一种 Web 2.0 环境下互联网热点挖掘算法

李东方　俞能海　尹华罡
(中国科学技术大学电子工程与信息科学系多媒体计算与通信教育部—微软重点实验室　合肥　230027)

摘　要：利用 Web 2.0 下用户丰富的反馈信息进行互联网热点挖掘具有重要的应用价值。该文将 Web 2.0 下用户在互联网上的信息活动看作为热度活动，并利用热量传递模型对其建模，然后基于该模型提出适用于 Web 2.0 环境下的话题抽取与热度评价算法。实验结果表明热量传递算法有效地利用了用户反馈信息，适用于 Web 2.0 下互联网环境。

关键词：互联网；热点话题发现；话题排序；Web 2.0；热度扩散模型
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2010)05-1141-05
DOI: 10.3724/SP.J.1146.2009.00641

Mining Hot Topics on Internet under Web 2.0

Li Dong-fang　Yu Neng-hai　Yin Hua-gang
(*MOE-MS key Laboratory of Multimedia Computing and Communication*, *Department of Electronic Engineering and Information Science*,*University of Science and Technology of China*, *Hefei* 230027, *China*)

Abstract: It is a valuable task of mining hot topics on Internet utilizing user's feedback under Web 2.0 environment. Motivated by heat diffusion phenomena, the information activities on the Internet are treated as heat flow, and heat diffusion model is used to simulate these acitivities. Based on heat diffusion model, a novel topic detection and ranking algorithm is proposed. The experiment results demonstrate that the algorithm works well under Web 2.0 environment.

Key words: Internet; Hot topic detection; Topic rank; Web 2.0; Heat diffsusion model

1　引言

互联网目前已经进入 Web 2.0 时代，用户是互联网的主角，每时每刻互联网上都有数十亿条的信息被用户发布、获取、评价与传播。如何在这动态而分散信息活动中挖掘其中的热点话题成为互联网的研究热点。传统的互联网热点话题挖掘方法主要针对新闻数据进行挖掘。文献[1]中算法引入了时间衰减因子，对随着时间变化的新闻数据流和新闻来源网站进行排序，保证排序符合新闻的时效性。文献[2]提出了一种半自动化的热点挖掘算法，通过计算新闻事件在一段时间内的频率分布及所持续的时间单元，对新闻事件进行排序。以上两种算法没有考虑新闻的特殊属性如新闻出现的位置信息等对新闻重要性的影响。文献[3,4]建立了 3 部图模型将新闻网站与新闻和新闻事件联系起来，并利用了新闻的位置信息区分新闻的权重。但这种算法没有考虑到新闻在时间和内容类别分布。文献[5]详细地分析了热点的概念，并通过计算新闻在时间与类别上的分布，提取具有代表性的词语，计算热度。这种算法仅仅对单一的新闻数据集进行计算，没有考虑到 Web 2.0 环境下新闻的网站及事件的关联性。

以上这些算法在 Web 2.0 的环境下不能很好地对热点信息进行挖掘，因为这些算法本质上是从信息发布者(“创造者”)的角度来衡量信息的重要性，而没有考虑到用户(“消费者”)对互联网上新闻信息的需求。文献[6]分析对比了用户对新闻关注信息(User Attention)，并对比新闻焦点(Media Focus)，发现二者并不是完全相吻合，即互联网用户(“消费者”)并不完全认可媒体发布网站(“创造者”)对新闻的排序。文献[7]通过构建用户浏览网页的路径图并结合用户在每一个网页上的浏览时间，引入了用户的反馈信息，来对网页的重要性进行排序，效果要比传统的排序方法效果要好。

通过以上分析可知，用户作为 Web 2.0 下信息活动的重要参与者在信息重要性评价中起着非常重要的作用。进一步分析，更多的用户的创造与反馈信息数据存在于 Blog，Forum 等社区环境中，综合这些信息能更好地挖掘网络热点。针对这个特点，本文提出了一种 Web 2.0 下互联网热点挖掘算法，

2009-04-30 收到，2009-09-25 改回
国家自然科学基金(60672056)和国家 863 计划项目(2008AA01Z117)资助课题
通信作者：李东方　dfli3@mail.ustc.edu.cn

将信息活动看作发生在互联网上的热量活动，用户的信息活动则表示为热量的传递，而互联网则是承载热量活动的系统，信息的热度则可以测量其在互联网中热量的分布来进行评价。

2 Web 2.0 下信息热度模型

本文将互联网看作一个开放的通信网络，对互联网系统进行建模；并将信息活动看作是热量活动，使用热量传递模型进行热度评价。

2.1 互联网系统模型

互联网本质上是一个开放的信息系统，是用户进行信息活动的一个平台。从信息传播的角度来看，在互联网这个虚拟的空间中，用户、信息和互联网络是互联网的基本构成。从关联角度来讲，互联网上用户与信息及信息与信息间通过潜在的话题存在内在的关联，我们将互联网建模为一幅 3 部图 G。

$$G=(V,E) \tag{1}$$

其中顶点 $V=U\cup M\cup O$ 由用户空间 U (User Space)、信息空间 M (Media Space)和隐含的话题空间 O (Topic Space) 3 部分构成。边集 $E=E_{MO}\cup E_{MS}$ 包含信息空间与话题空间的边集 E_{MO} 和信息发布者 U_S 与信息间的边集 E_{MS}。

用户空间被分为两个部分，$U=U_S\cup U_C$，其中 U_S 代表信息的发布者，U_C 表示接收的信息的用户(信息的消费者)，实际中无法准确表示互联网上每一个用户，但可以通过收集如页面访问记录、网站流量或者搜索引擎记录等反馈信息来估计信息的传播性质，对信息的热度进行度量，并将这种度量信息在 G 上扩散，进而挖掘互联网上的热点，这与 DiffusionRank [8−10] 中利用模拟热量扩散来研究信息的 Rank 有相同的出发点。

2.2 互联网热度评价模型

热量传播是自然界现象，本文将信息活动看作热度活动，用户为热源，采用热度模型 [8−10] 来对互联网信息活动进行建模(图 1)，来综合评价互联网中信息和网站的热度。

首先定义 $O=\{o_1,o_2,\cdots,o_L\}$ 为互联网的话题集合，$M=\{m_1,m_2,\cdots,m_N\}$ 为互联网上的信息集合，$S=\{s_1,s_2,\cdots,s_K\}$ 为信息来源集合，$\boldsymbol{w}^o,\boldsymbol{w}^m,\boldsymbol{w}^s$ 分别表示三者的热度。

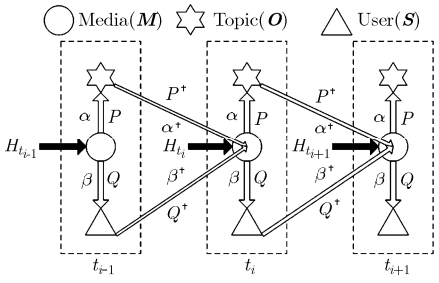

图 1 热量传递模型

定义 M 与 O 的关系矩阵为 $\boldsymbol{P}$：

$$p_{ij}=\text{probability}(o_j\mid m_i) \tag{2}$$

定义 O 与 M 的关系矩阵为 $\boldsymbol{P}^{\dagger}$：

$$p_{ij}^{\dagger}=\text{probability}(m_j\mid o_i) \tag{3}$$

定义 M 与 S 的关系矩阵为 $\boldsymbol{Q}$：

$$q_{ij}=\text{probability}(s_j\mid m_i) \tag{4}$$

定义 S 与 M 的关系矩阵为 $\boldsymbol{Q}^{\dagger}$：

$$q_{ij}^{\dagger}=\text{probability}(m_j\mid s_i) \tag{5}$$

定义用户反馈信息为一个“热量”流，$\boldsymbol{H}=\{\boldsymbol{H}_{t0},\boldsymbol{H}_{t1},\cdots,\boldsymbol{H}_t\}$，其中 $\boldsymbol{H}_{ti}$ 表示 t_i 时刻系统新进入的用户反馈信息。

定义 $\boldsymbol{H}_{ti}$ 为

$$\boldsymbol{H}_{ti}=H(M,t_i)=\{h_{ti}(m_1),h_{ti}(m_1),\cdots,h_{ti}(m_N)\} \tag{6}$$

然后定义热量传播模型。基本的热量传播模型为

$$\left.\begin{aligned}&\frac{\partial f(x,t)}{\partial t}-\Delta f(x,t)=0\\&f(x,0)=f_0(x)\end{aligned}\right\} \tag{7}$$

其中 $f(x,t)$ 是时刻 t 在位置 x 时的温度。本文将互联网建模为一个图，图中各个部分的点 M，S 和 O 建立起概率联系，并将互联网上的热量在图上随着时间进行传递(图 1)。

假设在时间点 t_i 上，互联网用户信息活动产生的热量为 $\boldsymbol{H}_{ti}$。我们有如下假设：(1) $\boldsymbol{H}_{ti}$ 与 $\Delta t=t_i-t_{i-1}$ 的长度是成正比的；(2)互联网上信息空间是无限的，互联网不会“过热”任意时刻的信息热度值收敛；(3)图中任意两点间 m_i 与 s_j 或者 m_i 与 o_k 的热量传输与两点间的权重成正比；(4)如果两点间没有联系，那么传输的热量为 0；(5)同一时段内，M，S 和 O 热量传导时间可以忽略，三者是热平衡的。

Δt 的时间内，图中任意一点的获得的热量是其输出到邻居点热量和从邻居点输入的热量之差。

$$\begin{aligned}w_i^m(t+\Delta t)-w_i^m(t)=&h_{t+\Delta t}^i-\gamma w_i^m(t)\Delta t\\&+\gamma\alpha^{\dagger}\alpha(\boldsymbol{P}^{\dagger})_i^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{w}^m(t)\Delta t+\gamma\beta^{\dagger}\beta(\boldsymbol{Q}^{\dagger})_i^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{w}^m(t)\Delta t\end{aligned} \tag{8}$$

求解式(8)可得

$$\boldsymbol{w}^m(1)=e^{\gamma\boldsymbol{R}}\boldsymbol{w}^m(0)+\boldsymbol{H} \tag{9}$$

其中

$$\begin{aligned}&\boldsymbol{R}=\boldsymbol{A}+\boldsymbol{B}-\boldsymbol{I},\\&\boldsymbol{A}=\alpha^{\dagger}\alpha(\boldsymbol{P}^{\dagger})^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}},\\&\boldsymbol{B}=\beta^{\dagger}\beta(\boldsymbol{Q}^{\dagger})^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\end{aligned} \tag{10}$$

$\boldsymbol{H}$ 为热量产生的平均速率；γ 为热量扩散因子；$0<\alpha,\beta<1$，$\alpha+\beta=1$，表示热量在 S 和 O 分配的比例，可以看做媒质内在属性和用户属性对热量传递影响的权重；$0<\alpha^{\dagger},\beta^{\dagger}<1$ 是热量传递因子，类似于热导因子，用以模拟时间对信息热量的影响。定义 $\alpha^{\dagger}$，$\beta^{\dagger}$ 为作用在 m 上热量传递算子。

$$\alpha^{\dagger}(\bullet,t)=e^{(-(t-t_0-hl_{\alpha})\bullet u(t-t_0-hl_{\alpha})/hl_{\alpha})} \tag{11}$$

$$\beta^{\dagger}(\bullet,t)=e^{(-(t-t_0-hl_{\beta})\bullet u(t-t_0-hl_{\beta})/hl_{\beta})} \tag{12}$$

其中 t_0 为媒质的出现时刻，hl_{α}，hl_{β} 为对应话题和源的参数，当 $x>0$ 时 $u(x)=1$，其他情况为一个很小的数 $\xi>0$。

实际中，由于图的规模比较大，直接计算式(9)运算量太大，本文采用下面近似方法来计算式(9)：

$$\boldsymbol{w}^m(1)=\left(\boldsymbol{I}+\frac{\gamma}{N}\boldsymbol{R}\right)^N\boldsymbol{w}^m(0)+\boldsymbol{H} \tag{13}$$

在系统建模的假设中，系统是一个热度均衡的系统，即在任意时刻，系统中节点的热度是收敛的。

收敛性证明：

令

$$\boldsymbol{\Gamma}=\left(\boldsymbol{I}+\frac{\gamma}{N}\boldsymbol{R}\right)^N \tag{14}$$

$$\lim_{i\to\infty,\forall k>0}\left|\boldsymbol{w}^m(i+k)-\boldsymbol{w}^m(i)\right|$$

$$=\lim_{i\to\infty}\|\boldsymbol{\Gamma}\|^i\left\|\sum_{j=0}^{k-1}\boldsymbol{\Gamma}^j\cdot\boldsymbol{H}+\left(\boldsymbol{\Gamma}^k-\boldsymbol{I}\right)\boldsymbol{w}^m(0)\right\|$$

易知：当 $\|\boldsymbol{\Gamma}\|<1$ 即 $\left\|\boldsymbol{I}+\frac{\gamma}{N}\boldsymbol{R}\right\|<1$ 时，系统收敛。

根据 Perron–Frobenius 定理可知，矩阵中每个元素为非负且每一行元素和 Δ_i 小于 1 时，矩阵最大特征值小于 1，即行列式小于 1，易知 $\boldsymbol{I}+(\gamma/N)\boldsymbol{R}$ 满足此条件。N 在合适的范围内系统能很好地近似理想的传递效果[8]，考虑到模型随时间的步进的特点，实验中我们采用范围(2~4)。

3　热点挖掘算法

通过以上对互联网信息热度进行建模，本文基于热度传播模型提出了互联网热点挖掘算法，并将该模型用于在线环境中。

3.1 话题抽取算法

在图 1 的模型中，如何表示话题并进行准确的抽取会影响到算法整体的效果。本文将互联网上的话题定义为一个多元组 $o=<z_1,\cdots,z_k>$，z 是包含文档中特定语义的词语，每个文档可以包含为一个或者多个话题。本文提出了一种基于话题语义的短语抽取与选取方法。

首先，定义文档中的词语权重为

$$\text{weight}(w)=\frac{\text{tf}\cdot\text{idf}}{\text{pEntropy}(w)}\text{Eng}(w,t) \tag{15}$$

$$\text{pEntropy}(w)=-\sum_{c\in C}p(c\mid w)\log_2 p(c\mid w) \tag{16}$$

$$\text{Eng}(w,t)=[1+\text{Var}(w,t)]\cdot\chi^2(w,t) \tag{17}$$

其中 tf 和 idf 为关于词频的统计量；pEntropy(w) 代表词语 w 在信息类别 C 上的分布的熵；Eng(w,t) 是根据文献[5]中 Aging Theory 和词语突发特性综合计算得到的权重因子，用来衡量词语的突发特性。

然后，从这些关键词(Candidate Set T)筛选出与文档语义相关的词语集 R。利用文献[11]中对标注质量的评价方法来筛选出能覆盖文档语义的标注词语。定义词语 w 对文档 d 的标注质量为 $S(w,d)$，$D(w)$ 为词语 w 出现的文档集合，$p(w_i\mid w_j)$ 为所有文档中 w_j 出现的情况下 w_i 出现的条件概率；$p(w_i\mid w_j,d)$ 为给定文档 d 和 w_j 后 w_i 的概率；计算公式为

$$p(w_i\mid w_j)=\frac{D(w_i)\cap D(w_j)}{D(w_j)} \tag{18}$$

$$p(w_i\mid w_j,d)=\sum_{k\neq i,j;w_k\in d}p(w_i\mid w_k)p(w_k\mid w_j) \tag{19}$$

$$S(w,d)=S(w,d)-\sum_{t\in T-R}p(w\mid t_i)\cdot S(t_i\mid d)+\sum_{t\in T-R}p(w\mid t_i,d)\cdot S(t_i\mid d) \tag{20}$$

式(20)从全局的角度来对剩余词语去除冗余信息，从文档局部关联角度来增强词语的关联信息，从而选取一定数量的词语作为文档语义覆盖的关键词集 R。最后，根据词语在文档中句子的共发频率和相似性来聚类 R 成独立语义的话题集合 $o(d)$。

定义式(2)为

$$p_{ij}=\sum_{t_k\in o_j}S(t_k,m_i)\Big/\sum_{t_k\in R}S(t_k,m_i) \tag{21}$$

表示话题对文档的语义覆盖程度。假设每个文档都是独立出现的，则定义式(3)为

$$p_{ij}^{\dagger}=\log_2(p_{ji})\Big/\sum_i\log_2(p_{ji}) \tag{22}$$

实验中，我们建立词语与话题的对应的倒排表进行快速查询，并利用 KeyGraph[12]来扩展词组，进而发现相关的话题。

3.2 热度评价算法

用图 1 中互联网热量传递模型来对互联网中信息和用户反馈进行建模。定义式(4)为

$$q_{ij}=1/\left|S(m_i)\right| \tag{23}$$

其中 $\left|S(m_i)\right|$ 为包含 m_i 的源的数目。定义式(5)为

$$q_{ij}^{\dagger}=w(m_j)/\sum_{m_k\in M(s_i)}w(m_k) \tag{24}$$

其中 $M(s_i)$ 为源 s_i 包含的信息，$w(m_j)$ 表示信息的对应的热度。

基于图 1 和以上分析和定义，将热度评价算法

描述如下：

参数： hl_α， hl_β， α， β， γ

输入：数据集 M 及相应的站点 S

输出：话题集合 O，与对应的 $\boldsymbol{w}^o, \boldsymbol{w}^m, \boldsymbol{w}^s$；

(1)对于每一个时间段 t_i，输入新的信息 ΔM 和更新信息 $\boldsymbol{H}_{ti}$；

(2)根据文献[11]与式(20)中的标注算法，提取话题；

(3)更新 M、S 和 O，$\boldsymbol{P}$、$\boldsymbol{P}^\dagger$、$\boldsymbol{Q}$ 与 $\boldsymbol{Q}^\dagger$；

(4)按照式(13)更新 $\boldsymbol{w}^o, \mathrm{w}^m, \boldsymbol{w}^s$；

可得到 t_i 时段对应的话题、媒质与话题源的热度。

实际中，每条信息都有一定的时效性，一段时间后，这些信息的热度会衰减为 0，可以将这些“过时”的信息对热度计算的影响忽略不计，进而可以减少算法的计算量。在实验中，从时间和热度两个方面设定了简单的门限，将过时的信息从更新中滤去。

4 算法评测与分析

从互联网上抓取了新闻、博客与论坛 3 类数据，分析并获取用户对这些信息做的反馈如回复数、点击数等，量化输入系统，进而来评价话题的热度，挖掘热点话题。对算法的有效性和话题挖掘的结果进行评测和分析。

4.1 实验设置

数据集包括新闻、博客和论坛 3 类数据，发布日期从 2009 年 2 月 11 日到 2009 年 4 月 9 日。这些数据来自于国内著名的站点如人民网、搜狐博客和天涯社区等。首先对实验数据进行预处理，包括去除错误信息、新闻数据去重和过滤低质量的论坛和博客数据等。最后，拥有如下数据集(表 1)。

表 1 数据集统计

	新闻	论坛	博客
总量(条)	23.4 万	18.7 万	12.1 万
最多(每天)	6237	4128	2453
最少(每天)	1812	2955	1830

算法中的参数包括 hl_α， hl_β， α， β， γ，根据文献[6,13]中的研究表明，对于绝大部分网页来讲，80%以上的用户关注度(点击)会在网页产生的 36 个小时内，设定 hl_α 为 36 h。类似地，根据人们记忆特性，将源对信息的影响 hl_β 设定为 48 h。α 和 β 表示信息内容和源对信息热度影响的比例，这是一个经验式的设定，设定 $\alpha = 0.73$ 和 $\beta = 0.27$。参数 γ 代表信息重要性的扩散程度，根据文献[8,10]中的分析，本文采用与 PageRank 算法中相同的配置 $\gamma = 0.85$。

4.2 实验结果与分析

首先对部分热点挖掘结果进行分析。表 2 是 2009 年 2 月 11 日到 2009 年 4 月 9 日时间段内，按照话题热度值所达到的历史最高点选取的 Top 10 的热点话题。分析这些话题可知，他们的共性是在新闻、博客和论坛等媒体信息中同时出现，其中有典型的突发事件等，也有网络热点议题。

表 2 Top 10 热点话题

1–5	圆明园 兽首	躲猫猫	法国 西藏	陕西 丹凤	胡锦涛 金融
6–10	小沈阳	南海 南沙	姚明 火箭	俄罗斯 新星	中国 航母

然后，对话题评价质量进行分析。本文以文献[5]中算法为对比，采用文献[4]中人为评测的方法对本文算法进行评价。另外还参照了新浪的热闻回顾[14]信息作为客观对比。新浪的热闻回顾可以按照用户的点击和回复数目等反馈信息对新闻进行排序，可以用来与本文的算法结果进行客观的对比。我们使用文中的算法选取了 6 周的时间内每周的热点话题中的 Top 200，对比文献[5]中算法选出的 Top 200(代表信息发布者角度)，以及新浪的热闻回顾[14]中在对应时间段热点新闻 Top 20(按点击排行，代表信息消费者的角度)。请 7 个人来对这些话题和新闻就其是否热进行评分。对每条信息的评分进行相加平均后，计算 NDCG@10[15]，得出最后结果(图 2)：

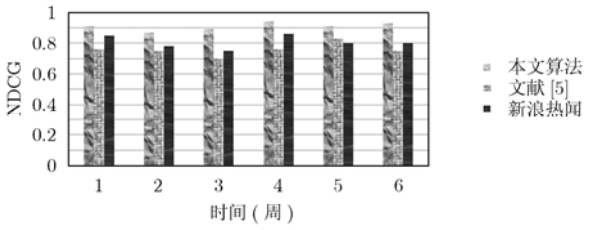

图 2 热度评价质量对比

通过对比可知，本文基于热量传递模型的热点发现算法比文献[5]仅仅依靠新闻特征性能好且稳定。对比图还说明媒体关注的焦点(文献[5]中的算法)与用户关注的焦点(新浪热闻)是有差别的，这与文献[6]中的研究是一致的。只有当媒体与用户感兴趣的话题共同出现时(如“躲猫猫”、“圆明园兽首”等)，文献[5]与新浪的热闻回顾[14]和本文算法效果才相接近。

实验结果说明了用户反馈信息对热点发现的重要性。在大部分情况下，基于用户反馈的新浪热闻回顾与本文算法的结果较为相近。实验结果还说明用户反馈有时不如文献[5]中完全依赖新闻的内容特征效果好。这是因为当没有明显的共同话题时，由于用户兴趣的多样性，用户反馈就会分散到互联网的话题空间中。这种情况下，新闻媒体的同一性就会取得优势，因而完全依赖新闻内容的文献[5]会取得较好的效果。

除此之外，本文算法的稳定性和良好性能也说明了综合利用新闻内容属性、媒体关注等信息有助于对评价信息热度。

5　总结与展望

本文针对 Web 2.0 下互联网信息活动中用户高度参与的特点，建立了基于热量传播的热点评价模型，并给出了互联网热度评价算法。实验结果表明算法能够综合利用用户反馈和网页等信息来准确的评价信息热度。未来我们还将进一步研究网络用户在虚拟社区的中的活动及其对热度评价的影响，另外将采用如文献[16]的策略方法改进热量传播算法。

参 考 文 献

[1] Del Corso G M and Gulli A, *et al.*. Ranking a stream of news[C]. Proceedings of the 14th international conference on World Wide Web, Chiba, Japan, May 10–14, 2005: 97–106.

[2] He T T, Qu G Z, and Li S W, *et al.*. Semi-automatic hot event detection[C]. Proceedings of the 2nd International Conference on Advanced Data Mining and Applications 2006, LNAI 4093: 1008–1016.

[3] Yao J Y, Wang J, and Li Z W, *et al.*. Ranking web news via homepage visual layout and cross-site voting[C]. Proceedings of the 28th annual European Conference on Information Retrieval, 2006, LNCS 3936: 131–142.

[4] Hu Y, Li M J, and Li Z W, *et al.*. Discovering authoritative news sources and top news stories[C]. Proceedings of 3rd Asia Information Retrieval Symposium, 2006, LNCS 4182: 230–243.

[5] Chen K Y, Luesukprasert Luesak, and Chou Seng-cho T. Hot topic extraction based on timeline analysis and multidimensional sentence modeling[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(8): 1016–1025.

[6] Wang C H, Zhang M, and Ru L Y, *et al.*. Automatic online news topic ranking using media focus and user attention based on aging theory[C]. Proceeding of the 17th ACM conference on Information and knowledge management, Napa Valley, California, USA, October 26–30, 2008: 1033–1042.

[7] Liu Y T, Gao B, and Liu T Y, *et al.*. BrowseRank: Letting Web users vote for page importance[C]. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, July 20–24, 2008: 451–458.

[8] Ma H, Yang H X, and King Irwin, *et al.*. Learning latent semantic relations from clickthrough data for query suggestion[C]. Proceeding of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, USA, October 26–30, 2008: 709–718.

[9] Yang H X, King Irwin, and Lyu Michael R. DiffusionRank: A possible penicillin for Web spamming[C]. Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23–27, 2007: 431–438.

[10] Song X D, Chi Y, and Hino Koji, *et al.*. Information flow modeling based on diffusion rate for prediction and ranking[C]. Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, May 8–12, 2007: 191–200.

[11] Xu Z C, Fu Y, and Mao J C, *et al.*. Towards the semantic Web: collaborative tag suggestions[C]. Proceedings of Collaborative Web Tagging Workshop at 15th International World Wide Web Conference, Edinburgh, UK, May 22–26, 2006. http://www.semanticmetadata.net/hosted/taggingws-www2006-files/13.pdf

[12] Mori Masaki, MIURA Takao, and Shioya Isamu. Topic detection and tracking for news web pages[C]. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA, 2006: 338–342.

[13] Wang Y, Liu Y, and Zhang M, *et al.*. Identify Temporal websites based on user behavior analysis[C]. Proceedings of 3rd International Joint Conference on Natural Language Processing, Hyderabad, India, 2008: 173–180.

[14] 新浪网热闻回顾，http://news.sina.com.cn/hotnews/. Hotnews on sina.com.cn, http://news.sina.com.cn/hotnews/.

[15] Jarvelin K and Kekalainen J. Cumulated gain-based evaluation of IR techniques[J]. *ACM Transactions on Information Systems*, 2002, 20(4): 422–446.

[16] Jiang Y C. Concurrent collective strategy diffusion of multiagents: the spatial model and case study[J]. *IEEE Transactions on Systems, Man and Cybernetics-Part C*, 2009, 39(4): 448–458.

李东方：　男，1985 年生，博士生，研究方向为大规模信息检索、数据挖掘.

俞能海：　男，1964 年生，教授，博士生导师，研究方向为信息检索、信息安全、模式识别等.

尹华罡：　男，1985 年生，博士生，研究方向为个性化信息检索.