

TRENDS IN AN INTERNATIONAL INDUSTRIAL ENGINEERING RESEARCH JOURNAL: A TEXTUAL INFORMATION ANALYSIS PERSPECTIVE

J.W. Uys¹ *, C.S.L. Schutte² and W.D. Van Zyl³

¹*Indutech (Pty) Ltd., Stellenbosch, South Africa, +27(0)218871180, wilhelm@indutech.co.za*

²*Dept. of Industrial Engineering, Stellenbosch University, South Africa, +27(0)218083617, corne@sun.ac.za*

³*Dept. of Industrial Engineering, Stellenbosch University, South Africa, +27(0)218084234, wernervanzyl@sun.ac.za*

Abstract:

Industrial engineering (IE) is a multi-disciplinary field, with its research borders broadening into a wide range of sub-disciplines. The Computers & Industrial Engineering (CaIE) journal is one of the prominent journals in the world to publish IE research and has done so since 1977. In the interest of evaluating research scope, it is worth determining how IE fields have been covered by this journal. What are the current topics in IE, and how are these positioned over time? This article attempts to investigate these issues in an objective way by using a text analytical technique to analyze the CaIE publication collection. Due to the growth in the quantity of accessible textual information, and the growing importance of this type of information to business people and industrial engineers alike, the basic methodological premise is provided for the topic modeling process. For the study presented in this paper, the Latent Dirichlet Allocation (LDA) topic modeling technique was applied to the CaIE corpus from 1977 to 2011 (Vol 60).

The focus of this article is thus twofold: first, on interpreting the underlying topic trends in the CaIE's publication history; and second, on introducing the concept of topic modeling whilst highlighting its value to the modern industrial engineer and researcher.

Keywords: Statistical Topic Models, Text Mining, LDA, Publication Trend Analysis

1. INTRODUCTION

The Computers & Industrial Engineering (CaIE) journal has been in existence from 1977 and it targets an audience of researchers, educators and practitioners of industrial engineering and related fields. The journal publishes original contributions to the development of new computer-enabled methodologies for solving industrial engineering problems, and applications of these methodologies to problems of interest to the associated communities [26]. One way of establishing the position of IE globally is to review its research scope. The past 34 years have seen IE researchers produce in excess of 4, 400 papers published in the CaIE journal. In the interest of evaluating research scope, it is worth determining how IE fields were covered by the journal. Determining which themes are longstanding can indicate established research fields, whilst emerging topics point toward development in the global IE research community, signifying possible new trends in the years ahead. The borders of IE as an interdisciplinary subject are broadening. Knowing what the role and function of IE will be in the future is a point of interest for IE researchers and practitioners alike. Dastkhan and Owlia [8] found that, in an international publication context, most future IE research will be focused on subjects like information technology, intelligent systems, optimization, quality, and supply chain management. Their research further highlights that the proportion of publication outputs on production management has decreased during the last decade. Research on topics like intelligent systems, supply chain management, and information technology has increased, with IE research topics spreading to other management and engineering departments. In order to investigate the CaIE research scope trends for comparison with the finding of the above-mentioned study [8], an objective method serves to determine the themes covered in CaIE articles. The effective analysis of such a substantial collection requires a structured approach: finding articles similar to those of interest, and exploring the collection by means of prevailing underlying themes. The problem is that the structure,

consisting of the index of ideas (or themes) contained in the articles as well as of those covering similar kinds of ideas (or themes), is not readily available in the CaIE or in most contemporary collections. The ability to analyze large quantities of textual information is not only relevant to this study: its importance is constantly growing since scientists (and engineers) are now confronted with access to millions of articles in their fields of interest, for which simple search capabilities may not yield satisfactory results. Blei and Lafferty [5] support this, stating that scientists (engineers) require new tools to explore and browse large collections of scholarly literature, since digital libraries now span a significant number of decades. The effort involved to read and understand large collections of unstructured information – or natural language text – remains a challenge in spite of many technological advances in the field of information and communication technology [13].

Unstructured information in the form of natural language text is a convenient and common way to capture and store a variety of information types. Currently, unstructured information is found in physical objects (e.g. books, reports, etc.) as well as in virtual objects (e.g. web pages, computer files, e-mails, etc.). New tools are required for automatically organizing, searching, indexing, and browsing the increasing number of collections of electronic documents [5] as the time available for an individual to collect, read, interpret, and act upon appropriate natural language text is limited in both corporate and research environments. The limitation of traditional information retrieval approaches, such as the well-known desktop search engines, is that such approaches are less efficient when the user is uncertain about what precisely is sought, or is not familiar with the content being searched. This limitation is even more severe when the content of the document collection is largely new territory for the searcher [17] – a situation that is not strange to the industrial engineer. The finding-a-needle-in-a-haystack approach of knowledge retrieval is not always optimal for answering all types of questions [16]. A multi-dimensional index is required to help the user to find the core themes addressed in a set of documents, as well as how these themes relate to individual documents, and vice versa. As such, topic modeling is proposed as a useful and time-efficient tool for scholars and professionals in analyzing their particular areas of interest in terms of prevailing themes, the relative time periods associated with such themes, and authoritative authors within certain areas of interest for a given corpus. The increasing electronic availability of articles makes computer-aided analysis of such articles more attractive.

In this article the basic methodological premises are provided for the topic modeling process. Then the method is applied to the CaIE corpus. The results and discussion shed light on the current state of IE research publication topics in the world (from a CaIE perspective), while simultaneously providing a comparison to the findings of another IE publication trend study.

2. METHOD

Statistical topic models are useful mechanisms for identifying and characterizing themes in a document collection. They allow users to explore a collection of documents in a topic-oriented manner [3], and prove to be a powerful method for discovering structure in otherwise unstructured data such as text data [10]. Topic models are useful for: distinguishing the gist of a set of documents, grouping documents based on their content, predicting the likely words of a theme, identifying the different senses or meanings of words (disambiguation), aiding information retrieval, and facilitating collaborative filtering [6];[11]. In essence, a topic model is a machine learning technique that can be applied to count data to find the underlying (or latent) structures for a given input dataset. It is a generative model since it models a process that gives rise to given sets of observed data items (e.g. the occurrence of words in documents). Statistical methods are applied to the generative model to make the underlying structures explicit. In a sense, topic models attempt to reverse the authoring process, where the author of a given text selects each significant written word from a mental, conceptual topic. The topic model seeks to identify and quantify these latent topics using a generative model and statistical inference to resolve the most likely set of latent structures, given the observed words in the actual text. In applying topic models to text data (e.g. the words of the article you are reading), the underlying structures correspond to the topics or themes indirectly addressed

in the text, while the observed data corresponds to the actual words found in the text. Topics are comprised of words associated with the individual topics. The relationship between a topic and a word is quantified by a probability representing the likelihood that the respective word corresponds to the topic in question. Topics can be seen as abstractions of the input text data in that they are more concentrated and aim to explain all the actual words found in the actual text analysed. Several topic modeling techniques exist (as described in [6];[15];[1];[3]), but all statistical topic modeling techniques are based on generative models. These techniques are differentiated by the parameters that the analyst has to specify (e.g. number of topics), the statistical methods used to infer the underlying structure given the observed words in documents, whether the similarity and hierarchy of the calculated topics are included in the resulting model (e.g. flat versus hierarchical models), and whether the technique assumes that topics are independent of time or caters for topic evolution. One of the attractive features of using topic models to analyze text documents is that a document is represented as a mixture of topics, while most other document-clustering techniques only represent a document using a single topic – an unrealistic restriction [25]. For the study presented in this paper, the Latent Dirichlet Allocation (LDA) topic modeling technique [6] was used. The LDA model assumes that words are generated by (a mixture of) topics and that such topics are infinitely exchangeable within a document [6];[3]. Moreover, documents are represented as random mixtures over latent (i.e. underlying) topics where each topic is defined or characterized by a distribution of (corpus vocabulary) words [17]. The LDA model's latent multinomial variables are referred to as topics. LDA is a true generative probabilistic model of a document corpus and potentially overcomes the drawbacks of earlier topic models (e.g. pLSI, Mixture of Unigrams, etc.) due to its generative properties and the possibility of assigning multiple topics per document [23]. It further caters for synonymy [11]), polysemy [20], represents a very useful model for deriving structure in otherwise unstructured data as well as for generalizing new data to fit into that structure [3]. On a high level, the generation of a document corpus in LDA is modeled as a three step process. The first step entails sampling a distribution over topics from a Dirichlet distribution for each document. Subsequently, 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 [22]. More specifically, the following generative process is assumed for each document w in a corpus D [6]:

1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each of the N words w_n :
 - a. Choose a topic $z_n \sim \text{Multinomial}(\theta)$
 - b. Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on topic z_n .

Blei et al. [6] simplifies this process to arrive at the following:

1. Choose $\theta \sim \text{Dir}(\alpha)$
2. For each of the N words w_n :
 - a. Choose a word w_n from $p(w_n|\theta, \beta)$

The simplified process defines a marginal distribution of a document w as the continuous mixture distribution:

$$p(\mathbf{w}|\theta, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^N p(w_n|\theta, \beta) \right) d\theta$$

In the equation above $p(w_n|\theta, \beta)$ represents the mixture components and $p(\theta|\alpha)$ the mixture weights. There are three levels to the LDA representation. The highest level, say the corpus level, involves corpus level parameters α and β which are assumed to be sampled once in the process of generating a corpus. The second level, say the document level, is linked to document-level variables θ_d which are sampled once per document. The third and most detailed level, the word-level, is represented by word-level variables z_{dn} and w_{dn} which are sampled once for each word in each document. LDA posits that each word of observed and unseen documents is generated by randomly selecting a topic from a distribution

with a randomly chosen parameter which is sampled once per document from a smooth distribution on the topic simplex. The major inferential problem to be solved in order to use LDA is that of calculating the posterior distribution of the hidden variables given a document. Regrettably, this distribution is intractable for calculating exact inference. However, an assortment of approximate inference algorithms, e.g. Laplace approximation, Expectation-Maximization, Markov Chain Monte Carlo, Variational Bayes, or Variational Expectation-Maximization, Expectation Propagation and collapsed Gibbs sampling can possibly be applied to LDA [12];[17];[9];[6]. In this study, Gibbs sampling was applied for inference and parameter learning. A shortcoming of LDA, shared by most probabilistic topic models, stems from the bag-of-words assumption which allows words that should be generated by the same topic to be allocated to various different topics. LDA has wider applicability than only text as it may be applied to address problems associated with other data collections such as content-based image retrieval and bioinformatics [6]. The input to the topic modeling process is a word-to-document co-occurrence matrix, where each row represents a word, each column represents a document, and the entries indicate the frequency with which the specific word occurs in the specific document. Topic modeling software is used to construct the matrix by extracting all unique words from the corpus and counting how often a given word occurs in each of the documents analyzed. In the analysis process, stop words are eliminated from the extracted text using a user-supplied list. For the majority of topic modeling techniques, the number of topics to be formulated is specified in advance by the user. Some non-parametric topic modeling techniques exist where the number of likely topics is calculated as part of the analysis process [12]. The word-to-document co-occurrence matrix is used to infer the latent structures (topics) by means of statistical inference algorithms. The result of the inference is a set of topics that includes the words associated with individual topics, as well as the strength of association for each word-topic combination. The minimum outputs of the topic modeling process are a topic-to-word and document-to-topic matrix. The topic-to-word matrix presents the identified topics, each topic having a list of associated words, quantified with individual probabilities for individual words.

The generated topics give an overview of the content of the documents analysed. Each topic can be supplemented with a descriptive label by a subject matter expert to aid the interpretation of such topics. The document-to-topic matrix presents the mixtures of topics associated with each of the documents analysed. The level of association of a document to a given topic is quantified by a mixture ratio (a number between 0 and 1). Using these two matrices, other useful outputs (e.g. the similarity between all topic or document pairs) can be calculated. By extracting the author information and year of publication from each of the analysed documents and supplementing the document-to-topic matrix with this information, it is possible to calculate author-to-topic and publication-year-to-topic relationships. These relationships can be used to find the topics associated with a given author, as well as the time span of a given topic. Although the analysis of journal articles using topic models has been presented in research literature (e.g. [3]; [10]), it is still considered a novel approach. In the case of [3], 16,351 articles (resulting in 19,088 unique words) obtained from the JSTOR on-line archive were analysed using LDA and other topic modeling techniques. The goal was to find the underlying topic structure of the corpus. The resulting topic model was then applied to aid the browsing of articles, searching, and making similarity assessments.

3. RESULTS AND DISCUSSION

The available electronic versions of the articles published in the CaIE during the period 1977-2011 were obtained and the full text of 4497 documents (which included actual articles, editorial board lists, forewords and editor's notes) was analysed using a software program named CAT (Content Analysis Toolkit by Indutech) which employs the LDA topic modeling technique. The analyses run were configured to use English stop word lists to eliminate frequently occurring words with little meaning; numbers were also ignored. Four runs were performed on the CaIE corpus using different numbers of topics (i.e. 30, 50, 80 and 100 topics) to obtain the desired level of generality versus specificity of the topics. The 80-topic analysis was subsequently selected. This analysis run required 38 hours to complete

on an Intel i7 PC. During the analysis 9, 745, 564 words were extracted from the articles, of which 145, 918 were unique. No text could be extracted from 41 of the 4497 documents. The authors then individually supplied labels for each of the 80 topics, by inspecting the words characterizing the topic and reading some of the article abstracts assigned to a given topic, upon which a final label for each topic was specified by means of group consensus. Eight topics were labeled as so-called noise topics since no sensible label could be specified for these. Also, four topics were given dual labels (separated by the “|” character), indicating that the specific topic encompasses two distinctive themes combined into one topic.

3.1 Topic Time Trends

Figure 1 shows how each of the 72 topics (the eight noise topics are not shown) is positioned in the time period 1977-2011 (derived from the document-to-topic matrix), with the earliest topics shown first and more recent topics shown last. The number in parentheses at the end of each topic label indicates the number of articles associated with the topic in question. The start and end of the thin black line of each topic bar respectively represent the year of publication of the earliest article and the most recent article associated with this topic. The middle of the purple rectangle of each topic bar represents the average year of all articles associated with the given topic, and the length of the rectangle is determined by the standard deviation of all publication years of articles associated with the topic in question. In an effort to confirm the calculated time trends of the various topics, certain topics were selected to check them against experience and investigate how they correspond to a timeline of current and historic events. For example, Topic 53: *Early applications of Microcomputers* was covered from 1983 to 1992 (having 289 articles associated), more or less coinciding with initial adoption of microcomputers. Topic 35: *Ant Colony Algorithm Applications*, representing a much more recent topic, was covered between 2006 and 2010 (having 12 articles associated) which can be defended since the first ant colony algorithm was only proposed in 1992 and it took some years to mature before being applied to solve industrial engineering related problems.

3.2 International Trends and CaIE

Dastkhan and Owlia [8] analysed the trends in IE research over the past three decades in an attempt to predict future research developments in this field. The publications on different IE topics from four main international publishers (Pergamon, Elsevier, Springer, and Emerald) were studied. Selecting journals more relevant to IE, data derived from 7,114 papers were analysed according to specified categories. These categories were defined through a survey of keywords in the publications, themes of IE conferences, and ideas of experts in this field. They highlight that the proportion of publication outputs on production management has decreased over the last decade, whilst research on topics like intelligent systems, supply chain management, and information technology has increased. As result they postulate that, in an international publication context, most future IE research will likely be focused on subjects like information technology, intelligent systems, optimization, quality, and supply chain management. It is argued that the trend of research on the subjects categories has varied during different periods (Figure 2), possibly due to changes in research demand by industry and society (pull factor) as well as the scientific interest of researchers (push factor).

To relate the IE research trends in CaIE to the trends in international IE research publications, each of the 72 topics, attained through the topic modeling approach, was allocated to one of the ten main categories defined by [8]. The four topics with dual labels were treated as eight individual topics in the mapping exercise for improved allocation accuracy. Note that the study presented in this article, and the study presented by [8], approach the same problem from different angles. In this article prevailing themes within a dataset were deduced using a topic modeling approach, whilst the latter study was done using predefined categories and keywords. Thus, this study takes a topic cue from underlying themes in a body of text (i.e. clustering), while in the other study the categories used to categorize articles in their dataset were specified in advance (i.e. classification).



Figure 1: Time trends of topics prevalent in the Computers & Industrial Engineering Journal (1977-2011)

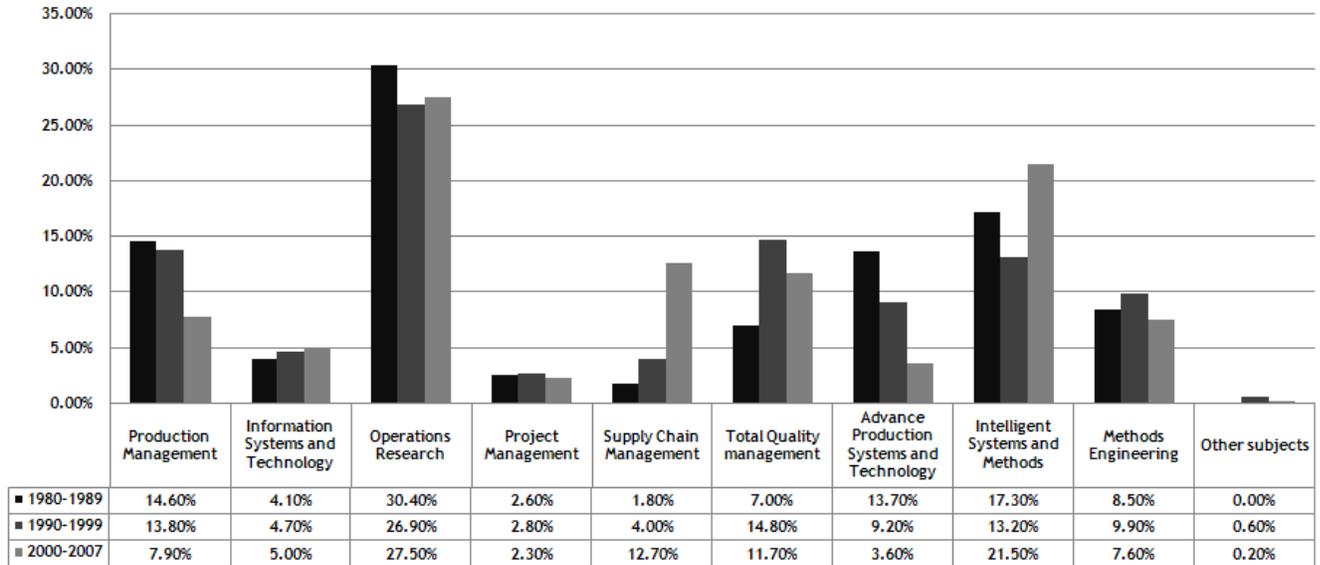


Figure 2: Proportion of research in different IE topics - international publications (adapted from [8])

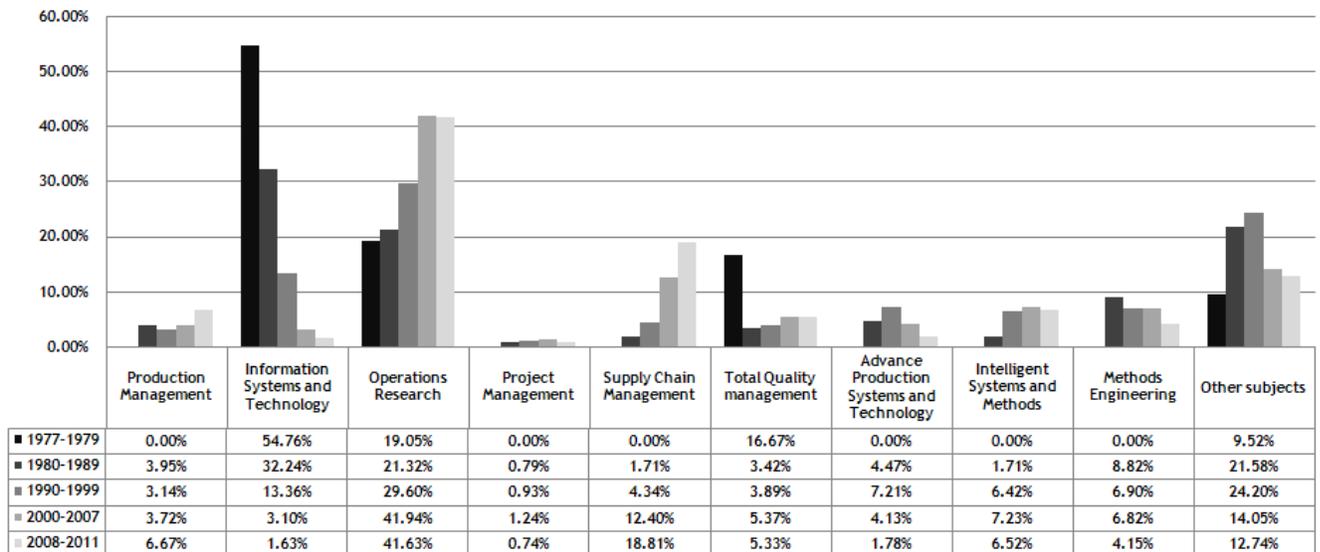


Figure 3: Proportion of research in different IE topics - CaIE (topic model approach)

A successful comparison of the results of the two studies therefore lies in the ability to relate and interpret Figure 2 and Figure 3. For instance, Figure 2 indicates that 26.9% of all (international) articles written between 1990 and 1999 were on the subject of Operations Research. From a CaIE perspective, Figure 3 indicates that during the same period 29.6% of the CaIE articles corresponded to the subject of Operations Research. Thus the first aspect to clarify is the high percentage of ‘Other subjects’ found in the 1990-1999 and 2000-2007 periods in Figure 3. Each of the 72 CaIE topics not allocated to one of the nine specific categories determined by [8] was grouped under ‘Other subjects’. Some of the most significant topics grouped under ‘Other subjects’ are: Topic 14: *Product Engineering*, Topic 18: *Risk & Safety Assessment Management Models*, Topic 24: *Dock Management Systems*, Topic 26: *Product Design for Manufacturing, Reuse, Recycling & Disposal*, Topic 42: *Product Remanufacturing & Resource Reuse Systems*, Topic 43: *Business Management Approaches & Strategies*, Topic 46: *Prevalent Computer Coordinate Measuring Techniques & Applications*, Topic 48: *Robotic Control Systems*, Topic 49: *Order Picking System Optimization*, Topic 50: *Economic Evaluation Models & Techniques*, Topic 76:

Engineering Management (Industrial), Topic 77: Automated Guided Vehicle Models & Systems, Topic 78: Machining & Tool Management and Topic 80: Enterprise Collaboration Models and evaluation thereof / Infrastructure Assessment Models. To a large extent, these topics correspond to the more recent topics found in Figure 1 (with Topics 48, 50, 76 and 78 being exceptions to this observation). Dastkhan et al. [8], who included CaIE in their analysis, used a manual categorization process, and while the details of their allocations are not known, the articles which fall into these categories were probably allocated to existing categories due to other reasons that were not apparent in this analysis.

Relating the results from the two studies can be simplified by removing the ‘Other subjects’ category from the comparison. This category in itself only constitutes 0.2% of international IE trends from 2000-2007; so its omission for the purpose of comparison is justified. Figure 4 shows the recalculated CaIE percentages across the nine specific categories, pairing the 2000-2007 series from Figure 2 and Figure 3, thereby objectively comparing the corresponding recent articles of this study with the recent articles of the study done by [8].

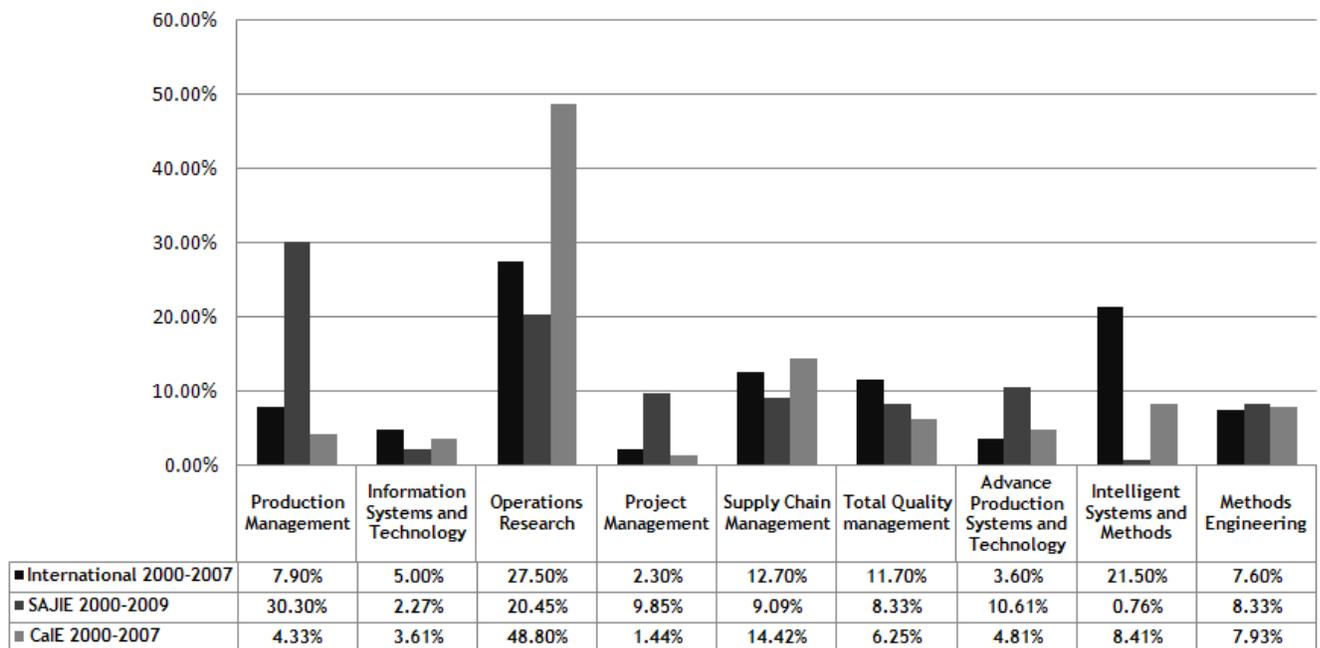


Figure 4: Proportion of research in different IE topics - CaIE and international

Figure 4 also shows the comparative percentages for the South African Journal of IE (SAJIE) that was determined as part of the results of a similar study (refer to [23]). By comparing the respective percentages of CaIE to that of the international study, it can be seen that CaIE closely corresponds to the international study in terms of the categories used by Dastkhan et al. [8], with the exception that CaIE seems to have a much stronger coverage in the Operations Research category and less coverage in the Total Quality Management and Intelligent Systems and Methods categories. This is probably not a surprising observation – each journal builds up a reputation regarding a focus area over its life, determined by the preferences of the editorial board and authors who published in the journal, and a casual analysis of CaIE articles confirms a stronger OR focus than other IE journals. An interesting observation is that Information Systems and Technology (that should actually be a prominent focus area considering the name of the journal), received significantly less attention in later years in CaIE, while the international trend was a slight increase.

4. CONCLUSION AND FUTURE RESEARCH

Electronic text is a convenient and popular way to capture and store information. Traditional information retrieval approaches are limiting in that they mostly cannot provide an overview of which subjects are covered in a collection of textual documents, making it difficult to find appropriate information. This limitation is even more severe when the content of the document collection is largely new territory for the searcher. New tools are required for automatically organizing, searching, indexing, and browsing the ever-growing electronic document corpora as the time available for an individual to collect, read, interpret, and act upon textual information is limited in both corporate and research environments.

Latent Dirichlet Allocation (LDA), a statistical topic model technique, is presented as a novel way to organize and explore a collection of electronic documents. When packaged in a software tool, the technique may be useful for improving the accessibility of the personal digital document libraries of researchers and practitioners alike. Benefits include being able to use calculated similarities between topics, documents, and individual words to navigate more easily through a document collection at different levels of aggregation.

Dastkhan et al. [8] reported on their study aimed at analyzing the international trend of IE research for the past three decades. In an attempt to compare the CaIE research trends with corresponding international trends, the authors first used the LDA technique to identify 72 research themes in the CaIE articles published from 1977 to 2011. These topics were subsequently mapped to the categories used by [8]. It was found that the CaIE topic spread leans more towards Operations Research based techniques, and when comparing the trends of CaIE with other journals in industry, it is clear that Operations Research remains a strong focus area (where other journals declined slightly). An interesting observation is that Information Systems and Technology (that should actually be a prominent focus area considering the name of the journal), received significantly less attention in later years in CaIE, while the international trend was a slight increase. These findings may be a convenient starting point for an open discussion around the future positioning of IE research in CaIE.

Due to space constraints, this paper currently provides a limited view on what is possible in this field. It is proposed that this research be extended to compare CaIE to other journals directly, and not via the Dastkhan categorization where the detail categorization process is not known. In addition, there are other topic modeling techniques available allowing for a certain degree of pre-specification of the desired topics (e.g. CTM [7], z-LDA [1]). Such techniques may be applied to the text of CaIE and other IE journals, using an industry accepted set of IE sub-disciplines (e.g. those identified in Salvendy [20]) as the predefined categorization structure, to determine the relative coverage of such sub-disciplines in the respective IE journals.

5. REFERENCES

1. Andrzejewski, D. & Zhu, X. 2009. Latent Dirichlet Allocation with Topic-in-Set Knowledge. NAACL 2009 Workshop on Semi-supervised Learning for NLP (NAACL-SSLNLP 2009).
2. Blei, D., Griffiths, T., Jordan, M. & Tenenbaum, J. 2004. Hierarchical topic models and the nested Chinese restaurant process, Neural Information Processing Systems Conference (NIPS).
3. Blei, D. & Lafferty, J. 2006, Modeling Science.
4. Blei, D. & Lafferty, J. 2007. A correlated topic model of science, *The Annals of Applied Statistics*, 1(1), pp 17-35.
5. Blei, D. & Lafferty, J. 2009. Topic models, In A. Srivastava and M. Sahami, eds, *Text mining: Theory and applications*. Taylor and Francis, in press.
6. Blei, D., Ng, A., & Jordan, M. 2003. Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3, pp 993–1022.
7. Chemudugunta, C., Smyth, P., Steyvers, M. (2008), *Combining Concept Hierarchies and Statistical Topic Models (CIKM'08)*, Napa Valley, California, USA.

8. Dastkhan, H. & Owlia, M.S. 2009. Study of trends and perspectives of Industrial Engineering research, *South African Journal of Industrial Engineering*, 20(1), pp 1-12.
9. Griffiths, T. & Steyvers, M. 2002. A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*.
10. Griffiths, T. & Steyvers, M. 2004. Finding scientific topics, *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 101(suppl. 1), pp 5228-5235.
11. Griffiths, T., Steyvers, M. & Tenenbaum, J. 2007. Topics in semantic representation, *Psychological Review*, 114(2), pp 211-244.
12. Jordan, M. editor. 1999. *Learning in Graphical Models*, MIT Press, Cambridge, MA.
13. Le Grange, L. 2006. The changing landscape of the contemporary university, *South African Journal of Higher Education*, 20(4), pp 367-371.
14. Li, W., Blei, D. & McCallum, A. 2007. Nonparametric Bayes Pachinko allocation, *Conference on Uncertainty in Artificial Intelligence (UAI)*.
15. Li, W., & McCallum, A. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations, *International Conference on Machine Learning (ICML)*.
16. Lieberman, J. From Metadata to megadata: Deriving broad knowledge from document collections, Collexis, Inc., Whitepaper. 2007.
17. Mimno, D. & McCallum, A. 2007. Expertise Modeling for Matching Papers with Reviewers, *Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*, San Jose, California, USA.
18. Minka, T.P. & Lafferty, J. 2002. Expectation-propagation for the generative aspect model, In *Uncertainty in Artificial Intelligence (UAI)*.
19. Nasukawa, T. & Nagano, T. Text analysis and knowledge mining system, *IBM Systems Journal*, vol. 40, no. 4, 2001.
20. Salvendy, G. 1992. *Handbook of Industrial Engineering*, John Wiley & Sons, Inc., New York.
21. Steyvers, M. & Griffiths, T. 2006. Probabilistic Topic Models, in *Latent Semantic Analysis: A Road to Meaning*, Trends in Cognitive Science. vol. 10, issue 7, pp. 327 – 334.
22. Steyvers, M., Smyth, P., Rosen-Zvi, M. & Griffiths, T. 2004. Probabilistic Author-Topic Models for Information Discovery, In: *10th ACM SigKDD conference knowledge discovery and data mining (Seattle, 2004)*.
23. Uys, J.W., Du Preez, N.D., & Uys, E.W. 2008. Leveraging Unstructured Information Using Topic Modelling, *PICMET 2008 Proceedings*, 27-31 July, Cape Town, South Africa. PICMET. pp. 955-951.
24. Uys, J.W., Schutte, C.S.L. & Esterhuizen, D. 2010. Trends in a South African Industrial Engineering Research Journal: A Textual Information Analysis Perspective, *South African Journal of Industrial Engineering (SAJIE)*, 21(1), pp.1-16.
25. Wei, X. & Croft, W.B. 2006. LDA-based document models for ad hoc retrieval, *Proceedings of the 29th SIGIR Conference*, pp 178-185.
26. http://www.elsevier.com/wps/find/journaldescription.cws_home/399/description