ISSN:0975-9646

Mining Accurate Information on Web Using Truth Algorithm

K.Vijaya Lakshmi , B.Jayanag, Dr. V.Srinivasa Rao

Dept. Computer Science & Engineering, V.R Siddhartha Engineering college, Kanuru, Vijayawada-520 007,

Abstract **-The World Wide Web has become the most important information source for most of us. Unfortunately, there is no guarantee for the correctness of information on the Web. Moreover, different websites often provide conflicting information on a subject, such as different specifications for the same product.**

In this project , we propose a new problem, called Veracity, i.e., 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. We design an algorithm, called Truth algorithm, which utilizes the relationships between websites and their information, i.e., a website 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 websites. It is used to infer the trustworthiness of websites and the correctness of information by using the Truth algorithm iteratively. Truth algorithm successfully finds true facts among conflicting information and identifies trustworthy websites better than the popular search engines.

1. INTRODUCTION

The World Wide Web has become a necessary part of our lives and might have become the most important information source for most people. Everyday, people retrieve all kinds of information from the Web. For example, when shopping online, people find product specifications from websites like Amazon.com or ShopZilla.com. When looking for interesting DVDs, they get information and read movie reviews on websites such as NetFlix.com or IMDB.com. When they want to know the answer to a certain question, they go to Ask.com or Google.com. “Is the World Wide Web always trustable?” Unfortunately, the answer is “no.” There is no guarantee for the correctness of information on the Web. Even worse, different websites often provide conflicting information, as shown in the following examples.

Example (Height of Mount Everest). Suppose a user is interested in how high Mount Everest is and queries Ask.com with “What is the height of Mount Everest?” Among the top 20 results, 1 he or she will find the following facts: four websites (including Ask.com itself) say 29,035 feet, five websites say 29,028 feet, one says29, 002 feet, and another one says 29,017 feet. Which answer should the user trust?

The trustworthiness problem of the Web has been realized by today’s Internet users. According to a survey on the credibility of websites conducted by Princeton Survey Research in 2005, 54 percent of Internet users trust news websites at least most of time, while this ratio is only 26 percent for websites that offer products for sale and is merely 12 percent for blogs. There have been many studies on ranking web pages according to authority (or popularity) based on hyperlinks. The most influential studies are Authority-Hub analysis, and Page Rank, which lead to search engines. However, does authority lead to accuracy of information? The answer is unfortunately no. Top-ranked websites are usually the most popular ones.

In this project, we propose a new problem called the Veracity problem, which is formulated as follows: Given a large amount of conflicting information about many objects, which is provided by multiple websites (or other types of information providers), how can we discover the true fact about each object? We use the word “fact” to represent something that is claimed as a fact by some website, and such a fact can be either true or false. In this project , we only study the facts that are either properties of objects (e.g., weights of laptop computers) or relationships between two objects (e.g., authors of books). A fact is likely to be true if it is provided by trustworthy websites.

A website is trustworthy if most facts it provides are true. Because of this interdependency between facts and websites, we choose an iterative computational method. At each iteration, the probabilities of facts being true and the trustworthiness of websites is inferred from each other. This iterative procedure is rather different from Authority-Hub analysis. Thus, we cannot compute the trustworthiness of a website by adding up the weights of its facts as in, nor can we compute the probability of a fact being true by adding up the trustworthiness of websites providing it.

2. NEED AND IMPORTANCE OF PROJECT PROBLEM

2.1 *Existing system*

Page Rank and Authority-Hub analysis to utilize the hyperlinks to find pages with high authorities.These two approaches identifying important web pages that users are interested in, Unfortunately, the popularity of web pages does not necessarily lead to accuracy of information

2.2 *Disadvantages of existing System*

The 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.

2.3 *Proposed System*

We formulate the Veracity problem about how to discover true facts from conflicting information. Second, we propose a framework to solve this problem, by defining the trustworthiness of websites, confidence of facts, and influences between facts. Finally, we propose an algorithm called Truth algorithm for identifying true facts using iterative methods.

2.4 *Advantages of Proposed System*

Our experiments show that Truth algorithm achieves very high accuracy in discovering true facts. It can select better trustworthy websites than authority-based search engines such as Google

3. OBJECTIVE

We formulate the Veracity problem about how to discover true facts from conflicting information.Second, we propose a framework to solve this problem, by defining the trustworthiness of websites, confidence of facts, and influences between facts. Finally, we propose an algorithm called Truth Algorithm for identifying true facts using iterative methods.

4.METHODOLOGY

System Architecture

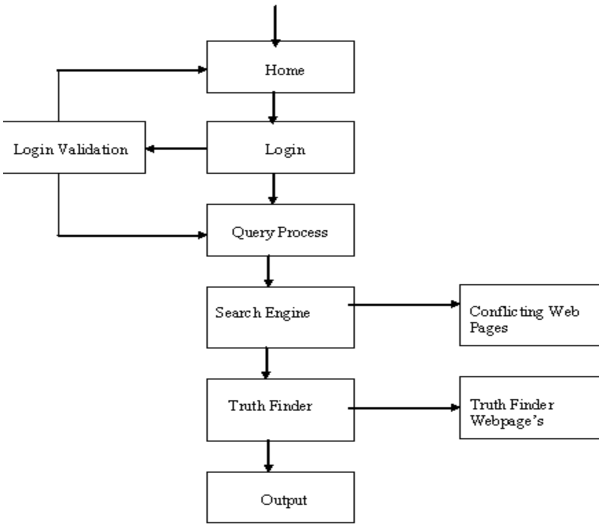

Figure 1 : System Architecture

The user logins into the system with the user name and password. Then he can search for a string. The system will search for that word from all the websites stored in the data base. The user can search for the string in the database by three ways.

When user can search for the string in normal mode all the websites which containing that word will be displayed.

When the user searches the string in page ranking mode then all the websites that contain the search string will be displayed in the descending order of their page ranking. The website which has more number of visitors will be displayed first.

When the user searches for the string in truth algorithm mode then the website which has more trustworthiness will be displayed first. All the websites will be displayed in the decreasing order of the trustworthiness.

Modules:

- *Collection of data*

  First we have to collect the specific data about an object and it is stored in related database. Create table for specific object and store the facts about a particular object.

- *Data search*

  Searching the related data link according to user input. In this module user retrieve the specific data about an object.

- *Truth algorithm*

  We design a general framework for the Veracity problem, and invent an algorithm called Truth Finder, which utilizes the relationships between web sites and their information, i.e., 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.

5. INPUT PARAMETERS

Initially we store all the websites and the facts that are stored in the website in the database.

We take all the websites and the corresponding facts in the database. The input diagram of the system is shown in the figure2.

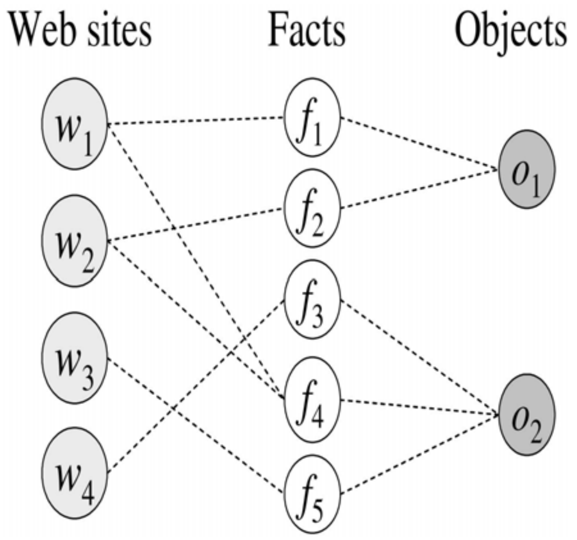

Figure 2. Input of Truth algorithm

The database contains the 5 fields. They are about, description, filename, rank and the trust. The about field contains the keyword of the website and the description will contain the information that is stored in the website. Filename conatins the name of the website. Rank will be numeric attribute which will be the rank of the website. The rank of the website will be determined by the number of visitors. The trust field will be the numeric attribute which ranges from 0 to 1. The website which has stored more fact information will have more trustworthiness. The website which has less facts stored in it will have low trustworthiness.

The Result will be the table which contains the website and trustworthiness. The trustworthiness will range from 0 to 1. All the websites which are stored in the database will be displayed in the result table.

The result table will be displayed in the graph . on the x - axis , websites will be displayed and the on the y-axis the trustworthiness will be displayed

6.CONCLUSION

In this paper, we introduce and formulate the Veracity problem, which aims at resolving conflicting facts from multiple websites and finding the true facts among them. We propose Truth algorithm, an approach that utilizes the inter dependency between website trustworthiness and fact confidence to find trustable websites and true facts. Experiments show that Truth algorithm achieves high accuracy at finding true facts and at the same time identifies websites that provide more accurate information.

.

Table 1: Database

About	Description	Filename	Rank	Trust
c	Desc1	Website1	3	0.5
C++	DESC2	Website2	2	0.8
java	desc	Website3	1	0.4
.net	description	Website4	4	0.99
c	Desc3	Website5	8	0.99
C++	Desc5	Website6	5	0.98
c	desc7	Website7	6	0.8
java	desc	Website8	7	0.7

Table 2: Result

Website	trustworthiness
Website1	0.5
Website2	0.8
Website3	0.4
Website4	0.99
Website5	0.99
Website6	0.98
Website7	0.8
Website8	0.7

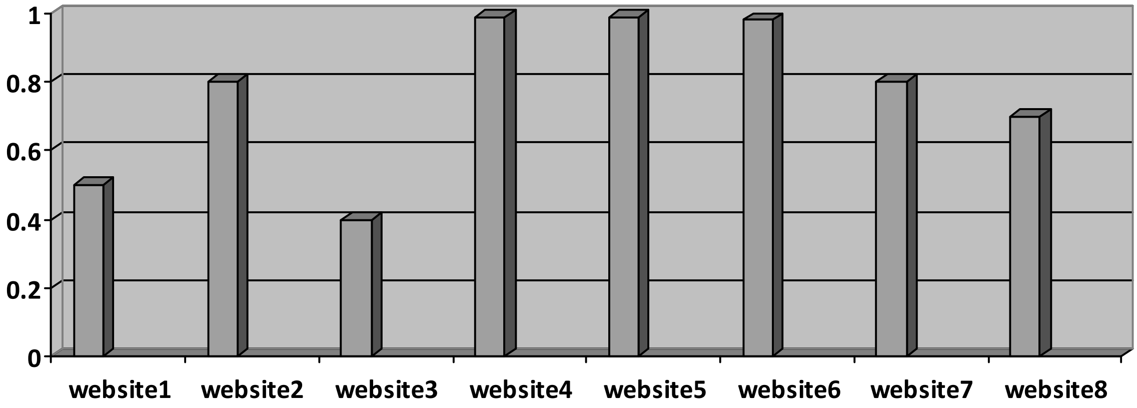

Figure 3: Output Graph

REFERENCES

[1] B. Amento, L.G. Terveen, and W.C. Hill, “Does ‘Authority’ Mean Quality? Predicting Expert Quality Ratings of Web Documents,” Proc. ACM SIGIR ’00, July 2000.

[2] M. Blaze, J. Feigenbaum, and J. Lacy, “Decentralized Trust Management,” Proc. IEEE Symp. Security and Privacy (ISSP ’96), May 1996.

[3] A. Borodin, G.O. Roberts, J.S. Rosenthal, and P. Tsaparas, “Link Analysis Ranking: Algorithms, Theory, and Experiments,” ACM Trans. Internet Technology, vol. 5, no. 1, pp. 231-297, 2005.

[4] J.S. Breese, D. Heckerman, and C. Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering,” technical report, Microsoft Research, 1998.

[5] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, “Propagation of Trust and Distrust,” Proc. 13th Int’l Conf. World Wide Web (WWW), 2004.

[6] G. Jeh and J. Widom, “SimRank: A Measure of Structural-Context Similarity,” Proc. ACM SIGKDD ’02, July 2002.

[7] J.M. Kleinberg, “Authoritative Sources in a Hyperlinked Environment,” J. ACM, vol. 46, no. 5, pp. 604-632, 1999.

[8] Logistical Equation from Wolfram MathWorld, http:// mathworld.wolfram.com/LogisticEquation.html, 2008.

[9] T. Mandl, “Implementation and Evaluation of a Quality-Based Search Engine,” Proc. 17th ACM Conf. Hypertext and Hypermedia, Aug. 2006.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web,” technical report,Stanford Digital Library Technologies Project, 1998.