

doi:10.3772/j.issn.2095-915x.2015.06.008

主题模型可视化研究综述

孙国超, 徐硕, 乔晓东

(中国科学技术信息研究所 北京 100038)

摘要: 在大数据背景下, 以 LDA 为代表的主题模型数据挖掘技术得到了飞速的发展。随着研究的深入和科技工作者对结果可理解要求的提高, 主题模型的可视化也成为了研究的热点。本文将主题可视化分为两类: 基于文档集内容的主题模型可视化和融合外部特征的主题模型可视化, 本文在前人的基础上, 总结了上述两类的主题模型可视化, 分析了各个可视化工具的优缺点, 并对其进行了客观的评价。也对以后的主题模型可视化提出了一些建议并进行了展望。

关键字: 主题模型 LDA 模型词项可视化

Review on Visualization of Topic Models

SUN Guochao, XU Shuo, QIAO Xiaodong

(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: In the big data background, LDA topic model as one of the most popular technology of data mining has been rapid developed. With further research and scientists to improve the results understandable requirements, visualization of topic model has become a hot research. This article will divided the visualization of topic model into two categories: Visualization of topic model based on document and integration of external characteristics relating to topic model visualization, this article on the basis predecessors, summed up topic model above two types of visualization, analysis of the various visualization tools advantages and disadvantages, and its objective evaluation. Also give some suggestions and prospects about the future visualization of topic model relating to the future model visualization.

key words: Topic model, latent dirichlet allocation token visualization

作者简介: 孙国超, 男 (1991.6-), 硕士研究生, 主要研究方向: 数据挖掘、数据可视化等。电话: 18811506911, E-mail: sungc2014@istic.ac.cn。徐硕, 男, (1979-), 博士, 通讯作者, 副研究员, 主要研究方向: 数据挖掘, 机器学习和知识工程等。E-mail: xushu@istic.ac.cn。乔晓东, 男, (1965-), 英国谢菲尔德大学硕士, 研究员, 主要研究方向: 信息服务, 信息资源管理等。
基金项目: 国家自然科学基金项目: 基于论文和专利资源的技术机会发现研究。ID: 71403255。

1 引言

伴随着互联网和信息技术的飞速发展,科技工作者需要处理的数据规模也成几何倍增长。现在的科技工作者常常需要分析和研究上千篇文献。如果快速、高效的研究和理解大规模的数据集已经成为工作者迫切需要解决的问题。在这样的背景下,数据挖掘和自然语言处理等技术得到了飞速的发展,从基于 TF-IDF 的词频统计模型到后来的主题模型,其中最基础也是应用最广泛的是基于 LDA 的主题模型。LDA 模型并不是简单的统计词频,它将文档看作是主题的多项式分布,而每个主题又是词项的多项式分布,模拟文档的生成过程,计算出文档中隐藏的主题组成和词项的主题标签。LDA 不仅降低了文档的维度,而且可以从语义上分析文档,有效地解决了文档中同义词和多义词的问题^[1]。然而,如何理解主题模型对文档集建模生成的主题和从全局上深刻把握文档集的主要内容,已经成为以 LDA 为代表的主题模型检测的重要问题。主题模型的可视化有效地解决了上述问题。主题模型的可视化用图形展示给工作者大规模文档集中文档中主题的组成,主题中所包含的词项,文档与文档、主题与主题的相似度等关系。一个好的可视化方法可以将主题模型从文档集中的提取的信息全面的展示出来,使工作者从全局上理解和研究文档集。

2 LDA 主题模型

早期的对大规模文档集的数据挖掘是基于词频的 TF-IDF 模型。TF-IDF 是用来评估一个词项对文档集的重要程度,词项的重要性与它在某篇文档中出现的次数成正相关,同时与他在文档集中出现的总次数成负相关。由于是基于词频统计,导致文档的维度过高,并且无法解决一次多义和一义多词。1988 年 Scott Deerwester 等随后

提出了 LSA (Latent Semantic Analyze) 隐语义索引^[2]。LSA 通过对词项 - 文档矩阵进行奇异值分解。LSA 大大降低了文档的维度,低维空间表示可以刻画同义词,即部分解决了同义词的问题。但 LSA 没能解决同义词的问题,也不是一种概率模型,没有对应的物理解释。1999 年 Halfmann 随后提出了 PLSA (probabilistic LSA)^[3]。PLSA 是 LSA 的概率形式,对其中的每个变量以及相应的概率分布和条件概率分布都有明确的物理解释。PLSA 在文档层面没有提供合适的概率模型,使 PLSA 模型随文档数的增加而变得越来越庞大。2002 年由 Blei 等提出了 LDA (Latent Dirichlet Allocation) 模型^[4]。LDA 不仅降低了文档的维度,而且从语义层面分析文档,解决了同义词和多义词的问题。

LDA 模型是一种由文档、主题、单词组成的三层贝叶斯生成模型。每篇文档由服从多项式分布的主题组成,每个主题由服从多项式分布的词项组成。LDA 模拟了文档的生成过程,如图 1。

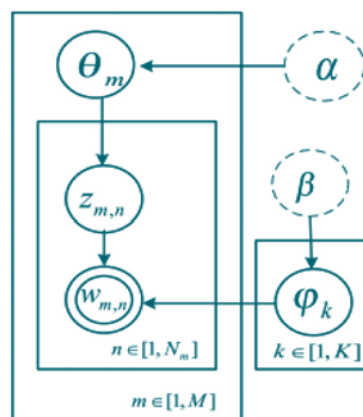


图 1 LDA 模型的概率图模型表示

α 是每个文档中主题多项分布的 Dirichlet 先验参数, θ_m 是第 m 篇文档的主题分布, $z_{m,n}$ 是第 m 个文档中第 n 个词的主题, β 是每个主题中单词多项分布的 Dirichlet 先验参数, ϕ_k 是第 k 个主题的词项分布, $w_{m,n}$ 是第 m 篇文档第 n 个单词。

生成过程如下: 对于每篇文档, ①从泊松分

布中取一个值作为文档的长度。②从参数为 α 的 Dirichlet 先验分布中取出一个主题多项式分布 θ_m , 表示文档中出现某主题的概率。③从主题的多项式主题中取出一个主题 $z_{m,n}$ 。④从参数为 β 的 Dirichlet 先验分布中取出一个词项多项式分布 ϕ_k , 表示对应主题的词项分布。⑤根据选出的主题, 从对应的词项分布中选出一个词作为文档的一个词项。重复以上过程直到把文档写完。其中 θ_m 和 ϕ_k 是主题模型中需要估计的参数^[5]。LDA 主题模型参数的近似估计主要有两类: 变分期望最大化 (Variational Expectation Maximization)、吉布斯取样 (Gibbs Sampling)。期望最大化方法的速度快, 但模型参数不够准确, 吉布斯取样模型参数准确, 但收敛比较慢。由于吉布斯取样方法描述简单, 比较容易实现, 成为了主题模型参数估计的常用方法。

3 主题模型的可视化

主题模型的可视化大都是基于 LDA 主题模型的可视化, 即基于文档的内容的主题模型可视化^[6]。根据是否融入文档外部属性可以将主题模型可视化分为两类: 基于文档内容主题可视化和融入外部属性的主题模型可视化^[7]。基于文档内容的模型可视化是研究整个文档集中文档、主题、词汇之间关系的可视化, 融入的外部属性包括文档的作者、时间属性等。融入作者属性的主

题模型可视化可以研究作者的兴趣和作者之间的关联^[8]。融入时间属性的主题模型可视化可以研究主题模型随时间的演变过程和趋势。

3.1 基于文档内容主题模型可视化

基于文档内容主题模型可视化是研究文档集中每个文档的主题分布, 每个主题的词项分布, 文档间的关系、主题间的关系, 并通过可视化将这些信息展现出来, 让科技工作者从宏观上快速了解和掌握大规模文档集的主题和主要思想。

虽然主题模型能够从语义上对文档进行表示, 但对大规模文档集基于 LDA 主题建模以后, 生成的主题由一组单词和该单词属于主题的概率组成, 然而, 对科技工作者来说, 理解由一个个单词组成的主题的主要思想并不容易。2009 年 David M. Blei 和 John D. Lafferty 提出 Turbo topic^[9]。其基本过程是: 利用 LDA 模型对文档集建模, 得到 k 个主题; 对已经处理好的文档中的单词标注其所属的主题, 如 $(w_1, z_1), (w_2, z_2)$, 形成单词主题对; 对其中的一个单词 w 和主题 z , 运行一个假设检验过程, 决定此单词前后的几个单词是否可以组成词组表示主题 z 。如图 2 Turbo topic 分词结果, 左侧已经标注好单词所属主题的文档, 右上侧 LDA 模型可视化结果, 右下侧是 turbo topic 可视化结果。大大提高了主题的可理解能力。但是 Turbo topic 的输出结果仍然以词或短语的排序来表示主题。

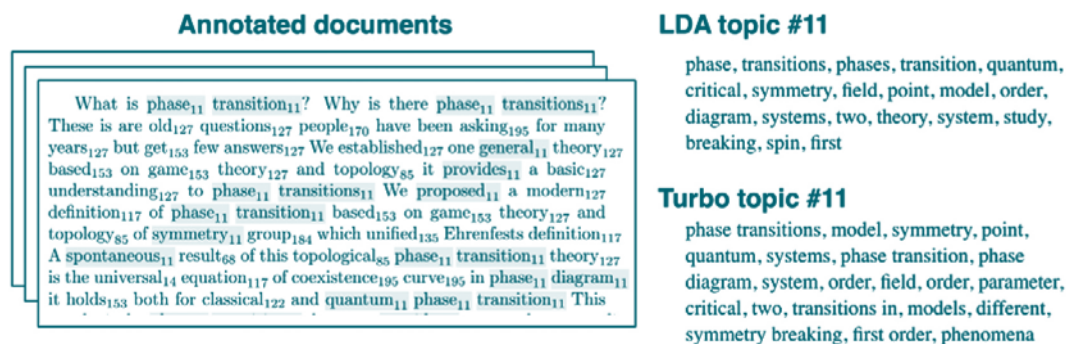


图 2 Turbo topic 分词结果

2012年由Chuang J, Manning C D, Heer J提出Termite^[10]用图形的方式来展示对大规模文档集主题建模的结果。如图3Termite主题可视化,横轴表示文档集中词汇表中所有词项,纵轴是经过主题模型建模后产生的主题编号,圆圈的大小表示词汇在相应主题中的权重。在选择表示主题的词汇时,Termite并不是单纯的取决于词汇在主题中的频率,而是增加了一个权重 $\text{distinctiveness}(w)$ 。w表

示单词。 $\text{distinctiveness}(w)$ 大小表示词汇对于主题的区别能力。包含此词汇的主题越多, $\text{distinctiveness}(w)$ 的值越小,此单词对主题的区别能力越差。反之, $\text{distinctiveness}(w)$ 越大,此单词对主题的区别能力就越大。此外,Termite还引入键能算法(the bond energy algorithm),把经常一起出现的词汇排列到一块。但是Termite缺乏一种对LDA模型的全局视角,更没有展示主题、文档间的联系。

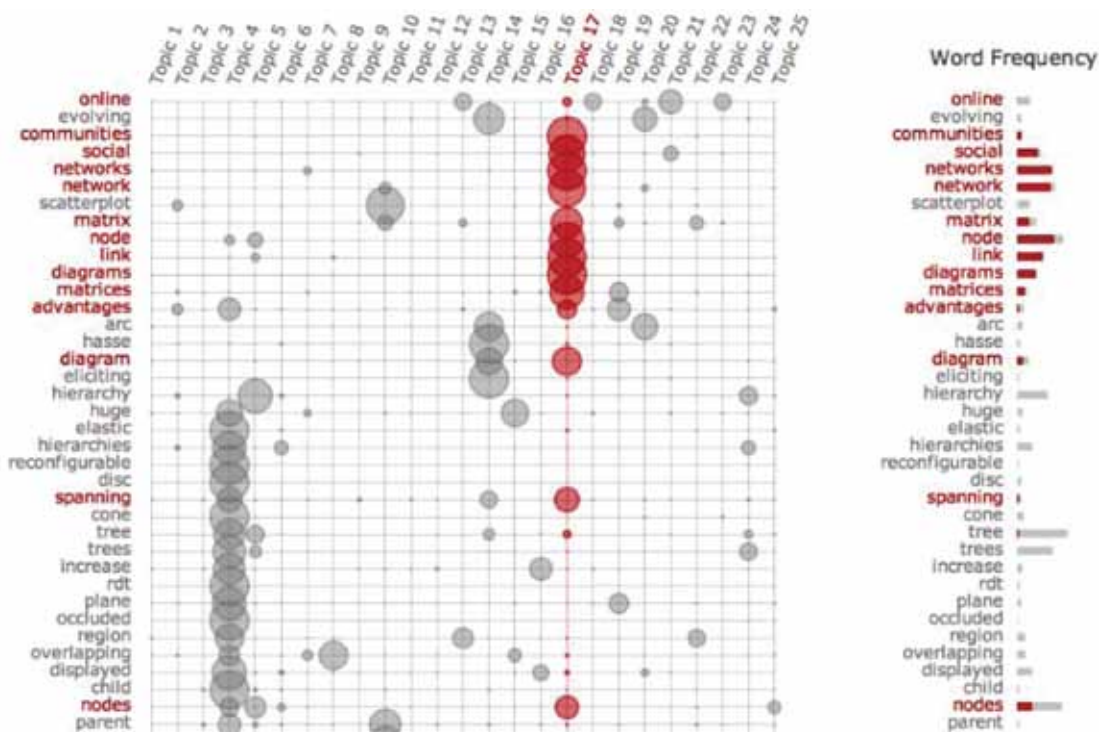


图3 Termite 主题可视化

Matthew J. Gardner、Joshua Lutes 等人提出了The Topic Browser^[11]。The Topic Browser除了展示了主题模型的输出结果和文档的信息外,还结合展示其他三方面的数据信息:①文档属性,如文档的作者,时间等。②主题指标,从简单的属于该主题的词频统计,到主题在文档集中的分布,还有主题间的相似度。③文档指标,和主题指标相似,包括词频统计、文档中的主题分布和文档相似度。如图4 topic browser,科技工作者可以

利用相关性过滤一些没意义的主题,利用文档熵(主题的文档分布)排序主题。当用户点击其中一个主题,进入该主题的详细介绍界面包括:①该主题的简单统计信息,如主题中包含词项的个数、主题的熵、词项的熵等。②该主题的两个词云。一个是按照属于该主题的概率大小排序,取前100个词做成词云,词项的大小由属于该主题的大小决定。另一个词云是经过由Blei和Lafferty提出的Turbo topics方法加权的词云。③列出了

该主题的相关主题。主题的相关性由主题的文档分布向量或主题的词项分布向量决定。④词项属于该主题的概率。⑤按概率大小选出前十个词项，取词项的前后 50 个字符在文档中显示出来。⑥按包含该主题词项的数量对文档排序，取前 10 篇文档作为和该主题最近似的文档。⑦设置文档的属

性，揭示主题中隐含的信息，让科技工作者更深刻的了解主题的内容。点击和该主题相近似的文档，进入文档的详细信息界面：①该文档的简单统计信息，如文档的熵，包含词项的数量。②各个主题在文档中所占的比率。③该文档的近似文档。④该文档的详细内容。

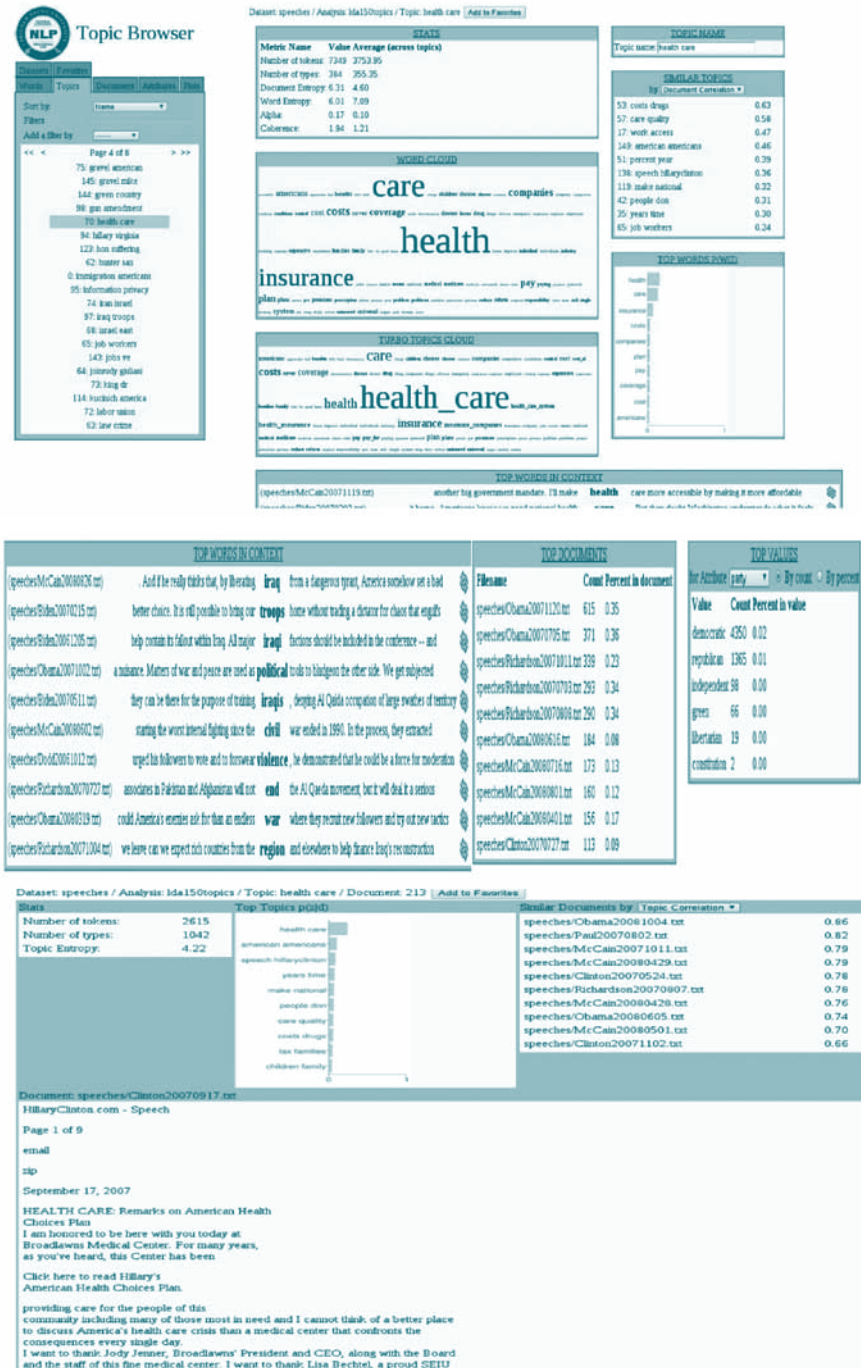


图 4 topic browser

2014 年 Carson Sievert 和 Kenneth E. Shirley 提出了 LDAvis^[12]：一种深刻理解和可视化主题的方法。LDAvis 结合了 R 语言和 D3，基于 LDA 主题模型对文档集主题建模，对主题进行可视化的方法。这种可视化方法对主题和主题之间的关系提供了一种全局的视角帮助科技工作者理解主题间的联系。另外 LDAvis 认为以词项属于主题的概率大小来选取表示主题的词项，并不是最优的结果，并提出了一种新的选取表示主题的词项的计算办法：词项和主题的相关度。按照词项和主题的相关度对属于主题的词项重新排序，并选取前

几个词项表示主题。如图 5 LDAvis 可视化，左图表示基于 LDA 模型的大规模文档集的主题建模结果。每一个圆圈表示一个主题，圆圈的大小表示包含该主题的文档的规模，圆圈的距离表示主题之间的相似度大小。科技工作者点击左图其中一个主题，右图显示的是该主题中排名前 n 的词项。灰色长条表示词项属于该主题的概率大小，红色长条表示词项和该主题的相关度大小。科技工作者还可以根据需要调整主题间相关度计算的方法（Jensen-Shannon divergence），每个主题展示的

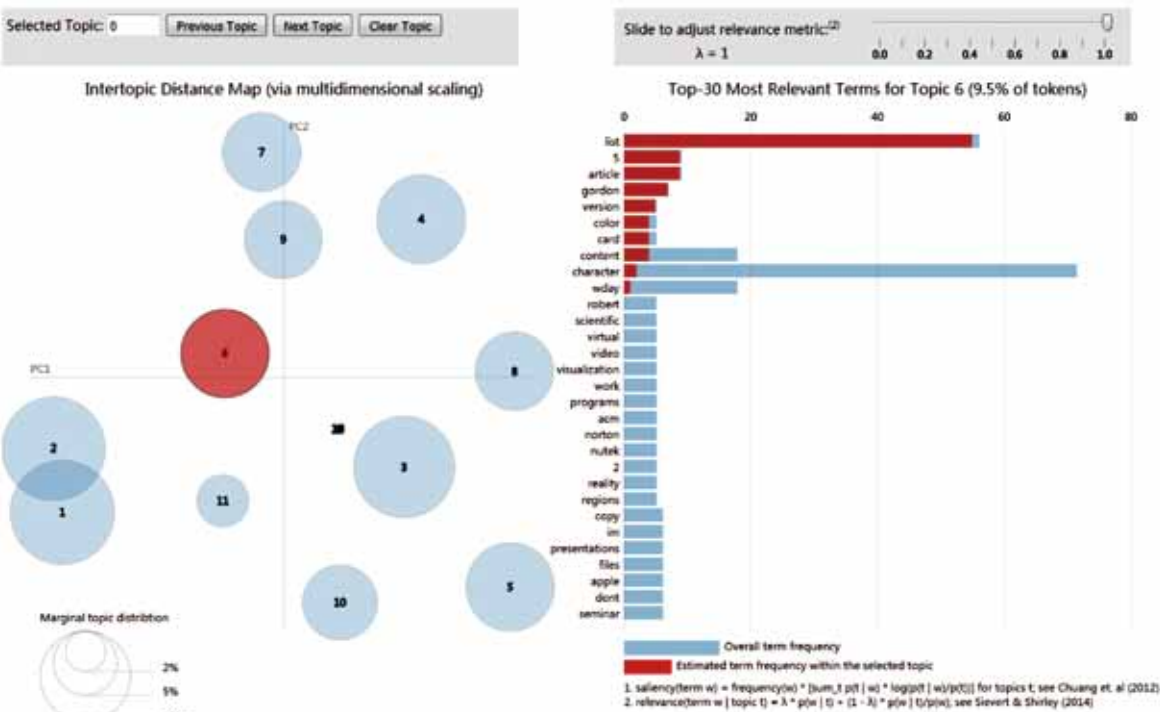


图 5 LDAvis 可视化

2012 年，Chaney A J B, Blei D M 提出了 VTM (Visualizing topic models)^[13]。VTM 不仅展示了主题 - 主题间的关系，主题 - 文档的关系，还展示了主题 - 词项间、词项 - 词项间的关系。如下图 6VTM 主题模型可视化，第一部分图是 VTM 基于 LDA 模型对大规模文档主题建模后产生的主题，每个长条代表一个主题；科技工作者

点击其中一个主题，就会进入进入该主题介绍页面：包括属于该主题的词项、包含该主题的文本、与该主题相关的主题。选择文本，就会进入文本的介绍页面：①文本的主题分布比例，以饼状图显示，点击饼状图上的主题，又进入相应的主题介绍页面。②文本的详细内容。③该文档的相似文档。每个长条代表一篇文本。

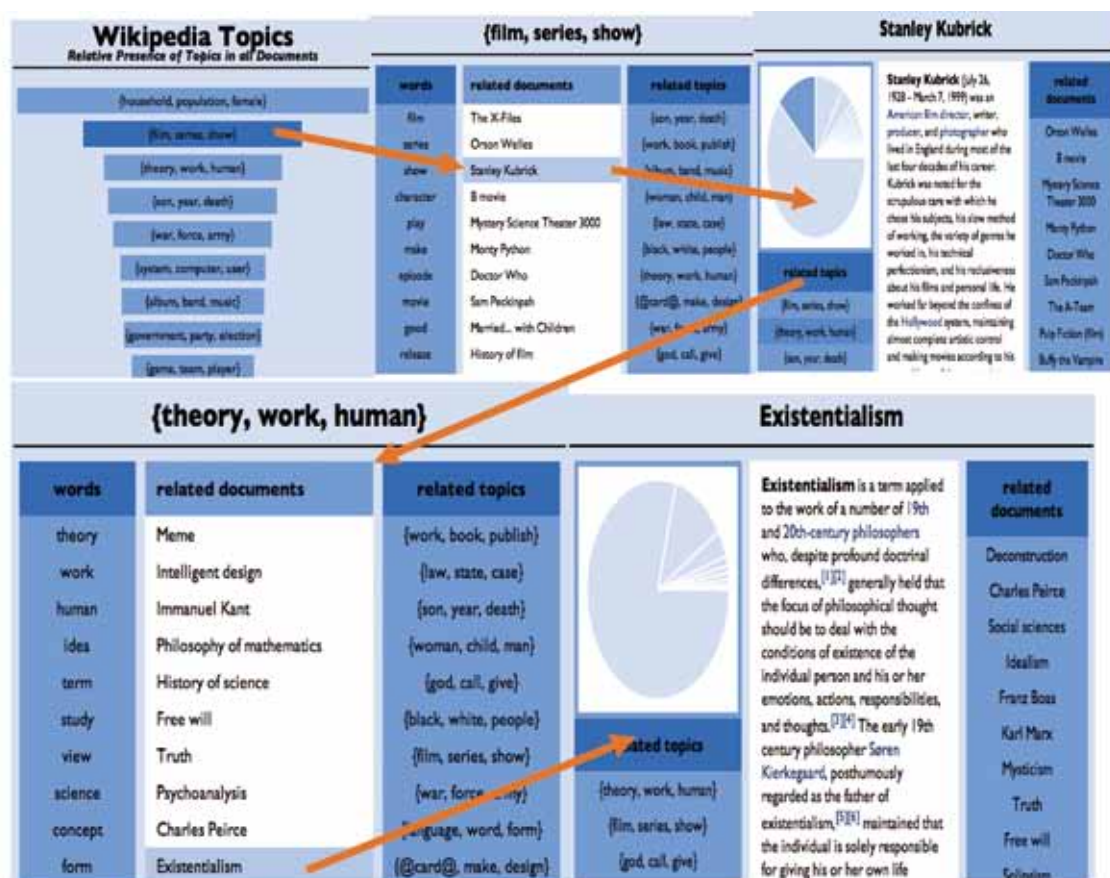


图 6 VTM 主题模型可视化

2015 年 Jaimie Murdock and Colin Allen 提出了 Topic Explorer^[14]。Topic Explorer 可以利用多窗口比较不同主题数目时，主题 - 文档、文档 - 文档间的区别。如图 7 Topic Explorer，用 LDA 模型对斯坦福哲学百科全书（SEP）的文章建模，主题数分别是 20、40。图中每一行代表一篇文献，

每一行的不同颜色代表文献的主题分布。每一行的宽度代表与选中文献的相似度（本图选中的是 Turing Machines）。将鼠标悬浮于主题上方，图中会显示该主题的前 10 个属于该主题的词项，并且在各行突出显示该主题。当点击某一主题，文档，即各行会根据该主题的权重重新排序。

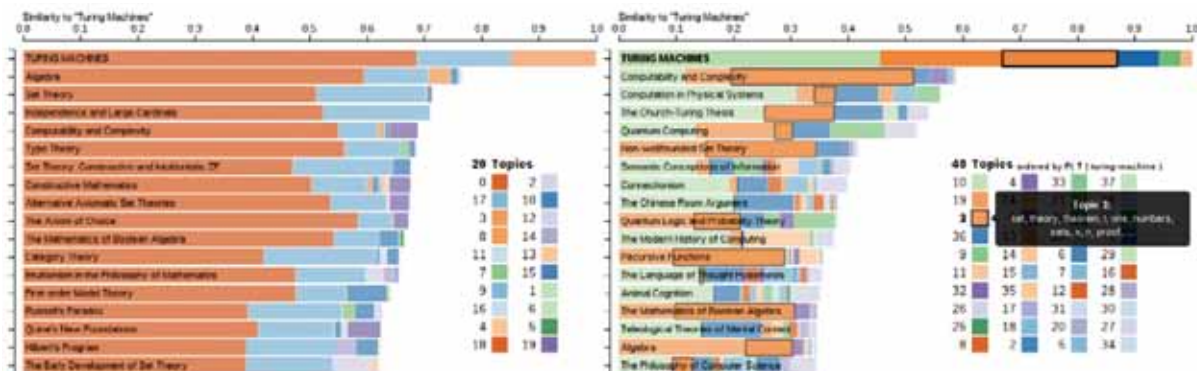


图 7 Topic Explorer

3.2 融合外部特征的主题模型可视化

文档的外部特征包含着重要的数据信息,是文档的重要组成部分。大多数文本都是随时间而变化的,不同时间段的主题也会发生变化。因此,时间属性是文档的一个重要属性^[15]。融入时间属性的主题模型可视化,在帮助科技工作者研究大规模文本内容的同时,也揭示文档主题随时间的演变过程 and 变化趋势。因此,许多学者将时间属性纳入 LDA 主题模型,并产生了不少研究性成果^[16]。

2002 年, Susan Havre 和 Lucy Nowell 提出了 Themeriver^[17]用来展示文档集的主题强度随时间的变化过程。Themeriver 在可视化上标出不同

时间发生的重大事件,帮助科技工作者分析主题变化和外部事件的因果关系。此外, Themeriver 还将 LDA 模型产生的主题分成不同的组,并用一种颜色家族表示不同的主题组,帮助科技工作者根据颜色区分和理解主题间的关系^[18]。如图 8 Themeriver 可视化。Themeriver 用河流的隐喻表示主题随时间的变化,上方文字是不同时间点发生的外部事件,下方是主题的时间轴,河流从左到右的流动代表文档集中的主题随时间的变化。中间河流每一条不同颜色的彩带代表一个主题。彩带的垂直宽度代表主题强度,即文档集中包含该主题的文档的数量。彩带随时间而变宽或变窄,表示主题强度随时间的增强或减弱。

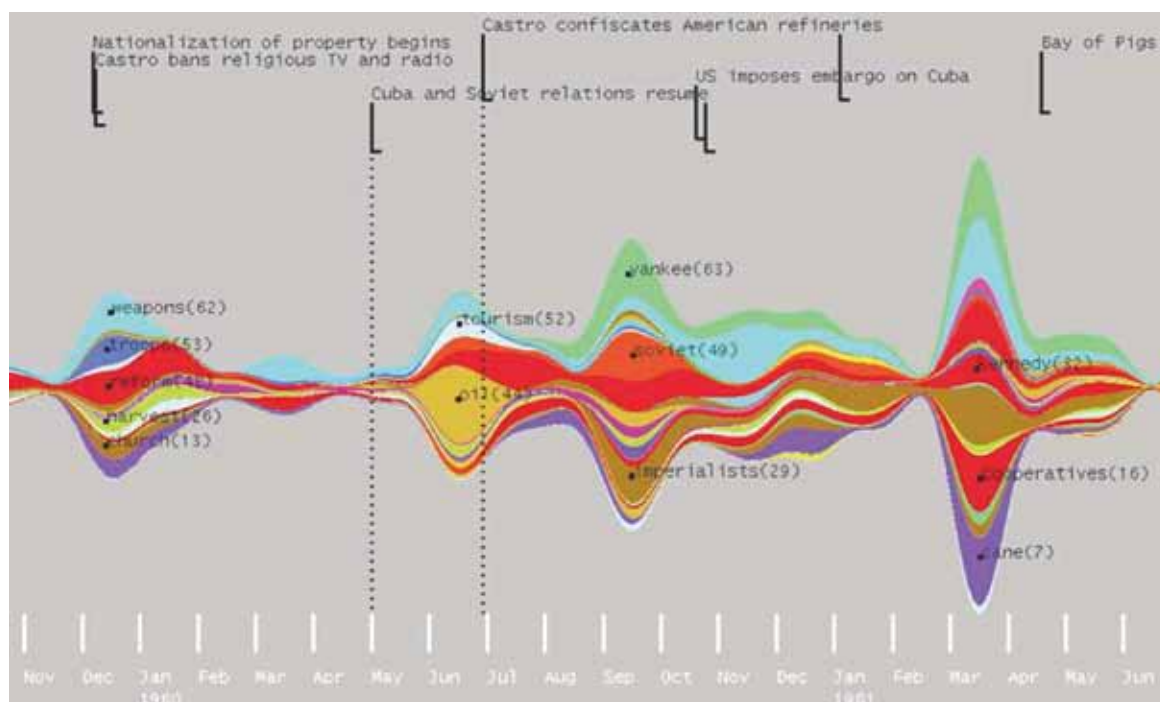


图 8 Themeriver 可视化

虽然 Themeriver 很好的展示了文档集的主题随时间的变化过程和演变趋势,但没有展示不同主题包含的词汇。而这些词汇可以帮助科技工作者理解主题的主题思想。2010 年, Wei F R, Liu S X, Song Y Q 提出了 TIARA (text insight via automated responsive analytics)^[19], TIARA 对 Themeriver 进行了扩展,在

Themeriver 中引入了标签云。如图 9 TIARA 可视化,用 LDA 对文档集进行主题建模后, TIARA 在不同的时间切片中选取若干个属于主题的词汇组成标签云,放在河流中相应的位置上。词汇的大小表示词汇在文档中出现的频率。TIARA 不仅展示了主题强度随时间的变化,还展示了主题语义的变化。

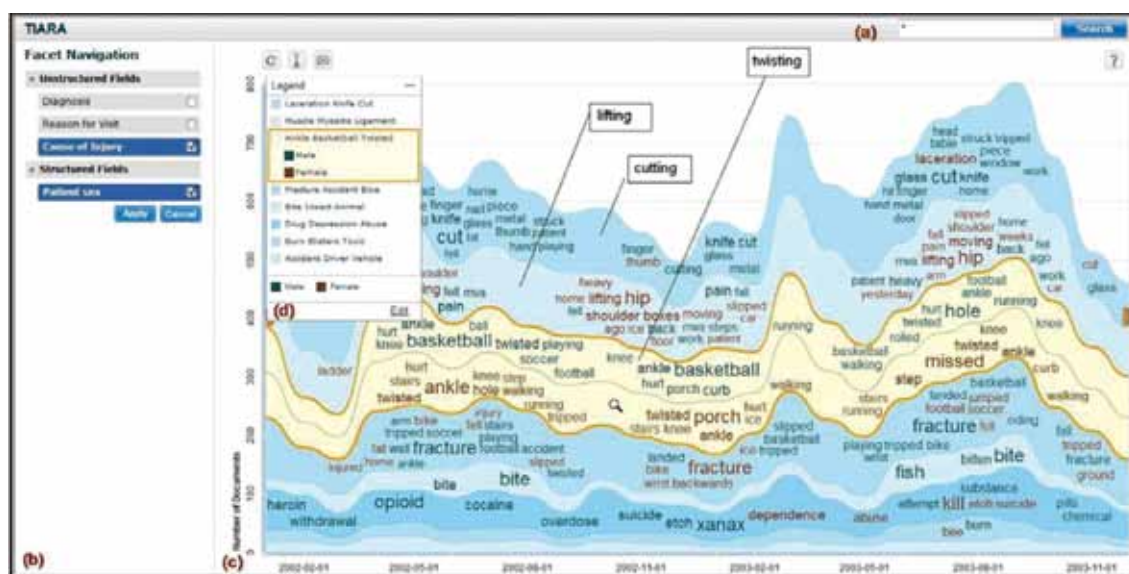


图 9 TIARA 可视化

2012 年, Alsakran J, Chen Y, Luo D N 提出了实时可视化文本流 RVST (real-time visualization stream text)^[20]。RVST 是模拟了力学系统的粒子运动来展示主题模型间及主题模型随时间的变化。用 LDA 模型对大规模文本进行主题建模, 得出每篇文档中包含的主题, 根据每篇文档中包含的主题计算文本间的相似度, 根据文本相似度计算出这个力学系统的总能量、粒子趋于平衡时的位置、

粒子的速度和粒子间的相互作用力。如图 10RVST 可视化, 每一个圆圈代表一个文档, 圆圈之间的距离代表文档间的相似度, 类之间的距离代表主题间的相似度。这些圆圈根据粒子间的作用力自动聚类, 每一类表示一个主题, 拖动位于下方的时间条可以查看每个聚类中新文档的出现和原文档的消失, 即主题随时间的演变趋势。科技工作者还可以在右侧查看主题中包含的词条等。



图 10 RVST 可视化

4 展望

主题模型作为一种能够从语义上对大规模文档集抽取主题模型,提供了一种帮助科技工作研究文档集主题的方法。中本文在详细介绍了LDA主题模型的基础上,分别综述了基于文档内容的主题模型可视化和融入外部特征的主题模型可视化^[21]。还分析了各个主题模型可视化的特点,并客观评价了它们的优缺点,为后续的可视化研究提供了依据。

通过对本文的综述研究可以发现,基于LDA主题模型的可视化已经非常成熟,优秀的可视化工具可以很好的展示文档、主题、词项三者间的关系^[22]。LDA模型没有考虑文档的外部特征,现在的可视化工具也仅仅只融入了时间属性,而其他的文档属性,如文档的作者,文档所属的项目等也包含着重要的信息。而这些重要信息并没有被很好的开发利用起来。而目前的主题模型大都是基于LDA主题模型的可视化。因此,本文认为,基于LDA模型的改进模型的可视化是未来面向科技情报分析领域的发展方向之一。不仅是融合单个外部属性的改进模型的可视化,融合多个外部属性的模型的可视化更具有广阔的前景^[23]。另外,随着数据更新的加快,工作者对信息的时效性要求的提高,用于在线动态更新的数据库信息的主題提取,很好的满足了工作者的需求^[24]。因此,在线实时更新的主题模型可视化也是未来发展方向之一。

参考文献

- [1] 徐戈,王厚峰.自然语言处理中主题模型的发展[J].计算机学报,2011(08):1423-1436.
- [2] 何锦群.LDA在信息检索中的应用研究[D]:天津:天津理工大学,2014.
- [3] 张小平,周雪忠,黄厚宽等.一种改进的LDA主题模型[J].北京交通大学学报,2010(02):111-114.

- [4] 邹晓辉,孙静.LDA主题模型[J].智能计算机与应用.2014,4(5):104-106.
- [5] 李勇,安新颖.基于LDA的主题演化研究[J].医学信息杂志,2013,34(2):57-61.
- [6] 施乾坤.基于LDA模型的文本主题挖掘和文本静态可视化的研究[D].广西:广西大学,2013:23-26.
- [7] 樊儒昆,傅晶,程司雷等.文本可视化研究综述.计算机辅助设计与图形学学报[J],2013,25(3):274-177.
- [8] 杨彦波,刘滨,祁明月.信息可视化研究综述[J].河北科技大学学报,2014,35(1):91-102.
- [9] David M. Blei, John D. Lafferty. Visualizing Topics with Multi-Word Expressions[EB/OL]. (2009-07-06)[2016-03-08]. <http://arxiv.org/abs/0907.1013v1>.
- [10] Chuang J, Manning C D, Heer J. Termite: visualization techniques for assessing textual topic models[C]//Proceedings of the International Working Conference on Advanced Visual Interfaces. New York: ACM Press, 2012: 74-77.
- [11] Matthew J. Gardner, Jeff Lund, Joshua Lutes, Josh Hansen, Dan Walker, Eric Ringger, Kevin Seppi. The Topic Browser: An Interactive Tool for Browsing Topic Models[C]. NIPS Workshop on Challenges of Data Visualization. USA: NIPS, 2010: 67-71.
- [12] Carson Sievert, Kenneth E. Shirley. LDAvis: A method for visualizing and interpreting topics[J]. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. 2014: 63-70.
- [13] Chaney A J B, Blei D M. Visualizing topic models[C]// Proceedings of the 6th International AAAI Conference on Weblogs and Social Media. Palo Alto: AAAI Press, 2012: 419-422.
- [14] Jaimie Murdock, Colin Allen. Visualization Techniques for Topic Model Checking[C]. Association for the Advancement of Artificial Intelligence. 2015.
- [15] 李湘东,张娇,袁满.基于LDA模型的主题演化研究[J].情报杂志,2014(07):115-121.
- [16] 崔凯.基于LDA的主题演化研究与实现[D].北京:

国防科学技术大学, 2010:36-39.

[17] Susan Havre, Lucy Nowell. ThemeRiver: Visualizing Thematic Changes in Large Document Collections[J], IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS. 2014, 8(1):9-19.

[18] 张龙飞, 姚中华, 宋汉辰等. 基于 ThemeRiver 的可视化技术发展综述 [J]. 系统仿真学报, 2013(09): 2091-2096.

[19] Furu Wei, Shixia Liu, Yangqiu Song, et al. TIARA: A Visual Exploratory Text Analytic System[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010:153-162.

[20] Alsakran J, Chen Y, Luo D N, et al. Real-time

visualization of streaming text with a force-based dynamic system[J]. IEEE Computer Graphics and Applications, 2012, 32(1): 34-45.

[21] 范云满, 马建霞. 基于 LDA 与新兴主题特征分析的新兴主题探测研究 [J]. 情报学报, 2014, 33(4):698-710.

[22] 汤斯亮, 程璐, 邵健等. 基于概率主题建模的新闻文本可视化综述 [J]. 计算机辅助设计与图形学学报, 2015 (05) : 771-782.

[23] 王晓光, 程齐凯. 基于 NEViewer 的学科主题演化可视化分析 [J]. 情报学报, 2013, 32 (9) : 900-911.

[24] 崔凯, 周斌, 贾焰等. 一种基于 LDA 的在线主题演化挖掘模型 [J]. 计算机科学, 2010 (11) : 156-159.