EVALUATING THE MAXIMUM CONFIDENCE OF FACTS FOR INFORMATION PROVIDED ON THE WEB

[1]B.Venkata Praveen Babu Reddy and [2]K.P. Supreethi

[1]*Software Engg, Dept. of Computer Science, JNTUCEA, Anantapur*
[2]*Asst professor, Dept. of Computer Science, JNTU.*
[1]*praveenbv.mtech@gmail.com* [2]*supreethi.pujari@gmail.com*

ABSTRACT

The World Wide Web is the important information source for us. But, there is no guarantee for the relevance of information retrieved from the Web. The information provided by one Website may be conflicting with the information on the property of an object from other website. In this paper, we study a concept, called Veracity, means, i.e. conformity to truth, which studies how to find true facts from a large amount of conflicting information of an object, provided by various websites. The TCRC (Trustworthiness and Confidence based on Ratio Contribution) algorithm, which consider the relationships between websites and their information and evaluate the confidence of facts and trustworthiness of websites from Ratio Contribution of the facts for that website.

Index Terms :* *WebMining, Graph Theory, Relevant Information.

1. INTRODUCTION

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 information that has accumulated on WWW represents an enormous knowledge base that may prove useful for numerous applications.

Everyday, people retrieve all kinds of information from the Web. For example, 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?"

Example: (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, 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?

In this paper, we study a problem called the Veracity problem, and influence of one fact on other.

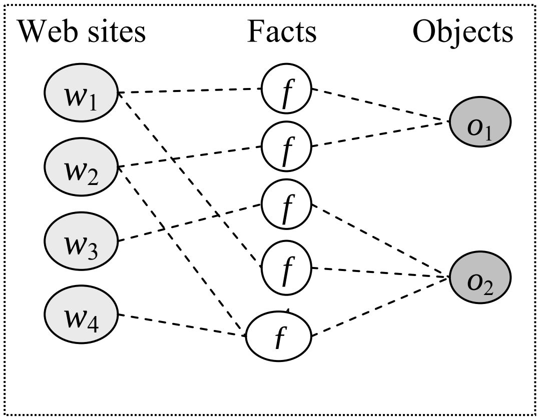

Fig1: Input to TCRC

Given the conflicting information about many objects, which is provided by multiple websites, how can we discover the true fact about each object?

Example: considering the names in the table1, resultant for the query "top ranking batsman in cricket ODI".

The facts in table 1 are conflicting with each other, as some websites provide some players name while others provide some other players name including some names in common with other websites. Thus considering the facts and influence between the facts, trustworthiness of websites and confidence of a fact can me maximized.

Table 1: Conflicting information about the players

websites	*players*
Sheetudeep	**R.T.Pointing, A.Symonds**
Headlinesindia	**M.S.Dhoni, M.E.K Hussey,Yuvraj Singh**
Altiusdirectory	**M.S.Dhoni, C.H.Gayle,M.E.K Hussey**
Thatscricket.oneindia.in	**M.S.Dhoni, M.E.K.Hussey Yuvraj Singh**
All37	**Ricky Pointing, Micheal Hussey, Gambir**

2. PROBLEM DEFINITIONS

In this paper, we study the problem of finding true facts and trustworthiness for websites.

2.1 Basic Definitions

Trustworthiness and Confidence Definition1 (Confidence of facts). The confidence of a fact f is the sum of websites trustworthiness pointing towards it and is denoted by cf(f).

Definition2 (Trustworthiness of websites). The trust worthiness of a website w is the ratio contribution of confidence of the facts which the websites point to and is denoted by tw(w).

The Influence between the facts may exist and is indicated as, if 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 -> 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->f_2) should be high. So the value of imp (f1->f2) lies between 0 to 1 for symmetric facts and negative for conflicting information.

Imp(f_1 -> 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.

So the website providing facts may be having implication influence which may be positive or negative influence. So considering all the facts, influences and using Ratio Contribution we find a solution to solve conflict information,

2.2 Basic Assumptions

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

Example: The captain of Indian cricket team is M.S.Dhoni.

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

Example : “Sachin Tendulkar” and “S.Tendulkar”.

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

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

Example: Wiki

3 COMPUTATIONAL MODEL

If a fact is provided by many trustworthy websites, it is likely to be true; and the website is trustworthy if it provides facts with high confidence.

Table 2 : Variables and parameters

Name	Description
tw(w)	Trustworthiness of website w
cf(f)	Confidence of a fact f.
w	Website
F(w)	Set of facts provided by w
f	Fact
in(f)	Influence score
in*(f)	Adjusted influence score
W(f)	The set of websites providing f
O(f)	The object f is about
Imp(fi->fj)	Influence between the facts
ρ	Weight of object about same object
d	Max difference between two iterations
cf*(f)	Adjusted confidence of fact

3.1 Fact Confidence and Website Trustworthiness

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

3.1.1 Basic Inference

As defined in Definition 2, the trustworthiness of a website is the contribution of facts at instant pointed by the website. For website w, we compute its trustworthiness tw(w) by calculating the contribution of facts for that website:

$$\mathbf{tw(w)=\textstyle\sum_{f\epsilon F(w)} cf(f)^{\wedge 2}/\sum_{f\epsilon F(w)} cf(f)} \qquad (1)$$

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

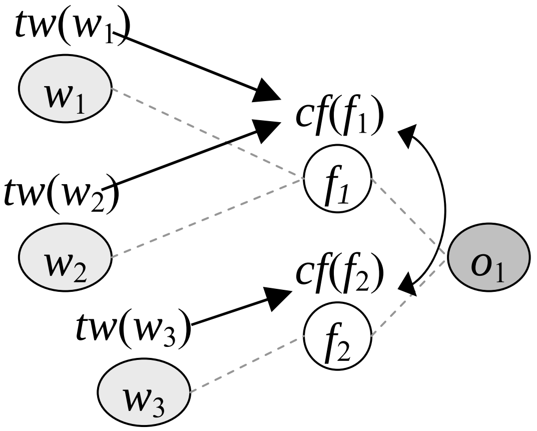

Fig. 2. Computing confidence of a fact.

The confidence of a fact can be calculated based on

$$\mathbf{cf\ (f) = \textstyle\sum_{w\ \epsilon\ W(f)} tw(w).} \ (2)$$

Where W(f) is the set of websites providing f and value of tw(w) in initial is taken as any positive value greater than one.

3.1.2 Influences between Facts

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.

We define the influence score of a fact as

$$\mathbf{in(f)=ln(cf(f))} \qquad (3)$$

Any value of cf(f) if less than 1, we discard the fact as it is having too less confidence value.

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 of websites providing f_2. We define the adjusted influence score of a fact f as

$$\mathbf{in^*(f)= in(f)+\rho. \textstyle\sum_{o(f')=o(f)} in(f'). imp(f'\text{->}f)} \qquad (4)$$

ρ is a parameter between zero and one, which controls the influence of related facts. We can see that in*(f) is the sum of the influence scores of f, and a portion of the influence score of each related fact f' multiplies the implication from f' to f. Please notice that imp(f'-> f) < 0 when f is conflicting with f'.

We use cf*(f) to represent this confidence for in*(f) :

$$\mathbf{cf^*(f)=e^{in^*(\ f)}} \qquad (5)$$

cf*(f) is the adjusted confidence of fact which is assigned back to cf(f).

In each step of the iterative procedure, TCRC first uses the website trustworthiness to compute the fact confidence and then recomputes the website trustworthiness from the fact confidence.

TCRC stops iterating when it reaches a stable state. The stableness is measured by how much the trustworthiness of websites changes between iterations. If tw(w) vector only changes a little after an iteration (measured by cosine similarity between the old and the new tw(w) vector), then TCRC will stop.

Algorithm:

Input: set of Websites, Facts and links between them.

Output: Trustworthiness and Confidence of websites and facts respectively.

```
For each websites
tw(w) =x,  // where x is default initial value;
repeat
cf(f)= $\sum_{w\ \epsilon\ W(f)}$ tw(w).//confidence
  if influence between facts exist then
repeat
in(f)=ln(cf(f))
in*(f)= in(f)+ρ. $\sum_{o(f')=o(f)}$in(f'). imp(f'->f)
until (number of facts)
   cf*(f)=$e^{in^*(f)}$     //trustworthiness
   For each fact
cf(f)=cf*(f)  // adjusted confidence
  tw(w)=$\sum_{f\epsilon F(w)}$ cf(f)$^{\wedge 2}$ /$\sum_{f\epsilon F(w)}$cf(f)
until (cosine similarity of iterations greater than 1-d)
```

4. IMPLEMENTATION DETAILS

Let us consider an example as in fig 3, and assume the initial trustworthiness of every website equally likely to be 20, the value of imp (f2->f1) be 0.5, difference between two iterations be less than 0.05 and value of ρ=0.5.

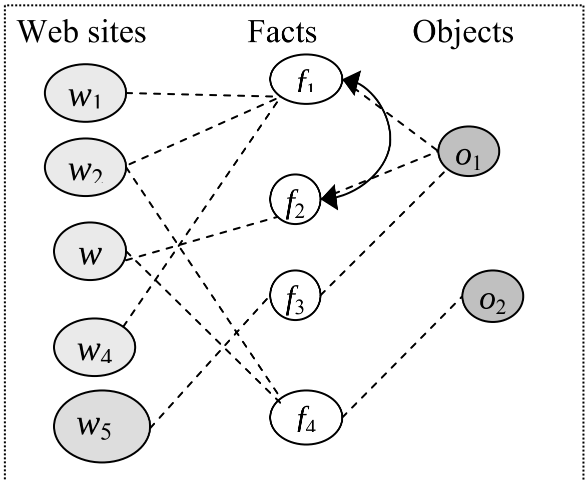

Fig: 3 An example for TCRC algorithm

Then calculating the trustworthiness of websites gives the values as tw(w1)=60, tw(w2)=52, tw(w3)=49.253, tw(w4)= 60, tw(w5)=20, and confidence of facts as cf(f1)= 60,cf(f2)=55.877, cf(f3)=20, cf(f4)=40.where as if average is taken we may get tw(w3)=47.938, using TCRC we can increase the trustworthiness and confidence of both websites and facts.

Example : Considering Table1, the websites.

Thatscricket.oneindia,com, Altiusdirectory.com, Headlinesindia, gives much relevant information for the facts they provide.

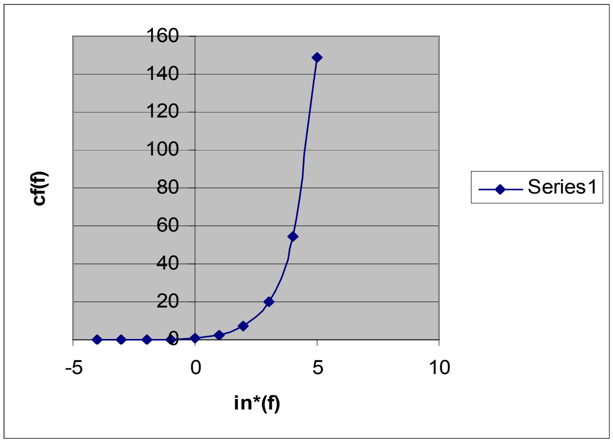

Fig4: Confidence vs. Adjusted Influence score.

For any value of influence score the confidence of a fact is greater than zero. So even for negative influence score the confidence score is positive and greater than but not zero.

5. CONCLUSION

In this paper, we study and formulate the Veracity problem, which aims at resolving conflicting facts from multiple websites and finding the true facts among them. We propose TCRC, an approach that utilizes the interdependency between website trustworthiness and fact confidence to find trustable websites with ratio contribution and true facts.

6. REFERENCES

[1] Xiaoxin Yin, Jiawei Han, Philip S.Yu .”Truth Discovery with Multiple Conflicting Information Providers on the Web”

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

[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.