URL 模式与 HTML 结构相结合的平行网页获取方法

刘奇，刘洋，孙茂松

（清华大学 计算机科学与技术系 智能技术与系统国家重点实验室，北京 100084）

摘要： 平行语料库是对机器翻译、跨语言信息检索等应用技术具有重要支撑作用的基础数据资源。虽然互联网上的平行网页数量巨大且持续增长，但由于平行网站的异构性和复杂性，如何快速自动获取高质量的平行网页进而构造平行语料库仍然是巨大的挑战。本文提出了一种 URL 模式与 HTML 结构相结合的平行网页获取方法，首先利用 HTML 结构实现平行网页的递归访问，其次使用 URL 模式优化遍历平行网站的拓扑顺序，从而实现高效准确的平行网页获取。在联合国与香港政府[1]两个平行网站上的实验表明，我们的方法相对传统获取方法在获取时间上减少 50%以上，准确率提高 15%，并显著提高了机器翻译的质量（BLEU 值分别提高 1.6 和 0.7 个百分点）。

关键词： 平行网页获取；平行语料库；URL 模式；HTML 结构

A Parallel Pages Mining Approach: Combining URL Patterns and HTML Structures

Qi Liu, Yang Liu and Maosong Sun

(Department of Computer Science and Technology, State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

Abstract: Parallel corpus are important supporting basic resource for statistical machine translation, cross-lingual information retrieval and others information processing technologies. Although the amount of parallel data on the web is continually increasing, the heterogeneity and complexity of parallel website make it still a large challenge to build parallel corpus by automatically extracting parallel text. This paper presents a new parallel web pages mining approach, which combines URL patterns and HTML structure together. First, we use HTML structure to recursively visit parallel pages. Then, URL patterns are used to optimize the traverse sequence of parallel web site topology. Thus we realized a both efficient and accurate parallel pages mining system. Compared with traditional approach, experiments on two parallel web sites(www.un.org and www.gov.hk[1]) show that this approach reduces more than 50% processing time, improves 15% accuracy and significantly improves the translation quality of MT System(BLEU score achieved 1.6 and 0.7 percent point increase, respectively).

Key words: Parallel pages mining; Parallel corpus; URL pattern; HTML structure

1 引言

平行语料库是平行句对的集合。所谓平行句对，是指互为翻译的两种语言的句子对。例如，下面是一个中文和英文的平行句对：

中国建筑业对外开放呈现新格局。

The Opening of China's construction industry to the outside presents a new structure.

平行语料库是对机器翻译和跨语言信息检索等应用技术具有重要支撑作用的基础数据资源。以统计机器翻译为例，无论是基于短语的方法[1]、基于层次短语的方法[2]，还是基于句法的方法[3]，都依赖于平行语料库获取翻译知识。因此，平行语料库的质量对于机器翻译的质量具有直接的影响[4]。

[1]香港政府网站包括 39 个不同机构的子网站，这里统称为香港政府网站。

随着互联网的蓬勃发展，包含大量平行网页的平行网站成为平行语料库构建的重要来源。所谓平行网页，是指相互翻译的两种语言的网页对，而平行网站就是包含大量平行网页的网站。例如，联合国网站(www.un.org)是一个典型的平行网站，包含中文、英文、法文、西班牙文等多个版本。不同语言版本的网页内容也往往相互对应。

因此，国内外研究人员相继提出一系列基于互联网的平行语料库挖掘方法([5]，[6]，[7]，[8]，[9]，[10]，[11]，[12])。虽然这些方法差异较大，但大体可分为两个步骤：

1. 发现平行网页；
2. 从平行网页中抽取平行句对。

由于互联网平行网站的异构性与复杂性，第一个步骤相对而言更为重要和困难，相应的研究工作大体可以分为两类：

1. 基于 URL 模式的方法；
2. 基于 HTML 结构的方法。

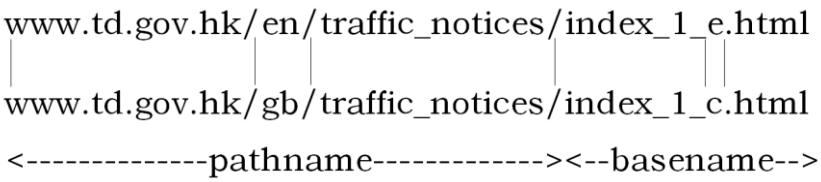

图 1：平行 URL 示例

基于URL模式的方法主张利用URL的命名规律来发现平行网页。图1给出了来自香港政府网站的一个英文网页的URL和一个中文网页的URL。 一般而言，URL 可以分为路径名（pathname）和基本名（basename）两部分。通过观察可以发现，两个URL的差异有两处：路径名的”en”'和”gb” 和基本名的’e’和’c’。如果大量的平行网页的URL都呈现这种差异模式，便有可能使用这种URL模式来发现平行URL对，进而定位平行网页。

早期的URL模式主要依赖人工预定义([7]，[10])，近年来一些研究人员提出自动发现URL模式的方法([11]，[12])，大幅提高了URL模式发现的覆盖率。基于URL模式的方法忽略平行网站的内部结构，直接通过发现命名符合模式的URL来定位平行网页。然而，正如[9]所指出，基于URL模式的方法需要下载平行网站的全部URL，获取时间长、效率低；另外平行网站的命名方式千差万别，仅仅依赖于预定义的URL模式难以确保平行网页的覆盖度和质量。而自动发现URL模式的方法，也需要首先下载全站网页、区分语种，然后在所有可能的URL对中计算模式，不仅浪费了大量带宽下载无关网页，而且全枚举计算URL模式计算和存贮的开销都很高，难以在大型网站中应用。

基于HTML结构的方法主张利用HTML的结构信息来发现平行网页。如图2，以联合国网站中文首页和英文首页作为入口，利用HTML的结构信息发现该网页对可能包含的指向其他平行网页的超链接，如“和平与安全”与”Peace and Security”以及“发展”和”Development”等。接下来访问“和平与安全”与”Peace and Security”对应的平行网页，又可以发现新的超链接对。这样通过不断迭代访问来发现更多的平行网页。由于网站中的所有网页通过超链接构成一个巨大的有向图（节点是网页，边是超链接），基于HTML结构的方法实际上通过同步访问两个有向图来实现平行网页发现。由于基于HTML结构的方法能够充分利用网页内容信息，相对于基于URL 模式的方法更容易找到高质量的平行网页[9]。然而，即使是同一个平行网站，中文网站与英文网站也不完全是同构的，即存在较大数量的非平行网页。基于HTML结构的方法可能处理大量的候选网页对却只发现少量的平行网页对。更严重的是，处理网页的计算量远大于处理URL，所产生的候选网页对数量也将持续膨胀，因此基于HTML结构的方法几乎无法完全处理平行网站的所有候选网页对就被迫中止。

图2 平行网页实例

针对上述挑战，本文提出一种URL模式与HTML结构相结合的平行网页发现方法。该方法利用HTML结构实现双语网页有向图的同步访问，通过动态生成带频度的URL 模式对未来待访问的节点对进行排序，从而实现遍历拓扑顺序的优化，尽可能避免访问非平行网页。我们进一步通过提前终止（early stopping）技术判定当前是否抽取完毕所有平行网页，从而大幅减少平行网页发现进程的运行时间。在联合国与香港政府两个平行网站上的实验表明，我们的方法相对只使用HTML结构的方法在获取时间上减少了50%以上，准确率提高15%，并显著提高了机器翻译的质量（BLEU 值分别提高1.6 和0.7 个百分点）。

2 URL 模式与 HTML 结构相结合的平行网页获取方法

我们的方法概述如下：如果将网站内的网页作为节点，网页间存在的链接关系作为边，双语网站的不同语种的网页集合就组成了两个树形拓扑结构，且结构间存在相似性。

给定双语网页的两个单语首页作为种子，从种子节点出发，利用页面 HTML 结构序列对齐的技术获取种子节点包含的双语文本以及指向的下级候选平行网页 URL 对。利用分类器对候选平行页面对进行验证，对确实为平行网页的网页对，用与对种子节点相同的处理方法获取双语文本和下级候选平行网页 URL 对，将下级候选平行网页 URL 对放入队列，同时对 URL 所对应的命名模式进行学习。随着发现进程运行，系统能够学习到多次匹配已验证平行网页 URL 对的模式，对其中频度超过一定阈值的 URL 模式我们称为可信赖模式。

利用可信赖 URL 模式频度对候选平行网页 URL 对组成的队列排序，即可形成一个基于 URL 模式的候选平行网页 URL 对优先队列。从而实现对符合可信赖 URL 模式的候选平行网页提前获取。随着发现进程进行，URL 模式的频度不断更新，并不断学习到新的模式，最后队列中符合可信赖 URL 模式的 URL 对数量降至为零。此时不能判定队列中已不存在平行页面对，一个基于 URL 命名模式发现的提前终止检测模块被调用，以此来做出是否停止发现进程的判断，以避免对大量的非平行页面进行处理。

下面将分别介绍递归获取流程、URL 模式优先队列和提前终止检查模块。

2.1 递归获取流程

URL 模式和 HTML 结构结合的平行网页获取机制包含相互结合的两个部分：递归获取流程及模式学习过程和基于 URL 模式的候选平行网页 URL 对优先队列构造过程。两部分同时进行，相互结合，可分为四步，如图 3：

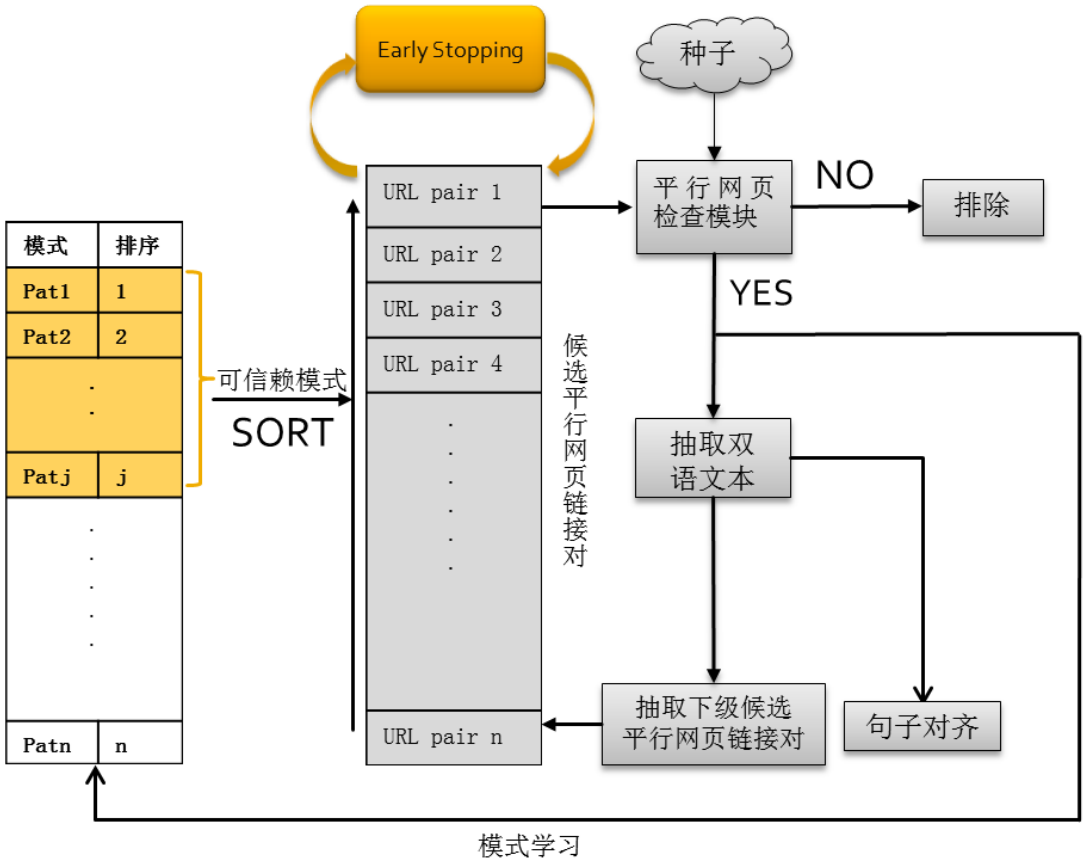

图 3：基于 URL 模式优先队列的平行网页发现机制

1. 给定网站的中英文首页作为种子平行页面对，下载种子页面，由于种子页面是先验的平行页面，可以通过平行页面检查模块。利用基于HTML 标记序列对齐的方法从中抽取双语文本，并获取种子页面的下级候选平行网页URL对，存入队列。
2. 对给定的候选平行网页URL对，下载对应网页，调用检查模块来判断两个网页的内容是否具有翻译等价性和HTML结构相似性。如果验证为非平行页面对，则不作处理。
3. 对验证为平行的网页对，使用基于HTML标记序列对齐的方法获取双语文本和下级候选平行网页URL对。调用句子对齐模块对双语文本进行对齐，并将对齐句子存入平行语料库。同时调用URL模式学习模块对中英文网页URL所对应的命名模式进行学习，如果为新发现的模式，则该模式频度为1，反之如果是已经学习到的模式，则模式频度加1。
4. 定期利用可信赖模式对候选平行网页URL队列排序，构成平行网页URL对优先队列，获取排序第一的URL对，回到第二步，如此循环进行。如果符合可信赖模式的候选平行网页URL 数量为零，调用提前终止检查模块，若发现队列中不可能存在平行页面，则系统停止并退出。

2.2 URL 模式优先队列

	pathname		basename
	主机	路径	文件
域	www.td.gov.hk/	en/traffic_notices/	index_1_e.html
token	www.td.gov.hk	en,traffic_notices	index,1,e,html

表 1：URL 字符串中的域及 token

URL命名模式为URL中路径名和基本名两部分模式的结合。在路径名中使用"/"进行分割，

而在基本名中我们使用{._=&-:?}进行分割，如表1所示，URL经过分割后URL由若干token组成。而由图1 可知，两个URL之间的命名模式为路径名中的token “zh”转换为”en”和基本名中的’c’转换为’e’。模式可形式化定义如下：

$$p = \{pn, bn, w\} \tag{1}$$

$$pn = \{tc_1 \rightarrow te_1, \cdots \quad te_n\} \tag{2}$$

$$bn = \{tc_1 \rightarrow te_1, \cdots \quad te_n\} \tag{3}$$

其中，p 是URL模式，pn 表示路径名所对应的模式，bn 表示基本名所对应的模式，w 是模式的频度，$tc_i \rightarrow te_i$ 表示使用 tc_i 替换 te_i 。tc_i 与 te_i 的不同出现对应着三种命名变换动作，如表2所示。一个URL可以在URL命名模式的作用下变换为另一个URL。

Tc	Te	动作
非空	非空	Tc替换te
非空	空	中文URL中删除对应的tc
空	非空	英文URL中删除对应的te

表2：URL模式代表三种动作

随着系统发现越来越多的平行网页，这些平行网页URL 所对应的模式频度将增加并超过某个阈值（我们设为20），对于超过阈值的URL模式，我们认为模式为可信赖模式。符合可信赖模式的所有候选平行网页URL对不需要使用平行页面验证模块即可直接认定为平行网页。同时利用可信赖模式对队列中的候选平行网页URL对进行排序，符合可信赖模式且模式频度大的排序靠前，不符合的排序靠后，以此构造出基于URL 模式的候选平行网页URL对优先队列，我们简称URL模式优先队列。

3 提前终止检查模块

平行网页对发现系统的基本出发点是：

1. 从种子节点出发可以遍历整个网站的所有网页，候选平行网页 URL 对可以从上级平行网页对中获取；
2. 平行网页间存在着 URL 命名模式；
3. 不在非平行网页中抽取可能的平行网页 URL 对。

如何找到一个恰当的停止点对提高平行网页发现系统速度和效率至关重要，最理想的结果是：在停止点之前获得了所有可能的平行网页，而在这个点之后候选平行网页 URL 队列中不存在平行网页。对于普通队列发现机制，随着发现进程进行，队列中的候选 URL 对越来越多，其中包含的平行网页对的密度将逐步降低，但系统停止的前提是队列为空，因此将经历一个较为漫长的长尾过程。而本方法的优先队列遍历机制，利用可信赖模式定期对候选平行页面 URL 对队列进行排序，将符合可信赖 URL 模式的待验证平行网页对排序靠前。随着过程的进行，符合可信赖 URL 模式的候选平行页面对数量将减少为 0，这为系统自动检测停止条件并实现提前终止提供了进入点。进入检查模块时，虽然队列中符合可信赖模式的候选数量为 0，但不能确定不存在平行页面，需要对可能存在的模式进行发现，并最终确定是否应当停止。提前终止检测模块流程如下：

1. 遍历候选平行页面 URL 对队列，对所有 URL 对对应的命名模式 p 进行频度计算，

频度计入模式的 w 频度域，对出现频度大于等于 2 的所有模式的总频度相加，得到一系列 URL 命名模式及使用这些模式命名的 URL 对的数量，记为 count。

$$count = \sum_{i=0}^{n} (w(p_i)) \tag{4}$$

2. 如果 count>0，则使用所有计算得到频度大于等于 2 的 URL 模式对队列进行排序，返回 false，信号量 shoud_stop 置为 0。

3. 如果 count=0，则 should_stop 值加 1。 由于进程定期（每处理一定数量页面对后）对 URL 队列进行排序，因此后续将再次遇到提前终止检查，如果 count 仍然为 0，则 should_stop 值会增加，当 should_stop 值累加到阈值（如 5）时，返回 ture。判定系统此时到达了停止点，认为后续所有 URL 对中不存在平行页面对，所有系统进程停止抓取。

4 平行页面验证模块

对候选平行网页 URL 对，需要下载所对应的网页，调用平行网页验证模块来对该网页对是否平行进行验证。这里我们借鉴了[13]和[7]中的方法，两种方法中使用了三个特征：(1) 中英文文档长度比，(2) HTML tag 相似度，(3)句子对齐分数。由于句子对齐工具假定的前提是两篇文章是相互翻译的，因此两篇不相互对译但主题相似的文章也能利用句子对齐工具取得一定数量的对齐句子，另外运行句子对齐工具的计算开销较大，对长文本文档处理较慢，因此用句子对齐分数评价存在一定的限制因素。本文采用线性分类器并利用三种特征进行分类：

1.长度比：$\dfrac{len(c)}{len(e)}$ (5)

2.HTML 结构相似度：$\dfrac{com(seq(c), seq(e))}{len(seq(e))}$ (6)

3.内容翻译等价性：$\dfrac{\sum_{i=1}^{n} \min(wd_i(e), \sum_{j=1}^{m} chiwd_j(wd_i(e)))}{words(e)}$ (7)

长度比定义为页面内中文字符串 c 长度占英文页面内英文字符串 e 长度的比率。HTML 结构相似度计算如下：将 HTML 的 tag 标记树序列化，将所有的文本节点标记为#text，然后使用类似于 UNIX Diff 的对齐算法得到两个 HTML 文件 tag 标记序列 seq(c),seq(e)的差异，除去差异，匹配成功的部分长度除以总长度。内容翻译等价性计算如下，英文页面中的英文词 $wd_i(e)$ 在词典中对应所有中文词 $chiwd_j(wd_i(e))$ 在中文页面中出现次数之和，与 $wd_i(e)$ 在英文页面中出现次数取较小数，相加得到对应中文词出现的总次数，除以所有英文词出现总数，即得到内容翻译等价性评分。这里需要对中文进行分词。

我们使用了 600 对随机挑选自香港政府及其 8 个子网站和联合国网站的中英文平行网页对线性分类器 LibLinear[14]进行训练，这些网页对都是人工判定的平行网页对。

5 双语文本和下级平行网页 URL 对抽取

平行网页挖掘系统的核心目的在于取得平行网页中包含的高质量双语语料，对于已经验证的平行网页对，我们使用 HTML tag 标记序列对齐来获得双语文本。平行网页具有相互对译的内容以及相似的页面结构，大部分平行页面的 HTML 页面结构从 tag 标记树来说具有很高的相似性。本方法对 HTML 页面的 tag 标记树进行序列化，对文本内容的节点，我们将文本内容本身视为 tag，使用 Diff 算法对齐，差异部分将能得到可能对译的双语文本。如将联合国中英文首页的 tag 标记序列处理后对齐，得到的片段形如下图 4:

```
30,30 2 28,28
<Skip to resources
—
>跳转到相关资源
36,36 2 34,34
<Welcome to the United Nations. It's your world.
—
>欢迎来到联合国，您的世界！
```

图 4 HTML 标记序列对齐获得双语文本

```
398,398 2 383,383
<http://www.un.org/zh/siteindex/
—
>http://www.un.org/en/siteindex/
402,402 2 387,387
<http://www.un.org/zh/aboutun/
—
>http://www.un.org/en/aboutun/
410,410 2 395,395
<http://www.un.org/zh/contactus/index.jsp
—
>http://www.un.org/en/contactus/index.jsp
```

图 5：HTML 标记序列对齐获得下级 URL 对

为了让系统递归的深入下一级网页进行平行网页发现，必须从已经确认的平行网页中抽取下级候选平行网页 URL 对。与获取双语文本相同，我们再次使用 HTML tag 标记序列对齐的技术，将文本节点统一标记为#text，以排除干扰，将 URL 链接替换对应 tag 标记，使用 Diff 算法对齐，就可以获取候选平行网页 URL 对，如将联合国中英文首页的 tag 序列对齐，得到的片段对如图 5 所示：

6 实验

我们的实验系统在联合国网站和香港政府网站上运行（香港政府网站包含 39 个不同机构的子网站，以下均简称为香港政府网站）。我们按照[9]Shi et al. (2006)的机制搭建了不使用 URL 模式优先队列的普通队列平行网页发现系统作为基准系统[2]。为了节约篇幅，下面将基准系统简称为**普通队列**系统，我们的系统简称为**优先队列**系统。

优先队列系统及普通队列系统均同步在 www.un.org 和 www.gov.hk 及其子网站进行运行。由于联合国网站较小，优先队列系统实现了智能检测并提前终止，并未处理全部网页，学习到可信赖 URL 命名规则 15 条，而普通队列系统实现了全部处理。由于香港政府网站规模较大，耗时较长，优先队列系统和普通队列系统分别进行了 96 小时和 192 小时后强制中断，优先队列系统学习到可信赖 URL 命名规则 39 条（如表 3 所示）。相关统计数据如表 4 所示：

www.gov.hk	www.un.org
{{“sc”->”en”}, ,15405}	{{“zh”->”en”}, ,2234}
{{“sc.lcsd.gov.hk/gb”->””,”b5”->”en”}, ,8077}	{{“zh/documents”->”ga/search”}, ,792}
{{“sc_chi”->”english”}, ,4561}	{{“chinese”->””}, ,712}
{{“gb_chi”->”eng”}, , 4398}	{{“zh”->””}, ,133}
{{“sc.epd.gov.hk/gb”->””},{“2”->”1”} ,1444}	{{“chinese”->”en”}, ,80}

表 3：学习到的 URL 模式（前五）

[2]为了两个系统公平比较平行页面发现算法的优劣，普通队列系统的实现与[9]在细节略有不同：使用本文中介绍的基于 HTML 结构标记的对齐模块获取下级平行网页 URL 对，并采用本文分类器使用的特征。

网站	方法	处理网页对数	判定平行网页对数	比率
www.un.org	普通队列	21,145	4,735	4.466
	优先队列	5,233	4,372	1.197
www.gov.hk	普通队列	116,121	47,045	2.468
	优先队列	57,923	48,045	1.206

表 4：获取平行网页数量对比

从以上数据我们可以看出，优先队列系统的获取单位平行网页所需处理的网页数要远小于普通队列系统。基于以上数据，我们对平行网页发现的准确率、效率和获取的双语质量进行了比较。首先我们对两个系统在不同网站上挖掘到的平行网页对的准确率进行比较。400 对平行网页随机从两个系统得到的所有平行网页对中挑出，人工判断是否为真正的平行网页。人工评判的准确度数据表明，优先队列系统准确度高达 98%以上，如表 5 所示。

对普通队列的挖掘系统，平行网页对的准确率完全依赖于平行网页对验证模块的准确度，由于双语网站中存在着大量的具有相同主题的非平行网页(如联合国安理会这个主题下就有安理会组织，新闻，热点等众多网页),分类器对这样的主题相同，网页结构也相同但文章不对译的网页对的准确度有限，因此降低了总体的分类器准确度。而由于结合了 URL 模式，利用可信赖模式直接对网页对判定，就直接减少了分类器分类错误可能导致的下级候选平行网页 URL 对无法获取的问题。直接帮助优先队列系统的发现覆盖度提高。另外，使用模式构成优先队列，对符合可信赖模式的页面对优先处理，系统每获取一个平行网页对所需处理的网页对数量大幅下降，效率大幅提高；而使用提前终止模块，避免系统在最后阶段检查大量的不可能存在平行页面的队列长尾，节省了系统处理的时间。

在联合国网站 www.un.org 上实现了平行页面全站发现，效率对比如图 6:

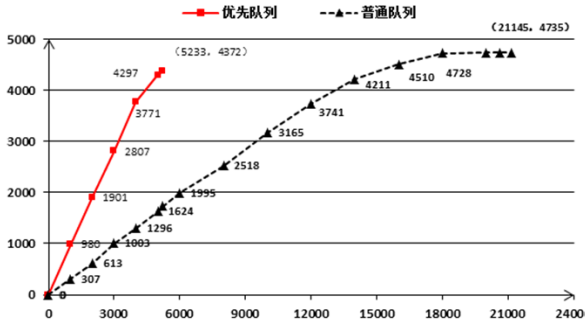

图 6：www.un.org 平行网页获取效率对比

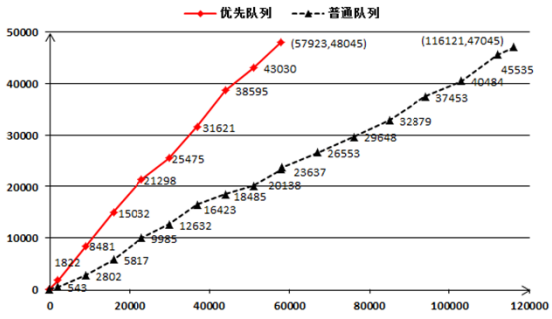

图 7：www.gov.hk 平行网页获取效率对比

在香港政府网站 www.gov.hk 上的实验在两种队列系统下分别进行，基于优先队列和普通对了的发现进行分别执行了 96 小时和 192 小时停止，分别获得 48045 和 47045 个平行页面对。两个系统的效率对比如图 7 所示，从图中可以看出，香港政府网站未能进行全站发现，普通队列发现机制的长尾效应未能体现。

我们对系统获取的平行网页中抽取双语文本，使用 Fast-Champollion[15]进行句子对齐，对获取的平行句对去重，得到了语料库，其统计数量如表 6 所示：

网站	方法	判定平行网页对数	准确度	平行网页数量
www.un.org	普通队列	4,735	82%	4,735
	优先队列	4,372	98%	4,372
www.gov.hk	普通队列	47,045	84%	47,045
	优先队列	48,045	99%	48,045

表 5：获取平行网页准确度对比

网站	方法	句对数量
www.un.org	普通队列	49,975
	优先队列	45,642
www.gov.hk	普通队列	713,766
	优先队列	695,521

表 6：获取双语句对（去重）对比

为了衡量所获得双语语料的质量，我们分别在从两个网站获得的语料上使用 MOSES 训练统计机器翻译模型，并分别使用 NIST02 和 NIST05 的标准测试集作为模型的开发集和测试集。语料库处理中滤除了中英文句子词数大于 100 的句子，测得 BLEU 值结果如表 7 所示。需要强调的是，由于联合国和香港政府网站文本内容领域与测试集相差较大，因此整体 BLEU 值相对较低。

网站	方法	句对数	中英词数	BLEU
www.un.org	普通队列	45,642	1,744,557	0.1129
	优先队列	45,642	1,566,593	**0.1291**
www.gov.hk	普通队列	705,489	22,017,223	0.1608
	优先队列	684,898	22,199,657	**0.1680**

表 7：语料 BLEU 分数对比

为了进一步检验优先队列系统获得的语料比普通队列系统获得语料质量更好，我们使用了由美国卡内基梅隆大学开发的显著性检验工具 paired_bootstrap_v13a 计算统计显著性[3]。结果都显示：构建于优先队列系统获得的双语语料上的统计机器翻译系统在显著性水平 0.01 及 0.05 上都是显著好于构建于普通队列获得的双语语料上的统计机器翻译系统[4]。

7 总结

由于互联网提供的双语资源具有内容丰富、领域平衡、总量大的特点，从互联网获取平行语料库来缓解机器翻译和跨语言信息检索系统所面临的高质量、领域平衡语料库匮乏问题是一个重要的研究方向，而其中首先需要解决的挑战是如何定位平行网页。本文方法将基于 URL 模式与基于 HTML 结构的方法有效结合起来，既充分利用 HTML 结构实现平行网站同步迭代访问，又通过 URL 命名规律避免访问非平行网页，实现优势互补。在联合国网站和香港政府网站上的实验数据表明，本方法对获取的双语平行网页对准确度达到 98%以上。所提出的提前终止技术能够选择恰当的停止点，避免系统处理大量无效网页对，这样不仅可以获得高质量（通过机器翻译的 BLEU 值分别提高 1.6 和 0.7 个百分点得到验证）平行网页双语语料，同时少处理 75% 的无效网页对，显著节省了带宽和流量，从而实现高质量、高效、高覆盖度的平行网页获取。

致谢

[3] http://www.ark.cs.cmu.edu/MT/，该工具使用[16]的方法。

[4] 两个网站上本方法系统与 baseline 系统的显著性测试结果都为：1000 次迭代中所有迭代的测试都是本方法的系统好于 baseline 系统。

本文受国家科技支撑计划课题(批准号：2009BAH41B04)和国家自然科学基金项目(批准号：60903138)及国家 863 计划项目(批准号：2012AA011102)资助。

参考文献

[1] Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proceedings of NAACL 2003.

[2] David Chiang. 2007. Hierarchical phrase-based translation. Computational Linguistics.

[3] Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of contextrich syntactic translation models. In Proceedings of COLING-ACL 2006.

[4] Dragos Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting nonparallel corpora. Computational Linguistics.

[5] Xiaoyi Ma and Mark Liberman. 1999. Bits: A method for bilingual text search over the web. In Proceedings of MT Summit 1999.

[6] Jiang Chen and Jian-Yun Nie. 2000. Automatic construction of parallel english-chinese corpus for crosslingual information retrieval. In Proceedings of ANLC 2000.

[7] Philip Resnik and Noah Smith. 2003. The web as a parallel corpus. Computational Linguistics.

[8] Jisong Chen, Rowena Chau, and Chung-Hsing Yeh. 2000. Discovering parallel text from the world wide web. In Proceedings of DMWI 2004.

[9] Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A dom tree alignment model for mining parallel data from the web. In Proceedings of COLING/ACL 2006.

[10] Ying Zhang, KeWu, Jianfeng Gao, and Phil Vines. 2006. Automatic acquisition of chinese-english parallel corpus. In Proceedings of ECIR 2006.

[11] Chunyu Kit and Jessica Yee Ha Ng. 2007. An intelligent web agent to mine bilingual parallel pages via automatic discovery of url pairing patterns. In Proceedings of Web Intelligence/IAT Workshops 2007.

[12] Shani Ye, Yajuan Lv, Yun Huang, and Qun Liu. 2008. Automatic parallel sentences extracting from web. Journal of Chinese Information Processing.

[13] Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. 1999. Cross-language

[14] R Fan, K Chang, C Hsieh, XWang, and C Lin. 2008. Liblinear: A library for large linear classification. Journal of Machine Learning Research.

[15] Peng Li, Maosong Sun, and Ping Xue. 2010. Fastchampollion: a fast and robust sentence alignment algorithm. In Proceedings of COLING 2010.

[16] Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Proceedings of EMNLP 2004. information retrieval based on parallel texts and automatic mining of parallel texts from the web. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

作者联系方式：姓名　地址 邮编　电话（最好手机） 电子邮箱
刘奇 北京清华大学 FIT 楼 4-506　100084　13466338393　flaminglq@gmail.com