文章编号：1003-0077(2018)03-0091-10

基于双语 URL 匹配模式可信度的平行网页识别研究

章成志[1,2]，马舒天[1]，揭春雨[2]，姚旭晨[2,3]

(1. 南京理工大学 信息管理系，江苏 南京，210094；
2. 香港城市大学 翻译及语言学系，香港；
3. 百度在线网络技术(北京)有限公司，北京 100085)

摘　要：平行语料是自然语言处理中一项重要的基础资源，在双语平行网页中大量存在。该文首先介绍双语 URL 匹配模式的可信度计算方法，然后提出基于局部可信度的双语平行网页识别算法，再依据匹配模式的全局可信度，提出两种优化方法：即利用全局可信度，救回因低于局部可信度阈值而被初始算法滤掉的匹配模式；通过全局可信度和网页检测方法，挖出深层网页。进一步，结合网站双语可信度、链接关系，侦测出种子网站周边更多较具可信度的双语网站。除了双语 URL 匹配模式自动识别，还利用搜索引擎，依据少数高可信度的匹配模式快速识别双语网页。为了提高以上五种方法识别候选双语网页对的准确率，计算了候选双语网页对的双语相似度，并设置阈值过滤非双语网页对。通过实验验证了所提方法的有效性。

关键词：平行网页获取；平行语料库；双语 URL 匹配模式；双语文本挖掘

中图分类号：TP391　　**文献标识码**：A

Detection of Parallel Web Pages Based on the Automatically Discovered Bilingual URL Pairing Patterns

ZHANG Chengzhi[1,2], MA Shutian[1], KIT Chunyu[2], YAO Xuchen[2,3]

(1. Department of Information Management, Nanjing University of Science & Technology, Nanjing, Jiangsu 210094, China;
2. Department of Linguistics and Translation, City University of Hong Kong, Hong Kong, China;
3. Baidu Online Network Technology (Beijing) Co. Ltd., Beijing 100085, China)

Abstract: Parallel corpora are one of the most important resources for natural language processing, a large volume of which can be mined from bilingual parallel web pages. This paper formulates a practical algorithm for recognizing parallel web pages based on the credibility of automatically discovered bilingual URL pairing patterns (or keys), then this paper extends it in two ways to find more parallel web pages, namely, rescue weak keys of low local credibility in terms of their global credibility, and unearth bilingual parallel deep web pages by means of applying strong keys of high global credibility. Furthermore, we detect more bilingual web sites according to their credibility in terms of their link relationship with the seed set of web sites in use, and also utilize search engines to recognize bilingual web sites efficiently with only a small set of URL pairing patterns of high credibility. To further enhance the recognition accuracy on top of these five methods, we calculate cross-lingual similarity of candidate parallel web pages and filter out weak ones with a threshold. The effectiveness of our approaches is confirmed by a series of experiments.

Key words: parallel webpage mining; parallel corpora; bilingual URL pairing pattern; bilingual text mining

收稿日期：2016-02-29　**定稿日期**：2017-10-12

基金项目：香港城市大学 SRG-Fd 项目(7008003)；香港研资局 GRF 项目(CityU 144410，11600415)；国家自然科学基金(70903032)

0 引言

平行语料库是指两种或多种语言在段落、句子甚至单词短语层面上互为翻译的语料。作为自然语言处理领域中的宝贵资源,平行语料在统计机器翻译[1]和跨语言检索[2]等任务中扮演着重要的角色。已有的平行语料库,无论在语种数量、语料规模、质量还是覆盖领域等方面,都仍需不断完善扩充,以满足实际需求。

过往的研究利用双语或多语网站来获取平行语料(包括双语平行和双语混合网页),并搭建了一些双语网页获取系统,如 STRAND[3]、BITS[4]、PTMiner[5]、PTI[6]及 WPDE[7]等。另外一种代表性方法则依据 URL 组成的模式,通过启发式规则从双语网站上自动发现双语网页,相比手工制定启发式规则,通过机器自动发现规则,能在一定程度上减少计算资源的开销[8-9]。

本文基于后一种方法,对双语 URL 匹配模式探测、模式可信度计算及应用等方面,进行比较全面的设计和实验[8-10]。首先,计算双语 URL 匹配模式的可信度;其次,在此基础上提出四种双语网页识别方法;然后,利用搜索引擎以及少量的高可信度双语 URL 匹配模式快速识别双语网页,以降低对匹配模式的过分依赖;最后,利用网页链接与高可信度的 URL 匹配模式计算候选网页对的双语相似度,由此来过滤非双语网页对,以进一步提高候选双语网页对的准确率。通过一系列实验,我们验证了所提方法的有效性。

1 相关研究概述

STRAND[3]是最早用于识别双语平行网页的系统之一,该系统通过搜索引擎检索指向不同语种版本链接的网页,然后将文本语种比较、URL 配对以及文本长度作为判别特征,生成候选平行网页对,最后利用网页结构进行过滤。PTMiner[5]首先利用链接锚文本来识别候选双语网站,通过搜索引擎得到这些网站下的网页,并利用 URL 模式找出平行对,最后通过网页内外部特征进行过滤。类似的挖掘系统还有 BITS[4]、PTI[6]、WPDE[7]等。另外,平行网页的识别方法也在不断更新,例如通过 DOM 树对齐模型来识别互译文本和两个平行 DOM 树之间的链接[11],利用 HTML 结构实现平行网页的递归访问,使用 URL 模式优化遍历平行网站的拓扑顺序,来获取平行网页[12]。另外,网页之间的链接关系也被用于计算网页之间的相似程度,迭代挖掘出平行网页[13]。

这些方法大多独立于语言,具体步骤为:抓取和识别候选双语网站、提取候选平行网页对,进而验证。其中,平行网页网址的先验知识常用于网页抓取或过滤。已有研究主要依靠两类信息来获取平行网页:一是单个网页信息,包括网址和网页内容;二是多个网页信息,主要是网页之间的链接关系。也有很多研究者利用搜索引擎检索表示语言类别的锚文本来定位候选双语网站。此外,网址中是否含有预先定义的双语 URL 模式也常被用来判断候选平行网页。然而,这些预定义的规则不可能涵盖所有情况,很多网站甚至没有任何关于语言类别的锚文本标记。因此,我们试图通过机器自动发现规则,来降低基于双语 URL 匹配模式的方法对外部先验知识的依赖性[8-9]。同时,我们还依据少量匹配模式,快速识别双语网页[10]。另外,为进一步提高这些方法所识别出的候选双语网页对的准确率,我们提出非双语网页对过滤算法。

2 研究总体框架

如图 1 所示,双语网页在双语网站上有多种出现模式,根据源语言与目标语言网页结构对应强度的不同,可以分为强、弱和无对应关系的双语网页(深层网页)。我们根据网页的 URL 结构,计算双语 URL 匹配模式可信度,并据此提出五种识别双语网页的算法,开发了相应的双语网页获取与评估系统 Pupsniffer①。该系统基于先前工作[8]并对其算法进行了优化,是一个很有用的多语网页自动挖掘工具[9]。

① https://code.google.com/p/pupsniffer/

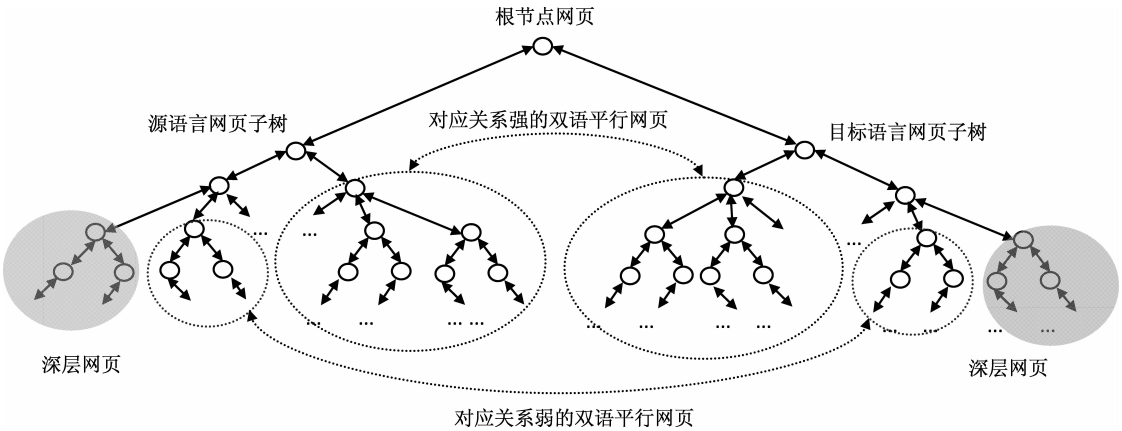

图 1　候选双语网站的网页对应结构示意图

如图 2 所示，Pupsniffer 系统分为三个模块，第一个模块是双语网页挖掘，根据所给的种子网站列表进行网页爬取，结合链接分析与双语 URL 匹配模式，利用五个主要算法获取双语网页，即：基于模式局部可信度的双语网页发现算法[8]和两个优化方法，分别是弱匹配模式救回算法和深层双语网页检测算法，以及深层双语网页发现增量算法[9]和仅考虑少量先验知识的双语网页获取方法[10]。第二个模块是非双语网页的过滤，利用网页链接，以及双语 URL 匹配模式进行过滤（图 2⑥）。第三个模块是候选双语网页测评，即对所得到的双语网页 URL 进行随机抽样并人工测评，最后得到测评结果。

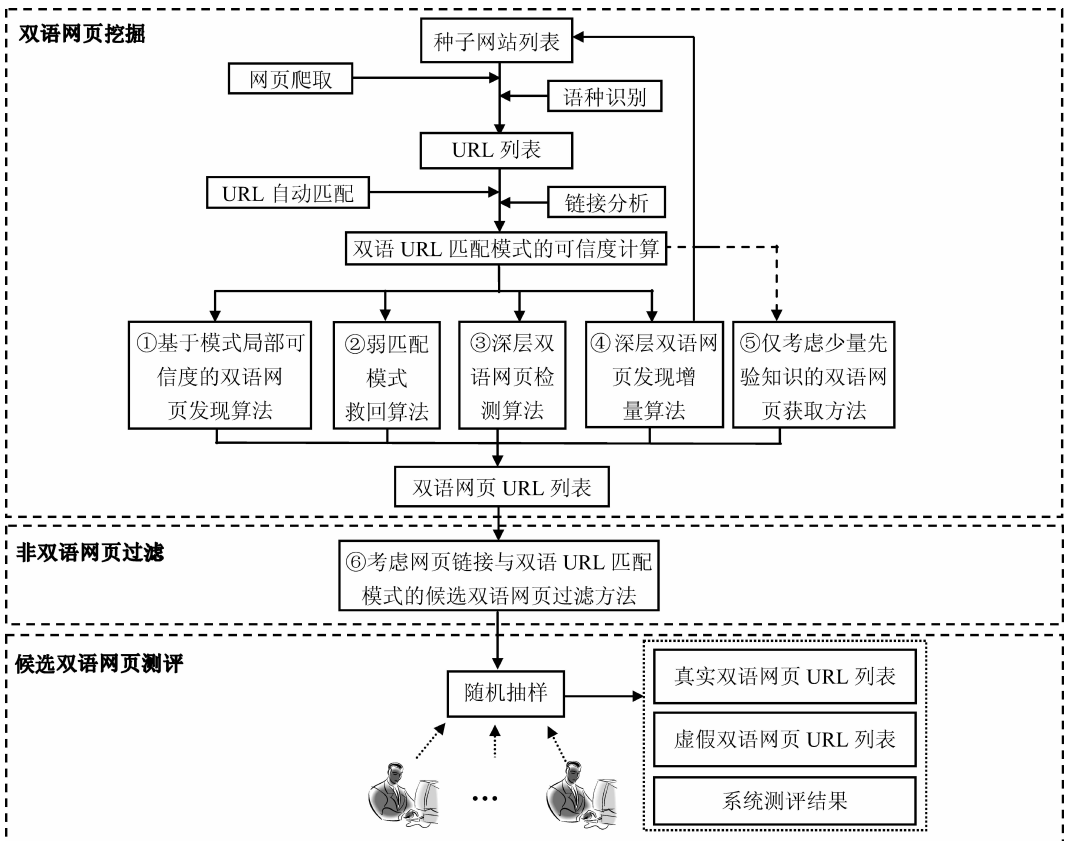

图 2　双语网页获取与评估系统总体框架图

3 双语URL匹配模式的可信度计算方法

针对某个网站下采集得到的网页，我们首先对其内容进行简单的语言识别，即：网页内容中超过50%的字符为英文字母，则判断该网页为英文网页，否则为中文网页[8]。然后，我们对网页URL进行切分等预处理，得到两个字符串单元集合，即网址路径的单元集合和网址文件名的单元集合，接着分别对这两个集合及其总集合进行双语URL匹配模式的识别[8]。

定义1(双语URL匹配模式)：给定一个双语网站的源语言与目标语言网页URL集合为U和U'，相应的字符串单元集合为T和T'，若从一个候选双语URL对$\pi=\langle u,u'\rangle\in U\times U'$中抽去一个单元对$k=\langle t,t'\rangle\in T\times T'$后，剩下的单元集合相同，即$u-\{t\}=u'-\{t'\}$，则该单元对$k$记为一个候选的双语URL匹配模式。

相应地，一个双语URL匹配模式$k=\langle t,t'\rangle$的得分计算可形式化为：

$$f(k,\pi)=\begin{cases}1, & \text{若 } u-\{t\}=u'-\{t'\}\\ 0, & \text{否则}\end{cases} \tag{1}$$

其中，$u-\{t\}$和$u'-\{t'\}$分别为从网址u和u'中抽去模式$\langle t,t'\rangle$中的字串t和t'后剩下的单元集合。举例来说，给出如下一对网址：

英文URL：http://www.legco.gov.hk/yr99-00/*english*/fc/esc/minutes/es061099.htm

中文URL：http://www.legco.gov.hk/yr99-00/*chinese*/fc/esc/minutes/es061099.htm

其中所含的“english”和“chinese”两个字符串显示出这两个网址所对应的语种及平行关系，根据以上定义，我们将“＜english，chinese＞”这样的字符串单元对选为一个候选双语URL对的匹配模式，或称匹配键(key)。遍历一个双语网站中所有的候选双语URL对后，每个匹配键得到一个总得分，即其在该网站中可能匹配上的双语URL对的总数。

定义2(双语URL匹配模式的频次)：双语URL匹配模式k(简称模式k)的频次为遍历给定网站w中所有的候选双语URL对后模式k的总得分，即其在w中可能匹配上的双语URL对的总对数，计算如式(2)所示。

$$p_{\pi\in U\times U'}(k,w)=\sum_{\pi\in U\times U'}f(k,\pi) \tag{2}$$

定义3(双语URL匹配模式的局部可信度)：模式k的局部可信度为给定网站w中k可能匹配上的双语网页数与w中URL总数的比值，计算如式(3)所示。

$$C(k,w)=\frac{N(k,w)}{|w|} \tag{3}$$

其中，$N(k,w)$为网站w中k可能匹配上的双语网页数，是双语URL对数目的两倍，即：$N(k,w)=2*p_{\pi\in U\times U'}(k,w)$，$|w|$为网站$w$的网页总数。

通常，在某一个网站上可信度高的双语URL匹配模式，不一定在所有的网站上都具有较高的可信度，而在大多数网站上都出现的匹配模式一般来说其可信度都较高。基于这个假设，我们给出双语URL匹配模式的全局可信度概念。

定义4(双语URL匹配模式的全局可信度)：对候选网站集合W中每个网站，将模式k可能匹配上的URL总数归一化后，与k的局部可信度相乘，然后对所有乘积求和，该乘积和称为模式k的全局可信度，计算如式(4)所示。

$$\begin{aligned}C(k)&=\sum_{w_i\in W}\frac{N(k,w_i)}{N}\times C(k,w_i)\\&=\sum_{w_i\in W}\frac{N^2(k,w_i)}{N*|w_i|}\end{aligned} \tag{4}$$

其中，N为候选网站集合W中所有网站网页总数，w_i为候选网站集合中第i个网站。由于N为常量值，不影响模式k全局可信度的排序结果，实验中无需加入计算。

定义5(网站的双语可信度)：网站w的双语可信度为其中所有双语URL匹配模式的局部可信度最大值，计算如式(5)所示。

$$C(w)=\max_k C(k,w) \tag{5}$$

4 基于双语URL匹配模式可信度的双语网页识别方法

在双语URL匹配模式可信度计算的基础上，我们提出四种适用于不同场景的双语网页识别方法。

4.1 基于双语URL匹配模式局部可信度的双语网页识别算法

基于双语URL匹配模式局部可信度的双语网页发现算法(图2①)假设双语网站中的双语平行网页对通常包含固定的URL匹配模式。该算法原理

如下：

给定从双语网站 w 采集到的所有 URL 地址，若其中的一对网址 u 与 v 只有一处不同，则此不同处为可能的双语 URL 匹配模式。然后，我们计算网站 w 中的双语匹配模式的局部可信度，给定阈值（实验中设为 0.1），得到双语匹配模式局部可信度超过该阈值的候选双语匹配模式，最后根据候选模式，得到候选双语网页[8]。

4.2　弱匹配模式救回算法

在初始算法中，设置局部可信度阈值显然会过滤掉局部可信度低但全局可信度可能较高的双语匹配模式及其对应的双语网页。为此，我们提出两种方法来解决这一问题。

首先，对于这样的匹配模式，我们设定一个全局可信度阈值 θ（实验中设为 500）①，若其可信度不低于 θ，则仍保留该匹配模式及其对应的双语网页。

其次，对于两种可信度都较低但当前网站对应域名的可信度较高②的情况，由于这种类型的网站可能包含大量的双语 URL 对，例如“gov.hk”域名，我们降低局部可信度阈值，从而获取更多可能的双语网页。

4.3　深层双语网页检测算法

有些网页只有通过数据库检索才能临时生成，这类网页称为深层网页③。在双语网站中，深层网页包括如下几种情况：(1)全子树深层网页，即网站的单语子目录无法被抓取；(2)部分子树深层网页，即部分子树对应网页不能被抓取；(3)部分节点深层网页，即双语网站的某些网页无法被抓取，尤其是动态创建的网页。

我们利用全局可信度高的双语 URL 匹配模式，生成深层网页 URL 对应的另一语种的网页 URL。实验中我们取全局可信度前 10 位的双语匹配模式进行深层双语网页检测。例如，中文网页 http://www.fehd.gov.hk/tc_chi/LLB_web/cagenda_20070904.htm 所对应的英文网页如果爬虫爬不到，则选择全局可信度高的双语匹配模式“<english, tc_chi>”，用“english”替换“tc_chi”来生成英文 URL④，然后用超文本传输协议检查生成的 URL 是否有效，若有效，则收集到候选双语网页对中。目前，大多数双语网页发现方法都是基于网页结构和内容的相似度计算，而没有事先获得候选 URL 对，因此都无法发现深层双语网页。我们把深层双语网页检测整合到双语网页发现方法中，可挖掘出更多高可信度的双语网页。

4.4　深层双语网页发现增量算法

双语网站往往与其他的双语网站存在链接关系。因此，如果给定双语网站列表，可以通过解析网站中的网页来采集外部网站，从而发现更多的双语网站。基于该想法，我们利用链接分析，结合网站可信度获得更多的候选双语网页。

定义 6（网站的链出数）：给定种子网站集合 $W_{seed}=\{w_1,w_2,\ldots,w_i,\cdots,w_N\}$，其中网站 w_i 的链出数是指从网站 w_i 链接到 W_{seed} 中其他网站的数量总和，记为 Linkout(w_i)。

定义 7（网站的权威度）：网站 w_i 的权威度为其 PageRank 值[14]，记作 PR(w_i)。

定义 8（考虑可信度的网站权威度）：考虑可信度的网站 w_i 权威度为 w_i 可信度与其 PageRank 值的乘积，即加权的（weighted）PR 值，记作 WPR(w_i)，计算公式如式(6)所示。

$$WPR(w_i) = C(w_i)PR(w_i) \tag{6}$$

为了减少系统开销，Linkout(w_i) 和 PR(w_i) 的计算仅依据种子网站之间的链接关系。根据定义 6～8，网站 w_i 包含 Linkout(w_i)、PR(w_i) 和 WPR(w_i) 三个量值。依此，我们分别使用这三个指标来度量一个相关外部网站的可信度，即其各指标的总和：$\sum$Linkout、$\sum$PR 和 $\sum$WPR 值。

使用这些指标的双语网页获取增量算法的具体步骤如表 1 所示。在每次迭代中，计算相关参数并得到新的候选种子网站及其网页。其中，预设的从外部网站选取候选网站的个数 K 可以换成一个适当的比例值，或为所用遴选指标的一个经验阈值。在我们的实验中，为了简化处理过程，该算法一次运行中同时计算三个遴选指标并输出结果，K 设定为

---

① 我们给出 θ=100 时对应的双语匹配模式及其全局可信度：http://mega.lt.cityu.edu.hk/~czhang22/pupsniffer-eval/Data/Pattern_Credibility_LargeThan100.txt

② 我们通过双语 URL 匹配模式的可信度与域名进行关联统计，得到 URL 集合中每个域名的可信度。

③ https://en.wikipedia.org/wiki/Deep_web_%28search%29

④ 对应英文 URL 为：http://www.fehd.gov.hk/english/LLB_web/cagenda_20070904.htm，目前该网页已失效。

500，迭代次数设定为1次。

我们邀请了两位硕士研究生分别评估这样获得的候选相关双语网站的前500个。依照上述三个指标，图3显示所识别的前N个候选网站中真正双语网站的数量走势，图4显示所识别的双语网站的正确率。可以看出，$\sum$WPR指标优于其他两种指标，在前500个候选网站中，识别出为真双语网站的准确率接近50%。

表1 深层双语网站发现增量算法描述

算法名称：深层双语网站发现增量算法

输入：已知种子网站列表$w_i(i=1,\cdots,N)$和这些种子网站的相关外部网站列表；
一个预设的外部网站候选数K和迭代次数L。

遴选指标：一个外部网站的$\sum$Linkout、$\sum$PR或$\sum$WPR值。

输出：相关外部候选双语网站和网页。

步骤：

1. 计算每个种子网站w_i的双语可信度$C(w_i)$、链出数Linkout(w_i)、基于PageRank值的权威度PR(w_i)和加权PageRank值WPR(w_i)；
2. 获取种子网站w_i的所有相关外部网站，计算每个相关外部网站的遴选指标值；
3. 根据遴选指标值对相关外部网站排序；
4. 选取该遴选指标排行榜前K个相关外部网站，加入到种子网站列表；
5. 重复步骤1～4，直至预设的迭代次数L或无法发现新的候选相关外部网站；
6. 从新发现的候选相关外部种子网站中，依据初始算法识别出双语网页。

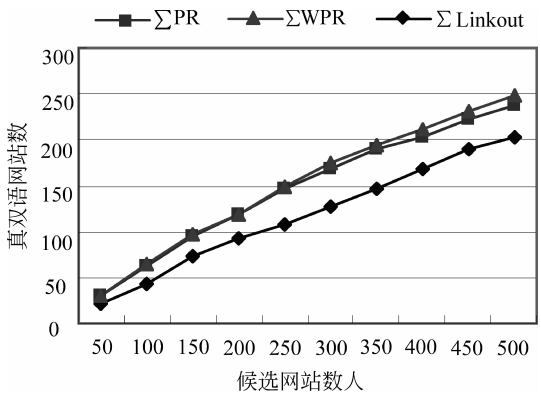

图3 前N个候选网站中真正双语网站的数量走势

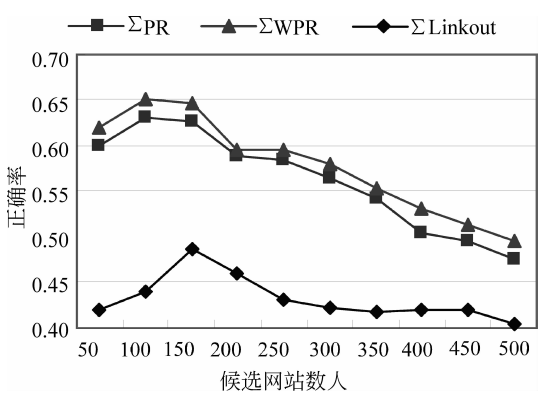

图4 前N个候选双语网站的正确率

4.5 基于少量先验知识的双语网页获取算法

为降低对初始种子网站和双语URL匹配模式的过度依赖，我们利用搜索引擎的优势，仅依据少量的高可信度双语URL匹配模式，快速识别双语网页[10]，具体步骤如下：

(1) 获取双语URL匹配模式中目标语言的标识符

URL中标识语种类型的字符串通常为该语言的英文单词或缩写，例如英文网页URL中可能包含“english”“eng”“en”等字符串。为此，我们可从双语URL匹配模式中获取目标语言的标识符。根据双语匹配模式及其全局可信度的计算结果，得到可信度排名靠前的双语URL模式，如“<en，tc>”“<eng，tc>”“<english，tc_chi>”等，其英文标识分别为“en”“eng”“english”。

(2) 依据搜索引擎快速获取候选双语种子站点

通过搜索引擎的搜索规则，构造查询式，我们可以快速获取候选的双语种子站点。例如：通过“site:”限定方式，可将搜索范围限定在香港政府(gov.hk)、教育(edu.hk)等类型的网站；通过“inurl:”来保证URL中含有“en”“eng”“english”等语言标识符；此外通过“filetype:”限定URL对应的文件类型。通过查询式“inurl：en site：gov.hk filetype：html”，我们能在Google上快速得到香港政府相关网页，在此基础上得到候选双语种子站点列表。

(3) 获取候选双语网页

依据双语匹配模式的全局可信度计算结果，我们得到与目标语言标识对应的排名前N(实验中设

为 5)的双语 URL 匹配模式。对候选双语网站的目标语言网址，按照可信度由高到低的顺序，将目标语言标识符替换为源语言标识符，从而得到候选的源语言网页 URL。根据 HTTP 协议判断源语言网页 URL 是否有效，将有效的 URL 对作为候选的双语网页 URL。

5　基于网页链接与双语 URL 匹配模式的非双语网页对过滤方法

一对平行双语网页所具有的网页链接往往互为平行网页。我们还可以根据识别出的候选平行网页对中各自的网页链接，借助少量高可信度双语 URL 匹配模式计算候选网页对中源语言与目标语言网页的双语相似度。然后，通过阈值进一步从候选网页对中过滤出非双语网页，以提高准确率。

定义 9(候选双语网页对的双语相似度)：给定一对候选双语网页对(目标语言网页 w_T 和源语言网页 w_S)，其双语相似度定义为它们的网页链接(分别为 L_T 和 L_S)中共同网页的相似度与利用双语 URL 匹配模式匹配上的双语网页相似度之和：

$$\mathrm{Sim}(w_T,w_S)=\alpha\times\mathrm{Sim\_Same}(L_T,L_S)+(1-\alpha)\mathrm{Sim\_Key}(L_T,L_S) \quad (7)$$

其中，α 是两者的相对权重(实验中，设为 0.5)，Sim_Same(L_T,L_S)为 L_T 和 L_S 中共同网页对的总网页数与 L_T 和 L_S 总网页数的比值：

$$\mathrm{Sim\_Same}(L_T,L_S)=\frac{2\times p(L_T,L_S)}{|L_T|+|L_S|} \quad (8)$$

对 L_T 和 L_S 中所有能够利用双语 URL 匹配模式匹配得上的双语网页对 π，将其匹配模式 k 匹配上的 URL 总数 $N(k,\pi)=2\times p(k,L_T\cup L_S)$ 与 k 的全局可信度 $C(k)$ 相乘，将所有这样的乘积和与总网页数的比值记作：

$$\mathrm{Sim\_Key}(L_T,L_S)=\frac{\sum_{\pi\in L_T\times L_S}C(k)N(k,\pi)}{|L_T|+|L_S|} \quad (9)$$

这个基于双语相似度的非双语网页对过滤算法适用于以上所有的双语网页对发现算法的输出。

6　结果评估与分析

我们对上面提出的四种双语网页发现方法、基于少量先验知识的双语网页发现算法以及基于网页链接及匹配模式的非双语网页对过滤方法，进行一系列实验，本节报告试验结果，并进行评估与分析。

6.1　基于四种不同双语网页发现

实验中，我们基于 12 800 个种子网站分别对以上四种方法所发现的双语网页进行质量评估。这些种子网站来源于香港，从如下两个途径获得：一个是香港网站目录[①]，截止 2010 年 7 月 17 日，该目录列出了 9 922 个网站；另一个是香港万维网数据库[②]中的 4 230 个网站列表。删除无效网站后，共获得大约 12 800 个候选种子网站[③]。

我们开发了双语网页的质量评估网站[④]，通过随机抽样方式对双语网页识别方法进行评估。我们邀请了五人(一位博士和四位硕士生)参加评估。评估人员需要判断候选双语网页对是否为真实的双语网页对。

经过实验，我们共发现 348 058 对候选双语网页。表 2 给出了不同方法的统计数据和正确率。可以看出，四个方法的整体正确率为 94.72%，基于双语 URL 匹配模式局部可信度的双语网页发现算法的正确率为 94.06%，利用弱匹配模式救回算法、深层双语网页检测算法以及深层双语网页发现增量算法，能额外多发现 21.82%的高可信度双语网页。

我们也分析了实验结果中 910 个的虚假双语 URL 对，将它们分为五类，其中："语言识别错误"，是由于 Pupsniffer 语言识别模块存在识别结果错误而造成的；"无效 URL"，是指由于网页采集时网站正在维护或者它们本身就不存在，造成源语言或目标语言 URL 无效；"只有单语"，是指 URL 对所对应的候选双语网页实际上都是同一语种网页；"内容提取错误"，是指有些候选网页是非纯文本文件；"虚假双语文本"，是指从网页内容来看候选双语网页不是真实的双语网页。经过统计发现，约 80%的虚假双语 URL 对是由于语言识别错误造成的，因此从理论上来说，如果能够解决这种类型的错误，识别出的双语网页正确率将提高至 98.79%。

① http://www.852.com/

② http://www.cuhk.edu.hk/hkwww.htm，注：该网页现已失效。

③ http://mega.lt.cityu.edu.hk/~czhang22/pupsniffer-eval/Data/All_Seed_Websites_List.txt

④ http://mega.lt.cityu.edu.hk/~czhang22/pupsniffer-eval/

表2 不同双语网页发现算法的质量评估

方法类型	参数设置	总对数	增量率/%	抽样率/%	随机抽样		
					真平行对	假平行对	正确率/%
基于模式局部可信度的双语网页发现算法	局部可信度阈值为0.1	290 247	—	3.50	9 541	603	94.06
弱匹配模式救回算法	全局可信度阈值为500	10 015	3.45	14.98	1 339	161	89.27
深层双语网页检测算法	检测依据：全局可信度前10位的双语URL匹配模式；检测范围：种子网站集合中域名为“gov.hk”的网站	15 825	5.45	16.26	2 445	128	95.03
深层双语网页发现增量算法	发现范围：每次迭代过程中种子网站的相关网站总数量K为500；迭代次数：1次	37 491	12.92	8.02	2 988	18	99.40
总计	—	348 058	21.82	4.95	16 313	910	94.72

6.2 基于少量先验知识的双语网页获取

我们依据目标语言的标识符（如“english”“eng”“en”等）及其对应全局可信度排名前五的双语匹配模式，利用Google搜索引擎检索到88 915对中英文URL①。同样，我们通过随机抽样来评估所发现的双语网页，结果如表3所示：4 460个中英文URL对中，有4 051对为真实的中英文双语网页对，双语网页发现的正确率为90.83%。虽然该方法相比于[8-9]的结果较低，但该方法仅考虑少量先验知识、以较少的系统开销即可发现一定规模的双语网页。

表3 不同双语网页发现方法的结果比较

方法类型	总对数	抽样率/%	随机抽样		
			真实平行对	虚假平行对	正确率/%
文献[8]方法	290 247	3.50	9 541	603	94.06
文献[9]方法	348 058	4.95	16 313	910	94.72
文献[10]基于少量先验知识的方法	88 915	5.02	4 051	409	90.83

对虚假双语URL对的错误进行统计，发现虚假双语URL对的错误主要集中在“只有单语”与“内容提取错误”这两种类型。

6.3 非双语网页过滤

基于网页链接与双语URL匹配模式的双语网页过滤方法，可以对以上各双语网页发现算法的候选结果进行进一步过滤。本节仅报告针对7.2节中的候选双语网页对所进行过滤的结果。根据该节得到的候选网页对，我们爬取到69 847②对有效的链接网址③。在进行高可信度双语URL模式匹配时，我们首先排除双语匹配模式中非中文英文对的模式，然后选择了全局可信度排名前30的双语匹配模式来进行双语候选网页对页面链接的匹配。

在计算候选双语网页对双语相似度时，为了降低计算复杂度，我们在实验中没有考虑各个模式的可信度，不同模式可信度均为1。我们将候选双语网页对的双语相似度阈值设置为0，即相似度为0时将该候选对滤掉。69 847对候选对中一共有2 664对的双语相似度为0。这些过滤掉的网页中，2 275对确实为非双语候选网页对，过滤的正确率达85.40%，它们的类型分布如表4所示。该方法仅利用网页链接和部分高可信度URL匹配模式，即可过滤掉一定规模的非双语网页，显然能进一步提高候选双语网页对的准确率。

① 检索日期为2014年2月。

② 有部分网页对未爬取到，原因是其中一个网页失效，或两个网页均失效，爬虫爬取时无反应。

③ 网页抓取日期为2016年9月。

表 4　非双语网页对的类型分布

非双语网页对类型	无效的 URL	只有单语	内容提取错误	虚假双语文本
总数	1 392	35	591	257
比率/%	61.19	1.54	25.98	11.30

7　结论与未来工作

本文对基于 URL 组成模式的双语网页发现方法进行了比较全面的设计和实验：(1)计算双语 URL 匹配模式的可信度；(2)在可信度计算的基础上，提出四种不同的双语网页识别算法；(3)利用搜索引擎的优势、仅依据少量的高可信度双语 URL 匹配模式，快速识别双语网页；(4)最后，利用双语候选网页的双语相似度，进一步过滤非双语网页对。通过实验，我们验证了所提方法的有效性。

今后的主要研究方向包括：(1)获取更多候选双语种子网站：一方面可以通过提出的增量算法寻找双语网站和网页；另一方面我们可以从网上公开目录得到候选网站列表；(2)进一步优化双语 URL 匹配模式可信度以及网站的双语可信度计算方法，比如：利用候选双语网页的链接关系来计算每个页面的 PageRank 值，然后利用 PageRank 值对双语 URL 匹配模式可信度进行加权；优化双语匹配模式全局可信度的计算方法；另外，在同一网站中考虑更多的双语匹配模式作为双语网站可信度计算依据。(3)研究在不需要双语种子网站或者尽量少的双语种子网站的情况下，获取大规模双语网页的方法。(4)在本文基础上，进一步抽取双语平行网页的正文、生成平行句对，最后利用标准数据集测试机器翻译结果的 BLEU 值，从侧面来评估本文最终生成的平行语料的质量。

参考文献

[1] Brown P F, Pietra V J D. Pietra S A D, etal. The mathematics of statistical machine translation: Parameter estimation[J]. Computational linguistics, 1993, 19(2), 263-311.

[2] Davis M W, Dunning T E. ATREC evaluation of query translation methods for multi-lingual text retrieval [C]//Proceedings of the TREC-4, 1995: 483-498.

[3] Resnik P. Parallel strands: A preliminary investigation into mining the web for bilingual text[C]//Proceedings of the AMTA 1998: MachineTranslation and the Information Soup, 1998: 72-82.

[4] Ma X, Liberman M. Bits: A method for bilingual text search over the web[C]//Proceedings of the Machine Translation Summit VII, 1999: 538-542.

[5] ChenJ, NieJ-Y. Parallel web text mining for cross-language IR[C]//Proceedings of the RIAO2000, 2000: 62-77.

[6] Chen J, Chau R, Yeh C-H. Discovering parallel text from the WorldWideWeb[C]//Proceedings of the 2nd Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation, 2004(32): 157-161.

[7] Zhang Y, Wu K, Gao J, etal. Automatic acquisition of Chinese-English parallel corpus from the web [C]// Proceedings of the 2006 European Conference on Advances in Information Retrieval. 2006: 420-431.

[8] Kit C, Ng J Y H. An intelligent web agent to mine bilingual parallel pages via automatic discovery of URL pairing patterns[C]//Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence & Intelligent Agent Technology Workshops, 2008: 526-529.

[9] Zhang C, Yao X, Kit C. Finding more bilingual webpages with high credibility via link analysis [C]// Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, 2013: 138-143.

[10] Ma S, Zhang C. Automatic collection of the parallel corpus with little prior knowledge[C]//Proceedings of the 2014 China National Conference on Computational Linguistics, 2014: 95-106.

[11] Shi L, Niu C, Zhou M, etal. A DOM tree alignment model for mining parallel data from the web[C]// Proceedings of the 2006 International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. 2006: 489-496.

[12] 刘奇，刘洋，孙茂松. URL 模式与 HTML 结构相结合的平行网页获取方法[J]. 中文信息学报，2013，27(3)，91-99.

[13] Liu L, Hong Y, Lu J, Lang J, Ji H, &Yao J. An iterative link-based method for parallel web page mining. [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014: 1216-1224.

[14] Brin S, Page L. The anatomy of alarge-scale hypertextual web search engine[J]. Computer networks and ISDN systems, 1998, 30(1), 107-117.

章成志(1977—),博士,教授,主要研究领域为信息检索、科技文本挖掘等。
E-mail: zhangcz@njust.edu.cn

揭春雨(1964—),通信作者,博士,副教授,主要研究领域为计算语言学、机器翻译、文本挖掘、计算术语学等。
E-mail: ctckit@cityu.edu.hk

马舒天(1992—),博士研究生,主要研究领域为平行网页发现、多语言文本聚类。
E-mail: mashutian0608@hotmail.com

(上接第 90 页)

[16] Kong L, Dyer C, Noah A. Segmental recurrent neural networks[C]//Proceedings of ICLR, 2016.

[17] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of Workshop at ICLR, 2013.

[18] Lai S, Liu K, He S, et al. How to generate a good word embedding [J]. IEEE Intelligent Systems, 2016, 31(6): 5-14.

王蕾(1992—),硕士研究生,主要研究领域为自然语言处理。
E-mail: wangleinlp@163.com

周俊生(1972—),博士,教授,主要研究领域为自然语言处理、人工智能。
E-mail: zhoujs@njnu.edu.cn

谢云(1993—),硕士研究生,主要研究领域为自然语言处理。
E-mail: 1316480114@qq.com