Baba Ghulam Shah Badshah University
Rajouri (J&K)

(SYNOPSIS FOR REGISTRATION FOR DEGREE OF DOCTOR OF PHILOSPHY IN COMPUTER SCIENCE)

Proposed Topic	:	**"Social Network Extraction Using Web Mining Techniques"**
Name of the Candidate	:	Mr. Tasleem Arif
Name of the Supervisor	:	**Supervisor:** Prof. M. Asger Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri. **Co-Supervior:** Dr. Rashid Ali Department of Computer Engineering, Aligarh Muslim University, Aligarh.
School/Department	:	Computer Sciences, School of Mathematical Sciences & Engineering.
Date of Submission	:	30 -11 -2011

Signature of the Supervisor

Signature of the Candidate

Signature of the Co-Supervisor
(Dr. Rashid Ali)

1. Background

Introduction:

Exponential growth of public information present on the Web in a set of interlinked heterogeneous sources is a consequence of widespread internet expansion and usage [1]. Search engines are the most widely used tools for searching information from the Web, but the general approaches to analyze the information so extracted cannot integrate different sources. The advent of Web 2.0 has added another dimension to the way the data and information is being populated and shared. E-mail had been a major communication platform since long but Web 2.0 has provided other platforms like instant messaging and social networking websites (such as Blogs, wikis, web albums), etc. to cater to instant communication needs. These interactions generate a lot of data about personal communications which when combined with search engine results can be used for extraction of social networks in more efficient ways.

Understanding and modeling network structures has been a focus of attention in a number of diverse fields, including physics, life sciences, computer science, statistics, and social sciences. Applications of network analysis include friendship and social networks, marketing and recommender systems, the World Wide Web, and disease models, among others.

A *social network* is a structured representation of the social *actors* (nodes) and their *interconnections* (ties) [2]. Social network is an aggregation of social *groups* (communities) that share common interests and therefore include different relationships such as positions, betweenness and closeness among individuals or groups [3]. These communities on the web are steadily emerging and the demand for forming an on demand social network is immense. They help people to find other people sharing the same interest and are available for discussion and collaborations. For example, *if a person is searching for specific information, he can look at the interests of people in his social network and get quite relevant references*.

Social networking services (SNSs) like Facebook, Orkut, LinkedIn and others have become very popular on the Web [4]. An SNS that manages and stores social networks can become a base of information infrastructure in the future [5]. The potential of Social Networking has been utilized to a good extent in the area of personal communication, evident from the growing popularity of these SNSs, but in the area of professional and academic interaction their potential has not been exploited as much. Social networking will

play an important role in future personal, organizational and commercial online interaction as well as location and organization of information and knowledge [4].

Social Networks Analysis (SNA) is an interesting research direction to analyze the structures and relationships of social networks, such as analyses of density, centrality and cliques in social network structures and has been an area of focus for the researchers for quite some time [6]. SNA is an essential and important technique to understand the social structures, social relationships and social behaviors. In SNA the main task is usually about how to extract *social network* from different communication resources (e.g. web, e-mail communication, Internet relay chats, organizational events, conferences, web usage logs, event logs, instant messenger logs, etc.).

Construction of the researcher network, for example can benefit many Web mining and social network applications [7]. For example, in this case, *if all the profiles of researchers are correctly extracted, we will have a large collection of well-structured data about real-world researchers*. The profiles so extracted can help in expert finding for *research guidance for new scholars, potential speakers* and *contributors for conferences, journals, workshops* etc. The academic network so extracted can be used in many services, such as finding an appropriate person to introduce or negotiate someone, who one should talk in order to expand his/her network efficiently [8]. The information so extracted can be used for various competitive intelligence tasks like searching for experts, gathering of information, organizations collaborations analysis, countries collaborations analysis or products impacts analysis and determining topics of interest in an academic community [1].

The extracted academic network can also be used for research *trend detection/prediction*. Trend detection can help a researcher to analyze the thrust area of research in a particular field, what other researchers are doing in that or related field. Trend prediction can help research community to have an idea of the potential research topics/areas in a particular field.

Related Work

The dramatic increase of popularity of social networks has attracted a lot of research. Factors like social interaction, knowledge exchange, knowledge discovery, ability to capture data about various types of social interactions at a very fine granularity with practically no reporting bias, and availability of data mining techniques for building descriptive and predictive models of social interactions have become key drivers for computer science research in SNA.

Discovering knowledge from these networks is a challenging and primary research issue because of their size, reachability and diversity. Several research studies have been conducted on social network analysis and two main approaches have been studied: one is influential user discovery [9-12] and the other is social network construction [5, 7, 8, 13-21]. The aim of the first approach is finding most influential users in communities by analyzing their relationships and activities and the second approach concentrates on discovering social network of users [22]. This section presents the work conducted so far in the latter approach.

Data from different sources like Web, e-mail communication logs, instant messenger logs, blogs, etc. has been used either individually or in combination for the purpose of social network extraction. In the section below classification of the various techniques used for network extraction has been presented.

a) Co-occurrence of Names on the World Wide Web:

Several studies have been undertaken to use a search engine to extract social networks from the Web [8, 13, 17, 20, 23]. Co-occurrence of names on the web, is obtained by posing a query including two names to a search engine, is commonly used as proof of relational strength. Referral Web [13] was the first attempt of this kind to develop an automated interactive tool for social network extraction and finding shortest referral chains to experts. It uses a search engine (Altavista) to extract social networks through co-occurrence of names in close proximity in any documents e.g. personal homepages, lists of co-authors in technical papers, citations of papers, and organizational charts publicly available on the WWW taken as evidence of a direct relationship. The network obtained is an egocentric network, in that it is focused on a specific person. The input to the system is name of the person (X) whose social network is to be obtained and the system extracts a list of related people (L). Jaccard coefficient [24] is used to measure the significance between X and Y, where $Y \in L$. The process is repeated for each $Y \in L$. The goal here is to find series of links i.e. referral chain from the requester node to the expert node (information hub). A path from a person to a person is obtained automatically using the system. With increasing usage of Internet and development of WWW large amount of information about our daily lives is available online, making automatic extraction of social relations more demanding than when Referral Web was developed.

Referral Web [13] has influenced many studies for automatic extraction of social networks. Tombe et al. [8] proposed a system for social network extraction of conference participants from the Web. The idea behind this study is that: at academic conferences, a participant registers a brief profile with fields like *Name*, *E-mail*, *Affiliation*, etc. well before

the conference which means that there is enough time to gather information about the participants from the Web. The relationships between any two participants are determined using the Web information gathered by posing a query to a search engine in a similar fashion of [13]. An edge exists between two nodes if the Jaccard Co-efficient between those two nodes is larger than a threshold value and the weight of that edge is set equal to the Jaccard Co-efficient. To alleviate the problem of ambiguity, it [8] labels the relationships between nodes and uses machine learning to identify them. Social network with 650 conference participants of WWW2002 has been extracted. The drawback of [8] is that it assumes that nodes in the network are predetermined and does not consider unregistered participants.

P. Mika developed *Flink* [23], a system for extraction, aggregation, and visualization of online Semantic Web community. The Web mining component of *Flink* similar to that of [13] obtains hit count from a search engine (Google) for both the persons X and Y individually as well as hit count for co-occurrence of these two names with the target being the Semantic Web community. It also performs the additional task of associating a researcher with a given topic of interest. The Web information source in this case being Web pages, e-mail messages, publication archives, and self created profiles (FOAF files). In [23], the strength of relations among individuals is calculated using the Jaccard coefficient $nX \cap Y / nX \cup Y$, where $nX \cap Y$ represents the number of hits yielded by the query X AND Y and $nX \cup Y$ represents the number of hits by the query X OR Y. The two researchers are considered to share a relation if the value is greater than a certain threshold. The term "Semantic Web OR Ontology" is added to the query for name disambiguation. Although, [23] has tried to remove the problem of ambiguity in identification of entities with similar names, the system still has certain problem because of data collection (general noise, errors in the extraction of specific cases) in this respect.

Matsuo et al. developed *POLYPHONET* [17], which also uses a search engine (Google) to measure the co-occurrence of names. In their study, several co-occurrence measures [25] have been compared, including the matching coefficient ($nX \cap Y$), mutual information, Dice coefficient, Jaccard coefficient, and overlap coefficient. The overlap coefficient $nX \cap Y / \min(nX, nY)$ performs best according to the experiments. In addition, POLYPHONET was operated at several AI conferences in Japan and a couple of international conferences to promote participant's communication. For disambiguating personal names, key phrases such as affiliations are added to queries.

The fundamental idea behind [8, 13, 17 and 23] is that *the strength of a relation between two entities can be estimated by co-occurrence of their names on the web*. The criteria to recognize a relation, such as the measure of co-occurrence and a threshold, are determined beforehand. An edge will become a part of edge set '*E'* when the relation strength determined by the co-occurrence measure is higher than the predefined threshold. Although the approach is effective for extracting a social network of researchers, studies [20] indicate that it does not perform well for various entities on the web.

Co-occurrence-based methods become ineffective when two entities co-occur universally on numerous web pages and function ineffectively when applied to inhomogeneous communities which mean that co-occurrence of names on the web is not always available to represent precisely the relational strength of two entities and for inhomogeneous entities it is difficult to precisely recognize the relation using a single criterion [20]. It [20] expands the existing social-network mining techniques using a search engine to obtain various social networks from the web. Two improvements: relation identification and threshold tuning have been made to specifically focus on complex and inhomogeneous communities respectively. Two social networks: artists of contemporary art, and famous firms in Japan are extracted.

The efficiency and accuracy of an extracted social network depends primarily on whether it has been able to address well the problems associated with profile extraction and name disambiguation. *Arnetminer* [7, 21] focuses primarily on *profile extraction* and *name disambiguation* for academic researchers. The system constructs a *semantic* based social network of academic researchers by extending the Friend-Of-A-Friend (FOAF) ontology [26] as the profile schema, proposes a unified approach based on Conditional Random Fields to extract researcher profiles from the Web using a search engine and integrating the extracted researcher profiles and the crawled publication data from the online digital libraries. A unified probabilistic framework for dealing with the name ambiguity problem has been proposed for integration. It proposes three generative probabilistic models for simultaneously modeling topical aspects of papers, authors, and publication venues. Based on the modeling results, it implements several search services such as expertise search and association search.

It [8] proposes a unified approach to profiling consisting of three steps: relevant page identification, preprocessing, and extraction. In relevant page identification, given a researcher name, list of web pages is obtained by a search engine (Google API) and then homepage/introducing page are identified using a binary classifier (SVM [27]).

Preprocessing has two steps (a) separating the text into tokens and (b) assigning possible tags to each token using Conditional Random Fields (CRFs) [28] as the tagging model. After each token is assigned with several possible tags, profiling is performed. For name disambiguation, [21] uses five types of relationships: *Co-Author*, *Citation*, *Co-PubVenue*, *Constraints*, and *τ-Co-Author*, have been used with each type of relation having an impact on F1-score and the relationship of Co-Author having the highest impact (+24.38% by F1). The draw back in this case is that *k* (actual number of researchers having same name, say *'a'*) has to be provided manually.

b) E-mail Communications:

Email is a valuable and pervasive mean of communication in the information society, is one of the primary ways that people use to communicate and access their widespread social networks, and as such it is a highly relevant area for research on communities and social networks [29]. It is the number one online activity for most users and there are few advanced email technologies that take advantage of the large amount of information present in a user's inbox [30]. Maintaining and using contacts is an essential and challenging task. Unfortunately, the task of manually maintaining contact information is tedious and error-prone and a system that extracts contact information automatically from email messages itself has limited coverage because of limited to the data present in email.

Due to rapid development of electronic communications, email data becomes a powerful information source for studying social networks because of a number of advantages: availability of large amount of data on personal communications in a standard electronic format; ubiquity of email usage; frequency, longevity, and reciprocity of email communications; type (content) of communication; temporal data; and availability on both sender and receiver side. In addition to the advantages, accessing email communications has certain issues as well. Privacy issues like compromising personal privacy and organizational confidentiality concerns are the biggest barriers for email related social research which can be alleviated by accessing only header information but ignoring information carried in the message significantly limits the potential of using email as source of information for analyzing social relationship. Although the format of email messages is relatively standard and it is easy to generate a communication links from email archives, automatic extraction of social network is not easy because of issues like: multiple identities of same person; spam and group aliases; categorization of social relations by email content; weighting ties by different indicators such as reciprocity, frequency, and longevity of discussion; and aggregation of "*To, Cc, Bcc*" [31].

Several studies [29-32] have used email communication as data source for social network extraction and tried to leverage the associated benefits and address the issues concerning its usage.

Communities of practice are the informal networks of collaboration that naturally grow, collaborate, coexist with the formal structure within organizations, serve many purposes, such as resolving the conflicting goals of the organization to which they belong, solving problems in more efficient ways, and furthering the interests of their members [29]. Any organization that provides opportunities for communication among its members is eventually threaded by communities of people who have similar goals and a shared understanding of their activities. Therefore a fast and accurate method of identifying these communities of practice is desirable. Manual inspection [33, 34] or an Internet-centric [35] approach to construct links and communities are accurate but time-consuming and labor-intensive in the context of a very large organization. Alani et. al [36] proposes a semi-automated utility that uses a simple algorithm to identify nearest neighbors to one individual within a university department and relies on previously collected relational data that may be difficult to obtain for a given organization.

Automatic construction of a network of correspondences and community detection from email data keeping privacy concerns in mind using only "to:" and "from:" fields from each email has been discussed in [29]. It [29] automatically identifies communities within an organization in two basic steps: (a) uses the headers of email logs to construct a graph where the vertices are senders or recipients of email messages and the edges denote an email communication between the nodes they connect, and (b) finds the communities embedded in the graph using the concept of betweenness centrality [37] to partition the graph obtained in first step into discrete communities of nodes. The authors claim that the method was able to identify small communities within a 400-person organization (HP Labs.) in a matter of hours, running on a standard Linux desktop PC and identifies leaders within these communities through network of correspondences.

In [29] population size is predefined and corporate directory (of HP Labs.) is used to remove name ambiguity. This is much easier as compared to an informal network where membership is not clearly defined and similar names pose ambiguity problem. In [29] only header information is used to extract the link structure ignoring information carried in the message which significantly limits the potential of using email as a research proxy for social relationship.

EmailNet [31] uses information both from the header and message to extract the link structure and addresses the concerns of privacy and confidentiality by hashing each email to make the messages unreadable to the human. It extracts mails both from personal email clients as well as organizational mail servers, uses filters to handle the issues such as, spammers, duplicate identities, etc., and employs a text clustering technology based email categorization function to categorize emails into several given types of social connections. Email usage pattern analysis functions in [31] help user investigate email interactions in detail, such as time distribution across hours and days, response thread visualization. An email oriented network visualization interface helps users explore the email social network. R SNA package is used as the social network analysis engine and its data can be directly exported to other social network analysis and visualization tools, such as UciNet and Pajek.

Extracting social networks and contact information from email and the Web and combining this information is discussed in [30]. The input to the system [30] is the set of email messages in a user's inbox and the output is an automatically-filled address book of people and their contact information, with keywords describing each person, and links between people defining the user's social network. The six modules of the system are *person name extraction, name co-reference, homepage retrieval, contact information and person name extraction, expertise keyword extraction,* and *social network analysis.*

After extracting people names from email messages, it [30] works to find each person's Web presence, and then extracts contact information from these pages using a probabilistic model (CRFs). In addition, the system uses an information-theoretic approach to extract keywords for each person that act as a descriptor of his or her expertise. It then obtains social links by extracting mentions of people from Web pages and creating a link between the owner of the page and the extracted person. The entire system is called recursively on these newly extracted people, thus building a larger network containing "friends of friends of friends". The network so obtained contains a significantly wider array of expertise and influence, and represents the contacts that the user could efficiently make by relying on current acquaintances to provide introductions, perform expert finding, and make new relevant connections.

Bird et. al [32] introduces the problem of identifying email users' aliases. It [32] construct social networks of Open Source Software (OSS) developers through the email archives of OSS projects which provide a useful trace of the communication and co-ordination activities of the participants and consider the developers related if there is evidence of email communication between them. It employs a hybrid (automated/manual)

approach to resolving aliases (name disambiguation). The automated approach executes in two steps: (a) automatically crawling messages and extracting all message headers to produce a list of <*name*, *email*> identifiers (IDs) and (b) executing a clustering algorithm that measures the similarity between every pair of IDs. IDs that are sufficiently similar are placed into the same cluster. Once clusters are formed, they are manually post-processed. Communication links between pairs of individuals are extracted from message headers, the sender, the receiver, the sent time, and the identifier of the message (if any) to which this message was a reply. Three measures, *in-degree*, *out-degree* and *betweenness* are taken as indicators of the importance of an individual in a network. Apache developer mailing list with 2544 separate IDs was used to empirically study the proposed techniques. It concludes that the most active developers play the strongest role of communicators, brokers, and gatekeepers.

c) Instant Messaging:

Instant messaging (IM) or Internet Relay Chat is a popular form of computer-based communications service that enables individuals or groups of people to collaborate and chat from anywhere in the world in real time over the Internet. The instant messaging system alerts its users whenever somebody on their private list is online. Users can then initiate a chat session with that particular individual in personal messaging environment or with group of people in a chat room environment. Relationship extraction/identification is a central problem in the analysis of such large-scale social networks in their study as social networks as there is no clear measure of relationship strength. In [38] several such measures, obtained from the status log of an IM user, have been proposed that describe the link information between any pair of members. Resig et al. show [39] that, in spite of their simplicity, status logs contain a great deal of structure and relationship identification from this data is not easy. The problem can be alleviated by obtaining acquaintances (e.g. buddies in AOL) list for each user but unfortunately, such lists are not published, so in order to obtain a collection of them one has to contact each author of each list making it an impractical solution. The solution lies in constantly tracking the status (online, busy, away, offline etc.) of each user relative to the IM service. This status data, along with the time at which a given client transitions from one state to another, is published electronically, making it possible to track the state of a population of IM users over a period of time. In [38] these status logs are used as a measure of the degree to which any two AOL IM users are related on the IMSCAN framework. The IMSCAN framework does not have content monitoring capabilities.

As pointed in [38], there are two major types of link data related to instant messaging networks; buddy lists: most useful but hardest to acquire, and 3rd party social-networking web sites: readily available and easily accessible. In [38], over 200,000 user names and their associated IM names were collected from LiveJournal out of which a group of 4878 users with a minimum in and out link degree of 15 were chosen. Two types of link-based discovery mechanisms viz. co-relation and clustering have been employed. For co-relation based link discovery two experiments were performed in [38]. *First*: for each user, in the instant messaging activity log the amount of time in seconds the user was online and the number of times the user changed his or her status to online from some other state is counted. *Second*: the degree to which each pair of users is, according to the IM status log, online at the same time is measured.

[38] concludes that further efforts are required to accurately model the relationship of two users being linked if they are online at the same time and clustering techniques for recovery of information about social communities in IM networks.

In [40], an IRC bot called PieSpy [41] is used to monitor channels and infer the social network structure. Measures like, direct addressing of users, temporal proximity, temporal density, and private message monitoring have been used to infer relationship strength. After inferring social relations the authors in [40] have used modified spring embedder force model based on those of Fruchterman and Reingold [42] for connected network components and m-limited force model [40] for disconnected components networks for drawing social network.

d) Hybrid Approaches:

The Web considered as the biggest database in the world has been used in various studies to extract social networks. Most of the researches mentioned above are focusing on a single source for extraction of social networks and the issue of social network extraction from different sources on the Web has not been discussed well in related literatures [18]. Ting et. al. [43] proposed a system to extract social networks from instant messages and e-mails.

The system proposed by Ting et. al. [43] has two major components offline data collection and online data processing. Related communication data from e-mail and instant messenger is collected, the data so extracted is filtered by the data extraction engine and relevant data is stored in the database. Data collected, processed and stored by the offline data collection module is used in online processing module for social network construction and visualization.

The relationship (communication frequency) is the most important element to form a social network. In [61] the strength of a relationship (Ri) from a specific node to a node 'i' is calculated using Ri={W1*Ei+W2*Mi+W3*Bi}, where W1, W2 & W3 are weights assigned to e-mail, messaging and blogging relationships respectively.

Study Area

The primary study area of this proposed work is "Social Network Extraction" with the focus being profile extraction and name disambiguation. Since the data for the proposed work has to be obtained from Web only, Web mining techniques for Social Network extraction will be studied as well.

In this proposed work, relationship information between various entities obtained from multiple data sources have to be extracted which may lead to various ambiguities. Soft Computing techniques can be used to solve ambiguity problems. We intend to develop Soft Computing techniques based person profile extraction and name disambiguation algorithms. We also intend to perform a comparative study to evaluate the effectiveness of these algorithms in social network extraction, particularly in academic social network extraction.

Justification of the proposed work

The majority of academic researchers present their research work on the Web. This trace can be used to derive useful information about their past & present activities and forecast their future intentions. It also provides an insight into what is going on presently in a particular research area and can help answering questions like:

Which are the experts in a specific area?

Which are the most influence groups?

How the organizations are structured?

Which are the most influential countries?

Which are the goal networks?

Which particular communities are working on thrust area?

Thus, extraction/construction of social network of academic researchers can be of important value for scientific community.

For an extracted network to be useful all the possible relations have to be analyzed. This relationship information can be obtained from various data sources available on the Web (Web links, e-mail archives, citations, Blogposts, instant message logs). From each of these data sources a single-relational network can be extracted and for a network to be multi-relational, relationship information from various data sources have to be considered. The techniques available so far use one or two of these data sources. Thus there is a need to

develop a system that will be multi-relational with the relationship information obtained from various data sources.

To the best of our knowledge, the spectral and temporal considerations for extraction of social networks have not been made so far, in any of the related studies. Social networks based on spectral and temporal considerations can help answering some of the questions raised above.

Very little work has been done in the area of Soft Computing techniques based social network extraction. In this work, we intend to study the existing baseline models, analyze them, and propose a new model based on Web mining techniques and soft computing techniques that shall be an improvement over the existing ones. Thus, we hope to improve the efficiency and usefulness of the extracted social network.

2. Aims & Objectives

The researcher intends to undertake the work with the following aims and objectives:-

1. Developing Soft Computing techniques based profile extraction algorithm(s).
2. Developing Soft Computing techniques based name disambiguation algorithm(s).
3. Developing a Social Network Extraction system that will be using multiple data sources, multi-relational in nature, based on spectral and temporal considerations where possible.

3. Material proposed to be used for the investigation

There are two primary requirements for the proposed study, *datasets* and *simulation packages*. The datasets for the proposed work are mostly available online and are free for research purposes. The intended work will use the data available from the below mentioned sources.

1. Organization/Institution Websites.
2. Digital Citation Libraries.
3. Digital Theses Libraries (Australian Digital Theses Program).
4. Web Links.
5. Blog Posts.
6. E-mail Communication Logs.

Network data for online communities are based on materials that are intentionally public and therefore considered a form of publication or mass communication.

For simulation purposes packages like MATLAB will be used and, where necessary required software/simulation package may be downloaded from freely available sources from the Internet for the study at hand.

4. Methodology, Action Plan and Time Line proposed to be followed

Methodology:

Figure 1 outlines the key processes in the extraction and mining of social networks.

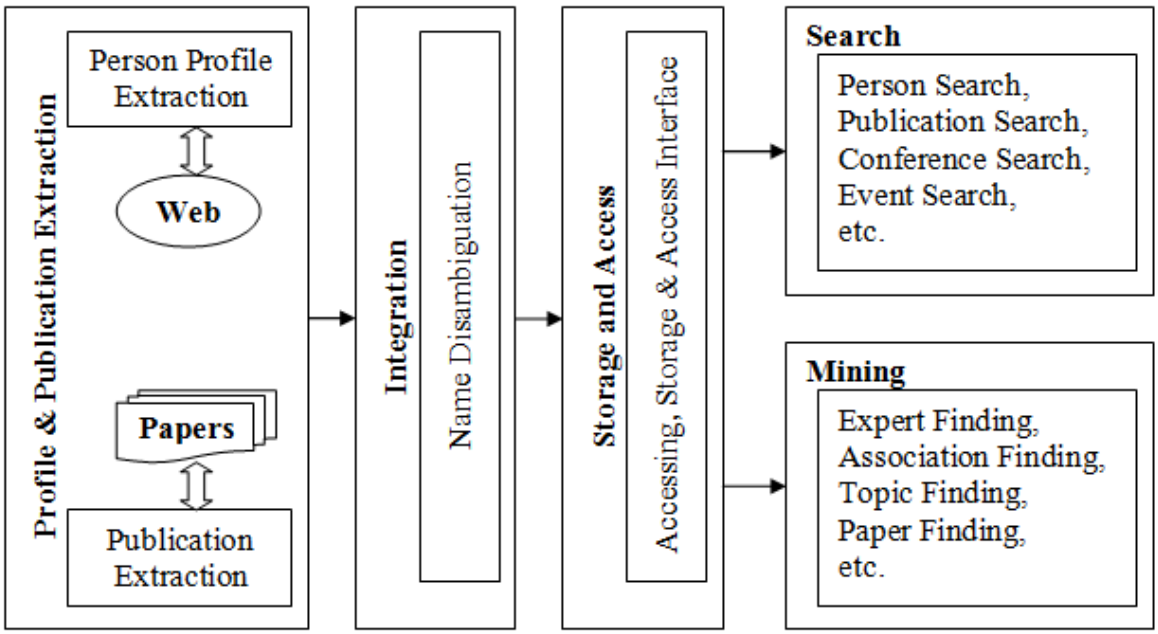

Figure 1: ***Steps in the Extraction of an Academic Social Network.***

Outline of the key steps to be followed in the proposed study:

1. Extraction of the most recent publications from the Web, search engines and even those which the search engines are not able to crawl/index or which don't find any mention in the citation digital library, will be extracted from whatever source they are present in. Blogposts, wiki edits, etc. can also be included in a researcher's publication list.
2. Integration of the extracted publications etc. and converting it into a format which can act as input to the next step for further processing. Name disambiguation is to be performed for the purpose of publications integration.
3. Developing a system that shall act as a decision support system for expert finding and recommendation purposes. Association mining, pattern detection and pattern discovery play an important role in this case. It is proposed that techniques of soft computing for web mining shall be used in order to propose a system that shall extract a social network based on the given keywords, provide expertise search and even act as a recommender system. Temporal and spectral considerations are proposed to be added as new attributes while profiling is being done. Based on these two attributes,

networks and sub-networks of interest may be extracted, which may incorporate geographic and time considerations.

4. Simulation of the proposed algorithms and system for validation and performance evaluation.

Use of the Soft Computing techniques in social network extraction will be analyzed. We intend to develop new algorithms based on the soft computing techniques for *person profile extraction* and *name disambiguation*.

Tools and Techniques proposed to be used for the proposed work:

Social network analysis software (SNA software) facilitates quantitative or qualitative analysis of social networks, by describing features of a network, either through numerical or visual representation. SNA software generates these features from raw network data formatted in an edgelist, adjacency list, or adjacency matrix (also called sociomatrix), often combined with (individual/node-level) attribute data. Network analysis software generally consists of either packages based on graphical user interfaces (GUIs), or packages built for scripting/programming languages. Packages like *UCINet*, *Pajek*, *NodeXL*, etc. will be used for the purpose of simulations.

Since web mining is a specialized branch of data mining, traditional data mining techniques such as *clustering*, *classification*, *association rule mining*, and *visualization* can also be used for web mining. In web mining, classification and clustering can be used to create different classes of users, the difference between the two is that, in *classification* classes are predefined (*supervised*) and in clustering they are not predefined (*unsupervised*). *Association rule mining* technique can be used to discover direct or indirect relationships between web entities. *Visualization* is a special technique to present data and information in graphical, understandable manner and plays an important role in web structure mining.

For this and most of the other on-line social network analyses, the three web mining techniques can't work alone and for some applications like the one under investigation all the three different web mining techniques viz. web content mining, web structure mining and web usage mining may be used altogether.

Action Plan and Time Line:

Of the steps in extraction of an academic social network discussed earlier, storage, access and searching have got quite a lot of attention from the research community over the years. Well reputed and efficient methods for the same exist. The emphasis of this proposed work shall be *extraction* and *integration*. It is proposed to carry out the research work over a period of three years and the tentative work plan for each year is as follows:

First Year:

During the first year the main task would be to study and understand various techniques proposed so far in the proposed field of investigation. What methodologies and techniques have been followed in their design and what are the loopholes and shortcomings in these approaches. An effort would be made to implement some existing social network extraction techniques to understand well the process in some other area (other than the proposed area of original author).

In addition all efforts will be made to have a comprehensive literature collection on different aspects of social network extraction such as extraction, integration, searching, mining etc.

Second Year:

Once the study of various conventional techniques proposed so far for extraction and integration is over soft-computing techniques based social network extraction will be investigated. Spectral and temporal considerations will be made and the effect of the usage of these two attributes on social network extraction will be investigated. Based on the study evaluation parameters for the proposed techniques/algorithms will be studied. Then, the few good evaluation parameters will be selected for comparative study of different methods.

In addition use of soft-computing techniques in social network extraction will be studied and the relevant techniques will be short listed for the design of proposed algorithms.

We will develop few soft-computing based social network extraction systems. The proposed system will be subjected to experimental testing for their performance. Based on the evaluation parameters decided, an effort will be made to propose the best possible system/solution.

Third Year:

The resultant system/solution will be tested for performance with the available benchmark datasets and comparison shall be drawn with the baseline methods.

Final Reporting of facts and details of background will be presented in the form of book as thesis for completion of research work. Around six months of the third year will be devoted for thesis writing and checkup.

5. Possible Outcome of the investigation

1. **Instance Unification:** Mapping various entity instances on the Web to a single real world entity.

2. **Expert findings and recommendations:** Facilities for finding experts, relevant research studies, research groups & organizations, etc. will be provided which will be an improvement over the existing models.
3. **Social Networks:** The use of data from multiple information sources along with geographical and time considerations will allow us to present a more complete picture of network under investigation.
4. **Novel Social Network Extraction Algorithms:** Novel algorithms based on web mining and soft computing techniques will be developed and used for social network extraction.
5. Any other suggestion/recommendation that will be made as per findings.

References

1. Troyano, R., Lopez, G. and Gasca, M. 2010, "***Competitive Intelligence Based on Social Networks for Decision Making.***", *International Journal of Software Engineering and its Applications*, Vol. 4, No. 4, October 2010, pages 93-104.
2. Scott, J. 2000, "***Social Network Analysis: A Handbook.***" Sage, London, 2nd Edition.
3. Ting, I-Hsien 2008, ***"Web Mining Techniques for On-line Social Networks Analysis."*** In *Proceedings of 2008 International Conference on Service Systems and Service Management,* July 2008, pages 1-5.
4. Mislove, A., et al. 2007, "***Measurement and Analysis of Online Social Networks.***", In *Proceedings of the 5th ACM/USENIX Internet Measurement Conference-IMC'07*, San Diego, CA, October 2007, pages 29-42.
5. Hope, T., Nishimura, T., and Takeda, H. 2006, ***"An integrated method for social network extraction.***" In *Proceedings of the 15th International Conference on World Wide Web-WWW '06,* Edinburgh, Scotland, May 23 - 26, 2006, ACM Press, New York, pages 845-846.
6. Nazir, F., Takeda, H., and Seneviratnel, A. 2008, ***"Comparison of Community Identification Techniques for Two-Mode Affiliation Networks Using Wikipedia Data.***" In *Proceedings of 2008 IEEE international symposium on technology and society*, June 26-28, 2008, Fredericton, Canada.
7. Tang, J., Zhang, D., and Yao, L. 2007, "***Social Network Extraction of Academic Researchers."*** In *Proceedings of International Conference on Data Mining-ICDM'07*, October 28-31, 2007, Omaha, Nebraska, USA, pages 292-300.

8. Tomobe, H., et al. 2003, "***Social Network Extraction of Conference Participants***." In *Proceedings of 12th International Conference on World Wide Web-WWW2003*, May 20-24, 2003, Budapest, Hungary.
9. Nauerz, A. and Groh, G. 2008, "***Implicit Social Network Construction and Expert User Determination in Web Portals.***" In *Proceedings of Twenty-Third AAAI Conference on Artificial Intelligence-AAAI'08*, July 13-17, 2008, Chicago, Illinois.
10. Shin, H., Xu, Z., and Kim, E. 2008, "***Discovering and Browsing of Power Users by Social Relationship Analysis in Large-scale Online Communities.***" In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, December 9-12, 2008, Sydney, NSW, Australia, pages 105-111.
11. Java, A., et al. 2006, "***Modeling the Spread of Influence on the Blogosphere.***" In *Proceedings of the 15th International Conference on World Wide Web-WWW '06*, Edinburgh, Scotland, May 23 - 26, 2006.
12. Agarwal, N., et al. 2008, "***Identifying the Influential Bloggers in a Community***." In *Proceedings of International Conference on Web Search and Web Data Mining-WSDM'08*, February 11 - 12, 2008, Palo Alto, California, USA, pages 207-218.
13. Kautz, H., Selman, B., and Shah, M. 1997, "***The Hidden Web.***" *American Association for Artificial Intelligence magazine*, 18(2), pages 27–35.
14. Furukawa, T., Matsue, Y., and Ohmukai, I. 2007, "***Social Networks and Reading Behavior in the Blogosphere.***" In *Proceedings of First International AAAI Conference on Weblogs and Social Media-ICWSM'07*, March 26-28, 2007, Colorado, USA, pages 51-58.
15. Karamon, J., Matsuo, Y., and Ishizuka, M. 2008, "***Generating Useful Network-based Features for Analyzing Social Networks.***" In *Proceedings of Twenty-Third AAAI Conference on Artificial Intelligence-AAAI'08*, July 13-17, 2008, Chicago, Illinois, pages 1162-1168.
16. Lin et al. 2006, "***Discovery of Blog Communities based on Mutual Awareness*.**" In *Proceedings of the 15th International Conference on World Wide Web-WWW '06*, Edinburgh, Scotland, May 23 - 26, 2006.
17. Matsuo, Y., et al. 2006, "***POLYPHONET: an advanced social network extraction system from the web.***" In *Proceedings of the 15th International Conference on World Wide Web-WWW '06*, Edinburgh, Scotland, May 23 - 26, 2006, ACM Press, New York, pages 397-406.

18. Ting, I., Wu, H., and Chang, P. 2009, ***"Analyzing Multi-Source Social Data for Extracting and Mining Social Networks."***, In *Proceedings of International Conference on Computational Science and Engineering-2009*, pages 815-820.
19. Matsuo, Y., Tombe, H., and Nishimura, T. 2007, ***"Robust Estimation of Google Counts for Social Network Extraction."***, In *Proceedings of Twenty Second Conference on Artificial Intelligence (AAAI-07)*, Vancouver, Canada, July 2007, pages 1395-1401.
20. Jin, Y.Z., Matsuo, Y., and Ishizuka, M. 2007, ***"Extracting Social Networks among Various Entities on the Web."*** In *Proceedings of the Fourth European Semantic Web Conference*, Innsbruck, Austria, June 2007, pages 251-266.
21. Tang J., et al. 2008, ***"Arnetminer: Extraction and Mining of an Academic Social Network.***" In *Proceedings of 1^{7th} International World Wide Web Conference-WWW'2008*, April 21-25, 2008, Beijing, China, pages 990-998.
22. Song, M., Lee, T., and Kim, J. 2010, ***"Extraction and Visualization of Implicit Social Relations on Social Networking Services.***" In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence-AAAI'10*, July 11–15, 2010, Atlanta, Georgia, pages 1425-1430.
23. Mika, P., 2005, ***"Flink: Semantic web technology for the extraction and analysis of social networks."*** Journal of Web Semantics, 3(2), 2005.
24. Salton, R., 1989, ***"Automatic Text Processing. Reading, Mass."*** Addison Wesley.
25. C. D. Manning and H. Schutze, 2002, ***"Foundations of statistical natural language processing."*** The MIT Press, London, 2002.
26. Brickley, D. and Miller, L., 2004, ***"FOAF Vocabulary Specification."*** In *Namespace Document, http://xmlns.com/foaf/0.1/*, September 2004.
27. Cortes, C. and Vapnik, V. 1995, "***Support-Vector Networks***", *Machine Learning*, 20, 1995: pages 273-297.
28. Lafferty, J., McCallum, A. and Pereira, F. 2001, ***"Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data."***, In Proceedings of ICML'01, 2001.
29. Tyler, J. R., Wilkinson, D. M., Huberman, B. A. 2003, "***Email as spectroscopy: automated discovery of community structure within organizations.***" In *Proceedings of C&T2003*.

30. Culotta, A., Bekkerman, R., and McCallum, A., 2004, ***"Extracting social networks and contact information from e-mail and the web."*** In *Proceedings of Conference on Email and Spam*.
31. Van Alstyne, M., and Zhang, J. 2003, "***EmailNet: A system for automatically mining social networks from organizational email communication.***" In *NAACSOS2003*.
32. Bird, C., Gourley, A., Devanbu, P., Gertz, M., and Swaminathan, A. 2006, ***"Mining Email Social Networks."***, In *Proceeding of MSR 2006*, Shanghai, China, May 22-26, 2006, pages 137-143.
33. Allen, T. 1984, ***"Managing the Flow of Technology."*** MIT Press.
34. Hinds, P. & Kiesler, S. 1995, "***Communication across boundaries: Work, structure, and use of communication technologies in a large organization."*** Organization Science, 6 , pages 373-393.
35. Garton, L., Haythornthwaite, C. and Wellman, B., 1997, ***"Studying online social networks."*** Journal of Computer-Mediated Communication, Vol. 3, No. 1, http://jcmc.indiana.edu/vol3/ issue1/garton.html.
36. Alani, H., O'Hara, K., and Shadbolt, N. 2002, ***"ONTOCOPI: Methods and Tools for Identifying Communities of Practice, Intelligent Information Processing Conference"***, IFIP World Computer Congress (WCC), Montreal, Canada, 2002.
37. Wilkinson, D. and Huberman, H., 2002, ***"A Method for Finding Communities of Related Genes."*** Submitted for publication, http://www.hpl.hp.com/shl/papers/communities/index.html.
38. Mutton, P., 2004, ***"Inferring and Visualizing Social Networks on Internet Relay Chat."*** InfoVis, Austin, TX, USA, pages 35–43.
39. Resig, J. and Teredesai, A., 2004, ***"A framework for mining instant messaging services."*** In *Proceedings of the 2004 SIAM Workshop on Link Analysis, Counter-terrorism, and Privacy*, Lake Buena Vista, Florida, April 24, 2004.
40. Resig, J., Dawara, S., Homan, C., and Teredesai, A., 2004, ***"Extracting social networks from instant messaging populations."*** In: Workshop on Link Analysis and Group Detection (LinkKDD2004), USA.
41. Mutton, P., 2001, "***PircBot Java IRC Bot Framework***", http://www.jibble.org/pircbot.php

42. Fruchterman, T.M.J. and Reingold, E.M., 1991, ***"Graph Drawing by Force-Directed Placement."*** Software Practice and Experience Vol. 21(11), pages 1129-1164.
43. Wang, K., Ting, I., Wu, H., and Chang, P. 2010, ***"A Dynamic and Task-Oriented Social Network Extraction System Based on Analyzing Personal Social Data."*** In *Proceedings of 2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 464-469.