Information Retrieval and the Virtual Document

Carolyn Watters
Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

After many years building the foundation of the digital library, information retrieval has emerged from its roots in bibliographic retrieval to establish its role in hypertext and Web applications. We are now in the midst of a "seismic" shift of digital content and context from the global repository [i.e., *memex* (Bush, 1945)] through *docuverse* (Nelson, 1981) to the global resource. From a repository of archived documents with embedded links we have seen the emergence of CGI, JavaScript, Java, ActiveX, and XML, all tuned to providing dynamic and application oriented use of information resources. The digital library is merging with the digital workplace and the digital marketplace.

Many of the "documents" retrieved are virtual documents. That is they are created on the way to the user rather than copied from a digital repository. This has a significant impact on the notion of information storage and retrieval. What does it mean to search nonrepository data? How is information authenticated or referenced when it comes from different sources and processes? Who owns it? Although the wide availability of Web resources makes this problem obvious, it is not restricted to the Web.

Virtual Documents

A virtual document is simply a document for which no persistent state exists and for which some or all of each instance is generated at run time. We define an electronic document as both content and links associated with that document. A virtual document can then be pages, Java applets, or application results, and may or may not have associated links. The content may be defined by tags, a template, a program, a database query, or by some application. Virtual documents have grown out of a need for interactivity and individualization of documents, particularly on the Web.

The paradigm of the Web has quickly shifted our expectations for information from that of retrieval of electronic copies of documents from a large repository of relatively static information to manipulation of a large collection of information resources. Some of these resources are documents and some of these resources are processes that create documents. In addition, the role of user is shifting from reader to active participant and author. Users expect hypertext functionality to be available with digital documents: to be able to make comments and annotations, to be able to initiate discussion, and to be able to add content and links while reading, both individually and collaboratively.

Categories of Virtual Documents

Virtual documents can be categorized, roughly, on how they are created, such as: using templates, from computations, as composites, and as metadata.

A simple category of virtual documents results from the use of *templates* in which much of the content is inserted at run time. For example, reports can be generated from a standard template in which the structure is in persistent storage and the actual content is pulled from a database. Documents can be generated that include *computational* results and visualizations based on current results or user interaction, such as the MathResource (www.MathResources.com) interactive math dictionary. *Composite* documents can be generated by putting together content from multiple sources at run time and presenting this to the user as a single unit, such as we see for personalized electronic news editions, such as PointCast (www.pointcast.com). *Metadata* may be generated on the fly by *extraction* and summarization programs to produce virtual documents that may or may not be stored and may or may not be reproducible, depending on the nature of the underlying data. Furthermore, structural standards, such as the proposed XML (www.w3.org), may deliver different representations to different users from the same document.

Concerns and Research Issues

The emergence of virtual documents reveals some very interesting information retrieval problems.

Search—How do you search for virtual documents? What is the domain? Will the document exist by the time the user requests it?

Revisiting—Users have an expectation that documents found one day will be available on a subsequent search. The notion of *bookmark* does not apply to virtual documents in its normal simplistic way. Bookmarks need enough information to recreate the document as it was. Users then need to be able to go forward and backward in time through changes to that virtual document.

Authentication—Who is responsible for the quality of the contents of a virtual document where components may come from a variety of sources and/or processes?

Reference—How do authors cite virtual documents?

Version—Version control has long been a concern of Information Retrieval research and is now a central issue for management of virtual documents.

Annotation—The roles of user of information and the supplier of information are merging. Readers expect to be able to add data, such as, comments, annotations, paths, and links, as well as content, while they are reading.

Summary

The Web has not only increased the scale of information retrieval systems and applications but also introduced a new variation of the notion of document. Basic research is required to provide the same level of understanding and measures of effectiveness and efficiency to virtual documents as has been achieved for persistent documents.

References

Bush, V. (July, 1945). As we may think. Atlantic Monthly, pp. 101–108.

Nelson, T.H. (1981). Literary machines. Swathmore, PA.

www.MathResources.com.

www.pointcast.com.

www.w3.org.