Custom Fusion Based on Borda

https://doi.org/10.3991/ijes.v4i3.6300

I. Abdelbaki, E. Benlahmar and E. Labriji
University Hassan II, Casablanca, Morocco

***Abstract*—Searching for information on the Internet is not only an activity newly rediscovered, but also a strategic tool to achieve a wide variety of information. Indeed, it's extremely important to know how to find the information quickly and efficiently. Unfortunately, the Web is so huge and so little structured, that gathering precise, fair and useful information becomes an expensive task. In order to define an information retrieval tool (meta search engine) that brings together multiple sources of information search, interest must be credited to the merger phase of search engines results. On the other hand, information search systems tend primarily to model the user with a profile and then to integrate it into the information access chain, to better meet its specific needs.**

This paper presents a custom fusion method based on Borda method and values retrieved from the user profile. We evaluated our approach on multiple domains and we present some experimental results.

***Index Terms*—Information search system, meta-search engine, Fusion, Borda, User profile.**

I. INTRODUCTION

Search engines are the most visited sites on the web, they are used by 85% of users (Schwartz, 1998). However, they index only a fraction of all available information and their coverage does not increase as rapidly as the size of the Web. Thus, the user is quickly lost in finding relevant information. Meta search turns out be a powerful way to work around this problem, by bringing together multiple sources of information search (search engine) in a single unified tool (meta search engine). However, among all the problems related to the meta research, lies the fusion and the classification of the search engines results.

Having obtained an ordered list of documents from each engine, meta engines should then merge these responses in order to present a single list to the user. The response quality of meta-search depends on the classification strategy. To solve this merger problem, several works have emerged. (Selberg, 1999) proposed a strategy named "everyone has his turn", it builds the final list by taking an element of each list in the different engines by descending order. (Yager and Rybalov, 1998) suggest a policy named "everyone has his turn" giving greater importance to the lists longer than to the rank of documents. Sometimes, search engines provide a score representing the similarity degree between the request and the document. This strategy is called "fusion by score"'. However, search engines apply heterogeneous classification algorithms, so we cannot normalize the score provided by the search engine.

Analysis of user's behavior reveals a particular importance. Indeed, it is by knowing perfectly how the user is going to develop its information retrieval strategies that it will be possible to propose significant information for his research. Modeling profiles and how to adapt them to different users who do not have a specific idea of the information that they are looking for, allows us to offer personalized access to scientific papers based on the user profile exploitation. The user profile is extracted from the history of requests from users, our goal is to know the search engines and the documents in which there was consultation, these two elements are called the request "assets".

We propose a fusion method based on Borda voting system, it regards the fusion of search engines results as a poll to vote for a collection of candidate documents by order, which voters are search engines. it was already processed in this area, on the other hand, the population of voters was not considered. In order to properly apply the method of Borda, we need to know the popularity of each search engine which represents its weight relative to the request. This value depends heavily on the specific user needs, so we assign for each search engine a weight based on the user profile.

The first section presents the most used fusion method, the second section presents some work in personalized information retrieval, the third section presents our approach with its different axes, the fourth section presents some experimental results evaluating the performance of our approach and finally in the last section, was completed by a conclusion and an overview on our perspectives.

II. GENERAL INFORMATION SEARCH

Having obtained an ordered list of document from each engine, the meta engines must merge these responses in order to present a single list to the user. The quality of the meta engine response depends strongly on the ranking strategy.

To solve this merger problem, several works have emerged. (Selberg, 1999) proposed a strategy named "everyone has his turn", it builds the final list by taking an element of each list in the different engines by descending order. (Yager and Rybalov, 1998) suggest a policy named "everyone has his turn" giving greater importance to the lists longer than to the rank of documents. Sometimes, search engines provide a score representing the similarity degree between the request and the document. This strategy is called "fusion by score". However, search engines apply heterogeneous classification algorithms, so we cannot normalize the score provided by the search engine. WebSum (ALO Jeanne El Jed, 2005) applies new criteria to the results provided by the search engine to reclassify the pages in relevance order for the request after checking the language and information form.

The merger may also take place under the probability estimated by logistic regression (Bookstein et al., 1992) on the basis of rank and score obtained by this document (Le Calvé & Savoy 2000). Yet, (Glover et al. 2001) use a decision theory to classify the results from various search engines.

Other methods are based on scores combination. CombSUM operator introduced by (Fox & Shaw, 1994), combines scores linearly. Indeed, the different sets considered in the merge receive the same weight. The operator CombMNZ is an extension of CombSum. The documents scores that have been found by more than one system are reinforced by being multiplied by the agreements number. However, is the reasoning of the CombMNZ operator is beneficial even if the systems share a significant number of non-relevant documents? To remedy this problem, the operator CombHMEAN combines the scores by taking the harmonic average. Finally, the Borda method proves to be a conventional method in the theory of collective choice.

III. CUSTOM INFORMATION SEARCH

Implementation of customized information research systems mainly consists of two main phases: the user modeling in a pattern that is the learning phase, and the integration of this profile in one of the access to information phases. We present in this section the main approaches used in these two phases.

A. User profile representation

The user center of interest is represented by his application submitted to IRS (Information Retrieval System), there are several interests representation techniques to constitute the user profile. A naive interests representation is based on key words, such as the web portals case My-Yahoo, InfoQuest, etc.. There are more elaborated representations to illustrate the user interests. (Gowan, 2003 and Sieg et al., 2004) represent the interests according to weighted vectors terms, and (Sieg et al., 2005 Challam et al., 2007) present them semantically following weighted concepts of general ontology, or as concepts matrices (Liu et al. 2004).

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

A matrix representation of the user profile is adopted in (Liu et al., 2004), the matrix is

constructed from the user search history to incrementally establish categories representing the user interests and associated weighted terms reflecting the interest degree of the user for each category.

B. User profile exploitation

Integrating user profile in the Information Retrieval process returns to operate in the reformulation and calculation of relevance score or search results ranking. (Sieg et al., 2004) offers a personalization based on queries refinement to describe a richer query translating the proper search context using a variant of the Rocchio algorithm. Indeed, the research context is represented by a pair of classes in the hierarchy of "Yahoo" categories, the first is the correct query category and similar to one of the user's interests, the second represents the category to be excluded during the search.

Other works include the user profile in the matching function query-document. (Tamine et al., 2007a) exploit interest centers in the pairing of the IR model. The value relevance of a document to a query is no longer based on the query alone but in addition to focusing on the user who submitted it.

Finally we find the personalization approaches (Challam et al, 2007.) (Ma et al, 2007.) (Liu et al, 2004) based on the search results: they are based on the combination of the initial document rank and the rank resulting from a similarity between the document and the user profile.

IV. CUSTOM FUSION BASED ON BORDA

Our approach is an adaptation of Borda model to search meta-engines, the fusion of search engines results can be looked as an election in which the search engines are the voters, each search engine suggest an ordered list of documents , whose documents are the candidates. To properly apply the Borda method, we need to know the popularity of each search engine (voter), it's the search engine weight regarding the user query. This value depends highly on the specific user need, it's the number of users who voted for this search engine, in other words, is the number of users who visited his returned pages. Therefore, we can recover it from the user model, this model will contain the history of visited pages and the search engines which gave in results these pages. Thus, we distinguish two main phases, the feeding phase of user model (called the learning phase) and the phase of fusion.

A. Learning phase

In this section we are presenting our user model so that the meta search engine can use it in the classification phase. In our case we need two information, namely the relationship between the query terms and documents and the relationship between the query terms and search engines. At first, the query terms are extracted, and then we save the user interaction in our knowledge base.

1) Terms extraction

To retrieve the query terms, we chose to implement a form study as follows:

- Segmentation: Find basic units corresponding to words.

Example: You're (We need to identify the separator, in this case «'» is not a separator).

- Recomposition : Find compound words.
- Lexical Analysis: Bring words to a morphological base form (conjugation, gender, number).
- Stemming: is to group words that have the same origin.

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

2) User profile construction

Based on user interactions, we recover information about the application, i.e. the query identifier, the query terms, the consulted documents and search engines associated with these documents. Indeed, when the user enters a query, it consults some documents, and search engines that gave these documents as result are deducted. These search engines and documents are called active in relation to the query.

User profiles are stored in our knowledge base so they can be used in the classification phase.

Example:

A request R contains terms (T1-T2-T3) with multiple results, the user has viewed a set of documents (D1-D2), the search engines (M1-M3-M4) gave results in these documents, so these search engines and these documents are considered assets in relation to the request R.

Specifically, each query has an identifier and has a list of weighted terms and a set of active search engines and active documents in relation to the search query.

B. Classification phase

Our approach is a Borda adaptation model to the metasearch engine, merging the results of search engines can be seen as an election in which search engines are the voters, each search engine provides a list of documents (which makes documents candidates).

Furthermore, we intend to give a score symbolized by "SdR" to documents related to the query, The score represents the documents level of occurrence among the old active documents of the query concepts based on the knowledge base fed during the learning phase. On the other hand, we also intend to give weight to the search engine. In other words, knowing the score of the search engine compared to the query "SeR". By examining our knowledge base, the weight of the search engine is the importance of search engine compared to the query concepts. The overall score of a document D(i) compared to the query (symbolized SG(Di)) is calculated as follows:

$SG(Di) = SdR(Di) * \sum_{j=1}^{N} SeR(Mj) * R(Di, Mj)$

- SdR(Di) : Score document Di compared to the query,
- SeR(Ej): Score of search engine Ej compared to the query,
- R(Di, Mj) : Nr(Mj) – rang(Di, Mj),
- Nb(Ej): Number of documents resultant from search engine Ej + 1,
- rank(di, Ej) : The rank of the document Di in the search engine Ej
- N: Number of search engine that responded to the request.

Example: Considering four search engines E1, E2, E3 and E4 with the SeR scores 30, 22, 23, 25 calculated from the knowledge base We chose to work only on the first 4 request results given by each search engine, each document Di will have a score SdR(Di) calculated from the knowledge base (the documents have the same score for all the search engines), we assume that we only have 4 documents D1, D2, D3 and D4, their scores for the query are respectively 34, 20, 24, 10.

TABLE I.
RESULTS OF VARIOUS SEARCH ENGINES

M1(SmR=30)	M2(SmR=22)	M3(SmR=23)	M4(SmR=25
D1(SDR=34)	D3(SDR=24)	D2(SDR=20)	D1(SDR=34)
D3(SDR=24)	D2(SDR=20)	D1(SDR=34)	D2(SDR=20)
D2(SDR=20)	D1(SDR=34)	D3(SDR=24)	D3(SDR=24)
D4(SDR=10)	D4(SDR=10)	D4(SDR=10)	D4(SDR=10)

This leads to the following counting points:

TABLE II.
SCORE CALCULATION

Documents	1er	2ème	3ème	4ème	Points
D1(SdR=34)	30 + 25	23	22	0	(30*4+23*3+22*2+25*4)*34=11322
D2(SdR=20)	23	22+25	30	0	(23*4+47*3+30*2)*20=5860
D3(SdR=24)	22	30	22+25	0	(22*4+30*3+47*2)*24=6528
D4(SdR=10)	0	0	0	100	(100*1)*10=1000

Therefore, the classification would be:

V. EVALUATION

To experimentally evaluate the performance of our fusion method results described in this article, we chose the web page collection used in the ninth TREC conference corpus named TREC9. We relied on two measures commonly used in classification, recall and precision.

A. Measures used

We use two measures, recall and accuracy, 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" which is the ratio between the relevant documents number found during a search and the total documents number found in response to the question. These two concepts are often used because they reflect the user point of view: if precision is low, the user will be dissatisfied because he'll waste time reading information that is not interesting. If the recall is low, the user will not have access to information they wished to have.

B. TREC Collection

Since there is currently no standard framework to evaluate a personalized access model to information, we propose an evaluation framework based on "TREC Collections"(Text Retrieval Conference), it is a American conference whose purpose is to allow comparison between the performances of information retrieval systems that exploit large volumes of data, it brings together toolkits and software information retrieval (in full text) designers. It has become a reference and an international standard in the field of information evaluation.

We chose to evaluate our model using the TREC9 collection, it includes 1692096 web pages written in English for a volume of 11,033 MB.

C. Learning phase

As a first step, we need to enrich our knowledge base. For this, we launched 10,000 applications to build the knowledge base.

D. Experimental results

We measured our approach by 1000 query of several areas, 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 knowing that our knowledge base is powered over water so the results will be progressively more relevant.

The relevance evaluation of our meta-search engine is being developed. In this article context, we conducted preliminary experiments to give a rough idea about the quality of our fusion method.

TABLE III.
EVALUATION ACCURACY

Number of request	Accuracy CombMNZ	Accuracy RankcomMNZ	Accuracy FPB
100	0,8402	0,8657	0,8793
200	0,8511	0,8693	0,885
300	0,8497	0,8697	0,8765
400	0,8483	0,8567	0,8793
500	0,8596	0,8657	0,886
600	0,8545	0,8697	0,8765
700	0,8593	0,8687	0,8783
800	0,8580	0,8677	0,8793
900	0,8585	0,8677	0,8810
1000	0,8599	0,8697	0,8820

TABLE IV.
EVALUATION RECALL

Number of request	recall CombMNZ	recall RankcomMNZ	recall FPB
100	2,1562	2,1657	2,1793
200	2,1571	2,1693	2,1820
300	2,1537	2,1697	2,1795
400	2,1543	2,1677	2,1793
500	2,1596	2,1693	2,1810
600	2,1575	2,1697	2,1795
700	2,1593	2,1687	2,1810
800	2,1580	2,1687	2,1810
900	2,1585	2,1687	2,1810
1000	2,1599	2,1687	2,1810

VI. CONCLUSION AND PERSPECTIVES

We presented through this paper a method that represents a custom merge using the Borda method in the results ranking in meta search engine. Thus, we took into account all factors, namely the document score, the search engine score and rank document proposed by the various search engines. We also conducted experiments to evaluate the performance of our meta search engine.

Various improvements can still be proposed, one of our goals is to extract query concepts to treat user query semantically, so we can enrich the query concepts extracted before it is sent to search engines.

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] R. Mghirbi, K. Arour, Y. Slimani et B. Defude, « Un modèle comportemental d'interclassement de résultats dans un système de recherche d'information P2P », Actes du XXVIII° congrès INFORSID, Marseille, mai 2010.

[4] YeeW. G., Frieder O., « On search in peer-to-peer file sharing systems », SAC '05 : Proceedings of the 2005 ACM symposium on Applied computing, ACM, New York, NY, USA, p. 1023-1030, 2005. http://dx.doi.org/10.1145/1066677.1066913

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

[6] 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. (Cité page 32.).

[7] Beitzel, S. M., E. C. Jensen, A. Chowdury, D. Grossman, O. Frieder et N. Goharian. 2004, On fusion of effective retrieval strategies in the same information retrieval system _, Journal of the American Society of Information Science & Technology, vol. 50, no 10, p. 859_868.

[8] Wahlster W. et Kobsa A.,Dialogue-based user models. In Proceedings of IEEE, Vol. 74(7), pp. 948-960, 1986. http://dx.doi.org/10.1109/PROC.1986.13574

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

[10] Gowan J., A multiple model approach to personalised information access, Master thesis in computer science, Faculty of science, Université de College Dublin, February, 2003. TREC, ≪ Text REtrival Conference ≫, 2008.

[11] Shen D., Chen Z., Yang Q., Zeng H., Zhang B., Lu Y., MaW., «Web-page classification through summarization », In Proceedings of the 27th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, Sheffield, South Yorkshire, UK, p. 242-249, 2004.

[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] H. Zargayouna, S. Salotti, et C. Golbreich. Raisonnement par similarité pour l'indexation et la recherche dans des documents multimédia. In Complément des actes des journées francophones d'Ingénierie des Connaissances (IC'2001), 2001.

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

[16] Ma Z., Pant G., Sheng, « Interest-based personalized search », ACM Transactions on Information Systems, 2007. http://dx.doi.org/10.1145/1198296.1198301

[17] Schwartz C. Web Search Engines, Journal of the American Society for Information Science. vol. (49):973-982, (1998).

[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).

AUTHORS

Issam Abdelbaki, El habib Benlahmar and **El Houssine Labriji** are with University Hassan II, Department of Mathematics and Informatics, Casablanca, Morocco (e-mail: najeh_khadija@yahoo.com).

Submitted 26 September 2016. Published as resubmitted by the authors 23 October 2016.