Text Mining: The next step in Search Technology

Finding without knowing exactly what you're looking for, or finding what apparently isn't there.

Inaugural speech given for the acceptance of the position of Adjunct Professor in the Knowledge Engineering Department of the Humanities and Science Faculty at the University of Maastricht

Maastricht, 23 January, 2009

Dr. ir. Jan C. Scholtes

1 Contents

2 Welcome and Introduction

Chancellor, fellow professors, ladies and gentlemen, welcome and thanks for your presence at this inaugural speech in which serves as my acceptance of the position of Adjunct Professor in *Text Mining* at the university of Maastricht.

I am both honoured and grateful that we are here this afternoon and that I have the opportunity to speak to you about the extremely interesting subject of *text mining*. I hope to be able to maintain your attention for the coming hour in a subject that keeps me occupied for the whole day ☺.

Within the speciality subject of *text mining*, sometimes also called text analytics, several interesting technologies meet, such as computers, IT, computational linguistics, cognition, pattern recognition, statistics, advanced mathematic techniques, artificial intelligence, visualisation, and not forgetting information retrieval. These are all subjects which I have in the last 25 years studied with interest and pleasure.

In the coming years I hope to be able to further study those subjects with students and colleagues from the University of Maastricht, so that in a couple of years we can say we've made further progress.

This afternoon my task is to give you all a short tour through my speciality subject, and give you an idea of its future direction. Using a few search examples, I'll first demonstrate in which situations *text mining* technology becomes relevant. Then I'll give more detailed explanation of the various technologies that are of importance to this speciality subject, supported by examples of successful *text mining* applications. There will also be a short discussion on areas within *text mining* that require more research.

The information explosion of recent times will continue at the same rate. You are undoubtedly aware of Moore's Law, named after Gordon Moore, co-founder of Intel and co-inventor of the computer chip; according to Moore, computer processor and storage capacities will double every 18 months. This law has proved true since the 1950s. Because of this exponential growth we could double the amount of information stored every 18 months, resulting in ever-greater *information overload* with ever more difficult information retrieval on one side, but at the same time the development of new computer techniques to help us control this mountain of information on the other. Naturally, these new techniques still have to be developed.

Text mining techniques shall play an essential role in the coming years in this continuing process.

3 What is *Text Mining*

The field of *data mining* is better known than that of *text mining*. A good example of data mining is the analyzing of transaction details contained in relational databases, such as credit card payments or debit card (PIN) transactions. To such transactions various additional information can be provided: date, location, age of card holder, salary, etc. With the aid of this information patterns of interest or behaviour can be determined.

However, 90% of all information is unstructured information, and both the percentage and the absolute amount of unstructured information increase daily. Only a small proportion of information is stored in a structured format in a database. The majority of information that we work with every day is in the form of text documents, e-mails or in multimedia files (speech, video and photos). Searching within or analysis using database or data mining techniques of this information is not possible, as these techniques work only on structured information.

Structured information is easier to search, manage, organize, share and to create reports on, for computers as well as people, hence the desire to give structure to unstructured information. This allows computers and people to manage the information better, and allows known techniques and methods to be used.

Text mining, using manual techniques, was use first during the 1980s. It quickly became apparent that these manual techniques were labour-intensive and therefore expensive. It also cost too much time to manually process the already-growing quantity of information. Over time there was increasing success in creating programs to automatically process the information, and in the last 10 years there has been much progress.

Currently the study of *text mining* concerns the development of various mathematical, statistical, linguistical and pattern-recognition techniques which allow automatic analysis of unstructured information as well as the extraction of high quality and relevant data, and to make the complete text more searchable.

High quality refers here, in particular, to the combination of the relevance (i.e. finding a needle in a haystack) and the acquiring of new and interesting insights.

A text document contains characters that together form words, which can be combined to form phrases. These are all syntactic properties that together represent defined categories, concepts, senses or meanings. *Text mining* must recognize, extract and use all this information.

Using *text mining*, instead of searching for words, we can search for linguistic word patterns, and this is therefore searching at a higher level.

Due to continuing globalisation there is also much interest in multi-language *text mining*: the acquiring of insights in multi-language collections. The recent availability of machine

translation systems is in that context an important development. Multi-language *text mining* is much more complex that it appears as in addition to differences in character sets and words, *text mining* makes intensive use of statistics as well as the linguistic properties (such as conjugation, grammar, senses or meanings) of a language.

We make daily use of *text mining* and other previously-mentioned techniques more often than is realised, often subconsciously. An example: if you search the internet using a search engine, then it is possible that targeted advertisements are shown when you read a certain article. These are, for example, advertisements that have an association with the text in the article, or if you use certain free e-mail providers, the advertisements are associated with the text in the e-mail message. *Text Mining* techniques are used for this.

Text mining therefore goes further than just the computer knowing where you are, what interests you, or your age. It would also be possible to see on a social network which information similar individuals find interesting. This is all still structured information. *Text mining* is about analysing unstructured information and extracting relevant patterns and characteristics.

Using these patterns and characteristics better search results and deeper data analysis is possible, giving quick retrieval of information that otherwise would remain hidden.

There are many application areas using these basic principles, but the most important is “searching for and finding information”, and it is in this area that we will concentrate today.

4 Searching with Computers in Unstructured Information

What happens exactly when someone uses a computer program to search unstructured text? I'll give a quick explanation. Computers are digital machines with limited capabilities. Computers cope best with numbers, in particular whole numbers, also known as *integers*, where speed is important. People are analogue, and human language is also analogue, full of inconsistencies, interference, errors and exceptions. If we search for something then we often think in concepts, senses and meanings, all areas in which a computer cannot directly deal with.

For computers to be able to make a computationally efficient search in a large amount of text, the problem needs first to be converted to a numerical problem that a computer can deal with. This leads to very large containers containing many numbers in which numbers representing search terms are compared with numbers representing documents and information. This is the basic principle that our specialism concerns itself with: how can we translate information that we can work with into information that a computer can work with, and then translate the result back into a form that people can understand.

This technology has existed since the 1960s. One of the first scientists working in this field was Gerard Salton, who together with others made one of the first text search engines. Each occurrence of a word in the text was entered in a keyword index. Searching was then done in the index, comparable to the index at the back of a book but with many more words and much quicker. With techniques such as *hashing* and *b-trees*, it was possible to quickly and efficiently make a list from all documents containing a word or a Boolean (*AND*, *OR* and *NOT* operators) combination of words.

Documents and search terms were converted to vectors and compared using the cosine distance between them: how smaller the cosine distance, the more the search term and the document corresponded. This was an effective method to determine the relevance of a document from the search term. This was called the *vector space* model, and is still used today by some programs.

Later, various other methods for searching and relevance were researched. There are many search techniques with good-sounding names such as: (*directed* and *non-directed*)-*proximity*, *fuzzy*, *wildcards*, *quorum*, *semantical*, *taxonomies*, *conceptual*, etc. Examples of commonly-known relevance defining techniques are: *term-based frequency ranking*, the *page-rank* algorithm (popularity principle), and *probabilistic ranking* (*Bayes classifiers*).

Salton's first important publication was in 1968, now 41 years ago. Have all problems related to searching and finding still not been resolved?, you may ask.

The answer is *no*. Because these days there is so much information digitally available and because it is now often imperative to directly (pro-actively) react on current happenings, new techniques are necessary to keep up with the continuously growing quantity of unstructured information. Furthermore, people will have different reasons for searching large quantities of data and different objectives to find, and those differences require alternative approaches.

5 *Text Mining* in Relation to "Searching and Finding"

The title of this lecture is "*Text Mining*: The next step in Search Technology", with the subtitle "*Finding without knowing exactly what you're looking for"* and *"Finding what apparently isn't there*". How do we do that? Who wants to do it? In other words, what is the social as well as the scientific relevance of this?

And that was also the question I was asked during my application for this professorship: "We already have Google, so why should we need anything else"? "A very good question", was my answer, "because this is exactly what so many others think too". Unfortunately the search problem is not solved and Google does not give the complete answers to your questions. If I can convince you of this in the next 45 minutes then I will have already succeeded in a part of my mission!

The questions I asked can also be asked in another way:
"Do you want to find the best or do you want to find everything?" or "Do you want to find that which does not want to be found?"

5.1 Everything

We are getting closer to the heart of the problem. Internet search engines only give the best answer or the most popular answer. Fraud investigators or lawyers don't only want the *best* documents, they want *all possible* relevant documents.

Furthermore, in an internet search engine everyone does his or her best to get to the top of the results list; search engine optimalization in itself has become an art. Criminals and fraudsters do not want to be at the top of the results list in a search engine; they actively try to hide what they do.

5.2 Finding someone or something that doesn't want to be found

How do they do that? They use synonyms and code names, and quite often these are common words which are used so often that a search cannot be done without returning millions of hits. *Text mining* can offer a solution to finding that relevant information.

5.3 Finding, when you don't know exactly what you want

Fraud investigators also have another common problem: at the beginning of the investigation they do not know exactly what they must search for. They do not know the synonyms or code names, or they do not exactly know which companies, persons, account numbers or amounts must be searched for. Using *text mining* it is possible to identify all these types of entities or properties from their linguistic role, and then to classify them in a structured manner to present them to the user. It then becomes very easy to research the found companies or persons further.

Sometimes the problems confronting an investigator go a little deeper: they are searching without really knowing what they are searching for. *Text mining* can be used to find the words and subjects important for the investigation; the computer searches for specified patterns in the text: "who paid who", "who talked to whom", etc. These types of patterns can be recognised using language technology and *text mining*, and extracted from the text and presented to the investigator, who can then quickly determine the legitimate transactions from the suspect ones.

An example: If the ABN-AMRO bank transfers money to the FORTIS bank then that is a normal transaction. However, if "Big Tinus" transfers money to Bahamas Enterprises Inc. then that is suspicious. *Text mining* can identify these sorts of patterns, and further searches can be made on the words in those patterns using normal search techniques to further indentify and analyse details.

The obtaining of new insights is also called serendipity (finding something unexpected while searching for something completely different). *Text mining* can be adapted very effectively to obtain new but frequently essential insights necessary to progress in an investigation.

We can therefore say the *text mining* helps in the search for information by using patterns for which the values of the elements are not exactly known beforehand. This is comparable with mathematical functions in which the variables and the statistical distribution of the variables are not always known. Here the core of the problem can be seen as a translation problem from human language to mathematics, and the better the translation, the better the quality of the *text mining*.

5.4 Text mining and information visualisation

Text mining is often mentioned in the same sentence as information visualisation. This is because visualisation is one of the technical possibilities after unstructured information has been structured.

An example of information visualisation is the figurative movement chart by M. Minard from 1869 that represented Napoleon's march to Russia. The width of the line represented the total men in the army during the campaign. The dramatic decrease in the army's strength over the advance and retreat can be clearly seen.

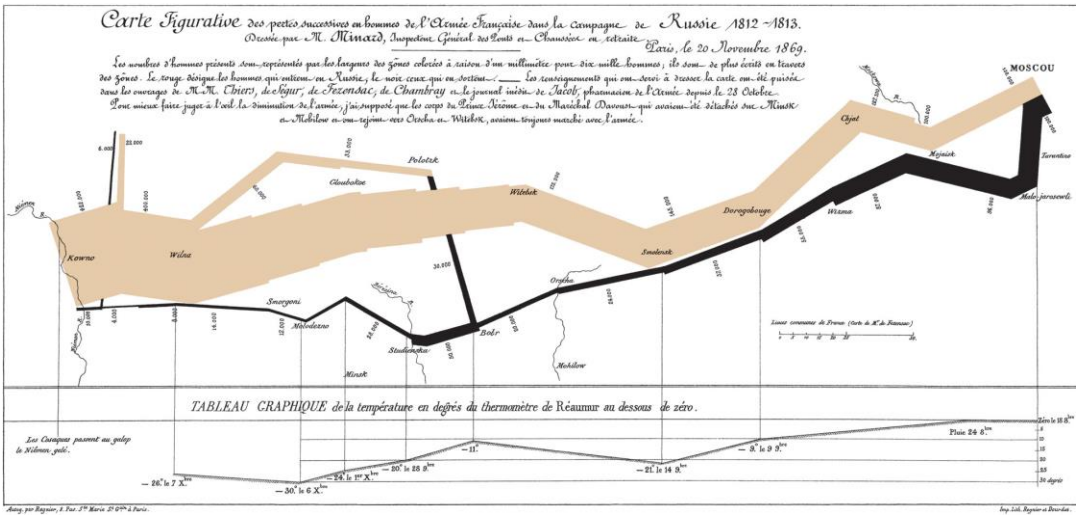


Figure 1: M. Minard (1869): Napoleon's expedition to Russia (source: Tufte, Edward, R. (2001). The Visual Display of Quantitative Information, 2nd edition)

This chart presents a quicker and clearer picture than would just a row of figures. That is a concise summary of information visualisation: a picture says a thousand words.

To be able to make these sorts of visualisations the details must be structured, and that is exactly the area in which *text mining* technology can help: by structuring unstructured information it is possible to visualise the data and more quickly obtain new insights.

An example is the following text:

ZyLAB donates a full ZylMAGE archiving system to the Government of Rwanda

Amsterdam, The Netherlands, July 16th, 2001 -ZyLAB, the developer of document imaging and full-text retrieval software, has donated a full ZylMAGE filing system to the government of Rwanda.

"We have been working closely with the UN International Criminal Tribunal in Rwanda (ICTR) for the last 3 years now," said Jan Scholtes, CEO of ZyLAB Technologies BV. "Now the time has come for the Rwanda Attorney General's Office to prosecute the tens of thousands of perpetrators of the Rwanda genocide. They are faced with this long and difficult task and the ZyLAB system will be of tremendous assistance to them. Unfortunately, the Rwandans have scarce resources to procure advanced imaging and archiving systems to help them in this task, so we decided to donate them a full operational system."

"We greatly thank you for this generous gift," says The Honorable Gerald Gahima, the Rwandan Attorney General. "We possess an enormous evidence collection that will require scanning so we can more effectively process, search and archive the evidence collection."

A demonstration of the ZyLAB software was done for the Rwandans by David Akerson of the Criminal Justice Resource Center, an American-Canadian volunteer group: "The Rwandans were greatly impressed. They want and need this system as they currently have evidence sitting in folders that is difficult to search. This is one of the major delays in getting the 110,000 accused persons in custody to trial."

"My hope and belief is that ZylMAGE will enable Mr. Gahima's office to process, preserve and catalogue the Rwandan evidence collection, so that the significance and details of the genocide in Rwanda can be preserved," Scholtes concludes.

In that text, the following entities and attributes can be found:

Places	Amsterdam
Countries	The Netherlands, Rwanda
Persons	Jan Scholtes, Gerald Gahima, Mr. Gahima's, David Akerson, Scholtes
Function titles	CEO, Rwandan Attorney General
Data	July 16th, 2001
Organisations	UN International Criminal Tribunal in Rwanda (ICTR), Government of Rwanda, Rwanda Attorney General's Office, Criminal Justice Resource Center, American-Canadian volunteer group
Companies	ZyLAB, ZyLAB Technologies BV
Products	ZyIMAGE

Let's assume that we have various documents containing this type of automatically-found structured properties; then the documents could not only be presented in table form, but also for example in a tree structure in which the document could be organised on occurrences per land and then on occurrences per organisation. This could then be loaded into, for example, a *Hyperbolic Tree* or in a so-called *TreeMap*

Both give the possibility to zoom in on the part of the tree structure that is of interest, without losing the whole picture.

A good example of a reproduction of a hyperbola (the principle on which the *Hyperbolic Tree* is based) can be found in the work of the Dutch artist M.C. Escher. Here a two-dimensional object is placed on a sphere where the centre is always zoomed-in and the edge is always zoomed-out.

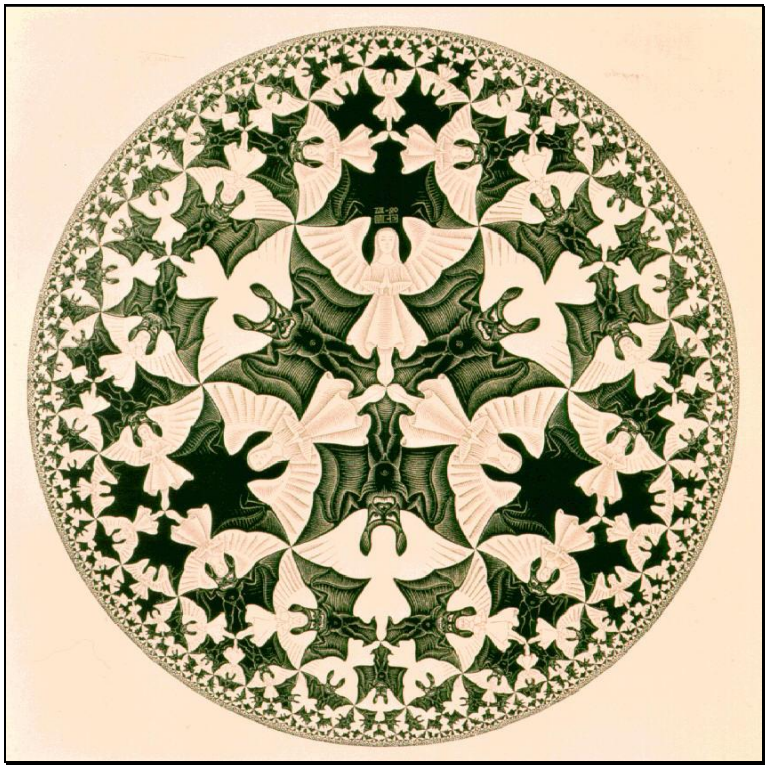


Figure 2: M.C. Escher: Circle Limit IV 1960 woodcut in black and ochre, printed from 2 blocks (source: http://www.mcescher.com/)

That principle can also be used to dynamically visualise a tree structure, which would then appear as follows:

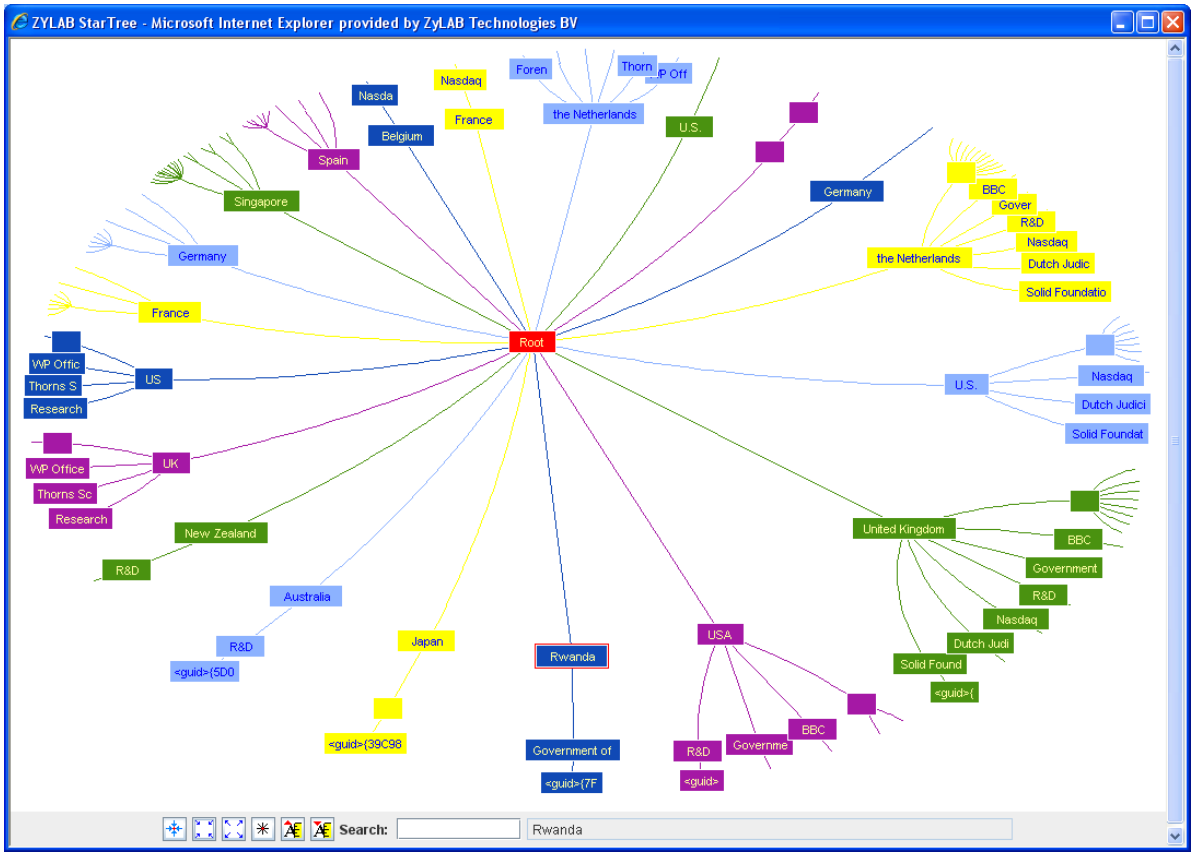


Figure 3: Hyperbolic Tree visualisation of a tree structure (source: ZyLAB Technologies BV)

Another method of displaying a tree structure is in a TreeMap, introduced by Ben Schneiderman in 1992. Here a tree structure is projected on an area, and the more leaves a branch has then the greater the area is allocated to it. This allows you to quickly see the area with the most entities. A value can also be allocated to a certain type of entity, for example the size of an e-mail or a file.

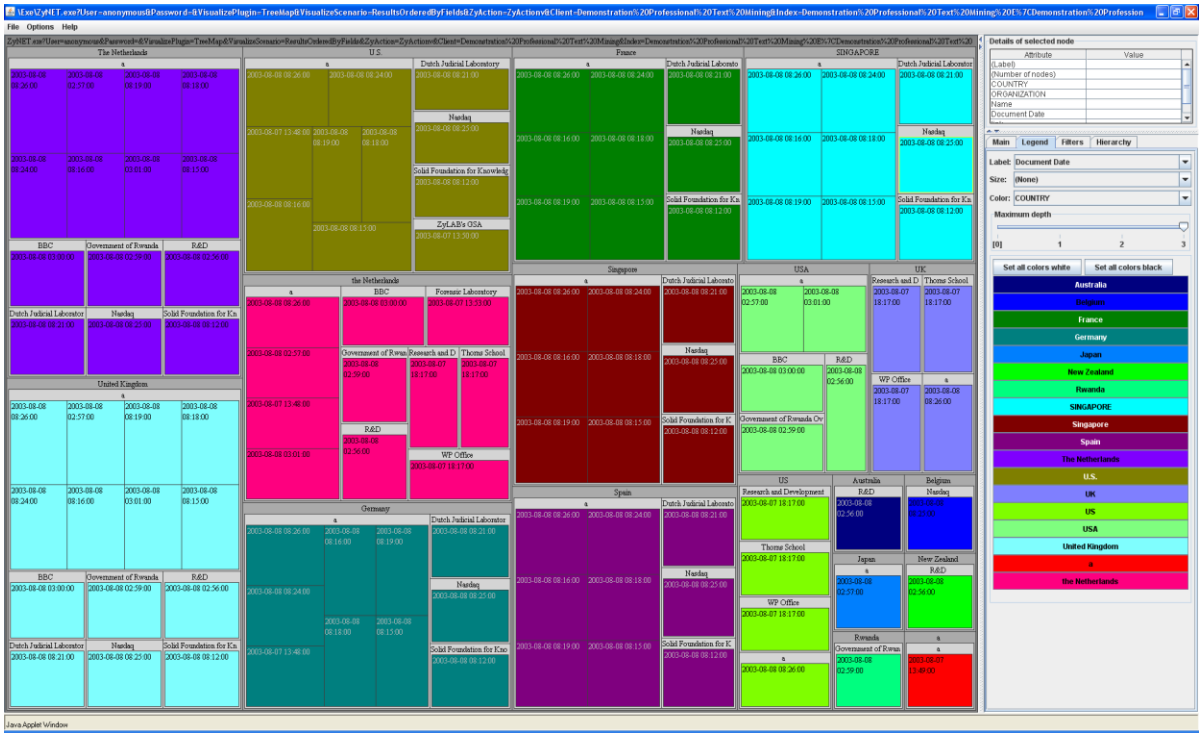


Figure 4: TreeMap visualisation of a tree structure (source: ZyLAB Technologies BV)

These types of visualisation techniques are ideal for allowing an easy insight into large e-mail collections. Alongside the structure that *text mining* techniques can deliver, use can also be made of the available attributes such as “Sender”, “Recipient”, “Subject”, “Date”, etc. Below, a number of possibilities for e-mail visualisation are shown.

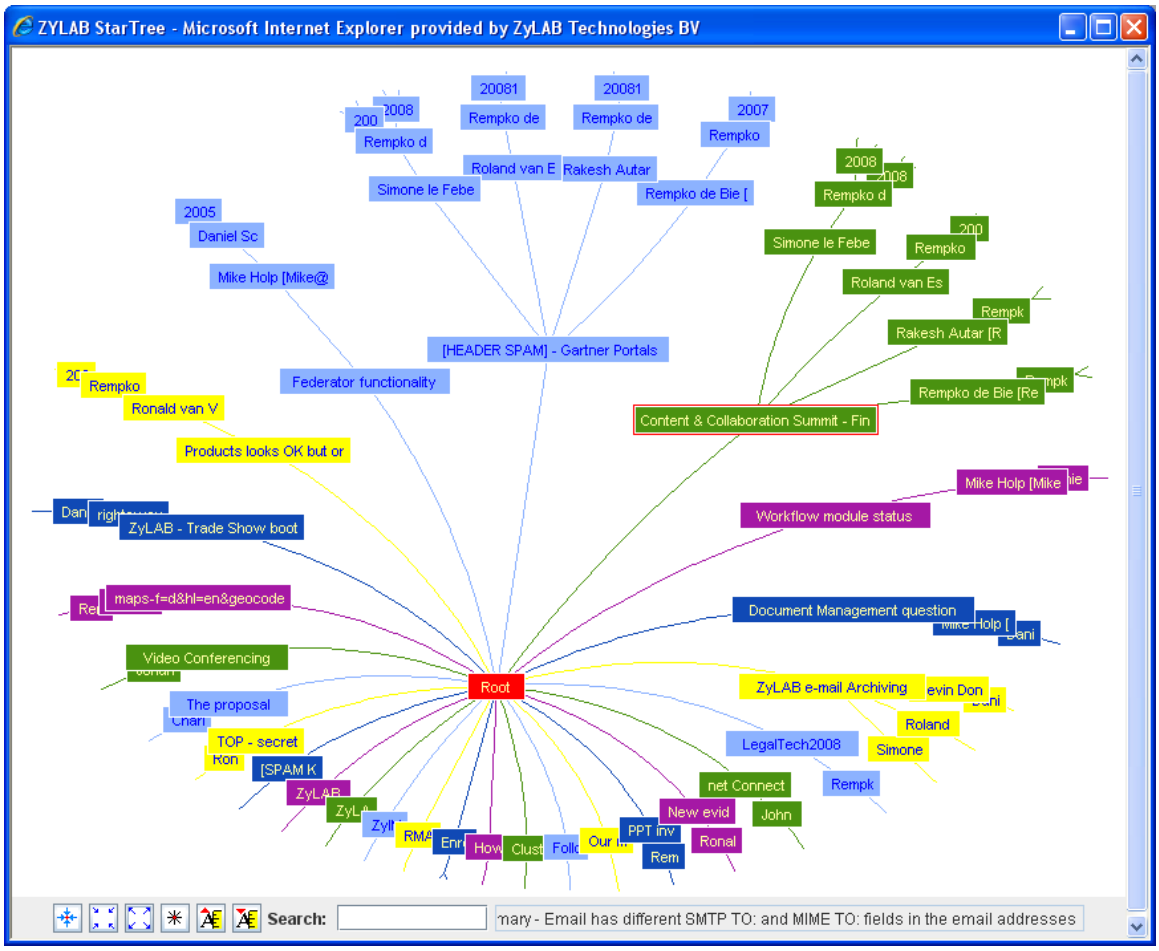


Figure 5: E-mail visualisation using a Hyperbolic Tree (source: ZyLAB Technologies BV)

With the help from these types of visualisation techniques it is possible to gain a quicker and better insight into complex data collections, especially if it involves large collections of unstructured information that can be automatically structured using data mining.

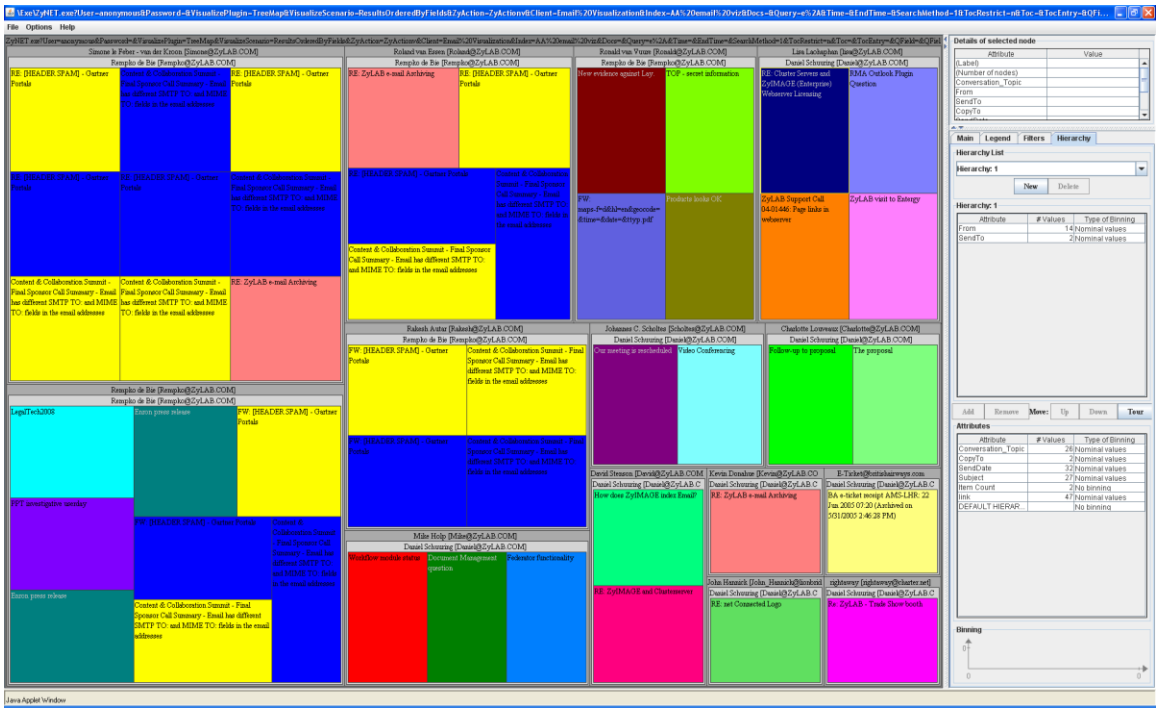


Figure 6: E-mail visualisation using a TreeMap (source: ZyLAB Technologies BV)

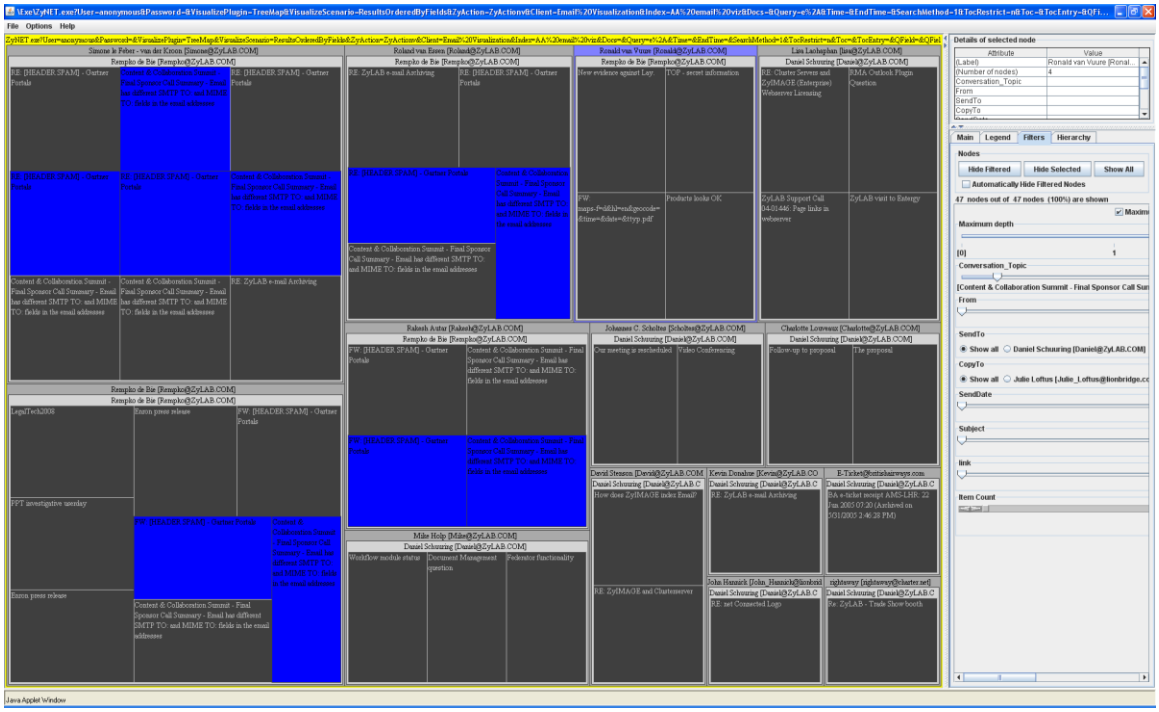


Figure 7: E-mail visualisation using a TreeMap in which all messages from one e-mail conversation are marked in the same colour: it can be immediately seen who was involved in that conversation (source: ZyLAB Technologies BV)

5.5 Other advantages of structured and analysed data

In addition to the visualisation mentioned above, various other applications are possible when the data has been structured and has meta-details. Here is a brief summary.

- Details are easier to arrange in folders.
- It is easier to filter data on specified meta-details when searching or viewing.
- Details can be compared, and linked using the meta-details (vector-comparison of meta-details)
- It is possible to sort, group and prioritise the documents using any of the attributes.
- Details can be clustered using the meta-details.
- With the help of meta-details duplicates and almost-duplicates can be detected. These can then be deleted or relocated.
- Taxonomies can be derived from the meta-details.
- So-called *topic* analyses and *discourse* analyses can be created using the meta-details.
- Rule-based analyses can be made on the meta-details.
- It is possible to search the meta-details from already-found documents.
- Various (statistical) reports can be made on the basis of the meta-details.
- It is possible to search for relationships between meta-details, for example: “who paid who how much”, in which the “who” and the “how much” are not previously known.

There are applications for these techniques in various speciality fields.

The following section examines everyday applications for *text mining* technology. Afterwards, we shall briefly examine the techniques necessary for successful *text mining* applications.

6 Examples of *Text Mining* Applications

6.1 Fraud, crime detection, and intelligence analyses

It is obvious that *text mining* has important application possibilities in fraud investigations, crime detection, intelligence analyses, and other similar areas. It must also be stated that *text mining* had its origins in these sorts of applications, and that these days it is not possible to work successfully and efficiently without *text mining* technology.

A nice example is given below. Both individual entities (first six, from *Person* to *Weapon*) and patterns of entities (last seven) are recognized and highlighted in specific colours. The last seven recognised are especially interesting. For these it was possible to recognise certain interesting linguistic patterns using *text mining* without first having to know the exact values of those entities that occur in the text. Thus, this is "searching without first knowing exactly what you are looking for".

FAWIZ AL (RABBATI) PURCHASED TEN 1-TON TRUCKS (NFI) AND GETS SMUGGLERS TO CROSS THE BORDER APPROXIMATELY 10-15 KILOMETERS OUTSIDE OF KHASON. RABBATI RECRUITED JAN ANTON KRACZEWKI (AL-KIELBASA) TO WORK FOR HIM. KRACZEWKI IS APPROXIMATELY 53 YEARS OLD, AND 180 CENTIMETERS (CM) TALL. HE DRIVES A FOUR-DOOR 1984 GREEN SUBARU. KRACZEWKI USES HIS BACKGROUND AS AN ELECTRICIAN TO CREATE SOPHISTICATED BOMBS.

Person	FAWIZ AL (RABBATI), RABBATI, JAN ANTON KRACZEWKI (AL-KIELBASA), KRACZEWKI
Vehicle	TEN 1-TON TRUCKS, FOUR-DOOR 1984 GREEN SUBARU
Person_Common	SMUGGLERS, ELECTRICIAN
Measure	10-15 KILOMETERS, 150 CENTIMETERS (CM)
City	KHASON
Weapon	SOPHISTICATED BOMBS
Buy Artifact	FAWIZ AL (RABBATI) PURCHASED TEN 1-TON TRUCKS (NFI)
Travel across Border	SMUGGLERS TO CROSS THE BORDER APPROXIMATELY 10-15 KILOMETERS OUTSIDE OF KHASON
Recruit	RABBATI RECRUITED JAN ANTON KRACZEWKI ((AL-KIELBASA))
Person Appearance: Age	KRACZEWKI IS APPROXIMATELY 53 YEARS OLD
Person Appearance: Height	KRACZEWKI IS 180 CENTIMETERS (CM) TALL
Person Attributes: Vehicle	HE (KRACZEWKI) DRIVES A FOUR-DOOR 1984 GREEN SUBARU
Make Artifact	KRACZEWKI USES HIS BACKGROUND AS AN ELECTRICIAN TO CREATE SOPHISTICATED BOMBS

Figure 8: Example of an analysis of entities and patterns in a typical anti-terrorism application (source: Inxight Software, Inc.)

Similar examples can be presented for fraud investigation, analysis of large international and complex criminal organisations, and for the investigation and prosecution of war crimes at international Courts of Justice.

There is a wide interest in *text mining* for these sorts of applications, and they more or less indispensible in a modern society.

6.2 Sentiment mining and business intelligence

There are, however, also other areas in which *text mining* is of relevance. Consider sentiment mining: for companies and organizations, it is increasingly important to know what positive and, especially, negative material is being written about them on the internet. Simply searching on the company name is now not sufficient: there are too many results. Searching on every possible negative phrase is also not feasible: there are just too many combinations. *Text mining*, and especially sentiment mining, offer a solution: define patterns of positive and negative phrases and let *web crawlers* search for them. At present these techniques are often used to give an early indication of potential PR problems following a product introduction.

The following example shows how using word patterns found in film reviews can be used to create a positive, negative or neutral score for a film.

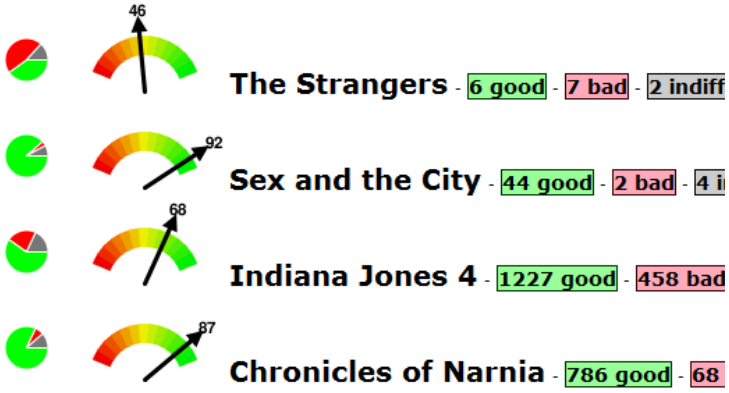


Figure 9: Source: Twitter Movie Reviews. www.twittercritic.com

Various *open source* word lists and semantic models are available that depict the sentiment of words and phrases.

This form of *text mining* is also frequently used for analysing opinions in *blogs*, news groups, websites, social networks and other internet sources, for example for company shares, new products, measuring customer service quality, and analysing the level of satisfaction of hotel guests.

Naturally, alongside information about your own company it is also possible to collect information about colleagues and market competitors: this is called *business intelligence*. In recent years, *text mining* technology has seen an increasing application in this area.

6.3 Clinical research and other biomedical applications

The need to be able to search for patterns where it is not known in advance exactly how they appear occurs frequently in the pharmaceutical industry. During research into the effect of new medicines or treatments, patterns in reported side effects must be found in tens of thousands of medical observations (often in a large amount of documents). Also here it is impossible to know in advance all occurrences of words that must be searched for or for which an *alert* must be given.

There are many reports in which early discoveries of new treatments were possible due to *text mining* technology, saving time and money on fruitless research.

Another application is for the analysing of medical scientific publications. This provides the possibility to analyse trends and to predict or determine who the most important authors, and therefore leaders, are in a given medical specialism. A slightly less ethical application is for some pharmaceutical companies to canvas the “leaders” in a field to act as lobbyists for their medicines and treatments.

6.4 Preventing guarantee problems

One of the first successful commercial applications of *text mining* in the business world was the analysing of guarantee problems in the car industry, and in consumer electronics. In this application, the dealer repair reports were analysed to get an early indication of repetitive patterns in guarantee repairs. These types of problems result in free repairs, or even free replacement of the product. Therefore the quicker the production process is changed to prevent these problems the better.

Often, patterns from internal repair reports are combined with patterns from consumer opinions from the internet and with e-mail communication from the helpdesk or a users’ internet forum. There are many success stories in which *text mining* has saved millions in guarantee costs.

6.5 Spam filters

Text mining technology is also used in spam filters which employ various e-mail characteristics to determine if an e-mail is spam or other unwanted material. Because filtering on a few words is often insufficient, and because spammers continuously develop new techniques to get around spam filters, *text mining* has become a powerful new tool.

6.6 The credit crisis: e-discovery, compliance, bankruptcy and data rooms

The next few years will see the most extensive application of data mining in two relatively new areas: *e-discovery* and *compliance*. Associated with these are the cognate areas of bankruptcy settlements, *due diligence* processes, and the handling of *data rooms* during a takeover or a merger.

6.6.1 E-discovery

At the present time, financial institutions have many problems due to the credit crisis. *Text mining* can help in two of those by limiting the costs of investigation and legal procedures.

Firstly, the administrators will want to know exactly what went wrong and who were responsible. Did companies know at an early stage, for example, what the situation was and that they willingly continued in the wrong direction?

The greatest problem when answering questions from administrators is that it must be exactly known what occurred in the organisation, and frequently information about specific types of transactions or constructions on specific dates is requested, under threat of high fines or prison sentences. Because it is problematic to determine where to search, there is often little choice but to have a specialist read all available information. This is, of course, very expensive and can take a long time.

With the help of *text mining* technology it is easier to present, within the requested time limit, relevant information obtained by letting a computer identify patterns of interest, which, when found, can be further searched.

Furthermore, shareholders, affected larger financial institutions and other involved organisations will also be filing charges and claims. Under American laws, it is permitted for opposing parties to request all potentially relevant information: this is called a *subpoena*, after which a *discovery* process occurs. This law is not only applicable to American companies, but also to *every* organisation that directly or indirectly conducts business in the United States.

10 to 20 years ago there was not nearly as much electronic information in existence, and in many instances it was sufficient during a *discovery* to supply a limited amount of paper information.

These days organisations have hundreds of gigabytes, and sometimes tens of terabytes, of completely unstructured electronic data on hard disks, back-up tapes, CDs, DVDs, USB sticks, e-mail, telephone systems (*voice mail*), etc. *E-discovery* is spoken of instead of just *discovery*. In recent years the costs related to this sort of investigation have, just like the quantity of information, seen an enormous growth.

An extra complication in *e-discovery* is confidential data: before information can be transferred to a third party all confidential and so-called *privileged* data must first be removed or made anonymous (*redaction*). For this, it is often not known what type of information must be searched for: social security numbers, employees' medical files, correspondence between lawyer and client, confidential technical information from a supplier or customer, etc.

Thus, documents must be searched when it's not exactly known what the content is or where it can be found. Often a resort was found in a linear *legal review* by an (expensive) lawyer, and the costs associated with that run quickly into millions.

Great savings can be made using *text mining*. A considerable part of the *legal review* can be done automatically. Additionally, with the help of *text mining* it is possible to make an *early-case assessment* to estimate the real extent of the problems, which can be important when the parties want to make a quick settlement.

6.6.2 Due diligence

In this context is the application for *due diligence* (analysing relevant company data before a takeover) is also of interest. For a *due diligence* process, frequently *data rooms* are created containing many hundreds of thousands of pages of relevant contracts, financial analyses, budgets, etc.

In many cases a buyer must, in a very short space of time, make a decision to take over a company or not. It is often not possible to analyse all data in a data room in the allotted time, and *text mining* technologies can help here.

6.6.3 Bankruptcy

Another application that is seen more and more is for its use in support of an administrator after a large *bankruptcy*. In many situations an administrator must determine whether the board of a bankrupt company has handled all creditors (including the company itself) equally (for example, having paid a board member's salary, but not those of the employees), and the administrator must investigate if there are other irregularities.

Also with bankruptcies, more and more frequently the greatest quantity of information is in the form of unstructured e-mails, hard disks full of data, and other similar data.

6.6.4 Compliance, auditing and internal risk analysis

We shall see the final application in this context in the future as major legislation changes and stricter control systems that will undoubtedly take place in the short term; companies will have to carry out on a more regular basis (*real time*) internal preventative

investigation, deeper audits, and risk analyses. *Text mining* technology will become an essential tool to help process and analyse the enormous amount of information on time.

7 The technology behind *Text Mining*

7.1 Introduction

A typical *text mining* application consists of the following parts:

1. The first sub-system in which *pre-processing* of the data occurs before it can be processed, enriched, visualized, and analysed. These are parts such as *full text* indexing, Natural Language Processing (*NLP*) techniques, statistical techniques, and corpus-based analysis techniques.
2. A second subsystem in which the actual (*core*) *text mining* operations, such as clustering, searching for patterns, categorisation, and information extraction, occur. This is also known as *knowledge extraction* or *knowledge distillation*.
3. A third sub-system comprises the *presentation layer* which allows users of the system to actively further enrich and organise data using navigation, visualisation and other techniques (and also manually), as well as performing quality control.

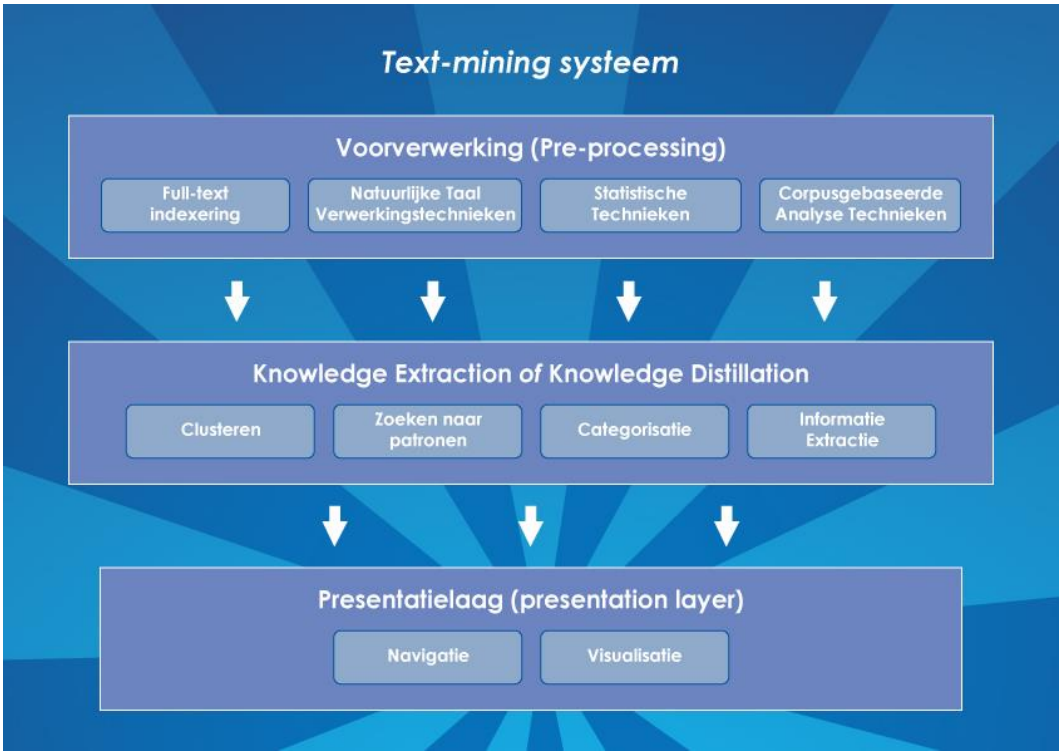


Figure 10: A typical text mining system

Below are detailed descriptions of the three sub systems.

7.2 Pre-processing

In the first phase of *text mining* a few basic steps must first be performed. The goal of pre-processing is to bring the fully unstructured and often multi-format information back to a common denominator. A number of forms of pre-processing are (in a non-specific order):

- Scanning paper documents to digital files.
- Creating text from a bitmap (*Optical Character Recognition: OCR*)
- Extraction of compressed and grouped files (ZIP, e-mail).
- Format and character set recognition of electronic files (PDF, HTML, ASCII, ANSI, UNICODE, MS-Word, etc.).
- Recognition of the spoken parts in multimedia files.
- Extraction of text (in the correct sequence) from a document.
- Recognition of the language used in a document, page, or paragraph.
- Machine translation of a file's content.
- Automatic creation of summaries.
- Deletion of words (*tokens*), noise words, punctuation, and other marks
- Building of the *inverted-file full-text* index.
- Recognition of linguistic sentences.
- Identification of word derivations and removal of conjugations.
- Determine a sentence's grammatical structure using statistical or linguistic techniques.
- Recognition of linguistic references ("*the man walks with his dog, he sees an aeroplane flying*").
- Identification of *named entities*.
- Identification of synonyms, homonyms, abbreviations, expressions (idioms), and other linguistic variants.

Often different techniques can be used in each sub-task. Thus it is possible to do grammatical analysis using grammatical techniques (classical sentence structure with lines), statistical techniques (*hidden-Markov* sequences and other techniques from *machine learning*), and also corpus-based techniques, or even a combination of the three techniques.

For *text mining* a very in-depth analysis is often not necessary; a reasonably light analysis can be sufficient to identify the most important elements of a sentence: the subject clause, the verb clause, potential proper nouns, references, and other relationships. In many cases *finite-state parsers* or *shallow parsers* can be used with the support of dictionaries. These analyses are also commonly known as a *part-of-speech* (*POS*) analysis.

With help from a corpus (a large collection of already-organized data) it is possible to yield statistical probabilities for linguistic manifestations.

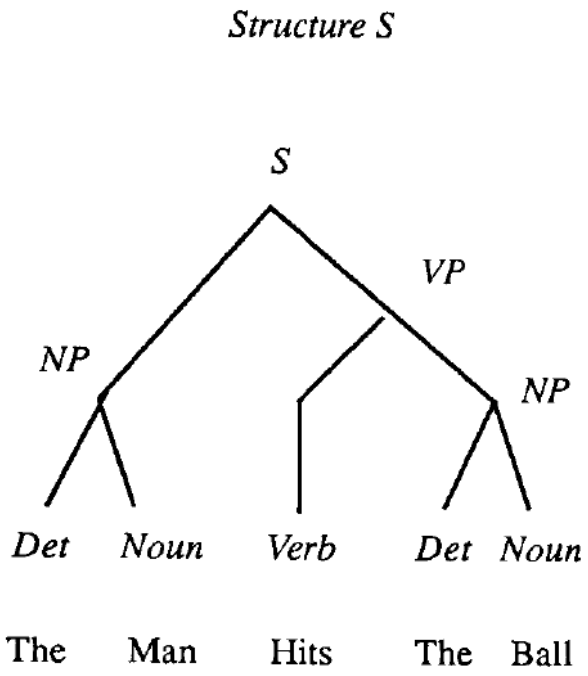


Figure 11: A part-of-speech analysis of a simple sentence (source: Scholtes, 1993)

The result from the *pre-processing* phase is a uniform data format in which the text has been given various statistical, line-based and linguistic attributes which are required in the next phase: the *core text mining.*

In this phase a balance must always be found between the quality and quantity of the addition enrichments on one side, and the speed at which this occurs on the other. Often gigabytes or even terabytes of text must be processed, and this must stay within an acceptable processing time.

It is also important that the processes used are robust (so that they can handle data which contains errors and obscurities: spelling errors, scan errors, write errors, typing errors, new jargon, unknown abbreviations, etc.). Statistical and corpus based techniques are frequently preferred above line-based grammatical techniques as they are both more robust and quicker. The availability of large (*open source*) corpora and the enormous amount of digital text on the internet makes the choice of those techniques over grammatical techniques only more logical and easier.

7.3 Core text mining

The *knowledge discovery*, also called *knowledge distillation*, occurs in the *core text mining* phase. Here, entities (principally user-specific words) are classified, patterns or sentiments are identified, and documents can be clustered or categorised. The following is a short description of the various techniques not discussed earlier.

Many of these techniques originate from classical pattern recognition, only the input in *text mining* is from data in textual form that must first be converted to numerical values. This conversion from other forms of data to numerical values is known as *feature selection* and *feature extraction*. The trick is often to select and measure the best features from a given object, then extract the most noteworthy features from those features (this is often a dimension reduction).

This is a problem that affects not only *text mining* but also speech recognition, image manipulation, and other applications in which a computer cannot handle the original analog data: the data must first be converted to a mathematical representation using advanced techniques. Then the traditional pattern recognition algorithms such as *probabilistic classifiers*, *tree classifiers*, *decision rule classifier*, neural networks, clustering (*k-means*, *nearest neighbour*, *gather/scatter*), *hidden Markov models*, etc., can be used.

In *text mining* these pattern recognition techniques are therefore used on the original textual information, on linguistic features collected in the pre-processing phase, and on other contextual information that is of importance. This can be information that is implicitly as well as explicitly present; this is also called domain knowledge: complimentary knowledge about a specific problem, specialism or (temporary) situation.

7.3.1 Information extraction

In the extraction of information, the following basic text elements can be identified:

1. *Entities*: the basis units that can be found in a text; for example: people, companies, locations, products, medicines, and genes.
2. *Attributes*: these are the properties of the found entities: consider function title, a person's age and social security number, addresses of locations, quantity of products, car registration numbers, and the type of organisation.
3. *Facts*: these are relationships between entities, for example, a contractual relationship between a company and a person.
4. *Events*: these are interesting events or activities that involve entities, such as: "one person speaks to another person", "a person travels to a location", and "a company transfers money to another company".

7.3.1.1 Entities and attributes

The first research into *named entity extraction* was subsidized by an American *Defense Advanced Research Project Agency* (DARPA) in 1995, and was conducted under the name of the *Message Understanding Conference* (*MUC-6*). One of the tasks of this research was to find all occurring persons, locations, organisations, times, and quantities in random text (often, messages or press releases were used). Because it was not known beforehand which features the text would contain, it was necessary to first make a linguistic analysis of the text from which the attributes (*named entities*) could be identified, and then to classify those features into possible categories using different techniques.

One of the ways of doing this is with the help of regular expressions, which allow data, telephone numbers, internet addresses, bank account numbers, and social security numbers to be fairly accurately identified.

A good example of a regular expression to find an e-mail address is:

```
\b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b
```

Figure 12: Example of a regular expression (source: http://www.regular-expressions.info/examples.html)

It is quite complex, and also quite a lot of work, to define these types of regular expressions, especially because there are many pattern variations; it is not always possible to cover everything with one simple regular expression: either very complex patterns are created, or there are always entities that, due to many irregularities, cannot be identified using regular expressions. There have to be many similar (regular) patterns (the name says it all) in entities for regular expressions using this technology to be able to successfully classify. A recent book with the title "*Practical Text mining with Perl"*, describes how the use of regular expressions were taken to the extreme, and how much was achieved [Bilisoly, 2008].

Another logical approach is to compare the longest occurring form of a *named entity* with known *named entities* in dictionaries that contain up-to-date information about what type of entity a specific word or word-combination is.

The linguistic probability that a specific word has a specific meaning can also considered. This is where *hidden-Markov* models (*HMM*) play an important role.

Within *text mining* a *hidden-Markov* model is used to determine what the chance is that, for example, a surname follows a title such as Mr., Mister, Mrs., or Dr. Various relationships between words and their contexts can be formally fixed in this manner. The probabilities can be automatically derived from large corpora using example texts. Using a similar model it is possible to determine the linguistic probability that, for example, an entity is a location or a person's name. A good example of this context is the entity "Mr. Holland", which is immediately obvious to people that it refers to a person and not a

location. With help from a *hidden-Markov* model, an algorithm can also quickly make the same correct decision.

The great advantage of this technique is that only a minimal knowledge of the underlying language is necessary.

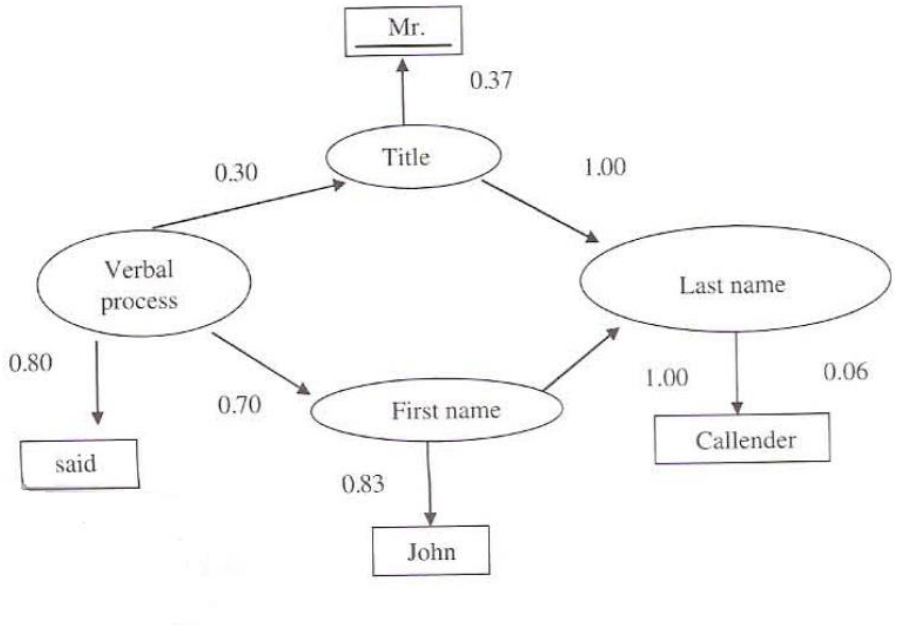


Figure 13: An example of a hidden-Markov model for named-entity recognition (source: Moens, 2006)

Naturally it is also possible to combine these techniques and to identify entities using the most probable classification. Frequently, more than 90% of the available attributes and entities are identified using a combination of the techniques.

7.3.1.2 Facts

Rules can be useful in the recognition of facts (relationships between entities and their attributes). The figure below shows a model that, using all identified entities as personal names, can identify synonyms or aliases from the linguistic context.

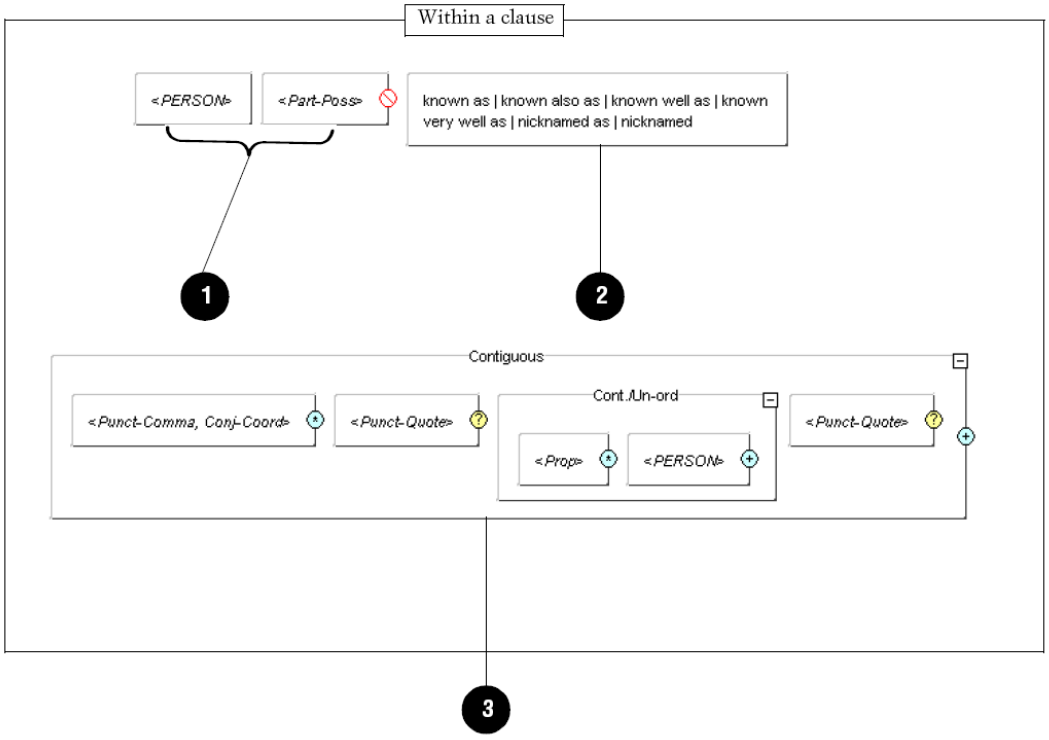


Figure 14: A rule for identifying aliases from personal names (source: Inxight Software Inc.)

7.3.1.3 Events

For the discovery of an event the relationships between entities are identified. This is one of the most interesting forms of pattern recognition because the very complex patterns can identify events such as “a person talks to another person”, “a person travels to a location”, and “a company transfers money to another company”.

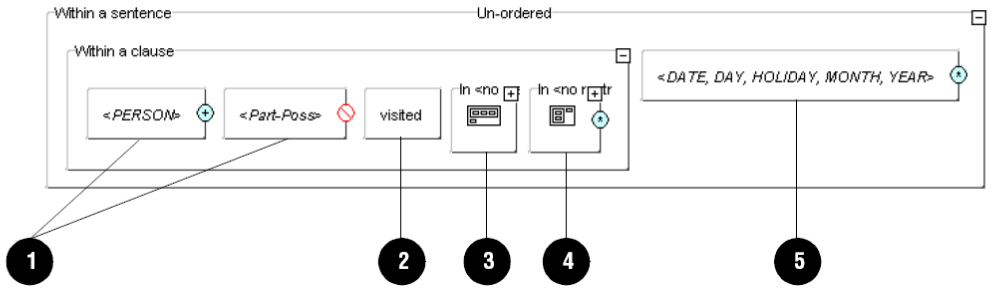


Figure 15: A rule to discover who visits whom on which day (source: Inxight Software Inc.)

One of the biggest problems in the discovery and identification of events is the resolving of the so called *anaphora* and *coreferences*. This is the linguistic problem to associate pairs of linguistic expressions that refer to the same entities in the real world. Research into this problem was first done in MUC-6 (1995) and MUC-7 (1998).

Consider the following text:

> “A man walks to the station and try to catch the train. His name is Jan Jansen. Later he meets his colleague, who has just bought a card for the same train. They work together at the Rail Company as technical employees and they are going to a meeting with colleagues in Utrecht.”

The text contains are various references and coreferences. Various *anaphora* and coreferences will have to be disambiguated before it is possible to fully understand and extract the more complex patterns of events. The following list shows some examples of these (mutual) references:

- *Pronominal Anaphora*: he, she, we, oneself, etc.
- *Proper Name Coreference*: For example, multiple references to the same name.
- *Apposition*: the additional information given to an entity, such as “Jan Jansen, the father of Piet Jansen”.

- *Predicate Nominative*: the additional description given to an entity, for example "Jan Jansen, who is the chairman of the football club".
- *Identical Sets*: A number of reference sets referring to equivalent entities, such as "Ajax", "the best team", and the "group of players" which all refer to the same group of people.

There are various ways of approaching these problems: (i) with an in-depth linguistic analysis of a sentence, or (ii) using a large already-annotated corpus. Both techniques have their advantages and disadvantages. Much research is required in this area in the coming years to improve the quality of these types of analyses.

Also, analysing over periods of time and the tracing of subjects across multiple documents and larger collections can be placed within this context. The analysing of e-mail collections can especially be of interest.

7.3.1.4 Sentiments

The concept of *sentiment mining* was introduced earlier. With the aid of the adjectival adjectives and verbs used, the sentiment of a document can be determined as positive, negative, or neutral. This is mostly done by comparing the words used in the text with a table containing the sentiment values of the words.

Unfortunately this technique is neither as reliable nor so advanced as the previously mentioned extraction techniques, so in the coming years, therefore, there is sufficient reason to bring the quality of *sentiment mining* up to the level of that of entity extraction.

7.3.2 Categorisation and classification

Previously, the categorisation and classification of so-called *named entities* was examined in depth, but this principle can also be extended to the categorisation and classification of whole documents or parts of documents. In this context it is better to call it the clustering of documents.

In general the categorisation, classification, and clustering of algorithms can be split in two main groups: *supervised* and *non-supervised* (also called self-organizing).

7.3.2.1 *Supervised* techniques

Supervised techniques are pre-trained using a representative training set and thereafter can be used on other data. Possible categories must be made known in advance and are explicitly trained on the system in combination with the associated input data. The previously-mentioned *hidden-Markov* model is a good example of this. Other examples are *supervised* neural networks (*back-propagation neural networks*), stochastic context-free grammars, maximum entropy models, and Support Vector Machines.

In all circumstances it is necessary first to obtain the relevant document properties to be able then to use them to train the above-mentioned algorithms.

7.3.2.2 *Un-supervised* techniques

With *un-supervised* or self-organising techniques a large quantity of representative data is presented to the system, whereupon the related algorithm or model itself recognises the data, analyses and organises it, and learns so that in the future new data can be automatically classified using the same model. Categories are not known in advance; the system recognises them itself. The advantage of self-organising models is that no training is required. The disadvantage is that convergence to a stable, correct or self-optimal state is not always guaranteed.

With all of these techniques a specific mathematical distance is first defined (Euclidisch, Cosinus, Levenstein, City Block, etc.) that is subsequently applied to a set of vectors using numbers, and which will later represent the properties of documents. In some cases these are word stems, and in other cases these are semantic groups (*Latent Semantic Indexing: LSI*), or for example the previously-described identified entities, attributes, facts, or events. LSI is a good technique to reduce the dimension of the vectors, just like *principle component* analysis and other comparable mathematical dimension-reduction techniques.

The vectors (and thus indirectly the documents or concepts) are then organised and classified in accordance with the chosen mathematical distance from each other. This is

possible by clustering the vectors with cluster and self-organising algorithms and from this, for example, to derive automatic relationships between ideas, entities or concepts.

Clustering is also very useful for identifying groups of duplicates, even though it is in many cases not really practical because the computational complexity of cluster algorithms is in general the square of the number of documents in the set to cluster. Therefore, due its complexity, it is not really suitable for clustering, for example, a set of 10 million e-mails.

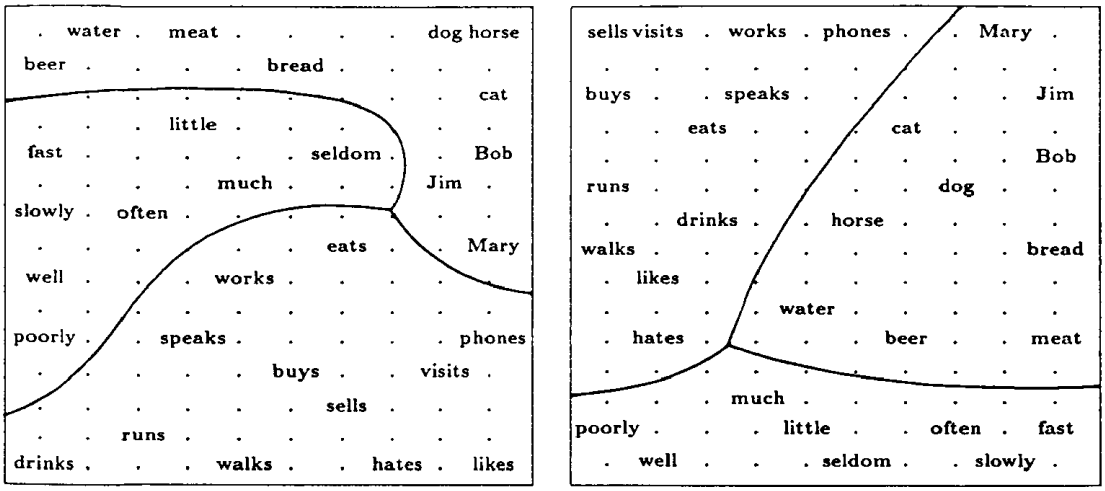


FIGURE 11. The semantotopic map for input manually tagged with contextual information (left) and the map for a restricted context of the immediate predecessor only (right) (reprinted from [Ritter et al., 1989b]).

Figure 16: An example showing clustering of words in semantic groups with a self-organising neural network (source: Scholtes, 1993)

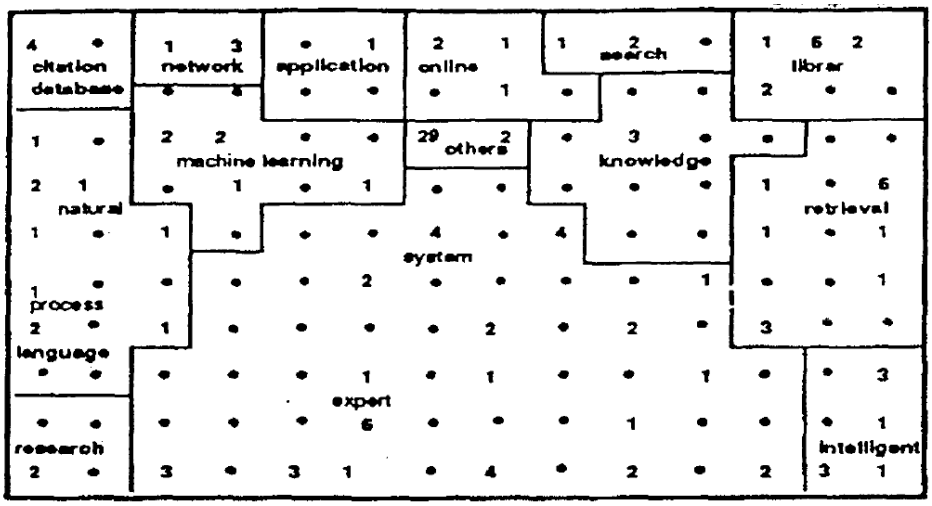
Comparable techniques can be used to automatically deduce taxonomies from unstructured document collections.

Examples of self-organising techniques are *Kohonen* self-organising neural networks and various other cluster techniques such as *N-Nearest Neighbour*, *K-Means*, and within *text mining*, the *Scatter/Gather* algorithm is popular.

The use of the *Scatter/Gather* algorithm is optimal when it is used in combination with manual selection and machine clustering. In the first case, documents are found using words in a *full-text* index. However if more general queries are made then a logical table of contents is reverted to and "neighbouring" documents shall be presented.

For each iteration in a *Scatter/Gather* session a document collection is firstly spread (*scatter*) over sets of clusters, and the short description of the clusters is presented to users. A new sub-collection is created (*gather*) from those short descriptions. The *Scatter/Gather* process is continually repeated over each new collection for as long as

there is sufficient resolution. This method creates a dynamic table of contents that can be used during navigation though the documents.



A self-organizing semantic map of AI literature. 140 documents from LISA database are used as input to produce the map. The areas on the map are automatically generated, their relative positions, neighbors, and sizes are determined by the input data. The numbers on the map represent the number of documents mapped to each node.

Figure 17: Example of a self-organising neural network in which Artificial Intelligence publications are automatically organised by subject (source: Scholtes, 1994b)

In practice the clustering of documents and the automatic deduction of a taxonomy is not very successful. This is primarily because the quality is often only 50% correct and the remainder requires manual adaptation. There are also many examples of systems that one time converge well and the next time do not, or where each iteration gives a different result. *Un-supervised* classification, clustering and automatic taxonomy generation systems require in all cases at least some human intervention, such as with the *Scatter/Gather* algorithm, and can then often attain a reasonable success.

7.4 Presentation layer of a *text mining* system

After all the processes such as those described in the preceding sections, the original data is furnished with diverse supplementary properties. Now, a completely new range of analysis and search techniques can be applied.

Use can be made of the presentation layer of the *text mining* system.

Using the enriched data it is possible to visualise data (see previously), make complex statistical analyses, call-up similar documents which searching, search on specific features, cluster on attributes, navigate using the complete text of a document and using the available document attributes, etc.

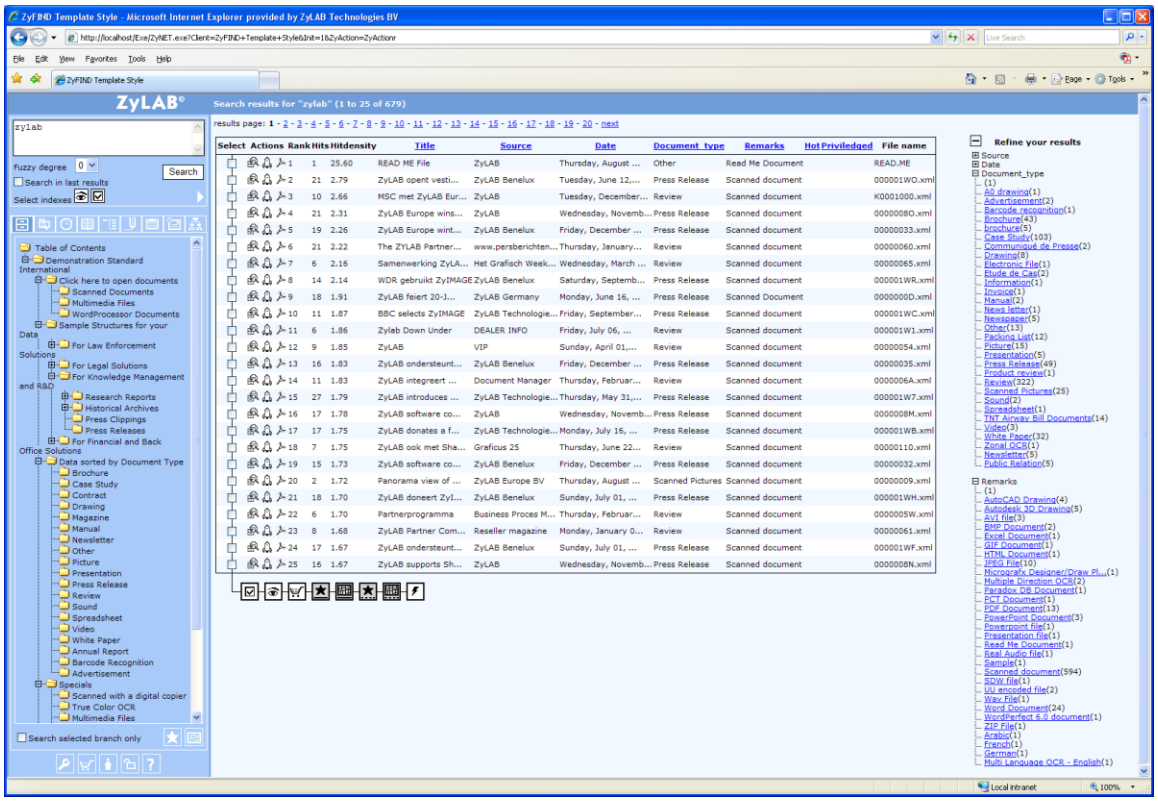


Figure 18: Presentation, and the possibilities to organize, navigate, and search (automatically found) attributes on previously found documents (source: ZyLAB Technologies B.V.)

Within this context it is of course very important to have a good, intuitive and clear user interface to be able to use the new search possibilities.

Special methods of visualisation are the following of links between documents (especially e-mails), or the creation of timelines on which documents are presented, or the generation of *heat maps* to present changes or underlying relationships between documents.

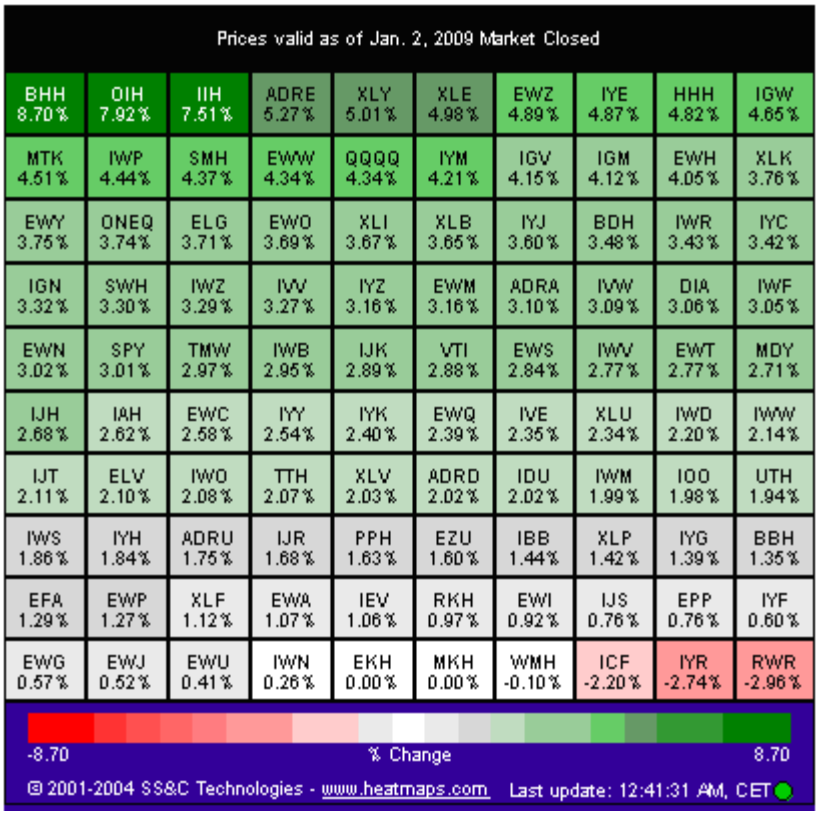


Figure 19: Heat map from the NASDAQ Stock Exchange on 2 January, 2009 (source: www.nasdaq.com)

8 Onderwijs en Onderzoek

Een kort woord over het onderwijs en onderzoek: onderwijs en onderzoek zijn de kerntaken van de leerstoel *text mining*.

De leerstoel zal zich hierbij richten op het onderwijzen van *text mining* methodieken voor taalafhankelijke *feature selection* en *feature extraction* zodat documenten kunnen worden voorzien van diverse extra entiteiten, attributen, feiten, gebeurtenissen, sentimenten en relaties die met behulp van geavanceerde gebruikersinterfaces goed doorzocht, gevisualiseerd, geanalyseerd en gefilterd kunnen worden.

In februari 2009, volgende maand, beginnen we met een college voor de *Masters Course on Knowledge Engineering* opleiding, genaamd *text mining.* Op termijn ligt het in de bedoeling om ook onderwijstaken te verzorgen voor *text mining* gerelateerde (gast)colleges bij de andere faculteiten van de Universiteit Maastricht. Zoals ik eerder heb aangegeven zijn er diverse aanknopingspunten met *life sciences*, *governance*, forensische en vanzelfsprekend juridische toepassingen.

De colleges zullen gericht zijn op het begrijpen van de hier eerder besproken technieken en de praktische toepasbaarheid daarvan binnen diverse vakgebieden.

Hiervoor zal onder andere gebruik gemaakt worden van diverse *open-source text mining* bibliotheken waarmee studenten zelf alledaagse problemen kunnen oplossen door de toepassing van *text mining*.

In de nabije toekomst zullen samen met *Masters* en eventueel ook *Ph.D.* studenten relevante onderzoeksonderwerpen gedefinieerd worden. Zoals eerder aangegeven is het vakgebied van *text mining* een jong vakgebied met vele deelgebieden waar onderzoek mogelijk en ook gewenst is. In samenspraak met de Universiteit en gekwalificeerde derde partijen zal de komende jaren gewerkt worden aan het verder brengen van het vakgebied door het uitvoeren van relevant onderzoek.

Ook zijn er diverse internationale *text mining* onderzoeksactiviteiten zoals de Legal TREC van de *University of Maryland* en diverse initiatieven binnen de Europese Unie waarbij aansluiting gezocht zal worden.

Onderzoek naar nieuwe technieken, evenals het uitbreiden van bestaande technieken voor andere of meer talen of het maken van applicatie-templates voor snellere toepasbaarheid zijn mogelijke onderzoeksonderwerpen.

9 Conclusies en Vooruitblik

9.1 Van lezen naar zoeken en vinden

In Delft maakte ik in 1982 kennis met de beginselen van de informatica. Ik verdiepte me in statistiek, patroonherkenning, artificiële intelligentie en leerde in 1987 samen met anderen een computer programma 'lezen' via zogenaamde *optical character recognition (OCR)* technologie. Met de parallelle verwerkingskracht van de NCUBE computer die we tot onze beschikking hadden (vergelijkbaar met 16 Digital VAX computers uit die tijd), was het op dat moment mogelijk om één pagina per minuut te verwerken. Nu, 22 jaar later, krijgt men bij een scanner van vijftig euro een gratis OCR pakket dat op een standaard PC per seconde bijna foutloos een volle pagina tekst leest.

Eind jaren tachtig, tijdens mijn militaire diensttijd bij de inlichtingendienst van de Koninklijke Marine was een paar gigabyte tekst per dag veel. Om daar zinnige dingen mee te doen, had men voor die tijd gigantische computers en opslagcapaciteit nodig. Nu lachen we daarom.

Tussen 1989 en 1993 heb ik getracht om met behulp van op neurale netwerken gebaseerde algoritmes, computers de structuur van taal te leren, structuren van taal af te leiden en in grote collecties tekst te zoeken. De toepassingen waren divers: van machinaal vertalen tot spraakherkenning en zoekmachines. Maar commercieel toepasbaar waren de technieken op dat moment nog niet.

Wel werden op dat moment de algoritmes die Gerald Salton eind van de jaren zestig ontwikkeld had voor het zoeken in grote hoeveelheden tekst net beschikbaar op het PC-platform. Dit was het begin van een commerciële revolutie en ook het begin van het succes van ZyLAB.

In de jaren negentig begon ook het onderzoek naar *text mining*. Op universiteiten en in samenwerking met DARPA (*TREC* en de *Message Understanding Conferences*) werden slimme algoritmes gemaakt om teksten samen te vatten, entiteiten te extraheren, documenten te clusteren, data te visualiseren, etc. Er werd veel onderzoek gedaan naar machinaal vertalen en spraakherkenning. Maar ook hier was de commerciële toepasbaarheid toen nog ver te zoeken.

Sinds 2000 is *full-text* zoeken een groot gemeengoed geworden. Iedereen gebruikt internet zoekmachines en de onderliggende technologie is algemeen geaccepteerd. Hoe anders was dat begin jaren negentig, toen het aan de man (of vrouw) brengen van *full-text retrieval* technologie nog gelijk stond aan evangeliseren. Menig bibliothecaris rilde van het idee om een eindgebruiker *full-text* toegang te geven tot een collectie: dat kon niet en was veel te gevaarlijk!

Tegenwoordig kunnen we niet meer zonder: we hebben toegang tot de inhoud van volledige bibliotheek en min of meer tot bijna alles wat ooit in de wereld geschreven is via het internet. We kunnen razendsnel zoeken en (denken alles te) vinden wat we zoeken. Nu zien we dat de balans zelfs is doorgeslagen: veel mensen denken dat we met de beschikbare internet zoekmachines alle zoekproblemen hebben opgelost. Ik hoop dat ik vandaag duidelijk heb kunnen maken dat dat niet altijd zo is.

We kunnen in ieder geval vaststellen dat op dit moment het vakgebied van de *text mining* en machinaal vertalen duidelijk bezig zijn met een commerciële doorbraak. Successen uit de inlichtingenwereld vinden nieuwe toepassingen in juridische-, medische- en industriële toepassingen. Zelfs marktonderzoek gaat tegenwoordig voor een groot deel door het internet af te zoeken naar consumenten meningen.

9.2 De generatiekloof

Een interessante observatie in deze context is dat nieuwe technologie vaak ongeveer twintig jaar nodig heeft om tot volle wasdom te komen, we hebben dat kunnen waarnemen bij de PC (1982-2002), het internet (1990-2010), en GSM telefoons. Men kan deze lijn zelfs terugtrekken tot de uitvindingen van de telegraaf, telefoon, de radio en de TV. Dit is ook heel logisch, want die twintig jaar dat is precies een generatie. Wij mensen hebben waarschijnlijk een generatie nodig om aan nieuwe technologie te wennen en ook om het “aan te passen” aan wat we acceptabel vinden.

9.3 Gevolgen van nieuwe informatie technologie

Als gevolg van al deze nieuwe informatie technologie zijn een aantal zaken de laatste tijd radicaal veranderd:

- Neem bijvoorbeeld marketing: men stuurt tegenwoordig niet meer willekeurig folders rond, maar men wacht tot mensen op bepaalde woorden zoeken die men interessant vindt en dan koopt men daar advertentieruimte omheen.
- Opsporingsdiensten kunnen nu ook kijken wie waar op gezocht heeft en, als dat mag, preventief actie ondernemen.
- Bedrijven en organisaties laten vrijwilligers problemen oplossen waar hun eigen onderzoeksafdelingen niet uitkomen of waar geen geld voor is.

De meeste van deze hedendaagse successen zijn een gevolg van de technologie die twintig jaar geleden voor het eerst ontwikkeld werd. We worden met zijn allen efficiënter, dat staat vast. Dat moet ook, want het is onmogelijk om zonder dit soort nieuwe technieken en principes de hedendaagse informatiestromen te verwerken of te controleren.

9.4 Andere te verwachten ontwikkelingen

De komende jaren staat ons nog veel meer te wachten:

- Computers zullen meer en meer automatisch data structureren en organiseren aan de hand van de inhoud van die data. Die structuren en eigenschappen zullen gebruikt worden om ons informatie op maat aan te leveren. Door vooruitgang op gebieden als de *text mining*, ons begrip van menselijke taal, het automatisch vertalen, spraakherkenning, beeldherkenning, e.d. zal dit alleen nog maar sneller gaan.
- Informatie die aanwezig is op sociale netwerken zal gekoppeld en geïntegreerd worden. Men wordt professioneel gekoppeld aan mensen met vergelijkbare interesses. Dat wordt in sommige gevallen zelfs al met routeplanners en mobiele telefoons geïntegreerd: men krijgt een SMS als er iemand binnen 10 meter is met vergelijkbare interesses!
- Interactieve websites zullen ons meer en meer gefilterde informatie presenteren waarbij rekening gehouden wordt met wie we zijn, wat onze interesses zijn en in welke informatie mensen met vergelijkbare interesses ook geïnteresseerd waren.
- Advertenties en andere commerciële aanbiedingen zullen gerichter aangeboden worden. Computers zullen steeds meer rekening gaan houden met: locatie, cultuur, ras, rijkdom, leeftijd, seksuele voorkeur, etc. Er is veel meer informatie over ons beschikbaar dan we denken.
- Taalbarrières zullen verdwijnen. De kwaliteit en binnenkort ook de gratis beschikbaarheid van hoge kwaliteit machinale vertaalsystemen zal het mogelijk maken om dynamisch informatie uit de hele wereld te vertalen en te gebruiken.
- De vorm van informatie wordt volledig transparant: telefoon, geluid, video, plaatjes, tekst, alles wordt geïntegreerd en alles wordt doorzoekbaar.

Een van de redenen waarom ik denk dat ons dit allemaal op redelijk korte termijn te wachten staat is de ontwikkeling van *The Grid*: dit is een groot gedistribueerd (vaak wereldwijd) netwerk van heel veel kleine computers. Denk aan honderdduizend tot meer dan een miljoen computers. Google en Microsoft bouwen al jaren aan een dergelijk eigen netwerk: miljarden geven ze er aan uit. Daarop is alles aan elkaar gekoppeld (ook al onze informatie). Een bekende formule is dat een netwerk even krachtig is als het kwadraat van het aantal gebruikers of computers in dat netwerk. Er staat ons dus nog heel wat te wachten!

The Grid is de infrastructuur achter een soort “wolk” (*the Cloud* wordt het ook wel in het Engels genoemd) die softwareprogramma’s host die op meerdere computers tegelijk en gedistribueerd kunnen draaien. Alle informatie is ook op meerdere computers tegelijk opgeslagen. Het is dus overal en nergens. Dit is een zeer krachtig concept. Er is in theorie geen eigen opslag meer nodig. Men kan overal bij met iedere computer die aan het internet gekoppeld is, maar ook met bijvoorbeeld een mobiele telefoon. Tien jaar gelden begon het onderzoek naar software die nodig was om gebruik te maken van dit soort

hardware. Vele manjaren zijn besteed aan het praktisch toepasbaar maken van een *Cloud* op een *Grid* architectuur. Het lijkt erop dat dit binnenkort op grote schaal gaat lukken.

En als we daarna verder kijken (of dromen), dan is er heel ver aan de horizon nog meer. Het meest revolutionair wordt waarschijnlijk de doorbraak van de kwantum computer (als die er ooit komt). Een kwantum computer is voor bepaalde toepassingen exponentieel zo krachtig vergeleken met gewone computers. De eerste kwantum algoritmes om massaal informatie te verwerken en te doorzoeken zijn er al: zoeken in grote hoeveelheden informatie kan daarmee nog beter en vooral sneller.

Kwantum computers zijn vooral goed in het zoeken in complexe hoogdimensionale data, die ook vaak nog onvolledig is en vol ruis zit, zoals geluid, spraak, taal, video, beeld, en DNA. Het probleem is alleen dat we nog geen echte kwantum “chips” kunnen maken, we komen nog niet verder dan simulaties. Maar bij de TU-Delft doen ze revolutionair onderzoek naar echte kwantum computeronderdelen. Wellicht dat dit nog 50-100 jaar duurt, maar dat het eraan komt is zeer waarschijnlijk.

Verder onderzoek naar het omzetten van algoritmes en principes zoals we die nu kennen op binaire computers naar algoritmes die geschikt zijn voor kwantum computers is één van de meer interessante uitdagingen voor de informatica wetenschap.

9.5 De komende twintig jaar

Zoals ik eerder aangaf, vonden we eind van de jaren tachtig twee gigabyte per dag veel informatie. Nu begint een paar terabyte (1.024 gigabyte) per dag een probleem te worden, toch is dat ook best te overzien. Maar als Moore's wetten van kracht blijven, dan hebben we over twintig jaar picabytes (1.024 terabyte) per dag te verwerken. En dat vinden we nu wel een probleem, want om daarmee om te gaan hebben we nog niet echt de technieken in huis. Ook zal de aard van de informatie de komende jaren veranderen: al die data zal, naast het bevatten van oneindig veel duplicaten, ook vooral transactioneel, real-time en multi-mediaal van aard zijn.

We hebben nu vaak al problemen met de complexe structuur van email bestanden, wat dus te denken van al die gigantische hoeveelheden *instant messaging*, chat sessies, sociale netwerken, geluid, foto's en video die op ons afkomen?

Daarvoor zullen we nieuwe technieken en principes moeten gebruiken die wetenschappers, studenten en bedrijven nu ontwikkelen.

We zullen verder door moeten gaan met relatief eenvoudig en dom werk door computerprogramma's te laten uitvoeren. Daar moet hoogwaardiger werk voor terug komen. Dat moeten we dan natuurlijk wel met zijn allen kunnen oppakken. Onderwijs van hoge kwaliteit en toegankelijk voor iedereen, blijft dus noodzakelijk in de toekomst!

Want, mijns inziens zullen mensen altijd noodzakelijk blijven voor kwaliteitscontrole en kwaliteitswaarborging. Ook zullen ze moeten zorgen voor "serendipiteit": nieuwe interesses aanwakkeren en verrassingen tonen, want daar is meer voor nodig dan willekeurige selecties gemaakt door computerprogramma's, zoals dat nu soms gaat.

Ook moeten we blijven begrijpen hoe computerprogramma's werken en wat de beperkingen zijn, zoals u nu weet wat de beperkingen van internet zoekmachines zijn. Ik heb er echter vertrouwen in dat mensen altijd de beperkingen van machines snel genoeg in de gaten zullen krijgen.

Als altijd blijft onze privacy belangrijk, we moeten controle blijven houden over onze informatie. Hoewel dat meer lijkt te gelden voor ouderen dan voor jongeren: de laatsten lijken al meer gewend aan het relatieve gebrek aan privacy of ze laten zich minder snel in de maling nemen.

En die twintig jaar die het duurt voor we allemaal dagelijks één of meer picabytes te verwerken krijgen, die zullen we hard nodig hebben om nieuwe technieken te ontwikkelen, deze toepasbaar te maken, en ze op voldoende data te testen en te verbeteren. Maar we hebben die tijd ook vooral nodig om met zijn allen aan al die nieuwe technieken en nieuwe manieren van werken te wennen!

10 Dankwoord

Ten eerste wil ik het College van Bestuur van de Universiteit van Maastricht hartelijk danken voor hun bereidheid te investeren in de leerstoel *text mining*. Zoals aangegeven investeert zij hiermee niet alleen in onderwijs en onderzoek in de technologie zelf, maar ook in het toegankelijk blijven van informatie in de toekomst. Volgende maand zal het vak *text mining* als verplicht onderdeel van de *Masters* opleiding van start gaan, en ik verheug me op de samenwerking met studenten en nieuwe collega's om de eerste versie van de cursus tot een succes te maken!

Ik wil ZyLAB Technologies BV, en mijn collega's in het bijzonder danken voor het financieren van de bijzondere leerstoel.

In deze gaat speciale dank uit naar de collega hoogleraren Jaap van den Herik en Eric Postma die met veel doorzettingsvermogen geholpen hebben deze bijzondere leerstoel te realiseren. Het gehele proces begon een aantal jaren geleden in Delft op initiatief van Jaap, die het vervolgens niet heeft losgelaten. Het is jammer dat jullie Maastricht verlaten hebben en naar Tilburg zijn gegaan, maar zoals ik aangegeven heb, zal ik in Maastricht de honneurs voor jullie waarnemen!

Jaap verdient een extra dankwoord, want onder hem ben ik indertijd mijn wetenschappelijke carrière begonnen, in 1985 aan de TU-Delft wel te verstaan. Jaap: na mijn afstuderen vertelde ik je dat ik een punt zette achter mijn wetenschappelijke carrière. Toen ik later toch besloot om te gaan promoveren aan de Universiteit van Amsterdam, kwam ik ermee weg door je te vertellen dat het puntkomma was geworden. In het kader van deze laatste stap was het wellicht toch een komma, maar daar heb je zelf dan ook hard aan meegewerkt. Jaap, bedankt voor alles!

Ook gaat dank uit naar Professor Remco Scha, bij wie ik in 1993 promoveerde aan de Universiteit van Amsterdam. Dank voor je flexibiliteit die mij toestond om te promoveren en tegelijk aan mijn eigen bedrijf te werken. Ook dank voor de vele sturing en in het bijzonder voor de tip die me nu nog bij staat: "een promotie is iets anders dan een verkoopverhaal, het was waarschijnlijk de enige keer in mijn leven dat ik ergens echt hard over na kon denken". Je had gelijk, het kan geen kwaad om ergens echt de tijd voor te nemen en af en toe eens heel diep na te denken.

Verder gaat mijn dank uit naar de Koninklijke Marine, en in het bijzonder de Marine Inlichtingen Dienst, voor de *hands-on* training en de unieke ervaring op een bijzonder gebied van de informatie technologie die voor weinigen toegankelijk dan wel bekend is.

In deze context ben ik ook dank verschuldigd aan alle klanten van ZyLAB, in het bijzonder diegenen die hun grootste informatie technische problemen met ons wilden delen en samen met ons aan nieuwe oplossingen wilden werken. Zelfs als dat betekende dat we het risico moesten nemen om nieuwe programma's en nieuwe technieken te ontwikkelen. Door mee te werken aan het oplossen van deze problemen, hebben ZyLAB

en ikzelf in het bijzonder altijd vooraan gelopen bij nieuwe inzichten in nieuwe technieken om de meest complexe informatie technische problemen van de laatste twintig jaar op te lossen met innovatieve, maar vooral ook praktische en gebruikersvriendelijke oplossingen. De lijst is waarschijnlijk te lang om op te noemen, maar de UN Oorlogstribunalen, de Europese Commissie en diverse verder niet bij naam te noemen (internationale) opsporings- en inlichtingendiensten verdienen een speciale vermelding. Ik hoop in de toekomst verder met hen te werken bij het ontwikkelen van vele nieuwe toepassingen van informatie technologie.

Mijn familie en ouders hebben altijd een belangrijke rol voor me gespeeld. Ik wil ze hierbij allemaal bedanken dat ze altijd achter me gestaan hebben. Ria, wat zou Kees trots geweest zijn, helaas kan hij er niet bij zijn. Het was mijn vader die mij min of meer inschreef voor de informatica opleiding aan de TU-Delft. Hij was het ook die, vanwege mijn dreigende militaire dienst als soldaat bij de landmacht, ervoor zorgde dat de Marine Inlichtingen Dienst in het bezit kwam van mijn afstudeerverslag, wat al snel leidde tot een plaatsing als officier bij de Koninklijke Marine in Amsterdam. Als zelfstandig ondernemer heeft hij mijn ZyLAB activiteiten altijd volledig ondersteund, maar ook mijn wetenschappelijke activiteiten heeft hij altijd gestimuleerd.

Tot slot wil ik mijn vrouw Frédérique en mijn kinderen Josefien, Stefanie en Loek ook bedanken voor hun onvoorwaardelijke steun en liefde. Frédérique heeft meer dan wie dan ook dit hoogleraarschap gestimuleerd en ervoor gezorgd dat ik er de nodige tijd en aandacht aan heb kunnen besteden, wat vaak ook ten koste ging van tijd en aandacht voor de familie! Frédérique, Josefien, Stefanie en Loek: vandaag is ook jullie feestje!

Dames en heren, dank voor uw aandacht.

Ik heb gezegd.

11 Verwijzingen en noten

In deze paragraaf zijn per onderdeel noten en verwijzingen naar aanvullende literatuur opgenomen.

Wat is Text Mining?

Scholtes, J.C. (2008d) geeft een kort overzicht over technieken die gebruikt worden in *text analytics* en *text mining*. In Witten, I.H. and Frank, E. (2005) staat een uitgebreid overzicht van het vakgebied *data mining*, de gestructureerde variant van *text mining*.

Andere basis boeken op het gebied van *text mining* worden vermeld in de sectie over *text mining* technieken. De meest toonaangevende op dit moment zijn: Feldman, R., and Sanger, J. (2006), Berry, M.W., Editor (2004) en Berry, M. W. and Castellanos, M. Editors (2006).

Zoeken met Computers in Ongestructureerde Informatie

Blair, D.C. and Maron, M.E. (1985) was het eerste onderzoek dat de effectiviteit van puur Booleaanse zoeksystemen in twijfel trok in 1985. De conclusies worden nog steeds bevestigd. Recent weer door het LEGAL-TREC onderzoek en door Baron, Jason R. (2005).

Andrews, Whit and Knox, Rita (2008) geeft een goed overzicht van commercieel verkrijgbare *Information Access* systemen: systemen die een combinatie van zoeken, visualisatie, *text mining* en integratie met andere business applicaties bieden.

Voor meer informatie over de werking van zoekmachines wordt verwezen naar een groot aantal klassiekers uit de *informatie retrieval* literatuur. Deze publicaties geven een uitgebreid overzicht geven van de diverse zoek-, relevance ranking- en programmeertechnieken die in de loop der jaren ontwikkeld zijn voor het zoeken binnen grote hoeveelheden tekst. Soms puur wiskundige technieken, maar in de loop der jaren ook meer en meer technieken die gebruik maakten van *artificial intelligence* en taaltechnologie: Crestani, F., Lalmas, M. and Rijsbergen, C.J. van, (Editors), 1998, Croft, W.B. and Harper, D.J. (1979), Croft, Bruce (Editor), (2000), Dominich, Sándor (2008), Grefenstette, Gregory (1998), Kowalski, Gerald (1997), Kruschwitz, Udo (2005), Losee, R.M. (1998), Manning, Christopher D., Raghavan, Prabhakar, and Schütze, Hinrich (2008), Meadow, C.T., Boyce, B.R., Kraft, D.H. and Barry, C. (2007), Rijsbergen, C.J. van (1979), Rijsbergen, C.J. van (2004), Salton, G., Wong, A. and Yang, C.S. (1968), Salton, Gerard (1971), Salton, Gerard, (1975), Salton, Gerard, and McGill, Michael (1983), Salton, Gerard, (1989), Scholtes, J.C. (1995), Scholtes, J.C. (1996), Scholtes, J.C. (2007g), Spink, Amanda and Cole, Charles (Editors), (2005), Tait, John I. (Editor), (2005), White, Martin (2007), en Wilkingson, R., Arnold-Moore, T., Fuller, M., Sacks-Davis, R., Thom, J. and Zobel, J. (1998).

Knuth, D.E. (1998) en Knuth, D.E. (2008) geven een goed overzicht van de onderliggende algoritmes die zoekmachines gebruiken.

Text Mining in Relatie tot "Zoeken & Vinden"

Scholtes, J.C. (2005a) en Scholtes, J.C. (2009) gaan in meer detail in waarom en wanneer het relevant is om "alles" te vinden in plaats van alleen de meest relevante documenten. Ook wordt ingegaan op technieken om zaken te vinden die niet gevonden willen worden en hoe men zaken vindt terwijl men niet precies weet waar men op zoekt.

Ingwersen, Peter and Järvelin, Kalervo (2005) beschrijft diverse technieken om gebruik te maken van de context van een document of kennis over een domein om beter en efficiënter te zoeken.

Over *text mining* en informatie visualisatie is een keur aan literatuur beschikbaar.

Een van de meest toonaangevende en volledigste is Card, Stuart K., Mackinlay, Jock D., and Shneiderman, Ben, Editors (1999), waarin een overzicht wordt gegeven van bijna alle visualisatie technieken die tot 2000 beschikbaar waren. Ook de herdruk van Tufte, Edward, R. (2001) is een absolute aanrader.

Andere referenties zijn: Bimbo, Alberto del (1999), Chen, Chaomei (2006), Fry, Ben (2008), en Scholtes, J.C. (2005b).

Meer over andere voordelen en toepassingen van *text mining* om van ongestructureerde data gestructureerde data te maken zijn te vinden in: Chakrabarti, S. (2003) en in Chan, G., Healey, M.J., McHugh, J.A.M., and Wang, J.T.L., (2001).

Voorbeelden van Toepassingen van Text Mining

In Knox, R. (2008) staat een uitgebreid overzicht van commerciële toepassingen van *text mining*, speciaal gericht op de strategische toepassing van *text mining* binnen een IT-organisatie.

Miller, Thomas W. (2005), Prado, Hercules Antonio Do (Editor), Ferneda, Edilson (Editor), (2008), Spangler, Scott and Kreulen, Jeffrey (2008), en Sullivan, Dan (2001) bevatten meerdere goede beschrijvingen van de praktische en commerciële toepassingen van *text mining* technologie.

Voor de toepassing van *text-mining* binnen fraude- en criminaliteitsopsporing en inlichtingen analyses wordt verwezen naar Scholtes, J.C. (2007a), Scholtes, J.C. (2007b), Scholtes, J.C. (2007c), Scholtes, J.C. (2007d), Scholtes, J.C. (2008b) en natuurlijk DARPA: Defense Advanced Research Project Agency (1991).

Voorbeelden van de toepassing van *text mining* voor *business intelligence* kunnen gevonden worden in: Halliman, Charles (2001) en in Inmon, William H. and Nesavich, Anthony (2008).

Technieken en toepassingen die gebruikt worden bij *sentiment mining* zijn te vinden in Shanahan, J.G., Qu, Y., and Wiebe, J. (Editors), (2006) en in Scholtes, J.C. (2008).

Meer over *text mining* bij klinisch onderzoek en andere biomedische toepassingen is te lezen in Herron, Patrick (2008), Zvelebil, M. and Baum, J.O. (2008) en in Ananiadou, Sophia (Editor), Mcnaught, John (Editor), (2006).

E-discovery is in potentie één van de meest veelbelovende toepassingsgebieden van *text mining*, zeker binnen de context van de kredietcrisis en alle onderzoeken en rechtszaken die gegarandeerd gaan volgen.

Meer over de wet- en regelgeving van *e-discovery* en de *Federal Rules of Civil Procedure* kan gevonden worden in Dahlstrom Legal Publishing (2006), op EDRM (Electronic Discovery Reference model): http://www.edrm.net, in Paul, G.L. and Nearon, B.H. (2006) en in *The Discovery Revolution. E-Discovery Amendments to the Federal Rules of Civil Procedure.* American Bar Association.

Debra Logan, John Bace, and Whit Andrews (2008) geeft een zeer volledig overzicht van commerciële leveranciers van *e-discovery* software oplossingen.

Meer referenties voor advocaten en juristen over *e-discovery* kunnen gevonden worden in Lange, M.C.S. and Nimsger, K.M. (2004), op de Sedona Conference website: http://www.thesedonaconference.org/ en in Socha, George (2009). Dit laatste rapport gaat over het in-huis uitvoeren van delen van het *e-discovery* proces met de bijbehorende risico's en voordelen.

Meer gedetailleerde beschrijvingen van *e-discovery* technieken en toepassingen in relatie tot *text mining* en *information retrieval* kunnen gevonden worden in Scholtes, J.C. (2006c), Scholtes, J.C. (2007f), Scholtes, J.C. (2007h), Scholtes, J.C. (200i7), Scholtes, J.C. (2007j), en Scholtes, J.C. (2008c).

In de komende jaren zal nieuwe regelgeving en *compliance* een belangrijk onderwerp worden. Meer hierover en over de toepassingen van *text mining* en *information retrieval* in relatie tot email, records management en fraude opsporing kan gevonden worden in: Manning, George A. (2000), Scholtes, J.C. (2004a), Scholtes, J.C. (2004b), Scholtes, J.C. (2005c), Scholtes, J.C. (2005d), Scholtes, J.C. (2006a), Scholtes, J.C. (2006b), Scholtes, J.C. (2007e), Scholtes, J.C. (2008f), en Scholtes, J.C. (2007k).

De Technologie achter Text Mining

De core-technologie achter *text mining* is zeer uitgebreid en gedetailleerd na te lezen in: Feldman, R., and Sanger, J. (2006), Berry, M.W., Editor (2004), Berry, M. W. en Castellanos, M. Editors (2006) en Weiss, et al. (2005). De eerste referentie zal gebruikt worden als tekstboek bij het college.

Een bijzonder boek is Bilisoly, Roger (2008), hierin wordt met behulp van *Perl* het gebruik van reguliere expressies tot het uiterste doorgevoerd. Zeker interessant voor fans van de programmeertaal *Perl*.

Er is veel geschreven over natuurlijk taalverwerking, oftewel *natural language processing* (NLP). Mitkov, Ruslan (2003*). The Oxford Handbook of Computational Linguistics,* is een van de meest complete overzichtswerken. Een van de eerste werken over *discourse analysis* kan gevonden worden in: Scha, R. and Polanyi, L. (1988). Kay, Martin (1986) en Woods, W.A. (1970) geven meer inzicht in snelle technieken om een grammaticale analyse te maken (*parsing*). Manning, Christopher D. and Schütze, Hinrich, (1999) is het grote standaardwerk op het gebied van statistische taalverwerking. In Scha, R., Bod, R. and Sima'an, K. (1999) en in Bod, R., Scha, R., and Sima'an, K. (Editors), (2003) wordt het parsen van taal aan de hand van een geanoteerd corpus beschreven: *Data-Oriented Parsing.* En Kao A., Poteet, S. R. (Editors), (2007) beschrijft de rol van natuurlijke taalverwerking binnen *text mining* in detail.

Meer over machinaal vertalen kan gevonden worden in: Goutte, C., Cancedda, N., Dymetman, M. and Foster, G. (Eds.). (2009). En tot slot beschrijft Moens, Marie-Francine, (2000) de diverse technieken die beschikbaar zijn voor het automatisch samenvatten van teksten.

Als standaardwerken over patroonherkenning gelden Devijver, P.A. and Kittler, J. (1982), Duda, R.O. and Hart, P.E. (1973) en in de recente bijgewerkte 2e editie: Duda, R.O. and Hart, P.E. (2001*).* Andere goede bronnen zijn: Bishop, C.M. (2006) en Chen, Y., Li, J., and Wang, J. (2004).

In Moens, Marie-Francine (2006) vinden we een zeer volledig en overzichtelijke uiteenzetting van de bekendste informatie extractie technieken.

Zoals eerder is aangeven, was het de Amerikaanse overheid die een eerste aanzet heeft gegeven voor de extractie van *named entities* uit vrije tekst. Meer hierover kan gevonden worden in een van de weinige openbare publicaties: DARPA: Defense Advanced Research Project Agency (1991).

Sparck-Jones, K. (1971) en Allan, James (Editor), (2002) geven een goed overzicht van de visie van traditionele *information retrieval* specialisten op entiteit-extractie.

Meer over *machine learning* kan gevonden worden in de klassieke werken van Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Editors), (1986a) en Michalski, R.S., Carbonell,

J.G. and Mitchell, T.M. (Editors), (1986a), en in Mitchell, Tom, (1997). *Machine Learning.* McGraw Hill.

Ikonomakis, M., Kotsiantis, S., and Tampakas, V. (2005) geeft een goed overzicht van het gebruik van *machine learning* technieken voor tekst classificatie.

Meer over *Support Vector Machines* (SVM) kan gevonden worden in Cristianini, N. and Shawe-Taylor, J. (2000).

Een andere klassieker is de wetenschappelijke publicatie met de titel *Latent Semantic Indexing* van Dumais, S.T., Furnas, G.W., Landauer, T.K. , Deerwater, S. and Harshman, R. (1988). Ook Voorhees, Ellen M. (1985) is een interessant overzicht van de toepassing van cluster technieken binnen *information retrieval*.

Vervolgens is er veel onderzoek gedaan in het begin van de jaren negentig door ondergetekende naar toepassingen van zelforganiserende neurale netwerken voor taalverwerking en *information retrieval*. Meer kan gevonden worden in: Kohonen, T. (1984) en in Scholtes, J.C. (1990a). Scholtes, J.C. (1990b). Scholtes, J.C. (1991a), Scholtes, J.C. (1991b). Scholtes, J.C. (1991c). Scholtes, J.C. (1991d). Scholtes, J.C. (1991e). Scholtes, J.C. (1991f). Scholtes, J.C. (1991g). Scholtes, J.C. (1991h). Scholtes, J.C. (1991i). Scholtes, J.C. (1991j). Scholtes, J.C. (1991k). Scholtes, J.C. (1992a). Scholtes, J. (1992b). Scholtes, J.C. (1992c). Scholtes, J.C. (1992d). Scholtes, J.C. (1992e). Scholtes, J.C. (1992f). Scholtes, J.C. (1992g). Scholtes, J.C. and Bloembergen, S. (1992a). Scholtes, J.C. and Bloembergen, S. (1992b). Scholtes, J.C. (1992h). Scholtes, J.C. (1993). Scholtes, J.C. (1994a). Scholtes, J.C. (1994b).

Onderwijs en Onderzoek

In Baron, Jason R. (2005) is meer te vinden over de grote voortrekker van het LEGAL-TREC initiatief om zoektechnieken te evalueren zodat ze betrouwbaar in rechtszaken kunnen worden ingezet. Details over het Legal-TREC Research Program zijn hier te vinden: http://trec-legal.umiacs.umd.edu/.

Jason Baron is ook betrokken bij de *Sedona Conference*, een initiatief van diverse advocaten, bedrijfsjuristen en rechters om standaarden te definiëren op het gebied van *e-discovery*: Sedona Conference: http://www.thesedonaconference.org/.

LEGAL-TREC was een voortzetting van TREC, meer over de geschiedenis, doelstellingen en resultaten van TREC kan hier gevonden worden: Voorhees, Ellen M. (Editor), Harman, Donna K. (Editor), (2005).

Konchady Manu, (2006) is een goed praktisch werkboek dat in combinatie met de nodige *open source text mining* software gebruikt gaat worden tijdens het praktische gedeelte van het *text mining* college.

Conclusies en Vooruitblik

Meer over *optical character recognition (OCR)* kan gevonden worden in Henseler, J., Scholtes, J.C., and Verhoest, C.R.J. (1987) en Herik, H.J. van den, Scholtes, J.C. and Verhoest, C.R.J. (1988).

Jurafsky, D. and Martin, J.H., (2009) geeft een breed overzicht over spraakherkenning technologie.

Een intrigerend boek is Kurzweil, Ray (2005). Hierin wordt door de maker van één van de eerste commerciële OCR machines een bijzondere visie gegeven over de gevolgen van de informatie maatschappij en de convergentie van mensen en machines.

Andere boeiende publicaties over de maatschappelijke impact die recente ICT technieken tot gevolg gehad hebben voor massa collaboratie, het oplossen van complexe problemen, het verschijnsel dat ook wel *Wishdom of Crowds* wordt genoemd en het "concurreren door te analyseren" kunnen gevonden worden in: Tapscott, D. and Williams, A.D. (2006), Ayers, Ian (2007), Davenport, T.H. and Harris, J.G. (2007), Segaran, T. (2007), en Surowiecki, James (2004). Al deze nieuwe maatschappelijke en economische principes zijn mogelijk geworden door de toepassing van *text mining* technieken.

Meer over het zoeken op inhoud in multimediale bestanden kan gevonden worden in: Postma E.O. and Herik, H.J. van den (2000), Wu, J.K., Kankanhalli, M.S., Lim, J.H., and Hong, D. (2000) en in Wang, James Z. (2001).

Uitleg over de architectuur en de algoritmes die gebruikt worden in *The Grid* zijn te vinden in: Berman, F., Fox, G. and Hey, T. (Editors), (2003), Li, Maozhen and Baker, Mark (2005) en Liu, Bing (2007).

Meer over kwantum computers en kwantum algoritmes kan gevonden worden in: Kaye, P., Laflamme, R. and Mosca, M. (2007) en in Steeb, W.H. and Hardy, Y. (2006).

En Scholtes, J.C. (2008e) geeft een korte visie op de toekomst en over de mogelijke risico's en voordelen van de moderne informatie maatschappij.

12 Literatuurlijst

Allan, James (Editor), (2002). *Topic detection and tracking: event-based information organization.* Kluwer Academic Publishers.

Ananiadou, Sophia (Editor), Mcnaught, John (Editor), (2006). *Text mining for biology and biomedicine.* Artech House.

Andrews, Whit and Knox, Rita (2008). *Magic Quadrant for Information Access Technology.* September 30, 2008. Gartner Research Report, ID Number: G00161178. Gartner, Inc.

Ayers, Ian (2007). Super Crunchers. *Why Thinking-by-Numbers is the New Way to be Smart.* Bantam Books.

Baron, Jason R. (2005). Toward a Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery. *Sedona Conference Journal*. Vol. 6, 2005.

Berman, F., Fox, G. and Hey, T. (Editors), (2003). *Grid computing: making the global infrastructure a reality.* John Wiley and Sons.

Berry, M.W., Editor (2004). *Survey of text mining: clustering, classification, and retrieval.* Springer-Verlag.

Berry, M. W. and Castellanos, M. Editors (2006). *Survey of Text Mining II: Clustering, Classification, and Retrieval.* Springer-Verlag.

Bilisoly, Roger (2008). *Practical Text Mining with Perl* (Wiley Series on Methods and Applications in Data Mining). John Wiley and Sons.

Bimbo, Alberto del (1999). *Visual Information Retrieval*. Morgan Kaufmann.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.

Blair, D.C. and Maron, M.E. (1985). An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. *Communications of the ACM*, Vol. 28, No. 3, pp. 289-299.

Bod, R., Scha, R., and Sima'an, K. (Editors), (2003). *Data-Oriented Parsing.* Center for the Study of Language and Information, Stanford, CA.

Card, Stuart K., Mackinlay, Jock D., and Shneiderman, Ben, Editors (1999). *Readings in information visualization: using vision to think.* Morgan Kaufmann Publishers.

Chakrabarti, S. (2003). Mining the Web. *Discovering Knowledge from Hypertext Data.* Morgan Kaufman.

Chan, G., Healey, M.J., McHugh, J.A.M., and Wang, J.T.L., (2001). *Mining the World Wide Web, an information search approach.* Kluwer Academic Publishers.

Chen, Chaomei (2006). *Information Visualization: Beyond the Horizon.* Springer-Verlag.

Chen, Y., Li, J., and Wang, J. (2004). *Machine Learning and Statistical Modelling Approaches to Image Retrieval.* Kluwer Academic Publishing.

Crestani, F., Lalmas, M. and Rijsbergen, C.J. van, (Editors), 1998. *Information Retrieval: Uncertainty and Logics. Advanced Models for the Representation and Retrieval of Information.* Kluwer Academic Publishers.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines: and other kernel-based learning methods.* Cambridge University Press.

Croft, W.B. and Harper, D.J. (1979). Using Probabilistic Models of Information Retrieval without Relevance Information. *Journal of Documentation*. Vol. 35, No. 4, pp. 285-295.

Croft, Bruce (Editor), (2000). *Advances in Information Retrieval. Recent Research from the Center for Intelligent Information Retrieval.* Kluwer Academic Publishers.

Dahlstrom Legal Publishing (2006). *The New E-Discovery Rules. Amendments to the Federal Rules of Civil Procedure Addressing Discovery of Electronically Stored Information (effective December 1st, 2006).*

DARPA: Defense Advanced Research project Agency (1991). Message Understanding Conference (MUC-3). *Proceedings of the Third Message Understanding Conference (MUC-3)*. DARPA.

Davenport, T.H. and Harris, J.G. (2007). *Competing on Analytics. The New Science of Winning.* Harvard Business School Press.

Devijver, P.A. and Kittler, J. (1982). *Pattern Recognition: A Statistical Approach.* Prentice Hall.

Dominich, Sándor (2008). *The Modern Algebra of Information Retrieval.* Springer-Verlag.

Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis.* John Wiley and Sons.

Duda, R.O. and Hart, P.E. (2001*). Pattern Classification (2nd Edition).* John Wiley and Sons.

Dumais, S.T., Furnas, G.W., Landauer, T.K. , Deerwater, S. and Harshman, R. (1988). Using Lantent Semantic Analysis to Improve Access to Textual Information. *ACM CHI'88*. pp. 281-285.

EDRM: Electronic Discovery Reference model: http://www.EDRM.net

Escher, M.C. Official M. C. Escher Web site, published by the M.C. Escher Foundation and Cordon Art B.V. http://www.mcescher.com/

Feldman, R., and Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press.

Fry, Ben (2008). *Visualizing Data. Exploring and Explaining Data with the Processing Environment.* O'Reilly.

Grefenstette, Gregory (1998). *Cross-Language Information Retrieval.* Kluwer Academic Publishers.

Grossman, D.A. and Frieder, O. (2004). *Information Retrieval: Algorithms and Heuristics (The Information Retrieval Series).* Springer-Verlag.

Goutte, C., Cancedda, N., Dymetman, M. and Foster, G. (Eds.) (2009). *Learning Machine Translation.* MIT Press.

Halliman, Charles (2001). *Business Intelligence Using Smart Techniques. Environmental Scanning Using Text mining and Competitor Analysis Using Scenarios and Manual Simulation.* Information Uncover.

Henseler, J., Scholtes, J.C., and Verhoest, C.R.J. (1987). *The Design of a Parallel Knowledge-Based Optical Character-Recognition System*. M.Sc. Thesis. Delft University of Technology, Department of Mathematics & Computer Science, 1987.

Herik, H.J. van den, Scholtes, J.C. and Verhoest, C.R.J. (1988). The Design of a Knowledge-Based Optical-Character Recognition System. *Proc. of the SCS*, June 1-3, 1988, pp. 350-358. Nice, France.

Herron, Patrick (2008). *Text Mining for Genomics-based Drug Discovery.* VDM Verlag Dr. Müller.

Ikonomakis, M., Kotsiantis, S., and Tampakas, V. (2005). Text Classification Using Machine Learning Techniques, *WSEAS Transactions on Computers*, Issue 8, Volume 4, August 2005, pp. 966-974.

Ingwersen, Peter and Järvelin, Kalervo (2005). *The Turn: Integration of Information Seeking and Retrieval in Context.* Springer-Verlag.

Inmon, William H. and Nesavich, Anthony (2008). *Tapping into Unstructured Data. Integrating Unstructured Data and Textual Analytics into Business Intelligence.* Prentice Hall.

Jurafsky, D. and Martin, J.H., (2009). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd Edition.* Pearson, Prentice Hall.

Kao A., Poteet, S. R. (Editors), (2007). *Natural Language Processing and Text Mining.* Springer-Verlag.

Kay, Martin (1986). *Algorithm schemata and data structures in syntactic processing. Readings in natural language processing*. Morgan Kaufmann Publishers Inc.

Kaye, P., Laflamme, R. and Mosca, M. (2007). *An Introduction to Quantum Computing.* Oxford Press.

Konchady Manu, (2006). *Text Mining Application Programming (Programming Series).* Charles River Media.

Kowalski, Gerald (1997). *Information Retrieval Systems. Theory and Implementation.* Kluwer Academic Publishers.

Knox, R. (2008). Content Analytics Supports many Purposes. *Gartner Research Report*, ID Number: G00154705, January 10, 2008.

Knuth, D.E. (1998). *Art of Computer Programming, Volume 1-3 (2nd Edition).* Addison Wesley Professional.

Knuth, D.E. (2008). *The Art of Computer Programming, Volume 4, Fascicle 0: Introduction to Combinatorial Algorithms and Boolean Functions.* Addison Wesley Professional.

Kruschwitz, Udo (2005). *Intelligent Document Retrieval. Exploiting Markup Structure.* Springer-Verlag.

Kohonen, T. (1984). *Self-Organization and Associative Memory.* Springer-Verlag.

Kurzweil, Ray (2005). The *Singularity is Near, when Humans Transcend Biology.* Viking (Penguin Group).

Logan, Debra, Bace, John, and Andrews, Whit (2008). *MarketScope for E-Discovery Software Product Vendors.* Gartner Research Report ID Number: G00163258. Gartner, Inc.

Lange, M.C.S. and Nimsger, K.M. (2004). *Electronic Evidence and Discovery: What Every Lawyer Should Know.* American Bar Association.

Legal-TREC Research Program: http://trec-legal.umiacs.umd.edu/.

Li, Maozhen and Baker, Mark (2005*). The Grid: Core Technologies.* John Wiley and Sons.

Liu, Bing (2007). *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data.* Springer-Verlag.

Losee, R.M. (1998). Text-*Retrieval and Filtering. Analytic Models of Performance.* Kluwer Academic Publishers.

Manning, Christopher D. and Schütze, Hinrich, (1999). *Foundations of statistical natural language processing.* MIT Press.

Manning, Christopher D., Raghavan, Prabhakar, and Schütze, Hinrich (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Manning, George A. (2000). *Financial Investigation and Forensic Accounting.* CRC Press.

Meadow, C.T., Boyce, B.R., Kraft, D.H. and Barry, C. (2007). *Text Information Retrieval System (3rd Edition).* Academic Press, Elsevier.

Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Editors), (1986a). *Machine Learning, an Artificial Intelligence Apporach. Volume 1*. Morgan Kaufmann.

Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Editors), (1986b). *Machine Learning, an Artificial Intelligence Apporach. Volume 2*. Morgan Kaufmann.

Miller, Thomas W. (2005). *Data and text mining: a business applications approach.* Pearson Prentice Hall.

Mitchell, Tom, (1997). *Machine Learning.* McGraw Hill.

Mitkov, Ruslan (2003*). The Oxford Handbook of Computational Linguistics.* Oxford University Press.

Moens, Marie-Francine, (2000). *Automatic Indexing and Abstracting of Document Texts.* Kluwer Academic Publishers.

Moens, Marie-Francine (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Springer-Verlag.

Paul, G.L. and Nearon, B.H. (2006). *The Discovery Revolution. E-Discovery Amendments to the Federal Rules of Civil Procedure.* American Bar Associaton.

Postma E.O. and Herik, H.J. van den (2000). Discovering the visual signature of painters. In N. Kasabov (Editor), *Future Directions for Intelligent Systems and Information Sciences. The Future of Speech and Image Technologies, Brain Computers, WWW and Bioinformatics, pp. 129-147.* Heidelberg: Physica Verlag (Springer Verlag).

Prado, Hercules Antonio Do (Editor), Ferneda, Edilson (Editor), (2008). *Emerging technologies of text mining: techniques and applications.* Information Science Reference.

Rijsbergen, C.J. van (1979). *Information Retrieval*. Butterworths, London.

Rijsbergen, C.J. van (2004), *The Geometry of Information Retrieval.* Cambridge University Press.

Salton, G., Wong, A. and Yang, C.S. (1968). A Vector Space Model for Automatic Indexing. *Communications of the ACM*. Vol. 18, No. 11, pp. 613-620.

Salton, Gerard (1971). *The Smart Retrieval System.* Prentice Hall.

Salton, Gerard, (1975). *Dynamic information and library processing.* Prentice Hall.

Salton, Gerard, and McGill, Michael (1983). *Introduction to modern information retrieval.* McGraw-Hill.

Salton, Gerard, (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Addison-Wesley.

Scha, R. and Polanyi, L. (1988). An Augmented Context Free Grammar for Discourse. *Proceedings of the 12th International Conference on Computational Linguistics.* Budapest, August 1988, pp. 573-577.

Scha, R., Bod, R. and Sima'an, K. (1999). A Memory-Based Model of Syntactic Analysis: Data-Oriented Parsing. *Journal of Experimental and Theoretical Artificial Intelligence (Special Issue on Memory-Based Language Processing, edited by Walter Daelemans).* Vol. 11, Nr. 3 (July 1999), pp. 409-440.

Scholtes, J.C. (1990a). Trends in Neurolinguistics. *Proceedings of the IEEE Symposium on Neural Networks*, June 21, Delft, Netherlands, pp. 69-86.

Scholtes, J.C. (1990b). *Neurolinguistics*. Computational Linguistics Project, CERVED S.p.A., Italy, 1990.

Scholtes, J.C. (1991a). Recurrent Kohonen Self-Organization in Natural Language Processing. *Artificial Neural Networks* (T. Kohonen, K. Makisara, O. Simula and J. Kangas, Eds.), pp. 1751-1754. Elsevier Science Publishers, Amsterdam, The Netherlands.

Scholtes, J.C. (1991b). Using Extended Kohonen-Feature Maps in a Language Acquisition Model. *Proceedings of the 2nd Australian Conference on Neural Nets*. February 2-4. Sydney, Australia, pp. 38-43.

Scholtes, J.C. (1991c). Learning Simple Semantics by Self-Organization. *Worknotes of the AAAI Spring Symposium Series on Machine Learning of Natural Language and Ontology*. March 26-29. Palo Alto, CA, pp. 146-151.

Scholtes, J.C. (1991d). Learning Simple Semantics by Self-Organization. *Worknotes of the AAAI Spring Symposium on Connectionist Natural Language Processing*. March 26-29. Palo Alto, CA, pp. 78-83.

Scholtes, J.C. (1991e). Unsupervised Context Learning in Natural Language Processing. *Proceedings of the International Joint Conference on Neural Networks*. July 8-12. Seattle, WA, Vol. 1, pp. 107-112.

Scholtes, J.C. (1991f). Self-Organized Language Learning. *The Annual Conference on Cybernetics: Its Evolution and Its Praxis*. July 17-21. Amherst, MA.

Scholtes, J.C. (1991g). Unsupervised Learning and the Information Retrieval Problem. *Proceedings of the International Joint Conference on Neural Networks*, November 18-21, Singapore., pp. 18-21,

Scholtes, J.C. (1991h). Filtering the Pravda with a Self-Organizing Neural Net. *Working Notes of the Bellcore Workshop on High Performance Information Filtering*. November 5-7, Chester, NJ.

Scholtes, J.C. (1991i). Kohonen Feature Maps and Full-Text Data Bases: A Case Study of the 1987 Pravda. *Proceedings of Informatiewetenschap 1991*. December 18. Nijmegen, The Netherlands, pp. 203-220. STINFON.

Scholtes, J.C. (1991j). Kohonen's Self-Organizing Map in Natural Language Processing. *Proceedings of the SNN Symposium*. May 1-2. Nijmegen, The Netherlands, p. 64.

Scholtes, J.C. (1991k). Kohonen's Self-Organizing Map Applied Towards Natural Language Processing. *Proceedings of the CUNY 1991 Conference on Sentence Processing*. May 12-14. Rochester, NY, p. 10.

Scholtes, J.C. (1992a). Neural Data Oriented Parsing. *Proceedings of the 2nd SNN*. April 14-15, Nijmegen, The Netherlands, p. 86.

Scholtes, J. (1992b). Neural Data Oriented Parsing. *Proceedings of the 3rd Twente Workshop on Language Technology*. May 12-13, Twente, The Netherlands.

Scholtes, J.C. (1992c). Filtering the Pravda with a Self-Organizing Neural Net. *Proceedings of the First SHOE Workshop.* February 27-28, Tilburg, The Netherlands, pp. 267-277.

Scholtes, J.C. (1992d). Resolving Linguistic Ambiguities with a Neural Data-Oriented Parsing (DOP) System. *Proceedings of the First SHOE Workshop*, February 27-28, Tilburg, The Netherlands, pp. 279-282.

Scholtes, J.C. (1992e). Resolving Linguistic Ambiguities with a Neural Data-Oriented Parsing (DOP) System. *Artificial Neural Networks 2* (I. Aleksander and J. Taylor, Eds.). Vol. 2, pp. 1347-1350. Elsevier Science Publishers, Amsterdam, The Netherlands.

Scholtes, J.C. (1992f). Neural Nets for Free-Text Information Filtering. *Proceedings of the 3rd Australian Conference on Neural Nets*. February 3-5, Canberra, Australia.

Scholtes, J.C. (1992g). Filtering the Pravda with a Self-Organizing Neural Net. *Proceedings of the Symposium on Document Analysis and Information Retrieval*, March 16-18, 1992, Las Vegas, NV, pp. 151-161.

Scholtes, J.C. and Bloembergen, S. (1992a). The Design of a Neural Data-Oriented Parsing (DOP) Model. *Proceedings of the International Joint Conference on Neural Networks*, June 7-10, Baltimore, MD.

Scholtes, J.C. and Bloembergen, S. (1992b). Corpus Based Parsing with a Self-Organizing Neural Net. *Proceedings of the International Joint Conference on Neural Networks*, November 3-5, Beijing, P.R. China.

Scholtes, J.C. (1992h). *Neural Nets versus Statistics in Information Retrieval*. A Case Study of the 1987 Pravda. Proceedings of the SPIE Conference on Applications of Artificial Neural Networks III, April 20-24, Orlando, FL.

Scholtes, J.C. (1993). *Neural Networks in Natural Language Processing and Information Retrieval*. Ph.D. Thesis, January 1993, University of Amsterdam, Department of Computational Linguistics, Amsterdam, The Netherlands.

Scholtes, J.C. (1994a). *Neural Networks in Information Retrieval in a Libraries Context*. State-of-the-Art report, PROLIB/ANN, DG XIII, European Commission, Luxembourgh.

Scholtes, J.C. (1994b). *Neural Networks in Information Retrieval in a Libraries Context*. Final report, PROLIB/ANN, DG XIII, European Commission, Luxembourgh.

Scholtes, J.C. (1995). *Report on knowledge and experience sharing for the International Fund for Agriculteral Development* (United Nations), Rome, May 1995.

Scholtes, J.C. (1996). *New Developments in Full-Text Retrieval*. Document 96. Birmingham, September 1996.

Scholtes, J.C. (2004a). What is the single most challenging Sarbanes-Oxley issue today? Ongoing vigilance. *Sarbanes-Oxley Compliance Journal*.

Scholtes, J.C. (2004b). XML for Archiving and Record Management. *The 25th Global Conference and Exhibit*. October 24-28, 2004, Dallas, TX, USA.

Scholtes, J.C. (2005a). Usability versus Precision & Recall. What to do when users prefer a high level of user interaction and ease-of-use over high-tech precision and recall tools. *Search Engine Meeting,* Boston, April 11-12, 2005.

Scholtes, J.C. (2005b). How end-users combine high-recall search tools with visualization. *Intelligence Tools: Data Mining & Visualization*, Philadelphia, June 27-28, 2005.

Scholtes, J.C. (2005c). Affordability in Content Management and Compliance. *Knowledge Management World. Best Practices in Enterprise Content Management,* May 2005.

Scholtes, J.C. (2005d). From Records Management to Knowledge Management. *Knowledge Management World. Best Practices in Records Management & Regularity Compliance*.

Scholtes, J.C. (2006a). Searching large E-mail collections: the next challenge. *The International Conference for Science & Business Information*, ICIC, Nimes, France. 22-25 October, 2006.

Scholtes, J.C. (2006b). A View on e-mail management. Balancing Multiple Interests and Realities of the Workplace. *Knowledge Management World. Best Practices in E-mail*, February 2006.

Scholtes, J.C. (2006c). Comprehensive eDiscovery and eDisclosure technologies. Next generation deployment of enterprise search tools. *Knowledge Management World. Best Practices in Enterprise Search*, April 2006.

Scholtes, J.C. (2007a). Finding Fraud before it finds you: Advanced Text Mining and other ICT techniques. *Fraud Europe 2007, Brussels, April 24, 2007.*

Scholtes, J.C. (2007b). E-Discovery and e-Disclosure for Fraud Detection. *Fraud World 2007, London, September, 2007.*

Scholtes, J.C. (2007c). Advanced eDiscovery and eDisclosure techniques. *Documation, The Olympia, London, October 2007.*

Scholtes, J.C. (2007d). Roundtable discussion, eDiscovery. David Cooper, J.C. Scholtes, Barry Murphy and Judith Lamonth. *Knowledge Management World,* September 2007.

Scholtes, J.C. (2007e). Efficient and Cost-Effective Email Management with XML. *Knowledge Management World, Best Practices in Email and IM Management*. February 2007.

Scholtes, J.C. (2007f). Mandated e-Discovery Requirement. Comliance Requires Optimal Email Management and Storage. *Today Magazine, the journal of Work Process Improvement*. March/April 2007. pp. 37.

Scholtes, J.C. (2007g). The Evolution of Enterprise Search. *Knowledge Management World. Best Practices in Enterprise Search*, May 2007.

Scholtes, J.C. (2007h). How to make eDiscovery and eDisclosure easier. *AIIM e-Doc Magazine*. Volume 21, Issue 4. July/August 2007. pp. 24-26.

Scholtes, J.C. (2007i). Where Records Management meets Enterprise Search and Knowledge Management: the bundle that optimizes discovery capabilities and supports profitability. *Knowledge Management World. Best Practices in Enterprise Search*, October 2007.

Scholtes, J.C. (2007j). Legal Ease. eDiscovery and eDisclosure. *DM Magazine UK*. November December 2006. pp, 26.

Scholtes, J.C. (2007k). Efficient and Cost-effective Email Management With XML. *Email Management*. (Ms.E jyothi and Elizabeth Raju Eds). Institute of Chartered Financial Analysts of India (ICFAI) Books.

Scholtes, J.C. (2008a). Is there a role for Sentiment Mining in Robot-Human Communications? *First International Conference on Human-Robots Personal Relationships.* June 12-13, 2008.

Scholtes, J.C. (2008b). Finding More: Advanced Search and Text Analytics for Fraud Investigations. *London Fraud Forum, Barbican, London*. October 1, 2008.

Scholtes, J.C. (2008c). Maintain Control During eDiscovery. *Knowledge Management World. Best Practices in eDiscovery*, February 2008

Scholtes, J.C. (2008d). Text Analytics—Essential Components for High-Performance Enterprise Search. *Knowledge Management World*. *Best Practices in Enterprise Search*, May 2008.

Scholtes, J.C. (2008e). De Hypothese, kolom in Vrij Nederland. 26 Augustus 2008.

Scholtes, J.C. (2008f). Records Management and e-Discovery: Why we need to re-learn Art of Information Destruction. *Knowledge Management World. Best Practices in Records Management and Compliance*. November 2008.

Scholtes, J.C. (2009). Understanding the difference between legal search and Web search: What you should know about search tools you use for e-discovery. *Knowledge Management World. Best Practices in e-Discovery.* January, 2009.

Sedona Conference: http://www.thesedonaconference.org/.

Segaran, T. (2007). *Programming Collective Intelligence, Building Smart Web 2.0 Applications.* O'Reilly.

Shanahan, J.G., Qu, Y., and Wiebe, J. (Editors), (2006). *Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series).* Springer-Verlag.

Socha, George (2009). *What does it take to bring e-Discovery in-house: risks and rewards.* Published at www.EDRM.org.

Spangler, Scott and Kreulen, Jeffrey (2008). *Mining the talk: unlocking the business value in unstructured information.* IBM Press/Pearson plc.

Sparck-Jones, K. (1971). *Automatic Keyword Classification for Information Retrieval*. Butterworths.

Spink, Amanda and Cole, Charles (Editors), (2005). *New Directions in Cognitive Information Retrieval.* Springer Verlag.

Steeb, W.H. and Hardy, Y. (2006). *Problems and Solutions in Quantum Computing and Quantum Information, 2nd edition.* World Scientific Publishers.

Sullivan, Dan (2001). *Document warehousing and text mining.* John Wiley and Sons.

Surowiecki, James (2004). *The Wisdom of Crowds.* Anchor Books.

Tait, John I. (Editor), (2005). *Charting a New Course: Natural Language Processing and Information Retrieval. Essays in Hobour of Karen Spärck Jones.* Springer-Verlag.

Tapscott, D. and Williams, A.D. (2006). *Wikinomics. How Mass Collaboration Changes Everything.* Portfolio (Penguin Group).

Tufte, Edward, R. (2001). *The Visual Display of Quantitative Information, 2nd edition.* Graphics Press.

Voorhees, Ellen M. (1985). The Cluster Hyphothesis Revisited. *Proceedings of the 8th Annual ACM-SIGIR Conference on Research and Development in Information Retrieval*. June 1985, pp. 188-196.

Voorhees, Ellen M. (Editor), Harman, Donna K. (Editor), (2005). *TREC: experiment and evaluation in information retrieval.* MIT Press.

Wang, James Z. (2001). Integrated *Region-Based Image Retrieval.* Kluwer Academic Publishers.

Weiss, et al. (2005). Sholom Weiss, Nitin Indurkhya, Tong Zhang, Fred Damerau. *Text mining: predictive methods for analyzing unstructured information.* Springer-Verlag.

White, Martin (2007). *Making Search Work. Implementing Web, Intranet and Enterprise Search.* Information Today, Inc.

Wilkingson, R., Arnold-Moore, T., Fuller, M., Sacks-Davis, R., Thom, J. and Zobel, J. (1998). *Document Computing. Technologies for Managing Electronic Document Collections.* Kluwer Academic Publishers.

Witten, I.H. and Frank, E. (2005). *Data Mining, Practical Machine Learning Tools and Techniques, 2nd. Edition.* Morgan Kaufman.

Woods, W.A. (1970). Transition Network Grammars for Natural Language Analysis. *Communications of the ACM*. Vol. 3, Nr. 10, pp. 591-606.

Wu, J.K., Kankanhalli, M.S., Lim, J.H., and Hong, D. (2000). *Perspectives on Content-Based Multimedia Systems.* Kluwer Academic Publishers.

Zvelebil, M. and Baum, J.O. (2008). *Understanding Bioinformatics.* Garland Science, Taylor and Francis Group LLC.

13 English Summary

This text is an extended version of my acceptance speech for the special chair of text mining at the department of knowledge engineering of the University of Maastricht, the Netherlands, presented Friday January 23rd, 2009 at 16.30 hours.

Text-mining refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining encompasses several computer science disciplines with a strong orientation towards artificial intelligence in general, including but not limited to pattern recognition, neural networks, natural language processing, information retrieval and machine learning. An important difference with search is that search requires a user to know what he or she is looking for while text mining attempts to discover information in a pattern that is not known beforehand.

Text mining is particularly interesting in areas where users have to discover new information. This is the case, for example, in criminal investigations, legal discovery and due diligence investigations. Such investigations require 100% recall, i.e., users can not afford to miss any relevant information. In contrast, a user searching the internet for background information using a standard search engine simply requires any information (as oppose to all information) as long as it is reliable. In a due diligence, a lawyer certainly wants to find all possible liabilities and is not interested in finding only the obvious ones.

Increasing recall almost certainly will decrease precision implicating that users have to browse large collections of documents that that may or may not be relevant. Standard approaches use language technology to increase precision but when text collections are not in one language, are not domain specific and or contain variable size and type documents either these methods fail or are so sophisticated that the user does not comprehend what is happening and loses control. A different approach is to combine standard relevance ranking with adaptive filtering and interactive visualization that is based on features (i.e. meta-data elements) that have been extracted earlier.

Other useful applications of text mining can be found in life sciences, compliance, consumer opinions on the internet and product guarantee analysis.

The extra-ordinary chair on text mining of the department of knowledge engineering of the University of Maastricht will focus on text mining methods for language-independent feature extraction for labeling text documents so that large result lists can be filtered and visualized in a meaningful way. Education and research will focus on the application of existing and development of new document classification techniques supporting multi-dimensional and sometimes hierarchical (taxonomy-based) classifications. Meaningful document classification will enable users to dynamically visualize and filter results so they can iteratively refine their search queries.