[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
H04L 12/24 (2006.01)
G06F 17/30 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200710175668.X

[43] 公开日 2009 年 4 月 15 日

[11] 公开号 CN 101409634A

[22] 申请日 2007.10.10
[21] 申请号 200710175668.X
[71] 申请人 中国科学院自动化研究所
地址 100080 北京市海淀区中关村东路 95 号
[72] 发明人 杨伟杰 戴汝为 崔 霞 王春恒
[74] 专利代理机构 中科专利商标代理有限责任公司
代理人 梁爱荣

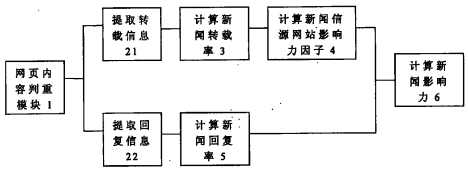
权利要求书 3 页 说明书 14 页 附图 3 页

[54] 发明名称

基于信息检索的互联网新闻影响力定量分析工具及方法

[57] 摘要

基于信息检索的互联网新闻影响力定量分析工具及方法，工具包括：网页内容判重模块判断网页是否为近似网页；相关信息提取模块提取网页中相关信息；新闻转载率计算模块计算转载网站权威度值；找出新闻源网站和新闻转载率；新闻信源网站影响力确定模块判断新闻源网站的人气指数，获取新闻信源网站影响力因子；新闻回复率计算模块确定网络新闻回复率；新闻影响力计算模块计算新闻影响力值。 方法是判断网页是否为近似网页，提取新闻网页转载信息和回复信息，计算新闻转载率，计算新闻的回复率，计算新闻源网站的权威度，计算新闻源网站的影响力因子，计算新闻影响力。 本发明定量计算结果与用户的定性分析结合，帮助用



1、一种基于信息检索的互联网新闻影响力定量分析工具，其特征在于，包括：

网页内容判重模块，接收网页内容用于判断网页是否为近似网页；

信息提取模块，接收近似网页信息，从近似网页中抽取后续计算需要的相关信息；

新闻转载率计算模块：接收信息提取模块中提取到的相关信息，计算转载网站的权威度值，之后把权威度最高的那个网站作为新闻的源网站，并把此网站的权威度值作为新闻转载率；

新闻信源网站影响力确定模块：用于判断新闻源网站的人气指数 CIIS 值，并把此指数归一化之后作为新闻信源网站影响力因子；

新闻回复率计算模块：用于确定网络新闻的回复率；

新闻影响力计算模块：用于通过以上模块得到的网络新闻转载率值、新闻信源网站的影响力因子值和新闻回复率值计算新闻影响力值。

2、根据权利要求 1 所述的互联网新闻影响力定量分析工具，其特征在于：所述网页内容判重模块：对整篇文档采用 MD5 散列值方法判重，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他们是否为近似网页。

3、根据权利要求 1 所述的互联网新闻影响力定量分析工具，其特征在于：相关信息提取模块还包括：

网页重复信息提取模块：如果判定两个网页为相似网页，则由此模块提取转载重复信息；主要是源网站以及转载网站之间的关系，包括直接转载和间接转载关系；

网页回复信息提取模块：用于提取源网站与转载网站中对新闻的回复次数，然后去除转载或相似的网页。

4、根据权利要求 1 所述的互联网新闻影响力定量分析工具，其特征在于：新闻转载率计算模块：利用相关信息提取模块中提取到的新闻转载网站之间的关系，利用 HITS 算法，计算转载网站的权威度值；计算之后将入链最多的那个网站作为新闻的源网站，并将此网站的权威度值作为新闻转载率。

5、根据权利要求 1 所述的互联网新闻影响力定量分析工具，其特征在于：新闻信源网站影响力确定模块：用于判断新闻源网站的人气指数 CIIS 值，并把此指数归一化之后作为新闻信源网站影响力因子。

6、根据权利要求 1 所述的互联网新闻影响力定量分析工具，其特征在于：所述新闻回复率计算模块，在浏览网页之后，根据新闻回复次数的相对数量总结一个回复率表，通过查找表中对应范围的回复率作为新闻的回复率。

7、根据权利要求 1 所述的互联网新闻影响力定量分析工具，其特征在于：所述新闻影响力计算模块，用于利用网页内容判重模块、相关信息提取模块、新闻转载率计算模块、新闻信源网站影响力确定模块、新闻回复率计算模块得到的数据计算新闻影响力，

$$NF = D(t_s,t) \times W_s \times (a \times \mathrm{Trans} + \mathrm{b} \times \mathrm{Rep})$$

其中，NF 为新闻的影响力；W_s 为新闻信源网站的影响力因子；Trans 为新闻转载率；Rep 为新闻回复率；$D(t_s,t)$ 为新闻发布时间与它的影响力之间的关系；a=0.8；b=0.2。

8、一种基于信息检索的互联网新闻影响力定量分析方法，其特征在于，包括：

（1）根据网页内容判断网页是否为转载或者重复网页；

（2）提取重复网页中的相关信息；

（3）计算新闻网页转载率；

（4）计算新闻信源网站的影响力；

（5）计算新闻回复率；

（6）使用上述步骤所得数据计算新闻影响力。

9、根据权利要求 8 所述互联网新闻影响力定量分析方法，其特征在于：所述新闻网页转载率计算步骤还包括：利用 HITS 算法把一个转载网站作为一个节点，网站之间存在的转载关系和原来算法中的 hub 属性相对应，计算转载网站的权威度值；计算之后把入链最多的那个网站作为新闻的源网站，并把求得源网站的权威度值作为新闻转载率。

10、根据权利要求 8 所述互联网新闻影响力定量分析方法，其特征在于：所述新闻信源网站的影响力计算步骤还包括：利用中国互联网指数系统中的网站人气指数 CIIS 值，确定对应信源网站的人气指数，然后归一化之后作为新闻信源网站的影响力因子。

基于信息检索的互联网新闻影响力定量分析工具及方法

技术领域

本发明涉及网络信息内容安全领域，具体地说是涉及网络信息内容安全领域中网络新闻影响力分析的实现方法。

背景技术

作为一种新兴的信息传播的方式，网络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，它不是国家权力，但随着新闻媒体在社会政治、经济、文化生活中的作用日益增强而变得越来越重要，发挥着重要影响力。因而确定新闻的影响力对把握社会舆论的动向，从而确定新闻对社会安全的影响具有重要意义。

在此之前，对网络新闻的分析主要为社会科学领域进行的一些定性分析，没有一个定量的工具来验证定性分析的正确性。因而我们提出了一种借助于信息检索的相关技术，获取相关的信息，对新闻影响力进行定量分析的方法。

此方法主要是通过对网页进行判重处理以及提取网页中的相关信息。然后利用这些信息判断互联网新闻影响力。主要思路为：第一步，对新闻网页进行去噪，提取内容块，然后对其进行相似性判断。如果判断为重复网页则提取网页相关信息并记录重复信息，以备以后计算时使用。第二步，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认可率计算。第三，将中国互联网指数系统对新闻的源网站的 CIIS 值进行归一化之后作为新闻影响力判断的一个比例因子。第四，根据新闻转载网站之间的

链接关系，利用 HITS 算法对新闻源网站进行权威度计算，最终对以上信息进行综合计算得出新闻的影响力。

发明内容

为了解决现有技术对网络新闻的分析主要为社会科学领域进行的一些人为的定性分析，没有一个定量的工具来验证定性分析的正确性的缺陷，本发明的目的在于提供一种基于信息检索技术、有效衡量互联网新闻影响力的定量分析工具或称为装置及方法，衡量新闻影响力结果与用户的定性分析相结合，可以帮助用户对新闻影响力大小进行有效的判断。

为了实现所述目的，本发明一方面，提供一种基于信息检索技术的互联网新闻影响力定量分析工具，包括：

网页内容判重模块，用于判断网页是否为近似网页；

信息提取模块，接收近似网页信息，从近似网页中抽取后续计算需要的相关信息；

新闻转载率计算模块：接收信息提取模块中提取到的相关信息，计算转载网站的权威度值,之后把权威度最高的那个网站作为新闻的源网站，并把此网站的权威度值作为新闻转载率；

新闻信源网站影响力确定模块：用于判断新闻源网站的人气指数 CIIS 值，并把此指数归一化之后作为新闻信源网站影响力因子；

新闻回复率计算模块：用于确定网络新闻的回复率；

新闻影响力计算模块：用于通过以上模块得到的网络新闻转载率值、新闻信源网站的影响力因子值和新闻回复率值计算新闻影响力值。

根据本发明的实施例，所述网页内容判重模块：对整篇文档采用 MD5 散列值方法判重，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他们是否为近似网页。

根据本发明的实施例，相关信息提取模块还包括：

网页重复转载信息提取模块：如果判定两个网页为相似网页，则由此模块提取转载重复信息；主要是源网站以及转载网站之间的关系，包括直接转载和间接转载关系；

网页回复信息提取模块：用于提取源网站与转载网站中对新闻的回复次数，然后去除相似网页。

根据本发明的实施例，新闻转载率计算模块：利用相关信息提取模块中提取到的新闻转载网站之间的关系，利用 HITS 算法，计算转载网站的权威度值；计算之后将入链最多的那个网站作为新闻的源网站，并将此网站的权威度值作为新闻转载率。

根据本发明的实施例，新闻信源网站影响力确定模块：用于判断新闻源网站的人气指数 CIIS 值，并把此指数归一化之后作为新闻信源网站影响力因子。

根据本发明的实施例，所述新闻回复率计算模块，在浏览网页之后，根据新闻回复次数的相对数量总结一个回复率表，通过查找表中对应范围的回复率作为新闻的回复率。

根据本发明的实施例，所述新闻影响力计算模块，用于利用网页内容判重模块、信息提取模块、新闻转载率计算模块、新闻信源网站影响力确定模块、新闻回复率计算模块得到的数据计算新闻影响力为：

$$NF = D(t_s,t) \times W_s \times (a \times \mathrm{Trans} + \mathrm{b} \times \mathrm{Rep})$$

其中，NF 为新闻的影响力；W_s 为新闻信源网站的影响力因子；Trans 为新闻转载率；Rep 为新闻回复率；$D(t_s,t)$ 为新闻发布时间与它的影响力之间的关系；a=0.8；b=0.2。

为了实现所述目的，本发明另一方面，提供一种基于信息检索技术的互联网新闻影响力定量分析方法，包括步骤如下：

（1）根据网页内容判断网页是否为转载或者重复网页；

（2）提取重复网页中的相关信息；

（3）计算新闻网页转载率；

（4）计算新闻信源网站的影响力；

（5）计算新闻回复率；

（6）使用上述步骤所得数据计算新闻影响力。

根据本发明的实施例，所述新闻网页转载率计算步骤还包括：利用 HITS 算法，利用 HITS 算法是把一个转载网站作为一个节点，网站之间存在的转载关系和原来算法中的 hub 属性相对应，计算转载网站的权威度值；计算之后把入链最多的那个网站作为新闻的源网站，并把求得源网站的权威度值作为新闻转载率。

根据本发明的实施例，所述新闻信源网站的影响力计算步骤还包括：利用中国互联网指数系统中的网站人气指数（CIIS 值），确定对应信源网站的人气指数，然后归一化之后作为新闻信源网站的影响力因子。

本发明提供了一种基于信息检索技术的互联网新闻影响力定量分析工具装置及方法，本发明的计算可以得到一个对网络新闻影响力的定量评估，通过把此定量分析结果与人为定性分析结果进行比较，可以有效判断网络新闻影响力大小。本发明解决了现有技术对网络新闻的分析主要为社会科学领域进行的一些人为的定性分析，没有一个定量的工具来验证定性分析的正确性的缺陷，有效衡量新闻影响力的定量分析，衡量新闻影响力的结果与用户的定性分析相结合，可以帮助用户对新闻影响力大小进行有效的判断。

附图说明

图 1 是本发明的原理示意图；

图 2 是本发明中相关信息提取模块框图；

图 3 是本发明方法的实施例流程图；

图 4 是本发明方法的实施例的回复人次规律统计；

图 5 是本发明时间因素对新闻影响力影响曲线图。

具体实施方式

下面结合附图对本发明作进一步详细的描述。

为了能够有效的确定新闻影响力，我们充分利用了新闻网页的一些特性。我们通过判重处理发现新闻网页的转载或者相似网页，然后抽取其中的转载信息和回复信息，并计算得到新闻的转载率和回复率，最后利用新闻信源网站的 CIIS 值作为最终的比例因子，利用公式计算得到新闻的影响力。以图 1 为例：

本发明系统的结构包括：

网页内容判重模块 1：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他们是否为近似网页。

如图 2 所示信息提取模块 2 还包括：

网页重复信息提取模块 21：如果判定两个网页为相似网页，则由此模块提取重复信息。主要是源网站以及转载网站之间的关系。包括直接转载和间接转载关系。

网页回复信息提取模块 22：用于提取源网站与转载网站中对新闻的回复次数。然后去除转载或相似的网页。

新闻转载率计算模块 3：利用相关信息提取模块 2 中提取到的新闻转载网站之间的关系，利用 HITS 算法是把一个转载网站作为一个节点，网站之间存在的转载关系和原来算法中的 hub 属性相对应，计算转载网站的权威度值。计算之后把权威度最高的那个网站作为新闻的源网站。并把求得源网站的权威度值作为新闻转载率。

新闻信源网站影响力确定模块 4：利用中国互联网指数系统中的网站人气指数（CIIS 值），确定对应信源网站的人气指数，然后归一化之后作为新闻信源网站的影响力因子。

新闻回复率计算模块 5：用于确定网络新闻的回复率。然而网页中点击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是很难得到的。但是回复次数是很容易就可以得到的。因而我们

在浏览了大量网页之后，根据新闻回复次数的相对数量总结了一个回复率表，通过查找表中对应范围的回复率作为新闻的回复率。

新闻影响力计算模块 6：用于结合网页内容判重模块、相关信息提取模块、新闻转载率计算模块、新闻信源网站影响力确定模块、新闻回复率计算模块得到的数据，根据公式计算新闻影响力。

图 3 是本发明所述方法的实施例流程图。按照图 3，本发明包括六个主要部分：

一是新闻网页判重；

二是提取新闻网页中的信息；

三是计算新闻转载率；

四是计算新闻源网站的影响力因子；

五是计算新闻的回复率；

六是计算新闻影响力。

首先在步骤 1 判断获得的一个新网页是否为转载网页，如果是执行步骤 3，否则执行步骤 2；

步骤 2：判断新网页是否为相似网页，如果是转步骤 3，否则重新获得一个新网页并返回步骤 1；

步骤 3：提取相似网页和转载网页的转载关系信息并执行步骤 4；

步骤 4：提取转载关系信息的回复信息并执行步骤 5；

步骤 5：根据网页之间的转载关系，计算各个网站权威度，确定源网站，并执行步骤 6；

步骤 6：计算回复信息，获取新闻回复率，并执行步骤 7；

步骤 7：计算新闻源网站影响力因子，并执行步骤 8；

步骤 8：计算新闻影响力因子，然后结束；

在图 3 的实施例中，对网页的判重及重复信息记录和利用主要方法如下：对于新闻来说，重复一般源于转载或对同一事件的不同报道，而且重复网页在净化之后进行信息提取得到的信息在结构和内容方面能够保持高度的一致性。这一部分我们主要是提取网络新

闻转载相关的信息。在对网页进行净化之后，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他们是否为近似相同，而网页主体内容采用向量空间模型（VSM）进行表示。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经验阈值时认为它们为近似相同，为重复网页。网页 U_i（$i \in [1,n]$）使用特征向量进行表示，其关键词权值 W_e 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：

$$W_e = \begin{cases} idf_e * 5 & \text{，e为命名实体} \\ idf_e & \text{，e为其他} \end{cases}$$

最后选取 m 个权值较大的词项生成网页特征向量，通过比较两个网页的特征向量中共现词项数量作为比较相似性的依据，如果共现个数大于预先设定的阈值，则认为这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，然后从网页集中去掉重复网页。

对网页进行判重之后，需要记录的主要信息有：

（1）转载网站（2）转载网站的信源网站（3）转载网站中的回复次数（4）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源网站和转载网站。最后的源网站在下一步中确定。

利用上一步中提取到的信息我们可以计算新闻转载率。一般情况下，新闻转载率（记为 Trans）＝转载次数/源网站点击次数，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是保存在服务器端，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关系，与 HITS 算法的初衷及其相似， HITS 算法中的 Authority 和 Hub 属性很自然地对应着网页自身的内容质量和它所链接指向的网页的质量。同样，本文中是把一个网站的内容质量和它的转载网站的质量与 HITS 算法中的 Authority 和 Hub 属性对应。把一个转载网站作为一个节点，网站之间存在的转载关系和原来算法中的 Hub 属性相对应，因而可知本文中 HITS 算法的应用与原 HITS 算法完全一致。而且本文利用 HITS 算法可以更加准确的计算新闻转载网站之间的关系。

具体算法如下：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。

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

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

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

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

按以上公式迭代更新每个节点的属性 $A_0(pt), A_1(pt)$。

利用提取到的转载信息，首先提取新闻转载网站之间的关系，此处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。

在图 3 的实施例中，源网站 CIIS 值的确定过程如下：

中文网站排行榜是中国互联网指数系统（CIIS）的重要组成部分，是互联网实验室的核心产品。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（CIIS，China Internet Index System）由互联网实验室与国家统计局于 2004 年联合发布。中国互联网指数系统（CIIS）由四大指数体系组成，分别是：

1.中国互联网基础指数

2.中国互联网满意度指数

3.中国互联网表现指数

4.中国网络股指数

其中表现指数是在互联网表现层描述互联网经济，利用 Alexa.com 作为第三方监测机构。又细分为三个重要指数：

1.网站人气指数（CIIS 值）

2.网站综合指数

3.网站结构指数

其中的人气指数是以 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。我们此文中利用的正是新闻源网站人气指数（CIIS 值），在把此指数归一化之后作为新闻重要性评估的又一个参数。

新闻回复率确定过程如下：

回复率直接体现了人们对网络新闻产生的反应。一般情况下，

回复率＝回复次数/点击次数

然而通过观察我们发现，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，而且网页中点击/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是很难得到的。回

复次数通过对网页进行信息抽取是很容易就可以得到的。因而我们在浏览了大量网页之后，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，我们使用的回复次数是源网站回复次数和转载网站回复次数的总和。新闻回复次数分布图 4 是本发明方法的实施例的回复人次规律统计所示， 从图 4 我们可以得出：大多新闻的回复次数是在 1000 人次以内的。极少数是在 3000 人次以上。根据上图统计规律得出下面的相对回复率比值。举例说明：其中回复次数（0－100）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（0－100）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 10%。如果回复人数超过了 5000，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%。

相对回复率列表如下：

回复次数（人次）	相对回复率比值（%）
5000-	100
3000-5000	80
2000-3000	70
1000-2000	60
500-1000	50
300-500	40
200-300	30
100-200	20
0-100	10

计算时，根据新闻的回复人次，通过查找表中对应范围的回复率作为新闻的回复率。

时间因素对新闻影响力大小也有很大影响。人们对新闻的关注

程度变化一般为两种，如图 5 所示。第一种是缓慢增长型，例如新闻关注度模型 a，对国家政策类新闻等知识类的关注度。这些类别的新闻的时效性不强，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，例如新闻关注度模型 b。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，新闻影响力与发布时间成反比关系。

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:新闻发布时间与它的影响力之间成反比关系。时间参数定义如下：

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

其中t_s为新闻的发布时间，并且有$t\geq t_s$。α的确定取决于新闻它所属于新闻类别的衰退时间，衰退时间指新闻从发布到无人关注中间经历的时间，此处定义α与新闻衰退时间之间的关系为：

$$\alpha=\begin{cases} e^{-24\alpha}=0.5,\text{时事类新闻} \\ e^{-72\alpha}=0.5,\text{非时事类新闻} \end{cases}$$

此处定义时事类新闻的衰退时间为 24 小时，而非时事类新闻的衰退时间为 72 小时。

在图 3 的实施例中，新闻影响力判断具体过程如下：

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

我们认为对新闻进行转载和回复即为人对新闻的认可，所以此

处我们把网络新闻认可率（Rec）定义为：

新闻认可率＝a×转载率＋b×回复率；

为了保证认可率为小于 1 的数值，此处的 a 和 b 的关系我们定义为 a＋b＝1；b 的确定借助于 80/20 法则而得到。此处理解为：浏览新闻的人也许很多，但是做出回复的人是极少的，大约仅占浏览人次的 20％。

最后综合以上信息，定义新闻的影响力（NF）如下：

$$NF = D(t_s,t) \times W_s \times (a \times \mathrm{Trans} + \mathrm{b} \times \mathrm{Rep})$$

其中 a=0.8,b=0.2。

下面是一个具体实施例。从网络上选择几个主题的新闻，然后利用网络搜索引擎把新闻主题作为关键字搜索相关的页面，从查询结果中选取前 100 个按照以上计算步骤进行统计计算它们的影响力值，得到一个定量分析的结果。然后对这些值进行排序从而得到一个新闻影响力排序结果。然后通过调查多个人对这些新闻影响力的排序结果，综合之后得到一个人为定性排序结果，比较这两个结果可以发现排序结果基本一致。举例说明比较结果如下：

列表 1 人为对新闻影响力排序的结果

序号	新闻标题	影响力值
1	陈良宇被依法罢免全国及上海人大代表职务	大
2	河南陕县发生煤矿透水事故 70 人被困	大
3	2010 年基本医保有望覆盖全国城镇非从业居民	大
4	陈水扁签署公约呈送联合国被潘基文退回	一般
5	发改委回应 6 月份房价上涨,要采取措施停	一般

	止炒房	
6	山东济南遭受特大暴雨袭击	一般
7	全国多地猪肉价格涨至历史最高点	一般
8	中国海军新型舰艇编队赴欧洲参加联合军演	一般偏小
9	塔利班绑架 23 名韩国人包括 15 名妇女	一般偏小
10	亚洲杯尤尼斯头槌定乾坤,伊拉克 1-0 沙特首捧冠军	小

列表 2 对相同新闻通过影响力排序的结果

序号	新闻标题	发布时间	影响力值
1	陈良宇被依法罢免全国及上海人大代表职务	7.27	0.7936
2	河南陕县发生煤矿透水事故 70 人被困	7.27－7.30 跟踪报道	0.7619
3	2010 年基本医保有望覆盖全国城镇非从业居民	7.24	0.110
4	陈水扁签署公约呈送联合国被潘基文退回	7.19	0.095
5	发改委回应 6 月份房价上涨、要采取措施停止炒房	7.25	0.045
6	亚洲杯尤尼斯头槌定乾坤、	7.29	0.0058

	伊拉克 1-0 沙特首捧冠军		
7	山东济南遭受特大暴雨袭击	7.19－7.20 跟踪报道	0.0056
8	外交部就台申请以“台湾名义加入联合国”答问	7.20	0.005
9	中国海军新型舰艇编队赴欧洲参加联合军演	7.25	0.00487
10	塔利班绑架 23 名韩国人包括 15 名妇女	7.21－7.31 跟踪报道	0.0047

总之，在新闻影响力分析过程中采用本发明可以帮助专家评估自己定性分析的正确性，解决分析过程中只有定性分析没有定量衡量工具的问题。

以上所述，仅为本发明中的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉该技术的人在本发明所揭露的技术范围内，可理解想到的变换或替换，都应涵盖在本发明的包含范围之内，因此，本发明的保护范围应该以权利要求书的保护范围为准。

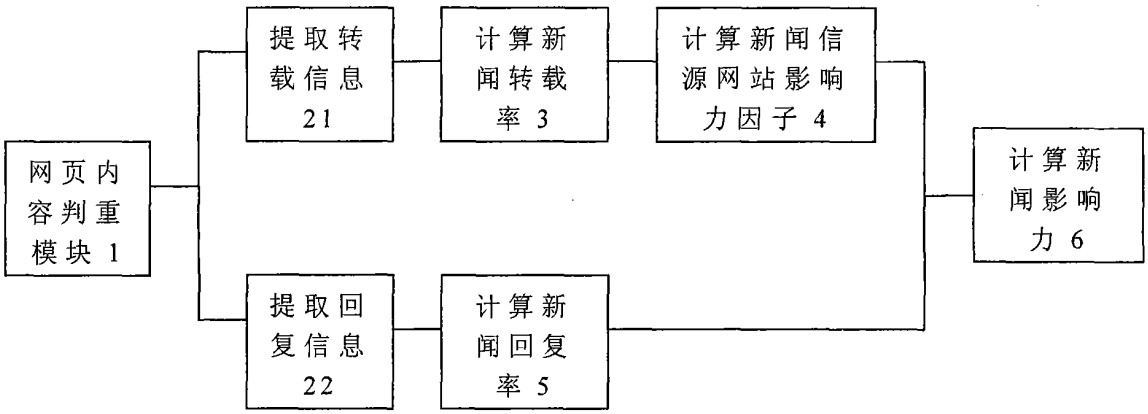


图 1

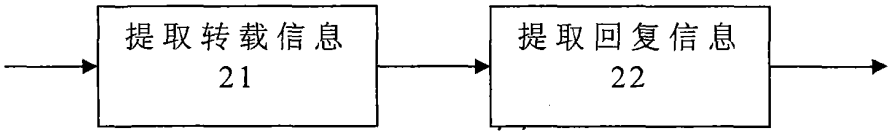


图 2

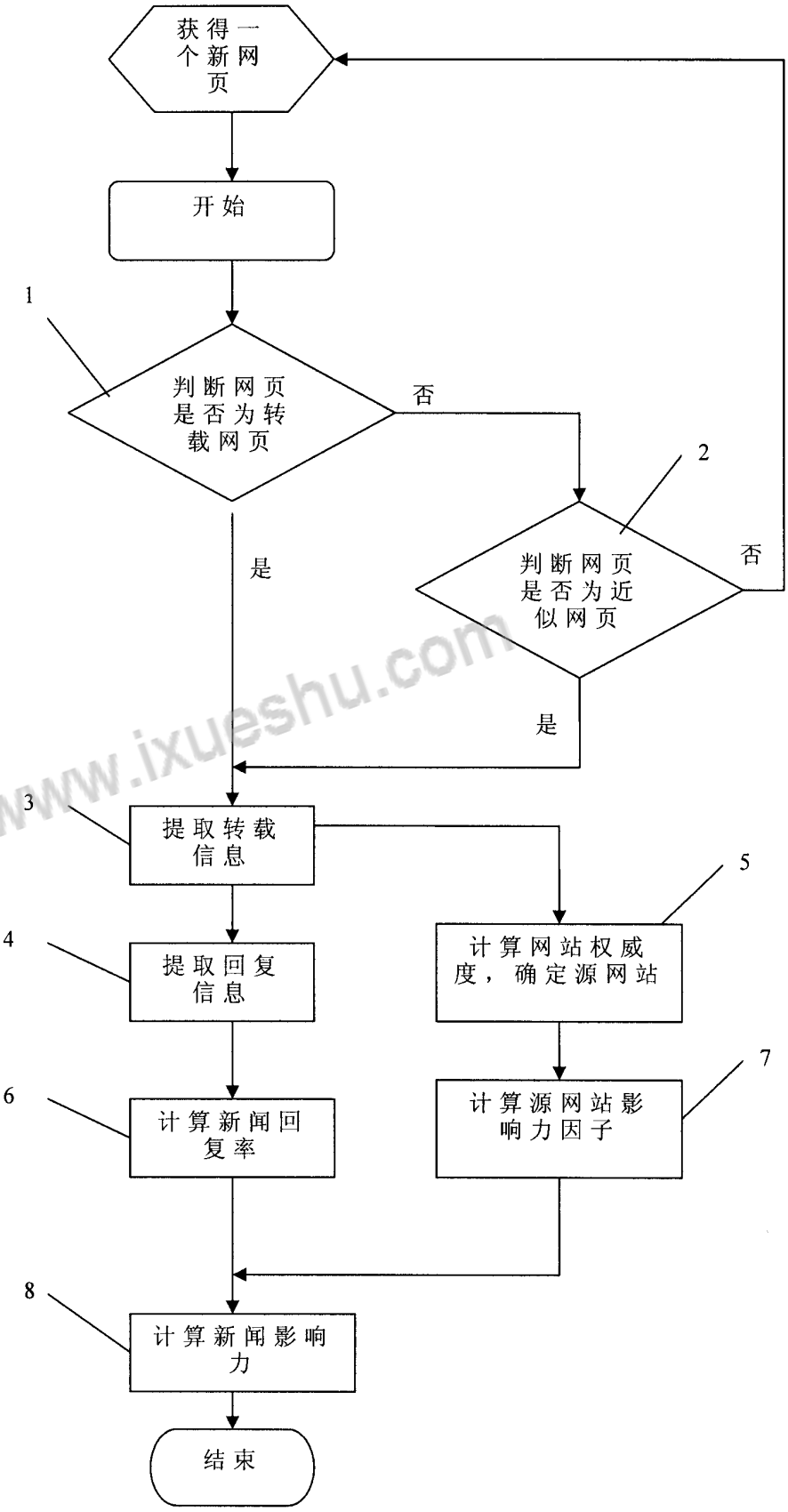


图 3

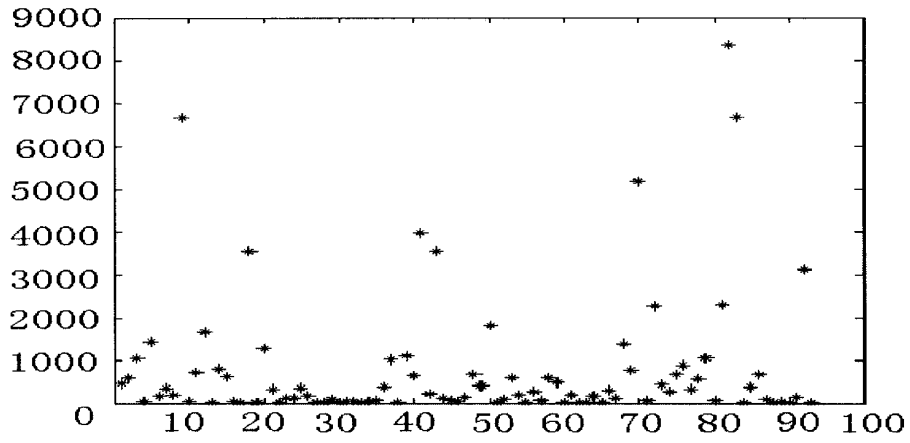
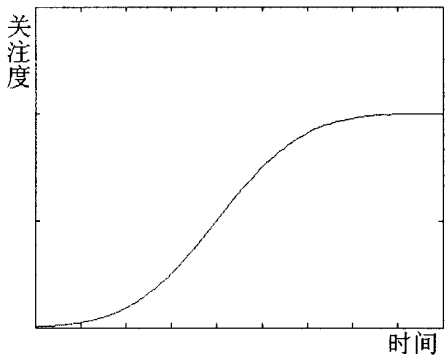
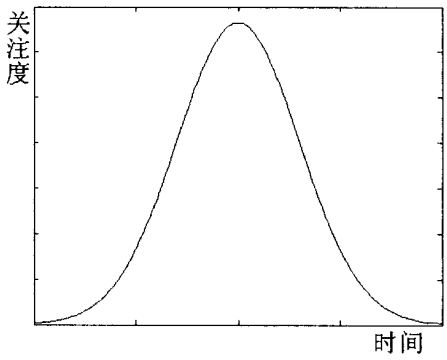


图 4



新闻关注度模型 a



新闻关注度模型 b

图 5

---

26. 网络信息检索工具的热门类目

27. 网上信息检索工具研究

28. 因特网上的信息检索工具

29. 搜索代理工具与信息检索

30. 网络信息检索工具探讨

31. 利用Gopubmed检索工具对禽流感研究的信息分析

32. 集合型网络信息检索工具

33. 法律信息检索工具研究

34. 近五年国内基于云计算的信息检索研究定量分析

35. 光学文献的几种检索工具方法

36. 网络信息检索工具研究

37. 网络信息资源检索工具和技巧

38. 基于本体的农业信息检索

39. 基于本体的信息检索

40. 浅析档案信息检索的类型及检索工具现行趋势

41. 基于本体的网络信息检索

42. 一种改进的基于文档结构的信息检索方法

43. 论各种文献信息检索工具及如何选择正确的检索工具

44. 基于本体的文本信息检索

45. 从网络信息检索工具的现状看其发展特点

46. 基于网络的信息检索与信息检索能力的培养

47. 基于语义网的信息检索

48. Yahoo开始测试新的信息检索工具

49. 浅述专题信息的检索工具和检索方法

50. 论各种文献信息检索工具及如何选择正确的检索工具