Linked Data in Crowdsourcing Purposive Social Network*

Priyanka Singh
University of Southampton
Southampton, UK
ps1w07@ecs.soton.ac.uk
WAIS Group, ECS

Prof. Nigel Shadbolt
University of Southampton
Southampton, UK
nrs@ecs.soton.ac.uk
WAIS Group, ECS

ABSTRACT

Internet is an easy medium for people to collaborate and crowdsourcing is an efficient feature of social web where people with common interest and expertise come together to solve specific problems by collective thinking and create a community. It can also be used to filter out important information from large data, 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. Semantic web technologies can be used to structure the community data so it can be combined, decentralized and be used across platform. Using such tools knowledge can be enhanced and easily discovered and merged together. This paper discusses the concept of a 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. The StackOverflow website is chosen to study the purposive network, different network ties and roles of user is studied. Linked Data is used for name disambiguation of keywords and topics for easier search and discovery of experts in a field and provide useful information that is otherwise unavailable in the website.

Keywords

Social Media; Social Machine; Crowdsourcing; Q&A; Name Entity Disambiguation; Linked Data

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Documentation; Human Factors

1. INTRODUCTION

The Web 2.0 social web has provided a platform for people to collaborate, solve problems, form communities based on similar interest and use the collective intelligence for distributed problem solving and create knowledgebase. Experts and people with similar interest connect with each other, create relationships and communities with strong and weak social bonds. A social network that is formed by people with common interest, goal, objective or purpose with explicit or implicit relation is studied in this case and named purposive social network. Messaging boards, question-answer forums, wikis are example of this type of social communities where people come together to create an emerging knowledge.

The issues with the large communities are to search for relevant informations and experts. In question and answering forums many new users are discouraged when they do not receive appropriate answers to their queries or when their questions are buried down in the large volume of new inputs. Also when the questions are repeated experts are less inclined to answer them and this reduces the efficiency, quality and user experience of the whole system. Another problem with these communities are the strict categorization of the data. The system allows user to use tags or categories from a specific list that binds the users to a smaller group of experts in the field and when users do not have a conclusive idea of which category they belong, it makes them harder to find the right solutions to their problems. Also, it is harder to find experts in multiple overlapping field because of the community structure and this increases the long tail of users who do not receive solutions to their problems.

StackOverflow [1] website is studied to analyse purposive social network. This website is used by computer programmers to ask questions and other expert programmers answer the questions, vote the questions and answers and keep quality control. 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 and issues of the purposive social network is studied alongside the incentive model of the website that encourage people to collaborate and contribute. The website data is collected and analyzed using Wikipedia-miner [2] [9] and OpenCalis [3] 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 Wikipedia articles and Drupal vocabulary and linked with the knowledgebase. This allows the broaden the categories and tags of the questions and answers and allows wider field to search for information and experts.

*

WWW, SOCM Workshop, '13, Rio de Janeiro
ACM 978-1-4503-2038-2/13/05.

[1] http://stackoverflow.com/
[2] http://wikipedia-miner.cms.waikato.ac.nz/
[3] http://www.opencalais.com/

2. STACKOVERFLOW AS PURPOSIVE SOCIAL NETWORK

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. The six-degree of separation theory says 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 or professionally to interact with one another [10]. In online world, where geographical boundaries are dissolved and have a certain degree of anonymity and privacy, people can reach to others with similar background and interest. StackOverflow can be considered as purposive social network where experts in the field of computer programming and software development ask questions and provide answers. The users also do the quality control in the website where they up vote good questions and answers and down vote bad content and moderate the community because of strong incentive. The motive for users to join such network varies from interaction with users with similar expertise and interest to solving the problem and gaining reputation for their knowledge. It is a purposive community where people create a symbiotic relationship and help other users by providing knowledge beneficial to the group [12]. The lifespan of any communicative channel is small, once the question is answered, users do not participate in the thread any longer.

StackOverflow website uses many of the crowdsourcing and social features of Web 2.0 to create a purposive social network ecosystem. It is like a forum and messaging board where user broadcast their questions to whole community of experts in a field and are not directed toward any one user. Any user who has the answer can response, the questioner can accept the answer or could wait for a better response from others. User create knowledge using the wiki function of the website and other users can add to the knowledgebase. Moderators also edit the answers provided to improve the quality. They also tag their question from the list of tags used in the website, this allows all the users subscribed to a tag to see the questions and answers, hence folksonomy help to categorize the content. Users also rate the quality of the questions and answers by voting it up or down to remove spam and repeated questions and help in sustaining high quality of content. Human computation is done when user asks questions about the errors they have in their program, generated by the compilers, and provide solutions and algorithms to solve it.

3. CROWDSOURCING IN PURPOSIVE SOCIAL NETWORK

Crowdsourcing system is 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. This system depends 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 [14]. It is also used to manage and moderate the community and do quality control. StackOverflow uses all these crowdsourcing technique to sustain the system and below are the following issues it faces.

3.1 Recruiting and Retaining Users

StackOverflow depends on large amount of user participation to create an active community. There are several methods to get users to contribute, like making users to play games 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 and to make the system easy to use, free and open so people can contribute 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 [2].

3.2 Incentive Model

Designing an incentive model for a large-scale crowdsourcing system is complex. It should be easier for people to contribute but also keep track of the quality of content created, 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. StackOverflow works on the users desire to be recognized and 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 [11]. Users who generate quality questions and promptly answers are rewarded for their contribution with badges or points as positive reinforcement. Similarly, when people are spamming or creating poor quality content or are asking repetitive question are given negative points and their contents are 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 [5]. Another 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 [13]. 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 [6].

3.3 Quality Control

Collective voting to rank large user generated content 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. By using the thumbs-up or thumbs-down style ratings by the users, questions on StackOverflow higher quality contributions are prominently displayed 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 [7]. 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 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 [4]. 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 [1].

3.4 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. 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. The drawback of such system is that the users are bound by the tags available to use in the website. Creation of new tags requires moderators approval and large reputation points, this makes harder for new users who are not sure about the category of their question to reach the right experts, hence get the right answers. Also, the users are categorized as experts in individual tags in the website, there is no means to find the experts who overlap in two or more categories. If the expert is not subscribed to the correct tags, they will not be able to see the questions and help other users.

4. STACKOVERFLOW DATA ANALYSIS

4.1 Network Linkage and Social Ties

The StackOverflow questions, answers and user profile data and meta data is collected using it's API. The tags data and it's number of instances is collected by screen scraping and all the data analysed in this paper are publicly available. In this analysis, the users asking and answering questions, voting the responses and commenting are the main entities of the social network. Their network ties are measured

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

Table 1: StackOverflow at glance as of June 2012

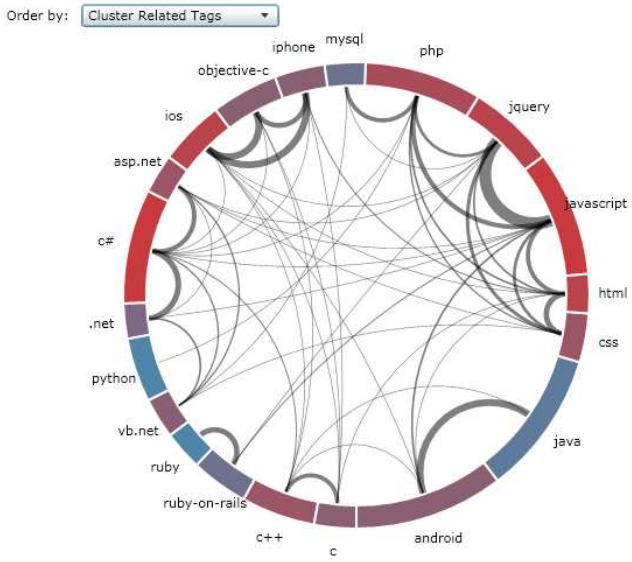

Figure 1: Related tags clustered together [3]

by their communication and interaction between them i.e. 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 [10]. The amount of their contribution is measured by crowdsourcing where people give up votes or down votes to their posts and the badges they receive for their contribution. 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 answers. Users have casted more than 26 million votes to mark the good questions and answers. In the year 2012, each questions have on average 1.645 number of answers. 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.

The questions and answers are provided with tags to categorize and arrange for easy search and discovery. The analysis of questions shows that 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. [3] 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 Figure 1 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.

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 Incentive Design and Quality Control

There are more than 1.2 million registered users in StackOverflow and they ask the questions, answer it, vote it and moderate the community. 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. There are 669554 users with 1 reputation point and there is 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. Although the website had 17 million unique visitors in the month of January 2013 [8], most of the users of website do not register or contribute to the knowledgebase.

The website uses an elaborate point system and contribution badges to encourage user participation. The reputation of the user has a direct correlation with the trust in the community. There are 77 different types of badges given to the user based on their contribution from the badge for user who asks questions with 1 reputation point (Student), to the user who edits the answers to make posts better (Editor) and there are badges even for 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 top contributor and user with highest reputation are featured on the question page, giving the user more visibility and acknowledgement of their expertise. This encourages participants to accumulate more points and contribute to get recognition.The other method that encourages the users to participate is the promptness of the response. 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.

When an answer is voted up user gains 10 reputations and 5 points when the question is voted up. When an answer is accepted the user receives 15 points and looses 2 point when a question or an answer is voted down. Negative point system keeps the spamming in check and repeated questions and answers are avoided. Higher reputation points gain more privileges as 15 point are required to up vote and 50 points allows users to comment. To stop harassment and spam, user requires 125 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 user's action and moderation. Users vote up the good questions and answers and vote down the bad quality content or repeated posts. There are more than 6 million votes casted in the website and the user with enough reputations are allowed to cast 40 votes per day. The analysis of the questions and votes shows that question receives 3.06 votes and an answer receives 0.99 votes on average. One question received negative 115 votes and the highest vote received to a question is 2499 and an answer received negative 57 votes and the highest vote is 4432.

5. USE OF LINKED DATA IN STACKOVERFLOW

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. or properly annotated. A sample (for the month of June 2012) of the question, answer and tag data is annotated with the links from Wikipedia datasets using Wikipedia Miner and Drupal datasets using Open Calais 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 returned data is further transformed into RDF and linked with the DBpedia dataset using special scripts.

Table 3 shows that the 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 information 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 scenarios.

5.1 Expert Finder

StackOverflow shows Jon skeet as the top user or an expert of C# with more than 80 thousand points and Alex Martelli in Python with more than 19 thousand point, without the tag disambiguation they only appear as an expert on a particulate tag, not the joint concept of the topic. 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 120 thousand points. Similarly, when the programming languages are further categorized as server side scripting 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 12 thousand points.

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. Table 3 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

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 2: Top users of top tags in StackOverflow

Disambiguated Keywords	Top User
Object Oriented prog. (C#, Java)	Jon Skeet (120.3k)
Programming lang.(C#, Java)	Jon Skeet (120.7k)
Serverside Scripting lang. (Python, PHP)	Alex Martelli (20.2k)
Clientside Scripting lang. (Javascript, AJAX)	CMS (12.3k)

Table 3: Top users of top disambiguated topics in StackOverflow

in bigger set of users and help in better search and discovery of experts and information.

6. CONCLUSION AND FUTURE WORK

Social networking is part of human interaction and communication process and the WWW has made it easier and simpler for people to connect and interact. In this paper StackOverflow website, a question and answer forum for programmers, is studied to see the creation and framework of purposive social network that thrives on user contribution and crowdsourcing for creating emergent knowledgebase.This type of system requires a strong framework to support engagement and incentive for people to contribute. The post of the website is analyzed to see the network ties and user interaction, 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. A sample of data is converted into RDF and Linked Data 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. The analysis shows that using Linked Data helps in better categorization and search and discovery of information and topic. It helps in finding the right experts and related questions and answers that cannot be previously done in the StackOverflow website. This shifts the long tail of the power graph where more information and experts are accessible to users and creates a better social machine. In the future, this framework is to be applied to other question-answering system like Reddit or Quora to integrate the knowledgebase across websites and platforms. This would make more information and experts available to new users, provide them with wider knowledgebase and help in solving problems quickly and efficiently.

7. ACKNOWLEDGMENTS

This work was supported by the EnAKTing project under grant EP/G0088493/1 and Web Science Doctoral Training Centre under grant EP/G036926/1, funded by the Engineering and Physical Sciences Research Council and the Research Innovation Services at the University of Southampton.

8. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, pages 183–194. ACM, 2008.

[2] A. Doan, R. Ramakrishnan, and A. Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96, 2011.

[3] C. Eberhardt. Visualising stackoverflow tag relationships with silverlight. Technical report, Scott Logic, 2012.

[4] A. Ghosh and P. Hummel. A game-theoretic analysis of rank-order mechanisms for user-generated content. In *12th ACM Conference on Electronic Commerce (EC)*, 2011.

[5] A. Ghosh and P. McAfee. Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*, pages 137–146. ACM, 2011.

[6] S. Jain, Y. Chen, and D. Parkes. Designing incentives for online question and answer forums. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 129–138. ACM, 2009.

[7] S. Jain and D. Parkes. The role of game theory in human computation systems. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 58–61. ACM, 2009.

[8] B. Marzewski. 2012 stack overflow user survey results. Technical report, StackOverflow, January, 2013.

[9] D. Milne and I. Witten. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 2012.

[10] P. Monge and N. Contractor. *Theories of communication networks*. Oxford University Press, USA, 2003.

[11] M. Richardson and P. Domingos. Building large knowledge bases by mass collaboration. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 129–137. ACM, 2003.

[12] C. Ridings and D. Gefen. Virtual community attraction: Why people hang out online. *Journal of Computer-Mediated Communication*, 10(1):00–00, 2004.

[13] V. Singh, R. Jain, and M. Kankanhalli. Motivating contributors in social media networks. In *Proceedings of the first SIGMM workshop on Social media*, pages 11–18. ACM, 2009.

[14] C. Treude, O. Barzilay, and M. Storey. How do programmers ask and answer questions on the web?: Nier track. In *Software Engineering (ICSE), 2011 33rd International Conference on*, pages 804–807. IEEE, 2011.