# 信息检索领域主题研究
## ——基于SIGIR邮件列表和会议论文的比较研究*

赵忠伟，程齐凯
（武汉大学信息管理学院，武汉 430072）

摘要：传统学科主题研究主要基于学术文本题录数据，研究对象单一。本文以SIGIR (Special Interest Group on Information Retrieval) 邮件列表为切入点分别构建SIGIR邮件列表数据集和同期会议论文数据集，并在两个数据集的基础上对信息检索的主题结构和主题演化进行对比分析。研究发现，信息检索领域存在研究内容不断深入、研究方法不断增多和核心主题逐渐分裂的规律；同时还发现，SIGIR邮件列表研究主题较会议论文而言，在时序上存在一定的“领先性”，通过该研究旨在揭示SIGIR邮件列表在信息检索领域的学术价值。

关键词：领域主题；主题结构；主题演化；共词分析；SIGIR邮件列表
中图分类号：G353.4
DOI：10.3772/j.issn.1673-2286.2017.06.007

## 1 引言

电子邮件作为一种较正式的交流方式，常被科研人员作为问题交流和科研合作的介质。在开源软件开发过程中，众多开发者通常使用电子邮件进行沟通。国外已有很多学者对开源软件的邮件列表展开相关研究，包括社会网络挖掘和内容挖掘等。

Ducheneaut利用社会网络分析法对开源软件邮件列表中的人物关系网络进行挖掘[1]，将开源软件社区人物分成核心开发者、维护者、Bug修复者、Bug报告者、用户文档撰写者和用户。Elsayed等对Enron Collection邮件列表进行挖掘，揭示邮件列表中人物间社会关系[2]。Bird等认为开源软件开发过程的每个阶段都包含一个由核心开发人员组成的小组[3]，并在后续研究中针对其他开源软件的邮件列表进行了社会网络分析[4-6]。

SIGIR（Special Interest Group on Information Retrieval）是信息检索领域的顶级国际学术会议，自1963年，SIGIR一直专注于信息搜索和信息获取技术的研究和教育。其主办方通过邮件发布会议通知，网站所有注册者均收到邮件。SIGIR官网保存了2007年10月—2017年2月的邮件列表。受国外开源软件邮件列表研究成果启发，本文认为SIGIR邮件列表对揭示信息检索领域的发展具有重要意义。本文希望通过对SIGIR邮件列表和同期会议论文进行比较，揭示信息检索领域主题结构和主题演化趋势。

目前，对于领域主题结构和主题演化研究的方法主要有词频分析法和共词分析法。词频分析法是通过分析领域内主题词历年走势来揭示领域研究主题的演化情况；共词分析法是通过构建主题词共现矩阵，进行主题词聚类分析。共词分析法在很多学科主题研究中得到应用，如高聚物化学[7]、信息检索[8]、软件工程[9]、生物医学[10-11]、图书情报[12-13]等。

## 2 研究设计

数据是研究的基石，因此本文构建SIGIR邮件列表和会议论文的数据集。具体包括三方面。（1）邮件数据获取。利用网络爬虫抓取SIGIR官网的邮件列表，获得邮件列表数据。（2）确定会议论文。获取同期发表在SIGIR会议和ECIR（European Conference on

* 本研究得到国家自然科学基金面上项目“面向词汇功能的学术文本语义识别与知识图谱构建”（编号：71473183）资助。

Information Retrieval）会议上的论文。（3）关键词抽取。从邮件列表的正文中对关键词进行抽取。统计发现，SIGIR、ECIR会议论文平均每篇文章有3.5个关键词，相对于邮件列表来说，会议论文的关键词数量过少，因此，本文从会议论文摘要中抽取部分词汇作为对原有关键词的补充。

利用共词分析法对信息检索领域的主题结构和主题演化进行研究。陈必坤等将学科知识网络的研究单元分为作者、机构、城市、国家/地区、专业术语（来自标题、摘要和关键词）、引文等[14]。本文选取关键词作为信息检索领域主题研究的研究单元。

本文的原始数据包括两部分：第一部分数据来自SIGIR官网的邮件列表，包含2007年12月—2017年2月的所有邮件数据，考虑到2007年和2017年数据不全，因此剔除2007年和2017年的邮件数据，只选取2008—2016年的邮件数据作为研究对象；第二部分数据为2008—2016年SIGIR和ECIR会议的会议论文。

从SIGIR官网的邮件列表共得到7 419封邮件，其中2008年1月—2016年12月共计7 218封，邮件格式相对统一，一般包括标题（header block）、正文（content block）和脚注（footer block）。

通过调研，发现SIGIR邮件列表中涉及的会议主要有SIGIR、ECIR、ICTIR（ACM International Conference on the Theory of Information Retrieval）、NTCIR（NACSIS Test Collections for IR）等。但由于未能获取到ICTIR和NTCIR会议论文的数据。因此，本文以SIGIR和ECIR会议论文作为对比研究对象。最终获取SIGIR会议论文1 968篇和ECIR会议论文828篇。

## 3 信息检索领域主题结构研究

### 3.1 信息检索领域关键词的抽取和选择

本文采用基于词表的方法对邮件列表邮件正文和会议论文摘要进行关键词抽取，该词表包含领域内8万多条相关关键词，关键词覆盖面广。

由于从摘要中抽取出的关键词数量巨大，若不对其进行筛选，则会造成构造的共词矩阵过大，给分析带来困难，造成维度灾难。TF-IDF是信息检索和数据挖掘常用的一种词语加权技术，因此本文考虑使用TF-IDF算法进行关键词筛选。

### 3.2 信息检索领域高频关键词分析

利用TF-IDF算法选取邮件列表和会议论文高频词进行聚类分析，在选取热点关键词时人工剔除与主题关联度较小的关键词，合并意义相近的关键词，最终选取在邮件列表数据集和会议论文数据集中出现频率排名前40的关键词，其中部分高频关键词如表1和表2所示。

表 1 邮件列表高频关键词（部分）

| 关键词 | 词频/次 |
|---|---|
| Social Media | 3 490 |
| Social Network | 3 162 |
| Recommender System | 2 888 |
| Machine Learning | 2 674 |
| Digital Libraries | 2 617 |
| Data Mining | 2 374 |
| Text Mining | 2 182 |
| Natural Language Processing | 2 046 |
| Artificial Intelligence | 1 337 |
| Sentiment Analysis | 1 237 |

表 2 会议论文高频关键词（部分）

| 关键词 | 词频/次 |
|---|---|
| Learning to Rank | 171 |
| Language Model | 162 |
| Information Needs | 138 |
| Recommender System | 134 |
| Query Log | 130 |
| User Study | 122 |
| Relevance Judgements | 120 |
| Retrieval Performance | 117 |
| Social Network | 117 |
| Retrieval Effectiveness | 98 |

### 3.3 信息检索领域热点关键词相异矩阵的构建

为分析信息检索领域的研究主题情况，对热点关键词进行聚类。将选取的40个关键词分别在两个数据集上构建共现矩阵，生成两个“40×40”的领域主题关

键词共现矩阵。单一的词频信息对反映关键词间的联系存在一定局限性，高频关键词与其他关键词共现的概率需大于低频关键词，为消除这种由词频带来的影响，需要构造共现矩阵[15]。本文采用*Equivalence*系数将共现频次转换成“[0，1]”的相似矩阵。关键词*A*和*B*的*Equivalence*系数计算方法如下。

$$E=\frac{F_{AB}^2}{F_A\times F_B}$$

其中$F_{AB}^2$是关键词*A*和关键词*B*在文档集中共现的次数平方，$F_A$，$F_B$分别代表关键词*A*和关键词*B*的词频[16]。*E*代表关键词*A*和*B*的*Equivalence*系数，*E*值越大代表关键词间的关联性越强。

通过分析，发现大多数*E*值较低，因此本文考虑将相似矩阵转换成相异矩阵，即用1减去相似矩阵中的值得到相异矩阵。本文利用Java语言自编程序计算热点关键词的相异矩阵，具体如表3和表4所示（篇幅有限，仅列举部分）。与相似矩阵不同，相异矩阵中的值越大，关键词间的关联性越弱，反之亦然[17]。

表 3 邮件列表热点关键词相异矩阵（部分）

| | Social Media | Social Network | Recommender System | Digital Libraries | Data Mining |
|---|---|---|---|---|---|
| Social Media | 0.00 | 0.87 | 0.92 | 0.93 | 0.93 |
| Social Network | 0.87 | 0.00 | 0.89 | 0.92 | 0.90 |
| Recommender System | 0.92 | 0.89 | 0.00 | 0.94 | 0.95 |
| Digital Libraries | 0.93 | 0.92 | 0.94 | 0.00 | 0.96 |
| Data Mining | 0.93 | 0.90 | 0.95 | 0.96 | 0.00 |

表 4 会议论文热点关键词相异矩阵（部分）

| | Learning to Rank | Language Model | Information Needs | Query Expansion | Recommender System |
|---|---|---|---|---|---|
| Learning to Rank | 0.00 | 0.99 | 0.99 | 0.99 | 0.99 |
| Language Model | 0.99 | 0.00 | 0.99 | 0.99 | 1.00 |
| Information Needs | 0.99 | 0.99 | 0.00 | 0.99 | 0.99 |
| Query Expansion | 0.99 | 0.99 | 0.99 | 0.00 | 1.00 |
| Recommender System | 0.99 | 1.00 | 0.99 | 1.00 | 0.00 |

### 3.4 邮件列表和会议论文热点关键词聚类分析

为进一步揭示不同关键词间的关联性，需对关键词进行聚类分析，将关联性较强的关键词聚成一个词簇，相同词簇内部的关键词具有较强的关联性，不同词簇间关键词的相异性较大。将上述生成的两个相异矩阵导入SPSS进行聚类分析，得出不同方式下的研究主题。

从邮件列表聚类结果来看，其研究主题主要分布在七个方面：（1）情感分析和意见挖掘的主题词包括“Sentiment Analysis”“Opinion Mining”；（2）自动问答的主题词包括“Natural Language Processing”“Computational Linguistics”“Question Answering”“Machine Translation”“Information Extraction”；（3）数字图书馆和交互式信息检索的主题词包括“Digital Libraries”“User Study”“User Interfaces”“Recommender System”“User Modeling”“Collaborative Filtering”；（4）人工智能与人机交互的主题词包括“Artificial Intelligence”“Human-Computer Interaction”；（5）社会网络和文本挖掘的主题词包括“Social Media”“Social Network”“Machine Learning”“Data Mining”“Text Mining”“Knowledge Discovery”；（6）机器学习排序和自然语言处理的主题词包括“Learning to Rank”“Language Model”“Topic Model”“Semantic Technologies”“Content-Based Recommendation”“Named Entity Recognition”“Named Entities”；（7）深度学习和图像检索的主题词包括“Image Retrieval”“Multimedia Retrieval”“Deep Learning”。

从会议论文聚类结果看，其研究主题主要分布在

七个方面：（1）推荐系统的主题词包括“Recommender System”“Collaborative Filtering”；（2）社会网络和社交媒体的主题词包括“Social Media”“Social Network”；（3）检索模型与评价的主题词包括“Evaluation Metrics”“Implicit Feedback”“Evaluation Measures”“Learning to Rank”“Ranking Model”“Machine Learning”；（4）文本挖掘和自然语言处理的主题词包括“Sentiment Analysis”“Named Entities”“Question Answering”“Knowledge Base”“Text Mining”“Retrieval Model”“Document Ranking”“Experimental Results”；（5）相关反馈的主题词包括“Relevance Judgements”“Retrieval Effectiveness”“Language Model”“Document Retrieval”“Query Expansion”“Pseudo-Relevance Feedback”“Retrieval Performance”；（6）用户为中心/交互式信息检索的主题词包括“User Behavior”“Click Model”“User Study”“User Satisfaction”“Query Log”“Query Suggestion”“Search Behavior”“User Interaction”“Information Needs”；（7）图像检索的主题词为“Image Retrieval”。

从列表和会议论文的主题词聚类结果可以看出，两个数据集聚类结果中有很多相似主题，如自然语言处理、文本挖掘、社会网络、社交媒体、交互式信息检索等。这说明高频主题词在两个数据集上的分布存在一定的相似性，但也存在一定差别。如自动问答、数字图书馆、人工智能、深度学习等主题词出现在邮件列表数据集的聚类结果中，但未在会议论文数据集的聚类结果中出现；检索模型与检索评价出现在会议论文数据集聚类结果中，但并未在邮件列表数据集的聚类结果中出现，具体如表5所示。

表 5 邮件列表和会议论文研究主题异同

| 数据集 | 不同主题 | 相同主题 |
|---|---|---|
| 邮件列表 | 情感分析、意见挖掘、自动问答、数字图书馆、人工智能、人机交互、机器学习排序、深度学习 | 社会网络、文本挖掘、自然语言处理、交互式信息检索、图像检索 |
| 会议论文 | 推荐系统、检索模型、检索评价、相关反馈 | |

## 4 信息检索领域主题演化研究

### 4.1 信息检索领域主题演化网络分析

受地理学冲积图影响，Rosvall等提出一种社区演化分析方法[18]，可用于研究主题演化，但该方法不能反映主题在当前时间段的活跃程度。王晓光等对此方法进行了改进，其通过对主题进行排序，将排名靠前的主题放在图形顶端，并在此基础上开发了一款学科主题演化可视化工具NEViewer[19]，该工具以时间为维度根据关键词共现关系绘制冲积图来表示领域主题的演化。

本文利用NEViewer对信息检索领域的主题演化进行可视化分析。从两个数据集中分别选取前2 000个高频关键词作为共现网络中的节点。将时间划分成三个阶段：第一个阶段为2008—2010年，第二个阶段为2011—2013年，第三个阶段为2014—2016年。在各时间段构建高频词共现矩阵并导入NEViewer，绘制以邮件列表为主题的信息检索研究主题演化冲积图（见图1）。

由图1可见，社交网络、推荐系统、文本挖掘、机器学习等主题拥有较高的中心度；同时，也可以看出社交网络、推荐系统、文本挖掘和机器学习持续处于图形比较靠近顶端的位置，说明社交网络、推荐系统、文本挖掘和机器学习是近年来的研究热点。

同样方法对会议论文数据集绘制如图2所示的主题演化冲积图。

由图2可知，机器学习排序、语言模型、查询扩展和推荐系统等主题拥有较高的中心度，是信息检索领域会议论文的研究热点。总体来看，信息检索领域研究主题演化存在以下规律。

（1）信息检索研究内容不断深入。传统信息检索研究主题主要集中在文档表示、查询分析、检索模型、检索结果排序和检索结果评价等。随着Web 2.0的发展和以Twitter、Facebook为代表的社交网络的兴起，社交网络、推荐系统正成为信息检索领域的研究热点，对传统研主题的研究正逐渐减少。

（2）信息检索研究方法不断增多。随着神经网络、深度学习、自然语言处理、人工智能等研究被引入信息检索领域，信息检索的研究方法更加丰富和多元化。

（3）信息检索存在核心主题演化现象。从邮件列表反映的研究主题演化来看，机器学习主题分裂出决

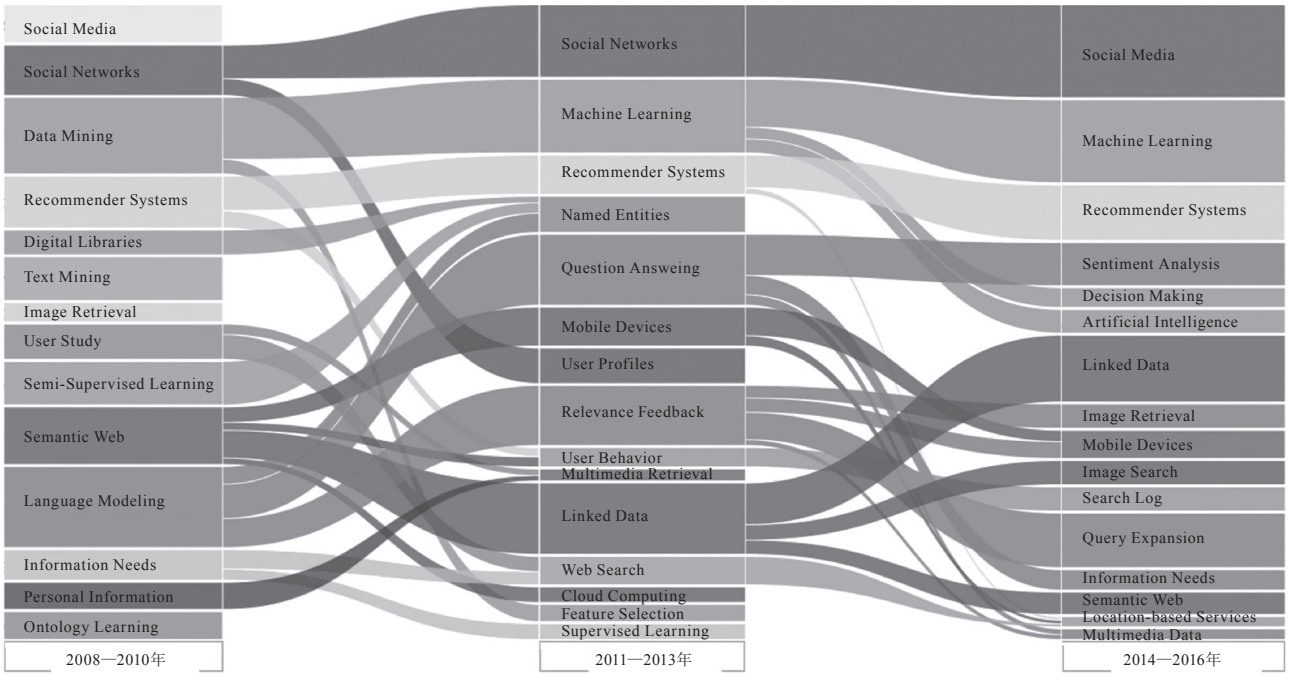


图 1 邮件列表主题演化冲击图

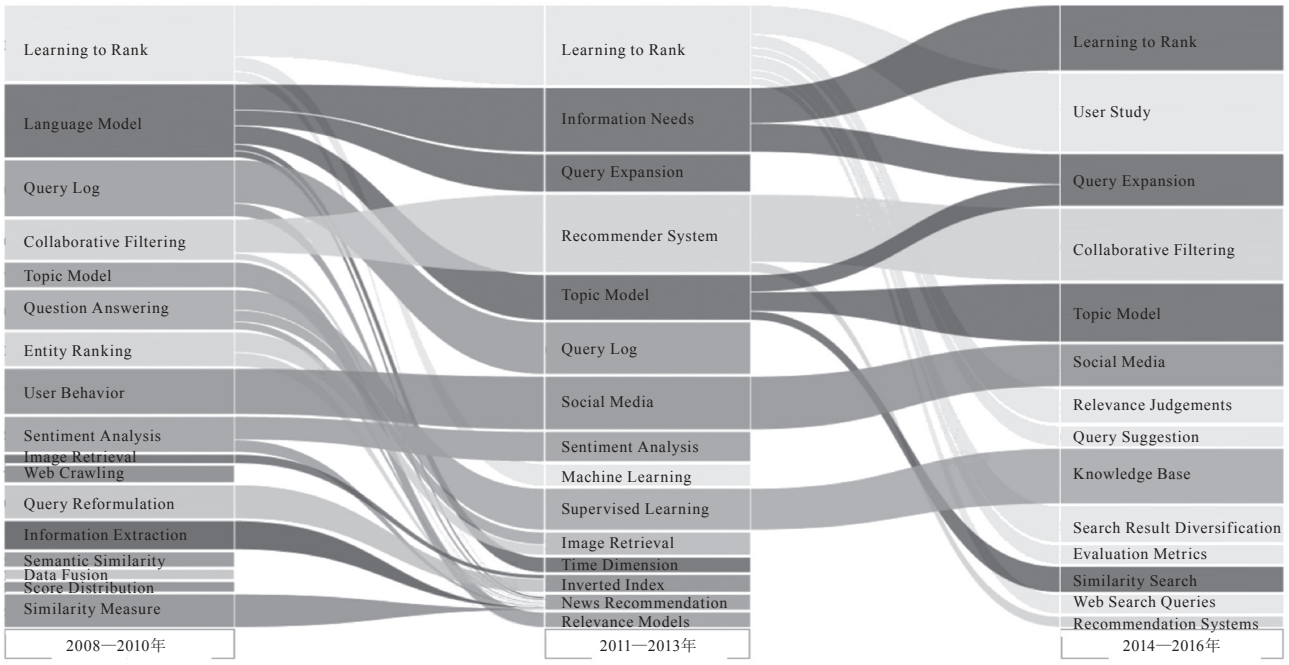


图 2 会议论文主题演化冲击图

策支持和人工智能；会议论文中机器学习排序分裂出相关性判别、查询建议、查询结果多样化等主题。但是，没有明显的主题合并现象出现。这说明随着研究方法的多样化，信息检索领域的研究更加精细化和专业化。另外，部分主题没有后续研究，这说明部分信息检索研究主题存在消亡现象。

## 4.2 信息检索领域主题词-时间共现网络分析

为研究主题词2008—2016年分布情况，进一步构建主题词-时间的共现网络，将40个主题词分别在两个数据集上构建主题词-时间共现网络。利用Ucinet子群分析法，将主题词按时间分成9个子类别，具体如表6和表7

所示。

通过对邮件列表的主题词-时间共现网络和会议论文主题词-时间共现网络的对比分析，可以发现邮件列表在8个研究主题存在“领先”会议论文研究主题的现象。如“Collaborative Filtering”在邮件列表中是2008年的代表性主题词，但在会议论文中是2015年的代表性主题词；“Social Network”在邮件列表中是2008年的代表性主题词，而在会议论文中是2010年的代表性主题词；“Learning to Rank”是邮件列表2014年的代表性主题词，而在会议论文数据集中是2016年的代表

表 6 邮件列表2008—2016年代表性主题词

| 时间/年 | 主题词 |
|---|---|
| 2008 | Machine Translation；Natural Language；Collaborative Filtering；Social Network；Information Extraction |
| 2009 | Information Access；Recommender System；Computational Linguistics；Named Entity Recognition；Knowledge Base |
| 2010 | Human-Computer Interaction；Opinion Mining；Multimedia Retrieval；Machine Learning |
| 2011 | Digital Libraries；Question Answering；User Interaction；Semantic Technologies |
| 2012 | Data Mining；Topic Model；Knowledge Management；Knowledge Discovery |
| 2013 | Text Mining；Deep Learning；User Modeling；Content-Based Recommendation |
| 2014 | Natural Language Processing；Image Retrieval；Learning to Rank；Named Entities；Image Retrieval |
| 2015 | User Interfaces；Language Model；Artificial Intelligence；Social Computing；User Study |
| 2016 | Social Media；User Model；Relevance Feedback；Sentiment Analysis；Part-of-Speech Tagging |

表7 会议论文2008—2016年代表性主题词

| 时间/年 | 主题词 |
|---|---|
| 2008 | Pseudo-Relevance Feedback；Named Entities；Language Model；Relevance Feedback；Machine Learning |
| 2009 | Question Answering；Document Ranking；Retrieval Performance；Information Needs |
| 2010 | Query Expansion；Document Retrieval；Social Network；Latent Dirichlet Allocation；Click Model |
| 2011 | Ranking Model；Evaluation Metrics；Recommender System；Retrieval Effectiveness |
| 2012 | Implicit Feedback；Relevance Feedback；Topic Model；User Behavior |
| 2013 | Evaluation Measures；Experimental Results；Query Log；Search Behavior |
| 2014 | Image Retrieval；User Interaction；Social Media；Text Mining |
| 2015 | Collaborative Filtering；Knowledge Base；Query Reformulation；Query Suggestion；Active Learning |
| 2016 | User Satisfaction；User Study；Learning to Rank；Retrieval Model；Sentiment Analysis |

主题词。主题词“User Interaction”“Recommender System”“Text Mining”“User Study”等在邮件列表和会议论文中也存在类似现象。因此，SIGIR邮件列表研究主题较会议论文而言，在时序上存在一定的领先性。

## 5 结语

本文以SIGIR邮件列表为切入点分别构建邮件列表和会议论文数据集。提出将SIGIR邮件列表作为信息检索领域研究对象，通过与同期会议论文进行比较分析，证明SIGIR邮件列表作为研究主体的价值，这是信息检索领域研究对象的创新。但该研究过程存在一定的局限性。

（1）SIGIR邮件列表的会议通知包含众多信息检索相关会议（如ICTIR、NTCIR等），但由于数据获取方面的原因，本文在构建会议论文数据集过程中未将这些会议论文纳入，因此分析结果可能存在一定的片面性。

（2）在关键词选取时，首先利用TF-IDF算法初步识别了一些高频关键词，其次对一些意义相近的关键词进行合并，人工剔除与主题关联度较小的关键词。在关键词合并和剔除过程中难免存在一定主观性，因此

分析结果可能存在一定的局限性。

## 参考文献


[1] DUCHENEAUT N.Socialization in an open source software community:a socio-technical analysis[J].Computer Supported Cooperative Work,2005,14(4):323-368.

[2] ELSAYED T,OARD D W.Modeling identity in archival collections of email:a preliminary study[C]//Ceas 2006-the 3rd Conference on Email and Anti-Spam,July 27-28,2006,Mountain View,California. DBLP,2006:95-103.

[3] BIRD C,GOURLEY A,DEVANBU P,et al.Open borders? Immigration in open source projects[C]//International Conference on Software Engineering Workshops.[S.1.]:IEEE Computer Society,2007:6.

[4] BIRD C,GOURLEY A,DEVANBU P,et al.Mining email social networks in Postgres[C]//International Workshop on Mining Software Repositories.[S.1.]:ACM,2006:185-186.

[5] HONG Q, KIM S,CHEUNG S C,et al.Understanding a developer social network and its evolution[C]//IEEE International Conference on Software Maintenance.[S.1.]:IEEE,2011:323-332.

[6] BIRD C,PATTISON D,D'SOUZA R,et al.Latent social structure in open source projects[C]//ACM Sigsoft International Symposium on Foundations of Software Engineering,November Atlanta, Georgia,2008.[S.1.]:DBLP,2008:24-35.

[7] CALLON M,COURTIAL J P,LAVILLE F.Co-Word analysis as a tool for describing the network of interactions between basic and technological research:the case of polymer chemsitry[J].Scientometrics,1991,22(1):155-205.

[8] DING Y,CHOWDHURY G G,FOO S.Bibliometric cartography of information retrieval research by using co-word analysis[J]. Information Processing & Management,2001,37(6):817-842.

[9] COULTER N,MONARCH I,KONDA S.Software engineering as seen through its research literature:a study in co-word analysis[J]. Journal of the Association for Information Science and Technology, 1998,49(13):1206-1223.

[10] 郑华川,于晓欧,辛彦.利用共词聚类分析探讨抗原CD44研究现状[J].中华医学图书情报杂志,2002,11(2):1-3.

[11] 郑华川, 崔雷. 胃癌前病变低频被引论文的共词和共篇聚类分析[J].中华医学图书情报杂志,2002,11(3):1-3.

[12] 马费成,望俊成,陈金霞,等.我国数字信息资源研究的热点领域:共词分析透视[J].情报理论与实践,2007,30(4):438-443.

[13] 洪凌子,黄国彬,于洋.基于CiteSpace的国内外数字图书馆研究论文的比较分析[J].图书馆论坛,2014(6):91-100.

[14] 陈必坤,王曰芬.学科结构与演化可视化分析的内容研究[J].图书情报工作,2016,60(21):87-95.

[15] 姚强,张士靖.国际健康素养研究热点与前沿文献计量分析[J].中国健康教育,2012,28(1):36-39.

[16] 李信,李旭晖,陆伟.大数据驱动下的图书情报学科热点领域挖掘——面向WOS题录数据的实证视角[J].图书馆论坛,2017(4):49-57.

[17] 段春雨,蔡建东.我国教育信息化研究热点知识图谱——基于2003—2013年硕士及博士学位论文的关键词分析[J].华北水利水电大学学报(社会科学版),2015,31(1):129-131.

[18] ROSVALL M,BERGSTROM C T.Mapping change in large networks[J].Plos One,2010,5(1):e8694.

[19] 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.


## 作者简介


赵忠伟，男，1990年生，硕士研究生，研究方向：信息检索、知识挖掘，E-mail：2009302330014@whu.edu.cn。
程齐凯，男，1989年生，博士研究生，研究方向：信息检索、机器学习，E-mail：chengqikai0806@163.com。


Research on the Subject of Information Retrieval: A Comparative Study Based on SIGIR Mailing List and Conference Papers


ZHAO ZhongWei, CHENG QiKai
(School of Information Management, Wuhan University, Wuhan 430072, China)



Abstract: Traditional discipline topic research is mainly based on the scholar papers, but the research object is too stereotyped. In this paper, the SIGIR mailing list dataset and the conference paper dataset are constructed with the SIGIR mailing list as the starting point, and the discipline topic structure and discipline topic evolution of information retrieval are compared and analyzed in two datasets respectively. We found that the research content of information retrieval is deepening, and the research methods continue to flourish, and the core research topics are splitting gradually, at the same time, we found that the SIGIR mailing list's research topics keep ahead than conference papers'. The academic value of the SIGIR mailing list in the discipline topic research is revealed.



Keywords: Discipline Topic; Topic Structure; Topic Evolution; Co-word Analysis; SIGIR Mailing List

（收稿日期：2017-04-28）