Text Mining:
The state of the art and the challenges

Ah-Hwee Tan

Kent Ridge Digital Labs
21 Heng Mui Keng Terrace
Singapore 119613
Email: ahhwee@krdl.org.sg

Abstract

Text mining, also known as text data mining or knowledge discovery from textual databases, refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. Regarded by many as the next wave of knowledge discovery, text mining has very high commercial values. Last count reveals that there are more than ten high-tech companies offering products for text mining. Has text mining evolved so rapidly to become a mature field? This article attempts to shed some lights to the question. We first present a text mining framework consisting of two components: *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 products/applications and align 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 [3] or knowledge discovery from textual databases [2], refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. It can be viewed as an extension of data mining or knowledge discovery from (structured) databases [1,4].

As the most natural form of storing information is *text*, text mining is believed to have a commercial potential higher than that of data mining. In fact, a recent study indicated that 80% of a company's information is contained in text documents. Text mining, however, is also a much more complex task (than data mining) as it involves dealing with text data that are inherently unstructured and fuzzy. Text mining is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining.

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 Framework of text mining

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. 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 is usually domain-independent and concept-based IF is domain-dependent.

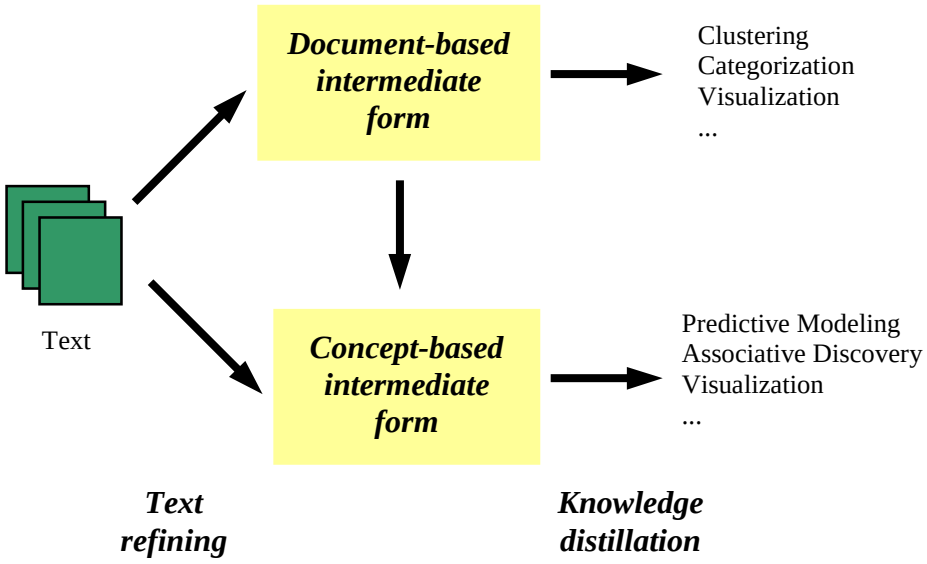

Figure 1: A text mining framework. *Text refining* converts unstructured text documents into an *intermediate form (IF)*. IF can be *document-based* or *concept-based. Knowledge distillation* from a *document-based IF* deduces patterns or knowledge across documents. A *document-based IF* can be projected 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. A Survey of text mining products

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. One group of products focuses on document organization, visualization, and navigation. Another group focuses on text analysis functions, notably, information retrieval, information extraction, categorization, and summarization. While we see that most text mining systems are based on natural language processing, none of the products has integrated data mining functions for knowledge distillation across concepts or objects.

3.1. Document visualization

There are a good number of text mining products that fall into this category. The general approach is to organize the documents based on their similarities and present the groups or clusters of the documents in certain graphical representation. The following list is by no means exhaustive but is 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 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. Text analysis and understanding

The second group of the text mining products is 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.

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 summary of illustrative text mining products and applications based on the text refining and knowledge distillation functions as well as the intermediate form adopted.

4. Open problems and future directions

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. It is essential to develop text refining algorithms, that process multilingual text documents and produce language-independent intermediate forms. While most text mining tools focus on processing English documents, mining from documents in other languages allows access to previously untapped information and offers a new host of opportunities.

4.3. Domain knowledge integration

Domain knowledge, not catered for by any current text mining tools, could play an important role in text mining. Specifically, domain knowledge can be used as early as in the text refining stage. It is interesting to explore how one can take advantage of domain information to improve parsing efficiency and derive a more compact intermediate form.

Domain knowledge could also play a part in knowledge distillation. In a classification or predictive modeling task, domain knowledge helps to improve learning/mining efficiency as well as the quality of the learned model (or mined knowledge) [5]. It is also interesting to explore how a user's knowledge can be used to initialize a knowledge structure and make the discovered knowledge more interpretable.

4.4. Personalized autonomous mining

Current text mining products and applications are still tools 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 automatically perform the appropriate mining operations. Text mining tools could also appear in the form of intelligent personal assistants [6]. 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

[1] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., MIT Press, Cambridge, Mass., 1-36.

[2] 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.

[3] 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.

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

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

[6] 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.