Collection Fusion for Distributed Image Retrieval

S. Berretti, A. Del Bimbo, P. Pala*

Dipartimento Sistemi e Informatica
Università di Firenze
via S.Marta 3, 50139 Firenze, Italy
{berretti,delbimbo,pala}@dsi.unifi.it

Abstract. Searching information through the Internet often requires users to contact several digital libraries, author a query representing the information of interest and manually gather retrieved results. However, a user may be not aware of the content of each individual library in terms of quantity, quality, information type, provenance and likely relevance, thus making effective retrieval quite difficult.
Searching distributed information in a network of libraries can be simplified by using a centralized server that acts as a gateway between the user and distributed repositories. To efficiently accomplish this task, the centralized server should perform some major operations, such as *resource selection*, *query transformation* and *data fusion*. Resource selection is required to forward the user query only to the repositories that are candidate to contain relevant documents. Query transformation is necessary in order to translate the query into one or more formats such that each library can process the query. Finally, data fusion is used to gather all retrieved documents and conveniently arrange them for presentation to the user.
In this paper, we introduce an original framework for collection fusion in the context of image databases. In fact, the continuous nature of content descriptors used to describe image content, makes impractical the applicability of methods developed for text. The proposed approach splits the score normalization process into a learning phase, taking place off-line, and a normalization phase that rearranges scores of retrieved images at query time, using information collected during the learning. Fusion examples and results on the accuracy of the solution are reported.

1 Introduction

Nowadays, many different document repositories are accessible through the Internet. A generic user looking for a particular information should contact each repository and verify the presence of the information of interest. This typically takes place by contacting the repository server through a client web browsing application. Then, the user is presented a web page that guides her/him in the process of retrieval of information (in a multimedia scenario this can be in the form of text documents, images, videos, audio and so on).

This framework has two major drawbacks. First of all, it is assumed that the user is aware of the presence of all potentially relevant repositories, in terms of quantity, quality, information type, provenance and likely relevance. Internet search engines such as *Altavista*, *Google* and *Lycos*, to say a few, can provide only limited support to accomplish this task. A second drawback is related to the fact that, in order to find the information of interest, the user typically contacts several repositories. Each one is queried and even for a limited

* This work is supported by EU Commission under grant IST-2000-26061. Detailed information about this project is available at http://www.mind-project.org

number of repositories, the user is soon overwhelmed by a huge and unmanageable amount of (probably irrelevant) retrieved documents.

Federated digital libraries have been recently proposed as a solution to these problems. A possible architecture for a federation of digital libraries is characterized by the presence of a central server acting as a gateway between the user and all federated resources. In particular, a generic user looking for some information sends the query to the server. This one is in charge to forward the query to all (or a subset of all) the networked resources. All retrieved documents (that is the set of all documents returned by each resource) are gathered by the central server and conveniently arranged for presentation to the user.

Implementation of this solution requires the availability of some modules that take care of summarizing the content of a resource (*resource description*), use resource summaries to identify which resources contain documents relevant to a query (*resource selection*), translate the query in the specific format accepted by each resource (*query transformation*), issue the query to the resources identified as relevant, merge retrieval results in decreasing order of relevance so that they can be presented to the user (*collection fusion*).

Resource description, resource selection and collection fusion for distributed libraries have been widely investigated in the past focussing on libraries of text documents [2], [3], [4], [5], [8], [9], [10].

However, solutions developed so far for libraries of text documents exploit the peculiar characteristics of content descriptions of textual materials. Typically, these descriptions are in the form of term and document frequencies. The former represent, for each vocabulary term, the percentage of documents that contain at least one occurrence of the term. The latter represent the occurrence of a term within one document. Since the vocabulary is composed of a finite number of elements, variables describing the content of textual materials take values over discrete sets.

Differently, content descriptors of multimedia documents—such as audio files, images, videos, Flash presentations—are in the form of feature vectors taking dense values over continuous spaces. This complicates (if not makes it impossible) the applicability of solutions developed for text to the domain of multimedia libraries.

In this paper, we present a novel method to accomplish collection fusion for distributed libraries of images. Although the proposed method is presented in the context of image libraries, its domain of applicability includes generic multimedia libraries. Indeed, the proposed solution can be applied in all those contexts where document content is in the form of feature vectors, representing instance values of metric attributes.

In general, the fusion process is intended as a mean to overcome difficulties related to the assessment of the relative relevance of retrieved items. These difficulties are particularly challenging if retrieved documents are not associated with matching scores—quite a common case for text retrieval engines but not as common for image retrieval engines. However, the availability of matching scores for all retrieved documents doesn't solve the problem completely. In fact, matching scores are not comparable across different collections since different collections may use retrieval engines with different similarity metrics and/or different content descriptors.

The proposed solution relies on a model based approach where, for each library, a normalization model is defined to map scores assigned by the library into normalized scores. Once

documents returned by all libraries have been associated with their normalized scores, the relevance of each document with respect to each other can be assessed. The normalization model is identified by a set of parameters (the model parameters) that determine the function mapping original scores into their normalized counterparts. Use of the normalization model is articulated in two distinct phases: *learning* and *normalization*. During learning, each library is processed through a set of *sampling queries* so as to learn values of model parameters for that particular library. Once values of model parameters have been learned for all federated libraries, normalization can be accomplished. Normalization takes place during a query session and allows, given the user query and results returned by all libraries, original scores assigned to retrieved documents to be transformed into normalized scores.

This paper is organized as follows. In Sect.2, we briefly review previous work on data fusion. Sect.3 addresses the proposed solution, introducing the learning and normalization phases for data fusion. Experimental results are reported in Sect.4. Finally, Sect.5 gives conclusions and future directions of work.

2 Previous Work

Several merging strategies have been proposed to deal with merging results returned by distributed libraries (mostly of text documents). In general, a common distinction used by researchers in this area is the difference between data fusion and collection fusion [11]. The former takes place in a setting where all of the retrieval systems involved have access to the same text collection. The latter is used when the collections searched by all retrieval systems are disjoint.

One of the most known approaches is called *Round-Robin* [10]. In this approach, it is assumed that each collection contains approximately the same number of relevant items that are equally distributed on the top of the result lists provided by each collection. Therefore, results merging can be accomplished by picking up items from the top of the result lists in a round-robin fashion (the first item from the first list, the first from the second list, ..., the second from the first list and so on).

Unfortunately, the assumption of uniform distribution of relevant retrieved items is rarely observed in practice, especially if libraries are generic collections available on the Web. This drastically reduces the effectiveness of the round-robin approach.

A different solution develops on the hypotheses that *i*) for each retrieved document, its matching score is available and *ii*) the same search model is used to retrieve items from different collections. In this case, document matching scores are comparable across different collections and they can be used to drive the fusion strategy. This approach is known as *Raw Score Merging* [6], but it is rarely used due to the large number of different search models that are applied to index documents even in text libraries.

To overcome limitations of approaches based on raw score merging or Round Robin, more effective solutions have been proposed developing on the idea of score normalization. These solutions assume that each library returns a list of documents with matching scores. In this case, some technique is used to normalize matching scores provided by different libraries. For instance, score normalization can be accomplished by looking for duplicate documents in different lists. The presence of one document in two different lists is exploited to normalize

scores in the two lists. Unfortunately, this approach cannot be used if one or more lists do not contain documents retrieved also by other retrieval engines.

A more sophisticated approach, based on cross similarity of documents, is presented in [7]. Retrieved documents are used to build a local collection. Then, the similarity of each retrieved document to each document in the local collection is evaluated. Cross similarities between documents are used to normalize matching scores. The main limitations of this approach are related to its computational complexity (cross similarities between all pairs of retrieved documents should be computed) and to the fact that it requires the extraction of content descriptors from each retrieved document. This latter requirement can be easily accomplished for text documents, but implies critical computational requirements if applied to images.

In order to lessen these requirements, in [9] a new approach is presented that is based on the use of a local archive of documents to accomplish score normalization. Periodically, libraries are sampled in order to extract a set of representative documents. Representative documents extracted from all the libraries are gathered into a central data fusion archive. When a new query is issued to the libraries, it is also issued to the central search engine that evaluates the similarity of the query with all documents in the central data fusion archive. Then, under the assumption that each retrieved list contains at least two documents that are also included in the central data fusion archive, linear regression is exploited to compute normalization coefficient (for each retrieved list) and normalize scores. This approach has been successfully tested for text libraries for which resource descriptions were extracted using *Query based Sampling* [2]. However, if resource descriptions are not built using query based sampling, the assumption that retrieval results of each library contain at least two documents of the central archive is rarely verified. In these cases, linear regression cannot be applied and matching scores are left un-normalized. Thus, results cannot be merged.

3 Data Fusion for Image Libraries

The proposed solution develops on the method presented in [9]. Improvements are introduced to address two major limitations of the original method: applicability of the solution to the general case, regardless of the particular solution adopted for resource description extraction; reduction of computational costs at retrieval time. These goals are achieved by decomposing the fusion process in two separate steps: *model learning* and *normalization*.

Model learning is carried out once, separately for each library: it is based on running a set of *sampling queries* to the library and processing retrieval results. In particular, for each sampling query, the library returns a list of retrieved images with associated matching scores. These images are reordered using a fusion search engine that associates with each image a normalized score. Then, the function mapping original scores onto normalized scores is learned.

Differently, normalization takes place only at retrieval time. It allows, given a query and results provided by a library for that specific query, to normalize matching scores associated with retrieved results. This is achieved by using the normalization model learned for that library during the model learning phase.

In doing so, collection fusion is achieved through a model learning approach: during the first phase parameters of the model are learned; during the second one the learned model is used to accomplish score normalization and enable fusion of results based on normalized scores merging. Model learning and normalization are described in detail in the following sections.

3.1 Score modelling

Let $\mathcal{L}^{(i)}(q) = \{(d_k, s_k)\}_{k=1}^{n}$ be the set of pairs document/score retrieved by the i-th digital library as a result to query q. We assume that the relationship between un-normalized scores s_k and their normalized counterpart σ_k can be approximated to a linear relationship. In particular, we assume $\sigma_k = a * s_k + b + \epsilon_k$, being ϵ_k an approximation error (*residue*), a and b two parameters that depend on the digital library (i.e. they may not be the same for two different digital libraries, even if the query is the same) and on the query (i.e. for one digital library they may not be the same for two different queries). The approximation error ϵ_k accounts for the use of heterogeneous content descriptors (the digital library and the data fusion server may represent image content in different ways) as well as for the use of different similarity measures (even if the digital library and the data fusion server use the same image content descriptors—e.g. 64 bins color histogram—they may use two different similarity measures—e.g. one may compare histograms through a quadratic form distance function, the other may use the histogram intersection).

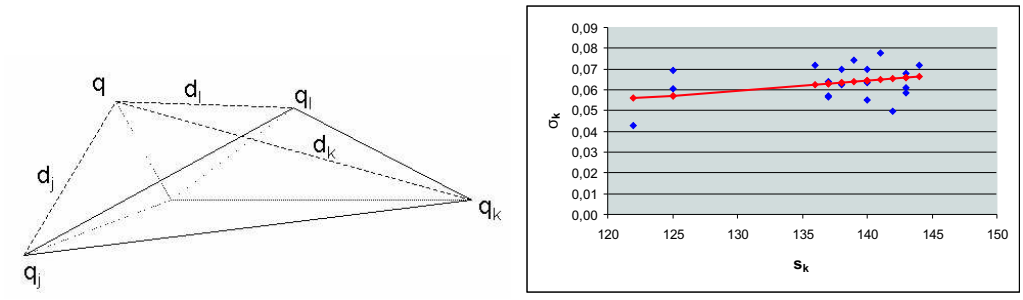
Values of parameters a and b are unknown at the beginning of the learning process. However, during learning, values of these parameters are computed, separately for each library. For each sample query, the value of parameters a and b is estimated. Computing the value of parameters a and b for different queries is equivalent to sampling the value distribution of these parameters on a grid of points in the query space. Knowledge of parameter values on the grid points is exploited to approximate the value of parameters for new queries. This is accomplished by considering the position of the new query with respect to grid points.

Approximation of parameters a and b for a new query is carried out as follows. If the new query matches exactly one of the sample queries (say q_k) used during the learning process, then the value of a and b is set exactly to the value of a and b that was computed for the sample query q_k. Otherwise, the three grid points q_j, q_k, q_l that are closest to the new query are considered. The value of a and b for the new query is approximated by considering values of a and b as they were computed for queries q_j, q_k, q_l. In particular, if these values were a_j, b_j, a_k, b_k, a_l and b_l, then values of a and b are estimated as:

$$\begin{aligned} a &= \frac{d_j}{D}a_j + \frac{d_k}{D}a_k + \frac{d_l}{D}a_l \\ b &= \frac{d_j}{D}b_j + \frac{d_k}{D}b_k + \frac{d_l}{D}b_l \end{aligned} \tag{1}$$

being $D = d_j + d_k + d_l$ and d_j, d_k and d_l the Euclidean distance between the new query and grid points q_j, q_k, q_l, respectively (see Fig.1(a)).

Using Eqs.1 is equivalent to estimate values of parameters a and b as if the new query lain to the hyperplane passing by q_j, q_k and q_l. For this assumption to be valid, the new query should be close to the hyperplane. The closer it is, the more precise the approximation.



(a) (b)

Fig. 1. (a) Position of a new query wrt the approximating points on the grid; (b) Normalized scores σ_k with respect to un-normalized ones s_k for a sample query, and the straight line approximating their relationship.

In the proposed approach, linear regression is the mathematical tool used for the estimation of normalization coefficients, given a sample query and the corresponding retrieved documents. In our case, data points are pairs $(\sigma_i,\ s_i)$ being s_i the original matching score of the i-th document and σ_i its normalized score. The application of this method to a collection of pairs $(\sigma_i,\ s_i)$ results in the identification of the two parameters a and b such that $\sigma_i = a * s_i + b + e_i$.

Fig.1(b) plots, for a representative sample query, the values of normalized scores σ_k with respect to un-normalized ones s_k. The straight line is that derived during the learning phase by computing parameters a and b of the linear regression between scores σ_k and s_k. This shows that the linear model assumption is quite well fulfilled, at least for the first few retrieved images.

3.2 Selection of Sample Queries

Through the learning phase. distribution of parameters a and b is sampled for a set of queries. Selection of these *sample queries* affects the quality of the approximation of a and b for a generic query in the query space during normalization.

Basically, two main approaches can be distinguished to select sample queries. The first approach completely disregards any information available about the distribution of documents in the library. In this case, the only viable solution is to use sample queries uniformly distributed in the query space. A different approach relies on the availability of information about the distribution of documents in the library. This should not be considered a severe limitation since this kind of information is certainly available to accomplish the resource selection task. In particular, this information is available in the form of a *resource descriptor* capturing the content of the entire library. Typically, in the case of image collections, resource descriptors are obtained by clustering library documents and retaining information about cluster centers, cluster population and cluster radius. Clustering is performed at several resolution levels so as to obtain multiple cluster sets, each one representing the content of the library at a specific *granularity* [1].

Resource descriptors can be used to guide the selection of sample queries by using the cluster center themselves as sample queries. In this way, the distribution of sample queries conforms to the distribution of documents in the library.

Once the resource description process is completed, each cluster center c_k is used as a query to feed the library search engine. This returns a set of retrieved images and associated matching scores s_i. Then, c_k is used as a query to feed the reference search engine (i.e. the search engine used by the data fuser) that associates with each retrieved image a normalized (reference) matching score σ_i. Linear regression is used to compute regression coefficients (m_k, q_k) for the set of pairs (σ_i, s_i). In doing so, each cluster center c_k is associated with a pair (a_k, b_k) approximating values of regression coefficients for query points close to c_k. The set of points (c_k, a_k, b_k) can be used to approximate the value of regression coefficients (a, b) for a generic query q by interpolating values of (m_k, q_k) according to Eqs.1.

4 Experimental Results

The proposed solution to data fusion for distributed digital libraries has been implemented and tested on a benchmark of two different libraries each one including about 1000 images. It is assumed that both libraries represent image content through 64 bins color histograms. However, they use two different similarity measures, namely L_1 and L_2 norm, in order to compare image histograms.

In Fig.2 results of the fusion process are shown for the two sample libraries.

The query image is shown in Fig.2(a). Results returned by the first library and their original scores are shown in Fig.2(b). The same information is shown in Fig.2(c) for the second library. Fused results, ordered based on their normalized scores are shown in Fig.2(d). It can be noticed that, based on the un-normalized scores returned by the two libraries, all the images in Fig.2(c) would be ranked before images in (b), with a potential loss of relevant results if, for example, only the first eight images are returned to the user. Differently, application of the proposed data fusion technique allows a more effective re-ranking of retrieved images lists, which better conforms to the user expectation.

Experimental results report on both the quality of normalization coefficients approximation and the overall quality of the data fusion process.

Each library was processed separately. Libraries were first subject to the extraction of resource descriptors according to the procedure presented in [1]. Cluster centers extracted through the resource description process were used as sample queries. For each sample query, values of normalization coefficients were computed, as explained in Sect.3.1. Values of normalization coefficients for sample queries were used to approximate the value of normalization coefficients for new queries. In particular, the quality of the approximation was measured with reference to a random set of test queries, not used during the learning phase. For each test query, values of normalization coefficients were approximated according to Eqs.1. Actual values of normalization coefficients were also computed by running the query and following the same procedure used for learning (Sect.3.1). Comparison of actual and approximated values of normalization coefficients gives a measure of the approximation quality. Let a and b be the estimated values of normalization parameters and $\hat{a}$ and $\hat{b}$ their actual values. Quality of the approximation is measured by considering the expected values

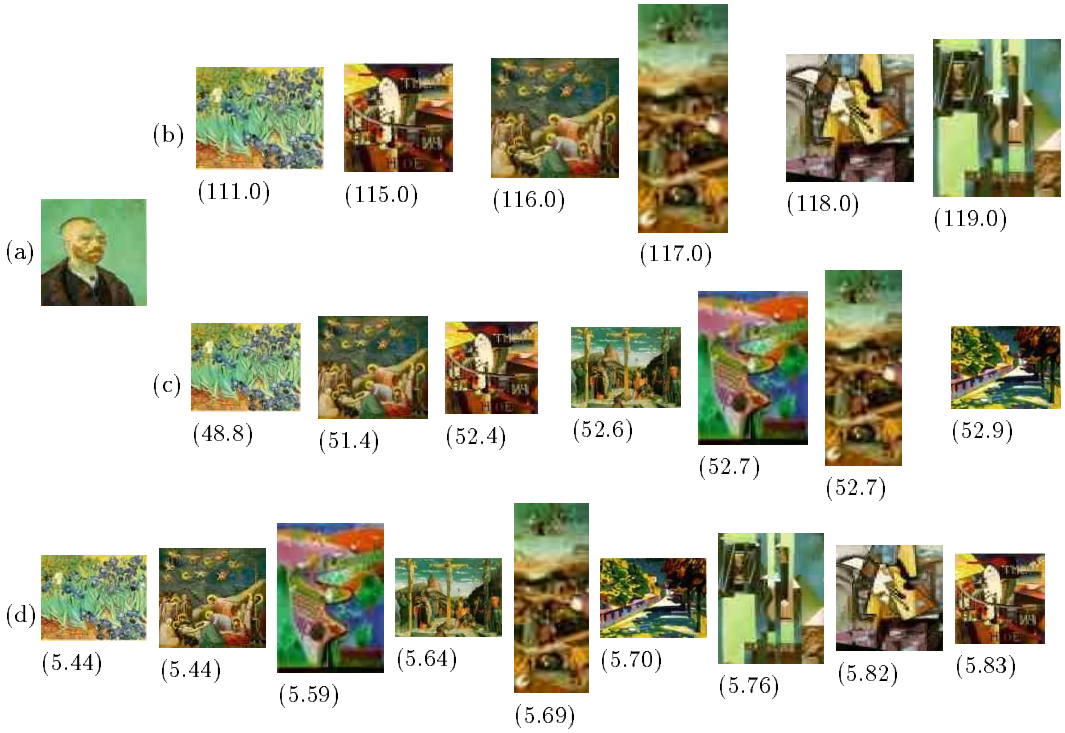


Fig. 2. Fusion of results for two distinct collections. (a) The query image; (b) images and matching scores retrieved from the first library; (c) images and matching scores retrieved from the second library; (d) fused results with normalized matching scores.

of the normalized error:

$$\epsilon_a = E\left[\frac{|a-\hat{a}|}{1+|\hat{a}|}\right], \quad \epsilon_b = E\left[\frac{|b-\hat{b}|}{1+|\hat{b}|}\right] \tag{2}$$

Plots in Fig.3(a) report values of the expected normalized error for different values of the granularity of sampling queries. Granularity is expressed as percentage of the number of query points with respect to the number of documents in the digital library. As expected, the error decreases when the number of sample points used in the learning phase is increased (i.e. finer granularity).

To represent the overall quality of the data fusion process we considered the presence of duplicated images in the retrieved lists. For a generic query, each library returns a list of retrieved documents. The presence of images that are included in both libraries can be used to assess the quality of the data fusion process. In particular, we measure the quality of data fusion as the capability of assigning the same *normalized* score to duplicate images. Let $\mathcal{L}^{(1)}(q) = \{(d_k^{(1)}, s_k^{(1)})\}_{k=1}^{n}$ and $\mathcal{L}^{(2)}(q) = \{(d_k^{(2)}, s_k^{(2)})\}_{k=1}^{m}$ be the set of pairs document/score retrieved by the two test digital libraries as a result to query q. Let the presence of duplicate elements be represented through a function $f : \{1,\ldots,n\} \mapsto \emptyset \cup \{1,\ldots,m\}$ that associates with the index of a generic document in the first list either the index of the same document in the second list (if it exists) or the null value.

We assume that the optimal data fusion should associate with duplicate images the same normalized score, that is: $\sigma_i^{(1)} = \sigma_{f(i)}^{(2)},\ \forall f(i) \neq \emptyset$. Thus, a measure of the overall quality of the data fusion process is expressed through the following functional:

$$\mathcal{F} = \frac{1}{\#\{i = 1,\ldots,n \mid f(i) \neq \emptyset\}} \sum_{i=1,\ldots,n \mid f(i)\neq\emptyset} 2\frac{|\sigma_i^{(1)} - \sigma_{f(i)}^{(2)}|}{\sigma_i^{(1)} + \sigma_{f(i)}^{(2)}} \tag{3}$$

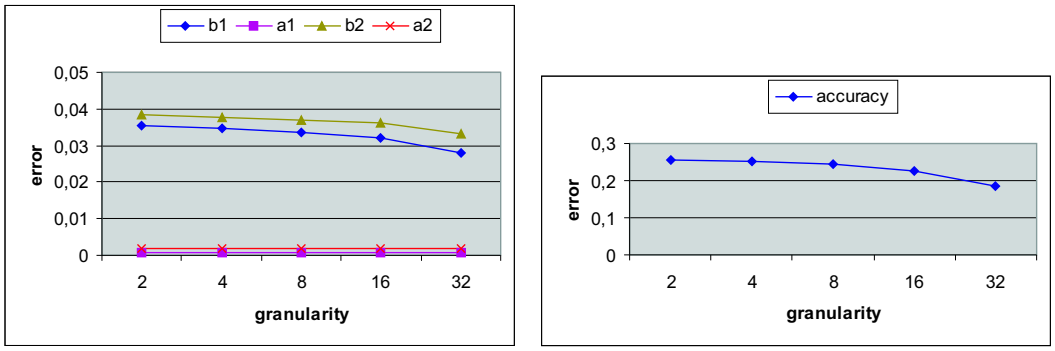
that is, the sum of the differences of the normalized scores with respect to their average values. The lower the value of $\mathcal{F}$ the higher the quality of the data fusion. The value of $\mathcal{F}$ is 0 for the optimal data fusion.

Plots in Fig.3(b) show the value of $\mathcal{F}$ averaged on 100 random queries. Values are reported for 5 distinct granularity levels.

It is worth noting that both plots in Fig.3(a)-(b) are affected by an error component originated by the heteronegeous descriptors used by the data fuser engine and the digital libraries under test.

5 Conclusion and Future Work

In this paper, an approach is presented for merging results returned by distributed image libraries. The proposed solution is based on learning and approximating normalization coefficients for each library. Since the major computational effort is moved to the learning phase, the approach has a low processing time at retrieval time, making it applicable to archives of images. In addition, differently from other fusion methods which rely on the presence of common documents in different result lists, the proposed fusion strategy is completely general, not imposing any particular constraint on the results. Preliminary results are reported to demonstrate the effectiveness of the proposed solution.



(a) (b)

Fig. 3. (a) Normalization error for approximated parameters of the two digital libraries. b_1, a_1 and b_2, a_2 correspond to the digital libraries using the L_1 and L_2 metric, respectively. (b) Accuracy of the data fusion measured on the two archives.

Future work, will address experimentation and comparison of the effect of the learning technique on fusion effectiveness. Moreover, issues related to the relationships between granularity of sample queries and accuracy of fusion will be investigated.

References

1. S. Berretti, A. Del Bimbo, P. Pala. Using Indexing Structures for Resource Descriptors Extraction from Distributed Image Repositories. *In Proc. IEEE Int. Conf. on Multimedia and Expo*, vol.2, pp.197-200, Lousanne, Switzerland, August 2002.
2. J. Callan and M. Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems*, vol.19, n.2, pp.97-130, 2001.
3. W. Chang, G. Sheikholaslami, A. Zhang, T. Syeda-Mahmood. Efficient Resource Selection in Distributed Visual Information Retrieval. In *Proc. of ACM Multimedia'97*, Seattle, 1997.
4. N. Fuhr. Optimum Database Selection in Networked IR. *In Proc. of the SIGIR'96 Workshop on Networked Information Retrieval*, Zurich, Switzerland, August 1996.
5. L. Gavarno and H. Garcia-Molina. Generalizing Gloss to Vector-Space Databases and Broker Hierarchies. *In Proc. of the 21st Int. Conf. on Very Large Data Bases*, pp.78-89, 1995.
6. K.L. Kwok, L. Grunfeld, D.D. Lewis. TREC-3 Ad-hoc, Routing Retrieval and Thresholding Experiment using PIRCS. *In Proc. of TREC-3*, 1995, pp.247-255.
7. S.T. Kirsch. Document Retrieval Over Networks wherein Ranking and Relevance Scores are Computed at the Client for Multiple Database Documents. US patent 5659732.
8. H.V. Jagadish, N. Koudas, S. Muthukrishnan, V. Poosala, K.C. Sevcik, T. Suel. Optimal Histograms with Quality Guarantees". VLDB 1998, pp.275-286.
9. L. Si and J. Callan. Using sampled data and regression to merge search engine results. *In Proc. of International ACM SIGIR Conference on Research and Development in Information Retrieval* pp.19-26, Tampere, Finland, 2002.
10. E.M. Vorhees, N.K. Gupta, B. Johnson-Laird. Learning Collection Fusion Strategies. *In Proc. ACM-SIGIR'95*, 1995, pp.172-179.
11. E.M. Vorhees, N.K. Gupta, B. Johnson-Laird. The Collection Fusion Problem. *In The Third Text REtrieval Conference (TREC-3).*