Technology of Text Mining

Ari Visa

Tampere University of Technology
P.O. Box 553, FIN-33101 Tampere, Finland
Ari.Visa@tut.fi

Abstract. A large amount of information is stored in databases, in intranets or in Internet. This information is organised in documents or in text documents. The difference depends on the fact if pictures, tables, figures, and formulas are included or not. The common problem is to find the desired piece of information, a trend, or an undiscovered pattern from these sources. The problem is not a new one. Traditionally the problem has been considered under the title of information seeking, this means the science how to find a book in the library. Traditionally the problem has been solved either by classifying and accessing documents by Dewey Decimal Classification system or by giving a number of characteristic keywords. The problem is that nowadays there are lots of unclassified documents in company databases and in intranet or in Internet.
First one defines some terms. Text filtering means an information seeking process in which documents are selected from a dynamic text stream. Text mining is a process of analysing text to extract information from it for particular purposes. Text categorisation means the process of clustering similar documents from a large document set. All these terms have a certain degree of overlapping.
Text mining, also know as document information mining, text data mining, or knowledge discovery in textual databases is an merging technology for analysing large collections of unstructured documents for the purposes of extracting interesting and non-trivial patterns or knowledge. Typical subproblems that have been solved are language identification, feature selection/extraction, clustering, natural language processing, summarisation, categorisation, search, indexing, and visualisation. These subproblems are discussed in detail and the most common approaches are given.
Finally some examples of current uses of text mining are given and some potential application areas are mentioned.

1 Introduction

Nowadays a large amount of information is stored in intranet, internet or in databases. Customer comments and communications, trade publications, internal research reports and competitor web sites are just a few examples of available electronic data. The access to this information is many times organized through the World Wide Web. There are already some commercial tools available that

P. Perner (Ed.): MLDM 2001, LNAI 2123, pp. 1–11, 2001.

are defined as knowledge solutions. The reason is clear; everyone needs a solution for handling the large volume of unstructured information. This is either in intuitive way clear but before a more detailed discussion it is useful to define some phrases and concepts.

We should keep in mind the distinction between data, information and knowledge. These terms can be defined in several ways but the following definitions are useful in Data and Text Mining purposes. The lowest level, data, used to be clear. It is a measurement, a poll, a simple observation. The next level, information, is already more diffuse. It is an observation based on data. We have for instance noticed a cluster among the data or a relation between data items. The highest level, knowledge, is the most demanding. It can be understood as a model or a rule. We know from theory of science that lots of careful planning and experimentation are needed before we can state to know something, we have knowledge.

The phrase document is a more complicated term. It is clear that work gets done through documents. When a negotiation draws to a close, a document is drawn up, an accord, a law, a contract, an agreement. When research culminates, a document is created and published. The knowledge is transmitted through documents: research journals, text books and newspapers. Documents are information and knowledge organized and presented for human understanding. A typical document of today is either printed or electrical one. The printed documents are transferable to electrical ones by optical scanning and Optical Character Recognition (OCR) methods. Tables, figures, graphics, and pictures are problematic under this transform process. The electrical documents are either hierarchical or free. The hierarchical documents use some kind of page description language (PDL), for instance Latex, and imager programs, which take PDL representations to a printable or projectable image. Free documents may contain only free text or free text with tables, figures, graphics, and pictures. Besides the mentioned document types two new types are coming popular: multimedia documents with voice and video in addition to text and pictures, and hyper-media documents that are non-linear documents. In the continuation one concentrates mainly on free text without any insertions as, tables, figures, graphics, pictures, and so on.

The need to manage knowledge and information is not a new one. It has existed as long as the mankind or at least the libraries have existed. Roughly we can say that the key questions are how to store information, how to find it and how to display it. Now we concentrate on the information seeking [3]. First it is useful to overview different kinds of information seeking processes, see Table 1. The presentations are general but we concentrate on the electrical form existing documents. Please, keep in mind that any information seeking process begins with the users' goal. Firstly, information filtering systems are typically designed to sort through large volumes of dynamically generated information and present the user with sources of information that are likely to satisfy his or her information requirement. By information source we mean entities which contain information in a form that can be interpreted by the user. The information

filtering system may either provide these entities directly, or it may provide the user with references to the entities The distinguishing features of the information filtering process are that the users' information needs are relatively specific, and that those interests change relatively slowly with respect to the rate at which information sources become available. Secondly, a traditional information retrieval system can be used to perform an information filtering process by repeatedly accumulating newly arrived documents for a short period, issuing an unchanging query against those documents, and then flushing the unselected documents. Thirdly, another familiar process is the process of retrieving information from a database. The distinguishing feature of the database retrieval process is that the output will be information, while in information filtering, the output is a set of entities (e.g. documents) which contain the information which is sought. For example, using an library catalog to find the title of a book would be a database access process. Using the same system to discover whether any new books about a particular topic have been added to the collection would be an information filtering process. Fourthly, the information extraction process is similar to the database access in that the goal is to provide information to the user, rather than entities which contain information. In the database access process information is obtained from some type of database, while in information extraction the information is less well structured (e.g. the body of an electronic mail message). Fifthly, one variation on the information extraction and database access processes is what is commonly referred to as alerting. In the alerting process the information need is assumed to be relatively stable with respect to the rate at which the information itself is changing. Monitoring an electronic mailbox and alerting the user whenever mail from a specific user arrives is one example of an information alerting process. Sixthly, browsing can be performed on either static or dynamic information sources, browsing has aspects similar to both information filtering and information retrieval. Surfing the World Wide Web is an example of browsing relatively static information, while reading an online newspaper would be an example of browsing dynamic information. The distinguishing feature of browsing is that the users' interests are assumed to be broader than in the information filtering or retrieval processes. Finally, there is a case when one tumbles over an interesting piece of information.

According to ANSI 1968 Standard (American National Standards Institutes, 1968), an index is a systematic guide to items contained in, or concepts derived from, a collection. These items or derived concepts are represented by entities in a known or stated searchable order, such as alphabetical, chronological, or numerical. Indexing is the process of analysing the informational content of records of knowledge and expressing the information content in the language of the indexing system. It involves selecting indexable concepts in a document and expressing these concepts in the language of the indexing system (as index entries) and an ordered list.

Natural Language processing [20] is a broad topic and an important topic for Text Data Mining but here I give only some terms. Stemming is a widely used method for collapsing together different words with a common stem [16]. For

Table 1. Examples of different information seeking processes.

Process	Information Need	Information Sources
Information Filtering	Stable and Specific	Dynamic and Unstructured
Information Retrieval	Dynamic and Specific	Stable and Unstructured
Database Access	Dynamic and Specific	Stable and Structured
Information Extraction	Specific	Unstructured
Alerting	Stable and Specific	Dynamic
Browsing	Broad	Unspecified
By Random Search	Unspecified	Unspecified

instance, if a text includes words Marx, Marxist, and Marxism, it is reasonable to observe the distribution of the common stem Marx instead of three separate distributions of these words. Accordingly, synonymy, hyponymy, hypernymy, and other lexical relatedness of words are detected by using thesauruses or techniques that define semantic networks of words.

Information or in this case text categorisation requires that there are existing categories. There have been several approaches but nowadays in libraries books are classified and accessed according to Dewey Decimal Classification (DCC) [7,6,9]. DCC defines a set of 10 main classes of documents that cover all possible subjects a document could be referring to. Each class is then divided into ten divisions and each division is divided into ten sections. In cases when we do not have existing categories we talk about text clustering. We collect similar documents together, the similarity is defined by a measure.

Data Mining contains the metaphor of extracting ore from rock. In practice Data Mining refers to finding patterns across large datasets and discovering heretofore unknown information. In the same way, as Data Mining can not be accessing data bases, Text Mining can not be finding documents. The emphasis in Information Retrieval is in finding document. The finding patterns in text collections is exactly what has been done in Computational Linguistics. [20]. Text mining, also know as document information mining, text data mining, or knowledge discovery in textual databases is an merging technology for analysing large collections of unstructured documents for the purposes of extracting interesting and non-trivial patterns or knowledge [12]. The aim is to extract heretofore undiscovered information from large text collections.

2 Text Mining Technology

Keeping in mind the evolution to Data Mining and to Text Mining one can state that there are need for tools. In generally tools for clustering, visualisation and interpretation are needed. For instance, under the document exploration tools to organise the documents and to navigate through a large collection of documents are needed. Typical technologies are text categorisation, text clustering, summarisation, and visualisation. Under the text analysis and domain-specific knowledge discovery technologies as question answering, text understanding, in-

formation extraction, prediction, associative discovery, and trend analysis are adequate. I will review most important steps and techniques in Text Mining.

In text mining, as in data mining, there are some initial problems before the work itself can be started. Firstly, some data pre-processing is needed [20]. The key idea is to transform the data into such a form that it can be processed. It might be removal of pictures, tables or text formation or it might contain replacement of mathematical symbols, numbers, URLs, and email addresses with special dummy tokens. If OCR techniques are used the pre- processing step may consist spelling checking. However, during the pre-processing stage some caution is needed, hence it is easy to destroy some structural information. The pre-processing might also be a language related process. In languages as Germany, Finnish, or Russian stemming might be needed. The information of the actual language of the text is very useful. The processing of mono linguistic collection is more straight forward than cross language or the multilingual processing [27,21,22].

It is common that long documents are summarised. The first attempts were made in the 1950's, in the form of Luhn's auto-extracts [19], but unfortunately since then there has been little progress. The reason is easy to understand by defining a summary. A summary text is a derivative of a source text condensed by selection or generalisation on important content. This broad definition includes a wide range of specific variations. Summarising is conditioned by input factors categorising source form and subject, by purpose factors referring to audience and function, and by output factors including summary format and style. The main approaches have been the following: Source text extraction using statistical cues to select key sentences to form summaries [24]. Approaches using scripts or frames to achieve deeper representations and an explicitly domain-oriented kind motivated properties of the world [35]. There has been research combining different information types in presentation. Thus combines linguistic theme and domain structure in source representations, and seeks salient concepts in these for summaries [8].

After the possible summarisation and the pre-processing some kind of encoding is needed. The key questions is the representation of text documents. This question is closely related to feature selection but here the term feature has a broader meaning than in pattern recognition. There are the two main approaches: use of index terms or the use of free text. These approaches are not competing each other but completing. It is common that natural language processing is used to reach index terms. It is possible to proceed directly with index terms and Boolean algebra as one does in information retrieval systems with queries. This is known as Boolean model [2]. The model is binary, the frequency of term has no effect. Due to its uncomplicated semantics and straightforward calculation of results using set operations, the Boolean model is widely used e.g. in commercial search tools. The vector space model is introduced by Salton [29,28] encode documents in a way suitable for fast distance calculations. Each document is represented as a vector in a space, where the dimension is equal to the number of terms in vocabulary. In this model the problem of finding suit-

able documents to a query becomes that of finding the closest document vectors for a query vector, either in terms of distance or of angle. Vector space models underlie the research in many modern information retrieval systems. The probabilistic retrieval model makes explicit the Probability Ranking Principle that can be seen underlying most of the current information retrieval research [20]. For a given query, estimate the probability that a document belongs to the set of relevant documents and return documents in the order decreasing probability of relevance. The key question is, how to obtain the estimates regarding which documents are relevant to a given query. These simple search approaches are similar to association and associative memories. The method is to describe a document with index terms and to build a connection between the index terms and the document. To build this connecting function among other methods artificial neural networks have been used [14].

The use of free text is more demanding. Instead of using index terms it is possible to use other features to represent a document. A common approach is to view a document as a container of words. This is called bag-of-words encoding. It ignores the order of the words as well as any punctuation or structural information, but retains the number of times each word appears. The idea is based on the work of Zipf [36] and Luhn [19]. The famous constant rank-frequency law of Zipf states that if the word frequencies are multiplied by their rank order (i.e. the order of their frequency of occurrence), the product is approximately constant. Luhn remarks that medium-frequency words are most significant. The most frequent words (the, of, and, etc.) are least content-bearing, and the least frequent words are usually not essential for the content of a document either. A straightforward numeric representation for the bag of words-model is to present documents in the vector space model, as points in a t-dimensional Euclidean space where each dimension corresponds to a word of a vocabulary. The i:th component di of the document vector express the number of times the word with index i occurs in the document. The described method is called term frequency document. Furthermore, each word may have an associated weight to describe its significance. This is called term weighting, The similarity between two documents is defined either as the distance between the points or as the angle between the vectors. To consider only the angle discards the effect of document length. Another way to eliminate the influence of document length is to use inverse document frequency this means that the term frequency is normed with document frequency. A variation of the inverse document frequency is the residual inverse document frequency is defined as the difference between the logs of the actual inverse document frequency and inverse document frequency predicted by Poisson model. Another main approach is term distributions models. They assume that the occurrence frequency of words obeys a certain distribution. Common models are the Poisson model, the two-Poisson model, and the K mixture model. The third main approach is to consider the relationships between words. A term-by-document matrix that can be deducted from their occurrence patterns across documents. This notation is used in a method called Latent Semantic Indexing [20], which applies singular-value decomposition to the document-by-word ma-

trix to obtain a projection of both documents and words into a space referred as the latent space. Dimensionality reduction is achieved by retaining only the latent variables with the largest variance. Subsequent distance calculations between documents or terms are then performed in the reduced-dimensional latent space.

The feature selection which means developing richer models that are computationally feasible and possible to estimate from actual data remains a challenging problem. However, facing this challenge is necessary if harder tasks related e.g. to language understanding and generation are to be tackled.

When we have produced either suitable models or gathered the features we are ready to the next step, clustering. Clustering algorithms partition a set of objects into groups or clusters. The methods are principally the same as in Data Mining, but the popularity of algorithms vary [20]. The clustering is one of the most important steps in Text Mining. The main types of clustering are hierarchical and non- hierarchical. The tree of a hierarchical clustering can be produced either bottom-up, by starting with the individual objects and grouping the most similar ones, or top-down, whereby one starts with all the objects and divides them into groups so as to maximise within-group similarity. The commonly used similarity functions are single-link, complete link, and group-average. The similarity between two most similar, or two least similar and average similarity between members is calculated. Non-hierarchical algorithms often start out with a partition based on randomly selected seeds (usually one seed per cluster), and then refine this initial partition. Most non-hierarchical algorithms employ several passes of reallocating objects to the currently best cluster whereas hierarchical algorithms need only one pass. A typical non-hierarchical algorithm is K-means that defines clusters by the centre of mass of their members. We need a set of initial cluster centres in the beginning. Then we go through several iterations of assigning each object to the cluster whose centre is closest. After all objects have been assigned, we recomputed the centre of each cluster as the centroid or mean of its members. The distance function is Euclidean distance [20]. In some case we also view clustering as estimating a mixture of probability distributions. In those cases we use EM algorithm [20]. The EM algorithm is an iterative solution to the following circular statements: Estimate: If we knew the value of a set of parameters we could compute the expected values of the hidden structure of the model. Maximize: If we knew the expected values of the hidden structure of the model, then we could compute the maximum likelihood value of a set of parameters.

Depending on the research task some text segmentation may be needed. It is also called information extraction. In information extraction also known as message understanding, unrestricted texts are analysed and a limited range of key pieces of task specific information are extracted from them. The problem is many times how to break documents into topically coherent multi-paragraph subparts. The basic idea of the algorithm is to search for parts of a text where the vocabulary shifts from one subtopic to another. These points are then interpreted as the boundaries of multi-paragraph units [13].

Finally, it is still important to visualise the features, the clusters, or the documents. Quantitative information has been presented using graphical means [32] since 1980s but during 1990s the scientific visualisation has developed a lot. This development helps us in information seeking, and in document exploration an in management. Quite a lot of has been done in connection to the project Digital Library. Properties of large sets of textual items, e.g., words, concepts, topics, or documents, can be visualised using one-, two-, or three- dimensional spaces, or networks and trees of interconnected objects, dendrograms [31]. Semantic similarity and other semantic relationships between large numbers of text items have usually been displayed using proximity. Some examples of that are the Spire text engine [34], document maps organised with Self Organized Map (SOM) [26,18,17,30], using coloured arcs in Rainbows [10], and with colour coordination of themes in the ET-map [23]. Another approach is to use the visual metaphor of natural terrain that has been used in visualising document density and clustering in ThemeView [34], in WEBSOM [15], and in a map of astronomical texts [25]. Relationships, e.g. influence diagrams between scientific articles, have been constructed based on citations and subsequently visualised as trees or graphs in BibRelEx [5]. Citeseer [4] offers a browsing tool for exploring networks of scientific articles through citations as well as both citation- and text-based similarity between individual articles. Searching is used to obtain a suitable starting-point for browsing. Term distributions within documents retrieved by a search engine have been visualised using TileBars [11]. Visualisation is rapidly developing field also in Text Mining.

3 Some Applications

I introduce briefly three applications of Text Mining.

In the first case we treated the annual reports that contained information both in numerical and in textual form [1]. More and more companies provide their information in electronic form this is the reason why this approach was selected. The numerical data was treated by Data Mining and the textual data by Text Mining. A multi level approach based on word, sentence, and paragraph levels were applied. The interesting point was to find out that the authors seems to emphasis the results even though the numerical facts are not supporting their attitude.

In the second case Text Mining has been used to identify the author [33]. A multi level approach based on word and sentence levels has been applied on database containing novels and poems. For authorship attribution purposes the authors William Shakespeare, Edgar Allan Poe, and George Bernard Shaw were selected. The interesting point was to identify and separate the authors.

In the third case a different approach is taken. In this approach WEBSOM [15] is used to visualise a database with 7 million patent abstracts. This is an typical exploration example and the map offers additional information regarding the results that cannot be conveyed by the one-dimensional list of results.

4 Discussion

Text mining is best suited for discovery purposes, learning and discovering information that was previously unknown. Some examples how text mining are used are: exploring how market is evolving, or looking for new ideas or relations in topics. While a valuable tool, text mining is not suited to all purposes. Just as you would not use data mining technology to do a simple query of your database, text mining is not the most efficient way to isolate a single fact. Text mining is only a support tool. However, text mining is relevant because of the enormous amount of knowledge, either within an organisation or outside of it. The whole collection of text is simply too large to read and analyse easily. Furthermore, it changes constantly and requires ongoing review and analysis if one is to stay current. A text mining product supports and enhances the knowledge worker's creativity and innovation with open-ended exploration and discovery. The individual applies intelligence and creativity to bring meaning and relevance to information, turning information into knowledge. Text mining advances this process, empowering the knowledge worker to explore and gain knowledge from the knowledge base. The text mining delivers the best results when used with information that meets the following criteria: The information must be textual. Numerical data residing within a database structure are best served by existing data mining technologies. The value of text mining is directly proportional to the value of the data you are mining. The more important the knowledge contained in the text collection, the more value you will derive by mining the data. The content should be explicitly stated within the text. Scientific and technical information are good examples of explicitly stated material. It seems that highly structured information already resides within a navigable organisation. Text mining is not as valuable in those cases, provided the structure of the information makes some sense. Text Mining is most useful for unorganised bodies of information, particularly those that have an ongoing accumulation and change. Bodies of text that accumulate chronologically are typically unorganised, and therefore good candidates for text mining.

There are already some commercial techniques and tools for text mining purposes. However, the text mining field is rapidly evolving, the following will guide users in what to consider when selecting among text mining solutions. One should consider the requirements of manual categorisation, tagging or building of thesauri. It is useful if long, labor-intensive integrations are avoided. The automatic identification and indexing of concepts within the text will also save a great deal of work. It is also nice if the tool can present visually a high level view of the entire scope of the text, with the ability to quickly drill down to relevant details. It is also nice if the tool enables users to make new association and relationships, presenting paths for innovation, or exploration and integrates with popular collaborative workflow solutions. Finally, if the tool scales to process any size data set quickly and it handles all types of unstructured data formats and runs on multiple formats.

Acknowledgments

The financial support of TEKES (grant number 40943/99) is gratefully acknowledged.

References

1. B. Back, J. Toivonen, H. Vanharanta, and A. Visa. Toward Computer Aided Analysis of Text. *The Journal of The Economic Society of Finland*, 54(1):39–47, 2001.
2. R. Baeza-Yates and B. Ribeiro-Neto, editors. *Modern Information Retrieval*. Addison Wesley Longman, 1999.
3. D. C. Blair. *Language and Representation in Information Retrieval*. Elsevier, Amsterdam, 1990.
4. K. Bollacker, S. Lawrence, and C. L. Giles. Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications. In *Proceedings of 2nd International ACM Conference on Autonomous Agents*, pages 116–123. ACM Press, 1998.
5. A. Brüggemann-Klein, R. Klein, and B. Landgraf. BibRelEx – Exploring Bibliographic Databases by Visualization of Annotated Content-Based Relations. *D-Lib Magazine*, 5(11), Nov. 1999.
6. M. Dewey. *A Classification and subject index for cataloguing and arranging the books and pamphlets of a library*. Case, Lockwood & Brainard Co., Amherst, MA, USA, 1876.
7. M. Dewey. Catalogs and Cataloguing: A Decimal Classification and Subject Index. In *U.S. Bureau of Education Special Report on Public Libraries Part I*, pages 623–648. U.S.G.P.O., Washington DC, USA, 1876.
8. U. Hahn. Topic parsing: accounting for text macro structures in full-text analysis. *Information Processing and Management*, 26(1):135–170, 1990.
9. S. P. Harter. *Online Information Retrieval*. Academic Press, Orlando, Florida, USA, 1986.
10. S. Havre, B. Hetzler, and L. Nowell. ThemeRiver^{TM}: In search of trends, patterns, and relationships. In *Proceedings of IEEE Symposium on Information Visualization (InfoVis'99)*, San Francisco, CA, USA, Oct. 1999.
11. M. A. Hearst. TileBars: Visualization of Term Distribution Information in Full Text Information Access. In *Proceedings of the ACM Conference on Human Factors in Computing Systems, (CHI'95)*, pages 56–66, 1995.
12. M. A. Hearst. Untangling text data mining. In *Proceedings of ACL'99, the 37th Annual Meeting of the Association for Computational Linguistics*, June 1999.
13. M. A. Hearst and C. Plaunt. Subtopic Structuring for Full-Length Document Access. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 59–68, 1993.
14. V. J. Hodge and J. Austin. An evaluation of standard retrieval algorithms and a binary neural approach. *Neural Networks*, 14(3):287–303, Apr. 2001.
15. T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585, May 2000.
16. T. Lahtinen. *Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods*. PhD thesis, Department of General Linguistics, University of Helsinki, Finland, 2000.

17. X. Lin. Map displays for information retrieval. *Journal of the American Society for Information Science*, 48(1):40–54, 1997.
18. X. Lin, D. Soergel, and G. Marchionini. A Self-Organizing Semantic Map for Information Retrieval. In *Proceedings of 14th Annual International ACM/SIGIR Conference on Research & Development in Information Retrieval*, pages 262–269, 1991.
19. H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165, 1958.
20. C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, USA, 1999.
21. P. Nelson. Breaching the language barrier: Experimentation with Japanese to English machine translation. In D. I. Raitt, editor, *15th International Online Information Meeting Proceedings*, pages 21–33. Learned Information, Dec. 1991.
22. D. W. Oard and B. J. Dorr. A Survey of Multilingual Text Retrieval. Technical Report CS-TR-3615, University of Maryland, 1996.
23. R. Orwig, H. Chen, and J. F. Nunamaker. A Graphical, Self-Organizing Approach to Classifying Electronic Meeting Output. *Journal of the American Society for Information Science*, 48(2):157–170, 1997.
24. C. Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26(1):171–186, 1990.
25. P. Poinçot, S. Lesteven, and F. Murtagh. A spatial user interface to the astronomical literature. *Astronomy and Astrophysics Supplement Series*, 130:183–191, 1998.
26. H. Ritter and T. Kohonen. Self-Organizing Semantic Maps. *Biological Cybernetics*, 61(4):241–254, 1989.
27. G. Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3):187–194, 1970.
28. G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.
29. G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
30. J. C. Scholtes. *Neural Networks in Natural Language Processing and Information Retrieval*. PhD thesis, Universiteit van Amsterdam, Amsterdam, Netherlands, 1993.
31. B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of IEEE Symposium on Visual Languages, (VL)*, pages 336–343, Sept. 1996.
32. E. R. Tufte. *The Visual Display of Quantitative Information*. Graphic Press, 1983.
33. A. Visa, J. Toivonen, S. Autio, J. Mäkinen, H. Vanharanta, and B. Back. Data Mining of Text as a Tool in Authorship Attribution. In B. V. Dasarathy, editor, *Proceedings of AeroSense 2001, SPIE 15th Annual International Symposium on Aerospace/Defense Sensing, Simulation and Controls. Data Mining and Knowledge Discovery: Theory, Tools, and Technology III*, volume 4384, Orlando, Florida, USA, Apr. 16–20 2001.
34. J. A. Wise. The Ecological Approach to Text Visualization. *Journal of the American Society of Information Science*, 50(13):1224–1233, 1999.
35. S. R. Young and P. J. Hayes. Automatic classification and summarisation of banking telexes. In *Proceedings of The Second Conference on Artificial Intelligence Applications*, pages 402–408, 1985.
36. G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, Massachusetts, USA, 1949.