Knowledge Mining in Manufacturing and Management

Boonyasopon P.

KMRC, Knowledge Management Research Center, KMUTNB, Bangkok, Thailand
G-SCOP Laboratory, Grenoble Institute of Technology, Grenoble, France

Riel A.

G-SCOP Laboratory and EMIRAcle AISBL, Grenoble Institute of Technology, Grenoble, France

Abstract

This paper presents intermediate results of a research study which investigates the potential of the use of text mining based approaches to capitalize on knowledge contained in research publications in the product development and manufacturing domain. The ultimate research target is to conceive a system which is able to motivate and facilitate researchers to collaborate, to help them get their publications cited, to improve their bibliographies, and thus to better capitalize on their own and related research. The capabilities of such a system shall go far beyond currently available full-text search based approaches. Departing from results obtained by the application of a particular text mining tool based on Latent Dirichlet Allocation (LDA) on a vast set of manufacturing research publications, the paper investigates alternative algorithmic approaches which promise to get rid of the shortcomings of the LDA-based implementation. It gives an outlook on further research steps that shall lead to an answer which approach is suitable for the application of knowledge mining in the product development and manufacturing research domain.

Keywords*: Knowledge Management, Knowledge Sharing, Information Technology, Text Mining, Topic Identification*

1 Introduction

Knowledge is considered as a key to success in any organization. Knowledge management is the process which helps capture an organization's knowledge residing in internal information sources such as people, databases and exchanged documents, as well as external sources, most notably the internet. The fast proliferation quantity of electronic text documents, however, results in information overload. Therefore, in order to capitalize on existing knowledge, people have to exploit knowledge by turning explicit and implicit knowledge to valuable and sustainable knowledge.

One way to support this is document classification. Documents are typically classified by title, keywords, abstract, or other specific parts of the whole document. This approach is very limited and fails to take into account the actual content [1]. Moreover, the process of categorization is usually done by manual work which is a time-consuming and error-prone task. Methodologies and tools are required that help automate this task, working on the complete content of documents. Text mining based approaches can provide an answer to helping people replace or supplement human readers and classifiers with automatic systems.

The target of this research is to propose and conceive a knowledge mining system based on text mining that facilitates and supports researchers in finding relevant and useful papers based on their actual interests. This paper presents intermediate results of the application of a particular text mining tool for the classification of manufacturing research papers based on automatic topic identification. The tool can automatically extract and cluster the papers in to different topics and provide results in visualize and statistic based view. Departing from the analysis of these results, it investigates the algorithm underlying this tool in order to find out the reasons for certain

limitations of the tool with respect to the research target.

The paper is divided into five sections as follows: Section 2 introduces the concept of text mining. Section 3 describes the text mining tool that is used in this research. Section 4 investigates the main algorithm underlying this tool and its limitations with respect the envisaged application. Motivated by these limitations, it also looks at alternative algorithmic approaches. Finally, section 5 summarizes and concludes this paper, and section 6 gives an outlook on the next steps in the authors' research.

2 A brief introduction to Text Mining

Text mining is the process of extracting interesting, new, non-trivial, undetected, and unstructured knowledge hidden from text documents. The major functionalities of text mining consist of [2]:

- Information extraction: the task to analyze unstructured text documents and identify key phrases and relationships within text by a process called pattern matching to provide the user with meaningful information.
- Topic detection and tracking: prediction and presentation of documents relevant to the interest of the user based on user profiles or documents viewed.
- Summarization: reduction of the length and detail of a document while retaining its main points and overall meaning.
- Categorization: classification of documents into pre-defined categories and identification of relationships based on words appearing in the document.
- Clustering: grouping of similar documents and representing concepts embedded in text document without having pre-defined categories. 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: connection of related documents by identification of shared concepts, to enable users to find information that they perhaps would not have found using traditional search methods.
- Information Visualization: Putting large textual sources in a visual map to facilitate understanding of user while exploring the results.
- Questioning and answering: search and find the best answers to a given query.

These functionalities can enable the user to better understand information and to discover useful information hidden in huge unstructured collections of textual documents. They also help analyze information sources effectively, and therefore provide knowledge to researchers in order to support their research and publication work, notably in terms of finding related research publications, and to increase their overall productivity and insight.

3 Mining Knowledge from Research Papers

3.1 *Problem definition*

Research literature is a highly important source of knowledge giving access to novelties, advances, inventions and innovations, developments, trends, and ideas. With the growing amount of papers published on the World Wide Web in electronic form, it is increasingly difficult to find actually relevant documents. Researchers are often not able to keep track of all new relevant documents from their domains, or to find relationships among documents. Moreover, documents are normally classified and indexed manually and subjectively. Researchers often categorize documents by just reading the title, the keywords, and the abstract. However, this highly limited investigation it is not sufficient to understand the actual key ideas of the complete publication. It not only hinders the correct, accurate classification, but it may also mislead subsequent document searches.

3.2 *Overview of CAT*

The text mining tool CAT (Content Analysis Toolkit [3]) by Indutech Ltd in South Africa [4] has been used as a point of departure in this research. The major capabilities of CAT are information extraction, clustering, concept linkage, and information visualization. It can thus help users exploit explicit and tacit knowledge which is hidden in unstructured electronic text documents.

CAT can extract key information from electronic text documents. Users can easily find the topic clusters underlying a collection of documents analyzed. The tool can automatically analyze and categorize documents into different topics. Users can get an idea of the content of the documents without actually spending time reading them. Relevant topics and related documents can easily be identified.

Figure 1 shows the main process of a text corpus analysis using CAT. The user has to specify the files

to be analyzed, to indicate the number of expected topics to be extracted from the pool of documents, and to define the number of times a word has to appear in order to be considered in the analysis (the word frequency). A so-called "stop-list" of words specifies words that have little or no semantic value, and are thus to be excluded from the analysis. Based on these inputs, CAT is able to automatically analyze all the documents provided. At the end of this process, CAT comes up with a results visualization, which essentially allows for the following operations:

- Visualization of word clouds associated with identified topics. Each topic is specified by the three most significant words associated with it.
- Mapping each document to related topics.
- Clustering documents based on their similarities.
- Visualization of relationships among documents and topics.

CAT can automatically generate topics based on word frequencies. The results can thus reveal that a specific document in the corpus relates to one or more of the discovered topics. However, human interpretation is needed in order to decide whether the results are useful in terms of both the identified topics, and the assignment of documents to topics. Details about CAT, its functions and associated activities can be found in [3] and [5].

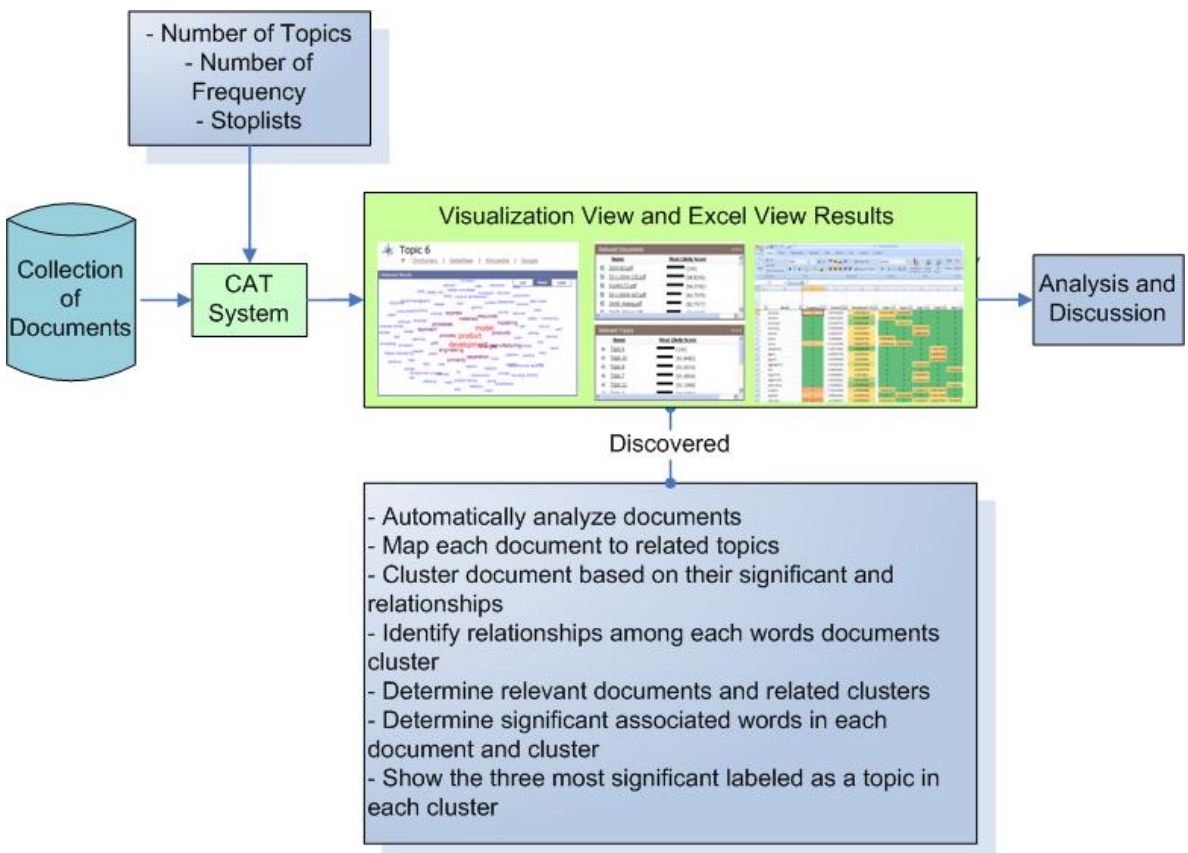

Figure 1: Document corpus analysis with CAT

3.3 *Case study in Manufacturing Research*

As is pointed out in [6], CAT has not been conceived for the specific purpose of performing an unsupervised reliable assignment of documents to automatically identified topics. Its main purpose is to give the user insight into the main subjects of very huge document corpora, without the explicit need of high accuracy and exact repeatability. In order to evaluate the performance of the CAT text mining tool for the envisaged research knowledge mining system, CAT was applied to different large and small sets of manufacturing research papers. The principal results of one of these studies have been published in [1]. In that publication, the targets of the envisaged knowledge mining support system for researchers are also described in greater detail. Other publications on case study results are in work. The purpose of this paper is to summarize the main findings from all of those studies with respect to the target system, and to investigate the principal properties of the algorithm implemented in CAT in order to understand the reasons for major limitations.

3.4 *Limitations of CAT*

The studies confirmed that CAT has indeed a lot of functionalities that can help researchers retrieve explicit or tacit knowledge from collections of research papers. However, in the application for this specific purpose, CAT has certain limitations. The limitations considered as most important are the following:

1. CAT is based on a probabilistic model, which leads to the fact that the results of several analyses of a given document collection may differ more or less significantly from one another. This can present a serious problem in terms of the repeatability as well as of the assessment of the quality and the reliability of a specific analysis.
2. CAT does not support incremental analysis and document fold-in operations. Therefore, whenever one or several new documents are added to the corpus, a complete analysis of the updated corpus has to be done. Apart from the fact that computation times for analyses are in the order of several hours or days for huge corpora, this limitation make it impossible to determine the relevance of a new document with respect to an existing corpus and topic structure.
3. CAT does not support a fully automatic and unsupervised process. A considerable amount of expert knowledge is required in order to set initial parameters such as the number of expected topic to configure the analysis.
4. CAT has been conceived for analyses of very huge document corpora. However, there are no specific rules that allow determining the minimum number of documents which should lead to optimal and reliable results.
5. From a semantic point of view, CAT does not use 'stemming' techniques which provide a way of treating different declinations, singular and plural, prefixes etc. of a specific word as one single word. Also, compound words are not recognized by CAT as such.

Number one of the issues listed above, i.e., the randomness of results from different runs represents the principal limitation of the use of CAT for the envisaged application. Consequently, the origin of randomness of the results obtained by CAT has to be investigated, and a potential alternative algorithmic approach has to be found. An overview of the authors' findings so far will be presented in the following section.

4 Investigation of Knowledge Mining Algorithms

4.1 *Essential concepts implemented in CAT*

CAT is essentially based on statistical topic modelling, which is used to distil and organize the content of text documents. Topic models are a form of unsupervised learning since there is no need for humans to give input or classify in order to learn the latent topics from the document corpus. Topic modelling is well suited to solve the problem of synonymy (i.e., multiple words with similar meaning) and polysemy (i.e., one word with multiple meanings).

The second principal concept implemented in CAT is clustering. Clustering is a major functionality applied for the purpose of the analysis of unstructured information coming from different sources in order to discover hidden topics. The core algorithm of CAT is based on the Latent Dirichlet Allocation (LDA) approach. According to [6] this algorithm has been selected in order to fulfil the clustering requirement because of its simplicity and its ability to formulate topics as semantic representations of the contents from a set of documents.

In the following subsection a very brief overview of LDA, with the target to pinpoint the source of the randomness of results of CAT analyses is presented.

4.2 *Latent Dirichlet Allocation (LDA)*

LDA is an unsupervised learning algorithm that discovers and extracts the underlying semantic topics structure from discrete data such as text corpora. It uses a generative probabilistic model which postulates a latent structure consisting of a set of topics. Each document is produced by choosing a distribution over topics, and each word is generated at random from a topic chosen by using this distribution [7, 8]. The LDA model assumes that the words in a document are generated by a mixture of topics, and these topics are infinitely exchangeable within a document. By labelling each word with a topic, it allows representation of a document in the form of its semantic topic content rather than the words or vocabulary [8].

The LDA model is a full probabilistic generative model that can capture human understandable semantic topics, which are represented as distributions over vocabulary. By representing a document in the topic space instead of in the vocabulary space, the LDA model effectively reduces the dimension of the texts while maintaining the semantic content of the document.

The output of the LDA analysis for a given dataset is a list of hidden topics each consisting of numerous terms ranked by relevance. The underlying idea of LDA based feature selection framework is that a good term should be highly ranked in only a few topics to be more discriminative for classification. The topics are used to illustrate the relationships between different scientific disciplines, assessing trends and hot topics by analyzing topic dynamics and using the assignments of words to topics to highlight the semantic content of documents.

On a high level, the generation of a document corpus in LDA is modelled as a three step process. The first step entails sampling a distribution over topics from a Dirichlet distribution for each document. Second, a single topic is selected from this distribution for each word in the document. The last step involves sampling each word from a multinomial distribution over words corresponding to the sampled topic.

The randomness in this process is rooted in the statistical inference algorithm that is used in order to compute the posterior distribution of the hidden variables given in a document [8]. This distribution is intractable to compute for exact inference. Markov Chain Monte Carlo (MCMC) was chosen and applied as a way to guide a random walk through parameter space of the model to numerically estimate the posterior probability of the parameters. MCMC requires little memory and is competitive in speed and performance compared to other inference algorithms. MCMC integration draws samples from the required distribution and then forms sample averages to approximate expectations.

In order to get rid of the randomness of this approach, the exact inference by complete enumeration needed to be performed. It means that each point in the associated probability space needs to be evaluated instead of using random walk by means of MCMC technique. Therefore, the number of calculation can be estimated by:

$$\text{Number of evaluations} = k^M,$$

where k is the number of topics and M is the number of words in all documents in the corpus. With M appearing in the exponent of this equation, this operation is of exponential complexity.

The illustration of this is given by the following example: If taking into consideration k=10 topics with the total of 100 documents that have 1000 words in each, then

$$K^M = 10^{100\,000}$$

which is infeasible in terms of calculation time even if powerful processors and parallel computing were used.

By offering the possibility to the user to manually specify the seed number for the randomization algorithm calculation, the random results could be avoided at the cost of the quality and the reliability of a particular result. Alternatively, running CAT several times on the same corpus, and apply an algorithm to the consolidation of all the results achieved to one unique result could also be possible. This, however, would imply a significant increase of computation effort, as well as the implementation of a sophisticated consolidation algorithm.

The extensive study in [6] shows that other probabilistic approaches to topic modelling are of similar complexity, and thus require randomized algorithms to perform the selection of potential

solutions. The study in [6] reveals that Latent Semantic Indexing (LSI) is the only non-probabilistic modelling technique that has been used for topic modelling. For this reason, this approach is investigated further in the following subsection.

4.3 *Latent Semantic Indexing (LSI)*

Latent Semantic Indexing (LSI) [10] is a well known information retrieval algorithm, which has been applied to a wide variety of learning tasks, such as search and retrieval, classification and filtering [11]. It is an approach based on matrix algebra and Singular Value Decomposition (SVD) that focuses on scalability and performance.

In order to implement LSI, a matrix of terms by documents must be constructed. The elements of the term-document matrix are the occurrences of each word which appears in a particular document [12]. A term-document matrix is an M x N matrix; the rows represent the M words found in the document set and the columns represent each of the N documents.

LSI can overcome the deficiency of lexical matching (vocabulary mismatch and match query to the terms in documents) because it uses a statistical technique to create a semantic analysis to derive conceptual indices instead of individual words for retrieval [13]. Its key feature is the ability to extract conceptual content of a body of text by establishing associations between those terms that occur in similar contexts by projects queries and documents into a space with latent semantic dimensions [14]. The words that appear in similar contexts tend to have similar meaning. LSI finds and fits a useful model of the relationships between terms and documents. It uses a matrix of observed occurrences of terms in documents to estimate parameters of that model with the resulting model.

A page related to concept search can be considered relevant to a particular keyword even if it does not contain that keyword considered relevant for the search criteria. Therefore, LSI can overcome the problem of synonymy and polysemy (see section 4.1) [10]. LSI is also a method for dimensionality reduction of the term-document matrix. The benefits of dimensionality reduction are to improve the interpretability of data, to reduce the time and storage requirement, to speed up the learning process, and to improve classification accuracy since it can prevent over fitting by eliminating the useless terms. It can choose the mapping that is optimal in the sense that it minimizes the distance. It aims to discover the most representative feature rather than the most discriminative feature for text representation.

In principle, LSI works in the following way [11]:

- Documents and terms are placed in a multidimensional vector space;
- each dimension in that space corresponds to a concept existing in the collection;
- thus underlying topics of the document is encoded in a vector;
- common related terms in a document and queries will pull document and query vector close to each other.

Although the memory and computation power requirements of LSI are high (in the order of M x N), they are *not* exponential as they are in the LDA case, and thus feasible if sufficient computation power and memory are provided. The core algorithm, SVD, has been very well researched and used over years, and numerous libraries are available that provide implementations, also for massively parallel computing environments. A very extensive overview of different kinds of implementations is available in [15]. LSI is thus a highly interesting alternative to LDA and other probabilistic modelling approaches. Due to the lack of access to a suitable LSI implementation and to a parallel computing infrastructure, the authors have not been able to study the performance of LSI with respect to the target requirements at this point of their research. It was thus decided to look for a completely different approach, which would completely avoid the complex step of topic modelling from the input document corpus.

4.4 *Deploying external Encyclopaedic Knowledge*

LDA, LSI and related approaches perform the document corpus analysis in two main steps:

1. Build a knowledge model of words contained in the document corpus.
2. Identify topics based on the model built.

All the knowledge available for the essential topic identification process is thus derived from the document corpus, which is both a limitation to semantic performance and a computationally complex task. The question is if it was possible to replace this step totally by capitalizing on some kind of existing body of semantic knowledge which grows independently of the document corpora submitted to the analysis. Ideally, this knowledge body would be

available for different languages. In this context the idea came up that the required external knowledge body essentially represents encyclopedic knowledge. The authors' subsequent research revealed that there is in fact a research community which has succeeded in using the digital encyclopedia Wikipedia exactly for this purpose. Particularly interesting and relevant contributions from this community can be found in [16], [17], and [18].

Basically, the Wikipedia-based approaches use a tool called Wikify, which is an unsupervised system to automatically identify the important encyclopaedic concepts in an input text that are relevant to the input document, and to link them to Wikipedia concepts. Otherwise stated, Wikify finds all semantically important words (including compound words) in the input text, and links them to related articles in Wikipedia. Currently its main application is the semantic annotation of webpages, however the fact that it makes available practically the whole semantic intelligence underneath Wikipedia, opens up a wide and yet largely unexploited field of applications.

Using Wikify for knowledge mining applications conceptually has the potential of completely replacing the step of building a semantic model of words, as it uses the semantic model of Wikipedia. This implies that the completeness of this model with respect to the vocabulary of the document corpus under investigation is in direct relation with the content of the digital encyclopaedia. Thus one would expect that the analysis of research documents with this approach could be problematic, as terms of cutting-edge research may not yet be explained in Wikipedia. It will be the next challenge of this research to investigate this assumption in the manufacturing and modern product development domain. In any case, this point of potential weakness is likely to be of minor importance, as the speed of growth of Wikipedia is unequalled. Moreover, for the purpose of identifying semantically important words in a text, the quality of the articles corresponding to the identified words is not at all an issue. This is also essential, as often in Wikipedia, words are added without an explaining article but instead with a call for an article. Other articles exist but have not yet been reviewed by experts.

It should also be mentioned at this point that the use of Wikify would at the same time provide a solution to issues that are highly problematic in LDA and related probabilistic semantic modeling approaches, such as compound words, polysemy, synonymy, and multi-language. Also, there is no need for stop-lists, which are language-specific and can be incomplete and outdated.

In terms of the second step, the identification of topics, a very interesting approach has been published in [19]. They present an unsupervised method for topic identification based on a biased graph centrality algorithm applied to a large knowledge graph built from Wikipedia. The relevant topics that may not even be mentioned in the document corpus can be obtained from external knowledge. Moreover, the topics are not known or predefined before. The whole process consists of the two following main steps:

1. Build a knowledge graph of encyclopedic concepts based on Wikipedia so that it can be efficiently used for topic identification of new documents.
2. Identify the important encyclopedic concepts in the text and create links between the content of the document and external encyclopedic graph and run a biased graph centrality algorithm on the entire graph so that categories are ranked based on their relevance to input document.

A limitation of this approach could be that only existing Wikipedia categories can be proposed as topics by the algorithm.

In [19] Coursey et al. present results of analyses of Wikipedia articles that are highly promising in terms of performance and quality of results. The authors' current target is to verify if the performance is equally well if the tool suite is applied to research documents of a specific domain.

5 Summary and Conclusions

Based on the application of the corpus analysis toolkit CAT from Indutech for the purpose of picking and relating research papers in the manufacturing domain, some problematic tool characteristics were identified. The issue of apparently random variations in corpus analysis results, which severely limit the usability of CAT for the targeted unsupervised research paper classification system, was investigated. It was found that this problem is due to a Markov Chain Monte Carlo based inference algorithm that is used for complexity reduction in the Latent Dirichlet Allocation (LDA) approach implemented in CAT. As LDA without this random element would be intractable, alternative non-probabilistic algorithms have been studied.

Another well-established approach in topic identification is Latent Semantic Indexing (LSI), whose main characteristic is that it is based on a non-probabilistic approach to semantic modelling. LSI is tractable without random elements, but it requires considerable computing power.

Very recent approaches that totally avoid the step of semantic modelling on the basis of the input document corpus have also been studied. All the works found use the digital on-line encyclopaedia Wikipedia as the source of external semantic knowledge, which is subject to an unequalled speed of growth and completeness. It seems that these approaches have many advantages over the more traditional LDA, LSI and related algorithms. Due to the fact that they have come up only very recently, few studies on performance and quality of results are available. The authors are very interested in contributing a study in the domain of research in manufacturing and modern product development. This contribution would at the same time help them advance in the decision for a topic identification approach suitable for the knowledge mining system for researchers that is the ultimate target of this research.

The major past and future steps in the authors' research that have been addressed in this paper are summarized in a flowchart diagram presented in Figure 2. The requirements specification for a new Knowledge Mining system for knowledge sharing in research has been established at the beginning, and forms the basis of roadmapping and evaluations all along the research process.

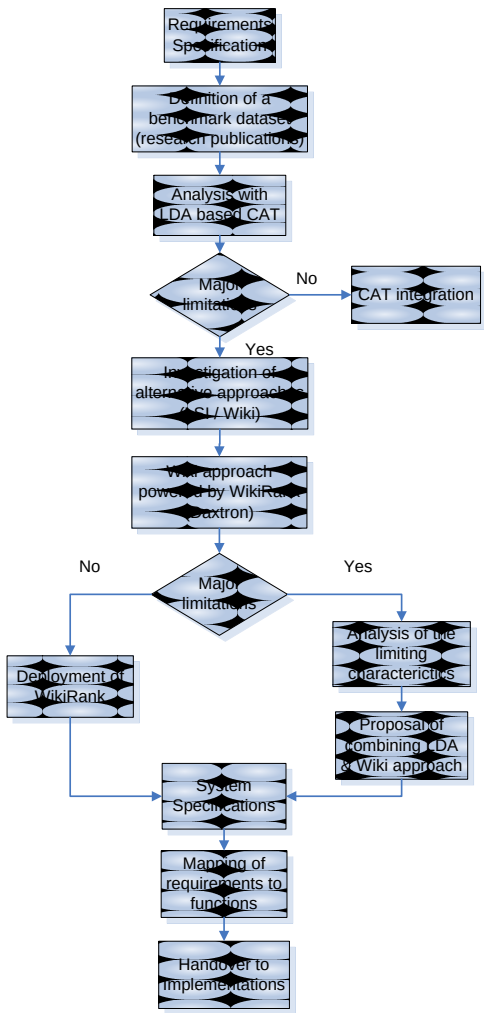

Figure 2: Research process

A set of research publications in the manufacturing and product development domain was defined to serve as benchmark for all studies. These documents were first analysed with the text mining tool CAT, which is based on Latent Dirichlet Allocation. Due to some intrinsic limitations of CAT and LDA with respect to the requirements, it was decided to investigate alternative approaches to text mining and topic identification. Most importantly, the necessity of a probabilistic algorithm to identify topics should be circumvented. These considerations resulted in the decision to continue the research using a completely different approach which is based entirely on external encyclopaedic knowledge contained in Wikipedia and a dedicated tool suite called "WikiRank" kindly made available to the authors by Daxtron lab, a spin-off of the University of North Texas, USA.

First studies carried out with this tool suite on the basis of the benchmark documents are highly promising. However, they also indicate that some work will need to be done in terms of adding terms and categories (topics) from the product development and manufacturing domain to Wikipedia in order to improve the specificity of the document classification proposed by WikiRank. The ultimate target is to come up with a system specification that integrates WikiRank or components of it. The functions provided by that system will be mapped against the requirements in order to assess how well the specified system fulfils the initial targets.

If WikiRank shows major limitations with respect to the requirements, the authors will direct their research towards the establishment of an approach which combines the advantages of LDA with Wikipedia-based knowledge capitalisation.

6 Outlook

The target of this research is to conceive a knowledge mining system that helps researchers capitalise in a very efficient way on knowledge hidden in existing research papers. The system uses automatic topic identification to enable users to discover the essential semantic elements in papers without having to read those. It is important to mention that the authors are themselves researchers in the manufacturing and modern product development domain, and therefore can look at this subject from the user's perspective only. Otherwise stated, to conceive the system on the basis of existing approaches, tools for text mining and topic identification have to be available. As was pointed out in the conclusion, this is effectively the background of the particular activity that has been presented in this paper. The authors' next steps in this research project will be to evaluate a Wikipedia-based approach when applied to different sets of research papers in the manufacturing domain. These studies shall be used to show and validate the potential of the proposed application, as well as the used algorithms. The ultimate target is to come up with a system specification that can serve as a basis for experts to actually implement the system.

Acknowledgements

The authors want to thank the whole team from Indutech in Stellenbosch [4], South Africa, in particular Niek du Preez and Ernst and Wilhelm Uys, for freely providing their commercial CAT tool for this research, and for their valuable support and suggestions. They also owe particular thanks to Kino High Coursey and Rada Mihalcea for providing access to their Wikipedia-based tool suite WikiRank, and for their very helpful and immediate support.

References

[1] Riel A., Boonyasopon P., 2009. A Knowledge Mining Approach to Document Classification (Keynote paper). *The Asian International Journal of Science and Technology in Production and Manufacturing*, 2(3):1-10.

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

[3] www.analyzecontent.com, last accessed on 19/07/2010.

[4] www.indutech.co.za, last accessed on 19/07/2010.

[5] Indutech Content Analysis Toolkit User Guide Version1, 27/3/ 2009.

[6] Uys J.W., 2010. *A Framework for Exploiting Electronic Documentation in Support of Innovation Processes.* PhD thesis. Stellenbosch University, Stellenbosch, South Africa.

[7] Griffiths T.L., Steyvers M., 2004. *Finding Scientific Topics*, in Proceedings of the National Academy of Sciences of the United States of America, April 2004, 5228-5235.

[8] Blei D., Ng A.Y., Jordan M.I., 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3: 993-1022.

[9] Mackay D.J.C. (2003). *Information Theory, Inference, and Learning Algorithms*. ISBN 978-0521642989. Cambridge University Press.

[10] Deerwester S., Dumais S.T., Harshman R., 1990. Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, 41(6): 391-407.

[11] Kontostathis A., Pottenger W.M., 2006. A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing and Management*, 42:56–73.

[12] Berry M.W., Dumais S.T., O'Brien G.W., 1995. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review archive*, 37(4): 573-595.

[13] Rosario B., 2000. *Latent Semantic Indexing: An Overview*, in INFOSYS 240, Final Paper. Spring 2000

[14] Wikipedia: Latent semantic indexing: last accessed 21/07/2010

[15] Wikipedia: Singular value decomposition: last accessed 17/07/2010

[16] Mihalcea R., Csomai A., 2007. *Wikify! Linking Documents to Encyclopedic Knowledge*, in Proceedings of the 16th ACM conference on Information and Knowledge Management (CIKM), Lisbon, Portugal, 233-242.

[17] Coursey K., Mihalcea R., Moen W., 2009. *Using Encyclopedic Knowledge for Automatic Topic Identification*, in Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL), Boulder, Colorado, June 2009, 210-218.

[18] Medelyan O., Witten I.H., Milne D., 2009. *Topic indexing with Wikipedia*, in Proceedings of the 1st AAAI WikiAI workshop on Wikipedia and Artificial Intelligence (WIKIAI'08), Chicago, I.L., CD-ROM.

[19] Coursey K., Mihalcea R., 2009. *Topic Identification Using Wikipedia Graph Centrality*, in Proceedings of NAACL HLT 2009: Short Papers, Boulder, Colorado, June 2009, 117–120.