Analysis of Conflicting Information and a Study on Truth-Finder Algorithm

Rathod Maheshwer[1] and Ch. Kishore Kumar[2]
[1,2]Intrinity Engineering College, India

Abstract– **The world-wide web is the most important source of information for most of us. Unfortunately, if there is no guarantee for the correctness of information on the web. The information which we need for us the different web sites often provide conflicting information on a subject, such as different specifications for the same product. Our analysis provides a new problem called Veracity, truth confirmation on the information, which studies how to find true facts from a large amount of conflicting information on many subjects that is provided by various web sites. A general framework for the Veracity problem, and invent an algorithm called TruthFinder, which utilizes the relationships between web sites and their information that a web site is trustworthy if it provides many pieces of true information, and a piece of information is likely to be true if it is provided by many trustworthy web sites. Our study show that TruthFinder successfully finds true facts among conflicting information, and also identifies the disadvantages on existing trustworthy web sites better than the popular search engines.**

Keywords– Veracity, Fact, Web Mining and TruthFinder

I. INTRODUCTION

Information quality is task dependent user might consider the quality of a piece of information appropriate for one task but not sufficient for another task. Information quality is quality attribute concerned user might consider the quality of the same piece of information appropriate for both tasks. Which quality dimensions are relevant and which levels of quality are [4] required for each dimension is determined by the specific task at hand and the subjective preferences of the information consumer. Earlier researchers generated compelling list of web attributes that engender trust worthiness for example one commonly cited study has identified six features of web sites that enhance consumer perceptions of the markets trust worthiness. These web features include i) safeguard assurances, ii) marketers reputation, iii) ease of navigation, iv) robust order fulfillment, v) professionalism of the website, and 6) the use of state-of-the-art Web page design technology.

The world-wide web has become a necessary part of important information source for most people everyday people retrieve all kinds of information from the web for example when online shopping people find product specifications from web site like ShopZilla.com looking for interesting [1] DVD they get information and review on web sites such as NetFlix.com or IMDB.com. Web services are the new industrial standard for distributed computing and are considered, for the first time, a real opportunity to achieve universal interoperability. Besides enabling such interoperability, web services can also be used as communication protocols for efficient and effective business application integration. At the same time, just with any new technology, these web services also bring with them some computational complexities and business challenges. For instance, while it is easy to generate a few web services, transforming business processes into web services and harnessing the integration of many hundreds of such web services for effective application development and integration remain an open. While the web services technology has created a new industrial standard for business application integration, its role in the broader area of services computing is least understood. We need to understand the differences between service technology and services computing, its elemental form, service technology is [2] any information technology that enables a business function or process to act as a "service," which can be called up and executed on demand. Obviously, "web services" is one illustration of a service technology.

The goal of services computing, however, is the use of information technology (IT) to allow an enterprise to act like a "service provider." For example, an enterprise may want to offer its procurement functions such as procuring an item as "services on [3] demand." This will require that many business processes (e.g., select a vendor, place an order, and check shipment) to be rendered as "services." This implies that these services must be integrated dynamically to meet changing customer demands. Service Oriented Architecture is an architectural style that guides all aspects of creating and using business processes, packaged as services, throughout their lifecycle, as well as defining and provisioning the IT infrastructure that allows different applications to exchange data and participate in business processes regardless of the operating systems or programming languages underlying those applications.

II. SURVEY ON TRUTH INFORMATION

A Conformity to truth which studies how to find true facts from a large amount of conflicting information on many subjects that is provided by various websites, [5] which help us to find trustable websites and true facts. TRUTHFINDER is existing algorithm for veracity utilizes the relationships between websites and their information, if the website provides many pieces of true information is likely to be true if it is provided by many trustworthy websites is a trustworthy website. For selecting trustworthy information

the TRUTHFINDER uses two parameters website trustworthiness and fact confidence with some limitations. The initial assumption of website Trustworthiness is taken as 0.9 in all cases like popular, authoritative and untrustworthy websites. For specific queries trustworthy websites are retrieved based on single object or [6] property Ex: height, width and also recalculation of trustworthiness of websites for each query given by the user reduces the performance of the system.

The quality of the search results from the web search engines varies as information providers have levels of knowledge and intentions. Users of web based systems are therefore confronted with the increasingly difficult task of selecting high quality information from the vast amount of web accessible information.

A. Overview of Web Site Information Filter System

The WebSIFT system is designed to perform usage mining from the serverlogs in the extended NSCA format. Preprocessing algorithms include identifying users' server sessions and inferring cached page references through the use of the referrer field. WebSIFT system performs content and structure preprocessing and provides the option to convert server sessions. The server session files can be run through sequential pattern analysis.

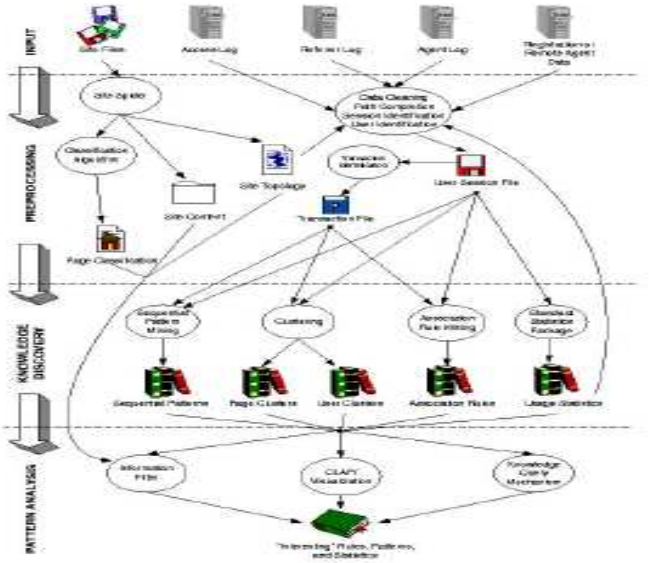

Fig. 1: Mining web sites

Input of the mining process includes three server logs access, referrer and agent the HTML files that make up the site and the optional data such as registration files remote agent logs. In the preprocessing process the input data is used to construct a user session file to derive a site topology and to classify the pages of a site. The user session file will be converted to the transaction file and output to next phase – pattern discovery. Both the site, topology and page classifications are fed into the information filter which belongs to the pattern analysis process and makes use of the preprocessed content and structure information to automatically [7] filter the results of the knowledge discovery algorithms for patterns that are potentially interesting.

III. PROBLEM DOMAIN

Information provided by different websites usually have multiple conflicting facts from different websites for each object. A user unable to find out the correct information certain query. To avoid the multiple conflicts for each object we design an algorithm TRUTHFINDER. The input of TRUTHFINDER is a large number of facts are provided by many websites, the goal of TRUTHFINDER is to identify the true fact among them.

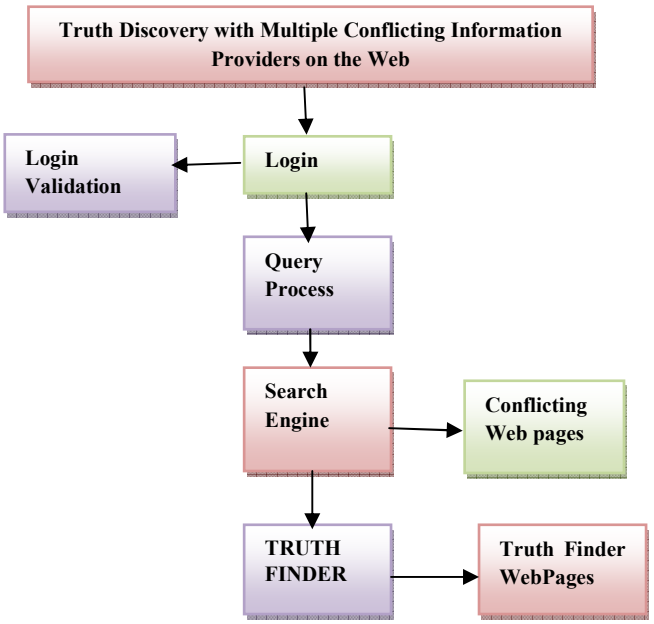

Fig. 2: Framework of TRUTHFINDER

A. Analysis Conflicting Information

In this framework problem proposes the Veracity that is Given a large amount of conflicting information about many objects which is provided by multiple web site or other types of information providers discovers the true fact about each object. The word fact to represent something that is claimed as a fact by some web site and such a fact can be either true or false.

Fig. 3: Analysis of the TruthFinder

Conflicting facts on the web such as different sets of authors for a book, many web sites some of which are more trustworthy than some others. A fact is likely to be true if it is provided by trustworthy websites.

Exact common sense and observations on real data we have four heuristics that serve as:

- Usually there is only one true fact for a property of an object
- This true fact appears to be the same or similar of an object
- The false facts on different web sites are less likely to be the same or similar

In a certain domain, a web site that provides mostly true facts for many objects will likely provide true facts for other objects.

B. Algorithm for TruthFinder

Algorithm 1: TRUTHFINDER

Input: The set of web sites W, the set of facts F, and links between them.

Output: Web site trustworthiness and fact confidence.

Calculate matrices **A** and **B**

for each $w \in W$ /* *setting initial state* */

$t(w) \leftarrow t_0$

$\tau(w) \leftarrow -\ln(1 - t(w))$

repeat /* *iterative computation* */

$\vec{\sigma} \leftarrow \mathbf{B}\vec{\tau}$

compute $\vec{s}$ from $\vec{\sigma}$

$\vec{t}' \leftarrow \vec{t}$ /* *make a copy of* $\vec{t}$ */

$\vec{t} \leftarrow \mathbf{A}\vec{s}$

compute $\vec{\tau}$ from $\vec{t}$

until cosine similarity of $\vec{t}$ and $\vec{t}'$ is greater than $1 - \delta$

IV. EXPECTED REAL DATASET RESULTS

A real dataset which shows the effectiveness of TruthFinder we compare it with approach called voting which chooses the fact that is provided by most websites. TruthFinder with Google by comparing the top websites found by each of the them experiments are performed on an Intel PC with a 1.66GHz dual-core processor 1 GB memory running Windows XP professional, implemented in Visual Studio.net (C#).

Dataset contains the authors of many books provided by many online bookstores it contains 1265 computer science books published by Addison Wesley, McGraw Hill, Morgan Kaufmann, or Prentice Hall. For each book we use ISBN to search on *www.abebooks.com,* which returns the book information on different online bookstores that sell this book. The dataset contains 894 bookstores, and 34031 listings (i.e., bookstore selling a book). On average each book has 5.4 different sets of authors. TruthFinder performs iterative computation to find out the set of authors for each book. In order to test its accuracy, we randomly select 100 books and manually find out their authors. Here find the image of each book, and use the authors on the book cover as the standard fact. Compare the set of authors found by TruthFinder with the standard fact to compute the accuracy. For a certain book, suppose the standard fact contains x authors, TruthFinder indicates there are y authors, among which z authors belong to the standard fact. The accuracy of TruthFinder is defined as z max (x,y). Sometimes, TruthFinder provides partially correct facts. The accuracy of the TRUTHFINDER is 1.

A. Comparative Study

Websites often provide conflicting information, suppose a user is interested to know the height of Mount Everest among the top 20 results he or she will find the facts say 29,035 feet, five websites say 29,028 feet one says 29,002 feet and another one says 29,017feet. To avoid this veracity problem in the existing system provides the Page Rank and Authority-Hub analysis is to utilize the hyperlinks to find pages with high a authority which identifies important web pages that users are interested in that unfortunately the popularity of web pages does not necessarily lead to accuracy of information. We have more disadvantages such as popularity of web pages does not necessarily lead to accuracy of information. Even the most popular website may contain many errors, where as some comparatively not-so-popular websites may provide more accurate information. Compare to existing system our proposed system formulate the veracity problem about how to discover true facts from conflicting information and framework to solve the problem by identifying the trustworthiness of websites, confidence of facts and influences between facts. Finally, TRUTHFINDER algorithm identifying true facts using iterative methods, which achieves high accuracy on finding both true facts and high quality websites.

V. CONCLUSION

Our proposed system formulate the Veracity problem, which aims at resolving conflicting facts from multiple web sites, and finding the true facts among them and this approach that utilizes the inter- dependency between web site trustworthiness and fact confidence to find trustable web sites and true facts. Comparative study show that TruthFinder achieves high accuracy at finding true facts and at the same time identifies web sites that provide more accurate information. In future we our on real time data which shows all the best results for an every search object for complex database. Our system also provides the best featured multiplayer games, funny videos, crazy tags, photos and protection from malwares, attacks, unofficial information.

REFERENCES

[1] Anonymous, Net users distrust corporate privacy policies-study, available at http://www.newsbytes.com/cgi-bin/udt/im..ble?client.id newsbytesandstory.id – 174596

[2] Ba, S., Establishing online trust through a community responsibility system. Decision Support Systems, Vol. 31, pp. 323–336, Feb 2002.

[3] A. Borodin, G. Roberts, J. Rosenthal, P. Tsaparas. Link analysis ranking: Algorithms, theory, and experiments. ACM Transactions on Internet Technology, Vol. 5(1), pp. 231–297, 2005.

[4] Brown, G., Kanungo, T., Carey, M., Kumar, A., Tanniru, M., & Zhao, J. L., Services Science: Services Innovation Research and Education, Proceedings of the IEEE International Conference on Services Computing, July 11–15, Orlando, Florida, 2005.

[5] Fang Liu, Clement Yu, Weiyi Meng, Personalized Web Search for Improving Retrieval Effectiveness, IEEE transactions on knowledge and data engineering, Vol. 16 (1), January 2004.

[6] Xiaoxin Yin, Jiawei Han, Yu, P.S, Truth Discovery With Multiple Conflicting Information Providers On the Web, IEEE Transactions on Knowledge and Data Engineering, Vol. 20(6), pp. 796-808, June 2008.

[7] Christian Bizera, Richard Cyganiakb, Quality-driven information filtering using the WIQA policy framework,

[8] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the World Wide Web. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997.

Rathod Maheshwer B.Tech from Jyothismathi Engineering College, Karimnagar, Master of Business in HR from Kakatiya University, M.Tech Software Engg. from Ramappa Engineering College, Warangal. Currently, he is working as Asst. Prof. in Intrinity Engineering College. His research includes Web Technology, Databases Information Security and Artificial Neural Networks.

Ch. Kishore Kumar Master of Computer Application from Jayamukhi Institute of Technology & Science, M.Tech CSE from Nishitha Engineering College. He is currently working as Asst. Prof. in Intrinity Engineering College. His research areas include Databases Object Oriented Analysis & Design Web Technology and Information Security.