TEXT MINING ON Q&A WEBSITES

VARUN JAIN
Information Technology, Singhad College of Engineering Pune, India

PROF. KAUSTABH A. HIWAREKAR
Information Technology, Singhad College of Engineering Pune, India

ABSTRACT:

A well-known online Q&A forum where software developers post and answer questions related to programming. We perform the analyses on a long list of factors in the raw data and identify those that have a clear relation to response time.

The main search is for the tag-related factors, such as their "count" (frequency of the tag used) and the cardinality of their "subscribers" (number of users can answer questions containing the tag), provide much stronger confirmation that factors not related to tags. Finally, we learn models using the identified evidential features for predicting the response time of questions, which also manifest the significance of tags chosen by the questioner.

The process of defining the factors incorporates different steps for the analysis of documents like a social post or raw data (extraction of keywords and their matching to the related concepts) and their weighting. The concept of weighting involves different scores, such as statistical patterns and sentiment analysis which attempt to measure the proficiency of the user in the relative field.

1. INTRODUCTION:

Question and Answer internet websites, such as Quora or Yahoo Answers, use social media to function knowledge exchange between backend engineers and fill archives with millions of records that contribute to the structure of knowledge in software development. Understanding the role of Q&A websites in the documentation will enable us to make recommendations on how individuals and organizations can leverage this knowledge effectively. In this, we analyse data from a raw database to categorize the type of questions that are asked, and to explore which questions are answered well and which ones remain unanswered. Our preparatory findings indicate that Q&A websites are particularly effective at code reviews and conceptual questions. We present research questions and propose future work to search the incentive of programmers that contribute to Q&A websites, and to understand the conclusion of turning Q&A exchanges into technical mini-blogs through the editing of questions and answers.

Q&A websites such as Facebook Questions or Yahoo! Answers built around an "architecture of participation" where user data is aggregated as a side-effect of using Web 2.0 applications. Questions and answers on Q&A websites represent archives with numerous of records that are of value to the community. For the sector of software development, the website of knowledge between coders connected via the Internet. In the 2 years since its foundation in 2008, more than a million questions have been requested on Stack Overflow, and more than 2.5 million answers have been provided.

In all these cases, the Q&A portal becomes a platform for technical mini-blogs despite such widespread use of Q&A websites, the role of Q&A websites in the software documentation landscape is not well understood.

It is unclear what kinds of questions get answered well and which ones remain unanswered, as well as how to phrase questions effectively. We are not aware of how helpful the questions and answers are to a wider scope of people nor what the suggestions are from the duality of a Q&A website as a question and documentation gateway.

In this project, we present research questions and report the preparatory outcome to recognize the role of Q&A websites in software development using qualitative and quantitative research methods.

1.1AIM

Extract meaningful knowledge from micro posts shared on social platforms.

To comprehend the role of Q&A websites in the documentation which will enable us to make recommendations.

To understand the collective production of Q&A sites by studying the typical contributor's behavior.

1.2MOTIVATION:

Question-and-answer (Q&A) websites have shown to be a resource for helping people to solve their everyday problems.

The main motivation of our work is the amalgamation of traditional content analysis techniques (entity, keyword extraction) with semantic web technologies.

The analysis of shared data on a social surface may give a new input for advanced recommendation strategies, as it provide sprecious insights into people's interests, searches and information required.

1.3OBJECTIVE:

To parse the questions given by the users and tokenize the words. We will use the concept of stemming from reducing the tokens to their root or parent words. The creation of a domain wise cluster according to these words for data sorting and arrangement. Finally to give possible suggestions to the user.

2. DESIGN AND IMPLEMENTATION:

We have 3 different modules

1. Question Dataset: This is simply the data set of the Q&A websites on which we would perform text mining through miners.
2. Tokenizing: Parsed string will be tokenized according to relevance.
3. Stemming: Stemming include the processing of the words modify them according to root word

CODE AND CONCEPT IMPLEMENTATION

2.1 TOKENIZATION:

Tokenization is fragmenting text into lexemes. Tokenization can be stated as the task of breaking a stream of characters into words separating futile tags remaining from type-setting information in newspaper archives.

Tokenizing involves:

Deleting the “non-textual” items such as horizontal or vertical line and page-break or paragraph tags in HTML documents, or in electronic mail smileys — :-) or :-(— and quotation representation (such as >> at the starting of the line).

Removing pointless tags remaining from type-setting information in newspaper archives.

Eliminating parts which are out of natural languages: mathematical or chemistry-related formulae, programs.

```
public class Tokenizer {public Array List <String> to
kanize Question(String
question){StringTokenizerdefaultTokenizer = new String
Tokenizer (question);
ArrayList<String>tokensArrayList
=newArrayList<String>();
while (defaultTokenizer.hasMoreTokens())
{Stringtr1=defaultTokenizer.nextToken().toLowerCase()
.replaceAll("[^\\w]","");//remove special character on
word tokens Array List.add(str1);//all word add into
array list//System.out.println(str1);}return
tokensArrayList;}
```

2.2 STEMMING:

High precision IR is often difficult for a variation of rationale; one of these is a vast number of morphological variants for any specified term. To addressa couple of the problem arising from a mismatch between different word forms used in the question and the relevant documents, researchers have long proposed the use of several stemming algorithms to lessen terms to a restricted base form.

```
public String Stemming Words(String words)
{English Stemmer english = new English Stemmer();
english. Set Current( words);
english. stem();
String str=english. Get Current();
str = str.substring(0, str.length() - 1) + 'y';return words;}
```

3.CONCLUSION:

We examine the tags of each question as the representative words of the subtopic of each question. In other words, we study the user expertise under tags rather than topics. We analyses the composition and productivity of groups formed by different types of contributors. How much and how well users contribute towards the purpose.

Our in-depth data examination and forecasting experiments manifest the capability of tag-based features as well as their dominance over more obvious, previously studied non-tag based factors. We conjecture that the tags chosen by the asker’s influence how their questions are answered

ACKNOWLEDGMENT:

We feel great joy in bestowing our deepest sense of gratitude to our guide Mr Kaustubh A. Hiwarekar for his guidance during the project work, without which it would have been very difficult task. We deeply express our sincere thanks for the valuable guidance, extreme assistance and cooperation extended to all the Staff Members of the department.

This acknowledgement would not be arrant without expressing special thanks to Dr N. J. Uke Head of the Department (Information Technology) for his support during the work.

We would wish to extend heartfelt gratitude to our Principal, Dr S. D. Lokhande who provided a lot of precious support.

Last but not least we would wish to thank all the Teaching, Non- Teaching staff members of the department, our parents and colleagues those who helped us directly or indirectly for completing of this project successfully.

REFERENCES:

1) A. Rechavi and S. Rafaeli, "*Not all is gold that glitters: Response time & satisfaction rates in Yahoo! answers*." in SocialCom/PASSAT. IEEE, 2011, pp. 904–909.

2) S. Letovsky. *Cognitive processes in program comprehension*. In Proc. of the 1st Workshop on empirical studies of programmers, pages 58–79, Norwood, NJ, USA, 1986. Ablex Publishing Corp.

3) D. Avrahami and S. E. Hudson, "*Responsiveness in instant messaging: predictive models supporting inter-personal communication.*" in CHI. ACM, 2006, pp. 731–740.

4) M. R. Morris, J. Teevan, and K. Panovich, "*What do people ask their social networks, and why? a survey study of status message Q&A behavior,"* in SIGCHI, 2010.