Automatic knowledge extraction from manufacturing research publications

P. Boonyasopon, Andreas Riel, Wilhelm Uys, Louis Louw, S. Tichkiewitch, N. Du Preez

Contents lists available at ScienceDirect

CIRP Annals
Manufacturing Technology
Journal homepage: www.elsevier.com/locate/cirp

Automatic Knowledge Extraction from Manufacturing Research Publications

P. Boonyasopon[a,b], A. Riel[b,d], Wilhelm Uys[c,d], Louis Louw[c,d], S. Tichkiewitch (1)[b,d], N. du Preez (1)[c,d]

[a]KMRC, Knowledge Management Research Center, KMUTNB, Bangkok, Thailand
[b]G-SCOP Laboratory, Grenoble Institute of Technology, Grenoble, France
[c]Department of Industrial Engineering, University of Stellenbosch, Stellenbosch, South Africa
[d]EMIRAcle AISBL Research Association, Brussels, Belgium

Knowledge mining is a young and rapidly growing discipline aiming at automatically identifying valuable knowledge in digital documents. This paper presents the results of a study of the application of document retrieval and text mining techniques to extract knowledge from CIRP research papers. The target is to find out if and how such tools can help researchers to find relevant publications in a cluster of papers and increase the citation indices their own papers. Two different approaches to automatic topic identification are investigated. One is based on Latent Dirichlet Allocation of a huge document set, the other uses Wikipedia to discover significant words in papers. The study uses a combination of both approaches to propose a new approach to efficient and intelligent knowledge mining.

Knowledge Management, Decision making, Document Retrieval Technique

1. Introduction

Modern information technology provides access to a rapidly increasing amount of information sources inside and outside of organisational boundaries. Both in industry and academia the task of turning the vast amount of structured and unstructured information into useful and sustainable knowledge has become one of the biggest challenges, as knowledge is considered as a key to success in any organization. To face this challenge, knowledge management processes and tools aim at capturing and capitalizing on implicit and explicit organization knowledge residing in internal information sources such as people, databases and exchanged documents, as well as in external sources, most notably the internet.

Researchers are confronted with a very similar problem. Their access to publications is increasingly facilitated by internet and dedicated on-line data vaults. Currently deployed document retrieval techniques in related environments are mostly based on full-text search and categorization by title, keywords, abstract, or other specific parts of the whole documents. This approach is very limited and fails to take into account the actual content [1]. Moreover, the process of categorization is usually done manually, which is time-consuming and error-prone.

The target of the research underlying this paper is to propose and conceive a knowledge mining system based on text mining that facilitates and supports researchers in finding relevant papers based on their actual interests. This paper presents a completely new approach to this target by the combination of two fundamentally different tools and algorithms for text mining and automatic topic identification.

Section 2 explains the target in greater detail, section 3 introduces a particular text mining tool for automatic topic identification. Section 4 introduces a Wikipedia-based keyword extraction tool. Section 5 proposes a new approach to relating research papers using a combination of the two tools introduced in the previous sections. Section 6 briefly reports about a case study which is entirely based on CIRP publications. Finally, section 7 concludes this paper with an outlook on the next steps.

2. Target Description

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. It should help them find publications which are relevant to a particular paper. The interest of such a system is manifold. Firstly, it can help authors cite papers that are indeed related to their works. Secondly, it can help them discover relevant papers which they normally would not have looked for (e.g. papers from another domain). Thirdly, it can help them increase citation indices of their papers (by means of putting their papers into the system, and thereby making them available to be proposed as relevant papers to other authors). Many more useful specific applications can be envisaged [1], but the authors considered the three above-mentioned ones to provide key added value.

The system should use text mining and automatic topic identification to enable users discover the essential semantic elements in papers without having to read those. It is important to mention that the authors are researchers in the manufacturing, modern product development and innovation domains, and

therefore can look at this subject from the user's perspective only. The very first step was thus to do an extensive investigation in this field [2], and to select and procure appropriate tools. In a second step, a reference document corpus was specified.

3. Categorization by Automatic Topic Identification

One fundamental element of relating documents and of giving a basic idea of what documents are about is their categorization according to a defined set of topics. The targeted system should be able to automatically identify such topics.

3.1. Overview of CAT

The text mining tool CAT (Content Analysis Toolkit [3]) by Indutech (Pty) 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 from electronic text documents. In particular the tool can automatically analyze and categorize a huge number of documents into different topics.

To use CAT, the user has to specify the files to be analyzed, indicate the number of expected topics to be extracted from the pool of documents, and define the number of times a word has to appear in order to be considered in the analysis (the minimum word frequency). One or more so-called "stop-lists", which specify words that have little or no semantic value, can be further selected or created to exclude such words 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 as well as an Excel file, which essentially allows for the following operations:

- Visualization of word clouds associated with identified topics. Each topic is also 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.

3.2. Characteristics and Limitations of CAT

CAT is largely based on Latent Dirichlet Allocation (LDA), a well-proven and well-established algorithm for knowledge mining from unstructured document sources. As it is based on statistical modelling, it requires very huge text corpora to function correctly, and is thus computationally expensive.

Studies linked to this research confirmed that CAT has indeed a lot of functionalities that can help researchers retrieve explicit or tacit knowledge from collections of research papers [1]. 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 for huge corpora, this limitation makes 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 unsupervised process. A considerable amount of expert knowledge is required especially for the assessment of the result quality.
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.

A more detailed analysis of CAT, its use and its underlying algorithm has been published in [2] and [5].

4. Encyclopaedic Keyword Extraction

LDA 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 implies 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 encyclopaedic knowledge. The authors' subsequent research revealed that there is in fact a research community which has succeeded in using the digital encyclopaedia Wikipedia exactly for this purpose. Particularly interesting and relevant contributions from this community can be found in [6] and [7].

4.1. Overview of Wikify

Basically, the published Wikipedia-based approaches use a tool called Wikify, which is an unsupervised system to automatically identify the semantically important encyclopaedic terms (including compound words and acronyms) in an input text, and to link them to Wikipedia articles [8]. Currently its main application is the semantic annotation of webpages, however the fact that it makes available practically the whole semantic intelligence of 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 inherent semantic model of Wikipedia. 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 important, 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 would at the same time provide a solution to issues that are highly

problematic in LDA and related probabilistic semantic modelling approaches, such as compound words, polysemy, synonymy, and multi-language. Also, to eliminate totally insignificant words for the analysis, there is no need for stop-lists, which are language-specific and prone to be incomplete and outdated.

4.2. Characteristics and Limitations of Wikify

As Wikipedia is a universal encyclopaedia, its knowledge model poses two fundamental challenges to the automatic identification of significant keywords in research publications from a particular domain:

1. Terms can only be identified as being significant if they are known to Wikipedia, which is not necessarily the case for terms of cutting-edge research. However, due to the unequalled speed of growth of Wikipedia, missing terms are likely to be added very rapidly.
2. Far more than the potentially lacking depth in a particular domain, the immense breadth of knowledge contained in Wikipedia definitely creates a big challenge for the targeted knowledge mining application: Wikify identifies keywords and phrases in a text as significant irrespectively of their relevance to a chosen domain. Although this can help reveal unexpected information and knowledge in specific documents, it primarily leads to a high amount of keywords related to general knowledge and thus renders the analysis of domain-specific keywords more difficult and cumbersome.

Inspired from the characteristics of CAT and Wikify, which both represent cutting-edge text mining technology, the authors have found a way of combining the advantages and overcoming the limitations of both approaches to conceive a knowledge mining system according to the targets described in section 2.

5. Combining CAT and Wikify

CAT identifies topics in *huge document corpora without external knowledge*. All operations work on the entire corpus and a *globally spanning statistical model*, so the computation effort is very *high*. Wikify uses *external knowledge* from Wikipedia to do keyword extraction *per document* in a *very efficient* manner.

The key operation of the target application outlined in section 2 is relating a new document (e.g. a new research paper) to a given corpus of documents (e.g. published research papers from the manufacturing domain). Based on the findings above, the following procedure is proposed:

Step 1: Use CAT to automatically identify the keywords that characterise the document corpus, and group them into suggested topics, which can then be used to categorize the documents in the corpus. For each of these documents, the internal keywords are stored.
Step 2: Use Wikify to extract the keywords of a new document.
Step 3: Filter the keywords found by Wikify using the keywords identified by CAT.
Step 4: Match the list of filtered keywords found by Wikify with the list of internal keywords for each document of the document corpus and store the number of common keywords
Step 5: Extract the most relevant documents relatively to the new document with the higher score of common keywords.
Step 6: Add the new document and the list of filtered keywords to the document corpus.
Step 7: Repeat Step 2 for any other new document, keeping unchanged the corpus-specific knowledge basis that was established in Step 1.
Step 8: Repeat Step 1 from time to time to improve the keywords of the new document corpus.

Step 6 assures that new keywords (that come up due to new research) become part of the document (knowledge) base whenever the latter is updated using CAT in Step 8. This makes sure that the knowledge base—including the keyword list—evolves with the application of the system to new publications.

6. Case Study

This procedure was applied with a reference document corpus from the CIRP community composed of all the 240 papers of the CIRP-STCs (Scientific Technical Committee) Assembly (STC-A), Design (STC-Dn), and Optimization (STC-O) which have been published in the CIRP Annals from 2003 and 2008.

CAT was asked to identify 5 different significant topics in the CIRP document corpus. The computation time on an average PC was approximately 10 hours. The complete set of analysis results is available in an Excel file containing several different worksheets, which allow most notably to relate documents to topics, and documents to documents. From this analysis, Topic 1 was chosen as the reference topic. Its three most significant keywords were *product*, *process*, and *knowledge*, but it was characterised by a total of 73 keywords. Let the set of these keywords be *KT1*. For each of the 240 papers in the corpus, the global CAT analysis result file shows the degree of relevance of the respective paper to the topic. The 38 most relevant papers (with a degree of relevance >65%) were chosen from there, and for each of them the two most similar documents of the corpus were determined from the results file.

6.1. Algorithm used for the Study

In the next step each of these 38 papers was considered as Document Reference (*DR*) and analysed with Wikify using the following procedure:

For all i [1,38] $DR \leftarrow D_i$

1. Wikify the reference paper *DR*, and select only those (compound) words which match semantically with *KT1*. The result *KDR* is the subset of *KT1* which can be found in *DR*.
2. Build a binary representation of the relationship between *KT1* and *KDR* such that
   for all k [1..73] $BDR_k \leftarrow 1$ if $KT1_k \in KDR$, else 0
3. For all j [1..38], $i \neq j$, $DC \leftarrow D_j$
   3.1. Wikify the comparison paper *DC*, and select only those words and compound words which match semantically with *KT1*. The result *KDC* is thus the subset of *KT1* which can be found in *DC*.
   3.2. Build a binary representation of the relationship between *KT1* and *KDC* such that
   for all k [1..73] $BDC_k \leftarrow 1$ if $KT1_k \in KDC$, else 0
   3.3. Build a binary representation of the relationship between *KDC* and *KDR* such that
   for all k [1..73] $BDD_{j,k} \leftarrow 1$ if $BDR_k = BDC_k = 1$, else 0
   3.4. Calculate $SDD_j = \sum BDD_j$
4. The documents with the highest *SDD* are considered most relevant to the reference document *DR*.

This algorithm can be fully automated and calculated with few computing resources. It determines the documents from the selected corpus which are considered most relevant to a given reference document from this corpus.

6.2. Results

Figure 1 shows an excerpt of the result table of the above algorithm for one particular reference paper. The column entitled "Keywords/Documents" contains the keywords identified by CAT from the corpus. Only an excerpt of the 73 keywords is shown. The grey-shaded column to the right indicates the binary pattern of the reference document, determined according to the algorithm shown in 6.1. The other columns which have a filename in their first line indicate the binary pattern of the respective document from the corpus. The line corresponding to a particular keyword contains "1" if this keyword appears at least once in the respective document, "0" otherwise. The red (dark grey) column to the right of each document column contains the result of applying the binary "AND" operation between the pattern of the corpus document on the left with the reference document pattern. The number on the bottom of each such column indicates the number of "1" in it, and corresponds to the number of keywords which appear both in the reference document AND the respective document from the corpus.
The bottom part shows the filename and the title of the reference document, below in the "Wikify" section the filenames and titles of the corpus documents which have the highest number of keyword matches with the reference document. The "CAT" section shows the filename and the title of the corpus document which CAT considers most similar to the reference document.

	Keywords/ Documents	A-2008-1-9	A-2008-2-21		Dn-2006-2-769		A-2008-1-17		Dn-2007-1-159		A-2003-1-25		Dn-2004-1-151	
2202	activity	0	0	0	1	0	0	0	0	0	1	0	0	0
2203	analysis	1	0	0	0	0	1	1	0	0	0	0	1	1
2204	application	0	1	0	1	0	0	0	1	0	1	0	1	0
2205	approach	0	0	0	0	0	0	0	1	0	0	0	1	0
2206	architecture	0	0	0	1	0	0	0	1	0	0	0	0	0
2207	aspect	0	0	0	0	0	0	0	0	0	0	0	0	0
2273	technology	1	1	1	1	1	1	1	1	1	1	1	1	1
2274	time	0	1	0	1	0	1	0	1	0	1	0	1	0
2275	tool	1	1	1	1	1	1	1	1	1	1	1	1	1
2276	value	1	1	1	1	1	1	1	0	0	1	1	1	1
2277				36		42		31		34		27		39
2278														
2279	Reference *DR*	A-2008-1-9			Integration of a service CAD and a life cycle simulator									
2280														
2281	Wikify:	Dn-2006-2-769	42		Virtual Research Lab: A New Way To Do Research									
2282		Dn-2005-1-117	40		Mapping Knowledge about Product Lifecycle Engineering for Ontology Construction via									
2283		Dn-2004-1-151	39		An Approach to Life Cycle Oriented Technical Service Design									
2284		A-2005-1-9	39		Product Redesign Using Value-Oriented Life Cycle Costing									
2285		A-2004-1-13	39		Development of a Productive Service Module Based on a Life Cycle Perspective of Mai									
2286		Dn-2008-1-133	39		Modularization as an enabler for cycle economy									
2287		O-2003-1-389	39		A Novel Digital Enterprise Technology Framework for the Distributed Development and									
2288	CAT:	A-2005-1-9			Product Redesign Using Value-Oriented Life Cycle Costing									
2289		Dn-2003-1-109			Integrative Design Environment to Improve Collaboration between Various Experts									

Figure 1. Excerpt from the analysis for one particular reference document

The complete result table reveals that for a very large part of the 38 papers there is at least one document from the corpus that is among the most relevant papers indicated by both CAT and Wikify. The other similar papers identified via Wikify are in fact also very relevant to the respective reference papers. These results clearly indicate that Wikify can lead quickly to accurate results when filtered by document-corpus specific word lists.

It has to be emphasized that this case study has been limited to three STCs for reasons of efforts and clarity. The presented approach can be applied without modification to a much larger CIRP document base, including publications from all STCs. This will enable users to find relevant papers from unexpected STCs, which can leverage inter-STC referencing and collaboration. A good example for a paper which is relevant to all STCs is [8].

7. Conclusion

This paper suggests a system that supports researchers in finding publications that are relevant to their papers, and that consequently helps them build good bibliographies and increase citation indices. It introduces a novel approach to relating documents in a very efficient, reliable and accurate way. Its key characteristics can be summarized as follows:

1. The LDA-based CAT tool is used to identify topics which characterize the documents in the corpus. Each topic is defined by a list of keywords, sorted by their relevance to the topic. By providing a homogeneous corpus, where all documents are from one particular domain, one obtains topics and keywords that are characteristic and/or specific for a given domain (CAT itself is domain-independent).
2. The Wikipedia-based Wikify tool is used to identify significant words in a given document, sorted by the number of occurrences in this document. For this operation, Wikify uses sophisticated semantic concepts and the broad knowledge contained in Wikipedia.
3. Combining these two approaches and tools allows finding very quickly and automatically documents in the corpus that are similar or relevant to any given new document.

A method was shown that allows the programmatic implementation of this approach. However, this method does not yet exploit the latter's full potential. Valuable and already available information, such as the keyword frequencies determined by Wikify, or the more complex semantic relationship between keywords and compound terms determined by Wikify and the individual keywords found by CAT, are not yet used. They may provide a very useful means to further improve the fidelity of the similarity measure. To investigate these improvements is the subject of the authors' future research.

Acknowledgements

The authors owe particular thanks to Kino 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] Uys, J.W., 2010. A Framework for Exploiting Electronic Documentation in Support of Innovation Processes. PhD thesis. Stellenbosch University, Stellenbosch, South Africa.

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

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

[5] du Preez, N., Perry, N., Candlot, A., Bernard, A. (1), Uys, W., Louw, L., 2005, Customised high-value document generation. Annals of the CIRP, Vol. 54/1/2005, pp 123-126.

[6] 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.

[7] 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.

[8] Tichkiewitch, S.(1), Shpitalni, M.(1), Krause, F.(1), 2006. Virtual Research Lab: A New Way To Do Research. Annals of the CIRP, Vol. 55/2/2006, pp 769-792.