Ranking SERP of a Specialty Search Engine based on Probability of Correctness of Facts

Srikantaiah K C\textsuperscript{a}, Srikanth P L\textsuperscript{a}, Tejaswi V\textsuperscript{a}, Shaila K\textsuperscript{a}, K R Venugopal\textsuperscript{a} and L M Patnaik\textsuperscript{b}

\textsuperscript{a}Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore-560 001, India. Contact: srikantaiahkc@yahoo.com

\textsuperscript{b}Vice Chancellor, Defence Institute of Advanced Technology (Deemed University), Pune, India.

The World Wide Web (WWW) is the repository of large number of web pages which can be accessed via Internet by multiple users at the same time and therefore it is Ubiquitous in nature. The search engine is a key application used to search the web pages from this huge repository, which uses the link analysis for ranking the web pages without considering the facts provided by them. A new application called Probability of Correctness of Facts (PCF)-Engine is proposed to find the accuracy of the facts provided by the web pages. It uses the Probability based similarity function (SIM) which performs the string matching between the true facts and the facts of web pages to find their probability of correctness. The existing semantic search engines, may give the relevant result to the user query but may not be 100\% accurate. Our algorithm probes for the accuracy among the facts to rank the web pages. Simulation results show that our approach is efficient when compared with existing Voting and Truthfinder\cite{1} algorithms with respect to the trustworthiness of the websites.

\textbf{Keywords} : Data mining, Page Rank, Search Engine, Trustworthiness, Web Content Mining.

1. INTRODUCTION

World Wide Web (WWW) is a collection of interconnected web pages accessed via internet offers information and data from all over the world. When searching for a topic in the WWW, it returns many links or web sites related on the browser to a given topic. The important issue is to determine the website that gives the accurate information. There are many related web sites that give unauthoritative information. While the information in other repositories like books, library and journals is evaluated by scholars, publishers, and subject experts. We have no mechanism to evaluate the information on WWW. Hence, it is necessary to consider some criteria\cite{2} to evaluate the information hosted on WWW.

Web search engines are programs used to search information on the WWW and FTP servers and to check the accuracy of the data automatically. Web search engines are classified into following categories: Crawler based, Directories, Hybrid engines, Meta engines and Specialty search engines. Crawler based search engines use crawler to survey and categorize the web pages (example google.com), Directories use manual editors to survey and categorize the web pages. Human editors map web sites to specific categories in the directory database, based on the information they find and using a predefined rules, for Example yahoo directory(www.yahoo.com), and open directory (www.dmoz.org).

Hybrid search engine makes use of properties of crawler based search engine and directories i.e., use a combination of crawler based results and directory results, example modern search engines like yahoo (www.yahoo.com) and google (www.google.com). Meta search engine combines the results from all other search engine into one large listing, example meta crawler (www.metacrawler.com), and specialty search engines are used to search in specific area, such
Ranking SERP of a Specialty Search Engine based on Probability of Correctness of Facts

8. CONCLUSIONS

In this paper a new approach called PCF-Engine which uses Probability based similarity (SIM) function is proposed for resolving the conflicts between the facts provided by the different information providers in web. The Probability based similarity (SIM) function finds the implication between the facts. If the websites provides the fact which is exactly similar to that of true fact in the knowledgebase the PCF-Engine computes its trustworthiness value as 1 on a fly in a single iteration. The work can be extended by dynamically fetching the true facts to the knowledge base and removing the domain specific dependency of true facts.

REFERENCES


Srikantaiah K C is an Asst. Professor in the Department of Computer Science and Engineering at S J B Institute of Technology, Bangalore, India. He obtained his B.E and M.E degrees in Computer Science and Engineering from Bangalore University, Bangalore. He is presently pursuing his Ph.D programme in the area of Web mining in Bangalore University. His research interest is in the area of Data mining, Web mining and Semantic Web.

Srikantaiah K C, et al.,

Tejaswi V is a Student of Computer Science and Engineering from Rastrriya Vidyalaya College of Engineering, Bangalore. Her research interest is in the area of Wireless Sensor Networks.

Data mining, Web mining and Semantic Web.

Shyla K is an Asst. Professor in the Department of Electronics and Communication Engineering at Vivekananda Institute of Technology, Bangalore, India. She obtained her B.E and M.E degrees in Electronics and Communication Engineering from Bangalore University, Bangalore. She is presently pursuing her Ph.D programme in the area of Wireless Sensor Networks in Bangalore University. Her research interest is in the area of Sensor Networks, Adhoc Networks and Image Processing.

K R Venugopal is currently the Principal, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Masters degree in Computer Science and Engineering from Indian Institute of Science Bangalore. He was awarded Ph.D. in Economics from Bangalore University and Ph.D. in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored 31 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems etc.. During his three decades of service at UVCE he has over 250 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Net-
works, Parallel and Distributed Systems, Digital Signal Processing and Data Mining.

**L M Patnaik** is a Vice Chancellor, Defense Institute of Advanced Technology, Pune, India. He was a Professor since 1986 with the Department of Computer Science and Automation, Indian Institute of Science, Bangalore. During the past 35 years of his service at the Institute he has over 700 research publications in refereed International Journals and Conference Proceedings. He is a Fellow of all the four leading Science and Engineering Academies in India; Fellow of the IEEE and the Academy of Science for the Developing World. He has received twenty national and international awards; notable among them is the IEEE Technical Achievement Award for his significant contributions to High Performance Computing and Soft Computing. His areas of research interest have been Parallel and Distributed Computing, Mobile Computing, CAD for VLSI circuits, Soft Computing and Computational Neuroscience.