Semantic Indexing

Ratchata Peachavanish

During the past years, there has been an extraordinary growth in the amount of digital information available to common end-users. Factors contributed to this growth include the world-wide proliferation of the Internet, the economical cost of digitizing information contents into computerized forms, and a general increase in computer literacy and accessibility to a large and growing number of people around the world. For instance, the Web has made information about almost any imaginable topic available to anyone with an Internet access; digital cameras have allow ordinary consumers to create large libraries of digital images; various audio/video encoding techniques have enabled music and films to be digitized, stored, and distributed using information technology infrastructures; even common textual documents have extensively been digitized thanks to improvements in various technologies.

Because of the sheer quantity and diversity of digital documents available to end-users, mechanism for their effective and efficient retrieval is of paramount importance. One crucial aspect of this mechanism is *indexing*, which serves to allow documents to be located quickly. In the field of Information Retrieval, an *index* for textual documents traditionally contains selected terms (a vocabulary), where each term indicates the locations where it occurs [2]. For example, suppose a user wishes to find all documents containing the term "laptop". Without an index, the system must examine each and every available document in the repository to determine whether it contains the term. With an index, however, the system simply searches through the index data structure for "laptop" to identify and locate the documents containing it.

Traditionally, indexing is a manual task done by professional *indexers* where they have the flexibility in choosing which terms to use in the indexes. However, the large volume of digitized text documents has necessitated the use of automated indexing by computer systems. In automatic indexing, algorithms are used to determine which terms to use in the indexes. For instance, a content-bearing term that occurs frequently within a document is more likely to be important in that document and should be included as an index term.

While using purely lexical approach to indexing is adequate in most situations, there are significant limitations. For example, suppose a user wants to do a search on portable computers, he would go on to the system and issue a query to search for documents containing the term "notebook". The problem is that if the index is strictly keyword-based, then documents containing the term "laptop" would not be included in the search results. This is a fundamental problem of *synonymy* where two different words describe the same concept. To make matter worse, documents related to writing stationery (paper notebook) would also be included in the search results even though they are not relevant to the user's need. This is a fundamental problem of *polysemy* where the same word has multiple *conceptual* meanings.

When the user issues the query "notebook", what he actually wants is not documents containing the word "notebook"; rather, he wants documents *pertaining to* or *about* the *concept* of notebook computers. The inability for the system to understand user's

intention leads to low-quality search results. As a consequence, the user must rely on manual search strategies obtained through experience to improve search quality. For example, he may issue multiple queries using synonyms or he may use exclusion terms to filter irrelevant search results.

Hence, in order to improve search quality and make searching an easier task for users, there needs to be an integration of *semantics* (human meaning) into the process of document indexing. One simple approach is to use a *thesaurus*, which contains a list of terms and their corresponding synonyms, during the indexing process. For example, "laptop" is a synonym to "notebook" so they are treated as the same concept in the index. Thus, a query on "notebook" would also return documents containing the term "laptop". However, this approach does not really solve the problem because, based on keywords alone, the system still does not know whether the user actually means a computer notebook or a writing stationery notebook. It is quite evident that *contextual* information is crucial in determining what the user means by "notebook".

The problem of determining the intended meaning of users' queries is a difficult one and has many facets to it. How does the system determine the context of a query? Similarly, how does the user specify the query context to the system? The question essentially becomes: how do the system and the user agree that "notebook" is a computer and not a writing stationery? Would it help if the query is modified to "notebook computer"? How does the system know then, that "notebook" is related to "computer", and hence related to "laptop"? One straightforward approach is to restrict the search domain to computer hardware and make explicit the relationships between "notebook", "laptop", and "computer". This approach makes use of *formal ontology*, where the domain of computer hardware is described by controlled vocabulary specifying that "notebook" and "laptop" are both "computers".

To see another approach to the problem, consider how the system determines that a document is *semantically similar* to another document (or to the concept specified in the query). In other words, how to index documents using their semantic contents rather than simply using their lexical features? An effective method, called *Latent Semantic Analysis* or *Latent Semantic Indexing* (LSA/LSI), has been proposed [1] and is now a standard algorithm in Information Retrieval. LSI is a statistical analysis method applied to a large set of textual documents to determine semantic similarity among them. A system with LSI would be able to retrieve documents with semantic similarity to the query concept even if the query term does not appear anywhere in the documents. For example, a search in a news database for "Saddam Hussein" would return documents on the Gulf War, oil embargo, and Iraq, even if those documents do not contain the term "Saddam Hussein" anywhere [5]. It is interesting that LSI does not use any human-constructed dictionary or knowledge base on domain concepts like ontologies [3], it accomplishes its task solely through statistical data mining.

Although the discussion so far has focused on text-based documents, many digital documents today and in the near future are *multimedia* in nature. Images, audio, video, even executable programs are considered documents for the purpose of indexing, searching, and retrieval. Textual information which describes those documents is generally called *metadata* and forms the basis of the index. For example, song files may be described and indexed using a musical genre ontology (e.g., "Jazz", "Classical", "Rock"). For a more advanced scenario, consider a TV on-

demand system that stores video segments of all football matches. A semantically-indexed system would be able to answer a user query like, “Give me the segments where the player Ronaldo appears and a goal is scored” [4].

A more advanced and difficult approach to semantic indexing of multimedia documents is through automated content analysis. This approach involves automatic extraction of semantic contents from non-text documents. For example, image processing techniques for object recognition may be used to identify real-world objects contained in the image (e.g., a car, a person). The identity of objects would then be used to index the image documents.

To conclude, the future of document indexing will involved the use of semantic contents to improve both the search quality and the ease-of-use of the systems. Technologies, such as XML, RDF, OWL, and MPEG-7 are now being developed and used both in research and real systems. Many opportunities are now present in the quest of realizing the vision of semantic-based multimedia document retrieval.

References

1. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41* (6). 391-407.
2. Korfhage, R.R. *Information Storage and Retrieval*. Wiley, 1997.
3. Landauer, T.K., Foltz, P.W. and Laham, D. Introduction to Latent Semantic Analysis. *Discourse Processes* (25). 259-284.
4. Tsinaraki, C., Polydoros, P., Kazasis, F. and Christodoulakis, S. Ontology-Based Semantic Indexing for MPEG-7 and TV-Anytime Audiovisual Content. *Multimedia Tools and Applications*, *26* (3). 299-325.
5. Yu, C., Cuadrado, J., Ceglowski, M. and Payne, S. Patterns in Unstructured Data Discovery, Aggregation, and Visualization, National Institute for Technology and Liberal Education.