Text Mining as Integration of Several Related Research Areas: Report on KDD'2000 Workshop on Text Mining

Marko Grobelnik
J. Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
Marko.Grobelnik@ijs.si

Dunja Mladenic
J.Stefan Institute, Ljubljana, Slovenia and
Carnegie Mellon University
Pittsburgh, PA, USA
Dunja.Mladenic@ijs.si, cs.cmu.edu

Natasa Milic-Frayling
Microsoft Research Ltd.
Cambridge CB2 3NH
United Kingdom
natasamf@microsoft.com

ABSTRACT

In this paper we give an overview of the KDD'2000 Workshop on Text Mining that was held in Boston, MA on August 20, 2000. We report in detail on the research issues covered in the papers presented at the Workshop and during the group discussion held in the final session of the Workshop.

Keywords

Text mining, information extraction, information retrieval, natural language processing, KDD workshop report.

1. INTRODUCTION

The growing importance of electronic media for storing and disseminating text documents has created a burning need for tools and techniques that assist users in finding and extracting relevant information from large data repositories. Information management of well organized and maintained structured databases has been a focus of the Data Mining research for quite sometime now. However, with the emergence of the World Wide Web, there is a need for extending this focus to mining information from unstructured and semi-structured information sources such as on-line news feeds, corporate archives, research papers, financial reports, medical records, e-mail messages, etc.

The objective of Text Mining is to exploit information contained in textual documents in various ways, including the type of analyses that are typically performed in Data Mining: discovery of patterns and trends in data, associations among entities, predictive rules, etc. Dealing with free text, in contrast to relatively 'clean' and well-organized data, presents many new challenges.

Over the past three decades a number of research areas have been addressing various aspects of processing textual information. Among those are linguistics and computational linguistics, information retrieval, machine learning, statistical learning, etc. We recognize that the advances made in these research areas are essential for approaching the problem of text mining. Thus the main inspiration for organizing this Workshop came from our desire to facilitate communication among researchers and practitioners from related and complementary research areas who are working on Text Mining and similar problems with possibly quite different approaches. The Workshop had the objective to enable presentation and exchange of ideas, stimulate discussions of important issues, and identify promising strategies and research directions.

2. PRESENTATIONS AND DISCUSSIONS

We were very pleased with the response to the Workshop's Call for Papers and the range of topics covered by the accepted papers. The Workshop Notes [1] include two extended abstracts from invited speakers, 11 full papers, and 15 short papers covering various topics from four research areas: Text Mining (or Text Learning), Information Retrieval, Natural Language Processing, and Information Extraction.

In the following sections we briefly discuss the main research directions taken by the authors of the presented full papers, list the topics presented in the poster papers, and outline the main points of the group discussion held at the end of the Workshop. All the accepted papers are available at the Workshop Web site[2].

2.1 Full Papers

2.1.1 Information Extraction and Text Mining

One way to approach Text Mining problem is to combine Information Extraction techniques with Data Mining of Knowledge Bases. Within this framework Information Extraction is an essential phase in text processing. It facilitates the automatic or semi-automatic creation of knowledge bases, more or less domain specific. Such knowledge bases are further processed using standard Data Mining techniques. In this manner we leverage to a great extent the effectiveness of the Data Mining techniques.

The Web→KB project from Carnegie Mellon University [Ghani & alt.: *Data Mining on Symbolic Knowledge Extracted from the Web*] is an illustration of this integrated approach to Text Mining. The approach taken involves Information Extraction from both free texts of Web site pages, thus completely unstructured textual data, and more or less structured Web information resources such as on-line directories (e.g., www.hoover.com).

An important issue that arises in this approach is the accuracy of Information Extraction and the impact that the noise in the knowledge base has on the effectiveness of Data Mining techniques. It is encouraging to learn that for a system configuration like DISCOTex [Nahm & Mooney: *Using Information Extraction to Aid the Discovery of Prediction Rules from Text*] the difference in the accuracy of the knowledge discovered from an automatically extracted database is close to that discovered from a manually constructed database, at least for the tested domain.

The issue of Information Extraction accuracy naturally leads to questions about how to improve it and thus reduce the effort required for cleaning up automatically created knowledge bases. [Caruana & Hodor: *High Precision Information Extraction*] present a High Precision Information Extraction Workbench (HPIEW) designed to facilitate a machine assisted information extraction by domain experts. The objective of this and similar efforts is to establish an appropriate trade off between the scalability and accuracy of Information Extraction. While manual extraction of information is very accurate but does not scale well to large data sets automatic information extraction scales well but is typically far less accurate. In the presented HPIEW experiment the workbench allowed the experts to review 5,000 files in an afternoon and extract data with an estimated precision and recall greater than 99.9%.

2.1.2 Text Mining applications: Finding themes/topics in text

In addition to Text Mining problems that can be handled by combining Information Extraction and Data Mining techniques there are others that benefit from the creation of appropriate language models and rely on direct statistical analysis of text representing features. These problems often involve discovering and exploiting the relationship between the document text and an external source of information such as time stamped streams of data (e.g., stock market quotes), topic hierarchy from encyclopedias or Web portals, document placement into folders (Bookmarks/Favorites) by a group of users, etc. Text processing here typically involves natural language processing, information retrieval, computational linguistics, and statistical analysis.

Into this category of problems falls the attempt by Chakrabarti & Batterywala [in *Mining Themes from Bookmarks*] to create document taxonomy suitable for a given community of users. Discovery of such taxonomies involves the discovery of document themes. This is based on the (collective) user classification of documents into Favorites and the occurrences of tokens in the document text.

Mather & Note [in *Discovering Encyclopedic Structure and Topics in Text*] focus on discovering encyclopedic structure and topics in unstructured document text. The objective is to design a method by which electronic publications could be analyzed for the presence of information related to encyclopedic topics. This information would be automatically extracted, labeled, and inserted into encyclopedia type knowledge resources.

2.1.3 Text Mining applications: Mining time-tagged text

Two systems that explore the time dimension in relation to the document content are TimeMines [Swan & Jensen: *TimeMines: Constructing Timelines with Statistical Models of Word Usage*] and EAnalyst [Lavrenko & alt.: *Mining of Concurrent Text and Time-Series*] from University of Massachusetts at Amherst.

TimeMines automatically generates timelines by detecting documents that relate to a single topic (an event in the 'history'). This is achieved solely by selecting and grouping semantic features in the text based on their statistical properties.

The EAnalyst on the other hand uses two types of data and attempts to find the relationship between them: the textual documents and numerical data, both with time stamps. The system discovers the trends in time series of numeric data (e.g., stock prices) and attempts to characterize the content of textual data (e.g., news articles) that precede these events. The objective is to use this information to predict the trends in the numeric data based on the content of textual documents that precede the trend. For example, to predict the trend in stock prices based on the content of new articles published before that trend occurs.

2.1.4 Text Mining applications: Mining e-mail data

Direct applications of text categorization techniques to managing e-mail documents is presented in [Rennie: *ifile: An Application of Machine Learning to E-Mail Filtering*] and [DeVil: *Mining E-mail Authorship*]. [Rennie] describes ifile, a system for filtering e-mail messages to the user created folders using a variant of the Naïve-Bayes classification algorithm. Ifile's distinct feature is a continuous updating of the classifiers based on newly classified e-mail messages. [DeVil] on the other hand uses SVM to identify the authorship of a document based on its structural and linguistic characteristics. The focus of this particular study is on the specific issues related to the authorship of e-mail messages.

In the above mentioned Text Mining problems and approaches it is often the case that text categorization plays an important role, typically as an intermediate step. For example, Information Extraction and automatic building of knowledge bases may be facilitated by categorizing documents into pre-defined categories and labeling them appropriately. For that reason, efficient and accurate text categorization represents an important problem. The following two Workshop papers focus on the text classification algorithm itself.

2.1.5 Text categorization methods using machine learning

Zhang [in *Large Margin Winnow Methods for Text Categorization*] shows how the large margin versions of the Winnow algorithm can be successfully applied to text categorization and achieve text classification performance comparable with Support Vector Machine (SVM) on the Reuters data. The result of this evaluation confirms that both the Perceptron and the large margin Winnow family of algorithms perform well for text categorization problems.

Shankar & Karypis [in *A Feature Weight Adjustment Algorithm for Document Categorization*] implemented and evaluated an iterative (and fast) algorithm for adjusting weights of the category features. In that manner they achieved an efficient classification scheme that improves the performance of the centroid-based classifier and is competitive with SVM.

2.2 Posters

In addition to the full paper presentations there were two Poster Sessions covering 15 poster presentations. Authors of the poster papers had an opportunity to inform the Workshop participants of the main ideas of their work in short presentations. Detailed discussion of their work took place during a separate Poster Viewing session.

The first poster session included 8 papers covering a wide variety of topics: documents clustering, databases merging, mining textual associations, modeling distribution of noun phrases, disambiguation of annotated text from audio data, and fuzzy semantic typing. In particular, document clustering was addressed by [Bao Ho, Nguyen, and Kawasaki: *Tolerance Rough Set Model*

Approach to Document Clustering], [Shin & Zhang: *Extracting Topic Words and Clustering Documents by Probabilistic Graphical Models*], and [Steinbach, Karypis, and Kumar: *A Comparison of Document Clustering Techniques*]. Zhu & Ungar presented their work on *String Edit Analysis for Merging Databases*. Dias, Guillore, and Lopes talked about *Mining Textual Associations in Text Corpora*. Various aspects of linguistic analyses of text were presented in three papers: [Corston-Oliver: *Modeling the Distribution of Noun Phrases*], [Khan & McLeod: *Disambiguation of Annotated Text of Audio Using Ontologies*] and [Subasic & Huettner: *Calculus of Fuzzy Semantic Typing for Qualitative Analysis of Text*].

The second poster session included 7 papers. Six of these papers were in the area of text categorization covering the issues from user modeling [Kim, Hall, and Keane: *A Hybrid User Model in Text Categorisation*] to use of knowledge resources [Wilcox, Hripcsak, and Friedman: *Using Knowledge Sources to Improve Classification of Medical Text Report*] and various conceptual and contextual features [Jensen & Martinez: *Improving Text Classification by Using Conceptual and Contextual Features*] to co-training [Nigam & Ghani: *Understanding the Behavior of Co-training*]. The remaining paper was on feature extraction [Fontaine & Matwin: *Feature Extraction Techniques for Unintelligible Texts*].

2.3 Discussion

During a 30-minute group discussion the Workshop participants had an opportunity to share their views on a number of important issues related to the current and future developments in the text mining research.

It was recognized that Text Mining is an emerging research area that addresses a class of problems that require expertise and resources from various complementary areas of research. At present, it appears to draw techniques and methodologies mostly from the areas of Machine Learning, Information Retrieval, Natural Language Processing, Information Extraction, and Data Mining.

It was noted by the Workshop participants that these individual areas of research stay separated and researchers rarely present and publish at the conferences that are not in their primary area of expertise. There are various reasons and factors that contribute to that. Text Mining Workshops and Conferences thus offer the opportunities for researchers who do work on problems that require interdisciplinary effort to present their work and obtain constructive and valuable feedback. We were pleased that the KDD'00 Text Mining Workshop was a contributor to that cause.

Furthermore, it has been noted that the linguistic resources and tools represent the enabling technologies for the text analyses and thus more of these, readily available for experimentation are needed. Similar conclusion holds for the availability of data for experimentation. The researchers from the Encyclopedia Britanica informed us of their intention to make some of their data publicly available for use by researchers.

The participants also raised the issue of standards that would enable more effective deployment of test mining tools and techniques (as it has been already achieved for data mining).

Finally, the discussion touched on the need for teamwork. As many text mining problems require interdisciplinary knowledge the logical answer to that challenge is an interdisciplinary team of experts in various research areas. Indeed, the collaboration of experts opens the possibility of addressing problems in Text Mining that none of them individually could undertake successfully. Researchers from Boeing confirmed that indeed their successful text mining projects typically involve experts working as a team.

The discussion was closed with the great optimism that the area of Text Mining will continue to develop, becoming further recognized for its importance and that the events similar to our Workshop would become more common in the near future.

3. Program Committee

-Helena Ahonen, University of Helsinki, Helsinki, Finland

-Simon Corston-Oliver, Microsoft Research, Redmond, WA, USA

-Mark Craven, University of Wisconsin, USA

-Walter Daelemans, Tilburg University, Tilburg, Netherlands

-Susan Dumais, Microsoft Research, Redmond, WA, USA

-David Elworthy, Microsoft Research Ltd., Cambridge, UK

-Ronen Feldman, Instinct Software, Israel

-Marko Grobelnik, J.Stefan Institute, Ljubljana, Slovenia

-Thorsten Joachims, Universitaet Dortmund, Dortmund, Germany

-Rosie Jones, Carnegie Mellon University, Pittsburgh, PA

-Natasa Milic-Frayling, Microsoft Research Ltd., Cambridge, UK

-Dunja Mladenic, CMU, Pittsburgh, PA, & J. Stefan Institute, Slovenia

-Jason Rennie, Massachusetts Institute of Technology, MA, USA

-Stephen Robertson, Microsoft Research Ltd., Cambridge, UK

-Sean Slattery, Carnegie Mellon University, Pittsburgh, PA, USA

-Ian Witten, University of Waikato, Hamilton, New Zealand

4. ACKNOWLEDGMENTS

We would like to take this opportunity to thank all the authors who submitted their work for consideration for the Workshop and the members of the Program Committee for their great assistance in reviewing the submitted publications.

5. REFERENCES

[1] The Workshop notes for KDD-2000 Workshop on Text Mining, August 2000, M. Grobelnik, D. Mladenic, N. Milic-Frayling (eds.). http://www.cs.cmu.edu/~dunja/KDDpapers/ProcTMKDD00.zip

[2] Web site for the KDD-2000 Workshop on Text Mining: <http://www.cs.cmu.edu/~dunja/WshKDD2000.html>

About the authors:

Marko Grobelnik works at the Department of Intelligent Systems of the Jozef Stefan Institute since 1984, first as a high-school student and since 1997 as a researcher. Most of his research work is connected with the study and development of data mining techniques and their applications to different problems in economy, medicine, manufacturing, and game theory. His current research focus is on data mining with particular emphasis on learning from text for large text data sets. <http://www-ai.ijs.si/MarkoGrobelnik>.

Dunja Mladenic has been associated with the Department of Intelligent Systems of the Jozef Stefan Institute, Ljubljana, Slovenia since 1987, first as an undergraduate student and since 1992 as a researcher. In 1996/97 she spent a year working on her PhD thesis at Carnegie Mellon University, Pittsburgh, PA, USA where she is currently serving as a visiting faculty (2000/01). She obtained her MSc and PhD in Computer Science at the University of Ljubljana in 1995 and 1998 respectively. Most of her research work is connected with the study and development of machine learning techniques and their applications to real-world problems that arise in various areas, e.g., medicine, pharmacology, manufacturing, economy. Her current research focus is on using machine learning for data analysis, in particular, on learning from text in the context of the Web and intelligent agents
<http://www.cs.cmu.edu/~dunja>
<http://www-ai.ijs.si/DunjaMladenic>.

Natasa Milic-Frayling is a researcher at the Microsoft Research lab in Cambridge, UK. She holds a PhD in Applied Mathematics from Carnegie Mellon University, Pittsburgh, PA. Her research interest ranges from identifying and solving specific information management problems that arise in highly distributed environments such as the WWW to answering various fundamental questions related to combining linguistic and statistical models in information management systems.
<http://research.microsoft.com/users/natasamf>