

基于关键词相关性的有害信息爬虫系统研究

苏金波,朱剑宇,杨柳,刘跃
(合肥市公安局网安支队,安徽合肥 230039)

摘要:传统的互联网有害信息发现方法是依据 Google、百度等元搜索工具,用户输入关键词进行检索,然后对获取的结果进行研判,但是用户经常无法准确地描述所需的资料,给出的关键词不准确,搜索结果常有用户不关心的垃圾数据,而一些有用的数据却不能列出。文中探讨了一种基于元搜索,引入关键词扩充技术的爬虫方法。该方法在网页抓取,用户检索的时候能扩充输入的关键词,从而提高搜索覆盖率和精度。该方法投入小,效果好,还可通过扩展应用到其他领域。

关键词:元搜索;爬虫;关键词扩充;索引

中图分类号:TP302.1

文献标识码:A

文章编号:1673-629X(2014)03-0143-04

doi:10.3969/j.issn.1673-629X.2014.03.036

Research on Harmful Information Crawler System Based on Keywords Correlation

SU Jin-bo, ZHU Jian-yu, YANG Liu, LIU Yue

(Cyber Security Department Bureau of Public Security of Hefei, Hefei 230039, China)

Abstract: Traditional approaches to harmful information detection on the Internet are based on Google, Baidu etc., users enter keywords for search, and then need to study the results obtained, however users often do not accurately describe the information they want, the keywords given by users are inaccurate, the search results often include what users don't want, some data which users care cannot be listed. It explores a crawler method based on meta-search, which introduces technology of keyword expansion. The method expands keywords in the Web capture and user query to improve information coverage and accuracy, with low cost and good effect, which can be applied to other domain with some extension.

Key words: meta-search; crawler; keywords expansion; index

0 引言

随着 Web2.0 互联网技术的快速发展,用户已经广泛参与到数字资源的组织和描述活动中,用户不仅是资源的创造者,还是新一代资源描述者和组织者。网民们每天在互联网上要创造和分享海量的信息,这些信息鱼龙混杂,其中包含大量类似色情、反动、六合彩等未经审查核实的有害信息、非法信息,这些信息进入网络渠道,自由快速的传播造成信息污染,严重影响社会的稳定。虽然我国从多方面加强对互联网信息的监管,但因信息量巨大,技术上还存在一定困难,成效甚微。

互联网信息的监管,首要问题是如何快速准确地找到有害信息^[1],目前,发现有害信息的方法通常是靠公共通用搜索引擎(如百度、Google、搜狗、网站自身

的搜索功能)进行内容检索。通用搜索引擎根据用户输入的“查询串”与索引页面匹配程度的高低返回页面,然后再由监管人员对这些结果进行二次研判,找出其中有害信息。但实际上,普通监管人员大都不能十分准确地描述需要的资料,给出的关键字不准确,导致搜索引擎返回的结果动辄上百万,而真正能用到的结果却很少,尤其是遇到某些需要专业背景的资料时(比如搜索毒品名称),更是大海捞针。

为了提高有害信息发现能力,文中提出一种基于元搜索^[2],增加关键词相关性计算^[3]的改进型爬虫方法,目的在于提高关键词的覆盖率、准确性,从而提高搜索结果的信息覆盖率和检索精度,这种方法投入小,见效快,具有推广性。而且该方法可通过扩展应用到其他领域。

收稿日期:2013-06-03

修回日期:2013-09-08

网络出版时间:2014-01-07

基金项目:公安部重点研究项目;国家“863”高技术发展计划项目(2008AA01Z408)

作者简介:苏金波(1980-),男,研究方向为数据库与 Web 技术。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140107.1511.004.html>

1 系统功能和体系结构

有害信息爬虫系统的搜索引擎采用元搜索技术,所谓元搜索引擎并不是自己搜索网页,而是将用户的查询请求同时向多个搜索引擎递交,将返回的结果进行去重、排序等处理后,以统一的格式返回给用户。服务方式为面向网页的全文检索,它省去了多次查询的麻烦,提高了单次查询的查全率。

基于关键词扩充技术的元搜索基本思想是根据初始关键词扩充关键词库,增加相关性强的其他关键词,并将关键词发送到通用搜索引擎,对得到的结果进行判别分析,若属于该主题,则保留该网页为后面建立索引做准备;若不属于该主题则抛弃,避免占用更多空间。

1.1 功能设计

该系统采用基于 B/S 的服务器模式^[4],功能模块主要有互联网网页信息采集功能;原始网页分析功能,主要包括文本内容的抽取、标点符号的过滤、内容分词等;为网页文本建立高效的全文索引库;提供查询接口;实现网页定位;还能通过链接分析、关键字关联性等算法为用户提供尽可能精确的结果,并能对检索结果进行排序。

1.2 体系结构

该系统的体系结构如图 1 所示。

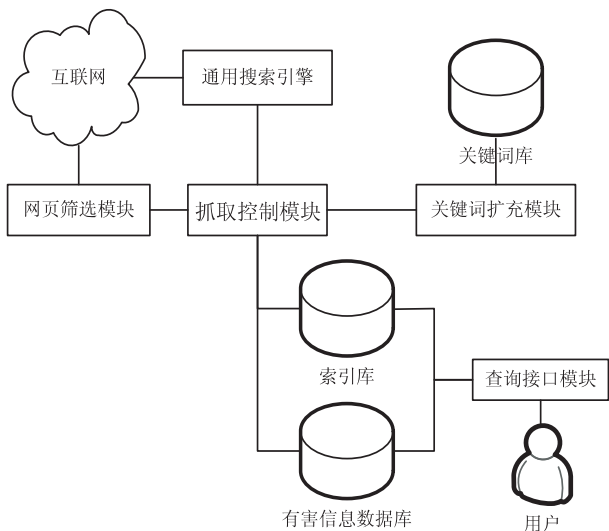


图 1 有害信息爬虫系统的体系结构

(1) 关键词扩充模块:对关键词库进行扩充,从关键词库循环取出关键词作为输入,输入到 Google 关键词工具^[5],对查询结果按本地搜索量进行倒序排列,取前十个,重新存入关键词库,并与原始关键词建立对应关系。

(2) 网页筛选模块:对通用搜索引擎的检索结果,进行页面信息分析,利用 HtmlParser^[6]将网页转换成文档对象模型树(DomTree)^[7],对模型树进行剪枝,去除页面中重复部分和与主题无关部分,从剩下的模型

树中获取信息块,以及获取指向其他页面的 URL 链接,通过这个步骤的循环可以继续选择访问该 URL 链接的页面,继续采集其页面信息。

(3) 抓取控制模块:该模块是整个爬虫系统的核心控制模块,主要功能为从关键词库中取出关键词,调用通用搜索引擎模块进行检索,检索结果通过网页筛选模块返回,然后对返回结果采用开源技术 Lucene^[8]建立索引,并将索引存入索引库,同时将返回的网页提取内容存入到有害信息数据库。

2 系统关键技术

2.1 种子 URL 列表的生成

所谓种子 URL 列表是指系统初始化时,爬虫开始抓取的起始 URL 集合,通过这些种子 URL 链接开始依次发现其他的网页,若其他网页中有超链接,则提取这些链接,然后再重复进行上述过程,从而完成自动爬取。有害信息爬虫监控系统属于主题爬虫^[9]类型,它关注于一些特定的网站,如论坛、微博等,为了节省存储、网络资源等成本,并没有重新设计一个爬虫系统,而是采用人工输入和元搜索提取相结合的种子 URL 列表生成策略。人工输入是指将关注度高的重点网站在系统中进行预设;元搜索策略是从关键词库中取出主体关键词从元搜索引擎中下载相关的结果,得到的结果中包含大量的链接地址,通过对地址的分析,可以得到这些地址的上级目录或网站,将与主题相关的网站或地址加入到种子 URL 列表中以进一步发现更多主题网站。爬虫依据种子 URL 列表,采用广度优先或深度优先算法^[10]进行扩展搜索,可不断地将新增 URL 加入到列表中。对于一些关注度大、有害信息多的重点网站,可以在 URL 列表库里对其赋予较高的权重。

2.2 扩充关键词库

系统初始化时,需要用户设置一组关键词,但用户绝不可能将涉及某个主题的关键词设全,为了尽可能提高信息覆盖范围,提高信息精度,需要在用户提供的关键词基础上,对关键词库进行扩展。

文中采用的扩充关键词词库的方法,理论思想主要依据关键词相关性程度计算方法,通过判断主题 Web 页面里关键词与跟它共同出现的其他关键词的相关性来确定该关键词与该主题的相关性程度的方法。如果两个关键词共同出现在同一 Web 页面中,则认为它们是共同出现的。下面给出共同出现的两个关键词相关性权重计算方法:

$$W_{ij} = \frac{P(a_i \wedge a_j)}{\sum P(a_i)}$$

其中, $P(a_i \wedge a_j)$ 表示主题页面 Web 集中同时有关键词 a_i 与 a_j 出现的页面个数; $P(a_i)$ 表示主题页面 Web 集中关键词 a_i 出现的所有页面个数; w_{ij} 反映了关键词 a_i 与关键词 a_j 之间的相关性权重。依据共同出现关键词之间的相关度模型可以推导出关键词本身与某主题领域的相关度。

目前,很多元搜索工具,依据关键词相关性理论已经开发出很多不错的关键词扩充方法,主要有:

- (1) 关键词工具,如谷歌关键词工具和百度指数。
- (2) 搜索查询,在百度和谷歌搜索框中输入关键词,搜索框中会自动显示与此相关的其他关键词,可以根据这样的方式得到搜索量较多的关键词,这些都是利用价值较高的关键词。

(3) 相关搜索,通用搜索引擎搜索结果页面下面通常会有其他的相关搜索关键词。

文中采用 Google 关键词工具对关键词库进行扩展。查询任何一个关键词,谷歌工具都会列出至少几十个相关关键词。取出搜索量排名前十的放入关键词库。如表 1,例如通过关键词“汽狗”可扩充出“气枪”、“汽鎗”、“汽槍”、“枪网”、“铅彈”等一系列相关性极强的关键词。

表 1 Google 关键词工具示例表

关键词	全球每月搜索量	本地每月搜索量
气鎗	14 800	14 800
气枪	14 800	14 800
气槍	14 800	14 800
枪网	8 100	6 600
槍網	8 100	6 600
枪专卖店	4 400	3 600
出售槍支	2 900	2 400
铅彈	1 900	1 900

2.3 页面识别抽取技术

当前的互联网广泛采用了动态页面生成技术,不同的页面往往是利用同一个页面模板动态生成的,因此在确定一个有害信息页后,可以将其作为信息抽取模板,其他同一个网站内的页面可通过计算与模板页面间的相似度来判断是否为有价值的目标页面。

这些页面中常常伴有噪音,如广告、导航、标题等非目标区域,为了提高信息抽取的效率和准确性,文中采用一种基于 MDR^[11] 的算法,结合 DOM 树和视觉特征的多证据网页信息抽取方法(DVF)^[12],该方法采用搜索候选目标数据区域的(Search Data Regions, SDRs)的算法。在把页面解析成 DOM 树后,利用 SDRs 从根节点开始向下递归搜索数据区域,具体如下:

SDRs(Node, ST, WT)

Input: Node——任意 DOM 树节点

ST——相似度閾值

WT——宽度閾值

Output: DataRegions——树节点 Node 下所有候选目标数据区域集合

```
1 if TreeDepth(Node) >= 3 then
2   DataRegions = IdenDRs(Node, ST, WT)
3   for each Child ∈ Node.Children && Child ∉ DataRegions
4     DataRegions = DataRegions ∪ SDRs(Child, ST, WT)
5   return DataRegions
6 else return ∅
7 end
```

以上第 1 行表示如果 Node 的子树深度为 1 或 2 时,算法将不会搜索数据区域,因为数据记录节点不太可能少于两层标签。函数 IdenDRs(Node, ST, WT) 返回 Node 的孩子节点子树构成的数据区域集合。该函数首先通过判断节点的宽度是否满足一定閾值,和相邻的兄弟节点的宽度是否一致,预先过滤掉相当一部分的噪音,以减少后面节点子树的比较,接着通过计算字符串编辑距离不断比较相邻兄弟节点子树是否相似,得到候选目标数据区域集合。SDRs 算法的时间复杂度和空间复杂度均为 $O(n)$, n 是根为 Node 的树的节点个数。

IdenDRs(Node, ST, WT)

Input: Node——任意 DOM 树节点

ST——相似度閾值

WT——宽度閾值

Output: DataRegions——由 Node 的孩子节点组成的候选目标数据区域集合

```
1 n = NumberOf(Node.Children)
2 c = Node.Children
3 LastNode = null
4 DataRegions = null
5 for i = 1 to n - 1
6   if c[i].width < WT || c[i].width != c[i + 1].width
7     continue
8   if TreeSimilar(c[i], c[i + 1], ST) != false
9     continue
10  if c[i] != LastNode
11    DR = (c[i], c[i + 1])
12    DataRegions = DataRegions ∪ DR
13    LastNode = c[i + 1]
14  else
15    DataRegions = DataRegions - c[i].DR
16    c[i].DR = c[i].DR ∪ c[i + 1]
17  DataRegions = DataRegions ∪ c[i].DR
```

第 6、7 行表示利用视觉特性进行噪音过滤;第 8、9 行利用字符串编辑距离判断相邻兄弟节点子树是否

相似,从而对数据区域进行过滤;第 10~13 行表示发现了一个新的候选目标数据区域;第 14~17 行是把一条新的数据记录添加到已有数据区域,并更新该区域至区域集合。

从视觉上看,一般页面中所占的版面面积最大的即为目标数据区域,假设由前述 SDRs 算法得到 n 个候选目标数据区域 DR_1, DR_2, \dots, DR_n , 根据各区域的长、宽计算区域面积分别为 S_1, S_2, \dots, S_n , 排序得最大面积为 $S_t = \text{Max}\{S_1, S_2, \dots, S_n\}$, $1 \leq t \leq n$, 则对应的数据区域 DR_t 即为目标数据区域,其余视为噪音。

2.4 用户搜索

考虑到用户常常无法全面地描述自己想要的资料,当用户利用该系统进行检索的时候,查询接口模块应根据用户输入的关键词,采用关键词工具进行扩展得到一组相关的关键词集合,再通过索引库进行查找。采用此方法,用户仅需要输入少数几个关键词,基本上就能得到全部信息,大大提高了信息覆盖率,减少了用户检索次数,从而提高效率。

图 2 为用户检索数据流程图。

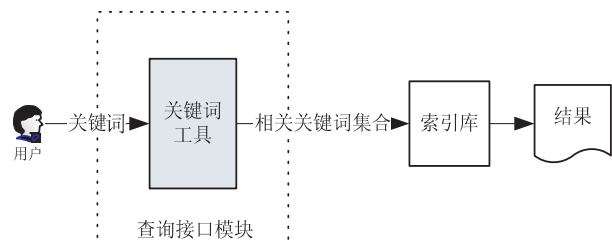


图 2 用户检索数据流程图

3 结束语

传统的爬虫搜索方法不考虑搜索关键词的相关性问题,搜索结果常带有垃圾文档,而用户关心的文档却

未找到,信息覆盖率、准确性都有待提高。文中对传统元索引进行扩展,引入关键词相关性技术,采用关键词扩充工具,从原始关键词关联出更大的一组关键词库,利用这些关键词库进行爬虫搜索,不仅提高了信息的覆盖率和精度,还大大节省网络、硬件等资源,而且很容易通过扩展应用到其他主题搜索领域。

参考文献:

- [1] 马民虎. 互联网信息内容安全管理教程[M]. 北京:中国人民公安大学出版社,2008.
- [2] Metasearch engine [EB/OL]. 2010. http://en.wikipedia.org/wiki/Metasearch_engine.
- [3] 刘耕,方勇,刘嘉勇. 基于关联词和扩展规则的敏感词库设计[J]. 四川大学学报(自然科学版),2009,46(3): 667-671.
- [4] Deitel H M. Java Web services for experienced programmers [M]. 北京:机械工业出版社,2003.
- [5] Google AdWords [EB/OL]. 2013. <https://adwords.google.cn/o/KeywordTool>.
- [6] Html Parser 2.0 [EB/OL]. 2010. <http://htmlparser.sourceforge.net/>.
- [7] XML Path Language (XPath) 2.0 (Second Edition) [EB/OL]. 2010-12-14. <http://www.w3.org/TR/2010/REC-xpath20-20101214/>.
- [8] Apache Lucene [EB/OL]. 2013. <http://lucene.apache.org/>.
- [9] Focused crawler [EB/OL]. 2013. http://en.wikipedia.org/wiki/Focused_crawler.
- [10] 王能斌. 数据库系统教程[M]. 北京:电子工业出版社,2002.
- [11] 关毅璋,郝志峰. 随机变点统计的MDR边缘检测算法[J]. 计算机应用研究,2009,26(1):384-386.
- [12] 安增文,徐杰峰. 基于视觉特征的网页正文提取方法研究[J]. 微型机与应用,2010(3):38-41.

(上接第 142 页)

- [8] Stütz T, Uhl A. Efficient format-compliant encryption of regular languages: Block-based cycle-walking [C]//Proc of the 11th IFIP TC 6/TC 11 Int'l Conf. Linz: Springer-Verlag, 2010:81-92.
- [9] Schneier B, Kelsey J. Unbalanced Feistel networks and block cipher design [C]//Proc of the fast software encryption'96. Cambridge: Springer-Verlag, 1996:121-144.
- [10] Spies T. Feistel finite set encryption mode [EB/OL]. 2008. <http://csrc.nist.gov/groups/ST/toolkit/BCM/documents/proposedmodes/ffsem/ffsem-spec.pdf>.

- [11] Bellare M, Rogaway P, Spies T. The FFX mode of operation for format-preserving encryption [EB/OL]. 2010. <http://csrc.nist.gov/groups/ST/toolkit/BCM/documents/proposedmodes/ffx/ffx-spec.pdf>.
- [12] Eric B, Thomas P, Jacques S. BPS: A format-preserving encryption proposal [EB/OL]. 2010. <http://brutus.ncsl.nist.gov/groups/ST/toolkit/BCM/documents/proposedmodes/bps/bps-spec.pdf>.

基于关键词相关性的有害信息爬虫系统研究

作者：[苏金波](#)，[朱剑宇](#)，[杨柳](#)，[刘跃](#)，[SU Jin-bo](#)，[ZHU Jian-yu](#)，[YANG Liu](#)，[LIU Yue](#)

作者单位：[合肥市公安局网安支队, 安徽 合肥, 230039](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

ISTIC

年，卷(期)：2014(3)

本文链接：http://d.wanfangdata.com.cn/Periodical_wjfz201403036.aspx