each edge* $(t_i, t_j, fco(t_i, t_j))$ *of* G_d **do**
  $fco_G(t_i, t_j) = fco_G(t_i, t_j) + fco_{G_d}(t_i, t_j)$
end
$fco_G(t_i, t_j)$ represents the frequency of co-occurrence between terms (t_i, t_j) in the graph G.

[1] documents and requests are represented by vectors.

The co-occurrence frequency (or co-frequency) between two words is defined as the frequency of both words occurring within a given textual unit. A textual unit can be k words window, a sentence, a paragraph, a section, or the whole of document. In the framework of our user profile, we consider that textual unit corresponds to a sentence, thus $fco(t_i, t_j)$ represents co-occurrence frequency between terms t_i and t_j in the set of sentences of the documents consulted by the user. As shown in algorithm 1, for each new consulted document d, a graph G_d is built, then G_d is added to the graph G representing the user profile.

3 Information Retrieval Model

We consider in this section that a request q was sent to a Web search engine, and that we have a set X of returned documents, and let p be a user profile. Our information retrieval model can be presented as a tuple $< X, Q, P, T, s, f >$, where X represents the set of documents (i.e. document collection), Q stands for the set of requests, P is the set of user's profiles, T represents the term set indexing, s is a similarity or distance function and f is the term set construction function. For a given request q and a profile p we have $T = f(p, q)$.

Our motivation is to integrate effectively the user interests in the information retrieval process. Thus, the construction of the indexing term set T is done in a dynamic way and depends both on the user profile p and on the user request q (i.e. $T = f(p, q)$). For each new user request q, a new term set T is rebuilt. After the determination of the indexing term set T, the request q and each document of the collection X are represented by vectors according to the indexing term set T. To better adapt to the user's needs, the initial request vector q is transformed into q'. The transformation of q to q' requires the construction of the profile-request matrix (Sect. 3.2).

3.1 Indexing Term Set Construction

The choice of the indexing terms takes into account user profile as well as information retrieval request. Our motivation is to choose indexing terms reflecting the knowledge of the user in the domain of his search. As shown by the algorithm 2, the indexing terms are selected among the terms of the user profile which are in co-occurrence with the terms of the initial request.

Algorithm 2: Indexing Term Set Construction
Input: user request q,
the user profile $p =< id, G >$
Output: indexing term set T
begin
1. $T \leftarrow$ terms contained in the request q;
2. **for** *each term* t_i *of* q **do**
  for *each term* t_j *of* G *such as* $fco(t_i, t_j) > 0$ **do**
    if $\frac{(fco(t_i,t_j))^2}{ft_i \times ft_j} > \beta$ **then**
      $T = T \cup \{t_j\}$
end
β : constant representing the threshold of term selection.

3.2 Profile-request matrix

From the indexing terms obtained previously, we extract from the user profile p, the co-occurrence frequency matrix of the indexing term set T. This matrix represents semantic bonds between the various indexing terms. Let T_p be the set of terms contained in the user profile $p =< id, G >$. We call matrix *profile-request*, noted M_T, the square matrix of dimension $\mid T \times T \mid$ such that $T \subset T_p$, where each element m_{ij} of M_T is defined by:

$$m_{ij} = fco(t_i, t_j) \quad \text{where } (t_i, t_j) \in T^2$$

3.3 Request and document representation

From the profile-request matrix M_T, we can calculate the new request q' in order to adjust it according to the user profile. This request aims to reflect, as well as possible, the user interest in his search domain.

$$q' = (1 - \alpha) \times \frac{q}{|q|} + \alpha \times \frac{q \times M_T}{|q \times M_T|}$$

$|q|$ (repec. $|q \times M_T|$) is the Euclidean length of vector q (repec. $q \times M_T$),
α: threshold such that $0 \leq \alpha \leq 1$, allowing hybridation between initial request $\frac{q}{|q|}$ and the enriched request $\frac{q \times M_T}{|q \times M_T|}$, the higher α is the more the user profile is considered.

4 Conclusion

We proposed in this paper a new approach for personalized information retrieval. The proposed model allows a better consideration of the user's interests in the information retrieval process by: (i) A choice of indexing terms which reflects as well as possible the user knowledge in his search domain. (ii) An enrichment of the user request by the matrix of profile-request.

In the models where the user is represented by vectors of terms, an iterative process of user profile re-indexing is necessary to take into account of new indexing terms. In our model none re-indexing of user profile is necessary, therefore it is very adapted to the Web, where information is very heterogeneous. The first experimental results carried out with *PAWebSearch* confirm the relevance of our approach. One of the prospects for research, is the application of the indexing term set construction method in the framework of a standard information retrieval model.

References

1. D.C. Blair and M.E. Maron. An evaluation of retrieval effectivness for a full-text document retrieval system. *Communication of the ACM*, 28(3):289–299, 1985.
2. C. Danilowicz and H.C. Nguyen. Using user profiles in intelligent information retrieval. In *Proceedings of the 13th International Symposium on Foundations of Intelligent Systems*, pages 223–231. Springer-Verlag, 2002.