GRAPH ANALYSIS AND COMMUNITY DETECTION IN CITATION NETWORKS

Ayushi Dalmia

GRAPH ANALYSIS AND COMMUNITY DETECTION IN CITATION NETWORKS

A PROJECT REPORT
Submitted By
Ayushi Dalmia (095173)

Under the Supervision
Of
Prof. (Dr.) Niloy Ganguly
Indian Institute of Technology, Kharagpur
Department of Computer Science and Engineering

as part of
Summer Internship, 2012
for partial fulfilment of the requirements
for the degree
of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING
HERITAGE INSTITUTE OF TECHNOLOGY, KOLKATA

November, 2012

DECLARATION

I certify that

a. The work contained in this report is original and has been done by me under the guidance of my supervisor.
b. The work has not been submitted to any other Institute for any degree or diploma.
c. I have followed the guidelines provided by the Institute in preparing the report.
d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
e. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Signature of the Student

(AYUSHI DALMIA)

ACKNOWLEDGEMENT

I would take this opportunity to thank Prof. (Dr.) Niloy Ganguly, Associate Professor, Dept. of Computer Science and Engineering,IIT Kharagpur for constantly supporting me and guiding me with his valuable insights.

I thank my mentor Mr Tanmoy Chakraborty, Phd Scholar, CNeRG Group, IIT Kharagpur for his constant support and guidance during the project. I do not know where would I have been without him.Thank You Sir !

I will be thankful to Prof Subhashis Majumder, HOD,Dept. Of Computer Science and Engineering, Heritage Institute of Technology for allowing me to do my summer internship at the prestigious Indian Institute of Technology, Kharagpur.

Last but not the least I thank all my friends for their cooperation and encouragement that they have bestowed upon me.

ABSTRACT

This project report addresses two problem statements with regard to citation networks.

Part A

The structural modularity and compartmentalization of a complex network is closely related to the dynamics toward clustering. Most of the community detection algorithms are inefficient at capturing overlaps in-between communities, detecting communities having disparities in size and density, and taking into account the modular structure of multipartite networks. In this experiment, for the first time, citation network has been visualized as a tripartite hypergraph. We detect the overlapping community structure simultaneously from the three partitions. Our algorithm modularizes the hyper-edges using unipartite community detection algorithm after converting citation hypergraph into its corresponding weighted line-graph, and hereby produces the overlapping communities of three partitions. We illustrate its efficiency through extensive experiments on synthetic as well as large scale real citation data of computer science and physics domains, and then compare it with existing state-of-the art unipartite overlapping community detection algorithm performing on dynamic networks.

Part B

In the field of research and analysis the importance of a paper is heavily dependent on the number of citations received by the paper. Authors, in general have the tendency to cite those papers which are written by them or one of their collaborators. Doing this increases the impact factor of papers .Also, if we consider the most cited author it happens that the popularity of the author increases due to these self citations and collaboration citations. In this work we studied and delved into answering several questions pertaining to this issue.

TABLE OF CONTENTS

Title	i
Declaration	ii
Certificate	iii
Acknowledgement	iv
Abstract	v
Part A	
Chapter 1.Introduction	1
Chapter 2.Proposed Algorithm	4
Chapter 3.Conclusion and Future Work	8
Chapter 4.Bibliography	9
Part B	
Chapter 1.Introduction	10
Chapter 2.Proposed Algorithm	12
Chapter 3.Dataset	14
Chapter 4. Conclusion and Future Work	17

PART A

Chapter 1

Introduction

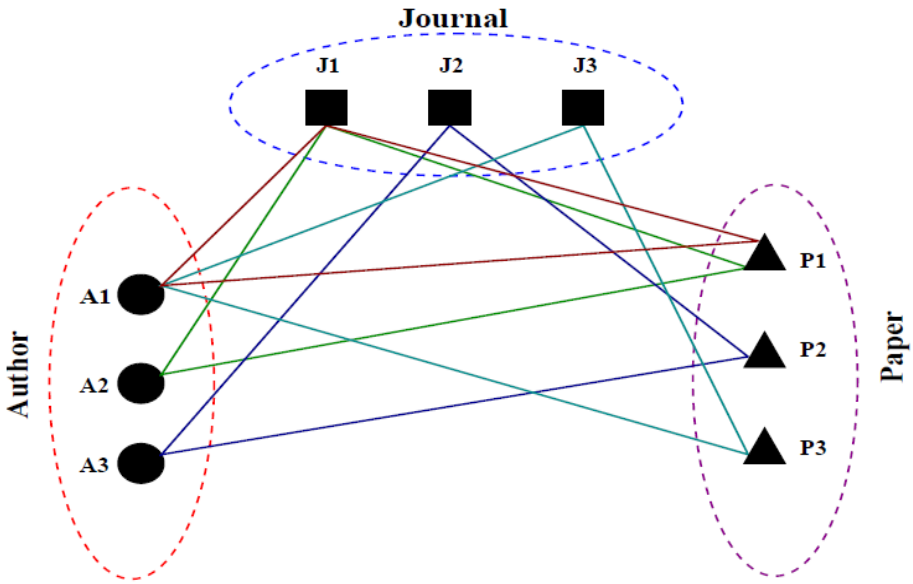
Many large scale dynamic complex networks can be described through the intricate web of connections among the units they are made of. The interpretation of the global organization of such networks as the coexistence of their a priori structural subunits (communities) associated with more highly interconnected parts has always been a demanding question to the researchers. The existing deterministic methods used for large networks find separated communities, while most of the actual networks are made of highly overlapping cohesive groups of nodes. Here we introduce an approach to detect the interwoven sets of overlapping communities making a step towards the uncovering of the modular structure of complex systems. We apply an efficient technique to explore overlapping communities on large scale citation network. The phenomenon of community overlapping is quite obvious in social perspective. For example, in personal network, a person may be a member of multiple communities like his family, his co-workers, his friend circle etc which may raise the problem of network of networks. In the field of research, this tradition is also observed now-a-days where the most of the researchers' field of interest cross a single frontier and spread into multiple research directions. In citation network, whatever may be the granularity of analysis, research domains may be identifiable, but their boundaries are quite difficult to characterize, due to the existence of such overlaps or to the existence of scientific communities at the border of the well identified domains or even at the crossroad in-between different domains in terms of interdisciplinary activities. To capture the multi-domain interests of the researchers, journals/conferences, and versatility of research articles into multiple research fronts, we visualize citation network as a tripartite hypergraph consisting of authors, publications and publication venues (journals/conferences) in three partite sets respectively with unequal size. Though this hypergraph structure is a general visualization of citation network, the in-depth composition of its components may analytically suggest its true terminology as Authorship Network where there is no intrinsic or extrinsic realization of citation in the actual tripartite network rather only the authorship information along with publication venue is involved with each hyper edge. Each hyper edge denotes a paper written by an author and published in a journal. Though a node in a network can be associated tomultiple semantic topics, a link is usually associated with only one semantics - for instance,

an author can have multiple research interest, but his every publication generally focuses on single direction, which in turn may act as an important resource for several research communities later. Link clustering algorithms utilize this notion to detect overlapping communities, by clustering links instead of the more conventional approach of clustering nodes. Though each link is placed in exactly one link cluster, this automatically associates multiple overlapping communities with the nodes since a node inherits membership of all the communities into which its links are placed. The advantage of overlapping community detection in hypergraph structure of citation network through edge classification via line graph rather considering traditional directed citation network structure is manifold. Firstly, multipartite hypergraph structure of citation network can be able to detect overlaps simultaneously from all partitions, which is not possible in any unipartite overlapping detection algorithm. Secondly, it directly relies on the authorship information of an author rather than the citations to his papers, which we think more straight forward consideration for detecting author's multiple research-domain overlaps. Thirdly, general overlapping detection algorithms in dynamic network are prone to modularize vertices using link density information, where the underlined semantics of the links may suffer from exploring even more valuable information of compartmentalization. Finally, developing error-free citation networks for various research domains are very challenging as the data set has to be bounded between some fixed time interval, which indirectly discards several citation information (mainly the citing information of the older articles and cited information of newer articles) from the data set. Along with this, proper indexed and well-structured citation data set construction is itself computationally tedious. On the other hand, authorship network development is subsequently easy task. Though in this experiment, co-citation and bibliographic coupling information are used to tag the weights of the links in the line graph, it may not be essential for other applications where only the graph-neighbour intersection is sufficient calculate weights from the hypergraph.

TRIPARTITE HYPERGRAPH

Generally, citation network structure can be visualized as graph G = (V, E) comprising a set V of nodes, which each node N_i representing a research paper R_i and a set E of directed edges, with each edge E_{ij} directed from the citing node N_i to the cited node N_j . There are two common and significant features of any typical citation network: first, it is directed and

cyclic; second, when it evolves over time, only new nodes and edges are added, and none are removed. It is formally represented as a list of edges comprising tuples of two end nodes of an edge. The detailed description of this network is given below. As shown in the following figure, a citation network is modelled as a tripartite hypergraph (more specifically 3 uniform tripartite hypergraph) G = (V, E) where the vertex set V consists of 3 partite sets V_a, V_p and V_j . Each hyper edge e (represent a paper) in hyper edge set E connects a triple of nodes (a, b, c) where a ε V_a, b ε V_p and c ε V_j .It indicates that the paper named as b is written by author a and published in journal c. It is important to note that if one paper is written by multiple authors then it is represented by multiple edges having common vertex in paper and journal partitions



Some of the typical properties of this network are as follows:

• The size of each partite set is uneven with the relation |Va| ≤ |Vp| ≤ |Vj | follows generally in every citation network (sometime equality in both side may be applicable).

• The relation between papers to journal partition is one-to-one, but the reverse may be one-to-many; and for author-paper and author-journal partitions, it is generally one-to-many.

• Edges which share a common vertex in paper partition, must share a common vertex in conference side also and vice-a-versa.

Chapter 2

Proposed Algorithm

In this section, we propose our overlapping community detection algorithm by clustering links of the hypergraph to detect the overlaps of the vertices. Our proposed link clustering algorithm for detecting overlapping communities in tripartite hyper graphs, named as 'Overlapping Citation Hypergraph' algorithm (abbreviated to 'OCH') is as follows.

Taking direct reference of the hypergraph notation discussed in section 3, for a given hypergraph G, we compute the weighted line graph G' which is a unipartite graph in which the hyper edges in G are nodes, and two nodes e_1 and e_2 in G' are connected by an edge if e_1 and e_2 are adjacent in G (i.e. the two hyper edges have at least one common node in G). The weight of the edge (e_1, e_2) in G' represents the similarity between the two hyper edges e_1 and e_2 in the hypergraph G, which is computed as follows. Let $N^X(i)$, $N^Y(i)$ and $N^Z(i)$ denote the set of neighbours of node i of type V_X, V_Y and V_Z respectively (if $i \varepsilon V_X$, then $N^X(i) = \Phi$ since nodes in the same partite set are not linked). Similarity between two adjacent hyper edges $e_1 = (a, b, c)$ and $e_2 = (p, q, r)$ (where $a, p \varepsilon V_X$; $b, q \varepsilon V_Y$; $c, r \varepsilon V_Z$ and assumed a = p) is measured by the relative overlap among the neighbours of the non-common nodes of the same type:

$$\alpha(e_1, e_2) = \frac{|S \bigcap S'| + |N^Y(c) \bigcap N^Y(r)| + |N^Z(b) \bigcap N^Z(q)|}{|S \bigcup S'| + |N^Y(c) \bigcup N^Y(r)| + |N^Z(b) \bigcup N^Z(q)|}$$

where $S = N^X(b) \cup N^X(c)$ and $S' = N^X(q) \cup N^X(r)$. Non-adjacent hyper edges are considered to have zero similarity.

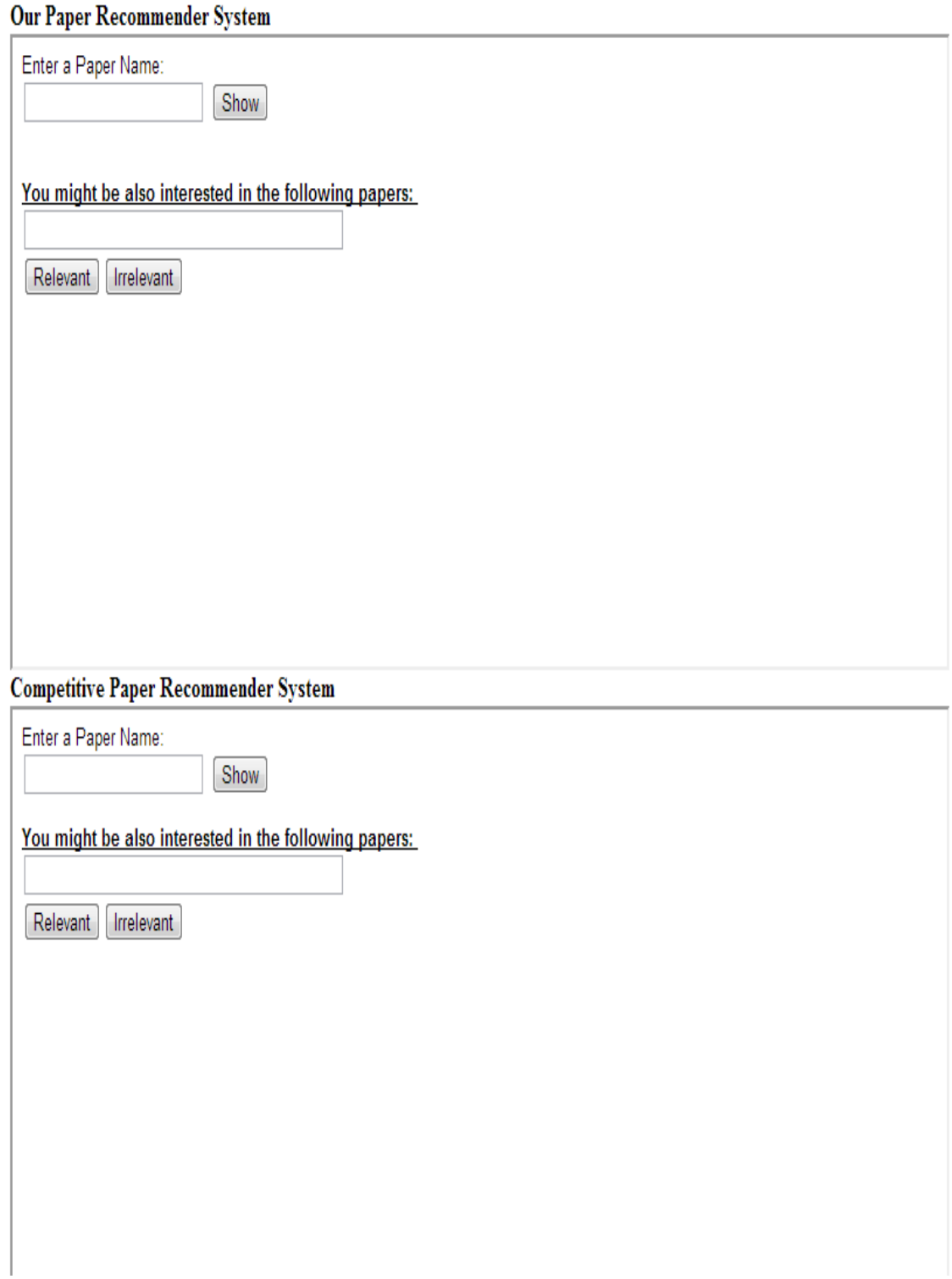
Once the weighted line graph G' is constructed from the given tripartite hypergraph G, any community detection algorithm for unipartite graphs (even the ones which do not produce overlapping communities) can be used to cluster the nodes in G'(i.e. the hyper edges in G).

We used the Infomap algorithm as this algorithm has been found to identify communities accurately as compared to several other algorithms. We are yet to compare this algorithm with other benchmark algorithm. This is because in this experiment, the problem lies in the ground of tripartite structure of citation network, where, to the best of our knowledge, there is no proposed algorithm which can detect overlapping communities simultaneously from tripartite partitions.

RECOMMENDER SYSTEM

We can use the metrics for testing our algorithm. Given a set of true clusters and the set of clusters found by an algorithm, these sets of clusters must be compared to see how similar or different the sets are. Various metrics have been proposed by previous researchers one of them being NMI. We also plan to design a simple recommender system which will help to test how the algorithm performs.

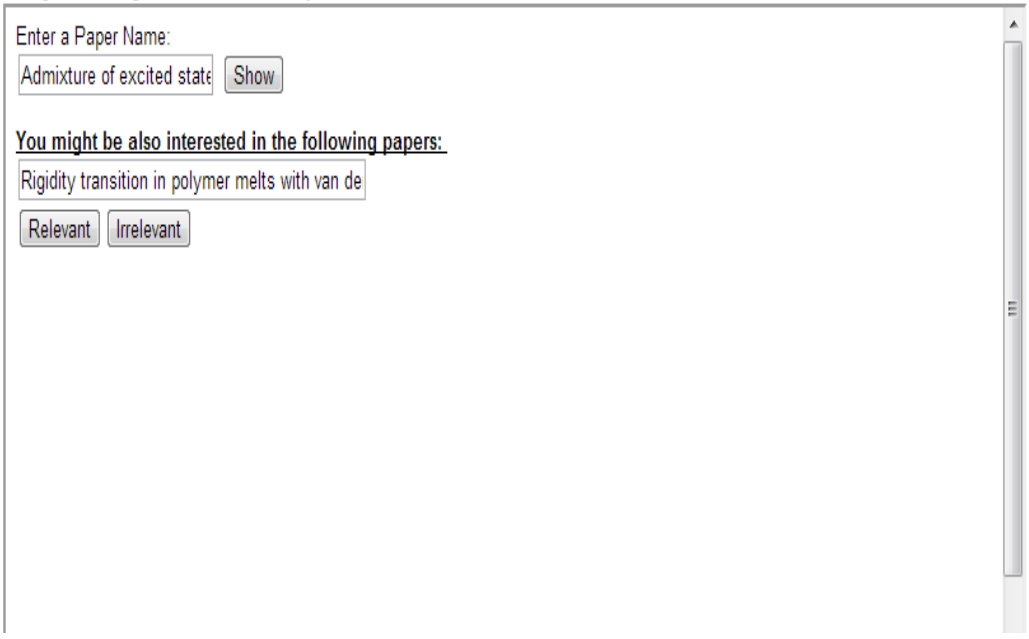
The following is the snapshot of the test recommender system built so far:


Our Paper Recommender System
Enter a Paper Name:
Show
You might be also interested in the following papers:
Relevant
Irrelevant
Competitive Paper Recommender System
Enter a Paper Name:
Show
You might be also interested in the following papers:
Relevant
Irrelevant

7

Our Paper Recommender System

Competitive Paper Recommender System



Chapter 3

Conclusion and Future Work

In this work, we proposed the first algorithm to detect overlapping communities considering the full tripartite hypergraph structure of citation networks. In this large reserve of research papers, it is difficult for an individual user to find resources of her interest. Our algorithm successfully groups nodes into multiple communities where each community represents a topic of interest. Based on these interests users can find out related resources. Thus the proposed algorithm can be effectively used in recommending interesting resources to users in citation networks. Building such a personalized recommendation system taking advantage of the effectiveness of the proposed algorithm comprises the future work. Also, we would like to test and compare how effective the algorithm is compared to other bipartite/unipartite community detection algorithms.

Chapter 4

Bibliography

[1] Abhijnan Chakraborty, Saptarshi Ghosh and Niloy Ganguly *Detecting Overlapping Communities in Folksonomies*

[2] Nicolas Neubauer and Klaus Obermayer. *Towards Community Detection in k-Partite k-Uniform Hypergraphs*, pages 1–9. 2009.

[3] T. S. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Phys. Rev. E*, 80:016105, 2009.

[4] X. Wang, L. Tang, H. Gao, and H. Liu. *Discovering overlapping groups in social media.* In ICDM, pages 569–578, 2010.

PART B

Chapter 1

Introduction

"A scientific paper does not stand alone; it is embedded in the 'literature' of the subject"- Ziman.

While writing a research paper, it has always been a tradition to acknowledge the papers referred during the research. This is called referencing a paper. The paper receives acknowledgement through this citation. The paper which is referring is called citing paper and the paper referred to is called a cited paper. If we study the scientific papers closely it can be observed that a complex network can be constructed from the citing and cited papers.

The impact factor (IF) of an academic journal is a measure reflecting the average number of citations to recent articles published in the journal. Therefore higher the number of citations received by a paper the higher is its impact factor. If we closely observe the citation network we find that the authors have the tendency to cite their own papers or one of their collaborators. This affects the impact factor of the paper highly. Hence in this work we try to build the network by removing those citations and later try to analyse and compare this network with the original citation network.

Before the algorithm let me first explain the terminologies used which will help to understand the algorithm in an even better manner.

Definition 1:

SELF CITATION: Let A denote the set of authors who have written a paper, say X and B denote the set of authors of the paper, say Y.A citation is considered to be self citation if paper X cites paper Y and $A \cap B \neq Ø$

Definition 2:

CO-AUTHOR: An author A is said to be a *Co-author* of another author B, if A and B have worked together at any point of time. Thus for every author we will have a *Coauthor Set*

Definition 3:

PAPER CO-AUTHOR: Authors A and B are considered *Paper Coauthor* of *paper* X, if A and B have worked together to write the paper. Thus for every paper we will have a *Paper Coauthor Set.*

Definition 4:

COLLABORATIVE CITATION: Consider that paper Y is cited by paper X. Let A denote the set of authors who have written a paper, say paper X. Let B denote the set of co-authors of all the authors of another paper, say Y. Now if $A\text{-}B = \Phi$, we say that that the citation is collaborative citation. This is because if $A\text{-}B \neq \Phi$, it implies that paper Y has an author who has never worked before with any of the authors of paper X.

Chapter 2

Dataset

We have two datasets, one form the field of Computer Science (CS) and another from the field of Physics. The Physics dataset was present in the xml format with inappropriate tags. It has been converted in the form of CS dataset. The following is the structure of the original dataset of Computer Science and the dataset for Physics:

#@:Denotes Author List

#index: Denotes Paper Id

#t: Denotes Year of Publication

#*: Denotes title of the paper

#c: Denotes the conference

#%: Denotes Cited Paper-Id

In addition to this information we have the following additional tags in the CS dataset:

#!: Denotes Abstract

In addition to this information we have the following additional tags in the Physics dataset:

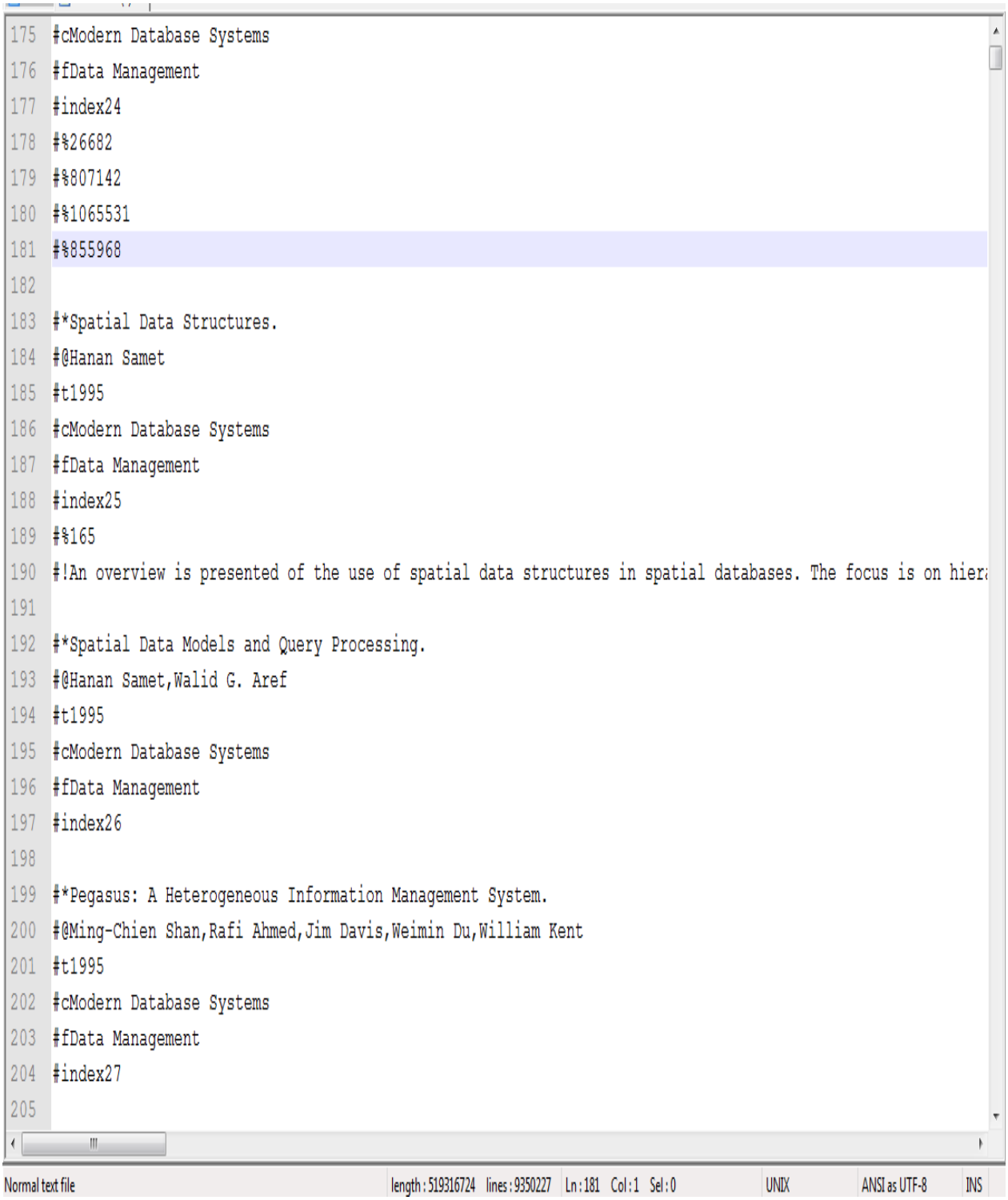
#np: denotes number of pages

#fp: denotes first page

#lp: denotes last page

#$: denotes affiliation

#&: denotes article type



```
175 #cModern Database Systems
176 #fData Management
177 #index24
178 #%26682
179 #%807142
180 #%1065531
181 #%855968
182
183 #*Spatial Data Structures.
184 #@Hanan Samet
185 #t1995
186 #cModern Database Systems
187 #fData Management
188 #index25
189 #%165
190 #!An overview is presented of the use of spatial data structures in spatial databases. The focus is on hier
191
192 #*Spatial Data Models and Query Processing.
193 #@Hanan Samet,Walid G. Aref
194 #t1995
195 #cModern Database Systems
196 #fData Management
197 #index26
198
199 #*Pegasus: A Heterogeneous Information Management System.
200 #@Ming-Chien Shan,Rafi Ahmed,Jim Davis,Weimin Du,William Kent
201 #t1995
202 #cModern Database Systems
203 #fData Management
204 #index27
205
```

Normal text file | length : 519316724 lines : 9350227 | Ln : 181 Col : 1 Sel : 0 | UNIX | ANSI as UTF-8 | INS

Figure 1 Snapshot of the Dataset

Chapter 3

Proposed Algorithm

STEP 1: PREPROCESSING

For our problem set, we need only the following information for each paper-id from the dataset.

i)Author List

ii)Paper Id

iii)Year

From the given dataset, we extract only this information from the dataset.

It is important to note that due to inefficient dataset, we need to remove those papers which do not have author list or do not have the year of publication. After removing such papers, it might so happen that a paper cites a paper which has been removed in the previous step. We need to remove such citations too.

STEP 2: BUILD AUTHOR LIST

In the next step we build the author list. In this step, we find out the list of unique authors in the entire dataset. It is important to note that the same author may have different representation of his/her name.

For example, the author Jose A. Berkley may write his name as Berkley A.J. in one paper, J.A.Berkley in another and might use his pen name in some other paper.

No measure has been taken to remove such duplicity. These three authors are considered three unique authors. This might be a problem as it might reduce the importance of an author .Maybe some measure will be taken later to remove this duplication .We proceed to the next step with this AUTHOR LIST considering that the author names are uniquely identified.

STEP 3: BUILD CO-AUTHOR LIST:

In this step we scan the entire dataset for each unique author in the AUTHOR LIST and then create the COAUTHOR LIST. This is done as follows:

We pick an author from the AUTHORLIST with id say A1.Next we find out the author names with whom author A1 has worked with in the entire dataset, find their ids from the AUTHOR LIST and write these ids against author A1.This is done for all the authors in the AUTHOR LIST. So a line in the COAUTHORLIST will look like this

A0 A1, A2, A5, A39

The above line will imply that the author with ID A0 has worked with authors having ids A1, A2, A5, A39. In this way we create the COAUTHORLIST for all the authors in the AUTHORLIST.

STEP 4: PAPER AUTHOR LIST

Since every paper has a unique id in the paper we use this information in the PAPER AUTHOR LIST. The PAPER AUTHOR LIST is created as follows:

For every paper we use the index of the paper in preparing the paper id. So, for example if a paper X has paper index #index2301,then its paper id is P2301.Now for every paper we find out the name of the authors of the paper, map it to the AUTHOR LIST and find out the author id. Thus a line in the PAPER AUTHOR LIST will look like this:

P1 A3, A5, A9

This means paper with index 1 is written by authors with author id A3,A5 and A9.

STEP 5: PAPER AUTHOR UNION LIST

In this step for every paper in the dataset, we know the authors from the PAPER AUTHOR LIST. Also, for every author for a paper we know the co-authors from the CO-AUTHOR LIST. Therefore now we create another list which will have the paper id and the list of co

authors of each author of the paper. Thus a line in the PAPER AUTHOR UNION LIST will look like this:

P1 A1, A2, A5, A7, A8, A9

Thus A1, A2, A5, A7, A8, A9 are those authors who have either written paper P1 or worked with the authors of the paper P1.

STEP 6: BUILD THE REQUIRED CITATION NETWORK

Now using the lists of AUTHORS, CO-AUTHORS and PAPER AUTHOR UNION we can easily build the desired citation network. This is as follows:

We scan the entire dataset as follows:

Consider the following instance of the data:

#*Specification and Execution of Transactional Workflows.
#@Marek Rusinkiewicz, Amit P. Sheth
#t1995
#cModern Database Systems
#fData Management
#index24
#%26682
#%807142
#%1065531
#%855968

We scan the papers which have been cited by the instance. If any of the authors of any of the citations is in the paper author union list of this instance we remove that citation.

In this way we successfully build the citation network without any self citation or collaborative citation.

Chapter 4

Conclusion and Future Work

Calculation of the actual impact factor of a paper is important. It adds to the quality of the paper. In this work we try build a network from the citation network such that it contains no self or collaborative citations. Understanding the nature of this network and later comparing this with the original citation network would be our future work.