

基于 Nutch 的就业垂直搜索引擎研究

肖红玉, 贺 辉, 黄灼东, 蔡昭阳

(北京师范大学珠海分校 信息技术学院 广东 珠海 519087)

摘 要: 针对通用搜索引擎专业性不够、查准率较低的问题, 基于 Nutch 开源搜索引擎, 采用基于本地词库和动态加载词库的正向迭代最细粒度切分算法实现中文分词。基于特征词和元数据标签的空间向量模型实现就业领域主题相关性判定, 基于 MapReduce 引入网页链入链接权重因子和时间衰减因子改进 LinkRank 排序算法等对 Nutch 进行二次开发, 并在网页信息抓取和过滤、就业信息搜索和特征词推荐等环节引入就业领域本体信息, 采用 Java 框架技术对用户查询接口进行了二次开发, 提供了如关键字智能提醒、定制爬虫、二次查找、设定查询结果日期、订阅查询等扩展查询接口, 设计并实现了基于 Nutch 的就业垂直搜索引擎。实验结果表明, 基于 Nutch 的就业垂直搜索引擎具有较高的查准率, 可以满足用户专业检索的需求。

关键词: 垂直搜索引擎; LinkRank 算法; 就业; Nutch

中图分类号: TP302

文献标识码: A

文章编号: 1673-629X(2019)02-0207-05

doi: 10.3969/j.issn.1673-629X.2019.02.043

Research on Employment Vertical Search Engine Based on Nutch

XIAO Hong-yu, HE Hui, HUANG Zhuo-dong, CAI Zhao-yang

(School of Information Technology, Beijing Normal University Zhuhai, Zhuhai 519087, China)

Abstract: Aiming at the problems that the general search engine has poor profession and low precision rate, based on Nutch, an open source engine, we use forward iteration and fine-grained segmentation algorithm based on local word lexicon and dynamically loaded word lexicon to achieve Chinese word segmentation. Vector space model based on feature words and metadata tags is used to determine topic relevance in employment field. The LinkRank sorting algorithm supporting MapReduce which is introduced the link weight factor and time decay factor is improved to make a secondary development of Nutch and employment domain ontology is applied to web information crawling and filtering, employment information retrieval and feature word recommendation stages. Spring MVC technology is used to develop the user query interface, which provides the extended query interface such as keyword intelligent reminder, customized crawler, secondary search, setting query result date, subscription query and so on. At last, the employment vertical search engine based on Nutch is designed and implemented. Experiment shows that the employment vertical search engine based on Nutch has a high precision and can meet the professional needs of user retrieval.

Key words: vertical search engine; LinkRank algorithm; employment; Nutch

0 引 言

随着高校不断扩张, 毕业生人数屡创新高, 2016 年高校毕业生 765 万^[1], 2017 年 795 万^[2], 2018 年将达 820 万^[3]。在线求职是毕业生就业环节的重要部分, 目前国内求职招聘相关的网站已经发展到近千家, 大学生们比较熟悉的有智联招聘、前程无忧、58 同城招聘、赶集网招聘等。各大招聘网站网罗的用人单位和发布的招聘职位众多, 但是各自为政, 数据无法共享, 使用百度、搜狗等通用搜索引擎搜索招聘信息时,

搜索结果有大量的无效信息。为了帮助毕业生快速、准确地检索招聘信息, 就业领域垂直搜索引擎应运而生, 其主要目标是提高毕业生检索招聘信息的查准率。

Nutch 是一个 Java 实现的开源搜索引擎^[4], 主要包括 Web Crawler(网页爬虫)和索引技术 Lucene 两部分。文中基于 Nutch 开源搜索引擎开发就业垂直搜索引擎, 鉴于 Nutch 本身没有中文分词器、缺少主题相关性判断、内置网页排序算法简单化、用户检索接口单一

收稿日期: 2018-04-06

修回日期: 2018-07-17

网络出版时间: 2018-11-15

基金项目: 广东省自然科学基金-博士启动(2014A030310415); 广东省教育研究课题(GDJY-2015-C-b048)

作者简介: 肖红玉(1976-), 女, 博士, 副教授, 研究方向为分布式水文系统、软件工程。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20181115.1051.094.html>

化等情形,基于 Nutch 进行二次开发,完成就业垂直搜索引擎。

1 系统框架

以 Nutch 为基础,借鉴了 TUCUXI (intelligent hunter agent for concept understanding and LeXical chaining)^[5]和 SHOE (simple HTML ontology language)^[6]的构建思路、元搜索引擎 (meta-search engine) 和

ifWeb 原型系统^[7]以及 LinkRank^[8]排序算法的设计思想,在基于 Nutch 的就业垂直搜索引擎设计上引入就业领域本体网页爬取和过滤,采用就业领域本体计算网页的就业相关性并改进 Nutch 自带的 LinkRank 网页评分算法,并结合 Spring boot、Spring-data-jpa、Shiro 等 Java 开发框架开发了系统管理后台,为用户提供高级检索、关键词定义、搜索词高亮显示等辅助查询接口。就业垂直搜索引擎体系结构如图 1 所示。

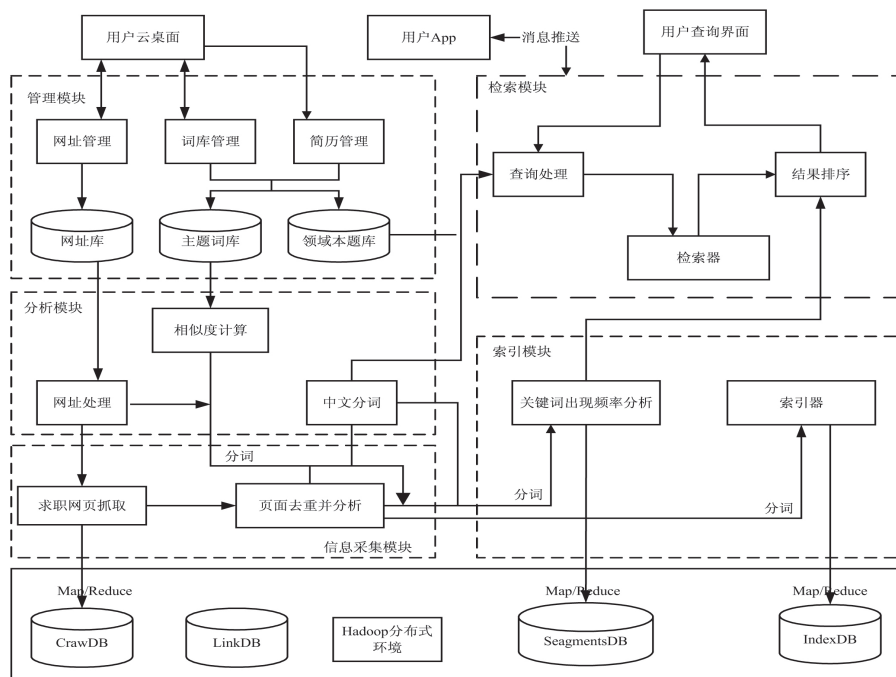


图 1 就业垂直搜索引擎体系结构

从业务逻辑流程分析,基于 Nutch 的就业垂直搜索引擎的工作流程分为 10 个阶段:

- (1) 创建 WebDb;
- (2) 将抓取初始 URLs 写入 WebDb;
- (3) 根据 WebDb 生成抓取列表 (fetchlist) 并写入相应的 segment;
- (4) 读取 fetchlist 中的 URL 信息,启动爬虫爬取网页;
- (5) 网页爬取结束后更新 WebDb;
- (6) 根据系统预设的爬取深度,循环第 3~5 步;
- (7) 采用 LinkRank 网页评分算法给爬取的网页打分,更新 segments;
- (8) 采用 Lucene 建立索引;
- (9) 把索引中重复的网页和 URL 丢弃;
- (10) 将 segments 中的索引进行合并并生成用于检索的最终索引。

2 系统关键技术

2.1 中文分词

中文分词是中文搜索引擎的关键技术,分词结果

会明显地影响检索结果。Nutch 本身针对英文检索,没有自带的中文分词器,因此需要基于 Nutch 进行二次开发。

采用基于本地词库和动态加载词库的正向迭代最细粒度切分算法的 ik-analyzer^[9]分词系统,对 Nutch 中文分词进行了改进。ik-analyzer 对待切分的字符串采用先最大词再最小词的迭代方式进行切分,以被切分的字符串“北京师范大学珠海分校开学了”为例,首先在词库中检索最大切分词后,分割为“北京师范大学珠海分校”和“开学了”;然后将“北京师范大学珠海分校”切分为“北京师范大学”和“珠海分校”,以此类推,最后,“北京师范大学珠海分校开学了”切分为“北京师范大学|北京|京师|师范|大学|珠海分校|珠海|分校|开学|了”,ik-analyzer 默认的切词方式是细粒度切分。若按最大词长切分,切分结果为“北京师范大学|珠海分校|开学了”。

2.2 主题相关性判别

国内外学者对网页主题相关性判断方法进行了研究,总体而言,有效的判断方法有以下 5 类:元数据判定法、扩展元数据判定法、链接分析法、语义判定法及

基于特征词的向量空间模型判定法^[10],各有优势与不足。文中采用基于特征词的向量空间模型算法。

传统的基于特征词的向量空间模型算法^[11]没有充分考虑网页的元数据标签,主题判别的准确度不够高。文中对此进行改进,充分使用元数据标签改进传统的基于特征词的向量空间模型主题相关性判别方法,再结合就业领域本体实现主题相关性判别功能,实验证明改进后的方法能够提高主题判别的准确度。算法的具体流程如下:

(1) 获取就业领域主题特征词向量 T 和特征词权重向量 W , 其中 $T = \{t_1, t_2, \dots, t_n\}$, $W = \{\omega_1, \omega_2, \dots, \omega_n\}$, ω_i 表示特征词 t_i 所对应的权重。

(2) 通过就业领域概念间关系获取就业主题特征词 t_i 的上下位词和同位词,保存在数组 A_i 中。

(3) 解析网页 D , 分析网页结构, 确定是否含有 `<meta>` 标签, 如果有进一步分析是否有 `<keywords>` (关键词)、`<description>` (描述) 等标签。如果没有, 直接跳到步骤 5。

(4) 提取内容并分词。如果网页中有 `<keywords>`、`<description>` 等标签, 提取该部分内容并分词, 对每个单词 n 进行判断, 如果 n 在数组 A_i 中, 将 n 替换成 T_i , 并将相应的权重 ω_i 加 1, 反之则丢弃该网页。

(5) 提取网页 `<title>` 标签中的内容对每个名词 n 进行判断, 如果 a 在数组 A_i 中, 则将 n 替换成 A_i , 并将对应的权重 ω_i 加 1。

(6) 采用 $TF * IDF$ (term frequency - inverse document frequency)^[12] 算法计算网页 D 中所有就业领域主题特征词的权重 $D = \{d_1, d_2, \dots, d_n\}$ 。

(7) 根据式 1 对网页 D 进行主题相关性判别。

$$\text{Sim}(D, S) = \frac{\sum_{k=1}^n \omega_k \times d_k}{\sqrt{\left(\sum_{k=1}^n \omega_k^2\right) \left(\sum_{k=1}^n d_k^2\right)}} \quad (1)$$

2.3 检索结果排序

2.3.1 基于 MapReduce 的 LinkRank 并行排序算法

基于 Nutch 的就业垂直搜索引擎通过 WebClawer 抓取网页, 如果在 WebClawer 抓取网页结束后再通过 LinkRank 算法计算网页得分, 将会消耗大量时间。Nutch 平台本身支持 MapReduce 并行操作, 即抓取网页、网页解析、索引建立等操作都可以并行进行, 以提高效率。借鉴 MapReduce 并行编程思想, 把耗时较长的网页抓取和解析、建立索引操作转化为 Key-Value 的并行处理, 实现基于 MapReduce 的 LinkRank 并行算法。图 2 为具体工作流程。

2.3.2 网页的就业相关度计算

网页的就业相关度用来衡量网页和就业领域的相

关程度, 是指网页和就业主题的相关度。白晓丹^[13]通过层次分析理论以及模糊分析法构建网页相关性评价体系, 该方法具有一定的优势, 尤其在提取 PDF 文件时准确度较高, 但是 URL 的评价结果不太理想。文中结合就业领域本体计算网页的就业相关度。具体步骤如下: 对网页进行中文分词, 去掉敏感词等停用词; 提取网页特征词, 根据就业领域本体库获取各特征词的上下位词和同位词, 用一个三元组表示; 统计三元组中每组词在网页中出现的次数, 以及在多少个网页中出现, 根据前文描述的 $TF * IDF$ 算法计算它们的权重, 计算所得的权重为各特征词的权重; 使用向量空间模型表示网页。网页 D 可表示为: $D = (\omega_1, \omega_2, \dots, \omega_n)$, 其中 ω_i 表示网页 D 中第 i 个特征词的权重。同时将就业领域主题词用向量 K 表示为: $K = (k_1, k_2, \dots, k_n)$, 其中 k_i 表示主题词 i 的权重。

把网页的就业相关度计算, 转换为求 D 和 T 向量的距离。根据式 1 计算网页的就业相关度, $\text{Sim}(D, T)$ 越大则网页与就业的相关程度越高。

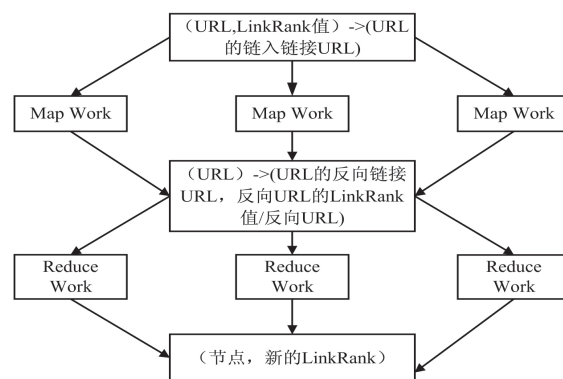


图 2 基于 MapReduce 的 LinkRank 并行算法工作流程

2.3.3 改进 Nutch 原有结果排序

Nutch 采用的 LinkRank 是一个 PageRank-like 的算法, 该算法通过构建 inlinks (链入链接数), outlinks (链出链接数) 和节点的列表, 然后由 LinkRank 类计算网页得分。LinkRank 非常接近最初的页面排序公式, 类似于:

$$(1 - \text{dampingFactor}) + (\text{dampingFactor} * \text{totalInlinkScore}) \quad (2)$$

其中, dampingFactor 是垃圾网页因子; totalInlinkScore 是页面链入链接数。

该算法有明显不足: 没有区分链入链接的重要性, 每个链入链接的权重一样; 偏重历史页面; 没有考虑页面内容与主题的相关程度。以上因素对网页的排序质量均有较大影响。因此做了如下改进:

(1) 增加链入链接权重因子。LinkRank 本身不区分链入链接的权威程度, 每个链入链接的权值一样, 然而, 网页中链入链接的权威度参差不齐, 如果采用相同

的权值,无法客观反映链入链接的重要性。甚至某些网页会故意添加广告链入链接进行作弊,以提高网站的排名。

文中给链入链接增加了权重因子,改进后的 total-InlinkScore 为:

$$\text{totalInlinkScore} = \sum_{i=1}^n \alpha_i * (\text{inlinkScore})_i \quad (3)$$

其中, α_i 为控制链入链接的权重因子。

通过控制链入链接权重控制各网页的权威度,避免了原算法中相同权重的不客观性,对防止通过作弊提高网页的得分有一定的效果,提高了检索质量。

(2) 增加时间衰减因子。LinkRank 算法通过网页间的相互链接关系计算网页得分,在互联网上放置时间越长的页面将得到更多的链入链接,而新的网页就算内容比较重要,但是由于链入链接数较少,得分则会相对较低。针对这一不合理情况,引入权重衰减时间因子对旧网页进行衰减处理。

在式 3 的基础上进行改进:

$$\text{totalInlinkScore} = \sum_{i=1}^n \alpha_i * (\text{inlinkScore})_i * 2^{\frac{\lambda}{t}} \quad (4)$$

其中, t 是页面创建至统计时的时间; λ 是常数因子。通过时间衰减因子 λ ,可以有效地调节新旧网页的权重,使得网页得分更接近真实情况。

2.4 用户查询接口扩展

2.4.1 关键字智能提醒

借助 WEB2.0 技术,就业垂直搜索引擎实现了自动记忆每次搜索信息的记录,同时通过就业专用词库,记录求职招聘特有的关键词,在用户输入查询关键字时在搜索框下方显示历史查询记录和该关键词的历史查询次数。通过关键字智能提醒扩展功能,搜索引擎不但可以记录用户输入的搜索关键词,了解用户习惯,并且可以形成与该搜索关键词相似的参考词组,简化用户操作。系统还可以根据关键词的搜索次数等因素,决定是否将此关键词自动识别为中文分词库中的关键词。

2.4.2 定制爬虫

利用 Java 后台开发技术 spring boot、spring data jpa 和 shiro,开发了求职招聘垂直搜索引擎的管理后台,通过后台可以对爬虫进行设置与管理,包括设定爬虫爬取时的深度与广度、设定爬取起始时间和间隔时间、Hadoop 集群管理参数设置(主从节点数量及状态设置)、索引库状态设置与管理等。

2.4.3 搜索辅助接口

除了基本的检索功能之外,还有其他的一些辅助功能,以提升用户体验增加用户粘度。比如:支持多关

键词同时搜索,如同时输入“Java”“MySQL”“软件工程师”;支持二次查找;设定查询结果的日期,比如最近一天、最近一周、最近一月、最近一年等;可指定搜索关键词在网页出现的位置,比如可设置为网页内的任何地方、标题或者网址;可设定搜索域,比如指定仅搜索指定网站或是将指定网站排除在搜索之外;提供搜索指引,在搜索首页与搜索结果列表页以多级分类方式提供用户搜索指引;提供历史搜索,在用户浏览器中保存用户最近搜索的 5 个关键词及其对应搜索时间,节省用户输入时间;订阅查询,在用户浏览器中可保存用户订阅的 5 个关键词,当用户进去搜索引擎首页或搜索结果列表页时,依次列出各关键词新增查询结果条数;个性化设置用户习惯,可根据用户个人使用习惯设置相关参数,如:每页显示结果数目、是否显示搜索框、设置网页语言搜索范围等;其他扩展功能如相关职位名推荐、热门词推荐、检索词高亮显示、按专业的关键词推荐等辅助功能。

3 测试与分析

3.1 测试数据准备

基于以上研究,从猎聘网下载职业词库以及从中国教育在线下载专业词库构建就业领域本体库,并选取十多个求职招聘网址,依次为智联招聘、前程无忧、58 同城招聘、赶集网招聘、猎聘网、中华英才网、卓博人才网、智通人才网、我的工作网、大街网、中国人才热线、应届生求职网、应届毕业生网、过来人求职网、拉勾网、大谷打工网、招聘信息、智通人才网等作为初始 URL 种子。

3.2 测试及结果分析

查全率和查重率是评价搜索引擎的 2 个重要指标^[14],理想状况下搜索引擎既有良好的查全率又有良好的查准率,而实际情况往往是两者不可兼得。文中从主题相关性、与其他搜索引擎比较这 2 个方面分析实验结果。

3.2.1 主题相关性分析

与百度谷歌等通用搜索引擎不同,基于 Nutch 的就业垂直搜索引擎设计了主题相关性判别,检索到的网页与主题相关程度,体现了垂直搜索引擎的针对性和专业性,主题相关率越高,专业性越强。主题相关率 = 与主题相关的网页数 / 检索出的网页总数量。文中选取“软件工程师”、“数据分析师”、“幼师”、“日语翻译”、“Java 开发”5 个关键词作为检索词,对检索结果的前 40 项进行分析,采用人工干预方式对改进 LinkRank 算法前后进行相关性分析,结果如表 1 所示,表中的数据“值 1-值 2”分别表示 LinkRank 算法改进后与改进前得到的值。

表1 主题相关性分析结果比较

关键词	结果总数	选取数	相关数	不相关数	主题相关率
软件工程师	847	30	30-29	0-1	100%-97%
数据分析师	230	30	29-28	1-2	97%-93%
幼师	200	30	30-27	0-3	100%-90%
日语翻译	2000	30	27-25	3-5	90%-83%
Java 开发	1 129	30	28-20	5-10	93%-67%
平均主题相关率	-	-	-	-	96%-86%

3.2.2 不同搜索引擎的比较

查准率、死链率、重复率、响应时间等是搜索引擎好坏的评价指标,查准率尤为重要。继续选用前文所述的“软件工程师”、“数据分析师”、“幼师”、“日语翻译”、“Java 开发”5 个搜索关键词进行检索并结合以上评价指标与百度、职友集进行对比分析,其中查准率、死链率、重复率、响应时间都采用平均值,由式 5 计算而得。

$$S = \frac{\sum_{i=1}^N \frac{n_i}{n}}{N} \quad (5)$$

其中, N 表示搜索关键词数量; n 表示网页数量(文中取 40); n_i 表示与用户需求有关的网页数量。

详细数据如表 2 所示,表 2 中的数据“值 1-值 2”表示改进 LinkRank 算法后和改进前的值。

表2 不同搜索引擎比较分析

搜索引擎	查准率/%	死链率/%	重复率/%	响应时间/ms
文中方法	94-92	0-0	0-0	65-60
职友集	90	1	1	55
百度	55	3	3	40

由表 2 数据可知,通用搜索引擎在响应时间方面有较大的优势,响应时间短,但是在查准率、死链率和重复率方面的劣势也较明显。文中基于 Nutch 的就业垂直搜索引擎虽然响应时间比其他搜索引擎长,但是在可接受范围内,在查准率、死链率和重复率方面有较明显优势。

4 结束语

文中设计并实现了基于 Nutch 的就业垂直搜索引擎,根据行业特色和需要对 Nutch 进行二次开发,加入 ik-analyzer 中文分词,同时运用就业领域本体改进基于特征词的向量空间模型,进行主题性判别,结合 LinkRank 算法改进排序结果。此外运用 Java 开发技术及框架,对用户查询接口进行了二次开发,增加了许多便于用户操作的功能,如关键词智能提醒、爬虫定制、高级检索、关键字订阅、定制用户显示界面、热门词推荐、关键词高亮显示等。基于 Nutch 的就业垂直搜

索引与通用搜索引擎相比方便了用户检索招聘信息,提高了求职招聘信息的查准率。

参考文献:

- [1] 人社部:高校毕业生人数创新高鼓励去基层就业[J].中国研究生,2016(8):64.
- [2] 荆德刚.2017 年高校毕业生就业的新特点与新机遇[J].中国高教研究,2017(7):27-30.
- [3] 柯进.2018 年高校毕业生将达 820 万[N].中国教育报,2018-02-27(1).
- [4] 袁威,薛安荣,周小梅.基于 Nutch 的分布式爬虫的优化研究[J].无线通信技术,2014,23(3):44-47.
- [5] BENASSI R, BERGAMASCHI S, VINCINI M. TUCUXI: the intelligent hunter agent for concept understanding and LeXical Chalning[C]//IEEE/WIC/ACM international conference on web intelligence.Beijing, China: IEEE, 2004: 249-255.
- [6] LIN Qingfeng, SCOTT S, SETH S C. A machine learning framework for automatically annotating web pages with simple HTML ontology extension(SHOE)[C]//Proceedings of IAWTIC 2001.[s.l.]: [s.n.], 2001.
- [7] MICARELLI A, GASPARETTI F, SCIARRONE F, et al. Personalized search on the world wide web[M]//The adaptive web.Berlin: Springer, 2007: 195-230.
- [8] 夏树倩.基于 Nutch 的学术搜索引擎的研究与实现[D].沈阳:东北大学,2011.
- [9] 纪晓阳.基于 Nutch 搜索引擎系统数据处理的中文分词技术的研究[D].成都:成都理工大学,2014.
- [10] 王超,李书琴,肖红.基于文献的农业领域本体自动构建方法研究[J].计算机应用与软件,2014,31(8):71-74.
- [11] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J].Communications of the ACM, 1975, 18(11):613-620.
- [12] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J].Information Processing and Management, 1987, 24(5):513-523.
- [13] 白晓丹.搜索引擎网页相关性评价及检索效率评价研究[D].北京:北京交通大学,2015.
- [14] 裴一蕾,薛万欣,赵宗,等.基于用户体验视角的搜索引擎评价研究[J].情报科学,2013,31(5):94-97.