A Decentralized Recommender System for Effective Web Credibility Assessment

Thanasis G. Papaioannou, Jean-Eudes Ranvier, Alexandra Olteanu and Karl Aberer
School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
Email: firstname.lastname@epfl.ch

ABSTRACT

An overwhelming and growing amount of data is available online. The problem of untrustworthy online information is augmented by its high economic potential and its dynamic nature, e.g. transient domain names, dynamic content, etc. In this paper, we address the problem of assessing the credibility of web pages by a decentralized social recommender system. Specifically, we concurrently employ i) item-based collaborative filtering (CF) based on specific web page features, ii) user-based CF based on friend ratings and iii) the ranking of the page in search results. These factors are appropriately combined into a single assessment based on adaptive weights that depend on their effectiveness for different topics and different fractions of malicious ratings. Simulation experiments with real traces of web page credibility evaluations suggest that our hybrid approach outperforms both its constituent components and classical content-based classification approaches.

Categories and Subject Descriptors

H.3.3 [**Information Storage & Retrieval**]: Information Search & Retrieval

General Terms

Theory

Keywords

collaborative filtering, social networks, similarity metrics

1. INTRODUCTION

People employ online information more and more in their everyday lives for shopping, business, social life, relationships, etc. However, the overwhelming amount of online content as well as the great economic potential involved in influencing people's knowledge render the problem of web credibility assessment a non-trivial one. Credibility is a rather subjective term depending on multiple factors, such as information authenticity, authority, objectivity, freshness and topic coverage [12]. Authors of non-credible online content employ dynamic web page content, easy domain name changes and web link graph modifications to hide their traces and appear at a high position in the ranking of the search results, i.e. web spamming.

Several approaches, e.g. [13, 18], have been proposed that combine different page content features to assess credibility with limited effectiveness. Since many users perceive high ranking position of a page in the search results as high credibility, reordering the search results based on page credibility would greatly facilitate people's access to credible information. To this end, several approaches [2, 6] have been proposed against web spamming. Also, some system initiatives have arisen, such as Google+ "+1" button[1] or Facebook[2] "like" button, in order to evaluate credibility of web pages based on user ratings. However, as these approaches collect ratings that are subject to data mining, valid privacy concerns are raised by the users that may deter participation to these systems.

In this paper, we propose a decentralized privacy-friendly social recommender system for web page credibility assessment. Our recommendation scheme for credibility assessment of a web page includes a user-based collaborative filtering component for taking into account the user ratings for the page, an item-based collaborative filtering component for considering the content-based features of the page and a third component that is based on the ranking of the page in the search results. To the best of our knowledge, none of the existing approaches combines these aspects in a generic way. Along the way of the content-based component description, we define new versions of existing similarity measures that are convenient for the different content features employed. Moreover, our approach is generic-enough to include additional web page features. The different components of the recommendation scheme are combined in a weighted average, while the weights are properly updated to reflect the effectiveness of each individual component for credibility assessment. In this decentralized social network, friends should be opportunistically added, so that ratings exist for a large number of web pages. Since malicious users may infiltrate into the social network, we propose a friend banishment approach that progressively expels them. We verify the effectiveness of our recommendation scheme by simulation experiments with two real corpus of web page

[1] http://www.google.com/+1/button/
[2] http://www.facebook.com

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

credibility assessments, while assuming the presence of varying fractions of malicious users in the social network. Finally, we implemented a working prototype of the proposed system and our design choices were validated in practice.

The remainder of this paper is organized as follows: In Section 2, we overview our system for credibility assessment. In Section 3, we present our multi-component approach for web page credibility assessment. In Section 4, we explain how to determine the weights of our multi-component recommendation scheme and then in Section 5, we describe our banishment scheme for malicious friend exclusion from the recommendation algorithm. In Section 6, we present our prototype system implementation, while in Section 7, we experimentally evaluate our web page credibility assessment based on two page corpus. In Section 8, we overview the related work and finally, in Section 9, we conclude our work.

2. SYSTEM DESIGN

In our system, users are connected to each other on a Peer-to-Peer (P2P) overlay. A P2P node resides in a plug-in at the user browser. The user inserts queries in the web search engine and the plug-in fetches the search results, assesses the credibility of the 100 top-ranked pages according to our credibility estimation scheme (described in the next section) and modifies the ranking of the search results accordingly. The user is able to assign a binary vote to any page that she visits or sees in the search results, i.e. $\{-1, 1\}$. Each user rating for a page is stored at a local database. Periodically, batches of user ratings are broadcasted to all user friends. Upon receiving a rating for a web page from a friend, the rating is stored at the local database and it is employed for credibility evaluation in subsequent web searches. The overall architecture of our approach is shown in Figure 1.

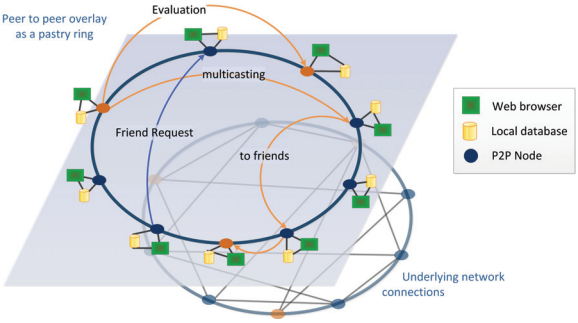


Figure 1: Overview of our system.

3. RECOMMENDATION SCHEME

In our system, users are connected to their friends in the "social graph", while pages can be implicitly linked with other pages based on their content-based similarity in a "content-similarity graph" or explicitly linked with other pages based on their incoming/outgoing links in the web, as depicted in Figure 2. Also, note that a user may rate or edit/own a web page. Based on the different semantics of the node associations in each graph, we derive a separate credibility assessment by each graph for a web page under evaluation and later we combine them into a single credibility assessment. In the social component, the credibility assessment is based on the recommendations of the friends in the social graph. Moreover, we evaluate the credibility of the user friends by evaluating the consistency of their recommendations with those of the user herself and banish them according to the approach of Section 5. In the content component, we evaluate credibility of a page based on its content-based similarity with other pages. Lastly, in the search-ranking component, we evaluate the credibility of the page based on its associations in the web graph. We describe these components for credibility evaluation and their combination in the following subsections.

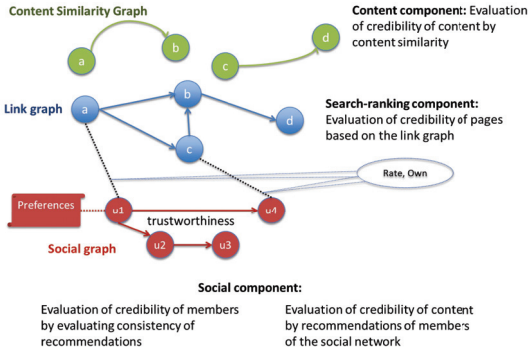


Figure 2: Our recommendation scheme.

3.1 Social Component

The social component ρ_u estimates the credibility of a web page based on the ratings of the user friends for this page. It is calculated according to the user-based collaborative filtering algorithm introduced in [7], which is given by the formula below:

$$\rho_{u,j} = \overline{r_u} + \frac{\sum_{v \in U_{u,i}} (\overline{w}_{u,v}(r_{v,i} - \overline{r_v}))}{\sum_{v \in U_{u,i}} (\overline{w}_{u,v})}, \quad (1)$$

where $\overline{r_u}$ is the average rating over all ratings of user u, $\overline{w}_{u,v}$ is the similarity between user u and user v based on their ratings for mutually rated pages (given by the mean-adjusted Pearson correlation metric), $r_{v,i}$ is the rating of user v for the page i, $\overline{r_v}$ is the average rating over all ratings of user v, and $U_{u,i}$ is the "neighborhood" of user u defined as her k nearest friends in terms of $\overline{w}_{u,v}$. The mean-adjusted Pearson correlation for users u, v that mutually rated pages in $I = I_u \cap I_v$ is given by:

$$\overline{w}_{u,v} = \frac{\sum_{i \in I} (r_{u,i} - \overline{r_u}) * (r_{v,i} - \overline{r_v})}{\sqrt{\sum_{i \in I} (r_{u,i} - \overline{r_u})^2 * \sum_{i \in I} (r_{u,i} - \overline{r_u})^2}} \quad (2)$$

Note that Pearson correlation is not defined for $|I| = 1$ and when the standard deviation for the rating vector of one user is 0, i.e. $\sigma_{R_u} = 0$ or $\sigma_{R_v} = 0$, where R_u, R_v are the rating vectors of users u, v. In such a case, a very low positive standard deviation is assumed in the calculations. Only the top-k most similar friends are considered in Eq. (1), which is referred to as the *neighborhood of the user* for this collaborative filtering algorithm.

3.2 Content Component

This component assesses the credibility of the web page based on content-based features, such as semantic ones (e.g. category, entities, keywords etc.), NLP ones (sentiments,

subjectivity, etc.), syntactic ones (part-of-speech tag multiplicities, punctuation marks, spelling errors, etc.), advertisements, page layout, etc. The credibility of a page is calculated by the item-based collaborative filtering [16] approach is used, which is defined as follows:

$$\rho_{c,i} = \frac{\sum_{j \in I_u} s_{i,j} \times r_{u,j}}{\sum_{j \in I_u} |s_{i,j}|} , \quad (3)$$

where I_u is the set of pages rated by u, $s_{i,j}$ is the similarity between pages i and j, and $r_{u,j}$ is the rating of user u for page j. Notice that only the pages rated by user u are considered in the item-based algorithm, as opposed to considering pages rated by friends as well. This is done in order for the user to have "personalized" means to assess the credibility of the page as she can be assumed to assign maximum trust to her own ratings. Moreover, we do not use the ratings of different users for calculating the similarity between pages, proposed by [16]. The calculation of the page similarity depends on the specific content features employed, as described in the following paragraphs. Only the top-k most similar pages are considered in Eq. (3), which is referred to as the *neighborhood of the page* for this collaborative filtering algorithm.

Syntactic and lexical features.

The syntactic features that we consider include part-of-speech and punctuation marks, while the lexical ones include text complexity and spelling errors. Part-of-speech feature extraction refers to the categorization of the page content to different word categories, e.g. nouns, verbs, adjectives, adverbs, based on their definition and their context. Text complexity defined in [8] is given by the entropy of different words in the text of the web page. Each of these features is single-valued, and thus, we can compare two different pages based on their corresponding feature vectors A, B, by using the cosine similarity, defined as follows:

$$\cos(\theta) = \frac{A \cdot B}{||A|| \times ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4)$$

Semantic features.

We employed as semantic features the page category, the entities, the keywords and the informativeness of the keyword to the page content. For comparing two pages based on their entities and keywords, we propose a Jaccard-like similarity, as explained below. The Jaccard similarity metric $J_{A,B}$ between pages A and B is given by:

$$J_{A,B} = \frac{A \cap B}{A \cup B} , \quad (5)$$

where $A \cap B$ is the set of elements in A and B that are strictly equal to each other and $A \cup B$ is the union of all elements. However, the requirement for strict equality between entities or keywords between pages is very restrictive. Thus, we relax this requirement and use a normalized Levenshtein[3] distance for comparing strings between pages. Specifically, the distance $\hat{L}_a(B)$ of a keyword/entity a in page A from page B is given by:

$$\hat{L}_a(B) = \min_{b \in B} \frac{L(a,b)}{\max\{length(a), length(b)\}} , \quad (6)$$

where $L(a,b)$ is the Levenshtein distance between strings a and b, and $length(.)$ provides the length of its string parameter. This function expresses the closest normalized distance of a keyword/entity a in page A with any of the keywords/entities in page B and takes values in $[0,1]$. Otherwise, it would not be obvious to select the keywords/entities based on which two pages would have to be compared. The Jaccard-like similarity $\hat{J}_{A,B}$ of two pages A, B is defined by the formula below:

$$\hat{J}_{A,B} = \frac{\sum_{a \in A} (1 - \hat{L}_a(B))}{\sum_{a \in A} (1 + \hat{L}_a(B))} \quad (7)$$

$\hat{J}_{A,B}$ takes values in $[0,1]$ and it is a loose version of the intersection (i.e. common characteristics) over the union (i.e. all characteristics) ratio of Eq. (5). Indeed, the numerator in Eq. (7) is the sum of similarities among keywords/entities between two pages, while the denominator expresses the summed distance between the pages.

The page category (elsewhere referred to as category cohesion [8]) refers to the proximity of a page to a specific category based on the frequency of appearance of its terms to the pages of various categories, e.g. sports, politics, religion, entertainment, etc. The page category was only used to evaluate the effectiveness of restricting the comparison between pages the same category, assuming no similarity between pages of different categories.

The informativeness of a keyword to the page (calculated by the $tf * idf$ metric in [8]) is also considered in page similarity. In the $tf * idf$ metric, tf is the term frequency, while idf is the inverse document frequency that measures how common is the term across all documents. The lower the informativeness of a keyword to the page where it belongs, the less significant it should be for comparing the page to another one. Therefore, we define an informativeness weight $m_{a,b}$ for the distance between two keywords/entities a and b belonging to pages A and B respectively, as the minimum of their informativeness to their corresponding pages, i.e.:

$$m_{a,b} = \min\{tf * idf(a), tf * idf(b)\} , \quad (8)$$

We integrate the keyword informativeness in the normalized Levenshtein distance of Eq. 6 as follows:

$$\hat{J}'_{A,B} = \frac{\sum_{a \in A} (1 - \hat{L}_a(B)) m_{a,\hat{b}}}{\sum_{a \in A} (1 + \hat{L}_a(B)) m_{a,\hat{b}}} , \quad (9)$$

where $\hat{b} = \arg\min_{b \in B} \frac{L(a,b)}{\max\{length(a), length(b)\}}$.

NLP features.

Two pages can very similar in terms of keywords and entities, but they may have different sentiments (i.e. positive/negative opinions) associated with them. Therefore, we also consider the sentiment associated with each keyword/entity (as provided by tools, such as LingPipe[4] and AlchemyAPI[5]) in the page similarity. The sentiment associated to a keyword/entity takes values in $[-1, 1]$, where -1

[3]Levenshtein (or edit) distance between strings a, b is the number of single character edits (i.e. insertion, deletion, substitution) in order for string a to transform to b.

[4]http://alias-i.com/lingpipe/

[5]http://www.alchemyapi.com/

means totally negative, 1 means totally positive and 0 means neutral sentiment. For two keywords/entities a, b, we define the *sentiment similarity* $\varphi_{a,b}$ by the formula below:

$$\varphi_{a,b} = 1 - |sentiment(a) - sentiment(b)| \qquad (10)$$

Note that $\varphi_{a,b}$ again takes values in $[-1, 1]$. We integrate sentiment similarity into Eq. (9) and we define the combined Jaccard-like similarity metric $\hat{J}^*_{A,B}$ between pages based on their keywords/entities, including their informativeness (or relevance) to the page and their sentiment, by the formula below:

$$s_{A,B} \equiv \hat{J}^*_{A,B} = \frac{\sum_{a\in A}(1 - \hat{L}_a(B))m_{a,\hat{b}}\varphi_{a,\hat{b}}}{\sum_{a\in A}(1 + \hat{L}_a(B))m_{a,\hat{b}}}\,, \qquad (11)$$

where $\hat{b} = \arg\min_{b\in B} \frac{L(a,b)}{\max\{length(a),length(b)\}}$. In Eq. (11), we multiply the pages similarity by the sentiments similarity to make sure that pages with opposite sentiments would have opposite ratings. Note that in Eq. (11) $s_{a,b}$ appears only in the numerator. This is due to the definition of the similarity as a Jaccard-like one: the numerator expresses the summed similarity between pages, while the denominator expresses the union of compared terms. Therefore, the maximum sentiment (i.e. 1) is implicitly considered in the denominator. The complexity of the combined Jaccard-like similarity metric is $O(n^2)$, where n is the maximum number of keywords/entities per page. To this end, proactive calculation of this metric for potentially-targetted pages is foreseen for adequately fast online web credibility assessment.

3.3 Search-Ranking Component

The search ranking component estimates the credibility of a web page based on the page ranking in the search results. According to [13], many people tend to believe that a higher ranking in the search results means a higher credibility, which is not always true. Even worse, spammers exploiting this belief, have targeted the ranking mechanisms of popular search engines with relative success [2, 6]. In general, page ranking algorithm employ the link associations among pages and page popularity, which we deem as important features for assessing credibility. Our search-ranking component employs the ranking of the page in the search results as a credibility metric. This can be done by using directly the Google PageRank metric normalized by its maximum value as a credibility score, i.e. $\rho_{g,i} = pagerank(i)/\max_j\{pagerank(j)\}$. Another approach would be to estimate the page credibility rating based on its search ranking by a Zipf distribution, i.e. $\rho_{g,i} = \frac{1/k^s}{\sum_{j=1}^{N}(1/j^s)}$ for a page i in the k^{th} position of the search results, where $s > 0$ is an exponent characterizing the steepness of the distribution.

3.4 Aggregate estimation

Each individual component of the recommendation scheme described above provides a (potentially different) estimation of the credibility rating that should be assigned to the web page under credibility evaluation. These credibility estimates can be presented to the user individually; this choice would be transparent to the user and she would be able to conclude on the page credibility given all information. However, individual assessment by the different components may be conflicting to each other and they may confuse the user. A simple alternative would be to provide the user with the average of the estimated ratings of the individual components as the credibility assessment. A criticism on this choice would be that not all components are so influential for assessing the credibility of specific categories or so informative for specific users. To this end, we assign a different weight to each individual component for credibility assessment and calculate the estimated overall credibility rating, as follows:

$$\rho_i = \omega_u\rho_{u,i} + \omega_c\rho_{c,i} + \omega_g\rho_{g,i}\,, \qquad (12)$$

where $\rho_{u,i}$, $\rho_{c,i}$, $\rho_{g,i}$ are the estimated ratings for page i based on the social component, the content component and the search results ranking component, and ω_u, ω_c, ω_g are the respective weights of these components in the aggregate credibility estimation. These weights satisfy that $\omega_u + \omega_i + \omega_g = 1$ and they are properly set, so as to reflect the significance that has to be put to the individual components of the credibility estimation of the web page, as explained in the following section.

4. ADAPTIVE WEIGHT UPDATE

A natural question on the aggregate credibility estimation is how to properly determine the weights of the different components of the recommendation mechanism, so as to maximize its effectiveness for credibility assessment of different web pages. For different pages, different aspects may be more dominant than others for the estimation of the page credibility, e.g. for scientific pages content-based recommendation may be preferable than community-based one, while for music-related pages the social component could be much more useful than others. To this end, we maintain different credibility component weights per page category at each user.

Also, as explained in [5], different aspects of the web page affect differently its perceived credibility by different users and the prominence of these aspects in the web page determines the estimated page credibility. In other words, the a posteriori (i.e. after visiting it) credibility assessment of a web page by a user is rather subjective. We respect this inherent subjectivity by maintaining individual credibility component weights per page category at the user premises.

We propose that the weight of each credibility assessment component, given a history of s agreements of the posterior credibility assessment by the user with the prior recommendation of the respective component of the web credibility mechanism out of t recommendations in total, is updated according to the following Beta-distribution-based reputation formula [9]

$$\omega = \frac{s'}{t'}\,, \qquad (13)$$

where $s' = \beta s + \mathbb{1}(agreement)$ and $t' = \beta t + 1$, while $\mathbb{1}(.)$ is the indicator function, *agreement* denotes the consistency of the a priori credibility estimation of the new web page by the credibility component with the a posteriori credibility assessment by the user, and $0 \le \beta \le 1$ a discount factor for past credibility assessments. This weight updating approach guarantees convergence of each weight to the mean effectiveness of its respective credibility component per page category, as opposed to asymmetric TCP-like approaches (i.e. additive increase upon agreement, multiplicative decrease upon disagreement). The discount factor β is chosen,

so as to reflect the dynamic effectiveness that may arise for a credibility component, i.e. for relatively stable effectiveness β should be close to 1 and vice versa. Note that after updating the weights based on Eq. (13), the weights are normalized so that they sum up to 1.

Recall that the contacts of the user in the social network are not necessarily trusted friends, as the social network has to opportunistically expand to large user communities, in order to effectively collect ratings for arbitrary web pages. Thus, the community composition may dynamically change; more malicious users may infiltrate the neighborhood of a target user, while certain pages may become targets of malicious ratings. On the other hand, the effectiveness of the content and the search-ranking components are assumed to be rather stable. Thus, a low discount factor (e.g. $\beta \sim 0.6$) can be chosen for the social credibility component, while a high one (e.g. $\beta \sim 0.9$) can be used for the content and the search-ranking components. However, although the overall ineffectiveness of the social component can be captured by the described weight updating approach, malicious ratings are still taken into account in the credibility assessment. We deal with this issue in the next section.

5. FRIEND BANISHMENT

As the social network of the user may opportunistically increase, malicious users may become part of the user neighborhood in the user-based collaborative filtering algorithm for a web page. In such a case, the weight of the social component will reflect its ineffectiveness according to the weight updating approach described in the previous section. However, given the low weight of the social credibility component, the overall credibility estimation will not be able to benefit from the social credibility component, while the malicious ratings will still be included in the credibility evaluation of future web pages. To this end, we propose a *friend banishment* mechanism that is inspired by [14], as explained below. A user maintains for each friend j a local variable $nd_j \geq 0$ for the number of disagreements with her and a banishment period $bp_j \geq 0$. After visiting a page, the user u compares her credibility evaluation $\hat{r}$ to the rating r_j for this page of every user j that is *eligible* to be part of her collaborative-filtering neighborhood (i.e. most similar users) for this page and updates nd_j, bp_j as follows:

- If $\hat{r} = r_j$, then $bp_j = \max\{bp_j - 1, 0\}$ and $nd_j = nd_j - y$, where $y > 0$ is a certain *forgiveness* factor.
- If $\hat{r} \neq r_j$, then $nd = nd_j + x$ and $bp_j = bp_j + b^{nd_j}$, where $b > 1$ is the base of the banishment period and determines the *harshness* of the punishment for disagreement.

In the user collaborative-filtering neighborhood for a page are only considered her top-k most similar friends that are not currently banished (i.e. those with banishment period equal to 0). All users anticipated in the construction of the neighborhood for a page are considered eligible and their nd_j, bp_j values are updated, as already explained. This approach considerably limits the number of users whose banishment values have to be updated per web page visit. This banishment mechanism satisfies two desirable properties: i) it quickly excludes altogether malicious users, ii) it allows the consideration of the credibility assessment of people that only occasionally disagree with the user.

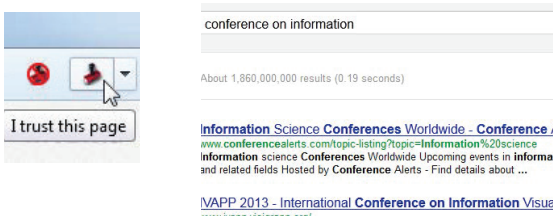
6. IMPLEMENTATION

Our recommendation mechanism has been implemented as a Firefox add-on. The interactions with the browser are made through Javascript, while computation-intensive operations and network communications are handled using Java, in order to benefit from its multithreading capabilities.

The peer to peer overlay management is made by a Java implementation of the peer-to-peer overlay Pastry [15]. Pastry is used to manage social relationships among friends in a decentralized manner (i.e. friend list is stored locally and managed through Pastry messages) and to exchange their evaluations.

All the data manipulated by the mechanism is stored locally in an embedded H2 SQL database[6] and then used to compute the credibility evaluations. The database is structured using the following tables:

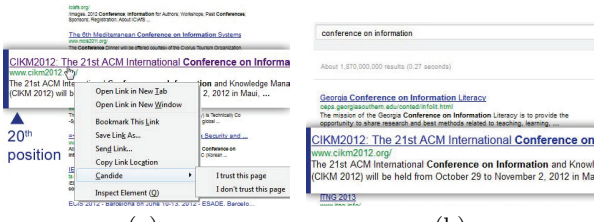
- Document: Represents a page that has been evaluated and the collection of its single-valued features.
- Entity: Represents an entity defined in a page, its relevance and its sentiment.
- Keyword: Represents a keyword defined in a page, its relevance and its sentiment.
- Evaluation: Contains the evaluations of pages made by friends of the user and the user himself.
- User: Contains the list of the friends of the user.



(a) (b)

Figure 3: (a) User interface in Firefox navigation bar. (b) Initial results returned for a query.

As we can see in Figure 3(a), the user interface is mainly composed of two buttons to rate the displayed page and a contextual menu (Figure 4(a)) to evaluate pages referenced by a link, i.e. without visiting the page.



(a) (b)

Figure 4: (a) Evaluation of the page by the user via the contextual menu. (b) Modified results.

A typical use case would be to query a search engine (in our case Google) for "conference on information". Assume that the initial ordering of the results places the CIKM 2012 website outside of the top 10 results (Figure 3(b)). If the user thinks that the website is credible and relevant to her

[6]http://www.h2database.com

query, she can manually rate the page as credible (Figure 4(a)). Upon resubmission of the same query by the user or her friends, her evaluation will be taken into account and the website will appear in the top results (Figure 4(b)). Notice that if her friends have already rated this page or if she has rated pages with similar content, the ordering of the results would also be affected accordingly.

7. EVALUATION

7.1 Simulation setup

Content-based component.

In order to assess the performance of our content-based component for web credibility evaluation, we compare our collaborative filtering approach to a machine learning approach (executed in Weka 3[7]). The Support Vector Machine (SVM) classifier is known for having good performance, in general, and thus we employ it as a base comparison against content-based collaborative filtering. We use a leave-one-out cross-validation to evaluate the performance of both approaches. This method consists in testing each page of the corpus independently, while training the model with the rest of the corpus.

Social component.

The social component requires several users that exchange their ratings. Although in the Reconcile data set (see Subsection 7.2) different users rate the same page, the Microsoft corpus provides only one credibility evaluation per document. We handle this issue by simulating multiple virtual users who rate pages from the Microsoft corpus, according to different behavioral types. We consider two types of behavioral profiles for the virtual users: A trustworthy profile according to which the user rates the pages with their corresponding ratings in the Microsoft data set. A malicious profile according to which the user rates most of the pages correctly except a specific subset of the corpus, namely the target pages, which represent the pages that the malicious user evaluates oppositely to the evaluation in the Microsoft data set. In our experiment, we gradually increase the percentage of malicious users and see how the weighted average and the social component react to these malicious users.

In practice, to simulate this multi-user environment, we create 100 users and, for each of them, we randomly select 100 pages across all categories to be stored in her history of visited pages; thus, we populate the neighborhoods of the collaborative filtering algorithm and avoid cold-start effects. Another 100 randomly selected pages are evaluated by each user for training her component weights and obtaining an optimal weighted average for each category. The Beta-distribution-based reputation formula for updating the weights is parameterized using $\beta = 0.6$, while the component weights are initially equal to $\frac{1}{3}$. Finally, 50 random trustworthy users randomly select one unvisited page from the subset of pages targeted by the malicious users and evaluate it with their social-based, content-based and search-ranking-based web credibility components. The weighted average of the component recommendations provide the credibility assessment for the web page. This prediction is finally compared with the ground truth represented by the Microsoft evaluation.

[7]http://www.cs.waikato.ac.nz/ml/weka/

Banishment system.

In order to assess the usefulness of the banishment approach, three different banishment configuration are tested: a) No banishment, in which case there is no penalty for a friend that provides a rating that is different than the user's a posteriori evaluation. b) Temporary banishment according to the approach of Section 5 with both base and initial exponent equal to 2. c) Lifetime banishment, which implies that if a friend provides an inaccurate rating, her ratings will not be used by the user for any future web credibility assessments.

7.2 Data sets

In order to evaluate the system, we use two different data sets: a) The Microsoft one that contains a single evaluation per page and b) the Reconcile one with fewer pages, but with the advantage of multiple evaluations per page. We describe these data sets in detail below.

Microsoft corpus.

This corpus comes from a study performed by Microsoft as part of [17]. It contains a set of 1000 pages belonging to 5 categories, namely: health, politics, environmental science, celebrities, personal finance.

Using Google zeitgeist[8], the authors in [17] were able to define for each topic, the 5 most popular queries performed on Google. Finally, each query has been provided to a search engine and the first 40 URLs for these queries have been used, i.e. 40 URLs x 5 queries x 5 topics = 1000 URLs. The original corpus has been provided with the credibility rating for each page given as a 5-point Likert-scale evaluation, with 0 denoting that the page is not credible at all, while 5 denoting high credibility.

The whole corpus has been entirely evaluated by one of the researchers. Then, several subsequent researchers as well as experts evaluated a subset of this corpus and verified that these evaluations are correlated to the initial rating. For convenience in the credibility evaluation, the ratings of this corpus have been transformed into a binary system, i.e. ratings 4 and 5 have been mapped to 1, ratings 1 and 2 have been mapped to -1, while ratings with 3 Likert points have been considered as neutral and extracted from the corpus. Thus, we end up with a corpus of 773 pages with binary credibility ratings.

Reconcile corpus.

This corpus has been created by researchers of the Polish Japanese Institute of Information Technology (PJIIT) of Warszawa in the framework of the Reconcile project. It is an aggregation of the evaluations of 90 students of PJIIT that have rated 9 web pages each from a corpus of 85 polish documents (all related to health topics).

The pages have been evaluated on a Likert 6 point scale but for the purpose of our experiments, as in the Microsoft corpus, each evaluation have been transformed to a binary scale, i.e. -1 (for ratings 0, 1 and 2) and 1 (for ratings 3, 4 and 5). This corpus allows us to evaluate the social component of our approach with real user evaluations instead of artificial ones by virtual users of different static behaviors.

7.3 Results

We first experimentally find the content features that maximize the effectiveness of our item-based collaborative filter-

[8]http://www.google.com/zeitgeist/

ing (CF) algorithm. At the same time, we verify the correctness of our CF content-based credibility evaluation approach by comparing its effectiveness with a machine learning (ML) classifier, namely SVM. As depicted in Figures 5, 6, 7, where we compare the precision and recall of CF and ML approaches per category of the Microsoft corpus for different feature sets, employing syntactic and lexical features has similar or better performance, as compared to using semantic ones. Also, as illustrated in Figures 5, 6, 7, our approach is a valid credibility assessment approach as the precision and recall of the CF algorithm are very close to those of ML for all different feature sets. Moreover, as depicted in Table 1, the area-under-the-(ROC)-curve (AUC) for the CF approach is higher than that of the ML one when the syntactic features are employed for all page categories, as opposed to using all features. Therefore, for simplicity in the computations and without loss of generality, in the rest of the experiments we only employed the syntactic and the lexical features in the content-based component. However, the AUC values in Table 1 are still not high enough. This means that our content component cannot perform adequately well when individually employed.

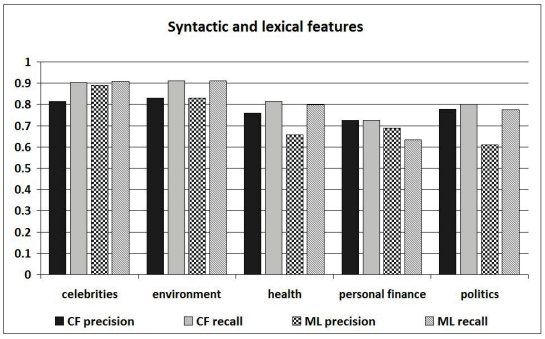


Figure 5: Syntactic and lexical features: Content component (CF) vs. SVM classification (ML).

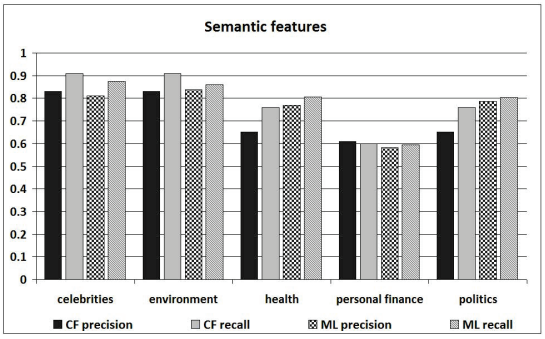


Figure 6: Semantic features: Content component (CF) vs. SVM classification (ML).

Indeed, combining the 3 components of our recommendation scheme by simple average, we vastly improve the performance of the content component when individually employed, as illustrated in Figure 8. In this experiment, we artificially varied the percentage of the malicious users in the population from 0 to 100; our recommendation scheme remains highly effective for fewer than 70% malicious users in the population.

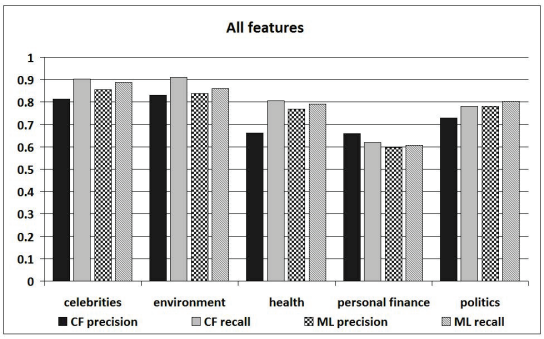


Figure 7: All features: Content component (CF) vs. SVM classification (ML).

Category	CF(a)	ML(a)	CF(b)	ML(b)
Celebrities	0.746	0.568	0.808	0.556
Environment	0.605	0.5	0.5	0.501
Health	0.675	0.491	0.409	0.601
Pers. Finance	0.755	0.567	0.638	0.58
Politics	0.776	0.495	0.796	0.586

Table 1: AUC of content component (CF) vs. SVM classification (ML) for different page categories when (a) only syntactic and lexical or (b) all content features are employed.

Since simply averaging the estimated ratings of the components of our recommendation scheme, one could argue that there is no need for finding the appropriate weights per page category. However, comparing Figures 9 and 10, we observe that different recommendation components can perform differently per page category: in this example, for the category "celebrities", the content component performs equally or better than the search-ranking one, and vice versa for the category "finance". Also, note that the relative performance of the social component depends on the fraction of malicious friends in the system. Therefore, a normalized weight approximating the relative performance of each component per category needs to be found.

As depicted in Figure 11, the weighted average component aggregation outperforms the simple average aggregation in terms of precision and recall when the component

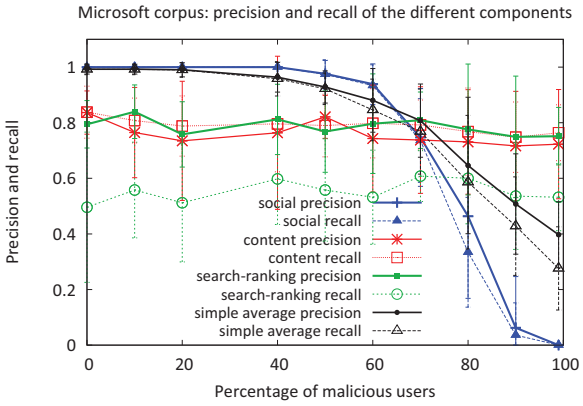


Figure 8: Microsoft corpus: Effectiveness of different components of the recommendation scheme for increasing fraction of malicious friends.

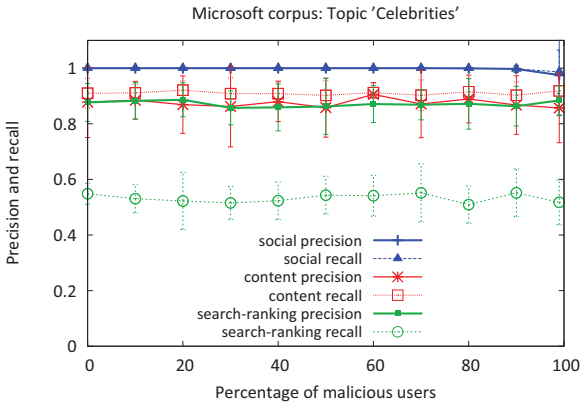


Figure 9: Microsoft corpus: Performance of different components for the category "celebrities". The content component has higher recall and almost equal precision as compared with the search-ranking component.

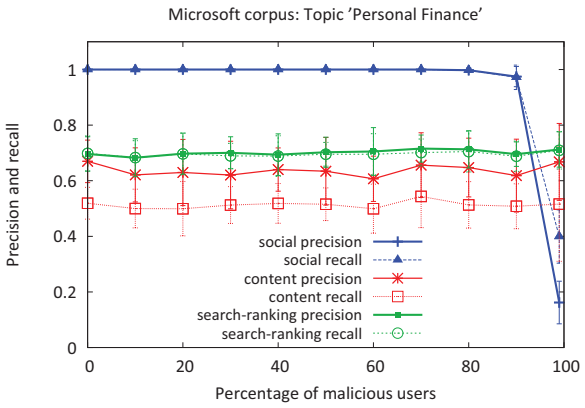


Figure 10: Microsoft corpus: Performance of different components for the category "finance". The search-ranking component has higher precision and recall than the content component.

weights are calculated according to the approach of Section 4. Similarly for the Polish corpus, we can easily see that the weighted average of the 3 components using different weights per category achieves higher effectiveness as compared to their simple average, both in terms of precision (cf. Figure 12) and of recall (cf. Figure 13).

Finally, we evaluate the effectiveness of our banishment mechanism for excluding the malicious users from the credibility assessment, while including all trustworthy ones at the same time. As depicted by Figures 14, 15, banishment increases the performance of our recommendation scheme for the Microsoft corpus, as compared to no banishment (cf. Figure 11). This was expected, since in this corpus, any disagreement in the credibility evaluation (cf. Section 5) means the discovery of a malicious user that has to be banned. However, as depicted in Figure 16 (as compared to Figure 12), lifetime banishment deteriorates the effectiveness of our recommendation scheme for honest -but occasionally subjective- friends, as is the case for the Reconcile corpus.

8. RELATED WORK

There is significant amount of related work in the literature dealing directly or indirectly with web credibility assessment. A number of them deal with defining web credibility and the main features involved in its assessment [12, 4, 13, 17, 18]. Kapoun in [12] perceived credibility as a journalism quality depending on accuracy of information, authority of the source, objectivity, currency and topic coverage. Fogg in [5] added in those factors the impact of prominence of different page features on personalized credibility perception.

Closer to our work, several approaches have been proposed for enhancing the search results with page credibility information [13, 17, 18]. Nakamura *et al.* [13] developed a system prototype augmenting the search results for each page with three attributes, namely topic majority (i.e. number of retrieved pages sharing same topic with the target page), the topic coverage and the locality of supporting pages ((i.e. pages linked to each search result). Schwarz *at al.* in [17] additionally take a social aspect by visualizing a different set of attributes, including the overall web page popularity, the popularity among experts, the location origin of the page hits, any awards or certifications of the page and the PageRank metric. However, both approaches in [13] and [17] had limited effectiveness for improving the page credibility assessment. Yamamoto *et al.* in [18] proposed an iterative approach based on which users evaluate the importance of different visualized page aspects in their credibility assess-

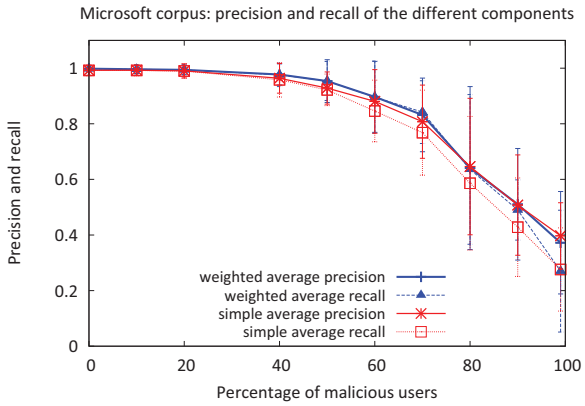


Figure 11: Microsoft corpus: Effectiveness of component weight updating per page category vs. simple average.

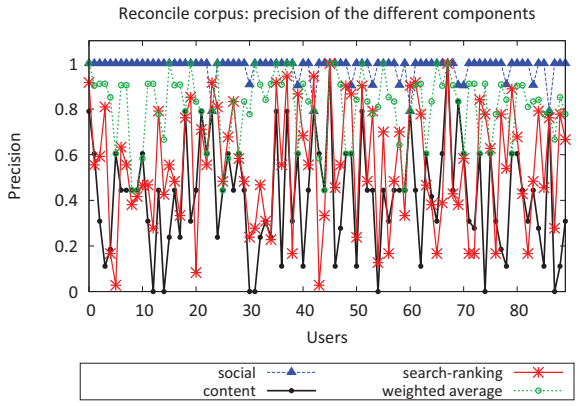


Figure 12: Reconcile corpus: Precision of different components for different users.

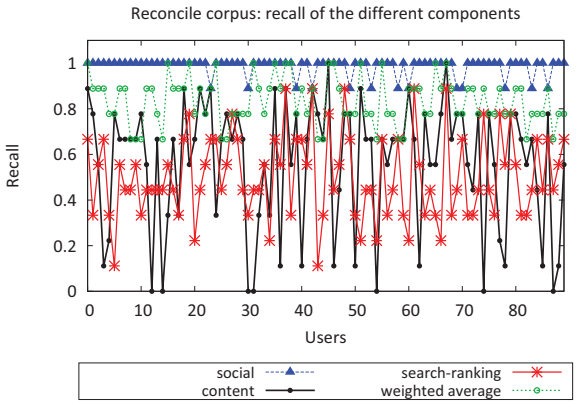


Figure 13: Reconcile corpus: Recall of different components for different users.

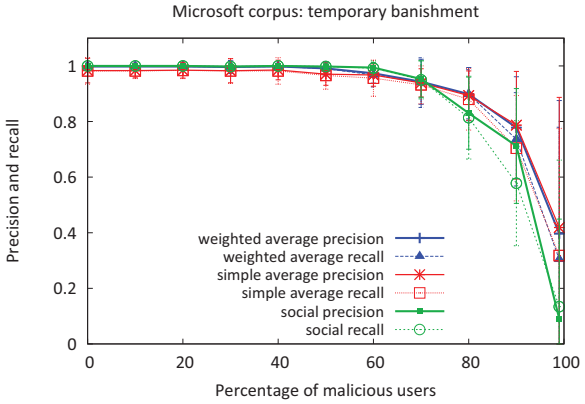


Figure 14: Microsoft corpus: Credibility assessment with temporary friend banishment upon wrong evaluation.

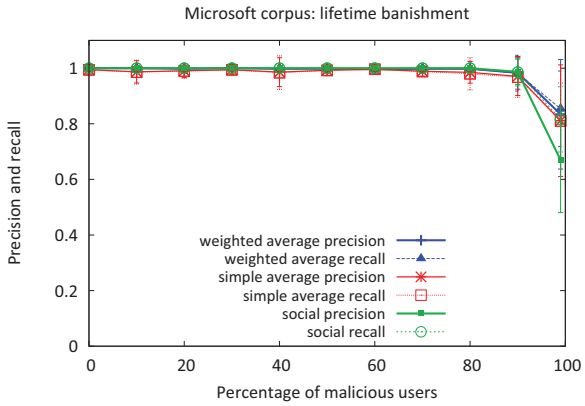


Figure 15: Microsoft corpus: Credibility assessment with lifetime friend banishment upon one wrong evaluation.

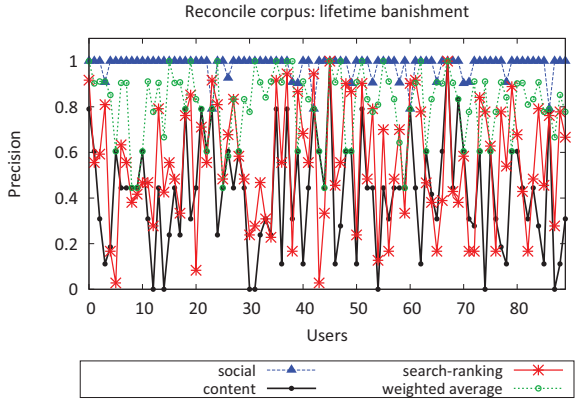


Figure 16: Reconcile corpus: Precision of credibility assessment with temporary friend banishment upon wrong evaluation. The effectiveness does not drop significantly due to the unfair banishments of honest friends that occasionally disagree.

ment, the system updates their significance in the credibility estimation and the search results are re-ranked. The visualized page aspects in [18] were the referential importance, the social reputation, the content typicality, the topic coverage, the freshness and the update frequency. Our system also takes into account the user feedback for updating the weights of the components of the recommendation scheme and for the banishment of friends from future credibility evaluations.

Additional approaches attempt automatic credibility evaluation based on the aggregation of different sets of features. For example, type of website, date of update, sentiment analysis and PageRank metric are utilized in [1], while information commonality, source independence, prestige of the source and experience with the source have been considered in [11]. However, none of these approaches employs a holistic approach such as ours to the problem of credibility assessment that combines a social-based, a content-based and a link-based component. Note that additional features that fall in the semantic scope of these three components can also be integrated in our approach.

Also, the robustness of web page ranking algorithms has been studied in [2, 6]. Both works study the resilience of web page ranking algorithms against web spam via link structure and credibility analysis. The use of trust and reputation mechanisms to minimize the influence of adversarial attacks in ranking systems has also attracted much effort [10]. Also, [6] uses reputation-based trust management techniques to improve the robustness of ranking systems, yet with little analysis on the impact of trust mechanisms to the adversarial cost for strategic manipulation of the system. Our multi-component recommendation scheme is less sensitive to link spamming by properly updating the component weight and by banishing the malicious users.

Collaborative filtering techniques are used in [3, 19] to weight ratings in trust estimation proportionally to the similarity of preferences between the agent who computes the estimate and the raters. Only ballot stuffing and positive discrimination are dealt with in [3] for adversaries constituting up to 10% of the population, as opposed to our approach that deals effectively with higher fractions of malicious friends.

As already mentioned, there are some system initiatives towards higher web page credibility by employing a socially-informed component. A popular one is the inclusion of the Facebook "like" button in a web page. When someone "likes" a page, this information is attached to the page and the total

number of likes is displayed in the page. However, although useful for web page credibility assessment, this approach mostly measures the popularity of a web page. Another similar approach is the "+1" button of Google+. According to this mechanism, a user can positively rate a certain page and this information can be shared with her friends or everyone. This information is integrated with the PageRank algorithm (in a proprietary way) and ideally results to a personalized ranking of web search results. The same idea with Google+ was employed in the search engine SearchWinds[9], which is now integrated into Microsoft Bing search engine[10]. Initially, all pages are assumed to be credible and their credibility is updated based on ratings by the members of the search engine. However, these approaches enjoyed only limited user adoption so far, partially because of user privacy concerns, as the user ratings can also be employed for data mining purposes. Our recommendation scheme could be integrated to Google+ or Bing for ranking the search results, while our decentralized system facilitates user adoption.

9. CONCLUSION

In this paper, we proposed a decentralized recommendation system for credibility evaluation. Our recommendation scheme employs a social-based, a content-based and a search-ranking based component, which are combined in a weighted average. The weights are properly adjusted to reflect the effectiveness of each individual component for credibility assessment. Moreover, our approach is robust against malicious ratings; we proposed a mechanism that bans a friend from the user neighborhood for an adequate number of web page credibility evaluations according to its malicious or not nature. Our simulation results with 2 real data corpus of web page credibility evaluations suggest that our approach is promising. As a future work, we plan to integrate into the item-based component the semantic distance between pages based on the page category and a certain category ontology.

Acknowledgments

This work was partially funded by the grant *Reconcile: Robust Online Credibility Evaluation of Web Content* from Switzerland through the Swiss contribution to the enlarged EU.

10. REFERENCES

[1] S. Aggarwal1 and H. V. Oostendorp. An attempt to automate the process of source evaluation. In *Proc. of the Int. Conference on Advances in Computer Engineering*, July 2011.

[2] J. Caverlee, S. Webb, L. Liu, and W. B. Rouse. A parameterized approach to spam-resilient link analysis of the web. *IEEE Trans. Parallel Distrib. Syst.*, 20(10):1422–1438, 2009.

[3] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behaviour. In *Proc. of the ACM EC*, New York, NY, USA, 2000.

[4] B. J. Fogg. Prominence-interpretation theory: explaining how people assess credibility online. In *Proc. of the CHI*, New York, NY, USA, 2003.

[5] B. J. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, and M. Treinen. What makes web sites credible?: a report on a large quantitative study. In *Proc. of the CHI '01*, New York, NY, USA, 2001.

[6] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Proc. of the VLDB*, 2004.

[7] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proc. of the ACM SIGIR*, SIGIR '99, New York, NY, USA, 1999.

[8] C.-F. Hsu, E. Khabiri, and J. Caverlee. Ranking comments on the social web. In *Proc. of the CSE*, Washington, DC, USA, 2009.

[9] A. Jøsang, S. Hird, and E. Faccer. Simulating the effect of reputation systems on e-markets. In *Proc. of the 1st International Conference on Trust Management*, Berlin, Heidelberg, 2003.

[10] A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decis. Support Syst.*, 43(2):618–644, 2007.

[11] A. L. Kaczmarek. Automatic evaluation of information credibility in semantic web and knowledge grid. In *WEBIST*. INSTICC Press, 2008.

[12] J. Kapoun. Teaching undergrads web evaluation: A guide for library instruction. *College and Research Library News*, 59(7), July/August 1998.

[13] S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. Tezuka, S. Oyama, and K. Tanaka. Trustworthiness analysis of web search results. *LNCS*, 4675:38–49, 2007.

[14] T. G. Papaioannou and G. D. Stamoulis. An incentives' mechanism promoting truthful feedback in peer-to-peer systems. In *Proc. of the CCGRID '05*, Washington, DC, USA, 2005.

[15] A. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems. In *Proc. of the IFIP/ACM International Conference on Distributed Systems Platforms (Middleware)*, Heidelberg, Germany, November 2001.

[16] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proc. of the WWW '01*, New York, NY, USA, 2001.

[17] J. Schwarz and M. Morris. Augmenting web pages and search results to support credibility assessment. In *Proc. of the CHI*, New York, NY, USA, 2011.

[18] Y. Yamamoto and K. Tanaka. Enhancing credibility judgment of web search results. In *Proc. of the CHI*, New York, NY, USA, 2011.

[19] G. Zacharia, A. Moukas, and P. Maes. Collaborative reputation mechanisms in electronic marketplaces. In *Proc. of the HICSS*, Washington, DC, USA, 1999.

[9] http://www.searchwinds.com

[10] http://www.bing.com