A Knowledge Mining Approach to Document Classification

Andreas Riel, Pawinee Boonyasopon

Keynote Paper

A Knowledge Mining Approach to Document Classification

Riel A.
G-SCOP Laboratory, Grenoble Institute of Technology, Grenoble, France

Boonyasopon P.
G-SCOP Laboratory, Grenoble Institute of Technology, Grenoble, France
KMRC, King Mongkut's University of Technology North Bangkok, Bangkok, Thaïland

Abstract
In the information age, there exists a huge amount of electronic data and information worldwide. A great challenge is how to exploit those information and knowledge resources and to turn them into useful knowledge available to concerned people, since the value of knowledge increases when people can share and capitalize on it. Thus approaches that can help researchers to benefit from existing hidden knowledge are needed. To this aim, tools that can extract relevant and useful knowledge are required. Text Mining is considered to be a key idea in order to help researchers to be able to extract explicit or implicit knowledge from unstructured data which is mainly in text written form. This paper presents the application of a text analysis tool that is used to mine knowledge by analyzing electronic text written documents in order to categorize different knowledge domains, to identify significant terms related to the documents and clusters of related documents, to identify unknown information hidden in these documents, find the relationships among them, and to visualize clusters of related topics and documents.

Keywords*: Knowledge Mining, Text Mining, Knowledge Sharing, Document Classification*

1 Introduction

In the highly competitive global market today, every organization is concerned with how to improve their performance to be more effective, more productive, and more innovative in order to be successful in the market. Knowledge is considered to be one of the key assets of companies in the 21st century [1]. Every company tries to exploit its knowledge to differentiate itself and to gain competitive advantages over competitors. With the huge amount of documents available both from internal sources such as company databases and external sources such as information on the internet, one of the major current problems is how companies can capitalize on these data and turn them into useful knowledge available to decision makers.

Documents are normally classified by title, abstract, and/or a part of the whole document. With limited text and an incomplete analysis, people and/or tools may not be able to understand the key idea of the whole document which could result in erroneous classifications. For example, the title of the document which was used for topic classification may not really relate to the whole idea of the document. Moreover, the idea of a document could relate to more than one topic and once the document is classified into one topic, it may not cover another topic which can lead to a loss of information. Tools and methodologies are needed that can help people extract embedded and deployable knowledge from the vast amount of unstructured information sources. Text Mining is a key element of such tools, and it is increasingly popular in both academia and industry.

This paper presents the application of a text mining and document classification tool to engineering research publications, which are considered to be rich sources of useful knowledge for researchers to keep eye on advances in research and development. The tool provides features to extract and visualize information, topic maps and cluster groups [2] according to the significance of documents. Researchers, students, and professors can benefit from the tool since it can facilitate obtaining relevant sources of interest in relevant domains. The paper

will also show how to use this tool to improve knowledge sharing among researchers.
The rest of the paper is organized in sections as follows: Section 2 presents a short overview of text mining concepts, as well as of the state of the art in related areas; section 3 presents the application of a particular tool to mining knowledge from research papers; section 4 describes tool limitations and suggestions for tool improvements; section 5 gives an outlook on future research; section 6 summarizes and concludes the paper.

2 State of the Art in related Areas

2.1 *Data, Information, and Knowledge*

Data without context have no meaning. They form a set of simple, discrete, objective facts, or events out of context with no relation to other things and once relations are understood, it is considered as information. Thus, data becomes information when their creator adds meaning, for example by contextualizing, condensing or categorizing them. Information has meaning and it is represented by relationships between data and possibly other information and once arranged and understood in meaningful patterns, is considered as knowledge. Broadly speaking, knowledge is information integrated into a specific context [3,4].
Davenport et al [3] define knowledge as a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. In organizations, it is often embedded not only in documents or repositories but also in organizational routines, processes, practices, and norms.

2.2 Mining from structured Data Sources

This section gives an overview of some key concepts used to mine information and knowledge from structured data sources. A more complete overview can be found e.g. in [5].

Knowledge Discovery in Databases (KDD) is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. It is the overall process of discovering useful knowledge from structured data sources [5].

Data Mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns or models over the data [5]. It is the exploration and analysis process of large quantities of structured data in order to discover new meaningful patterns, rules that are hidden in the data in order to yield knowledge and to lead to action.

Knowledge Mining is characterized as developing and integrating data analysis methods, collections of data using relevant prior knowledge that can be able to derive new knowledge which match the user's goals. The goal is an encoding of knowledge needs of the user which has to be guided by input criteria to define the type of knowledge useful for users [6].

2.3 *Mining from Unstructured Data Sources*

This section gives an overview of some key concepts used to mine information and knowledge from unstructured data sources, which is the core subject of this work. A more complete overview can be found e.g. in [7].

Text Mining is one of the techniques to extract textual data and to deliver valuable information that can lead to actionable knowledge intelligence. Text Mining is about sifting through vast collections of unstructured or semi-structured data beyond the reach of data mining tools. Text mining tracks information sources, links isolated concepts in distant documents, maps relationships between activities, and helps answer questions [7]. It can work with unstructured or semi-structured data sets such as full-text documents, HTML files, web pages, emails, and newsgroup postings instead of only structured data from databases, spreadsheets and XML files. There are numerous applications of text mining including information extraction, topic detection and tracking, summarization, categorization, clustering, concept linkage, information visualization, and question and answer. These applications are briefly introduced below.

- Information extraction: analyze unstructured or semi-structured text documents and identify key phrases and relationships within text. Then, transform unstructured information in documents or web pages into a structured database. This can be applied to different types of text such as research papers, emails, web pages, etc.;

- Topic detection and tracking: filter and present only documents relevant to a particular user profile;
- Summarization: text summarization reduces the length and the amount of detail of documents by retaining only its main points and overall meaning;
- Categorization: automatically classify documents into predefined categories and identify relationships;
- Clustering: plays an important role in grouping and representing concepts embedded in text documents. It is defined as a technique for grouping or partitioning similar data so that each partition or cluster contains groups of related documents;
- Concept linkage: connect related documents by identifying their shared concepts, helping users find information that they typically cannot find using traditional search methods;
- Information visualization: represent documents or information in graphical formats to facilitate browsing, viewing, and searching;
- Question and answering: search and extract the best answer to a given question.

2.3.1 Works related to Text Mining

Text mining technologies can be considered as a tool to help companies capitalize on their knowledge to gain competitive advantages over competitors. Text mining has become a popular topic for researchers and companies. It has found various application areas in the product development process, as well as in customer relationship management, patent analysis, and others.

Chaudhary et al [8] propose text mining for analyzing post project review which is a rich source of knowledge that contains good or bad project experience, important insights, and reports. If information from past projects can be extracted and analyzed effectively, the company can uncover patterns, associations, and trends which can help in improving overall knowledge reuse, exploitation, and learning. It can improve the quality and reduce time of future project as well.

Kin Nam Lau et al [9] illustrate the use of text mining in hotel industry. They use the text mining technique to develop competitive and strategic alliances by analyzing textual information of hotel databases or external data sources which are mainly in unstructured or semi-structured text documents. They extract meaningful patterns from those information sources, and then build predictive customer relationship models in order to predict new offers or improvements that should lead to improved customer satisfaction.

Han Tong Loh et al [10] focus on mining of textual databases to obtain and capitalize on valuable information for the product development process starting from planning, design, production, to service and support.

Menon et al [11] study how to analyze textual databases and extract information by using data mining to enable fast product development processes in order to have high quality information needed in the right moment and at the right location.

Yoon et al [12] propose an integrated process of text mining and network analysis in generating a patent network and conducting the quantitative analysis. Three major dimensions of analysis are measured which are the “importance”, “newness”, and “similarity” of each patent.

M. de Miranda Santo et al [13] study analyses on scientific papers of nanoscience and nanotechnology by text mining to map a given area and to identify trends. It shows an updated vision of the evolution in the nanoscience area and in the worldwide production of articles, including competitor countries and their interests in this field.

Delen et al [14] present the use of text mining to identify clusters and trends of related research topics from journals relating to the management information systems field by using the text mining process called IDEF0.

Mack et al [15] presents tools and methods to discover text-based knowledge by managing information, discover facts, relationships, and implications in biomedical literature that can be used to help solve biotechnical problems.

Drewes [16] presents three industrial applications of text mining. One application uses a classification approach to filter documents relevant for personal profiles from an underlying document collection. Another application combines cluster analysis with statistical trend analysis in order to detect emerging issues in manufacturing. The third application is a combination of static term indexing and dynamic singular value computation that is used to drive similarity search in a large document collection. However, all of these applications require a knowledgeable human to be part of the process. The

goal is not an automatic knowledge understanding but using text mining technology in order to enhance the productivity of existing business processes.

All the above mentioned applications make available explicit and implicit information contained in several different text documents and use this information to derive valuable knowledge by putting it in context of prior knowledge about facts, rules and relationships. According to the definition of Knowledge Mining given in [6] they are thus examples for the use of text mining as a means to capitalize on automatically found information in the form of new knowledge.

2.3.2 Text Mining Tools on the Market

On the market there are many text mining tools available that have been developed by researchers, companies, and universities. Some tools which can mainly work on the analysis of document collections are listed below. A more extensive overview of related tools can be found in [17].

- The cMap text mining from Canis [18] can map documents into clusters by SOM algorithm. The result is shown in three dimensions with the relationships and the degrees of similarity between clusters;
- Vizcontrols from Inxight [19] provides a new way to understand large amounts of information by showing visualization of data points, documents, or information objects;
- SemioMap from Semio [20] is a text mining tool that creates cluster maps from a set of documents from internet and intranet. It posts results on intranet so that people in the organization can share information. However, information from the internet cannot be imported and analyzed.

To work with a huge amount of electronic research documents, the problem to find just the most relevant papers is an important issue. Google Scholar [21] can provide the first step to get related documents from keyword search. Very often however, the search results do not correspond exactly to what is expected. Users may miss some documents that use vague or unclear keywords, but which actually relate to their works. The major target of this research is thus to make available automatically a maximum of key information about the content of single document as well as document collections.

3 Mining Knowledge from Research Papers

3.1 Introduction to CAT

For the research presented in this paper, the text mining tool named CAT (Content Analysis Toolkit) by InduTech [22] was used. Its major capabilities are information extraction, clustering, concept linkage, and information visualization. It can help users to exploit explicit and tacit knowledge which is hidden in unstructured electronic text documents.

CAT can extract key information from electronic text written documents. It helps users to find the topic clusters underlying a collection of documents analyzed. The tool can automatically analyze and categorize documents into different topics. So, from the result, users can get an idea of the content of the

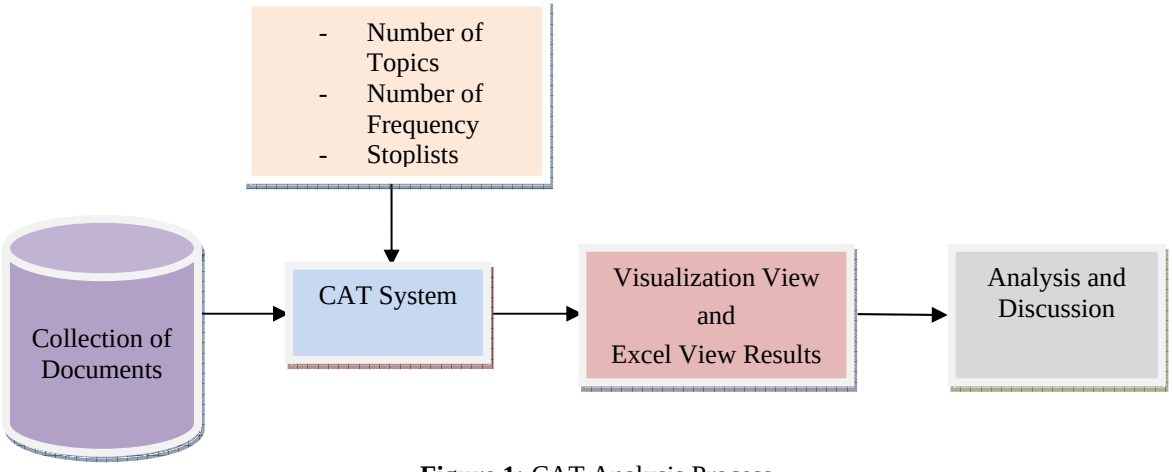


Figure 1: CAT Analysis Process

documents without actually reading them. Figure 1 gives an overview of the CAT analysis process.

From the pool of electronic documents analyzed, CAT will show the statistical analysis results in a dedicated visualization facility. The number of topics to be extracted needs to be specified by the user before starting the analysis. Furthermore, the threshold frequency to take into consideration for each word is required to be specified. For example, if a minimum frequency of 5 indicates that words which appear in any paper at least five times will be taken into account in the results. As a third parameter, the stop word list needs to be specified in order for the tool to be able to filter and exclude words that are not important or have little or no semantic value in the results of analysis.

The CAT text mining is able to

- automatically analyze documents;
- map each document to related topics;
- cluster documents based on their significance and relationships;
- identify relationships among each word, document, and cluster;
- determine relevant documents and related clusters;
- determine significant associated words in each document and cluster;
- show the three most significant words of each cluster that is labelled as a topic.

CAT represents the following results in the visualization view and on each sheet in Excel [23, 24]:

- Topics – it shows topics that have been discovered together with the words associated with them.
- Document Topics – it shows documents corresponding in each topic along with the value that show the strength of association between document and topic in focus.
- Topic Coverage – it shows how well each of the discovered topics are covered in the documents.
- Topic Similarities – it shows which topics are similar to one another.
- Topic Similarities Summary – it shows the three most similar topics of each topic.
- Document Similarities – it shows which documents are similar to one another so that closely related documents can be found easily.
- Document Similarities Summary – it shows the three most similar documents for each document.
- Vocabulary – it shows the list of all the words included in the topic model.

Detailed information about CAT and the associated research activities can be found in [23] and [24] respectively.

3.2 Application of CAT to Plan a Scientific Conference Schedule

In the context of the presented research, CAT was applied to the paper collection of the 42nd CIRP Manufacturing Systems conference in Grenoble in 2009. There were 73 accepted papers to be analyzed. It was decided to run analyses with 8, 10, and 12 topics in order to compare which analysis would best fit to the domains of the conference sessions. This choice corresponded to the number of different topics in the call for papers of this conference.

3.2.1 Conference Paper Analysis

The result obtained for 8 topics was already significant but considered not yet good enough. The tool does not sufficiently well classify the papers as the topics identified in the result seem to be too general. By contrast, the result obtained for 12 topics shows insignificant topics that are too specific. Hence, in this study, it was decided to put the focus on the results obtained for 10 topics. The analysis for 10 topics was run both using full papers and only their abstracts. The results of these two analyses along with the three most significant words are shown in Table 1.

By the analysis of the full text papers and just the abstracts, it was confirmed that the analysis of the full text of documents give more accurate analysis results of the actual content of papers than the analysis of the abstract parts only. The abstract of a paper is a short yet concise paragraph which is supposed to describe the overall key idea of the paper. However, authors may put words in the title, abstract, and keywords which are not really related to the whole idea of their works but rather to the relevant subjects that the authors would like their paper to be associated with. For this reason, the further study focuses on the analysis of full text papers.

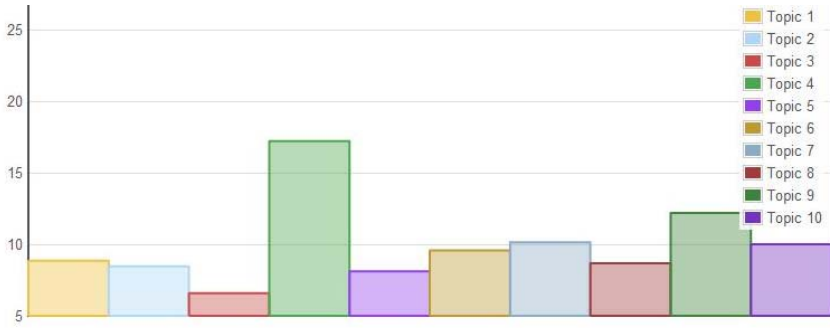
As shown in Figure 2, CAT can automatically categorize papers into topics. The top three words associated to each topic are shown.

Figure 3 shows the word clouds associated with selected topics. Word clouds are a representation of the frequency with which words appear in the context of this topic. The size and colour of words are an indication for their relevance to a specific topic. In the displayed example, the result of topic 1 is shown with the highest confidence. Word clouds help experts to get a good idea of what topics are about.

In the topic view of CAT, which is shown for topic 1 in Figure 4, the user can see for a selected topic all the relevant words, relevant documents, and the relationship to the other topics. Thus, this type of view allows the user to analyze the similarity of documents among one another, as well as their relevance to any of the topics. This analysis has been used to support to group papers into different sessions of the conference.

Table 1: Analysis results for full papers and abstracts

Group by	Full papers		Abstract only	
CAT	Topic 1	Part, machine, operations	Topic 1	Machining, processes, models
	Topic 2	Project, value, development	Topic 2	Engineering, process, grinding
	Topic 3	Flexibility, manufacturing, customer	Topic 3	Management, scheduling, problem
	Topic 4	Design, product, process	Topic 4	Systems, production, technology
	Topic 5	Service, customer, services	Topic 5	Service, keywords, university
	Topic 6	System, time , manufacturing	Topic 6	Tools, machine, tool
	Topic 7	Cutting, machining, process	Topic 7	Manufacturing, system, flexibility
	Topic 8	Tool, manufacturing, point	Topic 8	Production, factory, research
	Topic 9	Production, control, planning	Topic 9	Product, environmental development
	Topic 10	Product, environmental, production	Topic 10	Design, process, method
Expert	• **Cutting** (machining, process, tool) • **Design** (product, process, service) • **Environment** (product, production, design) • **Flexibility** (manufacturing, system, customer, time) • **Production & Planning** (control, system) • **Service** (customer, design) • **System** (manufacturing, time) • **Value** (development, cost, project) • **Part** (machine, operations) • **Tool** (manufacturing, point)		• **Design** (process, method) • **Environment** (product, development) • **Grinding** (engineering, process) • **Manufacturing** (system, flexibility) • **Scheduling** (management, problem) • **Service** (keywords, university) • **Systems** (production, technology) • **Matching** (processes, models) • **Production** (factory, research) • **Tools** (machine, system)	



Name	%
Topic 1 - part, machine, operations	8.86
Topic 2 - project, value, development	8.47
Topic 3 - flexibility, manufacturing, customer	6.6
Topic 4 - design, product, process	17.26
Topic 5 - service, customer, services	8.13
Topic 6 - system, time, manufacturing	9.59
Topic 7 - cutting, machining, process	10.16
Topic 8 - tool, manufacturing, point	8.69
Topic 9 - production, control, planning	12.23
Topic 10 - product, environmental, production	10.02

Figure 2: 10 topics analysis of full papers

3.2.2 Conference Schedule Creation

Although the tool can automatically identify topics and their relevance to papers, human knowledge is required to judge the quality and the usability of the results with respect to several criteria that cannot be taken into account by CAT. These criteria include constraints such as

- the balance of the number of accepted papers over all topics;
- the time constraints imposed by the overall timeframe of the conference;
- the number of keynote speeches;
- the number of papers per session;
- the overall session lengths, which ranged from 1.30 to 2.30 hours per session, as well as
- the potential absence of presenters.

For this reason, the creation of the actual conference program demanded expert intervention, which was however reduced to a minimum extent thanks to the consistent basis of the program provided by the CAT analysis. The latter gave the organizers a very good insight into the actually strong session topics, without having had to read the papers. It also revealed similarities and relationships among papers which would probably not have been detected during the normal human-only paper review and grouping procedure.

This also enabled organizers to establish and/or leverage links between researchers whose publication subjects were not obviously related.

Name	System Label	Confidence	Rating	Cloud
Topic 1	part, machine, operations			11 12 13 15 2006 38 45 agent agents algorithm algorithms approach assembled assembly assembly line assembly machine assigned autonomous balancing base block capabilities case change consists constraints cost created current cycle done es feasible features function generated grasp heuristic industrial job jobs journal line lines line balancing list machine machines machine agent machining manufacturing manufacturing system method methods min model operation operational operational research operations optimization order paper parallel part parts part agent part agents plan planning plans precedence problem problems procedure process process plan process planning product production references research schedule scheduling section sequence set single solution solutions step structure sub sub parts system time times tool transfer type

Figure 3: Word cloud view for Topic 1

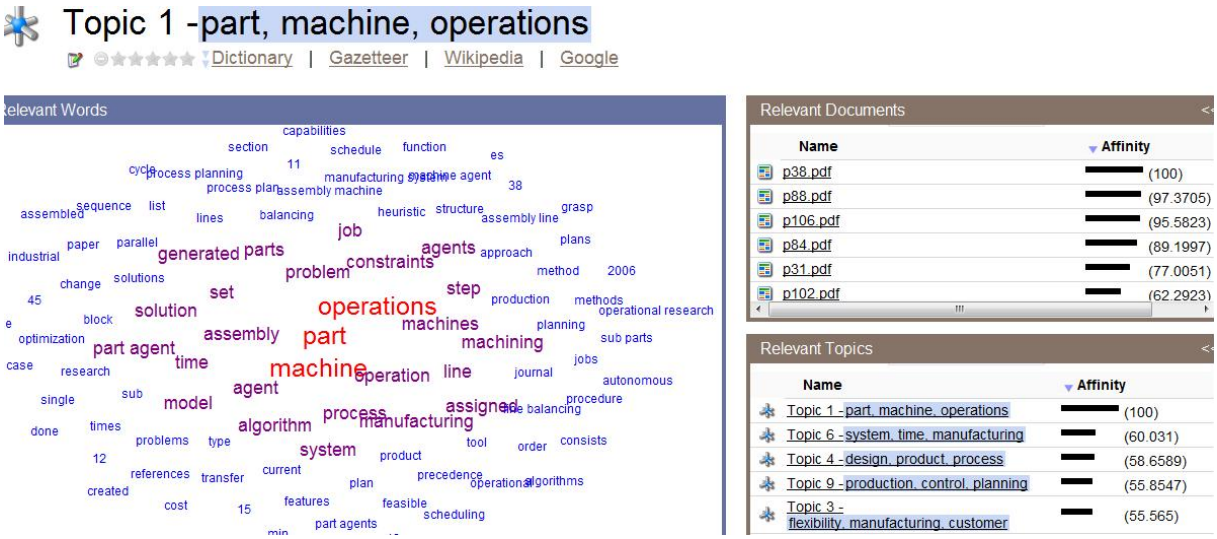


Figure 4: Documents-Topics relationships view for Topic 1

4 Tool Limitations and Improvement Suggestions

4.1 Problematic Issues of the Analysis Process

From the use of the content analysis toolkit for this study, some issues were identified to be especially problematic in terms of achieving a high-quality automatic identification of topics, and a good categorization of papers. They can be briefly summarized as follows:

- Papers with lots of formula, since formula cannot be recognized and interpreted by the tool.
- Papers with tools/methods descriptions in context of a more global topic (e.g., sustainability): if a paper focuses on tools being used or methods to reach a goal, it can be difficult to identify the core subject of the paper.
- Acronyms (e.g. PSS & IPS²), as they do not have any semantic meaning for the text mining tool.
- Compound words, as CAT cannot recognize compound words yet, and so it interprets each word individually. This can easily lead to the wrong results and misunderstandings.
- Words with multiple meanings: a typical case in point is the English word environment, which may signify natural space as well as a special technical infrastructure (e.g. for computing or production), depending on the context.
- "Insignificant" (technical) words appearing very frequently in the text body (e.g. point, tool, system, etc.), as they may "distract" the analysis result from the core subjects due to their frequent presence in a text in the explanation of those subjects.

In order to be able to overcome these difficulties by automated facilities, tools with significant semantic capabilities would be required. The authors are investigating such tools, which are coming up primarily in the domain of the Semantic Web [25].

4.2 Tool Limitations and Improvement Suggestions

Other than the general issues described in the previous section 4.1, the authors' studies with CAT also revealed some limitations of the tool itself in terms of its usability for the proposed purpose. The most important ones are listed below, along with relevant suggestions for improvement.

1. CAT does not use "stemming" techniques [26] in order to distinguish between similar words with the same root (e.g. singular/plural or gender of nouns).
2. Expert knowledge is required in order to set the number of expected topics to configure the analysis.
3. CAT has been conceived to be applied to a large number of documents. However, there is no specific rule to determine the minimum number of documents required to obtain a good and reliable analysis result.

4. As the core algorithm of CAT is based on a probabilistic model, the results of several analyses of a given document collection may differ more or less significantly. This can present a problem in terms of the repeatability as well as of the assessment of the quality and the reliability of a specific analysis.
5. CAT cannot do an incremental analysis, i.e., a content analysis of one or more new documents on the basis of an existing analysis of a document collection. Otherwise stated, it is impossible to determine the relevance of a new document with respect to an existing topic structure.

The limitations and suggestions cited above serve as appreciated inputs to the CAT management and development team. Some of these limitations render CAT in its current form inadequate for certain types of further studies in the area of Knowledge Mining from research publications.

5 Outlook

The target of the authors' research is to establish a Knowledge Mining based system that supports researchers in increasing their networks and in leveraging their citations. Based on the automated analysis of publications in terms of the subjects they treat, the references they make, and the relationships they have among one another and with other publications, such a system will be able to make available conveniently knowledge that provides answers to questions such as:

- Who is an expert in a specific domain?
- Which experts publish together?
- Who are the experts in a specific field in a specific country, organization, etc.?
- Which industrial partners are associated to the work of specific researchers?
- Which publications relate to or are similar to a publication newly submitted to the system?
- Etc.

Such a system would not only significantly leverage knowledge sharing and capitalization among researchers, but it would also help increase the reputation of researchers in terms of helping them to increase the citation indices of their publications. Currently the authors are preparing an extensive study based on the vast corpus of the publications of the CIRP [27] community with the above major targets in mind.

6 Conclusions

In view of the rapidly increasing amount of unstructured data available in electronic form, the need to help people access and extract useful knowledge, uncover hidden knowledge, and capitalize on knowledge is increasingly important. This paper presented text mining as a fundamental method to extract new explicit and implicit knowledge hidden in a large collection of text documents. It discusses the results obtained by applying a specific a text mining tool named CAT to obtain an "objective" repartition of research publications in the engineering domain into conference sessions with the objective to support conference organizers in creating the conference schedule. CAT can automatically analyze text documents, cluster documents into topics, determine the relevance of documents to topics, find relationships among documents and topics, and discover patterns and trends as hidden knowledge. It has, however, certain limitations that need to be partly or completely eliminated in order to apply it reliably and extensively for some of the envisaged Knowledge Mining applications. The authors are extending their research studies on other types of documents and on other applications in the strong belief in the capability of Knowledge Mining to leverage knowledge sharing and capitalization in organizations and communities.

References

[1] Bernard A, Tichkiewitch. S. (2008). Methods and Tools For Effective Knowledge Life-cycle-management, Springer, ISBN 978-3-540-78430-2.

[2] Riel A., Tichkiewitch S., Uys W., Du Preez N. (2008). Mining Knowledge in the Digital Enterprise. International Conference on Digital Enterprise Technology, Nantes, France, CD-ROM Proceedings.

[3] Davenport T.H., Prusak L. (1998). Working Knowledge How Organization Manage What They Know, Harvard Business School Press.

[4] Alavi M., Leidner D. E. (2001). "Review: Knowledge management and knowledge management systems: conceptual foundations and research issues." MIS Quarterly 25(1): 107-136.

[5] Fayyad U., Piatetsky.-S. G., Smyth P., (1996). "From Data Mining to Knowledge Discovery:

An Overview", in Advances in Knowledge Discovery and Data Mining. Fayyad U., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R. (Eds.), Cambridge, Mass.: MIT Press/AAAI Press.

[6] Michalski R.S. (2003). Knowledge Mining: A proposed New Direction. Sanken Symposium on Data Mining and Semantic Web. Osaka University, Japan.

[7] Fan W.L., Rich S., Zhang Z. (2006). "Tapping the Power of Text Mining." Communications of the ACM 49(9): 77-82.

[8] Choudhary A.K., Oluikpe P.I., Harding J.A., Carrillo P.M. (2009). "The needs and benefits of Text Mining applications on Post-Project Reviews." Computers in Industry 2125: 13.

[9] Lau K-N, Lee K.-H., Ho Y. (2005). "Text Mining for the Hotel Industry." Cornell Hotel and Restaurant Administration Quarterly 46(3): 344-362.

[10] Loh H.T., Menon R., Leong C.K. (2002). "Mining of Text in the Product Development Process." Innovation in Manufacturing Systems and Technology (IMST), DSpace@MIT Online Library, http://hdl.handle.net/1721.1/4033, last accessed on 09/11/2009.

[11] Menon R., Tong L. H., Sathiyakeerthi S. (2005). "Analyzing textual databases using data mining to enable fast product development processes." Reliability Engineering & System Safety 88: 171-180.

[12] Yoon B., Park Y. (2004). "A text-mining-basd patent network: Analytical tool for high-technology trend." The journal of High Technology Management Research 15: 37-50.

[13] De Miranda Santo M., C. G. M., Maria dos Santos D., Filho L.F. (2006). "Text mining as a valuable tool in foresight exerciss: A study on nanotechnology." Technological Forecating and Social Change 73: 1013-1027.

[14] Delen D., Crossland M.D. (2008). "Seeding the survey and analysis of research literature with text mining." Expert Systems with Applications 34: 1707-1720.

[15] Mack R., Hehenberger M. (2002). Text-based knowledge discovery: search and mining of life-sciences documents. Drugs Discovery Today (DDT). 7: 589-598.

[16] Drewes, B. (2005). "Some Industrial Applications of Text Mining" Studies in Fuzziness and Soft Computing 185: 233-238.

[17] Tan A.H., (1999), Text Mining: The state of art and the challenges, In proceedings, PAKDD'99 Workshop on Knowledge discovery from Advanced Databases (KDAD'99), Beijing, pp. 71-76, April 1999.

[18] http://www.canis.uiuc.edu/projects/interspace/technical/canis-report-0006.html, last accessed on 09/10/2009.

[19] http://www.dcc.uchile.cl/~rbaeza/cursos/visual/ix/index.html, last accessed on 09/10/2009.

[20] http://www.semio.com/, last accessed on 09/10/2009.

[21] http://scholar.google.com/, last accessed on 15/10/2009.

[22] http://www.indutech.co.za/, last accessed on 29/10/2009.

[23] http://www.analyzecontent.com/, last accessed on 29/10/2009.

[24] Indutech, Content Analysis Toolkit User Guide version 1, 27 March 2009.

[25] http://en.wikipedia.org/wiki/Semantic_Web, last accessed on 11/11/2009.

[26] http://en.wikipedia.org/wiki/Stemming, last accessed on 11/11/2009.

[27] http://www.cirp.net/.