基于时间特性的微博热门话题检测算法研究*

闫光辉，赵红运，任亚缙，陈 勇
（兰州交通大学 电子与信息工程学院，兰州 730070）

摘 要：以用户兴趣理论和用户之间的关注行为为基础，结合时间因素在微博热门话题检测中的重要作用，研究了如何有效获取微博中最新、最有价值的话题问题，基于 PageRank 经典算法提出了一种带时间参数的热门话题检测算法（TimePageRank）。算法首先使用投票机制抽取出用户感兴趣的话题并记录话题的生成时间；然后用权值计算公式计算每个话题的权值；最后使用 TimePageRank 算法对这些话题进行排名，从而检测出微博中的热门话题。真实数据集上的实验结果验证了该方法的高效性。

关键词：微博；热门话题；时间因素；TimePageRank 算法；用户兴趣；PageRank 算法
中图分类号：TP301.6 **文献标志码**：A **文章编号**：1001-3695(2014)01-0043-04
doi:10.3969/j.issn.1001-3695.2014.01.009

Detection algorithm research for hot topics in micro-blog based on time characteristics

YAN Guang-hui, ZHAO Hong-yun, REN Ya-jin, CHEN Yong
(*College of Electronic & Information Engineering*, *Lanzhou Jiaotong University*, *Lanzhou* 730070, *China*)

Abstract: Combined with the important role of the time factor in the detection of hot topics, this paper studied how to effectively get the latest and the most valuable topic issues in the micro-blog based on the theory of user interest and the behavior between users, and proposed a hot topic detection algorithm (TimePageRank, which modified the PageRank algorithm) with a time argument. First, the algorithm extracted topics which were interesting to users by using the voting mechanism and recorded the generation time of the topic. Then, it calculated the weight of each topic. Finally, this paper used the proposed algorithm to rank these topics to detect hot topics in the micro-blog. The experimental results over real data set illustrate the effectiveness and efficiency provided by the algorithm.

Key words: micro-blog; hot topic; time factor; TimePageRank algorithm; user interest; PageRank algorithm

0 引言

微博[1]是一个新兴的在线沟通平台，允许用户发布简短的信息更新。微博用户可以关注其他人，也可以被其他人关注，并且可以从自己所关注的用户那里获取到这个用户所有的信息，还可以点击自己感兴趣的话题，浏览与话题相关的内容。如果用户需要更全面地了解与话题相关的事件过程以及与事件相关的一些描述，就需要大量地浏览话题下其他用户发表的信息，不过这种方式获取的信息极容易出现不完整，既花费了大量的精力又得不到理想的结果[2]。另外，用户往往对有价值的、较热门的话题感兴趣。那么，怎样才能检测出微博中这些热门的话题呢？

近年来，国内外学者在热门话题检测方面作了大量的研究。Song 等人[3]基于用户的兴趣和用户之间的关注行为来检测微博热门话题。Peng 等人[4]结合各种热门话题的相关参数，针对突发事件提出了基于用户喜好的热门话题检测方法。Wu 等人[5]基于近似文本检测的方法，利用查询扩展技术对新闻话题进行了追踪和重排名。目前也有部分学者从微博用户传播影响力的角度对微博热门话题检测的方法进行研究[6~9]。例如，Anagnostopoulos 等人[6]在对大量数据进行统计分析的基础上确定了社会影响是个人行为与社会关系相关性的一个重要来源；Yeung 等人[7]提出了一种用户采纳行为的概率模型，推断出在微博传播过程中一个用户对另一用户的影响力。另外还有部分学者针对特定领域进行热门话题的检测[10~12]。纵观国内外研究，大多数学者主要集中在用户兴趣和用户传播影响力对热门话题检测的影响，都没有考虑时间因素。本文在考虑用户兴趣基础上，增加了时间参数，使较早发布但又没有得到及时更新的话题获得一个相对较低的权值，而使那些发布时间比较晚、质量却很高的话题获得一个相对较高的权值，从而更有效地检索出最新的、最有价值的信息。从以下原因可以看出时间参数的重要性：

a）用户往往对最新发生的事情感兴趣。针对发布信息的简洁、方便和草根性，微博中大部分内容都在不断地变化。理想的情况下，那些过时的内容应该被及时删除，然而在实际的使用中却并非如此[13]。

b）现有的热门话题排名技术基本还是青睐于链接关系和

收稿日期：2013-04-24；**修回日期**：2013-05-30 **基金项目**：国家自然科学基金资助项目(61163010)；新世纪优秀人才支持计划资助项目(NCET-10-0017)；甘肃省陇原青年创新人才扶持计划资助项目(252003)；兰州市科技计划资助项目(2008-1-28)；甘肃省电力信息通信中心资助项目(KJ[2012]80)

作者简介：闫光辉(1970-)，男，河南商丘人，教授，博士，主要研究方向为数据挖掘、数据仓库等(yangh9805@qq.com)；赵红运(1988-)，女，江苏徐州人，硕士研究生，主要研究方向为数据挖掘；任亚缙(1987-)，女，山西吕梁人，硕士研究生，主要研究方向为数据挖掘；陈勇(1989-)，男，河南三门峡人，硕士研究生，主要研究方向为数据挖掘.

用户之间的关注关系。因此,较早发布的话题往往最受欢迎。因为经过时间的积累,就会得到很多关于这个话题的链接,进而得到较高的 PageRank 值,从而获得较前的排名。与此相反,新的高质量话题往往获得的排名比较低。

1 相关理论

1.1 用户兴趣研究

用户兴趣检测有利于对用户的行为进行分析,因此,用户发布的信息和用户之间的关注关系可以很好地反映微博中用户的兴趣。假如一个用户对某些话题感兴趣,他可能就会发布有关这些话题的帖子或者关注发布了这些内容的其他用户。所以,主要从两个方面来检测用户的兴趣,一个是用户自身帖子的内容,一个是他们所关注的那些用户发布的帖子内容。这个过程可以用一个图形来形象地表示,如图 1 所示。

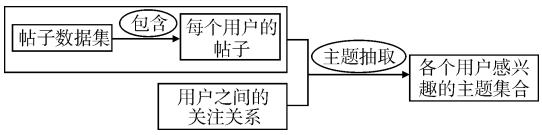

图 1 用户感兴趣的话题抽取图

1.2 时间维度

PageRank 和 HITS 算法是最知名的基于链接的排序算法,但是它们都没有考虑一个很重要的维度——时间维度。它们往往倾向于选择发布很早的网页,因为随着时间的积累有很多链接链向这些网页。但是过去有质量的网页在现在或将来可能就是没有价值的;新的网页可能是高质量的,但是由于指向它们的链接很少,所以排序结果往往靠后[14]。研究微博热门话题搜索也存在同样的问题。虽然目前的搜索引擎可能在它们的排序算法中考虑到了时间维度,但是它们并没有公开这方面的研究,所以公开地研究、解决排序的时间维度问题对于将来的搜索技术还是很重要的。

目前有部分学者在排序算法的研究中考虑到了时间特性。Li 等人[15]针对出版物搜索,研究了应用程序排序时的时间方面因素,并且基于 Markov 链的平稳概率分布提出了一个 TS-Rank 排序算法。但是针对微博热门话题检测的研究,还没有学者考虑到时间因素的影响。

1.3 PageRank 算法

PageRank[16]算法是由 Brin & Page 提出的,对一个排序向量进行预计算。这个排序向量对给定图形中的所有节点都提供了一个先验的权威估计[17]。该算法也适用于微博热门话题检测的研究。一个话题 A 的 PageRank(PR)值为

$$PR(A) = (1-d) + d \times \left(\frac{PR(p_1)}{C(p_1)} + \cdots + \frac{PR(p_n)}{C(p_n)} \right) \tag{1}$$

在这个等式里面,$PR(A)$ 代表话题 A 的 PageRank 值;$PR(p_i)$ 代表指向话题 A 的话题 p_i 的 PageRank 值;$C(p_i)$ 代表由话题 p_i 向外链出的数量;d 是一个阻尼因子,取值范围在 0 ~ 1 之间。本文仍然将阻尼因子的值设置为 0.85,这个数值在 Brin & Page 最初提出 PageRank 算法时被使用过,具有一定的可靠性。

2 基于用户兴趣和时间因素的热门话题检测算法

2.1 抽取用户感兴趣的话题

将话题信息用向量模型表示为 $HT_{U_m}(t_k) = (HW_1, HW_2, \cdots, HW_i; T_1, T_2, \cdots, T_i)$。在这个公式中,$HT_{U_m}(t_k)$ 表示用户 U_m 发表的帖子,HW_i 表示用户 U_m 发表的帖子 t_k 包含的话题,T_i 表示话题 HW_i 的生成时间。本文对文献[4]的投票方法进行了改进,增加了话题生成的时间,方便本文后半部分的研究。举一个简单的例子,如图 2 所示,用户 U_1 发布了不同的帖子 t_k,每一个帖子都包含了不同的主题。对每一个主题来说,包含它的帖子被看做是用户 U_1 对它的一次投票。本文通过投票数对这些主题进行排名,得到最高排名的主题就被认为是用户 U_1 感兴趣的主题。假如每个用户感兴趣的主题数量被限定为 3 个,那么 HW_2、HW_1、HW_3 就被认为是用户 U_1 感兴趣的话题。

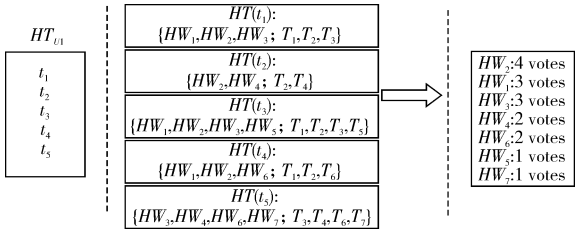

图 2 投票事例图

2.2 微博话题权重

微博用户可以根据自己的兴趣关注其他用户发布的某些话题,也可以自己发布有关这些话题的帖子。可以基于用户之间的关注关系,建立一个话题图形 $G(T,L)$。其中,T 是节点的集合,即话题的个数;L 是带方向的边的集合,即用户之间的关注行为。这样,给定两个话题 t_1 和 t_2,边 $\langle t_1, t_2 \rangle$ 只有在这两个话题相对应的用户之间存在关注行为的时候才存在。

因此,通过分析图 G 中的相连接性来度量每个话题的权值。每个话题的权值取决于链向它的那些话题的数量和权值。在这里,给定一个话题 $t_i \in T$,它的权值按式(2)计算。

$$\text{auth}(t_i) = d \times \sum_{t_j \in \text{linker}(t_i)} \frac{\text{auth}(t_j)}{|\text{linking}(t_j)|} + (1-d) \tag{2}$$

其中:$d \in (0,1)$,$\text{linker}(t_i)$ 是一个函数,返回值为链向话题 t_i 的话题集,$\text{linking}(t_j)$ 也是一个函数,返回值为话题 t_j 链出的数量。权值的计算采用迭代算法,其中,在最初时,每个话题的权值都被初始化为 $\text{auth}^0(t_i) = \frac{1}{|T|}$。在每一步,算法按下面公式重复计算权值:$\text{auth}^0(t_i) = d \times \sum_{t_j \in \text{linker}(t_i)} \frac{\text{auth}^{t-1}(t_j)}{|\text{linking}(t_j)|} + (1-d)$。当收敛条件满足时,该过程结束。

2.3 带时间参数的 PageRank 算法

本文基于文献[19]提出了带有时间参数的 PageRank 算法,即 TimePage-Rank 算法。本文注重一个话题当前的重要性,几个月之前的热门话题肯定比几年之前的热门话题重要。通过加权话题链接的产生时间,对 PageRank 算法进行了修改。系统按照式(3)计算每个话题链接基于时间权重的 PageRank(PRT)值。本文中,每个话题的 PageRank 值拟设为 1,按照如上迭代的方式进行计算,直到结果最终收敛。

$$PR^{\mathrm{T}}(A) = (1-d) + d \times \left(\frac{w_1 \times PR^{\mathrm{T}}(p_1) \times \text{auth}(p_1)}{C(p_1)} + \cdots + \frac{w_n \times PR^{\mathrm{T}}(p_n) \times \text{auth}(p_n)}{C(p_n)} \right) \tag{3}$$

在这个等式中,w_i 是每个话题链接基于时间的一个权重,它的值取决于话题 p_i 到话题 A 的链接时间,这也是话题 p_i 的

产生时间。话题间的链接产生得越早,这个权重就越小。由于在时间序列预测中经常使用指数的平均值,本研究规定权重随着时间呈指数递减,$w_i = \text{decayRate}^{(y-t_i)/12}$,考虑到时间跨度比较大的情况,将时间长度的单位由月转换成年,可以适当减小权重函数的计算值,便于 TimePageRank 算法的计算。其中 y 是当前的时间,t_i 是话题 p_i 的产生时间,$(y-t_i)$ 是时间差;decayRate 是一个参数。下面的例子使用 0.5 来说明这个概念。例如在本文的训练数据中,最新的话题产生于 2012 年 7 月份。2012 年 7 月和 2011 年 7 月出现的话题链接分别有一个权值 1 和 0.5。需要注意的是,如果 decayRate 为 1,那么时间加权的 PageRank 算法就和原始的 PageRank 算法一样。因此,这个参数可以根据数据集或话题的性质进行调整,当其值接近 1 时,权值随时间缓慢下降。它更适用于静态域或者是该域中新出现的话题。

本文注重对最近产生的链接进行加权,从话题的过去评估它的权重。同时本研究也对一个话题在将来的潜在重要度进行了探索。为了评估这一点,引入其他的参数——趋势因子。

就前面的例子来说,$PR^{\mathrm{T}}(A)$ 给出了截至 2012 年 7 月底话题 A 的重要性。在将来的时间,这个重要性会如何变化呢? 它将受截至 2012 年 7 月底话题链接变化的影响。因此,本文挖掘了一个话题 A 过去的一些特性来计算它的趋势因子 $\text{trend}(A)$:

a)数据预处理。过滤掉那些链接低于每月一次的话题。因为随着时间的推移,它们不可能持续。由于同样的原因,指定最小的趋势因子。

为了作出可靠的预测,使用每月链接数据的移动平均线抹平噪声。在特定的月份,一个话题的移动平均链接通过在那个月和前一个月它的链接取平均值计算。

b)计算截至 2012 年 7 月底链接变化的趋势因子。对一个话题 A 来说,假如 2012 年 3、4 月它的链接数是 nt,2012 年 5、6 月份它的链接数是 nf,那么,话题 A 的趋势比 $R(A) = nf/nt$。

如果一个话题的产生年龄小于 3 个月,就没有足够的数据来计算它的趋势比。

在对所有的话题计算 $R(A)$ 后,对它们进行标准化,因此标准化后的值处于最小趋势因子和 1 之间。最小趋势因子被设置为 0.5,因为对任何一个话题来说,每一个先前链接的权值通过一年的时间将会减少一半,如式(3)。$R(A)$ 的标准化值是话题 A 的趋势因子 $\text{trend}(A)$。话题 A 最后的 TimePageRank(TPR)是:

$$\text{TPR}(A) = \text{trend}(A) \times PR^{\mathrm{T}}(A) \tag{4}$$

其中:$PR^{\mathrm{T}}(A)$ 由式(3)计算可得。

3　实验结果与分析

3.1　实验数据

腾讯微博是国内比较著名的微博,它向人们提供了一种有趣和互动的方式来发现和讨论信息。用户使用微博可以容易地与朋友或其他人在线分享个人情感和意见。文献[18]采用爬虫技术爬取了 Twitter 数据集,本文也按照该文的方法用爬虫抓取了腾讯微博从 2011 年 7 月 1 号至 2012 年 7 月 1 号由 6 480个用户发布的 71 834 个帖子以及这些用户之间的关注联系,在此基础上,采用 IKAnalyzer 3.0 中文分词工具包[2] 进行分词,去掉停用词,过滤掉单词个数少于 5 个的消息后得到微博文本共 40 905 条。

3.2　评价标准

本文采用文献[5]中的评价方法,将基于词频的检测和提取方法及基于用户兴趣的检测和提取方法作为参照,利用精确度(accuracy)、召回率(recall rate)和 F-度量(f-measure)对热门话题检测的结果进行评价。具体的计算公式如下:

$$\text{accuracy} = \frac{\text{num}_i}{\text{num}_i + \text{num}_j}$$

$$\text{recall} = \frac{\text{num}_i}{\text{num}_i + \text{num}_k}$$

$$\text{F-measure} = \frac{2 \times \text{accuracy} \times \text{recall}}{\text{accuracy} + \text{recall}}$$

其中:num_i 表示抽取到的能够反映帖子主题和用户兴趣的热门话题数量;num_j 表示抽取到的不能反映帖子主题和用户兴趣的热门话题数量;num_k 表示没有被抽取到但却能反映帖子的主题和用户兴趣的热门话题数量。

3.3　实验结果

1)帖子内容分析

为了证明所提出方法的有效性,针对前文提到的数据集,使用本文 3.1 节提出的方法抽取了 10 条用户感兴趣的话题,然后使用 TPR 对这些话题进行排序,结果如表 1 所示。表中第一列表示抽取出的 10 个用户感兴趣的话题;第二列给出了这些话题的名称;第三列表示每个话题的生成时间;第四列给出了使用基于词频的检测方法得到的每个话题的排名;第五列给出了使用基于用户兴趣的检测方法得到的每个话题的排名;第六列给出了使用 TPR 得到的每个话题的排名。

表 1　抽取用户感兴趣的话题并排序

rank	topic	time	基于词频的检测方法	基于用户兴趣的检测方法	TPR
1	2012 奥运	2005-07	19	1	1
2	屌丝	2012-03	39	8	5
3	微公益	2010-12	716	6	3
4	中国好声音	2012-07	46	2	2
5	钓鱼岛是中国的	2010-09	323	3	6
6	最右	2011-04	614	7	5
7	舌尖上的中国	2012-05	576	9	7
8	那些年一起追的女孩	2012-02	620	20	9
9	三亚宰客	2012-01	78	12	8
10	神九升空	2012-06	69	13	11

通过表 1 可以得出,基于用户兴趣的检测方法在微博热门话题检测方面比基于词频的方法要优越很多,因为基于词频的方法主要是计算关键词的频度,并没有考虑到用户的兴趣,很多恶意的操作可以趁机人为地刷高所谓高质量话题的频度。本文提出的 TPR 算法的性能又比基于用户兴趣的检测方法要好,因为该算法考虑到了话题的生成时间,对那些发布时间早、一定时间内没有被关注的话题分配较低的权值,提高新生成话题的权重,从而有效地检索出微博中的热门话题。因此,该结果说明了 TPR 算法的高效性。

2)精确度、召回率和 F-度量对比

针对收集到的 40 905 条微博信息,本文将 TimePageRank 算法与其他传统的检测和提取算法(基于词频的检测和提取算法以及基于用户兴趣的检测和提取算法)进行比较。从上文获得的文本中提取热门话题,发现 TimePageRank 算法的准确性高于其他两种算法,初步的结果如表 2 所示。

表 2 TimePageRank 算法与其他算法的对比结果

比较项	TPR/%	基于词频的检测算法/%	基于用户兴趣的检测算法/%
accuracy	83.4	75.2	82.3
recall	77.6	66.7	75.9
F-measure	80.3	70.4	79.8

初步结果表明,TimePageRank 可以更有效地从微博中检测和提取热门话题,并削弱旧话题的权重,将较新的、有价值的话题推荐给用户。进一步证明了 TimePageRank 的检测和提取效率是较高的。

3)灵敏性分析

本文在引入 TimePageRank 概念的时候指出,对于一个给定的数据集,可以调节 decayRate 以达到最佳的结果。为了验证 TPR(A)的评分有效性,本文使用一定范围内的 decayRate 值对 TPR(A)进行测试,进而研究得分有效性和 decayRate 的关系。一组 decayRate 为{0.1,0.2,0.4,0.5,0.6,0.7,0.9,1.0},其结果如表 3 所示。

表 3 DecayRate 不同取值的影响

1	取值	No. of top topics		
		5	10	15
2	decayRate = 0.1	4387	5763	6679
3		75%	78%	81%
4	decayRate = 0.2	4417	5892	6798
5		76%	79%	82%
6	decayRate = 0.5	4442	5837	6862
7		75%	78%	81%
8	decayRate = 0.7	4312	5731	6881
9		77%	78%	82%
10	decayRate = 0.9	4289	5643	6281
11		76%	69%	68%
12	decayRate = 1.0	3609	4731	5351
13		59%	63%	64%
14	best tweet counts	5467	7251	8230

表 3 中第 1 行列出了排名较高的话题组;第 2 行给出了当 decayRate = 0.1 时,TPR 预测方法的结果;第 3 行给出了这种条件下总的话题提及量和理想排名的总提及量(第 14 行)。第 4-5、6-7、8-9、10-11、12-13 行和第 2-3 行含义相同,唯一不同的是,在这些实验中,decayRate 的值在 0.1 ~ 0.9 范围内变化。

结果表明,0.2 ~ 0.7 是 decayRate 最佳的取值范围。当 decayRate 低于 0.2 时,该系统在很大程度上侧重于最近讨论的话题;最近较少讨论的话题即便它们是重要的也不会出现在预测结果的前几名,这就降低了整体排名的质量。与此相反,当 decayRate 接近 1.0 时,系统不区分时差讨论。这样,受关注时间长、没有被更新的旧话题受青睐;新的话题在高质量排名中就被排除出靠前的名次。

4 结束语

本文研究了时间维度对热门话题搜索结果的影响,借鉴基于用户兴趣和用户之间关注行为的思想,分析了微博话题的生成时间、话题间的跟帖链接对最终热门话题检索结果的帮助,进而提出了一些方法来提高搜索的效率。实验结果表明,考虑话题生成时间后,检测出的微博话题是高度有效的。但是本文也存在不足之处,对于大文本数据集来说,该方法在抽取用户感兴趣话题过程中的时间复杂度会很高,如果能解决这部分的问题,检索的质量会进一步提高。

参考文献:

[1] KWAK H, LEE C H, PARK H, *et al*. What is Twitter, a social network or a news media? [C]//Proc of the 19th International World Wide Web Conference. 2010:591-600.

[2] 邱云飞, 程亮. 微博突发话题检测方法研究[J]. 计算机工程, 2012,38(9):288-290.

[3] SONG Shuang-yong, LI Qiu-dan, ZHENG Xiao-long. Detecting popular topics in micro-blogging based on a user interest-based model [C]//Proc of International Joint Conference on Neural Networks. 2012:1-8.

[4] PENG Fei-fei, QIAN Xu, LI Gao-ren. A research of hot topic detection through microblogging[C]//Proc of the 4th International Conference on Intelligent Human-Machine Systems and Cybernetics. 2012: 185-188.

[5] WU Xiao-meng, IDE I, SATOH S. News topic tracking and re-ranking with query expansion based on near-duplicate detection[C]//Proc of the 10th Pacific Rim Conference on Multimedia. 2009:755-766.

[6] ANAGNOSTOPOULOS A, KUMAR R, MAHDIAN M. Influence and correlation in social networks[C]//Proc of the 14th ACM International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2008:7-15.

[7] YEUNG C M A, IWATA T. Capturing implicit user influence in online social sharing[C]//Proc of the 21st ACM Conference on Hypertext and Hypermedia. New York: ACM Press, 2010:245-254.

[8] GOYAL A, BONCHI F, LAKSHMANAN L V S. Learning influence probabilities in social networks[C]//Proc of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2010:241-250.

[9] CRANDALL D, COSLEY D, HUTTENLOCHER D, *et al*. Feedback effects between similarity and social influence in online communities [C]//Proc of the 14th ACM International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2008:160-168.

[10] 李劲, 张华, 吴浩雄, 等. 基于特定领域的中文微博热点话题挖掘系统 BTopicMiner[J]. 计算机应用, 2012,32(8):2346-2349.

[11] ZHU Ming-liang, HU Wei-ming, WU Ou. Topic detection for discussion threads with domain knowledge[C]//Proc of International Conference on Web Intelligence and Intelligent Agent Technology. New York: ACM Press, 2010:545-548.

[12] ISHIKAWA S, ARAKAWA Y, TAGASHIRA S, *et al*. Hot topic detection in local areas using Twitter and Wikipedia[C]//Proc of ARCS Workshops. 2012.

[13] BENEVENUTO F, MAGNO G, RODRIGUES T, *et al*. Detecting spammers on Twitter[C]//Proc of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference. 2010.

[14] WANG Can-hui, ZHANG Min, RU Li-yun. *et al*. Automatic online news topic ranking using media focus and user attention based on aging theory[C]//Proc of the 17th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2008:1033-1042.

[15] LI Xin, LIU Bing, YU P. Time sensitive ranking with application to publication search[C]//Proc of the 8th IEEE International Conference on Data Mining. 2008:187-209.

[16] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine[C]//Proc of the 7th International Conference on World Wide Web. 1988:107-117.

[17] HAVELIWALA T H. Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search[J]. IEEE Trans on Knowledge and Data Engineering, 2003,15(4):784-796.

[18] WENG Jian-shu, LIM E P, JIANG Jing, *et al*. TwitterRank: finding topic-sensitive influential Twitterers[C]//Proc of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM Press,2010:261-270.

[19] YU P S, LI Xin, LIU Bing. Adding the temporal dimension to search: a case study in publication search[C]//Proc of IEEE/WIC/ACM International Conference on Web Intelligence. 2005.