This paper is published in Proc. of International Workshop on Multimedia Information Systems, pp. 81-88, Capri, Italy, 2001.

MULTI-MODAL RETRIEVAL FOR MULTIMEDIA DIGITAL LIBRARIES: ISSUES, ARCHITECTURE, AND MECHANISMS

Jun Yang[1], Yueting Zhuang[1] and Qing Li [2]

[1] Department of Computer Science, Zhejiang University, Hangzhou, CHINA
yangjun@acm.org; yzhuang@cs.zju.edu.cn

[2] Department of Computer Science, City University of Hong Kong, Tat Chee Ave., KLN, Hong Kong, CHINA
csqli@cityu.edu.hk

ABSTRACT

Supporting effective and efficient retrieval of multimedia data is a challenging problem in building a digital library. In this paper, we examine the issues related to accommodating multi-modal retrieval of multimedia data (text, image, video and audio), and propose *2M2Net* as a generic framework for such versatile retrieval in multimedia digital libraries. The retrieval is conducted based on the integration of multi-modal features including both semantic keywords and media-specific low-level features. This framework is capable of progressive improvement of its retrieval performance, by applying the *learning-from-elements* strategy to propagate keyword annotations, as well as the *query profiling* strategy to facilitate effective retrieval using historic information of the previously processed queries.

1. INTRODUCTION

As the most complex and advanced multimedia information systems, digital libraries are emerging at an increasingly fast rate throughout the world. One of the primary difficulties in building a digital library is to support effective and efficient retrieval of the media objects from the whole library, including text, image, video and audio. The current mainstream of the retrieval technologies in most digital libraries is keyword-based retrieval[5]. Although such technology works well with textual document, it cannot, by itself, accomplish the retrieval task in a multimedia digital library, mainly due to the limited expressive power of keyword to describe or index media objects. Feature-based retrieval, on the other hand, is proposed to index and search for media objects such as image, video and audio [1,2,3] based on their respective low-level features. This paradigm reflects to a certain extent the similarity between media objects at the perceptual level, such as visual and auditory similarity. However, in most cases it cannot achieve the retrieval accuracy that the keyword-based approach can reach, because low-level features cannot be easily associated with the intrinsic semantics of media objects, while keywords explicitly describe the semantics. Therefore, integrating feature-based retrieval with keyword-based approach provides great potentials of improved indexing and retrieval for digital libraries.

No matter which approach is adopted, two problems essential to the retrieval task in the context of digital libraries have to be addressed. First, *how to be multi-modal?* That is, the retrieval approach must be able to search for multimedia data of various modalities (text, image, video and audio), and it should exploit an integration approach to facilitate the multi-modal retrieval task. Second, *how to be progressive?* As a major impediment of retrieval performance, it is commonplace that content of a digital library is not adequately indexed either by semantics or by low-level features. The retrieval facilities must be therefore intelligent enough to improve its performance progressively by learning from the history of previously conducted queries.

To address the aforementioned issues, we propose *2M2Net* as a multi-modal framework for multimedia retrieval in digital libraries. It is characterized as being multi-modal in two aspects:

- Retrieval of various multi-modal data such as text, image and video can be conducted.
- Multi-modal features including both semantic keywords and low-level features are seamlessly integrated for retrieval purpose.

Moreover, this framework employs the following mechanisms to make itself progressive:

- *Learning-from-elements* for propagation of keyword annotation at the semantic level.
- *Query profiling* to facilitate feature-based retrieval based on querying history.

The rest of the paper is organized as follows. In Section 2, we provide an overview of the *2M2Net* framework. In Section 3 and 4, we discuss some key issues regarding semantic-level and feature-level retrieval respectively. We then demonstrate how the feature-level and semantic-level retrieval can be integrated in Section 5. The implementation issues of the prototype system are discussed in Section 6. Finally we present the concluding remarks in Section 7.

2. VERVIEW OF THE 2M2Net FRAMEWORK

The architectural framework of *2M2Net* is illustrated in Figure 1. In the context of this framework, a digital library is viewed as a collection of *multimedia documents*[1], which is recursively defined as a logical document consisting of several elements that are multimedia documents by themselves or individual media objects such as text, image, video and audio. A multimedia document is a semantic grouping of multimedia data, that is, all its elements share a common semantic subject. The concrete forms of a multimedia document can be a web page, a portion of a digital encyclopedia and other forms of multimedia data collection. Since multimedia documents can be constructed recursively, we are able to model many composite documents in real world, e.g. a newspaper, or a website. Multimedia document is internally represented by means of its *semantic skeleton*, which maintains the metadata of both high-level semantics and low-level features for each element in the document.

Multimedia documents are firstly pre-processed so that their various elements are extracted out and stored into the corresponding databases in the **Storage Subsystem**. The metadata of these elements, including semantic keywords as well as media-dependent low-level features, are extracted to constitute the semantic skeleton. User queries are handled by the **Query Processor** at either semantic level or feature level. In the **Feedback & Learning Subsystem***,* a set of feedback techniques specific to each media type is utilized for short-term refinement of retrieval results. For long-term improvement of retrieval performance, the *learning-from-element* strategy is applied to propagate and update semantic keywords, and the *query profile* is constructed to facilitate effective retrieval using querying history.

[1] If not indicated explicitly, document is referred to multimedia document in this paper

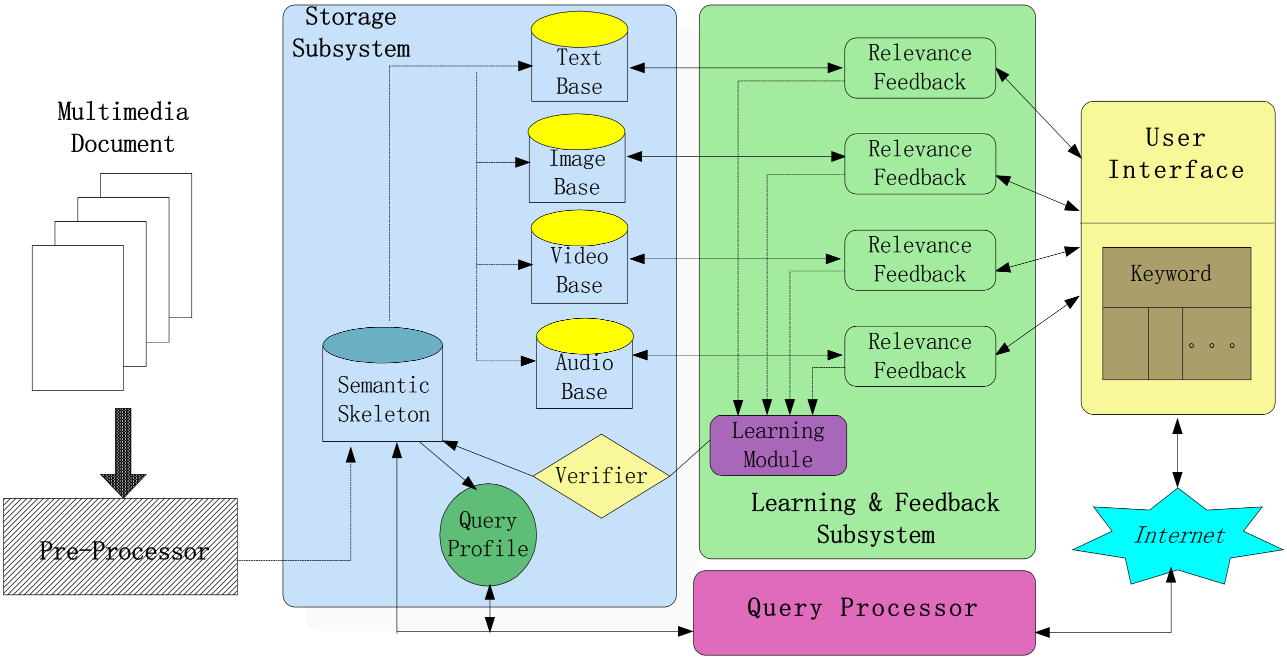


Figure 1: The *2M2Net* framework

3. SEMANTIC-LEVEL RETRIEVAL

In this section, we discuss two key issues regarding this semantic-level retrieval, as initial semantic analysis and the learning-from-element strategy.

3.1. Initial Semantic Analysis

In *2M2Net*, each document and media object has a list of weighted keywords attached to it as their semantic annotation. When a document is pre-processed, we perform semantic analysis to obtain the keyword annotation for the document and its elements. This analysis is straightforward for a textual element, as many traditional IR [5] techniques can be applied to extract representative keywords and calculate keyword weight from itself. The semantics of non-textual object such as image and video, however, cannot be extracted from its content by the current image (video) understanding techniques. Their semantics are acquired indirectly using the following heuristic method.

By our definition of multimedia document, the elements of a document are semantically correlated to each other, so that semantics of a non-textual element can be inferred from related textual elements. Such inference greatly depends on the concrete form of the document. In a digital encyclopedia, the accompanying text and captions of images and videos can be used to represent their semantics. In a Web environment, besides the above text sources, we can take the advantage of using URLs, link strings and other HTML tags to be the descriptions of image and video contained in the web page. The keyword weight is determined heuristically in this case. For example, among all text sources related to an image, we regard the image caption is of the highest relevance and thus give it a large weight. In this way, each keyword can be assigned an estimated weight. The keyword annotation of the whole document can be obtained similarly.

Besides the intra-document correlation used by this heuristic method, there is also inter-document semantic correlation that can be further explored for semantic analysis. Such inter-document correlations widely exist in a digital library, indicated by structural neighborhood and hyperlinks between documents or media objects. By analyzing the structure of these semantic links inside a digital library, we can induce the semantics of a document or a media object from other data directly or indirectly linked to it. However, currently we use solely the intra-document correlation in semantic analysis for the sake of simplicity.

After the keyword annotation become available, the semantic retrieval can be handled by matching the query with the annotation of each candidate document or media object. The matching can be conducted by simply counting the intersection between the query and object annotation in terms of common keywords, or by more sophisticated thesaurus-based semantic similarity measure.

3.2. Learning-from-Elements

The semantics obtained by the heuristic semantic analysis is likely to be incomplete, inaccurate or even non-existing, so that we propose the *learning-from-elements* strategy to propagate and improve the keyword annotations progressively and interactively during relevance feedback. This strategy can be thought as a kind of semantic feedback, because it is triggered when the user submits a set of the documents or media objects as the feedback examples for a given query. As illustrated in Figure 2, it propagate keywords along three directions, which are from user query to feedback examples of documents or media objects (scheme A), from a document to its elements (scheme B) and from a media object to its parent document (scheme C).

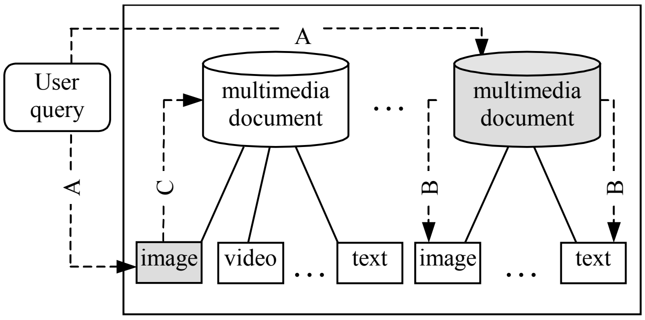


Figure 2: Learning-from-elements strategy

In scheme A, we adopt a simple voting algorithm to update the keyword list of the documents and media objects that are designated as feedback examples. This algorithm is described as follows. For a positive example, we add each query keyword into its keyword list. If the query keyword is already there, its weight is increased by a certain step. For a negative example, if there is any query keyword in its keyword list, we remove it from there. By applying this algorithm on each feedback example, the involved documents and media objects learn their semantics implicitly from users. Besides, the representative keywords with a majority of user consensus are likely to receive a large weight.

Scheme B and C are more ambitious propagation schemes that utilize the intra-document semantic correlation. We apply scheme B if the positive feedback example is a media object. In this case, the keywords inserted or updated (in terms of weight) by scheme A are propagated to its parent document. If the positive example is a document, we use scheme C to spread some of its keywords to its elements. To avoid spreading erroneous keywords, only the keyword with highest weight in the list may be propagated in both scheme B and C, because the top keyword is likely to represent the actual semantics of the document/object. Compared with scheme A, scheme B and C can spread the query keywords to more documents or media objects including those that are not designated as feedback examples, so that they are particularly advantageous when users are reluctant to give many feedbacks. But they are also not as reliable as scheme A. One likely concern of using these two schemes is a tradeoff between wide coverage of keywords and possible erroneous keywords. We argue that a rich set of keywords (perhaps imperfect) is more desirable than a small set of precise keywords for retrieval purpose, and erroneous keywords are relatively easy to identify and correct.

4. FEATURE-LEVEL RETRIEVAL

In this section, we discuss feature-level retrieval as an alternative to the keyword-based retrieval described in the previous section.

4.1. Feature Extraction and Matching

In *2M2Net* framework, the low-level feature of each media object is extracted in the pre-processing phase, such as the color and texture feature for image, structural and motion feature for video, etc. In feature-based retrieval, the user is required to submit a media object as the query example, and the results are retrieved based on the similarity of low-level features. We also incorporate into the framework a set of feedback techniques specific to each media type, including the image feedback technique proposed by Rui [4] and video feedback technique proposed by Wu [6]. Although these techniques are capable of improving retrieval results immediately, they make no contribution to long-term retrieval performance, because they do not memorize the previously conducted feedbacks and start from scratch for the new query.

4.2. Query Profiling

Query profiling is the counterpart of learning-from-elements strategy at the feature level. It can be regarded as an incremental feedback technique that memorizes the history of previous user feedbacks to assist the processing of future queries. The details of this strategy are given below.

For each media object O_i, we construct a query profile to record the objects that were designated as relevant or irrelevant to it in the past feedbacks. The profile is divided into two lists, one for relevant objects (denoted as L_{pi}) and the other for irrelevant objects (denoted as L_{ni}). Each object in the list has a weight attached to it. Initially, the query profile for each media object is empty. Later, when a media object O_i is selected as the query example and some objects are appointed as feedback examples to it, we update its profile using the voting scheme similar to that described in Section 3.2. For each object designated as positive example, we check to see if it is already in L_{pi}. If so, we increase its weight by a certain increment; otherwise we add it into L_{pi} with an initial weight. If we find this object in L_{ni}, it is immediately removed from there. The L_{ni} list of the query example can be updated in a similar way from the negative examples. The query profile can be utilized when O_i is used as query example again. In this case, we select the top *N* objects with the highest weights from L_{pi} and L_{ni} to be the positive and negative feedback examples, respectively. Then, the feedback techniques can be applied directly on these past feedback examples, without any real feedback conducted by the current user.

As more queries and feedbacks are submitted from users, the query profile for each media object can be constructed and improved incrementally. Higher retrieval performance is achievable by using query profiles, since the current retrieval can be brought to a higher level of accuracy achieved by many iterations of relevance feedbacks performed previously.

The query profile described above is associated with each media object and therefore domain-specific. On the other hand, we can construct user-specific profiles that model the different preferences and habits of performing query and feedback for individual users. Finally, personalized query profiles can be established by combing the query profiles and user profiles, which embody the characteristics of both the specific domain and the specific user.

5. INTEGRATION OF SEMANTICS AND FEATURES

In the previous sections, we discussed the retrieval approaches at the semantic level and feature level respectively. In this section, we show that the semantics and low-level features can be seamlessly integrated throughout the whole working flow of the framework to enhance its performance.

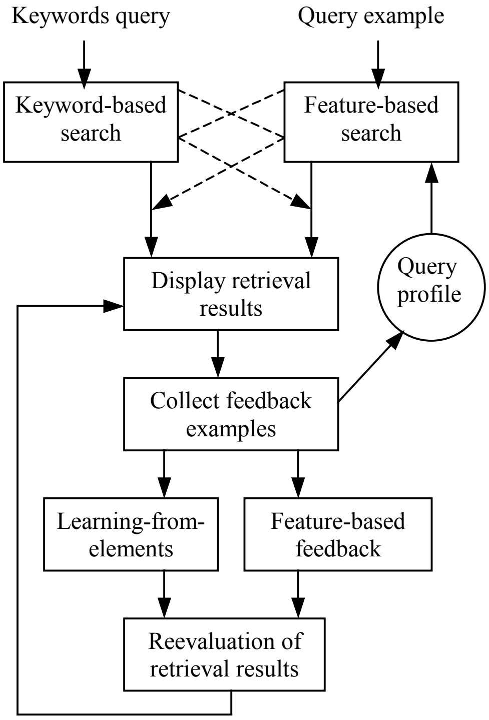


Figure 3: Integration of semantics and low-level features

As illustrated in Figure 3, the user can conduct either a keyword-based search by inputting a set of query keywords, or a feature-based search by submitting a query example. These two search paradigms can be combined to facilitate each other to achieve a better performance. For example, when the keyword annotations of media objects are insufficient, the matches returned by a keyword-based search are usually limited. In this case, we can start a feature-based search using the objects retrieved by keyword search in top ranks as the query examples, in order to find more media objects that resemble them. This second pass search provides many potential results, which, although not very precise, are important supplements to the keyword search. Similarly, feature-based search can benefit from keyword-based approach by accommodating the semantic similarity between the query example and other media objects.

After collecting the feedback examples from the user, the system conducts feedback process in parallel at the semantic level and the feature level. At the semantic level, the learning-from-elements strategy described previously is applied to propagate the query keywords among the related documents or media objects, given that the retrieval is initiated by a keyword-based search. If it is a feature-based search, the keywords in the annotation of the query example are propagated using the same strategy. At the feature level, many conventional relevance feedback techniques are utilized to improve the quality of low-level features, by means of adjusting their weights or revising the distance metric. The specific technique employed depends on the media type that is currently dealt with, including those for images [4], for videos [6] and possibly for audios (if any). For feature-based queries, we also need to update the query profile of the query example accordingly.

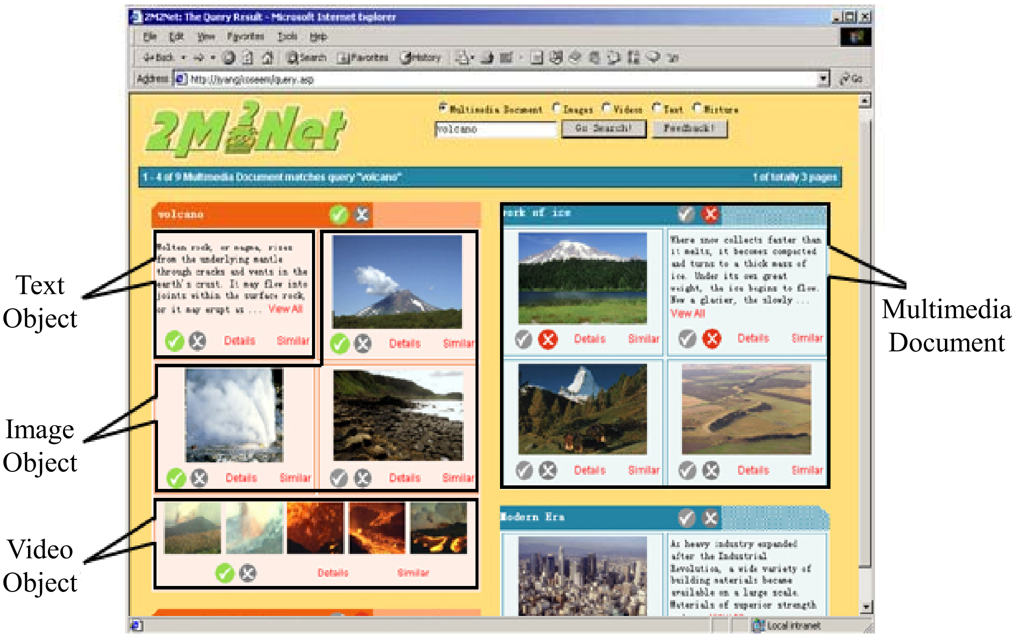
Finally, we recalculate the retrieval results based on the improved semantics and low-level features. Each candidate object is evaluated of its similarity to the query using a comprehensive metric that accommodates its similarity to the initial query, to the positive feedback examples and to negative examples. To achieve this, a uniform distance metric function that measures the similarity between a candidate object O_i and the query is given as follows:

$$S_i = \alpha R_i + \beta \left\{ \frac{1}{N_R} \sum_{k \in O_R} \left[(1 + R_{ik}) S_{ik} \right] \right\} - \gamma \left\{ \frac{1}{N_N} \sum_{k \in O_N} \left[(1 + R_{ik}) S_{ik} \right] \right\}$$

where α, β and γ are suitable constants, O_R and O_N are set of relevant and irrelevant media objects of a certain type, N_R and N_N are the number of objects in O_R and O_N. R_i is the semantic similarity between the object O_i and the initial query Q, calculated as the number of common keywords between Q and the annotation of O_i. If the retrieval is started with a content-based query, R_i is the similarity between O_i and the object designated as query example. R_{ik} is the similarity between O_i and the positive (or negative) feedback example with subscript k, calculated in the same way as R_i in the case of keyword query . S_{ik} is the their similarity in terms of low-level features. For textual object that has no low-level features, S_{ik} is simply set to 1.

6. IMPLEMENTATION ISSUES

A prototype system for multimedia retrieval in a digital encyclopedia has been built based on the proposed framework. The system is composed of a back-end and a front-end. The back-end is responsible for the processing, storage, authoring and retrieval of multimedia data, while the front-end is a web browser based interface that is in charge of all user-system interactions.



1 - 4 of 9 Multimedia Document matches query "volcano"

1 of totally 3 pages

Molten rock, or magma, rises from the underlying mantle through cracks and vents in the earth's crust. It may flow into joints within the surface rock, or it may erupt at ... View All

There snow collects faster than it melts, it becomes compacted and turns to a thick mass of ice. Under its own great weight, the ice begins to flow. Now a glacier, the slowly ... View All

As heavy industry expanded after the Industrial Revolution, a wide variety of building materials became available on a large scale. Materials of superior strength

2M2Net: The Query Result - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://syang/cosem/query.asp
Go
2M2Net
● Multimedia Document ○ Images ○ Videos ○ Text ○ Picture
volcano
Go Search!
Feedback!
volcano
Details
Similar
rock of ice
Modern Era
Local intranet
Text Object
Image Object
Video Object
Multimedia Document

Figure 4: The query results as multimedia documents

The system offers users with great flexibility to perform retrieval task. The user can choose to search for multimedia documents or media objects of a certain modality by a simple keyword-based approach. The main user interface shown in Figure 4 displays the multimedia documents retrieved for the query of "volcano". A multimedia document is rendered as its sketch, i.e. the abstracts of textual objects, thumbnails for image objects and key-frame lists for video objects that are shown together. For each media object, the user can click on the "Details" link below it to view the original media object and other related information. Besides keyword-based search, the user can conduct a feature-based search using a specific media object as the query example by clicking the "Similar" link below it. Each document and media object displayed as retrieval result has a "√" and a "×" icon attached to it that denotes positive and negative example respectively. The user can indicate feedback examples by clicking on the icons and signal the system to perform feedback by clicking the "Feedback" button.

7. CONCLUSIONS

In this paper, we have described *2M2Net* as a multi-modal framework for multimedia retrieval in digital libraries. Among others, *2M2Net* can accommodate retrieval of multi-modal data such as text, image, video and audio, based on the integration of multi-modal features including semantic keywords and the media-specific low-level features. This framework is capable of progressively improving its retrieval performance by applying the learning-from-elements and query profiling strategy. A prototype system has been built upon which the proposed framework is implemented.

Being the most popular operations, browsing and navigation of the multimedia documents are compulsory to be devised. The former is usually facilitated by their subject categories, whereas the latter is to traverse from one document to another related one either semantically or structurally. However, our current prototype system does not maintain any subject category, nor does it track the inter-document links through which the users navigate. In our future work, we plan to include the support of browsing and navigation functionalities into the current prototype system.

REFERENCE

[1] Chang, S. F., Chen, W., Meng, H. J., Sundaram, H., Zhong, D., "VideoQ: An Automated Content Based Video Search System Using Visual Cues", ACM Multimedia, 1997.

[2] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., "Query by image and video content: The QBIC system." *IEEE Computer*, 1995.

[3] Rui, Y., Huang, T. S, Chang, S.F. "Image Retrieval: Current Technologies, Promising Directions and Open Issues", Journal of Visual Communication and Image Representation, Vol. 10, pp39-62, 1999.

[4] Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S., "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval", IEEE Trans on Circuits and Systems for Video Technology*,* Special Issue on Segmentation, Description, and Retrieval of Video Content, Vol 8, pp644-655, 1998.

[5] Salton, G., Buckley, C. "Introduction to Modern Information Retrieval", McGraw-Hill Book Company, New York, 1982.

[6] Wu, Y., Zhuang, Y. T., Pan, Y. H., "Relevance Feedback of Video Retrieval", in Proc. of the first IEEE Pacific Rim Conference on Multimedia, pp 206-209, December, 2000.