IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB8931

## Chapter XII

# Mining Free Text for Structure

Vladimir A. Kulyukin
Utah State University, USA

Robin Burke
DePaul University, USA

## ABSTRACT

*Knowledge of the structural organization of information in documents can be of significant assistance to information systems that use documents as their knowledge bases. In particular, such knowledge is of use to information retrieval systems that retrieve documents in response to user queries. This chapter presents an approach to mining free-text documents for structure that is qualitative in nature. It complements the statistical and machine-learning approaches, insomuch as the structural organization of information in documents is discovered through mining free text for content markers left behind by document writers. The ultimate objective is to find scalable data mining (DM) solutions for free-text documents in exchange for modest knowledge-engineering requirements. The problem of mining free text for structure is addressed in the context of finding structural components of files of frequently asked questions (FAQs) associated with many USENET newsgroups. The chapter describes a system that mines FAQs for structural components. The chapter concludes with an outline of possible future trends in the structural mining of free text.*

## INTRODUCTION

When the manager of a mutual fund sits down to write an update of the fund's prospectus, he does not start his job from scratch. He knows what the fund's sharehold-

ers expect to see in the document and arranges the information accordingly. An inventor, ready to register his idea with the Patent and Trademark Office of the U.S. Department of Commerce, writes it up in accordance with the rules specifying the format of patent submissions. A researcher who wants to submit a paper to a scientific conference must be aware of the format specifications set up by the conference committee. Each of these examples suggests that domains of human activity that produce numerous documents are likely to have standards specifying how information must be presented in them.

Such standards, or presentation patterns, are a matter of economic necessity; documents whose visual structure reflects their logical organization are much easier to mine for information than unconstrained text. The ability to find the needed content in the document by taking advantage of its structural organization allows the readers to deal with large quantities of data efficiently. For example, when one needs to find out if a person's name is mentioned in a book, one does not have to read it from cover to cover; going to the index section is a more sensible solution.

Knowledge of the structural organization of information in documents[1] can be of significant assistance to information systems that use documents as their knowledge bases. In particular, such knowledge is of use to information retrieval systems (Salton & McGill, 1983) that retrieve documents in response to user queries. For example, an information retrieval system can match a query against the structural components of a document, e.g., sections of an article, and make a retrieval decision based on some combination of matches. More generally, knowledge of the structural organization of information in documents makes it easier to mine those documents for information.

The advent of the World Wide Web and the Internet have resulted in the creation of millions of documents containing unstructured, structured, and semi-structured data. Consequently, research on the automated discovery of structural organization of information in documents has come to the forefront of both information retrieval and natural language processing (Freitag, 1998; Hammer, Garcia-Molina, Cho, Aranha, & Crespo, 1997; Hsu & Chang, 1999; Jacquemin & Bush, 2000; Kushmerick, Weld, & Doorenbos, 1997). Most researchers adhere to numerical approaches of machine learning and information retrieval. Information retrieval approaches view texts as sets of terms, each of which exhibits some form of frequency distribution. By tracking the frequency distributions of terms, one can attempt to partition the document into smaller chunks, thus claiming to have discovered a structural organization of information in a given document. Machine-learning approaches view texts as objects with features whose combinations can be automatically learned by inductive methods.

Powerful as they are, these approaches to mining documents for structure have two major drawbacks. First, statistical computations are based on the idea of statistical significance (Moore & McCabe, 1993). Achieving statistical significance requires large quantities of data. The same is true for machine-learning approaches that require large training sets to reliably learn needed regularities. Since many documents are small in size, the reliable discovery of their structural components using numerical methods alone is problematic. Second, numerical approaches ignore the fact that document writers leave explicit markers of content structure in document texts. The presence of these markers in document texts helps the reader digest the information contained in the document. If these markers are ignored, document texts become much harder to navigate and understand.