Identifying trends and flows in Communication and Information Processing by means of keyword network analysis

Dotsika, F. and Watkins, A.

This is an electronic version of a paper presented at the *3rd International Conference on Communication and Information Processing,* Tokyo, Japan, 24 to 26 November 2017.

https://dx.doi.org/10.1145/3162957.3162990

Identifying trends and flows in Communication and Information Processing by means of keyword network analysis

Dr Fefie Dotsika
Westminster Business School,
University of Westminster,
London, UK
F.E.Dotsika@westminster.ac.uk

Andrew Watkins
Birkbeck,
University of London,
London, UK
andrew@dcs.bbk.ac.uk

ABSTRACT

The purpose of this paper is to identify influential themes and knowledge flows in the area of communications and information processing and suggest trends that are likely to make (or continue making) an impact. We applied keyword network analysis on articles whose keywords match the themes of the *International Conference on Communication and Information Processing*, collected through the Thompson Reuters' Web of Science and studied the articles' thematic interconnections and their dynamics. The keyword network was found to be clustered around the themes *cloud*, *data*, *mobile*, *security*, *semantic* and *social*. *Security* and *embeddedness* are found to be the most dominant topics, common to all groups. *Design* and *performance* are key influencers of thematic flows and *data mining/analysis* are close to all nodes/keywords and therefore most popular. *Big data*, *data fusion/integration* and *e-government* are themes identified as potentially strong future influencers.

CCS Concepts

Information systems → Information systems applications → Decision support systems → Data analytics
Human-centered computing → Collaborative and social computing → Collaborative and social computing design and evaluation methods → Social network analysis

Keywords

information processing, communication processing, data analysis, network analysis

1. BACKGROUND

Looking into journal articles as a complex system where researchers interact with scientific advancements in their subject areas, their environment and with one another, we approach the researchers' scientific output as a complex, dynamic environment of multiple variables and conceptualise it as a network of interconnected themes. The complexity of articles' thematic interconnection increases as the amount of overall publications within a thematic domain multiplies.

These thematic interconnections uncover technological advancements, flows of information, clusters and hotspots which, when studied, reveal influential developments and trends with significant effects on research, productivity and innovation. The dynamics of themes extend beyond the popularity and currency of subject areas, entering thus into the domain of emergence, trending and forecasting [8], [11].

The current paper focuses on journal publications in the area of communications and information processing. The research interest in the field is widespread as this subject area provides the foundations for sustainable development and has crucial implications in all aspects of life, economy and culture [19], [20]. The publications considered were those matching the themes addressed in the *International Conference on Communication and Information Processing*. The purpose of the research carried out is to identify clusters, influential themes and knowledge flows in the area, pinpoint current trends and suggest potential future trends that are likely to make an impact.

2. METHODOLOGY AND RESEARCH DESIGN

Bibliometric analysis has been used for the detection of technological trends by means of keyword co-occurrence and network analysis [16], [23], [18]. Visualisation techniques have been applied effectively to explore themes, monitor trends and locate interrelated fields [25], [24].

Keyword network analysis, a method best known for its application on social science research [5] has also been successfully employed in bibliometrics with author keywords representing the key points/core concepts of a publication [7], [15]. Network analysis centrality measures have been found efficient in identifying key leading themes in operations management research [3]. Network visualisation techniques and knowledge maps have been effective in theme and emergent trend discovery [14], [15]. Keyword network analysis has been successfully employed in the prediction and forecasting of emergent trends [9], [11]. In addition to these, the method was deemed most appropriate here due to its use of and focus on relational information to investigate and quantify structural properties [22].

Data from academic publications whose keywords match the main themes of the *International Conference on Communication and Information Processing* were collected through the Thompson Reuters' Web of Science API. The authors' own keywords were then used to create thematic networks by linking co-occurring keywords so that they form linked pairs. As a result,

the publications within the field covered by the ICCIP conference are represented as a network (graph G= (V, E)) whose nodes (V) correspond to keywords and edges (E⊆V×V) to relationships between them. The keywords used in the search can be seen in Table 1.

Table 1. International Conference on Communication and Information Processing Keywords

Access Controls	E-Technology	Real-Time Systems
Anti-cyberterrorism	Forensics,	Resource and
Assurance of Service	Recognition	Knowledge
Biometrics	Technologies and	Discovery Using
Technologies	Applications	Social Networks
Cloud Computing	Fuzzy and Neural	Self-Organizing
Computational	Network Systems	Networks and
Intelligence	Green Computing	Networked Systems
Computer Crime	Grid Computing	Scalability and
Prevention/ Detection	Image Processing	Performance
Computer Forensics	Indexing and	Semantic Web,
Computer Security	Query Processing	Ontologies
Confidentiality	for Moving	Sensor Networks and
Protection	Objects	Social Sensing
Critical Infrastructure	Information and	Signal Processing,
Management	Data Management	Pattern Recognition
Data Compression	Information	and Applications
Data Management in	Content Security	Social Networks
Mobile Peer-to-Peer	Information Ethics	Social Search
Networks	Information	Software Engineering
Data Mining	Propagation on	Ubiquitous
Data Stream	Social Networks	Computing, Services
Processing in	Internet Modeling	and Applications
Mobile/Sensor	Mobile	User Interfaces and
Networks	Networking,	Usability Issues form
Distributed and	Mobility and	Mobile Applications
Parallel Applications	Nomadicity	User Interfaces,
Digital Information	Mobile Social	Visualization and
Processing and	Networks	Modeling
Communications	Mobile, Ad Hoc	Web Services
E-Government	and Sensor	Architecture,
E-Learning	Network	Modeling and Design
Embedded Systems	Management	Web Services
& Software	Multimedia	Security
	Computing	Wireless
	Network Security	Communications
	Peer-to-Peer	XML-Based
	Social Networks	Languages
	Quality of Service	

The structure of the created network(s) was explored by means of basic network properties and their metrics and topologies compared to existing models of similar properties. This enabled us to explore thematic relationships and flows. The tools used for the analysis and visualisation were Gephi [2] and UCINET [5].

3. DATA ANALYSIS

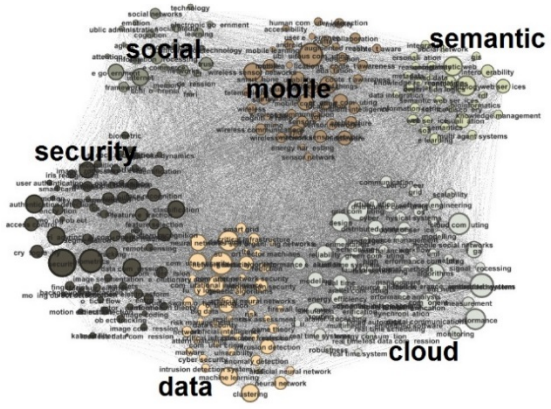
The (descriptive) basic network metrics calculated and their interpretation can be seen in Table 2. Clustering and sub-network metrics were also calculated to explore the existence of thematic clusters and their integration.

Table 2. Network metrics

	Metric	Interpretation	Value
Basic metrics	Size n=|V|	Thematic richness	284
	No of links m=|E|	Thematic connectivity	11079
	Density δ(G)=(2m/n(n−1))	Ratio of the number of keyword co-occurrences to the number of all possible co-occurrences	0.276
	Diameter	The longest of all the calculated shortest paths	3
	Average degree k=(2m/n)	Degree of a keyword is the number of edges connected to it	78.021
Clustering & sub-networks	Clustering coefficient	Ratio of existing links connecting a node's neighbours to each other to the maximum possible number of such links	0.488
	Average path length	Average number of steps along the shortest paths for all possible pairs of keywords	1.727
	Modularity	Strength of clustering calculated as the fraction of edges falling within the given groups minus the expected such fraction if edges were randomly distributed	0.395
	Communities	Thematic clusters	6

The modularity of the network is relatively high (modularity values range from -1 to 1) which suggests that there is a high strength of clustering. The short average path length and high clustering co-efficient metrics suggest that the network is of the small-world type characterised by short average path length and high clustering coefficient. Therefore the network contains interconnected thematic communities, each of which is internally strongly connected.

Based on this we used the Louvain method [4] for direct community detection and the network was found indeed to have a "community structure", that is, a clear distinction between sets of nodes densely connected to each-other and less well connected to the rest of the network. Six clear communities were identified, whose locality and main themes are depicted in Figure 1 .



cloud	25.7%
data	15.85%
mobile	16.2%
security	22.54%
semantic	10.92%
social	8.8%

Figure 1. Thematic communities

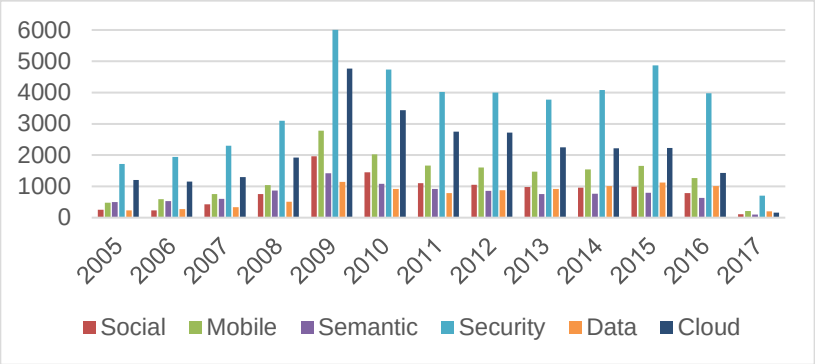


Figure 2. Thematic distribution of publications

The thematic communities revolve around six clearly defined themes: social (including social web, e-*, social media/networks, etc.), semantic (incl. semantic technologies, semantic web, ontologies, XML-based languages, etc.), data (incl. big data, data analytics, data mining, fuzzy/neural networks, etc.), mobile (incl. ubiquitous computing, P2P, mobile net. management, etc.), security (incl. cryptography/encryption, cyberterrorism, computer forensics, etc.), and cloud (incl. cloud computing, web services, grid computing, scalability etc.). The thematic distribution of number of publications (axis y) by year (2005-2017) can be seen in Figure 2.

Centrality metrics were calculated to aid the positional analysis of the network, identifying the keywords/themes that occupy distinguished positions among the nodes [22]. These were degree, closeness and betweeness centralities. Table 3 presents the metrics, their interpretation and the highest/lowest scoring keywords in each category.

Table 3. Centrality metrics

Metric	Interpretation	High scoring keywords	Low scoring keywords
Degree centrality (node i) $C_D(i) = \sum_{j=1}^{n} x_{ij}$ where x_{ij} is a tie from node *i* to node *j*	Number of links of a keyword.	security embedded systems semantic web cloud	data compression timed automata mimo (multiple i/o) schedulability
Betweenness centrality $C_B(i) = \sum g_{jk}(i)/g_{jk}, i \neq j \neq k$ where g_{jk} the no of shortest paths connecting *j* and *k* and $g_{jk}(i)$ the no of shortest paths between *j* and *k* that pass through *i*.	The number of times a keyword acts as a link along the shortest path between two other keywords.	security embedded systems performance design	data compression timed automata mobile learning description logistics
Closeness centrality $C_c(i) = \sum_{j=1}^{n} d_{ij}$ whered_{ij} is the distance connecting nodes *i* and *j*.	A keyword's distance to all other keywords.	security embedded systems data mining grid computing	data compression fmri (functional mri) description logistics moving object detection

Degree centrality is interpreted as a node's (keyword's) popularity and influence of its position in linking trends and controlling thematic flows. Closeness centrality is interpreted as the degree to which a keyword/theme is near all other keywords in the network. Betweeness is interpreted as a measurement of influence in the flow of information in the network. In order to further analyse the centrality metrics we use visualisation based on the ranking of the keywords for degree, betweeness and closeness centralities. In the graphic representations, nodes' sizes are depicted ranked according to their centrality scores: the higher the centrality, the larger the node. The results can be seen in Figure 3 (left to right: degree, betweeness and closeness).

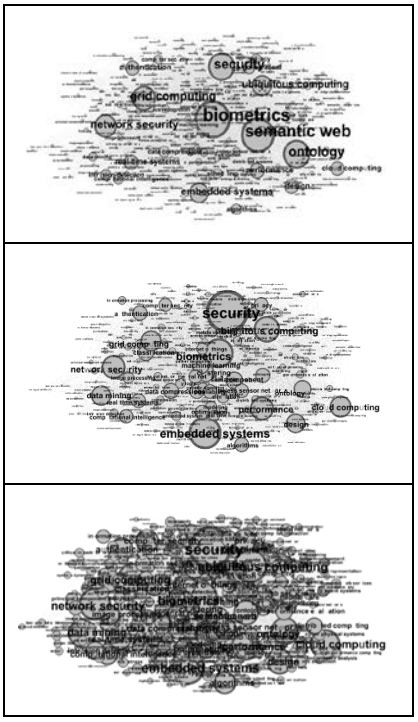


Figure 3. Degree, betweeness and closeness centralities

The degree centrality used here is weighted (weight of each edge has been taken into account). From the three graphs the closeness centrality is the most difficult to read, as the nodes overlap significantly. Instead of "spreading" the graph to avoid overlapping, we kept the format to enable cross-centrality comparisons. S*ecurity* and *embeddedness* are prominent in all graphs as expected. They represent well-established strongholds

of thematic flows. From the first graph (degree centrality) we identify b*iometrics*, *semantic web* and *ontology* as the most prominent/popular themes. In the second graph (betweeness) *security*, *biometrics* and *embedded systems* appear as the most influential themes. As expected, less connected themes of low centralities such as *data compression* and *description logistics* (identified as such in Table 3) are barely visible here. They signify topics of little influence within the group, or niche interest.

There is a special category of keywords that are identified through centrality measures that are in conflict with one another [17]. These keywords have at least one high centrality metric and one low. In our dataset, topics such as *semantic web* and *ontologies* are popular (i.e. have high degree centrality) but do not act as links of thematic domains (i.e. have low betweeness centrality). On the other hand, *machine learning* and *image processing* are influential (high betweeness) but not necessarily popular (low degree). From this special category, keywords with high closeness and low degree are central, noteworthy trends that have been recognised as either likely to keep trending, or having potentially significant future impact [11]. *Big data*, *data fusion/integration* and *e-government* have been identified as such themes here.

4. DISCUSSION AND CONCLUSIONS

Keyword network analysis has been used here to analyse knowledge flows and influential themes in a network of keywords from published articles matching the topics of the *International Conference on Communication and Information Processing*. The total of 75047 articles were collected through the Thompson Reuters' Web of Science API (publications are included as appearing at the end of June 2017). The network was created by linking co-occurring authors' keywords (a total of 281914).

The paper's implication for research is that the method was found suitable for theme analysis and for the identification of strong current and potential future trends in accordance with similar research [26]. Visualisation was possible due to the relatively small size of the network, a fact that further facilitated the study and analysis of data. Clear and distinct clusters were identified around six major themes: *cloud*, *data*, *mobile*, *security, semantic* and *social*. The topics of s*ecurity* and *embeddedness* were found to be the most dominant themes and were common to all groups. Thematic flows were found to be significantly influenced by the keywords d*esign* and *performance*. These results reflect and reinforce current research and are in line with top ranked market reports.

The topics identified as strong future influences were related to *big data*, *data fusion* and *data integration*. As before, our results reflect and are supported by recent academic research [6], [9] and industrial publications [13], [12]. *E-government* (and associated themes such as *e-government pathways*, *service innovation* and *smart cities*) was another theme that scored high as a future research trend and influencer. Research in the area follows critical developments in information processing and communication and is expected to generate significant research and innovation [21], [1].

Communication and information processing are, and will continue to be, heavily influenced by the themes of *big data*, *data fusion/integration*, and *e-government*. This has implications for practice as it creates new opportunities for timely research in these areas and identifies themes that ought to be considered, reflected and perhaps highlighted in the focus of future conferences in the field.

The results are indicative and require further research to confirm the nature of the identified trends. A potential study would be to carry out a longitudinal analysis that would help compare earlier trend identification. This is the area of future research.

5. REFERENCES

[1] Barrett, M., Davidson, E., Prabhu, J., & Vargo, S. L. (2015). Service innovation in the digital age: key contributions and future directions. *MIS quarterly*, *39*(1), 135-154.

[2] Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

[3] Behara, S., Babbar, R. S., & Andrew Smart, P. (2014). Leadership in OM research: a social network analysis of European researchers. *International Journal of Operations & Production Management*, *34*(12), 1537-1563.

[4] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

[5] Borgatti, S.P., Everett, M.G. and Freeman, L.C. (2002), Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies

[6] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171-209.

[7] Chiu, W. T., & Ho, Y. (2007). Bibliometric analysis of tsunami research. *Scientometrics*, *73*(1), 3-17.

[8] Choi, J., & Hwang, Y. S. (2014). Patent keyword network analysis for improving technology development efficiency. *Technological Forecasting and Social Change*, *83*, 170-182.

[9] Choi, J., Yi, S., & Lee, K. C. (2011b). Analysis of keyword networks in MIS research and implications for predicting knowledge evolution. *Information & Management*, *48*(8), 371-381.

[10] Dong, X. L., & Srivastava, D. (2015). Big data integration. *Synthesis Lectures on Data Management*, *7*(1), 1-198.

[11] Dotsika, F. and Watkins, A. (2017) Identifying potentially disruptive trends by means of keyword network analysis. Technological Forecasting & Social Change, 119. pp. 114-127. ISSN 0040-1625

[12] Forbes, (2016), 2017 predictions, online, https://www.forbes.com/sites/gilpress/2016/12/12/2017-predictions-for-ai-big-data-iot-cybersecurity-and-jobs-from-senior-tech-executives/#1c4c8e077a73, online, accessed July 2 2017

[13] Gartner, (2016), Gartner Survey Reveals Investment in Big Data Is Up, online, http://www.gartner.com/newsroom/id/3466117, accessed July 2 2017

[14] Kim, Y. G., Suh, J. H., & Park, S. C. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, *34*(3), 1804-1812.

[15] Lee, P. C., & Su, H. N. (2010). Investigating the structure of regional innovation system research through keyword co-occurrence and social network analysis. *Innovation*, *12*(1), 26-40.

[16] Li, H., An, H., Wang, Y., Huang, J., & Gao, X. (2016). Evolutionary features of academic articles co-keyword network and keywords co-occurrence network: Based on two-mode affiliation network. *Physica A: Statistical Mechanics and its Applications*, *450*, 657-669.

[17] Liang, Y., Chen, J., 2011. Group network centrality analysis of blogs in politics. Commun. Inform. Sci. Manag. Eng.

[18] Park, H., Kim, K., Choi, S., & Yoon, J. (2013). A patent intelligence system for strategic technology planning. *Expert Systems with Applications*, *40*(7), 2373-2390.

[19] Roy, S., Ahmed, A. M. M., & Abonamah, A. A. (2014). ICT and Economic Growth: Evidence from Twelve MENA Economies. *International Journal of Customer Relationship Marketing and Management (IJCRMM)*, *5*(1), 16-30.

[20] Schank, R. C. (2014). *Conceptual information processing* (Vol. 3). Elsevier.

[21] Snead, J. T., & Wright, E. (2014). E-government research in the United States. *Government Information Quarterly*, *31*(1), 129-136.

[22] Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

[23] Wu, C. C. (2016). Constructing a weighted keyword-based patent network approach to identify technological trends and evolution in a field of green energy: a case of biofuels. *Quality & Quantity*, *50*(1), 213-235.

[24] Yang, S., Han, R., Wolfram, D., & Zhao, Y. (2016). Visualizing the intellectual structure of information science (2006–2015): Introducing author keyword coupling analysis. *Journal of Informetrics*, *10*(1), 132-150.

[25] Yoon, B., Lee, S., & Lee, G. (2010). Development and application of a keyword-based knowledge map for effective R&D planning. *Scientometrics*, *85*(3), 803-820.

[26] Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, *15*(1), 37-50.