

# Text mining for portals:

Giving words a meaning

Master's Thesis  
MSc CoCo, cohort 98/01  
Student: Edwin van Dillen

Master of Science course Cooperative Computing

Kenniscentrum CIBIT  
Arthur van Schendelstraat 570  
Postbus 19210  
3501 DE Utrecht  
Tel: +31 - 230 23 08 900  
Fax: +31 - 230 23 08 999

Mentor:  
Timo Kouwenhoven

## Summary

Sources like the web provide people with a huge amount of information. Whatever the subject, information about it is stored somewhere. This makes it even hard to find the desired information. New techniques like text mining and collections of navigation forms like portals are introduced to optimise the search process.

In this thesis the role of text mining for personalised information portals will be discussed. This is done by determining the specifications of text mining techniques. They are related to the implementation requirements of portals.

Desk study shows that text mining techniques are based on information retrieval models in combination with linguistics. The semantic value of words in a particular document can be determined. Thereby, it becomes possible to automatic generate summaries of documents and add tags to particular parts of texts. Based on case based reasoning, a data mining related technique, it becomes possible to set-up an interaction with the user to retrieve one particular document.

There are many definitions of portals. Most of them differ in the perspective used to describe them. From a technical perspective, a component model can be defined that describes the components that portals exist of. The components that handle unstructured sources are of the most interest for text mining techniques.

Several user types for a portal can be classified: *Farmers* know exactly what they want; *Miners* determine whether a particular hypothesis can be supported; *Explorers* use a heuristically approach and *Tourist* are rather impulsive.

These user types can be offered a navigation form that is optimised to answer their question. A distinction is made in querying, browsing, categorising, localising, filtering and subscribing navigation forms.

By determining which types of users a portal will have to service the best navigation forms for that portal can be selected. The choice of an implementation of text mining techniques will depend on the navigation form that has to be offered.

Except for the localising navigation form, text mining can be used for all the requirements of the unstructured sources of a personalised information portal. For the localising navigation a simple Boolean information retrieval model will be sufficient.

Based on text mining techniques the users information desire can be personalised and profiles can be created. They can also create summaries of documents. Add tags to documents, chapters, paragraphs or any other particular part. By using proximities they can provide associative (i.e. browsing) navigation forms.

Autonomy is a tool that is intermediary of information retrieval and text mining. It can fulfil all the requirements of MYCIBIT.COM. MYCIBIT.COM is an example of a personalised information portal. In order to exploit Autonomy for MYCIBIT.COM successfully a knowledge steward is required. This person will fulfil an intermediary role between the users of the system and the technical operators. Thereby, determine how the facilities should be offered to the users.

## Preface

*“For years inventions have extended man's physical powers rather than the powers of his mind. Trip hammers that multiply the fists, microscopes that sharpen the eye, and engines of destruction and detection are new results, but not the end results, of modern science. Now instruments are at hand which, if properly developed, will give man access to and command over the inherited knowledge of the ages. The perfection of these pacific instruments should be the first objective of our scientists as they emerge from their war work.”* Dr. Vannevar Bush, July 1945 [Bush]

With these thoughts Dr. Bush encourage the development of techniques that help people to gain knowledge out of data.

Over the years a lot of experience has been gained of collecting new information or knowledge out of structured data sources. These techniques are referred to as data mining. Yet, they only work on structured sources, such as databases.

Text mining is a similar technique as data mining. Text mining techniques however can handle unstructured sources, such as textual documents.

What text mining techniques are and how they can be used for personalised information portals is the topic of this thesis.

### **Acknowledgements**

I would like to thank the people that helped me fulfilling this project: Cor Baars for the inspiration he gave me. Timo Kouwenhoven, my mentor during the project, for the discussions we had and his advices on the research process. Wilco Verdoold for the discussions we had and all the patience he had every time I was enthusiastic about the new things I had to discover for myself. I would like to thank all my colleagues who provided me with their inspiring input for this project.

In addition I would like to thank my friends, family, colleagues and acquaintances for their patience. In particular my diving buddies Josien and Wilco and my sister for excepting to miss my attention after the painful root canal treatment.

Special thanks to my parents for their support, motivation and the strength they gave me to fulfil all the educations. Dad: “This is it!”

Edwin van Dillen, October 2000

Edwin@van-dillen.com

Copyright © 2000, E. van Dillen

All rights reserved. No parts of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form by any means, electronically, mechanical, photocopying, recording, or otherwise, without the prior written permission from the author.

# Contents

<b>1. INTRODUCTION .....</b>	<b>6</b>
1.1 SUBJECT MOTIVATION .....	6
<b>2. RESEARCH .....</b>	<b>7</b>
2.1 INTRODUCTION .....	7
2.2 THE CHALLENGE .....	7
2.3 RESEARCH OBJECTIVES .....	8
2.4 RESEARCH METHOD .....	9
<b>3. TEXT-MINING .....</b>	<b>10</b>
3.1 INTRODUCTION .....	10
3.2 HISTORY .....	11
3.3 INFORMATION RETRIEVAL .....	12
3.4 DATA MINING .....	16
3.5 LINGUISTICS .....	19
3.6 TEXT-MINING .....	21
3.7 CONCLUSION .....	22
<b>4. PERSONALIZED INFORMATION PORTALS .....</b>	<b>23</b>
4.1 INTRODUCTION .....	23
4.2 WHAT IS A PORTAL? .....	23
4.3 COMPONENTS .....	25
4.4 USERS .....	26
4.5 CONCLUSION .....	28
<b>5. TEXT MINING AND PERSONALISED INFORMATION PORTALS .....</b>	<b>29</b>
5.1 INTRODUCTION .....	29
5.2 NAVIGATION AND TEXT MINING .....	29
5.3 CONCLUSION .....	32
<b>6. AUTONOMY .....</b>	<b>33</b>
6.1 INTRODUCTION .....	33
6.2 THE ARCHITECTURE .....	33
6.3 SOURCES .....	34
6.4 ABSTRACTS .....	36
6.5 PROFILES .....	37
6.6 CONCLUSION .....	39
<b>7. MYCIBIT.COM .....</b>	<b>40</b>
7.1 INTRODUCTION .....	40
7.2 HISTORY .....	40
7.3 MYCIBIT.COM USERS .....	41

<b>8. AUTONOMY AND MYCIBIT.COM.....</b>	<b>43</b>
8.1 INTRODUCTION .....	43
8.2 AUTONOMY FOR MYCIBIT.COM.....	43
8.3 EXPLOITATION OF AUTONOMY .....	45
<b>9. CONCLUSIONS AND FURTHER RESEARCH.....</b>	<b>47</b>
9.1 TEXT MINING.....	47
9.2 PERSONALISED INFORMATION PORTALS .....	47
9.3 TEXT MINING FOR PERSONALISED INFORMATION PORTALS .....	47
9.4 AUTONOMY AND MYCIBIT.COM.....	48
9.5 FURTHER RESEARCH.....	48
Appendix A: Project information.....	52
Appendix B: Questionnaire.....	53

# 1. Introduction

*"Obstacles are those frightful things you see when you take your eyes off your goal."*

Henry Ford (1863-1947)

## 1.1 Subject motivation

As the name of the study cooperative computing already reveals, a central aspect of the study is cooperation: between men, between man and machine and among machines.

One of the aspects dealt with during the study was the appliance of personal agents as assistants for humans. For instance, a personal agent can assist a user in order to find interesting information out of a large collection of documents like the Internet.

In order for the personal agent to find documents, the agent has to know what a document describes. Hence, it has to determine the topic of the document. Text-mining is a technology, which can be used to determine the documents topic, also referred to as the document's concept.

Beside that there was the uprising of the phenomenon portal. In the beginning a portal was not more or less then a website that a user could access in order to find the page describing the topic the user was interested in. Search engines like Yahoo became portals.

The concept of portals matured and a distinction was made between the types of users the portal was aiming at. Business portals arose serving their customers – B2C-, serving other businesses –B2B- and serving their own employees –B2E-.

In general portals offer their users two types of information, which is structured or unstructured information. In this sense plain textual documents are unstructured information. In order to know the topic the document describes, it needs to be interpreted. For this purpose text mining can be applied.

For personalised information portals text mining techniques can be used to provide a document summary to the user. They can determine the relation between the subject of documents and thereby retrieve the documents desired by the user.

This master thesis will describe the role of text mining for personalized information portals. Furthermore, the role of agents and navigation will be discussed in general.

Hence the focus of this thesis is mainly on the cooperation between, man and machine.

## 2. Research

*"Research is what I'm doing when I don't know what I'm doing."*

Wernher Von Braun (1912-1977)

### 2.1 Introduction

In this chapter the research project will be described. In paragraph 2.2 the challenge of the research project will be described. The objectives will be described in the form of the research question in paragraph 2.3. Finally, in paragraph 2.4 the research model will be described. Furthermore, the structure of this thesis will be described in the former paragraph.

### 2.2 The challenge

The growth of Internet technologies has a great influence on the accessibility of information. Whatever the subject, information about it is available. Nowadays there is so much information available that it becomes harder to find the specific information a person desires.

In order to find information a person desires, search facilities are provided. These facilities reach from key word search engines to directory services that are manual build. These services relieve the user by providing –groups of -documents that are of interest.

One of the disadvantages of these technologies is the low quality of their results. The reason for this is the approach they use to determine whether a document answers the user's question<sup>1</sup>.

New techniques are available that should be able to determine the concept of a document and thereby the relation between the document and the search question. These new techniques are referred to as Text-mining.

What text mining really is, how it works, whether it will deliver better results than conventional techniques and how these techniques can be applied for personalized information portals are the questions which will be addressed in this thesis.

---

<sup>1</sup> This is one of the disadvantages of key word searches, which came up during the research process. This will be elaborated in chapter: 3 "Text-mining"

## 2.3 Research objectives

The main research objective of this Master Thesis is to answer the question:

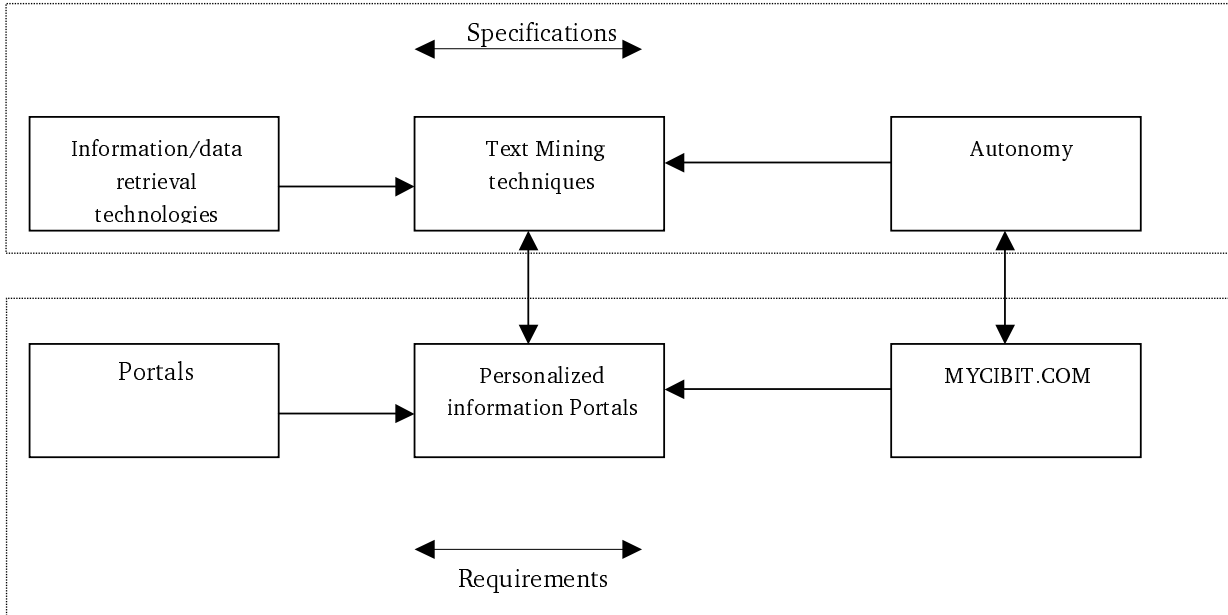
*Are Text-mining based technologies capable of fulfilling the implementation and exploitation requirements applied on personalized information portals?*

In order to answer this question the following detail questions will be answered:

1. What kind of technology is text mining? (Descriptive, Chapter 3 and 5)
  - 1.1. How does text mining relate to technologies as: case based and rule base reasoning, filtering, profiling, agents, indexing and data gathering? (Descriptive)
2. What requirements apply on a personalized information portal? (Descriptive Chapter 4 and 5)
  - 2.1. Which different types of portals can be distinguished? (Descriptive)
  - 2.2. Which tasks does the implementation and exploitation of a personalized information portal comprehend? (Descriptive)
3. Is Autonomy representative for an implementation of text mining technology? (Evaluative, Chapter 6 and 8)
4. Is MYCIBIT.COM representative for a personalised information portal in general? (Evaluative, Chapter 7)
5. To what extend does Autonomy fulfil the need of automated gathering and personalising of information for MYCIBIT.com? (Evaluative, Chapter 8)

## 2.4 Research method

In the figure below a visual representation of the research is given:



The first step will be desk research in order to define what text mining techniques are. This will be based upon techniques used for information retrieval data mining and linguistics in the broadest sense. See chapter 3.

Also in the form of desk research personalized information portals will be discussed in chapter 4. What personalized information portals are and their requirements will be described.

In chapter 5 the relation between text mining and personalized information portals will be discussed. The specifications of text mining techniques will be matched on the requirements of personalized information portals.

This ends the theoretical desk research of this thesis. In the next part the case Autonomy for MYCIBIT.COM will be evaluated. Autonomy is an example of a text mining tool, which can be used for portals. MYCIBIT.COM is an example of a personalised information portal.

In chapter 6 the architecture and functionalities of Autonomy will be described. The question whether Autonomy is a real text mining tool will be answered.

In chapter 7 MYCIBIT.COM will be discussed and the question will be answered whether MYCIBIT.COM is a personalized information portal.

In chapter 7 the specifications of Autonomy will be matched on the requirements of MYCIBIT.COM.

Finally, the end conclusion will be drawn in chapter 9.

## 3. Text-mining

*"I begin by taking. I shall find scholars later to demonstrate my perfect right."*

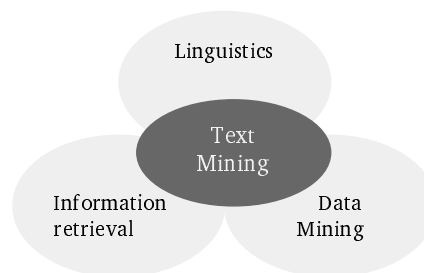
Frederick (II) the Great

### 3.1 Introduction

The objective of this chapter is to determine what kind of technique text mining is. There are several perceptions of text mining. Some think of it as an advanced technique to retrieve textual documents out of a large collection. [DWR]

Others think of gaining new information or knowledge from text documents by finding relations between them. This group focuses on the “mining” metaphor. *Finding the nuggets!* [Hearst-1]

A third group sees it from the – computational- linguistic perspective and wants to find patterns within collections of text documents. These patterns are used to form algorithms for various sub problems within natural language processing. [Hearst-1]



These three perceptions have in common that the meaning of a text document is determined. What is done with the results after the meaning has been determined differs per perception. Ranging from retrieving the document to generate new information or knowledge.

In paragraph 3.2 the history of information retrieval will be given. This will show how techniques have evolved over the years. In the following paragraph several information retrieval models will be discussed. Next data mining will be discussed. This will revile that a lot of the techniques used for data mining have a strong resemblance with the models used for information retrieval.

In paragraph 3.5 some linguistic techniques will be discussed. It is the combination of these techniques and the information retrieval and data mining models that enable Text Mining. Finally, in paragraph 3.6 real Text Mining will be described. Showing the relation between the former subject areas and the role they fulfil for unstructured sources will do this.

## 3.2 History<sup>2</sup>

Information retrieval was started from the idea of Vannevar Bush's in 1945. During that time, there was a tension between simple statistical methods and sophisticated information analysis [Bush].

In 1949, Warren Weaver thought about a feasibility of translating languages by computers. Where Bush talked of intellectual analysis, both by people and by machines. Weaver reacted to the success of mathematicians in cryptography. The analytical process, the Bush approach, can either use manual indexing or try for artificial intelligence programs that will archive a good accuracy of information identification. The accumulation of statistical detail in Weaver's approach can be done entirely with probabilistic retrieval techniques.

The 1960s were time of great experimentation in information retrieval systems. During this period the definition of *recall and precision* and the development of the technology for evaluating the performance of retrieval systems were also built. New retrieval techniques called "*relevance feedback*" were developed. The idea augmenting the user's query by adding terms from relevant documents.

In the 1970s the first real-time systems were introduced. Users could send their enquires directly from terminals and get answers within a few seconds. Through the upcoming of databases the amount of research on information retrieval declined in this period.

In the 1980s word processors became commonly used. Thereby, the amount of textual documents increased rapidly. On the other hand the cost of disk space declined and new storage devices were introduced, such as CD-Rom. The research on information retrieval relived again.

During the 1990s the popularity of the Internet exploded, with one million people signing up each month and still increasing. With services like the World Wide Web, where a lot of unstructured documents are presented, the need for full-text search algorithms increased.

At this time many information retrieval techniques are available. They range from the traditional ones as full text scanning, inversion, signature files and clustering to semantic information techniques as natural language processing, latent semantic indexing and neural networks. With semantic information techniques patterns of unstructured documents can be made which can be used to extract new information or knowledge from the documents. This technique is referred to as text mining.

---

<sup>2</sup> The historical facts of information retrieval, which are part of this chapter are taken from the work of Dinh van Dung [Dung].

### 3.3 Information Retrieval

#### 3.3.1 What is Information retrieval?

*“Information retrieval is the study of systems for indexing, searching, and recalling data, particularly text or other unstructured forms” [Weiss]*

*“Information retrieval (IR) deals with the representation, storage, organization of and access to information items.... The user of an IR system is concerned more with retrieving information about a subject than with retrieving data which satisfies a given query” [Baeza-Yates]*

The difference between these two descriptions of information retrieval is the precision. [Weiss] speaks of the retrieval of data. What the relevance of this data is for the user is not mentioned. [Baeza-Yates] focuses on the retrieval of information. That is, information that is relevant to the users query.

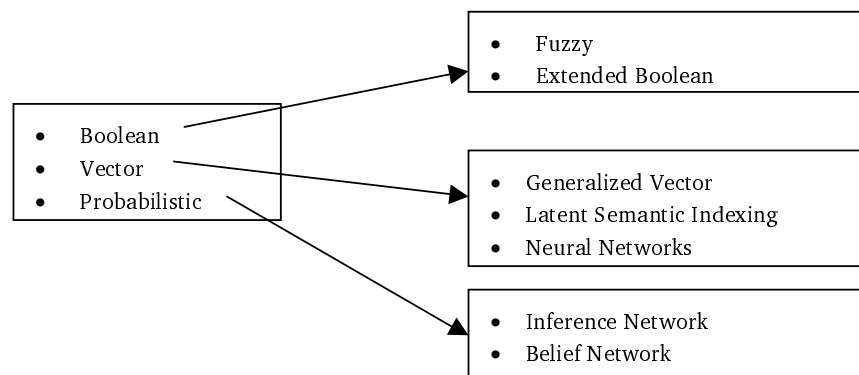
#### 3.3.2 Information Retrieval techniques

In order to retrieve a document a set of representative keywords are extracted from the document. These keywords are referred to as index terms. In general nouns are used as index terms. The noun has a semantic value on itself in contrast to adjectives and adverbs. However, it might be interesting to index all the words in the document in the form of a full text search. [Baeza-Yates]

In order to retrieve a document from the index, the user's query words are matched with the index terms. When the index is based on all the words of the documents (full text search) a query word may appear in all the documents. For the retrieval of a document this query word does not have any value. After all, the query word would result in all the documents. In order to overcome this problem two measures are taken:

1. *The introduction of a “stop list”*  
The stop list contains all the words which have no value on there own for an index. Articles (the, a, an) are typical words in the stop list.
2. *The introduction of numerical weights*  
By assigning a numerical weight to a query word and an index term a more precise match can be made.

There are three types of models to match the index terms and the query words. (See the figure below) The use of a stop list is for all of these models the same. This does not count for the way they handle the numerical weights.



### 3.3.2.1 Boolean

The Boolean models are the oldest of the three models. The Boolean models are easy to understand for a common user of a retrieval system. The queries are based on concepts from logic, i.e. Boolean algebra, with its terms joint together by logical operators. Typically the operators permitted are AND, OR and NOT. [Korfhage]

At one hand the results of a Boolean query can easily be predicted. The weight of an index term is either 1 or 0, which means it is in the document or not. So, in order for a document to be retrieved from a Boolean index the document must fully apply to the user's query. This is what makes the use of a Boolean retrieval system quite easy for a common user.

On the other hand Boolean models have some drawbacks. This concerns synonymy and polysemy of words:

- *Synonymy*  
Synonymy addresses the problem that there are many ways of referring to the same concept. [Deerwester] Furnas, Landauder, Gomez & Dumais showed that across people, the same word is used by two people to describe an object only 10 to 20% of the time. [Foltz]  
This suggests that Boolean keyword matching on its own in text documents may fail. A document may discuss the subject the user is interested in but due to the fact that the query word does not match with any of the index terms the document is not retrieved.
- *Polysemy*  
polysemy addresses the problem that one word can have more than one meaning. [Deerwester] The meaning of the word depends on the context it is used in. Searching on the isolated words could result in documents that match the user's query but have a totally different topic than the user might expect. For instance, a query with the word virus in it can result document about the latest computer viruses. If the user is a medical student these results will be disappointing.

By applying a thesaurus in combination with Boolean information retrieval the problem of synonymy can be reduced. Before the user's query word is matched on the index term, synonyms of the index term are retrieved from the thesaurus. The chance that the query word matches the index term and the results of the thesaurus is much higher. In this case a thesaurus is used for the index terms.

In order to reduce the problem of polysemy a thesaurus for the user's query word should be used in combination with feedback from the user. The query word of the user is retrieved from the thesaurus. The results from the thesaurus will be provided to the user. The user can choose the best fitting query words for his query. Finally, the extended query will be matched with the index term. For the virus example this would mean that the user would send the query word "virus" to the Boolean information retrieval system. The system will retrieve synonyms from the thesaurus and show them to the user. In this case the user will select "medical virus" and the extended query will be executed.

This shows that the problem of synonymy and polysemy can be reduced or even overcome by using a thesaurus. But in order to apply a thesaurus it has to be created and kept up to date. At this time this is still a manual process. This makes it rather complex to build a general retrieval system. Yet, for a retrieval system within a particular subject area it may be sufficient and feasible. The complexity of the thesaurus for a particular subject area is well manageable.

The application of the Boolean model has the some advantages and disadvantages:

**Advantage**

- *results can be predicted easily*  
The user can easily predict the results of the query. If the results do not match the need of the user, they can adjust the query so it will deliver better results.

**Disadvantage**

- *There is no good way to weight the terms for significance*  
In the Boolean algebra the term is either present or absent.
- *Wrong results through misstated queries*  
Due to the complexity of the queries statements and interpretation it can be hard for inexperienced users to created well-formed Boolean queries. For instance, to form a query which results documents about groupware and workflow on one hand and groupware and e-mail on the other hand the following query must be created.

Groupware AND (Workflow OR E-mail)

- *Controlling the size and composition of the retrieved set*  
The query may result a very small number or a very large number of results. In the latter case, the system may present the user with several hundred documents to examine, in no particular order.

3.3.2.2 Vector

In contrast with the Boolean model in the vector model the weight of a term is non-binary. This enables a more precise relation between two documents.

Based upon the index terms collected from a document a vector represents the document. Of all the documents that are indexed in a vector based retrieval system the vectors are placed in a vector space.

In order to find a document that fulfils the user's query, a vector of the users query is calculated. The distance between the user's vector and the document vectors in the matrix determines the relation between them.

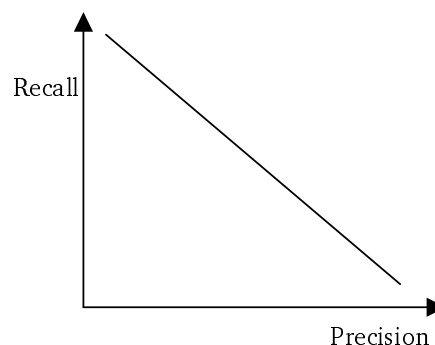
Due to the fact that the relation between the user's query and the indexed documents can be expressed as a value it is possible to add a threshold to the query. By changing the threshold the user can influence the measure in which the document matches the user's query. This approach has two main advantages:

1. *The query and the document can match just partially*  
This means that a user's query does not have to match 100% with the document in order to be retrieved. With the threshold the variance between the user's query and the document can be set.
2. *The user can control the amount of documents retrieved*  
By enlarging or decreasing the threshold the user can influence the amount of documents that are retrieved.

Of course the two aspects are related. Increasing the amount will decrease the precision of the results and vice versa. In the information retrieval theory this is expressed with the terms recall and precision. [Jansen]

- *Recall* = Relevant Documents Retrieved / Total Number of Relevant Documents
- *Precision* = Relevant Documents Retrieved / Total Number of Retrieved Documents

The application of the vector model has the some advantages and disadvantages:



#### **Advantage**

- *Non-binary weights*  
The vector models uses non-binary weights. Allowing partial matching, which means retrieval of documents that approximate the query conditions. The documents do not have to match 100% with the query, 90% may be enough to retrieve it.
- *Ranking results*  
Its ranking formula sorts the documents according to their degree of similarity of the query.

#### **Disadvantage**

- *Results must be within the Threshold*  
To determine whether the document is interesting to retrieve it should be above the threshold. In comparison with the probabilistic model this could be seen as a disadvantage of the vector model<sup>3</sup>.

#### 3.3.2.3 Probabilistic

*One concern about both Boolean-and vector-based matching is that they are based on “hard” criteria. In Boolean-based matching, either a document meets the logical conditions or it does not; in vector-based matching, a similarity threshold is set and either a document falls above that threshold or it does not. [Korfhage]*

The base of probabilistic matching is that it should be possible to calculate the probability that a document is relevant to the query. In order to calculate the probability the user is presented with a set of documents out of the document collection that have a relation with the query.

The user then selects the documents, which he finds the best. The probabilistic model then uses these documents to refine the query results and presents the refined results to the user. Again the user can select the documents, which he finds the best. After iterating this process several time the results will be increasingly better matching the users query.

The application of the probabilistic model has the some advantages and disadvantages:

---

<sup>3</sup> Within the probabilistic the “limited” whether a document is interesting is calculated and thereby based upon the available documents.

### **Advantage**

- *Ranking results*  
This is the same advantage as the vector model has. Though, the results are matched in decreasing order of their probability of being relevant.

### **Disadvantage**

- *Results are hard to predict*  
The probabilistic model is rather complex. This makes it harder for a user to predict the results and thereby to adjust the query words.
- *Guess the initial results*  
The need to guess the initial separation of documents into relevant and non-relevant sets is a disadvantage of the model. Yet, after a while the system will get “experience” and the initial results will become better.

### **3.3.3 Conclusion**

By comparing the three types of models the conclusion can be drawn that the results of a Boolean model are different than those of a vector model or probabilistic model.

Retrieval systems using the Boolean model are rather data retrieval systems than information retrieval systems. Within the Boolean models there is no attention for the correlation between the keywords and the document. That is, a Boolean model may retrieve a document in which the keyword is represented, but they are not necessarily the topic. For instance, in this document there is only referred to latent semantic indexing but is not discussed. A user who wants to know what latent semantic indexing is would be disappointed when this document is retrieved.

When using a vector model or a probabilistic model the index terms are related to the documents they are used in. By doing so the context in which the terms are used are determined. This increases the chance that the retrieved documents discuss the subject the user is interested in.

## **3.4 Data Mining**

### **3.4.1 What is Data Mining?**

*“Data mining enables the end user to extract useful business information out of large databases.” [Berson]*

*“Data mining is the iterative process of extracting patterns from your business and customer interaction. Above all, data mining is about leveraging artificial intelligence technology toward a strategic objective: competitive intelligence. It's about increasing your market share by knowing who your most valuable clients are, defining their features and then using that profile to target new customers.” [Mena]*

In these definitions two aspects are of particular interest:

- *“Out of large databases”*  
This indicates that data mining is concerned with a large collection of data. The fact that the data is stored in databases indicates that there is some sort of structure. That is, the database schema describes the structure of the data on a meta level.  
This makes it interesting to see whether the techniques used for data mining can also be used for unstructured data, such as textual documents.
- *“Process of extracting patterns”*  
With the extraction of patterns there is some form of relation described within the data.  
For unstructured documents pattern recognition could have two purposes. At first to discover a pattern within the documents themselves. This would have a strong resemblance with the vector and probabilistic models used for information retrieval. That is, by determining the pattern within the documents themselves they can be retrieved. Second, these patterns can also be used to determine the relation among textual documents. In this case new information or knowledge could be extracted from the relation of the text documents. Thereby, data mining techniques are used to mining collections of text documents.

### 3.4.2 Data Mining techniques

Based on the work of [ID] and [Thearling] a classification of data mining techniques as shown below can be made. Some of these techniques are based on the same mathematical algorithms as those that have been discussed in the paragraph “Information Retrieval techniques”. For the explanation of these algorithms there will be referred to that paragraph.

- **Logical**  
Rule-based systems and Decision Trees are examples of logical techniques. These are techniques that use logical algorithms, such as IF THEN statements, to determine a pattern within the dataset. Each condition has to be tested in order to take a next step in the process to build a pattern. This works fine for numerical systems and through the simplicity of the algorithms it is possible to solve rather complex patterns. Nevertheless, for the complexity textual documents offer these techniques are not sufficient.
- **Cross Tabulation**  
Belief nets are an example of cross tabulation techniques. They are part of the probabilistic model. (see paragraph “3.3.2 Information Retrieval techniques”)
- **Equational**  
Statistics and Neural Nets are equational techniques. These are respectively based on the probabilistic model and the vector model as discussed in 3.3.2 Information Retrieval techniques

- **Case-based reasoning (CBR)**

CBR is an approach to retrieve problem-solving experiences. When a problem is successfully solved, the experience is retained in order to solve similar problems in the future. When an attempt to solve a problem fails, the reason for the failure is identified and remembered in order to avoid the same mistake in the future. [Aamodt]

On an abstract level the CBR process can be described by four main steps: [Inreca]

1. Retrieve the most similar case(s)
2. Reuse the case(s) to attempt to solve the new problem
3. Revise the proposed solution if necessary
4. Retain the new solution as a part of a new case

In order for a case to be retrieved a CBR uses the nearest neighbour retrieval technique. This means that the stored cases are clustered based on the relation among each other.

The cases themselves can be stored in several forms. For instance, in a conversational CBR approach, cases are lists of questions and answers. The textual CBR approach is another form in which cases are represented in free textual format. The former approach is very interesting in relation with text mining. Instead of manually entering the cases in the form of lists of questions and answers. The cases can be stored in plain textual format. Text mining techniques can then be used to generate the necessary questions automatically from the text.

### 3.4.3 Conclusion

From the description of the Data Mining techniques two main conclusions can be drawn:

1. *Most of the data mining algorithms have a direct relation with information retrieval algorithms*

The same algorithms are used for two different kind or sources:

- a. unstructured –text- documents to retrieve with information retrieval
- b. structured sources that are used for data mining.

Besides the fact that the sources are different the objectives are also different. Is the focus of information retrieval to retrieve a document or a set of documents, the focus of Data Mining is to find patterns in a set of data.

Combining these two objectives results in a subject area, which focuses on finding patterns in a set of unstructured –text- documents.

2. *CBR-systems are related with text mining*

The combination of CBR-systems and text mining is an interesting development. With text mining it should be possible to determine the concept of a document (or when it is a large document, the concept of the chapter or paragraphs within the document). This means that entering of new cases in a CBR-system could take place automatically.

The learning aspects of the CBR-system and the interaction the system has with the user are unique. After all, the objective of the system is to retrieve a specific case. Hence, it has a strong relation with information retrieval. Yet, CBR not only retrieves the case but also learns during the process. (Feedback into the system)

### 3.5 Linguistics

#### 3.5.1 What is linguistics?

*“Linguistics, the study of language, concerns itself with all aspects of how people use language and what they must know in order to do so.”*[Nunberg]

This definition gives an impression of the extension of the study of linguistics. For this research project there are two main aspects, which are of interest. First the field of morphology and grammar. Second the field of computational linguistics.

#### **Morphology and grammar** [Nunberg]

Morphology is the study of internal structure of words, phrases and sentences. Many words are made up of smaller meaningful units, such as stems and suffixes; for example: stem 'glad' + suffix '-ly'.

In a text document a word can occur in many different forms. For example, mine, miner, mining, mined, etc. In order to index textual documents stemming algorithms are used to bring back the many different occurrences of the word to its stem. By doing so the complexity of the index and matching process are reduced.

Grammar is the study of the rules governing combinations of words cluster together into phrases, which combine to make sentences. These rules are also referred to as the semantics.

#### **Computational Linguistics** [Hobbs]

Within the computational linguistics a distinction is made in:

- *The technological perspective*  
to enable computers to be used as aids in analysing and processing natural language.
- *The psychological perspective*  
To understand, by analogy with computers, more about how people process natural language. This is out of the scope of this research project.

From the technological perspective, there are, broadly speaking, three uses for natural language in computer applications:

1. machine translation
2. natural language interfaces to software (also known as natural language processing)
3. document retrieval and information extraction from written text (Information Theory).

Of the former two applications some techniques will be discussed.

### **3.5.2 Linguistic techniques**

#### 3.5.2.1 Natural Language Processing

The field of natural language processing is divided into two sub fields [Turban]:

1. Natural Language *understanding* investigates methods of allowing the computer to comprehend instructions given in plain English so that computers can understand people more easily.
2. Natural Language *generation* strives to have computers produce ordinary English language so that people can understand computers more easily. These techniques can be useful to automatically generate summaries from text documents.

#### 3.5.2.2 Information theory

A text document is a collection of words in a particular combination (grammar). Words can have several appearances within that document (morphology). That is, the stem of the word with or without a suffix. Besides the appearances the context in which the word is used determines the meaning of it. In his so-called "information theory" Claude Shannon has described some mathematical rules, which can determine the signal-to-noise ratio. This ratio describes the weight of a word in relation to a document or collection of documents. In a simplified form this means that if a word is mentioned in most of the documents that are in the collection then the weight will be low. If a word is mentioned in just a few documents in the collection it will have a higher value.

### **3.5.3 Conclusion**

The most interesting role of computational linguistics is the weight determination of a word or combination of words in a textual document. That is, by applying information theory during the process of abstracting index terms of a textual document the relevance of the index term in relation to the document itself and the documents already indexed can be determined. As mentioned before the weight can be used to determine the relevance of a document for a particular query.

On one hand with natural language processing an application can communicate with the user in a natural language, for instance English. The user can query the system by applying a question in his/her own language. On the other hand a system will be capable to formulate questions in a natural language, which can be applied in for instance CBR-systems.

## 3.6 Text-mining

In the former paragraphs the three pillars on which text mining rests are discussed. The question remains "What is real text mining?".

### 3.6.1 Text mining and Information Retrieval

The first step of text mining, determining the concept of the document, is also known within information retrieval. As discussed in paragraph 3.3.2 information retrieval techniques such as vector-based and probabilistic models are capable of determining the –semantic- value of words within a document<sup>4</sup>.

In the terms of information retrieval value of the words within the document is used in order to determine whether the document should be retrieved or not. The second step of text mining –finding new information or knowledge based on the relation between different concepts- is not applicable for information retrieval. This leaves a gap in between real text mining a real information retrieval. In other words there is a gap between retrieving a document and gaining new information or knowledge for a document. This gap is being filled with new functionalities, which are being worked on now. [Hearst-2] These include:

- **Cut and Paste Tools** to help a researcher create summaries of material from multiple sources.
- **Document Intersection Finders** to help a researcher find where multiple documents (of a generally divergent content) have passages dealing with the same content.
- **Question Answering Tools** based on semantic parsing of documents.

### 3.6.2 Text Mining and Data Mining

The relation between text mining and data mining does not need much further explanation. Both techniques try to find new information or knowledge out of a collection of data. The difference is that text mining focuses on unstructured sources like textual documents and data mining on structured sources like databases.

There is one technique that should have some special attention and is often mentioned as a data mining technique: Case based Reasoning (CBR). CBR is especially interesting because the cases are often based on text. Text mining techniques can be used to enter –created- new cases, which can be used in the CBR system. CBR it self can be considered as an intermediary technique as discussed in the relation between text mining and information retrieval. CBR is an example of a questioning and answering tool. In contrast with information retrieval techniques Case based Reasoning can deliver one specific document, where information retrieval techniques would retrieve multiple documents.

---

<sup>4</sup> Within these models there are still some differences that have to be taken into consideration when choosing one for a particular implementation. Some are better than others in determining the semantic value of a word. This research project will the focus on the behaviour of the models.

### **3.6.3 Text mining and Linguistics**

Text mining uses on the one-hand linguistics techniques to determine the concept of a document and on the other hand to interact with the user in natural language.

As discussed above linguistic techniques as morphology, grammar and information theory fulfil an important role to determine the semantic value of words with in a document. The semantic value of words is the basis to determine the concept of a document.

### **3.7 Conclusion**

Real text mining is the process of gaining new information or knowledge from a textual document or a set of textual documents. In order to do this automatically, first the concept of a textual document must be determined. The concept describes the subject, which is discussed in the document. Based on this concept relations among documents can be discovered. These relations will deliver new information or knowledge about the subjects discussed in the documents.

## 4. Personalized information Portals

*"The important thing is to not stop questioning."*

Albert Einstein

### 4.1 Introduction

In this chapter personalised information portals will be discussed in such a way that at the end of this chapter the requirements to implement them can be given.

This will be done by first describing what portals are. In paragraph 4.3 the components out of which a personalised information portals consist will be described. By doing this it will become clear what the role of unstructured sources are within portals. In paragraph 4.4 the role of the user and navigation will be discussed.

The different types of navigation will form the requirements that text mining techniques have to fulfil. This will be discussed in the next chapter.

### 4.2 What is a Portal?

There are many definitions for portals. Yet, there are many different perspectives from which portals are described. Two of them are the user and functional perspective:

1. *User perspective*  
From the user perspective the following portal types can be distinguished: [Ruber]
  - *B2B*  
The B2B (business to business) portals are set-up for the exchange of products, services, or information between businesses. For instance a portal of a supplier with stock information that can be used by a retailer.
  - *B2C*  
The B2C (business to consumer) portals are set-up by a business to sell products to consumers and provide them with the information to do so. For instance a portal where books are soled.
  - *B2E*  
The B2E (business to employee) portals are set-up by an organisation to server the employees. The portal is the entrance into the organisation for the employees. Typical services are discussion facilities and learning applications.
  - *Vertical portals*  
Vertical Portal are also addressed as Vortals. Vortal refers to web site that aggregates disparate content and services of interest to a particular industry and makes it available to industry members.[Spitzer]

2. *Functional perspective*

From the functional perspective of enterprise portals Mark Davydov distinguishes: [Davydov]

- Enterprise Business Intelligence Portals (EBIP) provide the central launching point for corporate decision-processing and content-management applications. Their primary focus is to connect users with structured and unstructured content relevant to them.
- Enterprise Collaborative Processing Portals (ECPP) connect users not only with all the information they need, but also with everyone they need. ECPPs consolidate groupware, email, workflow, and critical desktop applications under the same gateway as decision-processing and content-management applications. ECPPs are characterized by “virtual project areas” or communities.
- Enterprise Mission Management Portals (EMMP) provide a “digital expertise-oriented workplace,” a highly specialized and personalized Web site where everything a user team needs (such as access to ERP applications, productivity and analysis tools, and relevant internal and external content) to effectively manage mission-critical management activities such as customer relationship management (CRM) is consolidated and made accessible via the Web.
- Enterprise Extended Services Portals (EESP) do everything the first three types do, but focus on providing comprehensive job support from the standpoint of “virtual enterprises” by creating communities and “virtual service spaces” of channel partners, suppliers, distributors, and customers

These two kinds of portal perspectives have some commonalities that are of interest:

- *Components*  
Portals offer information to users, whether it is about products or services which are offered or the location where they can be found. In order to offer this information the portals consist of some particular components. For instance personalization, navigation, etc.
- *Different type of users*  
Portals are used by different kind of users, who have their own information needs. The users can range from a consumer who wants to buy a product from an auction (B2C) to a purchaser who needs to buy office supplies for his enterprise (B2B).

In the context of this thesis portals will be discussed from the perspective of their users and the components they encompass. This will result in the requirements that have to be fulfilled in order to set them up.

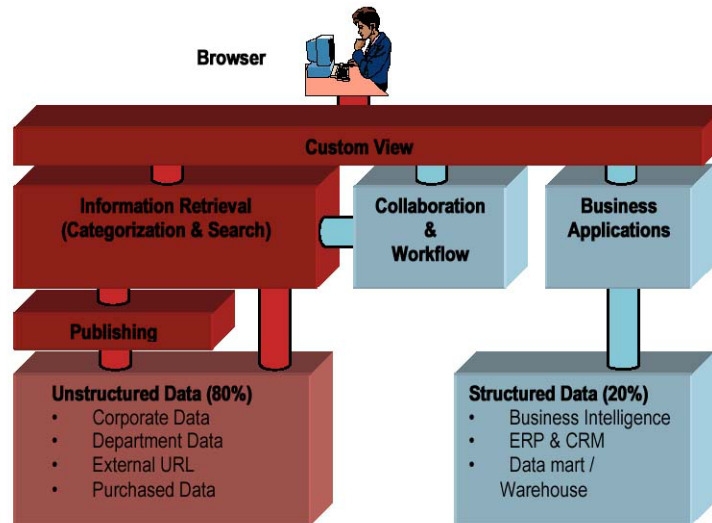
In the following paragraphs first several portal components will be discussed. Next a user classification will be discussed.

### 4.3 Components

Describing the components that functional portals consist of as discussed in the former paragraph results in the following list [Davydov]:

- Data filtering and analysis
- Information brokering
- Data mining
- Workflow management
- Document management
- Mission and task management
- Simulation and gaming
- Collaborative application integration
- Personal assistance

In order to implement these functional components several techniques are required. Another approach to describe the components a portal can consist of is from the technical perspective. This is done by Semio in the figure below. [Semio]



Depending on the objective of the portal a mixture of these components should be applied. For this research project the distinction between the components will be sufficient. In order to determine how the mixture of the components should be optimised for a particular portal requires further research.

For the remainder of this thesis the technical perspective of portals will be used. In particular the components that are concerned with unstructured data sources. Text mining techniques can be used for these data source.

## 4.4 Users

In this paragraph the following question will be answered, “What kind of questions need to be answered in order to fulfil the information desire of a user who uses unstructured sources of a portal?” This will reveal the functionalities that have to be provided by a portal.

In order to answer this question a distinction has to be made of the types of users of a portal. Followed by the types of navigation, which should be offered in order to set-up the interaction with the users. Finally, the relation between the type of users and the navigation forms can be determined.

### 4.4.1 Types of users

In the world of business intelligence Bill Inmon distinguishes four types of users<sup>5</sup>: Farmers, Explorers, Miners and Tourist. [Inmon]

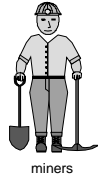
Business intelligence technologies are typically applied on structured sources and as shown in the former paragraph portals contain both structured and unstructured sources. The user classification as described by Bill Inmon also applies on the unstructured sources. The user's need for information does not differ depending on the type of source. The difference between these sources is the way the information is extracted from them (see chapter 3) and the type of navigation forms the user is offered. Navigation forms will be discussed in the next paragraph.



- **Farmers**  
The farmer is a person who knows exactly what he/she wants. These users are able to state their information need in a concrete question.



- **Explorers**  
Explorers have a need to look at details. They operate by a heuristic approach, which means that the next step analysis is dictated by the results obtained in the former step. An explorer may go six months with no queries and then submit a query every hour for the next week.



- **Miners**  
The miner is an individual who examines the truth of a hypothesis by looking at a lot of data and determine whether the data supports a hypothesis. The miner is fully capable of doing the work of the explorer. However, the miner normally operates on very large, very focused mass of data.



- **Tourists**  
Tourists have the least structure of all of the other types of users. The tourist concentrates on breadth and not on depth. The tourist never becomes an expert in any single subject matter.

---

<sup>5</sup> The user types should not directly be related to an individual. An individual can have different roles for different tasks.

#### 4.4.2 Types of navigation<sup>6</sup>

The objective of this paragraph is to discuss the different kinds of navigation forms a user of a portal could require. These types of navigation forms are discussed from a users perspective.

- **Querying**  
By submitting a query existing out of key words or in the form of natural language an amount of relevant documents are shown to the user. After the first results have been presented the user can optimise the query to narrow or broaden the results.
- **Categorising**  
Documents are categorised on a common subject they discuss. A list of categories is shown to the user, by choosing a category the user can drill down.
- **Browsing**  
Documents are shown as icons and the relation between them is shown by a line between them. The user can find an entry point in this network of documents and browse through the document collection. An example implementation of this navigation form is [The Brain].
- **Localising**  
The user submits a query as in the query navigation. Yet, when localizing the user knows exactly which document should be found. Its location is just unknown.
- **Filtering**  
The user enters a query or selects a predefined query and the results will be shown. The filtering process can take place by either selecting a document that is desirable or by selection the documents that are not desirable and should be removed from the results. This process iterates until the user has found the documents of interest.
- **Subscribing**  
Based upon a profile of the user desirable information is send by e-mail or in any other form. (WAP, SMS, etc). In this type of navigation there can be a time gap between the user entering the query or building the profile and the presentation of the results.

---

<sup>6</sup> The types of navigation that will be discussed in this paragraph, are the results of two workshops held with several colleagues. Among them were experts in Business Intelligence, E-Business and a former IT-manager. The results are described in a working paper [Dillen]

#### 4.4.3 Users and navigation forms

The complex part for setting up a portal for unstructured sources is the choice of the navigation forms. Implementing all the navigational forms is costly and it will depend on the type of users whether they all will be used. This makes it interesting to combine the classification of users and the navigation types.

In the matrix below the user classification and the navigation forms are presented. For each user type is determined which navigation forms are interesting<sup>7</sup>.

	Farmer	Explorer	Miner	Tourist
Querying	X	X	X	X
Categorising			X	X
Browsing		X	X	X
Localising	X			
Filtering	X	X	X	X
Subscribing				X

**Remark:** The matrix has been completed based on the input the participants of the workshop have provided. Although, there were several experts present in that group, further research is required to determine whether this can be used as a generic model.

For this research project the completion of the matrix is not a main objective. The types of navigation forms and the application of the matrix are.

#### 4.5 Conclusion

Personalised information portals can be described by the technical components they encompass. In this thesis the focus is on the components concerned with the unstructured sources. They can consist of text mining techniques.

The requirements that have to be fulfilled in order to implement functionalities for the unstructured sources of a portal are defined by the navigation forms. In other words, to determine whether text mining techniques can fulfil the requirements of a personalised information portal they have to implement the navigation forms.

---

<sup>7</sup> In order to determine which type of navigation is interesting for a user type, the workshop participants have been asked to fill in the table. Based up on their responses the table has been created.

## 5. Text Mining and Personalised Information Portals

### 5.1 Introduction

In this chapter the question whether text mining can fulfil the functional requirements of a personalised information portal will be answered.

Text mining techniques apply to textual documents. This means they only work for unstructured sources. Portals mostly consist out of a mixture of structured and unstructured sources. Thereby, with the requirements of a personalised information portal are meant the requirement concurring the unstructured sources.

In the following paragraphs there will be focussed on the role Text Mining can fulfil for unstructured sources. This will be done by first describing several techniques that can be used to implement the navigation forms of a portal. Finally, the role of Text Mining for portals will be made explicit.

### 5.2 Navigation and Text Mining

#### 5.2.1 Querying

Querying is the most basic navigation form of all. The user can submit a query and the system will retrieve matching documents. To do so, several techniques can be used. They range from simple Boolean queries to natural language interfaces.

Using a Boolean query the user can determine which words should be in the document that will be retrieved. The document and the query must fully match in order to be retrieved. That is, the document is retrieved will contain the words that were part of the query.

In a natural language interface the user can enter the query in his own language. The system will convert the query to any form necessary to match the query on the document collection. This can either be a Boolean form, but also a vector or probabilistic form. If the query is converted in one of the former two models, the documents that will be retrieved do not have to match 100%. They can approximate the users query. An even more interesting advantage of these two models is that they take in consideration the semantic value of the words. That is, they try to determine what the "meaning" of a query word is in the document to be retrieved. (Think of the virus example of paragraph 3.3.2.1). Hence, the quality of the retrieved documents by these models will in general be better than retrieved by a Boolean model.

Another aspect is the amount of results that should be delivered. Is the user interested in getting several documents from which can be chosen or is the user interested in the best document? With the vector and probabilistic based systems it will be hard to determine the best document to retrieve. To do so, the user's query must be precise enough to select that document. In such a case the user would enter a query, the system should match the query on the document collection. If more documents are retrieved the system should find the discriminating properties among these documents and ask the user which properties are of the most relevance. Paragraph 3.4.2 shows that this type of interaction is used for Case based Reasoning system.

### 5.2.2 Categorize

In the category navigation the entire collection of documents is ordered in categories of related documents. This can be done using two different approaches:

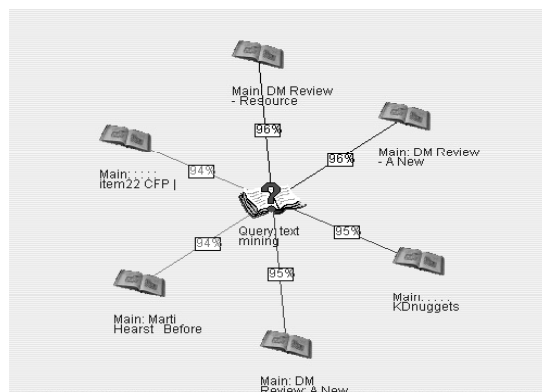
- *User's interest is leading (Top down)*  
In this approach the category structure is build by the user organization. That is, a taxonomy of the interest of the users is made as first. Based upon this taxonomy the document collection is ordered.  
The taxonomy in many cases is made manually. A more interesting approach would be to generate the taxonomy. In order to do this a profile of the user's interest should be created first<sup>8</sup>.  
  
In this approach the navigation is based on the user's interest. This type of categorization is of particular interest when the document collection is build up out of a volatile document collection. For instance, a document collection of daily news articles.
- *Documents are leading (Bottom up)*  
By determining the concepts of the documents that are part of the collection, the relation between the documents can determined. Some documents will have a stronger relation than others and they will form a cluster/category. The categories themselves will also have a relation with each other.  
In this approach the taxonomy is based on the document collection itself. The taxonomy represents the concepts of the documents that are part of the document collection. If a category is not in the taxonomy than there are no documents of that type in the collection.  
This approach is of particular interest for a "closed" document collection. For instance, to categorize the available documents in a library.

---

<sup>8</sup> In order to create the profile of a user it would be useful to log the queries a user executes.

### 5.2.3 Browsing

The user enters a query, which will result in documents that have a relation with that query. As shown in the picture displayed at the right the query is the centre of the “web” and the related documents are shown around the query. On the relation between the query and the documents a matching indication is given.



Autonomy Visualiser

That is, the chance that the document matches the query (when probabilistic models are used) or the distance between the query and documents (when vector models are used.)

The user can now read a document and determine whether it is satisfying. The document will be selected and become the centre of the web. Other documents that have a relation with the selected document will be shown and the process starts over.

The user is browsing through the document collection and is retrieving associated documents. This is particularly interesting for a researcher or a student who has to define a particular research subject.

### 5.2.4 Localising

Localising is special form of query navigation. The user knows up front which document should be retrieved, but does not know where to find it. Typical queries in this case are the title or author of an article. For this type of question a simple Boolean full text search would be sufficient to retrieve the document.

### 5.2.5 Filtering

Filtering is the process of finding an interesting document by including or excluding particular properties of a document. The user will get a list of properties to select from. For instance, the names of authors who wrote documents. The user can select a particular author name and the documents of that author will be retrieved. A new list of properties can be shown and the user can again select one, etc.

In order to show a list of properties some meta information of the documents is needed. This meta information can be added to the document during the creation of it<sup>9</sup>. But for a large existing document collection manually adding meta information will not be acceptable. In this situation automatic tagging –the process of adding meta information- is desirable.

### 5.2.6 Subscribing

With subscribing navigation is meant that the results of a query are send on a periodical or event basis<sup>10</sup> to the user for instance by e-mail. In the e-mail users will find a summary of the document and a reference to the document itself.

The differences between subscribe and query navigation is the time elapsing between submitting the query and presenting the results. Time is elapsing between the moment the query is created and the results are shown. This makes it desirable to combine the subscribe navigation with for instance the query. The user can create a query and give direct feedback on its results. If the results are not satisfying the query can be adjusted<sup>11</sup>. When the results of the query are desirable, a subscription can be made.

## 5.3 Conclusion

In order to implement the navigation forms for a portal a lot of different techniques are required. They range from standard information retrieval to Text Mining. As shown in chapter 3 it is hard to draw a line between these two techniques. The fact that a lot of techniques, which are used within information retrieval, are also used within Text Mining is the main reason for this difficulty.

It is clear that for the implementation of the localising navigation form, information retrieval techniques are sufficient. By using a Boolean based system the requirements can be fulfilled. The implementation of tagging functionalities, categorising documents and creating summaries Text Mining techniques are required.

From this perspective the main conclusion is that Text Mining techniques can fulfil the requirements of a personalised information portal.

---

<sup>9</sup> The Extensible Markup Language (XML) will have a large role for content filtering. XML –based documents will contain the meta information and when necessary the user can even be forced to provided the meta during the creation process.

<sup>10</sup> An event can be that new documents are added to the document collection that match the users query.

<sup>11</sup> There are several techniques that can be used to adjust the query. For instance, the user can expand or increase the query words. But also the system could make these adjustments. That is, if a vector or probabilistic based system is used. It could abstracted several index terms from the documents that the user desires.

## 6. Autonomy<sup>12</sup>

*“Probable impossibilities are to be preferred to improbable possibilities.”*

Aristotle

### 6.1 Introduction

In this chapter the functionality of the Autonomy product “Portal in the Box”<sup>13</sup> will be described. (See <http://www.autonomy.com>) This will be done with multiple purposes:

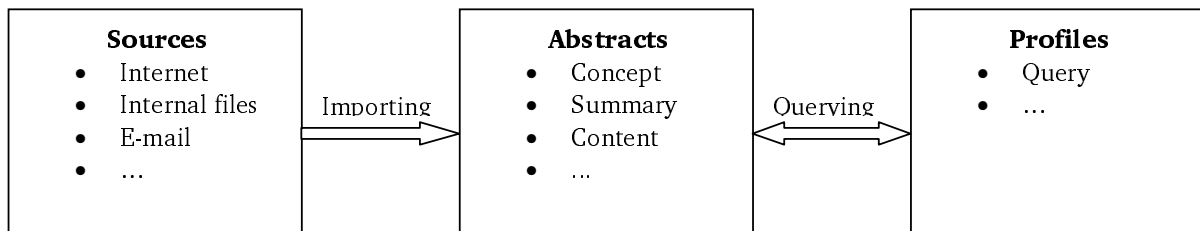
- to determine whether Autonomy is a Text Mining technology;
- to determine the activities necessary to implement and exploit Autonomy;
- finally to determine whether Autonomy can be used to set-up a personalized information portal.

The latter two objectives will be discussed in detail in chapter 8. This chapter will be the basis for this discussion.

In paragraph 6.2 an architectural overview of Autonomy will be given. In this paragraph three subject areas sources, abstracts and profiles will be distinguished. These will be discussed in respectively paragraph 6.3, 6.4 and 6.5

### 6.2 The architecture

Autonomy provides a portal tool in which information that is **gathered from specific sources** can be shown **in an instance** on the **user's personal interest**. In order to do so three main process steps are distinguished:



- *Sources:*  
The information shown in the portal can be gathered from several different sources such as the internet, intranet, internal file structures, specific databases (Lotus Notes, Oracle), etc. Documents from these sources are fetched and filtered so they can be imported into the Autonomy system.

---

<sup>12</sup> In this chapter the architecture of Autonomy will be discussed. Due to poor documentation of the product and the mysteriousness Autonomy keeps regarding the used algorithms some of the details have been gathered by studying the behaviour of the product in practice (Reverse engineering).

<sup>13</sup> When in this document is spoken of Autonomy the product Portal in the Box is referred to.

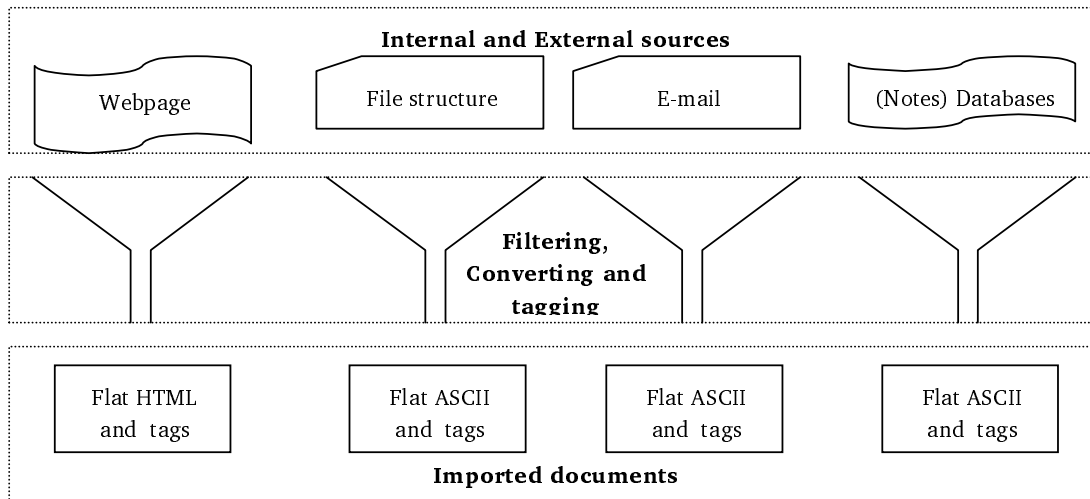
- *Abstracts:*  
 Importing the documents into the system means that a concept<sup>14</sup> of the documents meaning will be extracted. The concept together with a summary of the document, the document itself and some meta information can then be stored in the system.  
 Now the user is able to retrieve the documents, which are stored in the system by querying the document's concept.
- *Profiles:*  
 The user can query the system directly using natural language (a sentence or even by providing a complete document). A more interesting option is the use of an agent. This agent can query the system on behalf of the user and retrieve documents. By telling the agent which of the retrieved documents the user likes the agent can be trained to deliver better results. Finally, by combining the agents a profile of the user's needs can be created.

The figure above shows the three main subject areas and their relation. In the next paragraphs these will be described in depth.

### 6.3 Sources

As shown above the users are not directly querying the sources, which they are interested in. They query the concepts of the documents, from the sources, that are imported into the system. In order to import a document into the system it has to be fetched from its source. When the document is not an ASCII or HTML file it has to be converted to one of those formats. The document must meet the filtering requirements and finally meta information of the document can be tagged to it.

This process is shown in the figure below




---

<sup>14</sup> With the concept of the document is meant the subject of the document in such a form that it can be processed mathematically.

### 6.3.1 Fetching

Fetching is the activity of retrieving a document from its source. From Autonomy's perspective the following sources can be distinguished:

- Websites:  
These can either be Internet, intranet or extranet pages. By using a so-called spider<sup>15</sup> the fetch process retrieves documents from websites by following the hyperlinks on a specific website. When necessary the spider has the ability to login to a secured website.
- Databases:  
In order to retrieve for instance e-mail documents from a database, connections to ODBC, Lotus Notes, Oracle and other database systems have been build to fetch the documents they contain.
- File systems:  
Via the autoindexer documents from a specific file system can be fetched.

### 6.3.2 Converting

Fetching a source can deliver several different types of documents. For instance, on a website PowerPoint presentations, PDF's and Word documents can be shown. These files have to be converted to flat ASCII or HTML in order for the system to process them.

### 6.3.3 Filtering

Not all the documents fetched from a source, will be interesting enough to import into the system. By setting up filter requirements the interesting documents can be gathered from a source. The filtering requirements extent from the document type, date, size to specific requirements of the document's content.

### 6.3.4 Tagging

Tags can be added to documents that came through the filtering process. These tags can contain meta information about the document. This can be the source where the document originates from, the date on which it has been fetched or any other information. When the source is build-up out of XML documents the XML tags can be used.

Applying tags to a document enables the user to query the system based on these tags and on the other hand they can be used to provide the user with detailed information about the results the system will deliver (What is the documents source? From what date is it?).

---

<sup>15</sup> Spider also known as a crawler, ant, robot ("bot") and intelligent agent, a spider is a program that searches for information on the World Wide Web. It is used to locate new documents and new sites by following hypertext links from server to server and indexing information based on search criteria. [Techweb]

### 6.3.5 Importing

To import the documents into the system some final choices have to be made, the two most important choices are:

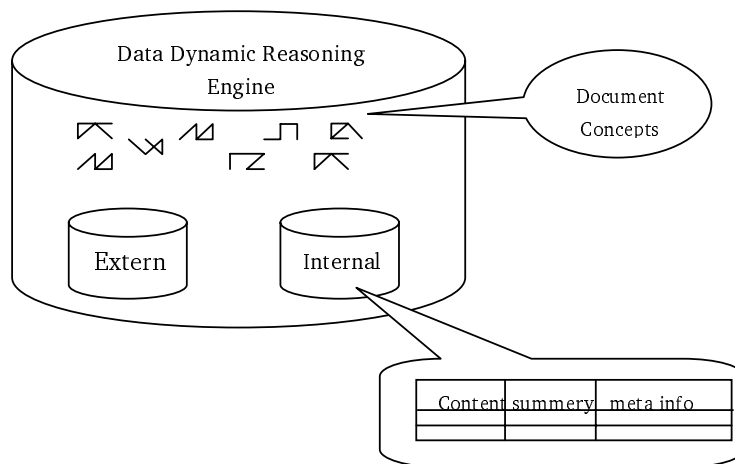
1. *Whether or not the content of the document should be stored?*  
 Storing the content of the document in the system makes it possible to show the document to the user without consulting its original source. This is in particular interesting for volatile sources. When the original document does not exist anymore it still can be presented. Of course by storing the content the system's need for disk space will be huge.
2. *In which database the document should be stored?*  
 In the system the documents will be stored in databases. Users can be granted access to these databases. In order to use the portal for several different users (with different access rights) it is important to set up the right databases.

By setting these last attributes the document is ready to be imported into the system. Importing in this sense means that the flat ASCII and HTML documents will be provided to the Data Dynamic Reasoning Engine (DRE).

### 6.4 Abstracts

The Data DRE is the core of the Autonomy system. In the DRE the imported documents will be stored. The DRE is also responsible for determining the concept of the document and the relation between the documents (see topic based clustering).

The figure below shows the structure of the DRE. As mentioned in the former paragraph, the documents are stored in a database within the DRE. Yet the documents concepts are stored separate from the databases. The advantage of this approach will become clear when describing the topic based clustering method.



#### **6.4.1 Concepts**

The concept of a document is a representation of the subject the document is describing. To determine the concept of a document, techniques based upon Shannon's information theory are used. In particular, the less frequently a word or phrase<sup>16</sup> occurs in a document the more information it conveys [Autonomy]. By using this method several words are extracted from the documents and are weighted. The combination of the words and their weights represent the concept of the document.

#### **6.4.2 Topic based Clustering**

Now that the concepts of the documents are determined in the form of words and their weights it is possible to use statistics to determine the relationships between the concepts. Algorithms of the probabilistic model, in particular Bayesian inference networks<sup>17</sup>, are used for that purpose. They calculate the probability that documents have a relation. Documents with a similar topic are clustered together. By calculating the probability that the user's query has a relation with a document, entry point within these clusters can be determined in an instance and the resulting documents can be delivered. The optimisation of the user's query is part of profile management, which will be discussed in the next paragraph.

#### **6.5 Profiles**

In order to retrieve documents from the Data DRE the user can send a query directly to it. The concept of the search query, which consists of words and their weight, will be determined. As a result the documents will be sent to the user by matching the concept of the search query with the concepts of the documents in the Data DRE.

Sending a search query consisting of words and their weights is not the user-friendliest interface for such a system. Even worse it is very hard for an average user to optimise the query in this way. Therefore agents were introduced.

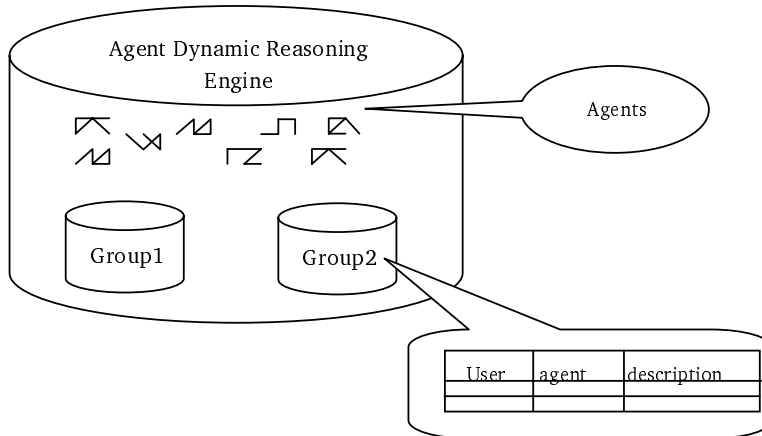
---

<sup>16</sup> From this point on when there is spoken of words in this documents both words and phrases are meant.

<sup>17</sup> Applying the Bayesian inference networks makes it possible to calculate these concepts very quickly. [Niedermayer]

### 6.5.1 Agents

Basically an agent is a representation of the users search query. In order to store the agent of the user an Agent DRE is set-up. This DRE has the same structure as the Data DRE. The concept of the search query is determined and stored in it. Within the Agent DRE several databases can be created where agents of different users (groups) can be stored. Each database in the system forms its own community. (See paragraph "Profiling")



The first time a user uses an agent it is still blank and does not contain a search query. The users can either enter some words, which they are interested in or provide a document from which the agent can create a concept. The agent will then send the concept to the Data DRE and the results will be shown to the user. By telling the agent which of the resulted documents fulfil the user's need the most, the user is able to train the agent.

Training the agent means that the agent will adjust the concept of the search query by adding the concept of the documents that fulfil the user's needs. Hence, from the users perspective training the agent will learn it to bring more relevant results.

### 6.5.2 Profiling

One of the advantages of using an agent is the ability to retrain it. Another advantage is the fact that it stores the user's query in the system. By combining the agents of a particular user a profile of that user is made. By monitoring the documents the user is reading, the profile is refined.

In general this means that the agents form the base of a user's profile. In order to adjust the wide spread of results the agents can deliver, the browsing behaviour of the user is logged and the profile is refined. Hence, a profile of the user is constructed by using the user's interest.

Due to the fact that the agents of all the users are stored they can also be compared with each other. When there is a large resemblance between the agents of two users they can be notified and use the agents of each other. The same counts for the user's profiles. Two users who have the same interest can be brought together by informing them about each other.

## 6.6 Conclusion

Now that it is clear what the main functionalities of Autonomy are and how they are implemented the question remains: Is Autonomy a text mining application?

As discussed in paragraph 3.6 real text mining is the process of gaining new information or knowledge from a textual document or a set of textual documents. Autonomy does not gain new information or knowledge.

Does this mean that Autonomy is an information retrieval system? An information retrieval system has the objective to retrieve information/documents that match with the user's query. If the agent of a user is defined as a query then Autonomy does retrieve documents that the user desires. But that is not all. There are several aspects that makes Autonomy more than just an information retrieval system:

- based on the feedback a user provides on the results Autonomy delivers the system "learns" and adjusts the results;
- a profile is created of the information desired by the user. Based on this profile Autonomy suggests new documents that might be of the user's interest. In this sense the system is pro-active;
- in order to show the results of a user's agent a –semantic- summary is made of the document's concept. Based on this summary the user can decide whether the document is to be retrieved or not;

This shows that Autonomy is typically a product that is intermediary of information retrieval and real text mining. It fills the gap between these two techniques.

## 7. MYCIBIT.COM

*"Make everything as simple as possible, but no simpler. "*

Albert Einstein

### 7.1 Introduction

In this chapter MYCIBIT.COM will be discussed. The objective of this chapter is to determine the requirements that MYCIBIT.COM has for Autonomy. This will be done by introducing MYCIBIT.COM, starting with its history.

The requirements that Autonomy has to fulfil will be discussed in paragraph 7.3. Describing some situations, which MYCIBIT.COM users will encounter, will do this. Although, these situations are just a selection of the total set of facilities MYCIBIT.COM will comprehend. They are representative for MYCIBIT.COM regarding the unstructured sources.

In chapter 8 the requirements of MYCIBIT.COM will be compared with the specifications Autonomy has to offer.

### 7.2 History

At the end of 1997 CIBIT started to build Virtual-CIBIT with the objective to support the cooperation of their students independent of time and location. Facilities as discussion groups, publication boards and a small library were offered to the students. Also lecturers could participate in the environment and support the students.

This was the first step in the lifecycle of Virtual-CIBIT. At this moment, several years later, the environment has evolved into a full-blown e-learning space. Facilities like video presentations, virtual project rooms and an extensive library are provided. Combined with extensive knowledge and experience of the organization of such facilities.

All the facilities Virtual-CIBIT offers are provided in the same way. All the users of a particular course get the same interface to access their facilities. As soon as the user has logged on, a lot of choices have to be made to get the desired facility. Offering a form of personalization would increase the efficiency of the environment.

On the other hand Virtual-CIBIT can be very useful during consultancy activities that CIBIT offers. During these activities Virtual-CIBIT can be used as a communication and collaboration platform and also as a learning environment for the customer.

The next step in the lifecycle of Virtual-CIBIT will be the introduction of personalization for its users. Besides that, the introduction of one main area with an overview of the facilities and their latest content will be offered. Virtual-CIBIT will become a personalized information portal for its users. It will become MYCIBIT.COM<sup>18</sup>.

In the next paragraphs several functionalities of MYCIBIT.COM will be discussed from the perspective of its further users. The objective of the discussion is to gather the requirements, which have to be implemented. In the next chapter these requirements will be held against the specifications of Autonomy, in order to see which requirements Autonomy can fulfil.

### 7.3 MYCIBIT.COM users

Just a selection of all the functionalities MYCIBIT.COM will provide are discussed in this paragraph. They represent a broad range of different functionalities that might be expected from MYCIBIT.COM in relation with unstructured sources.

The functionalities will be discussed from the perspective of the user. The following MYCIBIT.COM users are distinguished for this thesis: students; former customers/alumni; professionals and support staff.

#### 7.3.1 Students

People who attend at a course provided by CIBIT. This can range from a one-day course to a two-year Master of Science programme.

##### **Selecting a course**

In order for a person to select a course to take, that person will be advised by the on-line course advisor. The course advisor will ask the person several questions, which will enable the application to advise the course(s) that can fill in the learning goals. The advisor has to be able to advise one specific course and should not advise a course if the learning goals of the person cannot be fulfilled.

##### **Preparing for a course**

When someone decides which course to take, the person can select one of the subject areas related to the course to prepare. These subject areas will provide the student with several documents, which are pre-selected by the lecturer of the course.

##### **Writing a paper (thesis)**

As part of the course the student has to write a paper on a subject related to the course. To define the subject matter the student has to determine the context of the subject area. To do so, the student needs to find documents related to this subject area. There are two options to fulfil this task:

1. The student can go to a library and select pre-set agents. These will help him to find the related documents. The student is even capable to optimise the delivered results, by optimising the training of the agent.

---

<sup>18</sup> In paragraph 4.2 is shown from different perspectives what a personalized information portal is. In combination with the components they contain, discussed in paragraph 4.3, it is concluded that MYCIBIT.COM applies to these specifications.

2. The student browses through the results. A document, related with the subject area, is shown. The user can now select a related document, which has relations on its own, etc. By walking through these related documents the user will find relations between the subject areas.

### **7.3.2 Former customers/alumni**

People who have taken a course from CIBIT in the past.

#### **Stay informed after the course**

The student has finished the course successfully. It is up to the student to stay up to date with the subject area. As part of a service the student can receive e-mails periodically, containing the most interesting documents related to the subject.

### **7.3.3 Professionals**

Those employees of CIBIT that advise customers and lecture during courses.

#### **Gathering course material**

In order to provide the students with background information on the course subject, the professionals select several documents that are of particular interest. To do so, the professionals can create an agent or browse through several categories with documents related to the subject area.

### **7.3.4 Support staff**

Employees who support the professionals in order to fulfil their tasks.

#### **Created course file**

The professionals have selected several articles that should be supplied to the students of the course. The support staff creates the files for each student, which contain these articles. In order to find the articles, the professionals have suggested, the support employee can use a locate service. The name of the article can be entered and the article will be retrieved.

## 8. Autonomy and MYCIBIT.com

*“Problems cannot be solved at the same level of awareness that created them.”*

Albert Einstein

### 8.1 Introduction

In this chapter will be determined whether Autonomy can fulfil the requirements of MYCIBIT.COM. If so, it will be shown how and what this means for the exploitation. If Autonomy cannot fulfil the requirements, it will be explained why not and when possible alternatives will be given.

At first the situations MYCIBIT.COM will encounter as described in paragraph 7.3 will be reflected on Autonomy. Finally, in paragraph 8.3 some exploitation aspects of Autonomy will be discussed.

### 8.2 Autonomy for MYCIBIT.COM

#### 8.2.1 Student selecting a course

In order for a student to select a course without the on-line course advisor the student would have to read the course brochures. These brochures are texts that at least describe the course content, the knowledge needed to start the course and the objectives of the course.

In theory with real text mining the brochures would be enough to start a “question and answer”-dialog in order to select the right course (especially, when combined with a CBR system). But as mentioned above Autonomy is not a real text mining application. If Autonomy would be applied to search for the best course two problems would arise:

1. *The user query*  
The query the user creates should contain all the needed information in order to match with the right brochure. In order to guarantee that all this information is in the query a special kind of interface should be created.
2. *Quality of the results*  
In this case the best course should be retrieved. Autonomy will calculate the probability that the document (course) that will be retrieved answers the question. Moreover, there is not enough control to determine whether this is the right course for the student.

In this case a rule based or a case based system would deliver better results<sup>19</sup>. The quality of answers these systems deliver can be controlled better. In this particular case the choice for a rule based system would be preferable. The amount of courses is rather small and the amount of rules necessary to select a particular course are relatively small.

---

<sup>19</sup> At the moment the real text mining tools are still in an experimental phase, as far as the author knows.

### **8.2.2 Student preparing for a course**

The student will select one of the subject areas related to the course. This will result in the retrieval of several documents out of the sources indexed by Autonomy. The subject area the student selects, is the representation of an agent, which has been trained by for instance the lecturer of the course. This agent will retrieve the desired documents. In this case the agent is nothing more than a pre-defined query.

### **8.2.3 Student writing a paper/thesis**

Two approaches will be discussed to find related documents. Both can be fulfilled with Autonomy:

1. *The agent library*

The user selects an agent from a library, which has been trained by an expert on that particular subject. Assuming that the expert has trained the agent well, the results retrieved will be strongly related to the subject matter. The agent can be cloned allowing the user to train this agent for its particular desires. By training the agent, new related documents and even subject matter may be retrieved.

2. *Associative search*

In the associative search the user browses through the documents. This means that the user will enter a query that will result in the retrieval of a particular document. This document will have relations with other documents. By selecting a related document it will be retrieved. This document will also have relations with other documents, which can be selected. This process can be iterated until the user has found all the necessary related subject matter.

### **8.2.4 Former customer is kept informed after a course**

In order to be kept informed the customer can subscribe himself to a mailing service. This means that the customer selects a subject matter to be kept informed on. This subject matter is represented by an agent, which for instance could be trained by a CIBIT expert. The results of the agent could then be sent via e-mail to the customer. This could be done on a periodical basis.

### **8.2.5 Professional gathering course material**

The professional can use an agent to find the articles that are of his interest. This would be a standard approach. The professional will define a query (agent) to start. The agent will retrieve documents and the user can train the agent to optimise the results.

The second form would be to use a category to order the document collection. With Autonomy categories can be created based upon the user's interest (see paragraph 5.2.2 Categorize). Categories can be created based upon the user's agents. Autonomy stores the user's agents in a central place (the community DRE) thereby it is possible to use the pre-trained agents to create the categories.

### **8.2.6 Support staff create course file**

To find the articles the professionals have suggested the support staff need a localise tool. Autonomy is not a tool to use for this kind of question (see paragraph 5.2.4) .

### 8.3 Exploitation of Autonomy

During the implementation and introduction of Autonomy facilities within CIBIT, experiences have been gained. The role of the knowledge steward in relation with Autonomy is the most interesting. One of the roles of the knowledge steward within CIBIT is to bring together knowledge supplier and demander.

In order to exploit Autonomy successfully a knowledge steward is needed to optimise the use of the product. For the exploitation three main process steps must be distinguished: profile management, abstract management and source management. Within these process steps there are several issues that need special attention.

#### 8.3.1 Profile management

##### *Agent Management*

Users of the system can create agents, which will retrieve desired documents. Agents of different users can be cloned. Cloning the agent means that a copy of the original is made. The new –copied- agent does not have any relation with the parent agent any more. This means that training the parent agent will not influence the results of the cloned agent. Hence, it is not possible to create an agent hierarchy.

When implementing Autonomy for a user group of 10.000 users who make 10 agents each, the lack of an agent hierarchy might introduce some problems to manage the agents. Especially when the concept of agent academy is to be implemented.

##### *Agent Academy*

Within Autonomy users will create several agents. Based upon these agents Autonomy is capable of extracting an overall profile of the user's information desire.

The users can save a lot of time by cloning an agent of another user and optimise the training for their own need. Still, the agent has to be found before the user can clone it. In a large organisation where more than 10.000 agents have been built it will be hard to find the best. Introducing an Agent Academy will be desirable.

In the academy agents will be categorised, allowing the user to select one. The easiest way to set-up the academy is to select the most similar topics that the users are interested in. These can be determined by Autonomy<sup>20</sup>. The role of the knowledge steward in this process would be to determine whether the quality of the agent is satisfying. Further research is required to determine how the quality of the agents can be assessed.

---

<sup>20</sup> As shown in paragraph 6.5 the agents of users as stored in a DRE. Within this DRE the agents are represented as concepts. The probability of related concepts can be calculated. By doing so they can be clustered and topics can be determined.

### 8.3.2 Abstract management

The user's agent queries the abstract index (DRE) in order to retrieve documents. Within the DRE the document's concepts are stored in separate databases. A database is filled with documents from a particular source.

By selecting a particular database the query should use, the user can put focus on the documents that will be retrieved. For instance, if a database only contains documents from academic sources the user will not retrieve commercial product white papers<sup>21</sup>.

Dividing the entire document collection in databases gives the user some control over the document collection to retrieve from. The division of the document collection into databases will be different for each organisation. It will be the role of the knowledge steward to determine which databases should be distinguished.

### 8.3.3 Source management

Source management is the process of importing documents from their source into the Autonomy system. There are many types of sources that can be used to import documents from. Within these sources it is interesting to make a distinction between self controlled sources and external sources.

Self controlled sources are those sources controlled by the user organisation itself. If any change to these sources is made this can be communicated. With external sources this is more complicated. When an external source changes then the user organisation will have to find this out themselves.

The change of the source, that is the structure of the source, will reflect on the importing functionalities of Autonomy. Depending on the change the import functionalities will have to be adjusted.

Another aspect that is of importance regarding external sources is the links they have with other sources. To import the external –web- sources spiders are used. They will spider through the source and retrieve all the relevant documents. To determine which pages to spider through the spider follows the hyperlinks on the pages. If the source is a link portal<sup>22</sup> control over the documents returned by the spider will be very low.

It is up to the knowledge steward to determine which sources can be imported into the Autonomy system<sup>23</sup>.

---

<sup>21</sup> Taken in consideration that the academic sources do not provide commercial white papers.

<sup>22</sup> With a link portal is meant a portal that collects links to sites of a particular subject area. The link portal will be address by its users to find a site with the information they desire. They go to the link portal and directly leave it to visit the site they searched for.

<sup>23</sup> The knowledge steward within CIBIT is at this moment creating a checklist with points to consider when importing a source.

## 9. Conclusions and further research

*"I don't have any solution, but I certainly admire the problem."*

Ashleigh Brilliant

### 9.1 Text mining

Real Text Mining is the process of gaining new information or knowledge from a textual document or a set of textual documents. Text mining techniques based on information retrieval models in combination with linguistics can be used to determine the concept of a document. In particular the vector model and probabilistic model in combination with information theory can be used to determine the semantic value of words in a document. Based on case based reasoning, a data mining related technique, particular documents can be provided to the user.

### 9.2 Personalised information portals

There are many definitions of portals. However, portals can be segmented based on the components they consist of and the different types of users they sever. The components that handle the unstructured sources are of the highest relevance for text mining.

Users of portals can be classified into four types: *Farmers* know exactly what they want. *Miners* determine whether a particular hypothesis can be supported. *Explorers* use a heuristically approach and *Tourist* are rather impulsive.

These users can be offered several forms of navigation. These forms answer different kind of questions that users have. Distinguished navigation types are: querying, browsing, categorising, localising, filtering and subscribing.

The user types and the navigation forms can be placed in a matrix. This matrix can be used to determine in which navigation form should be invested when implementing a portal for a certain user group.

### 9.3 Text mining for personalised information portals

Except for the localising navigation form, text mining can be used for all the requirements of the unstructured sources of a personalised information portal. For the localising navigation a simple Boolean information retrieval model will be sufficient.

Based on text mining techniques the users information desire can be personalised and profiles can be created. They can also create summaries of documents. Add tags to documents, chapters, paragraphs or any other particular part. By using proximities they can provide associative (i.e. browsing) navigation forms.

#### **9.4 Autonomy and MYCIBIT.COM**

Autonomy does not gain new information or knowledge from unstructured documents. Hence, Autonomy is not a real text mining tool. But Autonomy is more than just an information retrieval system. It can retrieve documents that fulfil the information desire of the user. Furthermore, it can learn from the users feedback, it can create a profile of the user's information desire and it is capable of creating –semantic- summaries. This makes Autonomy typical a product that is intermediary of information retrieval and text mining.

MYCIBIT.COM is a personalised environment, which contains both structured and unstructured sources. For the unstructured sources at least the querying and browsing navigation forms will be implemented. Thereby, MYCIBIT.COM is an example of a personalised information portal.

Autonomy can fulfil all the requirements regarding the unstructured sources of MYCIBIT.COM. Yet, to exploit Autonomy successfully a knowledge steward is needed to optimise the use of the product for MYCIBIT.COM. The knowledge steward will fulfil an intermediary role between the users of the system and the technical operators. Thereby, determine how the facilities should be offered to the users of MYCIBIT.COM.

#### **9.5 Further research**

In this thesis the user classification and navigation matrix for unstructured sources of portals is introduced. This matrix can be used to determine which type of navigation will added the most value for a portal in a particular implementation. In order to generalise the completion of the matrix for all types of portals further research is required.

To optimise the re-use of agents within Autonomy it is import to determine the quality of an agent. Further research is required to determine how the quality of the agents can be measured.

## References

[Aamodt]

“Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches”, Agnar Aamodt and Enric Plaza, <http://citeseer.nj.nec.com/252387.html>

[Autonomy]

“Technology White Paper, Paper by Autonomy  
<http://www.autonomy.com/tech/whitepaper.pdf>

[Baeza-Yates]

“Modern Information Retrieval”, Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison Wesley, ISBN 0-201-9829, page 21, 24-64

[Berson]

“Data warehousing, datamining en OLAP”, Alex Berson and Stephen J. Smith, Academic Service, ISBN 90 395 1014 8, page 333

[Bush]

“As we may think”, Dr. Vannevar Bush, July 1945,  
<http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>

[Butler]

“Souped-up search engines”, Declan Butler, article of Nature 405 page 112 to 115, 11 May 2000  
[http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v405/n6783/full/405112a0\\_fs.html](http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v405/n6783/full/405112a0_fs.html)

[Davydov]

“EIP: The second wave”, Mark M. Davydov, Ph.D, Intelligent Enterprise Magazine, March 2000  
<http://www.intelligententerprise.com/000301/supplychain.shtml>

[Deerwester]

“Indexing by Latent Semantic Analysis”, Scott Deerwester, Susan T. Dumais, Richard Harshman  
<http://lsi.research.telcordia.com/lsi/LSIpapers.html>

[Dillen]

“Zoeken en navigeren in ongestructureerde gegevens”, Edwin van Dillen, Oktober 2000, working paper, Available on request.

[Dung]

‘A review of the evolution of information systems to retrieve archived information’, Dinh van Dung <http://ise.ee.uts.edu.au/ise/homepages/hmpgs96a/vdinh/subject/hypermd/assgn1.htm>

[DWR]

“Text Mining, Not Data Mining”, article of Data Warehouse Report, March 16 1999  
<http://datawarehouse.dci.com/Articles/990316mining.htm>

[Foltz]

“Using Latent Semantic Indexing for Information Filtering” Peter W.Foltz  
<http://www-psych.nmsu.edu/~pfoltz/cois/filtering-cois.html>

[Hearst-1]

‘Untangling Text Data Mining’, Marti Hearst, paper  
<http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html>

[Hearst-2]

Mining in Textual Mountains, Marti Hearst, Article of Mappa Mundi,  
<http://mappa.mundi.net/trip-m/hearst/>

[Hobbs]

“Computers and language”, Jerry R. Hobbs, <http://www.lsadc.org/web2/flfdr.htm>

[ID]

“A Characterization of Data Mining Technologies and Processes”, Information Discovery  
<http://www.datamine.aa.psiweb.com/dm-tech.htm>

[Inmon]

“The DSS Community: Understanding different user’s reporting and analytical needs”, Bill W. H. Inmon, <http://www.billinmon.com/library/whiteprs/dss.htm>

[Inreca]

“Detailed description of CBR”, <http://www.inreca.org/data/cbr/what/detials.html>

[Jansen]

“Using an intelligent agent to enhance search engine performance”, James Jansen, Firstmonday,  
[http://www.firstmonday.dk/issues/issue2\\_3/jansen/](http://www.firstmonday.dk/issues/issue2_3/jansen/)

[Korfhage]

“Information storage and retrieval” Robert R. Korfhage, Wiley Computer Publishing,  
ISBN 0-471-14338-3, page 79-97

[Mena]

“Data Mining FAQ’s”, Jesus Mena, DM Review January 1998,  
[http://www.dmreview.com/editorial/dmreview/print\\_Faction.cfm](http://www.dmreview.com/editorial/dmreview/print_Faction.cfm)

[Niedermayer]

“An introduction to Bayesian Networks and their contemporary Applications” Daryle Niedermayer, Paper of December 1 1998,  
[http://www.gpfn.sk.ca/~daryle/papers/bayesian\\_networks/bayes.html](http://www.gpfn.sk.ca/~daryle/papers/bayesian_networks/bayes.html)

[Nunberg]

“An overview of linguistics”, George Nunberg, <http://www.lsadc.org/web2/flfdr.htm>

[Ruber]

“Portals on a Mission”, Peter Ruber, Article form KM Magazine,  
<http://www.kmmag.com/km200004/feature1.htm#topofpage>

[Semio]

“Elevate your state of find: Strategies for building an effective corporate portal”, White paper of Semio corporation, March 2000, <http://www.semio.com>

[Spitzer]

“Vertical Horizon: Surveying the landscape of online industry”, Tom Spitzer, February 2000  
<http://www.webtechniques.com/archives/2000/02/spitzer>

[Techweb]

Online encyclopaedia <http://www.techweb.com/encyclopedia>

[Thearling]

“An Introduction to Data Mining and Advanced DSS Technology” Kurt Thearling, Ph.D,  
<http://www3.shore.net/~kht/dmintro/dmintro.htm>

[The Brain]

<http://www.thebrain.com>

[Turban]

‘Decision Support and Expert Systems’, Prentice Hall, ISBN 0-02-421701-8 Fourth Edition, page 454

[Whatis]

<http://www.whatis.com>

[Weiss]

“Glossary for Information Retrieval”, Scott Weiss <http://www.cs.jhu.edu/~weiss/glossary.html>

## Appendix A: Project Information

*Author:* Edwin van Dillen  
Marsstraat 2A  
4105 JL Gulemborg  
The Netherlands

*E-mail:* [edwin@van-dillen.com](mailto:edwin@van-dillen.com)

*Institute:* Kenniscentrum CIBIT  
Arthur van Schendelstraat 570  
Postbus 19210  
3501 DE Utrecht  
Tel: 030-2308900

*Mentor:* Timo Kouwenhoven  
*E-mail:* [tkouwenhoven@cibit.nl](mailto:tkouwenhoven@cibit.nl)

*Company:* Kenniscentrum CIBIT  
Arthur van Schendelstraat 570  
Postbus 19210  
3501 DE Utrecht  
Tel: 030-2308900

*Supervisor:* Cor Baars  
*E-mail:* [cbaars@cibit.nl](mailto:cbaars@cibit.nl)

## Kenniscentrum CIBIT

Kenniscentrum CIBIT is an independent institute. It is a Dutch centre of excellence for Business and ICT, Knowledge Management and Innovative System Development. As such, it provides a bridge between advanced research and practical use of the results of this research by industry.

CIBIT organises three types of MSc courses: Co-operative Computing, Knowledge Management Technology, and ICT Management.

The MSc programme Co-operative Computing is carried out under the full license of Middlesex University, London. Middlesex University, London, assures the quality of the course.

## Appendix B: Questionnaire

### **The first steps...**

In the following paragraphs a list of questions is presented. The questions are abstracted for the subjects discussed in this thesis and the experience gained during the project. Answering these questions is the first step in understanding the functionalities that are required to provide a portal user with the desired information.

This list is provided as a starting point to help organizations during their journey to find the best technology open their unstructured sources. I do release that in its current form it can only be used by a person whom understands both the users problems and the technology (As discussed in this thesis!). In order to be used in general I am working on a paper describing the problem from a user perspective.

#### **User**

- What kind of navigation should be provided to the user?
- On which frequency does the user launches a search?
- What type of question does the user launches in order to find the answer? (Precision of the question)
- Is the use of a Thesaurus required?
- What is the added value of the results the search will deliver?
- Can a categorization be made of the question the user launches?
- Can >60% of the answers be found by launching the 5 main questions?
- What is the experience level of the users using the system?
- How precise should the results match the search?
  1. Should it be one specific document?
  2. Should it be several documents nearby the search?

#### **Index**

- Is there any control of the content creation process?
  - Can meta-data be added to the documents during the creation process?
- Is the data, which is searched, volatile? (It is only interesting of a short moment, think of news)
- Is the data, which is searched, multi-language?

### **Sources**

- What kind of sources should be searched? (Web, file-system, etc)
  - Are they internal → Can the content creation process be altered?
  - Are they external → Are the sources secured or can they be accessed freely?
- What is the size of the sources, which are searched?
- Is there any evidence that the documents searched are in the sources?
- How often does the sources change? (Both the content and the structure of the source)
- What type of documents does the source contain? (HTML, XML, Microsoft Word, PDF, etc)

### **Technology**

- What is the available budget to buy a system?
- How many hours are available to spend on maintenance of the –search-system?
- What is the IT-knowledge level of the employees whom maintain the system?
- Is there an internal IT-development department available?