智能信息处理与知识服务丛书

丛书主编 / 何婷婷

基于概念的信息检索方法

涂新辉 / 著

华中师范大学出版社

智能信息处理与

丛书

基于概念的信息检索方法

涂新辉 / 著

華中師範大學出版社

新出图证(鄂)字 10 号
图书在版编目(CIP)数据

基于概念的信息检索方法/涂新辉著. —武汉：华中师范大学出版社，2015.3
ISBN 978-7-5622-6929-8

Ⅰ.①基… Ⅱ.①涂… Ⅲ.①概念—情报检索 Ⅳ.①G252.7

中国版本图书馆 CIP 数据核字(2015)第 060355 号

基于概念的信息检索方法

责任编辑：陈 艳 袁正科　　责任校对：刘 峥　　封面设计：罗明波
编 辑 室：第二编辑室　　电 话：027—67867362
出版发行：华中师范大学出版社
社 址：湖北省武汉市珞喻路 152 号　　邮 编：430079
销售电话：027—67863426/67863280
邮购电话：027—67861321　　传 真：027—67863291
网 址：http://www.ccnupress.com　　电子信箱：hscbs@public.wh.hb.cn
督 印：王兴平　　印 刷：虎彩印艺股份有限公司
开 本：710mm×1000mm 1/16　　印 张：8.25　　字 数：140 千字
版 次：2015 年 4 月第 1 版　　印 次：2015 年 4 月第 1 次印刷
定 价：18.00 元

前　言

随着互联网的快速普及以及廉价大容量存储设备的不断出现，人类社会已经产生了海量的数字化文档信息。这些数量惊人的数字化文档可谓是人类知识的一个重要的宝库，同时也使得人们越来越依赖信息检索系统来找到所需要的信息。在传统的信息检索模型中，通常使用“词袋”模型来表征文档和查询。然而，人类的自然语言是一个异常复杂的符号系统，词语之间存在包括同义、歧义、语义相关等错综复杂的关系。简单的“词袋”模型忽视了词语之间的丰富的语义关系，远远不能够表征自然语言中所蕴含的复杂的语义信息。

本书中的概念是指描述意义的一个基本单元。人类理解自然语言的过程是一个语义概念的联想和关联的过程，这种功能是由人类大脑中几百亿个神经元构成的复杂生理组织所提供的。把和文本主题相关的概念中所蕴含的语义知识融合到文档和查询表征模型中将是构建基于语义的信息检索模型的一个途径，也是本书中重点要解决的问题。本书针对基于概念的文本信息检索系统的各个方面进行了系统的研究，包括文本的概念标注的构建、概念的语义表征模型、基于概念的文档和查询模型平滑方法以及基于概念间语义相关性的检索模型。本书中的主要内容概括如下：

一、提出了构建文本的概念标注的方法。根据所处理文本对象的不同，提出了几种不同的概念标注方法。针对某些专业领域文档集合，可以直接利用其中专家标注的概念。在通常情况下，文档中没有直接可以利用的概念标注。我们提出了一种利用维基百科文档建立通用概念库，并利用基于排序学习的方法来实现文档的维基百科概念自动标注的方法。另外，中文维基百科的质量和数量均难以满足通用概念库的要求，上面提出基于维基百科概念的方法难以应用于中文文本集，因此我们还提出了一种自动概念抽取的方法为中文文本标注概念。

二、提出了构建概念的语义表征模型的方法。针对不同类型的概念系统，分别提出了不同的解决方法。针对专业辞典中的概念，提出了一种基于互信息的概念表征方法。针对维基百科概念，提出了一种基于混合模型的表征方法和一种基于互信息的概念表征方法。针对中文文本中自动抽取的概念，提出一种基于概念间语义相关度的方法。

三、提出了一种新的基于概念的文档模型平滑方法。通过把概念的语义信息有机地整合到基于“词袋”的文档模型来建立基于语义平滑的文档表征模型。为了验证方法的有效性，在包括专业领域文献和新闻文本的几种标准信息检索测试集上进行了测试。实验表明，相对于传统的语言模型方法，这种方法的检索效果有了显著的提高。

四、提出了一种新的基于概念的查询模型平滑方法。在为查询标注相关概念的过程中，分别提出了两种不同的方法：基于伪相关反馈文档中标注的概念来建立查询的概念表征模型的方法，以及基于交互式信息检索的概念选择的方法。在包括专业领域文献和新闻文本的几种标准信息检索测试集上进行的测试表明，相对于传统的语言模型方法，这种方法的检索效果有了显著的提高，进一步验证了这种方法的有效性。

五、提出了一种利用概念间语义关系的中文检索模型。利用排序学习方法有机地整合包括概念间语义相关度等各种特征项，建立更加有效的信息检索模型。在包含不同来源新闻报道的国际标准中文文本信息检索测试集 NTCIR 上进行了测试。实验结果表明，相对于传统的基于 BM25 模型的方法，这种方法的检索效果有了显著的提高。

涂新辉

2015 年 1 月 20 日

于武汉桂子山

目　录

第1章 绪 论

1.1 研究目的及意义

随着互联网的快速普及以及廉价大容量存储设备的不断出现，人类社会已经产生了海量的数字化信息。世界上最大的搜索引擎公司 Google 于 2008 年在官方网站上公布的数据表明，其索引的网页数已经超过了一万亿，并且每天都在以几十亿的数量递增。这还只是冰山一角，Web 上还有大量的网页是搜索引擎爬虫无法访问的。这些数量惊人的数字化信息可谓是人类知识的一个重要的宝库。而另一方面，这些数字化信息的数量如此之大，以至于即使是从一个经过筛选的非常小的子集中寻找答案也变成一项让人望而生畏的任务。因此，如何利用计算机技术更加智能地获取需要的信息成为一个十分重要的议题。

信息检索就是专门研究信息的结构化、分析、组织、存储和获取的领域。信息检索处理的对象可以是文本、图像、音频和视频。信息检索领域包括很多不同的任务，如文本检索、文本分类和聚类等。文本检索是信息检索中最重要的任务，也是本书中将研究的核心问题。文本信息检索的目的是帮助用户获取和查询相关的文本信息。文本检索系统的任务可以简单地描述如下：系统的输入是用户通过自然语言描述的信息需求；系统的输出为多个与查询相关的文档列表，列表中的文档根据其查询的相关度排列，相关度较高的在前面。

传统的信息检索方法通常使用基于“词袋”的文本表征模型。在这种模型中，文本通常被表示为词语项的集合，词语项是互相独立的无序的单位，构成文档的基本特征。因为这种模型实现起来非常简单，所以已经被大量应用于各种文本处理系统中。然而，这种模型在表征文本时存在很明显的缺陷。人类的自然语言是一个异常复杂的符号系统，词语之间存在包括同义、歧义、语义相关等错综复杂的关系。简单的“词袋”模型忽视了词语之间丰富的语义关系，远

远不能够表征自然语言中所蕴含的复杂的语义信息。人类理解自然语言的过程是一个语义概念的联想和关联的过程，这种功能是由人类大脑中几百亿个神经元构成的复杂概念网络所提供的。把文本中的概念所蕴含的语义知识融合到文本表征模型中将是构建基于语义的信息检索模型的一个途径。图 1.1 和图 1.2 分别给出了人脑中基于概念的文本表征和目前检索系统中基于“词袋”的文本表征方法的示例。

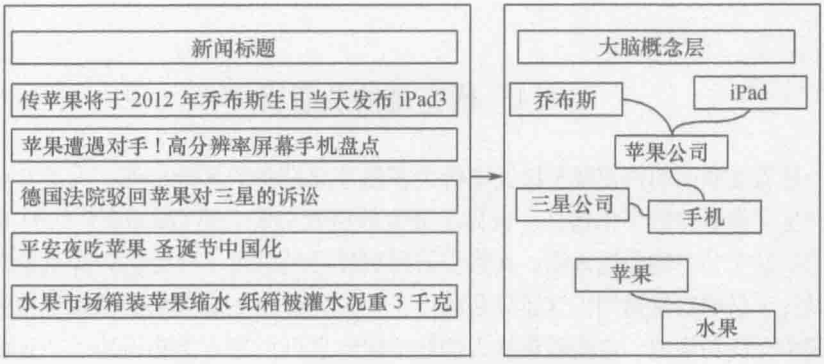

图 1.1　人脑中基于概念的文本表征

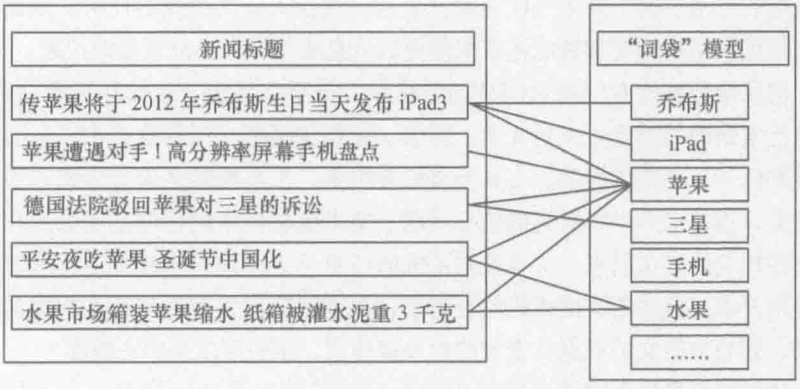

图 1.2　目前文本检索系统中的“词袋”模型

本书中的研究目的包括两个方面：(1)构建文本和查询的概念表征模型，能捕获语言中的语义信息，建立比“词袋”模型更加精确的文本表征；(2)把概念表征模型中蕴含的语义信息平滑融入传统的信息检索方法中，提高信息检索系统的性能，使得查询检索过程更接近于人类信息检索的本质。

在互联网“信息大爆炸”的大背景下，人们将越来越依赖信息处理系统帮助发现和组织信息。因此，对信息处理系统中的基于概念的表征模型以及检索方法进行深入的研究具有重要的理论意义和实用价值。虽然构建一种能够描述自然语言中蕴含的语义信息的文本表征模型将是一项非常具有挑战性的任务，但是这个领域40多年的研究工作为这个任务打下了坚实的基础，也给了我们很多灵感和启发。本书中的工作主要是对基于概念的信息检索系统的各个方面的问题进行了一些分析和探索，希望能够为信息检索领域贡献一份自己的力量。

1.2 主要内容及基本结构

本书中的主要内容是基于概念的检索系统的构建，图1.3给出了其基本结构。根据系统的结构，本书中的内容又可以分为以下几个部分：(1)如何构建文本的概念标注；(2)如何构建概念的语义表征模型；(3)如何利用基于概念的文档表征提高信息检索的效果；(4)如何利用基于概念的查询表征提高信息检索的效果；(5)如何利用概念间语义相关性提高信息检索的效果。

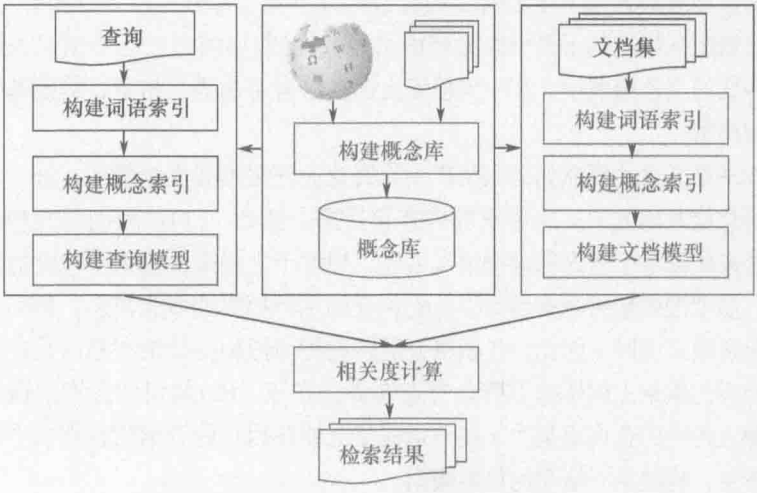

图1.3 基于概念的检索系统的结构

本书第3章将研究如何利用文本主题密切相关的概念构建文本的概念标注。根据所处理文本对象的不同，研究内容包括以下几个方面：(1)在一些如医学、法律之类的专业领域的英文文档集中，领域专家可以使用若干个概念来对每一

篇文档进行标注，这些专家标注的概念可以直接被作为文档的概念标注；(2)通常情况下的英文文档集中并没有可以直接利用的标注好的概念，因此我们将研究文档的概念自动标注问题。主要研究内容包括两个部分：如何利用英文维基百科文档集构建一个通用的概念库，以及如何利用机器学习方法自动为文本标注和文本主题密切相关的概念；(3)中文维基百科的质量和数量均难以满足通用概念库的要求，因此上面提出基于维基百科的概念标注方法难以应用于中文文本集，我们将研究如何从中文文本中自动抽取概念来构建文本的概念标注。

本书第4章将研究如何利用统计模型方法计算得到概念的相关词语及其相关度，构建概念的语义表征，这些工作是构建基于概念的信息检索模型的基础。针对不同类型的概念系统，研究内容包括以下几个方面：(1)如何建立领域本体中概念的表征方法；(2)如何利用统计方法构建维基百科概念的语义表征；(3)针对中文文本中自动抽取的概念，如何表征这些概念间的语义关联度。

本书第5章将研究如何利用文档的概念表征来增强文档模型，进一步提高文本信息检索的效果。主要研究内容包括以下三个部分：(1)如何利用文档中标注的概念来构建基于概念的文档语义表征；(2)如何把文档的语义表征平滑融入传统的文档模型中，在一定程度上捕获语言中的语义信息，构建基于概念的检索模型。

本书第6章将研究如何利用查询的概念表征来增强文档模型，进一步提高文本信息检索的效果。主要研究内容包括两个部分：(1)如何利用文档中标注的概念来构建基于概念的查询语义表征。相对于文档来说查询所包含的词语更加少，很难准确地将查询自动映射到和查询主题相关的概念集合，特别是当查询中包含歧义词时。因此，我们将分别研究利用伪相关反馈文档以及在交互式信息检索的框架上构建基于概念的查询语义表征。(2)如何把查询的语义表征平滑融入传统的查询模型中，在一定程度上捕获用户的查询信息需求中蕴含的语义信息，构建基于概念的检索模型。

本书第7章将研究如何把概念间相关性信息融入传统检索模型中以提高文本信息检索的效能。主要研究内容包括两个部分：(1)如何利用概念间相关度构建文本-查询的相关性特征；(2)如何有机地整合包括汉字、二元和概念相关度的多种特征构建更好的检索模型，进一步提高检索的精度。

1.3 组织结构

本书的组织结构如下：

第1章 绪论——介绍了本书中相关研究的目的和意义，以及本书的主要内容。

第2章 相关研究工作概述——对本书中相关的研究背景进行了详细的介绍，并进行了分析。

第3章 文本的概念标注方法——根据所处理文本对象的不同，介绍了三种不同的为文本标注概念的方法：基于领域本体概念的方法，基于维基百科概念的方法，以及基于自动抽取概念的方法，本章中的研究工作是构建基于概念的文档语义表征模型的基础。

第4章 概念的语义表征模型——针对各类概念系统，介绍了几种不同的概念语义表征方法。其中包括基于相关词群的专业领域概念表征模型，基于相关词群的维基百科概念表征模型，以及基于概念间相关性的中文概念表征模型。

第5章 基于概念的文档平滑方法——介绍了一种利用文档中标注的概念以及概念的语义表征建立基于概念的文档语义表征模型，并平滑地融入传统的基于“词袋”的文档模型方法中以进一步提高信息检索效能的方法。

第6章 基于概念的查询平滑方法——介绍了几种为查询标注相关概念的方法，并提出了一种利用查询标注的概念以及概念的语义表征建立基于概念的文档语义表征模型，并平滑地融入传统的文档模型方法中以进一步提高信息检索效能的方法。

第7章 基于概念相关度的中文检索模型——介绍了一种基于排序学习思想，把概念间相关性信息有机地融入传统检索模型以提高文本信息检索效能的方法。

第8章 结论与展望——对全书的主要研究成果进行了总结，并对本书中的工作进行了展望。

第 2 章　相关研究工作概述

本书的主要研究内容是基于概念的检索系统的构建，这个任务涉及信息检索研究许多方面的工作。这些信息检索领域相关的研究工作为这个任务打下了坚实的基础，也给了我们很多灵感和启发。本章将对和本书相关的研究工作进行全面的介绍。

2.1　传统信息检索方法

信息检索系统的关键是实现一个检索模型，这个检索模型可以根据用户提供的查询生成一个按相关度排序的文档列表。这个检索模型对查询和文档匹配的过程通常使用正规的形式化的数学模型来描述。

布尔模型是最早使用的信息检索模型，这种模型不能产生排序的文档列表，而是产生一个相关文档的集合，这个缺陷使布尔模型难以应用于大规模文档。向量空间模型出现后很快就取代了布尔模型，并且在很长的时间内都是这个领域的统治者。在向量空间模型中，文本和查询都被表示为由词项构成的向量，这个向量的维度由词表的规模决定，文档和查询的相关度可以由两个向量在这个欧式空间中的距离来确定。常见的距离计算方法是基于两个向量在高维空间的夹角余弦。向量中每个元素的值可以是简单的布尔值，也可以是复杂的浮点数。常见的浮点数类型的统计项包括词频（*TF*）和倒文档频率（*IDF*）。*TF* 和 *IDF* 这两个统计项并不是向量空间模型中特有的，实际上基本上所有的检索模型都使用它们。一个词项的 *TF* 值是指这个词项在文档中出现的频率，*IDF* 是由整个文档集合中包含这个词项的文档数的倒数计算获得。直觉上可以看出，一个文档中某个词项的 *TF* 越高，则它和包含这个词项的查询的相关度越高。另外，文档中有一些高频的常用词，它们通常不包含文档的主题信息，*IDF* 值可以降低它们的影响。在检索模型中，一个常见的统计项就是 *TF* 和

IDF 的乘积，表示为 $TF \cdot IDF$。下面是通过夹角余弦计算文档和查询相关性的公式：

$$sim(d_i, q) = \frac{\vec{d}_i \cdot \vec{q}}{|\vec{d}_i| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}, \tag{2.1}$$

其中 $|\vec{q}|$ 和 $|\vec{d}_i|$ 分别表示查询和文档向量的模。因为 $w_{i,j} \geqslant 0$ 且 $w_{i,q} \geqslant 0$，所以 $sim(d_i, q)$在 0 和 1 之间，它可以表示文档和查询之间的部分匹配关系。

图 2.1 和表 2.1 中给出了一个利用向量空间模型计算文档和查询相关度的例子。D_1 和 D_2 分别为两个不同的文档，Q 为一个查询。T_1、T_2 和 T_3 为文档和查询中出现的词语。文档和查询中词语的频率信息见表 2.1。通过向量空间模型计算后可以得出结论，文档 D_1 和查询更相关。

表 2.1　文档和查询中词语的频率信息

	T_1(电脑)	T_2(人脑)	T_3(检索)
D_1	2	3	5
D_2	3	7	2
Q	0	0	2

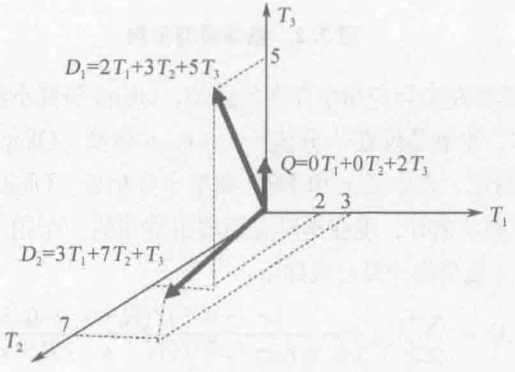

图 2.1　向量空间模型示例

除了向量空间模型外，还有很多种不同的检索模型，其中有一些现在依然被广泛使用。Maron 和 Kuhns 最早提出了利用概率模型计算查询和文档的相关度的方法。他们在布尔形式的词项表征模型的基础上引入了概率形式的表示词

项权重的方法。他们认为信息检索的过程可以看做一个统计推理的过程，检索系统预测哪些文档和查询相关并按照相关度对文档进行排列。这种思想在本质上和一种流行的机器学习方法——朴素贝叶斯模型非常相似。Maron 和 Kuhns 最早提出了按相关度排序的检索结果表征方法，这种思想对信息检索领域有很深远的影响。

在 Maron 和 Kuhns 工作的基础上，Robertson 和 Jones 提出了一种结合 *IDF* 和相关性反馈的 PRP (Probability Ranking Principle，也被称为概率排序原则)模型。PRP 模型同时利用了两个概率模型，分别对应所有相关文档和所有不相关文档，概率模型示例如图 2.2 所示。PRP 模型的关键部分是计算文档和用户相关的概率。这个概率值越高，文档和用户越相关。Robertson 证明在准确估计各项概率的情况下，利用 PRP 模型对文档排序可以优化检索效能。但是，在实际应用中想要准确估计这些概率是不太可能实现的。

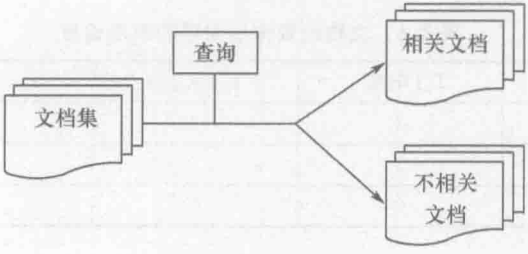

图 2.2　概率模型示例

基于 PRP 模型在实际应用中存在的缺陷，Okapi 研究小组提出了一种 PRP 模型的改进版本，也就是现在十分流行的 Okapi 模型。Okapi 模型充分利用了 *TF* 和文档长度特征，在本质上和 PRP 模型十分相近。Okapi 模型在许多测试文档集合上表现都非常好，现在是研究领域中常用的一个用于比较检索效能的基准模型。Okapi 模型的计算公式如下：

$$sim(Q,D)=\sum_{q\in Q}\log\frac{(r_i+0.5)/(R-r_i+0.5)}{(n_i-r_i+0.5)/(N-n_i-R+r_i+0.5)}\cdot\frac{(k_1+1)\cdot f_i}{K+f_i}\cdot\frac{(k_2+1)\cdot qf_i}{k_2+qf_i},\tag{2.2}$$

其中的 f_i 表示词项 i 在文档 D 中出现的频率，qf_i 表示词项 i 在查询 Q 中出现的频率，k_1，k_1 和 K 都是根据经验设定的值。