

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

结合知识组织体系的突发主题监测研究*

摘 要: 文章结合知识组织体系, 构建了基于 Kleinberg 突发监测算法的突发主题发现方法, 并在肿瘤学领域利用科技文献对该方法进行了测试。结果表明, 知识组织体系的应用可以有效提高突发主题的识别效率。

关键词: 知识组织体系; 突发事件; 理论研究

Abstract: Combining with the knowledge organization system, this paper develops a burst topics discovery method based on the Kleinberg's burst detection algorithm. And this method is tested in the field of oncology by the use of scientific literatures. The results show that the application of the knowledge organization system can effectively improve the identification efficiency of burst topics.

Keywords: knowledge organization system; emergency; theoretical study

文本数据挖掘的一个基本问题是从按时间连续到达的文本流中提取出有意义的结构。例如在连续到达的新闻中, 其包含的主题会经历从出现到逐渐受关注然后消失的过程。在特定研究领域中发表的文献在一个更长时间范围内也有类似的现象。

目前许多文本挖掘工作都是基于如下假设: 如果一个主题在出现频率方面具有显著增加的特征, 则这个主题在文本流中的出现被认为是一个事件的突发。自 2002 年 Kleinberg 提出突发监测算法之后, 出现了不少关于 Kleinberg 算法的应用、改进及扩展。但在当前的科技监测研究和实践中, 仍存在以下问题: 监测输出的结果一般为单个自由词或短语, 监测结果仍需要大量的人工筛选和解读, 对事件的挖掘意义有限。而知识组织体系 (如本体、叙词表等) 能精确地表征研究主题, 从而有效提高监测精度, 因此本文尝试将知识组织体系应用到科技监测流程的选词和主题聚类环节, 希望知识组织体系的应用可以提高突发主题的识别效率。

1 相关概念及研究现状

1.1 突发主题

本文主要关注科技文献领域中的突发现象, 并将突发主题定义为: 在科技领域中, 伴随着某一科技事件 (如重大科学实验进展、重要发明出现等) 的发生, 在短时间内

被突然广泛关注的研究领域。突发主题在某一时间点之前相关报道量较少, 在某一科技事件发生后的一个较短时期内被突然大量报道。突发主题的持续时间较短, 并且随着时间推移, 突发主题有可能转化为热点主题, 也可能消逝或退化为普通主题。主题的构成形式为一组相互关联的词或概念。

1.2 突发监测

2002 年, Kleinberg 提出的突发监测算法基于概率机对不同时间段上词出现的频次进行建模^[1]。概率机的状态确定了某时间点上词出现频次的期望值, 而概率机的状态改变由概率模型控制。词突发时, 概率机处于高频状态。给定文档集合, 确定状态的个数、状态差异的大小, 以及状态改变的成本, 对状态改变的模型求最优解, 得出概率机状态变化的最优时序序列, 最终得出词的突发权重。

在 Kleinberg 提出突发监测算法之后, 出现了不少关于 Kleinberg 算法的应用、改进及扩展。陈超美在其科学图谱软件 CiteSpace 中使用 Kleinberg 算法计算突发词^[2]。Toshiaki 在 BLOG 和 BBS 中监测突发现象, 并针对 Kleinberg 算法没有考虑到文本本身到达的不均匀所造成的伪突发现象, 当文本流的到达速率有较大波动时, 容易造成错误的检测现象, 提出使用某一类文本在总体文本流中所占的相对比例来标志该类文本到达速率的快慢, 因而能够较好地处理非均匀分布的数据^[3]。Mane 利用 Kleinberg 的突发检测算法对 1982—2001 年的 PNAS 文献进行挖掘, 得到多个关键词的共同演化过程, 对这些关键词进行分类和聚类, 能够发现该领域的热点与态势变化及预测未来的趋势^[4]。

* 本文为国家科技支撑计划课题“基于 STKOS 的科技监测应用示范” (项目编号: 2011BAH10B06-02) 和中央公益性科研院所基本科研业务费课题“基于改进 Kleinberg 算法的突发监测研究” (项目编号: 12R0117) 的成果。

此外, Russell Swan 等也提出了基于卡方统计的计算方法, 用来发现词在某个时间窗内的热度, 并对识别出的热点词进行排序^[5-7], 该方法适用于计算词在限定时间窗内的热度, 在用于词的动态跟踪时有一定的局限性。

1.3 知识组织体系

应用知识组织体系对科技文献进行深入的加工处理已经开展了很多年, 如美国国立医学图书馆利用 MeSH 词表对医学文献进行主题标引, 并赋予每篇文献主题词, 通过主题词将属分关系词、相关关系词以及等同关系词进行了有效的组织, 从而深入地揭示文章的主要内容。美国国立医学图书馆 (NLM) 研制的一体化医学语言系统 (Unified Medical Language System, UMLS) 是一个计算机化的可持续发展的生物医学检索语言集成系统和机读情报资源指南系统, 可以提高计算机程序“理解”用户提问中生物医学词汇含义的能力。英国普林斯顿大学认知科学实验室建立的 WordNet 是一个语言学本体库^[8], 涵盖所有领域概念及其关系, 在自然语言处理研究方面应用很广。美国肿瘤研究所开发的 NCI 词表涵盖了肿瘤领域的临床治疗、转化和基础研究、公共信息和管理活动, 可以准确地定位与肿瘤相关的概念^[9], 该词表也是 UMLS 的来源词表之一。

在将自由文本映射至概念方面, 许多科研人员已经开发了不同的程序, 如 H. J. Lowe 等人研发的 MicroMeSH^[10], D. A. Evans 等人研发的 Clarit^[11], R. A. Miller 等人研发的 Chartline^[12], W. R. Hersh 等人研发的 Sapphire^[13] 及 A. R. Aronson 等人研发的 MetaMap^[14] 等。这些系统都得到了较为广泛的应用并且取得了不同层次成果, 其中应用最为广泛的是 MetaMap, 它是一个基于 UMLS 的文本映射工具。在本文中将应用 MetaMap 和 NCI, 以便将肿瘤学领域的自由文本映射为概念。

2 结合知识组织体系的突发主题识别方法设计

本文针对突发主题的定义, 设计了结合知识组织体系、基于 Kleinberg 突发监测算法的识别方法, 基本思路为: 将文本中的自由词映射为概念, 提取每个时间片内出现每个概念的文档频次, 使用 Kleinberg 算法计算得到突发概念, 选出包含突发概念的所有文档, 使用这些文档中的所有概念聚类, 提取包含突发概念的类, 获得突发主题。流程如图 1 所示。

其过程描述如下:

步骤 1: 文档集合为 D , 将所有文档按时间片进行划分:

$$D = \{D_1, D_2, \dots, D_n\}$$

n 为以年为单位的时间片。

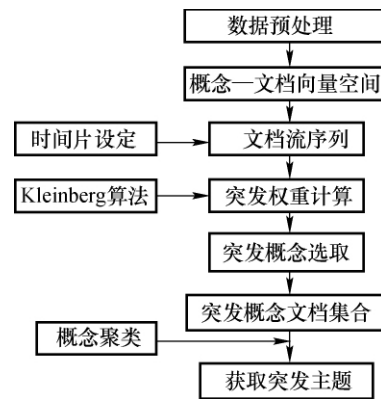


图1 突发主题识别过程

步骤 2: 统计每个时间片中每个概念 w_i 的文档频次 f_i 。

步骤 3: 利用 Kleinberg 算法计算得出突发概念 burst。

步骤 4: 选取包含有 burst 的所有文档:

$$D_{burst} = \{d \mid w_i = burst; w_i \in d\}$$

步骤 5: 对 D_{burst} 中的所有概念进行聚类, 形成类簇 $C = \{C_1, C_2, \dots, C_m\}$ 。

步骤 6: 判定突发主题 T_{burst} :

If $\forall w_j \in C_m$;

$w_j = burst$;

then $C_m = T_{burst}$;

得到突发主题 T_{burst} 。

2.1 概念映射

对大数据集的处理是在一个高维空间内进行的, 而高维空间点之间的数据点差别较小, 运行处理效率较低^[15]。本文使用 MetaMap 对文献中的自由词进行概念映射, 一方面可以有效地实现降维效果; 另一方面, 概念相对于自由词可以更精确地表达文章主题。

2.2 Kleinberg 突发监测算法

本文使用了 Kleinberg 突发监测算法的二状态模型, 该模型是由一个具有两个状态 q_0 和 q_1 的概率自动机 A 构成的, 这两个状态与文档发出速率的高低相对应。当 A 处于状态 q_0 , 文档以较慢的速率进行发送, 连续发出的两个文档间的间隔 x 满足概率密度方程 $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ 。当 A 处于状态 q_1 , 文档以较快的速率进行发送, 连续发出的两个独立的文档间的间隔 x 满足概率密度方程 $f_1(x) = \alpha_1 e^{-\alpha_1 x}$, 且 $\alpha_1 > \alpha_0$ 。在不同文档之间, A 以概率 $p \in (0, 1)$ 发生状态变化, 以概率 $1-p$ 保持当前状态, 而且与之前发出的文档和状态变化相互独立。

即 A 的初始状态为 q_0 , 在每个文档 (包括第一个) 发送之前, A 先以概率 p 改变状态, 然后发出一个文档,

根据当前状态所确定的分布, 确定下一个文档的时间间隔^[5-7]。

通过给定的文档集, 可以利用这个文档生成模型, 找到一个可能的状态序列。假设给定 $n+1$ 个文档, 以及它们精确的到达时间。这样就确定了 n 个文档的到达间隔 $X = (x_1, x_2, \dots, x_n)$ 。利用贝叶斯公式确定状态序列 $q = (q_{i_1}, q_{i_2}, \dots, q_{i_n})$ 的条件概率, 每个状态序列 q 都有一个关于间隔序列的概率密度方程 $f_q, f_q(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_{i_i}(x_i)$ 。若 b 代表序列 q 中状态转移的次数, 即满足 $q_{i_i} \neq q_{i_{i+1}}$ 的 i_i 的数量, 则概率 q 为:

$$\left(\prod_{i_i \neq i_{i+1}} p \right) \left(\prod_{i_i = i_{i+1}} 1-p \right) = p^b (1-p)^{n-b} = \left(\frac{p}{1-p} \right)^b (1-p)^n$$

令 $i_0 = 0$ 则:

$$Pr[q|x] = \frac{Pr[q]f_q(x)}{\sum_q Pr[q]f_q(x)} = \frac{1}{2} \left(\frac{p}{1-p} \right)^b (1-p)^n \prod_{i=1}^n f_{i_i}(x_i)$$

其中 Z 是 $\sum_q Pr[q]f_q(x)$ 的标准化常数。找到一个状态序列 q 使得这个概率最大化等价于找到一个序列使下式最小化:

$$-\ln Pr[q|x] = b \ln \left(\frac{1-p}{p} \right) + \left(\sum_{i=1}^n -\ln f_{i_i}(x_i) \right) - n \ln(1-p) + \ln Z$$

该式等价于找到一个状态序列 q 使得如下的成本函数最小化:

$$c(q|x) = b \ln \left(\frac{1-p}{p} \right) + \left(\sum_{i=1}^n -\ln f_{i_i}(x_i) \right)$$

通过最优的状态序列可以揭示词突发的层次结构, 而词的突发程度可以通过计算其从非突发状态跃迁到突发状态的成本得到, 即:

$$b_{\text{weight}} = \sum c_0 - c_1$$

通过以上方法, 最终得出所有概念在每个时间片上的突发权重, 设置阈值, 即得到每个时间片的突发概念。

2.3 突发主题获取

对突发词或突发概念的挖掘并不是本文研究的最终目标, 把突发词作为特征来挖掘与其相关联的信息才更能体现突发监测的意义。做法是选取某一年度中包含有当年突发概念的文档, 提取这些文档中所有的概念进行聚类, 使用目前应用较为广泛的层次聚类方法对概念进行聚类。在类簇成员间相似度的计算上, 目前常用的方法是根据词的共现关系进行计算, 该方法执行效率较高, 但是不能挖掘词间的内在语义关系, 因此引入知识组织体系计算其语义关系。概念间的语义关系由其所在本体决定, 首先选择普遍存在的层次关系来表征其语义关系, 然后分别根据概念间的共现关系、本体的层级关系计算其混合相似度, 将二

者按照预定的权重进行加权计算, 公式如下:

$$S(a, b) = w_w S_w(a, b) + w_n S_n(a, b)$$

其中, w_w 和 $w_n \geq 0$, 分别表示共现关系、层级关系对整体相似度的贡献大小, 且 $w_w + w_n = 1$ 。函数 S_w 和 S_n 是用来计算概念 a 和 b 的共现关系、层级关系。

在聚类得到的类簇中选取包含有突发概念的类簇, 再经过人工筛选, 得到突发主题。

3 实验

为了验证突发主题发现方法的可行性, 使用 C 和 R 实现了突发主题识别程序, 对突发主题发现过程中的突发权重进行了计算, 最终识别出热点主题。

3.1 数据源选择

在实验数据源的选择上, 根据 2011 年 SCI 影响因子排名, 选取了 PubMed 数据库中肿瘤领域的前 10 种期刊 2001—2010 年的 36 229 篇文献, 存储为 Medline 格式。

3.2 概念映射

使用 MetaMap, 将来源词表限制为 NCI Thesaurus (NCIt)。将每篇文献的标题映射为 UMLS 中的 NCIt 概念, 生成 NCIt 概念 11 123 个, 并建立文档—概念矩阵。

3.3 突发权重计算

使用 Kleinberg 算法计算每个概念的突发权重, 以 2008 年和 2009 年为例, 每年突发权重最大的 10 个概念, 如表 1 所示。

表 1 2008—2009 年突发概念

概 念	权重	起始年份	结束年份
Polyacrylamide Gel	281.569	2009	2010
TACSTD2 Gene	67.062	2009	2010
Interleukin 12 Receptor	16.784	2009	2010
Caspase Gene	9.374	2009	2010
Chromosome 18q	4.478	2009	2010
Chronic Hepatitis	3.885	2009	2010
Caveolin 2	1.311	2009	2010
Major Late Transcription Factor 1	1.169	2009	2010
Translational Activation	1.169	2009	2010
Tubular Epithelial Cell Count	1.169	2009	2010
FMR1 Gene	350.586	2008	2008
MSC Gene	17.541	2008	2008
Polysomy 17	10.985	2008	2008
MLN518	2.591	2008	2008
RNA-Induced Silencing Complex	1.370	2008	2010
Gene Function	1.251	2008	2010
Multiple Organ Failure	1.251	2008	2010
Benzyl Isothiocyanate	1.219	2008	2010
Bordetella	1.219	2008	2010
Mucopolidosis Type IV	1.219	2008	2010

3.4 突发主题获取及识别

1) 突发概念扩展。选取某一年度中包含有当年突发概念的文档,提取这些文档中所有的概念。

2) 聚类。根据共现关系计算概念间的共现相似度,并计算概念间的语义距离,对两者进行加权计算出最终概念相似度,两者权重 w_w 、 w_n 均为 0.5。

3) 结果分析。在实验结果中对聚类结果进行分析,选择包含有突发概念的类,列举几个类进行分析。

2009 年突发主题 1 为癌症基因机制研究(黑体为突发概念),如表 2 所示。

表 2 2009 年突发主题 1

概念名称	类型	中文译名
Polyacrylamide Gel	Manufactured Object	聚丙烯酰胺凝胶
TACSTD2 Gene	Gene	TACSTD 2 基因
B Lymphocyte	Anatomic Structure, System, or Substance	B 淋巴细胞
Malignant Neoplasm	Disease, Disorder or Finding	恶性肿瘤
Caspase Gene	Gene	切冬酶基因
Colorectal Carcinoma	Disease, Disorder or Finding	结肠癌

2008 年突发主题 1 为抗肿瘤药物研发(黑体为突发概念)如表 3 所示。

表 3 2008 年突发主题 1

概念名称	类型	中文译名
MSC Gene	Gene Product	MSC 基因
MLN518	Drug, Food, Chemical or Biomedical Material	坦度替尼
RNA Induced Silencing Complex	Anatomic Structure, System, or Substance	RNA 诱导沉默复合体
Benzyl Isothiocyanate	Drug, Food, Chemical or Biomedical Material	异硫氰酸苄酯
Adenylate Cyclase	Gene Product	腺苷酸环化酶

该主题主要是与抗肿瘤药物研发相关的研究,其中坦度替尼(MLN518)是 Millennium 制药公司研制开发的一种新型口服小分子多靶点受体酪氨酸激酶抑制剂,属于一种合成的嘧啶咪唑类化合物,主要治疗急性骨髓系白血病。对受体酪氨酸激酶有抑制作用,包括 FMS 样酪氨酸激酶-3 (FLT-3)、血小板衍生长因子受体 (Platelet-Derived Growth Factor Receptor, PDGFR) 和干细胞因子受体 (c-KitR)。

2008 年 Rossetto 等研究发现,热带水果番木瓜中含有丰富的具芳香烷侧链的 ITCs 成分——异硫氰酸苄酯 (Benzyl Isothiocyanate, BITC) [16]。BITC 有效抑制致癌因

子——DNA 加和物的形成,从而抑制由致癌因子诱导的肿瘤形成。此外,BITC 在一些人类细胞系中还可以诱导肿瘤细胞程序化凋亡。

4 存在问题及下一步工作

本文对知识组织体系在突发监测中的应用进行了初步的尝试,设计并实现了基于 Kleinberg 算法的突发主题发现方法,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识别时间窗内的突发主题,但仍存在一些问题需要在下一步的工作中继续改进和完善。

1) 突发主题获取方法。本实验在突发概念扩展的基础上聚类以获得突发主题,虽然聚类对突发概念所表达的含义进行了扩展,但是形成的类簇中往往会包含有大量无意义的概念,因此在下一步的研究中需要改进突发主题的获取方法。

2) 知识组织体系的应用。知识组织体系,如本体和叙词表,其本质是专家经验的载体,在具有精确性的同时,还具有滞后性的确定,知识组织体系中收录的概念是落后于科技文献中新知识、新主题的产生的。因此结合知识组织体系进行突发监测在新知识、新主题的探测上必然存在缺陷,虽然新主题也被认为是突发主题的一种特殊状态,但是笔者认为这种方法对于长时间序列上的科技文献研究主题的突发状态的挖掘是有效的。而对于如何应用知识组织体系以提高新主题的挖掘将作为后续研究展开。

3) 结果分析和展示。后续的工作中,将根据知识组织体系分别对识别出的各种类别的概念进行单独分析,从而为各种不同需求的科研工作者提供服务。□

参考文献

- [1] KLEINBERG J. Bursty and hierarchical structure in streams [C] // Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.
- [2] CHEN Chaomei. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature [J]. Journal of the American Society for Information Science and Technology, 2005, 57 (3).
- [3] TOSHIKAKI F. Identification of bursts in a document stream [C] // First International Workshop on Knowledge Discovery in Data Streams. Pisa, Italy, 2004.
- [4] MANE K K. Mapping topics and topic bursts in PNAS [C] // Proceedings of the National Academy of Sciences, 2004.
- [5] SWAN R, ALLAN J. Extracting significant time-varying features from text [C] // Proc. 8th Intl. Conf. on Information Knowledge Management, 1999.
- [6] SWAN R, ALLAN J. Automatic generation of overview time-lines [C] // Proc. SIGIR Intl. Conf. Information Retrieval, 2000.

(下转第 128 页)

- and behavior: a meta-analysis [J]. Computers in Human Behavior, 1997, 13 (1): 1-22.
- [19] HONG T. The internet and tobacco cessation: the roles of internet self-efficacy and search task on the information-seeking process [J]. Journal of Computer-Mediated Communication, 2006, 11 (2): 536-556.
- [20] TSAI M J, TAISE C C. Information searching strategies in web-based science learning: the role of internet self-efficacy [J]. Innovations in Education and Teaching International, 2003, 40 (1): 43-50.
- [21] 陈祖琴, 葛继科, 郑宏. 自我效能感与用户信息查寻行为 [J]. 图书情报工作, 2007 (7): 54-56, 119.
- [22] ROBERT G C, MANUEL P Q. Factors and evaluation of refinding behaviors [EB/OL]. [2012-05-30] <http://pim.ischool.washington.edu/pim06/files/capra-paper.Pdf>.
- [23] DAVID P, et al. A cyclic model of information seeking in hyperlinked environments: the role of goals self-efficacy and intrinsic motivation [J]. International Journal of Human-Computer Studies, 2007, 65 (2): 170-182.
- [24] WOOD R, et al. The impact of enactive exploration on intrinsic motivation strategy and performance in electronic search [J]. Applied Psychology, 2000, 49 (2): 263-283.
- [25] MUN Y Y, HWANG Y. Predicting the use of web-based information systems: self-efficacy, enjoyment, learning goal orientation, and the technology acceptance model [J]. International Journal of Human-Computer Studies, 2003, 59 (4): 431-449.
- [26] USLU E L Y K. Can ITC usage make a difference on student teachers' information literacy self-efficacy [J]. Library & Information Science Research, 2007, 29 (1): 92-102.
- [27] 刘小燕. 上海大学生网络自我效能的实证研究 [D]. 上海: 上海师范大学, 2005: 35.
- [28] STAMATOPOULOS A, MACKOY R. Effects of library instruction on university students' satisfaction with the library: a longitudinal study [J]. College & Research Libraries, 1998, 59 (4): 323-334.
- [29] REN H W. Library instruction and college student self-efficacy in electronic information searching [J]. Journal of Academic Librarianship, 2000, 26 (5): 323-328.
- [30] ROLDAN M, WU Y D. Building context-based library instruction [J]. Journal of Education for Business, 2004, 79 (6): 323-328.
- [31] 李富峰. 青少年互联网自我效能感、搜索策略和信息焦虑的关系 [D]. 北京: 首都师范大学, 2009.
- 作者简介: 谈大军, 女, 1966年生, 博士, 副教授, 硕士生导师。研究方向: 信息检索, 信息伦理。
任淑宁, 女, 1987年生, 硕士。
张新兴, 男, 1982年生, 博士。
- 收稿日期: 2012-10-29

(上接第123页)

- [7] SWAN R, JENSEN D. TimeMines: constructing timelines with statistical models of word usage [C]. KDD-2000 Workshop on Text Mining, 2000.
- [8] <http://www.cogsci.princeton.edu/~wn>.
- [9] NCI Thesaurus. [EB/OL]. [2012-10-24]. <http://ncit.nci.nih.gov>.
- [10] LOWE H J, BARNETT G O. MicroMeSH: demonstration of a microcomputer system for searching and exploring the national library of medicine's medical subject headings (MeSH) vocabulary [C] // Proc. of Eleventh SCAMC, Washington, D. C., November, 1987: 81-82.
- [11] EVANS D A, HERSH W R, et al. Automatic indexing of abstracts via natural-language processing using a simple thesaurus [J]. Medical Decision Making, 1991, 11 (4): 410-425.
- [12] MILLER R A, GIESZCZYKIEWICZ F M, VRIES J K, et al. Chartline: providing bibliographic references relevant to patient charts using the UMLS metathesaurus knowledge sources [C] // Proceedings Annual Symposium on Computer Application in Medical Care, 1992: 86-90.
- [13] HERSH W R, GREENES R A. SAPHIRE: an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships [J]. Computers and Biomedical Research, 1990, 23: 410-425.
- [14] MetaMap Protal [EB/OL]. [2012-09-26]. <http://metamap.nlm.nih.gov>.
- [15] BEYER K, GOLDSTEIN J, RAMAKRISHNAN R, et al. When is "nearest neighbor" meaningful [C] // Proceedings of the 7th International Conference on Database Theory. London: Springer-Verlag, 1999: 217-235.
- [16] ROSSETTO M R, et al. Benzylglucosinolate, benzyliothiocyanate, and myrosinase activity in papaya fruit during development and ripening [J]. Agric Food Chem, 2008, 56 (20): 217-235.
- 作者简介: 李勇, 男, 1982年生, 博士, 助理研究员, 发表论文10余篇。
安新颖, 女, 1978年生, 博士, 副研究员, 发表论文10余篇。
赵迎光, 男, 硕士。
贾晓峰, 男, 1984年生, 博士, 助理研究员, 发表论文10余篇。
- 收稿日期: 2012-12-14