# Combining Multiple Sources of Evidence in XML Multimedia Documents: An Inference Network Incorporating Element Language Models

Zhigang Kong and Mounia Lalmas

Department of Computer Science, Queen Mary, University of London
{cskzg, mounia}@dcs.qmul.ac.uk

**Abstract.** This work makes use of the semantic structure and logical structure in XML documents, and their combination to represent and retrieve XML multimedia content. We develop a Bayesian network incorporating element language models for the retrieval of a mixture of text and image. In addition, an element-based collection language model is used in the element language model smoothing. The proposed approach was successfully evaluated on the INEX 2005 multimedia data set.

## 1 Introduction

We believe that structure can play an essential role in the retrieval of multimedia content in XML multimedia documents. The proposed approach makes use of the semantic structure and logical structure in XML documents, and their combination for representing and retrieving XML multimedia document content.

This work develops a general framework for combining multiple sources of evidence from various structured elements in XML multimedia documents. The multimedia content here refers to any type of multimedia data or a mixture of text and multimedia data. An element language model is applied upon each XML element. The framework combines the language models associated with the elements used to perform the retrieval of the multimedia content, using the Inference network model. An element-based smoothing method for the element language model is proposed.

The general framework has been applied in the context of a retrieval task based on a mixture of text and image retrieval on the INEX 2005 Multimedia collection [4].

## 2 The Proposed Approach

XML documents are composed of structured elements that are nested within each other in a hierarchical tree. There are *logical* relationships between these nested elements. The elements logically surrounding a given element can be used to provide additional sources of evidence for representing this given element. In addition, the elements closer to it in the document hierarchy could provide more accurate representation than those further away.

XML document contents are surrounded by text markups. These, which here refer to the element names, can provide meaningful semantics that can be viewed as metadata describing the nature of the element. This semantics can also be used for the

G. Amati, C. Carpineto, and G. Romano (Eds.): ECIR 2007, LNCS 4425, pp. 716 – 719, 2007.

representation and retrieval of elements. We call the structure based on the meaningful markup *semantic structure*.

The two types of structures are based on different characteristics. The former is identified according to document logical hierarchy and the latter is classified according to the semantics of the markups (the name of the elements). They work in different ways to represent multimedia contents in XML documents. It would be expected that the combination of these two types of representations could lead to better effective retrieval of multimedia content than that based on only one of them.

We can directly query the semantically structured elements as well as query the multimedia data. This can be viewed as a retrieval of a mixture of text and multimedia data. Each semantically structured element (or the multimedia data) is represented by its logically surrounding elements (as shown in the left network of figure 1).

Furthermore, there could be other semantic structures (not the queried structures) in the XML documents. We can further structure the elements logically surrounding a queried element into semantic ones and non-semantic ones (as presented in the right network of figure 1). This could improve the representation based on the surrounding elements.

We use the first method (the left network of the figure 1) in this work as the test collection is not suitable for evaluating the second method. Due to the nested tree structure of XML documents, those structured elements are disjoint from each other so that they have non-overlapping content.

## 3 An Inference Network Incorporating Element Language Models

The inference network framework was explicitly designed for combing multiple representations and retrieval algorithms [1]. However, the heuristic estimation formulas (such as tf-idf) used in [3] do not correspond well to real probabilities. To address this [1, 2] incorporate the language model in the Bayesian network. We follow this idea but apply it to XML multimedia retrieval.

Figure 1 shows the Bayesian network combining the structured elements. In the networks, the node *D* models a document. The one or more nodes *S* model the queried elements, the element containing multimedia data and/or the semantically structured elements. The nodes between *D* and *S* model the elements logically surrounding the queried elements (the *S* nodes), where *own* is the queried element itself, *1$^{st}$* is its parent element, and so on. The *Q* node models a query and the *I* node models the information need.

The probability of a structured element can be estimated as the probability of it generating the query. In this work, the Dirichlet prior smoothing is used for the smoothing:

$$P(q \mid e)=\prod_{t\in q}\left(\lambda_1 P(t \mid e)+\lambda_2 P(t \mid C)\right) \quad (1); \quad P(t \mid e)=\frac{f(t,e)}{|e|} \quad (2); \quad \lambda_1=\frac{|e|}{|e|+\mu} \quad \lambda_2=\frac{\mu}{|e|+\mu} \quad (3)$$

where $\mu$ is a parameter set to the average length of the same type of structured elements in the collection; for example, it is the average length of <history> elements in the collection when querying //history. $|e|$ is the length of the element. $f(t,e)$ is the occurrences of a query term in the element.

$P(t \mid C)$ is a collection language model used for smoothing. However, we focus on estimating the structured elements instead of the entire document. Therefore, we compute the $P(t \mid C)$ as the probability of observing the term in the collection of the same type of queried elements. For our previous example, it is the probability of observing the term in all <history> elements in this collection. This is called an element-based collection language model.

We use the #WAND operator in the combination as using #WSUM can result in the smoothing component having no influence on the ranking [2].

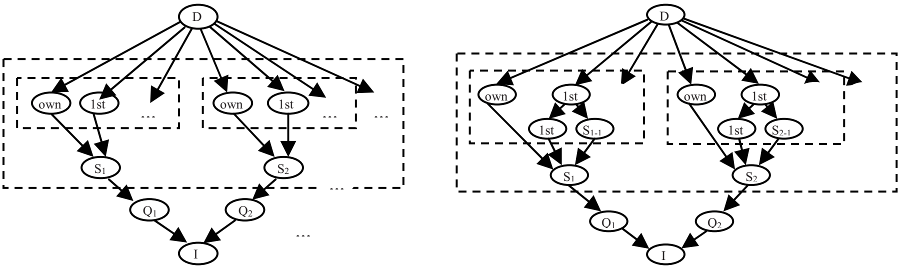

**Fig. 1.** The Bayesian networks for combining structured elements

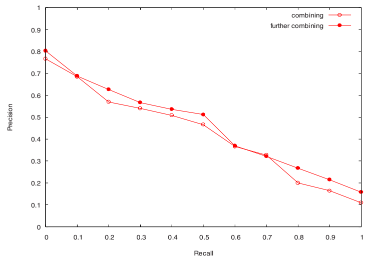

**Fig. 2.** Combining structured elements

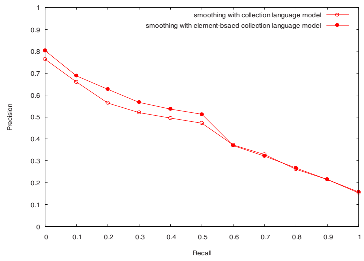

**Fig. 3.** Using different reference models

## 4 Retrieving a Mixture of Text and Image

We evaluate our approach using the INEX 2005 multimedia collection, which consists of 2633 images in 462 XML documents. It contains INEX CAS topics made of content and structural constraints. This work directly queries the image element and the structural constraints (viewing them as semantic structures).

Figure 2 shows the results of our approach. At first, we combine the semantically structured elements using their own contents. Then we further combine the logically surrounding elements to estimate their probabilities. The MAP and precision@10 are 0.4079 and 0.4105 in the former and 0.4408 and 0.4158 in the latter. Compared with the best official submission of the INEX 2005 multimedia track, the latter increases by 58.62% (MAP) and 33.91% (precision@10).

Further combining the logically surrounding elements of each semantically structured element achieves 8.07% improvement. We perform experiments to study each semantically structured element (as a query contains several structural constraints). The elements are grouped according to their depth. The results show that combining surrounding elements increases by 56.38% (MAP) and 31.86% (precision@10) for 2 depth elements, and 16.46% (MAP) and 30.73% (precision@10) for 3 depth elements.

The overall improvement of combining the logically surrounding elements is much lower than improvement of each structured element. This is due to most (12 out of 19) of the topics query for the document root element (/destination), which can not be improved. For those not querying for root element, the overall improvement is 16.00%, which increases 98.27% over that (8.07%) of all topics. Therefore, we can expect further improvement in the overall performance as there are only 7 topics not querying the document root element.

Figure 3 shows the results using different reference models in the smoothing. The first uses the standard collection language model and the second uses our element-based collection language model. The MAP of the former is 0.4184 and that of the latter is 0.4408. The latter increases 5.35% over the former.

As discussed above, most topics query for a document root element. In this situation, the element-based collection language model is the same as the standard collection language model. When restricted to topics not querying for a root element, the results show that the MAP of using element-based collection language model increases by 16.10% over that of using standard collection language model. Therefore, we expect the element-based collection language model to lead to improved effectiveness.

## 5 Conclusions and Future Work

This work makes use of the combination of semantic structure and logical structure in XML documents to represent and retrieve XML multimedia content. An inference network incorporating element language models was developed. In addition, we used an element-based collection language model. The experiments performed on the INEX 2005 multimedia collection showed promising results. Future work needs to be carried out into the use of the framework within a larger XML multimedia document collection.

## References

1. Croft , W. B. (2000). Combining approaches to information retrieval. In W. B. Croft, editor, Advances in Information Retrieval, pages 1--36. Kluwer Academic Publishers.
2. Metzler, D. and Croft, W. B. (2004). Combining the language model and inference network approaches to retrieval, IP&M, 40(5):735-750.
3. Turtle, H. and Croft, W. (1991). Evaluation of an inference network-based retrieval model. ACM TOIS, 9(3):187–222.
4. van Zwol, R., Kazai, G, and Lalmas, M. (2006). INEX 2005 multimedia track. Advances in XML Information Retrieval and Evaluation, INEX 2005 Workshop.