DECEPTIVE ANSWER PREDICTION

Minal U. Rajankar, Aanchal R. Ganvir, Sakshi R.Satankar, Sneha R. Naik
Department of Information Technology, Yeshwantrao Chavan College Of Engineering,
Nagpur, India
minalrajankar32@gmail.com, sakshisatankar15@gmail.com, aanchal.ganvir@gmail.com

Abstract-In Community question answering (QA) sites, malicious users may provide deceptive answers to promote their products or services. It is important to identify and filter out these deceptive answers. On the user side, the deceptive answers are misleading to users. If the asker was cheated by the provided answers, he will not trust and visit this site again. Therefore, it is a fundamental task to predict and filter out the deceptive answers. Thus our project proposes a technique for deceptive answer prediction using sequence comparison method. In this we are going to do the verification of data from general database with expert database. Then comparative, analytically data analysis gives the user the accuracy percentage of search data. Then generating an alert remark after comparing data with expert database. In this paper data of computer language like C,C++ and Java is considered, comparison algorithm is applied and whether the searched answer is correct or not is predicted.

I. Introduction

Currently, Community QA sites, such as Yahoo! Answers and WikiAnswers, have become one of the most important information acquisition methods. In addition to the general-purpose web search engines, the Community QA sites have emerged as popular, and often effective, means of information seeking on the web. By posting questions for other participants to answer, users can obtain answers to their specific questions. The Community QA sites are growing rapidly in popularity. However, some answers may be deceptive. As the answers can guide the user's behaviour, some malicious users are motivated to give deceptive answers to promote their products or services. There are at least two major problems that the deceptive answers cause. On the user side, the deceptive answers are misleading to users. If the users rely on the deceptive answers, they will make the wrong decisions. Or even worse, the promoted link may lead to illegitimate products. On the Community QA side, the deceptive answers will hurt the health of the Community QA sites. A Community QA site without control of deceptive answers could only benefit spammers but could not help askers at all. If the asker was cheated by the provided answers, he will not trust and visit this site again. Therefore, it is a fundamental task to predict and filter out the deceptive answers. Our project proposes a technique for deceptive answer prediction using sequence comparison method. In this we are going to do the verification of data from general database with expert database . Then comparative, analytically data analysis gives the user the accuracy percentage of search data. Then generating an alert remark after comparing data with expert database. We present algorithm RAPID, to be found and assessed extremely rapidly. RAPID is a word search algorithm which uses probabilities to modify the significance1 attached to different words.

II. Basic concepts

There are five main steps used in deceptive answer prediction. They are listed below

1. Generation of two databases one is general database and other is expert database.
2. Creating the user profile login session.
3. Collecting the user query for the verification
4. By algorithm comparing the data with expert database
5. Generating the result with answer prediction

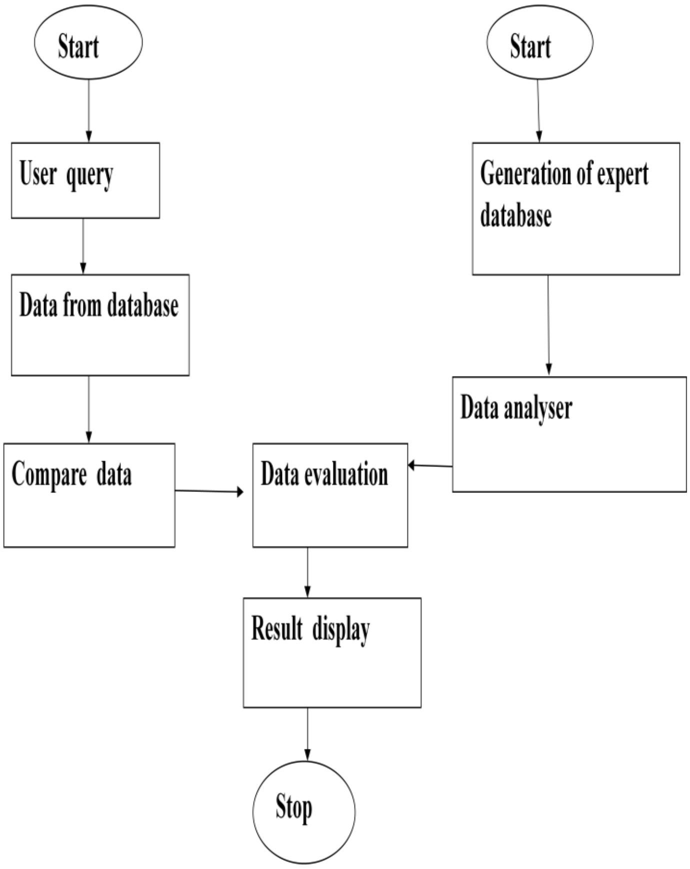

Fig 1. System Architecture

Fig 1. Shows system architecture for deceptive answer prediction. First the user will fire the query .According to it the user will get the required data from the general database. Then this data will be compared and evaluated with the data from expert database by sequence comparison Generation of two databases one is general database and other is expert database.

Expert database contains relevant information about the programming languages ie. C, C++ and JAVA. The topics covered in these languages are in the expert database. General database contains relevant as well as irrelevant information about the topics covered in the above mention programming languages. The information present in the general database will be matched with the information present in the expert database

technique. Then comparative, analytically data analysis gives the user the accuracy percentage of search data. Then generating an alert remark after comparing data with expert database.

with the help of sequence comparison technique.

Creating the user profile login session

This is the user interface module. We will create a website , user have to login in this site and user will select his course of interest. We will create a login session of each user so that different users' interest do not mix with other users. This will ease the database matching process.

Collecting the user query for the verification

This module acts as data collector. It collects the data from the user. User will fire the query ie. what information he has to search. Depending upon the query of the user , database will be searched.

By algorithm comparing the data with expert database

In this we are going to do the verification of data from general database with expert database. Then comparative, analytically data analysis gives the user the accuracy percentage of search data with the help of RAPID and sorting algorithm. RAPID is a word search algorithm which uses probabilities to modify the significance attached to different words; RAPID is a frequency matching algorithm. A sorting algorithm is an efficient algorithm which performs an important task that puts elements of a list in a certain order or arranges a collection of items into a particular order

Generating the result with answer prediction

This module will tell the user if the answer to his query is correct or not. This will generate an alert message regarding the content of the answer that the verified content is correct or not. This will tell the user regarding relevance of the answer.

III. Algorithm

RAPID compares two sequences a and b, by counting the number of words, N, occuring one or more times in a which also occur one or more times in b. This is compared to estimate E, of the number of such 'Matches' we would expect to occur by chance.

The number of matches to be expected by chance

Let W^a & W^b, sizes La & Lb, respectively, be the sets of words which occur one or more times in two sequences, a & b.

The total number of matches E between a sequences, a & an unrelated sequences b is estimated using equation

$$E= \sum_{i=0}^{La} \mathrm{LbP}\binom{W^a}{i} = \mathrm{Lb} \sum_{i=0}^{La} \mathrm{P}\binom{W^a}{i}$$

IV. Conclusion

We have discussed the deceptive answer prediction task in Community QA sites. Our project proposes a technique for deceptive answer prediction using sequence comparison method. The Community QA sites are growing rapidly in popularity. Currently there are hundreds of millions of answers and millions of questions accumulated on the Community QA sites. These resources of past questions and answers are proving to be a valuable knowledge base. From the Community QA sites, users can directly get the answers to meet some specific information need, rather than browse the list of returned documents to find the answers. Hence, in recent years, knowledge mining in Community QA sites has become a popular topic in the field of artificial intelligence. Thus our project will generate an alert message regarding the content of the answer that the verified content is correct or not. This will tell the user regarding relevance of the answer.

V. References

Papers

[1] A. Figueroa and J. Atkinson. 2011. Maximum entropy context models for ranking biographical answers to open-

domain definition questions. In Twenty-Fifth AAAI Conference on Artificial Intelligence.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, March.

[3] Daisuke Ishikawa, Tetsuya Sakai, and Noriko Kando, 2010. Overview of the NTCIR-8 Community QA Pilot Task (Part I): The Test Collection and the Task, pages 421–432. Number Part I.

[4] F. Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A. Konstan. 2008. Predictors of answer quality in online q&a sites. In Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, CHI ’08, pages 865–874, New York, NY, USA. ACM.

[5] Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein, and Hongyuan Zha. 2009. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In Proceedingsof the 18th international conference on World wide web, WWW ’09, pages 51–60, NY, USA. ACM.

[6] Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In Proceedings of the 14th ACM CIKM conference, 05, pages 84–90, NY, USA. ACM.

[7] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008. Knowledge sharing and yahoo answers: everyone knows something. In Proceedings of the 17th international conference on World Wide Web, WWW ’08, pages 665–674, New York, NY, USA. ACM.

[8] Peter F. Brown, John Cocke, Stephen A. Della Pietra,Vincent J. Della Pietra, Fredrick Jelinek, John D.Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. Comput. Linguist., 16:79–85, June.

[9] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. Journal of the American society for information science, 41(6):391–407.