

基于 XML 与自然语言处理的智能化资源检索

王民¹, 肖君¹, 高少琛²

¹上海远程教育集团, ²上海交通大学软件学院

摘要: 在海量信息系统中如何快速、智能地获得所需资源是非常重要的。本文以上海教育资源库为背景, 给出了一个基于 XML 与自然语言处理的智能化资源检索方案。文中给出了系统地体系结构, 阐述了支撑系统实现的多个关键技术, 并就实际系统的运行情况和进一步的工作做了说明。

关键词: 资源检索, 元信息, XML, 自然语言处理

1. 引言

上海教育资源库^[1]是由上海远程教育集团于 2004 年开始建设的, 计划经过三年的时间建设成一个为上海乃至全国用户提供服务的教育资源中心。目前, 上海教育资源库已具备了相当规模, 主要包含了上海市中小学和其他教育层次的各类资源。

该资源库存储了大量的、类型各异的资源, 包括音频、视频、纯文本、Word、HTML 等格式的文件, 显然这是一个海量的信息管理系统。而在海量信息系统中, 如何让用户快速、准确地获得所需资源是必须解决的一个问题。目前, 系统采用树型结构管理这些教学资源, 在资源数量不大、分类信息明确的情况下用户能够较方便地找到所需资源。但是当系统的资源数量扩展到一定规模之后, 用户将迷失在系统的树型结构中而无法快速、准确的获得所需资源。因此, 有必要设计实现一个高效的、智能化的资源检索系统。

本文以上海教育资源库这一海量信息系统为背景, 给出了一个基于 XML 和自然语言处理的智能化资源检索方案, 以支持快速准确地发现客户所需的资源。本文第二部分描述了系统的体系结构以及各组成部分的主要功能, 第三部分阐述了系统实现过程中用到的若干关键技术, 最后对原型系统的主要特点和进一步的工作做了说明。

2. 体系结构

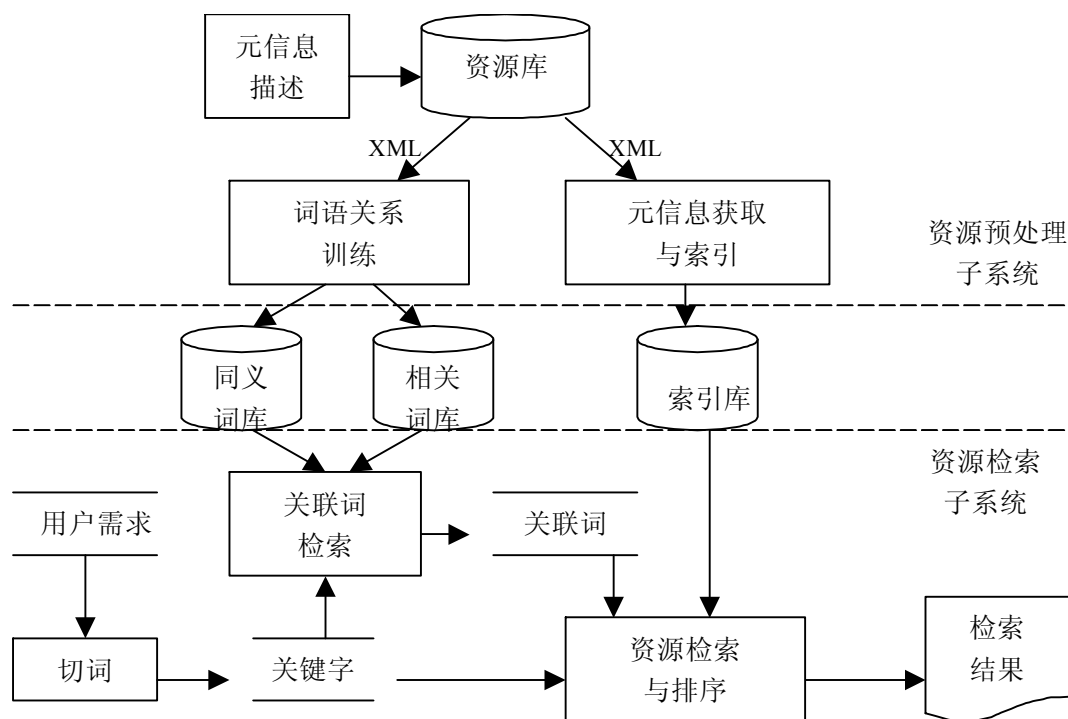


图 1 智能化资源检索系统的体系结构

图 1 描述了一个智能化资源检索系统的体系结构。整个系统由两部分组成：资源预处理子系统和资源检索子系统。

资源预处理子系统一方面负责获取各类资源的元信息并加以描述，并结合中文信息处理技术和索引技术建立资源元信息的索引库；另一方面负责构造同义词库和相关词库，用于关键词扩展以实现智能化检索。资源预处理子系统是（为在海量的、多样化的信息系统中获取所需资源的基础）整个系统的基础，为后续的检索功能提供了多个层面上的数据源。

具体的，资源预处理子系统的功能包括，

- 获取每个资源的元信息，并采用 XML 文件描述这些元信息。也就是说，系统给出了用于描述资源元信息的 XML 文件的格式定义，每个资源都有对应的 XML 描述。
- 结合 XML 格式定义，利用自然语言处理技术和索引技术从元信息描述文件中获得信息，并由此建立索引库。
- 以海量的资源元信息为依据，构造相关词库和同义词库。

资源检索子系统负责分析用户提交的请求，以索引库、同义词库和相关词库为依托，完成智能化的资源检索。本系统应用了中文信息处理的分词技术，使得用户可以提交自然语言描述的查询请求，而不必按关键词方式输入查询条件。这种友好的人机交互方式使用户获得了全新的、自然的用户体验。此外，系统按照精心设计的算法对查询的结果进行了排序，使得用户更容易获得最需要的资源。

具体地，资源检索子系统的功能包括，

- 对用户按自然语言描述的查询请求作分词处理，以获得关键词。系统丢弃了语句中的辅助性词汇，如连接词、助动词、指代词等。例如，用户输入“我要关于植物的资源”，系统在分词处理后得到“植物”“资源”这两个关键词。
- 依托同义词库和相关词库，得到上述关键词的同义词和相关词，为后续的智能化检索提供基础数据。

- 结合关键词及其扩展词汇，构造模糊查询条件。
- 根据查询条件，从索引库中检索符合条件的资源。
- 依据权重指标计算出来的匹配度，对检索的资源进行排序。将最满足用户需求的资源信息排列在最前面。

3. 关键技术

3.1 元信息描述

元信息是用于描述资源属性的信息。为了正确有效的组织、管理系统中的海量资源，必须有能力获取资源的元信息并进行结构化的描述。

在上海教育资源库中，资源的元信息包括标题、科目、摘要、作者、类型、URL、关键词等内容。系统采用 XML 文件描述资源的元信息。为了使所有的 XML 描述文件满足相同的规范以便系统实现自动化的信息处理，我们定义了该 XML 文件的格式文件，即后缀为 xsd 的 XML Schema。根据 XMLSchema 中的标签定义，本系统中的每个资源对应于 XML 文档一个 resource 节点，而资源的各种元信息通过 resource 节点包含的多个子节点（如 Title、Subject 等）描述出来。

因此，在每一个资源加入系统时，要同时提交一个按特定规范生成的 XML 文件以描述资源的元信息。而这些 XML 文件的集合就构成了资源检索系统的语料库，成为实现资源智能化检索的数据基础。

3.2 元信息获取与索引

为了实现检索功能，必须根据描述元信息的 XML 语料库建立索引。而建立索引的前提是要有能力让系统自动获得 XML 文件中的各种元信息。本系统采用 Apache 的 Digester 工具包解析 XML 文件以获取资源元信息，采用 Apache 的 Lucene 工具包建立索引。

Commons Digester^[2] 是 Apache Jakarta 旗下的一个开源软件项目，Digester 工具包提供了简单的高层用户接口以便将 XML 文档转变成 Java 对象。在本系统中，我们根据 XMLSchema 中的标签定义，在 Digester 中定义相应的解析规则，将 XML 文档中的每一个 resource 节点转换为 JavaBean 对象，再将 resource 节点中的内容抽取到该 JavaBean 对象中。由此，资源的各种元信息就被封装在 JavaBean 对象中了，后续对元信息的处理就可以通过处理该 JavaBean 对象来实现。

Lucence^[3]是由 Apache 资助的一个纯 Java 的全文检索工具包。Lucence 本身并不是一个检索系统，它只是提供了一组 API，开发者必须进行二次开发才能构造其检索系统。利用 Lucence 可以对各种类型的文件数据作索引，如 xml、doc、pdf、txt、html 等。本系统提供了创建新索引和按增量方式扩充索引两种方式，以满足用户的不同需求。

但是，一个突出的问题是 Lucence 本身的 API 并不支持中文，因为中文的词与词之间的边界并不是像英文那样以空格进行划分。因此，必须集成中文信息处理的能力，而且中文切词的效果也将直接影响到检索的效果和性能。目前已有第三方为 Lucence 提供了中文分析和检索的 API，但对中文的处理采用的是基于单字和二元切词方式，对大规模的海量数据作精确检索时效果并不尽人意。因此，在本系统中我们集成了自主开发的一个基于词表的、采用正向最大匹配算法的分词器，并提供了多达 13 万词、32 万余字的分词词表以及 2.3 万词、10 余字的用户自定义词表作切词支持，完全可以满足中文检索的切词需要。

实际系统共支持三种创建索引的策略，分别是分词索引，二元索引和单字索引。系统维护人员可以根据具体需求来决定建立哪几种类型的索引。

3.3 同义词库与相关词库

引入同义词和相关词的处理可以支持基于概念的检索方法,提高检索系统对用户需求的语义理解能力。其核心的问题就是如何扩展查询条件,以帮助用户获得其需要的信息。

本系统的同义词库是从 2000 版的知网数据库中抽取的同义词集。从检索的实际需求出发,为了减少作词条扩充而导致的噪声数据增加,系统只对查询中含有最大语义信息的名词和动词作同义扩充,而对其他词性的词汇不做处理。

对于相关词库,本系统提供了相应的工具由系统维护人员进行维护。在这种面向特定领域的海量信息系统中,由于维护人员熟知本领域内各种信息之间的关系,相关词库的内容会得到精心的组织,从而使得系统的查询能力得到大大提高。

3.4 资源检索与排序

本系统通过改造 Lucene 的查询策略,实现了更灵活有效的检索功能。系统的检索原则是,在保证检索结果拥有最大召回率的前提下,将最符合用户需求的文档排在结果集的最前面,并通过参数设置确定实际返回给用户的命中资源数量(省缺值为 50)。也就是说,返回给用户的资源将按照与查询请求的匹配度依次排列。

在 Lucene 中,查询条件被组织成布尔模型,关键词之间的地位是相等的。而本系统精心设计了相应的查询策略,将查询条件组织成向量空间模型,即根据用户输入关键词的顺序和获得同义词/相关词的情况,赋予查询词不同的权重。由于篇幅原因,具体的策略本文不予详述。实际运行效果表明,应用本系统的检索策略使得用户能更便捷的获得其所需资源。

此外,由于系统提供了三种不同模式的索引库,系统维护人员可以根据实际需要通过对参数设定的方式,确定利用哪种索引作检索。默认情况下系统综合利用分词索引与二元索引两个索引库进行检索。

[1] 系统特点

本系统具备以下几个主要特点:

- 易用性。本系统采用 Web/Server 结构,客户端通过浏览器即可访问教育资源库,并使用该智能化资源检索功能。且界面简洁,操作方便。
- 跨平台特性。整套系统全部采用 Java 语言实现,由于 Java 语言的平台无关性,本系统可以运行于多种不同的操作系统平台。目前该系统运行于 Linux 平台。
- 智能性。通过引入同义词库和相关词库,扩展了该检索系统对用户需求的语义理解能力,使得系统能支持模糊检索,并提高了资源的召回率,以帮助用户尽快找到其真正想要的资源。
- 支持用户按自然语言描述方式提交的查询请求。系统应用中文信息处理技术,自动获取关键词。使得用户可以以更为自然的方式表达查询需求,从而获得了全新的用户体验。
- 可集成性好。系统结构设计清晰,并提供了丰富的 API 和灵活参数设置功能,使得系统具有良好的可集成性。

[2] 结束语

本文以上海教育资源库为背景,给出了一个海量信息系统的智能化资源检索方案,并分

析了系统的体系结构、关键技术及主要特点。

目前，该信息检索系统已开发完毕，并投入试运行。系统对 5000 多个资源文件的元信息进行了处理，建立了相应的索引库，经过三个月的应用测试，我们认为设计的功能指标均已达到。下一步将对检索的效率作进一步优化。

参考文献

- [1] Shanghai Education Resource Center, www.sherc.net
- [2] Apache Jakarta Commons Digester, <http://jakarta.apache.org/commons/digester/>
- [3] Apache Lucene, <http://lucene.apache.org/>
- [4] L. M. de Campos, J. M. Fernandez, J. F. Huete. Building Bayesian Network-Based Information Retrieval Systems, Proceedings of the 11th International Workshop on Database and Expert Systems Applications, Page 543-553
- [5] Luis M. de Campos, J. M. Fernandez, J. F. Huete, Improving the efficiency of the Bayesian network retrieval model by reducing relationships between terms, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Page101-116