PIX: Exact and Approximate Phrase Matching in XML

Sihem Amer-Yahia
AT&T Labs–Research
sihem@research.att.com

Mary Fernández
AT&T Labs–Research
mff@research.att.com

Divesh Srivastava
AT&T Labs–Research
divesh@research.att.com

Yu Xu
UC San Diego
yxu@cs.ucsd.edu

1. INTRODUCTION

XML permits the interleaving of text with structural and semantic markup in documents. Markup is added to a document by *tagging* a portion of the text or by augmenting the text with *annotations*. The example below is inspired from the XML documents published by the Library Of Congress (`www.loc.gov`). It describes legislative bills where bill sponsors are marked up using the tag `<sponsor>`, and an annotation `<footnote>` is added to demarcate parenthetical remarks in the text.

```
<bill bill-stage = 'Introduction'>
   <congress>110th CONGRESS</congress>
   <legis-num>H.R. 133</legis-num>
   <action-desc>
      <sponsor>Mr.  English</sponsor>
      <footnote>For himself and
      Mr.Coyne</footnote> introduced
      this bill.
   </action-desc>
</bill>
```

In the absence of markup, *phrase matching* is a common technique to search text and identify relevant documents. Phrase matching typically requires that words in a phrase be contiguous or in close proximity. For example, searching for the phrase "Mr. English introduced this bill" could return very different results than searching for the same set of words as individual keywords. Most information retrieval (IR) systems support phrase matching on text and on HTML documents, ignoring all HTML tags to match phrases.

In extending phrase matching to XML, which permits any user-defined markup, it should be possible to specify the individual tags and the complete annotations (i.e., elements and their content) to ignore. For example, to match the phrase "Mr. English introduced this bill" in the above XML document fragment, it is necessary to ignore the `<sponsor>` tag, as well as the entire `<footnote>` annotation. See [1] for an extensive list of such examples.

We present PIX, a system for Phrase matching In XML documents. The key features of PIX are:

- *flexibility*: allowing users to specify which tags and annotations to ignore when matching a phrase,
- *approximation*: permitting both exact and proximity phrase matching, and returning ranked results, and
- *efficiency*: relying on inverted indices and novel evaluation algorithms.

PIX's functionality is fully integrated into XQuery and naturally combines XPath navigation (to identify context nodes, and also the tags and annotations to ignore) with phrase matching.

SIGMOD 2003, June 9-12, 2003, San Diego, CA

2. DEMONSTRATION OVERVIEW

PIX extends GALAX (`db.bell-labs.com/galax`), a full-fledged XQuery engine, with functions that implement novel stack-based algorithms that rely on inverted indices on all words and tags in a document. These indices are built offline (statically). The phrase to be matched and the ignored markup are specified at query time (dynamically).

Query Specification: Users can visualize and select from a variety of input DTDs and documents. They can write XPath queries to identify context nodes as well as markup to ignore. They can choose among a set of sample queries and modify them. Users can ask for exact or proximity phrase matching.

Answer Ranking: Users can choose among a variety of scoring functions. They can also specify individual weights for the ignored markup that ranking functions will take into consideration.

Answer Explanation: The set of query answers are displayed to the user as an HTML document in which the matched phrase and the ignored markup are highlighted. All phrase matches found in a context node are highlighted.

A preliminary version of PIX has been demonstrated at `ICDE 2003` [2].

3. REFERENCES

[1] S. Amer-Yahia and P. Case. XQuery and XPath full-text use cases. W3C Working Draft. Available from http://www.w3.org/TR/xmlquery-full-text-use-cases/, Feb. 2003.

[2] S. Amer-Yahia, M. Fernández, D. Srivastava, and Y. Xu. PIX: A system for phrase matching in XML. In *Proceedings of the IEEE International Conference on Data Engineering*, 2003. Demonstration.