文章编号:1000-5862(2013)02-0142-06

基于组合验证的 Web 页面抽取算法研究

耿 耘[1,2],蒋严冰[1*],郭 岩[2],刘 悦[2],余 钧[2],程学旗[2]
(1. 北京大学软件与微电子学院,北京 100190;2. 中国科学院计算技术研究所,北京 100101)

摘要:通过研究抽取算法的本质和抽取算法之间的关系,对抽取算法的互补性进行分析,提出了一种多算法组合验证机制,该机制能检测出抽取算法的错误,并通过结合动态阈值调整的方法,提高抽取算法的抽取准确率.

关键词:信息抽取;组合验证;阈值;多算法
中图分类号:TP 391 **文献标志码**:A

0 引言

自从 Web 中的信息所蕴含的价值被人们所发现以来,针对 Web 信息抽取的研究一直未停滞. 在这期间涌现出了许多针对 Web 信息抽取的优秀算法,这些算法各有各的优缺点,但它们实质上都是人们对页面理解的一种抽象化和公式化.

在论坛中的回复记录,由于大脑拥有自然语言理解的能力,可以通过文字内容对页面结构进行判断,因此可直观地判断出哪里是作者,哪里是日期等. 但计算机无法较好地理解自然语言,即假设页面的文本全部由乱码构成,人们对其的理解就会较为困难. 此时,人虽然读不懂其文本内容,但有经验的人仍能通过一系列经验规则综合地得出对页面结构的理解.

一个抽取算法便可以看作是一条经验规则,即人们对页面结构与内容之间关联的一种理解. 比如在针对如论坛、新闻评论等多记录类型页面的抽取算法中,有利用"重复结构"这一经验规则进行抽取的算法,这类算法中又有基于视觉信息的、基于 DOM 树的、基于 HTML 字符串的算法等,它们都符合在长期观察中所得到的页面排版结构规律. 又如,在针对如新闻正文等单记录的页面抽取算法中,有利用文本/链接比的算法、利用文本信息量的算法以及利用文本区域面积的算法等,这也符合"正文应当在最显眼最重要的地方"这一规律. 可见这些算法与人们对页面理解的经验规则是分不开的,而单独应用这些算法的效果就如同只通过单个经验规则来判断页面内容一样,在特定结构的页面上抽取效果较好,而一旦页面结构发生变更就会失效. 因此,本文认为只有将多个算法相结合,并根据这些算法的结果综合判断和推理得出页面结构的分析结果,才更符合人们对 Web 页面的理解方式,从而得到更好的结果. 基于以上想法,本文对抽取算法的关系进行了分析,并基于利用算法的结果进行组合验证的方式,提出了一个可支持任意符合标准的算法进行组合验证的框架,并给出了一个多算法组合验证进行抽取的实例及所得到的实验数据.

1 相关工作

在 Web 页面信息抽取算法中有比较主要的 2 大类,即针对多记录页面的抽取算法和针对单记录页面的抽取算法. 针对多记录页面的算法一般拥有一个前提假设,即页面中各个纪录的结构是相似的,即"重复模式"的概念. 基于这一概念而衍生出的算法都在寻求找到"页面中重复度最高的那一系列元素",并默认这些元素即为主体记录. 这类算法有基于字符串比较的 RoadRunner 算法[1,4],以及文献[3]中所介绍的基于 DOM 树重复结构的算法,还有如文献[2]中的基于视觉上重复的算法,这些算法应用十分广泛,且在一些比较规则的多记录页面上取得了不错的效果. 相比多记录页面的抽取算法,针对单记录页面的抽取算法则更多地从页面内容的特点上出发去寻找抽取特征,比如利用文本/链接比值进

收稿日期:2012-11-15
通信作者:蒋严冰(1975-),男,山东泰安人,副教授,博士,主要从事面向对象的方法与技术、建模语言和建模工具等方面的研究.

行正文抽取的算法,如文献[5]中所介绍的方法,以及利用文本信息量进行正文抽取的算法等. 这类算法和多记录抽取算法有一个共同的特点,就是它们都有一个前提假设,而一旦有些页面不符合这个特点,就会产生抽取失败的结果,并且这种失败常常无法被自动检测出来.

不少研究者在设计新算法和改进旧有算法时,已经开始尝试使用多种信息结合的思想来进行抽取,如文献[12]尝试将 DOM 树信息和视觉上的信息结合起来进行抽取. 但是,将多种算法结合起来进行抽取的研究并不多见. 而 Thomas Gottron 在文献[11]中,尝试将文本对齐信息和文本信息量结合起来进行正文内容抽取,之后他又在文献[4]中尝试将多种成熟的算法结合起来使用,并且较为系统地提出了几种通过多个算法结合进行抽取的框架,但他没有对算法的结合方式进行深入研究. 本文除了提出了一种更加合理的算法组合抽取机制,且对算法的本身特质进行了分析和研究.

2　多算法组合验证机制

2.1　抽取算法描述

为了更好地去理解和表达抽取算法,首先提出抽取流程中的 4 个要素: 抽取算法、抽取项、抽取对象和抽取结果.

表 1　抽取流程中的 4 个要素

要素名称	含义
抽取对象	一个由 HTML 编写的 Web 页面,有多种表现形式,如 DOM 树、文本串、视觉模型等
抽取项	人们所关心的抽取对象中所包含的内容,如“作者”、“正文”、“日期”、“产品价格”、“回复记录”等
抽取算法	对抽取对象按照抽取项进行选择得到抽取结果的一个规则
抽取结果	将抽取对象通过抽取算法规则进行选择所得到的结果

如一个新闻页面,整个 Web 页面就是一个抽取对象,其中可能包含有{正文,作者,标题,摘要,日期,广告}等抽取项,假设按照正则表达式匹配的抽取算法对“日期”项进行抽取,那么得到的正确抽取结果为“2012 年 07 月 06 日 01:49”.

2.2　抽取算法的组合验证

算法的组合验证,就是用多种算法对同一篇网页进行抽取得到许多结果,然后利用这些结果之间的异同对算法的正确性进行验证.

组合验证规则 1: 设 n 个对同一抽取项 k 进行抽取的抽取算法 $F_1, F_2, F_3, \cdots, F_n$ 组成的算法集合 G. 分别对同一个抽取对象进行抽取,得到了 n 个结果集合 $R_1, R_2, R_3, \cdots, R_n$. 将结果按照相同的为一类进行聚类从而获得 m 个类($m < n$) $C_1, C_2, C_3, \cdots, C_m$. 设结果数 $|C_x|$ 最多的类为正确的结果.

这是组合验证中最简单也是最基础的一个规则,它假设了这样一个前提“多数算法得到的结果便是正确的结果”,这也是进行算法组合验证的基础假设,符合人们本身对页面的理解过程,即多数线索所指向的目标最有可能是正确的目标. 利用规则 1 可能产生的结果有 3 种情况:

(Ⅰ)最完美的情况下所有算法得到的结果都相同,那么就认为该结果就是正确的结果.

(Ⅱ)少数算法返回的结果不同,但大多数算法返回的是同样的结果,就可以认定大多数算法所产生的是正确结果.

(Ⅲ)在最差的情况下,结果平均分为了数量相同的几个类. 此时,则只能随机选择一类作为正确结果.

在第(Ⅲ)种情况下多算法组合验证已经失去了它本来的意义. 即如果算法集合 $G(k)$ 选择不当,可能会导致大多数算法得到的结果并不是正确的结果. 规则 1 实际上就是求结果集合的交集.

考虑如下一种情况,假设由算法 A_1, A_2, A_3 所组成的算法集合 G 中,算法 A_1 和 A_2 分别是对 DOM 树重复结构和对 HTML 字符串重复结构进行判断的 2 个对多记录页面进行记录切分的算法,这 2 个算法在本质上是相同的,因而它们对任何页面进行抽取所得到的结果也是一样的. 此时对这 3 个算法的组合验证就失去了意义,因为验证的结果完全取决于 A_1 和 A_2,与 A_3 无关.

为了解决这个问题,首先提出算法同质的定义.

定义 1　抽取算法的同质关系: 设有抽取算法 A 和抽取算法 B,(Ⅰ)若算法 A 和算法 B 针对同一抽取项;(Ⅱ)对于同一个抽取集合,算法 A 和算法 B 抽取且获得的抽取结果 R_a, R_b,有 $R_a = R_b$.

若算法满足条件(Ⅰ)和条件(Ⅱ),则称算法 A 和算法 B 是同质的.

算法之间的同质可以通过推理证明来判定. 如对于前面的算法 A_1 和 A_2,可轻易用反证法证明其为同质的. 以此作为基础可对算法 1 进行扩展.

组合验证规则 1: 设有 n 个对同一抽取项 k 进行抽取的互不同质的抽取算法 $A_1, A_2, \cdots, A_n$ 组成的

算法集合 G. 分别对同一个抽取对象进行抽取,得到了 n 个结果集合 $R_1,R_2,\cdots,R_n$. 按结果相同的为一类进行聚类从而获得 m 个类($m<n$) $C_1,C_2,\cdots,C_m$. 设结果数$|C_x|$最多的那一类则为正确的结果.

这样,可以保证在不同质的算法之间进行组合验证,其结果必然是有意义的. 可以进一步对同质进行扩展.

定义 2 算法的互同与互异关系: 设同一抽取项的算法 A 和 B 对抽取对象集合 T 进行抽取并得到的结果集合为 R_a 和 R_b. 设集合 T 的元素个数为 n. 若 $\lim_{n\to\infty} Diff(R_a,R_b)=\infty$,此时称抽取算法 A 和 B 在抽取对象集合 T 上是互异的,否则称算法 A 和算法 B 是互同的.

这里可以使用求集合最小编辑距离的方法来衡量 R_a 和 R_b 的差距 $Diff(R_a,R_b)$,具体的算法有很多,这里不详述. 如果这一差距值随着抽取对象数量的增长而增长得较大,那么就说明这 2 个算法的"区别"比较大. 应当尽量选择区别比较大的算法作为算法集合中的元素,这样可以保证组合验证更具有意义.

刚才讨论的是对同一抽取项进行抽取的算法之间进行组合验证. 试考虑这样一种情况: 算法 A 是对一种对论坛页面的回复记录进行切分的算法,而算法 B 是一种对论坛页面回复作者进行抽取的算法,而已知在这种论坛页面中,每个回复记录必然会有一个作者. 那么假设算法 A 抽取出了 10 个记录,而算法 B 抽取出了 9 个作者,则可知必然有一个算法出现了错误,再者如果判定作者标签必定在视觉上包含在记录标签中,那么同样可以依次对其进行组合验证. 这 2 个算法虽然不是针对同一抽取项的,但仍可以进行组合验证,为此提出同构关系的概念.

定义 3 抽取算法的同构关系: 设算法 A 和算法 B. (Ⅰ)若算法 A 针对抽取项 A 进行抽取,算法 B 针对抽取项 B 进行抽取($A!=B$);(Ⅱ)使用算法 A 和算法 B 对同一批网页进行抽取得到结果集合 R_a 和 R_b. 如果 $\forall r_a\in R_a$,都存在一个一一映射 f,使得 $f(r_a)=r_b$,其中 $r_b\in R_b$.

若算法满足条件(Ⅰ)和条件(Ⅱ),称抽取算法 A 和抽取算法 B 为同构的.

之后可得到同构算法之间的组合验证规则 2.

组合验证规则 2: 设 n 个互相同构的算法 F_1,$F_2,F_3,\cdots,F_n$ 组成的算法集合 G,分别使用这些算法对同一抽取对象进行抽取,得到结果集合 R_1,$R_2,\cdots,R_n$,根据结果集合的个数$|R_x|$相等进行聚类而获得 m 个类 $C_1,C_2,\cdots,C_m$. 设结果数最多的一类为正确抽取结果.

易知互相同质的算法必然是互相同构的,且同构关系具有传递性. 只要是互相同构算法,都可以将其归并到算法结合中来进行组合验证,这样一来就大大增加了抽取算法集合的元素数.

2.3 组合验证的准确性问题

组合验证的最终目的是通过多种算法结合而提高算法的整体准确率,越多的算法会带来更多的验证因素,这会提高最终结果的准确度,但同时也会带来更多的验证失败. 设有算法集合 $G=\{F_1,F_2,F_3,\cdots,F_n\}$,假设其对于一个抽取对象集合 T 进行抽取,得到的正确的抽取结果为$\{R_1,R_2,R_3,\cdots,R_n\}$. 但算法 F_m 可能只能对 T 中的某一部分页面成功抽取,得到部分的正确结果集合 R_m,易知算法集合 G 所能得到的最大正确结果集合 $R_{\max}$为

$$R_{\max}=R_1\cup R_2\cup\cdots\cup R_n.$$

而对应地其最小正确结果集合 $R_{\min}$为

$$R_{\min}=R_1\cap R_2\cap\cdots\cap R_n.$$

易知算法集合 G 的抽取结果必然在 $[R_{\min},R_{\max}]$ 的范围之内,并且希望该抽取结果能尽可能地接近 $R_{\max}$. 假设使用某个抽取算法 F 对抽取对象集合 T 进行组合验证抽取后,结果中正确的算法数为 N 个,那么可以将抽取准确率 $Acu(F,T)$ 定义为

$$Acu(F,T)=N/|F|, \tag{1}$$

那么由抽取算法 $F_1,F_2,\cdots,F_n$ 所组成的算法集合 G,其平均抽取准确率为

$$Acu(G,T)=\sum_{i=1}^{n}Acu(F_i)\Big/n. \tag{2}$$

由(2)式可知准确率和算法集合 G 的大小是成反比的. 而对于一个抽取集合 T,设使用算法集合 G 抽取后得到的结果集 R 中正确页面数为 S,那么抽取成功率为

$$Suc(G,T)=S/|R|. \tag{3}$$

在实际应用中,当算法集合扩大后,由于多算法的加入,抽取结果的成功率 $Suc(G,T)$ 会提升,但 $Acu(G,T)$ 往往却会降低,导致抽取出的结果非常少. 这是因为在实际应用中,大部分页面往往只符合少数几个算法的前提假设,而大多数其他算法对这些页面的抽取效果都不好,因此如果算法结合中的算法数量越多,那么结果出现错误的算法也就越多,算法之间整体组合验证的最终结果便会变得很差.

这便使得多算法组合验证产生了“木桶效应”,使得抽取结果越来越接近 $R_{\min}$,这并不能提高抽取算法的效果. 为了解决这个问题,提出一种基于组合验证的阈值调整算法,来改善算法集合整体的平均准确度值.

2.4　基于组合验证的阈值调整算法

阈值调整的基本思路为: 在算法组合验证失败的时候,对算法进行一些调整,使之获得“更正确”的结果. 这更加符合人们对问题的理解,譬如人们在对 Web 页面的结果分析时总会带有一些主观的概念,而当这些主观的概念发生冲突时,就会对其中一些概念进行调整,从而使它们尽可能地相互符合,得到最终的结果. 比如认定“文本含量最多的标签中包含的是正文”,同时又认定“正文标签所占的面积应当最大”,但此时如果页面中广告图片占据了绝大部分面积,就会导致这 2 个概念产生冲突,则可以对第 2 条概念进行调整,变为“面积相对较大”的标签可能含有正文,尝试在规则间获得一个均衡点.

在当今的许多成熟抽取算法模型中都带有“阈值”的概念. 比如在页面切分算法的粒度问题上,以及在文本噪音过滤的界限问题上,都会需要通过使用一个人工配置的阈值. 这个阈值必须要人们手工去指定,且在算法执行过程中保持不变. 然而这个阈值并不应当是固定的某个值,而是允许其在一个连续的范围内变化. 在单算法情况下这种阈值的调整不容易实现,因为无法得知算法的当前阈值是否失效. 然而在多算法的组合验证过程中,完全可以通过组合验证的结果,对各个算法的阈值进行一定幅度的调整,使其更符合其他算法的“预期”. 组合验证算法的流程图如图 1 所示.

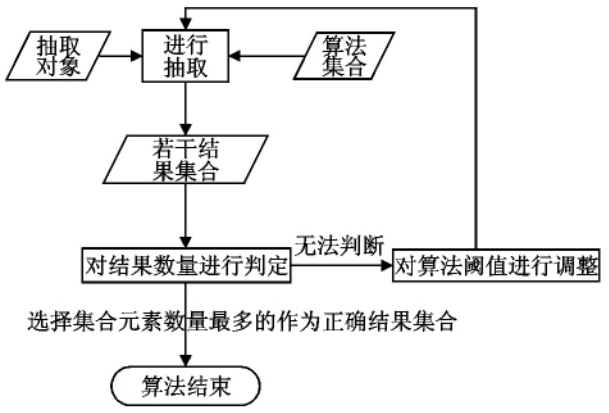

图 1　组合验证算法流程图

设有互相同构的带阈值的算法 A 和算法 B,他们在对同一页面进行抽取后发现结果不同,此时说明算法 A 和算法 B 中至少有一个算法出现错误了. 此时可以对算法 A 的阈值进行范围内的调整,看能否使其更符合算法 B 的“预期”. 如前面提到的文本/链接比抽取正文的算法 A,此时便可以降低文本/链接比的阈值,如果算法 B 也存在阈值,那么同样也可以调整算法 B 的阈值. 在大多数算法中都存在着一个人工设定的阈值,因此这在实际应用中的范围也非常广,许多算法的效果都受到阈值选择的影响. 将带有阈值并且带有阈值变化范围,以及阈值增量的页面抽取算法叫做可调整的阈值算法. 在实际应用中,可以对阈值算法的阈值变化范围,以及阈值增量进行人工定义,这一过程实现起来代价很小,且很简单. 或者也可以将该算法投入一个同类型的成熟算法集合 G 中并进行组合验证,使其自动训练得到一个平均最佳的阈值.

实验证明,通过调整算法的阈值,可以使得算法的综合抽取准确率达到更高的水平. 在实际应用中,只需要通过有限的几次的组合验证所得到的结果,即可让抽取器简单地学出一个规则模板,使得抽取剩下的大部分页面以及验证剩下的算法变得更容易,但这并不属于组合验证的范畴,本文不予讨论.

3　实验

3.1　实验数据

实验中选取了针对多记录型网页进行抽取的 4 个自动化算法组成算法集合,并对具有代表性的几个论坛中的几百个页面进行了抽取. 抽取目标是所有回复作者的作者名,这是进行一般舆情监测和进行人人关系分析的重要数据.

抽取前提已知主楼作者的作者名和所在标签,因为 Web 采集器在进行论坛板块页面采集时已经通过帖子入口页面处获得了帖子的作者名称. 另外,由于帖子的后几页实际上和第 1 页的结构是完全相同的,对的算法测试没有效果,因此只针对帖子的第一页进行实验.

3.2　实验算法选择

实验中的算法集合共由 4 个简单的算法组成,主要为了说明组合验证的机制和效果. 这几个算法分别基于 DOM 树、视觉信息、字符串等信息,通过不同手段对多记录网页的回复作者信息进行了抽取,这 4 个算法如表 2 所述.

表 2 算法结合介绍

算法名称	算法基本思路	是否带阈值
IRA-DOM	对 DOM 树的重复结构进行判定,假定所有记录都在同一层,且记录和记录之间的 DOM 树差异度最小,且记录的 DOM 树内容应当最多,从而对页面记录进行划分	否
IRA-VISUAL	利用视觉信息对页面标签的对齐,面积等进行判定,假定记录和记录之间在视觉上具有相似性,通过阈值设置切分的细度,从而对页面记录进行切分,细节可参考 [2]	是
IRA-CSS	利用标签的 CSS"内联样式"属性进行聚类,假定所有回复作者的内联样式都是同一个样式	否
IRA-LINK	假定所有作者的标签必为超链接类型标签,且从楼作者的超链接地址类似,对主楼作者链接的 href 属性进行解析,找到类似的超链接标签	否

3.3 实验结果

为了对比,首先列出单独使用这些算法对文章进行抽取所得到的结果的准确率,其结果见表 3.

为了让试验结果更具有说明性,将这 4 个算法通过几种不同的组合构成若干算法集合并用来进行抽取,并统计其准确率. 其中包括包含所有算法的算法集合 G_A:

G_1 = { IRA-DOM, IRA-VISUAL },

G_2 = { IRA-DOM, IRA-VISUAL, IRA-LINK },

G_3 = { IRA-DOM, IRA-CSS, IRA-LINK },

G_A = { IRA-DOM, IRA-VISUAL, IRA-CSS, IRA-LINK }.

除了 G_3 外,其它算法集合有可调整阈值的算法 IRA-VISUAL. 这些算法集合的组合验证抽取效果见表 4.

表 3 单算法实验数据结果

算法名	tieba. baidu. com	club. sohu. com	www. tianya. cn	ns2. awebring. com	Average
IRA-DOM	82%	42%	50%	87%	62.50%
IRA-VISUAL	82%	72%	51%	71%	69.50%
IRA-CSS	81%	–	–	–	20.25%
IRA-LINK	81%	100%	100%	–	70.25%

注: 每列中准确率最高的项用灰色底纹进行了标示.

表 4 算法集合实验数据结果

算法集合	tieba. baidu. com	club. sohu. com	www. tianya. cn	ns2. awebring. com	Average
G_1	52%	43%	54%	73%	55.50%
G_2	68%	59%	51%	73%	62.75%
G_3	81%	67%	51%	71%	67.50%
G_A	82%	82%	63%	73%	75.00%

3.4 实验结果分析

在单算法抽取时,由于 IRA-DOM 和 IRA-VISUAL 这 2 个算法都是基于页面重复结构的,由于它们默认页面会由多条记录组成,一旦遇到单纪录页面,譬如没有回复的帖子时就会出错,并且会抽取出边上的广告等信息认为是正确结果. 而 IRA-CSS 和 IRA-LINK 算法的假设性都偏强,比如百度贴吧中有部分匿名回复作者的作者名采用的并不是超链接标签,没有 href 属性且内联样式的值也不同. 此时的抽取结果中就会出现遗漏的现象,又如天涯论坛的网站,包含作者的标签中并没有 CSS 内联样式一项,因此 IRA-CSS 并不适用.

根据算法集合的试验数据,发现算法的选择与组合至关重要. 由于 G_3 组合中不包含可调整阈值的算法,因此其准确率并没有达到提升. 而其他包含了可调整的阈值的算法组合都取得了一定程度的准确度提升,且 G_A 组合中的算法在实验数据中达到了比较不错的效果.

4 总结与未来工作

本文提出了一种组合验证并调整阈值的抽取方式,对组合验证的规则进行了一些分析,并通过一个小例子来对组合验证的应用方法进行了阐明. 得出的结论就是只要按照一定规则谨慎的选择抽取算法,可以将许多算法结合起来进行抽取,并提高它们的平均抽取准确度. 由于获取成熟抽取算法的源代码比较困难,因此在进行实验时只按照成熟算法的思路编写了几个不同原理的算法,如果能选择质量更高的算法,便能够进一步提高算法抽取的准确度. 另外多算法会导致效率的降低,如何让多算法并行化的执行以提高效率也是要考虑的问题.

5 参考文献

[1] W3C. W3C document object model [EB/OL].

[2012-10-11]. http: // www. w3. org/DOM.

[2] Fabio Fumarola, Tim Weninger, Rick Barber, et al. HyLiEn: a hybrid approach to general list extraction on the Web [EB/OL]. [2012-09-16]. http: // www. cs. uiuc. edu/ ~ hanj/pdf/www11_ffumarola. pdf.

[3] Gupta S, Kaiser G, Stolfo S. Extracting context to improve accuracy for html content extraction [EB/OL]. [2012-09-16]. http: // wwwconference. org/www2005/cdrom/docs/p1114. pdf.

[4] Gottron T. Combining content extraction heuristics: the CombinE system [EB/OL]. [2012-09-22]. http: // citeseerx. ist. psu. edu/viewdoc/summary? doi = 10. 1. 1. 140. 3709.

[5] Sun Fei, Song Dandan, Liao Lejian. DOM based content extraction via text density [EB/OL]. [2012-09-22]. http: // disnet. cs. bit. edu. cn/DOM% 20Based% 20Content% 20Extraction% 20via% 20Text% 20Density. pdf.

[6] Liu Wei, Yan Hualiang, Xiao Jianguo, et al. Solution for automatic Web review extraction [J]. Journal of Software, 2010, 21(12): 3220-3236.

[7] Nitin Jindal, Liu Bing. A generalized tree matching algorithm considering nested lists for web data extraction [EB/OL]. [2012-09-26]. https: // www. siam. org/proceedings // datamining/2010/dm10_081_jindaln. pdf.

[8] Xia Yingju, Yu Hao, Zhang Shu. Automatic web data extraction using tree alignment [EB/OL]. [2012-09-27]. http: // dl. acm. org/citation. cfm? id = 1646194.

[9] Tim W, William H H, Han Jiawei. CETR-content extraction via tag ratios [EB/OL]. [2012-09-29]. http: // web. engr. illinois. edu/ ~ weninge1/pubs/WHH _ WWW10. pdf.

[10] Valter C, Giansalvatore M, Paolo M. Roadrunner: towards automatic data extraction from large web sites [EB/OL]. [2012-09-29]. http: // www. vldb. org/conf/2001/P109. pdf.

[11] Thomas Gottron. Content code blurring: a new approach to content extraction [EB/OL]. [2012-09-30]. http: // citeseerx. ist. psu. edu/viewdoc/download? doi = 10. 1. 1. 140. 5138&rep = rep1&type = pdf.

[12] Luo Ping, Fan Jian, Sam Liu, et al. Web article extraction for Web printing: a DOM + Visual based approach [EB/OL]. [2012-09-30]. http: // www. hpl. hp. com/techreports/2009/HPL-2009-185. html.

Research of Information Extraction Algorithm Based on Compositional Verification

GENG Yun[1,2], JIANG Yan-bing[1*], GUO Yan[2], LIU Yue[2], YU Jun[2], CHENG Xue-qi[2]

(1. School of Software and Microelectronics, Peking University, Beijing 100190, China;
2. Institute of Computing Technology, Chinese Academy of Science, Beijing 100101, China)

Abstract: The nature of universal web-information retrieval algorithm has been investigated, and a frame of cross-validation mechanism which could detect failure of the retrieval process has been proposed. After then, the performance by dynamically adjust threshold value of each algorithm has been improved.

Key words: information extraction; cross validation; threshold value; multi-algorithm

(责任编辑: 冉小晓)