**DISSERTATION ABSTRACT**

# Using Topic Shifts in Content-Oriented XML Retrieval

**Elham Ashoori**
Department of Computer Science
Queen Mary University of London
UK, E1 4NS
*elham@dcs.qmul.ac.uk*
*http://www.dcs.qmul.ac.uk/~elham*

**Abstract**

Content-oriented XML retrieval systems support access to XML repositories by retrieving, in response to user queries, XML document components (XML elements) instead of whole documents. The retrieved XML elements should not only contain information relevant to the query, but also should be specific to the given query (i.e. do not discuss other irrelevant topics).

To score XML elements according to how relevant and specific they are given a query, the content and logical structure of XML documents have been widely used. This thesis aims to examine a new source of evidence deriving from the semantic decomposition of XML documents. We consider that XML documents can be semantically decomposed through the application of a topic segmentation algorithm. Using the semantic decomposition and the logical structure of XML documents, we define the notion of topic shifts in an XML element. We then formalise the number of topic shifts to reflect the element's relevance, and more particularly its specificness, to the given user's query.

This thesis investigates the use of topic shifts in content-oriented XML retrieval, which is mainly involved in retrieving information from semi-structured (XML) documents. First, we examine the characteristics of XML elements reflected by their number of topic shifts. Second, we use the number of topic shifts to estimate the relevance of the elements in the collection. Finally, we use topic shifts to provide a *focused access* to XML documents, which aims to determine not only relevant elements, but those at the right level of granularity.

The main contributions of this thesis are the introduction of topic shifts in the context of content-oriented XML retrieval and the extensive evaluation of the ways this evidence can be employed in retrieving XML elements. This thesis demonstrates that topic shifts in XML elements constitute a useful source of evidence for both improving the ranking of XML elements, and determining elements at the right level of granularity in content-oriented XML retrieval.

The thesis is available online at http://elham.ashoori.org/publications/phd-thesis.pdf