北京外国语大学211工程建设学术成果系列
北京外国语大学2009年博士文库系列

被引内容分析
——探究领域知识结构的新方法尝试

马晓雷 著

外语教学与研究出版社

北京外国语大学211工程建设学术成果系列
北京外国语大学2009年博士文库系列

被引内容分析
——探究领域知识结构的新方法尝试

马晓雷 著

外语教学与研究出版社
北京

图书在版编目(CIP)数据

被引内容分析：探究领域知识结构的新方法尝试／马晓雷著．— 北京：外语教学与研究出版社，2011.2
(北京外国语大学2009年博士文库系列)
ISBN 978-7-5135-0631-1

Ⅰ．①被… Ⅱ．①马… Ⅲ．①语言学－研究方法
Ⅳ．①H0-3

中国版本图书馆CIP数据核字(2011)第025069号

出 版 人：于春迟
责任编辑：李旭洁
封面设计：袁 璐
出版发行：外语教学与研究出版社
社　　址：北京市西三环北路19号 (100089)
网　　址：http://www.fltrp.com
印　　刷：北京九州迅驰传媒文化有限公司
开　　本：850×1168 1/32
印　　张：9
版　　次：2011年3月第1版 2011年3月第1次印刷
书　　号：ISBN 978-7-5135-0631-1
定　　价：32.90元

*　　*　　*

购书咨询：(010) 88819929　**电子邮箱**：club@fltrp.com
如有印刷、装订质量问题，请与出版社联系
联系电话：(010) 61207896　**电子邮箱**：zhijian@fltrp.com

物料号：206310001

摘　要

研究一个知识领域的结构和发展趋势具有重要的意义。在文献计量学中，两种最主要的分析领域知识的方法是共引分析和共词分析。共引分析以引文为分析对象，认为引用是科学知识传播主要的方式。共词分析以关键词为统计单位，强调词语是知识概念最直接的载体。这两种方法各有优点，又都存在着各自难以解决的问题。有学者指出，两种方法应该互相结合以优势互补。

针对这种情况，本研究提出了一种新的领域知识分析方法——被引内容分析法，即通过分析文献正文中的被引内容来反映某一领域的知识结构。这种方法的优点在于既考虑到了引用在科学知识传播中的作用，又是以代表具体概念的语言作为分析对象。因此，它既综合了共引分析和共词分析的优点，又能在一定程度上解决两者各自存在的问题。同时，被引内容中还包含着可以反映某一研究点发展程度和发展趋势的语言特征，因而可以提供比引文和关键词更为丰富的信息。

为了验证这种方法的可行性，本研究选择第二语言教育为分析领域，重点回答以下三个研究问题：

1. 以被引内容为分析对象，是否能够反映出领域内主要的知识结构和发展轨迹？

2. 被引内容中的语言特征，是否能够预测某一研究点的发展程度和发展趋势？

3. 被引内容分析是否能够解决引文分析存在的部分问题？

为了回答以上问题，本研究首先收集了第二语言教育的四种权威期刊在 1999—2001、2002—2004 和 2005—2007 三个年段发

表的 732 篇研究论文，并从每篇论文中提取出包含被引用者姓名的句子作为被引内容。在此基础上，本研究分别进行了基于被引内容的聚类分析、语言特征分析和作者分析。

在聚类分析中，首先借助 CLUTO 软件将三个年段的论文分别聚类，接着从各类团的被引内容中提取出代表核心概念的特征词和特征短语，然后分析所有这些特征在三个年段的分布变化情况。最后综合以上分析结果，对第二语言教育近九年的主要研究方向和发展轨迹进行总结。

在语言特征分析中，以两个发展状态截然不同的研究点——动词论元结构研究和超音段特征研究——为分析对象，进行包括引用格式、转述动词、时体和立场标记语等语言特征在内的对比分析。

在引文分析中，首先从被引内容中提取出各个研究方向中被引用频率最高的十位作者，然后再分析作者分布与研究方向之间的对应关系。

以上分析的主要结果为：

1. 通过被引内容聚类，可以形成比较清晰的类团。借助特征词和特征短语，能够比较顺利地推测出各类团所代表的研究方向。通过比较特征词和特征短语在不同年段的分布差异，可以发现第二语言教育的发展变化轨迹。内部验证和外部验证的结果可以证明被引内容聚类分析的有效性。

2. 发展程度和发展趋势不同的研究点，在语言特征的使用上有差异。具体来说：在即将消失的研究点中，更倾向于出现句中引用、话语类动词和过去时、现在时等语言特征；而在刚刚诞生的研究点中，句外引用、研究类动词和完成时等语言特征的比例会大幅增加。同时，后者还会出现大量能够帮助引用者营造研究空间的立场标记语。

3. 作者类团并不适合用来解释被引内容聚类分析的结果，主要表现在：作者类团的解释需要分析者事先对其研究领域有所了解；在个别研究方向中，不存在具有代表性的高频被引作者；个别作者的研究兴趣可能涉及多个研究方向。

根据以上结果，本研究得出以下结论：

1. 以被引内容为分析单位，可以发现领域的知识结构和发展轨迹。

2. 被引内容中的语言特征，可以预测某一研究点的发展程度和发展趋势。

3. 被引内容分析可以部分解决引文分析存在的问题。

本研究的意义在于：提出了一种新的分析领域知识结构和发展态势的方法；实现了多个学科引文研究方法的结合；可以为被引内容数据库的构建提供参考；可以揭示第二语言教育的研究状况。

关键词：被引内容　知识域　趋势分析　语言特征　引文分析　第二语言教育

Abstract

Research on the structure and trends of a certain knowledge domain has always been of great importance. In Bibliometrics, the two most widely used techniques for mapping a research area are Co-citation Analysis and Co-word Analysis. Co-citation Analysis, which puts citation in the central place, holds that citation is the principal way of knowledge communication, whereas Co-word Analysis places great emphasis on key words, stressing that key words are the most important and direct carrier of scientific concepts and ideas. Despite their respective merits, both methods are embedded with unsolved problems. Some scholars have, therefore, suggested that the two methods be combined so that they can complement each other. To respond this suggestion, the present study proposes a new method named Cited Content Analysis, in which the cited content appearning in full texts is anaylzed to reflect the topical structure of a research field. The advantage of the method is that it takes into consideration both citation as a vehicle for knowledge communication and linguistic content as a conceptual carrier, thus combining the strengths of both Co-citation Analysis and Co-word Analysis while solving their inherited problems to some extent. Moreover, the analysis on the content in citation taps the linguistic features which provide clear clues of the developmental phase and trend of a certain research area, thus revealing much richer information than the analysis relying on either citation only or key words alone.

The area of Second Language Education chosen for the purpose of testing the feasibility of the new method, this study addresses the following questions:

(1) To what extent can Citation Content Analysis identify the structure and trends of the area of Second Language Education?

(2) To what extent can Citation Content Analysis predict the developmental maturity and trends of certain research topics within the area of Second Language Education?

(3) To what extent can Citation Content Analysis solve the problems inherited in traditional Citation Analysis?

The corpus built for the purpose of this study consists of 732 research articles published in four globally prestigious journals in the area of Second Language Education in three periods: 1999-2001, 2002-2004 and 2005-2007. The citation content of each article was collected by extracting from the full text all the sentences that contain a cited author name. Based on the corpus, three analyses were conducted: the citation-content-based cluster analysis, the linguistic analysis, and finally the citation analysis.

In the cluster analysis, the CLUTO software was used to divide articles into clusters in the three periods respectively. A set of features (words and phrases) extracted from the cited content were used to reflect upon the major topics and concepts of each cluster. Then a frequency analysis was conducted for all the features over the three periods. Based on the findings of the above analyses, major research topics and their developmental trends of Second Language Education as a discipline were summarized.

The linguistic analysis compared citation contents on *argument structure* and *suprasegmental*, which are two research topics at very

different developmental stages. The comparison was made in terms of citation type, reporting verb, tense and aspect, and stance markers.

In the citation analysis, the top ten most frequently cited authors were identified in each cluster of research trend, which was followed by a corresponding analysis between author distribution and research area mapping.

The main findings of the above three analyses are summarized as follows:

1. The citation-content-based cluster analysis can produce clear-cut clusters. Features extracted from the cited content make it easier to predict research topics each cluster represents. The frequency analysis of cluster features can reveal the developmental trends in Second Language Education to some extent. Both internal and external verification yielded results supporting the validity of Citation Content Analysis.

2. There exist significant differences in the use of linguistic features between two research topics of dramatically different developmental phases. A maturely developed research topic is more likely to favor integral citation, discourse verbs, present tense and past tense, while a newly developed research topic tends to display more cases of non-integral citation, research verbs and perfect aspect. Besides, greater use of stance markers, which help citers indicate potentials for future research, were found on newly developed research topic.

3. It was found that author-based Citation Analysis did not yield convergent results with Citation Content Analysis. Possible reasons might be: (1) Citation Analysis creates clusters of authors rather than topics, and these can hardly be interpreted if the analyst knows little

about the research interests of those authors; (2) it happens in some rare cases that representative authors of a certain research topic are not frequently cited; and (3) some authors may not just stick to a single topic.

Based on the above findings, this study concludes with answers to the three main research questions:

1. Citation Content Analysis is a reliable method for mapping the domain structure and trends of a discipline.

2. The analysis of the linguistic features in citation content can reliably predict the developmental stage and trend of a research topic.

3. Citation Content Analysis can at least partly solve the problems that Citation Analysis cannot overcome.

The significance of this study lies in that it explores a new method for analyzing the structure and maturity of a research area, realizes the methodological integration of citation studies in different disciplines, touches on some technical issues concerning the construction of cited content database, and finally gives an insightful picture of the research progress within Second Language Education.

Key words: cited content, knowledge domain, trend analysis, linguistic feature, citation analysis, Second Language Education

目 录

表 目

第三章

第四章

第六章

第七章

第八章