A Unified Retrieval Method of Multimedia Documents

Yu Suzuki, Kenji Hatano, Masatoshi Yoshikawa, Shunsuke Uemura
Nara Institute of Science and Technology
Graduate School of Information Science
8916-5 Takayama, Ikoma, Nara, 630-0101, Japan
{yu-su, hatano, yosikawa, uemura}@is.aist-nara.ac.jp

Abstract

In this paper, we propose a retrieval method for multimedia documents. In our method, features of each medium such as text, image, and their layout information are extracted from documents. For a given query, similarity values are calculated for each media, then those values are integrated into one similarity value per document. We propose four methods for calculating the integrated similarity values. Also, we performed experiments using PDF files, and verified the effectiveness of our method.

1 Introduction

The dissemination of the Internet technology have allowed us to deal with electronic documents which consist of not only text data but also multimedia data on current computers. HTML [1] which is used for the format of documents on the WWW (World Wide Web) is typical of it, it includes not only text data but also image, video, and sound data. For that reason, the information retrieval techniques which can treat these multimedia data are looked upon as important in recent research field of information retrieval, because we can see that electronic documents will include many kinds of media in the future. That is to say, we believe that IR techniques of the electronic documents should be able to deal with not only text data but also multimedia data; these media have to be treated equally. In order to realize these IR techniques, we have to input not only keywords but also more information as clues into information retrieval systems to retrieve electronic documents which we want to. To sum up, we have to develop information retrieval systems that can handle the documents contained multimedia data, and that can treat queries consisted of many kinds of information.

Zhang et al. suggested document retrieval system called WEBSSQL [3] for HTML to integrate multimedia retrieval technique. This system can retrieve multimedia documents, however, there is no strict layout on HTML format, it is difficult for users to describe layout information as queries.

In this paper, we propose a retrieval method for multimedia documents to integrate each media retrieval technique. In our method, we decompose documents into media objects such as terms and images, and extract documents' features such as term frequencies of text data, color histograms of image data, and layout information of these data as feature vectors. When users retrieve documents, our retrieval system calculates similarities of users' query in terms of each medium and also calculates scores of electronic documents to integrate these similarities. Furthermore, we performed experiments using our methods to PDF files, and verify the efficiency of our proposed methods.

2 Unified Retrieval Method of Multimedia Documents

In this section, we describe a way to extract features from a PDF file. In the description of PDF format, there are one or more objects, and an object contains a text or an image data. An object has some information such as the object's layout and the object's content. An overview of our proposed system are shown in Figure1. First, we extract PDF document's features from an PDF document. Extracted document's features are text, images, and their layout information. In order to get these information, we decompose the PDF document into each feature.

In our retrieval method, we extract as many PDF document's features as possible. This is because we use many kinds of PDF document's features to describe users' interests exactly. If we can express users' interests using extracted document's features, we can develop an information retrieval system that reflects users' interests using these features. In this paper, we develop the retrieval system for PDF files consisting of text and image data. In the rest of this section, we provide a more detailed description of extracted PDF document's features and generation of its feature vectors and indices.

Proceedings of the Seventh International Conference on Database Systems for Advanced Applications (DASFAA '01)
0-7695-0996-7/01 $10.00 © 2001 IEEE

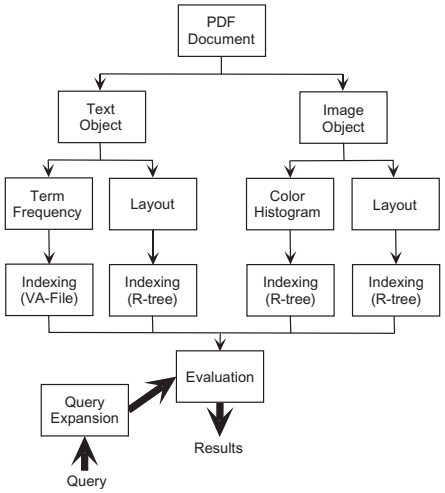

Figure 1. An Overview of Our Proposed System

This method output results as documents, not objects. However, it has already been mentioned that evaluation values are calculated to each object. For that reason, we define documents' evaluation value unifying evaluation values of each medium. Let X_i be an evaluation value of a document D_i. We can denote that evaluation value X_i is composed of evaluation values of each medium, so we define text and image evaluation values, X_i^{term} and X_i^{image}, of document.

- the arithmetic mean value

$$X_i = \frac{X_i^{term} + X_i^{image}}{2} \quad (1)$$

- the geometric mean value

$$X_i = \sqrt{X_i^{term} \cdot X_i^{image}} \quad (2)$$

- the maximum value

$$X_i = \max\{X_i^{term}, X_i^{image}\} \quad (3)$$

- the probability of consisting high value evaluation

$$X_i = 1 - ((1 - X_i^{term}) \cdot (1 - X_i^{image})) \quad (4)$$

In experiments for evaluation of our retrieval system, we cannot compare our system with current digital library system, because the query of our system is composed of not only keywords but also image and their layout information. Hence, we check the efficiency of our proposed method and decide calculation method of documents' similarities to do the following two experiments:

- Compare the retrieval accuracies of our proposed system and that of current digital library system if we use only keywords as a query.
- Compare the retrieval accuracies of our proposed system when we use only keywords as query with when we use all information extracted from electronic documents as a query.

First experiment shows that retrieval accuracy of our retrieval system is higher than that of ACM and IEEE Digital Libraries. This is because information retrieval model of these digital library systems are boolean model, and that of our retrieval system is vector model[2].

In second experiment, we found that evaluation function which handles each evaluation value equally is better than the evaluation function which pretend to handle one evaluation value. In summary, we can say that evaluation function (4) and (1) are relevant for our retrieval system. Especially, the evaluation function (4) is the most useful in this experiment.

3 Conclusion

In this paper, we presented a retrieval method for multimedia documents to unify the similarities calculated by retrieval methods for many kinds of medium. Our method can search multimedia documents because the method handle feature of many kind of medium equally. Moreover, we verified the efficiency of our proposed method through some experiments, and found the efficiency of our proposed method over current retrieval methods.

Acknowledgements This work was partially supported by Grant-in-Aid for Scientific Research from Japanese Ministry of Education, Culture, Sports, Science and Technology (No. 11480088, 12680417, 12780309, 12208032), and by CREST of JST (Japan Science and Technology).

References

[1] D. Raggett, A. L. Hors, and I. Jacobs. HTML 4.01 Specification. http://www.w3.org/TR/html4/, December 1999.

[2] G. Salton. *Automatic Text Processing: The Transformational, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1988.

[3] C. Zhang, W. Meng, Z. Zhang, and Z. Wu. WebSSQL – A Query Language for Multimedia Web Documents. In *Proceedings IEEE Advances in Digital Libraries 2000 (ADL2000)*, 2000.