UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL AND APPLIED SCIENCES

Electronics and Computer Science

Purposive Social Network and Linked Data

by

Priyanka Singh

Thesis for the degree of Doctor of Philosophy

February 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCES
Electronics and Computer Science

Doctor of Philosophy

PURPOSIVE SOCIAL NETWORK AND LINKED DATA

by Priyanka Singh

In the era of social web people communicate, collaborate and share online, the social media and social networking is amalgamated together. People live their private and professional lives on the web and connect to their friends, colleagues and peers with common interest using different online mediums and services. Some common means of online communication are emails, instant chat, messaging, forums and content specific social networking websites like Facebook, Twitter and YouTube.

Internet is also an easy medium for people to collaborate and solve problems by collective thinking and effort. Crowdsourcing is an efficient feature of social web where people with common interest and expertise come together to solve specific problems and create a community. Crowdsourcing techniques can also be used to filter out the important information from the collection of large data and remove spams, and gamification techniques are used to reward the users for their contribution and keep a sustainable environment for the growth of the community and network.

Semantic web technologies can be used to structure and community data so it can be decentralized and be used across platform. Using semantic web tools and ontologies different networks can be combined and knowledge can be enhanced and easily discovered and merged together.

This report discusses the concept of purposive social network where people with similar interest and varied expertise come together, use crowdsourcing technique to solve a common problem and build tools for common purpose. StackOverflow and Reddit websites are chosen to study the purposive network, different network ties and roles of user is studied and also Linked Data and semantic web technologies is used for name disambiguation of keywords and topics. This also helps in easier search and discovery of experts in a field and provide useful information that is otherwise unavailable in the website.

Contents

List of Figures

List of Tables

Acknowledgements

I want to take this opportunity to thank my family, especially my mother and my sisters for their support that helped me through the tough times and encouraged me to keep going with my research. I also want to thank my supervisors for their guidance and understanding.

I also want to acknowledge all the online forums and question and answering websites that helped me when I had problems with my programs and providing me with insightful information.

Nomenclature

WWW	World Wide Web
SNS	Social Networking Services
RDF	Resource Description Framework
RDFa	Resource Description Framework-in-attributes
HTTP	Hyper Text Transfer Protocol
FOAF	Friend Of A Friend
SIOC	Semantically Interlinked Online Community
OWL	Web Ontology Language
SSL	Secure Socket Language
RDFS	RDF Schema
SCOT	Social Semantic Cloud of Tags
MOAT	Meaning Of A Tag
OPO	Online Presence Ontology

1‘11q X 6GGYY5T4RTTT NMM./]

87

Chapter 1

Introduction

In the beginning of World Wide Web (WWW) the documents were linked with each other and with the advent of browsers and search engines, it became easier to search for documents. People created static web page to share information and opinions and used emails to communicate with other people. Web 2.0 has changed how people use and interact in Internet, they moved from static content to more dynamic sharing of information and started to connect with people and objects.

Now in WWW, not only documents are linked with each other using hyperlinks, people are also connected with each other using explicit relationship or through data, creating a personal social network. Data is not shared mere through webpages, it is shared using the social network of people and information travels faster using the social networking effect and creates larger impact.

There are many different Social networking Service (SNS) available fore people to sign up, create a profile by adding personal information and create links and relationships with friends, families or colleagues. Different SNS provides different functionalities, for e.g., Facebook [1] is for people to connect with their friends, LinkedIn [2] to connect with their colleagues and Twitter [3] for microblogging. There are also content specializing services, YouTube [4] for videos, Flickr [5] for pictures, GoodReads [6] for books.

The main drawback of current services is that they are all closed network and the SNS own all the users data i.e. users cannot take their data away with them once they delete their account and leave that website or the website closes down. Another drawback of this system is that if people want to sign up to different website to use their functionality, they have to undergo the same cycle of creating their user profile and add friends to

[1] http://www.facebook.com/
[2] https://www.linkedin.com/
[3] http://twitter.com/
[4] http://www.youtube.com/
[5] http://www.flickr.com/
[6] http://www.goodreads.com/

their network again, they cannot migrate their data and social network from one SNS to another. All their data, activities, relationships are trapped inside a silo and the service provider is the owner and in control of the data. This whole cycle of recreating profile and reconnecting with friends and colleagues in different SNS, Brad Fitzpatrick referred as Social Network Fatigue **?**.

The Web 2.0 social web has also given a distributed platform for people to come together from different part of the world and collaborate together to solve problems. People form groups and communities based on similar interest and purpose and use the collective intelligence for distributed problem solving and create knowledge. This crowdsourcing technique is different from human-based computing, here people broadcast any problems they have and other users and experts answers and submit solution. Experts and people with similar interest connect with each other, create relationships and communities with strong and weak social bonds. Messaging boards, question-answer forums, wikis are example of this type of social communities where people come together to create an emerging knowledge.

Semantic Web, as envisioned by Sir Tim Berners-Lee is the extension of the World Wide Web that enables people to share content beyond websites, applications and platform. It is an intelligent web of structured data where each resource has a URI and is represented in RDF triples. It is machine-readable and ontologies are implemented to describe the data and to give it meanings, understood by both machines and humans **?**. Semantic Web technologies provides appropriate tools to represent and structure the social networking data to make it portable and it also connects data from different portals to form a decentralized system that can be easily queried across platform and reused.

1.1 Research question and challenge

1.1.1 Problem with current social networking services

There are many social networking websites available on the WWW and each of them has a particular purpose and they target a particular user group. These website are centralized where people sign up, create a profile and invite their friends and create friendships to communicate with each other and share information. These websites can be open, where all the data is visible to all the users, or closed, where only the user sign-in to the website can view any information. In most of the cases, the websites are semi-open where the registered user can see everything and the other user can see partial information on the website. So, all the information is not accessible to everyone.

The other issue with these websites is as previously mentioned is that user does not own the content they created, it is controlled by the website. If any user wishes to delete it's account, they loose all their data, they cannot export their data, their interaction with

their friends and migrate to another website. They loose all their data and information once they leave the website. These websites are centralized and have their own APIs, structure and people cannot work across the platform and share information across different social networks. So, it is difficult for a user to maintain multiple accounts and creating different accounts and friendships causes social network fatigue.

Another issue with these services is the same as with the web, search and discovery of information. The search engines can only retrieve information when explicitly asked, it does not return the solution of a problem if it doesn't exist on a webpage. Using the search function of the website only returns result from their own network. People who want advice and opinion of an expert can join a forum or messaging board or join and email-list but they can only find the subset of experts registered to that service, they loose a whole community of expert in other network or community.

The amount of data created everyday on the web has increased exponentially in the recent years and a fast access to accurate information and key people is essential in the fast moving life and also in the knowledge web. The social networking websites in today's web are like small and separate islands and they need bridges to join together, have a common structure so the heterogeneous datasets could be integrated together into a homogeneous view to get the full potential of all the services.

1.1.2 Research challenges

The main challenge with the current structure of Word Wide Web and the social networking website is opening the closed data by any users or on any topics. These data then are required to be represented in a common structure so it can be integrated together to give a complete set of information. The main research question and challenges encountered are discussed below.

1.1.2.1 Data Harvest and Data Integration

The study of online social networks requires user data from the online social networks, forums and communities. Some of forums and boards are open to public and data can be harvested freely using the API and simple screen scraping but most of the social networking websites are closed due to privacy policies and data is not readily available to general public for consumption.

Data from different sources are harvested and structured using the semantic web technology and integrated together. Integration of data from multiple sources causes name entity problem for main topics and also a user might have multiple accounts in different system, so the information from all these different accounts need to be integrated to form a complete and homogeneous user profile. In order to do so, it is possible that the

same entities may be retrieved by different data providers that adopted different URIs and names, making paramount to solve problems of coreference and name ambiguity **?**.

1.1.2.2 Purposive Social Network formation

The concept of a network is very broad that can be associated with any kind of relation between people that can be identified by the set of entities. A social network that is formed by people with a common interest, a goal, an objective or purpose with explicit or implicit relation is studied in this case and named purposive social network.

This network can be a small, temporary community of people solving a problem and then dispersing once the task is finished. The relation between people and objects are varied with different attributes.

In this report, this type of purposive social network is studied in detail, how it is formed, why it is created, it?s different attributes and characteristics are identified. Also, the motivation of people to contribute and collaborate in such network is studied and what type of emergent knowledge and results are generated with the community is analyzed.

1.1.2.3 Network Analysis and initial result

To analyze a purposive social network an open online question and answer forum for programmers and developers, StackOverflow [7], is studied. This website is used by computer programmers to ask questions on any topics, other expert programmers answer the questions and other users of the website vote the questions and answers and keep quality control.

This website is open to the public and the API is easy to use to collect the data. This is also a good example of purposive social networks where people come together and share knowledge and help solve a problem and gain reputations. Crowdsourcing is used to create an emergent knowledgebase, to filter spam and keep the quality of questions and answers high and the incentive system keeps user motivated to contribute and solve problems.

The public data of the website is analyzed and different characteristics of the purposive social network is studied. Attributes like relationship/links between users, and users and objects that connect every person and entity in the website is analyzed. The incentive model of the website that encourage people to collaborate and contribute is also studied. The structure of the communication and network growth over time is also analyzed **?**.

[7] http://stackoverflow.com/

1.1.2.4 Linking the Data

The website data is collected and analyzed using Wikipedia-miner [8] **?** and OpenCalis [9] toolkit. These tool uses natural language processing to find the main keywords from the text and use machine learning algorithm to match the keywords to topics and particular vocabulary. Then the posts are structured into RDF and the main entities are matched with Wikipedia articles and Drupal vocabulary. This data is linked with other knowledgebase after the name entity recognition, categorized and integrated with other topics **?**.

1.2 Structure of the report

This report is structured into 5 chapters, this is the first chapter which introduces the topic of the research and the questions that need to be answered.

Second chapter gives a background and history of social networking and collective intelligence in the area. It also describes the role of Semantic Web and its technologies in this area. It explains the evolution of the Semantic Web and Linked Data in the area of Social Networks briefly introducing some of the most important and widely used social semantic technologies and applications used nowadays.

Third chapter describes in more detail the topic of purposive social networks and community formation and the motivation of the research explaining why forming a quick and purposeful community is important and how linked data can help in achieving this goal.

Fourth chapter provides analysis of StackOverflow, a question and answer forum for programmers, a purposive social network. The questions, answers and user information is analyzed to describe the network ties and relationships. The individual role of users is also studied and the incentive model of the website is discussed that motivates user for huge quality contribution. Also, the data is analyzed and shown how link data can help in name and topic disambiguation, community formation, integration and search and discovery of better information and expert.

Finally, the last chapter concludes this report and describes the future work to be done.

[8] http://wikipedia-miner.cms.waikato.ac.nz/
[9] http://www.opencalais.com/

Chapter 2

Evolution of Collective Intelligence and Social Network in Web

In the early 1969 when Internet was invented by ARPANET, it was created as a robust network that allowed military computers to communicate quickly and freely. This network was soon improved when TCP protocol was developed and large data packets could be transmitted through long distance. With the advent of Ethernet, TCP/IP and SMTP protocols and individual personal computers more people were connected to each other, they used Emails to communicate and shared documents to collaborate with each other **?**.

In the early 1990, development of WWW and HTML at CERN and launch of the first commercial browser, Netscape, Internet had spread to the masses and as of 2011, there are more than 360 million Internet users in the world **?**. The Internet and World Wide Web has become ubiquitous in our everyday life, it has become a vital source of information, means of communication and a channel of self-expression.

The process of globalization has been changed by the WWW, initially only military and academic institutions has the technology to share documents and create mailing lists to communicate, now the web has revolutionized the communication process. People connect to their friends, families and colleagues using different social networking services. They are always in touch with each other using emails, instant messaging services, mobile devices and also using websites like Facebook and Twitter. They interact with complete strangers from all over the world who share same interest as them. The Web 2.0 has made it easier for people to share their personal information and interest with their friends and network, they use different services for different purpose. One person has more than one account to stay connected with the people they know and to share different type

of information. WWW has bridged the gap and brought people from different places, socio-economic status and cultures to share and create resources.

The social networking in the web moved forward to create community of like minded people and many tools were used for social collaboration and people contributed to create a collective knowledgebase. This large number of human contribution or crowdsourcing is used to solve many problems that are harder for computers to do and human computation is used to solve many natural language processing, speech recognition and image-processing tasks. The combination of Social Web and human computation provides many opportunities to solve difficult computational problem and create an emergent knowledge to solve problems.

In this section, the report is going to discuss the emergence of Social Web and how the Web has evolved to Web 2.0 that connects all the objects and peoples. In the recent years, the Semantic Web technologies are also incorporated with the web to create a new era of Web 3.0 and it is discussed how these technologies could extend the scope of current social web and social media.

2.1 The emergence of Social Web

The invention of World Wide Web, HTTP and web browsers allowed more people to interact with each other, share documents and create static web page easily, without having a detailed technical and specialized knowledge. Although Tim Berners-Lee had envisioned a read-write web, where the very first browser also worked as an editor, the initial WWW was a read-only web for majority of people. In the early days, the web was mostly a collection of webpages with a phone book like directory to look up individual websites that were connected using hyperlinks.

The passive attitude towards the web changed when the web browsers and search engines made the webpages easily accessible to users and the business world adopted the technology and started using it for e-business. More communication tools like IRC and MS net meeting were developed for organizations to communicate and collaborate. Organization started to have their personal intranet and portals as a collaborative tool and to integrate information on certain topics **?**.

As most of the business world and workplace moved to the digital world, a new phase in the software development emerged, a highly collaborative and alternative model of open source softwares. This was the first wave of crowdsourcing where many programmers and developers came together to build a fully working system and solved problems, Linux software is an important example of this working model.

The web was adopted by more organization, especially the news and entertainment industry and more and more information was put on the web as RSS. People moved

from a static personal web page to opinionated blogs and wikis and started the social transformation of the Internet from web of data to web of people. The next section describes this transformation in detail **?**.

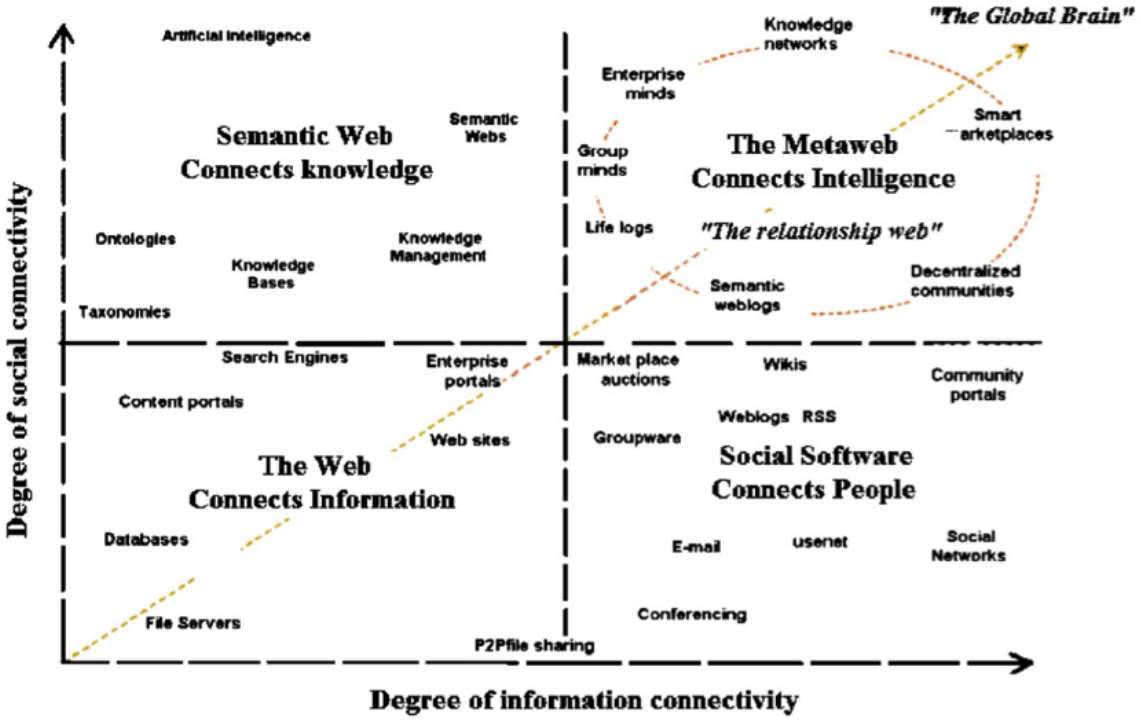


Figure 2.1: A taxonomy for collaboration alternatives **?**.

2.2 Web 2.0 and Social Media

Web 2.0 is the term coined by Tim OReilly for the new phase of web where not only the technology but also the business revolution in the computer industry is caused by using web as a platform. The main characteristics of the Web 2.0 websites are that it creates object-centred networks where people connect via objects of interest.

The first wave of socialization of Web began with the appearance of blogs and wikis, as it was easier to create content without the knowledge of HTML and web development. The easy to use "What you see is what you get" editor helped creation of large amount of contents and people could communicate and collaborate easily with each other, the original idea of Berners-Lee came into effect when the web was both read and write.

2.2.1 Blogosphere

Services like LiveJournal [1] and Blogger [2], both started in 1999, created a network of people with similar interest and purpose and create a collaborative knowledgebase for the community of like-minded people. The blogosphere soon recognized as a densely interconnected social network through which news, information, ideas and opinion travel rapidly formed by professional and corporate bloggers as well as hobbyists. As of 2011, NM Incite studies tracked over 181 million blogs growing from 36 million in 2006 and there are 1 million new posts everyday **?**.

Nowadays, there are blogs in every category and topic and it has turned into the means to share news and information instantly, like the Huffington Post[3]. These blogs are densely interconnected and are networked together based on region, language, topics and categories. Blogs have made it easier for user to create new content and easily share it with the masses. The simple editor does not require any technical knowledge and it is a large-scare platform for people to create network with similar interest and expertise.

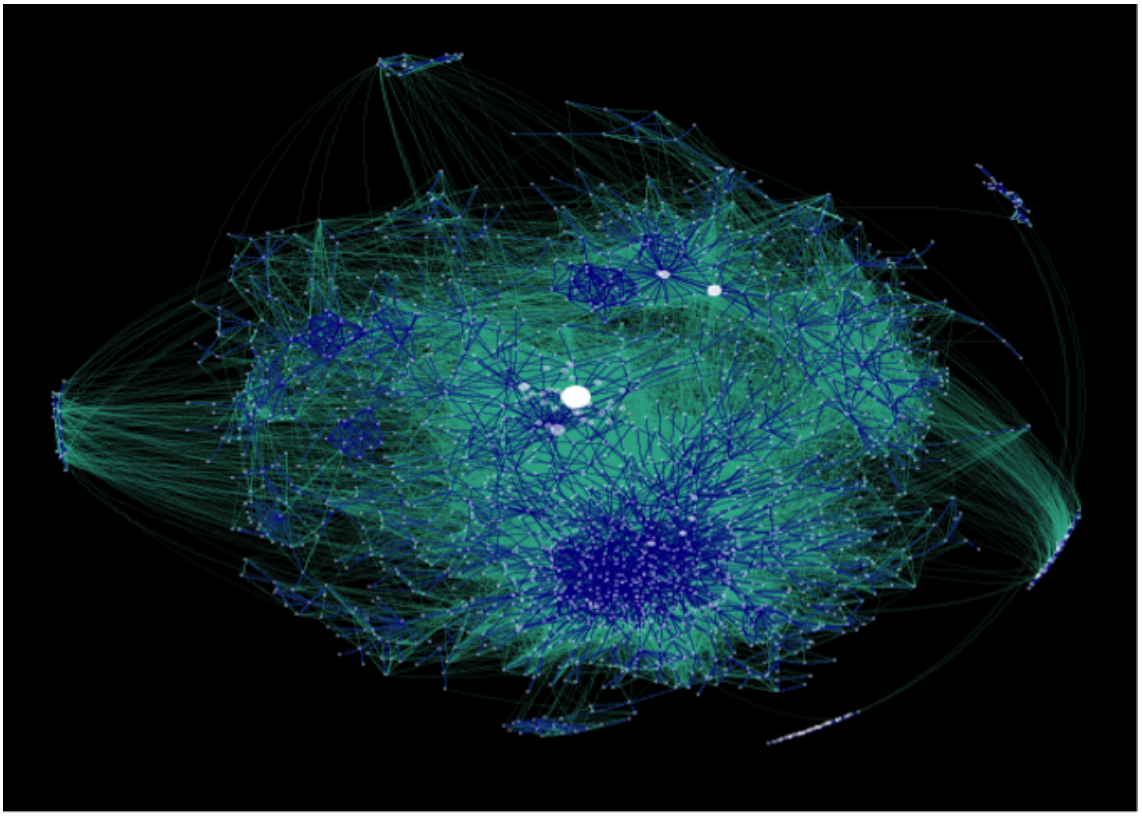


Figure 2.2: Blogosphere as social network **?**

[1] http://www.livejournal.com/
[2] www.blogger.com/
[3] http://www.huffingtonpost.com/

2.2.2 Social Networking Services

In 2003, one of the earlier social networking website Friendster [4] had attracted 5 million users that created profile and connected with friends and family, soon many other similar website emerged of which Facebook was the most successful with more than 900 million user as of March 2012. People create explicit connect with friends or colleagues (LinkedIn) to communicate and share contents with them.

People use these websites in every aspect of their life, they use it to share personal information with friends and family, use it to share their work and interest with colleagues and play games with other users. Now, business enterprise have also adopted these social networking website to connect to their consumers and advertise the products directly to the people who are interested in it **?**.

These websites now, not only connect people together through personal connection, it also connects people through shared objects. A good example of this is Facebook 'Like' button, this is used to connect people who like the same music or movies or share similar interest. This creates a vast community of people who share same interest and also share similar information. This giant global graph of people and object connected together is similar to the idea of Linked Data.

2.2.3 Microblogging

There are also microblogging services like Twitter and Tumblr [5] makes sharing information quicker and faster. These services create implicit relationship between people by follower-following method. The social network grows quickly as people do not need to approve any friendship and the content is distributed widely and rapidly by the resharing feature they have. Users can retweet the same tweet in Twitter and reblog the same posts in Tumblr multiple times creating a viral distribution of information.

There are more than 500 million twitter users as on April 2012 and there are 400 million tweets everyday. Analysis of the tweets has suggested that 40% of the tweets are pointless babbles and only 13% have some value or important information or news **?**. The rest of the tweets are spams, conversations or self-promotion and marketing. Twitter and Tumblr allow users to interact, create hash tags and share pictures, videos and other media.

[4] www.friendster.com/
[5] https://www.tumblr.com/

2.2.4 Peer to Peer network

The other method of collaborative information sharing is peer-to-peer network where peers are the computers system connected through Internet. Some software applications like Bit Torrent [6] are used to share any content between the users and large documents can easily be shared by reducing the load on one server. There are many controversy and legal issue with sharing copyright files like music, films and books but this social network allows millions of people to share large amount of data in distributed format.

2.2.5 Content Specific Social Networking services

After Facebook, a new trend emerged for content specific social networking websites like YouTube to watch videos, Flickr to share pictures, Delicious [7] to share bookmarks, LastFM [8] to listen to music, etc. In these websites people connect with other people with similar interest and passion. They form groups and communities and use user participation rate the content for quality control and user behaviour recommendation.

These are open networks where people can upload their content and it is visible to the whole world. It gives a platform to broadcast creative content to the large audience. These networks are completely depended on user contribution to sustain the community.

2.3 Collective Intelligence and Crowdsourcing

The World Wide Web provided a platform for people to share, discuss ideas and information, and collaborate with each other. The Web 2.0 main trend was using the collective intelligence of people to create a knowledgebase, get user preference and do recommendations.

2.3.1 Wiki

Web 2.0 provided easy to use tools for collaborative authoring and ownership of information. A wiki allows users to add, remove, edit or modify any content on a webpage and keep tracks of different versions changes. Wikipedia [9], a collaborative encyclopaedia, is an excellent example where people with similar interest and expertise came together to create this vast knowledgebase. Crowdsourcing is used for quality control and to stop vandalism spamming. These collaborative networks strive because they enforce a strong sense of community with people through collaboration and creation of content where

[6] www.bittorrent.com/
[7] http://delicious.com/
[8] http://www.last.fm/
[9] http://www.wikipedia.org/

individual contribution was counted and it became an efficient knowledge management tool.

2.3.2 Folksonomy

The other feature of Web 2.0 was the ability to add tags to any content and user generated taxonomy is utilized to categorize the pictures, videos, music and every kind of data. The tags make sit easier to classify and categorize information and also to search and discover similar content on any topic. Collaborative tagging and bookmarking is used to add metadata and categories to pictures in Flickr and links in Delicious and it makes it easier for people to search and discover information. Using multiple tags also connects categories and geo-tagging helps in identifying any object based on location.

2.3.3 Open Source Software

Linux open source software project gave an alternative to the traditional software development and the programmers formed a community to build the software and tools. The system provided the developers with a platform to collaborate and create where individual authorship is recognized but they dont get exclusive intellectual rights. Creative common license is used and it allows the creator to easily mark their creative work and share it with the world.

This type of large-scale software development requires lots of co-ordination and communication with the community and a project is divided into various phases of development. There are many tools available, like version control, bug tracking, task management and testing tools, for people to collaborate and communicate. Many software products like Firefox, Android, etc. are successfully developed and used by general mass.

2.3.4 Forums and messaging boards

Forums and boards gave a place for people to meet anonymously and have a discussion, it was the beginning of virtual community where face to face communication was replaced by messaging boards, chat rooms, question and answer forums. These online communities were formed based on interest and common objective and people from all around the world could come together and communicate on a particular topic of interest **?**.

2.3.5 User Recommendation system

The power of collective behaviour and knowledge was utilized by the websites like Amazon [10] and Pandora [11] for recommendations of books and music respectively. Users in these websites also rate and review the things and crowdsourcing is used for predicting users needs and contents can be recommended accordingly.

2.3.6 Human Computation

Humans are good at solving the AI complete problems like natural language processing, image analysis that computers still have problem solving. Luis von Ahn came up with various systems to utilize human computation and use crowdsourcing to solve these problems. The ESP games **?** are used to annotate images on the web **?**, reCaptcha **?** is used to digitize millions of book and Duolingo [12] is used to translate the web into various languages.

These systems are used by users to play games, identify humans from machines to stop spams, and to learn foreign language and they utilize the tasks easily done by human to solve computational problems. The same task done by many people give a common consensus to knowledge and prevent errors **?**.

2.4 Semantic Web and Social Networking

The idea of a Semantic Web was coined by sir Tim Berners-Lee in his Scientific American article **?** that described it as an extension of the current web, where all data is structured and has a meaning accessible by both machines and people. Linked Data is a set of best practices for publishing reusable structured contents using the existent Web as a sustaining framework. In Linked Data every resource is represented by a URI and HTTP URIs are used in order to retrieve documents that describe those resources. Documents are described with RDF where links to other resources are then used.

There are many Semantic Web technologies and application available in the area of social network and social media. These technologies help with data representation, data portability and cross platform interoperability. Using Linked Data principles a decentralized system can be created that helps in social network integration. Every object and entity is identified using a URIs and it is deferenceable by using HTTP **?**. Also, useful metadata about the entities can be added in structured format using RDF and it could be linked with other related data to improve discovery. Some of the social semantic technologies and application is discussed below.

[10] http://www.amazon.com/
[11] www.pandora.com/
[12] http://duolingo.com/

2.4.1 Social Semantic Technology

As previously discussed there is already a large amount of linked data available on Web called Open Linked Data cloud and there are many useful ontologies to represent this data, including social data. The evolution and formation of different and most widely used ontologies to describe people, communities, and their data are discussed below.

2.4.1.1 FOAF

FOAF ontology is used to describe people and the relationship between them. Each person has a unique identifier and a specific vocabulary is used to create personal profile of the users and describe their social network. It can be easily integrated with other Semantic Web ontologies and a person can describe one or more of their online social network **?**.

Many websites like LiveJournal, FriendFeed, identi.ca uses FOAF to describe their users. Many other websites has plug-in or external application that create their FOAF profile like FOAF generator, WordPress plug-in etc. It helps to solve distributed identity problem when one user has different account in different website, using FOAF, a user can combine all the information from various sources into one file and interlink his different identity and social network. This also helps in creating a complete user profile and cross-site content recommendation **?**.

Privacy and digital identity protection issues can also be resolved in FOAF by using SSL protocol that provides distributed authentication system to different user and create different policy for private and public information **?**.

2.4.1.2 SIOC

SIOC ontology is used to describe online communities activities and interlink them together. It can be easily integrated with FOAF profile of a user, so the combined document contains users information, their social network and all the data user created with the comments and metadata provided by others **?**.

SIOC can be used for blogs, forums, discussion board, mailing lists, etc. Many websites uses SIOC as well as ma enterprise and E-government uses SIOC to freely available their data.

SIOC is the next step in integrating different social networks together, it describes the content of the website with a structured format so the information is easier to access, search, and discover. The posts can be browsed in many innovative ways by using specialized queries. It creates connection between people from different networks and

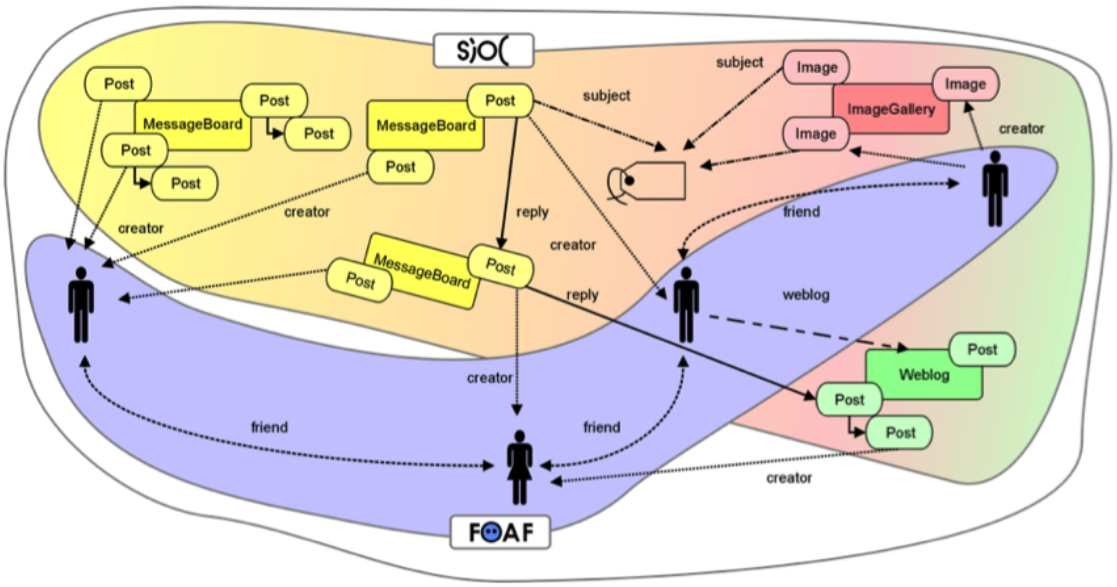


Figure 2.3: Creating a FOAF profile and SIOC file of a user.

people with different account can use it with FOAF to prevent identity problems or co-referencing. It makes it easier for people to connect the data in decentralized system ad enable data integration **?**.

The above figure describes how FOAF and SIOC together can help interlinking and integrating different communities and different account of the same person together using Semantic Web technology for easy reusability.

2.4.1.3 OPO

The OPO ontology is used to describe the online presence of a user who is using instant messaging service. It stores the state of user, if they are online or offline and who they are communication with and what they are communication.

2.4.2 Social Semantic Applications

Social web and semantic web has come a long way in past ten years. There are many social semantic application available present on the web that solves many underlying problem present in the SNS. They are described below **?**.

2.4.2.1 Semantic Tagging

People in Web 2.0 use tags to classify, categorize or group their content using their own vocabulary. This has enabled users to generate more data but this also ambiguous and

imprecise. People use different terms in the same context meaning the same concept and same terms in different contexts for different meanings.

The main issues with tags are the tag ambiguity problem where it is not clear in what context a tag is used. For example if a post tagged with Apple is describing a fruit or a computer brand. Thats why is so important to reconcile lexical tags to URIs that unambiguously identify the meaning of a term **?**. The same tag actually differs when people use different spelling or space or hyphen between words. And there is also a lack of organization and hierarchy between tags **?**. It is hard to comprehend if a tag marked Apple is a subclass of a tag marked Fruit.

Most of these problems can be solved by the MOAT ontology that helps to describe the meaning of a tag semantically and connect it with the object and related tags. Also the co-occurrence of tag can be used to group similar tags together. It is user defined interlinking and the contents are linked to the Linked Data Cloud. It uses the collaborative approach to share the meaning of tag in a community.

The Tag Ontology helps to solve the ontological and hierarchical problem that exists between groupings of tags. The SCOT (Social Semantic Cloud of Tags) ontology creates a model to describe a tag cloud and it makes the tag cloud portable so it can be exported from one service to another without losing the data.

Many research is carried out in the area of extracting ontologies from tags, FolksOntology is one of those project that extracts relationships between tags. FLOR is another project that automatically identifies the meaning of a tag. These ontologies can be combined with other ontologies to create a complete model for semantic tagging in the area of social network **?**.

2.4.2.2 Semantic blogging and microblogging

Blogosphere is a huge part of the web where people are generating data in exponential rate, but most of this data is not structured or categorized, it is in plain text **?**. With the advent of Twitter, microblogging is the new trend and this also lacks context and structure.

Semantic web technologies can be used to semantically annotate useful content and keywords of the blog posts and micro blogs to add meaning and structure to it. Drupal [13] website provides an easy to use platform to add annotation to the blog posts and it embeds RDFa into them. Twitter Annotation service takes the tweet metadata of location, time, etc. and annotates them to provide semantic metadata.

Another important project that helps in semantic blogging is the Open Graph Project and Facebook Open Graph that helps users to express preferences increasing in this way

[13] http://drupal.org/

the quality of the information provided in the blogs or indeed in any website **?**. People can add reviews and ratings to a movie or use the Facebook Like button to show their preference and the Open Graph project connects the objects and resources with people, their friends and people with similar interest to provide better recommendation. These projects follow an Open Graph Protocol and create the data in RDF and the data can be queried to provide useful information.

The main use of these technologies is that it is easier for users to publish their data in structured format and they can publish it cross-site because it is portable. The developers can quickly create mash-ups to provide better data visualizations by integrating data from different sites because search and discover is easy in structured data. This data can also be mapped to peoples FOAF profile and SIOC data files. If the dataset has a SPARQL endpoint, querying and retrieval of data can be done **?**.

2.4.2.3 Semantic Wiki

Wiki is an excellent example of collaborative creation and edition for emergent knowledge. There is a large number of people editing Wiki pages and helping in maintaining the quality of the information and preventing spam and irrelevant information **?**. The structured information from the Wikipedia info-boxes has already been converted into linked data within the DBpedia [14] project, and this data set is already one of the largest and most linked dataset in the Open Linked Data cloud.

DBpedia is very useful because it uses Wikipedia data and it is easy for anyone to read or write a Wikipedia article, and it also solves the problem machines cannot by using Crowdsourcing. But DBpedia only uses part of the data because Wikipedia lacks proper structure and agreed semantics of the data and its categories. This problem can be solved by semantic wiki where all the data is structured and properly categorized. This provides better search and discovery of information like in OntoWiki **?**.

[14] http://dbpedia.org/

Chapter 3

Purposive Social Network and Community Formation

A community is created when people collectively create and share information providing a common source of knowledge and WWW provides an easy platform for people to come together and create a knowledgebase by crowdsourcing. Wikipedia is a good example where thousands of people come together and create and edit a shared knowledgebase. The most efficient online communities depend on user participations, easy moderation, interest in topic and a common tie between communities members.

In this report a question and answer forum for the programmer, StackOverflow, is studied where people with a common interest come together and solves problems, and create a self-sustaining and self-regulating purposive social network.

3.1 What is Purposive Social Network (Different types of Social Network)

There are many reasons to join and form communities; human are social creature and forming a social tie is a natural instinct. In the online world, where geographical boundaries are dissolved and people from any part of the world can connect together, and also have a certain degree of anonymity and privacy, people can reach to others with similar background and interest.

A network with a purpose, as name suggests, is created when people come together with a common objective to get information, build knowledgebase, solve problems or achieve some common goal.

A purposive social network can be typically categorized into following types. It should be noted that a community can be of one or many categorizes mentioned below **?**.

3.1.1 Information based community

They are communities that share some particular type of information or resources, communities where people come to share their particular knowledge and gather more information or news on any matter by IRC channel, forums or just mailing list is one of the examples. This kind of community generally has many new people joining to get some information but the rate of returning back is low **?**.

3.1.2 Interest based community

They are communities where users come together to share their hobbies or interests and form a strong connection with each other. The number of people in the network is generally stable and they actively participate in any discussion. Online role-playing and gaming community is one example of this type of community **?**.

3.1.3 Location based community

They are communities where people come together for a geographical reason and where proximity is important. These types of communities form because of social issues concerning a region: neighbourhood watch, Yoga and jogging groups or support groups are examples of this type of community **?**.

3.1.4 Expert based community

They are formed mostly by experts in their field in order to discuss a common matter or to solve a common problem. Mostly academics or experts in a field form this kind of community.

3.2 Why Purposive Social Network is formed

People are spending more and more time online, they connect with their friends and families, do their work and shopping online and use other peoples recommendation and experience to solve their own problem. The social network analysis proves the six-degree of separation theory, that each of us is connected to any random person in the world through the right six people we know. But in the world of blogs, forums and messaging board this barrier is broken. One does not have to know anyone personally, professionally or at all to interact, comment or reply to one another **?**. The main motive for people to join such communities is discussed below.

3.2.1 Information exchange and self interest

With people spending more and more of their time on the web, this became a place where they can discuss ideas, problems, issues and, being a social creature, they would like to do it with like-minded people with similar interest and issues **?**.

Online communities are the best place to share and exchange information because it is available to everyone, anywhere in the world, people come together and join a forum or a discussion board to share knowledge about their hobbies, health or politics. People share different kind of information, from restaurants at Yell.com to websites in LiveJournal, two completely different websites for different purposes take advantage of user needs.

In the area of E-Government, the government (e.g. USA has data.gov [1], UK has data.gov.uk [2]) is publishing statistical and other Public Sector Information (or Open Government Data, OGD). People with different objectives come together and use this open government data and collaborate to prevent crime using crime statistic of their geographic area or create support groups based on the health and morbidity statistic of their nearby hospitals, websites like FixMyTransport [3], PatientsLikeMe [4], WhereDoesMyMoneyGo [5] are example where people come together to utilise public data.

3.2.2 Symbiotic relation and social exchange

People also communicate with their friends and family members using online communities. They would also like to come together to form a neighbourhood watch program or start a political protest. The online communities, with the use of microblogging or text service, would like to be up-to-date with current information, share their resources. People often go to forums to exchange recipes or to swap expensive tools they need for a short time. It is a mutual symbiotic relationship that helps them. BookMooch [6] and BarterPalace [7] web services are example of this mutual beneficiary relationship between people in online community **?**.

Also in SNS, people often measure their importance by the number of friends they have or by the number of famous or important people are parts of their network. People often also use this platform for advertising their product and skills and the more people they have access to, the more benefit they get from the communities and their social network, Twitter and Facebook are good example for this.

[1] http://www.data.gov/
[2] http://www.data.gov.uk/
[3] www.fixmytransport.com/
[4] www.patientslikeme.com/
[5] wheredoesmymoneygo.org/
[6] bookmooch.com/
[7] www.barterpalace.com/

3.2.3 Recommendation System

People often leave reviews and recommendations of products in their blogs or forums on an E-Business website like Amazon. This is also true for other recommendation websites (e.g. for music, movies or restaurants). People come together to rate and discuss the quality of places or products and also to get recommendation from each other.

The Linked Data and Semantic Web technology is very useful in this case as it provides a structure to publish this information so that all the ratings, reviews and recommendations can be done cross platform and cross-site.

3.2.4 Expert Finder

People often go online to find a solution to their problems; this is especially true of the software developing community where people go to ask for help and solution to their bugs in their program

People go to specific forums for programmers, like StackOverflow, that have an extensive community of developers behind it for helping people with their coding, system analysis, design, etc. Online communities are the best place to find experts but it is also hard to find the right expert and sometimes questions gets buried below the amount of data that is created everyday **?**.

3.2.5 Social recognition and personal satisfaction

As previously mentioned, people go to online communities to find experts or recommendations, hence the person with the most knowledge and up-to-date information gets social recognition and is considered therefore an expert **?**.

People also get pride and satisfaction from helping others, showing and sharing their expertise with others and their peers. Hence, online communities are a good place for people to come together to achieve a goal because its easy to share and accessible to all the people in global stage.

3.3 Characteristic of Purposive Social Network

The main characteristics of a purposive community are same as any other community is the social network structure but some attributes make them different . These special characteristics of purposive social network are as follows **?**:

3.3.1 Small number of people

This type of community is created by people with similar objective and purpose and it does not have a large number of people but a small amount of people coming together to share knowledge and solve a problem.

3.3.2 Focused interest

Small groups of people with shared interest come together to solve a very specific problem and create a focused knowledge.

3.3.3 Direct communication

Each individual member in the community share information and communicate directly with other members, they communicate directly with one another.

3.3.4 Active participation

Since purposive communities have focused goals and interest and small number is people, each member actively participates to solve the common problem.

3.3.5 Short lifespan

The purposive micro community has a focused interest and purpose and it exist for a small period of time and once the problem is solved, the community dissipates.

3.3.6 Strong incentive

Since the micro community exist for a small amount of time and only interested people participate focusing on solving a particular problem, strong incentive exists to motivate people to join and participate.

3.4 Crowdsourcing in Purposive Social Network

The main feature of a purposive social network is building a network with the people with similar interest and objective and motivating the people to contribute and collaborate together to create a problem solving system. Crowdsourcing systems are a good example of a purposive social network where people are contributing and creating a

knowledgebase and utilizing the power of network to achieve common goal. A crowdsourcing system is depended on the user contribution and self-sustaining systems are difficult to model. An efficient crowdsourcing social network requires motivated user who contribute and finding and retaining users and motivating them with enough incentive to contribute is a major part **?**. Crowdsourcing is also used to manage and moderate the community and do quality control.

3.4.1 Recruiting and retaining users

A crowdsourced content specific purposive social network utilizes user participation and expertise to create a community. The question and answer websites, collective knowledge generating website depends on large amount of user participation to create an active community. There are several methods to get users to contribute like paying the users or making it a requirement for users to contribute, as in reCAPTCHA where user have to digitize the image to finish the task. The popular option is asking for volunteers.

When the social network is specific for a special kind of group, the system can be made easy to use, free and open so people can contribute. Websites like Wikipedia, StackOverflow, YouTube are content specific and are free, people volunteer and contribute to these websites and create a vibrant community with like-minded people. The downside of this is that it is hard to predict how many people will actually participate and contribute in the whole process and this type of system requires a good incentive model that keeps user motivated enough to contribute and maintain the quality of knowledge **?**.

3.4.2 Incentive Model

Designing an incentive model for a large-scale crowdsourcing system is complex. The system should make it easier for people to contribute but also keep track of the quality of content created. The game theory approach is maintained in such scenario with proportional mechanism where the good content is rewarded to generate more active participation and the asymmetrical cost of participation is avoided where a low cost contributor is deterred because of the contribution made by higher cost contributor. A trade-off is done where it is made for people to participate and appropriately rewarded for good or bad behaviour and quality of the content.

The game industry uses the instant gratification model to incentivize and motivate user for more participation and contribution. They make the whole process of creating content an enjoyable experience as a game playing scenario, in the case of ESP **?** and the user gets motivated to perform more task.

Other question and answering system works on the users desire to be recognized and provides different methods to measure and show a reputation or expertise in a particular

field. When users establishes reputation and is recognized as an expert in the area, the user generates more quality content and is motivated to participate in the community **?**.

To sustain the constant quality of the content, there should be positive reinforcement for good quality contribution. Users who generate quality questions and promptly answers them rewarded for their contribution with badges or points are motivated and are active in the community. Similarly, when people are spamming or creating poor quality content or are asking repetitive question should be given negative points and their contents should be eliminated for the lack of quality with the entry restrictions. The maintaining of high quality knowledgebase brings back the users and they are more careful with the quality of their submissions **?**.

The other way to encourage user participation is to give the ownership of the content to its creator, this entitles the user and they become responsible with the maintenance and quality of the product. Also, creating a competitive environment where more contribution makes the user the top contributor of the category, this ensures higher rate of returning and contribution from the participants **?**.

An approval-voting scoring rule and a proportional-share scoring rule can enable the most efficient equilibrium with complements information, under certain conditions, by providing incentives for early responders as well as the user who submits the final answer **?**.

3.4.3 Quality Control

Many websites with large amount of user-generated content depend on the joint community action to rank the content according to the quality by collective voting. This controls the quality of information displayed on the web page, the higher quality content is displayed more prominently and the lower quality content is surpassed and spams are removed. Using thumbs-up or thumbs-down style ratings by the users, questions on StackOverflow, reviews on Amazon, and posts on Reddit [8] are shown **?**.

These websites display higher quality contributions more prominently by placing them near the top of the page and pushing lower quality ones to the bottom. Since content displayed near the top of the page is more likely to be viewed by a user, ranking good content higher leads to a better user experience. Another benefit is that it also provides an incentive to produce high quality content that might appeal to a contributors desire for attention **?**.

Rank order mechanism is used to influence the quality of the content and research has shown that the game theory model is used to motivate the attention driven users and

[8] http://www.reddit.com/

generate higher quality content and create a better environment for information distribution and sharing. The users that generate higher quality are featured prominently on the page and the proportional mechanism distributes the attention in proportion to the positive votes received. This creates a game theory equilibrium that facilitates higher quality posts and accordingly rewards the users creating a large incentive to participate in voting and contributing **?**.

A text analysis of the user content also determines the quality of the posts. A post with punctuation, grammar and typos can be easily analyzed to create an estimate of the knowledge and expertise of the contributor. Also, the syntactic and semantic complexity if the texts give an approximation of the overall knowledge of the user and their proficiency with the topic **?**.

3.4.4 Search and Discovery of Quality Content

The amount of content generated in user generated knowledge system is large and it is difficult to find the high quality content in a large-scale community. There are many algorithm and models used to filter the best content from the masses and they are discusses below.

Link analysis and link based ranking is used in blogs and other social networking systems to form an estimate of the quality of the information. PageRank **?** and HITS (Klienberg, 1999) are the prominent ranking algorithms used in this method. The network graph formed by the analysis shows the flow of information and relationship between the people. The mutual reinforcement facilitates good blogs connecting to other good blogs and good questions getting good answers. A learning framework that follows the factoid QA benchmark is efficient in getting the facts and trivia in a large-scale system **?**. These can be used to filter high quality materials from the large collection of information available.

User voting and tagging is another use of crowdsourcing to search and discover appropriate information. Users vote the best questions and answers to the top of the web page and make it easier for people to discover the information. People also tag the content with appropriate keywords and categorize information that makes it easier to browse related content.

User rating and recommendation is also used to search and discover high quality content as in Amazon where books and other items are recommended based on user who bought an item also bought other items. Also, in IMDB [9] and Rotten Tomatoes [10], movies recommended are given based on users ratings and reviews. These website tracks users popular behaviour and utilize the information to give recommendation to the rest of the population.

[9] www.imdb.com/
[10] www.rottentomatoes.com/

3.5 Role of Semantic Web in Purposive Social Network

3.5.1 Linked Data to support purposive communities

Linked Data and Semantic Web as previously mentioned helps in formatting and structuring the data. It eases the sharing and the portability of users data and the forming of an open decentralized social network across platforms and websites. The benefits of Linked Data in forming quick micro-communities are discussed below.

3.5.2 Multidimensional Network

All the data in the linked data cloud are linked with each other based on semantic equivalence and reuse of information by means of URIs. It forms a multidimensional network, a network formed of people with their friendships and other relationships, and of relationships of data by means of properties or other attributes. This creates an overlapping network that can be queried across dimensions and time to find important information.

3.5.3 Structured Data

In Linked Data, every resource is represented by a dereferenceable URI, which can be resolved in a document described in RDF format with metadata and ontology to describe it properties and provide definition. It provides the context of data and people and their interest. It provides a large knowledge base that is useful in finding correct information and people while forming a community.

3.5.4 Linking People to People and People to Data

In FOAF profiles, people are linked with their friends and colleagues, SIOC profiles link people to their data and users data sets are linked with each other. The links people to people, people to data and data to data, when queried, can provide useful relations and information.

3.5.5 Smart Query

The linked data sets can be queried using SPARQL **?** endpoints, or browsed using a semantic enabled browser. Querying those data sets it is easy to create many useful applications, mash-ups and discover relationships and patterns. Even if the resource or clusters of resources are decentralized, SPARQL can be used to measure the strength

of relationship in multiple dimensions (like social network, expert field) over time, with the growth of network and data, many interesting applications and communities can be generated.

3.5.6 Social Network Analysis

With the integration of multiple social networks and the elimination of identity fragmentation, social networks and their data can be analyzed and visualized in many ways. Experiments can be performed to learn and understand how networks come to be the way they are, or how the data flows within the network. It can also be used to measure the strength of a relationship or the degree of recognition of experts in networks and much more.

3.5.7 Integrated Knowledgebase

Semantic Web technologies can be used to describe different community and network and the data can later be integrated together to form a large knowledgebase.

Each system would be responsible for their won database but it will be open and interlinked with other datasets and can be queried. Relationships can be created between the datasets from different sources and it could be integrated like the Linked Data Cloud.

Chapter 4

Analysis of Purposive Social Network

A purposive social network is a broad and general concept that can be applied to many different social networks. At its core, it is a network with a specific goal and purpose where people come together to solve a particular problem.

For this research StackOverflow website is analyzed where programmers and developers ask questions and experts in the field answer them and solve the problem. In this analysis, the user asking and answering questions, voting the responses and commenting are the main entities of the social network. Their network ties are measured by their communication and interaction between them. The amount of their contribution is measured by the crowdsourcing where people give up votes or down votes to their posts and the badges they receive for their contribution.

4.1 Network Linkage and Social Ties

StackOverflow is not a typical social networking website per se as users cannot create an explicit friendship or follow other peoples work, they cannot send private messages or form groups.

The social ties are form implicit by user interaction with each other through asking questions and answering them, voting on the posts and commenting on them. The communication network is studied to see the social ties of the individuals **?**.

As of June 2012, there are over one million registered users in StackOverflow and more than 3.2 millions questions asked by users. The questions are categorized using tags and individual users can subscribe to tags to receive daily email of all the question asked in the tag. There are more than 30 thousand tags associated with various questions and

Post Type	Number
Questions	3279233
Answers	6578079
Registered Users	1225580
Tags	30408
Unanswered Questions	780535
Badges	3454994
Votes	26184363
Comments	12526162

Table 4.1: StackOverflow at glance as of June 2012

answers. Users have casted more than 26 million votes to mark the good questions and answers.

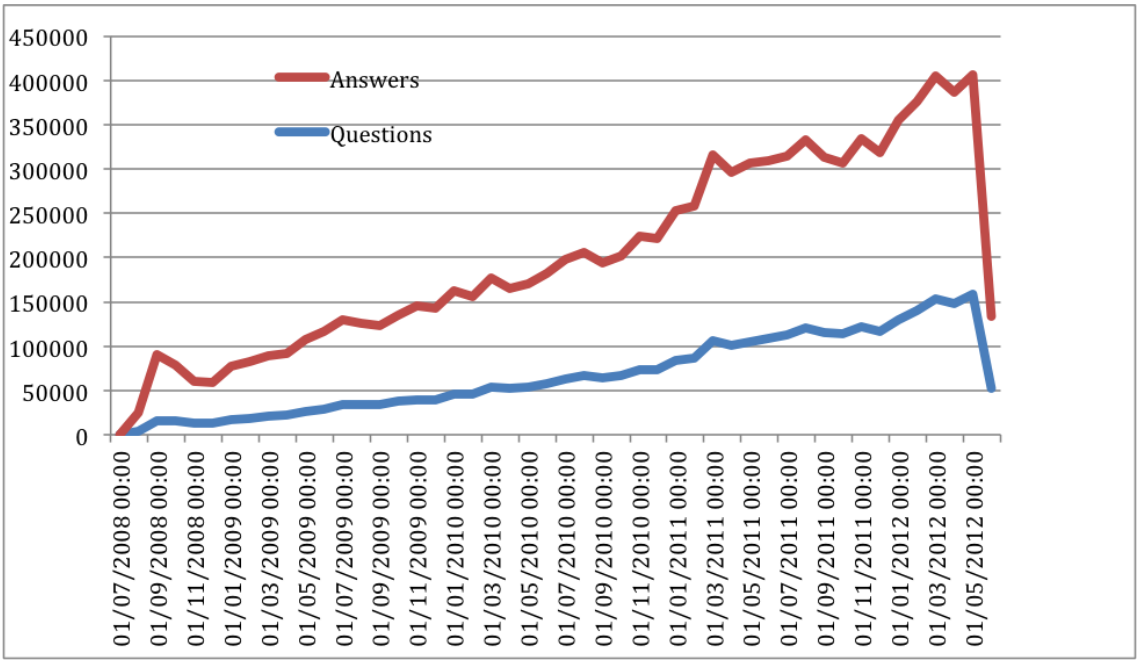


Figure 4.1: Questions and answers posted per month on StackOverflow

The Figure 4.1 shows the number of questions asked and answered by users each month. In the year 2012, each questions have on average 1.645 number of answers.

This community is made of programmers and their motivation is to solve the problem they had encountered and to provide answers to gain reputations. Thousands of questions and answers are posted everyday. The analysis of posts shows that the programmers prefer to ask the questions and answers on weekdays.

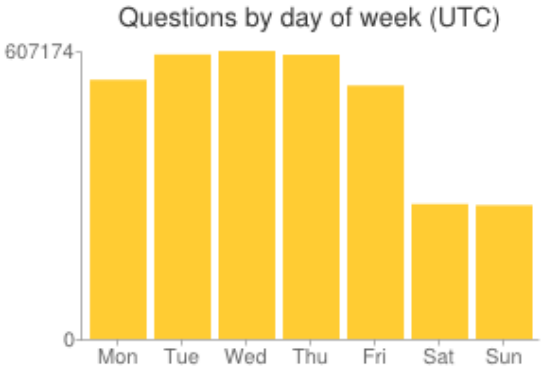


Figure 4.2: Questions posted by the day of week

Despite high user feedback and participation, 23.79% of questions are not answered or the answers do not receive any up votes. On average a question receives 2.006 answers and .12% of questions receives more than 15 answers.

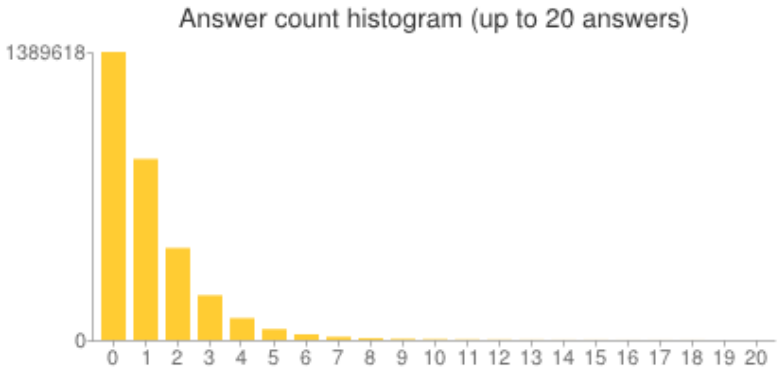


Figure 4.3: Answer count to the questions

The questions are answers are provided with tags to categorize and arrange for easy search and discovery. The entire website is categorized using the tags and the list of most popular tags are shown in the table. The figure shows the weekly use of the popular tags. The number of questions asked for each tags also provides an insight on the popular language used by developers at the time.

Tags	**Number of instance**
C#	370074
JAVA	315488
PHP	293755
JavaScript	278592
Android	244791

Table 4.2: Five most popular tags and its instances

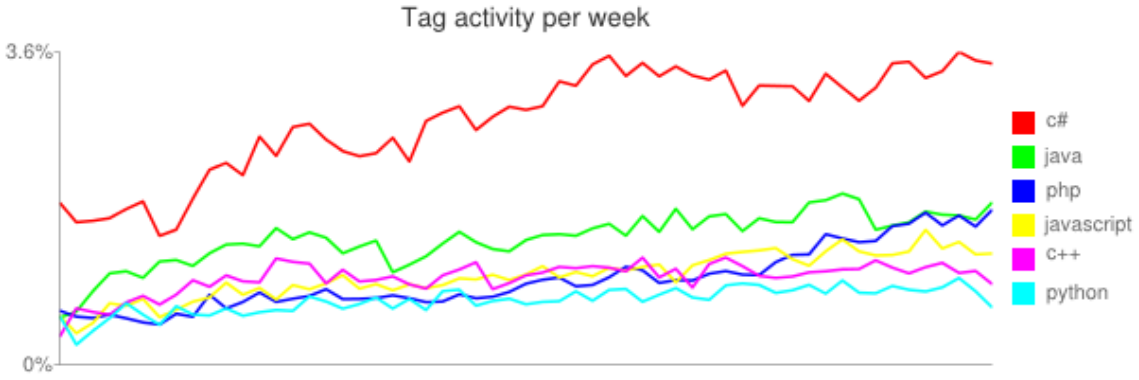


Figure 4.4: Tag trends per week of most popular tags

The analysis of questions shows that each question has between one to five tags associated with it. Most questions (70.30%) have 2 to 4 tags associated with it. The relationship between the tags shows the overlapping of networks and how it is tied with one another.

? provided an interactive graph in his website to show the relationships between the most popular tags and how closely they are related to each other. In the following graph each segment size is directly proportional to the number of instance it is used and the connection between the tags indicate the times they have been used together in a question. The thickness of the connection shows the strength of the relations. The segment is colour coded by the frequency of connections, red segments are strongly connected and blue segments are weakly connected.

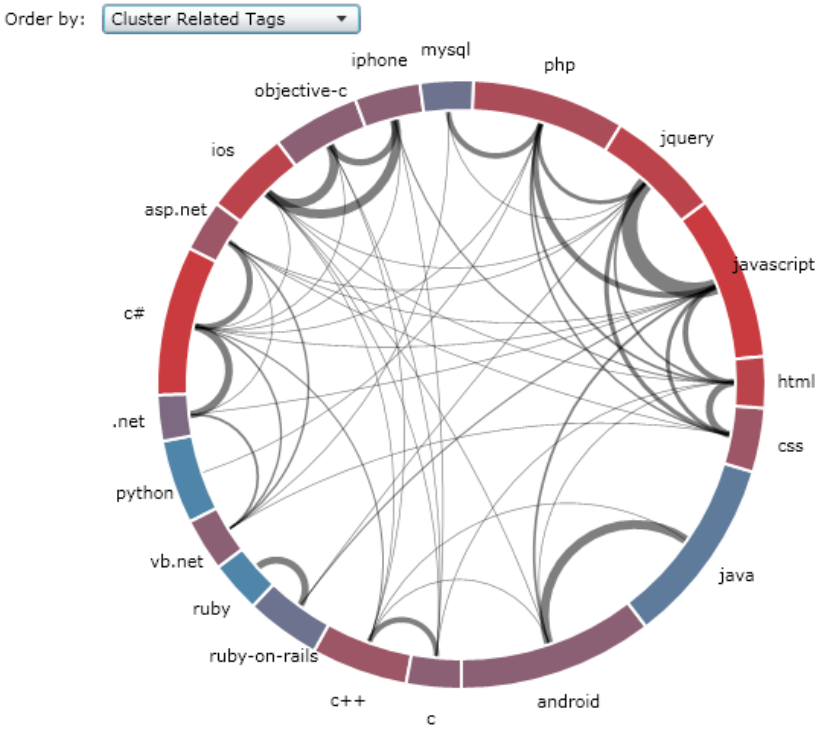


Figure 4.5: Related tags clustered together

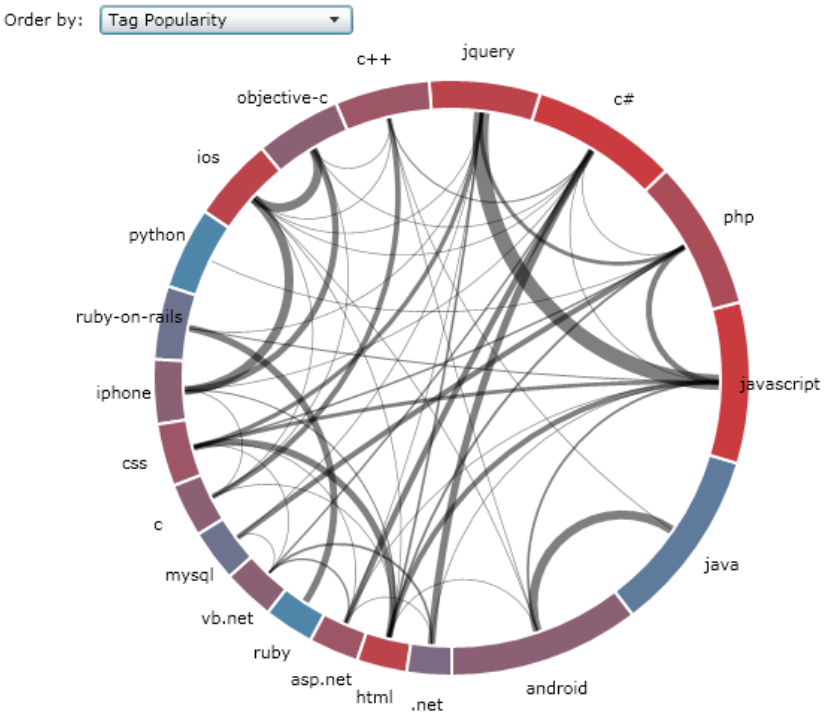


Figure 4.6: Popular tags clustered together

The clustering of the tags shows the relationship between the tags and technologies. The two popular tags JAVA and Android are closely related to each other but are scarcely joined with other tags. The strongest relationship is between jQuery and JavaScript because the overlapping framework of the two programming languages. C, C++ and C# are also a closely related groups as well as iOS, Objective-C and iPhone. However, sometimes Objective-C is also tagged with C, C++ and C#, if by mistake or deliberately can be argued. There is a large cluster of connected web development languages, CSS, HTML, JavaScript and jQuery, indicating the close knit use of these technologies in development of website and web applications. The interesting thing is the relationship between the scripting langue PHP and Python, they are popular tags but are sparsely connected with other tags and are weakly linked with database related tags.

4.2 Role of Individual Actors

Users who contribute to the website are the main actors of this purposive social network. There are more than 1.2 million registered users in StackOverflow and they ask the questions, answer it, vote it and moderate the community. The users are not directly linked to each other to create relationships; in this network the relationship is formed by their interaction and their contribution. The user behaviour, their motivation to use the website and incentive to contribute is described below.

Despite the high content generation by the users, 56.02% do not interact or contribute to the website, they have 1 reputation point that they receive while joining the website.

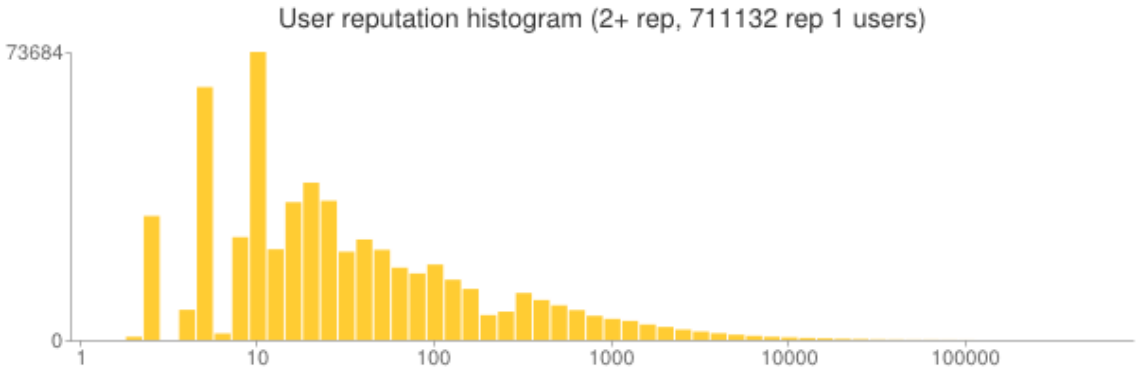


Figure 4.7: User reputation histogram

Reputation	Number of users
1	669554
2-10	126235
11-100	295389
101-1000	3161130
1001-10000	170993
10001-20000	1437
20001-100000	895
100001-200000	49
less than 200000	11

Table 4.3: Number of users with reputations

As Table 4.3 shows, there are 669554 users with 1 reputation point and one user with 452951 reputation points. The distribution of the users reputation shows that more than half of the users are lurkers and the elite users with the most reputation points are the editors and moderators of the community and are considered the expert in their field.

The reputation of the user has a direct correlation with the trust in the community. StackOverflow has designed an excellent reward program to motivate and incentivize the users to contribute and gain more reputations and badges.

Currently, there are 77 different types of badges given to the user based on their contribution. There are badges given to the user who asks questions with 1 reputation point (Student), to the user who edits the answers to make posts better (Editor) and to even an active user for a year (Yearling). This type of virtual acknowledgement of efforts encourage the user to participate and contribute to the website.

The other method that encourages the users to participate is the promptness of the response. The asker prefers to receive information sooner rather than later, and will stop the process when satisfied with the cumulative value of the posted information. The analysis of the posts shows that half of the questions get an answer within an hour of the posting and within a day the questions receives an accepted answer. When the answers are delayed, the questioners look for alternative websites to get a response.

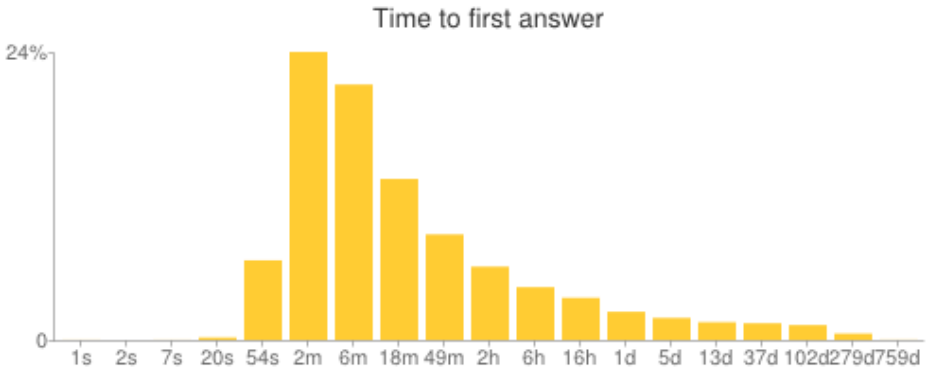


Figure 4.8: Time to receive the first answer

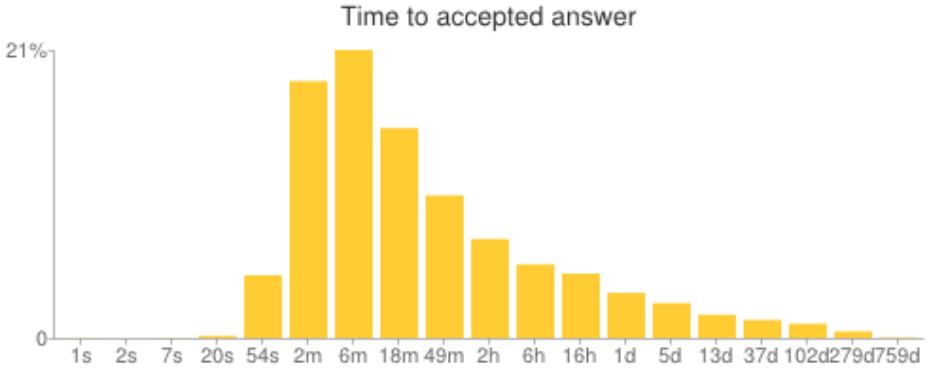


Figure 4.9: Time to get the accepted answer

4.3 Incentive Design and Quality Control

The StackOverflow website uses a game theoretic model to encourage user participation and activity. Participation is encouraged through an elaborate point system and users also receive badges for participation. Also, the top contributor and user with highest reputation are featured on the question page, giving the user more visibility and acknowledgement of the users expertise. This encourages participants to accumulate more points and contribute to get recognition.

When an answer is votes up, the user gains 10 reputations and 5 points when the question is voted up. When an answer is accepted the user receives 15 points and there is also negative point system, a user looses 2 reputation point when a question or an answer is

voted down. This keeps the spamming in check and repeated questions and answers are avoided.

The system also encourages users to participate as the higher reputation points gain more privileges. When a user has 15 reputation points, only then they can up vote and 50 points allows users to comment. To stop harassment and spam, user requires 125 reputation points to vote down and it costs the user 1 reputation point. The incentive model is thorough and higher reputation points open more gates for users to interact and contribute and be acknowledged as the expert in their field.

The community thrives because of the high quality of content and it is possible by the users action and moderation. Users vote up the good questions and answers and vote down the bad quality content or repeated posts. There is more than 6 million votes casted in the website and the user with enough reputations are allowed to cast 40 votes per day.

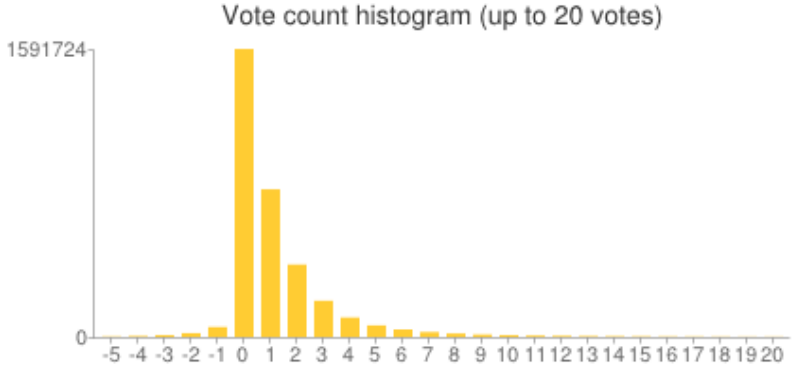


Figure 4.10: Vote count histogram

Vote	Vote Count
less than 20	25
0	504754
1-10	7802690
11-100	718650
101-1000	6460
1000-5000	11

Table 4.4: Questions votes count

The analysis of the questions and votes shows that every question receives 3.06 votes on average. One question received negative 115 votes and the highest vote received to a question is 2499. Similar analysis of answers and their votes shows that, on average an answer receives 0.99 votes and the lowest vote to an answer is negative 57 and the highest vote is 4432.

4.4 Using Semantic Web Technologies

The data extracted from StackOverflow website is in the form of simple text, some of the posts contain HTML codes but they are snippets of code and are represented as text in the database. The users categorize the posts by using tags and it gives information about the topic and the programming language the question is being asked. The answers do not have special tags but the tags of the questions are applied to the answers as well.

4.4.1 Using RDF and Linked Data

All the user data, post data, votes and badges are transformed in RDF data by applying simple RDF schema and ontologies.

The website only shows basic user profile information due to the privacy reasons and FOAF ontology is used to describe the data. An example of the simple user profile information is as follows:

```
<foaf:Person>
    <foaf:name> Geoff Dalgas </foaf:name>
    <foaf:mbox_sha1sum> b437f461b3fd27387c5d8ab47a293d35 </foaf:mbox_sha1sum>
    <foaf:based_near> Corvallis, OR </foaf:based_near>
    <foaf:age> 35 </foaf:age>
    <foaf:OnlineAccount> http://stackoverflow.com/users/2/geoff-dalgas
    </foaf:OnlineAccount
</foaf:Person>
```

Similarly, the posts created by users, the questions and answers are described using SIOC ontology. The content is described and linked with the user RDF using the similar URIs.

```
<sioc:Post rdf:about=" http://stackoverflow.com/questions/89228/calling-an-external
-command-in-python">
    <dcterms:title>Calling an external command in Python</dcterms:title>
    <dcterms:created> 2008-09-18T21:42:52.667 </dcterms:created>
    <sioc:has_container rdf:resource=" http://stackoverflow.com/questions/tagged
    /python"/>
    <sioc:has_creator>
       <sioc:UserAccount rdf:about=" http://stackoverflow.com/users/170339/bludger "
       rdfs:label="bludger"> </sioc:UserAccount>
     </sioc:has_creator>
     <sioc:content>How can I call an external command in Python</sioc:content>
```

```
<sioc:topic rdfs:label="python" rdf:resource=" http://stackoverflow.com
  /questions/tagged/python"/>
<sioc:topic rdfs:label="command" rdf:resource=" http://stackoverflow.com
  /questions/tagged/command"/>
<sioc:has_reply>
   <sioc:Post rdf:about=" http://stackoverflow.com/a/89243/1313327">
       <sioc:content>Look at the subprocess module in the stdlib: from subprocess
        import call call(["ls", "-l"]) The advantage of subprocess vs system is that
        it is more flexible (you can get the stdout, stderr, the "real" status code,
        better error handling, etc...). I think os.system is deprecated, too, or will
        be: http://docs.python.org/library/subprocess.html#replacing-older-functions
        -with-the-subprocess-module For quick/dirty/one time scripts, os.system
        is enough, though.</sioc:content>
        <dcterms:created>2008-09-18T23:42:52.667</dcterms:created>
        <sioc:has_creator>
           <sioc:UserAccount rdf:about=" http://stackoverflow.com/users/11465
            /david-cournapeau" rdfs:label=" david-cournapeau "> </sioc:UserAccount>
         </sioc:has_creator>
      </sioc:Post>
 </sioc:has_reply>
</sioc:Post>
```

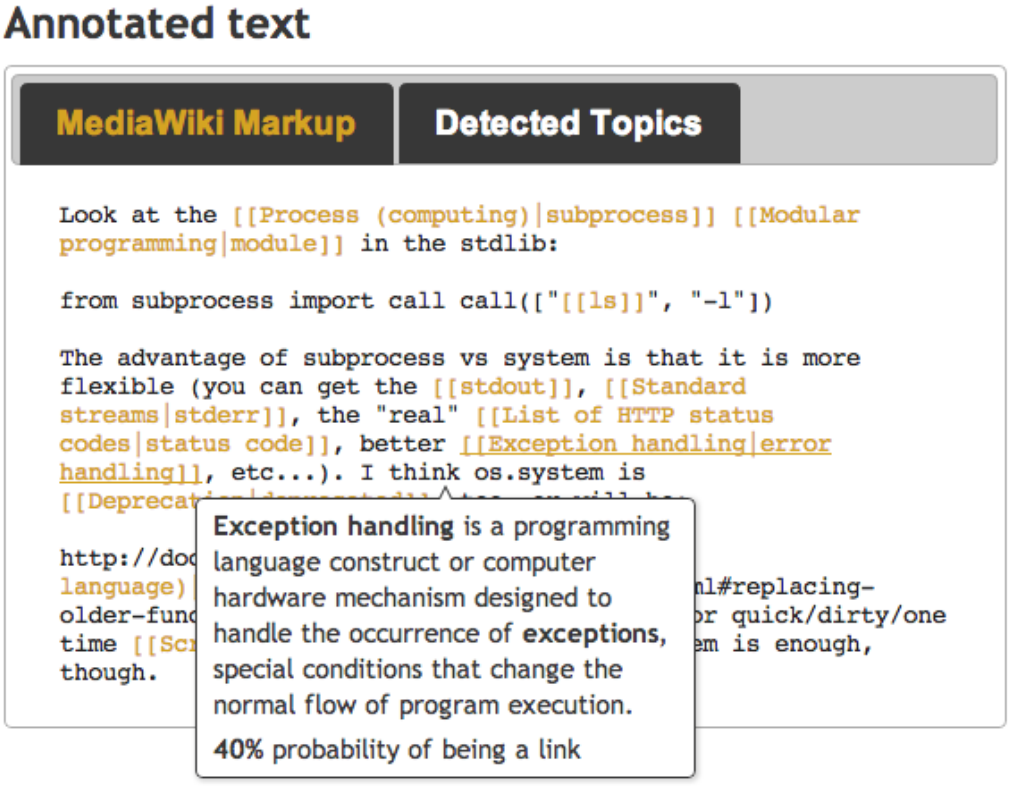
4.4.2 Topic Disambiguation

The StackOverflow dataset is sparsely annotated by user-generated tags and it is not linked with any other datasets. When user creates a question, they add tags to it to categorize into different topics but the answers have the tags from the questions. Also, all the main topics inside the text of question or answer is not clearly stated. The topics are ambiguous and not linked to any vocabulary or properly annotated.

A sample of the question, answer and tag data is annotated with the links from Wikipedia datasets and Drupal datasets to resolve the name and topic disambiguation. These services do the name entity recognition and match the entities with the appropriate topics and categories. The service does not convert the text into Linked Data or RDF, the returned data is further transformed into RDF and linked with the DBpedia dataset.

Wikipedia Miner service is used to annotate a small sample of posts, the service returns the text with annotated topics embedded into the text. The service accepts simple text or HTML and one can specify the density of links to be added and the level of accuracy required from the service. Below is an example annotated text of a question and answer posted.

(a) Annotating a text question



(b) Annotating a text answer

Figure 4.11: Wikipedia Miner web service annotating a text question and an answer

The Wikipedia miner service uses a word sense disambiguation based machine learning algorithm and where it detects key terms in a text excerpt and disambiguating then against Wikipedia article. It provides a JAVA API to access the Wikipedia database, including all the categories and it can be searched, browsed and iterated over **?**.

OpenCalais is another web service used to annotate the StackOverflow posts with the Drupal dataset. This tool creates a semantic rich metadata for the content using the natural language processing, machine learning and name disambiguation algorithm. It provides many services; it provides tag integration with different taxonomy and vocabulary, geo-mapping of location and semantic annotation of keywords. An example annotation of text using OpenCalais is below.

Text sample question: "Does Python have a ternary conditional operator? If not available, is it possible to simulate one concisely using other language constructs?"

```
<SocialTags>
  <SocialTag importance="2"> Conditional
<originalValue>Conditional (programming)</originalValue>
  </SocialTag>
  <SocialTag importance="2"> Python
   <originalValue>Python (programming language)</originalValue>
  </SocialTag>
  <SocialTag importance="2"> C
   <originalValue>C (programming language)</originalValue>
  </SocialTag>
  <SocialTag importance="2"> Ternary operation
   <originalValue>Ternary operation</originalValue>
  </SocialTag>
  <SocialTag importance="1"> Software engineering
   <originalValue>Software engineering</originalValue>
  </SocialTag>
  <SocialTag importance="1"> Computing
   <originalValue>Computing</originalValue>
  </SocialTag>
  <SocialTag importance="1"> Computer programming
   <originalValue>Computer programming</originalValue>
  </SocialTag>
</SocialTags>
```

As seen from above snippet, the OpenCalais service finds the keywords and matches it with a taxonomy or vocabulary and assigns the importance to the tag that it is disambiguating.

Both the services do the name entity recognition and match it to a known vocabulary and taxonomy. They do a natural language processing of the text and annotate the keywords. This annotation is then matched with the Wikipedia topics and the StackOverflow data is linked to the Wikipedia data.

These links when analyzed tell the type of categories the keywords matched to give additional information that the StackOverflow tags do not provide. The analysis shows the most asked question is asked from the following categories of programming languages:

Annotated keyword	**StackOverflow Tags**
Programming Language	C#, JAVA, Python
Framework	jQuery, ASP.net
Environment	Android, iPhone
Database	MySQL, SQLite

Table 4.5: Keyword analysis of StackOverflow question tags

It can be seen from the above example, name entity recognition, creating vocabulary and matching the keywords to a topic and linking it to another knowledgebase provides additional information. This leads to better search and discovery of information and using this an expert in a particular field can also be determined. JAVA being a programming language is also an Object Oriented language and the expert of JAVA also has a good grasp of Object Oriented programming concept and hence can help users in both the scenario.

4.4.3 Expert Finder

According to the StackOverflow website the top user or an expert of C# is Jon Skeet with more than eighty thousand reputation points and Python is Alex Martelli with more than nineteen thousand reputation point. These users appear on the individual pages of the tags as the top users and without the tag disambiguation they only appear as an expert on a particulate tag, not the joint concept of the topic.

Tag	**Top User with reputation point**
C#	Jon Skeet (80.6k)
Java	Jon Skeet (39.7k)
Python	Alex Martelli (19.8k)
PHP	Pekka (9k)
Javascript	CMS (12.3k)

Table 4.6: Top users of top tags in StackOverflow

When the tags are disambiguated and the keywords are matched to the topics, both Java and C# is categorized at the Object Oriented programming language and here Jon Skeet is considered as an expert in the whole area with more than one hundred and twenty thousand reputation points. Similarly, when the programming languages

are further categorized as server side script ion language with Python, PHP and Perl as main languages, Alex Martelli is considered as an expert and the user CMS is expert in the clients side languages such as Java and AJX with twelve thousand reputation points.

Disambiguated Keywords	**Top User with reputation point**
Object Oriented programming (C#, Java)	Jon Skeet (120.3k)
Programming language(C#, Java, Python)	Jon Skeet (120.7k)
Server side Scripting language (Python, PHP, Perl)	Alex Martelli (20.2k)
Clientside Scripting language (Javascript, AJAX)	CMS (12.3k)

Table 4.7: Top users of top disambiguated topics in StackOverflow

Semantic web and linked data helped in topic recognition and disambiguation and experts in broader concept and also specialized field can be ascertained even though these information is not present in the main website. The Table 4.7 only shows the experts in StackOverflow domain, when the data from multiple website and question/answer forums are combined, the linked data can help find experts in across domain in bigger set of users and help in better search and discovery of experts and information.

Chapter 5

Conclusions

Social networking is part of human interaction and communication process and the World Wide Web has made it easier and simpler for people to connect and interact. People not only connect to their friends and families, they also interact with strangers from all over the world, they create a network and community with people with similar interest and expertise.

The social semantic web technologies and their different examples are discussed as well. Some of the examples provided earlier show that the data integration across platforms is possible but to create a unified knowledgebase from different networks of web has limitations. Some of the data in the web is freely available but most of the data is still bound behind the closed walls of different websites. Using the APIs of the websites and with the proper encouragement and initiative of users, they can free their own data. These different knowledgebase can be integrated using the semantic web technologies and can be linked with the Linked Data Cloud.

In this report different types of social networking services available in the Web is analyzed and the motivation of creating communities is seen. In the current web, an agile approach is taken to create a network of people based on a topic and people come together with common purpose and solve problems and create purposive social network. This network is small, agile and thrives on the user contribution. Different types of crowdsourcing system are also described where people come together to solve problems and create a knowledgebase. This type of system requires a strong framework to support engagement and incentive for people to contribute.

StackOverflow website, a question and answer forum for programmers, is studied to see the creation and framework of purposive social network. In this network people ask questions and other experts in the field provide answers and gain reputation points. This system has a strong incentive design that motivates users to contribute. The system utilizes the user to moderate the community and to control the quality of the content by community voting.

The post of the website is analyzed to see the structure of the community, the network is not form by explicit connection of users but by studying the user interaction and how the knowledge is connected with each other. The network ties, user interaction and the incentive model is studied to see how the website with a small community of programmers created a self-sustaining environment for user to participate and continuously create high quality questions and answers and solve problems.

The text analysis on a small sample of data is done and Wikipedia miner and Open Calais tools is used to solve the name entity problem. These tools does a natural language processing on the text and uses machine learning algorithm to match the name with Wikipedia topic and Drupal vocabulary. The keywords and topics are categorized and linked with other knowledgebase.

Semantic web technologies and Linked Data is used to solve the data integration problem of the current web and it can be used to integrate heterogeneous systems and create a platform where user can generated knowledge, consume it and utilize it solve problem and help the community.

5.1 Future Work

In this report, StackOverflow website is analyzed to see the use of Linked Data and semantic web technologies in a purposive social network and how the use of linked data and solve the categorization and name entity ambiguities problem. These technologies and similar framework can be used to integrate different knowledgebase and communities together and can be used to create a bridge between isolated communities. People from one network can discover experts from other network and help solving the problem faster and easier.

For the future work, this framework will be applied to other question and answer forums like Reddit and Quora [1]. This will include integrating all the questions and answers together, doing a name entity recognition to disambiguate topics and linking the similar topics and categories across different websites. Then using the Linked Data all the knowledgebase will be linked to the Linked Data Cloud to generate richer information.

A simple user interface will also be designed so users can search for experts on a particular topic and to solve particular problems. The system will also allow users to discover information from the integrated knowledgebase and create a purposive community to interact and communicate with experts.

[1] https://www.quora.com/

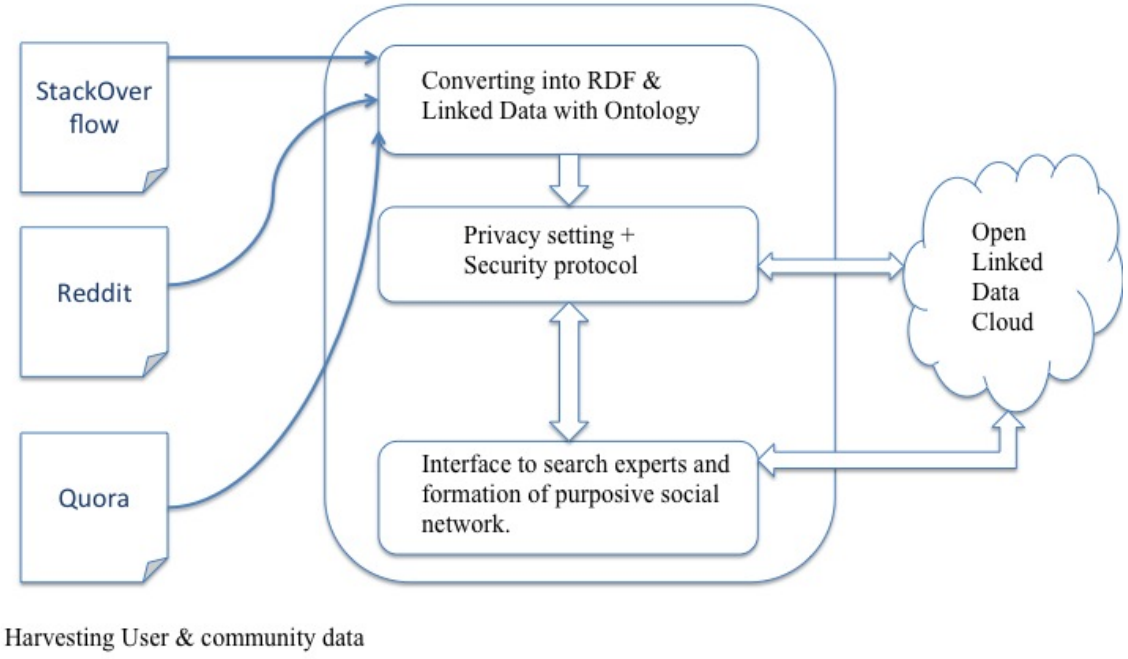


Figure 5.1: System design and framework for the formation of purposive social network using Linked Data

5.1.1 Gantt chart

The Figure 5.2 of Gantt chart outlines the timeline and the tasks required for the next ten months (40 weeks) to complete my thesis. Below it is a proposal to create a prototype of the system that harvest the online user and community data from Reddit and Quora, convert it into Linked Data and link it to the Linked Data Cloud. The system will also provide a user interface to browse and search the linked data so users can find right people with appropriate expertise to form purposive community and interact with each other.

The tasks to implement the prototype of the framework are as follows:

5.1.1.1 Reddit Data mining (Week 1-4)

The first month will be used mine the Reddit website data, especially from the programming sub-reddits. This will include the questions asked, users comments with the up votes and down votes of each post. The user profile information will also be collected to create a knowledgebase of experts in a field.

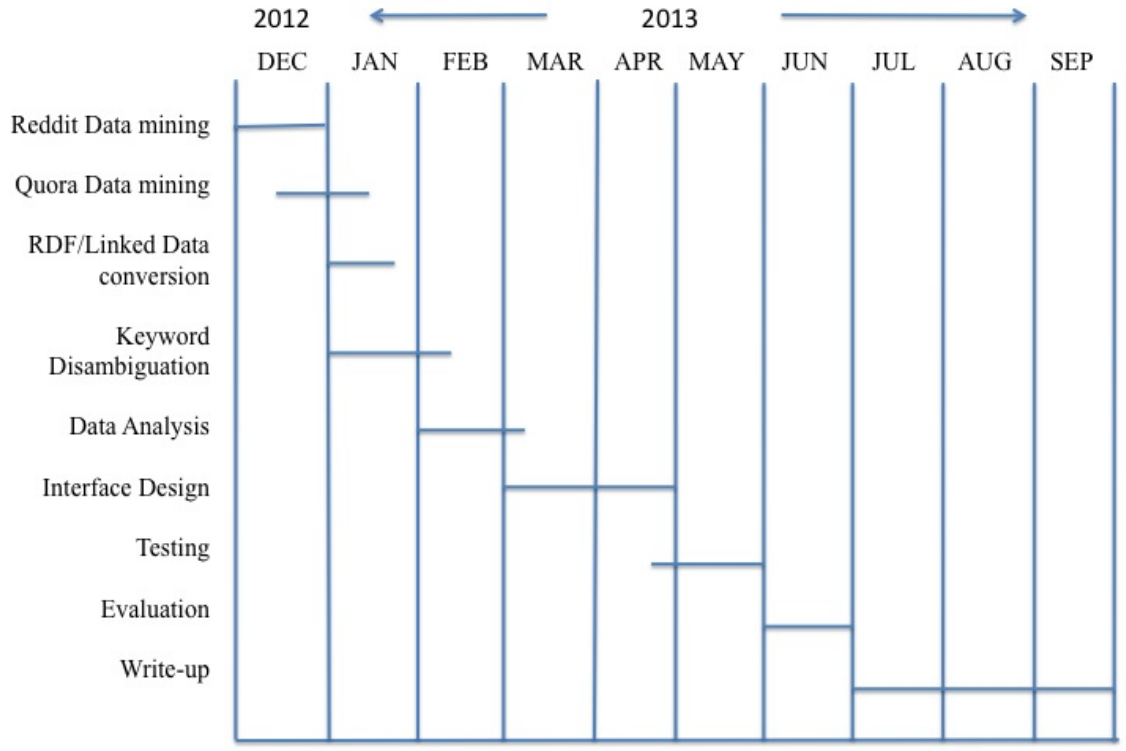


Figure 5.2: Gantt chart showing the timeline of task to do for writing PhD thesis

5.1.1.2 Quora Data mining (Week 3-6)

The data from Quora website can be simultaneously harvested with the Reddit data as it requires the same program with slight alteration of code. The programming categories will be used to gather questions and answers with the votes and user information.

5.1.1.3 RDF/Linked Data conversion (Week 5-7)

The program that converted StackOverflow data can reuse the same ontology to convert the Reddit and Quora data into RDF and Linked Data. The program can run in the background so when the Reddit and Quora data is being harvested, it will also be converted into RDF and N-Triples and saved in a Triplestore.

5.1.1.4 Keyword disambiguation (Week 4-10)

As the previous task, WIkipedia miner program and OpenCalis web service can run in the background and simultaneously get the website posts, use the web service to do natural language processing, find the main topics and keywords and link it to the

appropriate Wikipedia topic and Drupal vocabulary. This phase will also include linking the data to the Open Linked Data Cloud and creating a integrated knowledgebase.

5.1.1.5 Data Analysis (Week 9-13)

This task requires analysing the Linked Data, RDF and keywords data to find the main topics, remove the noise and find the best sample set to perform the final analysis to prove the concept of purposive social network and using the linked data to improve search and discovery of expert.

5.1.1.6 Interface Design (Week 13-20)

The main developments of the front end, the algorithm to query and search the user data and community data, the design of user interface and the final deployment of the framework will be done in the span of two month. A simple user interface will be designed so people can search for experts on any topic and also search for information from the integrated knowledgebase of the three website used in the experiments.

5.1.1.7 Testing (Week 19-24)

The final week of programming will coincide with the first week of testing of the system and the next three week will be utilized in fixing the bugs in any of the features and design.

5.1.1.8 Evaluation (Week 25-28)

Once the testing is finished and the system is working properly, the system shall be evaluated. It could be a user evaluation or statistical evaluation and the generated data is analyzed to gather the final result.

5.1.1.9 Thesis write-up (Week 29-40)

The last three month will be used for writing the thesis incorporating all the procedures, algorithms and results generated from the system. The literature currently read and review will be included with addition literature reviewed in the mean time.

Appendix A

Appendix

The StackOverflow website had be analyzed to see the formation of purposive social network and users communication network is studied. Chapter 4 includes many of the important graph, charts and histograms showing user interactions and activities. Some more of user activity has been analyzed but it was not included in the chapter. Some of the charts are as follows:

A.1 User activity

StackOverflow is a USA based website and since it is an english language website, many of the users are from english speaking nations and from western countries. Analysis of users activities by hour shows when user interact the most.

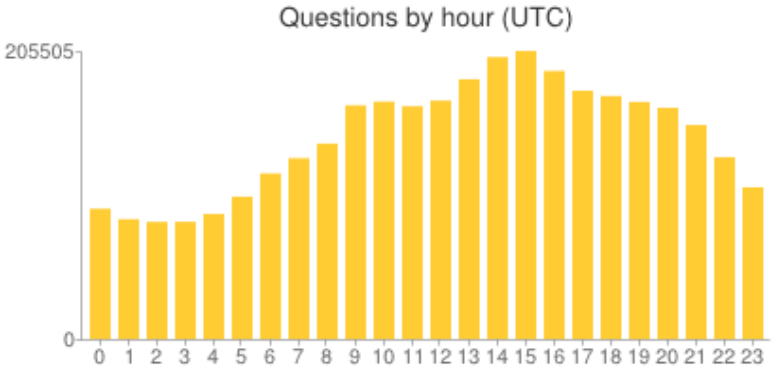


Figure A.1: Questions posted by the hour

As can be seen from the figures above, afternoon is the most active time for the users to post questions and answers in the USA time zone. Also, 9 and 10 am in the morning, the afternoon time for some part of Asia and Europe also shows some rise in user activity. Geographical analysis of the website suggests that most of the registered users are from USA and then from Europe.

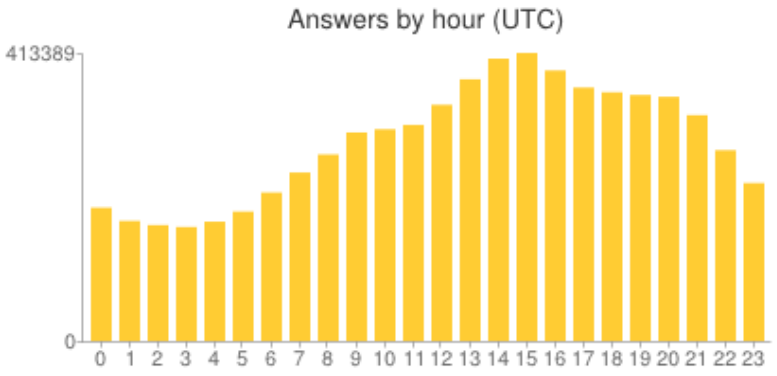


Figure A.2: Answers posted by the hour