中文图书分类号: TP391.1
密　　级: 公开
UDC: 004
学 校 代 码: 10005

硕 士 学 位 论 文

MASTERAL DISSERTATION

论 文 题 目: 基于聚类约束的高质量微博检索方法研究与应用

论 文 作 者: 王凯

学　　科: 计算机技术

指 导 教 师: 杨震 教授

论文提交日期: 2018 年 5 月

UDC: 004
中文图书分类号: TP391.1
学校代码: 10005
学　　号: S201507104
密　　级: 公开

北京工业大学硕士专业学位论文

题　　目: 基于聚类约束的高质量微博检索方法研究与应用

英文题目: High Quality Microblogging Retrieval Method Based on Clustering Constraints

论 文 作 者: 王凯
学 科 专 业: 计算机技术
研 究 方 向: 信息安全
申 请 学 位: 专业硕士
指 导 教 师: 杨震 教授
所 在 单 位: 信息学部计算机学院
答 辩 日 期: 2018 年 5 月
授予学位单位: 北京工业大学

独 创 性 声 明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京工业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名：_____

日期：2018 年 5 月 20 日

关于论文使用授权的说明

本人完全了解北京工业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

社交媒体的兴起，不仅降低了人们沟通的成本，而且改变了人们消费信息的习惯，人们不再满足于被动的消费信息，转而成为制造和传播信息的主体，使得数据传播迅猛，数据量空前巨大。以微博媒体为例，微博媒体的短文本特性，如长度较短，广泛使用特殊字符，表达口语化等，使得传统长文本检索方法在微博检索中性能退化，甚至完全不可用。

但是，主流社交媒体平台，如微博、Twitter 和 Facebook 等，又迫切希望构建快速、智能的信息过滤系统，为用户提供更加有效的信息推送服务。这就需要对适用于微博短文本检索的方法进行深入研究。

现有的许多改进短文本检索性能的方法中，通过引入外部信息提升微博检索性能的方法，简单易行且性能优异，得到了研究者的广泛关注。但随着对引入外部信息方法研究的深入，研究者发现面对微博检索问题，现有方法仍存在以下问题亟待解决：

1. 相关微博排序困难。通常检索方法能检索出大量相关微博，但是如何排序，使有限的推送中包含更多的信息，如何去掉冗余的信息，使推送服务质量更高，仍然有待研究。
2. 微博文本有效聚类困难。由于微博数据量大，文本短，表达口语化等特性，通常的聚类方法效果较差。

为了解决上述问题，本文提出了一种微博检索方法，通过结合微博文本的聚类信息，达到理解用户实际搜索意图，提高检索性能的目的。本文的主要成果总结如下：

1. 提出了一种微博检索框架，探究了几种基本查询扩展方法对检索性能的影响。
2. 提出了一种多元检索模型，比较验证了该多元检索模型的检索性能。
3. 提出了一种基于非负矩阵分解的聚类方法（BNMF，Basic Non-negative Matrix Factorization），在聚类约束下提升了检索模型的检索性能。
4. 提出了一种基于相关约束的聚类方法（RNMF，Relevance Non-negative Matrix Factorization），对比于 BNMF，验证了该聚类方法的性能。

本文在 TREC（Text REtrieval Conference）提供的 Microblog 数据集上进行的实验表明，基于聚类约束的高质量微博检索方法，相比较于基本检索方法，能够有效提升微博检索性能。同时，基于相关约束的聚类方法，相比较于基本非负矩阵分解的聚类方法，有性能上的提升。

关键词：微博检索；多元检索模型；非负矩阵分解；微博聚类

Abstract

The rise of social media not only reduces the cost of people's communication, but also changes people's habit of consuming information. People are no longer satisfied with passive consumption information and become the main body of manufacturing and disseminating information, making the data spread rapidly and the amount of data is unprecedented. Taking microblogging as an example, the short text features of microblogging, such as short length, wide use of special characters, and colloquial expression, make the traditional long text retrieval methods in microblogging retrieval neutral degradation, or even completely unavailable. However, the mainstream social media, such as Microblogging, Twitter and Facebook, are eager to build a fast and intelligent information filtering system to provide users with more effective information push services. This requires in-depth study of microblogging short text retrieval methods.

The method of introducing external information to improve retrieval performance is simple and effective, which has attracted wide attention from researchers. However, with the in-depth study of the method of introducing external information, researchers find that there are several problems in solving short text retrieval:

1. Relevant microbloggings are difficult to sort. Generally, a large number of relevant microbloggings can be retrieved, but how to sort the relevant microbloggings, making the limited push contain more information and pushing the quality of service higher, still remains to be studied.
2. The effective clustering of microbloggings is difficult. Because of the large amount of data, short text and colloquial expression of microblogging, the usual clustering method is not effective.

In order to solve the above problems, this paper proposes a microblogging retrieval method. By combining the clustering information of microblogging, it can achieve the purpose of understanding users' actual search intention and improving retrieval performance. The main achievements of this paper are summarized as follows:

1. A microblogging search framework is proposed, and exploring the impact of several basic query expansion methods on retrieval performance.
2. A multiple retrieval model is proposed, comparing and verifying the retrieval performance of the multiple retrieval model.
3. A clustering method based on non-negative matrix factorization (BNMF, Basic Non-negative Matrix Factorization) is proposed, improving the retrieval perfor-

mance of the retrieval model with the clustering constraints.

4. A clustering method based on relevant constraints (RNMF, Relevance Non-negative Matrix Factorization), which is compared to BNMF, verifying the performance of the clustering method.

The experiments on the microblogging data set provided by TREC (Text REtrieval Conference) show that the high quality microblogging retrieval method based on clustering constraints can effectively improve the performance of micro-blog retrieval compared with the basic retrieval method.

Key words: microblogging search; multiple retrieval model; non-negative matrix factorization; microblogging clustering

目 录

插图索引

表格索引

第 1 章　绪论

随着社会的发展，人们的社交需求逐步倾向于实时沟通和实时分享。社交媒体正是解决人们实时沟通和实时分享的桥梁，其中微博使用广泛，广受欢迎。由于微博本身是为了快速传播设计的，因此必须短小精干易于传播，并且随着传播范围扩大，人们不断的转发分享，使得微博包含的信息不断扩大，同时也使得人们获取特定领域特定信息的难度越来越大。本文提出的基于聚类约束的高质量微博检索方法期望能解决这一问题。对于如何提高检索质量，针对微博本身的特点，大量的研究工作集中在增加外部信息和改进模型结构。本文将对微博检索问题基于外部数据做查询扩展和聚类重排序，查询扩展能有效提高检索性能，并且检索结果将通过排序呈现，聚类重排序是根据微博聚类约束改进检索结果的排序。基于以上两点，本文提出了一个基于聚类约束的微博检索框架，期望利用丰富的外部信息来探寻用户的真实查询意图，满足用户的检索需求，提高信息获取的效率。

1.1　研究背景和意义

互联网的广泛使用使得存储信息的数量呈现指数增长，并且这些信息需要能通过网络快速访问。社交媒体的出现，已经深刻地改变了人们在线产生和使用信息的方式。国外这类社交媒体发展较早，如 Twitter，FaceBook 等。推特目前拥有 3.3 亿用户，平均每分钟能产生 100000 条信息，面对如此庞大的信息量，需要更先进的检索技术，帮助人们获取有效信息。国内这类媒体如微信朋友圈，微博等，同样面对这样的问题，中国互联网络信息中心（China Internet Network Information Center, CNNIC）最新发布的《中国互联网络发展状况统计报告》显示，截至 2017 年 12 月，我国网民规模达 7.72 亿。其中，微博凭借用户基数大，易于传播分享的特点，截至 2017 年第三季度，新浪微博月活跃用户达到 3.76 亿，同时微博用户使用率持续增长，如图1.1所示达到 40.9%，对新闻，网络话题影响力不断提升。

随着人们对实时沟通和实时分享的需求越来越强烈，当热点问题发生时，人们会在微博中检索，表达他们的意见或者转发分享给更多的人。如微博这类媒体和主流新闻媒体网站（如 CNN 或 nytimes.com）最大的区别在于社交网络中的人们是信息生产者和消费者，而不是单纯的接收信息。这种特性也使得社交网络中的信息杂乱无章，增加了用户获取感兴趣信息的难度。因此人们对精准信息获取的需求不断增长，如信息推荐，用户关系预测和准确性广告推送等，微博检索技

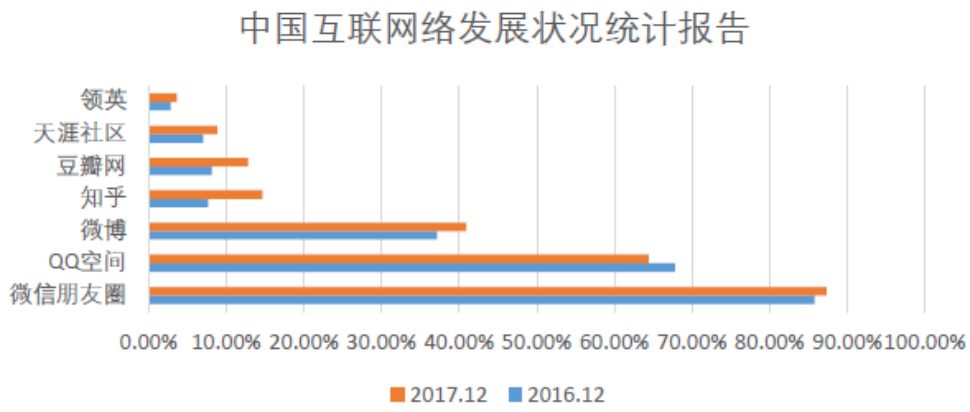

图 1.1　中国互联网络发展状况统计报告

Fig 1.1　Statistical report on the development status of Internet in China

术应运而生。

然而，由于微博通常只包含几个词，所以传统的信息检索（Information retrieval, IR）模型（包括向量空间模型的许多变体，概率模型和文本分类）在解决微博检索问题时遇到了困难。首先，由于微博非常简短，词汇不匹配问题变得相当严重。其次，如果没有足够的单词样本，检索模型预测变得非常困难。最后，面对大量短文本很难准确提取关键信息，信息推荐容易发生主题偏移。在过去的几年中，人们提出了许多方法，包括利用各种数据源，平台和知识库进行查询扩展，改进检索模型结构等，来提高微博过滤的性能。同时为了使用户更高效的获取信息，如何对现有信息排序也是一个亟待解决的问题。因此，研究者一直在深入研究微博检索问题，期望获得更好的检索性能。

1.2　国内外研究现状

在传统检索领域，许多检索模型提出较早，例如向量空间模型，概率模型和语言模型等，不仅作用于文本检索，并且在近年来已经被成熟用于其他领域的工业化生产中。例如，Pasquier, C 和 Gardes, J[1] 根据向量空间模型，在高维向量空间中表示了 RNA（核糖核酸）和疾病信息，通过计算相似性定义 RNA 和疾病之间的关系，为疾病诊断提供了新的方法。Anastasopoulos. et al.[2] 提出了一种无监督概率模型，应用于小语种的语音翻译工作。Richard, Alexander 和 Gall, Juergen[3] 基于统计语言模型，提出了一种对视频动作进行检索分类的方法，来准确获取视频片段。

在面对微博检索任务时，为了准确匹配微博和用户，必须克服微博本身具有的局限性。例如，Teevan J. et al.[4] 研究发现，通常用户不会直接检索感兴趣的内容，而是通过人名等其他信息检索感兴趣的内容。并且微博通常只包含有限数量的词，包括向量空间模型[5,6]，概率模型[7,8] 和语言模型[9] 等在内的传统的信息

检索（IR）模型遇到了困难。此外，由于微博非常简短，词汇不匹配问题变得相当严重。最后，因为没有足够的单词样本，构建词典进行分析和构建训练数据变得十分困难。在过去几年中，针对以上问题研究者们提出了许多方法，主要包括利用外部信息和改进检索模型两个方向。

查询扩展[10] 是利用外部信息的典型方法，其中，伪相关反馈（Pseudo relevance Feedback, PRF) 技术[11] 应用广泛，首先假定关联文档中的大多数文档项是有意义的，然后根据相关文档进行主题扩展，最后使用扩展的查询来检索内容，得到更有价值的微博检索结果。Ganguly. et al.[12] 提出了一种通过检索信息生成查询扩展的方法：将查询提交给搜索引擎，并将返回的结果视为信息源的扩展以生成查询扩展。同时还有一些其他利用外部信息的方法，Yue. et al.[13] 提出利用社交信息和位置信息丰富用户样本，根据用户潜在兴趣，检索相关微博。这表明使用外部信息不仅限于查询文档，还可以利用更多其他类型的信息。Kalloubi. et al.[14] 提出了一种利用知识库的方法，通过将微博关联外部语义，并且结合微博内容实体之间关系的概念图，解决数据稀疏和语义不匹配的问题。同样，Chen, Yan. et al.[15] 也使用了概率图方法解决微博自动分类和摘要生成。

改进检索模型同样也是一种有效的方法。Yashen. et al.[16] 结合了伪相关反馈改进概率模型，有效的优化了微博检索任务的性能。Koustav. et al.[17] 提出了一种基于低频关键词汇的自动分类方法，用于微博的检索分类。Wu, Fangzhao. et al.[18] 提出了一种在微博中使用分类器来寻找垃圾邮件发送者帐户的有效方法。Soichiro. et al.[19] 使用训练数据来训练支持向量机（SVM）来判断微博是否相关。Albishre, Khaled. et al.[20] 使用伪相关反馈结合语言模型，获得了很好的检索效果。这些研究证明了以上两种方法的有效性。

在微博检索的现有研究中也有应用聚类方法，例如，新闻工作者有时候需要从社交媒体中提取某个事件，包括人物，时间，地点，内容。但目前微博事件的检索，通常使用关键字，事件实体或者选定的微博来表示事件，无法提取事件的细节。Peiquan. et al.[21] 提出了一种事件聚类方法，将所有相关事件汇总到一个类簇中，然后从中提取整个事件，取得了较优的性能。Yulu. et al.[22] 使用了将整个时间区间划分为时间片段的方法，并且在每个片段内执行了选择性搜索：这些片段可以使用批处理或在线算法进行聚类。Zhang, Shunxiang. et al.[23] 提出了使用用户聚类的方法，在用户类簇内进行微博推荐，具有良好的准确率。这些工作证明了聚类方法在文本检索中的作用。

矩阵分解算法是聚类任务中常用的一种方法，在许多领域都取得了良好的效果。例如，在推荐方向，Lei et al.[24] 针对新用户推荐开发了一种有效的联合优化模型，称为局部代表性矩阵分解。Zhang, Guoying. et al.[25] 采用非负矩阵分解方法[26]，分别构建了用户关系矩阵和项目类型矩阵，给出了更有效的内容推荐。

同时，可解释性对于用户推荐系统具有重要意义，一份可解释的推荐或建议对于用户来说更易于接受。Behnoush et al.[27] 以矩阵分解模型为基础，并进一步假定没有任何额外的数据源，例如项目内容或用户属性，提出可解释矩阵分解技术。

通常的研究都是在矩阵分解过程中添加外部信息，Apurva et al.[28] 的研究工作表明，与传统矩阵分解技术相结合的外部特性可以有效的优化结果，对推荐和聚类都有启发作用。Jun et al.[29] 使用矩阵分解来完成协同排序，在矩阵分解中加入用户特征来改进推荐的排序结果，取得了良好的效果。Felix et al.[30] 提出了一种新方法，利用一些额外的数据维度，例如用户的重复购买行为或者其他可推广的研究结论，提高了聚类任务的性能。

此外，还有一些针对微博自身特点而进行的研究。Basu. et al.[31] 提出了一种针对微博语言特点的词干还原方法，并应用与检索系统。Liu, Mengchen. et al.[32] 提出了一种针对微博社交网络，提取热门用户，用户热门微博，话题标签等的图模型方法。You, Sukjin. et al.[33] 提出了一种针对微博的时间敏感性，通过事件识别算法对检索到的微博进行重排序的方法，改进了检索性能。Li, Haojie. et al.[34] 提出了一种针对微博图像内容，运用文本和图片线索的多图半监督学习方法。Bansal. et al.[35] 通过研究微博话题标签，将话题标签内容扩展到检索系统中，有效改进了检索性能。Huifang. et al.[36] 提出了一种针对微博用户的检索方法，将用户兴趣以标签的形式构建用户兴趣矩阵，并研究标签之间的相互关系，最终获得用户的兴趣，同样提高了检索性能。

在众多的检索方法中，添加外部信息的方法简单有效，同时，采用聚类方法有助于获得微博事件情况，因此本文提出基于聚类约束的高质量微博检索方法，并将探索微博聚类方法，期望提高检索性能。

1.3 论文组织结构

本文各章节主要内容如下：

绪论。本章从微博检索技术领域现状出发，提出了本文的研究背景与研究意义，并对国内外现有的研究成果进行一个简单的介绍，最后对本文的研究内容进行了说明。

第二章，相关研究基础综述。本章首先介绍了传统检索模型的优缺点。然后介绍查询扩展技术在微博检索中的应用。最后介绍本文所涉及非负矩阵分解方法及 KKT 条件。

第三章，基于聚类约束的高质量微博检索方法研究。本章详细阐述了非负矩阵分解方法应用与微博聚类，完成建模理论基础，对面对的问题进行描述，结合微博检索任务的基本检索结果，完成建模理论证明和算法分析，最后给出了具体实现的算法流程。

第四章，基于相关约束的微博聚类方法研究。本章详细阐述了正则化非负矩阵分解方法，并应用于微博聚类，结合微博检索相关信息，完成建模理论证明和算法分析，最后给出了具体实现的算法流程。

第五章，对比实验设计以及性能分析。本章主要简述系统设计方案，对比实验设计方案，根据实验结果证明算法性能，和可行性，并解释了算法中参数对最终结果的影响。

第六章，总结和展望。

第 2 章　相关研究综述

上一章已经对基于聚类约束的高质量微博检索方法的背景、意义以及国内外现状进行了一个简单的介绍。本章主要就现阶段微博检索技术取得的成果做介绍，同时对现有的 NMF 及其扩展方法，问题求解方法进行简单说明。

2.1　信息检索基本方法

同其他信息检索问题一样，随着信息的累积增加，微博数据信息的存储和检索都成为难题，每时每刻新的微博又在大量的产生，并不断传播，即微博的信息爆炸。如何从这些不断扩张的信息中提取出具有特定意义的信息，成为当今微博面对的一大难题。为了解决这一问题，微博检索系统应运而生。在开始的阶段，微博检索模型借鉴了传统的信息检索模型，但是由于微博检索相对传统信息检索具有独特的特征，同时需要更高的准确性和实时性，于是出现了很多针对微博检索问题的方法。

信息检索在工业界有很多成熟的应用，基本的检索模型主要包括：布尔检索模型、向量空间检索模型[37]、概率检索模型[38]、语言模型[39]，Learning To Rank(LTR) 等。下面将简要介绍这些基本模型。

2.1.1　布尔模型

布尔模型是一种以布尔表达式为依据衡量查询和文档相似程度，以关键词匹配数量确定匹配结果的检索模型。布尔模型简单快速，在一些简单过滤问题中比较有效，并且在一些早期的商业系统中有所应用。由于布尔模型是仅考虑关键词的精准匹配，在关键词准确的情况下，很容易返回精确的结果，但是构建这样的关键词或关键词表达式是很困难的。同时，由于无法区分不同的词项对相关性的贡献，使得这一方法缺少实际引用场景。

布尔模型是通过布尔逻辑表达式：逻辑与（AND），逻辑或（OR），逻辑非（NOT）组合表示用户的检索需求。例如，我们希望查找与比特币相关的信息，并且使用“比特币 AND （区块链 OR 数字货币）”这样的表达式作为查询，即希望查找一篇文档，包含了“比特币”并且也包含了“区块链”或“数字货币”。如表2.1所示，根据查询表达式文档 1 和文档 3 符合条件，即检索结果为文档 1 和文档 3。

表 2.1　布尔检索模型示例
Tab. 2.1　Boolean retrieval model example

关键词文档矩阵	文档 1	文档 2	文档 3	文档 4
比特币	1	0	1	1
区块链	1	1	1	0
数字货币	0	1	1	0

2.1.2　向量空间模型

向量空间模型是由 Salton 在 1971 提出[5,40] 的检索模型。向量空间模型的基本思想是将对文本的处理转化为向量空间中的向量运算，将查询和文档用向量表示，再计算向量之间的相似度表示语义的相似度。

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j}) \tag{2-1}$$

$$q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q}) \tag{2-2}$$

如上式2-1，对于向量空间模型，如何计算词项权重（即 $(w_{1,j}, w_{2,j}, \ldots, w_{t,j})$）是影响检索性能的关键因素。常用表示方法包括词频（Term Frequency, TF），逆文档频率 (Inverse document frequency, IDF) 以及综合考虑他们的关系构建的词频-逆文档频率（TF-IDF）。词频权重计算方法如下式2-3。$tf_{i,j}$ 表示文档 d_j 对于单词 t_i 的词频权重，$n_{i,j}$ 表示单词 t_i 出现在文档 d_j 中的频率，$\sum_k n_{k,j}$ 表示文档总词数。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2-3}$$

单纯的词频表示方法往往不能表达不同词项对文档区分能力的差异，因此有研究提出逆文档频率，如公式2-4所示，式子中 $|D|$ 代表文档集合中的文档总数，$|\{d \in D : t \in d\}|$ 表示包含词项 t_i 的文档数量（即 $tf_{i,j}$ 不为 0 的文件数目）。有时在匹配过程中会出现匹配缺失的情况，因此需要平滑处理，使用 $|\{d \in D : t \in d\}|+1$ 进行修正。

$$idf_i = \log \frac{|D|}{|\{d \in D : t \in d\}| + 1} \tag{2-4}$$

在实际使用中，通常采用 TF-IDF 计算词项权重，表示某单词在某文档中出现次数较多，而在其它文档中出现次数较少，则该单词应具有更大的权重。计算

方法如下 (2-5) 所示。

$$tf - idf_{i,j} = tf_{i,j} \times idf_i \tag{2-5}$$

通过计算得到词项权重，就将查询和文档表示为向量2-1，查询和文档之间的相似度就可以通过计算向量之间的相似度来度量，向量中的一个维度代表一个词项。常用距离计算方式是余弦相似度，如公式2-6，其中 $d_j \cdot q$ 是查询向量和文档向量的内积。$\|\mathbf{d_j}\|$ 是向量 $\mathbf{d_j}$ 的模，$\|\mathbf{q}\|$ 是向量 $\mathbf{q}$ 的模。

$$\text{sim}(d_j, q) = \frac{\mathbf{d_j} \cdot \mathbf{q}}{\|\mathbf{d_j}\| \cdot \|\mathbf{q}\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \tag{2-6}$$

向量空间模型简单有效，但是文档转换为向量的过程中损失了很多语义信息，没有考虑各个词项之间的联系。

2.1.3　概率模型

概率模型[41,42]的理论基础是贝叶斯决策理论，应用于信息检索领域。对于一个查询来说，将所有的文档分为了两类，相关文档和不相关文档，这样就转为了一个相关性的分类问题。对于某个文档 D 来说来说，$P(R|D)$ 表示该文档属于相关文档的概率，$P(NR|D)$ 表示该文档属于不相关文档的概率，如果 $P(R|D) > P(NR|D)$ 则该文档是与查询相关的。应用贝叶斯公式则：

$$P(R|D) > P(NR|D) \tag{2-7}$$

$$\Leftrightarrow \frac{P(D|R)P(R)}{P(D)} > \frac{P(D|NR)P(NR)}{P(D)} \tag{2-8}$$

$$\Leftrightarrow \frac{P(D|R)}{P(D|NR)} > \frac{P(NR)}{P(R)} \tag{2-9}$$

这样问题就转化为计算 $P(D|R)$ 和 $P(D|NR)$ 的值，然后按照 $\frac{P(D|R)}{P(D|NR)}$ 降序排列即可得到最终的相关性排序。通常，为了计算 $P(D|R)$ 和 $P(D|NR)$ 的值，基于二元独立模型做出了两个假设：

- 二元假设
  假设单词只有出现和不出现的两种情况，而不考虑词频等其他因素。
- 单词独立性假设
  假设单词之间没有任何联系，即每个单词的出现和不出现均与其他单词无关。

对于文档 D，用 p_i 表示第 i 个单词在相关文档中出现的概率，用 s_i 表示在不相关文档中出现的概率，则 $\frac{P(D|R)}{P(D|NR)}$ 可表示为：

$$\frac{P(D|R)}{P(D|NR)} = \prod_{i:d_i=1} \frac{p_i}{s_i} \times \prod_{i:d_i=0} \frac{1-p_i}{1-s_i} \tag{2-10}$$

$$= \prod_{i:d_i=1} \frac{p_i}{s_i} \times (\prod_{i:d_i=1} \frac{1-s_i}{1-p_i} \times \prod_{i:d_i=1} \frac{1-p_i}{1-s_i}) \times \prod_{i:d_i=0} \frac{1-p_i}{1-s_i} \tag{2-11}$$

$$= (\prod_{i:d_i=1} \frac{p_i}{s_i} \times \prod_{i:d_i=1} \frac{1-s_i}{1-p_i}) \times (\prod_{i:d_i=1} \frac{1-p_i}{1-s_i} \times \prod_{i:d_i=0} \frac{1-p_i}{1-s_i}) \tag{2-12}$$

$$= \prod_{i:d_i=1} \frac{p_i(1-s_i)}{s_i(1-p_i)} \times \prod_{i} \frac{1-p_i}{1-s_i} \tag{2-13}$$

$$= \prod_{i:d_i=1} \frac{p_i(1-s_i)}{s_i(1-p_i)} \tag{2-14}$$

在公式2-10中 $d_i = 1$ 表示在文档中出现的单词，$d_i = 0$ 表示没在文档中出现的单词。$\prod_i \frac{1-p_i}{1-s_i}$ 表示各个单词在所有文档中的计算结果，与具体文档无关，因此可以消去。接下来只需要统计文档 D 中各个单词在相关文档和不相关文档中出现的概率即可。设文档数量为 N，其中相关文档数量为 R，不相关文档数量为 $N-R$，文档中出现第 i 个单词的数量为 n_i，其中相关文档出现第 i 个单词的数量为 r_i。在加入平滑处理以后，可得如下公式2-15：

$$p_i = \frac{r_i + 0.5}{R + 1} \tag{2-15}$$

$$s_i = \frac{n_i - r_i + 0.5}{N - R + 1} \tag{2-16}$$

最终可得如下公式2-17：

$$score(Q, D) = \sum_{q_i=d_i=1} \log \frac{(r_i + 0.5)(N - R) - (n_i - r_i) + 0.5}{(n_i - r_i + 0.5)(R - r_i + 0.5)} \tag{2-17}$$

公式2-17表示，累加查询 Q 和文档 D 中的单词估值，就可以得到查询 Q 和文档 D 的相关性度量。通常，并不能确定哪些文档是相关的，哪些文档是不相关的，可以通过给公式的估算因子赋值，则该公式将会退化为 IDF。

由于概率检索模型只考虑了单词的出现与不出现两种情况，不符合实际应用的场景，因此，在考虑了单词在查询中的权值及单词在文档中的权值衍生出 okapa BM25 模型，该模型具体计算方法如下2-18：

$$score(Q, D) = \sum_{i=1}^{n} idf(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \tag{2-18}$$

其中 $f(q_i, D)$ 是 q_i 在文档 D 中出现的频率，$|D|$ 是文档 D 的长度，$avgdl$ 是文档集合中文档的平均长度。k_1 和 b 是经验参数，通常 $k_1 \in [1.2, 2.0]$，$b = 0.75$。$IDF(q_i)$ 是逆文档频率，如公式2-4所示。

综合来看，BM25 模型综合考虑了 IDF 因子，文档长度因子，文档词频，和查询词频。概率模型的最大优点是其选择的最优性，通过文档索引系统的信息可以计算文档的相关概率并按照降序排列。其缺点主要是假设了各个单词是相互独立的，实际上每个单词并不是独立的。

2.1.4　语言模型

语言模型是描述自然语言规律的模型。语言模型可以分为传统的文法语言模型和统计语言模型。文法语言模型需要利用语言专家掌握的语言学知识，但是这种模型不能适用于大规模文本处理。因此，产生了基于统计的语言模型，借助于统计语言模型的概率参数，估计单词和句子出现的可能性。

统计语言建模的目的是获得一个语言中的单词序列的联合概率函数，即，已知前面 $(i-1)$ 项的词项序列，计算第 i 个词项是 W 的概率。假设单词是一个句子的最小的结构单位，并假设一个语句 s 由词 $w_1, w_2, \ldots, w_n$ 组成，那么 $p(s)$ 可由公式2-19计算：

$$p(s) = p(w_1)p(w_2|w_1)p(w_3|w_1w_2)\ldots p(w_n|w_1w_2\ldots w_{n-1}) \tag{2-19}$$

$$= \prod_{i=1}^{n} p(w_i|w_i\ldots w_{i-1}) \tag{2-20}$$

由于上式中的参数过多，通常采用近似计算方法。常见的方法有 n-gram 模型方法、决策树方法、最大熵模型方法、最大熵马尔科夫模型方法[43]、条件随机场方法、神经网络方法，等等。

这里我们介绍 n-gram 模型，n-gram 模型也成为 $n-1$ 阶马尔科夫模型，假设一个单词是否出现仅与它前面 $n-1$ 个词有关，也就是 n 元语言模型。因此公式2-19可以近似为：

$$p(s) = p(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} p(w_i|w_{i-n+1}, \ldots, w_{i-1}) \tag{2-21}$$

假设词表的大小为 $100,000$，那么 n-gram 模型的参数数量为 $100,000^n$。n 越大，模型越准确，但是也越复杂，需要的计算量越大。因此考虑到计算可行性，常用的是 $n = 2$ 的 Bi-Gram 模型和 $n = 3$ 的 Tri-Gram 模型。

一般采用极大似然估计（Maximum Likelihood Estimation，MLE）对模型的参数进行估计：

$$p(w_i|w_{i-n+1},\ldots,w_{i-1}) = \frac{c(w_{i-n+1},\ldots,w_{i-1},w_i)}{c(w_{i-n+1},\ldots,w_{i-1})} \tag{2-22}$$

$c(w_{i-n+1},\ldots,w_{i-1},w_i)$ 表示在训练集中出现的次数，训练集越大，参数估计的结果越准确。但即使使用很大的训练集，也会有大量低频词覆盖不到，会出现 $p(w_i|w_{i-1}) = 0$ 的情况，从而导致 $p(s) = 0$，这在模型中无法解释。因此，语言模型经常需要结合数据平滑技术使用，常用的平滑算法有以下几种：

- 加法平滑

  通过加入一个常数 δ，避免概率为 0 的问题：

$$p(w_i|w_{i-n+1},\ldots,w_{i-1}) = \frac{c(w_{i-n+1},\ldots,w_{i-1},w_i)+\delta}{c(w_{i-n+1},\ldots,w_{i-1})+n\delta} \tag{2-23}$$

- 插值平滑

  插值平滑是为了利用不同阶的信息。在 n-gram 语言模型中，n 越大利用的上下文也越多，但也更容易遇到概率为 0 的情况，因此提出的方法是回退到 $n-1$。

$$p(w_i|w_{i-n+1},\ldots,w_{i-1}) = \lambda p(w_i|w_{i-n+1},\ldots,w_{i-1}) + (1-\lambda)p(w_i|w_{i-n+2},\ldots,w_{i-1}) \tag{2-24}$$

- Good-Turing 平滑

  Good-Turing 平滑的思想是对于任何一个出现了 r 次的 $n-gram$，都假定为 r' 次：

$$r' = (r+1)\frac{n_{r+1}}{n_r} \tag{2-25}$$

  其中 n_r 是训练集中发生了 r 次的 $n-gram$ 的数量。则平滑算法如下公式2-26：

$$p(w_i|w_{i-n+1},\ldots,w_{i-1}) = \frac{c'}{c(w_{i-n+1},\ldots,w_{i-1})} \tag{2-26}$$

$$c' = (c(w_{i-n+1},\ldots,w_{i-1},w_i)+1) \times \frac{n(c(w_{i-n+1},\ldots,w_{i-1},w_i)+1)}{n(c(w_{i-n+1},\ldots,w_{i-1},w_i))} \tag{2-27}$$

- Katz 平滑

  Katz 平滑是对 Good-Turing 平滑的扩展，认为对于所有的计数都采用近似

估计的方式并不可靠，采取的方式是当一个 $n-gram$ 的出现次数足够大时，用极大似然进行参数估计；当 $n-gram$ 的出现次数不够大时，采用 Good-Turing 平滑；当 $n-gram$ 的出现次数为 0 时，模型回退。

传统的 n-gram 语言模型在垃圾邮件识别、中文分词、词性标注、机器翻译等领域得到了成功应用。但是 n-gram 由于本身局限性导致计算复杂度大并且只能学习到 n 个词汇的上下文，面对长序列文本建模效果较差。John Lafferty 和 C Zhai[44] 通过改进现有语言模型，基于贝叶斯风险理论，提出了一种结合文档语言模型和查询语言模型的信息检索框架，在性能上获得了较大的提升。

2.1.5　Learning to rank

Learning to rank（LTR）是一种监督学习（Supervised Learning, SL）的排序方法。Learning to rank 已经被广泛应用到文本检索的很多领域，比如信息检索中排序返回的文档，推荐系统中的候选产品，用户排序，机器翻译中排序候选翻译结果等等。传统的排序方法一般是构造相关度函数，然后按照相关度进行排序。实际上影响到相关度的因素有很多，比如上面提到的 TF，IDF 等，但是对于传统的排序方法，很难融合多种信息，比如向量空间模型以 $tf-idf$ 作为权重构建相关度函数，就很难利用其他信息了，于是就有了 Learning to rank。机器学习方法很容易融合多种特征，而且有成熟深厚的理论基础，参数是通过迭代优化出来的，有一套成熟理论解决稀疏、过拟合等问题。

Learning to rank 算法系统框架如图2.1所示：

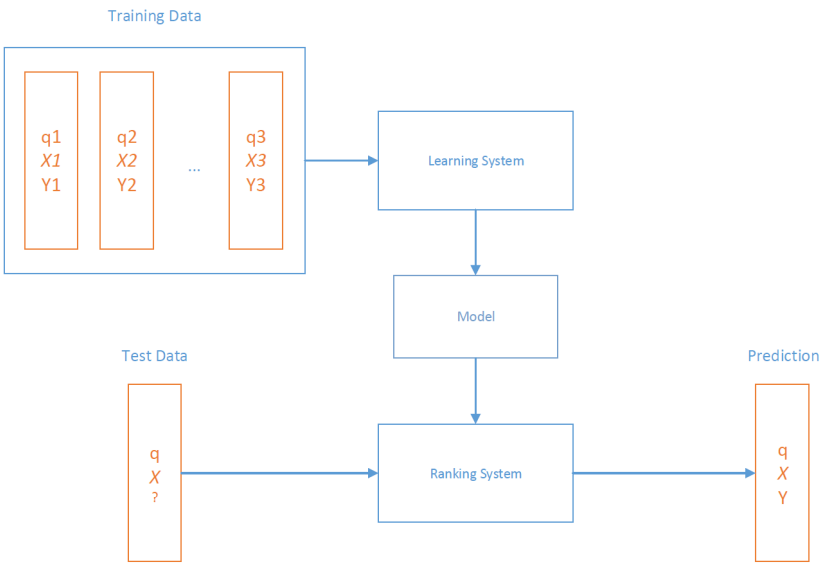

图 2.1　Learning to rank 算法系统框架
Fig 2.1　Learning to rank system framwork

对于已经标注的训练集，选定 Learning to rank 方法，确定损失函数，以最小化损失函数为目标进行优化即可得到排序模型的相关参数，这就是学习过程。预测过程将待预测结果输入学习得到的排序模型中，即可得到结果的相关得分，利用该得分进行排序即可得到待预测结果的最终顺序。

2.2 查询扩展技术

通常查询较短且描述不精确而文档较长，使得查询和文档之间出现词汇不匹配问题，造成语义丢失，无法理解查询意图，这是文本检索领域一个长期存在的问题。查询扩展（QE）通过将用户的查询增加额外的语义和相关的术语，来给原始查询一个更具体的描述和理解，从而增加查询和相关文档之间匹配的可能性。结合上文提到的检索模型使得微博检索取得了更好的成绩。Carpineto et al.[45] 在 2012 年对当时主流的查询扩展技术进行了总结，提供了一个较为全面的查询扩展技术调查报告。

2.2.1 伪反馈查询扩展技术

直接取得用户对查询的扩展是最直接有效的方式，但是这种方式成本很高也不实际，因此研究者们往往通过其他方式获得替代用户的真实反馈的扩展，即伪反馈查询扩展技术 (pseudo-Relevance Feedback)。结合不同的检索模型，伪反馈查询扩展技术已经证明了其有效性[46,47]，由于不需要用户主动参与反馈，减少了交互步骤，使得用户获得了更好的检索体验。

伪相关反馈也存在自身的问题，伪相关反馈一般会根据一个文档集扩展查询，因此查询扩展的结果跟文档集密切相关。如果不能很好的处理文档集上的噪声，保持原始查询的主题不发生偏移，伪反馈查询扩展的最终结果就不是可靠的。实际上这个问题普遍存在，Reliable Information Access（RIA）的相关研究表明[48] 伪相关反馈的性能非常不稳定，受到文档数量、来源、原始查询和扩展的权重等参数的影响。Lee. et al.[49] 研究发现伪查询扩展对文档集中高相关排名的文档敏感，通过聚类筛选出文档集中的核心文档，改进了检索模型的性能。因此合理利用反馈、降低噪声和保持主题是查询扩展研究的重点。

2.2.2 使用相关反馈查询扩展技术

将用户的初次检索结果作为反馈信息，通过提取排序最前面的文档集合中出现词语次数较多的词作为查询词的相似词集合，从而扩展原始查询。这种方法被证明在交互式信息检索中是有效的[50]，但是由于检索本身效果的不稳定，造成扩展效果得不稳定，尤其是由于文档集中存在相关文档较少的情况下很容易引入噪声。

2.2.3　基于外部知识的查询扩展方法

基于外部知识的查询扩展对原始查询进行扩展的过程中引入了外部知识，外部信息较原有语料更丰富，覆盖范围更广，时效性更强，可以为查询提供更多扩展词。较为优秀的外部知识库有 WordNet[51]、维基百科 (wikipedia)[52] 和各种专家知识库[53,54]，这些知识库能很方便的提供一个词的同义词集合。这些外部知识库集合了很多大公司的力量，并且任何人都可以加入进来，贡献自己的知识，创建至今，他们的内容变得越来越丰富和权威。为了方便使用，WordNet 和维基百科不仅提供了离线版本，而且开放了在线 API，可以直接获取信息。但是它们也有一定的局限性，范围宽泛的同时也会在某些领域没有足够的信息，甚至可能会存在噪声。缺乏足够的专业信息意味着构建的扩展模型在相关概念检测中失效，引入噪声可能会造成主题偏移，使扩展失效。相对而言，专家知识库具有足够的专业信息并且噪声少，但是需要我们付出大量的时间和人力来维护，特别是面对当前快速发展的大数据时代，专家知识库需要的成本过高。

2.3　非负矩阵分解方法

非负矩阵分解方法 (Non-negative Matrix Factorization, NMF)[55,56] 是在分解过程中保证矩阵中所有元素均为非负数的矩阵分解方法。非负矩阵分解（NMF）是目前使用较为广泛方法，可以用来解决唯独灾难，聚类等问题，该方法首先将原词项矩阵分解成两个或更多的子矩阵[57]，子矩阵既可以保持原矩阵中的数据特征，又可以完成聚类工作。

非负矩阵分解方法在计算机图形领域应用最早取得成果也最为丰富，在图像的各个子领域均有涉猎，例如，用于发现数据库中的图像特征，构建快速自动识别应用[58]，去除数字水印，图像降噪[59]，图像恢复，图像分割[60]，图像融合，图像分类[61]，图像检索，面部幻觉，面部识别[62] 等。

随着非负矩阵分解方法在图像工程领域不断的成功，有研究者提出可以将非负矩阵分解方法引入到文本检索领域，最初研究者发现非负矩阵分解方法本身矩阵聚类的特性，有研究者已经证明了 NMF 方法与 K-means 方法之间的性能基本等价[63][64–66]。通常情况下，矩阵分解过程中即使矩阵元素全部非负，矩阵分解的结果仍然可能存在负值。这样的矩阵分解结果不具备实际意义，例如在图像或文本检索等领域，出现负值时我们无法做出合理的解释。非负矩阵分解方法解决了负值问题，该方法通过不等式约束项将分解的结果限制为非负矩阵，使得矩阵分解具有了可解释性，例如在词项文档矩阵中正数恰好可以表示词频或逆文档频率[67]，这些特征仅在正数范围内有意义，并且在分解的过程中解决了原始矩阵的稀疏性问题，这使得非负矩阵分解技术在文本聚类领域具有更好

的适用性。在 NMF 方法在聚类工作中取得了成功之后，研究者们也越来越多的将该方法应用于文本检索的各个领域之中。

在其他应用领域中，非负矩阵方法也有很多成功应用，例如音频模式分离[68]，音乐流派分类[69]，语音识别，微阵列分析，盲源分离[70]，光谱学[71]，基因表达分类[72]，细胞分析，EEG 信号处理[59]，病理诊断，电子邮件监控[73]，在线讨论参与预测，网络安全，自动个性化总结，大气分析化合物识别[74]，地震预测，股票市场定价[75] 等。

本节将对 NMF 进行简单的介绍，之后将对 NMF 的一部分扩展方法进行介绍。

2.3.1 基本非负矩阵分解方法

基本非负矩阵分解方法的基本思想原理可以用图2.2表示。NMF 方法通过将原始矩阵 W 转换为两个子矩阵 U 和 V 相乘的形式，以达到矩阵分解的目的，即将原始矩阵 W 分解为子矩阵 U 和 V。矩阵 W 是一个高维稀疏矩阵，对于 W 中的列向量来说，低秩矩阵 U 被称为基矩阵，低秩矩阵 V 被称为权重矩阵，也叫做系数矩阵。为了便于说明，假设非负矩阵 U 的列向量为某一个低维空间的一组基向量，那么非负矩阵 V 中的列向量就是高维稀疏矩阵 W 的列向量在这一低维空间上的投影，这一假设在一定程度上也反映了“局部构成整体”的概念，符合人们对事物的认知过程。

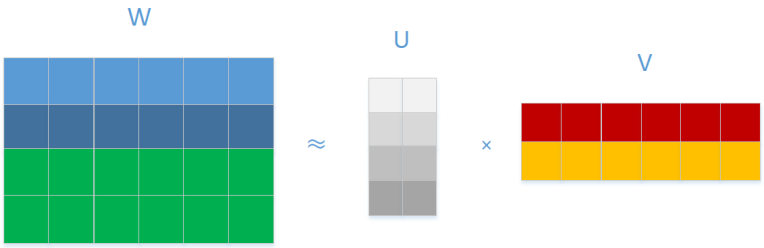

图 2.2　非负矩阵分解过程
Fig 2.2　Process of non-negative matrix factorization

在求解 U 和 V 过程中，可能存在多个解，为了得到最优解，通常需要在某些约束下，使得 $W - U \cdot V$ 误差最小，即 $W \approx U \cdot V$，这样就可以将一个稀疏的高纬矩阵分解为两个相对稠密的矩阵相乘的形式了，即公式2-28，其中 $\mathbf{W}||\mathbf{UV}$ 表示 W 分解为 U 和 V 两个子矩阵。

$$D_F(\mathbf{W}||\mathbf{UV}) = \frac{1}{2}||\mathbf{W} - \mathbf{UV}^T||_F^2 \tag{2-28}$$

2.3.2　正则化非负矩阵分解方法

针对非负矩阵分解方法的求解问题，即目标函数在求解过程中可能存在着多个解，这也是最优化问题中普遍存在的问题。通常情况下，最优化问题通过梯度下降方法的求解结果只是目标函数的解之一，但该函数可能仍然存在着其他的最优解，为了解决这一问题，我们通过利用正则化约束条件来找到我们希望求解的最优解。

在基本非负矩阵分解方法中，求解目标的约束条件只有非负约束，也就是两分解矩阵 $U \geq 0$，$V \geq 0$，这显然无法使目标函数获得唯一的最优解，因此我们需要对分解矩阵 U 和 V 继续引入更多的约束条件。不仅如此，通过这些约束条件，可以方便的引入外部信息，以更具体、更全面的方式反映出问题的特征，并且获得唯一的最优解。如公式2-29所示, 其中 αJ_1 和 βJ_2 表示正则约束项，α,β 是平衡优化目标与各适惩罚项之间的规则化参数。

$$D_F(\mathbf{W}||\mathbf{UV}) = \frac{1}{2}||\mathbf{W} - \mathbf{UV}^T||_F^2 + \alpha J_1(\mathbf{U}) + \beta J_2(\mathbf{V}) \tag{2-29}$$

正则约束项的引入一般分为以下四类：

- 稀疏约束

  稀疏约束的非负矩阵分解方法应用广泛，应用矩阵的稀疏特征作为约束。
- 正交约束

  正交约束限定分解的子矩阵 U 或 V 是正交的。在矩阵元素非负的约束下，正交必然会导致稀疏。因此，可以将正交约束视为稀疏约束的特殊情况，这稀疏约束和正交约束之间的优化模型有一定的差异。
- 判别约束

  判别约束是将分类作为外部信息，从模式识别的角度来看，基础非负矩阵分解方法可以被视为无监督学习，通过将判别信息与分解耦合，将生成模型和分类任务统一建模，可以应用于基于分类的任务中。
- 拓扑约束

  拓扑约束主要在分解过程中保持原始矩阵中固有的几何结构，可以增强学习性能。应用原始矩阵中的几何信息作为约束。

标准的非负矩阵分解方法在求解过程中，单纯的通过一定的组合来近似的还原原矩阵，这样的做法很难表示出原数据矩阵中的内在结构，并且当面对较大的训练数据集时，问题的解会变得不稳定。因此在目标函数的基础上添加约束，如稀疏约束、正交约束、判别约束、拓扑约束等正则约束项，对矩阵稀疏程度和元素分布进行控制，这样在优化过程中既考虑了问题特有的先验信息，又将问题的解限定在特定空间中。

因此，将问题的优化公式表示为 $w_L = argmin f(w) + \gamma g(w)$，其中，$f(w)$ 为损失函数，$g(w)$ 为正则约束项，表示对解的范围进行约束。常用的正则化约束项为 L_p 范数的形式，主要分为以下三种：

(1) L_0 范数表示矩阵中非 0 元素的个数，通过 L_0 范数可以约束矩阵的稀疏度，能够过滤一些噪声，因此 L_0 范数还可以用来做特征选择，但是求解较为困难。

(2) L_1 范数表示矩阵中所有元素的绝对值之和。L_1 范数是 L_0 范数的最优凸近似，求解复杂度也低于求解 L_0 范数的复杂度，通常使用 L_1 近似代替 L_0 作为约束项。

(3) L_2 范数表示矩阵中所有元素平方和的平方根，L_2 范数可以使目标矩阵中的每个元素都接近于 0 而不等于 0，可以解决过拟合问题。L_2 范数为凸函数，降低简化计算复杂度的同时近似保证矩阵的稀疏性和解的稳定性。

2.3.3 结构化非负矩阵分解方法

与正则化非负矩阵分解方法不同，结构化非负矩阵分解方法是通过优化目标项体现先验知识的，通常直接修改优化公式而不是引入额外的惩罚项作为优化约束。具体形式如公式2-30。

$$\mathbf{W} \approx \mathbf{F}(\mathbf{UV}) \tag{2-30}$$

通常优化函数 F 有以下三种：

- 通过在目标函数中增加权重矩阵，可以解决非负矩阵分解权重统一，不能体现各自信息的重要程度和噪声多的问题。
- 为了描述时间关系，在分解过程中引入卷积生成模型，可以在矩阵分解过程中保持原始矩阵的时间空间特性。
- 将原始矩阵分解为三个子矩阵相乘的形式，即 $W \approx U \cdot S \cdot V$。

2.3.4 广义非负矩阵分解方法

广义非负矩阵分解方法与以上方法不同，是非负矩阵分解方法的理论延伸，一般的可以将广义非负矩阵分解方法分为以下几种：

- 半约束非负矩阵分解方法

  非负矩阵分解方法将数原始矩阵 W 中的每个元素限制为非负数，当不对 W 中所有元素作约束时，分解过程中就会有负数出现。但是半约束非负矩阵分解方法仅约束子矩阵 V 中所有元素为非负，而不约束子矩阵 U 中的元

素。这种形式在针对某些实际问题时是有意义的，因为数据在某些情况下并不总是非负的，因此分解后的潜在特征维度也需要反映相应信息的情况。

- 非负张量分解方法

  Welling 和 Weber[76] 基于非负矩阵分解提出了非负张量分解 (Positive Tensor Factorization) 的概念。非负矩阵分解方法通常将多维度数据组织成二维矩阵，这样做虽然会简化建模过程，但是在面对复杂问题时会损失数据的真实性。因此将二维矩阵分解方法扩展为多维张量分解。

- 核非负矩阵分解方法

  非负矩阵分解方法及其扩展方法都是线性模型，如果存在非线性结构，则这些方法无法有效提取到这部分信息，这限制了非负矩阵分解方法的应用范围。为了解决这个问题，提出核矩阵分解方法，在模型中引入核函数将原始数据映射到高维空间。Zhang[77] 和 Buciu[78] 的研究已经证实了这种思路的有效性。

2.4　不等式约束的非线性优化

KKT 条件（Karush-Kuhn-Tucker Conditions, KKT）[79] 是解决不等式约束的非线性优化问题经常用到的一种方法。这里的非线性优化问题通常是指给定目标函数，求在定义域上的全局最小值。在使用非负矩阵分解方法及其扩展方法时，通常需要使用 KKT 条件求解。

非线性优化问题通常分为以下三类：

1. 无约束条件

   无约束优化问题的解决方法比较简单，通常是将目标函数对变量求导，而后令导函数为 0 的解可能是问题的解，最后将结果进行验证即可。

2. 等式约束条件

   等式约束问题形式如公式2-31所示，其中 $f(x)$ 为目标函数，$h_i(x)$ 为等式约束，在解决这一类优化问题时，我们通常使用的方法是拉格朗日乘子法（Lagrange Multiplier）。

$$\begin{aligned} &\min f(x) \\ &s.t. \quad h_i(x) = 0 \qquad i = 1, 2, \ldots, n \end{aligned} \tag{2-31}$$

拉格朗日乘子法将等式约束 $h_i(x)$ 并入目标函数，称为拉格朗日函数，如公式2-32所示，其中系数 λ_i 称为拉格朗日乘子，且 $\lambda_i \neq 0$。将拉格朗日函

数对各个变量求导，令导函数为零后求解，然后验证得最优解。

$$L(x,\lambda_i)=f(x)+\sum_{i=1}^{n}\lambda_i h_i(x) \tag{2-32}$$

3. 不等式约束条件

含有不等式约束的优化问题，通常形式如公式2-33所示，除等式约束 $h_i(x)$ 外，存在着多个不等式约束 $g_j(x)$。

$$\begin{aligned}&\min f(x)\\ &s.t.\quad h_i(x)=0 \qquad i=1,2,\ldots,n\\ &\qquad g_j(x)\le 0 \qquad j=1,2,\ldots,n\end{aligned} \tag{2-33}$$

$$L\left(x,\lambda_i,\mu_j\right)=f(x)+\sum_{i=1}^{n}\lambda_i h_i(x)+\sum_{j=1}^{n}\mu_j g_j(x) \tag{2-34}$$

这类问题通常使用 KKT 条件求解。首先求得拉格朗日函数，如公式2-34所示，KKT 条件是指，该公式的求解过程中需要满足以下条件：

(a) $L\left(x,\lambda_i,\mu_j\right)$ 对 x 求导为 0

(b) $h_i(x)=0$

(c) $\mu_j g_j(x)=0$

在满足上述三个条件的情况下求解，可得到函数的最优解。由于 $g_j(x)\le 0$，因此可以确定，此时应满足条件 $\mu_j=0$ 或 $g_j(x)=0$。

2.5 本章小结

本章综述了微博检索和非负矩阵分解的理论知识，并对非负矩阵分解方法求解时涉及的不等式约束的非线性优化问题做了简单的介绍，在后序的章节中，我们将对我们提出的微博检索算法进行介绍，并通过实验验证算法的性能。

第 3 章　基于聚类约束的高质量微博检索方法研究

在微博检索系统中，通常都可以检索得到相关微博，但是在推送给用户时，面临着推送选择的问题：如何选择相关微博，才能高效的利用有限的推送列表，达到用户满意的效果，即信息相关度高且不重复，则推送质量较高。在本章中，将对基于聚类约束的高质量微博检索方法进行详细的解释。首先，将介绍本文的问题设定和检索系统框架，然后对基于非负矩阵分解的微博聚类方法进行介绍，最后，将对聚类约束下的高质量微博检索方法进行分析，并给出求解过程。

3.1　问题设定

本文背景主要基于 Text REtrieval Conference（TREC）微博检索会议，该会议诞生于 1992 年，是微博检索领域最为权威的国际测评会议，由美国国家技术标准局 (NIST) 和美国国防部 (DOD) 联合组织举办，旨在交流信息检索领域内相关的技术并对这些技术在统一的标准平台上进行评测。由于 TREC 所设置的所有任务很接近实际问题，因此在 TREC 中所运用的检索技术都有重要的实用价值。

本文研究以 TREC2016 年微博检索任务为背景，并根据官方数据集对多种方法进行测评。微博检索任务如图3.1所示，参与者将在 Twitter 官方 API 接收实时数据，为期 10 天（2016 年 8 月 2 日至 2016 年 8 月 11 日，UTC），并依照用户兴趣信息（类似信息检索中的查询词，代表用户的信息需求）检索相关微博，并具体分为以下两个任务：

(1) 实时微博推荐。将根据用户兴趣信息进行相关微博推荐，并及时推送给手机端用户。

(2) 邮件列表微博推荐。根据用户兴趣信息生成邮件微博列表，以天为单位检索出 100 条相关内容后进行推送。

在此基础上，定义用户兴趣信息，即原始查询为 Q，微博数据集为 C，微博数据集对应的词项文档矩阵为 W。接下来，将以邮件列表微博推荐任务展开讨论。

3.2　检索系统框架

根据问题设定，构建了基于聚类约束的高质量微博检索系统，如图3.2所示，该框架主要包括以下几个部分：

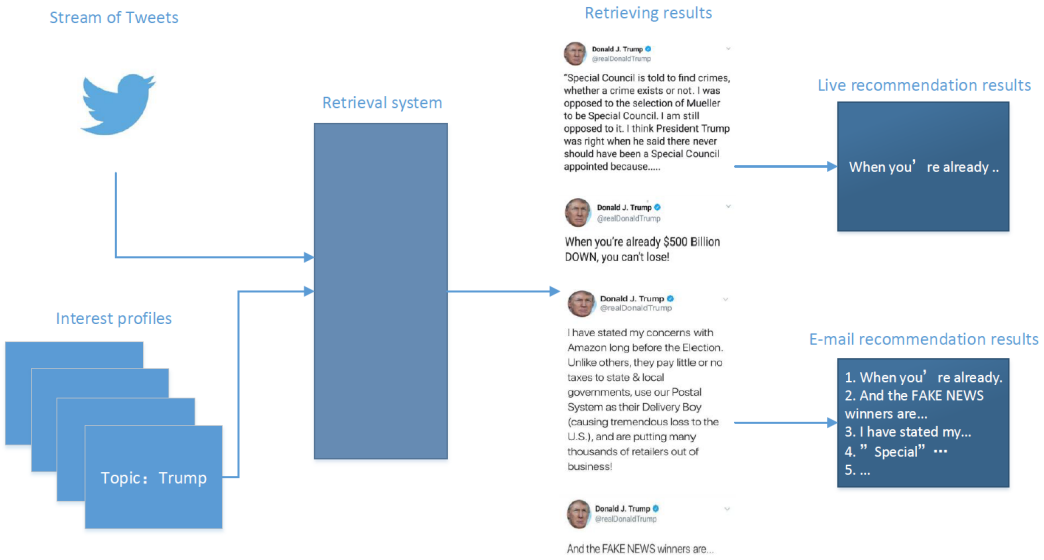

图 3.1　微博检索任务示意图
Fig 3.1　Microblogging retrieval task map

3.2.1　文本预处理模块

微博文本较短但内容形式复杂多样，在对微博进行检索之前要先对微博文本做预处理，去掉无意义的部分，并得到每条微博的关键词向量，在后续的检索中，将使用关键词向量进行检索。包括以下几个步骤：

(1) 过滤非英文微博，仅对英文内容做检索。

(2) 将微博中的大写字母全部转换为小写，并进行词干还原。

(3) 去掉停用词、数字、表情符号等无法识别的内容。

(4) 过滤掉词数小于 3 的微博，这些微博通常不包含实际的内容。

3.2.2　查询扩展模块

查询和微博之间的文本不匹配问题一直存在，例如，查询“苹果”，而微博内容中以"iPhone" 为关键字，这样的情况下相关微博无法被正确检索。因此，为了解决查询和微博之间文本不匹配的问题，引入了查询扩展技术。在这里选用了基于谷歌搜索和维基百科的查询扩展。将用户兴趣信息的标题和描述，分别作为查询项使用谷歌搜索 API、维基百科 API 进行检索，将返回结果的前 20 条或前 50 条作为扩展文档，同时为了正确匹配文本内容，同样使用上文的文本预处理模块对扩展文档进行处理，最后使用 TF-IDF2-5作为关键词权值，提取权值前 10 的关键词作为用户兴趣信息的扩展词。具体分为以下几种扩展方式：

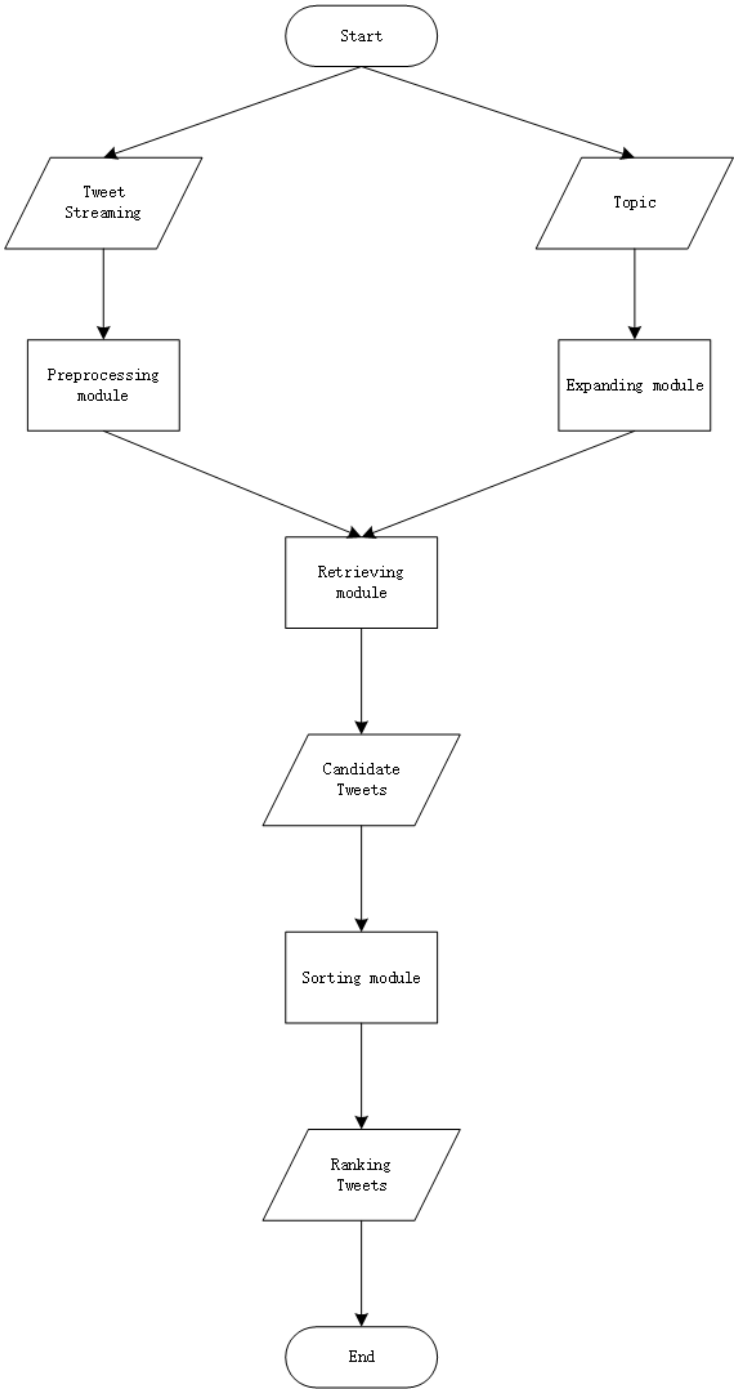

图 3.2　基于聚类约束的高质量微博检索系统框架

Fig 3.2　High quality microblogging retrieval system framework based on clustering constraints

3.2.3　文本检索模块

在检索模块中，本文使用了 BM25 概率检索模型，具体计算方法见2-18。BM25 概率检索模型根据每条文本中包含的查询词项来为文本相关度评分，因此也是一个排序方法。在获得扩展词后，结合原始查询，应用 BM25 概率检索模型，得到每条文本的相关度评分，在根据相关度评分排序后，可以得到初步的检索结果。

表 3.1 查询扩展方式
Tab. 3.1 Kinds of query expand

名称	扩展方式	具体描述
Timeliness top50 web extend	时效性网页内容扩展	采用评测前一天（2016 年 8 月 1 日）的谷歌网页前 50 条搜索结果作为扩展文档
Timeliness top50 news extend	时效性新闻内容扩展	采用评测前一天（2016 年 8 月 1 日）的谷歌新闻前 50 条搜索结果作为扩展文档
Non-Timeliness top50 web extend	网页内容扩展	采用谷歌网页前 50 条搜索结果作为扩展文档，不具备时效性
Non-Timeliness top50 news extend	新闻内容扩展	采用谷歌新闻前 50 条搜索结果作为扩展文档，不具备时效性
Non-Timeliness top20 web extend	网页内容扩展	采用谷歌网页前 20 条搜索结果作为扩展文档，不具备时效性
Non-Timeliness top20 news extend	新闻内容扩展	采用谷歌新闻前 20 条搜索结果作为扩展文档，不具备时效性
Non-Timeliness wiki extend	维基百科内容扩展	采用维基百科检索结果作为扩展文档，不具备时效性

BM25 检索模型是一个词袋检索方法，没有考虑文本中的查询词项之间的相互关系 (例如, 它们的相似度)，也没有考虑相似文本对结果的影响，因此在检索得到的相关微博中，会存在大量的高相似的微博，因此需要对这些相似微博进行处理，避免冗余微博过多的问题。

在这里本文采用 simhash[80] 算法做文本去重。simhash 算法是一种相似度计算算法，simhash 原理如图3.3，算法流程如下：

(1) 将 Doc 文档进行关键词抽取 (其中包括分词和计算权重)，抽取出 n 个 (关键词，权重) 对，即图中的 (word, weight)。

(2) 根据 hash 函数生成图中的 (hash,weight)，即图中的（110110，w_1）

(3) 哈希权值对进行位的纵向累加，如果该位是 1，则 +weight, 如果是 0，则-weight，最后所有结果累加，如图所示是 [3, 28, -2, 32, -32, 44]。这里产生的值和 hash 函数所用的算法相关。

(4) 最后将 [3, 28, -2, 32, -32, 44] 转化为 110101，正数为 1 负数为 0。

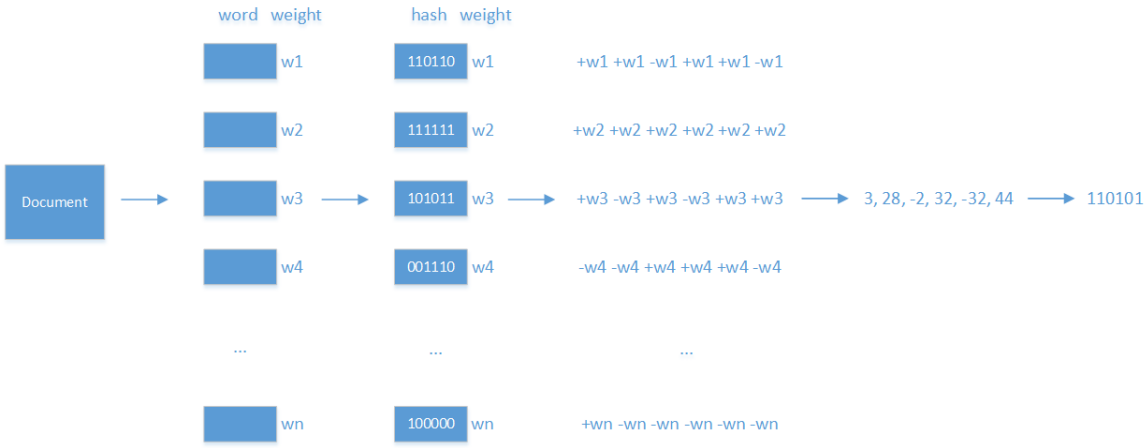

图 3.3　simhash 方法原理图
Fig 3.3　Simhash map

最后要计算两篇微博之间的海明距离，即可度量微博之间的相似度，通过这种方法来去掉相似度过高的微博。

3.2.4　微博排序模块

通常，检索得到的相关微博仅按检索相关程度排序，并不能满足所有用户的需要。因为在检索过程中，检索系统需要具体化用户的检索意图，即将用户的检索意图转化为可以处理的词向量；并且，检索系统的也不能完备的检索出所有符合检索需求的文本，即存在相关微博但检索失效。在这两个过程中，检索结果的准确率和召回率都会受到影响。基于这样的原因，考虑通过文本聚类，探究文本中潜在的事件，并根据事件对文本相关程度的影响对文本进行重新度量，具体内容在下文展开讨论。

3.3　基于聚类约束的微博检索方法研究

在对检索得到的相关微博结果进行排序的过程中，既要保证推送结果的相关性，又需要提高信息量去掉冗余信息，在这里，假设用户的查询结果中包含多个事件，依据事件对微博结果进行排序，属于高相关事件的微博取得较高的排名，同一事件中的重复微博取得较低的排名，即希望得到代表关键信息的微博，而去掉包含冗余信息的微博。

为了得到隐含的事件情况，考虑对微博文本内容进行聚类，通过聚类结果的每个文档类簇代表每个事件。同时微博整体数量较大、文本短和内容多样的特点，微博数据维度高、稀疏度高，因此使用非负矩阵分解方法解决维度过高的问

题，同时对微博文本进行聚类。最后，利用聚类信息对检索得到的相关微博进行排序。

基于这样的思路，本节中使用非负矩阵分解方法提取微博聚类信息。如下公式3-1所示，对于一个特定的查询 Q，将在发生查询前一段时间内所有微博作为检索候选集合记作 $\mathbf{C}$，用矩阵 W 代表 C 对应的词项文档矩阵，用 Clu 代表微博文档类簇。这里的目标将 $\mathbf{W}$ 分解为大小为 $m \times k$ 的非负矩阵 $\mathbf{U}$ 和大小为 $k \times n$ 的非负矩阵 $\mathbf{V}$ 相乘的形式，也就是最小化以下目标函数:

$$\min_{\mathbf{U},\mathbf{V}} ||\mathbf{W} - \mathbf{U}\mathbf{V}^T||_F^2 \tag{3-1}$$

公式 (3-1) 中 $||*||_F$ 表示矩阵的 Frobenius 范数，$\mathbf{U}$ 为系数矩阵，包含了矩阵分解后每条微博的类别信息，$\mathbf{V}$ 为基矩阵，包含了矩阵分解后的文档类簇信息。因此，微博的聚类信息可以通过求解 $\mathbf{U}$ 和 $\mathbf{V}$ 得到，用下式表示:

$$\min F = ||W - UV^T||_F^2 + \alpha||U||_F^2 + \beta||V||_F^2 \tag{3-2}$$

$$s.t. U \geq 0, V \geq 0 \tag{3-3}$$

基于非负矩阵分解方法获得微博聚类信息，通常，每个查询下都会包含多个文档类簇，每个文档类簇中包含多条微博，并且同一文档类簇中的微博共同代表了查询中的某一方面的信息。对每条微博来说，它既构成查询下的某种信息，也在文档类簇中发挥作用，代表文档类簇中的某种信息。认为当一个文档类簇代表的信息比另一个文档类簇代表的信息对查询更重要时，那么该文档类簇中的微博也应当比另一个文档类簇中的微博具有更高的价值。因此，认为当一条微博与查询紧密相关，并且它所在的文档类簇也代表了查询中较重要的信息，则这条微博应当是当前查询下的高质量微博。

基于这样的思路，首先需要表示每个文档类簇。采用以下两种方式表示文档类簇:

(1) 通常文档类簇由多条微博共同构成，因此，可以将这些微博构成的文档类簇作为一个整体看待。即将文档类簇内所有微博整合为一个长文本，长文本代表文档类簇。

(2) 在非负矩阵分解结果中，基矩阵 $\mathbf{V}$ 包含了各个文档类簇的信息。即将基向量作为不同类簇中心，基矩阵代表所有文档类簇。

然后，为了衡量不同文档类簇的作用，将提取的文档类簇在检索模块内进行检索，可以得到每个文档类簇的相关性关系，考虑通过类簇的相关性来补充检索相关性。

同时基于这种类簇相关性补充检索相关性的思路，定义原始查询的扩展文档为 E，通常 E 能较全面的描述查询。因此将扩展文档整合为一个长文本，利用长文本代表查询类簇，通过查询类簇与微博文本的相关性补充检索相关性。

最后，基于以上内容，本文提出以下多元检索模型，如图3.4所示：

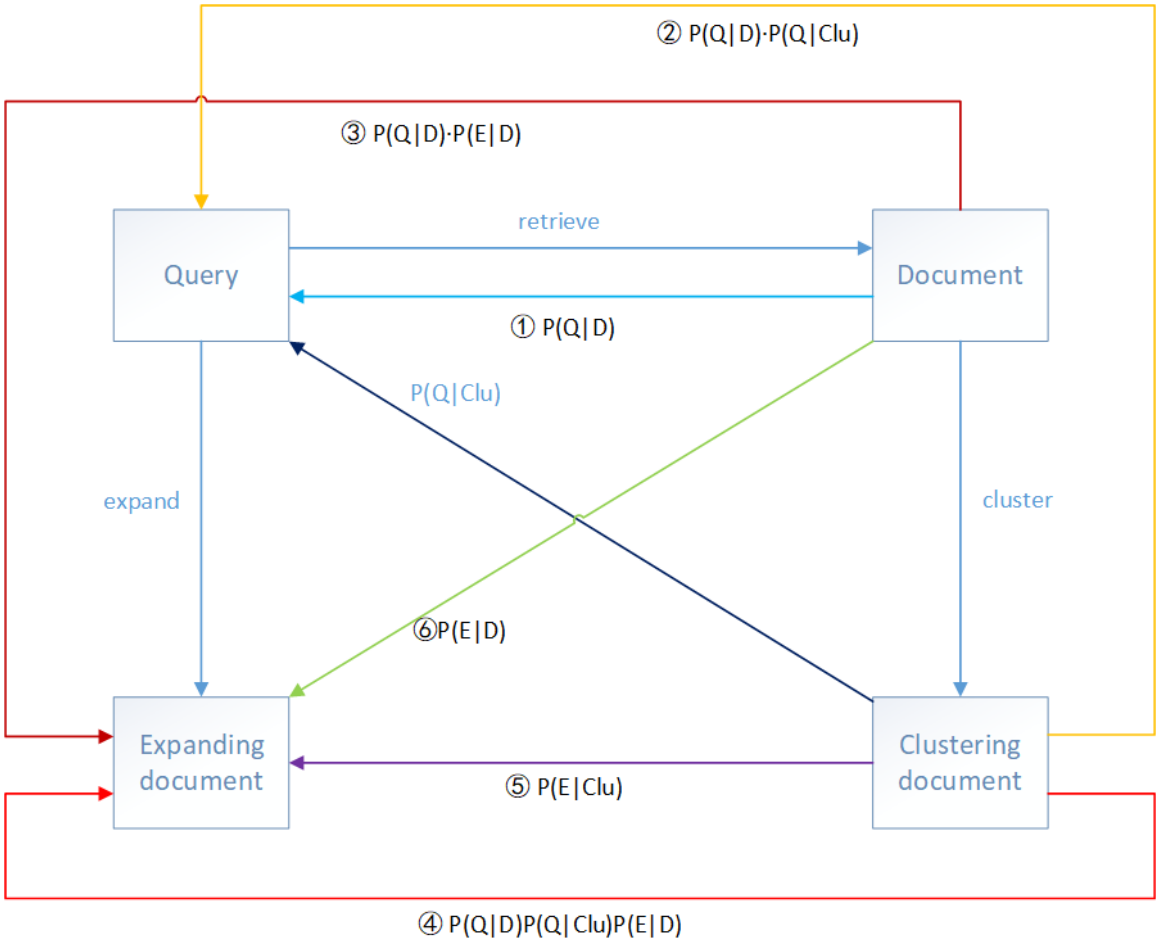

图 3.4　微博多元检索模型

Fig 3.4　Multiple retrieval model of Microblogging

其中：

1. $P(Q|D)$ 表示微博 D 属于查询 Q 下的相关微博的概率。
2. $P(Q|D)P(Q|Clu)$ 表示在聚类约束中，微博 D 属于查询 Q 下的相关微博的概率，Clu 代表文档所属于的类簇。
3. $P(Q|D)P(E|D)$ 表示在扩展文档约束中，微博 D 属于查询 Q 下的相关微博的概率，$P(E|D)$ 表示微博 D 属于查询扩展文档 E 下的相关微博的概率。
4. $P(Q|D)P(Q|Clu)P(E|D)$ 表示在聚类和扩展文档的共同约束下，微博 D 属于查询 Q 下的相关微博的概率。
5. $P(E|Clu)$ 表示类簇 Clu 属于扩展文档 E 下的相关微博的概率。
6. $P(E|D)$ 表示微博 D 属于扩展文档 E 下的相关微博的概率。

3.4 基于聚类约束的高质量微博检索方法求解

根据上文中所述，求解的目标函数是一个包含不等式约束的最优化问题，因此将利用 KKT 条件对公式进行求解。首先需要将目标函数转换为矩阵的迹的形式。

$$F = Tr(-2W^T UV^T + VU^T UV^T) + \alpha Tr(UU^T) + \beta Tr(VV^T) \tag{3-4}$$

$$s.t. U \geq 0, V \geq 0 \tag{3-5}$$

根据定理，F-范数可以转换为矩阵的迹的形式，转换公式如3-4所示。将 F-范数转换为矩阵的迹的形式的原因在于，F-范数无法进行求导。接下来，通过 KKT 条件对上述公式中的未知变量 U、V 进行求解，首先需要求解其偏微分导数，其结果如下：

$$\frac{\partial F}{\partial U} = -2WV + 2UV^T V + 2\alpha U \tag{3-6}$$

$$\frac{\partial F}{\partial H} = -2W^T U + 2V^T U + 2\beta V \tag{3-7}$$

$$\tag{3-8}$$

在满足 KKT 条件的情况下，可以得到以下等式：

$$-WV + UV^T V + \alpha U = 0 \tag{3-9}$$

$$-W^T U + V^T U + \beta V = 0 \tag{3-10}$$

$$\tag{3-11}$$

根据上述等式可以得到迭代更新公式：

$$U(i,k) \leftarrow U(i,k)\sqrt{\frac{WU}{UV^T V + \alpha U}} \tag{3-12}$$

$$V(i,k) \leftarrow V(i,k)\sqrt{\frac{W^T U}{VU^T U + \beta V}} \tag{3-13}$$

至此，就可以对分解项 U 和 V 进行求解了。在后面的章节中，将对算法的求解过程进行详细的说明，以便读者对算法有一个更加明确的认识。

3.5　算法伪代码

本节将对基于聚类约束的高质量微博检索算法过程进行说明。在算法过程中，将微博聚类结果矩阵 U 所代表的信息融入微博检索排序中，除此之外，还有查询扩展文档信息。首先，要对待求解的矩阵 U 和 V 进行初始化，之后开始进行迭代工作，这里迭代应直到收敛结束，但由于到后期迭代时间消耗很长，且影响很小，因此在实际工作中，常迭代固定次数，如 3000 次。

Algorithm 1: 基于聚类约束的高质量微博检索方法

Data: $\mathbf{W}, \alpha_n, \beta_n,$

Result: 基于聚类约束的检索值 $P'(Q|D)_n$

begin

基于检索模型计算每篇微博的检索值 $P(Q|D)_n$

随机初始化 $\mathbf{U}_n, \mathbf{V}$

while 反复更新直至收敛 **do**

计算如下算式

$\mathbf{T1} = \mathbf{WU}$

$\mathbf{T2} = \mathbf{UV}^T\mathbf{V} + \alpha\mathbf{U}$

$\mathbf{T3} = \mathbf{W}^T\mathbf{U}$

$\mathbf{T4} = \mathbf{VU}^T\mathbf{U} + \beta\mathbf{V}$

更新 $\mathbf{U}(i,j) \leftarrow \mathbf{U}(i,j)\sqrt{\frac{\mathbf{T1}(i,j)}{\mathbf{T2}(i,j)}}$

更新 $\mathbf{V}(i,j) \leftarrow \mathbf{V}(i,j)\sqrt{\frac{\mathbf{T3}(i,j)}{\mathbf{T4}(i,j)}}$

根据矩阵 $\mathbf{U}_n$，得到聚类结果 $\mathbf{clu}_n$

整合类簇 $\mathbf{clu}_k$，计算类簇检索值 $P(Q|Clu)_k$

根据查询扩展文档，计算检索值 $P(E|D)_n$

根据图3.4计算检索值 $P'(Q|D)_n$

返回 $P'(Q|D)_n$.

3.6　本章小结

本章着重对本文研究的基于聚类约束的高质量微博检索方法进行了详细的介绍，首先介绍了检索系统框架，然后介绍了基于非负矩阵分解的微博聚类方法，最后给出结合聚类约束的微博检索方法及求解过程。在之后的章节中，将利用实验验证该方法的性能，并得出结论。

第 4 章　基于相关约束的微博聚类方法研究

在本章中，将对基于相关约束的微博聚类方法进行详细的解释。首先，将对第三章中的微博聚类方法进行扩展，引入额外的先验信息，期望改进聚类算法的性能，然后，对基于相关约束的微博聚类方法进行分析，最后给出求解过程。

4.1　基于非负矩阵分解的微博聚类

在第三章中，已经介绍了应用非负矩阵分解对微博聚类的方法，在本章中，将对该方法进行扩展。如下公式4-1所示，对于一个时间段内所有的微博集记作 $\mathbf{W}$，目标时将 $\mathbf{W}$ 分解为大小为 $m \times k$ 的非负矩阵 $\mathbf{U}$ 和大小为 $k \times n$ 的非负矩阵 $\mathbf{V}$ 相乘的形式，也就是最小化以下目标函数:

$$\min_{\mathbf{U},\mathbf{V}} ||\mathbf{W} - \mathbf{U}\mathbf{V}^T||_F^2 \tag{4-1}$$

公式4-1中 $||*||_F$ 表示矩阵的 L_2 范数。W 中的行向量是每条微博的词向量，因此对于矩阵 W 中的行向量，矩阵 U 为系数矩阵，矩阵 V 为基矩阵。在矩阵 U 中的行向量表示每条微博在矩阵 V 为基矩阵的特征空间中的投影，因此求得系数矩阵 U，就能得到微博聚类信息，具体优化公式如下4-2所示。

$$\min F = ||W - UV^T||_F^2 + \alpha||U||_F^2 + \beta||V||_F^2 \tag{4-2}$$

$$s.t. U \geq 0, V \geq 0 \tag{4-3}$$

4.2　正则化方法在非负矩阵分解中的应用

如第二章所说，非负矩阵分解方法在分解过程中常常存在有多个局部最小值的情况，通常采用添加正则项来解决。Jiliang Tang. et al.[81] 提出可以利用社会网络同质性构建正则项来提升非负矩阵分解方法的性能，并在实验中证明了正则化非负矩阵分解方法性能更优，本文将利用该方法对算法性能进行提升。

在正则化非负矩阵分解方法中，需要在公式4-1的基础上添加正则项。为了便于说明，以微博聚类为例，微博词向矩阵 W 可以分解成 U 和 V 两矩阵，其中 U 矩阵即微博的聚类矩阵，当具有一些先验信息，假设存在聚类信息矩阵 I 时，则可利用 I 矩阵去限制 U 矩阵，在分解过程中保持矩阵 U 的特征。因此，可将

公式4-2扩展如下公式4-4。

$$\min_{U,V} F = \|W - UV\|_F^2 + \alpha \|U - I\|_F^2 \tag{4-4}$$

在公式4-2中，$\alpha \|U - I\|_F^2$ 即为正则项，其中 α 为参数，表示正则项的权重。在正则项的约束下，分解矩阵 U 就会向聚类信息矩阵 I 进行逼近，因此保持了矩阵 U 的特征。

当然，为了便于说明，将先验信息直接定义为聚类信息，并具体设为矩阵 I，在实际应用中，并不存在这样直观的情况，要根据具体应用场景调整正则项。

4.3 基于相关约束的微博聚类方法

在本节中，将根据前文所述，详细介绍基于相关约束的微博聚类方法研究。在聚类方法中，总是希望相同类别的样本能够足够聚拢，而不同类别的样本能够足够分散，即类内离散度小，类间离散度大。应用非负矩阵分解方法做聚类，也希望能达到这样的效果。

在第三章中，使用了 BM25 概率检索模型，来计算每条微博的检索相关度，该模型基于词袋，考虑了包含哪些单词而不考虑词之间的关系，因此，假设检索相关度相近的微博，在聚类中具有相同的类别。基于这样的假设，引入相关约束正则项，期望能提高非负矩阵分解方法的性能。

如图4.1所示，首先，定义了一条微博 u_i 和另一条微博 u_j，u_i 和 u_j 之间的相关约束系数 $\varepsilon(i,j)$ 满足以下三个条件[81]：

(1) $\varepsilon(i,j) \in [0,1]$

(2) $\varepsilon(i,j) = \varepsilon(j,i)$

(3) $\varepsilon(i,j)$ 的值越大，则表示 u_i 与 u_j 之间的相关约束越大

然后，利用公式，根据相关约束系数添加相关约束正则项，其中 $U(i,:)$ 代表微博词向矩阵中的行向量。

$$\min \sum_{i=1}^{n} \sum_{j=1}^{n} \varepsilon(i,j) \|U(i,:) - U(j,:)\|_F^2 \tag{4-5}$$

在向量空间中，当 $\varepsilon(i,j)$ 越大，则意味着 u_i 和 u_j 两条微博词向量之间的距离越近，反之 $\varepsilon(i,j)$ 越小就意味着 u_i 和 u_j 两向量之间的距离越远。

进一步对公式4-5进行推导，便可得到如下4-6结果。

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \varepsilon(i,j) \|U(i,:) - U(j,:)\|_F^2$$

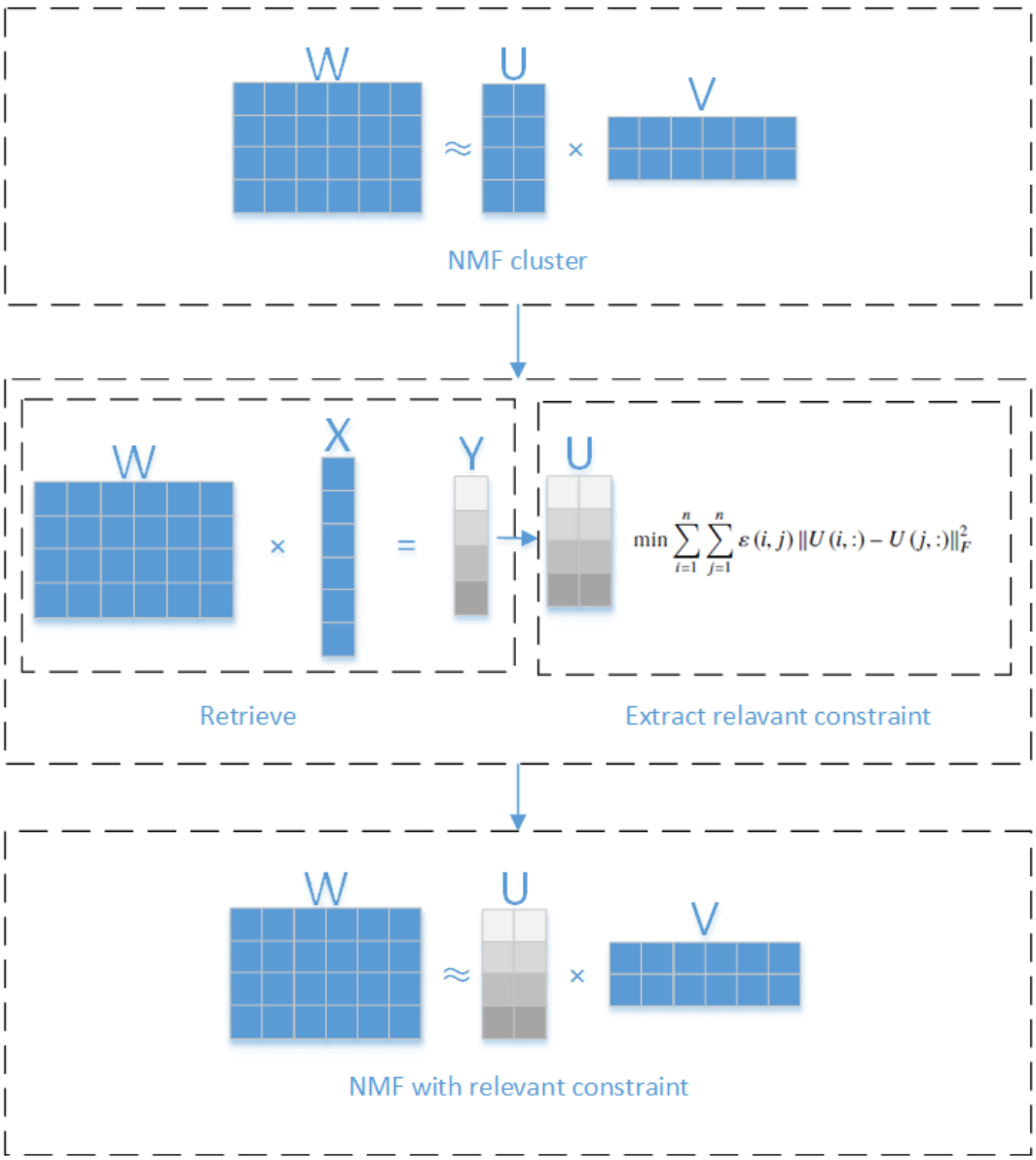

图 4.1　基于相关约束的微博聚类方法

Fig 4.1　Microblogging clustering method based on correlation constraints

$$
\begin{aligned}
&=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\varepsilon(i,j)\left(U(i,k)-U(j,k)\right)^2 \\
&=\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\varepsilon(i,j)\,U^2(i,k)-\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\varepsilon(i,j)\,U(i,k)\,U(j,k) \\
&=\sum_{k=1}^{d}U^T(:,k)\,(D-Z)\,U(:,k) \\
&=Tr\left(U^T L U\right)
\end{aligned}
\tag{4-6}
$$

其中 $L = D - Z$ 是拉普拉斯矩阵 (Laplacian matrix)，D 是对角矩阵，对角线上的元素是 $D(i,i) = \sum_{j=1}^{n} Z(j,i)$。Z 是相关约束系数矩阵，其中相关约束系数

$\varepsilon(i,j)$ 表示的是微博之间的相关程度，依据前文，将使用微博之间检索值之差的倒数表示相关系数。它定义为：

$$Z=\begin{bmatrix}\varepsilon(1,1) & \cdots & \varepsilon(1,n)\\ \vdots & \ddots & \vdots\\ \varepsilon(n,1) & \cdots & \varepsilon(n,n)\end{bmatrix} \tag{4-7}$$

最后，将该正则项加入到优化目标函数中后，可得如下公式4-8。

$$\min F=||W-UV^T||_F^2+\alpha||U||_F^2+\beta||V||_F^2+\gamma Tr(U^TLU) \tag{4-8}$$

$$s.t.U\geq 0,V\geq 0 \tag{4-9}$$

其中，α、β 和 γ 是常数，表示各项之间的权重。现在，已经得到了最终的优化目标函数，通过对该公式进行求解，得到的矩阵 U 将表示最终聚类的结果。

4.4 基于相关约束的微博聚类算法求解

在上述内容中，已经推导出了基于相关约束的微博聚类方法的最终待求解的表达式，如公式4-8所示，但通过该公式无法直接得出分解项 W 的解，因此本节将对求解进行说明，并给出求解的算法。

这是一个包含不等式约束的最优化问题，因此将利用 KKT 条件对公式进行求解。首先，因为 F-范数无法进行求导，所以将目标函数转换为矩阵的迹的形式。

$$||W||_F^2=Tr\left(W^T\cdot W\right) \tag{4-10}$$

F-范数可以转换为矩阵的迹的形式，转换公式如4-10所示。因此可以将最终带求解的公式4-8转换为矩阵的迹的形式，去除其中的常数项后，形式如下：

$$F=Tr(-2W^TUV^T+VU^TUV^T)+\alpha Tr(UU^T) \tag{4-11}$$

$$+\beta Tr(VV^T)+\gamma Tr(U^TLU) \tag{4-12}$$

然后，通过 KKT 条件对上述公式中的未知变量 U 和 V 进行求解，首先需要求解偏微分导数，其结果如下：

$$WV+\gamma ZU=UV^TV+\alpha U+\gamma DU \tag{4-13}$$

$$W^T U = VU^T U + \beta V \tag{4-14}$$

在满足 KKT 条件的情况下，根据上述公式可以得到迭代更新公式：

$$U(i,k) \leftarrow U(i,k)\sqrt{\frac{WV + \gamma ZU}{UV^T V + \alpha U + \gamma DU}} \tag{4-15}$$

$$V(i,k) \leftarrow V(i,k)\sqrt{\frac{W^T U}{VU^T U + \beta V}} \tag{4-16}$$

其中 α、β、γ 均为公式中各项对应的权重。至此，就可以对分解项 U 和 V 进行求解了。

4.5　算法伪代码

本节将对基于相关约束的微博聚类算法过程进行说明。首先，要对待求解的矩阵 U 和 V 进行初始化，之后开始进行迭代工作，这里迭代应直到收敛结束，但由于到后期迭代时间消耗很长，且影响很小，因此在实际工作中，常迭代固定次数，如 3000 次。最后，在求得矩阵 U 后，微博聚类的工作就结束。

4.6　本章小结

本章着重对本文研究的基于相关约束的微博聚类方法进行了详细的介绍，在之后的章节中，将利用实验验证该方法的性能，并得出结论。

Algorithm 2: 基于相关约束的微博聚类算法

Data: $\mathbf{W}$

Result: 聚类结果 $\mathbf{clu}_n$

begin

基于检索模型计算每篇微博的 BM25 检索值 $P(Q|D)_n$

计算得到矩阵 $\mathbf{Z}_n, \mathbf{D}_n$

随机初始化 $\mathbf{W}_n, \mathbf{U}_n, \mathbf{V}_n$

while 反复更新直至收敛 **do**

计算如下算式

$\mathbf{T1} = \mathbf{WV} + \gamma\mathbf{ZU}$

$\mathbf{T2} = \mathbf{UV}^T\mathbf{V} + \alpha\mathbf{U} + \gamma\mathbf{DU}$

$\mathbf{T3} = \mathbf{W}^T\mathbf{U}$

$\mathbf{T4} = \mathbf{VU}^T\mathbf{U} + \beta\mathbf{V}$

更新 $\mathbf{U}(i,j) \leftarrow \mathbf{U}(i,j)\sqrt{\frac{\mathbf{T1}(i,j)}{\mathbf{T2}(i,j)}}$

更新 $\mathbf{V}(i,j) \leftarrow \mathbf{V}(i,j)\sqrt{\frac{\mathbf{T3}(i,j)}{\mathbf{T4}(i,j)}}$

根据矩阵 $\mathbf{U}_n$，得到聚类结果 $\mathbf{clu}_n$

返回 $\mathbf{clu}_n$

第 5 章　对比实验设计以及性能分析

本文设计并实现了一个基于聚类约束的高相关微博检索系统，提出基于非负矩阵分解，融合聚类信息的方法，以提高检索结果的有效性。本章将对提出的基于聚类约束的微博检索方法进行实验并评测他们的性能。

5.1　对比实验目标

在本节中，对文章中提出的方法进行对比实验，以评估该微博的检索系统的有效性。通过这些实验，希望以下几组对比试验能解释以下几个问题：

- 系统查询扩展模块中的几种扩展方式对比效果如何?
- 系统对类簇的两种不同处理方式对比效果如何?
- 系统中多元检索模型对比效果如何?
- 系统中参数设置对最终实验性能的影响?
- 基于相关约束的微博聚类方法是否有效?

5.2　实验数据集

微博任务数据集包括了十天内的数据。数据集包括 8936348 条微博和 55 条用户兴趣信息，微博数据样例如下表5.1所示：

用户兴趣信息样例如下表5.2所示：

测试数据如表5.3所示，主要包含四个部分，查询信息 ID 表示测试数据对应的查询，微博 ID 对应具体的微博，测评得分表示微博和查询的相关性情况，分为三个等级：0 代表和查询无关，1 和 2 代表和查询相关，分数越高表示与查询主题越相关。

5.3　评测指标

本文使用 NDCG 作为评价标准，评估本文所提出的基于聚类约束的高质量微博检索方法。DCG (Discounted cumulative gain) 是对检索排序质量的衡量指标。在检索过程中，通常需要评估搜索结果列表信息质量情况，较为相关的结果是否排在相对靠前的位置，而相关性稍差的信息是否排在相对靠后的位置。基于这种考虑，将数据集中的每一个结果项赋予一个评分值，表示相关度的大小，结果列表中的项目所代表的值越大，那么结果越有意义，即增益（gain）。对于与查询无关的项，通常设置增益为 0。结果列表中的每一项增益相加，即得到累计增益

表 5.1　微博样本
Tab. 5.1　Sample of Microblogging

时间	微博 ID	内容
Tue Aug 02 00:00:43 +0000 2016	760264026756558849	RT @melanieusn1979: The truth of #DNCinPHL #DNCleak .@DemConvention .@People4Bernie .@BernieSanders https://t.co/POdLVFdelz
Tue Aug 02 00:00:43 +0000 2016	760264026777530368	#viral Malia Obama Twerks At Lollapalooza: 5 Times First Daughters... https://t.co/f1szCLgoom https://t.co/FPo3AlX9c2
Tue Aug 02 00:00:44 +0000 2016	760264030950821888	Michael Bennett clarifies comments on social issues https://t.co/HncBDuBLWM
Tue Aug 02 00:00:44 +0000 2016	760264030963380224	Love in Seoul #BOT

(cumulative gain)，同时，在这些增益累加之前，要赋予一定的权值，通常设置权值为与该项位置相关的对数值，即 DCG 值。在实际计算中，通常要将 DCG 值做标准化处理，以便在不同的场景下做比较，因此采用实际 DCG 值与理想 DCG 值之比作为最终结果，NDCG 计算公式如下：

$$\mathrm{NDCG} = Z_i \sum_{j=1}^{R} \frac{2^{r(j)} - 1}{\log(1 + j)}. \tag{5-1}$$

5.4　实验设计

在本节中，通过设计对比实验，评估本文所提出的基于聚类约束的高质量微博检索方法。针对本章第一部分所提出的问题，主要设计以下五组实验解决：

1. 查询扩展方法性能分析
   针对查询扩展方法的性能进行实验，根据表3.1中所列查询扩展方式，使用 BM25 检索模型计算微博与话题的相关度，即 $P(Q|D)$，对比所有话题下的实验性能。并将最优扩展方法作为后续实验的基准方法。
2. 类簇处理方法性能分析
   针对类簇的不同处理方法进行实验，通过计算聚类约束下的检索相关度比

表 5.2　用户兴趣信息样本
Tab. 5.2　Sample of user interest information

用户兴趣信息 ID	标题	描述	具体场景叙述
RTS1	transgender bathrooms	Find information on different sides of the debate on which bathroom can be used by a transgender individual	The user is interested in the politics of the transgender bathroom debate, including current and proposed bills, as well as backlash and economic implications (for example, boycotts).
RTS2	Zika in Ecuador	Find updates on the current Zika crisis in the country of Ecuador.	The user has family in Ecuador and wants to see how her family might be affected by the Zika crisis. She's interested in reports of new cases as well as measures being taken to control the outbreak.

表 5.3　检索结果样本示例
Tab. 5.3　Sample of retrieval result

话题 ID	保留字	微博 ID	测评得分
RTS1	Q0	760275863107698688	1
RTS1	Q0	760293772773105664	2
RTS1	Q0	760308251518918656	1
RTS1	Q0	760403139254038528	2
RTS1	Q0	760471758071824384	2
RTS1	Q0	760490208819380224	2
RTS1	Q0	760586426127937536	2
RTS1	Q0	760663081252954112	0
RTS1	Q0	760667510412718084	2
RTS1	Q0	761154624292200453	0

较性能，即 $P(Q|D)P(Q|Clu)$，不同类簇处理方法具体如下：

(a) 将文档类簇内所有微博整合为一个长文本，长文本代表文档类簇。

(b) 将基向量作为不同类簇中心，基矩阵代表所有文档类簇。

3. 多元检索模型性能分析

以最优查询扩展方法为基准方法，通过对比多元检索模型的性能，即计算 $P(Q|D)P(Q|Clu)$、$P(Q|D)P(E|D)$、$P(Q|D)P(Q|Clu)P(E|D)$ 和 $P(E|Clu)$，对比 $P(Q|D)$，验证基于聚类约束的高质量微博检索方法是否有效。

4. 系统中参数设置性能分析

在本文提出的检索方法中，聚类数量 k 的设置对结果的影响最为突出，因此在不同聚类数量下进行实验，期望得到最优的 k。

5. 基于相关约束的微博聚类方法性能分析

针对本文提出的基于相关约束的微博聚类方法进行实验，对比无相关约束情况下的实验性能。

5.5 实验结果分析

在本节中将展示实验结果，并对实验结果进行分析，得出结论。

5.5.1 查询扩展方法实验结果与分析

根据表3.1中所有查询扩展方法，将实验每组扩展方法，使用检索模块计算 $P(Q|D)$，并对比实验结果。实验结果如下表所示5.4。

表 5.4 查询扩展实验结果

topic	wiki	Tweb	web20	web50	news50	news20	Tnews
MB226	0.13562	0.09464	0.14307	0.13155	0.03333	0.03333	0.00000
MB229	0.07838	0.16035	0.10405	0.10762	0.15137	0.11222	0.17909
MB230	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
MB239	0.00000	0.03973	0.03904	0.04144	0.03462	0.02463	0.11681
MB254	0.00000	0.00000	0.00000	0.00000	0.00000	0.03562	0.00000
MB256	0.00000	0.15842	0.10226	0.10108	0.12729	0.07700	0.12700
MB258	0.04471	0.10221	0.09736	0.14107	0.10324	0.14180	0.11434
MB265	0.12891	0.04307	0.13333	0.12891	0.08010	0.08175	0.12619
MB267	0.31107	0.18663	0.27300	0.31107	0.21685	0.21331	0.15717
MB276	0.00000	0.01095	0.06810	0.06810	0.07798	0.06810	0.06021
MB286	0.00000	0.01672	0.01461	0.06790	0.07488	0.07947	0.04307
MB319	0.24297	0.23420	0.28181	0.28343	0.30919	0.34287	0.24177
MB320	0.35490	0.41244	0.37565	0.41622	0.45478	0.52804	0.43937
MB332	0.02346	0.12085	0.00000	0.03973	0.01391	0.03854	0.13008
MB351	0.03284	0.10559	0.07301	0.07402	0.01036	0.01095	0.01170

topic	wiki	Tweb	web20	web50	news50	news20	Tnews
MB358	0.08656	0.09317	0.08656	0.08656	0.04455	0.04053	0.07961
MB361	0.00000	0.12824	0.14237	0.13964	0.12501	0.13964	0.04307
MB362	0.32670	0.30530	0.31264	0.32064	0.27048	0.29338	0.37338
MB363	0.02634	0.24540	0.07034	0.08733	0.03996	0.04158	0.06373
MB365	0.00000	0.02044	0.02184	0.06309	0.23167	0.15803	0.38519
MB371	0.12143	0.07731	0.18870	0.11276	0.13935	0.14678	0.07847
MB377	0.20803	0.13603	0.20803	0.20803	0.10665	0.12387	0.11709
MB381	0.07854	0.16566	0.12775	0.17125	0.06008	0.05001	0.14437
MB382	0.05000	0.06131	0.06934	0.00000	0.06131	0.00000	0.10000
MB391	0.38514	0.38272	0.38702	0.38514	0.40869	0.24551	0.38141
MB392	0.08032	0.34512	0.11543	0.07943	0.03974	0.08207	0.23959
MB409	0.16169	0.18792	0.24982	0.16169	0.20928	0.16431	0.15690
MB410	0.00000	0.00000	0.00000	0.06309	0.00000	0.00000	0.10000
MB414	0.07087	0.18520	0.13110	0.17154	0.20791	0.20359	0.18086
MB420	0.13117	0.16672	0.16672	0.13513	0.09385	0.11451	0.15901
MB425	0.00000	0.03155	0.00000	0.00000	0.00000	0.00000	0.00000
MB431	0.30000	0.38155	0.36879	0.28155	0.30000	0.24307	0.20178
MB436	0.01510	0.02824	0.01815	0.05277	0.00000	0.00000	0.10000
MB438	0.00000	0.03333	0.03333	0.00000	0.00000	0.00000	0.03869
MB440	0.25231	0.23980	0.33206	0.34728	0.20681	0.20834	0.24088
RTS1	0.03704	0.18761	0.14269	0.12288	0.22723	0.14261	0.21444
RTS10	0.27759	0.36953	0.32449	0.34722	0.35864	0.26202	0.30085
RTS13	0.08308	0.24403	0.28131	0.22483	0.22866	0.23371	0.24042
RTS14	0.15000	0.07197	0.13010	0.25000	0.10000	0.09871	0.04307
RTS19	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
RTS2	0.00000	0.00000	0.10000	0.10000	0.00000	0.00000	0.00000
RTS21	0.11254	0.17372	0.13353	0.13681	0.11679	0.11679	0.17372
RTS24	0.00000	0.04347	0.01036	0.02044	0.00000	0.06131	0.05711
RTS25	0.08896	0.08853	0.08522	0.15480	0.15426	0.09318	0.14350
RTS27	0.00000	0.10517	0.01091	0.03714	0.02937	0.04865	0.07184
RTS28	0.01357	0.11815	0.11053	0.02921	0.05803	0.06572	0.14803
RTS31	0.05000	0.10332	0.08164	0.07090	0.05000	0.05000	0.09883
RTS32	0.00000	0.13339	0.20295	0.02258	0.07391	0.21488	0.13230
RTS35	0.10212	0.08597	0.13702	0.23454	0.12652	0.17197	0.11216
RTS36	0.07346	0.17642	0.13990	0.14033	0.04192	0.13274	0.17924
RTS37	0.00000	0.06127	0.15256	0.14285	0.06695	0.16530	0.06455
RTS4	0.00000	0.08066	0.19809	0.18702	0.08681	0.06309	0.06628
RTS43	0.09155	0.11784	0.13528	0.15322	0.17194	0.17043	0.09366
RTS5	0.02963	0.03975	0.15885	0.18363	0.14888	0.04294	0.19720
RTS6	0.06564	0.07915	0.08083	0.15774	0.03883	0.07954	0.08059
avg	0.08768	0.13056	0.13366	0.13628	0.11476	0.11375	0.13361

通过对比实验结果，可以得出以下实验结论：

1. 使用维基百科扩展文档的实验结果较差，在本文中不适合用作查询扩展文档。
2. 使用搜索引擎的搜索结果作为扩展文档，采用前 50 条搜索结果要优于前 20 条搜索结果。
3. 同时，采用时效性新闻搜索结果要优于时效性网页搜索结果。
4. 非时效性前 50 条网页搜索结果作为扩展文档，在本文中效果最好

最终，采用非时效性前 50 条网页搜索结果作为扩展文档，并且，这种方法将作为后续实验的基准方法。

5.5.2 类簇处理方法实验结果与分析

通过计算 $P(Q|D)P(Q|Clu)$ 对微博进行排序，根据实验结果，对比长文本类簇和基矩阵类簇对实验结果的影响，实验结果如下表5.5。

表 5.5 类簇实验结果对比

topic	$P(Q\|D)$	**text** $P(Q\|D)P(Q\|Clu)$	**matrix** $P(Q\|D)P(Q\|Clu)$
MB226	0.13155	0.12218	0.13267
MB229	0.10762	0.10830	0.10808
MB230	0.00000	0.00000	0.00000
MB239	0.04144	0.04724	0.05159
MB254	0.00000	0.00000	0.00000
MB256	0.10108	0.09097	0.10786
MB258	0.14107	0.13691	0.13553
MB265	0.12891	0.13327	0.13721
MB267	0.31107	0.31252	0.30882
MB276	0.06810	0.05553	0.03494
MB286	0.06790	0.06790	0.06790
MB319	0.28343	0.27316	0.30479
MB320	0.41622	0.51866	0.45769
MB332	0.03973	0.05829	0.02889
MB351	0.07402	0.07201	0.08854
MB358	0.08656	0.08394	0.07541
MB361	0.13964	0.15002	0.14794
MB362	0.32064	0.32143	0.31639
MB363	0.08733	0.08883	0.08899
MB365	0.06309	0.03155	0.05779
MB371	0.11276	0.12283	0.06220
MB377	0.20803	0.15956	0.21038

topic	$P(Q\|D)$	**text** $P(Q\|D)P(Q\|Clu)$	**matrix** $P(Q\|D)P(Q\|Clu)$
MB381	0.17125	0.11834	0.17848
MB382	0.00000	0.04736	0.04953
MB391	0.38514	0.38501	0.40100
MB392	0.07943	0.09678	0.08953
MB409	0.16169	0.17008	0.17419
MB410	0.06309	0.08762	0.07417
MB414	0.17154	0.20224	0.20224
MB420	0.13513	0.13513	0.13513
MB425	0.00000	0.00000	0.00289
MB431	0.28155	0.21833	0.21833
MB436	0.05277	0.05814	0.06102
MB438	0.00000	0.00867	0.00000
MB440	0.34728	0.34728	0.34728
RTS1	0.12288	0.11863	0.12205
RTS10	0.34722	0.35219	0.35288
RTS13	0.22483	0.22193	0.22352
RTS14	0.25000	0.25000	0.25000
RTS19	0.00000	0.00000	0.00000
RTS2	0.10000	0.10000	0.07000
RTS21	0.13681	0.17054	0.14442
RTS24	0.02044	0.02372	0.00000
RTS25	0.15480	0.11798	0.11289
RTS27	0.03714	0.01780	0.03714
RTS28	0.02921	0.01775	0.01700
RTS31	0.07090	0.06578	0.07122
RTS32	0.02258	0.11122	0.01000
RTS35	0.23454	0.23550	0.23584
RTS36	0.14033	0.19128	0.19128
RTS37	0.14285	0.14181	0.14719
RTS4	0.18702	0.20282	0.19804
RTS43	0.15322	0.19757	0.19674
RTS5	0.18363	0.18773	0.18773
RTS6	0.15774	0.14476	0.15118
avg	0.13628	0.13998	0.13775
chg	-	2.72%	1.09%

根据实验对比结果，长文本类簇更为有效，后续实验将采用长文本类簇。

5.5.3 多元检索模型实验结果与分析

为了对比多元检索模型的性能，将计算 $P(Q|D)P(Q|Clu)$、$P(Q|D)P(E|D)$、$P(Q|D)P(Q|Clu)P(E|D)$、$P(E|Clu)$ 和 $P(E|D)$，按计算结果排序后，对比 $P(Q|D)$ 原始排序，实验结果如下表所示5.6。

表 5.6 多元检索模型实验结果

topic	$P(Q\|D)$	$P(Q\|D)\cdot P(E\|D)$	$P(Q\|D)\cdot P(Q\|Clu)$	$P(Q\|D)P(Q\|Clu)\cdot P(E\|D)$	$P(E\|Clu)$	$P(E\|D)$
MB226	0.13155	0.10000	0.12218	0.10000	0.13499	0.04307
MB229	0.10762	0.13151	0.10830	0.13768	0.08145	0.14409
MB230	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
MB239	0.04144	0.05414	0.04724	0.05436	0.05252	0.02463
MB254	0.00000	0.03333	0.00000	0.00000	0.00000	0.00000
MB256	0.10108	0.14655	0.09097	0.12360	0.09097	0.10187
MB258	0.14107	0.16839	0.13691	0.17177	0.13464	0.14285
MB265	0.12891	0.13155	0.13327	0.11525	0.04746	0.11717
MB267	0.31107	0.30577	0.31252	0.30864	0.29941	0.27575
MB276	0.06810	0.06810	0.05553	0.06810	0.04405	0.07859
MB286	0.06790	0.23247	0.06790	0.23247	0.06790	0.21602
MB319	0.28343	0.27336	0.27316	0.27692	0.15156	0.23236
MB320	0.41622	0.25857	0.51866	0.38620	0.51600	0.17001
MB332	0.03973	0.05135	0.05829	0.05294	0.05856	0.03191
MB351	0.07402	0.25725	0.07201	0.25725	0.06379	0.25725
MB358	0.08656	0.02021	0.08394	0.02021	0.08853	0.01564
MB361	0.13964	0.25002	0.15002	0.25002	0.15002	0.13067
MB362	0.32064	0.32077	0.32143	0.35362	0.21869	0.18720
MB363	0.08733	0.11496	0.08883	0.11496	0.08899	0.09743
MB365	0.06309	0.17737	0.03155	0.13777	0.02500	0.17259
MB371	0.11276	0.13824	0.12283	0.18084	0.06220	0.17321
MB377	0.20803	0.10907	0.15956	0.09940	0.07470	0.18272
MB381	0.17125	0.17125	0.11834	0.17083	0.16967	0.17125
MB382	0.00000	0.01846	0.04736	0.01846	0.00527	0.00000
MB391	0.38514	0.40378	0.38501	0.41889	0.34977	0.33480
MB392	0.07943	0.14126	0.09678	0.15033	0.09678	0.09207
MB409	0.16169	0.17435	0.17008	0.16825	0.14756	0.07593
MB410	0.06309	0.00000	0.08762	0.01505	0.00000	0.00000
MB414	0.17154	0.15890	0.20224	0.19368	0.20139	0.09055
MB420	0.13513	0.15194	0.13513	0.16115	0.13513	0.09793
MB425	0.00000	0.02891	0.00000	0.00000	0.00431	0.06309
MB431	0.28155	0.23626	0.21833	0.19880	0.21833	0.19650
MB436	0.05277	0.04307	0.05814	0.06609	0.05647	0.05000
MB438	0.00000	0.00000	0.00867	0.00000	0.01787	0.00000

topic	$P(Q\|D)$	$P(Q\|D)\cdot P(E\|D)$	$P(Q\|D)\cdot P(Q\|Clu)$	$P(Q\|D)P(Q\|Clu)\cdot P(E\|D)$	$P(E\|Clu)$	$P(E\|D)$
MB440	0.34728	0.30351	0.34728	0.30351	0.34728	0.21516
RTS1	0.12288	0.19443	0.11863	0.17372	0.21052	0.19962
RTS10	0.34722	0.35583	0.35219	0.35710	0.34808	0.30974
RTS13	0.22483	0.24427	0.22193	0.25900	0.16596	0.25373
RTS14	0.25000	0.17640	0.25000	0.17564	0.22500	0.09871
RTS19	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
RTS2	0.10000	0.03333	0.10000	0.03277	0.10000	0.00000
RTS21	0.13681	0.26131	0.17054	0.21533	0.18193	0.14178
RTS24	0.02044	0.02882	0.02372	0.03010	0.02372	0.04091
RTS25	0.15480	0.11951	0.11798	0.12244	0.09445	0.11585
RTS27	0.03714	0.02641	0.01780	0.00000	0.01780	0.00000
RTS28	0.02921	0.06564	0.01775	0.01235	0.02020	0.11244
RTS31	0.07090	0.18255	0.06578	0.18255	0.07122	0.18749
RTS32	0.02258	0.04620	0.11122	0.03844	0.10000	0.09821
RTS35	0.23454	0.15096	0.23550	0.15291	0.21948	0.08203
RTS36	0.14033	0.17870	0.19128	0.23194	0.19128	0.12742
RTS37	0.14285	0.12090	0.14181	0.12329	0.13998	0.12439
RTS4	0.18702	0.26047	0.20282	0.29285	0.22326	0.32018
RTS43	0.15322	0.11559	0.19757	0.11771	0.19757	0.07617
RTS5	0.18363	0.11560	0.18773	0.11699	0.19639	0.07054
RTS6	0.15774	0.19033	0.14476	0.19033	0.14131	0.18353
avg	0.13628	0.14622	0.13998	0.14786	0.12853	0.12227
chg	-	7.29%	2.72%	8.50%	-5.68%	-10.28%

通过对比实验结果，可以得出以下实验结论：

1. 实验性能对比:$P(Q|D)P(Q|Clu)P(E|D) > P(Q|D)P(E|D) > P(Q|D)P(Q|Clu)$。在聚类约束和相关文档约束下，$P(Q|D)P(Q|Clu)P(E|D)$ 性能最优。
2. $P(E|Clu)$ 和 $P(E|D)$ 没有取得性能提升。
3. 对比 $P(Q|D)$，$P(E|D)$ 和 $P(Q|D)P(E|D)$ 可以发现，在相关文档约束下，实验性能获得提升。
4. 对比 $P(Q|D)$ 和 $P(Q|D)P(Q|Clu)$，$P(Q|D)P(E|D)$ 和 $P(Q|D)P(Q|Clu)P(E|D)$ 两组实验结果可以发现，在聚类约束下，实验性能获得提升。

5.5.4　参数比较实验结果与分析

在聚类任务中，聚类数目是非常重要的参数。本节将对比聚类数目对实验结果的影响，聚类数目 k 设置为 2、3、4、5、10。在多元检索模型中，

$P(Q|D)P(Q|Clu)P(E|D)$，$P(Q|D)P(E|D)$ 和 $P(Q|D)P(Q|Clu)$ 实验性能提升，但是 $P(Q|D)P(E|D)$ 与聚类数目无关，因此，将对比聚类数目对 $P(Q|D)P(Q|Clu)$ 和 $P(Q|D)P(Q|Clu)P(E|D)$ 的影响。

聚类数目对 $P(Q|D)P(Q|Clu)$ 影响的实验结果如下表5.7。

表 5.7　聚类数目对 $P(Q|D)P(Q|Clu)$ 实验结果影响对比

topic	$P(Q\|D)$	**k=2**	**k=3**	**k=4**	**k=5**	**k=10**
MB226	0.13155	0.12218	0.10774	0.10356	0.10387	0.09891
MB229	0.10762	0.10830	0.12479	0.11080	0.09534	0.08390
MB230	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
MB239	0.04144	0.04724	0.05796	0.05925	0.05508	0.03564
MB254	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
MB256	0.10108	0.09097	0.08747	0.08933	0.08494	0.09118
MB258	0.14107	0.13691	0.13964	0.12721	0.13738	0.10474
MB265	0.12891	0.13327	0.13248	0.13538	0.13841	0.13762
MB267	0.31107	0.31252	0.30368	0.30895	0.28598	0.29081
MB276	0.06810	0.05553	0.06810	0.06810	0.06810	0.06810
MB286	0.06790	0.06790	0.06266	0.06493	0.07092	0.04498
MB319	0.28343	0.27316	0.17569	0.19737	0.27987	0.31487
MB320	0.41622	0.51866	0.38787	0.40321	0.44249	0.37714
MB332	0.03973	0.05829	0.05585	0.03522	0.04160	0.02508
MB351	0.07402	0.07201	0.09480	0.09275	0.10490	0.09853
MB358	0.08656	0.08394	0.07642	0.07208	0.07188	0.08215
MB361	0.13964	0.15002	0.10273	0.12246	0.13906	0.14068
MB362	0.32064	0.32143	0.32064	0.33065	0.32564	0.31451
MB363	0.08733	0.08883	0.07645	0.07289	0.06931	0.06902
MB365	0.06309	0.03155	0.00000	0.00177	0.00631	0.02762
MB371	0.11276	0.12283	0.10818	0.11137	0.11056	0.12787
MB377	0.20803	0.15956	0.17869	0.19961	0.20719	0.10946
MB381	0.17125	0.11834	0.09360	0.15565	0.17125	0.14983
MB382	0.00000	0.04736	0.05036	0.03200	0.01933	0.02843
MB391	0.38514	0.38501	0.38171	0.38386	0.38434	0.37544
MB392	0.07943	0.09678	0.11168	0.09343	0.08909	0.06749
MB409	0.16169	0.17008	0.16143	0.23255	0.25808	0.21244
MB410	0.06309	0.08762	0.09262	0.05276	0.05644	0.10640
MB414	0.17154	0.20224	0.20260	0.21375	0.17243	0.17634
MB420	0.13513	0.13513	0.11878	0.13367	0.15129	0.16348
MB425	0.00000	0.00000	0.00000	0.00000	0.00000	0.02809
MB431	0.28155	0.21833	0.19273	0.25008	0.26290	0.29368
MB436	0.05277	0.05814	0.05818	0.05689	0.04836	0.01897
MB438	0.00000	0.00867	0.03569	0.03357	0.03255	0.00000
MB440	0.34728	0.34728	0.34737	0.34679	0.35247	0.41138

topic	$P(Q\|D)$	k=2	k=3	k=4	k=5	k=10
RTS1	0.12288	0.11863	0.12353	0.14919	0.14826	0.10801
RTS10	0.34722	0.35219	0.37876	0.36802	0.38206	0.36434
RTS13	0.22483	0.22193	0.21857	0.22171	0.22025	0.22127
RTS14	0.25000	0.25000	0.25500	0.24063	0.24580	0.22025
RTS19	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
RTS2	0.10000	0.10000	0.10000	0.10000	0.10000	0.07155
RTS21	0.13681	0.17054	0.05707	0.12045	0.16209	0.17443
RTS24	0.02044	0.02372	0.01123	0.02061	0.01622	0.01897
RTS25	0.15480	0.11798	0.11801	0.13780	0.12212	0.13901
RTS27	0.03714	0.01780	0.02140	0.01091	0.01091	0.01427
RTS28	0.02921	0.01775	0.01408	0.02111	0.01838	0.02812
RTS31	0.07090	0.06578	0.07064	0.06368	0.05874	0.05623
RTS32	0.02258	0.11122	0.01741	0.03118	0.03128	0.01818
RTS35	0.23454	0.23550	0.23957	0.23767	0.23954	0.21413
RTS36	0.14033	0.19128	0.18730	0.18853	0.19560	0.14396
RTS37	0.14285	0.14181	0.13298	0.11721	0.10669	0.12863
RTS4	0.18702	0.20282	0.20282	0.19995	0.19827	0.19230
RTS43	0.15322	0.19757	0.19453	0.18795	0.18381	0.16831
RTS5	0.18363	0.18773	0.13341	0.13307	0.14947	0.16459
RTS6	0.15774	0.14476	0.14105	0.13802	0.11321	0.12042
avg	0.13628	0.13998	0.12956	0.13417	0.13709	0.13167
chg	-	2.72%	-4.93%	-1.54%	0.60%	-3.38%

聚类数目对 $P(Q|D)P(Q|Clu)P(E|D)$ 影响的实验结果如下表5.8。

表 5.8　聚类数目对 $P(Q|D)P(Q|Clu)P(E|D)$ 实验结果影响对比

topic	$P(Q\|D)$	k=2	k=3	k=4	k=5	k=10
MB226	0.13155	0.10000	0.09262	0.08893	0.08000	0.05754
MB229	0.10762	0.13768	0.14705	0.12450	0.11844	0.09985
MB230	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
MB239	0.04144	0.05436	0.05568	0.05720	0.05755	0.05267
MB254	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
MB256	0.10108	0.12360	0.12066	0.11774	0.11789	0.11814
MB258	0.14107	0.17177	0.17061	0.19614	0.20299	0.17440
MB265	0.12891	0.11525	0.13953	0.13297	0.13359	0.16468
MB267	0.31107	0.30864	0.30379	0.30778	0.30295	0.28510
MB276	0.06810	0.06810	0.06810	0.06810	0.06810	0.06810
MB286	0.06790	0.23247	0.23171	0.22823	0.23368	0.22559
MB319	0.28343	0.27692	0.28581	0.29499	0.29855	0.31872
MB320	0.41622	0.38620	0.26650	0.27043	0.32163	0.29167

topic	$P(Q\|D)$	k=2	k=3	k=4	k=5	k=10
MB332	0.03973	0.05294	0.05585	0.05500	0.05394	0.05464
MB351	0.07402	0.25725	0.25725	0.25725	0.24356	0.24417
MB358	0.08656	0.02021	0.04045	0.02933	0.03074	0.05138
MB361	0.13964	0.25002	0.19655	0.17136	0.21634	0.21529
MB362	0.32064	0.35362	0.32613	0.35897	0.34278	0.30342
MB363	0.08733	0.11496	0.11252	0.11199	0.11150	0.12533
MB365	0.06309	0.13777	0.09937	0.10677	0.14254	0.12096
MB371	0.11276	0.18084	0.16110	0.16273	0.16236	0.13753
MB377	0.20803	0.09940	0.11897	0.12185	0.11735	0.10664
MB381	0.17125	0.17083	0.16743	0.17013	0.17125	0.16694
MB382	0.00000	0.01846	0.01945	0.01193	0.01503	0.00593
MB391	0.38514	0.41889	0.42096	0.42167	0.41426	0.42948
MB392	0.07943	0.15033	0.15240	0.14619	0.15196	0.15165
MB409	0.16169	0.16825	0.15653	0.15630	0.14851	0.14783
MB410	0.06309	0.01505	0.00000	0.00720	0.01204	0.02428
MB414	0.17154	0.19368	0.18291	0.18505	0.15169	0.16389
MB420	0.13513	0.16115	0.17141	0.18081	0.19717	0.21160
MB425	0.00000	0.00000	0.00000	0.00000	0.02313	0.02639
MB431	0.28155	0.19880	0.18574	0.21337	0.22823	0.31740
MB436	0.05277	0.06609	0.06939	0.06517	0.06596	0.03775
MB438	0.00000	0.00000	0.02911	0.02197	0.01852	0.00000
MB440	0.34728	0.30351	0.30198	0.30198	0.31487	0.37448
RTS1	0.12288	0.17372	0.18177	0.19435	0.19647	0.17855
RTS10	0.34722	0.35710	0.38433	0.38500	0.39276	0.36825
RTS13	0.22483	0.25900	0.26182	0.25988	0.26297	0.26352
RTS14	0.25000	0.17564	0.18160	0.16341	0.18107	0.22713
RTS19	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
RTS2	0.10000	0.03277	0.03481	0.03562	0.03593	0.03652
RTS21	0.13681	0.21533	0.11395	0.11587	0.15445	0.18068
RTS24	0.02044	0.03010	0.02475	0.02833	0.03133	0.02050
RTS25	0.15480	0.12244	0.12021	0.11841	0.11750	0.11604
RTS27	0.03714	0.00000	0.01056	0.00000	0.00000	0.00555
RTS28	0.02921	0.01235	0.03128	0.03928	0.03812	0.04526
RTS31	0.07090	0.18255	0.18255	0.17990	0.17849	0.17497
RTS32	0.02258	0.03844	0.03675	0.03659	0.04398	0.04273
RTS35	0.23454	0.15291	0.15472	0.15412	0.15898	0.16765
RTS36	0.14033	0.23194	0.20348	0.21271	0.21175	0.20234
RTS37	0.14285	0.12329	0.13241	0.12807	0.11904	0.12484
RTS4	0.18702	0.29285	0.30996	0.30574	0.30246	0.27723
RTS43	0.15322	0.11771	0.11786	0.11805	0.11819	0.11931
RTS5	0.18363	0.11699	0.13551	0.11546	0.10831	0.10889

topic	$P(Q\|D)$	k=2	k=3	k=4	k=5	k=10
RTS6	0.15774	0.19033	0.19033	0.19206	0.19033	0.20961
avg	0.13628	0.14786	0.14393	0.14413	0.14748	0.14805
chg	-	8.50%	5.62%	5.76%	8.22%	8.64%

为了便于比较，聚类数目对 $P(Q|D)P(Q|Clu)$ 影响的实验结果如下图5.1。

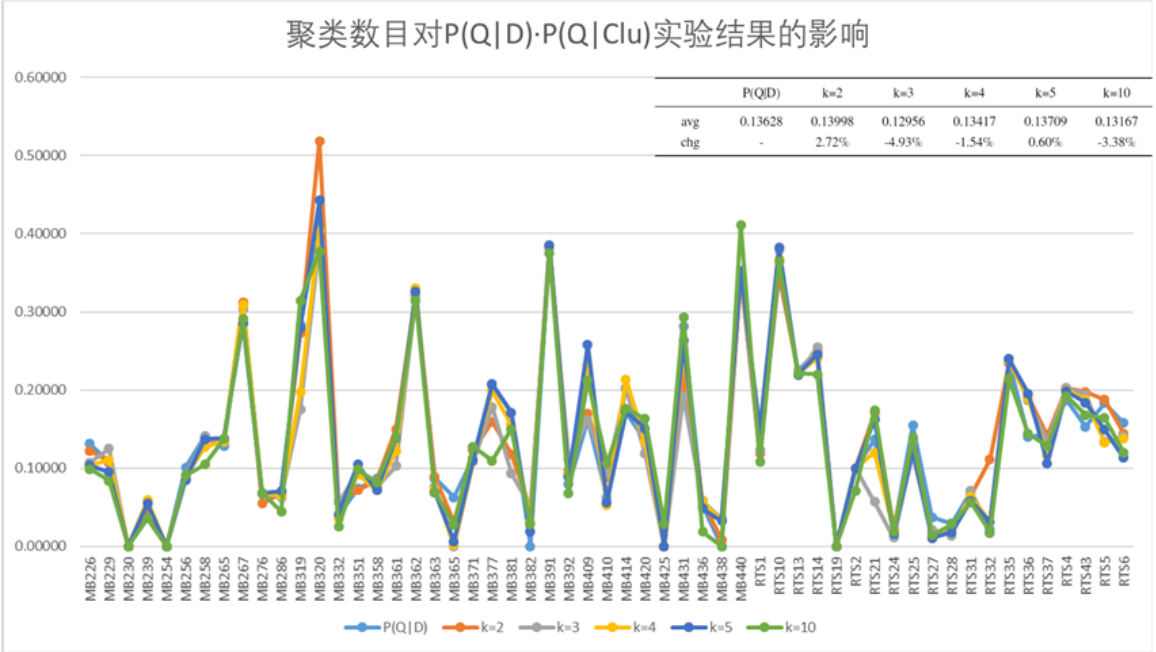

图 5.1　聚类数目对 $P(Q|D)P(Q|Clu)$ 实验结果的影响

Fig 5.1　Influence of clustering number on experimental results of $P(Q|D)P(Q|Clu)$

聚类数目对 $P(Q|D)P(Q|Clu)P(E|D)$ 影响的实验结果如下图5.2。

通过对比实验结果，可以得出以下实验结论：

1. 聚类数目对 $P(Q|D)P(Q|Clu)$ 和 $P(Q|D)P(Q|Clu)P(E|D)$ 的影响具有一致性。
2. 聚类数目在 $k = 2$ 和 $k = 10$ 取得了较优性能。
3. 由于不同主题之间的事件差异，不同主题取得最优性能的聚类数目一般不同。

5.5.5　基于相关约束的微博聚类方法实验结果与分析

为了对比基于相关约束的微博聚类方法，首先，将基于相关约束的微博聚类方法应用于多元检索模型，实验结果如下表5.9。

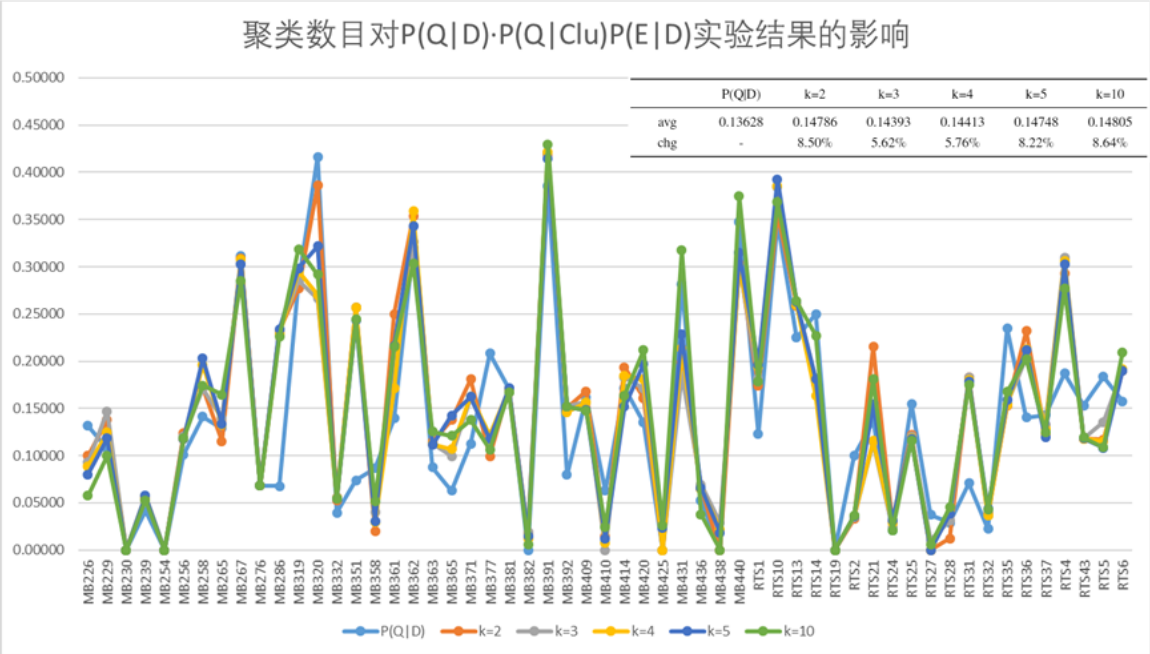

图 5.2　聚类数目对 $P(Q|D)P(Q|Clu)P(E|D)$ 实验结果的影响

Fig 5.2　Influence of clustering number on experimental results of $P(Q|D)P(Q|Clu)P(E|D)$

表 5.9　基于相关约束的多元检索模型实验结果

topic	$P(Q\|D)$	$P(Q\|D)\cdot P(E\|D)$	$P(Q\|D)\cdot P(Q\|Clu)$	$P(Q\|D)P(Q\|Clu)\cdot P(E\|D)$	$P(E\|Clu)$	$P(E\|D)$
MB226	0.13155	0.10000	0.10934	0.10000	0.06781	0.04307
MB229	0.10762	0.13151	0.10686	0.14031	0.06665	0.14409
MB230	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
MB239	0.04144	0.05414	0.04415	0.05414	0.03904	0.02463
MB254	0.00000	0.03333	0.00000	0.00333	0.00000	0.00000
MB256	0.10108	0.14655	0.10108	0.13503	0.10108	0.10187
MB258	0.14107	0.16839	0.14332	0.17977	0.14300	0.14285
MB265	0.12891	0.13155	0.13102	0.12917	0.13500	0.11717
MB267	0.31107	0.30577	0.31252	0.31057	0.31252	0.27575
MB276	0.06810	0.06810	0.06307	0.06810	0.06239	0.07859
MB286	0.06790	0.23247	0.06790	0.23247	0.06790	0.21602
MB319	0.28343	0.27336	0.27092	0.28929	0.11212	0.23236
MB320	0.41622	0.25857	0.49423	0.34378	0.50033	0.17001
MB332	0.03973	0.05135	0.05479	0.05294	0.04685	0.03191
MB351	0.07402	0.25725	0.07120	0.25725	0.02072	0.25725
MB358	0.08656	0.02021	0.08717	0.02343	0.08717	0.01564
MB361	0.13964	0.25002	0.15002	0.25002	0.15002	0.13067
MB362	0.32064	0.32077	0.32103	0.33431	0.28189	0.18720
MB363	0.08733	0.11496	0.08142	0.11303	0.08142	0.09743
MB365	0.06309	0.17737	0.03786	0.14569	0.03000	0.17259
MB371	0.11276	0.13824	0.12248	0.17915	0.06220	0.17321

topic	$P(Q\|D)$	$P(Q\|D)\cdot P(E\|D)$	$P(Q\|D)\cdot P(Q\|Clu)$	$P(Q\|D)P(Q\|Clu)\cdot P(E\|D)$	$P(E\|Clu)$	$P(E\|D)$
MB377	0.20803	0.10907	0.18549	0.10068	0.11532	0.18272
MB381	0.17125	0.17125	0.17125	0.17125	0.17125	0.17125
MB382	0.00000	0.01846	0.00315	0.00185	0.00000	0.00000
MB391	0.38514	0.40378	0.37459	0.41431	0.34869	0.33480
MB392	0.07943	0.14126	0.09626	0.14745	0.09626	0.09207
MB409	0.16169	0.17435	0.17796	0.17365	0.17796	0.07593
MB410	0.06309	0.00000	0.10000	0.03025	0.00000	0.00000
MB414	0.17154	0.15890	0.20985	0.18152	0.20985	0.09055
MB420	0.13513	0.15194	0.13513	0.16115	0.13513	0.09793
MB425	0.00000	0.02891	0.00000	0.01734	0.00000	0.06309
MB431	0.28155	0.23626	0.23333	0.19464	0.23333	0.19650
MB436	0.05277	0.04307	0.05908	0.06697	0.05981	0.05000
MB438	0.00000	0.00000	0.00774	0.00590	0.03297	0.00000
MB440	0.34728	0.30351	0.34728	0.30351	0.34728	0.21516
RTS1	0.12288	0.19443	0.12292	0.19506	0.21200	0.19962
RTS10	0.34722	0.35583	0.36132	0.37567	0.36113	0.30974
RTS13	0.22483	0.24427	0.21932	0.25372	0.15240	0.25373
RTS14	0.25000	0.17640	0.25000	0.17640	0.25000	0.09871
RTS19	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
RTS2	0.10000	0.03333	0.10000	0.03195	0.10000	0.00000
RTS21	0.13681	0.26131	0.18805	0.22644	0.21064	0.14178
RTS24	0.02044	0.02882	0.02184	0.02993	0.02372	0.04091
RTS25	0.15480	0.11951	0.12163	0.12634	0.10349	0.11585
RTS27	0.03714	0.02641	0.01780	0.00000	0.01780	0.00000
RTS28	0.02921	0.06564	0.01408	0.00469	0.01408	0.11244
RTS31	0.07090	0.18255	0.07064	0.18255	0.07122	0.18749
RTS32	0.02258	0.04620	0.01171	0.04021	0.10000	0.09821
RTS35	0.23454	0.15096	0.23485	0.15170	0.22450	0.08203
RTS36	0.14033	0.17870	0.18692	0.23563	0.18692	0.12742
RTS37	0.14285	0.12090	0.14285	0.12149	0.14285	0.12439
RTS4	0.18702	0.26047	0.20282	0.29211	0.22315	0.32018
RTS43	0.15322	0.11559	0.19725	0.11771	0.19725	0.07617
RTS5	0.18363	0.11560	0.18773	0.12254	0.19123	0.07054
RTS6	0.15774	0.19033	0.14373	0.19033	0.14839	0.18353
avg	0.13628	0.14622	0.13940	0.14885	0.13139	0.12227
chg	-	7.29%	2.29%	9.23%	-3.58%	-10.28%

通过对比实验结果，可以得出与无约束情况下的多元检索模型一致的结论，并且，在 $P(Q|D)P(Q|Clu)P(E|D)$ 中，基于相关约束的多元检索模型实验性能提

升更多。

为了更加全面的对比基于相关约束的聚类方法，将在不同聚类数目下进行实验，对比 $P(Q|D)P(Q|Clu)$ 和 $P(Q|D)P(Q|Clu)P(E|D)$ 的实验结果。基于相关约束的聚类方法表示为 $RNMF$，无约束情况下的聚类方法表示为 $BNMF$，实验结果如下表5.10和5.11。

表 5.10　聚类数目对 $P(Q|D)P(Q|Clu)$ 对比实验的影响

Tab. 5.10　Influence of clustering number on $P(Q|D)P(Q|Clu)$ comparative experiment

	$P(Q\|D)$	k=2	k=3	k=4	k=5	k=10
BNMF	0.13628	0.13998	0.12956	0.13417	0.13709	0.13167
chg	-	2.72%	-4.93%	-1.54%	0.60%	-3.38%
RNMF	0.13628	0.13940	0.12931	0.13114	0.13446	0.13182
chg	-	2.29%	-5.11%	-3.77%	-1.3%	-3.27%

表 5.11　聚类数目对 $P(Q|D)P(Q|Clu)P(E|D)$ 对比实验的影响

Tab. 5.11　Influence of clustering number on $P(Q|D)P(Q|Clu)P(E|D)$ comparative experiment

	$P(Q\|D)$	k=2	k=3	k=4	k=5	k=10
BNMF	0.13628	0.14786	0.14393	0.14413	0.14748	0.14805
chg	-	8.50%	5.62%	5.76%	8.22%	8.64%
RNMF	0.13628	0.14885	0.14419	0.14511	0.14694	0.14865
chg	-	9.23%	5.80%	6.49%	7.82%	9.08%

通过对比实验结果，可以得出以下实验结论：

1. 在 $P(Q|D)P(Q|Clu)$ 的实验结果中，基于相关约束的聚类方法与无约束的聚类方法差别不大。
2. 在 $P(Q|D)P(Q|Clu)P(E|D)$ 的实验结果中，基于相关约束的聚类方法要优于无约束的聚类方法。

5.6　本章小结

本章首先定义了实验的目标，数据集和评测指标，之后从实验出发，对基于聚类约束的高质量微博检索方法的性能加以验证，同时对其中的主要参数进行实验，从中分析找出参数的最优值。通过本章的实验可知，基于聚类约束的高质量微博检索方法性能优于基本检索方法。

结 论

现如今，整个社会高速发展，信息洪流不断席卷，利用数据挖掘技术对信息进行检索和排序成为必然趋势。检索问题主要受到用户意图理解困难的影响，因此，主要着重解决语义鸿沟，便可以有效提升检索性能。排序问题主要受到事件概括困难和文本去重困难的影响，因此，主要着重解决事件概括和从语义层面去重文本，才可以有效提升排序性能。并且，在面对长文本和短文本时，具体方法又会有所不同。因此，本文提出了基于聚类约束的高质量微博检索方法，主要贡献如下：

1. 提出了一种微博检索框架，探究了几种基本查询扩展方法对检索性能的影响。
2. 提出了一种多元检索模型，比较验证了该多元检索模型的检索性能。
3. 提出了一种基于非负矩阵分解的聚类方法（BNMF，Basic Non-negative Matrix Factorization），在聚类约束下提升了检索模型的检索性能。
4. 提出了一种基于相关约束的聚类方法（RNMF，Relevance Non-negative Matrix Factorization），对比于 BNMF，验证了该聚类方法的性能。

最后，通过各方面的实验，验证了本文算法性能，实验证明，基于聚类约束的高质量微博检索方法的性能优于基本方法，但是仍有很大的提升空间。在未来工作中，希望通过改进算法考虑更多因素，增加先验信息的利用，提升算法的性能和稳定性。

攻读硕士学位期间发表的学术论文

1. Kai Wang and Zhen Yang. BJUT at TREC 2016 Real-Time Summarization Track. NIST Special Publication 500-321: The Twenty-Fifth Text REtrieval Conference Proceedings (TREC 2016), 2016.
2. 杨震，王凯. 基于聚类信息的高质量微博检索方法. 专利申请号: 201810057738.X

参考文献

[1] Pasquier C, Gardès J. Prediction of mirna-disease associations with a vector space model. Sci Rep, 2016, 6:27036.

[2] Anastasopoulos A, Chiang D, Long D, et al. An unsupervised probability model for speech-to-translation alignment of low-resource languages. Conference on Empirical Methods in Natural Language Processing, 2016. 1255–1263.

[3] Richard A, Gall J. Temporal action detection using a statistical language model. Computer Vision and Pattern Recognition, 2016.

[4] Teevan J, Ramage D, Morris M R. #twittersearch: a comparison of microblog search and web search. Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February, 2011. 35–44.

[5] Salton G. A vector space model for automatic indexing. Communications of the Acm, 1975, 18(11):613–620.

[6] Castells P, Fernández M, Vallet D. An adaptation of the vector-space model for ontology-based information retrieval. IEEE Transactions on Knowledge & Data Engineering, 2007, 19(2):261–272.

[7] Bartell B T, Cottrell G W, Belew R K. Automatic combination of multiple ranked retrieval systems. International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994. 173–181.

[8] Robertson S E, Rijsbergen C J V, Porter M F. Probabilistic models of indexing and searching. ACM Conference on Research and Development in Information Retrieval, 1980. 35–56.

[9] Zhai C. Statistical language models for information retrieval. Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA, 2008. 3–4.

[10] Xu Y, Jones G J F, Wang B. Query dependent pseudo-relevance feedback based on wikipedia. 2009. 59–66.

[11] Yu S, Cai D, Wen J R, et al. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. Proc. International World Wide Web Conference, 2003. 11–18.

[12] Ganguly D, Bandyopadhyay A, Mitra M, et al. Retrievability of code mixed microblogs. International ACM SIGIR Conference on Research and Development in Information Retrieval, 2016. 973–976.

[13] Yan S, Yan S, Yan S, et al. Event classification in microblogs via social tracking. Acm Transactions on Intelligent Systems & Technology, 2017, 8(3):35.

[14] Kalloubi F, Nfaoui E H, Beqqali O E. Microblog semantic context retrieval system based on linked open data and graph-based theory. Expert Systems with Applications, 2016, 53:138–148.

[15] Chen Y, Zhang X, Li Z, et al. Search engine reinforced semi-supervised classification and graph-based summarization of microblogs. Neurocomputing, 2015, 152(C):274–286.

[16] Wang Y, Huang H, Feng C. Query expansion based on a feedback concept model for microblog retrieval. International Conference on World Wide Web, 2017. 559–568.

[17] Rudra K, Sharma A, Ganguly N, et al. Classifying information from microblogs during epidemics. International Conference on Digital Health, 2017. 104–108.

[18] Wu F, Shu J, Huang Y, et al. Social spammer and spam message co-detection in microblogging with social context regularization. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015. 1601–1610.

[19] Hirota S, Sasano R, Takamura H, et al. Real-time tweet selection for tv news programs. the International Conference, 2017. 299–305.

[20] Albishre K, Li Y, Xu Y. Effective pseudo-relevance for microblog retrieval. Australasian Computer Science Week Multiconference, 2017. 51.

[21] Jin P, Mu L, Zheng L, et al. News feature extraction for events on social network platforms. International Conference on World Wide Web Companion, 2017. 69–78.

[22] Wang Y, Lin J. Partitioning and segment organization strategies for real-time selective search on document streams. 2017. 221–230.

[23] Zhang S, Zhang S, Yen N Y, et al. The recommendation system of micro-blog topic based on user clustering. Mobile Networks & Applications, 2016. 1–12.

[24] Shi L, Zhao W X, Shen Y D. Local representative-based matrix factorization for cold-start recommendation. Acm Transactions on Information Systems, 2017, 36(2):1–28.

[25] Zhang G, Cai G, Wu H, et al. A nonnegative matrix tri-factorization technique for recommendation in microblog. International Conference on Machinery, Materials, Environment, Biotechnology and Computer, 2016.

[26] Xu B, Lu J, Huang G. A constrained non-negative matrix factorization in information retrieval. IEEE International Conference on Information Reuse and Integration, 2003. 273–277.

[27] Bing L, Cohen W W, Dhingra B. Using graphs of classifiers to impose declarative constraints on semi-supervised learning. Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017. 1454–1460.

[28] Pathak A, Gupta K, Mcauley J. Generating and personalizing bundle recommendations on steam. The International ACM SIGIR Conference, 2017. 1073–1076.

[29] Hu J, Li P. Decoupled collaborative ranking. International Conference on World Wide Web, 2017. 1321–1329.

[30] Sommer F, Lecron F, Fouss F. Recommender systems: the case of repeated interaction in matrix factorization. the International Conference, 2017. 843–847.

[31] Basu M, Roy A, Ghosh K, et al. A Novel Word Embedding Based Stemming Approach for Microblog Retrieval During Disasters. 2017.

[32] Liu M, Liu S, Zhu X, et al. An uncertainty-aware approach for exploratory microblog retrieval. IEEE Transactions on Visualization & Computer Graphics, 2016, 22(1):250.

[33] You S, Huang W, Mu X. Using event identification algorithm (eia) to improve microblog retrieval effectiveness. IEEE / Wic / ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2016. 122–125.

[34] Li H, Guan Y, Liu L, et al. Re-ranking for microblog retrieval via multiple graph model. Multimedia Tools & Applications, 2016, 75(15):8939–8954.

[35] Bansal P, Jain S, Varma V. Towards semantic retrieval of hashtags in microblogs. 2015, 58(58):7–8.

[36] Huifang M A, Jia M, Xiaohong L I, et al. A microblog recommendation method based on label correlation relationship. Computer Engineering, 2016..

[37] Manwar A B, Mahalle H S, Chinchkhede K D, et al. A vector space model for information retrieval: A matlab approach. Indian Journal of Computer Science and Engineering, 2012, 3(2).

[38] Kraft D H. Journal of the American Society for Information Science. ASIS, 1900: 421–421.

[39] Ponte J M. A language modeling approach to information retrieval. 1998. 275–281.

[40] Salton. The smart retrieval system—experiments in automatic document processing. Prentice-hall, Inc Upper Saddle River, 1971.

[41] Robertson S E, Jones K S. Relevance weighting of search terms. Journal of the Association for Information Science and Technology, 1976, 27(3):129–146.

[42] Robertson S, Zaragoza H. The probabilistic relevance framework: Bm25 and beyond. Foundations & Trendső in Information Retrieval, 2009, 3(4):333–389.

[43] 岑咏华, 韩哲, 季培培. 基于隐马尔科夫模型的中文术语识别研究. 现代图书情报技术, 2008, 24(12):54–58.

[44] Zhai C, Zhai C. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. ACM, 2017: 111–119.

[45] Carpineto C, Romano G. A survey of automatic query expansion in information retrieval. Acm Computing Surveys, 2012, 44(1):1.

[46] Zhai C, Lafferty J. Model-based feedback in the language modeling approach to information retrieval. Proceedings of the tenth international conference on Information and knowledge management. ACM, 2001. 403–410.

[47] Lv Y, Zhai C. A comparative study of methods for estimating query language models with pseudo feedback. Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009. 1895–1898.

[48] Harman D, Buckley C. The nrrc reliable information access (ria) workshop. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004. 528–529.

[49] Lee K S, Croft W B, Allan J. A cluster-based resampling method for pseudo-relevance feedback. International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008. 235–242.

[50] Salton G, Buckley C. Improving retrieval performance by relevance feedback. Morgan Kaufmann Publishers Inc., 1997.

[51] Liu X H, Smith A C. Semantic understanding and commonsense reasoning in an adaptive photo agent. Department of Electrical Engineering & Computer Science, Massachusetts Institute of Technology, 2002.

[52] Hassan S, Mihalcea R. Semantic relatedness using salient semantic analysis. AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, Usa, August, 2011.

[53] Bagdanov A D, Bertini M, Bimbo A D, et al. Semantic annotation and retrieval of video events using multimedia ontologies. International Conference on Semantic Computing, 2007. 713–720.

[54] Tu K, Meng M, Lee M W, et al. Joint video and text parsing for understanding events and answering queries. IEEE Multimedia, 2014, 21(2):42–70.

[55] 余先川, 任嘉勉, 张婷, et al. 非负矩阵分解及其应用研究综述. 全国地图学与 gis 学术会议, 2006.

[56] Packer H S, Samangooei S, Hare J S, et al. Event detection using twitter and structured semantic query expansion. International Workshop on Multimodal Crowd Sensing, 2012. 7–14.

[57] 李冰锋. 基于非负矩阵分解的图像聚类和标注方法研究. 2016.

[58] Liang F, Qiang R, Yang J. Exploiting real-time information retrieval in the microblogosphere. Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, 2012. 267–276.

[59] Cichocki A, Lee H, Kim Y D, et al. Non-negative matrix factorization with α-divergence. Pattern Recognition Letters, 2008, 29(9):1433–1440.

[60] Sandler R, Lindenbaum M. Nonnegative matrix factorization with earth mover's distance metric for image analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(8):1590–1602.

[61] Gupta M D, Xiao J. Non-negative matrix factorization as a feature selection tool for maximum margin classifiers. Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011. 2841–2848.

[62] Zafeiriou S, Tefas A, Buciu I, et al. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. IEEE Transactions on Neural Networks, 2006, 17(3):683–695.

[63] Sun P, Cha B R, Kim J W. Document summarization using nmf and pseudo relevance feedback based on k-means clustering. 2016, 35(3):744–760.

[64] Türkmen A C. A review of nonnegative matrix factorization methods for clustering. Computer Science, 2015, 1:V1–405–V1–408.

[65] Ding C, He X, Simon H D. On the equivalence of nonnegative matrix factorization and spectral clustering. Proceedings of the 2005 SIAM International Conference on Data Mining. SIAM, 2005. 606–610.

[66] Xu W, Liu X, Gong Y. Document clustering based on non-negative matrix factorization. SIGIR 2003: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada, 2003. 267–273.

[67] Xu B, Lu J, Huang G. A constrained non-negative matrix factorization in information retrieval. IEEE International Conference on Information Reuse and Integration, 2003. 273–277.

[68] Wang W. Squared euclidean distance based convolutive non-negative matrix factorization with multiplicative learning rules for audio pattern separation. Signal Processing and Information Technology, 2007 IEEE International Symposium on. IEEE, 2007. 347–352.

[69] Benetos E, Kotropoulos C. Non-negative tensor factorization applied to music genre classification. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(8):1955–1967.

[70] Zdunek R, Cichocki A. Non-negative matrix factorization with quasi-newton optimization. International Conference on Artificial Intelligence and Soft Computing. Springer, 2006. 870–879.

[71] Van Benthem M H, Keenan M R. Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. Journal of chemometrics, 2004, 18(10):441–450.

[72] Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics, 2007, 23(12):1495–1502.

[73] Berry M W, Browne M, Langville A N, et al. Algorithms and applications for approximate nonnegative matrix factorization. Computational statistics & data analysis, 2007, 52(1):155–173.

[74] Ulbrich I, Canagaratna M, Zhang Q, et al. Interpretation of organic components from positive matrix factorization of aerosol mass spectrometric data. Atmospheric Chemistry and Physics, 2009, 9(9):2891–2918.

[75] Fréin R, Drakakis K, Rickard S, et al. Analysis of financial data using non-negative matrix factorization. International Mathematical Forum, volume 3. Journals of Hikari Ltd, 2008. 1853–1870.

[76] Welling M, Weber M. Positive tensor factorization. Pattern Recognition Letters, 2001, 22(12):1255–1261.

[77] Zhang D, Zhou Z H, Chen S. Non-negative matrix factorization on kernels. PRICAI 2006: Trends in Artificial Intelligence, 2006. 404–412.

[78] Buciu I, Nikolaidis N, Pitas I. Nonnegative matrix factorization in polynomial feature space. IEEE Transactions on Neural Networks, 2008, 19(6):1090–1100.

[79] Kuhn H W, Tucker A W. Nonlinear programming. Berkeley Symposium on Mathematical Statistics and Probability, 1976. 481–492.

[80] Caitlin Sadowski G L. Simhash: Hash-based similarity detection. 2007..

[81] Tang J, Gao H, Hu X, et al. Exploiting homophily effect for trust prediction. ACM International Conference on Web Search and Data Mining, 2013. 53–62.

致 谢

至此，我的毕业论文已经进入收尾阶段，并且，我的研究生生涯也即将结束。人们在总结中前行，也在前行中总结。

首先，我要感谢我的导师杨震。在这三年中，杨老师给与了我很大的帮助，杨老师不仅为我提供了科研设备的支持，还在我的科研工作中提供了无与伦比的帮助。在杨老师的指导下，我的科研工作才能有效推进，并且在这个过程中，培养了我对科研工作的兴趣和能力，养成了思维严谨和工作认真的态度。杨老师在科研工作中硕果累累，指导我们的工作总能提出建设性的意见，不仅是良师，更是益友，在生活中，杨老师平易近人，言传身教，使我们耳濡目染于杨老师的生活态度。一个好的导师对研究生工作的影响是非常重大的，我为我能在杨老师的指导下完成研究生工作感到幸运和自豪。

然后，我要感谢我的家人，他们是我坚实的后盾，是我疲惫时候的动力，是我难过时候的支持。每次回家时候都会为我准备最喜欢的饭菜，在饭桌上，为我的开心而开心，为我的难过而担心。他们虽然都是普通的不能再普通的人们，但是却为我倾尽所有，谁言寸草心，报得三春晖。

最后，我要感谢我实验室的师兄师姐，对我入学以来的帮助，并且陪伴我度过了一段愉快的实验室时光。

时间如白驹过隙，在这三年中，不仅有科研，还有生活，不仅有成功，还有更多的挫折。生活当如此，希望在研究生阶段培养的科研兴趣不会止步于此。