DESIGN OF AN INTEGRATED QUERY PROCESSING SYSTEM FOR SOCIAL WEB

THESIS

submitted in fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

YMCA UNIVERSITY OF SCIENCE & TECHNOLOGY

by

CHARU VIRMANI

Registration No: Ph.D.-04-2K12

Under the Supervision of

Dr. ANURADHA PILLAI
ASSISTANT PROFESSOR

Dr. DIMPLE JUNEJA
DEAN, R&D

Department of Computer Engineering
Faculty of Engineering and Technology
YMCA University of Science &Technology
Sector-6, Mathura Road, Faridabad, Haryana, India

September 2018

Dedicated to my kids….

DECLARATION

I hereby declare that the thesis entitled “**DESIGN OF AN INTEGRATED QUERY PROCESSING SYSTEM FOR SOCIAL WEB”** by **CHARU VIRMANI,** being submitted in fulfillment of requirement for the award of Degree of Doctor of Philosophy in the Department of Computer Engineering under Faculty of Engineering and Technology of YMCA University of Science and Technology, Faridabad, during the academic year March 2013 to September 2018, is a bonafide record of my original work carried out under the guidance of **Dr. Anuradha Pillai, Assistant Professor, Department of Computer Engineering, YMCA University of Science and Technology, Faridabad and Dr. Dimple Juneja, Dean, Research and Development, Poornima University, Jaipur** has not been presented elsewhere.

I further declare that the thesis does not contain any part of any work which has been submitted for the award of any other degree either in this university or in any other university.

CHARU VIRMANI

Ph.D-04-2K12

CERTIFICATE

This is to certify that the thesis entitled **“DESIGN OF AN INTEGRATED QUERY PROCESSING SYSTEM FOR SOCIAL WEB”** by **CHARU VIRMANI,** being submitted in fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Computer Engineering under Faculty of Engineering and Technology of YMCA University of Science and Technology, Faridabad, during the academic year March 2013 to September 2018, is a bonafide record of work carried out under our guidance and supervision.

We further declare that to the best of our knowledge, the thesis does not contain any part of any work which has been submitted for the award of any degree either in this university or in any other university.

Dr. Anuradha Pillai
Assistant Professor,
Department of Computer Engineering
YMCA University of Science and Technology
Faridabad, Haryana

Dr. Dimple Juneja
Dean,
Research and Development
Poornima University
Jaipur, Rajasthan

Dated:

ACKNOWLEDGEMENTS

I would first like to thanks my supervisors **Dr. Anuradha Pillai, Assistant Professor, Department of Computer Engineering, YMCA University of Science and Technology, Faridabad and Dr. Dimple Juneja, Dean, Research and Development, Poornima University, Jaipur** for their mentoring and guidance during this research work. They have not only inspired, guided and supported but also made my dream of pursuing and completing the research work true. I am grateful to both of them for providing a cultivating environment that ropes liberty, open consultations, and constructive criticism, without which this work would not have been possible. They have always encouraged me to explore innovative things and were always there to guide me for improvisation by giving their invaluable suggestions. They have taught me to be strong and push forward when it is hard. I thank both of them from all my heart for the help in all aspects of my Ph.D. I am indebted to the long 'wisdom talks', which will benefit me throughout my future.

I would like to express my sincere gratitude to Prof. Komal Bhatia and Prof. Atul Mishra, Professors, Department of Computer Engineering, YMCAUST for trusting and believing in me. Both of them inspired me to set the research standards high and not to compromise with them. They have cheered for me during my ups and cheered me up during my downs. I am extremely honored to have worked and interacted with both of them. I am also thankful to all my students, faculty and staff members of the department who helped me directly or indirectly in completing my work.

I would like to thank my husband, Vikas Pujara for inspiring me and showing me that there are no limits and I can pursue whatever I am passionate about. He is the most amazing person I have met and I am very blessed to have him near me. He constantly cheered me when I was down and helped me when I needed advice and aid on my research.

Finally, I would like to thank my supportive and affectionate family for their constant support. I am grateful to my Mother for a wonderful education, my Brother, and my Mother-in-law for supporting my education.

I wish to thank my *Daughter Manya and Son Ayan* for their unconditional love, affection and the tolerance they have shown all these years.

Charu Virmani

ABSTRACT

Social Web is a matrix of interconnected relations and connections that connects people. In this fast-growing world of technology, online social networks (OSN's) have become most important sources of information and communication. These networks have not only affected many individuals but also companies and organizations including business, education, healthcare, and politics. Nowadays, almost every person has an account on different social media sites like Google+, Facebook, Twitter, Instagram etc. Managing these social media accounts can be easy but much more complicated at other times. For instance, posting something in the news feed is quite straightforward and less time-consuming. However, performing the same task on multiple social media sites could be time consuming; it has become a serious issue to manage OSN’s in many big organizations. Subsequently, it has become very difficult for users to manage different online social networks data files without using any particular tool (which can help users to manage their data with ease) because of factors such as dynamicity of social networks, the amount of data being managed regularly (data being added or deleted). In addition, it is also difficult for a user to monitor his/her data every time and remain updated with the latest information among multiple social networks. The most promising solution to the above listed reservations is Social Network Aggregator (SNA). SNA collates data spread across multiple social network services. The idea is to organize and ease the information retrieval process for a user maintaining multiple social networks actively. SNA consolidates the various social activities/data in such a way that user is not required to login to each site exclusively and performs the same social activity. Aggregator helps the user to perform the social activity at one site and the information gets synchronized to all the social networks that the user specifies. However, a lot of research is still going on aggregators to provide better integration of social data.

Conventionally, most of the query engines for the social network aggregator respond to the user’s request by using keyword search which in turn returns a huge lot of information comprising of both relevant and irrelevant information. Although there are various social network aggregators available, however; to the best of our knowledge and understanding, none of the aggregators have efficiently executed the

search of a user across multiple social networks. Thus the need for a processing system which allows searching a query in a user-friendly way and also returns more relevant results is highly apparent. Hence, ***Integrated Query Processing System for Social Web (QPSSN)*** is being proposed that exploits natural language techniques for query processing and extracting information from the social web. QPSSN offers an edge over other mechanisms as it not only retrieves more user-centric results as compared to traditional way of keyword based searching but also with more relevant results. Although natural language techniques are finding space in semantic search engines, however; the same has not been effectively applied against the response of queries executed on social networks by any other researcher to the best of our knowledge. Thus the motivation for QPSSN is to find a viable solution that can provide the intelligent and integrated result of user's free form query to get more user centric results.

Literature was grilled to explore the barriers to the design of QPSSN and it was discovered that social networks have diverse network structures that make the task of linking profiles difficult. Moreover, it is an evident challenge that many users may exist with identical usernames and can provide false information across their profile in order to masquerade. There is no easy mechanism available that can extract and map the entities from the query entered by the user in the social environment. Each Social Network Services (SNS) have their own syntaxes and terms for representing social data specific to their network and above all, there is a different meaning attached to the same term among different social networks. Last but not the least; user satisfaction has been a critical factor in determining the output of the query. To make the integration of user conceivable and generating succinctly results to the query, a need for an optimal query processing technique is highly apparent. There is a need to bridge the gap between user representation and intelligence of query processing system to provide more reliable and relevant results to the user's query.

On the basis of literature grilled during the initial phase of the research work, the need for a novel and efficient algorithm for integrating user's profile scattered over multiple social networks is unavoidable. Hence, the work proposes ***Hybrid Integrated Autonomous Social Network (HIASN),*** a novel architecture for integrating the profiles of the user in an effective manner. A clustering mechanism termed as ***Hybrid***

Ensemble k-Means Hierarchical Agglomerative clustering (HEKHAC), is also being proposed which uses user's publicly available attributes to make optimal clusters to retrieve the desired information from the query written in a natural language. Another significant contribution is made through ***Query Processing in Social Networks Aggregator (QPSNA)*** which includes four modules namely, ***Query Processing System (QPS), Content Based Semantic Matcher Maker (CBSMM), Machine Learning Mechanism (MLM) and Ranking of results*** to answer user's query**.** QPSNA extracts entity that is then mapped it to its semantic meaning, identifies user preferred profiles and improves upon the user's preference by ranking the profiles. The proposed system integrates several social websites together and responds to a user's query; extracting the relevant data as specified in query written in a natural language from multiple social networks and presenting data appropriately as result; thereby helping users who belong to multiple networks manage diverse profiles across multiple social networking sites. The proposed system will maintain several accounts at one place and extracts the relevant publicly available data. It also aims to offer an improvement over keyword searching by using NLP techniques.

The effectiveness of the proposed work has been practically established by evaluating the model/algorithms on various measurements and claims the accuracy. The results obtained have been analyzed and compared on parameters such as Precision, Recall, Receiver Operating Characteristic (ROC) graph, similarity measure and classifiers like Naïve Bayes, Logistic Research and Support Vector Machine- Linear as well as Kernel with existing popular mechanisms. It is worth mentioning that the results obtained are competitive and offers a significant breakthrough in the field.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATION	DESCRIPTION
APIs	Application Programming Interface
BOW	Bag Of Words
CBSMM	Context Based Semantic Match Maker
CID	Cluster Id
DJIA	Dow Jones Industrial Average
DOAC	Description of a Career
FOAF	Friend Of A Friend
GUMO	General User Model Ontology
HEKHAC	Hybrid Ensemble K-Means Hierarchical Agglomerative Clustering
HIASN	Hybrid Integrator for Autonomous Social Networks
HMM	Hidden Markov Models
IDF	Inverse Document Frequency
IRM	The Identity Resolver Module
JSON	Javascript object Notation
LDA	Latent Dirichlet Allocation
MLM	Machine Learning Mechanisms
NER	Named Entity Recognition
NLP	Natural Language Processing
NP	Noun Phrase
OPM	Open Provenance Model
OSN	Online Social Network
PIM	Profile Integration Module
POS	Parts-of-Speech
QPS	Query Processing System
QPSSN	Query Processing System for Social Network
ROC	Receiver Of Characteristics
RDF	Resource Description Framework
RSS	Rich Site Summary
SCOT	Social Semantic cloud of Tags

SIOC	Semantically Interlinked Online Communities
SNA	Social Network Aggregator
SNS	Social Network Services
SQL	Structured Query Language
SVM	Support Vector Machine
TF	Term Frequency
TGQ	Temporal Group Query
TSN	Temporal Social Network
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
WI	Weighted Interest
XFN	XHTML Friends Network
VSM	Vector Space Model

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

Social Web [1][2] is a matrix of interconnected relations and connections that connects people. In this fast-growing world of technology, online social networks (OSN's) [2] have become one of the most important sources of information and communication. These networks have not only affected many individuals but also companies and organizations including business, education, healthcare and politics. Now a day, almost every person has an account on different social media sites like Google+[1], Facebook[2], Instagram[3], Twitter[4] etc. Managing these social media accounts can be easy but much more complicated at other times. For instance, posting something in the news feed is quite straightforward and less time-consuming. However, performing the same task on multiple social media sites could be time-consuming; it has become a serious issue to manage Online Social Networks (OSN) in many big organizations.

Subsequently, it has become very difficult for users to manage different online social networks data files without using any particular tool (which can help users to manage their data with ease) because of factors such as dynamicity of social networks, the amount of data being managed regularly (data being added or deleted) [3][4][5][6][7]. In addition, it is also difficult for a user to monitor his/her data every time and remain updated with the latest information among multiple social networks.

The most promising solution to the above-listed reservations is Social Network Aggregator (SNA) [8][9][10[11] which provides a platform where a user can login to any social networking site and manage his/her account. It saves time and cost by

[1] https://plus.google.com/discover
[2] https://www.facebook.com
[3] https://www.instagram.com/?hl=en
[4] https://twitter.com/

integrating multiple social networks on the global platform. Conventionally, most of the query engines for the social network aggregators respond to the user's request by using keyword search [8][42] which in turn returns a huge lot of information comprising of both relevant and irrelevant information. Although there are various social network aggregators available, however; to the best of our knowledge and understanding, none of the aggregators have tried effectively to execute the search of a user in a natural language across multiple social networks. Thus the need for a processing system which allows searching a query in a user-friendly way and also returns more relevant results is highly apparent. Hence, Integrated Query Processing System for Social Web (QPSSN) is being proposed that exploits natural language techniques for query processing and extracting information from the social web.

QPSSN offers an edge over other mechanisms as it not only retrieves more user-centric results as compared to traditional way of keyword based searching but also produced more relevant results from user perspective. Although natural language techniques are finding space in semantic search engines, however; the same when will be applied to the response of queries executed on social networks will provide more favorable results. Thus the motivation for QPSSN is to find a viable solution that can provide the intelligent and integrated result of user's free form query.

The upcoming section briefs about social web laying the foundation of current work.

1.2 SOCIAL WEB AND SOCIAL NETWORK AGGREGATORS

Social web [1][2] deals with identifying and establishing connections among individuals through the social networking sites. A social web works by connecting technologies that enable people and community to generate and share content online with one another [12]. Online social web sites like LinkedIn, Instagram, Facebook and Twitter are the most prevalent sites on the internet. The social characteristic of communication over the social network is to facilitate interaction between users with similar tastes that vary depending on who the user is, and what they are interested in. Its influence on the people is large and ever changing [1]. The people (users of Internet), the community (the network of organization or groups of friends) and the content generated by users (like posts, videos, and images) are the core components of the OSN as depicted in figure 1.1.

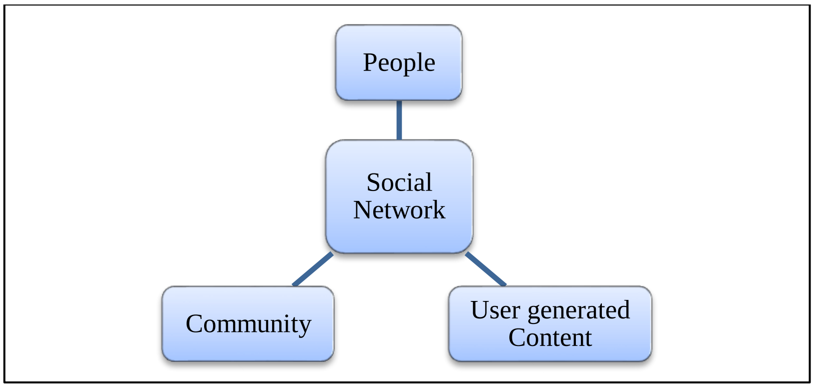


Figure 1.1 Core Components of Social Network

Further, reports [13] indicate the existence of users having multiple accounts at multiple online social network services (shown in Table 1.1).

Table 1.1 Survey of Users having Multiple Accounts

Site	Linked In	Facebook	Frien dster	Bebo	Orkut	Plaxo	Ning	MyS pace	Hi5
LinkedIn	100	42	8	4	3	3	8	32	2
Facebook	2	100	2	4	1	9	1	64	2
Friendster	6	23	100	5	1	0	2	49	4
Bebo	1	25	2	100	0	0	1	65	3
Orkut	8	35	4	3	100	1	2	29	7
Plaxo	54	48	8	5	4	100	14	34	2
Ning	19	35	6	6	2	2	100	44	1
MySpace	0	20	1	3	0	0	0	100	1
Hi5	1	24	4	7	2	0	0	69	100

The analysis of Table 1.1 reflects that the majority of users on MySpace, LinkedIn, and Twitter also have an account on Facebook. Having multiple accounts on various social networks is not an issue. However, organizing and managing of the content/contact or other social activities which is generated by the user is a major concern. In order to resolve the stated concern, Social Network Aggregator (SNA) [8][9][10[11] is a favorable answer. It is the process of collating/aggregating/ organizing data spread across multiple social network services. The idea is to organize and ease the information retrieval process for a user maintaining multiple social networks actively. SNA consolidates the various social activities/data in such a way that user is not required to login to each site exclusively and performs same social activity.

Aggregator helps a user to perform the social activity at one site and the information gets synchronized to all of the social networks that the user specifies. Aggregation tools are in place that provides users to consolidate messages, track social data across networks. Content Aggregators [14], Comparison Analytics [15], Relationship Aggregation [16] and Process Aggregation [17] are typical ways to integrate social data across multiple social networking sites. However, a lot of research is still going on aggregators to provide better integration of social data.

A social network aggregator aggregates the social-network members and social data to share social-network activities. The very rationale for having an aggregation is to let the user have one unified window to manage his social interaction and activities without hopping on each Social Network Services (SNS) [18] separately. All content appears in real time (or abstracts to be appearing), which eliminates the need to hop from one SNS to other. Justification of having aggregation lies in the fact that not every SNS can be the best place for a user having varying interest and hobbies.

Social data across multiple sites can hence be integrated on a common platform or protocol [8] as a representation of abstraction of user’s preferences. Since each SNSs have their own syntaxes and formats for representing social data, open web communities have developed standard representations of social data and the most important methods are OpenID, Activity stream etc. [19]. These standards provide the basic foundation and building blocks for social aggregators. IT companies and SNS have implemented these standards. For example, Facebook's News Feed is an activity stream. Keeping context of the user’s data in integrating social networks is another point of consideration where context implies social bound, the relationships, common interests, etc. [20] that actually forms the basis for connecting users on different SNS.

The volume, velocity, and variety of these users vary with the OSN. This results in the evolution of multiple profiles of the same user scattered across the Internet with no platform to detect the presence of one another. These disparate unlinked user raises concern for various patrons. For Example, it is very difficult for multinational companies and non-profit organizations to authenticate attributes across unlinked users and construct holistic footprints of their customers. Thus, there is a need for better aggregation and extraction algorithms that can provide user's publicly available information which is distributed over the OSN and may be related to individuals associated with a given username.

Social network aggregators are available that rely on Application Programming Interface (APIs) provided by the SNS for accomplishing the aggregation process but the users have to authenticate and give suitable permissions so that it can access and collect data from user's account. Once the access is gained, all recent information will be pulled from OSN into the aggregator.

Hootsuite[5], TweetDeck[6], and FriendFeed[7] are some known examples of OSN. Hootsuite and TweetDeck are social network management tools that aggregate user's information used for the professional purpose [21]. It also includes advanced features like scheduling information analysis, bookmark, posts, share in advance, RSS feeding [22] allows businesses and organizations to effectively lead their marketing campaign across SNSs. FriendFeed aggregates posts, updates, and photos of the user from multiple OSN so that the user can perform all services of social network in real-time..

Social Network offers a large number of applications in various forms that enable people to create public profiles and build social relations with friends who share similar interest, backgrounds or real life connections like Facebook, LinkedIn[8], and Google+, Twitter etc. The collecting body of information embodied by the community, networks or social circle is another aspect which helps the organizations to improve knowledge access and sharing for higher user-productivity and performance in the social media called as social knowledge like Knowledge Plaza. Numerous social networks utilize online social collaboration to make a scaffold to real life interaction. Connections are shaped between people by means of the social web and afterward turn out to be more personal through other forms of correspondence to share social data.

[5] https://hootsuite.com/#
[6] https://tweetdeck.twitter.com
[7] https://www.facebook.com/friendfeed
[8] https://www.linkedin.com/

Figure 1.2 depicts various channels supporting the existence and execution of social web.

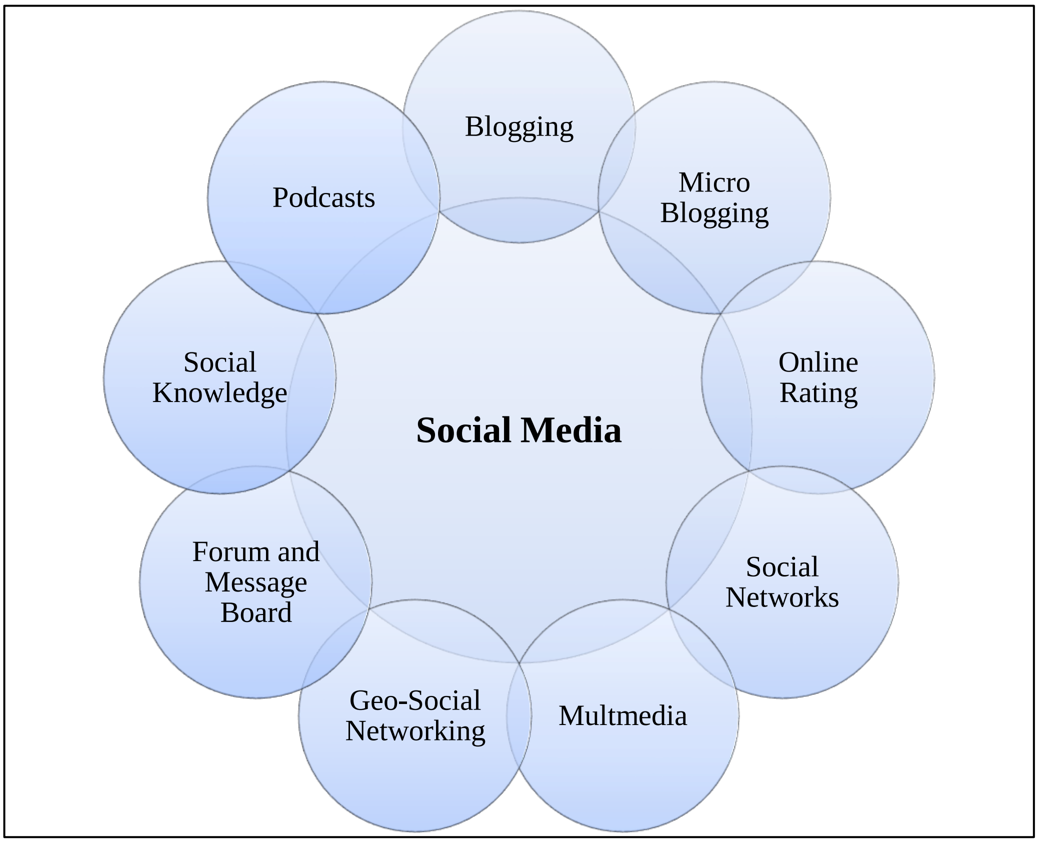


Figure 1.2 Types of Social Media

For instance, an informational journal called Blogs for discussion with discrete diary style posts allows posts to appear in reverse chronological order i.e. the most recent displayed first. Wordpress[9], Blogger[10], and Tumblr[11] are the popular blog services [23] that are used for small discussions.

Another broadcast medium of Blog is Microblogging [24] which allows users to exchange small sentences, images or video links to promote public relations, websites, services and products, and to promote collaboration within an organization like Twitter. Podcasts [25] are the audio or video files which helps a subscribed user to automatically download the new episodes using web syndication to the users own computer, mobile or any portable medium like Apple's iTunes[12].

[9] https://wordpress.com/
[10] https://www.blogger.com
[11] https://www.tumblr.com/
[12] http://www.apple.com/itunes/

Another popular application like Forum [26] is an online discussion platform on some specialized topic and interest. Online Rating [27][28] is an online attempt used for rating and reviewing movies, comic books or games to the users whereas Geo-Social Networking [29] is a Geo-based social Networking in which geographic services and capabilities are used to empower additional social dynamics using geocoding and geotagging. Geolocation systems [30][31] allow social networks to link and synchronize users with local people or events as per their interests. It makes use of texted location information, IP or hotspot trilateration or mobile phone tracking to enrich social networking like Geofeedia.

Multimedia based SNS allows users to organize, share and embed personal content like images, videos and corporate media videos like video clips, music videos or short documentary etc. like YouTube[13]. It offers an online community which is widely used by researchers and bloggers to host images and videos that they embed in blogs and social media. The above listed applications are the most popularly used and considered to be the conventional part of the OSN. It is to be noted that the distinctions among the different categories of social media are getting blurred. For example, social network sites and Multimedia based OSN overlap more and more.

Alongside the exploding popularity of the social networks, it raises some issues and challenges as well. It is worth mentioning that the socio-cultural ecosystem of the social media is complex as new services are created dynamically and further, constant changes to communication between people, groups and organizations also add complexity to the system. The upcoming section raises the concerns pertaining to the versatility of social networks.

1.2.1 Social Network Concerns

Over the years, many of the children have started misusing social networks to create discomfort and become victims of cyber bullying or cyber-stalking by creating a fake account and performing activities like threats, intimidation messages, and rumors that can be sent to the masses without being traced and may lead to occasional suicides.

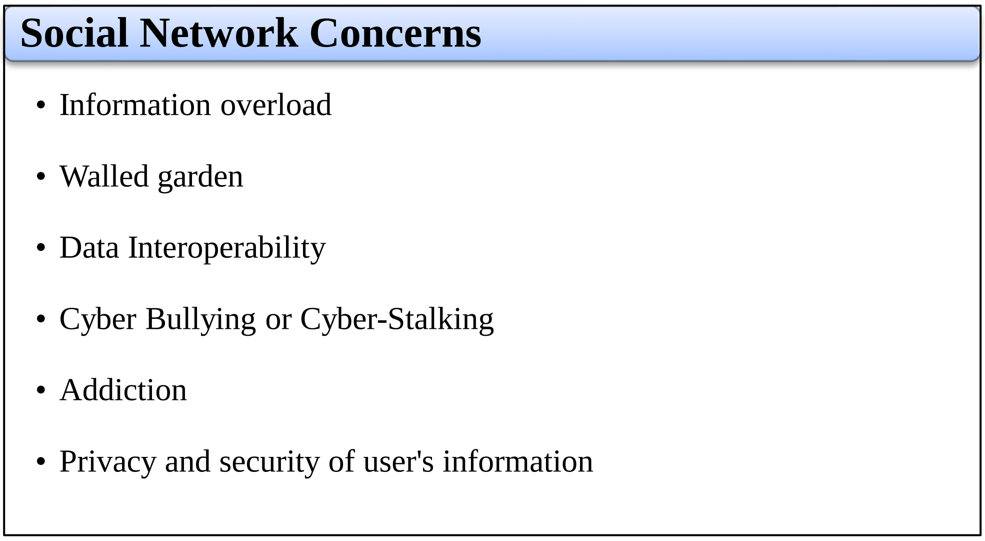
[13] https://www.youtube.com/

Addiction is another major problem of OSN. The youth are the most affected and spends extensive time on the social network which can be used by dynamic tasks and activities and thus cut off from the society. Another major concern is privacy and security of the user’s information; it is easy for a hacker to harm a user’s financial assets or personal life by fringing the user’s data. One’s personality can easily be ruined by compromising privacy using Identity theft or other known threats available like spreading false story across OSN. The above discussed adverse facets of misusage social network by the user are only some illustrative examples of issues.

Given the high volume and the veracity of today SNSs, it deserves energies of personal, social, government as well as the SNS providers to overcome the challenges of such issues. There are some other major concerns such as “Information overload” [32][33][34][35][36] and “Walled gardens” [8][37][38][41][64] that have originated from OSN that prevents user’s behavior from efficiently exploiting services of social network. The former is concerned with the continuous increase in the volume of the social network and later is concerned with the disconnected landscape that multiple social networks offer [8].

Users on the social network are overwhelmed by the huge volume of the incoming data related to their friends, organizations or companies. A user receives enormous information per day like profile updates, posts, video and so forth which is beyond their capability to process. Some social networks have provided filters like keyword/hashtags search as a solution to the above raised problem but have failed to provide a complete personalized solution for a user.

The Streamlined presentation like Newsfeed in case of Facebook is another attractive feature which suffers a major drawback when comes to search for the information of interest as the user has to follow the stream to discover the information of his/her interest. As a result, much valuable and interesting information continue to be neglected [33][35][36][41]. Isolation of SNS due to lack of interoperability between the provided services is another challenge that leads to the walled garden problem [37][38][41]. As a consequence, a user has to create a new profile, connect with friends and reuse their data which already exist at multiple social sites. Thus, the user profiles are scattered over multiple social networks and make inconvenient to handle multiple accounts. Moreover, it requires user intervention to remain updated resulting in either stopping using SNS or killing his/her precious time in the search for important information. Figure 1.3 summarizes the barriers to the success of social networks.



Social Network Concerns

- Information overload
- Walled garden
- Data Interoperability
- Cyber Bullying or Cyber-Stalking
- Addiction
- Privacy and security of user's information

Figure 1.3 Barriers to the Success of OSN

1.2.2 Design Challenges

Although it is evident that social networks are extremely popular among naive as well trained users, however; these are possessed with certain inherent design challenges such as informal language, short contexts, noisy sparse content etc. as briefly illustrated in figure 1.4 and discussed as follows.

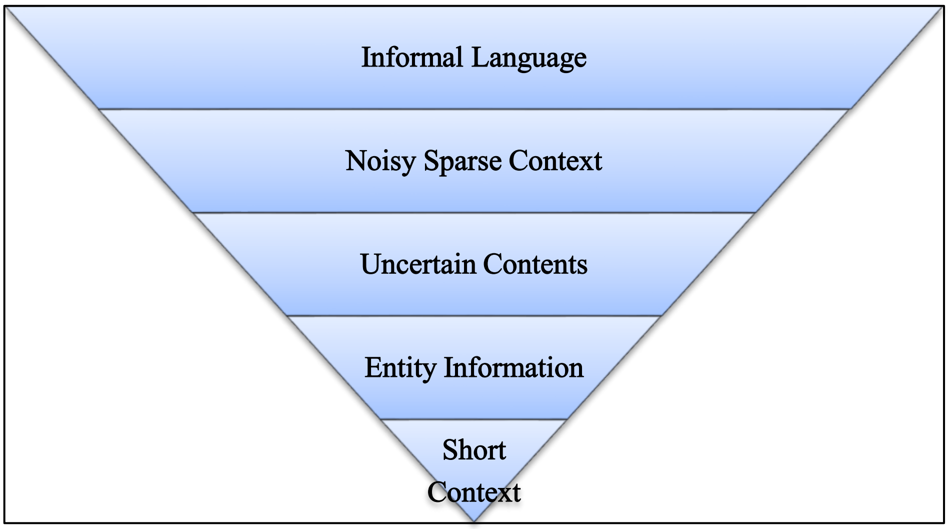


Figure 1.4 Design Challenges of Social Networks

- ***Informal Language***

Social Network users post texts in an informal language which is not only noisy but also lack in punctuation, misspelled, uses non-standard abbreviations, capitalization, shorthand’s, and do not contain grammatically correct sentences.

- ***Short Contexts***

OSN poses minimum length like Twitter; the user uses more abbreviations to precise more information in their posts. It is difficult to disambiguate mentioned entities due to the shortness of the posts and to resolve co-references among the feeds.

- ***Noisy Sparse Contents***

The users’ post on a social network does not always contain useful information. Thus, filtering is required as a pre-processing step to purify the input posts stream. The significant purpose of this step is to classify raw sentences into sentences which can be read by the machine.

- ***Information About Entities***

People use the social network to express information about their daily routine, happenings or about events and thus the entities are not stored in the Knowledge Base. The information extraction approach is to link the entities involved in the extracted information to a knowledgebase. Thus, there is a strong need for new suit of information extraction from social network posts.

- ***Uncertain Contents***

Not all information is trustworthy on the social network. The information contained in the users’ contributions is in conflict when confirmed with other sources and sometimes untrustworthy. The uncertainty involved in the extracted relations/facts is difficult to handle.

Literature [43][44][46][47] was thoroughly grilled to explore the enablers and barriers to the success of SNS forming the foundation of upcoming section.

1.3 RESEARCH CHALLENGES

With the growing popularity of the social web and expansion of the information on the social web, various social network aggregators have been designed to aggregate user's information and provide a single platform to access the services of multiple social networks. Although various SNA is performing the jobs assigned, however, social networks demands a more intensive query processing system that can integrate user's queries across the globe. While this research study commenced with the understanding of requirement of novel query processing system, the engraved study of literature [8][43][46][47] revealed that integrating various social networks is a stimulating task as it is associated with various design issues listed as follows:

- ***Diverse Network Structures***

  Social Networks have diverse network structures and profile attributes (like name, location etc.) that makes the task of linking profiles difficult.

- ***Multiple Profiles of the Same User***

  Users may choose their username depending upon the functionality and service of the social network that may not be associated with their real identity. It is an evident challenge that many users may exist with identical usernames and can provide false information across their profile in order to masquerade.

- ***Scattered Profiles***

  There is no effective solution that can integrate the profiles of the user available across multiple social networks.

- ***Extraction and Mapping of Entries***

  There is no effective mechanism available that can extract and map the entities from the query entered by the user in the social environment. Each Social Network Services have their own syntaxes and terms for representing social data specific to their network and above all, there is a different meaning attached to the same term among different social networks.

- ***Optimal Query Processing***

User satisfaction has been a critical factor in determining the output of the query. To make the integration of user conceivable and generating succinctly results to the query, a need for an optimal query processing technique is highly apparent.

- ***Gap Between User Representation and Intelligence of Query Processing System***

There exists gap to discover the output that a user expects by the representation of the query only. There is a need to bridge the gap between user representation and intelligence of query processing system to provide more reliable and relevant results to the user's query.

1.4 RESEARCH OBJECTIVES

On the basis of literature grilled during the initial phase of the research work, following objectives are being identified in the light of challenges already stated in the previous sections:

- *To design a novel algorithm for integrating user's profile scattered over multiple social networks.*
- *To develop intelligent clustering and sorting method to establish an effective result of the query.*
- *To design a novel algorithm which can provide intelligent answers to the user's free form query.*
- *To design an efficient strategy for giving appropriate ontology to the keywords of the user's query and ranking the result of the user's query considering user's interest.*
- *Evaluation & comparison of proposed work with existing conventional techniques used for the information retrieval.*

1.5 THE PROPOSED SOLUTION

The dwelling of literature clearly indicates the fact that aggregating the profiles of the social network provides a large amount of information about the user and the existing literature lacks to extract the information from this pool in a user-friendly way. The work proposes a novel architecture for integrating the profiles of the user in an effective and efficient manner.

A clustering mechanism is also being proposed which uses user's publicly available attributes to make optimal clusters to retrieve the desired information from the query written in a natural language. It extracts entity that is then mapped it to its semantic meaning, identifies user preferred profiles and improves upon the user's preference by ranking the profiles.

The proposed system integrates several social web sites together and responds to a user's query; extracting the relevant data as specified in a query written in natural language from multiple social networks and presenting data appropriately as result; thereby helping users who belong to multiple networks manage diverse profiles.

The proposed system will maintain several accounts at one place and extracts the relevant publicly available data. It aims to offer an improvement over keyword searching. The high level abstract view of the proposed work is depicted in figure 1.5. The next section presents the structure of the thesis.

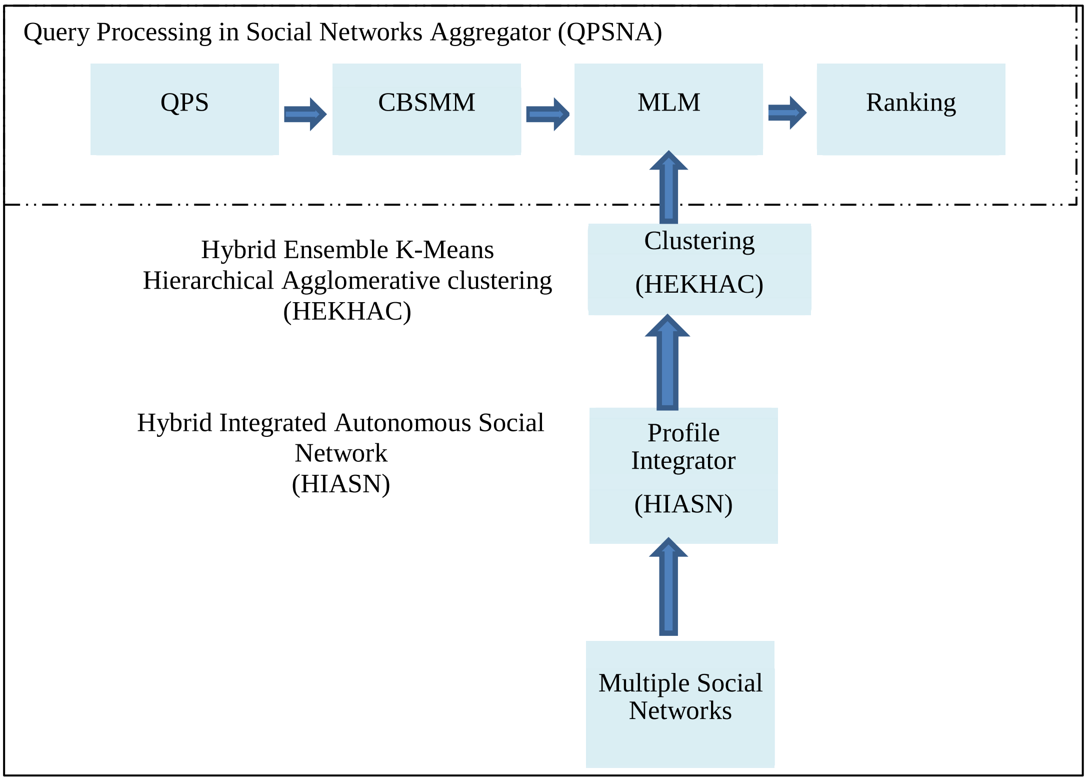

Query Processing in Social Networks Aggregator (QPSNA)
QPS
CBSMM
MLM
Ranking
Hybrid Ensemble K-Means Hierarchical Agglomerative clustering (HEKHAC)
Clustering (HEKHAC)
Hybrid Integrated Autonomous Social Network (HIASN)
Profile Integrator (HIASN)
Multiple Social Networks

Figure 1.5 Abstract View of QPSSN

1.6 STRUCTURE OF THESIS

The research work is principally carved into seven chapters as listed below:

Chapter 1 discusses the motivation for the research work and presents a brief idea of background concepts necessary for commencing the research work. It also illustrates the design issues serving as hurdles to the success of social network aggregators and also presents the major research objectives to be achieved during this course of work.

Chapter 2 begins by presenting detailed information about the social web, social networks, and social network aggregators. It also enlists a comparison of existing social media management tools. Further, the chapter also throws light on the challenges associated with the field and concludes by exploring the feasibility of deploying natural language search in social networks. It also discusses the various techniques of NLP that can be employed to extract meaningful information from the social networks.

Chapter 3 provides an insight into the existing techniques which motivated this research work. The very nascent idea of searching the social network using natural language has emerged because of a thorough study of the available literature which indicated that research should be carried forward in three different phases as listed in chapter 4. This chapter provides the backdrop of existing works pertaining to the mentioned phases and further explores the possibility of improvements.

Chapter 4 furnishes three phased QPSSN model, a novel approach which is presented in the light of drawbacks in the existing work. This chapter discusses only the first two phases of the proposed approach. The first phase proposes that extracts user profile from multiple social networks, aggregates and provides an integrated user profiles, ***Hybrid Integrated Autonomous Social Network (HIASN)*** is being proposed. HIASN has been analyzed on various vectors of public profiles attributes. The extracted profiles are later clustered using novel algorithm ***Hybrid Ensemble K-Means Hierarchical Agglomerative clustering (HEKHAC)*** which forms the Phase 2 of the proposed work. Phase 3 of the proposed work which helps the user to extract useful information is being described in depth in chapter 5.

Chapter 5 is based on the fact that the integrated user’s profile is available and contains useful information about the user like interest, location, connections etc. This determines the third phase of the work. **Qu*ery Processing in Social Networks Aggregator (QPSNA)*** which is developed in four modules which include ***Query Processing System (QPS), Content Based Semantic Matcher Maker (CBSMM), Machine Learning Mechanism (MLM) and Ranking of results.*** It starts with pre-processing of the query entered by the user to extract the keywords and providing the semantic meaning to the keywords. It further discusses the implementation and analysis to enhance the semantic meaning for providing the appropriate ontology and rule map to identify the right cluster for the retrieval of information. Context based semantic match maker has been proposed to enhance the semantic meaning of the extracted ontology which forms the basis of the QPSNA. This enhanced semantic information with machine learning mechanisms extracted the group of users using HEKHAC and later sorted the users as per the interest of the user’s free form queries, accounting the third and final phase of the proposed work.

Chapter 6 evaluates and analyzes the proposed work on various measurements and claims the accuracy of the proposed work. The results obtained have been analyzed and compared on parameters such as Precision, Recall, Receiver Operating Characteristic (ROC) graph, similarity measure and classifiers like Naive Bayes, Logistic Research and Support Vector Machine- Linear as well as Kernel with existing popular mechanisms. It is worth mentioning that the results obtained are competitive.

Chapter 7 concludes the outcome of the work. It summarizes the major achievements of the research work and elucidates the scope for future work in this domain.

1.7 CONCLUSION

The chapter began by presenting the motivation for carrying out the research work highlighting the potential of SNS. It is now understood that SNAs have a vital role to play in maintaining social data and various social activities from all of the social networking sites. The research issues pertaining to the aggregation of data across social networks like “Walled Garden” and “Information Overload” were detailed and an idea regarding the feasible solutions to overcome the constraints was also offered.

Summarizing the study presented, it can be concluded that a number of social media aggregators have revealed up in recent years, but social media services still require to be researched upon and hence implement more effective and efficient ways of aggregation. To the best of our knowledge, no such efficient example was found in the research that collects and excavates the data. Thus there is a need for an aggregator which can extract the profiles of a user existing at multiple social sites and provides information after combining the different profiles.

Next chapter presents an in-depth study of social networks and existing social network aggregators. It also considers highlighting the true as well as false promises made by existing SNAs.

CHAPTER 2

SOCIAL NETWORK AGREGATORS AND NATURAL LANGUAGE PROCESSING IN ONLINE SOCIAL NETWORKS: A PREFACE

2.1 ONLINE SOCIAL NETWORKS

The Advent of web 2.0 [48][49] resulted into the evolution of online social networks which in turn has become an integral and popular part of the modern Internet whose aim is to share and connect with people. The growth and evolution of social media have been in the world since the late 70s providing services like newsgroups, Internet Messengers (IMs), blogs and chat rooms. Moreover, the "Golden era" of social media started in early 20's that caught immediate attention of innovation like LiveJournal[14], encyclopedia, Wikipedia[15] that gave massive popularity among internet users all around the world. However, the huge boom of social media was followed by the emergence of LinkedIn in 2002, MySpace[16] in 2003, Facebook in 2004, and Twitter in 2006 [50]. Since then, it became an ever-demanded medium of communication having a larger user base and giving birth to user generated content which is growing exponentially over the years. It is a diverse and easily accessible platform serving as a source for building communities, sharing events of interest around the world, meeting the new acquaintance, getting updates, consume news and discuss various topics.

The social aspect introduced by OSN services caught immediate attention and made them immensely popular among internet users all around the world in a very short span of time. Today, large number of users around the world access to the Web and a large number of users have an account and uses services provided by OSN. For instance, Facebook (728 million) [51], 540 million on Google+ [52], 259 million on LinkedIn [53], and Twitter (over 200 million) [54] lead the way in terms of the number of monthly active users for a single OSN. A study by 11000 users in 2009

[14] https://www.livejournal.com/
[15] http://www.wikipedia.org/
[16] https://myspace.com/

exhibited that majority of LinkedIn users and Twitter users also have a Facebook account [12]. There is an expectation of the exponential increase in this overlap by 2018. Users share all kinds of information on social networks at an enormous rate. For example, 4.5 billion of likes are generated daily; every 60 seconds, users post 510 comments, update 293,000 statuses and upload 136,000 photos on Facebook [56]. Users often engage in different activities and reveal information about different aspects of their lives on different social networks.

On Facebook, users communicate with their friends and families and share facets of their personal lives. On LinkedIn, users give details about their professional evolution and aspirations. On Twitter, the users tend to post things they are passionate about. Such widespread reach and popularity make OSNs a powerful tool for communication, especially during national and international events of interest, like sports, natural calamities, political events, etc. Users around the world use OSNs as primary sources to assimilate news, updates, and information about events around the world. A majority of Twitter and Facebook users, for example, say that each of these platforms serves as a source for news about events and issues outside the domain of family and friends [39]. Some of the most popular social networks are categorized on the basis of their services in Table 2.1

Table 2.1 Popular Social Networks

Category	Example
Wiki	Wikipedia, Scholarpedia
Blogging	Blogger, LiveJournal
Social News	Digg, Mix
Micro Blogging	Twitter, Google Buzz
Opinion & Reviews	Yelp, ePinions
Question Answering	Yahoo! Answers
Media Sharing	YouTube, Flicker
Social Bookmarking	Delicious
Social Networking	Facebook, Twitter, LinkedIn

Considering the high volume and diversity of such information, it is difficult to track the profile information and events that users post on these OSNs, especially about the users across multiple social networks. This lack of control and inability to monitor information among multiple social networks enable hostile entities to exploit the techniques of aggregating user profiles and further, generate and promote various sorts of events over one platform. Such enormous information pertaining to user profiles and their posts pollute the information stream, making the aggregation of information a challenging task. The major factors that lead to the emergence of QPSSN proposed in this research are similar profile attributes and cross posting across the social networks.

Social Network Aggregators (SNAs) [8][9][10][11] aggregates the social information of users across many social networks. Owing to differences in the privacy policies (which in fact keep on evolving also) of all social networks, the existing SNAs fall short in various aspects such as resolving the identity of user i.e. ensuring that only the legitimate user profile is being integrated. The upcoming section throws light on the role of SNA in OSN and it also briefs about few popular SNAs.

2.2 SOCIAL NETWORK AGGREGATORS

Social Network Aggregators provides a wise solution to the problem of "walled gardens" of disconnected social networks [8][9][10][11]. It provides a platform where the social data of a user is collected, aggregated and organized from multiple social networks to streamline user's experience. Over 150 solutions for monitoring multiple accounts, out of which 30 are used for managing multiple accounts, were identified by Altimeter Group with an observation that there is no standard followed on their functionalities [58]. Indeed, SNA is required to manage multiple accounts well and efficiently. However, the diversity and lack of standards lead to a problem for the end users to determine which platform should be adopted. The major challenge in SNA is to provide a simulated appearance to the user so that user can access various social data, services, and activities without logging to each OSN and yet distinctly perform the same social activity. It is a three step process which consists of user's identification, the collection of data, and its representation. The first step will identify the user's multiple accounts followed by retrieving and representing the heterogeneous information available across social networks. OSN like Facebook or Twitter uses OAuth[17] framework, an open standard for authentication management

[17] http://oauth.net/

instead of OpenID[18] which raises the question of user's identification across multiple social networks as the user has to create a new public profile to use the services of the respective social networks.

To accomplish the first step, search engines like Peekyou[19] and Pipl[20] which allows tracing the footprints of the user's multiple accounts using the username, location or email might be helpful. However, the identification will be impossible if the user has specified different values or left the attributes blank. Social Graph API[21] provides an alternative of extracting links which referred to the other profiles of a user by crawling the user's Google profile information like URI account. Another possible solution is to implement authentication protocols by SNS and probe users to link their social identities.

User's social data can be collected by either crawling the user's profile with an automated script or using API's provided by SNS. Crawling and extracting user's social data will only be possible if it is allowed in the terms and conditions of SNS providers. A very small set of information is available for use using this methodology if permitted by SNS providers. The second mechanism necessitates registration of an application with suitable permissions to send relevant queries to SNS via API to collect publicly available data and much more than that. This technique seems to be a promising solution but it also suffers from two drawbacks. First, SNS providers can restrict a user to the minimum number of API calls. Secondly, the feature requires the user to learn a variety of API's as variable range of API's is available. To solve this problem, Google has developed openSocial as a hosting environment which has a set of common API's for web applications and supported by more than 80 social networks. GNIP[22], Datasift[23] or Topsy[24] are other commercially available solutions that provide real-time aggregated user's data from multiple SNS. The social networks can be visualized by its own structure and attribute to display user's activities. The major goal is to add a pragmatic approach which relies on descriptive language for

[18] http://opened.net/
[19] https://www.peekyou.com
[20] https://pipl.com/
[21] https://developers.facebook.com/docs/graph-api
[22] http://support.gnip.com/apis/
[23] https://datasift.com/
[24] http://topsy.com/

browsing using heuristic classification and semantic ontologies. The properties and representation used by the social network may vary across multiple SNS to elicit the crucial information through ontologies, folksonomies or taxonomies.

DBpedia[25] is a community effort to mine structured content from Wikipedia and to make this data presented over the Web. It is a revolutionary step for the interaction of the user which served as linked data that allows navigation on the web using browsers, automation of crawlers and posing queries. Semantically Interlinked Online Communities (SIOC)[26] ontology is developed by Bojars et al. [63] as depicted in Figure 2.1 for users, implicit friends and social contents to solve data interoperability for representing the information available on discussion platforms such as blogs, forums and mailing lists using RDFS[27], FOAF[28] such as foaf:maker property and RSS for describing post.

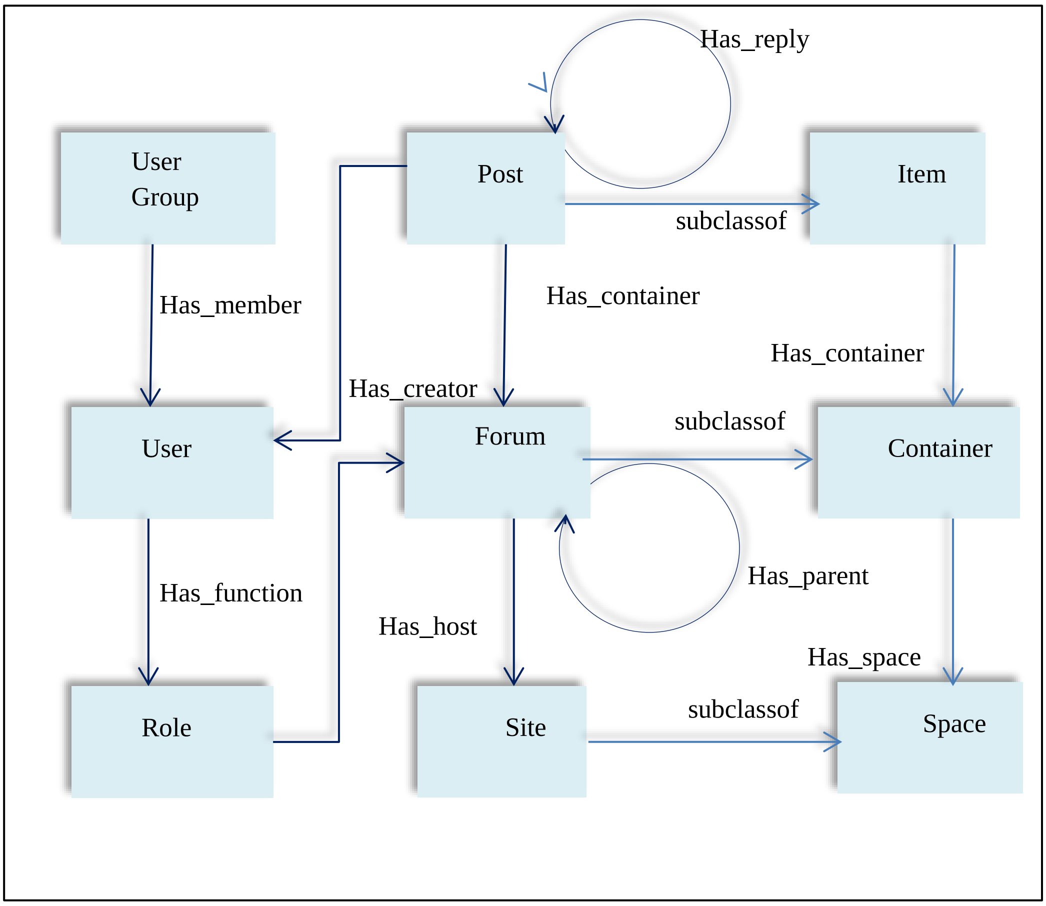


Figure 2.1 SIOC Ontology

[25] http://dbpedia.org/About
[26] http://www.sioc-project.org/
[27] http://rdfs.org/scot/spec/
[28] http://www.foaf-project.org

General User Model Ontology (GUMO) [59] is one of the known solutions which covers demographics of users like name, username or email but lacks in attributes like users interest. Semantic web community has provided the wide range of standards to represent the user and their activities. Many Researchers [33][36][60][61][62] have used these ontologies to represent semantic counterparts of the social data. Interlinked datasets is another representation used to enrich the semantic information that involves an extra effort which isolates entities from the text and links them to URIs like DBpedia, SIOC to name a few.

In addition to serving instantiation of semantic web vision, Friend Of A Friend (FOAF) is a descriptive vocabulary that makes up a significant space of all the data on semantic web that helps in describing users, their activities and relations to each other as depicted in figure 2.2. It eliminates the need of a centralized database and enables to describe OSN. It is used in OSN to find users profile by defining relations between people, it also helps in profile merging across OSN.

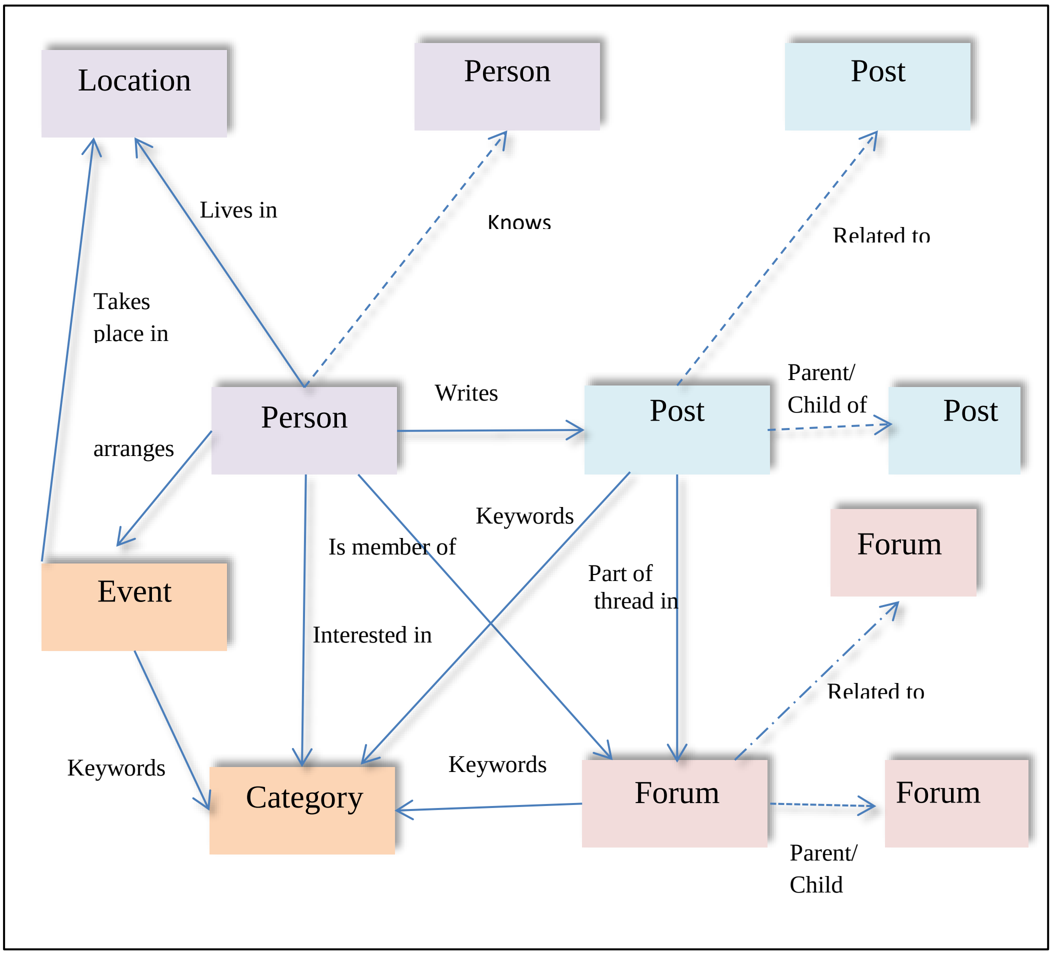


Figure 2.2 FOAF Ontology

There are other well-known proposed techniques to merge user's social information from multiple OSN. Some authors aggregated information into a unique entity whereas others inserted the provenance data into each user's interest using Open Provenance Model[29] (OPM) [36][46]. Brojas et al. [63] suggested utilizing two semantic properties owl:sameAs and rdfs:seeAlso to associate a user with existing profiles as shown in figure 2.3.

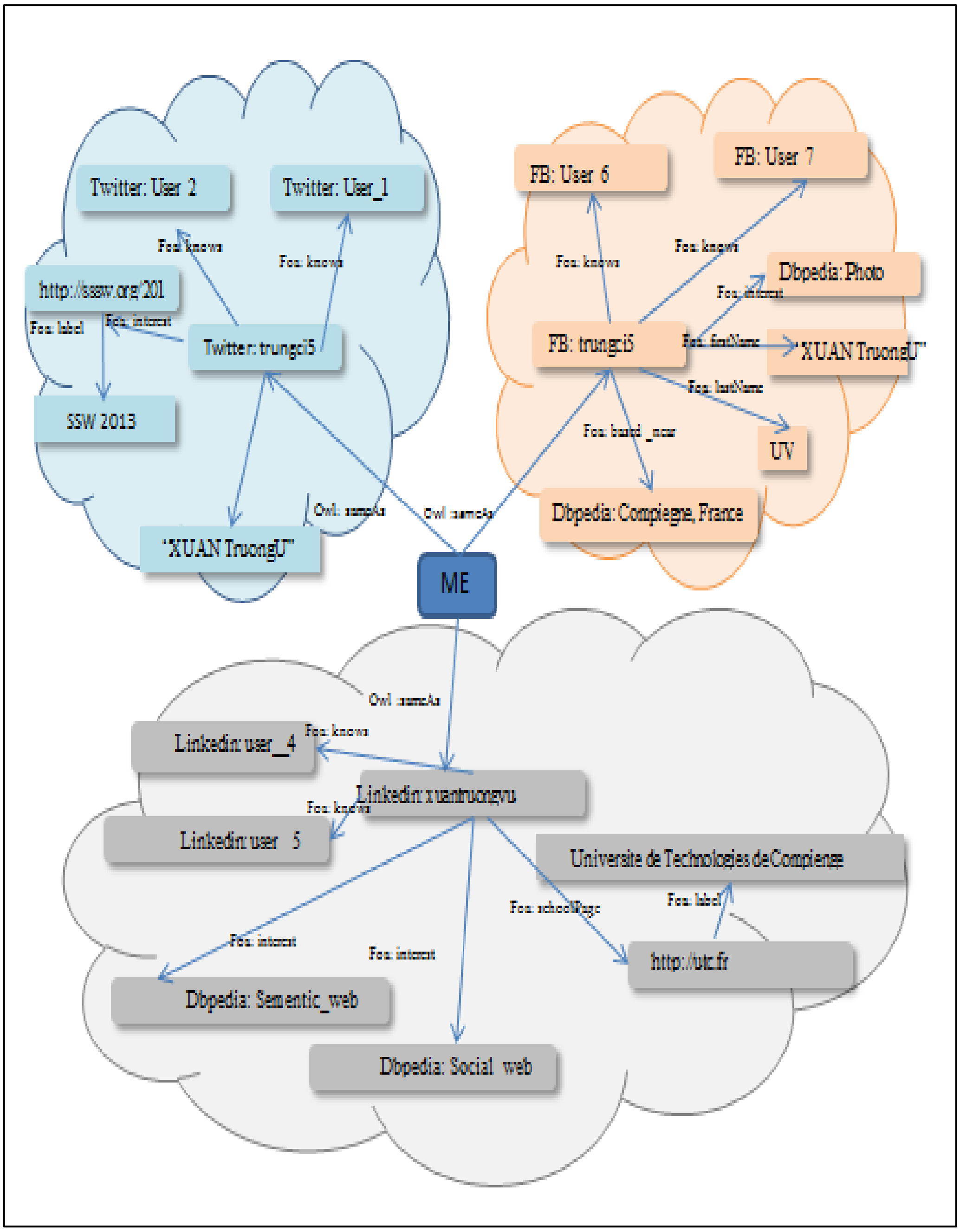


Figure 2.3 Aggregation of User's Profiles using OWL:sameAs

[29] http://openprovenance.org/

Another OSN Aggregator proposed by Zhang et al. [8] not only pulls the social information from multiple networks but also group, rate and notifies about the activities of friends. However, the system failed to integrate the networks. In fact, numerous models have been advanced to outline a collective objective model for assimilating a user [8][10][11][43]. Abel et al. [33] aggregated user profiles on the limited set of properties like name, photos etc. using the most popular solution FOAF from online social networks by applying rules. Another possible implementation of aggregating user profiles is by using Activity Stream protocol that consists of an actor, a verb, an object, an optional target and syndicates the activities of users across OSN. Table 2.2 lists known vocabulary to represent social data that is widely used in web 2.0 for providing interactive medium to users.

Table 2.2 Summary of Standards of Social Web

Ontology	Description
FOAF	Friend Of A Friend ontology is used to represent people, their relationships and activities.
Relationship[30]	Relationship ontology is used for specifying the type of relationship between people.
DOAC[31]	Description of a Character ontology represents the working experience and cultural background
GeoNames[32]	GeoNames for providing geospatial location of the user.
SIOC	Semantically Interlinked type of Communities is used for representing blogs and forums of the user.
DBpedia	DBpedia is a community effort to extract structured information from Wikipedia.
SCOT[33]	Social Semantic Cloud of Tags is used for representing tags
WI[34]	Weighted Interests Vocabulary is used for representing user interest.
OPM	Open Provenance Models ontology represents interest from specific website
GUMO	GUMO is an OWL based ontology to describe user’s demographics
XFN	XFN is XHTML Friend’s Network to describe user’s relations on the web.
Media RSS	MediaRSS is a RSS based schema to represent rich media like images, video s etc.
Activity Streams[35]	Atom based standard format to describe social activities of a user.

[30] http://vocab.org/relationship/
[31] http://ramonantonio.net/doac/1.0/
[32] http://www.geonames.org/ontology
[33] http://rdfs.org/scot/spec/
[34] http://purl.org/ontology/wi/core#
[35] http://www.activitystrea.ms/

The above listed standards have already been adopted by OSN and other IT companies such as the activity stream have been incorporated by Facebook and MySpace. Thus, the standards or ontologies allow users to represent the framework of the social data which has pre-defined sets of vocabularies to describe different types of social contexts. However, it is hard for a user to select appropriate value as every aspect of social data has too many possible values corresponding to too many dimensions. However, the syntax differences may still exist across SNS and require translation to aggregate user profiles across OSN. The services which allow the user to consolidate services of multiple OSN aggregates information of users across social network and provides the same experience as of social network are called social network aggregators. Figure 2.4 depicts few of the SNAs and each one of these is being discussed as follows:

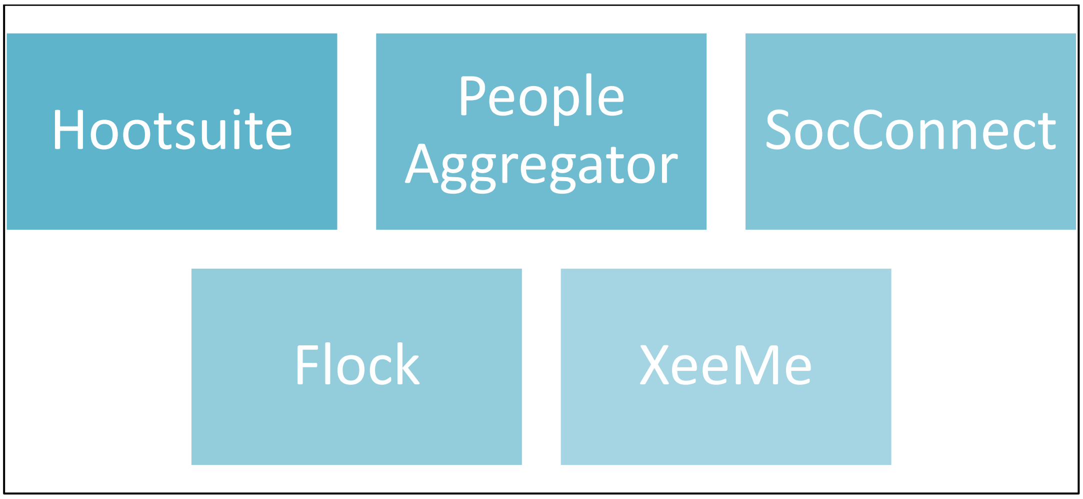
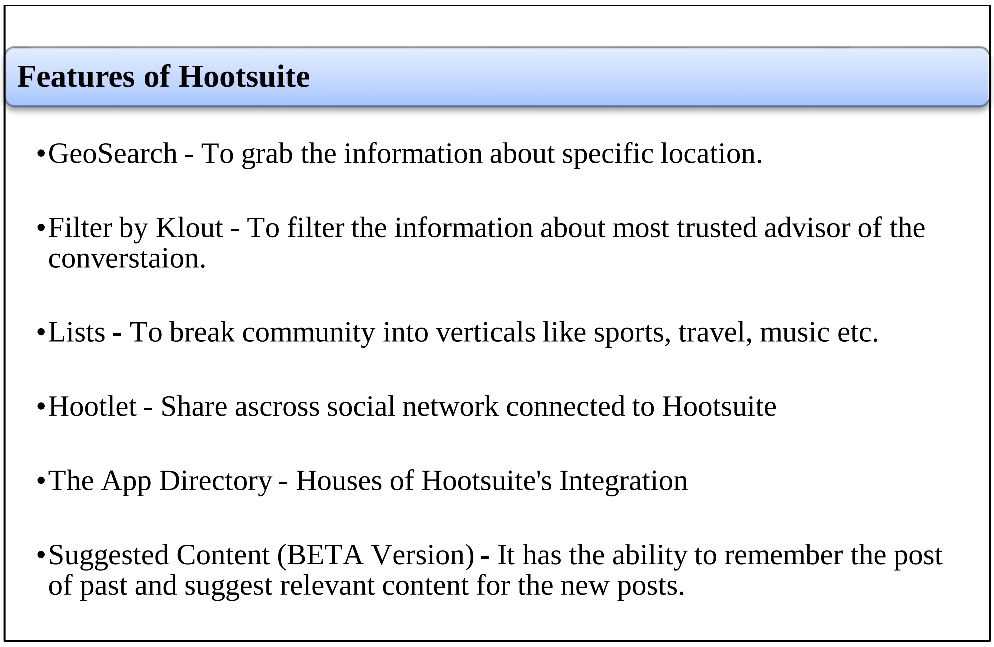


Figure 2.4 Social Network Aggregators

A) Hootsuite

Hootsuite is a robust tool for providing an experience of web dashboard to the business companies to improve upon the marketing promotions, identifying the new users and their needs to dispense target messages by applying various social networking strategies. It enables users to manage, bookmark, handhelds integration, RSS feed and publish the post to multiple social media accounts at one place. With Hootsuite, users can post updates, connect with the client base, and review responses on more than thirty-five popular social networks such as Foursquare, Facebook (including Events, Groups, Profiles, and Fan Pages), Twitter, LinkedIn (including Pages, Profiles, and Groups) etc. Hootsuite finds space in a business organization due to monitoring and listening as the key components of its success strategy. Though both the keywords are used interchangeably but have different space in the world of social media. Monitoring is an observance of the ongoing conversation about some brand across OSN, while listening is about ruling out the new opportunities to seam conversations that may not be related to your brand or products. Being able to monitor multiple channels using search by location or language from one dashboard eliminated the need for switching between web browsers.

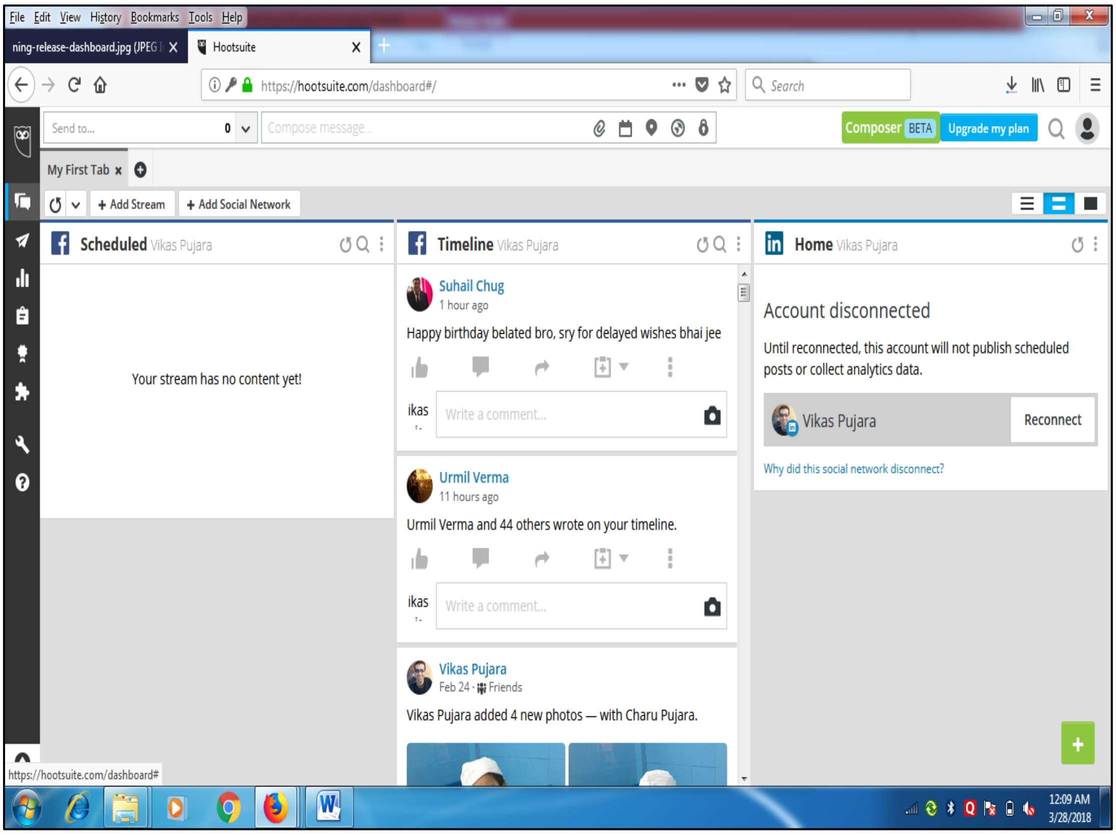
Additionally, the user can also gauge the sentiment, schedule their posts up to 350 posts at once to balance social messaging, carefully curate third-party content, and anticipate both seasonal and release-specific messaging. One single click from the Hootsuite dashboard helps users to respond to messages, mentions, and comments across OSN. A user can also create, import, track interaction history and share lists of those people whose engagements are more important or can affect them. Hootsuite has a rich content library and sources that can store and organize their assets like images and message templates across multiple channels with content storage solution in which storage solutions can be integrated such as Google Drive, Dropbox, Microsoft OneDrive to name a few. Major features of Hootsuite are summarized in figure 2.5 and figure 2.6 represents screenshot of Hootsuite.



Features of Hootsuite

- GeoSearch - To grab the information about specific location.
- Filter by Klout - To filter the information about most trusted advisor of the converstaion.
- Lists - To break community into verticals like sports, travel, music etc.
- Hootlet - Share ascross social network connected to Hootsuite
- The App Directory - Houses of Hootsuite's Integration
- Suggested Content (BETA Version) - It has the ability to remember the post of past and suggest relevant content for the new posts.

Figure 2.5 Features of Hootsuite



Your stream has no content yet!

Happy birthday belated bro, sry for delayed wishes bhai jee

Urmil Verma and 44 others wrote on your timeline.

Vikas Pujara added 4 new photos — with Charu Pujara.

Account disconnected

Until reconnected, this account will not publish scheduled posts or collect analytics data.

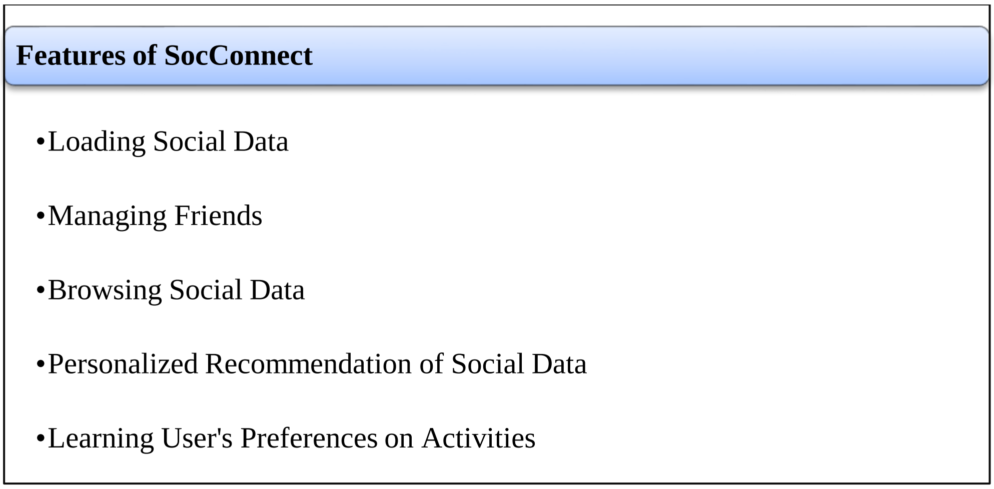
Why did this social network disconnect?

Figure 2.6 Hootsuite Screenshot

B) SocConnect

SocConnect [8] is a dashboard that communicates with the server that processes data and generates recommendations. It uses personalized suggestions and semantic contexts for aggregating the social data across multiple social networking sites. It models an intelligent system to manage friends, rate friends, activities and personalized suggestion of friend's activities using machine learning techniques. It learns the preference of the user and suggests new unread information to a user on the basis of their historic preference. Users can emerge as a new group of friends by combining friends from various social networks. Users can unify different accounts of a friend across SNS and create a single blended account for this friend.

SocConnect allows users to define their individual perspective of social data aggregated from different social networks that may indicate their presence for each environment. It provides content-based recommendations for social updates in social network services by incorporating rating of activities as "favorite", "neutral" or "disliked". The user's ratings of their friends are also used in predicting user’s interest in activities posted by their friends. Users can add a tag to enrich the context of the description to their friends and groups that later display the information related to the tag. It provides search on the basis of these tags to view the activities of few friends for whom the user doesn’t want to create a separate group and browse social data on the basis of groups as well. To represent the heterogeneous social context, it implemented a unified ontology for interlinking based on the combination of FOAF and activity stream using URI information of each user depending upon OSN. It uses authentication methods and API’s provided by multiple SNS to extract user’s information and their activities. Figure 2.7 represents the major features of SocConnect.



Features of SocConnect

- Loading Social Data
- Managing Friends
- Browsing Social Data
- Personalized Recommendation of Social Data
- Learning User's Preferences on Activities

Figure 2.7 Features of SocConnect

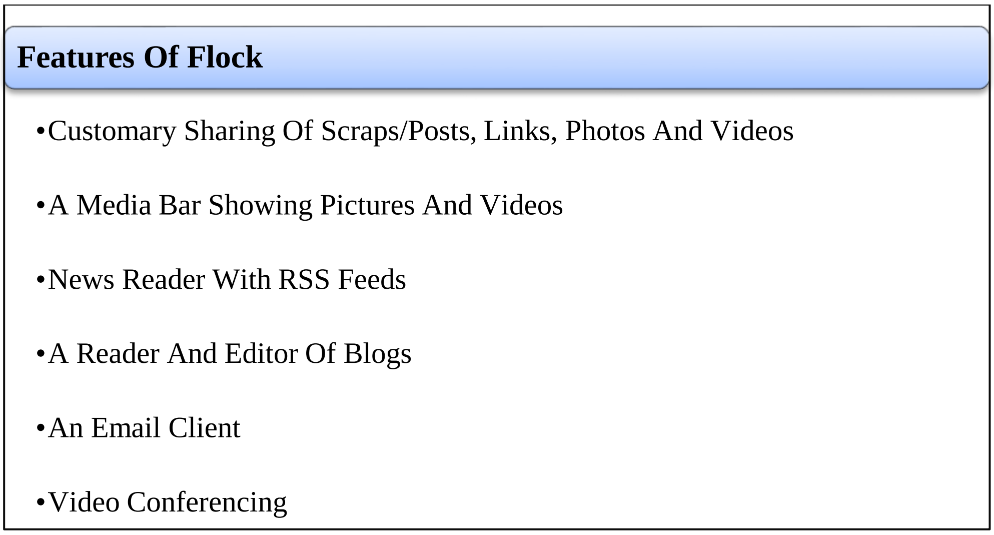
Based on extensive estimation, it provides a set of user preferences that can provide the best performance on the personalized recommendation. It has applied machine learning techniques for learning and prediction like Decision Trees, Support vector Machine, Bayesian Networks, and Radial Basis Function. However, the system did not fairly well in integrating multiple social networks.

C) Flock

Flock[36] is an online collaboration and communication platform tool that provides management tools for social networking and other Web 2.0 services. It provides a personal experience of the web by integrating the status updates and other social data like photos, friend’s update etc. from other popular social networking sites

[36] https://flock.com/

like Facebook, MySpace, Twitter etc. In addition, Flock can also search on Twitter to update multiple services at once and also uses Facebook chat service from the browser. It runs polls for the feedback and decision making, provides sharing of rich notes and automatic updates from 40+ tools, sends reminders and thus helps in taking decision faster. Figure 2.8 enlists few of the important features offered by Flock.



Features Of Flock

- Customary Sharing Of Scraps/Posts, Links, Photos And Videos
- A Media Bar Showing Pictures And Videos
- News Reader With RSS Feeds
- A Reader And Editor Of Blogs
- An Email Client
- Video Conferencing

Figure 2.8 Features of Flock

Flock doesn’t require us to provide authentication to any other site that maintains online security. However, the system was discontinued form April 2011.

D) People Aggregator

People Aggregator[37] is a service based social network aggregator. It amalgamates distributed profiles of users spread over several social networks and provides a centralized service to manage all content like blogs, media galleries, forums etc. It also disambiguates the identity of the user using “connective tissue” between the profiles to provide a marriage of different profiles into one unique profile. It also provides a unique summary of the user credentials to the recruiter which is a more trustworthy source of information than the conventional networks. People Aggregator access photos via API’s provided and establish the connection between two systems. Figure 2.9 throws light on some of the major features of People Aggregator.

[37] https://recruitee.com/hiring-resources/recruitment-dictionary/what-is-people-aggregator/

Features Of People Aggregator

- Customary Sharing Of Scraps/Posts, Links, Photos
- Disambiguation Of User Profiles To Unify Disparate Profiles.
- Thumbnail
- Connection Path
- Resume Databases
- Talent Pools

Figure 2.9 Features of People Aggregator

People Aggregator offers enterprise blogging to connect, create and collaborate on project management, helpdesks, searching and extracting experts in the specialized domains. It allows active talent sourcing by using resume databases with the searchable resumes packaged with the job brands. It transfers the information from one site to another for organizing or preparing data reports. It is used especially for streamlining the hiring process and building talent pools. It provides the connection path to map all the connection between the candidate and employer. It also provides thumbnail dossier of a user that consists of summary including skills, demographics, education etc. It ensures a consistent and coherent social presence across SNS and extracts suitable candidates that match specific resourcing requirements, returning results much faster than a manual consecutive search from multiple SNS using indexing of keywords like programming languages, job skills etc. However, the system was focused and implemented for recruiters only.

E) XeeMe

XeeMe [7] lets users/brands manage their entire social identity, identify new networks and people and nurture their presence as well. It helps increasing connections, raising popularity and strengthen relationships across SNS It has a long number of supported networks and it offers useful analytics. With Telegraph the user has a point of reference about his presence value and network relevance.

It offers the user the possibility of organizing all social networks at a single point and shares their social presence with one URL with friends, customers, partners, and people. Through the application, the user can discover new networks or people who are in other networks and offer the possibility of connecting with them. By sharing the URL on each post, the number of visits to the social site of the user increases. XeeMe offers a unique social address book, Social Media Time and Relation Management as listed in figure 2.10.

Features Of Xeeme/ Appearo

- Xeegraph
- Social Media Time Management
- Social Media Relation Management
- Social Address Book
- Style Sets
- Customize Profile Pages
- Email And Comment Signature

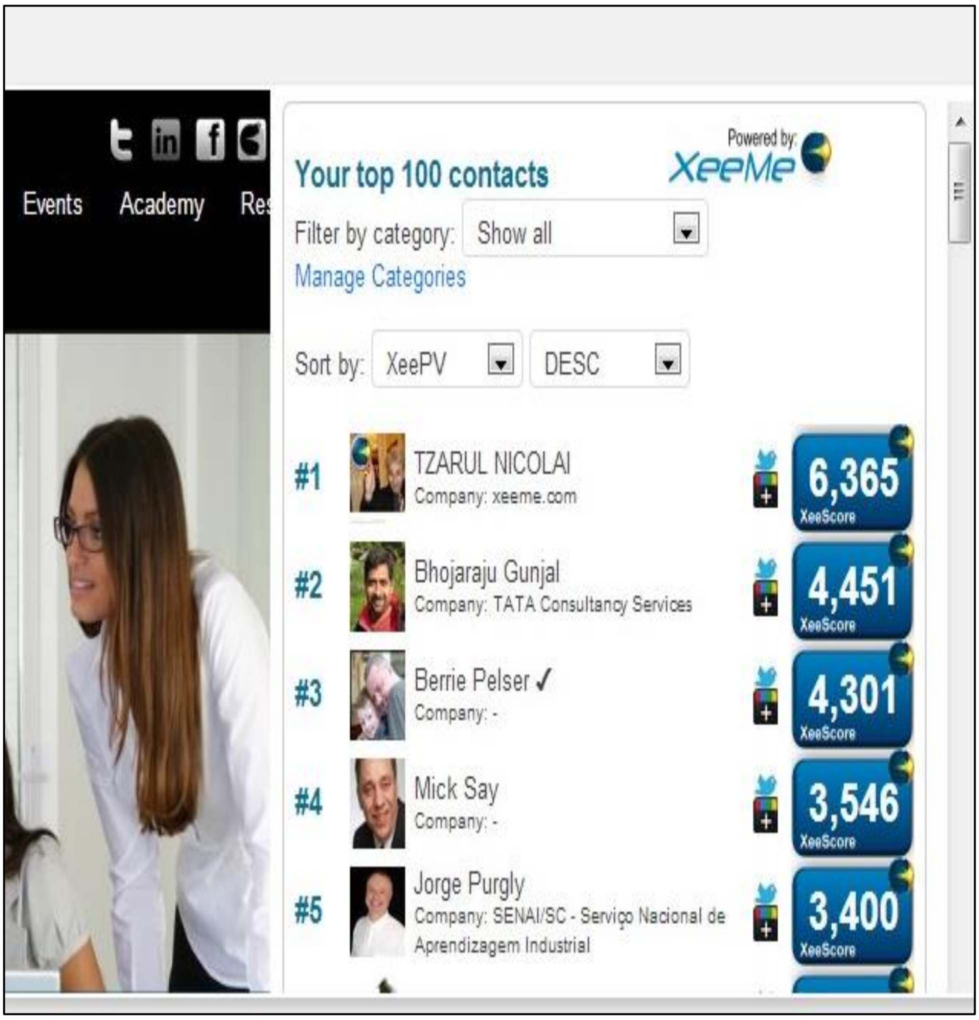
Figure 2.10 Features of XeeMe

It is now known as Appearoo[38] and provides the analysis to know the available connections that can be more valuable to grow your network easily by analyzing the number of visitors who visited your profile. It has added more security features to help users to build more trust in reaching out other networks. It provides one email signature as well as comment signature to reach all other connection over the SNS and can drive followers, following, referrals and traffic to your online appearance.

XeeMe has the ability to provide customize tab names, customize profile pages and different style sets for better UI experience. If a user visits profile page then it triggers multiple visits to the SNS. It represents XeeScore that depicts the social

[38] http://appearoo.com/yourname

presence value of the user by calculating inbound links and also helps the user to track the social data over 200 SNS as depicted in figure 2.11.



Your top 100 contacts

Filter by category: Show all

Manage Categories

Sort by: XeePV DESC

#1	TZARUL NICOLAI Company: xeeme.com	6,365 XeeScore
#2	Bhojaraju Gunjal Company: TATA Consultancy Services	4,451 XeeScore
#3	Berrie Pelser ✓ Company: -	4,301 XeeScore
#4	Mick Say Company: -	3,546 XeeScore
#5	Jorge Purgly Company: SENAI/SC - Serviço Nacional de Aprendizagem Industrial	3,400 XeeScore

Figure 2.11 XeeMe Contact Manager

It provides a one stop platform to organize the profile into multiple verticals to distinguish between Social Media Platforms versus Content Sharing Sites and Communication sites. It also generates reports for statistics of incoming traffic and thus providing valuable insight to grow the network and a platform to collaborate with multiple SNS.

While Flock gets refreshes from companions, notices, and photographs submitted at multiple social systems, SocConnect users make customized social and semantic settings for their social information. Users can join and group companions and rate network. On the other hand, Hootsuite entireties associations and organizations to cooperatively execute advancements over various informal organizations and XeeMe sorts out social nearness, finds new system and individuals. In fact, it sorts out the whole social nearness of the client, decides new systems and individuals and builds up their nearness and impact. People Aggregator bridges the gap between recruiter and job by unifying user's profile and indexing using keywords. As reflected by comparative view presented in Table 2.3, none of the aggregators have considered mining the numerous interpersonal organizations and extracting some helpful data subsequent to gathering information from various social networks.

Table 2.3 Comparison of different Social Networks

Network → Features ↓	ScoConnect	Flock	XeeMe	Hootsuite	People Aggregator
Analytics	No	No	Yes	Yes	No
Scheduling	No	No	Yes	Yes	Yes
Team Collaboration	Yes	Yes	Yes	Yes	No
Group Friends	Yes	No	No	No	No
Rating of Activities and Friends	Yes	No	No	No	No
Social Networks	Facebook LinkedIn Twitter	Facebook LinkedIn Twitter	Facebook LinkedIn Twitter	Facebook LinkedIn Twitter	Facebook LinkedIn Twitter
Analysis and Extraction of Social Data	No	No	No	No	No

All the social network aggregators can integrate Facebook, LinkedIn and Twitter account but to the best of our knowledge and understanding none of these have tried to assimilate the user's publicly available information within multiple social networks and in addition, none of the network aggregators can handle the query written in a natural language. The upcoming section presents a brief introduction to natural language processing and also explores the feasibility of employing NLP in social network.

2.3 NATURAL LANGUAGE PROCESSING

The need of extracting meaningful information and discovering knowledge from the huge amount of unstructured data on the web especially OSN elevated exponentially with the evolution of web 2.0 which allows interaction of users using text, images, videos etc. Working on the unstructured data requires a better understanding of data where NLP techniques find a better space [65][66][67][68] in order to discover patterns, unlike the traditional methodology that focuses on providing information access only. NLP is concerned with extracting the structure and meaning of free form language by computer using Artificial Intelligence Methodology [68][69].

There has been excessive use of NLP and web mining techniques to study Social Network. The main characteristic of data in OSN is its sparsity and huge dimension i.e. the valuable information of huge data could be represented by a bag of words whose semantic representation can help in analysis and mining [68][70][71][72]. NLP techniques map human language to machine language i.e. models the way user requests information so that computer or software understands it. However, simply searching for keywords is not an appropriate method in social network communication. Therefore, one can observe that the challenge in social network monitoring is to extract and interpret user communication.

Social networks are highly dynamic objects which grow and change rapidly over time. This evolving mechanism forms the motivation of the proposed work. Few social network analysis systems [65][68][73][74] also uses NLP methods with statistical techniques to ensure the extracted information to be correct and precise. NLP techniques are used in OSN to measure market sentiment and news data such as used for trading. For instance, Bollen et al. [75] measured sentiment of random sample of Twitter data, finding that Dow Jones Industrial Average (DJIA) prices are correlated with the Twitter sentiment 2–3 days earlier with 87.6 percent accuracy. The author predicted prices of individual NASDAQ stocks for forecasting prices 15 min in the future by training Support Vector Regression (SVR) model on the Twitter data [97]. Other applications include the following but are not limited to:

- Monitoring and analyzing responses of users to announcements
- Speeches and events especially of some celebrity or political comments
- Insights into the behavioral aspect of the community
- Early detection of embryonic events such as with Twitter.

As an illustration, Karabulut [76] found that Facebook's Gross National Happiness (GNH) exhibits peaks and troughs in-line with major public events in the United States. Lerman et al. [77] automatically predicted the influence of news on the public perception of political candidates. Lastly, Yessenov and Misailovic [78] stated various approaches to extract textual features by applying machine learning algorithm on the movie review comments such as Naive Bayes, Decision Trees, Maximum Entropy and k-Means clustering.

News Analytics, Opinion Mining, Sentiment Analysis and Text Analytics are the common techniques related to analyzing noisy and unstructured social textual data. Cleansing the social textual data poses numerous research problems and challenges [67][72]. Languages, slangs, misspelled words and the disruption of language faces myriad of challenges for analysis of social data. Figure 2.12 depicts few of the prominent NLP techniques which are being described in the upcoming subsections. Furthermore, Text mining issues and solutions are discussed to extract the useful information using NLP techniques in OSN.

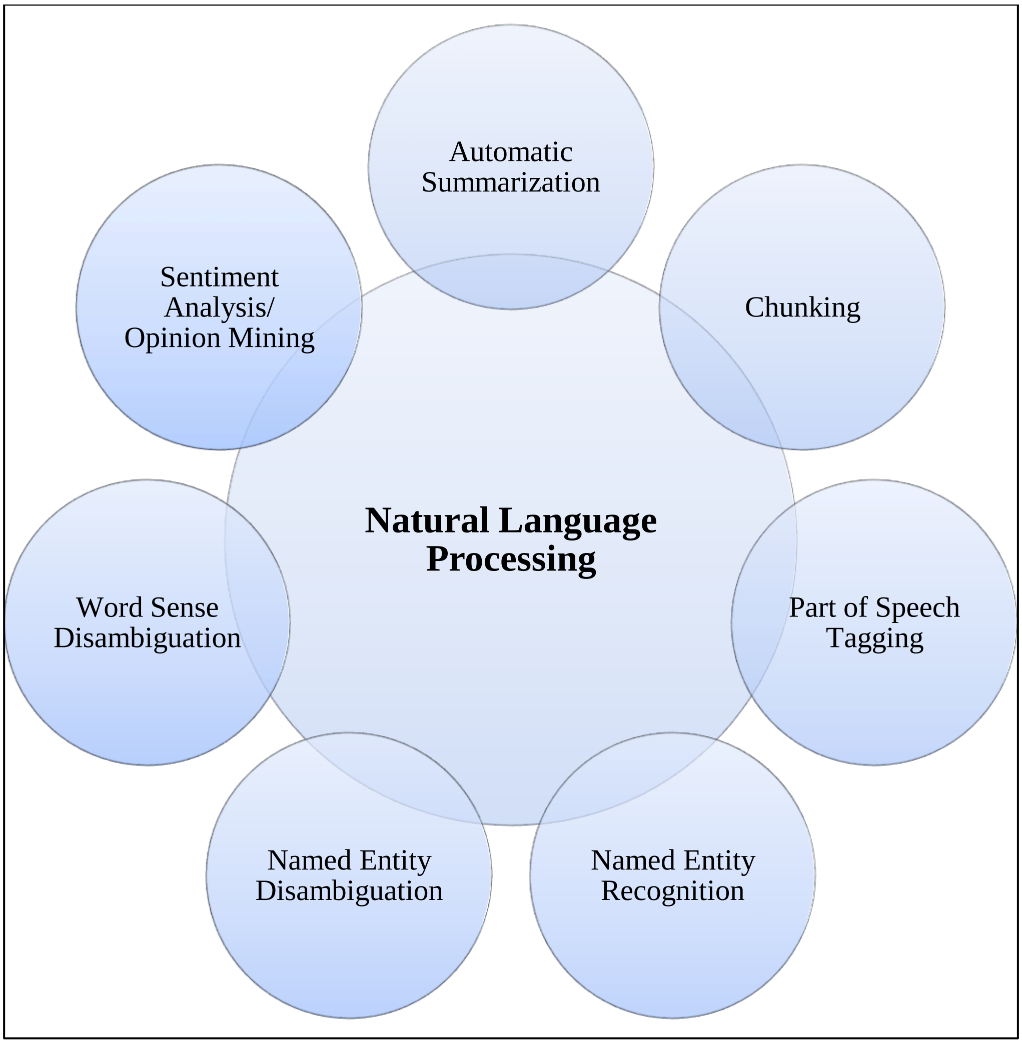

Automatic Summarization
Chunking
Sentiment Analysis/ Opinion Mining
Natural Language Processing
Word Sense Disambiguation
Part of Speech Tagging
Named Entity Disambiguation
Named Entity Recognition

Figure 2.12 NLP Techniques

Given the phenomenal data involved, analysis and visualization of social data is becoming increasingly important. To address these challenges posed by OSN, this section explores NLP approaches essential for analysis to carry out the semantic analysis of the unstructured content available on OSN.

A) Automatic Summarization

Automatic Summarization [79] is the process of reducing a text document with the help of a computer program such that it creates a summary retaining the most significant points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style, and syntax.

The main notion of summarization is to find a representative subset of the data, which contains the information of the entire set. The summarization bounces the solution to the problem of response generation that generates an automatic response to respond to OSN posts like Twitter status posts.

During crisis events, text summarization can be considered in an incremental and temporal manner. Incremental text summarization refers to generating a summary given: (1) the set of documents to be summarized, and (2) a reference set of documents which the user has read. The objective of the system is to produce a summary only of data the user has not already read [79]. Temporal text summarization refers to creating an extractive summary from a set of time-stamped documents, usually in retrospect [79][70].

Automatic summarization is categorized on extraction approach and abstraction approach [79]. While extraction refers to selecting a subset of existing words, phrases, or sentences in the original text to form the summary whereas abstraction builds an internal semantic representation and uses natural language generation techniques to create a summary that is closer to what a human might generate. The Automatic summarization is depicted in figure 2.13 and follows three basic steps:

- Analysis
- Transformation
- Realization

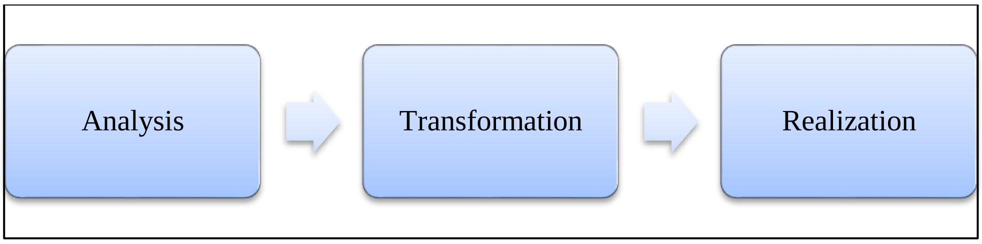


Figure 2.13 Phases of Automatic Summarization

In the analysis phase, a concise and fluent summary of the most significant information is produced in the input. It requires the capability to reorganize, modify and merge information expressed in different sentences in the input. Transformation pertains to the generation of an ordered text by manipulating the internal representation post analysis phase. At last, realization phase deals with generating an analyzed summary of text using scores of transformation.

Automatic Summarization technique is beneficial in building “chatbots” [80] for entertainment or companionship in social media and deals with the “information overload” problem that involves presenting users with a text representation of the upcoming events. It provides immediate assistance, human like engagement and is efficient in terms of service, moreover, save cost and time. It helps to find the precise answer to the customer needs without human intervention and offers a personalized one to one experience which enables the system to achieve real time interaction [79][70][80].

Automatic summarization framework is also exploited in Twitter that generates a phenomenal volume of information for most real-world events on regular basis to generate the coherent and concise summary of the events from an unfiltered twitter stream [81].

B) Chunking

Chunking is to identify the chunks from the words and their morphological syntactic class. The main goal is to divide the sentence into non-overlap syntactic units. Chunking [82] is the basic technique used for entity detection in OSN. Chunking selects a subset of the tokens rather than tokenization that omits whitespaces. The data source is monitored and occurrence of the events is detected from the selected source. It can be used for detecting real-time events like deaths due to Blue Whale Game on Twitter [83]. This is accomplished by training the classifier with messages, features to identify positive or negative attributes and applying a probabilistic model to search the user/location of the event [70] or analyzing the relation between user and tags on photos to detect event [73]. It can also be considered as tagging task [84]. The main task in chunking is to search noun phrase (NP) and identifying arbitrary chunks. Base NP chunks play an important role in knowledge discovery and question answering [85]. The Penn Treebank parser [87] annotates naturally occurring text for linguistic structure producing skeletal parses depicting rough syntactic and semantic information using bracketing style which enables it to extract simple predicate/argument structure with a set of over one million words of text. Simple and non-recursive NP methods [70] helps in recognizing base NP chunks whereas prepositional phrases, adverb phrases, adjective phrases and verb phrases are other promising methods to detect other types of the chunk. Rule based learning [88], Transformation based learning [89], Hidden Markov Models (HMM) [90], Memory based learning [92], Maximum entropy [93], Support Vector Machine (SVM) [94] are the known techniques for uniting linguistic information with chunk detection for dealing with various text data.

C) Parts-of-Speech Tagging

Parts-of-Speech (POS) Tagging [86] is a piece of software that reads the text in some language and assigns parts of speech to each word such as noun, verb, adjective to name a few. It is the fundamental step in NLP pipeline which identifies the role of a token in the sentence [70]. Generally, computational applications utilize more fine-grained Parts of speech tagging including tags like 'noun-plural'. Dictionaries have category or categories of a particular word which implies that a word may belong to more than one category [73]. For example, ‘Run’ is both a noun and verb. Taggers employ ‘Probabilistic Information’ [75] to solve this ambiguity. It is often used in machine learning techniques as a feature for further classification and popularly used in text preprocessing pipelines. POS taggers can be broadly classified as rule based and statistical based. POS taggers models are mostly implemented using statistical methods like HMM [90], SVM [94], Graph based [72] and perceptron based [73] training to a generalization of data. The major challenges that are faced in the quality of learning and the performance of end system are the corpus size, unknown words, lack of context, quality of corpus, traceability and tractability in machine learning algorithms. Data-driven POS tagging has been benefitted a lot from machine learning techniques as these are language and tag set dependent that makes its applicability easy to new language and domains [68]. Table 2.4 illustrates an example of POS tagger used from Penn Treebank parser.

Table 2.4 Example of POS tagger

Input of POS tagger	**Ram is 9 years old**
Output of POS tagger	Ram_NNP is_VBZ 17_CD years_NNS old_JJ
List of POS Tags used in above example	
NNP	Proper Noun, Singular
VBZ	Verb, 3rd person singular present
CD	Cardinal Number
NNS	Noun, plural
JJ	Adjective

The average accuracy of the state of the art methods is 96% contingent on the corpus and language. However, as it is evaluated on keyword basis that turns out on an average one tagging error per sentence. Even though it is limited, the information provided by the tagger is quite useful [72].

D) Named Entity Recognition (NER)

NER [66] is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. For an instance, Robert bought 500 shares of Accenture Corporation in 2008. In this, a person name consisting of one token, a two-token company name, and a temporal expression have been detected and classified. Hand-crafted grammar-based systems typically obtain better precision. Currently, statistical models are preferred as this approach initially uses training data against the model, followed by preparation of statistics [68]. These statistics are then used against real documents. NER is also offered as a solution to NLP problems in various organizations like Stanford University [98]. Moreover, it is also employed in libraries and Java platform to identify names and Entities. For Example, the newsfeed “enjoying U.S. weather at Texas with Monalisa” will extract entities like weather, Texas and Monalisa.

Rule based approach [95] and statistical learning approach [86] were two main approaches used in NER. Rule based learning represents rules like Ms. + Capital letter -> Female Name, these rules should be catalyzed first and then machine learning rules can be applied. NER methods have 90% accuracy on an average in longer texts whereas 40% on an average in short tweets/posts. This is mainly because of the shortness of length (maximum of 140 characters/tweets) makes it hard to interpret. Ambiguous text, low amount of discourse information, language variation, emoticons, abbreviations, and hashtags makes entity extraction a challenging task. A plethora of hybrid techniques [70][73][72] has been proposed and implemented in the past to overcome these challenges and extract information from text.

E) Named Entity Disambiguation (NED)

The task of linking the identity of entities available in the text is referred as Named Entity Disambiguation (NED) [23]. However, it is distinctive from named entity extraction as it identifies not the occurrence of names but their reference. It needs a Knowledge Base of entities to which names can be linked. OSN user profiles can be used for disambiguating entities mentioned in the user generated content, activities on web and interaction among entities. It can be used for constructing a part of user personality. Microposts like tweets in case of Twitter are short texts posted by the user which contains contextual information that can be extracted using disambiguation techniques [24].

Once named entities have been identified in a text, we can then extract the relations or facts that exist between specified types of named entity. The objective of the fact extraction is to detect and distinguish the semantic relations between entities in text or relations and fill it in a predefined template using the entities. The relations can be categorized as physical, social (family relation) or employment/affiliation relation.

F) Word Sense Disambiguation

This is an open NLP and ontology subject that identifies the correct sense of the word in a sentence where multiple meanings of the word exist. It’s easy for a human to understand the significance of a word based on the basis of its background knowledge of the subject. However, identifying the aspect of the word is difficult for a machine to understand. This methodology provides a mechanism to diminish the ambiguities of words in the text [72][73]. For example Word Net is a free lexical database in English that contains a large collection of words and senses.

G) Sentiment Analysis/ Opinion Mining

Sentiment Analysis/Opinion Mining is an NLP process which identifies, extracts, enumerates the attitude, opinion and emotions of the user towards a user, events, topics, and products. Sentiment Analysis or Opinion Mining can be used interchangeably as sentiment analysis is analyzing the sentiments expressed after identifying it from the text whereas opinion mining is extracting and analyzing user’s opinion about an entity [75][78]. Thus, sentiment analysis is ruling out opinions, categorizing sentiments and classifying their polarity.

Sentiment Analysis is extensively used in processing survey form, online reviews, and social media monitoring. It returns the identified sentiment with a numeric score from 1.0 to -1.0 where 1.0 means strongly positive and -1.0 means strongly [72][92]. For example, “I love it” with score 0.8 means a strongly positive analysis for the newsfeed or blog. A practical application of this can be in a typical e-commerce website. Famous or ‘Top Rated” products are likely to attract thousands of reviews and this may make it challenging for prospective buyers to track relevant reviews that may assist in making the decision. Sellers use sentiment analysis to decide relevant review and ignore the misleading reviews present to reviewers. A 5-star scale rating with five signifying best rated while one signifies poor rating.

The upcoming section discusses text mining which is an application of NLP as it lays the foundation for the proposed work to extract useful information from OSN.

2.4 TEXT MINING

NLP is an attempt to extract meaningful information from free text and text mining received attention due to its wide application in information retrieval, machine learning etc. Among the different data formats available in OSN, text plays an important role. For example, tweets, blogs, hashtags to name a few use textual data for representation and to extract information from the text provides a great opportunity for researchers to research.

Information is gathered from large scale databases with the help of traditional data mining commonly known as warehouses. Then this data mining aids in extracting information automatically discovers and extracts information from unstructured text documents and services. Searching with the help of text mining is a way of retrieving and searching on a social search engine that mainly searches user-generated content such as news, videos, and images related search queries on social media like Facebook, Twitter etc. Some applications of text mining in social Network are keyword search, Classification, Clustering, Linkage based Cross domain learning [72][85][93].

Keyword search identifies the social network nodes using a set of keywords which are close to the query result. Content and Linkage behavior plays an important role in order to determine the query output. It provides an effective method for accessing structured data. Query Semantics, Ranking Strategy, and Query Efficiency are the major concerns to perform the keyword search on social networks [8][64]. The nodes in the social network are associated with labels which are used for classifying the network. There are numerous algorithms available for classification of text from the content [65][68][69][71].

Clustering is the area where a set of nodes are used to determine the similar content for the evolution of clusters. Linkage of clusters is an important factor and when combined with content can classify the network which results provide better clusters.

The linkage information between multiple domains of social networks provides transfer of knowledge across various kinds of links. The text mining approach consists of three steps, which include Text Pre-processing, Text Representation and Knowledge Discovery as shown in figure 2.14.

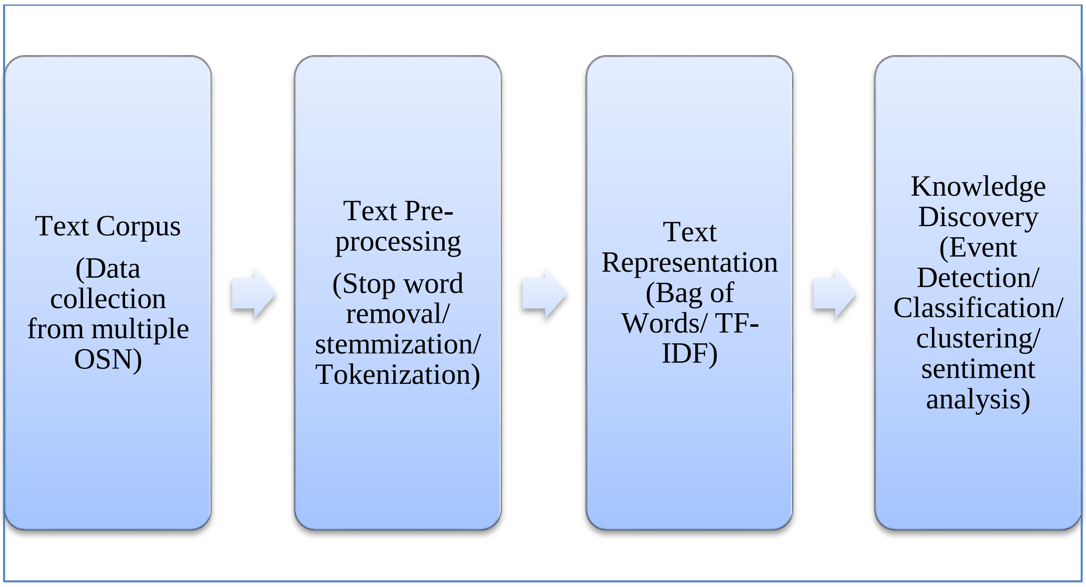

Text Corpus
(Data collection from multiple OSN)
Text Pre-processing
(Stop word removal/ stemmization/ Tokenization)
Text Representation (Bag of Words/ TF-IDF)
Knowledge Discovery (Event Detection/ Classification/ clustering/ sentiment analysis)

Figure 2.14 Phases of Text Mining

A) *Text Pre-processing*

This step refers to the processing of raw data to deliver a podium for data analysis. The significant purpose of this step is to classify raw sentences into sentences which can be read by the machine. The text is cleaned and delimiters are removed with the help of some pre known list of stop words which are not useful to classify the meaning of the sentence. The text and its characteristics are pointed in an attribute value table.

Users enter the social text in a free form and therefore it is a challenging task to classify that data. Just to be sorted out for this challenge, part-of -speech tagging and Named Entity Recognition are used [23]. Traditional methods for preprocessing consist of stop word removal and stemming. Stop word removal eliminates words using a stop word list, in which the words are considered more general, meaningless and stemming [23] reduces inflected (or sometimes derived) words to their stem, base or root form. For example, “watch”, “watching”, “watched” are represented as “watch”, so the words with variant forms can be regarded as the same feature.

Given the corpus of a microblog:

“Watching the Modi’ speech”
“I Like the Modi’ speech”
“ they decide to watch a movie”

The output of text pre-processing for the three blogs are:

“watch Modi’ Speech”
“Modi’ Speech”
“decide watch movie”

Pre-processing methods depend on the specific application. In many applications, such as Opinion Mining or NLP, they need to analyze the message from a syntactical point of view, which requires that the method retains the original sentence structure. Without this information, it is difficult to distinguish “Which university did the president graduate from?”, “Which president is a graduate of Harvard University?”, and “which have overlapping vocabularies?”. In this case, there is a need to avoid removing the syntax-containing words.

B) Text Representation

The most common way to model documents is to transform them into sparse numeric vectors and then deal with them with linear algebraic operations. This representation is called “Bag Of Words” (BOW) or “Vector Space Model” (VSM) [4]. In these basic text representation models, the linguistic structure within the text is ignored and thus leads to “structural curse” [5]. In BOW model, a word is represented as a separate variable having a numeric weight of varying importance. The most popular weighting schema is Term Frequency / Inverse Document Frequency (TF-IDF)[64] Where former deals with correlating the term frequency which is calculated as the number of times terms appears in a document and later is defined as the correlated value to the inverse of the number of documents in which the term appears.

C) Knowledge Discovery

When we successfully transform the text corpus into numeric vectors, we can apply the existing machine learning or data mining methods like classification or clustering. For example, in machine learning, a similarity is an important measure for many tasks. Knowledge discovery deals with developing algorithms to ascertain stimulating, unforeseen and unusual information form the patterns in the text document. One of the common tasks that occur is referred to as Apriori [71]. Frequent behaviors of persons or entities are recognized in the dataset. It identifies the inherent regularities in the data. This method was initially introduced in order to analyze customer buying behaviors from retail transaction databases.

Association, correlation, classification and cluster analysis form the strong foundation of data mining chores [68][72]. For example, finding a strong correlation between two users A and B, of the connection A ⇒B, indicates that user that likes a product were also likely to be preferred by his friend B, so using this rule company can make decision to sell product to B who hold strong friendship relation with A. Finding the user's opinion about a topic is another example. This can be done by using sentiment analysis to determine how the topic is discussed on Twitter or other social networking sites.

The efficiency and performance of the program depend on the type of indices chosen in this step. It is the central part of search engines and its primary aim is to optimize the speed to recognize the residence of output enquired by the query fired by the user. The efficiency of the impacting factors differs for various industry and organization. The speed at which the information retrieved and the cost to store the information is always a concern in any commercial application. In various situations, indexes are created in order to reduce storage. An inverted index is the most known and communal index method used in Information retrievals [65]. An example of worth mention is the index of books; with this, the location of the output of the query can be given by identifying the ID in the inverted index.

The above discussed techniques provide the platform to mine latent and valuable information from the social media like event detection, social tagging, collaborative question answering and helps in bridging the semantic gap. Time sensitivity, short length, unstructured phrases and abundant information are key challenges to analyze the textual information available on social media [55][92]. Information on the social web like news or user's posts update frequently, the response to the user's query should be the latest one. With the rapid evolution of user generated content on the social media, the text also suffers from the change and thus evolves the problem of time sensitivity in mining the information on OSN. Short context as already mentioned brings up the new challenge in text mining like text clustering and classification to name a few. The variance in the quality of text poses another challenge that originates from user's attitude of writing which makes filtering and ranking difficult to interpret. The rich variety of information posted by OSN gives rise to a wide array of non-content information available to the user as well which makes the system more complex. Recently, many authors have proposed techniques to handle the textual data with new features to extract useful information from posts, links and tags [33][84][95], identify influential users [29][34][44][47], aggregate the user's profile [8][36][38][64], understand the user's behavior[9][93], analyze the user's intention [43][82], measures the sentiments of user towards entity [72][75][78][92], predicts the popularity of news [39][77]etc. The techniques provided by text mining using concepts of NLP forms the foundation of the proposed model of QPSSN.

2.5 CONCLUSION

This chapter discussed Social Network Aggregators and NLP techniques for Social Network that can enhance the experience of the user in a more interactive way. The study presented reflects that traditional web text mining techniques are not popularly used in social network analysis. The combination of text mining and web mining techniques should be incorporated to analyze a social network system. NLP Techniques will help to aggregate user profiles using semantic vocabulary and enhance a user friendly search by the Social Network user while web mining encompasses the intelligence in the Social Network.

The next chapter throws light on the related work done in the field of social media by eminent researchers to fulfill the objectives of QPSSN, layouts gap in earlier research and motivates our intention to aggregate the user profile and extract useful information from it, which we will explore in this thesis work.

CHAPTER 3

THE RELATED WORK

3.1 INTRODUCTION

The social media network and its online services have become the most advanced system in this modern era [1][3]. In this virtual age, the services offered by the social communication networks are the important components of the digital image. Due to massive growth in the online social media, the size of the user footprint in online services is also increasing [6] as the user continues to interact with the friends, post the updates, write blogs, tag online resources and so on, just to list a few. The online digital footprints capture the user online identity and help in providing the identity based on the works done in the network.

Gross and Acquisti [99] were the first who studied user's sharing behavior on Facebook and its privacy implications. The authors observed a variety of information enthusiastically provided by the users ranging from their names, location, photos to interests (books, music, and movies), political views and sexual orientation, including personal information such as date of births, phone numbers, and email addresses. Similar analysis about the personal information of users on Twitter was taken up by Humphreys et al. [54] which provided evidence about the user's physical presence or activity using tweets that were publicly accessible.

The study of the literature reveals various works and application of information systems in social networks [62][65][67]. Various online SNS reveals various features and techniques to connect, interpret social data like sensing real-world events [30][13], detecting key users [100][109] etc. Organizations and enterprises use social data for various SNS to enhance brand exposure, brand community, link prediction, attribute inference and acceptance of the product by users [68][92][95]. Most of the work in literature has focused on to addressing either the 'walled garden' problem [37][38][41][64] or the information overload problem [32][33][34][35][36]. However, no significant work grinding the benefits and/or proposing solutions for helping users to provide a platform to extract the contents of their various social streams is found.

It is worth mentioning that the identified works utilized various techniques like data portability, Natural Language Processing (NLP), recommenders, mining, the structure of social network graphs and patterns in users activities [65][66][67][68]. Given the varsity of domains, it would be impossible to cover all domains of SNS. Therefore, for the purpose of our study, this literature study has been categorized into four major sections catering to the research objectives already stated in chapter 1.

Section 1 highlights the work of eminent researchers serving the need of extracting information across SNS. Section 2 analyzes the techniques suitable for identifying and integrating user's profiles across various SNS and extract useful information from the pool of SNS. Section 3 cites the efforts of renowned researchers advocating the need for clustering while gathering information over SNS. Section 4 analyzes and throws light on the algorithms suitable for sorting the information thus processed. Finally, a summary of the related work highlighting the most dominating works forming the foundation of this work is being presented.

3.2 INFORMATION EXTRACTION

Information extraction is one of the most important aspects of mining data from OSN. It aims at identifying data with relevant information or experience for a given area. Many authors have studied the need for information extraction and extracted useful information; this section presents a few the approaches used for information extraction.

Traditionally, information retrieval had been achieved by representing people via documents they are associated with. However, such an analogy does not hold good for real time cases in which objects are people and not documents. These people have social relationships, connections, friends etc. essentially it is qualitatively more complex to rank people than textual or object documents. Literature review suggests the concept of social matching systems such as referral web [102], expertise recommender [29][34][[44][47] and aardvark [4]. These matching systems extract out people based on the social similarity between candidate result and searcher. Some popular social matching systems are Recommender, Aardvark and referral web. These systems focus on the interaction of two user accounts called social similarity.

Further research has suggested that certain characteristics like Data completeness and hierarchical structure. Therefore the social search tools should provide a larger view of data within/outside existing social networks so as to give a broader view thereby emphasizing search within the online social network. A recent example of such a search (Exploratory people search) has been studied under PeopleExplorer project [9]. This exploratory people search allows users to create search preferences in form of defined models. An important finding of this project had been attempted to model task differences and user variants in people search. Literature also mentioned a dedicated search on web search engine usage in which researchers studied query logs, length, tropical distribution and temporal patterns of a web search query to understand search session.

People and Blog search [42] on a particular topic of interest issues a block search engine, this phenomenon has been researched and concluded that people are more likely to search for named entities. Similarly, a news blog would often be searched by news query which will refer to people for a relevant content. A study of search query log has also revealed that many users search for a single query; this means that such a people search will have a lower click through rates (CTR) as compared to a web search. A recent research on the social behavior of people suggested that search of queries often is limited to a primary click of one of the results which lead to higher CTR for a named query [96].

Matsuo et al. [103] have developed a system “POLYPHONET” to extract the information from the social network that detects relationships of person, groups of person and obtain keywords for a person. In the environment of the semantic web, social networks and semantics are the dualistic sides of the coin as pointed by Mika [104]. There exist several ways such as relation extraction, event detection etc. to extract the information from the social network [79][82].

Kautz [102] developed a Referral Web by extracting the measurement of the co-event of the names on the web. Research done by Kautz et al. [102] utilizes the network analysis of people to model the network of AI researchers. They use the name entity data found in close proximity in any public web pages such as the hyperlinks from home pages, co-authorship and citation of papers, exchange of information between individuals found in net-news archives, and organization charts.

In the past, the enhanced development of online social Networks (OSNs) derives the dispersion of a large amount of profile information inside corresponding social networks. Thus, sharing and reusing user's information accessible crosswise over OSNs is an emerging challenge.

Zhou et al. [109] have retrieved the information about the user using historical usage of information of the user. Yang S. [20] have employed the collection of operators on the graph and explored the Social Network Graph Query Language for performing a search on social media in a natural language. The system was implemented by building database management system from scratch that can make the control of components easy rather than taking inputs from existing social networks.

Tang et al. [30] introduced a uniform framework for efficient query processing and evaluation with an inexpensive storage and light deliverables of Points of Interest (POI) on large scale road networks. In order to further elevate the query efficiency a hot-zone based watchtower framework by incorporating mobile users movement information was provided into the physical framework of the construction of watchtowers.

Mukhopadhyay et al. [111] described the approach of searching in web was evolving constantly but the growth rate of the improvements was not that fast. The search engine proposed works efficiently and retrieves relevant web pages than reachable. In this study, a prototype that used multiple ontologies to perform multiple domains specific crawling for businesses to identify their clients in the market was proposed. The proposed research study works effectively and handles the challenges of relevant not reachable web pages.

Sun et al. [112] described the problem of protecting the user privacy in location sharing services such as nearby friends query and strangers query. A new framework and a new query algorithm (UDPLS) were proposed to protect user location privacy on the social network server and user's social network privacy on location privacy.

The user can share the location with specified friends instead of all friends. The query time of the framework almost has no effect on the number of friends in the friend query. The pseudonyms of user's friends and ID in the user terminals were matched. The extensive simulation experiments evaluated the performance of the proposed algorithm. The proposed research work resulted in additional traffic overhead.

According to Chen et al. [113], with the continuously developing applications more temporal social network groups will be paid attention. This research study focused on temporal analytics on social group query and the superiority of Temporal Group Query (TGQ). In order to effectively address the query two indexing structure to accelerate the query processing and two processing algorithms to accomplish the processing were deployed. The experimental results showed the research method was capable and the optimized method was efficient. The authors described that the time axis in social network was an important and useful tool to provide insight into the retrieval or statistics and augmenting the temporal query capability in such context was meaningful. Three different kinds of queries were proposed and Temporal Social Network (TSN) with users, relationship, and activity as well as corresponding temporal labels were modeled. A storage model was designed to logically and physically represent the TSN, and then proposed two index structures for accelerating the query process. The query processing algorithms were proposed for the three queries and evaluated the idea on a dataset which was synthetically generated from a real dataset, and experimental results showed that the indexes and query processing was effective and scalable.

Zhang et al. [8] personalized the search results by incorporating user's interest from multiple social networks using social activities of the user for Facebook and Twitter. The authors have exploited manual Friend Grouping and used various machine learning techniques on the textual features and the non-textual features (e.g. actor, activity type, source, etc.) of social data as well. The users are not expected to share the interesting information and useful contents with each other.

Bernstein et al. [55] suggested the use of search engines for identifying topics in a tweet. This includes a transformation of tweets into keywords, then creating a query to hit the database for final retrieval of results. Most frequent noun phrases are analyzed from the results as the topics of the tweets.

DBpedia, an external knowledge source can also be used to enrich tweets [4]. The technique first extracts names entities from tweets which are correlated to corresponding DBpedia entities. These entities are then assigned to original tweets as topics.

Irfan et al. [114] portrayed that despite the fact that various innovations were developed for the extraction of data from huge accumulations of the textual data, the extraction of useful information still turns out to be challenging when the textual information is not structured as per the grammatical convention. There have been numerous attempts to search for the structured data but none of the attempts is appropriate for unstructured data search engine [64]. However, the structured database supports text indexing but they agonize from poor performance [36][47]. The upcoming section details the related work in identifying user's digital footprints and about its aggregation across various social networks.

3.3 IDENTIFICATION AND AGGREGATION OF USER'S PROFILES ACROSS VARIOUS SNS

Identification and integration of user profiles is an inevitable task in social networks. Study of literature reveals that researchers have been putting days and night for improving the algorithms catering to the identification and aggregation of user profiles.

Lampe et al. [115] deliberated the impact of different types of profile attributes that users provide on Facebook. It was discovered that the profile attributes play an important role in the users profile to share common references (e.g., school, employer) and recommendation for friends. It took almost 10 years for users' online privacy concerns that caused users to the frontier the access to some of their profile attributes [3]. However, an enormous amount of information about the user still remains accessible to the public. So far, many researchers have only observed at SNS separately and did not pay much attention to integrate the available information and then extract relevant data from multiple SNS.

Prior work also premeditated that the users leave footprint across multiple social networks which can be useful in the aggregation of the user [8][9][10][11]. The creation of integrated profiles of users has many applications in industry and promotions of the products. While on a per-site basis, a user may seem fine what information is available to his/her from Facebook, Twitter, and LinkedIn accounts, she might be interested in much more than she realizes when considering them in integrating. As an example, one could first identify employees for an organization on LinkedIn, and then examine their interest from Facebook accounts for a personal background check to exploit other attributes of one’s personality as well to understand humanity patterns.

To accomplish this challenging task, there is a need to identify the information from unifying multiple accounts that correspond to a single individual. First, organizations are interested in correlating user activities and aggregating information across multiple social networks to develop a complete profile of individual users than the profile provided by any single social network. Second, social networks are interested in finding all the accounts corresponding to a single individual inside a single social network. Users are supposed to open only one account in a social network (as stipulated in the Terms of Service), however, some users create multiple accounts.

Through SNS, the user creates his/her profile by adding attributes, for example, his/her name, pictures, and friends; hence, diverse profiles are distributed over the network. These profiles incorporate significant data about the user for promoting, user-driven undertakings, and a user's individual verification. The worldwide data about the amassed user profiles should influence the client to comprehend the protection and security issues of his open data.

The connection of the user's profile remains a conceptual strategy for a credulous user. The developing requirement for coordinating user profiles and connecting his/her character won't just keep the user educated yet in addition is the premise of new headways in mining data about the user for customized errands. Thus, there is a need for integrated user’s profile that can construct heterogeneous social data into unique profile gathered from multiple social networks. Figure 3.1 depicts the integrated user who overlaps at two different social networks.

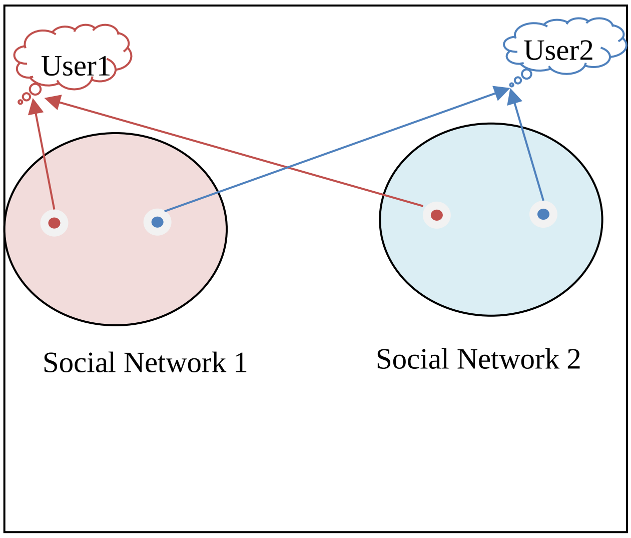


Figure 3.1 Desired Integration

A) *Identifying User*

On different social networks, a user puts a variety of his personal attributes; therefore, the challenge is to map a set of these attributes with high precision and accuracy. The above discussion brings up the fact that resolving an identity is a major challenge. In order to identify a user uniquely, initially a set of publicly available attributes which can find the similar account across multiple social networking sites with the claimed precision, accuracy, and recall were explored. Some of the important attributes common to most social networking sites [116][117] are user name, display name, profile image, description, location, age, sex, group of interest and connections. Although the major attributes that distinguish a user across multiple social networks are publicly available information fields, however, users may provide different information on different social networks for the same attribute. For example, the same user may use the name John on Facebook and Jon on Twitter. Thus, different information about the users' same attribute from multiple social networks requires learning about the mapping of these attributes to know more about the users [118][119]

In fact, while using social network services, the user creates his/her profile by adding, for example, his/her name, pictures, and friends; hence, diverse profiles are distributed over the network. These profiles include valuable information about the user for advertising, customer centric tasks, and a user's background check. The global information about the aggregated user profiles shall make the user understand the privacy and security issues of his open information [130]. The linking of the user’s profile remains an abstract procedure for a naive user. For instance, PleaseRobMe.com integrated information from tweets and FourSquare to discover that the user was not at home [28]. The growing need for matching user profiles and linking his/her identity will not only keep the user informed but also is the basis of new advancements in mining information about the user for personalized tasks.

Profile matching algorithms have been considering the ‘username’ and ‘name’ attributes to map a link between a pair of user. Literature review suggests various methodologies which compare inter-site attributes which are common or similar user profile based on defined metrics of these attributes. Limitations of such methods are discordant social platforms with some overlapping attributes and heterogeneity of some attributes. There are tools that compare common attributes between user identities which evaluate corresponding values based on defined metrics. Similarity on text attribute like name can be compared using the Jaro Similarity [118][119][126][136], while media attributes like profile picture can be compared using the histogram and other advanced matching techniques like face detection. Since these techniques compare the similarity of the current image to one on your profile picture, so there are chances that poses low similarity index because of an attribute being changed over the time.

There are some limitations though as the methodologies mentioned above consider current values of attributes or the attributes may have evolved over time. Attribute evolution studies have shown that the temporal nature of OSN lets user evolve some of the attributes over time. The situation becomes more complex when some attributes become outdated in some social networking sites for the user. This inherently gives more complexity to compare these attributes with precision and accuracy. Use of past values such as attribute history can be suggested solution of the problem. History gives an insight on a choice of length, characters, lexical and morphological structures, frequency of reuse of attribute values [118]. The attribute history is an important parameter to extract similarities in identifying a user. On different social networks, a user puts a variety of his personal attributes; therefore, the challenge is to map a set of these attributes with high precision and accuracy.

Recent studies [114] that examine temporal nature of OSNs suggest that users exhibit a tendency to evolve their attributes over time. Consider the following scenario: a user registers on Twitter and Facebook with the same username value; she favors Twitter and updates her Twitter profile more frequently than her Facebook profile. After a few weeks, she chooses a new username on Twitter, not similar to the old one but makes no such changes on Facebook. Due to the evolution of username over time on a favored social network, she now owns dissimilar usernames on her profiles. On observing dissimilarity, existing methods that match only the username falsely conclude that Twitter and Facebook profiles refer to different users.

To validate if a significant section of users changes attributes, an automated system is deployed to track Twitter users every fortnight and record changes to their attributes. A significant number of users have changed their attribute that evolves over time and holds distinct values for their attributes. On a two-month period, it has been observed by Jain et al. [126] that 63.21% users change their attributes and assign distinct values.

Further, the attribute test if evolution causes dissimilar current values across profiles of users and hence, filter users who evolve their usernames. Jaro Similarity and Edit Distance is computed between current usernames on their profiles and observed that 78% users have usernames with Jaro similarity less than 0.7 and 62% users with Edit distance greater than 0.7 implying dissimilar current usernames across profiles for a majority section of users due to username evolution. Thus, a low similarity between current usernames can be falsely manipulated by existing methods as different users.

A user profile is composed of multiple attributes; each signifies a unique characteristic of the user. Among the attributes, the literature suggests name, username, and location be an essential and discriminating attribute for profile linking [119]. In addition to availability and uniqueness, usernames can only contain alphanumeric and special characters irrespective of the preferred language of the user profile, thereby allowing clean string comparisons. History of other attributes like user posts can further help in identifying user profiles of the same user.

Singla et al. [90] have explored inference rules to identify if the records belong to the same entity using first order logic to recognize if a record predicate or reverse predicate are alike on two citation databases – Cora and BibServ.

Chen et al. [120] and Bhattacharya et al. [121] exploited the entity resolution problem by mapping each reference as a node and its relation with other references as an edge in a graph (co-occurrence). Chen et al. approached network structure on movie and citation databases and proposed that high confidence of two similar references should be considered as one [120] while Bhattacharya et al. [121] proposed to use network structures with arXiv and Elsevier BioBase citation databases to find the references between common entities co-occurring with each reference. The two references tend to point the same real-world entity if there is a large common network between two references.

Motoyama et al. [122] are the first that exposed profiles attributes (e.g. username, location, and school) to match user profiles to help users to identify friends when the user joins OSN. The proposed algorithm considered profile attributes as bags of words and measured the similarity between two user profiles as the number of common words between deliberated attributes. The methodology grieved from low recall as it failed to identify user's accounts for common attributes that have marginally different names. For instance, it failed to identify that the Georgia and Atlanta refer to the same area in the United States. Other researchers [123][124][125] defined more focused text based metrics for computing the similarity between different attributes and used classifiers to differentiate between dissimilar user accounts.

These methodologies [93][122][123][124][125] have only been evaluated on small datasets which are susceptible to provide many deceitful results of accounts when used at scale. A few recent studies identified user's accounts across multiple SNS at scale [119][126] and pointed out that identifying user's digital footprints at large scale produces a large number of false results. However, the system failed as the performance drops drastically when matching are between similar looking user profiles.

Irani et al. [127] exhibited that user's accounts can be matched by searching for accounts where the users have considered their names as screen names (the user login). A similar list is depicted for other OSN like Delicious,[39] Flick[40]r, LinkedIn, MySpace, LiveJournal, Twitter and, YouTube. Common attributes like home-town are consistently available only on OSN like Delicious, Flickr, MySpace and YouTube, while birthday is available on LiveJournal, MySpace and YouTube. Complementary attributes like gender and birthday are available only on Facebook but not on Twitter. Further, few OSNs enforce similar policies on the veracity of the information. The research indicated that on an average name, school, location are types of personal information that user willingly provide on one social network. Moreover, one can learn user's behavior by matching different value for same attributes on multiple SNS as user has a tendency to provide values to SNS as per its functionality.

Zafarani et al. [93] presented more refined techniques to identify the linking on the basis of user names on the assumption that it is a general practice that users choose same user names across social networks. Based on users' limited vocabulary words to create screen names or geographic origin to use local language, authors learned a supervised classifier and detect if a username belongs to the user who owns the username set. However, it failed to prove on SNS that generates usernames automatically from real names of users, like Facebook and LinkedIn.

To digitize the user's identity, various attributes (public and private) [129] are being exploited to match users across social networks. The unification of accounts using graph-based techniques is discussed in [57][93][127][129]. The graph-based technique, such as Friend of Friend (FOAF), links multiple user accounts based on identifiers such as email ID, Instant Messenger ID. The

[39] https://delicious.com/
[40] https://www.flickr.com/

graph thus generated across different social networks is compared and a score is assigned to it. If the threshold score is the same of every network, the identity is considered to be of the same user. However, this technique is not scalable and relies on private information, thereby raising question on the privacy policy of different social networks [6].

Narayanan et al. [129] de-anonymize Twitter users with the use of Flickr network using graph theoretic methodology. Authors iteratively matched each node network using a set of seed users (pre-deanonymized users) to find to the most similar node with the similar friend network and claimed 30.8% accuracy. However, the method needed 150 seed users in anonymized network and Flickr network, each having more than 80 friends. Recent research improvises on seed selection techniques by using unsupervised clustering methods on profile attributes [127], graph and subgraph matching methods to find social structure similarity [93]

Bilge et al. [128] proposed linking multiple identities by "searching and linking" as depicted in figure 3.2.

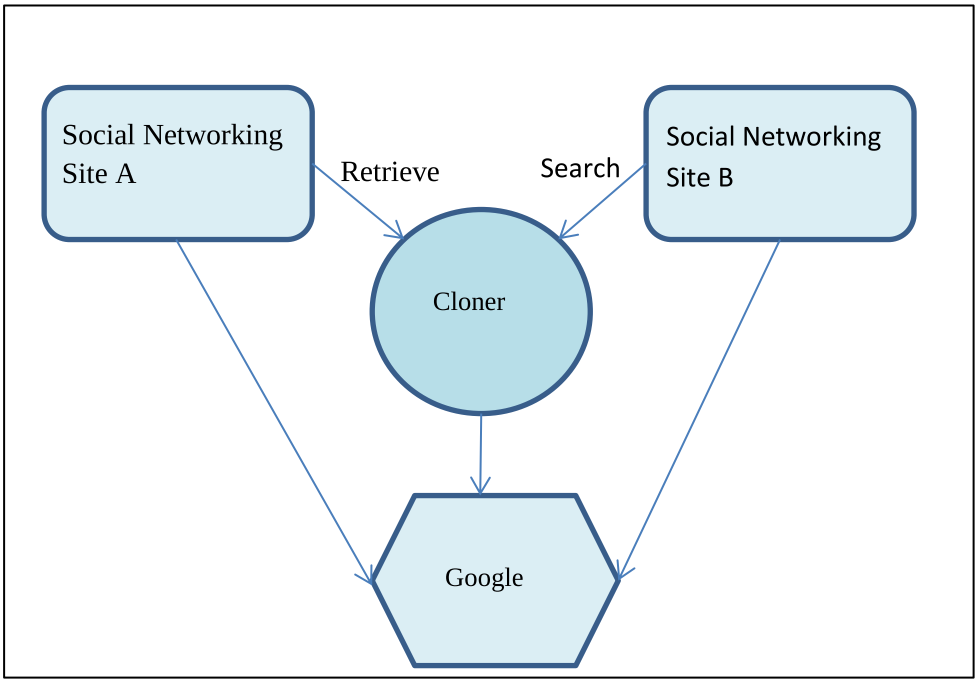


Figure 3.2 Searching and Linking by Bilge et al.

A few identity attributes can be picked such as first name, last name, education from each identity and compare it with a search engine. This is an identity resolution technique like Pipl which is extensively used. So based on attributes the online portals are searched which browse into criminal records, IM, court records etc. each portal being unique to attribute provide good matching results. However, due to vast data these portals often produce lists rather match the identity.

Another popular approach connecting various user accounts is the tagging [108]. The average established accuracy for the above method is around 64.5% [108]. Correlating user id and user names is another popular option to establish a single identity. However, the accuracy rate is just 66% [93]. The user identification algorithm [132] computing the weighted score of various attributes of the user profile is one of the most successful approaches in the domain.

Labitzke et al. [133] followed a different approach of matching mutual friends between two identities (to be matched). Authors used string matching methods to link names of common friends of two identities. The two identities were marked as linked (belonging to the same person), if there exists more than three mutual friends with the same name. However, the approach had a gap of understanding that in the real world, there could be multiple mutual friends between two users, or no mutual friends (in case when user used different social networks for different purposes).

It used a similar technique by comparing the friends on different SNS. It assumes that a user in one SNS will have many overlapping friends in another SNS. A similar approach has been used by Korula and Lattanzi [131] by using friendship graphs to match accounts across different SNS.

In a different scenario, Srivatsa and Hicks [134] explored how mobility traces can be utilized to match their contact graph with the friendship graph of a social network. While we use data about users’ friends to match individual user accounts, we do not leverage the social network graph as a whole. The structure of the social network graph is certainly a very powerful feature to match accounts. However, assuming that we gain access to the whole social graph is not practical if we want to build a real-time and on-demand service that takes as input one account on a social network and searches for the matching accounts on other social networks. Nevertheless, combined with other features, these techniques might improve the matching accuracy.

Acquisti et al. [135] used face recognition to detect user on dating site. The match percentage, though poor, let to the conclusion that face recognition algorithms are not scalable for practical purpose with 10% success rate. Face recognition work well if are able to train the classifier with multiple photographs, however, in a practical scenario we often have access to just one instance. Another way to identify similar user profile is to detect the similarity in photo by photo similarity.

Perito et al. [136] displayed that user's profile can be matched by measuring the similarity between their user names. The similar technique to map users' accounts across different forums were proposed and implemented by Liu et al. [137]. Since the same user name can resemble dissimilar users, both Liu et al. [137] and Perito et al. [136] explored the exclusivity of user names to increase the precision of proposed technique for matching.

Bartunov et al. [138] used a structure of networks to match two SNS. This technique which is called joint link-attribute used to infer the attributes that are missing based on network structure and link matching node by using profile attributes. This has been proved to be a powerful feature and improve overall matching accuracy.

Malhotra et al. [119] exploit user name, name, description, location, image, connections for mapping user profiles listed on Twitter and LinkedIn. Identity search by Jain et al. [126] performs matching on the basis of the user's profile, content, network, and a self-mentioning mechanism to map the user to Twitter and Facebook.

Zhang et al. [139] presented holistic supervised learning using the user's public and private information from the social network for resolving the identity. However, with the advent of Web 2.0, a user can prevent the visualization of his connections and other features which is used to identify and disambiguate users [6].

You et al. [140] used schemes to link user name to social identity. User's social identity is matched by a relational graph of co-occurrence of names, extracted from entitycube to friend graph. This technique has a drawback in a sense that it assumes the user is active on SNS thereby limiting the scalability.

Nie et. al. [141] represented a strategy to recount user's identities across social networks by mining user's behavior data and attributes. The approach utilized two components: the formal component recognizes diverse users by analyzing user's behavior and discovering robust divergent types, while the later component constructs a prototypical of behavior attributes that acquire to find the distinction of users across OSN.

Zhou et al. [142] presented a novel technique for user identification using the friendship structure and topology of the social networks which makes it an expensive approach for scant online social networks.

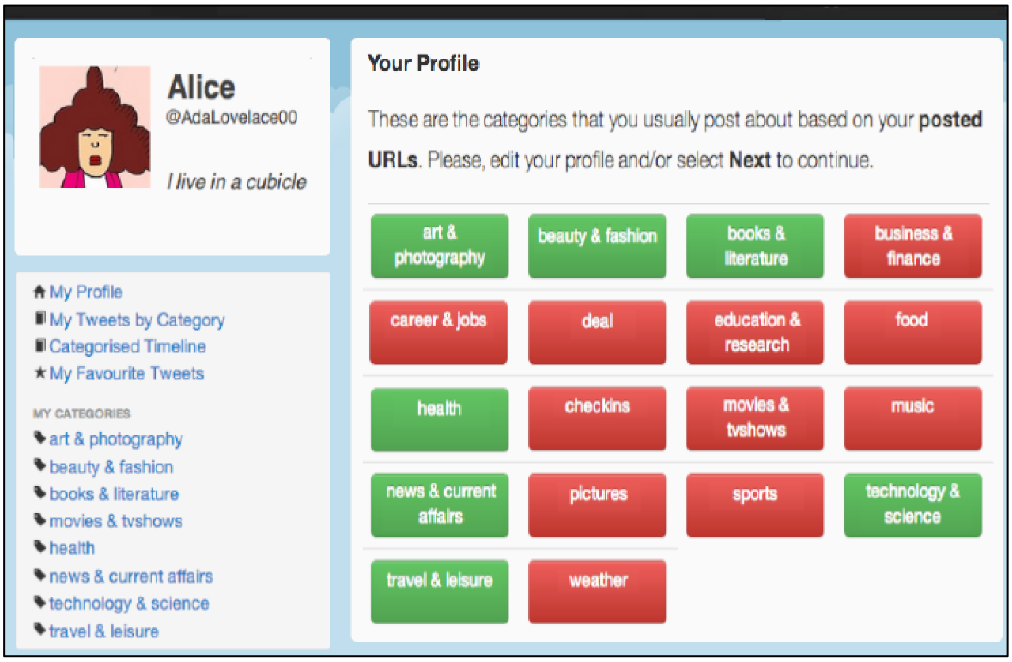
Liu et al. [137] proposed a semi-supervised embedding algorithm which uses the capability of the network to learn the follower/follower of each user. Despite the accuracy of above discussed algorithms, the researchers have not considered the timestamp for resolving the identity and have applied the techniques concerning the network factors.

Many researchers considers profile attributes in the criteria to match the identity using syntactic [127][136][142], semantic [143][144], and graph-matching techniques [57][93][126].

Further, the same set of researchers proposed an iterative resolution methodology where a set of references are resolved given that the references that they are connected to (or with which they co-occur) gets resolved first [143]. The process is iterative to start with the references of most confident similar references and then continue with resolving the entire database.

Like Chen et al. [121], the point of the research was on tweets containing URL and was assigned to one or more topics that are categorized based on contents of referred web pages using Bayes Multinomial Classifier [145]

Another user profiling technique suggested by Garcia-Esparza et al. [38] implemented a stream filtering system where users are embodied in categories as depicted in figure 3.3. A user's interest can be interpreted using filtering a timeline and prioritize tweets that encompasses information about user's own interests based on the categories of the posted URL's which corresponds to 18 general topics such as politics, movies or health.



Your Profile

These are the categories that you usually post about based on your **posted URLs**. Please, edit your profile and/or select **Next** to continue.

Figure 3.3 CatstreamTimeline

The dwelling of literature clearly indicates the fact that user profile disambiguation is achieved by using a large set of public and private attributes and in general, the three-step matching scheme [118] exists for mapping the users who deliberately create isolated profiles on different social networks.

In brief, the social network allows a user to opt out of the public display of the friend's list and other several attributes which is mostly used in the above techniques [119][121]. Since connections and the friend's list information can be restricted by a user, they cannot be used as matching criteria. The exact matching of a user's profile may not be possible as users tend to isolate their identity across social networks.

Table 3.1 throws light on the limitations of the work of the eminent researchers on identifying digital footprints across multiple social networks.

Table 3.1 Summary of Literature to Identifying User's Across OSN

Author	Description	Limitation
Szomszor et al. [108]	Used content of 502 users using syntactic methods	Experimented on limited number of users
Carmagnola et al. [101]	Used profile attributes to map MySpace and Flicker accounts of the user profile over 300 datasets using probabilistic methods achieving precision of 86.9%	This approach requires advancements to be fully functional and productive. Tested on low number of users
Irani et al. [116]	Covered 12674 sets of profile attributes using syntactic methods	The approach depends upon the screen name and same screen name can correspond to different users. Low in accuracy
Goga et al. [118]	Used profile and content attributes using syntactic and probabilistic methods on public and private information achieving accuracy of 29.8%	The identities are linked in a passive way, the probability of miss-linking of user identity is high.
Malhotra et al. [119]	Used user profile over 29,129 datasets using syntactic methods.	Used too many profile attributes.
Motoyana et al. [122]	Used profile and Network attributes over 900 sets of user's profile using syntactic methods achieving accuracy of 72% for MySpace and Facebook users	Low Recall and Fails to match significant different names
Jain et al. [126]	Used profile, content, network and self-mention attributes to match user profiles	Low Precision and Recall
Bilge et al. [128]	Searching and linking using first name, last name and education	Lacks in stability
Narayana et al. [129]	Graph theoretic based methods over 27k datasets providing accuracy of 30.8%	The probability of error rate is high
Labitzke et al. [133]	Used StudiVZ, Facebook, MySpace using Network attributes to map over 300 sets of users profile using string methods to link names of common friends of two users	Lacks in accuracy as there is a high probability of same name to multiple mutual friends or there may be no mutual friends between two users. Lacks in accuracy.
Acquisti et al. {135]	Used Face Recognition	Low success rate
Perito et al. [136]	Used profile attributes to map Google and ebay accounts over 10,000 datasets using syntactic methods achieving accuracy 71 %	This approach is based on the entropies of two strings to be associated with two usernames
Nie et al. [141]	Used user's behavior data and attributes	User's behavior is variable across OSN.
Zhou et al. [142]	Presented a novel technique for user identification using the friendship structure	Expensive approach

B) Aggregating User's Profile

Given the two linked profiles of the user, an interesting opportunity is to develop a unique profile that commendably empowers systems to benefit from the disseminated knowledge about users, preferring the exchange and reuse of user information.

Madnick and Siegel [146] projected that the usage of aggregation applications will be increased at pace due to an enormous growth of the content on the OSN. The aggregated applications will help the personalization mechanisms to improve the overall recommendation accuracy and the weight of information it carries [147]. Many researches have explored this area focusing on the issues such as the aggregation and management of diverse user profiles.

Berkovsky et al. [148] proposed a model to integrate user content from other personalization systems. The framework utilized specialized mediator components for interpreting the information between different prototypes using inference and reasoning mechanisms. There exist several challenges as already discussed to user data integration, such as formats used for representation, multiple meanings of procurement, privacy risks, etc., user aggregated profile can nonetheless be helpful as a provision to the personalization web.

SONAR [149] an interesting API for gathering and sharing user's content with respect to the aggregation from OSN. Specifically, it is centered on recognizing and exploiting connections between people, who might be connected in several ways. However, it failed to address the issue of identification of user across social networks.

Carmagnola et al. [150] have proposed a mechanized coordinating calculation which under the set of user characteristics like sexual orientation, birthday, city, can register the likely comparability between these starting characteristics and creep information from social sites. More is the information crept, more precise is the calculation. In any case, ambiguous information is freely accessible because of the closeness of the majority of prominent OSNs.

Another OSN Aggregator proposed by zhang et al. [8] not only pulls the social information from multiple networks but also group, rate and notifies about the activities of friends. However, the system failed to integrate the networks. In fact, numerous models have been advanced to outline a collective objective model for assimilating a user [64].

Singh et al. [151] also suggested extracting the user's birthday value resulting in exact identification when it is cross-matched with the users' name.

Abel et al. [152] aggregated user profiles on the limited set of properties like name, photos etc. using the most popular solution FOAF from online social networks by applying rules. They presented a related approach for generating RDF based profiles of the user as per the frequency of the entities mined from user tweets and later modeled using FOAF ontology.

An examination of various temporal patterns and dynamics for Twitter profiles is additionally given by [196] that concentrate on a conglomeration of profiles of the user by and large. Consequently, they propose an approach for consolidating diverse Social Web profile attributes, for example, email, phone number, home page, and so forth. Additionally, tag based profiles of inclinations aggregated from various Social networks are assessed in a tagging recommender framework

Vu et al. [64] have presented a primary social user aggregation based on the FOAF ontology. However, the model has neither kept trace of the provenance nor adding a time of any information. Moreover, there may be conflicting values for a given property and it is left to the user to decide if information should be kept or deleted.

In order to create users' aggregated profiles, Pontual et al. [153] designed a crawler that collects information from different social networks and sites using a real name. However, the same lacked in correlating the accounts that can achieve maximum accuracy.

Orlandi et al. [36] model the user interest by combining the profile information and semantic web using FOAF ontology and DBpedia resources.

The proposed weighting scheme used to generate semantic user profiles using an aggregated score with a temporal decay. Other techniques constructed Hierarchical Interest Graph as a named entity extractor to connect user's contents to DBpedia resources in order to extract the DBpedia categories associated with each tweet. It has been observed that user profiles based on DBpedia resources are more accurate than the profiles based on DBpedia categories

CUMULATE [154] and PersonIs [155] are the generic server's framework for modeling of user's profile information that handles the user's aggregated information. Given these advancements, it turns out to be increasingly imperative to create approaches that adventure the connected profile data in the setting of the present web view.

Last decade has seen SNAs aggregating the social information of users across many social networks like Hootsuite, ScoConnect, flock [7][8][64]. Owing to differences in the privacy policies (which in fact keep on evolving also) of all social networks, the existing SNAs fall short in various aspects such as resolving the identity of user i.e. ensuring that only the legitimate user profile is being integrated. Users need to register and authenticate themselves on each social network on the aggregator by providing their user-id and password to be syndicated. This section discussed mechanisms that connects and aggregates the user profile from various social networks.

Information needs to be aggregated in such a way that it is more than a trivial impression, and yet overcome the problem of information overload and walled garden. The presentation needs to provide user's information, allow associations and be easy to access. The Aggregators struggle to keep up with fast stridden social ecosystem as the information is evolving overtime.

Table 3.2 depicts the existing techniques implemented on Social Network to aggregate the user's profile but there is a need to propose effective novel algorithm that can be used to integrate user's profile from public available attributes among multiple social networks with high accuracy.

Table 3.2 Summary of Literature to Aggregate User across OSN

Author	Description	Limitation
Zhang et al. [8]	Proposed and implemented personalized SNA using user's public and private information	This has a limitation of centralized user's data and not integrated them
Orlandi et al. [36]	Models the user interest by combining the profile information and semantic web using FOAF ontology and DBpedia resources	It requires an overhead of analyzing that identifies entities to link
Vu et al. [64]	Aggregates user profiles and maintains links between these profiles	The approach failed to make the difference between various activities of user.
Zafarani et al. [93]	Correlates user id and user names to establish a single identity	Low Recall The assumption made in the approach is not suitable for the OSN that inevitably spawns screen name like Facebook
Matsuo et al. [103]	Designed "POLYPHONET" to extract the information from the social network that detects relationships of person, groups of person and obtain keywords for a person.	The approach is restricted for recognizing relations and groups
Madnick and Seigal [146]	Aggregates the available social content from various personalized systems and thus improves the recommendation.	This approach didn't kept trace of extracted data for further analysis
Berkvosky et al. [148]	Integrates user information from multiple systems	Low Reliability of the scheme The OSN have limited access to user's information
Singh et al. [151]	Extracted the user's birthday value to identify with the users' name.	User's Birthday is an optional attribute so there is a possibility that user provide no value to this attribute
Abel et al. [152]	Used FOAF for integrating user profiles	Limited to a less number of public user attributes
Pontual et al. [153]	Designed a crawler to collect the information	Failed to correlate the users.
CUMULATE [154] and PersonIs [155]	Designed generic user's framework for aggregation of user's information	Did not integrate the profile of user

3.4 CLUSTERING USER’ PROFILES

As the number of social network users increases, a tremendous amount of data is generated by the sharing of information. The intuitive nature of these social networks is the creation of related groups (or clusters) [156].This has become an area of interest in the discovery of communities in recent times. These patterns can be used to mine a variety of information, which can be used in various fields such as search, influence discovery, marketing etc.[157]. The emerging field of social analysis uses data mining as the key input for analyzing data. Clustering is an important factor in this analysis. It is the process of creating related social actors in a set of meaningful subclasses which will later help the ranking mechanisms to improve results. The aggregated profiles will aid in discovering the user profiles based on different attributes.

Good clusters were defined by various cluster conditions with numerous attempts to the multitude of algorithms. A good cluster has a maximum weight associated with the group and minimum weight between the groups. In a social network, users assigned within the group should be similar on some attributes and users assigned to different clusters should be highly dissimilar on the taken attributes. The clusters can be evaluated on the high intra cluster similarity, low inter cluster similarity measures and external criteria like Rand Score, F measure etc.

Numerous attempts were made to improve the quality of clusters using ensembling techniques [157][158][159]. The main concern of many of these algorithms is to elucidate label correspondence problem. The limitation of many of these algorithms is the assumption of the same number of the cluster in each partition and may perform poorly when the information about output cluster is not known in advance.

It is approached by various clustering algorithms, including k-means, fuzzy c-mean, and table modeling [160][161]. While k-means is very fast, its center value depends on the value of k. Different values of k will result in different clusters [160]. Tang et al. [162] observed that the k-means learning algorithm requires specification of the number of cluster centers. If two highly-overlapping data exist, then k-means will not be able to resolve the presence of two clusters and also it is not invariant to non-linear transformations.

Likewise, Armentano et al. [24] proposed an algorithm for recommending followees in Twitter where users' profiles contingent from the tweets and an extra selection step is added to the progression that limited the users to select one’s extended social network(friend or friends). It was based on the assumption that if a user u_F follows a user that is also followed by u_T , then other people followed by u_F can be of interest to u_T.

Groh and Hauffa [110] have characterized the social relationships using unsupervised learning and natural language techniques for the purpose of linguistic analysis on the classification of sentiment polarity.

Rohani et al. developed a robust recommender system for academic social networks to recommend and analyze the products that meet the user’s preferences [107]. Further to model the interest of the user, various weighted concepts using the semantic web are listed in [108]. Twitter and Facebook are the domains used for extracting information for the exploration of the text. Few other models have considered expertise and relationship of a user into consideration and developed a social search engine [4][111].

Tyler et al. [105] have explored the detection of relations on the basis of information. This algorithm relies on the notion of betweenness centrality [106]. Given all shortest paths between all vertices, the betweenness of an edge is the number of shortest paths that traversed it. The idea is that edges of high betweenness connect people from two distinct communities, while edges of low betweenness connect people within one community.

The modification consists in calculating the contribution to edge betweenness only from a limited number of vertex pairs, chosen at random, deriving a sort of Monte Carlo estimate [45]. The procedure induces statistical errors in the values of the edge betweenness. As a consequence, the partitions are in general different for different choices of the sampling pairs of vertices. However, the authors showed that, by repeating the calculation many times, the method gives good results, with a substantial gain of computer time. In practical examples, only vertices lying at the boundary between communities may not be clearly classified, and be assigned sometimes to a group, sometimes to another. The method has been applied to a network of people corresponding through email and to networks of gene co-occurrences.

Zhang et al. [163] proposed the mapping of network nodes to identify the overlapping community by Euclidean space and fuzzy c-means clustering. Many researchers have sought community in social networks, as well as proposed metrics for evaluating the structure [160][161][162]. Yang et al. [164] proposed finding people by using mobile phone usage patterns in a social network. Another researcher proposed a hybrid study to retain customers using clustering [165]. Shapira et al. [47] developed a collective recommender system by exploiting user inclinations.

Gao et al. [60] and Eslami et al. [37] proposed clustering friends using Graph based techniques where each node represents a friend in the group that related to a user, the chain builds up as friends of friends gets added on as nodes. Specific clustering algorithm on the graph is proposed to bunch its internal nodes. It uses three levels of clustering techniques are used by researchers to demonstrate multi-level structure with subsetting groups within groups. A disjoint clustering lets a friend be in one group only while overlapping cluster algorithm allows a friend to be in multiple groups and a hierarchical clustering algorithm to demonstrate a multilevel structure.

Qu et al. [166] utilize Social links and textual information as provided by Twitter (tweets based on our example of Twitter) for suggesting group members. This information captured and modeled user’s topical interests using Later Dirichlet Allocation [167] (LDA) to extract out topics from user tweets. For the purpose of the research between two sets of entities, the system utilized group seed to calculate similarity and analyze the likeliness of a user to belong to the group derived from the tweets.

Clustering of friends can be made in multiple ways in a SNS – in form of links like friends or friend of friends or topical list like “Professional Developers” or “Movie Actors”. This, however, suffers from an inherent problem to maintain the lists based on a user’s ever changing priorities or interests and thus generating automated lists did not pose a viable solution. Table 3.3 throws light on existing work of the eminent researchers to use clustering in the domain of users of OSN.

Table 3.3 Summary of Clustering Information from OSN

Author Name	Description	Limitation
Armentalo et al. [24]	Proposed method for recommending followees	Limited the user's to one's own extended medium of communication
Kapanipathi et al. [46]	Built a centralized repository and provided topics of interest	Limited to Twitter users and low accuracy
Shapira et al. [47]	Exploited user's preference from Facebook profile for collaborative recommendation	This approach didn't kept trace of extracted data for further analysis
Gao et al [60]& Eslami et al. [37]	Implemented graph based methodologies to explore group of friends	Failed to provide personalized user's experiences on various preferences
Tyler et al. [105]	Designed a system to detect relations on the basis of information	This model is used to extract relationships of users only
Groh and Hauffa [110]	Characterized the social relationships using unsupervised learning and natural language techniques for the purpose of linguistic analysis on the classification of sentiment polarity.	Lacks in interoperability and reliability Used for analyzing sentiments only
Sun et al. [161]	Used fuzzy C-means clustering to identify overlapping and non-overlapping community	Unable to identify good clusters
Tang et al. [162] **Yang et al. [164]**	Employed k-means clustering to extract the information	Accuracy depends upon the value of k
Qu et al. [166]	Proposed and implemented a recommender for new friends using social links and tweets	Lacks in stability and suffered from overload problem
Blei et al. [167]	Extracted topics from tweets using LDA	The model doesn't work for correlated topics
Rakesh et al. [168]	Developed a personalized Recommender	Works for Twitter users only and there is no evolution of topics over time

Numerous techniques for generating cluster results and combining them have been seen in the literature [160][162][163][167][168]. Generation of input partition followed by integration of all the partitions to obtain final partition is a two-way process given by vega-pons et al. [169]. Median partition and object co-occurrence are the two ways to generate a consensus. In median partition, the final partition maximizes the similarity with all the generated set in the ensemble. This approach is not considered for clustering as defining the Mirkin Distance [170] have been proven NP-hard and computationally expensive. Object co-occurrence is another approach that obtains the final partition from the generation set depending upon the frequency of occurrence of an object together or an object to one cluster followed by the similarity based clustering algorithm. Co-association Matrix followed by clustering mechanism is a way to generate the occurrence of an object. Relabeling and cumulative voting is another choice for attaining the final partition from the generation set depending upon the frequency of occurrence of objects. Relabeling solve label correspondence problem using Hungarian Algorithm [171] following voting process by using cumulative voting [172] to obtain final partition. Other final partitions can be obtained by Genetic algorithm [173], NMF [165] and kernel Method [174] under object co-occurrence that is beyond the consideration of this paper.

Different strategies have been utilized to recognize community and merge community structures. As data clustering and community detection are very comparative, it ought to be conceivable to merge community in an indistinguishable way from ensembles of clusters with great outcomes of the hierarchical approach.

3.5 SORTING USER PROFILES ACROSS SOCIAL NETWORK

It is interesting to realize that social stream ranking is another well researched method. Different SNS have a different ranking method like Facebook’s EdgeRank [179], Twitter’s most recent tweet etc. EdgeRank focus on textual information, however many other social networking sites also use source and target users as rank criteria like the freshness of tweet [176], an influence of authors, quality of tweets. The freshness of the tweet is calculated as a difference of time when user saw a tweet and the time when it was posted. An influence of authors is based on scoring computation of most followed author to be a high scored than others in line. Another criterion is quality of tweet, which takes into account the length, presence of a URL, number of hashtags, number of rewets in a tweet etc. Facebook on similar lines user features like – explicit clicks on a message received, no of shares, likes, and comments per hours etc. [175].

Usually, a ranking model is constructed on a training of a pre-defined set of features assembled with machine learning technique, and thus the corresponding rank is computed on the incoming information. Hannon et al. [176] introduced Twittermender, which builds up a weighted interest profile vector where weight is dependent on no of tweets published by the user and(or) friend(s). To analyze the most similar profiles on request, his/her profile would be matched with others' profiles.

Weng et al. [100] proposed a different algorithm called Twitterrank to measure the impact of tweets to the domain of Friend Recommendation. It extracted topic-sensitive users by taking the link structure between users based on interest and communities which are common. Lim and Datta [177] rank method are similar to above, however, they identified the popular users that relate to interest or community. It first categorized celebrities that were representative of an interest category and then detect communities based on linkages among followers of these celebrities. It provided the user with open choices and do not directly suggest suitable friends to a user.

SNS use ranking to prioritize information relevancy and get away from ‘information overload’ problem. One of the most popular ways to achieve is to use the social stream of data in chronological order of occurrence. This method has not been effective as we may not always have an important post as most recent. This lets researchers and SNS devise an alternate mechanism to rank. EdgeRank is Facebook’s own ranking algorithm [179]. Facebook ranks the user activities to determine which status update, comments etc. will be displayed on homepage based on the said algorithm. An outline of the working is based on three scores – Affinity score (how ‘connected’ is a given user to Edge, Edge Weight (weights of importance assigned to comments) and Time Decay (a story is new or old in time line). The final rank is calculated from these three scores. Higher the score more is the probability of the story to be on the home page.

Facebook’s reranking feature lets sort the message posted on Facebook again based on certain criteria, for example, if a message receives a higher number of comments compared to other messages posted, it would be ranked higher and appeared on top. SBRank proposed another ranking method/page popularity measure which ranks according to the number of existing social bookmarks. Similarly, the sound generative model uses a model based on language to perform ranking. Further, in literature, Bao et al. [178] proposed two algorithms SocialSimRank and SocalPageRank based on connections between the pages, user’s social interpretations.

In general, ranking model is developed on certain desired features which are applicable on a training dataset with the help of certain machine learning techniques, the information that is input into the model can be ranked based on those features. Table 3.4 illustrates the review of ranking mechanism presented in the section above.

Table 3.4 Review of Ranking Mechanisms

Author Name	Description	Limitation
Shen et al. [40]	Considered freshness of tweet, influence of authors, quality and other social features.	Low-ranked results
Burke et al. [175]	Number of explicit clicks , share , comments and the mean number of message were the parameters considered for ranking	Not personalized enough for user’s preferences.
Hanon et. al. [176]	Proposed Twittermender to calculate rank on no of tweets published by user and(or) friend(s).	Risk of Information overload
Weng et al. [100]	Proposed Twitterrank to measure Friend Recommendation using tweets.	Not favorable results
Lim and Datta [177]	Rank method by identifying the popular users that relate to interest or community.	Limited features considered
Bao et al. [178]	Proposed SocialSimRank and SocialPageRank using connections	global relevance measures were not considered as the notion of relevance is highly subjective and dependent on the initiator of the user’s interest

3.6 CONCLUSION

All the above mentioned techniques have explored the identity of the user which relies on the assumption that the user doesn't actively obfuscate /hide her real social network attributes to avoid detection. However, the information scattered among multiple platforms have not been used to enhance the reliability and availability of the information. To detect such identities, researchers have devised methods for each social network; however, to our knowledge, no effective solution has proposed techniques to resolve multiple entities and identified features to extract the information in user friendly manner achieving higher accuracy. The next chapter discusses the proposed methodologies to extract meaningful information from multiple online social networks.

CHAPTER 4

DESIGN OF AN INTEGRATED QUERY PROCESSING SYSTEM FOR SOCIAL WEB: THE PROPOSED WORK

4.1 INTRODUCTION

Social Network Aggregators (SNAs) is used to maintain and manage manifold accounts over multiple online social networks. Displaying the activity feed for each social network on a common dashboard has been the status quo of social aggregators for long; however, retrieving the desired data from various social networks is a major concern. A user inputs the query desiring the specific outcome from the social networks. Since the intention of the query is solely known by the user, therefore the output of the query may not be as per user's expectation unless the system considers 'user-centric' factors. Moreover, the quality of solution depends on many factors like user-centric factors, user inclination and the nature of the network as well. Thus, there is a need for a system that understands the user's intent serving structured objects. Further, choosing the best execution and optimal ranking functions is also a high priority concern.

Although the popular social networks have enhanced the interaction among people registered on their respective sites, however, the existing interaction rules do not allow inter-site sharing of user profiles and their activities. In fact, a user creates public profile with the intention to share activities globally through social networks. Now, since the identity of a user registered on numerous social networking sites is not integrated globally, therefore the different profiles of the user usually remain in isolation. It is an obvious fact that a genuine user registers with his or her unique attributes to create an identity, consequently shall be identifiable with at least some of the common attributes across all popular social networks. The current work finds motivation from the above requirements and thus uniquely contributes a profile integrator which is able to generate a single unique profile from multiple profiles (of a user) available across different social networks. The integrator outstandingly disambiguates user profiles existing across different social networks using public attributes with decision to map the profiles using change in location of the user. However, clustering plays a vital role in this process. Therefore, a novel clustering mechanism has been proposed to analyze the relationships and psychology behind it. The proposed clusters can be spawned to envisage the discoverability of a user for a particular interest. The current work also proposes the design of a query processing system to retrieve the relevant information from these clusters to extract user's intent from various social networks on the request of a user. The proposed framework also contributes a user-centric query retrieval model based on natural language and it is worth mentioning that the proposed framework is efficient when compared to temporal metrics. It is an innovative approach to investigate the new aspects of the social network. The proposed model offers a significant breakthrough scoring up to precision and recall respectively.

As shown in figure 4.1, entire research work is being carried out primarily in three phases namely, The Profile Integrator (HIASN), The Clustering Mechanism (HEKHAC) and The Query Processing Mechanism.

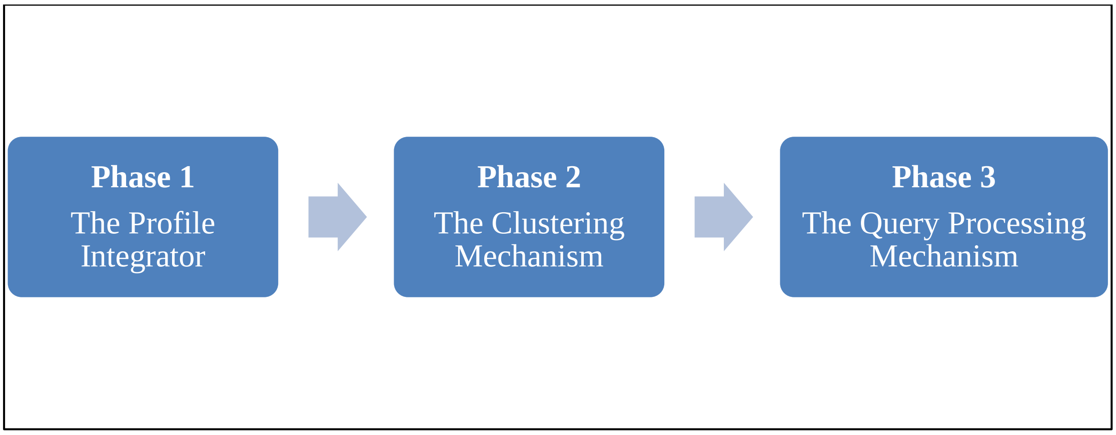


Figure 4.1 Phases of the Proposed Work

The chapter presents details pertaining to first two phases only i.e. The Profile Integrator and The Clustering Mechanism while the third phase is being addressed in next chapter. The Profile Integrator uniquely contributes to Identification of Contributing Attributes, Identity Resolver Module (IRM) and a Profile Integrator Module (PIM). In contrast to PIM which maps the unique identity of a user profile distributed across various social sites by correlating various public attributes, IRM performs user mapping based on strengthening certain attributes such as name and location. The proposed Clustering Mechanism is a hybrid approach and an abstraction of related groups interacting amongst social networks to analyze and develop relationships. The main goal of the proposed hybrid technique of clusters is to extract the entities and their corresponding interests as per the skills and location by aggregating user profiles across the multiple online social networks. The third and final phase proposes a query processing mechanism for the novel social network aggregator referred as Query Processing for Social Network Aggregator (QPSNA) is being detailed in the next chapter.

4.2 THE PROPOSED PROFILE INTEGRATOR

The discussion presented in the previous chapter brings up the fact that resolving the identity of the user is a major challenge. The current section thus uniquely contributes the profile integrator referred as *Hybrid Integrator for Autonomous Social Integrator (HIASN).*

The OSN enables the user to select their privacy settings and display the information that they want world to see [132]. Since associations and the companion's rundown data can be confined by a user, they can't be utilized as matching criteria A constrained characteristic set can be investigated for matching the user profile with a joining of the location trait of the newsfeed/tweets that is produced by the OSN or geo-labels that are created by the device when user refreshes through the status/tweets. Rather than utilizing various criteria to match and inquiry, this work discovers inspiration for a stepwise way to deal with settle the ambiguity of user profile giving better and more relevant outcomes.

HIASN is designed to aggregate the profiles extracted from multiple social networks. HIASN is an amalgamation of phonetic encoding score and the Levenshtein Algorithm to resolve the problem of matching user’s profile across multiple SNS and providing a unique integrated profile. While the former algorithm takes a keyword as input (person's name, location name etc.) and produces a character string that identifies a set of words that are (roughly) phonetically similar, later is being used to match user name spellings or pronunciations. The Levenshtein Algorithm is based on computing the Levenshtein distance between two strings where Levenshtein distance is defined as the minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character. This technique ensures that variations in user profile variables are handled correctly. While designing the HIASN, few challenges evolved which are being discussed in the next section.

4.2.1 Design Challenges

Although during the initial phase, designing HIASN seemed to be a simple task. However, following issue makes profile aggregation across social networks a stimulating task:

- Social Networks have diverse network structures and profile attributes for serving the functionality that makes the task of linking profiles difficult.
- Users may choose their username depending upon the functionality and service of the social network that may not be associated with their real identity.
- It is an evident challenge that many users may exist with identical usernames.
- Users may provide false information across their profile in order to masquerade.

In order to identify a user, the publicly available information is used conventionally. However, Goga et al. [118] depend on innocuous activity to identify users and used the location, timing, and writing patterns to enhance the quality of results. However, it has been observed that, rather than looking for all the locations, timing, and what the user has written, focus should be on the activity such as a change in the location of the activity at one social network and same should be mapped to another considering userid and name also.

In order to resolve the issues highlighted above, a solution addressing the needs is strongly desired. Hence, the literature was further grilled [118][119][124][126] and it was discovered that no best solution exists for mapping the user identity across the social network. Hence, hybrid solution exploiting phonetic encoding score and the Levenshtein algorithm has been proposed. It is worth mentioning that prior to the Levenshtein algorithm, the Jaro-Wrinkler algorithm has been used by various authors [118][119][126] playing a key role in computing string similarity. Jaro-Wrinkler is the modification of Jaro distance that calculates the string similarity as the sum of the number of common characters and the count of transposition as a weighted score for prefixes. The strings are more similar if the Jaro-Wrinkler distance is less.

Christen [180] has done an extensive study to compare the techniques for mapping string as personal name. It has been observed that choosing the right algorithm to match two short strings affects the performance of the system. In general, Jaro-wrinkler and Levenshtein distance are expensive algorithms as it involves an enormous number of evaluations because each string will be equated to every other string in the dataset. Identifying similar string using phonetic encoding and then applying sophisticated string matching algorithm will provide better performance and results. As illustrated in the previous section, Jaro-Wrinkler algorithm is expensive approach when applied for each name in millions of records and thus has not been considered while designing HIASN. HIASN determines user digital identity across several social networks targeting towards improving the search efficiently and precisely.

The next section uniquely contributes a profile integrator which is able to generate a single unique profile from multiple profiles (of a user) available across different social networks. The integrator outstandingly disambiguates user profiles existing across different social networks using public attributes with the decision to map the profiles using change in location of the user as one of the attributes. The proposed model is discussed in upcoming section that will increase the discoverability of the user, deriving new communities, and promotional activities among multiple domains.

4.2.2 High Level View of HIASN

HIASN determines user digital identity across several social networks targeting towards improving the search efficiently and precisely. Exploiting the fact that each social network provides various public attributes to identify the user’s digital footprints across an aggregated social network environment, the HIASN offers a three-phase solution i.e. identification of attributes contributing towards identifying user's profile, mapping of identified user profiles and finally produce a single integrated profile. The architecture of the system is shown in figure 4.2.

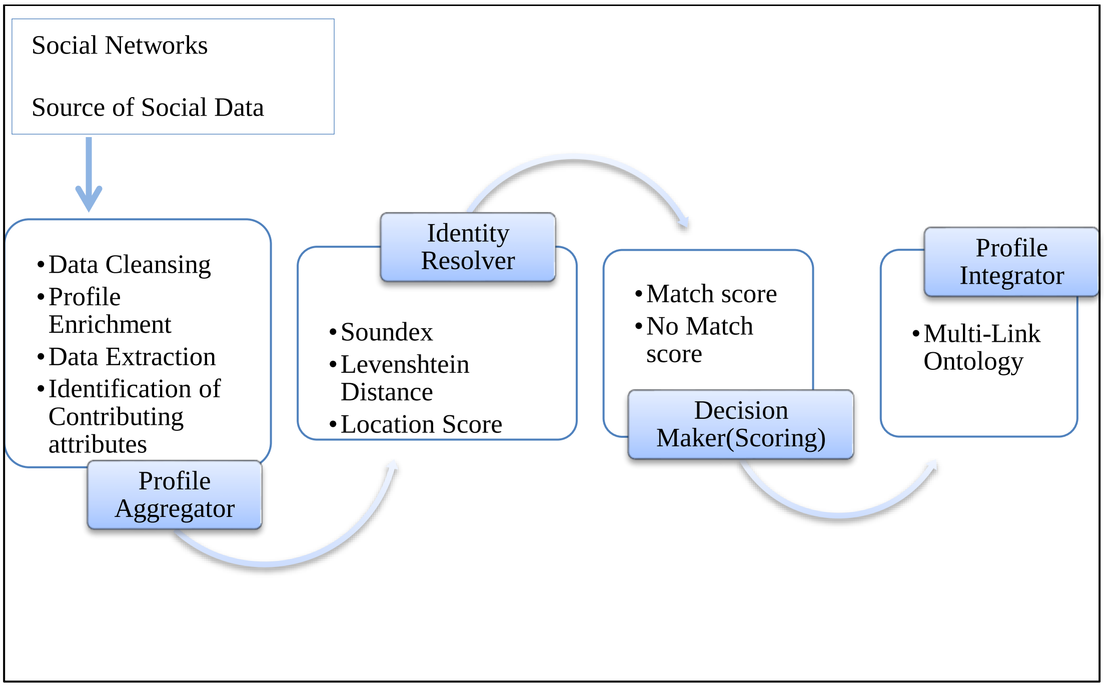


Figure 4.2 The Hybrid Integrator for Autonomous Social Networks

Prior work suggested the authentication protocols and APIs provided by the SNS providers are the most appropriate solution to HIASN for aggregating the users' social information. HIASN is dependent upon the services provided by SNS for aggregating the user's social data by requesting the APIs (e.g. Facebook Graph API, Twitter Rest API) for gathering the users' recent social data. In brief it takes user as input, and presented user possible social accounts among OSN. Then, upon user's grant, for each account as per the social network, the new social data is collected which is returned by the corresponding API using account's information, encrypted permissions and last request time to discard the already requested data. Upcoming section illustrates each of the above listed phases.

A) Identification of Contributing Attributes

A social network runs a set of services to ascertain a unique identity using publically available attributes. In order to identify a user uniquely, a set of personal information is identified and attempted to determine a set of publicly available attributes which can find the similar account across multiple social networks with the high precision and accuracy. Some of the important attributes common to most social networking sites are listed below [119]:

- ***User ID***

This refers to the unique username/userID or specific handle which identifies a user on a specific social network and allows him/her to sign in. However, most unique, this attribute cannot be used in isolation, especially when a user uses different IDs to distinguish his/her identity across different social networks.

- ***Display Name***

It is the name which displays on the profile information; however, a user may choose to display a phone name at times.

- ***Real Name***

It is the first name-last name pair which a user has used in his/her profile information. Again, the attribute alone cannot be used because two users may have the same names.

- ***Description***

It is the short write-up the user can exploit to introduce him/her. The text can be broken down in to tokens to identify the matching keywords across different descriptions.

- ***Location***

It is the location where a user resides. Location of profiles can directly be compared directly across different social networks.

- ***Profile Image***

This is a thumbnail image provided by the user to pictorially identify him/her on the social network. Nevertheless, this is not a true measure of comparison, as different images can be used to identify on different social networks.

- ***Connections***

Information about connections/friends or followers which are real identity for a network. This should not be an attribute to compare, as this information is no more public for some social networks.

The algorithm for identification of contributing attributes is depicted in figure 4.3.

```
Identification of Contributing Attributes

Input: User_Name to Social Network like Twitter
Output: Features_Extracted

Icr(User_Name)
{
users = searchTwitter(User_Name)
users = searchFacebook(User _Name)
matching_users = search LinkedIn(User_Name)
p_users = preprocessing(users)
p_matching_users = preprocessing(matching_users)
Features_Extracted = Extract_Feature(p_users)
Features_Mapped = Extract_Feature(p_matching_users)
}

Extract_Feature(processed user)
{
For each processed_user in processed_users
    UserID = Extract_UserID(processed_user)
    UName = Extract_UserName(processed_user)
    ULoc = Extract_UserLocation(processed_user)
    Tweets = Extract_Tweets(processed_user)
    For each Post in SN // For Ex. Tweets for Twitter
    P_Loc = Extract_PostLocation(Post)
    return Features
}
```

Figure 4.3 Identification of Contributing Attributes

However, more ‘mined’ attributes are indispensable to govern the search. In addition to the user ID and name which have proved to be most promising attributes to recognize a user, HIASN considers the location of the newsfeed/tweets of the user as an additional attribute to match the user. While considering the location, a weighted score of location and change in location of the user is evaluated. The current work thus uniquely contributes an *Identity*

Resolver which maps user's profiles across various social networks which in turn are correlated and integrated into a single profile by *Profile Integrator*.

B) Identity Resolver Module (IRM)

Since username is the unique attribute for each user across different social networks, it is possible to determine the mapping of user profiles using this as a major attribute. However, mapping the similarity for UserID is a challenging task as users may use different ID's to log on to the network such as email ID, name etc. The IRM employs phonetic encoding score and Levenshtein algorithm for computing similarity between usernames/userID.

The decision whether two profiles are the same or not is taken by the change in the location factor. HIASN extracts the feeds from one social network and finds the change of the location i.e. if the user has covered a distance on the basis of longitude and latitude more than a significant threshold value. The probability of matching a profile increases if the location of the user differs with the same value on another social network. This change in location is mapped to latitude and longitude using Google APIs. IRM computes combined weighted score to determine the location, based on the Euclidean distance [119] between two location using the latitude and longitudes. Any change in the location of the latest activity feed/tweet is mapped resolving the disambiguate user profiles. Thus, the resulting equivalence IRM vector is as given in (4.1) :

$$\mathrm{IRM}_{\mathrm{Vector}}\colon < User\ ID, Username\ , Location > \tag{4.1}$$

Where $\mathrm{IRM}_{\mathrm{Vector}}$ is the final Score for resolving the user's identity?

The raw data is obtained from multiple sources which are highly unstructured. The user profile variables, such as location and the posts, are the text variables and contain noisy information, such as common words, slangs, informal words, and keyword variations. Text cleaning is then performed on these variables.

The algorithm for resolving the identity is shown in figure 4.4.

```
The Identity Resolver Module (IRM)

Input : Features_Extracted of Potential Profiles of Twitter, Facebook, and LinkedIn
Threshold value, t
Output: Matched_Profiles

IRM()
{
 For Each Features_Extracted and
 every Feature_Mapped   MIDS = MatchIDScore (UserID, Feature_Mapped)
{
   MNS = MatchNameScore(UName, Feature_Mapped)
   MLS = MatchLocScore(ULoc, Feature_Mapped)
   Average = (MIDS+MNS+MLS)/3
  If Average > t
    Matched_profiles = Matched_Profiles + 1
return Matched_Profiles
}
}

MatchIDScore(UserID, Feature_Mapped)
{
  P_Score= PhoneticScore (UserID, Matching_UserID)
       if (P_Score > x)
  L_Score =LevenshteinScore(UserID, Matching_UserID)
 MIDS = (P_Score + L_Score)/ 2
 return MIDS
}

MatchNameScore(UName, Feature_Mapped)
{
   P_Score= PhoneticScore (UName, Matching_UName)
       if (P_Score > x)
   L_Score=LevenshteinScore(UName, Matching_UName)
MIDS = (P_Score + L_Score)/ 2
return MNS
}

MatchLocScore(UName, Feature_Mapping)


{
 LE_Score = Comp_Loc(ULoc,Matching_ULoc)
 For each post in SN.posts
   PL_Score = Comp_Loc(P_Loc[i],P_Loc[i+1])
   if PL_Score < y
    MPL_C = FindMatchingPLoc(PLoc[i])
  For each j in MPL_C
    if (PL_Score == Comp_Loc(MPL_C[j],MPL_C[j+1])
     MPL_C1 = Comp_Loc(PLoc[i],Matching_PLoc[j])
     MPL_C2 = Comp_Loc(PLoc[i+1],Matching_PLoc[j+1])
     Change_LocScore = (MPL_C1 + MPL_C2)/2
 MLS= (LE_Score+ Change_LocScore)/2
Return MLS
}
```

Figure 4.4 Algorithm of the Identity Resolver Module

Data reduplication is performed using flexible name matching techniques on entire raw data for the purpose of removal of duplicate content, unification of similar profiles, and enrichment of data metrics obtained from different sources. The name matching algorithms i.e. Phonetic Matching is applied to find names that are phonetically similar. Levenshtein distance technique ensured that variations in user profiles are handled correctly.

The HIASN is strengthened by matching the location field. Euclidean distance is applied to the location field to map the user location by extracting latitude and longitude from Google API. Several sites provide the location when a user posts/tweets on the social network. The module extracts and cleans the location attribute of the profile and the posts and computes the Euclidean distance between two locations of the profile. A change in the location of post more than a threshold value is observed and mapped to another network to verify the similar change. The Euclidian distance between these two locations of the post is calculated and a combined score is considered.

A similarity score is taken as equivalent to mean score of the proposed techniques on User ID, name, and location. A threshold value of 0.85 is derived from manual testing of results. This threshold implies that those pairs having the similarity score of greater than or equal to 0.85 are categorized as relevant matching candidates. In the algorithm 4.2, if x is more than 0.85 means the UName and UID are more similar whereas if y is less than 0.70 means the location is more dissimilar. This change is to be noted in another social network. The variation in the value of x and y is due to the fact that names are more similar than location. The list of the most expected profile of a user is identified and presented to the user. The user is asked to choose the profile which exactly matches an account on another network. Then, the chosen profile is integrated using the multilink structure as discussed in the upcoming section.

C) Profile Integration Module (PIM)

In order to develop a single unique profile for a user, the multiple ontology approaches is employed to model each user data source in combination for integration. It requires the mapping between multiple ontologies to provide a global view to profile. All public attributes of the profile from the different social network are now made visible to the user, the choice of displaying the value of attributes solely depends upon the user. PIM provisions the flexible modification of attributes.

General attributes used in most online social networking sites are personal characteristics, friends, interests, groups, studies, and user created content. A multilink data structure is used to store the information across different social networking sites and provide a global as view. Figure 4.5 provides multilink data structure across multiple online social networks.

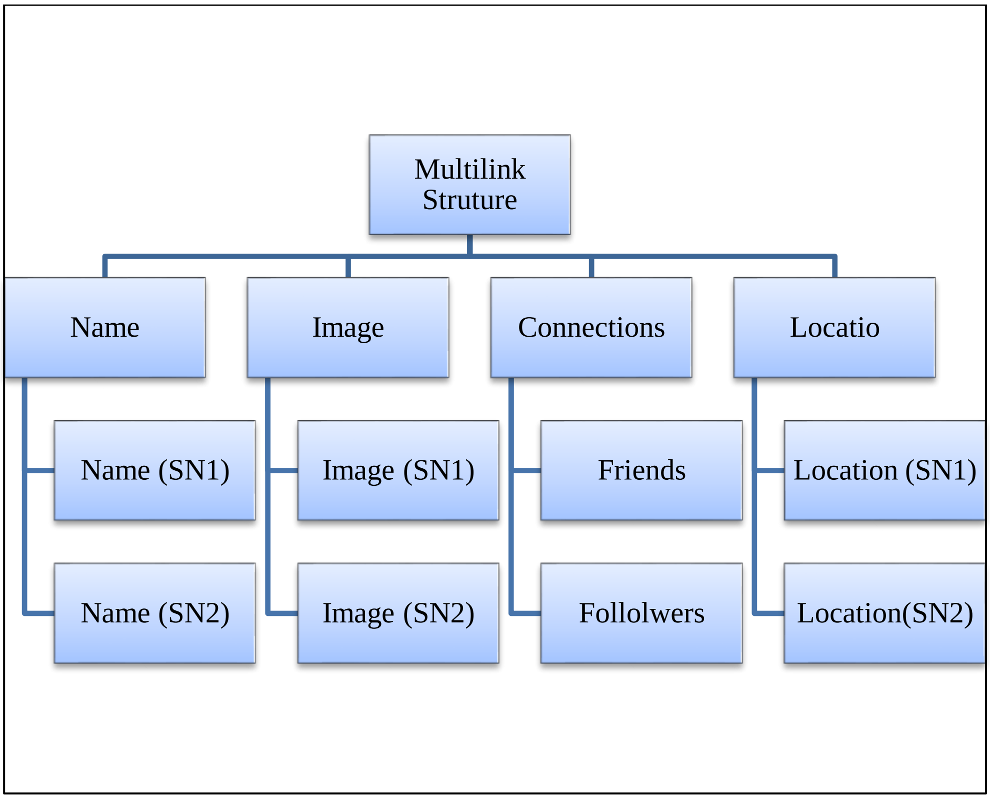


Figure 4.5 MultiLink Structure

If the location is the same for two different social networks, then the system will preserve only one location; if it is different, it will keep both the locations. Noticeably, a generalized identity is being kept by sub-grouping it with individual values of each social network. A user can select any medium of the social network to look for his/her profile and, above all, information is preserved at one place.

The algorithm for integrating the profiles is shown in figure 4.6

```
Profile Integration Module

Input: Matched_Profile
Output  Unique_Profile

PIM(Matched_Profiles)
{
  For Each attribute in Matched_Profiles
        If (att_SN1(value) = att_SN2(value))
 Store att_SN1
        Else
         Create Multi_link_att(att_SN1(value), att_SN2(value))
}
```

Figure 4.6 Algorithm of Profile Integration

The working Engine of HIASN is depicted in figure 4.7. The search vector used for the current search is used as of equation 4.1.

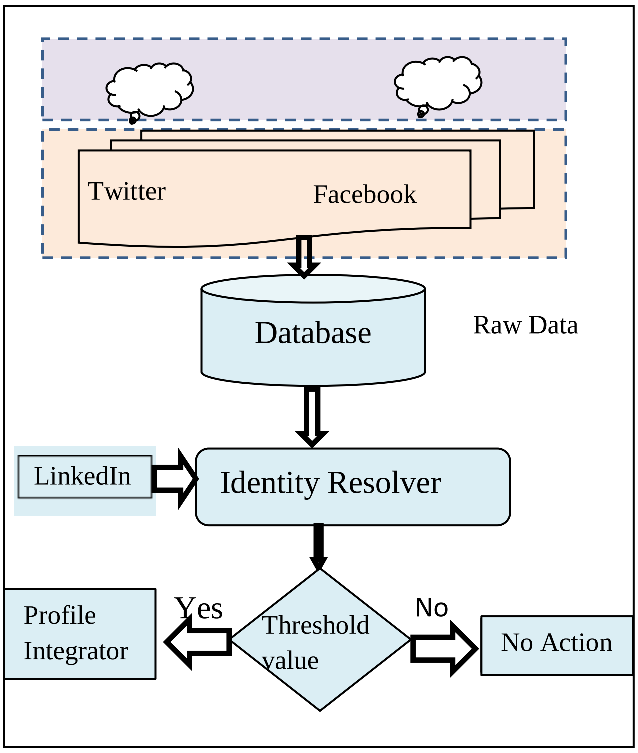


Figure 4.7 Working Engine of HIASN

The data flow for the aggregator explains the integration of social activities across networks and the role of the proposed aggregator in the same. Each SNS would have set of social activities which can be aggregated and collated at one place. Typical shared activities are blogs, newsfeed, applications, notifications, contacts etc. The proposed Aggregator would manage single signing in for all SNS connected by an open OAuth protocol and is represented in figure 4.8.

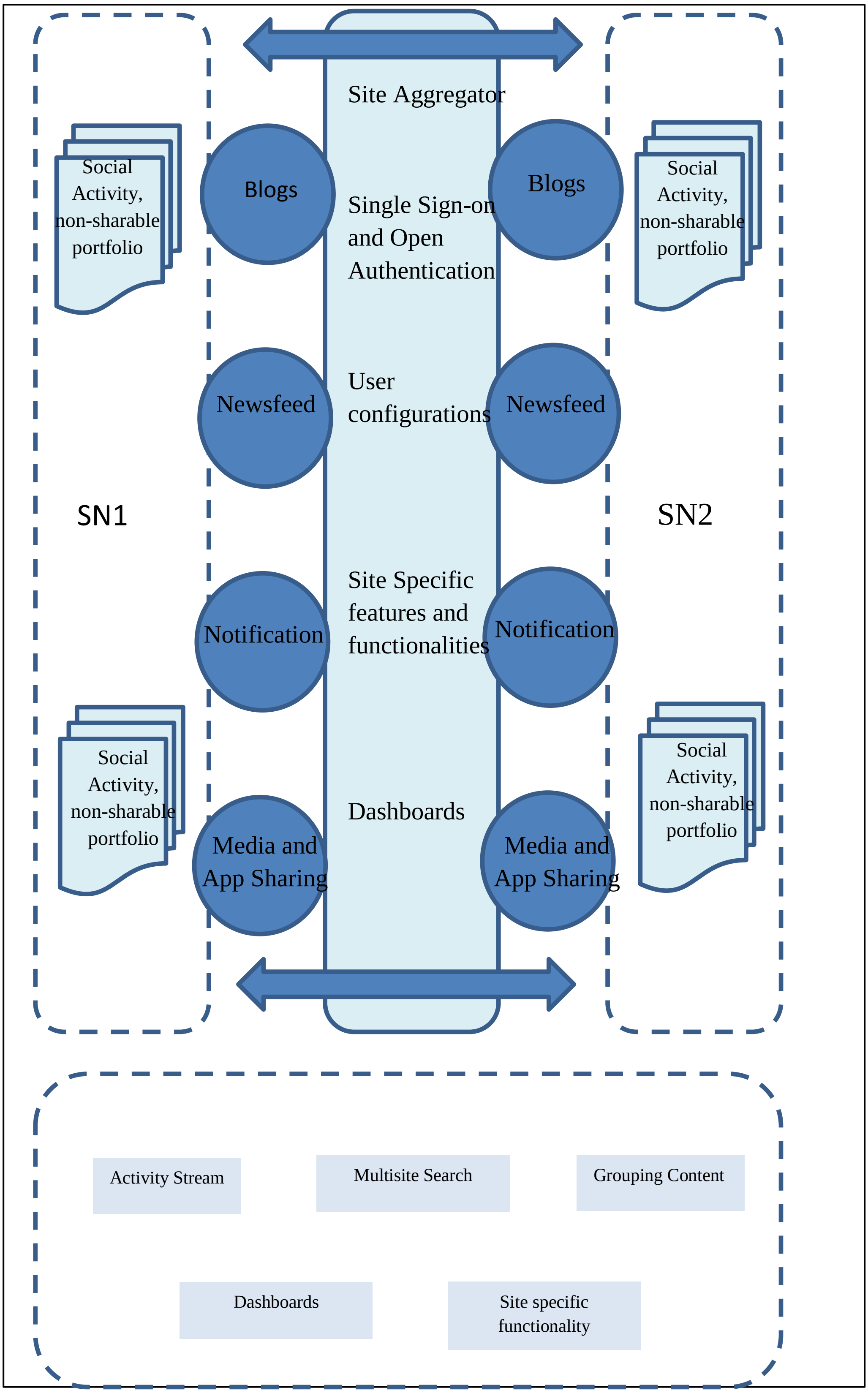

Site Aggregator
Single Sign-on and Open Authentication
User configurations
Site Specific features and functionalities
Dashboards
Blogs
Newsfeed
Notification
Media and App Sharing
Blogs
Newsfeed
Notification
Media and App Sharing
Social Activity, non-sharable portfolio
Social Activity, non-sharable portfolio
Social Activity, non-sharable portfolio
Social Activity, non-sharable portfolio
SN1
SN2
Activity Stream
Multisite Search
Grouping Content
Dashboards
Site specific functionality

Figure 4.8 Data Flow Diagram of HIASN

The OAuth 2.0 protocol is a delegation model used for authentication and authorization of web-enabled applications and APIs. OAuth protocol is used to authenticate each SNS configured on the SNA to track social activities being initiated from these sites. OAuth [181] which is essentially an open authorization protocol is supported by most of the SNS's. SNS's can be added as the trusted partner when user login to the aggregator thereby providing a 'single sign-on' ability. After the user has signed in the aggregator one can see the social networks at the machine. Three network tabs have been provided as default. Authentication to various SNS can be provided using the single sign-on functionality. This is achieved by trusting the SNS. For each SNS (Facebook, Twitter, and LinkedIn) in this case, the user has to click on the tab to authorize them for this SNS. This essentially means that the aggregator is actually marking the said site as 'trusted' and will not ask for the password again to login.

HIASN utilizes the combination of MongoDB and SQL databases where MongoDB is a free and open source that offers flexibility in storing data in JSON-like documents where fields can vary in structure and data structure can be changed over time. The document model maps to the objects of the application and makes it easy to work with Ad-hoc queries, indexing, and real time aggregation. The model has the ability to epitomize hierarchical relationships, to store arrays, and other more complex structures easily. MongoDB has a query language, highly-functional secondary indexes (including text search and geospatial), a powerful aggregation framework for data analysis, and more. Thus, it provides powerful ways to access and analyze the social data with high availability and scalability. Sample queries are depicted in figure 4.9.

Query	**Syntax**
To insert a user's data	db.users.insert({ user_id: 'abc001', age: 35, status: 'D' })
SELECT * FROM users	db.users.find()
To update a user's status to D where age is greater than 30	db.users.update({ age: { $gt: 30 } }, { $set: { status: 'D' } }, { multi: true })

Figure 4.9 MongoDB Queries

MongoDB databases are often recommended instead of SQL databases when dealing with the data portability and the interoperability among different databases. Multi-link structure of user’s profile having multiple social data with different attributes; is a good fit for MongoDB’s flexible data model. SQL databases were built to quickly set up a reliable database and to reduce the development time of group-based content organization model that contain a small number of entities and relationships to provide support in terms of data insertion and request. Various packages/libraries have furthermore been proposed to greatly facilitate the development of MongoDB to support data portability and interoperability over the SQL databases.

JSON is a language independent open standard for text based lightweight data-interchange format which is used for human readable and derived from JavaScript. The results of OSN API’s are stored in a JSON array of objects matching the supplied filters and the search string, in case of Twitter where each object is a tweet and its structure is clearly specified by the object’s fields, e.g., ‘created_at’ and ‘from_user’. The output of Twitter API’s will include both popular and real-time results in the response.

Facebook’s privacy issues require ‘open authorization’ status from users that makes it more complex as a lot of status messages are harder to obtain than tweets in case of Twitter. Facebook has APIs ranging from the graph and public feed APIs to keyword insight API [183] that stores all data as objects which can be accessed by its unique ID that must be known in advance to request the API for the response. The Facebook Graph API search queries require an access token included in the request. Searching for pages and places requires an ‘app access token’, whereas searching for other types requires a user access token. Replacing ‘page’ with ‘post’ in the search URL returns all public statuses containing the search term. Facebook also returns data in JSON format and so can be retrieved and stored using the same methods as used with data from Twitter, although the fields are different depending on the search type.

With increasingly advanced mobile devices, notably smartphones, the content (photos, SMS messages, etc.) has geographical identification added, called ‘geotagged.’ These geospatial metadata are usually latitude and longitude coordinates, though they can also include altitude, bearing, distance, accuracy data or place names. There are four different types of social media feeds discussed in Table 4.1 to specify ‘geospatial’ social data containing a location and time specifications which are generated generally from mobile devices.

Table 4.1 Types of Social Media Feeds

Types of Feeds	Description
Location and time sensitive	Transfer of posts/feeds specifying location and time.
Location sensitive only	Transfer of messages specifying location, which are tagged to a certain place and read later by others.
Time sensitive only	Transfer of posts/status updates to mobile devices specifying time only to increase immediacy
Neither location or time sensitive	Transfer of traditional social media applications to mobile devices specifying neither time nor locations

Figure 4.10 represents the basic schema for the proposed aggregator.

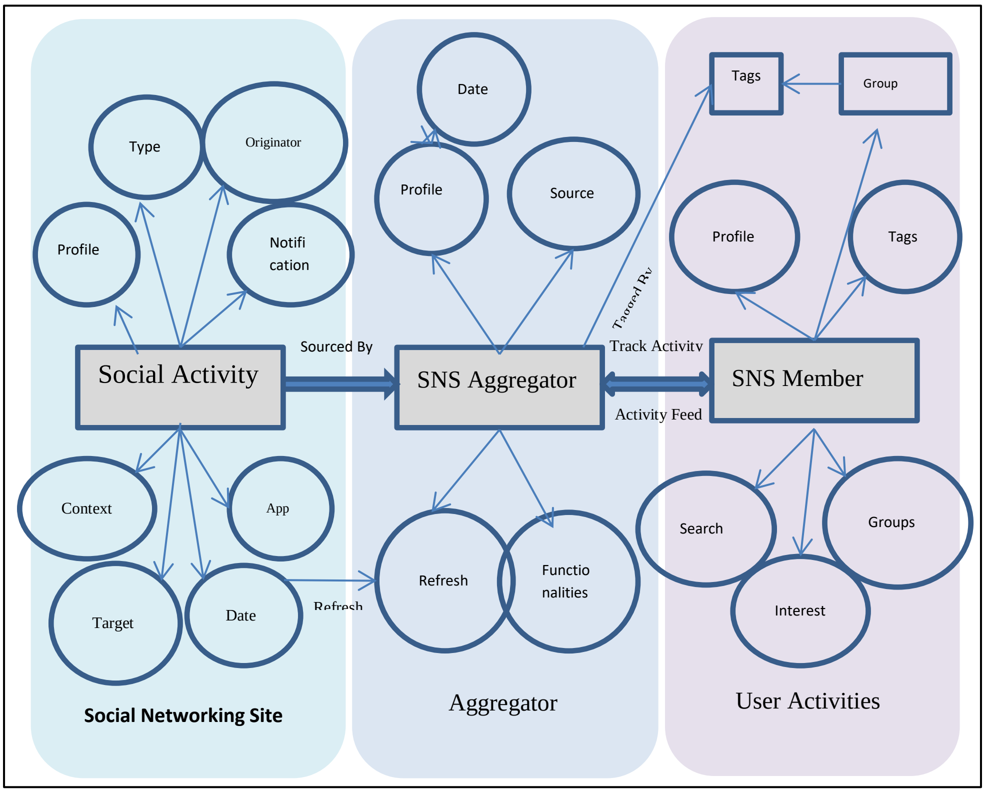
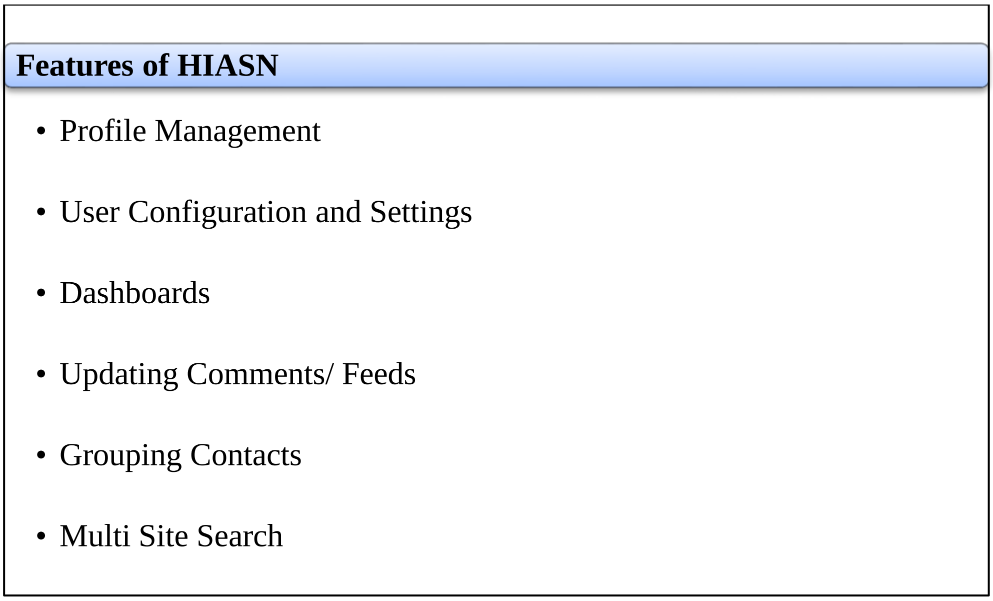


Figure 4.10 Schema Design of HIASN

It contains 5 entity types namely Social Activity, SNS Aggregator, SNS Member, Tags, and Groups. The SNS Aggregator entity is the heart of the schema that represents the data or the input is sourced from various social activities initiated by user profile in the social network site. The aggregator will connect with the activities of the SNS member or the group to which the user belongs in that SNS. HIASN will contain user specific interest and data and the available functionalities. There is another important attribute called source that links the aggregator to the site from which the social activity is generated. The data update will depend on the refresh rate of the aggregator. The Social Activity entity is defined as the actual social activity which is generated by the user and which is in the shared portfolio of the SNS in concern. The tag represents a user-generated mark. These represent contextual information of social data. The upcoming section is dedicated to highlighting the salient features of HIASN.

4.2.3 Salient Features of HIASN

HIASN aggregates the social-network members and social data to share social network activities. The very rationale for having an aggregation is to let the user have a one unified window to manage his social interaction and activities without hopping on each SNS separately. As shown in figure 4.11, HIASN consists of various features:-



Features of HIASN

- Profile Management
- User Configuration and Settings
- Dashboards
- Updating Comments/ Feeds
- Grouping Contacts
- Multi Site Search

Figure 4.11 Features of HIASN

- ***Profile management***

The component maintains the single sign-on for different social networking sites. The user needs a one-time authentication for each SNS; thereafter the data from those sites become trusted. The user needs to login to the aggregator and the data from these trusted sites will bypass the authentication. All content appears in real time (or abstracted to be appearing), which eliminates the need to hop from one social SNS to other. Justification of having aggregation lies in the fact that not every SNS can be the best place for a user having varied interest and hobbies.

- ***User configurations and settings***

This feature includes preferences and interests that can be further configured in the proposed SNA so that user can subscribe to the required set of activities (one wishes to see as activity stream). Various activity streams can be tweets, blogs, news publications etc.

- ***Dashboards***

Dashboards have been provided to display the data based on site of integration for all to have a unified view of all the configured and available functionalities for each SNS.

- ***Updating comments/feeds***

The feature allows regularly refreshing of the feeds from SNS and thus providing real time access to the information. The aggregator provides the real time update of user's status or notification input to the main SNS as well. This has been handled as the activity stream.

- ***Grouping contacts***

The feature groups various contacts from different sites at one place. The profile with same email id which has a presence in multiple sites will be encapsulated together to represent one contact. However, to distinguish the contact notification of the same across various networks, a tag field is used that marks the source of notification or social data. The fact that contact needs to be put together in one place is encapsulated and abstracted view of the SNS integrator as a whole. The system will also sense those profiles which have an overlap in more than one SNS and put them as one entity/contact.

- ***Multi-site Search***

Another unique feature of the proposed solution is a multi-site search functionality which essentially searches a keyword (contact or information) across many configured SNS for that user. This means if one can search for the data from various sites for which the search is configured. For example, a user can search other users of multiple networks at one place i.e. Ram can be searched on Facebook, Twitter, and LinkedIn as well. A similar search can be made for groups as well.

A social network is indeed an abstraction of related groups interacting amongst themselves to develop relationships. The intuitive nature of these social networks is the creation of related groups (or clusters) [184]. This has become an area of interest in the discovery of communities in recent times. These patterns are used to mine a variety of information, which is then used in various fields [191]. Moreover, to analyze any relationships and psychology behind it, clustering plays a vital role. Clustering enhances the predictability and discovery of like mindedness amongst users. The aggregated data requires clustering techniques to group the user’s information as per some interest or topic. The next section proposes Hybrid Ensemble k-means Hierarchical Agglomerative Clustering (HEKHAC) technique that will cluster the users to extract the entities and their corresponding interests as per the skills and location by aggregating user profiles across the multiple online social networks.

4.3 THE HYBRID ENSEMBLE K-MEANS HIERARCHICAL AGGLOMERATIVE CLUSTERING (HEKHAC)

As the number of social network users increases, a tremendous amount of data is generated by the sharing of information. The role of clustering can be observed as summarizing the social phenomenon of communication within a network which can be used in the discovery of patterns and relations. Its main aim is to identify set of users/nodes which have similar content for the purpose of clustering. Applications of clustering includes viral marketing [157][173][184], rating predictions [72][75][92], identifying influential users [100][109] etc. Many researchers have explored the problem and proposed variants of clustering algorithms by the data mining and text mining community [3][18][42] for multi-dimensional data. A people group or community is a subset of hubs inside a system such that associations between hubs in the subset are denser than associations with rest of the system. Detecting a community is a form of clustering of the information which is similar among neighbors. The aim of this section is to propose a method for combining several clusters and generalize this for the user's information. Latent Dirichlet Allocation (LDA) and k-means were implemented to achieve the objective of clustering. A brief description of each algorithm with their limitations is given below:

A) Latent Dirichlet Allocation (LDA)

LDA finds a pre-specified set of |C| clusters within |X| documents. Each term t in a profile with K_i terms then ends up correlated with a cluster C where C = {c1, c2, c3,.. } is the set of n latent clusters which exemplifies coarseness and resulting final set of clusters. The input to LDA is corpora of M documents, each representing i documents (user profiles) that will be a count of all words in corpora in a total of d documents and output is set of similar words of clusters.

B) K-means

As one of the simplest unsupervised clustering techniques, k-means discovers the degree of similarity among k groups assuming k centroids. K-centers are defined and placed spatially as far as possible. Each spatial point is marked to a given data set and associated to the nearest center. New centroids are calculated as barycenter of the clusters and rebounded between same dataset points to the nearest new center.

K-means suffers from major difficulty to predict the value of k and moreover, different initial values of K will result from different clusters. However, the performance of the cluster is good at local but global cluster, didn’t work well. It was observed that k-means do perform better than LDA, overall, both algorithms produce clusters of very poor quality (with respect to user profile). This suggests that profile do not tend to naturally cluster together along topic based lines, and the problem of user aggregated profile clustering is not inherently easy. It has been observed during the research that no work has been devoted to applying ensemble clustering methods in analyzing a user’s publicly available information. However, different strategies have been utilized to recognize community and merge community structures [121]. As data clustering and community detection are very comparative, it ought to be conceivable to merge community in an indistinguishable way from ensembles of clusters with great outcomes.

C) Ensemble K-means

The k-means ensemble clustering emerged as a prominent and viable method that combines multiple partitions generated by different values of k into single clustering solution for improving robustness, stability, and accuracy of unsupervised classification solutions. Many researchers explored consensus clustering [157][158][159] and faced consensus function as a major design challenge in collaborating the clusters. Ensemble clusters provide more robust and stable solutions with lower sensitivity to noise and high scalability. It can also be used in multi-objective clustering as a compromise between individual clustering with conflicting objective functions. Blends of clusters using multiple sources of data or features become increasingly important in the diverse structures of OSN. Several recent independent studies [191][192][193][194] have pioneered clustering ensembles as a new branch in the conventional taxonomy of clustering algorithms. Other related work includes [103][183][195][196][197] but is not limited to these only. Ensemble clustering has few inherent design challenges listed as follows:

- ***Consensus function***

The major design issue is to decide how to combine different clustering solutions? How to resolve the label correspondence problem? How to ensure symmetrical and unbiased consensus with respect to all the component partitions?

- ***Diversity of clustering***

Another point worth stressing is how to generate different partitions? What is the source of diversity in the components? The diversity of the individual clusters of a given dataset can be achieved by a number of approaches. Applying various clustering algorithms, using one algorithm with different built-in initialization and parameters, projecting data onto the different subspace, choosing different subsets of features, and selecting different subsets of data points are instances of these generative mechanisms.

- ***Strength of constituents/components***

How "weak" could each input partition is? What is the minimal complexity of component clustering to ensure a successful combination is one of the other challenges pertaining to ensemble clustering?

Owing to the limitations and challenges presented above, this work proposes a novel algorithm Hybrid Ensemble K-Means Hierarchical Agglomerative Clustering, henceforth referred as HEKHAC and the working of the same is being discussed as follows.

4.3.1 *Working Algorithm*

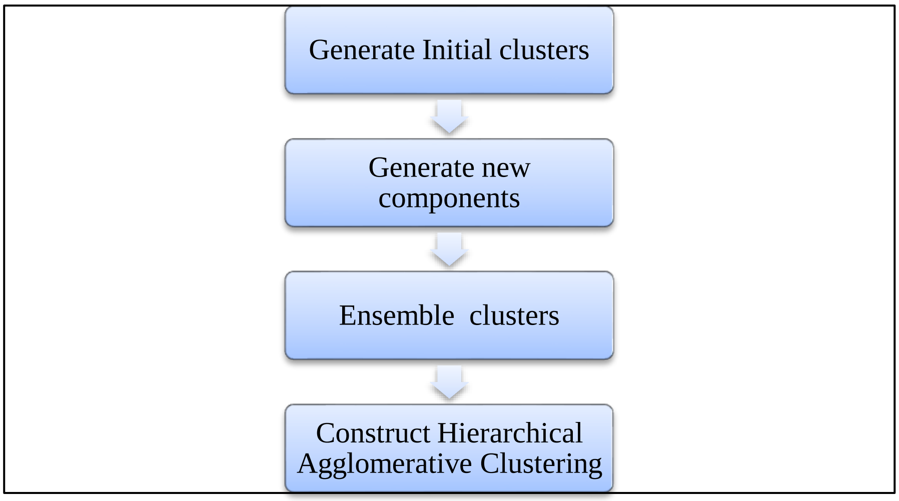

Generate Initial clusters
Generate new components
Ensemble clusters
Construct Hierarchical Agglomerative Clustering

Figure 4.12 The Proposed Architecture of HEKHAC

As shown in figure 4.12, phase 1 is focused on generating initial clusters using k-means for varying value of k while phase 2 considers generating new components using Hungarian algorithm [171], phase 3 ensembles final clusters on the newly generated components and later on phase 4 construct Hierarchical clusters on the ensemble clusters generated during the previous phase to align them with one to one correspondence. Unsupervised training is used to partition data on the basis of similarity using k-means. More similar users are grouped into a cluster using Euclidean distance in this technique across all the profiles aggregated by the network. This results into the creation of clusters belonging to a particular set of attributes.

Basically, ensemble clustering is a two-step process. The first step stores the results of some independent runs of k-means and then specific consensus function is applied to identify the final partition from stored results. Given the multiple clusters provided by k-means, ensemble cluster provides a combined cluster with better and scalable quality of results. The proposed strategy creates a new feature space utilizing the yields of initial k means algorithm.

A particular skill is found and applied for that location. k-means clustering models are applied to the converted list where k = 3 to 12 for skill and by-variance clusters for skill and location to generate input partitions. These techniques are applied separately to the different variables, thus resulting partitions into a different number of clusters. The results of clusters are then combined using Hungarian algorithm and cumulative voting for each cluster. Hungarian algorithm is a multi-objective clustering comprising of multiple clustering partitions with objective functions. It ensembles multiple partitions by combining individual clustering partition and giving a final partition. Final partitions of clusters can be found by applying the voting scheme [172]. This results in a cluster belonging to a particular variable. However, a weighted Euclidean distance is used to cluster the data with more similar attributes. A weight for the user is assigned to one parameter and group.

Further, Confusion matrix [172] is used to compute the similarity between clusters. To compute the confusion matrix of two different numbers of clusters, the remaining cluster of the smaller number of the cluster will be kept as empty.

Confusion matrix for two clusters (A, B) is of size A x B. The (i,j)th index of the matrix corresponds to the object that are in cluster i of A and in cluster j of B. Maximum element is selected using Hungarian Algorithm. Integration of element is done by aggregating the aligned partitions by selecting the element that takes the majority cluster label for each observed partition. Majority voting and plurality voting are the methods to generate the final clusters that involve selecting an object whose count is greater than a threshold value whereas plurality voting considers the majority cluster label for each observed value. The proposed algorithm for generating ensemble clusters comprising phase 1, 2 and 3 is shown in figure 4.13.

Ensemble Clustering

1. *Pass the entire dataset and identify the point with the weight assigned to it.*
2. *Compare the objects and consider it as per k (k = 3 to 12).*
3. *Check the similarity and calculate the mean value from each centroid to the cluster for the object.*
4. *Each object may reside in the cluster it wins the similarity.*
5. *Repeat steps 2 to 4 if there is no change.*
6. *Repeat step for another value of k until K=12*
7. *Compute confusion matrix based on multiple data partitions from step 5.*
8. *Find its maximum element, associate the two cluster as per the maximum object. Thus, reduce the matrix upon removal of these clusters.*

Figure 4.13 Algorithm of Ensemble Clustering

The proposed algorithm improves the accuracy and robustness but the clusters do not have one to one correspondence. The phase 4 of the algorithm performed clustering on aggregated user profiles from various social networks by taking input from the phase 3 to improvise the clusters. The limitation to input the value of k is removed by considering input value of k for different parameters and generating ensemble cluster by combining the results of clusters using Hungarian algorithm and cumulative voting for each cluster. This generated ensemble is given as an input to Hierarchical agglomerative clustering technique to overcome the cluster instability and improving the error rate using.

It consists of grouping data points (or network nodes) iteratively on the basis of the smallest distance measure over all the pairwise distances among the data points. At every pass, the distance between the formed clusters is calculated. The HEKAC algorithm maintains an active set of clusters such that each stage decides which two clusters should be merged. When the two clusters are merged then they are removed from the active set and their union will be added to the active set. This iteration continues until there is only one cluster remains in the active set. The algorithm tree is formed by tracking the clusters that are merged. The clustering in HEKHAC can be evaluated using dendrogram which is a visualization that highlights the kind of exploration which is enabled by the hierarchical clustering over the flat approaches such as k-means. A dendrogram shows the set of data items in one axis and the distances along the other axis. A key point in dendrogram is that the vertical base is located along the x-axis in accordance with the distance between the two groups that are merged. In order to result in a sensible clustering and a valid dendrogram, these distances must be increasing. The distance between the two merged groups must be greater than the distance between the previously merged subgroups. The assemblage of data continues until a single constellation is formed. On the origin of dendrogram, the actual number of clusters can be found out.

Formally the HEKAC algorithm can be stated as follows.

i) Begin ensemble clustering.
ii) Update the distance matrix D by deleting the rows and columns corresponding to the clusters and addition of new rows and columns to the newly formed cluster.
iii) If all data points are in one cluster then stop or else repeat the steps from ii.

It is noticeable that among the advantages of HEKAC the number of clusters is not required to be known in advance. After the clustering is done the bitonic sorting is executed for prioritizing purpose. The aim of HEKHAC is to combine several clusters which are similar to neighbors and generalize this for the user’s information. The proposed strategy creates a new feature space utilizing the yields of initial k means algorithm.

This work significantly overcomes the limitation of providing value of K to K means cluster by taking number of sample values for K and providing an ensemble cluster considering input of different values of K. This states that Ensemble K means clustering had considered objects that were very close to each other into clusters and the Hierarchical clustering will put these objects is in the same direction, hence, overcomes the limitation of one to one correspondence.

4.4 CONCLUSION

The chapter presented the details about an Integrated Query Processing System for Social Web. The proposed work has been carried out in three parts namely, The Profile Integrator, The Clustering Mechanism and The Query Processing Mechanism. The Profile Integrator i.e. HIASN dealt with a complete solution pertaining to aggregation using the proposed hybrid integrator for an autonomous social network. HIASN is used to identify users from one social network to another using an efficient algorithm which strengthens name and location attributes. The system is applied to the real world user profiles extracted from various social networks and aggregators. An integrated profile is proposed that provides a global view to give a single profile to the user. The Clustering Mechanism introduced a hybrid ensemble k-means hierarchical agglomerative clustering mechanisms to group the user's interest. The HEKHAC algorithm offered a competitive rate of convergence. It detailed four clustering algorithms in the context of clustering social network data when aggregated using HIASN. This opens up the scope of further research in regards to efficient use of this information for business and marketing strategies.

The third and final phase of the proposed work is being discussed in the next chapter.

CHAPTER 5

QUERY PROCESSING IN SOCIAL NETWORK AGGREGATOR

5.1 INTRODUCTION

As already mentioned, the proposed work comprises of three main phases: HIASN, HEKHAC and QPSNA where each phase has the specified roles and functions within the design of QPSSN. Chapter 4 detailed the first two phases HIASN and HEKHAC. The first phase, HIASN is a hybrid approach which allows the users to aggregate the profile from multiple social networks. The second component, HEKHAC is meant to organize and cluster the profiles as per user's interest. It now remains to define the QPSNA for accomplishing the proposed design and revolving it into veracity.

The third and final component i.e. the Query Processing in Social Network Aggregator (QPSNA) is proposed to exploit the techniques of natural language for extracting the entities from the query and understanding the semantic meaning of the entities to extract relevant results which form the basis of this chapter.

QPSNA is a system that extracts intelligent information from diverse profiles, provides personalized search considering user's interest and ranks the user profiles using the weighted score in order to rank the most relevant user profile on top that exists among multiple social networks. It aimed at finding information from an unstructured data satisfying user's condition from a large collection of social data.

Technically, the user's profile aggregated from multiple social networks using HIASN which is then clustered as per user's interest using HEKHAC, QPSNA now processes the query (specified by the user), extracts and enhances the entities, identifies the rules to find clusterID (CID) to specify the respective cluster and hence, ranking mechanism has been applied to achieve the objective. In the experiment, QPSNA accomplished promising results.

5.2 QUERY PROCESSING IN SOCIAL NETWORK AGGREGATOR (QPSNA)

QPSNA is covetous systems that consume the information available at multiple social networks and capable of autonomously extracting high quality valuable information from the social web. To search friends, social activities, events satisfying a certain condition give a vision into retrieval of information; hence extending query proficiency into social network is significant. Collecting useful information by searching the social web is a non-trivial, tedious and a manual process. Consider, a list of friends living in the United States. from Twitter and Facebook and the result is an error-prone and fragmentary search. The proposed QPSNA provide a natural way of managing, processing, and analyzing the complex, heterogeneous unstructured data. Designing such a new system that accommodates the voluminous data requires rethinking all aspects of a DBMS, including data modeling, storage management, indexing, query processing and optimization. It shall allow the entire social web to give personalized content or recommendation to the formal queries. QPSNA extracts user's public information and preferences across their online presence. The result of this intelligent search is the direct answer to the user's query instead of the multiple social networks to follow as shown in figure 5.1.

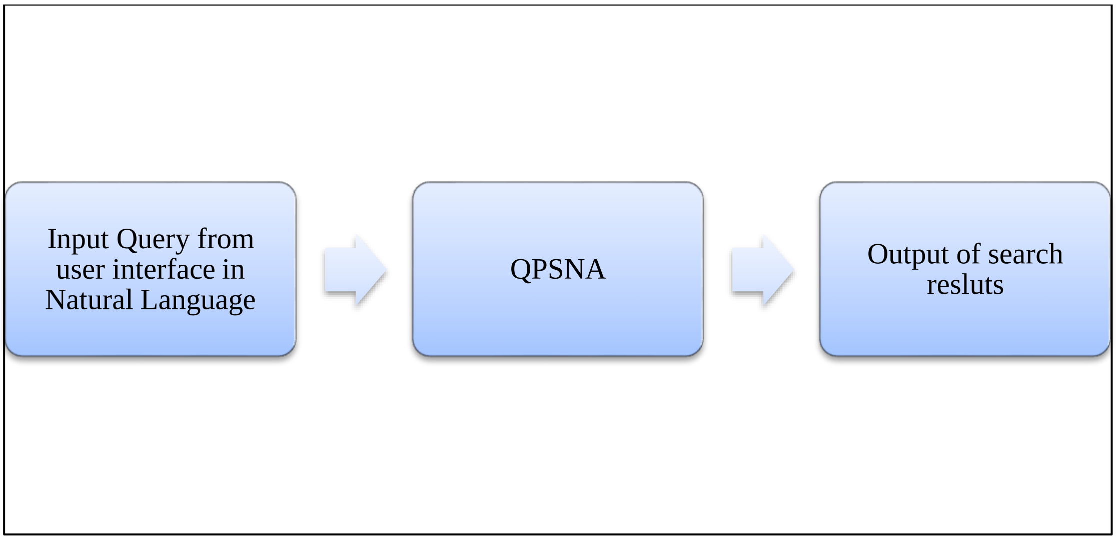


Figure 5.1 Abstract View of QPSNA

The user-centric search shall help users find information like places, skills, users or product that their friends or other people in the network have. It will improve the discoverability of a user in the social network for businesses and implication for many companies.

As outlined in figure 5.2, QPSNA primarily comprises of four modules namely:-

- ***Query Processing System (QPS):*** Extracts the possible entities from the input query.
- ***Content Based Semantic Matcher Maker (CBSMM):*** Maps the entity with the respective ontology to extract the desired output of the user.
- ***Machine Learning Mechanism (MLM):*** Extracts the information required.
- ***Ranking Algorithm***: Sorts the profiles as per user's preference

The high level design architecture is shown in figure 5.2. The details of components are discussed in the upcoming section and section 5.3 provides the detailed case study of QPSNA.

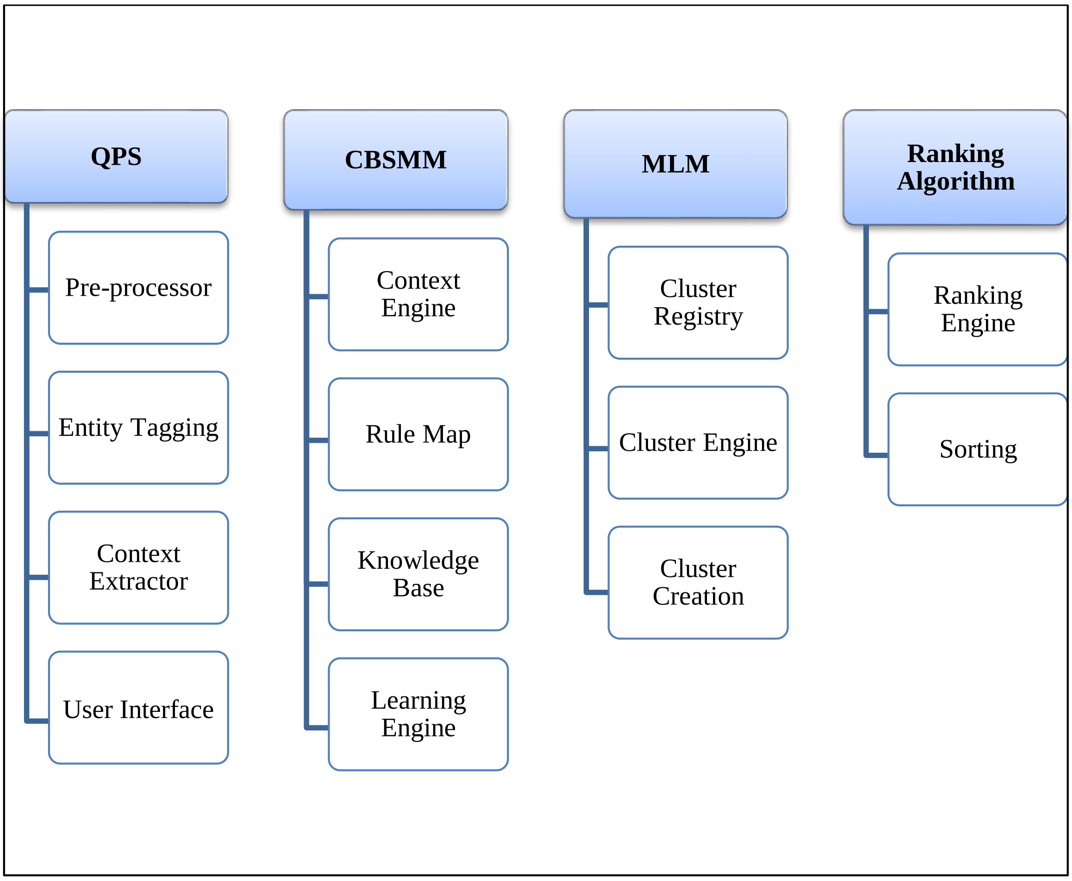


Figure 5.2 Architecture of QPSNA

5.2.1 Query Processing System

Queries written in human language are easy to use and a real world solution for the problem will provide the landscape for custom applications used for intelligent searching and generate user's interest profiles. The primary functionality of the proposed approach is to extract important information from the query by using QPS module. The Primary aim of QPS is to identify a variety of key notions, extract the possible entities from the input query and determine the context of each entity. QPS promises to produce precise entities from the queries by exploiting existing NLP techniques [7]. QPS majorly contributes towards entity recognition (noun phrase) and its extraction; executes a chain of individual phases namely user interface, preprocessing, entity tagging and context extractor as depicted in figure 5.3.

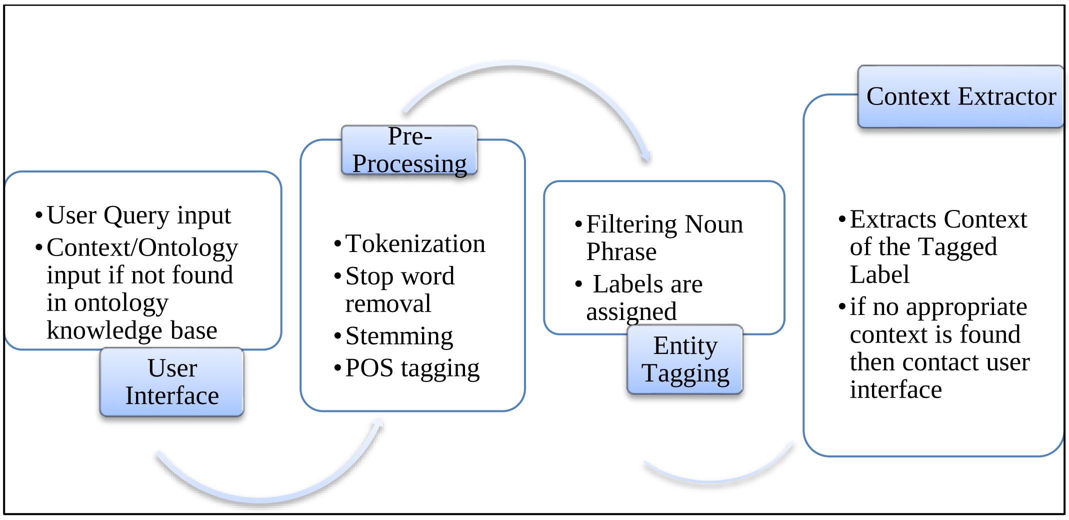


Figure 5.3 QPS Pipeline

- ***User Interface***

The QPS provides a simple and powerful query interface for specifying user queries. It takes user query as inputs from user interface which in turn are pre-processed using parsers and stemmers generating entities.

- ***Pre-Processing***

During pre-processing, query text is tokenized and cleaned (the determiners are characterized by stop words) by removing all stop words. This module finds its space from the list of stop rundown of words which are insignificant for the input [197]. QPS makes use of Morphological Analyzer [185] and Porter's Stemming algorithm [186] to establish a relation between the words and stemming of the words to its root. It utilizes a set of eight domain independent extraction patterns to generate the generic pattern of the entity that can identify a relation. For Example, "Friends living in countries such as United States and China" will consider U.S. and China as Countries. This module overcomes the gap faced in the keyword searching and from based search as a user should not be aware of any ontology or specific query language. The query interface provides simple and flexible way than form-based search as it is not limited to pre-defined query subjects and values. It supports complex queries rather than specifying keywords as input and thus provides more relevant and satisfying results to the user

- ***Entity Tagging***

In this phase, labels are assigned to the connected words produced during the previous phase. An entity tag differentiates the word as noun, pronouns, descriptors, determiners, and verb. About 70% of the query consists of noun phrases [68] which provide an index to the information. Entity tagging is based on Penn Tree Bank Parser [87] that interprets the structure of the phrase. The Penn Tree Bank is about a 4.5 million words dataset that has 15 words on an average in a sentence is which yields about 300,000 sentences providing 96% precision. Entity tagging results into tagged entities that serves as input to the context extractor which in turn returns the ontology of the entities. The robust entity tagging helped the system to improve the retrieval performance of user free form queries. It is also helpful for query expansion and substitution. A precise entity tagging mechanism is a challenging task because the same Part of Speech (POS) [70] tags can be different depending on the specific words involved. Thus, the lexical structure of the query needs an attention to be paid and thus computed to avoid sparsely.

- ***Context Extractor***

Usually, the social web users use informal language and slangs and have no prior knowledge of the underlying ontology. The context extractor tries to find a matching ontology from stored the ontology knowledge base. It is worth mentioning that context extractor of QPS is assisted with an ontology knowledge base containing ontological data to generate pragmatics. In case no matching ontology could be found, the context extractor requests the desired ontology from the user and in turn generates semantic information about the tagged entities. This is achieved by a user interface which does the required integration with the user to get the required context. This is then learned for future reference. Algorithm for QPS is as shown in figure 5.4.

```
Query Processing System
Input: Query in Natural Language
Output: Context

QPS(Query)
{
cleaned_Query = Activate PreProcessing(Query);
tagged_Entity = Activate EntityTagging(cleaned_query);
Context = Activate ContextExtractor(tagged_Entity)
return Context;
}

PreProcessing(Query)
{
Nonword = Identify nonword(query)
If nonword ≠ NULL
    delete nonword tokens
Words = Parse(text)
stopword = Identify stopwords(words)
If stopword ≠ NULL
        Remove stopwords
cleaned_Query = Stemming (!stopwords)//Porter’s stemming algorithm
Return cleaned_Query
}

EntityTagging (cleaned_query)
{
For each keyword from cleaned_query
       tagged_Entity = tag(keyword) // use Penn Tree Bank Parser
Return tagged_Entity
}

ContextExtractor(tagged_Entity)
{
For each tagged_Entity
       Context = search ontologyknowledgebase(tagged_Entity)
      If context ≠ NULL   then
         Return context
     Else
        Context  =  user_interface(tagged_Entity)
       Return Context
}
```

Figure 5.4 Algorithm of Query Processing System

The semantic information thus generated serves as input to CBSMM to enhance the semantic ontological information. The CBSMM promises to deliver more precise results based on semantic search rather than keyword search or form based search. The mapping of context given by QPS and enhanced context by CBSMM results in the resolution of heterogeneity in the ontology.

5.2.2 *Context Based Semantic Match Maker*

CBSMM plays a vital role in QPSNA as it refines the context obtained from the QPS to enhance and enrich the context. This essentially means that the scope of the context obtained will be intelligently increased so that more relevant and wider knowledge is retrieved. The domain knowledge is fed by an inbuilt knowledge base with learning capability. CBSMM follow the rule map to augment a complete and robust context. As shown in figure 5.5, CBSMM comprises of four modules as described next.

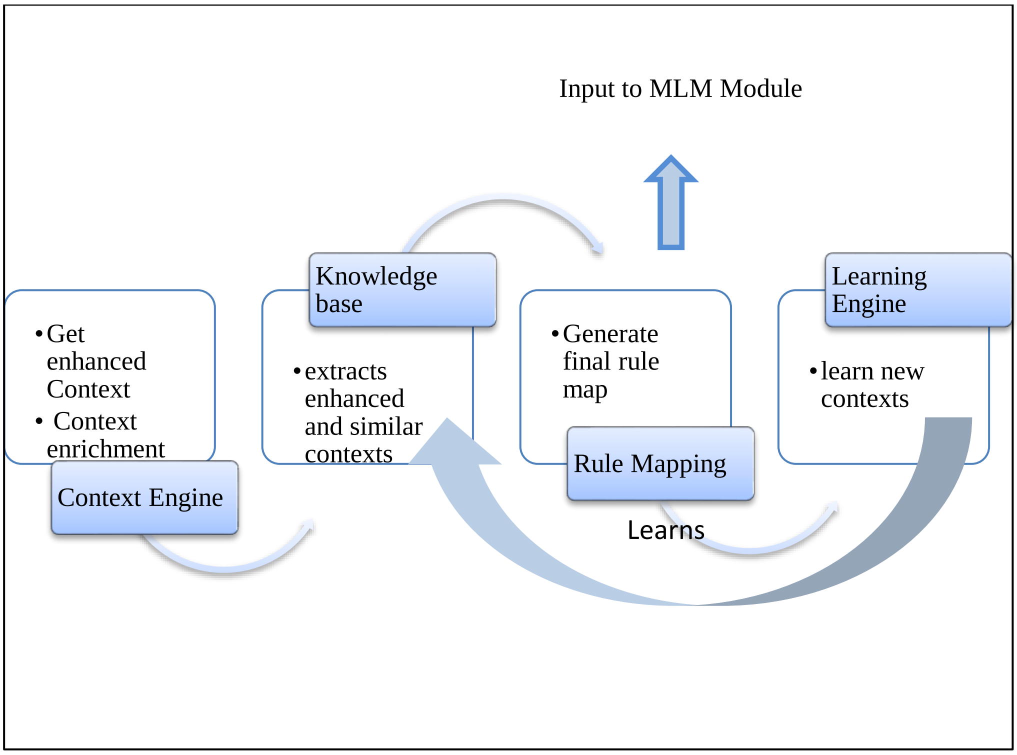


Figure 5.5 CBSMM Pipeline

- ***Context Engine***

Context engine automatically extracts domain independent ontologies from the knowledge base. It analyzes the context obtained from QPS module and fetches the relevant ontologies from the knowledge base. Unsupervised extraction of ontologies imparts with hand cataloging of training data. As there is no dependency on human intervention while using hand cataloging, it recursively develops new ontologies in a fully automated and scalable manner. This is called context enrichment and enhancement because the scope of context will now be improved to include other relevant ontologies as well.

The module measures the semantic orientation of the entities using PMI-IR statistics [187] as it secured a score of 74% when evaluated on Test of English as a Foreign Language (TOEFL) using 80 synonym test questions compared to Latent Semantic Analysis (LSA) that attained a score of 64% using statistical measure of word association on the same set of questions [188]. Based on the measurement, it assigns a probability to the entity to automatically manage the trade-off between the precision and recall. The measurement uses mutual information that indicates the strength of the semantic orientation of the entity.

- ***Knowledgebase***

It is a database where all context and ontologies which the system possess or have learned are stored. It is created using the entities in each rule, sends the query to the social web and applies the rule to extract the information from the resulting users.

- ***Rule Mapping***

The module maps the entities to provide all semantically related terms and context which will eventually increase the scope of the context search. It requires a set of manual training seed that uses a set of domain independent extraction rules to create its set of rules for the fully automatic extraction rules for each entity. This rule map will serve as the input to next module MLM.

- ***Learning engine***

Learning Engine feeds knowledge into the knowledge base for every new context-ontology pair, possibly obtained from the user inputs in QPS module. The working algorithm of CBSMM is illustrated in figure 5.6.

```
Context Based Semantic Match Maker

Input: Context
Output: Rule_Ontology

CBSMM(Context)
{
  Rule_ontology= Activate Context Engine(Context);
  Return Rule_ontology
}

Context Engine(context)
{
  semantic_info = Search KnowledgeBase(context);
  if semantic_info $\neq$ NULL then
  /for each context in semantic_info
  Rule_ontology = map ContextRuleMap(context,semantic_info);
  Return Rule_ontology
 }
```

Figure 5.6 Algorithm of Context Based Semantic Match Maker

5.2.3 Machine Learning Mechanism

QPSNA seems to be a wise solution to the problem of finding a cluster which is associated with an assemblage of relationships and constraints. MLM is the main module in QPSNA that returns the search results based on the refined ontologies as processed by CBSMM Module. The architecture of MLM is elaborated in figure 5.7.

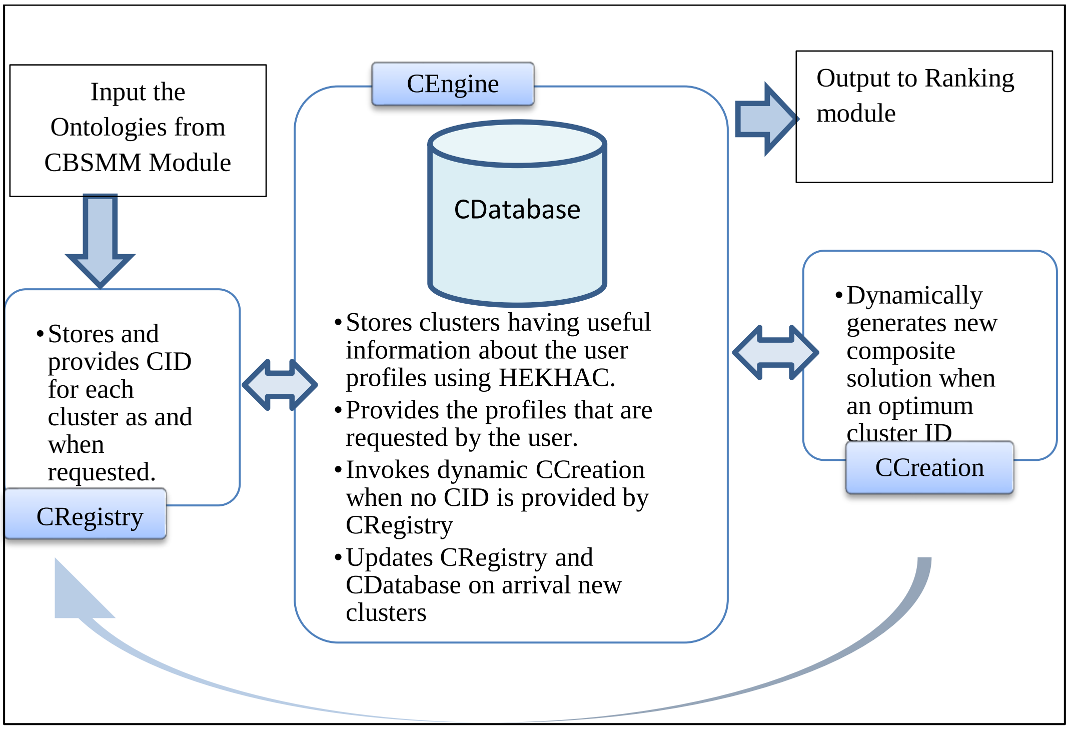

Input the Ontologies from CBSMM Module
CEngine
CDatabase
Output to Ranking module
• Stores and provides CID for each cluster as and when requested.
CRegistry
• Stores clusters having useful information about the user profiles using HEKHAC.
• Provides the profiles that are requested by the user.
• Invokes dynamic CCreation when no CID is provided by CRegistry
• Updates CRegistry and CDatabase on arrival new clusters
• Dynamically generates new composite solution when an optimum cluster ID
CCreation

Figure 5.7 MLM Architecture

Each context is searched in cluster registry to obtain appropriate matching cluster(s) that matched user's criteria. In case there is no appropriate cluster available, the MLM dynamically generates a cluster. The MLM phase is the third phase of QPSNA which supports the design of Intelligent Learning mechanisms. MLM has four components that return the set of user profiles. The following states the process of MLM.

- ***CRegistry***

It maps the context to cluster id (CID) that is related to a particular context. This CID is then used to fetch the information from the CDatabase. This ID is generated using an inverted index for each aggregated user and social information available from multiple SNS which will be helpful in extracting useful information. The primary step in the process of generating CID is to list the fields to analyze and transforming the attribute value into index terms. The fields which will be indexed are social network name, user name, social information, and timestamp. CID is stored to retrieve the user's information and social data.

- ***CEngine***

The CEngine is associated with CDatabase and invokes particular cluster from the CDatabase on the basis of CID given by CRegistry. It also entreats the CCreation if a relevant CID is not returned by the CRegistry. The cluster generated from the CCreation is then stored in the CDatabase, an id is allocated and stored into CRegistry.

- ***CDatabase***

It consists of clusters of user profiles having useful information about the user. Cluster analysis, or clustering in a social network context, is the grouping of a set of data objects (for example, friends, connections, communities, or personal information) in such a way that objects in the same group (or clusters) are more similar to each other than to those in other groups (or clusters). The identification of these patterns into clusters has numerous applications in the field of data science. There are various algorithms that can be used to cluster data

[192][193][194]. Popular clusters include groups with small distances between cluster members, dense areas of the data space, intervals, or particular statistical distribution [163]. Therefore, clustering can be formulated as a multi-objective optimization problem. A suitable clustering algorithm and parameter settings vary from the individual input and expected results. CDatabase is formed using HEKHAC. It also manages, synchronizes and collaborates with each other to find the composite solution when an optimum CID is not returned by the CRegistry. The algorithm of MLM is depicted in figure 5.8

```
Machine Learning Mechanism

Input: Rule_ontology,query
Output: Expected_Users

MLM(Rule_ontology,query)
{
  CID=search CRegistry(Rule_ontology)
  If CID ≠ NULL then
    Cluster_user_profiles = search CDatabase(CID)
  For each user = user1 in cluster_user_profiles
    Accuracy_Score = calculate accuracy(user1,query);
  If (threshold score>Accuracy_Score);
   Return user1;
  Else
   Dynamic_solution = Activate Dynamic_Cluster(users,query);
   Return Dynamic_solution_user1;
}
```

Figure 5.8 Algorithm of Machine Learning Mechanism

The algorithm for generating clusters is shown in figure 5.9.

```
Generate_Dynamic_Clusters

Input: User_Profiles
Output: Clusters

Dynamic_Cluster (users,query)
{
  Clusters = Activate HEKHAC_Clusters(users,query);
  Clusters_info = Extract Info(Clusters);
  Register Clusters and Clusters_info in CDatabase and assign a CID;
  Register CID in CRegistry
          If CID ≠ NULL then
                    For each Dynamic_Solution_user = user1 in Clusters;
                          Accuracy_Score = calculate accuracy(user1,query);
                          Threshold_score=0.85;
                  If (threshold score>Accuracy_Score);
                Return Dynamic_Solution_user1;
}
```

Figure 5.9 Algorithm of Generate_Dynamic_Cluster

- ***CCreation***

It creates the clusters dynamically to fetch the relevant data if the CDatabase does not have a cluster corresponding to the ontology, cluster is created and an entry is also made to CDatabase and CRegistry.

5.2.4 Ranking Algorithm

Ranking of user profiles is required to retrieve a specific user from pool of users. For example, if a user is interested in "*friends living in the United States*" then it is an essential property for the searched user to be his/her friend and living in the United States with some additional attribute such as New York or the number of mutual friends on top of the results. Thus, it is important to rank users as per the importance of user profile. This module ranks the user profiles resulting from the MLM on the basis of the weighted score so that most relevant profile results first to accomplish the phase 4 of QPSNA. The weight can be observed as the number and quality of all attributes that are linked to the user. The ranking architecture encompasses two major components to rank the desired documents and is shown in figure 5.10.

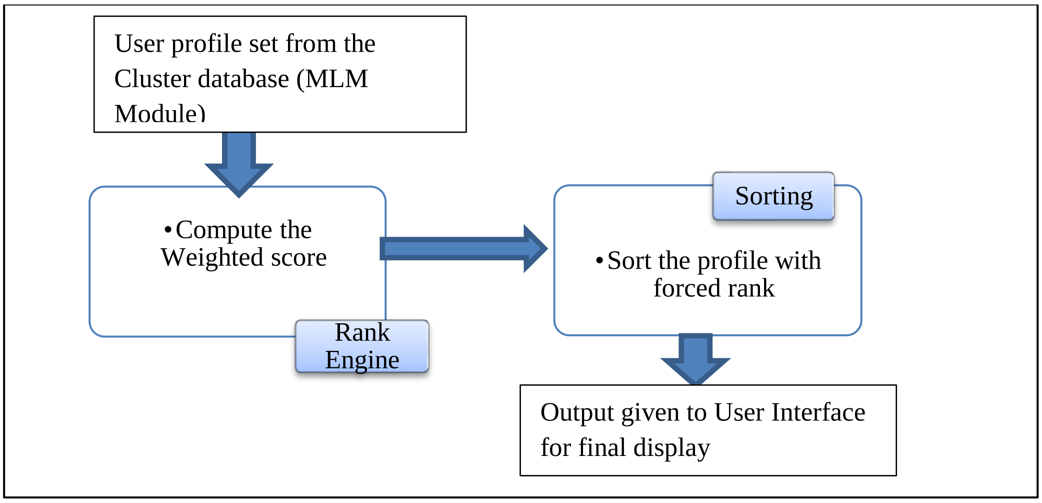

User profile set from the Cluster database (MLM Module)
•Compute the Weighted score
Rank Engine
Sorting
•Sort the profile with forced rank
Output given to User Interface for final display

Figure 5.10 The Ranking module

- ***Rank Engine***

The Rank Engine computes the rank or weighted score based on the user-to-user interaction. The user interaction could be further categorized as user-group, user-user and mutual friend's interaction as depicted in figure 5.11.

```
Ranking Algorithm

Input: Expected_Users
Output: Sorted profiles

Rank(Expected_Users
 {
            user1 = user1_info();  // the user who input the query
For each user= user2 in Expected_Users  c.C
            User_Interaction = Activate User_User_Interaction(user1,user2);
           Group = Search_Common_Group(user1,user2)
           For each Group.i in Group
            Group_Interaction = Activate Group_User_Interaction(Group,user1,user2);
               Group_Interaction_Score=(Group_Interaction1 + Group_Interaction2 + ---
                                          + Group_Interaction n)/n
           Mutual_friends = Activate Mutual_Friends(user1,user2);

Weighted_score=(User_Interaction+Group_Interaction_score+Mutual_Friends)/3;
             Sorted_Profiles = Sort(weighted_score);
Return Sorted_Profiles;
 }

  User_User_Interaction(user1,user2)
{
For each post or tweets in user1
If  Likes(user2) or follows(user2) then
          User_Interaction =User_Interaction+1
For each post or tweets in user2
If  Likes(user1) or follows(user1) then
          User_Interaction =User_Interaction+1
Return User_Interaction
}

   Group_User_Interaction(Group,user1,user2)
{
For each Group in user1 and user2
       For each post in Group
                If  Likes(user1) and Likes(user2) then
                            Group_User_Interaction =Group_User_Interaction+1
       Return Group_Uer_Interaction
}

   Mutual_Friends(user1,user2)
{
For each SN user1 belongs
                  score=Extract_Common_Friends(user1,user2)
Mutual_Friends=(score1+score2___+scoren)/n
Return Mutual_Friends
}
```

Figure 5.11 The Ranking Algorithm

Here, an interaction of users in a group determines the common groups between the searcher and users from the seed. If a common group exists, then the user who frequently interacts and has recently interacted with the group will be more important than the one who least interacts. User-to-User Interaction comprises of the user who interacts more with the searcher and is preferable to the user. The recent interaction between the searcher and the user will be given a high edge over another. In case of mutual friends, the searcher will be more interested in the target user who wins more number of common friends. The Rank of the resultant user's profile is sorted by using Bitonic Sort discussed later in this chapter.

- ***Bitonic Sorting Algorithm***

The sorting module sorts the user profile on the basis of weighted score discussed as above and provides the output to the user interface. The bitonic sorting [189] is a comparison based sorting algorithm that computes in parallel. This algorithm converts a random sequence of numbers into a bitonic sequence. The bitonic sort can be modeled as a kind of a sorting network. At first, the unsorted sequence is built into a bitonic sequence and then the series is split into smaller sequences till the inputs get aligned in sorted order. A bitonic sequence of n elements ranges from x_0 to x_{n-1} with the following characteristics:

i. The existence of the index, where i, $0 \leq i \leq n-1$ such that the increase or decrease of the sequence from x_0 to x_{n-1}
ii. The existence of the cyclic shift of indices by which the characteristics satisfy.

The bitonic sequence occurs after applying the above mentioned characteristics. The bitonic operation is applied for producing two bitonic sequences on which the merge and sort operation is applied. Sorting networks are the comparing networks for sorting out of inputs. The bitonic sorting network is the flexible and advanced approach way of inclusion network from the comparison elements that provides enormous speedup among parallel and sequential odd/even and rank sort algorithms. A single network can lodge the inputs lists of variable dimensions and modularity in which an outsized network can be split into several indistinguishable modules. The processing steps followed in the formation of a bitonic sequence in this current research are given as follows:

- The input database is extracted from the social network by appropriate query processing. The extracted database undergoes hierarchical agglomerative clustering in which, the data are grouped as per the attributes given.
- Now the merged data are sorted out by using bitonic sequence. In bitonic sequence generation, the starting sequence is fixed is predefined for Facebook profiles and twitter profiles. The rest of the combination is formed according to the binary combination of the social network and the attributes. For each of the four skills and locations, the respective binary combination is formed.
- After the bitonic sorting is done, the sorted profiles are aligned as per the priority level of the user profiles. The user profile which satisfies all the attributes given in the query will be prioritized first, and then follows the priority level. After the sorting is carried out, the prioritizing of the profile according to the attributes is done.

The QPSNA is the most important phase of the research as it has the responsibility for mining information about user's interest. QPSNA implemented text-valued natures of the social data available on OSN and indexed every piece of social data which makes it searchable by user's queries which can be keywords or phrase. The extracted information is then evaluated by weighted score using the ranking algorithm and ranked by the bitonic sort mechanisms such that the top scored information will be deliberated as contents of interest to be provided to the user. The upcoming section illustrates the working of QPSNA with a case study.

5.3 THE CASE STUDY

Before testing the QPSNA experimentally, an analytical study to find out the complexity and accuracy of the entire module is being carried out in this section. The algorithm was given the input query "Friends who know java" as depicted in Table 5.1 and the description of each step by step output thus obtained is as illustrated below.

Table 5.1 Input/Output of Case Study

Input	Output
Friends who know Java	Sorted User Profile set

- ***Processing in QPS Module (see Table 5.2)***

Table 5.2 Input/Output of QPS Module

Input	Output
Friends who know Java	Keyword-Ontology Pair as shown in Table 5.1.

Following steps are followed

- The input string in tokenized into keywords: Friends, who, know, Java
- Stemming is performed to remove the filler words. Resultant keywords are Friends, Java
- Keywords are tagged as Noun
- Friends is assumed to be a known keyword and Java being an ambiguous keyword is tagged to Skill, Location, Name ontologies
- The user was asked to suggest ontology for Java, and the user choose the ontology Skill as depicted in Table 5.3

Table 5.3 Keyword-Ontology Pair

Keyword	Ontology	User Input
Friends	Friends	N
Java	Skill	Y

- ***Processing in CBSMM Module (see Table 5.4)***

Table 5.4 Keyword/Ontology Pair

Input	Output
Keyword-Ontology Pair	Enhanced Related Terms as shown in Table 5.5 and Rule

Following steps are followed:

- Context Engine takes the relevant ontologies as input.
- Context Engine searches the refined ontologies from the Context Knowledgebase and results in some related ontologies.

- Rule map has created which maps the keyword to other Ontologies. Let us assume that our rule map defines the following for the keyword Friends (language) is as shown in Table 5.5**.**

Table 5.5 Enhanced Related Terms

Keyword	Ontology
Friends	Friends: Mutual Friends, Tagged friends, Shared links etc.
Java as a skill	Skill: Eclipse, Core Java, Beans , EJB, Sun Microsystems

- ***Processing in MLM Module (see Table 5.6)***

Table 5.6 Input/Output of MLM Module

Input	Output
Input: Enhanced Ontology and Rule: Friends(language)	*Output*: User Profile Set {P1, P2,P3,P4,P5}

Following steps are followed:

1) Search the CRegistry for CID matching with Ontologies and the result of CRegistry is as shown in Table 5.7.

Table 5.7 CID

CID	Ontology
C_Friends	Friend
C_Java	Java
C_Eclipse	Eclipse
C_Beans	Beans
C_sun	Sun Microsystems

2) CEngine retrieves the relevant user profile set from the CDatabase as shown in Table 5.8.

Table 5.8 Output Of Cluster Engine

Cluster ID	User Profiles
C_Friends	P1, P2
C_Java	P2, P3, P4
C_Eclipse	P2,P4
C_Beans	P2,P4
C_Sun	P2,P4,P5

- ***Processing in Ranking Module***

Table 5.9 Input/Output in Ranking Module

Input	Output
Input: User Profile Set {P1,P2,P3,P4,P5}	*Output*: Ranked User profile set {P2,P4,P3,P1,P5}

Following steps are performed

1) The weighted score for each user profile. Demo values are shown in Table 5.10

Table 5.10 Weighted Score

User Profile	Weighted scores
P1	0.2
P2	0.8
P3	0.5
P4	0.7
P5	0.2

2) Sorting of user profiles based on scores {P2,P4,P3,P1,P5}

QPSNA finally generated sorted profiles satisfying user's needs and interest of querying the system. It is worth mentioning that the proposed system succeeded in extracting the information of user's interest and provided encouraging results

5.4 CONCLUSION

The chapter presented the details about the third phase of the proposed work. The working model for the same will have a mechanism to input user query in natural language and produce result using QPSNA module. In fact, the proposed query processing system for social network aggregator is a complete solution pertaining to identifying the entities from the query, enriching the semantic meaning of the entities, applying machine learning techniques and finally ranking the results of the relevant profiles matching the criteria of the user. The ranking algorithm offers to sort the user profiles as per the preference of the user to ensure the maximum satisfaction of relevant results to the user. Next chapter presents the results thus obtained and also a detailed discussion concerning the results.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 INTRODUCTION

The social web plays an important role in identifying and establishing connections among individuals through the social networking sites and reports indicates that the users have multiple accounts at multiple online social network services. Having multiple accounts is not the primary issue however; keeping track of the content/contact or other social activities generated by the user is of major concern. In order to address the concern thus raised, an *Integrated Query Processing System for Social Web (QPSSN)* is designed with the intention to collate/aggregate/organize the data spread across multiple social network services. The idea is to organize and ease the information retrieval process for a user maintaining multiple social networks actively. Different modules of QPSSN are able to mine interpersonal information, gather profile information from various social networks and handle the query written in a natural language. The work is then empirically tested and results thus obtained are promising.

The chapter begins with highlighting the problem statement under consideration during the course of this research study and later it explores parameters suitable for evaluating the proposed algorithms and the entire model. The upcoming sections illustrate the implementation details followed by detailed discussion about results and also describe the experiments on the data sets collected from multiple social networks as discussed in the previous section.

6.2 THE PROBLEM STATEMENT

With the numerous benefits of OSN, various issues and challenges were raised in the literature review, some of which are an exceptionally convoluted and require an extensive variety of approaches and solutions. In this thesis, we tended two specific issues i.e. *Information Overload and Walled Gardens*. These two issues keep the users away from completely using and profiting by the abundance of the data accessible on OSN.

An aggregated social network should be able to identify and extract a unique user profile across social networks so that results can be fetched as per user's interest. Further to the need of extracting the information we realized a need of ontology based search so that the context of the search should be matched with user's preferences of results. A social search has to be intelligent and should be able to learn the user's need at the same time delivering results in real time. This motivated us to include machine learning capabilities in our architecture and clustering for faster and real time retrieval.

Finally fetched results have to be properly ranked to have a more relevant search at the top. The literature review suggested many issues in searching like user friendly and easily executed and abstracted. This is addressed in our implementation with the help of letting the user input the query using natural language.

The users have a ton of challenges ranging from an adjustment of all approaching data, finding extra data from outside of their companion cycles to impart the interesting contents to their diverse gatherings of interest. A novel and unique approach are thus proposed that can help the users to overcome such difficulties.

As mentioned already, QPSSN is divided into three phases namely, HIASN, HEKHAC and QPSNA to aggregate, cluster and extract user's information respectively from multiple social networks. The user is required to register in QPSSN. QPSSN then fetches his/her profiles from other social networks based on attributes like name, UserID and location attributes of the network. This then forms a unique view of the overall aggregated network of the user's existence at multiple networks. A user can perform search on this aggregated network using natural language.

Query processing system for social network aggregator extracts keywords from the query and matches a relevant ontology from knowledgebase (user may be asked to provide correct ontology if no relevant results are obtained). Ontology based search is executed to fetch out clustered results from cluster database finally ranking results in priority order as shown in Figure 6.1.

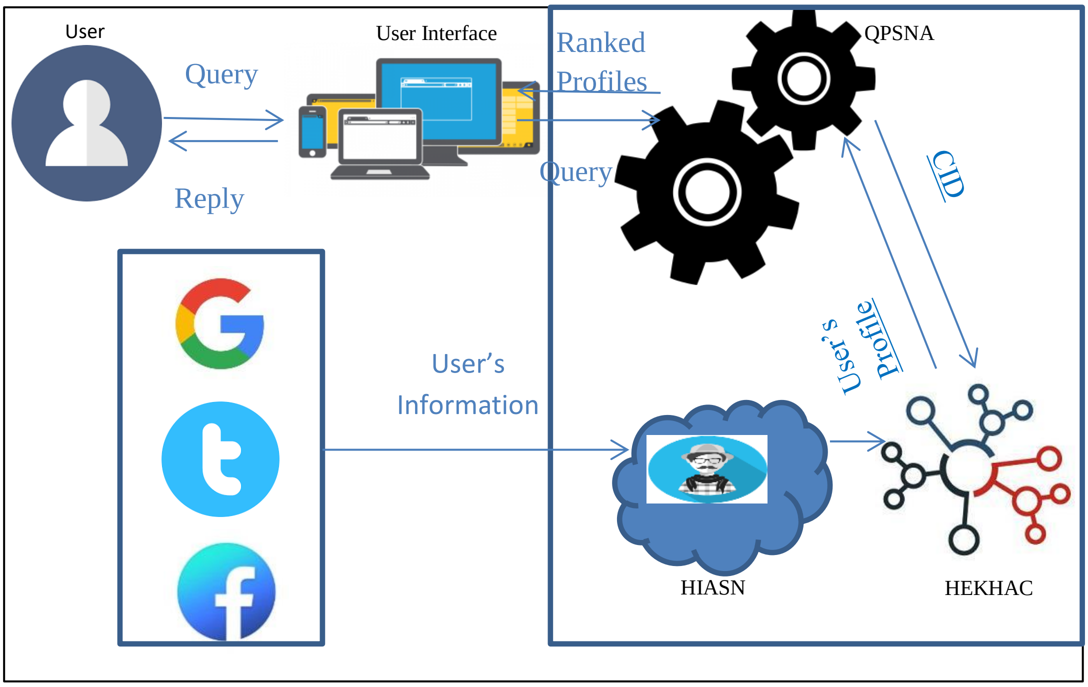


Figure 6.1 The Integrated QPSSN

The upcoming section illustrates the implementation details of the proposed work.

6.3 IMPLEMENTATION DETAILS

QPSSN is a system that extracts intelligent information from diverse profiles and provides a single image of the user by integrating profiles that exists among multiple social networks. QPSSN is implemented on Intel Core i5 with 8GB RAM using Windows 7 Operating System using MATLAB(R2014A). Data processing and data modeling, e.g., regression analysis, are straightforward using MATLAB, which provides time-series analysis, GUI and array-based statistics. MATLAB is significantly faster than the traditional programming languages and can be used for a wide range of applications. Moreover, the exhaustive built-in plotting functions make it a complex analytics toolkit.

The data provided by multiple SNS varies from one network to another such as Facebook provides more information than Twitter. Table 6.1 corresponds to the user's attributes that can be collected from multiple social media using API's provided by various SNS.

Table 6.1 Social Data Accessible from Multiple SNS using API’s

Attribute	Facebook	Twitter	LinkedIn
Nickname	✓	✓	✓
First Name	✓		✓
Last Name	✓		✓
Full Name	✓	✓	✓
Profile Photo	✓	✓	✓
About	✓	✓	✓
Email	✓		✓
Homepage	✓	✓	✓
Location	✓	✓	✓
Gender	✓		
Birthday	✓		✓
Relationship status	✓		
Language	✓	✓	✓
Affiliations	✓		✓
Education	✓		✓
Interest	✓		✓
Groups	✓		✓
Contacts	✓		✓
Social connections	✓	✓	✓
Posts	✓	✓	✓

The data collection system extracted different data metrics from different social media platforms with the help of input search queries. For instance, in this case, 20 different user-skills as the input queries such as “java”, “python” and “mongo” etc. are taken into consideration. The system collected different data metrics such as user-name, user-location, user-description, gender, birth, connections, tweets, etc. from multiple social media platforms by using user’s information from the online social media API‘s. User-level data and user demographics using Twitter public search are then gathered. The entire collection is carried out on 67,956 documents. The mixed inputs of user-variables in the Bing Search API’s are also used to collect the information of around 87,734 links out of which 35,413 are user profile links, however, only 26,543 of them had exact matches with the input-queries of the Twitter public search. QPSSN also used Facebook to extract the alternative user-profiles of a twitter user. The system has also used Facebook and LinkedIn to extract the user-profiles. The system collected total 24,341 user profiles from Facebook and 20,580 of LinkedIn users. Table 6.2 summarizes the ground truth set of user's accounts on OSN.

Table 6.2 Ground Truth Data

OSN	Crawled profiles
Twitter	26,543
Facebook	24,341
LinkedIn	20,580

In order to select a random set of user accounts in an OSN, methodology similar to [126] where random Twitter userId's are generated is used. For the purpose of research, we have just collected publically available data. Note that, the system collected only publicly available data available on social networks and does not engage in any user authorization asking for private data. The user's information is collected using access token and a prior approval from the users is taken to use the user's data for the research purpose only.

For the purpose of research, we choose a Twitter profile and searched out a data set of matching user profiles across other OSN. An equal number of negative instances, by randomly pairing a username set of the positive instances which are known to belong to different users. We extracted features from positive and negative instances and used features in an engineered framework that effectively classifies username sets as same or different users.

It is worth mentioning that due to the restriction on attributes offered by API's of different social networks, QPSSN could only use limited features of the profile. Social data corresponds to the data that user pushes on the network and the data that a user reads from friends which is a sum of the user's social stream and includes information like profile information, friends, posts to name a few. Figure 6.2 represents information which displays available information of attributes on OSN for users which exist only on individual networks and for overlapping users like

Facebook, LinkedIn, and Twitter. The figure 6.2 depicts that the availability of attributes varies among social networks and userID, Location, and Name are the most prominent available attributes across multiple social networks. The results unveil the reason of vector chosen for the aggregation of the profiles of overlapping users of who exists at multiple social networks.

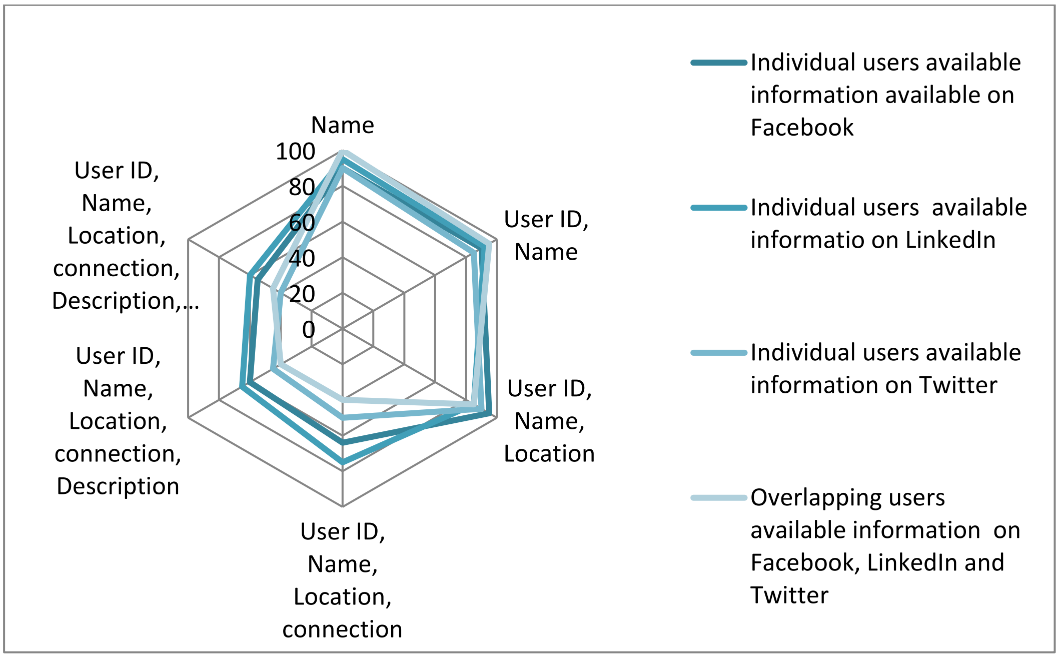


Figure 6.2 Information Available about Users

The training and testing dataset for evaluation of QPSSN is constructed on the basis of initial manual mapping for the user who overlaps on the networks. The users who didn’t match are taken as negative pairs for calculating precision and accuracy. For evaluating the result of QPSSN is applied to the dataset to establish the accuracy, precision, recall, and F1 respectively with the user’s aggregated profile. The precision is the fraction of retrieved users that are relevant to the query and Recall denotes the fraction of the relevant users that are successfully retrieved.

A) Evaluation scale

1) During the course of implementation of the QPSSN, the following three vectors could be observed and hence implemented to identify the overlapping profiles of the users. Here, the vectors represent:

$$\mathrm{IRM}_{\mathrm{Vector}}: < User\ ID, Username\ , Loc > \qquad (6.1)$$

$$\mathrm{Vector}_{\mathrm{v1}}: < User\ ID, Username\ , Loc, description, image > (6.2)$$

$$\text{Vector}_{\text{v2}}: < User\ ID, Username\ , Loc, description, email, connection > \quad (6.3)$$

The vectors are evaluated using classifiers like Naive Bayes, Logistic research, SVM-Linear, SVM-Kernel with parameters Accuracy, Precision, and Recall.

2) Partition effectiveness of clusters is evaluated using the score of Manhattan, Euclidean, and Cosine similarity measures.

3) Aggregated user profiles are also evaluated for various clustering techniques like Error rate, Jaccard Index, and RAND score.

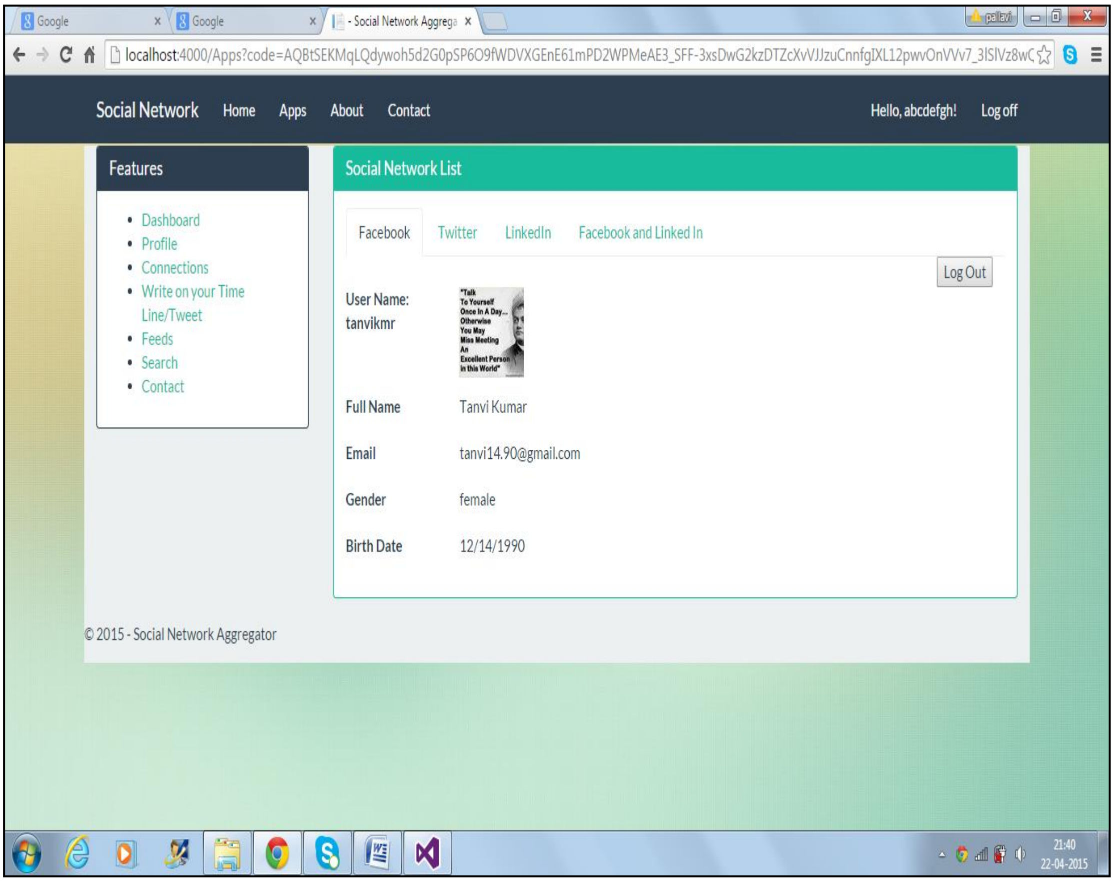
4) The QPSSN is evaluated on a sample that used 100 set of queries for testing the system precision and recall graphs are used to establish the effectiveness of the QPSSN system as a whole. The effectiveness of the search results is then compared with the keyword based and natural language search. The samples of input query are:
    a. Friends who live in the United States and knows DataScience
    b. Friends who are single and above 25 years
    c. Friends who work in Hays and is female
    d. People who live in London
    e. Looking for Java developer
    f. We require Java developer
    g. We require Java developer who lives in Delhi
    h. People who live in London and knows Python
    i. Looking for Java developer who lives in Gurgaon
    j. Friends who live in the United States and knows DataScience
    k. Friends who live in India and knows DataScience
    l. Friends who live in the United States and knows Database

Further, a Receiver Operating Characteristic (ROC) graph technique is used for visualizing QPSSN based on its performance. It represents a relationship between sensitivity (Recall) and specificity. It is a tool to evaluate the quality of cluster production, which shows the actual positive rate on the Y-axis and the curve showing the false positive rate on the X-axis.

6.4 RESULTS OBTAINED

QPSSN successfully aggregated information and tracks the social activities from Facebook, Twitter, and LinkedIn. The prototype proposed provides a comprehensive solution of SNS integration with some unique user friendly and powerful features like integrated profile management and integrated search capabilities.

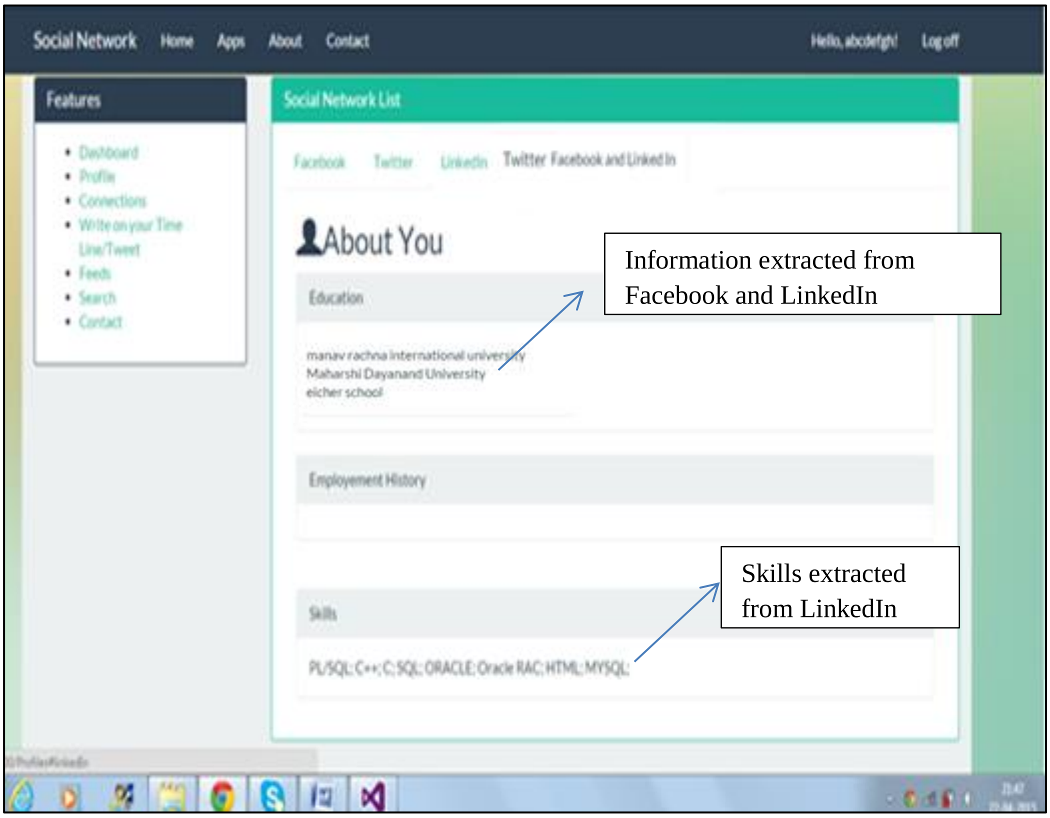
QPSSN is developed that integrated several social web sites together and extracted the useful information from multiple social networks. This has given an edge over the typical activity stream bases social activity implementations and is a step ahead to integrate the social data. We have implemented the proposed aggregator which has abstracted few features as an integrated solution with contacts and natural language search capabilities. Figure 6.3 displays the user profile of Facebook on our implementation of SNA. Similarly, Twitter and LinkedIn Tabs will display profiles of user available from Twitter and LinkedIn as shown in figure 6.3. The Integrated profile of Twitter, Facebook and LinkedIn is displayed in Figure 6.4 where Education is retrieved from Facebook and LinkedIn and Skill Set is extracted from LinkedIn.


Google
Google
- Social Network Aggrega
pallavi
localhost:4000/Apps?code=AQBtSEKMqLQdywoh5d2G0pSP6O9fWDVXGEnE61mPD2WPMeAE3_SFF-3xsDwG2kzDTZcXvVJJzuCnnfgIXL12pwvOnVVv7_3lSlVz8wC
Social Network
Home
Apps
About
Contact
Hello, abcdefgh!
Log off
Features
Dashboard
Profile
Connections
Write on your Time Line/Tweet
Feeds
Search
Contact
Social Network List
Facebook
Twitter
LinkedIn
Facebook and Linked In
Log Out
"Talk To Yourself Once In A Day... Otherwise You May Miss Meeting An Excellent Person in this World"
© 2015 - Social Network Aggregator
21:40
22-04-2015

User Name:
tanvikmr

Full Name	Tanvi Kumar
Email	tanvi14.90@gmail.com
Gender	female
Birth Date	12/14/1990

Figure 6.3 QPSSN View of Facebook Profile



About You

Education

manav rachna international university
Maharshi Dayanand University
eicher school

Employement History

Skills

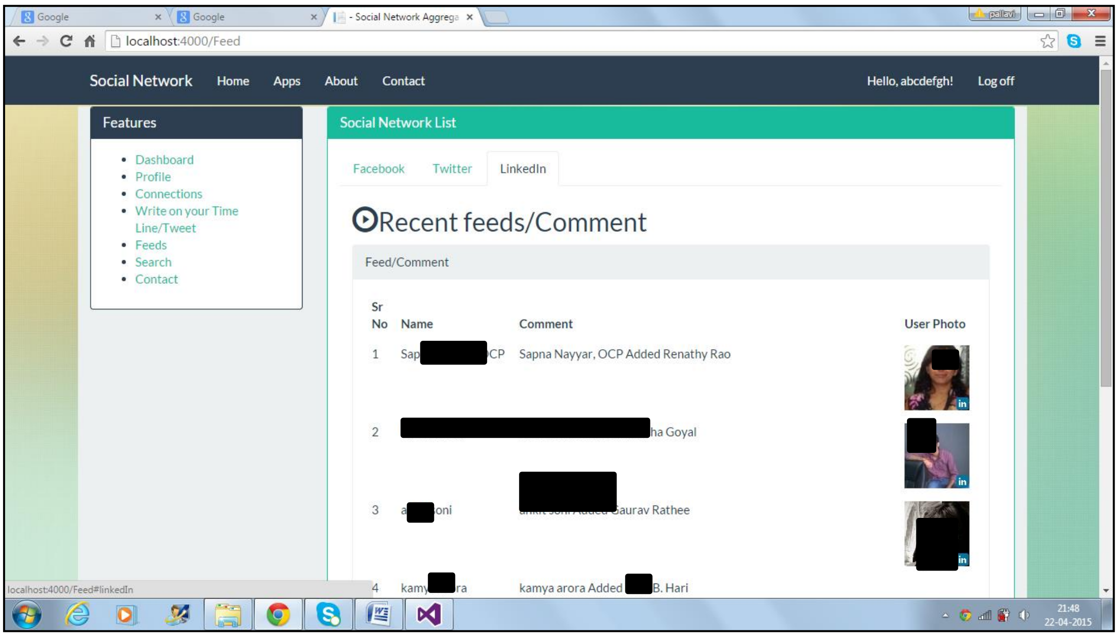
PL/SQL; C++; C; SQL; ORACLE; Oracle RAC; HTML; MYSQL;

Information extracted from Facebook and LinkedIn

Skills extracted from LinkedIn

Figure 6.4 Integrated Profile

A user can select the feature to display the relevant feed or information for Facebook as demonstrated in figure 6.5. The respective SNS to check the feeds/comments can be selected from the tabs.



Recent feeds/Comment

Feed/Comment

Sr No	Name	Comment	User Photo
1	Sap[redacted]CP	Sapna Nayyar, OCP Added Renathy Rao	[redacted]
2	[redacted]	[redacted]ha Goyal	[redacted]
3	a[redacted]oni	[redacted]aurav Rathee	[redacted]
4	kamy[redacted]ra	kamya arora Added [redacted]B. Hari	

Figure 6.5 Activity Stream of Feeds/comments

The aggregator provides the real time update of user's status or notification input to all social networks. This has been handled as an activity stream. An update in form of comment can be posted simultaneously in multiple sites by selecting the posed on checkboxes as depicted in figure 6.6.

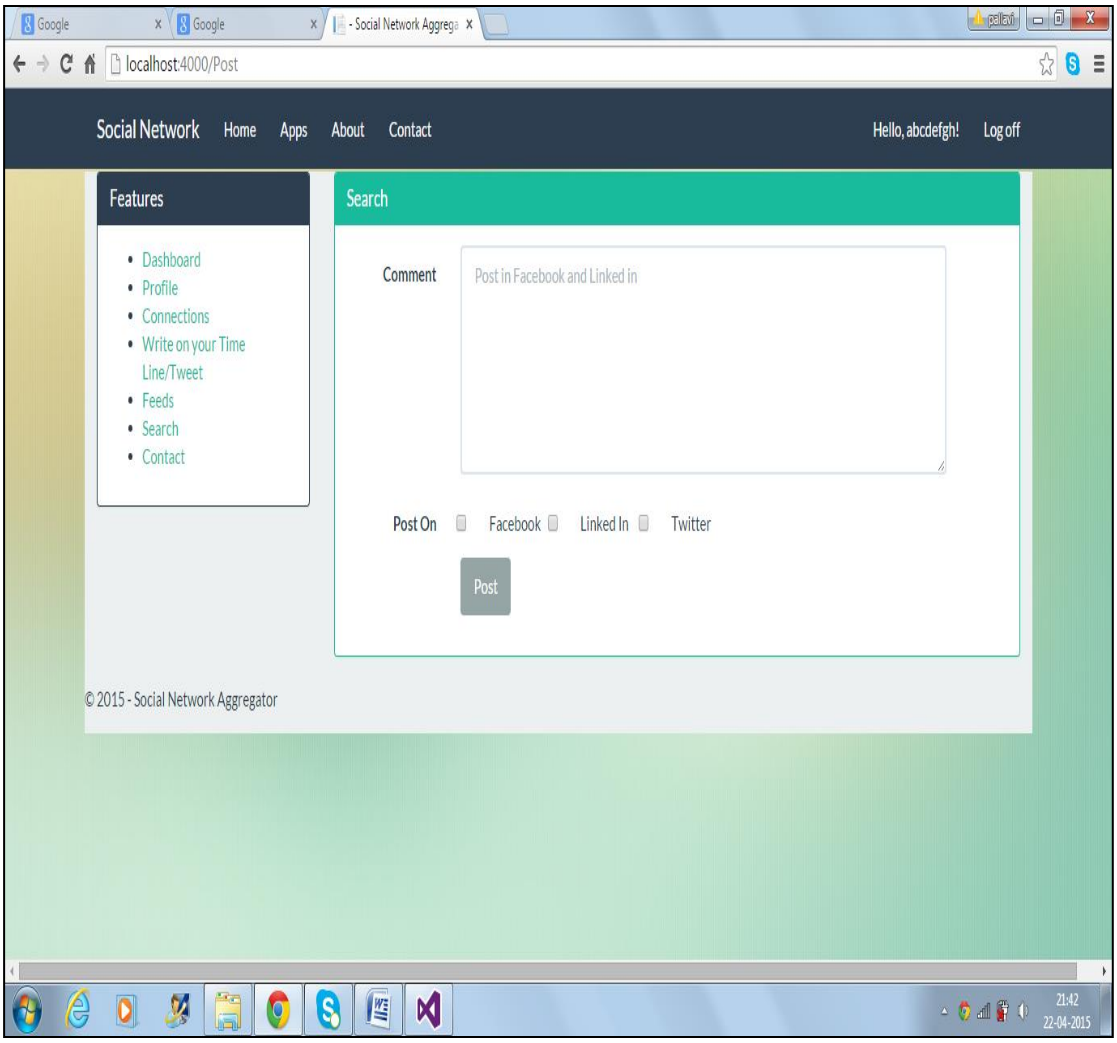
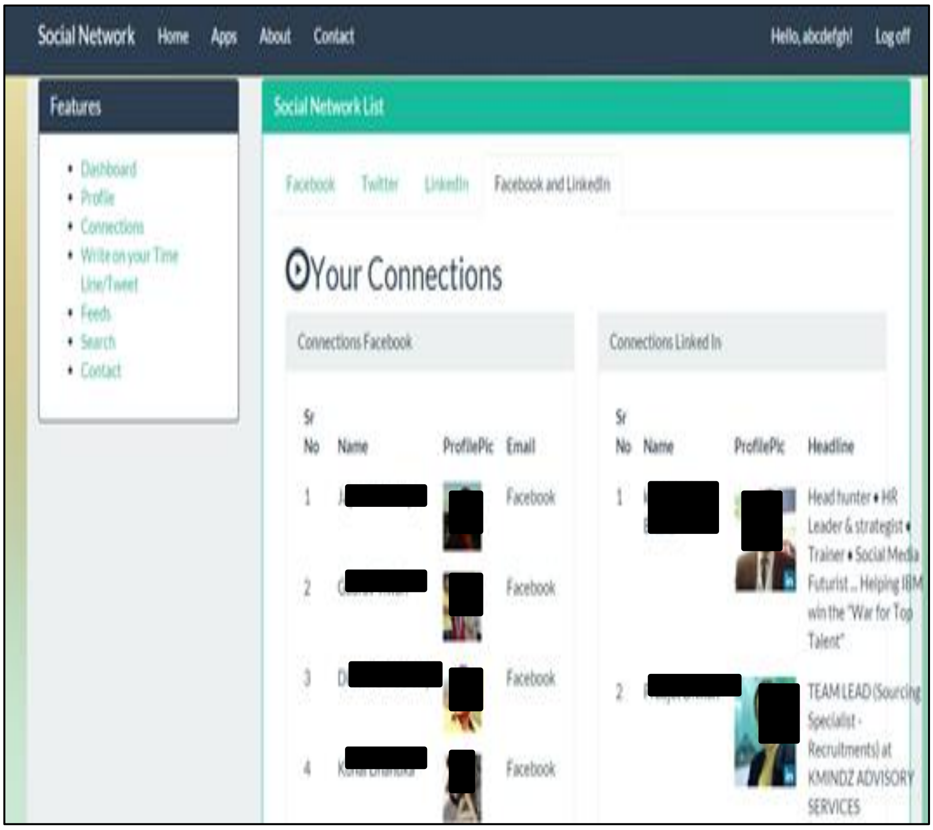


Figure 6.6 Updates of Post to Multiple Networks

Integrated contacts are a unique feature that makes it different from other aggregators. The fact that contact needs to be put together in one place is encapsulated and abstracted view of the SNS integrator as a whole. The system will also sense those profiles which have an overlap in more than one SNS and put them as one entity/contact as shown in figure 6.7.



Your Connections

Connections Facebook

Sr No	Name	ProfilePic	Email
1	[redacted]	[redacted]	Facebook
2	[redacted]	[redacted]	Facebook
3	[redacted]	[redacted]	Facebook
4	[redacted]	[redacted]	Facebook

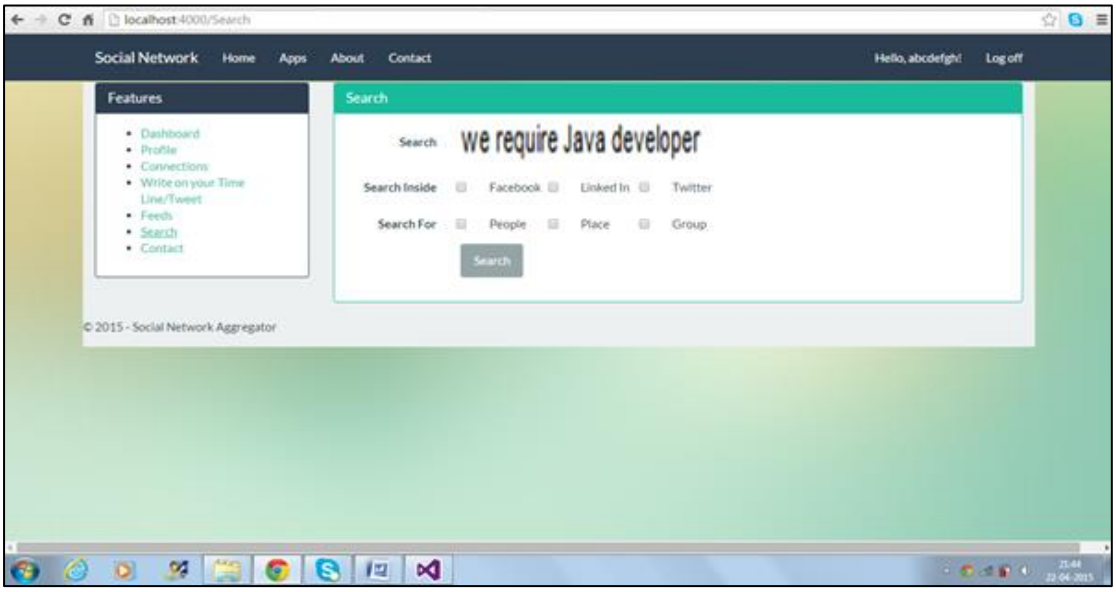
Connections Linked In

Sr No	Name	ProfilePic	Headline
1	[redacted]	[redacted]	Head hunter ♦ HR Leader & strategist ♦ Trainer ♦ Social Media Futurist .. Helping IBM win the "War for Top Talent"
2	[redacted]	[redacted]	TEAM LEAD (Sourcing Specialist - Recruitments) at KMINDZ ADVISORY SERVICES

Figure 6.7 Integrated contacts

The proposed solution enables an end to end social networking experience by integrating social data across multiple sites in one dashboard. With a user friendly interface, an attempt has been made to provide an abstraction of social activity by grouping the contacts, multi-site search to an extent.

The proposed system has opened doors for the ‘walled garden approach’ thereby extracting information across multiple SSN’s. This is a step ahead to integrate the social information. Multi-site search added an extra feature that gave an edge over other aggregators. A contact, keyword (place, contact for a demonstration in this prototype) can be searched from multiple sites (configurable) providing the integrated results. For Example, query based natural language search as shown in figure 6.8 and the output of Matlab is shown in figure 6.9.



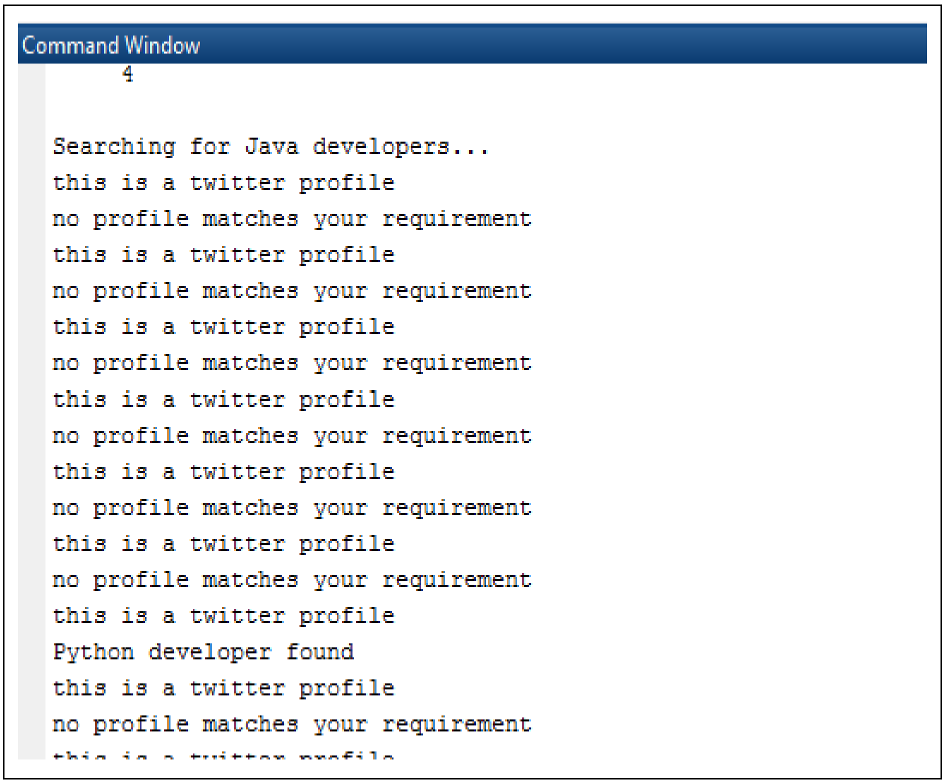
Search: we require Java developer

Search Inside □ Facebook □ Linked In □ Twitter

Search For □ People □ Place □ Group

© 2015 - Social Network Aggregator

Figure 6.8 Integrated Search

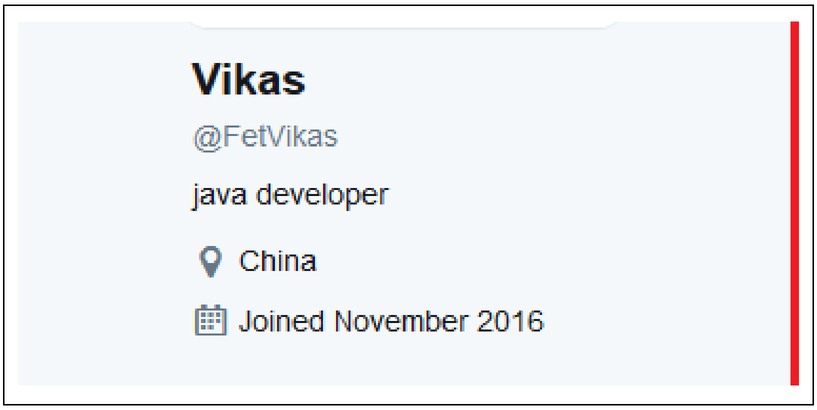


```
     4

Searching for Java developers...
this is a twitter profile
no profile matches your requirement
this is a twitter profile
no profile matches your requirement
this is a twitter profile
no profile matches your requirement
this is a twitter profile
no profile matches your requirement
this is a twitter profile
no profile matches your requirement
this is a twitter profile
no profile matches your requirement
this is a twitter profile
Python developer found
this is a twitter profile
no profile matches your requirement
this is a twitter profile
```

Figure 6.9 Output of Matlab

The result of the query is the first best option that meets the criteria as given in figure 6.10



Vikas

@FetVikas

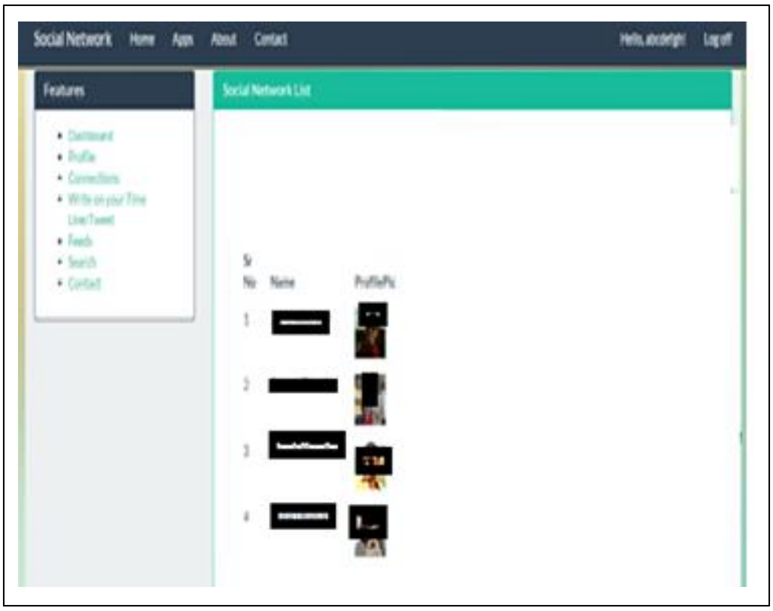
java developer

China

Joined November 2016

Figure 6.10 Result of the Query

The sorted list of output is as shown in figure 6.11



Sr No	Name	ProfilePic
1	[redacted]	[redacted]
2	[redacted]	[redacted]
3	[redacted]	[redacted]
4	[redacted]	[redacted]

Figure 6.11 Sorted List of Query results

6.5 DISCUSSION

On user’s profile dataset of Facebook, Twitter and LinkedIn, it has been observed that IRM$_{vector}$ holds its implication by performing well to the proposed HIASN algorithm for aggregating the profiles. It is observed that majority of users change their profile pictures and have variable friends among social network whereas UserID, UserName, and Location can be a promising attribute in identifying and linking profiles of the user. Further, the query processing system performed exceptionally well in case of natural language search.

The profile aggregated by QPSSN considered Facebook Profile and identified user's LinkedIn and Twitter Profile, the system was evaluated on Accuracy, Precision, Recall and F1 Score using five-fold cross validation. Accuracy depicted the user's which are correctly identified. The system is trained on supervised classifier using false negatives of the true positive set of true data with Naive Bayes which returned the probability that the IRM_{vector} was generated by v1 and v2 which belongs to the same user and sorted the profiles thus ensuring the accuracy in classification. For every similarity vector of the user profile, the possibility belonging to the same person is determined. Finally, we sort every one of the qualities of user's account in diminishing request to shape a rank R. The evaluation is taken for the vector on the four classifiers Naive Bayes, Logistic Regression, SVM-Linear, and SVM-Kernel to assess the accuracy, precision, recall, and F1 score.

The high quality examples contain all the similarity vectors for the profile pairs of the public profile dataset. The same wide variety of poor examples is synthesized by arbitrarily pairing profiles that don't participate in the same end user and calculating their similarity vectors. This yielded a complete list of instances. After training the classifier, output was tested by giving as input a profile pair of two social networks to be classified as a "Match" or a "Not Match". The functions set with the finest accuracy, precision, and recall using Naive Bayes is given in equation 6.1. Briefly, the Phonetic Encoding and Levenshtein distance to find the similarity between names, and the combined score of Euclidean distance to find the similarity between the location and change in location for the latest feed are the most promising attributes to resolve the identity of the user.

The matching score is calculated with the proposed algorithm and the results for each classifier are compared using the proposed vector achieving accuracy, precision, recall and F1 as 98%, 99%, 98% and 99% respectively. The high accuracy determined that the vector IRM_{vector} using Geo-Location are relevant attributes, helped in integrating the user profiles and is an essential feature for aggregating the profiles. Table 6.3 shows the result of multiple classifiers applied on the aggregated profile of the user of QPSSN for the vector IRM_{vector} with the proposed algorithm.

Table 6.3 Matching results

Classifiers	Accuracy	Precision	Recall	F1
Naive Bayes	0.987	0.998	0.983	0.990
Logistic research	0.965	0.995	0.946	0.97
SVM-Linear	0.982	0.992	0.966	0.979
SVM-Kernel	0.980	0.988	0.976	0.986

Figure 6.12 depicts that the vector IRM_Vector provides the best matching results. In 90% of the cases, the right profile was found at the top 5 ranks, while 50% if most of the public available attributes are used.

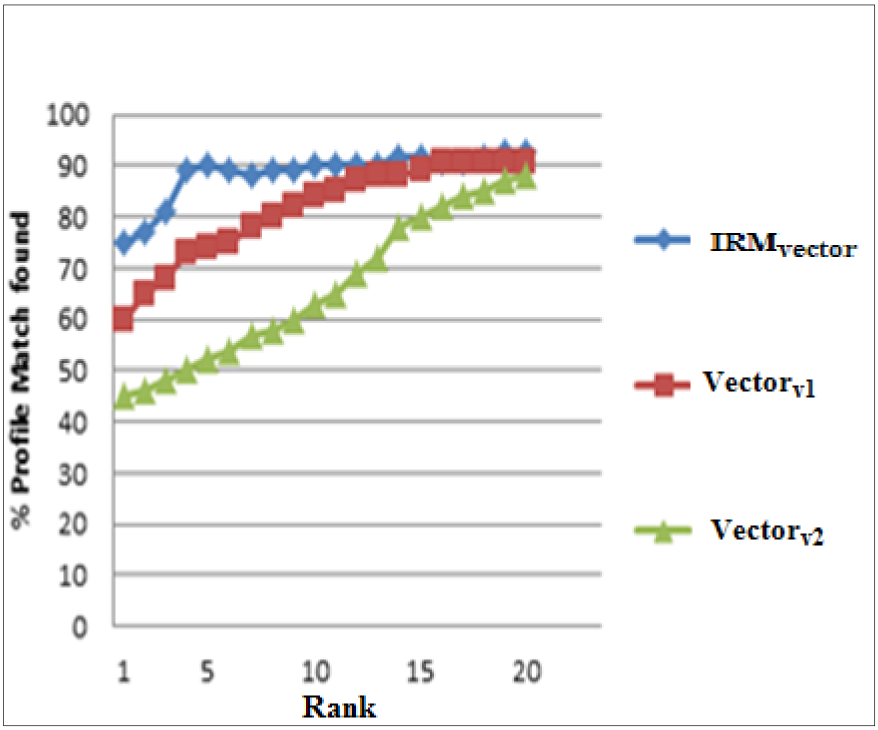


Figure 6.12 Comparison of various Vectors

The performance of the classifiers such as Naive Bayes and SVM-(Linear and Kernel) perform better than Logistic Research for false positive rates. In this scenario, one has 80% chance of finding the Twitter account associated with the Facebook/LinkedIn accounts as depicted in ROC curve of aggregated profile in figure 6.13.

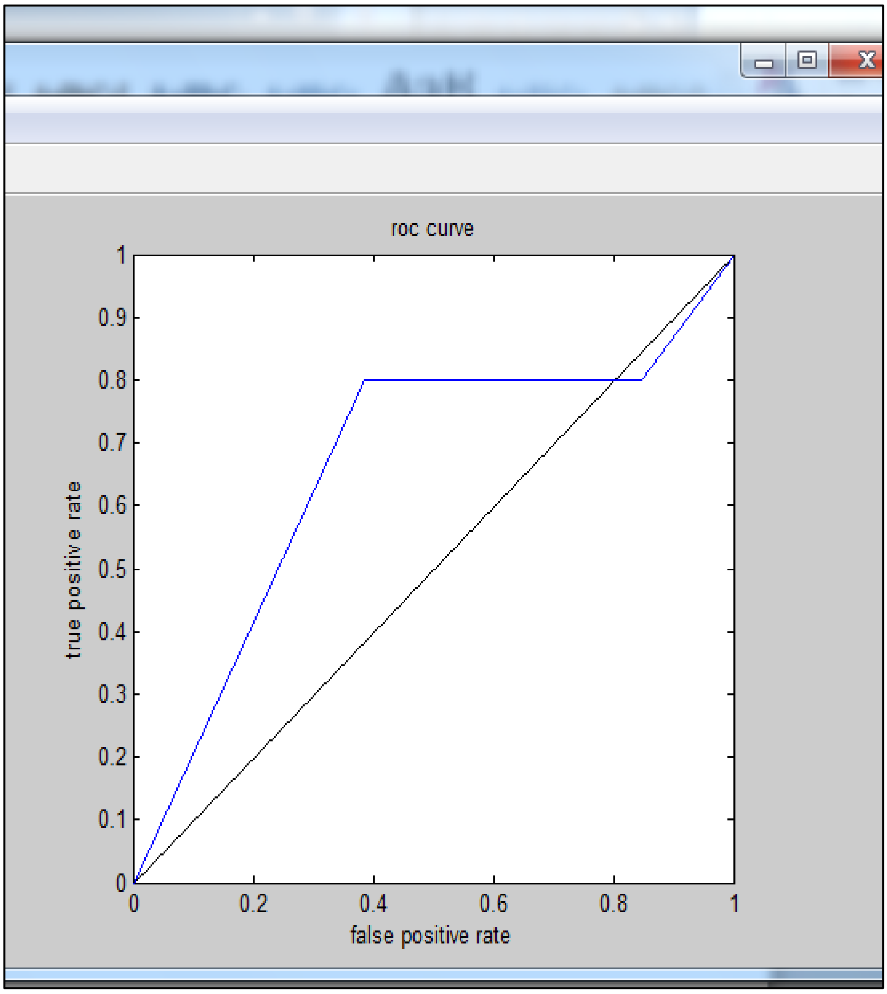


Figure 6.13 Roc Curve of Aggregated Profile

As per the obtained database, the linking of the user profile based on the several social networks is collected. As the same user can be available in various networks the overall profile details of the particular user is shown in above Figure 6.13. The clustering technique of QPSSN is evaluated on various similarity measures offering better clusters when compared with LDA, k-means and Ensemble k-means to create clusters and allocates a reference id to each cluster.

The clustered data is trained in the classifier. The entities placed to form the users are collected and fed to the classifier. The classifier compared the query with the trained data and finally classified results are obtained as the output.

The clustering technique is executed on 1000, 2000, 3000, 4000 and 5000 number of user profiles. Table 6.4 reflects the data values thus retrieved.

Table 6.4 Data values

Total Number of user profiles	Total Number of irrelevant user profiles	Total number of profile does not match the criteria	Total Number of Relevant user profiles found
1000	174	167	826
2000	336	326	1664
3000	492	378	2508
4000	560	504	3440
5000	675	612	4325

The system has initially chosen value of k varying from 3 to 12 to generate the partitions, first experiment is carried by passing value of k as 3 resulting in three clusters for each of the 12 queries: Node, NLP, Java, machine learning, database, Python, JavaScript, big data, deep learning, SQL, Hadoop and Datascience. These models identify repeating patterns in data and organize them into buckets known or "data clusters" and are depicted in figure 6.14. Similar results are obtained from k-mean clustering varying k from 4 to 12. Hence, the similar results are omitted.

```
database          2210
Top terms per cluster: database
Cluster 0:  job  administrator  sql  hire  database  server  derby  oracle  dba  disk
Cluster 1:  http  tungsten  dac  useful  ejnetwork  online  delete  8i  load  server
Cluster 2:  database  sql  look  dbm  nosql  9i  sanction  opm  expect  db2
javascript        22446
Top terms per cluster: javascript
Cluster 0:  javascriptinspirate  ebook  njavascript  kom  opensource  disponible  esta
Cluster 1:  javascript  developer  devops  job  library  jquery  know  use  linux  design
Cluster 2:  ncertification  dmoz  webmaster  leazysunny  php  javascript  javascriptd  fran
formvalidation
datascience       3636
Top terms per cluster: datascience
Cluster 0:  datascience  data  bigdata  machinelearning  analytics  iot  python  business  statistic
learn
Cluster 1:  bigdata  cancer  beat  use  artificialintelligence  deeplearning  datascience  iot  chatbot
fintech
Cluster 2:  ronald  vanloon  learn  machine  team  mix  expert  right  engineer  know
```

Figure 6.14 K-means Clusters for k = 3

For input queries, user’s information is collected and differentiated on the basis of interest and location. Data is collected for three different locations United Kingdom, United States and London. It is analyzed on the basis of java, nlp, Python, javascript, etc. Different parameters are analyzed to the model via k-means clustering on the data set (documents related to user-skills and user-level variables such as location, descriptions, etc.).

In order to identify that the user of a particular location has a particular skill, an approach must be found to identify the skill set of the user of the particular location. The particular location cluster can be created through the k-means algorithm because of its quick convergence to similarity. The skill cluster should define the boundaries of the skill set; this ends in a complex task. To obtain the skill set of the user, one needs to know the interest from the interest attribute (if available from the social network), as well as the user-generated post to mine information for the particular skill. In this study, clusters are obtained for k = 3 to 12 on skill wise user public data collected from various social networks. *K* partitions are generated optimally representing *M* partitions by voting scheme to generate a skilled public group for that particular location.

Input partitions to the confusion matrix are the clusters obtained from the previously discussed k-means (i.e., k = 3 to 12). In this phase, the clustering results are combined and the best cluster is chosen by computing similarity measure using confusion matrix and voting scheme. Table 6.5 shows the top five terms of each cluster by combining the results for a particular location London.

Table 6.5 Top five clusters

DataScience	JavaScript	Database
machineLearning	Jquery	NoSql
Datascience	FormValidation	Sql
BigData	Nodejs	MongoDB
DeepLearning	Library	Pymongo
Analytics	Reactjs	Database

Clusters of the entity are aggregated together based on their similarity in which high similarity is given more preference over low similarity ones. The formation of linkage of clustering is depicted in figure 6.15 using HEKHAC.

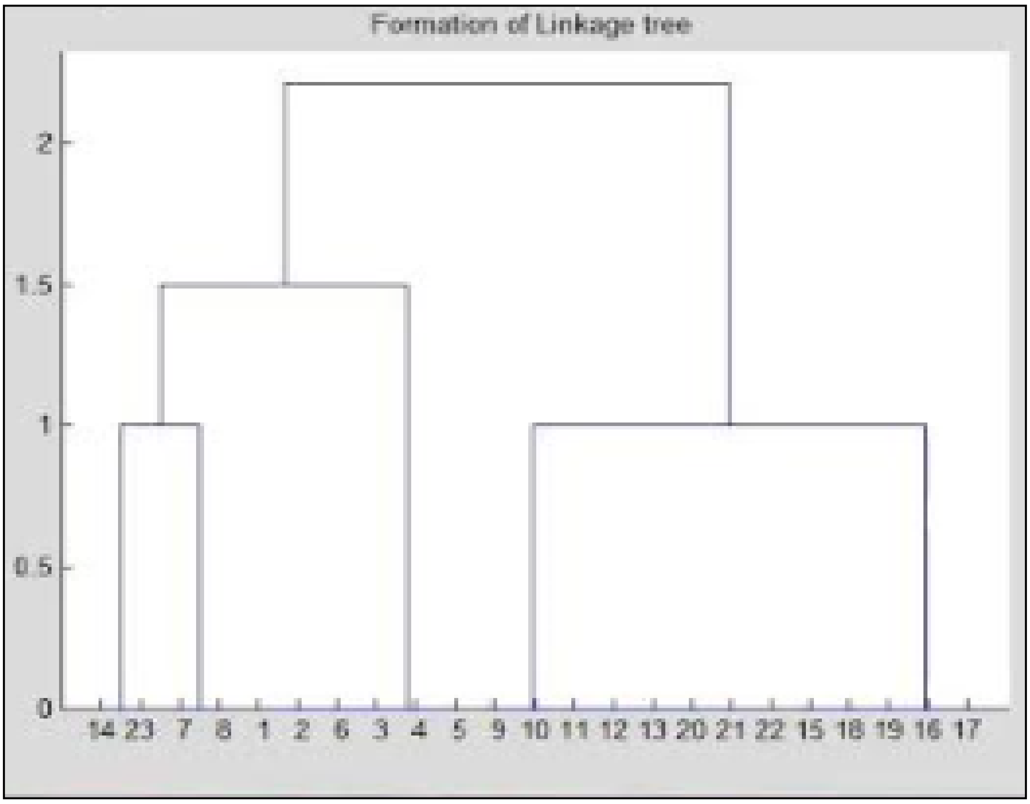


Figure 6.15 Linkage Graph of User Profile

The dendrogram is broken at different levels to obtain a different set of groups of user. This dendrogram is cut at 1.5 or less to obtain compact and well-separated clusters. The clustering obtained in this manner demonstrates that the users of integrated profile fall into several distinguishable clusters. The centroid of each of these clusters is determined by computing the mean of the IRM$_{\text{vector}}$ of the users falling into the cluster. On each cluster, five-fold cross validation is implemented to assess the improvement of clusters. It has been observed that the results produced by HEKHAC is 80% better than that can be produced by k-means or Ensemble technique as depicted in figure 6.16.

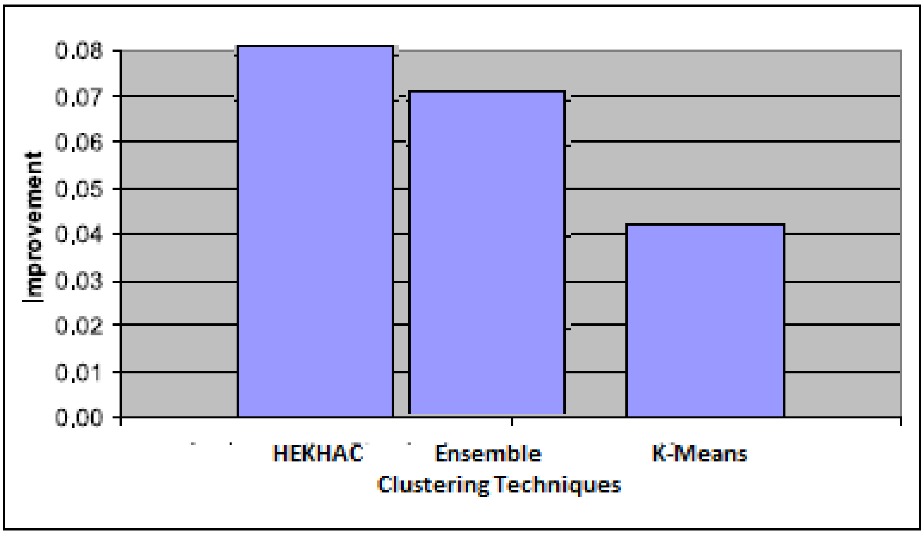


Figure 6.16 Improvement in clustering Techniques

LDA, k-means, Ensemble k-means and HEKHAC are evaluated with various similarity measures such as Error rate, Jaccard Index, Manhattan Distance, Euclidean distance, Cosine dissimilarity, and RAND index [198][199][200]. The error rate depicts the average number of misclassified elements. Partitions are more similar if the error rate is less. Error rate is used to validate the accuracy of the final partition. The purity of a cluster is one of the validation measures that quantify the coherence of the cluster where 0 indicates a bad clustering and 1 indicates a perfect clustering. The purity results of various evaluations on distance measures are depicted in Table 6.6 and 6.7 respectively.

Table 6.6 Evaluation on various measures

Data	Manhattan	Euclidean	Rand	Cosine
Twitter profile	0.91	0.83	0.46	0.10
Facebook profile	0.76	0.85	0.43	0.20
LinkedIn profile	0.66	0.65	0.33	0.15

Table 6.7 Evaluation of aggregated user profiles for various clustering techniques

Dataset	Method	Error rate	Jaccard Index	RAND score
Aggregated user's public information	LDA	42	0.45	0.65
Aggregated user's public information	K-means	45	0.49	0.68
Aggregated user's public information	Ensemble k-means	15	0.97	0.95
Aggregated user's public information	HEKHAC	10	0.98	0.97

K-means, Ensemble or LDA do not appear to be nearly as effective as HEKHAC. Bottom-up agglomeration appears to capture the similar users more effectively than a random selection of K seeds or ensembling k-means. It has been observed that clusters of user's profile attributes can be effectively used as a measure to ascertain the user's interest, thus providing an effective means to bridge the gap between users and entities of user's profile attributes.

The output of language cluster depicting the skills for particular location United States is shown in figure 6.17. Figure 6.18 represents the count of users for different cities of the United States for the skills data science, database and javascript. These clusters serve as an intermediate between user profiles and the results of the user query. It helped the system to achieve the desired output.

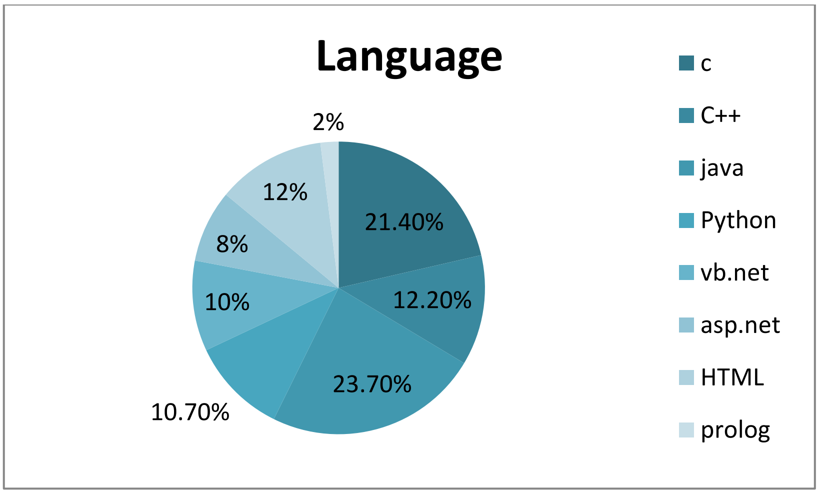


Figure 6.17 Clustering Output – Language

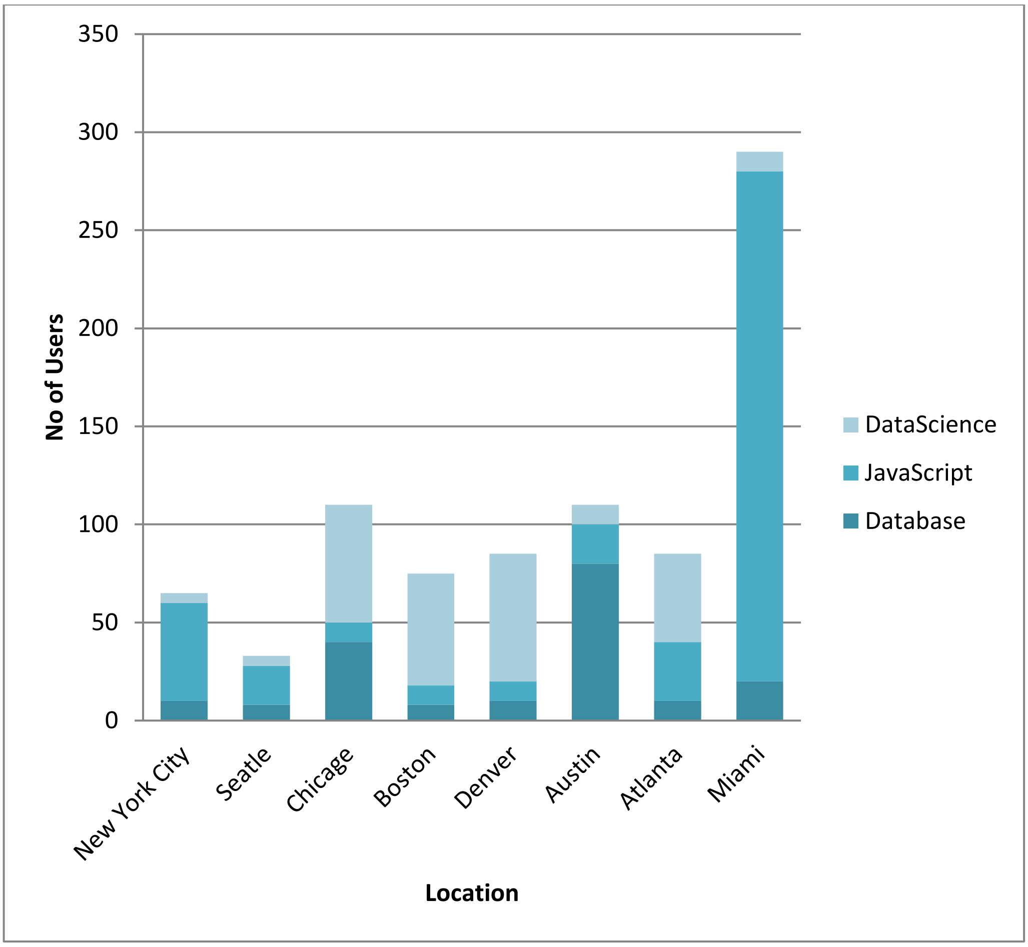


Figure 6.18 Count of User for Different Skills for Different Locations

For the input queries on QPSSN, the system is tested on 100 queries and is able to extract the relevant profiles from the clusters as per the cluster id**.** During the course of implementation, it has been observed that the system resulted in expected user profile up to the top 3 ranks. In information retrieval and query processing, the precision is the fraction of retrieved documents that are relevant to the query and Recall denotes the fraction of the relevant documents that are successfully retrieved. High precision depicts that the quality of retrieved results achieves the performance close to the expectations of the users.

Test 1

Initially, 1000 number of user profiles is supplied to the proposed system and the following is the data collected:

Total Number of user profiles = 1000

Total Number of irrelevant user profiles= 174

Total number of profile does not match the criteria= 167

Total Number of Relevant user profiles found= 826

Therefore, *Precision*= 826 / (826+174) = 0.826

Recall= 826 / (826+167) = 0.832

Similarly, tests are performed on the user profiles for 2000, 3000, 4000 and 5000 and analysis is obtained. Summarizing, the Precision is found in the range of 82.6% to 86.5%, Recall is found in the range of 83.2% to 87.6%. Table 6.8 shows the summarized result:-

Table 6.8 Accuracy Measure

No. of User Profiles	Precision (in %)	Recall (in %)
1000	82.6	0.832
2000	83.2	0.836
3000	83.6	0.869
4000	0.86	0.872
5000	86.5	0.876
Average	84.38	85.7

The precision-recall graph is depicted as follows in figure 6.19.

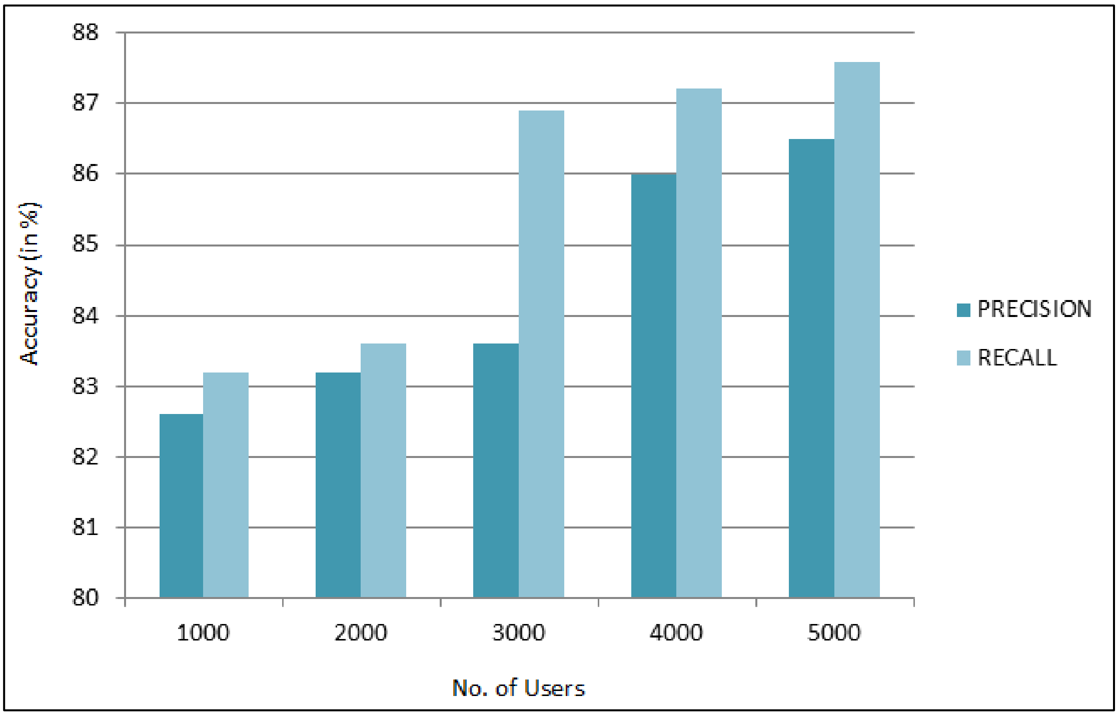


Figure 6.19 Precision Recall Graph

The precision value, true positive rate and false positive rate of the query search are obtained. Figure 6.20 depicts the precision of QPSSN to measure the accuracy and the traditional keyword search. It is clearly distinguishable that QPSSN performs cutting-edge results over keyword search.

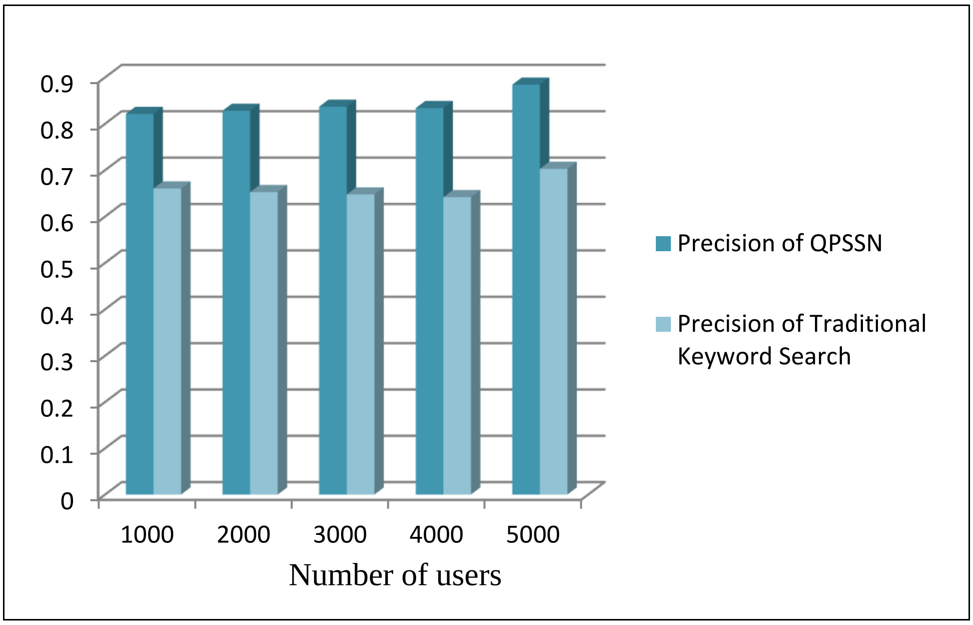


Figure 6.20 Comparison of Precision Value to Measure Accuracy

A Receiver Operating Characteristic (ROC) graph is a technique for visualizing, organizing and selecting classifiers based on its performance. It represents a relationship between sensitivity (Recall) and specificity. The threshold is set to be limited, resulting in a false positive rate of less than 4%. The ROC curve of the query is as shown in figure 6.21.

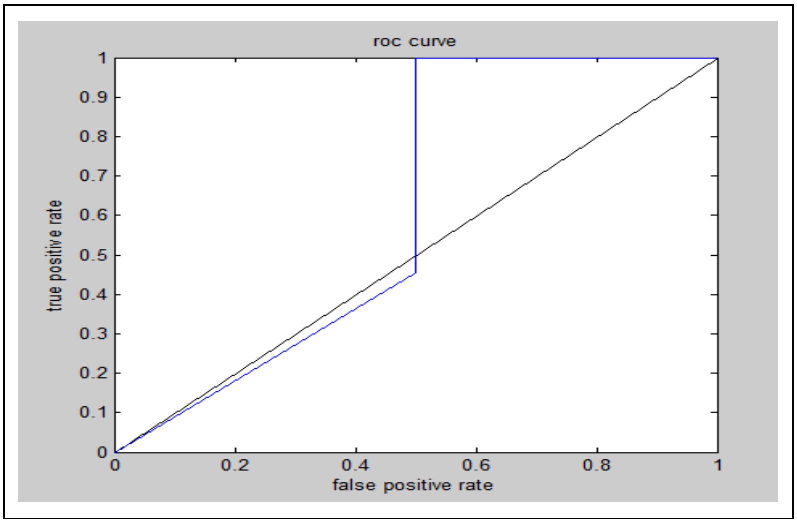


Figure 6.21 Roc Curve Depicting the True Positive Rate Vs False Positive Rate for Test Corpus

6.6 CONCLUSION

The chapter presented implementation and results pertaining to the proposed work. The research work bridged the gap of satisfaction of the user corresponding to his/her query to the social network and managing and organizing multiple social networks at one place. Clustering user’s aggregated profile using HEKHAC is a novel concept and the proposed approach could minimize the error rate maximizing RAND Score and Jaccard index. It also supported a hybrid integrated algorithm for the autonomous social network to provide an integrated profile of a user which otherwise remains in isolation among multiple social networks. The integrator outstandingly disambiguates user profiles existing across different social networks using public attributes with the decision to map the profiles using change in location of the user. An issue regarding keyword searching has been well addressed by QPSSN and appeared to be the most suitable approach as the same could overcome the limitations imposed by traditional search. Results are competitive and have an edge over existing works.

CHAPTER 7

CONCLUSIONS AND FUTURE SCOPE

7.1 CONCLUSIONS

The research commenced with exploring the potential of social media and discovered that social media is the new way of collecting and disseminating information by the means of social networks. Few of the most popular social networking sites like Facebook, LinkedIn, and Twitter provide social services ranging from scraps, updates, tweets to user stories, just to list a few. Communication, publishing information, and sharing of content have widely increased the scale of SNS leading to various issues and concerns which in turn demanded multidisciplinary solutions. The in-depth study of literature highlighted that ***"Information Overload"*** and ***"Walled Garden"*** are the most bulbous glitches dominant in the social web and are briefly discussed as follows.

- ***Information Overload***

Information is scattered across multiple social network sites. Users continue to spend a lot of time and effort to extract contents of interests across all the incoming information leading to unattended relevant information too. To add on, many sites have restrictions on what a user receives in his social stream. Essentially a user receives social data what has been shared with him/her or who are the member(s) thereof. A user may choose to add as many friends to access a vast range of information; however; the chances of information overload increases.

- ***Walled Garden***

The user can subscribe to different groups of a variety of interests, but again there is no guarantee that all members of the group are connected to the same social network and connected to each other.

These two problems have led to difficulties for users to explore the relationship to other networks and confined access the isolation barrier and also led to overlook some of the relevant information because of inability to prioritize the information.

Owing to the above stated high priority concerns, ***the current work of research proposed a novel social network aggregator with an easy to use query based search.*** Following modules are proposed and implemented to achieve the stated target:

- ***Hybrid Integrated Autonomous Social Network (HIASN)*** is a hybrid aggregator that uniquely identifies the presence of user across multiple OSN (Facebook and LinkedIn, in particular) and aggregated the social data and provided a single unique profile to the user. HIASN relies on FOAF and activity stream to retrieve the social data from multiple SNS's and makes use of OAuth protocol for Authentication. The proposed model could help the users to use multiple SNS normally with less effort and greater efficiency when compared to extract useful information. Only three publicly available attributes have been used and could achieve the best results in top ranks. UserID, Username, and location have been analyzed and resulted in being the most discriminative features for achieving the best results. The adoption of these publicly available features allowed achieving accuracy, precision, recall, and F1 score up to 98%, 99%, 98%, and 99% respectively as depicted in chapter 6.
- ***Hybrid Ensemble K-means Hierarchical Agglomerative Clustering (HEKHAC)*** is a clustering mechanism that could group the interests of aggregated user profiles. The bitonic sorting algorithm has been deployed to sort out the profiles as per the specified inputs and prioritize the output user profiles. The clustering mechanism proved to be an optimal algorithm as it could achieve minimum error rate, maximum Jaccard score, and RAND score.
- ***Query Processing in Social Network Aggregator (QPSNA)*** could extract the information from the user profile from various social networks using an efficient algorithm that takes the input query in a natural language. The system has been tested for a set of 100 queries and it has been derived that the expected result resulted in top 3 users by the system. The proposed research methodology incorporated HEKHAC and the Bitonic Sorting to cluster the

input datasets and prioritize the output data according to user's interest. The simulation resulted into the fact that natural language processing on the query using multiple social networks increased the discoverability of the user, helped the organizations and businesses to collaboratively execute promotions, and could determine new networks and people. The proposed strategy exhibited better efficiency and superiority when compared to the advanced user profile mapping techniques based on keyword searching. The proposed research method can be further used for background verification purpose, recruitment agencies, targeting a specific group of people. However, the system currently only considered the publicly available attributes only while in future; it can be extended to search for post/tweets to make provision for real time interaction.

Outcomes of the proposed work confirmed that the proposed framework of aggregating user's profile and extracting the information from this pool of information which is available at multiple social networks will return results in accordance to user-centric factors.

7.2 UNIQUE CONTRIBUTIONS

A number of social media aggregators have revealed up in recent years, but social media services still require to research and implement more effective and efficient ways to provide aggregation. The requirement of a new SNA is justified as at the forefront of study only, the existing SNAs were analyzed, compared and the same is reflected in the article titled. "***Study and Analysis of Social Network Aggregator***" published in IEEE *International Conference on Optimization, Reliability, and Information Technology (ICROIT), 2014,* 145-148.

Detailed study of Social Network and Social Network Aggregators led to the development of Profile Aggregator HIASN, Clustering Mechanism HEKHAC and Query Processing Mechanism QPSNA suitable for Social Networks. Following are the unique contributions:

1. The main contribution of ***HIASN*** is the development of automated identity resolution and aggregation methods, both for searching and linking user accounts that correspond to the same individual in popular social networks.

Matching accounts across OSNs allows marketers, enterprises, businesses and security professionals to work on comprehensive user profiles. The contribution is reflected through the article titled, "**Design of a Hybrid Integrator for Autonomous Social Networks.**" International Journal of Computer Information Systems and Industrial Management Applications, Volume 9,(2017), 241-248.

2. The aggregator is developed that provided a comprehensive solution of SNS integration with some unique user friendly and powerful features like integrated profile management and integrated search capabilities. It integrates several social websites together and extracting the useful information among multiple social networks. This has given an edge over the typical activity stream bases social activity implementations and is a step ahead to integrate the social data. The proposed aggregator has abstracted few features as an integrated solution with contacts and search capabilities. The contribution is reflected through the article titled, "**Content-Based Social Network Aggregation.**" In ICT Based Innovations, 185-194. Springer, Singapore, 2018.
3. The thesis proposed and implemented clustering mechanisms to group the aggregated user's profile as per their interest. The proposed ensemble clustering utilized known k-means algorithm to improve results for the aggregated user profiles across multiple social networks. The approach produced an ensemble similarity measure and provided better results than taking a fixed value of k or guessing a value of k while not altering the clustering method. This paper stated that good ensembles clusters can be spawned to envisage the discoverability of a user for a particular interest. This technique had then been wrapped over Hierarchical Agglomerative to align the clusters into one to one correspondence. The contribution is reflected through the article titled, "**Clustering in Aggregated User Profiles Across Multiple Social Networks.**" International Journal of Electrical and Computer Engineering (IJECE), Volume 7(6), 3692-3699.
4. The thesis also contributed a ***more realistic query processor*** that could answer the user's query in a natural language and overcame the traditional strategies adopting the high level of user's communication. In fact, it is a novel framework that could not only process the query using NLP but also has the

potential to integrate heterogeneous social networks. The proposed ***hybrid clustering algorithm*** could cluster the profiles optimally and hence contributed towards optimizing the entire framework. ***The novel ranking algorithm*** could rank the best profile on the top. The entire work has been evaluated experimentally on various parameters. The proposed framework is more efficient and superior to the existing user profile mapping techniques based on keyword searching. The contribution is reflected through the article titled, **"Design of Query Processing System to Retrieve Information from Social Network using NLP",** KSII Transactions on Internet and Information Systems, vol. 12, no. 3, pp. 1168-1188, 2018.

5. The research work also contributed three more articles reflecting the intensity of literature review that was dwelled upon to carry out this work. The review of literature was motivating as the challenges were too many to handle. The contribution is reflected through following articles:
   a. "***Extracting Information from Social Network using NLP***." International Journal of Computational Intelligence Research, 13(4), 621-630
   b. "***Characterizing User Demographics Across Social Network***", International Journal of Advanced Research, Volume 5(6), 2308-2312
   c. "***Major Encounters To Search The Social Network***", International Journal of Advanced Research in Computer Science and Software Engineering, volume 7, Issue 6, PP- 504-508

Besides the fact, that thesis has contributed significantly; the research remains a never ending pursuit. The future scope of the work is being laid down in the next section.

7.3 FUTURE SCOPE

Although the efforts made during this work tried to rejoin the open ends mentioned in literature. However; while bridging the gaps, some new concerns that are still challenging and can become the subject of research in the near future are identified and are mentioned below:

- Due to the enormous size of social network and the perpetual growth of the embryonic network, there is a need of an operative mechanism to demonstrate the closeness of two nodes, measuring centrality etc.
- To provide an optimal, flexible dynamic and minimal time for executing a query is another important task.
- The hypothetical questions, management of time and sources within a network, considering collaboration into the query language and updating the coordination of the networks is another hurdle to cross over.
- Displaying effective results of different statistics over the network is another challenge to face.
- The proposed system only considered the public available attributes, in the future; it can be extended to search for post/tweets to make provision for real time interaction.
- Increase the number of supported languages by the query processing system.
- Further research can be explored in the area on non-public information available which has been kept out of scope of this thesis.

In spite of the fact that there are numerous challenges still prevailing in the social network, yet with the distinct captivation of researchers and organizations in this domain, future will without a doubt convey new answers to improve this area worthy for better systems for better networking.

References

[1] Halpin, H., & Tuffield, M. “A standards-based, open and privacy-aware social web.” W3C Social Web Incubator Group Report 6th December2010, accessed on https://www.w3.org/2005/Incubator/socialweb/XGR-socialweb- 20101206/

[2] Gruber, T. “Collective knowledge systems: Where the social web meets the semantic web.” Web semantics: science, services and agents on the World Wide Web, Vol. 61, 2008, pp. 4-13.

[3] Tan, W., Blake, M. B., Saleh, I., & Dustdar, S. “Social-network-sourced big data analytics.” IEEE Internet Computing, vol. 175, september 2013, pp. 62-69.

[4] Horowitz, D., & Kamvar, S. D. , “The anatomy of a large-scale social search engine.”, In Proceedings of the 19th international conference on World wide web , ACM, Raleigh ,NC, USA, April 2010, pp. 431-440.

[5] Granovetter, M. “The impact of social structure on economic outcomes.”, Journal of economic perspectives, vol. 191, Stanford University, California, 2205, pp. 33-50.

[6] Bello-Orgaz, G., Jung, J. J., & Camacho, D. “Social big data: Recent achievements and new challenges.”, Information Fusion, vol. 28, 2016, pp. 45-59.

[7] Fadrique del Campo, Hector. "Design and development of a social network aggregator." PhD diss., 2012.

[8] Zhang, J., Wang, Y., & Vassileva, J. “SocConnect: A personalized social network aggregator and recommender.”, Information Processing & Management, vol. 49(3), 2013, pp. 721-737.

[9] Benevenuto, F., Rodrigues, T., Cha, M., & Almeida, V. “Characterizing user behavior in online social networks.” In Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, ACM, November 2009, pp. 49- 62.

[10] King, R.. When your social sites need networking. BusinessWeek, accessed on http://tinyurl.com/o4myvu

[11] Schroeder, S. 20 "Ways To Aggregate Your Social Networking Profiles.”, Mashable. Accessed on http://mashable. com/2007/07/17/social-

network-aggregators/accessed October, 9, 2011

[12] Perez, S. "Who Uses Social Networks and What Are They Like." (2009) accessed on https://readwrite.com/2009/12/31/ who_uses_social_networks_and_what _are_they_like_pa/.

[13] Patriquin, A. “Connecting the social graph: member overlap at opensocial and facebook.”, Compete. com blog 2007, accessed on http://blog.compete.com/2007/11/12/ connecting-the-social-graph-member-overlap-at-opensocial- and-facebook/

[14] Ghodsi, Ali, Teemu Koponen, Jarno Rajahalme, Pasi Sarolahti, and Scott Shenker. "Naming in content-oriented architectures." In Proceedings of the ACM SIGCOMM workshop on Information-centric networking, ACM, 2011, pp. 1-6.

[15] Madnick, S., & Siegel, M. “Seizing the opportunity: Exploiting web aggregation.” 2001, accessed on http://web.mit.edu/smadnick/www/ wp/2001-13.pdf

[16] Igoe, Patrick T., and Leonid Kravets. "Sending personal information to a personal information aggregator." U.S. Patent 7,966,647, issued June 21, 2011.

[17] Hansen, M., Madnick, S., & Siegel, M. “Process aggregation using web services.” In International Workshop on Web Services, E-Business, and the Semantic Web Springer, Berlin, Heidelberg., May 2012, pp. 12-27.

[18] Ellison, N. B. “Social network sites: Definition, history, and scholarship.” Journal of computer-mediated Communication, vol. 131, 2007, pp. 210-230.

[19] Burke, M., Marlow, C., & Lento, T. “Social network activity and social well-being.” In Proceedings of the SIGCHI conference on human factors in computing systems, ACM, April 2010 pp. 1909-1912.

[20] YANG, S. “Data Modeling and Query Processing for Online Social Networking Services” Doctoral dissertation, 2011.

[21] Geho, P. R., & Dangelo, J., “The evolution of social media as a marketing tool for entrepreneurs.”, The Entrepreneurial Executive, vol. 17, 2012, pp. 61-65.

[22] Burke, M., Marlow, C., & Lento, T. “Feed me: motivating newcomer

contribution in social network sites." In Proceedings of the SIGCHI conference on human factors in computing systems, ACM, April 2009, pp. 945-954.

[23] Imran, M., Castillo, C., Diaz, F., & Vieweg, S., "Processing social media messages in mass emergency: A survey." ACM Computing Surveys CSUR, vol. 474, 2015, pp. 67.

[24] Armentano, M. G., Godoy, D., & Amandi, A. A., "Followee recommendation based on text analysis of micro-blogging activity." Information systems, vol. 388, 2013, pp. 1116-1127.

[25] Java, A., Song, X., Finin, T., & Tseng, B. "Why we twitter: understanding microblogging usage and communities." In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, ACM, August 2007, pp. 56-65.

[26] Yusof, N., & Rahman, A. A. "Students' interactions in online asynchronous discussion forum: A Social Network Analysis.", In Education Technology and Computer, 2009. ICETC'09. IEEE, April 2009, pp. 25-29.

[27] Lemire, D., & Maclachlan, A. "Slope one predictors for online rating-based collaborative filtering." In Proceedings of the 2005 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, April 2005, pp. 471-475.

[28] Yang, J., & Mai, E. S. "Experiential goods with network externalities effects: An empirical study of online rating system.", Journal of Business Research, vol. 639-10, 2010, pp. 1050-1057.

[29] Bao, J., Zheng, Y., & Mokbel, M. F. "Location-based and preference-aware recommendation using sparse geo-social networking data.", In Proceedings of the 20th international conference on advances in geographic information systems, ACM November 2012, pp. 199-208.

[30] Tang, L., Chen, H., Ku, W. S., & Sun, M. T. "Exploiting location-aware social networks for efficient spatial query processing." GeoInformatica, vol. 211, 2017, pp. 33-55.

[31] Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P., & Zhao, B. Y. "User interactions in social networks and their implications." In Proceedings of the 4th ACM European conference on Computer systems, New York, NY, USA

.ACM, April 2009. pp. 205-218.

[32] Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis et al. "Life in the network: the coming age of computational social science." Science (New York, NY) vol. no. 323 (5915) ,2009, pp. 721-729.

[33] Abel, F., Henze, N., Herder, E., & Krause, D., “Linkage, aggregation, alignment and enrichment of public user profiles with Mypes”. In Proceedings of the 6th International Conference on Semantic Systems. New York, NY, USA, ACM, September 2010, pp. 11-19.

[34] Carmagnola, F., Osborne, F., & Torre, I. “User data distributed on the social web: how to identify users on different social systems and collecting data about them.” In Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender SystemsBarcelona, Spain , ACM, September 2010, pp. 9-15.

[35] Gomez-Rodriguez, Manuel, Krishna P. Gummadi, and Bernhard Schoelkopf. "Quantifying Information Overload in Social Media and Its Impact on Social Contagions." In ICWSM, 2014, pp. 170-179.

[36] Orlandi, F., Breslin, J., & Passant, A.” Aggregated, interoperable and multi-domain user profiles for the social web”. In Proceedings of the 8th International Conference on Semantic Systems New York, NY, USA, ACM, September 2012, pp. 41-48.

[37] Eslami, M., Aleyasen, A., ZilouchianMoghaddam, R., & Karahalios, K. G. “Evaluation of automated friend grouping in online social networks.” In CHI'14 Extended Abstracts on Human Factors in Computing Systems, ACM, April 2014, pp. 2119-2124.

[38] Sandra Garcia Esparza and Michael P. O Mahony and Barry Smyth “Catstream: categorising tweets for user profiling and stream filtering.” In Proceedings of the 2013 international conference on Intelligent user interfaces, ACM, 2013, pp. 25-36.

[39] Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, Jon Sperling “Twitterstand: news in tweets”. In Proceedings of the 17th acmsigspatial international conference on advances in geographic information systemsHUD Office of Policy Development & Research PD&R,

Washington, ACM, DC November 2009. pp. 42-51.

[40] Shen, K., Wu, J., Zhang, Y., Han, Y., Yang, X., Song, L., & Gu, X. , “Reorder user's tweets.” ACM Transactions on Intelligent Systems and Technology TIST, vol. 4(1), 2013, pp. 6:1 – 6:17.

[41] Helmond, Anne. "Identity 2.0: Constructing identity with cultural software." In Proceeding of Mini-conference initiative, University of Amsterdam (Amsterdam, ND–Jan 20-22 2010). 2010, pp 1-28.

[42] Spirin, Nikita V., Junfeng He, Mike Develin, Karrie G. Karahalios, and Maxime Boucher. "People search within an online social network: Large scale analysis of facebook graph search query logs." In Proceedings of the 23rd acm international conference on conference on information and knowledge management, ACM, 2014, pp. 1009-1018.

[43] Gauch, S., Speretta, M., Chandramouli, A., & Micarelli, A., “User profiles for personalized information access.”In The adaptive web lecture notes in computer science books. Springer, Berlin, Heidelberg, vol. 4321, 2007, pp. 54-89.

[44] Abel, F., Gao, Q., Houben, G. J., & Tao, K. “Analyzing user modeling on twitter for personalized news recommendations.” In International Conference on User Modeling, Adaptation, and Personalization Springer, Berlin, Heidelberg, July 2011, pp. 1-12.

[45] Doucet, A., De Freitas, N., & Gordon, N. , ”An introduction to sequential Monte Carlo methods. In Sequential Monte Carlo methods in practice”. Springer, New York, 2001, pp. 3-14.

[46] Kapanipathi, P., Orlandi, F., Sheth, A. P., & Passant, A. Personalized filtering of the twitter stream, 2011, pp 1-9 accessed on http://corescholar.libraries.wright.edu/knoesis/649.

[47] Shapira, B., Rokach, L., &Freilikhman, S. “Facebook single and cross domain data for recommendation systems.” User Modeling and User-Adapted Interaction, vol- 232-3, 2013,pp. 211-247.

[48] Tim, O. “ What is web 2.0? design patterns and business models for the next generation of software.” https://www.oreilly.com/ , 2005.

[49] Auvinen, A. M. “Social Media-The New Power of Political Influence”. Center for European Studies, 2012, accessed on

https://s3.amazonaws.com/academia.edu.documents/34569938/kansio-digital_democracy_final_en1.pdf.

[50] Edosomwan, S., Prakasan, S. K., Kouame, D., Watson, J., & Seymour, T. "The history of social media and its impact on business". Journal of Applied Management and entrepreneurship, vol. 16. No. 3, 2011. pp. 79-91.

[51] Protalinski, Emil. "Facebook passes 1.19 billion monthly active users, 874 million mobile users, and 728 million daily users." The Next Web (2013) accessed on https://thenextweb.com/facebook/2013/10/ 30/facebook-passes -1-19-billion-monthly-active-users-874-million-mobile-users-728-million-daily-users/

[52] Young, Bob. "Using Social Media for Client Development." Law Prac. Vol. 40., 2014, pp. 6.

[53] Hempel, J. LinkedIn:" how it's changing business and how to make it work for you." Fortune, vol.168 (1), 2013, pp. 68-74.

[54] Humphreys, L., Gill, P., Krishnamurthy, B., & Newbury, E. " Historicizing new media: A content analysis of Twitter.", Journal of communication, vol. 633, 2013, pp. 413-431.

[55] Bernstein, M. S., Suh, B., Hong, L., Chen, J., Kairam, S., & Chi, E. H., "Eddi: interactive topic-based browsing of social status streams". In Proceedings of the 23nd annual ACM symposium on User interface software and technology, ACM, October 2010, pp. 303-312.

[56] The Top 20 Valuable Facebook Statistics, October 2014, accessed on https://zephoria.com/social-media/top-15-valuablefacebook statistics/

[57] Rowe, M., &Ciravegna, F. "Harnessing the social web: The science of identity disambiguation." Web Science Conference, USA., April 2010, pp. 26-27.

[58] Owyang, J., Tran, C., & Webber, A, "The 8 success criteria for Facebook page marketing.", Altimeter Group. 2010, accessed on http://west17media.com/wp-content/uploads/2010/07/facebookreportfinal-100727110656-phpapp02.pdf.

[59] Heckmann, D., Schwartz, T., Brandherm, B., Schmitz, M., & von Wilamowitz-Moellendorff, M. " Gumo–the general user model ontology". In International Conference on User Modeling Springer, Berlin, Heidelberg. July

2005, pp. 428-432.

[60] Gao, Q., Abel, F., &Houben, G. J. " GeniUS: generic user modeling library for the social semantic web". In Joint International Semantic Technology Conference, Springer, Berlin, Heidelberg. December 2011, pp. 160-175.

[61] Mendes, P. N., Passant, A., &Kapanipathi, P. "Twarql: tapping into the wisdom of the crowd", In Proceedings of the 6th International Conference on Semantic Systems, ACM New York, September 2010, pp. 45 – 47

[62] Rowe, M., and Ciravegna, F. "Getting to me exporting semantic social network information from facebook", In The 7th International Semantic Web Conference, Citeseer, 2008, pp. 43.

[63] Bojars, U., Passant, A., & Breslin, J. "Data Portability with SIOC and FOAF." XTech 2008 conference, 2008, accessed on https://aran.library.nuigalway.ie/handle/10379/439.

[64] Vu, X. T., Abel, M. H., & Morizet-Mahoudeaux, P., "A user-centered and group-based approach for social data filtering and sharing.", Computers in Human Behavior, vol. 51, 2015, pp. 1012-1023

[65] Tuulos, V. H., & Tirri, H., "Combining topic models and social networks for chat data mining", In Proceedings of the 2004 IEEE/WIC/ACM international Conference on Web intelligence, IEEE Computer Society, September 2004, pp. 206-213

[66] Gilbert, E., & Karahalios, K., "Predicting tie strength with social media", In Proceedings of the SIGCHI conference on human factors in computing systems, ACM, April 2009, pp. 211-220

[67] Farzindar, A., & Inkpen, D., "Natural Language Processing for Social Media" Synthesis Lectures on Human Language Technologies, vol. 102, 2017, pp. 1-157.

[68] Cambria, E., & White, B. "Jumping NLP curves: A review of natural language processing research", IEEE Computational intelligence magazine, vol. 92, 2014, pp. 48-57.

[69] Ramos, C., Augusto, J. C., & Shapiro, D, "Ambient intelligence—the next step for artificial intelligence", IEEE Intelligent Systems, vol. 232, 2008, pp. 15-18.

[70] Bikel, D., & Zitouni, I., “Multilingual natural language processing applications: from theory to practice”, IBM Press, 2012, pp. 286.

[71] Agrawal R, Srikant R, “Fast algorithms for mining association rules”, In Proceedings of the 20th VLDB conference, 1994, pp. 487–499.

[72] Abdul-Mageed, M., Diab, M., &Kübler, S, “SAMAR: Subjectivity and sentiment analysis for Arabic social media”, Computer Speech & Language, vol. 281, 2014, pp. 20-37.

[73] Cambria, E., Schuller, B., Xia, Y., &Havasi, C, “New avenues in opinion mining and sentiment analysis”, IEEE Intelligent Systems, vol. 282, 2013, pp. 15-21.

[74] Asur, S., & Huberman, B. A, “Predicting the future with social media”, In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Volume 01, IEEE Computer Society, August 2010, pp. 492-499.

[75] Bollen, J., Mao, H., & Pepe, A, “Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena”, Icwsm, vol. 11, 2011, pp. 450-453.

[76] Karabulut, Yigitcan. "Can Facebook predict stock market activity?." (2013).

[77] Lerman, K., & Ghosh, R. , “Information contagion: An empirical study of the spread of news on Digg and Twitter social networks”, Icwsm, vol. 10, 2010, pp. 90-97.

[78] Yessenov, K., & Misailovic, S, “Sentiment analysis of movie review comments. Methodology”, vol. 17, 2009 , pp. 1-7.

[79] Chu, F. C. T., & Asur, S, “Automatic Summarization of Events from Social Media” In ICWSM, July 2013, pp. 81-90.

[80] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A, “The rise of social bots”, Communications of the ACM, vol. 597, 2016, pp. 96-104.

[81] Sharifi, B., Hutton, M. A., & Kalita, J, “Automatic summarization of twitter topics”, In National Workshop on Design and Analysis of Algorithm, Tezpur, India, January 2010, pp. 4-14.

[82] Selvan, M. P., & Selvaraj, R, “Monitoring Fishy activity of the user in social networking”, In Information Communication and Embedded Systems

ICICES, 2017 International Conference on IEEE, 2017, February, pp. 1-5

[83] Mukhra, R., Baryah, N., Krishan, K., &Kanchan, T, “Blue Whale Challenge: A Game or Crime?”, Science and engineering ethics, 2017, pp. 1-7.

[84] Zhou, J., Tang, M., Tian, Y., Al-Dhelaan, A., Al-Rodhaan, M., & Lee, S, “Social Network And Tag Sources Based Augmenting Collaborative Recommender System”, IEICE transactions on Information and Systems, vol. 984, 2015, pp. 902-910.

[85] Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G, “Finding high-quality content in social media”, In Proceedings of the 2008 international conference on web search and data mining, ACM, February 2008, pp. 183-194

[86] Collins, M., “Head-driven statistical models for natural language parsing”, Computational linguistics, vol. 294, 2003, pp. 589-637.

[87] Taylor, A., Marcus, M., & Santorini, B, “The Penn treebank: an overview”, In Treebanks, Springer Netherlands, 2003, pp. 5-22

[88] Carminati, B., Ferrari, E., & Perego, A, “Rule-based access control for social networks”, In OTM Confederated International Conferences On the Move to Meaningful Internet Systems, Springer, Berlin, Heidelberg, October 2006, pp. 1734-1744

[89] Dabbagh, N., & Kitsantas, A, “Personal Learning Environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning”, The Internet and higher education, vol. 151, 2012, pp. 3-8.

[90] Singla, P., & Domingos, P, “Entity resolution with markov logic In Data Mining”, ICDM'06. Sixth International Conference on IEEE, December 2206, pp. 572-582

[91] Han, S., He, D., Jiang, J., & Yue, Z., “Supporting exploratory people search: a study of factor transparency and user control”, In Proceedings of the 22nd ACM international conference on Information & Knowledge Management, Vol.6, October 2013, ACM, pp. 449-458.

[92] Gilbert, C. H. E. ,Vader: “A parsimonious rule-based model for sentiment analysis of social media text”. In Eighth International Conference on Weblogs and Social Media ICWSM-14, Georgia Institute of Technology,

Atlanta,, 2014., Available at 20/04/16 http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf. pp. 216-225.

[93] Zafarani, R., & Liu, H., “Connecting users across social media sites: a behavioral-modeling approach.” ,In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Computer Science and Engineering Arizona State University 2013, August, pp. 41-49. ACM.

[94] Fan, W., & Gordon, M. D. “The power of social media analytics”. Communications of the ACM, New York, NY, USA, vol.576, 201, pp. 74-81.

[95] Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., & Doan, “A. Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach”. Proceedings of the VLDB Endowment, University of Wisconsin-Madison 2013, vol.611, pp.1126-1137.

[96] Weerkamp, Wouter, Richard Berendsen, Bogomil Kovachev, Edgar Meij, Krisztian Balog, and Maarten De Rijke. "People searching for people: Analysis of a people search engine log." In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, 2011, pp. 45-54.

[97] Wolfram, M. Sebastian A. "Modelling the stock market using Twitter." Scotland, UK: University of Edinburgh (2010), accessed on https://pdfs.semanticscholar.org/506f/f2711e234466b46a2207d8f16d7e9f324189.pdf.

[98] Surdeanu, M., Tibshirani, J., Nallapati, R. and Manning, C.D., Multi-instance multi-label learning for relation extraction. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning , July 2012, pp. 455-465.

[99] Ralph Gross and Alessandro Acquisti. “Information revelation and privacy in online social networks”. In Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society, WPES, vol.05, New York, NY, USA, 2005, pp.71–80.

[100] Weng, J., Lim, E. P., Jiang, J., & He, Q. “Twitterrank: Finding topic-sensitive influential twitterers”,. In Proceedings of the third ACM international conference on Web search and data mining, Pennsylvania State University,

February 2010, pp. 261-270.

[101] Carmagnola, F., Osborne, F., & Torre, I.. “User data distributed on the social web: how to identify users on different social systems and collecting data about them”. In Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems, Barcelona, Spain, September 2010, ACM, pp. 9-15.

[102] Kautz, H., Selman, B., & Shah, M. “Referral Web: combining social networks and collaborative filtering”. Communications of the ACM, vol.403, 1997, pp. 63-65.

[103] Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K., & Ishizuka, M, “POLYPHONET: an advanced social network extraction system from the web”. Web Semantics: Science, Services and Agents on the World Wide Web, vol. 54, 2007, pp.262-278.

[104] Mika, P. “Social networks and the semantic web”. In Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence . IEEE Computer Society, September 2004, pp. 285-291. .

[105] Tyler, J. R., Wilkinson, D. M., &Huberman, B. A., “Email as spectroscopy: Automated discovery of community structure within organizations.”, In Communities and technologies Springer, Dordrecht., 2003, pp. 81-96.

[106] Freeman, L. C., “A set of measures of centrality based on betweenness. Sociometry” , JSTOR, American Sociological Association, vol. 40 , No. 1, 1977, pp 35-41.

[107] Rohani, V. A., Kasirun, Z. M., Kumar, S., & Shamshirband, S., “An effective recommender algorithm for cold-start problem in academic social networks.” Mathematical Problems in Engineering, 2014, pp 1-11.

[108] Szomszor, M., Alani, H., Cantador, I., O’Hara, K., &Shadbolt, N. “Semantic modelling of user interests based on cross-folksonomy analysis.” In Proceedings of the International Semantic Web ConferenceSpringer Berlin Heidelberg, 2008., pp. 632-648

[109] Zhou, M., Zhang, W., Smith, B., Varga, E., Farias, M., &Badenes, H. 2012, “February. Finding someone in my social directory whom i do not fully remember or barely know.” In Proceedings of the 2012 ACM international

conference on Intelligent User Interfaces pp. 203-206.

[110] Groh, G., &Hauffa, J. “Characterizing Social Relations Via NLP-Based Sentiment Analysis.” In ICWSM, July, 2011, pp 502-506.

[111] Mukhopadhyay, D., &Kulkarni, S., “An Approach to Design an IoT Service for Business Domain Specific Web Search.”, In Proceedings of the International Conference on Data Engineering and Communication Technology, Springer Singapore, 2017, pp. 621-628.

[112] Sun, G., Xie, Y., Liao, D., Yu, H., & Chang, V., “User-defined privacy location-sharing system in mobile online social networks.” Journal of Network and Computer Applications, Vol. 86, 2017, pp. 34-45.

[113] Chen, X., Zhang, C., Hu, Y., Ge, B., & Xiao, W. ,”Temporal Social Network: Group Query Processing.”, 27th International Workshop on Database and Expert Systems Applications DEXA”, IEEE, September 2016, pp. 181-185.

[114] Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., & Tziritas, N. “A survey on text mining in social networks.”, The Knowledge Engineering Review, 2015, vol 30(2), 157-170.

[115] Lampe, C., Ellison, N., &Steinfield, C., “A Face book in the crowd: Social searching vs. social browsing.” In Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, November 2006, ACM, pp. 167-170.

[116] Irani, D., Webb, S., Li, K., &Pu, C. “Large online social footprints--an emerging threat”. International Conference on Computational Science and Engineering, 2009. CSE'09. Vol. 3, pp. 271-276.

[117] Zheleva, E., &Getoor, L. “To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles.” In Proceedings of the 18th ACM International Conference on World Wide Web, 2009, pp. 531-540.

[118] Goga, O., Lei, H., Parthasarathi, S. H. K., Friedland, G., Sommer, R., & Teixeira, R. “Exploiting innocuous activity for correlating users across sites.” In Proceedings of the 22nd ACM international conference on World Wide Web, May 2013, pp. 447-458.

[119] Malhotra, A., Totti, L., MeiraJr, W., Kumaraguru, P., & Almeida, V.

"Studying user footprints in different online social networks." 2012 IEEE/ACM International Conference on In Advances in Social Networks Analysis and Mining ASONAM, 2012, pp. 1065-1070.

[120] Chen, Z., Kalashnikov, D. V., & Mehrotra, S. Adaptive graphical approach to entity resolution. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries ,University Of California, Irwin, ACM, 2007, pp. 204-213.

[121] Bhattacharya, I., & Getoor, L. "Collective entity resolution in relational data." ACM Transactions on Knowledge Discovery from Data TKDD, Vol. 1(1), 2007.

[122] Motoyama, M., & Varghese, G. "I seek you: searching and matching individuals in social networks." In Proceedings of the ACM eleventh international workshop on Web information and data management, Indraprastha Institute Of Technology,Delhi.India, University Of Maryland, Baltimore County, Baltimore, MD,USA, 2009, pp. 67-75

[123] Peled, O., Fire, M., Rokach, L., & Elovici, Y. Entity matching in online social networks. In Social Computing SocialCom, 2013 International Conference on IEEE, September 2013, pp. 339-344.

[124] Raad, E., Chbeir, R., & Dipanda, A.. User profile matching in social networks. In Network-Based Information Systems NBiS, 2010 13th International Conference on. IEEE, September 2010,, pp. 297-304

[125] Vosecky, J., Hong, D., & Shen, V. Y. , User identification across multiple social networks, In Networked Digital Technologies, 2009. NDT'09. First International Conference on IEEE, Volume 2 Number 1, Hong Kong University of Science and Technology Hong Kong, July 2009, pp. 360-365.

[126] Jain, P., Kumaraguru, P., & Joshi, A. , @ i seek'fb. me': Identifying users across multiple online social networks, In Proceedings of the 22nd international conference on World Wide Web ACM, Indraprastha Institute of Information Technology IIIT-Delhi, India ,University of Maryland, Baltimore County UMBC, USA, May 2013, pp. 1259-1268.

[127] Irani, D., Webb, S., Li, K., & Pu, C., "Large online social footprints--an emerging threat." IEEE International Conference on In Computational Science and Engineering, CSE'09. Vol. 3, August 2009, pp. 271-276.

[128] Bilge, L., Strufe, T., Balzarotti, D., & Kirda, E. All your contacts are belong to us: automated identity theft attacks on social networks. In Proceedings of the 18th international conference on World wide web ACM, April 2009, pp. 551-560.

[129] Narayanan, A., &Shmatikov, V., "De-anonymizing social networks." 30th IEEE Symposium on Security and Privacy, May 2009, pp. 173-187

[130] Kabay, M. E. "Privacy issues in social-networking sites." Network World, 27, 2010, accessed on https://www.networkworld.com/article /2237560/collaboration-social/privacy-issues-in-social-networking-sites.html

[131] Korula, N., & Lattanzi, S., "An efficient reconciliation algorithm for social networks", Proceedings of the VLDB Endowment, vol.75, 2014, pp. 377-388.

[132] Carmagnola, F., &Cena, F. "User identification for cross-system personalisation." Information Sciences, vol. 1791, 2009, pp.16-32.

[133] Labitzke, S., Taranu, I., & Hartenstein, H., What your friends tell others about you: Low cost linkability of social network profiles, In Proc. 5th International ACM Workshop on Social Network Mining and Analysis, San Diego, CA, USA, Steinbuch Centre for Computing & Institute of Telematics Karlsruhe Institute of Technology KIT, Germany, August 2011, pp. 1065-1070.

[134] Srivatsa, M., & Hicks, M., Deanonymizing mobility traces: Using social network as a side-channel. In Proceedings of the 2012 ACM conference on Computer and communications security , ACM, IBM T.J. Watson Research Center, University of Maryland, October 2012, pp. 628-637.

[135] Acquisti, A., Gross, R., & Stutzman, F.. Faces of facebook: Privacy in the age of augmented reality, 2011 accessed on http://marchiondelli.com/Blog/wp-content/uploads/2013/10/acquisti-face-BH-Webinar-2012-out.pdf

[136] Perito, D., Castelluccia, C., Kaafar, M. A., &Manils, P.. "How unique and traceable are usernames?". In International Symposium on Privacy Enhancing Technologies Symposium Springer Berlin Heidelberg., July 2011,, pp. 1-17.

[137] Liu, L., Cheung, W. K., Li, X., & Liao, L. "Aligning Users across

Social Networks Using Network Embedding.” In IJCAI, July 2016, pp. 1774-1780.

[138] Bartunov, S., Korshunov, A., Park, S. T., Ryu, W., & Lee, H., Joint link-attribute user identity resolution in online social networks. In Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. ACM, 2012.

[139] Zhang, H., Kan, M. Y., Liu, Y., & Ma, S. “Online social network profile linkage.” In Asia Information Retrieval Symposium, Springer, Cham., December 2014, pp. 197-208.

[140] You, Q., Bhatia, S., Sun, T., & Luo, J. , The eyes of the beholder: Gender prediction using images posted in online social networks. In Data Mining Workshop ICDMW, 2014 IEEE International Conference on IEEE, December 2014, pp. 1026-1030.

[141] Yuanping Nie and Jiuming Huang and Aiping Li and Bin Zhou.,"Identifying Users Based on Behavioral-Modeling across Social Media Sites," Web Technologies and Applications, vol. 8709, 2014, pp. 48-55,

[142] Zhou, X., Liang, X., Zhang, H., & Ma, Y. “Cross-platform identification of anonymous identical users in multiple social media networks.” IEEE transactions on knowledge and data engineering, vol.282, 2016, pp. 411-424.

[143] Raad, E., Chbeir, R., & Dipanda, A. “User profile matching in social networks.” 13th IEEE International Conference on In Network-Based Information Systems NBiS, September 2010, pp. 297-304.

[144] Cortis, K., Scerri, S., Rivera, I., & Handschuh, S. “Discovering semantic equivalence of people behind online profiles.” In Proceedings of the Resource Discovery RED Workshop, ser. ESWC, 2012, pp, 104- 118.

[145] Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. “Multinomial naive bayes for text categorization revisited.” In Australasian Joint Conference on Artificial Intelligence, Springer, Berlin, Heidelberg, December 2004, pp. 488-499.

[146] Madnick, S., & Siegel, M. ,“Seizing the opportunity exploiting web aggregation.” MIT Sloan Working Paper No. 4351-1, 2001, pp. 1-15 accessed at: https://ssrn.com/abstract=303827 or http://dx.doi.org/10.2139/ ssrn.303827

.

[147] Arazy, O., Kumar, N., & Shapira, B. , “A theory - driven design framework for social recommender systems.” Journal of the Association for Information Systems, vol.11(9), 2010, pp. 455 – 490.

[148] Berkovsky, S., Kuflik, T., Ricci, F.,:”Mediation of User Models for Enhanced Personalization in Recommender Systems.” Journal of User Modeling and User-Adapted Interaction, vol. 183, 2008, pp. 245–286,

[149] Guy, I., Jacovi, M., Shahar, E., Meshulam, N., Soroka, V., & Farrell, S. , “Harvesting with SONAR: the value of aggregating social network information.” In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems ACM, April 2008, pp. 1017-1026.

[150] Carmagnola, F., Vernero, F., & Grillo, P. , Sonars: “A social networks-based algorithm for social recommender systems.” In International Conference on User Modeling, Adaptation, and Personalization, Springer, Berlin, Heidelberg, June 2009, pp. 223-234.

[151] Singh, L., Yang, G. H., Sherr, M., Hian-Cheong, A., Tian, K., Zhu, J., & Zhang, S. “Public information exposure detection: Helping users understand their web footprints.” In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM, 2015, pp. 153-161.

[152] Abel, F., Herder, E., Houben, G. J., Henze, N., & Krause, D. “Cross-system user modeling and personalization on the social web”. User Modeling and User-Adapted Interaction, vol. 232(3), 2013, p. 169-209.

[153] Pontual, M., Gampe, A., Chowdhury, O., Kone, B., Ashik, M. S., & Winsborough, W. H. “The privacy in the time of the Internet: Secrecy vs transparency”. In Proceedings of the Second ACM Conference on Data and Application Security and Privacy, 2012, pp. 133-140.

[154] Yudelson, M., Brusilovsky, P., & Zadorozhny, V. “A user modeling server for contemporary adaptive hypermedia: An evaluation of the push approach to evidence propagation”. In International conference on user modeling, Springer, Berlin, Heidelberg, vol. 4511, July 2007, pp. 27-36.

[155] Assad, M., Carmichael, D. J., Kay, J., & Kummerfeld, B. , “PersonisAD: Distributed, active, scrutable model framework for context-

aware services". In International Conference on Pervasive Computing, Springer, Berlin, Heidelberg, Vol. 4480, May 2007, pp. 55-72.

[156] Otte, E., & Rousseau, R. “Social network analysis: A powerful strategy, also for the information sciences.” Journal of information Science, vol. 286, 2002, pp. 441-453.

[157] Ayad, H. G., & Kamel, M. S. “Cumulative voting consensus method for partitions with variable number of clusters.” IEEE transactions on pattern analysis and machine intelligence, Vol. 301, 2008, pp. 160-173.

[158] Singh, V., Mukherjee, L., Peng, J., & Xu, J. “Ensemble clustering using semidefinite programming with applications.” Machine learning, Vol.79, Issue 1-2, 2010, pp. 177-200.

[159] Bhatnagar, V., & Ahuja, S. “July. Robust clustering using discriminant analysis.” In Industrial Conference on Data Mining, Springer Berlin Heidelberg., Vol. 6171, 2010, pp. 143-157.

[160] Jain, A. K. “Data clustering: 50 years beyond K-means.” Pattern recognition letters, Vol. 31, Issue 8, 2010, pp. 651-666.

[161] Sun, P. G., Gao, L., & Han, S. S. “Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks.” Information Sciences, Vol. 1816, 2011, pp. 1060-1071.

[162] Tang, L., & Liu, H. “Scalable learning of collective behavior based on sparse social dimensions”. In Proceedings of the 18th ACM conference on Information and knowledge management, ACM, November 2009, pp. 1107-1116.

[163] Zhang, S., Wang, R. S., & Zhang, X. S., “Identification of overlapping community structure in complex networks using fuzzy c-means clustering.” Physica A: Statistical Mechanics and its Applications, Vol. 3741, 2007, pp. 483-490.

[164] Yang, X., Wang, Y., Wu, D., & Ma, A., “K-means based clustering on mobile usage for social network analysis purpose.” In 2010 6th International Conference on Advanced Information Management and Service IMS, Seoul, South Korea, IEEE, November 2010, pp. 223-228.

[165] Li, X., Huang, Y., Li, S., & Zhang, Y., May. Hybrid retention strategy formulation in telecom based on k-means clustering analysis. In 2011

International Conference on E-Business and E-Government ICEE, Shanghai, China, 2011, pp. 1-4.

[166] Qu, Z., & Liu, Y. , “Interactive group suggesting for Twitter.”, In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-, Association for Computational Linguistics, Vol. 2, June 2011, pp. 519-523.

[167] Blei, D. M., Ng, A. Y., & Jordan, M. I. “Latent dirichlet allocation.”, Journal of machine Learning research, Jan 2003, pp. 993-1022.

[168] Rakesh, V., Singh, D., Vinzamuri, B., & Reddy, C. K. , “Personalized Recommendation of Twitter Lists using Content and Network Information.” In ICWSM, June 2014, pp. 416-425.

[169] Vega-Pons, S., Correa-Morris, J., & Ruiz-Shulcloper, J., September. “Weighted cluster ensemble using a kernel consensus function.” In Iberoamerican Congress on Pattern Recognition, Springer Berlin Heidelberg, Vol. 5197, 2008, pp. 195-202.

[170] Mirkin, B., “Mathematical Classification and Clustering, Nonconvex Optimization and Its Applications”, Vol. 11, Pardalos, P. and Horst, R., editors, 1996.

[171] Kuhn, Harold W. "The Hungarian method for the assignment problem." Naval Research Logistics (NRL) vol. 52, No. 1, 2005, pp. 83-97.

[172] Dimitriadou, E., Weingessel, A., &Hornik, K., “Voting-merging: An ensemble method for clustering.” In International Conference on Artificial Neural Networks, Springer Berlin Heidelberg, Vol. 2130, August 2001, pp. 217-224.

[173] Yoon, H. S., Ahn, S. Y., Lee, S. H., Cho, S. B., & Kim, J. H., “Heterogeneous clustering ensemble method for combining different cluster results.” In International Workshop on Data Mining for Biomedical Applications, Springer Berlin Heidelberg, Vol. 3916, April 2006,pp. 82-92.

[174] Weingessel, A., Dimitriadou, E., & Hornik, K. “An ensemble method for clustering.” In Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vol. 62, 2003, pp.143-151.

[175] Bourke, Steven, Michael O'Mahony, Rachael Rafter, and Barry Smyth. "Ranking in information streams." In Proceedings of the companion

publication of the 2013 international conference on Intelligent user interfaces companion, ACM, 2013, pp. 99-100.

[176] Hannon, J., Bennett, M., & Smyth, B. , “Recommending twitter users to follow using content and collaborative filtering approaches.” In Proceedings of the fourth ACM conference on Recommender systems, September 2010, pp. 199-206.

[177] Lim, K. H., & Datta, A. “Finding twitter communities with common interests using following links of celebrities.” In Proceedings of the 3rd international workshop on Modeling social media, June 2012, pp. 25-32.

[178] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su., “ Optimizing web search using social annotations. In Proceedings of WWW, New York, NY, USA, 2007, pp. 501–510.

[179] Widman, J. Edgerank. Retrieved August 3, 2014 accessed on http://edgerank.net

[180] Christen, P. “A comparison of personal name matching: Techniques and practical issues.” IEEE International Conference on Data Mining Workshops, 2006. ICDM Workshops, December 2006, pp. 290-294.

[181] Hardt, D. 2012. The OAuth 2.0 authorization framework, accessed on https://www.rfc-editor.org/rfc/pdfrfc/rfc6749.txt.pdf.

[182] Virmani, Charu, Anuradha Pillai, and Dimple Juneja. "Clustering in Aggregated User Profiles Across Multiple Social Networks." International Journal of Electrical and Computer Engineering (IJECE) Vol. 7, no. 6, 2017, pp. 3692-3699.

[183] Arun, K., and M. Gomathy Nayagam. "Building Applications with Social Networking API's." International Journal of Advanced Networking and Applications vol. 5, no. 5, 2014 pp. 2070- 2075.

[184] Rapaport, Jeffrey A., Seymour Rapaport, Kenneth Allen Smith, James Beattie, and Gideon Gimlan. "Social network driven indexing system for instantly clustering people with concurrent focus on same topic into on-topic chat rooms and/or for generating on-topic search results tailored to user preferences regarding topic." U.S. Patent 8,539,359, issued September 17, 2013.

[185] Karp, Daniel, Yves Schabes, Martin Zaidel, and Dania Egedi. "A

freely available wide coverage morphological analyzer for English." In Proceedings of the 14th conference on Computational linguistics-Volume 3, Association for Computational Linguistics, 1992, pp. 950-955.

[186] Porter, M., “An algorithm for suffix stripping, Program”, vol. 143, 1980, pp. 130–137.

[187] Turney, Peter D., "Mining the web for synonyms: PMI-IR versus LSA on TOEFL." In European Conference on Machine Learning, Springer, Berlin, Heidelberg, 2001, pp. 491-502.

[188] Landauer, Thomas K., and Susan T. Dumais. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." Psychological review 104, vol. 2 ,1997, pp. 211.

[189] Shirahata, Koichi, Hitoshi Sato, Toyotaro Suzumura, and Satoshi Matsuoka. "A scalable implementation of a mapreduce-based graph processing algorithm for large-scale heterogeneous supercomputers." In Cluster, Cloud and Grid Computing CCGrid, 2013 13th IEEE/ACM International Symposium, IEEE, 2013, pp. 277-284.

[190] Estivill-Castro, V. “Why so many clustering algorithms: A position paper.”, SIGKDD Explorations Newsletter, 41, 2002, pp. 65-75.

[191] Topchy, A., Minaei-Bidgoli, B., Jain, A. K., & Punch, W. F. “Adaptive clustering ensembles.”, ICPR 2004. Proceedings of the 17th International Conference on Pattern Recognition, Vol. 1, August 2004, pp. 272-275.

[192] Zimek, Arthur, Ricardo JGB Campello, and Jörg Sander. "Ensembles for unsupervised outlier detection: challenges and research questions a position paper." ACM SIGKDD Explorations Newsletter 15, vol. 1 2014, pp. 11-22.

[193] Aggarwal, Charu C., and Chandan K. Reddy, eds. Data clustering: algorithms and applications. CRC press, 2013, accessed on http://charuaggarwal.net/clusterbook.pdf.

[194] Bhat, Sajid Yousuf, Muhammad Abulaish, and Abdulrahman A. Mirza. "Spammer classification using ensemble methods over structural social network features." In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence WI and Intelligent Agent Technologies

IAT, IEEE Computer Society, 2014., Volume 02, pp. 454-458.

[195] Leskovec, Jure, and Julian J. Mcauley. "Learning to discover social circles in ego networks." In Advances in neural information processing systems, 2012, pp. 539-547.

[196] Abel, Fabian, Qi Gao, Geert-Jan Houben, and Ke Tao. "Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web." In Proceedings of the 3rd International Web Science Conference, ACM, 2011, p. 2-9.

[197] Carenini, G., Cheung, J. C. K., & Pauls, A. “Multi Document summarization of evaluative text”, Computational Intelligence, 29(4), (2013), pp. 545-576.

[198] Rand, William M. "Objective criteria for the evaluation of clustering methods." Journal of the American Statistical association vol. 66, Issue 336, 1971, pp. 846-850.

[199] Ben-Hur, Asa, Andre Elisseeff, and Isabelle Guyon. "A stability based method for discovering structure in clustered data." In Biocomputing 2002, pp. 6-17.

[200] Chen, Sheng, Bernard Mulgrew, and Peter M. Grant. "A clustering technique for digital communications channel equalization using radial basis function networks." IEEE Transactions on neural networks 4, vol. 4, 1993, pp. 570-590.

BRIEF PROFILE OF RESEARCH SCHOLAR

Charu Virmani did her M.Tech. (Computer Science and Engineering) from Maharishi Dayanand University, Rohtak in 2007 and B.Tech. (Information Technology) from Maharishi Dayanand University, Rohtak in 2005. Ms. Virmani has over 13 years of experience in teaching B.tech and M.tech courses. Her areas of interest include Computer Networks, Semantic Web, Web Mining and Social Networks. She has published 21 research papers in various journals and conferences of international fame. Currently, she is working as Associate Professor in the department of Computer Science & Engineering at Manav Rachna International Institute of Research and Studies, Faridabad.

List of Publication out of Thesis

List of Published Papers

S.No.	Title of the paper	Name of the Journal Publisher	No.	Volume & Issue	Year	Pages
1	“Extracting Information from Social Network using NLP”, UGC Approved	International Journal of Computational Intelligence Research (IJCIR), RIP – INDIA	4	13	2017	621-629
2	“Characterising User Demographics across social network”, UGC Approved	International Journal of Advanced Research	6	5	2017	2308 – 2312.
3	“Major Encounters to search the Social Network”, UGC Approved	International Journal of Advance Research in Computer Science and Software Engineering	6	7	2017	504-508
4	“Clustering in Aggregated User Profiles Across Multiple Social Networks”, SCOPUS and UGC Approved ,	International Journal of Electrical and Computer Engineering, IAES Journal	6	7	2017	3692 – 3699
5	“Design of a Hybrid Integrator for Autonomous Social Networks”, SCOPUS and UGC Approved	International Journal of Computer Information systems and Industrial Management, MIR Labs	7	9	2017	241-248
6	“Design of Query Processing System to Retrieve Information from Social Network using NLP”, SCIE, SCOPUS AND UGC Approved	KSII Transactions on Internet and Information Systems, KSII	3	12	2018	1168-1188

List of Accepted Papers

7. Virmani, C., Pillai, A., & Juneja, D. (2018) "Design of A Novel Query System for Social Network", Journal of Information Technology and Research (SCOPUS AND ESCI INDEXED)

List of papers in Scopus/Web of Science in Conference

8. Virmani, C., Pillai, A., & Juneja, D. (2014, February). "Study and analysis of Social Network Aggregator". In Optimization, Reliability, and Information Technology (ICROIT), 2014 International Conference on (pp. 145-148). IEEE. (SCOPUS AND WEB OF SCIENCE INDEXED).
9. Virmani, C., Pillai, A., & Juneja, D, "Content-Based Social Network Aggregation." In ICT Based Innovations, pp. 185-194. Springer, Singapore, 2018. (SCOPUS INDEXED)

List of papers in National/International Conference

10 Virmani, C., Pillai, A., & Juneja, D, "Major Challenges in Intelligent Social Network Database", National conference on New Horizons in Technology for Sustainable Energy and Environment (NHTSEE 2017), YMCA University of Science and Technology, March, 2017.

11 Virmani, C., Pillai, A., & Juneja, D, "Comparison of various methods to recognize the digital impression of the user across social network", Advances in Mathematics and Computing, YMCA University of Science and Technology, May, 2017.

12 Virmani, C., Pillai, A., & Juneja, D, "A Novel method to Filter relevant users from online social networks", Advances in Mathematics and Computing, YMCA University of Science and Technology, May, 2017.

13 Virmani, C., Juneja, D, & Pillai, A., "Social Networks- A REVIEW", International conference on Science in Hindi, NIT Kurukshetra, August 2017.