结合编辑距离和 Google 距离的语义标注方法 *

张玉芳, 艾东梅, 黄 涛, 熊忠阳
(重庆大学 计算机学院, 重庆 400044)

摘 要: 提出了一种在领域本体指导下对网页进行语义标注的方法。该方法利用编辑距离和 Google 距离从词语的语法和语义两方面综合度量词汇与本体概念之间的语义相关度,从而在网页与本体之间建立映射关系。此外,对网页进行语义标注后,利用标注结果对本体进行有效扩充,使本体更趋于领域化。实验结果表明该方法是行之有效的。

关键词: 语义网; 本体; 语义标注; 编辑距离; Google 距离
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)02-0555-03
doi:10.3969/j.issn.1001-3695.2010.02.042

Approach for semantic annotation based on edit distance and Google distance

ZHANG Yu-fang, AI Dong-mei, HUANG Tao, XIONG Zhong-yang
(*College of Computer*, *Chongqing University*, *Chongqing* 400044, *China*)

Abstract: This paper presented a methodology for semantic annotation of Web pages with domain-specific ontology. This new approach used the edit distance and the Google distance to measure the relativity between words and ontology concepts in terms of syntax and semantic, so as to link Web pages with ontology concepts. In addition, it enriched the ontology with the results of annotation to make the ontology greatly cover the domain knowledge. The experimental result shows that this method is significantly effective.

Key words: semantic Web; ontology; semantic annotation; edit distance; Google distance

语义网是当前 Web 的延伸,它提供了一个通用的框架,对现有 Web 进行扩展。其核心思想是通过增加语义信息,让计算机参与到自动处理 Web 信息的过程,改变它在 Web 中的角色,使其可理解 Web 上的信息,使 Web 应用具有一定的智能,从而实现信息的自动化和智能化处理[1]。在语义网上,信息通过本体以及与本体一致的元数据以结构化的形式描述,并定义了良好的语义,使计算机能够理解网页内容,从而实现计算机之间的智能交互,让互联网真正成为一个全球化的信息共享和智能服务平台。

语义标注是通过本体描述网页中的概念或概念实例,为网页添加语义信息的。它在现有 Web 的基础上增加语义,产生语义元数据,辅助计算机从语义层次上处理信息。通过语义标注的作用,可将 Web 的状态从机器可读提高到机器可理解,这是发展和实现语义网的基础。因此,通过对网页进行语义标注,为网页添加语义元数据、提供机器可处理的语义在语义网的发展中起着至关重要的作用。

本文在传统语义标注的基础上引入了机器学习中的概念学习思想,并结合编辑距离[2]、Google 距离[3]等,提出了一种新的语义标注方法。

1 本体和语义标注简述

1.1 本体和语义标注

在 Berners-Lee 等人[1]提出的语义网模型中,本体层是最核心的组成部分。Studer 等人[4]在进行深入研究之后,给出了本体的定义:本体是共享概念模型的明确的形式化规范说明。这包含四层含义,即概念模型(conceptualization)、明确(explicit)、形式化(formal)和共享(share)。这表明本体是一种能在语义和知识层次上描述系统的概念模型,通过获取某个领域内的共有知识,确定该领域内共同认可的词汇,从不同层次的形式化模型上给出领域内的概念和概念间相互关系的明确定义,帮助不同主体之间进行知识交换,使各个主体对该领域的概念达成一致的理解,实现知识的共享和重用。

语义网中的信息是结构化的,允许人们将领域内的知识表示成概念、概念属性以及概念之间存在的各种关系。当信息用本体进行语义标注后,机器才能理解其含义,从而可以自动地完成互联网上的信息收集和集成。因此,本体不仅是语义网的核心,也是语义标注的基础。完善本体的构造,进行本体扩充[5]也是语义标注的目的。

语义标注即在领域本体的指导下,为网页添加语义元数据的过程,也就是网页中的实体与其语义描述相关联的过程。语

收稿日期: 2009-07-03; **修回日期**: 2009-08-23 **基金项目**: 中国博士后科学基金资助项目(20070420711);重庆市科委自然科学基金计划资助项目(2007BB2372)

作者简介:张玉芳(1965-),女,副教授,硕导,主要研究方向为数据挖掘与网络入侵检测等;艾东梅(1985-),女,硕士研究生,主要研究方向为数据挖掘、语义网(aidongmei1106@yahoo.com.cn); 黄涛(1982-),男,硕士研究生,主要研究方向为统计关系学习、机器学习、数据挖掘; 熊忠阳(1962-),男,教授,博导,主要研究方向为网格与并行处理技术、数据挖掘技术与应用、互联网应用关键技术.

义标注的目的不仅在于让人更好地理解 Web 上的信息,更重要的是让机器解释和理解信息。通过语义标注,建立起 Web 资源与元数据之间的桥梁,对 Web 文档的内容进行扩充,使 Web 资源以更加形式化的方式表示出来。

由于语义标注通常是建立在领域本体的基础上,为文档中的知识及其在领域中的语义提供精确的描述,使本体实例化,可通过语义标注完善本体,有效地进行本体扩充,使本体尽可能多地涵盖该领域的概念、属性和关系。同时,本体的层次结构也可使语义标注模式充分的形式化,有助于约束其结构,减少标注过程中的歧义和错误,辅助人们对标注本身所蕴涵的语义达成一致的认识。

1.2　传统语义标注方法

传统的语义标注方法主要分为手动标注、半自动标注和自动标注三类。手动标注方法大多集中在建立标注及分享标注的工具上。这方面考虑因素主要有标注的表达、有效分享标注的设计以及对标注的评估等[6],但是手动标注存在着效率低、标注的一致性难以保证等问题。半自动语义标注是利用词汇语义分析对网页进行标注的。利用信息抽取(information extraction,IE)技术抽取 Web 页面中的实例,在词汇抽取和分析技术的基础上,建立待标注词汇集合与本体概念类别之间的映射关系[7]。由于当前很少有 IE 是基于本体实现的,这类基于 IE 技术的半自动语义标注方法在提取到数据后就采用启发式方法对抽取结果进行后期处理和映射到本体[8],而从 IE 结果到本体这个语义映射过程需要大量人工干预,降低了自动化程度。自动语义标注主要利用机器学习的方法从统计学的角度为网页建立标注。

在对语义标注系统的研究中,使用得较多的方法是采用半自动技术完成,目前国外已经存在一些标注工具[9],如 KIM、GATE、Onto_mat、MnM、SHOE、Annotea、Annozilla、SMORE、Yawas、Melita、Briefing Associate、Semantic Word、Semantic Markup。现有工具普遍存在以下的不足[10]:多数工具不支持本体词汇扩充,这与语义标注的应用环境相悖;语义标注过程中本体查询、辅助推理支持以及元数据产生的自动化程度还不够。

标注效果的好坏很大程度上取决于标注过程中所使用的方法,不同的标注平台间可以通过其所使用的方法来进行区分。因此,从这个角度来看,语义标注方法主要有两大类,即基于模式的方法和基于机器学习的方法。一般情况下,基于机器学习的标注方法效果优于基于模式方法的标注。

2　基于编辑距离和 Google 距离的语义标注方法

2.1　基于领域本体的标注流程

本文采用的语义标注方法流程如图 1 所示。

标注过程主要分为以下三部分:

a)网页搜集和预处理。其包括网络爬虫搜集网页、网页存储到数据库、GATE(general architecture for text engineering)[11]分句、分词、词性标注等处理,词汇过滤处理,并将预处理结果存储到数据库中。

b)本体预处理。其包括本体获取、本体转换为三元组(subject, predicate, object)形式,从三元组中提取类、实例、属性、关系,并存储到关系数据库中。

c)语义标注。将待标注词汇与本体中的概念计算其编辑距离和 Google 距离,获得综合语义相关度,并通过排序和设定阈值进行取舍。再通过人工验证,将标注结果加入到本体中,对本体的概念、属性及关系进行扩充;对于舍弃掉的词汇,填充到词汇过滤集。

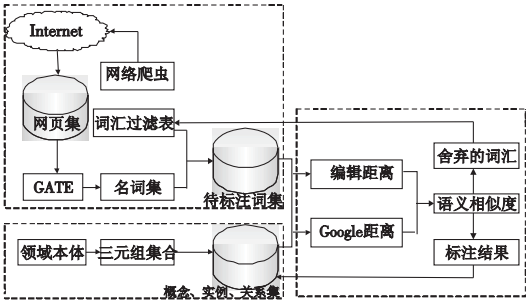

图1　基于语义相关度的语义标注系统框架

2.2　编辑距离和 Google 距离

本文主要从编辑距离和 Google 距离两方面来计算文档词汇与本体概念间的相关度。其中,编辑距离是从词汇的语法构成上来计算词汇间的相似程度,而 Google 距离是从语义的角度反映词汇间的关联程度。

2.2.1　编辑距离

编辑距离最早由俄国科学家 Levenshtein 提出,又称 Levenshtein distance,指两个字符串之间,由一个转成另一个所需的最少编辑操作代价。这里的编辑操作包括替换、插入和删除。

将两个词 x 和 y 的编辑距离用 EDS(x,y) 表示,那么 EDS(x,y) 的取值为$[0,\infty]$。两个词语之间的编辑距离常用动态规划方法来计算,假设词语 x 是一个长度为 m 的字符串 $x_1x_2\cdots x_m$,y 是一个长度为 n 的字符串 $y_1y_2\cdots y_n$,那么词语 x 到 y 的编辑距离即将词语 X 转换为 Y 的最简便的转换序列的代价,可递归地表示为[12]

$$\begin{aligned}&\mathrm{ED}(x,y)=\mathrm{ED}(X_m,Y_n)=\\&\min\begin{cases}\mathrm{ED}(X_{m-1},Y_{n-1})+C_s(X_m,Y_n)\\\mathrm{ED}(X_{m-1},Y_n)+C_d(X_m,\varepsilon)\\\mathrm{ED}(X_m,Y_{n-1})+C_i(\varepsilon,Y_n)\end{cases}\end{aligned}\tag{1}$$

其中:ED(X_{m-1},Y_{n-1}) 表示编辑(替换或删除或插入)x 的前 $m-1$ 个字符,y 的前 $n-1$ 个字符的最小代价;$C_s(X_m,Y_n)$ 表示将 x 的第 m 个字符替换为 Y 的第 n 个字符的代价;$C_d(X_m,\varepsilon)$ 表示将 x 的第 m 个字符删除的代价;$C_i(\varepsilon,Y_n)$ 表示在 x 的第 m 个字符后插入 y 的第 n 个字符的代价。

将上面的编辑距离归一化处理后,得到的归一化编辑距离(normalized edit distance):

$$\mathrm{NED}(x,y)=\frac{\mathrm{ED}(x,y)}{\max\{m,n\}}\tag{2}$$

显然,当 x 和 y 完全相同时,NED = 0;当 x 与 y 完全不相同时,NED = 1,即 NED$(x,y)\in[0,1]$。

归一化编辑距离反映了两个词间靠近或相似的程度。距离越大的两个词被认为越不相似,反之则越靠近。将归一化编辑距离转换为词语间的语法相似度,可用下式表示:

$$\mathrm{sim}(x,y)=1-\mathrm{NED}(x,y)\tag{3}$$

2.2.2　Google 距离

在 Google 中输入词汇进行查询,利用 Google 返回的匹配记录数来计算两个概念间的语义距离,称为 Google 距离,用 NGD(normalized Google distance)表示,它是从语义上分析两个词语的相似性的。任意两个词 x 和 y 的 Google 距离 NGD(x,y) 可表示为

$$\begin{aligned}\mathrm{NGD}(x,y)&=\frac{G(x,y)-\min(G(x),G(y))}{\max(G(x),G(y))}=\\&\frac{\max\{\log f(x),\log f(y)\}-\log f(x,y)}{\log N-\min\{\log f(x),\log f(y)\}}\end{aligned}\tag{4}$$

其中:$f(x)$表示在 Google 中搜索 x 时返回的匹配记录数;$f(y)$表示在 Google 中搜索 y 时返回的匹配记录数;$f(x,y)$表示在 Google 中搜索词组(x,y)时返回的匹配记录数;N 表示 Google 索引的 Web 页面数。NGD(x,y)是词条 x 和 y 共现的对称的条件概率:假设给定一个页面含有 x(或 y),那么 NGD(x,y)就表示这个页面同时含有 y(或 x)的概率。Google 距离有以下几个性质:

a)NGD 的取值范围为$[0,\infty]$。如果 $x=y$ 或 $x\neq y$,而 $f(x)=f(y)=f(x,y)>0$,那么 NGD$(x,y)=0$。这说明词汇 x 和 y 在 Google 中的语义是相同的。如果 $f(x)=0$,那么对于任何搜索词条 y 都有 $f(x,y)=0$,即 NGD$(x,y)=\frac{\infty}{\infty}$。

b)通常情况,NGD 为非负数且对任意 x 有 NGD$(x,x)=0$。对任意 x,y 有 NGD(x,y)=NGD(y,x),它们是相互对称的。

c)概念间的语义距离越大,则相似度越小,表明这两个概念的相关度越小。

由 Google 距离可知,词语与其本身的距离为 0;语义距离为 0 时,相似度为 1;语义距离为无穷大时,相似度为 0。

对于词语 x 和 y,语义距离为 NGD(x,y),Google 相关度(Google relevancy)记为 GR(x,y),那么可定义一个满足以上条件的转换关系:

$$\mathrm{GR}(x,y)=\frac{\lambda}{\mathrm{NGD}(x,y)+\lambda} \tag{5}$$

其中:λ 是一个可调节的参数,且 $\lambda\in(0,1)$。

本文将归一化编辑距离和 Google 相关度结合起来衡量两个词汇间的相关程度,表示如下:

$$\mathrm{weight}(x,y)=\alpha\times\mathrm{sim}(x,y)+\beta\times\mathrm{GR}(x,y) \tag{6}$$

其中:$\alpha+\beta=1$。

3　实验

3.1　数据集

3.1.1　领域本体

本实验采用 http://www.w3.org 提供的 wine 本体,这是一个 OWL DL(Web ontology language description logic)本体,涵盖了 wine 领域内定义的类、属性、实例以及实例间关系。由于本体是对共享概念模型的明确的形式化规范说明,对概念和属性的定义用语都是精确化和概念化的,且大多采用名词形式进行表述。

实验先对本体进行预处理,将本体从 OWL 形式转换为形为(subject, predicate, object)的三元组形式。再将三元组按照概念、实例、关系进行分解,并存储到数据库中,以提高语义标注和本体扩充的效率。

3.1.2　待标注页面

待标注页面来源于互联网,用爬虫工具下载与 wine 相关的英文网页,共 300 篇。由于网页内包含了很多导航栏、广告、图像等噪声信息,需要在下载网页后进行网页预处理,主要包括去除 HTML 标签、文本编码转换、噪声信息剔除、使用 GATE 的 Splitter、Tokeniser、POSTagger 模块进行分句、分词、词性标注预处理。从预处理结果中提取名词,经词汇过滤表筛选后形成待标注词汇列表,并存储到数据库中。

3.2　实验方法

在本实验中,将从抓取的页面中随机抽取 30 篇作为训练集,15 篇作为测试集。先利用编辑距离和 Google 距离的综合权重进行预标注,并根据设定的阈值和人工判断的方式进行取舍,其标注结果扩充到领域本体中,其余的作为反例添加到文档词汇过滤表中。在本体扩充前后分别对测试集中的 15 篇网页进行测试,并给出了不同条件下的标注结果。

3.3　实验结果及分析

表 1 给出了不同实验条件下的结果,分为以下几种情况:只用编辑距离与本体概念计算权重进行标注;只用 Google 距离与本体概念计算权重进行标注;结合编辑距离和 Google 距离与本体概念计算综合权重进行标注。

对于语义标注,传统的评价标准是通过人工判断标注准确率 P、召回率 R 和 F_1 值进行测评,其定义如下:

$$P=\frac{\text{标注正确的词汇数}}{\text{已标注的总词汇数}}\times100\%,\ R=\frac{\text{标注正确的词汇数}}{\text{待标注的总词汇数}}\times100\%$$

$$F_1=\frac{2\times P\times R}{P+R}\times100\%$$

表 1　本体扩充前后的标注性能比较

标注方法		P	R	F_1
扩充前	基于编辑距离标注	0.360 7	0.372 7	0.366 6
	基于 Google 距离标注	0.272 7	0.281 8	0.277 2
	编辑距离 + Google 距离标注	0.378 3	0.390 9	0.384 5
扩充后	基于编辑距离标注	0.668 6	0.690 9	0.679 6
	基于 Google 距离标注	0.513 2	0.530 3	0.521 6
	编辑距离 + Google 距离标注	0.686 2	0.709 0	0.697 5

由以上结果分析可知:

a)利用编辑距离和 Google 距离与扩充后的本体概念计算综合权重进行语义标注的方法整体性能上最优,其 F_1 平均值最高,为 69.75%,准确率也最高,达到 68.62%,这与预期效果是一致的。

b)对于仅用编辑距离或 Google 距离来确定词汇与本体概念的相关度方法,都具有各自的优势和不足。编辑距离仅能从词汇的语法构成上来获得与本体相关的概念和实例,执行效率较高,而通过 Google 距离计算词汇相关度的标注方法具备一定的语义分析和暗示能力。

从表 1 中的数据可以看到,单独用 Google 距离标注的方法的各测评指标值都最低。经过分析发现,这主要是由于用 Google 距离计算语义相关度的基础是整个互联网,而信息量庞大的互联网上存在太多的不相干信息,对计算两个词语之间的语义相关度存在干扰影响。所以,将编辑距离和 Google 距离有机结合,从词语的语法构成和语义两方面综合度量词语间的语义相关度来进行标注,能取得更好的效果。

c)从实验结果可知,将编辑距离与 Google 距离相结合后,与领域本体越相关的词汇,语义权重越大。而对本体进行有效扩充,增加反馈的过程,使本体尽可能多地包含本领域的概念、属性和关系,不但可使本体更具有领域性,更广泛地涵盖该领域的知识,而且可使标注的准确率、召回率和 F_1 值都明显提高,标注效率也得到改善,与语义标注就是本体扩充过程的实质也更加贴切。

4　结束语

本文给出了在领域本体下,利用编辑距离和 Google 距离计算语义相关度以进行语义标注的方法。通过对实验结果的分析表明,该标注方法是切实有效的。网页进行语义标注后得到的数据结构和语义可辅助 Web 发挥其巨大的潜能,不仅将隐含的语义信息显式地表达出来,有助于机器理解和处理,而且也反映了网页内容与领域相关类别的关联情况,是实现智能检索、信息抽取及语义推理的基础。

(下转第 562 页)

示。引擎对每个并发量档次的成功率,如图 4 所示。

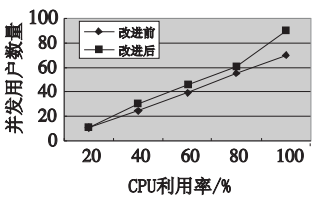

图2　改进前后引擎的最大负载量变化情况

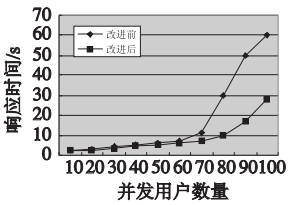

图3　改进前后引擎对每个并发量档次的响应时间

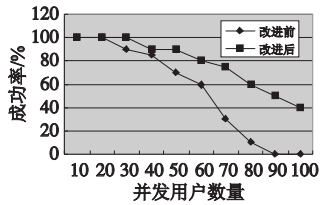

图4　改进前后引擎对每个并发量档次的成功率

通过测试,得到以下结论:

a)当 CPU 利用率稳定在 80% 左右的时候,改进后引擎的最大负载量比改进前大约提高了 13%,而满载负荷(CPU 占用率达到 100%)大约提高了 29%。

b)当虚拟用户的并发数目依次增加达到 100 个的时候,改进后引擎在每个并发量档次的响应时间和成功率分别比改进前都有了明显的提高。改进前引擎的并发用户达到 70 的时候就会出现响应异常的情况,改进之后的引擎对不超过 100 个的并发用户可以正常响应。

以上性能对比测试表明,本文将非阻塞双传输异步调用以及 cache 机制引入到 WebJetFlow 中,显著地提高了引擎的吞吐量,增强了引擎应付高负载请求的能力;引擎在每个并发量档次的响应时间和成功率分别都有了明显的提高。

5　结束语

在 Internet 上,网络流量和距离是导致基于 Web 服务的应用程序性能下降的主要原因之一,因此在遇到任何可能耗时较长的 Web 服务请求时,异步调用模式都是一个很好的方法。

本文讨论了 BPEL4WS 引擎 WebJetFlow 的 Web 服务异步调用机制,在引擎的服务调用代理中,对 Web 服务的调用采用非阻塞双传输异步调用,调用线程无须等待服务的响应,所有响应结果由统一的服务接口接收,调用线程可以继续去处理其他的调用请求,提高了调用线程利用率。同时引入了 cache 机制并设计了相应的 cache 替换算法,保证了那些调用外部响应时间短的服务流程实例可以很快被从内存中取出并往下执行,而对于等待时间较长的流程实例将会被持久化到数据库中,cache 中的实例保证了引擎对异步调用结果消息的匹配效率,数据库中的实例副本提供了数据安全性,即使掉电也不会丢失数据。

在以后的高负载实际测试环境中继续完善该异步调用机制,是笔者下一步要做的工作。

参考文献:

[1] IBM. BPEL4WS, business process execution language for Web services version 1.1[EB/OL]. (2007-02-08)[2009-05-20]. http://ww-w-128.ibm.com/developerworks/library/specification/ws-bpel/.

[2] CARDOSO J, SHETH A, MILLER J, *et al*. Quality of service for workflows and Web service processes[J]. Journal of Web Semantics, 2004, 1(3):281-308.

[3] NORRIS J, COLEMAN K, FOX A, *et al*. OnCall: defeating spikes with a free-market application cluster[C]//Proc of the 1st International Conference on Autonomic Computing. Washington DC: IEEE Computer Society, 2004: 198-205.

[4] 陈伟安,高春鸣. 一种基于反馈控制的 Web 服务组合执行引擎设计[J]. 计算机工程与应用, 2007,43(29):154-158.

[5] 袁小娟,高春鸣. Web 服务组合执行引擎中服务代理运行机制研究[J]. 计算机工程与应用, 2007, 43(28):103-106.

[6] The Axis2 Development Team. Apache Axis2 DOCS [DB/OL]. (2009-05-08)[2009-05-20]. http://ws.apache.org/axis2.

[7] BOSWORTH A, KALER C, FREY J, *et al*. Web services addressing (WS-Addressing)[EB/OL]. (2003-03-13)[2009-05-20]. http://ww-w.microsoft.com/taiwan/msdn/library/2003/Apr-2003/ws-addressin-ng.htm.

[8] 刘志明,彭宇行. 集群 VOD 系统中磁盘 cache 替换算法研究[J]. 计算机工程, 2004, 30(7):139-140,180.

[9] RONALD R P, CHASE J S, GADDE S, *et al*. The trickle-down effect: Web caching and server request distribution[J]. Computer Communications, 2002, 25(4):345-356.

(上接第 557 页)

参考文献:

[1] BERNERS-LEE T, HENDLER O L J. The semantic Web[J]. Scientific American, 2001,284(5):37-43.

[2] LEVENSHTEIN I V. Binary codes capable of correcting deletions, insertions, and reversals[J]. Soviet Physiscs Doklady, 1966,10(8):707-710.

[3] RUDI L, CILIBRASI P M B V. The Google similarity distance[J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(3):370-383.

[4] STUDER R, BENJAMINS V R D F. Knowledge engineering, principles and methods[J]. Data and Engineering,1998,25(1-2):161-197.

[5] FAATZ A, STEINMETZ R. Ontology enrichment with texts from the WWW[C]//Proc of WS'02. 2002.

[6] STABB S H A S. Authoring and annotation of Web pages in CREAM [C]//Proc of WWW2002. 2002.

[7] KIRYAKOV A, POPOV B, TERZIEV I, *et al*. Semantic annotation, indexing, and retrieval[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2004, 2(1):7-14.

[8] PAOLOZZI S, ATZENI P. Interoperability for semantic annotations [C]//Proc of the 18th International Conference on Database and Expert Systems Applications. 2007:445-449.

[9] [EB/OL]. http://annotation.semanticweb.org/tools.

[10] 邹亮,廖述梅. 基于本体的语义标注工具比较与分析[J]. 计算机应用,2004,24(S1):328-330.

[11] PORTOBELLO H C, CUNNINGHAM H, HUMPHREYS K, *et al*. GATE:a general architecture for text engineering[R]. [S.l.]:University of Sheffield, 2002.

[12] BILENKO M, MOONEY R J. Adaptive duplicate detection using learnable string similarity measures[C]//Proc of the 9th ACM SIGKOD International Conference on Knowledge Discovery and Data Mining. New York:ACM Press,2003:39-48.