South African Journal of Information Management
[DoE accredited]

Peer Reviewed Article

Vol.6(4) December 2004

Differentiating between data-mining and text-mining terminology

J.H. Kroeze
Department of Informatics, University of Pretoria, Pretoria, South Africa
jan.kroeze@up.ac.za

M.C. Matthee
Department of Informatics, University of Pretoria, Pretoria, South Africa
machdel.matthee@up.ac.za

T.J.D. Bothma
Department of Information Science, University of Pretoria, Pretoria, South Africa
theo.bothma@up.ac.za

Contents

1. Introduction
2. Plethora of definitions of text mining
3. Clarifications by Hearst
4. Clarifying Hearst
5. New proposal: *intelligent* text mining
6. Conclusion
7. References

Key words: Text mining; data mining; knowledge creation; knowledge discovery in databases (KDD); information retrieval (IR)

1 Introduction

When a new discipline emerges, it usually takes some time and a great deal of academic discussion before concepts and terms become standardized. Text mining is one such new discipline. In a groundbreaking article, *Untangling text data mining*, Hearst (1999) tackled the problem of clarifying text-mining concepts and terminology. This article is aimed at building on Hearst's ideas by pointing out some inconsistencies and inaccuracies, and suggesting an improved and extended categorization of data-mining and text-mining approaches.

Until recently, computer scientists and information system specialists concentrated on the discovery of knowledge from structured, numerical databases and data warehouses.

However, much, if not the majority, of available business data are captured in text files that are not overtly structured, for example memoranda and journal articles that are available electronically. Bibliographic databases may contain overtly structured fields, such as author, title, date and publisher, as well as free text, such as an abstract or even full text. The discovery of knowledge from database sources containing free text is called 'text mining'.

Web mining is a wider field than text mining because the World-Wide Web also contains other elements, such as multimedia and e-commerce data. As the Web continues to expand rapidly, Web mining becomes more and more important (and increasingly difficult). Although text mining and Web mining are two different fields, it must be borne in mind that a great deal of the content on the Web is text-based. 'It is estimated that 80% of the world's on-line content is based on text' (Chen 2001:18). Therefore, text mining should also form an important part of Web mining.

This article is a conceptual study. A brief overview of the problems regarding text-mining concepts and approaches are given. This is followed by a summary and critical discussion of Hearst's attempt to clarify the terminology. To conclude, an improved and expanded differentiation of data-mining and text-mining concepts and methods are proposed. The different kinds of data and text mining are also descriptively named.

top

2 Plethora of definitions of text mining

The nominal compound *text mining* suggests that it is either the *discovery* of texts or the *exploration* of texts in search of valuable, yet hidden, information. However, a few definitions are quoted below to indicate that things are not quite that simple:

- Text mining 'performs various searching functions, linguistic analysis and categorizations'. Search engines focus on text search, especially directed at 'text-based web content' (Chen 2001:5,9).
- 'Text mining is the study and practice of extracting information from text using the principles of computational linguistics' (Sullivan 2000).
- Text mining is 'to prospect for nuggets of new knowledge in the mountains of text which have become accessible to computer-based research thanks to the information and internetworking revolution' (Lucas 1999/2000:1).
- Text mining is 'a way to examine a collection of documents and discover information not contained in any individual document in the collection' (Lucas 1999/2000:1).
- Text mining as exploratory data analysis is a method of (building and) using software systems to *support* researchers in deriving new and relevant information from large text collections. It is a partially automated process in which the researcher is still involved, interacting with the system. 'The interaction is a cycle in which the system suggests hypotheses and strategies for investigating these hypotheses, and the user either uses or ignores these suggestions and decides on the next move' (Hearst 1999:6–7). Similar to the idea of academic hypotheses is the identification of business ideas: 'Text mining is ideal … to … inspect changes in the market, or to identify ideas to pursue' (Biggs 2000).
- Text mining is the establishing of 'previously unknown and unsuspected relations of features in a (textual) data base' (Albrecht and Merkl 1998).
- 'We define text mining to be data mining on text data. Text mining is all about extracting patterns and associations previously unknown from large text databases' (Thuraisingham 1999:167; compare Nasukawa and Nagano 2001:967 for a similar definition).
- Zorn, Emanoil, Marshall and Panek (1999:28) regard text mining as a knowledge

> creation tool: 'Text mining offers powerful possibilities for creating knowledge and relevance out of the massive amounts of unstructured information available on the Internet and corporate intranets.'

Like these text-mining definitions, the different text-mining products that are available also vary widely: 'Given the immaturity of the text-mining tool market, each of these tools takes a slightly different track' (Biggs 2000).

Thus, it is not that easy to decide what qualifies as text mining and what does not. Is it an advanced form of information retrieval, or is it something else? Most scholars agree that text mining is a branch or a sibling of data mining (e.g. compare Nasukawa and Nagano 2001:969; Chen 2001:5). Therefore, it will be useful to define and characterize data mining before Hearst's attempt to clarify the text-mining concepts is discussed.

Data mining is a step in the process of knowledge discovery from data (KDD). KDD concerns the acquisition of new, important, valid and useful knowledge from data. Berson and Smith (1997:341–342) maintain that: 'In the case of large databases sometimes users are asking the impossible: “Tell me something I didn't know but would like to know.”'

This type of knowledge is *what you don't know you don't know*. This is the most difficult type of knowledge to discover. It is easier to discover *what you know you don't know*, or even *what you don't know you know*. Data-mining tools promise to discover these types of knowledge and to transform them into *what you know you know*, by measuring how strong, unexpected and often-encountered the associations between elements in the data are (compare Westphal and Blaxton 1998: 62–65).

Data mining is a proactive process that automatically searches data for new relationships and anomalies on which to base business decisions in order to gain competitive advantage (Rob and Coronel 2002:654). Although data mining might always require some interaction between the investigator and the data-mining tool, it may be considered as an automatic process because 'data-mining tools automatically search the data for anomalies and possible relationships, thereby identifying problems that have not yet been identified by the end user', while mere data analysis 'relies on the end users to define the problem, select the data, and initiate the appropriate data analyses to generate the information that helps model and solve problems they uncovered' (Rob and Coronel 2002:654).

The assumption that this proactive characteristic is the essence of data mining and that text mining is a branch of data mining, implies that text mining should also be proactive, that is, the automatic finding of anomalies and relationships in texts that indicate trends or problems not yet discovered by anyone. Therefore, it is not surprising that Biggs (2000) mentions proactive business decisions as one of the benefits of text mining: 'Moreover, you can potentially step ahead of your competition by having more complete information to proactively make better-informed decisions.'

top

3 Clarifications by Hearst

In an attempt to clarify the concepts and terminology regarding this new field, Hearst (1999) wrote a paper, *Untangling text data mining*, that differentiates between information retrieval and text mining, which Hearst calls 'text data mining' (TDM). This section provides a short summary of her paper. However, because any summary or paraphrase contains interpretation, some conscious, and maybe even unconscious, assumptions were made, especially regarding a few unclear matters in Hearst's paper, as will emerge in the section that follows.

According to Hearst (1999:3–5), text data mining is sometimes confused with information retrieval. She maintains that this should not be so if one uses the level of novelty of the information being searched for as a parameter. In information retrieval, the desired information is already known (at least to the author of an existing text), and the problem is to locate it. Text data mining, however, should strive to find new information. According to Hearst, the essence of real text data mining is the discovery of 'heretofore-unknown information' or the finding of (new) answers to old questions. Therefore, Hearst does not regard the following techniques as examples of text data mining, but as advanced information retrieval, etc.:

- 'Text clustering to create thematic overviews of text collections'
- Automatic generation of 'term associations to aid in query expansion'
- 'Co-citation analysis to find general topics within a text-collection'
- Text categorization, which she defines as merely classifying the content of texts according to 'a set of pre-defined labels'
- The compilation of a summary of information 'that is already known'.

Hearst (1999:3) metaphorically refers to information retrieval as 'looking for needles in a needle stack', that is, finding relevant information between other valuable, but irrelevant, pieces of information. She puts it on the same level as database queries from numerical databases (see the right-hand column of Tables 1 and 2).

To bring text data mining in line with numerical data mining, the information in a textual database (which may be on the Web) should be treated 'as a large knowledge base from which we can extract new, never-before encountered information' (Hearst 1999:3–4). Hearst defines data mining as the '(semi) automated discovery of trends and patterns across very large datasets, usually for the purpose of decision making'. It is not 'discovering new factoids within … inventory databases'. She regards 'corpus-based computational linguistics' as similar to standard (numerical) data mining – statistics are computed over large text collections to discover useful linguistic patterns (see left-hand column of Tables 1 and 2). She also refers to the connection between corpus-based computational linguistics and natural language processing: computational linguistics improves language analysis and text analysis itself, but does not tell us anything about the outside world. Examples of computational linguistics are 'part-of-speech tagging, word sense disambiguation, and bilingual dictionary creation'.

Hearst (1999:3–5) proposes a third category, finding novel nuggets in 'otherwise worthless rock', that is, finding new, relevant information between otherwise worthless data (see the middle column of Tables 1 and 2). There is no comparable form of numerical data mining but, when it comes to text data, Hearst calls this approach 'real text data mining'. Hearst is of the opinion that although numerical data mining cannot be compared to finding novel nuggets of information, real text data mining can.

Table 1 Classification of data mining and text data mining approaches by Hearst (1999:5)

	Finding patterns	**Finding nuggets**	
		Novel	**Non-novel**
Non-textual data	Standard data mining	?	Database queries
Textual data	Computational linguistics	Real TDM	Information retrieval

A few examples of mining-for-novel-nuggets in text or real text data mining are (Hearst 1999:4–5):

- 'Augmentation of existing lexical structures', for example discovering lexical and syntactic features in texts ('data-mining-as ore-extraction')
- Using text category assignments (an element within a metadata set) to find unexpected patterns among text articles, for example 'distributions of commodities' among countries
- Discovering new themes or trends among texts, for example new news threads (also using metadata).

However, Hearst (1999:5–7) is not sure whether the mining of metadata should be regarded as standard data mining or (real) text data mining. She then gives two examples of exploratory data analysis as more pure forms of real text data mining:

- 'Using text to form hypotheses about disease', for example a magnesium deficiency may cause migraine
- 'Using text to uncover social impact', for example 'the technology industry relies more heavily than ever on government-sponsored research results'.

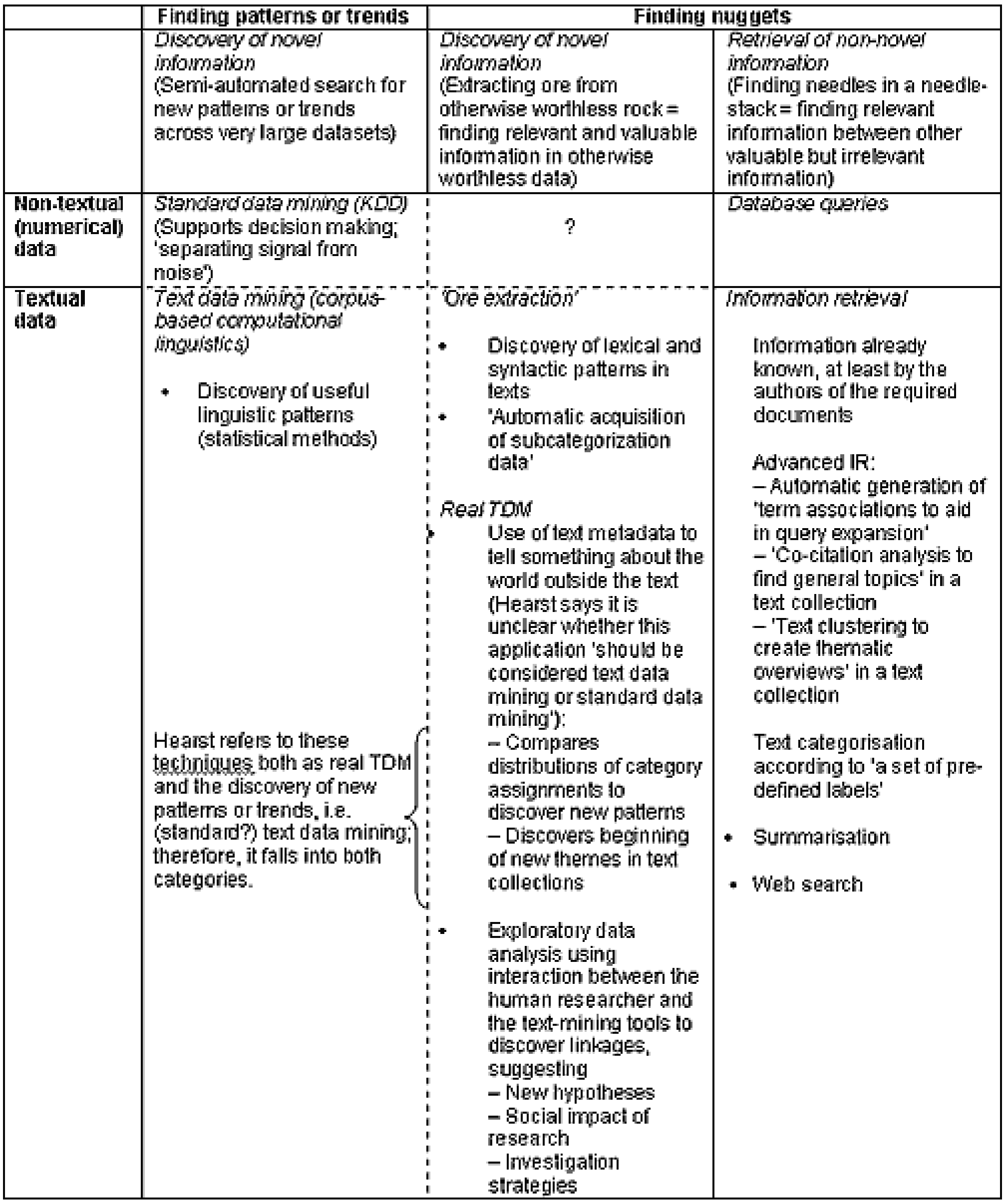
To summarize her views, Hearst (1999:5) classifies and names different data-mining and text-mining approaches, using the parameters of finding patterns *vs* finding nuggets, the novelty-level of the nuggets, and non-textual data *vs* textual data (Table 1).

As a practical example of the application of real text data mining, Hearst (1999:7) refers to the so-called LINDI project. Tools that provide 'support for issuing sequences of queries' are used, as well as 'tightly coupled statistical and visualization tools for the examination of associations among concepts that co-occur within the retrieved documents' to suggest hypotheses and strategies which 'the user either uses or ignores'.

Hearst's 'real text data mining' does not automate human intelligent behaviour (1999:8): 'I suggest that to make progress we do not need fully artificial intelligent text analysis; rather, a mixture of computationally-driven and user-guided analysis may open the door to exciting new results.'

Before a critical discussion of Hearst's paper can be given, it is necessary to come to a clear understanding of her views. Therefore, her paper has been summarized by expanding her own table (Table 1) to include other information and judgements in the paper (Table 2).

Table 2 A summary of Hearst’s paper, *Untangling text data mining* (1999)



	Finding patterns or trends	Finding nuggets	
	Discovery of novel information (Semi-automated search for new patterns or trends across very large datasets)	*Discovery of novel information* (Extracting ore from otherwise worthless rock = finding relevant and valuable information in otherwise worthless data)	*Retrieval of non-novel information* (Finding needles in a needle-stack = finding relevant information between other valuable but irrelevant information)
Non-textual (numerical) data	*Standard data mining (KDD)* (Supports decision making; 'separating signal from noise')	?	*Database queries*
Textual data	*Text data mining (corpus-based computational linguistics)* • Discovery of useful linguistic patterns (statistical methods) Hearst refers to these techniques both as real TDM and the discovery of new patterns or trends, i.e. (standard?) text data mining; therefore, it falls into both categories.	*'Ore extraction'* • Discovery of lexical and syntactic patterns in texts • 'Automatic acquisition of subcategorization data' *Real TDM* • Use of text metadata to tell something about the world outside the text (Hearst says it is unclear whether this application 'should be considered text data mining or standard data mining'): – Compares distributions of category assignments to discover new patterns – Discovers beginning of new themes in text collections • Exploratory data analysis using interaction between the human researcher and the text-mining tools to discover linkages, suggesting – New hypotheses – Social impact of research – Investigation strategies	*Information retrieval* Information already known, at least by the authors of the required documents Advanced IR: – Automatic generation of 'term associations to aid in query expansion' – 'Co-citation analysis to find general topics' in a text collection – 'Text clustering to create thematic overviews' in a text collection Text categorisation according to 'a set of pre-defined labels' • Summarisation • Web search

top

4 Clarifiying Hearst

Hearst's paper is innovative and groundbreaking because it distinguishes between different types of data mining and text mining (*vs* database queries and information retrieval). Her use of the parameters of novel *vs* non-novel information and finding nuggets *vs* finding patterns or trends is very useful. There are, however, some problems in her paper that are discussed below. This article is aimed at clearing up these problems by clarifying Hearst's parameters, rearranging her classifications and expanding upon both.

Hearst subdivides the 'finding nuggets' column into a 'non-novel' and a 'novel' column. The finding of non-novel nuggets is metaphorically called 'finding needles in a needle-stack', while the finding of novel nuggets is compared to finding valuable information nuggets in otherwise worthless rock (that is, finding needles in a haystack). However, the 'nuggets' and searched-for 'needles' already exist and are already known by someone. The problem is to locate them. Finding them cannot be regarded as novel information. In other words, there is no such thing as novel information *nuggets*; it is a contradiction in terms. With regard to the novelty of the information to be found, there is in principle no difference between finding needles in a haystack and finding needles in a needle-stack. Therefore, these two columns should be merged and the process could then be called *non-novel investigation*. Hearst refers to the separation of signal from noise as an aim of standard data mining. However, this is actually a synonymous definition of finding nuggets or non-novel investigation.

Hearst is uncertain about the position of metadata mining (is it standard data mining or real text data mining?). Adding a separate horizontal category for metadata (between numerical data and textual data) could have solved this problem (Table 3).

In addition to these problems, the following contradictions and obscurities in her paper necessitate further conceptual research about the different approaches to data and text mining:

- In her introduction Hearst says that 'in the case of text, it can be interesting to take the mining-for-nuggets metaphor seriously', but does not explain why text data mining should be different from numerical data mining in this regard.
- It is not clear why 'identifying lexico-syntactic patterns' is regarded as 'data-mining-as-ore extraction' (middle column), while computational linguistics is regarded as the discovery of useful linguistic patterns (first column). In both cases the patterns already exist, but still have to be discovered. There seems to be no or little difference between part-of-speech tagging and the identification of syntactic patterns with regard to the novelty of the information or the discovery of patterns.
- Hearst places some techniques in two categories: the comparison of distributions of text category assignments to discover *new patterns or trends*, as well as the discovery of *new themes* in text collections, is called *real TDM*. However, the discovery of patterns and trends (or themes) implies that it should be put in the same column as standard data mining and computational linguistics, that is (standard?) text data mining.
- Both computational linguistics (finding patterns) and real TDM use statistical methods, implying that they are similar.
- The fact that no form of data mining exists that is comparable to real text data mining suggests that there may still be a gap, or even a flaw, in the argument.
- Why is the discovery of sub-categorization data regarded as novel information 'nuggets' ('ore-extraction'), but text categorization according to a set of pre-defined labels is not? The sub-categorization data are already present in the texts and can be discovered – this is semi-novel investigation. The classification can also be regarded as semi-novel investigation – although the labels or classes may already be known, the categorization itself is also semi-novel.

Taking a closer look at Hearst's examples of real text data mining (which she classifies as finding novel nuggets) reveals that new information is indeed discovered. Although the lexical and syntactic features in texts themselves, and the patterns, trends or new themes regarding the outside world already exist in the text data, they are as yet unknown and the discovery thereof is new. The same applies to the discovery of linkages that enable exploratory data analysis to take place. This is similar to standard data mining. Thus, Hearst's suggestion that 'the mining-for-nuggets metaphor' should be taken seriously for text mining is not acceptable. All her examples of real text data mining should be moved to the finding patterns or trends column, which implies that 'real text data mining' is actually *standard text (data) mining*. The fact that she puts computational linguistics in both columns (finding patterns and finding novel nuggets) supports this conclusion. Automatic text clustering, generation of term associations and co-citation analysis should also be regarded as semi-novel investigation because the associations and the themes or topics already exist in the text collection but still have to be discovered as new knowledge. This article suggests that this process be called *semi-novel investigation*.

top

5 New proposal: *intelligent* text mining

This gives rise to the question: what then is novel text-mining investigation? If non-novel text-mining investigation is information retrieval, and if semi-novel text-mining investigation is knowledge discovery, then novel text-mining investigation should be knowledge creation, the deliberate process of creating new knowledge that did not exist before and cannot simply be retrieved or discovered. This is a process that is usually done by humans and is very difficult to automate. Hearst also refers to the interaction between the human researcher and the text-mining tools. She does not discuss the use of artificial intelligence (AI) to analyse the discovered patterns and trends, because she is convinced that progress can be made without it. However, because AI simulates human intelligence and behaviour, it may be used to facilitate automatic *novel investigation*. Such an automatic process could be called either *intelligent data mining* (for overtly structured numerical data, that is, strictly formatted numerical and alphanumerical fields in databases) or *intelligent text mining* (for covertly structured data, that is, inherently structured text data). Surprisingly, the term 'intelligent text mining' has not yet become part of information systems jargon.

Intelligent data and text mining should tell us something about the world, outside the data collections themselves (expanding Hearst's words). For example: What do the patterns and trends mean and imply? Which business decisions are prompted by them? How can the linguistic features of text be used to create knowledge about the outside world? How should the patterns or trends that are signalled by distributions of category assignments be described? Is a new theme that is discovered in a text collection valid (does it reflect reality)? How should the hypotheses prompted by linkages that are found be refined and formulated? What investigation strategies do the linkages found imply, and are they feasible and relevant? What social impact do the linkages found suggest? Therefore, the discovery of patterns should be separated from the analysis and interpretation of the patterns. The former is semi-novel investigation and the latter is novel investigation, which will be facilitated either by the interaction between the human researcher and the data-mining or text-mining tools, or by AI. Mack and Hehenberger (2002:S97) regard the automation of 'human-like capabilities for comprehending complicated knowledge structures' as one of the frontiers of 'text-based knowledge discovery'.

Artificial intelligence, and especially natural language processing, plays an important role in intelligent text mining. Natural language processing can be used to discover the inherent structure of free texts. This form is difficult to decipher and therefore Hearst tries to find ways of doing 'real text data mining' without taking the linguistic structures into account. Natural language processing may be used to analyse the underlying linguistic structures, and to build syntactic and semantic representations of the texts. Intelligent text summarization is an example of the use of natural language processing to simulate human reasoning. According to Hovy and Lin (1999), the difference between extracts and abstracts lies in the novelty level of the phrasings. While an extract is a mere collection of verbatim phrases from the original, an abstract interprets and describes the content in other words, requiring topic fusion and text generation. Their SUMMARIST system, for example, combines natural language processing with existing semantic and lexical knowledge sources. However, in the proposed differentiation of data and text-mining approaches below (Table 3), intelligent text summarization will be another example of semi-novel investigation because the knowledge is already known, while the formulation of the summary is new. The use of natural language processing in text mining should be subjected to further scrutiny in a separate research project.

The concepts of data, information and knowledge are also relevant to drawing a distinction between non-novel, semi-novel and novel investigation. Data consist of raw facts that have no intrinsic meaning; they have to be sorted, grouped, analysed and interpreted in order to become information. Information, combined with context and experience, becomes knowledge. Zorn *et al.* (1999:28) say:

'Data is only useful when it can be located and synthesized into information or knowledge, and text mining looks to be the most efficient and effective way to offer this possibility to the Web.'

However, owing to the fluidity of these concepts and terminology, it is not possible to link each of the concepts of data, information and knowledge exclusively to one of the text investigation approaches (non-novel, semi-novel or novel). Yet, it seems that both non-novel and semi-novel investigation are more on the level of data and information (even though semi-novel investigation is called *knowledge* discovery!), while novel investigation is more on the level of knowledge.

A further extension of Hearst's categories could be the definite distinction of (text) metadata as a separate category, in addition to 'non-textual' data (i.e. overtly structured, mainly numerical data) and textual data (covertly structured data). The use of keywords for information retrieval, theme discovery and the comparison and interpretation of the distribution of text category labels can be dealt with in this section. However, there is no fundamental difference between data mining and metadata mining because both deal with overtly structured data.

The discussion above is summarized in a new table (Table 3) to compare the different approaches to data and text mining and related fields. The parameters used for the distinction are the novelty level of the investigation (non-novel, semi-novel and novel) and the level of textual structure (mainly numerical: overtly/visibly structured; text metadata: overtly structured bibliographical fields; textual: covertly/inherently/opaquely structured). In the cells (the intersections of rows and columns), the headings (in italics) refer to techniques while the asterisks refer to the kind of problems that are solved by the techniques.

Table 3 Overview of data and text mining and related fields

	Non-novel investigation – Data/information retrieval (*Finding/retrieving* already existing and known information)	**Semi-novel investigation – Knowledge discovery** (*Discovery* of existing patterns – the patterns/trends already exist in the data, but are yet unknown and the discovery thereof is new)	**Novel investigation –** Knowledge creation (*Creation* of new important knowledge – tells something about the world, outside of the data collection itself)
Numerical data (including strictly formatted alpha-numerical fields) (Overtly structured)	*Database queries* (Uses database operations such as SQL queries) • Retrieves specific	*Standard data mining* (Often uses statistical methods, e.g. link analysis; on-line analytical processing) • Reveals business	*Intelligent data mining* (Uses interaction between investigator and computerized tool; AI)

	(mainly) numerical data	patterns in numerical data	• What do the patterns and trends mean and imply? • Which business decisions are suggested?
Text metadata (Overtly structured bibliographical fields, e.g. author, date, title, publisher, keywords; excluding free-text sections, such as abstracts)	*Information retrieval of metadata* (Uses exact match and best match queries) • Retrieves references to specific documents	*Standard metadata mining* (Uses mainly statistical methods) • Discovers the start of a new theme or trend in a chronological series of documents, based on metadata • Compiles distributions of category assignments	*Intelligent metadata mining* (Uses interaction between investigator and computerised tool; AI) • Compares and interprets distribution of text category labels within subsets of the document collection
Textual data (Inherently/covertly structured)	*Information retrieval of full texts* (Uses exact match and best match queries) • Finds full texts of articles etc.	*Standard text mining* (Uses mainly statistical methods) • Discovers lexical and syntactic features in texts (computational linguistics) • Finds beginning of new themes in text collections • Discovers linkages between entities in and across texts • Identifies term associations and co-citations • Compiles thematic overviews	*Intelligent text mining* (Uses interaction between investigator and computerized tool; AI) • How can the linguistic features of text be used to create knowledge about the outside world? • Does a new theme reflect reality? • How should

		• Groups texts according to inherent characteristics (text- • clustering) • Discovers subcategorization data • Categorizes texts into pre-existing classes • Summarizes text intelligently	the hypotheses prompted by linkages found be refined and formulated? • What are the investigation strategies implied by the linkages found, and are they feasible and relevant? • What is the social impact suggested by the linkages found? • Which business decisions are implied?

top

6 Conclusion

Text mining is a new field and, as a result, concepts and approaches to it still vary considerably. It has become important to differentiate between advanced information retrieval methods and various text-mining approaches. Hearst (1999) used the parameters of novel/non-novel information to draw that distinction. An attempt has been made in this article to expand the scope of the topic from 'real text data mining' (as opposed to information retrieval) to *intelligent* text mining.

The essence of text mining is the discovery or creation of new knowledge from a collection of documents. The new knowledge may be the statistical discovery of new patterns in known data (standard text mining) or it may incorporate AI abilities to interpret the patterns and provide more advanced capabilities such as hypothesis suggestion (intelligent text mining). Artificial intelligence, and especially natural language processing, can be used to simulate human capabilities needed for intelligent text mining. Further research should be conducted into the necessity for natural language processing in order to facilitate intelligent text mining. The parameters of non-novel, semi-novel and novel investigation were used to differentiate between full-text information retrieval, standard text mining and intelligent text mining. The same parameters were also used to differentiate between related processes for numerical data and text metadata. These distinctions may be used as a road map in the evolving fields of data and information retrieval, knowledge discovery and the creation of new knowledge.

top

7 References

Albrecht, R. and Merkl, D. 1998. Knowledge discovery in literature data bases. *Library and Information Services in Astronomy III.* (*ASP conference series* , vol. 153.) [Online]. Available WWW: http://www.stsci.edu/stsci/meetings/lisa3/albrechtr1.html (Accessed 20 August 2002).

Berson, A. and Smith, S.J. 1997. *Data warehousing, data mining, and OLAP*. New York: McGraw-Hill.

Biggs, M. 2000. Resurgent text-mining technology can greatly increase your firm's 'intelligence' factor. *InfoWorld* 11(2):52.

Chen, H. 2001. *Knowledge management systems: a text mining perspective.* Tucson, Arizona: University of Arizona (Knowledge Computing Corporation).

Cornford, T. and Smithson, S. 1996. *Project research in information systems: a student's guide.* Houndmills: Macmillan. (Information system series.)

Halliman, C. 2001. *Business intelligence using smart techniques: environmental scanning using text mining and competitor analysis using scenarios and manual simulation.* Houston, TA: Information Uncover.

Han, J. and Kamber, M. 2001. *Data mining: concepts and techniques.* San Francisco, CA: Morgan Kaufmann.

Hearst, M.A. 1999. Untangling text data mining. In: *Proceedings of ACL'99: the 37 th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, June 20–26 (invited paper). [Online]. Available WWW: http://www.ai.mit.edu/people/jimmylin/papers/Hearst99a.pdf (Accessed 20 August 2002).

Hovy, E. and Lin, C.Y. 1999. Automated text summarization in SUMMARIST. In Mani, I. and Maybury, M.T. (eds.) *Advances in automated text summarization.*MIT Press, MA:81–94. [Online]. Available WWW: http://www.isi.edu/~cyl/ (Accessed 24 June 2003).

Kontos, J., Malagardi, I., Alexandris, C. and Bouligaraki, M. 2000. Greek verb semantic processing for stock market text mining. In *Proceedings of Natural Language Processing: 2nd International Conference, Patras, Greece, June 2000*, edited by D.N. Christodoulakis. Berlin: Springer:395–405. (Lecture notes in artificial intelligence, no. 1835).

Lucas, M. 1999/2000. Mining in textual mountains, an interview with Marti Hearst. *Mappa Mundi Magazine, Trip-M* 5:1–3. [Online]. Available WWW: http://mappa.mundi.net/trip-m/hearst/ (Accessed 20 August 2002).

Mack, R. and Hehenberger, M. 2002. Text-based knowledge discovery: search and mining of life-science documents. *Drug Discovery Today* 7(11) (Suppl.)*:*S89–S98.

Nasukawa, T. and Nagano, T. 2001. Text analysis and knowledge mining system. *IBM Systems Journal* 40(4):967–984.

New Zealand Digital Library, University of Waikato. 2002. *Text mining.* [Online]. Available WWW: http://www.cs.waikato.ac.nz/~nzdl/textmining/ (Accessed 25 April 2003).

Perrin, P. and Petry, F.E. 2003. Extraction and representation of contextual information for knowledge discovery in texts. *Information Sciences* 151:125–152.

Ponelis, S. and Fairer-Wessels, F.A. 1998. Knowledge management: a literature overview. *South African Journal of Library and Information Science* 66(1):1–9.

Rajman, M. and Besançon, R. 1998. Text mining: natural language techniques and text mining applications. In Spaccapietra, S. and Maryanski, F. (eds.) *Data Mining and Reverse Engineering: Searching for Semantics.* London: Chapmann and Hall:50–64.

Rob, P. and Coronel, C. 2002 . *Database systems: design, implementation, and management* , 5 th ed. Boston, MA: Course Technology.

Stair, R.M. and Reynolds, G.W. 2001. *Principles of information systems: a managerial approach* , 5 th ed. Boston, MA: Course Technology.

Sullivan, D. 2000. The need for text mining in business intelligence. *DM Review,* Dec. 2000. [Online]. Available WWW: http://www.dmreview.com/master.cfm.

Sullivan, D. 2001. *Document warehousing and text mining: techniques for improving business operations, marketing, and sales.* New York: John Wiley.

Thuraisingham, B. 1999. *Data mining: technologies, techniques, tools, and trends.* Boca Raton, Florida: CRC Press.

Westphal, C.R. and Blaxton, T. 1998. *Data mining solutions: methods and tools for solving real-world problems.* New York: Wiley.

Zorn, P., Emanoil, M., Marshall, L. and Panek, M. 1999. Mining meets the Web. *Online* 23 (5):17–28.

Disclaimer

Articles published in SAJIM are the opinions of the authors and do not necessarily reflect the opinion of the Editor, Board, Publisher, Webmaster or the Rand Afrikaans University. The user hereby waives any claim he/she/they may have or acquire against the publisher, its suppliers, licensees and sub licensees and indemnifies all said persons from any claims, lawsuits, proceedings, costs, special, incidental, consequential or indirect damages, including damages for loss of profits, loss of business or downtime arising out of or relating to the user's use of the Website.

top

ISSN 1560-683X

Published by InterWord Communications for the Centre for Research in Web-based Applications,
Rand Afrikaans University