

参赛作品报告

作品名称：新闻可信度的评判与分析系统

参赛队伍：aidc0082

参赛学校：上海理工大学

学院 /系：光电信息与计算机工程学院/智能科学系

指导教师：傅迎华

组 长：李江

组 员：柴勤、谷志博

通信地址：上海市杨浦区军工路 516 号上海理工大学

电 话：15800957518

电子邮箱：1035742612@qq.com

目录

1	摘要	3
2	作品介绍	4
2.1	背景	4
2.2	功能与应用	4
3	系统方案	5
3.1	方案设计	5
3.2	软件流程	6
4	实现原理	6
4.1	微信交互	6
4.2	新闻检索	6
4.3	语义向量空间生成	7
4.4	新闻坐标点定位	8
4.5	观点聚类与分析	8
5	性能测试	8
5.1	环境配置	8
5.2	测试过程	9
6	创新性	13
7	总结	13
8	附录	14

1 摘要

当下国际上新闻媒体众多，但是由于政治立场或观察角度的不同，许多新闻媒体在报导同一件新闻的时候，往往会产生自己的立场与倾向，甚至有时因为政治立场而遮蔽或者歪曲新闻的内容，给人们带来困扰并左右群众的立场。

由于新闻结构的多样，需要较为准确地理解事件对机器而言并不容易。目前绝大多数的 NLP 项目也并未涉足新闻可信程度的判断。新闻的真实性与客观性在信息爆炸的时代尤为重要，有了客观准确的判断，新闻将不再成为有意者操纵舆论的手段。

我们通过比较大量新闻数据来判断新闻的可信度。在程序的检测过程中，需要用到的主要方法有：段落中心句的提取或生成，不同新闻里相似中心句的语义相似度计算，建立语义向量坐标系以囊括所有新闻数据点，新闻数据点的聚类与分析。用户通过微信与程序后台交互，输入想查询的新闻内容，系统将自动搜索相关内容并进行新闻的可信度分析，最终返回给用户相关新闻及观点归类、可信度等信息，可以让用户方便而快速地浏览分析新闻要点，快速而客观地形成辩证的态度。

2 作品介绍

2.1 背景

一些最近的自然语言理解产品，例如 Google 云自然语言 API、百度自然语言处理、玻森中文语义开放平台等，采用了机器学习的方法，可以实现词性分析、情感分析、新闻摘要、文本归类等功能。对中文分析的准确度可以达到 80% 上下。将自然语言理解与数据挖掘结合的实例多数应用于观点舆情分析，而对新闻本身内容的数据分析并不常见。

然而，在信息流通迅速尤其是自媒体蓬勃发展的当今社会，不同人对同一件事情有多重角度的解读、立场不一，又甚至有新闻刻意提供虚假信息或者回避重要细节，最终使读者变得毫无头绪，或是轻易相信一方之词。此时，快速分析新闻观点及可信度、为用户推荐最为中肯的新闻材料意义非凡。

2.2 功能与应用

- **新闻检索**

用户输入关键字，程序自动搜索近期新闻。

- **新闻观点分类**

根据新闻之间观点不同进行分类，推荐各个观点具有代表性的文章，避免用户阅读重复新闻，节省用户时间精力。

- **分析新闻情感、可信度**

对新闻的可信度做出大致估计，为用户是否应该相信该报道提供参考。

3 系统方案

3.1 方案设计

系统的功能架构见下图：

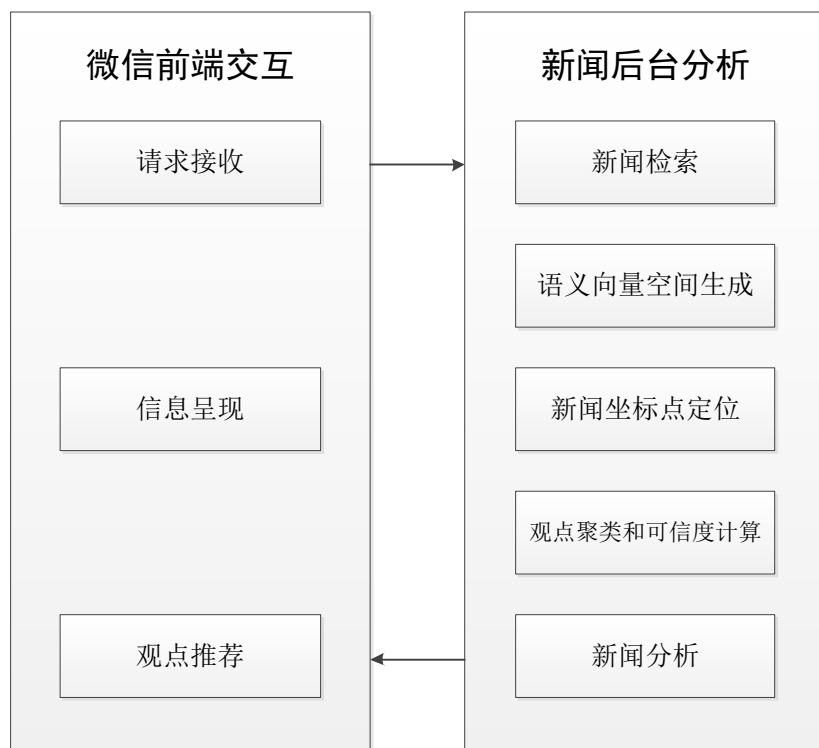


图 1 功能框架图

系统依据功能整体划分为 2 个部分：

- 微信前端交互：请求接收、信息呈现、观点推荐；
- 新闻后台分析：新闻检索、语义向量空间生成、新闻坐标点定位、观点聚类 and 可信度计算、新闻分析。

3.2 软件流程

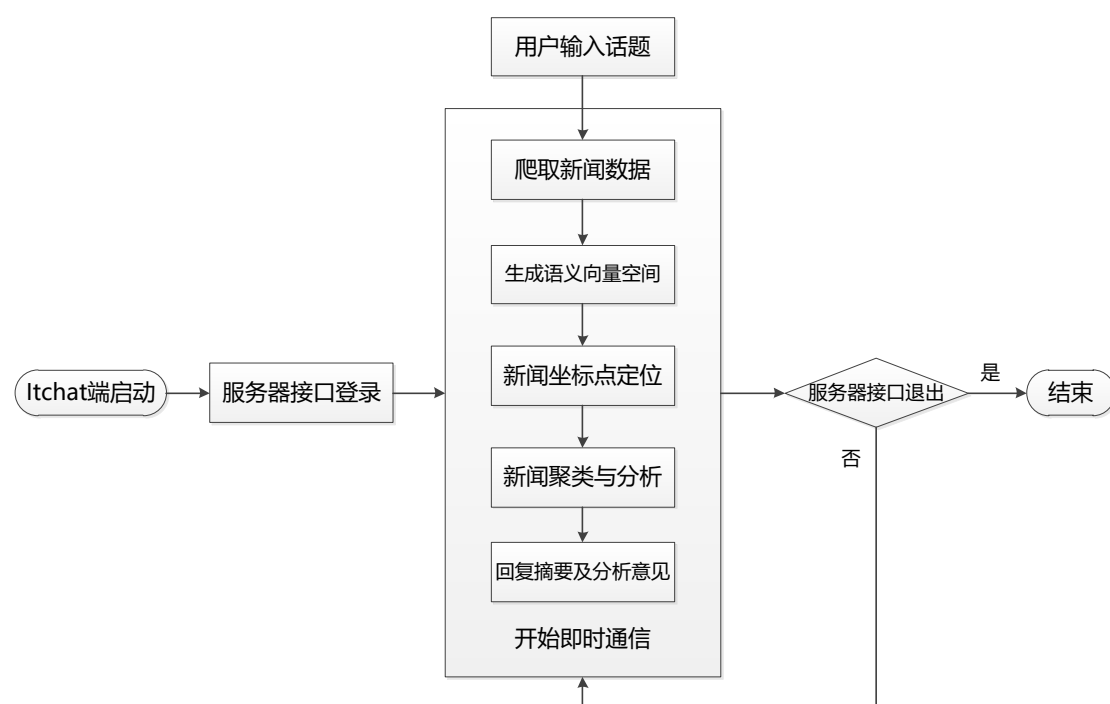


图 2 总系统流程

4 实现原理

4.1 微信交互

本系统使用 Itchat^[1]进行微信交互，测试者扫描程序弹出的二维码，作为程序后台与用户交互的媒介。用户向测试者微信发送感兴趣的新闻话题，程序后台将收到指令并进行相应新闻的检索。

4.2 新闻检索

程序后台收到微信交互传输进来的待搜索内容，先从百度新闻爬取相关内容，获取新闻标题、作者、时间，并获取新闻网页链接，再在任意新闻链接中爬取其中的正文部分，存储在程序中进行后续操作。

4.3 语义向量空间生成

为了给每个新闻定位坐标并分析，我们需要建立语义向量空间以表示他们。我们对每个新闻进行关键句提取^[2]，每个新闻可以有多个关键句，所有新闻的所有关键句生成两两间的语义相似度关系^[3]，得到一个全局语义相似度矩阵。我们希望通过向量来表示每个关键句，新闻的坐标点通过向量归一化相加得到，因此语义相似度的作用则是反映向量之间的角度关系。

由于相似度矩阵的数值不具有传递性，因此如果直接将相似度映射到 $0^\circ \sim 90^\circ$ 的空间，当向量超过两维，将有可能无法产生正确的角度关系。例如，假设一共只搜索到三个关键句，生成 3×3 的相似度矩阵 $S = \begin{pmatrix} 1 & 0.1 & 0.9 \\ 0.1 & 1 & 0.8 \\ 0.9 & 0.8 & 1 \end{pmatrix}$ ，如果每个关键句代表一个向量，那么向量之间的角度关系 $A = (1 - S) \times 90^\circ$ ，得到如下角度矩阵 A 。

$$A = \begin{pmatrix} 0 & 81 & 9 \\ 81 & 0 & 18 \\ 9 & 18 & 0 \end{pmatrix}$$

其中，第三个向量与另外两个向量的角度分别为 9° 和 18° ，那么另外两个向量之间的角度应该在 $9^\circ \sim 27^\circ$ 之间，然而另外两个向量之间的角度为 81° 。显然，这样的角度关系无法在坐标系中画出。

我们发现，如果直接把相似度矩阵中的每一行代表每个向量，则可以避免生成角度关系，从而直接不矛盾地得到向量的值。每个向量的维数等于所有向量的个数。

4.4 新闻坐标点定位

生成了语义向量空间后，每则新闻将作为一个坐标点映射在向量空间内。如果某一新闻有三个关键句即三个向量，那么这个新闻的坐标则是这三个向量的之和，值得注意的是，每加一个向量需要进行归一化处理。考虑到标定新闻的坐标不仅要看新闻的内容相似性，也需要考虑情感，每个新闻的坐标将再添加一维情感^[2]。

4.5 观点聚类与分析

每个新闻映射到向量空间后，可采用传统数据挖掘方法进行观点归类与分析。我们采用的是基于密度的 Mean Shift^[4]聚类方法，可以通过密度阈值自动决定聚类的簇数。在本程序中，每个簇代表一类相似的新闻集合，有几个簇就有几种角度或观点。在新闻聚类完成之后，每一个簇中最接近中心点的新闻最可信，也最有代表性。如果某一新闻 n 属于簇 C ，这条新闻到该簇中心点的距离记为 D_n ，簇 C 有 m 个新闻，所有该簇新闻到该簇中心点的距离记为 $\sum_{i=1}^m D_i$ ，则该新闻的可信度 $R = 1 - \frac{D_n}{\sum_{i=1}^m D_i}$ 。

5 性能测试

5.1 环境配置

- Python 3.6: 安装并配置 pip
- Itchat: 微信端框架，安装命令\$ pip install Itchat
- Bosonnlp: 玻森中文语义开放平台，安装命令\$ pip install -U bosonnlp
- Aip: 百度开放平台，安装命令\$ pip install baidu-aip

5.2 测试过程

- 用户输入关键字后，实时更新运行状态

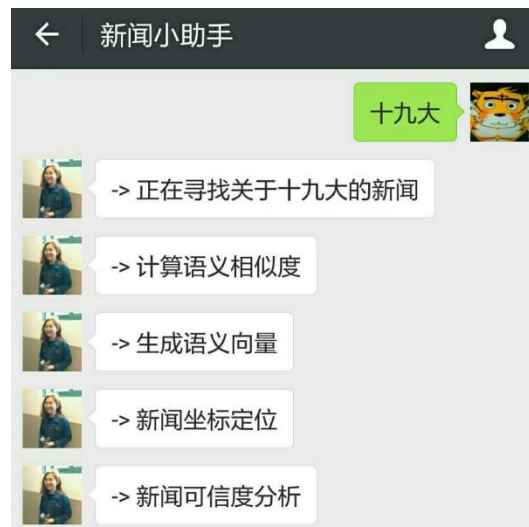


图 3 用户界面

- 程序后台运算过程

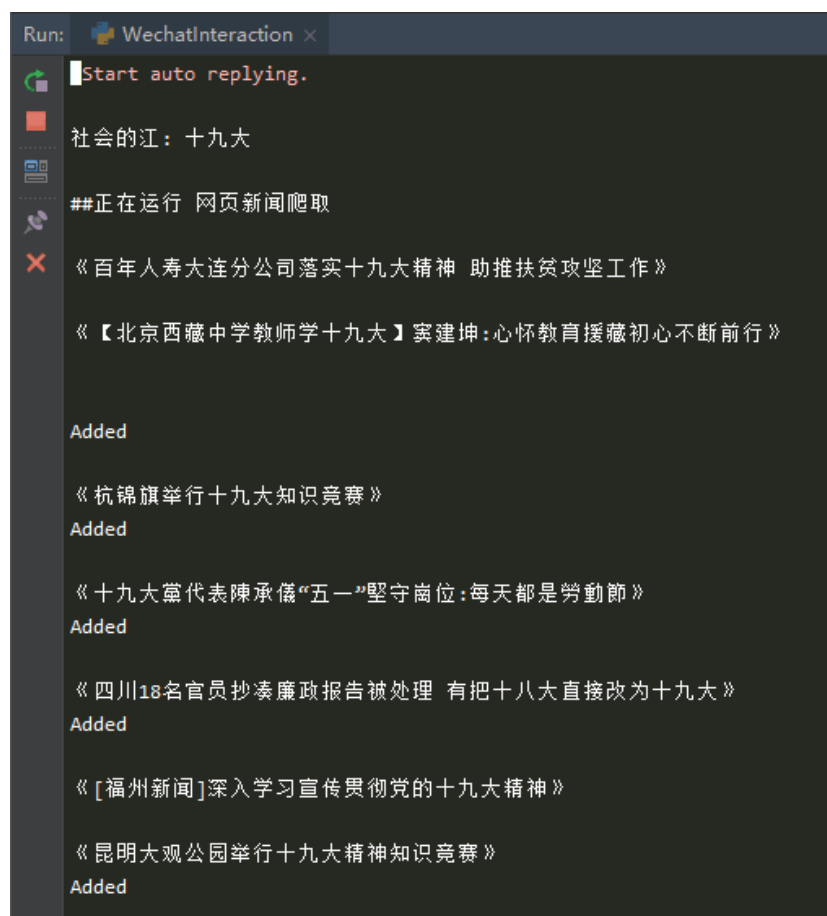


图 4 查找新闻

```
Run: WechatInteraction x
##正在运行 语义相似度计算
0.5263157894736842%
1.0526315789473684%
.....
1.5789473684210527%
2.1052631578947367%
.....
2.631578947368421%
3.157894736842105%
```

图 5 计算语义相似度

```
Run: WechatInteraction x
##正在运行 语义向量生成
最终向量
[[ 1.          0.487889  0.328525  0.572091  0.316969  0.476407
  0.492652  0.         0.255542  0.0528478  0.312951  0.         0.
  0.126195  0.100606  0.         0.0240546  0.0564436  0.209396
  0.135588 ]
[ 0.487889  1.          0.382722  0.62802  0.29609  0.616082
  0.491911  0.181721  0.302042  0.0119172  0.24806  0.124827  0.
  0.222096  0.120756  0.0718138  0.356145  0.200784  0.369797
  0.255743 ]
[ 0.328525  0.382722  1.          0.456673  0.31333  0.389706
  0.401465  0.0910425  0.         0.11041  0.206258  0.292931  0.
  0.26966  0.201  0.324263  0.0421355  0.241514  0.128625
  0.0616912]
[ 0.572091  0.62802  0.456673  1.          0.391691  0.433852
  0.482594  0.195042  0.273192  0.0761484  0.229786  0.0752032  0.
  0.167342  0.0642444  0.0276813  0.0763225  0.0610442  0.351548
  0.188009 ]
[ 0.316969  0.29609  0.31333  0.391691  1.          0.638528
  0.478831  0.         0.140548  0.10435  0.172196  0.243603  0.
  0.404936  0.3737  0.162332  0.34448  0.307915  0.424364
  0.374762 ]
[ 0.476407  0.616082  0.389706  0.433852  0.638528  1.          0.648027
  0.0168222  0.16839  0.165873  0.284352  0.309153  0.         0.273461
  0.220757  0.053973  0.317909  0.253577  0.422583  0.352256 ]
[ 0.492652  0.491911  0.401465  0.482594  0.478831  0.648027  1.
  0.117283  0.266271  0.129154  0.219343  0.273308  0.         0.32694
```

图 6 生成语义向量

```

Run: WechatInteraction x
##正在运行 新闻坐标点定位

中国西藏网 [ 0.1070646 0.12475958 0.12280261 0.17021306 0.10044804 0.16095316
0.18411371 0.05526497 0.10317336 0.03431209 0.14027974 0.07886314
0.
0.13454294 0.09383085 0.0884078 0.11186005 0.13750327
0.29033349 0.22259901 8.18220565]
新华网内蒙古站 [ 0.06233774 0.07235186 0.06110796 0.07680831 0.13401122 0.16549433
0.18185338 0.01375374 0.05973033 0.04381719 0.07716479 0.10122919
0.
0.13520774 0.12002401 0.06873004 0.17904152 0.16640029
0.28752256 0.37756576 9.97956565]
人民网 [ 0.0290214 0.04172308 0.03504858 0.04395198 0.03567561 0.05278647
0.05912921 0.10519685 0.12114878 0.15697103 0.20027276 0.19336626
0.10378611 0.13891853 0.14130053 0.10124582 0.16662561 0.1372998
0.21063913 0.3470289 9.7820295 ]
中国经济网 [ 0.0106319 0.03396373 0.03898097 0.01489448 0.06070445 0.04530314
0.05374935 0.05705591 0.02315967 0.0438387 0.01793512 0.07349712
0.0764434 0.19028409 0.20869542 0.24607551 0.32280047 0.46015838
0.14604618 0.37360434 0.45796581]
新华网云南频道 [ 0.03163751 0.05906378 0.01898521 0.05476003 0.08530089 0.08953024
0.10391952 0.06723247 0.08455041 0.04649863 0.04466953 0.04495195
0.01219497 0.06327382 0.05980731 0.02698875 0.10888251 0.09135599
0.3691764 0.5133511 9.95272112]

```

图 7 新闻坐标

```

Run: WechatInteraction x
新闻距离矩阵
[[ 0. 1.8131636 1.64322709 7.74630517 1.81509741]
[ 1.8131636 0. 0.3859844 9.53353221 0.26902396]
[ 1.64322709 0.3859844 0. 9.33674877 0.42587507]
[ 7.74630517 9.53353221 9.33674877 0. 9.51343493]
[ 1.81509741 0.26902396 0.42587507 9.51343493 0. ]]

```

图 8 新闻距离

```

Run: WechatInteraction x
##正在运行 新闻可信度分析

n_clusters_ 2
centroids [[ 6.50224747e-02 7.81569426e-02 5.43842774e-02 9.08554737e-02
7.97093015e-02 1.36278326e-01 1.01378898e-01 7.46118105e-02
7.75009203e-02 9.96346093e-02 1.35258125e-01 1.14322634e-01
3.22403876e-02 1.14882464e-01 1.21540359e-01 8.39789808e-02
1.32134152e-01 1.33267305e-01 2.98573075e-01 3.95182015e-01
9.54202874e+00]
[ 2.53784915e-02 5.48603575e-02 8.38066819e-02 -6.24297104e-03
5.76637320e-02 7.66439043e-02 7.15902146e-02 6.38578638e-02
1.90685504e-02 2.46229029e-02 1.05250542e-02 1.06927752e-01
7.97407276e-02 1.56806928e-01 2.09930987e-01 2.50677548e-01
2.86208701e-01 4.48984431e-01 1.80065899e-01 3.97696787e-01
4.86114357e-01]]

```

图 9 可信度分析

● 用户界面显示结果



图 10 新闻摘要、观点分类、可信度分析

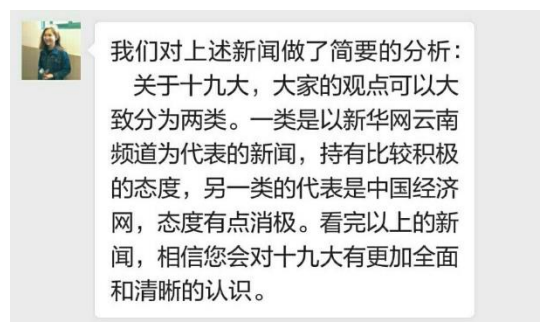


图 11 新闻观点简要分析

6 创新性

- **新闻的可信度评判与分析**

本系统由用户输入感兴趣的新闻话题，自动无监督地计算新闻的可信度、对相似的新闻归成一类，同时给予用户适当的新闻分析。

- **将新闻作为数据点映射到语义空间进行聚类分析**

新闻整体将作为语义空间里的一个数据点，所有同一话题的新闻都能在语义空间中得到表示。通过 Meanshift 聚类方法，可以自动对新闻进行聚类，并找到每一类新闻的近心点与离群点。

- **语义向量空间通过文本语义相似度生成**

对搜索到的所有同话题新闻进行关键句提取，所有新闻的所有关键句计算两两之间的语义相似度，得到的语义相似度来进一步构建语义向量空间以囊括新闻数据点。

- **采用方便自然的微信交互**

用户在微信端对相应公众号输入感兴趣的话题，系统将回复对应的新闻及分析结果。用户无需在手机或电脑下载任何软件或访问网页，所有操作方便而简约。

7 总结

首先要感谢主办方提供的此次学习机会，让我们可以充分运用课内外学习的知识，同时也开拓了视野，了解到在人工智能领域里最前沿的发展。我们也实际体会到了一个产品从最初的设计，初步的实现，以及后期的调试和修改，

一个完整的产品开发过程。整个团队在这个比赛中也颇有感想，具体有以下几点：

1.在开发过程中，团队的成员必须要保持充分沟通交流。充分的沟通能保证开发的质量以及一致性，提高开发的效率。

2.这次比赛促使我们不断的去了解和探索学习新技术，让我们感受到人工智能的强大知识内涵和高速的发展。我们在报名时讨论并确定的技术方法，在2个月之后就出现了新的、更好、更强的方法，这确实给我们开发带来了一定的困难，但我们没有放弃，组织了多次学习讨论会，努力去学习这些新方法、新技术，既提高了自身的能力，也加快了开发的进度，取得了很好的效果。

3.有志于从事人工智能的学生必须要做到：跟踪最新科技成果、不断学习、勇于尝试。

8 附录

[1] Itchat 微信框架，一个开源的微信个人号接口，可以方便实现对微信。

[2] 调用玻森中文语义开放平台 API

[3] 调用百度自然语言处理 API，采用 CNN 卷积神经网络模型，模型语义泛化能力介于 BOW 词包/RNN 之间，对序列输入敏感，相较于 GRNN 循环神经网络 模型的一个显著优点是计算效率会更高些。

[4] Mean Shift 算法：指一个迭代的步骤，即先算出当前点的偏移均值，移动该点到其偏移均值，然后以此为新的起始点，继续移动，直到满足一定的条件结束。