基于主题的文本可视分析研究

巫英才, 崔为炜, 宋阳秋, 陈 杨, 刘世霞
(微软亚洲研究院 北京 100190)
(yingcai. wu@microsoft. com)

摘 要: 文本可视分析是一个跨学科的领域,涉及文本数据挖掘、计算机图形图像以及人机交互等各方面的知识和技术,可以帮助用户以可视分析的手段交互地分析海量的文本数据内容,提供及时的反馈,发现异常和规则,提取知识以及获取洞察. 已被应用在国土安全、商业智能分析以及金融分析等很多不同的领域,受到国内外学术界、工业界以及政府部门越来越多的重视. 文中首先简要地介绍了文本可视分析的一般流程;然后系统地介绍了典型的文本分析和可视化技术,并着重讨论这 2 类研究的最新技术以及发展;最后对全文进行了总结并展望了文本可视分析面临的四大研究挑战:海量数据规模、复杂数据的不确定性、数据融合以及人机交互.

关键词: 文本信息可视化;文本数据挖掘;信息可视化;可视分析;人机交互
中图法分类号: TP391

A Survey on Topic-Based Visual Text Analytics

Wu Yingcai, Cui Weiwei, Song Yangqiu, Chen Yang, and Liu Shixia
(*Microsoft Research Asia, Beijing* 100190)

Abstract: Visual text analytics is a multi-discipline research area involving text mining, computer graphics and imaging, and human computer interaction. It helps users analyze a large amount of text documents interactively and visually to provide timely feedback, find anomalies, detect patterns, and gain insight. It has been used in many fields such as homeland security, business intelligence, and financial analysis and has received much attention from academia, industry, and government. This paper introduces a general pipeline of visual text analytics, followed by a survey of existing text mining and visualization techniques. We finally suggest a few potential research challenges, namely, the challenges of scalability, uncertainty, heterogeneity, and user interactions.

Key words: text visualization; text mining; information visualization; visual analytics; human computer interaction

在纷繁复杂的信息世界里,文本是最为重要的信息传递及交流沟通手段. 随着计算机网络和信息技术的广泛应用,文本信息越来越海量化、多样化和即时化. 例如,目前互联网上的网页数量已达数十亿,并且呈快速增长的趋势;据估计,每天有数十万的网页更新,数百万新的网页加入. 因此如何帮助人们迅速分析海量文本信息,获取洞察,已成为当前十分关键也颇具挑战性的问题.

为了帮助用户更好地理解和处理日益增长的文本信息,研究人员设计开发了多种工具和技术,并利用它们有效地组织和管理这些丰富而又复杂的文本集,进而帮助用户快速、准确、全面地从中找到他们所需要的信息. 这些工作主要来自于文本挖掘和信息可视化 2 个研究领域.

收稿日期:2012-07-31. 巫英才(1983—),男,博士,副研究员,主要研究方向为信息可视化、可视分析、科学可视化;崔为炜(1982—),男,博士,副研究员,主要研究方向为信息可视化、可视分析;宋阳秋(1981—),男,博士,副研究员,主要研究方向为机器学习、文本数据挖掘及可视化;陈 杨(1984—),男,博士研究生,主要研究方向为可视分析;刘世霞(1974—),女,博士,主管研究员,主要研究方向为信息数据可视化、可视分析.

在过去的几十年中，研究人员在文本挖掘领域提出了一系列基于文本主题的分析技术[1-3]，可以在一定程度上帮助用户理解这些复杂的文本信息. 然而主题分析技术的结果通常非常复杂，一般的用户难以理解. 例如，一个主题通常由一组关键词来描述，其中每个关键词有一个概率值用于衡量它在该主题中的重要性，而每个文档又有一定的概率属于不同的主题. 另一方面，不同的用户有不同的需求，一个主题模型很难满足不同用户的信息需求.

为了解决这一问题，研究人员将文本分析技术和交互式可视化技术结合在一起，设计了多种文本可视分析技术. 这些技术充分利用人类与生俱来的对图形的迅速辨识及分析能力，将文本挖掘结果及相应的文本数据转换成直观的、可交互的展现形式，使人们可以通过视觉迅速获得有用信息，从而达到对大文本数据集进一步分析、推理以及理解的目的. 已有的可视分析技术主要包括静态和动态两大类方法：静态可视方法不关心文档的时间属性，着重研究文档以及内容直接的静态关系；而动态方法则研究文档集合中随着时间变化的内容以及相应关系，用于找出一些关键的时刻和事件，并进一步推导相应事件产生的原因.

文本可视分析技术有很多实际的应用. 在日常工作中，用户经常需要通过分析大量的文本信息了解文档集里面的主要内容和主题以及主题之间的关系，在此基础上进一步找出哪些主题和任务相关. 例如，酒店管理人员非常关心客户对酒店的评价，为此他们会收集大量的用户反馈，并从中分析出客户的主要意见和建议以及相应的改进方案. 在有时间属性的文档集中，用户需要理解主题随着时间变化的主要趋势以及这些变化之间的关系. 通常用户会关心如下问题：最近产生了哪些有趣的主题？哪些被关注的主题最近消失了？哪些主题发生了合并？哪些主题分裂出多个新的主题？例如科研人员需要阅读大量的科技文献，并从中发现他们所关心学科的研究热点、趋势及相关学科之间的演变.

本文系统地分析了已有的主要可视分析技术，帮助读者更好地理解这些技术的优点及不足. 在此基础上，总结了该技术研究中的主要难点以及未来的研究热点. 本文中尽量避免一些枯燥而复杂的技术细节介绍，从实际问题出发帮助读者更好地理解基于主题的文本可视分析技术研究中的主要方法和技术难点，为进一步设计和应用相应的可视分析技术做一个铺垫工作.

1　文本可视化流程

文本可视分析紧密结合文本挖掘技术以及交互式的数据可视化，并充分利用用户丰富的背景知识、高带宽的视觉处理能力以及强大的推理能力对海量文档数据进行分析，从中提取知识、寻找规则、获取洞察和发现异常，以辅助用户进行决策. 图 1 所示为文本信息可视分析的通用流程，其中从左到右显示了原始无结构的文本数据在经过各种挖掘技术的一系列变换处理之后，浓缩成有意义的结构化信息；然后通过可视化技术的转换生成利于交互的直观图表，从而帮助用户进行分析推导.

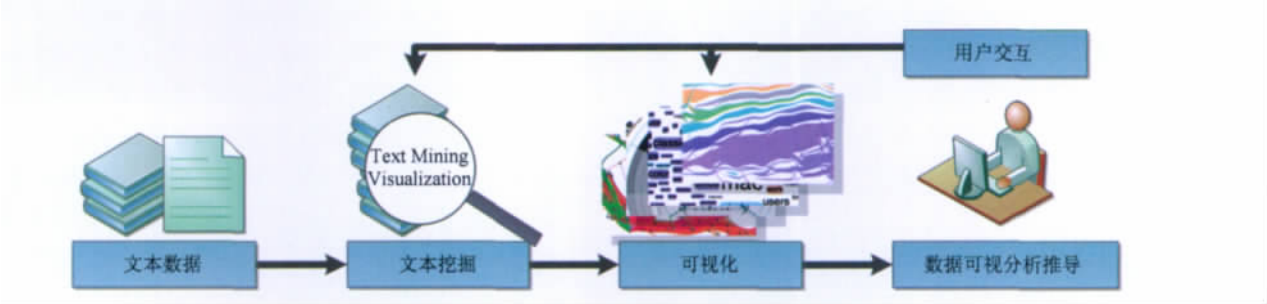

图 1　文本信息数据可视化流程图

图 1 中，文本挖掘和可视化是 2 个主要的核心技术. 有效的文本数据挖掘技术往往可以帮助我们从大规模的文本集合中过滤大量的冗余和无用的信息，将海量的文本数据浓缩成为适合可视分析的结构化数据. 文本挖掘技术可以自动从大量的新闻文本中提炼出富含信息的关键词或者新闻主题[1]，这些精炼的信息可以用来比较准确地刻画和反映原来的数据内容. 例如，可以通过一组关键词了解相应文档的大概内容，依靠这些信息可以极大地压缩文本集合的数据规模同时保留关键信息. 可视化技术可以将这些提炼出来的结构化信息直观地呈现出来，在高度交互的人机图形界面以及充满美感的可视化

图表的帮助下,用户可以利用其自有的领域知识非常容易地设定文本挖掘的各种参数,引导文本挖掘的过程.此外,用户也可以交互地生成各种有效的可视化图像,以多个角度去分析经过提炼的结构化文本数据信息,从而发现数据中蕴含的有用信息.

2 主要研究方法和发展现状

2.1 文本分析方法

文本主题是具有某一类主题的文档集合,例如新闻中谈论政治、经济和娱乐的文章就涉及到不同的主题.如果已知需要哪几类主题,就可以使用针对带监督信息的分类技术进行判断;如果给定的文本集合未指定主题的类别,则需要使用非监督的聚类或者主题模型来进行分析.本节主要讨论将文本表示成词袋(bag of words)的相关方法.

文本分类的研究已有较长的历史,其中经典的方法有朴素贝叶斯(naive Bayes)[4]、最大熵(maximum entropy)[5]和支持向量机(support vector machine)[6]等.这些方法均把文本表示成词袋,使用高维空间中的向量来计算文本的相似度,在高维向量空间中寻找在不同准则下的判别函数来对文本进行分类.当有特定的目标去寻找某类主题时,只要有足够多的标注样本,分类模型是最好的选择.

现实问题中,很多文本数据是没有类别标签的,例如在互联网上可以得到大量的新闻、博客、BBS讨论版和微博数据.因此尤其是在可视化领域,非监督的聚类和主题模型会受到更多的关注.文本的聚类可以看成是一种混合模型的概率密度估计问题[7],其中的关键就在于如何设计一个分布函数以能够很好地描述文本数据的特点,例如 K-means 就可以看成是一种概率混合模型的特殊形式[8].在实际中,文本的每个词都可以由不同的主题产生,因此一个文本可以看作是多个主题的混合.但是聚类方法使用一个分布来刻画某个文本中词的分布,从而导致对文档的产生模型的概率描述不够精确.

为了解决这个问题,主题模型,如 PLSA (probabilistic latent semantic analysis)[9]和 LDA (latent Dirichlet allocation)[1]等把文本看成主题的概率密度混合,进而把主题表示为词的分布.主题模型和聚类的关键区别是主题产生文本的过程.聚类认为每个文本是由一个主题产生的,而主题模型则认为每个文本是由多个主题产生的.最近,人们在聚类或者主题模型上进一步引入狄利克莱过程(Dirichlet process)先验或者层次化的狄利克莱过程(hierarchical Dirichlet processes)先验,使得模型可以通过自动推理得到主题个数[2].

主题模型有很多变形和改进,可以用于处理各种不同的问题.例如,半监督的主题模型[10]将人的先验知识引入主题模型中,使得训练的主题可以更好地和人的先验知识结合到一起;处理词之间跳转关系的隐马尔科夫主题模型[11]考虑到词与词之间不是互相独立的,突破了词袋模型的局限,从而得到更好的主题模型;另外还有对主题模型进行加速的方法[12].这些改进扩展了主题模型的适用范围,提高了建模能力.

随着互联网的发展,大量的文档不断地被产生出来,因此分析文本随着时间变化的主题模型受到很大重视.尤其在可视化领域,分析主题的动态变化可以帮助用户更好地理解和分析数据.对文本主题的动态建模可以简单地分成两大类:一类是将时间视为随机变量,从而进行连续时间的建模[13];另一类是把时间离散化为一系列时间戳,进而对离散化的时间点构建动态贝叶斯网络[14].随后,Wang 等又把第二类方法推广到极限情况来处理连续时间的动态主题跟踪问题[15].此外,还有一些研究工作是使用狄利克莱过程构建动态演化主题模型进行主题跟踪[16].由于狄利克莱过程作为先验的聚类或主题模型在某种意义上可以自动确定主题个数,因此这类模型自然地可以发现主题的出现和消亡,从而帮助用户自动检测关键兴趣点.

对主题的动态跟踪还有 2 个方向值得引起重视.1)对不同数据源的跟踪和建模,包括发现同一个主题在不同源之间的比例、出现顺序等[3].由于在不同的数据源中对某个主题的讨论是不同时的,例如新闻中对大型事件的报道就会比博客和 BBS 中快,而与某些小道消息相关的主题则有可能在博客或 BBS 中先被讨论.能够自动地分析这些信息传播并将结果展示给用户,对更有效率的数据理解很有好处.2)寻找并跟踪随着时间变化的主题之间的关系,如分析主题的分裂和融合等[17].由于主题之间并不是完全不相关的,例如新闻中讨论经济的主题也会讨论与政治相关的事件,或者某个领域学术文章的子主题之间都是互相关联的,因此我们不仅希望用计算机自动找到主题,还希望计算机能够帮助用户迅速地发现主题之间的联系和演变关系.

2.2　可视化方法

根据是否着重研究文档的时间属性，文本可视分析方法可大体分为静态和动态的两大类. 例如，在对一个新闻文档集进行可视分析时，如果只研究其中包含几个主题以及分析各个主题内容之间的关系，即属于静态可视化一类；而如果研究的是不同主题随时间的变化，那么属于动态可视化一类（即使其生成的可视化结果也为一张静态的图片）.

在静态可视化类中，最常见的方法是使用投影的方式来显示主题（或是文档）的分布以及它们之间的关系[18-19]. 投影类方法的最大作用是将文档的聚类信息直观地展示给用户，以帮助他们找到特别的或是所关心的主题，进而找到相应的文档. Cutting 等[20]的研究结果表明，当人们的需求或者任务并不明确时，这类方法往往比传统的关键词搜索或者浏览的方法更加有效. InfoSky[18] 就是投影方法的一个典型例子，其将文档集合以及文档比作银河系以及其中的恒星，将它们投影到二维平面上来显示其中各个主题的层次结构关系，如图 2 所示. 除了投影之外，topic island[21] 还借鉴了小波分析技术将文本的主题结构转化成“热度图”，用颜色来表示各个主题的强度和分布. 但是这些投影及类似的方法往往也会丢失大量的信息，造成投影出的结果难以理解或解释. 针对这个问题，最近人们开始使用其他技术来加强这种投影方式的表达能力，FacetAtlas[22] 就是其中的一种，其将投影方法和点线图的绘制方法相结合来表达数据中的不同侧面的信息，帮助分析它们之间潜在的关联模式，如图 3 所示.

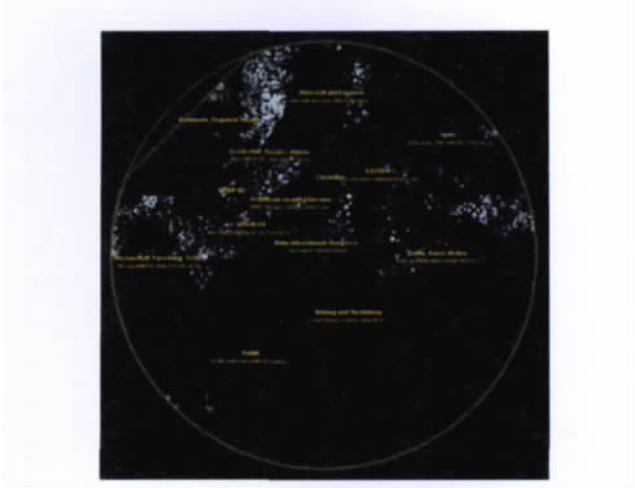

图 2　InfoSky[18]

与静态可视分析相比，动态可视分析是一个相对比较新的研究方向，并且大有后来居上之势，这主要归功于带时间标记的文档变得越来越重要且容易

图 3　FacetAtlas[22]

收集. 其中一种方式是直接扩展上述静态可视分析方法：首先按照时间顺序将文档集合分割成许多个时间片，并且针对每个时间片生成一个静态的可视化结果；然后将所有结果集中起来，使用动画等技术来表现不同结果之间的变化. 如 Hetzler 等[23] 在每个时间片上将文档投影到一个二维平面，并在其中着重画出新出现或是即将消失的内容；通过控制一个滑动窗口，帮助人们向前或是向后浏览不同的时间片，从而快速地找到关键的时间点以及具体内容. Erten 等[24] 将每个时间片中的文档主题以及它们之间的关系归纳为一个点线图，然后使用一种图的布局算法来保证图在从一个时间片变化到另一个时间片时能够保持一定的稳定性.

虽然动画技术（或是滑动窗技术）能够快速地给用户展示不同时间片的文本主题信息，但是由于人类大脑的短暂记忆特点[25]，用户通过动画很难记住或者比较各个主题在时间维度上的具体变化，或是发现不同主题在不同时间上的具体关联. 因此，当前研究也越来越多地倾向于使用静态的方式来表示动态的文档集合，其中最流行的是基于 ThemeRiver[26-28] 的方法. 在这类可视化方法中，时间被表示为从左往右的一条水平轴，然后用不同的色带代表不同的主题. 这种方式可以直观地描述主题是随着时间发展的，如图 4 所示，每条色带在不同时间的宽度代表该主题在该时间的一个度量（例如主题的强度）. 使用这种表示方式的最大优点是人们可以很容易地跟踪任何一个主题在该度量上随时间发展的变化；此外，也能容易地比较不同的主题在同一个时刻的相对大小. 但是，这类方法在表达能力上也有很大的局限性，例如传统的 ThemeRiver 只能够将每个主题在每个时间刻度上概括为一个简单的数值，而随时间的变化主题并不能被一个简单的度量所完整地描述. 为了弥补这个缺陷，人们对 ThemeRiver 做了进一步

的扩展，以表达更多的信息. 例如 TIARA[29-30] 将 ThemeRiver 和 Word Cloud 技术相结合，用来描述文本主题在内容上随时间变化的规律. 具体而言，TIARA 为每个文本主题在每个时间点上提取出不同的关键词，然后将这些词排布在相应色带上的相应位置，并用词的大小表示关键词在该时刻出现的频率. TIARA 能帮助用户快速分析文本的具体内容以及其随时间变化的规律，而不是仅仅限于某一个度量的变化，如图 5 所示. TextFlow[31] 作为 ThemeRiver 的另一种扩展，它不但关心主题在内容上的变化，而且也关心各个文本主题之间在时间维度上的关系. 例如某个文本主题在某个时刻分裂为 2 个或是更多个主题，或是若干个主题在某个时刻融合成了一个单独发展的主题，如图 6 所示.

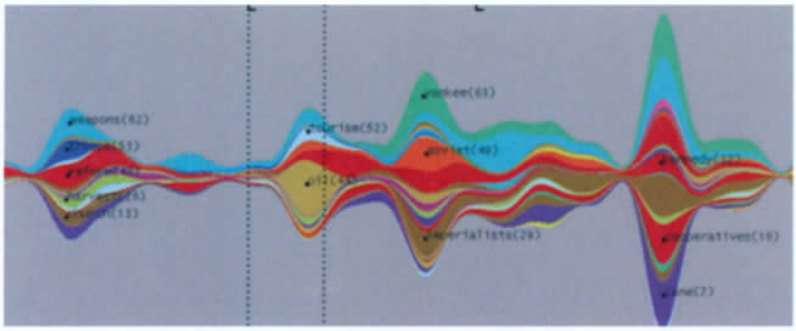

图 4 ThemeRiver[28]

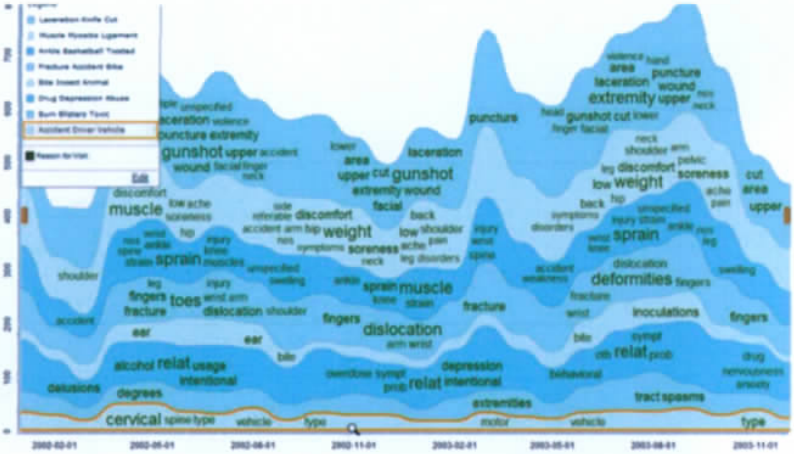

图 5 TIARA[29]

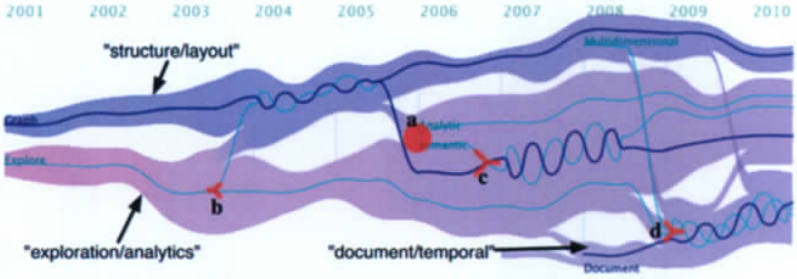

图 6 TextFlow[31]

当然，ThemeRiver 并不是分析文本主题演化的唯一静态方式. Viégas 等提出了 Themail 技术[32] 来分析两个人之间的邮件主题演化. 在这个工具中，主题的变化被表现为一组关键词列表，在每一个时间片上用关键词的大小来表示其出现的频率，借此告诉人们各个主题的迁演. CAST[33] 将新闻文档集合中的主题抽象成一组随时间变化的关键词，在一个主题内关键词之间用宽度不一的线条相连以表示新闻故事的发展主线，如图 7 所示.

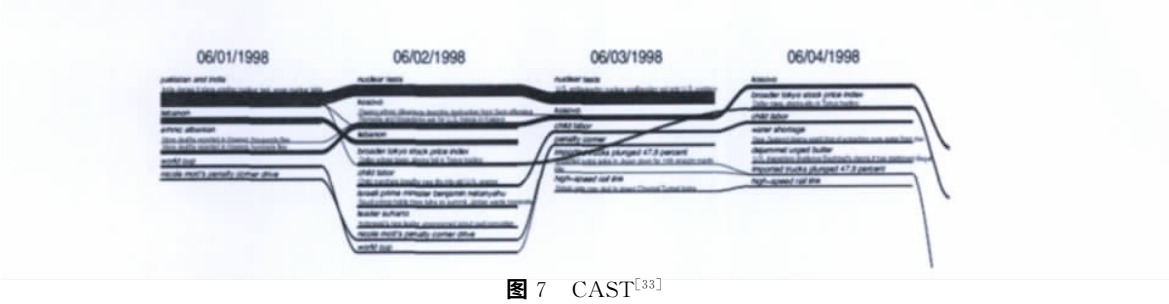

图 7　CAST[33]

3　总结与展望

文本可视分析是一种非常有效的数据分析方法，已被运用在各行各业. 本文介绍了文本可视分析中的主要问题以及一般流程，并着重分析了该领域当前的研究和发展现状. 虽然近年来文本可视分析受到了学术界、工业界以及政府部门的重视，但是它的进一步发展和推广却受到了一些技术方面的制约. 我们认为，当前的研究与应用面临着海量数据规模、复杂数据的不确定性、数据融合以及用户交互等一系列重大挑战.

海量数据. 在实践中我们经常遇到海量的、实时的文本信息流，如微博信息流和实时的新闻信息聚合(RSS). 对这类信息进行分析跟踪可以帮助我们及时了解舆论风向，对于理解信息传播规律以及发现突发性事件有非常重大的意义. 但是由于当前技术的限制，大多数文本可视分析系统均不具备处理和分析大规模实时数据流的能力. 主要瓶颈在于文本信息可视化的各个过程，如缺乏高效可靠的文本数据挖掘技术、高度可扩展的可视化图像表达方法以及在大规模数据下的有效的人机交互方法等. 如何适应大数据时代的海量实时数据，是文本可视分析的一个主要挑战.

数据不确定性. 文本信息是一种复杂抽象的数据. 当前的数据挖掘技术并不能够完全正确地理解文本数据所包含的内容(特别是语义方面的)，因此由数据挖掘所提取出来的信息往往含有不确定性. 如何在数据挖掘中准确地描述这种不确定性，并且在分析中将这种不确定性忠实地展现给用户，避免误导用户得出错误的结论或者做出错误的决策，也是一个主要的研究方向.

数据融合. 文本数据经常需要和其他相关的非文本数据融合在一起进行关联分析. 例如微博数据既包含了无结构的文本微博信息，也包含了用户的一般性资料，如地理信息、年龄段、性别等等非文本的结构数据. 文本可视分析的一个主要的优点是它允许用户融合多种异构的数据信息，以不同的角度去关联、分析以及理解数据. 但是，如何有效地将文本数据与其他数据在可视化的图像上面进行融合是一个非常有挑战性的问题.

用户交互. 直观易用的交互机制可以帮助用户更好地将他们已有的领域专业知识反馈给文本挖掘模型，以此帮助改进知识获取的过程. 另一方面，也可以在交互的过程中更有效地展现信息和传达知识给用户. 因此，它可以帮助消除人与机器之间的鸿沟，从而允许用户更深入地去分析、理解以及探索数据. 但是，由于文本数据本身的抽象复杂性、大规模以及融合等方面的问题，给直观易用的人机交互方法带来了一定的挑战.

参考文献(References)：

[1] Blei D M, Ng A Y, Jordan M Y. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022

[2] Teh Y W, Jordan M I, Beal M J. Hierarchical Dirichlet processes [J]. Journal of the American Statistical Association, 2006, 101(476): 1566-1581

[3] Zhang J, Song Y, Zhang C, *et al*. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora [C] // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 1079-1088

[4] McCallum A, Nigam K. A comparison of event models for naive Bayes text classification [C] // Proceedings of AAAI/ICML Workshop on Learning for Text Categorization. Palo Alto: AAAI Press, 1998: 41-48

[5] Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification [C] // Proceedings of IJCAI Workshop on Machine Learning for Information Filtering. Palo Alto: AAAI Press, 1999: 61-67

[6] Joachims T. Text categorization with support vector machines: learning with many relevant features [C] // Proceedings of the European Conference on Machine Learning. Heidelberg: Springer, 1998: 137-142

[7] Zhong S, Ghosh J. Generative model-based clustering of documents: a comparative study [J]. Knowledge and Information Systems, 2005, 8(3): 374-384

[8] Bishop C M. Pattern recognition and machine learning [M]. 2nd ed. Heidelberg: Springer, 2007

[9] Hofmann T. Probabilistic latent semantic indexing [C] // Proceedings of the Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Press, 1999: 289-296

[10] Andrzejewsk D, Zhu X, Craven M. Incorporating domain knowledge into topic modeling via Dirichlet forest priors [C] // Proceedings of the Annual International Conference on Machine Learning. New York: ACM Press, 2009: 25-32

[11] Gruber A, Rosen-Zvi M, Weiss Y. Hidden topic Markov models [C] //Proceedings of Artificial Intelligence and Statistics. Cambridge: MIT Press, 2007: 163-170

[12] Porteous I, Newman D, Ihler A, *et al*. Fast collapsed Gibbs sampling for latent Dirichlet Allocation [C] // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2008: 569-577

[13] Wang X, McCallum A. Topics over time: a non-Markov continuous-time model of topical trends [C] // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006: 424-433

[14] Blei D M, Lafferty J D. Dynamic topic models [C] // Proceedings of the Annual International Conference on Machine Learning. New York: ACM Press, 2006: 113-120

[15] Wang C, Blei D M, Heckerman D. Continuous time dynamic topic models [C] // Proceedings of the Conference in Uncertainty in Artificial Intelligence. Arlington: AUAI Press, 2008: 579-586

[16] Ahmed A, Xing E P. Timeline: a dynamic hierarchical Dirichlet process model for recovering Birth/Death and evolution of topics in text stream [C] // Proceedings of the Conference in Uncertainty in Artificial Intelligence. Arlington: AUAI Press, 2010: 20-29

[17] Gao Z, Song Y, Liu S, *et al*. Tracking and connecting topics via incremental hierarchical Dirichlet processes [C] // Proceedings of the IEEE International Conference on Data Mining. Los Alamitos: IEEE Computer Society Press, 2011: 1056-1061

[18] Andrews K, Kienreich W, Sabol V, *et al*. The Infosky visual explorer: exploiting hierarchical structure and document similarities [J]. Information Visualization, 2002, 1(3/4): 166-181

[19] Chen Y, Wang L, Dong M, *et al*. Exemplar-based visualization of large document corpus [J]. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6): 1161-1168

[20] Cutting D R, Karger D R, Pedersen J O. Constant interaction-time scatter/gather browsing of very large document collections [C] // Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1993: 126-134

[21] Miller N E, Wong P C, Brewster M, *et al*. Topic islands—a wavelet-based text visualization system [C] // Proceedings of the IEEE Conference on Visualization. Los Alamitos: IEEE Computer Society Press, 1998: 189-196

[22] Cao N, Sun J, Lin Y R, *et al*. FacetAtlas: multifaceted visualization for rich text corpora [J]. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 1172-1181

[23] Hetzler E G, Crow V L, Payne D A, *et al*. Turning the bucket of text into a pipe [C] // Proceedings of the IEEE Symposium on Information Visualization. Los Alamitos: IEEE Computer Society Press, 2005: 89-94

[24] Erten C, Harding P J, Kobourov S G, *et al*. Exploring the computing literature using temporal graph visualization [C] // Proceedings of the Conference on Visualization and Data Analysis. Bellingham: Society of Photo-Optical Instrumentation Engineers Press, 2003: 45-56

[25] Nowell L, Hetzler E, Tanasse T. Change blindness in information visualization: a case study [C] // Proceedings of the IEEE Symposium on Information Visualization. Los Alamitos: IEEE Computer Society Press, 2001: 15-23

[26] Byron L, Wattenberg M. Stacked graphs-geometry & aesthetics [J]. IEEE Transactions on Visualization and Computer Graphics, 2008, 14(6): 1245-1252

[27] Dork M, Gruen D, Williamson C, *et al*. A visual backchannel for large-scale events [J]. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 1129-1138

[28] Havre S, Hetzler E, Whitney P, *et al*. ThemeRiver: visualizing thematic changes in large document collections [J]. IEEE Transactions on Visualization and Computer Graphics, 2002, 8(1): 9-20

[29] Liu S, Zhou M X, Pan S, *et al*. TIARA: interactive, topic-based visual text summarization and analysis [J]. ACM Transactions on Intelligent Systems and Technology, 2012, 3(2): 25: 1-25: 28

[30] Wei F, Liu S, Song Y, *et al*. TIARA: a visual exploratory text analytic system [C] // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2010: 153-162

[31] Cui W, Liu S, Tan L, *et al*. TextFlow: towards better understanding of evolving topics in text [J]. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2412-2421

[32] Viégas F B, Golder S, Donath J. Visualizing email content: portraying relationships from conversational histories [C] // Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2006: 979-988

[33] Butner S, Cowley W, Gregory M, *et al*. Describing story evolution from dynamic information stream [C] // Proceedings of the IEEE Symposium on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2009: 99-106