Manchester Metropolitan University

https://e-space.mmu.ac.uk

Extracting scientific trends by mining topics from Call for Papers

Noor Arshad and Abu Bakar
Information Technology University, Lahore, Pakistan
Saira Hanif Soroya
University of the Punjab, Lahore, Pakistan
Iqra Safder
Information Technology University, Lahore, Pakistan
Sajjad Haider
Capital University of Science & Technology, Islamabad, Pakistan
Saeed-Ul Hassan
Department of Computer Science,
Information Technology University, Lahore, Pakistan
Naif Radi Aljohani
Faculty of Computing and Information Technology, Jeddah, Saudi Arabia
Salem Alelyani
Center for Artificial Intelligence (CAI),
King Khalid University, Abha, Saudi Arabia and
College of Computer Science, King Khalid University, Abha, Saudi Arabia, and
Raheel Nawaz
Manchester Metropolitan University, Manchester, UK

Abstract

Purpose – The purpose of this paper is to present a novel approach for mining scientific trends using topics from Call for Papers (CFP). The work contributes a valuable input for researchers, academics, funding institutes and research administration departments by sharing the trends to set directions of research path.
Design/methodology/approach – The authors procure an innovative CFP data set to analyse scientific evolution and prestige of conferences that set scientific trends using scientific publications indexed in DBLP. Using the Field of Research code 804 from Australian Research Council, the authors identify 146 conferences (from 2006 to 2015) into different thematic areas by matching the terms extracted from publication titles with the Association for Computing Machinery Computing Classification System. Furthermore, the authors enrich the vocabulary of terms from the WordNet dictionary and Growbag data set. To measure the significance of terms, the authors adopt the following weighting schemas: probabilistic, gram, relative, accumulative and hierarchal.
Findings – The results indicate the rise of "big data analytics" from CFP topics in the last few years. Whereas the topics related to "privacy and security" show an exponential increase, the topics related to "semantic web" show a downfall in recent years. While analysing publication output in DBLP that matches CFP indexed in ERA Core A* to C rank conference, the authors identified that A* and A tier conferences not merely set publication trends, since B or C tier conferences target similar CFP.
Originality/value – Overall, the analyses presented in this research are prolific for the scientific community and research administrators to study research trends and better data management of digital libraries pertaining to the scientific literature.
Keywords Text mining, Topic modelling, Call for Papers (CFP), DBLP, Hot topics, Trending conferences
Paper type Research paper

1. Introduction

With the rapid growth of science and technology in recent years, there has been an explosion of electronic information published on the web in the form of millions of articles, conference papers, books and blogs (Shardlow *et al.*, 2018; Safder and Hassan, 2019). However, the sheer amount of such information makes it practically impossible for users to grasp the depth or the extent to find out the crunch of their interest (Batista-Navarro *et al.*, 2013; Nawaz *et al.*, 2013; Thompson *et al.*, 2017; Jahangir *et al.*, 2017). Generally, the scientists and decision-making communities are more concerned with the advanced information monitoring systems to determine their right goals and the best decisions to take at a right stage (Nawaz *et al.*, 2012; Ananiadou *et al.*, 2013). However, the timely decision-making can be accomplished by keeping themselves conversant with the ever-changing scientific trends, which has become nearly impossible due to an exponential increase in scholarly publications. Therefore, there is a need to design a system that can automatically discover these relevant and ever-changing exciting research trends to support researchers, decision makers and scientific communities. In recent years, an attempt, termed as Topic Detection and Tracking (TDT), is made to find the solution for the problem of "well-awareness" on this dynamic data (Yeh *et al.*, 2016). For topic detection, we have used TDT theme in order to identify the topics of conferences from their calls and corresponding topics from DBLP using Association for Computing Machinery (ACM) classifications.

Since the scientific topics of interest are announced in Call for Papers (CFP) of conferences managed by scientist and researchers almost every year, this CFP is also considered as the theme of the conference for that year, and the researchers only consider CFP for those conferences in which they have aimed to submit their research. In this paper, we have mined CFP topics to detect the hot topics, ever-changing research trends and prestige of a conference that set these research trends.

The major contributions of this paper are as follows:

- The first contribution of this work is the compilation of CFP data set. We manually collected CFP for 146 conferences mapped with data format field (Field of Research (FOR): 804) code for the last 12 years (on average). It should be noted that all these conferences are ranked by Australian Research Council (ARC) from Core A* to Core C rank. To the best of our knowledge, we are the first to compile this kind of data set and to use it for hot topics' evaluation.

- The second contribution is the identification of topics and keywords from CFP corpus. We rigorously evaluated these topics and keywords on temporal basis to find trending topics.

- The third and the main contribution is to measure the impact of extracted topics from CFP. Using ACM Computing Classification System (CCS), we mapped 1.3m publications indexed by DBLP to related conferences into their thematic areas by matching the keywords appeared in their titles.

The rest of the paper is organized as follows: the upcoming section presents related studies, followed by the data and methods employed. Next, the results are discussed succeeded by the concluding remarks.

2. Related work

There are millions of scientific articles being published to digital archives on a monthly basis (Khabsa and Giles, 2014; Olson *et al.*, 2015; Safder and Hassan, 2018). Thus, it is a challenging task for researchers and scientists to mine the trends and topics from this large number of scientific documents (Hofmann, 2017; Hassan *et al.*, 2018; Wang *et al.*, 2011).

In past years, researchers have proposed different kinds of methods for textual topic detection (Al-Yahya, 2018; Petkos, Papadopoulos and Kompatsiaris, 2014; Safder *et al.*, 2018). Normally, topic detection falls under three categories, namely, document-pivot method,

feature-pivot method and probabilistic topics models (Petkos, Papadopoulos and Kompatsiaris, 2014; Petkos, Papadopoulos, Aiello, Skraba and Kompatsiaris, 2014), where Wu *et al.* (2008) presented a document-pivot method to cluster documents on the basis of terms' similarity using the TF–IDF weighting schemes. Similarly, the feature-pivot method is proposed to build cluster of terms on the basis of the terms' co-occurrence patterns. However, the probabilistic topics modelling techniques are used to represent the joint distribution of topics and terms using the generative probabilistic model (Fang *et al.*, 2018).

Furthermore, Hofmann (2017) used PLSI (Probabilistic Latent Semantic Indexing) technique for automated indexing of the document by using the probability of word weights in addition to the Latent Semantic Analysis (LSA) approach. In this work, co-occurred frequent terms (TF–IDF) are utilized for retrieval of similar documents by catering the weights of the related terms. Although LSA has shown noise reduction and success in many different domains for the document indexing, it lacks the satisfactory statistical foundation. Therefore, Hofmann (2017) added statistical foundation via PLSI, which cannot generate new documents that are not available in the training stage. Thus, in order to address this limitation, Latent Dirichlet Allocation (LDA) by Blei *et al.* (2003) was proposed to introduce a Dirichlet prior for the topic distribution of the documents. Furthermore, Bernoulli Process Topic (BPT) model was developed as a general framework in the sense that LDA is special case of BPT. BPT is used to discover latent topics from a corpus of documents at the citation level; it incorporates the link information present in the corpus to model the relationship among the documents (Guo *et al.*, 2009; Kou *et al.*, 2015).

Similarly, a reviewer assignment system (RAS) was demonstrated by (Patel *et al.*, 2011) to automatically extract the profiles of reviewers and their submissions in the form of topic vectors (Patel *et al.*, 2011). Next, these profiles are used to automatically assign reviewers to papers without relying on a bidding process. The RAS also includes the assignment model that maximizes, for each paper, the coverage of its topics by the profiles of its reviewers.

A document analysis tool using topic hierarchy and context-based document analysis tool is proposed by Chen *et al.* (2016), which allows the users to explore any multi-topic document based on fine-grained and hierarchical topics automatically mined from it. The supply chain management has been applied on a structured literature review from 1991 to 2015, published in eight academic journals, to understand important insights (Swanson *et al.*, 2018).

The state of the art describes the way to topic model; however, no one has applied it in CFP data set for topic classification along with the mapping under ACM classes. First, we have indexed the CFP data set for finding the research trends by mapping it with the ACM classes. Second, for topic clustering/document classification of CFP and DBLP publications, we have utilized ACM classification by matching the weighted keywords (probability weights, gram depth weights, relative weights and hierarchical weights) explained in the methodology section, since both our data source (CFP) and methodology of classification using the ACM standard with the help of weighted keywords are different to the best of our knowledge.

2.1 Trending topics from news and social media

Unsupervised approach has been used by Zhang *et al.* (2017) to generate stories of evolutionary topics in news and Twitter data sources using an incremental algorithm based on the alternative direction method of multipliers. TF–IDF scheme was used to extract topics by dividing the feature sets into shared features, news features and twitter features. To track the stories, topics were categorized into emerging, evolving and fading ones.

The user-generated contents are collected from microblogs and sub-topics are extracted to analyse emotion of users of the event using a topic model given by Zhou and Zhang (2017) and Hassan and Haddawy (2015). They classified subjective microblogs by finding the adjectives and applied LDA for sub-topic model along with the use of weighted frequent terms (TF–IDF and TF–RDF). Here, the work discussed in the context of social media topic

modelling resembles our methodology in using the weighted frequent terms, but it does not use the gram depth weights and hierarchical weights.

In recent years, a topic modelling approach has been designed to treat tweets as semi-structured text. A novel hashtag graph-based topic model has been proposed to discover the topics of tweets. This model utilizes hashtag relation information in hashtag graphs and potent to understand the word semantic relations, even if the words have not co-occurred in a tweet (Wang *et al.*, 2016).

Hassan *et al.* (2014) found a very useful keyword-based approach and utilized it for mining the research trends in sustainable development at the country and institute level. Direct citation clustering and co-citation threading-based models have been utilized to identify the emerging topics for helping the decision makers (Small *et al.*, 2014). Keywords collected from domain experts (seed keywords), author-defined keywords, publication titles and abstracts have been used for bibliometric analysis.

The related work discussed above shows that the topic detection and trend analysis from the news document and social media have been done using the document clustering, TF–IDF, terms' co-occurrence patterns and probabilistic techniques. We have used the same probabilistic keyword-based technique, but our work for mining trends is based on the conferences called topics in relation to the publications, which has not been studied earlier in this context.

3. Data and methods

In this section, we discussed our data sets and proposed methodology for this comprehensive research trends study.

Kindly note that the data and code used in this study can be downloaded from: https://github.com/slabitu/research_trends/

3.1 Data sets

3.1.1 Call for Papers data set. Generally, CFP is considered as the theme of the conference decided by the board members of the conference. In order to discover the relation of trending topics with CFP, we compiled a temporal CFP corpus of 146 data format conferences (FOR: 804) ranked by the ARC (www.universityrankings.com.au/research-excellence-rankings.html). We manually collected CFP topics by exploring different URLs over the web search and built a CFP corpus of conferences with ranking A* to C. It is noteworthy that we only examined CFP topics over the period of last 12 years (on average). However, for some conferences, CFP data are present for the window of 1993–2014. The CFP corpus is the main and novel data set for our proposed research. We also maintained different characteristics for CFP corpus such as basic information of conferences with their ranks according to CORE2014, history of conferences with location and date, and where and when conferences were held. Also, CFP topics are also listed for each historical event.

3.1.2 DBLP data set. This data set contains 1.3m publications related to computer science conferences, originally downloaded from DBLP bibliography. DBLP data set is available in the form of XML file (http://dblp.org/xml/). DBLP does not include the field of keywords of the documents, and the title of the article represents the article content's description. Therefore, we use the title's keywords for topic identification from the DBLP articles data set. The "Year" field of DBLP data set (from 1936 to 2016) has also been extracted for comparison with CFP years.

3.1.3 ACM Computing Classification System data set. ACM CCS 2012 is built as poly-hierarchal ontology for standard classification system in the computing field. There are 11 top-level classes in this data set, from A to K, as given in Table I, and all 1,474 classes including top-level classes are distributed in four levels of hierarchy. We classify the CFP and DBLP articles against the ACM classification using the high weighted keywords by summing the probabilistic weights along with the TF–IDF weights.

| Node ID	Label
A	General Literature
B	Hardware
C	Computer Systems Organization
D	Software
E	Data
F	Theory of Computation
G	Mathematics of Computing
H	Information Systems
I	Computing Methodologies
J	Computer Applications
K	Computing Milieux

3.2 Approach

This section contains the details of our proposed methodology. Figure 1 shows the high-level architecture of our proposed system for research trends discovery. First, the model extracts the publication's metadata information such as article title and year for each paper in DBLP data set. Meanwhile, we also extracted topic keywords with year details from CFP data set. Furthermore, the pre-processing techniques of text mining like stop words, punctuation marks' removal and stemming/lemmatization are applied on both titles of DBLP articles and CFP topics by using Natural Language Processing Toolkit in Python.

Afterwards, the system extracts the keywords up to four grams from both DBLP titles and CFP topics. In order to enrich the keywords dictionary with their synonyms words, we employed WordNet and Growbag data set (Diederich and Balke, 2007). The Growbag data set also helped to add related/co-occurred terms against the keywords. The duplicate entries were removed from the data sets before further processing. Finally, the system classified the DBLP articles and CFP topics by employing ACM CCS using their titles, and a similar method was employed for topic assignment.

The following sub-sections contain the details of different weighting schemas, that is probabilistic, gram, relative, accumulative and hierarchal, to measure the significance of terms.

3.3 Keyword weights

In order to measure the importance and impact of the keyword, we deployed multiple weight schemes. We carefully assigned a weight to each keyword on the basis of gram and relevance.
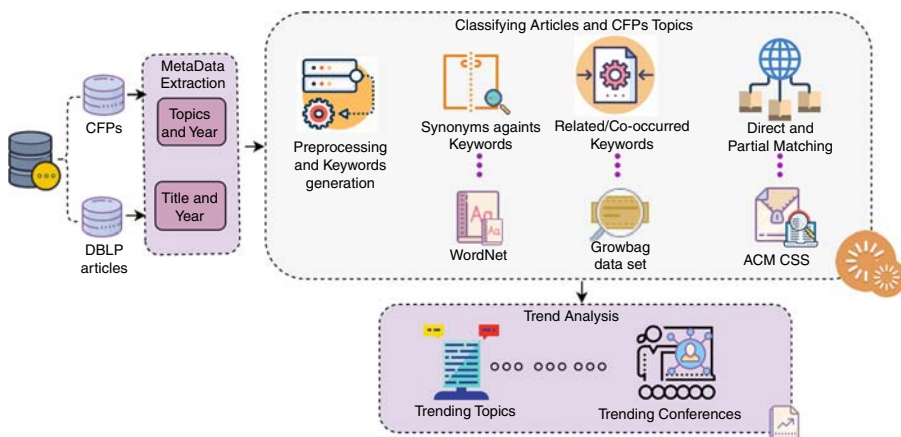
Since we used the probabilistic topic detection method, probability weights are assigned to ACM CCS tokens. TF–IDF is also calculated for each keyword extracted from article's titles and CFP topics. An example of weights is given in Table AI.

3.3.1 Probabilistic weights. In order to map the DBLP article titles and CFP topics to a related class, we have assigned probability score to each keyword of ACM CCS label. These probabilistic weights are calculated, as presented in the following equation:

$$\text{Probability of } w \text{ in } d = \text{frequency of } w \text{ in } d/\text{frequency of } w \text{ in } D, \tag{1}$$

where w is the keyword from ACM CCS class, d is the class label from which keyword is extracted and D is the set of all class labels from ACM CCS.

3.3.2 Gram depth weights. Gram depth is the number of words in a keyword: the higher the value of grams, the higher is the weight of the keyword, as shown in the following equation:

$$G1 < G2 < G3 < G4, \tag{2}$$

where G1 is for uni-gram and G4 is for four grams. We also used full-text search using CONTAINS function to match n-gram keywords, where $n > 1$. For instance, by using CONTAINS function, we can match "knowledge discovery" and "discovery of knowledge" keywords interchangeably. The sample tokenization for a topic is given as follows:

- CFP topic example: foundations and principles of data mining;
- uni-gram tokens: foundations, principles, data, mining;
- bi-gram tokens: foundations principles, principles data, data mining;
- tri-gram tokens: foundations principles data, principles data mining; and
- four-gram tokens: foundations principles data mining.

3.3.3 Relative weights. The keywords that are matched directly from DBLP articles title or CFP topics have more weights than synonyms and semantically related/co-occurred terms from WordNet library and Growbag data set. Thus, the precedence is mapped, as shown in the following equation:

$$R1 > R2 > R3, \tag{3}$$

where R1 is weight for direct keywords, R2 for synonyms and R3 for related terms.

3.3.4 Cumulative weight. Since the multiple keywords can be matched to the same class in ACM CCS and a keyword may have multiple weights, we deployed the following score summation scheme, as shown in the following equation:

$$w_c = \sum_{i=1}^{n} \sum_{j=1}^{t} w_j, \tag{4}$$

where w_c is the cumulative weight, n denotes the number of keywords matched, t represents the type of weight (like gram depth) and w_j is the keyword weight for the selected type.

3.3.5 Hierarchical weight. The ACM CCS implements the classification on the hierarchically structured data. Therefore, to avoid the situations in which "no keyword is matched in the top hierarchy but multiple keywords are matched with a sub-class in the lower hierarchy" or "there is a possibility that very few keywords can get matched with a class in the low hierarchy but more keywords are matched in top hierarchy", we need to apply some prioritization mechanism for these cases, as keywords matching in second case

should have more weight. The formula to calculate the hierarchical weights of the class is shown in the following equation:

$$w_h = \sum_{i=1}^{n} w_{c_i}, \qquad (5)$$

where w_h is the hierarchical weight, n shows the level of class in the hierarchy and w_c is the cumulative weight. It should be noted that, using hierarchical weights, most of the results are classified as lower levels in the hierarchical tree of the ACM CCS, since the weights from upper level classes are cumulated in lower level classes.

3.4 Classification
We have used the naive-Bayes supervised methodology for topic assignment on the basis of ACM CCS. First, we have assigned the topics to the titles of CFP and DBLP articles by using the ACM classification, based on the weighted n-gram keywords, probabilistic model and the nearest neighbour techniques by applying LDA methodology. Moreover, the classes are ranked with respect to hierarchical weights and cumulative weights for both titles of DBLP articles and CFP topics by using multiple ranking thresholds.

As LDA is probabilistic model for a set of documents, which are represented as a group of latent topics, with each topic being distributed over keywords, in this model, a fixed set of topics is specified before data are generated (in our case it is ACM CCS, the fixed set of topics); it represents documents as collection of topics that separate keywords with their probabilities. Higher probability means it is more likely to be similar to that topic. The LDA method maps all the documents to the topics in a way that the keywords in each document are mostly captured by a given fixed set of topics. We have assigned topics to the titles of CFP and DBLP articles using cumulative and hierarchical weights by applying nearest neighbour techniques. Weights of keywords are cumulated against ACM CCS and then ranked for classification.

3.4.1 Evaluation. We have evaluated our model using ACM data set given by Santos and Rodrigues (2009); there are 86,116 ACM publications in this data set, and 54,994 publications are classified by using keywords. Moreover, these keywords are extracted from title, abstract, keywords and general terms. We have applied our classification model on the same data set and achieved 77.3 per cent similar results as compared to given classification by Santos and Rodrigues (2009).

3.5 Trend detection
This section describes our trend detection mechanism. As illustrated in Figure 1, there are two types of trend identifiers: identification of trending topics from DBLP articles and the discovery of trending conferences. For the identification of trending topics based on DBLP articles, the trending topics are ranked using the frequency of publications against the topics identified on a temporal basis. The trending conferences are identified by linking trending topics against the conferences having the same interest in the same period of time.

4. Results and discussion
In this section, the achieved results are illustrated in detail. The following sub-sections contain the discussion related to trending topics based on CFP and DBLP articles. The arguments for the trendsetter conferences have also been added, based on the trending topics.

4.1 Analyses of CFP
This section illustrates the CFP hot topics trends appearing in multiple conferences.

Figure A1[1] demonstrates the hot topic trends constantly appearing in multiple conferences' CFP from the year 2006 to 2015. The frequency bars are showing how many times a topic appears in multiple conferences (year wise). The results clearly depict that many times the same topics have appeared in the same conference over the years. For example, the topic "data mining" in the top bar of Figure A1 shows that it appeared in more than 350 conferences in 10 years (2006–2015 time window). Each colour (from dark blue = 2015 to dark green = 2006) in the bar differentiates the year, and length of each coloured part denotes the number of conferences. The same topic is repeated in the same/different conferences over the years. Moreover, CFP topical segmentations reveal that the two main categories of ACM, namely, C (Computer Systems Organization) and H (Information Systems), have remained under the focus of CFP.

Similarly, the results of most frequent bi-gram keywords from CFP topics are shown in Figures A2 and A3. Figure A2 represents the overall frequency of keywords and the number of conferences in which a keyword was being used in their "topics of interest (CFP)". Likewise, Figure A3 demonstrates the same results on the temporal basis from 2006 to 2017. In both figures, "Data Mining" research trend has appeared as dominant topic because the CFP data set is compiled only from data format-related conferences (FOR: 804). It is also evident from the graph that the most popular topics in addition to "Data Mining" are "Data Warehouse", "Privacy and Security", "Data Management" and "Semantic Web".

In Figure A3, it is observed that the term "Big Data" became frequent after 2011. It is also observed that the rise of the term "Big Data" in 2012–2013 pushed the popularity of the term "Data Warehouse" down during 2012–2017 along with the decline of another hot term "Data Mining" during this era. Also, the same trend has been observed in the case of "Semantic Web" term. It should be noted that the term "Big Data" appeared in 2012–13, and it immediately received attention from research community. Moreover, the term "Big Data" has received consistent attention of researchers since 2012. Along with this, another hot area, "Security and Privacy", is constantly attracting researchers since 2006.

4.2 Analyses of DBLP

In order to compare our CFP trending keywords results with some benchmark, we used DBLP data set for finding hot topics with respect to the number of publications on a temporal basis. In Figure A4, hot topics are shown from 2006 to 2015. The term "Modelling Methodologies" under "Computing Methodologies" appeared as the most dominant topic. However, we ignored this topic in our results because modelling is applicable in almost every field and it was stretching down other results. Topics are also filtered with respect to CFP trending topics.

We can clearly observe that the Computer Systems Organization (SO) is dominant with 5 topics out of 13 top topics. Likewise, another prominent topic is from Information Systems (IS), namely, "information filtering".

4.3 Conferences impact on publications

In order to identify conference impact on publications, we combined trending topics from DBLP data set and conference topics from CFP. Figure A5 represents all those conferences that are calling the same topics and having a high number of articles in DBLP corpus. It is worth mentioning here that we have shown only those conferences that contain at least seven hot topics in their CFP from 2006 to 2015 and have at least 45K articles against these topics. It has been observed that EJC, which is a "C" ranked conference, comprises 14 trending topics with 110K publications in the last 10 years. However, SDM, which is an "A" ranked conference, captures only seven topics, but the number of publications against its topics is almost the same. Moreover, trending conferences and hot topics are also combined in Figure A6 where different colour bars are representing the number of conferences that are

calling the same hot topic in a year. The bar graph shows that trending topics are appearing in multiple conferences in the same year. Figures A7 and A8 are demonstrating the trending conferences against the top 9 trending topics with their number of articles against each year from 2006 to 2015. It has been noticed that the same topics are coming from all "A" to "C" ranked conferences. We can infer from these graphs that the conference VLDB encompassed the trending topics over the last 10 years majorly as compared to all other conferences of A* and A categories. Similarly, the vertical analysis of the graph shows that MobiDE, a C-ranked conference, has covered most of the trending topics continuously over the span of last 10 years, and almost all the trending topics are coming in CFP list of MobiDE every year. It should be noted that the topics covered by some conferences are also less prominent in publications set. These graphs also illustrate the contribution of conferences from all ranks to the respective FOR.

5. Concluding remarks
In this paper, we mine scientific trends based on conference CFP topics and DBLP publications. We show rise of "big data analytics" in CFP topics in recent years; in contrast, topics such as "semantic web" and "intrusion detection" show downfall. Findings of the study also confirmed that top tier conferences not necessarily set research trends. Overall, the analyses presented in this research are vital for the scientific community and research administrators to study research trends for better data management of digital libraries pertaining to scientific literature. We also believe that analysing scientific trends using CFP data sets could be a better way that will help the early career researchers to select more relevant research topics. Last but not the least, the proposed approach could help to identify the trending conferences with respect to the contribution in emerging topics. Moreover, the presented approach allows researchers to analyse in-depth calls to conferences, thematic analysis, etc. It also provides knowledge regarding the different themes that occur in conferences.

Note
1. Figures related to results (Figures A1–A8) are given in Appendix 1.

References

Al-Yahya, M. (2018), "Stylometric analysis of classical Arabic texts for genre detection", *The Electronic Library*, Vol. 36 No. 5, pp. 842-855.

Ananiadou, S., Thompson, P. and Nawaz, R. (2013), "Enhancing search: events and their discourse context", in Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing*, CICLing 2013, Lecture Notes in Computer Science, Vol 7817, Springer, Berlin and Heidelberg.

Batista-Navarro, R.T., Kontonatsios, G., Mihăilă, C., Thompson, P., Rak, R., Nawaz, R., Korkontzelos, I. and Ananiadou, S. (2013), "Facilitating the analysis of discourse phenomena in an interoperable NLP platform", in Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing*, CICLing 2013, Lecture Notes in Computer Science, Vol. 7816, Springer, Berlin and Heidelberg.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, Vol. 3, January, pp. 993-1022.

Chen, J., Wang, T.T. and Lu, Q. (2016), "THC-DAT: a document analysis tool based on topic hierarchy and context information", *Library Hi Tech*, Vol. 34 No. 1, pp. 64-86.

Diederich, J. and Balke, W.T. (2007), "The semantic Growbag algorithm: automatically deriving categorization systems", in Kovács, L., Fuhr, N. and Meghini, C. (Eds), *Research and Advanced Technology for Digital Libraries*, ECDL 2007, Lecture Notes in Computer Science, Vol. 4675, Springer, Berlin and Heidelberg.

Fang, D., Yang, H., Gao, B. and Li, X. (2018), "Discovering research topics from library electronic references using Latent Dirichlet Allocation", *Library Hi Tech*, Vol. 36 No. 3, pp. 400-410.

Guo, Z., Zhang, Z., Zhu, S., Chi, Y. and Gong, Y. (2009), "Knowledge discovery from citation networks", *Ninth IEEE International Conference on Data Mining*, pp. 800-805.

Hassan, S.U. and Haddawy, P. (2015), "Analyzing knowledge flows of scientific literature through semantic links: a case study in the field of energy", *Scientometrics*, Vol. 103 No. 1, pp. 33-46.

Hassan, S.-U., Haddawy, P. and Zhu, J. (2014), "A bibliometric study of the world's research activity in sustainable development and its sub-areas using scientific literature", *Scientometrics*, Vol. 99 No. 2, pp. 549-579.

Hassan, S.U., Safder, I., Akram, A. and Kamiran, F. (2018), "A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis", *Scientometrics*, Vol. 116 No. 2, pp. 973-996.

Hofmann, T. (2017), "Probabilistic Latent Semantic Indexing", ACM SIGIR Forum, pp. 211-218.

Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K. and Nawaz, R. (2017), "An expert system for diabetes prediction using auto tuned multi-layer perceptron", *2017 Intelligent Systems Conference (IntelliSys), IEEE, September*, pp. 722-728.

Khabsa, M. and Giles, C.L. (2014), "The number of scholarly documents on the public web", *PLoS One*, Vol. 9 No. 5, p. e93949.

Kou, N.M., Hou, U.L., Mamoulis, N., Li, Y., Li, Y. and Gong, Z. (2015), "A topic-based reviewer assignment system", *Proceedings of the VLDB Endowment*, Vol. 8 No. 12, pp. 1852-1855.

Nawaz, R., Thompson, P. and Ananiadou, S. (2012), "Identification of manner in bio-events", *LREC*, pp. 3505-3510.

Nawaz, R., Thompson, P. and Ananiadou, S. (2013), "Negated bio-events: analysis and identification", *BMC Bioinformatics*, Vol. 14, p. 14, doi: 10.1186/1471-2105-14-14.

Olson, N., Nolin, J.M. and Nelhans, G. (2015), "Semantic web, ubiquitous computing, or Internet of Things? A macro-analysis of scholarly publications", *Journal of Documentation*, Vol. 71 No. 5, pp. 884-916.

Patel, A., Bakhtiyari, K. and Taghavi, M. (2011), "Evaluation of cheating detection methods in academic writings", *Library Hi Tech*, Vol. 29 No. 4, pp. 623-640.

Petkos, G., Papadopoulos, S. and Kompatsiaris, Y. (2014), "Two-level message clustering for topic detection in Twitter", NOW-DC@ WWW, pp. 49-56.

Petkos, G., Papadopoulos, S., Aiello, L., Skraba, R. and Kompatsiaris, Y. (2014), "A soft frequent pattern mining approach for textual topic detection", *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14), New York, NY*, 10pp., available at: https://doi.org/10.1145/2611040.2611068

Safder, I. and Hassan, S.U. (2018), "DS4A: deep search system for algorithms from full-text scholarly big data", *2018 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, November*, pp. 1308-1315.

Safder, I. and Hassan, S.U. (2019), "Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications", *Scientometrics*, Vol. 119 No. 1, pp. 257-277.

Safder, I., Hassan, S.U. and Aljohani, N.R. (2018), "AI cognition in searching for relevant knowledge from scholarly big data, using a multi-layer perceptron and recurrent convolutional neural network model", *Companion of the Web Conference 2018 on the Web Conference 2018, International World Wide Web Conferences Steering Committee, April*, pp. 251-258.

Santos, A.P. and Rodrigues, F. (2009), "Multi-label hierarchical text classification using the ACM taxonomy", *14th Portuguese Conference on Artificial Intelligence (EPIA)*, pp. 553-564.

Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J. and Ananiadou, S. (2018), "Identification of research hypotheses and new knowledge from scientific literature", *BMC Medical Informatics and Decision Making*, Vol. 18, p. 46, doi: 10.1186/s12911-018-0639-1.

Small, H., Boyack, K.W. and Klavans, R. (2014), "Identifying emerging topics in science and technology", *Research Policy*, Vol. 43 No. 8, pp. 1450-1467.

Swanson, D., Goel, L., Francisco, K. and Stock, J. (2018), "An analysis of supply chain management research by topic", *Supply Chain Management: An International Journal*, Vol. 12 No. 3, pp. 100-116.

Thompson, P., Nawaz, R., McNaught, J. and Ananiadou, S. (2017), "Enriching news events with meta-knowledge information", *Language Resources and Evaluation*, Vol. 51 No. 2, pp. 409-438.

Wang, X., Rak, R., Restificar, A., Nobata, C., Rupp, C.J., Batista-Navarro, R.T.B., Nawaz, R. and Ananiadou, S. (2011), "Detecting experimental techniques and selecting relevant documents for protein-protein interactions from biomedical literature", *BMC Bioinformatics*, Vol. 12, p. S11, doi: 10.1186/1471-2105-12-S8-S11.

Wang, Y., Liu, J., Huang, Y. and Feng, X. (2016), "Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28 No. 7, pp. 1919-1933.

Wu, H.C., Luk, R.W.P., Wong, K.F. and Kwok, K.L. (2008), "Interpreting TF-IDF term weights as making relevance decisions", *ACM Transactions on Information Systems*, Vol. 26 No. 3, 37pp., available at: https://doi.org/10.1145/1361684.1361686

Yeh, J.-F., Tan, Y.-S. and Lee, C.-H. (2016), "Topic detection and tracking for conversational content by using conceptual dynamic Latent Dirichlet Allocation", *Neurocomputing*, Vol. 216 No. 1, pp. 310-318.

Zhang, X., Zhao, L., Chen, Z., Boedihardjo, A.P., Dai, J. and Lu, C.T. (2017), "Trendi: tracking stories in news and microblogs via emerging, evolving and fading topics", *2017 IEEE International Conference on Big Data (Big Data), IEEE, December*, pp. 1590-1599.

Zhou, Q. and Zhang, C. (2017), "Emotion evolutions of sub-topics about popular events on microblogs", *The Electronic Library*, Vol. 35 No. 4, pp. 770-782.
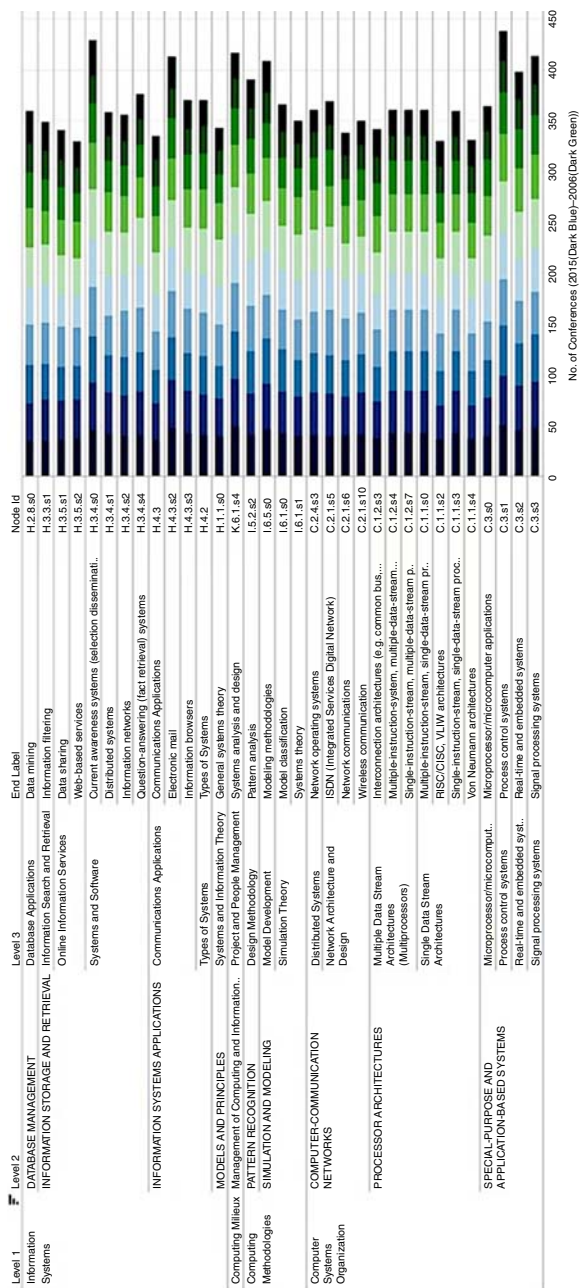
Further reading

Zhai, C. and Lafferty, J. (2017), "A study of smoothing methods for language models applied to *ad hoc* information retrieval", *ACM SIGIR Forum*, Vol. 51 No. 2, pp. 268-276
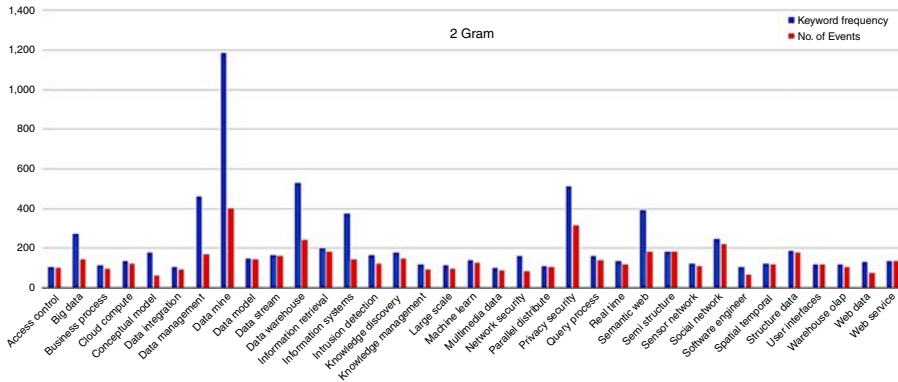
Corresponding author

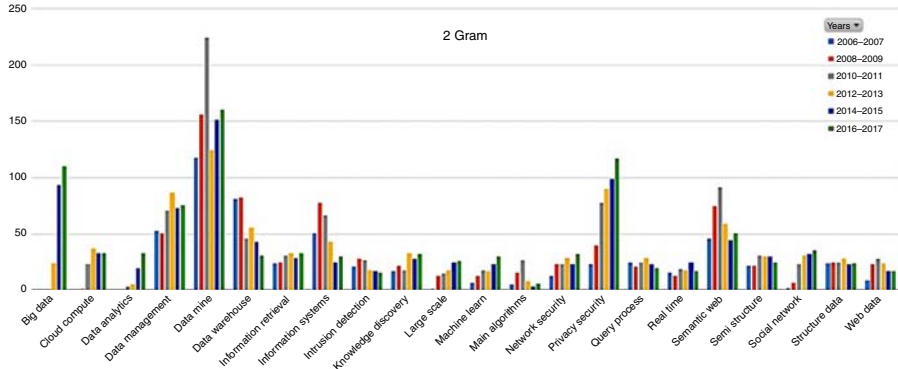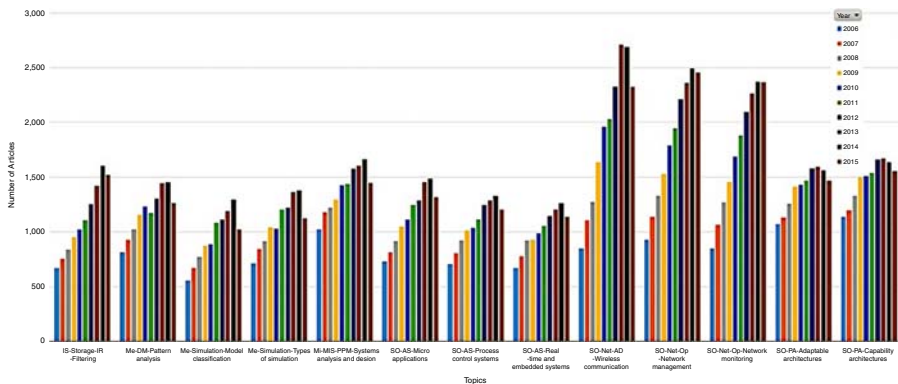Saeed-Ul Hassan can be contacted at: saeed-ul-hassan@itu.edu.pk

Appendix 1

Figure A1.
Top topics of CFP – from 2006 to 2015
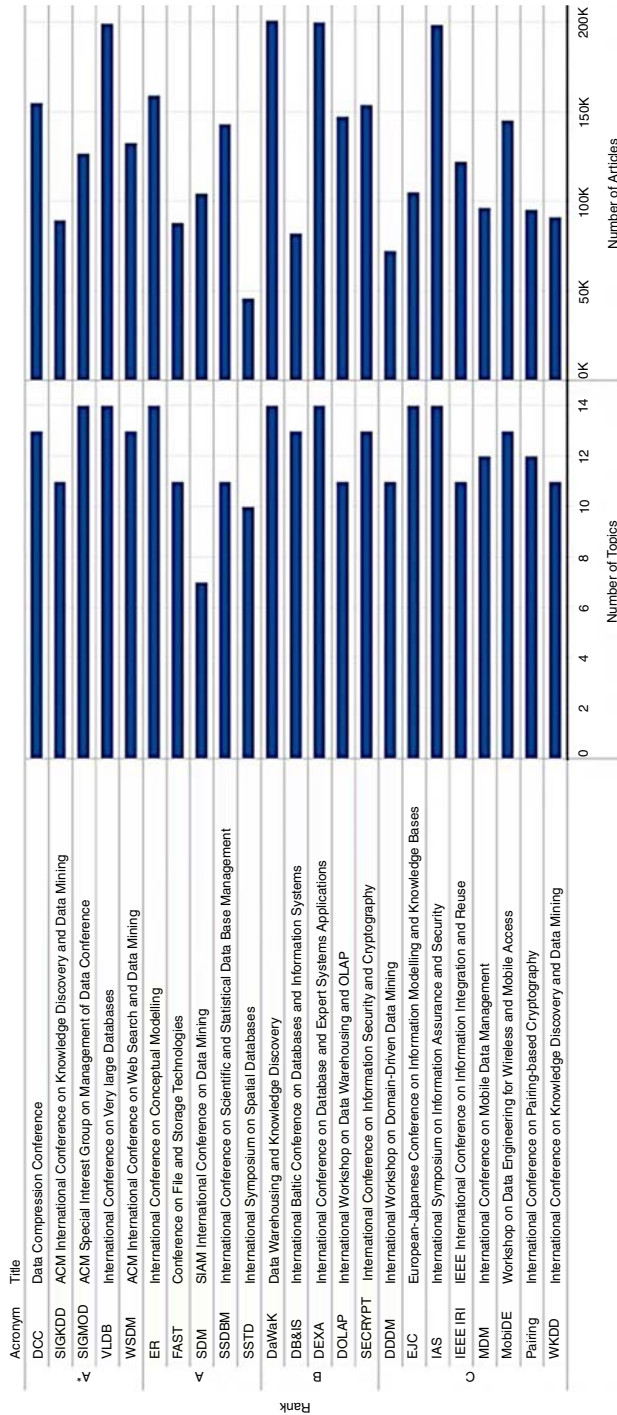
Figure A2.
Frequent keywords
from CFP data
set – two grams

Figure A3.
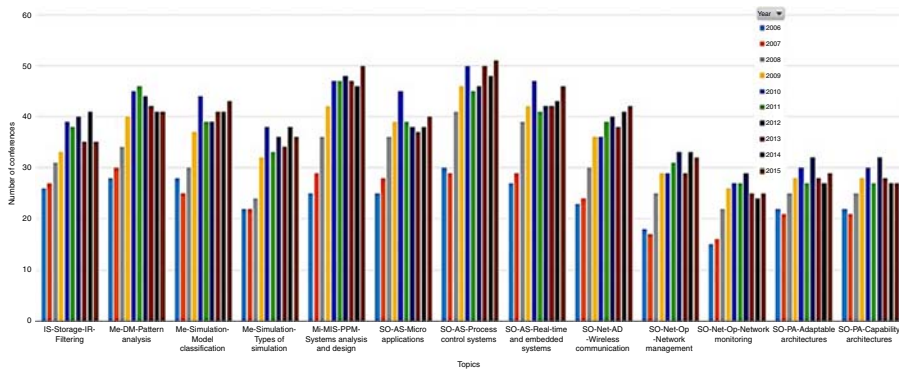Frequent keywords
from CFP data
set – year
wise – two grams

Figure A4.
Hot topics from DBLP
data set – 2006–2015

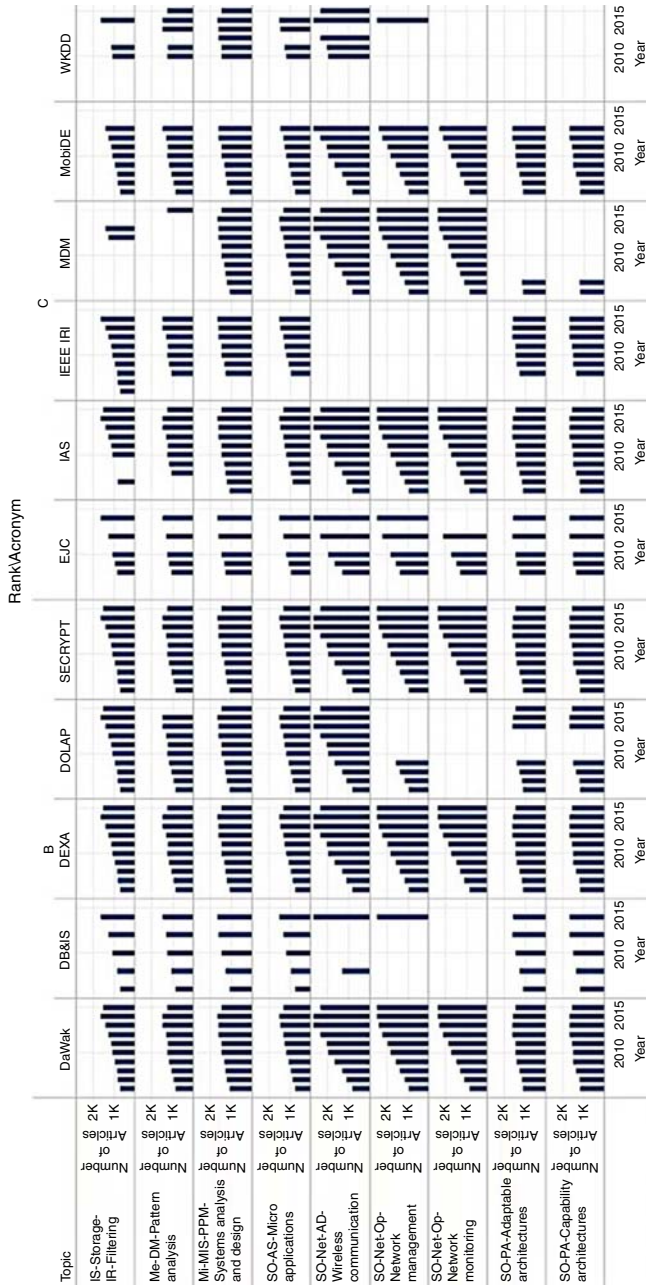**Figure A5.**
Trending conferences
with respect to
hot topics

Figure A6.
Hot topics frequency
in CFP from
conferences

Figure A7.
Summary of trending conferences vs hot topics (Rank A* and A)

Figure A8.
Summary of trending conferences vs hot topics (Rank B and C)

Appendix 2

Token ID	Token	Node ID	ACM class	Probabilistic weight	Node token TF–IDF	Title token TF–IDF	Gram weight (G)	Relative weight (R)	Agr. weightage
4954	data mining	D.3.2	Language Classification	1	1.5842	0.0032	2	3	3.8566
1593	mining	H.2.8.s0	Data Mining	1	1.5842	0.0026	1	1	3.7459
125	applications	G.1.10	Applications	0.0588	1.938	0.0068	1	1	3.1627
4954	data mining	H.2.8.s0	Data Mining	0.5114	1.1759	0.0032	2	1	3.0996
125	applications	I.3.4.s0	Application Packages	0.25	1.2832	0.0068	1	1	2.6991

Table AI.
Example for weights