Decision Model Based on Semantic Similarity in Internet Information Recognition

Han Jing and Ma Jing
School of Foreign Languages, Handan College, Handan City 056005, China

Abstract—In order to solve the problems like the credibility of internet information, this paper proposes the credible decision model based on semantic similarity. The model determines the corresponding relevant text information by capturing and analyzing facts statement and realizes the credible decision to determine the facts. In the process of facts credible decision, it considers the semantic similarity between related text information and facts statement and the credibility of related factors. The contribution of related text information is measured, and the fact statement determination is conducted. Simulation experiments show that the reliable determination of the model is correct and accurate.

Index Terms—Internet; Information Identification; Sentence Extraction; Similarity

I. INTRODUCTION

At present, the similarity calculation between information object is mainly on the basis of describing the external form of the object, such as the similarity calculation of traditional document is mainly on the basis of whether key is all in the two documents, key words must be “exactly the same” in the same concept under the cases of the different expression or the same word in different contexts have different semantic, seriously affecting the accuracy of the similarity calculation, the same words in the description of different objects can express different meanings; you can use the same Structure or different vocabulary with different Structure but have the same or similar semantic to describe the same object, in order to solve the problem of semantic conflict between this data, need to fundamentally eliminate the information under the network environment, to realize the connectivity of information resources and resource sharing. With the development of semantic section touch and ontology, this problem has been solved, the semantic web is a concept put forward by TimBemers-Lee in 1998, its core idea is to add semantics which can be understood by computer on documents of the world wide web, and then, makes the whole Internet become a general medium of exchanging information [1].

False facts information describe on the Internet affected people obtain information effectively. Therefore, how to recognize the fact that the Internet credible describes information becomes a problem to be solved. On the internet, the description is mainly in the form of facts, has the branch of true and false, do not include emotional and degree of description [2-3]. If a fact’s message is believable, the facts as credible facts; On the contrary, the facts are not to be trusted. According to the characteristics of the facts shows that for any negative facts, there is a certain facts and the matching, and negative facts credible decision can realize through the credible determination implementation of corresponding certain facts, this article only for credible decision of certain facts.

People in determining a facts should or shouldn’t be trusted, will specify the fact statement doubt unit, the user needs to determine the fact that the narrative part [4]. In this case, the facts can be seen as the answer about a topic, when facts’ corresponding theme has only one correct answer, the facts as the only answer; On the other hand, is called the more answer facts statement [5-7].

For example: specific facts "Obama is American president "'s suspect unit is "Obama", then the facts corresponding theme: "Who is American president ", the answer of the theme is unique, therefore, facts statement " Obama is American president " is the only answer to the facts; When the suspect unit has been designated as " American president ", the corresponding theme is "What is Obama", have more than one correct answer, this theme including "Obama is Christian", " Obama is American president" and so on, in this case, the facts is the more answer facts statement [8-12].

Unrelated facts statement credible decision in the existing field, facts credible decision mainly achieve through the analysis of the relevant text information of the facts. Some methods through the analysis of the number of relevant text information and emotional tendency to realize facts judgment, however this kind of method shortcomings are as follows: (a) The relevant text information do not always support the corresponding facts, so the related text information quantity is unreliable as a judge; (b) Related theme of the text information is not necessarily to determine facts, so emotional tendency of relevant text information does not take the place of its treatment for determining facts emotion tendency; (c) Related text information on difference of the credibility is not considered reasonable [13-14].

Another method found comparable of determine facts from the text database, through treatment the determine facts and comparable facts statement, select the highest score as reliable facts, so as to realize the reliable determination of facts. However, these methods have the following drawbacks: (a) In the process of obtaining the fact statement, it needs suspected unit of specified
determine facts, when doubt unit contains too much information, however, is difficult to find suitable contrast facts; (b) This kind of method applies to the only answer determination of facts and in determining the more answer facts, in many cases do not make the accuracy judgment [15].

Aiming at the shortcomings of the above research work, this paper proposes a new field of unrelated facts credible evaluation model MFSV (Multi answer Fact Statements Verification). The model through search engines to get related text information of determine facts; In the process of the determination of facts, the model takes the relevant text information and the corresponding relation between the support of the facts and relevant text information in the differences of credibility into account, to make up for the deficiency of the category 1 methods; In addition, the facts determination process in the use of the model, do not need to specify the unit to facts, do not need to search and analysis comparable facts, so this model can also be applied to many answers facts statement credible determine, makes up for the class 2 methods in many facts judgment on the limitations of the answers.

This paper mainly makes the development and innovation of work in the following aspects:

(a) In view of the Internet information credibility problem, this paper put forward a credible decision model MFSV based on semantic similarity. The model using the search engines to get text information related determine facts; Analysis the semantic similarity between related text information and determine of facts, on the basis of measuring text information and corresponding to the support of the relationship between facts; Through analyzing the importance of the source text information and popularity, get order for text information credibility; according to semantic similarity and text information credibility, measure contribution of relevant text information for the facts credible judgment, and on this basis, put forward the two facts credible methods: basic judgment method and decision method based on SVM classification, to realize the reliable determination of facts;

(b) To measure the facts and the semantic similarity between the relevant text information. According to the text information not every words to make contribution to the determination of facts, extracted from the text information is proposed for facts to determine the method of meaningful sentence, so as to extract the facts determining meaningful sentences; Extends the sentence similarity computing method based on semantic and word order, measure extract sentences in the text information and semantic similarity between facts, thus get the text information and the semantic similarity between facts; According to relevant information’s popularity and importance to obtain relevant information’s sort of the web page credibility; Using reliable sources of information are more likely to pass the principle of credible information, measure the credibility of the related text information, obtain the relevant information credibility sort; The judge the rationality of the model is verified by a series of experiments with accuracy.

(c) In order to further verify the proposed credible decision model based on semantic similarity, the correctness and effectiveness of the relevant text information, information classification parameter estimation, the text information credibility sorting and credible decision model based on semantic similarity methods accuracy evaluation of simulation experiment. Simulation experiments show that the reliable determination of the model is correct and accurate.

II. CREDIBLE DECISION MODEL

Fact credible evaluation model of MFSV as shown in figure 1, the model consists of four modules: related text information retrieval; Semantic similarity measure; Text information credibility sort; Facts credible decision. The input of the model is to be determined facts; the output is the reliable determination result of the facts.

![Figure 1. Fact credible evaluation model of MFSV](image)

Related text information acquisition module and related text information of determine facts. Because the model for the field of unrelated facts judgment, so the text information obtained through search engines. The input for the module to determine facts, text output statements to the corresponding related information collection. Semantic similarity measurement module is used to calculate related text information and semantic similarity between the corresponding facts.

First, extracted meaningful sentences for the facts determining from the text information; second, get the similarity between sentence and corresponding facts by measuring extraction, thus get the text information and semantic similarity between the corresponding facts. The module input S the related text information and the corresponding facts, output text information with the corresponding state of semantic similarity. Credibility sort module implements the text information related to the credibility of the sort. In the module, through access to relevant source text information corresponding pager **ank** (site) and its position in the **Ale xa** ranking, implement the relevant documents information source’s importance sorting and popularity sorting, merging the two sorts, namely for text information related to the credibility of the sort. The module input source for related text information (station), the output to text information credibility. Facts credible decision module implements the facts credible decision. In this module, according to the text information and corresponding similarity of facts and text information credibility sorting, measure the text information corresponding to the facts.
determining contribution, comprehensive considerations contribution of text information, achieve the credible determine facts. The module inputs related text information corresponding semantic similarity and credibility, the output to determine the results.

A. Semantic Similarity

1) Basic Knowledge

**Definition 1** the relevant text information collection

\[ R = \{ r_1, r_2, ..., r_n \} \]

represent n collection of text information related to determine facts. Among them, the \( r_i \) represents the i text message of \( R(1 \leq i \leq n) \), set its from the web PI. \( R_i \) can be expressed as

\[ r_i = < t_i, s_i, urli > \]

including:

- \( t_i \): Represents the title of the PI;
- \( s_i \): Represents text content of related corresponding facts in the \( P_i \);
- \( urli \): Represents the website of \( P_i \).

In fact, credible decision model, relevant text information through search engines, so the \( s_i \) by a search engine drawn from \( P_i \). Relevant text information in accordance with the rules searches engine ranking order, according to the sort order for related text information.

**Definition 2** Key words set k and \( s_i \). \( K \) represents facts \( (f_s) \) words set after removing the stop, called the \( f_s \), keywords set. \( k_i \) Represents in \( r_i f_s \), keywords \( f_s \) related set of text information, if regard \( r_i \) as a collection of words, there is

\[ s_i = \{ \sigma | \sigma \in l \land \sigma \in r_i \} \]

**Definition 3** The shortest continuous sentence combination is \( c_i \). The given text information \( r_i \) and \( k_i \), said \( c_i \) for the shortest continuous sentence combination of \( k_i \) if and only if the \( c_i \) completely contains \( k_i \) and there is no other contains \( k_i \) in \( r_i \) and length less than continuous sentence combination of \( c_i \).

Stanford Parser is grammar dependent relationships analysis tool between words, use Stanford Parser can obtain the syntax of the dependencies between words in a sentence. In Stanford Parser, grammatical dependencies is expressed as the following format: dependencies (dependent, depends), shows that in the sentence "dependent" and "dependent" between "dependencies" refers to grammatical dependencies. In dependency grammar, "dependent" attached to the "dependent", in a subordinate position. For example, dependency grammar \( a_{mod} \) (great, president) represent dependent adjective modifiers between great and president, great modifies president. 52 kinds of grammatical dependencies between each word is given by Stanford Parser, including the subject-predicate relationship \( n_{subj} \), verb-object \( dobj \), \( a_{mod} \) adjectives modify \( a_{mod} \), etc. This paper according to Stanford Parser relation with the backbone of the sentence, it can be divided into two categories: the necessary dependencies GRE and unnecessary dependencies GRO.

For a dependency grammar, if the dependence and is dependent on all is the main part of the sentence, that it is necessary to rely on relationship, otherwise is a necessary dependencies. For example, represent grammatical dependencies in the sentence \( (n_{subj}, x_{subj}) \), verb-object dependencies \( (dobj, obj) \) are necessary dependencies; Qualified modified role in sentence grammar dependencies, such as adjectives modify relationship \( (a_{mod}) \), adverbs \( (adv_{mod}) \), etc., is a unnecessary dependencies.

2) Sentence Extraction

Related text information’s correlation between \( r_i \) and \( f_s \), mainly through \( r_i \) contained in the \( r_i \) sentence, so the shortest continuous sentence \( c_i \) combination with \( r_i \) contained in \( r_i \) as the foundation to realize sentence extraction. The by acquiring \( c_i \), grammar dependent relationships between words, combined with heuristic rules to get rid of \( c_i \) in the contribution to facts judgment without words, extract sentence which is meaningful for \( f_s \) determination. In this article, use \( st_i \) to express meaningful sentence of credible decision for \( f_s \) drawn from \( r_i \). Extraction of \( st_i \) steps are as follows:

1) access the shortest continuous combination \( c_i \) contains \( k_i \) in \( r_i \) sentence; (2) access to \( c_i \), grammatical dependencies set; (3) by heuristic rules, combining with \( c_i \) grammatical dependencies set of \( c_i \), get credible determines necessary word set \( f_s \) in the \( c_i \); (4) the word in the Ni, according to the position they appear in the \( c_i \) into the \( st_i \). Set \( Di \) for \( c_i \), middle grammar dependent relationships set, \( d_{yj} \) for \( c_i \) grammatical relations between words, \( d_{yj} \in d_i \) produced in the process of obtaining \( Ni \) using heuristic rules is as follows: (1) if the \( d_{yj} \) is necessary dependencies, dependents of \( d_{yj} \) (\( d_{yj} \). Dependent) and is dependent on (called\( d_{yj} \)) into the \( Ni \); (2) if the \( d_{yj} \) is not necessary dependencies, and, \( d_{yj} \). Dependent belongs to \( Ni \), will put \( d_{yj} \). Governor in \( n_i \), algorithm 1 describes the process of extracting the \( st_i \) from \( r_i \).

**Algorithm I** The sentence extraction

Input: Text information \( r_i \), keyword \( r_i \) contains \( f_s \), set \( k \)
Output: Sentence \( st_i \), extraction of text information

1) \( c_i = s_i, j \) that \( c_i \), reset is \( s_i \)
2) IF \( s_i \) contains all words \( in k \), then
3) \( c_i = t_i \)
ELSE FOR EACH \( s_i \) in \( s_i \) AND \( s_i \) contains
4) \( k, DO/1/s_i \). Composed of continuous sentences to include \( k_i \), \( s_i \) belongs to the \( s_i \).
5) IF \( len(s_i) < len(c_i) \) THEN

© 2014 ACADEMY PUBLISHER
Making sentence s1 is sentence st, abstracted from r, s2 for the r corresponding facts f. The following content calculates the similarity of s1 and s2 for example to explain similarity calculation process.

Correlation calculation of semantic vector, by generating sentence si and s2 corresponding semantic vector, computing the cosine similarity between semantic vector, implement semantic vector correlation calculation. Set s1, s2 respectively represent word set after remove stop.

w1 = {w10, w11, ..., w1n}
w2 = {w20, w21, ..., w2n}

Make

w = w1 ∪ w2, w = {w1, w2, ..., wn}

Set up corresponding semantic to si

v1 = {v10, v11, ..., v1m}

In the process of getting component v1ii is as follows:

1) w1 w w , if w1 w2 1;
2) w1 w w , find the best matching word wbest of w1 in s1, such as wbest = s1 (w1, WBM) , otherwise v2 = 0. For w1 to obtain the best matching word wbest process of algorithm 2.

**Algorithm 2** Get the best matching words

Input: w1(wi ∈ ei), ei, w2, w1 threshold δ

Output: w1 ins, to the best matching word wbest

1) Tenoset ← ∅;
2) FOR EACH w1 ∈ w1 DO
3) IF S(w1, w2) ≥ γ then
4) Tempest ← w1;
5) ENDF
6) ENDF
7) w1 = findle fword(w1, si);
8) w2 = findle rightword(w1, si);
9) Mvalue = 0
10) FOR EACH w1 ∈ Temset DO
11) IF(S(w1, w2) + s1(w1, w2)) + (w1, w2) = Mvalue) then
12) IF s1(w1, w2) ≤ s1(w1, w2) THEN
13) wbest = w1, ENDF
14) IF(S(w1, w2) + s1(w1, w2)) + (w1, w2) > Mvalue) then
15) Mvalue = S(w1, w2) + S(w1, w2) + S(w1, w2)
16) wbest = w1;
17) ENDF
18) ENDF
19) RETURN wbest

Similar to the method for v1, s2’s corresponding semantic vector v2 can be obtained. The s1 and s2 semantic vector similarity retrieval obtained by calculating cosine v1, v2. As shown in type (2)
\[ S_i(s_1, s_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|} \] (2)

Word order vector correlation calculation, by generating the corresponding sentence word order vector, according to the type (3) to calculate word order vector correlation between sentences

\[ t_m(t_1, t_2) = 1 - \frac{d_1 - \alpha}{d_1 + \alpha} \] (3)

The calculation of semantic similarity, according to the semantic vector correlation, and word order vector correlation, \( s_1 \) and \( s_2 \) semantic similarity can be calculated by type (4)

\[ s(s_1, f_1) = \begin{cases} \theta s_1(st_1, f_1) + (1 - \theta) s_2(st_1, f_1) & r_i \text{ does not exist negative phenomena in } f_i \\ -\theta s_1(st_1, f_1) + (1 - \theta) s_2(st_1, f_1) & \text{The existence of } f_i \text{ negative } r_i \end{cases} \] (4)

In the process of obtaining \( st_i \), according to it is involved in negative dependency grammar and whether there is a negative adverb and in \( c_i \), whether exist adverb or not such as hardly, rarely, few, seldom to determine whether the \( r_i \) has negative tendencies to \( f_i \). The Stanford Parser shows that sentence grammar dependent relationships between the corresponding word can say the sentence clearly exists negative tendencies. For example, when sentence appears the negative word not, corresponding grammatical dependencies for \( \text{neg} \), negative turn conjunction, such as rather than reflected in grammatical dependencies for \( \text{conj}_\text{negec} \) etc. So you can determine whether there is a negative sentence tend by dependency grammar relations; On the other hand, when there is a negative adverb in the sentence, such as hardly, rarely, few, seldom is also a tendency to determine whether there is a negative sentence. Therefore, whether the \( r_i \) has negative tendencies to \( f_i \) can be determined according to the following rules: (a) drawn \( st_i \) from \( c_i \) when involved in dependency grammar including no grammatical dependencies, such as \( \text{neg} \) and \( \text{conj}_\text{negec} \) etc. (b) Are there negative adverb included in \( c_i \), such as hardly, rarely, few, seldom, scarcely, never, little, etc. If you meet one of the rules above, thinks that \( r_i \) has negative tendencies to \( f_i \). Because this article is a credible decision on certain facts, therefore in the process of judging the negative tendencies, without considering the \( f_i \) as negative facts of the case.

B. The Sorting of Text Information Credibility

Whether facts statement is trusted or not is related to the credibility of the relevant text information. Usually, if the more credible text information supports the facts, the fact is more believable. Because trusted sites may always be carrying credible information, therefore these articles measure the credibility of the source text information (website), to realize the credibility sorting of text information. The reputation of the website is the important factors influencing the site whether can be trusted or not. If website is important, so the credibility of the site is higher; if the site in a long time to keep high traffic, namely the high popularity, so the site has high credibility. The importance of the site has certain relevance with popularity, but cannot replace each other. For example, Org website is often important, but the traffic is not high.

Website pagerank’s value describes the importance of the site Alexa ranking reflects the popularity of the website. Alexa ranking is based on site visits to site sorting, can be used as a site measure of popularity, the ranking in the top one million websites, Alexa rankings can more accurately reflect the popularity of the website. Because the related source text information corresponding to the Alexa ranking within the majority is in the top one million, therefore Alexa rankings can well reflect the popularity of source text information; outside sources of information for rankings in one million, the one millionth Alexa ranking. Based on getting pagerank’s value of the text information corresponding website and its position in the Alexa ranking, measure the importance and popularity of the source text information, and realize the credibility sorting of related information.

Prank represents the sorting of corresponding site’s importance degree site related text information, prank, represents \( r_i \) corresponding information sources on a Prank sort.

Using the BORDA. Footrule 2 sorts merge algorithm respectively merged prank and Arank, Prank Arank, and GArank, getting four kinds of text information credibility sorting CBrank, CFGrank, CFGrank, four sort’s ways of access as shown in table 1.

<table>
<thead>
<tr>
<th>Table 1: Access Obtain of Credibility Sorting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Footrule Algorithm is sorted merge algorithm based on distance. Assuming that exist for the same object collection s’s m full sort \( \text{rank}_1, \text{rank}_2, ..., \text{rank}_m \), Footrule sort merge method is looking for the sort of rank in a collection of objects, make the sorting to Footrule’s minimum distance. Footrule distance calculation formula from rank to \( \text{rank}_1, \text{rank}_2, ..., \text{rank}_m \) is shown below

\[ F(\text{rank}_1, \text{rank}_2, ..., \text{rank}_m) = \sum_{i=1}^{n} \frac{F(\text{rank}, \text{rank}_i)}{m} \] (5)
By building completed two parts graph \( G(V, E, W) \) to achieve rank. \( V \) consists of object collection \( s \) and ranking position \( p \), \( p = \{1, ..., |S|\} \). \( E \) for edge set from element of \( s \) in the element of \( P \). \( W \) for edge weight collection, computation formula of edge weight \( w(s, p) \) from node \( S \in S \) to \( P \in P \) is as follows

\[
\sigma(s, p) = \sum_{i=1}^{\infty} |\text{rank}(s) - p|
\]

Through looking for minimum cost exactly match in the figure to get rank.

### C. Reliable Judgment Models

Because the facts statement only have true and false, facts statement can be divided into two categories, the credible and uncertainty, so you can use the method of classification of facts credible screening. By choosing a certain number of facts as training set, the classification model, this model to determine the facts of the category to achieve the credible determine facts. Because the support vector machine (SVM) method is based on VC dimension theory and structure risk minimum principle in statistical learning theory, according to the limited sample information to find the best compromise between the complexity of the model and the ability of learn, has good generalization ability. So this paper presents facts statement credible decision method based on SVM classifier (SVM methods).

And basic methods are different, in the SVM method; think of all the relevant text information will affect the trusted facts. Facts corresponding related text information mainly contains the text features and text information source. Due to the characteristics above are reasonably considering contribute to the text information to the facts, so in the SVM classification method, \( R_{pos}, R_{neg} \) and \( R_{neu} \) text information on facts contribution value (\( \Delta_{pos}, \Delta_{neg} \) and \( \Delta_{neu} \) ) as classification feature. In order to avoid larger scope of characteristic value to influence the accuracy of classification, the characteristic value scope is mapped to \([-1, 1]\). For any facts, get the corresponding classification eigenvectors according to the related text information collection.

In use of the SVM classifier, the classification and the selection of kernel function can affect the classification accuracy, in this article the classification feature vector dimensions involved are relatively low, and the Gaussian RBF function is adopted as the kernel function of SVM classification, RBF function as shown in type (8). In the process of using RBF kernel function for classification model, involves the penalty parameter \( c \) and kernel parameter \( \sigma \) value, \( c \) and values will affect the classification of sigma new sample classification accuracy and generalization ability of the models. The optimal values of parameters \( c \) and \( \sigma \) through grid search method. By using a \( c \) and \( \sigma \) for \( M \) and \( N \) values, the \( M \ast N(c, \sigma) \) combination, training different support vector machine (SVM) respectively, to estimate the learning accuracy, select a highest combination of learning accuracy as the optimal parameters.

\[
K(y_i, y_j) = \exp\left(-\frac{\|y_i - y_j\|^2}{2\sigma^2}\right)
\]

In basic decision method and the SVM decision method, the facts of the determination results for reliable facts or uncertain credible facts (Uncertain). The facts have been judged to be uncertain credible reasons: (1) the facts itself is not credible; (2) the facts corresponding related text information on the quality and quantity can't support to its accurate judgment, for example, facts contain more information and for long-tail query. So for facts which has not been judged to be reliable, not arbitrarily think it’s not credible, but that need to be further determine.

### III. EXPERIMENTS AND SIMULATIONS

Obtain facts to constitute the experimental data from TRE2007. Randomly selected 30 trusted only answer facts and more than 20 credible answer facts constitute part data set credible facts in the TRE2007. Because in TRE2007, any credible truth has close and untrusted correspond to the facts, so select 50 corresponding unreliable facts as text information retrieval way: the facts as a search engine query, get relevant text information, referred as FQ way; With keyword set of facts as a search engine query, get relevant text information, referred as KQ. For any facts, through yahoo boss 2 respectively on FQ, KQ two ways to obtain 150 search results (related text message). In addition, 13 graduate students have a long time experience of the Internet conduct annotation of facts according to the text information and corresponding relation of text information. According to the annotation text information can be divided into support corresponding statement, against the corresponding statements and 3 classes which have nothing to do with the corresponding statement.

By conducting a series of experiments to verify the fact that the rationality and accuracy of MFSV credible evaluation model:

(a) Through experimental analysis in different ways of text information retrieval (FQ and KQ), containing corresponding facts on the semantic of the distribution of text information;

(b) Threshold \( \kappa \) values influence the classification of the relevant text information, thus affecting facts determination accuracy, obtained \( k \)’s optimal values through the experiment;

(c) The analysis under the different way of related text information retrieval, text information credibility sort distribution;

(d) Analysis text information quantity, threshold delta, credibility and determine ways of basic text information retrieval accuracy method;
(e) Analysis, reliability sorting and related text information acquisition way to influence the accuracy of the SVM decision method;

(f) Analysis and compare the differences of decision accuracy method with basic decision method and the SVM determination. The experiment is conducted on the Windows 7 platform with IntelCore2Quad2.66GHz processor, 2 GB of memory.

A. Text Information Distribution

This experiment analyzed the related text information acquisition in different way, in the text of the related information; include the corresponding facts ‘percentage of the text information. Due to the possibility of unreliable facts is lower than its corresponding text information, so only to the credible facts. The experiment in the figure, \( f_m \), \( f_m' \) respectively represent credible only answer facts and more answer facts, shooting rate represents text information contained its corresponding facts’ percentage in the total text information. Figures 2 and 3, respectively represent in the FQ (with facts as the search query, get relevant text message), KQ (keyword set of facts as a search query, get relevant text information) way for related text information, shot with the change of number of text messages shape. In figure 2, along with the increase of \( n \), the corresponding ratio of \( f_m \) and \( f_m' \) decreases, namely in the related text information, sort of information is more likely to contain the corresponding facts; Due to answer more facts have multiple credible comparing facts, so the \( f_m' \) corresponding ratio is always higher than that of \( f_m \) corresponding shooting. Figure 3 said related text information in the form of KQ, shooting change with the shape; it reflects the change tendency is similar to figure 2. As can be seen from the figure 2 and figure 3, in the text of the relevant information in KQ way, contains text information corresponding to the facts of the percentage significantly lower than the FQ mode contains the percentage accounted for the corresponding facts’ percentage is lower than in FQ mode. Because under the FQ, query involves more semantic information, the text quality of information the search engine return is higher. The experiment shows that the obtained text information quality under FQ way is a higher than under KQ way.

Figure 2. Change trend of shooting rate under FQ mode with \( n \)

B. Information Classification Parameter Estimation

According to the parameters \( \kappa \) and text information to make a text information classification of its corresponding facts statement’s similarity, thus combining text information annotation, \( k \)’s \( F_1 \) - measure values of different text information classification is available. When \( K \) value makes \( F_1 \) predominate - measure the maximum, then the value is optimal values of \( K \).

Figure 4 show that the \( \kappa \) takes different values in FQ and KQ for obtaining text information, the changes of values. From the figure 4 shows that the \( \kappa \) value is less than 0.8, \( F_1 \) - measure value under FQ mode is higher than KQ mode, because text information quality obtained under FQ mode is higher than the KQ mode for text information quality; under KQ, when \( \kappa = 0.55 \), corresponding \( F_1 \) – measure maximum is 0.52; under FQ, \( K = 0.51 \), the maximum value is 0.65, the \( K = 0.56 \), \( F_1 \) - measure value close to 0.65. Since the FQ or KQ way for text information, \( K = 0.56 \), the corresponding \( F_1 \) - measure value is close to or to the maximum, therefore, in the next experiment, take 0.55 as \( K \) values.

Figure 4. Change trends of \( F_1 \)-measure with \( \kappa \) parameters

C. The Text Information Credibility

The experiment analyze the distribution of credibility sort under two kinds of information access, FQ and KQ. Figure 5 and 6 respectively mean that under FQ and KQ mode, the number of text information is 150 (\( n = 150 \)), the distribution of \( CB_{Rank}, CB_{Grank}, CF_{Rank}, \)
When credibility sorting respectively $C_{Brank}$, $CB_{Grank}$, $C_{Fr ank}$, $CF_{Grank}$, the relationship between SVM method to determine accuracy and $n$.

We can see from the figure 7, in FQ way, the accuracy of the $C_{Fr ank}$ is higher than $CF_{Grank}$, $CB_{Grank}$ accuracy better than in $CB_{Grank}$ at the same time, the reason is that the SVM method considering the semantic similarity of its corresponding facts absolute value is less than the $K$ text information influence on facts judgment; When $n=90$, credibility sort is $C_{Fr ank}$, SVM methods accuracy reached its highest 0.83, from figure 8 shows that under the KQ way, $C_{Fr ank}$ credibility sorting still brings higher accuracy; Similarly, $CB_{r ank}$ performance will be better than $C_{Br ank}$; When $n=112$, credibility sort is $C_{Fr ank}$, high accurate to 0.74. From the experiment above shows that because the way’s quality of KQ mode for text information is low, FQ’S corresponding determine accuracy is relatively high of getting text information.

D. The SVM Methods Accuracy Assessment

Experiments using libsvm software package, choose RBF kernel function and the 3-fold cross-validation method to obtain the precision of the SVM method. Figure 7 and figure 8 respectively under the FQ and KQ, credibility sorting respectively $C_{Br ank}$, $CB_{Gr ank}$, $C_{Fr ank}$, $CF_{Gr ank}$, the relationship between SVM method to determine accuracy and $n$.

From table 2 shows that the SVM decision method and relative basic decision method has higher precision, the reason is that the SVM method considering the neutral text information (with the corresponding facts of absolute value is less than the similarity of kappa) predominate influence; FQ way can always bring a higher precision degree of determination, the reason is that FQ way for text information on quality is better than KQ mode for text information; By SVM methods considering the more information of text, making the SVM decision method and the basic methods to achieve the highest accuracy credibility of the sorting ’s order is not the same. Relative to decision method based on SVM, basic methods can be changed determination conditions by parameters adjustment, make the decision stricter or loose, better able to assist users to trust determine facts.

IV. CONCLUSION

This paper proposed facts MFSV credible evaluation model based on semantic similarity areas unrelated. The model gets and analyses relevant text messages of the fact waiting for determination by acquiring and analysis, to realize the credible decision to determine the facts. In the process of facts credible decision, considering the related
text information and facts, the semantic similarity between the text information and related factors such as the credibility, and on this basis, measuring the related text’s contribution to the degree of determining the facts, has realized the trusted fact determination. The judgment model does not need to specify the suspected unit of the facts, which makes the decision model, is suitable for the only answer to the facts and more credible answer facts.

Due to the fact statement does not include the degree of emotional and description, to deny the facts of trusted judgment, can pass on their corresponding facts certainly credible decision, thus helping credible decision to deny the facts, so this paper focuses on the judgment of certain facts. The quality of the relevant text information is to the premise accurately determine of the truth, when to determine facts is more complicated, the corresponding information quality is relatively low, it is difficult to make the right decision for the facts. So, in the future work, hope that through using decomposition of facts, such as rewriting technology, high quality related text information, so as to make the facts credible decision more accurate.

REFERENCES


