ISSN 1512-1232

VERACITY FINDING FROM INFORMATION PROVIDE ON THE WEB

[1]D.Vijaya Kumar, [2]B.Srinivasa Rao

[1]Jawaharlal Nehru Technological University Kakinada, India
[2]Department of Computer Science and Engineering of PVPSiddhardha Engineering college,Vijayawada, India

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 uti- lizes the relationships between web sites and their information, i.e., a web site is trustworthy if it provides 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.

Keywords*: Data quality, Web mining, Link analysis.*

1. INTRODUCTION

The world-wide web has become a necessary part of ourlives, 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 interesting 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 ∗The work was supported in part by the U.S. National Science Foundation NSF IIS-05-13678/06-42771 and NSF BDI05-15813. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agencies. ¤Xiaoxin Yin has joined Google Inc. 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 com- parison, 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.

ISSN 1512-1232

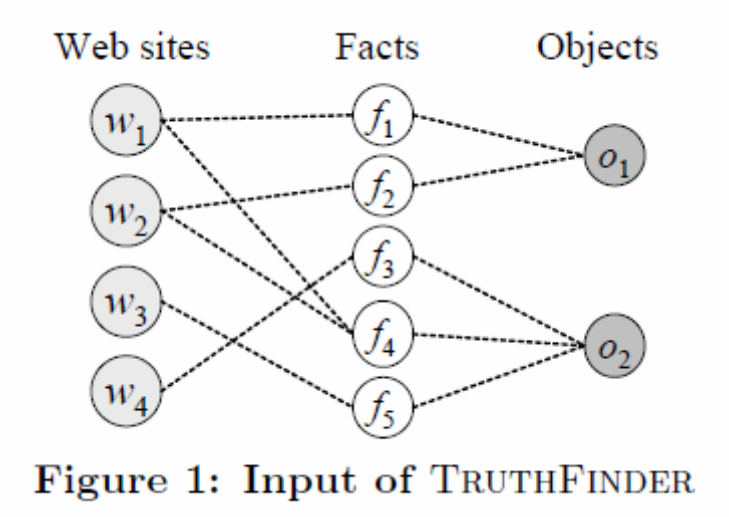

Figure 1: Input of TRUTHFINDER

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 AuthorityHub 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 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 BurnsWendell", 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 describethe problem in Section 2, and propose the computational model in Section 3. xperimental results are presented in Section 4, and we conclude this study in Section 5.

ISSN 1512-1232

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 f1 to f2, imp(f1 ! f2), is f1's influence on f2's confidence, i.e., how much f2's confidence should be increased (ordecreased) according to f1's confidence. It is required that imp(f1 ! f2) is a value between −1 and 1. A positive value indicates if f1 is correct, f2 is likely to be correct. While a negative value means if f1 is correct, f2 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 avail- able, the user can define imp(f1 ! f2) = sim(f1, f2) −base sim, where sim(f1, f2) is the similarity between f1 and f2, and base sim is a threshold for similarity. 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: Ina 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 trust- worthy 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 calcu- lating the average confidence of facts provided by w.

ISSN 1512-1232

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

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

Table 1:Variables and Parameters of Truth Finder

In comparison, it is much more difficult to estimate the confidence of a fact. As shown in Figure 2, the confidence of a fact f1 is determined by the web sites providing it, and other facts about the same object.

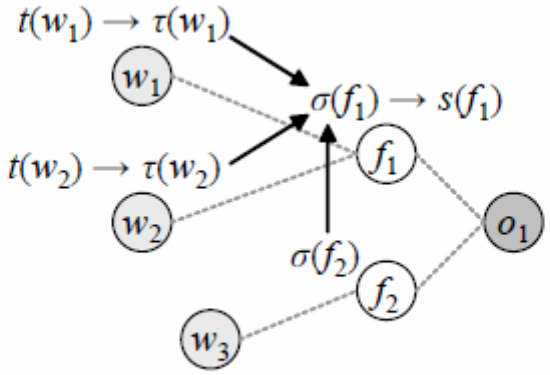

Figure 2: Computing confidence of a fact

Let us first analyze the simple case where there is no related fact, and f1 is the only fact about object o1 (i.e., f2 does not exist in Figure 2). Because f1 is provided by w1 and w2, if f1 is wrong, then both w1 and w2 are wrong. We first assume w1 and w2 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 $(1 - t(w1)) \cdot (1 - t(w2))$, and the probability that f1 is not wrong is $1 - (1 - t(w1)) \cdot (1 - t(w2))$. In general, if a fact f is the only fact about an object, then its confidence s(f) can be computed as

ISSN 1512-1232

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

where W(f) is the set of web sites providing f. In Equation (2), 1 − t(w) is usually quite small and mul-tiplying 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

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

(w) is between and 0 and +1, which better characterizes how accurate w is. For example, suppose there are two web sites w1 and w2 with trustworthiness t(w1) = 0.9 and t(w2) = 0.99. We can see that w2 is much more accurate than w1, but their trustworthiness do not differ much as t(w2) = 1.1 × t(w1). If we measure their accuracy with trustworthiness score, we will find ¿ (w2) = 2×¿ (w1), 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

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

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

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 t0. (t0 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 −! t . If −! t only changes a little after an iteration (measured by cosine similarity between the old and the new −! t), then TruthFinder will stop.

ISSN 1512-1232

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 ½ = 0.5 and 0.3. The maximum difference between two iterations, ±, 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 list- ings (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 cer tain book, suppose the standard fact contains x authors, TruthFinder indicates there are y authors, among which z authors belong to the standard fact. The accuracy of TruthFinder is defined as z max(x,y) .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 differ- ent 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 f1 and f2 is defined in a very similar way as the accuracy of f2 with respect to f1. One important observation is that many bookstores provide incomplete facts, such as only the first author. For example, if a web site w1 says a book is written by “Jennifer Widom”, and another web site w2 says it is written by “Jennifer Widom and Stefano Ceri”, then w1 actually supports w2 because w1 is probably providing partial fact. Therefore, If fact f2 contains authors that are not in fact f1, then f2 is actually supported by f1. The implication from f1 to f2 is defined as follows. If f1 has x authors and f2 has y authors, and there are z shared ones, then imp(f1 ! f2) = z/x − base sim, where base sim is the threshold for positive implication and is set to 0.5.

ISSN 1512-1232

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.

Total Number of Figures: 2
Total Number of Tables: 1

Article received: 2010-08-07