蒙古语语言学研究主题分析

那日松[1]，齐力格尔[2]

(1.杭州师范大学 国际,浙江 杭州 311121;2.内蒙古大学 蒙古学学院,内蒙古 呼和浩特 010021)

摘要： 本文以中国知网中所有与蒙古语语言学研究相关的 899 篇论文题录为实例，基于论文关键词共词分析法，对关键词聚类与多维尺寸分析结果，可以呈现出蒙古语语言学的研究主题，并利用社会网络分析方法，通过绘制共现网络知识图谱与战略坐标图，进一步揭示该研究领域结构的内部联系及其特征。目的是更好地了解和掌握蒙古语语言学研究的现状及趋势，为今后把握蒙古语语言学研究方向提供必要的数据支持。

关键词： 蒙古语;语言学;主题分析;共词分析;社会网络分析方法

中图分类号码: H212　　**文献标识码：** A

随着自然语言处理、人工智能等与语言学相关领域的迅速发展，面向少数民族语言文字的研究也越来越活跃起来。为了更科学和深入地了解蒙古语语言学的研究现状、结构与发展动态，我们以在线文献为实例，通过计量、社会网络分析等方法来展开该项研究。文献统计分析是了解一个学科领域研究现状的一种途径。Law 等学者曾指出文献对于把握学科研究结构和发展的作用与优势，大部分研究领域的主要学者都将研究成果贡献于科技文献。[1][2]

一、数据采集

中国知网①是当前国内最大的具有垄断地位的集各种全文学术信息于一体的网站，包括中国学术期刊网络出版总库、中国博士学位论文全文数据库、中国优秀硕士学位论文全文数据库、中国重要会议论文全文数据库、中国重要报纸全文数据库、中国年鉴网络出版总库、中国标准数据总库等在内的中国知识资源总库（China Integrated Knowledge Resources Database)。

我们在中国知网中通过全文高级搜索“蒙古文、蒙古语、蒙文、蒙语”，并对获得的论文再进行人工剔除不属于语言学研究领域的文章后，共获得了 899 篇论文，时间跨度在 1961 年到 2015 年之间。论文按年份分布如图 1 所示。由于 2015 年到目前为止只搜索到 6 篇论文，因此图 1 中未作考虑。从论文数的年度分布来看，论文数在逐年增加。

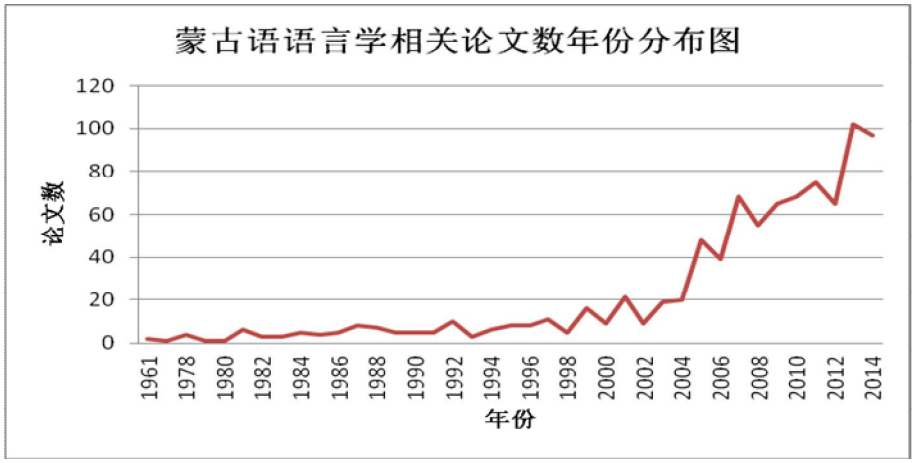

图 1 蒙古语语言学研究相关论文数年份分布图

由于蒙古语语言学研究相关论文的数量不太高，因此我们在此考虑了所有期刊和硕博论文，并不只是局限于核心期刊。再者，我们在国际期刊内也进行了相关搜索，但是论文数量较少，所以未做考虑。下表 1 中列出了相关论文数排在前 10 的期刊。

表 1 蒙古语语言学期刊论文采集处理结果

期刊名称	论文数
民族语文	47
中文信息学报	31
内蒙古大学学报（自然科学版）	28
满语研究	24
内蒙古大学学报（哲学社会科学版）	21
内蒙古师范大学学报（哲学社会科学版）	21
内蒙古民族大学学报（社会科学版）	20
中央民族大学学报（哲学社会科学版）	17
内蒙古师范大学学报（自然科学汉文版）	16
内蒙古大学学报（人文社会科学版）	16
中央民族大学学报	13
论文量/关键词总个数/关键词总频次/平均每篇关键词数	899/2505/4577/5.09

从以上期刊信息来看，蒙古语语言学的研究在人文社科、自然科学、交叉学科（中文信息学报）都有涉及。

二、数据分析

以采集到的 899 篇论文题录为实例，基于论文关键词共词分析法，对关键词聚类与多维尺寸分析结果，可以呈现出蒙古语语言学的研究主题，并利用社会网络分析方法，通过绘制共现网络知识图谱与战略坐标图，可进一步揭示该研究领域结构的内部联系及其特征。

共词分析法是内容分析法的一种，其认为两个能够表达论文主题内容的词条在一篇论文中同时出现，则表明二者具有一定的共现关系，共现次数越多，则关系越强[2]。聚类分析和多维尺度分析法，用来构建聚类图和多维尺度图谱，聚在一起的若干关键词可构成一个研究主题领域。利用社会网络分析方法，可绘制网络知识图谱并呈现出各个研究主题在相互作用下的分布情况（核心与边缘），因知识图谱并不能很好的反映主题领域的成熟度，难以判定某研究领域的成长趋势，还将基于共现矩阵构建战略坐标图，进一步解析各个研究领域的特征以验证结论。

（一）关键词频率分析

我们将蒙古语语言学研究领域的所有论文（899 篇）集合起来，对该集合内所有关键词进行词频的统计分析，认为词频较高的关键词为用来确定该研究领域的研究热点主题词。在操作过程中对于论文关键词并没有进行删减或对同义相似词的词频进行合并等操作。表 2 所示的是中国知网中蒙古语语言学研究相关论文中高频率关键词列表。我们列出了出现频率较高的 30 个关键词。

表 2 中国知网蒙古语语言学高频关键词列表

	高频关键词（共 2505 个）			高频关键词（共 2505 个）	
序号	关键词	频次	序号	关键词	频次
1	蒙古语	327	16	八思巴文	12
2	蒙古文	166	17	语言态度	12
3	长元音	41	18	偏误分析	12
4	汉语	26	19	阿尔泰语系	12
5	蒙古人	24	20	英语	12

6	附加成分	23	21	元音字母	12
7	蒙语	23	22	编码	12
8	突厥语	21	23	语音合成	11
9	机器翻译	18	24	喀尔喀	11
10	语料库	17	25	词干	11
11	八思巴字	15	26	比较	11
12	语言模型	15	27	语音	11
13	Unicode	13	28	对比	10
14	维吾尔语	13	29	喀喇沁	10
15	词缀	13	30	蒙古文字	10

从以上关键词的频率列表能大致发现蒙古语在“长元音、附加成分、机器翻译、语料库”等方面成果相对较多。也能大致看出当前蒙古语言学研究的一些热点，例如：机器翻译、语料库、语言态度、偏误分析等，这些研究主题也与当前国内语言学研究方向较一致。

关键词的出现频率只是从一个很小程度上反映蒙古语语言学研究的一面。下面我利用共现分析法来获得关键词的共现频率，再以聚类树状图和多维尺度图谱来揭示蒙古语语言学的研究主题及结构等。

（二）聚类树状图和多维尺度图谱

聚类分析通过聚类算法将关联密切的关键词聚集在一起形成类别（研究主题）的过程，可以用来揭示蒙古语语言学的研究主题结构。我们选择共现词频较高的 30 个单元，得出高频关键词共现相似矩阵，将相似矩阵导入 Ucinet②进行层次聚类分析，得到如图 1 所示的蒙古语语言学高频关键词聚类树状图。

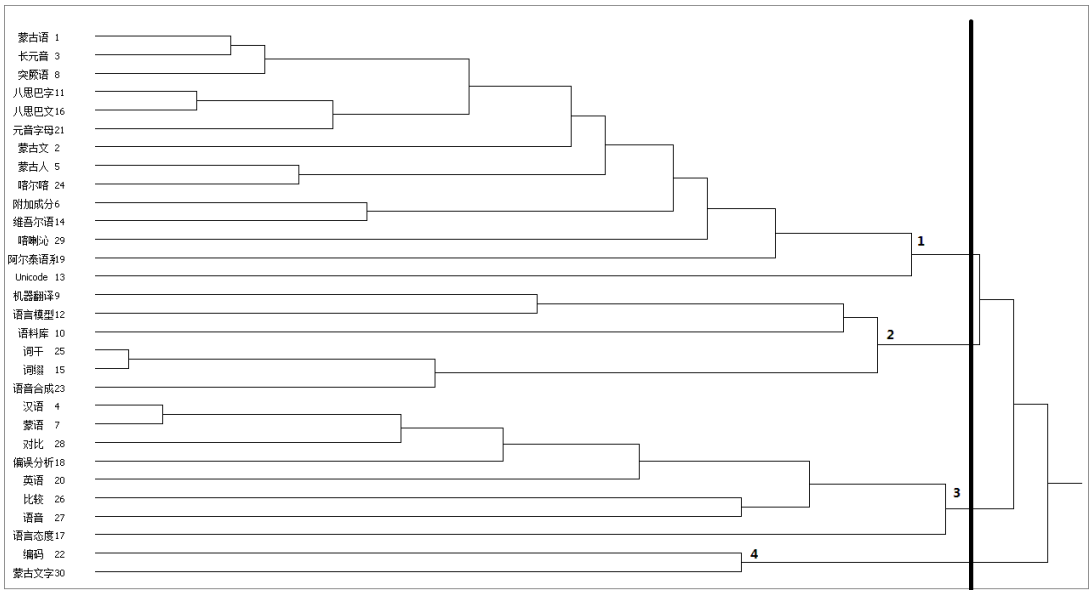

图 2 蒙古语语言学高频关键词聚类树状图

多维尺度分析通过计算关键词之间的距离来挖掘主题结构。与聚类树图相比，多维尺度分析可以在较低维空间中直观地判断出某研究领域在学科内的位置。将相异矩阵导入 SPSS③进行多维尺度分析，得到如图 2 所示的多维尺度图谱。

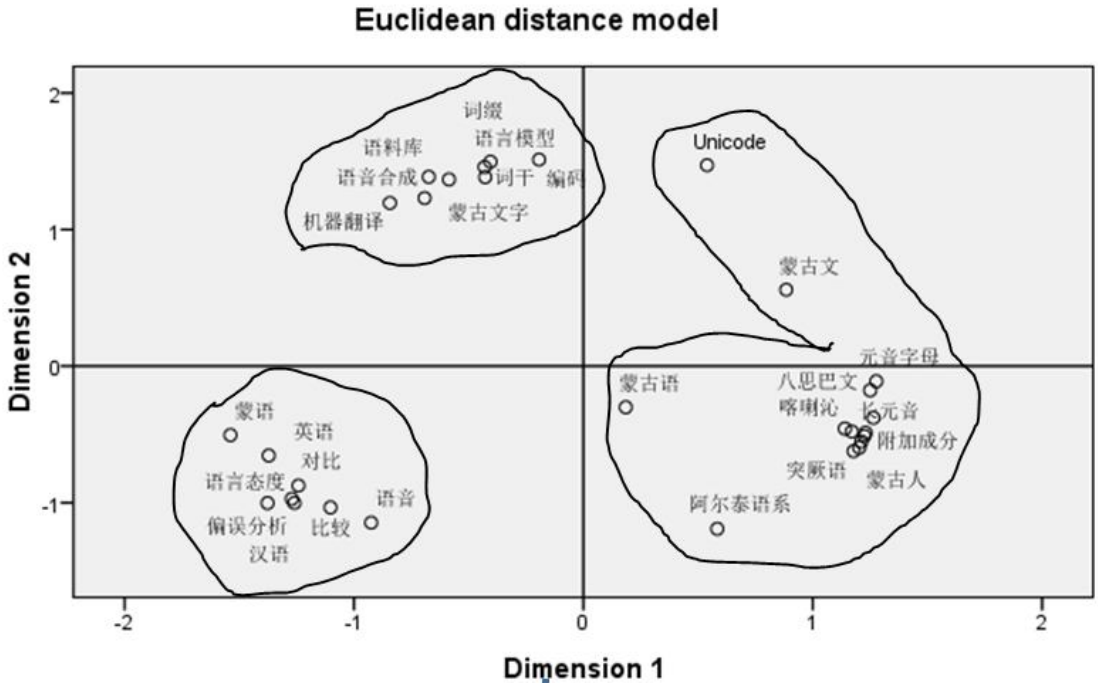

图 3 蒙古语语言学高频关键词多维尺度图谱

从以上聚类图和多维尺度图谱的一致性出发，图 1 中数字 1～4 为聚类分析后所获得的 4 个研究主题，结合图 2 多维尺度图谱结果，将主题 2 与主题 4 进行了合并，并给出蒙古语语言学研究可大致分为三大主题结构：1）蒙古语特点研究：包括蒙古语、蒙古文、长元音、蒙古人、附加成分、突厥语、八思巴字、Unicode、维吾尔语、八思巴文、阿尔泰语系、元音字母、喀尔喀、喀喇沁；2）蒙古语与其他语言对比研究：包括汉语、蒙语、语言态度、偏误分析、英语、比较、语音、对比；3）蒙古语语言信息处理：包括机器翻译、语料库、语言模型、词缀、编码、语音合成、词干、蒙古文字。

（三）网络知识图谱与战略坐标图

网络知识图谱以关键词为结点，关键词之间的共现为连线，结点位置越居中认为该关键词越核心，连线越粗认为关键词之间的关系越紧密。我们选择词频较高的 30 个单元，得出高频关键词共现相似矩阵。在 Ucinet 中计算每个结点的中心度（Degree Centrality），以此来控制网络结点大小，其中不同颜色（红色：主题 1；蓝色：主题 2；绿色：主题 3）的结点表示属于不同类别，即不同研究主题，如图 3 所示。图 3 网络知识图谱展示了蒙古语语言学研究主题结构的内部关系。图中主题 1（红色）较集中，内部联系紧密；而主题 2（蓝色）和主题 3（绿色）较分散，但是“蒙语”和“汉语”相关的研究、“词缀”和“词干”相关的研究较多（连线较粗）。其中“Unicode、蒙古文字、编码”3 个关键词位置分散，与 3 个主题之间都有一定距离。图中结点边缘颜色越重越代表在近 5（不包括 2015 年）间被研究的次数越是频繁，从图上来看，3 个主题研究都比较频繁，主题 1 中有些关键词在近几年出现频率有所降低。

下面我们制作战略坐标图来解释每个研究主题的重要程度及其特性。

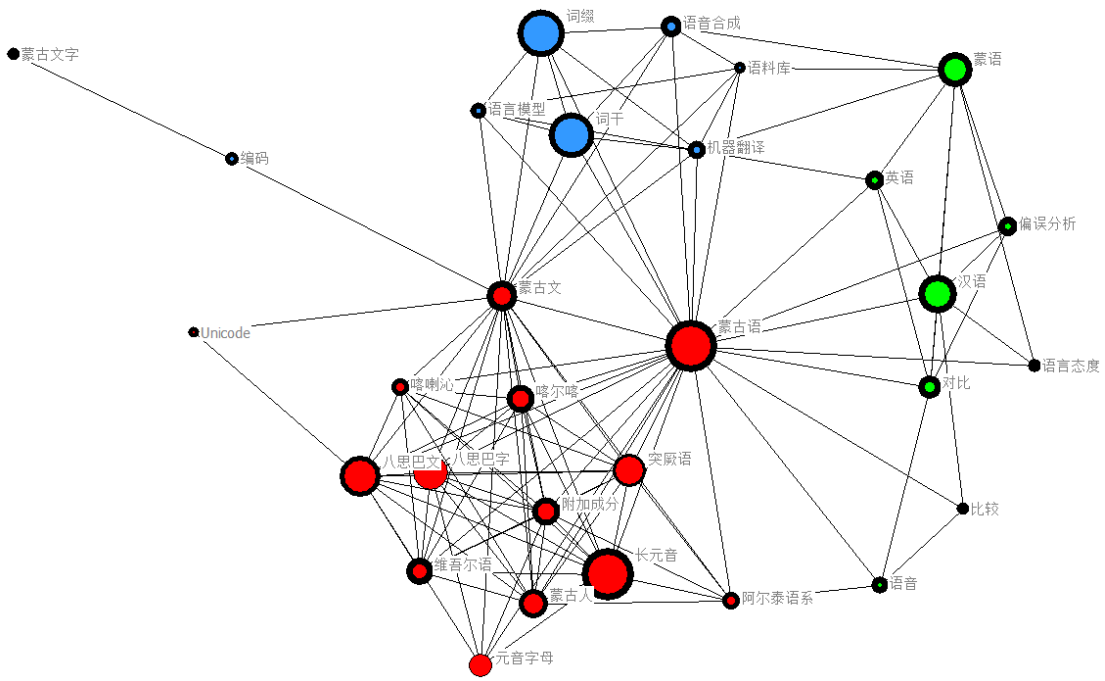

图 4 蒙古语语言学关键词共现网络知识图谱

战略坐标图由 Law[1]等人提出，用来揭示各主题聚类内容的强度和类间的关系。其中横轴代表向心度（Centrality），即某研究领域在整个学科的核心程度，揭示研究领域与其他主题领域之间的关联程度，纵轴代表密度（Density），即某研究领域的内部强度，揭示某研究领域维持和发展自身的能力。

基于向心度与研究领域核心程度一致，密度与研究领域成熟程度一致的思想，从图 4 可以得出，三大研究领域中，蒙古语语言本身特点相关研究(主题 1)向心度低、密度高，说明该主题研究处于整个研究的边缘（非核心）位置，已经受到关注，且被很好的研究过；蒙古语与其他语言对比研究（主题 2）向心度和密度都低，说明也是处于整个研究的边缘，研究尚不成熟；蒙古语语言信息处理研究（主题 3）向心度高、密度低，说明该主题与其余各主题有广泛的联系，即处于所有研究主题的核心，但不成熟。

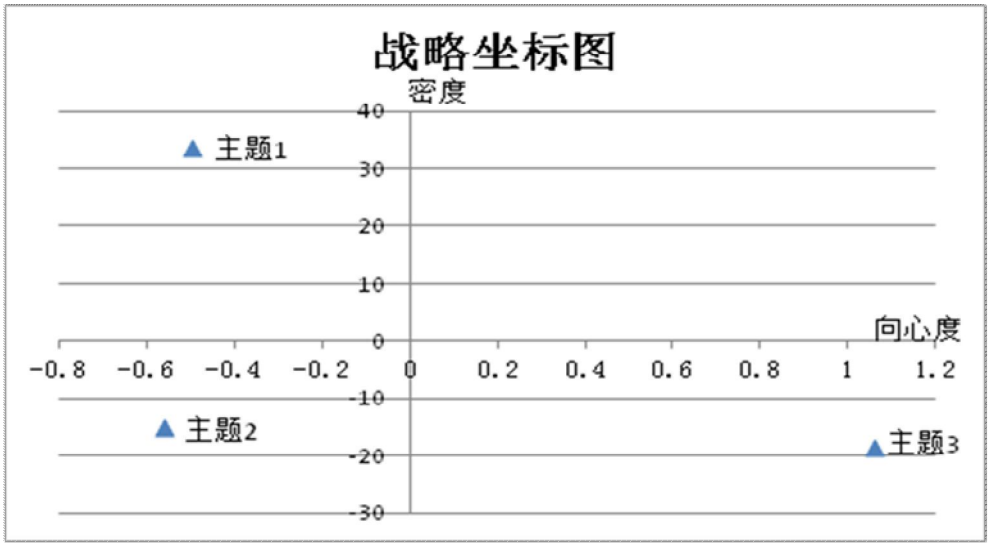

图 5 战略坐标图

三、讨论及结论

从中国知网所获得的所有蒙古语语言学相关论文的数量及所涉及期刊数来看，蒙古语语言学研究的成果相对于其他语种（英语、汉语等）的较少。虽然中国知网中的论文无法覆盖蒙古语语言学研究的所有成果，但是可以理解为抽样数据的话，对该数据的分析结果也能在一定程度上反映该研究领域的情况。从以上的研究数据中，我们得出在蒙古语语言学研究领域有 3 个研究主题比较突出；其研究特点和趋势为：蒙古语语言信息处理研究（主题 3）与其他主题研究联系紧密备受关注，但是研究不够成熟，需增强研究和成果的输出；蒙古语语言本身特点研究(主题 1)和蒙古语与其他语言对比研究（主题 2），处于蒙古语语言学研究的边缘位置，但是成果的成熟度不同。但是随着今后蒙古语语言学研究成果的不断增加，三类主题之间的关系会发生变化，且向心度和密度也会有所改变，但是从目前数据分析的结果来看，主题 3 属于研究的核心位置。

总结本文的研究：以中国知网论文为数据源，检索和获取了所有与“蒙古语语言学”相关的学术论文，通过高频关键词的共词聚类和多维尺度分析，确定了蒙古语语言学主流研究主题，通过绘制网络知识图谱与战略坐标图考察了蒙古语语言学研究各个主题的发展现状及研究地位，最后在此基础上给出了加强研究的建议。

注释

①http://zh.wikipedia.org/wiki/CNKI

②http://baike.baidu.com/view/2343008.htm

③http://baike.baidu.com/view/130328.htm

参考文献

[1]Law J, Bauin S, Courtial J, et al. Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification[J].Scientometrics,1998,14(3-4):251-264.

[2]刘启元,叶鹰. 文献题录信息挖掘技术方法及其软件 SATI 的实现[J].信息资源管理学报,2012,(1):50-58.

[3]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 chemistry[J]. Scientometrics, 1991, 22(1):155-205.

[4]Cottrill C A, Rogers E M, Mills T. Co-citation analysis of the scientific literature of innovation research traditions: Diffusion of innovations and technology transfer [J]. Knowledge, 1989, 11(2):181-208.

[5]杨颖,崔雷.基于共词分析的学科结构可视化表达方法的探讨[J].现代情报, 2011,31(1):91-96.

Analysis on the subject of Mongolian Linguistics

Narisong[1],Qiligeer[2]

(1.School of International Education, Hangzhou Normal University;2. School of Mongolia Studies, Inner Mongolia University)

Abstract: Taking 899 articles published in CNKI（China Knowledge Resource Integrated Database）as the sample, this paper revealed three potential research fields of Mongolian Linguistics research area based on the consistency between clustering analysis and multidimensional scaling analysis results, and figured out relations and features of subject areas by interpreting the network knowledge map and strategic diagram. The goal is to better understand and grasp the status and trends of Mongolian linguistics research, provide necessary data support for the future research direction of Mongolian linguistics.

Key words: Mongolian; Linguistics; subject analysis; co-word analysis; social network map

收稿日期: 2015-03-19;

作者简介: 1.那日松（1980—），女，内蒙古兴安盟人。主要从事蒙古文信息处理、计量语言学方面的研究；2.齐力格尔（1990—），女，内蒙古通辽市库伦旗人。主要从事蒙古文信息处理方面的研究。