Investigating the Evolving Knowledge Structures in New Technology Development

J. Gopsill, P. Shakespeare, C. Snider, L. Newnes, B. Hicks

Investigating the Evolving Knowledge Structures in New Technology Development

J. A. Gopsill[1], P. Shakespeare[2], C. M. Snider[3], L. Newnes[2] and B. J. Hicks[3]

[1] University of Bath, UK
J.A.Gopsill@bath.ac.uk
[2] National Composites Centre, UK
[3] University of Bristol, UK

Abstract. The development of new technology has been identified as one of the key enablers to support business and economic growth in developed countries. For example, the United Kingdom (UK) has invested £968 Million into the creation of Catapult centres to provide 'pull through' of low Technology Readiness Level (TRL) research and science. While these Catapults have been instrumental in developing new technologies, the uptake of new technology within industry remains a considerable challenge.

One of the reasons for this is that of skills and competencies, and in particular, defining the new skills and competencies necessary to effectively apply and operate the new technology within the context of the business. Addressing this issue is non-trivial because the skills and competencies cannot be defined a priori and will evolve with the maturity of the technology. Therefore, there is a need to create methods that enable the elicitation and definition of skills and competencies that co-evolve with new technology development, and what are referred to herein as knowledge structures.

To meet this challenge, this paper reports the results from a dynamic co-word network analysis of the technical documentation from New Technology Development (NTD) programmes at the National Composites Centre (NCC). Through this analysis, emerging knowledge structures can be identified and monitored, and be used to inform industry on the skills & competencies required for a technology.

Keywords: Knowledge Management, Competency Mapping, Knowledge Structures, Graph Theory, Dynamic Network Analysis, Co-Word Analysis

1 Introduction

New Technology Development (NTD) has been identified as one of the key enablers to support business and economic growth in developed countries. This is one of the reasons that many developing countries are investing in specialist centres to support the growth of NTD. One example of this is in the United Kingdom (UK), where the government has invested £968 Million into the creation of Catapult centres to provide 'pull through' of low Technology Readiness Level (TRL) research and science [1].

adfa, p. 1, 2011.

These centres provide state-of-the-art facilities to further research and innovation, and provide a nexus for blue-skies University research to be developed into commercially viable technologies. Although the provision of state-of-the-art facilities is paramount to NTD, the uptake of NTDs within industry remains a considerable challenge. A critical barrier to this uptake is not in the access and/or cost of new equipment but in the development of the skills, competencies, and knowledge structures around NTD.

Given the largely digital nature of modern-day NTD, there now lies an opportunity to investigate how knowledge structures evolve through the dynamic co-word analysis of technological terms within NTD reports. Understanding the evolution of these knowledge structures could support the adoption of NTD through enhanced identification of the skills & competencies pertaining to an NTD and the development of best practice in ensuring the appropriate knowledge structures around NTD's are built.

To investigate this potential, this paper reports the initial findings from a dynamic co-word network analysis of a set of reports generated from NTD projects at the UK's National Composites Centre (NCC). The centre is a specialist facility aimed at developing low TRL research into high TRL commercially viable technologies.

The paper first provides an overview of dynamic co-word network analysis with a discussion of the types of insight and information that can be generated (Section Two). This is followed by a discussion of the context in which the reports have been generated and statistics of the resulting dataset (Section Three). Section Four then discusses how dynamic co-word network analysis has been applied to the dataset and the results that will be produced. This continues into Section Five, which presents the results and discusses the insights it has brought to understanding the evolving knowledge structure around NTD. The paper then concludes by highlighting the key findings, limitations and future work.

2 Dynamic co-word network analysis

Dynamic co-word analysis is the investigation of the semantic structure of a corpus of textual data through the co-occurrence of terms over time. By analysing the co- occurrence of terms, a network of connected terms (a.k.a. nodes) is generated, which enables the application of algorithms developed in graph theory to uncover underlying structures within the network and examine the nature of the connections behind the terms. For example, centrality measures are often used to identify the most important and influential terms within the network structure. In addition, clustering algorithms, such as Louvain community partitioning [2, 3], seek to identify groups of highly connected nodes within a network. In the context of co-word analysis, the clustering of terms is often referred to as the identification of topics.

The dynamism of the network comes from the continual addition of new documents. As new documents are added to the corpus, the connected nature of the terms is updated, and leads to a change in the structure of the network. The analysis of these temporal networks can reveal patterns in how knowledge around NTD evolves and matures. It is the hypothesis of this paper that NTD's that have been widely adopted will contain particular patterns in the development of the associated knowledge structure.

Identifying these patterns would then enable the development of new processes and best practice to encourage the development of appropriate knowledge structures in future NTDs.

Dynamic co-word analysis has been particularly successful in identifying research topics within scientific communities [4–6]. Whilst [7] has demonstrated the potential of the technique to support engineering project management through the monitoring of topics being discussed and potential requirements/scope creep. It is this ability of identifying and monitoring the evolution of these topics in real-time that is the current state-of-the-art within research [6].

Also, it is not only the quantitative metrics afforded by this analytical technique but the ability to aggregate and visualise a large corpus of information into a more manageable form for users to interpret and make decisions on that makes dynamic co-word network analysis an attractive proposition. Example visualisation techniques include: re-arranged matrices in relation to the clustering of the terms [8]; force-based network diagrams to reveal the connected nature of the terms [6]; and, quadrant diagrams that show the movement of clusters of terms (i.e. topics) and how their influence evolves over time [9].

It is for these reasons that dynamic co-word analysis has been selected as the technique to elicit and characterise the Knowledge Structures of NTD.

3 Context and dataset

The National Composites Centre (NCC) is a world-leading research & development centre for UK composites. Established in 2009 as a result of the UK Composite strategy, it is now part of the UK government's CATAPULT programme to develop world- leading centres designed to transform the UK's capability for NTD and help drive future economic growth. The NCC currently provides R&D support for over 40 companies.

This analysis looks at the research projects performed over a four-year period with projects typically lasting between 6-12 months. Each research project results in a set of reports detailing methods, tools and key findings pertinent to NTD. Due to the sensitivity of some of their projects, the analysis has been performed on a sub-set of documents generated from publicly funded projects. This represents approx. 20% of the total projects by the NCC. Table 1 provides a detailed breakdown on the number of final project reports produced year-on-year from publicly funded projects.

Table 1. Dataset statistics

Year	No. of Documents	Sum No. of Words	Mean Words	Max. Words	Min. Words
2012	5	31.9×10^3	6.4×10^3	9.4×10^3	3.6×10^3
2013	6	56.2×10^3	9.4×10^3	17.8×10^3	2.8×10^3
2014	7	83.8×10^3	12.0×10^3	20.4×10^3	999
2015	8	68.1×10^3	8.5×10^3	19.0×10^3	4.0×10^3
Combined	26	240.0×10^3	9.2×10^3	20.4×10^3	999

4 Dynamic co-word analysis of technical reports

In order to apply co-word analysis to full-text engineering reports, this paper applies a four-step process.

Step 1. Keyword extraction
Step 2. Co-word network generation
Step 3. Topic identification
Step 4. Dynamic analysis through Quadrant Charts

Compared to previous research where co-word analysis has been applied to keywords within academic reports and/or short texts such as e-mails, instant messaging or tweets, the application of co-word analysis on full-text documents necessitates a greater amount of pre-processing [4, 6, 9]. In particular, it is necessary to include keyword identification and extraction to identify and define topics pertaining to NTD. Following this, Step 2 covers the generation of the co-word network. The co-word network is formed of nodes that represent the key terms identified in Step 1, and edges and associated weightings representing the number of times terms co-occur.

With the terms connected to one another within a network, clustering techniques can then be applied to generate the topics relating to NTD (Step 3). These clusters are then further post-processed in Step 4 to analyse how the topics connectedness and density evolve over time. These metrics are then plotted over time using quadrant charts where insights into the knowledge structures of NTD can be drawn. Each step is now discussed in detail.

4.1 Keyword Extraction

All documents were archived in portable document format (PDF) and required parsing to extract the text in UTF-8 format. The majority of the documents were purely digital whilst a few had been scanned and required Optical Character Recognition (OCR) to parse the text. Regular expressions were used to extract the year in which the report was created and generate the list of n-grams contained within the report. The n-grams are further post-processed where any n-grams containing terms from the Natural Language Toolkit (NLTK) [10] stopwords list for the English Language as well as stopwords list for commonly used technical terms within the organisation were removed. This left 1.2×10^3 terms whose frequencies and number of documents they featured in are shown in Fig. 1.

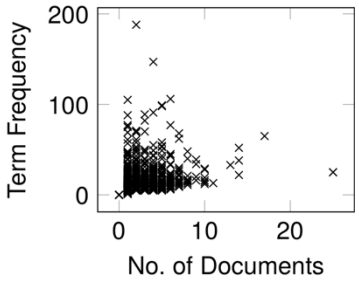

Fig. 1. Keyword statistics

4.2 Co-word network generation

With the final set of terms being identified, the document set is parsed further to identify the co-occurrence of terms within the documents. First, a network is generated

containing nodes that represent each of the terms. The process iterates through each document and identifies the number of terms that exist within the document. Edges are then made between two nodes (terms) if they exist within the same document. The edges are weighted based on the number of documents that the two terms co-occur. This is further normalised to adjust for the effects of the range of occurrences for the different terms as shown in Equation 1. Where $n_{i,j}$ is the normalised edge weighting and is determined by the number of times the two terms have co-occurred ($w_{i,j}$) divided by the minimum occurrence f_i or f_j for terms i and j respectively.

$$n_{i,j} = \frac{w_{i,j}}{\min(f_i, f_j)} \qquad (1)$$

4.3 Topic identification

As the measurements of co-occurrence are transactional and continuous, and there exists a degree of error in relating terms to one another given the length of the reports (i.e. a report may be discussing two studies where terms would be related in Step 2 but do not actually feature in the same context), Louvain community clustering algorithm has been selected as the method for identifying topics in the network.

The objective of the Louvain community algorithm is to generate a set of topic for the network that returns the highest modularity value. Modularity (Q) is an assessment of the quality of the network clustering and is defined as [11]:

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \qquad (2)$$

Where $m = 1/2\sum_{i,j} A_{ij}$ is the number of co-occurrences within the network. δ is the Kronecker delta function and is 1 if a co-occurrence exists between two models and 0 otherwise. $k_i k_j / 2m$ is the probability that a co-occurrence may exist between two terms, where k_i is the number of terms that have co-occurrences with term i and k_j is the number of terms that have co-occurrences with term j. And, A_{ij} is the normalised weighted co-occurrence between the two terms.

In order to obtain the highest modularity, the algorithm iterates between two modes. The first assigns each term to its own topic. This is then followed by the algorithm sequentially moving one term to a different topic and calculating the change in modularity. From this, the maximum modularity change can be identified.

The second mode merges the terms together to form a topic of terms and combines the co-occurrences of the terms to form single value for the co-occurrence that links the topic to the rest of the network. In addition, the edge weightings within the topic are combined to identify the strength of the internal connection within the topic. The aim is to achieve a clustering whereby each topic is highly connected internally and weakly connected to one another.

Thus, it can be considered a form of hierarchical clustering and the algorithm iterates until the modularity can no longer be increased by further aggregation of the terms. This paper uses the community API implementation of the Louvain community partitioning algorithm within the NetworkX python package [3].

4.4 Quadrant Charts

With the topics defined, one can now investigate the features of these topics and how they change over time. To achieve this, Quadrant Charts can be used. Quadrant Charts use normalised axes with lines drawn along the mid-axes to form four equal quadrants. A and third measure can be represented by altering the size of the markers on the chart (Fig. 2).

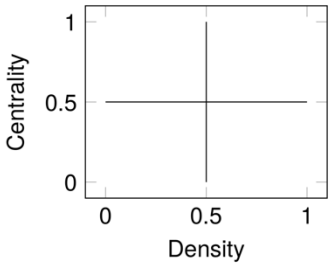

Fig. 2. Quadrant chart example

Quadrant charts have been widely used in business to support decision makers [12] and new market identification, and in network analysis to investigate potential correlations in network characteristics [6]. A particular affordance of quadrant charts is to provide a fixed space to monitor and observe the evolution and movement of partitions within a network. For example, [9] analysis of topics within an engineering project highlighted the chart's potential to provide 'actionable' information with respect to how topics are emerging, declining and/or becoming core to the projects activities. This enabled the project managers to compare their hypotheses on project activities with the actual topics being discussed in the project.

In this case, the metrics used are the cluster density and eigenvector centrality. Cluster density (D_c) is the number of edges connecting the terms within a cluster (m) divided by the possible number of edges connecting all the terms within the network ($n(n-1)/2$). Thus, 1 shows that all terms are connected with one another within the cluster and is considered a measure of how well-defined a topic is.

$$D_c = \frac{2m}{n(n-1)} \tag{3}$$

Eigenvector centrality is used to measure used to measure the influence of a cluster within the network. Relative scores are assigned to all clusters in the network based on the idea that edges to high-scoring clusters contribute more to the score of the cluster in question than equal edges to low-scoring clusters. A high eigenvector score means that a node is connected to many nodes who themselves have high scores. Equation 4 is the Eigenvector equation where $\mathbf{A}$ is the weighted co-occurrence matrix, $\boldsymbol{x}$ are the nodes and λ are the eigenvectors.

$$\mathbf{A}\boldsymbol{x} = \lambda\boldsymbol{x} \tag{4}$$

In addition, as this analysis is monitoring the development of topics over time, one can also measure the similarity of the topics from one year to the next. To do this, the terms within each topic of the previous year is compared to the terms within the topic of the current year. A ratio of the number of terms that occurs in both topics can then be generated and a full pair-wise comparison leads to a matrix of ratios for the topics. This pair-wise comparison can then indicate where the majority of the terms from the previous topics have moved to and is used to show how topics have merged and developed across the years. Table 2 provides an example of comparing the evolution of topics

from 2012 to 2013. The values in bold highlight the topics that are most similar based on the occurrence of terms.

Table 2. Pair-wise comparison of topics between years to identify mergence and growth of NTD topics

		2012 Topics					
		1	2	3	4	5	Notes
2013 Topics	a	0.00	0.04	0.03	0.03	0.17	New Topic
	b	**0.87**	**0.52**	0.00	0.06	**0.67**	Mergence of 1, 2 and 5
	c	0.07	0.07	**0.84**	0.30	0.04	Expansion of 3
	d	0.03	0.19	0.10	**0.56**	0.05	Expansion of 4
	e	0.03	0.19	0.03	0.05	0.08	New Topic
	Sum	1.00	1.01	1.00	1.00	1.01	

It can be seen that there is significant movement of terms between the topics generated in the years. Topics *a* and *e* appear to be new topics with little relation to the previous years, whilst *b* is a mergence of topics *1*, *2* and *5* from 2012. Topics *c* & *d* show expansion of topics *3* & *4*. It is these dynamic behaviours that are of interest for determining the maturity of NTD. Table 3 provides further insight into the mergence of topics *1*, *2* and *5* from 2012 to form topic *b* of 2013 as well as evidence to show the effectiveness of method to group terms. One such example of this is the term u-shape male tool, u-shape male and u-shape cut, which are all related terms to a technique used in carbon composite manufacture.

Table 3. Terms within topics

2013 - a	2012 - 1	2012 - 2	2012 - 5
u-shape male tool	**u-shape male tool**	**ceramic block material**	end effector
ceramic block material	first ply	bond line	link arms
woven preforms	**woven preforms**	carbon fibre	system integration
in-plane shear	**in-plane shear**	mechanical performance	**flexible membrane**
tool surface	**male tool**	mechanical properties	**core automation**
male tool	**u-shape cut**	**ceramic block**	**ply cutting**
u-shape cut	**fibre misalignment**	high temperature	wide range
u-shape male	**ud tape**	tests conducted	surface area
woven fabric	**woven material**	surface finish	laser system
fibre misalignment	**single ply**	block material	robot arm
flexible membrane	**u-shape male**	residual stresses	robot cell
ud tape	**woven fabric**	paste adhesive	kuka robot
core automation	4ply stack	material surface	initial concept
woven material	fibre direction	relatively low	press schemes
single ply	ambient temperature	bonded together	robotic pick
ceramic block	red line	material suppliers	dry fabric
ply cutting	elevated temperatures	epoxy filler	scoring matrix
…	…	…	…

5 Results

Table 4 reveals the development of topics over the four-year period. As one would expect, more terms are being added to the knowledge structure as more research and studies are being performed. In addition, there is an increase in the number of edges connecting the terms within the evolving network. Reviewing the network density (Table 4), it reveals that there is little change over the areas and is very low. This indicates that there is a high-level of structure relating these terms to one another.

Table 4. Dynamic network statistics

Year	No. of Terms	No. of Edges	Network Density	No. of Clusters	Modularity Score
≤ 2012	239	1.1×10^3	20×10^{-3}	7	0.644
≤ 2013	557	3.0×10^3	10×10^{-3}	15	0.660
≤ 2014	847	6.1×10^3	10×10^{-3}	16	0.638
≤ 2015	1009	8.1×10^3	10×10^{-3}	19	0.614

Further evidence of this is provided by the number of topics being generated by the analysis and the high modularity score of 0.6. Scores of > 0.3 are considered typical of highly-structured networks [13].

Continuing from the descriptive statistics, Fig. 3 shows the topics plotted on the quadrant chart of density against centrality and the node size relating to the number of terms within the topic. New topics are generated each year and the bracketed values indicate similarity with a topic in the previous year. In 2012 (Fig. 3a), it can be seen that all the topics are of density 1 highlighting that all the terms are connected with all the other terms within the topic. Given this is the starting point of the analysis, this would seem a logical finding as it is most likely that these topics will represent individual reports with little transfer of knowledge between them.

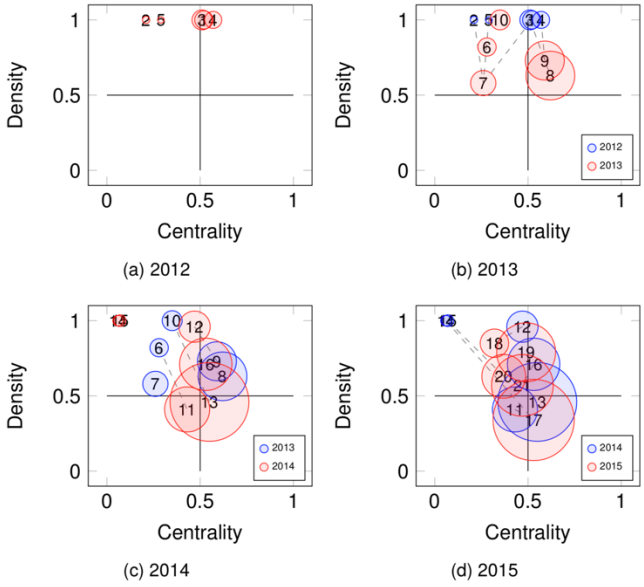

Fig. 3. Quadrant chart of evolving technology knowledge structures

Moving to 2013 (Fig. 3b) and the addition of further reports, the results show an increase in size of the topics and a decline in the density of the topics. It is interesting to note that topic 7 is a combination of 1, 2 & 5 from 2012 demonstrating there has been some work on relating these topics. The decrease in density shows that these topics have yet to be fully combined and the low centrality highlights that this topic has little influence on the rest of the NTD. In addition, a further separation of the topics based on the centrality begins to occur. The topics 8 & 9 are indicative of core NTD's as they

have grown in size and have built-up from previous work (3 & 4, respectively) whilst maintaining a high-density and centrality with other topics within the network. Topics 6 & 10 represent new topics that are being introduced into NTD.

In 2014 (Fig. 3c), it can be seen that there is the addition of new topics into NTD (14 & 15). It also appears that topics 12, 16, 11, & 13 represent existing topics as they are linked to one previous topic respectively and show continued growth through the addition of more terms. It is also interesting to see an alignment of the topics along the midline of centrality indicating that neither has more influence than the other. This may be a key feature of topics surrounding NTD.

This trend continues in 2015 (Fig. 3d) where the topics maintain an equal influence with one another. In addition, previous topics 14 & 15 have now been integrated into the main group of topics by their mergence with topics 17 & 20, respectively. At this stage of NTD, it is the density and size that are key differentiating factors. It could be that these indicate the topics that represent the NTD with the other topics representing key features of the NTD.

6 Discussion & Future Work

The results from the dynamic co-word analysis of NTD reports has highlighted the complex dynamics surrounding the generation of topics and their relations to one another. The high-level of dynamism observed shows there is potential for patterns to be identified that would relate to best practice structuring of knowledge around NTD.

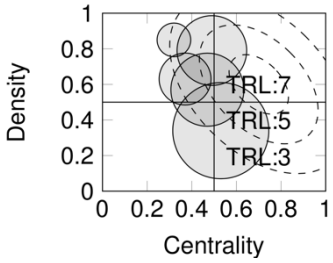

Fig. 4. Setting TRL regions based on density & centrality

However, to reach this stage, further work is required on capturing the secondary data in terms of identifying NTDs that have been easily adopted by industry. This is currently ongoing work at the NCC. With this, patterns within the co-word analysis can then be correlated to easy adoption of NTD.

In addition, it is a further hypothesised that Technology Readiness Levels (TRLs) could be mapped to these quadrant charts, which would then enable project managers to identify the readiness of their R&D (Fig. 4). The authors are actively working on this through questionnaires with experts at the NCC to qualify the TRL of different topics. The aim is to then correlate these results with the metrics from the co-word analysis.

The last aspect that is being expanded upon is in the contents of the network itself. Examples include in mapping topics to individuals, projects, companies and equipment in order to gain a better understanding of how the distribution of skills and competencies within an organisation may help or hinder the adoption of NTD.

7 Conclusion

New Technology Development (NTD) has been identified as one of the key enablers to support business and economic growth in developed countries. Facilities have now been created to ensure developed countries maintain their advantage of being at the

forefront of NTD. Although these facilities provide state-of-the-art tools and equipment, challenges exist in the development of the skills and competencies, and the body of knowledge concerning NTDs.

This paper has presented results from a dynamic co-word analysis of NTD reports within the NCC and has demonstrated the viability of this technique to provide insights into the evolution and growth of the knowledge structure surrounding NTD. The evolution of the topics concerning a matured NTD have equal influence on one another with the topic density and size being the differentiating factor on the role within NTD.

Acknowledgements

The work reported in this paper has been funded by the Engineering and Physical Sciences Research Council (EPSRC). Grant references EP/K014196/2, EP/R513556/1 & EP/R013179/1.

References

1. Catapult. https://catapult.org.uk. Accessed: 2018-02-320.
2. V. D. Blondel et al. “Fast unfolding of communities in large networks”. In: Journal of Statistical Mechanics: Theory and Experiment 2008.10 (2008), P10008.
3. A. A. Hagberg et al. “Exploring network structure, dynamics, and function using NetworkX”. In: Proceedings of the 7th Python in Science Conference (SciPy2008). Pasadena, CA USA, Aug. 2008, pp. 11–15.
4. Y. Ding et al. “Bibliometric cartography of information retrieval research by using co-word analysis”. In: Information processing & management 37.6 (2001).
5. N. Coulter et al. “Software Engineering as seen through its Research Literature: A Study in co-word Analysis”. In: JASIST 49.13 (1998).
6. Y. Liu et al. “CHI 1994-2013: mapping two decades of intellectual progress through co-word analysis”. In: Proceedings of the 32nd annual ACM conference on Human factors in computing systems. ACM. 2014, pp. 3553–3562.
7. J. Gopsill et al. “The Evolution of Terminology within a Large Distributed Engineering Project”. In: ICED. 2015.
8. J. A. Gopsill et al. “Automatic generation of design structure matrices through the evolution of product models”. In: Artificial Intelligence for Engineering Design, Analysis and Manufacturing 30.4 (2016). doi: 10.1017/S0890060416000391.
9. S. Jones et al. “Subject Lines as Sensors: Co-Word Analysis of E-Mail to Support the Management of Collaborative Engineering Work”. In: ICED. 2015.
10. E. Loper et al. “NLTK: The Natural Language Toolkit”. In: Proceedings of the ACL-02 Workshop on ETMTNLP. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 63–70. doi: 10.3115/1118108.1118117.
11. M. E. Newman. “Analysis of weighted networks”. In: PRE 70.5 (2004), p. 056131.
12. J. Vargo et al. “Crisis strategic planning for SMEs: finding the silver lining”. In: IJoPM 49.18 (2011), pp. 5619–5635. doi: 10.1080/00207543.2011.563902.
13. M. E. J. Newman. “Modularity and community structure in networks”. In: Proceedings of the National Academy of Sciences 103.23 (2006), pp. 8577–8582.