ARTICLE

Truth discovery with multiple conflicting information providers on the web

XIAOXIN YIN, University of Illinois Urbana-Champaign, Urbana, IL, United States

J. HAN, University of Illinois Urbana-Champaign, Urbana, IL, United States

PHILIPS S YU, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, United States

Truth Discovery with Multiple Conflicting Information Providers on the Web[*]

Xiaoxin Yin[*]
UIUC
xyin1@uiuc.edu

Jiawei Han
UIUC
hanj@cs.uiuc.edu

Philip S. Yu
IBM T. J. Watson Res. Center
psyu@us.ibm.com

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 web sites often provide conflicting information on a subject, such as different specifications for the same product. In this paper 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 web sites. We design a general framework for the Veracity problem, and invent an algorithm called TRUTHFINDER, 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.* Our experiments show that TRUTHFINDER successfully finds true facts among conflicting information, and identifies trustworthy web sites better than the popular search engines.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database applications – *Data mining*

General Terms: Algorithms.

Keywords: data quality, web mining, link analysis.

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 web sites like Amazon.com or ShopZilla.com. When looking for inter-

esting DVDs, they get information and read movie reviews on web sites such as NetFlix.com or IMDB.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 web sites often provide conflicting information, as shown below.

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.

Web site	Authors
A1 Books	Karen Holtzblatt, Jessamyn Burns Wendell, Shelley Wood
Powell's books	Holtzblatt, Karen
Cornwall books	Holtzblatt-Karen, Wendell-Jessamyn Burns, Wood
Mellon's books	Wendell, Jessamyn
Lakeside books	WENDELL, JESSAMYNHOLTZBLATT, KARENWOOD, SHELLEY
Blackwell online	Wendell, Jessamyn, Holtzblatt, Karen, Wood, Shelley
Barnes & Noble	Karen Holtzblatt, Jessamyn Wendell, Shelley Wood

Table 1: Conflicting information about book authors

The trustworthiness problem of the web has been realized by today's Internet users. According to a survey on credibility of web sites, 54% of Internet users trust news web sites at least most of time, while this ratio is only 26% for web sites that sell products, and is merely 12% for blogs.

There have been many studies on ranking web pages according to authority based on hyperlinks, such as Authority-Hub analysis [2], PageRank [4], and more general link-based analysis [1]. But does authority or popularity of web sites lead to accuracy of information? The answer is unfortunately no. For example, according to our experiments the bookstores ranked on top by Google (*Barnes & Noble* and *Powell's books*) contain many errors on book author information, and some small bookstores (*e.g.*, *A1 Books*) provide more accurate information.

In this paper we propose a new problem called *Veracity* problem, which is formulated as follows: Given a large amount of conflicting information about many objects, which is provided by multiple web sites (or other types of information providers), how to discover the true fact about each object. We use the word *"fact"* to represent something that
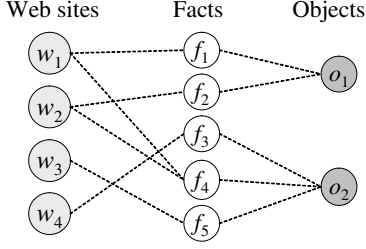
Figure 1: Input of TRUTHFINDER

is claimed as a fact by some web site, and such a fact can be either true or false. There are often conflicting facts on the web, such as different sets of authors for a book. There are also 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 web sites (especially if by many of them). A web site is trustworthy if most facts it provides are true.

Because of this inter-dependency between facts and web sites, we choose an iterative computational method. At each iteration, the probabilities of facts being true and the trustworthiness of web sites are inferred from each other. This iterative procedure is rather different from Authority-Hub analysis [2]. The first difference is in the definitions. The trustworthiness of a web site does not depend on how many facts it provides, but on the accuracy of those facts. Nor can we compute the probability of a fact being true by adding up the trustworthiness of web sites providing it. These lead to non-linearity in computation. Second and more importantly, different facts influence each other. For example, if a web site says a book is written by "Jessamyn Wendell", and another says "Jessamyn Burns Wendell", then these two web sites actually support each other although they provide slightly different facts.

In summary, we make three major contributions in this paper. First, 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 web sites, confidence of facts, and influences between facts. Finally, we propose an algorithm called TRUTHFINDER for identifying true facts using iterative methods.

The rest of the paper is organized as follows. We describe the problem in Section 2, and propose the computational model in Section 3. Experimental results are presented in Section 4, and we conclude this study in Section 5.

2. PROBLEM DEFINITIONS

The input of TRUTHFINDER is a large number of facts about properties of a certain type of objects. The facts are provided by many web sites. There are usually multiple conflicting facts from different web sites for each object, and the goal of TRUTHFINDER is to identify the true fact among them. Figure 1 shows a mini example dataset. Each web site provides at most one fact for an object.

We first introduce the two most important definitions in this paper, the confidence of facts and the trustworthiness of web sites.

Definition 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.

Definition 2. **(Trustworthiness of web sites.)** The trustworthiness of a web site 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, one web site claims the author to be "Jennifer Widom" and another one claims "J. Widom". If one of them is true, the other is also likely to be true.

In order to represent such relationships, we propose the concept of ***implication*** *between facts.* The implication from fact f_1 to f_2, $imp(f_1 \rightarrow f_2)$, is f_1's influence on f_2's confidence, *i.e.*, how much f_2's confidence should be increased (or decreased) according to f_1's confidence. It is required that $imp(f_1 \rightarrow f_2)$ is a value between -1 and 1. A positive value indicates if f_1 is correct, f_2 is likely to be correct. While a negative value means if f_1 is correct, f_2 is likely to be wrong. The details about this will be described in Section 3.1.2.

Please notice that the definition of implication is domain specific. When a user uses TRUTHFINDER on a certain domain, he 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 $imp(f_1 \rightarrow f_2) = sim(f_1, f_2) - base\_sim$, where $sim(f_1, f_2)$ is the similarity between f_1 and f_2, and $base\_sim$ is a threshold for similarity.

Based on common sense and our observations on real data, we have four basic heuristics that serve as the bases of our computational model.

Heuristic 1: *Usually there is only one true fact for a property of an object.*

Heuristic 2: *This true fact appears to be the same or similar on different web sites.*

Heuristic 3: *The false facts on different web sites are less likely to be the same or similar.*

Heuristic 4: *In a certain domain, a web site that provides mostly true facts for many objects will likely provide true facts for other objects.*

3. COMPUTATIONAL MODEL

Based on the above heuristics, we know if a fact is provided by many trustworthy web sites, it is likely to be true; if a fact is conflicting with the facts provided by many trustworthy web sites, it is unlikely to be true. On the other hand, a web site is trustworthy if it provides facts with high confidence. We can see that the web site trustworthiness and fact confidence are determined by each other, and we can use an iterative method to compute both. Because true facts are more consistent than false facts (Heuristic 3), it is likely that we can distinguish true facts from false ones at the end. In this section we discuss the computational model.

3.1 Web Site Trustworthiness and Fact Confidence

We first discuss how to infer web site trustworthiness and fact confidence from each other.

3.1.1 Basic Inference

As defined in Definition 2, the trustworthiness of a web site is just the expected confidence of facts it provides. For web site w, we compute its trustworthiness $t(w)$ by calculating the average confidence of facts provided by w.

$$t(w) = \frac{\sum_{f \in F(w)} s(f)}{|F(w)|}, \qquad (1)$$

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

Name	Description
M	Number of web sites
N	Number of facts
w	A web site
$t(w)$	The trustworthiness of w
$\tau(w)$	The trustworthiness score of w
$F(w)$	The set of facts provided by w
f	A fact
$s(f)$	The confidence of f
$\sigma(f)$	The confidence score of f
$\sigma^*(f)$	The adjusted confidence score of f
$W(f)$	The set of web sites providing f
$o(f)$	The object that f is about
$imp(f_j \to f_k)$	Implication from f_j to f_k
ρ	Weight of objects about the same object
γ	Dampening factor
δ	Max difference between two iterations

Table 2: Variables and Parameters of TRUTHFINDER

In comparison, it is much more difficult to estimate the confidence of a fact. As shown in Figure 2, the confidence of a fact f_1 is determined by the web sites providing it, and other facts about the same object.
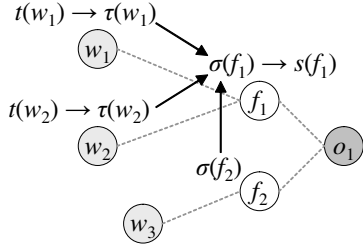

Figure 2: Computing confidence of a fact

Let us first analyze the simple case where there is no related fact, and f_1 is the only fact about object o_1 (*i.e.*, f_2 does not exist in Figure 2). Because f_1 is provided by w_1 and w_2, if f_1 is wrong, then both w_1 and w_2 are wrong. We first assume w_1 and w_2 are independent. (This is not true in many cases and we will compensate for it later.) Thus the probability that both of them are wrong is $(1 - t(w_1)) \cdot (1 - t(w_2))$, and the probability that f_1 is not wrong is $1 - (1 - t(w_1)) \cdot (1 - t(w_2))$. In general, if a fact f is the only fact about an object, then its confidence $s(f)$ can be computed as

$$s(f) = 1 - \prod_{w \in W(f)} (1 - t(w)), \qquad (2)$$

where $W(f)$ is the set of web sites providing f.

In Equation (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 logarithm and define the *trustworthiness score* of a web site as

$$\tau(w) = -\ln(1 - t(w)). \qquad (3)$$

$\tau(w)$ is between and 0 and $+\infty$, which better characterizes how accurate w is. For example, suppose there are two web sites w_1 and w_2 with trustworthiness $t(w_1) = 0.9$ and $t(w_2) = 0.99$. We can see that w_2 is much more accurate than w_1, but their trustworthiness do not differ much as $t(w_2) = 1.1 \times t(w_1)$. If we measure their accuracy with trustworthiness score, we will find $\tau(w_2) = 2 \times \tau(w_1)$, which better represents the accuracy of web sites.

Similarly, we define the *confidence score* of a fact as

$$\sigma(f) = -\ln(1 - s(f)). \qquad (4)$$

A very useful property is that, the confidence score of a fact f is just the sum of the trustworthiness scores of web sites providing f. This is shown in the following lemma.

LEMMA 1.

$$\sigma(f) = \sum_{w \in W(f)} \tau(w) \qquad (5)$$

PROOF. *According to Equation (2),*

$$1 - s(f) = \prod_{w \in W(f)} (1 - t(w)).$$

Take logarithm on both side and we have

$$\ln(1 - s(f)) = \sum_{w \in W(f)} \ln(1 - t(w))$$
$$\iff \quad \sigma(f) = \sum_{w \in W(f)} \tau(w)$$

∎

3.1.2 *Influences between Facts*

The above discussion shows how to compute the confidence of a fact that is the only fact about an object. However, there are usually many different facts about an object (such as f_1 and f_2 in Figure 2), and these facts influence each other. Suppose in Figure 2 the implication from f_2 to f_1 is very high (*e.g.*, they are very similar). If f_2 is provided by many trustworthy web sites, then f_1 is also somehow supported by these web sites, and f_1 should have reasonably high confidence. Therefore, we should increase the confidence score of f_1 according to the confidence score of f_2, which is the sum of trustworthiness scores of web sites providing f_2. We define the *adjusted confidence score* of a fact f as

$$\sigma^*(f) = \sigma(f) + \rho \cdot \sum_{o(f')=o(f)} \sigma(f') \cdot imp(f' \to f). \qquad (6)$$

ρ is a parameter between 0 and 1, which controls the influence of related facts. We can see that $\sigma^*(f)$ is the sum of confidence score of f and a portion of the confidence score of each related fact f' multiplies the implication from f' to f. Please notice that $imp(f' \to f) < 0$ when f is conflicting with f'.

We can compute the confidence of f based on $\sigma^*(f)$ in the same way as computing it based on $\sigma(f)$ (defined in Equation (4)). We use $s^*(f)$ to represent this confidence.

$$s^*(f) = 1 - e^{-\sigma^*(f)}. \qquad (7)$$

3.1.3 *Handling Additional Subtlety*

One problem with the above model is we have been *assuming different web sites are independent with each other*. This assumption is often incorrect because errors can be propagated between web sites. According to the definitions above, if a fact f is provided by five web sites with trustworthiness of 0.6 (which is quite low), f will have confidence of 0.99! But actually some of the web sites may copy contents from others. In order to compensate for the problem of overly high confidence, we add a *dampening factor* γ into Equation (7), and redefine fact confidence as $s^*(f) = 1 - e^{-\gamma \cdot \sigma^*(f)}$, where $0 < \gamma < 1$.

The second problem with our model is that, *the confidence of a fact f can easily be negative* if f is conflicting with some facts provided by trustworthy web sites, which makes

$\sigma^*(f) < 0$ and $s^*(f) < 0$. This is unreasonable because even with negative evidences, there is still a chance that f is correct, so its confidence should still be above zero. Therefore, we adopt the widely used Logistic function [3], which is a variant of Equation (7), as the final definition for fact confidence.

$$s(f) = \frac{1}{1 + e^{-\gamma \cdot \sigma^*(f)}}. \tag{8}$$

When $\gamma \cdot \sigma^*(f)$ is significantly greater than zero, $s(f)$ is very close to $s^*(f)$ because $\frac{1}{1+e^{-\gamma \cdot \sigma^*(f)}} \approx 1 - e^{-\gamma \cdot \sigma^*(f)}$. When $\gamma \cdot \sigma^*(f)$ is significantly less than zero, $s(f)$ is close to zero but remains positive, which is consistent with the real situation. Equation (8) is also very similar to Sigmoid function [6], which has been successfully used in various models in many fields.

3.2 Iterative Computation

As described above, we can infer the web site trustworthiness if we know the fact confidence, and vice versa. As in Authority-hub analysis [2] and PageRank [4], TRUTHFINDER adopts an iterative method to compute the trustworthiness of web sites and confidence of facts. Initially, it has very little information about the web sites and the facts. At each iteration TRUTHFINDER tries to improve its knowledge about their trustworthiness and confidence, and it stops when the computation reaches a stable state.

We choose the initial state in which all web sites have a uniform trustworthiness t_0. (t_0 is set to the estimated average trustworthiness, such as 0.9.) In each iteration, TRUTHFINDER first uses the web site trustworthiness to compute the fact confidence, and then recomputes the web site trustworthiness from the fact confidence. It stops iterating when it reaches a stable state. The stableness is measured by the change of the trustworthiness of all web sites, which is represented by a vector $\overrightarrow{t}$. If $\overrightarrow{t}$ only changes a little after an iteration (measured by cosine similarity between the old and the new $\overrightarrow{t}$), then TRUTHFINDER will stop.

4. EMPIRICAL STUDY

In this section we present experiments on a real dataset, which shows the effectiveness of TRUTHFINDER. We compare it with a baseline approach called VOTING, which chooses the fact that is provided by most web sites. We also compare TRUTHFINDER with Google by comparing the top web sites found by each of them.

All experiments are performed on an Intel PC with a 1.66GHz dual-core processor, 1GB memory, running Windows XP Professional. All approaches are implemented using Visual Studio.Net (C#). The two parameters in Equation (8) are set as $\rho = 0.5$ and $\gamma = 0.3$. The maximum difference between two iterations, δ, is set to 0.001%.

4.1 Book Authors Dataset

This 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 its 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. We find the image of each book, and use the authors on the book cover as the standard fact.

We 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 $\frac{z}{\max(x,y)}$.[1]

Sometimes TRUTHFINDER provides partially correct facts. For example, the standard set of authors for a book is "Graeme C. Simsion and Graham Witt", and the authors found by TRUTHFINDER may be "Graeme Simsion and G. Witt". We consider "Graeme Simsion" and "G. Witt" as partial matches for "Graeme C. Simsion" and "Graham Witt", and give them partial scores. We assign different weights to different parts of persons' names. Each author name has total weight 1, and the ratio between weights of last name, first name, and middle name is 3:2:1. For example, "Graeme Simsion" will get a partial score of 5/6 because it omits the middle name of "Graeme C. Simsion". If the standard name has a full first or middle name, and TRUTHFINDER provides the correct initial, we give TRUTHFINDER half score. For example, "G. Witt" will get a score of 4/5 with respect to "Graham Witt", because the first name has weight 2/5, and the first initial "G." gets half of the score.

The implication between two sets of authors f_1 and f_2 is defined in a very similar way as the accuracy of f_2 with respect to f_1. One important observation is that many bookstores provide incomplete facts, such as only the first author. For example, if a web site w_1 says a book is written by "Jennifer Widom", and another web site w_2 says it is written by "Jennifer Widom and Stefano Ceri", then w_1 actually supports w_2 because w_1 is probably providing partial fact. Therefore, If fact f_2 contains authors that are not in fact f_1, then f_2 is actually supported by f_1. The implication from f_1 to f_2 is defined as follows. If f_1 has x authors and f_2 has y authors, and there are z shared ones, then $imp(f_1 \rightarrow f_2) = z/x - base\_sim$, where $base\_sim$ is the threshold for positive implication and is set to 0.5.
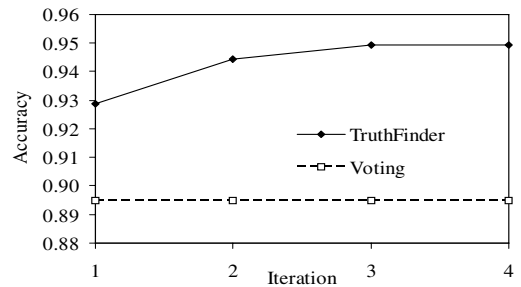

Figure 3: Accuracies of TRUTHFINDER **and** VOTING

Figure 3 shows the accuracies of TRUTHFINDER and VOTING. One can see that TRUTHFINDER is significantly more accurate than VOTING even at the first iteration, where all bookstores have the same trustworthiness. This is because TRUTHFINDER considers the implications between different

---

[1]For simplicity we do not consider the order of authors in this study, although TRUTHFINDER can report the authors in correct order in most cases.

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 to about 95% after the third iteration and remains stable. It takes TRUTHFINDER 8.73 seconds to pre-computes the implications between related facts, and 4.43 seconds to finish the four iterations. VOTING takes 1.22 seconds.
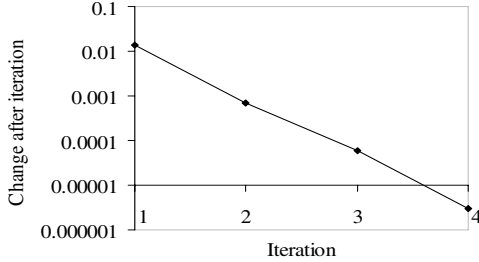

Figure 4: Relative changes of TRUTHFINDER

Figure 4 shows the relative change of the trustworthiness vector after each iteration, which is defined as one minus the cosine similarity of the old and new vectors. We can see TRUTHFINDER converges in a steady speed.

In Table 3 we manually compare the results of VOTING, TRUTHFINDER, and the authors provided by Barnes & Noble on its web site. We list the number of books in which each approach makes each type of errors. Please notice that one approach may make multiple errors for one book.

Type of error	VOTING	TRUTHFINDER	Barnes & Noble
correct	71	85	64
miss author(s)	12	2	4
incomplete names	18	5	6
wrong first/middle names	1	1	3
has redundant names	0	2	23
add incorrect names	1	5	5
no information	0	0	2

Table 3: Compare the results of VOTING, TRUTHFINDER, and Barnes & Noble

VOTING tends to miss authors because many bookstores only provide subsets of all authors. On the other hand, TRUTHFINDER tends to consider facts with more authors as correct facts because of our definition of implication for book authors, and thus makes more mistakes of adding in incorrect names. One may think that the largest bookstores will provide accurate information, which is surprisingly untrue. Table 3 shows Barnes & Noble has more errors than VOTING and TRUTHFINDER on these 100 randomly selected books.

Finally, we perform an interesting experiment on finding trustworthy web sites. It is well known that Google (or other search engines) is good at finding authoritative web sites. But do these web sites provide accurate information? To answer this question, we compare the online bookstores that are given highest ranks by Google with the bookstores with highest trustworthiness found by TRUTHFINDER. We query Google with "bookstore"[2], and find all bookstores that exist in our dataset from the top 300 Google results. The accuracy of each bookstore is tested on the 100 randomly selected books in the same way as we test the accuracy of

[2]This query was submitted on Feb 7, 2007.

TRUTHFINDER. We only consider bookstores that provide at least 10 of the 100 books.

TRUTHFINDER			
bookstore	trustworthiness	#book	accuracy
TheSaintBookstore	0.971	28	0.959
MildredsBooks	0.969	10	1.0
alphacraze.com	0.968	13	0.947
Marando.de Versandbuchhandlung	0.967	18	0.947
blackwell online	0.962	38	0.879
Annex Books	0.956	15	0.913
Stratford Books	0.951	50	0.857
movies with a smile	0.950	12	0.911
Aha-Buch	0.949	31	0.901
Players quest	0.947	19	0.936
average accuracy			0.925
Google			
bookstore	Google rank	#book	accuracy
Barnes & Noble	1	97	0.865
Powell's books	3	42	0.654
ecampus.com	11	18	0.847
average accuracy			0.789

Table 4: Compare the accuracies of top bookstores by TRUTHFINDER and by Google

Table 4 shows the accuracy and number of books provided (among the 100 books) of different bookstores. TRUTHFINDER can find bookstores that provide much more accurate information than the top bookstores found by Google. TRUTHFINDER also finds some large trustworthy bookstores, such as A1 Books (not among the top 10 shown in Table 4) which provides 86 of 100 books with accuracy of 0.878. Please notice that TRUTHFINDER uses no training data, and the testing data is manually created by reading the authors' names from book covers. Therefore, we believe the results suggest that there may be better alternatives than Google for finding accurate information on the web.

5. CONCLUSIONS

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

6. REFERENCES

[1] A. Borodin, G. Roberts, J. Rosenthal, P. Tsaparas. Link analysis ranking: Algorithms, theory, and experiments. *ACM Transactions on Internet Technology*, 5(1):231–297, 2005.
[2] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA*, 1998.
[3] Logistical Equation from Wolfram MathWorld. http://mathworld.wolfram.com/LogisticEquation.html
[4] 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.
[5] Princeton Survey Research Associates International. Leap of faith: using the Internet despite the dangers. Results of a National Survey of Internet Users for Consumer Reports WebWatch, Oct 2005.
[6] Sigmoid Function from Wolfram MathWorld. http://mathworld.wolfram.com/SigmoidFunction.html