ISSN:0975-9646

Merge the Results of Search Engine

Abdelbaki Issam, Benlhmar El Habib, Labriji Elhoussin

Faculty of Sciences BEN M'SIK, University Hassan II – Mohammadia
Casablanca Maroc

Abstract— **Owing to a progressive quantity of the document in the web, the user feels lost to find the most relevant information. The metasearch turns out to be a powerful way to help the user by combining several sources of information (search engine) into a single unifying tool (metasearch engine). However, the metasearch can face some obstacles and the most critical is the fusion and the rearrangement of search engine results.**

This paper presents a fusion approach based on the moderate voting system Borda, we see the fusion of results from different search engines as an election to elect a set of candidate documents in a certain order, and thus it defines the search engines as electorates. To properly apply the Borda Method, we need to know the popularity of each research engine (number of electors for each search engine), it represents the weight of the search engine to the query. This value heavily depends on the specific need of the user, so we can give a weight to each search engine relative to the user query according to his profile. We have evaluated our approach on the collection of TREC documents and we will present some experienced results.

Keywords— **Merge, Ranking, User profile, Borda , Information retrieval system, Metasearch engine, Vote.**

I. INTRODUCTION

Searching for information on the Internet is not only an essential way to get information, but also a strategic tool to obtain a wide variety of information. Unfortunately, Web is so huge and so unstructured that the collection of specific, fair and useful information becomes a daunting ant time-wasting task. To define information research tool (metasearch engine) which has the aim to combine several sources of information, the main interest should be brought to the fusion step of results from search engines.

The common fusion methods are global, including the classical vector model which was the foundation of models based on scores, languages and the rank, especially tourniquet model also known by Round Robin model (Greengrass, 2000)) in addition to Borda Count model.

We envisage a fusion approach based on the moderate voting system Borda, that defines the fusion of results from different search engines as an election to elect a set of candidate documents in a certain order, and defines the search engines as electorates.

To apply the Borda Method, we need to make a distinction between the popularity of each research engine relatively to a special user need. As a conclusion, the weight given to research engine is relative to the user profiles. As a consequence, we intend to apply a supervised classification which is a process involving two phases: training and classification.

The first section shows an overview of the most known fusion and ranking methods, the second section represents the user concept profile and fusion methods based on profile, the third section represents our approach with different axes particularly its main phases, the fourth section represents some experimental results evaluating the performance of our approach, and finally in the last section, we end with a Conclusion and we provide an overview of our perspective.

II. CLASSIC MERGE METHODS

Based on an ordered document list from each search engine, metasearch engines must merge these data in order to put on view one list to the user, the quality of data of the metasearch depends mainly on the ranking strategy.

In order to overcome the fusion obstacles, several studies have emerged. (Selberg, 1999) proposed a strategy called "each one in his turn", he built the final list by taking an element of each list of different engines in decreasing order. (Yager and Rybalov, 1998) suggest a strategy called "each one in his turn" by giving more importance to the longest lists than documents rank. Sometimes search engines provide a result representing a similarity between the request and the document, this strategy is called "merger by the score". However, search engines apply heterogeneous ranking algorithms, therefore we cannot standardize the score provided by the search engine.

WebSum (Olfa Jenhani El Jed, 2005) applies new criterions to sort results provided by the search engines, by classifying the pages, starting by more relevant ones, after verifying the meaning and the linguistic form of to the request.

The fusion may also be done under the probability estimated by the logistic regression (Bookstein and al. 1992) based on the rank and the score obtained by this document (Le Calvé and Savoy 2000). While, (Glover and al. 2001) use a decision theory to classify results coming from different search engines.

Other methods are based on a scores combination. For example, the CombSUM operator introduced by (Fox and Shaw, 1994), which combines scores thanks to linear method. In fact, the various entities used in the fusion receive the same weight. The operator CombMNZ is an extension of CombSum. As a matter of fact, documents scores that have been found by more than one system are strengthened by being multiplied by the number of

synchronization. It's a fact that the operator CombMNZ logic turned out to be very useful, but would it still if be useful if the systems share a number of meaningless documents? To overcome this issue, the operator CombHMEAN combines the scores by taking the adjusted average. Finally, the Borda method has proved that it is a classical method in a matter of choice theory.

III. CUSTOM MERGE METHODS

Implementation of customized research systems information depends mainly on two main phases: Giving the user a model according to a certain profile, in other words, the learning and integration phase of that profile in the classification phase. In this section, we will present the main used approaches in these two phases.

A. The representation of a user profile

The user center of interest is represented by his submitted query to the IRS. There are several techniques of the user profile representation of Center interests.

A Simple representation of interests is based on key words, such as the case of web portals MyYahoo, InfoQuest, etc…

There are other more highly structured representations to illustrate the interests of the user. (Gowan, 2003 and Sieg and al., 2004) have represented the center of interests as vectors of weighted words, on the other hand (Sieg and al., 2005 and Challam et al., 2007) have represented it semantically based on concepts of a general ontology, or as matrices of concepts (Liu et al., 2004).

(Gowan, 2003) (Sieg and al., 2004) have proposed a model of the user profile according to classes of vectors. Each class represents the center of interest to the user, the "centroids" of classes thus represent the interests of the user.

Semantic representation approaches uses reference ontology that allows representing the interests of the user based on concepts vectors of the ontology used. We take as references the hierarchy of concepts of "Yahoo" or ODP 3 as evident sources which are most often used in this type of approach. (Challam and al., 2007) have built the user profile on a supervised classification documents technique recognized relevant according to a measure of similarity vector with ontology concepts of the ODP. This classification allows in multiple research sessions to associate to each concept of the ontology, a weight calculated by aggregating the similarity scores of documents classified under that concept.

The user profile will be made up of all concepts with the highest importance representing the interests of the user. On the other hand (Sieg and al., 2005) have simultaneously used the interests of the user represented by vectors of weighted words and the concepts hierarchy of "Yahoo". The user profile will consist of contexts, each formed by a pair of concepts of the hierarchy: one is the appropriate concept for research, and the other is the concept to be excluded in the search.

A matrix representation of the user profile has been adopted by (Liu and al., 2004). The matrix is constructed from the search history of the user in incrementally way to develop categories representing the interests of the user and the associated words reflecting the degree of interest of the user for each category.

B. Use of a user profile in the classification phase

Integrating a user profile in the IR process means using it in one of the following major phases: reformulation, calculating relevancy score or ranking of search results.(Sieg et al., 2004) offers a personalization based on the refinement of queries to describe a richer query translating the proper context to search 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 appropriate category to the query similar to one of the centers of interest of the user and the second represents the category that must be excluded during the search.

Other works include the user profile in the matching function query-document. (Tamine and al.2007a) exploits "field of interests" in the matching function of the IR model. The relevance of a document towards a query is no longer only based on the query itself but also on the field of interest of the user who submitted it.

Finally we find customization approaches (Challam and al., 2007) (Ma and al., 2007) (Liu and al., 2004) based on the reordering of search results, it is based on the combination of the initial rank and the rank of the document resulting from a measure of similarity between the document and the user profile.

IV. OUR APPROACH

Our approach is based on two main phases, namely the learning phase used to create a knowledge base so that the ranking algorithm may merge results of search engines at the ranking phase.

A. Learning phase

In this phase, there are three main steps. The first one is an extraction of information from the user navigation background in a XML format log file. The second is the construction of the formal context from the log file generated in the previous step. The third consists on a profile creation using the previously generated formal contexts.

1) The generation and update of a log file

Based on user interactions, we get information about the query: the query identifier, the terms, the documents consulted and the search engines associated to documents. In fact, when the user enters a query, he consults some documents, that exposed their sources: search engines. Those search engines are active compared to the request.

Example

A query 'R' contains terms (T1-T2-T3) which has several results; the user has selected a set of active documents 'D' (D1-D2) associated to a set of search engines 'M' (E1-E3-M4).

```
<Requêtes>

<Requête identifiant = "R1">
    <Concepts>
        <concept>T1</concept>
        <concept>T2</concept>
        <concept>T4</concept>
        <concept>T6</concept>
    </Concepts>
    <Moteurs>
        <moteur>M1</moteur>
        <moteur>M2</moteur>
    </Moteurs>
    <Documents>
        <document>D1</document>
        <document>D2</document>
        <document>D3</document>
    </Documents>
</Requête>

<Requête identifiant = "R2">
    <Concepts>
        <concept>T1</concept>
        <concept>T2</concept>
        <concept>T4</concept>
    </Concepts>
    <Moteurs>
        <moteur>M1</moteur>
        <moteur>M10</moteur>
        <moteur>M5</moteur>
    </Moteurs>
    <Documents>
        <document>D1</document>
        <document>D2</document>
        <document>D7</document>
    </Documents>
</Requête>

<Requête identifiant = "R3">
    <Concepts>
        <concept>T1</concept>
        <concept>T2</concept>
        <concept>T3</concept>
    </Concepts>
    <Moteurs>
        <moteur>M1</moteur>
        <moteur>M3</moteur>
        <moteur>M4</moteur>
    </Moteurs>
    <Documents>
        <document>D1</document>
        <document>D2</document>
    </Documents>
</Requête>

<Requêtes>
```

Illustration 1: log file generated

Each query has an identifier and has as a subset list of: terms and an active set of search engines and active documents.

2) Generation of formal contexts

'O' is a set of objects, 'P' a set of property and 'R' a binary relation between P and O.

A formal context is defined by the triplet (O, P, R). The elements of 'O' are called objects and the elements of 'P' are called the properties of the context.

To express that an object o of 'O' is related to a property p of 'P', we write oRp. This means that object o has property p.

In our case, the terms are the objects, properties are either active documents or active search engines, and we define two types of context:

- Context Document Term (CDT): defines a relationship between a set of terms (objects) and a set of documents (property).
- Context Document Engine (CDE): defines a relationship between a set of terms (objects) and a set of search engines (property).

In our case, we say that an object Oi has the property Pj when this latter is always present in the presence of the object Oi. It can be represented by a matrix where 1 means the object Oi has the property Pj and 0 otherwise.

Example:

	O1	O2	O3	O4	O5
P1	1	1	1	0	0
P2	1	0	0	1	1
P3	0	1	1	1	1
P4	1	1	0	1	0

Chart 1: relationship between object and property

3) Generation of profiles

From CDT and CDE contexts we extract two kind of profile, the first represents the link between the ran requests and the active search engines, which is called Profile Engine Term (PET), the second represents the link between the ran requests and actives documents , which is called Profile Document term (TDP), defined as: ({m1, ..., mi}, {t1, ..., tj}), respectively, ({d1, ..., dt} ; {t1, ..., tk}) knowing that {m1,. . . , Mi} is a set of search engines that have in common the set of terms {t 1,. . . , Tj} and {d1,. . . , Dt} a set of documents that have in common all of the terms.

All the profiles represent a cover, in our case, we have two types of cover for a PET symbolized by 'C1' and the other for PDT symbolized by 'C2', both covers represent a knowledge base generated during the learning phase symbolized by B (C1, C2).

In Chart 1, the objects {O1, O2, O4} have properties {P2, P3, P4}. In this case we can define a profile P = ({O1, O2, O4}, {P2, P3, P4}).

B. Ranking 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 presence rate of the document 'result' among the old documents of similar requests from the cover C2. The similarity is calculated as follows:

Ta and Tb is the set of query terms a and b. The similarity is defined using the following formula (by Salton, 1989):

$$\text{Sim (Ta, Tb)} = \frac{|\text{Ta} \cap \text{Tb}|}{|\text{Ta} \cup \text{Tb}|}$$

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 SMR. By examining our knowledge base, specifically the cover C1, the weight of the search engine is the importance of search engine compared to the query.

The overall score of a document D(i) compared to the query is calculated as follows:

$$\text{SG(Di)} = \text{SdR(Di)} * \sum_{i=0}^{N} SeR(Ej) * (Nb(Ej) - rank(Di, Ej))$$

- SdR(Di): Score document Di compared to the query
- SeR(Ej): Score of search engine Ej compared to the query
- rank(di, Ej) is the rank of the document Di in the search engine Ej
- Nb = Number of documents resultant from search engine Ej + 1

Example:

Considering four search engines M1, M2, M3 and M4 which have a 30%, 22%, 23% and 25% of popularity calculated through the cover C1. Each engine provides four results of a given query, each document has a score SdR (Di) calculated through the cover C1.

We admit having only 4 documents: D1, D2, D3 and D4, and their scores compared to the query are respectively 34, 20, 24 and 10.

M1(30%)	M2(22%)	M3(23%)	M4(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)

Illustration 2: results of different search engines

This leads to the following counting points:

Documents	1re	2e	3e	4e	Points
D1	30 + 25	23	34	0	(55*4+23*3+34*2)*34 = 357 * 34 = 12138
D2	23	22+25	30	0	(23*4 +47*3 + 30*2)*20 = 293 * 20 = 5860
D3	22	30	22+25	0	322 = (22*4 + 30*3 + 47*2) * 24 = 272 * 24 = 6528
D4	0	0	0	100	(100*1)*10 = 1000

Therefore, the classification would be:

Final Classification	D1	D3	D2	D4

V. EVALUATION

In order to validate our proposals, we have conducted experiments to evaluate the impact of our Borda fusion method - based on the user profile - on the system performance. On the other hand, we compared our personalized “Borda fusion method” with two types of fusion methods, the first is based on the score ComMNZ and the second on the rank RankcomMNZ.

We used two measures as basic indicators to test the effectiveness of the methods, it is the "rate of return", ie the ratio between the number of relevant documents found during a search and the total number of relevant documents existing in the system. The other indicator is the "accuracy rate" which corresponds to the ratio between the number of relevant documents found during a search and the total number of retrieved documents in response to the question.

A. *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 NIST Collection (discs 4-5) of the TREC evaluation with a size of 741,670 documents.

B. *Learning Phase*

At first, we need to expand our knowledge base. To this end, we launched the first 10,000 requests to build a log file for each initial peer. Subsequently, we launched the profile management module to build a knowledge base from its log file.

C. *Experimental results*

We measured our approach with both the ComMNZ and the RankcomMNZ method. Figure 2 shows the results for both "Precision" and "Recall" measures. The first tests represented in this figure are very encouraging. The comparison of our approach to the existing ones shows that ours is competitive knowing that our knowledge base is fed gradually so the results will undoubtedly be progressively more relevant.

Nombre of requests	Precision CombHMEAN	Precision Borda	Precision 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
100	0,8599	0,8697	0,8820

Chart 2: Precision evaluation

Nombre of requests	Recall CombHMEAN	Recall Borda	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
100	2,1599	2,1687	2,1810

Chart 3: Recall evaluation

VI. CONCLUSION AND PERSPECTIVES

The model we propose in this article is a model of supervised merger which is based on the method of adjusted Borda, Thanks to documents and search engines resulting from the user's search query, we fuel our knowledge based on his background clicks.

Diverse improvements can be suggested concerning the terms of requests. For instance, we can consider objects as much as concepts instead of words by using ontology.

Indeed, there may be several terms that have the same meaning. On the other hand, to meet the specific needs of users, we must know first their interest. In others words, for a given query, each user has its own interest therefore different needs.

REFERENCES

[1] Amini M., Usunier N., Laviolette F., Lacasse A., Gallinari P., « A Selective Sampling Strategy for Label Ranking », Proc. *ECML'06*, 2006, p. 18–29.

[2] Callan J. Distributed Information Retrieval. In W. B. Croft(Ed.), *Advances in Information Retrieval*. *Kluwer Academic Publishers*, 2000 pp. 127-150.

[3] Renda M. E., Straccia U., «Web metasearch : rank vs. score based rank aggregation methods », *SAC* 2003, *proceedings of the 18th Annual ACM Symposium on Applied Computing*, p. 841-846, 2003.

[4] 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.

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

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

[7] 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, *Proceedings of Symposium on Applications and the Internet*, pages 23–31, January 2001. (Cité page 32.).

[8] 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.

[9] Wahlster W. et Kobsa A.,Dialogue-based user models. In *Proceedings of IEEE*, Vol. 74(7), pp. 948-960, 1986.