Associative Query for Multi-version
Web Documents

Yu ZHANG    Kang ZHANG
Department of Computing, Division of ICS
{yzhang, kang@ics.mq.edu.au}

With the growing importance of the World Wide Web [1] as a truly global
distributed system for information sharing, there has been a significant increase in the
creation and modification of web documents. In order to produce and modify variants of
web sites efficiently, approaches to the creation of multi-version web sites which contain
different versions of web documents and allow all the different versions to share one
generic source page have been developed. Meanwhile, as more and more retrievals for
documents of interest are being done by searching the World Wide Web, more precise
queries are required so as to adequately extract the web-objects from millions of web
pages. Current practice of searching the World Wide Web is to use the existing
commercial search tools such as AltaVista [2], WebCrawler [3], OpenText [4], etc, or
search services such as Bigbook [5], Yahoo [6], etc. The search engines deployed by both
search tools and search service provide keywords indexing and some content indexing.
The resources that the search engines index are the registered URL databases. For a web
site where multi-version codes [7][8] share one generic source code, the URL that is
registered to the search tools can only refer to one version of the web site. Consequently
those conventional query methodologies are not adaptable to multi-version sites. For
instance, a web designer may intend to create a site that provides information such as
news, analysis, company introductions, historic and real time prices, on line ordering, etc
of each of the main stock markets of the world. For easy creation and modification, a
multi-version site in which each of the stock market source pages(a version) shares a
generic stock market information page would be the most appropriate. As one specific
version is only revoked at the time when clients are trying to obtain some
information in that specific stock market. Therefore, if the information that clients are
trying to retrieve resides on the specific version of the stock market source page,
querying through the existing search tools will definitely be unsuccessful. Another
example that illustrates the limitation of the existing searching tools is when they are
used in querying multiple language sites. Although most of the search tools provide
language options for clients to retrieve documents written in the language they favour,
search engines cannot return the URLs in which the language used to create the web
documents differs from the language selected for retrieving URLs with search tools. Thus it can be seen that the existing search tools are only available for one version sites. Furthermore, these search tools cannot support the queries that combine different versions of a site.

This paper proposes an associative querying approach for information retrieval on multi-version web sites. In a multi-version web site, a generic source page on the web acts as a template for various versions of HTML documents. The elements in the generic source page can be interpreted to point to an appropriate version. Querying multi-version sites in the approach combines both page generation and information retrieval. In order to facilitate this retrieval, we categorize objects in the actual HTML page generated by generic source code, which are called intra web page, into three hierarchical categories: root, children and nodes. The title of the page is known as the root. The children category includes links, includes, headings. Nodes are disjoingable objects such as graphic files with .gif, JPEG, MPEG, etc audio files with .au, real, integer, html file, java applets. A hierarchically structured tree is created for the intra web page according to the categorization during the querying procedures. The two integer version labels, V and M, representing the current version and modification in the generic source clauses are also extracted and assigned as attributes to the root object in the tree. In the query sentences that use our defined SQL-like language, variables of version labels are assigned with integers referring to specific versions in the where condition clause. The appropriate version of the page is invoked according to the label value and an intra web page is generated by the generic source page upon receipt of queries from a client. Operation of conditions such as logic and are supported in the query. The appropriate version of the site can be found by matching the label value in the where clause with the version attribute value in the root of the tree, so that more than one version of an intra web page tree can be simultaneously searched by the select operation in the query. For querying objects which need to cooperate with different versions, we create virtual version tables (VVT). Values of the VVT attributes are assigned according to version labels and keywords in the query. Primary keys are set up for the VVTs. Thus complex queries can be achieved by combining different VVTs. The associative query is being implemented in Java. The Java applet running on the web browser provides a visual interface to assist clients in specifying the queries as well as selecting versions. This approach integrates searching and browsing for the multimedia objects in multi-version sites.

Keywords: multi-version web, search engines, SQL-like query, virtual version table

Reference