Structural Text Mining

Vladimir A. Kulyukin
Utah State University, USA

John A. Nicholson
Utah State University, USA

INTRODUCTION

The advent of the World Wide Web has resulted in the creation of millions of documents containing unstructured, structured and semi-structured data. Consequently, research on structural text mining has come to the forefront of both information retrieval and natural language processing (Cardie, 1997; Freitag, 1998; Hammer, Garcia-Molina, Cho, Aranha, & Crespo, 1997; Hearst, 1992; Hsu & Chang, 1999; Jacquemin & Bush, 2000; Kushmerick, Weld, & Doorenbos, 1997). Knowledge of how information is organized and structured in texts can be of significant assistance to information systems that use documents as their knowledge bases (Appelt, 1999). In particular, such knowledge is of use to information retrieval systems (Salton & McGill, 1983) that retrieve documents in response to user queries and to systems that use texts to construct domain-specific ontologies or thesauri (Ruge, 1997).

BACKGROUND

Structural mining of texts consists of two related tasks: the task of partitioning text into components, for example, topics, sentences, terms, and so forth; and the task of finding relations among found components, for example, term and topic associations. Texts can be divided into three broad categories: free, structured, and semi-structured.

Free texts do not give the computer many road maps to the information they contain. To discover a road map in a free text requires a certain amount of data mining through parsing, statistical analysis, and/or machine learning. Novels and newspaper and journal articles are examples of free texts. Structured texts organize their content according to well understood road maps. Relational databases are structured texts where all of the relations between textual entities, that is, records, are known and can be readily obtained through well-defined queries. Semi-structured texts offer more structure than free texts but less than structured ones. HTML pages are semi-structured texts. While they offer a standard set of tags that point to the structural organization of information in them, they do not specify the types of information that the tags label or the relations among these types.

ISSUES IN TEXT MINING

The three fundamental problems in structural text mining are:

- Text Segmentation;
- Automatic Ontology (Thesaurus) Construction; and
- Information Extraction.

Text segmentation is a process of partitioning free texts into segments of content. The underlying assumption is that texts are intellectual artifacts that consist of words related to each other semantically in a number of complex ways (Bookstein, Kulyukin, Raita, & Nicholson, 2003). The intellectual process of producing texts incidentally leaves behind simple statistical regularities. Capturing those regularities through statistical analysis allows one to arrive at the structural organization of information in the texts.

The two most prominent approaches to text segmentation are statistical and qualitative. Statistical approaches to text segmentation (Hearst, 1997) first parse texts to identify primitive components, for example, sentences, and then combine those primitive components into larger segments by defining various similarity measures between pairs of components. For example, if components are represented as vectors of terms each of which is assigned a specific weight (1 or 0 in the basic case), the similarity between two components can be computed through a range of vector metrics: dot product, cosine of the angle between the vectors, a hamming distance, and so forth. Powerful as they are, statistical approaches to text segmentation have two drawbacks. First, statistical computations are based on the idea of statistical significance. Achieving statistical significance requires large quantities of data. Since many documents are small in size, the reliable discovery of their structural components using numerical methods alone is not always appropriate.

Second, numerical approaches frequently ignore the fact that text writers leave explicit markers of content structure in document texts. The presence of these markers in texts helps the reader digest the information contained in the texts. If these markers are ignored, the texts become much harder to navigate and understand. These intuitions are at the heart of qualitative approaches to text segmentation (Kulyukin & Burke, 2003). In these approaches, the structural organization of information in texts is discovered through mining free text for content markers left behind by text writers. For example, police crime reports and scientific journal papers have well defined structures that can be fruitfully mined for information. The ultimate objective of qualitative approaches is to find scalable data mining solutions for free text documents in exchange for modest knowledge engineering requirements.

Research in automatic thesaurus construction investigates ways to extract thesaurus relations from texts. A thesaurus is a set of terms plus a set of relations among them. Automatic thesaurus construction complements manual thesaurus construction, which, as the argument goes, is expensive in terms of expert time and effort and cannot respond in a timely manner to rapid changes in scientific fields. Automatic thesaurus construction is usually collection dependent, that is, it is done on a specific text collection at hand. Approaches to automatic thesaurus construction include statistical analyses of term co-occurrence data (Dagan, Lee, & Pereira, 1999), syntactic patterns used to extract semantic relations among terms (Grefenstette, 1994; Hearst, 1992), and ensemble methods that combine different information extraction techniques and rank their outputs by their utility to the task at hand, for example, manual query expansion during retrieval (Curran, 2002). Evaluation of automatic thesauri, that is, evaluation of the authenticity of found relations and their utility, remains a major challenge.

Information extraction goes hand in hand with automatic thesaurus construction. In information extraction, the problem of mining text for structure is cast in terms of extracting sets of facts, for example, a specific statistic in a crime report, and/or rules, for example, how to find a victim's name and age in crime reports, from the texts at hand. In particular, many researchers are concerned with the problem of extracting database-like structures from Web pages, in effect reverse-engineering the process of database-backed Web page generation.

Hammer et al. (1997) present a configurable tool for extracting semi-structured data from a set of HTML pages, given a declarative specification of where the data of interest are located. The machine learning approach to this problem has been labeled "wrapper induction" (Kushmerick et al., 1997). The extraction procedure, or wrapper, for a specific resource is learned from a set of representative pages from that resource.

Hsu and Chang (1999) describe a formalism to represent information extractors as Finite-State Transducers (FST). A finite-state transducer is a variation of a finite-state automaton (Hopcroft & Ullman, 1979). The input document is assumed to be tokenized before it is given to a finite-state transducer. The authors distinguish two types of transducers: single-pass and multi-pass. A single-pass transducer scans the text only once. A multi-pass transducer scans the text multiple times, each time focusing only on a specific type of object to extract. The ultimate goal of this approach is the automated construction of extractors from a set of training examples. However, the reported empirical evaluations assume that the space of possible graph structures, that is, finite-state automata, is restricted or that the structure is given to the learner in advance.

Freitag (1998) casts information extraction as a relational learning problem. Relational learning represents hypotheses as sets of if-then rules. Because sets of if-then statements can be viewed as programs in a logic programming language, such as PROLOG, relational learning is often called Inductive Logic Programming (Mitchell, 1997). Freitag describes a general purpose top-down relational learning algorithm for information extraction called "SRV". SRV takes as input a set of token-oriented features that encode most of the domain-specific information. For example, they may encode a standard set of questions that can be asked of someone's home page, such as the owner's name, affiliation, e-mail, and so forth. An answer to each question is assumed to be a text fragment from that home page. Thus, the algorithm solves the problem of finding the best unbroken fragment of text that answers a question from a given set of questions. The SRV algorithm makes no assumption about document structure. Instead, structural information is supplied as input to the system.

Jacquemin and Bush (2000) present a tool for the acquisition of named entities, for example, names of companies, from textual sources. The authors' approach combines lexical indices with formatting instructions. Lexical indices are discourse markers and formatting instructions are HTML tags. The system includes three shallow parsers for mining HTML texts for specific structures such as lists, enumerations, and anchors. The named entities are extracted from the found structures by analyzing discourse markers and HTML tags.

FUTURE TRENDS

The issues discussed in this article are likely to remain major challenges in structural text mining. The push to automation will bring an ever greater emphasis on the