

# HETWIN: Helping Evaluate the Trustworthiness of Web Information for Web Users Framework using Semantic Web Technologies

Jarutas Pattanaphanchai, Kieron O'Hara, and Wendy Hall

Electronics and Computer Science, Faculty of Physical and Applied Science,  
University of Southampton,  
Southampton, SO17 1BJ, United Kingdom  
[{jp11g09, kmo, wh}@ecs.soton.ac.uk](mailto:{jp11g09, kmo, wh}@ecs.soton.ac.uk)

**Abstract.** Assessing the trustworthiness of information found on the Web is challenging because of two factors. First, there is a little control over publishing quality. Second, Web users have little information available on which to base judgment of the trustworthiness of Web information while they are interacting with it. This work addresses this problem by collecting and presenting metadata about authority, currency, accuracy and relevance to evaluate the trustworthiness of Web information during information seeking processes. In this poster, we propose the *HETWIN* application framework and present a design as a prototype tool that employs this framework for academic publications.

**Keywords:** Trust, Credibility, Information Quality, Semantic Web

## 1 Introduction

It is known that ordinary Web users base decisions on whether to trust information on the Web on heuristic factors (pertaining to its presentation and layout). However, as these heuristic factors are mainly based on surface level characteristics of the Web page, and such characteristics are easily disguised, Web users can arrive in the wrong conclusions about the trustworthiness of information they consume [3]. However, a number of studies have suggested that additional information such as the identity of the author (e.g. name, position), and the expertise of the author could potentially increase the Web users' confidence and help them to make better assessments than by using their own heuristic criteria alone [2, 4–6]. In particular, Bizer *et.al.* [1] proposed the TriQL.P browser, a RDF browser, that presents recommended RDF datasets that should be trusted based on trust policies. However, in their work, the user needs to go to a certain Web page, from which the browser can extract Semantic Web content. On the contrary, it is more useful to provide Web users with a tool with which they can look for the information they need while it *automatically* gathers the supportive information to help them evaluate the trustworthiness of Web information.

In order to address this problem, we propose a framework to help users evaluate the trustworthiness of Web information, called *HETWIN*, which, with

information of the factors from the studies above, are selected to use in the framework. In addition, we propose a prototype tool, which employs the HETWIN framework implemented as a chrome extension. The prototype collects metadata using Semantic Web technologies and presents it in a useful way in the context of the users' search for information. In the following section, we explain the HETWIN architecture and display an example result from our prototype. Then, we present planned future work.

## 2 Helping Evaluate the Trustworthiness of Web Information for Web Users Framework

Our framework uses Semantic Web technologies to collect RDF data which is published alongside Websites and queried from SPARQL endpoints, which it then integrates to build metadata graphs. Then, these metadata graphs are used to create supportive data, which is presented to users in order to help them evaluate the trustworthiness of Web information. In this work, we assume that the RDF data on the Web or in the data store is accurate. Evaluation is based on a case study of the *ePrints* of the University of Southampton<sup>1</sup>, which is an online repository of academic publications, in which the accuracy of the RDF data published is verified by authorized staff<sup>2</sup>. Our application framework, as shown in Figure 1, consists of three main modules.

- **The input module** accepts the user's search keywords and the domain of interest, which affects the type of information returned by the search. In this work, we defined four domains of interest (business, informational, news and personal). The input module extracts any RDF linked to from the web page. Also, our model evaluates the trustworthiness of the information every time the user interacts with the system. Therefore, the system obtains the most recent information at the time at which the evaluation is performed.
- **The trustworthiness criteria and metacollection module** is composed of two main components. The trustworthiness criteria comprises of four basic criteria: *authority*, *currency*, *accuracy*, and *relevance*, which the assessment of trustworthiness in each domain of interest is based. Each criterion provides the basic predicate keys of RDF that should be used to collect metadata. For instance, in the informational domain, trustworthiness is evaluated based on the authority criterion, using predicate key, “*dct:creator*”, the currency criterion, using predicate key, “*dct:date*”, the accuracy criterion which is based on the predicate key, “*bibo:status*”, and the relevance criterion which is based on data returned from querying using the predicate key, “*dct:title*” and “*dct:abstract*”. Alternatively, in the news domain, which still evaluates the trustworthiness of the information based on the same criteria, different or additional predicate keys might be used. For example, the authority criterion might use the predicate key “*dct:publisher*” in addition to the “*dct:creator*”.

---

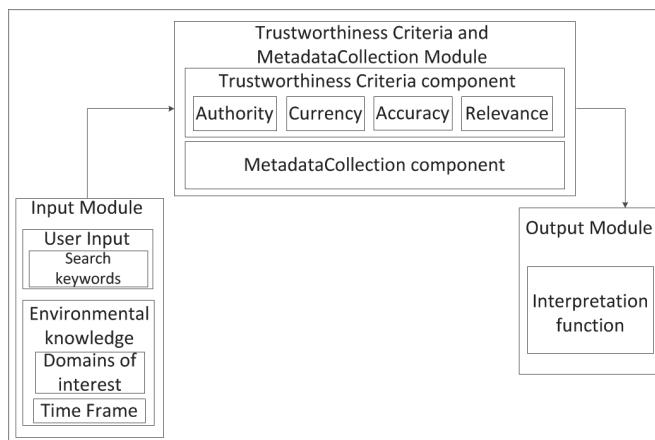
<sup>1</sup> <http://www.eprints.soton.ac.uk>

<sup>2</sup> <http://www.southampton.ac.uk/library/research/eprints/policies/eprints.html>

Our framework allows one to add additional predicate keys or new domains by adding them into its configuration file. Therefore, the framework can adjust for use in different domains and can extend to new domains.

The metacollection component gathers metadata based on the predicates which are defined in the trustworthiness criteria. The collected metadata will be aggregated in order to build metadata graph. The basic approach of aggregating metadata assumes that the metadata from the four basic predicates have the same level of important for assessing the trustworthiness of Web information. In the case that the system needs the additional data, the system will add the additional data into the metadata graph after the basic metadata has been added.

- **The output module** displays the metadata graph in a human readable format to help the users assess the trustworthiness of Web information. In addition, it orders the results based on the relevance of the information to the user's query which is computed based on the frequent of the appearance of search terms in the title and the abstract and the expertise of the authors or creators of the information.



**Fig. 1.** HETWIN architecture

### 3 Results

We present example results of the output of our tool, using data from ePrints at University of Southampton. Specifically, we consider the publications of the School of Electronics and Computer Science, and we focus our evaluation on the informational domain. The result in Figure 2 displays the identifying details of the publication including its title, its abstract and the name of its authors.

## HETWIN: Helping Evaluate the Trustworthiness of Web Information for Web Users

Moreover, the results are ordered by the relevance of the information to the user's interests. Specifically, it shows the authors' full names and also their appellations. These can indicate the authority of the author, which represents their reputation in producing this content. In addition, it displays the detail of each author's publications, which itself is indicative of the author's expertise in the area. For example, if the author has several publications that relate to privacy, this implies that the author is not only interested in that area but also has expertise in it. In addition, the more publications that exist by that author that relate to the keywords, the more likely the author is to be an expert in that topic. The publication date indicates the currency of the publication. The system gathers the status of the publication for determining the accuracy of the information within it because, in the case of publications, there is a review process, which can help to evaluate the accuracy of the content. For example, if the publication has been peer-reviewed or published in an academic publication, it is likely to be accurate, and therefore trustworthy. In current work, the relevance criterion analyses the abstract and the title of the publication. If they contain the user's search keywords, the document is more likely to meet the user's needs. However, we consider another potential approach to evaluate relevance more efficiently than matching exact keywords; adopting an ontology concept for finding terms related to the user's keywords to match in key areas of the content such as the title or the first paragraph of content. This allows the framework match the relevant information to user's needs better.

**Retrieved result for search keywords:**

**privacy**

**A (multi'domain'sional) Scrutable User Modelling Infrastructure for Enriching Lifelong User Modelling**

**Publication abstract:**  
A Scrutable User Modelling Infrastructure (SUMI) is proposed in this paper, which focuses on three key Lifelong User Modelling (LUM) requirements: interoperability, scrutability [more]

**The number of times the keyword, "privacy", appears in the abstract:**  
The author of publication:  
Demetris Kyriacou  
[more information]  
Professor Hugh Davis  
[more information]  
Dr Thanassis Tiropoulis  
[more information]

**Year of publication:** 2009  
**The type of publication:** ConferenceItem|Print, EPint, Article, AcademicArticle  
**The status of publication:** peerReviewed, published

**Making Public Media Personal: Nostalgia and Reminiscence in the Office**

**Publication abstract:**  
In this paper we explore the notion of creating personally evocative collective content from publicly available material. Compared to the personal media that look at, [more]

**The number of times the keyword, "privacy", appears in the abstract:**  
The author of publication:  
Paul Andre  
[more information]  
Abigail Sellen  
[more information]  
dr monica m c scherfel  
[more information]  
Ken Wood  
[more information]

**Year of publication:** 2011  
**The type of publication:** ConferenceItem|EPint, EPint, Article, AcademicArticle  
**The status of publication:** peerReviewed, published

**A (multi'domain'sional) Scrutable User Modelling Infrastructure for Enriching Lifelong User Modelling**

**Relevance Criterion**

[A Scrutable User Modelling Infrastructure (SUMI) is proposed in this paper, which focuses on three key Lifelong User Modelling (LUM) requirements: interoperability, scrutability [more]]

**The number of times the keyword, "privacy", appears in the abstract:** 3

**The author of publication:**  
Demetris Kyriacou

**Authority Criterion**

[Professor Hugh Davis  
Southampton ECS Role: Professor Hugh C. Davis as Academic staff in Web and Internet Science Hugh Davis has 216 publications  
13 of these publications relate to the keywords, as follows:  
① has relevant keywords in the title and the abstract.  
② have relevant keywords in the abstract.  
③ less]

**Currency Criterion**

[Dr Thanassis Tiropoulis  
[more information]]

**Accuracy Criterion**

[Year of publication: 2009  
The type of publication: ConferenceItem|EPint, EPint, Article, AcademicArticle  
The status of publication: peerReviewed, published]

**Fig. 2.** Example of results using “privacy” as keyword and “informational” as a domain (the four criteria are shown in bold)

#### 4 Conclusions

We proposed an application framework and prototype tool which helps users to evaluate the trustworthiness of Web information using Semantic Web technologies. The result from the prototype shows the supportive data in each criteria, which is explained to the user and can help them assess the trustworthiness of Web information. In future work, we will evaluate our framework by conducting a user survey. This survey will elicit information about how satisfied the users were with the system and how their approach to assessing trust has changed since using our system in comparison to an expert.

#### References

1. Bizer, C., Cyganiak, R., Gauss, T., Maresch, O.: The TriQL. P browser: Filtering information using context-, content-and rating-based trust policies. In: Proc. of the Semantic Web and Policy Workshop. vol. 7, pp. 12–20 (2005)
2. Fogg, B., Marshall, J., Laraki, O., Osipovich, A.: What makes Web sites credible?: a report on a large quantitative study. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems pp. 61–68 (2001)
3. Fogg, B., Soohoo, C., Danielson, D.R., Marable, L., Standford, J., Tauber, E.R.: How do users evaluate the credibility of Web sites?: a study with over 2,500 participants. Proc. of the 2003 conference on Designing for user experiences pp. 1–15 (2003)
4. Rieh, S., Belkin, N.: Understanding judgment of information quality and cognitive authority in the WWW. In: the 61th annual meeting of the Am. Soc. for Inf. Sci. and Technol. vol. 35, pp. 279–89 (1998)
5. Tate, M.: Web Wisdom: How To Evaluate and Create Information Quality on the Web, Second Edition. CRC Press (2009)
6. Wathen, C.N., Burkell, J.: Believe it or not: Factors influencing credibility on the Web. Journal of the Am. Soc. for Inf. Sci. and Technol. 53(2), 134–144 (2002)