便于快速信息融合的主题检测算法

施侃晟[1]，刘海涛[1]，白英彩[1]，宋文涛[1]，周书勇[2]

(1. 上海交通大学电子与电气工程系 上海 徐汇区 200030; 2. 中国孵化中心 杭州 310053)

【摘要】物联网要求对海量信息源里的不同主题，自动地高性能地进行检测和融合。目前大多数公开报道的中文主题检测算法时间复杂度是非线性的，在海量多信息源的信息融合方面缺乏可行性。该文采用高效能的一元语法模型结合全文检索的方法降低主题间的比较次数，理论上将算法效率提升到线性。通过新华社实际数据的实验证实，算法的时间复杂度确实为线性的。另算法应用于两项云计算的实际产品中，也验证了算法适用于物联网环境下的高速信息融合。

关 键 词 全文检索； 主题检测； 一元语法模型； 向量空间模型

中图分类号 TP311 **文献标识码** A doi:10.3969/j.issn.1001-0548.2012.06.014

Chinese Topic Detection Algorithm for Fast Information Aggregation

SHI Kan-sheng[1], LIU Hai-tao[1], BAI Yin-cai[1], SONG Wen-tao[1], and ZHOU Shu-yong[2]

(1.College of Electronic and Electric Engineering, Shanghai Jiaotong University Xuhui Shanghai 200030;

2. China Incubating Center Hangzhou 310053)

Abstract The most salient features of Internet of Things (IoT) are its ubiquitous large scale information gathering and intelligent processing to meet everyone's needs. Current solutions in Chinese topic detection and clustering have high time complexity such as $O(n^2)$ or $O(n^3)$. This paper presents an efficient and patented algorithm for defining topic detection and information clustering over the Internet of Things by combining an improved unigram language model and full text retrieval technique to reduce the time complexity. The experiments and real world applications show that the new method possesses much lesser time complexity.

Key words full text retrieval; topic detection; unigram language model; VSM

信息社会世界峰会(WSIS)正式确定了“物联网”概念，并发布了题为《ITU Internet Reports 2005—the Internet of Things》的报告，报告涵盖了物联网特征及相关技术[1]。由于物联网具有明显的“智能性”的要求和特征，因此智能信息处理的相关关键技术和研究基础对于物联网的发展具有重要的作用。物联网泛在网络环境中，需要有适用于物联网特点的数据挖掘、预测、内容服务、智能信息处理、信息聚合和融合、相似度特征提取、分类器、匹配引擎和智能交互[2-5]，支撑技术[4-5]。自动主题检测[6-12]是信息聚类融合的重要元素。然而，多级的、多方面的来自传感网中多个信息源的海量信息的汇聚，对现有的主题检测技术是一个挑战。目前公开报道的中文主题检测算法[7-14]大部分不是线性时间复杂度的，无法胜任物联网环境下多信息源的海量信息采集。本文提出了检测效果较好前提下的线性时间复杂度的主题检测算法。

1 改进的主题检测算法

1) 改进思路。

以往的主题检测方法，都是将每则新来信息和已有的主题系列进行一对一的比对，根据与主题的相似性判断该则信息属于哪个主题，这样造成了系统开销不是线性的。本文提出为该则信息在已有主题系列中，先查询推荐出最可能的主题集；然后将该则信息与最可能的主题集进行相似性比对，不必与剩余的其他主题比对，从而提高了系统性能。

本文采用性能优越的一元语法模型[15-16]，该模型以词作为特征项，特征项权重为词在主题中出现的频数。

2) 主题表示定义为 $\vec{T}=(f_{T1}, f_{T2},\cdots, f_{Tn})$，其中 $f_{Tj}(1\leqslant j\leqslant n)$ 表示主题 $\vec{T}$ 的特征。

3) 后续信息表示定义为 $\vec{d}=(f_{d1}, f_{d2},\cdots, f_{dm})$，其中 $f_{di}(1\leqslant i\leqslant m)$ 表示信息 $\vec{d}$ 的特征。

收稿日期：2010－02－25； 修回日期：2012－09－06

基金项目：国家自然科学基金(61073150)

作者简介：施侃晟(1966－)，教授，主要从事信息挖掘和云计算方面的研究.

4) 主题$\bar{T}$与后续信息$\bar{d}$的相似度，按以下公式计算：

$$S(\bar{d},\bar{T})=\frac{1}{L_d}\sum_{w\in d}\mathrm{tf}(w,\bar{d})\lg\frac{\lambda P(w|\bar{T})+(1-\lambda)P(w)}{P(w)} \quad (1)$$

式中，$S(\bar{d},\bar{T})$为主题$\bar{T}$与后续信息$\bar{d}$的相似度；w为信息和主题的特征项；$\mathrm{tf}(w,\bar{d})$为词w在信息$\bar{d}$中出现的频数；L_d为信息$\bar{d}$中的总词数；λ为平滑系数(0, 1)，在TDT3语料上训练，使得跟踪开销最小的值；$P(w|\bar{T})$为在主题$\bar{T}$中出现的概率，由给定几个样本训练，定义为：

$$P(w|\bar{T})=\frac{C(w,\bar{T})}{Nw(\bar{T})} \quad (2)$$

式中，$C(w,\bar{T})$为词w在主题$\bar{T}$中出现的次数；$Nw(\bar{T})$是主题$\bar{T}$的总词数；$P(w)$是词在背景语料中的先验概率，在背景语料中统计：

$$P(w)=\frac{C(w,\mathrm{background})}{N(\mathrm{background})} \quad (3)$$

式中，$C(w,\mathrm{background})$为词w在背景语料中出现的次数；$N(\mathrm{background})$为背景语料的总词数。

5) 主题检测方式。

计算每个后续信息与主题的相似度，若$S(\bar{d},\bar{T})>\theta$，则相关；否则不相关。

本文减少多源信息获取相关性模型时比较的次数，及主题检测时相关性比较的次数。因为大多数与某一则信息相似的信息是在查询返回结果的前k项中，与该主题模型相关的信息只应在前k项中，所以比较项最多为k项，而忽略与其余项的比较。

本文模型采用以下公式作为相关性度量：

$$D(M_1\|M_2)=\sum_w P(w|M_1)\lg\frac{P(w|M_1)}{P(w|M_2)} \quad (4)$$

6) 算法具体过程。

① 对供给来的信息进行预处理，计算信息中各个词的权值，并按前面形式化表达的公式建立信息模型。

② 如果信息是信息流中的第一则，成立一个以该信息为种子的主题，按照以上内容建立主题模型。将信息和主题双双存储。由此增加了系统的存储量，但配对关系类似于对比表，信息可以压缩，所以增加的空间并不多。

③ 如果信息不是信息流中的第一则，则先查询，返回k项信息。采用全文检索的方式。用式(1)计算信息和已存在主题的相似度，记录最高相似度以及取得最高相似度的主题T。全文索引建立的时间复杂度基本与信息量成线性增长。

④ 如果最高相似度超过了预设的阈值，则表示该则信息和主题T相关，将信息加入主题T，并更新主题T的模型；否则成立一个以该信息为种子的新主题并建立该主题的模型。

⑤ 重复上述过程直到所有的供给来的信息处理完毕。

2　实验与分析

实验采用两批数据：学习信息310则(带标记)和演示信息50 628则(由新华社提供)。实验内容为回溯事件检测，采用单遍聚类，设定适当的阈值检出主题。聚类过程中，采用分块进驻内存的方式，通过检索确定每则信息所需比较的比较对象。

回溯事件检测分为以下步骤：1) 数据准备(切词)；2) 特征提取；3) 回溯事件检测。通过对测试数据300、600、900、3 000、5 000、10 000、20 000、30 000、40 000、50 000则信息进行10次检测，时间代价如图2所示。

从图2的检测结果可以看出算法的时间复杂度是线性的，与理论一致。

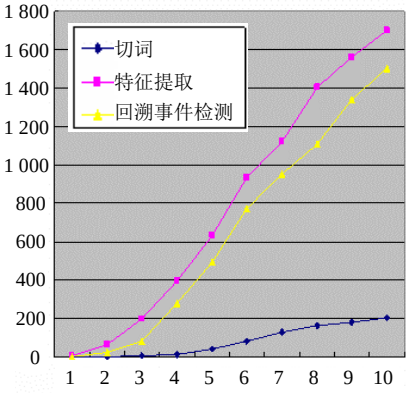

图2　体现线性性能的检测

本文为了克服时间代价问题，把多源头的信息切块，然后再合并，最终算法达到了线性的时间复杂度。

3　结　论

文献[17]提出基于物联网的“感知中国”国家战略。预示开始趋向物联网的智能化本质，迫切需要有本文相关的高性能技术自动地、智能化地进行基于用户指定主题的信息处理和信息融合，以体现物联网最终的使用价值。

目前中文主题检测算法大部分均不是线性时间复杂度的，而是$O(n^2)$和$O(n^3)$，甚至是指数级的。这样的算法在多源采集的大信息量下几乎无法适用。

本文提出的中文主题检测算法具有线性高性能并已成功应用到基于云计算的易合®系统，以及基于物联网的易智童®早期教育系统中，获得了国家发明专利[18]。

参 考 文 献

[1] International Telecommunication Union. Internet Reports 2005: The Internet of things[R]. Geneva: ITU, 2005.

[2] 刘强, 崔斌, 陈海明. 物联网关键技术与应用[J]. 计算机科学, 2010, 37(6): 1-5.
LIU Qiang, CUI Bin, CHEN Hai-ming. Key technologies and applications of internet of things[J]. Computer Science, 2010, 37(6): 1-5.

[3] 王桐, 赵春晖, 焉晓贞. 基于PML及Hedge的物联网异构信息集成处理模型[J]. 东南大学学报(自然科学版), 2011, 41(2): 301-304.
WANG tong, ZHAO Chun-hui, YAN Xiao-zhen. Heterogeneous information integrations from internet of things based on PML and Hedge automata[J]. Journal of Southeast University (Natural Science), 2011, 41(2): 301-304.

[4] 孙其博, 刘杰, 黎彝, 等. 物联网: 概念、架构与关键技术研究综述[J]. 北京邮电大学学报, 2010, 33(3): 1-9.
SUN Qi-bo, LIU Jie, LI Shan, et al. Internet of things: summarize on concepts, architecture and key technology problem[J]. Journal of Beijing University of Posts and Telecommunication, 2010, 33(3): 1-9.

[5] 陈志刚, 韩正君. 物联网信息聚合服务模式及运营商定位探析[C]//2010(第四届)移动互联网国际研讨会论文集. 北京: 中国移动通信集团公司, 2010: 22-24.
CHEN Zhi-gang, HAN Zheng-jun. Exploration of service model of information aggregation and operator of internet of things[C]// 2010 4th of China International Information and Telecommunication. Beijing: China Mobile, 2010: 22-24.

[6] 骆卫华, 刘群, 程学旗. 话题检测与跟踪技术的发展与研究[C]//全国第七届计算语言学联合学术会议(JSCL-2003)论文集. 北京: 清华大学出版社, 2003: 560-566.
LUO Wei-hua, LIU qun, CHEN Xue-qi. Development and analysis of technology of topic detection and tracking[C]// Proceeding of 7th National Joint Conference of Computational Language(JSCL-2003). Beijing: Tsinghua Press, 2003: 560-566.

[7] 洪宇, 张宇, 刘挺, 等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报, 2007, 21(6): 71-87.
HONG Yu, ZHANG Yu, LIU Ting, et al. Topic detection and tracking review[J]. Journal of Chinese Information Processing, 2007, 21(6): 71-87.

[8] HE Qi, CHANG Kui-yu, LIM Ee-peng, et al. Keep it simple with time: A reexamination of probabilistic topic detection models[C]//IEEE Transaction on Pattern Analysis and Machine Intelligence. 2010: 1795-1808.

[9] 刘嵩, 张先飞, 李弼程, 等. 基于概念相似度的话题自动检测方法[J]. 信息工程大学学报, 2010, 11(3): 303-306.
LIU song, ZHANG Xian-fei, LI Bi-cheng, et al. Automatic topic detection based on document concept similarity[J]. Journal of Information Engineering University, 2010, 11(3): 303-306.

[10] 张辉, 周敬民, 王亮, 等. 基于三维文档向量的自适应话题追踪器模型[J]. 中文信息学报, 2010, 24(5): 70-76.
ZHANG Hui, ZHOU Jing-min, WANG Liang, et al. An adaptive topic tracking model based on 3-Dimension document vector[J]. Journal of Chinese Information Processing, 2010, 24(5): 70-76.

[11] 张京阳, 张华平, 刘金刚. 基于聚团词的大规模文本转载识别算法[J]. 计算机应用, 2010, 30(6): 1661-1663, 1670.
ZHANG Jin-yang, Zhang Hua-pin, Liu Jin-gang. Large-scale document forward detection algorithm based on agglomerate-term[J]. Journal of Computer Applications, 2010, 30(6): 1661-1663, 1670.

[12] 赵华, 赵铁军, 张姝, 等. 基于内容分析的话题检测研究[J]. 哈尔滨工业大学学报, 2006, 38(10): 1740-1743.
ZHAO Hua, Zhao Tie-jun, ZHANG Shu, et al. Topic detection research based on content analysis[J]. Journal of Harbin Institute of Technology, 2006, .38(10): 1740-1743.

[13] 张阔, 李涓子, 吴刚, 等. 基于关键词元的话题内事件检测[J]. 计算机研究与发展, 2009, 46(2): 245-252.
ZHANG Kuo, LI Juan-zi, WU Gang, et al. Term-committee-based event identification within topics[J]. Jounal of Computer Research and Development, 2009, 46(2): 245-252.

[14] 王会珍, 朱靖波, 陈文亮, 等. 基于一元语法模型的中文话题追踪[C]// 第二届全国学生计算语言学研讨会. 2004: 422-427.
WANG Hui-zhen, ZHU Jin-bo, CHEN Wen-liang, et al. Unigram model based Chinese subject tracking[C]// Proceeding of 2nd National Conference of Computational Language. [S.l.]: [s.n.], 2004: 422-427.

[15] 王会珍, 朱靖波, 陈文亮, 等. 基于反馈学习自适应的中文话题追踪[J].中文信息学报, 2006, 20(3): 92-98.
WANG Hui-zhen, Zhu Jin-bo, CHEN Wen-liang, et al. Adaptive Chinese topic tracking based on feedback learning[J]. Journal of Chinese information Processing, 2006, 20(3): 92-98.

[16] 洪宇, 张宇, 范基礼, 等. 基于语义域语言模型的中文话题关联检测[J]. 软件学报, 2008, 19(9): 2265-2275.
HONG Yu, ZHANG Yu, FAN Ji-li, et al. Chinese topic link detection based on semantic domain language model[J]. Journal of Software, 2008, 19(9): 2245-2275.

[17 温家宝. 2010年政府工作报告[EB/OL]. (2010-03-15). http://www.gov.cn/2010lh／content_ 1555767.htm
WEN Jia-bao. Review of work in 2009[EB/OL]. (2010-03-15). http://www.gov.cn/2010lh／content_1555767.htm

[18] 施章祖, 施侃晟. 计算机辅助报告与知识库产生的方法: 中国, ZL200810063295.1 [P]. 2008-10-18.
SHI Zhang-zu, SHI Kan-sheng. Computer aided method of generating report and knowledgebase China, ZL200810063295.1 [P]. 2008-10-18.

编 辑 张 俊