ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.9, September 2009, pp.2397−2406
doi: 10.3724/SP.J.1001.2009.03434

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种基于信息检索技术的网络新闻影响力分析方法*

杨伟杰[1,2+] 戴汝为[2+] 崔 霞[2]

[1](北京工商大学 计算机与信息工程学院, 北京 100048)

[2](中国科学院 自动化研究所 复杂系统与智能科学重点实验室,北京 100190)

Model for Internet News Force Evaluation Based on Information Retrieval Technologies

YANG Wei-Jie[1,2+], DAI Ru-Wei[2+], CUI Xia[2]

[1](School of Computer Science and Information Engineering, Beijing Technology and Business University, Beijing 100048, China)

[2](Institute of Automation, The Chinese Academy of Sciences, Beijing 100190, China)

+ Corresponding author: E-mail: weijie.yang@ia.ac.cn, ruwei.dai@ia.ac.cn

Yang WJ, Dai RW, Cui X. Model for Internet news force evaluation based on information retrieval technologies. *Journal of Software*, 2009,20(9):2397−2406. http://www.jos.org.cn/1000-9825/3434.htm

Abstract: Evaluating the force of Internet news is an important aspect in the social security field. This paper establishes a quantitative analysis model for calculating the Internet news force with information retrieval methods based on the correlative information got through information retrieval technologies. This method can help people understand how great the news affects the social security. A lot of experiments, lead to a conclusion: this method can rank a stream of news effectively, and also can find most important news items from a number of news which belong to different types; its result is very close to people's judgment.

Key words: information retrieval; social security; the force of news; HITS (hypertext induced topic selection) arithmetic; China Internet index system (CIIS)

摘 要: 利用信息检索领域中的相关算法,分析研究通过信息检索相关技术得到的相关信息,建立了一个网络新闻影响力模型来定量地计算一则新闻的影响力,从而估计它对社会安全产生影响的程度.在对大量实验结果的统计分析中发现,此方法可以有效地对新闻文章进行排序,发现不同新闻类型中最值得关注的新闻,其结果与人的定性判断结果具有较高的一致性.

关键词: 信息检索;社会安全;新闻影响力;HITS(hypertext induced topic selection)算法;中国互联网指数系统(CIIS)

中图法分类号: TP391 **文献标识码**: A

作为一种信息传播的方式,新闻会对社会稳定产生很大的影响.新闻舆论监督的勃兴,肇始于美国大法官斯特瓦特创设的第四权力理论.所谓的第四权力就是指新闻舆论.事实上,它虽然不是国家权力,但随着新闻媒体

* Supported by the National Natural Science Foundation of China under Grant No.60602032 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA010106 (国家高技术研究发展计划(863)); the National Basic Research Program of China under Grant No.2007CB311007 (国家重点基础研究发展计划(973)); the CASIA Innovation Fund for Young Scientists of China under Grant No.08J10A1FZ1 (中国科学院自动化研究所创新基金)

Received 2007-10-29; Revised 2008-06-03; Accepted 2008-08-07

在社会政治、经济、文化生活中的作用日益增强,它发挥着越来越重要的作用.同时,随着网络媒体“议程设置”功能的减弱和“沉默的螺旋”作用的不断增强,网络新闻作为网络舆论和社会舆论形成的主要源泉,准确判断它的影响力从而准确、即时地把握社会舆论的动向变得尤为重要,因而确定新闻影响力对社会安全及其他相关方面具有重要意义[1].例如在一个社会舆论出现之后,社会安全调控部门可以根据相关新闻的某些指标来判断这个论断的影响范围,从而做出对可能的突发事件的预先反应.而这个过程中使用的新闻的相关指标就可以用来判断新闻的影响力,如果有了一定的新闻影响力的计算模型,这个判断的过程则会大为简化.所以,本文中提及的新闻的影响力可以理解为新闻所影响的人群、地域等范围的大小、对社会产生作用力的大小等因素的综合.

另外,可以利用新闻影响力来帮助新闻搜索引擎对新闻进行排序.最近几年,越来越多的人开始习惯在网上看新闻,据估计,Yahoo!新闻的用户占整个 Yahoo!用户总量的一半.因而如何及时地把最重要的新闻推荐给用户成为各大新闻搜索引擎的努力目标.这也促使对网络新闻的研究变得越来越重要.对网络新闻的研究,从社会科学的角度,主要是对网络新闻的传播方式和新闻对社会的影响方式等方面进行的一些定性分析.在信息领域,出现了很多商业性的新闻搜索引擎,例如 Google 新闻,Yahoo 新闻,MSNBot 等等,文献[2]中给出了一份完整的商业新闻引擎名单.另外,也出现了一些新闻信息处理的原型系统,例如 NewsInEssence[3,4],QCS[5].虽然在新闻检索和新闻信息处理方面都在进行不断的努力,但是真正涉及到网络新闻影响力排序的学术研究仍然很少.文献[6,7]是仅有的关于新闻排序的论文.文献[6]主要利用新闻的时效性和新闻转载信息来对新闻进行排序.文献[7]则是利用了网页的布局和新闻转载信息对网页进行排序,因为涉及到了新闻链接在网站首页中的位置信息,所以这种方法对单个网站中新闻之间的排序更加有效.这两篇文章利用的信息是新闻排序的主要信息,但是新闻网页中可以用来进行新闻排序的信息还不止这些,例如新闻的回复率,这是新闻影响力的一个很好的体现,但在这两篇论文中都没有提到.

本文提出一种对新闻影响力进行定量分析的算法模型.通过分析了新闻影响力有关的因素(第 1 节),借助于信息检索中的预处理等相关技术,有针对性地从新闻网页中提取相关的信息,利用相关的算法有效地综合这些信息得到新闻的影响力值(第 2 节).第 3 节我们将给出一部分实验结果,这些结果体现了此模型在新闻影响力排序方面的有效性.第 4 节是结论.

1 网络新闻影响力的要素分析及其计算模型

通常情况下,对信息影响力的评价需要考虑信源可靠性、传播源可靠性、发布时间、信息内容的性质(领域)等几个要素.新闻作为信息的一个重要特例,对其进行影响力排序也应该考虑类似的要素.另外,由于新闻有其特有的写作方式,并考虑到网络传播方式的特殊性,以及人们对网络新闻产生的感想也会通过一定的方式明确地显现在网络上,所以为了建立网络新闻影响力的计算模型,本文根据网络新闻的特性,首先对网络新闻影响力的几个要素进行分析.

(1) 新闻网页质量与新闻信源网站质量之间互相影响的关系.好的网站发出的新闻往往具有比较高的质量,而且好网站的浏览人次一般都会比较多,因而它对社会产生的影响就比一般网站更深远.同样,发布好的新闻网页会提升整个网站的影响力.比如,我们通常会认为新华网发布的新闻会比其他门户网站发布的新闻更加权威.

(2) 新闻传播速度和传播规模.传播速度快,而且传播范围广的新闻一般是比较受关注的新闻,会对社会舆论形成有比较大的贡献,网络新闻的传播主要是通过浏览和转载来完成的,即浏览人数越多说明新闻越重要,转载新闻的网站越多说明新闻越重要.而且,如果转载这则新闻的网站是新闻网站中质量比较高的网站,那么这则新闻就显得更加重要.但是浏览人数是在服务器端存储的,所以我们无法取得.因而我们判断新闻传播状态的时候主要是利用了新闻转载次数以及转载网站的质量.

(3) 新闻的回复次数.浏览者对新闻发出了回复,说明他对新闻产生了反应,回复人数越多,说明新闻对越多的人产生了影响,那么这则新闻的影响力就相应地变大了.

(4) 新闻的发布时间.由于新闻具有时效性,因而一般认为最新发布的新闻要比以前发布的新闻更加重要,

而且,新闻的回复次数和转载次数也与新闻发布的时间有很大关系.一般情况下,新发布新闻的回复次数和转载次数在新闻发布的初期会比它之前发布的新闻低一些.

(5) 新闻链接在新闻网站中所处的位置.如果是对单个网站中的新闻进行重要性排序,这点是很好的依据.因为按照习惯,每个网站会把当时比较重要的新闻的链接放在网页最显眼的地方,而且会加一些图片和文字的摘要说明或者采用比较特殊的字体.不重要的新闻则只是将链接罗列在相关新闻列表中.而且这些布局信息也反映了网站编辑人员对新闻排序的看法,对新闻网页的排序也有重要的指导意义.本文涉及的算法主要是针对任意网站、任意新闻网页,所以暂不考虑这个因素.

从以上分析可以看出,对新闻影响力进行计算需要考虑新闻信源网站及其质量、新闻转载网站及其质量、新闻回复人次、新闻发布时间等几大要素.融合这些要素,本文提出了新闻影响力的计算模型,框架见式(1):

$$N_F = D(t_s,t)\times W_s\times(a\times Trans+b\times Rep) \tag{1}$$

其中,N_F为新闻影响力大小, $D(t_S,t)$ 是时间因素,W_S为新闻信源网站的影响力因子,$Trans$为新闻转载率因素,Rep为新闻回复率因素,a 和 b 为待定的系数因子,它们之间的关系为 $a+b=1$ 且 $a>0,b>0$,它们的取值决定了转载率因素和回复率因素在决定新闻影响力大小时所起的作用.式(1)计算模型中各项影响因素的计算,在下文中将给出详细陈述.

2 新闻影响力定量分析算法实现

依据式(1)的新闻影响力计算模型,本文的排序算法实现流程如图 1 所示.第 1 步,对新闻网页进行相似性判断,如果判断为转载或相似网页则提取网页转载或重复信息;第 2 步,用新闻转载网站之间的关系,利用 HITS 算法[8]对各转载网站进行了权威度计算,确定最终的信源网站和新闻转载率;第 3 步,对新闻网页进行信息提取,并利用提取的信息和上步中得到的重复信息进行回复率计算;第 4 步,利用中国互联网指数系统对新闻的信源网站的质量进行判定,并将其作为新闻影响力判断的一个整体的比例因子;第 5 步,考虑时间因素对新闻影响力的作用;第 6 步,根据以上步骤得到的信息进行综合计算得出新闻的影响力.

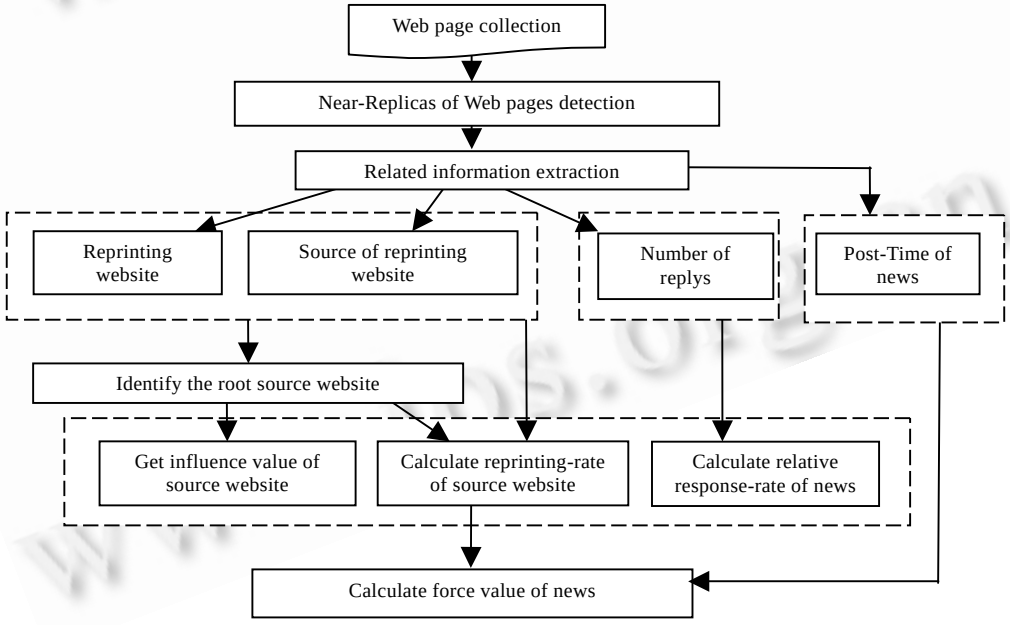

Fig.1 Framework of Internet news force evaluation

图 1 新闻影响力算法框架图

2.1 新闻网页的判重及相关信息提取

新闻网页的重复,一般源于转载或对同一事件的不同报道,导致新闻文档的完全一致或者部分一致.因而,新闻网页的判重需要进行两种方式的判别[9].

首先,对整篇文档进行 MD5 方法判重,如果文档完全一致,则直接确定网页之间的转载关系.如果文档并不完全一致,则进一步采用基于网页主体内容间的相似程度来判断它们是否为近似相同.

基于网页主体内容的判重,采用向量空间模型(VSM)表示网页主体内容,同时识别文章主体中的命名实体,因为命名实体最能体现新闻的特征,是新闻相似性判断的一个重要依据,此算法中需要识别的命名实体为人名、地名、机构名称和时间.当两个网页主体内容相似比例达到设定的阈值时,判别它们为近似相同,为重复网页.在计算过程中,网页 $U_i(i\in[1,n])$使用特征向量进行表示,其关键词权值 *We* 采用以 TF*IDF 方法来确定,如果判定词项为命名实体,权值适当加强.具体定义如下:

$$W_e=\begin{cases}idf_e\times\alpha, & e\text{为命名实体}\\ idf_e, & e\text{为其他}\end{cases} \tag{2}$$

其中,α为加权因子,本文实验中取值为 5.

最后选取 *m* 个权值较大的词项生成网页特征向量,以两个网页特征向量中共现词项数量为相似性判据,如果共现个数大于阈值,则两个网页为相似网页.

确定转载或近似关系之后,提取并记录相关的信息,需要记录的主要信息有:转载网站、转载网站的信源网站、转载网站中的回复次数以及新闻发布时间.此处的转载网站和信源网站只是对转载关系的一种记录,并非最后确定的真正的信源网站和转载网站.最后的信源网站将在下一步中得到确定.

2.2 新闻转载关系判断及新闻信源网站权威度计算

通常,

新闻转载率(记为 *Trans*)=转载次数/源网站点击次数 (3)

然而,由于网络新闻的转载关系存在直接转载和间接转载两种,使得源网站一开始不能确定,而且源网站的点击次数保存在服务器端,网页中一般不提供,所以很难得到.由于新闻网页与其源网站之间存在互相增强的反馈关系,应用 HITS 算法原理,本文把网站作为节点,网站拥有内容质量(权威)属性 Authority 和转载属性 Hub,应用迭代算法计算如下:

每个网站 *pt* 有内容质量属性值$A_0(pt)$和转载属性值$A_1(pt)$.首先,在网络整体层次上将所有节点的这两个属性值初始化为 1.然后,用 $pt\to qt$ 描述网站 *pt* 转载了网站 *qt* 的新闻,用下面的迭代公式计算内容质量属性值和转载属性值,每次迭代完成后将所有网页的属性值正则化为 1.

$$A_0(pt)=\sum_{qt\to pt}A_1(qt) \tag{4}$$

$$A_1(pt)=\sum_{pt\to qt}A_0(qt) \tag{5}$$

$$A_0(pt)=\frac{A_0(pt)}{\left[\sum_{\forall pt}(A_0(pt))^2\right]^{\frac{1}{2}}} \tag{6}$$

$$A_1(pt)=\frac{A_1(pt)}{\left[\sum_{\forall pt}(A_1(pt))^2\right]^{\frac{1}{2}}} \tag{7}$$

按以上公式迭代更新每个节点的属性$A_0(pt)$,$A_1(pt)$.

利用第 2.1 节中提取到的转载信息,首先提取新闻转载网站之间的关系,包括直接转载和间接转载关系,计算各个转载网站的权威度值,最终把被转载(类似于普通网页的入链)次数最多的那个网站作为源网站,把它的权威度值作为新闻的转载率值.

2.3 新闻源网站影响力因子(W_S)确定

对新闻网站质量的评价来自人们对这个网站的关心程度,浏览这个网站的人数多了,自然可以认为这个网站的质量比较高,它提供的新闻就比较有价值.因而,新闻源网站的质量好坏程度,也是对网络新闻影响力进行判断的一个重要依据.

中国互联网实验室与国家统计局联合发布了中国互联网指数系统(China Internet index system,简称CIIS)[10]对网站进行评估.CIIS 利用 Alexa.com 作为第三方监测机构,依托各监测网站的人气指数,将提供中文服务的网站按照所处行业、地域、提供服务等进行划分,并由此进一步揭示出中国互联网行业的行业发展及区域发展特征.

中国互联网指数系统中的人气指数是以 Alexa.com 的数据为基础进行计算,选取各个行业排名靠前的网站为成分网站,对其访问量(IP 值)及人均页面访问数(PV)进行加权计算得出平均值,其他网站与此值相比,得到各自的人气指数值.本文利用的正是新闻源网站人气指数(CIIS 值),再把此指数归一化作为新闻信源网站的质量评估值,即新闻影响力因子 W_S,也就是新闻影响力评估的一个整体参数.

2.4 新闻回复率计算

回复率(记为 *Rep*)直接体现了人们对网络新闻产生的反应.通常,

回复率=回复次数/点击次数 (8)

通过观察我们发现,大部分新闻网页只是提供了回复人次,而没有提供点击/浏览人次,并且网页中点击/浏览次数是在网页服务器端存储的,通过简单的抓取和信息抽取很难得到.在大量观察的基础上,根据新闻回复次数的相对数量总结了一个回复率比值,把这个比值作为新闻的回复率.此处,回复次数是源网站回复次数和转载网站回复次数的总和.新闻回复次数分布如图 2 所示.

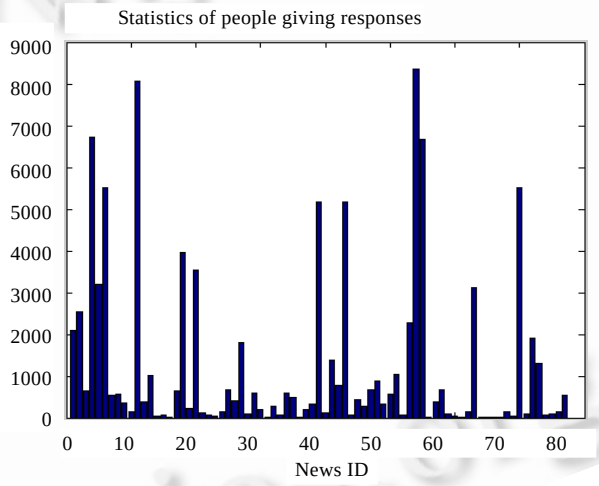

Fig.2 Statistics of number of replys

图 2 回复人次统计

从图 2 我们可以得出:大多新闻的回复次数是在 1 000 人次以内的.极少数是在 3 000 人次以上.根据上图统计规律得出相对回复率值如表 1 所示.举例说明:回复次数(0~500)表示对本条新闻发出回复的人数范围,相对回复率比值表示在发出回复人数为(0~500)之间时,我们可以认为对本条新闻发出回复的人数占浏览人数的 5%.如果回复人数超过了 5 000,表示浏览过本条新闻的人基本上都发出了回复,所以相对回复率为 100%.

Table 1 List of relative response-rate

表 1 相对回复率列表

Number of replys	Relative response-rate (%)
5 000~	100
4 500~5 000	90
4 000~4 500	80
3 500~4 000	70
3 000~3 500	60
2 500~3 000	50
2 000~2 500	40
1 500~2 000	30
1 000~1 500	20
500~1 000	10
0~500	5

2.5 时间要素对新闻排序的影响

人们对新闻的关注程度变化趋势一般有两种,如图 3 所示.此处关注程度用单位时间内浏览新闻的人次来衡量.第 1 种是缓慢增长型,例如对国家政策类新闻等知识类的关注度.这些类别的新闻时效性不强,人们对它们的关注度是随着时间的推移缓慢增长的.另外一种则是快速增长下降型.主要是针对时事类的新闻,这类新闻的时效性很强,人们对这类新闻的关注度在短时间内快速增长,经过一段时间之后,关注度快速下降.因而在对新闻排序时一定要首先进行类别判断,然后考虑时间要素产生的影响.从这方面看,新闻重要性与发布时间成反比关系.

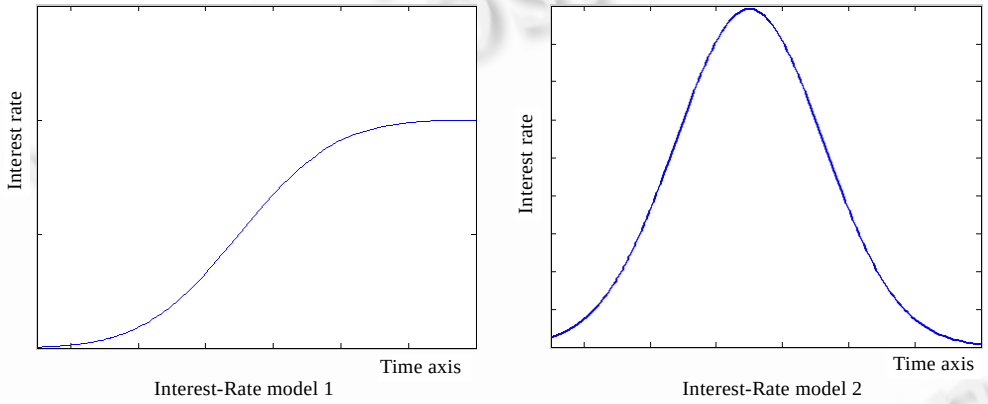

Fig.3 Model of news interest-rate

图 3 新闻关注度

另外,发布时间越长,被转载和被回复的几率越大,回复次数和转载次数越多.如果不考虑时间因素对新发布的新闻是不公平的.所以必须选定一个参数作为时间因素对新闻重要性产生影响的平衡.对发布时间长的新闻在回复次数和转载次数做一些削减.总结以上两点并结合文献[4]中对新闻衰退时间参数的定义,我们定义时间参数定义如下:

$$D(t_s,t)=\mathrm{e}^{-\alpha(t-t_s)} \tag{9}$$

其中,t_s 为新闻的发布时间,并且有 $t\geq t_s$.α的确定取决于新闻所属新闻类别的衰退时间,衰退时间是指新闻从发布到无人关注中间经历的时间,这里定义α与新闻衰退时间之间的关系为

$$\alpha=\begin{cases}\mathrm{e}^{-\beta\alpha}=0.5, & \text{时事类新闻}\\ \mathrm{e}^{-\gamma\alpha}=0.5, & \text{非时事类新闻}\end{cases} \tag{10}$$

其中,β为时事类新闻衰退时间,γ为非时事类新闻衰退时间.

2.6 新闻影响力判断

通过以上步骤,我们可以得到如下数据:新闻转载率(*Trans*),新闻回复率(*Rep*),新闻信源网站的影响力因子

(W_s),以及新闻发布时间参数 $D(t_s,t)$.

我们认为,对新闻进行转载和回复即为人对新闻的认可,所以把网络新闻认可率(记为 *Rec*)定义为

新闻认可率=a×转载率+b×回复率 (11)

为了保证认可率为小于 1 的数值,其中的 a 和 b 的关系定义为 $a+b=1$.因为没有合适的语料库,无法通过训练方法得到 a 和 b 的值,所以它们的确定借助于 80/20 法则而得到.这里可以理解为:浏览新闻的人也许很多,但是做出回复的人是极少的,而做出转载行为的人更少.所以我们认为转载率更能体现新闻的影响力.实验表明这种定义方法是可行的.

最后综合以上信息,定义新闻的影响力(NF)如下:

$$N_F = \mathrm{e}^{-\alpha(t-t_s)} \times W_s \times (a \times Trans + b \times Rep) \quad (12)$$

其中,a=0.8,b=0.2,*Trans* 计算过程见第 2.2 节,*Rep* 计算过程见第 2.4 节.

3 实验及结果

本节首先介绍如何收集数据集,然后给出一部分实验结果以及结论.

3.1 实验数据集

我们跟踪中文网站排行榜中在新闻资讯方面排名靠前的 9 个网站一周,因为在做算法研究时,总结的特征大多根据中文网页产生,所以实验中仍然采用中文网页作为实验对象.对这 9 个网站的首页新闻每一小时抓取一次.抓取的实验数据列表见表 2.

Table 2 Web pages sets for test

表 2 实验网页列表

Source	News page volume
Tencent News	913
Sina News	1 287
Xinhua	1 091
PRC News	1 798
Sohu News	1 103
NetEase News	1 193
Dongfang News	1 156
Tom News	1 556
China News	1 757

进行网站内部去重处理之后,这些网页按内容被分成了 6 类,各类别新闻分布情况见表 3.这里,新闻的时效性强弱是根据新闻所属关注度模型得出.

Table 3 Classification of test corpus

表 3 实验网页分类情况列表

Category	Percentage of total (%)	Timeliness
Events/News	57	Strong
Business	8	Strong
Health	2	weak
Sports/Entertainment	7	Strong
Military	3	Strong
Technology	13	weak

3.2 实验结果

3.2.1 2007 年 9 月 10 日~16 日一周新闻影响力排序

表 4 给出了 2007 年 9 月 10 日~16 日一周时间内排名前 10 位的新闻以及它们的影响力值.这里定义时事类新闻的衰退时间为 72 小时,而非时事类新闻的衰退时间为 120 小时.

图 4 给出了一周内新闻影响力值的分布图.

Table 4 Top ten news articles during all the observation period from 2007.9.10 to 2007.9.17

表 4 2007 年 9 月 10 日~17 日排名前 10 位的新闻列表

ID	News summary	Source	Category	Post time	Force value
1	The revenue growth gap between urban and rural residents expanded, the ratio of last year was 3.28:1	Xinhua	Business	9.13	0.682 0
2	U.S. officials said that the Taiwan authorities seek to change the national title by the referendum	Xinhua	Events/News	9.11	0.651 7
3	The launch of Japan's lunar exploration satellite “goddess of the moon” has been postponed	Xinhua	Technology	9.11	0.651 7.
4	Shanghai World Financial Center has been completed, and it becomes the highest building of China	Xinhua	Events/News	9.14	0.636 6
5	Chinese warships with the British aircraft carrier have held a joint maritime military exercises	Xinhua	Military	9.10	0.616 8
6	Chinese government asked for consultation as a means of settlement over US anti-dumping measures against Chinese coated paper	Xinhua	Events/News	9.14	0.616 8
7	Japanese Prime Minister Abe formally announced his resignation at 14:00 on the 12th	Xinhua	Events/News	9.12	0.610 8
8	A study of the United States shows that the use of anti-thrombosis drug leads to more bleeding for Asians	Xinhua	Health	9.14	0.610 8
9	Chinese military helicopters will implement the 5th launch for searching the six missing Russian	Xinhua	Events/News	9.14	0.610 8
10	The CPC improves its governing capability while handling of major emergencies	Xinhua	Events/News	9.10	0.609 3

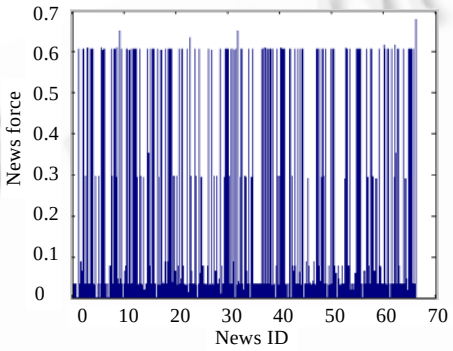

Fig.4 Distribution curve of news-force during all observation period from 2007.9.10 to 2007.9.16

图 4 2007 年 9 月 10 日～2007 年 9 月 16 日所收集新闻的影响力值分布图

3.2.2 指定新闻影响力排序

本算法同样适用于对指定新闻的排序,给定若干条主题不同的新闻,把这些新闻的主题作为关键词,利用现在比较流行的搜索引擎作为工具搜取相关的新闻网页.从检索结果中选取前 50 个进行统计计算.取前 50 个是因为它们基本上包含了所有的转载网页和大部分的近似网页,以及中文网站排行榜中在新闻资讯方面排名前 50 的网站.足以确定新闻的信源网站和新闻转载率.而且我们经过大量的浏览网页发现,所有的网友评论都是在排名靠前的几个网站中,其他网站网友回复几乎为 0,因而选取这些网页也足以得到较为准确的新闻回复率值.得到每个新闻主题及转载网页之后,抽取相关信息,然后对各个主题的新闻根据以上算法进行影响力分析,求得影响力系数,根据影响力系数排序,得到一个定量分析的排序结果.然后,通过调查多个人对这几个主题排序的结果,进行综合之后得到一个人工的定性分析的排序结果.最后比较两个结果的一致性.通过多次比较结果发现,由此方法计算得到的排序结果与人工排序结果基本一致.举例说明比较结果见表 5、表 6.这里为不相关话题新闻之间影响力的比较,实验得出此方法对相关话题同样适用.

Table 5 Ranking result of news-force for given news articles by man

表 5 指定新闻影响力人工排序的结果

ID	News summary	Source	Category	Post time	Force
1	NPC deputy Chen Liangyu was removed from office duties and functions of Shanghai people's congress deputies in accordance with the law	Xinhua News	Events/News	7.27	Strong
2	Seventy people were trapped in Shanxian coal mine flooding accident in Henan	Xinhua News	Events/News	7.27	Strong
3	Basic medical insurance is expected to cover the urban non-business residents in 2010	China News	Events/News	7.24	Strong
4	The convention signed and sent to the United Nations by Chen Shui-bian was returned by the Ban Ki-moon	Dongfang News	Events/News	7.19	Fair
5	The SDRC responses to the rising of house prices in June, and measures will be taken to stop “Chaofang”	China News	Events/News	7.25	Fair
6	Shandong Jinan was attacked by torrential rain	Qilu Evening News	Events/News	7.19	Fair
7	Pork prices rises to the highest point in some places of China	The Beijing News	Events/News	7.20	Fair
8	Chinese new navy fleet will go to Europe to participate in joint military exercises	PLA Daily	Military	7.25	Weak
9	The Taliban kidnaps 23 South Koreans, including 15 women	China Daily	Events/News	7.21	Weak
10	Iraq held Asian Cup champion for the first time	Sina Sports	Sports	7.29	Weak

Table 6 Ranking result of news-force for given news articles by evaluation model

表 6 指定新闻影响力定量分析排序的结果

ID	News summary	Source	Category	Post time	Force value
1	NPC deputy Chen Liangyu was removed from office duties and functions of Shanghai people's congress deputies in accordance with the law	Xinhua News	Events/News	7.27	0.601 4
2	Seventy people were trapped in Shanxian coal mine flooding accident in Henan	Xinhua News	Events/News	7.27	0.577 4
3	Basic medical insurance is expected to cover the urban non-business residents in 2010	China News	Events/News	7.24	0.083 3
4	The convention signed and sent to the United Nations by Chen Shui-bian was returned by the Ban Ki-moon	Dongfang News	Events/News	7.19	0.072 0
5	The SDRC responses to the rising of house prices in June, and measures will be taken to stop “Chaofang”	China News	Events/News	7.25	0.034 1
6	Iraq held Asian Cup champion for the first time	Sina Sports	Sports	7.29	0.004 3
7	Shandong Jinan was attacked by torrential rain	Qilu Evening News	Events/News	7.19	0.004 2
8	Pork prices rises to the highest point in some places of China	The Beijing News	Events/News	7.20	0.003 8
9	Chinese new navy fleet will go to Europe to participate in joint military exercises	PLA Daily	Military	7.25	0.003 7
10	The Taliban kidnaps 23 South Koreans, including 15 women	China Daily	Events/News	7.21	0.003 5

4 结论及展望

本文针对网络新闻对信息内容安全方面的影响,提供了一种定量计算新闻影响力的模型,力求客观地衡量新闻的影响力.它可以作为一种人们对新闻导向舆论能力判断的辅助工具,也可以用于新闻搜索引擎中改善结果排序.本文只是利用了新闻网站或网页中比较浅显的信息,对新闻影响力进行了定量计算.同时,因为此算法的复杂度与所要比较的新闻网页数量呈线性关系,所以可以用于在线的新闻排序.作为后续工作,如何加入对新闻事件内容的分析提高网络新闻影响力判断的准确性,以及如何把它应用到实际的新闻搜索引擎中进行在线排序,这都是值得继续深入研究的问题.

References:

[1] Huang L. On characteristics of communicational function of networking media. Journal of Huazhong University of Science and Technology (Social Science Edition), 2000,14(2):115−117 (in Chinese with English abstract).

[2] 2007. http://searchenginewatch.com/

[3] Radev DR, Blair-Goldensohn S, Zhang Zh, Raghavan RS. Interactive, domain-independent identification and summarization of topically related news articles. In: Constantopoulos P, Sølvberg I, eds. Proc. of the ECDL 2001. London: Springer-Verlag, 2001. 225−238.

[4] Radev DR, Blair-Goldensohn S, Zhang Zh, Raghavan RS. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In: Marcus M, ed. Proc. of the HLT 2001. Morristown: Association for Computational Linguistics, 2001. 1−4.

[5] Dunlavy DM, Conroy J, O'Leary DP. QCS: A tool for querying, clustering, and summarizing documents. In: Hearst M, Ostendorf M, eds. Proc. of the NAACL 2003. Tarrytown: Pergamon Press, Inc., 2003. 11−12.

[6] Del Corso GM, Gullí A, Romani F. Ranking a stream of news. In: Ellis A, Hagino T, eds. Proc. of the WWW 2005. New York: ACM, 2005. 97−106.

[7] Yao JY, Wang J, Li ZW, Li MJ, Ma WY. Ranking Web news via homepage visual layout and cross-site voting. In: Lalmas M, ed. Proc. of the ECIR 2006. Heidelberg: Springer-Verlag, 2006. 131−142.

[8] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999,46(5):604−632.

[9] Wang JY, Xie ZM, Lei M, Li XM. Research and evaluation of near-replicas of Web pages detection algorithms. Acta Electronica Sinica, 2000,28(11A):130−132 (in Chinese with English abstract).

[10] 2007. http://top.chinalabs.com/

附中文参考文献:

[1] 黄鹂.论网络媒体传播功能的特点.华中理工大学学报(社会科学版),2000,14(2):115−117.

[9] 王建勇,谢正茂,雷鸣,李晓明.近似镜像网页检测算法的研究与评价.电子学报,2000,28(11A):130−132.

杨伟杰(1980－),女,山东潍坊人,博士生,主要研究领域为信息检索,模式识别与智能系统.

戴汝为(1932－),男,研究员,博士生导师,中国科学院院士,主要研究领域为自动控制,模式识别,人工智能,复杂性科学及思维科学.

崔霞(1975－),女,博士,副研究员,主要研究领域为复杂性科学,模式识别与智能系统,信息检索.