Text Mining: Promises And Challenges

Article · February 2002

Source: CiteSeer

Ah-Hwee Tan
Singapore Mamagement University

TEXT MINING:

PROMISES AND CHALLENGES

Ah-Hwee Tan

Kent Ridge Digital Labs
21 Heng Mui Keng Terrace, Singapore 119613
Homepage: http://textmining.krdl.org.sg
Email: *ahhwee@krdl.org.sg*

ABSTRACT

Text mining, also known as knowledge discovery from text, and document information mining, refers to the process of extracting interesting patterns from very large text corpus for the purposes of discovering knowledge. Text mining is an interdisciplinary field involving information retrieval, text understanding, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining. Regarded by many as the next wave of knowledge discovery, text mining has a very high commercial value. This talk presents a general framework for text mining, consisting of two stages: text refining that transforms unstructured text documents into an intermediate form; and knowledge distillation that deduces patterns or knowledge from the intermediate form. We then survey the state-of-the-art text mining approaches, products, and applications by aligning them based on the text refining and knowledge distillation functions as well as the intermediate form that they adopt. In conclusion, we highlight the upcoming challenges of text mining and the opportunities it offers.

1. INTRODUCTION

Text mining, also known as text data mining (Hearst, 1997) or knowledge discovery from textual databases (Feldman & Dagan, 1995), is an emerging technology for analyzing large collections of unstructured documents for the purposes of extracting interesting and non-trivial patterns or knowledge. It can be envisaged as a leap from data mining or knowledge discovery from (structured) databases (Fayyad *et al*., 1996; Simoudis, 1996).

As the most natural form of storing and exchanging information is written words, text mining has a very high commercial potential. In fact, a recent study indicated that 80% of a company's information was contained in text documents, such as emails, memos, customer correspondence, and reports. The ability to distil this untapped source of information provides substantial competitive advantages for a company to succeed in the era of a knowledge-based economy. There are many possible applications of text mining technology. We briefly highlight a few below.

1. *Customer profile analysis*, e.g., mining incoming emails for customers' complaint and feedback.

2. *Patent analysis*, e.g., analyzing patent databases for major technology players, trends, and opportunities.

3. *Information dissemination*, e.g., organizing and summarizing trade news and reports for personalized information services.

4. *Company resource planning*, e.g., mining a company's reports and correspondences for activities, status, and problems reported.

Text mining is a challenging task as it involves dealing with text data that are inherently unstructured and fuzzy. The field is interdisciplinary, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining. To facilitate discussion, this article presents a general framework for text mining consisting of two components: *Text refining* that transforms free-form text documents into an *intermediate form;* and *knowledge distillation* that deduces patterns or knowledge from the intermediate form. We then use the proposed framework to study and align the state-of-the-art text mining products and applications based on the text refining and knowledge distillation functions as well as the intermediate form that they adopt.

The rest of this paper is organized as follows. Section 2 presents the proposed text mining framework that bridges the gap between text mining and data mining. Section 3 gives an overview of the current text mining products and applications in the light of the proposed framework. The final section discusses open problems and research directions.

2. A TEXT MINING FRAMEWORK

Text mining can be visualized as consisting of two phases: *Text refining* that transforms free-form text documents into a chosen *intermediate form*, and *knowledge distillation* that deduces patterns or knowledge from the intermediate form (Tan, 1999).

Intermediate form (IF) can be *semi-structured* such as the conceptual graph representation, or *structured* such as the relational data representation. Intermediate form can be *document-based* wherein each entity represents a document, or *concept-based* wherein each entity represents an object or concept of interests in a specific domain. Mining a document-based IF deduces patterns and relationship across documents. Document clustering/visualization and categorization are examples of mining from a document-based IF. Mining a concept-based IF derives pattern and relationship across objects or concepts. Data mining operations, such as predictive modeling and associative discovery, fall into this category. A document-based IF can be transformed into a concept-based IF by realigning or extracting the relevant information according to the objects of interests in a specific domain. It follows that document-based IF could be domain-independent whereas concept-based IF is always domain-dependent.

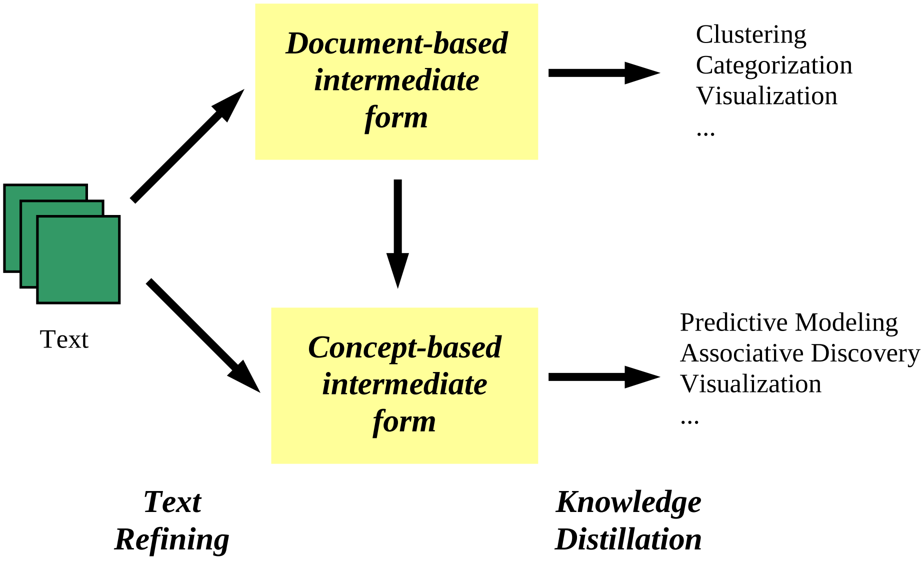

Figure 1: A text mining framework. *Text refining* converts unstructured text documents into an *intermediate form (IF)*. IF can be *document-based* or *concept-based. Knowledge distillation* from a *document-based IF* deduces patterns or knowledge across documents. A *document-based IF* can be projected onto a *concept-based IF* by extracting object information relevant to a domain. *Knowledge distillation* from a *concept-based IF* deduces patterns or knowledge across objects or concepts.

For example, given a set of news articles, text refining first converts each document into a document-based IF. One can then perform knowledge distillation on the document-based IF for the purpose of organizing the articles, according to their content, for visualization and navigation purposes. For knowledge discovery in a specific domain, the document-based IF of the news articles can be projected onto a concept-based IF depending on the task requirement. For example, one can extract information related to "*company*" from the document-based IF and form a company database. Knowledge distillation can then be performed on the company database (company-based IF) to derive company-related knowledge.

3. PRODUCTS AND APPLICATIONS

Table 1 shows an illustrative list of text mining products and applications based on the text refining and knowledge distillation functions as well as the intermediate form adopted. The text mining products/applications can be roughly organized into two groups. One group focuses on document organization, visualization, and navigation. The other group focuses on text analysis functions, notably, information retrieval, information extraction, categorization, and summarization. While we see that most text mining systems provide natural language processing (NLP) functions, few, if any, have integrated data mining functions for knowledge distillation across concepts or objects.

Company/ Organization	Product/ Application	Text Refining Functions	Intermediate Form	Knowledge Distillation Functions
Cartia	ThemeScape		Document-based	Clustering, visualization
Canis	cMap		Document-based word histograms	Clustering, visualization
IBM/ Synthema	Technology Watch		Document-based	Clustering, visualization
Inxight	VisControls		Document-based Hyperbolic tree	Visualization
Semio Corp	SemioMap		Concept-based	Visualization
Knowledge Discovery System	Concept Explorer	Info retrieval	Concept-based	
Inxight	Linguist	Info retrieval, text analysis, summarization	Document-based	
IBM	iMiner	Info retrieval, summarization	Document-based	Clustering, categorization
TextWise	DR_LINK CINDOR CHESS	Info retrieval Info extraction	Concept-based	
Cambio	Data Junction	Info extraction	Concept-based	
Megaputer	TextAnalyst	Info retrieval, summarization	Document-based semantic net	Classification

Table 1: A list of selected text mining products and applications based on the text refining and knowledge distillation functions as well as the intermediate form adopted.

3.1. Document exploration tools

Document exploration tools organize documents based on their content and provide an environment for a user to navigate and browse in a document or concept space. A popular approach is to perform clustering on the documents based on their similarities in content and present the groups or clusters of the documents in certain graphical representation. There are a good number of text mining products that fall into this category. The following list is by no means exhaustive but should be sufficient to illustrate the variety of the representation schemes available.

Cartia's ThemeScape is an enterprise information mapping application that presents clusters of documents in a landscape representation. Canis's cMap is a document clustering and visualization tool based on Self-Organizing Map. IBM’s Technology Watch, developed jointly with Synthema in Italy, is a text mining application in the scientific domain. It performs document clustering plus visualization in the form of maps for patent databases and technical publications. Inxight also offers a visualization tool, known as VizControls, that performs value-added post-processing of search results by clustering the documents into groups and displaying based on a hyperbolic tree representation. Semio Corp's SemioMap employs a three-dimensional graphical interface that maps the links between concepts in the document collection. Note that SemioMap is concept-based in the sense that it explores the relationships between concepts whereas most other visualization tools are document-based.

3.2. Document analysis tools

Document analysis tools analyze the content of the documents and discover relationships between concepts or entities described in the documents. They are mainly based on natural language processing techniques, including text analysis, text categorization, information extraction, and summarization.

Knowledge Discovery System's Concept Explorer is a visual search tool that helps to find precisely related content on the web. It "learns" relationships between words and phrases automatically from sample documents and visually guides you to construct searches. Inxight's LinguistX is another document retrieval tool with some text analysis and summarization capabilities. IBM’s Intelligent Miner is probably one of the most comprehensive text mining products around. It offers a set of text analysis tools, including a feature extraction tool, a set of clustering tools, a summarization tool, and a categorization tool. Also incorporated are the IBM’s text search engine, NetQuestion Solution and the IBM web crawler package. TextWise, an R&D company based in Syracuse University, offers various text mining products. DR-LINK is an information retrieval system based on automatic concept expansion. CINDOR is its cross lingual version. CHESS is a text analysis and information extraction tool. Also an information extraction tool is the Data Junction's Cambio, which extracts data in the form of relational attributes from text. Megaputer's TextAnalyst uses a semantic net representation of documents and performs automated indexing, topic assignment, text abstraction, and semantic search.

4. OPEN PROBLEMS AND FUTURE DIRECTIONS

Despite the great potential and the mushrooming of text mining products, there are technical issues to be overcome before text mining becomes a main stream technology.

4.1. Intermediate form

Intermediate forms with varying degrees of complexity are suitable for different mining purposes. For a fine-grain domain-specific knowledge discovery task, it is necessary to perform semantic analysis to derive a sufficiently rich representation to capture the relationship between the objects or concepts described in the documents. However, semantic analysis methods are computationally expensive and often operate in the order of a few words per second. It remains a challenge to see how semantic analysis can be made much more efficient and scalable for very large text corpora.

4.2. Multilingual text refining

Whereas data mining is largely language independent, text mining involves a significant language component. Multilingual text mining is the area we expect to see a lot of activities in the next few years due to the substantial competitive advantages and the huge commercial potential that one can obtain through mining in languages other than English. Languages that are of particular interests include European languages and Asian languages, in particular Japanese and Chinese. As each language has a different syntactic structure and requires specialized semantic interpretation, a systematic approach for bringing in language modeling is inevitable and will form an essential part of multilingual text mining.

4.3. Domain knowledge integration

Current text mining systems do not make use of domain knowledge. We expect it to be an integral component of the future text mining tools. Domain knowledge is useful in orientating and focusing attention so as to improve the text parsing efficiency and to help to derive a more compact representation. Domain knowledge also plays a major role in knowledge distillation tasks. In a classification or predictive modeling task, for example, domain knowledge helps to improve learning/mining efficiency as well as the quality of the learned model (or mined knowledge) (Tan, 1997). It is also interesting to explore how a user’s knowledge can be used to initialize a system’s knowledge structure and make the discovered knowledge more interpretable.

4.4. Personalized autonomous mining

Another important dimension of research is to make text mining tools more user friendly. Current text mining products/applications are designed for trained knowledge specialists. Future text mining tools, as part of the knowledge management systems, should be readily usable by technical users as well as management executives. There have been some efforts in developing systems that interpret natural language queries and perform appropriate mining operations automatically. Text mining tools could also embedded in intelligent personal assistants (Tan & Teo, 1998). Under the *agent* paradigm, a personal miner would learn a user's profile, conduct text mining operations automatically, and forward information without requiring an explicit request from the user.

REFERENCES

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), "From data mining to knowledge discovery: An overview", in U. Fayyad *et al.* (eds.) *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, Mass., 1-36.

Feldman, R. & Dagan, I. (1995), "Knowledge discovery in textual databases (KDT)", in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Montreal, Canada, August 20-21, AAAI Press, 112-117.

Hearst, M.A. (1997), "Text data mining: Issues, techniques, and the relationship to information access", Presentation notes for UW/MS workshop on data mining, July 1997.

Simoudis, E. (1996), "Reality check for data mining", *IEEE Expert*, **11**(5).

Tan, A.-H. (1997), "Cascade ARTMAP: Integrating neural computation and symbolic knowledge processing", *IEEE Transactions on Neural Networks*, **8**(2), 237-250.

Tan, A.-H. & Teo, C. (1998), "Learning user profiles for personalized information dissemination", in *Proceedings, International Joint Conference on Neural Networks (IJCNN'98)*, Alaska, 183-188.

Tan, A.-H. (1999), "Text Mining: The state of the art and the challenges", in *Proceedings, PAKDD'99 workshop on Knowledge Discovery from Advanced Databases*, Beijing, April, 1999.