基于本体相似度的中文科研论文信息抽取

徐 慧,杨学兵
(安徽工业大学 计算机学院,安徽 马鞍山 243002)

摘 要:随着大量的科研论文出现在互联网上,从中精确地抽取论文头部信息和引文信息显得十分重要。提出了基于本体相似度的信息抽取方法,该方法的关键在于用本体相似度判定某个行本体是正例还是反例,然后通过主动学习选择最有可能包含抽取信息的行本体集,再充分利用本体的语义推理能力找到正确的片断。从论文中提取头部信息和引文信息为进一步的语义检索和语义存储奠定基础。测试数据集的实验结果显示该方法比其他方法具有较高的准确率。

关键词:信息抽取;本体相似度;语义推理;主动学习
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2008)12-0203-04

Information Extraction from Chinese Research Papers Based on Ontology Similarity

XU Hui, YANG Xue-bing
(School of Computer, Anhui University of Technology, Maanshan 243002, China)

Abstract: Information extraction from Chinese research papers based on ontology similarity abstract as many research papers appear on the Internet, it becomes more and more important to extract paper header information and citations accurately from these papers. Presents a new information extraction algorithm which is based on ontology similarity. The key point of the algorithm is to divide the row-ontology samples into positive and negative instances, extract the most appropriate set of row-ontologes by active learning, and then retrieve the correct pieces lie in them by using the reasoning mechanism contained in the ontologies. It can get header information and citation from these papers, which assist the semantic searching and storage. Test results show that the algorithm is more precise than other approaches.

Key words: information extraction; ontology similarity; semantic reasoning; active learning

0 引 言

随着科学研究的快速发展和信息更新速度的加快,全世界发表的科研论文越来越多,这使得人们对科研论文的检索、抽取技术要求越来越高。获取科研论文的元数据信息,不仅可以有效地组织和管理这些论文,提高用户检索论文的效率,而且还能够进行大量的统计工作。如对期刊、科研单位、某篇论文或某个学者进行引用分析以及发现研究热点和研究趋势等。所以,从科研论文中自动抽取头部信息和引文信息有着重要的研究价值。

国内外对于论文的元数据信息抽取展开了相应的研究。文献[1]利用正则表达式规则提取出论文头部信息和引文信息。文献[2]采用同正则表达式构造的文本信息项特征模式对已经定位的信息块进行抽取。文献[3]利用领域 Ontology 里的概念关系产生的规则对文档进行标注与抽取。文献[4]结合本体技术,采用模式匹配方式从文档中抽取引文元数据信息;然而模式匹配具有局限性,对没有考虑到的引文模式,抽取精度低。文献[1]对引文是整体抽取,文献[4]只是单独对引文抽取,而没有对论文的头部进行抽取。文献[1]和文献[2]都用到正则表达式进行元数据抽取,但正则表达式不能像本体一样利用语义关系和语义推理,不能表示概念之间的联系。文中提出基于本体相似度的中文科研论文头部信息和引文信息抽取方法,充分利用本体关系及推理能力,并通过主动学习选择最有价值的训练文本,实验结果表明该方法明显优于上面几种方法。

1 本体描述

在语义 Web 中,本体是概念和关系的集合。本体描述语言能够描述复杂的关系并具有简单的推理能力。这里采用 W3C 推荐的 OWL 本体描述语言,定义

收稿日期:2008-04-23
基金项目:安徽省自然科学基金重点资助项目(2004KJ053ZD)
作者简介:徐 慧(1982-),女,安徽巢湖人,硕士研究生,研究方向为本体与信息抽取;杨学兵,教授,研究方向为数据挖掘。

了文档元数据本体(Document Metadata Ontology,简称DMO)。图 1 显示了文档元数据本体的框图。通过本体定义,可以规范概念内涵及其之间的关系,有利于机器处理。

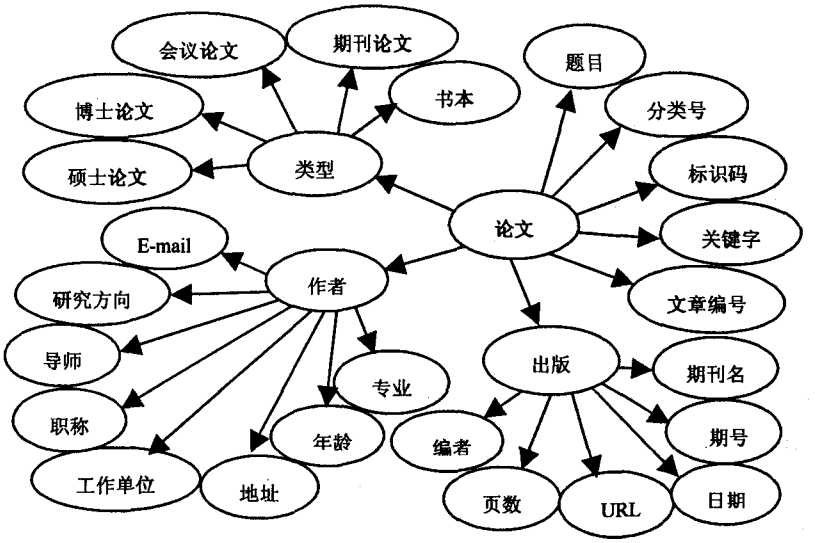

图 1 DMO 框架图

2 基于本体的信息提取流程

信息抽取是自然语言处理(Natural Language Process,NLP)和人工智能(Artificial Intelligent,AI)结合的结果。信息抽取系统主要从文本中抽取出特定的信息。通常被抽取出来的信息以结构化的形式存在,可以直接存入数据库中,供用户查询和进一步分析利用。信息提取的流程图如图 2 所示,对来自互联网上的各种无结构或半结构化的科研论文进行文档结构分析,然后利用本体库中已经建立的本体概念、关系和实例等信息,计算各概念之间相关性。为进一步提高准确率,建立了相应的人名数据库和常见期刊名称库。抽取出的元数据信息通过本体描述语言 OWL 进行描述和形式化,存储到知识库中,供下一步的基于语义的检索使用。

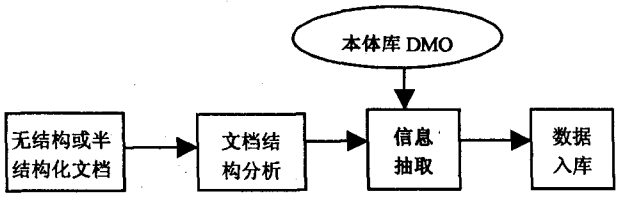

图 2 信息提取流程图

3 基于本体相似度的信息抽取

科研论文通常是自由格式的文本,从其中抽取元数据有一定的难度。但是根据观察和分析,它们还是有一定的固定结构。比如说,大多数论文由标题、作者、作者相关描述信息、摘要、关键字、主体和参考文献这七部分组成。当然一篇科研论文还有其他方面的属性,比如邮政编码仅包含六位数字,姓名包含 2~4 个汉字等。

现有的信息抽取算法基本上采用 NLP 工具,文中采用 Gate 对文档进行的标注和学习。为了深入挖掘标注后文档中的关系,提高信息抽取的效率,提出了用语义 Web 中的本体对自然语言处理后的文档进行描述,通过主动学习选择最有价值的训练文本,并将学习结果应用到未知文档中,此时可以认为主动学习工具为支持向量机(Support Vector Machine,SVM)。在形成对文本的本体描述的过程中,所有的描述本体都继承自定义了类、属性和关系的本体模板,这些描述本体之间仅仅在类的实例以及实例之间的关系上有区别,因此,可以采用文献[5]中的领域本体。另外若干本体合并后,其关系比原来的本体更加丰富。

3.1 行本体相似度计算

在完成对标注后信息的本体描述后,需要计算 2 个本体的相似度来判定某个行本体是正例还是反例。目前已有的本体相似算法有两种:一是利用同义词词典计算相似度;二是基于大规模的本体库统计信息,用词语的相关性来计算相似度[6],文中采用后一种方法。

设:A,B 是两个行本体,将 A,B 分词,得分词向量:$A=\{a_1,a_2,\cdots,a_k\}$。$B=\{b_1,b_2,\cdots,b_m\}$

定义(行本体相似向量):将行本体 A,B 的分词向量作并运算,得行本体的相似向量 $\boldsymbol{C}$,也就是 $\boldsymbol{C}=A\cup B=\{c_1,c_2,\cdots,c_n\}$,其中 $c_i\neq c_j,\in A$ 或者 $B,n\leqslant k+m$。则行本体 A,B 的相似向量 $\boldsymbol{A}^*=\boldsymbol{C},\boldsymbol{B}^*=\boldsymbol{C}$。

设 $Ta=\{t_1,t_2,\cdots,t_n\}$ 为 $\boldsymbol{A}^*$ 的词权重向量:集合 G 为 DMO 同义词集

词权重向量 TA^* 的计算方法如下:

(1) 如果词 $c_i\notin A$,若 c_i 在 DMO 中找到,并且存在 c_j 属于 A,使得 c_i 和 c_j 在 DMO 语义树中四层内有共同祖先,则 $t_i=0.8$,否则 $t_i=0$;

(2) 如果 $c_i\notin A$,且 c_i 在 DMO 中找不到,则 $t_i=0.5$;

(3) 如果词 $c_i\in A$,且 c_ic_{i+1} 没有同时出现在 A 和 B 中,则 $t_i=1$;

(4) 如果 $c_ic_{i+1}\in A$,且 $c_ic_{i+1}\in B$,若 $c_ic_{i+1}c_{i+2}$ 没有同时出现在 A 和 B 中,则 $t_i=2$;

(5) 如果 $c_ic_{i+1}c_{i+2}\in A$,且 $c_ic_{i+1}c_{i+2}\in B$,则 $c_i=$

3。

得到 Ta 和 Tb 的值,计算 A 和 B 的句子相似度,利用夹角余弦计算,公式如下:

$$\mathrm{Sin}(A,B) = \mathrm{Cos}(Ta,Tb) = \frac{\sum_{i=1}^{n} Ta_i * Tb_i}{\sqrt{\sum_{i=1}^{n} Ta_i^2} * \sqrt{\sum_{i=1}^{n} Tb_i^2}} \quad (1)$$

3.2 主动学习的基本思想

在文本信息抽取中,主动学习用于选择最有可能包含抽取信息的训练文本。对于训练集 D,给定一个未标注样本(输入)x,定义学习器对 x 的标注结果(输出)为 $\hat{y}(x)$,而期望标注结果为 $y(x)$,x 和 y 的未知联合分布为 $P(x,y)$,x 的已知边缘分布为 $P(x)$,则在训练集 D 上学习器标注结果和期望标注结果的方差为[7]:

$$\sigma_y^2 = E_D[(\hat{y}(x) - y(x))^2/x] \quad (2)$$

$$\sigma_y^2 = E_D[(y(x) - E[y/x])^2] + (E_D[\hat{y}(x)] - E[y/x])^2 + E_D[(\hat{y}(x) - E_D[\hat{y}(x)])^2] \quad (3)$$

学习的期望错误率表示为:

$$IV = \int_x \sigma_y^2 P(x)\mathrm{d}x \quad (4)$$

式(4)表示了在输入为 x 的情况下,学习器估计的不确定性。主动学习的目标是选择一个新的例子 $\tilde{x}$,使得当由此产生的实例结果 $(\tilde{x},\tilde{y})$ 被加入训练集时,方差 σ_y^2 的积分 IV 最小,即学习的期望错误率最小。此时,方差 σ_y^2 发生改变,用 $\tilde{\sigma}_y^2$ 表示新的方差:

$$\tilde{\sigma}_y^2 = E_{D\cup(\tilde{x},\tilde{y})}[\alpha_y^2/\tilde{x}] \quad (5)$$

如何选取 $\tilde{x}$ 使 IV 最小化,需要计算 $\tilde{\sigma}_y^2$。如果已知 $P(\tilde{y}/\tilde{x})$ 就可以估算 $\tilde{\sigma}_y^2$。

3.3 算法基本过程描述

本体关系推理算法分为学习过程和应用过程,学习过程描述如下:

1)采用 NLP 工具 Gate 中的 ANNIE 对训练文档进行分词以及命名实体识别(主要包括标题、作者、关键字、期刊名和日期等)。

2)为训练文档每行构建一个本体,描述行中词的顺序关系以及词的属性,称为行本体。

3)把训练文档中的所有包含待抽取信息的行本体合并成一个独立的特征本体。

4)构建一个支持向量机,计算特征本体和每个行本体的相似度。在整个训练集上训练这个支持向量机。

通过训练,一个行特征本体被构建出来,这个行特征本体的任务是从行本体列表中找到正确的行本体。对任何一个未知文档,应用的过程描述如下:

(1)把文档拆分成行本体的集合。

(2)通过主动学习过程构建的支持向量机和行特征本体找到最有可能包含抽取信息的行本体的集合。

(3)对每个候选的行本体,通过计算它与训练过程中每个正例行本体的距离,找到 k 个最接近的正例行本体。这 k 个正例行本体被合并到一个称为目标特征本体的本体中(这里参考了 k - 近邻算法[8])。目标特征本体的作用是从候选本体中把正确的文字片断区分出来。

(4)构建一个支持向量机,用目标特征本体训练这个支持向量机。

(5)应用这个支持向量机,把信息从待处理的行本体中抽取出来。

从上面的描述可以看出,算法的关键部分是本体推理能力以及支持向量机的应用。这个算法也是第一个通过两个步骤来完成信息抽取的算法。步骤(1)先找到正确的行,步骤(2)从正确的行中找到正确的片断。这主要是为了充分地利用本体对模糊信息的描述能力以及本体的推理能力,对信息内部关系的扩展。

4 性能测试

消息理解会议(Message Understanding Conference,MUC)为信息检索和信息提取领域内的算法性能测试制定了一系列的评估参数:设总共需要提取的信息数目为 N,提取正确的信息数目为 Nco,提取错误的信息数目为 Nin,那么信息查全率 Re 和提取精确度 Pr 为:

$$Pr = Nco/Nco + Nin, Re = Nco/N$$

实际评估时,应同时考虑 Pr 和 Re,但同时比较两个数值,很难做到一目了然。所以通常采用综合两个值(综合指标 F 值)进行评价,其计算公式[9]如下:

$$F = \frac{(\beta^2+1)PrRe}{\beta^2 Pr + Re}$$

式中 β 决定了查全率与精确度的比值,通常设定为 1、2 或 1/2。文中 β 取值为 1,即对二者一样重视。

从中国期刊全文数据库和引文数据库中检索计算机有关方面的论文整理标注而成。其中,科研论文头部信息数据 600 篇,引文数据 1500 条。采用 GATE - SVM 中提到的方式对数据集进行测试,用 Gate 中的 ANNIE 对文本进行标注,用 Wordnet 来计算 2 个词之间的相似度。在 SVM 的实现上,采用了 LibSVM。

对于论文头部信息数据集,把 600 篇中的 400 篇作为训练语料,剩下的 200 篇作为测试语料。这里把基于行本体相似度的中文论文信息抽取算法和文献[2]所提出的算法都对上面的语料进行语义项抽取,具

体情况见表 1。

对于中文论文引文信息数据集，把 1500 篇中的 1000 篇作为训练语料，剩下的 500 篇作为测试语料。同样在这里把文中提出的算法和文献[4]所提出的算法都对上面的语料进行语义项抽取，具体情况见表 2。

表 1 论文头部信息抽取结果

抽取域	文中结果			文献[2]结果		
	P	*R*	*F*	*P*	*R*	*F*
标题	0.972	0.984	0.957	0.914	0.846	0.878
作者	0.986	0.947	0.963	0.921	0.823	0.869
单位	0.978	0.915	0.945	0.917	0.820	0.871
地址	0.975	0.935	0.955	0.918	0.814	0.865
邮编	0.998	0.982	0.990	0.958	0.914	0.936
摘要	0.989	1.000	0.995	0.973	0.978	0.975
关键字	0.979	0.939	0.957	0.922	0.901	0.912
中图分类号	0.987	0.968	0.981	0.963	0.956	0.959
文献标识码	0.998	0.997	0.998	0.978	0.981	0.980

表 2 中文论文引文信息抽取结果

抽取域	文中结果			文献[4]结果		
	P	*R*	*F*	*P*	*R*	*F*
作者	0.988	0.973	0.980	0.904	0.842	0.872
标题	0.986	0.945	0.966	0.891	0.816	0.842
期刊名	0.998	0.982	0.990	0.958	0.914	0.936
刊号	0.954	0.955	0.955	0.862	0.860	0.862
日期	0.998	0.982	0.990	0.958	0.914	0.936
页码	0.998	0.997	0.998	0.978	0.981	0.980

从表 1 和表 2 可以看出所提出的信息抽取算法优于文献[2]和文献[4]所提出的算法。用本体对文本内容进行描述和用行本体相似度计算本体相似度，可以增强信息抽取的效率。

5 结束语

文中针对不能充分利用本体语义关系这一问题，提出一种基于本体相似度的中文论文头部信息和引文信息抽取方法。系统还处于初级阶段，本体库不够建全；在以后的工作中将完善本体库，改进本体相似度算法，比如在计算 2 个实例集合的相似度时，考虑更为合理的算法，以及对各种测试数据的适应性，都需要进一步研究和探讨。

参考文献：

[1] 李朝光，张 铭，邓志鸿，等. 论文元数据信息的自动抽取[J]. 计算机工程与应用，2002，38(21)：189－191.

[2] 李胜利，李昌清，袁平鹏，等. 基于 Web 的电子期刊元数据信息抽取方法[J]. 华中科技大学学报，2007，35(12)：13－15.

[3] 陈 兰，左志宏，熊 毅，等. 一种新的基于 Ontology 的信息抽取方法[J]. 计算机应用研究，2004，21(8)：155－157.

[4] 郭志鑫. 基于本体的文档引文元数据信息抽取[J]. 微计算机信息，2006，22(18)：304－306.

[5] Xavier C，Lluis M，Lluis P. Learning a Perceptron－based Named Entity Chunker via Online Recognition Feedback[EB/OL]. 2003－02. http://www.cnts.ua.ac.be/conll2003/pdf/15659car.pdf.

[6] 张承立，陈剑波，齐开悦. 基于语义网的语义相似度算法改进[J]. 计算机工程与应用，2006，42(17)：165－166.

[7] 周顺先，林亚平，王耀南. 基于主动学习隐马尔可夫模型的文本信息抽取[J]. 湖南大学学报，2007，34(6)：74－77.

[8] Yaoyong L，Kalina B，Hamish C. SVM Based Learning System for Information Extraction[M]. Berlin：Springer，2005.

[9] 李保利，陈玉忠，俞士汶. 信息抽取研究综述[J]. 计算机工程与应用，2003，39(10)：1－5.

(上接第 202 页)

图 4 系统实现界面

达到预期效果，只是在夜间的效果有待进一步的提高。实验表明文中提出的基于 DSP 的汽车牌照识别系统实时性好，系统稳定可靠、机动灵活，具有广泛的市场应用前景。

参考文献：

[1] TI. TMS320DM642 Video/Imaging Fixed－Point Digital Signal Processor Data Manual[M]. Dallas：Texas Instruments，2004.

[2] 李方慧，王 飞，何培琨. TMS3206000 系列 DSPs 原理与应用[M]. 第 2 版. 北京：电子工业出版社，2003.

[3] TI. TMS320C6000 DSP External Memory Interface Reference Guide[M]. Dallas：TI，2004.

[4] TI. TMS320C64x Image/Video Processing Library Programmer's Reference[M]. Dallas：TI，2003.

[5] TI. TMS320C6000 DSP/BIOS Application Programming Interface (API) Reference Guide[M]. Dallas：TI，2004.