一种基于生命周期理论的文献热点发现方法*

——以肿瘤领域为例

赵迎光　安新颖　李　勇　贾晓峰
(中国医学科学院医学信息研究所　北京 100020)

【摘要】针对文献热点发现方法存在的指标单一、高频常用词过滤效果不明显等问题,将 TDT 领域的生命周期理论和 TF* PDF 方法应用到文献热点发现中,通过跟踪词在时间上的变化率来发现热点词,并确定热点出现的具体时间。实验结果表明,该方法能够有效过滤掉高频常用词,对各时间窗内的研究热点有较高的识别率。

【关键词】生命周期理论　热点发现　文本挖掘
【分类号】G250

A Method for Detecting the Hot Topic of Literature Based on Lifecycle

——A Case Study of Neoplasm Field

Zhao Yingguang　An Xinying　Li Yong　Jia Xiaofeng
(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China)

【Abstract】There are some shortcomings of hot topic detection in literature, such as single index and the inefficient filtering of high - frequency common words. The paper applies lifecycle theory and TF* PDF algorithm to literature detection, which finds the hot words by tracking the variation of words over time, then locates the time hot words appeared. The results of the empirical tests show that this approach is effective in filtering high frequently used terms and identifying hot research topics in time windows.

【Keywords】Lifecycle theory　Hot topic detection　Text mining

1　背　景

科研领域中数字资源的急剧增加以及跨学科研究的快速发展,为科研人员及时、准确地获取其研究领域的热点和动向提出了更高的要求。同时,随着 TDT(Topic Detection and Tracking)技术的发展,越来越多的新闻和社会网络热点挖掘方法和工具被研究和应用,为用户从海量网络信息环境中获取有用信息提供了更加便捷的工具,但是在文献监测领域中热点发现方法则发展较为缓慢。目前国内的文献热点发现技术大都是通过词频、引文以及聚类等传统方法识别热点研究领域[1],存在指标单一和聚类结果模糊等问题。国外关于学科趋势的研究中,基于词频、文档频的方法也较多,例如 BioJournalMonitor[2] 文献趋势发现系统中采用了词频变化率来进行趋势的计算,Swan 等[3] 从研究主题所包含的文档数角度构建了 TimeMines 系统; Guo 等[4] 综合构建了多指标的学科趋势发现系统,其核心仍然集中于词频、共词及聚类,需要建立完善的停用词表来过滤高频停用词,否则热点信息容易被大

收稿日期: 2012 - 10 - 29
收修改稿日期: 2012 - 11 - 12
* 本文系国家"十二五"科技支撑计划基金项目"基于 STKOS 的科技监测应用示范"(项目编号: 2011BAH10B06 - 02)、中国医学科学院医学信息研究所中央级公益性科研院所基本科研业务经费基金项目"基于阈值自动设置的热点识别方法研究"(项目编号: 12R0118)和教育部人文社会科学青年基金项目"基于知识组织体系的科技文献新主题监测研究"(项目编号: 11YJC870001)的研究成果之一。

量高频常用词的噪音所淹没。

生命周期理论已经在 TDT 领域得到了广泛应用并表现出了较好的效果，本文将 TDT 领域的生命周期理论应用到文献监测领域，构建文献领域的热点发现方法，从而为科研和决策提供支持。

2　相关理论研究

2.1　关于热点的界定

Bun 等[5]在新闻热点话题监测方面做了较多的工作，认为新闻中的热点话题是指在一段时间内出现频率相对较高的话题，话题热度通过两个指标来衡量：话题在一篇新闻报道中出现的次数和包含该话题的新闻报道的数量。同时，Bun 等认为每个热点话题都不可能无限“热”下去，它们都要经历一个产生、增长、成熟和消亡的生命周期过程。Chen 等[2]在 Bun 等的基础上对热点话题进行了更详细的描述。

科技文献中的研究热点不同于新闻热点，新闻热点通常以新闻话题的形式出现，而在科研过程中，例如肿瘤研究中的新药物、基因以及治疗方法常常能够成为研究热点，因此本文将热点界定为独立的词，如果某个词在某一个时间段内的热度值相对较高，就将其作为该时间范围内的研究热点。借鉴 Chen 等对新闻热点话题的定义，本文认为文献领域的研究热点需同时满足以下条件：

(1) 在一个期刊中，包含该研究的文献数量相对较多；

(2) 关于该研究的文献分布在多个期刊中；

(3) 与该研究主题相关的其他研究在同时间段内出现较多；

(4) 该研究的热度随时间变化明显。

2.2　词权重计算

词的权重计算是热点发现过程中的重要阶段，通过计算每个词在文章或者期刊中的重要程度，就可以用这些词来表示文章或者期刊。文献[7－9]分别使用了 TF* IDF、词在文档中的分布与整个语料库中分布的差异、共现偏移量等方法来计算权重。在目前的词权重计算方法中，最常用的是 TF* IDF 方法[10]，将词频和逆文档频次的乘积作为词在集合中的权重，这种方法认为在一定文档中出现而不在另外文档中出现的词权重较大，而对在所有文档中都出现的词则赋予较低的权重。在带有时间信息的新闻报道流中，使用 TF* IDF 算法有一定的局限性，例如在大量报道开始报道该事件之前或者在事件发展前期，TF* IDF 算法可以起到较好的识别作用；当大量媒体和新闻都开始报道或者在事件发展中后期，出现该事件中一些关键词的文档越多，其 IDF 值就越小，所以使用 TF* IDF 算法将不能完整识别和跟踪一个热点事件或热点主题，在科研领域也是如此。

Bun 等[11]在 2001 年提出了 TF* PDF 算法，对 TF* IDF 算法进行了改进，避免了在 TF* IDF 算法中当一个重要的词在多个文档中出现时其权重减小的问题。

基于上述原因，本文中的模型采用 TF* PDF 算法作为词频权重计算方法，TF* PDF 算法对于在大量文档中出现频次较高的词赋予较高的权重，而对于出现次数很少的词赋予较低的权重。

2.3　生命周期理论

词权重根据词在文档中的分布情况反映其权重，属于空间上的属性；而词的生命周期反映的是它在时间上的变化趋势。2003 年，Chen 等[12]首先提出了新闻事件的生命周期模型，将新闻事件按照生命周期分为产生、增长、衰退和消亡 4 个阶段，并提出了能量函数的概念来跟踪事件的生命周期，通过新闻报道频次和衰退因子分别建模新闻事件在生命周期中的增长和衰退过程，之后在 TDT 领域得到了广泛的应用。文献[13－16]将生命周期理论应用于新闻、SNS (Social Networking Services)、博客等领域的热点发现，并取得了较好的效果。文献[6]在生命周期理论中引入了能量值和生命值两个指标，并将新闻站点数量也作为一个变量引入模型，在实验中取得了理想的结果。

本文基于这样一个前提：文献中的热点一定具备生命周期特征。如果一个研究主题在时间上的变化率越大，则其生命周期属性越强，成为热点的可能性就越大。对于高频常用词，例如肿瘤领域的 Human、Patients、Malignant Neoplasm(恶性肿瘤)等，虽然其出现频率很高，但是由于不具有生命周期特征，因此不能成为热点。同时，考虑到科研领域热点生命周期较长，因此文献覆盖的时间越长，生命周期值的计算就越准确。

3　模型构建

本文借鉴了 Chen 等[2]在新闻领域的热点监测算

法,将其用于文献监测领域。

3.1 使用 TF* PDF 计算词的权重

将 Bun 等[11]算法中的媒体用期刊来代替。则给定一个文献集合 C,对于词 t,其在期刊集合 c 中的 TF* PDF 值的计算方法为:

$$TF^* PDF = \sum_{c=1}^{|C|} |F_{t,c}| \exp\left(\frac{n_{t,c}}{N_c}\right) \quad (1)$$

$$|F_{t,c}| = \frac{F_{t,c}}{\sqrt{\sum_{k=1}^{K} F_{k,c}^2}} \quad (2)$$

其中,$F_{t,c}$是词 t 在期刊 c 中出现的频次,$|F_{t,c}|$则是对 $F_{t,c}$的标准化处理,$|C|$是文献集合中期刊数量,$n_{t,c}$是期刊 c 中出现词 t 的文档数量,N_c 是期刊 c 中的所有文档数量,K 是 c 中所有词的个数。因此如果一个词的词频越大,并且包含该词的期刊越多,TF* PDF 值就越大。

3.2 使用生命周期理论计算词的变化率

由于文献中的科学研究领域的主题生命周期一般较长,而且各种类型的期刊出版周期(月刊、季刊)不同,因此以年为单位进行热点词的识别,算法如下:

(1)计算词 t 在一个时间窗 s 内得到的能量,用 $E_{t,s}$表示:

$$E_{t,s} = nf \times \left(\sum_{c \in C} \chi_{t,c}^2\right) \quad (3)$$

其中,nf 是能量转换因子(Nutrition Transfer),是主观设定的常量。C 是所有期刊,$c \in C$,使用卡方检验的统计量 $\chi_{t,c}^2$计算时间对词 t 的影响作用:

$$\chi_{t,c}^2 = \frac{(A+B+C+D) \times (AD-BC)^2}{(A+B) \times (C+D) \times (A+C) \times (B+D)} \quad (4)$$

其中,A、B、C、D 均表示词在时间窗及集合中出现的个数,具体算法如表 1 所示:

表 1 公式(4)参数说明

	在期刊 c 中	不在期刊 c 中
在时间窗 s 中	A	B
不在时间窗 s 中	C	D

(2)计算词 t 在时间窗 s 内的生命值:

$$lifesupport_{t,s} = \ln(Eng_{t,s}) \quad (5)$$

其中,$lifesupport_{t,s}$是词 t 在时间窗 s 中的生命值;$Eng_{t,s}$是在时间窗 s 上词 t 获取的能量值。

(3)使用标准差计算词 t 的生命值变化率 V_t:

$$V_t = \sqrt{\frac{1}{N}\sum (lifesupport_{t,s} - \overline{lifesupport})^2} \quad (6)$$

其中,N 是在给定的时间段内时间窗个数,$lifesupport_{t,s}$是 t 在时间窗 s 内的生命值,$\overline{lifesupport}$是 t 在所有时间窗内的生命值的平均值。

3.3 选取具有生命特征的热点词

根据公式(1)和公式(6)分别计算出 TF* PDF 和 V_t 后,计算每个词的热点值:

(1)按照 TF* PDF 值对词进行降序排列,得到其权重顺序值 Weight Order(WO),TP* PDF 值最大的词其 WO = 1。

(2)按照变化率 V_t 顺序值 Variance Order(VO)进行降序排列,变化率最大的词其 VO = 1。变化率越大,生命周期特征越明显,而常用词则不会表现出生命周期特征。

(3)使用 WO 和 VO 的差值计算 Bonus Point(BP),BP 用于增加或减少词的最终权重,当 VO 小于 WO 时,BP 为正数,意味着使用该词的变化率较大:

$$BP = WO - VO \quad (7)$$

(4)计算权重修正系数:

$$WeightModifier = \frac{BP}{TermNum} \quad (8)$$

其中,TermNum 为所有词的总个数,WeightModifier 介于 -1 和 1 之间。

(5)最终权重为:

$$NewWeight = TF^* PDF_t + Var_t \times (1 + WeightModifier) \quad (9)$$

其中,$TF^* PDF_t$ 是 t 的 TF* PDF 值,Var_t 是 t 生命值的变化率。

经过上面的步骤,每个词都被赋予新的权重,然后通过新的权重对所有词进行排序,选取前 n 个词作为整个时间段内的热点词。

3.4 各时间窗中热点词的确定

在新闻热点话题识别中,通常以天为单位识别出短期内的热点事件;而以年为单位识别出的则是长时间段内的热点,因此还需确定各个热点在生命周期中出现的时间,具体算法如下:

热点词 t 在时间窗 s 内的生命值指数(Lifesupport Index)为:

$$lifesupportIndex_{t,s} = \frac{lifesupport_{t,s} - \overline{lifesupport}}{\overline{lifesupport}} \quad (10)$$

生命值指数是对生命值的标准化处理,从而可以实现不同词在同一时间窗内的比较。$lifesupportIndex_{t,s}$是热点词 t 在时间窗 s 内的生命值指数,将每个时间窗内的热点词按照 Lifesupport Index 降序排序,选取前 m

个热点词作为该时间窗内的热点词。

4　实验与分析

为了验证热点发现算法的可行性，笔者使用 Java 语言实现了热点识别程序，对热点发现过程中的 TF* PDF 值、生命值、生命周期变化率进行了计算，识别出每年的热点词。

4.1　数据源

实验数据来自于 PubMed 数据库，根据 2011 年 SCI 影响因子排名，选取肿瘤领域的前 10 个肿瘤期刊，时间范围为 2001 – 2010 年，最终得到 36 229 条 Medline 格式的数据作为实验数据源。

4.2　数据预处理

为了提高热点识别率，在数据预处理过程中，使用概念映射工具 MetaMap[17] 将每篇文章的标题映射为 NCI 主题词，NCI 词表覆盖了肿瘤领域的临床治疗、基础研究等领域的概念和术语。同时保留映射过程中的语义类型，获取 NCI 主题词 11 123 个，并建立文档 – 词矩阵。

为了减少计算时间，提高计算效率和准确度，在计算中只选取了频次大于等于 10 的词进行计算，共 3 136个。

4.3　词权重与生命周期计算实验

使用 TF* PDF 算法计算概念的词频权重，使用生命周期算法计算 Variation，并计算出最终热点值，表 2 为根据公式(1) TF* PDF 选出的前 5 个词与根据公式(9) NewWeight 选出的前 5 个词。

表 2　TF* PDF 与 NewWeight 分别选出的前 5 个词

No.	Term	TF* PDF	No.	Term	NewWeight
1	Malignant Neoplasm	3.779	1	MicroRNA	30.79
2	Patients	1.867	2	Cancer Stem Cell	28.02
3	Clinical Trials	1.623	3	Colony	25.08
4	Malignant Neoplasm of Breast	1.414	4	Sorafenib	23.38
5	New	1.318	5	Inflammatory	22.21

其中，Patients、Clinical Trials、New 等不具备明显热点特征的词 TF* PDF 值较大，这与 TF* PDF 算法对在越多的期刊中出现的词赋予越大的权重有关。为了更清晰地对比两种算法的差别，它们的生命周期曲线如图 1 所示。

从图 1 可以看出，尽管 TF* PDF 值最高的几个词都是肿瘤领域的常用词，出现频次和文档频次都较高，但却不具备生命周期特征，也不能反映出研究热点；而

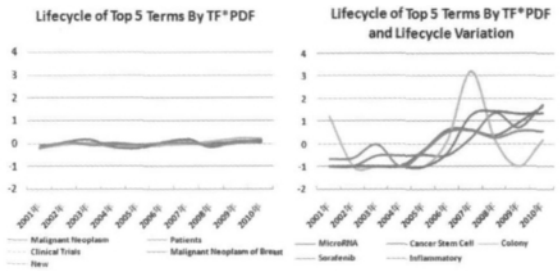

图 1　TF* PDF 与 TF* PDF + Lifecycle Variation 比较

使用 TF* PDF 和生命周期相结合的算法综合考虑了词频、文档频次和生命周期变化率的影响因素，结果中的 MicroRAN 为近年来肿瘤领域的研究热点[18]，Cancer Stem Cell(肿瘤干细胞) 为刚刚兴起的研究领域，有良好的发展前景[19]。因此使用生命周期能够有效过滤掉高频常用词，同时生命周期曲线准确反映了其随时间的变化趋势。

4.4　热点计算结果分析

根据公式(10) 计算热点词在每个时间窗内的生命值指数，获取每年热点词；同时使用 UMLS 提供的语义类型将热点词分为不同的类，得到每个细分领域的研究热点，2009 年与 2010 年肿瘤领域热点药物如表 3 所示：

表 3　2009 年与 2010 年肿瘤领域热点药物

2009 年	2010 年
Lapatinib	Methenamine
Azacitidine	PAZOPANIB
Panitumumab	Nilotinib
Sunitinib	DENOSUMAB
Omega – 3 Fatty Acids	mTOR Inhibitor
Nilotinib	Gold
PAZOPANIB	Lomustine
Ascorbic Acid	Niacin
LENALIDOMIDE	Panitumumab
Dasatinib	Formaldehyde

表 3 中的药物基本上都是肿瘤分子靶向药物，分子靶向药物也是近年来的研究热点，利用肿瘤细胞与正常细胞之间分子生物学的差异，抑制肿瘤细胞的生长增殖，最后使其死亡。近年来，肿瘤分子靶向治疗因具有疗效高、不良反应少且轻等特点而备受瞩目，各种新型分子靶向治疗药物成为近年来的研究热点，并逐步成为临床肿瘤治疗的重要部分。2009 年肿瘤领域热点研究药物生命周期如图 2 所示。

为了验证算法在时间维度上的准确性，将结果和 Google Trends 进行了对比。Google Trends 通过分析

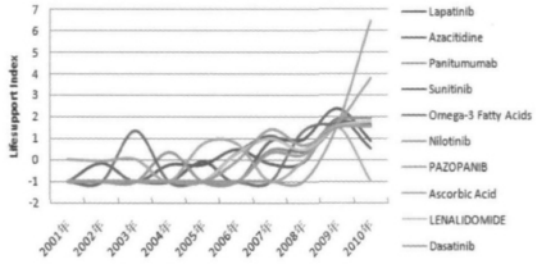

图 2　2009 年肿瘤领域热点研究药物生命周期

Google 网络搜索以计算用户输入的词被搜索的次数，并将其与 Google 上随时间推移的搜索总量相比较，用图表向用户显示结果。以 2009 年的 10 个热点药物为关键词，从 Google Trends 中下载完整数据（2004 年 - 2010 年，Google 目前只提供 2004 年之后的数据），并对其进行标准化处理，如图 3 所示：

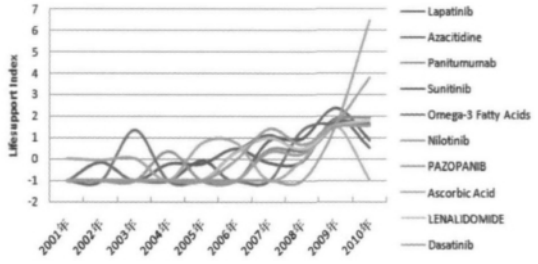

图 3　Google Trends 中用户对 2009 年科研热点药物的关注度

从图 3 可以看出，在 10 个所关注的热点词中，7 个热点词和图 2 的生命周期一致，均在 2006 年与 2007 年得到了用户关注度的峰值，而其余的三个词在这几年间则一直是平缓的下降趋势，分别是 Omega - 3 Fatty Acids、Ascorbic Acid、Azacitidine。而图 2 中的 Omega - 3 Fatty Acids 在 2003 年、Ascorbic Acid 在 2005 年都有一些峰值的出现。通过对语义类型的分析，发现是由于 MetaMap 中的语义类型将临床药物、有机化学物质和药理物质归为一类，因此造成了不属于药物的 Omega - 3 Fatty Acids、Ascorbic Acid 等并没有被过滤掉，这是算法需要改进的地方。

此外，图 3 和图 2 中的峰值相差了 2 - 3 年的时间，以 Panitumumab（帕尼单抗）为例进行分析。Panitumumab 是结直肠癌治疗的分子靶向药物，2005 年 7 月，Panitumumab 获得 FDA（Food and Drug Administration）快速通道审批资格。2005 年底，安进公司及其合作伙伴 Abgenix 公司共同向 FDA 提交了该品生物制剂许可申请，用于治疗化疗失败后转移性结直肠癌。通过对 Google Trends 的搜索发现，2006 年的用户关注度大都与该事件有关，相关新闻报道量增多，而此时对该药物的研究大都处于保密状态，研究文献较少。该药物投入市场后，在基础研究和临床中的研究逐渐增多，因此在科研领域，该药物研究的生命周期峰值出现在 2009 年。其他的几种药物均有类似的情况。

5　结　语

本文在生命周期理论的基础上，设计并实现了基于生命周期的文献热点发现方法，并在肿瘤领域进行了初步的实验，实验结果表明该算法能够有效地过滤高频常用词以及识别时间窗内的研究热点，但仍存在一些问题需要在下一步的工作中继续深入研究：

（1）算法中的阈值设置：算法中涉及多次根据阈值选取热点词的过程。在大规模、不同类别的数据处理中，依靠人工设置来选取存在可靠性和效率较低等问题，因此在未来研究中需要解决各个阶段阈值的设置问题。

（2）热点主题的范围：本文通过算法较为准确地识别出部分热点词，但是并没有发现和这些词相关的其他词，例如对于热点药物 Panitumumab，如果能够找出和它相关的疾病以及治疗方法，将会为科研人员提供更大的帮助，这也是下一阶段的研究方向。

（3）算法在其他领域的可扩展性：由于该算法目前只在肿瘤领域进行了测试，在其他学科的通用性还需继续试验。同时，本文使用了医学领域的概念映射工具对词进行降维并识别其语义类型，最后实现了热点药物的提取。在其他学科中如果没有类似工具的支持，则算法还需改进。

参考文献：

[1] 章成志，梁勇. 基于主题聚类的学科研究热点与研究趋势监测方法[J]. *情报学报*，2010，29（2）：342 - 349.（Zhang Chengzhi，Liang Yong. Detecting Hotspot and Trend of Disciplines Using Topic Clustering[J]. *Journal of the China Society for Scientific and Technical Information*，2010，29（2）：342 - 349.）

[2] Mörchen F，Dejori M，Fradkin D，et al. Anticipating Annotations and Emerging Trends in Biomedical Literature[C]. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*（*KDD'08*），Las Vegas，Nevada，

USA. New York: ACM, 2008: 954 – 962.

[3] Swan R, Jensen D. TimeMines: Constructing TimeMines with Statistical Models of Word Usage [C]. In: *Proceedings of the ACM SIGKDD* 2000 *Workshop on Text Mining*, Boston, MA, USA. ACM, 2000: 73 – 80.

[4] Guo H, Weingart S, Börner K. Mixed – indicators Model for Identifying Emerging Research Areas [J]. *Scientometrics*, 2011, 89(1): 421 – 435.

[5] Bun K K, Ishizuka M. Topic Extraction from News Archive Using TF* PDF Algorithm [C]. In: *Proceedings of the* 3*rd International Conference on Web Information Systems Engineering* (*WISE'* 02), Singapore. Washington, DC: IEEE Computer Society, 2002: 73 – 82.

[6] Chen K Y, Luesukprasert L, Chou S 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.

[7] Kumaran G, Allan J. Text Classification and Named Entities for New Event Detection [C]. In: *Proceedings of the* 27*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'* 04), Sheffield, UK. New York: ACM, 2004: 297 – 304.

[8] Pazienza M T. Information Extraction in the Web Era [M]. Springer, 2003.

[9] Hisamitsu T, Niwa Y. A Measure of Term Representativeness Based on the Number of Co – occurring Salient Words [C]. In: *Proceedings of the* 19*th International Conference on Computational Linguistics* (*COLING'* 02). Stroudsburg: Association for Computational Linguistics, 2002: 1 – 7.

[10] Holmes D E, Jain L C. Data Mining: Foundations and Intelligent Paradigms, Volume 1: Clustering, Association and Classification [M]. Springer, 2012.

[11] Bun K K, Ishizuka M. Emerging Topic Tracking System [C]. In: *Proceedings of the* 1*st Asia – Pacific Conference on Web Intelligence: Research and Development* (*WI'* 01). London: Springer – Verlag, 2001: 125 – 130.

[12] Chen C C, Chen Y T, Sun Y S, et al. Life Cycle Modeling of News Events Using Aging Theory [C]. In: *Proceedings of Machine Learning: ECML* 2003. Berlin, Heidelberg: Springer – Verlag, 2003: 47 – 59.

[13] Liu M, Liu Y, Xiang L, et al. Extracting Key Entities and Significant Events from Online Daily News [C]. In: *Proceedings of the* 9*th International Conference on Intelligent Data Engineering and Automated Learning* (*IDEAL'* 08), Daejeon, South Korea. Springer, 2008: 201 – 209.

[14] Wang C, Zhang M, Ru L, et al. Automatic Online News Topic Ranking Using Media Focus and User Attention Based on Aging Theory [C]. In: *Proceedings of the* 17*th ACM Conference on Information and Knowledge Management* (*CIKM'* 08), Napa Valley, California. New York: ACM, 2008: 1033 – 1042.

[15] Zheng D H, Li F. Hot Topic Detection on BBS Using Aging Theory [C]. In: *Proceedings of the International Conference on Web Information Systems and Mining* (*WISM'* 09). Berlin, Heidelberg: Springer – Verlag, 2009: 129 – 138.

[16] Lee Y, Jung H Y, Song W S, et al. Mining the Blogosphere for Top News Stories Identification [C]. In: *Proceedings of the* 33*rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'* 10), Geneva, Switzerland. New York: ACM, 2010: 395 – 402.

[17] MetaMap [EB/OL]. [2012 – 09 – 26]. http://metamap.nlm.nih.gov/.

[18] 王福飞. microRNA 在肿瘤诊治中的进展 [J]. 江西医药, 2011, 46(6): 580 – 582. (Wang Fufei. The Progress of microRNA in Cancer Diagnosis and Treatment [J]. *Jiangxi Medical Journal*, 2011, 46(6): 580 – 582.)

[19] 侯萍, 李剑平. 肿瘤干细胞的研究进展 [J]. 中国组织工程研究与临床康复, 2011, 15(14): 2629 – 2632. (Hou Ping, Li Jianping. Advances in Cancer Stem Cell Research [J]. *Journal of Clinical Rehabilitative Tissue Engineering Research*, 2011, 15(14): 2629 – 2632.)

(作者 E – mail: zhao.yingguang@imicams.ac.cn)