基于主题的自适应、在线网络热点发现方法及新闻推荐系统

吴永辉[1,2],王晓龙[1,2],丁宇新[2],徐 军[2],郭鸿志[2]
(1.哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001;
2.哈尔滨工业大学深圳研究生院,深圳市网络环境智能计算重点实验室,广东深圳 518055)

摘 要: 本文提出了一种基于改进 HotRank 算法的站点排序及种子 URL 选择方法,建立了在线主题发现系统信息采集自适应增量更新模型;结合 LDA 模型和仿射传播聚类算法(AP),提出了一种网络主题发现和热点新闻推荐方法,并在海天园知识服务平台热点新闻推荐系统中得到了应用.

关键词: 知识服务; 主题发现; 增量; 自适应; LDA 模型; 仿射传播聚类
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2010) 11-2620-05

Adaptive On-Line Web Topic Detection Method for Web News Recommendation System

WU Yong-hui[1,2], WANG Xiao-long[1,2], DING Yu-xin[2], XU Jun[2], GUO Hong-zhi[2]
(1. *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*;
2. *Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong* 518055, *China*)

Abstract: We put forward a new URL choosing method and an incremental refreshing model based on HotRank algorithm. Using this method, the adaptive incremental refreshing model for hot news recommendation system is developed. A new topic detection and topic based hot news recommendation algorithm is proposed combining LDA and Affinity Propagation(AP). This research supports the development of hot news recommendation system , which is an important part of Haitianyuan knowledge service platform.

Key words: knowledge service; topic detection; incremental; adaptive; LDA; affinity propagation

1 引言

互联网的飞速发展使信息检索服务成为网络上使用频率最高的服务之一.快速变化的信息使传统搜索引擎越来越难以满足用户更加专业和个性化的需求.依托于国家 863 计划"基于 NLP 的智能搜索引擎"中对信息检索相关技术的研究,结合自然语言处理方法和数据挖掘方法,我们提出了知识服务的概念,构建了海天园知识服务平台,利用精细划分的领域知识提供更加专业和个性化的服务.本文主要探讨网络知识服务系统的热点发现问题.网络热点推荐系统的性能主要由两个部分决定:网络信息采集的实时性、有效性和热点主题、热点新闻发现的准确性.

对于信息采集器的实现部分,已经有了非常成熟的研究[1~3].增量更新技术[4~7]是抓取新网页一种有效方法.增量更新策略中评价网页质量的一种有效方法是基于链接分析的 PageRank[8]方法,在此方法基础上,产生了多种改进策略[9].

网络热点话题发现算法是影响新闻推荐系统性能的另一个重要因素.有关主题发现的研究,主要源自由美国国防高级研究计划局(DARPA)支持发起的"话题检测于跟踪"(TDT)任务[10].同时,主题分析技术也被用于其他自然语言处理的任务[11].在最近的研究中,LDA 模型[12~14]作为有效的语义分析概率模型,成为主题检测问题研究热点之一.本文把仿射传播聚类算法(AP)[15,16]和 LDA 模型相结合,提出了一种自适应网络主题发现和热点新闻推荐方法.实验表明,本文提出的方法,具有更好的主题区分能力.

2 基于 HotRank 的自适应增量更新模型

2.1 传统增量更新模型存在的问题

传统增量更新模型在热点发现的应用中有以下缺

收稿日期:2009-02-23;修回日期:2010-08-31
基金项目:国家 863 高技术研究发展计划(No.2006AA01Z197);国家自然科学基金(No.60703015,No.60973076)

点:(1)缺少对更新的网页内容的考虑;(2)容易被某些动态 URL 和主题无关的垃圾网页欺骗;(3)除了网页链接和网页内容的因素,网页更新时间对热点发现更为重要.

本文针对上述问题进行了改进,提出了一种更有效的 *HotRank* 算法.

2.2 增量更新模型建立

定义增量更新模型中,O_t 表示 t 时刻网页集中过期网页的数目,F_t 表示 t 时刻爬虫可以抓取的网页数目;U_t 表示 t 时刻产生的新网页数目;B 为网络带宽.网页集 P 中网页 p 的重要程度由以下公式决定:

$$\underset{p\in S_i}{PW(p)} = rank(p) + freq(p) + SW(Si) \tag{1}$$

其中,$rank(p)$为网页 p 由链接分析和锚文本分析得出的权重,$freq(p)$为网页变化频率,$SW(S_i)$为网页 p 所属的网站 S_i 的领域信息权重.则增量更新模型可以看作如下的优化问题:

$$\left.\begin{array}{l} MIN\{\sum_{t=1}^{T}(O_t + U_t)\} \\ MAX\{\sum_{p\in S_i} PW(p)\} \\ \text{带宽和抓取速度限制} \end{array}\right\} \tag{2}$$

2.3 种子站点选择

网站 S_i 的时效性权威值 HotRank(S_i)可以描述为:

$$HotRank(S_i) = \lambda \times SR(S_i) + (1-\lambda) \times Freq(S_i) \tag{3}$$

其中 $SR(S_i)$为网站 S_i 为对 S_i 网页的链接和锚文本信息使用改进的网页转移概率后得到的网站 SiteRank 值.$Freq(S_i)$为网站 S_i 的更新频率度量值.改进的网页转移概率计算公式为:

$$p'_{ij} = \begin{cases} \alpha \times s(i,j) + (1-\alpha) \times 1/n, & L(i,j) \neq 0 \\ (1-\alpha) \times 1/n, & L(i,j) = 0 \end{cases} \tag{4}$$

其中 $S(i,j)$为网页 i 与网页 j 间存在链接时的跳转概率,计算公式为:

$$s(i,j) = \beta \times 1/d_i + (1-\beta) \times Sim(AT_j, CON_i) \tag{5}$$

$Sim(AT_j, CON_i)$为指向网页 j 的链接文本 AT_j 与网页 i 正文 CON_i 的相似度,使用向量空间模型来计算.

公式第二部分,考虑文本内容的站点更新频度计算公式:

$$Freq(s) = \delta \times \frac{N_a}{D} \times \frac{avgsize(N_a)}{avgsize(N)} + (1-\delta) \times \frac{N_{na}}{D} \tag{6}$$

其中,N_a 为站点内更新的主题型网页的数目;N_{na} 为站点内更新的非主题型网页的数目;N 为更新网页的总数目;D 为更新的网页的更新时间区间;$avgsize(N_i)$为网页集 Ni 的平均大小,δ 为权重参数,$0<\delta<1$.根据网页统计,选择为 0.85.

HotRank 将网页的更新区分为两个部分:网页数目更新和站点质量的更新.通过这种改进,既增加了算法精确度,又可以防止站点通过大量更新主题无关型网页的作弊现象.

2.4 网页变化规律及增量更新频率

我们对"HotRank"算法抽样出的一些比较权威而且更新比较频繁的网站作为种子网址,对这一集合中的站点进行了为期 3 个月的"集中采集".去除垃圾网页,共采集了 1818326 网页.对 10 月份的网页更新情况分别就每天更新情况和一天内的各个小时更新情况分别进行统计,得到了 2008 年 10 月份的"天"变化图和"小时"变化图.从图 1 和图 2 我们可以看出,网站的更新有明显的规律性.

图1　2008年十月份更新情况统计(天)

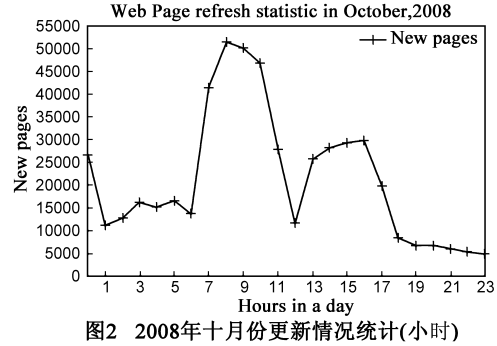


图2　2008年十月份更新情况统计(小时)

2.5 自适应增量更新模型

以"HotRank"方法选择种子站点,根据网页更新密度和频率的自适应的增量更新模型如图 3 所示.模型对 HotRank 抽取的 URL 进行高频更新,其他 URL 由周期较

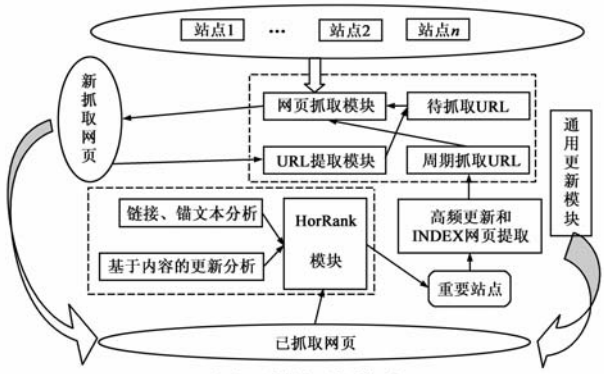


图3　增量更新模型

长的通用更新模块完成. HotRank模块根据网站更新质量和网络环境的变化情况,进行自适应的调整. 基于以上更新模型的自适应增量更新流程描述如表 2 所示.

表 1 10 月份数据统计表

新网页数目	时间跨度	网站数目
510806	30 天	70

表 2 基于 HotRank 的自适应增量更新流程

定义:已抓取网页集:P,增量更新的种子集合 HFURL
(1)在网页集 P 中使用"HotRank"算法选择排序值高的站点集合 S'.
(2)分别在 S'抽取 INDEX 型 URL 集合:HFURL(S')
(3)使用增量更新爬虫对 HFURL(S')进行增量抓取,得到更新网页集和 P'
$P \Leftarrow P' \cup P$
(4)到 HotRank 周期? goto:1,否则 goto:3

3 网络主题发现和热点新闻推荐方法

3.1 基于 LDA 的主题分解和主题距离

聚类是解决主题发现问题的有效方法. 样本点之间的距离度量是影响聚类结果的重要因素. 和基于"文档共现"特征词的方法相比,如夹角余弦和 LM 距离,主题分解后得到的文档距离,考虑了特征词在不同主题中出现的概率,是一种基于"主题共现"的特征,如图 4 所示. 文档对于 k 个主题:$\theta_1 \cdots \theta_k$ 的混合系数可以构成一个 k 维的向量:

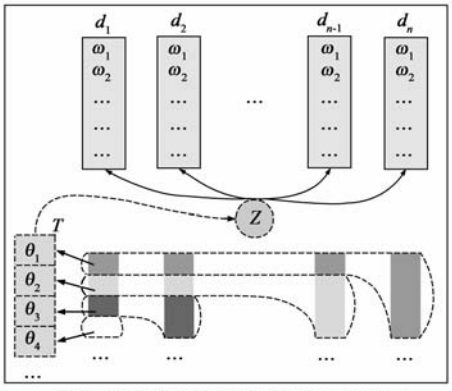


图4 主题模型对文档主题分解示意图

$$V(d_i) = (P(\theta_0 \mid d_i), \cdots, P(\theta_k \mid d_i)) \tag{7}$$

文档 d_i 和 d_j 的主题距离,定义为向量 $V(d_i)$ 和 $V(d_j)$的 KL 距离:

$$D_{TM}(d_i, d_j) = -\sum_{k=1}^{K} p(\theta_k \mid d_i) \ln\left(\frac{p(\theta_k \mid d_i)}{p(\theta_k \mid d_j)}\right) \tag{8}$$

Frey 在 AP 文本聚类的算法中,使用了一种基于信息熵的距离[15]:

$$D_{\inf}(d_i, d_j) = -\left(\sum_{w \in d_i \cap d_i} \log(\mid d_j \mid) + \sum_{w \in d_i - d_i} \log(\mid V \mid)\right).$$

以此距离作为基准方法,和本文提出的结合了 LDA 的算法进行对比,实验表明,本文提出的方法可以提高主题发现的准确率和新闻推荐的召回率.

3.2 基于 LDA 和 AP 的主题新闻推荐方法

结合 LDA 和 AP 的主题发现和推荐算法如表 3 中所示. 其中,ε 是迭代的误差下限,α 是阻尼因子,$\text{loop}_{\max}$ 为最大迭代次数,A 和 R 是临时矩阵. 结果集中的类别按照网页数目排序,把含有较多网页个数类别的中心网页推荐给用户.

表 3 自适应主题发现和热点新闻推荐算法

基于 AP + LDA 聚类的主题新闻推荐算法
输入:网页集 $W, n, \varepsilon, \alpha, loop_{\max}$
输出:推荐给用户的 n 条新闻,和每个主题包含的网页
算法:
(1)LDA 主题分解得到 $p(\theta_0 \mid d_i)$
(2)使用公式 2 计算主题距离矩阵 D
(3)选择矩阵 D 的偏向参数 p
(4)$A \Leftarrow 0$, $R \Leftarrow 0$
(5)While 迭代次数小于 $loop_{\max}$并且误差大于 ε,DO

$$R_{i,k} \Leftarrow \alpha R_{i,k} + (1-\alpha) \begin{cases} D_{i,k} - \max\limits_{k' \neq k} \{A_{i,k} + D_{i,k}\} (i \neq k) \\ D_{i,k} - \max\limits_{k' \neq k} \{D_{i,k}\} (i = k) \end{cases}$$

$$A_{i,k} \Leftarrow \alpha A_{i,k} + (1-\alpha) \begin{cases} \min\{0, R_{k,k} + \sum \max\{0, R_{i',k}\}\} (i \neq k) \\ \sum\limits_{i' \neq k} \max\{0, R_{i',k}\} (i = k) \end{cases}$$

END
$Centers = \arg\limits_i \{R_{i,k} + A_{i,k} > 0\}$
RETURN $\text{TOP}_n(Centers)$

4 实验和分析

4.1 自适应增量更新模型的有效性

实验数据集是由海天园搜索平台中的网页爬虫抓取的 100G 金融网页库,总共包含 2662 个站点,1712739 个网页. 为确定公式中的参数,分别就 λ 取值从 0.1 到 0.9,参数 β 在 0.1 到 0.9 的取值进行实验对比,发现 λ 和 β 分别取值 0.4 和 0.5 时,算法取得较好效果.

表 4 平均算法排序效果对比

排序算法	Kendall 距离	Spearman 距离
SiteRank	0.413495102	0.497268
AggregateRank	0.711938111	0.753809
HotRank	0.773753	0.863333
PageRank 总和	0.878028021	0.8653111

为验证本文提出的 HotRank 的有效性,我们以 alexa 中站点的排名情况作为基准集,分别抽取站点数目为:10,15,20,25,30,35. 求得相对于基准集 Kendall 距离的评价值. 不同站点数目下的四种算法的排序效果对比如表 4 所示.

通过以上的实验数据可以看出,本文的 HotRank 得到的站点排序效果比 AggregateRank 有着平均 6.2% 的性能提升,在不同抽取站点值的最好情况下甚至有着 10% 的提升.

4.2 热点主题新闻推荐算法的有效性

为验证本文方法的有效性,我们标注了包含 30 个主题、5030 个网页的数据集.随机抽取了包含主题数目从 10 到 30 个不等,间隔为 2 的 11 个测试集.

4.2.1 参数选择和算法有效性

首先,假设测试集中的主题数目 k 是已知的,在此前提下对 AP + LDA 和 AP + D_{info} 的主题检测结果进行对比.我们研究了不同的 p 值对聚类结果的影响,从中找到合适该领域数据集的参数值.在以下的实验中,迭代误差下限 ε 设为 0.0001,最大迭代次数 $loop_{\max}$ 设为 500,阻尼因子 α 为 0.5.

图 5 和图 6 是在 20 个主题的测试集中使用不同偏向参数 p 值得到的类别数目和相应的 F-Measure 值的关系.可以看到,聚类结果在预设的偏向参数 p 取得和主题数相同的结果时,聚类结果的 F-Measure 达到最大(如图 6 中所示).

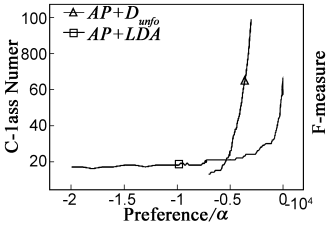


图5　不同预设值p得到的类别数目(20个主题)

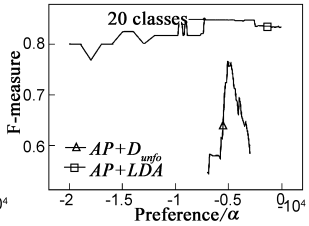


图6　不同预设值p对应聚类结果的F-Measure(20主题)

图 7 是使用 AP + LDA 和 AP + D_{info} 在得到相数目的类别时 F-Measure 的对比.可以看到,将 LDA 和 AP 结合的主题聚类方法,在不同数目的聚类结果中都可以得到主题性更强的结果.

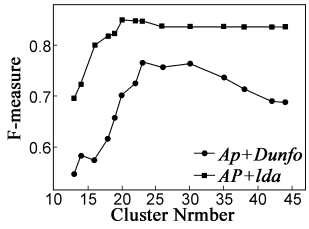


图7　不同聚类数目结果的对比(20个主题)

我们在上述 11 个包含主题数目不等的测试集中重复上述对比试验,寻找到了各个测试集中取得正确的主题数目时的 p 值.如图 8 所示,系统在 $P = 2.5P_m$ 左右得到最好结果.

图 9 是得到最好的聚类结果时,LDA + AP 和 AP + D_{info} 的结果对比.可以看到,在主题数不同的据集中,本文提出的自适应主题检测方法可以取得更好的结果.

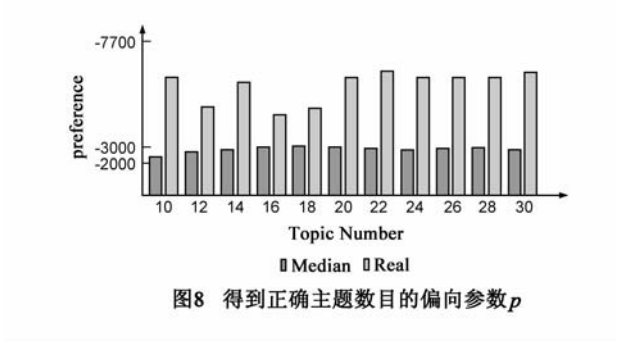


图8　得到正确主题数目的偏向参数p

4.2.2 基于主题的热点新闻推荐的结果分析

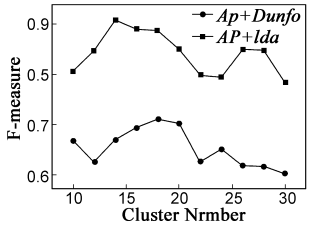


图9　不同数据集中的结果对比

在以上对比试验中,主题的数目被假设为已知的.而在真实网络环境下,主题数目是未知的.由于主题还可以根据语义粒度不同划分为更细的主题,因此,可以选择一个足够大的 k 值作为 LDA 进行主题分解的参数,从而得到文档的主题距离矩阵.

以上述 20 个主题为例,将每个主题随机抽取的网页数增加到 100,得到一个包含 20 个主题、2000 个网页

表 5　推荐结果(20 个主题)

编号	聚类中心网页标题	网页数	召回率	准确率	F-Measure
1	萨达姆在法庭上大发言论,要求伊民众抵抗侵略者	108	0.990000	0.916667	0.951923
2	联合国官员称防治禽流感需要 15 亿美元	101	0.980000	0.970297	0.975124
3	埃及渡轮沉没 1400 多人落入红海	99	0.990000	1.000000	0.994975
4	斯台普斯紧张施工确保悼念仪式万无一失	99	0.990000	1.000000	0.994975
5	核设施验证和对朝援助成为本轮会谈难题	98	0.980000	1.000000	0.989899
6	上海可能迎来雨夹雪	92	0.920000	1.000000	0.958333
7	姚明返沪观战或已物是人非兄弟友谊那刻已变	92	0.920000	1.000000	0.958333
8	巴勒斯坦领导人强调必须尽快实现停火	91	0.840000	0.923077	0.879581
9	中国人踏进太空历史	77	0.770000	1.000000	0.870056
10	布什结束巴以行中东和谈无突破	66	0.630000	0.954545	0.759036
11	阿尔卑斯奶糖被秘密召回	60	0.600000	1.000000	0.750000
12	贾庆林会见中国国民党主席吴伯雄一行	59	0.590000	1.000000	0.742138
13	关注力拓案,检察机关批准逮捕胡士泰等人	53	0.530000	1.000000	0.692810
14	福田康夫决定辞职	50	0.500000	1.000000	0.666667
15	绵阳九洲体育馆灾民返乡自救	47	0.390000	0.829787	0.530612
16	美国大选投票率超过上一届	43	0.430000	1.000000	0.601399
17	吴伯雄拜谒中山陵称振兴中华是共同目标	42	0.410000	0.976190	0.577464
18	上海白领 1300 元包车赴海宁追日	42	0.380000	0.904762	0.535211
19	福田辞职将加剧日本民众对自民党不满	42	0.410000	0.976190	0.577464
20	周正龙被控两宗罪或将当庭宣判	41	0.410000	1.000000	0.581560
21	蔡斌虚心向铁榔头学习,郎平望为国多培养人才	36	0.360000	1.000000	0.529412
22	力拓案震动业界外资矿企高管回国避风头	36	0.360000	1.000000	0.529412
23	邱兴华女儿称帮警察抓爸爸心里很矛盾	33	0.330000	1.000000	0.496241

的测试集.系统在参数 k 取 50,偏置参数设置为 $2.5P_m$ 时得到了一个包含 41 个类的结果.20 个主题在此结果中的 F-Measure 为 0.772041.系统最后的推荐结果如表 5 所示.在表 5 所示的推荐结果中,编号为 17、19、22 的类,分别和编号为 12、14、13 的类是来自同一主题的子主题.由于假设的主题数目 k 多于真实主题数目,这些主题被分解为两个独立的子主题.使用本文算法,向用户推荐 23 条新闻,就可以涵盖全部 20 个主题.

5 实践与应用

以上研究在海天园知识服务平台热点新闻推荐系统得到了应用.该平台运行在 IBM 刀片服务器上.服务器包含 128 个运算单元,100T 左右的存储空间,运算节点间通过带宽为 10Gb/S 的 InfiniBand 连接,操作系统为 Linux.目前检索到 2.4TB 约 85,000,000 个原始网页.热点新闻推荐系统使用了其中 4 个运算节点,采用 C++ 语言实现.

目前,该热点发现系统已经正式运行,自适应和实时性的特点得到较好的验证,根据第三方网络流量最近统计,访问者来自 14 个国家和地区.其中,来自中国大陆的城市为 171 个

参加本研究工作的还有博士生孟宪军、李露,硕士生王志勇、彭伟华等研究人员,他们在本研究工作中完成了极有建设性的工作,在此一并深表谢意.

参考文献:

[1] Hafri Y, Djeraba C. High performance crawling system[A]. In: Proc. of the 6th ACM SIGMM Int'l Workshop on Multimedia Information Retrieval[C]. New York: ACM Press, 2004. 299 – 360.

[2] A Heydon, M Najork. Mercator: a scalable, extensible web crawler[A]. International conference on World Wide Web[C]. New York: ACM Press, 1999. 219 – 229.

[3] Yan HF, Wang JY, Li XM, Guo L. Architectural design and evaluation of an efficient Web-crawling sysgem[J]. Journal of Systems and Software. 2002, 60(3): 185 – 193.

[4] J Edwards, K McCurl, J Tomin. An adaptive model for optimizing performance of an incremental web crawler[A]. International conference on World Wide Web[C]. New York: ACM Press, 2001. 106 – 113.

[5] 高凯.搜索引擎中信息动态采集策略的研究[J].电子学报,2007,35(10):1984 – 1988.
GAO Kai. Dynamic refresh strategy for crawler in search engine[J]. Acta Electronica Sinica, 2007, 35(10): 1984 – 1988. (in Chinese)

[6] 孟涛,王继民,闫宏飞.网页变化与增量搜集技术[J].软件学报,2006,5(17):1051 – 1067.
MENG Tao, WANG Ji-Min, YAN Hong-Fei. Web evolution and incremental crawling[J]. Journal of Software, 2006, 5(17): 1051 – 1067. (in Chinese)

[7] J Cho, H Garcia-Molina. Effective page refresh policies for web crawlers[A]. ACM Transactions on Database Systems[C]. New York: ACM Press, 2003. 390 – 426.

[8] Page L, Brin S, Motwani R. The PageRank Citation Ranking: Bring Oreder to the Web[R]. Technical report, 1998.

[9] Feng G, Liu TY, Wang Y, et al. AggregateRank: bring order to web sites[A]. Proceedings of the 29th annual international ACM SIGIR conference[C]. New York: ACM Press, 2006. 75 – 82.

[10] J Allan, J Carbonell, G Doddington. et al. Topic detection and tracking pilot study: Final report[A]. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop[C]. San Fransisco: Morgan Kaufmann Press Ltd, 1999. 194 – 218.

[11] 刘铭,王晓龙,刘远超.基于主题分析的文本分割技术研究[J].电子学报,2009,37(2):278 – 284.
LIU Ming, WANG Xiao-long, LIU Yuan-chao. Research on text segmentation based on topic analysis[J]. Acta Electronica Sinica, 2009, 37(2): 278 – 284. (in Chinese)

[12] D M Blei, A Y Ng, M I Jordan. Latent dirichlet allocation[J]. J. Mach. Learn. Res., 2003, 3(5): 993 – 1022.

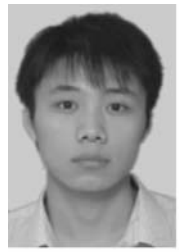
[13] 石晶,胡明,石鑫,戴国忠.基于 LDA 模型的文本分割[J].计算机学报,2008,31(10):1865 – 1873.
SHI Jing, HU Ming, SHI Xin, DAI Guo-Zhong. Text segmentation based on model LDA[J]. Chinese Journal of Computers, 2008, 31(10): 1865 – 1873. (in Chinese)

[14] R Arora, B Ravindran. Latent dirichlet allocation based multi-document summarization[A]. Proceedings of the second workshop on Analytics for noisy unstructured text data[C]. New York: ACM Press, 2008. 91 – 97.

[15] B Frey, D Dueck. Clustering by passing messages between data points[J]. New York: Science, 2007, 315(5814): 972 – 976.

[16] 肖宇,于剑.基于近邻传播算法的半监督聚类[J].软件学报,2008,9(11):2803 – 2813.
XIAO Yu, YU Jian. Semi-supervised clustering based on affinity propagation algorithm[J]. Journal of Software, 2008, 9(11): 2803 – 2813. (in Chinese)

作者简介:



吴永辉 男,1981 年生于河北正定.哈尔滨工业大学计算机科学与技术学院博士研究生.研究方向为网络信息采集、主题检测、自然语言处理. E-mail: yhwu@insun.hit.edu.cn

王晓龙 男,1955 年生于黑龙江.哈尔滨工业大学计算机科学与技术学院教授,博士生导师.研究方向为信息检索、文本挖掘、自然语言处理.