Web Services Integration with Regard to the Metrics of Data Believability

Adam L. Kaczmarek
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
ul. G. Narutowicza 11/12, 80-233 Gdansk, Poland
e-mail: adam.l.kaczmarek@eti.pg.gda.pl

***Abstract*—The paper is concerned with estimating the believability of data acquired from web services. In the paper, a new method for believability estimation is introduced. The method is designed for integrating web services. The believability estimation is based on the following metrics: quantity, reputation, approval, independence, traceability, maturity, authority and objectivity. In the method, data trustworthiness is determined by the credibility of the data source. In the believability estimation, information about data provenance is used. Moreover, the method is based on the consideration that it is possible to increase the data believability not only by finding more believable sources of data, but also by acquiring the same kind of data from many different sources and analyzing this data. It is possible in the field of web services because there are many vendors of the same kind of services.**

Keywords-data believability; web services; data processing

I. INTRODUCTION

When new data is acquired, the problem with data credibility occurs. Human beings practice various ways to estimate the trustworthiness of information and sources of information. There is a growing need to develop techniques, which can enable computer applications to automatically estimate the believability of information. These techniques should make it possible to identify and exclude unlikely information. It is particularly important in the case of applications collecting data from external sources through the Internet. These kinds of applications embrace software based on web services architecture [1].

This paper is concerned with data believability in web services. Applications using web services in order to achieve their goals receive data from services provided by vendors, which may not be reliable. It is particularly important for applications, in which integration of different kinds of web services is performed. If data collected by an application is wrong, the application will not produce proper results. Thus, it is necessary to verify the believability of data.

Data believability was defined by Wang and Strong as "the extent to which data are accepted or regarded as true, real and credible" [2]. The notion of believability is also referred to by the terms credibility, trustworthiness and plausibility. It needs to be stressed that data believability differs from data security. Security is related to problems of authentication, authorization and access to data. Data corruption may be caused by an undesirable influence of people or malicious programs as a result of poor data security. However, even if security is ensured and data is safely delivered from the source it is supposed to come from, problems with the truthfulness of this data may still exist: the source may spread incorrect information. Data believability is thus an issue, which goes beyond data security and it occurs even if problems with security are resolved.

The paper consists of five sections. The section following the introduction contains an overview of related work concerning data quality, believability and provenance. The third section presents the method for data believability estimation designed by the author of this paper. The fourth one contains an evaluation of the method. The last section refers to conclusion and future work.

II. DATA BELIEVABILITY

Estimation of data believability requires regarding of data not only by meaning, but also in the context of its provenance. Data provenance is an integral feature of data. On the basis of who created data, how it was stored, and how it was processed, conclusions can be drawn about data believability. Recently, there has been a significant amount of research on acquiring and storing information about data provenance in web services. Tsai et al. presented a profound description of requirements and solutions concerning data provenance problems [3]. Techniques for solving these problems include the use of metadata, databases and new types of protocols. Moreover, an in-depth description of data provenance problems was presented by Moreau [4]. Storing information about data provenance makes possible to use this information in order to estimate data believability.

Information about data provenance indicates a web service, which is the source of that data. The quality of this web service can be taken into account in estimating data believability. Some web services' parameters are objective and they can be determined on the basis of statistics about the behavior of web services. Such parameters include web service availability, fees and latency in data transmission [5]. The quality of a web service can be also evaluated by its users in the same way as on eBay or Amazon, where customers provide their feedback about products and suppliers. It is possible to prepare ratings concerning the quality of web services. Such ratings can be based on both objective metrics, referring to the performance of a web

 ISBN: 978-1-61208-181-6

service and the opinions of the service users. In the field of data believability estimation, the most important are ratings concerning users' feedback about trustworthiness and believability of data delivered by web services. Various methods for rating web service quality and managing trust in web services were described by Golbeck [6].

Another important problem concerning data believability estimation is determining the metrics of data quality. Wang and Strong wrote an influential paper, in which they presented a detailed list of data quality attributes [2]. They also specified 20 dimensions characterizing data quality. Problems with measuring data quality are also the main subject of a book by Khan [7]. Furthermore, a paper [8] wrote by the same author as this paper, presents a method for evaluating the credibility of information in semantic web and knowledge grid.

III. A NEW METHOD FOR RATING DATA BELIEVABILITY

This section presents a new method for rating data believability. It was designed by the author of this paper and it is intended to be used in the integration of web services. The method consists of determining the level of data believability. This level is calculated on the basis of multiple metrics.

Common methods for rating data believability focus on determining the believability of individual sources, in order to select the most believable information provided by one or other of those available sources. The method introduced in this paper represents a different approach. It is designed to acquire the same kind of data from many different sources. On this basis conclusions about data believability are made. In the method, a level of data believability is calculated. The value of this level can be higher than 1. The level of believability is not like the probability of data trustworthiness. It is a perceived believability estimated by the method. In the case of one source of information, the value of the level ranges from 0 to 1. When there are more sources the level can be higher than 1.

In order to define metrics, which indicate the level of believability, a distinction between a claim and data, needs to be made. When some source of information publishes data, it cannot be assumed that this data is definitely true. Data provided by sources will be, in this paper, called claims, similarly as in [9]. Claims are also a kind of data, but there are two additional features of a claim:

- The source of a claim is specified.
- It is not resolved whether a claim is true or false.

Claims will be, in this paper, denoted by the symbol ζ. The data corresponding to claim ζ, but without a specified source, will be denoted by d_ζ.

It also needs to be considered what the granularity of data, being a claim, is. In the method, it is assumed that a claim is a portion of data of any size. A claim can be either one logical sentence, a sequence of such sentences, or the whole portion of data received from a web service. The size of the data does not affect the process of determining its believability. The only difference is that the level of believability concerns different data. The general formula for calculating the data believability level is given by (1).

$$b(d_\zeta) = \sum_{i=1}^{m_0} \left(\left(\frac{\sum_{k=1}^{M} w_k m_k(\zeta_i)}{M} \right) |\zeta_i| \right) \quad (1)$$

where b is the believability level, m_0 is the number of sources supporting claim ζ, which corresponds to data d_ζ, symbol ζ_i refers to claim ζ announced by source i, symbol m_k denoted metrics used in calculating the believability level, w_k is the weight of a metric m_k, letter M stands for the number of considered metrics, letters i and k are indexes used in additions. Symbol $|\zeta_i|$ in (1) represents the influence of a single claim given by the source number i to the level of data believability without the use of any metrics. Expression $|\zeta_i|$ is similar to the cardinality of a set. Every claim stands for a single portion of data, so the value of $|\zeta_i|$ is always equal to 1. This kind of notation is used in order to indicate that only claims, which correspond to data d_ζ, have influence on the level of this data believability.

Equation (1) states that the level of believability is equal to the number of claims supporting data d_ζ with regard to the metrics. Each source providing claim ζ_i increases the believability level of data d_ζ. The extent of this increase is equal to the weighted average of all metrics concerning the claim. Weights used in calculating the weighted average correspond to the importance of the metrics.

The following metrics are taken into account: quantity, reputation, approval, independence, traceability, maturity, authority and objectivity. The values of all metrics used in the method presented in this paper, apart from metric quantity, range from 0 to 1.

In the method presented in this paper, data believability is estimated on the base of attributes of the data source. There is also a possibility to estimate data believability on the basis of the data content. Several attributes of the data itself can be taken into account; like data validity, its accuracy and the context to which the data applies. In the method presented in this paper it is not considered. Data is only evaluated on the basis of the believability of its source. Nevertheless, it is possible to enhance the method with these attributes.

A. Quantity

The quantity of claims acknowledging the data being under verification is denoted with m_0. Although this value is not present in (1) in the same way as the other values of the metrics, it in fact refers to one of the metrics considered in determining the believability level. The value of m_0 defines the upper bound of the first summation presented in (1). The metric quantity represents the principle that the more sources are announcing the data, the greater is the level of believability. An assumption was made that the relationship between the number of claims and the level of believability, is linear when no other metrics are taken into account. All sources are then treated equally. The believability based only on the metric quantity is presented by (2).

$$b_{m_0}\left(d_\zeta\right)=\sum_{i=1}^{m_0}\left(\left|\zeta_i\right|\right)=m_0 \quad (2)$$

where b_{m0} stands for the level of believability when quantity is the only metric considered. In this case, the believability level is equal to the number of sources providing claims corresponding to data d_ζ.

B. Reputation

In this method for determining the believability level, sources are not treated as if they were equal. Their influence on the level of believability is biased by various factors. First one is the reputation of the source. The value of the metric's reputation is denoted by m_1.

The reputation of a web service is defined as "a general opinion i.e., it aggregates the ratings of the given service by other principals. Typically, a reputation would be built from a history of ratings by various parties" [10]. Principals are understood here as service providers or requesters. In the method for estimating believability presented in this paper these kinds of opinions are taken into account in the form of a metric *reputation*. The value of this metric is calculated on the basis of a web services' rating system. There are a large variety of such systems. However, they are most often concerned with many web services' parameters, such as performance, reliability, latency, fees etc. The reputation concerned in this paper is based only on the opinion about the believability of data provided by a web service.

The level of reputation can be acquired from web services collecting data about the quality of other web services and web services' ratings. The quality of web service, in the context of data believability, needs to be given in the form of a numerical rating (e.g., 6 using a scale 0 to 10). Such a rating is converted to scale from 0 to 1 and it is directly used as a value for a metric reputation. In the presented method the metric reputation can also be based on ratings acquired from many sources and rating systems. In this case the value of the metric is equal to the average of ratings adjusted to a scale from 0 to 1.

C. Approval

Approval is a metric, which is similar to the metric *reputation* in the way that it is also concerned with the behavior of a web service in the past. Whereas metric reputation is indicated by third parties, metric approval is the own opinion of a customer of a web service. The customer can, and should, store data about cooperation with web services. In the case where a web service provided data, which appeared to be wrong, the believability of this web service is decreased. On the other hand, the believability of proven web services should be increased.

The metric approval has three parameters: *q, p* and *T*. Parameter *p* is the influence of providing by a web service appropriate data, parameter *q* represents the impact of publishing wrong data and parameter *T* is concerned with the time, after which the influences caused by wrong and right data are no longer valid. It is assumed that the change in the value of metric approval is not perpetual and after some time the impact of providing wrong, or right, data is eliminated. The impact is diminished linearly, starting from the initial level of parameters *p* and *q* until there is no influence when time T has elapsed. The formula for calculating the metric approval is presented by (3).

$$m_2=m_{2base}\prod_{k=1}^{N_q}\left(q+\frac{t_k}{T}\left(1-q\right)\right)\prod_{l=1}^{N_p}\left(p+\frac{t_l}{T}\left(1-p\right)\right) \quad (3)$$

where m_{2base} is the default value of metric approval, N_q is the number of valid influences of providing wrong data, N_p is the number of valid influences of providing right data, q is the initial level of influence for wrong data, p is the initial level of influence for right data, t_k and t_l are times since the event of providing wrong or right data occurred and T is the parameter indicating the time, after which occurrence of wrong or right data is no longer taken into account.

The values of m_{2base} p, q and T included in (3) need to be specified. Parameter m_{2base} is the value of the metric when there is no experience in cooperation with the web service. The attitude to such a web service is neutral. Thus, the value of that metric is then equal to 0.5. In setting the value of parameter q it needs to be stated how severely the believability of a web service should be diminished when a web service provided wrong data. The parameter q can have various values. For example, it can be assumed that when no other metrics are taken into account, two web services, which once provided wrong data are as believable as one unknown web service. In this case, the value of metric approval is reduced by a half when a web service provides wrong data. When data acquired from a web service is right, the value of metric approval can be increased by half of its previous value. Thus, possible values of parameters are q=0,5 and p=1,5. For example, a web service, which once provided wrong data and once right, would have the level of metric approval equal to $m_{2base}\times p\times q$=0,5×0,5×1,5=0,375.

The value of parameter T can be selected arbitrarily and it can be set to 365 days. Values of t_k and t_l can be then changed once a day. They would indicate the number of days since wrong, or right, data was extracted from a web service.

D. Independence

The metric *independence* arises from the remark that data confirmed by two independent sources is more believable than data provided by two sources when one of those sources obtains data from the other one. A similar rule is applied by press agencies, assuming that information is true when it is confirmed by two independent sources.

In the method presented in this paper, the metric independence indicates the number of independent sources of data. If there is only one source of some information and other web service providers supply data on the basis of that one source, the value of metric independence is as minimal as possible. It is then equal to 0. On the other hand, a maximum value of metric, i.e., 1, indicates that there is an unlimited number of independent sources. In order to satisfy these conditions, the metric's value is increased because of subsequent independent sources in a similar way as a

geometric progression. The value of metric independence is presented in (4).

$$m_3 = 1 - \frac{1}{a^{u-1}} \quad (4)$$

where m_3 stands for the metric independence, a is the parameter determining the level of the increase caused by the existence of subsequent independent sources and u is the number of independent sources providing data. The value of a can be set to 2. Then, the medium value of the metric would mean duplication of data by two independent sources.

The metric independence also applies to situations more complicated than repeating data provided by an independent source. Web services are based on acquiring data from various sources. When there is a group of web services, there is also a group of independent sources, from which data is acquired. In the group of web services some part of data may be derived from a smaller group of independent sources than the other parts. Different parts of data can be confirmed by a different number of independent sources. There is a part confirmed by the smallest number of sources: the number of a source confirming this part is assigned as the value of parameter u from (4). Thus, when there is some data, which all web services acquired from the same source, the value of parameter u is set to the same level as if there was only one independent source. The value of u would be equal to 1.

E. Traceability

Metric *traceability* is another metric used in the method presented in this paper. The value of the metric is denoted by m_4. It depends on whether a web service specifies sources of information, which were used to make the service available, or there is no such information. If web services base their results on data acquired from other parties, they should provide information about that. For web services preparing all data by themselves, there should also be a notice that no external sources were used. Providing data about sources of information is possible due to the researches concerning storing information about data provenance.

Providing source information is similar to the bibliographies presented at the end of scientific manuscripts. When bibliography is not present or it is poor, a paper is treated as less credible. When a web service does not provide any data about its sources of information, the value of metric traceability is the smallest possible, i.e., equal to 0. If full information is available, the value of metric traceability is equal to 1. It is also possible that information about sources is partly present. In this case, metric traceability corresponds to the extent of source information availability.

F. Maturity

Another metric used in the method presented in this paper is *maturity*. The metric is based on the premise that web services, which are operational for some period of time are more believable than those, which are new and not tested by customers. Similarly, companies with tradition are more respected than the new, and unproven, ones.

The value of the metric maturity depends on the time elapsed since the release of a web service. The time, after which a web service is treated as fully believable, can be determined differently. It can be assumed that the believability of web services, which are available for over one year, is no longer reduced by the metric maturity. Web services, which are absolutely new, have the lowest value of metric maturity equal to 0. The values of this metric are changed linearly, for those web services whose time of service is in between these limits. Thus, the value of metric maturity is given by (5).

$$m_5 = \min\left(\frac{t}{T}, 1\right) \quad (5)$$

where t is the time, which elapsed since the release of a web service and T is the period of time, after which the believability of a web service is not reduced by the metric maturity. The parameter T can be set to 365 days and the value of the parameter t can be changed daily.

G. Authority

The value of metric *authority* is denoted by m_6. This metric refers to the sources' competence to provide data. In particular, it concerns data that sources claim to have obtained themselves. In the case where there is a doubt that the source does not have qualifications to provide a certain kind of data, the value of metric authority is reduced. For example, when a web service provides data concerning the number of people on Earth, claiming that this data was acquired by itself, there is a reasonable basis to distrust such information.

It is problematic to estimate the value of metric authority. Methods of storing provenance data do not reach a complex enough level to correlate the possibility of sources of information with data they provide. The value of metric authority needs to be assessed partly manually. In particular, this metric would concern data that only some kinds of sources are able to obtain. For example, information about population in countries can mainly be derived only from government sources. The value of metric would be equal to 0 for sources, which do not satisfy the requirements and it would be equal to 1 otherwise. A list containing specific kinds of data with corresponding sources can be prepared. When a list is available, applications using the believability estimation method presented in this paper can automatically use this previously prepared list.

H. Objectivity

The value of metric *objectivity*, denoted with m_7, is in most cases equal to 1. This value is changed for data that can be biased by the source due to its own interests. The value of metric objectivity is lowered for information that vendors claim about their products and their quality. Companies, on the basis of marketing needs, tend to modify information in order to improve their image. In such cases the metric objectivity is set to 0, as the source is not objective. It is possible to determine the value of this metric automatically

 ISBN: 978-1-61208-181-6

on the basis of metadata concerning product, their manufactures and resellers. In case no such data is available, these metrics can be set on the basis of a manually prepared list similarly as with metric authority.

I. *Weights of metrics*

The influence of metrics on the level of believability is modified by the weights of these metrics. Some metrics are treated as more important than others. Apart from metric quantity there are seven metrics with weights. The sum of weights has to be equal to one, because when there is only one source of information and all metrics are equal to one, then the level of believability needs to be also equal to one.

The most important metrics are reputation and approval. In fact, conclusions about the web service believability can be drawn only on the basis of own opinion about a web service and opinion of others. Moreover the values of other metrics in many cases will be the same for different web services. The weight of metrics reputation and approval need to be higher than other weights. Weights of metrics reputation and approval can be set to 0,25 and weights of other metrics can be equal to 0,1. Thus, $w_1=w_2=0.25$ and $w_1=w_2=w_3=w_4=w_5=0.1$.

J. *Application of the method*

There are three types of information sources used when the method is applied: web services, third parties and own knowledge. Third parties provide information about the quality of web services, such as their reputations. When an application needs a certain kind of information it collects claims from different web services concerning this information. It is like stating a question and seeking for the answer. The application also collects information from third parties and takes into account its own knowledge. The level of believability of each kind of the answer is rated with the use of the method. As the result, the answer with the highest level of believability is regarded as truthful. Acquiring data from many sources is more expensive then taking into account only one source, however the idea of the method is to improve the quality of data despite increased cost.

IV. EVALUATION

The method is based on the assumption that in general information provided by web services are truthful. The method resolves the problem of excluding information from untruthful, low quality web services (on the basis of metrics reputation, approval, maturity, authority and objectivity). It also manage the problem of providing wrong data by a noble web service due to some accidental mistake (on the basis of metrics quantity). In such cases, without using the method, false information would be regarded as truthful.

However the method does not guarantee the truthfulness of information. In case false information is universally regarded as truthful the method will also accept the truthfulness of information. Nevertheless the method attempts to disregard such information (on the basis of metrics independence and traceability). In case of such information there are some sources, which published it. If information concerning data provenance were commonly provided the spread of untruthful information could be limited. The results of the method in case of this kind of untruthfulness depend on the availability of information about data provenance.

V. CONCLUSION AND FUTURE WORK

The method presented in this paper makes it possible to automatically estimate the data believability level on the basis of information about data provenance and web services' ratings. In further work, we are planning to enhance the method with metrics referring not only to the believability of the source of data but also to the data itself.

One of the significant problems related to the presented method is that web services should provide information about sources of data, which were used to make the service available. This would protect other applications from propagation of wrong data in case some source is publishing not truthful information.

ACKNOWLEDGMENT

This work was supported in part by the Polish Ministry of Science and Higher Education under research project N N519 172337.

REFERENCES

[1] D. Booth, H. Haas, F. McCabe, E. Newcomer, M. Champion, C. Ferris, and D. Orchard (eds.), Web Services Architecture, W3C Working Group Note 11 February 2004, W3C, 2004.

[2] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers" Journal of Management Information Systems Vol. 12, No. 4, , Spring 1996, M. E. Sharpe, Inc., pp. 5-34.

[3] T. Tsai, X. Wei, Y. Chen, R. Paul, J.-Y. Chung, and D. Zhang, "Data provenance in SOA: security, reliability and integrity," Service Oriented Computing and Applications, vol. 1, no. 4, Springer-Verlag, Dec. 2007, pp. 223-247.

[4] W. Moreau "The Foundations for Provenance on the Web" Foundations and Trends in Web Science, Now, 2009, submitted for publication.

[5] M. Ouzzani and A. Bouguettaya, "Efficient Access to Web Services" IEEE Internet Computing, Mar./Apr. 2004, IEEE Computer Society, pp. 34-44.

[6] J. Golbeck, "Trust on the World Wide Web: A Survey," Foundations and Trends in Web Science, Vol. 1, No. 2, Now, 2006, pp. 131-197.

[7] K. M. Khan (ed.), Managing Web Service Quality: Measuring Outcomes and Effectiveness, IGI Global, 2009.

[8] A. L. Kaczmarek, "Automatic Evaluation of Information Credibility in Semantic Web and Knowledge Grid, "Proc. of the 4th International Conf. on Web Information Systems and Technologies (WEBIST 2008), vol. 2, Funchal, Madeira-Portugal, INSTICC, May 2008, pp. 275-278.

[9] C.Bizer and R.Oldakowski, "Using Context- and Content-Based Trust Policies on the Semantic Web," WWW 2004, New York: ACM, May 2004, pp. 228-229.

[10] E. M. Maximilien and M. P. Singh, "Conceptual Model of Web Service Reputation" ACM SIGMOD Record: Special section on semantic web and data management, Vol. 31 , No. 4, Dec. 2002, pp. 36-41.

 ISBN: 978-1-61208-181-6