International Journal of Communication Networks and Security

Volume 2 | Issue 3 Article 10

January 2014

EXTRACTING ACCURATE DATA FROM MULTIPLE CONFLICTING INFORMATION ON WEB SOURCES

AKSHATA ANGADI
Computer Science and Engineering Department, K.L.E.I.T.,India, akshata_angadi@yahoo.co.in

KARUNA GULL
K.L.E.I.T. Hubli, India, karuna7674@gmail.com

PADMASHRI DESAI
Computer Science and Engineering Department, B.V.B.C.E.T. Hubli, India, padmashri@bvb.edu

EXTRACTING ACCURATE DATA FROM MULTIPLE CONFLICTING INFORMATION ON WEB SOURCES

AKSHATA ANGADI[1], KARUNA GULL[2], PADMASHRI DESAI [3]

[1,3]Computer Science and Engineering Department, [1]K.L.E.I.T. , [#3] B.V.B.C.E.T. Hubli, India
[2]K.L.E.I.T. Hubli, India
E-mail: [1]akshata_angadi@yahoo.co.in, [3]padmashri@bvb.edu, [2]karuna7674@gmail.com

Abstract- For The World-Wide Web has become the most important information source for most of us. As different websites often provide conflicting information there is no guarantee for the correctness of the data. Among multiple conflict results, can we automatically identify which one is likely the true fact?, In this paper our experiments show that Fact finder, a supporter for user to resolve the problem, successfully finds true facts among conflicting information, and identifies Trust worthy websites better than the popular search engines. In our paper we give ratings based on two things- popularity or the hits & number of occurrences of same data. As we can't give preference only to popularity, we have considered another rating i.e. about number of occurrences of same data in several other websites, which are less popular. This paper helps user to get resolved by conflicting facts from multiple websites on two basis. Further by considering few more relations we can develop a search engine that truly helps the user to resolve the Veracity problem.

I. 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. When we want to know the answer to any certain question, we go to ask.com or google.com."Is the World-Wide Web always trustable…?" Unfortunately the answer is "NO". Different Websites often provide conflicting Information, as shown in the following examples….
Example 1: Height Of The Mount Everest:

Suppose a user is interested in how high the 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 8850m, five websites say 8849.868 feet, one says 8848 feet. Each object has a set of conflictive facts. And each web site provides some facts. Which answer should the user trust…?

TABLE 1:
CONFLICTING INFORMATION ABOUT HEIGHT OF MOUNT EVEREST.

Website Name	Height (m)
en.wikipedia.com	8850m
www.britannica.com	8849.868m
geography.about.com	8849.868m
wiki.answers.com	8848m

Top ranked websites are usually the most popular ones. But popularity doesn't mean accuracy.

For example: According to above set of information about height of mountain, websites ranked on top by Google contain conflicts about the correct information. In comparison of websites, some small websites (i.e. britannica.com, geography.about.com) provide accurate information based on our experiments.
Example 2: Author of Books:

According to Table.2. an experiment on who wrote the book Rapid Contextual Design(ISBN: 0123540518), In set of authors information, bookstores ranked on top by Google i.e. (Powell's books) contains error on book author information. In comparison, some small bookstores (i.e. A1 books) provide accurate information.

We tried to find out we found many different sets of authors from different online book stores.

TABLE 2:
CONFLICTING INFORMATION ABOUT BOOK AUTHORS.

Websites	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

Trustworthiness of the Web

i) The trustworthiness problem of the web. According to a survey on credibility of web sites [1] as shown in fig.1.:

- 54% of Internet users trust news web sites most of time.
- 26% for web sites that sell products.
- 12% for blogs.

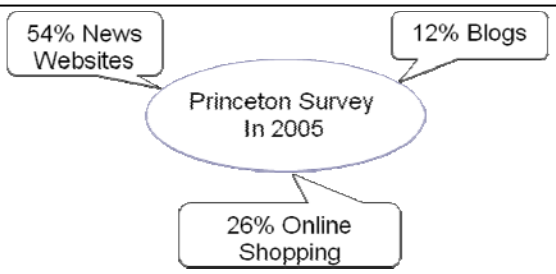

Fig.1. Survey on credibility of web sites

ii) The problem of Veracity: Conformity to truth

- Given a large amount of conflicting information about many objects, provided by multiple web sites.
- How to discover the true fact about each object?

A new problem called Veracity problem, which is formulated as follows:

Given a large amount of conflicting information about many objects as shown in fig.2., 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.

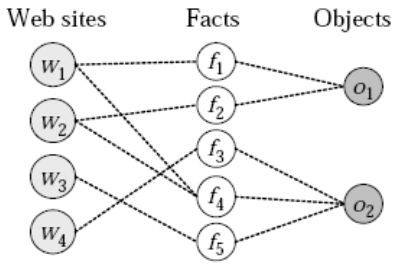

Fig.2. Input to the TruthFinder

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 trust worthiness of web sites are inferred from each other [2]. 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. 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.

Heuristics:
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. We assume that there is only one true fact for a property of an object. The case of multiple true facts will be studied in our future work.
Heuristic 2: This true fact appears to be the same or similar on different web sites. 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."

Heuristic 3: The false facts on different web sites 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. 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.

For example, Height of Mount Everest, the first real data set contains the set of website list which has been extracted from the Google. Table 1 contains a list of website names and the height information extracted from those websites. The proposed system extracts the values given in websites in one particular unit of measurement (in our e.g. meters). Ratings are calculated on two things i) popularity/hits ii) number of occurrence of the same value in different sites. Lastly we calculate average of those and give a rating for al websites.
In summary, we make three major distributions 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 websites, confidence of facts, and influences between facts. Finally, we propose an algorithm for identifying true facts using iterative methods.

Our experiments show that Fact Finder achieves accuracy in discovering true facts based on rating given to websites. In our experiment we mainly consider two ratings i.e. Based on websites popularity & Based on the number of occurrences. Here popularity means number of hits given by users. In our sample experiment we are going to take an average of both the ratings and specify which is having a high rating in tabular column. By which we can say our system can select better trustworthy websites than authority-based search engines such as Google.

A. Web Mining

Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. (Mining means extracting something useful or valuable from a baser substance, such as mining gold from the earth.) Web mining is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign.

The rest of the paper is organized as follows: We describe We discuss related work in Section 2. The problem statement in Section 3 and in Section 4 we added the system analysis. System Implementation is described in Section 5. In Section 6 Experimental results are presented and lastly we have concluded this study in Section 7.

II. RELATED WORK

The quality of information on the Web has always been a major concern for Internet users [1]. There have been studies on what factors of data quality are important for users [3] and on machine learning approaches for distinguishing high-quality and low-quality web pages [4], where the quality is defined by human preference. It is also shown that information quality measures can help improve the effectiveness of Web search [5]. In 1998, two pieces of groundbreaking work, PageRank [6] and Authority-Hub analysis [7], were proposed to utilize the hyperlinks to find pages with high authorities. These two approaches are very successful at identifying important web pages that users are interested in, which is also shown by a subsequent study [8]. In [9], the authors propose a framework of link analysis and provide theoretical studies for many link-based approaches. Unfortunately, the popularity of web pages does not necessarily lead to accuracy of information. Two observations are made in our experiments: 1) even the most popular website (e.g., Barnes & Noble) may contain many errors, whereas some comparatively not-so-popular websites may provide more accurate information, and 2) more accurate information can be inferred by using many different websites instead of relying on a single website. Truthfinder studies the interaction between websites and the facts they provide and infers the trustworthiness of websites and confidence of facts from each other. An analogy can be made between this problem and Authority- Hub analysis, by considering websites as hubs (both of them indicate others' authority weights) and facts as authorities. However, these two problems are very different, and Authority-Hub analysis cannot be applied to our problem. In Authority-Hub analysis, a hub's weight is computed by summing up the weights of authorities linked to it. This is unreasonable in computing the trustworthiness of a website, because a trustworthy website should be one that provides accurate facts instead of many of them, and a website providing many inaccurate facts is an untrustworthy one. Moreover, the confidence of a fact is not simply the sum of the trustworthiness of the websites providing it. Instead, it needs to be computed using some nonlinear transformations according to a probabilistic analysis. Another difference between truthfinder and Authority-Hub analysis is that truthfinder considers the relationships (implications) between different facts and uses such information in inferring the confidence of facts. This is related to existing studies on inferring similarities between objects using links. Collaborative filtering [10] infers the similarity between objects based on their ratings to or from other objects. There are also studies on link-based similarity analysis [11], [12], which defines the similarity between two objects as the average similarity between objects linked to them. In [13], the authors propose an approach that uses the trust or distrust relationships between some users (e.g., user ratings on eBay.com) to determine the trust relationship between each pair of users. Truthfinder uses iterative methods to compute the website trustworthiness and fact confidence, which is widely, used in many link analysis approaches [13], [11], [7], [6], [12]. The common feature of these approaches is that they start from some initial state that is either random or uninformative. Then, at each iteration, the approach will improve the current state by propagating information (weights, probability, trustworthiness, etc.) through the links. This iterative procedure has been proven to be successful in many applications, and thus, we adopt it in Fact finder

III. PROBLEM DEFINITION

To design a system which finds true facts among conflicting information, and identifies Trust worthy websites better than the popular websites. In this we assign ratings based on two things- popularity or the hits & number of occurrences of same data. As we can't give preference only to popularity, we have considered another rating i.e. about number of occurrences of same data in several other websites, which are less popular.

Further by considering few more relations we can design a search engine that truly helps the user to resolve the Veracity problem.

IV. SYSTEM ANALYSIS

A. Existing System

- Page Rank and Authority-Hub analysis is to utilize the hyperlinks to find pages with high authorities.
- These two approaches identifying important web pages that users are interested in, Unfortunately, the popularity of web pages does not necessarily lead to accuracy of information

B. Disadvantage

- The popularity of web pages does not necessarily lead to accuracy of information.
- Even the most popular website may contain many errors.
- Where as some comparatively not-so-popular websites may provide more accurate information.

C. Proposed System

- 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 websites, confidence of facts, and influences between facts.
- Finally, we propose an algorithm for identifying true facts using iterative methods.

The use case diagram of our proposed system is shown in Fig.3.

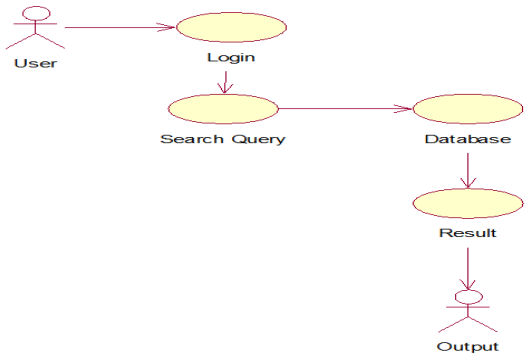

Fig.3. Diagram of Proposed System

D. Advantage

- Our experiments show that Fact Finder achieves very high accuracy in discovering true facts.
- It can select better trustworthy websites than authority-based search engines such as Google.

V. SYSTEM DESIGN

1) Login Module

This module validates the user name and password in login page. Here only the authorized user can use the Fact Finder.

2) Data Search:

Searching the related data link according to user input. In this module user retrieve the specific data about an object

3) Collection Of Data:

Next we have to collect the specific data about an object and it is stored in related database. Create table for specific object and store the facts about a particular object.

4) Truth Algorithm:

We design a general framework for the Veracity problem, and design an algorithm called Truth Finder, 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.

5) Result Calculation:

For each response of the query we are calculating the Performance. Using the count calculated find the best link and show as the output.

All these modules are shown in fig.4, fig.5 and fig.6 using detailed use case, collaboration and class diagrams.

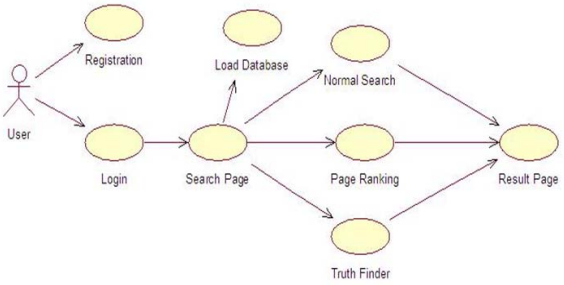

Fig. 4. Detailed Use Case Diagram

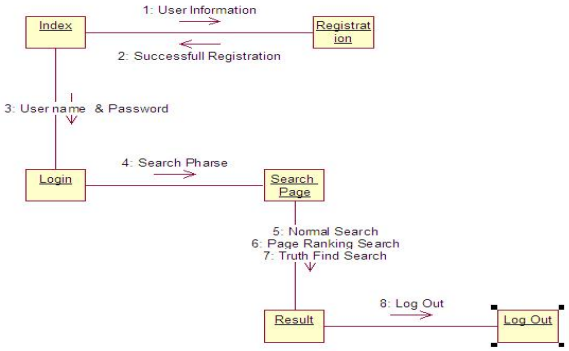

Fig. 5. Collaboration Diagram

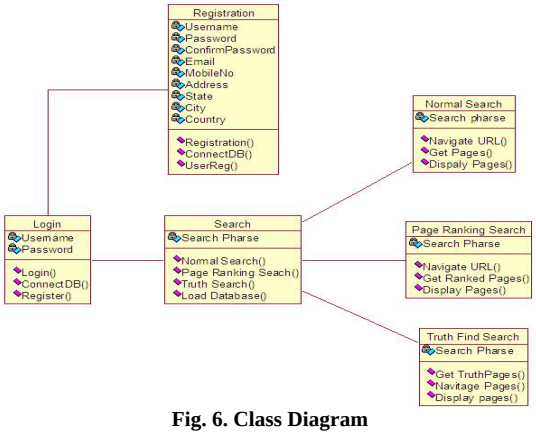

Fig. 6. Class Diagram

VI. SYSTEM IMPLEMENTATION

A. Experimental Setup

We are implementing using VB.net and running it on a Pentium – V with 1GB of RAM and 200 GB Hard disk. The Operating system used is Windows XP. The server side script is written in VB.net and database creator and connector used is MySQL 5.0 and ODBC connector

B. Implementation includes 5 modules/steps:

1) Login Module:

This module validates the user name and password in login page. Here only the authorized user can use the Fact Finder. The user-Id and password is authenticated, that is checked with stored user name and password to allow only the legitimate user to access the account. If the user is not legitimate a message box (or alert window) is displayed saying its "invalid user" and the value in the text box is cleared.

2) Search module:

The time the query is submitted to search, the query written in the text box gets copied into the Google search box .When the search button is clicked on the main page, the query written in the textbox gets executed. And the search results are obtained in the background of the main page.

3) Extract module:

When the extract button in the main page is clicked after search, the domain name and the values of the results are separated. This is done as follows, the search results will be in the form of lists a pre-ordered list rather, now the first list is extracted and the domain part is extracted from list and is split to get required URL copied into the rating page. Similarly the related information is copied into specified location in rating page.

4) Extract results module:

After the domain part and the values gets extracted we need to click on the extract value button. Here the domain name, which is extracted in the previous routine, the query entered, in the text box and the values of the results are displayed in appropriate columns created in the rating page.

```
Private Sub btnCalculate_Click(ByVal sender As System.Object, ByVal e As System.EventArgs) Handles btnCalculate.Click

Dim i, cnt As Integer
dgvRatings.Rows.Clear()
cnt = 1
For i = 0 To dgvExtracted.Rows.Count – 2
'column1=count column2=domainname column3=entered_query col4=result
dgvRatings.Rows.Add()
dgvRatings.Rows(i).Cells(0).Value = cnt
dgvRatings.Rows(i).Cells(1).Value = dgvExtracted.Rows(i).Cells(1).Value
dgvRatings.Rows(i).Cells(2).Value = Query
dgvRatings.Rows(i).Cells(3).Value = GetValue(dgvExtracted.Rows(i).Cells(2).Value)
cnt += 1
Next
End Sub
```

5) Rating module:

After the value gets extracted, the rating to individual website is provided based on the popularity and based the number of websites providing same fact about the object. Here we considered the 10 results from the search engine, as the Google search engine displays the results based on the popularity, popularity based rating is provided based on its occurrence in the results, that is first domain name in the results is given highest rating and rating decreases thereafter .Now the number of websites providing same fact is done by comparing fact with every other website's fact about same object and rating based on it provided with most occurrences given highest rating. For instance if 3 or more websites provide same fact, it is given the highest rating and the procedure continues for other website's fact also.

VII. EXPERIMENTAL RESULTS

Fig.7. shows the Login page. The user-Id and password is authenticated, that is checked with stored user name and password to allow only the legitimate user to access the account.

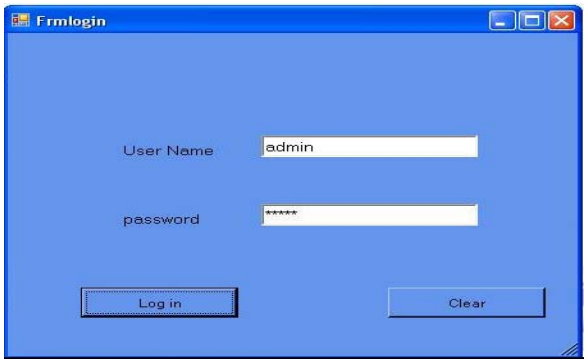

Fig.7. Login Page

Fig.8. shows how the query executes in the background, when we click the extract button on the main page the domain name and the value gets separated.

Fig.8. Search Page

Fig.9. shows the domain name, query and the results separated, when we click on the extract values.

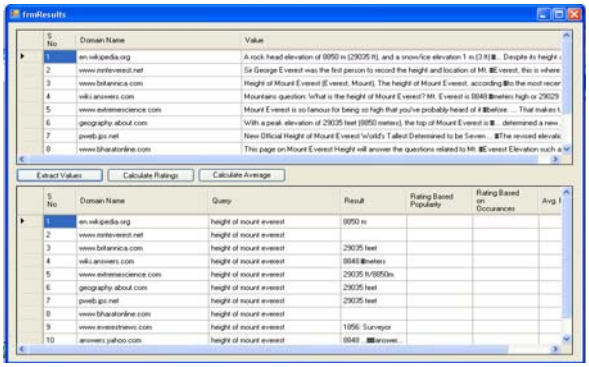

Fig.9. Extraction

Here in Fig.10. overall rating for the site is calculated by taking an average of two ratings based on the popularity of the website and based on number of occurrences of the value.

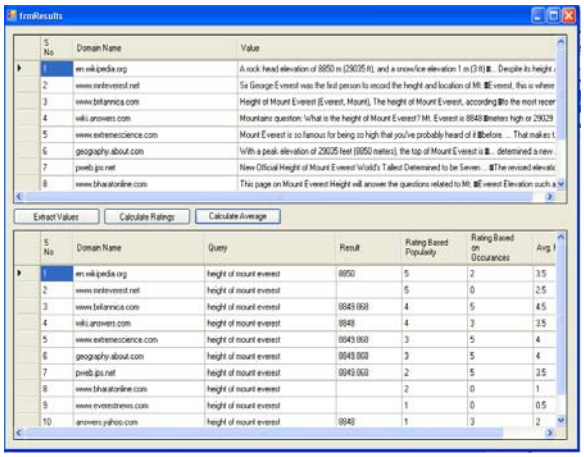

Fig. 10. Rating

VIII. CONCLUSIONS AND FUTURE SCOPE

In this paper, we have taken real dataset as of Height of Mount Everest , and experimented the Truth Finder [2] algorithm based on two facts i.e. popularity and occurrences. An attempt to write this paper is to throw a light of how we can work on this further in a better way. As we know we can't predict particular website is true enough, by its popularity alone, we thought of giving overall rating based on the average of popularity or the hits & number of occurrences of same data in many different websites. We have worked on numerical queries which gave successful result as shown in the Section 6. Further the work can be continued & make this as a better search engine than any popular ones by considering few more relations or the facts.

REFERENCES

[1] 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.

[2] X. Yin, J. Han, and P. 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.

[3] R.Y. Wang and D.M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," J. Management Information Systems, vol. 12, no. 4, pp. 5-34, 1997.

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

[5] X. Zhu and S. Gauch, "Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web," Proc. ACM SIGIR '00, July 2000.

[6] 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.

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

[8] 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.

[9] 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.

[10] J.S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," technical report, Microsoft Research, 1998.

[11] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity," Proc. ACM SIGKDD '02, July 2002.

[12] Yin, J. Han, and P.S. Yu, "LinkClus: Efficient Clustering via Heterogeneous Semantic Links," Proc. 32nd Int'l Conf. Very Large Data Bases (VLDB '06), Sept. 2006.

[13] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of Trust and Distrust," Proc. 13th Int'l Conf. World Wide Web (WWW), 2004.

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

[15] Logistical Equation from Wolfram MathWorld, http:// mathworld.wolfram.com/LogisticEquation.html, 2008.

[16] Sigmoid Function from Wolfram MathWorld, http://mathworld.wolfram.com/SigmoidFunction.html, 2008.

◈◈◈