

TRENDS IN A SOUTH AFRICAN INDUSTRIAL ENGINEERING RESEARCH JOURNAL: A TEXTUAL INFORMATION ANALYSIS PERSPECTIVE

J.W. Uys^{1*}, C.S.L. Schutte² and D. Esterhuizen³

^{1,2,3}Department of Industrial Engineering
Stellenbosch University, South Africa

¹wilhelm@indutech.co.za, ²corne@sun.ac.za, ³denele@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 South African Journal of Industrial Engineering (SAJIE) has been the official voice of IE research in South Africa for the past 22 years. In the interests 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 do they align with IE publication trends in other countries? This article attempts to investigate these issues in an objective way by using a text analytical technique to analyse the SAJIE 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 relevant text analytical method is also outlined.

OPSOMMING

Bedryfsingenieurswese is 'n multidissiplinêre veld, met navorsingsgrense wat strek oor 'n wye reeks sub-dissiplines. Die Suid-Afrikaanse Tydskrif vir Bedryfsingenieurswese (SAJIE) was oor die afgelope 22 jaar die amptelike mondstuk vir bedryfsingenieurswese-gerigte navorsing in Suid-Afrika. Dit is betekenisvol om te bepaal in watter hoedanigheid bedryfsingenieurswesevelde deur die SAJIE gedek word, ten einde die bestek van navorsing te evalueer. Wat is die huidige temas van bedryfsingenieurswese, en hoe vergelyk dit met bedryfsingenieurswese-publikasieneigings in ander lande? Hierdie artikel poog om die vraagstuk op 'n objektiewe wyse na te vors deur die toepassing van 'n teksanalitiese tegniek om die SAJIE publikasie versameling te ontleed. Weens die toename in die hoeveelheid beskikbare tekstuele inligting, asook die toenemende belang van hierdie tipe inligting vir besigheidspersone asook bedryfsingenieurs, word die betrokke teksanalitiese metode ook bespreek.

¹The author was enrolled for a PhD (Industrial Engineering) degree in the Department of Industrial Engineering, Stellenbosch University.

²The author was enrolled for a PhD (Industrial Engineering) degree in the Department of Industrial Engineering, Stellenbosch University.

³The author is currently enrolled for an MSc (Industrial Engineering) degree in the Department of Industrial Engineering, Stellenbosch University.

*Corresponding author

1. INTRODUCTION

The South African Journal of Industrial Engineering (SAJIE) publishes articles with an emphasis on research, development, and application within the fields of industrial engineering and engineering and technology management. In this way, it aims to contribute to the further development of these fields of study and to serve as a vehicle for the effective interchange of knowledge, ideas, and experience between research and training oriented institutions and application oriented industry [20]. The SAJIE has been the official voice of industrial engineering (IE) research in South Africa since 1987, encouraging the publication of articles with practical applications, original research, meaningful new developments, and state-of-the-art surveys.

One way of establishing the position of IE in a country is to review its research scope. The past 22 years have seen IE researchers produce 283 South African-published papers. In the interests 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 South African IE research community, signifying 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 [7] found that, in an international publication context, most future IE research will be focused on subjects like information technology, intelligent systems, optimisation, 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 discover similar South African research scope trends and benchmark these against those of the international study [7], an objective method serves to determine the themes covered in SAJIE 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 SAJIE or in most contemporary collections. The size and growth rate of modern digital libraries inhibit manual analysis. To develop the tools required for exploring and browsing such libraries, largely automated methods of organising, managing, and presenting their contents are required. The personal digital (document) libraries of researchers and practitioners reveal the same need. As such, statistical topic modelling is proposed as a useful and time-efficient tool to analyse particular areas of interest in terms of prevailing themes, timing of themes, and authoritative authors within areas of interest in a predetermined dataset.

However, the ability to analyse 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 [3] 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 associated with reading and understanding 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 [11].

Unstructured information in the form of natural language text is abundant in various types of organisation [3] and expands daily. Electronic text is a convenient and common way to capture and store a variety of information types such as facts, currencies, dates, persons, places, etc. Currently, unstructured information is found in physical objects such as books,

reports, academic dissertations, magazines, and newspapers. Unstructured information is also found in virtual objects such as text messages on cellular phones, web pages (including wikis, blogs, etc.), word processor and other computer files, e-mails, e-books, instant messaging messages, databases (issue tracking systems, customer relationship management systems, etc.) and many more. New tools are required for automatically organising, searching, indexing, and browsing the increasing number of collections of electronic documents [3] 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 [15] - 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 [14]. 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 modelling is also proposed as a useful and time-efficient tool for scholars and professionals in analysing 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.⁴ 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 modelling process. Then the method is applied to the SAJIE corpus. The results and discussion shed light on the current state of IE research publication topics in South Africa, while simultaneously providing a benchmarking guideline for international IE research publication trends. The focus of this article is twofold: first, on interpreting the underlying topic trends in the SAJIE's publication history; and second, on introducing the concept of topic modelling whilst highlighting its value to the modern industrial engineer and researcher.

2. METHOD

Statistical topic models are useful mechanisms for identifying and characterising themes in a document collection. They allow users to explore a collection of documents in a topic-oriented manner (Blei & Lafferty [2]), and prove to be a powerful method for discovering structure in otherwise unstructured data such as text data (Griffiths & Steyvers [9]). 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 filtering⁵ (Blei, Ng, & Jordan [4]; Griffiths, Steyvers & Tenenbaum [10]).

The value of a topic model as a mechanism to generate interpretable semantic representations, and its link to cognitive psychology, is thoroughly discussed in Griffiths et al. [10].

⁴ The term 'corpus' refers to a collection of texts.

⁵ Personalised recommendations of items (e.g. journal articles) to persons with similar preferences within given domain of interest.

In essence, a topic model is a machine learning technique that can be applied to count data⁶ 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 topics⁷ 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 abstractions⁸ 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 modelling techniques exist (as described in Blei et al. [4]; Li & McCallum [13]; Blei, Griffiths, Jordan & Tenenbaum [1]; Blei & Lafferty [2]), but all statistical topic modelling 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 cater for topic evolution. One of the attractive features of using topic models to analyse 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 (Wei & Croft [19]).

For the study presented in this paper, the Latent Dirichlet Allocation (LDA) topic modelling technique (Blei et al. [4]) was used. Although the analysis of journal articles using topic models has been presented in research literature (see Blei & Lafferty [2]; Griffiths & Steyvers [9]), it is still considered a novel approach. In the case of Blei & Lafferty [2], 16,351 articles (resulting in 19,088 unique words) obtained from the JSTOR on-line archive were analysed using LDA and other topic modelling techniques. The goal was to discover 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.

The input to the topic modelling 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 modelling 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 analysed. In the analysis process, stop words⁹ are eliminated from the extracted text using a user-supplied list. For the majority of topic modelling techniques, the number of topics to be formulated is specified in advance by the user. Some non-parametric topic modelling techniques exist where the number of likely topics is calculated as part of the analysis

⁶Count data is data in which the dependent variable can take only non-negative integer values.

⁷The subject matter of a conversation or discussion.

⁸The process of formulating general concepts by abstracting common properties of instances.

⁹High-frequency words with little meaning that occur in all writings, such as the, of, for, from, etc.

process (Li, Blei & McCallum [12]). The word-to-document co-occurrence matrix is used to infer the latent structures (topics). Statistical inference algorithms such as Expectation Propagation, Expectation-Maximisation, Markov Chain Monte Carlo, Variational Bayes, or Variational Expectation-Maximisation, are typically used. 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 modelling process are a topic-to-word matrix and a 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, for the given topic. 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. 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.

3. RESULTS AND DISCUSSION

The available electronic versions of the articles published in the SAJIE during the period 1987-2009 were obtained, and the remainder of the articles were scanned to create electronic versions using hard copies of the journal. The full text of 283 articles was analysed using a software program named CAT¹⁰ (*Content Analysis Toolkit* by Indutech). This article draws on the results from this analysis, illustrating the applications and benefits of this topic modelling technique to the industrial engineer and then highlighting the main SAJIE findings in terms of time trends, and benchmarking the current state of IE research to international trends.

The analysis run was configured to use English and Afrikaans stop word lists to eliminate frequently occurring words with little meaning in these languages. The eventual analysis run required three hours to complete on a multiprocessor PC. During the analysis 609,360 words were extracted from the articles, of which 17,038 were unique. A topic model consisting of 50 topics¹¹ was formulated using CAT, and each article was further automatically associated with one or more of these topics with individual strength of association values. Once the analysis run was completed, the authors assigned a descriptive label to each topic by inspecting the words characterising the topic and reading some of the article abstracts assigned to a given topic.

¹⁰www.analyzecontent.com

¹¹Topic models consisting of 30 and 100 topics were also formulated and labelled by the authors, but it was found that the quality of topics given the size of the data set was best illustrated by a topic specification of 50.

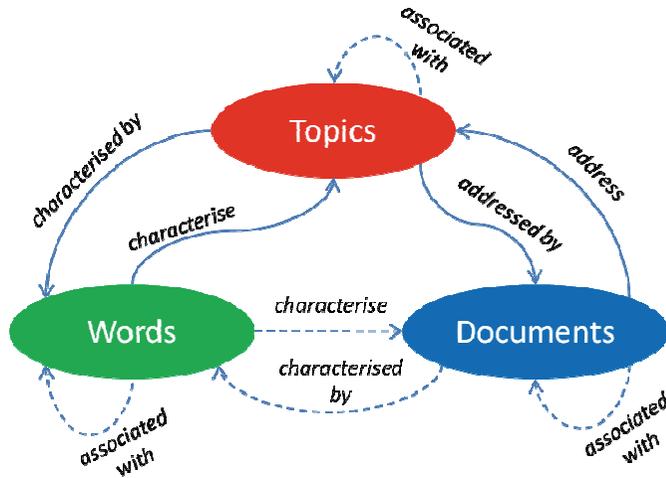


Figure 1: Different aspects of a topic model

As indicated in Figure 1, the output of a topic modelling analysis can be used to find similarities between topics and individual words, topics and individual documents, a given topic and other topics, a given document and other documents, and between a given word and other words. In this figure the dotted lines represent relationships that can be derived from the resulting topic model data. Table 1 provides an overview of the labels of the SAJIE topics¹² as interpreted by the authors. Each person individually supplied labels for the 50 topics, and then the resulting labels were compared and mutually agreed upon. Dual labels are represented by an asterisk (*), indicating that the specific topic encompasses two distinctive themes combined into one.

Each topic in Table 1 is characterised by 100 words or terms, each having an associated weight for a specific topic. The topic model may associate a given word or term with more than one topic.

3.1 Finding articles relevant to a given topic

The following figure (Figure 2) is an example of the words and terms characterising the topic *Project Management Methods and Tools* (Topic 2, see Table 1). As indicated by Figure 1, using a given topic as reference point is a useful method to find other relevant topics (e.g. Topic 31, 23, etc.) as well as relevant documents (e.g. *indeng_v17_n1_a5.pdf* [6], *indeng_v19_n1_a7.pdf* [8], etc.). The tag cloud representation in the left hand side of Figure 2 is a spatial representation of the words and terms associated with the given topic, with the terms and words in the middle corresponding to the words and terms having high likelihood, given the topic. The right hand side of Figure 2 shows the documents and topics associated with the given topic. Such an analysis may help researchers to analyse topics of interest for associated keywords and other linked areas of study, as well as the applicable literature.

¹²Topics are listed as returned by the software without any ranking or rating implied by topic numbering.

Topic 1 : Solving Scheduling Problems: Related Methods & Heuristics	Topic 26 : Capacity optimisation Applications
Topic 2 : Project Management: Methods & Tools	Topic 27 : Inventory Systems & Production Planning
* Topic 3 : Maintenance Scheduling Spent Fuel Management	* Topic 28 : Organisational Maturity Leveraging Organisational Intangible Assets
Topic 4 : Models for Software Reliability & Information Systems	Topic 29 : Approaches for Distribution Scheduling & Stock Replenishment
Topic 5 : Productivity Measurement	* Topic 30 : Lift Club Scheduling Reliability Centred Maintenance
Topic 6 : Holistic Business Management	Topic 31 : Project Success Factors & Governance Practices
* Topic 7 : Technology Management Grinding Process	Topic 32 : Applications of Statistical & Mathematical Modelling
Topic 8 : Payload Optimisation	Topic 33 : Fuzzy Logic & Mechatronics Controllers
Topic 9 : Measuring & Managing Organisational Performance	Topic 34 : Technology Management & Strategy
Topic 10 : Quality Management	Topic 35 : General industrial engineering Terminology
Topic 11 : Reliability of Electricity Supply	Topic 36 : Organisational Change
* Topic 12 : Technology Forecasting Optimal Facility Location	Topic 37 : Commercialisation & Transfer of Technology in South Africa
Topic 13 : Product & Enterprise Design	Topic 38 : Sampling Planning
Topic 14 : Manufacturing Process Optimisation	Topic 39 : Strategies & Models for Business Operations
Topic 15 : Simulation Modelling	Topic 40 : General Afrikaans Topic (1)
Topic 16 : Quality Management	Topic 41 : Machining Processes
* Topic 17 : Optimisation of Recycling Operations Financial Risk Analysis	Topic 42 : Supply Chain Management
Topic 18 : Manufacturing Process Improvement	Topic 43 : Business Management Approaches & Strategies
Topic 19 : Maintenance Performance	Topic 44 : Condition-based Maintenance & condition monitoring
Topic 20 : Legislation impacting the Engineering Profession	Topic 45 : General Afrikaans Topic (2)
Topic 21 : System Availability & Reliability	Topic 46 : Sustainable Development
Topic 22 : Robotic Control Systems	* Topic 47 : Productivity Improvement Human Capital
Topic 23 : Resource scheduling	Topic 48 : Noise Topic
Topic 24 : Risk Assessment & Management	Topic 49 : Methodologies (Various)
Topic 25 : Expert systems, Knowledge Based Systems & Manufacturing Simulators	* Topic 50 : Industrial Engineering Research & Education Entrepreneurship

Table 1: Overview of the topics included in the SAJIE topic model

* Dual label - topic encompasses two themes.

Topic 2 -Project Management: Methods and Tools

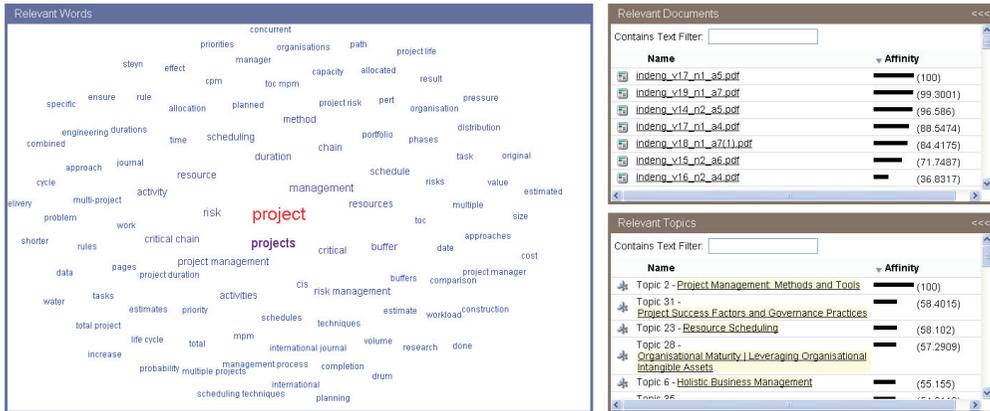


Figure 2: Characterisation of the topic *Project Management Methods and Tools*

3.2 Finding topics relevant to a given article

Figure 3 shows that, using a given document (e.g. *indeng_17_n1_a5.pdf* [6]) as reference point, the model conveniently indicates other relevant documents (e.g. *indeng_v19_n1_a7.pdf* [8], *indeng_v18_n1_a7(1).pdf* [18], etc.), associated topics (e.g. *Project Management Methods and Tools*) and words describing the specific document (e.g. *projects, management, risk, critical, duration, project management, resources*, etc.). The value of such topic analysis lies in linking the explicit level (one or more articles) to the abstract level (the associated topics). At the topic level, relationships to other related topics may be investigated, and the analyst can zoom in to the explicit level of such topics by investigating the associated articles.

indeng_v17_n1_a5.pdf

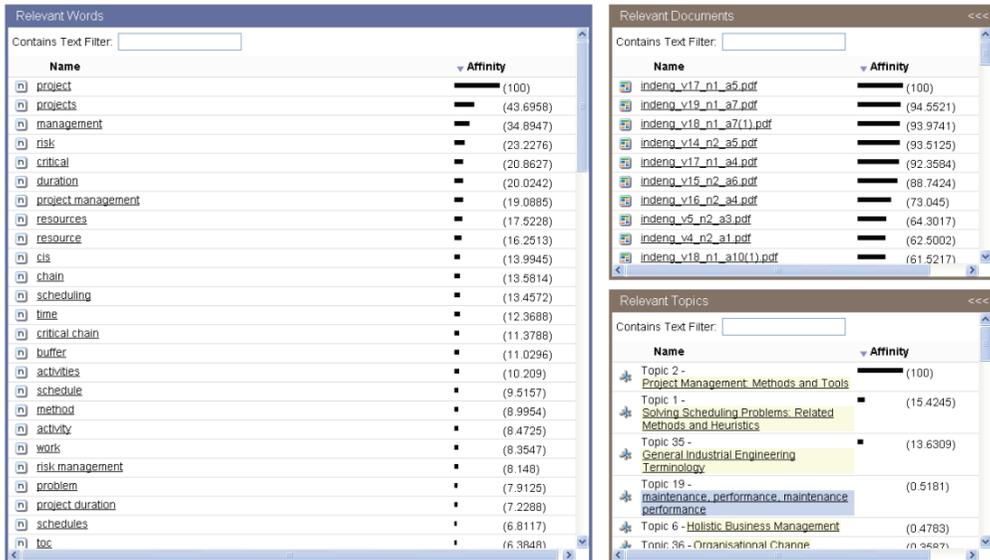


Figure 3: Characterisation of the document *indeng_v17_n1_a5.pdf* [7]

3.3 Finding articles, topics, and words associated with a specific key word or term

Using a specific word as a reference point (e.g. ‘scheduling’, Figure 4) one may utilise the model to find associated words (e.g. *maintenance*, *project*, *time*, *problem*, *constraint*, etc.), documents (e.g. *indeng_v16_n2_a8.pdf* [16], *indeng_v19_n2_a4.pdf* [17], etc.) and topics (e.g. Topic 3 - *Maintenance Scheduling | Spent Fuel Management*, Topic 1 - *Solving Scheduling Problems: Related Methods and Heuristics*, etc.).

3.4 Topic time trends in SAJIE

The calculated topic model was extended by manually including the year of publication of each article analysed. Using the topic-to-document relationships included in the topic modelling results, and the publication year of individual articles, it became possible to calculate the time trend of each topic. Figure 5 shows how each of the 50 topics is positioned in the time period 1987-2009, with the earliest topics shown first and the 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.

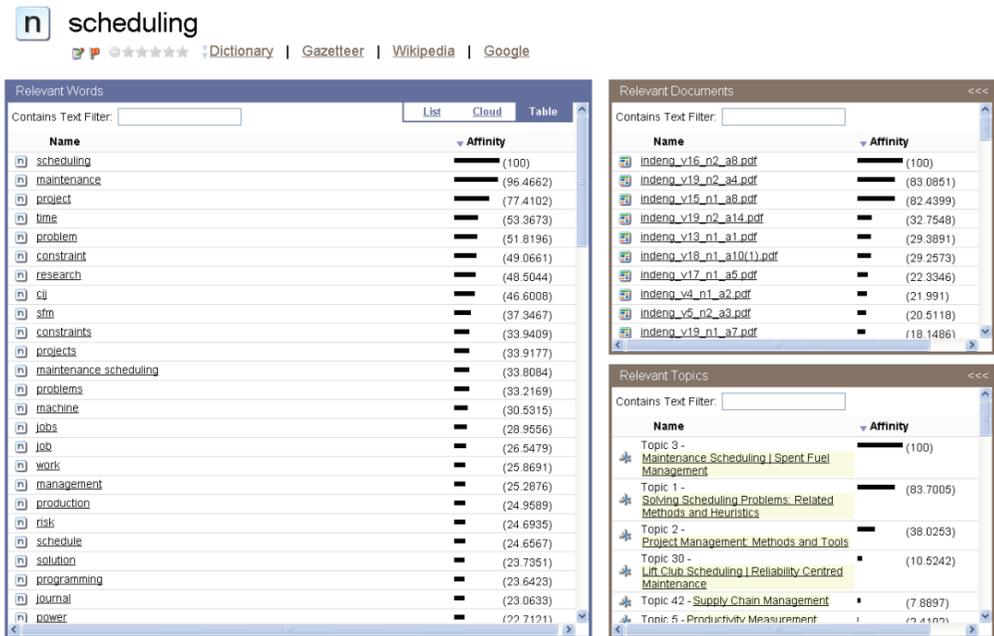


Figure 4: Characterisation of the word ‘scheduling’

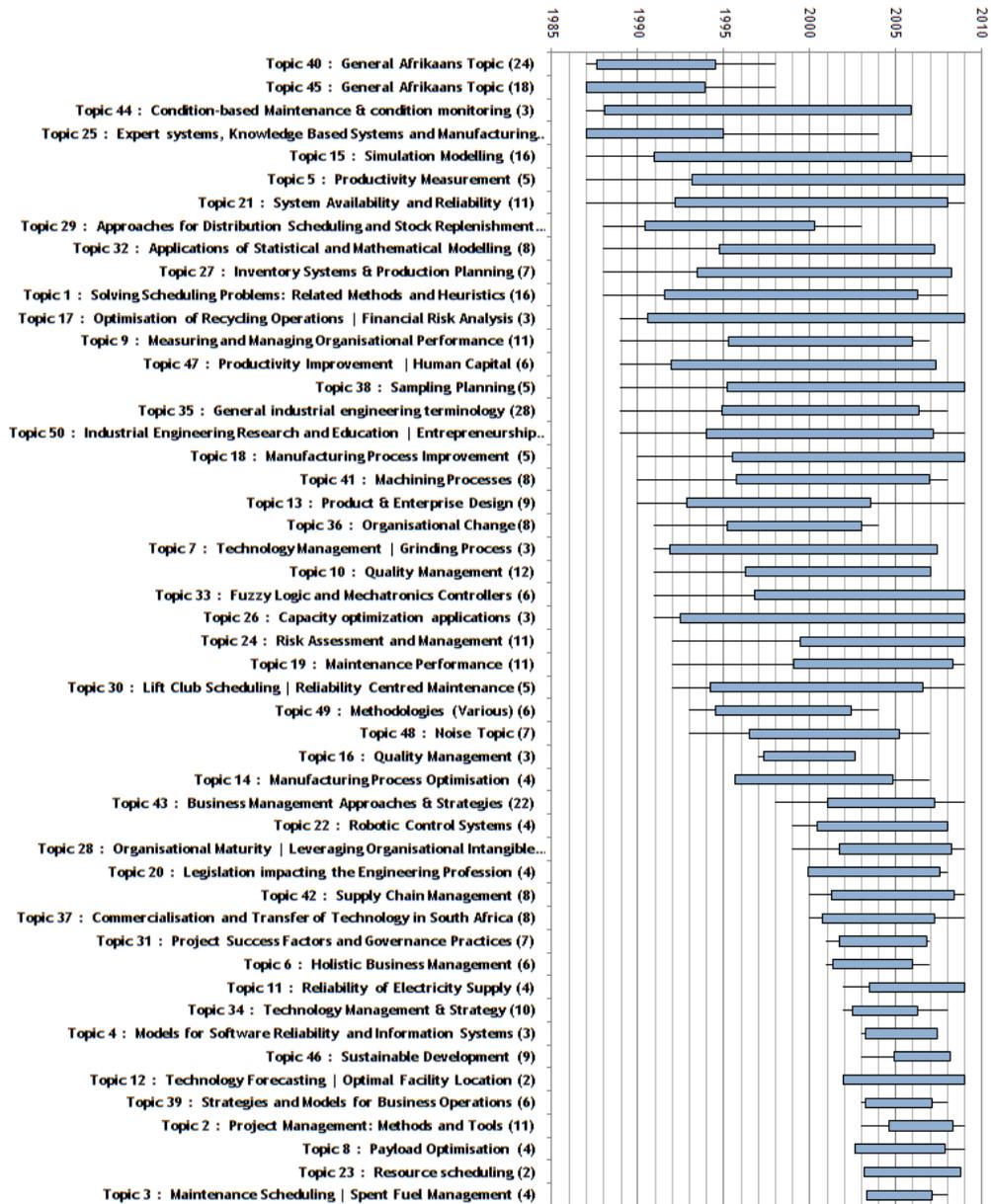


Figure 5: Time trends of topics prevalent in the South African Journal of Industrial Engineering (1987-2009)

In Figure 5 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 blue 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.

Figure 5 provides the possibility of further analyses beyond the scope of this article, but various interesting trends are briefly discussed under the headings of ‘longstanding topics’, ‘brief topics’, and ‘recent topics’.

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 11: *Reliability of Electricity Supply* was covered from 2003 onwards, more or less coinciding with the recent South African electricity crisis. SAJIE published selected full text articles in Afrikaans during the 1980s and 1990s, and this is affirmed by the time trends of two topics, namely Topic 45: *General Afrikaans Topic (1)* and Topic 40: *General Afrikaans Topic (2)*. The themes represented in section 3.4.2 also affirm the relevance of the topics calculated by the model and the timing of such topics as calculated by the authors.

3.4.1 Longstanding topics

Figure 5 illustrates the following topics as having the longest duration in years (regardless of how recently they have appeared) over the analysed period of time (1987-2009):

- Topic 21: *System Availability and Reliability*
- Topic 5: *Productivity Measurement*
- Topic 15: *Simulation Modelling*
- Topic 1: *Solving Scheduling Problems: Related Methods and Heuristics*
- Topic 27: *Inventory Systems & Production Planning*

Topics that reflect a longstanding interest may be particularly relevant to researchers in longitudinal studies, where reported data over a longer period may then be compared. This information may also be of interest to newcomers in a field, assisting them to grasp the scope and fundamental basis of the specific discipline.

3.4.2 Recent topics

The following are some of the significant topics identified as being recent (i.e. topics that were addressed from 2002 onwards):

- Topic 46: *Sustainable Development*
- Topic 39: *Strategies and Models for Business Operations*
- Topic 34: *Technology Management & Strategy*
- Topic 4: *Models for Software Reliability and Information Systems*
- Topic 23: *Resource scheduling*

Recent topics are influenced by recent developments in IE research, by 'hot topics' in the world arena (e.g. global warming and sustainable development), and by local engineering management researchers using the journal as a publication target, thereby extending the scope of the journal in recent years to include engineering management topics as well.

3.4.3 Topic coverage

Using the topic-to-document assignment scores of the topic model, it is possible to calculate the topics that are best represented in the analysed corpus. The analysis of topic coverage allows the analyst objectively to determine the prevalence of the respective topics in a given dataset. The results of this study indicate that more has been written about Topic 43: *Business Management Approaches & Strategies* (3.4%), the best-covered topic, compared with the least-covered topic, Topic 11: *Reliability of Electricity Supply* (0.85%). This seems intuitive, as the reliability of electricity supply in the South African context only became an issue in recent years, whereas approaches to business management represent a timeless topic.

3.5 International trends and the SAJIE

Dastkhan and Owlia [7] 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, optimisation, quality, and supply chain management. It is argued that the trend of research on the subjects categories has varied during different periods (Figure 6), possibly due to changes in research demand by industry and society (pull factor) as well as the scientific interest of researchers (push factor).

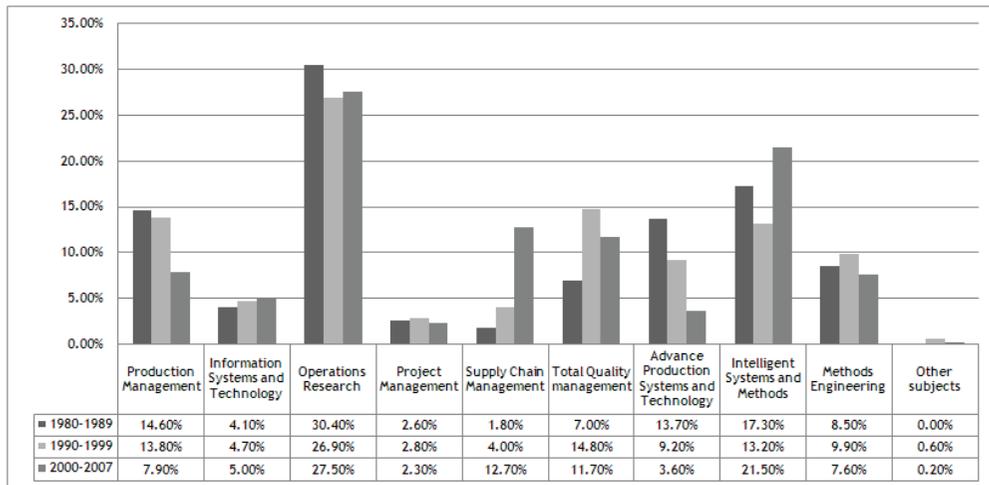


Figure 6: Proportion of research in different IE topics - international publications (adapted from Dastkhan and Owlia [7])

To relate the trends in IE research in the SAJIE to the trends in international IE research publications, each of the 50 topics attained through the topic modelling approach outlined in this article was allocated to one of the ten main categories defined by Dastkhan and Owlia¹³ [7]. The eight topics with dual labels were treated as 16 individual topics in the mapping exercise for improved allocation accuracy. It is important to note that the study discussed in this article, and the study presented by Dastkhan and Owlia, approach the same problem from different angles [7]. The authors of this article deduced prevailing themes within a dataset using a statistical topic modelling approach, whilst the latter study was done using predefined categories and keywords. In other words, this study takes a topic cue from underlying themes in a body of text (this corresponds to clustering), while in the other study the categories used to categorise all articles in their dataset were specified in advance (this corresponds to classification).

¹³The corresponding author will provide detailed information on the mapping process on request.

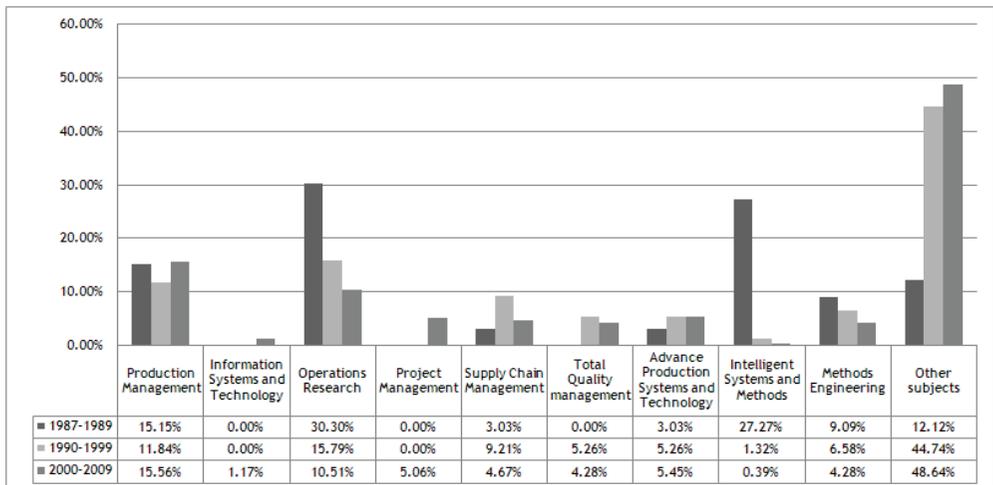


Figure 7: Proportion of research in different IE topics - SAJIE (topic model approach)

A successful comparison of the results of the two studies therefore lies in the ability to relate and interpret Figure 6 and Figure 7. For instance, Figure 6 indicates that 26.9% of all (international) articles written between 1990 and 1999 were on the subject of Operations Research. From a South African perspective, Figure 7 indicates that during the same period 15.79% of the SAJIE 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 2000-2009 periods in Figure 7. Each of the 50 SAJIE topics not allocated to one of the nine specific categories determined by Dastkhan and Owlia [7] was grouped under ‘Other subjects’. Most of these topics are more management-oriented, and may be attributed to the fact that the SAJIE also publishes articles in the fields of Engineering and Technology Management (ETM). Some of the most significant topics grouped under ‘Other subjects’ are: Topic 6: *Holistic Business Management*, Topic 13: *Product & Enterprise Design*, Topic 28: *Organisational Maturity | Leveraging Organisational Intangible Assets*, Topic 34: *Technology Management & Strategy*, Topic 39: *Strategies & Models for Business Operations*, Topic 43: *Business Management Approaches & Strategies*, and Topic 46: *Sustainable Development*. To a large extent, these topics correspond to the recent topics found in Figure 5.

Relating the results from the two studies can be simplified by reducing the ETM influences in the SAJIE research outputs by removing the ‘Other subjects’ category from the comparison. The ‘Other subjects’ category in itself only constitutes 0.2% of international IE trends from 2000-2007; so the omission of this category for the purpose of comparison is justified. Figure 8 shows the recalculated SAJIE percentages across the nine specific categories, pairing the 2000-2009 series from Figure 6 and Figure 7, thereby objectively comparing the recent articles of this study with the study done by Dastkhan and Owlia [7].

From Figure 8 it is evident that, aside from ETM research, the main South African IE research area in recent publication history was Production Management, followed by Operations Research and Advanced Production Systems and Technology. Internationally, however, Operations Research was the primary IE research publication area, followed by Intelligent Systems and Methods and Supply Chain Management.

The SAJIE topic spread leans toward Project Management and Production-oriented research in comparison with the relevant international trend. The proportion of South African IE publication outputs seem on a par with recent international trends in the subject areas of Information Systems and Technology, Operations Research, Supply Chain Management, and Total Quality Management.

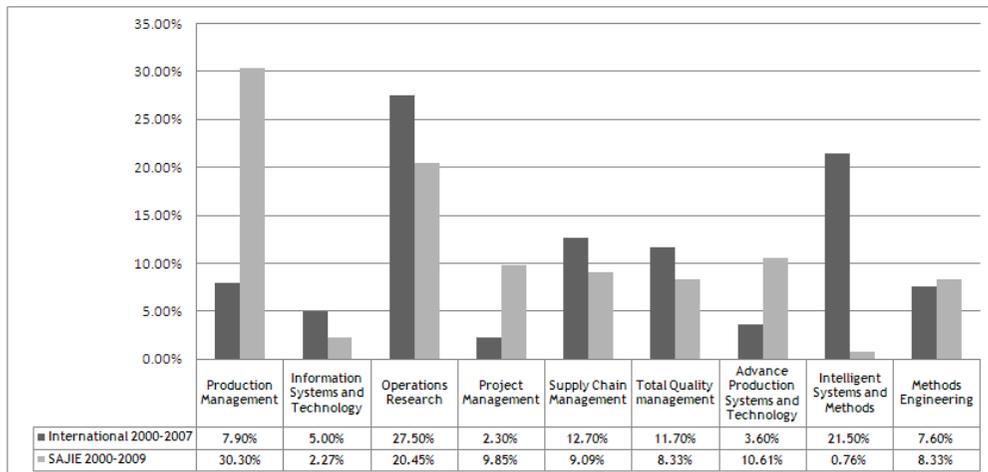


Figure 8: Proportion of research in different IE topics - SAJIE and international

The limitations of the comparative trend analysis presented in this article arise from the different approaches of the two studies, as well as differences in the scope and size of the corpora analysed in each study. The application of the Concept-Topic Modelling approach [5] (a statistical topic modelling technique that uses user-defined categories to classify a collection of textual documents) to the SAJIE article collection will be investigated in future.

4. CONCLUSION

Electronic text is a convenient and popular way to capture and store information. The amount of information available in this format expands daily. 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 - a situation that is very familiar to the industrial engineer. New tools are required for automatically organising, 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 organise 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.

In a recent article, Dastkhan and Owlia [7] reported on their study aimed at analysing the international trend of IE research for the past three decades. In an attempt to compare South African IE research trends with corresponding international trends, the authors first used the LDA technique to identify 50 research themes in the SAJIE articles published from 1987 to 2009. These topics were subsequently mapped to the categories used by Dastkhan and Owlia [7]. It was found that the SAJIE topic spread leans toward Project Management and Production-orientated research, compared with the relevant international trend. Furthermore, the proportion of South African IE publication outputs seems on a par with recent international trends in the subject areas of Information Systems and Technology, Operations Research, Supply Chain Management, and Total Quality Management. However, far less was published in South Africa on the topic of Intelligent Systems and Methods over

the past 10 years than abroad. Conversely, the subjects corresponding to Engineering and Technology Management received more attention in South African publications than abroad. These findings may be a convenient starting point for an open discussion around the future positioning of IE research in South Africa. The possible closer alignment of local IE research in support of the eight United Nations Millennium Development Goals [21], intended to be achieved by 2015, comes to mind.

The results of harnessing the full potential and power of the statistical topic modelling technique proposed in this essay will be magnified when analysing sample sizes larger than the 283 articles published by the SAJIE to date. Further research is proposed to illustrate this, with analyses of underlying trends in papers from longstanding international conferences - such as the Portland International Conference on Management of Engineering and Technology (PICMET) - being of particular interest.

5. REFERENCES

- [1] Blei, D., Griffiths, T., Jordan, M. & Tenenbaum, J. 2004. Hierarchical topic models and the nested Chinese restaurant process, *Neural Information Processing Systems Conference (NIPS)*.
- [2] Blei, D. & Lafferty, J. 2007. A correlated topic model of science, *The Annals of Applied Statistics*, 1(1), pp 17-35.
- [3] Blei, D. & Lafferty, J. 2009. Topic models, In A. Srivastava and M. Sahami, eds, *Text mining: Theory and applications*. Taylor and Francis, in press.
- [4] Blei, D., Ng, A. & Jordan, M. 2003. Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3, pp 993-1022.
- [5] Chemudugunta, C., Holloway, A., Smyth, P. & Steyvers, M. 2008. Modelling documents by combining semantic concepts with unsupervised statistical learning, *Proceedings of the International Semantic Web Conference 2008 (ISWC 2008)*, pp 229-244.
- [6] Dass, S. & Steyn, H. 2006. An exploratory assessment of project duration in multiple-project schedules where resources are allocated by the Theory of Constraints method, *South African Journal of Industrial Engineering* 17(1), pp 39-54.
- [7] 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.
- [8] Govender, S. & Steyn, H. 2008. Computerised information systems: Comparison of relative impacts on planned project duration within a multi-project environment, *South African Journal of Industrial Engineering* 19(1), pp 89-104.
- [9] 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.
- [10] Griffiths, T., Steyvers, M. & Tenenbaum, J. 2007. Topics in semantic representation, *Psychological Review*, 114(2), pp 211-244.
- [11] Le Grange, L. 2006. The changing landscape of the contemporary university, *South African Journal of Higher Education*, 20(4), pp 367-371.
- [12] Li, W., Blei, D. & McCallum, A. 2007. Nonparametric Bayes Pachinko allocation, *Conference on Uncertainty in Artificial Intelligence (UAI)*.
- [13] Li, W., & McCallum, A. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations, *International Conference on Machine Learning (ICML)*.
- [14] Lieberman, J. 2007. From Metadata to megadata: Deriving broad knowledge from document collections, Collexis, Inc., Whitepaper.
- [15] Nasukawa, T. & Nagano, T. 2001. Text analysis and knowledge mining system, *IBM Systems Journal*, vol. 40, no. 4.
- [16] Oke, S.A. & Charles-Owaba, O.E. 2005. An inflation-based maintenance scheduling model, *South African Journal of Industrial Engineering*, 16(2), 123-142.
- [17] Twala, V.G., Auf der Heyde, T., Bredell, P.J. & Pretorius, L. 2008. Developing a framework for evaluating spent nuclear fuel management options for South Africa, *South African Journal of Industrial Engineering*, 19(2), 41-60.

- [18] **Viljoen, P.J. & Steyn, H.** 2007. A conceptual model for improved project selection and prioritisation, *South African Journal of Industrial Engineering*, 18(1), pp 91-106.
- [19] **Wei, X. & Croft, W.B.** 2006. LDA-based document models for ad hoc retrieval, *Proceedings of the 29th SIGIR Conference*, pp 178-185.
- [20] www.saiie.co.za
- [21] www.undp.org.za/index.php/mdgs-in-south-africa