© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 13 Issue: 03 | NCASIT-2022 | April 18th 2022
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed

Review Paper On Web-Base Question Answering system By Using Credibility Assessment Algorithm On The Basis Of Various categories For Improving Accuracy.

Akshata N.Kumbhare
Department Of Computer Sci & Engg.
Bapurao Deshmukh College Of Engineering
Sewagram,Wardha

Akhil D. Gotmare
Department Of Computer Sci & Engg.
Bapurao Deshmukh College Of Engineering
Sewagram,Wardha

***Abstract*— Web-based question answering (QA) systems are effective in corroborating answers from multiple Web sources. However, Web also contains false, fabricated, and biased information that can have adverse effects on the accuracy of answers in Web-based QA systems. Existing, solutions focus primarily on finding relevant Web pages but either do not evaluate Web pages' credibility or evaluate two to three out of seven credibility categories. This research proposed a credibility assessment algorithm that uses seven categories, including correctness, authority, currency, professionalism, popularity, impartiality, quality, for scoring credibility, where each credibility category consists of multiple factors. The credibility assessment module is added on top of an existing QA system to score answers .**

Keywords- Credibility assessment, information processing, Natural language processing, web credibility, question answering

I. INTRODUCTION

QA is a complex form of Information Retrieval (IR) system where the information requested is partially e xpressed in natural language statements . This makes QA systems one of the most natural ways of communicating with computers. QA is a complex process and involves multiple domains including natural language processing (NLP), IR, Information Processing (IP), and machine learning . This is a complex process, in comparison to IR, because IR considers complete documents as relevant, whereas in QA, the only specific portion(s) of te xt within documents are considered as The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney. answers . In short, a user is only interested in concise, comprehensive, and correct information from QA systems . QA systems are seeing a revival, primarily due to the popularity of Web, which is one of the largest repositories of data, making it the primary source of information for users and systems . To find relevant information on the Web, users and systems make use of search engines . QA systems making use of the Web as an information resource using search engines are called Web-based QA systems . Although search engines are increasingly efficient at identifying best sources for a given question and answers within these sources, many of the Web sources on the Web are not trustworthy because they contain erroneous, false, misleading, biased, or outdated information Some studies. laim that one in every five Web pages on the Web is fake Unfortunately, most Web users are not aware of this and simply t rust the information provided without verifying its credibility . Credibility is defined as ''the quality of being convincing or believable'' . Many researchers have defined credibility, where it depends on two or more credibility categories consisting of one or more credibility factors. Here credibility factor is referred to as a characteristic that can be used to judge the credibility of a resource and credibility category covers a certain aspect of credibility such as quality or correctness . For e xa mple, Fogg and Tseng stated that credibility is based on two categories including trustworthiness and expertise. Meola e xpanded these to five categories instead including accuracy, objectivity, authority, currency, and coverage, allowing a lot of credibility factors to be mapped onto them. Shah and Ravana e xpanded these categories further after reviewing several credibility assessment systems and identified seven credibility categories that affect credibility. These include correctness, authority, currency, professionalism, popularity, impartiality, and quality, which are used for credibility assessment of Web pages in this research. A naive solution to filtering untrustworthy answers would be to aggregate answers found on multiple Websites, which may help in eliminating typos or promoting the popular answer. However, this solution fails to consider the fact that answers e xtracted from different Web pages are not equal, as some Web pages are more credible than others . Scammers or spammers take advantage of such systems, which rely on redundancy of answers for verification, by creating multiple copies of Web pages having the incorrect answer, thus jeopardizing the outcome . Therefore, there is a need to rate Web pages based on their credibility and rank the answers accordingly. Users require assistance in reaching a conclusive and correct answer by providing and using Web pages' credibility data. In its absence, questionable Web pages can mislead naïve Web users such as elementary and high school students. This is the reason why educators consider credibility as a topic of utmost importance. Jenkins, et al. in his book terms Web search as one of the major ‘‘new media literacies’’ for students and regards credibility assessment is an essential part of the process. Though guidelines are available for evaluating Web credibility effectively such as one written by

LITERATURE REVIEW .

Web-based QA systems have been effective in providing quick answers with adequate accuracy using NLP and IR-based techniques, in comparison to semantic-based techniques. However, their accuracy is severely affected due to the amount of incorrect information on Web pages . More details on QA systems types can be found in the review paper by Gupta and Gupta . Most of the research done in Web-based QA system has been focused primarily on improving answer e xt raction techniques and intelligent answer deduction, with limited emphases given to scoring answers based on the credibility of Web sources. Therefore, the research looks into the credibility factors used in e xisting Webbased QA systems and information systems shown .

A. CREDIBILITY ASSESSMENT IN W EB-BASED QA SYSTEMS.

Work on question answering systems dates back to the mid 1960s . Improvements have been made in QA systems in terms of improving the relevancy of the documents retrieved, e xtract ing answers using IR techniques, and answer scoring but little attention has been given towards the credibility of documents retrieved. It is the case with Web-based QA systems as well, where systems are either not using credibility assessment at all or are covering credibility assessment partially. Most Web-based QA systems using NLP and IRbased techniques focused primarily on either answer e xt raction or answer scoring for improving the accuracy of answers. These systems proposed methods like the use of information e xtraction from e xterna l resources, voting procedure, and probabilistic phrase re-ranking algorithm . While these methods do improve existing modules, the accuracy of answers cannot be improved greatly without considering credibility. However, some Web-based QA systems did suggest the use of credibility assessment but covering two or three credibility categories only. Systems like Corrob, Genre QA, and Honto?search not only provide a ranked answer list, but also evaluates the credibility of Web sources based on credibility factors such as document quality, reputation, search result rank, originality, and update frequency , user's credi bility judgment for ranking the search results .

Fogg , yet users are not educated enough in the area to use it properly or the process is time-consuming. Therefore, automatic tools for evaluating credibility are becoming increasingly popular . To address the issues highlighted, the study made the following contributions: Credibility Assessment Algorithm.

.Though these systems do consider credibility factors, yet they can be enhanced much further by covering all credibility categories. Since the literature on credibility-based Web QA systems is limited; therefore, this study also reviewed Web Information Systems (IS) conducting credibility assessment on Web pages .

(B) CREDIBILITY ASSESSMENT IN WEB

IS Shah, et al. lists nine different approaches that can be used for conducting credibility assessment. These approaches show Web credibility evaluations results by either 1) estimation or 2) computer-aided support systems. Cred ibility estimation techniques assign credibility scores to Web pages based on credibility factors and rank Web pages based on the scores assigned . These systems consider factors like the structure of the Web page, content quality, content analysis, author analysis, and domain type to determine its credibility. TrustRank, PageRank, and Credible are examples of such systems that estimate credibility based on the Web page's link structure, organization, amount of spam and advertisements , Banerjee and Han proposed system also estimates Web credibility by checking the relevance of answers found in a Web page and question asked context model. System by Tanaka, et al. also provided credibility results for Web pages, where the pages are assessed based on content analysis, social support analysis, and author analysis. This system was e xpanded by mapping scores into credibility categories and predicting user's credibility judgment for ranking the search results quality, content analysis, author analysis, and domain type to determine its credibility. TrustRank, Page Rank, and Credible are examples of such systems that estimate credibility based on the Web page's link structure, organization, amount of spam and advertisements Banerjee and Han proposed system also estimates Web credibility by checking the relevance of answers found in a Web page and question asked context quality, content analysis, author analysis, and domain type to determine its credibility. TrustRank, PageRank, and Credible are examples of such systems that estimate credibility based on the Web page's link structure, organization, amount of spam and advertisements Banerjee and Han proposed system also estimates Web credibility by checking the relevance of answers found in a Web page and question asked context model. System by Tanaka, et al. also provided credibility results for Web pages, where the pages are assessed based on content analysis, social support analysis, and author analysis. This system was e xpanded by mapping scores into credibility categories and predicting

Computer-aided credibility support systems only provide valuable credibility information to Web users to assist them in conducting credibility evaluation but do not rank the m assess the correct answer . Other systems provide credibility info like on-page, off-page, aggregate features to assist in Web credibility assessment instead . However, the above-mentioned systems do not cover all seven Credibility

Conclusion

This study made two important contributions including 1) developed a prototype Web-based system incorporating a credibility assessment module called Cred OMQA system and 2) providing evaluation results showing the impact of credibility assessment on answer accuracy in Web-based QA systems. Our findings show that four out of seven categories had a significant impact on improving perCorrect and MRR results. Moreover, the CredOMQA system covered all seven credibility categories in comparison to baseline systems that only covered limited credibility categories. The introduction of credibility assessment in Web-based QA systems would allow users to have greater confidence in the answer given by the system, making them more credible and accurate. Moreover, the credibility assessment model will improve the way Web users surf the Web, improve Web publishing standards, and will apply to multiple domains such as education, medical. Some systems like WISDOM provide the distribution of positive and negative opinions relating to a topic to allow them

Assessment Support Tool (WebCAST) developed by Aggarwal, et al. is one such system that covers a wide range of credibility factors and addresses all seven categories.

References

[1] L. Hirschman and R. Gaizauskas, ‘‘Natural language question answering: The view from here,’’ Natural Lang. Eng., vol. 7, no. 4, pp. 275–300, Dec. 2001.

[2] A. McCallum, ‘‘Information e xtraction: Distilling structured data from unstructured text,’’ Queue, vol. 3, no. 9, pp. 48–57, Nov. 2005.

[3] D. Mamgai, S. Brodiya, R. Yadav, and M. Dua, ‘‘An improved automated question answering system from lecture videos,’’ in Proc. 2nd Int. Conf. Commun., Comput. Netw., vol. 2019, pp. 653–659.

[4] P. Gupta and V. Gupta, ‘‘A survey of te xt question answering techniques,’’ Int. J. Comput. Appl., vol. 53, no. 4, pp. 1–8, Sep. 2012.

[5] M. Devi and M. Dua, ‘‘ADANS: An agriculture domain question answering system using ontologies,’’ in Proc. Int. Conf. Comput., Commun. Autom. (ICCCA), May 2017, pp. 122–127.

[6] K. Purcell, ‘‘Search and email still top the list of most popular online activities,’’ in Pew Internet & American Life Project, vol. 9. Washington, DC, USA: Pew Research Center’s Internet & American Life Project, 2011. [Online]. Available: https://www.pewresearch.org/internet/wpcontent/uploads/sites/ 9/m edia/Files/Reports/2011/PIP_Search-andEmail.pdf

[7] M. Wu and A. Marian, ‘‘A framework for corroborating answers from multiple Web sources,’’ Inf. Syst., vol. 36, no. 2, pp. 431–449, Apr. 2011.

[8] A. Abbasi, F. M. Zahedi, and S. Kaza, ‘‘Detecting fake medical Web sites using recursive trust labeling,’’ ACM Trans. Inf. Syst., vol. 30, no. 4, pp. 1–36, Nov. 2012.

[9] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, and J. F. Nunamaker, ‘‘Detecting fake Websites: The contribution of statistical learning theory,’’ Mis Quart., vol. 34, no. 3, pp. 435– 461, 2010.

[10] A. Abbasi and H. Chen, ‘‘A comparison of fraud cues and classification methods for fake escrow Website detection,’’ Inf. Technol. Manage., vol. 10, nos. 2– 3, pp. 83–101, Sep. 2009.

[11] Z. Gyongyi and H. Ga rcia-Molina, ‘‘Spam: It’s ot just for inbo xes anymore,’’ Computer, vol. 38, no. 10, pp. 28–34, Oct. 2005.

[12] D. Lewandowski, ‘‘Credibility in Web search engines,’’ in Online Credibility and Digital Ethos: Evaluating ComputerMediated Communication, M. Folk and S. Apostel,
Eds. Hershey, PA, USA: IGI Global, 2013, pp. 131–146. [Online]. Available: https://www.igiglobal.com/chapter/credibility-websearchengines/72626, 10.4018/978- 1-4666-2663-8.ch008.

[13] A. A. Shah and S. D. Ravana, ‘‘Evaluating information credibility of digital content using hybrid approach,’’ Int. J. Inf. Syst. Eng., vol. 2, no. 1, pp. 92–99, 2014.

[14] M. Meola, ‘‘Chucking the checklist: A conte xtual approach to teaching undergraduates Web-site evaluation,’’ Portal: Libraries Acad., vol. 4, no. 3, pp. 331–344, 2004.

[15] H. Jenkins, R. Purushotma, M. Weigel, K. Clinton, and A. J. Robison, Confronting the Challenges of Participatory Culture: Media Education for the 21st Century. Cambridge, MA, USA: MIT Press, 2009.

[16] B. J. Fogg, ‘‘Stanford guidelines for Web credibility,’’ Res. Summary Stanford Persuasive Technol. Lab., Stanford Univ., Stanford, CA, USA, Tech. Rep., May 2002. [Online]. Available: