Modeling and Evaluating Credibility of Web Applications

Sara Guimarães
Dept. of Computer Science
Federal University of Minas Gerais (UFMG)
Belo Horizonte, MG, Brazil
sara@dcc.ufmg.br

Adriano Pereira
Dept. of Computer Engineering
Federal Center for Technological Education of Minas Gerais (CEFET-MG)
Belo Horizonte, MG, Brazil
adriano@decom.cefetmg.br

Arlei Silva, Wagner Meira Jr.
Dept. of Computer Science
Federal Univ. of Minas Gerais
Belo Horizonte, MG, Brazil
arlei@dcc.ufmg.br
meira@dcc.ufmg.br

ABSTRACT

The popularization of the Web has given rise to new services every day, demanding mechanisms to ensure the credibility of these services. In this work we adopt a framework for the design, implementation and evaluation of credibility models. We call a credibility model a function capable of assigning a credibility value to a transaction of a Web application, considering different criteria of this service and its supplier. To validate this framework and models, we perform experiments using an actual dataset, from which we evaluated different credibility models using distinct types of information. The obtained results are very good, showing representative gains, when compared to a baseline and also to a known state-of-the-art approach. The results show that the credibility framework can be used to enforce trust to users of services on the Web.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services—*online reputation systems*

General Terms

Experimentation, Management, Security

Keywords

Credibility, ranking, trust management, e-markets, Web 2.0

1. INTRODUCTION

The task of evaluating and quantifying credibility in a Web application represents the major challenge of this research. Among the main difficulties of this task, we can highlight the large number of variables involved and the low reliability of the information available. Considering these, and other difficulties related to this task, we proposed, in a previous work, a framework, named *CredibilityRank*, which allows the development of credibility models that can generate crediblity rankings based on several different criteria.

In this work, we use a framework for the design, implementation and evaluation of credibility models. This evaluation is based on a representative sample of services provided by users, with their respective feedback, and a ranking that represent a scale of credibility generated by the model. The greater the capacity of the model to position vendors that offer satisfactory services (which are qualified as such from the user feedback) in the top positions on this scale, the higher its quality.

It is important to explain that, despite the problems of reputation systems [12], it is necessary to use feedback information to measure the user opinion of a service that can be described by different characteristics that we denote credibility attributes in our model. Moreover, there are specific researches that deal with improving the quality of reputation systems, such as the identification of fraudsters of these systems [10], which was also used in the same real application used in this research.

In the recent years, the concept of credibility has begun to be studied on the Web in order to measure whether a user relies on a service or information available. It is a consensus in the literature that credibility can be subjective to the user, but it also depends on objective measures. The credibility of Web applications has become a multidisciplinary subject, where researchers from communication have been focusing on a more qualitative (and subjective) analysis of credibility [4], while the area of computer science has focused on more objective metrics. The methods proposed in the area of computer science are strongly based on trust and reputation [6], and credibility rankings that take into account the source of information [2] and its content [9].

We perform experiments using an actual dataset of a Web application, from which we evaluated different credibility models using different types of information sources, such as attributes related to offer's characteristics, seller's expertise and qualification. We describe a methodology to combine different credibility models in order to maximize the credibility, validating it through a complete case study. The results show that the credibility framework can be very useful and promising. The obtained results were very good, showing representative gains, when compared to a baseline and also with a known state-of-the-art approach.

2. FRAMEWORK DEFINITION

In this section we briefly describe *CredibilityRank*, a new framework for the design and analysis of credibility models. This framework is based on a new modeling for the problem of credibility on the Web. It was first described in a previous work [1] and the version used in this work presents some improvements.

The credibility of a service is associated to some information that characterizes or describes this service. This information can be related to the service's configuration and the user that offers it. When someone uses a service, a transaction which contains information

WebQuality '11, March 28, 2011 Hyderabad, India.

about the service and its supplier, is registered. From each transaction of a service is possible to extract a set of information that can be used to model the credibility. Moreover, from the feedback that an user provides to a transaction it is possible to assess their satisfaction level with it.

A credibility model is a function that allows the comparison of services in terms of credibility. An effective credibility model is able to identify which are the most trustworthy services and point out which ones are less reliable for a given user.

Definition A credibility model $\mathcal{M}$ is a function that receives a set of services $S = s_1, s_2, \ldots s_n$, where n is the number of services and s_i is a tuple of attributes of the service i, and returns a ranking R, where services are positioned in terms of credibility based on their attributes. A ranking can be described by a function $R : i \rightarrow \mathbb{N}$, where $R(i)$ is the position of the service i and $0 \leq R(s_j) < n, \forall s_j \in S$. The higher is the credibility of a given service i according to M the lower is the value of $R(i)$.

A tuple s_i can be composed of different service attributes related to the service supplier and also to the specific service evaluated. For instance, in a video sharing site a service s_i should contain information about how long the user who uploaded the video has been subscribed on the site, the number of views and positive evaluations received by the video. Based on this information, a credibility model may enable the user to distinguish which videos are spams or disagree with the description displayed.

Figure 1 illustrates the steps followed by our framework. It reproduces the user interaction with the credibility model in order to assess the quality of the rankings generated in terms of credibility. First, the history of transactions is divided into service tuples (i.e., the attributes of the service that originated each transaction) and user feedback received. Then, the service tuples are ranked by the credibility model. The quality of the credibility model is evaluated using the users' feedback in a final step.

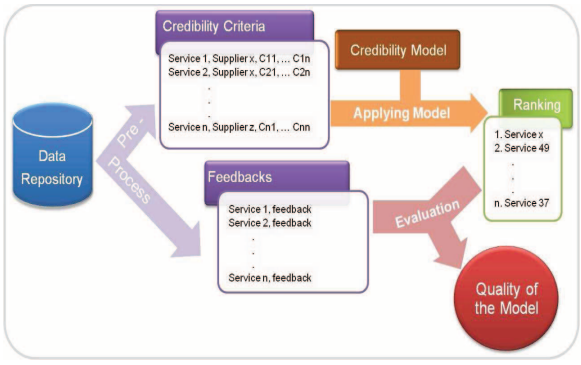

Figure 1: The *CredibilityRank* Framework Scheme

Algorithm 1 is a high-level description of *CredibilityRank*. Besides the extraction of the service tuple and feedback from a log of transactions $\mathcal{D}$, the pseudo-code describes how the framework iterates over the set of transactions in $\mathcal{D}$. For each transaction t, the algorithm identifies the set of services A that were active when t was performed. It is also necessary to identify the specific service s associated to t. Through the application of the credibility model $\mathcal{M}$ receiving A and the service s we get the rank R. Positions of the services in the rank ($position(s, R)$) and the feedback F_t received by each transaction are included in a set S, which will be used in the evaluation of $\mathcal{M}$.

Algorithm 1: **CredibilityRank Framework**

Input : $\mathcal{M}, \mathcal{D}$
Output: Evaluation of $\mathcal{M}$
$\mathcal{F} \Leftarrow get\_user\_feedback(D)$;
$\mathcal{T} \Leftarrow get\_transactions(D)$;
$R \Leftarrow \emptyset$;
for $t \in \mathcal{T}$ **do**
  $A \Leftarrow active\_services(t, T)$;
  $s \Leftarrow get\_service(t)$;
  $R \Leftarrow \mathcal{M}(s, A)$;
  $S \Leftarrow S \cup (position(s, R), F_t)$;
$Evaluate(S)$;

The framework does not define a specific evaluation procedure, since it may depend on the Web application analyzed. However, several rank-based evaluation metrics may be used in the framework depending on the application scenario.

3. CASE STUDY

This section presents our case study where we apply and evaluate some credibility models using the *CredibilityRank* framework and actual data from an electronic market. A dataset description is given in Section 3.1 and the experiments are detailed in Section 3.2.

3.1 Dataset Description

TodaOferta[1] [11], which is a marketplace developed by the largest Latin America Internet Service Provider, named Universo Online Inc. (UOL), is a website for buying and selling products and services through the web.

Table 1 shows a summary of the *TodaOferta* dataset. It embeds a significant sample of users, listings, and negotiations. Due to a confidentiality agreement, the quantitative information about this dataset can not be presented. The subset of this dataset that we have used in this research has some tens of thousands of transactions.

Coverage (time)	Jul/2007 to Jul/2009
#categories (top-level)	32
#sub-categories	2,189
Average listings per seller	42.48
Negotiation options	Fixed Price and Auction

Table 1: TodaOferta Dataset - Summary

3.2 Methodology

This section presents our methodology to evaluate credibility, which we are going to apply to the actual dataset previously described. Our methodology follows the process illustrated in Figure 1. We define some credibility criteria (attributes) from the original dataset source. As we are going to describe, we use a strategy to create credibility models based on these attributes and combinations of them, besides comparing them with a baseline and a state-of-the-art model.

Given a set of combined attributes c_i, we define a credibility model $\mathcal{M}_i$ that ranks the services according to c_i. The highest is the value of c_i for a given service s the highest is the rank of s. The results obtained by these combined-attribute models are important to motivate the design of new credibility models based on more accurate techniques to choose and automatically combine attributes in order to quantify more accurately the credibility of services. We selected 15 attributes to be used and combined in the ranking of services, such as:

[1]http://www.todaoferta.com.br

- **Price**: price of the product/service being offered.
- **Views**: the number of visualizations of the listing.
- **Percentage Positive Qualifications**: the relative amount of positive qualifications a user (seller) has received.
- **Global Score**: the seller reputation rating score, considering the different score types.
- **Average Negotiated Value**: the average price per transaction performed by the seller.
- **Retailer**: indicates whether the user is considered a powerful seller by *TodaOferta*.

For more detail about this dataset and attributes, refer to [1]. Each attribute can be used and/or be combined with other ones to generate different credibility models $\mathcal{M}_i$. A model ranks the active services, considering the transactions (i.e., sales performed in the e-market), according to the given set of combined attributes.

We adopt a method to combine these models $\mathcal{M}_i$, which have one or more credibility attributes, producing a new credibility model. Our proposal is to join them in pairs (two-to-two), choosing the best k models to continue this process until a stop condition, which we determine as:

- A fixed number of iterations N to be performed; or
- A minimum increase in the gain obtained from the best models of one interaction i to the previous one $(i-1)$.

In order to evaluate the credibility models we are going to use some baseline models, which are determined using some attributes that are usually presented by the Web application to its users: *Global Score*, *Percentage of Positive Feedback* and a balanced combination of them (i.e., the arithmetic mean). Despite of being presented to users, it is important to explain that these attributes are not showed to users as a ranking of credibility as we are proposing to do with the objective of comparing them to other credibility models.

Moreover, we are going to apply a state-of-the-art method to compare with our models, which is Support Vector Machine (SVM) applied for ranking - SVM-Rank [7, 8].

The metric applied in the evaluation of credibility models is the probability of receiving negative feedback at the X% top positions of the rank, which are used to generate a graph that has in axis x the position of the rank. The probability of negative feedback is an intuitive evaluation metric that presents as an important characteristic to be more trustable than the positive feedback in most of the e-markets, since typical attacks try to inflate the sellers' reputation through positive feedbacks [10]. As long as the selected metric considers only information about negative feedbacks, which can assumed to be trustworthy, it can provide a more accurate analysis of the quality of credibility models.

We can analyze credibility models using this graph representation, observing which one has the curve with best behavior in terms of credibility, that is, a smaller inclination during the top positions of the ranking and the opposite behavior at the last positions (bottom of the ranking). This analysis are going to provide a general view of the quality of the credibility models.

Moreover, it is important to measure with more confidence this quality, trying to determine an indicator that measures this quality. The area under the curve (AUC) represents the cumulative amount of negative feedback in a specific interval of the ranking [3]. In order to measure the credibility of a service (transaction), we define a simple indicator, called **CI** (Credibility Indicator), which can be calculated by the inverse area under the curve of a model, that is:

$$CI = 1/AUC \tag{1}$$

Thus, **CI** can be used to determine which are the best models considering both the top and the bottom of the rank. Such indicator is a good metric to compare the models, since it is inversely proportional to the amount of untrustable services at a specific part of the rank. Therefore, a good model to measure credibility for the top of the ranking is one that has high values of **CI** (and thus a small area under the graph). The opposite is true for models for the bottom of the rank.

3.3 Experiments and Results

In this section we apply the *credibilityRank* framework to evaluate some credibility models using *TodaOferta* dataset.

Table 2 presents the credibility models that are used in this case study, and describes how they were obtained from the set of the attributes presented. Some of them use different weights to a given attribute, which is the same as recombine it with a model that already had it. The table also shows the baselines and the SVM-Rank models, and how they are obtained. From now on, we are going to reference the models by the notation presented in this table.

Top	Top_1	2*Retailer + Duration + %Pos.Feedback
	Top_2	Retailer + Duration
	Top_3	6*Retailer + 5*Duration + 3*%Pos.Feedback + 2*Certified
Bottom	Bottom_1	Average Negotiated Value + Views
	Bottom_2	2*Average Negotiated Value + Views + Offer w/ SafePayment
	Bottom_3	Average Negotiated Value + Views + Safe Transaction
Baselines	BaseLine_1	Global Score + %Pos. Feedback
	BaseLine_2	%Positive Feedback
	BaseLine_3	Global Score
SVM-Rank	SVM_1	5.000 training set
	SVM_2	10.000 training set
	SVM_3	20.000 training set

Table 2: Notation and Description of Credibility Models

Figures 2 and 3 show the best results obtained by the credibility models considering the top and the bottom of the ranks, respectively. To evaluate which ones are the best models in each case, we use the discussed metric **CI**.

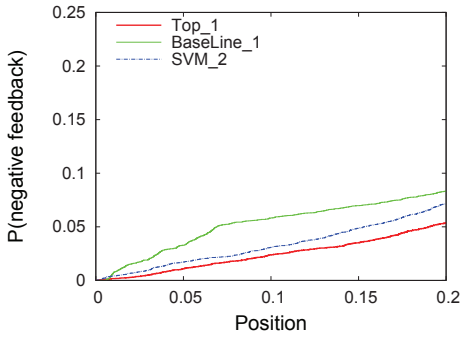

Figure 2: Credibility Models - Comparison of the best models at the TOP of the rank

Figure 2 shows a comparison among the best model of each group presented for the top positions of the rank. As discussed before, Top_1 is the best model from its group, but is also the best of all, presenting the lower probability of receiving negative feedback at the first 20% top positions. $BaseLine\_1$ is the best model considering only the baselines, but when compared to the bests of each category, it appears as the worst of them, showing a great difference in terms of the **CI** value to the others. Analyzing the model SVM_2, we can see that its curve is situated among the other two, even though it appears closer to the best model curve.

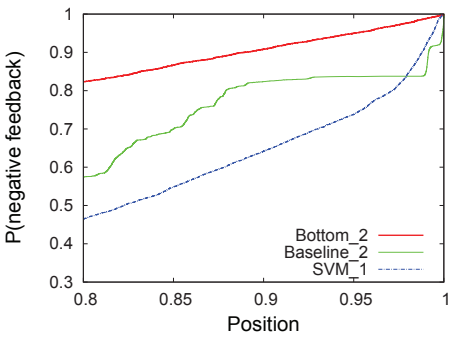

Figure 3: Credibility Models - Comparison of the best models at the BOTTOM of the rank

Figure 3 compares the best model of the three previous groups analyzed here. $Bottom\_2$ is clearly the best model, since it has a area under curve more than 30% higher than the second best model, which is $BaseLine\_2$. The model SVM_1 appears as the worst of the three, despite being better than the baseline model at the end.

Through this analysis we can conclude that the attributes *Duration* and *Retailer* are good indicators of the credibility of a service to the top positions of the rank, since they appear in all the best three models presented for the top. The attributes *Views* and *Average Negotiated Value* are also good indicators, but to measure credibility at the last positions, appearing in all the three models that considers the bottom of the rank.

4. CONCLUSION

In this work we presented and evaluated some credibility models for an e-business application (electronic marketplace), using actual data. Moreover we have compared these models with a baseline, which adopt information that users have available from the Web site of this marketplace. We also evaluate a simple method to combine credibility models to generate more accurate models, besides comparing them to a known state-of-the-art technique (SVM-Rank).

We perform experiments to evaluate different credibility models using different types of information sources and an actual dataset from an electronic market - the *TodaOferta*. The results show the applicability of the proposed framework.

The results were very good, showing significant gains for both top and bottom ranking positions. For the top (where the best transactions would be positioned), our best credibility model (Top_1) obtains a gain of 116.8% in comparison with the baseline and a gain of 36.4% over the SVM-Rank. For the bottom ranking position (where the worst transactions would be ranked), our best approach outperforms the baseline and SVM-Rank in 24.6% and 37.8%, respectively.

These results are promising and the research provides insights that contribute to understand the scenario used as case study and to identify directions for future work. The analysis also shows that a good model can be obtained from the combination of few attributes of credibility, and not necessarily the use of a greater number of attributes results in an increase in the quality of the generated models.

As ongoing work we want to improve the evaluation and analysis of the credibility models that we have presented in this work. Moreover, we want to implement new credibility models based on techniques of machine learning and genetic algorithms [5]. Finally, we also want to apply the framework in other Web scenarios, such as digital libraries and social networks.

5. ACKNOWLEDGMENTS

This work was partially sponsored by Universo OnLine S. A. - UOL (www.uol.com.br) and partially supported by the Brazilian National Institute of Science and Technology for the Web (CNPq grant no. 573871/2008-6), CAPES, CNPq, Finep, and Fapemig.

6. REFERENCES

[1] S. G. aes, A. Pereira, A. Silva, and W. Meira Jr. Credibilityrank: A framework for the design and evaluation of rank-based credibility models for web applications. In *Proc. of the 2010 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing / Symposium on Trusted Computing and Communications (TrustCom'2010)*, pages 504–510, Hong Kong, China, 2010. IEEE CS.

[2] A. Amin, J. Zhang, H. Cramer, L. Hardman, and V. Evers. The effects of source credibility ratings in a cultural heritage information aggregator. In *WICOW '09: Proceedings of the 3rd workshop on Information credibility on the web*, pages 35–42, New York, NY, USA, 2009. ACM.

[3] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861 – 874, 2006. ROC Analysis in Pattern Recognition.

[4] A. J. Flanagin and M. J. Metzger. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media Society*, 9(2):319–342, April 2007.

[5] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, 1 edition, January 1989.

[6] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 403–412, New York, NY, USA, 2004. ACM.

[7] T. Joachims. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, New York, NY, USA, 2002. ACM.

[8] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[9] A. Juffinger, M. Granitzer, and E. Lex. Blog credibility ranking by exploiting verified content. In *WICOW '09: Proceedings of the 3rd workshop on Information credibility on the web*, pages 51–58, New York, NY, USA, 2009. ACM.

[10] R. Maranzato, A. Pereira, A. P. do Lago, and M. Neubert. Fraud detection in reputation systems in e-markets using logistic regression. In *SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1454–1459, New York, NY, USA, 2010. ACM.

[11] A. M. Pereira, D. Duarte, W. M. Jr., V. Almeida, and P. Góes. Analyzing seller practices in a brazilian marketplace. In *18th International World Wide Web Conference*, pages 1031–1041, April 2009.

[12] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. *Commun. ACM*, 43(12):45–48, 2000.