垂直搜索引擎的信息抽取与索引

梁杰 3150102259

（浙江大学 计算机学院）

摘 要：相对于普通的全文搜索引擎，垂直搜索引擎为用户提供更专业、更深度、更具个性化的搜索服务；垂直搜索引擎的实现所需完成的几个主要内容包括网页信息获取与分类、建立索引并检索等；本文首先概述了垂直搜索引擎的优势与主要设计思路，之后阐述了主流的垂直搜索引擎的实现方式及改进点。

一、 垂直搜索引擎概述

1 与普通搜索引擎的比较

搜索引擎是帮助人们从互联网海量数据中迅速并较为准确地得到预期内容的有效工具，但随着互联网中数据规模的飞速增长，普通的搜索引擎有时很难做到提供精准的、具有专业性的搜索结果，特别是当用户需要针对某个领域、伴随有专业性需求来进行搜索时，普通搜索引擎泛而不精的弊端就尤为显现了。当其选择首先索求信息广度时，面对如此大的数据，搜索引擎显然无法对于内容信息再做更深的挖掘与提取了。在这样的背景下，垂直搜索引擎应运而生。垂直搜索引擎凭借明确的检索目标定位，对网页进行选择性收集，信息采集量小，更新及时，因而能有效解决通用搜索引擎的弊端。垂直搜索引擎正在以其日趋精准化、人性化的息检索服务提升着人们对搜索引擎的使用率和认同度，助推了搜索引擎的快速发展。

2 搜索引擎架构概述

主流的垂直搜索引擎的实现一般可以分为三个模块：信息采集与抽取、信息索引、信息查询。信息采集模块负责使用爬虫从互联网上抓取大量的相关信息；信息索引模块负责对采集到的数据进行整理、分类、索引等；信息查询模块负责根据关键词或语句对信息进行查询得到符合要求的结果集，并按相关性进行排序。其中的查询模块十分依赖于前两个模块，在此不再细讲，其他两个模块的具体功能在文章的之后部分将详细介绍。

二、 信息采集与抽取子系统

1 信息采集

信息采集的主要逻辑是：采用网络爬虫技术进行数据的采集，采用广度优先算法，采集处理 URL 链接，采集完的数据保存在本地文件中。而具体来讲的话还有以下问题需要解决。

对于垂直搜索引擎来说，首先要解决的是采集范围问题。 综合搜索引擎采集范围是广泛和无限延伸的互联网，垂直搜索引擎的采集范围相对比较专一。它的采集工具和网页库的爬虫相比应该更加专业并可定制化，可定向性地采集和垂直搜索主题相关的网站，忽略不相关的和不必要的网站，选择内容相关的以及适合做进一步处理的网站深度优先采集。如何确定哪个网站是采集目标，哪个不是采集目标，这就涉及网站分类问题，分类的结果集只有两类，一类为是“目标网站”，另一类为“不是目标网站”。 研究得比较多的是文本分类和网页分类。 网页的分类可以使用决策树的方法、统计的方法、支撑向量集的方法以及粗糙集的方法等。 网页分类还涉及到多方面的技术，如网页的清洗、分词、人工样本集分类等技术。 另外，在属性提取方面，与文本分类还有区别， 文本分类完全采用分词的方法进行属性提取，网页分类则还要考虑结构方面的属性，如超链接的个数，超链接与整个文字的比例，超链接类别个数等。第二个问题是采集频率问题， 即对目标有选择地调整更新频率。 垂直搜索对信息的更新有特别的要求，根据这些特点可以考虑以下几点：一是信息源的稳定性，不能让目标网站感觉到压力；二是抓取的成本问题， 因频繁的信息采集需要大量的资源；三是目标网站的信息更新速度，这是需要重点考虑的问题，它关系到搜索引擎的用户体验。在实际应用中， 垂直搜索的采集技术可以按需控制采集目标和范围、 按需支持深度采集及按需支持复杂的动态网页采集，采集技术要求具备聚焦、纵深和可管控性，以获取最新、及时的信息。

2 信息抽取

垂直搜索是以结构化数据作为最小单元，然后存储在数据库中。一般的 web 搜索引擎是最小的 web 页面或 web 块。因此，结构化信息抽取的技术水平是决定垂直搜索引擎质量的重要技术指标，因此结构化信息抽取技术已成为垂直搜索引擎的关键技术之一。

结构化信息提取技术有三种方式，一是使用模板;二是使用结构化信息提取;使用模板进行 web 模板设置或自动生成模板提取数据,抽样也有针对性的 web 页面,适合规模相对较小,更少的信息来源和稳定,优点是快速实现,成本低、灵活性

强,缺点是后期维护成本高,信息来源和少量的信息。结构化信息提取和模板提取最大的区别是,它独立于特定的 web 页面,可以从任何正常网页中提取信息,导致质变的数据容量法和模板法,但灵活性差、成本高。但这两种方法并不是对立的,它们在垂直搜索引擎中是互补的。

结构化信息提取的方法有许多，在这里简单介绍以下三种：基于包装器的信息抽取、基于隐马尔可夫模型(HMM)的信息抽取、DIPRE 抽取。

2.1 基于包装器的信息抽取

包装器（Wrapper） 是一种软件构件，一个 Wrapper 类一般针对某一单一数据源中的一类页面,负责将数据和查询请求由一种模式转换成另一种模式。在 WEB 环境下，Wrapper 负责将隐含在 HTML 文档中的信息抽取出来，并且转换成能够被进一步处理的以某种数据结构存储的数据。形式地，一个 Wrapper 类实际上是一类页面到该页面所含元组集合的函数。典型的 Wrapper 系统有 WIEN, SoftMealy, STALKER, WHISK, Wrapper-Up, T-Wrapper 等。对于基本包装器的改进是一项热门的工作，其中一些有意义的改进工作包括：将归纳学习方法引入 Wrapper 类的自动生成；使用监督学习，从手工标记的训练示例中推导出一个抽取规则集；应用归纳学习方法于 WEB 数据抽取，即用学习到的分层结构来表示抽取规则(包括开始规则和结束规则),自动构造 Wrapper 类的有效性和表达性。比较有趣的，有学者以文档对象模型 DOM 为基础，把所要抽取的信息在 DOM 层次结构中的路径作为信息抽取的“坐标”，并以这个基本原理为基础设计了一种归纳学习算法来半自动地生成抽取规则，然后根据抽取规则生成 Java 类。生成的 Java 类可以作为 WEB 数据源 Wrapper 类组成的重要构件。基于规则的抽取模型比较常用，在很多情况下其精度也非常令人满意。包装器是信息集成机制的重要组件，它易于建立，抽取精度高，对于含有较多半结构化信息的 WEB 页面是很合适的，因此具有一定的研究价值。

2.2 基于隐马尔可夫模型(HMM)的信息抽取

基于隐马尔可夫模型(HMM)的信息抽取可以有效地体现时域或空域上的随机概率过程,已经成功地应用于语音识别和手写体识别。在信息提取中,HMM 用状态对应提取域,状态的词表对应每个域出现的符号,用状态之间的转换对应各个域之间的位置关系。最初的 HMM 构造是通过观察样本手工构造。或用一个状态对应一个域,或用几个状态对应一个域,然后通过对概率的调整来获得较好的效果。现在的 HMM 构造一般是采用自动构造,如“Cora 计算机科学研究论文搜索引擎”,利用 HMM 提取每篇论文的头部信息,包括标题、作者、关键词等。用 HMM 来进行信息提取的一般途径是:每个域(提取的每个语义项)对应一个或多个状态,原始文本中的符号作为状态的输出符号,如果模型给定,那么信息提取过程就是搜索

最可能创建符号序列的状态序列,这个问题可以由 Viterbi 算法解决。

2.3 DIPRE 抽取

DIPRE(Dual Iterative Pattern Relation Extraction),方法由以下 5 步组成:①用户指定包含少量数据的样本。②在信息源中找到样本中所有数据在信息源中出现情况的上下文信息,并且保存这些信息。③根据前面所保存的信息来产生关系模式,这一步是所有步骤中最重要的,它既要能从与上下文信息相似的集合中产生模式,又要有较低的错误率,还要尽可能的增加模式的覆盖率。④根据产生的模式再次搜索信息源,以获得更多的要抽取的信息。⑤如果要抽取的信息的数量已经满足要求则停止算法的运行,否则重复 2 -5 步。

三、 索引子系统

索引模块主要完成对于整理、分类和索引建立。

传统的基于关键词检索的搜索引擎，检索过程是基于关键词机械匹配，因此查全率低、查准率低、相关排序效果不佳。为了弥补这些缺点，人们提出了许多改进的方式。雅虎是分类搜索引擎的代表，它有自己的分类目录，采用宽泛的主题领域建立分类索引，类目详尽，使得对网络信息的全面检索变成现实。但是由于分类目录完全由人工完成，随着信息的增多，维护工作越来越艰难。随后，人们引入自然语言处理技术、在传统倒排索引结构的基础上引进邻近位信息、设计了基于概念的信息检索系统、将异构语义引入索引，这些改进使得搜索的效率与准确性都大大提高。

垂直搜索引擎具有很强的领域针对性，能够排除冗杂信息，可以在很大程度上减少不相关的检索结果，提高检索效果。

1 索引逻辑结构设计

倒排索引是全文检索引擎中最常用的数据结构。所谓倒排索引(Invertedindex)，也常被称为反向索引，是一种被用来存储在全文搜索下某个单词在一个文档或者一组文档中的存储位置的映射。由于倒排索引具有优良的特性，许多垂直搜索引擎都采用了倒排索引作为分类索引库的主要数据结构。

2 索引物理结构设计

许多搜索引擎的索引的物理结构都是在 lucene 索引物理结构基础上进行的改进，lucene 原始的物理结构不存在单独存储类别信息的文件，要存入类别信息可以采用两种方式：①将类别信息存入专门存储关键词信息的.tis 文件中；②将文档按类别分别建立独立索引。第 1 种方式虽然可以满足要求，但本质上还是基于关键词查询实现对类别的匹配，不仅效率低而且在多类分类时可用性较差；第 2 种方式真正实现了分类索引，但是在对索引进行更新操作时，性能不佳。两种方法各有得失。

参考文献：

[1] 齐鹏,张俊,李冠宇 基于本体的垂直搜索引擎分类索引模型设计 计算机工程与设计 2010,31 (23)

[2] 季春,姜琴,吴铮悦 垂直搜索引擎关键技术研究综述 情报探索 第 14 期

[3] 张敏,杜华 垂直搜索引擎系统的设计与实现 情报科学 2011.3

[4] 王敬普，基于包装器模型的文本信息抽取算法研究 湖南大学硕士学位论文