(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104317881 B
(45)授权公告日 2017.11.24

(21)申请号 201410564145.4

(22)申请日 2014.10.21

(65)同一申请的已公布的文献号
申请公布号 CN 104317881 A

(43)申请公布日 2015.01.28

(66)本国优先权数据
201410144185.3 2014.04.11 CN

(73)专利权人 北京理工大学
地址 100081 北京市海淀区中关村南大街5号北京理工大学

(72)发明人 礼欣 李懿 翟艳梅

(74)专利代理机构 北京理工正阳知识产权代理事务所(普通合伙) 11639
代理人 唐华

(51)Int.Cl.
G06F 17/30(2006.01)

审查员 刘津

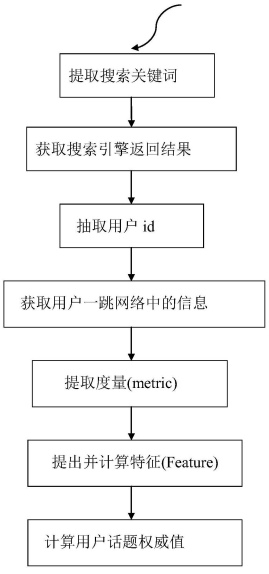
权利要求书2页 说明书8页 附图8页

(54)发明名称

一种基于用户话题权威性的微博重排序方法

(57)摘要

本发明涉及一种基于用户话题权威性的微博重排序方法，该方法通过获取用户搜索关键词信息，将用户搜索关键词划分到某个话题，然后对微博搜索引擎按照时间顺序返回来的近几天最新结果，再在该话题上对所有的用户计算话题权威值(表征该用户的话题权威性)，根据此话题权威值，再一次对搜索引擎返回的搜索结果进行重排序。本发明针对微博搜索领域，综合考虑用户话题权威性以及传统的话题权威度量，提出用户话题权威值的计算方法，并使用得到的用户话题权威值对搜索引擎按照时间顺序返回的结果进行调整，其意义在于，该方法能够提高排序后返回结果的质量，从而证明用户话题权威性在微博排序中的有效性，增强用户体验。



CN 104317881 B

1.一种用户话题权威性的计算方法,其特征在于,包括以下步骤:

步骤一、获取话题以及由微博搜索引擎返回的按照时间排序的结果集;

步骤二、在步骤一得到的结果集中抽取所有用户id;

步骤三、获取步骤二中每一个用户id一跳网络中的如下信息:

用户id的所有粉丝的id及其所有微博;

用户id所关注的所有人的id及其所有微博;

用户id的所有微博;

步骤四、从步骤三得到的结果集中提取步骤一获取的话题上的所有微博及其对应的用户信息;

步骤五、从步骤四得到的结果集中提取如下话题度量信息:

表示原创微博的度量:原创微博的数量OT1,分享链接的数量OT2、用户所有微博中的单词的重复度OT3和hashtag的数量OT4;

表示会话微博的度量:会话微博的数量CT1和由该用户发起的会话微博的数量CT2;

表示转发微博的度量:转发微博的数量RT1,原创微博中被不重复计算的其他用户转发的个数RT2和转发该用户的微博的不重复计算的所有用户的个数RT3;

表示提及的度量:该用户提及到的相同用户可重复计算的其他用户的次数M1,该用户提及到的相同用户不重复计算的其他用户的个数M2、其他用户提及到该用户的次数M3和提及到该用户的其他用户的个数M4;

表示与用户关系图相关的度量:该用户在该话题上活跃的粉丝数G1,该用户关注的人在该话题上活跃的个数G2、在该用户之后发布该话题微博的粉丝数G3和该用户关注的人中先于该用户发布该话题微博的数量G4;

表示用户总的受欢迎程度的度量:该用户总的粉丝的数量F1和该用户关注的人的总的数量F2;

步骤六、提出用户话题权威性的特征、特征计算公式并计算;

步骤七、提出用户话题权威性即用户话题权威值计算公式并计算,具体如下:

(1)、话题参与强度:$TS=\frac{OT1+CT1+RT1}{|tweets|}$;其中|tweets|表示该用户所有话题上的所有微博的数量;

(2)、原始话题强度:$SS=\frac{OT1}{OT1+RT1}$;

(3)、非会话话题强度:$\overline{CS}=\frac{OT1}{OT1+CT1}+\lambda\frac{CT1-CT2}{CT1+1}$;

其中λ用于表示用户倾向于进入微博会话的程度,较优的λ=0.9;

(4)、转推影响力:RI=RT2*log(RT3);

(5)、提及影响力:MI=M3*log(M4)-M1*log(M2);

(6)、信息传播度:ID=log(G3+1)-log(G4+1);

(7)、一跳网络得分:NS=log(G1+1)-log(G2+1);

(8)、超链接在原创微博中所占的比例:$OT21=\frac{OT2}{OT1}$;

(9)、关键词hashtag在原创微博中所占的比例：$OT41=\frac{OT4}{OT1}$；

(10)、作者微博所用词的相似度：$OT3=\frac{2*\sum_{i=1}^{n}\sum_{j=1}^{i-1}S(s_i,s_j)}{(n-1)*n}$；

其中n表示作者所有的微博数量，$S(s_i,s_j)=\frac{|s_i\cap s_j|}{|s_i|}$表示$s_i$和$s_j$的相似度，$s_i$和$s_j$是由作者的第i和第j条微博中通过去掉停用词以及做stem之后得到的单词的集合；在计算OT3之前，所有微博先按照时间排序，即times(s_i)＜times(s_j)：$\forall i<j$；

(11)、所有粉丝中该话题上有微博的粉丝所占的比例：$GF1=\frac{G1}{F1}$；

(12)、粉丝强度：$F12=\frac{F1}{F2}$；

步骤八、返回用户话题权威值计算结果。

2.根据权利要求1所述的一种用户话题权威性的计算方法，其特征在于，所述用户话题权威值计算公式如下所述：

$$AS(x_i)=[\prod_{f=1}^{11}F_f(x_i^f;\theta_f)]^{\beta}[F_{12}(x_i^{12};\theta_{12})]^{(1-\beta)}$$；

其中，x_i表示第i个用户，f表示第f个特征，x_i^f表示用户i在第f个特征上的值，F_f表示参数为θ_f的特征f的在其分布上的累积概率分布在x_i处的值，θ_f表示特征f的概率密度分布的参数，$\beta\in(0,1)$，表示在话题特征以及非话题特征之间做平衡的参数，其值由最大化皮尔逊相关系数求得。

一种基于用户话题权威性的微博重排序方法

技术领域

[0001] 本发明涉及一种微博排序方法，特别涉及一种基于用户话题权威性的微博重排序方法，属于微博搜索技术领域。

背景技术

[0002] 随着计算机技术的不断发展以及人民生活水平的不断提高，互联网越来越普及，网络资源极大丰富，这给网页搜索以及微博搜索技术提出了极大挑战。对于网页搜索，现有的比较有代表性的搜索引擎比如谷歌、百度，运用一定的策略搜集互联网上的信息，然后使用一定的方法根据用户查询关键字将检索到的信息展现给用户，而微博搜索引擎和传统的网页搜索引擎相似，区别在于检索的信息以及实用的排序机制不同。

[0003] 现有的技术中，微博搜索引擎所采用的主流技术是：当用户以关键词搜索微博时，搜索引擎会在数据库中进行查询，如果找到与该用户输入内容相符的微博，便采用一定的策略，比如说，根据该条微博被转发的次数、发表该微博的用户的权威值，以及该条微博与其他微博的内容相似度等特征，计算出每一条微博的对应值，并以此值为基础对搜索出来的微博进行排序，将得到的微博排序结果返回给用户。

[0004] 但是，上述微博搜索引擎采用的主流排序技术中，在考虑用户权威这个特征时，只是考虑的用户的比较一般化的特征，比如说，使用用户的粉丝数、用户的粉丝数与其关注的人的数量比、用户的微博被转发次数等来代表用户的权威值，并没有考虑用户在特定话题上的权威性，这些传统的衡量用户权威值的方法存在一定的弊端，因为它们这样做会使用户在所有话题上的权威值相同，而直观来讲，一个用户很有可能只对一个或几个领域精通，对其他的领域则不甚了解。

发明内容

[0005] 本发明的目的是在微博搜索领域提供一种用户话题权威性的计算方法以及一种基于用户话题权威性值的微博重排序方法，从而证明用户话题权威性值在微博搜索排序中的重要性。该方法能够根据用户输入的搜索关键词，在搜索引擎返回的按时间排序的结果集中，计算结果集中的每一个用户的话题权威值，并按照话题权威值对返回的微博进行重新排序，以此来提高返回结果的质量。

[0006] 本发明技术方案的思想是通过获取用户搜索关键词信息，将用户搜索关键词划分到某个话题，然后对微博搜索引擎按照时间顺序返回来的近几天最新结果，再在该话题上对所有的用户计算话题权威值(表征该用户的话题权威性)，根据此话题权威值，再一次对搜索引擎返回的搜索结果进行重排序。

[0007] 本发明的具体实现步骤如下：

[0008] 一种用户话题权威性的计算方法，该方法包括以下步骤：

[0009] 步骤一、获取话题以及由微博搜索引擎返回的按照时间排序的结果集；

[0010] 步骤二、在步骤一得到的结果集中抽取所有用户id；

[0011] 步骤三、获取步骤二中每一个用户id一跳网络中的信息；

[0012] 步骤四、从步骤三得到的结果集中提取步骤一获取的话题上的所有微博及其对应的用户信息；

[0013] 步骤五、从步骤四得到的结果集中提取话题度量；

[0014] 步骤六、提出用户话题权威性的特征、特征计算公式并计算；

[0015] 步骤七、提出用户话题权威性即话题权威值计算公式并计算；

[0016] 步骤八、返回用户话题权威值计算结果。

[0017] 一种基于用户话题权威性的微博重排序方法，该方法包括以下步骤：

[0018] 步骤一、按照用户话题权威值从大到小顺序对用户排序；

[0019] 步骤二、根据用户的排名顺序对搜索引擎返回的按照时间顺序排列的微博进行重新排序；对于一个用户多条微博的情况，微博按照时间先后排序；

[0020] 步骤三、将重新排序的微博结果返回给用户。

[0021] 有益效果

[0022] 本发明针对微博搜索领域，综合考虑用户话题权威性以及传统的话题权威度量，提出用户话题权威值的计算方法，并使用得到的用户话题权威值对搜索引擎按照时间顺序返回的结果进行调整，其意义在于，该方法能够提高排序后返回结果的质量，从而证明用户话题权威性在微博排序中的有效性，增强用户体验。

附图说明

[0023] 图1为本发明实施例中微博用户权威值计算流程图；

[0024] 图2为本发明实施例中微博重排序的流程图；

[0025] 图3为本发明实施例中特征ID、GF1、MI、TS以及NS的密度函数图；

[0026] 图4为本发明实施例中特征ID、GF1、MI、TS以及NS为高斯分布的QQ图；

[0027] 图5为本发明实施例中特征TS以及NS分别为对数正态分布(Lognormal)以及混合高斯分布(GMM)时的QQ图；

[0028] 图6(a)(b)(c)分别为本发明实施例中，以“google”数据集上特征ID、TS以及NS的密度函数拟合图；

[0029] 图7(a)(b)分别为本发明实施例中，数据集“google”以及“healthcare”在由前5～1000条微博计算的NDCG值。

具体实施方式

[0030] 图1是本发明第一实施例的流程图。该用户话题权威值计算方法可应用于微博用户。需要注意的是，本发明所提出的方法仅针对热门话题，因此在抽取用户关键词之后，还需要有一步用于判断用户输入的关键词是否属于热门话题。

[0031] 具体地，首先获取到用户输入的搜索关键词，根据获取到的关键词来判定其是否属于热门话题。其中，判定热门话题的方法是，统计最近一段时间的与搜索关键词相关的关键词标签(hashtag)数量，并对其进行排序。排在前20位的我们即可认定其为热门话题行列。若该搜索关键字不被认定为热门话题，则没有证明本方法的适用性。若被认定为是热门话题，则按照本发明所提的方法对搜索引擎所返回的微博用户结果进行计算得到相应的用

户权威值。

[0032] 计算微博用户权威值的流程图见附图1，具体流程如下：

[0033] 首先，获取由微博搜索引擎返回的按照时间排序的结果集，在结果集中抽取每一位用户的id，在此基础上获取每一个用户一跳网络中的信息：包括该用户所有的粉丝和该用户所关注的所有的人的id，以及他们和该用户的所有微博。在新获取的三个数据集上，根据关键字匹配(即字符串匹配)提取前面提到的所有微博中的该话题上的所有微博，从而得到一个子数据集。在该子数据集上，提取出所涉及到的用户话题度量，话题度量详情参见表1。

[0034] 表1.用户话题权威性度量列表

[0035]

编号	度量
OT1	原创微博的数量
OT2	分享链接的数量
OT3	用户所有微博中的单词的重复度
OT4	hashtag 的数量
CT1	会话微博的数量
CT2	由该用户发起的会话微博的数量
RT1	转推微博的数量
RT2	原创微博中被其他用户转发的个数（不重复计算）
RT3	转发该用户的微博的所有用户的个数（不重复计算）
M1	该用户提及到其他用户的次数（相同用户可重复计算）
M2	该用户提及到其他用户的个数（相同用户不重复计算）
M3	其他用户提及到该用户的次数
M4	提及到该用户的其他用户的个数
G1	该用户在该话题上活跃的粉丝数
G2	该用户关注的人在该话题上活跃的个数
G3	在该用户之后发布该话题微博的粉丝数
G4	该用户关注的人中先于该用户发布该话题微博的数量
F1	该用户总的粉丝的数量
F2	该用户关注的人的总的数量

[0036] 其中，OT，CT，RT，M以及G分别表示原创微博、会话微博、转发微博、提及和与用户关系图相关的度量。表1中所列出的特征中涵盖了微博的形态学特性(如微博中嵌有超链接、

hashtag的数量)，及其被使用的方式特征(如转发、提及、会话、原创微博等)，还有就是表示用户话题兴趣的特征。另外，针对微博搜索领域，基于人们对名人的观点感兴趣这一点，我们加入F1及F2两个度量，用于表示用户总的受欢迎程度。

[0037] 其次，根据表1所提出的用户话题权威性度量，我们提出12个相应的度量用户话题权威性的特征，如表2所示。

[0038] 表2.每个用户的话题权威性特征

[0039]

特征ID	计算方法	描述
TS	$\frac{OT1+CT1+RT1}{\lvert tweets\rvert}$	话题参与强度
SS	$\frac{OT1}{OT1+RT1}$	原始话题强度
$\overline{CS}$	$\frac{OT1}{OT1+CT1}+\lambda\frac{CT1-CT2}{CT1+1}$	非会话话题强度
RI	$RT2*\log(RT3)$	转推影响力
MI	$M3*\log(M4)-M1*\log(M2)$	提及影响力
ID	$\log(G3+1)-\log(G4+1)$	信息传播度
NS	$\log(G1+1)-\log(G2+1)$	一跳网络得分
OT21	$\frac{OT2}{OT1}$	超链接在原创微博中所占的比例
OT41	$\frac{OT4}{OT1}$	关键词 hashtag 在原创微博中所占的比例
OT3	$\frac{2*\sum_{i=1}^{n}\sum_{j=1}^{i-1}S(s_i,s_j)}{(n-1)*n}$	作者微博所用词的相似度
GF1	$\frac{G1}{F1}$	所有粉丝中该话题上有微博的粉丝所占的比例
F12	$\frac{F1}{F2}$	粉丝强度

[0040] 其中，TS表示作者参与一个特定话题的程度，其计算公式中|tweets|表示该用户所有话题上的所有微博的数量，SS用来衡量作者微博的原创性程度，同时也衡量作者的话题性强度。另外，$\overline{cs}$ 用来衡量作者在多大程度上在这个话题上发表微博，以及作者在该话题上跑题到会话的程度。我们使用$\overline{cs}$ 这个特征，主要是用于区别网络用户中的个体与组织或机构，因为一般来讲，个人用户更容易倾向于进入会话，而组织或机构则不会。再者，$\overline{cs}$ 这个特征是用来对那些不是由用户发起，是用户处于礼貌性的初衷而进入的会话，做一个折损。直观上来讲，$\overline{CS}<\frac{OT1}{OT1+CT2}$ ，这样，根据此不等式，有 $\lambda<\frac{OT1}{OT1+CT2}*\frac{CT1+1}{OT1+CT1}$，我们就求解出λ。根据经验值，我们取λ满足90％的用户，其中λ用于表示用户倾向于进入微博会话的程度。

[0041] 接下来,RI特征把作者的微博被转发的次数以及转发作者微博用户的个数考虑在内,用于衡量作者微博内容的影响力。与RI相似,特征MI通过考虑被提及的次数来衡量用户在话题上的影响力。特征ID主要是用来衡量由该作者引起的在他一跳网络上所散发的微博传播的影响力。NS综合考虑了在该话题上活跃的粉丝数与其关注的人中在该话题上活跃的数量,旨在估计在该作者周围该话题的活跃程度。对于OT21、OT41,他们是用来计算超链接以及hashtag在作者原创微博中的出现的比率。OT3用于计算作者在其所有的n条(包括该话题上以及该话题外)微博中,所使用的单词的重复度,其中,对于两个单词的集合s_i,s_j,其相似度被定义为$S(s_i,s_j)=\frac{|s_i\cap s_j|}{|s_i|}$,其中,$s_i$,$s_j$是由作者的第i,第j条微博中通过去掉停用词以及做stem之后得到的单词的集合,且在计算特征OT3之前,所有微博先按照时间排序,即$times(s_i)<times(s_j):\forall i<j$。

[0042] 直观上来讲,对于一个特定的话题领域,关注用户的人在该话题上的比率越大,该用户在该话题上的影响力就越大。特征GF1就是由用户在该话题的粉丝在总的粉丝中的比例,由此从粉丝角度来衡量话题上的权威性。考虑到微博搜索这种应用场景中,人们往往喜欢关注名人的在某事情上的看法,我们加入非话题权威度量,由F12表示。

[0043] 最后,对于以上提出的12个特征,我们给出部分具有代表性的特征的概率密度分布图(见附图3),由于我们需要将其拟合成连续状态下的函数,通过观察其概率密度函数图像,以及给出Q-Q图验证(见附图4,附图5),附图4是假定所有的特征均为高斯分布时给出的Q-Q图,通过观察,很明显的可以发现,只有特征ID以及GF1符合高斯分布,其余的特征都不能使用高斯分布进行很好的拟合(因为他们的Q-Q图中,有太多的点远离直线y=x);附图5是假定特征TS、NS分别服从对数正态分布(Lognormal)以及混合高斯分布(GMM)时给出的Q-Q图。我们对其潜在的分布分为4个类别,对于每一个类别,拟合的方法相同。其中,第一个类别包括特征ID,GF1,用高斯分布来拟合;第二个类别包括特征TS,F12,用对数正态分布(Lognormal)进行拟合;第三个类别包括特征MI,RI,OT41,由于其数据分布过于集中,我们将其值划分为n个区间,并在此基础上求得对应的累积概率值;第四个类别包括特征NS,OT3,OT21,CS以及SS,由于没有现有的分布能够很好的进行拟合,我们选用基于无监督学习的高斯混合模型(Gaussian Mixture Model,即GMM)进行拟合。其部分拟合效果见附图6(a)、6(b)、6(c)。

[0044] 基于以上特征拟合,我们提出基于累积概率分布(CDF)的话题权威值计算方法(参见表3)。下面详细介绍话题权威值的计算步骤:

[0045] 我们使用基于累积概率分布来计算每一个用户在该话题上的权威值,即CDF_10或CDF_12方法。对于用户x_i,其话题权威值计算公式如下:

[0046] $$AS(x_i)=\prod_{f=1}^{m}F_f(x_i^f;\theta_f)$$

[0047] 其中,其中x_i表示第i个用户,x_i^f表示用户i在表2中第f个特征上的值(f取值范围为1-12),F_f表示参数为θ_f的第f个特征的累积概率分布函数在x_i^f处的CDF值,m表示所用到的特征的个数,即方法CDF_10用到表2中的前10个特征,同理,方法CDF_12用到表2中的前12

个特征。对于参数θ_f，其对于不同分布表示不同的参数，例如，对于高斯(Gaussian)分布和对数正态(Lognormal)分布，其代表(μ_f,σ_f)，对于由K个高斯组件组成的混合高斯分布(GMM)(在我们的实验中，所有的K＝2)，其代表(π_k,μ_k,Σ_k)，k∈[1,K]。对于每一个分布函数中的参数，我们采用极大似然估计方法得到。

[0048] 为了更好的逼近真实话题特征值，我们在以上话题权威值计算公式的基础上又提出了一种基于加权的计算公式，即CDF_weighted方法，其话题权威值计算公式如下：

[0049] $$AS(x_i)=[\prod_{f=1}^{11}F_f(x_i^f;\theta_f)]^{\beta}[F_{12}(x_i^{12};\theta_{12})]^{(1-\beta)}$$

[0050] 其中x_i表示第i个用户，x_i^f表示用户i在第f个特征上的值；β∈(0,1)，用于在话题特征以及非话题特征之间做平衡，我们通过最大化皮尔逊相关系数来求得每一个数据集中对应的最优β值。

[0051] 根据微博用户权威值对微博重排序的流程图见附图2，具体流程如下：

[0052] 首先根据前面计算出的用户话题权威值按照从大到小的顺序对用户排序；

[0053] 其次根据用户的排名顺序对搜索引擎返回的按照时间顺序排列的微博进行重新排序，对于一个用户多条微博的情况，微博之间按照时间先后排序；

[0054] 最后将重新排序的微博结果返回给用户。

[0055] 为了进一步证明以上所提方法的有效性，我们使用几种计算权威值的方法作为对比，具体参见表3。

[0056] 表3.权威值计算方法列表

[0057]

方法名称	描述
Conv_based	只基于传统特征 F12 计算权威值的方法
Gaus_10	基于表 2 中前 10 个特征服从高斯分布时 CDF 值的乘法计算用户权威值
SUM_12	基于表 2 中所有的特征使用加法计算用户权威值

[0058]

MUL_12	基于表 2 中所有的特征使用乘法计算用户权威值
CDF_10	基于表 2 的前 10 个特征的 CDF 值的乘法计算用户权威值
CDF_12	基于表 2 中所有的特征的 CDF 值的乘法计算用户权威值
CDF_weighted	方法 CDF_12 的加权形式

[0059] 评价指标：为了评价排序的效果，我们采用NDCG(Normalized Discounted Cumulative)作为评价指标。其计算方法如下：

[0060] $$NDCG_n = Z_n \sum_{i=1}^{n} \frac{2^{G_i} - 1}{\log_2(i+1)} \quad ;$$

[0061] 其中，n表示经过重排序后的前n条微博，G_i是重排序后的微博列表的第i条微博的得分，Z_n是归一化因子，它使得NDCG的理想值为1。

[0062] 微博评分方法：

[0063] 本评分分为3个等级，分别为3、2、1分，其中，3分为最高等级，2分次之，1分为最低等级。

[0064] 对每一条微博，评分准则如下：

[0065] 1).如果它包含的信息与查询该微博的关键字相关，且带有很好的信息量，则可评为3分。

[0066] 2).如果它包含的信息与查询该微博的关键字相关，且附带有部分的信息量，则可评为2分。

[0067] 3).如果它包含的信息与查询该微博的关键字相关，且基本上不包含相关信息量；或者它基本与查询该微博的关键字无关，则评为1分。

[0068] 其中，判断微博含有信息量的标准包括：是否含有超链接(URL)、关键字Hashtag，以及提供与该关键字相关的其他信息。另外，评分时还需要考虑微博的语言表达部分，比如表达的是否完整、单词缩写情况，以及微博用语是否文明等等。

[0069] 数据集：

[0070] 关于数据集，我们使用的是Twitter上2009年6月到10月份的数据。所有的微博加上用户关系文件大概有65.8G。我们从中选择5个热门话题作为关键词，分别是：google，healthcare，iran，music以及twitter。对于每一个关键词，我们收集大概6千条最新且字符串匹配效果最好的微博，该数据集的大致情况参见表4.

[0071]

关键词	google	healthcare	iran	music	twitter
微博数量	5371	2919	4162	5175	5208
用户数量	4221	1949	1953	4446	4651
用户粉丝数量	788149	600355	917983	834016	832140
用户话题粉丝数量	131281	34292	57197	143870	321804
用户朋友数量	550980	347651	388208	426138	604472
用户话题朋友数量	114565	30401	39763	121119	272095

[0072] 备注：用户粉丝即关注用户的人，用户朋友即用户关注的人。

[0073] 实验结果：

[0074] 接下来我们给出我们的实验结果，下面是由权威值计算方法CDF_12在各个数据集上计算出来的前10名话题权威值最高的作者名称列表：

[0075] 表5.各个数据集上的前10名作者列表

[0076]

google	healthcare	iran	music	twitter

[0077]

programmableweb	healthcareintl	iranhr	showhype	dehboss
paulkbiba	hcrepair	jricole	nytimesmusic	chito1029
omarkattan	hcdmagazine	newscomirancvrg	variety music	Louer_voiture
morevisibility	notmaxbaucus	jerusalemnews	im_musiclover	twithority
wormreport	Bnet_healthcare	jewishnews	digitalmusicnws	trueflashwear
followchromeos	healthnewsblogs	dailydish	musicfeeds	twedir
digg technews	vcbh	haaretzonline	wemissmjblog	jointhetrain
webguild	presidentnews	guneyazerbaycan	411music	robbmontgomery
junlabao	chinahealthcare	ltvx	radioriel	youtubeprofits
redhotnews	ilgop	reuterskl	jobsinhiphop	thepodcast

[0078] 对于表5中的数据，我们手动对其进行了检查，发现他们主要是由名人、受欢迎的博客作者等组成，而且，我们的算法能够发现那些专注于特定领域并且粉丝数目很少的人(在表中以黑体字表示)。

[0079] 进一步的，我们随机选择两个数据集，即google和healthcare，并给出他们分别使用表3中列出的权威值计算方法得到的试验结果，见附图7(a)、7(b)。从图中可以看出，我们所提出的基于加权的权威值计算方法(CDF_weighted)比其他所有的计算方法排序效果都要好；附图7(a)中可以看出，Conv_based方法的性能随着k的增加而总体上迅速下降，在附图7(b)中，Conv_based方法的性能也不如我们提出的基于CDF的方法。因此，这种现象进一步的证明了话题权威值在微博排序中所起的作用。从附图7(a)、7(b)中，还可以看出我们提出的CDF_10方法的性能比其高斯版本的方法(Gaus_10)性能要好的多，从而进一步的证明了我们所提出的精确的拟合方法比只是简单的假设特征服从高斯分布的方法具有更好的性能。更进一步地，基于累加和累乘的方法(SUM_12以及MUL_12)的性能不如我们所提出的基于累积概率分布(CDF_based)的方法。总的来说，我们提出的CDF_weighted方法相对于传统的用户权威值度量(Conv_based)以及基于高斯的方法，性能提升20％以上。

[0080] 因此，本发明所提出的话题权威值的计算方法以及基于此特征进行的微博搜索排序是非常有实际应用价值的。

[0081] 为了说明本发明的内容及实施方法，本说明书给出了一个具体实施例。在实施例中引入细节的目的不是限制权利要求书的范围，而是帮助理解本发明所述方法。本领域的技术人员应理解：在不脱离本发明及其所附权利要求的精神和范围内，对最佳实施例步骤的各种修改、变化或替换都是可能的。因此，本发明不应局限于最佳实施例及附图所公开的内容。

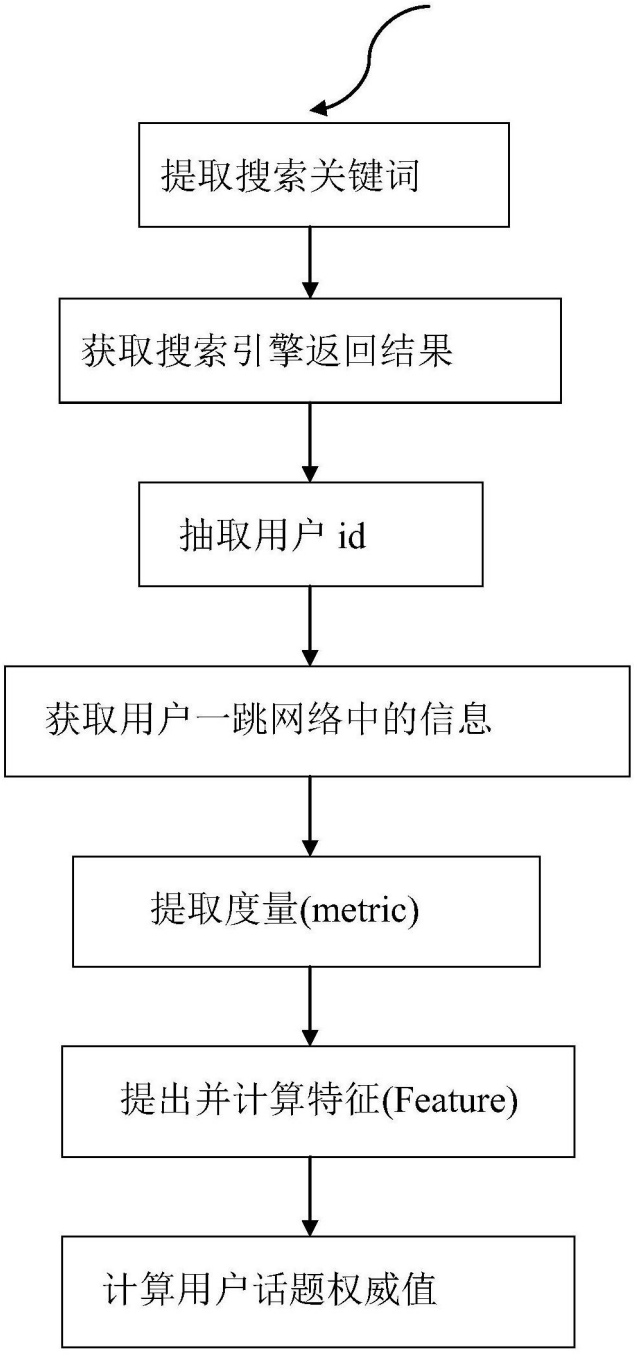

提取搜索关键词
获取搜索引擎返回结果
抽取用户 id
获取用户一跳网络中的信息
提取度量(metric)
提出并计算特征(Feature)
计算用户话题权威值

图1

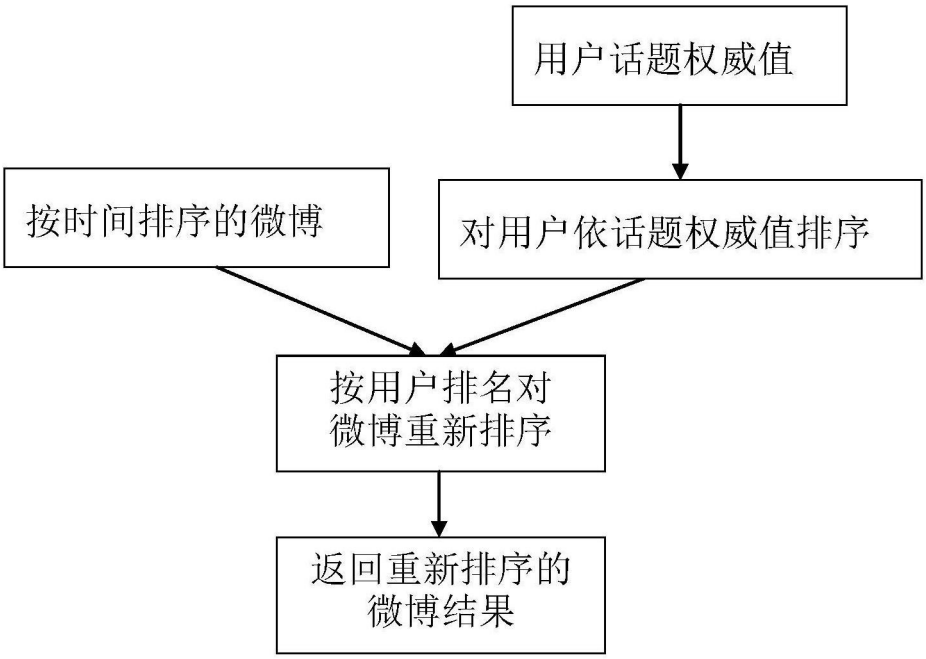


图2

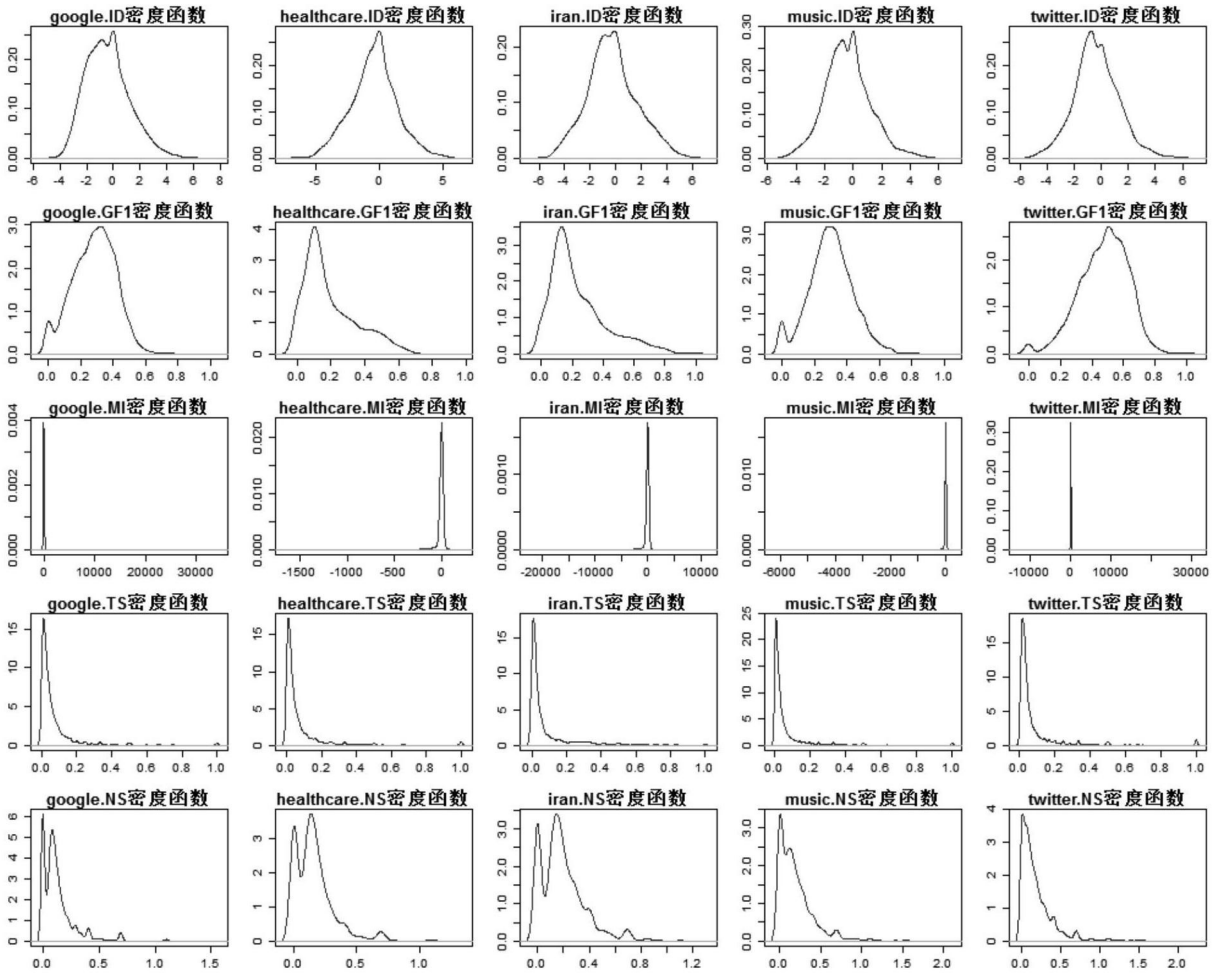

google.ID密度函数
healthcare.ID密度函数
iran.ID密度函数
music.ID密度函数
twitter.ID密度函数
google.GF1密度函数
healthcare.GF1密度函数
iran.GF1密度函数
music.GF1密度函数
twitter.GF1密度函数
google.MI密度函数
healthcare.MI密度函数
iran.MI密度函数
music.MI密度函数
twitter.MI密度函数
google.TS密度函数
healthcare.TS密度函数
iran.TS密度函数
music.TS密度函数
twitter.TS密度函数
google.NS密度函数
healthcare.NS密度函数
iran.NS密度函数
music.NS密度函数
twitter.NS密度函数

图3

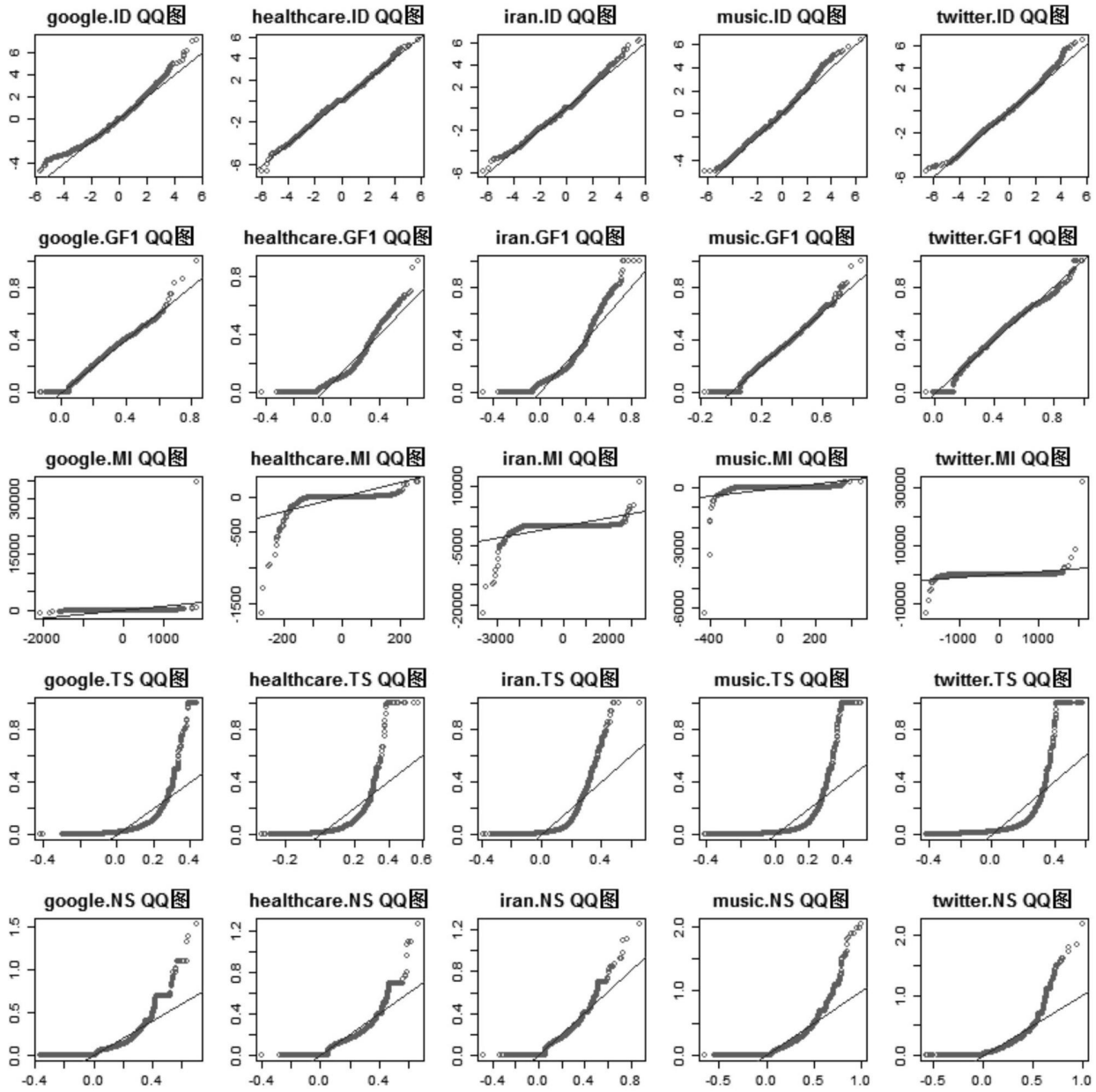

google.ID QQ图
healthcare.ID QQ图
iran.ID QQ图
music.ID QQ图
twitter.ID QQ图
google.GF1 QQ图
healthcare.GF1 QQ图
iran.GF1 QQ图
music.GF1 QQ图
twitter.GF1 QQ图
google.MI QQ图
healthcare.MI QQ图
iran.MI QQ图
music.MI QQ图
twitter.MI QQ图
google.TS QQ图
healthcare.TS QQ图
iran.TS QQ图
music.TS QQ图
twitter.TS QQ图
google.NS QQ图
healthcare.NS QQ图
iran.NS QQ图
music.NS QQ图
twitter.NS QQ图

图4

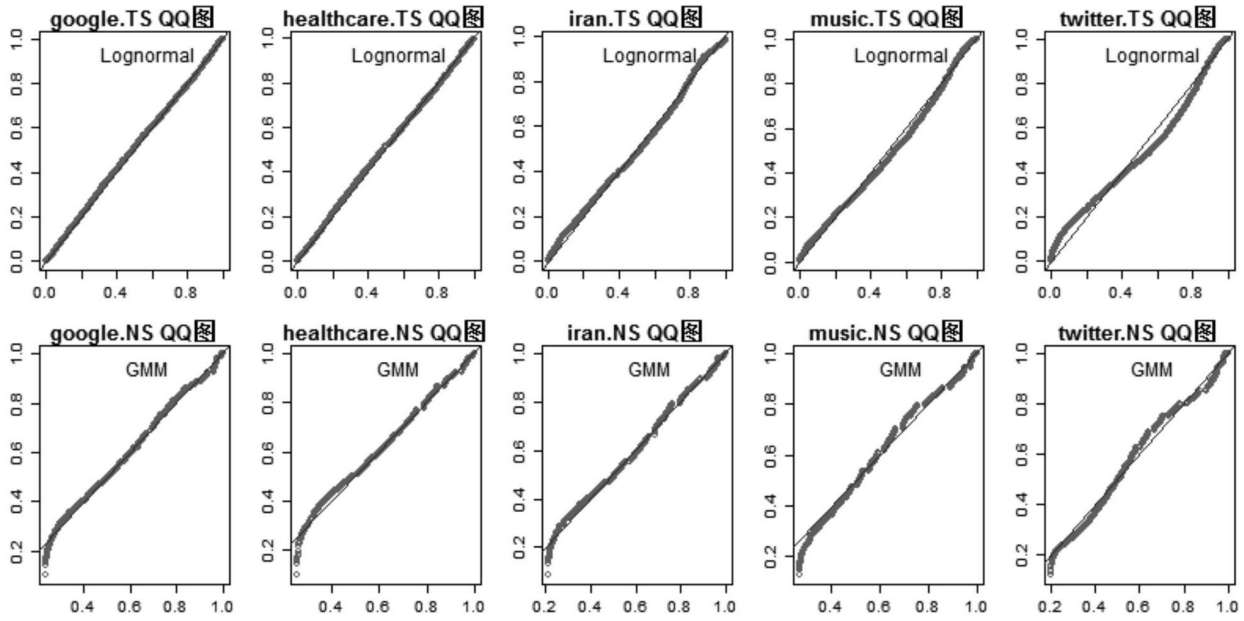
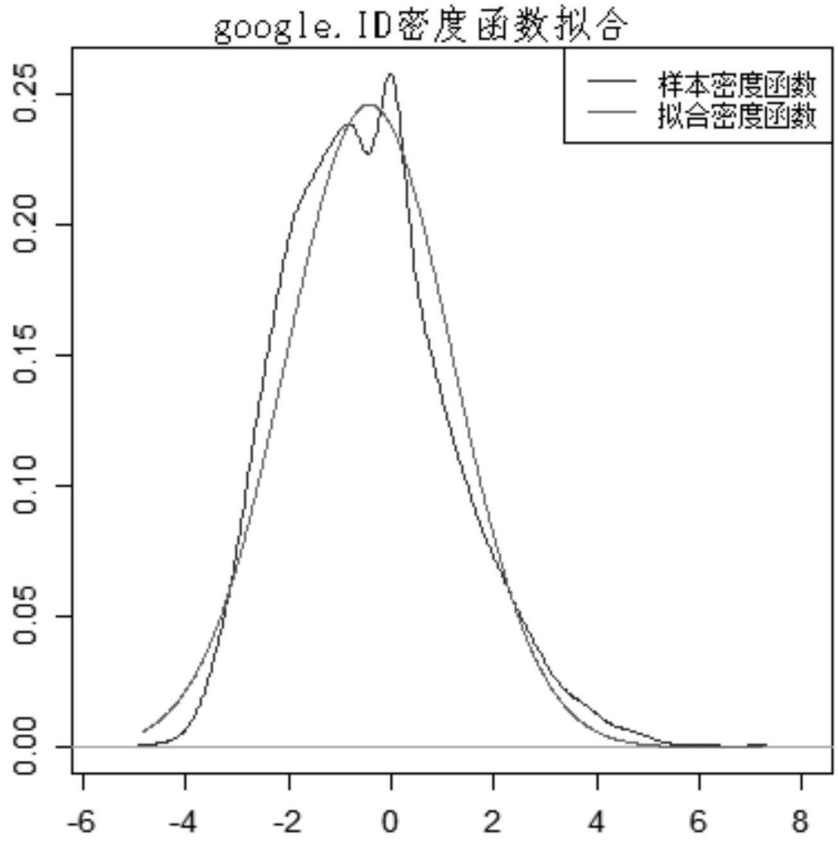
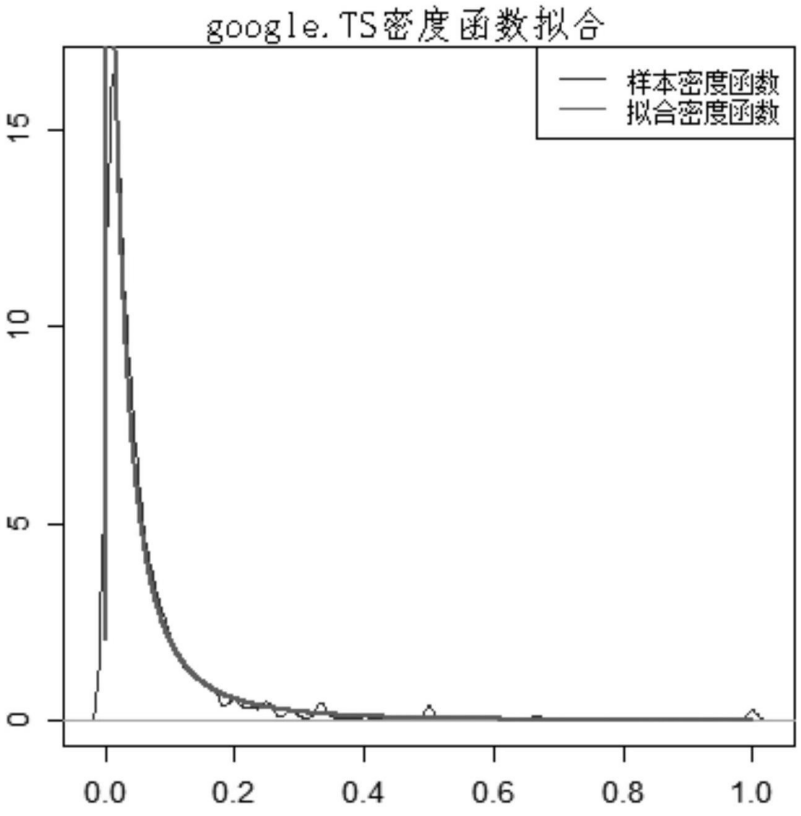

google.TS QQ图
Lognormal
healthcare.TS QQ图
Lognormal
iran.TS QQ图
Lognormal
music.TS QQ图
Lognormal
twitter.TS QQ图
Lognormal
google.NS QQ图
GMM
healthcare.NS QQ图
GMM
iran.NS QQ图
GMM
music.NS QQ图
GMM
twitter.NS QQ图
GMM

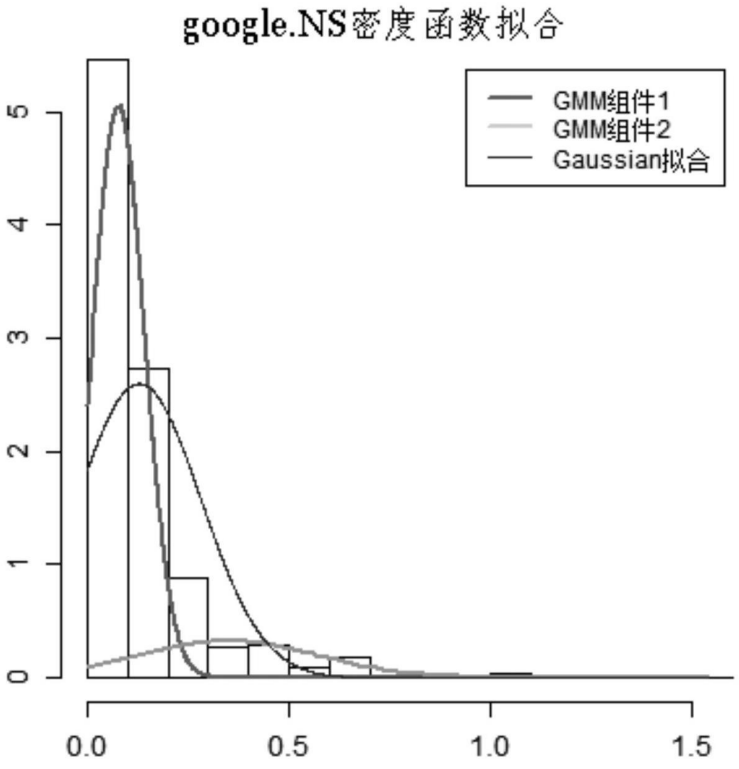
图5


google. ID密度函数拟合
样本密度函数
拟合密度函数
0.00
0.05
0.10
0.15
0.20
0.25
-6
-4
-2
0
2
4
6
8

（a）

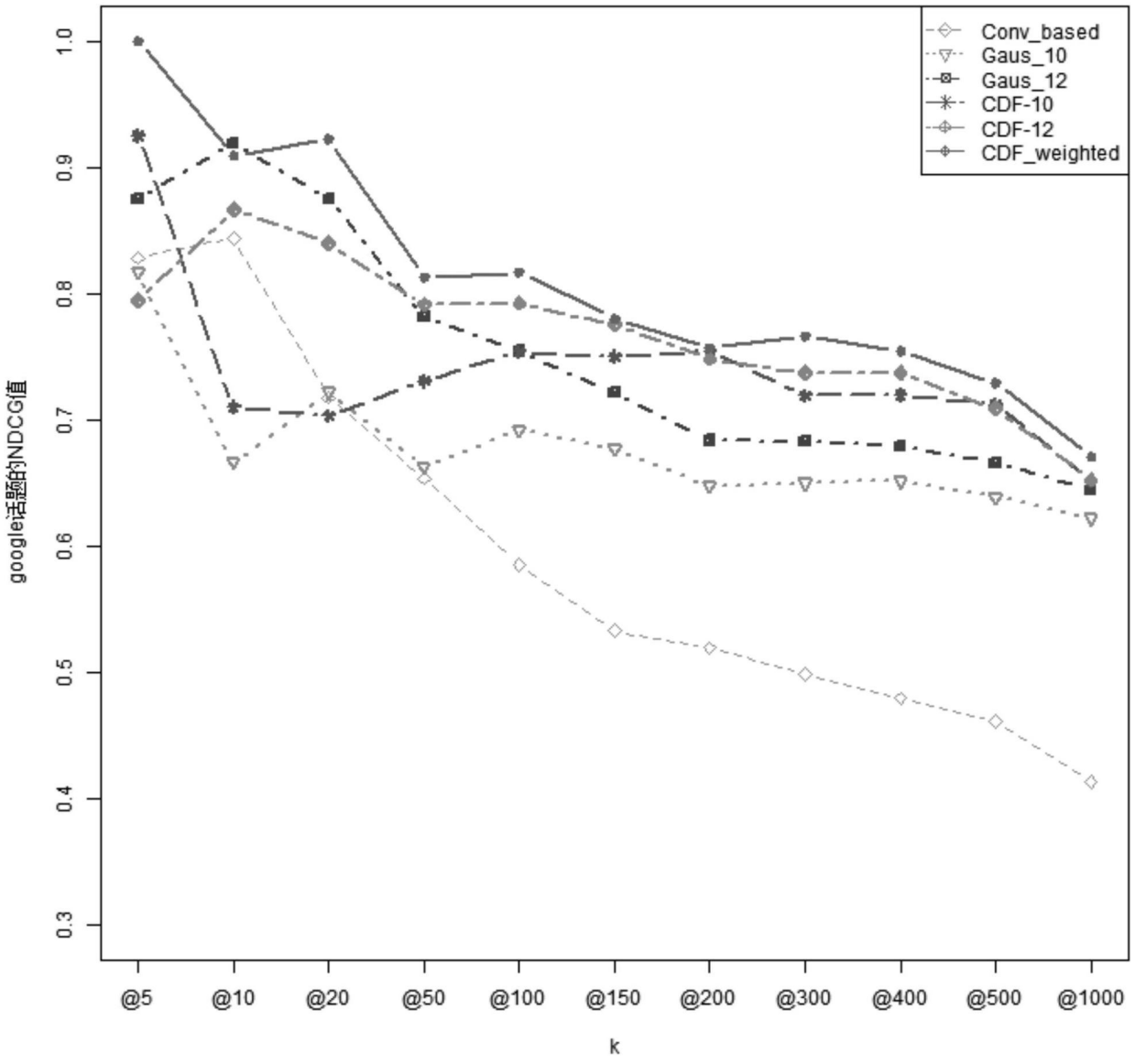

google.TS密度函数拟合
样本密度函数
拟合密度函数
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0

（b）

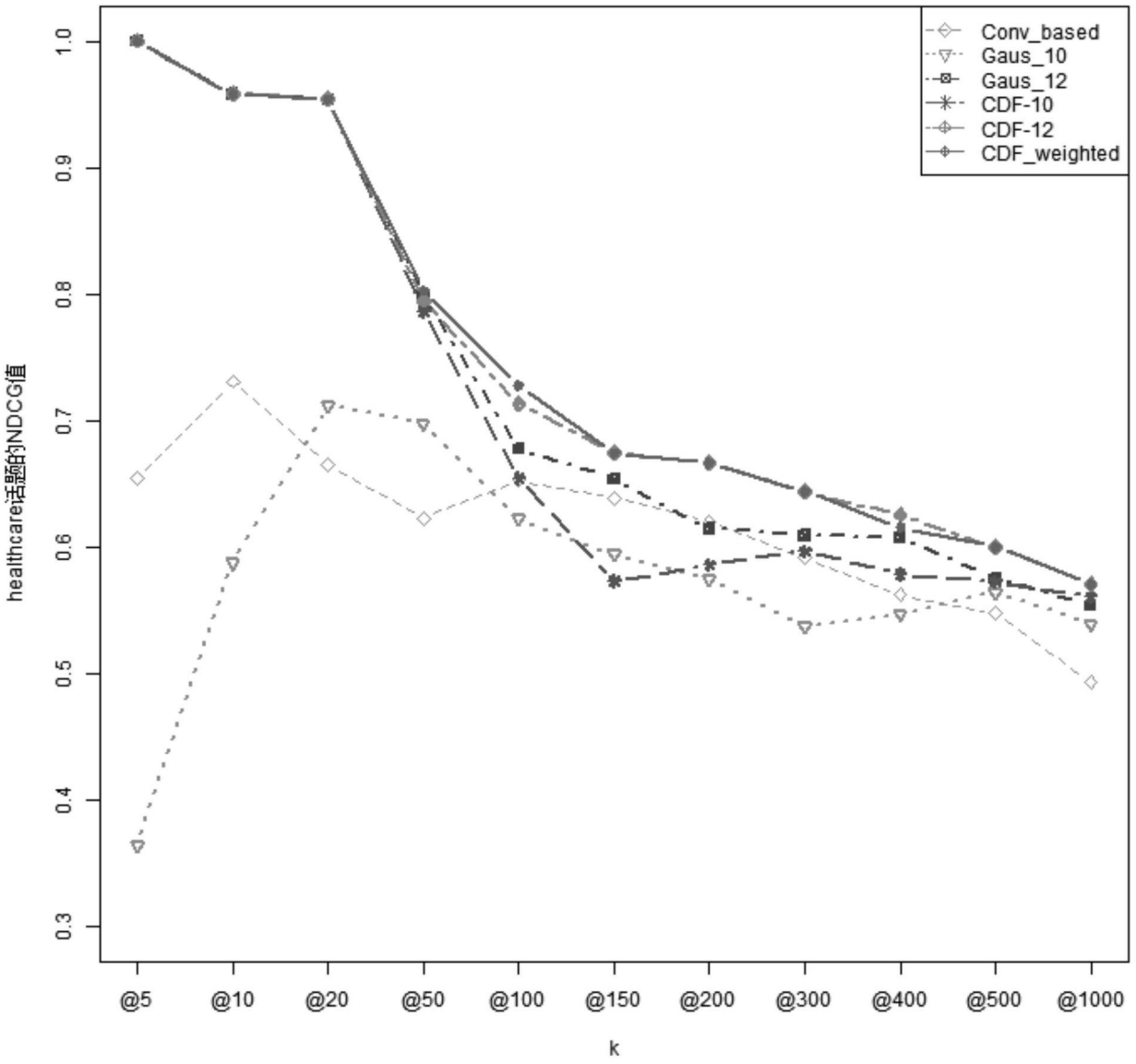

google.NS密度函数拟合
GMM组件1
GMM组件2
Gaussian拟合
0
1
2
3
4
5
0.0
0.5
1.0
1.5

（c）

图6


Conv_based
Gaus_10
Gaus_12
CDF-10
CDF-12
CDF_weighted
google话题的NDCG值
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
@5
@10
@20
@50
@100
@150
@200
@300
@400
@500
@1000
k

（a）


Conv_based
Gaus_10
Gaus_12
CDF-10
CDF-12
CDF_weighted
healthcare话题的NDCG值
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
@5
@10
@20
@50
@100
@150
@200
@300
@400
@500
@1000
k

（b）

图7