Document enrichment using semantic tags for effective XML retrieval

ABSTRACT

Using XML to mark up document contents with user-defined and self descriptive terms makes XML technology as one of the most widely used technology for information representation and exchanges over the Internet. As a result many documents are now represented and stored as XML documents on the web. Therefore, there is the need to develop precise, efficient and user-friendly search techniques. The existing systems that support Content Only (CO) queries can be categorized into three. The Lowest Common Ancestor (LCA)-based, Query structuring systems and document Structure based systems. The answers return by first group of systems are either irrelevant to user search intention or may not be meaningful or informative enough because of the restriction on the choice of the root node. The other group requires mostly the existence of data scheme for its query conversion which is not always available or complex and fast evolving. Most of the existing systems put their emphases on query side. In this paper, we focus on document side instead of query side. Our approach exploits document structure; we enriched Wikipedia XML documents text with annotated semantic tags presence in the document. The effect of enriching elements' text content is investigated through three retrieval experiments for which only the text content of document collection differ. The results of the experiments revealed that enriching elements' text content with the semantic tags could improve the effectiveness of CO queries.

Keyword: Content-Only Query (CO); Content and Structure Only Query (CAS); XML retrieval; Annotated semantic tag