汽车交易信息搜索引擎的设计与实现①

祝伟华 杨永毅 (重庆大学 软件学院 重庆 400044)

摘 要: 分析了汽车交易领域的特点，提出了汽车交易信息搜索引擎应具有的功能和系统架构。对 Lucene 的功能结构进行了研究，在此基础上给了系统的信息获取模块、索引创建模块、查询展现模块等实现的关键技术。并提出了一种改进的向量空间模型(VSM)，引入了汽车信息属性的加权值，实现了主动推荐相似汽车信息的功能，并通过实验进行了验证，完成了汽车交易信息搜索引擎的设计与实现。

关键词: 汽车；交易信息；搜索引擎；Lucene；向量空间模型

Design and Implementation of the Automobile Trade Information Search Engine

ZHU Wei-Hua, YANG Yong-Yi

(School of Software Engineering, Chongqing University, Chongqing 400044, China)

Abstract: This paper analyzes the characteristics of the automobile trade market and proposes the functions and system architecture of the automobile trade information search engine. Based on the research of Lucene's functions and structure, it presents the strategy to implement the information gathering modle, index creation module and information query module in detail. It improves the Vector Space Model (VSM) by using the weighted values of automobile information, realizes the similar automobile information recommendation function, and completes the design of the automobile trade information search engine.

Keywords: automobile; trade information; search engine; Lucene; vector space model (VSM)

1 引言

随着我国汽车产业的快速发展以及人民消费水平的提高，越来越多的消费者选择购买汽车。据资料统计，2008 年国内汽车销量为 938.05 万辆，同比增长 6.70%。即使遭遇金融危机，汽车交易市场也逆势上扬，显示出了巨大的发展潜力。但与此极不相称的是汽车交易市场的信息化程度[1]，虽然现在已经有大量的汽车交易网站，但其中的汽车交易信息不容易被普通用户获得。然而使用 Google，Baidu 等通用搜索引擎[2]对汽车交易信息进行搜索，也暴露出了诸多问题。首先通用搜索引擎的“通用”特点决定了它不能满足汽车交易这个特定领域的专业化的信息需求，而且其提供的信息存在大量的重复，搜索结果中混杂垃圾数据，同时也遗漏了大量通用搜索引擎的 Spider 不能获取的隐藏网页中的信息。

本文针对以上问题，提出了基于开源全文索引库 Lucene 的汽车交易信息搜索引擎的设计方法并进行实现，同时提出了一种改进的向量空间模型(Vector Space Model，VSM)，实现了向用户主动推荐汽车的目的。

2 Lucene简介

Lucene 是一个高性能的可伸缩的信息检索库，它可以为应用程序添加索引和搜索能力[3]。Lucene 最初是用 Java 编写的一个全文索引工具包，它于 2001 年 9 月加入了 Apache 软件基金会的高质量开源 Java 产品 Jakarta 家族。目前 Lucene 还出现了 C、.Net 版本。本课题使用的 Lucene 下载地址为：http://apache.etoak.com/lucene/java/lucene-2.9.0-src.zip。

2.1 Lucene 系统结构

Lucene 由七个模块组成，以下将通过分析这七个模块各自对应的功能来讨论其体系结构。

① 收稿时间:2009-09-28;收到修改稿时间:2009-11-01

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2.1.1 lucene.analysis

这个模块用于对要创建索引的文本进行分词，通过预先定义的分词规则，把输入的字符串切分成一个个词。Analyzer 是一个抽象类，完成对文本内容的分词规则。但Lucene提供的默认分析器只有StandardAnalyzer 和 SimpleAnalyzer 两个，它们只能完成对英文的分词。对中文而言，需要修改这个两个分析器，引入一个中文词库[4]，完成对中文的切词。

2.1.2 lucene.document

这个模块封装了索引存储的各个单元，其中 Document 类相对应于关系型数据库的记录对象，主要负责字段的管理，字段分两种，一是 Field，即文本型字段，另一个是日期型字段 DateField。

2.1.3 lucene.index

这个模块是整个系统的核心，为每个切出来的词创建索引。负责创建索引的类是 IndexWriter。它的主要作用就是接收新加入的 document，然后在内部为之生成相应的小段，最后再合并，并向索引文件中输出。IndexReader 类的功能是从索引中删除 document 对象。

2.1.4 lucene.search

这个模块定义了大量类，用于在索引中寻找符号条件的结果。IndexSearcher 类主要用于打开要被搜索的索引，并执行搜索。真正完成搜索的是 Query，类 Query 是一个抽象类，用于搜索。搜索的结果封装在 Hits 类中。

2.1.5 lucene.store

这个模块抽象出了和平台文件系统无关的存储操作，提供了数据的存储管理。

2.1.6 lucene.queryParser

这个模块用于解析查询语句，把可读性较好的搜索语句转化成 Lucene 内部的查询语句，实现查询关键词之间的运算，如与、或、非等。

2.1.7 lucene.util

这个模块包含一些工具性函数、一些常量以及几个经过优化的数据结构。

3 搜索引擎的设计与实现

搜索引擎主要由前台展现模块和后台数据模块组成[5]。汽车交易信息搜索引擎则再进一步细化，主要由汽车交易信息获取、索引创建更新、前台查询展现、用户兴趣主动推荐等模块组成。

汽车交易信息模块首先根据数据获取源URL列表，获取汽车交易网站上的网页，抽取其中的汽车交易信息，存入数据库中；然后由索引创建更新模块进行增量方式的索引创建；Lucene 索引创建好以后，用户就可以用前台查询展现模块对各种汽车交易信息进行查询了；同时在用户浏览某一辆汽车的具体信息时，向用户主动推荐其可能感兴趣的汽车信息。系统结构如图 1 所示。

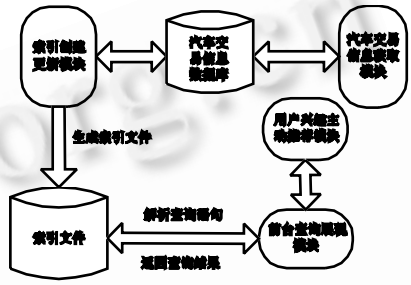

图 1　汽车交易信息搜索引擎结构

3.1 汽车交易信息获取

3.1.1 数据源的选取

本系统是专门针对汽车交易信息的垂直搜索引擎，所以 Spider 并不是像通用搜索引擎那样对所有的站点都进行爬取。本系统的 Spider 只对包含汽车交易信息的特殊站点进行爬行，同时保证爬取地范围不向外延展，只在指定的种子网站之内。

种子网站的获取主要通过两种途径，第一种途径是利用汽车交易信息中的各个关键词，在多个通用搜索引擎中进行搜索，把排名在前几页的若干网站选定为种子网站；第二种途径是根据用户的反馈信息，加入汽车交易行业的权威网站。通过对种子网站的严格选取，可以从源头提高搜索引擎的查准率，同时提高 Spider 的抓取效率，缩短搜索引擎的更新周期。

3.1.2 抓取任务的划分

通过分析，大部分的汽车交易信息网站都有一个通用的四层结构，即第一层为站点主页，第二层为按汽车品牌划分的分类界面，第三层为汽车交易信息的简要列表，第四层为具体一辆汽车的详细交易信息。由于数据源网站结构相对简单，本系统的 Spider 采用非递归的设计。进行抓取的时候，有多个 Spider 同时运行，并且每个 Spider 有多个线程同时运行。所以必须对抓取任务进行划分。抓取任务的划分，采取每个网

© 中国科学院软件研究所 http://www.c-s-a.org.cn

站按品牌划分任务的方法。总任务下，按网站划分，网站下，按照汽车品牌划分,每个线程运行一个抓取任务。

3.1.3 信息抽取

由 Spider 抓取的网页,并不需要保存在本地的硬盘上，而是直接抽取其中的汽车交易信息，保存到本地数据库中。抽取的信息包括汽车的车型、标价、发动机规格、颜色、制造时间、制动类型、行驶里程数、选配设备、卖家描述等。由于数据源来自多个不同的网站，网页中的数据呈现出多种异构的表达方式，很难通过简单的“属性-值”匹配方法抽取出来。现阶段主要的信息抽取模型有：基于正则表达式的信息抽取、SRV 模型、神经网络模型、隐 Markov 模型等[6]。

3.2 索引创建更新

索引创建更新模块主要负责从数据库中读取每辆汽车的交易信息，建立索引文件后，存储在硬盘上。Lucene 创建的索引，由多个段(Segment)组成，每个 Segment 又由多个文档(Document)组成。对于一个 Document 对象，其包含了一个或多个自命名的域(Field)。本系统是要为数据库中的数据建立索引，所以一个 Document 对应数据库中的一条记录，一个 Field 对应数据库中的一个属性。

在为 Document 添加 Field 时,有一个参数 Store 需要注意。当 Store 参数的值为 Field.Store.YES 时，表示这个字段的数据存入了索引文件，当对索引文件进行查询时，能查询到这个 Document，同时也能返回这个 Field 中的值；当 Store 参数的值为 Field.Store.NO 时，这个字段的数据不存入索引文件中，只能查询到这个 Document,但不能返回 Field 中的值。在本系统中，汽车车型、标价、汽车品牌等信息，因为其数据量小且被查询率高，所以直接存入索引文件中；而汽车选配信息、详细描述等数据，因为数据庞大，所以不存入索引文件，它们只能进行查询，根据返回的 Document 中的数据库 ID 字段，在数据库中读取这一部分的数据。这样可以有效的减小索引文件的大小，同时提高查询时的效率。

当有新的数据被 Spider 抓取存入到数据库时,需要对索引文件进行更新,Lucene 可以实现对索引文件的增量式更新,不需要对整个索引文件进行重新创建。

3.3 前台查询展现

汽车交易信息搜索引擎的前台展现主要有两种形式。第一种就是和通用搜索引擎一样的表现形式，通过用户在搜索栏中输入要查询的关键字，然后系统在索引文件中进行搜索，返回查询的结果列表给用户。

但这样的查询过于简单，对于专业性很强的汽车交易信息，这种方式并不是十分适合。所以本系统同时提供了另一种查询法方，就是按各种分类进行汽车交易信息的展现。其中分类包括：按汽车品牌分类、按汽车车型分类、按汽车交易地点分类等等。用户进入各个分类，可以看到该分类下的汽车概要信息列表，然后点取感兴趣的汽车，查看详细信息。

3.4 用户兴趣主动推荐

这个模块主要负责，在用户查看汽车的详细信息时，主动向用户推荐相似的其它汽车。这个模块的实质是提取用户正在查看的汽车信息的特征项,然后在此基础上设计出某种模型来找到与其相似的信息。

3.4.1 基于 TF-IDF 的向量空间模型

向量空间模型由康奈尔大学的 Salton 在 20 世纪 60 年代提出。此模型将文档转化为向量形式，将文档内容的比较简化为空间向量的计算，模型简单高效，得到了广泛的应用[7]。

模型设文档集合 D 中包含有 n 篇文档，记为 $D=\{d_1,d_2,d_3,\cdots,d_n\}$，并且 D 中可以提取出 m 个特征值，记为 $T=\{t_1,t_2,t_3,\cdots,t_m\}$。通过计算文档 d_i 中各个特征值 $T=\{t_1,t_2,t_3,\cdots,t_m\}$的权重，得到 d_i 的特征向量 $C=\{w_1,w_2,w_3,\cdots,w_m\}$。其中权重 w_j 的计算公式如下：

$$w_j = TF_j \times IDF_j \tag{1}$$

$$IDF_j = (\log_2^{(n/n_j)} + 0.01) \tag{2}$$

其中 TF_j 为特征值 t_1 在 d_i 中出现的频率，IDF_j 为反文档频率，n 为 D 中的文档总数，n_j 为 D 中出现了特征值 t_1 的文档数。由公式(1)可知 TF_j 越大，即特征值 t_1 在 d_i 中出现的次数越多，权重 w_j 越大；n_j 越小，即 t_1 在文档集合 D 中出现的文档数越小，w_j 越大，越能代表文档 d_i 的特征。

设有 D 中的两文档，特征向量分别为 $U=\{wu_1, wu_2,wu_3,\cdots,wu_n\}$,$V=\{wv_1,wv_2,wv_3,\cdots,wv_n\}$。两者的相似度可用向量之间的距离(用余弦值表示)来度量,相似度计算公式为：

$$sim(U,V) = \frac{\sum_{i=1}^{n}(w_{ui} \times w_{vi})}{\sqrt{\sum_{i=1}^{n}(w_{ui})^2} \times \sqrt{\sum_{i=1}^{n}(w_{vi})^2}} \tag{3}$$

为了判断两文档是否相似，需要选取一个阈值，当 sim(U,V)大于等于这个阈值时,即认为这两个文档相似。

© 中国科学院软件研究所 http://www.c-s-a.org.cn

3.4.2 改进的汽车信息向量空间模型

使用传统的向量空间模型，实现相似汽车信息主动推荐，还是存在一些问题。首先反文档频率 IDF_j 会导致一些判断汽车相似性的关键特征值的权重降低。比如汽车的车型字段的值是“轿车”，这个词在判读汽车相似性时起着至关重要的作用，一个在查看“轿车”信息的用户是极少关注“卡车”或其它车型汽车信息的。然而“轿车”这个词在整个数据集中出现的频率非常高，根据公式(2)可知其 IDF_j 是很低的，这样就导致了其权重值的下降，不能完成向用户推荐相似汽车的功能。同时，每当有新的汽车信息入库或删除时，就需要重新计算 IDF_j，导致系统效率下降。

针对以上问题，本文提出了一种改进的汽车信息向量空间模型。一条汽车交易信息记录由车型、品牌、标价、行驶里程数等属性构成。在这些属性中出现的词，应适当的增加其在特征向量中的权重。根据人工分辨以及用户反馈，我们得到了特征值在在汽车信息各个属性中出现的加权值，如表 1 所示：

表 1　汽车信息各属性加权因子

特征值出现的属性	加权值	表示符号
车型	10	α_1
品牌	8	α_2
发动机规格	8	α_3
标价	5	α_4
颜色	4	α_5
制造时间	6	α_6
制动类型	3	α_7
行驶里程数	5	α_8
其他属性	1	α_9

引入加权值后，当特征值在相应的属性中出现的时候，α_i 的值为其对应的加权值，否则为 0。如特征值在“品牌”属性中出现，则 $\alpha_2=8$，否则 $\alpha_2=0$。这样新的权重 w'_j 计算公式为：

$$w'_j = TF_j \times \sum_{i=1}^{9} \alpha_i \tag{4}$$

则两汽车交易信息的相似度为：

$$sim'(U,V) = \frac{\sum_{i=1}^{n}(w'_{ui} \times w'_{vi})}{\sqrt{\sum_{i=1}^{n}(w'_{ui})^2} \times \sqrt{\sum_{i=1}^{n}(w'_{vi})^2}} \tag{5}$$

以上改进在一定程度上提高了关键特征值的权重，拉大了相似与非相似汽车信息的相似度的差距，便于选取适当的阈值对是否相似进行判断。

3.4.3 改进向量空间模型测试结果

我们选取了数据库中 50 条记录进测试，其中一辆汽车的信息作为比较的基准，计算另外 49 辆汽车与它的相似度。同时我们进行了人工分类，49 辆车中，和比较车相似的有 21 辆，不相似的有 28 辆，实验结果如表 2 所示：

表 2　两种模型的实验结果

比较项目	传统向量空间平均相似度	改进向量空间平均相似度
相似 21 辆汽车信息	0.599426	0.791671
不相似 28 辆汽车信息	0.482489	0.300143

实验结果说明改进后的向量空间模型，成功的把相似汽车信息的相似度扩大，把不相似汽车信息的相似度缩小，相似与不相似的界限更加分明，能容易的选取一个合适的阈值判断汽车信息是否相似，实现了主动向用户推荐相似汽车的功能。

4　总结

本文提出了一个基于 Lucene 的汽车交易信息搜索引擎的设计实现方案，剖析了 Lucene 各模块的功能，详细阐述了本搜索引擎各模块的技术细节，并提出了一种改进的向量空间模型，克服了传统向量空间模型区分汽车相似信息模糊的缺点，实现了主动向用户推荐相似汽车的功能。此搜索引擎聚合了汽车交易领域的信息资源，为用户提供了准确、实时的汽车交易信息，能更好的推动汽车交易市场的信息化。

参考文献

1 且小钢.全国汽车交易市场动态分析.汽车与社会, 2003,z1:8.

2 Almpanidis G, Kotropoulos C. Combining Text and Link Analysis for Focused Crawling – An application for Vertical Search Engines. Information Systems, 2007,32(6):886 – 908.

3 Hatcher E, Gospodetic O. Lucene in Action. New York: Oreilly, 2004. 40 – 68.

4 张华平.计算所汉语词法分析系统ICTCLAS 3.0白皮书, 2006. [2008-7-10].http://www.i3s.ac.cn/Manual/

5 李刚,宋伟. Ajax+Lucene 构建搜索引擎.北京:人民邮电出版社, 2006. 53 – 73.

6 罗三定,黄勇.一个应用模糊方法的智能搜索引擎的构建.计算机工程, 2000,26(12):113 – 115.

7 庞剑峰,卜东波,白硕.基于向量空间模型的文本自动分类系统的研究与实现.计算机应用研究, 2001,18(9):23 – 26.

© 中国科学院软件研究所 http://www.c-s-a.org.cn