

Differentiating Data- and Text-Mining Terminology

JAN H. KROEZE, MACHDEL C. MATTHEE AND THEO J.D. BOTHMA

University of Pretoria

When a new discipline emerges it usually takes some time and lots of academic discussion before concepts and terms get standardised. Such a new discipline is text mining. In a groundbreaking paper, *Untangling text data mining*, Hearst [1999] tackled the problem of clarifying text-mining concepts and terminology. This essay aims to build on Hearst's ideas by pointing out some inconsistencies and suggesting an improved and extended categorisation of data- and text-mining techniques. The essay is a conceptual study. A short overview of the problems regarding text-mining concepts is given. This is followed by a summary and critical discussion of Hearst's attempt to clarify the terminology. The essence of text mining is found to be the discovery or creation of new knowledge from a collection of documents. The parameters of non-novel, semi-novel and novel investigation are used to differentiate between full-text information retrieval, standard text mining and intelligent text mining. The same parameters are 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/information retrieval, knowledge discovery and the creation of new knowledge.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications – *Data mining*; H.2.4 [Database Management]: Systems – *Textual databases*; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.3.6 [Information Storage and Retrieval]: Library Automation – *Large text archives*; I.2.7 [Artificial Intelligence]: Natural Language Processing – *Text analysis*; I.5.2 [Pattern Recognition]: Design Methodology – *Pattern analysis*; I.7.5 [Document and Text Processing]: Document Capture – *Document analysis*

General Terms: Algorithms, Documentation, Languages, Measurement, Theory

Additional Key Words and Phrases: Text data mining, TDM, text mining, text-mining, information retrieval, IR, knowledge creation, knowledge discovery, KDD, metadata, database queries, full-text retrieval; knowledge management

1. INTRODUCTION

When a new discipline emerges it usually takes some time and lots of academic discussion before concepts and terms get standardised. Such a new discipline is text mining.¹ In a groundbreaking paper, *Untangling text data mining*, Hearst [1999] tackled the problem of clarifying text-mining concepts and terminology. This paper aims to build on Hearst's ideas by pointing out some inconsistencies and inaccuracies and suggesting an improved and extended categorisation of data- 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.² However, much, if not the majority, of available business data are captured in text files that are not overtly structured, e.g. memoranda and journal articles that are available electronically.³ 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.⁴ 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 web also contains other elements, such as multimedia and e-commerce data.⁵ As the web continues to expand rapidly web mining becomes more and more important (and more

¹ 'Unlike search engines and data mining that have a longer history and are better understood, text mining is an emerging technical area that is relatively unknown to IT professionals' [Chen 2001:vi.]. Even data mining technology is considered to be still in its infancy [Rob and Coronel 2002:657].

² In this paper the concept of numerical data includes highly structured and limited text fields, often found in databases, also referred to as strings, characters or alphanumerical fields, but excludes free text, such as comments, memos, abstracts and full-text articles.

³ According to Rajman and Besançon [1998:50] 'only a small fraction (5-10%) of the collected data is ever analysed'.

⁴ These mixed databases are sometimes called partially structured databases. This is based on the assumption that free text is unstructured. However, free text is covertly structured.

⁵ According to Thuraisingham [1999:188] a web miner should integrate multimedia databases and data mining techniques: 'One needs to develop tools first to mine multimedia data and then we can focus on developing tools to mine such data on the web.' Multimedia data include text, images, video and audio data.

Author Addresses:

J.H. Kroeze, Department of Informatics, School of IT, University of Pretoria, Pretoria, 0002; jhkroeze@postino.up.ac.za.

M.C. Matthee, Department of Informatics, School of IT, University of Pretoria, Pretoria, 0002; mmatthee@hakuna.up.ac.za.

T.J.D. Bothma, Department of Information Science, School of IT, University of Pretoria, Pretoria, 0002; tbothma@postino.up.ac.za.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, that the copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than SAICSIT or the ACM must be honoured. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee.

© 2003 SAICSIT

and more difficult). Although text mining and web mining are two different fields,⁶ it has to be remembered that a lot of the content on the web is text based. ‘It is estimated that 80% of the world’s online content is based on text’ [Chen 2001:18].⁷ Therefore, text mining should also form an important part of web mining.⁸

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

2. THE MESS OF MINING TEXT

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 will be quoted below to indicate that it is not all 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 how to (build and) use software systems to *support* researchers to derive 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 et al. [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 available text-mining products also vary a lot: ‘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 not. Is it an advanced form of information retrieval, or is something else? Most scholars agree that text mining is a branch or a sibling of data mining [for example, compare Nasukawa and Nagano 2001:969; Chen 2001:5]. Therefore, it will be useful to define and characterise data mining before Hearst’s attempt to clarify the text-mining concepts is discussed.

Data mining is a step in the knowledge discovery from data process (KDD). KDD concerns the acquisition of new, important, valid and useful knowledge from data. Berson and Smith [1997:341–342] say:

‘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 it 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].

⁶ ‘Text mining is about looking for patterns in natural language text ... Web mining is the slightly more general case of looking for patterns in hypertext and often applies graph theoretical approaches to detect and utilise the structure of web sites’ [New Zealand Digital Library 2002].

⁷ According to Zorn et al. [1999:20] 90% of electronically available material is ‘unstructured’.

⁸ According to Han and Kamber [2001:428] ‘text mining has become an increasingly popular and *essential* theme in data mining’ (our italics).

Data mining is a proactive process that automatically searches data for new relationships and anomalies to make 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’ [ibid.].

Assuming 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, i.e. 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.’

3. CLEANING UP 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 (compare the following section).

According to Hearst [1999:3–5] text data mining is sometimes conflated with information retrieval. She rejects this possibility using as parameter the level of novelty of the searched-for information. 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 categorisation, 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 needlestack’, i.e. 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 right hand column of Table 1 and Table 2).

In order 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 Table 1 and Table 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 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’, i.e. finding new, relevant information between otherwise worthless data (see the middle column of Table 1 and Table 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 holds the opinion that, although numerical data mining cannot be compared to finding novel nuggets of information, real text data mining can.

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

Table 1. *Classification of data mining and text data mining approaches by Hearst [1999:5].*

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’, e.g. discovering lexical and syntactic features in texts (‘data-mining-as ore-extraction’)
- Using text category assignments (an element within a metadata⁹ set) to find unexpected patterns among text articles, e.g. ‘distributions of commodities’ among countries
- Discovering new themes or trends among texts, e.g. 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’, e.g., a magnesium deficiency may cause migraine
- ‘Using text to uncover social impact’, e.g., ‘the technology industry relies more heavily than ever on government-sponsored research results’.

To summarise her views, Hearst [1999:5] classifies and names different data- and text-mining approaches using the parameters of finding patterns vs. finding nuggets, the novelty-level of the nuggets, and non-textual data¹⁰ vs. textual data (see 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 summarised by expanding her own table (see Table 1) to include other information and judgements in the paper (see Table 2).¹¹

4. CLEARING UP HEARST

Hearst’s paper is innovative and groundbreaking because it distinguishes between different types of data mining and text mining¹² (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 will be discussed below. This paper aims to clear up these problems by clarifying Hearst’s parameters, rearranging her classifications and extending on both.

Hearst sub-divides 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 (= finding needles in a haystack). However, the ‘nuggets’ and searched-for ‘needles’ already exist and they are already known by someone, and 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

⁹ Metadata are data about other data. Text metadata are data about documents containing free text, such as author, title and keywords.

¹⁰ Hearst’s category of non-textual data refers to highly structured, mainly numerical, data in databases. (It excludes multimedia data.)

¹¹ Broken lines in Table 2 indicate unclear borders in Hearst’s article regarding the differentiation between standard data mining, computational linguistics and real text data mining (see the critical discussion below).

¹² Text mining is the more generally used term for what Hearst [1999] calls text data mining.

	<i>Finding patterns or trends</i>	<i>Finding nuggets</i>	
	<i>Discovery of novel information</i> (Semi-automated search for new patterns or trends across very large datasets)	<i>Discovery of novel information</i> (Extracting ore from otherwise worthless rock = finding relevant and valuable information in otherwise worthless data)	<i>Retrieval of non-novel information</i> (Finding needles in a needle-stack = finding relevant information between other valuable but irrelevant information)
<i>Non-textual (numerical) data</i>	<i>Standard data mining (KDD)</i> (Supports decision making; 'separating signal from noise')	?	<i>Database queries</i>
<i>Textual data</i>	<i>Text data mining (corpus-based computational linguistics)</i> * 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 in both categories.	<i>'Ore extraction'</i> * Discovery of lexical and syntactic patterns in texts * 'Automatic acquisition of subcategorization data' <i>Real TDM</i> * Use of text metadata to tell something about the world outside the text (Hearst says it is unclear if 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	<i>Information retrieval</i> * 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

Table 2. A summary of Hearst's paper, *Untangling text data mining [1999]*

difference between finding needles in a haystack or finding needles in a needle-stack. Therefore, these two columns should be merged and the process can be called *non-novel investigation*.¹³ 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 (see Table 3).

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

¹³ Novel investigation, semi-novel investigation and non-novel investigation refer to the novelty level of the information retrieved, discovered or created. It does not describe the novelty level of the processes used.

- In her introduction Hearst says that ‘in the case of text, it can be interesting to take the mining-for-nuggets metaphor seriously’, but she does not motivate 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 puts 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, *i.e.* (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 subcategorisation data regarded as novel information ‘nuggets’ (‘ore-extraction’), but text categorisation according to a set of pre-defined labels not? The subcategorisation 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 categorisation 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, the patterns, trends or new themes regarding the outside world already exist in the text data, they are yet unknown and the discovery thereof is new. The same applies to the discovery of linkages that enable exploratory data analysis. This is similar to standard data mining.¹⁴ Thus, Hearst’s suggestion that ‘the mining-for-nuggets metaphor’ should be taken seriously for text mining is not acceptable.¹⁵ 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.¹⁶ This paper suggests that this process be called *semi-novel investigation*.¹⁷

5. A NEW PROPOSAL: INTELLIGENT TEXT MINING

This leaves 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,¹⁸ 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 to analyse the discovered patterns and trends, because she is convinced that progress can be made without it. However, because artificial intelligence (AI) simulates human intelligence and behaviour it may be used to facilitate automatic *novel investigation*.¹⁹ Such an automatic process could be called either *intelligent data mining* (for overtly structured numerical data, *i.e.* strictly formatted numerical and alphanumeric fields

¹⁴ ‘The difference between text mining and information retrieval is analogous to the difference between data mining and database management’ [Thuraisingham 1999:167].

¹⁵ According to Nasukawa and Nagano [2001:969] their approach to text mining ‘is a text version of generalized data mining’. Therefore, if data mining is not regarded as mining for nuggets, neither should text mining be. Text mining should ‘focus on finding valuable patterns and rules in text that indicate trends and significant features about specific topics’ [ibid.:967].

¹⁶ Cf. Mack and Hehenberger [2002:S97]: ‘In all these examples the relationships between terms were all expressed in the text documents, and, theoretically, are available for researchers to read, remember and comprehend.’

¹⁷ Halliman [2001:7] also hints in the direction of a scale of newness of information: ‘Some text mining discussions stress the importance of “discovering new knowledge.” And the new knowledge is expected to be new to everybody. From a practical point of view, we believe that business text should be “mined” for information that is “new enough” to give a company a competitive edge once the information is analyzed.’

¹⁸ Zorn et al. [1999:28] also use the concept of knowledge creation. Also compare Rob and Coronel [2002:654]: ‘... data-mining tools ... *initiate* analyses to create knowledge’.

¹⁹ According to Rob and Coronel [2002:654] data-mining tools are based on algorithms that form the building blocks for artificial intelligence, neural networks, inductive rules and predicate logic.

in databases) or *intelligent text mining* (for covertly structured data, i.e. inherently structured text data).²⁰ Surprisingly, the term ‘intelligent text mining’ has not yet become part of information systems jargon.²¹

Intelligent data and text mining should tell something about the world, outside the data collections themselves (to extend Hearst’s words), e.g.: 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 signaled 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 found linkages be refined and formulated?²² What investigation strategies are implied by the found linkages, and are they feasible and relevant? What social impact is suggested by the found linkages? Therefore, the discovery of patterns should be separated from the analysis and interpretation of the patterns. First-mentioned is semi-novel investigation and last-mentioned is novel investigation, which will be facilitated either by the interaction between the human researcher and the data- or text-mining tools, or by artificial intelligence.²³ Mack and Hehenberger [2002:S97] regards the automation of ‘human-like capabilities for comprehending complicated knowledge structures’ as one of the frontiers of ‘text-based knowledge discovery’.

In intelligent text mining artificial intelligence, and especially natural language processing, plays an important role. 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 to do ‘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.²⁴ Intelligent text summarisation 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 is to be found 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 (see Table 3), intelligent text summarisation 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 examined further in a separate research project.

The concepts of data, information and knowledge are also relevant for the distinction between non-novel, semi-novel and novel investigation. Data are raw facts that have no intrinsic meaning; it has to be sorted, grouped, analysed and interpreted to become information. Information, combined with context and experience, becomes knowledge.²⁵ 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, due 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.

²⁰ It is a fallacy that text data are unstructured. Text is actually highly structured in terms of morphology, syntax, semantics and pragmatics. However, it is true that these structures are not directly visible. ‘... text represents factual information ... in a complex, rich, and *opaque* manner’ [Nasukawa and Nagano 2001:967] (our italics).

²¹ One exception is a publication by Kontos et al. [2000:395-396], who use the term in the context of reaching a conclusion through reasoning (‘deductive inference’). They created a computational lexicon by using a machine-readable dictionary and, through user interaction and feedback, teaching the system to correctly interpret polysemous Greek verbs in stock market texts. This lexicon can be used for ‘intelligent information extraction and text mining’. According to them intelligent text mining is the ‘deductive mining of knowledge from texts (about the behaviour of companies)’ or the extraction of information ‘from text either directly or after applying deductive inference’. Greek sentences are analysed morphologically, syntactically and semantically, and transformed into Prolog facts (or predicates), which are used for deductive inference [ibid.: 404]. Intelligence here refers to the ability of logical deduction.

²² Researchers often make a huge cognitive leap from meagre indications to a theory: ‘Just occasionally empirical data may scream out a theory, in the sense of a vivid relationship, but more often it whispers and some cognitive leap is required to make the generalization (positivism) or to achieve the understanding (interpretivism)’ [Cornfield and Smithson 1996:44].

²³ Intelligent behaviour is ‘the ability to learn from experience and apply knowledge acquired from experience, handle complex situations, solve problems when important information is missing, determine what is important, react quickly and correctly to a new situation, understand visual images, process and manipulate symbols, be creative and imaginative, and use heuristics’ [Stair and Reynolds 2001:421].

²⁴ Sullivan [2001:37] regards the representation of meaning by means of syntactic-semantic representations as essential for text mining: ‘Text processing techniques, based on morphology, syntax, and semantics, are powerful mechanisms for extracting business intelligence information from documents ... We can scan text for meaningful phrase patterns and extract key features and relationships’.

²⁵ Cf. Poneis and Fairer-Wessels [1998:3].

	<i>Non-novel investigation – Data/information retrieval</i> (Finding/retrieving already existing and known information)	<i>Semi-novel investigation – Knowledge discovery</i> (Discovery of existing patterns – the patterns/trends already exist in the data, but are yet unknown and the discovery thereof is new)	<i>Novel investigation – Knowledge creation</i> (Creation of new important knowledge – tells something about the world, outside of the data collection itself)
<i>Numerical data</i> (including strictly formatted alpha-numerical fields) (Overtly structured)	<i>Database queries</i> (Uses database operations such as SQL queries) * Retrieves specific (mainly) numerical data	<i>Standard data mining</i> (Often uses statistical methods, e.g. link analysis; on-line analytical processing) * Reveals business patterns in numerical data	<i>Intelligent data mining</i> (Uses interaction between investigator and computerised tool; AI) * What do the patterns and trends mean and imply? * Which business decisions are suggested?
<i>Text metadata</i> (Overtly structured bibliographical fields, e.g. author, date, title, publisher, keywords; excluding free-text sections, such as abstracts)	<i>Information retrieval of metadata</i> (Uses exact match and best match queries) * Retrieves references to specific documents	<i>Standard metadata mining</i> (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	<i>Intelligent metadata mining</i> (Uses interaction between investigator and computerised tool; AI) * Compares and interprets distribution of text category labels within subsets of the document collection
<i>Textual data</i> (Inherently/ covertly structured)	<i>Information retrieval of full texts</i> (Uses exact match and best match queries) * Finds full texts of articles, etc.	<i>Standard text mining:</i> (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 ²⁶ * Compiles thematic overviews * Groups texts according to inherent characteristics (text-clustering) * Discovers subcategorisation data * Categorises texts into pre-existing classes * Summarises text intelligently	<i>Intelligent text mining</i> (Uses interaction between investigator and computerised 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 the hypotheses prompted by found linkages be refined and formulated? * What are the investigation strategies implied by the found linkages, and are they feasible and relevant? * What is the social impact suggested by the found linkages? * Which business decisions are implied?

Table 3. An overview of data and text mining and related fields

²⁶ According to Perrin and Petry [2003] 'useful text structure and content can be systematically extracted by collocational lexical analysis' with statistical methods.

A further extension of Hearst 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 summarised in a new table (see 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/opaque 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.

6. CONCLUSION

Because of the newness of the field of text mining, concepts and approaches still vary a lot. 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 make that distinction. This paper tried to move further, 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 artificial intelligence abilities to interpret the patterns and provide more advanced abilities such as hypotheses suggestion (intelligent text mining). Artificial intelligence, and especially natural language processing, can be used to simulate human capabilities needed for intelligent text mining. The necessity of natural language processing to facilitate intelligent text mining should be researched further. 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.

7. REFERENCES

- ALBRECHT, R. AND MERKL, D. 1998. Knowledge discovery in literature data bases. In *Library and information services in astronomy III*. (ASP conference series, vol. 153.) <http://www.stsci.edu/stsci/meetings/lisa3/albrechtr1.html>.
- BERSON, A. AND SMITH, S.J. 1997. *Data warehousing, data mining, and OLAP*. McGraw-Hill, New York, NY.
- 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*. University of Arizona (Knowledge Computing Corporation), Tucson, Arizona.
- CORNFORD, T. AND SMITHSON, S. 1996. *Project research in information systems: a student’s guide*. Macmillan, Houndmills. (Information system series.)
- HALLIMAN, C. 2001. *Business intelligence using smart techniques: environmental scanning using text mining and competitor analysis using scenarios and manual simulation*. Information Uncover, Houston, TX.
- HAN, J. AND KAMBER, M. 2001. *Data mining: concepts and techniques*. Morgan Kaufmann, San Francisco, CA.
- HEARST, M.A. 1999. Untangling text data mining. In *Proceedings of ACL’99: the 37th annual meeting of the association for computational linguistics*, University of Maryland, June 20–26 (invited paper). <http://www.ai.mit.edu/people/jimmylin/papers/Hearst99a.pdf>.
- HOVY, E. AND LIN, C.Y. 1999. Automated text summarization in SUMMARIST. In *Advances in automated text summarization*. I. MANI AND M.T. MAYBURY, Eds. MIT Press, MA, 81–94. <http://www.isi.edu/~cyl/>.
- 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*, D.N. CHRISTODOULAKIS, Ed. Springer, Berlin, 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, 005*, 1–3. <http://mappa.mundi.net/trip-m/hearst/>.
- 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*. <http://www.cs.waikato.ac.nz/~nzdl/textmining/>.
- 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 *Data mining and reverse engineering: searching for semantics*, S. SPACCAPIETRA AND F. MARYANSKI, Eds. Chapman and Hall, London, 50–64.
- ROB, P. AND CORONEL, C. 2002. *Database systems: design, implementation, and management, 5th ed.* Course Technology, Boston, MA.
- STAIR, R.M. AND REYNOLDS, G.W. 2001. *Principles of information systems: a managerial approach, 5th ed.* Course Technology, Boston, MA.
- SULLIVAN, D. 2000. The need for text mining in business intelligence. *DM Review*, Dec. 2000. <http://www.dmreview.com/master.cfm>.
- SULLIVAN, D. 2001. *Document warehousing and text mining: techniques for improving business operations, marketing, and sales*. John Wiley, New York, NY.
- THURAISINGHAM, B. 1999. *Data mining: technologies, techniques, tools, and trends*. CRC Press, Boca Raton, Florida.
- WESTPHAL, C.R. AND BLAXTON, T. 1998. *Data mining solutions: methods and tools for solving real-world problems*. Wiley, New York, NY.
- ZORN, P., EMANOIL, M., MARSHALL, L. AND PANEK, M. 1999. Mining meets the web. *Online 23(5)*, 17–28.