Available online at www.ijpe-online.com

International Journal of Performability Engineering

Vol. 13, No. 6, October 2017, pp. 878-885
DOI: 10.23940/ijpe.17.06.p9.878885

Recognition of Opinion Leaders in Micro-Blog based on Linked Data

Zhiyun Zheng, Pengfei Li, Xingjin Zhang, Dun Li*

School of Information Engineering,Zhengzhou University, Zhengzhou 450001, China

Abstract

In view of the lack of subjectivity and accuracy in the traditional micro-blog opinion leader recognition method to measure the important factors of users, a new micro-blog opinion leader recognition method is proposed. This paper used the linked data to describe the micro-blog data, used the association rule mining algorithm to quantitatively determine the important factors that affected the users' ranking, and constructed the opinion leader recognition model according to the index scoring method. Experiments show that our method using linked data identifies the opinion leaders the same as the standard leaders, is more accurate and has better feasibility than that of traditional data.

Keywords: linked data; micro-blog; opinion leader; association rule

(Submitted on July 25, 2017; Revised on August 30, 2017; Accepted on September 15, 2017)

(This paper was presented at the Third International Symposium on System and Software Reliability.)

1. Introduction

Micro-blog [4] is a social network platform for users to publish, share, distribute and acquire information, and is a kind of innovation to the mode of information diffusion, which plays an important role in the promotion of information. Different from Web1.0's traditional network media, online users become another important source of information in the new Web2.0 media represented by micro-blog. Due to the explosive growth of micro-blog users and the divergence of micro-blog information, micro-blog plays an increasing important role in the formation and development of public opinion. Opinion leaders are the most important media for information dissemination in political election, emergency communication and other social events.

The recognition of micro-blog opinion leaders is the research focus in social networking, but the study is still in its infancy. Massive micro-blog data, complex user relationships and a wide range of data attributes ensure the perfection of the micro-blog functions, but they also bring many obstacles to expression of information accurately. Therefore, we intend to use linked data [2,3,7] to express information and mining association rules, identify key factors, and build micro-blog opinion leader recognition model.

The content of the paper is arranged as follows. Section 2 introduces the related research of the opinion leader recognition. Section 3 establishes micro-blog linked data. Section 4 builds micro-blog opinion leader recognition model based on the association rules mining algorithm. Section 5 obtains micro-blogging data and experimental results analysis. Section 6 summarizes and prospects.

2. Related Studies

Opinion Leader was first proposed by the American scholar Lazarsfeld [8]. He believed that information is transmitted according to the mode from media opinion leaders to audience. Opinion leaders are often able to provide information for others in interpersonal communication network, influence people and play an intermediary and filtering role in the mass communication.

* Corresponding author.
E-mail address: ielidun@zzu.edu.cn.

At present, the common research on opinion leaders recognition is divided into two categories: social network structure mining and index scoring.

- Social network structure mining methods are generally based on the relationship between users to build social networks, and then use the network structure algorithm to analyze the link relations between users, calculate the importance of the user ranking, and then the front in the sequence is the opinion leader. Weng proposed TwitterRank method of the influential user discovery in the specific topic based on the PageRank [11]. Xiao Yu presented LeaderRank's opinion leader discovery algorithm adding emotion weight in the PageRank [12]. These methods are the improvement of PageRank algorithm and have some limitations in the micro-blog platform.

- Index scoring method is based on the characteristics of opinion leaders. Through the characteristics analysis of the opinion leaders, the corresponding index system were established and marked. Vergani chose three indicators of social identity, expertise and social capital [5]. Ding Hanqing built some indicators including centrality, activity, cohesion, and infectivity [9]. According to the characteristics of micro-blog, Liu Zhiming established a relatively complex index system, including the influence of the two levels of active indicators, in the first level there were two indicators of influence and activity, in the second level there were 7 indicators of the number of repost, the number of comment, the number of mention, the number of original, the number of reply to own blog, the number of reply to others blog and the number of active days [6]. After the establishment of the index system, one method is that they determine the weight through hierarchical analysis [10] and calculate the weighted average to get the final result; another is to construct the scoring model directly and calculate the evaluation function to get the final result. However, the indexes in these methods are based on qualitative analysis of existing research results and lack of objectivity.

In this paper, the association rules mining algorithm and index scoring method are combined to determine the index quantitatively and identify the opinion leaders of micro-blog.

3. Linking the Micro-blog Data

With micro-blog's massive data and the complex relations, the linked data is simple and effective. The complete expression of micro-blog data in the semantic information and relations avoids loss of information and redundancy, and provides help for information sharing and other applications. At present, there is no unified, authoritative linked micro-blog data. Drawing on the idea of Seven Stages ontology construction method, this section establishes micro-blog linked data using Protege [10] ontology editing tools in four aspects of the concept, attributes, relationships and examples.

3.1. Concepts

Around the micro-blog's attention, point praise, forwarding, comments and other functions, the core contains three aspects such as user, blog and topics. Blogs and topics can be divided into 48 and 45 categories through labels. But these categories will make the blogs and topics classification too similar and too large, not the best classification. If we make the labels as the property of blogs and the topics, it will cause the inconvenience of data input. As a result, we take the labels as the micro-blog concepts and establish the relations between labels and blogs and topics to express explicitly. According to the different ways of certification, users can be divided into four categories: individual authentication users, the official certification users, self-certified users and ordinary users. According to the blog richness, blog can be divided into four categories, which contain pictures blogs, video blogs, pictures and video blogs and pure text blogs.

3.2. Attributes

Attributes in linked data are used to describe conceptual properties and relationships. Attributes can be divided into data attributes and object relationship attributes. The data attribute describes the characteristic and basic information of the concept. The object relational attribute describes the relationship between two concepts. In micro-blogs, the user's data attributes include age, gender, location, number of blog posts, educational information, job information, number of friends and number of followers. Blog's data attributes include the number of likes, the number of comments, the number of reposts and time. The data attributes of the topic include the number of discussions and the number of readings. Object relationship attributes include collection, comment, release, friend, follower, inclusion, like, mention, share, similar, create and join. The properties of micro-blog are shown in Figure 1, 2, 3, and 4.

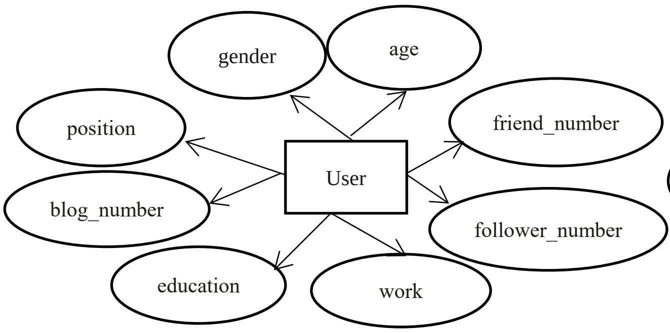


Figure 1. User data attribute graph

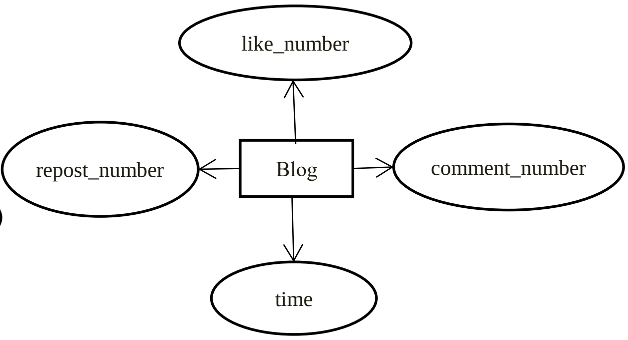
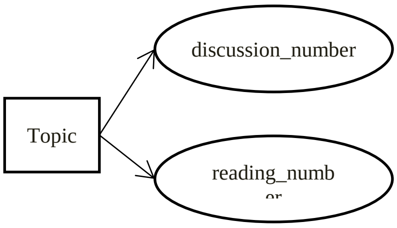


Figure 2. Blog data property diagram



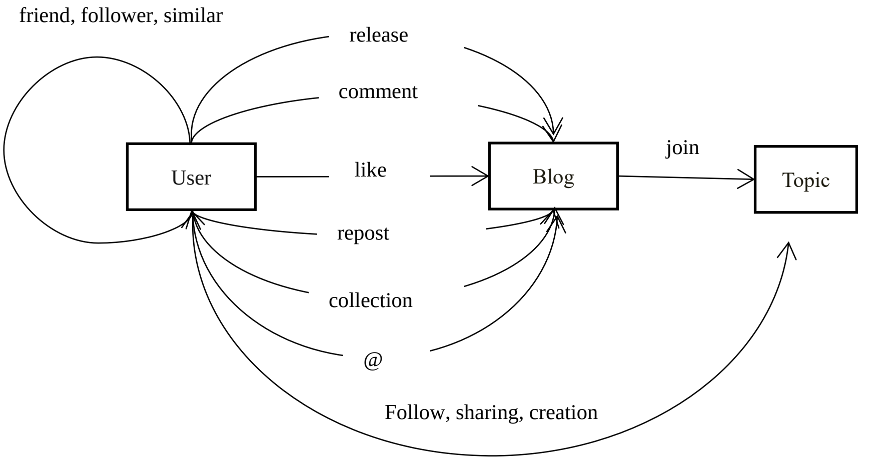
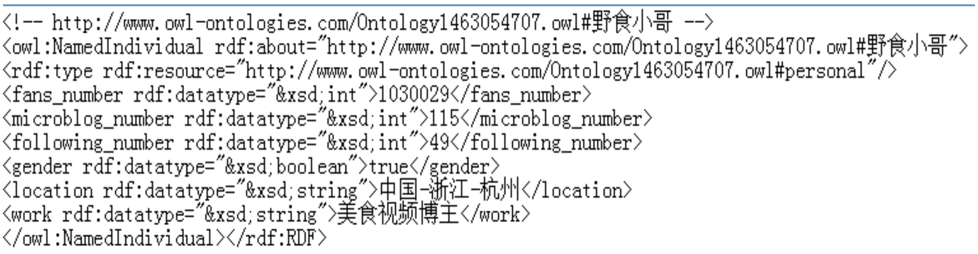


Figure 3. Topic data attribute graph

Figure 4. Micro-blog object relationship attribute map



```
<!-- http://www.owl-ontologies.com/Ontology1463054707.owl#野食小哥 -->
<owl:NamedIndividual rdf:about="http://www.owl-ontologies.com/Ontology1463054707.owl#野食小哥">
<rdf:type rdf:resource="http://www.owl-ontologies.com/Ontology1463054707.owl#personal"/>
<fans_number rdf:datatype="&xsd;int">1030029</fans_number>
<microblog_number rdf:datatype="&xsd;int">115</microblog_number>
<following_number rdf:datatype="&xsd;int">49</following_number>
<gender rdf:datatype="&xsd;boolean">true</gender>
<location rdf:datatype="&xsd;string">中国-浙江-杭州</location>
<work rdf:datatype="&xsd;string">美食视频博主</work>
</owl:NamedIndividual></rdf:RDF>
```

Figure 5. RDF data instance fragment

3.3. Examples and Associations

The object class or concept describes the abstract with the same characteristic, data attribute and object attribute description. The object attribute provides the range and type of the attribute, while the instance is a concrete manifestation of the concept. Many instances form a concept. Examples are concrete and clear. An instance has all the attributes of the delegated concept. The instance is concrete and explicit. For example, “野食小哥、男、中国-浙江-杭州、美食视频博主、1030029、49、115” is a user instance. The instance is the main part of the information and the concrete description which is our direct object of the research. The reasonable instances construction could give the real and effective meaning of the data. The constructed RDF instance fragment is shown in Figure 5.

4. Opinion Leaders Recognition

Different to the previous subjective opinion leader recognition model of index factors, we use the linked data and the association rule to mine the key factors, and combine with traditional index scoring method to identify opinion leaders of micro-blog.

4.1. Transaction of Linked Data

Focusing on RDF data, we extend the SPARQL syntax to define the mining model, generate the corresponding item sets according to the mining model, and the transaction following. Then, the association rules mining algorithm is used to generate the association rules. The mining model consists of two parts: the target concept set and the item sets. The target concept set refers to the set of the concepts that can be related through the relevant attributes in the model. The item set is used to specify the origin of the item set that is a set of attributes. Our mining model is Q = (*User*, *{Properties}*), the target concept *User* represents the micro-blog user, and the item sets concept *Properties* represents all the attributes in the micro-blog linked data.

With a series of instances, the project forms the item set and generates a transaction. The instance project comes from the stored information. A fragment of an instance corresponding to the micro-blog linked data is shown in Figure 6.

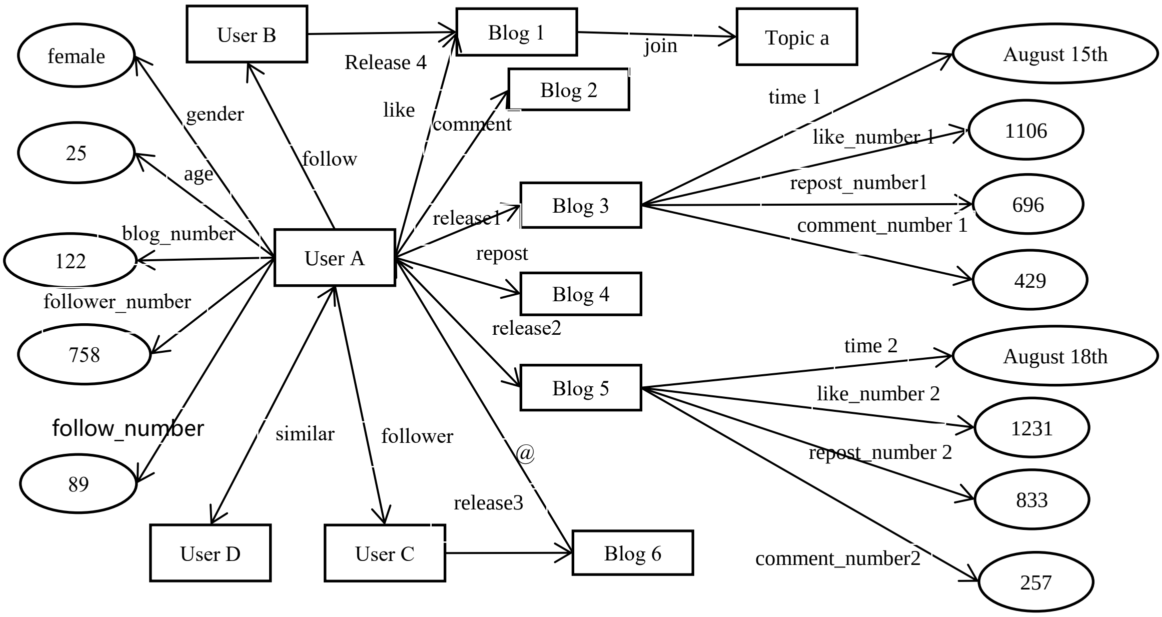


Figure 6. Micro-blog user instance fragment

The instance fragment shown in Figure 6 provides the example fragment to illustrate the process of data conversion. It includes all information about user *A*, such as the basic attributes of the user A, the relevant information of the user and blog that related with the user *A*. In the depth-first algorithm, we start from the user-specified target concept *A*, stop until meeting the user-specified attribute, record the transaction in triple, and then continue in another path.

The information in the graph is converted into triples of *subject-path-object*. As shown in Table 1, the tuple generation algorithm is described as follows.

Algorithm 1: triples generation

```
Begin
input the instance snippet
visit user A vertex (denoted as vertex v), visited [v] = 1
w = the first adjacent point of vertex v
while (w exists)
{ if (w belongs to the attribute && w is not visited) then
 execute the algorithm recursively beginning from the vertex w;
w = next adjacent point of vertex v;
}
End
```

Table 1. The information of example User A

ID	subject	composition path	object	feature
1	User A	（follow, release4, join）	Topica	join
2	User A	（follow, release4）	blog1	release
3	User A	（follow）	UserB	follow
4	User A	（like, join）	Topica	join
5	User A	（comment）	blog2	comment
6	User A	（release1, time1）	August 15th	time
7	User A	（release1）	blog3	release
8	User A	（release1, like_number1）	1106	Like_number
9	User A	（release1, repost_number1）	696	repost_number
10	User A	（release1, comment_number1）	429	comment_number
11	User A	（repost）	blog4	repost
12	User A	（release2, time2）	August 18th	time
13	User A	（release2）	blog5	release
14	User A	（release2, like_number2）	1231	like_number
15	User A	（release2, repost_number2）	833	repost_number
16	User A	（release2, comment_number2）	257	comment_number
17	User A	（follower, release3）	blog6	release
18	User A	（follower）	UserC	follower
19	User A	（similar）	UserD	similar
20	User A	（gender）	female	gender
21	User A	（age）	25	age
22	User A	（blog_number）	122	blog_number
23	User A	（follower_number）	758	follower_number
24	User A	（follow_number）	89	follow_number

The triple information in Table 1 produces from Figure 6. The first triple starts from the target user A through the path of follow-release-join, encounters the corresponding value that belongs to the attribute, and then the first triple is generated. The depth-first algorithm is to generate the longest path along the paths, and then generate its sub paths in turn. The second triple is sub-path of the first path and stopped the search until encountering user-specified properties. And so on, 24 triples are generated which forms the initial instance set. In the next step, the final transaction is generated based on the instance set. The process of generating the transaction set from an instance item set is to merge the instance items of the same sub-path in the triple according to the attribute set specified by the user, into a transaction. The transactions generated from Table 1 are shown in Table 2.

Table 2. The transaction of instance User A

ID.	transaction
1	{follow.release.join.topica,follow.release.blog1,follow.UserB}
2	{like.join.topica}
3	{comment.blog2}
4	{release. August 15th,release.blog3,release.like_number→1106,release.repost_number→696,release. comment_number→429}
5	{repost.blog4}
6	{release. August 18th,release.blog5,release.like_number→1231,release.repost_number →833, release.comment_number→257}
7	{follower.release.blog6,follower.UserC}
8	{similar.UserD}
9	{gender→female}
10	{age→25}
11	{blog_number→122}
12	{follower_number→758}
13	{follow_number→89}

In Table 2, the first transaction is the combination of first, second and third triples; the fourth transaction is the combination of the 6th to 10th triples; the sixth transaction is the combination of 12th to 16th triples; the seventh transaction is a combination of 17th, 18th triples.

4.2. Data Standardization

Since the information described in the linked data is very detailed, some data must be pre-processed to the data with statistical characteristics to reflect the relation with the whole, rather than individual's contingency. For example, the micro-blog number of the user is very large and the number of reviews to every micro-blog must be different, so a micro-blog's data does not illustrate the problem, and we need calculate the average and variance of all of the comments.

Table 3. Attributes indicators and corresponding methods

Attributes indicators	Processing method	Remark
Followers number	Subsection	
Number of authenticated users in followers	Count	
Follower interaction rate	Calculate the ratio	The proportion of followers who participated in the blog interaction
Friends number	Subsection	
Number of authenticated user in friends	Count	
Age	Subsection	
Micro-blog frequency	Calculate the frequency	Calculated by the time and quantity of Micro-blog
Number of Micro-blog repost	Calculate the average	The average number of repost numbers of all blogs issued by the user
Number of Micro-blog comment	Calculate the average	The average number of comment numbers of all blogs issued by the user
Number of Micro-blog like	Calculate the average	The average number of like numbers of all blogs issued by the user
Micro-blog richness	Calculate the ratio	Measured in the form of blogs content

4.3. Recognition of Option Leader

Using the Apriori algorithm [1] generates association rules from the attribute indicators that become opinion leaders, and searches in the frequent item set layer by layer. Firstly, we scan all the transaction records and find the frequent 1-item sets according to the defined support threshold, marked as "M1". And so on, we find frequent N-item sets until we find the largest frequent item sets. At last, association rules are generated and the irrelevant rules are filtered. The rules that meet the requirements are shown in Formula 1.

$$X \rightarrow \text{opinion leaders} \tag{1}$$

Where X is the index set, the rules in the set are the key indicators to measure the users, and the consequence is the opinion leader. We adopt the Analytic Hierarchy Process (AHP) to assign weight of each indicator: input the user data, mark users' attributes with indicator scoring method, calculate the weighted average, and recognize the opinion leaders. Algorithm 2 is shown below.

Algorithm 2: the opinion leader recognition

```
Begin
input the sample data
threshold←x
call Apriori algrithm
i←0
while(i< number of association rules)
{if consequent of association rules =' opinion leader'
then {key index set←key index set + rule antecedent
   i←i+1}}
call Analytic Hierarchy Process to weight values for each key index
input user data
call Indicate Scoring to score the user attributes and calculate the weighted average
if score > threshold then the user is the opinion leader
End
```

The efficiency of the algorithm is decided by the Apriori algorithm. We analysis the algorithm and get the computation complexity of Apriori algorithm including 3 parts:

a) The production of the frequent itemsets is O(Nw), where N is the number of the transactions and w is the average width of the transcations.

b) The production of the candidates is $O(C_{k=2}^{w}k(k-2)C_k)$, where k is the item number in the frequent itemsets.

c) The calculation of the support is $O(N_k C_w^k \alpha_k)$, where the transaction with |t| length could produce C_t^k k-itemsets, and α_k is the consuming of support count in updating the candidate k-itemset in the Hash tree.

5. Experiments and Results Analysis

The experimental environment is Jena 2.6, Eclipse, the hardware configuration is Intel (R), Core (TM), i3-2100, @3.10G Hertz CPU, 4GB memory. The experimental data come from Sina Microblogs. The raw data transform into experimental data in two stages of acquisition and screening. In the acquisition stage, Octopus is used to obtain the user's background information and interactive information between users. In the filter phase, the invalid data are deleted. There are 2000 users including a total of 593,909 attribute information. The core flow chart of the experiment is shown in Figure 7.

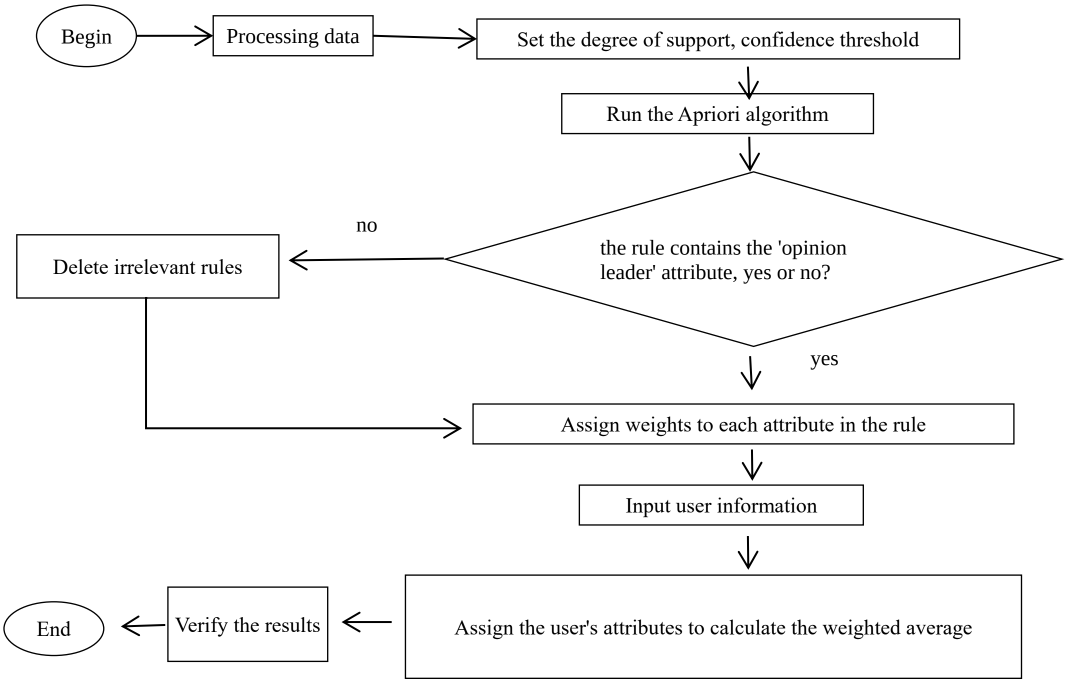

Begin
Processing data
Set the degree of support, confidence threshold
Run the Apriori algorithm
the rule contains the 'opinion leader' attribute, yes or no?
no
Delete irrelevant rules
yes
Assign weights to each attribute in the rule
Input user information
Assign the user's attributes to calculate the weighted average
Verify the results
End

Figure 7. the Core Process of Algorithm

The key indexes and their weights are shown in Table 4.

Table 4. Key indexes and weights

index	weight
Follower_number	0.2175
Number of authenticated users in followers	0.1537
users' category	0.1678
Follower interaction rate	0.1460
Number of Micro-blog repost	0.1372
Number of Micro-blog comment	0.0889
Number of Micro-blog like	0.0889

The *Micro-blog Fengyun list* (*Hall of Fame*) is user value and data services platform authorized by Sina micro-blog officials. This platform uses a large distributed system to store, update, and analyse micro-blog data, and adopts a set of complex evaluation algorithms to measure user value and influence. We use the data and rankings provided by the *micro-blog Fengyun list* as the experimental evaluation criteria and comparison. The top ten opinion leaders identified in the paper are shown in Table 5.

Table 5. Ranking of opinion leaders

User's Nickname	Our ranking	*Fengyun List* ranking
头条新闻	1	1
央视新闻	2	4
人民日报	3	5
何炅	4	2
新浪娱乐	5	6
微博搞笑排行榜	6	7
李开复	7	3
南方周末	8	8
当时我就震惊了	9	9
头条新闻	10	10

As shown in Table 5, the two ranking orders are nearly the same. Among them, there are 9 authentication users and 1 non-authentication user. That shows that certification has a certain impact on the user to become opinion leaders, but not entirely decisive role. Among the top-10, there are 5 individual users, 5 official users, and the order is basically consistent with the *micro-blog Fengyun List*. Only two users's rank declined, 何炅 and 李开复. The top-10 contains 3 public Figures recognized by the masses of the national and authoritative users because the dissemination of information has certain influence. Other non-public Figures, the majority of users of authoritative media, indicate the masses for the confidence of the official platform. The experimental results show the feasibility of the opinion leader recognition method that based on the associated data.

6. **Conclusions**

Aiming at subjective selection index and the lack of objectivity in the option leader recognition of the traditional method, we constructed Micro-blog linked data and proposed an opinion leader mining method based on the associated data. Firstly, the depth first algorithm is used to pre-process the linked data. Secondly, we used the Apriori algorithm to select the objective influencing factors. Finally, we combined with the index scoring method to recognize the opinion leaders. Experiment shows that the opinion leaders recognized in the paper are nearly consistent with the standard leaders, which is more accurate than the traditional data and possesses feasibility and accuracy.

Acknowledgments

The authors are grateful to the editors and reviewers for their helpful comments and suggestions. This work was partially supported by National Social Science Foundation project (17BXW065), Science and Technology Research project of Henan province (14430051007) and Science and Technology Research project of Zhengzhou(141PPTGG368).

References

1. T. Bernerslee, J. Hendler, O. Lassila. “The Semantic Web,” *Scientific American*, vol. 284, no. 5, pp. 34-43, 2001.
2. C. Bizer, T. Heath, T. L. Berners. “Linked Data-the Story so Far”. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pp. 205-227, 2009.
3. “China Internet Development Statistics Report”. http://www.cnnic.net.cn/gywm/xwzx/rdxw/201708/t20170804_69449.htm, Last accessed on August 4, 2017.
4. H. Q. Ding, Y. P. Wang. “Analysis of *Opinion Leaders* Characteristics in SNS Cyberspace: a Case Study of Douban.com”. *Journalism & Communication*, no. 3, pp. 82-91, 2010
5. J.Y. Guo, Z. B. Zhang, Q. Y. Sun. “Study and Applications of Analytic Hierarchy Process”. *China Safety Science Journal* (CSSJ), no. 5, pp. 26, 2008.
6. “Linked data”, Available at https://www.w3.org/DesignIssues/LinkedData.zh-cn.html, last accessed on September 23, 2017
7. P. F. Lazarsfeld, B Berelson, H Gaudet. “The Peoples’ Choice: How the Voter Makes up His Mind in a Presidential Campaign”. 1968.
8. Z. M. Liu, L. Liu. “Identification and Analysis of Opinion Leaders in Internet Public Opinion”. *Journal of Systems Engineering*, vol. 29, no. 6, pp. 8-16, 2011.
9. “protégé”. Available at https://protege.stanford.edu/, last accessed on September 23, 2017
10. M. Vergani. “Are Party Activists Potential Opinion Leaders?”. *Javnost-The Public*, vol. 18, no. 3, pp. 71-82, 2011.
11. J. Weng, E. Lim, J. Jiang, and Q. He, “TwitterRank: Finding Topic-sensitive Influential Twitters,” in Proceedings of the Third ACM International Conference on Web Search and Data Mining, 2010.
12. Y. Xiao, W. Xu, L. Xia. “A Network Group Opinion Leader Recognition Algorithm Based on Emotional Tendency Analysis”. *Computer Science*, vol. 39, no. 2, pp. 34-37, 2012.