情报工程 ISSN 2095-915X
第2卷 第4期 20-29 2016年8月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.4 20-29 Aug 2016

doi:10.3772/j.issn.2095-915x.2016.04.004

# AToT 模型可视化工具开发

孙国超，徐硕，乔晓东

（中国科学技术信息研究所 北京 100038）

摘要：随着科研人员需要处理的文献集规模的日益庞大，以 LDA 为代表的主题模型能够从语义层面挖掘大规模文献集中隐含的主题，因此，LDA 主题模型的应用越来越广泛。LDA 模型仅仅关注文献集的内容，而忽略了文献其他重要的外部信息，AToT 模型在 LDA 主题模型的基础上引入了文献作者和文献发表时间两个属性，使 AToT 模型不仅可以挖掘文献中隐含的信息，还可以分析文献作者的研究兴趣及文献主题随时间的变化。AToT 模型对文献集建模的结果是以概率矩阵的形式呈现，不能直观、全面、清晰的呈现挖掘出来的信息，特别是对数据挖掘不熟悉的科研人员，因此，本文开发了一个基于 AToT 模型的可视化系统，该可视化系统清晰、美观地展现了 AToT 模型中文献、主题、作者、时间、词项间的关系。如文档中的主题分布、主题的词项分布、作者的研究兴趣分布、主题的相似主题和主题的演化趋势等。

关键词：LDA 模型，AToT 模型，可视化， Django

中图分类号：G35，TP39

# Development of Visualization Tool for AToT Model

SUN GuoChao, XU Shuo, QIAO XiaoDong

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

**Abstract:** Since LDA (Latent Dirichlet Allocation) topic model could mine underlying topics fromthe collection

基金项目：本文受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（71403255），“十二五”国家科技支撑计划项目：面向科技情报分析的信息服务资源开发与支撑技术研究（2015BAH25F01）和中国工程科技知识中心建设项目“知识组织体系建设”（CKCEST-2016-2-10）资助。

作者简介：孙国超，硕士研究生；徐硕（通讯作者）（1979 – ），博士，副研究员，研究方向：智能情报分析，数据挖掘和大数据等，E-mail: xush@istic.ac.cn；乔晓东（1965 – ），英国谢菲尔德大学硕士，研究员，研究方向：信息服务、信息资源管理等，E-mail: qiaox@istic.ac.cn。

of large-scale documents in the semantic viewpoint, it has been applied successfully in various fields. However, LDA model was only focus on the contents of documents while ignoring other important external information, such as authorship, timestamp, et ac. In order to overcome this problem, AToT (Author Topic over Time) model was proposed by combined analysis of the authorships and the publication time of documents, which can improve the AToT model for mining the implicit information of the documents, and analyzing the research interest of the authors and the variation of the documents. However, it was difficult to understand the results of these models, especially for researchers unfamiliar with data mining. Therefore, this study developed a visualization tool for AToT model. The visualization system showed the relationship between topic,term,document,time and author clearly, for instance,the distribution of topic in document,the probability of words in topic,author's interest,the similar topics and the trend of topic over time.

**Keywords:** LDA model, AToT model, visualization , Django

## 1 引言

文献作为科研成果的主要表现形式之一，是科研人员的智慧结晶，展现了某一领域的发展现状、研究进展和发展趋势。对文献集合中科技主题的探测和跟踪有助于科技人员快速地了解特定领域，并对所研究内容做出进一步的调整。然而随着社会的快速发展，人们需要处理的数据越来越庞大，特别是科技工作者，动辄需要从几千甚至几万篇文献的数据中挖掘隐含的主题及其演化规律，这极大的增加了科研工作者的工作强度。通过人工阅读这些文献挖掘主题的方式显然是不现实的，于是 1983 年 Salton 和 McGill[1] 提出了基于词频的文档表示模型，而文档间的联系不仅仅决定于词语的重复，这种技术无法从语义层面判断词语或文档间的联系，也无法解决一词多义和一义多词的语言现象，这限制了基于词频统计的技术的效率。以 LDA（Latent Dirichlet Allocation）[2] 为代表的主题模型，能够揭示隐藏在大规模文档集中的主题，从语义层面分析比较文档间的联系，并且大大降低了文档的纬度，提高了计算的效率，日益成为文本挖掘的研究热点 [3]。

科技文献除了内容之外还包含大量的外部属性信息，比如作者、发表时间以及参考文献等 [4]，这些外部属性对隐性知识显性化也起着非常重要的作用。因此，以 LDA 模型为基础，国内外学者通过引入了相应的外部属性信息，提出了很多衍生模型，比如 AT[5]、DTM[6]、ToT[7] 等模型，但大多模型只能引入单一外部属性信息，同时融合多种外部属性特征的主题模型研究仍处于起步阶段，徐硕等人提出的 AToT（Author-Topic over Time）模型 [8、9] 尝试将作者和时间两种外部属性特征融入主题模型，该模型不仅可以挖掘隐藏在文档中的主题，文档的作者与主题间的关系，还可以分析作者的研究兴趣随时间的演化规律。

上述所有的主题模型尽管可以揭示科技文献中的隐性知识，来达到帮助科研人员深入分析文档集的目的，然而用户必须明白，理解并分析建模结果的数据分布结构，才能明白主题建模的实际含义 [10]，这为用户解读主题模型的结果增加了不少困难，因此本文研发了一种针对 AToT 模型建模结果的交互式可视化系统。AToT 模型通过对文档集的建模，总结和组织了整个文档集，挖掘出隐含的主题、作者、文档、词项及时间等五

要素间的相互关系，而本文所研发的可视化系统可以通过交互的方式将这些相互关系清晰地展现给科研人员，帮助科研人员轻松理解和分析整个文档集 [11]。该可视化系统可以直观揭示：①文档集中隐含的主题；②主题包含的词项、相似主题、包含该主题的文档、擅长该主题的作者；③文档的主题分布、相似文档、作者分布、文档内容；④作者的主题分布、作者兴趣随时间演化、相似作者。

## 2 作者主题演化（AToT）模型

徐硕等人在 AT(Author-Topic Model) 模型和 ToT(Topic over ime) 模型的基础上提出了 AToT 模型，不仅引入了文献作者外部属性，还引入了文献的时间属性，在挖掘文献主题、作者兴趣的同时，也从动态的角度分析了作者的兴趣随时间的变化。图 1 给出了 AToT 模型的概率生成图，其生成过程如下：（1）对于每篇文献，从作者的分布中随机均匀选择一个作者 Xm,n；（2）对于每个作者 Xm,n，从先验参数为 α 的 Dirichlet（α）中取样作者所关联的主题多项式分布 θa；（3）从作者所关联的主题多项式分布 θa 随机取样一个主题 Zm,n；（4）对于每个主题，从参数为 β 的 Dirichlet（β）中取样该主题所关联的词项多项式分布 φk；（5）从词项多项式分布 φk 中取出一个词项，从主题随时间的 Beta（Zm,n）中取出一个时间戳分配给词项（词项和时间戳的取样顺序无先后之分）。AToT 模型中所使用的符号含义如表 1：

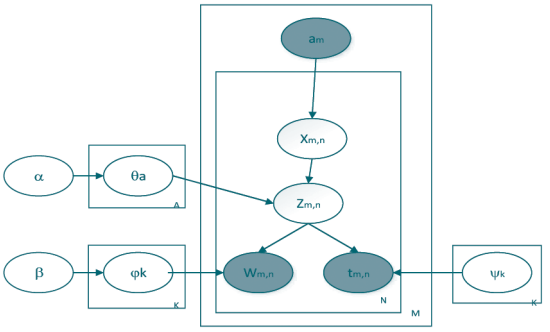


图 1 AToT 模型概率生成图

## 3 AToT 模型可视化工具主要功能模块的实现

### 3.1 Django 工作机制

考虑到浏览器和网络应用程序的流行，本文研发的可视化工具以 Python 中 Django 模块 [12] 为基础，采用 B/S 架构形式。Django 处理用户请求的工作机制如图 2 所示：（1）接收用户发送的 HTTP 请求；（2）将接收到的 HTTP 请求打包成 HttpRequst 对象，与网址视图管理器（URL.PY）进行模式匹配，匹配成功，则将匹配的视图传给视图管理器 VIEW.PY，反之，则返回没有找

表 1 AToT 模型符号含义

| 符号 | 含义 | 符号 | 含义 |
|---|---|---|---|
| 深色圆节点 | 观测值 | θa | 作者 a 的主题多项式分布 |
| 圆节点 | 隐含变量 | Φk | 主题 k 的词项多项式分布 |
| 方框 | 其中内容重复 n 次 | am | 第 m 篇文档的主题向量 |
| M | 文档的数量 | Xm,n | 第 m 篇文档中第 n 个词的作者分配 |
| K | 主题的数量 | Zm,n | 第 m 篇文档中第 n 个词的主题分配 |
| A | 作者的数量 | Wm,n | 第 m 篇文档中第 n 个词 |
| N | 文档的长度 | Tm,n | 第 m 篇文档中第 n 个词的时间戳 |
| α | θa 的先验参数 | Ψk | 第 k 个主题的 beta 分布 |
| β | Φk 的先验参数 | | |

到该网址的错误；（3）视图管理器搜索匹配的视图函数，匹配成功，执行该视图函数，反之，返回错误；（4）视图函数从 Model 模型（数据库）中查询用户所请求的数据；（5）视图函数从模板管理器 Templates 中调用所需要的网页模板文件，匹配成功，将查询得到的数据传入网页模板文件。反之，则返回错误；（6）将模板文件或错误打包成 Response 对象返回给用户。

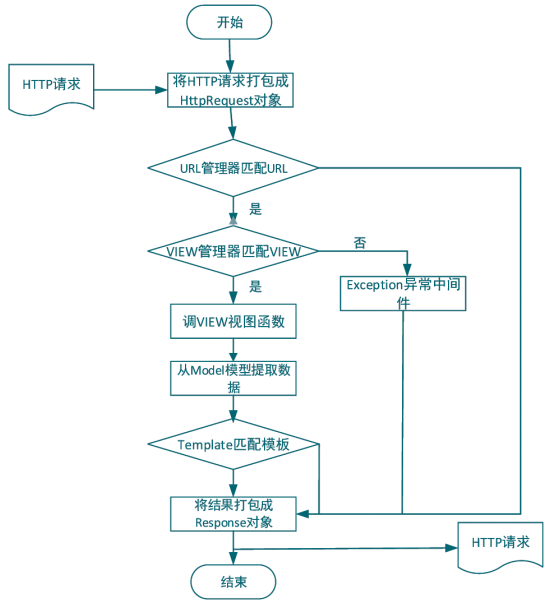


图 2　Django 工作原理

## 3.2 视图模块的实现

视图模块是该可视化工具的核心部分，视图模块根据用户的请求从数据库中检索相应的数据，将数据经过计算处理后转换为正确的格式传递给模板文件 Templates。AToT 模型数据可视化功能的实现是基于 JavaScript 脚本语言的 D3.js 类库实现的，D3.js 类库需要将数据存储成 JavaScript 的复合对象。因此，VIEW 视图模块需要将处理后的数据转换成 D3.js 类库方便处理的 JSON 格式，然后传递给模板文件 Templates。

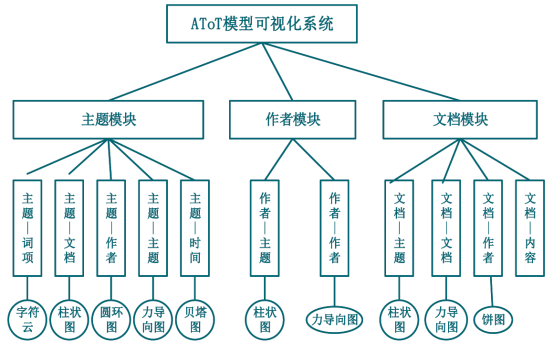


图 3　AToT 模型可视化系统模块框架

### 3.2.1 主题视图模块的实现

主题视图模块接收用户的 HTTP 请求后从数据库检索数据，通过计算处理主题与词项、文档、作者、其他主题间的关系。（1）主题内包含的 Top N 词项。通过文档内的词项属于主题的概率筛选出概率较大的前 N 个词项，通过字符云的方式呈现该主题所包含的词项，概率越大，词项的字体越大。（2）与主题相似程度较高的 Top N 主题。以词汇表中所有词项属于主题的概率组成主题向量，通过余弦相似度公式计算主题间的相似度，通过刻画力导向图鲜明、直观地展现主题与主题间的相似程度。（3）包含该主题比例较高 Top N 文档。通过文档内每个词项的主题标签，计算每个文档内各个主题所占的比例，并筛选出比例较高的前 N 篇文档通过柱状图的形式展示包含该主题的比例较高的文档。（4）该主题内排名较高的 Top N 科研人员 [13]。从文献作者的主题多项式分布中选取作者的研究兴趣是该主题的概率较高的前 N 个科研人员，通过圆环图展示作者的研究兴趣是该主题前 10 个作者。（5）主题随时间的变化趋势。主题随时间的变化服从贝塔分布，利用贝塔分布的两个参数，利用 Python 中的 Matplotlib 模块刻画贝塔分布图，展现主题随时间的发展规律。

余弦相似度的计算公式如下：

$$Cos(a,b)=\frac{(a*b)}{|a|+|b|}$$

其中 *a,b* 表示两个向量，*Cos(a,b)* 表示向量 *a,b* 间的相似度。

#### 3.2.2 文档视图模块的实现

文档视图模块接收用户的 HTTP 请求后从数据库检索数据，通过计算处理文档与主题、作者、其他文档间的关系。（1）文档内包含的主题。通过文档内每个词项的主题标签，计算每个主题在文献中所占有的比例，通过饼图展现文档内包含的主题及其所占有的比例。（2）文档合著作者的贡献程度。通过文档内每个词项的作者标签，计算每个作者在文献中所占有的比例，通过饼状图展示文档中包含的作者及其对文档的贡献程度。（3）与该文档相似程度较高的 Top N 文档。通过文献中隐含的每个主题在文档中占有的比例组成文档向量利用余弦相似度公式计算文档间的相似程度，利用力导向图刻画与该文档内容相似的文档。

### 3.2.3 作者视图模块的实现

作者视图模块接收用户的查询 HTTP 检索请求后，经过计算处理作者与主题、其他作者间的关系。（1）作者的研究兴趣分布。通过每个作者对应主题多项式分布计算作者的研究兴趣是各个主题的概率，通过饼状图直观、鲜明的展示作者的研究兴趣。（2）与该作者研究兴趣相似的 Top N 作者 [14]。利用作者的研究兴趣是各个主题的概率组成作者向量，通过余弦相似度公式计算作者间的相似程度，并通过力导向图展示作者的研究兴趣与其他作者的相似程度。

## 4 ATOT 模型的可视化

### 4.1 数据来源及预处理

本文采用的研究领域是情报学，数据源是从万方数据库中选取情报学领域发文量大于 50 的科研人员，并下载选中的科研人员从 2006—2015 年 10 年间发表的论文，共计 2536 篇，并将这些论文按年限分组。将下载的 PDF 格式的全文文献转换成 TXT 格式，然后进行数据清洗，删除一些不能识别的乱码，并利用斯坦福大学的 Stanford Segmenter 分词工具进行分词，并删除词频小于 3 的词项 [15]。从万方数据库下载的 50 位高发文量作者的文献集经过上述处理后，包含 D=2536 篇文献，K=1766 个作者和 V=16796 个单词的词汇表。

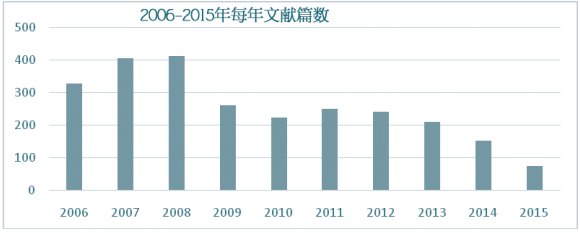


图 4 50 位高频作者每年论文数量

### 4.2 试验结果可视化

本文所实现的可视化系统主要包含五个界面：主题展示界面、主题（词项、作者、文档）随机检索界面、主题详细界面、文档详细界面、作者详细界面。这些页面挖掘和展现整个文档集中隐藏的结构，本文还通过链接的方式将这些页面贯穿起来。因此，可视化系统中每个页面的每个元素都可以让科研人员进入到新的页面，这样科研人员就能很容易的链接到系统中任何一个页面。例如：主题详细页面包含相关词项模块、相关作者模块。用户点击相关作者模块进入该作者的作者详细界面，点击作者详细界面中的相关主题模块，又重新进入到主题详细界面。

#### 4.2.1 主题模块可视化

“主题展现”页面：主题是文档集的隐含变量，概括和总结了文档集。主题热力图页面通过热力图的方式呈现了 AToT 模型对文档集建模后

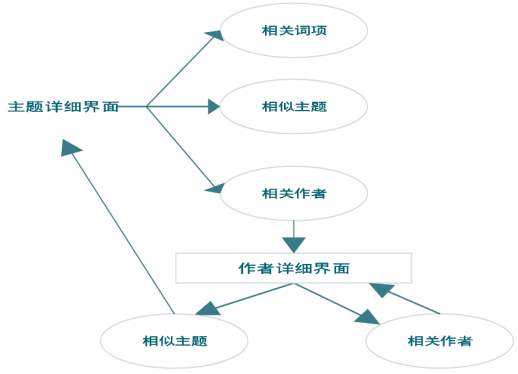


图 5　主题、作者链接机制示意图

挖掘的所有主题。热力图中每个点代表一个主题，主题颜色的深浅代表主题的权重，即该主题在文档集所有文档中所占比例之和，点与点之间的距离代表主题的相似度。从图 6“主题热力图”中可以看出，主题 25（“竞争情报”）、主题 26（“知识管理”）、主题 40（“文献计量”）、主题 55（“情报学基础理论”）等在整个文献集中所占的比例比较大，受到的关注度强。用户可以点击其中的任意主题进入相应的主题详细页面。

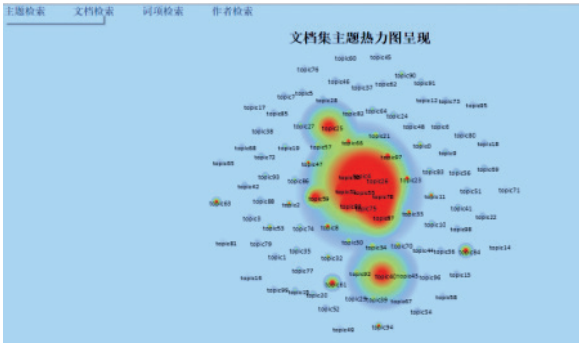


图 6　主题热力图

“主题详细”页面：该页面是对“主题展现”页面中某一主题的补充和进一步的描述。该页面通过五个模块，即五张图展示了 AToT 模型对文档集建模后挖掘的任意某个主题和文档、词项、作者、时间间的关系[16]。（1）主题与词项间的关系。“相关词项”模块通过字符云的方式展示了属于该主题的前 N 个词项列表。图 5 展示了主题 77（“数字图书馆服务”）包含的词项概率较高的有“咨询、服务、图书馆、参考、用户、数字、实时、方式、大学”等，检索框可以控制字符云的个数。（2）主题与文档的关系。“相关文档”模块通过柱状图展示了包含该主题的前 N 个文档。图 6 展现了主题 77（“数字图书馆服务”）在文档中所占比例较高的前 30 篇文档。当鼠标滑过每个文档时，页面的右上角会显示该文档的题目。前两篇文档分别为“我国高校图书馆网上实时咨询服务调查与分析”、“美国图书馆数字参考咨询服务成功因素探析”，其中“数字图书馆服务”主题在第一篇文献中所占比例为 90%，从这两篇文献的题目可以看出，这两篇文献与主题 77 非常相关。在检索框中可以控制文档的篇数。点击任意一篇文档进入该文档的详细页面。（3）主题与作者的关系。“相关作者”通过环形图展示了擅长该主题前 N 个作者列表。每个小环形表示作者擅长的领域是该主题的概率，图 7 展示了研究兴趣是“数字图书馆服务”且排名较高的 8 位作者。其中前两位作者是詹德优、李英剑。通过查看作者的相关主页可以得知这两位作者的研究兴趣与该主题高度相关。检索框可以控制作者的个数。点击相应的小环行进入相应的作者详细页面。（4）主题与主题的关系。通过力导向图的方式展示了与该主题相似的主题。每个点表示一个主题，点与点之间的距离表示主题与主题间的相似度。当鼠标滑过每个主题时，在页面的上方会显示该主题的名称，帮助科技工作者了解主题。图 10 展示了与主题 77（“数字图书馆服务”）主题内容相似的前 30 个主题。其中最相似的是主题 59、主题 68，主题 22 的名称是“图书馆、服务、知识、阅读、研究、论文、读者”。点击主题进入相应的“主题详细”页面。（5）主题与时间的关系。展示了该主题随时间的变化规律。图 9 展示了主题 77“数字图书馆服务”随时间增长呈下降趋势，即收到的关注度减少。


图 7　主题 77（“数字图书馆服务”）包含的词项

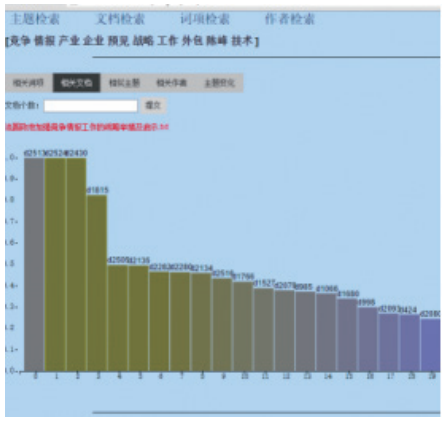


图 8　主题 77 在文档中占有比例较高的文档


图 9　研究兴趣是主题 77 排名较高的作者

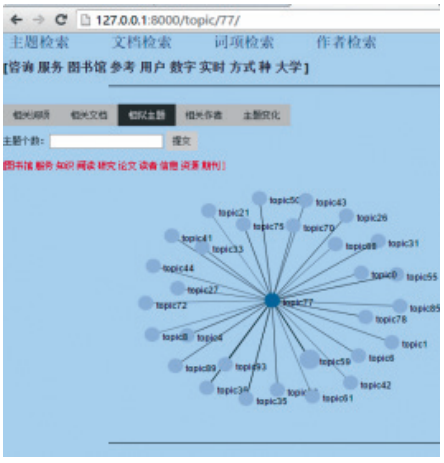


图 10　与主题 77 相似的主题

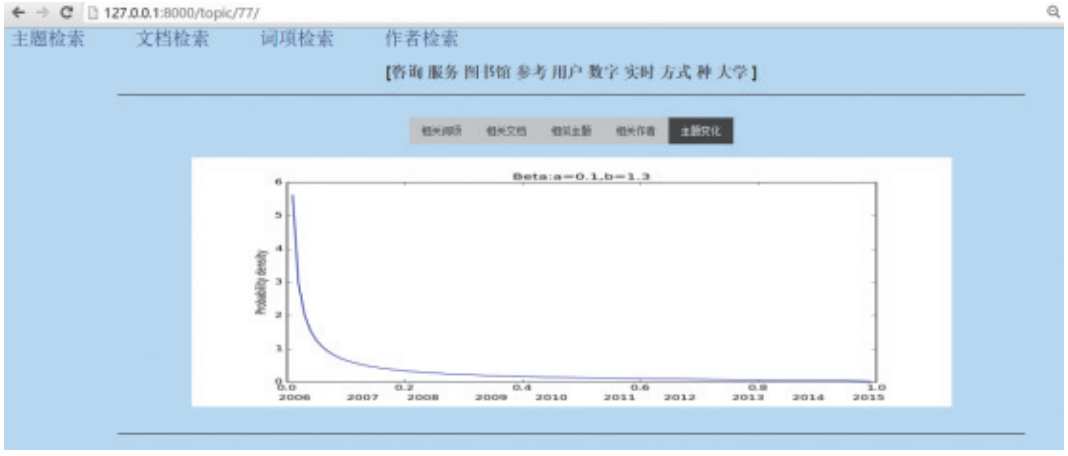


图 11　主题 77 随时间变化

### 4.2.2 文档模块可视化

文档详细页面：文档详细页面描述了文档与 AToT 模型对文献集建模后挖掘的主题、作者间的关系。（1）显示了文档的内容。“文档内容”模块展示文档的原文，帮助用户理解文档内的主题。图 12 展示了第 2093 篇文档“香港贸易发展局为中小企业提供竞争情报服务的方法”的原文内容。（2）文档与主题间的关系。“主题分布”模块通过柱状图显示了文档的主题构成，图 13 展示了第 2093 篇文档“香港贸易发展局为中小企业提供竞争情报服务的方法”的主题构成，其中主题 26“知识管理”占该文档近 70%，主题 25“竞争情报”占该文档近 25%。鼠标滑过主题会在上方显示该主题的名称，帮助用户理解该主题。点击主题进入相应的主题详细页面。（3）文档与作


图 12　文档 2093 原文

者间的关系。“作者贡献”模块通过饼状图显示了文档的作者构成。如图 14 所示第 2093 篇文档的作者对文档的贡献比例，其中陈峰所做的贡献占该文档近 45%，郑彦宁所做的贡献占该文档近 42%，赵筱媛 所做的贡献占该文档近 13%。点击作者进入相应的作者详细页面。（4）文档与文档间的关系。显示了与该文档相似的文档列表。图 13 展示了与第 2093 篇文档相似的前 30 个文档。

可以看出内容最相似的是主题文档 2092“风电产业竞争情报工作现状与需求实态调研”和文档 1907 “信息资源云与知识服务”。搜索框可以控制文档的个数。当鼠标悬浮于主题之上，右上方会显示该主题的名称，点击该主题进入相应的主题详细页面。

作者详细页面：作者详细页面描述了 AToT 模型对文献集建模后作者与作者、挖掘的主题的关系。（1）作者与主题的关系。“研究兴趣”模块通过柱状图显示作者的兴趣比率，图 16 展示了陈峰的研究兴趣分布，主题 25“竞争情报”是陈峰的兴趣的概率是近 0.5。搜索框可以控制研究兴趣的个数。鼠标悬浮于主题，上方会显示主题的名称，进入柱状图进入相应的主题详细页面。（2）作者与作者的关系。显示了与该作者研究兴趣相似的作者列表。图 15 展示了与陈峰研究兴趣相似

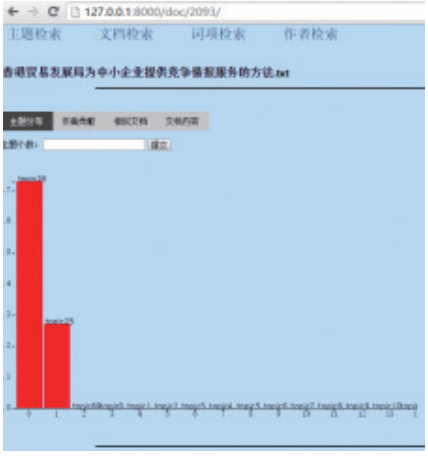


图 13　第 2093 篇文档主题分布

的前 30 位作者的姓名。可以查出与陈峰研究兴趣最相似的作者是郑彦宁。搜索框可以控制相似作者的个数，点击作者进入相应的作者详细页面。

主题检索、文档检索、作者检索、词项检索页面：如果用户想要检索任意一位科研人员、任意主题、任意文档，可以通过相应页面检索，该页面随机抽取了 150 个主题或文档或文献作者或词项，通过 3D 球星转动的方式展现出来，可以


图 14　第 2093 篇文档作者贡献比例


图 15　第 2093 篇文档的相似文档

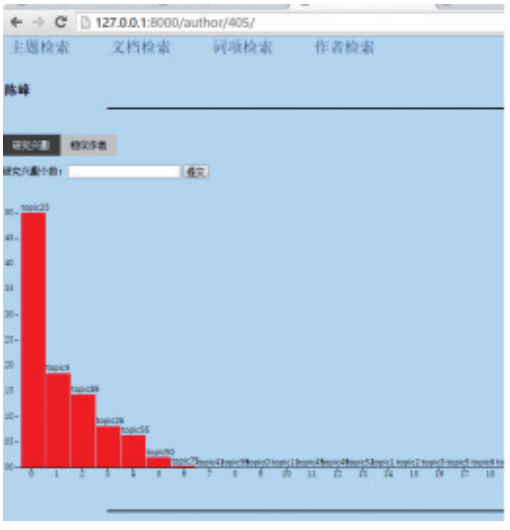


图 16　作者研究兴趣分布

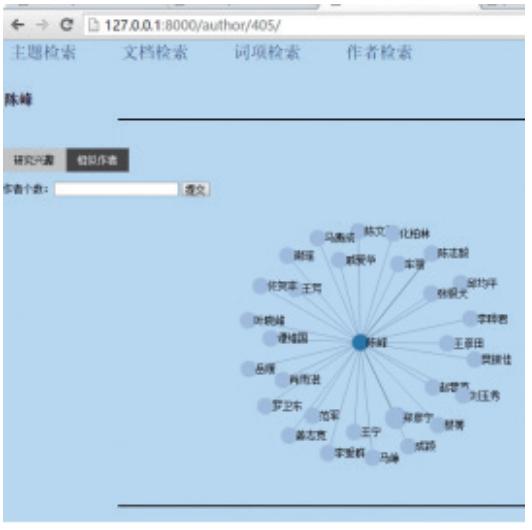


图 17　陈峰的相似作者

在页面中滚动的球状体中找到相应的内容点击进入，也在相应的检索框中输入检索内容即可，如想要检索邱均平的详细内容，在检索框中输入“邱均平”，进入邱均平的作者详细界面。

## 5 结论

本文从 LDA 主题模型的历史发展出发，分析了融入作者属性的主题模型和融入时间属性的主题模型的发展现状，并介绍了几种应用广泛的主题模型的特点及其优缺点，并着重介绍了融入作者、时间双属性的 AToT 模型，进而开发了基于 AToT 模型的可视化系统，帮助科研人员更清晰、直观地理解 AToT 模型对文档集建模的结果。本文所使用的数据集是从万方数据库中下载的 50 位高发文量和较高影响力的作者的文献集，这 50 位作者的合著作者发表的文献收录较少，所以，这 50 为合著作者的研究兴趣挖掘结果不是很理


图 18　作者随机选取页面

想，因此，接下来的工作会尽可能全地收录合著作者的发表的文献，进而准确挖掘合著作者的研究兴趣。

参考文献

[1]Deerwester S, Dumais S T, Furnas G W, et al. Indexing by Latent Semantic Analysis[J]. Journal of the American Society for Information Science, 1990, 41(6):391-407.

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

[3] 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, 34(8):1423-1436.

[4] 张晗，徐硕，乔晓东 . 融合科技文献内外部特征的主题模型发展综述 [J]. 情报学报 ,2014(10):1108-1120.

[5]Rosen-Zvi M, Griffiths T, Steyvers M, et al. The Author-topic Model for Authors and Documents[C]// Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2004:487-494.

[6]David M B,John D L.Dynamic Topic Models [C]// Proceedings of the23rd ICML International Co nference on Machine Learning , New York: ACM Press , 2006: 113-120.

[7]Wang X, Mccallum A. Topics Over Time: A Non-Markov Continuous-time Model of Topical Trends[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006:424-433.

[8]Xu S, Shi Q, Qiao X, et al. A Dynamic Users' Interest Discovery Model with Distributed Inference Algorithm[J]. International Journal of Distributed Sensor Networks, 2014 (2014):1-11.

[9] 史庆伟，乔晓东，徐硕，等 . 作者主题演化模型及其在研究兴趣演化分析中的应用 [J]. 情报学报 , 2013, 32(9):912-919.

[10] 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, 27(5):771-782.

[11]Blei D M, Lafferty J D. Visualizing Topics with Multi-Word Expressions[J]. Statistics, 2009:1-12.

[12] 刘班 . 基于 Django 快速开发 Web 应用 [J]. 电脑知识与技术 , 2009, 5(7):1616-1618.

[13]Sievert C, Shirley K E, Shirley K E. LDAvis: A Method for Visualizing and Interpreting Topics[C]// The Workshop on Interactive Language Learning, Visualization, and Interfaces at the Association for Computational Linguistics, 2014:63-70.

[14] Matthew J G, Joshua L, Jeff L,et al. The Topic Browser:An Interactive Tool for Browsing Topic Models[C]//NIPS Workshop on Challenges of Data Visualization, 2010:100-108.

[15] 黄昌宁，赵海 . 中文分词十年回顾 [J]. 中文信息学报 , 2007, 21(3):8-19.

[16] 李湘东，张娇，袁满 . 基于 LDA 模型的科技期刊主题演化研究 [J]. 情报杂志 , 2014(7):115-121.