WebTrust
Agent Platform for Identifying Trustworthy Commercial Websites

ir. W.F. Stoutjesdijk
Accenture Technology Labs
BP 99, Les Genêts
449 route des Crêtes
06902 Sophia Antipolis, France

drs. dr. L.J.M. Rothkrantz
Delft University of Technology
Mekelweg 4
2628 BZ Delft, The Netherlands
L.J.M.Rothkrantz@cs.tudelft.nl

August 7, 2001

Abstract

In this document we present a blueprint for agents that can label websites according to a personalised view of trustworthiness. This trustworthiness goes beyond the technical trust that is implemented in security related issues and covers the whole 'feeling' of a trusted website selling a product. A model of site properties that contribute to this idea of a (dis)trusted site was designed. Agents will learn how their user reacts to these properties and calculate a trust-value upon it. With enough trained sites the agents can start labeling new websites. When the user is not acquainted with the topic of the site he can ask the agent of an expert or a consumer magazine to help: agents can negotiate with each other to get a more profound rating.

1 Introduction

On the ever growing Internet more and more products are being sold. To get the best price for an item one has to search a lot of sites. Only some of them look as if they will actually ship your order, handle your credit-card number properly and not pass your personal information around: not all sites can be *trusted.*
In this document we present agents that can help decide whether or not to order, by calculating a personal trust-label for websites. These agents can negotiate with other agents to come up with a shared trust rating.

2 Related work

2.1 Human trust

In his "Formalising Trust as a Computational Concept" [5], Marsh describes trust as a human phenomenon in 3 layers:

Basic trust; inherent in the trusting person. One can be an easy truster or a hard one. As with more things in life we can distinguish between optimists and pessimists: "...optimists are general trusters whose disposition to trust is relatively inflexible in a downward direction, despite past experience. Thus their trust in others can only increase, never decrease. The opposite is true for pessimists" (page 56). It is the way they look at life (from a trust perspective). They are gullible, suspicious or something in between.

General trust; measured from one subject to another. In other words how much person x trusts person y in general as a person.
Seen from a commercial perspective this is a customer-company relation. It is calculated regardless of the medium (shop, Internet or telephone). It is purely what the customer thinks of the company.

Situational trust; the most specific of the three describes trust in a specific context. The example Marsh gives is very clear: "while I would trust my brother to drive me to the airport, I most certainly would not trust him to fly the plane!".
In other words one can have General Trust, but not in every situation. For trust on the Web the site is the context: although you trusted this company, their site changed this perception.

Temporal issues are also discussed in his model, but will not be taken into account in WebTrust. From here on, trust factors can be seen from these three perspectives.

2.2 Trusting websites

Since the start of Internet commerce, company websites have been under scrutiny of trust. People are suspicious towards these sites. Cheskin did a survey on trust amongst Internet users in the Americas [1] [2]. Two of their main findings are:

1. Some people take more risk in trusting than others: they trust more. Cheskin actually confirms Marsh's idea here that people have different Basic Trust.
2. Respondents said to look at specific properties of the site. They looked for the presence of seals of approval, brandnames and layout. These seals of approval are logos from 3^{rd}-parties that guarantee security (VeriSign and BBB Online) or insure the transaction (American Express and Visa). These parties benefit from the quality of the transaction and associate themselves with 'trusted' sites. They are also called trusted 3^{rd}-parties.

Egger [3] makes use of this information to advise companies on how to make trusted websites. He makes the companies ask their customers what they think of the company's reputation, the content on the site, which 3^{rd}-party to choose and what the privacy policy should say.

2.3 Web of trust

Consumers do not have to depend on trusted parties at eOpinions [4]. On this website, where consumers review products and sites, everyone is part of the *Web of trust*. Users have a portfolio of people they trust. They do this either because they have read their reviews and concur with them, or because they trust the people *this* user has in his portfolio of trusted users. The idea behind this web of people is based on the assumptions that

- friends have a proven track-record. If a friend consistently gives you good advise, you are likely to trust him the next time too;
- you and your friend share interests. If you both like the same kind of films, you are more likely to trust his recommendations.

3 Feature Model

In order to get a more complete picture of the properties of sites mentioned by Cheskin, a survey was done. This survey asked 19 respondents[1] their trust opinion on 50 websites. On a 5-point scale they indicated how much they would trust sites (or the companies behind them) to do what they promised to do. These promises range from correct handling of payment and credit-card related issues to shipping the order and not spreading around personal information to others.
Aside from this, respondents had to write down anything that made them trust this site the way they did: the properties of the site and related issues (the fact that the company is public on the NASDAQ for example).

The reasons people submitted for their trust opinion were split up into 22 categories. Some of these categories are regarded as not trust-related: respondents indicating that they were not interested in the product or service, for example. From the others, the following 5 scored best (in order):

1. Design; the layout of the site was often mentioned positively and negatively. Sites that did not live up to the standards set by the 'professionals' and did not have anything else to offer (e.g. brandname) were rated negatively. Whereas websites with no track-record at all had benefit from their appearance.
2. Reputation; this category of answers was used positively in every case. When a company or its website is well-known, people have an opinion on it. This is often positive because respondents reason that a well-known company has to deliver good service. This category does not include 'experience' with the site, the product or the company.
3. Company information; respondents looked for real-world information on the website. They wanted to know where the company was based, what its phone number was etc. Contact information other than Internet related, gives more confidence in the company.
4. Privacy Policy; many sites have a privacy policy that states how the company will treat all personal information supplied to it. Respondents never mentioned any statements made inside this policy, just its existence.
5. Trusted Seals; as mentioned before, these seals are logos from organisations that are regarded as trusted.

Apart from the company information, no text-related issues are mentioned here. People do not seem to *read* the site, but merely *look* at it. These findings are confirmed by Cheskin's survey mentioned earlier, in which people looked for trusted seals, brandnamens (a form of reputation) and layout (design).

The ratings respondents gave to the sites were very conservative. Even though some of the sites could be regarded as awkward, the answers average on the 'Neutral' score. In Table 1 the ratings can be seen per scale-point.

To confirm the fact that trust is a subjective issue the agreement amongst respondents was measured. The 5-point ratings were mapped onto 'Negative', 'Neutral' and 'Positive'. For every site, the number of these ratings were counted. The agreement is then defined as the maximum group of ratings divided by the number of ratings. If 5 people said 'Negative', 1 said 'Neutral' and 2 said 'Positive' then the largest group is 'Negative'. The agreement would be $\frac{5}{5+1+2} = 63\%$.
This agreement measure strongly suggests that trust is subjective: only 6 out of 50 sites score an agreement of more than 75%. People do not seem to have the same

[1] 10 of which filled in the complete survey and are used in further analysis.

Table 1: Spread of the trust ratings

Not at All	*A little*	*Neutral*	*A lot*	*Absolute*
13%	21%	29%	24%	14%

trust rating for sites.

From work presented earlier and this survey, trust on the Web was understood as a process: people investigate the site for properties they find important and supply these properties with their personal weights. The 'sum' of these weighted properties will determine their judgment. The investigation of a site can involve the following:

Experience; the user will see if he knows the company or site already. If he does, he already has an idea about the trust he can attribute this site.

Layout; the design of the site is taken into account. Especially for unknown sites this can be very important. It is a sign of professionalism when it looks well. This will be reiterated later in this section.

Interaction; part of the layout are links and buttons. These are susceptible to errors. If the designer is not careful they may not work: the link does no longer exist or the button generates some kind of error. All these things can be discovered by interacting with the site: clicking on it.

Other people; finally the opinion of other people is important as shown by the eOpinions website, discussed before. In general these people can be friends, consumer magazines, newspapers or any other trusted party with an opinion. The trust rating of these others with regards to this particular site is of importance to the user.

All these questions asked and tests performed to assess trust are used to measure the *features* on a website. A feature is defined as a measurable entity present on, or related to a site or the organisation behind it.
It is the basis of the *Feature Model.* Except for the 'Other people' item which is part of the negotiating agents principle, explained later on.
The Feature Model consists of a list of all features identified by this research together with their metric, split up into 4 categories defined for it:

Reputation is an easy way for people to decide on a site. If they know it already because they read about it in a magazine, saw the company on television or heard about it from a friend they have an initial idea about its trustworthiness. Sometimes this idea is so strong that they will decide without really taking the site into account or even contradicting their feelings about other features. Reputation can be personal as well: the fact that they know it because they have used the site before (experience) is also included[2].
Reputation of a company and its site are not found on the site itself but on others indexing it. The features in this category are all external sources of information that give an impression on how well-known it is and what people think of it.
Examples: Number of references to this site (reversed search in AltaVista) and trust of partners (trust rating for outward pointing links found on the site).

Professionalism one seeks in both brick-and-mortar companies and online ones. It is typically a vague subject people refer to. Even the Cheskin report talks about "Clarity of purpose" and "Craftsmanship" (ibid. page 11).
Examples: Presence of 3^{rd}-party seals, number of pictures on the site and presence of a privacy policy.

Technical issues are noticed when they do not work, or when one wants the site to conform to a certain technical standard. This can be true for more categories, but here it is really obvious. The site is not working, not very fast etc. Or one really wants the site to be secure before ordering anything on it. Examples: Number of dead links found, use of encryption for site (indicated by the browser) and speed of the site (average roundtrip-time of packets).

Other features mentioned by people were hard to classify (i.e. implement) and ended up in a separate category. The modeling of them will take more research.

Using this model, sites can be represented as vectors: a vector of feature ratings. This will be called the *feature vector* of a site.

4 Trust agents

The Feature Model for trust is used to define the *trust agent.* This agent has 2 basic functions. It can

1. give a trust rating to a site based on what it perceives as trusted features;
2. communicate this rating to other agents to come up with a *negotiated* trust rating.

Both of these functions will be discussed below.

[2] As opposed to the survey category, where experience was not included.

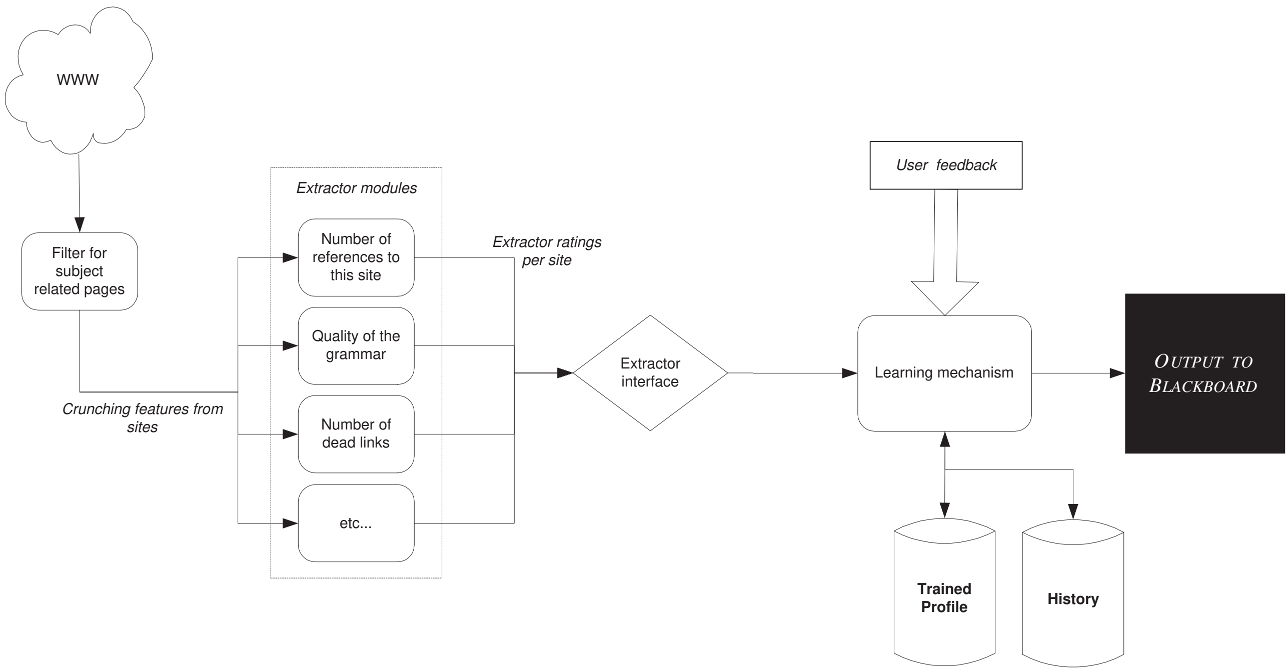

Figure 1: Architecture of a trust agent

4.1 A trust agent

The agent that was designed to rate sites is connected to a search engine that selects subject related websites. This is depicted in Figure 1. The list of websites that remain can now be labeled.
For every page the agent will extract features. This is done by extractors. Given the URL of the page, an extractor can calculate one feature rating (there is one extractor for every feature). For the 'quality of grammar' feature, for example, it will take all plain, visible text from the page and run it through a spell checker. The number of mistakes the checker finds (relative to the number of words) will be returned by this extractor. Some extractors use external sources. The 'number of references' for example, will go to AltaVista and do a reverse search for the website. This search will return the number of pages that point *towards* this site. This score will be returned as its rating. All these extractor ratings together make up the feature vector of the site.
The extractor interface is there to make the extraction process transparent to the rest of the application. This is done because in this way extractors can be added and removed in a modular way.
As stated before, trust is a subjective matter and the agents, therefore have to be user specific. The measuring of the feature vector was still done in an objective way, but the way the agent 'sees' these numbers is different for every user. In other words, the agent has to know which features contribute to the user's sense of trust and how much. It has to know the user's trust behaviour.

Calculating trust

Before the agent will make any recommendations it will show the user some selected sites and ask his trust rating for them. The assumption is, that in this way it will learn how the feature vector influences the final rating. For this learning process 3 methods have been implemented and tested:

- Feed-forward neural network; will take the feature vector as an input and train on the feedback given by the user.
- k-Nearest Neighbour; plots all trained sites in a space and will decide on the Euclidean k nearest ones.
- Decision Tree; takes every extractor as a leaf and calculates the order in which they are traveled down by looking at the most distinguishing extractors (algorithm from Russel & Norvig [6, page 534]). When, for example, the presence 3^{rd}-party seals always leads to a 'Absolute' trusted site, this will be a very distinguishing feature placed high in the tree.

When the feature vector is presented to one of these mechanisms, the agent will first look in the history file. If the URL is already present here, the user has already rated it and his own answer will be returned. If this is not the case, the mechanism will have to infer the answer from the trained user profile, which was generated by the first selected sites. If the user does not agree with the

answer, he can supply the mechanism with feedback in order for it to learn more.
Together with this answer comes a confidence level. This level indicates how 'sure' the agent was given its inference mechanism and data. For the k-NN, for example, the confidence is partly based on the diversity in the 'neighbourhood' out of which the answer was generated. When $k = 5$, there are 5 sites close to the tested one that have ratings. These ratings can be very diverse, but one answer has to be chosen from them. The diversity is then defined as the number of answers in this set that were different from the one inferred. Confidence levels are expressed as percentages. Once the mechanism has inferred an answer and calculated its confidence, it will be written to the blackboard for other agents to see.

4.2 Negotiating on trust

Based on eOpinions' idea of the Web of trust, the user can select more agents to help him decide on trust. When looking for second-hand cars for example, he may want to ask a friend to take a look at the site or maybe even a consumer magazine. With WebTrust, these two parties can have their own agents. In this way, the user will profit from the trust knowledge put into the consumer magazine's agent together with his own to select sites that are trusted by both of them. The user's agent for example focuses on quality of grammar, whereas the consumer magazine's focuses on well-known cardealer names.
All agents selected by the user will write their ratings to the blackboard together with their confidence level. This data forms the input to the overall measurement, which is depicted in Figure 2.
This process models a basic form of human negotiation: when one enters a room with experts that all have an opinion one may ask them all for their answer and their confidence. The decision maker has to generate one single answer from all these experts.
In the current implementation all agents have equal weights, but for the negotiation to be more realistic, it is important to make these different. These weights can be used to turn the anonymous agents into agents that have a reputation for their field of expertise. Some agents will have a very low confidence level (below 30%), they are discarded. After this, the system looks for agents that are very confident (above 95%). This level is also equal to the agents that have found their rating in their history. Since this answer was supplied by the user himself, they are very confident. If confident agents are found then all the others are discarded. The weighted average (to confidence level) of the remaining agents will decide the final answer of the system.

5 Test results

The trust ratings yielded from the survey were used to test the quality of the agent mechanisms and the negotiation process. From the 50 sites rated by respondents a trainset was taken and the others had to be predicted. At the time of testing only 6 extractors had been implemented due to time constraints. This means that the predictions are not based on the full set of features that a respondent had taken into account when assessing his trust rating.
From the quality of this prediction a measurement could be made: taken over the average of all the predictions that had to be made, the quality measurement was defined as the Average error * (1 - Precision[3]). For a trainset of 40 sites and 10 predictions, the system's scores are displayed in Table 2. Presented in this table are the numbers for the 3 primitive methods and their 4 possible combinations. The lowest score is the best agent in this metric, the k-NN therefore scores best. The ANN combined with k-NN scores very well too. This is because the ANN most of the time, is not very confident and will not play an important role in negotiation, i.e. most of the answers are still determined by the k-NN method.

It can also be seen from this table that the combined methods do not score significantly better than the primitive ones. This is partly due to the calculation of the confidence level, which turned out not to work very well. The decision tree, for example, was always very confident, while its answers were very bad. All combinations with a decision tree have therefore been blurred by this.
Tests have also been done to see if the methods increase the quality of their predictions when the trainset was increased. Both the neural network and the k-NN managed to increase (lower quality measure), whereas the decision tree decreased when training more. For k-NN the average quality increase was 17% when doubling the trainset.
Finally benchmarks were setup to see if the agents actually scored better than (smart) random generators. We made an agent that gave scores uniformly and one that scored according to a normal distribution (which is a lot like the respondents of the survey did). They scored a quality measurement of 1.17 and 0.94 respectively. Even the worst method implemented, (combination of ANN and decision tree, 0.76) scored better than these benchmarks. Further improvements can be related to these scores.

6 Future Work

The quality measurements show that agents score better than the benchmark. However, they do need more extractors to have 'full vision'. The modular approach of the application facilitates adding extractors. This will therefore be relatively easy to do.
A more hybrid approach in which the extractors are measured by different methods may further improve results.

[3] relative number of correct answers

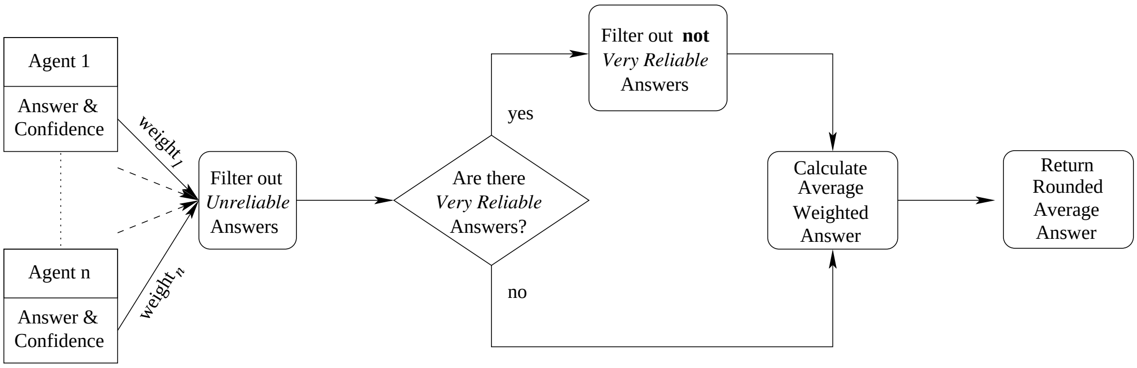

Figure 2: The process of negotiating agents

Table 2: Quality measurement for prediction of 10 sites

ANN	*k-NN*	*Tree*	*A & k*	*A & T*	*k & T*	*All*
0.70	0.51	0.74	0.52	0.76	0.68	0.57

This means that methods specialise in certain features on a site. Neural networks, for example, are known to work well in visual recognition whereas the decision tree may work better when rule-based features are taken into account.
Furthermore, the negotiation process needs revision. Especially the decision tree needs a a confidence level measurement that better fits the actual quality of its rating. Other ways of inferring the answer can also be tried. The k-NN method can, for example, be 'upgraded' towards a full Case-based reasoning system with more knowledge.

7 Conclusions

We found strong support for our Feature Model describing properties of sites that influence trust behaviour. It presents a full listing of features that appear on, or are related to websites. Using this Feature Model, it proved possible to automate the labeling of commercial websites with respect to trust. The agents that were implemented, automated the way humans assess trustworthiness of a website by learning from their behaviour. k-Nearest Neighbour is the best method of the three tested to learn from human trust behaviour. Even combinations of these methods (by negotiation) did not perform better.

Acknowledgments

This work has been done as a Master thesis project for Delft University of Technology, the Netherlands in cooperation with Accenture Technology Labs, France. We would like to thank everyone who participated in the course of this project.

References

[1] Cheskin. eCommerce trust study. Technical report, Cheskin Research, Jan. 1999.

[2] Cheskin. Trust in the wired americas. Technical report, Cheskin Research, July 2000.

[3] F. Egger. Human factors in electronic commerce: Making systems appealing, usable & trustworthy. In *Graduate Students Consortium & Educational Symposium, 12th Bled International E-Commerce Conference*, June 1999.

[4] eOpinions.
http://www.eopinions.com.

[5] S.P. Marsh. *Formalising Trust as a Computational Concept.* PhD thesis, University of Sterling, UK, Apr. 1994.

[6] S. Russel and P. Norvig. *Artificial Intelligence, a modern approach.* Prentice Hall, 1995.