# 一种基于搜索策略的多主题信息采集方法

仲兆满[1],李存华[1],刘宗田[2],管 燕[1]
(1.淮海工学院计算机工程学院,江苏连云港 222000;2.上海大学计算机学院,上海 200072)

**摘 要:** 本文针对多主题信息采集效率低下的问题,调研了主题规则在内置搜索引擎和通用搜索引擎上搜索结果的差异,提出将主题规则拆分成原子规则的思想,分析了原子规则间的相同、互换、包含三种关系.在原子规则之间关系的基础上,设计了针对内置搜索和通用搜索不同的原子规则分配策略,这样做一方面提高主题信息采集的准确率,另一方面减少搜索采集的次数.针对原子规则直接搜索结果的准确率不高的问题,提出了基于句群的主题与信息相关性的过滤方法.设置 138 条主题规则(拆分后的原子规则为 8223 条),14 个内置搜索引擎和 4 个通用搜索引擎,在单位时间内采集到的信息总条数与采集到的相关信息的条数两个方面进行了实验比较.结果表明,所提方法在信息采集数目及相关信息采集数目方面均具有较好的性能.

**关键词:** 多主题信息采集; 原子规则; 内置搜索; 通用搜索; 相关性计算
**中图分类号:** TP311 **文献标识码:** A **文章编号:** 0372-2112 (2014)12-2352-07
**电子学报 URL:** http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.12.003

# A Method of Multi-Topic Crawling Based on Search Strategy

ZHONG Zhao-man[1], LI Cun-hua[1], LIU Zong-tian[2], GUAN Yan[1]
(1. *School of Computer Engineering, Huaihai Institute of Technology, Lianyungang, Jiangsu* 222000, *China*;
2. *School of Computer, Shanghai University, Shanghai* 200072, *China*)

**Abstract:** Aiming at the low efficiency of multi-topic crawling, the difference between built-in search engines (BSEs) and general search engines (GSEs) is investigated. The idea and method of dividing topic rules into atomic rules are proposed respectively, and three relations (equating relation, exchanging relation and containing relation) between atomic rules are analyzed. Based on atomic rule relations, the different allocation strategies for BSEs and GSEs are designed, which can not only improve the precision of topic-specific crawling, but also reduce crawling times. Furthermore, a method of sentence cluster-based relevance computing between topics and documents is proposed to solve the low precision problem of directly crawling information by atomic rules. We conduct an experiment with 138 topic rules (containing 8223 atomic rules), 14 BSEs and 4 GSEs for evaluating the number of crawling information and related information in unit time. The results show that the proposed method offers more effective performances.

**Key words:** multi-topic crawling; atomic rules; built-in search engines; general search engines; relevance computing

## 1 引言

面对海量的互联网信息,借助搜索引擎获取相关信息已是人们日常生活中的常态行为.1997 年和 1998 年相关的研究表明,即使大型的信息采集系统对 Web 的覆盖率也只有 30%左右[1,2].美国华盛顿大学的研究者认为大多数搜索引擎对于同一个查询请求返回的结果很不相同,质量也参差不齐[3].Hsu 等[4]通过实验发现 2005 年 Google 对互联网上 URL 的覆盖率为 30 ~ 40%,2006 年覆盖率为 60%.随着 WWW 信息的爆炸性增长,采集的速度也越来越不能满足实际应用的需要.

主题信息采集是指有选择性地采集那些与预先定义好的主题相关信息的行为.自 1999 年 Chakrabarti 等发表了第一篇面向主题的信息采集论文以来[5],迄今为止,主题信息采集一直是互联网信息处理研究的热点问题.本文研究了基于规则的主题表示方法、规则关系判别方法及主题与信息相关性判定方法,主要的创新点体现在:(1)针对互联网信息监测系统经常需要面对多个

收稿日期:2013-09-24;修回日期:2013-12-19;责任编辑:覃怀银
基金项目:国家自然科学基金(No.60975033,No.61273328)

用户定义多个主题,主题对应大量的规则,主题之间的规则又有着内在本质的关系,提出了主题原子规则的概念,研究了主题规则到原子规则的拆分方法,原子规则之间的相同、互换、包含三种关系的判定方法;(2)对主题信息采集而言,揭示了内置搜索与通用搜索的不同,在原子规则关系的基础上,提出了针对内置搜索引擎和通用搜索引擎不同的采集调度策略;(3)围绕具体主题采集,提出了基于句群的主题与信息相关性的过滤方法.

本文的结构安排如下:第二章介绍了相关的研究现状,重点是主题的表示、主题与信息的相关性计算;第三章详细的阐述了本文提出的方法;第四章进行了实验分析与比较;第五章对全文进行了小结.

## 2　研究现状

### 2.1　基于规则的主题表示方法

基于规则的主题表示方法使用若干关键字及其"与"、"或"关系.基于关键字的主题表示方法是基于规则的表示方法的简化,仅罗列若干关键字(可以包含权重),而不强调关键字之间的关系,比如文献[5~7].著名的 Fish-Search 算法[8]以用户输入的查询关键字为主题,通过字符串匹配来判定哪些信息包含主题关键字,体现了关键字之间的"与"关系.对信息检索而言,很多主流的搜索引擎都支持关键字的"与"、"或"关系,比如百度、谷歌、雅虎中文、搜狗、新浪等,一些典型的网站提供的内置搜索也支持关键字之间的"与"、"或"关系,比如新浪微博、搜狐微博、腾讯微博、人民网、百度贴吧等.文献[9]针对用户获取事件类信息的需求,讨论了事件要素之间的约束关系,可以理解为关键字之间的"与"关系.

为了将规则中的关键字爬升到概念的语义级别,Rodrigo 等[10]提出了一个分布式式的主题采集器的框架,使用了基于本体的知识表示方法驱动采集器从 Web 上获取特定的信息;Punam 等[11]以本体为主题采集的扩展语义源,提高了信息采集器的覆盖率;与一般概念本体不同的是,Yang[12]提出了本体支持的网站模型,一个网页包含三个属性:基本信息,统计信息和本体信息.

### 2.2　主题与信息的相关性计算

为了高效地抓取与主题相关的信息,研究者们提出了许多主题与信息的相关性计算方法,大体可以分为两类:基于网页内容的方法和基于 Web 链接分析的方法.

基于网页内容的方法主要是利用网页标题、正文、锚文本等文字信息.比较有代表的是 De Bra 等提出的 Fish-Search 算法[8],该算法以用户输入的查询关键字为主题,通过字符串匹配来判定哪些页面包含主题内容,包含就认为是主题相关的,但是这种方法无法对页面的相关度高低进行排序.Hersocvic 针对 Fish-Search 算法的问题对其进行改进,提出了 Shark-Search 算法[13],对链接价值的计算采用连续值的相似度函数,这样就能通过数值来判断网页与主题的相关性大小.向量空间模型是广泛使用的主题与信息相似度的计算方法,比如文献[14]将 Web 页面和用户的查询表示为向量,应用余弦距离计算页面与查询主题的相似度.文献[9,15]将事件的诸要素表示为向量,考虑了事件要素的不同作用对要素的权值做了优化调整,取得了较好的相似度计算效果.

基于 Web 链接分析的方法主要是依据文献计量学的引文分析理论,认为入链接和出链接比较多的页面价值也比较高.Martinez 等[16]借鉴锚标签、URL 及页面包含的链接指导发现新的采集页面.Liu 等[17]使用爬取信息路径序列的概率模型,给出了跳到目标距离的计算方法.Du 等[18]提出使用概念上下文图存储基于用户点击历史的知识,进而指导爬虫采集的方向.Torkestani[19]设计了带有学习机制的采集器,通过计算信息与主题的相似度,决定是否沿着该条路径继续爬取信息.高凯[20]研究了搜索引擎中信息动态采集策略,采用基于网页从属关系和内容分析的相关性来调节信息采集的过程.

Mohsen 等[21]采取链接分析和文本相似度结合的方法提出了一种"主题"爬虫的方法;Yuvarani 等[22]也有相似的思路,把链接中的关键字和链接周围文本的语义作为提取本体的语料,设计了一个 Java 语言实现的多线程"主题"爬虫.Melanie 等[23]提出了使用 HTML 元数据信息构建分类器,从社交媒体上收集 Web 评论数据.相关的研究文献[19,24]表明,融合内容与链接分析的方法能显著的改善主题与信息相关度计算的效果.

## 3　基于搜索策略的多主题信息采集方法

### 3.1　相关定义

**定义 1(主题信息采集)**　主题信息采集指从互联网上有选择性地采集那些与预先定义的主题相关的信息的行为,定义为:$D=(W,T,F,R(T,w_i))$,其中,$D$ 指根据主题从互联网上获取的最终信息集合,$W$ 代表互联网上所有的信息,$w_i\in W$,$T$ 代表主题,$F$ 指主题及信息的表示框架,$R(T,w_i)$指主题 $T$ 与一篇信息 $w_i$ 的相关性计算方法.可见,$D\subset W$.

**定义 2(主题内置搜索采集)**　主题内置搜索采集指使用网站自带的用于对网站自身内容进行搜索的引擎而采集的与预先定义的主题相关的信息的行为,定义为:$D_i=(W_i,T,F,R(T,w_i^j))$,其中,$D_i$ 代表根据主

题从网站 $W_i$ 上获取的最终信息集合.可见,$D_i \subset W_i$.

**定义 3(主题通用搜索采集)** 主题通用搜索采集指从多个独立的搜索引擎的结果中获取那些与预先定义的主题相关的信息的行为,定义为:$D_m = (W_s, T, F, R(T, w_s^j))$,其中,$D_m$ 代表最终从多个搜索引擎获取的信息集合,$W_s$ 代表 $n$ 个独立的搜索引擎 $S_1, S_2, \cdots, S_n$ 的信息集合.可见,$D_m \subset W_s$.

**定义 4(基于规则的主题表示)** 基于规则的主题表示指使用关键字之间的"与"、"或"关系描述主题,定义为:$T = C(K, U)$,其中 $K$ 代表主题 $T$ 包含的关键字集合,$K = \{k_1, k_2, \cdots k_m\}$,$U$ 代表关键字之间的关系集合,$U = \{*, +\}$,$*$ 代表"与"关系,$+$ 代表"或"关系,约定"$*$"关系运算级别高于"$+$"关系,$C$ 代表 $K$ 与 $U$ 之间的组合方式.

**定义 5(原子规则)** 原子规则指对主题的表示规则进行拆分,拆分后的关键字之间仅仅存在"与"的关系.比如主题规则 $T = 2008 * (汶川 + 四川) * 地震$,拆分后得到两个原子规则分别是 $R_1^a = 2008 * 汶川 * 地震$ 和 $R_2^a = 2008 * 四川 * 地震$.可见,单个关键字是原子规则的特殊情况.

### 3.2 系统模型

基于搜索策略的多主题信息采集方法的模型如图 1 所示.

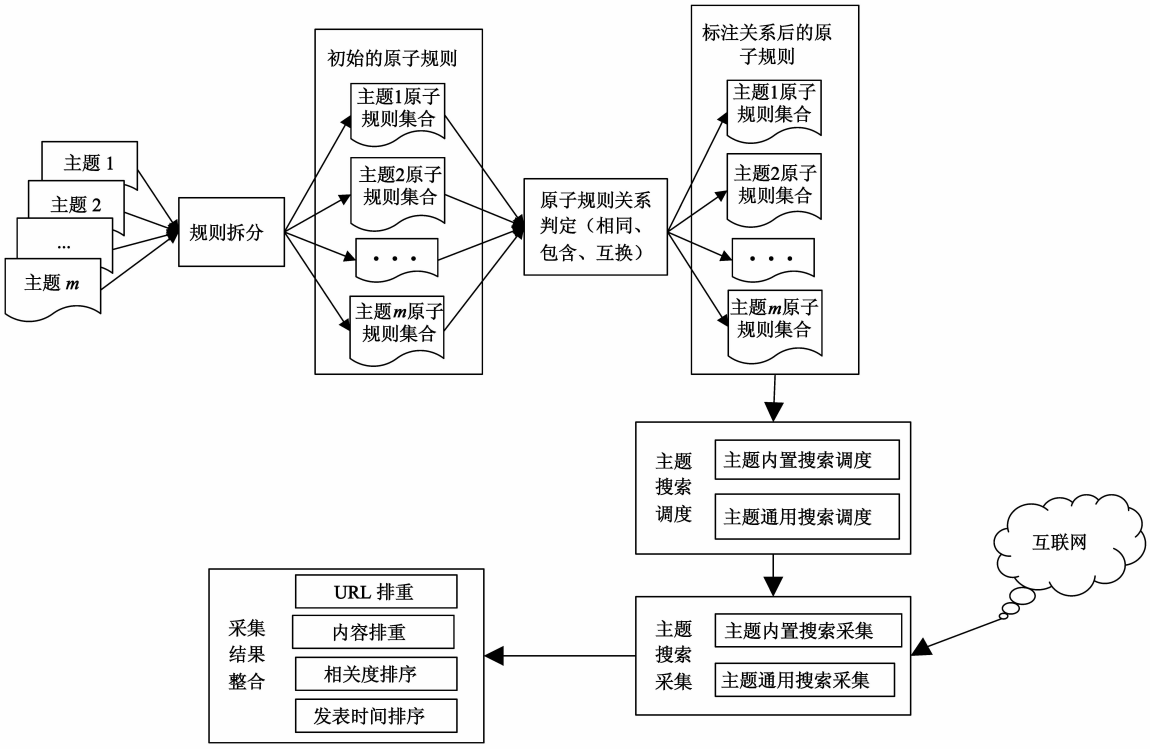


**图1 多主题信息采集系统模型**

图 1 所示模型的关键组件有:(1)规则拆分,用于把主题规则拆分成原子规则;(2)原子规则关系判定,将原子规则之间的关系区分为三种情况:相同关系、包含关系和互换关系;(3)主题搜索调度,根据原子规则之间的关系,分别选取原子规则调度至主题内置搜索和主题通用搜索;(4)主题搜索采集,执行主题内置搜索采集和主题通用搜索采集;(5)采集结果整合,包括 URL 排重、信息过滤、信息排序等步骤.以上五个组件中,(4)中的技术、(5)中的信息排重、信息排序等技术已经比较成熟,但根据原子规则对采集信息进行过滤还不精准.因此本文重点研究的是组件(1)、(2)、(3)及(5)中的信息过滤,共计 4 个核心技术点.

### 3.3 主题规则拆分

(1)为何要拆分主题规则?

如定义 4 所述,主题规则是由主题的关键字及其"与"、"或"关系构成的.比如一条主题规则 $R_i$ 为"南京 * 检察院 * (冤案 + 打人)",这条规则的意思是从互联网上采集与"南京检察院冤案"或"南京检察院打人"相关的信息.将 $R_i$ 放到通用搜索引擎,浏览返回的结果.在百度搜索引擎中,检索规则格式变为"南京检察院(冤案|打人)",空格代表"与"关系,"|"代表"或"关系.浏览后发现,返回的总共 760 条信息中,约 700 条都是不相关的,准确率只有 10% 左右.而将此主题规则拆分成两条完全"与"关系的规则"南京 检察院 冤案"和"南

京 检察院 打人”,分别用这两条规则进行搜索,在返回的 1560 条信息中,约有 600 条有一定的相关性,准确率近 40%.同样的,将规则 $R_i$ 放入到谷歌搜索引擎中,格式变为“南京 检察院 冤案 OR 打人”,空格代表“与”关系,“OR”代表“或”关系.浏览后发现,返回的总共 750 条信息中,约 600 条都是不相关的,准确率约 20%.而将两条仅包含“与”关系的规则分别进行搜索,在返回的 1500 条信息中,约有 650 条有一定的相关性,准确率约 43%.

经过对谷歌、百度、必应等主流的通用搜索引擎调研后发现,主题规则拆分成原子规则后可以显著的提高检索结果的准确率.

(2)主题规则拆分算法

Input: 主题规则集 $R$

Output: 原子规则集 $R^a$

算法流程:

**Step 1** 依次从 $R$ 中取出每条主题规则 $R_i$, $R^a$ 置空;

**Step 2** 判段 $R_i$ 中是否包含“+”关系.如果有,则转 Step 3;否则 $R^a = \{R_i\}$,转 Step 5;

**Step 3** 判断 $R_i$ 中是否存在“与”分配律.如果有 $m$ 个($m>0$),则循环执行 $m$ 次“与”分配律运算,得到 $R_i'$;否则,$R_i' = R_i$,转 Step 4;

**Step 4** 依据“+”关系将 $R_i'$ 切分成 $n$ 个原子规则,将 $n$ 个原子规则放入 $R^a$ 中,转 Step 5;

**Step 5** 输出 $R^a$,得到原子规则集.

### 3.4 原子规则关系及搜索结论

(1)原子规则关系

原子规则之间的关系包括三种情况:

**关系 1(相同关系)** 两条原子规则 $R_1^a$, $R_2^a$,经过“*”关系切分后,如果关键字出现的顺序完全一致,则这两条规存在相同关系,记作 $R_1^a = R_2^a$.比如,$R_1^a = \{A * B\}$, $R_2^a = \{A * B\}$,则 $R_1^a = R_2^a$.

**关系 2(互换关系)** 两条原子规则 $R_1^a$, $R_2^a$,经过“*”关系切分后,如果关键字完全相同,但出现的顺序不一致,则这两条规则存在互换关系,记作 $R_1^a \approx R_2^a$.比如,$R_1^a = \{A * B * C\}$, $R_2^a = \{A * C * B\}$,则 $R_1^a \approx R_2^a$.

**关系 3(包含关系)** 两条原子规则 $R_1^a$, $R_2^a$,经过“*”关系切分后,如果 $R_1^a$ 的关键字是 $R_2^a$ 关键字的真子集,则这两条规则存在包含关系,记作 $R_1^a \subset R_2^a$.比如,$R_1^a = \{A * B\}$, $R_2^a = \{A * C * B\}$,则 $R_1^a \subset R_2^a$.

(2)不同关系的原子规则搜索结论

使用不同关系的原子规则,经过对 14 个内置搜索引擎和 4 个通用搜索引擎进行搜索,得到的结论如下:

($a$)如果 $R_1^a \approx R_2^a$,则用 $R_1^a$ 或 $R_2^a$ 在内置搜索引擎中得到的结果是一样的,区别仅在于不同的排序.对于内置搜索引擎,使用 $R_1^a$ 或 $R_2^a$ 搜索一次即可,勿需搜索两次.

($b$)如果 $R_1^a \approx R_2^a$,则用 $R_1^a$ 或 $R_2^a$ 在通用搜索引擎中得到的结果是不一样的.对于通用搜索引擎,需要使用 $R_1^a$ 和 $R_2^a$ 搜索两次.

($c$)如果 $R_1^a \subset R_2^a$,用 $R_1^a$ 在内置搜索引擎中得到的结果集记作 $D_{R_1^a}$,用 $R_2^a$ 在内置搜索引擎中得到的结果集记作 $D_{R_2^a}$,则 $D_{R_2^a} \subset D_{R_1^a}$.对于内置搜索引擎,使用 $R_1^a$ 搜索一次即可,勿需搜索两次.

($d$)如果 $R_1^a \subset R_2^a$,则用 $R_1^a$ 或 $R_2^a$ 在通用搜索引擎中得到的结果是不一样的.对于通用搜索引擎,需要使用 $R_1^a$ 和 $R_2^a$ 搜索两次.

($e$)如果 $R_1^a = R_2^a$,则用 $R_1^a$ 或 $R_2^a$ 在内置搜索引擎、通用搜索引擎中得到的结果是一样的,使用 $R_1^a$ 或 $R_2^a$ 搜索一次即可,勿需搜索两次.

### 3.5 搜索调度

搜索调度的算法如下:

Input: 原子规则集 $R^a$

Output: 内置搜索队列 $Q^I$、通用搜索队列 $Q^C$

算法流程:

**Step 1** 遍历所有原子规则,没有关系的原子规则直接分别放入 $Q^I$、$Q^C$.如果某些原子规则有关系,记有关系的原子规则集为 $R^{a\prime}$,转 Step 2;否则转 Step 6;

**Step 2** 从 $R^{a\prime}$ 取出一条原子规则,判断它与其他原子规则之间的关系.如果两条原子规则 $R_1^a = R_2^a$,则选 $R_1^a$ 或 $R_2^a$ 分别放入 $Q^I$、$Q^C$,转 Step 5;否则,转 Step 3;

**Step 3** 如果 $R_1^a \approx R_2^a$,则选取选 $R_1^a$ 或 $R_2^a$ 放入 $Q^I$,选 $R_1^a$ 和 $R_2^a$ 放入 $Q^C$,转 Step 5;否则转 Step 4;

**Step 4** 如 $R_1^a \subset R_2^a$,则选取 $R_1^a$ 放入 $Q^I$,选 $R_1^a$ 和 $R_2^a$ 放入 $Q^C$,转 Step 5;

**Step 5** 判断 $R^{a\prime}$ 中的原子规则是否全部取完,完成则转 Step 6;否则,转 Step 2;

**Step 6** 分别输出 $Q^I$ 和 $Q^C$.

### 3.6 基于句群的主题与信息相关性的过滤方法

依据原子规则从内置搜索引擎或通用搜索引擎采集到的信息必然包含了原子规则中的每个关键字,但包含了原子规则中的每个关键字的信息并不一定是与主题密切相关的.

文档信息在对具体主题描述时,一般都会在连续的几个句子中交代清楚该主题涉及的关键要素,比如“时间”、“事件”、“地点”和“人物”等.本文通过实验的方法选取的句群单位是三个连续的句子,句子的分隔符号是“.”、“?”和“!”,文档标题看作是一个句子.如果

文档的最后一句话,不管标点符号是哪种,统一作为一个句子看待.在文档正文的开头或结尾处如果有媒体来源的标识,这些部分不作为文档信息句子的内容.假设,当前句子为 $S_i$,则基于句群的主题与信息相关性的过滤算法定义如下:

$$\mathrm{Sim}(R^a, d_i) = \begin{cases} 1, & \text{if } \mathrm{Contain}(S_{i-1}S_iS_{i+1}, K(R^a)) = 1 \\ 0, & \text{else} \end{cases} \tag{1}$$

其中,$S_{i-1}S_iS_{i+1}$指文档 $d_i$ 中的句子 $S_i$、$S_i$ 的前一个句子($S_{i-1}$)及 $S_i$ 的后一个句子($S_{i-1}$)共计三个连续的句子组成的句群.如果句子个数少于三个,直接取实际句子的数目,比如文档的第一句话,取连续的两个句子即可.$K(R^a)$指原子规则 $R^a$ 的关键字,Contain 是包含运算,$\mathrm{Contain}(S_{i-1}S_iS_{i+1}, K(R^a))$指 $S_iS_{i+1}S_{i+2}$是否包含 $K(R^a)$.

式(1)直接采用了句群中是否包含原子规则特征词的过滤方法,不需要分词,在实践中证明速度要比分词快很多.

## 4 实验结果及分析

### 4.1 实验环境及评测指标

目前,研究者多是自行收集整理语料用于主题信息采集的评测.比如,文献[9]围绕突发事件设置了 10 个查询项用于评测扩展查询的效果,文献[18]围绕运动设置了 14 个主题,从 Yahoo 上下载了 5000 个 Web 页面.本文围绕互联网公职人员信息监测的热点制定主题.截止实验分析时,内置搜索目标为 14 个,通用搜索引擎为 4 个.制定的主题规则共有 138 条,拆分后的原子规则 8223 条,这些原子规则中存在包含关系的有 4146 条,存在互换关系的有 717 条,存在相同关系的有 427 条.经过关系判定后,推送到内置搜索引擎队列的原子规则个数为 4009 条,减少了 4214 条,减少比例为 51%,对内置搜索目标而言,减少了约一半的访问工作量.推送到通用搜索引擎队列的原子规则个数为 8002 条,减少了 221 条,减少比例为 3%.对通用搜索引擎而言,只有原子规则具有了相同关系才能减少访问次数,所以采集次数减少并不明显.

对获取信息量的比较使用的评测指标:$P = \mathrm{Num}(t)$,其中,$t$ 指单位时间,$\mathrm{Num}(t)$指单位时间内采集到的信息数目.对获取相关信息量的比较使用的评测指标:$F = (2 \times P \times R)/(P + R)$.$P$ 是获取信息的准确率,$P = (\sum_{i=1}^{m} R(t_i)/N(t_i))/m$,其中,$R(t_i)$指在单位时间 $t_i$ 内采集到实际相关信息条数,$N(t_i)$指单位时间 $t_i$ 内返回相关信息条数,$m$ 指选定的时间范围.$R$ 是获取信息的召回率,$R = (\sum_{i=1}^{m} R(t_i)/\mathrm{Num}(t_i))/m$,其中 $\mathrm{Num}(t_i)$指单位时间 $t_i$ 内返回的所有信息条数.

### 4.2 结果分析

#### 4.2.1 采集信息量的比较

使用两种不同的采集方法,在不同的单位时间内统计采集的信息量.两种方法如下:(1)进行主题规则到原子规则的拆分,不考虑原子规则之间的关系,将所有的原子规则调度到内置搜索引擎和通用搜索引擎采集,该方法记作 $M_1$;(2)进行主题规则到原子规则的拆分,考虑原子规则之间的关系,依据原子规则之间的关系调度不同的原子规则调度到内置搜索引擎和通用搜索引擎采集,该方法记作 $M_2$.

为了统计获取的信息量,方法 $M_1$ 和 $M_2$ 都仅进行 URL 排重,不进行内容排重.不同的单位时间(h 表示小时)采集到的信息量见表 1 所示.

**表 1 不同的单位时间获取的信息量**

| 方法 | 1(h) | 4(h) | 8(h) | 16(h) |
|---|---|---|---|---|
| $M_1$ | 78302 | 96443 | 133722 | 133766 |
| $M_2$ | 89976 | 133688 | 133731 | 133799 |

从表 1 可见,对方法 $M_2$ 而言,4 个小时已经完成了大多数搜索引擎的信息采集,4 个小时后信息量变化已经不太明显.如果是热点主题,信息量的变化会更明显一些.而对于方法 $M_1$,8 个小时后,基本完成了一个轮次的信息采集,信息量已经和 $M_2$ 方法获取的信息量相当.可见,方法 $M_2$ 显著的提高了单位时间内采集的信息量,尤其是针对一些实时性要求较高的主题监测,方法 $M_2$ 能在较短的时候内获取大量信息.

#### 4.2.2 采集相关信息量的比较

对方法 $M_2$ 获取的信息,使用了四种主题规则与信息相关性的计算方法进行了实验比较.四种方法如下:(1)使用经典的余弦相似度计算方法,原子规则的关键字权重统一设置为 1,信息中的特征词权重设置为 TF(词频),具体的计算公式略,该方法记为 Base-1.通过实验确定相关性的阈值是 0.6;(2)使用整个文本中,包括标题、正文,是否包含原子规则中的所有特征词的方法.如果包含,就认为相关,否则认为不相关,该方法记为 Base-2.该方法类似于文献[8]介绍的 Fish-Search 算法;(3)文献[15]构建了新闻的表示模型 $d_i = \{T, K, D, F\}$.在计算检索项与信息的相关性时,认为检索项之间的距离与相关性成反比.由于本文所提方法收集的是新闻、帖子、微博等各类信息,因此直接计算主题规则与信息正文的相关性,并考虑了规则关键字之间的距离与相关性成反比,该方法记为 Base-3;(4)使用本文提出的基于句群的相关性计算方法,该方法记为 Sen-Cluster.通过实验确定的句群单位为邻近的 3 个句子.

为方便评测,以每日新获取的信息量为参考,方法 $M_2$ 平均每天获取约 220 条信息.使用 $F$ 指标连续评测了 10 天.具体的评测结果见表 2 所示.

**表 2　三种方法的 $F$ 指标结果比较**

| 方法 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 平均 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Base-1 | 0.53 | 0.51 | 0.53 | 0.49 | 0.51 | 0.5 | 0.56 | 0.54 | 0.53 | 0.54 | 0.524 |
| Base-2 | 0.71 | 0.74 | 0.64 | 0.64 | 0.61 | 0.6 | 0.61 | 0.73 | 0.73 | 0.75 | 0.676 |
| Base-3 | 0.75 | 0.76 | 0.68 | 0.65 | 0.63 | 0.63 | 0.66 | 0.74 | 0.73 | 0.76 | 0.699 |
| Sen-Cluster | 0.83 | 0.81 | 0.77 | 0.79 | 0.8 | 0.84 | 0.76 | 0.79 | 0.83 | 0.85 | 0.807 |

由表 2 可见,方法 Sen-Cluster 得到的 $F$ 值为 0.807,比方法 Base-1 高出 0.283 个点,比方法 Base-2 高出 0.131 个点,比方法 Base-3 高出 0.103 个点.在向量空间模型(VSM)的基础上,使用经典的余弦相似度计算方法效果最差(方法 Base-1),主要原因是:该方法是一种模糊匹配方法,容易被部分特征词干扰,适合于广泛主题的采集.方法 Base-3 使用了类似余弦相似度的计算方法,不过考虑了规则中多个关键字之间距离的约束,距离越近认为相似度越高,和方法 Base-1 相比,提高了 0.175 个点.Base-1 和 Base-3 都基于 VSM,此类方法需要对文本信息进行分词,统计词频,在时间的消耗上明显要比 Base-1 和 Sen-Cluster 方法大许多.

方法 Base-2 实现起来最为简单,但效果比 Base-1 要好,主要原因是:依据主题规则中关键字间的约束关系匹配文档,不是模糊匹配,这对具体主题采集是非常适用的,尤其是当前流行的社交媒体,文档信息经常篇幅较短,很难准确从中统计特征词的权重,这种简单匹配的方法将会有较大的应用空间.方法 Base-2 和 Base-3 相比,相差 0.023 个点,结果非常接近,但方法 Base-2 不需要分词、统计词频,效率要比 Base-3 高.

方法 Sen-Cluster 比 Base-2 好的原因主要是,该方法进一步的限定了主题规则中关键字在文档中出现的范围,不是以整篇文档为匹配单位,而是以 3 个连续的句子组成的句群为匹配单位,这样做可以有效的过滤掉不相关的信息.

## 5　总结及展望

本文针对多主题信息采集涉及的规则多、系统采集负担重、信息过滤算法不够精准等问题,提出了基于原子规则的信息调度、采集策略,取得的主要研究成果有:(1)给出了基于搜索策略的多主题信息采集的系统模型,讨论了该模型涉及的关键技术;(2)分析了主题采集规则、原子规则在内置搜索引擎、通用搜索引擎上搜索结果的差异,揭示了原子规则之间的相同、包含及互换三种关系;(3)给出了原子规则往内置搜索目标队列、通用搜索目标队列调度分配的算法;(4)提出了基于句群的主题与信息相关性的过滤方法.

在研究中,发现如下问题需要进一步探讨:(1)提出的方法更多的适合于具体主题信息的采集,该方法对广泛主题信息的采集有哪些是可以借鉴应用的;(2)重点解决了提高单位时间内主题信息采集的数量、主题相关信息采集的数量的问题,对信息的判重、排序等没有做研究,使用了已有的 URL 排重算法,基于内容的主题信息排重及主题信息的动态展示等也有很多方面值得进一步研究;(3)已有的语义资源如何应用到所提方法中,进一步优化提升主题采集的效果有待研究;(4)主题信息采集的可视化展示方面,应超越目前的纯文本信息的展示模式,探讨融合图像、视频等内容,给用户更加友好的使用体验,比如文献[25,26].

**致谢**　我们向对本文提出中肯修改建议的审稿人表示衷心地感谢.

### 参考文献

[1] Steve Lawrence, C Lee Giles. Accessibility of information on the Web[J]. Nature, 1999, 400(6740): 107 – 109.

[2] Steve Lawrence, C Lee Giles. Searching the world wide web [J]. Science, 1998, 280(5360): 98 – 100.

[3] Selberg E, Etzioni O. The Metacrawler architecture for resource aggregation on the Web[J]. IEEE Expert, 1997, 12(1): 11 – 14.

[4] Hsu C C, Wu F. Topic-specific crawling on the Web with the measurements of the relevancy context graph. Information System, 2006, 31(4 – 5): 232 – 246.

[5] Chakrabarti S, Berg M, et al. Focused crawling: a new approach to topic-specific Web resource discovery[J]. Computer Networks, 1999, 31(11 – 16): 1623 – 1640.

[6] Hersovici M, Jacovi M, et al. The shark-search algorithm and application: tailored web site mapping[J] Computer Networks and ISDN Systems, 1998, 30(1 – 7): 317 – 326.

[7] Bergmark D, Lagoze C, et al. Focused crawls, tunneling, and digital libraries[A]. 6th European Conference on Research and Advanced Technology for Digital Libraries[C]. London, UK: Springer-Verlag, 2002. 91 – 106.

[8] De Bra PME, Post RDJ. Information retrieval in the world wide web: making client-based searching feasible[A]. 1st International World Wide Web Conference[C]. Geneva, Switzerland:

Elsevier Science BV,1994.183 – 192.
[9] 仲兆满,朱平,等.一种基于局部分析面向事件的查询扩展方法[J].情报学报,2012,31(2):151 – 159.
Zhong Zhong-man, Zhu Ping, et al. Research on event-oriented query expansion based on local analysis[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(2):151 – 159.(in Chinese)
[10] Rodrigo Campos, Oscar Rojas, et al. Distributed ontology-driven focused crawling[A]. 21st Euromicro International Conference on Parallel, Distributed, and Network-Based Processing[C]. Belfast, United Kingdom: IEEE Computer Society, 2013. 108 – 115.
[11] Punam Bedi, Anjali Thukral, et al. Focused crawling of tagged web resources using ontology[J]. Computers and Electrical Engineering, 2013, 39(2):613 – 628.
[12] Yang S Y. Ontocrawler: a focused crawler with ontology-supported website models for information agents[J]. Expert System Application, 2010, 37(7):5381 – 5389.
[13] Michael Hersovici, Michal Jacovi, et al. The shark-search algorithm: an application: tailored web site mapping[J]. Computer Networks and ISDN System, 1998, 30(3):256 – 264.
[14] Martinet J, Chiaramella Y, et al. A relational vector space model using an advanced weighting scheme for image retrieval[J]. Information Processing and Management, 2011, 47 (3):391 – 414.
[15] 仲兆满,李存华,等.面向 Web 新闻的事件多要素检索方法.软件学报,2013,24(10):2366 – 2378.
Zhong Zhong-man, Li Cun-hua, et al. Web news oriented event multi-elements retrieval[J]. Journal of Software, 2013, 24(10):2366 – 2378.(in Chinese)
[16] Martinez-Romo J, Araujo L. Updating broken web links: an automatic recommendation system[J]. Information Processing and Management, 2012, 48 (2):183 – 203.
[17] Liu H Y, Milios E. Probabilistic models for focused Web crawling[J]. Computational Intelligence, 2012, 28(3):289 – 328.
[18] Du Y J, Pen Q Q, et al. A topic-specific crawling strategy based on semantics similarity[J]. Data & Knowledge Engineering, 2013, 88(11):75 – 93.
[19] Torkestani J A. An adaptive focused Web crawling algorithm based on learning automata[J]. Appliance Intelligence, 2012, 37(4):586 – 601.
[20] 高凯.搜索引擎中信息动态采集策略的研究[J].电子学报,2007,35(10):1984 – 1988.
Gao Kai. Dynamic refresh strategy for crawler in search engine[J]. Acta Electronica Sinica, 2007, 35(10):1984 – 1988.(in Chinese)
[21] Mohsen J, Hassan S, et al. A method for focused crawling using combination of link structure and content similarity[A]. 2006 IEEE/WIC/ACM International Conference on Web Intelligence[C]. Hong Kong: IEEE Computer Society, 2006. 753 – 756.
[22] Yuvarani M, Iyengar N, et al. Lscrawler: a framework for an enhanced focused Web crawler based on link semantics[A]. 2006 IEEE/WIC/ACM International Conference on Web Intelligence[C]. Hong Kong: IEEE Computer Society, 2006. 794 – 800.
[23] Melanie N, Markus N, et al. Focused crawling for building Web comment corpora[A]. 10th IEEE Consumer Communications and Networking Conference[C]. Las Vegas, NV: IEEE Computer Society, 2013. 685 – 688.
[24] Almpanidis G, Kotropoulos C, Pitas I. Combining text and link analysis for focused crawling—An application for vertical search engines[J]. Information Systems, 2007, 32(6):886 – 908.
[25] Wang Meng, Li Guang-da, et al. When Amazon meets Google: product visualization by exploring multiple information sources[J]. ACM Transactions on Internet Technology, 2013, 12(4):Article 12.
[26] Nie Li-qiang, Wang Meng, et al. Beyond text QA: multimedia answer generation by harvesting Web information[J]. IEEE Transactions on Multimedia, 2013, 15(2):426 – 441.

## 作者简介

**仲兆满** 男,1977 年 8 月出生于江苏省赣榆县.现为淮海工学院副教授.主要研究领域为信息检索,文本信息挖掘,事件本体等.以第一作者身份在国内外期刊发表研究论文 30 余篇,其中 SCI/EI 检索 15 篇,主持省、市级自然基金项目各 1 项.
E-mail: zhongzhaoman@163.com

**李存华(通信作者)** 男,1963 年 9 月出生于江苏省徐州市.2004 年毕业于东南大学计算机系.现为淮海工学院教授.主要研究领域为数据挖掘,人工智能,图像处理等.发表研究论文 100 余篇,主持国家、省、市级项目 9 项.
E-mail: cli@hhit.edu.cn

**刘宗田** 男,1946 年 6 月出生于江苏省临沂市.现为上海大学教授、博士生导师.主要研究领域为人工智能,软件工程等.发表研究论文 300 余篇,主持国家自然科学基金项目 4 项.
E-mail: ztliu@shu.edu.cn

**管 燕** 女,1976 年 4 月出生于安徽省当涂县.现为淮海工学院讲师.主要研究领域为人工智能,图像处理等.发表研究论文 30 余篇,参与省市级项目 6 项.
E-mail: gy764@sohu.com