Spiros Sirmakessis (Ed.)

Text Mining and its Applications

Springer-Verlag Berlin Heidelberg GmbH

Studies in Fuzziness and Soft Computing, Volume 138

Editor-in-chief
Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series can be found on our homepage: springeronline.com

Vol. 119. Karmeshu (Ed.)
Entropy Measures, Maximum Entropy Principle and Emerging Applications, 2003
ISBN 3-540-00242-1

Vol. 120. H.M. Cartwright, L.M. Sztandera (Eds.)
Soft Computing Approaches in Chemistry, 2003
ISBN 3-540-00245-6

Vol. 121. J. Lee (Ed.)
Software Engineering with Computational Intelligence, 2003
ISBN 3-540-00472-6

Vol. 122. M. Nachtegael, D. Van der Weken, D. Van de Ville and E.E. Kerre (Eds.)
Fuzzy Filters for Image Processing, 2003
ISBN 3-540-00465-3

Vol. 123. V. Torra (Ed.)
Information Fusion in Data Mining, 2003
ISBN 3-540-00676-1

Vol. 124. X. Yu, J. Kacprzyk (Eds.)
Applied Decision Support with Soft Computing, 2003
ISBN 3-540-02491-3

Vol. 125. M. Inuiguchi, S. Hirano and S. Tsumoto (Eds.)
Rough Set Theory and Granular Computing, 2003
ISBN 3-540-00574-9

Vol. 126. J.-L. Verdegay (Ed.)
Fuzzy Sets Based Heuristics for Optimization, 2003
ISBN 3-540-00551-X

Vol 127. L. Reznik, V. Kreinovich (Eds.)
Soft Computing in Measurement and Information Acquisition, 2003
ISBN 3-540-00246-4

Vol 128. J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.)
Interpretability Issues in Fuzzy Modeling, 2003
ISBN 3-540-02932-X

Vol 129. J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.)
Accuracy Improvements in Linguistic Fuzzy Modeling, 2003
ISBN 3-540-02933-8

Vol 130. P.S. Nair
Uncertainty in Multi-Source Databases, 2003
ISBN 3-540-03242-8

Vol 131. J.N. Mordeson, D.S. Malik, N. Kuroki
Fuzzy Semigroups, 2003
ISBN 3-540-03243-6

Vol 132. Y. Xu, D. Ruan, K. Qin, J. Liu
Lattice-Valued Logic, 2003
ISBN 3-540-40175-X

Vol. 133. Z.-Q. Liu, J. Cai, R. Buse
Handwriting Recognition, 2003
ISBN 3-540-40177-6

Vol 134. V.A. Niskanen
Soft Computing Methods in Human Sciences, 2004
ISBN 3-540-00466-1

Vol. 135. J.J. Buckley
Fuzzy Probabilities and Fuzzy Sets for Web Planning, 2004
ISBN 3-540-00473-4

Vol. 136. L. Wang (Ed.)
Soft Computing in Communications, 2004
ISBN 3-540-40575-5

Vol. 137. V. Loia, M. Nikravesh, L.A. Zadeh (Eds.)
Fuzzy Logic and the Internet, 2004
ISBN 3-540-20180-7

Spiros Sirmakessis (Ed.)

Text Mining and its Applications

Results of the NEMIS Launch Conference

Springer

Dr. Spiros Sirmakessis
Assistant Professor
Research Academic
Computer Technology Institute
61 Riga Feraiou Str
26221 Patras
Greece
E-mail: syrma@cti.gr

ISSN 1434-9922

Library of Congress Cataloging-in-Publication-Data
International Workshop on Text Mining and its Application (1st : 2003)
Text mining and its applications : results of the NEMIS Launch Conference /
Spiros Sirmakessis (ed.).
p. cm. -- (Studies in fuzziness and soft computing ; v. 138)
Includes bibliographical references.
ISBN 978-3-642-05780-9 ISBN 978-3-540-45219-5 (eBook)
DOI 10.1007/978-3-540-45219-5

1. Data mining--Congresses. I. Sirmakessis, Spiros, 1969- II. Title. III. Series.
QA76.9.D343I59 2004
006.3'12--dc22

Originally published by Springer-Verlag Berlin Heidelberg New York in 2004
Softcover reprint of the hardcover 1st edition 2004

Typesetting: Camera-ready by author
Cover design: E. Kirchner, Springer-Verlag, Heidelberg
Printed on acid free paper 62/3020/M - 5 4 3 2 1 0

Foreword

The world of text mining is simultaneously a minefield and a gold mine. It is an exciting application field and an area of scientific research that is currently under rapid development. It uses techniques from well-established scientific fields (e.g. data mining, machine learning, information retrieval, natural language processing, case-based reasoning, statistics and knowledge management) in an effort to help people gain insight, understand and interpret large quantities of (usually) semi-structured and unstructured data.

Despite the advances made during the last few years, many issues remain unresolved. Proper co-ordination activities, dissemination of current trends and standardisation of the procedures have been identified, as key needs. There are many questions still unanswered, especially to the potential users; what is the scope of Text Mining, who uses it and for what purpose, what constitutes the leading trends in the field of Text Mining -especially in relation to IT- and whether there still remain areas to be covered.

The NEMIS project (http://nemis.cti.gr), funded by the IST framework, was set out to create a network of excellence (NoE), and bring together experts in the field of Text Mining to explore the grey areas relating to the status, trends and possible future developments in the technology, practices and uses of Text Mining. The main objectives of the project are to:

- Develop closer relations between European organizations for scientific and technological co-operation;
- Establish a homogenous system of scientific and technical reference;
- Present the current scientific research in the area and define open problems and new challenges;
- Create and maintain advanced communication facilities allowing optimum cooperation among interested actors;
- Bring together experts, researchers, academics and end-users who face common problems and share common needs in the field so as to create a critical mass in this area;
- Ensure that the needs of the users reach the scientific community and vice-versa (that research results and products reach the market and the potential users);
- Facilitate the transfer of knowledge including identified best practices, case studies, information sources, state-of-the-art reports, commercial tools, applications, etc;
- Identify future potential research areas and demonstrating how newly rising needs can be met;
- Examine in detail the relations between TM and official statistics and investigate the feasibility of applying TM to statistical processes;
- Support training opportunities in the field by organizing workshops;
- Provide a broad dissemination of research results in the area of TM;

For the efficient operation of the network, five preliminary topics have been identified, which constitute different scientific sections by themselves. Each topic forms a Working Group (WG):

- WG1 - Document processing and visualization techniques
- WG2 - Web mining
- WG3- Text mining applications, knowledge management and multilingual aspects
- WG4 - Market survey and comparative analysis of TM tools
- WG5 - User aspects and relations to Official Statistics

The *1st International Workshop on Text Mining and its Applications* was the first NEMIS public event and addressed not only network members but also other individuals interested in Text Mining. The event gave the opportunity to meet each other and provide detailed information for the network, its scope and expected results. Topics of interest included, but were not limited to:

Document processing & visualization techniques

- Document Representation & Storage
- Metadata Production
- Document Classification/Clustering
- Topic Detection Information/Entity/Relation Extraction
- Content Analysis
- Visualization Techniques

Web mining

- Web Content
- Structure & Usage Mining
- User Behaviour Modeling
- Machine Learning applied on the web
- Personalized Views
- Web Searching
- Usability Metrics and Assessment
- Business Intelligence in eCommerce
- Security & Privacy
- Semantic Web Mining
- RDF
- Ontologies

TM & knowledge management: Theory & applications

- Customer Relationship Management
- Human Resources
- Technology Watch
- Patent Analysis
- Lexicographic Analysis
- Linguistic Preprocessing

- Statistical Analysis of Textual Data
- Comparative Analysis of TM tools

User aspects & relations to Official Statistics

- Structures & Applications for Searching and Organising Metadata
- Catalogued Information vs. Free-Text Searching
- Evaluation of Existing Search-Functions for Statistics
- Discovery of Updates in Statistical Databases and Publishing Systems
- Tools and Applications for Tracing and Enumerating Official Statistics in Electronic mass media

The conference managed to maintain a balance between theoretical issues and descriptions of case studies and demonstrated the large margins of synergy between theory and practice in the field of Text Mining.

I would like to thank the members of the scientific and technical committee listed below, for their contribution to the success of the conference.

Scientific Committee
Driss AFZA, Training for European Statisticians Institute
Sergio BOLASCO, University of Roma 1
Albert PRAT, Polytechnic University of Catalonia
Martin RAJMAN, Ecole Polytechnique de Lausanne
Antonis SPINAKIS, Quantos Sarl
Bo SUNDGREN, Central Statistical Office of Sweden
Athanasios TSAKALIDIS Research Academic Computer Technology Institute

Technical Committee
Ana Nora FELDMAN, University of Roma 1
Alf FYHRLUND, Central Statistical Office of Sweden
Christos MAKRIS, University of Patras
Pia MARGERIT, Polytechnic University of Catalonia
Konstantinos MARKELLOS, Research Academic Computer Technology Institute
Penelope MARKELLOU, Research Academic Computer Technology Institute
Gina PANAGOPOULOU,Quantos Sarl.
Vivi PERISTERA, Quantos Sarl.
Maria RIGOU, Research Academic Computer Technology Institute
Bert FRIDLUND, Central Statistical Office of Sweden

This conference could not have been held without the outstanding efforts of Eleni Rigou at the Conference Secretariat. Finally, recognition and acknowledgement is due to all members of the Internet and Multimedia Research Unit at CTI

Spiros SIRMAKESSIS
Assistant Professor

July 2003

Table of Contents

Mining for Gems of Information

Spiros Sirmakessis

Research Academic Computer Technology Institute
61 Riga Feraiou Str., 26221 Patras, Greece
syrma@cti.gr
http://www.ru5.cti.gr

1 Introduction

The rapid progress in digital data acquisition has led to the fast-growing amount of data stored in databases, data warehouses, or other kinds of data repositories. [4] Although valuable information may be hiding behind the data, the overwhelming data volume makes it difficult, if not impossible, for human beings to extract them without powerful tools. In order to relieve such a *data rich but information poor* plight, during the late 1980s, a new discipline named data mining emerged, which devotes itself to extracting knowledge from huge volumes of data, with the help of the ubiquitous modern computing device, i.e. computer.
Due to its interdisciplinary nature, data mining has received contributions from a lot of disciplines such as databases, machine learning, statistics, information retrieval, data visualization, parallel and distributed computing, *etc.*

Text Mining (TM) is an exciting application field in the data mining domain and an area of scientific research that is currently under significant development. According to Grobelnik et al. [2] *"The objective of Text Mining is to exploit information contained in textual documents in various ways, including ...discovery of patterns and trends in data, associations among entities, predictive rules, etc"*. Hearst in [3] defines TM as *"another way to view text data mining is as a process of exploratory data analysis that leads to heretofore unknown information, or to answers for questions for which the answer is not currently known"*.

TM uses techniques from well-established scientific fields (i.e. data mining, machine learning, information retrieval, natural language processing, case-based reasoning, statistics and knowledge management) in an effort to help people gain insight, understand and interpret large quantities of (usually) semi-structured and unstructured data. Typically, TM involves pre-processing of documents, storing and indexing of intermediate results, analysis and visualization of the output [1].

Despite the advances made during the last few years, many issues remain unresolved. Proper co-ordination activities, dissemination of current trends and standardization of the procedures have been identified as key needs, while there are many questions still unanswered, especially to the potential users.

S. Sirmakessis (ed.), *Text Mining and its Applications*

2 The 1st International Workshop on Text Mining & its Applications

The mining community responded very enthusiastically to the 1st International Workshop on Text Mining & its Applications, with about 90 people attending. The workshop brought together experts in the field of TM to present the status, trends and possible future developments in the technology, practices and use of TM.

Session 1

The aim of the first session was to present different application areas of TM by surveying the existing literature and indicate future research initiatives and open problems.

Martin Rajman and Martin Vesely presented a TM approach to the extraction of knowledge from documents. Document processing and visualization is regarded as one of the key topics in the domain of TM. They reviewed the document processing techniques that are required for the natural language pre-processing and described the way data mining techniques are applied in the TM domain. They also presented the most common visualization techniques and discussed their applicability to textual data. At the end of the presentation the focus was set on identifying key issues and areas of current research activity.

Kostantinos Markellos, Penelope Markellou, Maria Rigou and Spiros Sirmakessis discussed the past, present and future of Web Mining. They reported the most representative scientific activities in the three main categories of web mining (content, structure and usage mining), investigated the foreseen future directions, and outlined the new and enticing challenges to be answered in the years to come.

Sergio Bolasco, Francesco Baiocchi, Alessio Canzonetti, Francesca Della Ratta and Ana Feldman provided an overall picture of the applications, sectors and strategies of TM. Text mining is examined following three main dimensions: types of applications (ranging from customer relationship management -CRM- and market analysis to technology watch and patent analysis); the sectors of activity (from financial domain to health sector, from media and communication to public administration); and the schemes of strategy (document pre-processing, lexical and TM processing). They also reported on the results of a survey on text analysis traditions in Italy and some of the most relevant Italian experiences of the business sector in the domain of linguistic technology production and TM solutions.

Session 2

The second session of the workshop consisted entirely of invited presentations describing state-of-art problems in the area of TM.

Support Vector Machines (SVM) was the subject of Gerhard Paas, Joerg Kindermann and Edda Leopold. They showed that although SVMs are effective

classifiers fir text documents, their performance may be increased by various parameter settings (i.e. lemmatization and part of speech tagging). In addition they investigated weighting of the raw frequencies and as it turned out usually raw or logarithmic frequencies combined with redundancy weights perform best. Finally, they demonstrated that n-grams of syllables and phonemes prove especially effective for classification as they reduce the error rates.

An exhaustive survey of web document clustering approaches available on the literature, classified into three main categories: text-based, link-based and hybrid was described by N. Oikonomakou and M. Vazirgiannis. Furthermore, a thorough comparison of the algorithms based on the various facets of their features and functionality was presented. From the review of the different approaches the authors concluded that although clustering has been a topic for the scientific community for three decades, there are still many open issues that call for more research.

Franca Debole and Fabrizio Sebastiani presented the idea of supervised term weighting (STW), a term weighting methodology specially designed for IR applications involving supervised learning, such as text categorization and text filtering. Supervised term indexing leverages on the training data by weighting a term according to how different its distribution is in the positive and negative training examples. They have also proposed that this should take the form of replacing idf by the category-based term evaluation function that has previously been used in the term selection phase; as such, STW is also efficient, since it reuses for weighting purposes the scores already computed for term selection purposes. STW was tested in all the combinations involving three different learning methods and three different term weighting functions, each tested in its local and global version and the results confirmed the overall superiority of gain ratio over information gain and chi-square when used as a STW function.

The need for the automatization of knowledge extraction from text in functional genomics grows with the development of large scale methods like DNA chips. This was the opening theme of Claire Nedellec. Two research lines are mainly explored: hand-coded information extraction patterns and statistics methods based on co-occurrence counting. They yield to low recall in the first case and to low precision in the latter case. Machine Learning applied to knowledge acquisition from corpora overcomes some of these limitations. In a first step, linguistic processing normalizes various textual expressions in order to highlight relevant regularities for learning knowledge extraction patterns in a second step. The syntactic-semantic normalization requires lexical, terminological and ontological resources that are learnable from corpora. An extended description of this process was presented.

Eric Gaussier addressed the problem of processing multilingual collections, for such TM applications as cross-language clustering, categorisation and information retrieval. He showed that in most cases it was not possible to guarantee equivalent processing of different languages and this implies that consistency and performance equivalence across languages are difficult to achieve but constitutes an interesting research topic. He also presented the conceptual differences between methods used to cross the language barrier in TM applications.

Session 3

The third session presented the industrial view and the needs in the area of TM. Antonis Spinakis and Paraskevi Peristera from Quantos presented the main axes of the comparative analysis of TM tools that will be performed in the framework of the NEMIS project. They presented the methodology for comparing software tools based on general evaluation criteria and applied the aforementioned methodology and criteria to compare two TM tools.

Alessandro Zanasi from TEMIS in his presentation claimed that only those that know how to retrieve, analyze and turn into actionable intelligence documents, web pages, emails, chat lines and, generally, public/open sources content, will be able to acquire and maintain this knowledge advantage. An introduction to Temis and to its TM technology, with some indications about Temis current research directions was briefly presented.

Luca Dini and Giampaolo Mazzini from CELI addressed a crucial topic in current CRM processes, i.e. constant monitoring of customer opinions. They used the label "Real Time Customer Opinion Monitoring" to denote the process of retrieving, analyzing and assessing opinions, judgments, criticisms about products and brands, from newsgroups, message boards, consumer association sites and other public sources on the Internet. They suggested that the use of Language Technologies and – more specifically – of Information Extraction technologies provides a substantial help in Customer Opinion Monitoring, when compared to alternative approaches, including both the "traditional" methodology of employing human operators for reading documents and formalizing relevant opinions/facts to be stored, and data mining techniques based on the non–linguistic structure of the page (web mining) or on statistical rather than linguistic analysis of the text (TM in its standard meaning). In the light of these considerations, a novel application (ArgoServer) was presented, where different technologies cooperate with the core linguistic information extraction engine in order to achieve the result of constantly updating a database of product or brand-related customer opinions gathered automatically from newsgroups.

Stefano Spaggiari from Expert System presented the architecture and the functionalities of Cogito®; a linguistic platform of Expert System comprising a set of technologies and proprietary resources. The system does not superficially manipulate a group of words, but relies on the query of a semantic network which contains millions of pieces of information about terms, concepts, abbreviations, phraseologies, meanings, domains and connections among terms thus allowing the retrieval of meanings, the comprehension of natural language, the translation, the sharing and circulation of knowledge.

Furio Camillo from Alma Mater Studiorum take a spatial approach -by transferring models of spatial statistics to the web scenario- to monitoring the behaviour of web users, according to which it is possible to determine, at different levels of depth, the important crossroads (the pages that feature as the centres of polarisation) of a portal. These crossroads are the places where the different *'internauts'* make their fundamental choices for the construction of their individual navigation path. Using TM and semantic/semiotic web mining it is possible to estimate the a-priori distance matrix between two or more pages of an Internet-portal.

Session 4

The last session of the workshop presented algorithmic and other problems related to official statistics. Ludovic Lebart discussed various validation techniques for TM with application to the processing of open-ended questions. Clustering methods and principal axes techniques as well, play a major role in the computerized exploration of textual corpora. However, most of the outputs of these unsupervised procedures are difficult to assess. He focused on the two following issues: *external validation*, involving external data and allowing for classical statistical tests, and *internal validation* based on re-sampling techniques such as bootstrap and other Monte Carlo methods. In the domain of textual data, these techniques can efficiently tackle the difficult problem of the plurality of statistical units (words, lemmas, segments, sentences, respondents).

Mónica Becue, Bert Fridlund, Alf Fyhrlund, Albert Prat and Bo Sundgren talked about the current challenges for statistical information systems and described the different types of such systems, as well as the major processes typically supported. They claim that due to the new options offered by the web technology, there is tremendous increase in the number of actors in the statistical arena (producers, distributors and users). These actors are not sufficiently informed about the ways in which they can benefit form the information technologies and more specifically TM. The presentation explored the possible applications of TM in the world of production and dissemination of official statistics, including advanced website warehouse querying, coding and processing answers to open-ended questions, sophisticated access to internal and external sources of statistical meta-information, and pulling statistical data and metadata from the web sites of sending institutions.

3 Conclusions

The area of TM seems to be an active research area, which also attracts the interest of the industrial sector. During the workshop, participants had the opportunity to attend several presentations from scientific and industrial representatives. The discussions following each session focused on specific problems and open issues. It was clear from the quality of the discussions that the scientific community has to provide feedback to the industrial sector and vice versa. New problems and needs were presented and different application areas of the scientific development were initiated. This is a strong endorsement of the high interest of all sides in the new developments and applications of TM.

References

1. Dörre , J., Gerstl, P., Seiffert, R.: Text Mining: Finding Nuggets in Mountains of Textual Data. *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California, United States (1999) 398-401.
2. Grobelnik, M., Mladenic, D., Milic-Frayling, N.: Text Mining as Integration of Several Related Research Areas: Report on KDD'2000 Workshop on Text Mining (2000).
3. Hearst, M.: "Untangling Text Data Mining,". *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (1999).
4. Zhou. Z.-H. Three Perspectives of Data Mining Artificial Intelligence, 143(1): 139-146 (2003)

From Text to Knowledge: Document Processing and Visualization: a Text Mining Approach

Martin Rajman[1], Martin Vesely[1]

[1] Laboratory of Artificial Intelligence (LIA),
EPFL, Swiss Federal Institute of Technology
Ch-1015 Lausanne, Switzerland

Abstract: Document processing and visualization was identified as one of the key topics in the domain of Text Mining. For this reason, the WG1 working group was created in the NEMIS project. In the areas it covers, this working group contributes to the production of a roadmap for follow-up research and technological development in text mining, which is the overall goal of the NEMIS project. Relevant research topics have been identified in the document processing and visualization domains. An analysis of the associated state-of-the-art techniques and state-of-the-software tools have been carried out. The results of this analysis are presented to foster discussion and contribute to a collaborative improvement of the roadmap content.

1 Introduction

Text Mining refers to the process of knowledge discovery from large collections of unstructured textual data [Feldman95][Hearst99]. In contrast, data mining was previously studied and practiced mainly in connection with structured data stored in databases. This type of data mining activity, usually known as knowledge discovery in databases (KDD) [Fayyad95], is often defined as a non-trivial extraction of implicit, previously unknown, and potentially useful information from data [Frawley91]. In principle, the process of text mining has a similar structure to the process of knowledge discovery, with, however, some additional requirements and constraints imposed by the specific properties of unstructured textual data expressed in natural language.

As natural language is inherently unstructured, there is a need for some pre-processing of the available documents in order to reconstruct the missing data structure. Traditionally, this structure has a form of a feature vector the dimensions of which are associated with the terms extracted from the full content of the documents. Once the targeted structured data representation is produced, the question arises to what extent it is possible to apply standard data mining techniques. We focus on these issues in the document processing section.

Another important issue is the high dimensionality of the vectorial representations that are produced and processed at several stages of the text mining process. To deal with this problem, various dimension reduction techniques need to be applied, typically reducing the representation space from thousands into hundreds of features.

S. Sirmakessis (ed.), *Text Mining and its Applications*

Dimension reduction is also a key issue for the visualization techniques that are required to support the interpretation of text mining results, as end users are naturally limited to two or three dimensions.

In the next section we review the document processing techniques that are required for the subsequent natural language pre-processing step. In addition to procedures specifically dedicated to the processing of the natural language, these techniques often also correspond to some combination of two complementary approaches: information retrieval, for example to select documents to be processed [Strzalkowski99] and information extraction, for example to extract the features composing the structured representations [Pazienza97][Gaizauskas98]. In the third section, we describe how data mining techniques are applied in the text mining domain. In the fourth section, we review the most common visualization techniques and discuss their applicability to the textual data. In the whole document we try to review the main approaches and techniques, aiming at identification of the key research issues.

2 Document Preprocessing

As mentioned in the introduction, Natural Language Processing (NLP) is the main focus for the document preprocessing phase. The required NLP techniques involve both statistical and machine learning [Daelemans02] approaches, extended by artificial intelligence and information theoretic approaches.

Firstly, as far as evaluation and diffusion of these techniques is concerned, it is important to notice the significant contribution of competitive research evaluation campaigns, such as the Message Understanding Conferences (MUC) [Poibeau03], the Text Retrieval Conferences (TREC), organized annually by NIST and DARPA, the Document Understanding Conferences (DUC) focusing on Text Summarization, or the SENSEVAL campaign evaluating Word Sense Disambiguation systems. The important impact of international conferences such as the Conference for Natural Language Learning (CoNLL) and the conferences and workshops organized under the auspices of the Association for Computational Linguistics (ACL, EACL) also has to be stressed. For this presentation, we suggest to adopt the generic information extraction model presented by [Hobbs93] dividing the pre-processing of documents into a sequence of distinct natural language processing tasks that are described in the subsequent subsections.

2.1 Data Selection and Filtering

Data selection and filtering aims at performing a first rough reduction of the vast amount of documents available from numerous information sources in order to avoid the overload related to the rather computationally intensive pre-processing and mining processes.

In the text mining field, data selection consists in the identification and retrieval of potentially relevant documents from available data sources. During this data selection step, the main focus is on the exogenous information associated with the documents, usually represented by explicit descriptive metadata attached to them, such as

keywords or descriptors. The main issues related with the task of data selection are therefore tightly connected with metadata interoperability that has been a subject of recent research initiatives, such as the interoperability frameworks developed within the Dublin Core Metadata Initiative (DCMI) and W3C, the Dublin Core and the Resource Description Framework (RDF) [Lassila99].

In contrast to data selection, data filtering focuses on the endogenous information (i.e. the actual content of documents) to evaluate the document relevance. The endogenous information is sometimes denoted as 'implicit metadata'. Basic concepts for textual data filtering are described in [Oard97]. Compared to the tasks in information retrieval, data filtering is quite close to the problem of document routing, where the focus is rather on the documents in the data source and their changes than on the queries. As for performance evaluation, the traditional IR measures such as precision, recall and their variants (e.g. the F-measure, the E-measure, non-interpolated and interpolated average precision and the average precision at recall level) are used [Yang99b]. One of the important current research issues is the definition of efficient relevance metrics that are applicable on large volumes of textual data streams.

2.2 Data Cleaning

The task of data cleaning is to remove noise from the textual data in order to improve its quality. This goal is also often referred to as "avoiding the GIGO" (Garbage-In-Garbage-Out) effect. Noise can be the consequence of various error sources, leading for example to data inconsistencies and outlying values. The importance of data cleaning also significantly increases when data comes from multiple heterogeneous sources, typically when transformed from one data structure into another or, as it is the case in text mining, when the associated structure is being created out of unstructured data. Among the important data cleaning tasks that are especially relevant in the scope of text mining, one can cite [Rahm00]:

- Spelling error correction.
- Reconstruction of missing accentuation.
- Letter case normalization.
- Abbreviation normalization.
- Language identification (to filter out parts that are not in the processed language(s)).
- Production of (meta-)linguistic information (PoS tagging, named entity tagging, identification of syntactic roles ,...).

Language identification can be understood as a text categorization problem, where the categories represent the considered languages. Several methods have been developed for the language identification problem. Simple methods include the small word approaches and the n-gram approach [Grefenstette95]. More sophisticated methods include the Rank Order, Monte Carlo methods and methods based on Relative Entropy that are considered to perform at the limit of theoretically achievable results.

In addition, when processing multiple heterogeneous data sources, data integration techniques are applied in order to obtain a homogeneous data input (including

resolution of value conflicts and attribute redundancy[1]). The general task of data integration is defined in more details in the domain of document warehousing [Sullivan01].

2.3 Document Representation

The most common document representation model relies on the notion of feature vectors mapped into n-dimensional vector space (Vector Space Model). In the simplest approach the dimensions of the full-scale feature vectors are associated with the words extracted out of the documents (collection vocabulary). This representation, often referred to as the "bag-of-words" approach, although very easy to produce, is not considered to be optimal for several reasons. One of its main drawbacks is the high dimensionality of the representation space (which grows with the size of the vocabulary used). Notice that the resulting representations with dozens of thousand of dimensions put severe computational constraints on the text mining process. Some dimension reduction is usually performed, leading to the selection of features strongly representative of the content of the documents. The criteria used for feature selection are usually based on word frequency[2] or on more sophisticated selection methods relying on criteria such as chi-square tests, information gain or mutual information. Varying importance for the individual features (integrating for example their discriminative power for the documents in the processed collection) can be taken in account through various weighting schemes, as it is done in the IR field with, for example, the tf.idf (term frequency x inverted document frequency) weighting scheme or some of its variants including Rocchio weighting or Ide weighting [Monz01].

The bag-of-words representations that can be easily generated are used for many content sensitive tasks. However, more sophisticated representations are being investigated for more complex tasks, including more structured semantic models. For example, the use of ontologies has been suggested in [Hotho01], where related terms such as synonyms can be aggregated resulting in a reduced document representation space. The resulting representations are then more oriented on concepts rather than on just words.

Finally, the Generalized Vector Space Model was developed for multi-lingual document representation [Carbonell97][Yang98]. The appropriateness of the vector space model in multi-lingual environments is discussed for example in [Besancon02]. Evaluation of Cross Language Information Retrieval (CLIR) was performed in the scope of TREC-6 [Gaussier98].

2.4 Morphological Normalization and Parsing

Morphological normalization refers to a group of natural language tasks such as stemming, lemmatization and Part-of-Speech tagging that aims at the production of canonical surface forms.

---

[1] An attribute is redundant if it can be derived from other attributes.

[2] The distribution of word frequencies in natural language texts tends to follow the Zipf's law with many *hapax legomena* (words occuring only once in the text collection).

The finite-state automata technology enables to define re-usable regular expressions that generalize over language patterns. Finite-state transducers (FST) were developed to enable the description of regular relations between the morphological base and the corresponding expression in some morphological form. Since these relations are regular, they allow the automated morphological normalization computation [Chanod96]. An example system using finite-state approach is the well known FASTUS system [Appelt93]. Incremental robust parsing (e.g. the XIP parser recently developed at the Xerox Research Center, Europe) and the use of finite state transducers for shallow syntactic parsing was studied and illustrated with examples of applications in [Ait97].

Parsing aims at assigning some syntactic structure to a morphologically normalized text. It mainly includes text segmentation, sentence chunking into smaller syntactic units such as phrases or syntagms, and the production of syntactic relations between the identified units. Approaches to parsing based on shallow techniques have reached good results for most of the above mentioned tasks, however more sophisticated techniques of parsing as well as incremental robust parsing are still investigated.

In general, parsing techniques are often divided into two main families of approaches: probabilistic and non-probabilistic. Probabilistic approaches include techniques such as memory-based shallow parsing [Daelemans99] or statistical decision trees [Magerman94].

An often discussed research topic is the use of deep (full) parsing as an alternative to shallow approaches. However, deep parsing has not yet been developed in full extent, partly because of the fact that shallow parsing techniques perform well enough for a large number of applications. Nevertheless, deep parsing is advocated, for example in [Uszkoreit02], emphasizing that there is a real potential for improvement, especially for applications requiring the resolution of more complex linguistic problems such as anaphora resolution. Attempts to integrate shallow and deep parsing have also been recently undertaken and are reported in [Daum03][Cryssmann02].

2.5 Semantic Analysis

One of the main objectives of semantic analysis in the domain of text mining is to resolve semantic ambiguities, for example generated by the presence of synonymous and polysemous expressions in the documents. In this perspective, the main areas for semantic analysis are word sense disambiguation (WSD) [Veronis98] and anaphora resolution [Mitkov02].

For the co-reference/anaphora resolution problem, one of the important current open questions is whether such kind of resolution can be performed purely with syntactic tools or if it also requires the integration of semantic or even pragmatic knowledge. A thorough scientific analysis in the domain of co-reference and anaphora resolution can be found in a monograph recently published by [Mitkov02].

In word sense disambiguation, the context of the expression to disambiguate is analyzed in order to assign the appropriate word meaning. The concepts have been introduced as shared word meaning independent from the lexical realization. Concepts are essentially semantically unambiguous and the main issue of WSD can therefore be formulated as the search for a mechanism that reliably assigns concepts