Journal of Systems Engineering and Electronics
Vol. 28, No. 3, June 2017, pp.617 – 626

EXCLUSIVELY AVAILABLE VIA
IEEE Xplore® DIGITAL LIBRARY

Truth finder algorithm based on entity attributes for data conflict solution

Xiaolong Xu[1,2,*], Xinxin Liu[1], Xiaoxiao Liu[3], and Yanfei Sun[4]

1. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. State Key Laboratory of Information Security, Chinese Academy of Sciences, Beijing 100093, China;
3. Institute of Big Data Research at Yancheng, Nanjing University of Posts and Telecommunications, Yancheng 224005, China;
4. Science and Technology Department, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: The Internet now is a large-scale platform with big data. Finding truth from a huge dataset has attracted extensive attention, which can maintain the quality of data collected by users and provide users with accurate and efficient data. However, current truth finder algorithms are unsatisfying, because of their low accuracy and complication. This paper proposes a truth finder algorithm based on entity attributes (TFAEA). Based on the iterative computation of source reliability and fact accuracy, TFAEA considers the interactive degree among facts and the degree of dependence among sources, to simplify the typical truth finder algorithms. In order to improve the accuracy of them, TFAEA combines the one-way text similarity and the factual conflict to calculate the mutual support degree among facts. Furthermore, TFAEA utilizes the symmetric saturation of data sources to calculate the degree of dependence among sources. The experimental results show that TFAEA is not only more stable, but also more accurate than the typical truth finder algorithms.

Keywords: truth finder, data reliability, entity attribute, data conflict.

DOI: 10.21629/JSEE.2017.03.21

1. Introduction

The rapid development of mobile networks and the Internet has brought the significant growth of Web traffic. The whole Internet has already become a large-scale platform with big data, which is still growing rapidly. Web data have clearly become an important source for users to obtain the needed information. Even though the Internet makes people's lives and works more convenient, the problem of data quality provided online has also become increasingly more prominent. The network medias, such as blogs, make it easier for users to publish and spread information. One entity may be described by multiple records in different data sets. All redundant records are not always the same due to expressive errors and different schemata of data sources [1]. As a result, people might get a lot of false or outdated information.

However, due to the inherent characteristics of the Internet, each website on the Internet and its databases have become subsets of the Web dataset, and these subsets usually only serve in a specific department or area, without considering data sharing or collaboration. Therefore, there are a lot of conflicting data in the Web dataset [2]. Conflicting information of one entity from different websites is especially common. For example, the information of flights, train routes and hotel accommodations offered by different websites may be conflicting. For example, five famous e-book websites provide conflicting information of the book *The Concept of Database System* (*Version 6*). Among them, China-pub, Jingdong online-mall and Huazhang publisher provide a complete list of authors, while the information of authors provided by Amazon and Dangdang is false. The numbers of page provided by different websites are also different. In a word, how to find the correct information among conflicting information has become a serious problem that needs to be solved.

In the same dataset, there is more than one type of information that can be used to describe the same entity attribute, and some may be consistent, while some may be contradictory and result in confliction. This is mainly caused by the existence of redundant information, false information, outdated information or duplicated data, etc. Given a set of data sources as well as a set of facts, the truth finder means to find a correct description of an entity [3]. The simple and intuitive solution is using the voting scheme. Each data source separately votes on the fact, and the accuracy of each fact is finally determined accord-

Manuscript received August 09, 2016.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (61472192) and the Scientific and Technological Support Project (Society) of Jiangsu Province (BE2016776).

ing to the result of vote. However, the voting mechanism treates each data source equally, without taking the differences among data sources into account. Therefore, the voting results are usually different from the reality.

In order to solve the data conflict problem, researchers have conducted a series of researches on various factors affecting the truth finder judgment. Yin et al. [3] proposed a truth finder algorithm, using the iterative mechanism similar to the authority-hub [4,5] to jointly derive the reliability of a data source and the accuracies of facts provided by the data source. Dong et al. [6,7] used the Bayes rules to infer the dependencies among data sources. Kao et al. [8] proposed the iteration vote (IVote) algorithm, the iteration-reputation vote (IRVote) algorithm and the iteration-reputation-duplication vote (IRDVote) algorithm. In addition, there are algorithms based on information retrieval [9], link analysis [10], semi-supervised learning (SSL) [11 – 13] and other methods [14] to improve the accuracy and efficiency of truth finding.

There are many factors affecting the performance of truth finder algorithms, such as the reliabilities of data sources, the accuracies of the facts provided by data sources, the mutual support degrees among the facts given to the same entity attribute, the authorities of data sources, and the interdependences among data sources. However, the degrees of these factors affecting the truth finder algorithms are different, and many truth finder algorithms consider these factors without combining specific problem models. The current truth finder algorithms generally have the following problems:

(i) By only considering a single factor, it will cause the algorithm to be inaccurate; or by considering too many unimportant factors, it will result in complexity and difficulty.

(ii) For the calculation of the mutual support degree among facts, the string similarity based on editing distance between the text is always utilized to replace the mutual support degree among the facts, which seriously affects the accuracy of truth finding.

(iii) Most truth finder algorithms treat data sources equally, but the characteristics of each data source determine its different support degrees to the facts provided by itself.

(iv) Most truth finder algorithms only provide factors, without the detailed analysis of calculation approaches of impact factors.

In order to address the above problems, we propose a novel truth finder algorithm based on entity attributes (TFAEA). The main contributions of this paper are as follows.

(i) We build a truth finder model. Based on the existing data sources' reliabilities and facts' accuracies iterative calculation system, the following two factors are considered in the model of TFAEA: the mutual support degree among the facts given to the same entity attribute and the interdependences among data sources. Efforts are also made to simplify the truth finder algorithm and enhance its accuracy.

(ii) The unidirectional text similarity and the degree of fact conflict are combined to calculate the mutual support degree among the facts. The data source symmetric saturation method is also proposed to calculate the interdependence degree among data sources. It has stronger adaptation ability and accuracy in logogram, abbreviation, haplography, more writing, reverse order and other complex situations, which can also enhance the accuracy of truth finder.

2. Related works

In order to eliminate data conflicts and find the truth for data entity, many researchers have conducted a serious of works to improve truth finder algorithms. As we mentioned above, typical algorithms include truth finder algorithms based on voting, link analysis, the Bayes rules, and the expectation maximization (EM) model, etc.

Kao et al. [8] proposed a series of truth finder algorithms based on voting, including IVote, IRVote and IRDVote. IVote adopted the probability voting method for the iterative calculation, and the description with the highest vote was chosen as the final result. IRVote further considered the authority of data sources, namely the voting proportion of data sources, and the more authoritative the data source was, the greater weight it had in the voting process. IRDVote utilized the Bayesian approach to calculate copies among data sources.

The premise of the truth finder algorithm based on the link analysis is that the reliability of a data source has a certain impact on the accuracy of information judgment [3]. The reliability of a data source and the accuracy of information are interactive, which inspires us to use the iterative algorithm to calculate the reliabilities of data sources and the accuracy of information. Different from the voting-based algorithm, the truth finder algorithm calculates the reliability of a data source based on the link analysis, and considers the accuracy of the information provided by the data source. By combining the two processes, a complete truth finder algorithm is ultimately formed based on the iterative algorithm.

Based on the relationship among data sources and entity descriptions, the truth finder algorithm conducts repeated iterations until the obtained description accuracy does not change anymore, or the change is within a specified range. The truth finder algorithm relies on the principle that "the

correct descriptions provided by different data sources for the same entity attribute are consistent, while the false descriptions they provide have different forms". Therefore, the more correct descriptions a data source provides, the more reliable it is. In turn, the greater reliability a data source has, the more accurate the description. When the truth finder algorithm determines the reliability of a data source, it does not rely on the number of descriptions it provides, but depends on the accuracy of each description. More important, the truth finder algorithm also considers the impact of the mutual support degree on the accuracy of each description. However, the truth finder algorithm calculates the mutual support degree among descriptions by only using the string similarity based on editing distance instead. The accuracy of the truth finder algorithm for complex information is even more disappointing. And the truth finder algorithm does not consider the impact of the copy relationship among data sources on the accuracy of each description either.

Through repeated iterations used in a common link analysis method, it can make each node achieve a stable credibility score. The description with the highest score from all the descriptions of each entity is chosen as the true value of the entity. It is clear that the entity description with the highest score is most like the correct description of the entity, while the data source with the highest score is also most like the data source which accurately describes the entity. However, the score of an entity description cannot represent the probability whether it is correct or not, and the score of a data source cannot describe the probability whether it provides a correct value. To solve this problem, the Bayes rules are used to solve this truth finder problem [6,7]. The correction probability of each entity description is provided quantitatively. The truth finder algorithm based on the Bayes rules makes it possible to quantitatively evaluate the qualities of data sources and entity descriptions. The algorithm requires that the input data sources are independent of each other. However, in the real environment, the data sources have the relationships with each other sometimes, which affects the performance of the algorithm.

Some other researchers try to use statistical theories and probability models to solve the truth finder problem, and get many reasonable and efficient truth finder results. One classic example is the truth finder algorithm based on EM proposed by Dong et al. [15]. The EM-based truth finder algorithm consists of two steps: first, it calculates the logarithmic expectation of the global observation probability under the condition of a desired variable; second, it calculates the relation equation between the variable of the current round and the variable of the last round when the first-step reaches the maximum value. The algorithm can further improve the accuracy of truth finding. However, it does not take the relationship between data sources into account, either.

3. Proposed truth finder algorithm for data conflict solution

3.1 Model

Truth finder means organically combining different sources, formats and characteristics of data in logic, so as to provide the accurate data value for users. The truth finder mechanism searches or receives data provided by different data sources at first; then, it goes through four stages of processing: pattern matching, collision detection, truth finding and data fusion; finally, it outputs the correct and complete data to the main storage system for users. The truth finder model is shown in Fig. 1, where s_1 to s_4 are the four data sources, and the main DB represents the main storage system.

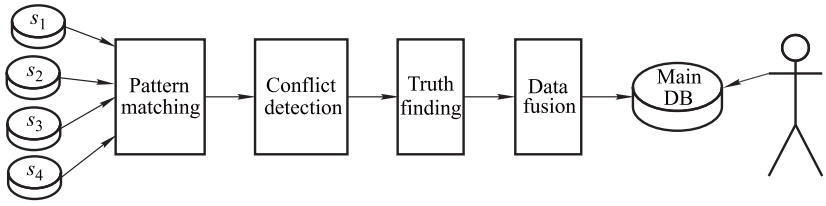


Fig. 1 Truth finder model

The four stages of the truth finder mechanism are described as follows:

Stage 1 Pattern matching. In a distributed storage system, each subsystem is allowed to operate the local data independently, which may cause each subsystem to provide different results of the same data. During the truth finding process, each subsystem is likely to be inconsistent on the storage format, naming rules and expressions for the same data. The data provided by these subsystems needs to be processed uniformly first, that is, to provide the mapping mechanism from the subsystem data to the main system data.

Stage 2 Conflict detection. The conflict detection strategy is utilized to determine whether the data provided by

subsystems conflict or not.

Stage 3 Truth finding. If the conflicting data is detected, i.e., different data sources provide different data for the same entity, the truth finder mechanism will call the truth finder algorithm to find the correct data value from these conflict data.

Stage 4 Data fusion. The data fusion of all right data values is arranged, and the final true value is provided to users.

3.2 Concepts and definitions

Data conflict is mainly manifested in that each data source provides different data for the same entity attribute. Some of these values can correctly reflect the objective fact, while some cannot. The following are the related concepts.

Data source A data source may be a database, a website, or a terminal, etc. A data source is denoted as a set S, namely $S = \{s_1, s_2, s_3, \ldots, s_n\}$, and s_i $(1 \leqslant i \leqslant n)$ represents the ith data source.

Entity Entity refers to a substance that actually exists in the real world, and it is characterized by a number of attributes, such as people, book, and car. An entity set is denoted as E, namely $E = \{e_1, e_2, e_3, \ldots, e_n\}$, and e_i $(1 \leqslant i \leqslant n)$ represents the ith entity.

Entity attribute An entity attribute is used to describe an entity, such as author of book, and color of car. An entity attribute set is denoted as EA, namely $EA = \{ea_1, ea_2, ea_3, \ldots, ea_n\}$, and ea_i $(1 \leqslant i \leqslant n)$ represents the ith property of an entity.

Fact For an entity attribute, it represents a description provided by a data source. A fact set is denoted as F, namely $F = \{f_1, f_2, f_3, \ldots, f_n\}$, and f_i $(1 \leqslant i \leqslant n)$ represents the ith fact.

True value If facts correctly describe the properties of an entity, the value of the entity is denoted as v.

In the truth finder model, each data source has provided a large number of facts, especially for the same entity attribute. However, due to the complexity and independence of data sources, some of these facts are true values, while some are not, leading to data conflict. The relationships of data sources, facts, entity attributes, and entities are shown in Fig. 2. For example, s_1 and s_2 provide two different facts f_1 and f_2 for the entity attribute ea_1, and these two facts will bring data conflicts. The truth finder algorithm is used to find the fact from f_1 and f_2 that can correctly describe the attributes of entities, i.e. the true value.

The definitions related to the truth finder algorithm are presented as follows.

Data conflict Different data sources provide different facts about the same entity attribute.

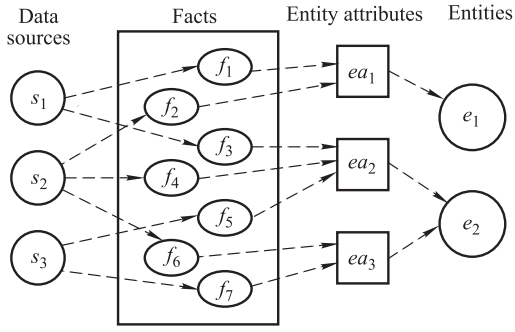


Fig. 2 Relationships between data sources, facts, entity attributes, and entities

Data conflict solution For an entity attribute $ea \leqslant EA$, the corresponding true value f_i $(f_i \leqslant F_{ea})$ is selected from the fact set $F_{ea} = \{f_1, f_2, f_3, \ldots, f_n\}$ provided by each data source of $S = \{s_1, s_2, s_3, \ldots, s_n\}$.

Accuracy of a fact According to the current judgment accuracy condition, the probability of fact f determined as a true value is denoted as the probability $t(f)$.

Reliability of data source The reliability of a data source s is determined by the true value in the fact set F provided by s, denoted as the probability $t(s)$.

According to the analysis of real datasets, the truth finder model proposed in this paper is based on the following four principles:

Principle 1 For an entity attribute, if there is only one value, it will be taken as the true value.

Principle 2 The true values provided by different data sources for the same entity attribute are the same.

Principle 3 The probability that different data sources provide the same false value for the same entity attribute is very low.

Principle 4 A data source can provide no more than one value for an entity attribute, while can provide values for multiple entity attributes.

3.3 Algorithm

In order to solve the problem of data conflict, we propose TFAEA.

(i) Reliability of data source

The reliability of a data source is closely related to the facts it provides. We use the average accuracies of all facts provided by the data source to represent the reliability of the data source.

$$t(s) = \frac{1}{m}\sum_{i=1}^{m} t(f_i) \tag{1}$$

where $f_i \in F(s)$, and $F(s)$ represents the set of fact provided by the data source s; m is the total number of facts provided by s.

(ii) Accuracy of fact

Assume that under the initial condition, both s_1 and s_2 provide f_1 for ea. If f_1 is not a true value, s_1 and s_2 are likely to be unreliable data sources. Assume that each data source is independent of others, so the probability that s_1 and s_2 are not reliable at the same time is $(1-t(s_1))(1-t(s_2))$. Then the probability that f_1 is the true value is $1-(1-t(s_1))(1-t(s_2))$. In general, if k data sources $S=\{s_1,s_2,s_3,\ldots,s_k\}$ all provide the fact f for the same attribute ea, the accuracy of f is

$$t(f)=1-\sum_{i=1}^{k}(1-t(s_i)). \tag{2}$$

Supposing $t(s_1)=0.9$ and $t(s_2)=1.1t(s_1)$, the discrimination between the reliabilities of s_1 and s_2 is not obvious, and $(1-t(s_1))(1-t(s_2))$ is very small, which tends to cause the underflow of calculations. Thus, in order to effectively use the data and for the convenience of calculation [3], we amplify the data with

$$t_g(s)=-\ln(1-t(s)) \tag{3}$$

where $t_g(s)$ is used to replace $t(s)$ to denote the reliability of the data source s, $t_g(s)\in(0,+\infty)$.

Similarly, in order to denote the accuracy of the fact f, we use $t_p(f)$ to denote the reliability of the fact f instead of $t(f)$.

$$t_p(f)=-\ln(1-t(f)) \tag{4}$$

where $t_p(f)\in(0,+\infty)$. The accuracy of the fact f is equal to the sum of the reliabilities of all data sources that provide the fact f, that is

$$t_p(f)=\sum_{i=1}^{k}t_g(s_i) \tag{5}$$

where $s_i\in S(f)$, and $S(f)$ is the set of data sources which provide the fact f.

(iii) Degree of mutual support among facts

For each entity attribute, there are always a lot of data sources providing a lot of facts, and there is a certain correlation between these facts. Suppose there exist two facts f_1 and f_2, f_1 is a fact provided by many data sources with very high reliabilities, and f_1 and f_2 have a strong correlation. Then, it is reasonable to consider that f_2 also gets support from these data sources with very high reliabilities. Thus, it is necessary to increase the accuracy of f_2.

For the calculation of mutual support degree among facts, current truth finder algorithms usually use the string similarity algorithm based on editing distance instead of the algorithm based on the mutual support degree among facts. Because the calculation of mutual support degree among facts has an important impact on the effectiveness and correctness of the whole truth finder algorithm, by improving the calculation method of mutual support degree among facts, it can also improve the accuracy of the algorithm. However, the string similarity algorithms based on editing distance can only literally measure the relationships between different facts, but cannot exactly measure the mutual support degree among facts. Therefore, in this paper, we combine the one-way text similarity and the degree of fact conflicts among descriptions of facts to calculate the mutual support degree among facts, which can better improve the accuracy of the algorithm.

(a) Degree of fact conflict

The set of different conflicting facts provided by each data source for an entity attribute $ea\in EA$ is descried as $F(ea)=\{f_1,f_2,f_3,\ldots,f_k\}$. Then, the conflict degree of f_i, denoted as $conflict(f)$, is expressed as

$$conflict(f_i)=1-\frac{|f_i|}{\sum_{j=1}^{k}|f_j|} \tag{6}$$

where $|f_i|$ represents the times that the fact f_i appears.

Obviously, for a fact, the more data sources it provides, the lower conflict degree it has and the more accurate it is.

(b) Similarity degree of one-way text among descriptions of facts $sim(f_i,f_j)$

For the facts provided by each data source, the stop word list is designed here to remove useless stop words and special symbols in strings, such as punctuations, word segmentations, messy codes, and a keyword which can express the meaning of original strings will be extracted. The keyword is called a lemma. In this paper, we define four kinds of relations: inclusion, equivalence, equality and independence. For example, the data sources s_1 and s_2 provide the facts of the authors of three book as shown in Table 1.

Table 1 Author information of three books provided by s_1 and s_2

Book	s_1	s_2
Book 1	Jennifer Widom	J. Widom
Book 2	Yang J M, Wang Y, Zhu J	Yang J M, Wang Y, Zhu J
Book 3	Chris Grover, E. A. Vander Veer	E. A. Vander Veer, Chris Grover

For Book 1, the fact about authors provided by s_2 is the abbreviation of the fact provided by s_1. Another example is Book 3: these two facts are consistent in content, while the formats are different. The relationship between these two facts is called the equivalence relationship. For Book 2, s_1 provides a little less author information than s_2, that is, the facts that s_2 provides contains the facts that s_1 provides, and this relationship is called the inclusion relationship. If

the facts provided by two data sources do not have any intersection and are exactly the same, we call it the independence relationship and equality relationship respectively. As a result, we define the similarity degree of one-way text between facts f_i and f_j as

$$sim(f_i, f_j) = \begin{cases} \dfrac{c}{len(f_i)}, & i \neq j \\ 0, & i = j \end{cases} \tag{7}$$

where c represents the number of common lemmas that f_i and f_j contain, $len(f_i)$ represents the number of lemmas in fact f_i. Therefore, if the relationship between f_i and f_j is an equivalence relationship or an equality relationship, $sim(f_i, f_j) = 1$; if the relationship between f_i and f_j is an independence relationship, $sim(f_i, f_j) = 0$; otherwise, the relationship between f_i and f_j is an inclusion relationship, i.e. $sim(f_i, f_j) = \dfrac{c}{len(f_i)} \neq sim(f_j, f_i) = \dfrac{c}{len(f_j)}$.

Taking the impact of the mutual support degree among facts for the same entity attribute into account, the accuracy of the fact f is adjusted as

$$t'_p(f) = t_p(f) + (1 - conflict(f))* \sum_{i=1}^{n}[t_p(f_i) * sim(f_i, f)] \tag{8}$$

where z represents the number of facts that each data source provides for an entity attribute $ea \in EA$.

(iv) Dependence relationships among data sources

When calculating the accuracies of facts, we assume that the data sources are independent of each other. However, the data sources show some common features during the process of providing various facts, which indicates an interdependent relationship among them. In this paper, we use the symmetrical inclusion of data sources to measure the dependences among data sources.

If two data sources provide consistent facts for many entity attributes, there are dependences between these two data sources. As a result, the facts provided by them for other entity attributes are also very likely to have the same reliability.

Assume that s_i and s_j provide m and n facts respectively, and the number of same facts (including the two relationships of equality and equivalence) is t. $p = m - t$ is used to indicate the number of facts provided by s_i which are different from the facts provided by s_j; $q = n - t$ indicates the number of facts provided by s_j which is different from the facts provided by s_i. Then, the symmetrical inclusion between s_i and s_j is

$$depend(s_i, s_j) = \begin{cases} \dfrac{t}{p + q + t}, & i \neq j \\ 0, & i = j \end{cases}. \tag{9}$$

Taking the dependence relationships among data sources into account, the accuracy of the fact f is adjusted as

$$t'_g(s_i) = t_g(s_i) + \frac{1}{l}\sum_{j=1}^{l}[t_g(s_j) * depend(s_i, s_j)] \tag{10}$$

$$t''_p(f) = \sum_{i=1}^{k} t'_g(s_i) \tag{11}$$

$$t'''_p(f) = t''_p(f) + (1 - conflict(f))* \sum_{i=1}^{z}[t_p(f_i) * sim(f_i, f)] \tag{12}$$

where l represents the number of data sources, k represents the number of data sources that provide fact f, and z represents the number of facts provided by each data source for an entity attribute.

(v) Normalization

As shown in (12), the accuracies of facts are greater than 0 and approach the infinity during the iteration process. Based on the iterative mechanism and for the convenience of calculation, the normalization of the accuracies of facts and the reliabilities of data sources is necessary.

$$t^*_p(f_i) = \frac{t'''_p(f_i)}{\sum\limits_{f_i \in F(ea)} f'''_p(f_i)} \tag{12}$$

where $F(ea)$ represents the set of all the conflicting facts provided by each data source for an entity attribute. Equation (1) can be further rewritten as

$$t^*_g(s) = \frac{1}{m}\sum_{i=1}^{m} t^*_p(f_i) \tag{13}$$

where $f_i \in F(s)$, $F(s)$ represents the set of facts provided by the data source s, and m is the number of facts provided by s.

3.4 Workflow

The iteration mechanism is used to calculate the reliabilities of data sources and the accuracies of facts. The workflow of TFAEA is shown in Fig. 3.

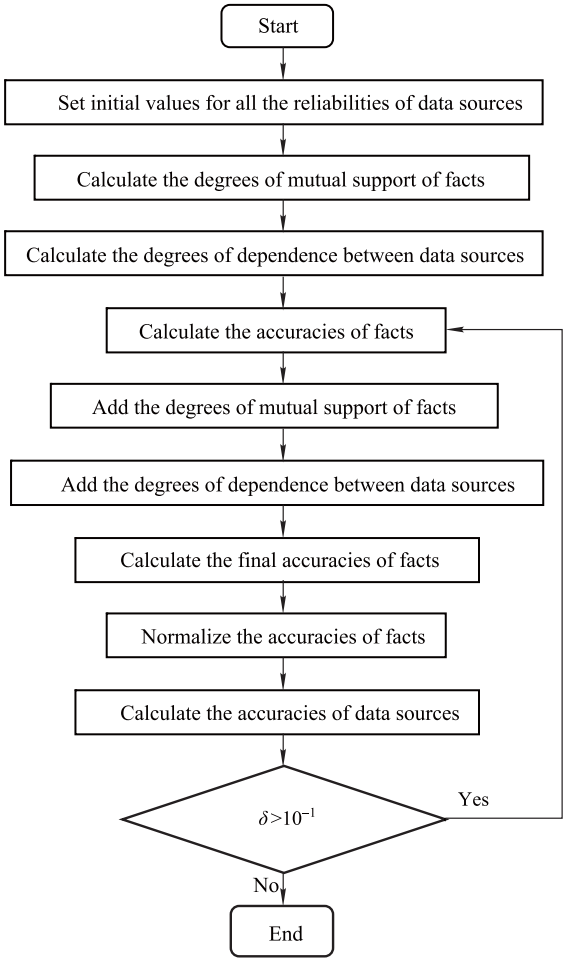


Fig. 3 Workflow of TFAEA

Firstly, initial values for all reliabilities of data sources are set. The calculation of the mutual support degree among facts and the dependences among data sources are decided by the given static dataset. Therefore, it only needs to calculate once during the whole iteration process. In each iteration process, successively add the factors of the mutual support degree among facts and the dependences among data sources, and recalculate the accuracies of facts and the reliabilities of data sources. Repeat the calculation in this way, and the iteration process will end when the reliabilities of data sources in two successive iterations are smaller than or equal to the preset difference (such as $\delta = 10^{-5}$).

3.5 Computational complexity

Suppose that the number of data source is m, the total number of facts provided is n, and n is much greater than m.

In the first stage of TFAEA, before the iterative calculation, the mutual support degree between facts and the dependent relationship between data sources are calculated. The mutual support degree are calculated with the conflict degree of facts and the one-way text similarity. Therefore, both the time complexity and the space complexity of calculating the conflict degree of fact are $O(n)$. The time complexity of calculating the one-way text similarity is $O(n(n-1))$, and the space complexity is $O(n^2)$. The time complexity of calculating the dependent relationship between data sources is $O(m(m-1)/2)$, and the space complexity is $O(m^2)$.

In the second stage of TFAEA, the time complexity of calculating the reliability of data source is $O(m^2)$. Supposing the iteration is conducted i times, the time complexity of calculating the accuracy of fact is $O(2im^2)$. The space complexity is $O(n+n^2+m^2)$.

TFAEA does not need to obtain the initial value via the pre-training. The mutual support degree between facts and the dependent relationship between data sources are determined by the data set itself, and can be obtained before the iteration, which improves the computational efficiency of the algorithm.

4. Experiments and performance analysis

4.1 Experimental environment

The *Octopus Collector* is obtained from the website www.amazon.com to capture the ISBNs of 1 422 books related to computer science and technology. ISBN is used as the keyword for searching on the website www.abebooks.com and capturing all the data sources which provide information of these books. These data sources and the book information they provide are used as the experimental dataset in this paper, including ISBN, title, author, publisher and data source. The dataset contains 27 965 books from 752 different data sources, in which there are 6 084 different pieces of information about authors and 6 501 different pieces of information about titles. Obviously, there are multiple data sources in this dataset that provide different information for the same book, which has generated data conflict. Since the ISBN of each book is unique, this experiment will solve the conflicts for two different types of information (book author and title) to verify TFAEA. Compared to the information of title, the information of author is more complex and diverse. The information of author consists of several parts, and its problems include logogram, abbreviation, haplography, reverse order, etc.

The experimental data are preprocessed to remove the noise information, such as punctuation, word segmentation, messy code, unification of upper and lower case, so that it will be able to accurately reflect the algorithm's validity and accuracy. The dataset contains 752 data sources.

In this paper, 20 data sources $s_1, s_2, s_3, \ldots, s_{20}$ provide the most number of books selected for experiments.

4.2 Experiments

First, the initial value of the reliability of each data source is set to 0.8, and the iterative computations of the reliabilities of 20 data sources are implemented.

As shown in Table 2 and Fig. 4, in accordance with the data presented during each iteration, the reliability of each data source increases gradually with the increasing number of iterations, and then approaches a stable state. Most data sources reach a stable value in the 9th iteration, that is, the difference between two successive iterations of the data source is 0. Several data sources have reached a stable value with less iterations. For example, s_6, s_{10}, s_{12} and s_{20} reach a stable value in the 8th, 6th, 7th and 8th iterations respectively. Table 2 shows that TFAEA can achieve the state of convergence after finite iterations, and it can also effectively calculate the reliability of each data source.

Table 2 Data in each iteration

Data source	Iteration time				
	1	2	3	4	5
S_1	0.640 33	0.768 59	0.792 84	0.797 79	0.798 95
S_2	0.720 81	0.811 46	0.828 40	0.832 45	0.833 52
S_3	0.578 21	0.696 52	0.725 00	0.731 99	0.733 79
S_4	0.586 53	0.689 83	0.711 78	0.716 94	0.718 26
S_5	0.559 43	0.637 81	0.651 44	0.654 37	0.655 08
S_6	0.663 05	0.775 17	0.794 47	0.798 44	0.799 39
S_7	0.529 08	0.617 73	0.636 87	0.641 39	0.642 55
S_8	0.672 49	0.767 49	0.783 76	0.787 14	0.787 95
S_9	0.304 95	0.310 88	0.309 08	0.308 23	0.307 94
S_{10}	0.344 49	0.364 48	0.367 38	0.367 90	0.368 00
S_{11}	0.577	0.658 89	0.685 18	0.694 96	0.698 62
S_{12}	0.313 74	0.295 31	0.281 44	0.276 13	0.274 13
S_{13}	0.439 62	0.567 91	0.605 07	0.618 34	0.623 26
S_{14}	0.871 77	0.812 46	0.783 77	0.773 12	0.769 14
S_{15}	0.593 16	0.625 21	0.638 83	0.644 05	0.646 01
S_{16}	0.342 49	0.412 25	0.436 2	0.445 02	0.448 31
S_{17}	0.398 42	0.548 38	0.579 21	0.589 42	0.593 14
S_{18}	0.737 36	0.796 59	0.819 33	0.827 81	0.830 98
S_{19}	0.426 07	0.537 62	0.568 39	0.579 53	0.583 69
S_{20}	0.574 6	0.706 86	0.733 45	0.741 47	0.744 14
Data source	Iteration time				
	6	7	8	9	10
S_1	0.799 24	0.799 32	0.799 34	0.799 35	0.799 35
S_2	0.833 81	0.833 89	0.833 91	0.833 92	0.833 92
S_3	0.734 26	0.734 39	0.734 43	0.734 44	0.734 44
S_4	0.718 62	0.718 71	0.718 74	0.718 75	0.718 75
S_5	0.655 27	0.655 32	0.655 33	0.655 34	0.655 34
S_6	0.799 64	0.799 70	0.799 72	0.799 72	0.799 72
S_7	0.642 85	0.642 93	0.642 95	0.642 96	0.642 96
S_8	0.788 15	0.788 21	0.788 22	0.788 23	0.788 23
S_9	0.307 84	0.307 81	0.307 81	0.307 80	0.307 80
S_{10}	0.368 01	0.368 01	0.368 01	0.368 01	0.368 01
S_{11}	0.699 47	0.699 59	0.699 61	0.669 62	0.669 62
S_{12}	0.273 67	0.273 65	0.273 65	0.273 65	0.273 65
S_{13}	0.624 4	0.625 1	0.625 19	0.625 2	0.625 2
S_{14}	0.768 22	0.767 65	0.767 58	0.767 55	0.767 53
S_{15}	0.646 47	0.646 75	0.646 76	0.646 76	0.646 76
S_{16}	0.449 07	0.449 54	0.449 58	0.449 61	0.449 61
S_{17}	0.593 99	0.594 52	0.594 57	0.594 58	0.594 58
S_{18}	0.831 72	0.832 17	0.832 25	0.832 26	0.832 26
S_{19}	0.584 66	0.585 25	0.585 31	0.585 32	0.585 32
S_{20}	0.744 73	0.745 07	0.745 1	0.745 1	0.745 1

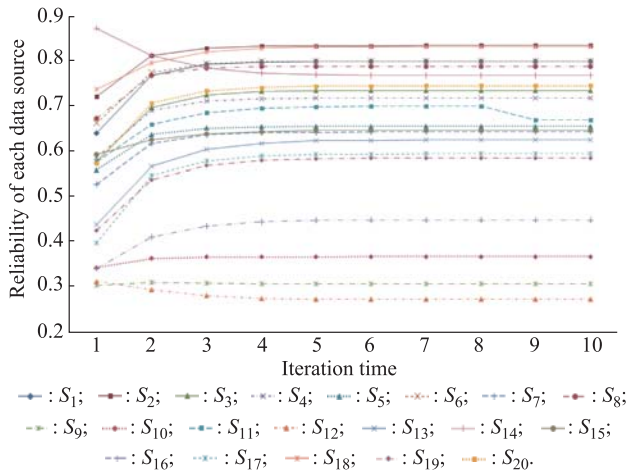


Fig. 4 Reliability of each data source

The initial value of the reliability of each data source is set from 0.1 to 0.9 respectively. Experiments are implemented on the data sources s_1, s_2 and s_3 respectively, with different initial values.

As shown in Table 3, different initial values of the reliability of a data source have no impact on either the reliability of the data source or the speed of the algorithm's convergence, which validates the stability and reliability of TFAEA.

Table 3 TFAEA's iteration data with different initial values

Initial value	Data source	Iteration time				
		1	2	3	4	5
0.1	S_1	0.631 13	0.586 05	0.572 96	0.568 79	0.567 31
	S_2	0.470 13	0.560 77	0.591 90	0.603 05	0.607 19
	S_3	0.571 59	0.518 92	0.501 41	0.495 2	0.492 92
0.2	S_1	0.627 10	0.585 79	0.572 92	0.568 79	0.567 32
	S_2	0.461 94	0.560 58	0.591 90	0.603 06	0.607 2
	S_3	0.568 86	0.518 74	0.501 33	0.495 16	0.492 9
0.3	S_1	0.623 78	0.585 64	0.572 91	0.568 79	0.567 32
	S_2	0.456 01	0.560 45	0.591 89	0.603 07	0.607 2
	S_3	0.567 28	0.518 67	0.501 30	0.495 15	0.492 89
0.4	S_1	0.621 09	0.585 53	0.572 90	0.568 79	0.567 32
	S_2	0.451 59	0.560 38	0.591 90	0.603 07	0.607 2
	S_3	0.566 31	0.518 62	0.501 28	0.495 14	0.492 88

Continued

Initial value	Data source	Iteration time				
		1	2	3	4	5
0.5	S_1	0.618 92	0.585 45	0.572 89	0.568 79	0.567 32
	S_2	0.448 15	0.560 33	0.591 90	0.603 07	0.607 21
	S_3	0.565 74	0.518 58	0.501 26	0.495 13	0.492 88
0.6	S_1	0.615 72	0.600 01	0.572 88	0.568 79	0.567 32
	S_2	0.443 08	0.524 43	0.591 89	0.603 07	0.607 21
	S_3	0.565 40	0.537 46	0.501 25	0.495 13	0.492 88
0.7	S_1	0.615 72	0.585 35	0.572 88	0.568 79	0.567 32
	S_2	0.443 08	0.560 22	0.591 89	0.603 07	0.607 21
	S_3	0.565 40	0.518 56	0.501 25	0.495 13	0.492 88
0.8	S_1	0.614 54	0.585 32	0.572 88	0.568 79	0.567 32
	S_2	0.441 15	0.560 17	0.591 88	0.603 07	0.607 21
	S_3	0.565 47	0.518 56	0.501 25	0.495 13	0.492 88
0.9	S_1	0.613 55	0.585 29	0.572 88	0.568 79	0.567 32
	S_2	0.439 50	0.560 12	0.591 87	0.603 07	0.607 21
	S_3	0.565 59	0.518 57	0.501 25	0.495 13	0.492 88

Initial value	Data source	Iteration time				
		6	7	8	9	10
0.1	S_1	0.566 98	0.566 77	0.566 75	0.566 74	0.566 74
	S_2	0.608 15	0.608 33	0.608 35	0.608 36	0.608 36
	S_3	0.492 39	0.492 2	0.492 17	0.492 15	0.492 15
0.2	S_1	0.566 98	0.566 77	0.566 75	0.566 74	0.566 74
	S_2	0.608 16	0.608 32	0.608 33	0.608 35	0.608 36
	S_3	0.492 37	0.492 21	0.492 16	0.492 15	0.492 15
0.3	S_1	0.566 98	0.566 78	0.566 76	0.566 74	0.566 74
	S_2	0.608 16	0.608 32	0.608 35	0.608 36	0.608 36
	S_3	0.492 37	0.492 22	0.492 17	0.492 15	0.492 15
0.4	S_1	0.566 98	0.566 78	0.566 75	0.566 74	0.566 74
	S_2	0.608 16	0.608 32	0.608 35	0.608 36	0.608 36
	S_3	0.492 36	0.492 22	0.492 17	0.492 15	0.492 15
0.5	S_1	0.566 98	0.566 78	0.566 75	0.566 74	0.566 74
	S_2	0.608 17	0.608 34	0.608 36	0.608 36	0.608 36
	S_3	0.492 36	0.492 22	0.492 17	0.492 15	0.492 15
0.6	S_1	0.566 98	0.566 78	0.566 75	0.566 74	0.566 74
	S_2	0.608 17	0.608 34	0.608 36	0.608 36	0.608 36
	S_3	0.492 36	0.492 22	0.492 17	0.492 15	0.492 15
0.7	S_1	0.566 98	0.566 78	0.566 75	0.566 74	0.566 74
	S_2	0.608 17	0.608 34	0.608 36	0.608 36	0.608 36
	S_3	0.492 36	0.492 22	0.492 17	0.492 15	0.492 15
0.8	S_1	0.566 99	0.566 78	0.566 75	0.566 74	0.566 74
	S_2	0.608 17	0.608 34	0.608 36	0.608 36	0.608 36
	S_3	0.492 36	0.492 19	0.492 16	0.492 15	0.492 15
0.9	S_1	0.566 99	0.566 78	0.566 75	0.566 74	0.566 74
	S_2	0.608 17	0.608 34	0.608 36	0.608 36	0.608 36
	S_3	0.492 36	0.492 22	0.492 17	0.492 15	0.492 15

We randomly select 100 books from the dataset, record the information of titles and authors on the cover of books and set them as true values. The information of book titles and authors is used as the experimental object. As shown in Fig.5, TFAEA has the higher accuracy compared with Voting and TruthFinder, and its accuracy is higher than 80% on both title and author datasets. Followed by TruthFinder, Voting has the lowest accuracy. From the dataset of title to the dataset of author, the accuracies of all three algorithms have declined, but TFAEA has the least decline, which also indicates the stronger stability and adaptability of TFAEA, and it has higher efficiency especially when handling diverse and complex text information. Voting only considers that the minority is subordinate to the majority without taking the dependences among the data sources and their reliabilities into account. It also fails to consider the mutual support degrees among facts and other important factors, which makes it have the lowest accuracy among the three algorithms. TruthFinder and TFAEA both consider the reliabilities of data sources and the mutual support degrees among facts. The difference between them is that TruthFinder only uses the similarity of text based on editing distance to replace the mutual support among facts, which causes low adaptation of the algorithm, and the accuracy is especially not ideal when handling complex datasets. As shown in Fig. 5, from book titles to book authors, the accuracy rate of TruthFinder has a significant decline. However, TFAEA combines the degree of fact conflict and the degree of one-way similarity among the descriptions of facts to calculate the mutual support degree among facts, and also considers the impact among data sources, which makes TFAEA more accurate.

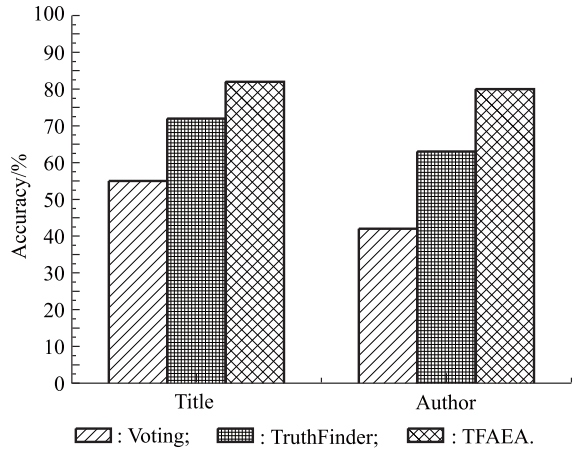


Fig. 5 Comparison of the accuracy of TFAEA, Voting and TruthFinder

5. Conclusions

In this paper, based on the reliabilities of data sources and the iterative calculation mechanism of fact accuracy, TFAEA fully considers the mutual support degrees among the facts and the dependences among data sources. During the iteration process of the algorithm, we present the specific method to calculate the mutual support degree among facts and the dependences among data sources, which realizes higher stability and accuracy of TFAEA. In future, we will consider the data copy situations among data sources

and the qualities of data sources in order to improve the performance of TFAEA.

References

[1] C. Sun, D. Shen, Y. Kou, et al. A genetic algorithm based entity resolution approach with active learning. *Frontiers of Computer Science*, 2016, 10(4): 1 – 13.

[2] C. Sun, D. Shen, Y. Kou, et al. A related data oriented joint entity resolution approach. *Chinese Journal of Computers*, 2015, 38(9): 1739 – 1754. (in Chinese)

[3] X. Yin, J. Han, P. Yu. Truth discovery with multiple conflicting information providers on the Web. *IEEE Trans. on Knowledges and Engineering*, 2008, 20(6): 796 – 808.

[4] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 1999, 46(5): 604 – 632.

[5] P. Lawrence, B. Sergey, M. Rajeev, et al. The PageRank citation ranking: bringing order to the Web. *Stanford Infolab*, 1998, 9(1): 1 – 14.

[6] X. Dong, L. Berti-Equille, D. Srivastava. Integrating conflicting data: the role of source dependence. *Very Large Data Base Endowment*, 2009, 2(1): 550 – 561.

[7] X. Dong, F. Naumann. Data fusion-resolving data conflicts for integration. *Very Large Data Base Endowment*, 2009, 2(2): 1654 – 1655.

[8] M. Kao, W. Zhang, H. Gao. Truth discovery methods in conflict data integration. *Chinese Journal of Computer Research and Development*, 2010, 47(S): 188 – 192. (in Chinese)

[9] B. Zhao, B. Rubinstein, J. Gemmell, et al. A Bayesian approach to discovering truth from conflicting sources for data integration. *Very Large Data Base Endowment*, 2012, 5(6): 550 – 561.

[10] J. Pasternack, R. Dan. Knowing what to believe (when you already know something). *Proc. of the 23rd International Conference on Computational Linguistics*, 2010: 877 – 885.

[11] Z. Zhang, L. Liu, X. Xie, et al. Information evaluation based on sources dependence. *Chinese Journal of Computers*, 2012, 35(11): 2392 – 2402. (in Chinese)

[12] A. Galland, S. Abiteboul, A. Marian, et al. Corroborating information from disagreeing views. *Proc. of the 3rd ACM International Conference on Web Search and Data Mining*, 2010: 131 – 140.

[13] W. Tan, X. Yin. Semi-supervised truth discovery. *Proc. of the 20th International Conference on World Wide Web*, 2012: 217 – 226.

[14] X. Zhang. Classification-based data integration method. Guangzhou, China: Guangdong University of Technology, 2013. (in Chinese)

[15] Z. Yang. *Research on the accuracy of the data quality*. Harbin, China: Harbin Institute of Technology, 2013. (in Chinese)

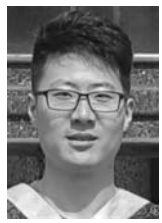
Biographies

Xiaolong Xu was born in 1977. He received his B.S. degree in computer and its applications, M.S. degree in computer software and theories and Ph.D. degree in communications and information systems from Nanjing University of Posts and Telecommunications, Nanjing, China, in 1999, 2002 and 2008, respectively. He worked as a postdoctoral researcher at Station of Electronic Science and Technology, Nanjing University of Posts and Telecommunications from 2011 to 2013. He is currently a professor in College of Computer, Nanjing University of Posts and Telecommunications. He is a senior member of China Computer Federation. His current research interests include cloud computing, mobile computing, intelligent agent and information security.

E-mail: xuxl@njupt.edu.cn

Xinxin Liu was born in 1994. She received her B.S. degree in software engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2016. She is now a postgraduate student majoring in software engineering, and work in a project team supported by State Key Laboratory of Information Security, Chinese Academy of Sciences, Beijing, China. Her current research interests include mobile computing and information security.

E-mail: B12040303@njupt.edu.cn



Xiaoxiao Liu was born in 1988. He received his B.S. degree in computer science and technology from Sanjiang University, Nanjing, China, in 2013, and M.S. degree in computer technology from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2016, respectively. He works as an engineer in Institute of Big Data Research at Yancheng, Nanjing University of Posts and Telecommunications, Yancheng, China, carrying out research in mobile computing and data synchronization technology.

E-mail: 1213043115@njupt.edu.cn

Yanfei Sun was born in 1976. He received his Ph.D. degree in communications and information systems from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2006. He is currently a professor and the director in Science and Technology Department, Nanjing University of Posts Telecommunications. His current research interests include communication network, mobile networks and big data.

E-mail: sunyanfei@njupt.edu.cn