Truth Discovery with Multiple Conflicting Information Providers on the Web

Xiaoxin Yin, Jiawei Han, *Senior Member*, *IEEE*, and Philip S. Yu, *Fellow*, *IEEE*

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 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 websites. We design a general framework for the Veracity problem and invent an algorithm, called TRUTHFINDER, 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*. An iterative method is used to infer the trustworthiness of websites and the correctness of information from each other. Our experiments show that TRUTHFINDER successfully finds true facts among conflicting information and identifies trustworthy websites better than the popular search engines.

Index Terms—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 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 1 (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 says 29,002 feet, and another one says 29,017 feet. Which answer should the user trust?

1. The query was sent on 9 February 2007.

- *X. Yin is with Microsoft Research, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052. E-mail: xyin@microsoft.com.*
- *J. Han is with the Department of Computer Science, University of Illinois at Urbana-Champaign, 201 N. Goodwin Avenue, 2132, Urbana, IL 61801. E-mail: hanj@cs.uiuc.edu.*
- *P.S. Yu is with the Department of Computer Science, University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607. E-mail: psyu@cs.uic.edu.*

Manuscript received 17 July 2007; revised 20 Nov. 2007; accepted 6 Dec. 2007; published online 19 Dec. 2007.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2007-07-0369.
Digital Object Identifier no. 10.1109/TKDE.2007.190745.

Example 2 (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. According to a survey on the credibility of websites conducted by Princeton Survey Research in 2005 [11], 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 [7], and PageRank [10], which lead to Google.com. However, does authority lead to accuracy of information? The answer is unfortunately no. Top-ranked websites are usually the most popular ones. However, popularity does not mean accuracy. For example, according to our experiments (Section 4.2), the bookstores ranked on top by Google (*Barnes & Noble* and *Powell's books*) contain many errors on book author information. In comparison, some small bookstores (e.g., *A1 Books*) provide more accurate information.

In this paper, 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 paper, we only study the facts that are either properties of

TABLE 1
Conflicting Information about Book Authors

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

objects (e.g., weights of laptop computers) or relationships between two objects (e.g., authors of books). We also require that the facts can be parsed from the web pages.

There are often conflicting facts on the Web, such as different sets of authors for a book. There are also many websites, some of which are more trustworthy than others.[2] A fact is likely to be true if it is provided by trustworthy websites (especially if by many of them). 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 are inferred from each other. This iterative procedure is rather different from Authority-Hub analysis [7]. The first difference is in the definitions. The trustworthiness of a website does not depend on how many facts it provides but on the accuracy of those facts. For example, a website providing 10,000 facts with an average accuracy of 0.7 is much less trustworthy than a website providing 100 facts with an accuracy of 0.95. Thus, we cannot compute the trustworthiness of a website by adding up the weights of its facts as in [7], nor can we compute the probability of a fact being true by adding up the trustworthiness of websites providing it. Instead, we have to resort to probabilistic computation. Second and more importantly, different facts influence each other. For example, if a website says that a book is written by "Jessamyn Wendell" and another says "Jessamyn Burns Wendell," then these two websites actually support each other although they provide slightly different facts. We incorporate such influences between facts into our computational model.

In summary, we make three major distributions 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 websites, confidence of facts, and influences between facts. Finally, we propose an algorithm called TRUTHFINDER for identifying true facts using iterative methods. Our experiments show that TRUTHFINDER achieves very high accuracy in discovering true facts, and it can select better trustworthy websites than authority-based search engines such as Google.

2. The "trustworthiness" in this paper means accuracy in providing information. It is different from the "trustworthiness" in the studies of *trust management* [2].

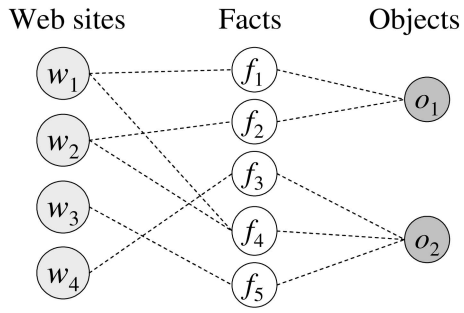

Fig. 1. Input of TRUTHFINDER.

The rest of the paper is organized as follows: We describe the problem in Section 2 and propose the computational model and algorithms in Section 3. Experimental results are presented in Section 4. We discuss related work in Section 5 and conclude this study in Section 6.

2 Problem Definitions

In this paper, we study 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 shows a miniexample data set, which contains five facts about two objects provided by four websites. Each website provides at most one fact for an object.

2.1 Basic Definitions

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

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 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. For example, one website claims that a book is written by "Karen Holtzblatt," whereas another claims that it is written by "Jessamyn Wendell." However, sometimes facts may be supportive to each other although they are slightly different. For example, one website claims the author to be "Jennifer Widom," and another one claims "J. Widom," or one website says that a certain camera is 4 inches long, and another one says 10 cm. If one of such facts 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 that if f_1 is correct, f_2 is likely to be correct. While a negative value means that if f_1 is correct, f_2 is likely to be wrong. The details about this will be described in Section 3.1.2.

We define implication instead of similarity between facts because such relationship is asymmetric. For example, in some domains (e.g., book authors), websites tend to provide incomplete facts (e.g., first author of a book). Suppose two websites provide author information for the same book. The first website indicates that the author of the book is "Jennifer Widom," which is fact f_1. The second website says that there are two authors "Jennifer Widom and Stefano Ceri," which is fact f_2. If f_2 is correct, then f_1 is incomplete and will have low confidence, and thus, $imp(f_2 \rightarrow f_1)$ is low. On the other hand, we know that it is very common for a website to provide only one of the authors for a book. Thus, f_1 may only tell us that "Jennifer Widom" is one author of the book instead of the sole author. If we are confident about f_1, we should also be confident about f_2 because f_2 is consistent with f_1, and $imp(f_1 \rightarrow f_2)$ should be high. From this example, we can see that implication is an asymmetric relationship.

Please notice that the definition of implication is domain specific. The implication for book authors should be very different from that for the number of pixels of camcorders. 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 $imp(f_1 \rightarrow f_2) = sim(f_1, f_2) - base\_sim$, where $sim(f_1, f_2)$ is the similarity between f_1 and f_2, and $base\_sim$ is a threshold for similarity.

2.2 Basic Heuristics

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

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

In this paper, we assume that there is only one true fact for a property of an object. The case of multiple true facts will be studied in our future work.

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

Different websites that provide this true fact may present it in either the same or slightly different ways, such as "Jennifer Widom" versus "J. Widom."

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

Different websites often make different mistakes for the same object and thus provide different false facts. Although false facts can be propagated among websites, in general, the false facts about a certain object are much less consistent than the true facts.

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

TABLE 2
Variables and Parameters of TRUTHFINDER

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
$imp(f_j \rightarrow f_k)$	Implication from f_j to f_k
ρ	Weight of objects about the same object
γ	Dampening factor
δ	Max difference between two iterations

There are trustworthy websites such as wikipedia and untrustworthy websites such as blogs and some small websites. We believe that a website has some consistency in the quality of its information in a certain domain.

3 COMPUTATIONAL MODEL

Based on the above heuristics, we know that if a fact is provided by many trustworthy websites, it is likely to be true; and, if a fact is conflicting with the facts provided by many trustworthy websites, it is unlikely to be true. On the other hand, a website is trustworthy if it provides facts with high confidence. We can see that the website 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 find and distinguish true facts from false ones at the end.

In this section, we introduce the model of iterative computation. Table 2 shows the variables and parameters used in the following discussion.

3.1 Website Trustworthiness and Fact Confidence

We first discuss how to infer website trustworthiness and fact confidence from each other. The inference of website trustworthiness is rather simple, whereas that of fact confidence is more complicated. We start from the simplest case and proceed to more complicated ones step by step.

3.1.1 Basic Inference

As defined in Definition 2, the trustworthiness of a website 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:

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

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

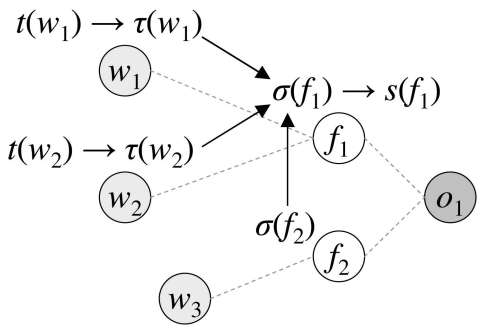

Fig. 2. Computing confidence of a fact.

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

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 Fig. 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 that 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

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

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

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 as

$$\tau(w) = -\ln(1 - t(w)). \tag{3}$$

$\tau(w)$ is between and zero and $+\infty$, and a larger $\tau(w)$ indicates higher trustworthiness.

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

$$\sigma(f) = -\ln(1 - s(f)). \tag{4}$$

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. This is shown in the following lemma, which is used to compute $\sigma(f)$ in TRUTHFINDER.

Lemma 1.

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

Proof. According to (2),

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

Take the logarithm on both sides, and we have

$$\ln(1 - s(f)) = \sum\nolimits_{w \in W(f)} \ln(1 - t(w))$$
$$\Longleftrightarrow \quad \sigma(f) = \sum\nolimits_{w \in W(f)} \tau(w).$$

□

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 Fig. 2), and these facts influence each other. Suppose in Fig. 2 that 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 websites, then f_1 is also somehow supported by these websites, 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 the trustworthiness scores of websites providing f_2. We define the *adjusted confidence score* of a fact f as

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

ρ is a parameter between zero and one, which controls the influence of related facts. We can see that $\sigma^*(f)$ is the sum of the confidence scores of f, and a portion of the confidence score of each related fact f' multiplies the implication from f' to f. Please notice that $imp(f' \rightarrow f) < 0$ 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 (4)). We use $s^*(f)$ to represent this confidence:[3]

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

3.1.3 Handling Additional Subtlety

Until now, we have described how to compute fact confidence from website trustworthiness. There are still two problems with our model, which are discussed below.

The first problem is that we have been *assuming that different websites are independent of each other*. This assumption is often incorrect because errors can be propagated between websites. According to the definitions above, if a fact f is provided by five websites with a trustworthiness of 0.6 (which is quite low), f will have a confidence of 0.99. However, actually, some of the websites may copy contents from others. In order to compensate for the problem of overly high confidence, we add a *dampening factor* γ into (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

3. With a similar proof as in Lemma 1, we can show that (7) is equivalent to $s^*(f) =$

$$1 - \prod_{w \in W(f)} (1 - t(w)) \cdot \prod_{o(f') = o(f)} \left(\prod_{w' \in W(f')} (1 - t(w')) \right)^{\rho \cdot imp(f' \rightarrow f)}.$$

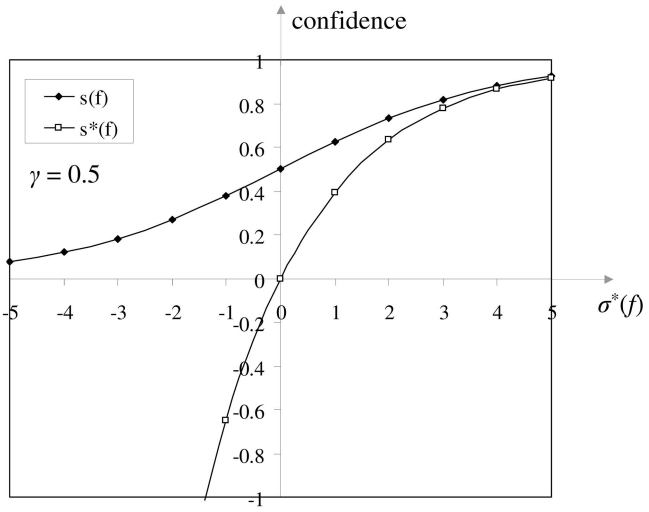

Fig. 3. Two methods for computing confidence.

facts provided by trustworthy websites, which makes $\sigma^*(f) < 0$ and $s^*(f) < 0$. This is unreasonable because 1) the confidence cannot be negative and 2) even with negative evidences, there is still a chance that f is correct, so its confidence should still be above zero. Moreover, if we set $s^*(f) = 0$ if it is negative according to (7), this "chunking" operation and the multiple zero values may lead to unstable conditions in our iterative computation. Therefore, we adopt the widely used Logistic function [8], which is a variant of (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. We compare these two definitions in Fig. 3. One can see that the two curves are very close when $\sigma^*(f) > 3$. $s^*(f)$ decreases very sharply when $\sigma^*(f) < 1$, which is not reasonable because it is always possible that the fact is true even with negative evidence. In comparison, $s(f)$ decreases slowly and is slightly above zero when $\sigma^*(f) \ll 0$, which is consistent with the real situation. Please notice that (8) is also very similar to the Sigmoid function [12], which has been successfully used in various models in many fields.

3.2 Computing Website Trustworthiness and Fact Confidence with Matrix Operations

We have described how to calculate website trustworthiness and fact confidence. It will be more convenient if we can convert the computational procedure into some basic matrix operations, which can be implemented easily and performed efficiently. To facilitate our discussions, we use vectors to represent the trustworthiness of all websites and the confidence of all facts. Let the vectors be

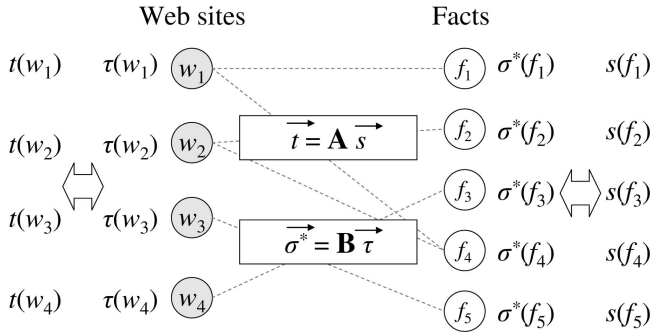

Fig. 4. Computing website trustworthiness and fact confidence with matrix operations.

$$\begin{aligned}
\overrightarrow{t} &= (t(w_1), \ldots, t(w_M))^{\mathrm{T}}, \\
\overrightarrow{\tau} &= (\tau(w_1), \ldots, \tau(w_M))^{\mathrm{T}}, \\
\overrightarrow{s} &= (s(f_1), \ldots, s(f_N))^{\mathrm{T}}, \\
\overrightarrow{\sigma^*} &= (\sigma^*(f_1), \ldots, \sigma^*(f_N))^{\mathrm{T}}.
\end{aligned}$$

We want to define a $M \times N$ matrix $\mathbf{A}$ for inferring website trustworthiness from fact confidence and a $N \times M$ matrix $\mathbf{B}$ for the reverse inference, i.e.

$$\begin{aligned}
\overrightarrow{t} &= \mathbf{A}\overrightarrow{s}, \\
\overrightarrow{\sigma^*} &= \mathbf{B}\overrightarrow{\tau}.
\end{aligned} \tag{9}$$

$\overrightarrow{t}$ and $\overrightarrow{\tau}$ can be converted from each other using (3), and $\overrightarrow{s}$ and $\overrightarrow{\sigma}$ can be converted using (8).

Matrix $\mathbf{A}$ can be easily defined according to (1), by setting

$$\mathbf{A}_{ij} = \begin{cases} 1 \big/ |F(w_i)|, & \text{if } f_j \in F(w_i), \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

In comparison, matrix $\mathbf{B}$ involves more factors because facts about the same object influence each other. From (6) and Lemma 1, we can infer that

$$\mathbf{B}_{ji} = \begin{cases} 1, & \text{if } w_i \text{ provides } f_j, \\ \rho \cdot imp(f_k \rightarrow f_j), & \text{if } w_i \text{ provides } f_k \text{ and } o(f_k) = o(f_j), \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Both $\mathbf{A}$ and $\mathbf{B}$ are sparse matrices. In general, the website trustworthiness and fact confidence can be computed conveniently with matrix operations, as shown in Fig. 4.

The above procedure is very different from Authority-Hub analysis proposed by Kleinberg [7]. It involves nonlinear transformations and thus cannot be computed using eigenvector computation as in [7]. Authority-Hub analysis defines the authority scores and hub scores as the sum of each other. On the other hand, TRUTHFINDER 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.

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 $\mathbf{A}$ and $\mathbf{B}$
for each $w \in W$ /* *setting initial state* */
$t(w) \leftarrow t_0$
$\tau(w) \leftarrow -\ln(1 - t(w))$
repeat /* *iterative computation* */
$\overrightarrow{\sigma^*} \leftarrow \mathbf{B}\overrightarrow{\tau}$
compute $\overrightarrow{s}$ from $\overrightarrow{\sigma^*}$
$\overrightarrow{t}' \leftarrow \overrightarrow{t}$ /* *make a copy of* $\overrightarrow{t}$ */
$\overrightarrow{t} \leftarrow \mathbf{A}\overrightarrow{s}$
compute $\overrightarrow{\tau}$ from $\overrightarrow{t}$
until cosine similarity of $\overrightarrow{t}$ and $\overrightarrow{t}'$ is greater than $1 - \delta$

Fig. 5. Algorithm of TRUTHFINDER.

3.3 Iterative Computation

As described above, we can infer the website trustworthiness if we know the fact confidence and vice versa. As in Authority-Hub analysis [7] and PageRank [10], TRUTHFINDER adopts an iterative method to compute the trustworthiness of websites and confidence of facts. Initially, it has very little information about the websites 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.

As in other iterative approaches [7], [10], TRUTHFINDER needs an initial state. We choose the initial state in which all websites have uniform trustworthiness t_0. (t_0 should be set to the estimated average trustworthiness, such as 0.9.) From the website trustworthiness TRUTHFINDER can infer the confidence of facts, which are very meaningful because the facts supported by many websites are more likely to be correct. On the other hand, if we start from a uniform fact confidence, we cannot infer meaningful trustworthiness for websites. Before the iterative computation, we also need to calculate the two matrices $\mathbf{A}$ and $\mathbf{B}$, as defined in Section 3.2. They are calculated once and used at every iteration.

In each step of the iterative procedure, TRUTHFINDER first uses the website trustworthiness to compute the fact confidence and then recomputes the website trustworthiness from the fact confidence. Each step only requires two matrix operations and conversions between $t(w)$ and $\tau(w)$ and between $s(f)$ and $\sigma^*(f)$. The matrices are stored in sparse formats, and the computational cost of multiplying such a matrix and a vector is linear with the number of nonzero entries in the matrix. TRUTHFINDER stops iterating when it reaches a stable state. The stableness is measured by how much the trustworthiness of websites changes between iterations. 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. The overall algorithm is presented in Fig. 5.

3.4 Complexity Analysis

In this section, we analyze the complexity of TRUTHFINDER. Suppose there are L links between all websites and facts. Because different websites may provide the same fact, L should be greater than N (number of facts). Suppose on the average there are k facts about each object, and thus, each fact has $k-1$ related facts on the average.

Let us first look at the two matrices $\mathbf{A}$ and $\mathbf{B}$. Each link between a website and a fact corresponds to an entry in $\mathbf{A}$. Thus, $\mathbf{A}$ has L entries, and it takes $O(L)$ time to compute $\mathbf{A}$. $\mathbf{B}$ contains more entries than $\mathbf{A}$ because $\mathbf{B}_{ji}$ is nonzero if website w_i provides a fact that is related to fact f_j. Thus, there are $O(kL)$ entries in $\mathbf{B}$. Because each website can provide at most one fact about each object, each entry of $\mathbf{B}$ involves only one website and one fact. Thus, it still takes constant time to compute each entry of $\mathbf{B}$, and it takes $O(kL)$ time to compute $\mathbf{B}$.

The time cost of multiplying a sparse matrix and a vector is linear with the number of entries in the matrix. Therefore, each iteration takes $O(kL)$ time and no extra space. Suppose there are I iterations. TRUTHFINDER takes $O(IkL)$ time and $O(kL + M + N)$ space.

If in some cases, $O(kL)$ space is not available, we can discard the matrix operations and compute the website trustworthiness and fact confidence using their definitions in Section 3.1. If we precompute the implication between all facts, then $O(kN)$ space is needed to store these implication values, and the total space requirement is $O(L + kN)$. If the implication between two facts can be computed in a very short constant time and we do not precompute the implication, then the total space requirement is $O(L + M + N)$. In both cases, it takes $O(L)$ time to propagate between website trustworthiness and fact confidence and $O(kN)$ time to adjust fact confidence according to the interfact implication. Thus, the overall time complexity is $O(IL + IkN)$.

4 EMPIRICAL STUDY

We perform experiments on two real data sets and synthetic data sets to examine the accuracy and efficiency of TRUTHFINDER. The first real data set contains the authors of many books provided by many online bookstores, and the second one contains the runtime of many movies provided by many websites. As will be explained in Section 5, there is no existing approach to the problem studied in this paper, and thus, we compare TRUTHFINDER with a baseline approach as described below.

4.1 Experiment Setting

In order to show the effectiveness of TRUTHFINDER, we compare it with a baseline approach called VOTING. 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. This is the simplest approach and 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 in the following iterations.

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

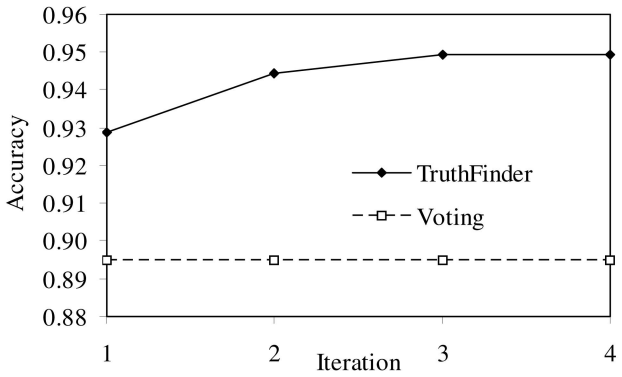

Fig. 6. Accuracies of TRUTHFINDER and VOTING.

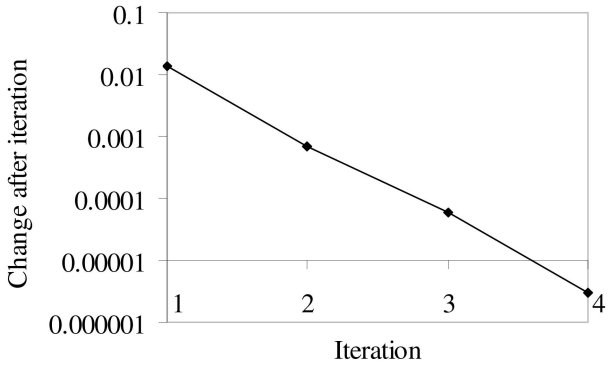

Fig. 7. Relative changes of TRUTHFINDER.

is set to 0.001 percent. We will show in our experiments that the performance of TRUTHFINDER is not sensitive to these parameters.

4.2 Real Data Sets

We test the accuracy and efficiency of TRUTHFINDER on two real data sets. The results are presented below.

4.2.1 Book Authors

The first real data set contains the authors of many books provided by many online bookstores. It contains 1,265 books about computer science and engineering, which are 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 online bookstores that sell the book and the book information from each store, including the price, the authors, and a short description. The data set contains 894 bookstores and 34,031 listings (i.e., bookstore selling a book). On the average, each book has 5.4 different sets of authors, which are indicated by different bookstores.

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 for each book with the standard fact to compute the accuracy of TRUTHFINDER. For a certain book, suppose the standard fact contains x authors; TRUTHFINDER indicates that there are y authors, among which z authors belong to the standard fact. The accuracy of TRUTHFINDER is defined as $\frac{z}{\max(x,y)}$.[4]

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 a total weight of 1, and the ratio between weights of the 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 a 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. Because many bookstores provide incomplete facts (e.g., only the first author), if f_2 contains authors that are not in f_1, the implication from f_1 to f_2 will not be decreased. If f_1 has x authors, 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.

Fig. 6 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 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 percent after the third 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.

Fig. 7 shows the relative change of the trustworthiness vector after each iteration. The change is defined as one minus the cosine similarity of the old and the new vectors. We can see that the relative change decreases by about one order after each iteration, showing that TRUTHFINDER converges at a steady speed. After the fourth iteration, the stop criterion is met, which is indicated by the horizontal axis.

In Table 3, we manually compare the results of VOTING and TRUTHFINDER and the authors provided by Barnes & Noble on its website. 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.

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

4. 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.

TABLE 3
Comparison of the Results of VOTING, TRUTHFINDER, and Barnes & Noble

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

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 according to our experiment. Table 3 shows that Barnes & Noble has more errors than VOTING and TRUTHFINDER on these 100 randomly selected books. It has redundant names for many books. For example, for a book written by "Robert J. Muller," it says the authors are "Robert J. Muller; Muller." Actually Barnes & Noble is less accurate than TRUTHFINDER even if we do not consider such errors.

We give an example mistake made by TRUTHFINDER. The book *"Server Storage Technologies for Windows 2000, Windows Server 2003, and Beyond"* is written by Dilip C. Naik. There are nine bookstores saying that its authors are "Dilip C. Naik and Dilip Naik," seven saying "Dilip C. Naik," and five saying "Dilip Naik." TRUTHFINDER infers that its authors are "Dilip C. Naik and Dilip Naik," which contains redundant names. However, Amazon makes the same mistake as TRUTHFINDER, which may explain why the mistake is propagated to so many bookstores.

Finally, we perform an interesting experiment on finding trustworthy websites. It is well known that Google (or other search engines) is good at finding authoritative websites. However, do these websites 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"[5] and find all bookstores that exist in our data set from the top 300 Google results.[6] The accuracy of each bookstore is tested on the 100 randomly selected books in the same way as we test the accuracy of TRUTHFINDER. We only consider bookstores that provide at least 10 of the 100 books.

The results are shown in Table 4. 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, which provides 86 of 100 books with an 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 that the results suggest that there may be better alternatives than Google for finding accurate information on the Web.

5. This query was submitted on 7 February 2007.

6. Our data set does not contain the largest bookstores such as Barnes & Noble and Amazon.com, because they do not list their books on www.abebooks.com. However, we include Barnes & Noble here because we have manually retrieved its authors for the 100 books for testing.

TABLE 4
Comparison of the Accuracies of Top Bookstores by TRUTHFINDER and by Google

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
Strand bookstore	40	0	N/A
Brett's books	140	0	N/A
Covenant bookstore	165	0	N/A
Awesome books	187	0	N/A
Sam Weller's	264	0	N/A
average accuracy			0.789

4.2.2 Movie Runtime

The second real data set contains runtimes of movies provided by many websites. It contains 603 movies, which are the movies with top ratings in every genre on IMDB.com.[7] We collect the runtime data of each movie using Google. For example, for the movie *Forrest Gump*, we query Google with "`forrest gump`" + `runtime` and parse the result page that contains digests of the first 100 results. If we see terms like "130 minutes" or "2h10m" appearing after "runtime" (or "run time"), we consider such terms as the runtime of this movie. We found 17,109 useful digests, which contain information from 1,727 websites. On average, each movie has 14.3 different runtimes provided by different websites.

Because of the authority of IMDB, we consider the runtime it provides as the standard facts (information from IMDB.com is excluded from our data set). We randomly select 100 movies and find their runtimes on IMDB, in order to test the accuracy of TRUTHFINDER. If the standard runtime for a movie is x minutes and TRUTHFINDER infers that its runtime is y minutes, then the accuracy is defined as $\frac{|y-x|}{\max(x,y)}$. If there are two facts f_1 and f_2 about a movie, where $f_1 = y$, and $f_2 = z$, then the implication from f_1 to f_2 is defined as $imp(f_1 \rightarrow f_2) = \frac{|y-z|}{\max(y,z)} - base\_sim$, where $base\_sim$ is the threshold for positive implication and is set to 0.75.

7. Please see http://imdb.com/chart/.

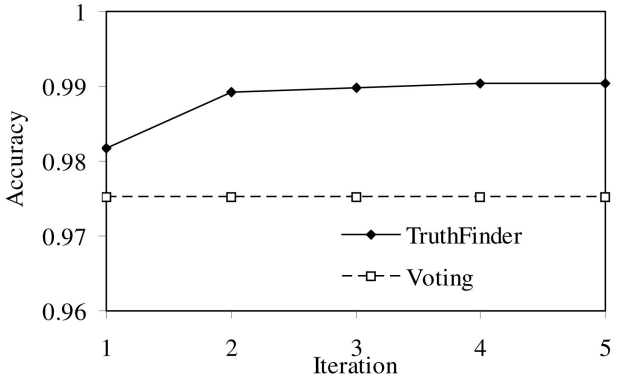

Fig. 8. Accuracies of TRUTHFINDER and VOTING.

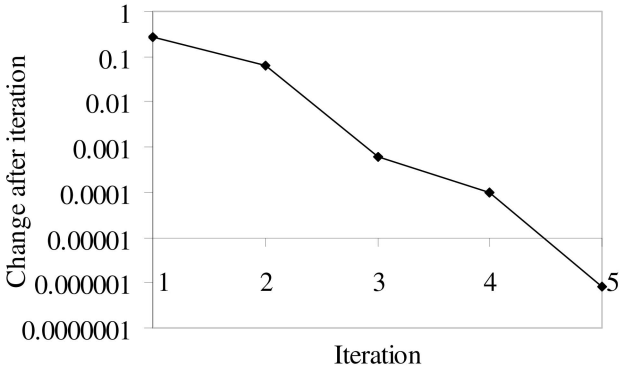

Fig. 9. Relative changes of TRUTHFINDER.

Fig. 8 shows the accuracies of TRUTHFINDER and VOTING on the movie runtime data set. TRUTHFINDER achieves very high accuracy on this data set, reaching 99 percent after three iterations. In comparison, the error rate of VOTING is more than twice that of TRUTHFINDER.

Fig. 9 shows the relative change after each iteration of the trustworthiness vector of TRUTHFINDER. It can be seen that TRUTHFINDER still converges at a steady speed, and it takes five iterations to meet the stop criterion. It takes TRUTHFINDER 2.89 seconds for initialization and 3.19 seconds for five iterations. VOTING takes 1.55 seconds.

In Table 5, we compare the numbers of errors of VOTING and TRUTHFINDER in different cases. Again, we find that TRUTHFINDER is much more accurate than VOTING.

In Table 6, we compare the most trustworthy websites by TRUTHFINDER and top ranked movie websites by Google (with query "movies"[8]). Again, we find that the trustworthy websites found by TRUTHFINDER provide much more accurate information than those ranked high by Google.

4.2.3 Parameter Sensitivity

There are two important parameters in the computation website trustworthiness and fact confidence—ρ and γ in (6) and (8) in Section 3.1. ρ controls the degree of influence between related facts (i.e., facts about the same object), and γ determines the shape of the confidence curve in Fig. 3. By default, $\rho = 0.5$, and $\gamma = 0.3$. The following experiments show that the accuracy of TRUTHFINDER is only very slightly affected by these two parameters.

Fig. 10a shows the accuracy of TRUTHFINDER with different ρs on the two data sets (with $\gamma = 0.3$). It can be seen that the accuracy of TRUTHFINDER is very stable when $\rho \geq 0.1$. Fig. 10b shows the accuracy with different γs on the two data sets (with $\rho = 0.5$). γ has more influence on TRUTHFINDER than ρ. However, the influence is still very limited, and the accuracy only varies by about 0.5 percent even when γ is changed significantly.

We also investigate how the parameters influence the convergence of the algorithm. Figs. 11a and 11b show the relative changes after each iteration of TRUTHFINDER, with different ρs and γs, when TRUTHFINDER is applied on the book author data set. (We do not show experiments on the movie runtime data set because of limited space.) In the figures, we can see that the algorithm always converges rapidly, with a relative change of less than 10^{-5} after five iterations. TRUTHFINDER converges faster when ρ is smaller, which means that the influence between different facts is smaller. This is reasonable because interfact influences add complexity to the problem and will probably make it slower to converge. TRUTHFINDER also converges faster when γ is smaller, probably because a smaller γ means a smaller influence of $\sigma^*(f)$ on $s(f)$ (there is no influence when $\gamma = 0$).

4.3 Synthetic Data Sets

In this section, we test the scalability and noise resistance of TRUTHFINDER on synthetic data sets. We generate data sets containing M websites and N facts. There are $N/5$ objects (i.e., each object has five facts on average), and there are four websites providing each fact on average (i.e., $4N$ links between websites and facts). The expected trustworthiness of each website is $\bar{t}$, and the trustworthiness of each website is drawn from a uniform distribution from $\max(0, 2\bar{t} - 1)$ to $\min(2\bar{t}, 1)$. For example, if $\bar{t} = 0.7$, then the range for website trustworthiness is [0.4, 1].

Each object has a true value, which is a random number drawn from a uniform distribution on interval [1,000, 10,000]. The facts of each object are real numbers that do not deviate too much from the true value. We want to generate the facts by randomly generating some numbers on an interval around the true value. However, we do not want the true value to be at the middle of the interval, which makes it very easy to find the true value by taking the average of all facts. Suppose the value for object o is $v(o)$. We create a *value range* for o that is an interval of length $v(o)/2$. The value range is randomly placed so that the true value $v(o)$ falls at any position in it with equal probability. Then, the facts of o are drawn from a uniform distribution

TABLE 5
Comparison of the Results of VOTING and TRUTHFINDER on Movie Runtime

Type of error	VOTING	TRUTHFINDER
exactly same	40	72
at most 1 min	16	14
2 to 5 min	37	10
more than 5 min	7	4

8. This query was submitted on 7 February 2007.

TABLE 6
Comparison of the Accuracies of Top Movie Websites by TRUTHFINDER and by Google

Most trustworthy movie web sites by TRUTHFINDER			
web site	trustworthiness	#movie	accuracy
dvddb.sparkyb.net	0.968	22	0.996
www.lacuracao.com	0.938	11	0.960
www.guilfordfamily.com	0.926	19	1.0
video.ils.unc.edu	0.925	12	0.953
www.fandango.com	0.920	17	0.964
www.rottentomatoes.com	0.902	80	0.968
www.starflix.com	0.886	20	0.968
www.reel.com	0.868	68	0.958
www.ew.com	0.866	16	0.961
www.bestprices.com	0.835	91	0.927
average accuracy			0.966
Top ranked movie web sites by Google			
web site	Google rank	#movie	accuracy
movies.aol.com	4	85	0.935
www.fandango.com	5	17	0.964
www.movieweb.com	24	43	0.843
www.msnbc.msn.com	25	20	0.863
www.film.com	33	12	0.876
www.austin360.com	42	14	0.891
www.mytelus.com	58	19	0.828
www.accessatlanta.com	64	33	0.882
www.reel.com	66	68	0.958
www.guidelive.com	79	11	0.830
average accuracy			0.887

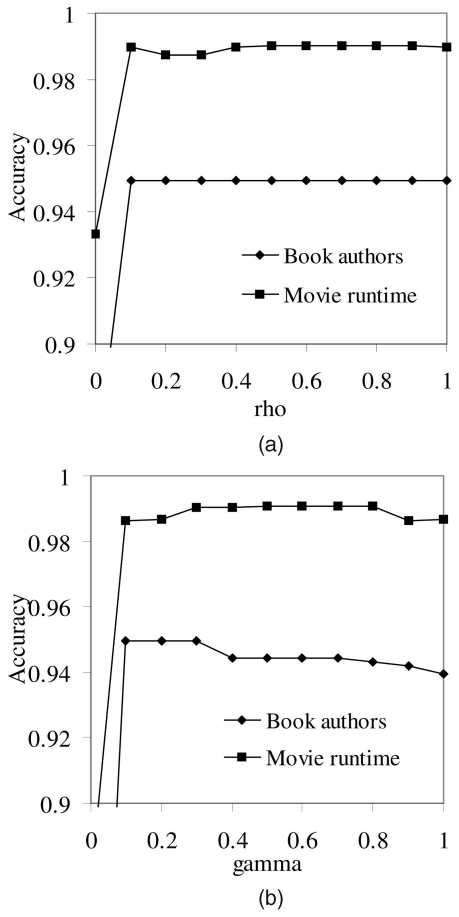

Fig. 10. Accuracy of TRUTHFINDER with respect to ρ and γ. (a) Accuracy with respect to ρ. (b) Accuracy with respect to γ.

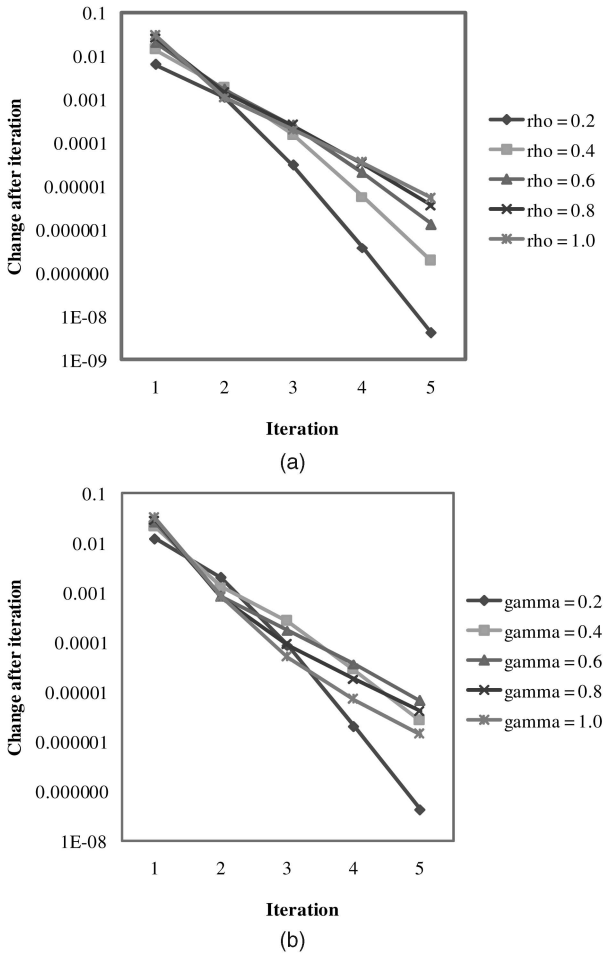

Fig. 11. Relative changes after each iteration with different ρs and γs. (a) Convergence with ρ. (b) Convergence with γ.

on this value range. In the following experiments, TRUTHFINDER does not precompute implications between facts (because they can be easily computed on the fly) and performs five iterations.

We first study the time and space scalability of TRUTHFINDER with respect to the number of facts. The number of websites is fixed at 1,000, and the number of facts varies from 5,000 to 500,000. The results are shown in Fig. 12. Its runtime increases 118 times as the number of objects grows 100 times, which is very close to being linearly scalable, and the minor superlinear part should come from the usage of dynamically growing data structures (e.g., vectors and hash tables). The memory usage is also linearly scalable. In fact, it grows sublinearly, possibly because of the fixed part of memory usage.

Then, we study the scalability with respect to the number of websites, as shown in Fig. 13. There are 10,000 objects and 50,000 facts, and the number of websites varies from 100 to 10,000. It can be seen that the time and memory usage almost remain unchanged when the number of websites grows 100 times, which is consistent with our complexity analysis.

Finally, we study how well TRUTHFINDER can do when the expected trustworthiness varies from zero to one. We compare the accuracies (as defined on the movie runtime data set) of TRUTHFINDER and VOTING in Fig. 14a and the

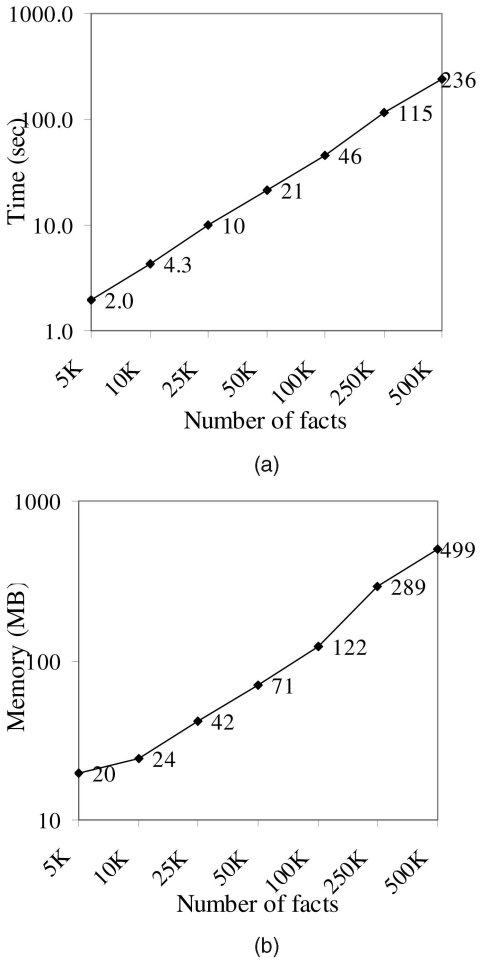

Fig. 12. Scalability of TRUTHFINDER with respect to the number of facts. (a) Time. (b) Memory.

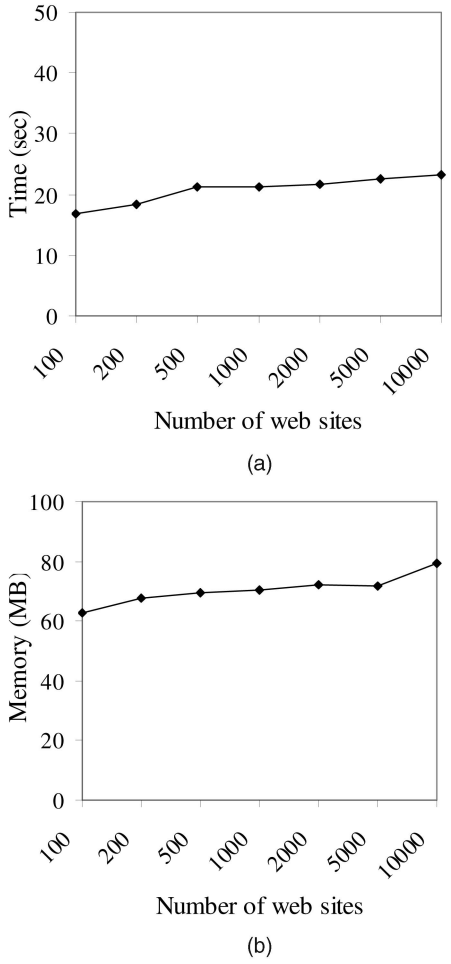

Fig. 13. Scalability of TRUTHFINDER with respect to the number of websites. (a) Time. (b) Memory.

percentages of objects for which exactly correct facts are found in Fig. 14b. When the expected trustworthiness is zero, TRUTHFINDER and VOTING are not doing much better than random. (It can be shown that even if the true value and the fact found are two random numbers between 1,000 and 1,500, the accuracy is as high as 0.87.) However, their accuracies grow sharply as the expected trustworthiness increases. When the expected trustworthiness is 0.5, their accuracy is more than 99.5 percent, and the percentage of exact match for TRUTHFINDER is 98.2 percent, and that for VOTING is 90.5 percent. This shows that TRUTHFINDER can find the true value even if there are many untrustworthy websites.

5 Related Work

The quality of information on the Web has always been a major concern for Internet users [11]. There have been studies on what factors of data quality are important for users [13] and on machine learning approaches for distinguishing high-quality and low-quality web pages [9], where the quality is defined by human preference. It is also shown that information quality measures can help improve the effectiveness of Web search [15].

In 1998, two pieces of groundbreaking work, PageRank [10] and Authority-Hub analysis [7], were proposed to utilize the hyperlinks to find pages with high authorities. These two approaches are very successful at identifying important web pages that users are interested in, which is also shown by a subsequent study [1]. In [3], the authors propose a framework of link analysis and provide theoretical studies for many link-based approaches.

Unfortunately, the popularity of web pages does not necessarily lead to accuracy of information. Two observations are made in our experiments: 1) even the most popular website (e.g., Barnes & Noble) may contain many errors, whereas some comparatively not-so-popular websites may provide more accurate information, and 2) more accurate information can be inferred by using many different websites instead of relying on a single website.

TRUTHFINDER studies the interaction between websites and the facts they provide and infers the trustworthiness of websites and confidence of facts from each other. An analogy can be made between this problem and Authority-Hub analysis, by considering websites as hubs (both of them indicate others' authority weights) and facts as authorities. However, these two problems are very different, and Authority-Hub analysis cannot be applied to our problem. In Authority-Hub analysis, a hub's weight is

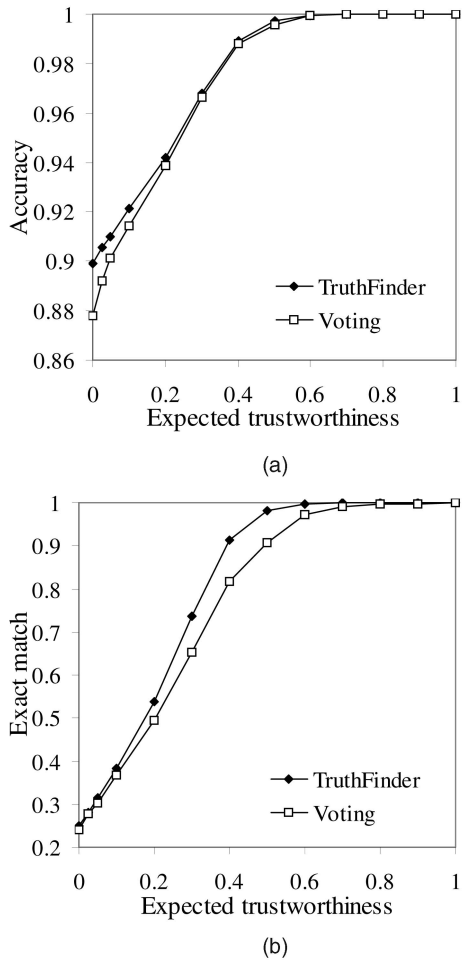

Fig. 14. Noise resistance of TRUTHFINDER and VOTING with respect to website trustworthiness. (a) Accuracy. (b) Percentage of exact match.

computed by summing up the weights of authorities linked to it. This is unreasonable in computing the trustworthiness of a website, because a trustworthy website should be one that provides accurate facts instead of many of them, and a website providing many inaccurate facts is an untrustworthy one. Moreover, the confidence of a fact is not simply the sum of the trustworthiness of the websites providing it. Instead, it needs to be computed using some nonlinear transformations according to a probabilistic analysis.

Another difference between TRUTHFINDER and Authority-Hub analysis is that TRUTHFINDER considers the relationships (implications) between different facts and uses such information in inferring the confidence of facts. This is related to existing studies on inferring similarities between objects using links. Collaborative filtering [4] infers the similarity between objects based on their ratings to or from other objects. There are also studies on link-based similarity analysis [6], [14], which defines the similarity between two objects as the average similarity between objects linked to them. In [5], the authors propose an approach that uses the trust or distrust relationships between some users (e.g., user ratings on eBay.com) to determine the trust relationship between each pair of users.

TRUTHFINDER uses iterative methods to compute the website trustworthiness and fact confidence, which is widely used in many link analysis approaches [5], [6], [7], [10], [14]. The common feature of these approaches is that they start from some initial state that is either random or uninformative. Then, at 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, we adopt it in TRUTHFINDER.

6 CONCLUSIONS

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 TRUTHFINDER, an approach that utilizes the interdependency between website trustworthiness and fact confidence to find trustable websites and true facts. Experiments show that TRUTHFINDER achieves high accuracy at finding true facts and at the same time identifies websites that provide more accurate information.

ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation Grants IIS-05-13678/06-42771 and NSF BDI-05-15813. Any opinion, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

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] X. Yin, J. Han, and P.S. Yu, "LinkClus: Efficient Clustering via Heterogeneous Semantic Links," *Proc. 32nd Int'l Conf. Very Large Data Bases (VLDB '06),* Sept. 2006.
[15] X. Zhu and S. Gauch, "Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web," *Proc. ACM SIGIR '00,* July 2000.

Xiaoxin Yin received the BE degree from Tsinghua University in 2001 and the MS and PhD degrees in computer science from the University of Illinois, Urbana-Champaign, in 2003 and 2007, respectively. He is a researcher at the Internet Services Research Center, Microsoft Research. He has been working in the area of data mining since 2001, and his research work is focused on multirelational data mining and link analysis, and their applications on the World Wide Web. He has published 14 papers in refereed conference proceedings and journals.

Jiawei Han is a professor in the Department of Computer Science, University of Illinois, Urbana-Champaign. He has been working on research into data mining, data warehousing, stream data mining, spatiotemporal and multimedia data mining, biological data mining, information network analysis, text and Web mining, and software bug mining, with more than 350 conference and journal publications. He has chaired or served on many program committees of international conferences and workshops. He also served or is serving on the editorial boards of *Data Mining and Knowledge Discovery*, the *IEEE Transactions on Knowledge and Data Engineering*, the *Journal of Computer Science and Technology*, and the *Journal of Intelligent Information Systems*. He is currently serving as the founding editor in chief of the *ACM Transactions on Knowledge Discovery from Data* and on the board of directors for the executive committee of the ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD). He has received many awards and recognitions, including the ACM SIGKDD Innovation Award in 2004 and the IEEE Computer Society Technical Achievement Award in 2005. He is a senior member of the IEEE and a fellow of the ACM.

Philip S. Yu received the BS degree in electrical engineering from the National Taiwan University, the MS and PhD degrees in electrical engineering from Stanford University, and the MBA degree from New York University. He is a professor in the Department of Computer Science, University of Illinois, Chicago, and also holds the Wexler chair in information and technology. He was the manager of the Software Tools and Techniques Group at the IBM T.J. Watson Research Center. His research interests include data mining, Internet applications and technologies, database systems, multimedia systems, parallel and distributed processing, and performance modeling. He has published more than 500 papers in refereed journals and conference proceedings. He holds or has applied for more than 300 US patents. He is an associate editor of the *ACM Transactions on the Internet Technology* and the *ACM Transactions on Knowledge Discovery from Data*. He is on the steering committee of the IEEE Conference on Data Mining and was a member of the IEEE Data Engineering steering committee. He was the editor in chief of *IEEE Transactions on Knowledge and Data Engineering* from 2001 to 2004, an editor, an advisory board member, and also a guest coeditor of the special issue on mining of databases. He had also served as an associate editor of *Knowledge and Information Systems*. In addition to serving as a program committee member on various conferences, he was the program chair or cochair of the IEEE Workshop of Scalable Stream Processing Systems (SSPS '07), the IEEE Workshop on Mining Evolving and Streaming Data (2006), the 2006 Joint Conferences of the Eighth IEEE Conference on E-commerce Technology (CEC '06) and the Third IEEE Conference on Enterprise Computing, E-commerce and E-services (EEE '06), the 11th IEEE International Conference on Data Engineering, the Sixth Pacific Area Conference on Knowledge Discovery and Data Mining, the Ninth ACM Sigmod Workshop on Research Issues in Data Mining and Knowledge Discovery, the Second IEEE International Workshop on Research Issues on Data Engineering: Transaction and Query Processing, the PAKDD Workshop on Knowledge Discovery from Advanced Databases, and the Second IEEE International Workshop on Advanced Issues of E-commerce and Web-Based Information Systems. He served as the general chair or cochair of the 2006 ACM Conference on Information and Knowledge Management, the 14th IEEE International Conference on Data Engineering, and the Second IEEE International Conference on Data Mining. He has received several IBM honors, including two IBM Outstanding Innovation Awards, an Outstanding Technical Achievement Award, two Research Division Awards, and the 93rd Plateau of Invention Achievement Awards. He was an IBM Master Inventor. He received a Research Contributions Award from the IEEE International Conference on Data Mining in 2003 and also an IEEE Region 1 Award for "promoting and perpetuating numerous new electrical engineering concepts" in 1999. He is a fellow of the ACM and the IEEE.