AUTOMATIC EVALUATION OF INFORMATION CREDIBILITY IN SEMANTIC WEB AND KNOWLEDGE GRID

Adam L. Kaczmarek
Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology
ul. Gabriela Narutowicza 11/12 80-952 Gdansk, Poland

Keywords: Credibility, Trustworthiness, Semantic Web, Knowledge Grid.

Abstract: This article presents a novel algorithm for automatic estimation of information credibility. It concerns information collected in Knowledge Grid and Semantic Web. Possibilities to evaluate the credibility of information in such structures are much greater than those available for WWW sites which use natural language. The rating system presented in this paper estimates credibility automatically on the basis of the following metrics: information commonality, source independence, prestige of the source, experience with the source and conclusions from related information.

1 INTRODUCTION

The Internet provides great amount of information, however this information is not always true. Some methods to rate the credibility of web pages have been designed. For example Google News ranks items according to the reliability of the news source (Google News Patent Application, 2003). Nagura took into account commonality, numerical agreement and objectivity of web news (Nagura, 2006). In estimating credibility Fogg focused on a design of a web page and its layout (Fogg, 2003). A large list of approaches to estimating credibility by different authors was presented by Abdula (Abdula, 2002). Some methods concerning estimation of credibility in news and articles are based on surveys. The aspect of data expiration was included by Breners-Lee (Breners-Lee, 1998). Contradictions can also be caused by the fact that some information applies in a different context (Palmer, 2001).

However all these methods focus on rating the credibility of an entire web site or an article, but not of a particular piece of information. Rating certain information or a sentence on a web page is more problematic as this information is presented in natural language in a form easily readable for human beings. The ability to process such data by computer systems is limited. However, in Semantic Web (Palmer, 2001) and in Knowledge Grid (Zhuge, 2004) information is presented in a form designed for computer systems. It opens new perspectives in verifying credibility of information.

Systems based on Semantic Web or Knowledge Grid usually assume that information provided by the Internet is true. However, because in WWW some web pages provide wrong information, it can be expected that also Semantic Web and Knowledge Grid will contain false data. It will lead to contradictions. The occurrence of contradictory data causes that in terms of classical Boolean logic all existing data should become useless. It is because of *ex falso quodlibet* principle. The consequence of contradiction is that every sentence can be proved to be true. There are two possibilities to overcome this situation: using Non-Boolean, paraconsistant reasoning methods (Schaffert, 2005) or assuming that some of the data is wrong. This article focuses on the second approach and presents a novel algorithm to verify the credibility of information automatically.

2 METRICS

The rating system presented in this paper takes advantage of new possibilities enabled by storing information in Knowledge Grid and Semantic Web. It applies metrics which so far were used in manual estimations of credibility in order to rate credibility in these structures automatically. The rating system takes into account the following metrics:

L. Kaczmarek A. (2008).
AUTOMATIC EVALUATION OF INFORMATION CREDIBILITY IN SEMANTIC WEB AND KNOWLEDGE GRID.
In *Proceedings of the Fourth International Conference on Web Information Systems and Technologies*, pages 275-278
DOI: 10.5220/0001516002750278

information commonality, source independence, prestige of the source, experience with the source and conclusions from related information.

It is necessary to distinguish here between source information and the information which the rating system believes that is true. In classical logic if there is a source claiming some sentence p then this p is assumed. However, if there are many sources and it is possible that some of them are lying, there are source information and accepted information concerning the same subject

2.1 Commonality

First, let us assume that the only aspect taken into account is the commonality. The commonality indicates that the more sources claim that a certain sentence is true, the greater is the probability that it is true indeed. Let us introduce a new notation here. For each source sentence p_i, a sign p_i with two bars over it will be equal *1*, (equation 1).

$$\underset{p_i}{\forall} \overline{\overline{p_i}} = 1 \tag{1}$$

p_i – source sentence

If there is a sentence p and source sentences p_i which are equivalent to p or $\neg p$, the rating system makes a decision on the basis of all p_i. It is based on the result of function C defined by the equation 2.

$$C(p_1,\ldots,p_n) = \sum_{i=1\ldots n} \overline{\overline{p_i}} \tag{2}$$

n – number of source sentences
C – function which estimates credibility
$p_1,\ldots,p_n$ – sentences such that $p_1 = p_2 = \ldots = p_n = p$

Decisions made on the basis of function C are expressed by the equation 3.

$$p^* = \begin{cases} p & if\ C(p_1,\ldots,p_n) - C(p_{n+1},\ldots,p_m) > 0 \\ \neg p & if\ C(p_1,\ldots,p_n) - C(p_{n+1},\ldots,p_m) < 0 \\ \varnothing & if\ C(p_1,\ldots,p_n) - C(p_{n+1},\ldots,p_m) = 0 \end{cases} \tag{3}$$

$p_1,\ldots,p_n$ – sentences such that $p_1 = p_2 = \ldots = p_n = p$
$p_{n+1},\ldots,p_m$ – sentences such that $p_{n+1} = \ldots = p_m = \neg p$
p – sentence under verification
p^* – sentence which is assumed as true by the rating system

If there is the same number of sources claiming that there is p and ¬p, the value p* remains empty. It is unknown whether p or ¬p is true.

2.2 Independence

Sources can base their knowledge on information published by other sources. Information provided by two independent sources is more credible than the information provided by two sources which depend on one another. It is often unknown from where the source derives information and whether it was copied from another source or the source prepared it by itself. Such notifications are neither present in Semantic Web nor in Knowledge Grid, although they should be included there. Similarly, as there are bibliographies at the end of scientific articles, there should be a source of information given in these structures. This would made truth verification easier and it would help to reveal mistakes in the future.

What is more credible: two independent sources or many sources which depend on a given one? In this case, the two independent sources would be assumed as more credible. It can be expressed by inequity (4). Finite number of dependent sources is less credible than two independent sources. Function C here is based on function (2), however it includes the aspect of dependences.

$$C(p_{0,1}, p_{0,2}) > C(p_{1,1}, p_{1,2},\ldots, p_{1,n}) \tag{4}$$

$p_{0,x}$ – independent source sentences
$p_{1,x}$ – dependent source sentences

Let us now consider what is more credible: one independent source or two sources, when one of these two sources depends on the other? It could be assumed that these cases are equally credible, however more accurate would be an assumption that those two sources are more credible than only one, even if they depend on each other. It is also assumed that $n+1$ dependent sources are more credible then n dependent sources. It can be expressed by the inequity (5).

$$C(p_{1,1},\ldots, p_{1,n}, p_{1,n+1}) > C(p_{1,1},\ldots, p_{1,n}) \tag{5}$$

$p_{1,x}$ – dependent source sentences

There are many ways of solving inequities (4) and (5). In order not to complicate the C function, the rating system uses a linear form. There needs to be a parameter which will decrease the significance of the dependent sources. The C function for dependent sources can have a form expressed by the equation (6).

$$C(p_{k,1},\ldots, p_{k,n}) = \sum_{i=1\ldots n} \alpha \overline{\overline{p_{k,i}}} \tag{6}$$

α –parameter which fulfills equations (4) and (5)

$p_{k,x}$ – dependent source sentences
k – index of a set of dependent sources, k≠0

The α is a decreasing parameter. As long as it is positive, the inequity (5) is fulfilled. To fulfill the equity (4) the parameter α can not be constant. It needs to decrease for each next dependent source. This is achieved when $\alpha=1/\beta^i$ and $\beta>1$. Inequity (4) can be solved as expressed by equations (7).

$$\begin{cases} C(p_{0,1},p_{0,2}) = \overline{\overline{p_{0,1}}} + \overline{\overline{p_{0,2}}} = 2 \\ C(p_{k,1},\dots,p_{k,n}) = \sum_{i=1\dots n} \alpha \overline{\overline{p_{k,i}}} = \sum_{i=1\dots n} \frac{1}{\beta^i} \end{cases} \quad (7)$$

β – component of the parameter α

Conclusion (8) based on (4) and (7) shows that *β* needs to be not less then 2.

$$\sum_{i=1\dots n} \frac{1}{\beta^i} < 2 \Rightarrow \sum_{i=1\dots n} \frac{1}{\beta^i} < \lim_{n\to\infty} \sum_{i=1\dots n} \frac{1}{2^i} \quad (8)$$

The parameter *β* can be assigned to a boundary value equal *2*. The *C* function for dependent sources adopts the form expressed by the equation (9).

$$C(p_{k,1},\dots,p_{k,n}) = \sum_{i=1\dots n} \frac{1}{2^i} \overline{\overline{p_{k,i}}} \quad (9)$$

2.3 Prestige

The prestige of a source is based on its authority. For example, information provided by universities would be of higher credibility than this delivered by unknown sites. A list of prestigious institutions can be prepared. Preparing it needs human interference and this list needs to be updated, however it can be used by the algorithm automatically. This list would be the main mechanism in the algorithm against intentional attempts of spreading false data. It is not likely that sources which are included in this list would participate in intentional propagation of false information. However, on the Internet it is possible to create a large number of not prestigious sources which would publish false data. The method to overcome this problem is to treat all sources without prestige as dependent ones. Also information from prestigious sources belongs to the same group of dependence as not prestigious ones when a prestigious source bases its knowledge on the source which is not prestigious. The aspect of prestige can be expressed by equations (10).

$$\begin{aligned} &G_x = \left(p_{x,1}, p_{x,2}, \dots, p_{x,a_x}\right) \\ &C(p_{0,1},\dots,p_{0,n},G_1,\dots,G_m) = \sum_{i=1\dots n} \overline{\overline{p_i}} + \sum_{i=1\dots m} C(G_i) \\ &C(G_x) = \sum_{i=1\dots a_x} \frac{1}{2^i} \overline{\overline{p_i}} \end{aligned} \quad (10)$$

G_x – set of dependent information
a_x – number of sentences in set x.
m – number of sets of dependent information

Groups of sentences G_x are those containing information from not prestigious sources and information from prestigious ones which depends on those not prestigious.

2.4 Experience

The experience is based on a history of cooperation with a source. It can be negative if in the past it turned out that a source provided false information. Positive experience with sources is not as important as negative one. It can be expected that most sources will provide truthful information. However, when some source publishes false data it is more likely that such an incident will happen again.

What is more credible: one source which always published truth or two independent sources which once lied? In the rating system it is assumed that these cases are of equal credibility. Each false piece of information provided by the certain source reduces the estimated credibility of this source by a half. Reduced credibility for source which lied is not perpetual. The rating system assumes that it expires after one year.

In equations (10) the order of source information within each group was unimportant. However, when some source has reduced credibility it becomes significant. The *C* function for groups of related sources where some of them have spoiled opinion has a form expressed by the equation (11)

$$\begin{aligned} &G_x = \left(p_{x,1}, p_{x,2}, \dots, p_{x,b_x}, p_{x,b_x+1}, \dots, p_{x,a_x}\right) \\ &C(G_x) = \sum_{i=1\dots a_x} \frac{1}{2^i} \frac{1}{2^{l_i}} \overline{\overline{p_{x,i}}} \end{aligned} \quad (11)$$

b_x – number of sources with reduced credibility
l_i – number of falsenesses found recently for the source which publishes information p_i
$p_{x,1},\dots,p_{x,bx}$ – sentences with reduced credibility arranged from the most to the least credible one.

If there is no negative experience with the source, then the l_i value is equal *0*. Sentences from sources with solid opinion are ordered such that the more spoiled opinion a source has, the smaller is the parameter *1/2^i^* which multiplies its influence. For sources without dependencies, this order is not necessary. The influence is only multiplied by the *1/2^i^* value as expressed by the equation (12).

$$C(p_{0,1},...,p_{0,n},G_1,...,G_m) = \\ = \sum_{i=1...n} \frac{1}{2^{l_i}} \overline{\overline{p_{0,i}}} + \sum_{i=1...m} C(G_i) \qquad (12)$$

2.5 Conclusions from Known Sentences

Apart from source information, the rating system also estimates the credibility of information on the basis of this which it already accepted as true. For example, when the rating system is verifying whether a sentence *p* is true or not and it already accepted as true sentences *r* and *t* such that $r \wedge t \rightarrow p$, this fact is included in evaluation of credibility. The question is how important such known sentences should be in truth verification. In the example given above $r \wedge t \rightarrow p$ will be treated by the rating system in the same way as the independent source claiming *p*. The value added to the *C* function will be equal *1*. When apart from sentences *r* and *t* there will be sentences *g* and *h* such as $r \wedge g \wedge h \rightarrow p$ then these sentences *g* and *h* will add ½ to the value of *C* function. It is because the sentence *r* has already been taken into account. The addition of aspect of known sentences to the function *C* is expressed by equations (13).

$$C(p_1,...,p_n,G_1,...,G_m,R) = \\ = \sum_{i=1...n} \frac{1}{2^{l_i}} \overline{\overline{p_i}} + \sum_{i=1...m} C(G_i) + C(R) \\ R = (r_1, r_2, ..., r_k) \qquad (13) \\ C(R) = \sum_{i=1...k} C(r_i) \\ C(r_j, ..., r_{j+v_j}) = 1 \text{ if } r_j \wedge ... \wedge r_{j+v_j} \rightarrow p$$

R – set of known sentences
$r_1,...,r_k$ *– known sentences*
k – number of known sentences
$r_j,..,r_{j+vj}$*–any known sentences which implicates p*

Function *C(R)* expressed by equations (13) can be calculated in various ways. In the rating system this function is not ambiguous. It first assigns values for relations with the smallest number of known sentences and then, iteratively, for those relations where the number of known sentences is greater.

3 CONCLUSIONS

Current researches on the computer science tend to focus rather on trust in terms of security and on reliability in the meaning of the systems' stability, than on credibility. Nevertheless, even if requirements of security and reliability are fulfilled, the matter of providing credible information still remains unresolved. Structures such as Knowledge Grid and Semantic Web where information is stored in a formalized manner create much more possibilities in detecting falseness automatically. It was used in the algorithm presented in this paper.

Truth verification is more effective when information copied from the other source is published with reference to the original source. Such notes can prevent from propagation of false data. Including data about the original source of information should be a common practice in Knowledge Grid and Semantics Web.

REFERENCES

Abdulla R. A., Garrison B., Salwen M., Driscoll P., Casey D., 2002. *The Credibility of Newspapers, Television News, and Online News,* In Proc. of the Association for Education in Journalism and Mass Communication Annual Convention, Miami Beach, FL

Berners-Lee T., 1998. *Semantic Web: Inconsistent data,* http://www.w3.org/DesignIssues/Inconsistent.html

Fogg B.J., 2003. *Persuasive Technology: Using Computers to Change What We Think and Do*, Morgan Kaufmann Publishers

Google News, 2003. *Google News Patent Aplication*, http://www.webpronews.com/topnews/2005/05/03/google-news-patent-application-full-text

Nagura R., Seki Y., Kando N., Aono M., 2006 *A Method of Rating the Credibility of News Documents on the Web,* In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval

Palmer S. B., 2001. *The Semantic Web: An Introduction*, http://infomesh.net/2001/swintro/

Schaffert S., Bry F., Besnard P., Decker H., Decker S., Enguix C., Herzig A., 2005. *Position Paper: Paraconsistent Reasoning for the Semantic Web,* In Proceedings of Workshop Uncertainty Reasoning for the Semantic Web, Galway, Ireland

Zhuge H., 2004. *The Knowledge Grid*, World Scientific Publishing Company