Merging Techniques for Performing Data Fusion on the Web

Theodora Tsikrika
Department of Computer Science
Queen Mary, University of London
Mile End Road, E1 4NS
London, UK
theodora@dcs.qmul.ac.uk

Mounia Lalmas
Department of Computer Science
Queen Mary, University of London
Mile End Road, E1 4NS
London, UK
mounia@dcs.qmul.ac.uk

ABSTRACT

Data fusion on the Web refers to the merging, into a unified single list, of the ranked document lists, which are retrieved in response to a user query by more than one Web search engine. It is performed by metasearch engines and their merging algorithms utilise the information present in the ranked lists of retrieved documents provided to them by the underlying search engines, such as the rank positions of the retrieved documents and their retrieval scores. In this paper, merging techniques are introduced that take into account not only the rank positions, but also the title and the summary accompanying the retrieved documents. Furthermore, the data fusion process is viewed as being similar to the combination of belief in uncertain reasoning and is modelled using Dempster-Shafer's theory of evidence. Our evaluation experiments indicate that the above merging techniques yield improvements in the effectiveness and that their effectiveness is comparable to that of the approach that merges the ranked lists by downloading and analysing the Web documents.

Keywords

Information Retrieval, Web data fusion, Dempster-Shafer's theory of evidence

1. INTRODUCTION

The advent of the World Wide Web was accompanied by an explosion of the amount of easily accessible information. The predominant means of searching the Web is through the use of search engines, which are query-based information retrieval (IR) systems that index and retrieve Web documents. However, search engines, though more effective than browsing, present several limitations, such as the significantly limited coverage of the publicly indexable Web [17]. Moreover, search engines index different, overlapping portions of the Web [19] and adopt different IR techniques for representing documents and queries, and for determining which Web documents to retrieve in response to the query being posed to them. Therefore, distinct search engines produce different retrieval results in response to the same query and their effectiveness may vary widely. The users, though, have usually neither the knowledge to select the most appropriate search engine with regard to their information need, nor the time to pose their query to all the available search engines and then extract the most appropriate and useful results [12].

In order to provide a more effective method of retrieving relevant Web documents, the application of the *data fusion*/*collection fusion* approach was considered. The combination of retrieval results generated by using multiple document or query representations or multiple retrieval strategies is referred to as either *data fusion* when all the collaborating IR systems operate on the same document collection or as *collection fusion* when the document collections are disjoint [20, 26, 25]. In the context of the Web, the process is still referred to as *data fusion*, even though the individual search engines operate on neither the same nor disjoint document collections, but on overlapping sets of Web pages [25].

IR research has shown, in many cases, that the application of *data fusion*, both in traditional IR environments [2, 5, 10, 26, 27] and on the Web, performed by metasearch engines [6, 11, 12, 21, 22, 19], yields improvements in the effectiveness over that of a single representation scheme or a single retrieval strategy.

However, current Web metasearch engines retain some of the limitations of their underlying search engines, such as their reduced precision [16]. The main sources of their deficiencies are the following: First of all, *data fusion* on the Web consists of the combination of the retrieval results produced by independent retrieval strategies that rely on possibly different weighting schemes, similarity measures and retrieval models. Moreover, they operate on multiple, overlapping, heterogeneous document collections that differ in size, cover diverse topics and have different retrieval characteristics. Furthermore, the fusion strategy applies a merging function to the answer lists associated with each of the participating search engines, without actually having access to their internal workings. It relies, therefore, only on the limited amount of information provided by the search engines and

accompanying the retrieved documents, such as their ranking in the list and their document scores.

This paper investigates merging techniques, which aim at improving the effectiveness of the metasearch engines by processing more of the information provided to them by the participating search engines. The proposed merging techniques utilise not only the *rank positions* of the retrieved documents, but also their *title* and the *summary* accompanying them, describing their content. Furthermore, the data fusion process is viewed as being similar to the combination of belief in uncertain reasoning and is modelled using *Dempster-Shafer's theory of evidence*. The lists of retrieved documents correspond to *bodies of evidence*, which are aggregated (merged) using *Dempster's combination rule*. Finally, it is investigated whether the effectiveness of the proposed merging strategies, which are based entirely on the information provided by the underlying search engines, is comparable to that of the approach that merges the ranked lists by downloading and analysing the retrieved Web documents.

This paper is organised as follows: *Section 2* presents a literature review of the application of data fusion both in traditional IR environments and on the Web. *Section 3* introduces our merging methods, whereas *Section 4* describes the system we implemented in order to evaluate them. *Section 5* presents the evaluation experiments and *Section 6* their results. Finally, in *Section 7*, our conclusions and suggestions for further work are discussed

2. RELATED WORK

In traditional IR environments, several *data fusion* approaches have been proposed. Turtle and Croft [24] introduced a model, which can combine different document and query representations using a Bayesian inference network as its underlying framework. This model was implemented by the *INQUERY* retrieval system [4] and demonstrated that this combination results in improvements in the effectiveness. Similar results were obtained when various merging methods for combining distinct query formulations or results generated by multiple retrieval strategies were introduced [2].

In IR research, the data/collection fusion process is viewed as being divided into 3 phases: the *collection selection*, the *document selection* and the actual *merging*. The *collection selection* phase corresponds to the identification of the document collections most likely to contain relevant documents to the submitted user query. The basis of this approach relies on the heuristic that merging results from the "*best of the best*" will possibly produce better results than merging the retrieval lists provided by all possible sources [12]. For instance, a ranking of the available document collections can be produced by using inference networks [5] or training queries [26] and then select the top-ranked ones.

The *document selection* phase corresponds to determining the number of documents to be retrieved from the selected document collections. The simplest case would be to select equal number of documents from each individual list. This strategy is based on the assumption that there is the same number of documents in each collection [20] and that there is an identical distribution of relevant and non-relevant documents [26]. Alternatively, the number of documents to be retrieved could be determined to be proportional to the quality of each document collection [5, 26].

The *merging* (fusion) phase corresponds to the actual combination of the individual ranked lists produced by the multiple retrieval strategies. The ranking of the documents in the final unified list can be determined by using, for instance, a probabilistic approach that considers the original rank positions of the documents [26, 27] or by considering the document scores computed by the participating retrieval strategies [20] and the quality of the document collection this document belongs to [5].

In the context of the Web, the *collection selection* phase determines which search engines should be chosen as the underlying implementation layer. *ProFusion* metasearch engine analyses the query being posed to the system, identifies its topic(s) and issues the request to the 3 search engines which have shown to perform better for this topic, in response to training queries [11, 12]. *SavvySearch* intelligently selects the most promising and accessible search engines using information from past searches and estimated network traffic [6]. These approaches, however, increase query time and another method, adopted by several metasearch engines [21, 19], is to treat equally all search engines.

In the *document selection* phase, Web metasearchers follow the simple approach of selecting the same number of documents from the selected search engines, ranging between 10 [21, 19] and 20 [11, 12] documents.

The *merging* algorithms of metasearchers, on the other hand, vary and their differences lie mostly on the elements of information they use in order to compute the final ranking. However, search engines only provide a limited amount of information such as the document rankings and scores, without exporting internal information, such as statistics about the documents they index and *tf* and *itf* values of the query terms. Although protocols [8, 9, 14] that determine what information should be made available to the metasearch engines (so that they can experiment with a variety of approaches) have been proposed, they have not been implemented. Therefore, *Metacrawler* [22], for instance, utilises the *document retrieval scores* and *Fusion* [19] the *rank positions*, so that duplicate documents have their ranks summed up, and documents are penalised if they are not retrieved by a particular search engine. Finally, *ProFusion* [11, 12] uses a weighted score merging algorithm, similarly to [5], where the final ranking is determined using both the initial retrieval score assigned by the search engine and the score expressing the quality of that search engine.

This paper focuses on the merging algorithms of the data fusion process on the Web and our aim is to utilise more of the information provided by the search engines, than just the *rank positions* and *documents scores* of the retrieved documents: the *title* and *summary*, which is generated by the search engine, accompanying each retrieved document. These pieces of information are not considered by the existing merging algorithms and our objective is to investigate whether their incorporation in the fusion process as surrogates of the documents' content can enhance retrieval effectiveness. Therefore, the *title* and the *summary* are used to index the retrieved documents by using traditional IR techniques, and by introducing a formal approach based on *Dempster-Shafer's theory of evidence*. This latter approach models the individual lists of retrieved documents as *bodies of evidence* associated with *term spaces* generated using the terms contained in the *titles* and *summaries* of the retrieved

documents. These *bodies of evidence* are aggregated (merged) into a single ranked list. Finally, we investigate whether our merging techniques, which use only the information provided to them by the search engines, can be as effective as the approach of downloading and indexing the full text of the retrieved documents.

3. MERGING TECHNIQUES

This section presents the merging strategies that aim at improving the effectiveness of data fusion operation on the Web by incorporating in their computations more of the information returned by the participating search engines in their ranked lists of retrieved documents. These merging techniques will take into account the *rank positions* of the retrieved documents, their *title* and the brief *summary* accompanying them. Although some merging functions use the *document scores* in their computations, these scores will not be taken into account by our merging techniques for the following reasons. First, d*ocument scores* are generated by the participating search engines, and they are usually computed using different IR models and therefore cannot be directly compared. Second, even if two search engines adopt the same IR model, the *document scores* are still considered as being incomparable, since, in their computation, there is an incorporation of statistics (*tf, idf*) dependent on the Web document collection indexed by each search engine.

The following section introduces our merging techniques. **Method 1** considers only the rank positions of the documents; **Method 2** their title and summary; **Method 3** their title and summary and models the data fusion process using *Dempster-Shafer's theory of evidence*; **Method 4** takes into account the *rank positions*, the *title* and the *summary* of the retrieved documents and finally **Method 5** produces the merged ranked lists by downloading and indexing the Web documents. These merging methods are described in detail next.

3.1 Method 1: Merging Using Rank Positions

The simplest method that can be employed in the process of merging the individual lists of retrieved Web documents is the one that takes into account only the *rank positions* of the documents. This method implicitly incorporates information about the content of the document, since the *rank positions* themselves are determined by the search engines that retrieve the document, and the ordering is decided based on the index terms of the document.

In this method, the duplicate documents have their ranks summed up and the rest of the documents are interleaved. Therefore, **Method 1** favours the documents retrieved by more than one search engine. This method is simplistic, since it relies on the minimum amount of information provided, and it will serve as a *baseline* to our experimental approach.

3.2 Method 2: Merging Using the Title and Summary of the Retrieved Documents

This method considers the title and the summary of the retrieved documents and attempts to investigate whether the effectiveness can be improved by indexing the retrieved documents using them. The title of the Web document as it appears in the list of retrieved documents provided by each of the participating search engines is the title of the actual Web page that this document corresponds to. The summary is generated by the search engine that retrieves the document, and it usually consists of extracts of the document that contain the terms of the query in response to which this document was retrieved. Therefore, these textual elements incorporate information about the document's content and can be used as its representation, since the full text of the retrieved documents is not directly indexed by the metasearch engine.

For this method, a single set of retrieval documents is formed, containing all the Web documents in the lists provided by the participating search engines. The documents in this set can be then indexed using the terms in their *titles* and *summaries* and be represented as vectors of these index terms. Weights are assigned to these terms in order to determine which are good discriminators of this Web document's content. The commonly used weighting scheme of *tf x idf* [1] cannot be applied here, because the terms that we consider as good discriminators of a Web document are the query terms and they are more likely to be present in most of the documents, since they were retrieved in response to this query. Consequently, *idf(t)*, when *t = query term*, is most likely to be equal to *0*.

A different weighting approach for each term *t* in document *d* is used instead. The documents are represented as $d = \{w_{1,d}, \ldots, w_{k,d}\}$, where $w_{i,d}$ is the weight of the *ith* index term *t* in document *d*, computed using its *term frequency tf(t, d)* only. The documents are then reranked using the similarity of the document representations and the query *Q*. The similarity function used is $\sum_{t \in Q} w(t,d)$.

The duplicate documents have the same URL, but different summaries since these summaries are generated by the search engines that retrieved them. For each document, the summaries of its duplicates are kept and concatenated together. In that way, the duplicate documents are associated with the longest summaries among the retrieved documents and the reranking mechanism favours them, since no normalisation is applied in the ranking function. This reflects the heuristic that documents retrieved by more than one search engine should be ranked higher.

3.3 Method 3: Merging Using Dempster – Shafer's Theory of Evidence

A formal approach can be introduced by defining a merging algorithm, which considers the *title* and the *summary* of the retrieved documents and is based on *Dempster–Shafer's theory of evidence*. This theory allows the explicit representation of imprecision, ignorance and the combination of evidence and will be used as a means to incorporate the uncertainty when merging the individual lists of retrieved documents produced by the participating search engines. This will result to a model of the fusion process, which is viewed as being analogous to the aggregation of belief in uncertain reasoning.

The basic concepts of Dempster-Shafer (D-S) Theory of Evidence will be briefly presented, based on the description given in [23].

3.2.1 Basic Concepts of Dempster-Shafer Theory of Evidence

Dempster-Shafer theory of evidence is an extension of the probability theory and it allows the explicit representation of uncertainty and the combination of evidence. This property makes the use of Dempster-Shafer theory particularly attractive in modelling the IR fusion process. The combination of evidence is captured as a fundamental property by the *Dempster's combination rule*, which allows the expression of aggregation. Aggregation is the basic concept underlying the fusion (merging) process and consequently this theory allows the modelling of both the representation of the retrieved documents and of the merging strategy itself.

According to the Dempster-Shafer framework, the set representing the domain of all possible values that propositions can take, is called the *frame of discernment*. Propositions are then represented as subsets of this set. For instance, if the frame of discernment is *U*, an example of a proposition is 'the value of *u* is in *A*' for some $A \subseteq U$. The proposition $A = \{a\}$ for $a \in U$ constitutes a basic proposition 'the value of *u* is in *a*'.

Beliefs can be assigned to propositions to express their uncertainty. The beliefs are usually computed based on a density function $m : \wp(U) \rightarrow [0,1]$, called *basic probability assignment* (bpa), such that:

$$m(0) = 0 \text{ and } \sum_{A \subseteq U} m(A) = 1$$

m(A) is the belief committed exactly to *A*, that is the exact evidence that the value of *u* is in *A*. If there is positive evidence for the value of *u* being in A then $m(A) > 0$ and A is called a *focal element*. The focal elements and their associated bpas define a *body of evidence*.

Given a body of evidence with bpa *m*, one can compute the total belief provided by the body of evidence for a proposition. This is achieved with a *belief function* $Bel : \wp(U) \rightarrow [0,1]$ defined upon *m* as follows:

$$Bel(A) = \sum_{B \subseteq A} m(B) .$$

Bel(A) is the total belief committed to *A*, that is the total positive effect the body of evidence has on the value of *u* being in *A*. The higher the value, the higher the total belief is.

When two *independent* bodies of evidence defined within the same frame of discernment exist, *Dempster's combination rule* can be used to combine them into one body of evidence, under the conditions that the *bodies of evidence* are independent of each other. Let m_1 and m_2 be the bpas associated to the two independent *bodies of evidence* defined in a frame of discernment *U*. The new body of evidence is defined by a bpa *m* on the same frame *U* as follows:

$$m(A) = m_1 \oplus m_2(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{\sum_{B \cap C \neq \emptyset} m_1(B) m_2(C)}$$

Dempster's combination rule computes a measure of agreement between two *bodies of evidence* concerning various propositions discerned from a common frame of discernment. The rule focuses on those propositions that the bodies support. The new bpa takes into account the bpa associated to the propositions in both bodies that yield the propositions of the combined body. The denominator of the above equation is a normalisation factor that ensures that *m* is a bpa.

3.2.2 Modelling Data Fusion on the World Wide Web Using Dempster-Shafer Theory of Evidence

To model the data fusion process on the Web using Dempster-Shafer theory of evidence, the lists of retrieved documents need to be defined and represented and then fused into a single ranked list.

Representation of the lists of the retrieved Web documents Each list of retrieved documents generated by each participating search engine is represented as *body of evidence* defined in a frame of discernment *T*. The frame of discernment *T* is defined as the set of the indexing terms used to index the Web documents in the lists. Each indexing term $t \in T$ corresponds to the basic proposition that 'the term *t* belongs to set of index terms that index the documents contained in the list'. In our approach only basic propositions, corresponding to single index terms, are considered, as generally in standard text-based IR where only single terms are used to index a document [1].

The indexing procedure will produce a set of index terms for each list of retrieved Web documents, by considering the *titles* and *summaries* of the documents. Once the index terms are extracted, the indexing procedure will assign a weight to each of them, representing its 'goodness' at discriminating document content. The 'goodness' of an index term is measured by its distribution in the documents forming the *body of evidence*. This *body of evidence* can be viewed as a *term space* associated with each individual list of retrieved documents. The distribution of a term is expressed by the term *document frequency*, $df(t) = \log n(t)$, where the document collection consists of all the retrieved documents in this particular list.

Therefore, this weight can be considered as evidence of how good an indexing term is, so the bpa can be defined as $m(t) = \log n(t)$ for basic propositions. From the definition of the bpa, each body of evidence must assign the same total amount of belief to the frame of discernment, i.e. to the entire set of terms in all the term spaces associated with all the lists of retrieved documents. The remaining belief is treated as *uncommitted belief* and is assigned to the frame of discernment *T*. Uncommitted beliefs express the uncertainty associated to the effectiveness of each search engine that provides the list of retrieved documents. The value of the uncommitted belief *m(T)* is defined as follows: $\frac{\sum \log n(t)}{k} + m(T) = 1$, where *k* is a normalising factor ensuring that *m* is a bpa. The value of *k* is estimated through a test-and-try approach and for *k* = 0.95 we obtained the best experimental results, which are presented in Section 6.

Fusion Dempster's *combination rule* aggregates two bodies of evidence into one, reflecting the propositions that both bodies support and computes the uncertainty that reflects this aggregation. In our case the bodies of evidence that are aggregated are the individual lists of retrieved documents merged into a single list.

The bpa assigned to the merged list is calculated by combining the bpas of the individual lists, using the Dempster's *combination rule*. Suppose that the ranked list of documents *l* is defined as the aggregation of two individual lists of ranked documents, generated by Web search engines, l_1 and l_2 with respective bpas m_1 and m_2. All terms that belong to the term space of lists l_1 or l_2, also belong to the combined term space of list *l*. The bpa *m* associated to *l* is:

$$m(t) = \frac{m'(t)}{K} \quad \text{and} \quad m(T) = \frac{m'(T)}{K}$$

where:

$$m'(t) = \begin{cases} m_1(t)m_2(t) + m_1(t)m_2(T) + m_1(T)m_2(t) & \text{if } t \in l_1 \text{ and } t \in l_2 \\ m_1(t)m_2(T) & \text{if } t \in l_1 \text{ and } t \notin l_2 \\ m_1(T)m_2(t) & \text{if } t \notin l_1 \text{ and } t \in l_2 \\ 0 & \text{otherwise} \end{cases}$$

and $m'(T) = m_1(T)m_2(T)$

where *K* is defined: $K = m'(T) + \sum_{t \in l_1 \text{ or } t \in l_2} m'(t)$

Once the combined term space constructed from all the individual term spaces is formed, the single merged ranked list containing all the initial retrieved documents has to be produced. Therefore, the relevance of each Web document to the query being posed to the system has to be calculated. The relevance of the Web document is estimated using the belief function: $Bel(q) = \sum_{t \in q} m(t)$.

Bel(q) is used to rank the Web documents according to their estimated relevance to the query. *Bel(q)* expresses relevance because it is based on all query terms that are supported by the document. It also takes into account the beliefs associated to their use; the higher the beliefs, the higher the relevance. The quantity *Bel(q)* indicates that the document contains information that concerns the query *q*. The higher *Bel(q)*, the more information is contained in the document. Therefore, when each Web document is assigned a value, an ordering is determined among them and a final ranking is produced.

3.3 Method 4: Merging Using Rank Positions, the Title and the Summary of the Retrieved Documents

So far, **Method 1** utilises the *rank positions* of the documents, whereas **Method 2** and **Method 3** utilise the information contained in the *title* and *summary* of the retrieved documents. The ranked lists generated by these methods can be further fused together, if they are provided as inputs to **Method 1**. Therefore, by merging the lists generated by **Method 1** and **Method 2**, the fusion operation takes into account more of the available information. The same is true in the case of the lists of **Method 1** and **Method 3**. By introducing this method, we aim at investigating whether the combination of more information returned by the search engines (*rank positions*, *titles*, *summaries*) leads to improvements in the effectiveness.

3.4 Method 5: Merging by Downloading the Web Documents

All the merging methods discussed in the previous sections rely on the information provided by the contributing search engines that retrieve the Web documents, without the methods accessing the full text of these documents. However, if the actual Web pages are downloaded and analysed, it is believed that the ranking of the final merged list can be improved. This approach is based on the fact that since the whole of the document's content is available, the merging function can take this information into account and generate a more effective merged ranked list. Furthermore, the downloading of the pages leads to identification of the pages that no longer exist [16].

The drawback of this approach is that it requires the real-time downloading and analysis of the Web documents. Consequently it requires additional bandwidth, places higher demands on computer performance and increases query time. However, *Inquirus* metasearch engine [15, 16] demonstrated that the real-time analysis of documents returned from Web search engines is feasible and therefore this approach can be applied.

This merging algorithm is similar to **Method 2** with the difference that the full-text of the document, instead of its summary, is used to represent its content**.** However, when a duplicate document is detected only one copy is kept. The ranking algorithms described in **Method 2** are applied, instead of a more sophisticated IR ranking function, so that the results are comparable to those of the other methods. **Method 5** will serve as a baseline and the comparison of its effectiveness to that of the other merging methods will lead to conclusions about whether considering only the information provided by the search engines is sufficient, without having to access the Web documents themselves.

4. IMPLEMENTATION

The above five merging techniques were implemented as part of a system, which performs data fusion by merging the lists of documents, which are stored locally, after being initially retrieved by search engines in response to a user query. These lists are then parsed so that each Web document and its associated *rank position*, *title*, *summary* and *URL* are extracted correctly.

In order to detect the duplicate documents, the following rules are used. Each document's URL distinguishes it from the rest of the documents and when two documents have the same URLs, they are duplicates. However, our system handles the case where an identical Web page is referenced by slight variations of the same address; for example, *http://www.dcs.qmw.ac.uk/* refers to the same page as *http://www.dcs.qmw.ac.uk/index.html*. This duplicate detection method is followed by the publicly available metasearch engines [11, 12, 21, 22].

To index the documents for **Method 2** and **Method 5** and to construct the frame of discernment of **Method 3**, conventional indexing techniques, such as *stop word removal*, using the list provided in [7], and *stemming* [18] are applied. Frequency statistics, such as *term frequency* and *document frequency* are computed and the merging techniques are applied, each providing a list of documents.

5. EXPERIMENTS

The purpose of our experiments is to compare the effectiveness of the different merging techniques, which vary in the amount of the available information they consider and in the way they process it.

Several decisions are made affecting the design of our experiments. First of all, 4 search engines were used, following the heuristic described in [25]. These are: *Google* (http://www.google.com), chosen because of its ranking mechanism which takes into account the Web link structure [3], *InfoSeek* (http://www.infoseek.com), *Northern Light* (http://www.northernlight.com) and *Webcrawler* (http://www.webcrawler.com), selected because they are part of publicly available metasearch engines [11, 12, 21, 22, 19]. The number of documents retrieved from each search engine was set to 30, since the more documents retrieved, the more likely an increase on the number of duplicate documents is.

There were 10 sample queries used in our experiments; some of them are taken directly from [19], some are our localised equivalents and some are real queries to Web search engines executed by the users, who participated in the experiments and provided the relevance assessments. Finally, the notion of *precision* (proportion of retrieved documents which are relevant) at each *rank position* [13] is used in the evaluation of the effectiveness. For instance, if there are 4 relevant retrieved documents in the first 10, then precision at rank position 10 equals to 0.4.

6. RESULTS

The results of the experiments, which were carried out in order to evaluate the effectiveness of the merging methods, are presented in this section.

First of all, it was observed that the number of duplicate documents within the lists of retrieved documents of all the participating search engines is relatively small (4.45%). This leads to the conclusion that there is very little overlapping among the Web pages that distinct search engines index, as it has already been suggested by similar studies [19]. This further implies that the data fusion approach on the Web is likely to increase the breadth of the retrieval process, because the number of documents returned in the merged list for a single query is much greater than the number returned by the individual search engines.

One of our merging functions, **Method 5**, requires the downloading of the Web pages, so that their full text is taken into account when determining the ranking of the documents. This approach leads to identification of the pages that no longer exist, have been moved to another location or are unreachable and further reveals whether the indices of the search engines are updated or not. Our experiments demonstrated that the participating search engines contain valid documents in their indices in an average percentage of 93.11%.

The average precision over all the queries for each merging method is presented in Table 1 and **Method 1** and **Method 5** act as our baselines.

These results suggest that the merging strategies **Method 2** and **Method 3** which take into account the *title* and the *summary* of the retrieved Web documents are more effective than **Method 1** which uses only their *rank positions*. The average precision of **Method 2** and **Method 3** displays an increase of 11.23% and 9.57% respectively compared to that of **Method 1**. This improvement in the effectiveness of the fusion algorithms is shown to be significant using a paired t-test (significance level was set at 0.05). Furthermore, this is a first indication that the information provided by the participating search engines is sufficient in order to develop more sophisticated techniques that can yield improvements in the effectiveness.

Although, both **Method 2** and **Method 3** are merging algorithms that take into account the *title* and the *summary* of the retrieved Web documents, they use different approaches in determining the final merged ranked list. One of the fundamental differences between these two methods lies in the way in which they assign weights in their index terms in order to determine how good discriminators of the document's content, these terms are. **Method 2** uses the *term frequency* of the index terms, whereas **Method 3** uses their *document frequency* in order to determine the value of the *basic probability assignment* (*bpa*), associated with each term. Therefore, the two methods are not strictly comparable and although **Method 2** performs better than **Method 3**, we cannot conclude, at this stage, which one is actually better.

Method 3 introduces a formal approach in the data fusion problem on the Web by applying *Dempster-Shafer's theory of evidence* in modelling the merging operation. The *uncommitted belief* that reflects the uncertainty associated to the effectiveness of a particular search engine is one of the fundamental properties of this theory. Our results indicate that the value of the *uncommitted belief* corresponds well with the effectiveness of the participating search engines. For instance, our experiments demonstrated that *Google* is the more effective of our search engines and that at the same time, it has the highest value of *uncommitted belief*. Therefore, if we rank the search engines according to their effectiveness, we observe that the same order applies to the values of *uncommitted belief* associated with them.

Table 1 Average precision over all queries for each method

	Method 1	**Method 2**	**Method 3**	**Method 4 (1&2)**	**Method 4 (1&3)**	**Method 5**
Average	56.18%	62.49%	61.56%	63.04%	62.19%	62.08%
%Change	0.00%	**+11.23%**	**+9.57%**	**+12.21%**	**+10.70%**	**+10.50%**
%Change	-9.5%	**+0.66%**	-0.84%	**+1.55%**	**+0.18%**	0.00%

Consequently, these values could be used to generate a ranking among the participating search engines that will reflect their 'quality' in terms of the number of relevant documents contained in their lists of retrieved documents. Therefore, the value assigned to each search engine and the ranking among them could be further exploited by the merging algorithm, as done in [5, 11, 12].

Our results further indicate that the merging technique of taking into account all of the information provided by the participating search engines – *rank positions*, *title* and *summary* of the retrieved Web documents – is more effective than the ones that take into account only the *rank positions* or only the *title* and *summary* of the retrieved Web documents. To be more specific, **Method 4(1&2)** that combines the results of **Method 1** and **Method 2**, is more effective than its components (i.e. **Method 1** and **Method 2**) and the same applies in the case of **Method 4(1&3)**.

Finally, the results of our experiments indicate that we can achieve a comparable performance on the effectiveness of the fusion operation, if we simply rely on the faster approach of taking into account the information provided by the participating search engines, without having to resort to the more sophisticated, more computationally demanding, approach of actually downloading and analysing the retrieved Web documents. As a matter of fact, these results suggest that **Method 1** is the only one that performs significantly worse than **Method 5**, whereas the effectiveness of all the other methods is comparable to that of **Method 5**.

7. CONCLUSIONS & FURTHER WORK

This paper seeks to investigate whether the application of the merging strategies proposed here, can yield improvements in the effectiveness of the fusion operation. The results of the experiments, which were carried out in order to evaluate the effectiveness of the proposed merging functions indicate that, first of all, the merging strategies that take into account the *title* and the *summary* of the retrieved Web documents are more effective than the one that uses only the *rank positions*. Furthermore, the merging algorithms that combine all the information provided (*rank positions*, *title*, *summary*) perform even better. Moreover, we can achieve similar performance on the effectiveness of the fusion operation by taking into account only the information provided by the participating search engines, without having to resort to the more sophisticated, yet slower, approach of downloading and analysing the documents. Finally, the use of the Dempster-Shafer theory indicates that the formal modelling of data fusion on the Web can lead to improvements in the effectiveness of the merging functions.

Further research is required on the formal modelling of the data fusion process on the Web, using the Dempster-Shafer theory of evidence. First of all, the *basic probability assignment* (*bpa*), associated with each of the index terms was modelled here using the *document frequency* of the terms. Further experiments are required using alternative approaches, which will define the bpa using other statistics, such as the *term frequency* and the *inverse document frequency* of the terms. Also, further experiments may lead to better estimation of the value of the uncommitted belief and this may result in further improvements in the effectiveness of the merging operation.

Finally, it should be noted here that only a small number of queries was used for the evaluation of the effectiveness of our merging strategies. Therefore, evaluation like the one reported here can only be taken as indicator of the effectiveness and consequently, future research in this area should perform larger scale evaluation experiments.

8. REFERENCES

[1] Baeza-Yates, R. & Ribeiro-Neto, B. *Modern Information Retrieval.* Addison & Wesley, 1999.

[2] Belkin, N. J., Kantor, P., Fox, E. A. & Shaw, J. A. Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management,* 31(3), pp. 431-448, 1995.

[3] Brin, S. & Page, L. The Anatomy of a Large-Scale HyperTextual Web Search Engine. *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, 1998.

[4] Callan, J.P., Croft, W.B., & Harding, S.M. The INQUERY Retrieval System. In the *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, Valencia, Spain, 1992, pp. 78-83.

[5] Callan, J. P., Lu, Z. & Croft, W.B. Searching Distributed Collections with Inference Networks. In the *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995.

[6] Dreilinger, D. & Howe, A. *Experiences with Selecting Search Engines Using MetasearchI.* ACM TOIS, 15(3), July 1997, pp. 195-222.

[7] Frakes, W. B. & Baeza-Yates, R. *Information Retrieval: Data Structures & Algorithms.* Prentice Hall, Englewood Cliffs, NJ, USA, 1992.

[8] Gravano, L., Chang, K., Garcia-Molina, H., Lagoze, C. & Paepcke, A. Digital Library Project, Stanford University. STARTS – Stanford Protocol Proposal for Internet Retrieval and Search. http://www-db.stanford.edu/~gravano/starts.html

[9] Gravano, L., Chang, K., Garcia-Molina, H. & Paepcke, A. STARTS – Stanford Protocol Proposal for Internet Meta-Searching. In the *Proceedings ACM SIGMOD International Conference on Management of Data*, May 13-15, 1997, Tucson, Arizona, USA.

[10] Gravano, L. & Garcia-Molina, H. Merging Ranks from Heterogeneous Internet Sources. In the *Proceedings of the 23rd VLDB Conference*, Athens, Greece, 1997.

[11] Gauch, S. & Wang, H. Information Fusion with ProFusion. In the *Proceedings of the WebNet96: The First Conference on the Web Society*, San Francisco, CA, USA, October 1996.

[12] Gauch, S., Wang, H. & Gomez, M. ProFusion: Intelligent Fusion from Multiple, Distributed Search Engines. *Journal of Universal Computing*, Springer-Verlag, Volume 2 (9), September 1996.

[13] Hawking, D., Craswell, N. & Harman, D. Results and Challenges in Web Search Evaluation. In the *Proceedings of the Eigth International World Wide Web Conference*, Toronto, Canada, 1999.

[14] Kirsch, S. T. Document retrieval over networks wherein ranking and relevance scores are computed at the client for multiple database documents, United States Patent #5,659,732, 1997.

[15] Lawrence, S. & Lee Giles, C. *NEC Research Institute*. Inquirus – The NECI Metasearch Engine. http://www.neci.nj.nec.com/~lawrence/inquirus.html.

[16] Lawrence, S. & Lee Giles, C. Inquirus – The NECI Metasearch Engine. In the *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, Elsevier Sience, pp. 95-105, 1998.

[17] Lawrence, S. & Lee Giles, C. *NEC Research Institute*. Searching the World Wide Web. *Science*, Volume 280, Number 5360, pp.98-100, 1998.

[18] Porter, M.F. An algorithm for suffix stripping. In K. Sparck Jones and P. Willet, editors, *Readings in Information Retrieval*, pages 313-316. Morgan Kaufmann Publishers Inc., 1997.

[19] Smeaton, A. F. & Crimmins, F. Using a Data Fusion Agent for Searching the WWW". Poster presented at the *Sixth International World Wide Web Conference*, Stanford, USA, April 1997.

[20] Savoy, J., Le Calvé, A. & Vrajitoru, D. Report on the TREC-5 Experiment: Data Fusion and Collection Fusion. *Proceedings TREC5*, 1996.NIST Publication 500-238, Gaithersburg (MD), 489-502.

[21] Selberg, E. & Etzioni, O. Multi-Service Search and Comparison using the MetaCrawler. In the *Proceedings of the 4th International World Wide Web Conference*, December 1995.

[22] Selberg, E. & Etzioni, O. The MetaCrawler Architecture for Resource Aggregation on the Web. *IEEE Expert*, January / February 1997, Volume 12 No. 1, pp. 8-14.

[23] Shafer, G. *A mathematical theory of evidence*, Princeton University Press, 1976.

[24] Turtle, H. & Croft, W.B. Evaluation of an Inference Network-Based Retrieval Model. *ACM Transactions on Information Systems*, 9(3), pp. 187-222.

[25] Vogt, C. C. How much more is better? Characterising the effects of adding more IR systems to the combination. In the *Proceedings of the Computer Assisted Information Retrieval International Conference (RIAO)*, Paris 2000.

[26] Voorhees, E. M., Gupta, N. K. & Johnson-Laird, B. The collection fusion problem. In the *Proceedings of the Third Text Retrieval (TREC-3) Conference*, pp. 95-104, 1994.

[27] Yager, R. R. & Rybalov, A. On the Fusion of Documents from Multiple Collection Information Retrieval Systems. *Journal of the American Society for Information Science.* 49(13), pp.1177-1184, 1998.