基于话题模型的科技文献话题发现和趋势分析

贺 亮，李 芳

上海交通大学 计算机科学与工程系，中德语言技术联合实验室，上海 200240

E-mail: mazailiang@sjtu.edu.cn

摘 要: 自动挖掘科技文献话题，总结研究领域的发展趋势及最新研究动态，能给科技工作者的研究工作提供帮助。本文提出一种话题发现和趋势分析的方法，该方法首先利用 LDA 话题模型抽取科技文献的话题，然后计算话题的强度和影响力，最后研究话题的趋势变化。本文提出了可以针对任何文集的话题强度和影响力的计算方法，对热门和冷门话题以及影响力高和影响力低的话题分别进行了趋势分析。对 ACL 论文集进行实验，结果显示了计算语言学领域的一些发展状况且验证了话题强度和影响力的计算方法是可行的。

关键词: 话题模型；趋势分析

Topic Discovery and Trend Analysis in Scientific Literature Based on Topic Model

He Liang, Li Fang

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240

E-mail: mazailiang@sjtu.edu.cn

Abstract: Automatically extracting effective information from scientific literature and sum up the research trends and shifts, which can provide a great convenience for scientists. In this paper, we use LDA model to generate topics from the scientific literature; then calculate the strength and impact of the topic; finally, find the trends of topics. The method is suitable for any documents to calculate topic strength and topic impact. The trends of hot topics , cold topics, high impact topics and low impact topics are analyzed. The experiments on ACL anthology have shown the trend of computational linguistics and also proved the proposed calculating method.

Keywords: topic model; trend analysis

1 引言

在这个信息爆炸的时代，科学技术的发展也日新月异，对于科技工作者来说，如何快速的获取相关领域的最新研究动态，是一个值得关注的问题。为了了解最新的研究工作，科技工作者会关注该领域的关键问题，这些问题都用到了什么样的技术，在众多的技术中，哪些是目前的研究热点，哪些逐渐被人们淡忘。因此，对于科学技术趋势的自动分析研究，旨在帮助科学工作者从大量的学术会议和科技文献中提取出有用的信息，具有重要的现实意义。

要进行趋势分析，首先需要从大量的语料集合中提取出潜在的语义信息。Blei[1]提出的 LDA 模型可以挖掘大规模语料的语义信息，是机器学习、信息检索等领域很流行的一个模型。因此，目前国内外相关研究中，对文本语义信息进行挖掘通常都使用 LDA 或其扩展的话题模型。

本文的主要工作包括两个方面。首先，利用话题模型即 LDA 模型对语料建模，挖掘出该领域中的研究热点及相关技术，并对话题的热门程度和影响力进行研究；然后，研究这些子领域以及技术在整个时间段上的趋势变化。

本文的组织结构如下：第二部分介绍相关的工作，第三部分是研究方法的描述，第四部分是实验结果和分析，第五部分为结论及展望。

2 相关工作

对于科技文献的研究，主要利用了科技文献的作者、文本信息、引用信息和时间信息，去进行话题的发现和趋势的分析工作。

首先，话题的发现，即是挖掘文献中的隐含的语义信息。目前主要有两类方法可以发掘话题。第一类利用话题模型进行话题发现，这里话题（topic）的定义是一组词的概率分布。根据文集的文本信息可以利用 LDA 以及其拓展模型（CTM、DTM 等）进行建模[2,3]，发现话题；如果结合作者信息，有作者话题模型（ATM）以及其拓展模型（ACT、TATM 等）[4-6]，通过对该模型的推导可以得到每个作者在话题空间上的分布，通过分析该分布就可以了解在某一特定领域（话题）都有哪些专家，以及这些专家关注的研究领域（话题）是什么；结合文献引用信息，既考虑到了文献间引用关系对生成过程中的影响，有继承话题模型（ITM）[7]。第二类方法则通过构造网络图，利用文献的文本信息以及文献间的引用信息进行话题发现。有学者使用词组（term）来表示话题，然后利用词组（term）在文集中的分布关系并结合文集之间的引用关系发现话题[8]。

从文集中发掘出话题信息后，就可以在这话题空间上进一步分析这些话题的特点。有学者利用 LDA 对文集建模得到的话题空间，再加入文献之间引用的信息，去研究话题的特性。这些特性有话题的影响因子，用于衡量话题对文档的影响；有话题的影响多样性，衡量话题的影响范围；有话题的年龄，衡量话题的新旧程度；还有话题的转移度，衡量话题之间相互的影响[9]。

更进一步，加入时间的信息，进行话题的趋势分析。有学者利用话题的后验概率去定义话题的强度，通过计算每个时间点上的强度得到其强度的趋势变化[10,11]，对这些话题的趋势变化进行分析，以获得科技发展的一些特点，例如一些技术的应用走向，是偏向理论性的研究还是偏向于实际应用等[11]。有学者使用分时间段进行话题建模，考虑各个时间段话题之间关联的方法，可以从内容上去分析话题的变化趋势[3,7,12]。有学者在作者话题模型的基础上，加入时间信息，利用话题与作者间对应关系，从而可以分析这些作者的研究兴趣如何随时间推移而变化[6]。

为了提出一种方法能够针对任何文集，例如新闻报道[12]，数字文献等，我们只考虑文献的时间和文本信息，忽略作者和引用信息。采用 LDA 话题模型，找到潜在话题，借鉴文献[9-11]对话题的强度和影响力这两个特性进行研究，提出了不同的计算公式，通过这两个特性的分析可以找到热点话题和有影响力的话题，然后根据话题的强度再对它们进行趋势分析。

3 研究方法

首先对文本集合应用 LDA 建模，挖掘潜在话题，然后，研究话题的强度和影响力，最后对热点话题和有影响力话题进行趋势分析。话题强度主要描述了话题的关注度，也就是说，讨论某话题的文章数越多，就说明该话题的强度越高，可以认为是热门话题。话题的影响力则是从影响的范围，即话题的多样性去衡量，如果一个话题对多个话题都有一定程度的影响，该话题可以认为是有影响力的话题。

3.1 话题建模

LDA 模型是一个生成概率模型，是三层的变参数层次贝叶斯模型，首先假设词由话题的概率分布混合产生，而每个话题是在词汇表上的一个多项式分布；其次假设文档是潜在话题的概率分布的混合；最后针对每个文档从 Dirichlet 分布中抽样产生该文档包含的话题比例，结合话题和词的概率分布生成该文档中的每一个词汇。

对文集进行整个时间段上的建模，根据 LDA 话题抽取的结果，定义话题的支持文档如下：假

设某一文档 d 中有至少 10%的词是由话题 z 生成的，那么该文档是话题 z 的支持文档。根据该定义，一篇文档可以支持多个话题。表 1 列出了本文使用的符号。

表 1　文中使用到的符号

符号	符号的描述
α	LDA 模型的 *Dirichlet* 先验参数，表示文档-话题分布的先验
β	LDA 模型的 *Dirichlet* 先验参数，表示话题-词分布的先验
K	话题个数
θ_d	文档 d 的话题多项式分布
D^t	t 时间内所有的文档集合
D_z^t	t 时间内话题 z 的支持文档集合
$S(z,t)$	话题 z 在 t 时间段的文档支持率
$I(z)$	话题 z 的影响力

3.2　话题强度计算

要从文集中找到热门的技术或研究领域，需要比较话题的强度。在这里使用话题的文档支持率作为话题的强度。时间间隔 t 的话题 z 的文档支持率计算公式如下：

$$S(z,t)=\frac{|D_z^t|}{|D^t|} \tag{1}$$

其中，分子表示 t 时间段话题 z 的支持文档个数，分母表示该时间段文档的总数。

3.3　话题影响力计算

要衡量一个话题的影响范围，可以使用其影响的多样性（Impact Diversity），作为它的影响力衡量标准。我们基于这样的假设，一个话题在某一时间段产生之后，可能会对之后时间段的话题有影响，这种影响将通过文档之间的关联来体现，如果前一时间段 t 话题 z 的支持文档 d 与后一时间段 t'话题 z'的支持文档 d'是关联的，那么可以认为话题 z 对话题 z'有一定的影响作用。

计算影响力时需要统计属于不同话题的文章之间的关联数量。每篇文章可表示为在话题空间上的分布，通过计算话题空间上分布的 JS 距离（Jensen-Shannon divergence）来判断文章之间是否关联。假设时间段 t 话题 z 的支持文档 d 与后一时间段 t'话题 z'的支持文档 d'，在话题空间中的分布分别为 θ_d 和 $\theta_{d'}$，则它们的 JS 距离计算公式如下：

$$D_{JS}(\theta_d \| \theta_{d'})=\frac{1}{2}(D_{KL}(\theta_d \| \theta_m)+D_{KL}(\theta_{d'} \| \theta_m)) \tag{2}$$

其中 $\theta_m=\frac{1}{2}(\theta_d+\theta_{d'})$

定义话题 z 对话题 z'的影响程度，计算公式如下：

$$P_z(z')=\frac{\Sigma_t\Sigma_{t'>t}\#relations\ between\ D_{z'}^{t'}\ and\ D_z^t}{\Sigma_t\Sigma_{t'>t}\#relations\ between\ D^{t'}\ and\ D_z^t} \tag{3}$$

其中，分子表示话题 z 的支持文档与后续话题 z'的支持文档关联数量，分母表示话题 z 的支持文档与后续文档关联数量。

话题 z 的影响力定义为话题 z 对所有话题的影响程度的熵，计算公式如下：

$$I(z)=H(P_z)=-\Sigma_{z'}P_z(z')\log P_z(z') \tag{4}$$

4 实验结果与分析

本文主要针对计算语言学领域的发展趋势进行研究。ACL 论文集（ACL Anthology）[1]，包括了所有计算语言学主流的会议论文，可以说涵盖了该领域的主流的研究内容和技术。因此，我们将 ACL 选集作为研究的语料数据集。

实验主要包括三个方面，一是找到热门的话题，二是研究话题的影响力，三是研究它们随时间变化的趋势。

本实验利用 Gibbs Sampling 方法进行参数的推理。实验使用了开源的 Gibbs Sampling 工具[2]，模型参数 α，β 分别设置为 50/K 和 0.01，话题个数 K 设为 100。

4.1 实验数据

实验数据是 1985 年至 2009 年的 ACL 论文集，该论文集包含了 ACL、COLING、EACL、EMNLP 等众多会议，总共 11072 篇文章。以上语料只取标题和摘要，并过滤停用词、低频词等。

4.2 热门话题分析

通过公式（1）计算话题每年的强度，比较话题的强度，可以发现每年的热门话题。表 2 展示了 2006 年至 2009 年每年最热门的五个话题，话题名称均为人工标签。

表 2　2006 年至 2009 年热门话题

2006		2007		2008		2009	
话题	强度	话题	强度	话题	强度	话题	强度
Stat. Parsing(72)	0.0627	Stat. MT(7)	0.1105	Stat. MT(7)	0.1111	Stat. MT(7)	0.1206
Stat. MT(7)	0.0617	Stat. Parsing(72)	0.0748	Sentiment(34)	0.0719	Stat. Parsing(72)	0.0622
Q&A System(77)	0.0552	WSD(78)	0.0614	Stat. Parsing(72)	0.0546	Sentiment(34)	0.0556
Named entity(28)	0.0543	Sentiment(34)	0.0480	Named entity(28)	0.0510	CRF(85)	0.0509
Info. Retrieval(42)	0.0496	Named entity(28)	0.0458	Summarization(14)	0.0455	Semantic Role(50)	0.0462

从表 2 可以看到，基于统计的机器翻译（Stat. MT）是近几年来最热门的话题。众所周知，自从统计技术在机器翻译领域取得成效后，人们对其的研究热情一直未减。统计技术也同样应用于计算语言学的其他方面，如基于统计的句法分析（Stat. Parsing），热门程度仅次于基于统计的机器翻译。值得一提的还有情感分析（Sentiment）在近年的研究热度迅速提升。

为了验证上述结果，我们参考了《2010 年 ACL 评述》[13]。结果显示，2009 年最热门的三个话题基于统计的机器翻译、基于统计的句法分析以及情感分析，在 2010 年 ACL 投稿论文数以及录用数都位于前列。主会议 646 篇投稿长文，上述三个话题的论文投稿数分别是 64、68 和 46 篇，在总共 164 篇录用长文中分别占了 15、16 和 9 篇。这说明该方法找到的热门话题是可信的。

4.3 话题影响力分析

文献[7]提出一种话题影响力的计算方法，它利用文档之间引用关系计算话题间影响概率，再计算这些影响概率的熵值，作为话题影响力。该方法将其作为 Baseline 与我们的方法进行对比。

首先使用公式（2）计算文档之间的关联度，阈值定为 0.07，然后，利用公式（3）、（4）计算话题的影响力。表 3 分别列出了影响力前五和后五的话题。

[1] http://www.aclweb.org/anthology/

[2] http://gibbslda.sourceforge.net/

表3 话题影响力得分情况

我们的方法		Baseline	
话题	影响力	话题	影响力
Kernel Method(52)	5.17	SVM(22)	7.00
Probabilistic Model(21)	5.14	Tagging(5)	6.99
SVM(22)	5.09	Probabilistic Model(21)	6.74
Sentiment(34)	5.08	Sentiment(34)	6.73
Ngram Model(27)	4.99	Kernel Method(52)	6.70
Stat. MT(7)	4.26	WSD(78)	5.94
Spell Correction(46)	4.24	Stat. MT(7)	5.86
Speech Recognition(25)	4.12	Summarization(14)	5.68
WSD(78)	3.40	Word Segmentation(68)	5.56
Word Segmentation(68)	3.05	Morphology(62)	5.48

结果显示了影响力高的话题都是一些使用比较广泛的技术，例如核方法（Kernel Method）、支持向量机（SVM）等在数据挖掘、机器学习领域很流行的分类技术，它们在计算语言学领域也发挥着很大的作用。而影响力最小的话题都是一些偏应用方面的领域，比如说机器翻译、词义消歧（WSD）以及分词（Word Segmentation）等，这些领域的特点是比较专一，影响面比较窄。

实验结果与 Baseline 方法的结果大体一致，虽然我们的方法计算量会比 Baseline 的要大，这是因为我们的方法需要计算文档之间的距离来判断它们是否关联，但是我们的方法不需要额外的文档之间相互引用的信息，可以应用任何文档集合。

4.4 话题趋势分析

本小节的实验是利用话题逐年的强度变化来分析话题的变化趋势，这些话题包括热门话题，冷门话题，影响力大的以及影响力小的话题，以此了解计算语言学领域近二十多年发展情况。

首先来看最近几年的热门话题的强度变化趋势。从图 1 可以看出基于统计方法的机器翻译技术作为最热门的话题从 1999 年开始，进入了一个飞跃上升的阶段。出现这个变化的原因，就是在 1999 年出现了一个机器翻译的热潮，其最主要的特征是基于统计的方法在这一领域开始占据主导地位，机器翻译的质量出现了一个跨越式的提高。这股热潮持续至今，仍未现衰减之势。同时，基于统计的句法分析的强度也随着这股热潮不断提升。而情感分析在 2000 年前一直都是比较冷门的话题，但现今研究者对它的青睐不断增加。

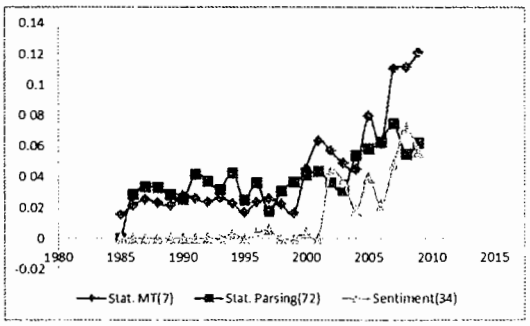

图1 热门话题强度变化趋势

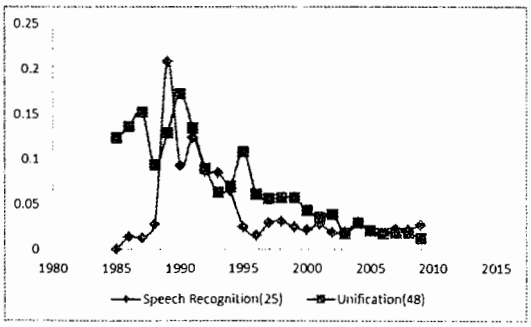

图2 冷门话题强度变化趋势

根据实验结果，图 2 列出了一些冷门技术的变化趋势，包括语言识别（Speech Recognition）和联并方法（Unification）。

联并方法是八十年代末九十年代初的研究热点，然后渐渐的淡出了研究者的视线。而语音识别技术的变化趋势比较奇特，它在 1989 年至 1994 年有一个爆发式的高峰。究其缘由，是因为这几年举办的 DARPA 语音及自然语言研讨会（DARPA Speech and Natural Language Workshop），这些研讨会产生了大量这方面技术的研究论文，而之后该技术的研究就进入低谷。

通过对热门话题和冷门话题的趋势分析，可以看到统计技术的兴起对这些热门话题的强度上升起了很大的推动作用；另一方面，冷门话题的下降趋势也有不同的表现形式，有的是缓慢下降，有的是急速下降。

接下来看影响力比较高的话题变化趋势情况，见图 3。

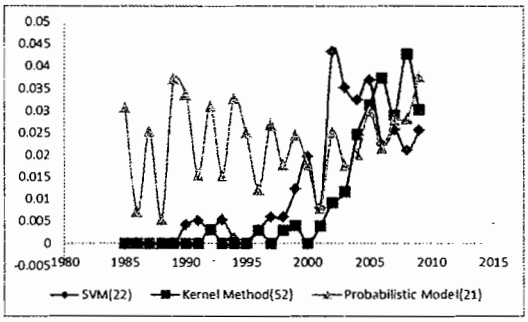

图 3　影响力高的话题强度变化趋势

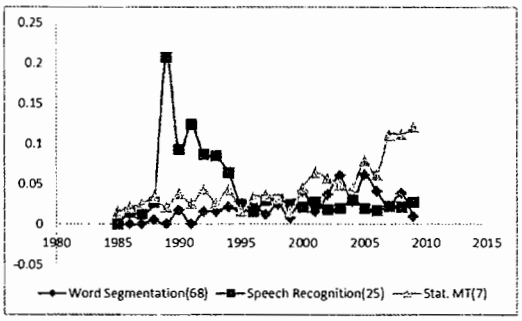

图 4　影响力低的话题强度变化趋势

这几个话题都是一些流行的技术，首先是概率模型（Probabilistic Model），它在计算语言学的领域一直都是比较主流的技术，它的强度变动在 2000 年前呈波动形式，之后呈上升趋势。而支持向量机和核方法在九十年代末开始兴起，此后也越来越受到研究者重视，保持着上升的形式，成为了计算语言学领域中比较重要的分析方法。

而影响力较低的话题，即比较偏应用的话题，它们的趋势变化也没有固定特点，从图 4 可以看到，有的呈现上升趋势，例如基于统计的机器翻译；有的呈现下降趋势，例如语音识别。

通过对影响力大的和影响力小的话题进行趋势分析，可以发现它们的强度变化趋势与影响力大小是无关的，这也说明了话题强度和话题影响力这两个指标是相互独立的两个标准，可以从不同方面去描述话题的特性。

5　结论与展望

本文利用话题模型的方法去对科技文献进行建模分析。首先使用 LDA 话题建模，发现文集中隐含的话题。接着，使用两个指标——话题强度和话题影响力去研究话题的特性。同时，对这些研究领域或技术受关注程度随时间变化的趋势进行分析，发现它们的变化特点。

通过分析实验结果，可以发现利用话题模型能够从大量文献中发掘出有意义的信息。使用我们的方法对话题特征以及趋势的分析得到的结果也是与实际情况相符合的，说明我们的方法对科技文献的分析是行之有效的。以下是对 ACL 论文集分析研究得到的一些结论。

首先，通过对热门话题的发现，可以得到最近比较热门的研究领域包括机器翻译、句法分析以及情感分析等；其次，通过话题影响力的分析，可以发现理论型的技术往往有较大的影响范围，可能会应用到多个子领域，而应用型的研究领域的影响范围比较窄；最后，通过趋势分析，可以了解计算语言学近二十多年来的发展情况，包括统计技术的流行大大促进了机器翻译和句法分析的研究，语音识别技术的研究热潮兴起与回落，联并语法研究的逐步衰落等。

今后的工作将考虑如何进一步挖掘话题的特点，更好的探索话题之间的关联。另外，从更多的角度去分析话题的变化趋势，比如从内容上分析话题在各个时间段的特点。

参 考 文 献

[1] D. M. Blei, A. Y. Ng, and M.I.Jordan. Latent Dirichlet Allocation. The Journal of Machine Learning Research, 2003, vol.3, pp.993-1022.

[2] D. M. Blei, J. D. Lafferty. A Correlated Topic Model of Science. The Annals of Applied Statistics 2007, Vol.1, No.1, 17-35.

[3] D. M. Blei and J. D. Lafferty. Dynamic Topic Model. In International conference on Machine Learning, 2006, pp.113-120.

[4] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The Author-Topic Model for Authors and Documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 2004.

[5] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD). 2008, pp.990-998.

[6] A. Daud, Juanzi Li, Lizhu Zhou, and F. Muhammad, Exploiting Temporal Authors Interests via Temporal-Author-Topic Modeling. Proceedings of ADMA, 2009, LNAI 5687, pp.435-443.

[7] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra and C. L. Ciles. Detecting Topic Evolution in Scientific Literature: How Can Citations Help. Proceeding of CIKM, 2009, pp.957-966.

[8] Y. Jo, C. Lagoze, C. L. Giles. Detecting Research Topics via the Correlation between Graphs and Texts. Proceedings of KDD, 2007. pp.370-379.

[9] G. S. Mann, D. Mimno, and A. McCallum. Bibliometric Impact Measures Leveraging Topic Analysis. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, 2006.

[10] T. L. Griffiths and M.Steyvers. Finding Scientific Topics. Proceeding of the National Academy of Science, 2004, pp.5228-5235.

[11] D. Hall, D. Jurafsky, and C. D. Manning. Studying the History of Ideas Using Topic Models. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp.363-371.

[12] 楚克明, 李芳. 基于 LDA 话题关联的话题演化. 上海交通大学学报, 2010, 44(11): 1501-1506.

[13] 吴甜, 田浩, 吴华, 王海峰. ACL2010 评述. 中文信息学报, 2010, 24(5): 127-128.