UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL AND APPLIED SCIENCES

Electronics and Computer Science

Using Linked Data in Purposive Social Networks

by

Priyanka Singh

Thesis for the degree of Doctor of Philosophy

May 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCES
Electronics and Computer Science

Doctor of Philosophy

USING LINKED DATA IN PURPOSIVE SOCIAL NETWORKS

by Priyanka Singh

This work is all about . . .

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Priyanka Singh , declare that the thesis entitled *Using Linked Data in Purposive Social Networks* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: (**?**), (**?**) and (**?**)

Signed:...

Date:...

Acknowledgements

Thanks to no one.

Nomenclature

w	The weight vector

Chapter 1

Introduction

1.1 Overview

In the beginning of World Wide Web (WWW), people created static HTML webpages to share information and opinion and the documents were linked using hyperlinks. People used emails to communicate with other people even before WWW was invented. Web 2.0 has changed how people use and interact on Internet. They moved from static webpages to share content to dynamic sharing of information and started to connect with people and objects. Now in WWW, not only documents was linked with each other, people are also connected with each other. The WWW provided a distributed and easy to use platform for people to create content and communicate with other people from all over the world. It also makes it easier for people to share information and collaborate with each other. It's easy to form communities and share knowledge. In these social networks people are connected with each other using explicit relationship (friendship) or through data. Information is shared directly with the personal social network or it cascades through the network effect more rapidly and have higher impact.

There are many different Social networking Service (SNS) created for a particular purpose and they target a particular usergroup. 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. Different websites provides different functionalities, for e.g., Facebook [1] is for people to connect with their friends and Twitter [2] for microblogging. There are also content specializing services, YouTube [3] for

[1] http://www.facebook.com/

[2] http://twitter.com/

[3] http://www.youtube.com/

videos, Flickr [4] for pictures, etc.

The Web 2.0 social web has also provided a distributed platform for people to come together from different part of the world and collaborate to solve common 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.

The main drawback of current services is that they are all closed or partially open networks 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 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 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 their 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. The text based search results only uses the keywords used in the search query to retrieve results, they don't expand the query to broader or narrower fields. Also, the search engines cannot access the closed communities and websites. 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

[4] http://www.flickr.com/

the subset of experts registered to that service, they loose a whole community of expert in other network or community.

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. 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. It is machine-readable and ontologies are implemented to describe the data and to give it meaning, understood by both machines and humans **?**. The added semantics to the data when linked can create a network of interlinked categories that can be used to search for extra information and provide more details to the queries. The same techniques can be used to create a semantic rich user profile and the experts can be linked with categories. In the scenario when a query has no appropriate answers then the correct experts can be recommended to help with the question.

This thesis discusses the concept of Purposive Social Network (PSN), a social network with a purpose, where people come together with a common objective to get information, build knowledgebase, solve problems or achieve some common goal. Here people with similar interest and varied expertise come together, use crowdsourcing technique to solve a common problem and build tools for common purpose. The term is defined in much more details in chapter 3.

The aim of this thesis is to show Semantic Web and Linked Data can be used to create an PSN where the data is open, linked and have added knowledge. This creates a distributed social network that improves the search and discovery of information across platform and recommend experts that can help answer queries with no results. StackOverflow [5] and Reddit [6] websites are chosen to study the purposive network and different network ties. The research focuses on question and answering systems for programmers and developers where users ask technical questions to the community and experts in the field provides solution using crowdsourcing techniques. Further on, Linked Data and semantic web technologies is used for name disambiguation of keywords and topics and create a keyword matrix. This also helps in improving search and discovery of answers for questions, and experts in a field and provide useful information that is otherwise unavailable in the website.

[5] http://stackoverflow.com/

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

1.2 Research Challenge

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.

The main research question and challenges encountered are discussed below.

1.2.1 Data Harvest and Data Integration

The study of purposive social networks requires user data from different 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. The focus of this research is QA systems and most of the information is crowdsourced and the knowledgebase is created by the contribution of multiple users. The details of each users contribution is also required to create a proper user models of the experts.

1.2.2 Name Entity Disambiguation

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 **?**.

1.2.3 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. The relationship are created by studying their communication networks (posts, questions, answers, etc.) if users don't create an explicit relationship. This also creates a network of experts based on topics and categories.

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 **?**.

1.2.4 Linking the Data

The website data is collected and analyzed using Wikipedia-miner [7] **?**, DbpediaSpotlight and OpenCalis [8] 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 **?**.

1.2.5 Search and Query

All the data is converted into RDF and stored into a triplestore. The recognized keywords are also given an accuracy score and stored into a matrix structure. When a query is made, it's analyzed and broken down into main keywords and these sets of keywords are matched with the listed keywords in the database. The best matched results with the same frequency of keywords are returned. The SPARQL endpoints also matches the keywords with the users associated with the keywords and returns the list of experts in the field.

1.3 Research Contribution

The primary aim of the research represented in this thesis is to show Linked Data and Semantic Web technologies can be used to improve the search of answers and experts in crowdsourced question and answering system. To achieve this, first Purposive Social Network is defined and datasets are collected from websites that fulfills the criteria.

A system is created that provides the tools to harvest data, clean it and then structure it into RDF. In this process the entity disambiguation problem is solved using the combination of available tools and keywords are categorized and mapped to the concepts.

[7] http://wikipedia-miner.cms.waikato.ac.nz/

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

The system also created weighted keyword matrix for each question and answers. The algorithm then searchers for answers to unanswered questions and lists experts in the field. The results are evaluated and analyzed in details.

Also, this thesis, purposive social network is studied in detail, how it is formed, why it is created, its 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.

1.4 Structure of Thesis

This thesis is structured into 7 chapters, this is the first chapter which introduces the research question and research contribution.

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 defines and 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 details of how the system was designed, data was collected, cleaned and structured. It also discussed different tools used to solve the keyword disambiguation problem and how it was linked to each other. This chapter also describes the algorithm that is used for concept mapping and creating the document term matrix to improve the search of queries and experts.

Chapter 5 provides the purposive social network analysis. 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.

Chapter 6 evaluates the system. It describes the experiment designed to test and evaluate the system bu users and how the data was collected. This is later analyzed by different statistical tests. This section shows 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 thesis and describes the future work to be done.

Chapter 2

Background

Hello

2.1 Emergence of Social Web

2.1.1 Web 2.0 and Social Networks

2.1.2 Content-specific Social Networking Services

2.2 Collective Intelligence and Crowdsourcing

2.2.1 Crowdsourcing Services

2.2.1.1 Wiki

2.2.1.2 Forums and Messaging Boards

2.2.1.3 Q&A Services

2.2.2 Human Computation

2.2.3 User Recommendation System

2.3 Semantic Web and Linked Data

2.3.1 Social Semantic Technology

2.3.1.1 FOAF

2.3.1.2 SIOC

2.3.1.3 OPO

2.3.2 Linked Data Technology

2.3.2.1 Linking the Datasets

2.3.3 Open Linked Data and Graph

2.4 Semantic Search

2.4.1 Keyword Disambiguation

2.4.2 Concept Mapping

2.4.3 SPARQL Queries

Chapter 3

Purposive Social Network

Hello

3.1 What is Purposive Social Network

3.1.1 Information Based Community

3.1.2 Interest Based Community

3.1.3 Expert Based Community

3.1.4 Location Based Community

3.2 Characteristics of Purposive Social Network

3.2.1 Community Size

3.2.2 Focused Interest

3.2.3 Direct Communication

3.2.4 Active Participation

3.2.5 Short Lifespan

3.2.6 Strong Incentive

3.3 Benefits of Purposive Social Network

3.3.1 Information Exchange and Self-interest

3.3.2 Symbiotic Relation and Social Exchange

3.3.3 Social Recognition and Personal Satisfaction

3.3.4 Recommendation System

3.3.5 Expert Finder

3.4 Crowdsourcing in Purposive Social Network

3.4.1 Recruiting and Retaining Users

3.4.2 Incentive Model

3.4.3 Quality Control

3.4.4 Search and Discovery of Quality Content

Chapter 4

System Design and Methodology

Hello

4.1 Data Mining

4.1.1 StackOverflow Data Mining

4.1.2 Reddit Data Mining

4.2 Data Structuring

4.2.1 Data Cleaning

4.2.2 Converting into N-Triples

4.2.3 Linking the Data

4.3 Keyword Disambiguation

4.3.1 OpenCalais Service

4.3.2 DBpedia Spotlight Service

4.4 Concept Mapping

4.4.1 Keyword Matrix

4.4.2 Document Term Matrix

4.5 Search and Query

4.5.1 Database Design

4.5.2 Searching Answers of Unanswered Questions

4.5.3 Expert Finder

Chapter 5

Purposive Social Network Analysis

Hello

5.1 Network Linkage and Social Ties

5.2 Role of Individual Actors

5.3 Incentive Design

5.4 Quality Control

5.5 Linked Data Graph

5.6 Discourse Analysis

Chapter 6

Evaluation

Hello

6.1 Experiment Design

6.1.1 Calculating Sample Size

6.1.2 Selecting Questions

6.2 Data Collection

6.2.1 Questionnaire Design

6.2.2 Dataset Frequency Distribution

6.3 Results

6.3.1 Keywords T-Test

6.3.2 Q&A Correlation Test

6.3.3 Summary

Chapter 7

Conclusions

It works.

7.1 Summary of Results

7.2 Limitation of System

7.3 Future Work

Appendix A

Stuff

The following gets in the way of the text....