This paper is published in the Proc. of the 2nd IEEE Pacific-Rim Conference on Multimedia, pp. 482-489, Beijing, China, 2001.

SEARCH FOR MULTI-MODALITY DATA IN DIGITAL LIBRARIES

Jun Yang [1] *Yueting Zhuang* [1] *Qing Li* [2]
[1] Microsoft Visual Perception Laboratory of Zhejiang University
Hangzhou, 310027, China
yangjun@acm.com, yzhuang@cs.zju.edu.cn

[2] Dept of Computer Science
City University of Hong Kong
Kowloon, HKSAR, China
csqli@cs.cityu.edu.hk

ABSTRACT

Developing effective and efficient retrieval techniques for multimedia data is a challenging issue in building a digital library. Unlike most previously proposed retrieval approaches that focus on a specific media type, this paper presents *2M2Net* as a seamless integration framework for retrieval of multi-modality data in digital libraries. As its specific approaches, a learning-from-elements strategy is devised for propagation of semantic descriptions, and a cross media search mechanism with relevance feedback is proposed for evaluation and refinement of user queries. Experiments conducted on a digital encyclopedia manifest the effectiveness and flexibility of our approaches.

1. INTRODUCTION

Digital libraries are becoming the most complex and advanced form of information systems that is used to store, access, share and disseminate multimedia data of various types including text, image, audio and video [1]. A challenging issue in building a digital library is to support effective and efficient retrieval of such multi-modality data in the whole library.

Currently, many digital library systems rely on text-based retrieval [4] technologies for retrieving information from the library. While such technologies could be tolerable for textual documents, they have limited applicability to digital libraries due to lack of effective means to specify queries for multimedia data. The problem is even amplified when the multimedia data are not well annotated. In contrast, content-based retrieval has been proposed to index and search multimedia data by their low-level features. This approach has been respectively adopted to image, video and audio retrieval [3][5]. Formulating a query in this approach is to create or select a representative media object as example and search for other objects that resemble to it in terms of low-level features, denoted as query-by-example. However, the performance of this approach is very low when the semantics of multimedia data cannot be readily represented by their low-level features.

Not surprisingly, most approaches currently available for multimedia retrieval are dedicated to a certain media type and thus inapplicable to a digital library that contains multi-modality data. Moreover, there are great constraints on the means by which a user can formulate his/her query. A user must know clearly which media type to search for, and has an appropriate way to express the query by submitting a good example. However, this is very inconvenient and inflexible for users of a digital library. Quite often, a user may have only a vague idea of his/her information need, and/or has no proper retrieval example at hand that can be directly mapped to his/her information need.

To address the limitations of the current retrieval technologies, we propose a seamless integration framework, *2M2Net*, for retrieval of multi-modality data in digital libraries. It features a learning-from-elements strategy for propagation of the semantic descriptions, as well as a cross media search mechanism that is tailored to multi-modality data. Within this framework, user queries can be (re-)formulated in a flexible and convenient way. Finally, a retrieval system using the proposed framework has been built on a digital encyclopedia.

This paper is organized as follows. In Section 2, we present the architectural framework of *2M2Net*. Two specific approaches of the framework are described in Section 3. In Section 4 we present the implementation issues and the experimental evaluations of the system. Conclusion and future works are given in Section 5.

2. THE ARCHITECTURAL FRAMEWORK

Our proposed framework does not simply put together the existing retrieval methods for each media to handle multi-modality data. Instead, it fully explores the semantic correlation existed among various media objects of a digital library to enhance the retrieval performance. It is named as *2M2Net* due to our intension to model the multi-modality (the first "2M"), multimedia data (the second "2M") in a digital library as a *Net*work at the semantic layer.

A digital library can be 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. Each document has a semantic subject that is shared with all its elements. The physical counterparts of multimedia document include web page(s), a chapter of encyclopedia and other forms of multimedia data collection.

The framework of *2M2Net* is illustrated in Figure 1. Multimedia documents are firstly pre-processed so that their various elements are extracted out and stored into the corresponding databases in the **Storage Subsystem**. In *2M2Net*, a multimedia document **D** is represented by means of its semantic skeleton, S^{D}*= (ID, Title, URL, Keyword-list, Element-set)*, where *Keyword-list* is a list of weighted keywords describing the document semantics, and *Element-set = (Texts, Images, Videos, Audios, …)* is a set of component media objects. Each media object is represented by its low-level features and semantics as descriptive keywords, e.g., *Image = (Keyword-list, Image-features)*. In the pre-processing phase, the semantics and low-level features of each media object

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

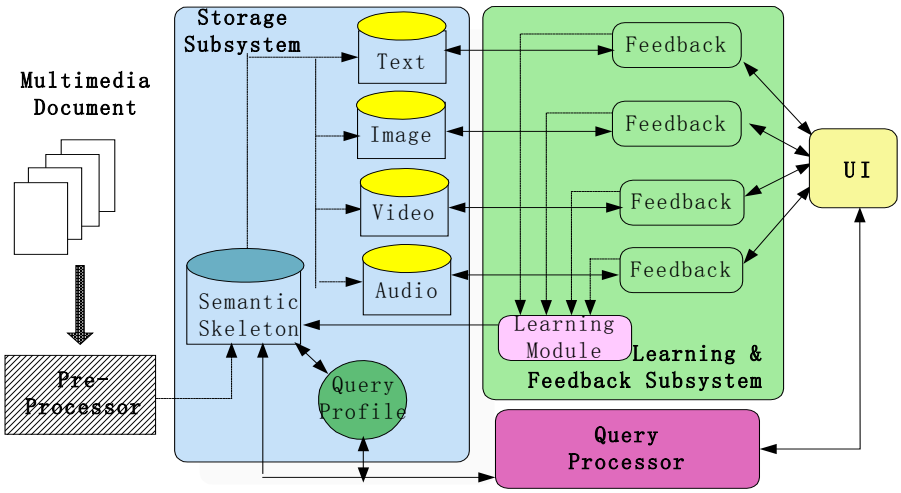

Figure 1: The *2M2Net* Framework

are also extracted. The semantics of text object is directly obtained from itself using the traditional IR techniques. For non-textual objects such as image and video, their semantics can be obtained from the accompanying textual descriptions (e.g. surrounding text, captions, HTML tags), depending on the physical form of the document. All the features and semantics extracted are stored into semantic skeleton base to construct the initial semantics skeleton.

User query is processed based on semantic skeleton by the **Query Processor**, which conducts a cross-media search to retrieve relevant documents or media objects. This mechanism allows a simple keyword-based search to induce a suite of more sophisticated content-based retrievals to be conducted. User feedback is accepted and handled by the **Feedback & Learning Subsystem**. It conducts a parallel session of relevance feedback on each media type to improve the retrieval results immediately. Meanwhile, a learning-from-element process is performed to propagate descriptive keywords among semantically related documents and media objects, which is likely to enhance the retrieval performance in a long term. This can be also regarded as an incremental construction process of the semantic skeleton by enriching its semantic part. The **Query Profile** is designed to expedite and optimize the retrieval process by memorizing the history of complex queries and their resolutions. However, it is not implemented in the current system.

3. THE SPECIFIC APPROACHES

In this section, we describe two key approaches employed by the proposed framework, which are learning-from-elements strategy and cross-media search mechanism with relevance feedback.

3.1 Learning-from-Elements

In practice, the initial semantics of documents and media objects acquired from pre-processing is inaccurate, incomplete or even non-existing. In view of this, we devise a machine learning strategy, learning-from-element, to supplement their semantics by propagating descriptive keywords among them. It is triggered whenever a user submits a set of the documents and media objects as feedback examples for a given set of query keywords. As illustrated in Figure 2, there are four directions in which keyword propagation can take place: from user query to feedback examples of documents or media objects (type A), from a document to its elements (type B), from a media object to its parent document (type C) and from a media object to visually similar ones of the same modality (type D).

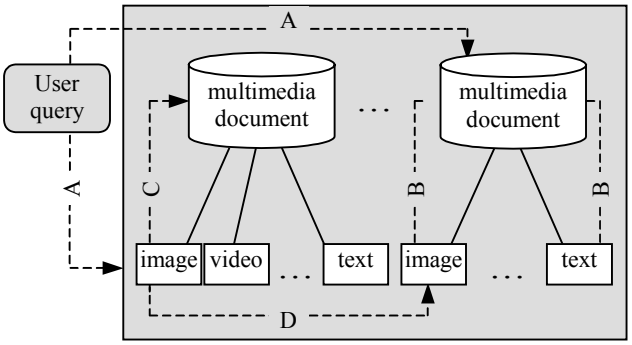

Figure 2: Keyword Propagation Scheme

Propagation from the query to the feedback examples (type A) serves as the starting point of the whole learning process. Its mechanism is described as follows. For a positive example, we add each query keyword into its keyword list. If the query keyword is already in the list, its weight is increased by a certain step. For a negative example, if any query keyword is in its keyword list, we remove it from that list. By conducting such propagation for each feedback example, the semantics of involved documents and media objects are implicitly learnt from users. Also, the correct and representative keywords with a majority of user consensus are likely to receive a large weight.

Propagation types B and C are more ambitious schemes that utilize the semantic correlation among the elements of a document. If the positive feedback example is a media object, the keywords inserted or affected (in terms of their weights) by type A scheme are propagated to its parent document (type B); otherwise if the positive example is a document, some of its keywords are pass to its elements (type C). Both propagations should be done with great

care to avoid spreading wrong keywords. Currently, we apply a simple verification process by examining if each keyword to be propagated is the one with largest weight in the keyword list. The top keyword is likely to represent the actual semantics of the document/object and therefore can be further propagated. The negative examples are not considered for such propagations, since otherwise correct keywords can be possibly removed from other documents or media objects by erroneous propagation.

Type D is applied each time a new image or video is registered into the system. Considering the difficulty of obtaining initial semantics for such non-textual object, this scheme attempts to make some guesses of its semantics based on low-level features. That is, for each new image or video, a content-based search is performed to find other images or videos that visually resemble to it. The keywords collected from the top N matches are inserted into the keyword list of the new object.

Among the four propagation schemes, type A is the most reliable one since the keywords are obtained directly from user queries. Type B and C intend to make the best of user feedbacks by further spreading the query keywords to related documents or media objects that are not designated as feedback examples. Thus, they are particularly advantageous when users are reluctant to give many feedbacks. A possible concern of these two schemes is the tradeoff between a wide coverage of keywords and the 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. Moreover, it would be relatively easy to identify and correct these imprecise or even erroneous keywords (if any). Due to the insufficiency of low-level features to address semantics, type D is the most ambitious scheme.

3.2 Cross Media Search and Relevance Feedback

In *2M2Net*, a user can choose to search for either multimedia documents or media objects of a certain type using a keyword-based approach. We use the document search as an example to describe the retrieval process. The similarity R_i of document D_i to the query is calculated as the total weight of keywords in common between the query and the keyword list of D_i:

$$R_i = \sum_{k=1}^{M} w_{ik} \qquad (1)$$

where M is the number of common keywords between D_i and the query, w_{ik} is the weight of the kth such keyword of D_i. All the candidate documents are compared with the query and returned in the descending order of their similarity. The retrieval of media objects can be conducted in a similar way.

Since the keyword descriptions are insufficient initially, the matches returned by the keyword-based search can be quite limited. In this case, a suite of more sophisticated keyword or feature-based search is triggered intending to find more promising candidates. If the query is for documents, an exhaustive keyword search is conducted. That is, for each document that is not matched previously, we merge the keywords of all its elements and reevaluate the similarity by comparing the query against this combined keyword list. Otherwise if the query is for a specific media, we conduct a content-based search to find more media objects that are visually similar to top match retrieved in the keyword-based search. The documents or media objects obtained in this second-pass search are ranked behind the matches of the keyword-based search.

The system proceeds to the feedback phase when the user submits a set of documents or media objects as feedback examples. The whole feedback process is described as follows. First, we collect the feedback examples from the user. Then, the learning-from-elements strategy described in Section 3.1 is applied to propagate the keywords among the involved documents or media objects. Finally, a parallel session of relevance feedbacks is conducted on each type of media or on document, subject to the user's choice.

A hybrid approach of semantics- and feature-based relevance feedback is proposed by Lu et al. [2] for image retrieval. We generalize this method so that it can accommodate all kinds of media. It is formulated as a uniform distance metric function that measures the similarity between a media object and the query:

$$S_i = \alpha R_i + \beta \left\{ \frac{1}{N_R} \sum_{k \in O_R} [(1 + R_{ik}) S_{ik}] \right\} - \gamma \left\{ \frac{1}{N_N} \sum_{k \in O_N} [(1 + R_{ik}) S_{ik}] \right\} \qquad (2)$$

where α, β and γ are suitable constants, O_R and O_N are the relevant and irrelevant media objects of a certain type, N_R and N_N are the number of media objects in O_R and O_N. R_i is the semantic similarity between the ith candidate object and the initial query defined in (1). R_{ik} is the similarity between the ith object and the kth positive/negative feedback example, which is also calculated using (1) by taking the keywords of the feedback example as the query keywords. S_{ik} is the their similarity on the low-level features. For textual object that has no low-level features, S_{ik} is set to 1. This function can be adapted and used for multimedia document, by regarding O_R and O_N as the relevant and irrelevant document collections and setting each S_{ik} to 1. Therefore, we can adopt (2) to calculate the improved retrieval results as either media objects of a specific type or the whole documents.

4. IMPLEMENTATION AND EXPERIMENTS

To show the effectiveness of the proposed framework, a prototype system using this framework has been established for conducting multimedia retrieval in a digital library.

4.1 The System Setup

The system consists of a back-end and a front-end. The back-end is responsible for the processing, storage, authoring and retrieval of multimedia data. The front-end is a web-accessible interface that handles all user-system interactions by communicating with the back-end. The main user interface is as shown in Figure 3, which displays the relevant documents retrieved for the query of "water". A multimedia document is visualized as its sketch, i.e. abstracts for text, thumbnails for images and key-frame lists for videos. The keyword search can be aimed at a certain media type as well. In addition to the keyword-based search mode, the user can perform a content-based search using a specific media object as the query example by clicking on the "Similar" link below it.

Figure 3: The query results of multimedia documents

In the main interface, each document or media object has a "√" and a "×" icon attached to it, denoting positive and negative example respectively. The user can indicate feedback examples by clicking on them and then press the "Feedback" button to signal the system to perform feedback. Despite what is retrieved initially, users can choose to refine the retrieval results as either multimedia documents or media objects by feedbacks. That is to say, the user can start with searching for documents and then switch to a specific media during feedbacks, or vice versa. This mode is particularly useful to the user who does not previously have a clear preference on which media to search, and/or cannot specify the initial query precisely. By roughly searching for documents firstly, relevant media objects of various types are returned, among which the user may find promising candidates and narrow the search range accordingly in the following feedbacks.

4.2 Experiments

Some simple experiments are conducted to show the effectiveness of the system. The test data are collected from Microsoft Encarta Interactive World Atlas 2000, which is a part of the Encarta Encyclopedia. The "World Tour" part of the atlas has 19 categories, with each category being further divided into several topics (totally 160 topics). In each topic, there are usually several images, text paragraphs and sometimes a video clip. We regard each topic as a multimedia document and feed all the media objects inside it into the system manually. The initial semantics are obtained from the title of category and topic (for document), and from the caption (for image, video and text paragraph).

In the experiments, we input some keywords to search for multimedia documents and perform feedbacks by marking relevant documents as positive examples. It is noticed that after an average of 4 iterations, the query keywords are spread from the involved documents to 90% of their elements, along with a remarkable rise of retrieval performance. A similar experiment is conducted on a specific media such as image, which shows that the keywords can be also efficiently propagated from media objects to their parent document.

A performance evaluation in terms of precision and recall is not conducted for two reasons. First, a ground-truth database required for performance evaluation is very hard to construct, since most digital libraries are not well annotated, indexed or classified. In addition, considering the variety of search and feedback modes supported by the system, a survey has to be made to study the user behaviors of posting queries and giving feedbacks. However, such a survey requires the help from a large number of human subjects, which is not currently available to us.

5. CONCLUSION AND FUTURE WORK

In this paper, we have presented a novel framework of *2M2Net* for retrieval of multi-modality data in digital libraries. As its specific approaches, the learning-from-elements strategy is devised for interactive propagation of keyword descriptions, and the cross media search mechanism with relevance feedback is proposed for evaluation and refinement of user queries. As demonstrated by the prototype system built on a digital encyclopedia, this framework allows users to (re-)formulate the query in a flexible and convenient way and provides satisfactory performance.

The major contribution of our work is providing a generic framework for retrieval of multi-modality data, instead of any specific retrieval technique for a certain media. This framework is general and open enough to accommodate other media types (such as audio), or incorporate any other retrieval or feedback algorithms for each media.

In the keyword-based search, the currently used exact keyword matching method is unable to address the similarity between relevant (but different) keywords. As our future work, we plan to utilize thesauruses to define a similarity metric between different keywords and therefore improve the accuracy of keyword-based search. Another interesting future work is to establish the query profile for a digital library. The query file intends to memorize the history of some complex queries as a sequence of user-system interaction, as well as the resolutions to them. The next time a similar query is encountered, the results can be directly deduced from the query profile without exhaustively searching the whole library, thereby optimize and expedite the retrieval process.

6. REFERENCES

[1] Communications of ACM, special issue on digital libraries, vol. 41, no. 4, ACM Press, 1998.

[2] Lu, Y. et al, "A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems", ACM Multimedia, 2000.

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

[5] Smoliar, S.W., Zhang, H.J. "Content-based Video Indexing and Retrieval", IEEE Multimedia, Vol.1, pp356-365, 1994.