Dealing with Uncertainty in Structured and Unstructured Information Integration

Raghu Krishnapuram
IBM India Research Lab
Block I, IIT, Hauz Khas
New Delhi 110016, INDIA
kraghura@in.ibm.com

Knowledge management covers a wide range of activities that are required to create knowledge assets that can be used to improve the organization's productivity, reduce costs, aid innovation, monitor the competition, track trends, and make business decisions. The information required to create the knowledge assets (such as information on products, technologies, people and competitors) typically reside on heterogeneous information sources such as online databases, the intranet, various document collections and the Web. The source of the information can either be structured (e.g. online databases), unstructured (e.g. text files, postscript files and e-mail), or semi-structured (e.g. HTML or XML documents). While some of the useful knowledge is captured in the form of structured data, much of it remains in the form of unstructured data.

According to recent estimates, the amount of unstructured data residing in enterprises is doubling every three months, and a very large percentage of business is conducted based on unstructured information. For example, information on sales, customers, competitors, products, suppliers and people is typically stored as unstructured information. Thus, there is a need for technologies that would facilitate the integration of traditional structured data sources and new generation unstructured information sources for information gathering, business process management or decision-critical business intelligence applications. For example, in customer relationship management (CRM) and biomedical domains, structured data is quite limited, and much of the interesting information lies in the form of unstructured data waiting to be analyzed and exploited. Therefore, the next big leap in knowledge management can only occur when there is a greater integration of structured and unstructured data.

Extracting useful information from heterogeneous unstructured information poses several challenges. This is because we cannot know with certainty that a particular piece of information we have exacted out of unstructured text is correct. For example, in a CRM database of an automobile company, a record pertaining to a complaint might contain the sentence: "Customer talked to Kevin last. Her Viking makes a strange noise whenever she drives uphill". We may conclude the following: "The customer owns a car model called Viking. The vehicle has an engine-related problem. The customer had already called at least once and talked to a person called Kevin". However, there is an uncertainty associated with each of these conclusions. Thus, when information of this type is extracted and stored (e.g. in a structured form), the associated uncertainty also needs to be recorded. There can also be uncertainty at the record or page level. This can arise when there is a question about the credibility, reliability, relevance, or authenticity of the record. Another form of uncertainty arises when there can be multiple records or reports pertaining to the same incident. In many business intelligence applications that require the capability to generate aggregate reports and perform 'slicing and dicing,' we would need to establish whether two or more similar reports are describing the same event, and if they are, combine the information extracted from these reports. If this is not done, the generated reports will overestimate the results. For example, in a CRM scenario involving automobiles, if we need to know the total number of instances of brake related problems by geography, we would obtain a wrong result if reports that describe the same problem are not combined. The combining process will typically require resolving incomplete and conflicting information, as well as computing the overall uncertainty of each of the items which is a function of the uncertainties associated with the same items in the original reports. The relevance or reliability of each of the reports needs to be taken into account while performing this operation.

Apart from information extraction, information search is also a serious issue when the data is is unstructured and heterogeneous. If we need to perform a search on a corpus that contains both structured and unstructured data, traditional information retrieval techniques cannot be used. The situation becomes even more complex when meta data pertaining to the unstructured documents needs to be generated automatically.

In this talk, we will present an overview of the issues involved in integrating structured and unstructured information and describe a few efforts that address them.