PCF-Engine: A Fact Based Search Engine

Srikantaiah K.C.[1], Srikanth P.L.[1], Tejaswi V.[1], Shaila K.[1], Venugopal K.R.[1], Iyengar S.S.[2], and Patnaik L.M.[3]

[1] Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore 560 001
[2] Head of Wireless Sensor Networks and Robotics Research Laboratory, Department of Computer Science, Lousiana State University, USA
[3] Vice Chancellor, Defence Institute of Advanced Technology, Pune, India
srikantaiahkc@yahoo.com

Abstract. The World Wide Web (WWW) is the repository of large number of web pages which can be accessed *via* Internet by multiple users at the same time and therefore it is *Ubiquitous* in nature. The search engine is a key application used to search the web pages from this huge repository, which uses the link analysis for ranking the web pages without considering the facts provided by them. A new application called *Probability of Correctness of Facts(PCF)-Engine* is proposed to find the accuracy of the facts provided by the web pages. It uses the Probability based similarity (SIM) function which performs the string matching between the true facts and the facts of web pages to find their probability of correctness. The existing semantic search engines, may give the relevant result to the user query but may not be 100% accurate. Our algorithm probes for the accuracy among the facts to rank the web pages. Simulation results show that our approach is efficient when compared with existing Voting [1] and Truthfinder [1] algorithms with respect to the trustworthiness of the websites.

Keywords: Data mining, Page Rank, Search Engine, Trustworthiness, Web Content Mining.

1 Introduction

World Wide Web (WWW) is a collection of interconnected web pages accessed *via* internet offers information and data from all over the world. When searching for a topic in the WWW, it returns many links or web sites related on the browser to a given topic. The important issue is to determine the website that gives the accurate information. There are many related web sites that give unauthoritative information. While the information in other repositories like books, library and journals is evaluated by scholars, publishers, and subject experts. We have no mechanism to evaluate the information on WWW. Hence, it is necessary to consider some criteria [2] to evaluate the information hosted on WWW. Web search engines are programs used to search information on the WWW and FTP servers and to check the accuracy of the data automatically. It operates in the following order: *Web Crawling*, *Indexing*, *Searching* and *Ranking*. Ranking is

K.R. Venugopal and L.M. Patnaik (Eds.): ICIP 2011, CCIS 157, pp. 113–122, 2011.

a process of arranging the retrieved WebPages of the Search Engine Result Page (SERP) based on the relevance of the query entered. Relevance of a webpage is calculated based on the contents of the web page, including title, Meta data, popularity, authority, facts, location and frequency of a term in a web page.

Motivation: The existing page rank algorithms such as $Authoritative - Hub$ analysis and *PageRank* uses the statistical analysis i.e., the rank for the page is calculated based on the number of links referring to the page and on the importance of the referring pages. The facts provided by the web pages are not considered while assigning the ranks to the pages.

Contribution: A new approach called PCF-Engine is proposed in this paper to find the probability of correctness of the conflicting facts by applying Probability based similarity (SIM) function between conflicting facts and true facts available in the knowlegde base and it is referred as trustworthiness of the website.

Organization: The rest of this paper is organized as follows: Section 2 describes Related Work. Section 3 descibes Background of our work, Section 4 defines problem, describes Mathematical Model and algorithm of PCF-Engine, Section 5 comprises of experimental evaluation along with experimental results and analysis. The concluding remarks are summarized in Section 6.

2 Related Work

Several algorithms have been proposed to rank the webpages retrieved by search engine and they are categorized into Authority Based and Fact Based. PageRank [3] and HITS [4] are authority based rank algorithms. PageRank Algorithm has been developed by Larry Page, ranks the webpages based on the indegree of a node in the web graph and it is a query independent algorithm. Further the original pageRank algorithm has been improved by considering weights to links [5], Cluster Prediction, Subgraphs [6], Timestamp, Extrapolation method, Index [7], and Machine Learning [8]. HITS and its variants [9, 10], rank the webpage based on both indegree and outdegree of a node in the web graph, and it is a query dependent algorithm. But most authority pages do not contain accurate information [11]. Truthfinder [1] is a fact based search engine; it ranks websites by computing trustworthiness score of each website using the confidence of facts provided by websites.

3 Background

Xiaoxin Y et al., [1] have proposed an algorithm Truthfinder to find true facts from conflicting information from different infomation providers on the web. This approach is applied on certain domain such as, book authors and Movie run time. For the books domain, the Truthfinder uses author name as the facts which assigns the weights for first, middle and last name of the authors to find the confidence of the facts and this is repeated for every fact to find trustworthiness of the website. It assigns the weight ratio of 2:1:3 for first, midddle and last name respectively.

Example: True fact says, the author of some book is Graeme C. Simsion, where weight 2 is assigned for Graeme, 1 for C and 3 for Simsion and if the fact obtained from book seller website is Grame Simsio, where it does not contain middle name C and the charater 'n' is missing in the last name, it is only partially correct. Truthfinder assigns the half of the weight allocated for last name, i.e., 3/2 and full weght of 2 to first name and zero to middle name, therefore confidence of the fact is (2+1.5)/6 which is 58.33%.

PCF-Engine performs string matching between author names provided by the book sellers with author names of the corresponding book in the knowledge base and hence, it searches for exactness of the fact. Therefore confidence of the above example is 93.75%. Hence it is more accurate than Truthfinder.

4 Proposed Model

4.1 Problem Definition

Given a set of objects, a set of websites providing conflicting facts for an object and a set of true facts for a specific domain, the main goal is to find the Probability of Correctness of the conflicting facts with respect to the true facts to rank the webpages providing the facts.

Assumption: The facts available in Knowledge base are 100% accurate and it is obtained from the trustworthy resource.

4.2 Mathematical Model

Basic Definitions:

(i)*Probability of Correctness of Fact (PCF)*- is defined as the probability by which the fact is similar to that of the true fact or in other words by the factor that the fact has minimal deviation from the true fact.

(ii)*Implication between the facts* - is defined as the extent by the facts has influence other facts of the same object, i.e., the deviation between the PCFs of the facts from the threshold (maximum allowable deviation between the PCFs of the facts).

Trustworthiness of website is directly proportional to confidence of all the facts provided by that website and implication between the facts [1]. The basic notations used in the model are shown in Table 1.

Table 1. Basic Notations

ε	: is the threshold, i.e., allowable deviation of PCFs between any two facts.
$p(f)$	: is the probability of correctness of the fact about an object in some
$Ob\{\}$	: Set of objects in certain domain.
$TF\{\}$	: Set of true facts indexed by objects.
$F^{'}\{\}$	: Set of facts provided by different websites indexed by objects.
$Web\{\}$	: Set of websites URLS indexed by objects.

Probability of Correctness of Facts (PCF): If $\exists\ f \in TF\{\}$, such that $f \leftarrow TF\{o\}$, provided by $Web\{o\}$, where, $\forall o \in Ob\{\}$ then,

$$p(F_i^{'}\{o\}) \leftarrow SIM(f, F_i^{'}\{o\}). \tag{1}$$

where, $1 \leq \mathrm{i} \leq \mid F^{'}\{o\} \mid$, *SIM(f, $F^{'}$)*$\leftarrow$ is defined as the factor by which $F^{'}$ is true with respect to f and it is based on the domain or context where it is used. If $F^{'}$ is completely true with respect to f, then the probability of $F^{'}$ is correct when f is considered as true fact is 1 *i.e.*, $F^{'}$ is 100% correct about an object, where $F_i^{'}\{\mathrm{o}\}$ is a i^{th} fact for an object o provided by some website as shown in Eq. (1).

It implies that facts obtained from the website about an object is exactly similar to that of the true fact of an object and therefore the *SIM(f, $F^{'}$)* is also used to find the initial trustworthiness of the website by applying this function between all the facts provided by the website and the corresponding objects true facts available in the knowledgebase.

Implication Between Facts: Let Δ represents the difference between the probability of two facts f_1 and f_2, i.e., $\Delta = p(f_1)$ - $p(f_2)$ and based on the value of Δ, there are three cases.

Case 1: If $(0 < \Delta < \varepsilon)$ or $(\Delta > 0$ and $\Delta > \varepsilon$ then, f_1 has low impact on f_2 by $|\varepsilon - \Delta|$.
Example: if $p(f_1)$=0.7 and $p(f_2)$=0.2 then Δ=0.5 which is greater than ε, i.e., 0.4, this implies f_1 is 70% correct and f_2 is 20% correct, the difference is 50% which is greater than 40%(threshold) which is preferably allowed deviation between any two facts, therefore f_1 is having low impact on f_2 by 10% (50-40)%.

Case 2: If $(\Delta > 0)$ and $(\Delta = \varepsilon)$; then, f_1 has impact of ε on f_2.
The difference between the probabilities of correctness of the facts is equivalent to the value of threshold and hence f_1 has the impact ε on f_2.

Case 3: If $(\Delta < 0)$ then, f_1 has high impact on f_2.

Example: if $p(f_1)$=0.2 and $p(f_2)$=0.7, then $\Delta = -0.5$ which is negative and less than ε, i.e., 0.4, this implies f_1 is 20% correct and f_2 is 70% correct, the difference is -50% which is less than 40%(preferably allowed deviation between any two facts), therefore f_1 is having high impact on f_2 by 90% (40-(-50))%. In otherwords, by adding 50% to f_1 gives f_2 correctness, therefore f_2 is having low impact on f_1. Hence impact or influence between any two facts f_1 and f_2 on the same object can be defined as,

$$Inf(f_1, f_2) = \begin{cases} |\varepsilon - \Delta| * s(f_2), & \text{for case 1 and case 3} \\ \varepsilon & \text{for case 2.} \end{cases} \tag{2}$$

Therefore adjusted confidence of a fact is defined as,

$$s^{'}(f_1) = s(f_1) + \sum_{o(f_1)=o(f_2)} Inf(f_1, f_2). \tag{3}$$

where, $s(f)$ is a confidence of a fact f defined in [1]

$$s(f) = 1 - \pi(1 - t(w)). \tag{4}$$

$$s'(f) = min \begin{Bmatrix} s'(f) * 10^{-\alpha} : s'(f) * 10^{-\alpha} > 1, \\ 1 \leq \alpha \leq \infty \end{Bmatrix} * 10^{-1} \tag{5}$$

and adjusted confidence score is defined in [1]

$$\sigma^*(f) = -ln(s'(f)). \tag{6}$$

In Eq. (5) dumping factor *i.e.*, $10^{-\alpha}$ is multiplied to the adjusted confidence $s'(f)$ to get the probability value less than or equal to 1.

SIM(TF,F') for Books Domain: Let Ob =$\{ob_1, ob_2, ob_3, \ldots, ob_n\}$, TF =$\{ TF_{11}, TF_{22}, TF_{33}, \ldots, TF_{nn}\}$ and F=$\{F'_{11}, F'_{22}, F'_{33}, \ldots, F'_{nn}\}$. where,

$$F'_{ij} = \{y_{ik} : 1 \leq k \leq n_b\}. \tag{7}$$

where, n_b is the number of authors in the i^{th} fact about the j^{th} object and i=j. F'_{ij} is again the set of authors for j^{th} object (book). For example : F'_{22} = $\{y_{21}$, y_{22}, y_{23}, $\ldots$, $y_{2n_b}\}$ is the set of authors of the second book(second object). The true fact can also contain only one author or a set of authors for a book as defined according to Eq. 8,

$$TF_{ij} = \{x_{ik} : 1 \leq k \leq n_a\}. \tag{8}$$

where, n_a is the number of authors in the i^{th} true fact about the j^{th} object and i=j and $x_{ik} = TF_i - X$, $X = x_l : 1 \leq l \leq n_a$ and $l! = k$. Therefore, the similarity function between i^{th} true fact and corresponding i^{th} fact provided for any object $o \in Ob$ is defined according to the Eq. (9),

$$SIM(TF_i, F'_{io}) = \sum_{j=1}^{|F'_{io}|} \sum_{k=1}^{|TF_{io}|} LEN(F'_{ioj})/LEN(TF_{iok}), if f F'_{ioj} \subseteq TF_{iok}. \tag{9}$$

Repeat the process for all object $o \in Ob$, therefore

$$SIM(TF, F') = \sum_{i,o=1}^{|F'|} \sum_{j=1}^{|F'_{io}|} \sum_{k=1}^{|TF_{io}|} LEN(F'_{ioj})/LEN(TF_{iok}). \tag{10}$$

$LEN(f)$: Gives the number of characters found in the author name F' in Eq. 9 and Eq. 10. Here, the probability of correctness is calculated depending on the number of characters matched in the first, last and middle of the author name of obtained facts from different website to the first, last and middle name of the author names taken in that order of the true fact about the object(book) available in the knowledge base.

Example: if TF={{Cay S Horstmenn,Gary Cornell}, TF_{22}, . . . , TF_{nn}} is a true fact about the book Core java Volume 1 with ISBN 8131701621 and TF_{11}={Cay S Horstmenn,Gary Cornell}, F'={{Cay S Horstmenn,Gary",
"Horstmenn","Corne}, F'_{22}, . . . , F'_{nn}} where, F'_{11}= {Cay S Horstmenn, Gary, Horstmenn, Corne}.

Consider first author F'_{111} *i.e.*, Cay S Horstmenn provided by w_1 which is same as the author name provided in true fact Cay S Horstmenn therefore $p(F'_{111})$=1. Similarly, consider the second author F'_{112} *i.e.*, Gary which is a subpart of true fact Gary Cornell therefore, $p(F'_{112})$= LEN(Gary)/LEN(Gary Cornell) which is 0.33 similarly $p(F'_{113})$= 0.6 and $p(F'_{114})$=0.416.

Therefore, initial trustworthiness of w_1, $t(w_1)$= (1+0.33)/2 *i.e.*, 0.625 on ISBN 8131701621; if w_1 provides F'_{111} and F'_{112}. Similarly, initial trustworthiness of w_2 on ISBN 8131701621 is $t(w_2)$=1.016/2, *i.e.*, 0.508 where, 2 indicates number of facts provided by websites on the object(ISBN 8131701621) if it provides F'_{113} and F'_{114}. This process is repeated for every object(every book) provided by the corresponding websites to get their respective initial trustworthiness and they are ranked accordingly.

Initially, it is assumed that none of the websites are trustworthy, therefore initial trustworthiness of all websites are assigned to zero. Therefore the trustworthiness of website $t(w)$ is redefined as,

$$t(w) = \begin{cases} SIM(TF, F'), \, if \; t(w) = 0 \\ \sum_{f \in F(w)} s(f)/|F(w)| : otherwise[1]. \end{cases} \tag{11}$$

where, F' is a set of facts provided by website w and $F' \subseteq F$. If trustworthiness is zero, then the website is added with the new data to the database whose trustworthiness is calculated using $SIM(TF, F')$, otherwise, trustworthiness is calculated by taking the average of confidence of all the facts provided by the website w in Eq. 11.

The proposed *Probability of Correctness of Fact*(PCF) engine ranks the page depending on the accuracy of the facts provided by the websites. The facts which are assumed as true about any object are stored in knowledge base. For example the true facts about the different books are taken from the respective coversheets of the books. Following properties are some of the facts taken for the book, ISBN: Uniquely identifies the fact, Author Names: Authors for the corresponding book, Publisher: publisher for the book, Price: cost of the book. Once knowledge base is constructed, the dataset containing conflicting facts for the various objects are populated using the website *www.abebooks.com*.

The $\in$ is set to 0.4 which indicates that 40% deviation in PCF between the facts are allowed. The algorithm behavior can be rendered by changing the value of threshold. The algorithm includes three important stages:(i) calculation of trustworthiness of all the websites, (ii) calculation of confidence of all the facts available in database and (iii) finding the influence between the facts. Since the algorithm operates on real dataset it is scheduled to run on every day to update the contents of the database.

Table 2. Algorithm: PCF-Engine

Input

$TF\{\}$: Set of true facts indexed by objects.
$F'\{\}$: Set of facts providedby different websites indexed by objects
$Ob\{\}$: Set of objects.
$Web\{\}$: Set of Websites URL'S providing the facts.
ε : Equal to 0.4, allowable deviation between any two facts on the same domain.

Output

Trustworthiness of the websites and confidence of the facts.

Process

begin
for each $w \in Web$; *Compute Trustworthiness for every Website*
do
if $t(w)$=0
then
$t(w)= SIM(TF, F')$; where F' is the set of facts provided
; by website w
else
$t(w)= \sum_{f \in F(w)} s(f)/|F(w)|$
end if
end for
for each $f \in F$; *Compute the Confidence of the facts*
do
$s(f)$=1-π(1-$t(w)$) ; for every website w providing a fact f
where, $w \in Web\{\}$
$\sigma(f)$=-ln(1-$s(f)$) ; confidence score of a fact f
end for
for each $f \in F$; *Compute the Implication between the facts*
do
for each $f' \; \varepsilon \; F$ and $f' \neq f$
do
Δ=$p(f)$ - $p(f^{'})$
if Δ=ε
then
$inf(f, f^{'})$+= ε
else
$inf(f, f^{'})$+= $\mid \varepsilon - \Delta \mid$ * s($f^{'}$)
end if
end for
$s(f) \leftarrow$ get the confidence of f from database
$s^{'}(f)$=$s(f)$+$\sum_{o(f')=o(f)} inf(f, f^{'})$
end for
end

The Initial trustworthiness is calculated depending on the PCF of all the facts provided by the website where the PCF for every fact is determined by using Probability based similarity (SIM) function. The facts provided by the different websites may be similar to the true fact and hence the PCF for those facts is 1. Which indicates the fact is 100% true and this is calculated on a fly in a single iteration. If PCF of all the facts provided by website is 1 and the deviations in the implications between the facts are low, then the trustworthiness closely approaches to 1 and hence the PCF engine always probes for exactness of the facts about an object.

The algorithm calculates trustworthiness for every websites by finding the PCF for the facts and it also computes the confidence of the facts by taking the trustworthiness of the corresponding websites providing the facts and hence *trustworthiness* and $confidence$ are totally depending on each other. The algorithm stops after computing the trustworthiness of all the websites and confidence for all the facts found in the database. It recomputes the trustworthiness and confidence values when it is scheduled for next execution by considering the new facts arrived after the previous Execution. The algorithm is presented in Table 2.

5 Experimental Results

The data set consists of facts for the books domain, where domain in this context corresponds to values of certain attributes of the book such as ISBN, Author Names, Publisher, Price, URL of Book seller website and quantity (availability). The data set consists of the above specified information for 26 websites with 47 facts. The initial trustworthiness of websites, confidence and confidence scores of all the facts are initialized to zero. The Author Name of the book is considered as the important fact for Probability based similarity (SIM) function to perform the relevance analysis. The result of the Probability based similarity (SIM) function for all the facts provided by a website is used to calculate the trustworthiness of the website and this is performed on all the websites to rank them accordingly.

The PCF-Engine is developed using ASP.NET with C# as the underlying language. The Visual Studio 2005 (IDE), Windows XP(OS) and the MySQL 4.0 for database forms the complete development environment. The initial implementation is done with the ε set to 0.4. As shown in Fig. 1, the search is made for the ISBN 8183330088 of the book titled with Web Enabled Commercial Applications Development using HTML DHTML Javascript Perl CGI , the webIds providing the facts about this book are 5, 7, 3, 10, 8, 9 . . . *etc.*, of which the most trustworthy websites with value 1 are 5 and 7 and hence they are occupying the first two positions in the searched result.

The graph is plotted for the websites providing the facts for the book Web Enabled Commercial Application devlopment using HTML, DHTML, Java Scrift, Perl, CGI" by Ivan Bayross. As it is observed from the Fig. 2, the trustworthiness of websites fall in the range 18%-25% for Voting, 26%-33% for Truthfinder

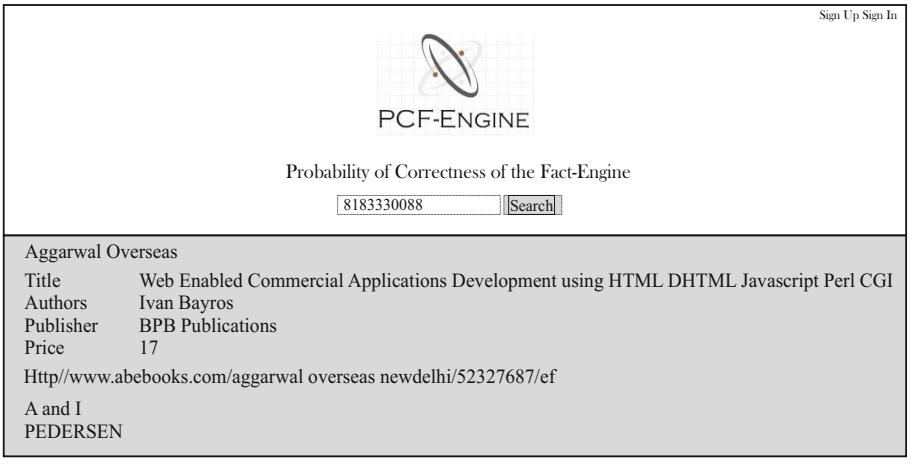

Fig. 1. Snapshot of the PCF-Engine

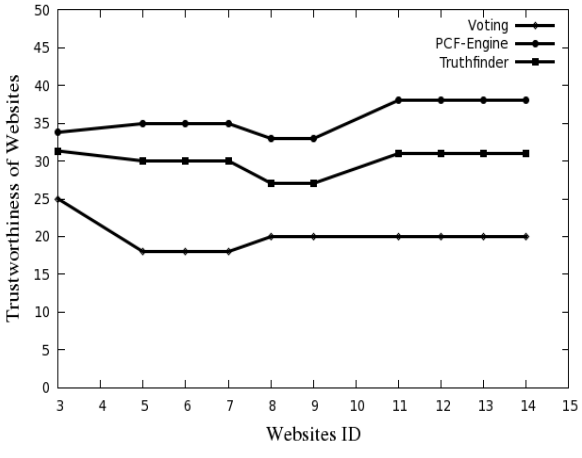

Fig. 2. Comparison between the Trustworthiness values of Truthfinder, Voting and PCF-Engine

and 33%-38% for PCF-Engine and the deviation between PCF Engine and Truthfinder with trustworthiness calculations is 0.058 and hence PCF-Engine is 5.8% more accurate than the Truthfinder. The probability values are normalized to two digits numbers in Y-axis. Since the Voting uses the facts count without considering the truthness of the facts provided by the websites, its accuracy is low compared to PCF-Engine and Truthfinder algorithms.

6 Conclusions

In this paper a new approach called PCF-Engine which uses Probability based similarity (SIM) function is proposed for resolving the conflicts between the

facts provided by the different information providers in web. The Probability based similarity (SIM) function finds the implication between the facts. If the websites provides the fact which is exactly similar to that of true fact in the knowledgebase the PCF-Engine computes its trustworthiness value as 1 on a fly in a single iteration. The work can be extended by dynamically fetching the true facts to the knowledge base and removing the domain specific dependency of true facts.

References

1. Xiaoxin, Y., Jiawei, H., Philip, S.Y.: Truth Discovery with Multiple Conflicting Information Providers on the Web. Journal of IEEE Transactions on TKDE 20(6), 796–808 (2008)
2. Johns Hopkins University, `http://www.library.jhu.edu/researchhelp/general/evaluating/`
3. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Journal of Computer Networks 30(7), 107–117 (1998)
4. Kleinberg, J.M.: Authoratative Sources in a Hyperlinked Environment. Journal of ACM 46(5), 604–632 (1999)
5. Xing, W., AliGhorbani: Weighted PageRank Algorithm. In: 2nd Annual Conference on Communication Networks and Services Research, pp. 305–314. IEEE Press, Los Alamitos (2004)
6. Heasoo, H., Andrey, B., Berthold, R., Erik, N.: BinRank: Scaling Dynamic Authority-Based Search using Materialized Subgraph. Journal of IEEE Transactions on TKDE 22(8), 1176–1190 (2010)
7. Amit, P., Chakrabarti, S., Manish, G.: Index Design for Dynamic Personalized PageRank. In: IEEE 24th International Conference on Data Engineering, pp. 1489–1491. IEEE Press, Los Alamitos (2008)
8. Sweah, L.Y., Markus, H., Ah Chung, T.: Ranking Web Pages using Machine learning Approaches. In: IEEE International Conference on Web Inteligence and Intelligent Agent Technology, pp. 677–680. IEEE Press, Los Alamitos (2008)
9. Matthew, H., Julie, S., Chaoyang, Z.: A Scalable Parallel HITS Algorithm for Page Ranking. In: First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2006), pp. 437–442. IEEE Press, Los Alamitos (2006)
10. Allan, B., Gareth, O.R., Jeffrey, S.R., Panayiotis, T.: Link Analysis Ranking Algorithms, Theory and Experiments. Journal of ACM Transactions on Internet Technology 5(1), 231–297 (2005)
11. Brian, A., Loren, T., Hill, W.: Does Authority Mean Quality? Predicting Expert Ratings of Web Documents. In: ACM SIGIR 2000, pp. 296–303 (July 2000)