Hypertext Databases and Data Mining*

Soumen Chakrabarti
Indian Institute of Technology Bombay
`soumen@cse.iitb.ernet.in`

The volume of unstructured text and hypertext data far exceeds that of structured data. Text and hypertext are used for digital libraries, product catalogs, reviews, newsgroups, medical reports, customer service reports, and the like. Currently measured in billions of dollars, the worldwide internet activity is expected to reach a trillion dollars by 2002. Database researchers have kept some cautious distance from this action. The goal of this tutorial is to expose database researchers to text and hypertext information retrieval (IR) and mining systems, and to discuss emerging issues in the overlapping areas of databases, hypertext, and data mining.

Keyword indices: The workhorse of text search is the so called "inverted index" data structure. Inverting a document collection involves building a mapping from each term in the collection to the set of documents containing the term, with details such as offset in the document. This enables boolean queries like `socks AND network AND NOT shoes`. Most keyword indexing systems also support phrase search, such as `"inverted index"`, and NEAR (proximity) queries, such as `warehouse NEAR mining`.

Vector space model and relevance ranking: Large responses must be ordered so that documents likely to fulfill the user's information need are ranked highly. To define similarity between a query and a document, they are represented as vectors where each dimension corresponds to a term. Suitable weighting can be used for each term. The cosine of the angle between the query and document vectors is usually used as similarity.

Similarity, clustering, and collaborative filtering: A limitation of the vector space model is that indirect evidence of similarity is overlooked. Latent Semantic Indexing (LSI) transforms the vector space into a lower dimensional space so that similar terms such as `car` and `auto` are mapped to similar vectors, rather than being orthogonal as in the original space. This enhances query capability beyond exact keyword match. An improved measure of document similarity also enables clustering them into related sub-collections, thereby constructing topical hierarchies. Just as terms induce similarity between documents (and vice versa), the relation between entities (movies, songs, books) and people liking them can be analyzed to cluster both the people and entities, thereby enabling collaborative recommendation of additional material.

Relevance feedback and supervised learning: Search systems can significantly improve their relevance ranking, if the user can give even a little feedback, say by marking some of the top ranking documents as relevant or irrelevant, thereby generating a two-class learning problem. More general learning programs can be trained to route documents automatically to appropriate nodes in a topic hierarchy such as Yahoo! Unlike classifiers for structured data, text classifiers must handle very high-dimensional (say 50,000) and noisy data; they must automatically extract those dimensions that are best for classification, and construct statistical topic models. Scaling to large taxonomies and collections is important not only in its own right, but also for improved accuracy.

Exploiting hyperlinks for better learning: Judging the topic of a hypertext document in isolation, without regard to its link neighborhood, gives poor accuracy. There is information in links, but it is noisy. A page about 'cardiology' may cite other pages on cardiology, but also possibly a page on 'swimming'. Such 'sociology' of citations between topics can also be learnt. Integrating this knowledge into text classifiers using a general model of citations results in significant improvements in classification accuracy.

Social network analysis for hypertext: Hypertext is a kind of *social network*. Social networks formed between people by co-authoring, citing, and advising have been extensively researched to find authoritative nodes that have high *prestige*. The prestige of a social network node may be recursively modeled as the sum of the prestige of nodes that cite it. Ordering web query results by prestige, as in the Google search engine, improves the searching experience. Another measure from social network is *reflected prestige*, as evident in topical *hubs* of resource links that are often found on the Web. Hubs can be found by a mutual recursion involving prestige and reflected prestige.

Distributed resource discovery: Given the social structure of the web and the tendency of linked communities to form spontaneously, it is possible to mine the Web to obtain a coherent view of a topical section of it. Care is needed to crawl the web in this goal-directed fashion, starting from a handful of relevant pages. The crawl frontier must be continually pruned to eliminate irrelevant links and paths, and good topical hubs detected for preferential expansion.

*The printed version will not show the many informative hyperlinks to be found in the PDF version available through `http://www.cs.berkeley.edu/~soumen`.