ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

Subgraph Mining of Seminal Papers in a Large Graph with Their Paper Genealogy

Monika Deshpandey[#1], Sonali Bodkhe

Department of Computer Science & Engineering, G.H. Raisoni Academy of Engineering &Technology, Nagpur, Maharashtra, India[#1]

Asst. Professor, Department of Computer Science & Engineering, G.H. Raisoni Academy of Engineering &Technology, Nagpur, Maharashtra, India [#]

ABSTRACT: In information technology there many types of graphs such as biological graphs ,social network graphs, Semantic Web graphs are presented in real world and, single vertex of any graphs always contain huge information, which can be created by a set of elements. In this paper, we will study a subgraph mining in a large graph over a large database, which creates subgraphwith help of some keywords for example name of domain and Subject. By putting all required keywords the main large graph will be created. That main graph contains rich information of different scientific searched papers, and those papers are connected in the form of graph according to year wise from which is ancient paper to new latest paper form, this connected papers show the genealogy of different seminal papers. By using the proposed method it is easy to find different seminal papers in sub graph from the main large graph. The genealogy algorithm is useful to find out quick result of searched topic, in which all founded papers are similar to given query. Other proposed methods are used to find the similar papers according to their citations and reference. In today's real world searching of data is the most important part of daily use, similar this when a freshers and researchers want to search any topic regarding their field and domain so they have to face so many problems regarding to search a papers and seminal paper related to that particular searched paper. Bythis proposed work, the searching process will be easier for userin which user will not expend their important time on searching different papers related their domain this proposed methods show the effectiveness efficiency with the accuracy of different data .

KEYWORDS*:* Information technology, Subgraph mining,searchingof similar paper, genealogy graph of papers, Analysis of data, Construction of Efficient Genealogy.

I. INTRODUCTION

When a newresearcher starts research on anew topic, he/she faces a lot of problem in searching the problem of searching in a particular domain, for literature survey, for similar papers etc. a considerable amount of time has been spent on these required query. There for a new graph mining technique is using for searching relevant papers from conferences, journals, or scholar search engines. First, they may consider good survey papers as a starting point, for that some keywords have to fill up in the beginning of proposed work. However, due to their static nature, new research results cannot be covered in them frequently.

Second, the researcher will consider all selected seminal papers which would be obtained by given query. While a researcher going to work on a literature survey, it is important to acknowledge present research trend in a selected topic. A research trend gives us insight in the topic, i.e., what kind of research has been done before and what the latest issues are at this time. Working on literature survey part is more challenging when fresher starts with unfamiliar domains such as physics, psychology, chemistry, statistics, and biology for interdisciplinary studies.

There are so many ways to search papers in the form of document, but the Graph visualization is the new way to research in a new topic. A bunch of research papers with their literature papers make confuse to new user, according to

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

proposed work subgraph mining process will be easier than document searching. Graph visualization can help to form an overview of relational patterns and detect data structure which may much faster than data in a tabular form.
In the Graph visualization all papers will be shown in the graph, a specific nodes and vertexes are used to show different papers. There are some problems arising while a new user goes to search a new topic that problems are 1) seminal papers finding, and 2) constructing automatic genealogy from given keywords. For working on first problem the Genetic algorithm has been used which helps to deals with the problem of matching of citations in different papers and that of automatically constructing a citation index. After that the large main graph has been generated which shows the genealogy of different seminal papers.
The large main graph contain huge number of paper for searching another reference paper the subgraph mining has been used which will be done by putting year in search box, the year wise paper selection is beneficial to get instant data in the form of graph . The paper Genealogy algorithmrecommended for finding seminal papers in a specific topic. However, this method mainly focuses on how to reflect targetkeywords for gaining the result.

Seminal papers result is quite different depending on the target keywords. Paper selection process is depends upon the user that's why selected domain is useful to work on this proposed work. There aresome folds: 1) we propose Genetic algorithm (G.A.) for SPF, and 2) we propose how to construct their genealogy. There are some points which would be declared the steps of proposed work

1) Selection of target keyword and domain*:*First have to put some target keywords and the name of domain for searching the data as a form of graph.

2) A Genetic Algorithm for seminal paper finding*:* The genetic algorithm is used to search the data.

3) KNN and Navie baise for paper genealogy*:* These methods are used to find the seminal papers in the form of graph as well as used to find subgraph from a main large graph.

II. RELATED WORK

Novel classificationtechnique incorporates graph structural dataduring a hybrid index structure. TALE method is using to match all the sub graphs. this is often a general tool for making graphs. this iseasilycustomizedto fulfillthe needof many applications. These project empirical evaluations denote the improved effectiveness and efficiency. this isuseful for searchinglarge nodes during a specific domain [1].Tuttle 1st introduced barycentric embedding, analysis of graph visualization techniques that continues to be a highly active field attracting a lot of attention [2].
When any inherent complexity arises, for this Finding Random graphs (FRG) is generating on their multiplicative linear random range generator algorithm used for finding a ho-momorphic image of a pattern graph in a very target graph for this some algorithms are applied which is useful to remove unsuccessful mapping which retrieves sub graphs that are structurally similarity to the query graph, and meanwhile satisfy the condition of vertex pair matching with weighted (dynamic) set similarity. [3]. one in all the fastest method Graph X-Ray (G-Ray), is used to find out sub-graphs that match the desirable question pattern.For graph indexingthere isone more recent method, willreturn no answer whenan exact instance of a pattern does not exist. The efficiency of this techniqueis very low whichis used for intermediate or non-intermediate node. The non-intermediate nodes will bereferred to as matching nodes. [4]. in the condition of huge graph queries there isa novel technique for approximate matching of large graph queries.

For showing an overview of relational patterns Graph visualizationis used. it'sbeneficial for fasterknowledgedetect -action whenknowledge is present in tabular kind. Understanding the way of generation of graph the formthe data sets node representationis verynecessarywhere time acknowledgement is also a representative partwhich graph is taking more timeto generate , that isunderstand by node representation. [5,6].Exact subgraph matching queryrequiresthat all the vertices and edges are matched exactly. The Ullman's subgraph isomorphismmethod [14] and VF2 [11] algorithmdon't utilize any index structure, thereforethey areusuallycostlyfor big graphs. Tree Pi [15] indexes graph databases usingfrequent sub trees as indexing structures.GADDI [16] could be a structure distance primarily based subgraph matching algorithmic program in a very giant graph. Zhao et al. [6] investigated the SPath algorithmic program, which utilizes shortest paths around the vertex as basic index units. Cheng et al. 18] this technique proposed a new two-step R-join algorithmic program to expeditiously find matching graph patterns from a

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

large graph. Zou et al. [1] proposed a distance primarily based multi-way join formula for answering pattern match queries over an oversized graph. dynasty et al. [17] .
Quick SI algorithmic rule for subgraph similarity optimized by choosing associate search order supported some features of graphs. SING [18] could be a novel categorization system for subgraph similarity during a large scale graph. Graph [19] is a query language for graph databases which supports graphs as the basic unit of information. Sun et al. [7] utilised graph exploration and parallel computing to method subgraph matching question on a billion node graph. Recently, associate degree efficient and strong subgraph isomorphism algorithmic program Turbots [12] was proposed.
RINQ [20] and GRAAL [21] area unit graph alignment algorithms for biological networks, which might be used to solve isomorphism problems. However, a query graph is much smaller than the data graph in subgraph isomorphism problems, while the two graphs sometimes have similar size in graph alignment problems. to resolve subgraph isomorphism problems, graph alignment algorithms introduce additional value as they must initial realize candidate subgraph of similar size from the large knowledge graph. in addition, existing exact subgraph matching and graph alignment algorithms don't take into account weight set similarities on vertices; this is often the high post process in set similarity computation.
Approximate subgraph matching question typically issues the structure data and allows a number of the vertices or edges not being matched exactly. the first graph index technique Closure-tree [22] supports both subgraph queries and graph similarity queries. Adventure story [25] is Associate in nursing approximate subgraph matching technique that finds sub graphs in the info that area unit similar to the query, allowing for node mismatches, node gaps, and graph structural differences.Torque [24] may be a topology working with genealogical graphs isn't any exception in this sense
The approaches of the first class follow the main idea behind the Apriori algorithmic rule [1] for mining frequent item sets. additional specifically, they consider the apriori property, consistent with which all the sub patterns of a frequent subgraph pattern are also frequent. Thus, to enumerate candidate patterns, they apply breadth-first search to generate sub graphs of size (k + 1), by change of integrity two sub graphs of the previous level. Bijan Ranjbar-Sahraeib, Gerhard Weiss projected the homophile principle for augmentation of the original input graph by connecting the potential similar references, and second, to use a random walk based approach to consider contextual information available for each reference in the augmented graph. Keyword query routing approach is used short out all the arise queries [19].
Routing approach is used for searching joined information for improving the quality of project multilevel ranking method is used in this proposed system used for finding relevant information. Some graph join methods like merge join are used to make efficient graph in this approach, those all approaches create the complexity of nodes when combination parts ar increased. Ranking Keyword routing arranges is helpful to show efficient detail graph form [19].

III. LIMITATIONS (OF EARLIERWORKS)

There huge amount information sets square measure presented in the database for searching any data classification is needed to match the data. Complexity is arising from the largest data. The sub graph matching is working with short structure based mostly data like bibliography and for some attribute (CEO, Manager, Customer), but not working with large structure based information.
Computational problem had been arising in finding subset of vertices, from all adjacent to each different, adjacent also called complete sub graph. The accuracy, efficiency, and effectiveness of information are low in genealogy of reference graph. The pruning technique used to classify the data, but it couldn’t work to identify efficient graph according to keyword.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

IV. PROPSED WORK

This proposed system is useful for new researchers who want to start research on the new topic in the particular domain. Researchers have to work on latest project or on that project where some works are left to complete.

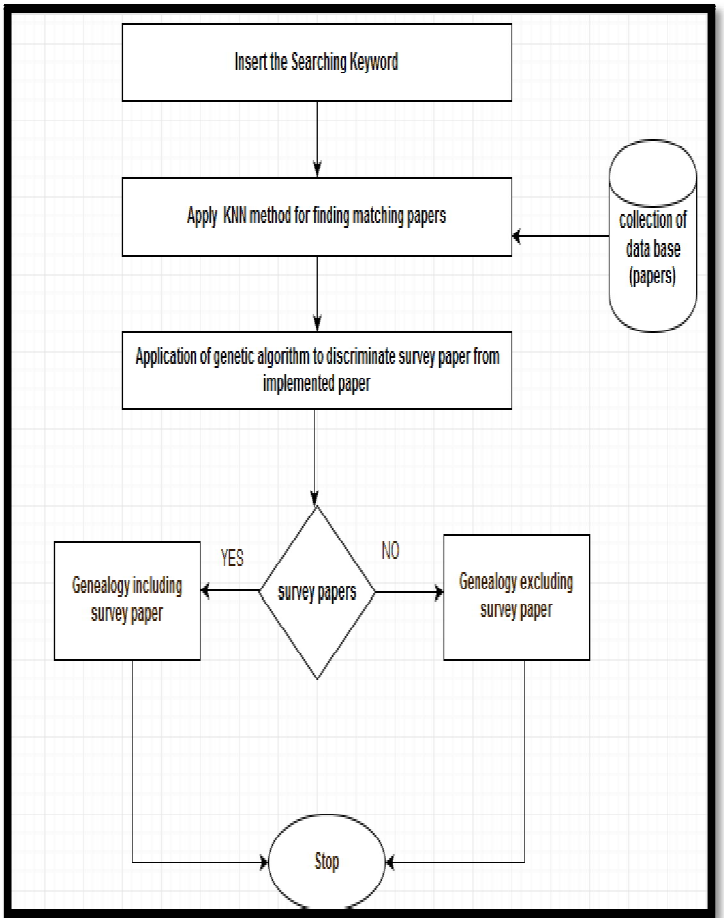

Fig1 – Block diagram of procedure

A paper genealogy term is used in this project; in which numbers of seminal papers are presented which would be constructed in the form of graph. There a search buttons have been used which are for taking training data for example Domain search and paper type search, there are many domain available like information technology, data mining, software engineering, biology etc. and in the papers there are review paper, design paper, implantation paper are used to get latest information regarding a particular topic. **G.A. algorithm**-G.A. algorithm stands for genetic algorithm. The use for G.A. algorithm in this project for searching the seminal papers from a folder of papers , an algorithm starts from step1 start- in the starting process number of training sets inserted, start steps shows suitable solution of training data sets. Step 2 fitness, it is using for data evolution process from paper folder. Step3 selection, after data evolution process, selection of seminal papers are depended on the matching of data, there is bigger chance to evolution of seminal papers depend on better fitness. Step 4 Cross over and mutation are used to find other seminal paper from there references and citations. After that testing process to be done which stops the G.A. algorithm and return the best solution from given population. The large main graph has been constructed after applying this algorithm where a main paper is connected to other child papers

. The **KNN and Navie baise** algorithm is used to find subgraph from a large main graph. year wise paper selection is used to find subgraph from a large main graph. In proposed system a subgraph is generated by using KNN and Navie

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

baise algorithm, this subgraph is the collection of many relevant papers, on one click the useful selected paper will be opened up by user.
This process is very easy to find out seminal papers in a form the of graphs, so many papers are constructed only the help of many vertexes and nodes, which consume time of user to search similar papers in a particular domain.

V. RESULT

According to the proposed system the folder of different seminal papers stored in desktop as a form of documentation. A large graph has been generated by the using of genetic algorithm which is the main graph of proposed system, which shows all seminal papers of same or different citations, this large graph contains some subgraph which shows the genealogy of those all papers which had been selected by users, the subgraph will be creating from the main graph.

- **Algorithm**

R and L are conditionally independent given M if for all x,y,z in {T,F}:
P(R=x | M=y ^ L=z) = P(R=x | M=y)
More generally:
Let S1 and S2 and S3 be sets of variables.
Set-of-variables S1 and set-of-variables S2 are conditionally independent given S3 if for all Assignments of values to the variables in the sets, P(S1's assignments S2's assignments & S3's assignments)= P(S1's assignments S3's assignments)
P (A|B) = P(A ^B)/P(B) '
Therefore
P (A^B) = P(A|B).P(B) – also known as
Chain Rule
Also P (A^B) = P (B|A).P (A)
Therefore P (A|B) = P (B|A).P (A)/P(B)
P (A, B|C) = P (A^B^C)/P(C)
= P (A|B, C).P (B^C)/P(C) – applying chain
rule
= P (A|B, C).P (B|C)
= P (A|C).P (B|C)
, If A and B are conditionally independent given CThis can be extended for n values as P (A1,A2...An|C) = P(A1|C).P(A2|C)...P(An|C) if A1, A2...An are conditionally independent given C.

Result Description

- **Data Collection-** All IEEE Researched papers have been selected in proposed method, which is taken as a data base in this proposed system.
- **Process phase -**Pre-processing part has been done in this step. Some specific numbers are given for all selected papers. According to their type of domain and title that numbers are concatenate to the title of papers. And at last .pdf has been connected to the name of papers name.
- **Keyword Extraction-**Keyword extraction is used to finding out the similarities between papers, where input key words are given by users.

Analysis

- **Scalability**- This proposed system is more scalable than earlier system. There is much capacity to put many papers in a folder of using system.
- **Time Complexity**-The time complexity of proposed method (navie bayse) is O(N) big O N, where N is the Number of training data.
- **Experimental Set up**-There Windows 8, Visual Studio 12,SQl 12, ASP.NET MVC are used to perform the proposed sytem.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

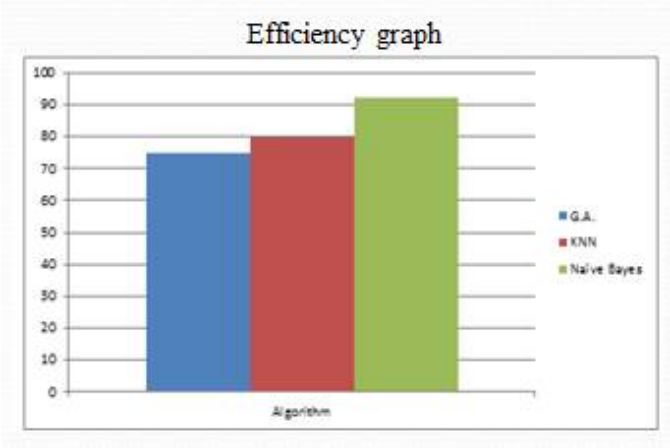

Fig 2-Efficiency of proposed algorithm

	A	B	C	D
1	Algorithm	G.A.	KNN	Naïve Bayes
2	Algorithm	75	80	92

Table 1- Efficiency in percentage

The efficiency of searching has been increased as compare to documentation searchand the degree of accuracy of navie baise algorithm is higher (by fig 2) to other proposed method which were applied in this proposed system and the visualization of graph is a lot of fine than earlierwork.

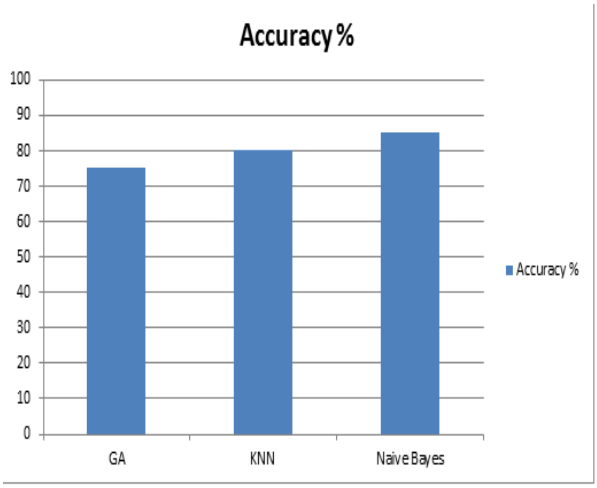

Fig 3- Accuracy comparison between proposed method

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

	A	B
1	Algorithm	Accuracy %
2	GA	75
3	KNN	80
4	Naive Bayes	85

Table 2- Accuracy in percentage

After processing no. of input papers in proposed methods, so these percentages have been created. This show G.A. algorithm gives 75% , KNN gives 80% and Naïve bayes gives 85% accuracy

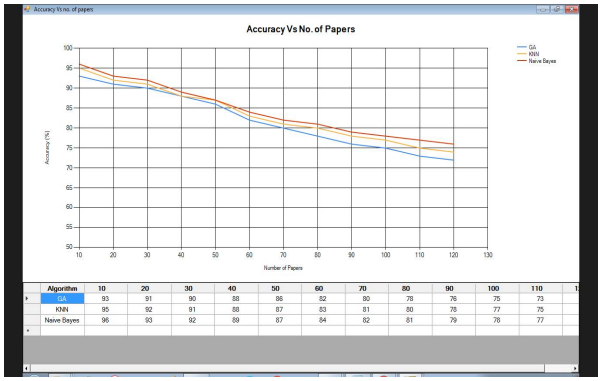

Fig 3- Accuracy Result of proposed work

This graph can show number of papers which contains their seminal paper from references papers as well as from the citation of that user selected paper.
Accuracy will be fine by this formula-

Accuracy =Matching with labled data set *100 / Total number of papers

When number of papers were 10 than accuracies of G.A. Algorithm, KNN algorithm, Navie Baise algorithm were 93%,95%,96% respectively. As well as when number of papers 110 than accuracies a of G.A. Algorithm, KNN algorithm, Navie Baise algorithm were 73, 75,77respectivly as shown in below screen shot.

VI. CONCLUSION

Anew innovation has been done from this project which is useful to search seminal papers in the form of graph visualization. Graph visualization is more easier than documentation searching, by this proposed work user can easily work on different domain and different seminal papers in a particular domain. Discovery of those relationships can benefit many interesting applications such as expert finding and research community analysis.
Our Proposed System is a windows based system in which all algorithm work in uploaded papers those uploaded papers collected in a folder of desktop. Those all papers become shown as a form of graph which would be very easier to new users and researchers to find out latest topic as per their requirement and interest. In upcoming time it would be connected to online authorise citations by which only in one click on this project user can get a lot of seminal papers of authorized citation. This proposed work gives efficient result and high performance to search the seminal papers as well as relevant papers year wise pattern in different citations.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

REFERENCES

[1]"Distance-join: Pattern match query in a large graphdatabase"PVLDB, vol.2,no.1, 2009.
[2]"Fast graph pattern matching," in Data Engineering,2008. ICDE 2008. IEEE 24th International Conference on. IEEE, 2008, pp. 913–922
[3]"Tale: A tool for approximate large graph … .matching," in ICDE, 2008.
[4]"Torque: topology-free querying of protein networks," Nucleic Acids Research, vol. 37, no. suppl 2, pp. W106–W108, 2009.
[5]"Saga: a subgraph matching tool for biological graphs," Bioinformatics, vol. 23, no. 2, pp. 232–239, 2007.
[6]"On graph query optimization in large . .networks",PVLDB, vol. 3, no. 1-2, 2010.
[7]"Efficient subgraph matching on billion node ...graphs" PVLDB, vol. 5, no. 9, 2012.
[8]"Node similarity in the citation graph" ,Knowledge ..and Information Systems, vol. 11, no. 1, pp. 105–.129, 2006.
[9]"DBpedia: A nucleus for a web of open data", in ISWC, 2007.
[10]"Tran. An efficient implementation of graph …grammars `based on the RETE matching … \algorithm." In Proc. 4th .Int. Workshop on …Graph-Grammars and their Application .to Computer Science and Biology, volume 532 of ….Lecture Notes in Computer Science, pages 174–….189. .Springe.r-Verlag, 1991.
[11]"Review on Natural Language Processing Tasks …..for Text .Documents", IEEE International…..Conference on on Computational Intelligence and …..Computing Research. (ICCIC), 2014.
[12]"An efficient algorithm for similarity joins with ….edit .distance constraints. PVLDB", 1(1):933–. …944, 2008.
[13]"Efficient merging and filtering algorithms for … ….approximate string searches", In ICDE, pages ….257–266, .2008.
[14]"Template Based Semantic Similarity for Security ....Applications", pages 621–622. Springer, 2005 .
[15]"Sub graph Matching with Set ….Similarity in a Large Graph Database" ,IEEE ….Transactions on Knowledge and Data ….Engineering, (Volume:PP , Issue: 99),12 January ….2015
[16]"On Constructing Seminal Paper Genealogy", ….IEEE .Tranacactions on Cybernetics …,VOL41,NO,1,January2014
[17]Shimul Sachdeva University of California, Berkeley," ..Family Tree Visualization" Washington, DC, University of California USA, p. 43, 2001 .
[18]"Keyword Query Routing" IEEE Transactions …on ..Knowledge and Data Engineering,, VOL. …26, NO. 2, ..February 2014 363.
[19]"Design of Hierarchical genealogy with subgraph mining" International Journal of Engineering Applied Sciences and Technology, 2017Vol. 2,Issue 3, ISSN No. 2455-2143, Pages 1-4
[20]Entity resolution in disjoint graphs: An .. .application on genealogical data Hossei...Rahmania,b,∗, Bijan Ranjbar-Sahraeib, ...GerhardWeissb and Karl TuylscIntelligent Data ....Analysis 20 (2016) 455–475 455 DOI …10.3233/IDA-160814 IOS Press.
[21]R. Agrawal and R. Srikant. Fast algorithms for ....mining association rules in large databases. In ...VLDB, pages 487–499, 1994.
[22]W. T. Tutte, "Convex representations of graphs," …Proceedings f the London Mathematical Society, …Third Series, no. 10, pp. 304–320, 1960.
[23]"How to draw a graph," Proceedings of the ….London Mathematical Society, Third Series, no. ….13, pp. 743–768, 1960.
[24]L. Zou, L. Chen, and M. T. Ozsu, "Distance- ....join: Pattern match query in a large graph ...database,"PVLDB,vol.2,no.1, 2009.
[25]"Fast graph pattern matching," in Data ….Engineering, J. Cheng, J. X. Yu, B. Ding, P. S. and H. ...Wang, 2008. ICDE 2008. IEEE 24th ..International Conference on. IEEE, 2008, pp. …..913–922
[26]"Implementation of Paper Genealogy in Subgraph Mining" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 5 Issue: 1 199 –202199.