

基于 Citespace 的文本检索相关性研究

张 倩

(华南农业大学珠江学院, 广东 广州 510900)

摘要: 运用 CiteSpace 可视化知识图谱分析,可对当前领域的研究现状及发展趋势进行直观的揭示。以中国社会科学引文索引(CSSCI)为资料来源,对相关研究论文进行检索。从文献数量的变化、研究热点与发展趋势、作者合作等多个方面进行了知识图谱的分析与论述。研究表明,文本检索相关研究的总体趋势日趋强大,但作者之间的合作程度、研究热点的凝聚度等仍有待加强,而跨媒体检索、微博检索、数据融合、文本检索、短文检索、文本检索、短文检索等相关研究有望成为研究热点或发展新趋势。

关键词: 文本检索; 相关研究; CiteSpace

一、引言

信息检索 (InformationResearch) 是指信息用户寻找、识别和获取有关事实、资料、知识的活动和过程,以帮助人快速准确的得到所需要的信息。文本检索技术目前已在政府、企业、图书情报机构等各类机构和组织中得到了深入应用,成为信息化建设和知识管理的重要工具,是一种最早也是最重要的信息检索形式,使人们在海量的语料中匹配和查询所需要的信息。为此,本文选取中国知网中文文本检索相关文献,利用 CiteSpace 可视化分析了该领域的研究现状。了解目前对文本检索相关性的研究情况,为后续研究起到借鉴和参考作用。

二、文本检索相关性的文献分布

(一) 研究工具和数据采集

CiteSpace 是一种科学文献分析工具,旨在通过绘制知识结构图谱,展示该范畴的研究特点和研究的前沿领域,探索学科领域

演变的关键路径和知识转折点。

为了提高查询结果的准确度,在中国知网(CNKI)高级检索界面上,采用逻辑与的检索方法,运用检索式主题=(文本检索 and 相关性)(精确匹配)检索相关文献,得出结果 771 篇,人工剔除无效与无关文献,共获取相关文献 439 篇。

(二) 文献时序分析

文章的时序分布如图 1 所示,从中可以看出,文本检索的研究始于 1985 年,研究的主要内容是文本检索的硬件实现和全文的自动标引,而 2002 年以前的文本检索相关研究一直处于初级阶段,发文的数量总是少之又少;2004~2013 年处于研究高峰,2007 年达到最高值,文本检索开始应用于语义检索、概念检索等新兴领域;之后的发文数量较之前有所减少,处于不断的波动变化中,可能是有些技术的研究处于瓶颈期,还没有突破。

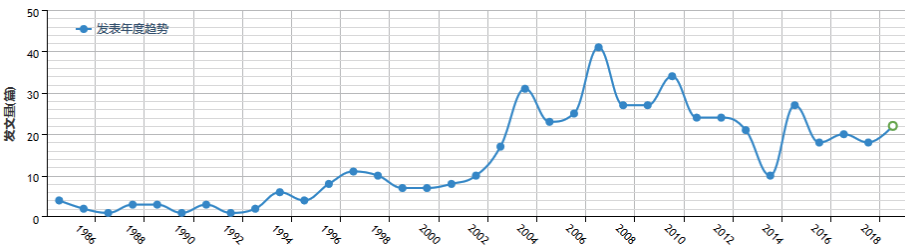


图 1 文献时序分布图

(三) 关键词的分布

利用 CiteSpace 的可视化功能对所选期刊文献进行关键词词频统计(如表 1 所示),共抽取了 92 个高频关键词,其中文本检索、信息检索、向量空间模型、图像检索、图像检索等关键词研究的文献数量较多,文本检索模型一直是研究的重要内容。在文本过

滤、超文本、聚类、位置加权、小世界现象等方面的研究还比较少,从齐普夫的定律可以看出,某学科相关论文关键词出现的频率和规律,可以看出该学科学术研究的热点,也可以看出该学科未来发展趋势的预兆。

表 1 计量研究(部分)具有较高频次的文本检索关联词

Freq	107	48	32	19	14	13	12	11	9	6	6	6	5	5
Centrality	0.72	0.34	0.28	0.14	0.04	0.06	0.09	0.02	0	0.02	0	0.01	0	0.01
Keyword	文本检索	信息检索	向量空间模型	图像检索	文本分类	基于内容的图像检索	查询扩展	检索模型	跨媒体检索	lucene	相关反馈	特征选择	文本过滤	聚类

利用时区视图从时间维度显示演进过程,统计 2000~2019 年的研究情况,如图 2 所示。可以看出从 2000~2002 的关键词主要是文本分类、向量空间模型、超文本等。效果相对较好的算法是程军根据前期调研情况分析得出的。林鸿飞等提出了易于操作,便于解释,能够不依赖具体领域的文本结构分析中的一种层次分析方法。2003~2006 的关键词是图像检索、文本过滤、检索模型。孙岩国对中文文本过滤方法进行了优化,给出了一种混合式的中文文本过滤实现方法。黄斌等采用以内容为主的文本检索技术实

现以内容为主的图像检索,从实现原理、算法流程、检索实现三个方面给出了以文本检索为主的解决方案,以图像映射为文本,以文本还原为图像,以实例系统的方式实现图像检索。

2007~2009 是 Lucene,语义检索,查询延伸。赵亚慧通过潜在语义模型处理文本向量空间,再利用遗传算法最优化的查找特征,找出与查询文本相匹配的结果,并据此提出用向量空间模型表示文本的文本检索算法(expressivespacemodel)。郑廷等实现了一种语义检索实验系统,采用 C/S 架构模式,基于 Lucene 的传统

