硕 士 学 位 论 文

基于主题模型的文本挖掘的研究

Research on Text Mining Based on Topic Model

作 者 姓 名：王　　亮

学科、 专业：计算机应用技术

学　　号：21209188

指 导 教 师：张 绍 武

完 成 日 期：2015-04-25

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年____月____日

摘　　要

随着大数据时代的到来以及互联网的不断发展，以文本资源为典型的各种资源呈爆炸式增长，从纷繁复杂的文本资源中挖掘有潜在价值的、用户感兴趣的信息变得愈加困难。研究人员钻研了各种算法、设计了各种工具以便用户能够帮助我们快速、有效地理解大量文本内容，这些工作归结在一起主要是文本主题挖掘技术。主题是文本的灵魂，发掘文本主题是用户去粗取精、去伪存真、从感性认识到理性认识的飞跃过程，是用户对文本深入开发的过程。本文首先利用 LDA 主题聚类技术挖掘期刊论文中的主题信息，发现主题模型对期刊推荐方法有较大的帮助，但是也存在一定的问题，例如 LDA 主题数目确定困难、主题随时间发生变化。因此本文又深入探索了如何挖掘主题随时间的变化并可视化展示，即主题演化信息及可视化展示问题。发现主题演化信息并展示对了解主题的研究热点、演变趋势以及对未来进行预测等有很大帮助。文本主要内容如下：

首先，本文研究了主题模型对期刊推荐的应用价值，以潜在狄利克雷分配（LDA）主题模型的结果为基础，结合 SVM 分类算法，大大提升了分类算法在期刊推荐的效果。论文投稿不仅牵扯到论文和期刊的研究方向，还牵扯到论文以及期刊的质量高低，为了在学者投稿时帮助学者选择合适的期刊，本文对 LDA 主题模型进行深入学习，结合 SVM 分类算法进行期刊推荐，实验发现基于 LDA 的期刊推荐算法明显优于基于 SVM 的期刊推荐方法、基于内容的期刊推荐方法、基于用户的期刊推荐、基于期刊相似度的推荐方法，而且本文在对推荐错误的论文进行研究发现有些期刊存在发表与自己研究主题不大相符的问题。

其次，本文利用分层的狄里克雷分布（HDP）主题挖掘算法研究了主题演化问题（主题的分流、合流，主题的渐增、渐减，主题的新生、消亡），并利用主题河将主题演化信息生动地展示出来。本文以汽车专利为出发点研究汽车产业中的主题演化信息，将 HDP 算法应用到汽车专利主题聚类中，通过当前主题以及加入历史信息之后的主题变化来发现主题之间的分流、合流等关系，然后将主题以及主题的分流、合流信息利用可视化技术直观展示出来。实验发现中文汽车专利有三个重要主题，而且各个主题之间有分流、合流，有逐年递增也有逐年递减，有新生主题也有消亡主题等各种形式。

关键词：主题挖掘；分层狄利克雷分布；潜在狄利克雷分配

Research on Text Mining Based on Topic Model

Abstract

With the continuous development of the Internet, a variety of resources, led to text resource are in explosive growth, finding the valuable information from them which the user interested is becoming more and more difficult. In order to do that, amount of researchers have studied a variety of algorithms and developed a lot of text tools to allow users to effectively organize and manage the text information, thus helping users to quickly, accurately and comprehensively to find the information they need. The work is mainly topic mining technology. Topic is the soul of the text, mining the topic information is the process of discarding the dross , selecting the essential , from perceptual to rational knowledge and further development of the text. At first, we used LDA topic model to mining the topic information of paper and found that the LDA topic model is useful to journal recommendation, but we also found some problems, such as the topic number is very difficult to determine and the topic changes over time. Therefore, we studied how to dig out the topic evolution and visualize it more vividly. Finding the topic evolution means a lot to understand topic hotspot, topic evolution trend and prediction of the topic. The main contents are as follow.

Firstly, we studied the value of topic model to journal recommendation. We combined the latent dirichlet allocation topic model and SVM classification model greatly enhanced the result of journal recommendation. Paper submission is a very difficult academic and practical problem, it not only involves the research topic but also the quality of the paper and the journal. To help scholars to choose the right journal when they submit for publication, the paper combines the result of LDA topic model and the SVM classification method to recommend the right journals. Compared with other models (svm-based journal recommendation, content-based journal recommendation, user-based journal recommendation, journal similarity-based journal recommendation), topic-based journal recommendation has a better performance. And, we discovered that some journal exist the problem that they published some paper which were not consistent with the research topic.

Secondly, the paper use the hierarchical dirichlet processes topic mining method to study topic evolution, such as the shrinking, the expanding, the newborn, the perishing , the increasing, the decreasing of the topic, and use the ThemeRiver visualization method to display them vividly. The paper takes vehicle patent as a starting point to study the topic evolution of automotive, and uses the HDP topic model to cluster the patent data and mine splitting and merging of the topics by comparing the topics of each year and the topics with history data clustered by HDP and then visualizes the relationship of the topic information using stacked

graph. The paper discovers that there are three major topics of the vehicle patent data and here are splitting and merging among different topics, shrinking of the topic, expanding of the topic, newborn of the topic and perishing of the topic.

Key Words：Topic Mining; Hierarchical Dirichlet Processes; Latent Dirichlet Allocation

目　　录

1 绪论

1.1 研究背景

随着大数据时代来临以及互联网的不断发展，以文本资源为首的各种资源呈现爆炸式增长。如何从纷繁复杂、杂乱无章的文本资源中挖掘有潜在价值的、用户感兴趣的信息变得愈加困难。为此研究人员钻研了各种算法并设计开发了各种工具以便用户能够对这些纷繁复杂的文本集进行管理，从而能够快速、全方面地帮助用户找到用户需要的内容，这些内容归结在一起是文本挖掘相关技术。最早由 Feldman 等人[1]于 1995 年正式提出的文本挖掘技术（Text Mining）是指在大规模文本中挖掘出潜在的、有用的模式的过程，亦被称为文本知识发现。此后，文本挖掘相关技术迅速发展，其研究内容主要有关联规则挖掘、文本分类、特征选择、文本聚类、主题挖掘等。到目前为止，文本挖掘技术已成为数据挖掘技术中一个非常重要的组成部分，使用文本挖掘算法来解决上述问题为我们的研究提供了一种思路。

在文本挖掘领域中，主题可以看成是词项的概率分布，可以作为知识发现的基本构建块，挖掘文本主题及其演化过程亦是比较重要的课题，也是本文的一个重要的研究内容。在海量文本信息中挖掘出主题信息以及主题演化信息有很大的作用，例如掌握汽车领域技术的主题分布以及主题发展演化情况，对汽车研发人员以及决策人员了解汽车领域的研究内容和研究热点以及技术发展趋势等有很大的帮助，可以为国家和企业的决策提供技术支持。

文本资源中的专利是非常有价值的内容，是集商业情报和技术情报于一体的技术载体，其格式规范、用语严谨，在技术内容方面具有系统详尽、分类科学以及及时可靠的特点。基于以上特点，对专利进行深度技术挖掘，从宏观层面来看可利用专利技术主题演化信息来预测未来技术发展方向，同样可用来进行竞争技术情报分析，而且对技术人员及决策人员掌握当前技术的研究状态、未来的研究方向都有很大的帮助；从微观来看可挖掘技术创新的方法细节，取长补短从而改善其它专利。汽车产业是国民经济的支柱产业之一，对我国国民经济和社会经济发展发挥着举足轻重的作用。基于以上特点本文以中国专利数据库中的汽车专利为出发点研究汽车产业中的主题演化信息，并结合文本可视化技术对主题演化进行直观、有效地展示以便研究人员从中挖掘出更多潜在有价值的信息。

文本资源中的论文同样是非常有价值的内容，论文是来进行科学研究和描述科研成果的文章。一篇论文从构思到写作完成需要少则一两个月多则一年半载甚至更长时间。

论文作者选择合适的期刊来发表文章，期刊编辑找专家进行审稿，快则几个月慢则一年半载，而且审稿之后给予各种理由拒稿的情况也是时有发生，然后论文作者需要重复以上工作，直到论文被录用为止。因此，将论文推荐到研究方向相近的期刊而且论文质量和期刊质量亦相近的期刊是非常重要的。因此，本文拟利用主题挖掘技术挖掘论文和期刊的主题信息，将论文推荐到相应的期刊中。

1.2 研究现状

1.2.1 主题挖掘

近年来在文本挖掘领域，统计主题模型（Statistical Topic Model）得到很好的应用与发展（如应用到文本分类、主题检索、话题发现、主题演化等领域），它的核心思想是通过调节参数对文本集的主题进行定位，从而对文本中深层的、隐含的语义信息进行挖掘，下面详细介绍主题模型的发展历程：

1990 年 Deerwester S C 等人[2]提出的隐性语义索引（Latent Semantic Indexing, LSI）模型是最早的主题模型，该模型虽然不是真正意义上的概率主题模型，但是模型的提出被认为是为主题模型发展奠定了坚实的基础；1999 年 Hofmann T 等人[3]改进了隐性语义索引模型并提出概率隐性语义索引（Probabilistic Latent Semantic Indexing, pLSI）模型，该模型被大多数研究人员公认为是第一个真正的概率主题模型，它的提出促进了概率主题模型的发展；在此之后的 2003 年，著名的 Blei D M 等人[4]在 pLSI 的基础上提出潜在狄利克雷分配（Latent Dirichlet Allocation, LDA）模型被众多专家学者认同并给予极高的评价，主题模型也受到越来越多研究人员的重视，主题模型也得到了广泛的应用；从此之后研究人员为了满足不同的需求提出了各种基于 LDA 的改进模型，如 Wang 等人[5]提出的动态狄利克雷分布（Dynamic Latent Dirichlet Allocation，dLDA）模型，将时间加入到主题模型中，解决了主题动态变化的问题； Lancichinetti 等人[6]在复杂网络中利用一致聚类进行研究，在动态网络聚类的研究中取得一定的研究成果；2006 年 Teh 等人[7]提出的层次狄利克雷过程(Hierarchical Dirichlet Processes, HDP)改进了 LDA 的主题数需要人工确定的缺点，并且可以将其应用于主题演化方面。

LDA 模型作为一种能够提取文本隐含主题的非监督机器学习算法拥有很好的泛化能力，是机器学习、文本挖掘领域很流行的一个算法。基于 LDA 模型的研究非常广泛[8-10]，截止 2015 年 3 月 11 日，在 Google Scholar 中搜索 LDA 的引用次数已经达到 10809 次，Web of Science 中引用次数达到 2398 次，这足以说明 LDA 在主题模型中举足轻重的作用，因此本文在挖掘期刊的选择时的主题模型使用的是 LDA 模型。

HDP 模型作为一种能够自动确定隐含主题个数的主题挖掘算法，拥有很好的主题适应能力，在话题演化、主题演化方面有广泛的应用[11-13]。基于 HDP 模型的研究同样很广泛，截止 2015 年 3 月 11 日，在 Google Scholar 中搜索 HDP 的引用次数已经达到 1940 次，Web of Science 中引用次数达到 413 次，这说明 HDP 算法也得到广大研究人员的认可，因此，本文在主题演化方面的研究使用 HDP 算法来进行研究。

1.2.2 期刊推荐方法

国内外关于期刊推荐方法的研究非常少，对于期刊的研究主要集中在期刊影响力评价等方面[14-15]；而对于论文的研究也很少，主要研究集中在论文评价[16-17]、论文合作关系网络[18-19]、引文分析[20-22]等。由于没有特意针对期刊推荐的研究现状，那么只能针对利用文本挖掘算法进行期刊推荐的方法探讨其研究现状，本文使用的文本挖掘算法包括 LDA 主题模型、支持向量机（Support Vector Machine，SVM）分类算法[23]等，LDA 主题模型已经介绍过，因此下面主要介绍分类算法的研究进展状况。

分类算法本质上属于监督学习算法，算法需要通过大量已知分类类别的训练数据对分类模型的各个参数进行训练，发现其中的分类规则，然后利用训练好的模型对测试数据的归属进行预测。不同的分类模型、算法有不同的特性，当然也有不同的分类效果，其效果的好坏往往由分类准确率、分类模型的稳定性、分类模型训练速度、分类模型识别速度、分类模型的鲁棒性等标准进行衡量。

比较常见的分类算法有 1997 年 Friedl M A 等人[24]提出的决策树算法，该算法对噪声数据不敏感，是使用最为广泛的算法之一；由于决策树算法选择分类特征时一般选用信息增益（ID3 算法）或者信息增益比（C4.5 算法）最大的节点作为分类节点，因此决策树算法亦被称之为贪心算法。

朴素贝叶斯算法也是比较常见的分类算法，其核心思想就是统计概率中的朴素贝叶斯定理，主要基于已知的先验概率和条件概率这两种概率来求最大的类别（后验概率）概率的过程，该算法适用于数据集较大的情况下，其先验概率准确率会比较高，这样最终计算出来的后验概率也就会更加准确。可想而知其最大的缺点也就是在数据量比较小或者数据不完整的情况下准确率会大大降低。

1998 年 Hearst M A 等人提出的 SVM（支持向量机）算法在机器学习领域受到了广泛关注，是分类算法中最经典、最有效的算法之一，可以解决线性以及非线性等分类问题，并且能够游刃有余地解决二分类和多分类问题[24-26]，其核心思想就是利用支持向量来最大化分类边界，SVM 算法运算的复杂度与支持向量的个数有直接关系，与数据样本空间的维度没有直接关系，因此 SVM 算法可以有效地避免维度灾难。

1.2.3 专利技术演化

专利技术演化分析分两种，一种是没有引入文本挖掘知识的分析方法，即专利分类分析法（利用专利分类号和人工阅读的方式），另一种是引入文本挖掘技术的主题演化分析方法。利用人工阅读的方式无疑是最差的方式，专利阅读者主观差异性太大，每个人分析的结果都不同，且费时费力，在大数据时代不可能奏效。利用专利分类号进行技术主题分析方法其缺点是在进行分析时过于依赖专利分类号，从而专利分类号的正确与否对其分析会产生很大的影响；而且单纯地通过专利分类号对专利主题信息进行划分并不能完全满足专利主题分析的需求，因此，借助文本挖掘技术等相关方法来进行专利技术主题演化分析无疑是最简单、有效的方式。

将文本挖掘应用到专利技术主题演化分析领域，最简单的方法是利用词频统计在专利的摘要、专利标题、关键字、专利主体内容等部分的专业术语的数量，用专业技术术语来反映专业技术主题，由于词频本身固有的难以反映词与词之间的关联的特性，其分析效果之差可想而知，但是对于分析主题领域内的热门技术来说还是有可取之处；第二种方法是共词分析法，该方法在文献计量学中常见，该方法大大弥补了词频分析法的一大不足之处（该方法可以反映词与词之间的联系，可以用来挖掘词与词之间的共现强度），其具体方法包括共词网络分析、共词聚类分析等；第三种方法就是本文使用的方法，专利文本聚类方法，利用主题聚类技术可以对专利进行技术聚类，每个聚类就是一个技术主题，然后将这个技术主题利用主题词的方式进行展示，另外可以对专利按照时间进行聚类，来发现聚类信息随时间的变化，从而发现专利技术主题的演化信息，这对于掌握技术演变有很大的帮助。范宇等人[27]将 LDA 主题聚类应用到专利信息聚类中，该论文的一个非常大的缺点是论文作者根据经验事先人工确定了主题个数，而我们都知道主题个数的选取对主题聚类的准确度、聚类效果有很大的影响。郝智勇等人[28]将 LDA 主题聚类与可视化方法结合，展示了 4 年中各个主题之间的相关性聚类散点图，并没有考虑到各个主题之间的变化，亦没有展示各个主题的发展变化趋势等。

从主题模型提出以后，研究人员提出了一些文本主题分析方法，可以在从不同方面利用不同方法挖掘出用户需要的主题信息，然而主题模型产生的结果往往比较复杂，一般对普通用户来说理解起来还是非常困难的。例如，LDA 或 HDP 主题分析的结果就很难理解，LDA 或 HDP 的结果中每个文档有不同的概率属于不同的主题，而一个主题又是由一些词语来表示，其中每个词语有一个属性值表示其属于当前主题的可能性大小。为此，研究人员将主题分析技术与可视化技术相结合，利用人类对图形、图像的敏感性以及人类固有的辨识和分析能力，将文本主题挖掘的结果利用更加生动、形象的方式进行展示，使得人们可以通过眼睛从可视化图示中迅速捕获有价值的信息，从而达到进一

步分析、推理的目的，而且可视化的结果也可以很方便地进行传播。将发现的主题信息以及各个主题之间的相关关系以友好的方式呈现给用户，研究人员也设计了很多主题可视化方面的算法，如 Havre 等人[29]提出主题河(ThemeRiver)；另外 Wei 等人[30]发表的论文 TIARA，以河流的形式详细地分析用户的邮件主题信息；Cui 等人[31]提出的 TextFlow 也是以河流的形式细致、连贯地展示了主题随时间的各种变化。

1.3 本文工作

本文主要利用主题挖掘方法对专利信息以及论文信息进行挖掘，主要内容分两方面，一方面是对汽车专利利用主题演化技术发现汽车主题的变化（主题分流、主题合流、主题生长、主题消亡等），从而发现潜在有价值的信息；另一方面是将主题聚类技术应用到期刊推荐方法中将论文推荐最有可能发表的期刊上去，从而帮助研究人员更好更快的发表文章。

本文利用 word2vec[32]工具对文献进行向量化，利用 LDA 主题聚类技术、支持向量机（SVM）、基于内容的推荐、基于用户的推荐、基于期刊相似度的推荐等方法进行期刊推荐，结果显示基于 LDA 的主题聚类技术的推荐效果明显优于其他方法，具有一定的实用价值。而且，本文对影响论文水平的因素进行了简要的探讨，并将论文水平添加到上述方法中，发现按照论文水平高低将论文推荐到相应水平的期刊中其准确率明显提高。此外，本文对推荐失败的结果进行深入研究发现有些期刊存在刊登与自己期刊主题不相符的论文的情况，这种现象的存在也是本文期刊推荐结果稍差的一个原因。

本文将分层的狄利克雷过程(HDP)应用到专利主题聚类中，通过当前主题与加入历史数据之后的主题变化来挖掘主题的分流与合流，最后对主题信息利用叠式图进行可视化展示。实验结合实际的汽车专利数据进行分析研究，发现汽车专利主要分为三个大主题，而且各个主题之间有分流、合流，有逐年递增也有逐年递减，有新生主题也有消亡主题等各种形式，并发现从 2006 年开始汽车安全领域和汽车新能源领域分别独立成为一个主题并呈逐年增长的趋势，这说明汽车安全和汽车新能源越来越受到人们的重视。

1.4 本文结构

本论文工作分为四章，详细阐述了本文的研究背景、研究现状以及本人使用 HDP 主题模型在专利主题演化中的应用以及 LDA 主题模型在论文期刊选择中的应用。

第一章绪论主要综述了本文的研究背景、以及主题挖掘、专利技术分析、期刊推荐方法的研究现状。

第二章主要讲解本文需要用到的相关技术，主要包括 LDA 主题模型、HDP 主题模型、Word2vec 词向量模型。

第三章主要讲解 LDA 主题模型在期刊推荐方法的应用。

第四章主要讲解基于 HDP 算法的汽车专利主题演化研究。

在论文总结部分，讨论了本文主要的研究内容、主要工作、并对下一步主要工作进行展望。

2 相关技术

2.1 LDA 主题模型

LDA（Latent Dirichlet Allocation）在向量空间模型（Vector Space Model）[33]和统计语言模型[34]（Statistical Language Model）的基础上，改进了 PLSI 模型的产生式语义的问题，利用更富有表现力的主题层来表示文本表达式，形成一种语义上一致的话题模型。LDA 模型是目前非常流行的主题模型之一，广泛应用于机器学习、自然语言处理、文本挖掘、知识发现等多个领域[35-38]。

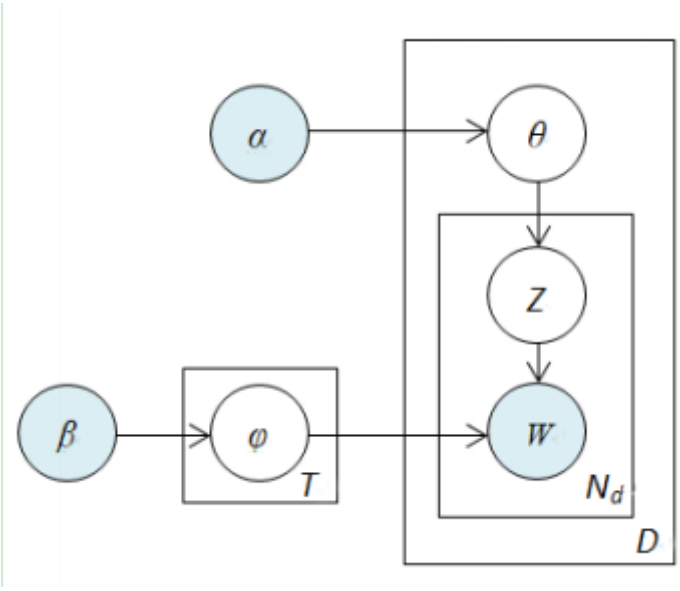

图 2.1 LDA 主题图模型

Fig. 2.1 Graphical model representation of LDA

LDA 是一种文档主题生成模型，其训练的目标是主题分布，而主题分布在训练集中是不能直接观测到的量，因此才被称之为潜在狄利克雷分布。LDA 每个待聚类文档都有一个属于各个主题的概率主题分布，其基本元素有文档（d）、文档词（w）、文档主题（z），这三个基本元素通过词袋（Bag of Words）的表示方法形成“文档-主题”、“主题-词”两层分布，“文档-主题”层的意思是一篇文档可以表示成不同主题所构成的概率分布，“主题-词”层的意思是每个主题可以表示成词汇所构成的概率分布。如图 2.1 所示，从图中可以看出 LDA 的生成过程与 PLSI 的生成过程非常相似，该模型设计的参数变量有主题数 T、文档数 D 以及第 d 篇文档的文档长度 N_d，其中参数 θ 表示“文档-主题”的分布，由 Dirichlet 先验知识 α 控制产生的在每篇文档中都是不同的，α、

β分别为利克雷先验参数，在 LDA 主题聚类过程中只采样一次，对所有的文档都是相同的，α一般取值为 $\alpha = 50/|T|$，β一般取值为 β=0.1。

LDA 的具体生成过程如下：

1.在 Dirichlet 分布中选择参数 β，利用参数 β 为每个主题 z 生成多项式分布 ϕ_z；

2.在 Dirichlet 分布中选择参数 α，利用参数 α 为每个文档 d 生成多项式分布 θ_d；

3.对于文档 d 中的每个单词 w，从多项式分布 θ_d 中选择一个话题 z∈{1,L,K}；

从上述步骤中可以看出，最重要的是求出参数 α、参数 β 以及多项式分布 ϕ_z 和 θ_d，对于模型中这几个参数的求解方法主要分为基于吉布斯采样[39]的方法和基于 EM 变分法求解。与基于 EM 变分法求解方法相比，吉布斯采样方法是一种迭代方法，其优点是易于实现而且在可以高效地在大规模文本集中抽取主题，因此在常用于 LDA 主题模型中参数的估计，本文亦使用该方法。

LDA 模型的一个缺点是需要人工确定到底整个文档分多少个主题，如果对整个文档有比较深刻的了解，那么对于主题数的确认有一定的帮助，当然这必须需要阅读大量的文档才能够有深刻的了解，费时费力。为了解决这个难题，研究人员一般通过反复计算困惑度（*Perplexity*），通过困惑度曲线来确定最优主题个数[40]。一般而言困惑度的值的大小能够反映出模型产生主题模型性能的高低也间接反映出模型推广性能的好坏，利用对主题数取值的不同来对困惑度进行计算从而得出困惑度变化曲线，从变化曲线中确定最优主题数目，其计算公式如（2.1）所示。

$$Perplexity(D) = exp\left\{-\frac{\sum_{i=1}^{M} lnP(d_i)}{\sum_{i=1}^{M} N_i}\right\} \tag{2.1}$$

其中 $p(w)$是测试集中出现的每个词的概率，具体到 LDA 模型中就是：$p(w) = \sum_z p(z|d) * p(w|z)$其中 z，d 分别指训练过的主题和测试集的各篇文档，N_i 是测试集中出现的所有词的文档长度。

2.2 HDP 主题模型

分层的狄利克雷过程是一种基于 Dirichlet 过程的层次化建模方法，与典型的主题文本聚类模型 LDA 相比，LDA 聚类过程需要根据经验人工设置聚类的个数，而对于实际情况来说一般很难知道到底有多少主题，而 LDA 的这个缺点恰恰是 HDP 的优点，HDP 算法不需要过多的人工干预，能够自动确定需要生成聚类的个数以及生成各个聚类的分布参数。

HDP 算法是基于 Dirichlet 过程的，而 Dirichlet 过程是关于分布的分布，即：Dirichlet 过程的每一个采样即为一个随机分布，而该随机分布的任意有限维边缘分布均是

Dirichlet 分布。如图 2.2 右图所示，该图为 HDP 算法的有向图表示，由图中信息可以看出，HDP 算法是基于 Dirichlet 过程的混合模型的多层形式。HDP 算法的参数包含如下几个部分：1、基分布 H，2、聚集度参数γ和α_0，基分布 H 为θ_{ji}提供先验分布。下面是 HDP 算法构造的具体过程：

第一步：基于基分布 H 以及参数γ构成 Dirichlet 过程；

第二步：以 G_0 为基分布，以α_0 为参数，利用数据构造 Dirichlet 过程混合模型；

第三步：根据该层 Dirichlet 过程为先验分布，构造 Dirichlet 过程混合模型。

与 Dirichlet 过程类似，有两种方法可以构造 HDP，不过构造关系稍微复杂一些。他们分别是：Stick-breaking 构造（图 2.2 左图）、Chinese restaurant franchise 构造。而 HDP 的采样方法主要有三种方式：基于 CRF 的后验采样算法、基于 Augmented representation 后验采样算法、直接分配采样算法，这三种方法各有优缺点，基于 CRF 的后验采样和基于 Augmented representation 后验采样算法这两个算法的缺点是实现过程相对比较繁琐，优点是算法的收敛速度快、结果比较好；直接分配采样算法收敛速度相对比较慢、效率低。

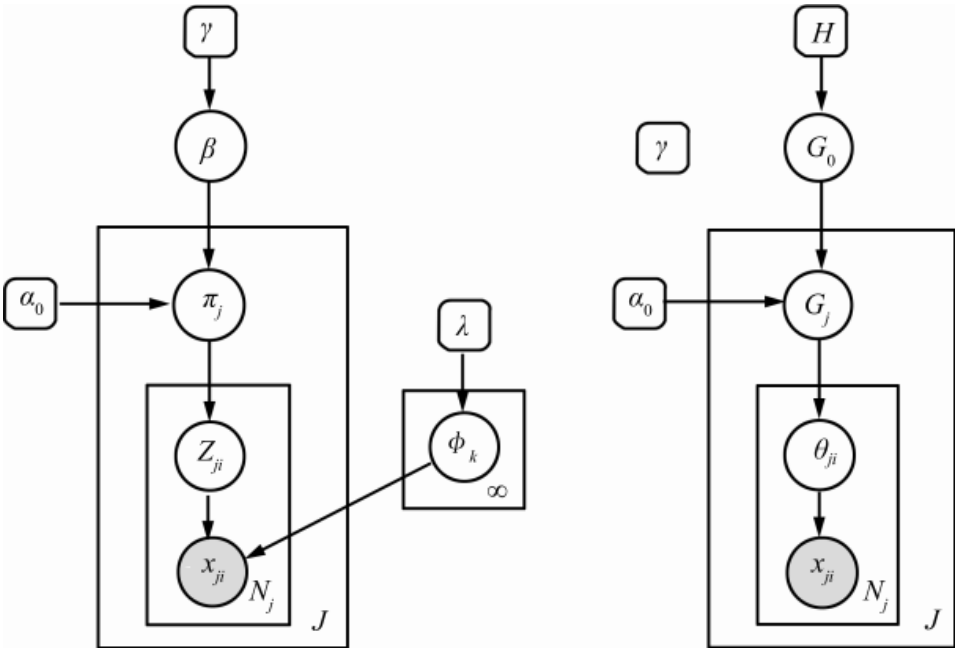

图 2.2　HDP 算法的有向图模型表示

Fig. 2.2　Directed graph of the hierarchical dirichlet process

2.3 Word2vec 词向量模型

Word2vec 和向量空间模型（BOW 经典模型）都是词向量[41]方法，也就是将一个单词、词组、句子、文章等表示成一个向量，这样通过向量之间的运算就可以计算词与词、句与句、文章与文章之间的关系。

传统的向量空间模型对词语进行向量化的一个问题就是向量空间模型由于没有考虑到词与词的上下文关系，因此无法捕捉词与词之间的相似度、也不能够表达语义内容，比如中国队战胜了日本队、日本队战胜了中国队，这两个句子词语完全一致、基于向量空间模型表示成的向量亦完全相同，但是句子本身却表达了完全相反的意思；另外，向量空间模型容易形成维灾难，单词数目越多词向量的纬度越大，当词语过多时算法效率会大大降低，尤其是在 Deep Learning 相关的一些应用中。

Word2vec 是 2013 年 Tomas Mikolov 等人提出的一种将词转换成向量形式的工具，主要是用于深度学习（Deep Learning）的工具，与向量空间模型相比具有良好的语义特征。word2ve 为计算向量词提供了一种有效的连续词袋(continuous bag-of-words)和 skip-gram[42]架构实现。利用 word2vec 对语句进行训练，可以把每个词语简化为 K 维向量空间中的一个向量，与 VSM 相比，word2vec 训练的词向量空间上的相似度可以用来表示文本语义上的相似度，而且 word2vec 能够捕获很多常见的语言规律，例如向量（“山东”）减去向量（“济南”）再加上向量（“湖北”）得到的向量结果和武汉的词向量非常相近；向量（“王子”）减去向量（“帅哥”）加上向量（“美女”）得到的词向量结果和公主的词向量结果非常相近。因此，很多数据挖掘、文本挖掘、自然语言处理等相关的工作都使用通过该模型训练得到的词向量，比如寻找反义词、主题聚类、词性分析等等。

Word2vec 有两种模型训练方法，CBOW 和 Skip_gram（如图 2.3 所示），由图中可以看出，CBOW 和 Skip_gram 两种模型都是神经网络中的一种形式，都包含输入、输出、以及映射层，只不过 CBOW 模型是利用上下文的词语来预测当前词语，而 Skip_gram 是利用当前词语来预测上下文词语，这是两种截然相反的思维。另外，对于这两种方法 word2vec 分别提供了两种提高词向量的训练效率的方法 HS（Hierachy Softmax）、NS（Negative Sampling），也就是说 word2vec 总共有四种模型训练方式。

Word2vec 训练过程中有几个参数是需要调节的，主要调节的是选择 CBOW 还是 Skip_gram 方法以及训练模式选择 HS 或者 NS 方法；需要确定词向量最终输出的纬度，一般选择 200 维左右；还有一个参数是 min-count 即词语出现的最小阈值，它的意思是词频低于该阈值的词将不在训练范围内；对于窗口大小这个参数也就是考虑的上下文窗

口的大小，就是考虑当前词的前后 N 各词，亦可使用随机窗口大小,另外学习率一般默认即可。

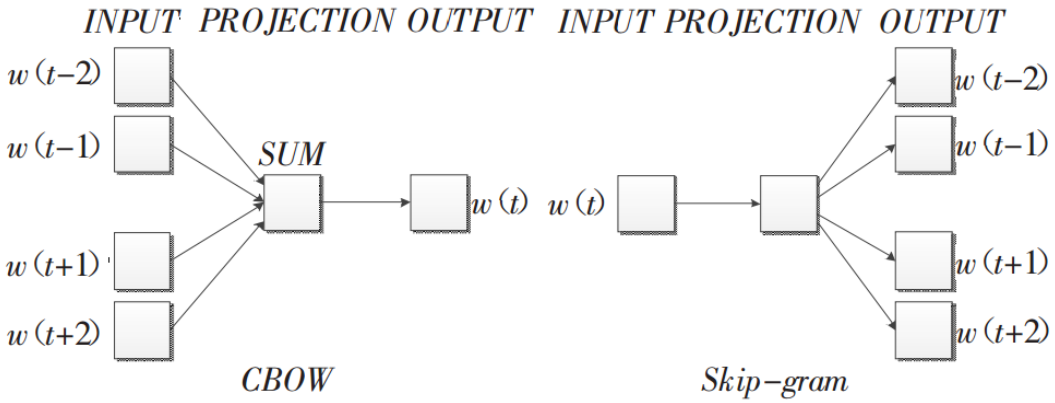

图 2.3　CBOW 和 Skip-gram 模型

Fig. 2.3　CBOW and Skip-gram model

2.4　SVM 分类方法

支持向量机（Support Vector Machine, SVM）是机器学习领域中重要的学习算法，属于有监督学习方式，通常用来进行样本分类。SVM 主要的思想就是最大化分类间隔，即寻找到的分类超平面能够将不同的类别分开，且此超平面附近的样本点同样有足够大的置信度将它们分开。

针对线性可分情况，一般会有很多分类直线可以将正负例分开，但只有中间红色直线将正负例分的彻底，即对正负实例中距离红色直线最近的点也能很好地区分开来。对于线性不可分的问题，一般使用映射算法将低维空间中不能线性区分的样本转变为高维空间中的线性可分问题，进而可以使用线性算法在高维特征空间解决非线性可分问题，此技巧称为核技巧（核函数）。

SVM 分类问题分为线性可分（硬间隔）、线性支持（软间隔）与非线性支持向量。当数据线性可分时，是三种情况中最简单的，可通过硬间隔最大化，学习一个线性模型，即线性可分支持向量机，此时也叫硬间隔支持向量机；当数据近似线性可分时，通过软间隔最大化，同样学习一个线性模型，称为线性支持向量机，也可叫作软间隔支持向量机；当数据线性不可分时，借助核技巧与软间隔最大化，学习非线性支持向量机。

SVM 进行分类的基本思想为分类间隔最大化，此处间隔指几何间隔最大。理解几何间隔首先需要了解函数间隔，一般地，一个点离超平面的距离一定程度上可以反映预测结果的可靠程度，在超平面 $w\cdot x+b=0$ 确定的情况下，$|w\cdot x+b|$能够相对地表示相对来说

可反应点 x 到超平面的距离远近，而 $w \cdot x+b$ 的符号与样本类别 y 的符号是否一致能够表示分类正确与否，因此可用 $y(w \cdot x+b)$表示衡量分类的正确性与可靠程度，即函数间隔。

即使函数间隔能够反应预测结果的正确性与可靠程度，在选择分类超平面时，只考虑函数间隔是不足够的，因为当 w 和 b 成比例改变时，即使函数间隔改变（与 w 和 b 的变化比例保持一致），但超平面并没有发生变化。这一事实告诉我们，需要限制分类超平面的法向量 w，比如规范化，即使得$||w||=1$，使函数间隔固定，此时函数间隔变为集合间隔。

对于线性分类问题，线性分类支持向量机是一种很有效的模型，但现实问题中，有很多问题并不是线性可分的，因此需要使用非线性支持向量机，而其主要的技巧是核函数。线性不可分时 SVM 采用核函数解决，即核函数可将部分线性不可分问题转化为线性可分的，将低维输入空间向高维空间转化，之后构建最优分类超平面。在 SVM 中常用的核函数主要有：多项式核、径向基核和 sigmoid 核。

SVM 算法起初是为二分类问题设计，为解决多分类问题，需要构造恰当的多类分类器。当前，构造多分类 SVM 可分为两类：一是直接法，直接修改目标函数，将多个分类器的参数合并到一个约束优化问题中，求解该约束优化问题实现多类分类。直接法看上去简单，但其计算量较大，实现比较困难，只适合应用于小型问题中；另一类是间接法，一般通过组合多个二类分类器解决多分类问题，常用方法有一对一（one-versus-one）和一对其余（one-versus-rest），一对其余一般也称之为一对多。

3 基于 LDA 的期刊推荐方法的研究

3.1 问题引出

随着硕士生、博士生、导师等数量的不断增长，每年都会发表大量论文。图 3.1 为 CNKI 中每年发表论文数量的统计数据，由图 3.1 可以看出从 1994 年开始每年发表论文数量增长的幅度都比较大，从 1996 年起每年都会发表 100 万篇论文以上。每年发表这么多论文，一般都需要发表到相应的期刊、杂志、会议等载体上。截止到 2014 年，全国总共有 8048 本期刊，总共发行了 1215544 期，总共 40237231 篇论文。那么，学者们如何从八千多期刊中选择最适合自己的期刊呢？学者们大都关注与论文本身的内容，却很少有人关注如何利用计算机帮助学者选择期刊来投稿的问题。通常投稿一般存在如下问题：论文写得很好，论文与期刊的研究方向不符的，论文被拒再重新投稿，浪费时间浪费精力；论文质量不好，却投了一个质量很好的期刊，导致拒稿。从上述问题可以看出，投稿时选择论文的方向和期刊方向相似的并且论文写作水平和期刊影响力水平相当的非常重要。

目前与期刊推荐研究直接相关的工作比较少，王超等人[43]对国内外学术信息推荐方法研究进展进行综述，指出对学术信息进行推荐主要集中在论文推荐和研究人员推荐：徐键等人[44]利用 PageRank 算法结合内容推荐和协同过滤推荐等方法有效地提高了推荐系统的命中率；倪卫杰等人[45]针对用户兴趣建立个性化的论文推荐系统；杜永萍等人[46]针对文献推荐问题，提出基于主题效能的学术文献推荐算法，实验结果可满足用户对个性化和文献质量两方面的需求。邓少伟等人[47]为使用户能够准确、高效地查找出关联的科研人员、学科知识及研究领域等信息，提出一种基于论文共同作者学术关系的推荐系统，实验表明与普通方法相比推荐系统的精准度能提高 5%左右。陆艳春[48]针对目前会议推荐的缺点提出了基于引用关系的学术会议推荐系统提高了学术会议推荐的准确性，从而提高学术资源信息的利用率，为学术研究人员提供更多，更有价值的数据服务。由于现有直接相关研究较少，本文尝试借助于通用的推荐方法进行期刊推荐，Adomavicius G 等人[49]指出推荐系统主要分为基于邻域的推荐（基于用户的推荐、基于物品的推荐）和基于内容的推荐，本文借鉴了这三种方法进行期刊推荐的研究。

学者选择期刊主要有四大因素，分别是学者的个人偏好（如某些学者经常投稿给某期刊）、期刊偏好信息（主要指期刊的研究方向）、学者论文写作水平（论文创新性）、期刊质量的高低（主要是指期刊的影响因子的高低）。本文主要目的就是利用上述四大因素帮助学者选择合适的期刊（简称期刊推荐），主要根据论文标题、关键字、摘要以

及作者以往发表期刊的历史信息来推荐某几个期刊中供作者选择，并对论文写作水平作为外部因素供作者选择，让作者选择其中最适合自己论文的期刊来进行投稿。

根据投稿时选择期刊的各种因素，主要有五种期刊推荐方法：前三种算法是按照上述的推荐领域基本算法，分别是基于内容的期刊推荐方法、基于用户的期刊推荐方法和基于期刊相似度的推荐方法。第四种是按照分类的思想，将期刊看作类别，论文看作待分类的原始数据；第五种是按照主题模型思想，每个期刊发表的论文都有一定的主题相关性，只要论文的主题与该期刊的这些主题相似即可推荐到该期刊上；投稿人以推荐的若干期刊作为参考，从中选择与论文质量、审稿速度、期刊类型等与自己最相符的期刊，这对于准备投稿的作者有一定的借鉴意义。另外，鉴于利用计算机自动评价论文质量的高低比较困难，因此本文方法并不直接对论文质量进行自动化评价，而是将论文质量高低与衡量期刊质量的期刊综合影响因子进行对应，质量高论文自动的推荐到期刊质量好的、综合影响因子大的期刊中，这样对投稿人选择更加准确，通过加入论文质量这一因素发现上述五种方法的准确率都有大幅提高，这说明根据论文的写作水平推荐相应的期刊对期刊推荐有较大作用。本文上述五种方法都进行了尝试，发现基于 LDA 的主题模型对于期刊推荐来说效果最好，准确率最高；分类模型比基于 LDA 的主题模型效果稍差些；基于用户的推荐效果比基于分类的期刊推荐稍微差些；基于内容的推荐效果最差，准确率最低。投稿人选择 TOP5 的期刊作为参考，从中选择与论文写作水平、审稿速度、期刊类型等与自己最相符的期刊，这对于准备投期刊的作者有一定的借鉴意义。

本文对影响论文水平的因素进行了简单的探讨，并将论文水平与欲推荐的期刊质量的高低挂钩，论文写作水平高的论文推荐到相应期刊质量好的论文中，这样对投稿人选择更加准确，通过加入论文写作水平这一因素发现上述五种方法的准确率都有大幅提高，这说明根据论文的写作水平推荐相应的期刊对期刊推荐有较大作用。

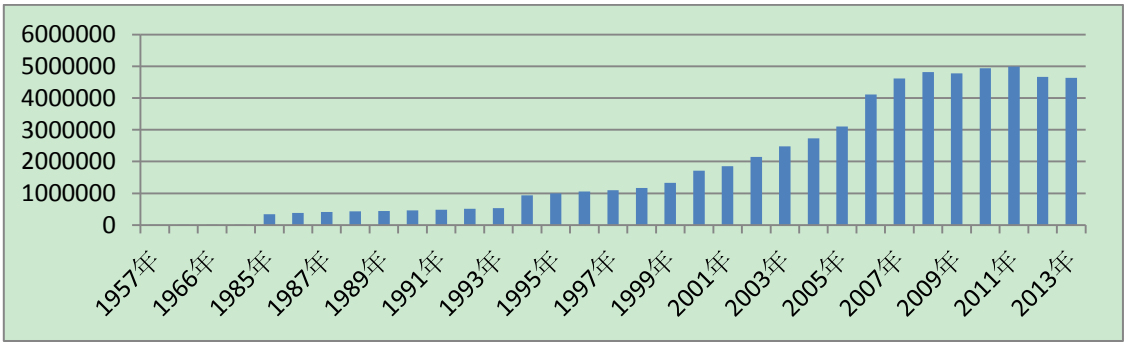

图 3.1　CNKI 中统计的 1957 年到 2013 年论文发表数量

Fig.3.1　The number of published paper between 1957 and 2013 in CNKI

3.2 期刊推荐方法的研究

到目前为止对期刊推荐方法的研究非常少，因此本文只能按照研究人员投递论文时选择期刊的思想来进行分析，主要分为基于分类的期刊推荐方法、基于 LDA 的期刊推荐方法、基于内容的期刊推荐方法、基于用户的协同过滤的期刊推荐方法、基于期刊相似度的推荐方法分别介绍具体方法，下面将详细介绍其算法思想。

3.2.1 基于分类的期刊推荐方法

从分类的角度看，期刊就是相应的类别，而论文就是待分类的数据，期刊推荐就是将准备发表的论文正确分到与之相符的期刊（类别）中。

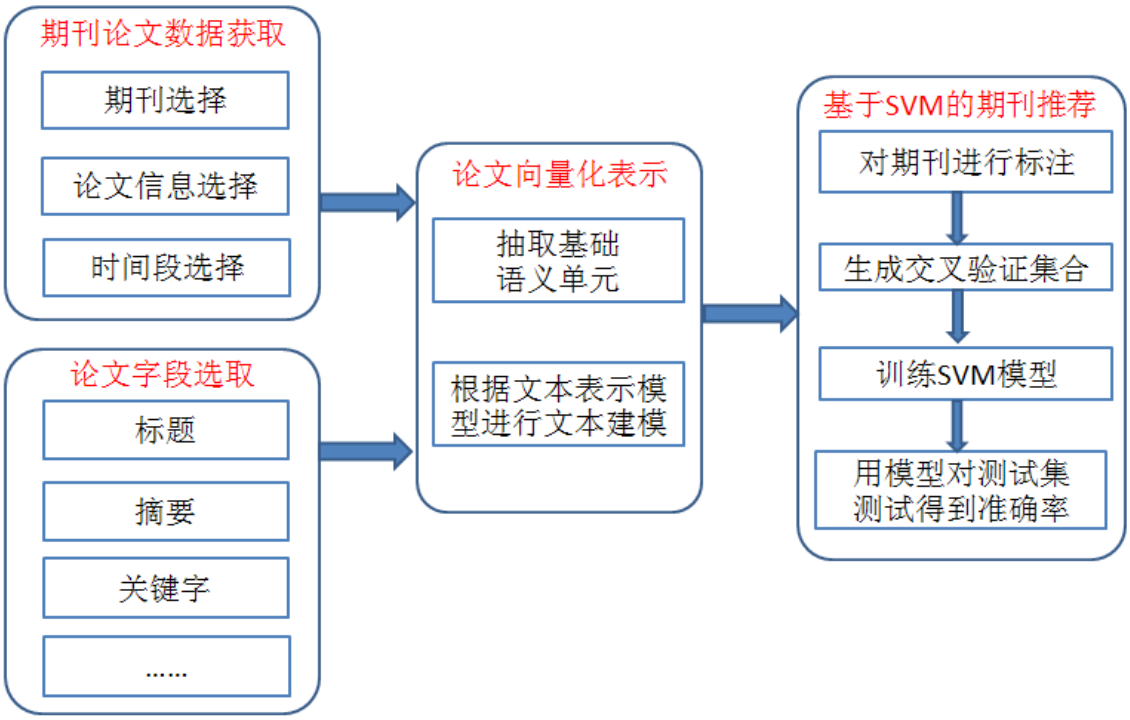

图 3.2 基于 SVM 的期刊推荐方法的步骤

Fig. 3.2 The method of SVM-based journal recommendation

分类问题又称为监督学习或者归纳学习，其学习思路类似于人类学习的方式，人类可以从过去的经验中获取知识以用于提高解决当前问题的能力。然而，由于计算机没有“经验”可用，计算机只能从自己所收集的过去的数据中获取知识，这些数据就代表了过去的经验。对于分类问题，现在研究得已经很成熟了，常见的分类方法有决策树、关联规则[50]、朴素贝叶斯[51]、支持向量机（SVM）等。支持向量机（SVM）算法的众多

优点使得它成为最流行的算法之一。它不仅有扎实的理论基础，而且在许多领域比大多数其他算法更准确，尤其在处理高维数据时。故本文采用支持向量机作为分类算法，本文使用的是台湾大学林智仁教授的 libsvm[52]工具包。

由图 3.2 可以看出，基于分类的期刊推荐方法的主要步骤如下：

第一步：期刊论文数据获取，这一部分是整篇论文的基础，而且由于论文数据属于商业信息，主要由 CNKI、VIP 网、万方等学术论文运营商保存，论文难以获取，这些论文数据是通过特制的网络爬虫费尽千辛万苦才爬取到的；

第二步：论文向量化表示（训练集和测试集都如此），主要将爬取的数据抽取出所需要的标题、摘要、关键字、作者信息等并表示成向量；

第三步：具体的支持向量机的分类方法，首先对期刊进行标注，SVM 算法要求的格式是“Lable 1:value1 2:value2 3:value3……”这种格式，因为论文信息已经利用 word2vec 表示成向量了，所以只需要将论文对应的向量前边加入自己的期刊所对应的 Lable 即可，其实 Lable 标签就是类别标签，可以对所有的期刊按照字母顺序从 1 开始编号，然后编号就是类别标志；

第四步：需要对整个数据集生成十倍交叉验证的训练集和测试集，生成十倍交叉验证的目的就是防止数据分布不均导致出现偶然结果的情况；

第五步：对训练集利用 libsvm 进行训练即可；

第六步：利用生成的模型对测试集进行测试，并计算得到准确率。

3.2.2 基于主题的期刊推荐方法

从主题的角度来看，每个期刊都有不同的研究方向，不同的期刊研究方向有可能相同，从而不同的期刊的研究内容同属于一个主题，首先利用 LDA 模型对期刊进行主题聚类，生成不同的聚类，同一聚类下的期刊研究内容相似却又不完全相同，同一聚类生成一个 SVM 分类模型，SVM 训练中的类别就是期刊编号。当新的文章来到时首先利用 LDA 主题聚类判断文献所属的主题类别，然后利用相应主题类别的 SVM 分类模型进行分类，SVM 分类概率最大的前 5 个就是本文推荐的期刊。

由图 3.3 可以看出，基于 LDA 的期刊推荐方法的具体步骤的前两步与基于 SVM 的期刊推荐方法完全一致，下面主要讲解第三部分，也就是具体的 LDA 模型和 SVM 集合的期刊推荐方法，基于 LDA 的期刊推荐方法主要分为四步：

第一步：对所有期刊进行 LDA 聚类，聚类个数利用困惑度计算获得最优聚类个数，确定每个期刊所属的聚类；

第二步：对同一个聚类内的期刊利用 SVM 分类模型进行训练，每个期刊是 SVM 训练类别中的一个类，最终第一步生成多少个聚类就生成多少个 SVM 分类方法；

第三步：对测试集中的论文信息利用 LDA 主题模型进行预测其主题类别；

第四步：利用对应主题的 SVM 分类方法对该论文预测其具体所属的期刊类别；

第五步：对预测结果按照从大到小排序得到最相近的几个期刊并计算整体的准确率。

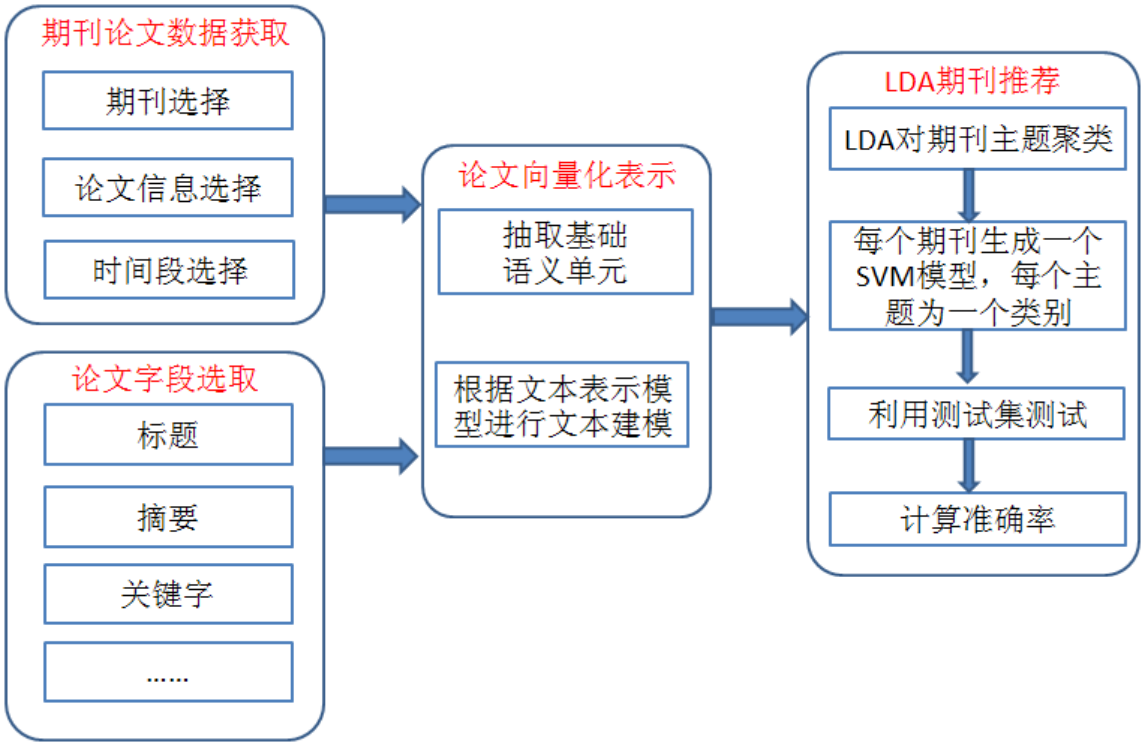

图 3.3　基于 LDA 的期刊推荐方法的步骤

Fig .3.3　The method of LDA-based journal recommendation

3.2.3　基于内容的期刊选择方法

基于内容的推荐是推荐系统中常用的做法，这种做法对于每个 item 基于其自身属性来表示这个 item 内容，从而推荐那些和当前 item 含有相同或者相近特征的 item。这种推荐的优点是算法容易实现、不需要用户关系数据、物品关系数据等，因此不存在协同过滤推荐算法的稀疏矩阵和冷启动的问题。基于 item 本身特征推荐不存在过渡推荐热门问题，所涉及的技术也是搜索引擎中应用比较成熟的技术。

基于内容的期刊推荐方法需要对内容进行向量化，经典的方法是向量空间模型（VSM），向量化之后期刊本身是一个向量表示，文献也是一个向量表示，通过计算期刊和文献的向量之间的相似度来判断该文献的内容与期刊整体内容相似的程度。

由图 3.4 可以看出，基于内容的期刊推荐方法的步骤如下：

第一步：将所有的论文的标题、摘要、关键字这些信息都整合到一个文本中，每篇论文占据一行；

第二步：利用中科院分词器[53]进行分词；

第三步：去除停用词、标点等无用信息；

第四步：对所有的分词之后的信息从 0 开始编号，从而对每个文献进行向量化；

第五步：期刊是期刊内各篇论文的集合，期刊向量化最简单的做法是所有文献向量相加再求平均值；

第六步：计算测试集中文献和各个期刊的向量的相似度，推荐最相似的几个期刊；

第七步：十倍交叉验证计算平均准确率。

另外，本文同样利用 word2vec 进行向量化与向量空间模型的算法进行对比，其算法基本步骤与向量空间模型的非常相似，只是对文献向量化的方法不一样。

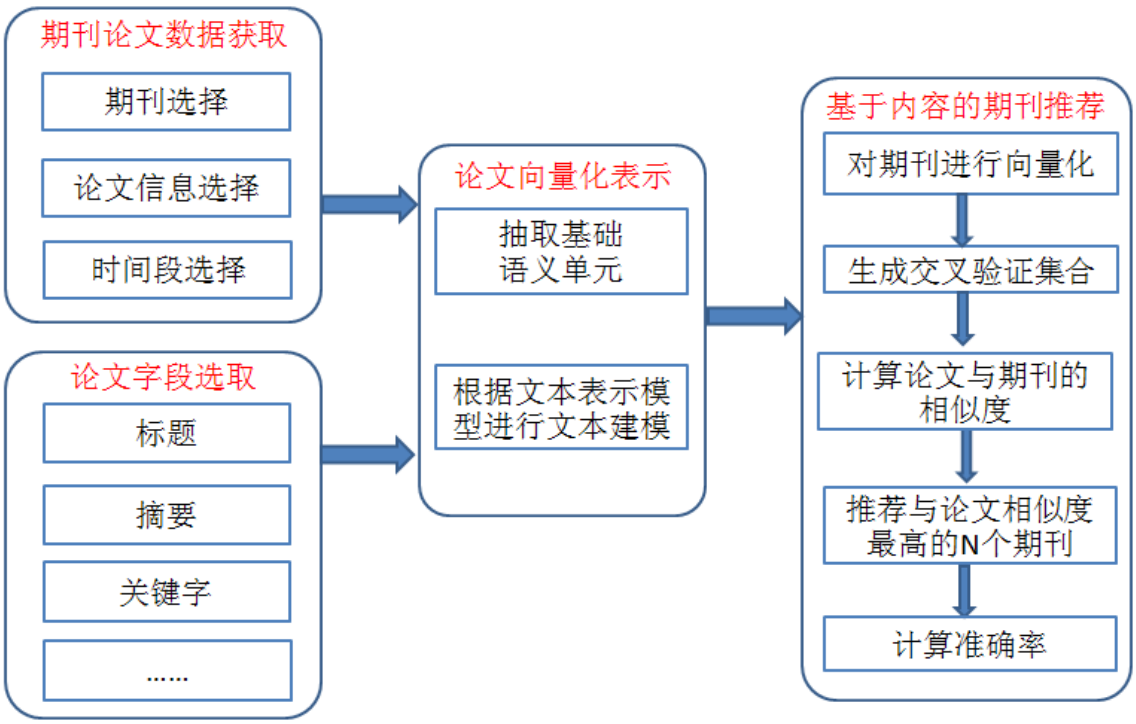

图 3.4　基于内容的期刊推荐方法的步骤

Fig. 3.4　The method of content-based journal recommendation

3.2.4 基于用户的协同过滤推荐方法

基于用户的期刊推荐方法是寻找自己及其合作者发表的论文所在的期刊的集合，将这些期刊推荐给该用户，如果该用户没有发表过论文而且合作者也没发表过论文，那么就按照基于内容的期刊推荐方法。

对于用户来说，用户对某些期刊有特殊的偏好，有些人 10 篇论文中有 8 篇是发表到 A 期刊中，另外 2 篇各有一篇发表到 B 期刊中，这说明这个人下一次发表文章更倾向于发表到 A 期刊。因此每个用户在每个期刊上发表论文的数量也是一个影响因素，本文对在基于用户协同过滤推荐的基础上对每个用户在每个期刊上发表的数量进行加权，发表文章数目多的权重大，发表文章数目少的权重小。其具体公式为公式（3.1）。

$$\text{Lable}_i = \begin{cases} \text{argmax}_{c_j \in c_{user}} Simlarity_{i,c_j} * tf_{user,c_j} & c_{\text{user}} \neq \emptyset \\ \text{argmax}_{c_j \in c} Simlarity_{i,c_j} & c_{user} = \emptyset \end{cases} \tag{3.1}$$

其中 Lable_i 是指编号为 i 的文档所属的期刊，c_j 是指期刊编号，c 是所有期刊的集合，c_{user} 是用户 user 及其合作者发表过的文章所在的期刊集合。

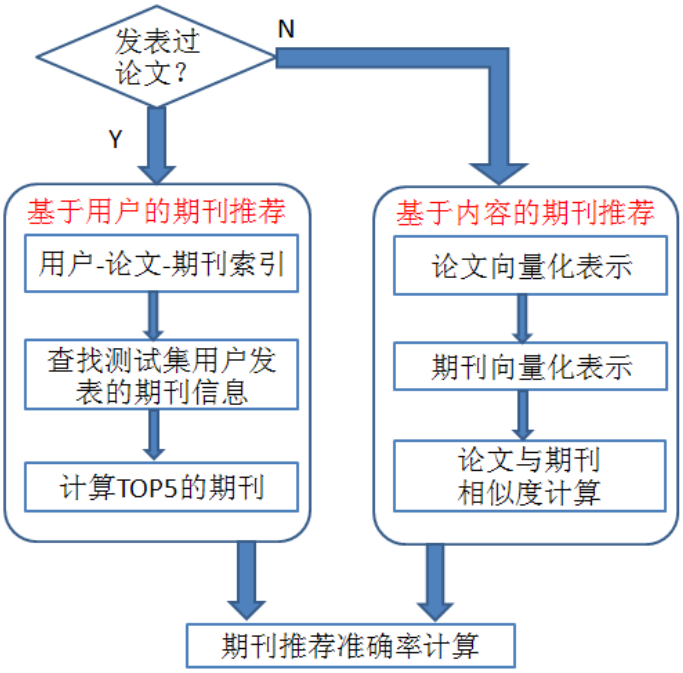

图 3.5 基于用户的期刊推荐方法的步骤

Fig. 3.5 The method of user-based journal recommendation

由图 3.5 可以看出，基于用户的协同过滤推荐的步骤如下：

第一步：对训练集生成用户-论文-期刊倒排索引表，方便查找用户发表过的论文以及该论文发表到哪个期刊上；

第二步：对测试集论文进行推荐，首先查找该论文的所有作者发表过的论文信息以及对应的期刊信息；

第三步：按照公式（3.1）计算得到可能性最大的 5 个期刊并计算推荐的准确率。

3.2.5　基于期刊相似度的推荐方法

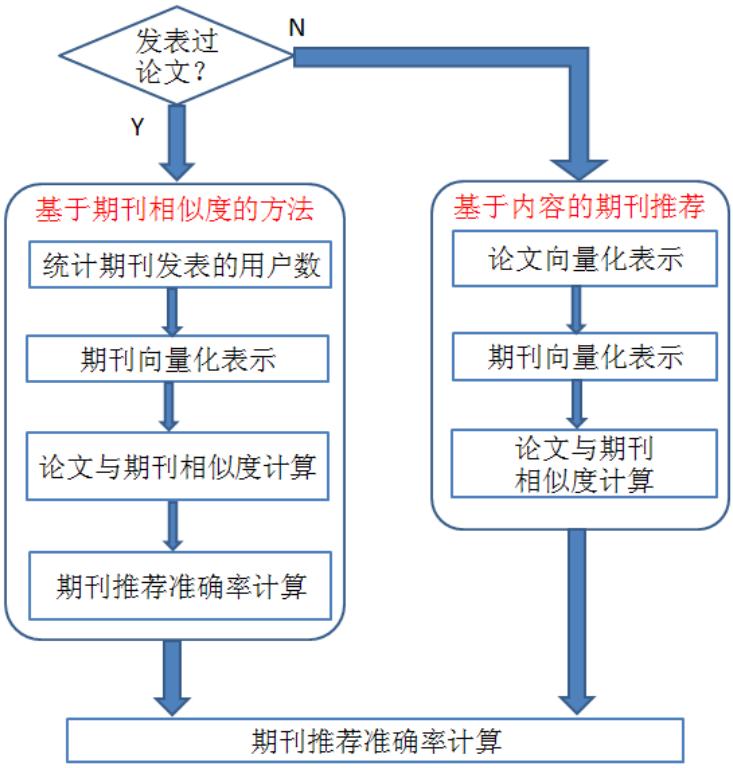

图 3.5　基于期刊相似度的推荐方法的步骤

Fig. 3.5　The method of journal-similarity recommendation

基于期刊相似度的期刊推荐方法的思想是：发表到该期刊的人同时也发表到另外一个期刊，用来计算两个期刊的相似度，推荐一个期刊时同时将与该期刊相似度高的期刊也推荐给用户。其物品相似度计算方法为公式（3.2）所示。

在进行期刊推荐时首先找到该用户及该用户的合作者发表过的期刊集合，然后加入与集合中期刊相似度比较高的期刊组成新的期刊集合，计算这些期刊与待发表论文的相似度，然后按照相似度进行排序，给用户推荐相似度最高的 5 个供用户进行选择，其计算公式与基于用户的期刊推荐公式（3.1）一致。

$$Simlarity_{i,j} = \frac{|N_i \cap N_j|}{\sqrt{|N_i||N_j|}} \tag{3.2}$$

$Simlarity_{i,j}$ 是指期刊 i 与期刊 j 的相似度，N_i 是指发表在期刊 i 上的用户数目，N_j 是指发表在期刊 j 上的用户数目，$N_i \cap N_j$ 是指同时发表到期刊 i 和 j 上的用户数目。

基于期刊相似度的推荐方法的主要步骤如下：

第一步：计算发表到每个期刊的用户数；

第二步：对期刊两两组合，分别计算同时发表到两个期刊的共同用户数；

第三步：利用公式（3.2）计算期刊的两两相似度；

第四步：对论文进行推荐，首先寻找到该论文的作者以前发表过的期刊然后寻找与这些期刊最相近的那些期刊进行推荐，选择最相似的 5 个期刊进行推荐，并计算推荐准确度。

3.2.6 影响论文水平高低的因素

论文的写作水平影响发表到期刊的水平，论文写作水平高的论文一般都发表到期刊影响力较大的期刊中，论文写作水平较差的文章发表到高质量期刊的可能性较小，一般来说写作水平与论文发表的期刊的水平是对应的。

由于影响论文的写作水平的因素非常多，例如：

方法的创新性：影响论文水平的最重要的因素，方法的创新性一般是通过相同研究方向的专家才能确定，该影响因素很难通过计算机自动评价；

实验的丰富性：影响论文水平重要的因素之一，实验的丰富性关乎论文方法的可靠性，实验越丰富说明该方法越可靠而不是偶然出现的结果；

参考文献的水平高低：最近几年参考文献数量比较多说明论文作者对该研究方向最新的研究状况比较了解，了解最新的研究成果。另外参考文献的水平能够反映论文作者的水平，如果引用的参考文献都是水平很差的期刊的论文，那么说明该作者并没有站在巨人的肩膀上。

作者的身份（硕士、博士）：很容易理解作者身份也是影响论文水平的一个很重要的因素之一，一般来说硕士的写作水平比博士差很多，据统计计算机三大权威学报（软件学报、计算机学报、计算机研究与发展）的作者很少出现硕士的情况，一般都是博士。

学校的科研水平：学校水平有很大的差距，985 的高校比普通高校的科研水平还是高许多的，很多普通高校都没有硕士点、博士点，这样科研能力一般都比较弱。

是否有国家级的基金的支持：根据文献[18]得出的结论，不论是论文的被引频次还是论文的被引率有基金支持的论文明显高于没有基金支持的论文，而论文的被引频次、论文被引率高低则反映了论文的水平的高低。

这些只是影响论文水平的一部分因素，还有很多因素都没有考虑到，而这些因素中有很多部分是不能利用计算机进行评价的，因此论文水平自动评价有非常大的困难，为此，本文将论文水平作为一个外部因素进行控制，由使用者进行控制，论文作者使用该方法进行期刊推荐时，可以手动选择期刊水平的高低。为了验证论文水平高低对期刊推荐的作用，本文将期刊水平按照 CNKI 中提供的期刊影响因子来分为高中低三档，利用测试集测试时，对高中低三档的期刊数量进行控制，例如对计算机学报这一期刊的论文进行推荐时自动推荐期刊水平属于高档期刊的那些期刊。

3.3 实验结果与分析

表 3.1 期刊名称及其综合影响因子

Tab. 3.1 Journal and integrated impact factor

期刊名称	期刊影响因子	期刊名称	期刊影响因子
电子学报	1.134	计算机应用研究	0.549
通信学报	0.917	信息安全与技术	0.115
软件学报	1.682	微电子学与计算机	0.335
计算机工程	0.421	智能计算机与应用	0.170
计算机应用	0.670	计算机科学与探索	0.414
计算机学报	1.886	计算机研究与发展	0.869
控制与决策	0.875	计算机应用与软件	0.352
电子测量技术	0.803	计算机与应用化学	0.294
中文信息学报	0.773	计算机工程与科学	0.334
微计算机信息	0.562	计算机工程与应用	0.456
微计算机应用	0.200	计算机工程与设计	0.452
微型机与应用	0.230	计算机集成制造系统	0.977
系统仿真学报	0.419	模式识别与人工智能	0.620
智能系统学报	0.564	小型微型计算机系统	0.342
电脑开发与应用	0.139	复杂系统与复杂性科学	0.495
计算机系统应用	0.296	计算机辅助设计与图形学学报	0.628

3.3.1 语料来源及预处理

本文语料来自 CNKI，本文将 1990 年开始到 2013 年 12 月之前所有的论文的标题、作者、机构、年份、关键字、摘要这些信息，32 个计算机相关的期刊，其涉及的方向有，电脑开发、电子、图形学、应用、控制与决策、模式识别、信息安全、软件、通信、微机、系统仿真、中文信息处理、人工智能、信息检索、文本挖掘等，具体内容如表 3.1，表中期刊所对应的影响因子是从 CNKI 中获取得到的。语料总共有 164017 名作者信息，217972 篇文章。

本文首先利用中科院分词器[53]对语料进行分词处理，然后利用 word2vec 对语料进行向量化。

3.3.2 实验结果以及分析

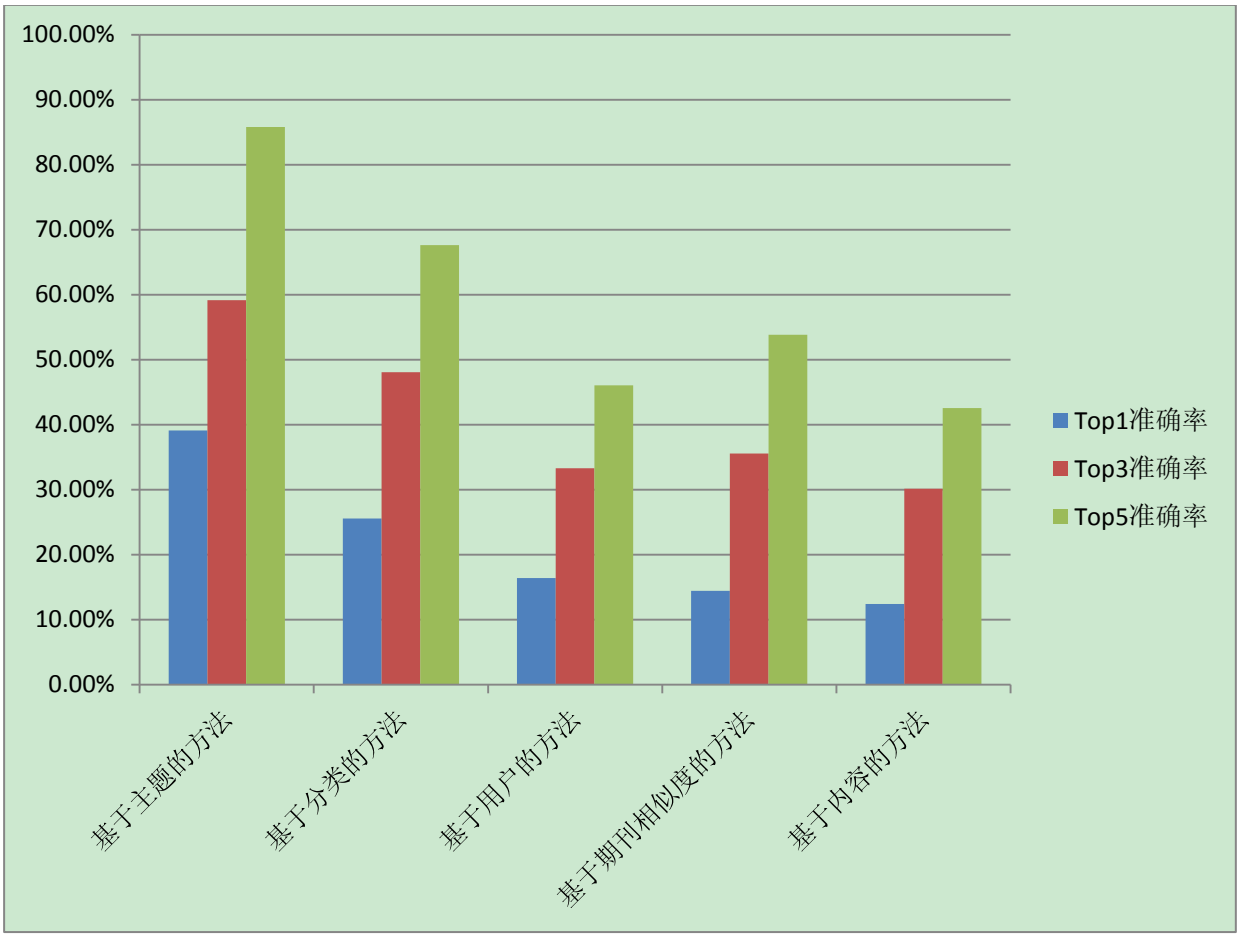

图 3.6 各种模型的结果比较

Fig. 3.6 The result of the various models

投稿是一个极其复杂的过程，投稿人存在很多考虑，它牵扯到论文的写作水平、期刊的影响因子、期刊的类别、期刊的录用率、期刊的审稿周期长短等各种因素，因此只给投稿人推荐一个期刊是不足以满足投稿人的需求的。因此我们的评价指标不再是TOP1 的准确率、召回率和 F 值，而是 TOP1、TOP3、TOP5 的准确率。TOP5 的准确率比较高，投稿人将推荐的这 5 个期刊作为参考，可以再详细根据自己论文和期刊的综合信息来加以考虑具体选择哪个期刊。图 3.6 是各模型结果对比。

由图 3.6 可以看出，对于期刊推荐来说，LDA 主题模型效果最佳，SVM 分类模型效果稍差一些，基于期刊相似度的推荐方法、基于用户的期刊推荐方法和基于内容的推荐模型结果很相近。

基于 LDA 的期刊推荐算法之所以比基于 SVM 的期刊推荐算法好是因为而基于LDA 期刊推荐方法对期刊又进行更为细致的主题划分，首先确定论文所属主题自然就缩小了分类的范围推荐结果就好了很多，该方法吸收了主题模型和 SVM 分类方法的优点，效果更好。

基于用户的期刊推荐方法效果较差的原因是因为数据过于稀疏，每个作者发表的论文数目不是很多，主要是因为大部分硕士生在研究生期间只能发表 1-2 篇论文，如果以后不再从事科研行业，那么可能这一生也就这一两篇论文了，而且研究生的数量大大多于博士生和老师的数量，虽然老师发表的论文数目稍多，但是整体平均情况下还是非常少的。

基于内容的推荐方法中基于向量空间模型的推荐比基于 word2vec 的推荐效果要差，所以在实验结果中只展示了最好的效果的，主要原因是向量空间模型没有考虑到词与词之间的相关关系以及上下文关系，其 Cosine 相似度并不能真正表示句子语义上的相似度，而 word2vec 由于考虑到这些信息因此效果要好一些，但是单纯地利用相似度来进行推荐还是比其他方法要差一些。

如图 3.7 所示，各个期刊的准确率都各不相同，有的期刊的准确率比较高，TOP5 都能到达 78%，而有的期刊的准确率比较低，只能达到 15%甚至更低。

准确率比较低的期刊有软件学报、计算机应用研究、计算机学报、小型微型计算机系统、计算机工程、计算机科学与探索等，这些期刊的 TOP1 的准确率只有百分之几，TOP10 的准确率都不足百分之五十，更有甚者不到百分之三十。通过分析我们发现，这些期刊有两个特点：一、期刊的研究方向太过于分散；二、期刊数量很少。

例如：计算机科学与探索的研究方向有高性能计算机、体系结构、并行处理、计算机科学新理论、算法设计与分析、人工智能与模式识别、系统软件，软件工程、数据库、计算机网络、信息安全等十多个研究方向。其研究方向太多，导致内容繁杂。另外由于

该期刊由 2007 年才开始创建，每年出版的文章数目也只有一百篇左右，数据比较稀疏。因此，不论是利用相似度还是利用基于用户的期刊推荐由于其数据稀疏性的原因最终导致整体期刊推荐的准确率比较低。小型微型计算机系统刊登文章的内容涵盖计算技术的各个领域(计算数学除外)。包括计算机科学理论、体系结构、计算机软件、数据库、网络与通讯、人工智能等各方面的学术论文，由于研究涉猎的范围非常广泛只要和计算机相关的论文基本上都可以投递，因此对于这种类型的期刊准确率会比较低。其他计算机应用研究、计算机学报、软件学报、计算机工程等研究方向也都非常广泛，最终其期刊推荐的准确率都比较低。

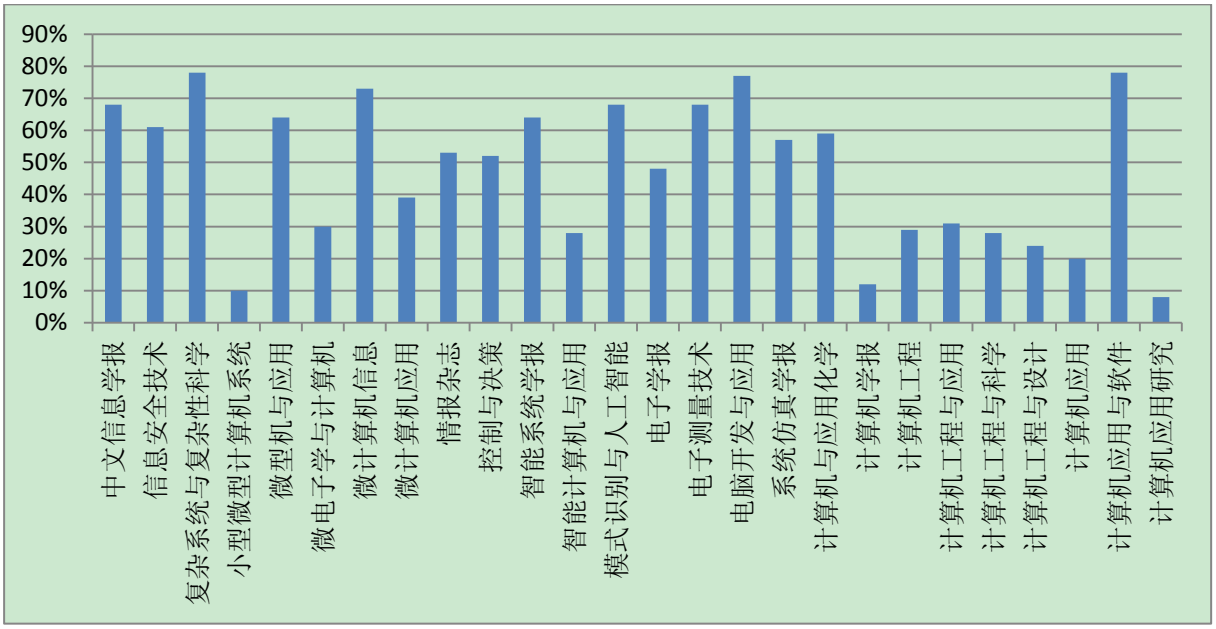

图 3.7　基于内容的推荐 TOP5 准确率

Fig. 3.7　Content-based journal recommendation TOP5 accuracy

准确率比较高的期刊都是大都是专业性比较强的期刊，例如：计算机辅助设计与图形学报和中文信息学报。计算机辅助设计与图形学报在国内一直是计算机图形学领域的权威期刊，主要研究图形与可视化、图像与视觉、虚拟现实与交互技术、数字化设计与制造、VLSI 设计与测试及电子设计自动化等，这些研究方向都与图形学和计算机辅助设计息息相关，其他研究方向的内容一律不涉猎。中文信息学报计算语言学、机器翻译、中文语音识别与合成、信息检索（IR）信息抽取（IE）及相关的语言技术、网上搜索引擎数据挖掘、知识获取、神经网络、人工智能（AI）技术等。这些方向都是相应领域非常专业性的研究方向，类似这种期刊的准确率一般都会比较高。

另外，我们都知道论文的写作水平影响发表到期刊的水平，论文写作水平高的论文一般都发表到期刊影响力较大的期刊中，论文写作水平较差的文章发表到高质量期刊的可能性较小，一般来说写作水平与论文发表的期刊的水平是对应的。

因此，本文对论文水平与期刊质量高低进行挂钩，本文按照表 3.1 所示的期刊的影响因子来划分期刊质量高低，分为高档期刊、中档期刊、低档期刊。为了更加充分说明论文质量对期刊质量的影响，本文默认高档期刊的论文为高水平论文，默认从高档期刊中进行期刊推荐；中档期刊的论文为中等水平的论文，默认从中档期刊中进行期刊推荐；低档期刊的论文为中等水平的论文，默认从低档期刊中进行期刊推荐。具体结果如图 3.8 所示。

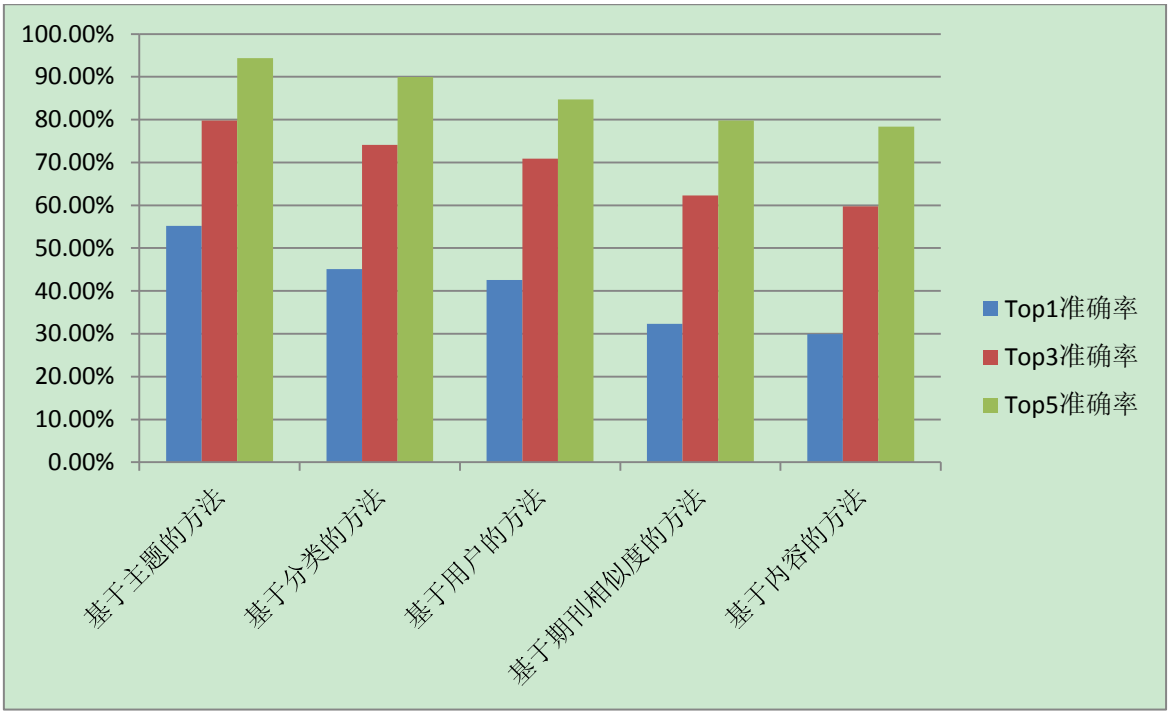

图 3.8　加入论文水平的期刊推荐方法各结果比较

Fig .3.8　The result of the various models added the factor of paper level

通过 3.6 和 3.8 的对比来看，加入论文水平这一因素明显提升了各个期刊推荐算法的准确率，这说明正确评价论文水平对期刊推荐有很大的帮助。

3.4　本章小结

本文深刻分析了投稿人投递期刊时的各种偏好因素，主要分为学者偏好因素、期刊偏好以及论文写作水平三个方面。然后将投稿问题按照不同的角度进行考虑，发现可以

将期刊投稿问题分为主题模型推荐问题、分类问题、基于内容的推荐、基于用户的推荐和基于期刊相似度的推荐方法这五类问题，通过实验分析可以得出，对于这几种考虑分类问题是一个很好的解决方案。按照 LDA 主题思想 TOP1 可以达到 35%的准确率、TOP5 可以达到 85%的准确率，投稿人最终可以从 TOP5 中选择最符合投稿人的期刊来进行投递，这对于投稿人都有一定的参考价值。另外，本文研究了论文水平和期刊推荐水平的关系，发现加入论文水平这一因素对期刊推荐的准确率的提升有很大的帮助，加入论文水平这一因素 TOP1 最高达到 48%，TOP5 最高准确率达到 90%以上，这说明正确地评价论文水平对期刊推荐有很大的帮助，论文所提出的方法具有一定的实用价值。

4 基于 HDP 的汽车专利主题演化研究

4.1 问题引出

近年来专利文献越来越受国内外各界的重视，专利数量也呈爆炸式增长，主要原因是专利固有商业技术情报价值以及其他特点诸如及时、可靠、内容详尽等。那么从浩如烟海的专利信息中准确的获取专利的各种主题演化信息对于各界人士都有很重要的作用。汽车产业是人们生产生活息息相关的产业，在国民经济以及社会发展中发挥着举足轻重的作用。基于以上特点本文以中国专利数据库中的汽车专利为出发点研究汽车产业中的主题演化信息，从而挖掘潜在有价值信息。

聚类作为一种有效的信息组织的方法，可以帮助人们从浩如烟海的专利数据中获得更多潜在的信息。在文本挖掘领域，研究人员已经生成各种主题挖掘方法，典型的代表是 Blei 等人提出的潜在狄利克雷分布(Latent Dirichlet Allocation，LDA)模型；Wang 等人提出的动态狄利克雷分布(Dynamic Latent Dirichlet Allocation，DLDA)；Teh 等人提出的分层的狄利克雷过程(Hierarchical Dirichlet Processes ,HDP)。

另一方面，将发现的主题信息以及各个主题之间的相关关系以友好的方式呈现给用户，研究人员也设计了很多主题可视化方面的算法，如 Havre 等人提出主题河(ThemeRiver)；另外 Wei 等人发表的论文 TIARA，以河流的形式详细地分析用户的邮件主题信息；Cui 等人提出的 TextFlow 也是以河流的形式细致、连贯地展示了主题随时间的各种变化。

在专利挖掘领域，范宇等人将 LDA 主题聚类应用到专利信息聚类中，该论文的缺点是根据经验事先假定了主题个数 K，而主题数目的确定对聚类准确度、精度都有一定的影响。郝智勇等人将 LDA 主题聚类与可视化相结合，展示了 4 年中各个主题之间的相关性聚类散点图，并没有考虑到各个主题之间的变化，亦没有展示各个主题的发展变化趋势等。

基于以上缺点，本文提出使用基于 HDP 算法的汽车专利主题演化方法，首先使用 HDP 算法解决了 LDA 算法的局限性，然后将专利主题演化信息以主题河的形式展示出来，以便用户以更简单的方式获得最有价值的信息。具体过程如下：首先通过 HDP 算法对每个时间段的专利数据进行主题挖掘，然后根据 Salton 等人提出的向量空间模型来计算每个主题与下一个时间段的每个主题的相似度，根据相似度来判断当前阶段各个主题和下一阶段各个主题之间的对应关系。本文通过当前阶段主题信息与添加历史数据

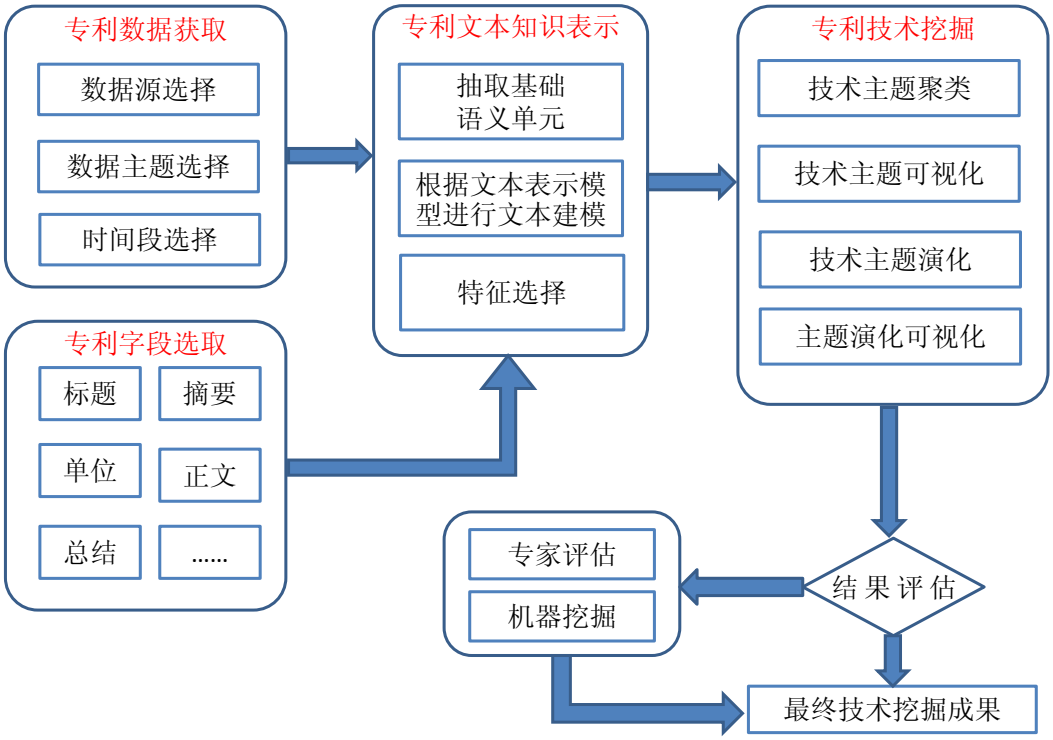

图 4.1　汽车专利挖掘主题演化研究思路

Fig. 4.1　The method of mining the topic evolution of automotive patent

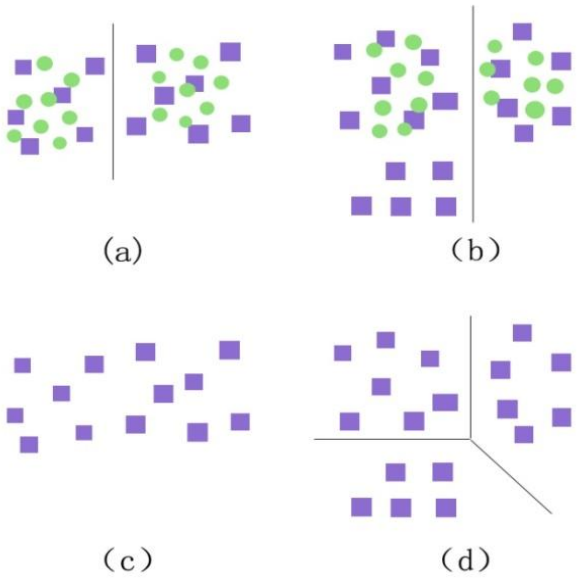

图 4.2　主题之间分流合流的例子

Fig. 4.2　An example of splitting and merging of clusters

的主题聚类信息之间的变化来发现各个主题之间的分流与合流等信息，然后利用三次样条曲线将各个主题组织成叠式图。实验结果显示，该算法能够发现各个主题之间的分流、合流关系以及主题的新生与主题的消亡等主题关系，本文将专利主题聚类与主题演化以及可视化结合在一起，能够直观地展示出专利主题之间的关系。图 4.1 是本文的研究思路。

4.2 基于分层的狄利克雷过程的主题演化

发现各个主题之间的主题演化关系（如主题产生、消亡，主题分流、合流），需要两个步骤，一是如何确定主题之间的如分流、合流等主题演化方式也称为主题抽取过程，二是如何将各个主题相关关系用主题河的方式呈现给用户也称为可视化过程。

4.2.1 主题抽取

主题是一系列的文档的组合，在主题抽取步骤，不仅仅主题本身被抽取出来，而且主题之间分流、合流等关系也被一起抽取出来。我们下面分为抽取单个主题和抽取各个主题之间的主题演化关系两方面分别进行介绍。

（1）抽取单个主题

本文对每一年的专利数据利用 HDP 进行主题聚类，因为 HDP 能够自动决定聚类的个数，所以有些主题在发展过程中可能就消亡了，也可能会新生成一些主题。这产生了一个新问题，对每年的专利数据进行 HDP 进行主题聚类，每年都会生成一系列的主题，那么时间段 *t-1* 的主题如何和时间段 *t* 的主题对应呢？本文使用经典的向量空间模型，HDP 算法生成的每个主题都是一系列词语的组合，通过将各个单词映射到 VSM 中的一维空间，各个主题生成 VSM 中的一个向量，主题之间的相似度的计算就变成对应向量之间夹角角度 Cosine 值的计算。相似度计算公式 (4.1)所示。

$$\mathrm{cosine}(\overline{T_i},\overline{T_j})=\frac{\overline{T_i}\,\overline{T_j}}{|\overline{T_i}|*|\overline{T_j}|}=\frac{\sum_{k=1}^{n} w_{i,k}*w_{j,k}}{\sqrt{(\sum_{k=1}^{n} {w_{i,k}}^2)(\sum_{k=1}^{n} {w_{j,k}}^2)}} \tag{4.1}$$

其中 T_i 、T_j 分别表示第 i、j 个向量，$W_{i,k}$、$W_{j,k}$ 示第 i、j 个向量中的第 k 个单词所占的权重。

我们将 *t-1* 时刻所有的主题都跟 *t* 时刻所有的主题计算相似度，然后根据主题之间的相似度判断两个主题是否是同一个主题。实验取相似度阈值大于 0.61 的两个主题为同

一个主题的可信度比较高，通过对每年的主题与下一年各个主题的对应发现主题整体的变化趋势。

（2）发现主题间分流合流

发现主题间的分流、合流主要使用三个步骤。第一步：对每年专利数据进行 HDP 主题聚类（即 4.2.1 中讲的内容）；第二步：添加历史数据后对其进行 HDP 主题聚类，发现主题的变化；第三步：根据主题之间的变化确定分流合流。

如图 4.2 所示，圆形代表 *t-1* 时主题中的文档，方块代表 *t* 时的主题中的文档，图中(c)、(d)是当前时间 HDP 算法生成的主题，图中(a)、(b)是指添加历史数据(*t-1* 时刻的数据)之后得到的 HDP 主题聚类结果。由图 4.2 可以看出，添加历史信息之后，主题的结构随之发生一些变化，而实验恰恰是根据这些主题的结构发生的变化来发现具体的分流和合流的主题演化关系。图中(a)、(c)为合流的情况，而(b)、(d)为分流的情况。

实际上，合流就是在时间 *t* 时一个聚类中有多少文档是来自时间 *t-1* 的不同的主题；而分流就是在时间 *t-1* 时一个聚类有多少文档被分到时间 *t* 时不同的聚类中去。

$$P_t(s->r)=\frac{\sum_i I(Z_{i,t,old}=s\,\&\,Z_{i,t,new}=r)}{\sum_{i=1}^{n_t} I(Z_{i,t,new}=r)} \tag{4.2}$$

合流的计算即：从 *t-1* 到 *t* 时刻，主题 *r* 由主题 *s* 合并来的文档比例为公式(4.2)所示。

$$P_{t-1}(s->r)=\frac{\sum_i I(Z_{i,t,old}=s\,\&\,Z_{i,t,new}=r)}{\sum_{i=1}^{n_t} I(Z_{i,t,old}=s)} \tag{4.3}$$

分流的计算即：从 *t-1* 到 *i* 时刻，主题 *s* 分化为主题 *r* 的文档比例为公式(4.3)所示。

其中，$Z_{i,t,old}$ 表示时间 *t* 的文档 *i* 在添加历史数据之后的主题编号，$Z_{i,t,new}$ 表示时间 *t* 的文档 *i* 的主题编号。$I(Z_{i,t,old}=s\,\&\,Z_{i,t,new}=r)=\begin{cases}1 & Z_{i,t,old}=s\,\&\,Z_{i,t,new}=r\\ 0 & \text{其它}\end{cases}$

4.2.2 主题可视化

（1）独立主题形成主题河

实验将时刻 *t-1* 时的主题通过计算主题相似度与时刻 *t* 时的主题相对应，主题的宽度是整个主题所有专利文献的个数所占当年所有主题的专利文献总数的比例（由于一篇专利文献可能属于不同的主题，所以这个数量可能会大于每年具体专利文献数量）。

实验将各个主题的变化趋势以 ThemeRiver（主题河）的形式展示给用户，根据图 4.3 所示，每个主题的变化趋势形成一个三次样条曲线高度函数 f_i，将各个曲线叠加在一起形成一个新的三次样条曲线高度函数 g_i，如果叠式图从时间坐标轴开始，那么 g_0=0，具体 g_i 的计算公式为公式(4.4)所示。

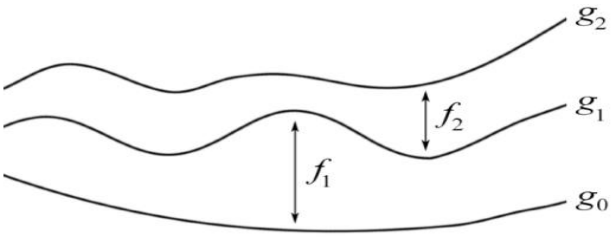

图 4.3　叠式图中曲线函数 f_i 和 g_i 的可视化展示

Fig. 4.3　A visual description of stacked graph functions f_i and g_i

$$g_i = g_0 + \sum_{j=1}^{i} f_j \tag{4.4}$$

（2）各个主题之间相关关系

实验中不仅仅挖掘出各个独立主题的发展变化趋势，而且发现了各个主题之间的分流、合流等关系，为了将分流、合流等更好地展示给用户实验将各个主题的分流、合流也在叠式图中显示。

根据公式(4.2)和(4.3)计算得到 *t* 时刻主题 *i* 中有 *m* 篇文档是由 *t-1* 时刻的主题 *j* 中的文档分流（或合流）而来，则分流（或合流）的宽度为文档总数 *m* 与该主题整个文档个数的比值。主题分流、合流的颜色的是由 *t-1* 时刻的主题 *j* 的颜色与 *t* 时刻主题 *i* 的颜色共同来确定，颜色逐渐由 *j* 的颜色渐变到 *i* 的颜色，可以更形象地表示出主题演化的过程。

4.2.3　主题词排序

主题词的选择不仅仅要考虑该主题词在当前时间段中出现的次数，而且要考虑该主题词在多少不同的主题中出现的次数。如果一个词出现次数很多而且该主题词仅仅在一个主题中出现，那么该词就很能反应该主题的内容；如果一个词出现次数很多，但是该主题词在所有的主题中都出现，而且次数都比较多，那么该主题词就不能很好的反应该主题的内容，公式(4.5)是在时间 *t* 时主题词 *w* 在主题 *k* 中的权重计算公式。

$$Weight_{w,t,k} = \frac{TF_{w,t,k}}{L_{w,t}} \tag{4.5}$$

其中 w 代表一个词，$TF_{w,t,k}$ 代表主题词 w 在时间 t 时在主题 k 中出现的次数，$L_{w,t}$ 代表主题词 w 在时间 t 时在多少主题中出现。

$$Weight_{w,t,k} = \frac{TF_{w,t,k}}{\sum TF_{w,t,k}} * \exp(-\lambda * Weight_{w,t-1,k}) \tag{4.6}$$

与 TextFlow[31]中方法相比公式(4.6)，TextFlow 考虑了词语在当前主题中出现的次数占该词在所有主题中出现的次数的比重，这对于词语比较少的情况下还是比较好用的，但是当每年都有很多新词出现且只在某个单一主题中出现此时所得的权重值为最大值 1，或者是有些词每隔一两个时间段出现一两次，这种情况下权重在中间没出现那个时间段的权重为 0，下一个时间段突然出现，而且只在某一个主题中出现，则权重最大值 1。

4.3 实验结果与分析

4.3.1 语料来源及预处理

实验中采用的专利数据来自中国知识产权局网站，实验选取 2000 年到 2011 年度公布的关于汽车方面的发明专利和实用新型专利，总共有 49791 篇专利文档（实验将所有标题或者摘要中包含“汽车”主题词的专利均下载下来），由于专利法中规定,专利在提出申请以后还需要经过审查程序才能公开,如发明专利从申请日开始一般需要 18 个月才能公开而 2012 年的专利还没有全部被公开，因此论文只使用了 2000 年到 2011 年的数据。

实验对语料做如下预处理：

1. 利用中科院分词工具 ICTCLAS[53]对专利实验数据进行分词（加载汽车专利辞典，该汽车辞典包含了汽车各个领域的专业词汇，总共 5597 个词汇）；

2. 将 HDP 算法得到每个阶段的主题与下一个阶段的主题内容进行标注，如果两个主题描述的内容基本一致则标记为同一个主题。实验选择了人文学院、汽车学院、以及本实验室研究生各一名对主题内容进行标注，如果三人对标注的内容有异议则进行讨论决定是否标记为同一个主题，实验最终标记了 43 对主题。

4.3.2 主题相似度阈值选择

本文使用人工标注的方式进行参数选择，首先标注所有 *t-1* 时间段和 *t* 时间段的主题是否讲述的内容是同一个主题内容，经过标注总共发现 43 对主题讲述的内容基本一致，计算大于阈值 *s* 的准确率、召回率以及 F 值如图 4.4 所示。

所谓准确率即：大于该阈值所有被标注为同一主题的数目和相似度阈值大于该阈值的所有主题数之间的比值。

$$Precision(s)=\frac{\text{大于阈值}s\text{的主题对中被标注为同一主题的数目}}{\text{相似度阈值大于}s\text{的个数}} \tag{4.7}$$

所谓召回率即：大于该阈值所有被标注为同一主题的数目和所有被标注为同一主题数目之间的比值。

$$Recall(s)=\frac{\text{大于阈值}s\text{的主题对中被标注为同一主题的数目}}{\text{所有被标注为同一主题的数目}} \tag{4.8}$$

所谓 F 值即：综合准确率和召回率的评估指标。

$$F(s)=\frac{2*\text{Recall(s)}*\text{Precision(s)}}{\text{Recall(s)}+\text{Precision(s)}} \tag{4.9}$$

因此我们选择 F 值最大的点即阈值为 0.6 时 F 值最大，然后查看大于 0.6 的最小的相似度为 0.61，故选择 0.61 为最终阈值。

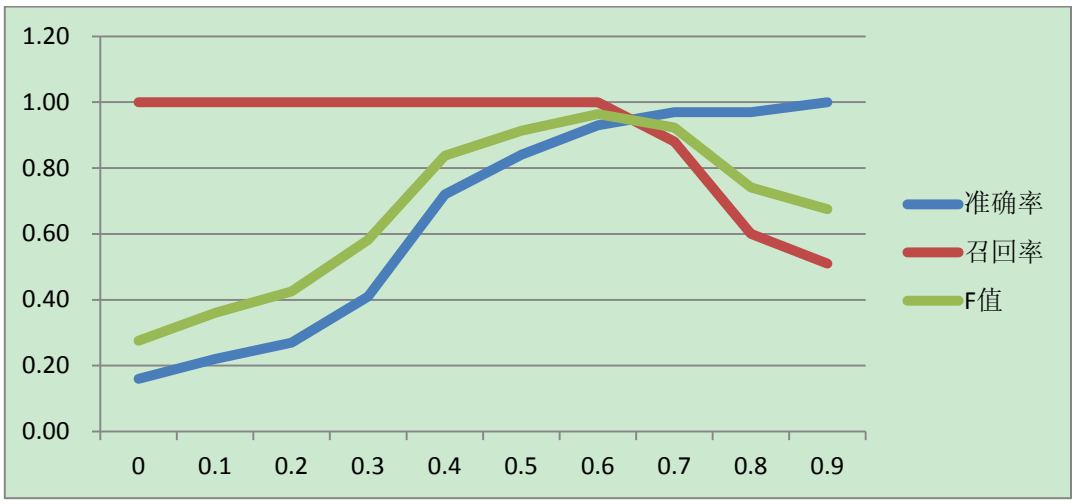

图 4.4　主题相似度和相对应的准确率和召回率

Fig 4.4　Topic similarity and its corresponding precision and recall rate

4.3.3　实验结果

实验总共得到 6 个主题，每个主题的变化趋势在图 4.5 中以主题流的方式显示，主题的分流、合流等变化在图 4.6 中显示，每个主题所对应的具体信息在表 4.1 中显示。

4.3.4　实验结论及分析

由图 4.5 和图 4.6 可知，该汽车专利数据主要包括三大主题：主题 1：车辆动力系统，供暖、制冷系统，节能、净化系统；主题 3：车辆配件、部件，车灯照明系统；主题 5：

智能防盗系统，信号控制系统。这三个主题占据了专利的大部分内容，其中主题 1 和主题 5 所占的比例基本不变，说明这主题 1 在稳步向前发展，主题 3 所占的比例首先减少，后来又增加，整体来说，所占的比例相差不是很大，也是在稳步向前发展。

主题 2（玻璃，车身内饰，车辆维修，舒适性等）在 2002 年开始出现，到 2006 年迎来比较大的发展，所占的比例越来越大，后来在 2007 年、2008 年迅速衰减一直整个主题消亡，而在图 4.6 中可以看出，在 2006 年有很大一部分主题是分流到主题 4 中，而剩下一少部分则逐渐消亡，这说明该主题研究得没有以前那么热。

主题 4（蓄电池，电瓶等，车辆照明，一般制动控制系统或其部件）在 2008 年开始形成一个单独的主题 2010 年达到顶峰然后开始平稳发展，这说明从 2008 年开始电动汽车这部分研究得越来越多，研究地越来越热，这同样与国家最近提倡的节能环保汽车相一致。这个结论与文献[54]的结论“说明电动汽车行业从 2008 年进入快速发展阶段”也是相符合的。

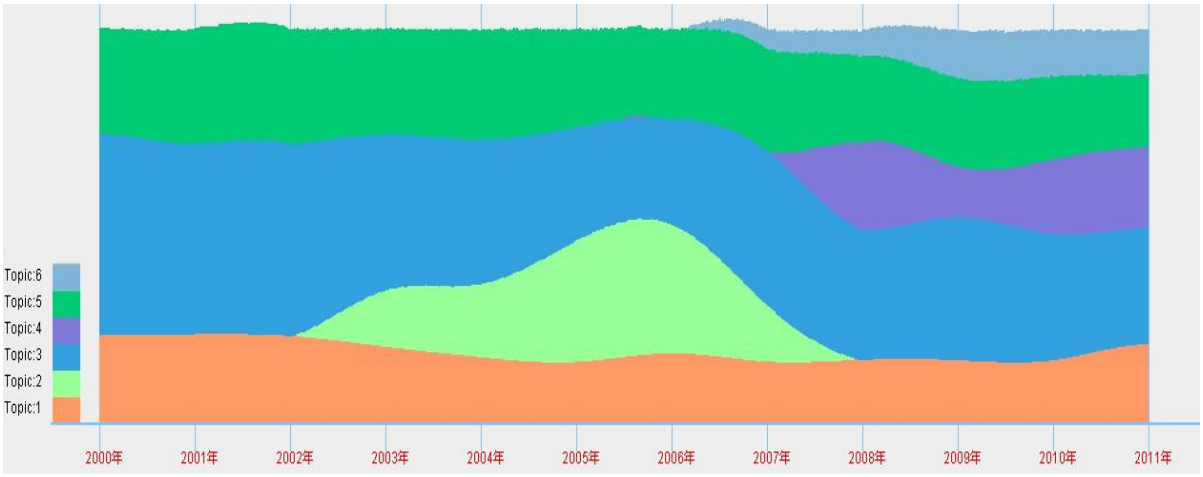

图 4. 5　2000-2011 年各个主题形成的叠式图

Fig. 4.5　Stacked graph of topics between 2000 and 2011

主题 6（安全性，防撞，护罩，防辐射，雾灯/倒车灯）在 2006 年开始发展，到 2011 年整个主题文档数目已经占用所有文档数目 10%左右，这个主题以逐年递增的趋势发展，而这个主题所代表的内容主要是汽车安全性这一块，这说明汽车安全领域越来越受到重视。这与实际情况也是相符合的：随着汽车购买门槛的降低，汽车保有量逐渐升高，越来越多的家庭购买汽车，汽车事故发生数量越来越多，伤亡人数越来越多。据新华社报道，到 2008 年为止，我国交通事故所造成的死亡人数连续十余年居世界之首，每年因交通事故死亡人数均超过 10 万人，这使得人们越来越重视生命，越来越重视汽车安全。

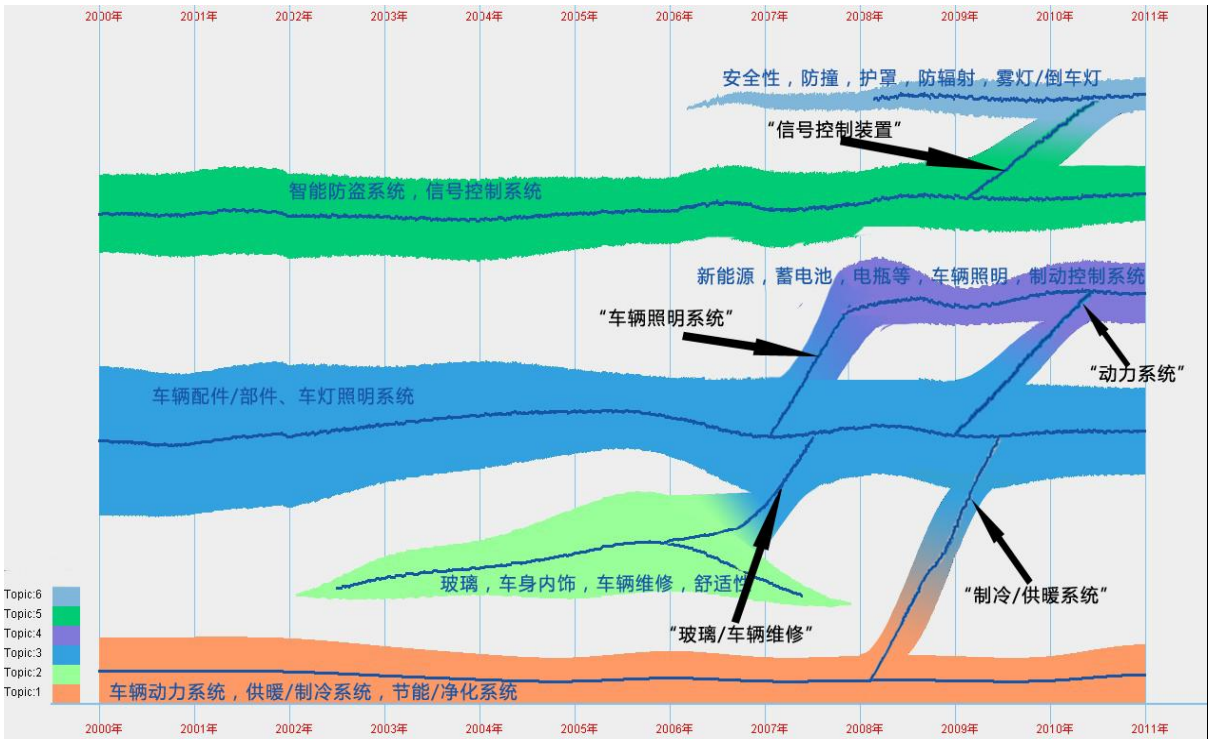

图 4.6　2000 年-2011 年主题分流与合流情况

Fig. 4.6　Stacked graph of different topics with splitting and merging between 2000 and 2011

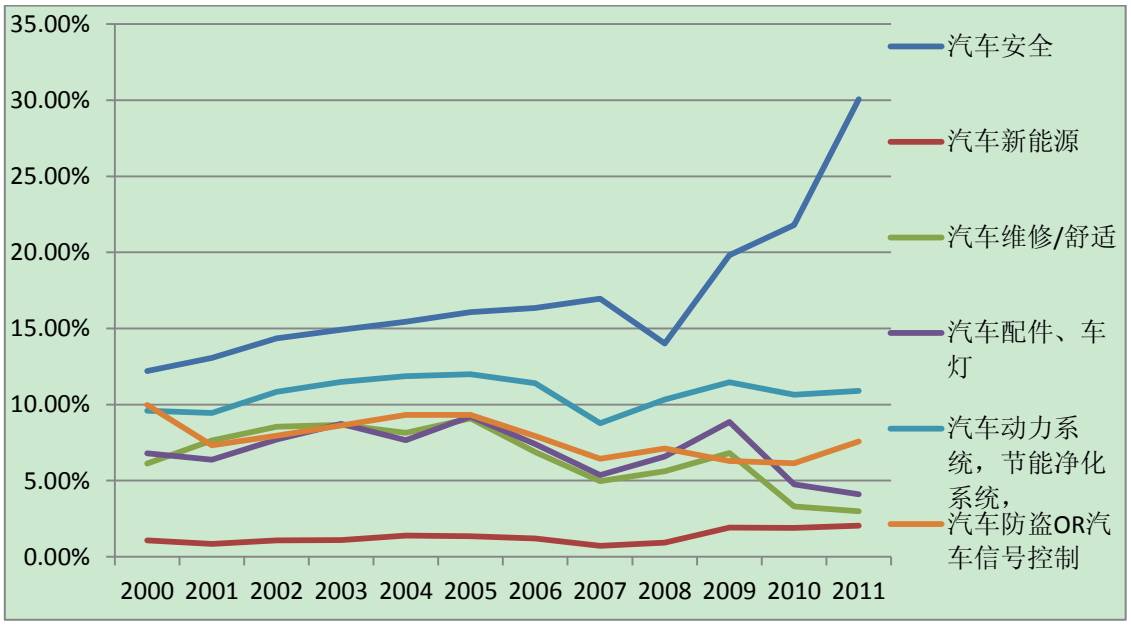

图 4.7　2000 年-2011 年各个主题在 google scholar 中所占比重的趋势图

Fig. 4.7　The proportion of the different topics in google scholar between 2000 and 2011

4.3.5 google scholar 中验证结论的正确性

为了证明实验结果的正确性，本文使用 google scholar 中的数据信息作为佐证。首先在 google scholar 中搜索全文中包含“汽车”主题词的每年文章数目（不包含引用和专利），然后将对应的每个主题的信息按照内容进行搜索（不包含引用和专利），得到每个主题每年的文章个数，然后将每个主题所占包含“汽车”主题词的文章总数的百分比形成主题发展趋势图，在图 4.7 中显示。

表 4.1 主题编号和其对应的主题内容

Tab. 4.1 Topic number and it’s corresponding topic content

主题编号	主题详细内容
1	车辆动力系统，供暖/制冷系统，节能/净化系统
2	玻璃，车身内饰，车辆维修，舒适性
3	车辆配件/部件，车灯照明系统
4	新能源，蓄电池，电瓶等，车辆照明，一般制动控制系统
5	智能防盗系统，信号控制系统
6	安全性，防撞，护罩，防辐射，雾灯/倒车灯

由图 4.7 可以看出，主题 1 在 google scholar 中所占的比重基本上在 10%到 12%之间，变化得不是特别大，这与本文得到的结论几乎如出一辙；

主题 2 从 2000 年到 2005 年一直处于增长状态，从 2005 年开始到 2007 年急速下降到 5%一下，此后已经不成为一个主题，剩下的部分已经并入主题 4 中，这与实验结果也相差无几。

主题 3 所占的比例变化比较明显，但是也不是特别大，基本上都在 5%以上 9%一下，而且所占比例也是首先减少然后增加，这也与结论有一定的相似性；

主题 4 从 2000 年到 2007 年所占比例基本不变，还有点轻微下降的趋势，但是到 2007 年开始，该主题所占比例逐渐增加，这与本文所得到的结论也是一致的。

主题 5 从 2000 年到 2011 年之间虽然有起伏变化，但是整体变化不大，所占比例基本都在一定范围内变化，都在 7%左右徘徊，与实验结果相类似。

主题 6 从 2008 年开始呈现突然增加的趋势，这恰恰说明了，人们对汽车安全的重视程度，这也侧面证明了该主题的正确性。

4.4 本章小结

本文利用 HDP 主题聚类技术对专利数据进行主题聚类，利用每年真实的主题信息与添加历史信息之后的主题结构的变化来发现各个主题之间的分流与合流等相关关系，然后利用三次样条曲线将各个主题组织成叠式图，最后经过调节各个主题之间的布局信息以得到优化的叠式图。本文将专利主题聚类与主题演化以及可视化结合在一起，以友好的方式展示给用户。本文提出的模型不仅仅适用于对专利数据信息的挖掘，对于学科之间的变化，交叉学科的产生过程等各个领域都适用。本文下一步可以将该方法应用到学科发展演化方面，来挖掘各个学科的发展趋势等。

结　论

随着网络技术的不断发展以及大数据时代的来临，以文本资源为首的各种资源充斥着我们生活中的各个角落，如何从浩瀚的文海中利用主题挖掘技术等寻找自己所需的真正有用的信息，对每个用户都是一个巨大的挑战，主题是文本中非常重要隐藏的信息，主题信息可以被当作词项的概率分布，利用文档集出现的词项共现信息可以抽取出语义主题信息。在 pLSI 的基础上改进的 LDA 模型的提出使得主题模型得到了极大的发展并获得大部分学者的认同、LDA 模型迅速在主题模型中占据重要地位，后来 HDP 模型的提出改进了 LDA 模型必须人工确定主题个数的缺点也得到了一定的发展。本文主要利用 LDA 和 HDP 这两个主题模型对主题挖掘技术进行了初步探讨，主要内容如下：

一、本文第三部分探讨了如何对待发表的论文推荐投稿的期刊的问题，将主题挖掘技术与分类技术相结合大大提高了期刊推荐的效果，有较大的实际应用价值。

论文投稿是一个很困难的问题，利用文本挖掘技术实现期刊推荐对投稿人发现最适合的期刊有一定的实用价值，能够有效地避免论文与投稿的期刊研究方向不一致等问题，从而缩短了投稿的时间。本文将 LDA 主题模型与 SVM 分类算法相结合，将论文推荐到主题相似的期刊中，利用基于主题模型的期刊推荐方法 TOP5 能达到 85%的准确率，效果最好，有一定的实用价值。同时实验与其它模型（基于 SVM 的期刊推荐方法、基于内容的期刊推荐方法、基于用户的期刊推荐、基于期刊相似度的推荐方法）进行对比，认真分析了各个模型在期刊推荐中的优缺点。为了研究正确的对论文质量进行评价对期刊推荐结果的影响，本文将论文水平作为一个外部因素对期刊推荐进行选择，用户可以根据自己的实际情况来选择高档、中档、低档三档期刊，实验发现，将论文水平加入到算法中，能大大提高各种期刊推荐算法的准确率，这说明正确评价论文水平对期刊推荐算法具有非常重要的意义。另外，本文发现有些期刊存在发表与自己期刊主题不相符的论文的情况，这种情况在多个期刊中也都出现过，对这种情况本文的方法是很难进行正确推荐的，这也是导致本文结果不是特别好的一个原因。

二、本文第四部分针对主题模型存在主题数目难以确定以及主题存在演化现象这种问题，利用 HDP 主题挖掘算法深入研究了汽车专利的主题演化现象，挖掘专利中的潜在价值对研究人员以及市场决策人员了解主题的研究热点、主题的演变趋势以及对未来主题发展趋势进行预测都有很大的价值。

本文利用 HDP 主题挖掘技术对专利数据进行主题聚类，利用每年真实的主题信息与添加历史信息之后的主题信息的变化来发现各个主题之间的分流与合流等相关关系，然后利用三次样条曲线将各个主题组织成叠式图，最后经过调节各个主题之间的布局信

息以得到优化的叠式图。本文将专利主题演化以及可视化有机的结合在一起，将主题演化信息以友好的方式展示给用户。本文深入研究了主题随时间变化而不断发生主题演化的现象，通过发现主题演化从而帮助用户发现主题研究热点以及主题未来的研究趋势，对研究人员以及市场决策人员未来的决策有很大帮助。

下一步工作可以根据实验中出现的问题来进行研究，主要内容有以下几个方面：

1、改进第三章中论文质量水平的评价方法，对论文的质量水平进行自动化评价，从而提高期刊推荐的准确率。

2、可以将第四章主题演化的方法应用到学科发展演化方面，来挖掘各个学科的发展趋势等。

参 考 文 献

[1] FELDMAN R, DAGAN I. Knowledge discovery in textual databases (KDT)[C].Proceedings of the 1st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Montreal, Canada, 1995: 112-117.

[2] DEERWESTER S C, DUMAIS S T, Landauer T K, et al. Indexing by Latent Semantic analysis[J]. Jasis, 1990, 41(6): 391-407.

[3] HOFMANN T. Probabilistic latent semantic indexing[C].Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, 1999: 50-57.

[4] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2003(3): 993-1022.

[5] WANG X, MCCALLUM A. Topics over time: A non-markov continuous-time model of topical trends[C].Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, 2006: 424-433.

[6] LANCICHINETTI A, Fortunato S. Consensus clustering in complex networks[J]. Scientific Reports. 2012, 2: 336-343.

[7] TEH Y W, JORDAN M I, BEAL M J, et al. Hierarchical dirichlet processes[J]. Journal of the American Statistical Association. 2006, 101(476): 1566-1581.

[8] YANG F, MAO K Z, LEE G K K, et al. Emphasizing minority class in LDA for feature subset selection on high-dimensional small-sized problems[J]. IEEE Transactions on Knowledge and Data Engineering. 2015, 27(1): 88-101.

[9] LEE S, TAE S, JEE N, et al. LDA-based model for measuring impact of change orders in apartment projects and its application for prerisk assessment and postevaluation[J]. Journal of Construction Engineering and Management. 2015(3): 160-168.

[10] FU R, QIN B, LIU T. Open-categorical text classification based on multi-LDA models[J]. Soft Computing. 2015, 19(1): 29-38.

[11] TANG Y W, WANG B, TANG H H, et al. Unsupervised sentiment orientation analysis on micro-blog based on dependency parsing and hierarchical dirichlet processes[C]. Applied Mechanics and Materials, Shenzhen, China, 2014: 1224-1232.

[12] ZHENG J, LIU S, NI L M. Effective mobile context pattern discovery via adapted hierarchical dirichlet processes[C]. 2014 IEEE 15th International Conference on Mobile Data Management (MDM), Brisbane, QLD, 2014: 146-155.

[13] MA T, SATO I, NAKAGAWA H. The hybrid nested hierarchical dirichlet process and its application to topic modeling with word differentiation[J]. Association for the Advancement of Artificial Intelligence 2015(5): 256-264.
[14] 陈慧敏. 中国区域学术论文影响力评价分析[J]. 福州大学学报: 哲学社会科学版. 2012 (3): 38-43.
[15] 马瑞敏. 学术期刊影响力评价研究——基于历时视角的新实践[J]. 中国科技期刊研究. 2014, 25(11): 1397-1403.
[16] 张玉华, 潘云涛. 科技论文影响力相关因素研究[J]. 编辑学报. 2007, 19(2): 81-84.
[17] 张静. 引文, 引文分析与学术论文评价[J]. 社会科学管理与评论. 2008 (1): 33-38.
[18] 刘睿远, 刘雪立, 王璞, 等. 基金论文比作为科技期刊评价指标的合理性-基于 SCI 数据库中眼科学期刊的实证研究[J]. 中国科技期刊研究. 2013, 24(3): 472-476.
[19] 王贤文, 丁堃, 朱晓宇. 中国主要科研机构的科学合作网络分析——基于 Web of Science 的研究[J]. 科学学研究. 2010, 28(12): 1806-1812.
[20] 马峥, 潘云涛. 基于引文分析方法的中国英文科技期刊的交流价值研究[J]. 编辑学报. 2012, 24(4): 307-310.
[21] 吴志荣. 对引文分析法方法论地位的重新思考 [J]. 图书馆杂志. 2012(5): 11-18.
[22] 赵蓉英, 曾宪琴, 陈必坤. 全文本引文分析——引文分析的新发展[J]. 图书情报工作. 2014, 58(9): 129-135.
[23] HEARST M A, DUMAIS S T, Osman E, et al. Support vector machines[J]. Intelligent Systems and their Applications. 1998, 13(4): 18-28.
[24] FRIEDL M A, BRODLEY C E. Decision tree classification of land cover from remotely sensed data[J]. Remote Sensing of Environment. 1997, 61(3): 399-409.
[25] RZENIEWICZ J, SZYMAŃSKI J. Selecting Features with SVM[M]. New York, USA: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 2013.
[26] YANG B, JANSSENS D, RUAN D, et al. A data imputation method with support vector machines for activity-based transportation models[M]. Paris, France: Atlantis Press, 2013.
[27] 范宇, 符红光, 文奕. 基于 LDA 模型的专利信息聚类技术[J]. 计算机应用. 2013, 33(A01): 87-89.
[28] 郝智勇, 贺明科, 谭文堂, 等. 基于多维标度法的专利文本可视化聚类研究倡[J]. 计算机应用研究. 2010, 27(12): 4608-4611.
[29] HAVRE S, HETZLER E, WHITNEY P, et al. Themeriver: visualizing thematic changes in large document collections[J]. IEEE Transactions on Visualization and Computer Graphics. 2002, 8(1): 9-20.
[30] WEI F, LIU S, SONG Y, et al. Tiara: a visual exploratory text analytic system[C]. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 2010: 153-162.

[31] CUI W, LIU S, TAN L, et al. Textflow: towards better understanding of evolving topics in text[J]. IEEE Transactions on Visualization and Computer Graphics. 2011, 17(12): 2412-2421.
[32] TOMAS MIKOLOV, KAI CHEN, GREG CORRADO, et al. Efficient estimation of word representations in vector space[J]. ArXiv Preprint ArXiv.2013,11(3):1301-1313.
[33] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM. 1975, 18(11): 613-620.
[34] CLARKSON P, ROSENFELD R. Statistical language modeling using the CMU-cambridge toolkit[C]. Fifth European Conference on Speech Communication and Technology, Rhodes, Greece,1997: 2707-2710.
[35] HUANG J, KALBARCZYK Z, NICOL D M. Knowledge Discovery from Big Data for Intrusion Detection Using LDA[C]. IEEE International Congress on Big Data (BigData Congress), Anchorage, Alaska, 2014: 760-761.
[36] SMITH D, MCMANIS C. Classification of text to subject using LDA[C]. IEEE International Conference on Semantic Computing (ICSC), Anaheim Marriot, USA, 2015: 131-135.
[37] SOUTO U T C P, BARBOSA M F, DANTAS H V, et al. Screening for coffee adulteration using digital images and SPA-LDA[J]. Food Analytical Methods. 2014: 1-7.
[38] WAN H X, PENG Y. Public opinion hotspot discovery algorithm based on fuzzy clustering LDA[J]. Applied Mechanics and Materials. 2014, 433: 626-629.
[39] CASELLA G, GEORGE E I. Explaining the gibbs sampler[J]. The American Statistician. 1992, 46(3): 167-174.
[40] CAO J, XIA T, LI J, et al. A density-based method for adaptive LDA model selection[J]. Neurocomputing. 2009, 72(7): 1775-1781.
[41] JOSEPH TURIAN， LEV RATINOV， YOSHUA BENGIO. Word representations：a simple and general method for semi －supervised learning[C]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala，Sweden，2010：384-394.
[42] GUTHRIE D, ALLISON B, LIU W, et al. A closer look at skip-gram modelling[C]. Proceedings of the 5th International Conference on Language Resources and Evaluation, genoa, Italy, 2006: 1-4.
[43] 王超, 吕俊生. 国内外学术信息推荐方法研究进展[J]. 情报杂志. 2013, 32(9): 142-147.
[44] 徐键. 基于 PAGERANK 的科技论文推荐系统[J]. 电子世界. 2013 (1): 104-105.
[45] 倪卫杰. 基于用户兴趣模型的个性化论文推荐系统研究[D].天津: 天津大学,2010.
[46] 杜永萍, 杜晓燕, 姚长青. 基于主题效能的学术文献推荐算法[J]. 北京工业大学学报. 2015, 41(2): 10-18.

[47] 邓少伟,罗泽,李树仁,等. 基于论文共同作者学术关系的学者推荐系统[J].计算机工程，2013，39(2) : 12-17.

[48] 陆艳春. 学术会议话题趋势分析与推荐方法研究[D].湖北:华中科技大学, 2012.

[49] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering. 2005, 17(6): 734-749.

[50] AGRAWAL R, IMIELIŃSKI T, SWAMI A. Mining association rules between sets of items in large databases[C]. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data，Washington, DC, 1993: 207-216.

[51] HECKERMAN D. Bayesian networks for data mining[J]. Data Mining and Knowledge Discovery. 1997, 1(1): 79-119.

[52] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST). 2011, 2(3): 27.

[53] Institute of Computing Technology, Chinese Academy of Sciences. ICTCLAS [EB/OL]. [2012-08-10]. http://www.ictclas.org (in Chinese)(中国科学院计算技术研究所. ICTCLAS [EB/OL]. [2013-08-10]. http://www.ictclas.org).

[54] 罗立国, 余翔, 周力虹, 等. 我国电动汽车技术领域专利网络研究[J]. 情报杂志. 2012,31(12):1-8.

攻读硕士学位期间发表学术论文情况

1　基于 HDP 的汽车专利主题演化研究.**王亮**，张绍武，丁堃，许侃，林鸿飞. 情报学报，2014 年，33 卷（9 期）：944-951. 主办单位：中国科学技术情报学会和中国科学技术信息研究所。（本硕士学位论文第四章）

2　基于词向量的期刊推荐方法研究.张绍武，**王亮**，林鸿飞.第二十一届全国信息检索学术会议（ccir2015）.主办单位：中国中文信息学会和中国计算机学会（本硕士学位论文第三章）(在投)

致　　谢

时光荏苒，岁月如梭，在大连理工大学的三年的研究生生活转瞬即逝，回首这三年，研究室生活依然历历在目，期间有收获、有感动、有坎坷、有困难。在论文完成之际，向那些曾经帮助我的老师、同学、朋友、亲人致以最诚挚的谢意！

首先，非常感谢我的指导教师张绍武副教授，张老师对待我们如同对待家人一样，对我们悉心教导，张老师思想开放、生活不拘小节、能够让我们在实验室愉快地搞科研。对我们生活中、科研中、学习中遇到的各种困难，张老师都能指导我们逐渐解决困难。从老师那里，我们学到了很多书本上学不到的知识。

同时，非常感谢林鸿飞教授，通过林教授我来到了梦寐以求的信息检索研究室，在实验室的老师、师兄师姐等人的帮助下我学到了搜索引擎、推荐系统、数据挖掘、文本可视化、情感分析等高深内容。在此，特别感谢许侃老师、孙晓玲老师、林原老师、杨亮师兄等人对我论文等方面的帮助和支持。

其次，非常感谢实验室的成员，尤其是“文本情感一家亲”组内的成员，和他们在一起可以一起探讨学术问题、一起游玩、一起聚餐、一起打羽毛球、一起参与元旦晚会表演等等，和他们在一起我感受到了家庭般的温暖，和他们一路同行，让我的研究生生活无比充实，我无比珍惜和他们在一起的每一分每一秒。

另外，非常感谢宿舍的三位同学赵明珍、张建伟、张虎，和他们在一起我养成了每天晚上睡前刷牙洗脸泡脚并及时洗袜子的好习惯，并学会了很多为人处世的道理，很荣幸和他们在一起。

最后，非常感谢我的家人，无论我在哪里，无论我的选择是什么，他们始终坚定地支持我，无形之中给予我最强大的力量。

临毕业之际，真心感谢每一个给我帮助的老师和同学。

衷心感谢对本篇论文进行审阅的各位专家教授！

大连理工大学学位论文版权使用授权书

本人完全了解学校有关学位论文知识产权的规定，在校攻读学位期间论文工作的知识产权属于大连理工大学，允许论文被查阅和借阅。学校有权保留论文并向国家有关部门或机构送交论文的复印件和电子版，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、或扫描等复制手段保存和汇编本学位论文。

学位论文题目：_____

作 者 签 名：_____ 日期：____年____月____日

导 师 签 名：_____ 日期：____年____月____日