Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2006 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2006

Text Mining Promise and Reality

Antonina Durfee
Appalachian State University

Text Mining Promise and Reality

Antonina V. Durfee
Appalachian State University
durfeea@appstate.edu

ABSTRACT

This paper provides taxonomy of common text mining tasks and approaches. We surveyed the market of modern text mining tools, compared their features and grouped them into information retrieval, standard or intelligent text mining categories in order to examine how theoretical promises materialized in modern technologies. The study is the first one in a series of studies trying to provide an understanding of impediments in the development of text mining products and users' expectations.

Keywords

Text mining, information retrieval

INTRODUCTION

A massive quantity of information continues accumulating in numerous text repositories held at news agencies, libraries, corporations, individual PCs, and the Web due to the cheap digital storage and fast processing. The amount of stored information proliferates at about 1.5 billion gigabytes per year. A large portion of all available information today exists in the form of unstructured texts. Plain text accounts for 24 terabytes of data growth per year (Lyman, 2000). About 80% of a company's information is saved in form of digital text (Tan, 1999) and about 80% of the world's online content is text based on free-form text. A few years ago, the Internet archive collected about five times more unstructured information than the Library of Congress of the USA, the largest library in the world. We are increasingly unable to meet the challenges of this growth.

Researchers, analysts, magazine editors, venture capitalists, lawyers, help desk specialists, and even students are faced by text analysis challenges. Analyzing large amount of textual information is often involved in making informed and correct decisions in a *timely manner.* A dynamic business environment does not allow decision makers and knowledge workers to spend sufficient time locating, reading and analyzing relevant documents to produce the most informative decisions. As a result, only a small fraction of the collected textual data is ever getting analysed (Rajman and Besancon, 1998). Managers and decision makers are searching for intelligent electronic assistance and help for automating different text analysis projects.

Text mining (TM) tools aim at knowledge discovery from textual databases by isolating key bits of information from large amounts of text, by identifying relationships among documents, and by inferring new knowledge from them. These new relationships and information can assist users in effective problem structuring and resolution. TM promises its users the ability to categorize, prioritize, understand and compare documents, and utilize the meaning of any particular document automatically skipping tedious searching, browsing and reading.

TM is an interdisciplinary field that comes under different names, such as text analytics or textual data mining, and is often confused with text processing, text management, natural language processing (NLP) (or computational linguistics) and information retrieval (IR). While great progress has been achieved in supporting fields of text processing and IR, the TM field remains fragmented with the ambiguous operational definition. As a result, the differentiation between tasks and approaches of TM or mere text processing remains unclear for TM solutions users. There is a necessity to identify which human information need can be met by available TM technology. Knowledge workers and decision makers want to know which modern TM tool is capable of performing what task. Software developers want to know which features should be included in their TM products to satisfy users' ever increasing analytic needs.

This paper provides an overview of features and processes used for general TM; presents the criteria for grouping TM tools into IR, standard TM or intelligent TM according to their tasks; investigates existing tools using proposed feature classification; and suggests a list of features to be included in modern TM tools to be more intelligent.

The paper is organized as follows. Section 2 provides a literature review to clarify the operational definition of TM, its tasks and intersection with related fields. Section 3 aggregates TM tasks with types of investigation and presents a model of TM

feature classification. Section 4 contains a methodological description of the study. Section 5 groups TM tools according to their features. Section 6 provides a discussion of the current state of existing TM tools and identifies desirable features and trends in TM development, imitations and intended future research. Section 7 draws some conclusions about future development.

LITERATURE OVERVIEW

Foundations of Text Mining

Text has richness of interpretation and meaning with a complicated and ambiguous multilevel structure of tens of thousands of dimensions (Fayyad and Uthurusamy, 2002). Structural principles exist in the formation of words (morphology of language), in the creation of grammatical sentences (syntax), and representation of meaning (semantics). The three components of text: word usage, grammatical construction, and content vary within every individual document and language. The authors and readers of the text often represent the same semantics using different words (synonymy) or describe different meanings using words that have various meanings (polysemy) (Manning and Shutze, 1999). Symbols, word usage, clause construction, content and reader backgrounds contribute to the detection of meaningful patterns that lead to text interpretation and understanding.

TM is a relatively new discipline that has generated academic discourse concerning its definition and tasks. The first definition of TM was suggested by M. Hearst in 1999 as an extension on knowledge discovery from databases (KDD). KDD is a process of "identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et al., 1996). Following Hearst's definition of TM as a process of discovering "patterns and associations useful for particular purposes from textual databases", TM is viewed as data mining on textual data ((Dörre et al., 1999), (Thuraisingham, 1999), (Nasukawa and Nagano, 2001)). Alternatively, Miller views TM as "the automated or partially automated processing of text". (Miler, 2005).

The concepts of "data" and "novel" in the definition of KDD greatly contribute to confusion around TM. Depending on a type of data to be mined, researchers differentiate data mining for numeric (structured) data and text mining for textual (unstructured) data. In reality not all text is unstructured data since meta textual data is inherently structured (e.g. name, abstract, keyword of a scientific paper). Data classification of overtly structured (numeric or alphanumeric) and inherently, covertly structured (textual characters) instead of numeric and textual data is more accurate (Kroeze et al., 2003). Depending on knowledge novelty that can be extracted from a database by querying, Berson and Sminth (1997) identified 3 types of discoveries: what I don't know I don't know (the most difficult knowledge to mine resulting in novel investigation), what I don't know I know (semi-novel investigation) and what I know I don't know (non-novel investigation). The determination of novelty of textual patterns is intricate since some researches argue that text is information (not data) and all patterns in it are known to at least an author.

This paper builds on the ideas of Kroeze et al. (2003) who classified mining according to types of data to be mined and the types of discovery to be performed as represented in Table 1. The researchers clarified the traditional definition of TM by further dividing it into IR, standard and (truthful) intelligent TM. *Information retrieval* (IR) is the process of locating the subset of the documents that are deemed to be relevant to a posed query(van Rijsbergen, 1979). *Standard mining* is a process of finding semi-novel useful patterns and is referred to as real text mining by Hearst (1999). Although lexical, syntactic patterns and new themes already exist in text, they are yet unknown and the discovery thereof is new. *Intelligent mining* can be regarded as human-like capability for comprehending complicated structures and "creating knowledge outside of data collection" (Mach and Hehenberger, 2002), e.g. "Which business decisions are prompted by discovered patterns? How can the linguistic features of text be used to create knowledge about the outside world? Does a newly discovered theme in a text collection reflect or validate the reality? Could the hypotheses prompted by found linkages be refined and formulated?"

Type of investigation/ Type of data	Non-novel investigation	Semi novel investigation	Novel investigation
Numeric data (overtly structured alphanumeric)	Database queries	Standard data mining	Intelligent data mining
Text metada (structured textual data)	Information retrieval of metadata	Standard metadata mining	Intelligent metadata mining
Textual data (inherently, covertly structured)	Information retrieval of full texts	Standard text mining	Intelligent text mining

Table 1. A classification of data and text mining (adopted and modified from (Kroeze et al., 2003))

TM Tasks

Standard TM uses statistical and natural language processing methods to explore patterns in text. This general task is accomplished by specific mathematical approaches: *clustering, feature extraction* and *thematic indexing* (Hand et al., 2002). Schutze and Silverstein (1997) state that speech recognition, language models, parsing, and machine translations are not TM tasks. The researchers consider clustering, information extraction, question answering as the typical TM tasks. *Clustering* in TM is the process of partitioning a given collection into a number of previously unknown groups of documents with similar content. Clustering allows for the discovery of unknown or previously unnoticed links in a subset of documents or terms in any particular document collection. *Feature extraction* refers to the extraction of linguistic items from the documents to provide a representative sample of their content. Distinctive vocabulary items found in a document are assigned to the different categories by measuring the importance of those items to the document content. *Thematic indexing* refers to the identification of the significant terms for a particular document collection. Indexing identifies a given document or a query text by a set of weighted or unweighted index terms or keywords obtained from a document or a query text. Intelligent TM combines the mathematical approaches of IR and standard TM together with machine learning (ML) and artificial intelligence methods to enable interaction between the TM tool and an investigator (knowledge worker, decision maker). Intelligent TM brings some learning component into analysis by, for instance, combining it with predictive data modelling (Kloptchenko, 2003). In an attempt to recognize the semantic peculiarities of text, standard and intelligent TM methods use more elaborated text encoding and representation algorithms (vector quantization, parsing)than simple bag-of-word method.

TAXONOMY OF TM

Based on the types of patterns discovered and approaches involved, we can distinguish functions and results of IR, standard and intelligent TM (Table 2). The IR system assumes that the user has a classification system in mind that separates the relevant documents from nonrelevant ones. Traditionally, IR systems are query-based, and they assume that users can describe their information needs explicitly and adequately in the form of a query. Modern IR conveniently relies on language representation as a "bag of words" which views a language as a fixed stock of words. Words interact in many ways: some words co-occur near certain words with higher probability than others. The product of the frequency of words and their rank (the order of importance) is, according to Zipf's law, approximately constant (Zipf, 1972). The extraction of important keywords or indexes from text does not guarantee the extraction of meaning from text. The danger of the keyword approach is in using different keywords by different individuals to describe the same concept (synonymy) while creating a query. A part of a document that does not include query-matching keyword is ignored by conventional IR systems. IR can be applied for text categorization, text routing and text filtering (Riloff and Hollaar, 1996). Standard TM performs feature extraction and text categorization based on features which enables summary creation and document comparison. Those features are formed not only by index terms or keywords but by their co-occurrences. Text categorization assigns documents to pre-existing categories, called "topics" or "themes"(Lewis, 1992). Intelligent TM discovers new patterns that enriches domain knowledge or validates already existing patterns against data domain. In other words, intelligent TM should be able to build predictive models or hypothesis.

Type of investigation	Non-novel investigation	Semi novel investigation	Novel investigation
Features	Information retrieval of full texts uses exact match and best match queries: Compose a query Index text collection Search text relevant to a query Retrieve relevant document Locate a (set) text/documents	Standard text mining uses statistical methods Feature extraction Thematic indexing Cluster and categorize text Discover link between themes in text (rule induction) Visualize themes/relationships in/among documents	Intelligent text mining uses interaction between investigator and a tool, AI Validate the discovered theme with reality What business decision are implied by Make inferences of textual content (Hypothesis formulation) Extends knowledge based on extracted features
Description of Tasks	Search and locate relevant to a query document/piece of a document, document extraction, text routing and filtering	Create automatic thesauruses, summary, topic hierarchy, automatic dictionary, classify new text in a new categories, author attribution	Create additional knowledge/ hypothesis about reality, predict future state of reality

Table 2. Features and tasks of IR, standard and intelligent TM systems

METHODOLOGY

The process of TM tool comparison and grouping includes two phases: selecting TM tools for comparison and grouping them according to their features. We intentionally omitted software whose primary focus is statistical or mathematical engines not TM (e.g. MATLAB, Statistica). We collected a list of most known TM products from websites of NEMIS (Network of Excellence in Text Mining and its Application in Statistics) and kddnuggets (forum of knowledge discovery from database professionals and academicians). In order to describe TM products we rely on viewing TM as a sequence of text representation and distillation, knowledge sophistication and representation (Figure 1), where:

1. *Text representation and distillation* transforms and represents free-form text in a chosen format and/or consolidates documents from various sources. Text from a string of symbols has to be encoded in some numeric format (numeric vectors or ranges) within a document or collection of documents. Encoded text from every individual document is transformed further into lower dimensional formats via word stemming[1], word disambiguation, constructing a dictionary of word senses[2], and removing stop words that occur very frequently in a document and do not contribute to overall meaning.
2. *Knowledge sophistication* deduces concepts and patterns from the distilled text by the utilizing knowledge discovery algorithms. A document or entire document collection can be clustered, categorized, or visualized to reveal inter-document or interterm relationships. The extracted features from a document or collection can be summarized to present new knowledge to a user.
3. *Knowledge (relationship) representation* delivers and presents the deduced knowledge to a user. The discovered relationships from the previous part are presented in some graphical or other visual form that a user can easily interpret, i.e. lists and tags, hierarchies, hypertext diagrams, semantic maps, tables or matrixes.

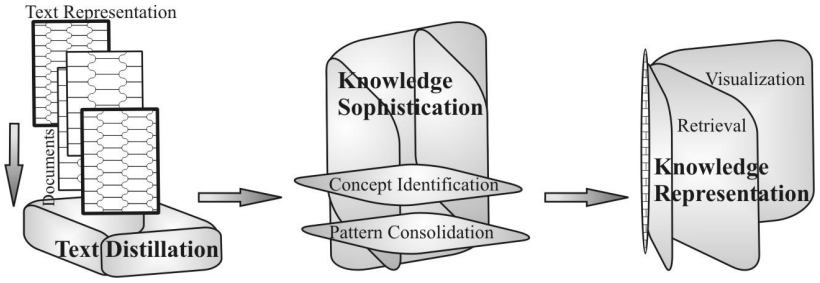

Figure 1. General TM framework

The evaluation criteria for grouping TM tools are based on the features and tasks that tools help to accomplish. The description of mathematic tasks of IR, standard TM and intelligent TM is presented in Table 3. The comparison of TM products in the second phase is performed by matching deliverable features with the theoretically desired ones.

RESULTS

Feature Comparisons of TM Tools

There are a number of tools from such moguls as IBM and more narrow focused SAS to academia-based *Text Miner* and *webSOM* that incorporate different mathematical algorithms to solve text related problems. Appendix 1 lists the best-known TM products by the end of 2005, whose feature and tasks are summarized in Table 3. The products are described based on the domain they can be used in, the status of their development, their knowledge sophistication and representation methods. The absolute majority of the presented products uses stemming, synonym list[3] composition, text parsing[4], and dimension reduction[5] for text filtering and representation.

---

[1] For instance, the canonical form of the words analyzed, analysis, analyzing is "analy".

[2] For instance, the word bank in a financial document means a depository or financial institution, but not a sloping land.

[3] Synonym table composition assigns one meaning to every term used in document vocabulary (Riloff, E. and Hollaar, L. (1996) *ACM Computing Surveys*, **28,** 133-134.

[4] *Text parsing* algorithms convert text into word, phrases or clauses and put them into short memory. Text parsing decomposes text and generates a quantitative representation suitable for DM (Mayes, M., Drewes, B. and Thompson, W. (2002) In *Distilling Textual Data for Competitive Business Advantage*Heldelberg, Germany, pp. 20.

[5] Dimension reduction algorithms treat every document as a vector where each dimension is a count of occurrences of a different word. It results into tens or hundreds of dimensions of every document (Isbell, C. L. (1998) *Advances in Neural Information Processing Systems).*

Tools (quantity)	Intended Domain	Knowledge Sophistication Approach	Knowledge Representation Approach	Additional Features
Group1 "IR" (12)	Any, e.g. Business Intelligence, Email Routing, E-commerce, Knowledge Management.	Indexing Categorization Feature Extraction Bayesian Inference Query-term expansion	Retrieval Keyword listing Searching Navigation Browsing	Language independent , criteria/sorting results, spider technology, multilanguage recognition/ relevancy
Group 2 "Standard TM" (24)	Any	Categorization Clustering classification Feature Extraction Fuzzy pattern matching	Semantic Retrieval Visualization (concept mapping) Summarization Tracking Routing	Multi-tier extracting of terms, online access, multiformat support
Group 3 "Intelligent TM" (12)	Any	Clustering Thematic Indexing Categorization Rule induction	Summarization Visualization Hypothesis creating	Language independent

Table 3. Aggregated tasks and features of TM products

Group 1 represents IR tools which index document collection and assist users in a process of non-novel discovery or navigation within single or multiple documents. Tools enable document retrieval by choosing satisfactory matches to a submitted query. IR systems are query-based methods, which rely heavily on the use of term (keywords, items, indexes) extraction, i.e. *SONIA*, *TextMiner, Sapere*. *dtSearch* offers desktop and network retrieval engine that includes a variety of forensics-oriented features, such as automatic parsing analogous to those recovered through an "undelete" process or partially recovered file fragments, proprietary filtering for scanning recovered text, and language recognition algorithms for detecting text in a large variety of languages. Some of the IR tools borrow ML techniques to help users in query formulation. *DataSetV* and *dSearch* use fuzzy logic for constructing a better formulated query and searching. As a trend, IR tools support multilingual retrieval from different file formats, e.g. *ISYS* supports 125 file types.

Group 2 represents standard TM tools which determine features in text, create themes based on those features, build links among different themes, categorize text, visualize text features and/or summarize text. Integration of clustering and categorization algorithms for TM with easy to interpret representation of the results are main features of these systems. An emerging feature is enabling systems to work on-line in real time with different text formats consolidated from various databases. For instance, *Copernic* searches corporate intranets, servers and public websites and uses vector representation of documents to create unparallel indexes that enable to launch federated search on many indexes. A user can track the appearance of the index in various sources and pinpoint the key concepts of texts to extract the most relevant sentences to produce a condensed version of the original text (summaries) and ignore irrelevant text. *VisualText*® is the first integrated development environment for NLP that integrates multiple strategies, including statistical, keyword, grammar-based, and pattern-based, as well as diverse information sources, including linguistic, conceptual, and domain knowledge, to quickly and efficiently develop text analysis applications. *Compare* suite uses a traditional "bag of word" model to compare documents word by word and infer knowledge of newer vs older versions of the same document. *Docyoument* is a multipart visualization tool that creates topic maps, landscapes, networks of themes and main concepts from a text. *Onix Text Summarizer* and *Copernic Summarizer* compose summaries that not only highlight the main sentences in a document but construct new ones based on the main ideas introduced in text. *Enkata* enables users not only to identify main concepts and summarize the meaning of a document but to track concept migration and evolution among the documents.

Group 3 combines intelligent TM tools, which are comprised of very few products: *SAS Text Miner*, SPSS Predictive Text Analytics (*Clementine*). These systems create new knowledge by discovering novel patterns in text and linking them to a specific domain. In order to be called intelligent, tools satisfy at least one of the following criteria: adapt in a functional way to a new situation presented by new data (produce new knowledge of outside world), offer a solution to a new situation (propose possible actions based on analyzed content), relate new situations to old ones (compare content of documents, build hierarchy of knowledge from documents), derive a decision on an asymmetric information or ill-defined context (learn from content of presented documents). Modern intelligent TM products are tool boxes with different ML, statistic and NLP algorithms that require high user proficiency. The tools can handle different types of data so a user can construct complex models for cross validation and verification. They offer great graphical capabilities (more than 20) which require an expert to interpret. These tools are employed by the majority of Fortune 500 companies.

From Text Related Tools to TM Solutions

The development of software solution for text analytics has been taking place in several generations over the course of a few decades. The first generation of content-base document management (pro-TM) systems appeared in the 1980s and consisted of research-driven tools focusing on single tasks. These tasks included IR and indexing text collections, clustering (for example, hierarchical clustering using Ward's method (El-Hamdouchi and Willett, 1986)). Intended users of the earlier systems had to be technically sophisticated and allocate a lot of time for transforming textual data between different systems to perform more than one analytical operation. The second generation of TM systems came around at the end of the 1990s, and was called TM suites. These systems recognized the complexity of KDD process and included not only text preprocessing but also mining and visualization capabilities. The suites like *SAS Text Miner* and *IBM Intelligent Miner for Text*, *Enkata*, *Entriva*, *InsightfulFact* allow a user to perform several discovery tasks, such as indexing, classification, clustering and visualization. The suites support data transformation and representation of the results visually in quite sophisticated ways (e.g. 3-D maps and landscapes).

An examination of the tools listed on a popular directory for knowledge discovery community kdnuggets.com maintained by G. Piatetsky-Shapiro and computerworld.com (KDD community blog) reveals the stabilization in the number of TM suites (40 in 2003 and 45 in 2005). The number of commercial suites has a tendency to decline once a market gets consolidated. At the same time, the number of academic or research suites for IR by content and TM are still increasing (TREC conference). The second generation of TM tools requires from its users significant knowledge and skills in statistical, NLP and ML theory, and, thus, those tools are not very appealing to business users. Business users led the development of vertical, third generation TM solutions in the beginning of the 2000, where a specific business problem, such as e-mail filtering or categorization (for example, *dtSearch*, *Klarity*), medical text summarization, or financial news organization (*Factiva*), are targeted. As estimated by Monash (2005) medical-discovery text mining industry is worth around a $10 million. Graphical interfaces of TM tools hide all mathematical complexity of TM and appear to be relatively user friendly. The fourth generation of TM development is a development of intelligent TM systems that discovered value added knowledge about the reality outside a text collection. There is an inherent difference between creating a summarization, or simply restating briefly the content of a document (standard TM) and comparing documents or creating predictive models. For instance, *DigimineRetail Advisor* creates knowledge and offers cross-selling recommendations by analyzing numeric and textual data for activities on the web. It appears that intelligent TM solutions of today are possible for vertical application but they require highly skilled professionals and lack user friendliness. They incorporate not only processes of TM (see Figure 1) but also include domain specific expertise in form of ML inference from domain specific data. To obtain user friendliness, TM needs to settle with fixed terminology and standards to indicate industry maturity.

DISCUSSIONS, LIMITATION AND FUTURE RESEARCH

TM products, being successors of longer maturing fields of NLP, statistics, machine learning and IR can be used as parts of advisory or decision-support systems and assist by exploring lines of analysis and problem structuring. TM products evolution led to the acceptance of the various data files (ASCII, doc, pdf, txt, ps, SQl, html, xml, rtf) and online support with emphasis on results visualization (mapping based on concepts, indexes, and keywords). The exploratory navigation in TM is enriched by drill down capabilities. Most of the proprietary algorithms used in TM products are improved combinations of well-established statistical retrieval models, traditional text representation and machine learning (i.e. neural network, support vector machine, fuzzy logic) methods. Trying to position themselves on the market companies either pursue specific text domain or offer suite solutions that allow a user to combine software products to satisfy specific TM needs. Some vendors, *Matchbox*, *InformationExtraction_and_TextClassification library*, *Klarity* offer tailored TM solution to a particular client, while others (*Mindserver*, *Inxight*) offer software suite that cover most text related operations. We see more and more suites being built where several approaches and algorithms are offered to summarize or retrieve text (*Megaputer*, *SAS*, *Enkata*).

There is an attempt to analyzing documents in a collection not only on word level but on higher morphological and semantic levels. Some of the surveyed tools from group 1 offer word or paraphrase level analysis. Tools from group 2 exhibit more sophistication and try to capture content of a document by mining it on sentence or paragraph level. *Insightful inFact*, for instance, treats documents on three levels: morphological (word root clusters represented as the number of shares n-grams[6]), semantic (sentences represented as the linguistic normalization maps) and syntactic (patented transformation rules to recognize semantic equivalence of multiple sentence structures). A multilevel approach to document representation and analysis should be required features of any TM solution.

[6] An n-gram is a sequence on a consecutive letters. The words "mine", "miner", and "miners" share three unique diagram (mi, in, and ne), two unique trigram (min, ine) and one unique quadram (mine) (Insightful Inc.).

Big corporations with commercially available software packages, such as IBM, SAS, and SPSS offer the combination of data and text mining solutions in the form of toolboxes or add-on modules of various applicable algorithms, i.e., hierarchical and k-means clustering, etc. Notably, TM modules in those packages were introduced only recently as an individual exploration tool for different types of unstructured data. TM discovers the patterns that can serve as hints to unlock the knowledge contained in text data so that it can be combined with data from numeric databases to build better models. Consequently, binding sophisticated data and text mining algorithms requires very specific mathematical and domain expertise from those who wish to apply them for effective problem solving.

The open question remains in all these applications: how to integrate domain knowledge with the results of TM tools. As Tan (1999) noticed, domain knowledge can be used in a process of knowledge discovery from text as early as in the text refining-distillation stage. The interpretation and evaluation of the discovered patterns are still cumbersome and include intensive human involvement. The requirements for well-trained users who can interact with TM systems are still obligatory. Managers - heavy consumers of textual information - rarely have the time or technical expertise to master complicated TM applications and to gain the experience to recognize valuable discovered patterns.

One can argue about the limitation of the chosen technologies and how they were included in the study. The information about TM products were gathered mostly from the webpages, white and technological papers of the companies, industry reports and scientific conferences proceedings. As a part of our future research we plan to survey the actual needs of users in performing text related operations.

CONCLUSION AND FUTURE TRENDS

The existing categorization of TM systems is not accurate due to multiple levels of operational definition problems. As a result many IR systems are misrepresented as TM systems, which can lead to a hazardous situation where users' expectations exceed what technology can deliver (Fayyad, October 05). In order to qualify as a standard TM tool, a system should be able to create automatic concept dictionary, to present the content of a document in a concise form (summary or abstract), to build a hierarchy of different topics/concepts presented in a text collection or in one specific document, to cluster text in groups with similar themes, to compare text based on pattern similarities, and to visualize the results using mapping for easier navigation. In reality, the systems perform more indexing and retrieval, summarization and categorization jobs.

In order to qualify as an intelligent TM tool, a system should be able to create additional knowledge about reality based on the content of a text in form of hypothesis or predicting models, categorize new documents according to the derived categorization scheme and improve it, and integrate and validate derived knowledge with domain. Intelligent TM systems of today are toolboxes with several mathematic algorithms (statistical clustering, PCA categorization, NLP feature extraction, neural networks) that can analyze textual and numeric data related to phenomena and build some predictive models on it. Those models are all domain and problem specific, unlike standard TM solutions which are problem specific but domain independent. As a future development, horizontal TM systems should offer easy graphical interface and ability to build KDD process based on relevant data in different languages and formats collected automatically and to result into models. Many tools currently available are generic tools from machine learning or statistical communities. The tools operate separately from the data source and require significant preprocessing. At the same time, realistic knowledge discovery process is iterative and interactive. Intelligent TM tools should include tight integration with database management system for data selection, preprocessing, and result validation. The capability to directly access different data sources from online as well as offline will greatly reduce data transformation task. In the light of increasing number of proposed algorithms and mathematical models for text mining, it is important to provide architecture for easy synthesis and adaptation of new methods for experienced users as well as novice ones.

This paper outlines common trends in product development and derives some explanation as to why modern TM products are in fact merely text processing IR or at the best standard TM products. We proposed a list of qualifying features that TM products should possess in order to be called intelligent. This list o features can be used as guidelines for software developers to create marketable applications.

REFERENCES

1. Baeza-Yates, R. and Ribeiro-Neto, B. (1999) *Modern Information Retrieval*, ACM Press, New York.
2. Berson, A. and Smith, S. J. (1997) *Data warehousing, data mining, and OLAP*.
3. Dörre, J., Gerstl, P. and Seiffert, R. (1999) In *KDD-99, Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* ACM, San Diego, USA, pp. 398-401.
4. El-Hamdouchi, A. and Willett, P. (1986) In *ACM Conference on Research and Development in Information Retrieval*ACM Press, pp. 149-156.
5. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) *Communications of the ACM*, **39,** 27-34.
6. Fayyad, U. and Uthurusamy, R. (2002) *Communications of the ACM*, **45,** 28-31.
7. Isbell, C. L. (1998) *Advances in Neural Information Processing Systems.*
8. Kloptchenko, A. (2003) In *First Annual Pre-ICIS Workshop on Decision Support Systems* Seattle, USA.
9. Kolenda, T. and Hansen, L. K. (2002) In *IEEE Workshop on Neural Networks for Signal Processing XII* IEEE Press.
10. Kroeze, J., Matthee, M. and Bothma, T. (2003) In *SAICSIT*, pp. 93 –101
11. Lewis, D. (1992) In *Speech and Natural Language Workshop*.
12. Lyman, P., and Varian, H. (2000) University of California at Berkeley.
13. Mach, R. and Hehenberger, M. (2002) *Drug discovery today*, **7,** S89-S98.
14. Manning, C. and Shutze, H. (1999) In *Foundations of Statistical Natural Language Processing* The MIT Press, Cambridge, MA, pp. 141-177.
15. Mayes, M., Drewes, B. and Thompson, W. (2002) In *Distilling Textual Data for Competitive Business Advantage.* Heldelberg, Germany, pp. 20.
16. Miler, T. (2005) *Data and Text Mining*, Prentice Hall, Upper Saddle River, NJ.
17. Nasukawa, T. and Nagano, T. (2001) *IBM Systems journal*, **40,** 967-984.
18. Rajman, M. and Besancon, R. (1998) In *6th Conference of International Federation of Classification Societies (IFCS-98)* Rome, Italy.
19. Riloff, E. and Hollaar, L. (1996) *ACM Computing Surveys*, **28,** 133-134.
20. Schutze, H. and Silverstein, C. (1997) In *SIGIR 97*, Vol. 3 ACM Press New York, NY, USA, Philadelphia, PA, USA, pp. 74-81.
21. Tan, A. (1999) In *PAKDD-99, Workshop on Knowledge Discovery from Advanced Databases (KDAD'99)*Beijing, China, pp. 65-70.
22. Thuraisingham, B. (1999) In *CRC Press*Florida.
23. van Rijsbergen, C. (1979) *Information Retrieval (Second Edition)*, Butterworths, London:.
24. Zipf, G. K. (1972) *Human behaviour and the principle of least effort. An introduction to human ecology*, 1st edn:Cambridge, MA: Addison-Wesley, 1949, New York: Hafner reprint.

Appendix 1. Text Mining Software

System Name (reference)	Purpose (C / P)[10]	Intended Domain	Knowledge Sophistication Approach	Knowledge Representation Approach	Additional Features
GROUP I: IR					
Autonomy (autonomy.com) , Autonomy Inc.,UK	C	BI, Email Routing, E-commerce, ERP, Knowledge Management	Bayesian Inference on pattern-matching	Retrieval	Conceptual Search, language independent
CINDOR(ConceptualINterlinguaDOcumentRetrieval (www.cindorsearch.com), TextWise Inc., NY, USA	C	Any	Indexing Query-term expansion	Query- based retrieval	NL retrieval in several languages
Compare Suite (www.comparesuite.com) ***File Search Assistant*** (www.aks-labs.com/products/fsa_home), AKS-Labs, Raleigh, NC, USA	C	Any/compare documents for research, translators and writers, monitor competitor website, find latest files	Indexing	Keyword listing Searching Navigation	Compare text by keywords, highlights common and unique keywords/flexible search (-,+) criteria/sorting results
DataSetV (www.ds-dataset.com), Intercon Systems Inc., Israel	P	Integrated Knowledge Management system, data cleansing	Record Matching Fuzzy record look up	Retrieval Ranking Visualization	statistical analysis to segment search strings to determine retrieval potential for each query character/SQL support/probabilistic retrieval/fuzzy search
dtSearch (www.dtsearch.com), dtSearch Corp.,MD, USA	C	Document management: email, filtering, legal, medical, financial, forensics	Indexing	Retrieval	fuzzy search/spider technology, multilanguage recognition/relevancy ranking/automatic parsing/proprietary filtering
Information Extraction library and Text Classification library (www.media-style.com), Media Style GmbH, Germany	C tailored	Any	Identify entities, Feature (tagger) Extraction, Categorization	Retrieval	library is trainable and supports different languages allows plug-able preprocessing filters
ISYS: search www.isysusa.com), ISYS Search Software Pty Ltd, Sydney, Australia	C	Government, legal, financial, healthcare and recruitment	Indexing, Categorization	Retrieval Hit-to-hit navigation	desktop, intranet, website, email and network search/ fully inverted index with optional fuzzy preconception and relevance ranking, supports 125 file types
LexiQuest Mine, LexiQuest Categorize, LexiQuest Guide (www.spss.com/home_page/wp130.htm), SPSS Inc., IL,USA	C	Legacy Systems, CRM, Investment Research, e-mail filtering, Combining Data and Text Mining for Business Intelligence	Categorization (statistical proximity matching), Feature Extraction	Retrieval	Proprietary Language Recognition based on 600000 word dictionary instead of keyword queries

[10] C refers to commercial, P to prototype. Commercial (C) tailored means that software is primary consulting solution.

Matchpoint (www.triplehop.com) TriplehopTechnology NY, USA	C tailored	Information discovery	Indexing Categorization (support vector machine)	Retrieval	context-sensitive search crawler technology for searches, collaborative searches and meta-searching technology, multi-criteria searches, user profiling and domain-specific focus
Sapere (ai.mit.edu/research/abstracts/abstracts2002/naturallanguage/04katz.pdf), AI Lab, MIT, MA, USA	P	Any	Indexing	Retrieval	Using ternary expression to facilitate easier indexing for storing knowledge and answer queries
WordStat	C	Any: open-ended questions, interviews, articles, public speeches	Indexing	Retrieval	Open-ended searching
GROUP II. Standard TM					
Atlas.ti (www.atlasti.de) Scientific software Development, GmbG, Germany	C	Any document based research areas (sociology, theology, psychology, etc)	Indexing Categorization	Semantic Retrieval Visualization ("mind-mapping')	Document-based analytical research
Copernic Agent: Summarizer and Tracker (www.copernic.com), Copernic TechnologiesInc.,Quebec, Canada	C	Any	Indexing Federated search	Summarization Retrieval Tracking	Tokenization/concept extraction/ incorporate agent technology
ClearForest Text Analitics, (www.clearforest.com) ClearForest Corp., MA, USA	C	Chemical, Manufacturing, Consumer Goods, Financial Services, Government Agencies, Insurance, Publishing,	Feature Extraction, Indexing, Categorization	Retrieval Visualization (tags)	Text tagging
digiMine Retail Advisors (www.digimine.com/solutions/howdigimineworks.asp), digiMine Inc.,NY, USA	C	Customer Analytics and Interaction Optimization from emails, databases and web logs	Collects data from various sources, parse and clean the data, produce analytical reports based on business profiles	Summarization (Analytical Assertion) Visualization	Provides cross-sell recommendations, runs on analytics of web activity
DocMiner (www-i5.informatik.rwth-aachen.de/lehrstuhl/projects/DocMINER/index.html), Institute for Applied Information Technology,Germany	P	newsgroup articles, software documentation	Indexing Clustering	Summarization Retrieval Visualization	Visualization via adaptable document maps based on topology of document dissimilarities
Docyoument (www.mediastyle.com/docyoument.html), Media Style, Hamburg, Germany	C tailored	Information management	Classification Clustering	Retrieval Visualization (topic map, landscapes, networks) Summarization	analyzes, categorizes, available information based on user-defined topics and questions
Enkata:The Statistical Text Mining (www.enkata.com/), Enkarta Technologies Inc.,CA, USA	C suite	Enterprise level solutions for Healthcare, Telecom, Customer Analysis	Categorization (probabilistic) Feature Extraction	Concept mapping	data enrichment using statistics/ proprietary fee form text analysis using Active Learning ™

Entriva (www.entrieva.com) successor of Semio Map (Tagger and, Discovery),Entrieva Corp.,VA, USA	C suite	Knowledge management base on Advanced Text Analytics Platform ™	Categorization Indexing	Retrieval Visualization(3D map)	multi layered concept map for searching, phrase based query, personalized notification, web search
Factiva.com (www.factiva.com) , NY, USA	C tailored	Dow Jones and Reuters newswires	Classification Feature Extraction	Summarization Visualization (taxonomy) Tracking	enhanced searching, taxonomy generation, multiple language interfaces from 9,000 sources
IBM Intelligent Miner for Text (Tkatch 1997), IBM Corp., CA, USA	C	Knowledge Discovery, Information Mining	Clustering Classification, Feature Extraction	Retrieval Browsing Visualization	language identification, finding similarities based on lexical affinities
Information Compiler and PointScope (www.insight.com.ru) InsightSoft-M company, Moscow, Russsia	P	Any	Feature Extraction	Summarization Retrieval	synopsis of topics/composes query related reports/extracts information from collection of documents and deliver exact answer to a desktop
InsightfulInFact (www.insightful.com), Insightful Corp.,WA, USA	C suite	Biopharmaceutical, financial, manufacturing, telecom	Automated incremental indexing Fuzzy matching	Retrieval Summarization Visualization	constructs semantic network of associated terms
Insight Discoverer: Categorizer, Clusterer, Extractor (www.capital-k.com/) Online Miner (temis-group.com), Autentica/Temis, France	C	Knowledge Management for Financial Market Analysis, Competition Intelligence	Categorization, Clustering, Information Extraction	Retrieval Visualization	Search Result Organization, Document Mapping using trained classification model of large online document collection,
Inxight (www.inxight.com), Inxight Software, Inc.,CA, USA	C suite	Information discovery for government organizations, pharmaceutical, financial; publishers; software vendors	Classification (topic/subject/ entity) Entity extraction	Summarization/ Retrieval/Navigation/Visualization (relationship, timeline and trends)	identifying, sorting and delivering text by word tokenization, stemming, de-compounding, part-of-speech tagging, noun phrase extraction, support 31 languages
Klarity (www.intology.com.au), Intology, Canberra, Australia	C tailored	Document management and resource discovery	Concept based categorization Keyword extraction Indexing	Retrieval enriched search, Summarization	keyword extraction to create metadata of a document to categorize it, build taxonomy of collection and rules to categorize documents
Kwalitan 5 (www.kwalitan.net), V.Peters, Department of Research Methodology Universityof Nijmegen, TheNetherlands	C	Analyzing qualitative data (text, pictures, audio)	Indexing Categorization	Retrieval Tree-like Visualization	uses codes for text fragments to facilitate textual search, display overviews or words and frequencies of their use, create memos referring to the codes, segment document
Leximanser (www.leximancer.com), The Universityof Queensland,Australia	C	Any, plagiarism detection	Clustering	Visualization of entities and properties on a map	automatic taxonomy discovery, concept maps, based on Bayesian theory, classification based on text tags using thesaurus, concept ontology
MindServer (www.recommind.com), Recommind Inc., CA,USA	C suite	Any	Feature Extraction Indexing Categorization	Retrieval	word tokenization, language-domain independent, Probabilistic Latent Semantic Analysis

Monarch (www.datawatch.com), Datawatch Corporation, MA,USA	C	Any	Feature Extraction	Retrieval Summarization	transforms any report into a live database based on user-defined data extraction template
Onix FullTextIndexing and Retrieval, ***X Document*** Classifier and Summarizer, (www.languageidentifier.com) Lextek International,UT, USA	C suite	Document and knowledge management	Full text indexing Rule-based Classification	Retrieval Routing Summarization	Identifies 260 languages, stemmers or morphological analyzers
SONIA (Service_for_Organizing_Networked_Information_Autonomously), Digital Libraries, StanfordUniversity(Sahami, Yusufali et al. 1998)	P	Digital Library	Categorization Clustering Feature Extraction	Navigation in topical information space	Multi-tier extracting of terms, statistical clustering, Bayesian classification
TextAnalyst 2.0 (www.megaputer.com/products), Megaputer Intelligence Inc.,USA	C	Engineering, Educational, Customer Documents	Clustering	Navigation/ Visualization Summarization Retrieval	Creating concept-based Semantic Hierarchical Neural Network
VisualText1.5 (textanalysis.com/body_index.html), Text Analysis International Inc., CA,USA	C	Knowledge Engineering	Information Extraction, Categorization	Information Extraction Visualization Summarization	Text Analyzer, NL query, converting text to XML, SQL, text from speech
WebSOM (Kohonen 1999), Helsinki University of Technology, Finland	P	Any	Clustering	Visualization/Navigation	(Online) Text clustering, mapping
GROUP III. Intelligent TM					
SAS Text Miner module in Complete Data Mining Solution (www.sas.com/technologies/textminer), SAS Institute, Germany	C suite	Any: e-mail filtering, news and document routing, predicting of stock prices, etc.	Categorization, Clustering, Feature Extraction	Visualization Retrieval	Concept-based analysis of document collections
SPSS Predictive Text Analytics: Text Mining Builder, TM for Clementine (www.spss.com/predictive_text_analytics), SPSS Inc., Chicago, IL, USA	C suite	Any:	Categorization, Clustering, Feature Extraction		Built predictive model Compare text
Text Analysis tool (readability, word counts, etc.)					
TextQuest1.5 (www.textquest.de/tqe.htm), Textquest, Germany	P	Any	Categorizing by key-word-in-context	Retrieval	Textual content analysis
Bullfighter Delloite Consulting Inc., NY, USA	C	Financial	Keyword		Readability Analysis

The description of academic prototypes come from academic publications, while the description of commercial technologies come mostly from promotional websites, white papers, news portals, and business journals.