An Efficient Query Processing with Approval of Data Reliability using RBF Neural Networks with Web Enabled Data Warehouse

K. Soundararajan
Associate Professor
Vivekanandha College of Engineering for women,
Tiruchengode, Tamil Nadu.
Email: soundar919@gmail.com

Dr. S. Sureshkumar
Principal
Vivekanandha College of Technology for women,
Tiruchengode, Tamil Nadu
Email: nice.ssk@gmail.com

P.Selvamani
Department of IT
Vivekanandha College of Engineering for Women,
Tiruchengode, Tamil Nadu.
Email: selvi30it@gmail.com

Abstract— **To rise above the limitation of the Traditional load forecasting method using data warehousing system, a new load forecasting system basing on Radial Basis Gaussian kernel Function (RBF) neural network is proposed in this project. Genetic algorithm adopting the actual coding, crossover and mutation probability was applied to optimize the parameters of the neural network, and a faster growing rate was reached. Theoretical analysis and models prove that this model has more accuracy than the traditional one. There are several methods available to integrate information source, but only few methods focus on evaluating the reliability of the source and its information. The emergence of the web and dedicated data warehouses offer different kinds of ways to collect additional data to make better decisions. The reliable and trust of these data depends on many different aspects and meta-information: data source, experimental protocol. Developing generic tools to evaluate this reliability represents a true challenge for the proper use of distributed data. In this project, RBF neural network based approach to evaluate data reliability from a set of criteria has been proposed. Customized criteria and intuitive decisions are provided, information reliability and reassurance are most important components of a data warehousing system, as their power in a while retrieval and examination.**

Keywords- RBF neural network, uncertainty, reliability, meta-information, query processing.

I. INTRODUCTION

One of the most important issues in building data warehouse is the evaluating its reliability during extracting the information from it. Due to the growth of the web, a large amount of data is distributed on the Web and large data sets are interlinked. Many new applications that emerge utilize this data in a useful and in innovative ways. The growth and the emergence of dedicated data warehouses offer different methods and models to collect additional data to build models or to make better decisions. The reliability and trust of these data depends on many different aspects and meta-information: data source, experimental protocol. However, the openness of the Web and the ease to combine Linked Data from different sources creates new challenges.

Developing tools to evaluate this reliability represents a true challenge for the proper use of distributed data. There are also very few works that propose a solution to evaluate data reliability. The data reliability can be assessed by comparing the data to the model exists and the values of the data should be within the range of the mode. But, such models are not always available and data reliability must then be estimated by some other means. This estimation is especially important in areas where data are scarce and difficult to obtain.

The huge amount of documents (technical, scientific...) available on the Web includes many data tables. In addition to local data sources, it provides possible external data sources for the data warehouse, to a given domain of application. Estimating data consistency is a most important, as these data are used in additional inferences. And also to reduce the burden of experts who selects the data from the data warehouse for a particular application, it is necessary to give them analytical reliability evaluations. In this paper, it presents a framework to estimate the reliability of data tables collected. Compared to more peculiar or particular estimation, the presented method can give insights to the expert, to why a particular data table is tagged as reliable or not reliable.

During the collection of data, data reliability is regularly ensured by adapting experimental design and by statistical recurrence. However, full traceability is no longer ensured when data are reused at a later time by others. Therefore reliability estimation is an essential part of the Semantic Web architecture, and many research works focus on issues such as source authentication, reputation, etc. We propose a method to evaluate data reliability from meta-information. Several criteria are used, each one providing a piece of information about data reliability. These pieces of information data are then aggregated into a global assessment that is sent back to the users after proper post-treatment. In Section 2 discusses related notions and works detail the proposed method to assess Web data table reliability. Section 3 describes the purpose and architecture of the ontological data warehouse In Section 4 shows the proposed method and how this reliability assessment is

presented and explained to the user. Finally, Section 5 describes the results and conclusion.

II. RELATED NOTIONS

In evidence theory, methods to evaluate reliability consist in choosing reliability scores that minimize an error function [11]. In spirit, the approach is similar to compare source assessments with reference values (as done to evaluate experts in probabilistic or possibilistic [16] methods). It requires the definition of an objective error function and a fair amount of data with known reference value. This is hardly applicable in our case, as data are sparse and can be collected and stored for later use, i.e., not having a specific purpose in mind during collection. Other approaches rely on the analysis of conflict between source information [6], assuming that a source is more reliable when it agrees with the others. This comes down to make the assumption that the majority opinion is more reliable. If one accepts this assumption, then the results of such methods could possibly complement our approach.

Reliability has strong connections with other notions such as relevance, truthfulness, trust, data quality. All these related concepts are, however, either different from or less specific than the notion of reliability. First, there is a difference between data reliability, i.e., the trust we can have in data values, and data relevance, i.e., the degree to which data can answer our needs. Indeed, some data can be totally reliable and totally irrelevant for a given situation.

Also, while the notion of reliability is rather objective (data are reliable or not, irrespectively of the situation), the notion of relevance only makes sense w.r.t. a particular question of interest. Note that in some occasion a same criterion can be used to evaluate the reliability and relevance, but the two notions can behave differently w.r.t. this criterion. Example is the data imprecision, to which the notions react in opposite ways: when data are more imprecise, data reliability increases (i.e., they are more likely to contain the true value) and data relevance decreases (i.e., they are less likely to answer a specific question). So, both notions of data reliability and relevance need to be considered in real-world cases, but they should not be confused and they require separate treatment and evaluation.

Similarly, the notion of source sincerity or truthfulness [13] is different and should be handled individually. This notion only makes sense if the source can be suspected of lying in order to gain some advantage, and is different for reliability. For instance, one may infer the truth from a lying source if the lying mechanism is known, while nothing may be inferred from an unreliable source. In our particular application area, sources are seldom untruthful, contrary to what may happen in other domains (e.g., multi-agent systems). Finally, the problem of evaluating data reliability can be paralleled with the one of trust evaluation that has recently received much attention, as it is an important aspect of multiagent systems [14] and of the semantic web [19]. Trust is an integral part of the Semantic Web architecture, and many research works [8] focus on issues such as source authentication, reputation or cryptography.

Indeed, data reliability can be seen as the amount of "trust" we have in the data validity. However, the notion of trust is often considered in a large sense and can include considerations related to reliability as well as to relevance or truthfulness. In multi-agent systems, the notion of trust as well as the information used to evaluate it can take many forms [14]. One can differentiate between individual-level and system-level trusts, the former concerning the trust one has in a particular agent, while the latter concerns the overall system and how it ensures that no one will be able to take advantage of the system.

The collected information about the trustworthiness of an agent may be direct, i.e. coming from past transactions one has done with this agent. It can also be indirect, i.e. provided by third-party agents. In that case, it can be based on an evaluation of the agent reputation or on information concerning some of its characteristics. This latter type of evaluation is the closest to the current work. However the information itself may have an impact on trust assignment, either due to its nature or to its content. As the Semantic Web provides tools to formalize and exchange the content of resources, a new potential arises for further innovative techniques that examine the information content when determining trust. A recent paper [7] describes at length several factors that affect how users determine trust in content provided by a Web resource.

Naturally these include source authority and direct experience. Among the remaining factors, one can find items like topic and criticality, which are somehow related to data relevance. The limitation of resources may play a role, as well as the incentive to provide good information or on the contrary to be biased or deceptive (elements which are related to the notion of truthfulness). Source agreement and user expertise also have an impact.

Another paper [21] supports for many aspects of approach to trust models in internet environments. The authors point out the great number of terms and intertwined meanings of trust, and the difficulty to capture the wide range of subjective views of trust in single-faceted approaches. Therefore OWL based ontology has been proposed which is of trust related concepts, such as credibility, honesty, reliability, reputation or competency, as well as a meta-model of relationships between concepts. Through domain specific models of trust, they can propose personalised models suited to different needs.

The idea is to provide internal trust management systems, i.e. the trust assessment being made inside the system, to collect reliable data. Among the several methods of proposing solutions to evaluate the reliability or data quality in web related applications, this method is close to our proposal, but uses possibility theory as a basis for evaluations rather than belief functions.

Another difference between this approach and ours is that global information acquired is not by a mixture of multiple uncertainty models, but by the propagation of uncertain criteria through an aggregation function (e.g., a weighted mean). Each method has its pros and cons: it is easier to integrate criteria interactions in aggregation functions, while it is easier to retrieve explanations of the final result in our approach.

The problem with using available trust evaluation systems is that they are often intended for a particular service or to a particular application. They propose relatively simple representation and aggregation tools, notably due to the fact that Semantic Web applications are confronted to scalability issues. Our situation is somehow different, since we aim for a general method applicable to situations where the number of items will seldom exceed tens of thousands, and will in fact be often limited to some dozens.

In [2], the impact of data quality on decision making is explored, and an experimental study about the consequences of providing various kinds of information (none, two-point ordinal and interval scale) regarding the quality of data is performed. They point out that the availability of the information is not enough, and that an important consideration is how data quality information is recorded and presented. Decision tasks (apartment or restaurant selection) were experimented using two groups of subjects, one group performing the tasks without data quality information and the other one with a given quality format.

Users were asked to explain their decision process, and issues like complacency, consensus and consistency were examined. Complacency is a measure of the degree to which data-quality information is ignored. The following interesting result was observed. For simple tasks, the smallest level of complacency, corresponding to the greatest impact of data quality information, was observed when comparing groups with interval scaled data quality information with groups with no data quality information. For complex tasks, there seems to be an information overload effect, and no statistically significant conclusions appear. This is an important point in favour of giving a lot of attention in presenting readable results to end users.

III. THE ARCHITECTURE DATA WAREHOUSE

A. Existing system

Data warehouses are designed to aggregate data and allow decision makers to obtain accurate, complete and up to date information, but a data warehouse is only as good as the data it contains. Most existing warehouses contain a subset of the available data and allow a narrow range of queries on that data, simply because the extra benefit is not considered sufficient to justify the cost of development of a more complete system. Data warehouse opened on the Web focuses on the aggregation of heterogeneous data tables extracted from web documents. The attention has been put on Web tables for the following two reasons: (i) experimental data are often summarized in tables and (ii) data are already structured and easier to integrate in a data warehouse than, e.g., text or graphics. The main steps of Web table integration are given in Fig 1. The key point of data integration in Web is the central role played by the domain ontology.

This ontology describes the concepts, their terminology and the relationships between concepts proper to a given application domain. Web can be instantiated for any application domain by defining the corresponding ontology including the domain knowledge. Once the ontology is built, data integration in the warehouse is done according to the steps of Fig. 1. Concepts found in a data table and semantic relations linking these concepts are automatically recognised and annotated, which allows interrogation and querying in an homogeneous way.

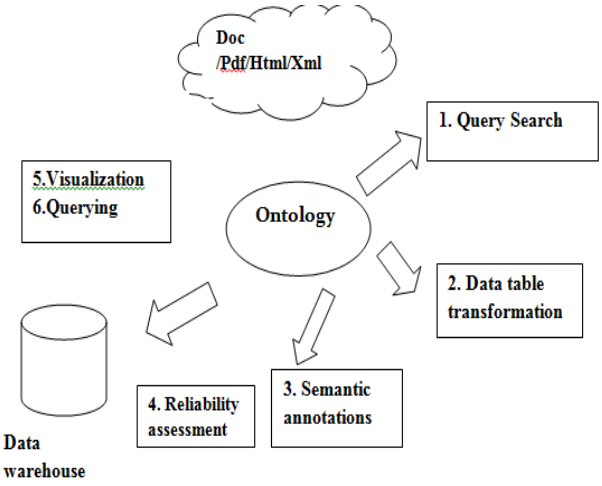


Figure 1. Ontological data warehouse

Web generic ontology: The current OWL ontology used in the @Web system is built by two main parts: a generic part, the core ontology, which contains the structuring concepts of the Web table integration task, and a specific part, the domain ontology, which contains the concepts specific to the domain of interest. The core ontology is composed of symbolic concepts, numeric concepts and relations between them. It is therefore separated from the definition of the concepts and relations specific to a given domain, the domain ontology. All the ontology concepts are materialized by OWL classes.

Web workflow: The first three steps of Web workflow are the following: the first task consists in retrieving relevant Web documents (in html or pdf) for the application domain, using key-words extracted from the domain ontology. It does so by defining queries executed by

different crawlers; in the second task, data tables are extracted from the retrieved documents and are semi-automatically translated into a generic XML format.

The Web tables are then represented in a classical and generic way – a table is a set of lines, each line being a set of cells; in the third task, the Web tables are semantically annotated according to the domain ontology. This annotation consists in identifying what semantic relations of the domain ontology can be recognized in each row of the Web table. This process generates RDF descriptions.

Validating the trust and reliability measures and criteria will hopefully be a better plus to justify the choices and to capitalize knowledge than the use of ad-hoc estimation. Tools development was carefully done, web recommended languages, so that created tools would be standard, universal and reusable in other data warehouses. When dealing with thousands of data and over five criteria groups, many tools may help users to do a quick analysis and retain the data that best serve their purposes.

Even though there are many methods available in existing System to integrate information source reliability in indecision form of representation, but these are not focusing on the problem of evaluating this reliability. The existing system consists of following disadvantages:

- Less reliable.
- Less relevance.
- No idea about feedback from user.
- Large test data volume and long test time
- Basically a process speed is very slow.
- Inability to recover from database corruption.
- Complexity

III. PROPOSED SYSTEM FOR RELIABILITY ASSESSMENT

A. Methodology

RBF is a three-layer neural network. The input layer is made up of the signal source node. The second tier is hidden layer, as described in its modules are based on the needs and problems. The third tier is the output layer, which responded to the role of imported models. RBF implied by the input space to transform the hidden space is non-linear, from the hidden layer and output layer space to transform the output space is linear. The transformation function of hidden units is Radial Basis Functions.

The hope we know how to enclose in data values, and data significance, the extent to which information can respond our require. Only makes intelligence but the spring be capable of be so-called of insincere in regulate to expand various benefit, and is discrete from reliability. The trouble of evaluating data reliability can be paralleled with the one of trust RBF based evaluation. It can be used on an evaluation of retrieving the information's and high reliability of data. The user input the query to search our related data's or information getting from database management or data warehousing system. Data warehousing concept is made that is collecting data .managing and retrieving of particular is made.

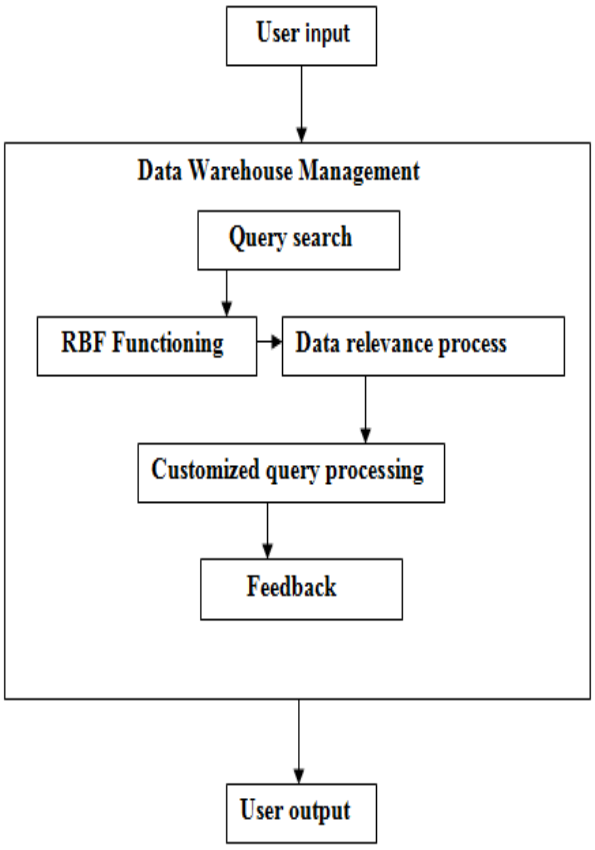


Figure 2. RBF function architecture

The emergence of the web and dedicated data warehouses offer different kinds of ways to collect additional data to make better decisions. The reliable and trust of these data depends on many different aspects and meta-information. The neural network function is made for data reliability.

Reliability has strong connections with other notions such as relevance, truthfulness, trust, data quality. Among the several methods of proposing solutions to evaluate the data quality in web related applications, this method is close to our proposal and is recommended, but uses possibility theory as a basis for evaluations rather than belief functions. It has following advantages:

- More reliable data functioning.
- Multi-layer support data relevance.
- Getting feedback from user.
- Reduces time in retrieving information.

In this project, the data reliability and confidence are added in the data ware house. This may be the short thing but this is very essential in real world. First step is to query of information by the user in the search text box then the process of query with data reliability and confidence with the help of user feedback and neural network. Feedback from user is used to easy identification for data reliability. Normally RBF neural network is used to extract the important feature from the given query or data with the priority wise increment by the usage of the user.

*i) Query search and Data management: T*he user query is taken as input and is searched from database management or data warehousing system. Searching is not limited for the user. Here data ware housing concept is used to collect the data and managing it, then retrieving of particular information is obtained. Due to the increase development in data warehouse, the information obtained by it, is used to make better decisions.

ii) RBF Neural network functions: The neural network function used in this project to obtain a reliable data. Reliability has strong connections with other notions such as relevance, truthfulness, trust, data quality. Among the several methods of proposing solutions to evaluate the data quality in web related applications, this method is very close to our proposal. Here RBF neural network is used for improving the data reliability.

iii) Data relevance: The data relevance is defined as the degree to which data can answer our need. Naturally these include source authority and direct experience. Among the remaining factors, one can find items like topic and criticality, which are somehow related to data relevance. The limitation of resources may play a role, as well as the incentive to provide good information or on the contrary to be biased or deceptive.

iv) *Customized query processing:* Here in this module, user required data is modified or customized according to their need. Here unwanted or unrelated search has been greatly reduced. It provides only the optimized search.

v) Feedback: The feedback is for improving data relevance concept. The user sends feedback to the administrator of the data warehouse management system after searching has completed. Then the feedback is collected by the admin, if any of the feedback is irrelevant for the given content, they can modify or update the data's. The admin have the rights to access and delete the feedback.

Finally the output is given to the user with data relevance and data reliability. Relevant data are the answers to the user query; some of these answers may be unreliable. The accuracy of information gets increased, based on the feedback which is provided by different user while searching for the data.

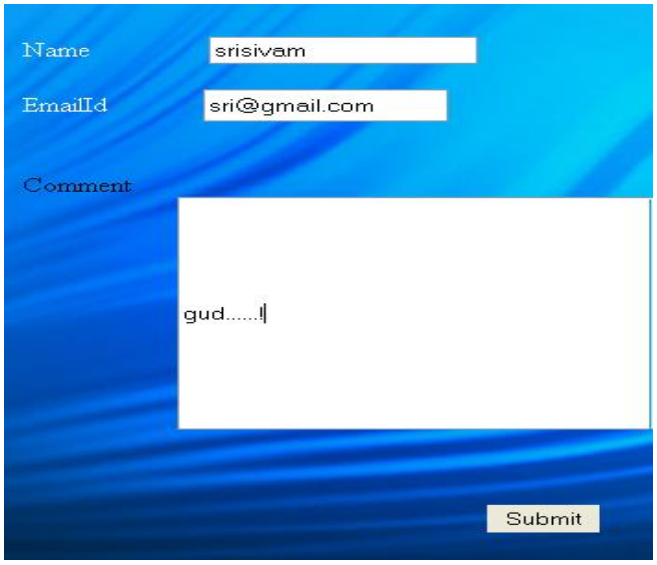


Figure 3 Feedback Form

Name	EmailId	Comment
poorni	poo@gmail.com	very good...
ananth	ananth@gmail.com	very very bad...
uis	uis@gmail.com	not bad....
united	unit@gmail.com	gud...
fdsdf	dfdf@gmail.com	xcdfgfdg
nam	d	dsf
sa	as	as
sada	asdasd	asdasd
srisivam	sri@gmail.com	gud.....!

Figure 4 Feedbacks



Emails	query	linkname	noofclick
sri@gmail.com	dotnet	http://www.Csharpdotnet.edu.aspx	2
sri@gmail.com	dotnet	http://www.msdnCsharpdotnet.edu.aspx	2
sri@gmail.com	java	http://www.Cjava2snet.edu.aspx	1
sri@gmail.com	java	http://www.javaenters.net.aspx	1
sri@gmail.com	java	http://www.javas2sserver.com	1

Figure 5 Rank List

IV. DISCUSSION

In this paper data reliability concept has been implemented in data warehouse for effective searching and retrieval of information. Feedback is given by user after every search of information, so that by using this feedback information the data can be updated. Therefore information accuracy gets increased; users can also obtained more relevant information to their need.

V. CONCLUSION

A new load forecasting system basing on RBF neural network was proposed, and genetic algorithm was applied to optimize the parameters of the neural network. The simulation results show that the forecasting accuracy and convergence speed of this model has been improved greatly. And it can meet the needs of load forecasting. Finally, dependability evaluations and ordering of data tables are achieved by means of higher opportunity, allowing us to mirror ambiguity in the evaluation. The importance display to end users is an ordered list, from the most to the slightest reliable ones, collectively with an interval-valued assessment.

REFERENCES

[1] *Benferhat.S, Dubois.D, and Prade.H (1996), 'Reasoning in inconsistent stratified knowledge bases'*, in Proc. of the 26th Inter. Symp. On Multi- Valued Logic (ISMVL), pp. 184–189.

[2] Chengalur-Smith.I.N, Ballou.D.P, and Pazer.H.L (1999), '*The impact of data quality information on decision making: An exploratory analysis'*, IEEE Transactions on Knowledge and Data Engineering, vol. 11, pp.853–864, 1999.

[3] Destercke.S, Dubois.D, and Chojnacki.E (2009), *'Possibilistic information fusion using maximal coherent subsets*," IEEE Trans. on Fuzzy Systems, vol. 17, pp. 79–92.

[4] Denoeux.T (2008), '*Conjunctive and disjunctive combination of belief functions induced by non-distinct bodies of evidence'*, Artificial Intelligence, vol. 172, pp. 234–264.

[5] Denoeux.T (2009), '*Extending stochastic ordering to belief functions on the real line*', Information Sciences, vol. 179, pp. 1362–1376.

[6] F. Delmotte and P. Borne, "*Modeling of reliability with possibility theory*," *IEEE Trans. on Syst., Man, and Cybern. A*, vol. 28, no. 1, pp. 78–88, 1998.

[7] Gil.Y and Artz.D (2006), '*Towards content trust of web resources*' in WWW'06 : Proceedings of the 15th international conference on World Wide Web, New York, NY, USA, pp. 565–574.

[8] Golbeck.J and Hendler.J (2004), '*Inferring reputation on the semantic web'* in Proceedings of the 13th international World Wide Web conference, New York.

[9] Jane M. Booker, Mark C. Anderson, Mary A. Meyer (2006), *'The role of expert knowledge in uncertainty quantification'*, IEEE transaction on knowledge and data engineering.

[10] Li Ma, Jing Mei, Yue Pan, Krishna Kulkarni, Achille Fokoue, Anand Ranganathan (2006), '*Semantic web technologies and data management*', IEEE transaction on knowledge and data engineering.

[11] Mercier.D, Quost.B, and Denoeux.T (2008), '*Refined modelling of sensor reliability in the belief function framework using contextual discounting*', Information Fusion, vol. 9, pp. 246–258.

[12] Patrice Buche, Juliette Dibie-Barthelemy, Liliana Ibanescu, and Lydie Soler(2011),*'Fuzzy web data tables integration guided by an ontological and termnological resource'*, IEEE transaction on knowledge and data engineering, 1041-4347/11.

[13] Pichon.F, Dubois.D, and Denoeux.T (2011), '*Relevance and truthfulness in information correction and fusion'*, in International Journal of Approximate Reasoning.

[14] Ramchurn.S, Huynh.D, and Jennings.N (2004), '*Trust in multi-agent Systems'*, The Knowledge Engineering Review, vol. 19, pp. 1–25.

[15] Smets.P (2006), *'Analysing the combination of conflicting belief functions,*" Information Fusion, vol. 8, pp. 387–412.

[16] Sandri.S, Dubois.D, and Kalfsbeek.H (1995), '*Elicitation, assessment and pooling of expert judgments using possibility theory*', IEEE Trans. On Fuzzy System*s*, vol. 3, no. 3, pp. 313–335.

[17] Sébastien Destercke, Patrice Buche, and Brigitte Charnomordic (2011*), 'Evaluating Data Reliability: An Evidential Answer with Application to a Web-Enabled Data Warehouse'*, IEEE Trans. Knowl. Data Eng. 25(1): 92-105.

[18] S´ebastien Destercke and Patrice Buche and Brigitte Charnomordic(2011),*'Data reliability assessment in a data warehouse opened on the web' Author manuscript, published in Flexible Query Answering Systems*, Ghent : Belgium (2011), DOI : 10.1007/978-3-642-24764-4_16.

[19] Sabater.J and Sierra.S (2005), '*Review on computational trust and reputation models*' Artificial Intelligence Review, vol. 24, no. 33-60.

[20] Zied Elouedi, Eric Lefèvre, David Mercier(2010)'*Discountings of a Belief Function using a Confusion Matrix'* International Conference on Tools with Artificial Intelligence. 1082-3409/10.

[21] K. Quinn, D. Lewis, D. O'Sullivan, and V. Wade, "*An analysis of accuracy experiments carried out over a multi-faceted model of trust," International Journal of Information Security*, vol. 8, pp. 103–119, 2009.