硕 士 学 位 论 文

面向生物医学领域的跨语言信息检索

Research on Cross-Language Information Retrieval for Biomedicine

作 者 姓 名：宁健

学科、 专业：计算机应用技术

学　　号：20809316

指 导 教 师：林鸿飞 教授

完 成 日 期：2010.12

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

随着互联网的快速发展，人们非常依赖于互联网来获取资源。由于世界语言的多样化和用户所掌握语言的差异性，导致了自由获取互联网中的不同语言的信息具有很大的困难，所以对跨语言信息检索具有重大的理论和实际价值。

潜在语义分析模型在跨语言信息检索领域的应用取得了良好的效果，因为该方法能够很好的解决同义词和多义词带来的歧义问题，因为潜在语义分析能够将同义的但是属于不同语言的词汇映射到语义空间中距离相近的点上，在语义空间对文本之间的关系进行分析。然而由于潜在语义分析需要对原始词-文本空间进行降维，所以选取降维因子k存在一定的风险：如果k值很高，则达不到理想的降维效果；如果k值过低，则会损失很有有用的特征，对检索精度造成影响。

本文采用基于SVD和NMF矩阵分解相结合的改进潜在语义分析的方法为生物医学文献双语摘要进行建模，该模型将英汉双语摘要映射到同一语义空间，不需要外部词典和知识库，自动处理不同语言之间的对应关系，在双语空间中进行检索，并综合考虑两种矩阵分解结果。充分利用医学文献双语摘要语料中的锚信息，通过不同的 k 值构建多个检索模型，计算每个模型的信任度，使得多个模型都对查询和文本的相似度做出贡献。在语义空间上进行项与项、文本与文本、项与文本之间的相似度计算，实现了双语摘要的交叉检索，取得了较好的实验效果。

以往的搜索引擎计算文本之间关系的方法往往是计算文本相似度，而实际上文本相关性更侧重于强调文档关系的内在特征，更能代表文本之间的关系。为了提高检索精度，本文采用一种基于主题的文档检索模型，基于 LDA 模型的主题分布，从主题层面上对文档和查询进行建模，并且从文本相关性方面考虑文本之间的关系。为了弥补 LDA 模型参数估计较为粗糙问题产生的噪声带来的影响，本文又采用模型平均化的思想，构建多个潜在语义模型和基于 LDA 的文本相关度模型，综合考虑多个模型的检索结果。实验结果证明，该方法使得潜在语义模型对 LDA 模型起到了良好的平滑作用，提高了检索的召回率。

关键词：改进潜在语义分析；语义空间；双语语料；交叉检索； LDA 模型

Research on Cross-Language Information Retrieval for Biomedicine

Abstract

With the rapid development of Internet，people are very dependent on the Internet to obtain information resources. Because of the diversity of language in the world and language differences between different users，it is of great difficulty for users accessing to the information in different language on the Internet. Therefore，research on cross-language information retrieval will benefit a lot.

Because the Latent Semantic Indexing can map synonymous words in different languages in the proximity point in the semantic space， Latent Semantic Indexing model can well solve the synonym and polysemy problems which are caused by ambiguity of words. However， Latent Semantic Indexing requires to reduce the dimension of the original word-text space， so there is risk to select a certain dimension reduction factor k.

This paper describes a bilingual biomedical space model in which both Chinese and English abstract are represented using improved Latent Semantic Indexing with combined SVD and NMF matrix factorization method. The improved LSI-based method combined with location information and anchor information which classify the programs and sentences is used to improve the function of similarity in semantic space and combine the results of different matrix factorization. A set of k-dimension models is set up，under the help of which we can achieve Bilingual cross-language indexing. The experiment gets a better result.

The traditional text search engines are usually computing the text similarity to calculate relationship between different texts， but in fact， the correlation between texts is more focused on the intrinsic characteristics of the document. This paper takes advantage of a document retrieval model which is based on LDA distribution model. This model builds model for query and documents at the subject level， and considers the relationship between the texts in aspects of the text correlation to improve the retrieval accuracy. In order to compensate for the impact of noise，this article uses model averaging idea to construct several latent semantic text models and LDA-based text correlation models and to take advantage of the search results in different models. Experimental results show that the Latent Semantic Indexing model plays an important role on the smoothing of the LDA model， and the recall rate is improved.

Key Words：Improved Latent Semantic Indexing; Semantic Space; Bilingual Corpora; Cross-Language Indexing; LDA model

目　　录

1 绪论

1.1 问题的提出

随着互联网的迅速普及和搜索引擎的迅猛发展，人们可以方便和快捷的获取到所需要的信息和资源。单语的搜索引擎发展到现在已经十分成熟，能够满足大多数的人们单一语种的检索需求。但是语言问题已经成为互联网用户进一步获取更多信息的障碍[1]，因为当时的信息检索系统并不能很好的满足人们在不同语言之间的检索的需求。虽然这个问题很早就存在，但是由于信息检索的开始阶段，单语搜索引擎不是十分完善，人们进行单语检索的需求还没有得到很好的满足，所以人们的需求更多的是如何获取母语中的所需信息，所以跨语言检索的需求不是最强烈的需求。但是随着搜索引擎不断发展，人们进行的单语检索需求得到了很好的满足，所以人们才会更多的被如何检索不同语言之间的信息这个问题困扰，而使得克服语言的障碍这个问题越发的重要起来，进行跨语言检索的需求也变得越来越强烈。

当前社会各个领域都获得了迅猛的发展，交通运输和信息通讯变得越来越方便快捷，在现实中和网络中各个国家的人们互相交流和沟通越来越便捷和频繁，这使得不同国家的人们能够通过各种渠道进行沟通和交流，导致人们接触非母语的机会和需求也迅速增长，而且在互联网中，人们接触非母语的机会更多[2]，因为人们利用互联网进行跨国，跨语言的交流成本更小，交流更便捷。但是一个问题不容忽视，当前互联网中的资源主要还是以英文为主，据统计，互联网用户中，母语为非英语的用户比例在不断增加，但是互联网中，除了英语以外的其他几大语种的网络信息资源所占比例却很小，这导致了一个明显的问题，即母语为非英语的互联网用户很难检索到所需的信息和资源，这给这些非英语的互联网用户查询信息，获取英文资料，充分网络信息带来了很大的不便。比如一个母语为汉语的互联网用户，如果他不具备十分完善的英文阅读能力的话，当他面对互联网中数量巨大的英语网页时，往往显得很迷茫，因为他不能很好的检索出他想要的英文文献，用户自然就产生了强烈的跨语言信息检索的需求。人们以往开发出来的各种信息检索系统和搜索引擎不会考虑这种多元化语言的检索需求，所以以前的搜索引擎基本上都是针对单一语种的，这就导致了以前的检索系统不能满足当前人们日益增长的跨语言信息获取的需求，即不能解决一种语言的查询与其他语种信息之间的互检问题。

当用户希望利用检索系统检索所需信息是，人们都希望利用自己最熟悉的语言来构造查询来进行检索，从而获得资源，人们希望获取不仅是母语的信息，而且很有可能希

望获取多种语言的相关资源。用户获取信息的需求呈现国际化和多样化的特点，人们已经不满足于输入一个查询得到单语的资源；而是希望输入一个查询时，得到自己设定的一种或者多种语言的信息和资源。在某些特定的领域，比如论文检索、科学调研、国际化市场研究，数字图书馆等领域，单语查询检索到其他语言的需求更为迫切。在单语的信息检索系统中，如果用户不仅想获得母语的文献，同时要检索关于某个主题的不同语言文献，那么用户不得不需要分别构造多种语言的查询词，并且找到他所需要的几种不同语言的单语的信息检索系统，然后在各个系统上分别输入对应语言的查询关键词，分别进行检索。也就是说当用户想要利用信息检索系统获取信息和进行检索时，就必须首先掌握不同的语言，并且获得不同语言的检索系统，同时由于各个搜索引擎的用法不尽相同，搜索引擎的用户友好程度也有好有坏，严重影响用户检索的体验。如何满足人们日益增长的跨语言信息检索的需求，成了很多专家学者和互联网企业的研究课题或增长动力，并同时激励人们不断寻找如何跨越语言障碍实现跨语信息检索。

1.2 跨语言信息检索研究的作用及意义

随着经济全球化、信息国际化的不断发展，各国之间的经济文化等交流越来越密切，同时使得不同语言的人们的交流和合作越来越不可避免。然而很多人对其他语言并不熟悉和了解，不同语言之间的人们的沟通和交流存在很大的困难，不同语言的人们很难进行信息的分享和交流，并且很难有效的利用其他语言的信息和资源。如果能够消除语言障碍给人们带来的困扰，使人们能够在不同语言之间方便快捷的进行信息检索，使人们能够获取不同语言的优质资源和信息，并在此基础上帮助人们理解这些信息，毫无疑问，这对于对不同语言的人们进行资源的共享，对人们获取信息进行学习和生产，对社会的发展和文明的进步，将会产生重大的促进作用。解决这个问题的一个方法是搭建一个跨语言信息检索系统，这个思路是搭建一个中英文交叉语言检索的平台，在这个平台中，用户只需要用他们最熟悉的语言进行检索，而不需要用户掌握多门语言，这个平台会根据用户的需求，自动将查询转换成不用语言的检索查询词，然后使用该查询来检索出系统支持的多语文档，最终根据用户的语言需求返回给用户特定语言的所需资源。该方法能够使用户克服语言的障碍，不需要掌握多种语言，就可以得到需要的特定语言的有用信息，这将非常有利于更多的用户进行信息的共享和利用。

跨语言检索有重要的意义，首先，跨语言检索可以提高人们获取信息的范围，用户不仅可以获取到本语言的资源，同时可以获得其他语言的优质资源。其次，跨语言信息检索可以满足用户日益增强的获取知识的需求；当前单一语言的信息检索技术已经相当

成熟，人们可以通过单一语言的搜索引擎，方便快捷的获取单语的信息和资源；但是，获取不同语言的信息和资源的需求不能得到很好的满足，所以跨语言信息检索系统能够更好的满足用户在这方面的需求，实现全球范围内的资源共享和资源检索。在某些特定的领域，比如论文检索、科学调研、国际化市场研究，数字图书馆等领域，跨语言信息检索可以更好的满足这些领域专门的需求，更好的促进该领域的发展。再次，跨语言信息检索也属于信息检索的一个方面，对跨语言信息检索的研究和发展，可以促进信息检索的技术的发展丰富和完善。因而，对中英文跨语言信息检索进行研究都具有十分重要的意义[1]。

1.3 跨语言信息检索的研究现状

跨语言信息检索区别于单一语言信息检索系统的最主要的特点是：允许提问语言与文献语言之间属于不同类型的语言。也就是说，跨语言信息检索不同于以往的单一语言检索系统，它是这样一种检索，它允许用户以一种语言构造检索查询语句，然后该信息检索系统利用该查询语句检索出跨语言信息检索系统所支持的其它语言写成的文本或者其他形式的文献。针对跨语言信息检索的研究，最早是在上个世纪 60 年代，有些专家和学者就已经提出跨语言信息检索的思想和方法，并且开始开发研究跨语言信息检索系统。为了提高国际联机检索的质量，使人们能够利用和理解国外的文献信息资源，跨语言信息检索技术开始被关注并应用到国际联机检索中。最初的跨语言信息检索的基本思想是利用一个词表，进行不同语言之间的对照，这个词表是进行跨语言信息检索的关键因素。因为跨语言信息检索需要解决不同语言之间的不匹配问题，所以如果构建一个包含不同语言的词汇之间相互对照的词表的话，就可以很好的对不同语言之间的词语进行配对，从而能够达到跨语言信息检索的目的。比如该词表将英文的关键词“computer”和中文的“电脑”进行了配对，也就是认为英文的关键词和中文的关键词的意思是相同的，所以当用户的查询为“computer”时并同时希望检索到中文的关键词，跨语言信息检索系统，会自动对该关键词进行匹配，从而找到“电脑”这个关键词，该系统会利用“电脑”这个关键词进行查找和结果的返回，这个返回给用户的文档为中文的结果。60 年代末 70 年代初，康奈尔大学的 Salton[2]教授利用手工编制的英语-德语双语词表，进行了英德跨语言检索的尝试。在同一时期，加州大学的 Pevzner[4]教授也进行了跨语言信息检索方面的尝试，他采用布尔模型的方法进行了英语-俄语之间的跨语言信息检索实验，他利用一个受控词表（Controlled Vocabulary）进行词语转换，达到了较好的试验效果。进入八十年代后，受控词表在跨语言信息检索系统中很快得到了应用。

受控词表的意义在于，通过界定其内涵，使得信息检索能够明确索引词之间的逻辑关系，该检索系统能够明确检索词和索引词之间的层次关系，该检索系统并不基于字符匹配，而是基于语义层次，从而能够克服基于字符匹配的一系列问题。由于受控词表本身的存在一定问题，具有一定的局限性，使得虽然基于受控词表的跨语言信息检索技术得到了较好的发展，但是基于该方法的跨语言信息检索系统不能得到很好的应用和推广。首先，用户的利用方便程度很低，因为该系统需要利用复杂的词表进行不同语言之间的对照，用户想要利用该系统进行跨语言检索，需要很多操作，一般的用户不经过学习和训练，很难迅速掌握检索的操作方法，并且不同系统的词表编制规则和使用规则不尽相同，大大影响了用户体验。其次，不同语言词表的编制主要由人工完成，而且词表本身非常复杂，使得词表的编制工作非常困难，而且很难利用计算机自动生成词表，人工编制复杂的词表需要耗费大量的人力，而且词典规模也非常有限，不可能涵盖所有的词汇。同时，人工编制的词表容易受到编制者文化水平和主管倾向的方面的影响，不能保证词表的客观性。再次，双语、多语受控词表由于由人工生成，所以导致更新速度上出现问题，人们很难每天都对该词表都进行更新，但是现在的互联网上每天都会出现很多新鲜的词汇，这就导致很多词汇不能被收录到词表中，从而影响了检索效果。此外，由于不同系统所编制的受控词表往往不尽相同[5]，用户往往不熟悉双语、多语受控词表的用法。在当前的跨语言信息检索研究中，查询扩展[6, 7]可以在一定程度上缓解这种情。

进入 90 年代，人们开始探讨跨语言自由文本信息检索的可能性，这种方法不需要受控的词表也能进行跨语言信息检索。经过跨语言信息检索领域和自然语言处理领域的研究人员的努力，跨语言信息检索领域已经发生很大的变化，取得了很大的进展。90 年代初，Littman 等人[8, 9]提出了一种技术“跨语言潜在语义索引”（Cross-Language Latent Semantic Indexing，简称 CL-LSI）技术，跨语言潜在语义分析可以自动处理对齐语料，将相同语义不同语言的词汇对应到潜在语义空间中距离相近的点上，不需要外部词典和词表就能够进行跨语言的信息检索。

1.4 重要国际会议

近年来，跨语言信息检索的研究受到了众多专家和学者的重视，随着检索技术的迅速发展以及不同国家不同语言的人们交流趋势的加快，跨语言信息检索的研究已成为信息检索领域研究的热点问题。为了更好的促进跨语言信息检索技术的发展，每年国际上都会举行几次关于跨语言信息检索的专题会议，下面介绍几个影响力最大的跨语言信息检索相关会议。

（1）文本检索会议（Text Retrieval Conference， 简称 TREC）。其旨在促进大规模文本检索领域的研究，加速研究成果向商业应用的转化，促进学术研究机构、商业团体和政府部门之间的交流与合作。TREC 是信息检索领域最具权威的年度测评活动。TREC 中跨语言信息检索专题的主要目标是：第一，为跨语言信息检索技术的研究成果提供评价标准，并对评价标准进行完善和维护；第二，TREC 会议的专家和学者一直在调研有效的跨语言检索的各种方法和评价体系；第三，TREC 会议为了促进跨语言信息检索的发展，为提供一个供世界各地跨语言信息检索研究者共同交流的平台。

（2）日本国家科学信息系统中心信息检索系统测试集会议，该会议的研究主要侧重英语与本国语言之间的跨语言信息检索问题，如何在完全不同结构和来源的语言之间的转换是研究的主要难点。该会议于 1999 年 8 月在日本东京召开第一次会议，然后每年举办一届，该会议除了跨语言信息检索之外，还包括一些其他的研究项目和问题，如专家系统、问答系统（Question Answering）、Web 检索等。NTCIR-1、2 开始日-英的跨语言信息检索，从 NTCIR-3 开始，主要研究面向多语言的信息检索系统和单语信息检索系统的性能比较等。

（3）跨语言信息检索评价论坛（Cross-Language Evaluation Forum，简称 CLEF）。该会议每年举办一次，侧重点为欧洲各个语言之间的跨语言信息检索问题，该会议为跨语言信息检索的测试和评估提供标准的测试集，同时对跨语言信息检索进行评估和研究。该会议从 2000 年起每年举办一次，至今已有九届，每次会议围绕该目标侧重几个不同的主题，主要侧重于欧洲范围内跨语言检索问题的评价，而且当前针对亚洲语系的研究也可以成为研究的一个热点。

1.5 跨语言信息检索实现方法

跨语言信息检索系统一般包含以下三个步骤：

（1） 查询翻译[10]：把查询从一种语言翻译成另一种或几种目标语言；

（2） 应用传统的单语信息检索技术实现查询与文档的匹配；

（3） 合并各个语言的结果，过滤掉阈值一下的结果，并将符合条件的结果，根据用户的语言需求，返回给用户。

其中，步骤（1）是实现跨语言信息检索系统的关键。目前，目标语言与源语言之间的统一主要有三种方式：根据翻译时所采用的资源不同现有的跨语言检索方法可分为基于词典基于语料和基于机器翻译模块的方法将目标语言表示的文档。

（1） 基于词典的方法[11]。基于词典的方法是最为常用的查询翻译方法，查询翻译方法是利用传统的单语检索技术在目标语言文档集上进行检索。其基本思想是自动从一部

在线双语词典中选择合适的翻译来替换每一个查询词。它可以很容易地与传统单语信息检索技术紧密地结合，这种方式是目前实现跨语言信息检索系统的主流思想，而且由于仅需对查询进行翻译，所以工作量较小。这种基于词典的方法存在一些问题，非常主要的问题是未登录词（OOV 词）的识别，解决这一问题的方法有：忽略掉查询中出现的所有 OOV 词[12]，音译方法 [13]，基于锚文本的方法 [14]和基于 Web 的方法 [15, 16, 17, 18, 19]等等。

（2） 基于语料库方法[20, 21]。早期研究使用语料库的目的主要是想从平行语料库中提取能够表示词的用法的统计信息从而自动构造词典。因为语料中存在大量的有用信息，如果能对语料进行有效的处理，并且提取出这些信息，或者利用这些信息的特征和统计特点进行训练，获得一个知识集，则能够提高见词典的宽度和广度。而统计检索技术的兴起则使研究者们放弃了这种舍近求远的方法而直接利用语料库知识来进行跨语言信息检索。

基于语料库的方法根据双语对其的程度可以分为词对齐句子对齐文档对齐和不对齐四种。词对齐的语料库是对齐精度最高的语料，这种语料能够比较容易的处理，方便的提取有用信息，但是基于词对齐的语料由于对齐精度非常高，在现实的情况下，较难获得。句子对齐的语料对齐精度也较高，利用该类语料进行处理和分析的结果也比较理想，同时这类句子对齐的语料比较容易获取，所以很多研究和应用都是针对句子对齐的语料，比如论文的中英文语言的摘要，属于句子对齐。

（3） 基于机器翻译模块的方法。实际上基于机器翻译模块的方法与前面两种方法没有本质区别我们可以根据机器翻译实现的方法将其归入前面两类中。在文档翻译方法中，机器翻译系统能够提取索引词而且能够将文档信息翻译成查询语种。总的来说，但由于机器翻译针对两种语言之间的自动翻译问题做了大量细致深入的研究工作因此从理论上来说机器翻译系统的效果要比现有的跨语言检索所采用的方法更好。

（4）基于本体的方法(Ontology-based Approach)[22, 23]。本体是对现实事物的一种抽象，通过建立本体关系，可以更好的帮助搜索引擎理解用户需求，更好的提高检索的效果，因为本体相当于一个知识库，搜索引擎可以利用这个知识库理解用户的需求，同时对用户的检索进行推荐等应用。本体中会包含很多关系，比如包含关系：动物—>老虎，狮子，羚羊；比如同位关系：诺基亚 5800—>诺基亚 5610，诺基亚 N97，诺基亚 7610。这些关系对搜索引擎理解用户需求能够起到良好的效果。例如用户的查询是“诺基亚 5800”，如果本体知识库中存在“诺基亚 5800”的对应关系，比如“诺基亚 5800 报价”，“诺基亚 5800 论坛”，此类关系的话，搜索引擎可以同时对这些关系进行检索，搜索

引擎所得到的检索结果不单单是“诺 基亚 5800”关于手机信息的结果，搜索引擎还能得到“诺基亚 5800”非常相关的结果和页面，比如“诺基亚 5800”论坛等等，因为搜索引擎返回的结果更丰富，所以提高了用户的满意度。同时搜索引擎还能对用户的检索查询词进行推荐，比如推荐“诺基亚 5610”，“诺基亚 N97”，“诺基亚 7610”，这些词作为查询词，因为用户如果对某一型号的诺基亚手机感兴趣的话，很有可能对诺基亚的其他型号的手机也感兴趣。这样就相当于搜索引擎，利用本体知识库，更好的了解了用户的需求，更好的满足了用户的需求。现在很多搜索引擎开始构建本体知识库，提高检索的效果，比如 WORDNET 等。

1.6 论文结构

本文的主要内容包括：

第二章：主要介绍了基于潜在语义分析的跨语言检索方法。详细介绍了基于 SVD 矩阵分解的潜在语义分析方法，同时引入了一种基于 NMF 的矩阵分解方法，该矩阵分解方法具有明确的物理意义，并且能够较好的弥补 SVD 矩阵分解方法造成的特征损失。

第三章：引入了一种模型平均化的潜在语义分析方法，该方法选取不同的 k 值，建立多个模型，对于特定查询来说，MALSI 允许每个模型都对相关文档的相似度”投票”，给每个模型都赋予一个信任度进行计算总的文档相似度，从而补偿了单一模型的风险。

第四章：提出了一种基于主题的文档检索模型，该模型从文本的相似性和文本相关性两个方面考虑文本之间的关系，得到了较好的试验效果。

第五章：总结全文并对以后的研究方向进行展望。

2 基于潜在语义分析的跨语言检索

2.1 向量空间模型的问题

随着互联网的迅猛发展，互联网上的信息和资源爆炸性的增长，如何在浩如烟海的信息之中，查找用户有用的信息，是搜索引擎的主要工作。为了满足用户日益增长的检索需求，越来越多的搜索引擎出现和不断发展晚上，来满足用户的需求。向量空间模型取得了很大的成功，基本上实现了搜索引擎的主体功能，即文档的查询，促进了搜索引擎和信息检索系统的迅速发展，当代大多数搜索引擎的基本框架和搜索理论都是基于向量空间模型进行的完善和改进。

向量空间模型最重要的部分是建立倒排索引，所谓倒排索引，就是关键词作为索引，每个关键词对应文档号或者网页号，关键词所对应的文档号表示该文档或者网页中包含这个关键词。倒排索引的好处是在用户进行查询之后，通过对查询进行分词和相关处理，可以得到几个关键词，即查询关键词，然后搜索引擎利用这些查询关键在倒排索引中进行关键词的匹配，如果匹配成功，则计算将查询向量化之后的向量和关键词所对应的文档向量之间的相似度，计算相似度的方法各有不同，最简单、有效的方法是计算两个向量的夹角的余弦值。通过计算得到所有查询向量和所有对应文档向量的相似度，然后根据相似度大小进行排序，返回相似度符合阈值的对应文档、网页，或者返回相似度最大的前多少个文档、网页。

通过对向量空间的基本过程进行分析，我们发现向量空间模型的一个很重要的步骤是匹配查询关键词和索引关键词。当前主流的搜索引擎系统，主要是基于查询词和索引词的匹配，由于多义词和同义词现象导致不可避免的产生不匹配现象的发生。而且即使在分词的时候，同样会产生歧义，而这种歧义同样会对查询词和索引词之间的匹配产生影响。比如“独立自主和平等互利的原则”，应该分成“独立自主”、“和”、“平等互利”、“的”，“原则”。但是不同的分词方式可能会产生不同的分词效果。所以这句话很有可能被切分成了“独立自主”，“和平”，“等”，“互利”，“的”，“原则”。如果这句话被切分成了下面的形式，则当用户检索“平等互利”的时候，是不能检索到该篇文档的。因为搜索引擎室基于查询词和索引词进行检索的，由于该篇文档在被切分后，没有“平等互利”这个词汇，在搜索引擎健索引的时候，“平等互利”这个索引词后面的对应的倒排索引中，并没有该篇文档，所以当用户输入“平等互利”时，并不能检索出该篇文档。虽然切词技术发展至今已经很成熟，不过由于自然语言的多样

性和语义的复杂性，对歧义词的切分产生错误是不课避免的，而这种错误也会影响基于查询词和索引词进行检索的搜索引擎的准确率和召回率。

通过上面的分析我们知道，基于查询词和索引词匹配的搜索引擎和信息检索系统，之所以准确率和召回率收到严重的影响，本质是因为没有考虑语义的信息，也没有考虑用户的检索需求，为了解决这种不可避免的问题，Dumais 等人提出了一种基于潜在语义进行检索的方法，该方法不是基于查询词与关键词进行匹配，不需要建立关键词和文档之间的倒排索引，而是构造一个关键词-文档对应的矩阵，矩阵的元素为关键词在文档中的权重，利用严谨的数学方法对矩阵进行处理，对关键词和文档在语义层面上进行分析和处理，构建了一个潜在的语义空间，将文档和关键词映射到的语义空间中对应的点上，具有相同语义的关键词和文档，不管是否具有相同的形式，都会在这个语义空间中，映射到位置相近的点上；语义不同的关键词和文档，即使具有相同的形式和结构，也会被映射到语义空间中位置相隔很远的点上；并且将查询词作为一个文本，也映射到语义空间中相应位置的点上，然后在语义的空间中，计算查询和文档之间的相似度，由于该方法考虑了语义的关系，而不是基于查询词和索引词的简单匹配，所以有效的避免了同义词和多义词问题，提高了检索的准确率和召回率，下面我们对这种方法进行讨论和分析。

2.2 潜在语义分析概论

潜在语义索引简称为LSI[8, 9](Latent Semantic Indexing)或者LSA(Latent Semantic Analysis)。潜在语义分析首先需要构造一个关键词-文档对应的高维度矩阵，矩阵的维度有关键词的个数和文档的个数共同决定。矩阵的行数为文档的个数，矩阵的列数为所有文档中所有关键词的个数。矩阵的元素表示关键词在对应文档中的权重。计算权重的方法有很多中，比较简单的方法是直接根据关键词在文档中出现的频率，还有一种简单有效的计算权重的方法是计算关键词的频率*逆文档频率值。关键词-文档构成的矩阵和矩阵处理技术是LSI的基本要素。LSI将关键词=文档矩阵进行矩阵的SVD分解从而获得潜在的语义结构模型。奇异值分解singular value decomposition (SVD)，是一种重要的矩阵分解技术，这种矩阵分解技术能够将任意矩阵分解成三个矩阵相乘的形式。奇异值分解SVD是数学和统计数学的众多矩阵处理方法的一个，在众多领域和其他严谨的数学方法有着广泛的应用。

关键词和文档构成的相关矩阵是矩阵分解的基础，不同的查询和文档之间可以通过关键词-文档矩阵中不同文档所共有的关键词数来衡量查询和文档之间的紧密程度。奇异值分解方法把一个关键词-文档矩阵分解为三个矩阵乘积的形式。中间的矩阵是一个

对角矩阵，每一行中的元素为一个奇异值。可以通过分解之后的三个矩阵进行数学处理，和相似度的计算来得到查询词和文档之间的相似度，这种方法比对原始的关键词-文档矩阵进行数学计算有一定的优势。优势之一就是可以通过降低维度来简化运算。由于信息检索系统所要处理的文档数量很大，通过奇异值分解之后，可以形成三个矩阵，其中中间的矩阵为对角矩阵，通过选择一个特定的维度，在这个维度之后的对角矩阵的元素都设置为0，通过这种方式，可以分解之后的另外两个矩阵中的元素的维度降低，从而将高维度的关键词-文档矩阵，对这些矩阵进行处理，可以大大提高检索的效率。

在关键词-文本矩阵进行降低维度后形成的三个新的矩阵中，这三个降低维度之后的新的矩阵相当于将原来的关键词-文本空间，转会为新的潜在语义空间，通过对这三个矩阵进行数学处理，比如计算夹角的余弦值，或者计算向量之间的内积，相当于是在新的语义空间中计算关键词和文档之间的距离。可以对查询向量进行处理，将其映射到降低维度之后的潜在语义空间中，在这个新的空间中，查询向量被映射为空间中的一个点，通过求与这个点距离相近的点的文档，相当于是在计算与查询向量相似度最高的文档号，就相当于进行了检索，而这个过程不需要建立关键词和文档的索引，不需要进行查询词和索引关键词的匹配，而是基于语义的信息，进行潜在语义空间的相似度的计算。

下面讨论详细的矩阵分解理论，用严谨的数学方法分析和证明，为潜在语义分析在各个领域的应用奠定了坚实的理论基础。

2.3 SVD 模型

2.3.1 基本理论

任何一个矩阵 X，潜在语义分析用到的关键词-文档矩阵都可以分解为下列形式，

$$X = T_0 S_0 D_0^T \tag{2.1}$$

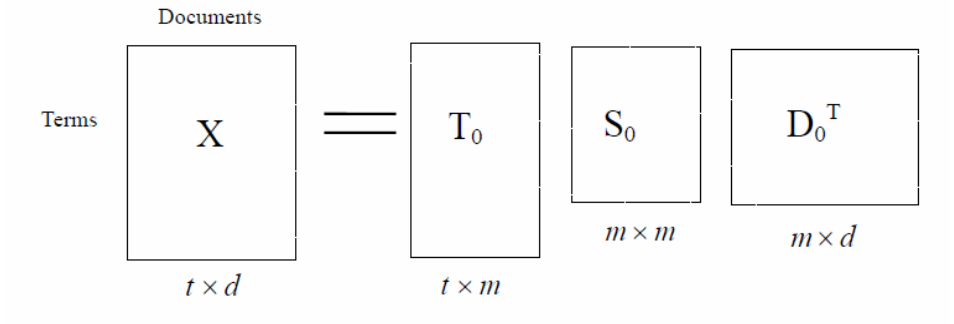

图2.1 奇异值分解

Fig.2.1 Singular Value Decomposition