International Journal of advancement in electronics and computer engineering (IJAECE)
Volume 2, Issue 11, January 2014, pp.271-275, ***ISSN 2278 -1412***

Extracting Trustworthy Data from Multiple Conflicting Information using Semi-Supervised Approach

Priti R. Sharma[1], Mr. Manoj E. Patil[2]

[1]*ME (CSE) scholar, SSBT's ,COET Bambhori , Jalgaon,pritirsharma@gmail.com, India;*
[2]*Dept of Computer Engineering, SSBT's , COET Bambhori ,, mepatij@gmaill.com, India;*

Abstract – *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 huge knowledge base that may prove useful for various applications. Now day for everything people depend on the Web. But information retrieved is not guaranteed trustworthy, not always trustable .Truthfinder is used to find trustworthy data using unsupervised approach. It is assumed that a fact provided by more sources is more likely to be correct. The proposed system based on the semi supervised approach. This uses some training data which will help for improving accuracy.*
Experimental results show that the proposed algorithm is more effective than unsupervised approach and which help to resolve veracity problem.

Keywords: *Truth discovery, semi supervised;*

I. Introduction

The information that has accumulated on WWW represents huge knowledge base that may prove useful for various applications. Now day for everything people depend on the Web. The World Wide Web has become a necessary part of our lives and might have become the most important information source for most people. Every day, 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 they want to know the answer to a certain question, they go to Ask.com and google.com. Now days for getting updates, reading news people always take the help of websites. e.g. if a person want to get the knowledge of 'C' language then also website play very important role."Is the World Wide Web always trustable?" Unfortunately, the answer is "no." There is no guarantee for the correctness of information on the Web. As website use is most popular then data copying on we also popular. Conflicting information present on web is the common thing now days. User cannot get the relevant accurate data most of the time. More problems has to face in searching data for facts changing with time, since out-of-date information often exists in more web sites than up-to-date information[10].

A new problem called the Veracity problem is recognized now days, 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 to discover the true fact about each object [1]. The trustworthiness problem of the web. According to a survey on credibility of web sites[1]:1.54% of Internet users trust news web sites most of time.2.26% for web sites that sell products.3. 12% for blogs Existing system tried to resolve this problem with unsupervised approach. TruthFinder is used to resolve veracity problem. Truthfinder is used to get

the true facts from conflicting information .But better solution is to go for some kind for supervision. i.e. result can be improved when we use semi supervise approach instead of unsupervised approach. Proposed system is based on semi supervised approach. We come to recognize that the even a small amount of training data also greatly help for improving the performance.

II. Related Work

Quality assessment is important is information retrieval.The research on Ranking algorithm is going on from many years.some researchers used link analysis I their ranking algorithms. The popularity of link analysis and the assumptions about its role in search engine rankings has led to great efforts by search engines optimization professionals[3, 5]. Many links are set only to gain influence on the ranking of certain pages. The objective of these approaches is to ground the quality evaluation on a broader knowledge base. Not only the pages often cited by Web page authors but also pages often visited should be regarded as being of high quality. The most popular algorithm is Page Rank. The basic assumption of Page Rank and similar approaches is that the number of in- or back-links of a Web page can be used as a measure for the popularity and accordingly for the quality of a page[5]. Page Rank assigns an authority value to each Web page which is primarily a function of its back links. Additionally, it assumes that links from pages with high authority should be weighed higher and should result in a higher authority for the receiving page. The algorithm is carried out iteratively until the result. However, link analysis has several serious shortcomings. The number of in-links for Web pages follows a power law distribution. In such a distribution, the median value is much lower than the average. This means, that many pages have few in-links while few pages have an extremely high number of in-links. This finding indicates that Web page authors choose the Web sites they link to without a thorough quality evaluation. The Page Rank technique, introduced by Page et al. [8], actually tried to mend this problem by looking at the importance of a page in a recursive manner: "a page with high Page Rank is a page referenced by many pages with high Page Rank". Drawback of page ranking: Any website having higher visits will be having higher page rank irrespective of the content in data, so there will be problem if website having higher page rank and giving wrong data, many user will get confused or blindly believe conflicting or wrong data from other website. Page ranking is use to find pages with high authorities. The user predicts the true information according to the ranking of page.

Finally observation is that it is not true that most popular website will provide the true fact. This assumption is totally wrong because out of date information is remain present on many websites. Hubs and Authorities (Kleinberg, 1999) gives each page a hub score and an authority score, where its hub score is the sum of the authority of linked pages and its authority is the sum of the hub scores of pages linking to it[2].After that the HITS algorithm is proposed by Kleinberg in 1988. HITS algorithm identifies two different forms of Web pages called hubs and authorities. Authorities are pages having important contents. Hubs are pages that act as resource lists, guiding users to authorities. Thus, a good hub page for a subject points to many authoritative pages on that content, and a good authority page is pointed by many good hub pages on the same subject. HITS algorithm is ranking the web page by using in-links and out-links of the web pages. The veracity problem i.e. the conformity of truth is not resolved in above algorithm. Research work was to find trustworthy data. Voting is the easiest way of computing the best fact related to an object is to choose the one with maximum number of votes. This method, however, does not provide good accuracy.Yin, Han [4] first propose Truth finder algorithm to find true facts. Truth Finder studies the interaction between website and the facts they provide and infer the trustworthiness of websites and confidence of fact from each other.

III. Existing System

III.1.Unsupervised Approach

Existing System uses unsupervised approach to resolve veracity problem. The first algorithm proposed to find true fact from false fact is Truth Finder which is unsupervised [7]. It is assumed that a fact provided by more sources (especially more trustworthy and more independent sources) is more likely to be correct. They all use iterative approaches, which start by assigning the same trust worthiness to all data sources, and iterate by computing the confidence of each fact and propagating back to the data sources. The input of TRUTHFINDER is a large number of facts in a domain that are provided by many websites specified in fig 1. 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.

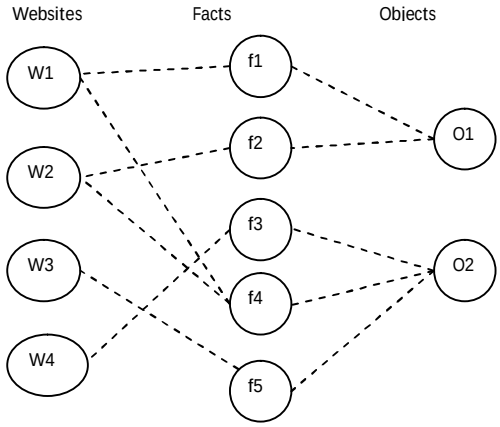

Fig 1: Facts, Objects and Information Providers (Websites)

III.2 Drawbacks of Truth finder

Truth finder makes some assumptions.
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.
It is assumed that an object is associated with only one type of fact.
The above assumption does not true every time. Accuracy depends on these assumptions.

Drawback
It is assumed that a provider either provides good facts for every object or bad facts for every object.
It is assumed that there is only one true value for an object. But often, multiple values (set) could be true sometimes with different degrees of truth.
It is assumed that number of providers providing the true fact is much more than the number of providers providing the bad fact.

IV. Proposed System

We can say that since Truth Finder is unsupervised approach no guaranteed accuracy is given by it. Some level of supervision can help the iterative fact finding algorithm in right direction. So its better solution to go for semi supervised approach.
The approach is based on three principles:
(i) facts provided by the same data source should have similar confidence scores, (ii) similar (and therefore mutually supportive) facts should have similar confidence scores and (iii) if two facts are conflicting, they cannot be both true.

IV.1 System overview architecture

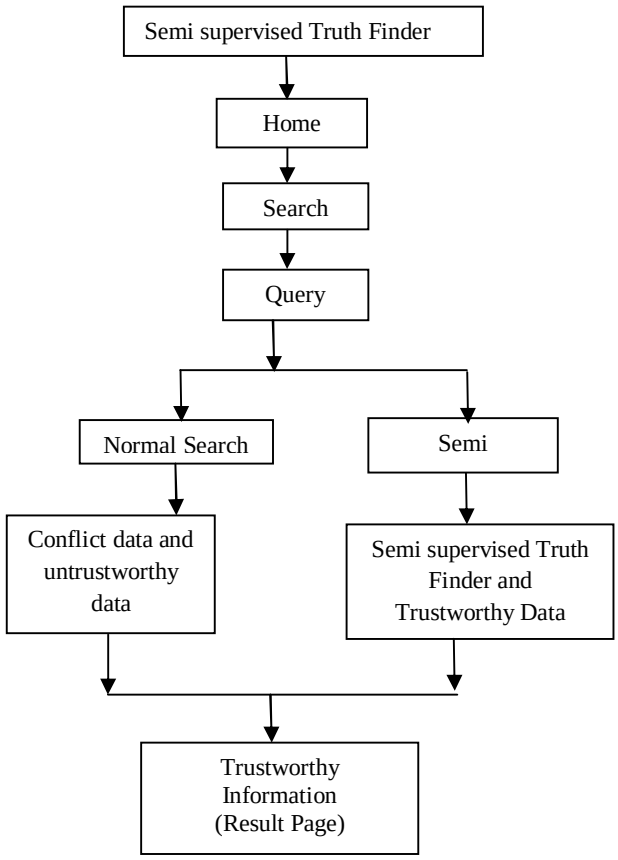

Fig 2: System overview architecture

In semi supervised Truth Finder, a confidence score is assigned to each fact, so that true facts have higher scores than false facts. Ground data are used to retrieve query result from Data Source. Maintaining the ground truth facts sites as a small data sets.
In normal search all related websites are listed without any priority.
In Semi supervised Search, we have to provide additional key words with the normal search keyword. The ranked websites are displayed after comparing with the ground truth fact data set. Ground truth data contain a small set of highly confident fact and use them to infer the trustworthiness of data sources and confidence of facts.
For semi supervised search, the confidence score will be calculated by the site visited by the user. Comparing with the ground truth facts sites, if the visited site is same as the ground truth fact site it'll be consider as true fact site and its value is 1. if it is not same then it'll be considered as false facts site

and its value is -1. If the user doesn't visit the site it'll be taken as unknown site and its value is 0.

V. Experimental result

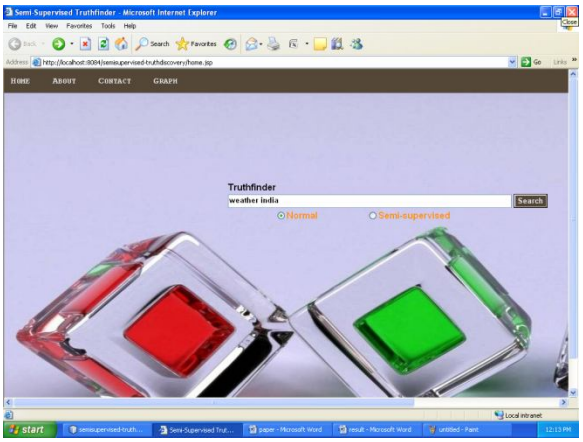

Fig 3 keyword searching for normal search

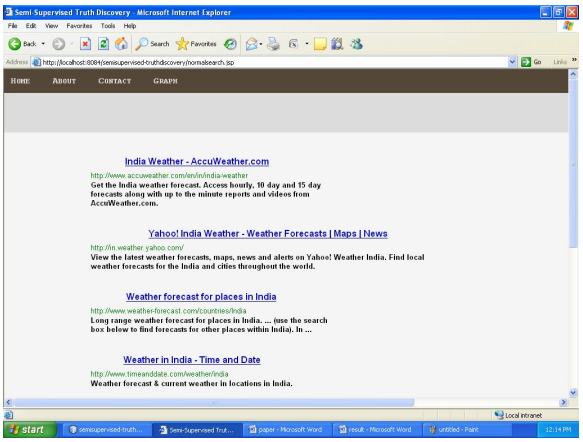

Fig 4 Result of normal search for 'weather india'

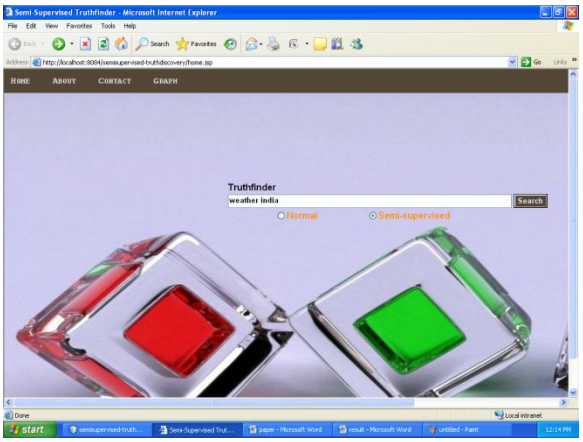

Fig 5 Keyword searching for Semi supervised search

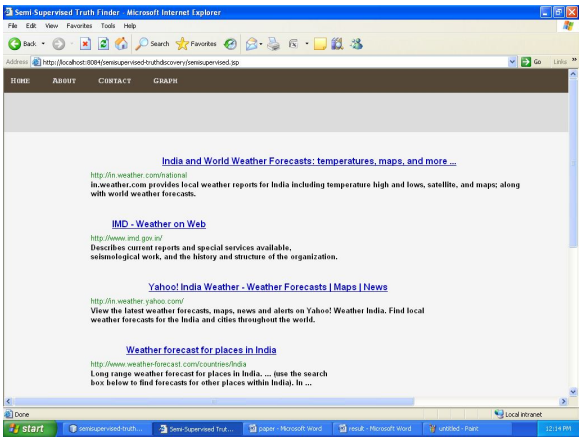

Fig 6 Result of semi supervised search for 'weather india'

VI. Conclusion

Existing system make use of unsupervised approaches. It is assumed that a fact provided by more sources is more likely to be correct. So, there is problem for differencing between true and false facts using only the data itself. In semi-supervised approach that finds true values with the help of a small amount of ground truth data. Using keyword label search we can find truth data from untruth data. In semi supervised search the result is improved as compared to unsupervised Truth finder, since it contain training data.

References

[1] Xiaoxin Yin, Jiawei Han, Philip S. Yu," Truth Discovery with Multiple Conflicting Information Providers on the Web" San Jose, California, USA KDD'07, August 12–15, 2007

[2] Jeff Pasternack Dan Roth, "Knowing What to Believe (when you already know something)." Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 877–885, Beijing, August 2010

[3] Jiawei Han, "Mining Heterogeneous Information Networks by Exploring the Power of Links." Springer-Verlag Berlin Heidelberg 2009

[4] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques."

[5] Lise Getoor ,"Link Mining: A New Data Mining Challenge", SIGKDD Explorations. Volume 4, Issue 2 - page 1

[6] T.Mandl, Implementation and Evaluation of a Quality-Based Search Engine, proc. 17th ACM conference hypertext and hypermedia, 2006.

[7] Xiaoxin Yin, Jiawei Han, Philip S. Yu, "Truth Discovery with Multiple Conflicting Information Providers on the Web", IEEE transactions on knowledge and data engineering, vol. 20, no. 6, June 2008

[8] Ricardo Baeza-Yates,Paolo Boldi,Carlos Castillo,"Generalizing PageRank: Damping Functions for Link-Based Ranking Algorithms", SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.

[9] Xiaoxin Yin, Wenzhao Tan" Semi-Supervised Truth Discovery" WWW 2011, March 28–April 1, 2011, Hyderabad, India.ACM 978-1-4503-0632-4/11/03.

[10] R. Guha , Ravi Kumar, Andrew Tomkins, "Propagation of Trust and Distrust", ACM 158113844X/04/0005.

Author's profile

Priti R Sharma *is research scholar at SSBT COET Bhambhori, Jalgaon (MS.).Her area of interest are data communication, information theory and computing techniques.*

Manoj E. Patil *is currently working as a assistant professor in SSBT COET Bhambhori, Jalgaon(MS.). his area of interest are data communication, information theory and computing techniques.*