International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-3)

Research Article | **March 2015**

Efficient and Flexible Method for Evaluating Data Reliability Based on Relevance Feedback

Vibha Sharma, Prof. A.D. Gujar
TSSM'S Bhivarabai Sawant college of Engineering and Research,
Pune, Maharashtra, India

Abstract—

Now days the use of web recommendation systems is increasing in order to provide the customized data for end users, hence this area becomes challenging for researchers. Basically there are two categories in which web recommendation systems are divided such as content based web recommendation system and collaborative web recommendation system. In collaborative recommendation systems, such systems trying to find out the users those are sharing the similar tastes for particular end user and also according to the linking of that user websites are recommended. On the other hand, content based recommendation systems recommending the websites those are same as websites liked by end user. We have studied recently the improved method of mining algorithm with aim of solving the problem of efficient and reliable web page recommendation. We have added user feedback as an additional (and parallel) source of information about reliability or relevance. The importance is changes according to number of visits to that particular web page as well as amount time user stay over the same page. We have used the feedback for researching in less time consumption. In this paper we are presenting the different types of web recommendation systems, different methods presented for web recommendation systems and then future work for the same.

Index Terms— collaborative web recommendation system, content based web recommendation, mining algorithm, efficient reliability and relevance.

I. INTRODUCTION

The huge amount of technical and scientific documents available on the Web includes many data tables. In addition to local data sources, they represent big potential external data sources for the data warehouse of a company dedicated to a given domain of application. To lighten the burden laid upon domain experts when selecting data from the data warehouse for a particular application, it is necessary to give them indicative reliability evaluations. In this paper, we present a framework to estimate the reliability of data tables collected from the Web. Compared to more ad-hoc estimation, the presented generic method can give insights to the expert as to why a particular data table is tagged as reliable or not reliable. Due to its generic nature, this method can be reused in other data warehouses using the semantic web recommended languages.

Reliability estimation is an essential part of the Semantic Web architecture, and many research works [1] focus on issues such as source authentication, reputation, etc. For example advocates a multi-faceted approach to trust models. They propose an OWL based ontology of trust related concepts. The idea is to provide systems using the annotation power of a user community to collect information about reliability. Our approach is different, as we do not rely on users but rather on information about the Web data table origins to compute reliability estimations. Among methods proposing solutions to evaluate trust or data quality in web applications, the method presented in [3] is close to the method presented in the paper. It uses possibility theory evidence theory, whereas we base our method on evidence theory. Another difference is that in our approach global information is obtained by a fusion of multiple uncertainty models, while in [3] global information results from the propagation of uncertainty models through a aggregation function. Each method has its pro and cons: it is easier to integrate interactions between criteria in aggregation functions, while it is easier to retrieve explanations of the final result in our approach.

Data reliability is mostly ensured by measurement device calibration while collection, by adapted experimental design and by statistical repetition. But full traceability is no longer ensured when data are reused at a later time by other scientists. If a validated physical model exists and data values fall within the range of the model validated domain, then data reliability can be assessed by comparing data to the model predictions. However, such models are not always available and data reliability must then be estimated by other means. This estimation is especially important in areas where data are scarce and difficult to obtain (e.g., for economical or technical reasons), as it is the case, for example, in life Sciences. The growth of the web and the emergence of dedicated data warehouses offer great opportunities to collect additional data, be it to build models or to make decisions.

In this paper we are presenting the extended method to overcome the limitations of previous methods. We focus on extend generic method to evaluate the reliability of data automatically retrieved from the web or from electronic documents by considering the additional features like dealing with multiple experts, dealing with criteria of nonequal importance as well as with uncertainly known criteria. In addition to this, we are combining the current approach with other notions or sources of information: in particular, relevance appears to be equally important to characterize

experimental data. In this project we are adding the user feedback as an additional source of information about reliability or relevance, as it is done in web applications.

In next section II we are presenting the literature survey over the various methods security at data sharing systems. In section III, the proposed approach and its system block diagram is depicted. In section IV we are presenting the current state of implementation and results achieved. Finally conclusion and future work is predicted in section V.

II. LITERATURE SURVEY

Existing Methods

There are many methods presented to evaluate the reliability of data automatically retrieved from the web or electronic documents. Recently we have studied the new method in [1]. In [1], generic method to evaluate the reliability of data automatically retrieved from the web or from electronic documents. This method overcomes the many limitations of previous methods; however still this approach needs to be extending further with complementing the current method with useful additional features. In this project we are extending the current method of data reliability using the additional features such as coping with multiple experts etc. as well as combining the current approach with other notions or sources of information.

- **Flexible Sparql Querying of Web Data Tables Driven by an Ontology :**

This paper concerns the design of a workflow which permits to feed and query a data warehouse opened on the Web, driven by domain ontology. This data warehouse has been built to enrich local data sources and is composed of data tables extracted from Web documents. We recall the main steps of our semi-automatic method to annotate Web data tables driven by domain ontology. The output of this method is an XML/RDF data warehouse composed of XML documents representing Web data tables with their fuzzy RDF annotations. We then present how to query simultaneously the local data sources and the XML/RDF data warehouse, using the domain ontology, through a flexible querying language. This language allows preferences to be expressed in selection criteria using fuzzy sets. We study more precisely how to retrieve approximate answers extracted from the Web data tables by comparing preferences expressed as fuzzy sets with fuzzy annotations using SPARQL.

- **Fuzzy Annotation of Web Data Tables Driven by a Domain Ontology :**

We propose an automatic system for annotating accurately data tables extracted from the web. This system is designed to provide additional data to an existing querying system called MIEL, which relies on a common vocabulary used to query local relational databases. We will use the same vocabulary, translated into OWL ontology, to annotate the tables. Our annotation system is unsupervised. It uses only the knowledge defined in the ontology to automatically annotate the entire content of tables, using an aggregation approach: first annotate cells, then columns, then relations between those columns. The annotations are fuzzy: instead of linking an element of the table with a precise concept of the ontology, the elements of the table are annotated with several concepts, associated with their relevance degree. Our annotation process has been validated experimentally on scientific domains (microbial risk in food, chemical risk in food) and a technical domain (aeronautics).

- **Inferring Reputation on the Semantic Web :**

The so-called "Web of Trust" is one of the ultimate goals of the Semantic Web. Research on the topic of trust in this domain has focused largely on digital signatures, certificates, and authentication. More social notions of trust which are reputation based are beginning to gain some attention in their own right, but have been traditionally overlooked. In this paper, we describe an algorithm for generating locally calculated reputation ratings from a semantic network. We present mathematical and experimental results that show the effectiveness of this algorithm to accurately infer the reputation of a node. We then describe Trust Mail, an application that uses the network for rating relevant emails.

- **Towards Content Trust of Web Resources :**

Trust is an integral part of the Semantic Web architecture. Most prior work on trusts focuses on entity-centered issues such as authentication and reputation and does not take into account the content, i.e., the nature and use of the information being exchanged. This paper defines content trust and discusses it in the context of other trust measures that have been previously studied. We introduce several factors that users consider in deciding whether to trust the content provided by a Web resource. Our goal is to discern which of these factors could be captured in practice with minimal user interaction in order to maximize the quality of the system's trust estimates. We present results on a study to determine which factors were more important to capture, and describe a simulation environment that we have designed to study alternative models of content trust.

III. PROPOSED APPROACH

There are few methods presented for evaluating the data reliability from meta information. However these methods are failing to achieve solution for conflicting information and hence such methods are not traceable. Some methods are not readable to end users, both in its different input parameters and results, as the method and the system it is implemented in will be used mainly by non-computer scientists. Hence to address such issues one needs to have efficient method in place. Recently we have studied the new approach for evaluating data reliability in order to achieve the solutions over above said limitations.

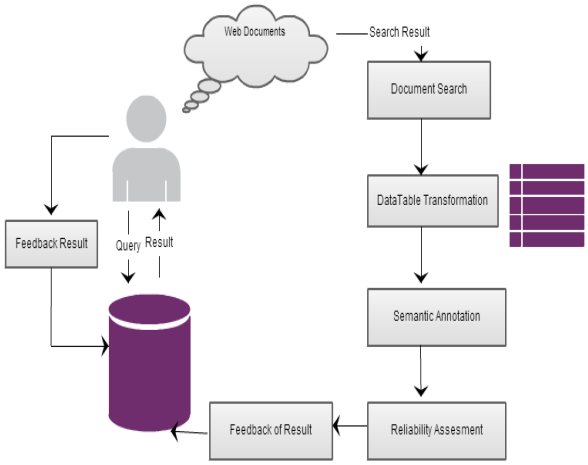


Fig 1: Proposed Systems

This method is presented in [1]. From experimental analysis, this method shows better performance. However this further needs to extend in order to make it more effective and dynamic with use of user feedback, use of this method with some other source of information.

Reliability with coherent subset

We assume that reliability takes its value on a finite ordered space

Input

$\Theta = \theta 1 \ldots\ldots\ldots\ldots\ldots\ldots \theta N$ $\{\theta i < \theta$ jiff $i<j\}$

θN= Total reliability

$I_{a,b} = \{\theta_a \ldots\ldots\ldots\ldots \theta_b\}$ $(a \leq b)$ these are the interval values for realibility

$S = \{A_1 \ldots\ldots\ldots\ldots\ldots A_S\}$ is group criteria, A group may be composed of multiple Criteria.

Each group Ai, i= 1 . . . ……S can assume Ci distinct values on a finite space Ai ={ai1; . . . ; aiCi}.

For each possible value of each criteria group A1. . ….AS, a domain expert is asked to give its opinion about the corresponding data reliability.

A fuzzy set μ defined on a space A is a function μ:→[0, 1] with μ(x) the membership degree of x. Recall that the support S(μ) and kernel K(μ) of a fuzzy sets are the sets S(μ)={x∈A|μ(x) >0} and K(μ)={x ∈A|μ(x)}

Mathematical model

Input: S1…….SN source data e.g. doc, web links etc.

Process:

Step 1: Extract source data from Google api by user query search
api= {s1→webdata, s2→weblink, s3→ data}

Step 2: Annotate that data

Step 3: Apply k-means document clustering algorithm

Step 4: Apply maximal coherent subset algorithm

Step5: Take feedback for corresponding data

Step6: Store feedback in database and use it next time for the same query

Output: Reliable source data

1. Data Table Transformation:

In this phase, we extract html tables from web documents like web page, pdf etc.

Input: D: DOM node (root node), T: set of block defining HTML elements
Output: B: Set of blocks

```
B ⇐ D
for all t ∈ T do
  for all b ∈ B do
    if b hasChildNode(t) then
      B^N ⇐ getBlocks(b, t)
      B ⇐ (B − b) ∪ B^N
    end if
  end for
end for
return B

function getBlocks(b, t);
B ⇐ ∅
C ⇐ descendants(b)
for all m ∈ C do
  if elementType(m) = t then
    B ⇐ B ∪ {m}
  end if
end for
```


2. Semantic Annotation using k-means clustering Algorithm:

Input: Web tables

Output: Similar type of tables

In the clustering problem, we are given a training set {x^{(1)}, ... , x^{(m)}}, and want to group the data into a few cohesive "clusters." Here, we are given feature vectors for each data point x^{(i)} \in \mathbb{R}^n as usual; but no labels y^{(i)} (making this an unsupervised learning problem). Our goal is to predict k centroids and a label c^{(i)} for each datapoint. The k-means **clustering algorithm** is as follows:

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.

2. Repeat until convergence: {

   For every i, set
   $$c^{(i)} := \arg\min_j \|x^{(i)} - \mu_j\|^2.$$

   For each j, set
   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }

3. Reliability Assesment:

So far the method used for the data reliability is the generic method where the Maximal Coherent Systems (MCS) and merging have been used. Merging is the method of Combine or cause to combine to form a single entity, esp. a commercial organization. MCS consists in applying a conjunctive operator within each nonconflicting (maximal) subset of sources, and then using a disjunctive operator between the partial results. With such a method, as much precision as possible is gained while not neglecting any source, an attractive feature in information fusion.

INPUT: k intervals Ia,b

OUTPUT: List of maximal coherent subsets Kj

List = ∅ ; j=1; k=∅;

Order by increasing value

{ai ,a=1,.....,k}U{bi ,i=1,.....,k}(in case of ties, ai is before bi);

Rename them {ci , i=1,.....2k} with type(i)=a if

Ci=am and type(i)=b if ci=bm;

for i=1,..... , 2k-1 do

if type(i) = a then

Add source m to K s.t. ci=am;

If type(i+1)=b then

Kj=K;

Add Kj

to list;

J=j+1;

Else

Remove source m from K s.t. ci=bm;

4. Feedback Assesment

Input: User satisfication

Output: Feedback for corresponding data

Step 1: System shows or gives reliable data to user as per query

Step 2: Calculate user satisfication

The formula and variable definitions for relevance feedback is as follows:

$$\overrightarrow{Q_m} = \left(a \cdot \overrightarrow{Q_o}\right) + \left(b \cdot \frac{1}{|D_r|} \cdot \sum_{\overrightarrow{D_j} \in D_r} \overrightarrow{D_j}\right) - \left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\overrightarrow{D_k} \in D_{nr}} \overrightarrow{D_k}\right)$$

Where

$\overrightarrow{Q_m}$	Modified Query Vector
$\overrightarrow{Q_o}$	Original Query Vector
$\overrightarrow{D_j}$	Related Document Vector
$\overrightarrow{D_k}$	Non-Related Document Vector

a	Original Query Weight
b	Related Documents Weight
c	Non-Related Documents Weight
D_r	Set of Related Documents
D_{nr}	Set of Non-Related Documents

Step 3: Store feedback in database
Step 4: If user search for same query, then system will shows feedback result which we stored in database.

IV. RESULT EVALUATION

We implemented reliable data warehouse system and collected various empirical performance numbers, which verify the linear (in the number of "entries" as described below) time and space costs of the various operations and data structures.

Implementation and Experimental Setup

We implemented reliable datawarehouse system using c#.net, Asp.net and SQL server 2008.
We evaluated our system on a 2.2 GHz Intel dual core with 4 GB of RAM.
For each experiment relating to protocol performance, we report the average of 10 runs. The evaluation of data structure sizes is the byte count of the marshaled data structures that would be sent over the network.

Experimental Results

Fig. 2 shows the size of the various data structures. The X-axis represents the number of entries in each data structure—of reliability of data and y-axis represent number of urls which are reliable or non-reliable.

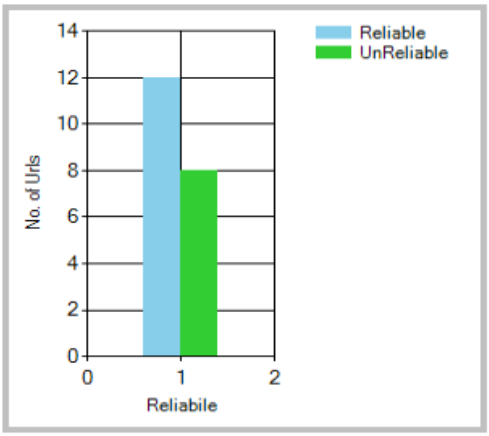


Fig 2: Reliability of System

We also check redundancy of web documents before reliability assessment, below graph in fig 3 shows redundancy of data.

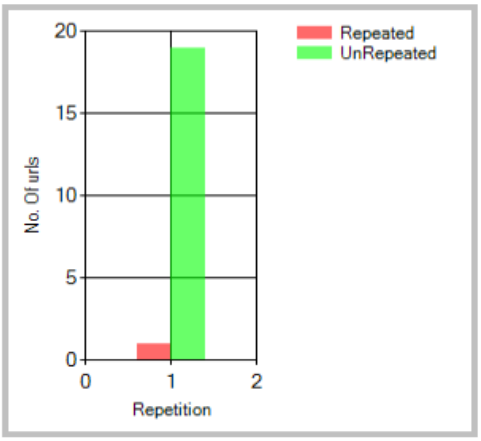
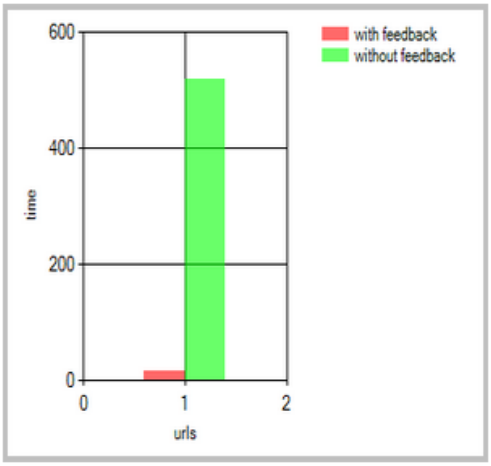
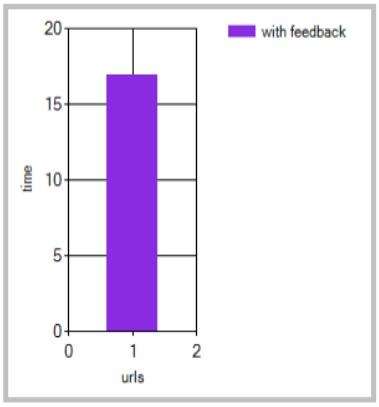


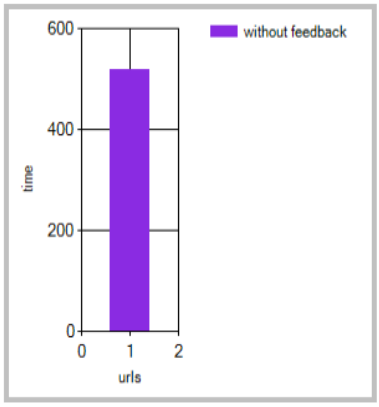
Fig 3: Redundancy of system

V. GRAPH RELEVANT TO FEEDBACK AND TIME

This graph clearly shows the comparison between feedback and without feedback result of data searching. The time consumption of searching a reliable and relevant data is much less with a user feedback.







VI. CONCLUSION

A generic method is proposed to evaluate the reliability of data automatically retrieved from the web. The method evaluates data reliability from a set of common sense (and general) criteria. It relies on the use of basic probabilistic assignments and of induced belief functions, since they offer a good compromise between flexibility and computational tractability. To handle conflicting information while keeping a maximal amount of it, the information merges excluding repeatitions of data information.

Finally, reliability evaluations and ordering of data tables are achieved by using lower/upper expectations, allowing us to reflect uncertainty in the evaluation. The results displayed to end users are an ordered list of tables.

Also, we have added user feedback as an additional (and parallel) source of information about reliability or relevance, as it is done in web applications.

The user feedback has helped to save time in searching the desired information asked by the user providing a reliable and relevant data.

As future works, we see two main possible evolutions:

- Reliability assessments presented in an ordered list can be achieved.
- Reliability can be evaluated with adding a constraint, completeness of the data.

REFERENCES

[1] Sebastien Destecke, Patrice Buche, and Brigitte Charnomordic, "Evaluating Data Reliability: an Evidential answer with Application to a web-enabled Data warehouse", *IEEE Trans. On knowledge and data engineering*, vol. 25, no.1, Jan 2013.

[2] P. Buche, J. Dibie-Barthe´lemy, and H. Chebil, "Flexible Sparql Querying of Web Data Tables Driven by an Ontology," *Proc. Eighth Int'l Conf. Flexible Query Answering Systems (FQAS)*, pp. 345- 357, 2009.

[3] G. Hignette, P. Buche, J. Dibie-Barthe´lemy, and O. Haemmerle´, "Fuzzy Annotation of Web Data Tables Driven by a Domain Ontology," *Proc. Sixth European Semantic Web Conf. The Semantic Web: Research and Applications (ESWC)*, pp. 638-653, 2009.

[4] D. Mercier, B. Quost, and T. Denoeux, "Refined Modeling of Sensor Reliability in the Bellief Function Framework Using Contextual Discounting," *Information Fusion*, vol. 9, pp. 246-258, 2008.

[5] R. Cooke, Experts in Uncertainty. Oxford Univ. Press, 1991.

[6] S. Sandri, D. Dubois, and H. Kalfsbeek, "Elicitation, Assessment and Pooling of Expert Judgments Using Possibility Theory," *IEEE Trans. Fuzzy Systems*, vol. 3, no. 3, pp. 313-335, Aug. 1995.

[7] F. Delmotte and P. Borne, "Modeling of Reliability with Possibility Theory," *IEEE Trans. Systems, Man, and Cybernetics A*, vol. 28, no. 1, pp. 78-88, 1998.

[8] F. Pichon, D. Dubois, and T. Denoeux, "Relevance and Truthfulness in Information Correction and Fusion," *Int'l J. Approximate Reasoning*, vol. 53, pp. 159-175, 2011.
[9] J. Sabater and S. Sierra, "Review on Computational Trust and Reputation Models," *Artificial Intelligence Rev.*, vol. 24, pp. 33-60, 2005.
[10] J. Golbeck and J. Hendler, "Inferring Reputation on the Semantic Web," *Proc. 13th Int'l World Wide Web Conf., 2004.*
[11] Y. Gil and D. Artz, "Towards Content Trust of Web Resources," Proc. 15th Int'l Conf. World Wide Web (WWW '06), pp. 565-574, 2006.
[12] K. Quinn, D. Lewis, D. O'Sullivan, and V. Wade, "An Analysis of Accuracy Experiments Carried Out over a Multi-Faceted Model of Trust," Int'l J. Information Security, vol. 8, pp. 103-119, 2009.
[13] A. Denguir-Rekik, J. Montmain, and G. Mauris, "A Possibilistic- Valued Multi-Criteria Decision-Making Support for Marketing Activities in E-Commerce: Feedback Based Diagnosis System," European J. Operational Research, vol. 195, no. 3, pp. 876-888, 2009.
[14] I.N. Chengalur-Smith, D.P. Ballou, and H.L. Pazer, "The Impact of Data Quality Information on Decision Making: An Exploratory Analysis," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 853-864, Nov./Dec. 1999.
[15] L. Zadeh, "The Concept of a Linguistic Variable and Its Application to Approximate Reasoning-i," Information Sciences, vol. 8, pp. 199-249, 1975.
[16] L. Zadeh, "Fuzzy Sets as a Basis for a Theory of Possibility," *Fuzzy Sets and Systems*, vol. 1, pp. 3-28, 1978.
[17] S. Ramchurn, D. Huynh, and N. Jennings, "Trust in Multi-Agent Systems," *The Knowledge Eng. Rev*., vol. 19, pp. 1-25, 2004.

