面向网络舆情的关联度分析

摘　要：

随着互联网的迅速发展，网络舆情对政治秩序秩序与社会稳定有着非比寻常的作用，因此挖掘网络用户信息与网络舆情信息的关联度的过程意义重大。编制 java 语言，在关键词词频统计的基础上，给予用户信息中的关键词赋予比重不同的权重，实现在网络舆情资料库中对 web 信息提取与关键词提取的功能。将用户信息中的关键词转化为向量 A，向量的分量为对应的关键词的权重，同时使用二值法将网络舆情信息库中的网页转为向量 B，向量的分量为对应关键词的出现与否，将向量 B 中的各分量分别乘于相应的权重值，得向量 C。求向量 A 与向量 C 的余弦值并由此构建关键词加权评分系统，完成对相关网页的评分高低排序并归档。

关键词：java，关键词赋权，余弦值

The thesis title

Abstract:，

With the rapid development of Internet,network public opinion on the political order and social stability has had an extraordinary effect，so excavating the association between the network users and the network public opinion information is significant. Preparing the java language,on the base of keywords frequency statistics,given to keywords in the information user different weights,in order to achieve the function of extracting web information and keywords in the network public opinion database. Changing the keywords which are in the user information into a vector A, the component of the vector A corresponding to the weight of the keywords.Meanwhile,using a binary method to change a web page which are in the network public opinion database into the vector B,the vector component of the corresponding keyword the presence or absence,and multiply corresponding weight with the vector B,we can get a vector C.Finding cosine vectors A and C,and thus the vector construct Keywords weighted scoring system,the completion of the relevant pages of score level sorting and archiving.

Key words: java，giving weight to keywords， cosine value

目　录

1. 挖掘目标

本次建模目标是利用网络爬虫工具采集的大量网络舆情信息，采用数据挖掘技术，给予用户信息中的关键词赋权，分析用户信息中不同权重大小的关键词与网络舆情信息间的关系，挖掘用户与网络舆情信息的关联规则，使用关键词加权评分系统构建反映与用户最相关的网络舆情信息。从而可以发挥网络对社会监督的巨大作用，以及及时有效的处理网络舆情突发事件。

2. 分析方法与过程

2.1. 总体流程

本用例主要包括如下步骤：

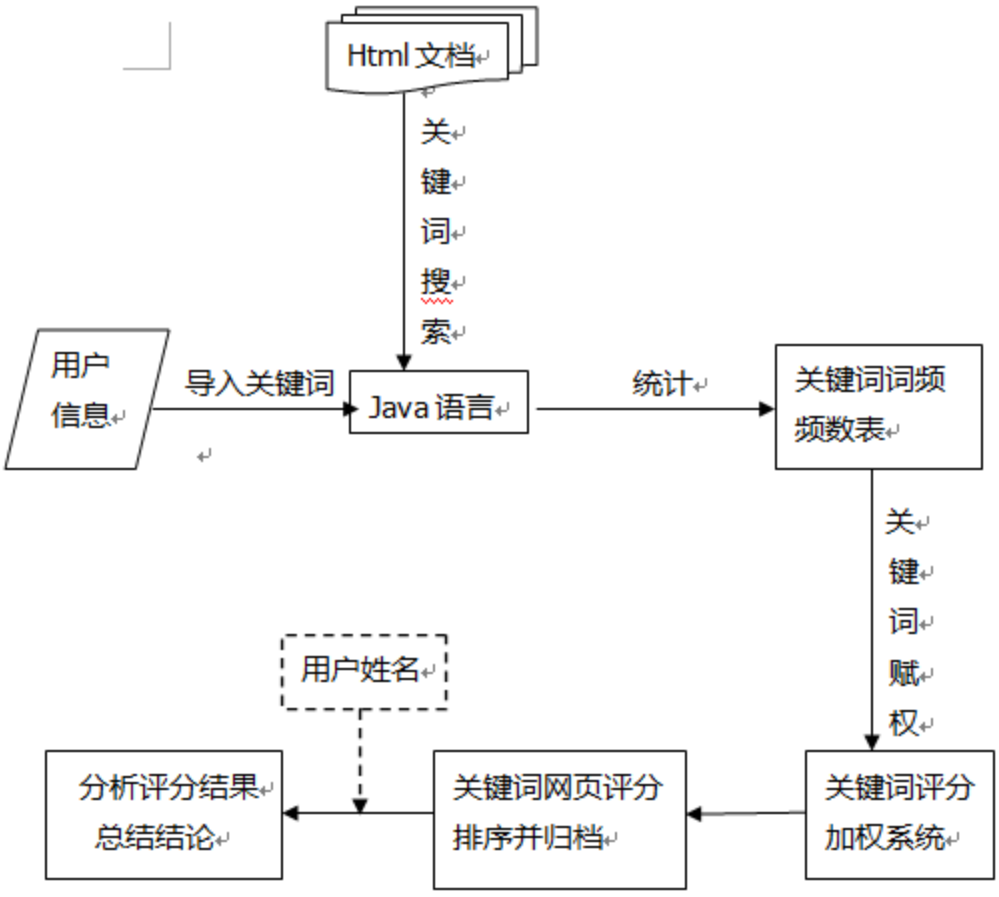

图1

步骤一：编写 java 语言，在网络舆情信息资料库中进行 web 信息提取与关键词提取，对十项关键词（去除二个样本无差别变量：国别，照片），进行“or”逻辑规则搜索，得到相应关键词的网页地址。

步骤二：在关键词词频统计的基础上，设置用户信息中十项关键词的权重。

步骤三：基于二值法优化构建关键词加权评分系统(详细流程见图 2)，将净化出关键词的网页进行评分。

步骤四：对网页的评分高低排序并归档，完成关联度分析。

2.2. 具体步骤

步骤一：web 信息提取与关键词提取

- **样本预处理**

在用户信息的十二项关键词中，“国别”关键词在样本中均为“中国”，无法区别不同样本之间的差别，所以予以剔除。“照片”关键词只在三个样本中出现，相对全部样本而言样缺失值过多，所以予以剔除。

编写 java 语句，使其可以输入关键词并遍历“网路舆情信息”文件夹。对关键词使用“or”逻辑规则完成对“网络舆情信息”的搜索。“or”逻辑规则即搜索时只要含有任意一个关键词即符合搜索条件，可输出。

搜索结果如下：

关键词	姓名	性别	住址	身份证号	电话号码
频数	94	201812	426	10	159

关键词	出生日期	QQ 号码	E-mail	MSN	附加关键字
频数	1	159	2	2	296

表一

由上表可知，在性别与住址频数中存在大量干扰信息，所以进一步修正搜索规则，对性别，住址单独出现的网址，予以剔除。

修正搜索结果如下：

关键词	姓名	性别	住址	身份证号	电话号码
频数	94	306	109	10	159

关键词	出生日期	QQ 号码	E-mail	MSN	附加关键字
频数	1	159	2	2	296

表二

步骤二：关键词的权重设置

根据表二中的关键词频数，计算关键词权重。

关键词出现的频数越多，提供的信息的干扰性越强，有效性越低，因此笔者将出现频数越大的关键词设置较小的权重。

计算方法如下：

1. 单个关键词权重系数=$\frac{\text{所有关键词频数之和-单个关键词频数}}{\text{所有关键词频数之和}}$

2. 关键词权重=$\frac{\text{单个关键词权重系数}}{\sum \text{单个关键词权重系数}}$

例：姓名权重系数=$\frac{\text{所有关键词频数之和-姓名频数}}{\text{所有关键词频数之和}}=\frac{1138-94}{1138}=0.9174$

姓名权重=$\frac{\text{姓名权重系数}}{\sum \text{单个关键词权重系数}}=\frac{0.9174}{9}=0.1091$

将其余关键词代入上式公式，得关键词权重表如下：

关键词	姓名	性别	住址	身份证号	电话号码
权重	0.1091	0.0812	0.1005	0.1101	0.0956

关键词	出生日期	QQ 号码	E-mail	MSN	附加关键字
权重	0.111	0.0956	0.1109	0.1109	0.0822

表三

步骤三：二值法优化构建关键词加权评分系统

1. 将十项关键词权重转换为十维向量：

∂=（姓名，性别，住址，身份证号，电话号码，
出生日期，QQ号码,E-mail,MSN,附加关键字）
=（0.1091，0.0812，0.1005，0.1101，0.0956，
0.111，0.0956,0.1109,0.1109,0.0822）

2. 将单个网页中的关键词频数转换为十维向量：

   β_{ij}=（姓名，性别，住址，身份证号，电话号码，

   出生日期，QQ号码,E-mail,MSN,附加关键字）

   $i=1, \cdots,27$，即用户信息.xls中的样本个数。

   $j=$“网络舆情信息”文件夹中网页个数

3. 运用二值法将向量 β_{ij} 各分量中频数大于 0 的值记为 1，等于 0 仍记为 0。得出向量 β_{ij}^{1}。

4. 将向量 β_{ij}^{1} 中的各分量分别乘于相应的权重值，得向量 β_{ij}^{*}。

   例：某网页含有关键词频数结果：

关键词	姓名	性别	住址	身份证号	电话号码
频数	10	1	0	0	0

关键词	出生日期	QQ 号码	E-mail	MSN	附加关键字
频数	0	0	0	0	0

   则转换为向量 β^{1}=(1,1,0,0,0,0,0,0,0,0)，与关键词权重表（表三）相应分量相乘得出向量 β*=(0.1091,0.0812,0,0,0,0,0,0,0,0)。

5. 对向量 ∂ 与向量 β_{ij}^{*} 求余弦值，即 $\cos\theta_{ij}=\dfrac{\vec{\partial}\bullet\vec{\beta}_{ij}^{*}}{|\partial|\bullet|\beta_{ij}^{*}|}$。

二值法优化构建的关键词加权评分系统流程图如下：

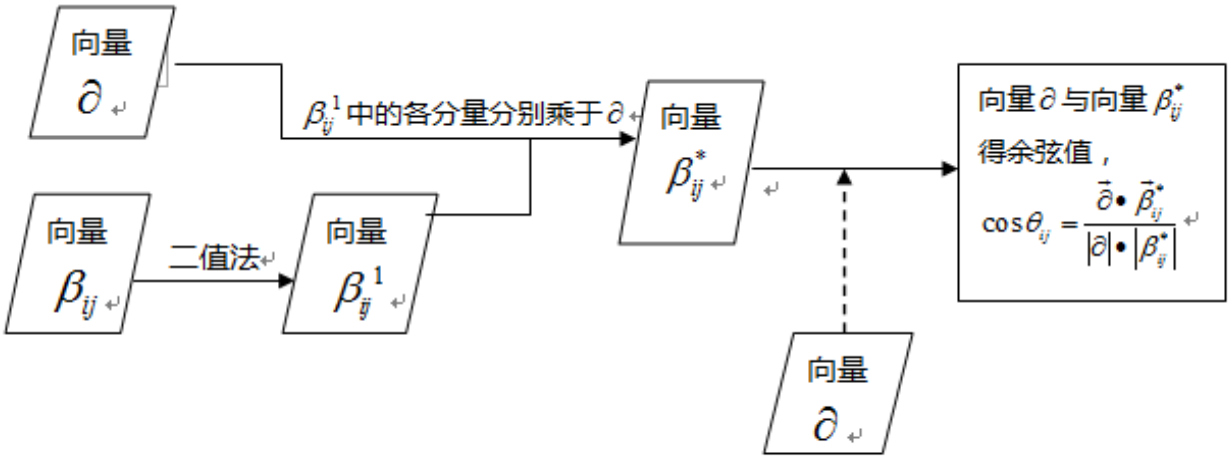

图2

步骤四：评分排序并归档

1. 将步骤三中的 $\cos\theta_i$ 计算出结果并从高至低排序，同时为了数据直观表现，笔者将 $\cos\theta_i$ 的数值放大 100 倍。
2. 建立分档规则：70 分及以上为优；50~70 分为良；30~50 分为中;低于 30 分为差
3. 将数据按分档规则进行分档。

2.3. 结果分析

经计算结果如下：

1. 用户信息中的“**高连岳，周茂名，周世涛，陈志祥，黄浩**”五个样本信息无法找到与“网络舆情信息”文件夹中的网址有相匹配的关键词出现，则算得的 $\beta^1=(0,0,0,0,0,0,0,0,0,0)$ [此处的 β^1 与步骤三中的 β_{ij}^1 意义相同]。

高连岳				
档次	向量β¹	原始等分	得分	网址
无符合条件的网址				
周茂名				
档次	向量β¹	原始等分	得分	网址
无符合条件的网址				
周世涛				
档次	向量β¹	原始等分	得分	网址
无符合条件的网址				
陈志祥				
档次	向量β¹	原始等分	得分	网址
无符合条件的网址				
黄浩				
档次	向量β¹	原始等分	得分	网址
无符合条件的网址				

表四

因此向量 ∂ 与向量 β^* 的之间余弦值为 0，无法对上述用户信息进行评分排序并归档。

2. 用户信息中的“**王林**”在“网络舆情信息”文件夹中找到 3 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表五所示，根据步骤四中的评分标准与归档规则，“**王林**”与该 3 个网页的关联程度为“**差**”等。

王林			
档次	向量β¹	得分	网址
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10833726.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10834540.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10835078.html

表五

3. 用户信息中的“**王力宏**”在“网络舆情信息”文件夹中找到 1 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表六所示，根据步骤四中的评分标准与归档规则，“**王力宏**”与该 1 个网页的关联程度为“**中**”等。

王力宏			
档次	向量β¹	得分	网址
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10835056.html

表六

4. 用户信息中的“**郑玉龙**”在“网络舆情信息”文件夹中找到 18 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表七所示，根据步骤四中的评分标准与归档规则，“**郑玉龙**”与其中 1 个网页的关联程度为“**良**”等；与 8 个网页的关联程度为“**中**”等；与 9 个网页的关联程度为“**差**”等。

郑玉龙			
档次	向量β¹	得分	网址
良	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10835056.html
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jn_10835736.html
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jn_10835802.html
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jn_10835803.html
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jn_10835848.html
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jn_10838168.html
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jw_1030976.html
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jw_1030993.html
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jw_1031005.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10831545.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10831893.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10832900.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10833435.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10834860.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10835563.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10836526.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10837787.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10838169.html

表七

5. 用户信息中的“**丁羽心**”在“网络舆情信息”文件夹中找到 4 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表八所示，根据步骤四中的评分标准与归档规则，“**丁羽心**”与其中 3 个网页的关联程度为“**中**”

等；与 1 个网页的关联程度为“**差**”等。

丁羽心			
档次	向量β[1]	得分	网址
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jw_1031190.html
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jw_1031192.html
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jw_1031214.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jw_1031082.html

表八

6. 用户信息中的“**胡万林**”在“网络舆情信息”文件夹中找到 13 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表九所示，根据步骤四中的评分标准与归档规则，“**胡万林**”与其中 1 个网页的关联程度为“**良**”等；与 12 个网页的关联程度为“**中**”等。

胡万林			
档次	向量β[1]	得分	网址
良	1 0 0 1 0 0 0 0 0 1	53.78547159	jn_10837828.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10832332.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10833027.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10833201.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10833576.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10834486.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10834660.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10834954.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10836487.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10837467.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jw_1031344.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jw_1031398.html
中	0 0 0 1 0 0 0 0 0 0	34.61759425	jn_10833665.html

表九

7. 用户信息中的“**李天**”在“网络舆情信息”文件夹中找到 13 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表十所示，根据步骤四中的评分标准与归档规则，“**李天**”与其中 1 个网页的关联程度为“**良**”等；与 11 个网页的关联程度为“**中**”等。

李天			
档次	向量β¹	得分	网址
良	1 0 1 0 0 0 0 0 0 1	51.89418503	jn_10835979.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jn_10833113.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jn_10833874.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jn_10833917.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jn_10834018.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jn_10834046.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jn_10834822.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jn_10835094.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jn_10836070.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jn_10836527.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jn_10837142.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jn_10837543.html

表十

8. 用户信息中的“**李江**”在“网络舆情信息”文件夹中找到 9 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表十一所示，根据步骤四中的评分标准与归档规则，“**李江**”与其中 5 个网页的关联程度为“**中**”等；与 4 个网页的关联程度为“**差**”等。

李江			
档次	向量β¹	得分	网址
中	1 0 0 0 0 0 0 0 0 0	32.0393538	jn_10837401.html
中	1 0 0 0 0 0 0 0 0 0	32.0393538	jw_1031338.html
中	1 0 0 0 0 0 0 0 0 0	32.0393538	jw_1031459.html
中	1 0 0 0 0 0 0 0 0 0	32.0393538	jw_1031472.html
中	1 0 0 0 0 0 0 0 0 0	32.0393538	jw_1031487.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10831673.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10833240.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10834822.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10837543.html

表十一

9. 用户信息中的“**黄明**”在“网络舆情信息”文件夹中找到 3 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表十二所示，根据步骤四中的评分标准与归档规则，“**黄明**”与这 3 个网页的关联程度为“**中**”等。

黄明			
档次	向量β¹	得分	网址
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jw_1030951.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jw_1030961.html
中	1 0 0 0 0 0 0 0 0 1	41.16429426	jw_1030972.html

表十二

10. 用户信息中的“**余晓明**”在“网络舆情信息”文件夹中找到 158 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表十二所示，根

据步骤四中的评分标准与归档规则，“**余晓明**”与其中 157 个网页的关联程度为“**良**”等；与其中 1 个网页的关联程度为“**中**”等。

余晓明			
档次	向量β[1]	得分	网址
良	0 1 0 0 1 0 1 0 0 1	55.91809	jn_10831435.html
良	0 1 0 0 1 0 1 0 0 1	55.91809	jn_10831436.html
良	0 1 0 0 1 0 1 0 0 1	55.91809	jn_10831437.html
…			
良	0 0 0 1 1 0 1 0 0 0	54.82158	jn_10833665.html
良	0 0 0 0 1 0 1 0 0 1	49.74944	jn_10831438.html
良	0 0 0 0 1 0 1 0 0 1	49.74944	jn_10831533.html
良	0 0 0 0 1 0 1 0 0 1	49.74944	jn_10832797.html
…			
良	0 0 0 0 1 0 1 0 0 1	49.74944	jn_10838241.html
良	0 0 0 0 1 0 1 0 0 1	49.74944	jn_10838243.html
中	0 0 0 1 0 0 0 0 0 0	34.61759	jn_10837828.html

表十三

11. 用户信息中的“**张望**”在“网络舆情信息”文件夹中找到 11 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表十四所示，根据步骤四中的评分标准与归档规则，“**张望**”与其中 7 个网页的关联程度为“**中**”等；与其中 4 个网页的关联程度为“**差**”等。

张望			
档次	向量β[1]	得分	网址
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10833311.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10834371.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10835091.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10835378.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10835494.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10836089.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10836444.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10837403.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10837693.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10837826.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jw_1031565.html

表十四

12. 用户信息中的“**方小明**”在“网络舆情信息”文件夹中找到 8 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表十五所示，根据步骤四中的评分标准与归档规则，“**方小明**”与这 8 个网页的关联程度为“**中**”等；

方小明			
档次	向量β[1]	得分	网址
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10833613.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10833697.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10834540.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10837599.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10837872.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10838239.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10838246.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jw_1031349.html

表十五

13. 用户信息中的“**张秋白**”在“网络舆情信息”文件夹中找到 9 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表十四所示，根据步骤四中的评分标准与归档规则，“**张秋白**”与其中 1 个网页的关联程度为“**优**”等；与其中 7 个网页的关联程度为“**中**”等；与其中 1 个网页的关联程度为“**差**”等。

张秋白			
档次	向量β[1]	得分	网址
优	1 1 1 1 1 1 1 1 1 1 1	100	jn_10835782.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10831761.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10835516.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10836599.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10836814.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10837495.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10837701.html
中	0 1 1 0 0 0 0 0 0 0 0	40.62428558	jn_10838018.html
差	0 0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10837826.html

表十六

14. 用户信息中的“**王五**”在“网络舆情信息”文件夹中找到 1 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表十四所示，根据步骤四中的评分标准与归档规则，“**王五**”与这 1 个网页的关联程度为“**中**”等。

王五			
档次	向量β[1]	得分	网址
中	1 0 0 0 0 0 0 0 0 0 1	41.16429426	jn_10837826.html

表十七

15. 用户信息中的“**李世民**”在“网络舆情信息”文件夹中找到 6 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表十八所示，根据步骤四中的评分标准与归档规则，“**李世民**”与其中 1 个网页的关联程度为“**中**”等；与其中 5 个网页的关联程度为“**差**”等。

李世民			
档次	向量β[1]	得分	网址
中	0 1 0 0 0 0 0 0 0 1	36.32910941	jn_10836454.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10833367.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jw_1031395.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jw_1031401.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jw_1031403.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jw_1031406.html

表十八

16. 用户信息中的“**钟建国**”在“网络舆情信息”文件夹中找到 7 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表十九所示，根据步骤四中的评分标准与归档规则，“**钟建国**”与其中 7 个网页的关联程度为“**中**”等。

钟建国			
档次	向量β[1]	得分	网址
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10833311.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10834371.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10835091.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10835378.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10835494.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10836089.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10836444.html

表十九

17. 用户信息中的“**李龙**”在“网络舆情信息”文件夹中找到 8 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表二十所示，根据步骤四中的评分标准与归档规则，“**李龙**”与其中 8 个网页的关联程度为“**中**”等。

李龙			
档次	向量β[1]	得分	网址
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10833613.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10833697.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10834540.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10837599.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10837872.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10838239.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10838246.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jw_1031349.html

表二十

18. 用户信息中的“**陈龙**”在“网络舆情信息”文件夹中找到 8 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表二十一所示，根据步骤四中的评分标准与归档规则，“**陈龙**”与这 8 个网页的关联程度为“**中**”等。

陈龙			
档次	向量β¹	得分	网址
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10831761.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10835516.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10835782.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10836599.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10836814.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10837495.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10837701.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10838018.html

表二十一

19. 用户信息中的“**马小龙**”在“网络舆情信息”文件夹中找到 1 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表二十二所示，根据步骤四中的评分标准与归档规则，“**马小龙**”与这个网页的关联程度为“**优**”等。

马小龙			
档次	向量β¹	得分	网址
优	1 0 0 1 1 0 1 1 1 1	84.4489047	jn_10835776.html

表二十二

20. 用户信息中的“**胡万林**”在“网络舆情信息”文件夹中找到 2 个网址中的关键词与此用户信息的关键词匹配，则向量 β^1，得分，网址等详细信息如表二十三所示，根据步骤四中的评分标准与归档规则，“**胡万林**”与其中 1 个网页的关联程度为“**良**”等；与其中 1 个网页的关联程度为“**中**”等。

胡万林			
档次	向量β¹	得分	网址
良	1 0 0 1 0 0 0 0 0 1	53.78547159	jn_10837828.html
中	0 0 0 1 0 0 0 0 0 0	34.61759425	jn_10833665.html

表二十三

3. 结论

从上述的结果分析中，笔者发现用户的得分结果并不理想，用户信息与网页关键词匹配程度达到“**中**”等，“**差**”等居多，达到“**优**”等占极少数，而其中同时只出现性别和住址两个关键词的网页与用户信息匹配居多，而性别与住址两者包含的无效信息多。因此笔者的结果分析不能提供高的精确度保证结论的正确性。

代表性用户信息分析：

- 在结果分析中，笔者发现用户信息中“王力宏”的信息量少，出现的匹配关

键词只有性别和住址，评分结果为“中”等，因“王力宏”为知名人士，而名人对自身的隐私保护相当严谨，以防止自身隐私泄露不应承担的负担。

- 在结果分析中，笔者发现用户信息中“余晓明”的信息量大，出现的匹配关键词有性别，电话号码，QQ 号码与附加关键字。评分结果多为“良”等，与之匹配的网页多为广告型网页，发布的信息多为联系方式，以方便销售产品，因此推论此人为广告商或是广告代理人员。
- 在结果分析中，笔者发现用户信息中“张秋白”的信息量大，原因在与之关键词匹配的网页中，有 1 个网页中的关键词完全与此用户信息的关键词匹配。则评分结果为“优”等。结果发现该网页内容为离婚协议。通过其他与此用户信息关键词匹配的网页，可以推此人为律师。
- 在结果分析中，笔者发现用户信息中“胡万林”的此人的信息重复出现两次，除了出生日期不同，其余关键词均相同。推论为用户为注册账号时，避免个人信息泄露过多，或是填写出生日期时选择错误，导致此种情况出现。

模型优点：笔者通过 java 程序，可以迅速完成对关键词的在“网络舆情信息”中的网页搜索，相比人工搜索省时省力，精准度大幅提高。同时也实现了关键词加权评分系统的一体化，即可以一步输出评分结果。不必分步计算结果，使用户可以更简单方便的取得评分结果，并在此基础上推导结论。

模型缺点：在关键词词频统计时，由于关键词性别与住址出现频数过多，导致其他关键词的权重区别不大，因此，为了区分关键词权重的不同，笔者将对性别，住址单独出现的网址，予以剔除。使得不同关键词的权重差异明显，但是按照笔者的思路，性别与住址这两个关键词的权重应该是在全部关键词权重中最小的两个，但是根据笔者的方法修改后，性别与住址的权重相对变大了，影响了在对关键词加权评分系统中的正常评分结果。

模型改进：在结果分析中，笔者发现许多用户信息中与之匹配关键词的网址都只出现了性别和住址，但这两个关键词的无效信息大，应予以删除，重新计算评分排名情况。

修正前评分排名情况：

张秋白			
档次	向量β^1	得分	网址
优	1 1 1 1 1 1 1 1 1 1	100	jn_10835782.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10831761.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10835516.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10836599.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10836814.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10837495.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10837701.html
中	0 1 1 0 0 0 0 0 0 0	40.62428558	jn_10838018.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10837826.html

修正后评分排名情况：

张秋白			
档次	向量β^1	得分	网址
优	1 1 1 1 1 1 1 1 1 1	100	jn_10835782.html
差	0 0 0 0 0 0 0 0 0 1	25.84528835	jn_10837826.html

删除了同时只出现性别和地址的网页，使得用户信息的评分系统更能真实反映用户与网页的关联度。

4. 参考文献

[1]张义忠. 基于内容的网页特征提取[J].计算机工程与应用，2001，10

[2]何佳.网络舆情监控系统的实现方法[J].郑州大学学报，2010，10

[3]华秀丽.语义分析与词频统计相结合的中文文本相似度量方法研究[J].计算机应用研究，2012，03

[4]李实.中文网络客户评论的产品特征挖掘方法研究[D].哈尔滨：哈尔滨工业大学，2009

[5]何新贵.中文文本的关键词自动抽取和模糊分类[J].中文信息学报，1999，01

附录：

词频统计的程序代码：

```
package html;


import java.io.BufferedReader;
import java.io.File;
import java.io.FileInputStream;
import java.io.FileNotFoundException;
import java.io.IOException;
import java.io.InputStreamReader;


public class main_html {
   static boolean containsAny(String str, String searchChars){
         return str.contains(searchChars);
   }


   public static boolean deleteFile(String sPath) {
       boolean flag = false;
       File file = new File(sPath);
       // 路径为文件且不为空则进行删除
       if (file.isFile() && file.exists()) {
           file.delete();
           flag = true;
```

```
        }
        return flag;
    }

    public static double cos(int []a){
        int s1=0,s2=0;
        for(int i=0;i<a.length;i++){
            s1+=a[i];
            s2+=a[i]^2;
        }
        return s1/(Math.sqrt(s2)*Math.sqrt(a.length));
    }

    public static void main(String[] args) throws IOException {
        File file = new File("D:/sjwj2");  //文件夹的位置
        File[] files = file.listFiles();    //得到文件夹下所有的文
件
        File f;
        FileInputStream ism;
        InputStreamReader isr;
        BufferedReader bs;
        String info;
        int [] count = {0,0,0,0,0,0,0,0,0,0};   //统计单个网页词频
的数组
        int sum=0;
```

```
int a=0;
int [] c = {0,0,0,0,0,0,0,0,0,0};        //标记单个网页词频的二值数组，用于筛选
int [] sum_count = {0,0,0,0,0,0,0,0,0,0}; //统计总词频的数组中

String [][] key ={{"王林","  男  ","  江  西  萍  乡  人  ","360321196109183330",
        "13338941845","1961-09-18","1961  年  9  月  18  日  ","1961.09.18","345552","tiantianshang@163.com","asd@live.com","贩毒"},
        {"高连岳","男","  大          连          理          工  ","210213198512034662","15846576767","1985-12-03","3567457",
          "122898868@qq.com","","凌分贝"},
        {"  王  力  宏  ","  男  ","  广  西  玉  林  ","450922199008194334","18674635914","1990-08-19","1990 年 8 月 19 日","1990.08.19",
            "63472457","wyq@126.com","","金曲奖"},
        {"  郑  玉  龙  ","  男  ","  广  西  玉  林  ","450922198501078990","15320230485","1985-01-07","1985年1月7日  ","1985.01.07",

  "383475421","yud@163.com","zhengyulong@live.com","郑玉龙","小子","假小子"},
        {"  丁  羽  心  ","  女  ","  山  西  沁  水  县  ","
```

440113197803028412","13640786439","2013-12-02","2013 年 12 月 2 日","2013.12.02",

"76295638","dingyuxin@163.com","dingyuxin@msn.com","丁于心","丁书苗"},

{"胡万林","男","四川","440113197803028412","13640786438","2013-12-02","2013 年 12 月 2 日","2013.12.02","76295638",

"dingyuxin@163.com","dingyuxin@msn.com","胡万林"},

{"周茂名","男","海南文昌","412393199012147650","13348762495","1990-12-14","1990 年 12 月 14 日","1990.12.14",

"76297621","zhoumaoming@163.com","maoming@msn.com","周茂名"},

{"周世涛","男","海南文昌","412953199012147650","13348762495","1990-12-14","1990 年 12 月 14 日","1990.12.14",

"76297621","zhoumaoming@163.com","maoming@msn.com","周世源"},

{"李天","男","北京市海淀区","210422196107197904","13348762496","1964-04-01","1964 年 4 月 1 日","1990.04.01",

"76295622","litianyi@163.c0m","litianyi@msn.com","李天一"},

{"李江","男","北京市海淀区","370285198509212000","13348762497","1939-03-10","1939年3月10日","1939.03.10",

"76295623","lishuangjiang@163.com","lishaungjiang@msn.com","李双江"},

{"陈志祥","男","江西省抚州市","440115198510085699","13348762498","1988-10-10","1988年10月10日","1988.10.10",

"76295624","chenzhixiang@163.com","chixiang@msn.com","陈志祥"},

{"黄明","男","江西省抚州市","440116198910094000","13640786426","1988-10-11","1988年10月11日","1988.10.11",

"76295625","huanging@163.com","huangming.@msn.com","黄明"},

{"黄浩","男","江西省抚州市","440114199105117595","13640786427","1988-10-12","1988年10月12日","1988.10.12",

"76295626","hanghao@163.com","hanghao.@msn.com","黄浩"},

{"余晓明","男","江西省抚州市","440113197803028412","13322838884","1988-10-13","1988年10月13日","1988.10.13",

"268033328","yuxiaoming@163.com","yuxiaoming.@msn.com","18924889850","汽车出售"},

{"张望","男","北京市","440184198212196 73X","13640786429","1988-10-14","1988年10月14日","1988.10.14",

"76295628","zhanshang@163.com","zhanshan@msn.com","张三"},

{"方小明","男","广州市","440183197602106276","13640786430","1988-10-15","1988年10月15日","1988.10.15",

"76295629","fxm@163.com","fxm@msn.com","方小明"},

{"张秋白","男","深圳市","440183197906082 71X","13640786431","1988-10-16","1988年10月16日","1988.10.16",

"76295630","ls@163.com","lisi@msn.com","李四"},

{"王五","男","江门市","440105198509036000","13640786432","1988-10-17","1988年10月17日","1988.10.17",

"76295631","wangwu@163.com","wangwu@msn.com","王五"},

{"李世民","男","海珠市","440114198611218000","13640786433","1988-10-18","1988年10月18

日","1988.10.18",

"76295632","maliu@163.com","maliu@msn.com","马六"},
{"钟建国","男","北京市","440111198505274000","13640786434","1988-10-19","1988年10月19日","1988.10.19",

"76295633","zjg@163.com","zjg@msn.com","钟建国"},

{"李龙","男","广州市","44011419851001515X","13640786435","1988-10-20","1988年10月20日","1988.10.20",
"76295634","lilong@163.com","lilong@msn.com","李龙"},
{"陈龙","男","深圳市","440113199311063000","13640786436","1988-10-21","1988年10月21日","1988.10.21",

"76295635","chenlong@163.com","chenlong@msn.com","陈龙"},
{"马小龙","男","江门市","440116198109195000","13640786437","1988-10-22","1988年10月22日","1988.10.22",

"76295636","malxiaolong@163.com","maxiaolong@msn.com","马小龙"},

{"胡万林","男","四川","440113197803028412","13640786438","1949-12-12","1949年12月12日","1949.12.12",

"76295637","huwanlin@163.com","huwanlin@msn.com","胡万林"},

{"丁羽心","女","山西沁水县","440113197803028412","13640786439","1955-01-01","1955年1月1日","1955.01.01",

"76295638","dingyuxin@163.com","dingyuxin@msn.com","丁于心","丁书苗"}

};

```
for(int ii=0;ii<key.length;ii++){    //ii 表示第 ii 个用户
   sum=0;
   for(int i=0;i<files.length;i++){   //i 表示第 i 个文件
      f = files[i];
      ism = new FileInputStream(f);
      isr = new InputStreamReader(ism);
      bs = new BufferedReader(isr);
      info=bs.readLine();
      while(info!=null){
         for(int jj=0;jj<key[ii].length;jj++){  //jj 表示第 jj 个关键字
```

```
                if(containsAny(info,key[ii][jj]) && jj<=4){
                      c[jj] = 1;
                      count[jj]+=1;
                }
                if((!key[ii][jj].equals(""))                    &&
containsAny(info,key[ii][jj]) && jj>=5 && jj<=7){
                      count[5]+=1;
                      c[5]=1;
                }
                if((!key[ii][jj].equals(""))                    &&
containsAny(info,key[ii][jj]) && jj>=8 && jj<=10){
                      count[jj-2]+=1;
                      c[jj-2]=1;
                }
                if(jj>10 && containsAny(info,key[ii][jj])){
                   count[9]+=1;
                   c[9]=1;
                }
             }
             info = bs.readLine();
          }

          for(int k=0;k<count.length;k++){
             sum+=c[k];
          }
```

```
    if(sum==1 && (c[1]==1 || c[2]==1)){
        c[1]=0;
        c[2]=0;
        sum=0;
    }
    if(sum>0){
        for(int jj=0;jj<count.length;jj++){
            sum_count[jj]+=count[jj];
            count[jj]=0;
            c[jj]=0;
        }
    }
    else{
        for(int jj=0;jj<count.length;jj++){
            count[jj]=0;
            c[jj]=0;
        }
    }
    sum=0;
    ism.close();
    isr.close();
    bs.close();
    sum=0;
}
```

```
            System.out.println("已对第"+ii+"个用户完成搜索统计");
        }
        for(int k=0;k<count.length;k++){
            System.out.print(sum_count[k]+"   ");
        }
        System.out.println("end");
    }
}
```

求权值的程序代码：

```
package html;
public class qz {
    public static void main(String[] args) {
        double []a={94,306,109,10,159,1,159,2,2,296};     //关键字词频统计结果
        double s_a=0;
        for(int k=0;k<10;k++){
            s_a+=a[k];
        }
        double [] b= new double[10];
        double [] c = new double[10];
        double s_b=0;
        for(int i=0;i<10;i++){
            b[i]=(s_a-a[i])/s_a;
```

```
            s_b+=b[i];
        }
        for(int j=0;j<10;j++){
            c[j]=b[j]/s_b;
        }
        for(int i=0;i<10;i++){
           System.out.print(c[i]+"  ");  //输出权值
        }
    }
}
```

求余弦值的程序代码：

```
package main_html;

import java.io.BufferedReader;
import java.io.File;
import java.io.FileInputStream;
import java.io.FileNotFoundException;
import java.io.IOException;
import java.io.InputStreamReader;

public class html_main {
   static boolean containsAny(String str, String searchChars){

         return str.contains(searchChars);
```

```
}

public static boolean deleteFile(String sPath) {
    boolean flag = false;
    File file = new File(sPath);
    // 路径为文件且不为空则进行删除
    if (file.isFile() && file.exists()) {
        file.delete();
        flag = true;
    }
    return flag;
}

public static double cos(int []a,double [] weights){
   double s1=0,s2=0,s3=0;

   double [] b =new double[a.length];
   for(int i=0;i<a.length;i++){
      b[i]=a[i]*weights[i];
   }
   for(int i=0;i<a.length;i++){
      s1+=b[i]*weights[i];
      s2+=b[i]*b[i];
      s3+=weights[i]*weights[i];
   }
```

```
        return s1/(Math.sqrt(s2)*Math.sqrt(s3));
    }

    public static void main(String[] args) throws IOException {

        File file = new File("D:/sjwj2");  //文件夹的位置
        File[] files = file.listFiles();    //得到文件夹下所有的文件
        File f;
        FileInputStream ism;
        InputStreamReader isr;
        BufferedReader bs;
        String info;
        int [] count = new int[10];          //统计词频的二维数组
        double [] weights = {0.1019,0.0812,0.1005,0.1101,
                0.0956,0.1110,0.0956,0.1109,0.1109,0.0822};     // 权值数组
        //录入关键字，日期有三种格式，附加关键字多值的做或搜索
        String [][] key ={{"王林","  男  ","  江  西  萍  乡  人  ","360321196109183330",
                "13338941845","1961-09-18","1961  年  9  月  18  日  ","1961.09.18","345552","tiantianshang@163.com","asd@live.com","贩毒"},
                {"高连岳","男","  大          连          理          工  ","210213198512034662","15846576767","1985-12-03","3567457",
```

"122898868@qq.com","","凌分贝"},
{"王力宏","男","广西玉林","450922199008194334","18674635914","1990-08-19","1990年8月19日","1990.08.19",
"63472457","wyq@126.com","","金曲奖"},
{"郑玉龙","男","广西玉林","450922198501078990","15320230485","1985-01-07","1985年1月7日","1985.01.07",

"383475421","yud@163.com","zhengyulong@live.com","郑玉龙","小子","假小子"},
{"丁羽心","女","山西沁水县","440113197803028412","13640786439","2013-12-02","2013年12月2日","2013.12.02",

"76295638","dingyuxin@163.com","dingyuxin@msn.com","丁于心","丁书苗"},
{"胡万林","男","四川","440113197803028412","13640786438","2013-12-02","2013年12月2日","2013.12.02","76295638",

"dingyuxin@163.com","dingyuxin@msn.com","胡万林"},
{"周茂名","男","海南文昌","412393199012147650","13348762495","1990-12-14","1990年12月14日","1990.12.14",

"76297621","zhoumaoming@163.com","maoming@msn.com","周茂名"},

{"周世涛","男","海南文昌","412953199012147650","13348762495","1990-12-14","1990年12月14日","1990.12.14",

"76297621","zhoumaoming@163.com","maoming@msn.com","周世源"},

{"李天","男","北京市海淀区","210422196107197904","13348762496","1964-04-01","1964年4月1日","1990.04.01",

"76295622","litianyi@163.c0m","litianyi@msn.com","李天一"},

{"李江","男","北京市海淀区","370285198509212000","13348762497","1939-03-10","1939年3月10日","1939.03.10",

"76295623","lishuangjiang@163.com","lishaungjiang@msn.com","李双江"},

{"陈志祥","男","江西省抚州市","440115198510085699","13348762498","1988-10-10","1988年10月10日","1988.10.10",

"76295624","chenzhixiang@163.com","chixiang@msn.com","陈志祥"},

{"黄明","男","江西省抚州市

","440116198910094000","13640786426","1988-10-11","1988年10月11日","1988.10.11",

"76295625","huanging@163.com","huangming.@msn.com","黄明"},

{"黄浩","男","江西省抚州市","440114199105117595","13640786427","1988-10-12","1988年10月12日","1988.10.12",

"76295626","hanghao@163.com","hanghao.@msn.com","黄浩"},

{"余晓明","男","江西省抚州市","440113197803028412","13322838884","1988-10-13","1988年10月13日","1988.10.13",

"268033328","yuxiaoming@163.com","yuxiaoming.@msn.com","18924889850","汽车出售"},

{"张望","男","北京市","44018419821219673X","13640786429","1988-10-14","1988年10月14日","1988.10.14",

"76295628","zhanshang@163.com","zhanshan@msn.com","张三"},

{"方小明","男","广州市","440183197602106276","13640786430","1988-10-15","1988年10月15日","1988.10.15",

"76295629","fxm@163.com","fxm@msn.com","方小明"},

{"张秋白","男","深圳市","440183197906082 71X","13640786431","1988-10-16","1988年10月16日","1988.10.16",

"76295630","ls@163.com","lisi@msn.com","李四"},

{"王五","男","江门市","440105198509036000","13640786432","1988-10-17","1988年10月17日","1988.10.17",

"76295631","wangwu@163.com","wangwu@msn.com","王五"},

{"李世民","男","海珠市","440114198611218000","13640786433","1988-10-18","1988年10月18日","1988.10.18",

"76295632","maliu@163.com","maliu@msn.com","马六"},

{"钟建国","男","北京市","440111198505274000","13640786434","1988-10-19","1988年10月19日","1988.10.19",

"76295633","zjg@163.com","zjg@msn.com","钟建国"},

{"李龙","男","广州市","440114198510015 15X","13640786435","1988-10-20","1988年10月20日","1988.10.20",

"76295634","lilong@163.com","lilong@msn.com","李

龙"},

{"陈龙","男","深圳市","440113199311063000","13640786436","1988-10-21","1988年10月21日","1988.10.21",

"76295635","chenlong@163.com","chenlong@msn.com","陈龙"},

{"马小龙","男","江门市","440116198109195000","13640786437","1988-10-22","1988年10月22日","1988.10.22",

"76295636","malxiaolong@163.com","maxiaolong@msn.com","马小龙"},

{"胡万林","男","四川","440113197803028412","13640786438","1949-12-12","1949年12月12日","1949.12.12",

"76295637","huwanlin@163.com","huwanlin@msn.com","胡万林"},

{"丁羽心","女","山西沁水县","440113197803028412","13640786439","1955-01-01","1955年1月1日","1955.01.01",

"76295638","dingyuxin@163.com","dingyuxin@msn.com","丁于心","丁书苗"}

};

```
int sum=0;
for(int ii=0;ii<key.length;ii++){   //遍历所有用户
   sum=0;
   System.out.println(key[ii][0]);
   for(int i=0;i<files.length;i++){ //遍历所有文件
      f = files[i];
      ism = new FileInputStream(f);
      isr = new InputStreamReader(ism);
      bs = new BufferedReader(isr);
      info=bs.readLine();
      while(info!=null){      //遍历网页所有行
         //关键字二值搜索，有为1，无为0
         for(int jj=0;jj<key[ii].length;jj++){  //遍历用户的所有关键字
            if((!key[ii][jj].equals(""))                    && containsAny(info,key[ii][jj]) && jj<=4){
               count[jj]=1;
            }
            if((!key[ii][jj].equals(""))                    && containsAny(info,key[ii][jj]) && jj>=5 && jj<=7){
               count[5]=1;
            }
            if((!key[ii][jj].equals(""))                    && containsAny(info,key[ii][jj]) && jj>=8 && jj<=10){
               count[jj-2]=1;
```

```
            }
            if(jj>10 && containsAny(info,key[ii][jj])){
                count[9]=1;
            }
        }
        info = bs.readLine();
    }
    //检查搜索结果，去除只有地址或性别被搜索到的网页
    for(int k=0;k<count.length;k++){
        sum+=count[k];
    }
    if(sum==1 && (count[1]==1 || count[2]==1)){
        count[1]=0;
        count[2]=0;
        sum=0;
    }
    //输出搜索和计算的结果
    if(sum>0){
        double c = cos(count,weights);
        System.out.print(c+"余弦值："+f.getName()+"   ");
        for(int k=0;k<count.length;k++){
            System.out.print(count[k]+" ");
            count[k]=0;
        }
        System.out.println();
```

```
                }
                ism.close();
                isr.close();
                bs.close();
                sum=0;
            }
        }
        System.out.println("end");
    }
}
```

改进后的算法：

```
//去除只有地址和性别被搜索到的网页
//count 数组标记各关键字的搜索情况，sum 表示搜索到的关键字数
if(sum==2 && (count[1]==1 && count[2]==1)){
count[1]=0; //性别改为没搜索到
count[2]=0; //地址改为每搜索到
sum=0; //搜索到的关键字数为0
}
```