ISSN:0975-9646
Sudhakar Katherapaka et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, 1505-1508

Estimating Maximum Confidence and Trustworthiness for Facts for Information Providers on the Web

Sudhakar Katherapaka [1],P.Pradeep Kumar [2,] Manjula Aakunuri[3]
[1]*Department of Computer Science & Engineering, Vivekananda Institute of Tech & Sciences, Karimnagar, JNTUH,Hyderabad, AP, INDIA*
[2]*Prof and HOD,Department of Computer Science & Engineering, Vivekananda Institute of Tech & Sciences, Karimnagar, JNTUH,Hyderabad, AP, INDIA*
[3]*Asst Prof,Department of Computer Science & Engineering, Jyothishmathi Institute of Tech & Sciences, Karimnagar, JNTUH,Hyderabad, AP, INDIA*

***Abstract*---Web mining is the application of data mining techniques to discover patterns from the Web according to analysis targets. Data quality is the quality of data of high quality "if they are fit for their intended uses in operations, decision making and planning" or refers to the degree of excellence exhibited by the data in relation to the portrayal of the actual phenomena.**

***Keywords*--- Data quality, Link analysis, Web mining, Page rank.**

I. INTRODUCTION

The World Wide Web has become the most important information source for most people. Unfortunately, there is no guarantee for the correctness of information on the Web. Moreover, different websites often provide conflicting information on a subject Called Veracity, i.e., conformity to truth, which studies the method of finding the true facts from a large amount of conflicting information on many subjects provided by various websites. A general framework for the Veracity problem solution is the Truth Finder, which uses confidence of facts and trustworthiness of websites. The confidence determines the trustiness of websites and facts it stated. By using the influence among the facts help in increasing the efficiency and reducing the time complexity. Putting this framework into border application like Mass Collaboration, the cooperation between the independent people on single project can be maximized.

Early search engines retrieved relevant pages for the user based primarily on the content similarity of the user query and the indexed pages of the search engines. The retrieval and ranking algorithms were simply direct implementation of those from information retrieval. It became clear that content similarity alone was no longer sufficient for search due to two reasons. First, the number of Web pages grew rapidly during the middle to late 1990s. Given any query, the number of relevant pages can be huge. This abundance of information causes a major problem for ranking, i.e., how to choose only 30–40 pages and rank them suitably to present to the user. Second, content similarity methods are easily spammed. A page owner can repeat some important words and add many remotely related words in the pages to boost the rankings of the pages or to make the pages relevant to a large number of possible queries.

The researchers in academia began to work on the problem. They resort to hyperlinks. Unlike text documents used in traditional information retrieval, which are often considered independent of one another (i.e., with no explicit relationships or links among them), Web pages are connected through hyperlinks, which carry important information. Some hyperlinks are used to organize a large amount of information at the same Website, and thus only point to pages in the same site. Other hyperlinks point to pages in other Web sites. Such out-going hyperlinks often indicate an implicit conveyance of authority to the pages being pointed to. Therefore, those pages that are pointed to by many other pages are likely to contain authoritative or quality information. Such linkages should obviously be used in page evaluation and ranking in search engines

The appearance of the World Wide Web (WWW) at the end of the last century led to a rapid growth in the Internet and in the quantity of accessible information for users. 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.

The 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.

There are often conflicting facts on the Web. There are also many websites, some of which are more trustworthy than others. 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.

There may be some websites which represent the common facts while other may represent some different facts. There may be some websites which provide same facts in different representations or provide partially similar facts as of the other websites which is considered as influence of one fact on the other facts.

1. Scope

Websites are the main source of information providers. The trustworthiness of websites and the confidence of facts are important attributes for considering how much the websites are trustable and how much the facts it provided is correct.

Trustworthiness and Confidence are the useful attributes for calculating

- The websites which provide the correct information.
- Facts that are been provided by many websites.

2. Objective

The objective is to obtain the websites that provide true facts instead of large number of facts and also to obtain facts which have the high confidence value than other facts. The consideration of the influence of one fact on the other which has the same means but in different representation is similar in partial.

II. PROBLEM STATEMENT

The problem can be simply addressed as: The Link-Based approaches considered only the hyperlinks between websites while they do not consider the Influence of one page on the other. One website may be providing facts which are been provided by many of the websites resulting in high confidence for the fact pointed by many websites and vice versa high trustworthiness of website providing it. But they do not consider the interdependency between different facts provided by many websites. The interdependency may be as if the same facts with different represent or one fact contain the other fact, i.e. one fact contains information which the other fact contains and some other information in extra.

The Link-based approach is not considering the interdependencies, the same facts with different representation may be considered as different facts instead as a single and dividing the trustworthiness between these facts result in decrease in trustworthiness of websites pointing to them even both the facts are same.

1. Veracity Problem

Given a large amount of conflicting information about many objects, which is provided by multiple websites (or other types of information providers), how can one discover the true fact about each object? The word "fact" represents something that is claimed as a fact by some website, and such a fact can be either true or false [1].

There are often conflicting facts on the Web. The conflicting information is the relationships between two objects (e.g., authors of books). There are also many websites, some of which are more trustworthy than others.

TABLE I CONFLICTING INFORMATION ABOUT BOOK AUTHORS

Online Store	***Authors***
Powell's books	Holtzblatt, Karen
Barnes & Noble	Karen Holtzblatt, Jessamyn Wendell, Shelley Wood
A1 Books	Karen Holtzblatt, Jessamyn Burns Wendell, Shelley Wood
Cornwall books	Holtzblatt-Karen, Wendell Jessamyn Burns, Wood
Mellon's books	Wendell, Jessamyn
Lakeside books	WENDELL, JESSAMYN HOLTZBLATT, KARENWOOD, SHELLEY

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, one uses an iterative computational method. At each iteration the probabilities of facts being true and the trustworthiness of websites are inferred from each other [2].

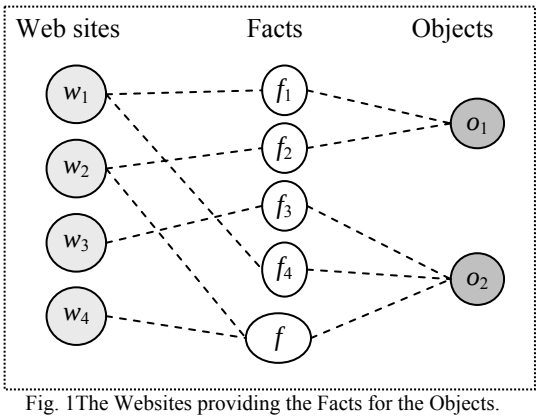

Fig. 1The Websites providing the Facts for the Objects.

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.

2. Trustworthiness and Confidence

The confidence of a fact f is the probability of f being correct, according to the best of our knowledge. The trustworthiness of a website w is the expected confidence of the facts provided by w [3].

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". In order to represent such relationships, the concept of implication between facts been stated. 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 according to f_1's confidence. The definition of similarity 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 [4].

3. Basic Heuristics

- Usually there is only one true fact for a property of an object.
- 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."
- 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.
- In a certain domain, a website that provides mostly true facts for many objects will likely provide true facts for other objects.

III. TRUTH FINDER ALGORITHM

Truth Finder algorithm is generalized framework for evaluating the confidence of facts and the websites

Trustworthiness. The Algorithm states a good websites points to true facts and the true facts are been pointed by many websites.

1. Website Trustworthiness and Fact Confidence

The trustworthiness of a website is just the expected confidence of facts it provides. For website w, its trustworthiness t (w) is the average confidence of facts provided by website.

Analyze the simple case where there is no related fact, and f_1 is the only fact about object O_1 .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. Assume that w_1 and w_2 are independent. Thus, the probability that both of them are wrong is (1- t (w_1)), (1- t (w_2)), and the probability that f_1 is not wrong is 1- (1- t (w_1)). (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 the multiple of all wrong probabilities that point to fact, which is subtracted from 1.

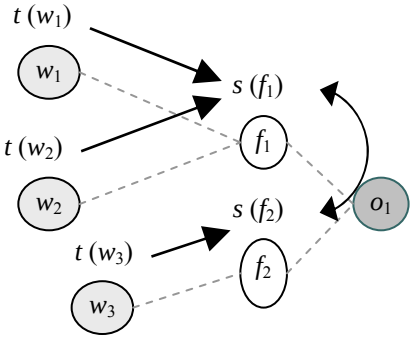

Fig. 2 Computing Confidence of a fact

As 1-t (w) is usually quite small and multiplying many of them may lead to underflow. In order to facilitate computation and veracity exploration, the logarithmic been used and define the trustworthiness score of a website as the negative of logarithmic value of (1-t (w)).

The negative of logarithmic value of 1-t (w) is between zero and infinity, and a larger value indicates higher trustworthiness. Similarly, the confidence score of a fact is defined as the negative of logarithmic value of (1-s (f)).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.

2. Influences between Facts

There are usually many different facts about an object (such as f_1 and f_2 in Fig 3.2), and these facts influence each other. Suppose in Fig 3.2, 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, the confidence score of f_1 should be increased according to the confidence score of f_2, which is the sum of the trustworthiness scores of websites providing f_2.

The adjusted confidence score of a fact f is the sum of Confidence score of the facts multiplied with the influence between the related facts, which totally is multiplied by a weight of an object over the other plus trustworthiness of the same fact [6].

Weight of an object over the other is a parameter between zero and one, which controls the influence of related facts. Adjusted confidence score is the sum of the confidence scores of f, and a portion of the confidence score of each related fact f^i multiplies the implication from f^i to f [5].

The confidence of 'f' based on adjusted confidence score is computed based on confidence score, i.e. the exponential function of the negative value of adjusted confidence score subtracted from 1.

TABLE II VARIABLES AND PARAMETERS

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

3. Handling Additional Subtlety

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, so adding a dampening factor value and redefine fact confidence as the exponential function of the negative value of adjusted confidence score multiplied with dampening factor value subtracted from 1. Where dampening factor value lies between 0 and 1.

The confidence of a fact f can easily be negative if f is conflicting with some facts provided by trustworthy websites with the above equation, which makes adjusted confidence score less than 0 and confidence value become 0. This is unreasonable because the confidence cannot be negative and even with negative evidences, there is still a chance that f is correct, so its confidence should still be above zero. Moreover, if confidence is set to zero, if it is negative according to the adjusted confidence exponential function, this "chunking" operation and the multiple zero values may lead to unstable conditions in iterative computation. Therefore, use of Logistic function, which is a variant of above equation, as the final definition for fact confidence as 1 divided by the exponential function of the negative value of adjusted confidence score multiplied with dampening factor value added with 1 [7].

If multiple of adjusted confidence score and the dampening factor value is significantly less than zero, the confidence value is close to zero but remains positive.

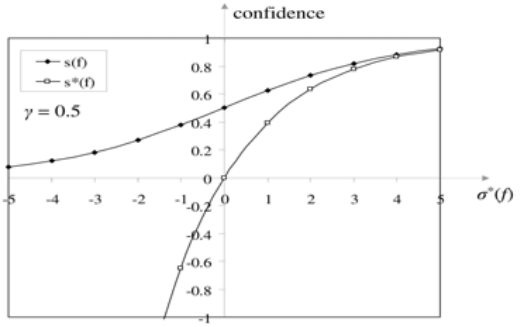

Fig. 3 Two methods for computing confidence

4. Iterative Computation

The website trustworthiness can be inferred from the fact confidence and vice versa. Truth Finder 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, Truth Finder tries to improve its knowledge about their trustworthiness and confidence, and it stops when the computation reaches a stable state. Truth Finder needs an initial state 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 Truth Finder can infer the confidence of facts, which are very meaningful because the facts supported by many websites are more likely to be correct.

In each step of the iterative procedure, Truth Finder first uses the website trustworthiness to compute the fact confidence and then re-computes the website trustworthiness from the fact confidence. Each step only requires two matrix operations and conversions between trustworthiness of website and trustworthiness score and between confidence of fact and adjusted confidence score. 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. Truth Finder stops iterating when it reaches a stable state. The stableness is measured by how much the trustworthiness of websites changes between iterations. If t vector only changes a little after an iteration (measured by cosine similarity between the old and the new vector t), then Truth Finder will stop [8].

5. Complexity Analysis

Analyzing the complexity of Truth Finder, suppose there is 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 two matrices A and B, Each link between a website and a fact corresponds to an entry in A. Thus, A has L entries, and it takes O (L) time to compute A.B contains more entries than A because B_{ji} is nonzero if website w_i provide a fact that is related to fact f_j. Thus, there are O (kL) entries in B. Because each website can provide at most one fact about each object, each entry of B involves only one website and one fact. Thus it still takes constant time to compute each entry of B, and it takes O (kL) time to compute 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. Truth Finder takes O (IkL) time and O (kL+M+ N) space [8].

If in some cases, O (kL) space is not available, discard the matrix operations and compute the website trustworthiness and fact confidence using the equations. If already pre-computed implication lies 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 the implication is not pre-calculated, 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 inter-fact implication. Thus, the overall time complexity is O (IL + IkN).

IV CONCLUTION & FUTURE WORK

The Veracity problem, that aims at resolving the conflicting facts from multiple websites and finding the true facts among them. The Truth Finder algorithm, an approach which uses the interdependency between website trustworthiness and fact confidence finds trustable websites and true facts. Truth Finder resolves the conflicting information provided by many websites and identifies true facts and at the same time identifies websites that provide more accurate information.

Truth Finder uses the interdependencies between the facts and calculates the trustworthiness of websites and the confidence of facts considering the inter-fact dependencies.

Putting this Truth Finder framework into broader application scope like mass collaboration help in resolving the conflict information between the people connected in the mass collaboration. And also collects the interdependencies among the facts if made automation the efficiency in retrieving the true fact also increase and make less involve of human in entering the interdependencies among facts.

REFERENCES

[1] M. Blaze, J. Feigenbaum, and J. Lacy, "Decentralized Trust Management," Proc. IEEE Symp. Security and Privacy (ISSP '96) May 1996.

[2] 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.

[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] J.M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," J. ACM, vol. 46, no. 5, pp. 604-632, 1999.

[6] T. Mandl, "Implementation and Evaluation of a Quality-Based Search Engine," Proc. 17th ACM Conf. Hypertext and Hypermedia, Aug. 2006.

[7] 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.

[8] 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 Web Watch, Oct. 2005.