Analyzing the Socio-Cognitive Structure of an Economic Research Community

Tiwana, Birinder and Sarkar, Sudeshna

Indian Institute of Technology Kharagpur -> Department of Computer Science and Engineering

6 May 2009

Analyzing the Socio-Cognitive Structure of an Economic Research Community

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF

BACHELOR OF TECHNOLOGY (HONOURS)

IN

COMPUTER SCIENDE AND ENGINEERING

By

Birinder Singh Tiwana
05CS1014

Under the guidance of
Prof. Sudeshna Sarkar

Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

May 2009

Acknowledgement

I offer my sincerest gratitude to my supervisor, Prof Sudeshna Sarkar, who has supported me throughout my project with her guidance and knowledge whilst allowing me the room to work in my own way. She exposed me to the research topic through proper counsel rigorous discussion and always showed great interest in providing timely support and suitable suggestions. Without the help, support and patience I received from my guide, this thesis would never have materialized.

Birinder Singh Tiwana
05CS1014
May 2009

CERTIFICATE

This is to certify that the thesis titled "**Analyzing the Socio-Cognitive Structure of an Economic Research Community**" is bona fide work carried out by **Birinder Singh Tiwana (05CS1014)** under my supervision and guidance for the partial fulfillment of the requirements for Bachelor of Technology Degree in Computer Science & Engineering during the academic session 2008-20009 in the Department of Computer Science & Engineering, Indian Institute of Technology, Kharagpur.

Prof. Sudeshna Sarkar
Department of Computer Science &
Engineering
Indian Institute of Technology, Kharagpur

ABSTRACT

Research papers published by authors in various fields give rise to huge datasets. A social network of the authors can be built up using these datasets by treating the references between papers as the links between the authors publishing those papers. These networks also have a cognitive structure to them based on the themes that various authors discuss. In this thesis we describe the procedure of building up a bibliographical tool for presenting a dataset of research papers in a particular field. We show the results that this tool produces on the dataset downloaded from the Repec website. We also describe a few mathematical measures that analyze the relationship between the cognitive and social structure of the network created from this dataset. The results produced by this analysis are in coherence with the results produced by different researchers in the past.

TABLE OF CONTENTS

CHAPTERS *PAGES*

CHAPTER1

INTRODUCTION

In this introductory chapter, we shall first give some basic definitions and concepts that are relevant or are used in the work presented in subsequent chapters. Then we shall define the problem statement. Finally we will give an overview of the work done in this thesis.

1.1 Social Network

A social network is a social structure that is a made of nodes (which are generally individuals or organizations) that are tied by one or more specific types of interdependency, such as values, visions, ideas, financial exchange, friendship, kinship, dislike, conflict or trade[1]. Social network analysis views social relationships in terms of nodes and ties. Nodes are the individual actors within the networks, and ties are the relationships between the actors. There can be many kinds of ties between the nodes. For ex, we can build up a social network from the people shopping at a camera shop. The nodes of the network will be the individuals doing the shopping and a link between two nodes will indicate that they have bought the camera of the same model. These concepts are often displayed in a social network diagram, where nodes are the points and ties are the lines. In our project we are interested in the social network that is formed by various researchers working in the field of economics. The nodes in our network will be the authors. A link from node A to node B will indicate that author corresponding to node A has referenced the author corresponding to node B in one of his papers.

1.2 Centrality in a Social Network

Measuring the network location of a node is referred to as finding the *centrality* of a node. These measures give us insight into the various roles and groupings in a network -- who are the connectors, mavens, leaders, bridges, isolates, where are the clusters and who is in them, who is in the core of the network, and who is on the periphery. There are three popular measures that help us in determining the centrality of a node in the network.

Betweenness centrality: - Betweenness centrality measures the extent to which a node lies between other nodes. It is defined as the ratio of the sum of the shortest paths connecting two nodes via the given node to the total number of existing shortest paths. This key figure shows the involvement of an actor in relation to other actors regardless of the direction of the relation. Actors with a high betweenness centrality have the potential to control communication in a network and to take the role of a coordinator in group processes [2].

Degree centrality: - Degree centrality of a node is a direct measure of the centrality of node. It is defined as the number of direct links that a node has to other nodes. The number of edges incident upon a node is referred to as its *indegree*. The number of edges a node directs towards its neighbors is referred to as its *outdegree*. The degree centrality of a node is defined as the sum of its indegree and its outdegree.

Closeness Centrality: - Closeness centrality of a node in a network is defined as the inverse of the sum of geodesic distance to all other nodes [3]. Nodes that are “shallow” to other nodes (that is, those that tend to have short geodesic distances to other nodes within the graph) have higher closeness. Closeness is referred in network analysis to mean shortest-path length, as it gives higher values to more central nodes.

1.3 Cognitive Analysis of the Network

In a network different authors discuss different ideas. For ex, in a network formed by the researchers of Computer Science, some authors might be discussing ideas about Data

Mining while the others might be discussing about Compilers. The division of the nodes in the network into various categories based upon the ideas that the authors corresponding to the nodes in each category discuss is called a cognitive structuring of the network. This structuring helps us in grouping the similar authors together and studying the characteristics of the network formed by them separately. The knowledge being created and discussed in a research field is reflected in the field's cognitive structure [4]. We can study the flow and development of new ideas in the network. The cognitive structure is constantly changing, as different topics become relevant for the discipline [5]. There are people in a network who bring in new ideas. We can find out whether there exist some relationships in the social position of an author in a network and the topic that he discusses or not.

1.4 Problem Statement

We use the dataset available online at RePEc website. This is one of the largest bibliographic dataset available on the research papers in the field of economics. Most of the papers can be downloaded freely from the internet. Some salient features of the dataset are as follows:-

- 258023 papers are available on it.
- 126354 papers have references to other papers that are available on this website.
- These papers consist of works from more than 18000 authors.

We have the following two aims:-

1. We want to build up a tool that will help the users to access various papers of the database in a convenient manner. We will build up a social network on the authors in the network. We will calculate the centrality value of each author in the network. We will then cluster the authors based on the topics they discuss. Then we will provide a website where the information about authors corresponding to various categories and their papers can be accessed. A user will also be able to

know the various topics that were important during various periods of time. The important authors during each period of time will also be shown on the website. In other words we will try to present a complete cognitive, social and temporal analysis of the dataset on a website. This tool can be used with minor changes in the parsing techniques for the data for any other dataset of research papers.

2. We will also try to find out if there is a relationship between the social and cognitive structure of the network. We will calculate the centrality values of the various clusters of the dataset formed on the basis of the topics that are being discussed in them. Then we will do some mathematical analysis that will help us to answer the questions like whether the central authors are discussing central topics, how does the social position of an author affect the likelihood of a topic that he will discuss etc.

1.5 Overview of The Thesis

The rest of the document in organized as follows. In chapter 2, we give review of the work done in our field in section 2.1 and then we discuss the motivation for our work in section 2.2. In chapter 3, we describe the procedure for building the social network in section 3.1. Section 3.2 discusses the method to calculate the centrality of authors in the network. Section 3.3 describes the procedure for building the cognitive structure of the network and section 3.4 presents the various feature of the bibliographic website. Then chapter 4 describes the mathematical techniques and results for socio-cognitive analysis of the network. Finally in chapter 5 we conclude this work and provide some indications about our future work.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we shall first give some brief review of the work that has already been done in this field. Then we will give the motivation towards our work.

2.1 Related Work

In the Repec Project [6] the database of economic papers has been presented in a well organized manner. The papers have been categorized into journal papers, working papers, book chapters etc. A user can also get the papers written by authors of a particular country. The papers have also been categorized on the basis of the topics they discuss. Whenever an author submits a paper he is asked to specify one out of a limited number of categories to which his paper belongs and that category is assigned to that particular paper. There is also a ranking system in place which ranks authors and institutions submitting the papers. They use several criteria for ranking and have a special technique for calculating the impact factors of publications [7].

A lot of research has been done to build the social structure of research communities. The most commonly used technique is the reference mining. This is a form of link based mining. If author A has referenced author B's paper in his own paper then there is a tie between these two authors. For building the cognitive structure of a network we need to get similarities between two publications. The most common approach for this is co-word analysis [8] [9]. Some set of keywords are extracted from each document and a whole set of keywords is formed. Similarity of two keywords in this set is calculated based on their co-occurrence patterns. Clusters of keywords are created based upon these co-occurrence patterns. These clusters are called themes [10] [11] [12]. Then documents and authors are assigned to these clusters. The relationship between the social and cognitive structure of a topic has been studied by several authors. Renner used LEXIMAPPE for examining the development of topics in various types of social science research fields [12] [8]. She also studied the relation between innovativeness of a topic and its social coherence [13]. Some authors suggest that new ideas are brought in by central authors and are carried upon by others from them. Peter Mutschke and Anabel Quan Haase [4] specifically built up a relationship between the social position of the authors and the topics that they discuss using some mathematical measures. They established that a direct relationship exists between centrality of a topic and the centrality of the authors that discuss this topic i.e, central authors discuss central topics.

2.2 Motivation of Our Work

The Repec website has presented the research papers in a well organized manner. But they are missing the temporal feature. We are taking this feature into consideration. We will implement most of the features shown on the Repec website. Apart from that we will show the important topics during different time periods ranging for over 100 years. We will also show the rankings of various authors during these time periods. The Repec website only shows these rankings for a period of 2 years. Also for a particular topic we will graphically show the rise and fall of the interest of authors for that particular topic.

The Repec project asks an author to mention the category to which his paper belongs when he submits that. Then they use that to cluster the authors into various categories. Instead of that we will use the co-word analysis technique to build clusters of authors. As far as finding the relationship between social and cognitive structure is concerned we will use the techniques implemented by Peter Mutschke and Anabel Quan Haase [4]. But the dataset that they used for producing these results was pretty small. Instead of that we are using a very large dataset here and we will try to verify the results that they produced on our dataset.

CHAPTER3

THE BIBLIOGRAPHIC TOOL

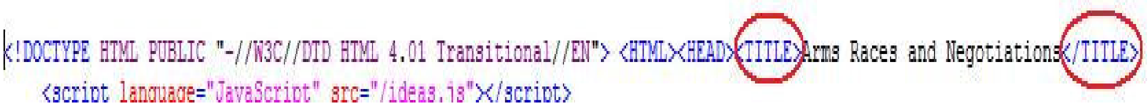
In this chapter we will discuss the procedure used in building up the tool and its features. We will also show the results obtained on out dataset.

3.1 Building the Social Network

Our first task was to build up a social network on the dataset obtained. We downloaded the html pages containing information about the papers from the Repec Website. We downloaded only those pages for which the actual paper was downloadable free and can be converted to a text file so that we can extract keywords during building up the cognitive structure of the network. The whole procedure of building the social network can be divided into three phases.

3.1.1 Parsing Phase

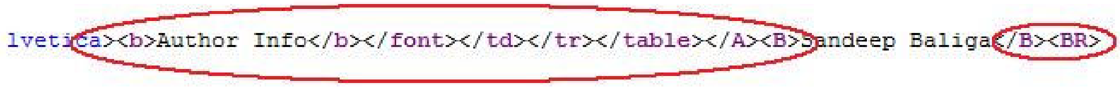
We needed to extract title, author, year of publication, references and citations from these html pages. We do this by parsing the html file using php script. For extracting the title we used the <TITLE> tags as shown in the figure below.



```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"> <HTML><HEAD><TITLE>Arms Races and Negotiations</TITLE>
   <script language="JavaScript" src="/ideas.js"></script>
```

Fig 1. Extracting the title

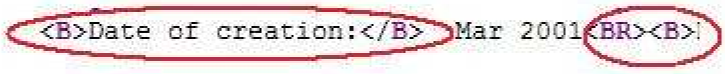
We extract the Author name from the line containing "Author Info" as shown in the figure below.



```
lvetica><b>Author Info</b></font></td></tr></table></A><B>Sandeep Baliga</B><BR>
```

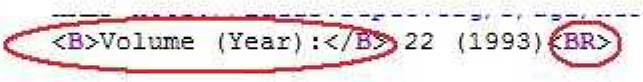
Fig 2. Extracting the Author Name

The year of Publication is present either in the line containing “Date of creation” or the line containing Volume (year). We can extract that from any one of these lines as shown in the figures below.



```
<B>Date of creation:</B> Mar 2001<BR><B>
```

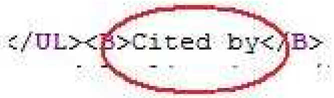
Fig 3. Date of Creation



```
<B>Volume (Year):</B> 22 (1993)<BR>
```

Fig 4. Volume (year)

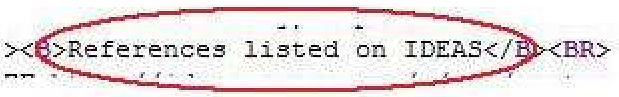
An html page can have references to which the paper references or citations to the paper or both. If the paper has citations then we get to know of their beginning by the line containing “Cited by” text as shown in the figure below.



```
</UL><B>Cited by</B>
```

Fig 5. Extracting Citations

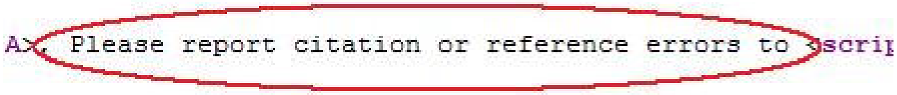
If the paper has references then we get to know of their beginning by the line containing “References listed on IDEAS” text as shown in the figure below.



```
><B>References listed on IDEAS</B><BR>
```

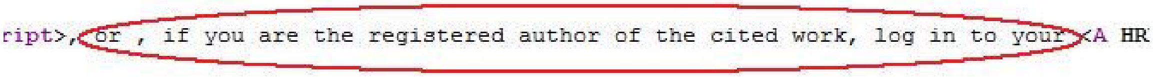
Fig 6. Extracting References

We use the following two more parts of text to confirm if whether we are extracting references or citations or not.



```
A>, Please report citation or reference errors to <scrip
```

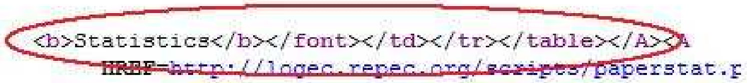
Fig 7. Extracting References and Citations 1



```
ript>, or , if you are the registered author of the cited work, log in to your <A HR
```

Fig 8. Extracting References and Citations 2

The line containing "Statistics" as shown in the figure below marks the end of the Citations and References.



```
<b>Statistics</b></font></td></tr></table></A><A
     HREF=http://logec.repec.org/scripts/paperstat.p
```

Fig 9. Extracting Citations and References 3

3.1.2 Improvement Phase

After the extraction of these features from the html pages we need to improve them. This is needed to avoid redundancy among the data. We try to explain this using the following examples: -

- Consider an Author Lou Kim. In one paper his name is written as "Lou Kim". In other paper his name is written as "lou kim". In another paper his name is mentioned as Kim, Lou.

- Consider two authors Lou Kim and “Sandip Baliga”. In one paper their names are written as “Lou Kim, Sandip Baliga”. In another paper their names are written as “Kim, Lou and Baliga, Sandip”.
- Consider a paper with the title “Sugar, Spice and Tea Industry in India”. In another version of the same paper the paper name is written as “Sugar; Spice and Tea Industry in India”. We need to treat both the versions of one paper as the same.
- A paper’s title may also be written in a slightly different manner in a reference to it. For example the author title may contain “ but the reference to it may contain ‘. We need to treat both of them as the same.

So, we make all the characters in author name, titles as well as references to the lower case. We also remove some common symbols like “, (, ‘, etc. from them. We also do rearrangement of author names around “,”. By doing this we obtain two unique lists of authors and papers. We use these two lists to remove the redundancies from the author names and from the papers as well as the references.

3.1.3 Link Building Phase

In this phase we build the links between the nodes from the data that we have obtained. Firstly we give numbers to both the authors as well as the papers. So, each author has a unique number and each paper has a unique number. We obtain 91,954 distinct authors in total and 1,52,145 distinct papers. We obtain the following lists from the above data: -

1. An adjacency list of 152145 papers and authors of each paper. An entry “25 45 -> 341” means that paper no. 25 was written by author numbers 45 and 341.
2. An adjacency list of the references of the papers whose html page contains the references. An entry “231 46 -> 567 -> 4561” means that paper number 231 has references to paper number 46, 567 and 4561.
3. An adjacency list of the citations of the papers whose html pages contain citations. An entry “2545 56 -> 123 -> 4567 -> 654” means that paper number 2545 was referenced by paper number 56, 123, 4567 and 654.

4. We combine the adjacency lists in 2 and 3 to build up a single reference list which has no redundancies.
5. Each paper has been references by one or more authors. We are interested in building up a network of authors and not papers. So, we use the adjacency lists in 1 and 4 to build up and adjacency list of authors. An entry “34 456 -> 9899” in this list means that author number 45 has referenced to author numbers 456 and 9899 in one or more of its papers.

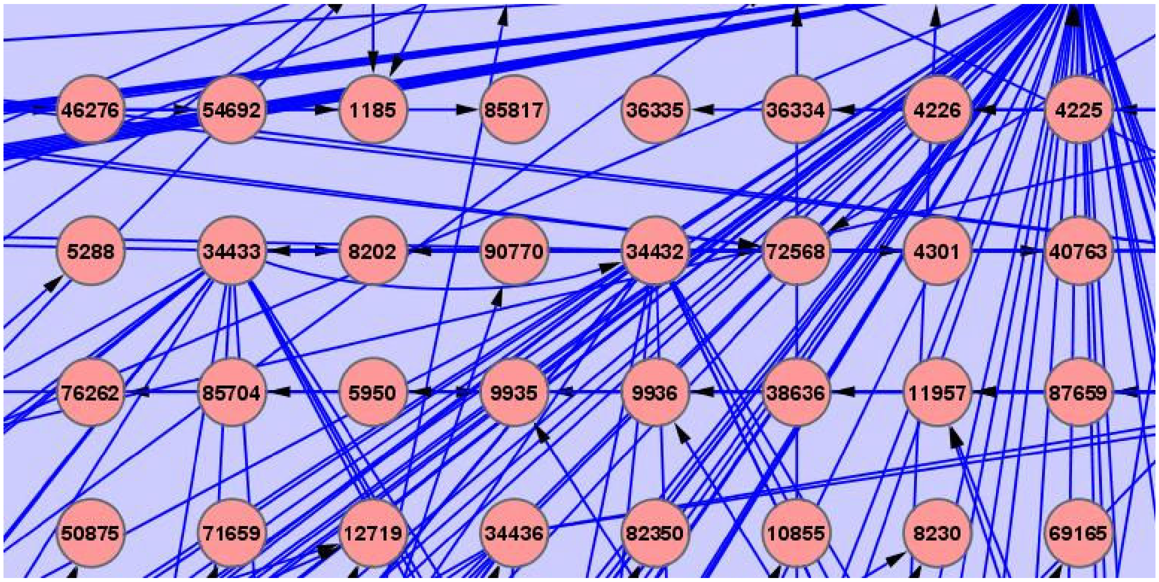


Fig 10. Small region of the network

So, the list obtained above in 5 completes the procedure of building up the social network of the authors. We have used adjacency lists in place of arrays because the network is very sparse. Since the dataset is very large so using arrays will also make the computations take a very long time to finish. The graph obtained has 91,954 nodes and 19,22,659 links between the nodes. The graph is a directed graph with the arrows indicating the direction of the reference between the papers. A small region of the graph is shown in figure above. As shown in figure below there are some dense regions in the graph and some sparse regions.

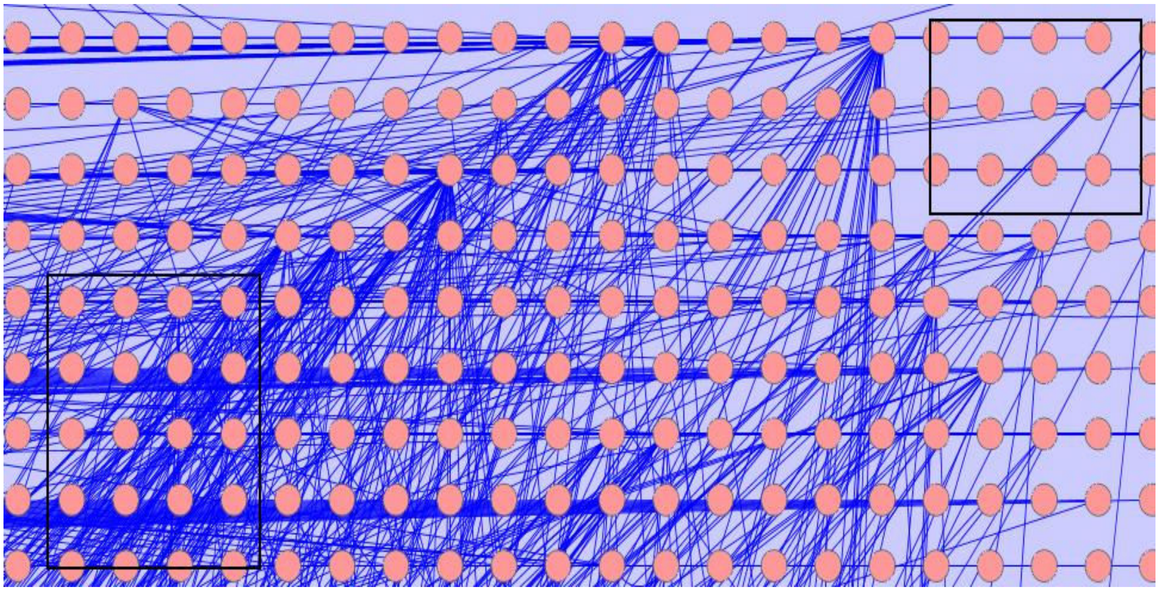


Fig 11. Dense and Sparse Regions in the network

3.2 Calculating the centrality of authors

As discussed in section 1.2 there are three common ways to measure the centrality of a node in a network. In this section we describe the results when these three methods are used to calculate the centrality values on the nodes in our network.

We use the JUNG libraries [14] for the calculation of these centrality values. JUNG stands for Java Universal Network/Graph Framework which is a software library that provides a common and extensible language for the modeling, analysis, and visualization of data that can be represented as a graph or network. JUNG provides a general, flexible and powerful API for manipulating, analyzing and visualizing graphs and networks in Java. Only a few Java methods have to be written in order to call up these libraries. It also supports a variety of representations of nodes and their edges, including directed and undirected graphs, multi modal graphs and graphs with parallel edges (multi graphs). Besides, it can load input data available in many formats. The data format that we used was in NET format [14] .

The closeness centrality of a node in a network requires one to calculate its mean distance to all the other nodes and take its inverse. But the network in our case is not fully connected i.e, there are nodes in the network which cannot be reached from each other by following edges along a path. So, we cannot calculate the closeness centrality of a node in this network.

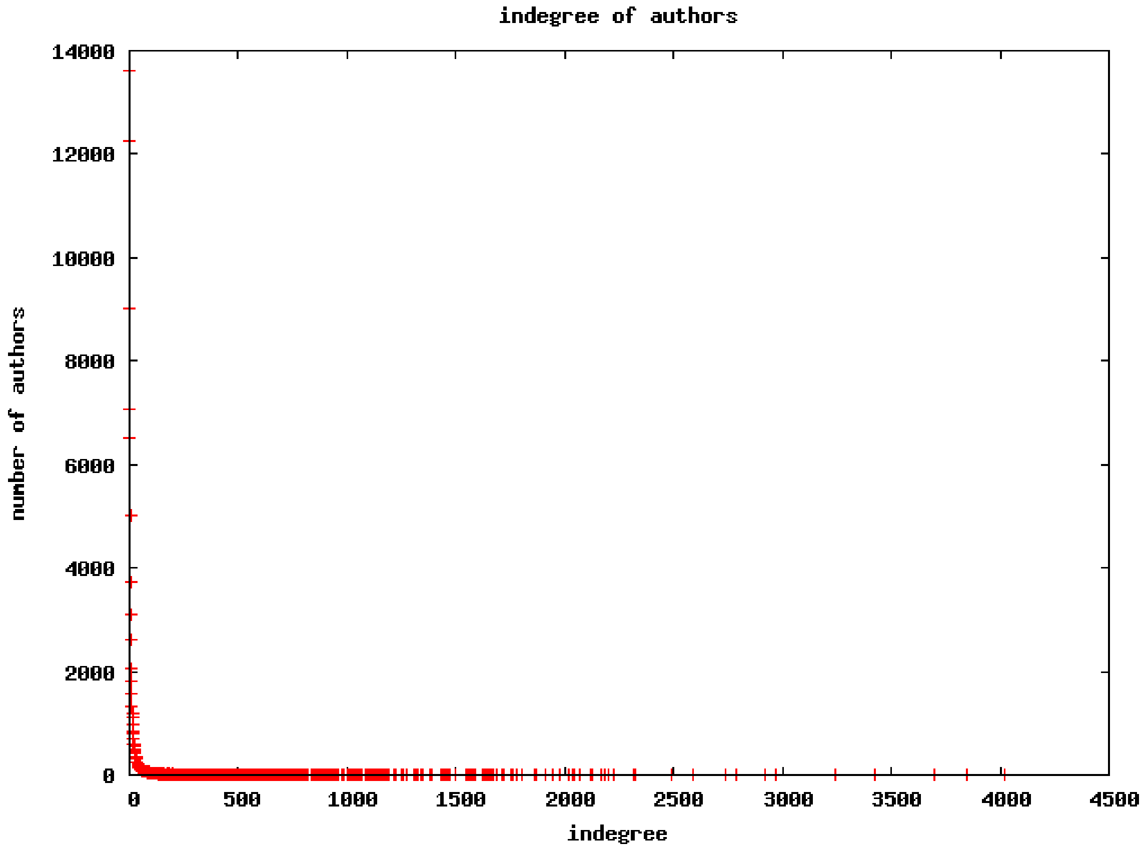


Fig 12. Range of Indegree Values

For the case of degree centrality, we only calculated the indegree values because calculating the outdegree values in this case was meaningless. If any author refers to a large number of authors in his papers then it suggests nothing valuable about the social position of that author. But if a node has many incoming edges then the author corresponding to that node is surely an important author because many other authors refer to his papers. We used JUNG libraries to calculate the indegree values of all the nodes in the network. The author Robert J. Barro had the highest indegree value followed by

Joseph E. Stiglitz. 9663 authors had indegree values greater than 35. A plot of the indegree values versus the number of authors is shown in the figure above.

Betweenness centrality of a node in the network shows the involvement of the author corresponding to that node in relation to other authors regardless of the direction of the relation. Actors with a high betweenness centrality have the potential to control communication in a network and to take the role of a coordinator in group processes. So, we calculated betweenness centrality of the nodes in our network. But only 88 nodes were found to have the betweenness centrality values greater than 0.01. This is because our network is actually very sparse.

We use the indegree values as the measure of centrality in our network. This is because in an author network with links between nodes defined by reference between the papers of the authors corresponding to nodes the most important factor is the reference itself and not the path to a referenced author.

3.3 Building the Cognitive Structure of the network

In this section we will discuss the procedure of building the cognitive structure of the network. The procedure can be divided into the following steps.

3.3.1 Keyword Extraction

We downloaded all the actual papers. We then convert them into text files. We removed the author names, title and references from these text files. We changed all the characters in the text files to lower case letters in order to prevent redundancy in the data. We also removed the stop words from the text files like “.”, “;” etc. Then we made a list of all the words in all the documents.

We used the term frequency – inverse document frequency (tf - idf) value of each of these words. The term count of a term (word) in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term t_i within the particular document d_j. Thus we have the *term frequency*, defined as follows.

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the number of occurrences of the considered term in document d_j, and the denominator is the sum of number of occurrences of all terms in document d_j [1]. For a given word we take the max of the term frequency values over all the documents. The *inverse document frequency* is a measure of the general importance of the term (obtained by dividing the number of all documents by the number of documents containing the term, and then taking the logarithm of that quotient) [1].

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

with

- $| D |$: total number of documents in the corpus
- $|\{d : t_i \in d\}|$: number of documents where the term t_i appears (that is $n_{i,j} \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to use $1 + |\{d : t_i \in d\}|$

Then tf-idf of a word is defined as the product of its tf and idf values [1]. We calculate the tf-idf values of all the words and took only the keywords having high tf-idf values. We were able to get 10,305 keywords.

3.3.2 Building the Clusters of keywords

After extracting the keywords we built clusters from the keywords. For this we needed to calculate the *equivalence* between the keywords. Equivalence of two keywords i and j is defined as

Equivalence of Keywords i and j = (Number of documents in which i and j occur together)/ ((number of documents in which i occurs)*(number of documents in which j occurs))

We calculated the equivalence values of all the keywords and created an adjacency list. An entry “sulphur -> emit 0.0019493177387914 -> soils 0.001194743130227 -> pollutants 0.0016339869281046” in the adjacency list means that keyword sulphur is related to keywords emit, soils, pollutants with equivalence values of 0.0019493177387914, 0.001194743130227 and 0.0016339869281046 respectively.

After finding the equivalence values we used the single linkage hierarchical clustering algorithm [1] to create the clusters of the keywords. Initially each word is declared as a single cluster. Then we use a metric for combining two clusters. This metric is also called the distance between the clusters. In single linkage, the distance between two clusters is computed as the distance between the two closest elements in the two clusters.

Mathematically, the linkage function — the distance $D(X,Y)$ between clusters X and Y — is described by the following expression :

$D(X,Y) = \min(d(x,y))$

where

- $d(x,y)$ is the distance between elements $x \in X$and $y \in Y$. This distance in our case is the inverse of equivalence value.
- X and Y are two sets of elements (clusters)

So, we find two clusters having the highest value of the metric and combine them into one cluster. We repeat the algorithm with the new set of clusters and keep on doing this until we reach one cluster. But in our case we do not need a single cluster. Instead we stop when the value of the metric falls below a specified value. In this way we were able to generate the clusters. The total number of clusters generated was 105. A plot of the cluster number against the size of clusters is shown below.

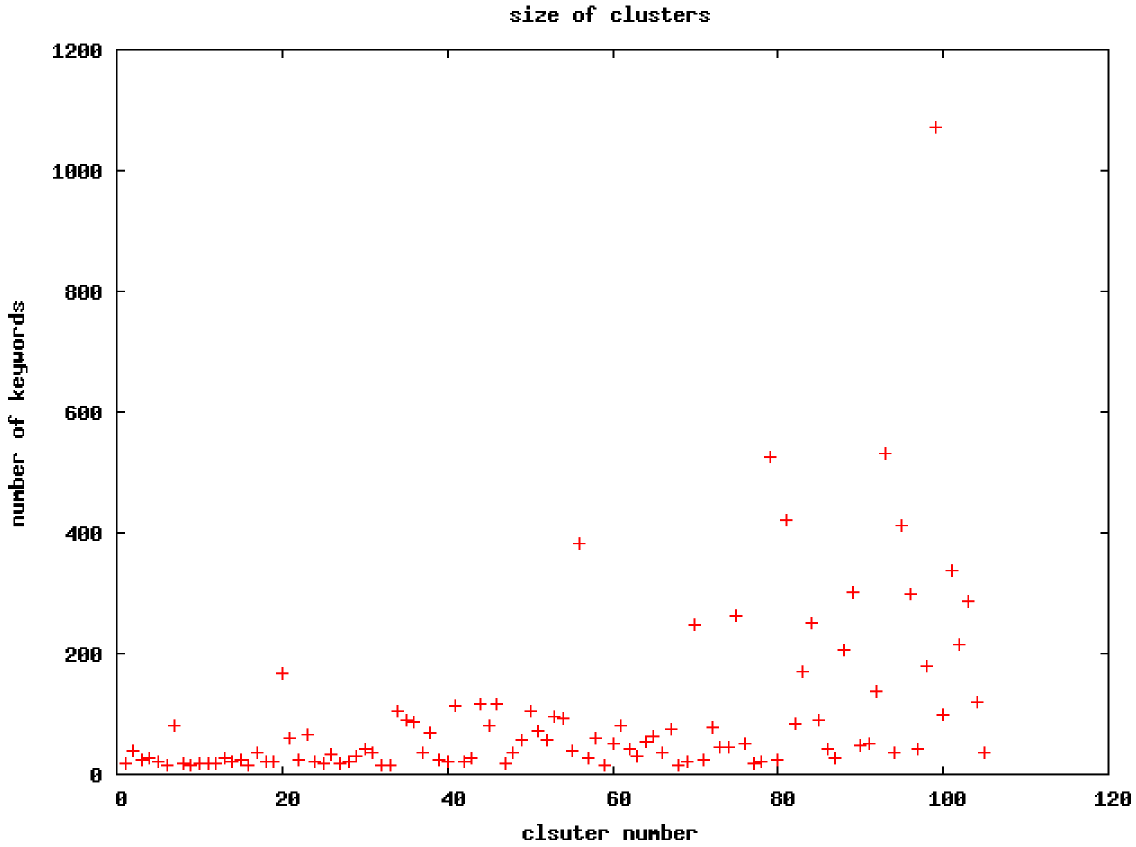


Fig 13. Size of Keyword clusters

3.3.3 Building the Clusters of authors

From the clusters we had to build the clusters of authors. The clusters of keywords are also called *themes*. The keywords in a cluster are said to be defining a particular theme. To determine the set of authors which belong to a cluster (theme), the relevance of authors, their documents respectively, for a particular theme are considered. The relevance of a document for a given theme is determined due to the degree of matching the terms which define the cluster [4]. Let D be the set of documents, P be the set of actors, K be the set of keywords, C be the set of clusters, $K_d \in K$ be the set of terms describing a document d $\in$ D and $K_c \in K$ be the set of terms defining a cluster c $\in$ C, the relevance of d for c is defined as $R_{dc}=|K_d \cap K_c|/|K_c|$, whereby R_{dc} = 1.0 indicates that c is completely matched by d. The relevance of an actor p $\in$ P for a cluster c $\in$ C is then defined as $R_{pc} = \sum R_{dc}$, for all d $\in$ D_p, D_p $\in$ D, $R_{dc} > t$, where D_p is the set of documents of p, and t the threshold which defines the minimum relevance index d should have to be considered as a typical document for c. For every author we take the maximum of its relevance values and assign it to the cluster for which this relevance value is the greatest. There are some authors which do not belong to any of the categories because all of their papers had R_{dc} values less than t for all the clusters. Such authors are called outliers and we place them in cluster 0. A plot of the cluster number against the size of each cluster is shown below.

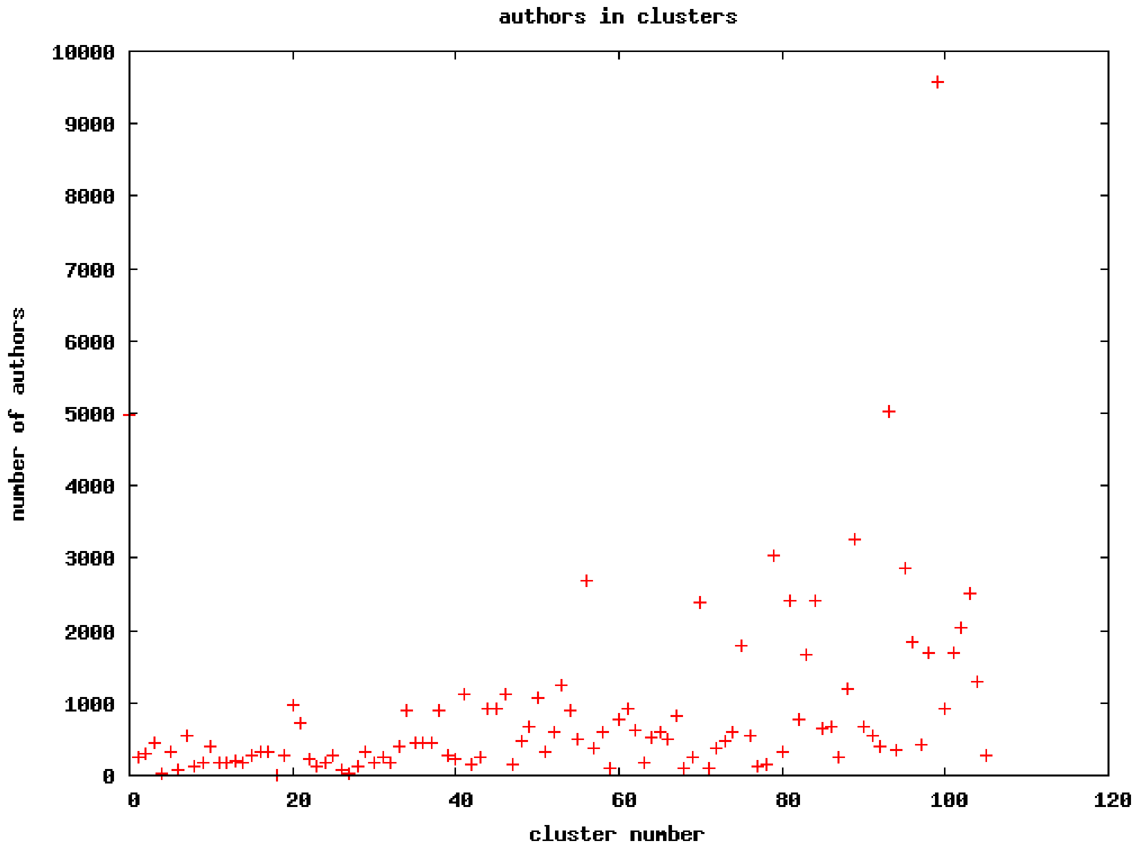


Fig 14. Size of Author Clusters

3.3.4. Calculating the Centrality Values of the Clusters

For a given cluster the centrality is defined as the number of ties it has to another clusters divided by the total number of nodes in the cluster [4]. Let C be the set of clusters and P be the set of authors. $P_c \in$ P is the set of authors belonging to the cluster c $\in$ C. Let T_{ic} be the number of ties between an author $P_{ic} \in P_c$ and the authors belonging to the set P-P_c i.e, all the other authors except those belonging to its own cluster. So, the centrality of the cluster c is defined as

$C_c = \sum T_{ic}/N_c$ for each author $P_{ic} \in P_c$

where N_c is the number of authors in cluster c.

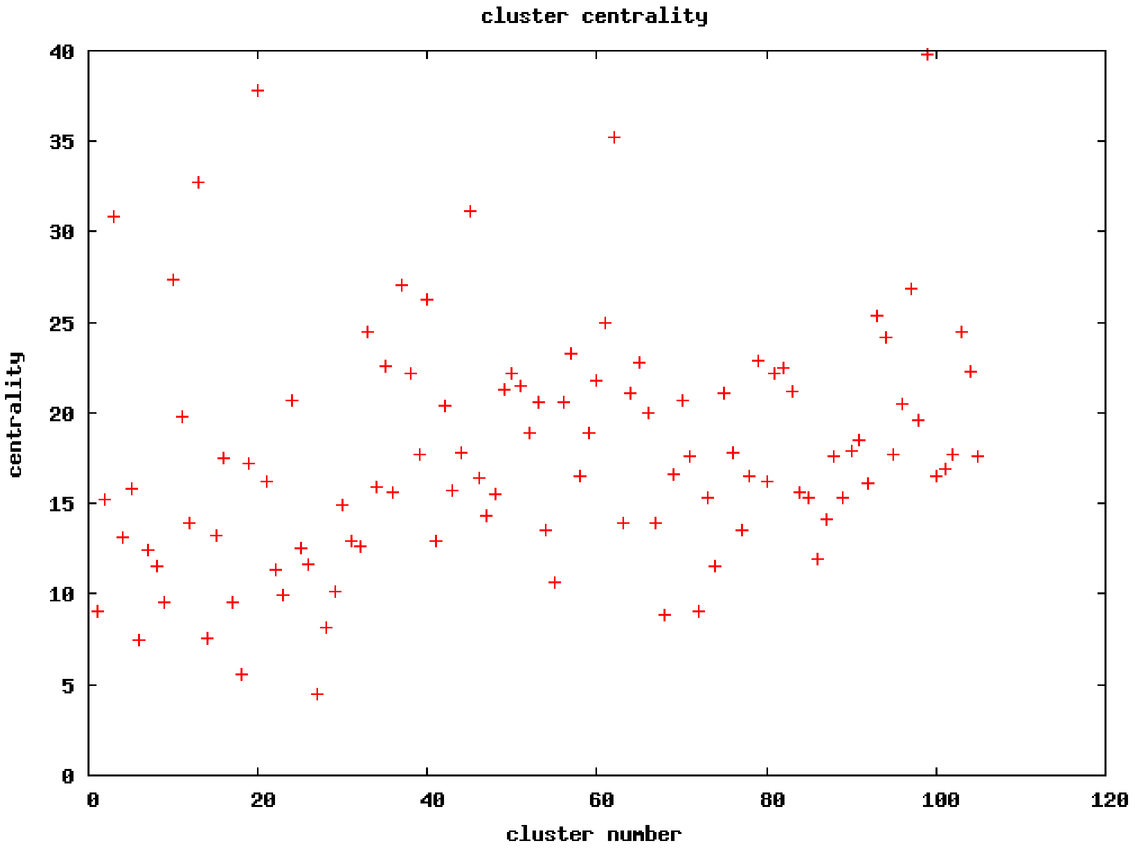


Fig 15. Centrality of Author Clusters

Using this approach we calculate the centrality values of all the clusters in the network. A plot of the cluster number versus the centrality of each cluster is shown in the figure above.

3.4 The Website showing the results

We stored all the data from the previous sections into a database and used that to build up a website having various web pages. In this section we discuss the results shown on these pages.

3.4.1 Home Page

As shown in the screenshot below the left side of the webpage contains the links to 106 categories into which we divide the papers. Each of these categories is also called a theme which is defined by certain keywords. All the other pages contain these links to all the category pages on the left side. On clicking on any of these links the user reaches the home page of that category.

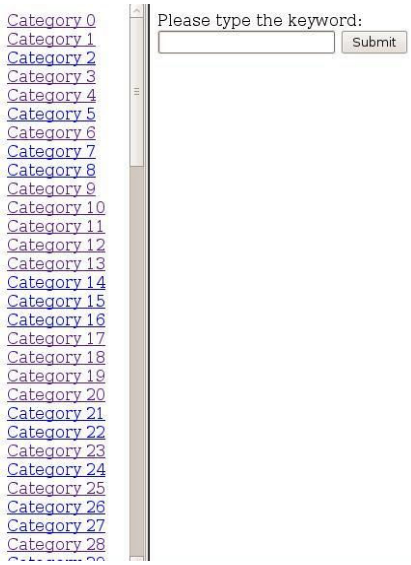
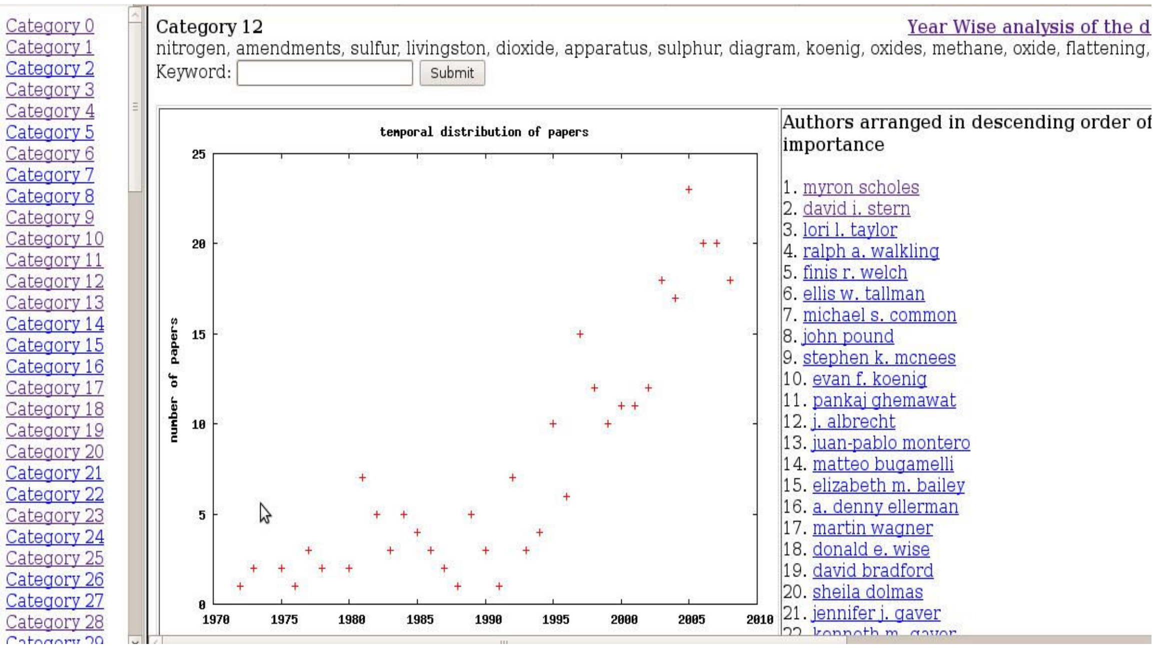


Fig 16. Home Page

The right side of the webpage contains a search box. The user searches with a keyword and a link appears to the theme that is defined by that keyword. If the keyword does not relate to any of the themes then a link to category 0 appears. As discussed in section 3.3.3 category 0 contains the authors that do not belong to any of the themes.

3.4.2 Category Page

A category page shows the information about a single category out of 106. A screenshot of the category page is shown below.



Category 12

nitrogen, amendments, sulfur, livingston, dioxide, apparatus, sulphur, diagram, koenig, oxides, methane, oxide, flattening,

Keyword:

Authors arranged in descending order of importance

1. myron scholes
2. david i. stern
3. lori l. taylor
4. ralph a. walkling
5. finis r. welch
6. ellis w. tallman
7. michael s. common
8. john pound
9. stephen k. mcnees
10. evan f. koenig
11. pankaj ghemawat
12. j. albrecht
13. juan-pablo montero
14. matteo bugamelli
15. elizabeth m. bailey
16. a. denny ellerman
17. martin wagner
18. donald e. wise
19. david bradford
20. sheila dolmas
21. jennifer j. gaver
22. kenneth m. gaver

Fig 17. Category Page

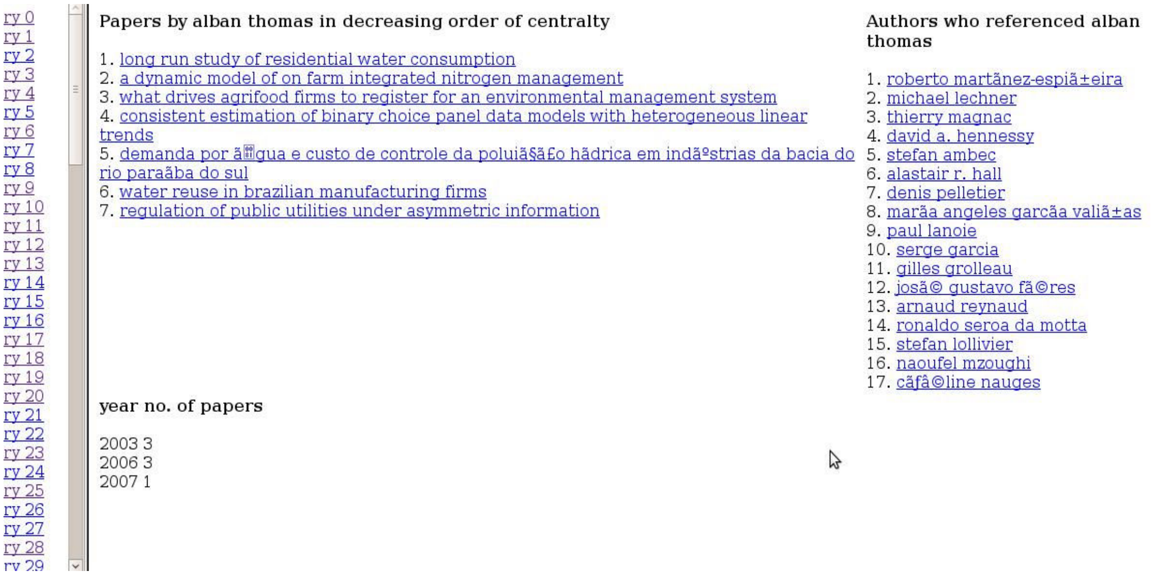
The right side of the category page contains the following: -

- Search box for finding the theme related to the searched keyword.
- A subset of the keywords defining the category that we are looking at.
- A link to the year wise analysis of the dataset.
- A plot with the title "temporal distribution of papers". This plot describes the rise and fall of the theme of the current category over the years. This plot shows the number of papers of the theme published in each year from 1970 to 2009.

- A list of the authors belonging to this category arranged in decreasing order of their centrality values. This list is actually a list of links. A user reaches an author page by clicking on any of these links.

3.4.3. Author Page

An author page shows the information about a single author out of 91954. A screenshot of the author page is shown below.



Papers by alban thomas in decreasing order of centralty

1. long run study of residential water consumption
2. a dynamic model of on farm integrated nitrogen management
3. what drives agrifood firms to register for an environmental management system
4. consistent estimation of binary choice panel data models with heterogeneous linear trends
5. demanda por ã□gua e custo de controle da poluiã§ã£o hãdrica em indãºstrias da bacia do rio paraãba do sul
6. water reuse in brazilian manufacturing firms
7. regulation of public utilities under asymmetric information

year no. of papers

2003 3
2006 3
2007 1

Authors who referenced alban thomas

1. roberto martãnez-espiã±eira
2. michael lechner
3. thierry magnac
4. david a. hennessy
5. stefan ambec
6. alastair r. hall
7. denis pelletier
8. marãa angeles garcãa valiã±as
9. paul lanoie
10. serge garcia
11. gilles grolleau
12. josã© gustavo fã©res
13. arnaud reynaud
14. ronaldo seroa da motta
15. stefan lollivier
16. naoufel mzoughi
17. cãƒâ©line nauges

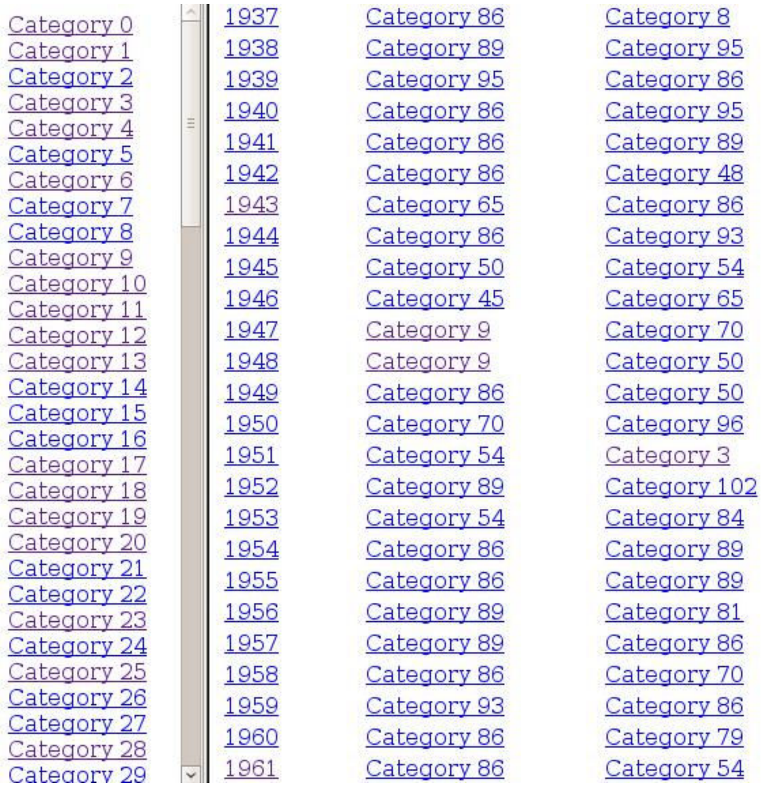
Fig 18. Author Page

The right side of the author page contains the following: -

- The papers of the author in decreasing order of centrality. The centrality of a paper is the indegree of the paper i.e, the number of papers who have references to that paper.
- A list of authors who referenced this author in one or more of their papers.
- A list containing the number of papers the author has written in each year. Nil values are eliminated.

3.4.4 Year wise analysis of the dataset

The right side of this page contains three types of links. For each year from 1879 to 2009, the top two themes discussed by the authors in that year are displayed. Also clicking on the link to that year will lead the user to the year page. A screenshot of the page is shown below.



Category 0
Category 1
Category 2
Category 3
Category 4
Category 5
Category 6
Category 7
Category 8
Category 9
Category 10
Category 11
Category 12
Category 13
Category 14
Category 15
Category 16
Category 17
Category 18
Category 19
Category 20
Category 21
Category 22
Category 23
Category 24
Category 25
Category 26
Category 27
Category 28
Category 29

1937	Category 86	Category 8
1938	Category 89	Category 95
1939	Category 95	Category 86
1940	Category 86	Category 95
1941	Category 86	Category 89
1942	Category 86	Category 48
1943	Category 65	Category 86
1944	Category 86	Category 93
1945	Category 50	Category 54
1946	Category 45	Category 65
1947	Category 9	Category 70
1948	Category 9	Category 50
1949	Category 86	Category 50
1950	Category 70	Category 96
1951	Category 54	Category 3
1952	Category 89	Category 102
1953	Category 54	Category 84
1954	Category 86	Category 89
1955	Category 86	Category 89
1956	Category 89	Category 81
1957	Category 89	Category 86
1958	Category 86	Category 70
1959	Category 93	Category 86
1960	Category 86	Category 79
1961	Category 86	Category 54

Fig 19. Year wise Analysis of the dataset

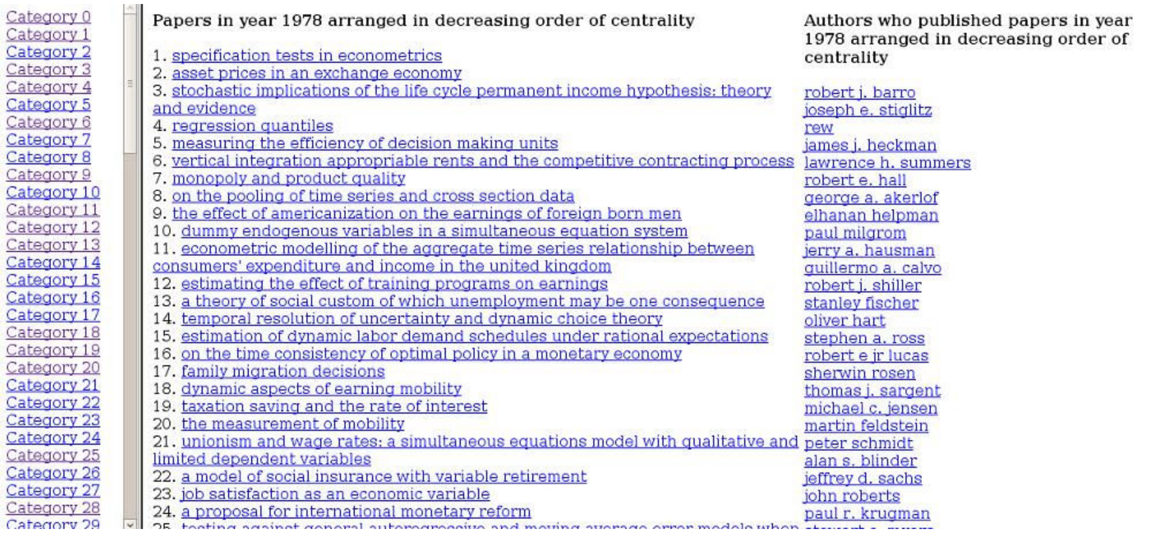
3.4.5 Year Page

A year page shows the information about a single year from 1879 to 2009. The right side of the year page contains the following: -

- A list of all the papers published during that year organized in decreasing order of centrality. The centrality of a paper is defined in the same way as in section 3.4.3.

- A list of all the authors who have published papers during that year in decreasing order of their centrality values.

A screenshot of the year page is shown below.



Category 0
Category 1
Category 2
Category 3
Category 4
Category 5
Category 6
Category 7
Category 8
Category 9
Category 10
Category 11
Category 12
Category 13
Category 14
Category 15
Category 16
Category 17
Category 18
Category 19
Category 20
Category 21
Category 22
Category 23
Category 24
Category 25
Category 26
Category 27
Category 28
Category 29

Papers in year 1978 arranged in decreasing order of centrality

1. specification tests in econometrics
2. asset prices in an exchange economy
3. stochastic implications of the life cycle permanent income hypothesis: theory and evidence
4. regression quantiles
5. measuring the efficiency of decision making units
6. vertical integration appropriable rents and the competitive contracting process
7. monopoly and product quality
8. on the pooling of time series and cross section data
9. the effect of americanization on the earnings of foreign born men
10. dummy endogenous variables in a simultaneous equation system
11. econometric modelling of the aggregate time series relationship between consumers' expenditure and income in the united kingdom
12. estimating the effect of training programs on earnings
13. a theory of social custom of which unemployment may be one consequence
14. temporal resolution of uncertainty and dynamic choice theory
15. estimation of dynamic labor demand schedules under rational expectations
16. on the time consistency of optimal policy in a monetary economy
17. family migration decisions
18. dynamic aspects of earning mobility
19. taxation saving and the rate of interest
20. the measurement of mobility
21. unionism and wage rates: a simultaneous equations model with qualitative and limited dependent variables
22. a model of social insurance with variable retirement
23. job satisfaction as an economic variable
24. a proposal for international monetary reform
25. testing against general autoregressive and moving average error models when

Authors who published papers in year 1978 arranged in decreasing order of centrality

robert j. barro
joseph e. stiglitz
rew
james j. heckman
lawrence h. summers
robert e. hall
george a. akerlof
elhanan helpman
paul milgrom
jerry a. hausman
guillermo a. calvo
robert j. shiller
stanley fischer
oliver hart
stephen a. ross
robert e jr lucas
sherwin rosen
thomas j. sargent
michael c. jensen
martin feldstein
peter schmidt
alan s. blinder
jeffrey d. sachs
john roberts
paul r. krugman

Fig 20. Year Page

CHAPTER4

THE MATHEMATICAL TECHNIQUES FOR SOCIO COGNITIVE ANALYSIS

In this chapter we will discuss three different analysis techniques for the socio-cognitive analysis of our dataset.

4.1 Centrality Technique

The *Pearson product-moment correlation coefficient* (denoted by *r*) is a common measure of the correlation (linear dependence) between two variables *X* and *Y* [1]. It is very widely used in the sciences as a measure of the strength of linear dependence between two variables, giving a value between +1 and -1 inclusive. It is defined as the sum of the products of the standard scores of the two measures divided by the degrees of freedom. Let there be n data points P and (X_i, Y_i) is such a point $P_i \in$ P. Based on a sample of paired data (X_i, Y_i), the sample Pearson correlation coefficient can be calculated as

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

where

$$\frac{X_i - \bar{X}}{s_X}, \bar{X}, \text{ and } s_X$$

are the standard score, sample mean, and sample standard deviation (calculated using $n - 1$ in the denominator) [15].

The result obtained is equivalent to dividing the sample covariance between the two variables by the product of their sample standard deviations:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

In our case we take the Pearson's correlation of the cluster centrality and the actor centrality [4]. So, in our case the variables can be defined as follows

1. n is 105 i.e., the number of clusters.
2. A point P_i = (X_i, Y_i) where
    - X_i is the cluster centrality value of cluster i as calculated in section 3.3.4.
    - Y_i is the average of the centrality values of the authors belonging to the cluster i.

We then calculate the Pearson's Correlation Coefficient on these values. The result comes out to be 0.73249125520754. Such a high value of Pearson's correlation coefficient shows that cluster centrality is strongly related to the author centrality. So, the important authors in the network tend to discuss the central themes. This result is in coherence with the result produced by Peter Mutschke and Anabel Quan Haase [4].

4.2 Occurrence Technique

Firstly we divide the set of authors S into three categories A, B and C according to their centrality values. The authors in category A have centrality values less than 10. The authors in category B have centrality values greater than or equal to 10 but less than 100. The authors in category C have centrality values greater than 100. In this technique we take the Pearson's coefficient of the occurrence of authors belonging to a particular category (A,B or C) in a cluster against the centrality values of the clusters [4]. So, in this case we calculate three Pearson's coefficient values. We consider the first one below.

The variables in this case are

1. n is 105 i.e., the number of clusters.
2. A point P_{ia} = (X_i, Y_{ia}) where
   - X_i is the cluster centrality value of cluster i as calculated in section 3.3.4.
   - Y_{ia} is the total number of authors of category A belonging to the cluster i.

We calculate the Pearson's correlation coefficient values using these variables and let the result be called PC_a. Similarly we calculate the Pearson's correlation coefficient values for the set of authors B and C and let these be called PC_b and PC_c respectively. The results that we get are as follows

I. PC_a = 0.41515484360162
II. PC_b = 0.44928892963813
III. PC_c = 0.48842337213292

These values are also positive. This proves that the authors in the network irrespective of their centrality values tend to engage in the themes having high centrality values. Moreover as the centrality values of authors increase their tendency to engage in central clusters increases. This result is also in coherence with the result produced by Peter Mutschke and Anabel Quan Haase [4].

4.3 Relevance Technique

In this case too we first divide the set of authors S into three categories A, B and C which are the same as in section 4.2. In this case we calculate the relevance values of the authors belonging to certain category (A, B or C) in a cluster against the centrality values of the clusters [4]. So, in this case too we calculate three Pearson's correlation coefficient values. The variables in this case are

1. n is 105 i.e., the number of clusters.
2. A point P_{ia} = (X_i, Y_{ia}) where
   - X_i is the cluster centrality value of cluster i as calculated in section 3.3.4.

- Y_{ia} is the mean of the relevance values of authors of category A belonging to the cluster i. The relevance value of the author to a cluster is calculated as in section 3.3.3.

We calculate the Pearson's correlation coefficient values using these variables and let the result be called PC_a. Similarly we calculate the Pearson's correlation coefficient values for the set of authors B and C and let these be called PC_b and PC_c respectively. The results that we get are as follows

I. PC_a = 0.43141822909613
II. PC_b = 0.4680054055587
III. PC_c = 0.49153809624806

These values are also positive. This proves that the relevance of authors towards the themes having high centrality values is positively correlated irrespective of the centrality values of the authors of the network. This result is also in coherence with the result produced by Peter Mutschke and Anabel Quan Haase [4].

CHAPTER5

CONCLUSION AND FUTURE WORKS

In this thesis we described the procedure for building up a bibliographical tool for presenting a dataset of research papers in a user-friendly manner. Any dataset of research papers can be presented using this tool. The only changes that we will have to make are in the parsing phase. We presented the results of building a social and a cognitive network on the dataset downloaded from the Repec website. We also presented some mathematical measure for relating the social position of authors with the centrality of the cognitive network and the results were in coherence with the results produced by other researchers. In the future we will like to analyze the other socio-cognitive feature like origin of innovativeness and flow of information in the network.

References

1. http://www.wikipedia.org/

2. L.C. Freeman “Centrality in social networks: Conceptual clarification”, *Social Networks*, Volume 1, 1979, pp. 215-239.

3. Kaiser, C.; Tiwana, B.S.; Bodendorf, F., "Bridging the Gap between Qualitative and Quantitative Analysis of Opinion Forums," *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*, vol.1, no., pp.120-126, 9-12 Dec. 2008

4. P. Mutschke, A. Q. Haase, "Collaboration and cognitive structures in social science research fields. Towards socio-cognitive analysis in information systems", *Scientometrics*, Vol. 52, No. 3 (2001) 487–502

5. M. Callon, J. P. Courtial, F. Laville, “Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry”, *Scientometrics*, 22 (1991) 155–205.

6. http://ideas.repec.org/

7. C. Zimmermann, 2007. "Academic Rankings with RePEc," *Working papers 2007-36, University of Connecticut, Department of Economics*, revised Mar 2009.

8. M. Callon, J. P. Courtial, W. A. Turner, S. Bauin, “From translations to problematic networks: An introduction to co-word analysis”, *Social Science Information*, 22 (1983) 191–235.

9. M. Hesse, “Revolutions and Reconstructions in the Philosophy of Science”, *Harvester Press, London*, 1980.

10. P. Mutschke, I. Renner, “Akteure und Themen im Gewaltdiskurs: Eine Strukturanalyse der Forschungslandschaft”, *In: E. MOCHMANN, U. GERHARDT (Eds), Gewalt in Deutschland: Soziale Befunde und Deutungslinien, Oldenburg Verlag*, 1995, pp. 147–192.

11. L. Grivel, P. Mutschke, X. Polanco, “Thematic mapping on bibliographic databases by cluster analysis: A description of the SDOC environment with SOLIS”, *Knowledge Organisation*, 22 (1995) 70–77.

12. K. M. Van Meter; W. A. Turner, “Cognitive mapping: The German FORIS database and Sociological Abstracts’ AIDS research”, *In: H. BEST et al. (Eds), Informations- und*

Wissensverarbeitung in den Sozialwissenschaften. Westdeutscher Verlag, Opladen, 1994, pp. 257–274.

13. I. Renner, "Soziale Kohärenz und Innovatität: Struktureffekte zur Akzpetanz neuer Themen in sozialwissenschaftlichen Forschungsfeldern", *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 49 (1997) 74–97.

14. http://jung.sourceforge.net/

15. Moore, David (August 2006). "4". "Basic Practice of Statistics (4 ed.)". *WH Freeman Company*. pp. 90–114. ISBN 0-7167-7463-1.

List of Figures