Improving Trustworthiness Of Websites Using Truth Finder Algorithm

Mrs.G.Kirubasri [#1], Mrs.A.Sathya Sofia [*2]

[#] *Assistant Professor,Department of Computer Science & Engineering*
PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.
[1]kiruba.me2010@gmail.com

[*] *Assistant Professor,Department of Computer Science & Engineering*
PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India.
[2]sathyasofia@gmail.com

Abstract — **The World Wide Web has become the most important information source for most of us. It provides hundreds or thousands of relevant documents of widely varying quality products. Unfortunately there is no guarantee for the correctness of information on web. Many websites often provide conflicting information on a subject, such as different specifications for the same product. The proposed system introduces 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. The veracity problem is solved by an algorithm, called TRUTHFINDER, which utilizes the relationships between websites and their information. An iterative method is used to infer the trustworthiness of websites and the correctness of information from each other. The proposed system shows that the TRUTHFINDER successfully find true facts among conflicting information and identifies trustworthy websites better than the existing methods Normal search and Page Rank.**

Keywords — **Data quality, Web mining, Link Analysis, Confidence of fact, Trustworthiness of Websites.**

I. INTRODUCTION

The World Wide Web is continuously growing and "collecting" all kind of resources. It has become a necessary part of our lives and it provides most important information source for most people. People find product specifications from websites. But there is not guarantee for the correctness of information. It gives different specifications for same objects, as shown in the following example.

Example 1(Authors of books). We tried to find out who wrote the book Rapid Contextual Design (ISBN: 0123540518). We found many different sets of authors from different online bookstores, and we show several of them in Table 1. From the image of the book cover, we found that A1 Books provides the most accurate information. In comparison, the information from Powell's books is incomplete, and that from Lakeside books is incorrect.

The trustworthiness problem of the Web has been realized by today's Internet users. The Existing system uses Page Rank [11] and Authority-Hub analysis [9] to utilize the hyperlinks to find pages with high authorities. These two approaches identifying important popular web pages that user are interested in. However, popularity does not mean accuracy. Noble and Powell's books) contain many errors on information. In comparison, some small bookstores (e.g., A1 Books) provide more accurate information.

In this paper, proposed a new problem called the Veracity problem, which is a large amount of conflicting information about many objects on multiple websites. To 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. The facts that are either properties of objects or relationships between two objects.

For example, according to this experiment the bookstores ranked on top by Google (Barnes & Noble and Powell's books) contain many errors on information. In comparison, some small bookstores (e.g., A1 Books) provide more accurate information.

TABLE I

Conflicting Information about Book Authors

Web site	Authors
A1 Books	Karen Holtzblatt, Jessamyn Burns Wendell, Shelly Wood
Powell's books	Holtzblatt, Karen
Cornwall books	Holtzblatt- Karen, Wendell -Jessamyn Burns, Wood
Mellon's books	Wendell, Jessamyn
Lakeside books	Wendell, Jessamynholtzblatt, Karen wood, Shelly
Blackwell online	Wendell, Jessamyn, Holtzblatt, Karen, wood, Shelly
Barnes & Noble	Karen Holtzblatt, Jessamyn Wendell, Shelly Wood

In this paper, proposed a new problem called the Veracity problem, which is a large amount of conflicting information about many objects on multiple websites. To 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. The facts that are either properties of objects or relationships between two objects.

A fact is likely to be true if it is provided by trustworthy websites. A trustworthy website provides only true facts. The iterative computational method has been chosen Because of interdependency between facts and websites. In every iteration, the trustworthiness of websites is inferred from each other.

There are three major distributions in this paper. Normal search, Page Rank Search, Truth Finder search and Performance Analysis of Truth Finder in comparison with above two searches.In order to evaluate the performance among these search methods, First, formulating the Veracity problem based on how to discover true facts from conflicting information. Second, propose a framework to solve this problem, by defining the trustworthiness of websites, confidence of facts, and influences between facts. Finally, Use an algorithm called TRUTHFINDER for discovering true facts using iterative methods. TRUTHFINDER gets very high accuracy in identifying true facts, and it can identified better trustworthy websites than NORMAL and PAGE RANK search.

Fig.1 shows the overall system design and three methods for retrieving information. They are Normal Search, Page Rank Search, and TRUTHFINDER search where input of each method is search object. The first method Normal search provides all the available conflicting information that matching with the user search query. The second method Page Rank identifies the pages with high usage based on how many times the user visiting that WebPages. The Proposed method TRUTHFINDER retrieves trustable WebPages with out conflicting information

The rest of the paper is organized as follows: The problem is described in section 2 and the basic heuristics, computational model in section 3 and 4.Algorithm is presented in section 5.Experimental settings in section 6 and conclusion in section 7.

II. PROBLEM DEFINITIONS

This paper describes the problem of finding true facts in a certain domain. Here, a domain refers to a property of a certain type of objects, such as authors of books or number of pixels of camcorders. The input of TRUTHFINDER is a large number of facts in a domain that are provided by many websites. There are usually multiple conflicting facts from different websites for each object, and the goal of TRUTHFINDER is to identify the true fact among them.

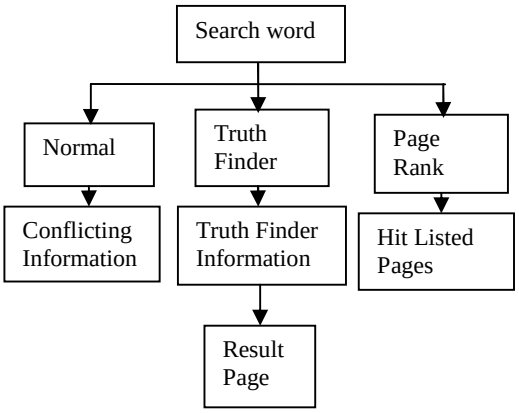

Fig. 1. System Design

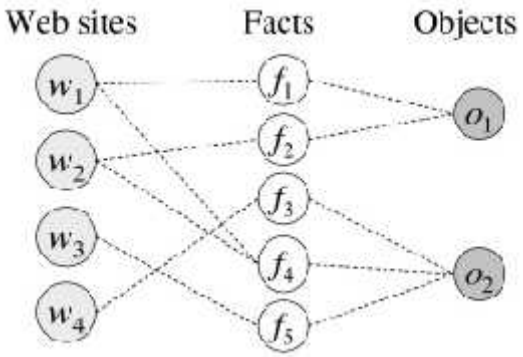

Fig. 2. Input of TRUTHFINDER.

Fig.2 shows a mini example data set, which contains five facts about two objects provided by four websites. Each website provides at most one fact for an object.

A. Basic Definitions

The two most important definitions in this paper are the confidence of facts and the trustworthiness of websites.

1). Confidence of facts: The confidence of a fact f (denoted by s (f)) is the Probability of f being correct, According to the best of our knowledge.

2). Trustworthiness of websites: The trustworthiness of a website w (denoted by t(w)) is the expected Confidence of the facts provided by w.

Different facts about the same object may be conflicting. However, sometimes facts may be supportive to each other although they are slightly different. For example "Jennifer Widom," and "J. Widom," If one of such facts is true, the other is also likely to be true.

B. *Concept of Implication between Facts.*

In order to represent such relationships between the facts, the proposed concept of implication between facts. The implication from fact f1 to f2, impf1 → f2), is f1's influence on f2's confidence. The f2's confidence should be increased (or decreased) according to f1's confidence. It is required that imp(f1→ f2) is a value between -1 and 1.A positive value indicates that if f1 is correct, f2 is likely to be correct. While a negative value means that if f1 is correct, f2 is likely to be wrong.

When a user uses TRUTHFINDER on a certain domain, he or she should provide the definition of implication between facts. If in a domain, the relationship between two facts is symmetric and the definition of similarity is available, the user can define imp(f1 → f2) = sim(f1, f2)- base_ sim, where sim(f1, f2) is the similarity between f1 and f2, and base sim is a threshold for similarity.

III. BASIC HEURISTICS

Based on observations on real data, we have four basic heuristics that serve as the base of computational model.

i. Usually there is only one true fact for a property of an object.
ii. This true fact appears to be the same or similar on different web sites
iii. The false facts on different web sites are less likely to be the same or similar.
iv. In a certain domain, a web site that provides mostly true facts for many objects will likely provide true facts for other objects.

IV. COMPUTATIONAL MODEL

This section, introduce the model of iterative computation. Table II shows the variables and parameters used in the following discussion.

A. *Website Trustworthiness and Fact Confidence*

The inference of website trustworthiness is rather simple, whereas that of fact confidence is more complicated.

1) ***The trustworthiness of a website*****:** It is just the expected confidence of facts it provides. For website w, we compute its trustworthiness t(w) by calculating the

Average confidence of facts provided by w:

$$t(w) = \frac{\Sigma f \in F(w) s(f)}{F(w)} \text{ -------- (1)}$$

Where F (w) is the set of facts provided by w.

2) ***The confidence of a fact f: s(f)*** **:** One minus the probability that all websites providing f are wrong

$$S(f)=1-\prod_{w \epsilon W(f)}(1-t(w)) \quad \text{----------} \quad (2)$$

In (2), 1-t(w) is usually quite small, and multiplying many of them may lead to underflow. In order to facilitate computation and veracity exploration, we use a logarithm and define the trustworthiness score of a website.

TABLE II
Variables and Parameters of Truth Finder

Name	Description
M	Number of web sites
N	Number of facts
w	A web site
t(w)	The trustworthiness of w
τ (w)	The trustworthiness score of w
F(w)	The set of facts provided by
f	A fact
s(f)	The confidence of f
σ (f)	The confidence score of f
σ *(f)	The adjusted confidence score of f
W(f)	The set of web sites providing f
o(f)	The object that f is about
imp(f1→ f2)	Implication from f1 to f2
P	Weight of objects about the same object
γ	Dampening factor
δ	Max difference between two iterations

3) Trustworthiness score:

$$\tau(w) = -\ln(1-t(w)) \quad \text{----------} \quad (3)$$

τ(w) is between and zero and +∞, and a larger τ(w) indicate higher trustworthiness.

Similarly confidence score of fact as

$$\sigma(f) = -\ln(1-s(f)) \quad \text{----------} \quad (4)$$

A very useful property is that the confidence score of a fact f is just the sum of the trustworthiness scores of websites providing f.

V. ITERATIVE COMPUTATION

Fig. 3 shows the algorithm of TRUTHFINDER And it makes use of iterative procedure .It studies the probabilities of websites being correct and facts being true, which cannot be defined as simple summations because the probability often needs to be computed in nonlinear ways. That is why TRUTHFINDER requires iterative computation to achieve convergence.

In each step of the iterative procedure, TRUTHFINDER first uses the website trustworthiness to compute the fact confidence and then recomputed the website trustworthiness from the fact confidence. The matrices are stored in sparse

formats, and the computational cost TRUTHFINDER stops iterating when it reaches a stable state. The stableness is measured by how much the trustworthiness of websites changes between iterations. If only changes a little after an iteration, then TRUTHFINDER will stop.

Algorithm 1: Truth Finder
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 */*
$t(w) \leftarrow t0$
$\tau(w) \leftarrow -\ln(1-t(w))$
repeat */* iterative computation */*
$\sigma^* \leftarrow \boldsymbol{B}\tau$
Compute s from σ^*
$t' \leftarrow t$ */* make a copy of t */*
$t \leftarrow \boldsymbol{A}\, s$
Compute τ *from t*
until *cosine similarity of t and t' is greater than* $1-\delta$

Fig.3. Algorithm of TRUTHFINDER.

VI.EXPERIMENT SETTING

In order to show the effectiveness of TRUTHFINDER, we compare it with a baseline approaches called VOTING and text comparison which is used in page Rank search and normal data search. When trying to find the true fact for a certain object, VOTING chooses the fact that is provided by most websites and resolves ties randomly. It only uses the number of websites supporting each fact. In comparison, TRUTHFINDER considers the implication between different facts from the first iteration and considers the different trustworthiness of different websites.

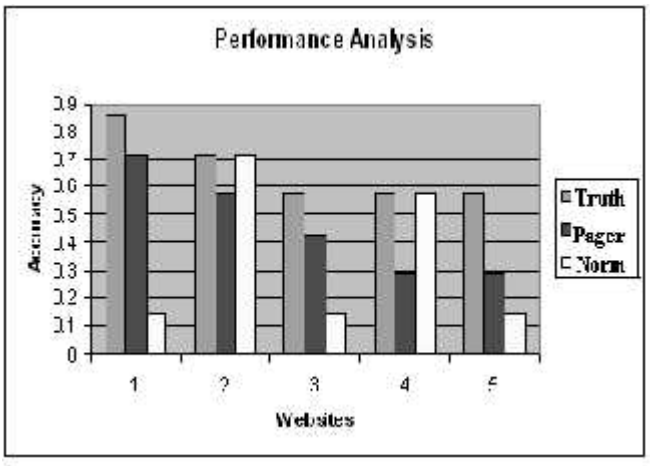

Fig.4. Accuracies of NORMAL, PAGE RANK and TRUTHFINDER SEARCH

Fig.4 shows that the accuracies of TRUTHFINDER is more than Normal and Page Rank even at the first iteration, where all bookstores have the same trustworthiness. This is because TRUTHFINDER considers the implications between different facts about the same object, while VOTING does not.

As TRUTHFINDER repeatedly computes the trustworthiness of bookstores and the confidence of facts, its accuracy increases at each iteration and remains stable. It takes TRUTHFINDER 8.73 seconds to precompute the implications between related facts and 4.43 seconds to finish the four iterations. VOTING takes 1.22 seconds.

VII.CONCLUTION

This paper introduces and formulates the Veracity problem, which aims at resolving conflicting facts from multiple websites and finding the true facts among them. The proposed system uses TRUTHFINDER, an approach that utilizes the interdependency between website trustworthiness and fact confidence to find trustable websites and true facts. In each iteration, the approach will improve the current state by propagating information (Weights, probability, trustworthiness, etc.) through the links. This iterative procedure has been proven to be successful in many applications, and thus, adopt it in TRUTHFINDER. The TRUTHFINDER is used to achieve high accuracy at finding true facts and at the same time identifies websites that provide more accurate information. Performance analysis has been made between these three approaches and the system is proved TRUTHFINDER provides more accuracy than Normal, Page Rank Searching.

In future, this system can be enhanced into broader application scope, such as mass collaboration and mass labelling. It combines the broad coverage system using unstructured query and low coverage system which is using structured query together. This integrated system can be utilized by the TRUTHFINDER algorithm to find true facts in larger systems. The system is working with text files it can be enhanced for retrieving information from html files with further modification.

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.

[11] Princeton Survey Research Associates International, *"Leap of faith: Using the Internet Despite the Dangers,"* Results of a Nat'l Survey of Internet Users for Consumer Reports WebWatch, Oct. 2005.

[12] *Sigmoid Function from Wolfram MathWorld*, http://mathworld.wolfram.com/ SigmoidFunction.html, 2008.

[13] R.Y. Wang and D.M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," J. Management Information Systems, vol. 12, no. 4, pp. 5-34, 1997.

[14] Xiaoxin Yin,Jiawei Han, senior Member, IEEE and Philip S.Yu, Fellow, IEEE. *"Truth Discovery with Multiple Conflicting Information Providers on the Web"*. IEEE Transactions on knowledge and DataMining, Vol.20, No.6, June 2008.