Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

Exploring Relevance as Truth Criterion on the Web and Classifying Claims in Belief Levels

Fairouz Zendaoui*

*Corresponding author, Laboratoire de la Communication dans les Systèmes Informatiques, Ecole Nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie. http://www.esi.dz. ORCID: 0000-0002-8270-1614. E-mail: f_zendaoui@esi.dz

Walid Khaled Hidouci

Laboratoire de la Communication dans les Systèmes Informatiques, Ecole Nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie. http://www.esi.dz. ORCID: 0000-0002-8290-1093. E-mail: w_hidouci@esi.dz

Abstract

The Web has become the most important information source for most of us. Unfortunately, there is no guarantee for the correctness of information on the Web. Moreover, different websites often provide conflicting information on a subject. Several truth discovery methods have been proposed for various scenarios, and they have been successfully applied in diverse application domains. In this paper, we have attempted to answer the question whether the truth is relevant. We conducted an experimental study in which we analyzed and compared the results of two different truth discovery methods: Relevance-based sources ranking and Majority vote. We have found that the truth is not always held by the most relevant sources on the web. Sometimes the truth is given by the majority vote of the crowd. In addition, we have proposed a method of presenting the results of truth discovery with gradual degrees of belief. A method that allows to configure and target the desired level of trust.

Keywords: Truth discovery; Source ranking; Relevance; Uncertainty; Belief degree.

DOI: 10.22059/jitm.2020.75786

Introduction

In the era of information explosion, data have been woven into every aspect of our lives, and we are continuously generating data through a variety of channels, such as websites, social networks, blogs, discussion forums, crowdsourcing platforms, etc. In these scenarios, data, even describing the same object or event may conflict with each other. This is due to the imperfection quality of the information involving subjectivity, uncertainty, imprecision, ambiguity and incompleteness. The research question is how to find the most plausible value from the noisy information available on the web and how to get closer to the truth.

“Is the world-wide web always trustable?” Unfortunately, the answer is “no”. There is no guarantee for the correctness of information on the web. Even more, different websites often provide conflicting information. For example, Table 1 illustrates the claimed values provided by different websites about the location of the heritage site "Timgad" which is an ancient Berber-Roman city. The truth is that this site is located on the territory of the eponymous municipality of Timgad, in the province of Batna in the Aurès region, in the North-East of Algeria. This is a very simple example which shows conflicting values describing the same property which is the location for the same object which is the heritage site named "Timgad". Imagine, then, the huge amount of information on the web and the rate of conflict that is generated by the multi-source context.

Several truth discovery methods have been proposed for various scenarios, and they have been successfully applied in diverse application domains (see section 2). They make different assumptions about aspects of truth discovery like input data, source reliability, identified truths, object, claimed value and output.

Table 1. Conflicting claimed values regarding the location of Heritage Site “Timgad”

Source	Claimed Value	Source	Claimed Value
UNESCO	Algeria	yelp	Ireland
worldatlas	Tunisia	Tripadvisor	Algeria
wikipedia	Algeria	bizdb	London
scribd	Algeria		

On the other hand, there have been many studies on ranking web pages according to relevance based among others on domain authority, page authority and citation flow (Roa-Valverde & Sicilia, 2014). But does relevance or importance of websites lead to accuracy of information? We found that none of the previous truth discovery works has considered sources ranking in their process, knowing that relevance is the main performance metric targeted by the majority of search engines. Our second observation concerns the presentation of the results of this process. All truth discovery methods present the inferred values with the same level of confidence, that is, the same degree of belief. This is not appropriate because each inferred value is associated with it a trust mass measured by the method used and differs from the masses of the other inferred values.

In this paper we explore the influence of the *website relevance* also named *website importance* on the results of truth discovery process. Our research question is: are true values held by the most important or popular sources? For this, we conducted an experimental study on real data from the web relative to the location of world heritage sites. For the first time, we implemented a sources-ranking based method for truth discovery as well as the voting classic method. We analyzed and compared the results of the two methods, and we even tried to merge these methods.

We also tackled another point which relates to the restitution of inferred values. Since each method of truth discovery follows an inference algorithm to deduce the truth and since the results are always qualified as uncertain, we propose a method of gradual qualification of inferred values. We based on degrees of belief interpreting the level of trust in these values. This vision is founded on previous works that we have done and in which we have proposed a multilayer representation of historical information distinguishing *information*, *source* and *belief*, we differentiated three levels of granularity (*attribute*, *object*, *relation*) to express belief, and we have identified 11 gradual degrees of certainty (Zendaoui & Hidouci, 2019a, 2019b).

This paper is organized as follows. In section 2, we present works which approached the truth discovery according to various aspects relating to the information sources. In section 3, we expose the experimental methodology by first defining the problem, presenting the data collection, establishing the truth discovery algorithms, and identifying the quality measures. Section 4 is devoted to the presentation, analysis and interpretation of the obtained experiments results. We conclude the paper in section 5.

Related Works

Several methods have been proposed for truth discovery among conflicting information from the Web. This problem of resolving conflicts from multiple sources has been extensively studied. Some methods consider the *dependencies between sources* to find the truths like the algorithm proposed by Dong, et al. (2009) which exploits Bayesian analysis to find the dependencies. This algorithm was extended by computing the distribution of the false values from the records while the first one assumes that it is uniform (Dong, et al., 2012). Other truth discovery approaches infer trustworthy information from conflicting multi-source data by taking *source reliability* into consideration. To estimate such source reliability, existing approaches make the assumption that each source's reliability is consistent over all the claims it makes. This assumption is made by Yin, et al. (2008) and further adopted in several works (Dong, et al., 2009; Li, et al., 2014; Li, et al., 2015; Pasternack & Pasternack, 2010; Yin & Tan, 2011; Zhao, et al., 2012; Li, et al., 2016). Gurjar and Moon (2016) presented a survey in which they focused on providing a comprehensive overview of truth discovery methods, and summarizing them from different aspects to offer some guidelines on how to apply these approaches in applications domains. Al-Araji, et al. (2019) proposed an algorithm which compares the *interaction found between sources* and the information hosted on these sources to discover the true facts from the vast quantity of contradictory information on Quran's explanation, provided on different internet websites. Recent research uses *crowdsourcing* to refine the result of the truth discovery since it is an efficient way to utilize human intelligence with low cost as Jung, et al. (2019) who combined a hierarchical truth inference with task assignment for developing a crowdsourced truth discovery algorithm.

In this paper, we approach another aspect in the discovery of the truth which is the *relevance or importance of sources*. This aspect has not been considered in any truth discovery research, while all internet users systematically use search engines, and the latter provide results by considering this aspect. So, we explore the accuracy level reached when we consider the relevance of the source.

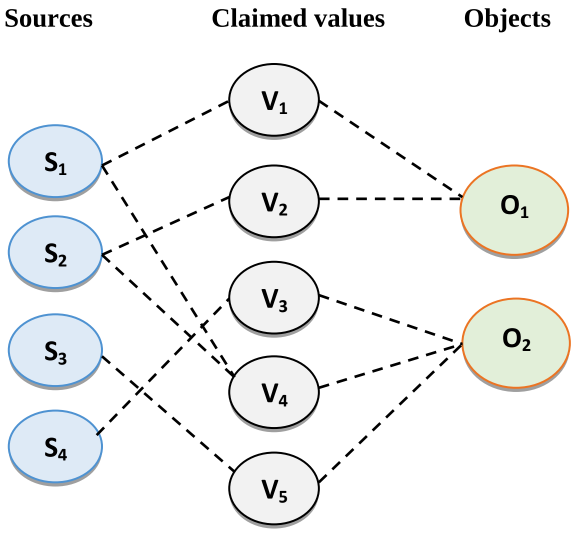
Another observation is that all the works dealing with the truth discovery have presented the inferred values with the same level of confidence, while the methods used, each time, provide scores that differ from one value to another. We see that it is more appropriate to reflect this aspect by distinguishing *gradual degrees of belief* distributed according to the scores obtained.

Experimentation Methodology

Problem Definition

The input of our experimentation is a large number of facts about properties of a certain type of objects. The facts are provided by many websites. There are usually multiple conflicting facts from different websites for each object, and our goal is to identify the true fact among them. Figure 1 shows a mini example dataset. Each website provides at most one fact for a given object.

Figure 1. Inputs of truth discovery problem



To make the following description clear and consistent, in this section, we introduce some definitions and notations that are used in this paper.

- An object **o** is a thing of interest, a source **s** describes the place where the information about objects can be collected from, and a value ***v*** represents the information provided by source **s** about object **o**.
- A claim **c**, also known as a record, is a 3-tuple that consists of an object, a source, and its provided value.
- The inferred truth for an object **v*** is the information selected as the most trustworthy one from all possible candidate values about this object.
- Each inferred value is associated with a score between 0 and 1 representing the mass of confidence granted to this value and which will be interpreted as the corresponding belief degree.
- Each inferred value is assigned to a specific belief level according to its confidence mass. Table 2 presents the different levels of belief distributed according to the confidence masses.

Table 2. Belief levels

Belief Level	Confidence Mass Interval
Certain	1
Presumed	[0.6, 1[
Doubtful	[0.45, 0.6[
Probable	]0, 0.45[
Completely uncertain	0

Based on these definitions, let's formally define the truth discovery task as following.

Definition 1. For a set of objects **O** that we are interested in, related information can be collected from a set of sources **S**. Our goal is to find the truth **v*** for each object $\mathbf{o} \in \mathbf{O}$ and its confidence mass by resolving the conflicts among the information from different sources and assuming that there is only one true fact for a property of an object.

Data Collection

Our experimentation uses a real-life dataset of the locations of World Heritage Sites. It was collected by Jung, et al. (2019) in order to experiment their proposed probabilistic model to utilize the hierarchical structures and an inference algorithm to find the truths. This dataset is publicly available at *http://kdd.snu.ac.kr/home/datasets/tdh.php*. It contains the locations of 785 Cultural and Natural World Heritage Sites queried about with Bing Search API and resulting in 4,424 claimed values from 1,577 distinct websites. We use as ground truths the values provided by UNESCO World Heritage Centre, available at *http://whc.unesco.org*. Regarding the ranking of sources by relevance, we retrieved the Page Rank of each source by querying the site *https://www.prepostseo.com/google-pagerank-checker*. The latter provides a Google Page Rank Checker which combines a wide range of metrics to calculate Page Rank.

Truth Discovery Algorithm

We have built our experimentation on two different methods of truth discovery. The first is the classic majority vote method and the second method is based on the ranking of sources by relevance. We are the first to have explored the second method.

The algorithm 1 represents the main approach followed to infer the truths. We implemented it for each of the two truth discovery methods that we explored. This algorithm has two main stages: The calculation of the confidence masses of the claimed values and the selection of the inferred values. We implemented different variants of the same algorithm according to the method of discovering truth considered. The implementation of this algorithm differs from one method to another in the step of calculating confidence masses.

Algorithm 1: General procedure of truth discovery

```
Input: Claimed values from multiple sources.
Output: Inferred values
{Calculation of confidence masses}
For each object o of the collection Do
    For each distinct instinct claimed value v relative to the object o Do
        Calculate the mass of confidence m of value v corresponding to the voting rate obtained for
        the value v of object o.
{Selection of inferred values}
For each object o of the collection Do
    Select the value having the highest confidence mass. This value is the inferred value for the object o.
```

We summarize below the calculation principle for each method tested.

A. **Majority Vote:** The principle of majority vote is to favor the values with the highest voting rate. So, the confidence mass of a given value is equal to this rate.

B. Relevance Source Ranking: In this method, we consider the ranking of sources by relevance. We implement two different variants for this same method:

B.1. Average Ranking: The belief mass of a value **v** corresponds to the average ranking of the claiming sources.

B.2. Maximum Ranking: The belief mass of a value **v** corresponds to the best ranking among the claiming sources rankings.

C. Combining Voting and Ranking: We also tried to combine majority vote and relevance ranking for discovering truth by considering separately each variant of ranking.

C.1. Merging Majority Vote and Average Ranking: The confidence mass of a given value in this case is equal to the average of the two masses of this same value obtained by calculating separately with the Majority Vote and Average Ranking.

C.2. Merging Majority Vote and Maximum Ranking: The confidence mass of a given value is equal to the average of the two masses of this same value obtained by calculating separately with the Majority Vote and Maximum Ranking.

Quality Measure

The inputs of the problem are the claimed values from multiple sources and the ground truths. The conflict resolution methods that we have implemented are conducted in an unsupervised manner in the sense that the ground truths will only be used in evaluation. We consider an inferred value to be true if it is a substring of the ground truth.

We analyzed and compared the results of the different truth discovery methods based on the following metrics:

A. The Accuracy: We calculated the accuracy rate which is equal to the ratio of the number of true inferred values on the total number of inferred values. We tried to find out the truth rates achieved by the implemented methods.

B. The Average Confidence Mass: We calculated the average confidence mass of true inferred values and that of false positives. The goal is to find a correlation between the average belief mass and the accuracy rate and to answer the question of what is the minimum belief mass that could lead to truth.

C. The Distribution of Inferred Values on Belief Levels: We seek to exploit the belief levels that we have previously defined to classify the inferred values according to their belief masses and give a more precise illustration and interpretation of the results of the truth discovery process.

Experimental Results

In this section, we present the results of our experiments. First, we analyze and compare the results of the majority vote with the ranking of sources mainly in terms of accuracy. Then, we study the results obtained by merging the majority vote with the two variants of ranking (average and maximum rank). Lastly, we show how to classify the inferred values in belief levels according to their computed confidence masses, and this for each of the implemented methods.

Voting versus Ranking

Table 3 shows the results obtained for the majority vote method as well as the two variants of the ranking method. Figure 2 schematizes the accuracy with sectors.

The majority vote outperformed in terms of accuracy. 88% of the values inferred by this method are true. Consequently, its rate of false positives is 12%, it is the percentage of the inferred values that are false. We deduced from this that the majority vote has led to the truth more than the two variants of the ranking method.

On the other hand, the maximum ranking method showed an accuracy equal to 79% higher than that of the average ranking which gave a percentage of 68%. This result means that it is not the average ranking of claiming sources which increased the plausibility of the inferred value, the maximum ranking obtained just by a single claimed source reinforced its plausibility.

Table 3. Comparison results of majority vote and source ranking

Method	Majority Vote	Average Ranking	Maximum Ranking
Accuracy	**88%**	68%	79%
False positives rate	**12%**	32%	21%
Average confidence mass	0.66	0,72	0,87
False positives average confidence mass	0,45	0,77	0,84
True inferred value average confidence mass	0,69	0,70	0,88

Furthermore, with an average majority vote of 66% (average confidence mass) we reached 88% accuracy. More precisely, the average confidence mass of true inferred values is 0.69 while that of false positives is 0.45. Rather, it shows a significant and logical distribution. The higher the voting rate for a claimed value, the greater the level of plausibility granted to this value.

This relativity is not noticed for the two variants of ranking, in contrast, they showed very close average confidence masses between the true and false positive inferred values:

(0.77, 0.70) for the average ranking and (0.84, 0.88) for the maximum ranking. Here, relevance was not the factor directly affecting accuracy and other parameters have to be considered.

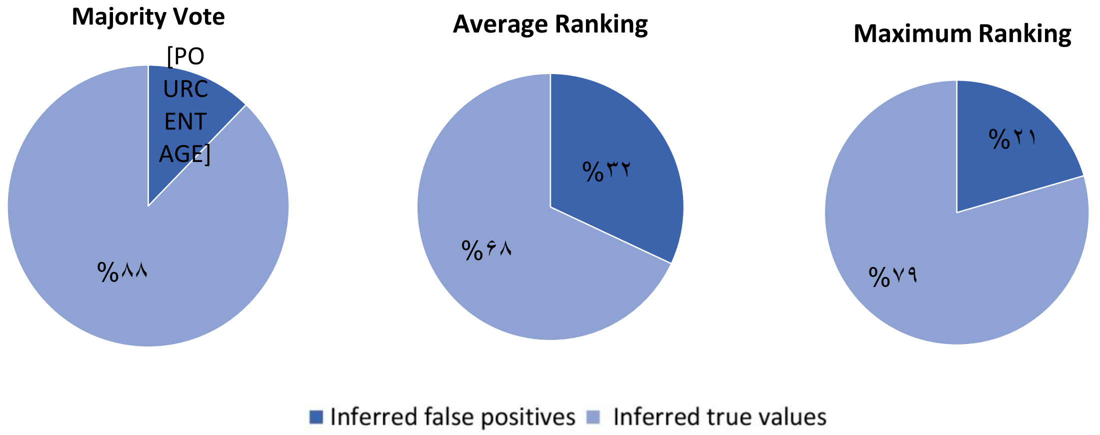


Figure 2. Accuracy of majority vote, average ranking and maximum ranking

Merging Voting and Ranking

Merging the vote with the ranking showed similar results with regard to accuracy which is 85%, although their results were not equal when we tested them separately (see table 4). Same for the rest of the measures relating to the average confidence mass, the two merges gave very close values. The majority vote has contributed considerably in improving the accuracy for both fusion variants. So, this reinforced our previous finds that the truth was affected by collective choices more than by the relevance sources quality.

Table 4. Results of merging vote and ranking

Method	Vote and Average Ranking	Vote and Maximum Ranking
Accuracy	**85%**	**85%**
False positives rate	**15%**	**15%**
Average confidence mass	0,65	0,61
False positives average confidence mass	0,55	0,51
True inferred value average confidence mass	0,66	0,62

Truth Discovery with Belief Degrees

In this part, we start from the principle that if we use a given truth discovery method it is because we believe that it gives us the truth. Therefore, the scores or masses of confidence of inferred values can express the degrees of belief granted to them. When we previously defined the problem of discovering truth, we identified five levels of belief that correspond to gradual degrees of uncertainty could be associated with beliefs. According to this vision, we illustrate how to classify the inferred values according to these belief levels. In this way we not only infer the value, but we qualify it with a belief level which makes it possible to distinguish between the inferred values and not to consider them as being equal in terms of belief. An inferred value with a belief mass of 0.9 belongs to the level "Presumed" while an inferred value with a belief mass of 0.3 belongs to the level "Probable".

Table 5 presents the distribution of the inferred values on the belief levels based on the inclusion of their belief masses in the belief intervals. This is also illustrated by sectors in figure 3.

Table 5. Classification of inferred values according to belief degrees

Belief Degree	Majority Vote	Average Ranking	Maximum Ranking
Certain	17%	2%	21%
Presumed	**44%**	**79%**	**76%**
Doubtful	20%	16%	3%
Probable	19%	3%	1%
Completely uncertain	0%	0%	0%

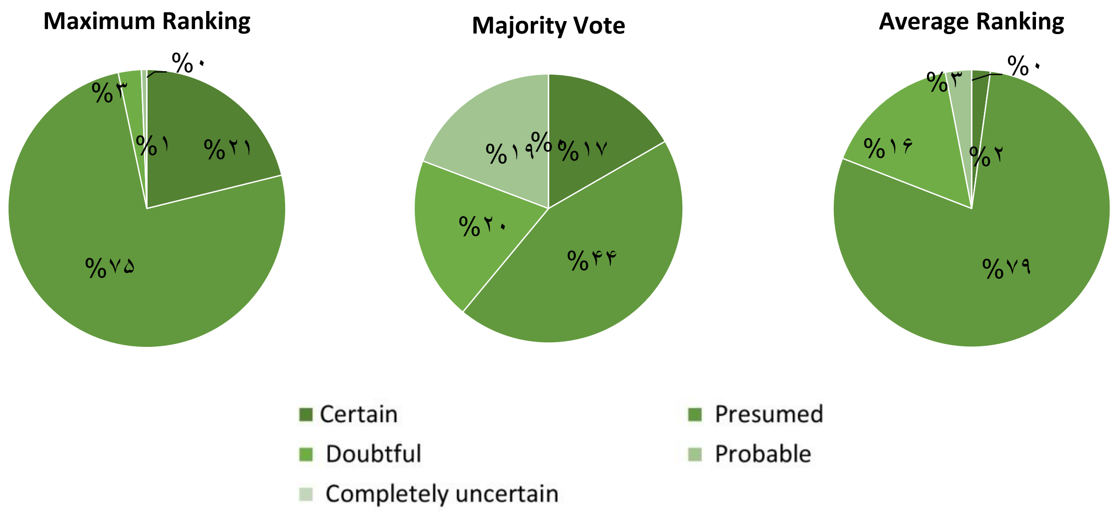


Figure 3. Distribution of inferred values according to belief degrees

The results showed that the largest percentages of the values inferred either by the vote or by the two variants of the ranking are classified in the level "Presumed" which corresponds to the interval of belief [0.6, 1 [. The result that interested us most is that of the majority vote as it has already shown significant results in terms of accuracy. When we analyzed its results, we noticed that 61% of the inferred values are either presumed or certain.

This classification allowed us to configure and target the desired belief level based on the corresponding degrees of belief. We could set the minimum belief degree necessary to accept an inferred value. For example, to see only certain results, we retained values with a belief mass equal to 1. To tolerate presumption, we just had to keep values with a belief mass greater than 0.6 neglecting any doubtful, probable or uncertain values.

Discussion

From the results of our study, we deduce that the majority vote could lead to the truth more than the relevance of sources and that the crowd on the web holds truths not obvious to target by methods of research by relevance. This probably means that relevance did not play a dominant role in finding the truth and that other aspects affect the veracity of the claims.

Although voting has outperformed both ranking variants ranking, the maximum ranking gave more accuracy than average ranking. This means that the maximum ranking obtained just by a single source could reinforce the plausibility of its claims (which is valid at least for the dataset we studied).

On the other hand, the confidence masses of inferred values help to quantify beliefs granted to them. We illustrated how to classify results of truth discovery process in five beliefs levels (certain, presumed, doubtful, probable and uncertain) according to the inclusion of each belief mass in a given belief interval.

This approach of classification allows to configure the belief level targeted by any truth discovery method, either restrict the results to presumed or certain values or release them by tolerating doubtful, probable or even completely uncertain values. In this way, we can set the minimum belief degree desired to accept an inferred value.

Conclusion

Through our study, we deduced that the majority vote could lead to the truth more than the relevance of sources and that the crowd on the web holds truths not obvious to target by methods of research by relevance. Relevance strengthens accuracy, but it is not the predominant criterion for discovering the truth. Also, we illustrated how to qualify and classify the inferred values according to belief levels. This allows configuring the minimum belief necessary to accept an inferred value. In future work, we will consider other criteria in truth discovery such as the age of the claim as well as the type of sources.

References

Al-Araji, Z. J., Ahmad, S. S. S., Al-Lamy, H. A., Al-Salihi, M. W., Al-Shami, S. A., Mohammed, H., & Al-Taweel, M. H. (2019). Truth Discovery Using the TrustChecker Algorithm on Online Quran Tafseer. *In Intelligent and Interactive Computing* (pp. 71-80). Springer, Singapore.

Dong, X. L., Berti-Equille, L., & Srivastava, D. (2009). Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment*, *2*(1), 550-561.

Dong, X. L., Saha, B., & Srivastava, D. (2012). Less is more: Selecting sources wisely for integration. *Proceedings of the VLDB Endowment*, *6*(2), 37-48.

Gurjar, K., & Moon, Y. S. (2016). Comparative Study of Evaluating the Trustworthiness of Data Based on Data Provenance. *Journal of Information Processing Systems*, *12*(2).

Jung, W., Kim, Y., & Shim, K. (2019). Crowdsourced Truth Discovery in the Presence of Hierarchies for Knowledge Fusion. *arXiv preprint arXiv:1904.10217*.

Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., & Han, J. (2014, June). Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. *In Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 1187-1198).

Li, X., Dong, X. L., Lyons, K., Meng, W., & Srivastava, D. (2015). Truth finding on the deep web: Is the problem solved? *arXiv preprint arXiv:1503.00303*.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., ... & Han, J. (2016). A survey on truth discovery. *ACM Sigkdd Explorations Newsletter*, *17*(2), 1-16.

Pasternack, J., & Roth, D. (2010, August). Knowing what to believe (when you already know something). *In Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 877-885). Association for Computational Linguistics.

Roa-Valverde, A. J., & Sicilia, M. A. (2014). A survey of approaches for ranking on the web of data. *Information Retrieval*, *17*(4), 295-325.

Yin, X., & Tan, W. (2011, March). Semi-supervised truth discovery. *In Proceedings of the 20th international conference on World wide web* (pp. 217-226).

Yin, X., Han, J., & Philip, S. Y. (2008). Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, *20*(6), 796-808.

Zendaoui, F., & Hidouci, W. K. (2019a). Considering Uncertainty in Modeling Historical Knowledge. *ISeCure*, *11*(3).

Zendaoui, F., & Hidouci, W. K. (2019b). Multi-version representation of historical event. *Periodicals of Engineering and Natural Sciences*, *7*(1), 141-147.

Zhao, B., Rubinstein, B. I., Gemmell, J., & Han, J. (2012). A bayesian approach to discovering truth from conflicting sources for data integration. *arXiv preprint arXiv:1203.0058*.

Bibliographic information of this paper for citing:

Zendaoui, F, & Hidouci, W.Kh. (2020). Exploring Relevance as Truth Criterion on the Web and Classifying Claims in Belief Levels. *Journal of Information Technology Management*, 12(2), 1-12.