文章编号：1004-4353(2015)01-0042-04

基于最大 Jaccard 相似度的互激励实体验证算法

刘宝超， 崔荣一*
（延边大学工学院 计算机科学与技术学科 智能信息处理研究室，吉林 延吉 133002）

摘要：针对基于规则的信息抽取技术提出了一种互激励实体验证算法.该算法兼顾了信息抽取过程中互激励算法的优点，并在此基础上引入了实体等待队列，用于存储未被成功验证的实体，并以最大 Jaccard 相似度为原则进行实体验证.实验结果表明，将该算法应用在基于规则的参考文献命名实体抽取中，其抽取的准确率要比 SermeX 系统高约 15%，比 Para Tools 系统高约 40%.

关键词：互激励；信息抽取；参考文献；实体验证
中图分类号：TP391.1 **文献标识码**：A

Mutual incentive entity verification algorithm based on the max Jaccard similarity

LIU Baochao， CUI Rongyi*
（*Intelligent Information Processing Lab.*, *Department of Computer Science & Technology*, *College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*）

Abstract: The technology of information extraction rules is proposed based on a mutual incentive entity authentication algorithm. The algorithm has both advantages of information extraction in the process of incentive algorithm, and on the basis of introducing the entity waiting queue, used to store has not been successfully verified entity, with the max Jaccard similarity principle of entity authentication. The experimental results show that, if the algorithm is applied in the reference named entity extraction, the extraction precision is higher than SermeX system about 15%, and is higher than Para Tools system about 40%.

Key words: mutual incentive; information extraction; reference; entity verification

基于规则的信息抽取是应用比较广泛的一种抽取方式，一般包括规则获取和规则应用两个过程，其中规则获取是该方式中最为关键的部分，只要能够获取规则，抽取工作就完成了一大部分，而且抽取效率极高[1-2].可是，就抽取的准确率而言，基于规则的抽取方式要明显低于基于 NLP(自然语言处理)和基于统计学习的方式，其原因在于该方式没有深入文本自身的含义，并且不考虑抽取结果的合理性，只要其符合抽取规则，就会被当作目标抽取出来[3].然而对于一些特殊领域的文本抽取，往往需要得到精确的抽取结果，因此鉴于上述情况，在抽取过程的最后阶段引入实体验证环节尤为重要[4].

科技论文中著录的文后参考文献属于半结构化的应用型文本，从众多样式的参考文献中自动提取出责任者、文献题名、出版地等信息是文献智

收稿日期：2014-12-07 **基金项目**：吉林省科技发展计划项目(20140101186JC)
* **通信作者**：崔荣一(1962—)，男，博士，教授，研究方向为模式识别、智能计算.

能分析的重要内容之一[5-6]. 采用基于规则的信息抽取方式不仅可以实现对参考文献中责任者、文献题名、期刊名、会议名、卷期、时间等命名实体的抽取，同时该方法操作简单，而且抽取的准确率较高，是目前大部分信息抽取系统使用的主流抽取方式. 例如CiteSeer系统就是采用启发式规则实现参考文献命名实体抽取的，并且该系统还能提供某一具体文献的“引用”和“被引用”情况以及文献的引用次数等信息[7]. 然而，这种仅按照启发式规则抽取的方式，其准确性仍依赖于被抽取文本自身的准确性和完整性[8]. 为了摆脱这种依赖和提高抽取准确率，本文在该抽取方法的最后阶段加入了实体验证环节，同时将改进的互激励算法(mutual bootstrapping)应用到该环节中，以进一步提高命名实体的抽取准确率.

1　互激励算法与原激励算法

互激励法无须指出所有实例与目标领域的相关性，但要给出一定量的种子词(关键词)进行整个过程的初始化. 初始化首先由种子词获取一定量的规则模式，由于规则具有一定的普遍性，因此规则中所隐含的种子词一般要多于原来的种子词，只要充分利用信息抽取的规则模式，即可得到更多的种子词. 在整个过程中，种子词推动规则模式的产生，而规则模式反过来又推动种子词的获取，形成相互激励的过程；反复该过程，直到没有新的符合要求的种子词或规则模式产生为止，即互激励过程结束[9-10].

当规则模式数量逐渐增多时，互激励法有可能将一定量的非种子词加入到种子词集中，使得算法效率和准确率降低，因此对新加入的种子词需要进行严格控制[11-12]. 多层激励法(multilevel level bootstrapping)又称为原激励法(meta bootstrapping)，它在互激励法的基础上对种子词进行评分，通过分数值控制其是否能够加入到种子词集中，从而保证所选种子词的合法性，提高算法的效率. 图 1 为互激励法与原激励法的关系，图 2 为改进的互激励法.

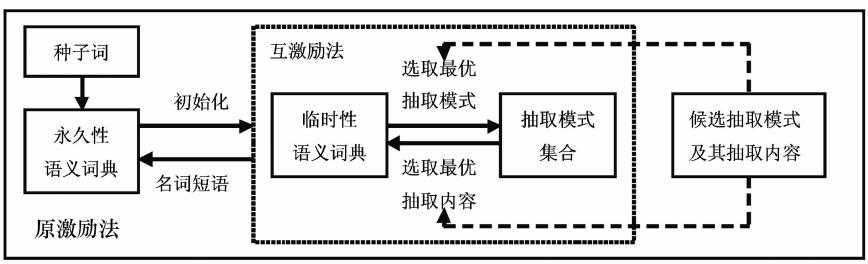

图 1　互激励法与原激励法的关系

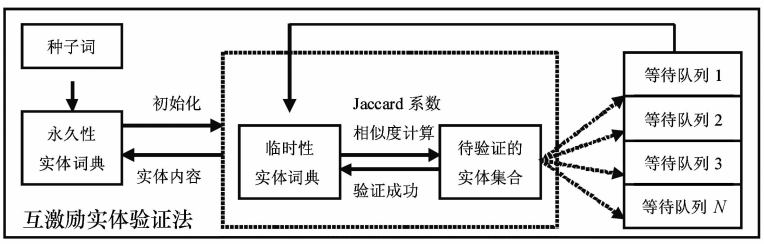

图 2　改进的互激励法

2　互激励实体验证算法

在基于规则的参考文献命名实体抽取中，对文献题名和期刊名的区分是最困难的，这是因为文献题名和期刊名文本长度类似又无标点符号，所以确定起来相对较难；而责任者存在“逗号和文本长度特征”，时间存在“数字特征”，期卷存在“括号”特征，出版地或学院存在专有名词，所以这些部分相对容易确定[13]. 对此，本文对原有的互激励验证算法进行改进，以解决文献题名和类型实体验证问题(|·|表示集合的元素个数). 具体算

法描述如下：

Step 1 建立初始期刊名关键词库，

$$Dictionary=\{w_1,w_2,\cdots,w_{n-1},w_n\},\quad(1)$$

其中 $w_i(i=1,2,\cdots,n)$ 为种子词，$n=|Dictionary|$ 为种子个数.

Step 2 将种子词拆分得

$$w_i=\{w_{i1},w_{i2},\cdots,w_{im_i}\}\ (i=1,2,\cdots,n),\quad(2)$$

其中 $m_i=|w_i|$ 为种子词 w_i 的长度.

Step 3 将待验证的文献题名 R_name 和期刊名 R_type 按照 Step 2 的方法进行拆分得：

$$R\_name=\{u_1,u_2,\cdots,u_n\},\quad(3)$$

$$R\_type=\{v_1,v_2,\cdots,v_t\},\quad(4)$$

其中 $n=|R\_name|$，$t=|R\_type|$.

Step 4 按(5)式定义分别计算 R_name 和 R_type 与 $Dictionary$ 的 Jaccrad 系数 $J(R\_name,Dictionary)$ 和 $J(R\_type,Dictionary)$，

$$J(A,B)=|A\cap B|/|A\cup B|,\quad(5)$$

其中 A 和 B 为两个集合.

Step 5 按下式计算文献题名相似度 S_{R_name} 和期刊名相似度 S_{R_type}：

$$S_{R\_name}=\max(J(R\_name,w_k),0\leqslant k\leqslant n);\quad(6)$$

$$S_{R\_type}=\max(J(R\_type,w_k),0\leqslant k\leqslant n).\quad(7)$$

Step 6 if ($S_{R\_name}>S_{R\_type}$)，判定文献题名与期刊名顺序书写颠倒，调整抽取内容，并将 R_name 加入到 $Dictionary$ 中，R_type 加入到文献题名数据库中.

Step 7 if ($S_{R\_name}<S_{R\_type}$)，判定文献题名与期刊名顺序书写正确，将 R_name 加入到文献题名数据库中，R_type 加入到 $Dictionary$ 中.

Step 8 if ($S_{R\_name}=\!=S_{R\_type}$)，无法确定抽取的内容是否准确，将 R_name 放入文献题名临时等待队列，R_type 放入期刊名临时等待队列.

Step 9 若文献未全部验证结束，返回 Step 3，否则如果验证队列空则转 Step 13.

Step 10 验证等待队列中的文献题名 R_name 和期刊名 R_type，返回 Step 6.

Step 11 在验证等待队列中的实体若出现 if ($S_{R\_name}=\!=S_{R\_type}$)，这时不再放入等待队列，而是利用文献题名数据库按(8)式和(9)式计算相似度，并记录循环次数 flag.

$$S_{R\_name}=\max(J(R\_name,w\_n_k),0\leqslant k\leqslant n);\quad(8)$$

$$S_{R\_type}=\max(J(R\_type,w\_n_k),0\leqslant k\leqslant n).\quad(9)$$

式中 w_n_k 是按照 Step 2 中的方法对文献题名数据库中已验证成功的文献题名进行拆分的结果.

Step 12 通过文献题名数据库验证时也出现 $S_{R\_name}=\!=S_{R\_type}$，则说明文献题名和期刊名相同.

Step 13 结束.

因为文献类型相对好确定，种子词获取相对容易，所以本文在 Step 1 中建立了一个初始文献题名关键词库，种子词 $w_i(i=1,2,\cdots,n)$ 的选取是通过统计足够多的学位论文文后参考文献之后确定的出现最多的前 n 个关键词. Step 2 中中文种子词拆分单位为汉字，而英文种子词拆分单位为空格.

在 Step 8 中导致无法确定的原因可能有两点：

1) $Dictionary$ 中的词数量过少，导致计算后 $S_{R\_name}=\!=S_{R\_type}=\!=0$；

2) R_name 和 R_type 本身很相似，如 R_name=“中文信息学报发展综述”，R_type=“中文信息学报”，使得 $S_{R\_name}=\!=S_{R\_type}\neq 0$.

3 实验结果及分析

本文通过准确率 P、召回率 R 和 F(measure)值这 3 个常用指标对实验结果进行评价，这样可以较好地与 SemreX 和 Para Tools 系统所得结果进行比较，其计算公式如下：

$$P=A/(A+C),\quad(10)$$

$$R=A/(A+B),\quad(11)$$

$$F=(\alpha^2+1)P*R/(\alpha^2P+R),\quad(12)$$

其中：A 表示提取的样本中抽取正确的文献数；B 表示未能正确提取的文献数；C 表示提取的样本中抽取错误的文献数；F 为综合评价指标；α 越大，R 对 F 的影响越大，本文中取 $\alpha=1$.

实验数据由某高校计算机类硕士学位论文中著录的文后参考文献构成，共计 741 条，其中中文参考文献 582 条，英文参考文献 159 条. 对每一条参考文献，根据参考文献信息抽取规则进行命名

实体抽取，并计算出 P，R，F 值，然后与 SemreX 系统和 Para Tools 系统进行比较，对比结果见表 1. 由表 1 可以看出，本文抽取的各项指标的平均值要高于 SemreX 系统约 15%，高于 Para Tools 系统约 40%. 另外，SemreX 系统和 Para Tools 系统中所用的抽取方法只适用于英文文献的抽取，而本文提出的方法可以适用于中 / 英文文献的抽取，扩大了抽取范围.

4　结论

本文针对基于规则的信息抽取技术提出了一种互激励实体验证算法，并将其应用在参考文献命名实体抽取中，实验结果表明：① 在信息抽取中引入实体验证环节，能有效减少对抽取文本自身含义准确性的依赖；② 在实体验证环节，将规则学习阶段的互激励法进行了改进，引入了实体等待队列，使得最终抽取的结果其 P，R，F 值 3 项指标远高于 SemreX 和 Para Tools 系统. 由于本算法没有对 *Dictionary* 进行优化，存在运行时间较长的不足，因此本文将在今后的研究中运用组合数学原理和遗传算法等对 *Dictionary* 进行优化，以提高抽取效率.

表 1　实验结果对比

项目	本文方法			SemreX			Para Tools		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
责任者	93.1	97.0	93.5	92.6	85.3	88.8	41.4	30.5	35.0
文献题名	96.3	98.0	97.1	89.1	78.3	83.4	42.3	23.8	30.5
期刊名	96.3	98.0	97.1	84.5	64.2	73.0	—	—	—
日期	99.2	99.8	99.5	84.4	79.1	81.7	78.2	43.1	55.6
卷号	99.2	99.8	99.5	74.7	61.6	67.5	73.4	61.5	66.9
起止页	99.2	99.8	99.5	72.3	65.9	69.0	62.3	30.4	40.9
网址	93.4	94.7	94.0	87.2	70.1	77.7	88.2	73.9	80.4
专著题名	91.3	92.0	91.6	73.1	50.9	60.0	—	—	—
版本项	91.0	91.8	91.4	—	—	—	—	—	—
出版社	91.2	91.9	91.6	74.3	61.6	67.3	44.2	32.8	37.7
院校	98.7	99.1	98.9	69.7	51.2	59.0	—	—	—
会议名	96.5	97.4	96.7	77.5	64.9	70.6	—	—	—
出版地	92.5	93.2	92.8	—	—	—	—	—	—
平均值	95.2	96.3	95.7	78.3	62.0	68.9	58.2	39.3	46.5

参考文献：

[1] 李洪亮，黄莉. 基于规则的百科人物属性抽取算法的研究[D]. 成都：西南交通大学，2013：11-25.

[2] 李湘东，霍亚勇，黄莉. 图书网页的自动识别及书目信息抽取研究[J]. 现代图书情报技术，2014(4)：71-74.

[3] 郭志鑫，金海，陈汉华. SemreX 中基于语义的文档参考文献元数据信息抽取[J]. 计算机研究与发展，2006，43(8)：1368-1374.

[4] Cheng Xianyi，Zhu Qian，Wang Jin. The Principle and Application of Chinese Information Extraction [M]. Beijing：Science Press，2010：181-182.

[5] 孙明，陆春生，徐秀星，等. 一种基于 SVM 和 AdaBoost 的 Web 实体信息抽取方法[J]. 计算机应用与软件，2013，30(4)：101-106.

[6] 张秀秀，马建霞. PDF 科技论文语义元数据的自动抽取研究[J]. 现代图书情报技术，2009(2)：102-106.

[7] Li Chaoguang，Zhang Ming，Deng Zhihong，et al. Automatic metadata extraction for scientific documents[J]. Computer Engineering and Applications，2002，21(10)：189-191.

[8] Liu Wei，Yan Hualiang. A unified and automatic web news object extraction approach[J]. Computer Engineering，2012，38(11)：167-169.

[9] Zhang M，Yin P，Deng Z H，et al. SVM＋Bi-HMM：a hybrid statistic model for metadata extraction[J]. Journal of Software，2008，19(2)：358-368.

[10] Wang Shuang. Research of web information extraction technology oriented to digital tourism website[D]. Xi'an：Xidian University，2012.

[11] 龚立群，马宝英，常晓荣. 科技文献元数据自动抽取研究综述[J]. 计算机系统应用，2013，22(3)：11-15.

[12] 杨春磊，邵堃基. 基于模式匹配的结构化信息抽取研究[D]. 合肥：合肥工业大学，2013：11-30.

[13] 陈先军. 文后参考文献引著质量及其审查方法[J]. 中国科技期刊研究，2014，25(9)：1145-1148.