情报学报　2017 年 9 月　第 36 卷　第 9 期
Journal of the China Society for Scientific and Technical Information, Sep. 2017, 36(9): 940-953

中英文科技主题排序相关性的比较研究：以计算机领域为例

蒋卓人[1]，高良才[1]，赵　星[2]，刘晓钟[3]，袁　珂[1]，汤　帜[1]

（1. 北京大学计算机科学技术研究所，北京　100871；2. 华东师范大学经济与管理学部信息管理系，上海　200241；3；印第安纳大学布鲁明顿分校信息和图书馆科学系，布鲁明顿　47405）

摘　要　主题排序是信息检索、信息组织等情报学核心领域的重要问题，本研究从静态全局角度和动态时间维度上对这一问题进行综合性探索，尝试以出现频率、被引次数、PageRank 值等文献计量学参量为基础，重点对中英文科技主题在各自数据集中的排序相关性进行比较研究。本文的实验研究以 ACM（Association for Computing Machinery，美国计算机学会）英文论文数据集和万方数据知识服务平台抽取的中文论文数据集为例，揭示和比较了不同排序方式下和不同时间阶段中的中英文科技主题排序的相关关系以及动态变化特征。实验结果显示，在不同的时间阶段，使用多种排序方法的中英文科技主题排序均存在正相关关系，并且某一历史阶段的英文科技主题排序，随着时间的发展，和不同历史阶段的中文科技主题排序的正相关关系成增强趋势，而不同的排序方式会使得这种正相关增强趋势的作用时间范围发生变化。

关键词　主题排序；比较研究；可视化

Comparative Study of Chinese-English Scientific Topic Ranking Correlations: Computer Science Perspective

Jiang Zhuoren[1], Gao Liangcai[1], Star Zhao[2], Liu Xiaozhong[3], Yuan Ke[1] and Tang Zhi[1]

(1. *Institute of Computer Science & Technology, Peking University, Beijing*　100871;
2. *Department of Information Management, East China Normal University, Shanghai*　200241;
3. *Department of Information and Library Science, Indiana University Bloomington, Bloomington*　47405)

Abstract: Topic ranking is an important issue for information retrieval and organization, and in this study, we conducted a comprehensive exploration of this issue from static global and dynamic historical perspectives, focusing on a comparative study of Chinese-English scientific topic rank correlations in each dataset, which was based on bibliometrics such as the frequency, number of citations, and PageRank value. In the experiment, which was conducted on an English dataset extracted from the Association for Computing Machinery (ACM) and a Chinese dataset extracted from the Wanfang knowledge platform, the correlation relationships of Chinese and English topic rankings were revealed and compared using different ranking methods in different historical stages, as well as a dynamic change in the features of different correlation relationships. The experiment results show that, at the different historical stages, the Chinese and English topic rankings using different ranking methods all have positive correlations, and that an English

收稿日期：2016-08-09；修回日期：2017-03-29

基金项目：中国博士后科学基金面上项目“面向需求变迁的异构学术信息推荐技术研究”（2016M590019），国家自然科学基金项目“版式文档中复杂异构对象的识别技术研究”（61573028）。

作者简介：蒋卓人，男，1986 年生，博士，博士后，主要研究方向为数字图书馆、信息检索、学术资源推荐；高良才，男，1980 年生，博士，副教授，通讯作者，主要研究方向为数字图书馆、文档分析与理解、信息抽取；赵星，男，1983 年生，博士，研究员，博士生导师；刘晓钟，男，1979 年生，博士，助理教授；袁珂，女，1992 年生，博士研究生；汤帜，男，1965 年生，博士，研究员，博士生导师。

topic ranking in a certain historical stage, with the development of time, will have increasing positive correlations with the Chinese topic rankings. Different ranking methods will change the time range of this increasingly positive correlation effect.

Key words: topic ranking; comparative study; visualization

1　引　言

随着科学研究范围的不断扩展和深入，科研工作者们正以惊人的速度创造出海量学术文献，尤其是进入 21 世纪后，学术文献数目的增加更是呈现出一种指数级的增长态势[1]。在此背景下，科研工作人员除了需要对自身专业知识有着更为深入的理解，也需要具有跨越多个研究主题的广博知识储备[2]。然而，面对主题多样的海量文献资源，科研人员很难高效地获取某一学科领域的研究热点和趋势、把握研究主题的演化情况，从而无法准确地弥补研究过程中的不足以及发现可能的新知识增长点，更进一步地，也会对科研管理和决策部门有效开展业务决策、研究布局和科研扶持等工作造成一定困难[3]；相较于以文献篇章为基本研究单元的传统科学计量学研究手段，以科技主题作为基本信息单元的研究能够实现对文献粒度更小、层次更深、更全面的分析和研究[4]。

目前，国内外已经有大量的关于科技主题的研究方法，既有词频分析、引文分析等传统的科学计量学的研究方案[5-8]，又有以概率主题模型为代表的机器学习、信息检索和文本挖掘方法[9-11]等，这些方法都能取得良好的研究效果，且存在以下特点：

（1）研究大都针对单一语种，对不同语种的科技主题之间的比较研究不足。事实上，对于类似中国这样的非英语母语国家而言，众多领域的科研人员在进行研究领域文献与主题检索、发现和分析时，通常需要同时考虑英语和当地母语两种数据源。目前较为成熟的多语言文本聚类方法[12]，对于本文探索不同语种的科技主题具有参考意义。

（2）对文献的排序研究较为成熟，但对主题排序的研究仍在发展中。而实际上主题的排序对于专业学术领域的研究热点把握和研究趋势的探索有着重要的作用。在信息检索领域，利用主题排序改进检索结果的研究并不鲜见，研究已经证明可以通过主题聚类技术提高文档检索的准确率和效率[13-14]。本文对不同语种主题排序相关性的研究，是对主题排序的有关研究的进一步延伸和探索。

（3）传统的科技主题研究往往针对数据集进行全局性的静态研究，忽视了时间要素的重要性，已经有越来越多的研究[15-18]开始从时间动态角度对科技主题进行研究。

本文通过对 ACM（Association for Computing Machinery，美国计算机学会）论文数据集和万方数据知识服务平台计算机专业学位论文数据集进行分析，对计算机领域的**中外科技主题排序**进行了**跨语言**定量比较研究，不仅从全局角度进行了**静态比较研究**，而且从时间维度上进行了**动态比较研究**。

本文的主要结果包括：

◆在多种排序方法下，中文科技主题排序和英文科技主题排序均存在**正相关关系**，并且不同的排序方式对中英文科技主题排序的正相关关系会产生影响。

◆**在不同的时间阶段**，在多种排序方式下，中文科技主题排序和英文科技主题排序均存在**正相关关系**。

◆某一历史时期的英文科技主题排序，**随着时间的发展**，和不同历史阶段的中文科技主题排序的**正相关关系成增强趋势**，并且不同的排序方式会使得这种**正相关增强趋势的作用时间范围发生变化**。

2　相关研究评述

针对学术文献中的科技主题，国内外已取得了一定的成果，特别是运用情报计量学方法进行主题研究的工作具有特色。例如，马费成等[5]利用关键词确定科技主题，用词频分析的方法，对比分析了国内外知识管理的科技主题及其相关信息；王林等[6]改进了作者同被引分析方法，提出施引关键词与被引作者的交叉共现分析，从而能够更为清晰地发现科学领域中的学术流派、重要研究人员及相关主题领域；Ozcinar[7]采用引文分析方法对结构设计领域的科学主题的趋势和演变进行了研究；Bharat 等[8]巧妙地利用链接分析来寻找和分析主题，从而有效定位到高质量文档；此外，引文分析等计量学方法的使用也在不断深化，陈仕吉等[19]提出了结合 C-value 和 TF-IDF 算法的文献簇主题识别方法，该方法可以更好地应用于引文分析中文献簇的主题识别。这些研究充分表明，诞生于图书情报学的文献计量及相关方法在主题分析中具有重要的价值与发展前景。

对于主题分析的另一类大类研究来源于计算机学科，并侧重模型与算法探索。章成志等[20]采用数据挖掘中的聚类算法，对主题信息资源进行聚类，从而为用户打开了一个新的信息资源浏览维度和视角。概率主题模型（Probabilistic topic models）是近年来在机器学习和统计学领域中被广泛研究和应用的新技术，这是一系列旨在发现和标记大规模文档的主题信息的算法[9-10]。Ramage 等[11]采用语料库中的元数据（如学术文献的关键词）对主题概率分布赋予标签，提高了主题的可解释性。范云满等[21]对利用主题模型方法进行主题探测的国内外研究进行了综述。相对于传统计量学手段，概率主题模型的计算成本较高，耗费时间较长，对语料库的文档数量有一定要求，且目前也尚未发现可以直接应用于主题排序的相关模型。

虽然关于科技主题的分析已有一定积累，但针对主题排序的研究尚处起步阶段。徐小龙等[22]对关键词进行排序，研究云计算领域的热点科技主题，但该研究主要针对中文研究，没有涉及国内外研究的对比；Wang 等[23]采用媒体聚焦和用户注意力的方式对主题进行排序，但该研究主要针对的是新闻主题而非科技学术主题；Cutting 等[13]和 Hearst 等[14]的研究通过文档聚类技术，把用户的信息需求和相关文档通过主题相关联，从而提高文档排序的准确率和效率；Shubhankar 等[24]研究了学术论文语料库中的科技主题排序和演化问题，该研究针对的是英文文献，没有进行跨语言的研究；Shuai 等[25]把学术主题和社交媒体中的对应主题进行比较研究，但也仅仅针对英文数据集开展实验和分析。章成志等[12]对多语言的文本聚类做了深入研究，综述了目前主要的多语言文本聚类方法，对于本文探索不同语种的科技主题有着重要的参考意义。

在主题的排序中，时间维度是值得密切关注的分析视角。沈洪洲等[15]对文献关键词进行抽取和统计，运用可视化工具展现科技主题随年份的演化情况，进而分析研究主题的发展趋势；章成志等[18]采用主题聚类方法，对包括时间信息的论文全文进行主题分析与主题聚类，归纳出某一特定学科的研究热点和趋势；动态主题模型[16]和有监督的动态主题模型[17]是对传统主题模型在时间维度上的扩展，但这两个模型的计算和训练非常耗时，很难直接在大数据量的数据集上使用；Jiang 等[26]利用科技主题在不同历史阶段的变化实现了基于时序的引文推荐；赵迎光等[27]对基于主题模型的演化方法进行综述，按照难易程度对相关模型进行分类并总结了各种模型的优缺点，讨论了结合时间要素的主题演化模型在情报分析领域的适用性。

综上所述，目前对科技主题的研究方法多种多样，既有情报学的计量学方法也有以概率主题模型为代表的机器学习方法，均可以取得良好的效果，但由于概率主题模型计算代价和资源消耗相对过高，且目前也没有成熟的可直接用于主题排序的主题模型；而文献计量学手段容易理解，便于操作，且效率较高，因此，本文采用基于出现频率、被引次数和引用关系网络分析的研究方法对科技主题的重要度进行度量，并在此基础上进行排序分析研究；本文针对不同语种，进行跨语言比较研究；并且在时间维度上进行扩展，进行了动态比较研究，具体的研究方法将会在下面的章节中进行详细阐述。

3 问题定义

本文的研究总体框架如图 1 所示，其中，**P** 是学术论文集合，P_C 是中文学术论文，P_E 是英文学术论文。T 是一系列的科技主题，和研究文献[5,11,15,17,22,25-26]类似，本文采用关键词来表示科技主题，其中，T_C 是中文科技主题，T_E 是英文科技主题。采用关键词作为学术主题的优势主要有：

◆关键词是由大量不同的作者提供，对文章所涵盖主题的概括更为准确和全面，由此确定的学术主题数目也更为自然、合理。

◆由关键词表示主题，主题在语义表现层面上也更有解释性。

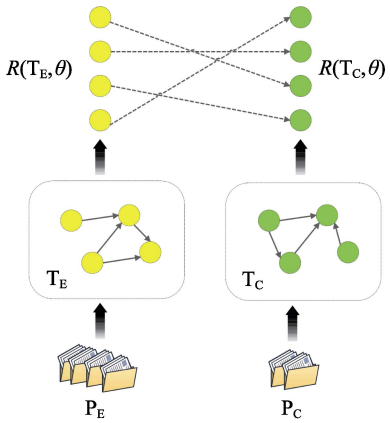


图 1 研究框架

为了清晰论述，对本文的关键概念进行定义说明如下：

[主题引用关系] $C_T=\{(t_i,t_j)\mid t_i\text{引用}t_j,t_i,t_j\in T\}$，

其中，主题引用关系可以由论文的引用关系确定：如果论文 p_a 引用了论文 p_b，且 t_i 是 p_a 的主题、t_j 是 p_b 的主题，则主题 t_i 引用了主题 t_j。

[主题引用网络] $\mathrm{G_T}=\{\mathrm{T},\mathrm{C_T},\mathrm{F_T}\}$，表示一个有向带权重网络，其中，$\mathrm{F_T}$ 是一个权重函数，为每一条边 (t_i,t_j) 赋予一个正整数值 $f(t_i,t_j)\in\mathrm{N}^+$，该正整数值表示从 t_i 到 t_j 的引用数目。

[排序] $R(X,\theta)$，能够把一组随机变量 $X=\{x_1,x_2,\cdots,x_i,\cdots\}$ 转换为一个有序的组合 X^*，其中 $\theta(\cdot)$ 表示排序方式，若 $\theta(x_i)>\theta(x_j)$，则 x_i 比 x_j 排序靠前，不同的排序方式能够生产不同的有序组合。本文主要研究三种不同的排序方式：

（1）$\theta_f(x_i)$，即随机变量 x_i 的出现频率（Frequency）。

（2）$\theta_c(x_i)$，即随机变量 x_i 的被引用次数（Cited Number）。

（3）$\theta_p(x_i)$，即随机变量 x_i 在引用关系网络中的 PageRank 值。

[主题对应映射] $\mathrm{TR}=\{(t_i,t_j)\,|\,t_i=\mathrm{translate}(t_j)\,\&\,\&t_j=\mathrm{translate}(t_i),t_i,t_j\in\mathrm{T}\}$ 表示不同语种科技主题经过机器翻译[28]之后的对应关系，其中，t_i、t_j 属于不同语种的科技主题集。TR 具有对称性，当且仅当 t_i 和 t_j 能够被互相翻译对应时，主题对应映射关系才成立。

本文的主要目标是研究中文科技主题排序 $R(\mathrm{T_C},\theta)$ 和英文科技主题排序 $R(\mathrm{T_E},\theta)$ 之间的关系。具体而言，本文主要探索以下三个问题：

（1）中文科技主题排序 $R(\mathrm{T_C},\theta)$ 和英文科技主题排序 $R(\mathrm{T_E},\theta)$ 之间的相关性如何？

（2）不同的排序方式 θ 如何影响 $R(\mathrm{T_C},\theta)$ 和 $R(\mathrm{T_E},\theta)$ 的相关性？

（3）从时间的动态角度看，$R(\mathrm{T_C},\theta)$ 和 $R(\mathrm{T_E},\theta)$ 的相关性是如何变化的？

4　研究方法

如前所述，本文采用三种排序方式对中英文科技主题分别进行排序，并对排序结果进行静态和动态的比较分析研究。本节将对本文的研究方法进行详细描述。首先介绍 PageRank 排序方式，然后对斯皮尔曼秩相关系数评估方法进行阐述，最后给出本文的静态和动态比较研究方案。

4.1　PageRank 排序方式

主题引用网络使用科技主题作为图上的节点，主题引用关系被视为引用和被引用主题之间的有向边，$\mathrm{G}=(\mathrm{V},\mathrm{E})$。每一个节点，$v\in\mathrm{V}$，代表了一个科技主题，每一条边，表示连接 v_i 和 v_j 的引用关系（v_i 引用 v_j）。

如果从基于引用关系的排序研究角度来看，大多数算法满足如下两个假设[29]：①当一个节点获得大量引用关系（入链接）的时候，这个节点是重要的；②当一个节点被重要的节点引用的时候，这个节点是重要的。

基于以上假设，PageRank[30]是一个可以循环计算引用网络图中的节点重要度的代表性算法：

$$\pi(v)^{i+1}=\mathrm{d}\left(\sum_{u=1}^{d_{\mathrm{in}}(v)}\frac{\pi(u)^i}{\left|d_{\mathrm{out}}(u)\right|}\right)+\frac{1-d}{N}$$

该计算公式展示了一条代表随机游走的马尔可夫链的众多时间步骤当中的一步。其中，$\pi(v)^{i+1}$ 表示了在时间步骤 $i+1$ 上节点 v 的重要性。$\pi(v)^{i+1}$ 由所有的引用该节点的节点 u 在时间步骤 i 上的重要性 $\pi(u)^i$ 决定，u 的重要性被 u 的出度 $\left|d_{\mathrm{out}}(u)\right|$ 所均分。参数 d 是阻尼系数（damping factor），N 表示所有的科技主题数目。已有大量研究显示 PageRank 是计算引用关系网络中节点重要性的一个有效方法[29,31-32]。本文也将采用 PageRank 作为一种排序方式，对中英文科技主题的重要性进行度量和排序。

4.2　比较研究的评估方法——斯皮尔曼秩相关系数

为了进行比较研究的量化评估，本文通过计算中文科技主题排序和其对应映射的英文科技主题排序之间的斯皮尔曼秩相关系数（Spearman's rank correlation coefficient）来进行定量分析。斯皮尔曼秩相关系数是一种非参数化的排序相关性评价指标。具体而言，它能够评估中英科技主题根据不同排序方式产生的主题排序的相似度。

斯皮尔曼秩相关系数可以被定义为[33]：

$$\rho=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i(x_i-\bar{x})^2\sum_i(y_i-\bar{y})^2}}$$

其中，x_i 是其中一个排序 $R(X,\theta)$ 中某随机变量的根据某种排序方式 $\theta(\cdot)$ 进行重要性排序后的秩，相似地，y_i 是另一个排序 $R(Y,\theta)$ 中对应的随机变量的排序后的秩，$\bar{x}$ 是 $R(X,\theta)$ 中所有随机变量重要性的秩的均值，类似的，$\bar{y}$ 是 $R(Y,\theta)$ 中所有随机变量重要性的秩的均值。

斯皮尔曼秩相关系数的值域为[–1,1]。如果两个排序完全相同，该相关系数的值为 1；反之，如果一个排序和另一个排序正好相反，那么它们的斯皮尔曼秩相关系数为–1；如果两个排序是完全随机独立的，那么它们的斯皮尔曼秩相关系数为 0。因此，如果斯皮尔曼秩相关系数越趋近于 0，那么这两个排序的相关度越低。

在本研究中，主题排序方式共有三种：分别为出现频率、被引用次数和 PageRank $(\theta_f(\cdot),\theta_c(\cdot),\theta_p(\cdot))$，其中，主题出现频率指主题的出现次数，主题被引用次数指该主题被其他主题的引用次数，主题 PageRank 指主题引用关系网络中每个科技主题的 PageRank 值。

4.3 计算机领域科技主题排序的跨语言比较研究方案

本文针对两种不同语种（中文、英文）的学术信息数据集进行分析研究：ACM（美国计算机学会）的英文学术论文数据集和万方中文学术论文数据集。

比较视角主要包括静态比较和动态比较研究，其中，静态比较研究的四个主要步骤如图 2（a）所示。

（1）从不同语种的学术信息集中，抽取不同语种的科技主题（关键词）信息。

（2）生成对应语种的学术主题引用信息。

（3）利用机器翻译技术[28]，对中文科技主题进行翻译，并根据翻译结果映射到对应的英文科技主题上。

（4）基于不同的排序方式 $(\theta_f(x_i),\theta_c(x_i),\theta_p(x_i))$，对能够产生映射的中英科技主题进行排序，并对排序结果进行比较研究。

如图 2（b）所示，动态比较研究分析是在静态研究的基础上，增加两个研究步骤：

（1）利用数据集中包含的时间属性划分不同的历史阶段，并依据历史阶段的划分来分别生成不同时期的主题引用网络。

（2）对不同的历史时期内的中英科技主题，基于不同的排序方式，对产生映射的中英科技主题进行排序，并对不同历史时期内的排序结果进行动态比较研究。

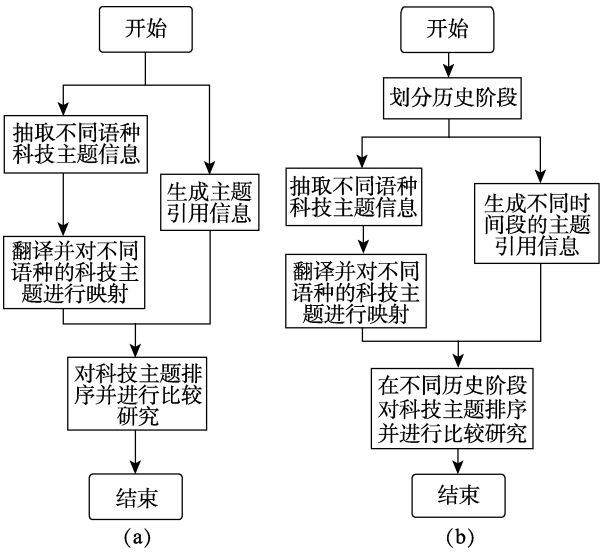


图 2 研究方法流程图

5 实验与结果分析

本章节将详细描述在美国计算机学会英文论文数据集和万方数据知识服务平台中文计算机专业学位论文数据集上进行的实验，并对实验结果进行分析，分为四个部分，首先介绍实验采用的数据，其次陈述静态比较实验结果，然后阐述动态比较试验结果，最后对实验结果进行分析。

5.1 实验数据

本文的实验数据包括如下两部分：

（1）本文从计算机科学技术领域的 6818 种国际学术期刊以及 4455 种国际会议论文集或者研讨会（Workshop）论文集中（来自 ACM，Association for Computing Machinery，即美国计算机学会论文数据库）抽取了 204427 篇英文学术论文（跨越的时间区间为 1990 年到 2011 年）作为英文数据集，包含 6003 个由作者提供的英文主题（关键词）以及 2247012 对主题引用关系。

（2）本文抽取了 217 所国内高校的 21550 篇有关计算机科学技术领域的硕士和博士中文学位论文（来自万方数据知识服务平台）作为中文数据集①，

① 由于 ACM 拥有其出版的大量期刊和会议论文集的版权，而中文计算机领域版权相对分散，没有类似单一机构拥有对应数量级的期刊和会议论文集，因此，为了避免数据收集不全而可能导致的偏歧，在中文计算机领域，本研究主要针对高校计算机专业的硕博学位论文进行数据收集。

该数据集的时间跨度为 13 年（2004 年到 2016 年），包含 5445 个由作者提供的中文主题（关键词）以及 28931 对主题引用关系。

由于中英文语料库的数量并不平衡，本文采用和研究文献[34]类似的方法对主题（关键词）进行归一化（Normalization）处理，从而消除潜在的偏歧：首先，计算不同语种的语料库的数量比例；然后，量级较大的语料库中的主题（关键词）数量除以该比例。在本实验中，该数量比例为 9.5（204427/21550），由于英文文本数量多于中文文本，因此，英语的主题（关键词）出现频率和被引次数需要除以 9.5。该归一化方法已被成功应用于谷歌公司的机器翻译相关研究[34]。

本文对科技主题（关键词）的处理主要包括以下步骤：

（1）首先通过分析将意义相近的主题合并作为同一个主题。

（2）对主题进行筛选，去除频率过高（在超过三分之一的文献中出现）和过低（少于两篇文献中出现）的主题。

（3）通过机器翻译 API[28]对中英文主题进行互相翻译，并各自建立索引，进行交叉对照匹配搜索，最终确定意义相同的对应中英文主题。

本文抽取的中英文科技主题数量如图 3 所示，在 5445 个中文主题和 6003 个英文主题中，共得到 2583 个可映射中英文主题。

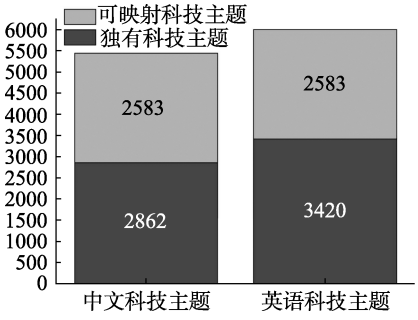


图 3　实验中的科技主题数量

中英文主题无法产生精确映射的原因主要有：

（1）部分主题代表的是中国独有的研究内容，如“中医面诊”、“中医临床诊疗”、“中国大豆网”、“广义中国剩余定理”、“中文垃圾邮件”等。

（2）缩写、冷僻，或者独创的概念，这些概念未被广泛使用，因此也无法找到可映射的科技主题。

（3）对于部分专业主题，通用机器翻译工具可能没有准确翻译到对应的英文主题。

本文的比较研究针对 2583 组可映射中英文科技主题展开。为了验证中英文科技主题的翻译质量，本文对该 2583 组可互相映射的中英文科技主题进行了抽样评估。评估实验在北京大学和美国印第安纳大学布鲁明顿分校（Indiana University Bloomington, IUB）两所高校进行，共有 6 名计算机相关专业的硕博士研究生①（北大 3 人、IUB 3 人）参与。每一名参与实验的研究生需要对 30 组中英文主题对（从 2583 组可映射中英文主题集中随机生成）打分，翻译正确打 1 分，翻译错误打 0 分。表 1 显示了最终得分，图 4 展示了得分分布，抽样评估显示建立映射的中英文主题平均正确率在 97%以上。

表 1　中英文主题翻译质量抽样得分

序号	正确	错误	总数
1	29	1	30
2	28	2	30
3	30	0	30
4	29	1	30
5	29	1	30
6	30	0	30

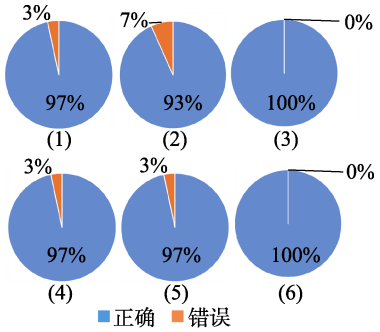


图 4　中英文主题翻译质量抽样得分分布

5.2　静态比较研究

本部分主要从全局角度对中英文科技主题进行比较研究，由于没有涉及时间要素，因此属于静态比较分析。

根据中英文主题在各自数据集中的出现频率，对中英文主题分别进行排序，表 2 展示了出现频率前 20 的中英文主题，图 5 是中英文主题排序的可视化，以主题词汇云的方式进行展示，主题的大小和出现频率成正比（为了显示清晰，只显示排序前 100 位的主题）。

① 由于需要判断中英文的翻译质量，参与该评估实验的美国高校研究生是中国留学生。

表 2 中英文主题出现频率

排序	出现频率	中文科技主题	英文科技主题	出现频率	排序
1	896	数据挖掘	algorithm	561.68	1
2	821	无线传输	distributed	225.47	2
3	712	计算机科学	secure	188.11	3
4	590	web 搜索	processor	177.79	4
5	491	特征提取	interfaces	177.47	5
6	417	云计算	control	177.26	6
7	416	分布式应用	context	173.37	7
8	411	支持向量机	standardization	138.84	8
9	391	网络安全	internet	133.26	9
10	349	遗传算法	semantic	132.42	10
11	289	嵌入式软件	web	130.11	11
12	258	软件开发	visualization	106.11	12
13	248	图像分类	cache	104.32	13
14	231	机器学习	genetic_algorithm	101.68	14
15	223	负载均衡	wireless_sensor_network	101.16	15
16	203	个性化编辑	cluster	96.95	16
17	195	入侵检测系统	java	92.42	17
18	192	自适应算法	sensor_networks	89.37	18
19	188	android 操作系统	machine_learning	85.58	19
20	188	功能模型	usability	81.16	20

注：英文科技主题出现频率已进行归一化处理。

从表 2 和图 5 可见，在计算机领域，英文主题出现频率最高的是“algorithm（算法）”、“distributed（分布性）”、“interfaces（人机界面）”等涵盖范围非常宽泛的主题，代表了计算机科学领域最为基础和重要的研究方向;其中,只有排名第 14 位的“genetic algorithm（遗传算法）”是比较具体的算法。与之对应的，中文主题也呈现相同的特征，除了排名第 8 位的“支持向量机”和排名第 10 位的“遗传算法”是具体的机器学习算法，其余出现频率最高的主题，如“数据挖掘”、“web 搜索”等都是范围较大的研究方向。出现这种现象的原因可能是宽泛的主题往往涵盖了一定数量的具体主题，因此，能获得更多的出现概率。在出现频率最高的前 20 个中英文主题中，同时出现的映射主题共有 3 个，分别是“genetic algorithm-遗传算法”、“wireless sensor network-无线传输”、“machine learning-机器学习”；由此可见，从出现频率排序的角度，中英文的主题呈现出一定的相关性。

另外，中文出现频率排名第 3 位的“计算机科学”（英文排名第 260 位）以及英文出现频率排名第 1 位的“algorithm（算法）”（中文排名第 1674 位）呈现出较大的差异性，这可能和本文采用的实验数据本身的特性有关：由于中文数据集是学位论文，因此作者可能更倾向于标记出论文的专业类别，所以容易造成“计算机科学”这个主题出现次数偏多，而英文数据集是 ACM 论文数据，ACM 分类系统中，“algorithm”是一个大的分类，因此，容易造成该主题在英文论文中出现频率较高。

表 3 和图 6 表示根据中英文主题在各自数据集中的被引用的次数对中英文主题分别进行排序的情况，其中，表 2 展示了被引用次数前 20 位的中英文主题，图 5 是中英文主题排序的可视化，主题的大小和被引用的次数成正比，主题间的连线表示引用，其粗细和引用次数成正比（为了显示清晰，只显示排序前 100 位的主题）。

从表 3 和图 6 可见，以被引用次数进行排序，英文主题排序最高的主题产生了变动，在被引次数最多的前 20 个主题中，60%的主题（12 个）和出现频率排序排名最高的 20 个主题相同，虽然被引次数最多的前 10 个主题都是出现频率前 20 位的主题，但是排名前 10 位至前 20 位的主题，大都不是出现频率最高的主题；并且，本文也发现，被引次数最多的前 20 个英文主题均为涵盖范围宽泛的主题。与之对应的是，中文主题也呈现相似的特征，在被引

图 5 中英文主题词汇云

表 3　中英文主题被引用次数

排序	被引次数	中文科技主题	英文科技主题	被引次数	排序
1	558	数据挖掘	algorithm	5555.26	1
2	445	云计算	distributed	2231.47	2
3	367	无线传输	processor	1980.42	3
4	364	计算机科学	interfaces	1701.58	4
5	348	web 搜索	context	1678.21	5
6	253	分布式应用	control	1636.00	6
7	243	协同过滤	internet	1311.68	7
8	196	文本分析	web	1259.89	8
9	185	机器学习	standardization	1210.63	9
10	181	特征提取	cache	1082.84	10
11	174	信息探测	semantic	1000.74	11
12	174	软件开发	precise	947.47	12
13	166	支持向量机	secure	918.11	13
14	163	android 操作系统	robust	793.79	14
15	158	网络安全	operating_system	723.89	15
16	136	搜索引擎	consistency	676.21	16
17	131	负载均衡	throughput	633.16	17
18	129	电子商务	libraries	585.58	18
19	121	Hadoop 架构	wireless_networks	543.89	19
20	116	特征选择	formalism	520.11	20

注：英文科技主题被引次数已进行归一化处理。

次数最多的前 20 个主题中，65%的主题（13 个）和出现频率排序排名最高的 20 个主题相同，被引次数最多的前 10 个主题有 9 个都是出现频率前 20 位的主题，而排名前 10 位至前 20 位的主题，大都不是出现频率最高的主题，但是和英文主题不同，如“协同过滤”、“支持向量机”等这样和具体算法模型有关的中文主题也有很高的被引次数。在被引次数最高的前 20 个中英文主题中，同时出现的映射主题数量为 0，因此和出现频率排序相比，以被引次数进行排序后，中英文主题排序的相关性下降了；另一方面，被引次数最高的前 20 个主题中，部分主题虽然不是直接映射主题，但依然具有一定的相关性，如“分布式应用”和“distributed（分布性）”等。

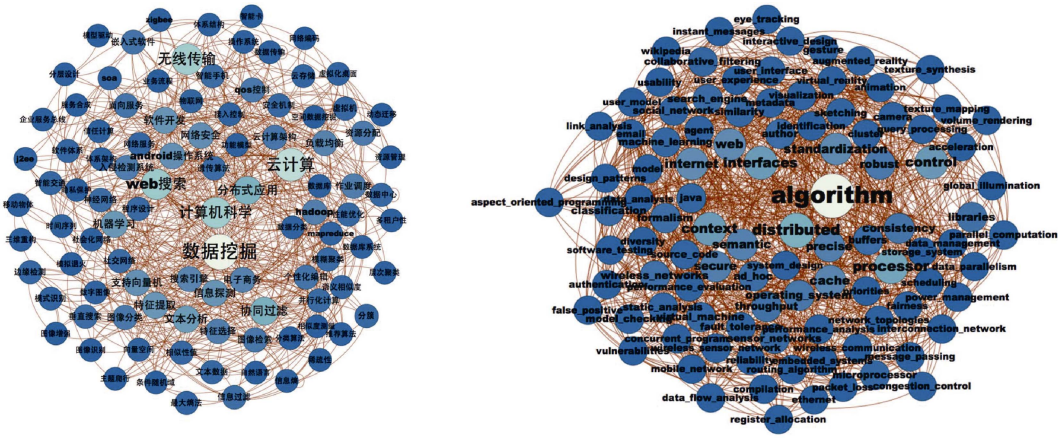
表 4 表示根据中英文主题在各自数据集中的 PageRank 值对中英文主题分别进行排序的情况。

正如前文所述，相较于单纯对被引次数进行计数，PageRank 是计算引用关系网络中节点重要性的一种有效算法。从表 4 可见，以 PageRank 值进行排序，英文主题排序最高的主题有微小的变动，PageRank 值最大的前 20 个主题和被引次数最多的前 20 个主题一致，只是排序发生了变化，与之对应的是，中文主题也呈现类似的特征，在 PageRank 值最大的前 20 个主题中，90%的主题（18 个）和被引次数排名最高的 20 个主题相同，其中，“图像分类”和“嵌入式软件”这两个主题是原先被引次数排名最高的 20 个主题中没有包括的；而且排序也出现了微小变化。和被引次数排序相同，在 PageRank 值最大的前 20 个中英文主题中，同时出现的映射主题数量为 0，但也有部分主题呈现出一定的相关性。总体而言，和出现频率排序相比，以 PageRank 值进行排序后，中英文主题排序的相关性也呈现下降趋势。

为了进一步进行定量的相关性的比较分析，本文对中英文科技主题通过三种不同的排序方式得到的总体排序的斯皮尔曼秩相关系数进行计算，得到的结果如图 7 所示。

5.3　动态比较研究

第 5.2 节从全局角度对中英科技主题排序进行了比较研究，但这种比较是静态的，忽视了时间因



(a)　　(b)

图 6　中英文科技主题被引用次数可视化

表 4 中英文主题 PageRank 值

排序	PageRank 值	中文科技主题	英文科技主题	PageRank 值	排序
1	0.007536	数据挖掘	algorithm	0.035716	1
2	0.005577	计算机科学	distributed	0.014537	2
3	0.005252	云计算	processor	0.012524	3
4	0.004753	web 搜索	interfaces	0.012049	4
5	0.004455	无线传输	control	0.011357	5
6	0.003826	协同过滤	context	0.011296	6
7	0.003390	分布式应用	web	0.008623	7
8	0.003223	软件开发	standardization	0.008549	8
9	0.002929	机器学习	internet	0.008283	9
10	0.002786	特征提取	cache	0.006851	10
11	0.002647	支持向量机	semantic	0.006747	11
12	0.002430	信息探测	precise	0.006388	12
13	0.002406	文本分析	secure	0.005488	13
14	0.002362	电子商务	robust	0.005250	14
15	0.002293	android 操作系统	consistency	0.004668	15
16	0.002213	网络安全	operating_system	0.004666	16
17	0.002173	图像分类	throughput	0.004177	17
18	0.001874	负载均衡	libraries	0.003950	18
19	0.001743	搜索引擎	formalism	0.003501	19
20	0.001725	嵌入式软件	wireless_networks	0.003457	20

素的重要性，由于科技文献具有时间属性，在不同的时间阶段，科研的热点、研究的趋势都会发生动

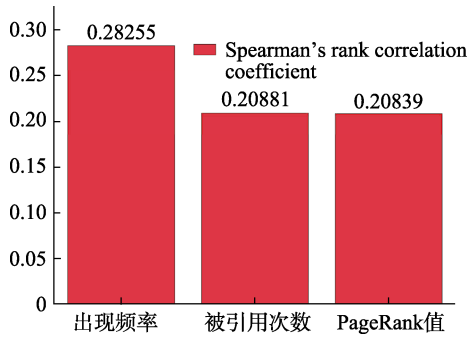


图 7 中英文主题排序的斯皮尔曼秩相关系数

态的变化，而时间维度对于科技领域的发展和研究人员的知识认知都有关键性影响。因此，有必要把时间要素纳入研究范围，从动态的角度对中外科技主题排序进行比较研究和分析。

本文将语料库各自分为 4 个时间片，这样共有 6 个时间片，其中重合的时间片有 2 个，如表 5 所示。

从表 5 可知，在各自的数据集中，在不同的历史阶段，论文数量的分布是不均匀的，早期的论文数量往往相对偏少。另外，由于实验数据集的限制，英文数据集的数量在各个时间片上的数量都超过中文数据集，因此对英文论文数进行归一化处理。

出现频率排序前 10 位的中英文科技主题在不同时间片上的变化如图 8 所示。

表 5 时间片划分

时间片	1990-1999	2000-2003	2004-2008	2009-2011	2012-2013	2014-2016
英文论文数	5151.05	3146.42	7456.53	5764.63	—	—
中文论文数	—	—	196	3456	11047	6851

注：英文论文数已进行归一化处理。

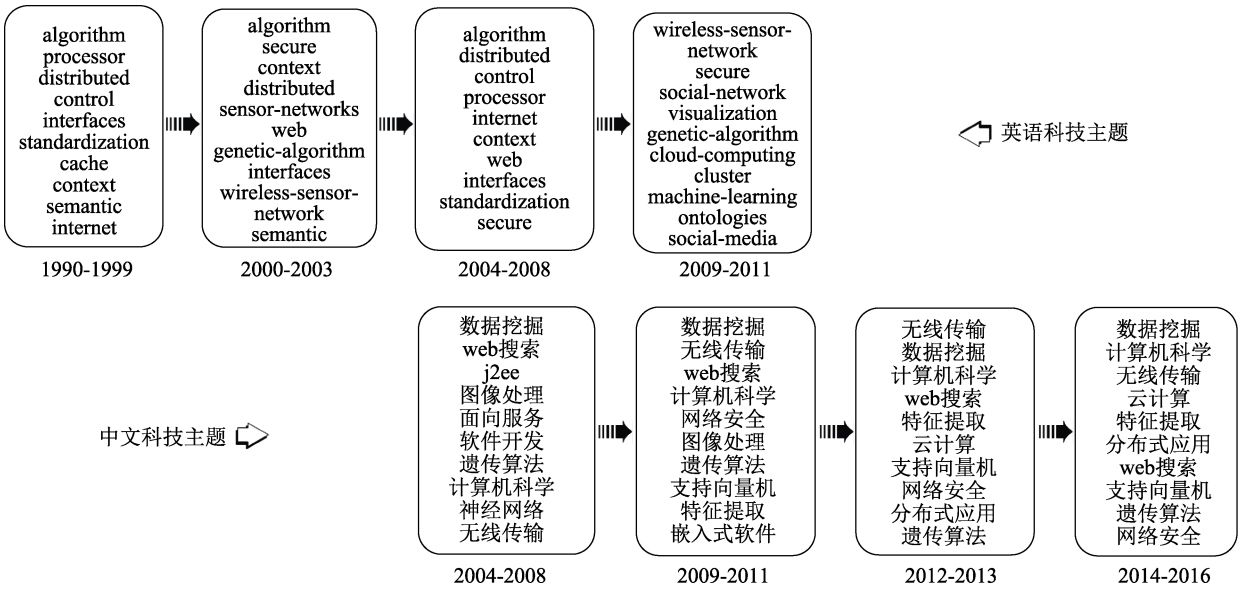


图 8 出现频率前 10 位的中英文主题在不同时间片上的变化

出现频率反映了主题出现的论文数，是研究热点的一个显著指标。由图 8 可知，从出现频率角度而言，1990-1999 时期，英文高频出现的主题大都是一些传统且基础的研究主题，如“algorithm（算法）”等。进入到 2000-2003 时期，“sensor networks（传感网络）”等主题的出现频率开始上升，表明这一阶段，对于网络相关的研究得到了国际研究领域的重视。在 2004-2008 时期，“Internet（因特网）”等主题的出现频率上升，表明该阶段对互联网和文本信息的研究热度在提高。前面三个历史阶段，高频出现的英文科技主题还是保持相对稳定，但到了 2009-2011 阶段，最高频出现的英文主题发了巨大变化，如“social network（社交网络）”、“machine learning（机器学习）”等科技主题迅速进入了前 10 位的行列。可见这一阶段，随着技术的进步、互联网的发展，国际计算机研究领域的热门研究主题发生了变化，社交媒体开始进入人们的研究视野，海量的数据使得机器学习等主题的研究变得更为迫切。

在中文计算机研究领域，早期（2004-2008 时期）除了和国际研究领域类似的主题，中文论文还有“软件开发”、“j2ee”等主题高频出现，表明这一阶段国内的计算机研究还有相当部分专注于开发等工程问题，研究的层次相对而言比较初级。在 2009-2011 时期，和开发有关的主题就退出高频主题的行列了，表明国内的计算机研究的研究层次正在不断进步。从 2012-2013 阶段，“云计算”、“分布式应用”等主题的出现频率上升，表明国内计算机研究也开始向机器学习、大数据、分布式方向迈进。

另外，本文还发现一个有趣的现象，从出现频率而言，某一时期的高频英文主题往往和本时期的中文主题不太相似，但和之后时期的中文主题较为相近，有明显的“滞后”特征。比较明显的例子是，2009-2011 时期的高频英文科技主题出现了“cloud computing（云计算）”，但在本阶段的中文高频主题中并没有出现该主题，需要等到 2012 以后的两个阶段，云计算才成为中文研究领域的热门主题。

图 9 展示了 PageRank 排序前 10 位的中英文科技主题在不同时间片上的变化。

PageRank 值是科技主题重要性的一种度量，不仅反映了主题在该历史阶段的热门程度，也表明主题在该历史阶段的影响力。从图 9 可知，在国际计算机研究领域，前三个历史时期，英文主题被引次数排序基本保持稳定，高 PageRank 主题均是较为传统和基础的科技主题，而到了 2009-2011 时期，“social network（社交网络）”、“machine learning（机器学习）”这些主题的 PageRank 值迅速上升，表明了国际计算机研究的重点发生了变化，朝着自动化、社交化、智能化的方向发展。而中文科技主题方面，高 PageRank 值主题从早期的工程相关主题朝着后期的分布式、智能化相关的主题不断变化。和出现频率排序类似，PageRank 值排序也出现了英文主题排序和中文主题排序相关性的“滞后相似”现象。

为了更好地对相关性进行比较分析，本文比较了在不同历史时期，采用不同排序方法生成的中英文主题排序的斯皮尔曼秩相关系数，具体比较方法是：

（1）采用某种排序方式 $\theta(\cdot)$ 对各个时间片上的

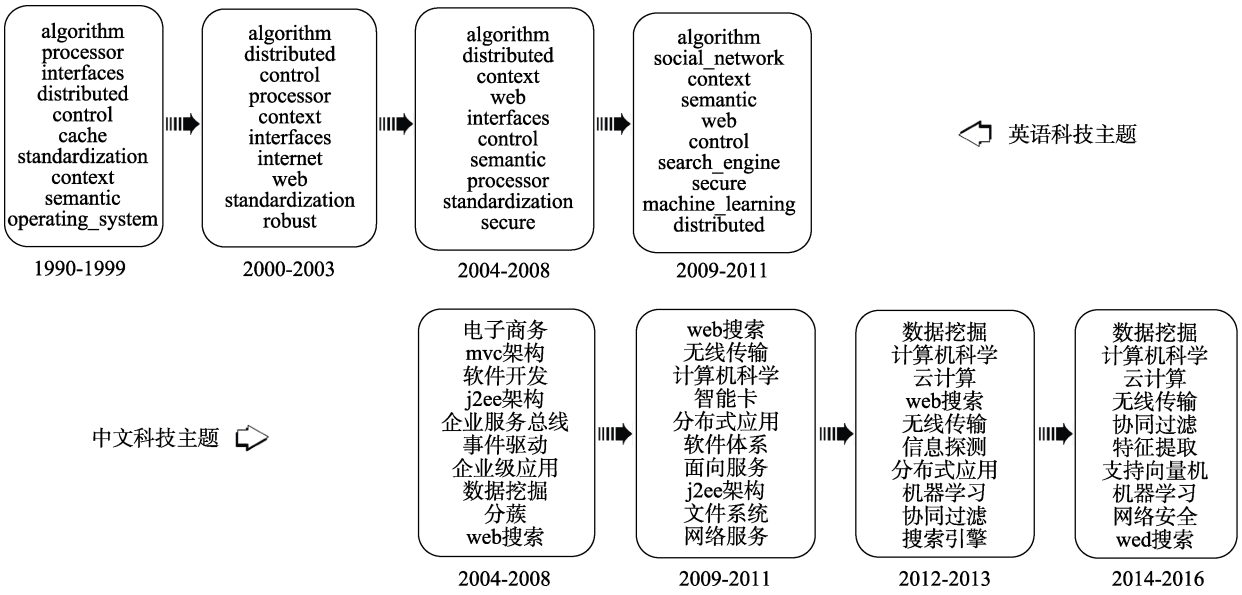


图 9　PageRank 前 10 位的中英文主题在不同时间片上的变化

中英文主题进行排序。

（2）以各个时间片上的英文主题排序为基准，分别计算其和各个时间片上的中文主题排序的斯皮尔曼秩相关系数。

最终得到的结果见表 6，其变化趋势的可视化由图 10 展示。

5.4 实验结果分析和讨论

在**静态比较试验**中，通过对采用出现频率、被引次数、PageRank 值（$\theta_f(\cdot)$，$\theta_c(\cdot)$，$\theta_p(\cdot)$）三种排序方式排序后中英文主题的排序结果分析，可以得到以下定性结论：

表 6 在不同时间片上的中英文主题排序的斯皮尔曼秩相关系数

英文主题排序所在时间片	中文主题排序所在时间片	排序方法	相关系数	排序方法	相关系数	排序方法	相关系数
1990-1999	2004-2008	出现频率	0.115955	被引次数	0.048691	PageRank	0.048883
	2009-2011		0.183231		0.088610		0.089577
	2012-2013		0.148447		0.097106		0.096375
	2014-2016		0.104112		0.111700		0.109600
2000-2003	2004-2008	出现频率	0.137631	被引次数	0.068059	PageRank	0.068205
	2009-2011		0.221008		0.128813		0.128499
	2012-2013		0.195169		0.138073		0.135899
	2014-2016		0.165214		0.170409		0.168936
2004-2008	2004-2008	出现频率	0.149407	被引次数	0.061332	PageRank	0.060096
	2009-2011		0.256914		0.128598		0.126391
	2012-2013		0.248096		0.176444		0.172225
	2014-2016		0.194751		0.178966		0.175569
2009-2011	2004-2008	出现频率	0.135415	被引次数	0.067036	PageRank	0.060559
	2009-2011		0.247490		0.157451		0.154412
	2012-2013		0.243302		0.178223		0.173388
	2014-2016		0.231571		0.207339		0.205197

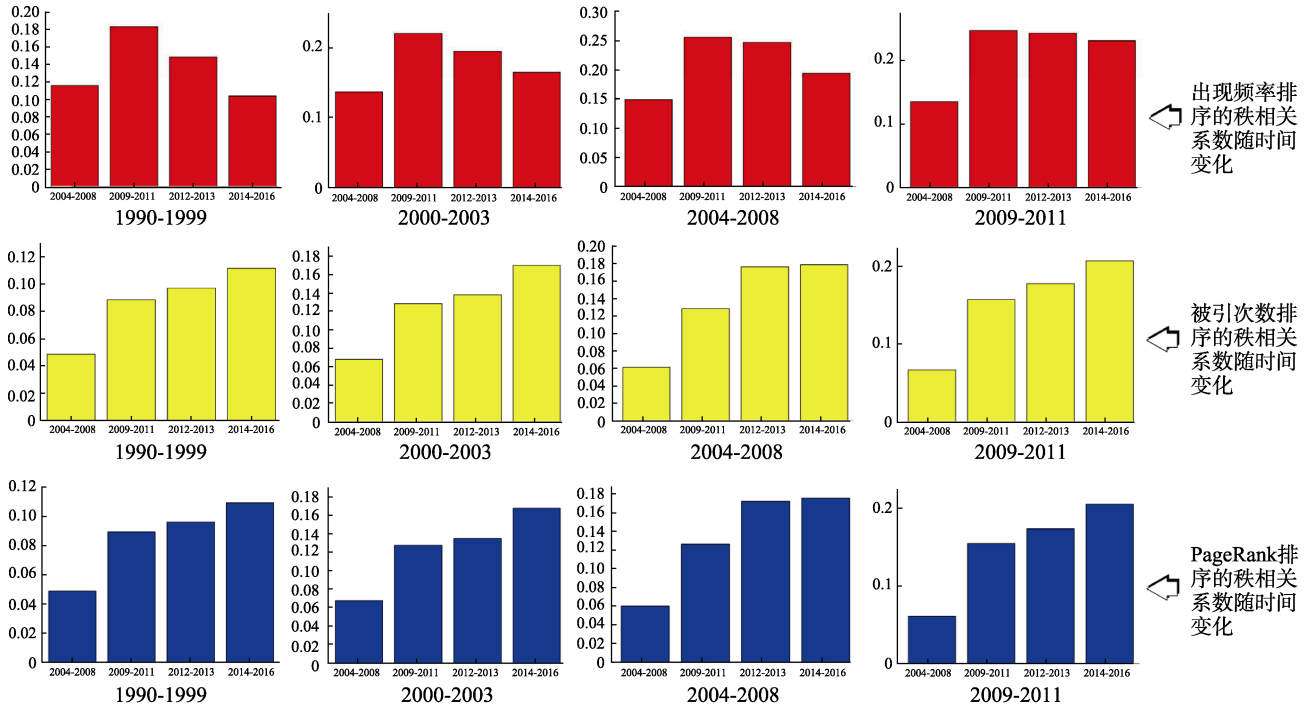


图 10 中英文主题排序的斯皮尔曼秩相关系数在不同时间片上的变化

◆在计算机科学领域，采用本文所涉及的排序方式，中英文主题排序均呈现出类似的特征，具有一定的相关性。

◆以出现频率进行排序后的中英文主题排序比其他两种排序方式呈现出的相关性要更强一些。

这些结果表明中文计算机研究领域和国际计算机研究领域会对一些共同的研究主题产生兴趣，从而形成类似的研究热点和研究趋势。

通过计算静态比较下的不同排序的斯皮尔曼秩相关系数，可以定量地得到中英文科技主题排序的相关性度量，从计算结果看：

◆不管采用何种排序方式，中英文主题排序的秩相关系数均大于 0，但没有超过 0.3，呈现出较弱的**正相关性**。这说明，在计算机学科，中文研究领域和国际研究领域的科技主题研究热点和趋势存在一定的类似性，但又各有研究侧重。

◆以出现频率进行排序后的中英文主题排序正相关性最强，而以被引次数和 PageRank 值进行排序后的中英文主题排序的正相关性相对较弱，且非常接近。这说明，从文章数量来看，中文计算机研究领域和国际计算机研究领域的**热门科技主题相对类似**，但从影响力和重要性而言，中外计算机研究的有影响力的**重点主题相关性相对较弱**，**具有一定差异性**。

在**动态比较实验**中，在不同时间片上的中英文主题排序的斯皮尔曼秩相关系数计算结果表明：

（1）不论在任何历史时期，采用何种排序方式，中英文主题排序的秩相关系数均大于 0，其中，最低的秩相关系数为 0.048691，最高的秩相关系数为 0.256914，因此，和全局比较类似，在时间动态比较中，中英文主题排序依然呈现出较微弱的**正相关性**。

（2）在各个历史阶段上，中英文主题出现频率排序正相关性最强，中英文主题被引次数排序和 PageRank 排序的正相关性相对较低，且非常接近。这说明，在各个历史阶段上，中文计算机研究领域和国际计算机研究领域的科技主题依旧是**数量排序更为相关而影响力和重要性排序相关性较弱**。

（3）如果把不同的历史阶段的不同排序方式得到的中英文主题排序进行混合比较，中英文主题被引次数排序和 PageRank 排序的相关系数也有可能超过出现频率排序。例如，2009-2011 时期英文主题 PageRank 排序和 2014-2016 时期中文主题 PageRank 排序的秩相关系数为 0.205197，这大于 2009-2011 时期英文主题出现频率排序和 2004-2008 时期中文主题出现频率排序的秩相关系数（0.135415）。这说明，时间因素对中英文主题的相关性影响是不可忽视的，有时候甚至可以超越排序方式的影响。

（4）从不同历史阶段的秩相关系数变化趋势来看，中英文主题排序的相关性**“滞后相似”**现象的确存在，即某一历史时期的英文科技主题排序，**随着时间的发展**，和后期中文科技主题排序的**正相关程度会增加**。这说明了国内的计算机科学研究一定程度上受到国际研究的影响，某一阶段国际研究的热点会逐渐带动国内的研究热潮。由于历史发展的原因，欧美发达国家的计算机研究比我国的有关研究起步早，进展快，先进程度高，目前我国的计算机科学研究一定程度上还处于“跟随者”、“追赶者”的状态，我们必须要正视这一客观情况，相信随着国内科研实力的不断进步，这种差距会越来越小。

（5）**不同的排序方法**对中英文主题排序的相关性**“滞后相似”现象的作用时间范围不同**，从图 10 可以明显看出，出现频率排序对相关性的影响时间较短，往往在后一个历史阶段中英文主题排序相关性会增加，但更往后阶段的相关性反而会下降；而与之对应的，被引次数排序和 PageRank 排序的正相关性是随着时间不断增加的趋势。这从一定程度上揭示了，对于国外的热门研究主题，短期内会有大量的类似国内研究跟进，但这种数量上带来的可持续性不强。而被引次数排序和 PageRank 排序代表了有影响力的重要主题排序，这种主题的研究能够经受时间的考验，正相关程度在时间维度上不断增强。

6　结论和展望

本文针对计算机领域国际和国内的科技主题进行比较研究，通过在 ACM 英文论文数据集和万方数据知识服务平中文学位论文数据集上的静态和动态对比分析实验，得到以下结论：

（1）从全局而言，中英文科技主题排序存在正相关关系；从时间维度而言，中英文科技主题排序在各个历史时期也存在正相关关系，且英文科技主题排序和不同历史时期的中文科技主题排序的正相关关系随着时间成增强趋势，不同的排序方式会造成正相关增强趋势的作用时间范围发生变化。

（2）中英文科技主题排序的正相关性说明了国内外计算机科学研究的热点方向存在一定的相似性，不同排序方式带来的正相关性变化说明了国内计算机科学的热点研究主题仅仅处于数量相关的阶段，影响力和重要度相关性相对较弱。

（3）中英文主题排序的相关性在时间维度上的“滞后相似”现象说明，国际计算机科学研究热门主题会带动国内有关研究的发展，由于历史发展的原因，国内的计算机研究还处于“跟随者”的角色。

本文对主题排序的比较研究可以帮助科研人员以及管理决策部门对研究热点和发展趋势进行准确预判，而且对提高科学研究生产率也能够起到良好的支持和辅助作用，是情报服务的重要方面，对信息组织和检索等情报学核心领域亦有基础性作用。其中，不同语种的研究又对我国的科技研究发展具有实践意义。当然，本文研究仅是针对跨语言主题排序及其比较的探索性工作，存有的不足之处将构成作者团队后续研究的探索重点。主要包括：由于数据源结构的问题，实验的中文数据集和英文数据集的数量和时间不完全平衡，寻求和积累更完善的实验和实践数据将是之后重要的工作；探索上下文环境下的主题对应翻译从而提高翻译质量也是未来工作的一个改进方向。本研究仅以计算机领域为例，之后需要在更多领域进一步地探索和验证。

参 考 文 献

[1] Bornmann L, Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references[J]. Journal of the Association for Information Science and Technology, 2015, 66(11): 2215-2222.

[2] Jensen S, Liu X, Yu Y, et al. Generation of topic evolution trees from heterogeneous bibliographic networks[J]. Journal of Informetrics, 2016, 10(2): 606-621.

[3] 王燕鹏. 国内基于主题模型的科技文献主题发现及演化研究进展[J]. 图书情报工作, 2016, 60(3): 130-137.

[4] 关鹏, 王日芬. 基于 LDA 主题模型和生命周期理论的科学文献主题挖掘[J]. 情报学报, 2015, 34(3): 286-299.

[5] 马费成, 张勤. 国内外知识管理研究热点——基于词频的统计分析[J]. 情报学报, 2006, 25(2): 163-171.

[6] 王林, 冷伏海. 施引关键词与被引作者交叉共现分析方法及实证研究[J]. 情报学报, 2012, 31(4): 362-370.

[7] Ozcinar Z. The topic of instructional design in research journals: A citation analysis for the years 1980-2008[J]. Australasian Journal of Educational Technology, 2009, 25(4): 559-580.

[8] Bharat K, Henzinger M R. Improved algorithms for topic distillation in a hyperlinked environment[C]// Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1998: 104-111.

[9] Blei D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4): 77-84.

[10] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[11] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2009: 248-256.

[12] 章成志, 王惠临. 多语言文本聚类研究综述[J]. 现代图书情报技术, 2009(6): 31-36.

[13] Cutting D R, Karger D R, Pedersen J O, et al. Scatter/Gather: A cluster-based approach to browsing large document collections [C]// Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1992: 318-329.

[14] Hearst M A, Pedersen J O. Reexamining the cluster hypothesis: scatter/gather on retrieval results[C]// Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1996: 76-84.

[15] 沈洪洲, 宗乾进, 袁勤俭. 国外社会化媒体研究主题演化分析[J]. 情报科学, 2013, 31(1): 99-105, 152.

[16] Blei D M, Lafferty J D. Dynamic topic models[C]// Proceedings of the 23rd International Conference on Machine Learning. New York: ACM Press, 2006: 113-120.

[17] 蒋卓人, 陈燕, 高良才, 等. 一种结合有监督学习的动态主题模型[J]. 北京大学学报(自然科学版), 2015, 51(2): 367-376.

[18] 章成志, 梁勇. 基于主题聚类的学科研究热点及其趋势监测方法[J]. 情报学报, 2010, 29(2): 342-349.

[19] 陈仕吉, 王小梅. 基于 *C-value* 与 *TF-IDF* 的文献簇主题识别研究[J]. 情报学报, 2009, 28(6): 821-826.

[20] 章成志, 张庆国, 师庆辉. 基于主题聚类的主题数字图书馆构建[J]. 中国图书馆学报, 2008, 34(6): 64-69.

[21] 范云满, 马建霞. 利用 LDA 的领域新兴主题探测技术综述[J]. 现代图书情报技术, 2012(12): 58-65.

[22] 徐小龙, 李永萍, 李涛. 云计算领域科技文献统计与研究热点分析[J]. 南京邮电大学学报(自然科学版), 2015, 35(4): 1-14.

[23] Wang C H, Zhang M, Ru L Y, et al. Automatic online news topic ranking using media focus and user attention based on aging theory[C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2008: 1033-1042.

[24] Shubhankar K, Singh A P, Pudi V. An efficient algorithm for topic ranking and modeling topic evolution[C]// Proceedings of the 22nd International Conference on Database and Expert Systems Applications. Heidelberg: Springer-Verlag Berlin, 2011: 320-330.

[25] Shuai X, Jiang Z R, Liu X Z, et al. A comparative study of academic and Wikipedia ranking[C]// Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. New York: ACM

Press, 2013: 25-28.

[26] Jiang Z R, Liu X Z, Gao L C. Chronological Citation Recommendation with information-need shifting[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2015: 1291- 1300.

[27] 赵迎光, 洪娜, 安新颖. 主题模型在主题演化方法中的应用研究进展[J]. 现代图书情报技术, 2014(10): 63-69.

[28] 百度翻译开放平台[EB/OL]. [2016.06]. http://api.fanyi.baidu.com/.

[29] Jiang Z R, Liu X Z, Chen Y. Recovering uncaptured citations in a scholarly network: A two-step citation analysis to estimate publication importance[J]. Journal of the Association for Information Science and Technology, 2016, 67(7): 1722-1735.

[30] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the web[J]. Stanford Digital Libraries Working Paper, 1998, 9(1): 1-14.

[31] Ding Y. Applying weighted PageRank to author citation networks[J]. Journal of the American Society for Information Science and Technology, 2011, 62(2): 236-245.

[32] Liu X Z, Zhang J S, Guo C. Full-text citation analysis: A new method to enhance scholarly networks[J]. Journal of the American Society for Information Science and Technology, 2013, 64(9): 1852-1863.

[33] Pirie W. Spearman rank correlation coefficient[J]. Encyclopedia of Statistical Sciences, 1988.

[34] Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation[J]. OALib Journal, 2013.

（责任编辑　宋　扬）