Asian Journal of Information Technology 15 (7): 1223-1231, 2016
ISSN: 1682-3915

Discovering Expert Communities from Ontology Based Document Classification: An Efficient Maximal Biclique Approach

K.P. Swaraj and D. Manjula
Department of Computer Science and Engineering, Anna University, Chennai, India

Abstract: As a consequence of advanced research occurring in various domains there is always a hunt for emerging hot topics and expert authors of such trending topics. An interesting phenomenon noticed is that the topics of publications of a group of researchers at a particular time interval have similarity and therefore exhibit similar multi hot topic interests. Efficiently finding such expert author communities having similar multi hot topic interests which are closed and complete on a temporal basis is the major contribution of this study. The novel framework proposed in this study and the maximal Biclique based SMTAC_Find algorithm effectively detects such communities with multi hot topic interests. At first top-k hot-topics of a time interval are discovered utilizing article clusters created by ontology based domain classification system. Subsequently for a chosen time slot, the novel maximal biclique based algorithm is employed to detect the expert author communities having maximal similar multi hot-topic interests. DBLP dataset and its associated metadata of computer science articles were used for getting bibliographic information. A synthetic dataset of varying number of authors and their hot topic interests was used to verify Proposed SMTAC_Find algorithm. Analysis and performance of the algorithm depict that it can efficiently find expert communities in polynomial time for a specified number of hot topics 'k' under scoring its effectiveness.

Key words: Classification, DBLP, expert community detection, hot topic interests, maximal biclique, ontology

INTRODUCTION

From its initial stages, the field of computer science has been very broad and ever growing. This fact is revealed by thousands of research study which are published in the field of computer science and engineering. These researchs study cover broad computer science topics like data mining, artificial intelligence, information extraction, etc., which can be further divided into several sub-topics, sub-sub-topics and so on . DBLP is a bibliographic dataset which provides metadata about computer science publications and is now used by many computer scientists for research purpose (Ley (2002). Rich information in this metadata has enticed considerable research interest. Some of them are:

- To figure out the experts in relevant fields of computer science towards consulting as well as for academic purpose (Deng *et al.*, 2008)
- To analyze an author's scientific career and have a study on the existing computer science communities Biryukov and Dong (2010)
- To extract author cooperation by measuring the strength of association between authors (Minks *et al.*, 2011)
- To extract knowledge from a dataset by applying different overlapping clustering methods (Obadi *et al.*, 2010).

Careful analysis of this dataset divulges certain interesting statistics. For example, analysis of dataset for a particular time interval, say most recent 12 months shows that, there will be certain hot topics which will have more research orientation than other topics. Such novel topics can be commonly termed as trending hot topics. Another key observation is that, many of the authors have multiple specialized research areas (Deng *et al.*, 2008). They do not confine constantly in one area of discipline but exhibit interdisciplinary interests and publish articles in more than one topic. Among these researchers, several may have overlapping topic interests which sometimes may emerge as trending hot topics in a particular time interval. The novel framework proposed in this study attempts to identify such expert authors who have multi-disciplinary hot-topic interests at a particular interval of time. Besides, it also proposes an efficient algorithm to identify all such expert author communities who have similar top-k hot topic interests in a specific time interval providing scope for collaborative work.

Corresponding Author: K.P. Swaraj, Department of Computer Science and Engineering, Anna University, Chennai, India

Literature review: The aim of this research is to identify maximal expert author communities or clusters if any having similar multi hot-topic interests in emerging hot topics. Hence, focus of this research is on topic discovery from publications, classification based on domain (Moitheen and Khader, 2014). Topic extraction and classification from document content have been studied based on content statistics and generative probabilistic models. A content statistics-based study has been proposed and evaluated, (Wartena and Brussee, 2008). They used most frequent nouns, verbs and proper names as keywords and clustered them based on different similarity measures adopting the induced k-bisecting clustering algorithm. A lot of works were based on a popular topic modeling technique called generative model and its variants (Rathore and Roy, 2014). Steyvers *et al.* (2004) proposed a new unsupervised learning technique for extracting information from large text collections based on Probabilistic Author-topic models (Steyvers *et al.*, 2004). A few other previous works focused on detecting bursty and hierarchical structure in streams (Kleinberg, 2003) and discovering evolutionary theme patterns from text (Mei and Zhai, 2005). Another unsupervised method of clustering documents is based on frequent itemsets (Krishna and Bhavani, 2010). Variants of this Apriori-based algorithm has been studied in the algorithms FTC and HFTC (Beil *et al.*, 2002). FTC creates flat clustering and HFTC, a hierarchical clustering. Both works consider documents as bags of words and has the limitation that the semantic information present in the document is lost. In our proposed research an attempt is made to use only titles of documents extracted from dblp for topic detection considering title as sequential group of phrases preserving semantics. Mapping of article titles to topics is based on dissociation of phrases into frequent keyword-sets which is very fast and highly scalable, (Shubhankar *et al.*, 2011). We extract phrases from the titles of the research study and derive frequent substrings as frequent keyword-sets, maintaining the underlying semantics.

Detection methods of various clusters from any static dataset vary according to the requirements and applications. Onli ne social network analysis reveals that people with similar interests tend to form groups and collaborate with one another. A series of approaches have been proposed to investigate about social networks and clusters alias communities (McCallum *et al.*, 2007; Newman, 2003; Adebiyi *et al.*, 2015). Radicchi *et al.* (2004) introduced two quantitative definitions of community and showed how they are implemented in practice in the existing algorithms. Wang *et al.* (2010) proposed a novel co-clustering framework utilizing the networking information between users and tags in social media in order to discover overlapping communities. Zhang *et al.* (2007) used social network analysis methods to identify Expertise networks in online communities from Java Forum, a large online help-seeking community.Two generative Bayesian models have been proposed by Zhou *et al.* (2006) for semantic community discovery in SNs. Certain studies have been conducted for community detection in xml based DBLP dataset (Alwahaishi *et al.*, 2011; Huang *et al.*, 2009). Left-Right-Oscillate algorithm was proposed by Drazdilova *et al.* (2013) for finding communities in a large co-author network from DBLP. Tulasi and Rao (2014) studied about the structural features in DBLP and LiveJournal to elucidate community formation, growth and how the overlaps among pairs of communities change over time (Backstrom *et al.*, 2006). But, the aforementioned techniques cannot be used in our scenario since we map authors and topics to a bipartite graph model. A recent approach for community detection from social network is SWG_Find algorithm based on bipartite graph and biclique generation (Chalil and Sendhilkumar, 2014; Pandra and Sendhikumar, 2013). However, drawback is that this method considers only all bicliques from the bipartite graph as same wavelength groups. This leads to overlapping and duplicate clusters and hence lack of precision. In our research network relationships are established from xml tags in xml based dblp dataset by building a Bigraph of 'Topics' and 'Authors'. The proposed SMTAC_Find algorithm is based on construction of all maximal bicliques from Bigraph and generates only maximal bicliques of multi-hot topic expert author clusters, thereby guaranteeing 100% precision towards identifying maximal similar Multi-hot topic interest expert author clusters.

MATERIALS AND METHODS

Proposed framework: Architecture of framework is shown in Fig. 1. Main components of the framework are Bibliographic dataset and a domain classification system. Metadata of publications are first extracted from dataset. Then based on titles of publication, publications are mapped to their corresponding subject topics and classified accordingly in ontology based domain classification system. Thereafter this novel framework tries to find out similar hot topic expert author clusters in the polynomial time.

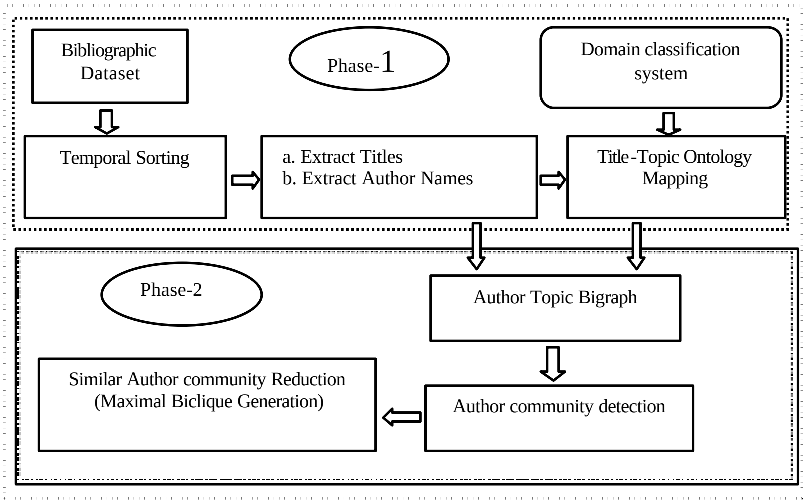

Bibliographic Dataset
Phase-1
Domain classification system
Temporal Sorting
a. Extract Titles
b. Extract Author Names
Title-Topic Ontology Mapping
Phase-2
Author Topic Bigraph
Similar Author community Reduction (Maximal Biclique Generation)
Author community detection

Fig. 1: Architecture

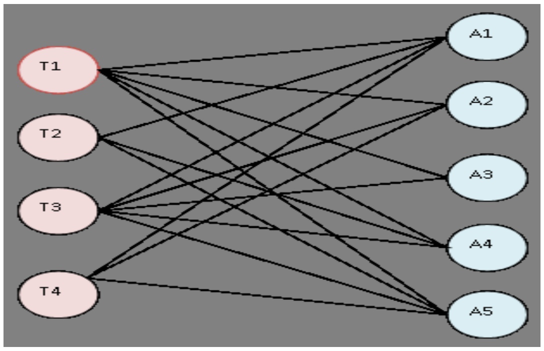


Fig. 2: Topic-author interest graph

There are 2 phases in this framework and this research is based on Computer science domain. In first phase, bibliographic dataset plays a major role. We have chosen DBLP dataset which list more than 2.1 million computer science publications and about 1.2 million authors. For computer science researchers, the DBLP web site is a popular tool to trace the research of colleagues and retrieve bibliographic details. Ranking and profiling of persons, institutions, journals or conferences is another sometimes controversial usage of DBLP. It is easy to derive several graphs like the bipartite person-publication graph, the person-journal or person-conference graphs or the co-author graph which is an example of a social network. In this framework titles and authors of journals are extracted from DBLP xml dataset in the first step of Phase 1.

The second step of Phase 1 is mapping titles of articles to topic classes in Computer science classification taxonomy, the 2012 ACM Computing Classification system. The full CCS classification tree is freely available for download in these formats: SKOS (xml), word and HTML for educational and research purposes. Mapping of article titles to topics is based on dissociation of phrases into frequent keyword-sets (Steyvers *et al*., 2004). Top-k trending topics of a particular time interval are found by statistical methods. In this study, we focus on algorithm to find similar topic interest expert author clusters and so the details regarding title to topic mapping and top-k hot topics identification are considered as out of scope.

At the end of the two steps in Phase-1, the study, is classified into its, corresponding topic based on phrases in title. Its research is already extracted from the dblp data set. We are considering only top-k hot topics and authors who have published in these topics. Therefore, a connection between the authors and hot-topics can be clearly established from Phase-1. Bipartite graph creation and author community generation is performed in Phase 2.

In Fig. 2 two set of vertices are presented. One represents set of 4 hot topics {T1, T2, T3, T4}. Other is the set of 5 authors denoted by {A1, A2, A3, A4, A5}.

Since, there are two classes of vertices such as authors and topics, a bipartite graph with author-topic relationship can be used to represent this type of relationship for a given period in the next phase. A bipartite graph is one whose vertices can be partitioned into a pair of non-empty, disjoint partitions such that no two vertices within the same partition are connected by an edge. An example of such a bipartite graph is given in Fig. 2.

Table 1: Topic-author interest matrix

	A Matrics				
T Matrics	A1	A2	A3	A4	A5
T1	Y	Y	Y	Y	Y
T2	Y	N	N	Y	Y
T3	Y	Y	Y	Y	Y
T4	Y	Y	N	N	Y

< {T1, T3}, {A1, A2, A3} > is a biclique, but is not maximal; < {T1, T3}, {A1, A2, A3, A4, A5}> is a maximal biclique; < {T1, T2, T3}, {A1, A4, A5} > is a maximal biclique

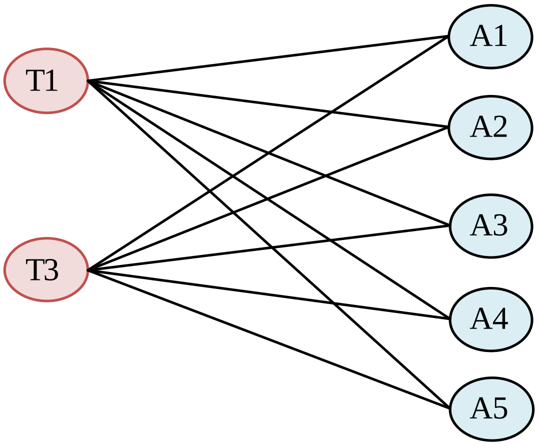


Fig. 3: Maximal similar topic-interest expert author cluster 1

As in Table 1, each row represents the hot topics in which a particular author has published articles. The Authors are represented in columns of the matrix. There can be many authors who publish documents in a particular time interval. In this research, top-n hot topics of a time interval are determined and then the authors corresponding to those topics are identified and represented as a bipartite graph. Once the Topic-Author relationship is represented in bipartite graph, the stage is set for finding the maximal similar multi-hot topic interest expert author clusters which can be accomplished through an enumeration of maximal bicliques. A biclique in a bipartite graph is a complete bipartite subgraph, that is, a bipartite subgraph containing all permissible edges. The notion is formalized as follows:

Definition 1:- Let G = (U V, E) denote a bipartite graph. A biclique BC = (U',V') is a subgraph of G induced by a pair of two disjoint subsets U' U, V' V, such that uU',vV', (u,v) E.

A maximal biclique is a biclique, which is not contained in any other larger biclique. From (2), set of authors {A1, A2, A3, A4, A5} is the maximal author set which is having similar interest topics {T1, T3} which is shown in Fig. 3.

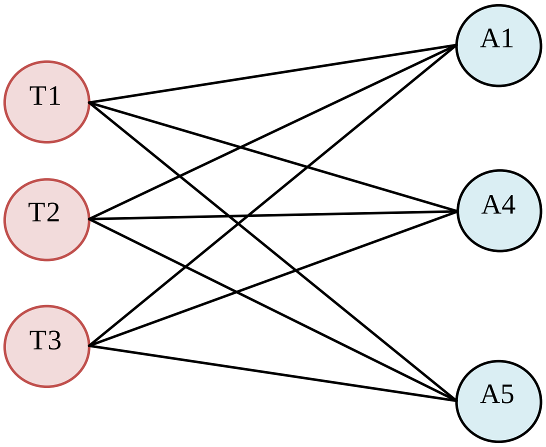


Fig. 4: Maximal similar topic-interest expert author cluster 2

From (2), set of authors {A1, A2, A3 } is an author set which is having similar interest in topics {T1, T3}. But, this is not a maximal one since there is an expert author cluster superset {A1, A2, A3, A4, A5} which is the maximal set of authors with interest in topic set {T1,T3}. This research attempts to detect all maximal clusters with interest in different subsets of topics.

From (3) A1, A4 and A5 form a maximal expert author cluster for a time period since they have published articles in similar topics{T1,T2,T3} which is shown in Fig. 4.

Set notation for bipartite graph representation: For finding maximal Bigraph, the problem is represented in set notation. Set notation can be directly derived for bipartite graph representation as follows. Consider Fig. 2 with hot Topic set 'T' having 4 hot topics (T1, T2, T3, T4) for the time interval considered.

Hot-Topic Set T = {T1, T2, T3, T4}

Power set of T = { {T1, T2, T3, T4},

{T1, T2, T4}, {T1, T2, T3}, {T1, T3, T4}, {T2, T3, T4},

{T1, T2}, {T1, T3}, {T1, T4}, {T2, T3}, {T2, T4}, {T3, T4}

{T1}, {T2}, {T3}, {T4},{ } }

Let Author Set A = {A1,A2,A3,A4,A5},the set of authors who have published in hot topics during the time interval considered. For Topic T1 the author set is {A1,A2,A3,A4,A5}; The above notation means that 5 authors A1, A2, A3, A4 and A5 have published articles in the hot topic T1 for the considered time interval. For Topic T2 the author set is {A1,A4,A5}; For Topic T3 the author set is {A1,A2,A3,A4,A5}; For Topic T4 the author set is {A1,A2,A5}.

In order to find the similar topic-interest expert author clusters, SWG-Find algorithm calculates expert author clusters for all the subsets with subset-size>1 of the topic set:

T = {T1, T2, T3, T4}.

Hence, in SWG-Find we need to find the reduced PowerSet from PowerSet [T].
Reduced Power set of T:

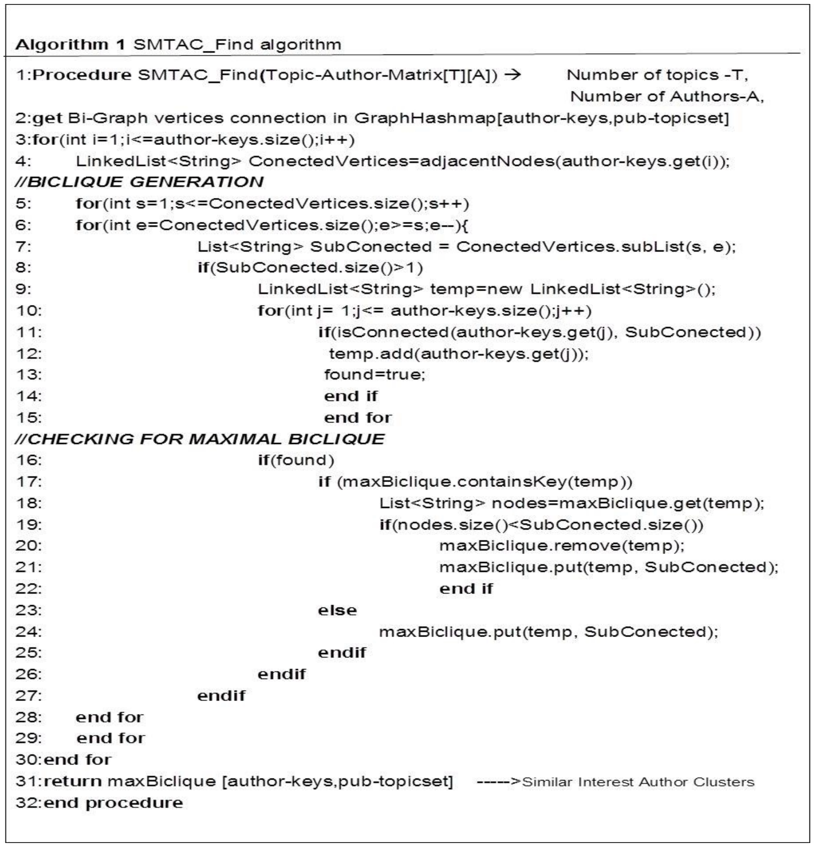
T = { {T1, T2, T3, T4}, {T1, T2, T4}, {T1, T2, T3}, {T1, T3, T4}, {T2, T3, T4}, {T1, T2}, {T1, T3}, {T1, T4}, {T2, T3}, {T2, T4}, {T3, T4} }.

Then SWG-Find calculates the intersection of all topic subsets of Topic set T with number of subset elements >1 and prints resulting expert author clusters as SWG.

There seems to be always a trend in writing styles of authors and the way hot topics are dealt in journals. Hence, at a particular time interval, the number of hot topics in articles will be always a finite value. On the basis of above observation, we can deduce that for detecting similar topic expert author clusters within a particular time interval, the number of hot topics in the topic set could be fixed to a finite minimum value and can be computed in polynomial time. The proposed SMTAC_Find( ) algorithm does not consider all subsets of topics like SWG-Find( ) nor does it return all bicliques as expert author clusters. Instead it considers only topic subsets (connected-topic vertices) in which authors have published articles and returns only maximal bicliques which improve computational time.

Algorithm and complexity analysis: Pseudo code of SMTAC_Find Algorithm for solving the similar maximal topic expert author cluster detection problem based on hash map data structure is given above in (Fig. 5).

Input is given as a Topic-Author [n x m] matrix .If value of Topic-Author [I][j] =1, it means in Hot Topic i, the author j has published an article in the selected time interval.



Algorithm 1 SMTAC_Find algorithm

```
Procedure SMTAC_Find(Topic-Author-Matrix[T][A]) → Number of topics -T,
                                                   Number of Authors-A,
get Bi-Graph vertices connection in GraphHashmap[author-keys,pub-topicset]
for(int i=1;i<=author-keys.size();i++)
    LinkedList<String> ConectedVertices=adjacentNodes(author-keys.get(i));
//BICLIQUE GENERATION
    for(int s=1;s<=ConectedVertices.size();s++)
    for(int e=ConectedVertices.size();e>=s;e--){
            List<String> SubConected = ConectedVertices.subList(s, e);
            if(SubConected.size()>1)
                LinkedList<String> temp=new LinkedList<String>();
                for(int j= 1;j<= author-keys.size();j++)
                    if(isConnected(author-keys.get(j), SubConected))
                     temp.add(author-keys.get(j));
                    found=true;
                    end if
                    end for
//CHECKING FOR MAXIMAL BICLIQUE
                if(found)
                    if (maxBiclique.containsKey(temp))
                        List<String> nodes=maxBiclique.get(temp);
                        if(nodes.size()<SubConected.size())
                            maxBiclique.remove(temp);
                            maxBiclique.put(temp, SubConected);
                            end if
                    else
                        maxBiclique.put(temp, SubConected);
                    endif
                endif
            endif
    end for
    end for
end for
return maxBiclique [author-keys,pub-topicset]    ----->Similar Interest Author Clusters
end procedure
```

Fig. 5: SMTAC-find algorithm

The algorithm begins with bipartite graph construction from matrix in step-1 as hash map [author-keys, pub-topicset] format. For n number of hot topics and m authors, this can be max O(n×m) steps which is in polynomial time for a low value of n. Expert author clusters are found from the first n hot topics in a domain after ranking the total number of hot topics detected during a time period. Hence, the number of hot topics ‘n’ is always a minimum, say top-50 topics. From bipartite graph, published hot-topic set is retrieved for each author key first. For example if author A1 has published in T1, T2 and T3 this connected list< T1, T2, T3 > is retrieved. In steps 4, 5 connected topic vertices list are retrieved for each author in outer loop. Then in steps from 6-12, PowerSet (set of subsets) of connected-topic vertices of author is generated in descending order of topic-subset size. It is verified with other author-keys to see if there is any other author who has published for this topic-subset. If it is there, then those authors are added to Linked-list ‘temp’. That is if along with A1,authors A2 and A3 have published in topic-list<T1, T2, T3>,then Linked-list ‘temp’ has values <T1_T2_T3>. This completes a biclique with Authors[A1,A2,A3] and Topics [T1, T2, T3].

Even though subset generation takes exponential time, here we are considering only subsets of connected top-n topics 1<n<50 of an author for a particular period. Using a hash map implementation for set, the running time will be in polynomial time. Now from steps 15-23, the generated biclique is checked to see whether it is maximal. If it is maximal, it is added to the result-hash map in O(m) steps. Hence overall the running time of the algorithm is O(m×n2) which is in polynomial time for top-n topics 2<n<50 at a particular interval of time. Algorithm returns all maximal author bicliques in a Hash Map data structure in format maxBiclique(author-keys, pub-topicset).

RESULTS AND DISCUSSION

The SMTAC_Find algorithm was implemented and evaluated using hash map data structure of java language in a system with a configuration of 2.3 GHz, Core i7, Intel processor and 8GB RAM. We used Eclipse IDE to implement and test the algorithm. We mapped top-50 topics to symbols T1, T2, T3…T50 and ‘n’ authors to A1, A2, A3…An. The number of authors can vary from 2 to many. We did not consider author size of 1 since, this case is a trivial case.

Input to the algorithm was given in a file with Topic-Author [n x m] matrix format. Two sample inputs were given. We have compared expert author cluster results with SWG-Find algorithm by Chalil and Sendhilkumar (2014).

Table 2: Topic-author interest matrix

	A Matrics				
T Matrics	A1	A2	A3	A4	A5
T1	1	1	1	1	1
T2	1	0	0	1	1
T3	1	1	1	1	1
T4	1	1	0	0	1

Sample inputs and results; case 1: Sample input of 5 authors and 4 topics is given in Table 2.

Result: SMTAC_Find algorithm:- 4 maximal expert author clusters returned which are all maximal bicliques:

- [A1, A5]_[T1, T2, T3, T4],
- [A1, A2, A5] _ [T1, T3, T4],
- [A1, A2, A3, A4, A5] _ [T1, T3],
- [A1, A4, A5] _ [T1, T2, T3]}

Maximal bicliques Returned by SMTAC_Find

Analysis:

- Authors A1 and A5 have common interests in hot topics T1, T2, T3, T4 which is maximal
- Authors A1, A2 and A5 have common interests in topics T1, T3 and T4 which is maximal
- Authors A1, A2, A3, A4, A5 have common interests in topics T1 and T3 which is maximal
- Authors A1, A4 and A5 have common interests in topics T1, T2 and T3 which is maximal

SWG-Find Biclique generation method by Chalil and Sendhklumar (2014). The 11 bicliques returned instead of 4 maximal which leads to lower precision values. Most of the results are overlapping and non-maximal (Table 3).

Case 2: Sample input of 10 and 10 topics is given below in Table 4.

Result:

- SMTAC_Find algorithm:- 39 maximal expert author clusters returned which are all maximal bicliques
- SWG-Find Biclique generation method by Chalil and Sendhikumar (2014). The 1013 researcher bicliques returned which leads to very low precision values (Table 5 and Fig. 6)

Comparison of precision values for case-1 and case-2:

$$\text{Precision} = \frac{\text{No.of maximalauthor clusters returned}}{\text{TotalNo.of author clusters returned}} \times 100$$

Analysis: SMTAC_Find algorithm returns maximal clusters with 100% precision always. On Comparison, the

Table 3: Bicliques returned by SWG-find

	Perameters	
1	2	3
A1, A5]→[T1, T2, T3, T4]	[A1, A5]→[T2, T3, T4]	[A1, A4, A5]→[T2, T3]
[A1, A5]→[T1, T2, T4]	[A1, A4, A5]→[T1, T2]	[A1, A5]→[T2, T4]
[A1, A4, A5]→[T1, T2, T3]	[A1, A2, A3, A4, A5]→[T1, T3]	[A1, A2, A5]→[T3, T4]
[A1, A2, A5]→[T1, T2, T3	[A1, A2, A5]→[T1, T4]	

Table 4: Topic-author interest matrix

					A matrics					
T matrics	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
T1	1	1	1	1	1	1	1	1	1	1
T2	1	0	1	0	1	0	1	1	1	1
T3	1	1	1	0	1	0	0	1	1	1
T4	1	1	1	1	1	1	1	1	1	1
T5	1	1	0	0	0	0	0	1	1	1
T6	1	1	1	1	0	0	1	1	0	1
T7	1	1	1	1	1	1	1	1	1	1
T8	0	0	1	1	0	1	0	1	1	1
T9	0	1	0	1	0	1	0	1	1	1
T10	1	1	1	1	1	1	1	1	1	0

Table 5: Comparison of precision

Case	SMTAC Find (%)	SWG Find (%)
1	100	(4/11) 100 = 36.36
2	100	(39/1013) 100 = 3.84

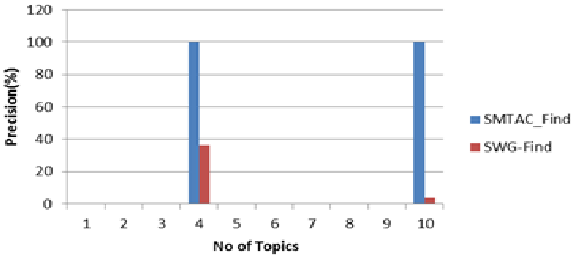


Fig. 6: Precision comparison

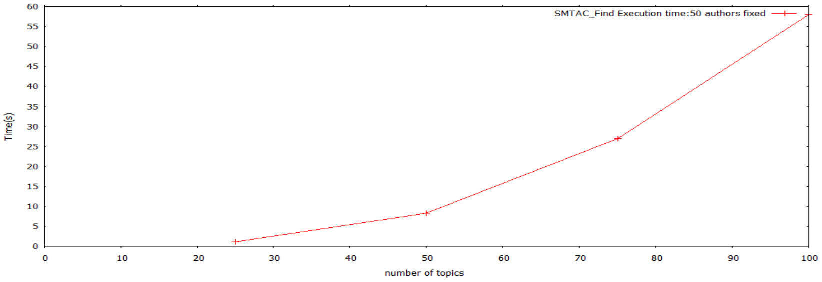


Fig. 7: Computation time (number of authors: constant)

precision value of SWG-Find decreases dramatically as number of topics and authors increases as it returns all bicliques which includes overlapping as well as non maximal bicliques instead of expected maximal bicliques.

Computation of execution time: For experimental purpose we considered two cases for calculating execution time. In both cases, algorithm returned all maximum bicliques in polynomial time as shown below. We have not compared execution time with SWG_Find algorithm as it generates only bicliques as same wavelength group rather than the expected maximal bicliques (Fig. 7, 8). Restricted the size of authors to 50 and considered topics in 25, 50, 75, 100 sizes.

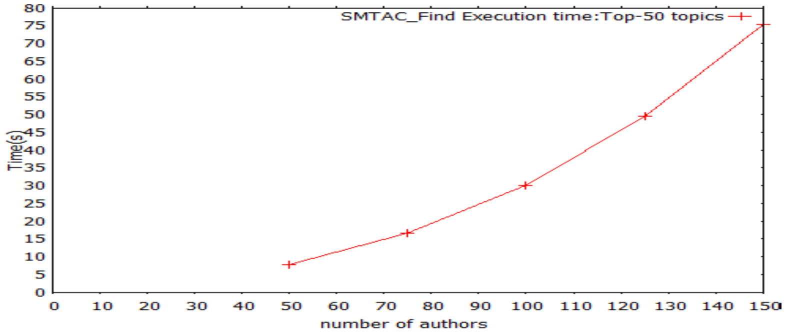


Fig. 8: Computation time (number of topics: constant)

CONCLUSION

The contribution of this study is the identification of expert author communities having similar hot topic interest which is maximal from bibliographic dataset. A novel polynomial time maximal biclique based SMTAC_Find algorithm is proposed to ascertain all the maximal expert author groups which have similar interest in more than one hot topic at a particular time period. The framework is mapped to a bipartite graph theoretical model. Proposed algorithm returns maximal expert author clusters without any overlaps and duplicates. We have implemented the proposed algorithm with hash map data structure and compared its performance with the SWG-Find algorithm. Results demonstrated that the proposed method has significant performance advantage and it outperformed SWG-Find algorithm in Precision values on all cases, demonstrating its effectiveness and efficiency. This can be very much helpful to researchers looking for maximal group of authors who have published in a set of hot topics at a particular interval.

RECOMMENDATIONS

In this research, we concentrated only on top-k (1<k<50) hot topics and its non-overlapping maximal expert author clusters. As a future work, cloud computing techniques and algorithms can be introduced to rapidly find out maximal expert author clusters from all topics and research in the dataset for any time interval, significantly reducing its computational cost.

REFERENCES

Adebiyi, A.A., S. Okuboyejo M. Akinbode M.G. Agboola and A.A. Oni, 2015. Exploring social networking and university students academic performance. Asian J. Inf. Technol., 14: 253-259.

Alwahaishi, S., J. Martinovic and V. Snasel, 2011. Analysis of the DBLP Publication Classification Using Concept Lattices. In: Digital Enterprise and Information Systems. Ariwa, E. and E.E. Qawasmeh (Eds.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-642-22602-1, pp: 99-108.

Backstrom, L., D. Huttenlocher J. Kleinberg and X. Lan, 2006. Group formation in large social networks: Membership, growth and evolution. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2006, ACM, New York, USA., ISBN:1-59593-339-5, pp: 44-54.

Beil, F., M. Ester and X. Xu, 2002. Frequent term-based text clustering. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002 Edmonton, Alberta, Canada, pp: 436-442.

Biryukov, M. and C. Dong, 2010. Analysis of computer science communities based on DBLP. Proceeding of the International Conference on Research and Advanced Technology for Digital Libraries, September 6-10. 2010, Springer Berlin Heidelberg, Glasgow, UK., ISBN: 978-3-642-15463-8, pp: 228-235.

Chalil, R.P. and S. Sendhilkumar, 2014. Same wavelength group identification from online social networks: A general framework. Comput. Sci. Inf. Syst., 11: 229-239.

Deng, H., I. King and M.R. Lyu, 2008. Formal models for expert finding on DBPL bibliography data. Proceeding of the for Eighth IEEE International Conference on Data Mining, December 15-19, 2008, IEEE, Pisa, Italy, ISBN: 978-0-7695-3502-9, pp: 163-172.

Drazdilova, P., J. Martinovic and K. Slaninova, 2013. Spectral Clustering: Left-Right-Oscillate Algorithm for Detecting Communities. In: New Trends in Databases and Information Systems. Mykola, P. and W. Marek (Eds.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-642-32518-2, pp: 285-294.

Huang, Z., Y. Yan, Y. Qiu and S. Qiao, 2009. Exploring emergent semantic communities from DBLP bibliography database. Proceedings of the ASONAM'09 International Conference on Advances in Social Network Analysis and Mining, 2009, July 20-22, 2009, IEEE, Athens, Greece, ISBN: 978-0-7695-3689-7, pp: 219-224.

Kleinberg, J., 2003. Bursty and hierarchical structure in streams. Data Min. Knowl. Discovery, 7: 373-397.

Krishna, S.M. and S.D. Bhavani, 2010. An efficient approach for text clustering based on frequent itemsets. Eur. J. Scient. Res., 42: 399-410.

Ley, M., 2002. The DBLP computer science bibliography: Evolution, research issues, perspectives. Proceedings of the 9th International Conference on Information Retrieval, September 11-13, 2002, Springer Berlin Heidelberg, Lisbon, Portugal, ISBN: 978-3-540-44158-8, pp: 1-10.

McCallum, A., X. Wang and A. Corrada-Emmanuel, 2007. Topic and role discovery in social networks with experiments on enron and academic email. J. Artif. Intell. Res., 30: 249-272.

Mei, Q. and C. Zhai, 2005. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, August 21-24, 2005, ACM, New York, USA., ISBN:1-59593-135-X, pp: 198-207.

Minks, S., J. Martinovic, P. Drazdilova and K. Slaninova, 2013. Author cooperation based on terms of article titles from DBLP. Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011) August 2011, July 17, 2013, Springer Berlin Heidelberg, Prague, Czech Republic, ISBN: 978-3-642-31602-9, pp: 281-290.

Moitheen, T.P. and P.S.A. Khader, T.P. 2014. Framework for personalized learning system using ontology. Asian J. Inf. Technol., 13: 368-374.

Newman, M.E.J., 2003. The structure and function of complex networks. Soc. Ind. Applied Math. Rev., 45: 167-256.

Obadi, G., P. Drazdilova, L. Hlavacek, J. Martinovic and V. Snasel, 2010. A tolerance rough set based overlapping clustering for the DBLP data. Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), August 31-September 3, 2010, IEEE, Toronto, ON., ISBN: 978-1-4244-8482-9, pp: 57-60.

Pandara, R. and S. Sendhilkumar, 2013. Identifying same wavelength groups from twitter: A sentiment based approach. Proceedings of the ACIIDS 2013 Conference on Intelligent Information and Database Systems, March 18-20. 2013, Springer Berlin Heidelberg, Lumpur, Malaysia, ISBN: 978-3-642-36542-3, pp: 70-77.

Radicchi, F., C. Castellano, F. Cecconi, V. Loreto and D. Parisi, 2004. Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA., 101: 2658-2663.

Rathore, A.S. and D. Roy, 2014. Performance of LDA and DCT models. J. Inf. Sci., 40: 281-292.

Shubankar, K., A. Singh and V. Pudi, 2011. A frequent keyword-set based algorithm for topic modeling and clustering of research papers. Proceedings of the 3rd Conference on Data Mining and Optimization (DMO) 2011, June 28-29, 2011, IEEE, New York, USA., ISBN: 978-1-61284-212-7, pp: 96-102.

Steyvers, M., P. Smyth, M.R. Zvi and T. Griffiths, 2004. Probabilistic author-topic models for information discovery. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 22-25, 2004, ACM, New York, USA., ISBN:1-58113-888-1, pp: 306-315.

Tulasi, R.L. and M.S. Rao, 2014. Tools and techniques for ontology interoperability: A survey. Int. J. Comput. Sci. Inf. Secur., 12: 93-98.

Wang, X., L. Tang, H. Gao and H. Liu, 2010. Discovering overlapping groups in social media. Proceedings of the IEEE 10th International Conference on Data Mining (ICDM) 2010, December 13-17, 2010, IEEE, Sydney, NSW, ISBN: 978-0-7695-4256-0, pp: 569-578.

Wartena, C. and R. Brussee, 2008. Topic detection by clustering keywords. Proceedings of the 19th International Workshop on Database and Expert Systems Application 2008 DEXA'08, September 1-5, 2008, IEEE, Turin, Italy, pp: 54-58.

Zhang, J., M.S. Ackerman and L. Adamic, 2007. Expertise networks in online communities: Structure and algorithms. Proceedings of the 16th International Conference on World Wide Web, May 08-12, 2007, ACM, New York, USA., ISBN: 978-1-59593-654-7, pp: 221-230.

Zhou, D., E. Manavoglu, J. Li, C. L. Giles and H. Zha, 2006. Probabilistic models for discovering e-communities. Proceedings of the 15th international conference on World Wide Web, May 22-26, 2006, Scotland Uk, pp: 173-182.