Truth Finder Algorithm for Multiple Conflicting Information Providers on the Web

Viral Panchal
B.E Computer Engg
University of Pune
viral.91@gmail.com

Shailesh Pillai
B.E Computer Engg
University of Pune

Ashutosh Singh
B.E Computer Engg
University of Pune

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.

General Terms
Authority Hub Analysis, Page Ranking Algorithm, Veracity Problem.

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 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 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.

Table 1. Conflicting Information About Book Authors

Website	Authors
A1 Books	Karen Holtzblatt, Jessamyn Burns Wendell, Shelley Wood
Powell's Book	Holtzblatt, Karen
Carnwall Books	Karen Holtzblatt, Jessamyn Burns Wendell, Shelley Wood
Mellon's books	Wendell, Jessamyn
Lakeside books	Wendell, Jessamyn,Holtzblatt, Karenwood, Shelley
Blackwell online	Wendell, Jessamyn, Holtzblatt, Karen, Wood, Shelley
Barnes & Noble	Karen Holtzblatt, Jessamyn Wendell, Shelley Wood

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 , PageRank , and more general link-based analysis. 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.

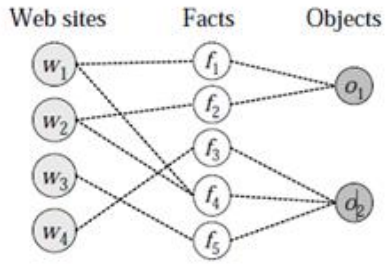

Fig 1.Input of Truthfinder

.

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. 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 describe the problem in Section 1, Experimental results are presented in Section 2, and we conclude this study in Section 3.

2. PROBLEM DEFINITION

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 bject.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 (or decreased) according to f1's confidence. It is required that imp(f1 ! f2) is a value between −1 and 1. A positive value indicates 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.

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

3. 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#).

3.1 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 z max(x,y) .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 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.

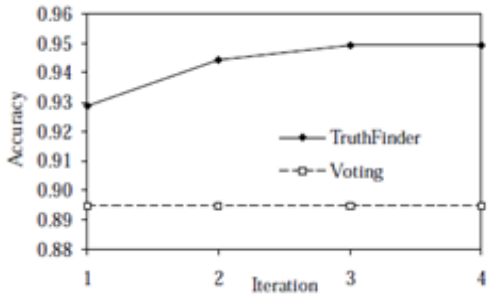

Fig 2. Accuracies of Truthfinder and Voting

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% 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.

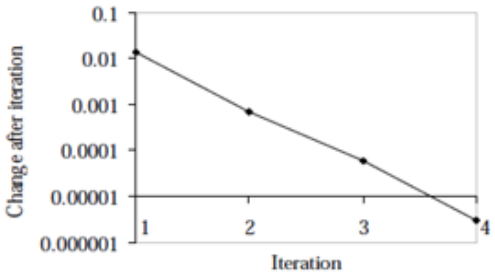

Fig 3.Relative changes of Truthfinder

Figure 3 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 2 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.

Table 2. Compare The Results Of Voting, TruthFinder And Barnes & Nobles

Type of error	VOTING	TRUTHFINDER	Barnes & Nobles
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

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 2 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 2This query was submitted on Feb 7, 2007. TruthFinder. We only consider bookstores that provide at least 10 of the 100 books.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 2 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 2007. TruthFinder. We only consider bookstores that provide at least 10 of the 100 books.

Table 3 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 3) 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.

Table 3. Compare 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
Marondo.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	1	97	0.865
Barnes & Noble	3	42	0.6554
Ecampus.com	11	18	0.847
Average accuracy			0.789

4. 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.

5. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

6. REFERENCES

[1] Xiaoxin Yin(UIUC), Jiawei Han(UIUC) and Philip S. Yu(IBM T. J. Watson Res. Center) www.cs.uiuc.edu/~hanj/pdf/kdd07_xyin.pdf

[2] 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.

[3] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In SODA, 1998.

[4] Princeton Survey Research Associates International. Leap of faith: using the Internet despite the dangers. Results of aNational Survey of Internet Users for Consumer Reports WebWatch, Oct 2005.