Text Metadata Mining: Exploring its Potential

Padmini Srinivasan

School of Library and Information Science
The University of Iowa
Iowa City, IA, 52242
padmini-srinivasan@uiowa.edu

The importance of metadata in digital libraries underlies the development of schemas such as Dublin Core and frameworks for interoperability such as RDF. Metadata offer important access points for information retrieval. Indeed the need to organize for retrieval has been a key motivating factor in developing the Library of Congress Subject Headings and the Medical Subject Headings (MeSH), two examples of seasoned metadata (classification) schemes. These controlled vocabularies serve as important bridges between the vocabularies of authors and those of end users.

Our position is that the utility of metadata goes beyond that of document and record retrieval. We suggest that when metadata schemes are sufficiently sophisticated these may be used to discover new knowledge, or in other words, to mine text collections. Statistical patterns on the distribution of individual metadata across the text collection or appropriate subsets can be revealing. The same can be said of comparisons made across metadata distributions. Finally, such distributions may be used to study phenomenon external to the text collection yet related to it conceptually. For example we may mine a text collection of company financial reports for the relative distribution of metadata referring to nonfinanancial performance measures (such as customer satisfaction and employee morale). This distribution may then be studied in relation to actual financial performance figures reported by these companies.

In this talk we outline several applications of text metadata mining that we are presently exploring. The goal is to promote discussion on similar applications of metadata in digital libraries that move beyond the realm of information retrieval. There are also implications on the design of metadata schemes. For example, we seek to demonstrate that there is merit to building information rich schemes.

Metadata Mining Applications

The first application that we present is one where we studied the global distribution of disease research in the MEDLINE database [2]. This was done by exploiting the fact that in addition to terms representing diseases, MeSH includes terms representing geographical

area (countries, cities etc.). We selected a group of 19 diseases including for example, breast cancer and cholera. For each disease we first mined a profile from MEDLINE representing the geographic distribution of research on that disease. We then compared this distribution to a second distribution representing the prevalence of the disease. Prevalence data were obtained from the World Health Organization web site. An example conclusion made from this study is that as expected, the two distributions are correlated. What is interesting is that this correlation does not typically hold for those diseases that are most prevalent in low income nations. In the same research we also looked at temporal trends in disease research and compared these trends across countries.

The second application that we present is one where MeSH metadata have been used to discover connections between bibliographically distinct topics. Two topics are considered bibiliographically distinct if their relevant documents do not overlap. That is, there are no documents that discuss the two topics together. In the mid 1980's Swanson and Smalheiser proposed what is now called a 'closed discovery process' that allows a researcher to explore potential connections between bibliographically distinct topics (see [3] for example). Thus if a researcher has an intuition that two disconnected topics may indeed be related in some indirect way, then the closed discovery process may be explored to test this intuition. Connections identified may form the basis for further research. This type of discovery process applied to the biosciences domain has recently been referred to as 'conceptual biology' [1]. We have recently implemented this closed discovery process using MeSH metadata and have succeeded in replicating the many discoveries made by Swanson and Smalheiser.

Examples of other applications that we are currently pursuing are (1) analysis of microarray data using metadata mining in MEDLINE and (2) mining financial reports submitted by companies. In the microarray work we are studying groups of genes that have similar expression patterns by examining their 'metadata profiles' in MEDLINE. In the finance domain we are looking at aspects such as the kinds of predictions made by companies and their relationship to actual performance data.

We present these applications as illustrative examples of the kinds of benefits that can accrue from exploiting metadata assigned to records. As mentioned before the aim is to promote discussion on other metadata mining applications and metadata applications in general.

References

[1] Blagosklonny, M. V., & Pardee, A. B. Unearthing the gems. (2002). Nature, 416, 373.

[2] Srinivasan, P., & Wedemeyer, M. (2003). Mining Concept Profiles with the Vector Model or Where on Earth are Diseases being Studied? In: Proceedings of Text Mining Workshop. Third SIAM International Conference on Data Mining. San Francisco, CA.

[3] Swanson, D.R., & Smalheiser, N.R. (1997). An interactive system for finding complementary literatures. Artificial Intelligence, 91, 183-203.