XML-Aided Phrase Indexing for Hypertext Documents

Miro Lehtonen
Department of Computer Science
P.O. Box 68
FI-00014 University of Helsinki
Finland
Miro.Lehtonen@cs.Helsinki.Fi

Antoine Doucet
GREYC CNRS UMR 6072 University of Caen
Lower Normandy
F-14032 Caen Cedex
France
Antoine.Doucet @info.unicaen.fr

ABSTRACT

We combine techniques of XML Mining and Text Mining for the benefit of Information Retrieval. By manipulating the word sequence according to the XML structure of the marked-up text, we strengthen phrase boundaries so that they are more obvious to the algorithms that extract multiword sequences from text. Consequently, the quality of the indexed phrases improves, which has a positive effect on the average precision measured by the INEX 2007 standards.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

General Terms

Algorithms

Keywords

XML, Phrase, Word sequence, Text mining, XML Retrieval

1. INTRODUCTION

XML Mining has previously been only applied to XML data and data-centric XML documents with the exception of clustering techniques that are also applicable to XML documents with textual content. XML-aware text mining has also been a quite neglected field of research, so far, despite its potential in indexing methods for XML Information Retrieval. In this paper, we show first how a whole class of text mining algorithms for multiword sequence extraction can be "aided" with the low level information coded in XML documents, and second, how an indexing method combined with one such algorithm yields a 14.9% improvement in average precision, according to the evaluation.

2. WORD SEQUENCE MANIPULATION

The document collection used for testing is the Wikipedia corpus in an XML format provided by the INEX evaluation initiative. It consists of nearly 660,000 articles of hypertext documents from the English Wikipedia. Most of the XML markup in the articles describes either the presentation of the content or the hyperlink structure of the corpus, both of which show as mixed content with inline level XML elements. In these cases, the start and end tags of the inline level elements denote the start and the end of a word sequence that we call an *inline phrase.* These phrases include the anchor texts of hyperlinks as well as phrases with added emphasis, e.g., italicized passages. An exact definition for the XML structures that qualify has been presented previously [7]. Intuitively, the inline phrases are highly similar to the multiword sequences that text mining algorithms extract from plain text documents. Therefore, the tags of the inline elements are strong markers of potential phrase boundaries. However, this observation has long gone wasted as the XML markup is ignored by sequence extraction algorithms.

```
...kernel trick has been applied to several algorithms in
<link>machine learning</link> <link>machine learning</link> and
<link>statistics</link> <link>statistics</link>, including...
```

Table 1: Two inline elements duplicated in place (XML attributes for link targets omitted).

Because text mining algorithms and, in particular, phrase extraction algorithms operate on word sequences without XML, we need to incorporate the explicit phrase marking tags into the word sequence — a task far from trivial. Our proposal is that *in-place duplication* of the inline element adds phrase boundary indicators of an appropriate strength to the word sequence. Examples of such duplication are shown in Table 1.

Considering the effect of duplication, we only look at the character data as the tags and other XML markup are removed before phrase extraction. The most obvious effect is the increase in phrase frequency of the duplicated inline phrases with a similar side effect on the individual words they compose of. Moreover, the distance between the words preceding and following the phrase increases, which makes the phrase boundaries more explicit to those phrase extraction algorithms that allow gaps in the multiword sequences.

3. PHRASE EXTRACTION

As a contrast to phrase extraction in general, we are only interested in providing support to those algorithms that define a phrase as a unit of text that consists of multiple words (a multiword sequence). The experiments in this paper are based on a definition called Maximal Frequent Sequence (MFS) [1], but we claim that the technique we propose will provide similar assistance to other relevant phrase extraction algorithms, as well.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

A frequent sequence is defined as a sequence of words that must occur in the same order more often than a given sentence-frequency threshold. MFSs are constructed by expanding a frequent sequence to the point where the frequency drops below the threshold. This way we obtain a compact phrasal description of a document collection [6].

We observe that phrase repetition also repeats the proximity of its components. As most phrase extraction techniques are based on the statistical analysis of co-occurrence phenomena, from simple adjacent n-grams [2, 3] to more advanced techniques [8], they benefit from the artificial repetition of the co-occurring words, should the repeated word sequences be true phrases.

Another approach to phrase extraction is based on text segmentation. The Voting Experts algorithm [4] draws a phrase boundary after a word where the entropy of the statistical distribution of the following words peaks. The text itself is used for learning the statistics. By duplicating a phrase, we lower the entropy inside the phrase which inherently creates a peak in the entropy immediately after the phrase and before the start of the duplication.

Hence, many phrase extraction techniques could have been chosen for our experiments. The MFS extraction algorithm is merely one way to test and demonstrate the XML-aided phrase indexing approach. This also implies that our evaluation focuses on assessing the added value of the XML structure and not that of any single phrase extraction algorithm.

4. SCORING OPTIONS AND RESULTS

In our test environment, the phrase index with MFS's is accompanied with a word index based on the vector space model. A Retrieval Status Value (RSV) is computed for each index. The Word RSV comes from the cosine of vectors with tfidf weights. The phrase RSV is calculated using the technique presented in [5]. It accounts for several variations of phrase usage, i.e., inversion of the constituents and discontinuous phrases. While our goal is not to evaluate the quality of the scoring technique, it is a necessary system component as we want to learn about the impact of duplicate phrases on phrase extraction and phrase RSV calculation. Both RSV values are later combined into a single value through normalisation and linear combination.

Regarding the word sequences, we test three different configurations for computing the combined RSV score: 1) Both RSVs are computed from the original word sequence, 2) Both RSV's are computed from the manipulated word sequence with inline elements duplicated, and 3) The Word RSV is computed from the original sequence, whereas the Phrase RSV is computed from the sequence with duplication ("Hybrid"). The result sets corresponding to the configurations (Runs 1–3) are computed for the CO topics of the INEX 2007 adhoc track. Altogether 107 topics are currently included in the official evaluation, the results of which are reported in Table 2.

According to a topicwise comparison of the results, the best configuration is Run 2 where the duplication affects both the phrase index and word index. However, both configurations involving duplication result in significantly higher Mean Average interpolated Precision (MAiP) values than Run 1 which can be considered a baseline: Run 2 for 56/107 topics (p=0.011) and Run 3 for 64/107 topics (p=0.004). Although the MAiP declines for 40–47% of the topics, the overall improvement of 11.4% in MAip and 14.9% in AiP @0.01 over the baseline indicate that we gain more precision than we lose by the word sequence manipulation. Because the difference between Run 2 and Run 3 is not significant (p=0.21), we cannot conclude whether term frequencies should be calculated from the original or the manipulated word sequence. However, phrase indexing clearly improves if the sequence with duplication is used.

	1) Orig.	2) w/Dupl.	3) Hybrid
iP 0.01 (107 topics)	0.3319	0.3773	0.3815
MAiP (107 topics)	0.0912	0.1024	0.1036

Table 2: Interpolated precision at 0.01 and Mean Average interpolated Precision of two official submissions (1,2) and an unofficial result set (3).

5. CONCLUSION

We have improved text mining in two ways by the XML-based manipulation of the word sequence. First, we are able to locate more high quality phrases in XML documents. Second, the correspondence between frequent multiword sequences and the high quality phrases of natural language is better: we now extract fewer unnatural phrases composed of words that just happen to co-occur frequently. Because most (83.8%) of the duplicated content come from the anchor texts of hyperlinks, we strongly believe that duplication when indexing phrases is also applicable to other hypertext documents such as HTML.

6. REFERENCES

[1] H. Ahonen-Myka. Finding all frequent maximal sequences in text. In *Proceedings of ICML-99 Workshop on Machine Learning in Text Data Analysis*, pages 11–17, 1999.

[2] S. Banerjee and T. Pedersen. The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the CICLing*, pages 372–383, 2003.

[3] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[4] P. Cohen, B. Heeringa, and N. Adams. Unsupervised segmentation of categorical time series into episodes. In *Proceedings of ICDM'02*, pages 99–106, Washington, DC, USA, 2002. IEEE Computer Society.

[5] A. Doucet and H. Ahonen-Myka. Non-contiguous word sequences for information retrieval. In *Proceedings of ACL-2004 Workshop on Multiword Expressions: Integrating Processing*, pages 88–95, July 2004.

[6] A. Doucet and H. Ahonen-Myka. Fast extraction of discontiguous sequences in text: a new approach based on maximal frequent sequences. In *Proceedings of IS-LTC 2006*, pages 186–191, 2006.

[7] M. Lehtonen and A. Doucet. Phrase detection in the Wikipedia. In N. Fuhr, M. Lalmas, A. Trotman, and J. Kamps, editors, *Focused access to XML documents, 6th INEX Workshop*, LNCS. Springer, 2008.

[8] O. Vechtomova. The role of multi-word units in interactive information retrieval. In *Proceedings of ECIR 2005*, pages 403–420, 2005.