

# 一种面向 B2B 垂直搜索的网页信息去噪方法

崔 阳<sup>1,2</sup>, 吴爱华<sup>2</sup>

(1. 北京科技大学 信息工程学院, 北京 100083;

2. 九城网络技术集团有限公司, 北京 100020)

**摘 要:** B2B 垂直搜索引擎是垂直搜索引擎在电子商务领域的应用。怎样更好地对互联网中海量的企业产品信息进行抽取和去噪, 是当前 B2B 垂直搜索引擎构建中所面临的重要问题。介绍了 B2B 垂直搜索引擎的特征; 分析了一般企业网站的基本结构, 在此基础上提出一种面向 B2B 垂直搜索引擎的企业站点产品信息去噪方法; 给出了该方法的实验结果。使用这种方法抽取到的产品信息可用于指导产品进一步的分类工作。

**关键词:** B2B 垂直搜索引擎; 信息抽取; 去噪; 企业站点树

**中图分类号:** TP393.09

**文献标识码:** A

**文章编号:** 1673-629X(2008)12-0070-04

## A Method of Eliminating Noisy Information in Web Pages Oriented B2B Vertical Searching

CUI Yang<sup>1,2</sup>, WU Ai-hua<sup>2</sup>

(1. Department of Computer Science, Beijing University of Science and Technology, Beijing 100083, China;

2. Ninetowns Internet Technology Group Co., Ltd, Beijing 100020, China)

**Abstract:** B2B vertical search engine is a kind of vertical searching engines and used for E-business. Now it is an important issue that how to eliminate noise and extract useful manufacture information from corporation websites. The characters of B2B vertical search engine is introduced briefly first, then the general structure of the corporation websites is analyzed, and a method of eliminating noisy information in corporation websites is proposed, at last the result of experiments is given. The information extracted by that method can help the manufacture classification.

**Key words:** B2B vertical search engine; information extraction; noise elimination; corporation website tree

### 0 引 言

计算机和互联网技术的快速发展导致了电子商务的出现, 而 B2B 是电子商务的重要组成部分之一。B2B(Business to Business)是指企业与企业间通过互联网进行产品、服务及信息的交换。传统的企业间交易通常要耗费企业的大量资源和时间, 使产品成本增加。通过 B2B 的交易方式, 买卖双方能够在网上完成整个业务流程, 减少许多事务性的工作流程和管理开支, 从而降低企业的运营成本。

从当前的发展趋势看, B2B 平台的进一步完善和发展与垂直搜索引擎技术密切相关。垂直搜索引擎, 又称专业搜索引擎, 是用于查询特定信息的工具, 专门收录某一领域、某一行业或某一主题的信息, 在解决特

定问题的查询时比综合搜索引擎更为快速有效。由于 B2B 是在企业之间展开的, 所以构建对各类企业的多种产品能够进行有效分类和高效查询的平台就显得尤为重要。而这正是 B2B 垂直搜索引擎的主要任务和目的。文中拟对 B2B 垂直搜索引擎构建中遇到的网页去噪及分类问题进行一些讨论。

### 1 B2B 垂直搜索引擎的特征分析

垂直搜索引擎与普通搜索引擎的最显著区别是对网页信息进行了结构化转换和抽取, 即垂直搜索引擎需要将非结构化、半结构化的网页数据转换为特定的结构化数据。结构化数据作为垂直搜索引擎的基本单位, 首先要存储到数据库, 进行去噪、分类、分词和索引等一系列处理, 最后以搜索的方式呈现给用户<sup>[1]</sup>。

具体来说, 垂直搜索引擎在信息的采集、加工和处理方面都有特殊之处。垂直搜索引擎的信息采集以深度优先为策略, 这是因为垂直搜索引擎以解决某一领

收稿日期: 2008-03-21

基金项目: 国家自然科学基金(60675030)

作者简介: 崔 阳(1979-), 男, 博士研究生, 研究方向为知识发现;

导师: 杨炳儒, 教授, 博士生导师, 研究方向为知识工程与知识发现。

域或专业问题的搜索为目的,必然要求信息采集时具有足够的深度,否则有可能遗漏有价值的重要信息。垂直搜索引擎对采集的信息在存储之前要进行加工,除将非结构化的网页数据转换为结构化数据外,还要对关键信息进行去噪,并在此基础上对信息进行分类,形成数据清洁、主题明确的数据库供用户查询。在信息检索方面,垂直搜索引擎不仅能够对网页信息中的结构化信息进行检索,还能提供结构化和非结构化信息相结合的检索方式。从检索结果的排序方式看,垂直搜索的排序可以更加多样化,如按时间排序、按相关度排序、按某个结构化字段排序等等<sup>[2]</sup>。

就 B2B 垂直搜索引擎而言,最重要的特征也是最主要的功能,是对企业产品信息的抽取和分类。这是因为 B2B 垂直搜索引擎作为进行 B2B 交易的企业双方获取信息的渠道,如果抽取的各类企业产品信息中包含大量噪音数据,则会造成查询速度下降、查准率降低,导致用户在查询中遗漏有价值的产品信息,造成商业损失。如果分类不够准确,可能会使用户在搜索某一类产品信息时,出现与该类产品毫不相关的产品信息。最典型的如在搜寻品牌名为“APPLE”的 MP3 产品时,搜索结果中除了包含销售该类产品的企业信息外,还出现了销售苹果这种水果的企业信息。正因为如此,文中将讨论的重点放在企业站点信息的去噪,及在此基础上进行的分类问题上。

## 2 企业站点网页去噪及应用

### 2.1 网页去噪方法概述

当前 Web 网页的数量和增长速度都呈爆炸性趋势。一个网页中除了表达主题的内容外,一般还包括维系页面间关系的超链接、为方便用户浏览或引起用户兴趣而设计的导航信息(如导航栏、列表栏、图片等),以及一些广告、声明等内容。这些内容从分析网页数据的角度都可视为噪音数据。此外,页面的噪音数据还应包括用标记语言或脚本语言定义的页面样式,如 CSS 等。

网页数据去噪主要有两种情况:第一种是基于单个页面,根据页面的 DOM 树和可视化信息等,应用一些规则将页面噪音去除;第二种是基于一个网站内所有页面,将生成这些页面的模板,也就是网站页面所共有的信息视作噪音去除。目前实现此类工作的方法有多种,如基于 DOM 树的去噪方法<sup>[3]</sup>、Site Style Tree 方法<sup>[4]</sup>、根据内容块熵值的去噪方法<sup>[5]</sup>、基于页面可视化信息的 VIPS 算法<sup>[6]</sup>等。但这些方法普遍存在通用性较好,但对某一类噪音或某一专题去噪能力不足的问题。

特别需要明确的是,对于垂直搜索引擎而言,并非网页中所有表达主题的内容都有价值。这是因为垂直搜索引擎总是针对特定领域、特定主题的,其抽取的信息也必定是与领域或主题密切相关的,除此之外的一切信息都可视为噪音。这一原则也是在去除企业站点网页噪音、构建 B2B 垂直搜索引擎时要注意的。

### 2.2 面向企业站点的去噪过程

一般来说,B2B 垂直搜索引擎在抽取一个企业站点的信息时,最关注的是该企业的产品名称、明细,以及企业介绍及联系方式等内容,而对诸如企业招聘信息、合作发展等信息较少或完全不关注。除在 2.1 中指出的各类噪音外,这类信息也应看作是噪音数据,在抽取时要予以去除。另外企业站点网页的去噪是基于整个站点进行的,因此去噪和抽取规则对于站点内所有页面都应该是适用的。

#### 2.2.1 低端噪音数据的清洗

诸如页面中的 CSS、脚本语言代码等数据,可称之为低端噪音数据,在大多数网页去噪过程中都要考虑。对于网页数据的分析和去噪是基于网页源代码进行的。在去噪前,首先将站点的所有页面源代码下载并以一定文件格式保存。通过识别文件中标记语言或脚本语言的相应标签,可以很容易地将这些噪音数据过滤,而将正文内容保留,从而得到过滤后“清洁”的页面代码文件。通常还应为这类文件建立索引文件。

#### 2.2.2 网站层次树的建立

企业站点的一个重要性是内容具有层次性。例如,站点首页作为入口,内容通常包括企业介绍、产品展示、联系方式、人才战略、招聘信息等主题,这些主题一般由导航栏列出。一个主题下可能又有多个子主题。这样就形成了企业站点的层次特征。通过对这种层次特征的分析研究,可以为去噪提供一种有效的方法。

如前所述,B2B 垂直搜索引擎中最关注的是某一企业的产品名称、描述和分类。因此应首先查找到那些主题为产品明细的页面,将这些页面清洗后,抽取和保存与搜索有关的数据,并基于这些数据进行产品分类。怎样识别企业站点中哪些页面为产品页面,是一个较为困难的问题。一种方法是对站点中所有页面的布局进行比较,将大量布局相同或相似的页面视为产品页面。但这种方法过于依赖站点自身的页面状况,如果某一企业站点各个产品页面布局较为混乱,则有可能无法正确识别产品页面。另外,这种方法有可能将与产品页面布局类似的无关页面,如产品列表页面,也识别为产品页面。另一种方法是通过人工确定一些识别规则,如认为页面的 URL 中含有“PRODUCT”字

样,即为产品页面,遍历站点中所有的 URL 进行识别。但这种方法的局限性更大,规则通常只能针对某一单独企业站点,不具有普遍性。

文中使用的是一种通过建立企业站点层次树的方法实现产品页面识别的问题。如前所述,企业站点呈层次性,这种层次性通过页面间的链接表现。分析企业站点,可得如图 1 所示的企业站点层次结构。

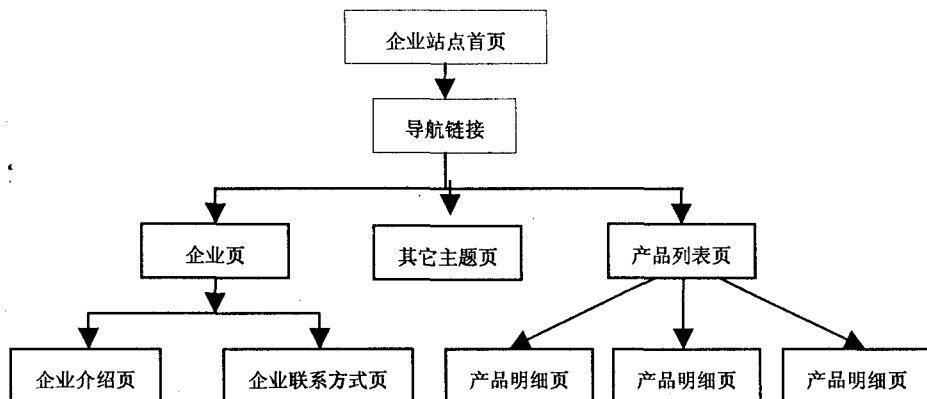


图 1 企业站点层次结构图

从图 1 可知,企业站点的第一层即为入口地址,也就是企业站首页。该层的主要内容导航栏,即站点各主题的链接。第二层为导航栏引出的各主题首页,如企业介绍、联系方式、产品介绍等。一些主题没有第三层链接,如企业联系方式。但对于产品介绍这样的主题,则第三层链接有可能是各种产品明细的链接,也有可能是产品各种子分类的链接。这类链接的数量通常是较庞大的,它们指向的正是所需的产品页面。如果能够自动识别出企业站点中指向产品介绍主题的链接,也就是找到了各种产品页面的入口地址。为此首先要建立企业站点层次树。通过对大量企业站点的分析可知,多数企业站点的产品页面位于第三层或第四层。因此层次树深度为 3 或 4 即可。

树的基本特征是无回路,但企业站点的全部链接实际组成的是图状结构,或称网状结构。这是因为某一页面上的某个链接,指向的可能是其上层;多个页面上包含有指向同一页面的链接。另外还有一些非企业站点内部链接的链接,如一些友情链接、广告链接等。在建树时都要进行过滤,这也是去噪工作的一部分。

层次树的建立仍以 2.2.1 中已经去除掉低端噪音数据的页面代码文件为基础。树的根节点即为企业站点的首页 URL,通过首页 URL 在文件中查询到相应的页面代码,将首页代码中所包含的 URL 依次提取出来作为第二层树节点,这些 URL 多为导航栏引出。递归此方法即可得到企业站点层次树。

在建树过程中,要注意 URL 的分析,过滤掉以下几类 URL:

- (1) 指向该层同层或上层的 URL;
- (2) 指向非页面的 URL,如指向 JPG、PDF、EXE 等文件的 URL;
- (3) 指向企业站点外部的 URL,如广告链接等。

对于(1)类 URL,可以通过比对当前 URL 与同层及上层 URL 进行判断;对于(2)类 URL,可以通过识别 URL 格式简单进行判断;对于(3)类 URL,则需要

判断其与企业站点各链接的相关性。另外在多个页面上出现的指向一个相同下层页面的 URL,只将其作为第一个页面的下层链接,其它页面中的忽略。

层次树的建立过程实际上完成了对企业站点页面的第二次去噪,即将大量与产品介绍无关的页面链接过滤掉,只保留各类产品的

明细页面,以及企业介绍、联系方式等页面。建好的企业站点层次树中存储的正是 B2B 垂直搜索引擎最关注的数

据,可以通过遍历层次树获取这些数据。

算法过程可简单描述如下:

输入:企业站点首页 URL

输出:企业站点层次树

步骤:

- 1) 设置层次树的最大深度;
- 2) 将企业站点首页 URL 作为站点层次树根节点;
- 3) 以 URL 值作为关键字,在保存“清洁”页面代码的文件中查询得到该 URL 对应的页面源代码;
- 4) 逐一分析源代码中的 URL,过滤掉 2.2.1 中提出的几类噪音 URL;
- 5) 将过滤后得到的所有 URL 进行保存,并作为根节点的子节点;
- 6) 层次树深度增 1;
- 7) 依次将当前根节点各子节点 URL 作为子树的根节点;
- 8) 若当前层次树深度未达到最大深度,则重复步骤 3;反之算法结束。

### 2.3 在网页去噪基础上进行产品分类

B2B 垂直搜索引擎构建中另一个关键问题是对产品的分类。分类的精确度不高,有可能对客户搜索产品信息造成影响,典型如前面提到的搜索“APPLE”牌 MP3 的例子。之所以造成这种情况,除分类技术和算法的局限外,网页数据量过于庞大也是原因之一。如

果将从不同企业站点抽取的网页数据,无差别地存储在数据库中,则在基于关键词刻画的产品分类过程中可能将某些关键词类似、但本质不同的产品错划为一类。通过运用企业站点层次树,可在一定程度上解决这一问题。

在一个企业站点中,产品总是按照其分类进行介绍的,而且这种分类的准确率非常高。可以考虑将这一知识应用于企业产品数据的预分类。定义层次树节点的出度为:节点对应链接指向的页面中包含的子链接数目(页面中的链接已经过去噪)。基于统计结果可知,出度较大的节点通常为产品列表(分类)页面,其下层节点即为某类产品页面的集合。这样在存储产品页面数据时,可通过其父节点获得该产品属性及在站点中的分类情况,以及企业性质等信息,同产品页面数据一并保存。在分类时,可将企业作为基本单位进行,从而在很大程度上避免了将性质完全不同的企业的产品页面进行比较,造成分类出现偏差的问题。

### 3 实 验

为检验算法效果,从互联网中随机搜索的 10 个不同企业站点作为实验数据来源。首先将某一企业站点中所有的页面予以下载保存,然后可以通过 DOM 树等方法将低端噪音数据去除,得到“清洁”的页面代码文件。为这些页面文件建立 URL 到相应代码的索引文件。接下来调用企业站点层次树算法,构建层次树。在实验中将层次树的深度设置为 3。

图 2 显示的是针对某一小型企业站点使用文中的去噪方法的结果。从图中可以看到,站点中大量无关链接已被过滤,在第二层节点 <http://www.xingyue-zskt.com/products.php> 下包含了较多子节点,而这些节点正是指向产品明细页面的链接。通过比对和观察,对其它企业站点的实验结果也取得了较好的效果。

### 4 结束语

B2B 垂直搜索引擎是搜索引擎技术发展中具有相当活力的一个研究方向,其性能优劣将直接影响到企业间通过 B2B 方式交易的快捷性和高效性。文中讨论的对企业产品网站信息抽取时的去噪方法,能够较好地过滤与 B2B 主题完全无关或无直接关联的数据和信息,使建立的 B2B 垂直搜索引擎更加专业、搜索和分类的准确性和全面性进一步提高。今后将进一步研究和改进这种方法的效率和适用性,以便将其更好地运用于具有海量数据和层次结构更复杂的企业产品

网站的去噪工作中。

```

Input the URL:
http://www.xingyue-zskt.com

The tree of the website is:
level 1 http://www.xingyue-zskt.com
level 2 http://www.xingyue-zskt.com/about.php
level 2 http://www.xingyue-zskt.com/products.php
level 3 http://www.xingyue-zskt.com/products.php?id=1
level 3 http://www.xingyue-zskt.com/products.php?id=3
level 3 http://www.xingyue-zskt.com/products.php?id=25
level 3 http://www.xingyue-zskt.com/products.php?id=107
level 3 http://www.xingyue-zskt.com/products.php?id=108
level 3 http://www.xingyue-zskt.com/products.php?id=109
level 2 http://www.xingyue-zskt.com/service.php
level 2 http://www.xingyue-zskt.com/feedblack.php
level 2 http://www.xingyue-zskt.com/contact.php
level 2 http://www.xingyue-zskt.com/enabout.php
level 3 mailto:george.wu@xingyue-zskt.com
level 2 http://zsdongying.cn.alibaba.com/athena/bizreflist/zs-
dongying.html
level 2 http://zsdongying.cn.alibaba.com
  
```

图 2 实验结果显示

### 参考文献:

- [1] 刘 畅. 综合搜索引擎与垂直搜索引擎的比较研究[J]. 情报科学, 2007, 25(1): 97-102.
- [2] 林文清. B2B 垂直搜索引擎在信息获取技术中的应用[J]. 情报杂志, 2007, 26(9): 120-121.
- [3] Gupta S, Kaiser G, Neistadt D, et al. DOM-based content extraction of HTML documents[C]//Proceeding of the 12th International Conference on World Wide Web. New York: ACM Press, 2003: 207-214.
- [4] Yi Lan, Liu Bing, Li Xiaoli. Eliminating noisy information in Web pages for data mining[C]//Proceeding of the 8th ACM SIG KDD International Conference on Knowledge Discovery and Data Mining. Washington, DC: ACM Press, 2003: 296-305.
- [5] Lin Shian-hua, Ho Jan-ming. Discovering informative content block from Web documents[C]//Proceeding of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2002: 588-593.
- [6] Cai Deng, Yu Shi-peng, Wen Ji-rong, et al. Extracting content structure for Web pages based on visual representation [C]//Proceeding of the 5th Asia Pacific Web Conference. Berlin: Springer-Verlag, 2003: 406-417.