Building a Document Corpus for Manufacturing Knowledge Retrieval

Y. Liu [1], H. T. Loh [1], S. B. Tor [2]
[1] Singapore-MIT Alliance, National University of Singapore, Singapore 117576.
[2] Singapore-MIT Alliance, Nanyang Technological University, Singapore 639798.

***Abstract*—When faced with challenging technical problems, R&D personnel would often turn to technical papers to seek inspiration for a solution. The building of a corpus of such papers and the easy retrieval of relevant papers by the user in his query is an area that has not been systematically dealt with. This is an attempt to build such a corpus for manufacturing R&D personnel. Manufacturing Corpus Version 1 (MCV1) is an archive of more than 1400 relevant manufacturing engineering papers between 1998 and 2000. In this paper, the origins and motivation of building MCV1 is discussed. The innovative coding process which is specially designed for manufacturing companies will be presented. All other relevant issues, like coding policy, category codes and input documents, will be explained. Finally, two quality indicators which integrate all concerns about coding quality will be examined.**

***Index Terms*—corpus, engineering paper resources, manufacturing**

I. Introduction

Intense global competition in recent years has reshaped the manufacturing industry dramatically. Faced with this pressure, manufacturing companies have responded by concentrating on core competencies, such as R&D capabilities, response time to the market, production planning and supplier management.

When faced with challenging problems, one source of inspiration to which the R&D personnel would turn to is archival papers in the research literature. Such documents are more and more becoming electronically accessible and growing at an explosive rate. This dramatic change has basically two implications for the aforementioned research personnel. On the one hand, these documents are a rich resource of information and knowledge which may be utilized by manufacturing companies to solve technical problems. On the other hand, the biggest challenge for any company today is how to handle such huge volume of textual data and information.

In the last decade data mining techniques have been gaining popularity as a tool to discover patterns and knowledge [1]–[4]. It has been successfully applied in many fields, starting with the finance sector. More recently, it has been applied to the manufacturing domain, especially in the area of design, quality control, and customer service [5]–[7]. It has proven to be useful in helping companies in the understanding of manufacturing process and equipments, as well as consumer behavior in the market. However, most of what has been done in the manufacturing domain is numerical data-mining. To handle textual data, we have to turn to text-mining, which is more complicated though it shares many common techniques with numerical data-mining. In text-mining, we aim at analyzing large sets of documents for the purpose of pattern and knowledge discovery by using statistical, machine learning based, information retrieval based and natural language processing based techniques [8]–[19].

In order to apply text-mining in the manufacturing research domain, at least one set of original documents which is called corpus in this area of research is needed. However, none of existing electronic corpora is manufacturing centered or mainly about manufacturing related issues, like manufacturing process, manufacturing equipments, design issues, materials, quality issues, etc. Existing electronic corpora are mainly about daily news or medical issues. Hence, there is motivation to begin building a manufacturing centered corpus and develop techniques suitable for knowledge mining in such a corpus. The benefits will be two-fold. The direct benefit is to assist researchers to be able to retrieve information and mine for knowledge in manufacturing applications. The other benefit is that experience gained can contribute towards creating such corpora for other specific areas of research.

Manufacturing Corpus Version 1 (MCV1) is an archive of 1434 English language manufacturing related engineering papers. It combines all engineering technical papers from Society of Manufacturing Engineers (SME) from year 1998 to year 2000. The final output of each document has been formatted as XML files. One advantage of using XML format is the contents of each engineering paper can be clearly separated, for example title, authors, abstract, full text and topic labels assigned to each document etc. Therefore, it is obviously helpful to researchers for data access, exchange and manipulation.

Having described the motivation of building MCV1, the

Y. Liu is a Ph.D. candidate in Innovation in Manufacturing Systems and Technology (IMST) of Singapore MIT Alliance (SMA).

H. T Loh is an Associate Professor in the Department of Mechanical Engineering, National University of Singapore.

S. B. Tor is an Associate Professor in the School of Mechanical and Production Engineering, Nanyang Technological University.

rest of this paper is organized as follows. The characteristics of common existing electronic corpora and the difference between MCV1 and these corpora will be discussed. The innovative coding process of MCV1 is presented and finally two quality indicators which have been used to measure the coding quality are explained.

II. Existing Electronic Corpora and MCV1

In the domain of text mining and information retrieval, there are several existing corpora available for research. OSHUMED contains 348,566 references, which are derived from the subset of 270 journals covered in the KF MEDLINE Primary Care product ranging from 1987 to 1991 [20]. Reuters-21578 was originally collected and labeled by Carnegie Group, Inc. and Reuters, Ltd. in the course of developing the CONSTRUE text categorization system [21], [8], [9]. It contains 21,578 articles appeared in the Reuters newswire in 1987 and the articles are marked in SGML tags. Reuters Corpus Volume 1 (RCV1) is one recently available from Reuters, Ltd [22], [21]. It is an archive of over 800,000 manually classified newswire articles. It covers the newswire produced by Reuters from 20/08/1996 to 19/08/1997, and is prepared in XML format and available on two CD-ROMs.

However, the following weaknesses of the existing corpora have been noted by the authors.

- Lack of the full document text (e.g. OSHUMED)
- Too fine granularity of categories (e.g. OSHUMED)
- Inconsistent or incomplete category assignment (e.g. Reuters-21578)
- Lack of documentation about the preparation process of documents collections (e.g. OSHUMED and Reuters-21578)
- Usually built up by using a serial process with a large number of operators involved (e.g. RCV1)

Besides the above considerations, the fact that most of the content of all these text collections are not manufacturing relevant is the biggest weakness for our needs. Therefore, with such motivation in mind, the following issues are taken into consideration when building MCV1.

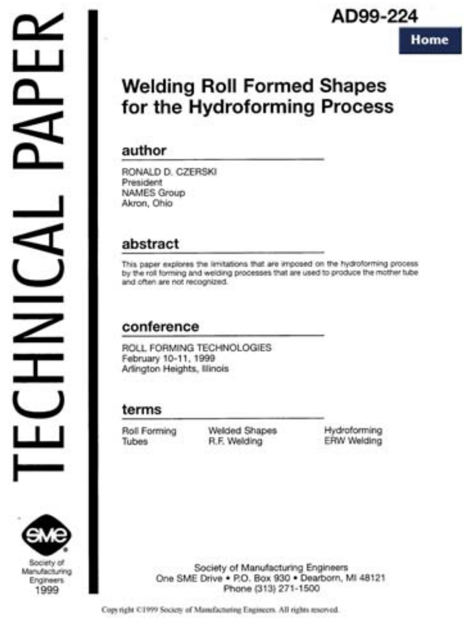
- This is the first corpus which aims at manufacturing industry. The selected documents sources represent knowledge in manufacturing context.
- An innovative coding process will be adopted to code the documents collection. It will address the concerns of availability of human labors, time, cost, etc. And it can be applied to build corpus for other industry.
- Details of how to build the MCV1 will be archived and attached to the corpus.
- A set of quality concerns will be addressed in the coding process, like coding inconsistency, incomplete category assignment, unbalanced documents distribution, etc. New and more meaningful coding quality indicators will be developed to indicate the quality of this manufacturing corpus.
- Full document text will be provided.
- Statistics information of corpus (e.g. distribution of labels, support of each label to a specific paper) will be provided

III. Coding the Manufacturing Corpus Version 1

More than 90 editors were dedicated to the creation of RCV1 [21, [22]. Bearing that in mind, one of the biggest concerns in building a corpus is whether a company has enough well trained and dedicated personnel to handle this. Consequently, a feasible and practical way specially designed for manufacturing companies is devised.

A. Input Sources – Engineering Papers and Coding Labels

The Society of Manufacturing Engineers (SME) has provided us with 1434 of its technical papers from year 1998 to year 2000. These 1434 papers have been used as the input documents for MCV1. Fig.1 shows the front page of a typical SME technical paper.



TECHNICAL PAPER

AD99-224

Home

Welding Roll Formed Shapes
for the Hydroforming Process

author

RONALD D. CZERSKI
President
NAMES Group
Akron, Ohio

abstract

This paper explores the limitations that are imposed on the hydroforming process by the roll forming and welding processes that are used to produce the mother tube and often are not recognized.

conference

ROLL FORMING TECHNOLOGIES
February 10-11, 1999
Arlington Heights, Illinois

terms

Roll Forming	Welded Shapes	Hydroforming
Tubes	R.F. Welding	ERW Welding

SME
Society of
Manufacturing
Engineers
1999

Society of Manufacturing Engineers
One SME Drive • P.O. Box 930 • Dearborn, MI 48121
Phone (313) 271-1500

Copyright ©1999 Society of Manufacturing Engineers. All rights reserved.

Fig. 1. The front page of a typical SME engineering paper provided

As for coding labels, basically we adopted the taxonomy provided by SME for manufacturing industry. It is called Manufacturing Knowledge Architecture (MKA) in our research.

In order to facilitate data processing all MKA items have been coded. A part of MKA labels are shown as follows:

……

C04. Finishing & Coating

C0401	Finishes, Curing
C0402	Finishing & Coating Fundamentals
C0403	Material & Part Handling for Finishing
C0404	Parts Cleaning, Degreasing
C0405	Quality & Inspection of Finishes
C0406	Substrate Selection & Pretreatment
C0407	Coating Specific Substrates
C040701	Painting, Metal Substrates
C040702	Painting, Plastics Substrates
C040703	Painting, Wood Substrates
C0408	Finishing Processes
C040801	Anodizing
C040802	Automated Coating
C040803	Dip Coating
C040804	Electrocoating (E-Coat)
C040805	Electrostatic Finishing
C040806	Metallizing
C040807	Painting
C040808	Plating & Electroplating
C040809	Powder Coating Processes
C040810	Robotic Finishing
C040811	Spray Finishing
C040812	Vapor Deposition
C04TH	Others

……

So in total, there are four levels of labels in MKA including manufacturing as the root.

B. Coding Policy

The coding policies serve as the rules to guide the coding operators during the coding process. These need to be explained explicitly at the beginning of the coding process. This will help to reduce the coding errors and maintain the good quality of coding.

Some coding policies have been mentioned in the literature [22] and they have grouped into two main policies in the literature [21]. The essence of these two policies has been adopted by us.

- Boundary Policy: Each article has to be assigned at least one topic label. If none of labels can be matched, then label <Others> will be chosen. Furthermore, since maximizing the information coverage is desirable, there is therefore no upper limit on the number of the most specific suitable labels (end leaf labels) assigned to any article.
- Hierarchy Policy: Coding operators are required to assign the most specific suitable labels (end leaf labels) to the articles. In order to save time and energy, all ancestors of one specific label are not required to be assigned by coding operators. The system can obtain them automatically.

In the meantime, the authors also note that it is not necessary to apply multi sorts of labels in our work besides the topic labels (e.g. industry codes and region codes in RCV1).

C. The Coding Process

Usually, a serial coding process which involves a large number of people is adopted by industry to build the corpus. E.g. in RCV1, one document is coded by an operator first with his results checked by another operator later, and altogether 90 editors are involved at its peak. This can create some potential problems.

Firstly, serial process can bring subjective bias from previous operators to the latter operators when the latter operators come to read and check the results assigned by the previous operators.

Secondly, quality indicators have been applied in order to ensure the quality of coding. However, by using serial process (e.g. RCV1) only partial coding data of operators have been investigated by its quality indicators. The data are mainly about the number of documents to which a given operator (editor) applied the final coding. Therefore, the picture to tell whether the operators have subjective bias towards certain labels is incomplete [22], [21], [23].

Furthermore it is hard for a typical manufacturing company to get enough personnel to build a corpus by using the serial process mentioned above. Thus it motivates us to establish a different coding process for one with only 4 to 8 operators, which is more realistic in manufacturing industry, bearing in mind that we want to maximize the output quality of manufacturing corpus.

We developed a parallel process to maximize the output coding quality from human operators. The idea is inspired by how information about customer requirements is collected in a Product Design and Development process (PDD). The authors believe operators communicating with each other can encourage thinking and agreement.

The process for the parallel coding process is visualized in Fig.2.

As indicated from the figure above, at least 4 coding operators are needed. A senior coding operator might be involved for the final verification step if needed. He can act as the final control for the quality of manual coded manufacturing corpus.

Another important issue to consider is how to control and to improve the coding quality. Basically, there are four phases available during the step of joint verification:

Phase 1: Right after all operators have finished the manual coding process, we can compute the most initial performance of operators and investigate their coding patterns.

Phase 2: If disagreement exists, operators have to sit down, exchange opinions and try to persuade each other. If there are changes, then run quality indicators again.

Phase 3: If for some documents after phase 2, the disagreements are still not resolved, then the labels are moved one level up, for example from the fourth level to the third level. However, this action is only valid for the disagreement about the labels in 4th level.

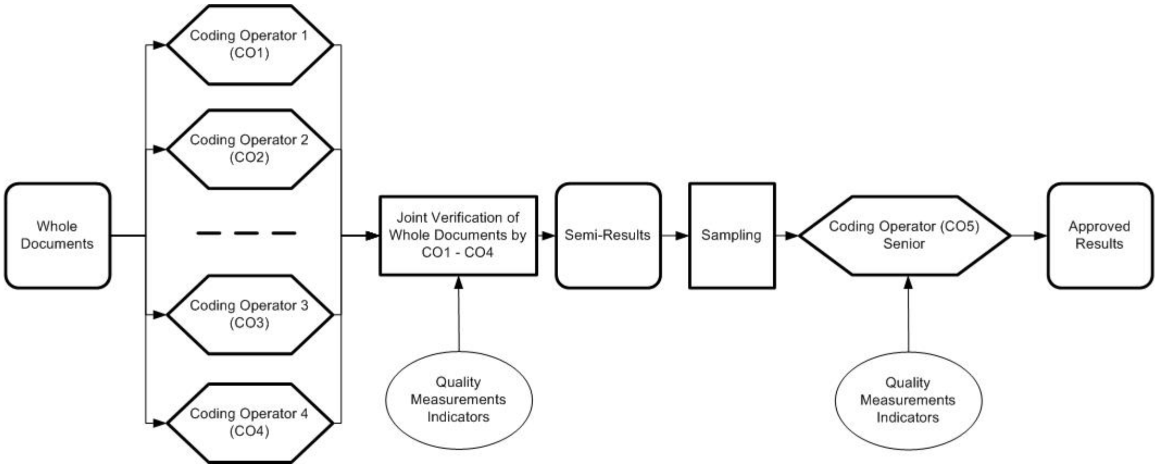


Fig. 2. The parallel coding process used by MCV1

In other words, moving labels one level up cannot go above 3rd level. This is to prevent the labels being too general for the documents.

Phase 4: For the rest of documents in which there are still disagreement among operators, simply assign all labels to the document. This is mainly to maximize the information coverage.

The whole step is visualized in Fig.3.

IV. CODING QUALITY INDICATORS

In order to ensure good coding quality of this manufacturing corpus, good and meaningful indicators for manual coding are a must. Here, a new and meaningful indicator has been created and one existing indicator used by industry has been modified to make it more suitable for this parallel coding process.

A. Coding Agreement Indicator (CAI)

The main purpose of Coding Agreement Indicators (*CAI*) is to indicate the coding agreement among different operators.

$$CAI = \frac{\sum_{i=1}^{n} \frac{L_i}{UL_i}}{n} \tag{1}$$

i donates to the number of documents in the corpus, which is equal to n.

Li donates to the number of identical labels assigned by every operator.

ULi donates to the unique labels assigned by all operators.

Here is an example to explain *CAI*. Coding operator 1 (CO1) and coding operator 2 (CO2) classify a corpus with only two documents. (In this case n is equal to 2.) The outcome is shown in table 1, A, B, C and D are four labels assigned by them.

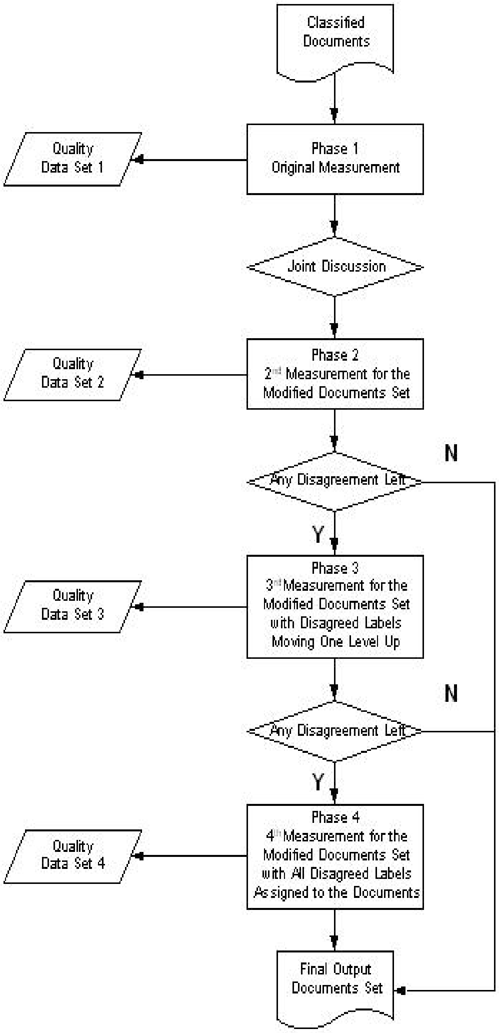

Classified Documents
Quality Data Set 1
Phase 1 Original Measurement
Joint Discussion
Quality Data Set 2
Phase 2 2nd Measurement for the Modified Documents Set
N
Any Disagreement Left
Y
Quality Data Set 3
Phase 3 3rd Measurement for the Modified Documents Set with Disagreed Labels Moving One Level Up
N
Any Disagreement Left
Y
Quality Data Set 4
Phase 4 4th Measurement for the Modified Documents Set with All Disagreed Labels Assigned to the Documents
Final Output Documents Set

Fig. 3. Illustration of joint verification

TABLE 1. AN EXAMPLE OF OUTCOME FOR CAI

	Doc1	Doc2
CO1	A, B	B, C
CO2	A, B, C	A, B, C, D

Then *CAI* is equal to:

$$CAI = \frac{\frac{2}{3} + \frac{2}{4}}{2} = \frac{1}{3} + \frac{1}{4} = \frac{7}{12} = 0.5833$$

The smaller the *CAI* is, the lower the uniformity of coding agreement. The lowest value is zero which means all operators completely disagree in their labels assigned to each documents. The idealized *CAI* is equal to one, which means operators are completely in agreement.

B. Coding Consistency Indicator (CCI)

There are two main functions of the Coding Consistency Indicator (CCI) - to indicate the main content of the corpus and to investigate whether the operators have subjective bias towards to different labels. It is similar to the idea of screening systematic bias in Reuters' indicator 2 for RCV1. Some simple statistics analysis will be put forward with the data. One example is shown in Fig.4.

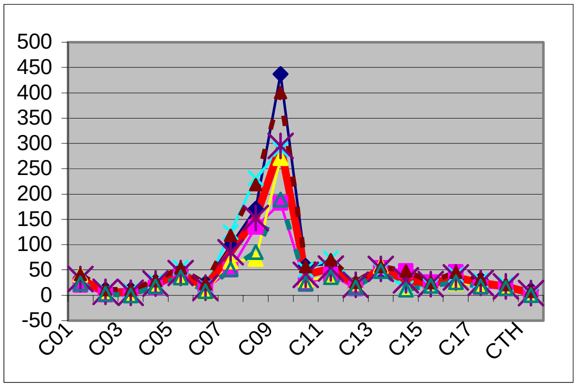


Fig. 4. An example chart of CCI output

The x axis represents the labels and y axis represents the frequency of a specific label assigned by operators.

V. CONCLUSION

The authors have reviewed the necessity to create a manufacturing centered corpus for the purpose of text mining, information retrieval and other knowledge discovery techniques to assist researchers in looking for information from research papers. A feasible and practical way to create such a corpus for manufacturing companies is presented. All the relevant concerns, including input source, coding labels, quality control are discussed and explained. As this is a very preliminary framework, much further research and application of text mining and information retrieval will be carried out.

ACKNOWLEDGMENT

The authors would like to thank the Society of Manufacturing Engineers (SME) for providing all the original engineering papers.

This research has also benefited greatly from the private communications with Tony Rose and Miles Whitehead from Reuters (Tony is a former employee of Reuters).

The authors also appreciate the suggestions and support by Ong Kok Koon from the library of National University of Singapore and Rolf Kwakkelaar from Elsevier.

This work also owes much to the suggestions and comments of text mining and information retrieval team of Design Technology Institute (DTI) under the National University of Singapore (NUS).

REFERENCES

[1] U. Fayyad and P. Stolorz, "Data mining and KDD: promise and challenges," Future Generation Computer Systems, vol. 13, pp. 99-115, 1997.

[2] M. J. Berry and G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support: John Wiley & Sons, Inc. New York, NY, USA, 1997.

[3] P. A. Flach, "On the state of the art in machine learning: a personal review," Artificial Intelligence, vol. 13, pp. 199-222, 2001.

[4] J. Han and C.-C. Chang, "Data mining for Web intelligence," Computer, IEEE, vol. 35, pp. 64-70, 2002.

[5] D. Braha, "Data Mining for Design and Manufacturing," Kluwer Academic Publisher, 2001.

[6] R. Menon, H. T. Loh, S. S. Keerthi, A. C. Brombacher, and C. Leong, "The Needs and Benefits of Applying Textual Knowledge Management within the Product Development Process," Quality and Reliability Engineering International, To be published.

[7] L. S. Larkey, "A Patent Search and Classification System," presented at Digital Libraries 99 - The Fourth ACM Conference on Digital Libraries, 1999.

[8] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys (CSUR), 2002.

[9] T. Joachims, Learning to Classify Text Using Support Vector Machines: Kluwer Academic Publishers, 2002.

[10] A. Visa, "Technology of Text Mining," presented at Proceedings of Machine Learning and Data Mining in Pattern Recognition, Second International Workshop, MLDM 2001, Leipzig, Germany, 2001.

[11] S. T. Dumais, D. D. Lewis, and F. Sebastiani, "Report on the Workshop on Operational Text Classification systems (OTC-02)," ACM SIGIR Forum 2002.

[12] W. J. Trybula, "Text mining and knowledge discernment: an exploratory investigation," The University of Texas at Austin, 2001.

[13] D. Sullivan, Document warehousing and text mining: John Wiley & Sons, 2001.

[14] D. D. Lewis and F. Sebastiani, "Report on the Workshop on Operational Text Classification systems (OTC-01)," 2001.

[15] W. J. Trybula and R. E. Wyllys, "An Evaluation of Text-Mining Tools as Applied to Selected Scientific and Engineering Literature," presented at Annual Meeting of the American Society for Information Science, Chicago, Illinois, 2000.

[16] W. B. Croft, "Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval," Kluwer Academic Publishers, 2000.

[17] M. A. Hearst, "Untangling Text Data Mining," presented at Proceedings of ACL'99, the 37th Annual Meeting of the Association for Computational Linguistics, invited paper, University of Maryland, 1999.

[18] D. Merkl, "Text Data Mining," in A Handbook of Natural Language Processing - Techniques and Applications for the Processing of Language as Text, R. Dale, H. Moisl, and H. Somers, Eds., 1st ed. New York: Marcel Dekker, 1998.

[19] M. Dixon, "An Overview of Document Mining Technology," Computer Based Learning Unit, University of Leeds 1997.

[20] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam, "OHSUMED: an interactive retrieval evaluation and new large test collection for research," presented at 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'94), 1994.

[21] D. D. Lewis and Y. Yang, "RCV1: a new benchmark collection for text categorization research," Forthcoming.

[22] T. Rose, M. Stevenson, and M. Whitehead, "The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources," presented at the third international conference on language resource and evaluation, 2002.

[23] T. Rose and M. Whitehead, Private communication: RCV1 building, 2003

[24] A. Salminen and F. W. Tompa, "Requirements for XML document database systems," presented at Proceedings of the 2001 ACM Symposium on Document engineering, 2001.

[25] J. A. Rydberg-Cox, A. Mahoney, and G. R. Crane, "Document quality indicators and corpus editions," presented at Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries, 2001.

[26] T. Joachims, "A Statistical Learning Model of Text Classification with Support Vector Machines," presented at Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States, 2001.

[27] J. Han and M. Kamber, Data mining: concepts and techniques: Morgan Kaufmann Publishers, San Francisco, 2001.

[28] L. Marquez, "Machine Learning and Natural Language Processing," Departament de Llenguatges i Sistemes Informatics (LSI), Technical University of Catalonia (UPC), Barcelona, Spain 2000.

[29] D. Pyle, Data preparation for data mining: Morgan Kaufmann Publishers, San Francisco, Calif., 1999.

[30] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing: The MIT Press, 1999.

[31] T. Mitchell, Machine Learning: The McGraw-Hill Companies, Inc., 1997.

[32] L. S. Larkey and W. B. Croft, "Automatic Assignment of ICD9 Codes to Discharge Summaries," CIIR Technical Report, IR-64, Dept. of Computer Science, University of Massachusetts, 1995.

[33] J. Broglio, J. P. Callan, and W. B. Croft, "INQUERY system overview," presented at Proceedings of the TIPSTER Text Program (Phase I), San Francisco, CA, 1994.