Analysis of Association Degree Algorithm Based on Complex Network Public Opinion

Jing Luo and Guoqing Xu

Abstract—**The network public opinion plays an irreplaceable role in today's society. The analysis and processing of data mining can effectively find out the sensitive network public opinion and prevent the intensification of social public opinion. In this paper, we propose an association degree algorithm between personal speech and user real name information in Chinese community. By defining the user's name, address, ID number, phone number, QQ number, e-mail, MSN and other information as keywords and setting different weights. The hot public opinion phrases in user comments are extracted, the word frequency is counted, the text semantic similarity classification model is constructed, the semantic tree of unknown text is automatically constructed, the semantic relevance based on the topic is calculated, and the relevance between the public opinion topic and the specified user information is obtained. The algorithm in this paper uses Chinese segmentation index technology and common algorithms of text de duplication technology. From the technical feasibility, economic feasibility and other aspects of the feasibility analysis; from the system response speed, scalability and security three main levels of the system function requirements analysis. The results show that the system design meets the requirements and improves the accuracy of online public opinion text collection and emotional analysis.**

Index Terms—**Application in classification, big-data analysis, clustering, FP-growth algorithms.**

I. Introduction

The purpose of this paper is to use web crawler tools to collect online public opinion information from some Chinese communities, and use data mining technology to find out the relationship between public opinion topics and designated users from the resource collection. Using the different association rules of different personal information and users of the same user, the association degree between these users and public opinion topics is calculated and sorted.

At present, classification algorithms at home and abroad mainly focus on how to improve the accuracy of nearest neighbor classification, reduce the computational complexity and select the appropriate distance metric function. [1] K-nearest neighbor algorithm can improve the performance of nearest neighbor classification, overcome the influence of outliers on the classification performance of small sample data set, and improve the classification accuracy of classification algorithm [2]. K-nearest neighbor algorithm, also known as KNN algorithm, is the simplest algorithm in data mining technology [3]. The working principle of KNN is that given a training data set of known label categories, after inputting new data without labels, K instances closest to the new data are found in the training data set [4]. If most of these K instances belong to a certain category, then the new data belongs to this category [5]. It can be simply understood as: the k points nearest to x determine which category x belongs to [6]. At present, the k-nearest neighbor classification algorithms at home and abroad mainly focus on how to improve the accuracy of nearest neighbor classification, reduce the computational complexity, and select the appropriate distance metric function [7]. It can improve the performance of nearest neighbor classification, overcome the influence of outliers on the classification performance of small sample data set, and improve the classification accuracy of classification algorithm [8]. When the number of samples is large or tends to be infinite, the nearest neighbor classification has higher classification accuracy, but when the sample size is limited, the nearest neighbor classification may produce larger classification error [9]. The classification performance of KNN classifier is very sensitive to the neighborhood size k value. To solve this problem, Mitani and Hamamoto proposed a local mean based k-nearest neighbor (lmknn) classification algorithm [10]. Its classification idea is to use the local mean of each nearest neighbor to determine the category of the sample to be tested [11]. Because the mean value of k nearest neighbors is used in the classification judgment of samples to be tested, the lmknn classification algorithm has achieved good results in the processing of outliers in small samples [12]. Since the lmknn algorithm was proposed, its idea has been successfully applied to distance measure learning, discriminant analysis and group classification [13]. But when the sample size is limited, the nearest neighbor classification may produce larger classification error [14]. The classification performance of KNN classifier is very sensitive to the neighborhood size k value [15].

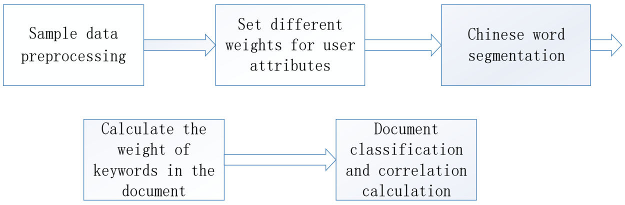

Sample data preprocessing
Set different weights for user attributes
Chinese word segmentation
Calculate the weight of keywords in the document
Document classification and correlation calculation

Fig. 1. Overall design process.

In order to solve this problem, TF-IDF (term frequency - inverse document frequency) is a common weighting technology for information retrieval and text mining [16]. TF-IDF is a statistical method to evaluate the importance of a word to a document set or one of the documents in a corpus [4]. The importance of a word increases in proportion to the number of times it appears in the document, but at the same

Manuscript received July 9, 2020; revised September 23, 2020.

Jing Luo is with the Wuhan Institute of Technology, China (e-amil: 971672513@qq.com).

doi: 10.18178/ijfcc.2020.9.3.566

time it decreases inversely with the frequency of its appearance in the corpus. The main idea of TF-IDF is: if the frequency of a word in one article is high and rarely appears in other articles, it is considered that the word or phrase has a good classification ability and is suitable for classification [17].

II. Sample Data Preprocessing

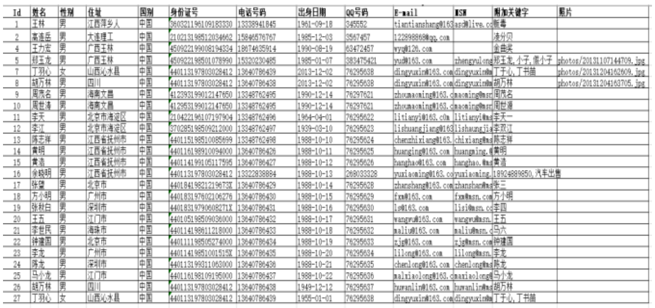
A. Review Stage

Sample data preprocessing is divided into two parts: the first part is the preprocessing of public opinion resources, the second part is the preprocessing of user's personal information form.

The preprocessing of public opinion resources mainly includes: (1) invalid HTML document filtering; (2) HTML document noise cleaning.

When filtering invalid HTML documents, because HTML documents are crawled by web crawler tools, invalid web pages may be crawled. If the page cannot be accessed, there are a large number of invalid HTML documents with the content of "404 < HTML >" in the given public opinion resources, which will be filtered by python programming. In addition, some HTML documents contain only numbers and no body content, but there are many numeric attributes in user information. For example: ID card number, telephone number, etc., we do not exclude the possibility of some user information in these HTML documents.When cleaning the noise of HTML documents, because the text in public opinion resources is usually described by segmented text, there is usually no large number of hyperlinks in the middle. Through the analysis of the captured HTML documents, the body information of the web page can be obtained. The specific purification operations are as follows: programming with python, removing the symbols such as ",", "|", ", N / R", ">", and "+" In addition, considering that English information of users may appear in the document, such as e-mail, MSN and numerical information such as ID number, telephone number, etc., we do not filter out the numerical information and English information in the document as noise information.

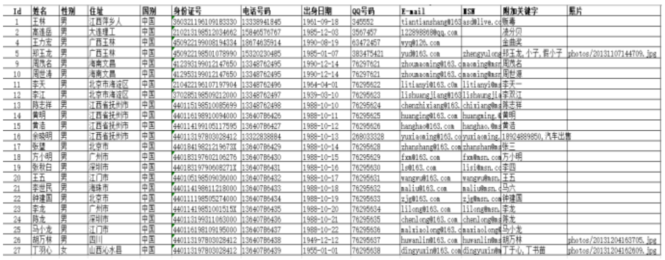
The preprocessing of user's personal information form mainly includes: (1) invalid information processing; (2) duplicate record processing. For invalid information processing, in the original Excel data table, the country attribute value of all users is "China". Therefore, this attribute does not work for the analysis of the association degree between a given user and public opinion resources, and it can be manually deleted.



Id	姓名	性别	住址	国别	身份证号	电话号码	出身日期	QQ号码	E-mail	MSN	附加关键字	照片
1	王林	男	江西萍乡人	中国	360321196109183330	13338941845	1961-09-18	345552	tiantianshang@163	asd@live.co	贩毒	
2	高连岳	男	大连理工	中国	210213198512034662	15846576767	1985-12-03	3567457	1228988680@qq.com		凌分贝	
4	王力宏	男	广西玉林	中国	450922199008194334	18674635914	1990-08-19	63472457	wyq@126.com		金曲奖	
5	郑玉龙	男	广西玉林	中国	450922198501078990	15320230485	1985-01-07	383475421	yud@163.com	zhengyulong	郑玉龙,小子,假小子	photos/20131107144709.jpg
7	丁羽心	女	山西沁水县	中国	440113197803028412	13640786439	2013-12-02	76295638	dingyuxin@163.com	dingyuxin@m	丁于心,丁书苗	photos/20131204162609.jpg
8	胡万林	男	四川	中国	440113197803028412	13640786438	2013-12-02	76295638	dingyuxin@163.com	dingyuxin@m	胡万林	photos/20131204163705.jpg
9	周茂名	男	海南文昌	中国	412393199012147650	13348762495	1990-12-14	76297621	zhoumaoming@163.c	maoming@msn	周茂名	
10	周世涛	男	海南文昌	中国	412953199012147650	13348762495	1990-12-14	76297621	zhoumaoming@163.c	maoming@msn	周世源	
11	李天	男	北京市海淀区	中国	210422196107197904	13348762496	1964-04-01	76295622	litianyi@163.c0m	litianyi@ms	李天一	
12	李江	男	北京市海淀区	中国	370285198509212000	13348762497	1939-03-10	76295623	lishuangjiang@163	lishaungjia	李双江	
13	陈志祥	男	江西省抚州市	中国	440115198510085699	13348762498	1988-10-10	76295624	chenzhixiang@163.	chixiang@ms	陈志祥	
14	黄明	男	江西省抚州市	中国	440116198910094000	13640786426	1988-10-11	76295625	huanging@163.com	huangming.@	黄明	
15	黄浩	男	江西省抚州市	中国	440114199105117595	13640786427	1988-10-12	76295626	hanghao@163.com	hanghao.@ms	黄浩	
16	余晓明	男	江西省抚州市	中国	440113197803028412	13322838884	1988-10-13	268033328	yuxiaoming@163.co	yuxiaoming.	18924889850,汽车出售	
17	张望	男	北京市	中国	44018419821219673X	13640786429	1988-10-14	76295628	zhanshang@163.com	zhanshan@ms	张三	
18	方小明	男	广州市	中国	440183197602106276	13640786430	1988-10-15	76295629	fxm@163.com	fxm@msn.com	方小明	
19	张秋白	男	深圳市	中国	44018319790608271X	13640786431	1988-10-16	76295630	ls@163.com	lisi@msn.co	李四	
20	王五	男	江门市	中国	440105198509036000	13640786432	1988-10-17	76295631	wangwu@163.com	wangwu@msn.	王五	
21	李世民	男	海珠市	中国	440114198611218000	13640786433	1988-10-18	76295632	maliu@163.com	maliu@msn.c	马六	
22	钟建国	男	北京市	中国	440111198505274000	13640786434	1988-10-19	76295633	zjg@163.com	zjg@msn.com	钟建国	
23	李龙	男	广州市	中国	44011419851001515X	13640786435	1988-10-20	76295634	lilong@163.com	lilong@msn.	李龙	
24	陈龙	男	深圳市	中国	440113199311063000	13640786436	1988-10-21	76295635	chenlong@163.com	chenlong@ms	陈龙	
25	马小龙	男	江门市	中国	440116198109195000	13640786437	1988-10-22	76295636	malxiaolong@163.c	maxiaolong@	马小龙	
26	胡万林	男	四川	中国	440113197803028412	13640786438	1949-12-12	76295637	huwanlin@163.com	huwanlin@ms	胡万林	
27	丁羽心	女	山西沁水县	中国	440113197803028412	13640786439	1955-01-01	76295638	dingyuxin@163.com	dingyuxin@m	丁于心,丁书苗	

Fig. 2. Original excel data sheet.

When processing duplicate records, ID8 and id26 represent the same user, id7 and id27 represent the same user. This duplicate information will affect the results of association analysis, so it is necessary to manually delete two lines of duplicate information. It is found that id7 and ID8 have the same ID card number, date of birth, QQ number, e-mail and MSN. It can be seen that these two lines of data can not represent the two users very well. However, id26 and id27 lack "photo" information, so the "photo" information in id7 and ID8 is manually added to id26 and id27, and the two lines of id7 and ID8 are manually deleted.



Id	姓名	性别	住址	国别	身份证号	电话号码	出身日期	QQ号码	E-mail	MSN	附加关键字	照片
1	王林	男	江西萍乡人	中国	360321196109183330	13338941845	1961-09-18	345552	tiantianshang@163	asd@live.co	贩毒	
2	高连岳	男	大连理工	中国	210213198512034662	15846576767	1985-12-03	3567457	1228988680@qq.com		凌分贝	
4	王力宏	男	广西玉林	中国	450922199008194334	18674635914	1990-08-19	63472457	wyq@126.com		金曲奖	
5	郑玉龙	男	广西玉林	中国	450922198501078990	15320230485	1985-01-07	383475421	yud@163.com	zhengyulong	郑玉龙,小子,假小子	photos/20131107144709.jpg
9	周茂名	男	海南文昌	中国	412393199012147650	13348762495	1990-12-14	76297621	zhoumaoming@163.c	maoming@msn	周茂名	
10	周世涛	男	海南文昌	中国	412953199012147650	13348762495	1990-12-14	76297621	zhoumaoming@163.c	maoming@msn	周世源	
11	李天	男	北京市海淀区	中国	210422196107197904	13348762496	1964-04-01	76295622	litianyi@163.c0m	litianyi@ms	李天一	
12	李江	男	北京市海淀区	中国	370285198509212000	13348762497	1939-03-10	76295623	lishuangjiang@163	lishaungjia	李双江	
13	陈志祥	男	江西省抚州市	中国	440115198510085699	13348762498	1988-10-10	76295624	chenzhixiang@163.	chixiang@ms	陈志祥	
14	黄明	男	江西省抚州市	中国	440116198910094000	13640786426	1988-10-11	76295625	huanging@163.com	huangming.@	黄明	
15	黄浩	男	江西省抚州市	中国	440114199105117595	13640786427	1988-10-12	76295626	hanghao@163.com	hanghao.@ms	黄浩	
16	余晓明	男	江西省抚州市	中国	440113197803028412	13322838884	1988-10-13	268033328	yuxiaoming@163.co	yuxiaoming.	18924889850,汽车出售	
17	张望	男	北京市	中国	44018419821219673X	13640786429	1988-10-14	76295628	zhanshang@163.com	zhanshan@ms	张三	
18	方小明	男	广州市	中国	440183197602106276	13640786430	1988-10-15	76295629	fxm@163.com	fxm@msn.com	方小明	
19	张秋白	男	深圳市	中国	44018319790608271X	13640786431	1988-10-16	76295630	ls@163.com	lisi@msn.co	李四	
20	王五	男	江门市	中国	440105198509036000	13640786432	1988-10-17	76295631	wangwu@163.com	wangwu@msn.	王五	
21	李世民	男	海珠市	中国	440114198611218000	13640786433	1988-10-18	76295632	maliu@163.com	maliu@msn.c	马六	
22	钟建国	男	北京市	中国	440111198505274000	13640786434	1988-10-19	76295633	zjg@163.com	zjg@msn.com	钟建国	
23	李龙	男	广州市	中国	44011419851001515X	13640786435	1988-10-20	76295634	lilong@163.com	lilong@msn.	李龙	
24	陈龙	男	深圳市	中国	440113199311063000	13640786436	1988-10-21	76295635	chenlong@163.com	chenlong@ms	陈龙	
25	马小龙	男	江门市	中国	440116198109195000	13640786437	1988-10-22	76295636	malxiaolong@163.c	maxiaolong@	马小龙	
26	胡万林	男	四川	中国	440113197803028412	13640786438	1949-12-12	76295637	huwanlin@163.com	huwanlin@ms	胡万林	photos/20131204163705.jpg
27	丁羽心	女	山西沁水县	中国	440113197803028412	13640786439	1955-01-01	76295638	dingyuxin@163.com	dingyuxin@m	丁于心,丁书苗	photos/20131204162609.jpg

Fig. 3. Excel table after preprocessing.

III. Set Different Weights for Each Attribute

The weight of user attributes is set based on analytic hierarchy process. Because the attributes of user such as name, address, ID card number, telephone number, QQ number, e-mail, MSN and other attributes have different degrees of association with users, the emergence mode of these information in public opinion resource collection also indirectly reflects the relationship between resources and users. Therefore, before analyzing the association degree between public opinion resources and users, it is necessary to set different weights to prove the association rules of various attributes of users. This paper uses analytic hierarchy process (AHP) to determine the weight of the user's name, address, ID number, telephone number, QQ number, e-mail, MSN and other attributes to the user through comparison. This is a quantitative process.

Analytical steps of analytic hierarchy process. First of all, the hierarchical structure model is established. In this paper, we only need to determine the weight of each user's attribute to the user. Therefore, the user's name, address, ID number, telephone number, QQ number, e-mail, MSN and other attributes are taken as the criteria layer, which are $C1$, $C2$, $C3$ ……$C11$. Take the user as the target layer 0.Then, a pair of comparison matrix A is formed to compare the attributes of users by using the relative size to prove the criteria C1, $C2$, $C3$……$C11$. Importance to goal 0. Where AIJ represents the relative importance of Ci to Cj. Like: Ci: $Cj \rightarrow aij$

$$A=(aij)\ n\times n,\ aij>0,\ aji=1/(aij) \tag{1}$$

The comparison scale is: Saaty *et al.* Proposed 1-9 scale: AIJ value 1,2 ……9 and its reciprocal number 1,1/2……1/9, which is convenient for qualitative to quantify transformation. Then, the arithmetic average method of weight is used to determine the weight of each influencing factor. The weight of each effective judgment matrix is calculated. This can be attributed to the problem of calculating the maximum eigenvalue and eigenvector of judgment matrix. The calculation methods include: sum, root and power. When the accuracy requirement is not high, the use of summation, the root of the square can meet the requirements of practical application. When the accuracy requirement is high, the power method and the square root method can be used to calculate the product *Mi* of each row of elements in the judgment matrix, as shown in the following formula:

$$Mi= \prod_{j=1}^{n} Cij \ (i=1,2,\ldots\ldots,n) \tag{2}$$

Calculate the *n-th* root of MI

$$Wi= \sqrt[n]{Mi} \tag{3}$$

Normalize the vector, as shown in the formula:

$$Wi=Wi/\sum_{j=1}^{n} Wj \tag{4}$$

That is to say, the feature vector is the corresponding weight coefficient

IV. Chinese Word Segmentation

Using the analysis technology based on ICTCLAS, word segmentation technology is the process of recombining continuous word sequences into word sequences according to certain specifications, so as to establish indexes for them. Chinese word segmentation is the basic work of script mining. In this paper, ictbras, a Chinese word segmentation tool developed by the software room of the Institute of computer science of Chinese Academy of Sciences, is used to segment HTML and TXT documents. ICTCLAS has the functions of fast segmentation, high precision, new word recognition and user dictionary support, which basically solves the problem of Chinese word segmentation.

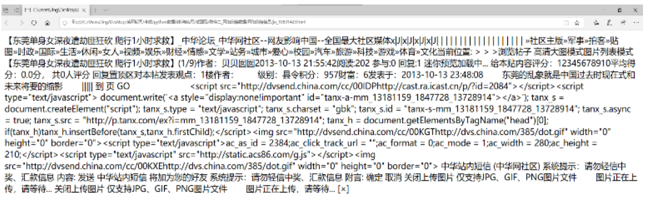
Specific steps of Chinese word segmentation. The set of part of speech and proper name category labels are shown in the table below, including 24 part of speech tags (lower case letters) and 4 proper name category labels (capital letters):

This paper uses Jieba 0.42.1 to download: https://pypi.org/project/jieba/

After decompressing, enter the directory and run python setup.py Install Python code, interface components only provide jieba.cut Method is used for word segmentation. The cut method takes two input parameters. The first parameter is the string to be segmented, cut_ The all parameter is used to control the word segmentation mode. The string to be segmented can be GBK string, UTF-8 string or Unicode, jieba.cut The returned structure is an iteratable generator. You can use the for loop to get each word (Unicode) after word segmentation, or you can use list（ jieba.cut (...)) to list. Chinese text word segmentation is realized, such as Chinese word segmentation for the following original HTML document:

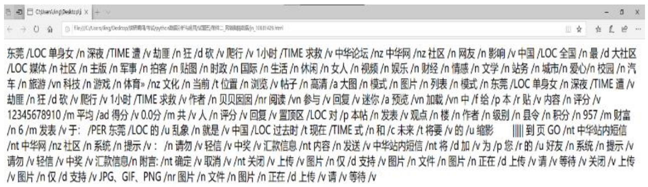
TABLE I. Part of Speech Category

tag	meaning	tag	meaning	tag	meaning
n	Common nouns	*f*	Location NOUN	*s*	Place NOUN
nr		*ns*	place name	*nt*	Organization name
nz	name	*v*	Common verbs	*vd*	verbal adverb
a		*ad*	Adverbial words	*an*	Noun form words
m	Other proper names	*q*	classifier	*r*	pronoun
c		*u*	auxiliary word	*xc*	Other function words
PER	adjective	LOC	place name	ORG	Organization
t	time	*vn*	Noun verb	*p*	preposition
nw	Title of the work	*d*	adverb	*w*	punctuation



```
【东莞单身女深夜遭劫匪狂砍 爬行1小时求救】_中华论坛_中华网社区--网友影响中国--全国最大社区媒体x|J|x|J|x|J|x|J||||||||||||||||||»社区主版»军事»拍客»贴
图»时政»国际»生活»休闲»女人»视频»娱乐»财经»情感»文学»站务»城市»爱心»校园»汽车»旅游»科技»游戏»体育»文化当前位置: > > >浏览帖子 高清大图模式图片列表模式
【东莞单身女深夜遭劫匪狂砍 爬行1小时求救】(1/9)作者: 贝贝囡囡2013-10-13 21:55:42阅读:202 参与:0 回复:1 迷你预览加载中... 给本贴内容评分: 12345678910平均得
分: 0.0分, 共0人评分 回复置顶区对本帖发表观点: 1楼作者:      级别: 县令积分: 957财富: 6发表于: 2013-10-13 23:48:08      东莞的乱象就是中国过去时现在式和
未来将要的缩影      ||||| 到 页 GO              <script src="http://dvsend.china.com/cc/00IDPhttp://cast.ra.icast.cn/p/?id=2084"></script><script
type="text/javascript"> document.write('<a style="display:none!important" id="tanx-a-mm_13181159_1847728_13728914"></a>'); tanx_s =
document.createElement("script"); tanx_s.type = "text/javascript"; tanx_s.charset = "gbk"; tanx_s.id = "tanx-s-mm_13181159_1847728_13728914"; tanx_s.async
= true; tanx_s.src = "http://p.tanx.com/ex?i=mm_13181159_1847728_13728914"; tanx_h = document.getElementsByTagName("head")[0];
if(tanx_h)tanx_h.insertBefore(tanx_s,tanx_h.firstChild);</script><img src="http://dvsend.china.com/cc/00KGThttp://dvs.china.com/385/dot.gif" width="0"
height="0" border="0"><script type="text/javascript">ac_as_id = 2384;ac_click_track_url = "";ac_format = 0;ac_mode = 1;ac_width = 280;ac_height =
210;</script><script type="text/javascript" src="http://static.acs86.com/g.js"></script><img
src="http://dvsend.china.com/cc/00KXEhttp://dvs.china.com/385/dot.gif" width="0" height="0" border="0"> 中华站内短信 (中华网社区) 系统提示: 请勿轻信中
奖、汇款信息 内容: 发送 中华站内短信 将加为您的好友 系统提示: 请勿轻信中奖、汇款信息 附言: 确定 取消 关闭上传图片 仅支持JPG、GIF、PNG图片文件      图片正在上
传, 请等待... 关闭上传图片 仅支持JPG、GIF、PNG图片文件      图片正在上传, 请等待... [×]
```

Fig. 4. Original web page



```
东莞 /LOC 单身女 /n 深夜 /TIME 遭 /v 劫匪 /n 狂 /d 砍 /v 爬行 /v 1小时 /TIME 求救 /v 中华论坛 /nz 中华网 /nz 社区 /n 网友 /n 影响 /v 中国 /LOC 全国 /n 最 /d 大社区
/LOC 媒体 /n 社区 /n 主版 /n 军事 /n 拍客 /n 贴图 /n 时政 /n 国际 /n 生活 /n 休闲 /n 女人 /n 视频 /n 娱乐 /n 财经 /n 情感 /n 文学 /n 站务 /n 城市/n 爱心/n 校园 /n 汽
车 /n 旅游 /vn 科技 /n 游戏 /n 体育» /nz 文化 /n 当前 /t 位置 /n 浏览 /v 帖子 /n 高清 /a 大图 /n 模式 /n 图片 /n 列表 /n 模式 /n 东莞 /LOC 单身女 /n 深夜 /TIME 遭 /v
劫匪 /n 狂 /d 砍 /v 爬行 /v 1小时 /TIME 求救 /v 作者 /n 贝贝囡囡 /nr 阅读 /vn 参与 /v 回复 /v 迷你 /a 预览 /vn 加载 /vn 中 /f 给 /p 本 /r 贴 /v 内容 /n 评分 /v
12345678910 /m 平均 /ad 得分 /v 0.0分 /m 共 /v 人 /n 评分 /v 回复 /v 置顶区 /LOC 对 /p 本帖 /n 发表 /v 观点 /n 楼 /n 作者 /n 级别 /n 县令 /n 积分 /n 957 /m 财富
/n 6 /m 发表 /v 于:  /PER 东莞 /LOC 的 /u 乱象 /n 就是 /v 中国 /LOC 过去时 /t 现在 /TIME 式 /n 和 /c 未来 /t 将要 /v 的 /u 缩影     ||||| 到 页 GO /nt 中华站内短信
/nt 中华网 /nz 社区 /n 系统 /n 提示 /v :  /n 请勿 /v 轻信 /v 中奖 /v 汇款信息 /nt 内容 /n 发送 /v 中华站内短信 /nt 将 /d 加 /v 为 /p 您 /r 的 /u 好友 /n 系统 /n 提示 /v
请勿 /v 轻信 /v 中奖 /v 汇款信息/n 附言: /nt 确定 /v 取消 /v /nt 关闭 /v 上传 /v 图片 /n 仅 /d 支持 /v 图片 /n 文件 /n 图片 /n 正在 /d 上传 /v 请 /v 等待 /v 关闭 /v 上传
/v 图片 /n 仅 /d 支持 /v JPG、GIF、PNG /nr 图片 /n 文件 /n 图片 /n 正在 /d 上传 /v 请 /v 等待 /v
```

Fig. 5. After Chinese word segmentation.

Weight of keywords in the document. In this paper, TF-IDF (term frequency – inverse document frequency) method is adopted, which is a common weighting technology for information retrieval and text mining [18]. A corpus is used to evaluate the importance of a word set in a document [19]. The importance of a word increases in proportion to the number of times it appears in the document, but at the same time it decreases inversely with the frequency of its appearance in the corpus [20]. Various forms of TF-IDF weighting are often used by search engines as a measure or rating of the correlation between files and user queries [21].

The main idea of TF-IDF is: if a word or phrase appears in one article with high frequency of TF and rarely appears in other articles, it is considered that the word or phrase has good classification ability and is suitable for classification. TF-IDF is actually: TF * IDF [22].

1) Term frequency (TF) refers to the frequency of a given word appearing in the file. The ratio of the number of times the word w appears in document D is count (W, d) and the total number of words (d) in document D.

$$tf(w, d) = \text{count}(w, d) / \text{size}(d) \tag{5}$$

This number is a normalization of term count to prevent it from skewing to long files. (the same word may have a higher number of words in a long file than in a short file, regardless of whether the word is important or not.)

2) Inverse document frequency (IDF) is a measure of word universal importance. The IDF of a specific word can be obtained by dividing the total number of files by the number of files containing the word, and then taking logarithm of the quotient. That is, the logarithm of the ratio of the total number of documents n to the number of documents in the word w (DOC (W, d)).

$$idf = \log(n / \mathrm{docs}(w, D)) \quad (6)$$

TF-IDF according to TF and IDF for each document D and keywords w [1] The query string Q composed of w [k] calculates a weight, which is used to represent the matching degree between query string Q and document D

$$tf\text{-}idf(q, d) = \mathrm{sum}\ \{\ i = 1..k \mid tf\text{-}idf(w[i], d)\ \} = \mathrm{sum}\ \{\ i = 1..k \mid tf(w[i], d) * idf(w[i])\ \} \quad (7)$$

The high word frequency in a particular file and the low file frequency of the word in the whole file set can produce TF-IDF with high weight. Therefore, TF-IDF tends to filter out common words and retain important words.

Specific steps of keyword weight calculation in the document. In Python, there is an API to calculate TF-IDF under scikit learn package, and the effect is also very good. First, scikit clean must be installed. For different system installation, please refer to: http://scikit-learn.org/stable/install.html

1) Install the scikit learn package (install the dependency package first, and then install sklearn). Install it through pip to verify whether the installation is successful,

2) Calculate TF-IDF. Scikit learn package mainly uses two classes for TF-IDF word segmentation weight calculation: countvectorizer and tfidftransformer. Where countvectorizer is through fit_ The transform function converts the words in the text into the word frequency matrix, and the matrix element a [i] [J] represents the word frequency of J word in the i-th text. That is, the number of times each word appears_ feature_ Names () can see the keywords of all texts, and toArray () can see the result of word frequency matrix.

Document classification and correlation calculation. The main goal of this mining is to get the association degree between each user and the whole public opinion resources. In this paper, we understand the relevance degree as: from the public opinion resource collection, we mine the number of documents that match each user. It is equivalent to taking each user related to public opinion as a category and classifying all public opinion documents. According to its proportion in the whole public opinion resource set, the association degree between each user and the resource set is represented and sorted.

Firstly, several concepts are introduced: support degree; confidence degree; promotion degree.

1) Support: support can be understood as the current popularity of an item. The calculation method is as follows:

Support = (number of records including item a) / (total number of records) (8)

Take the supermarket record above as an example. There are five transactions, and milk appears in three transactions, so the support of {milk} is 3 / 5. The support of {egg} is 4 / 5. The number of times milk and eggs appear at the same time is 2, so the support of {milk, egg} is 2 / 5.

2) Confidence: confidence means that if you buy item a, you are more likely to buy item b. The calculation is as follows:

Confidence ($a \rightarrow b$) = (number of records containing items a and b) / (number of records containing a)

For example: we already know that the number of times to buy (milk, eggs) together is two times, and the number of times to buy eggs is 4 times. Then the confidence (milk > egg) is calculated as confidence (milk > egg) = 2 / 4.

3) Lift: promotion refers to the increase in sales rate of another item when one item is sold. The calculation method is as follows:

Promotion ($a \rightarrow b$) = confidence ($a \rightarrow b$) / (support a) (9)

For example, we have calculated the confidence level of milk and eggs (confidence) = 2 / 4. If the support of milk is 3 / 5, then we can calculate the support of milk and eggs: lift = 0.83. When the value of promotion degree (a > b) is greater than 1, the more items a sell, the more B will sell. A promotion of 1 means that there is no correlation between product a and product B. Finally, if the promotion degree is less than 1, it means that the purchase of a will reduce the sales volume of B.

Specific steps of document classification and association calculation. This paper uses Apriori algorithm, we need to provide two parameters, data set and minimum support. Apriori recursively traverses all item combinations. First, traverse the case of one item combination, remove the data items whose support is lower than the minimum support, and then use the remaining items to combine. After traversing the two item combinations, the combinations that do not meet the conditions are eliminated. Recursion continues until there are no more items to combine.

We take the ID number of each user as its category number, and then read it into each document associated with the user in the public opinion resource collection. Reading the document is to calculate the score that belongs to each user category. If the highest score is higher than the threshold set by us, it belongs to the category of the user; otherwise, it belongs to the "other" category. Finally, the ratio of the number of documents contained in each category to the number of all documents in the whole public opinion resource collection is taken as the association degree of each user to the public opinion resources, and is arranged in descending order. When the threshold is set to 0.6, only the set with support greater than 0.6 is the frequent itemset.

V. Conclusion

The results show that the correlation between users and public opinion resources is: *id*16 > *id*17 > *id*18 > *id*26 > *id*5 > *id*9 > *id*11 > *id*12 > *id*21 > *id*27 > *id*14 > *Id*1 > *id*20 > *id*25 >

$id4 > id22 > id23 > id24$

Apriori algorithm finds out the frequent itemsets in the example, and several concepts that will be used in the processing, such as support degree, confidence degree and promotion degree. Then, this paper mainly introduces how to find the frequent itemsets of items efficiently by using Python based on the text tendentiousness association analysis.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Jing Luo, Guoqing Xu conducted the research; JingLuo analyzed the data; JingLuo wrote the paper; all authors had approved the final version.

References

[1] L. Yun, “Design and implementation of network public opinion analysis system based on crawler and text clustering analysis,” University of Electronic Science and Technology, pp. 22-26, may 2014.

[2] Z. H. Zhu, “Application research on network public opinion monitoring based on text tendentiousness association analysis,” North China Electric Power University, pp. 20-26, March 2018.

[3] Y. Hao, “One belt, one road social network public opinion spatial semantic association analysis,” China University of Mining and Technology, pp. 30-37, May 2018.

[4] J. X. Du and C. Jing, “Data collection and text classification technology analysis of network public opinion monitoring,” *Wireless Internet Technology*, pp. 123-124, September 2019.

[5] W. Chao, “Analysis of association network structure of unexpected network public opinion events in China,” *Modern Intelligence*, pp.121-130, December 2019.

[6] H. X. Ma, “Research on several pattern classification methods based on K-nearest neighbor criterion,” Shaanxi Normal University, pp. 49-58, June 2018.

[7] *Computer Learning Practice*, People's Posts and Telecommunications Press, pp. 562-571.

[8] Z. H. Zhou, *Machine Learning*, Tsinghua University Press, pp. 351-435.

[9] J. X. Zhang, H. X. Sun, Z. X. Sun, and W. C. Dong, “Reliability assessment of wind power converter considering scada multistate parameters prediction using fp-growth, wpt, k-means and lstm network,” *Journal of Engineering*, 2020.

[10] Y. L. Jia, L. Liu, H. Chen, and Y. H. Sun, “A Chinese unknown word recognition method for micro-blog short text based on improved FP-growth,” *Pattern Analysis and Applications*, 2020.

[11] L. Hamdad, Z. Ournani, K. Benatchba, and A. Bendjoudi, “Two-level parallel CPU/GPU-based genetic algorithm for association rule mining,” *International Journal of Computational Science and Engineering*, 2020.

[12] A. Augello, I. Infantino, G. Pilato, and F. Vella, “Sensing the web for induction of association rules and their composition through ensemble techniques,” *Procedia Computer Science*, 2020.

[13] Y. W. Fang, H. L. Huang, T. D. Li, and J. B. Wang, “Fast data mining algorithm based on FP growth association rules,” *Journal of Chongqing University of Technology*, pp. 191-194, 2020.

[14] Z. J. Zhang, J. Huang, J. G. Hao, J. X. Gong, and H. Chen, “Extracting relations of crime rates through fuzzy association rules mining,” *Applied Intelligence*, 2020.

[15] J. W. Li, N. Yu, J. W. Jiang, X. Li, Y. Ma, and W. D. Chen, “Research on student behavior inference method based on fp-growth algorithm,” *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2020.

[16] X. Gao, F. Q. Xu, and Z. M. Zhu, “The application of improved FP-growth algorithm in disease complications,” in *Proc. 2019 International Conference on Computational Modeling, Simulation and Optimization*, 2019.

[17] C. Yang, “Research on parallel acceleration algorithm of association rules based on Hadoop,” Nanjing University of Posts and Telecommunications, pp.15-23, 2019.

[18] W. E. Chao, “Research on frequent pattern mining algorithm based on compact pattern tree and multiple minimum support,” Xi'an University of Technology, pp. 3-7, 2019.

[19] X. L. Zhu and Y. G. Liu, “An efficient frequent pattern mining algorithm using a highly compressed prefix tree,” *Intelligent Data Analysis*, 2019.

[20] S. C. Hu, “Research and improvement of Apriori algorithm,” Qingdao University, pp.27-32, June 9, 2019.

[21] G. Quan, “Research and application of parallel FP growth mining algorithm based on cloud computing platform,” Nanjing University of Aeronautics and Astronautics, pp. 3-5, January 1, 2018.

[22] L. Afuan, A. Ashari, and Y. H. Suyanto, “Building the electronic evidence analysis model based on association rule mining and FP-growth algorithm,” *International Journal of Advanced Computer Science and Applications*, 2019.

Jing Luo was born in Hubei, China, in 1995. In 2017, he received a bachelor's degree in electrical information engineering from Wuhan University of Engineering. He is currently working on a master's degree in computer technology from Wuhan University of Engineering.

His current research interest is national language processing (NLP).