• 开发与应用 •

基于统计模型的主题爬虫的研究与实现

金明珠， 丁岳伟

(上海理工大学 光电信息计算机工程学院，上海 200093)

摘 要：在研究了现存的主题爬虫的基础上，提出了一种基于统计模型的主题爬虫，它对抓取过程中可获得的信息进行分析，并运用统计模型计算的结果过滤URL，有效地解决了偏好特定主题的用户检索和Web信息的索引等相关问题。实验结果表明，与基于链接和网页内容分析的主题爬虫相比，该主题爬虫能够在检索较少的网页时，抓取到较多的与主题相关的网页，提高了抓取精度。

关键词：统计模型；主题爬虫；URL过滤；特征信息；字段

中图法分类号：TP311 **文献标识码**：A **文章编号**：1000-7024 (2010) 16-3700-05

Research and implementation for topic crawler using statistic model

JIN Ming-zhu, DING Yue-wei

(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Based on the analysis of the existed topic crawlers, another topic crawler using statistic model is proposed, which analyzes the information available and filters the URLs by using the results calculated from the statistic model during crawling, and to effectively addresses the problem how to index the mass web resource and how to find specific topic pages fit user's interest. The experimental results show that, compared with topic crawlers based on linkage and page content analysis, topic crawler using statistical model can fetch more topic relevant web pages by retrieving less web pages, and improve the crawling accuracy.

Key words: statistic model; topic crawler; URL filtering; feature information; term

0 引 言

随着万维网中信息指数增长，有限的网络和存储资源给通用网络爬虫和搜索引擎带来了更多的挑战，主题爬虫的目标是遍历Web的一个子集，抓取其中和特定主题相关的网页，筛选网页中潜在的与主题资源相关的前向链接，避免进入偏离主题的网页分枝。

本文研究了一种基于统计模型的主题爬虫，主题爬虫根据当前网页中的前向链接的主题相关度对其进行过滤，使爬虫聚焦在一个特定主题相关的Web子集中。它是针对前向链接的主题相关度，利用人工建立的主题文档集合，对抓取过程中所得的相关特征信息(如网页内容，URL字段等主题相关性)进行分析，构建统计模型进行估值计算的方法。

1 相关研究

主题爬虫的主要目标是以较好的方式，高效地抓取Web中与主题相关的网页。它与传统的通用搜索引擎相比，减少了对资源的利用且支持扩张性的检索处理。对于主题爬虫而言，最重要的是如何过滤网页中的前向链接，使得爬虫聚焦在一个特定主题的Web子集中。根据Pant和Srinivasan在文献[1]中所描述，大多数主题爬虫在抓取网页时使用了不同优先级策略的Best-First算法，维护一个未被访问的URL的优先级队列，每个URL都有一个相关联的权值表示其优先级，它决定了爬虫下一个将要选择抓取的网页。文献[2]中提及的Naïve Best-First爬虫，它使用了基于页面内容分析的策略，通过余弦相似度算法计算当前页面的主题相关度，进一步估算页面中前向链接的主题相关度，并由此设定前向链接的优先级。Guilherme和Alberto在文献[3]中描述了利用文档风格和网页内容进行分析，估算页面相关度，并根据此相关度更新URL队列中当前页面的兄弟链接的优先级(即链接相关度)。文献[4]中提出了一种不同的主题爬虫，它利用HMM (hidden Markov model)和CRF(conditional random fields)两种不同的概率模型对链接结构和网页内容建模，通过分析用户浏览特定主题网页的方式和内容，估计链接的主题相关性。但是网页内容的主题交错性和相关页面中可能包含偏离主题的前向链接，这种基于页面内容估算链接主题相关度的算法在抓取主题相关页面时具有一定的偏离性，降低了全局网页数据存储中主题相关网页的比例。另外，一部分主题爬虫为了节省资源，单纯的

收稿日期：2009-08-11；修订日期：2009-10-15。

作者简介：金明珠 (1985－)，男，江苏泰州人，硕士研究生，研究方向为软件工程及知识挖掘； 丁岳伟 (1964－)，男，上海人，教授，硕士生导师，研究方向为信息安全、软件工程及知识挖掘。E-mail：dyuewei@hotmail.com

对URL本身进行分析，估算其主题相关度。文献[5]中描述了一种基于机器学习的URL分析方法，它从URL中抽取出特征向量，并对其运用支持向量机算法，计算主题相关度。Jun Li提出了一种利用决策树对超链接文本分析的算法[6]，通过提取超链接文本和超链接周围的字段，基于已建立的训练集，构建决策树，估算URL的主题相关度。上述的利用URL或者超链接文本进行分析的算法中，由于网页中存在超链接广告，以及超链接文本和URL中含有的信息量较少，不能从整体上对前向链接的主题相关度进行分析，因此，分析结果的误差相对偏大，会在很大程度上降低主题相关网页的抓取准确率。

2 统计模型和系统结构

我们研究了基于不同策略过滤URL的主题爬虫，它们的一个共同缺陷就是在URL的主题相关性的判断上不够全面。因此，我们在本章提出了一种基于统计模型的主题爬虫系统，它综合地考虑爬虫检索过程中可获得的各种可能与URL主题相关性的信息，全面地判断URL的主题相关性。

2.1 统计模型

主题爬虫在抓取主题相关的网页时，必须对抓取网页中的URLs进行过滤，并将主题相关的URLs添加到所要抓取的URL列表中，以便保证始终能够聚焦在某一特定主题上。在对URLs进行过滤时，主题爬虫通过分析在抓取过程中所能获得的信息，对URL的主题相关度进行估算，并利用估算值进行过滤。在文献[7]中提出了主题抓取过程中的两个重要特性：

(1)主题局域性：主题相关的网页中的前向链接所引网页更可能是主题相关的。

(2)兄弟局域性：若一个网页所引的某个网页是主题相关的，则它所引的其它网页也可能是主题相关的。

根据以上描述的两个特性和先前的相关研究，检索过程中所获得的相关特征信息包含如下几个方面：①网页内容；②兄弟URL；③URL字段；④超链接文本。

在基于统计模型的主题爬虫中，我们提出了一种结合上述相关特征信息对URL主题相关性进行分析的方法。在该方法中，通过对各个相关特征信息的分析，得出其主题相关性，构建统计模型，并对URL的主题相关度进行估算。为了形式化的描述URL主题相关度估算的统计模型，我们首先给出如下相关定义。

定义1 C表示某个URL所指向的网页的主题相关性。

定义2 P(C)表示URL所指向的网页是主题相关的可能性(即URL的主题相关度)。

定义3 E_i表示某个相关特征信息。

定义4 $R(E_i)$表示由某个相关特征信息推断出的URL的主题相关性。

在基于统计模型的主题爬虫中URL的主题相关性是根据特征信息的主题相关性进行估算的，公式(1)描述了它们之间的关系。

$$\bigcup_{i}^{n} E_i \Rightarrow C \tag{1}$$

由上述的推导关系，构建出对应于URL主题相关度的统计模型如公式(2)所示。

$$P(C)=\sum_{i=1}^{n} w_i * R(E_i) \tag{2}$$

式中：w_i——E_i对URL主题相关度的影响因子，它是由不同的特征信息对URL主题相关度的影响程度决定的，对于所有的w_i，它们满足公式(3)所描述的限制关系。

$$\sum_{i=1}^{n} w_i = 1 \tag{3}$$

2.2 系统结构

基于统计模型的主题爬虫系统主要由主题爬虫、特征信息分析模块、主题文档集合、主题字典库、网页集合存储5个部分组成，其中主题爬虫主要负责抓取网页和过滤链接的工作。特征信息分析模块是由网页内容分析、兄弟URL分析、URL字段分析、超链接文本分析4个小模块所组成。主题文档集合是特定领域的专家从一个基于网页内容分析的主题爬虫所抓取的网页中筛选后建立的。主题字典库是利用一种半自动的字段生成算法构建的[3]。网页集合存储则用于存储主题相关的网页。基于统计模型的主题爬虫系统的总体结构，以及各部分之间的交互如图1所示。

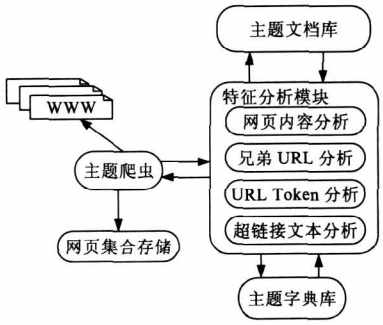

图1 系统结构图描述

3 特征信息分析

特征信息分析模块主要是对网页中的URL对应的特征信息进行分析，并将各个特征信息的分析结果返回给主题爬虫。本章将分析不同的特征信息对URL的主题相关性的估计方法。

3.1 页面内容分析

每个网页的主题特征是由页面中的关键字决定的。而在不同网页中的不同字段可能描述相同的概念。为了能够准确的判断页面内容的主题相关性，从全文的角度考虑，利用潜在语义索引[8-9](latent semantic indexing，LSI)对文档进行表述。在主题爬虫抓取到网页p后，计算p的LSI，利用余弦相似度函数(如公式(4)所示)对p和已建立LSI的主题文档集合计算它们之间的相似度。

$$cos(p,t)=\frac{\sum_{k\in p\cap t} w_{pk}\cdot w_{tk}}{\sqrt{\sum_{k\in p} w_{pk}^2 \cdot \sum_{k\in t} w_{tk}^2}} \tag{4}$$

式中：w_{dk}——在利用LSI算法计算后的文档d的向量字段k的权重，其计算公式如公式(5)所示。

$$w_{id}=\frac{tf_{id}(\log_2\frac{N}{df_{i\ 1}})}{\sqrt{\sum_{k=1}^{T}(tf_{id})^2(\log_2\frac{N}{df_i}+1)^2}} \tag{5}$$

在公式(5)中，tf_{id}表示字段 i 在文档 d 中频率，N 表示文档集合中的文档数，df_i表示包含字段 i 的文档数。

文中提到了主题局域性，若一个网页是主题相关的，则其所引用的网页也可能是主题相关的。因此，我们可以根据余弦相似度对网页p的主题相关性来估计网页中 URL 的主题相关性，给定一个阈值δ，若$cos(p,t)$的值大于等于δ，则说明网页中的 URL 也是主题相关的，反之，则说明是主题无关的。公式(6)描述了页面主题相关性$R(up)$的判定。

$$R(up)=\begin{cases}1 & cos(p,t)\geqslant\delta\\ 0 & cos(p,t)\leqslant\delta\end{cases} \tag{6}$$

式中：up——对应于 URL 的页面内容特征信息。

3.2 兄弟 URL 分析

在前文中，描述了在主题爬虫抓取网页时兄弟 URL 对应的网页存在一定的主题局域性，即对于任一个网页 p，它所引用的某个网页是主题相关的，则其所引用的其它网页也可能是主题相关的。因此，在分析网页 p 中的 URL 主题相关性时，需要考虑 URL 的兄弟所对应的网页的主题相关性，以提高 URL 主题相关性判定的准确率。一个网页可能被多个网页引用，故其兄弟页面对应的URL可能分散在多个引用网页(即父网页)中。因此，在对一个 URL 的兄弟 URL 进行分析时，必须对其所有兄弟 URL 进行分析，根据其中主题相关的网页所占比例对当前 URL 的主题相关性进行估计。定义 P 为网页 u 的父网页集合，其表达式如公式(7)。

$$P=\{p|(p,u)\in E\wedge p\in T\cup UT\} \tag{7}$$

在主题爬虫中，Web 被定义为一个图$G=\{E,V\}$，其中E表示网页之间的引用关系集合，V表示网页集合，T表示已抓取的主题相关网页集合，UT表示已抓取的主题不相关网页集合。由于无法确定未抓取的兄弟 URL 对应的页面的主题相关性，在计算兄弟 URL 对应的网页中主题相关的比例时，只考虑已抓取的网页。定义S为网页u的兄弟网页集合，其表达式如公式(8)所示。

$$S=\{s|(p,s)\in E\wedge p\in P\wedge s\in T\cup UT\} \tag{8}$$

故兄弟 URL 对应的网页中主题相关网页的比例$I(u)$计算表达式如公式(9)所示。

$$I(us)=\frac{|S\cap T|}{|S|} \tag{9}$$

而在整个 Web 中，特定主题相的网页在所有网页中是占有的比例一定的。因此，我们通过比较$I(u)$和所有已抓取的网页中主题相关网页的比例对 URL 的主题相关性$R(us)$进行判定，其判定形式如公式(10)所示。

$$R(us)=\begin{cases}1 & I(us)\geqslant\dfrac{|T|}{|T\cup UT|}\\ 0 & I(us)<\dfrac{|T|}{|T\cup UT|}\end{cases} \tag{10}$$

式中：us——对应于 URL 的兄弟 URL 特征信息。

3.3 URL 字段的分析

在根据URL字段对URL的主题相关性进行分析时，首先需要判断 URL 字段的主题相关性。我们通过对 URL 进行解析，去除 URL 中的无关字段和字符，如“http”，“www.”，“/”，“-”，“.com”等，从中提取有意义的字段f，构成一个字段集合F，利用 T 和主题字典库 TTD(topic terms dictionary)进行匹配，计算 F 的主题匹配度$M(F,TTD)$，其表达式如公式(11)所示。

$$M(F,TTD)=\frac{|F\cap TTD|}{|F|} \tag{11}$$

在此，给定一个阈值ε，通过对M和阈值ε的比较，对 URL 的主题相关性$R(ut)$进行估计，其估计函数如公式(12)所示。

$$R(ut)=\begin{cases}1 & M\geqslant\varepsilon\\ 0 & M<\varepsilon\end{cases} \tag{12}$$

式中：ut——URL 字段特征信息。

3.4 超链接文本分析

同样地，在根据超链接文本对 URL 的主题相关性进行分析时，从超链接文本中去除无用字段(如量词，连词等非名词字段)，提取出有意义的字段，构成字段集合F，使用和 URL 字段主题相关性分析的相同方法，对URL的主题相关性$R(ua)$进行估计，其中ua表示超链接文本特征信息。

4 主题爬虫

主题爬虫主要负责网页抓取和 URL 过滤，它从一个候选 URL 列表中按照优先级高低来选择 URL 抓取网页，并根据特征信息分析模块的分析结果和统计模型计算网页中的URL的主题相关度$P(C)$，利用$P(C)$和一个给定的阈值γ对比，过滤出主题相关度高的 URL 保存到候选 URL 列表中，其中 URL 对应的优先级是由$P(C)$决定的。下面给出了主题爬虫的核心算法。

```
01 Procedure TopicCrawling
02 Input: SeedURLs,ExamplesSet,γ,δ
03 Output: PagesCollection
04 Begin:
05  Let CandidateURLs be a table of URLs with their respective
parent page url and priority
    Let CrawledURLs be a table of URLs with their respective
relevancy and parent page url
    CandidateURLs ← SeedUrls
06  SetParentAndPriority(CandidateURLs,γ)
07  While CandidateURLs is not empty
08   Begin:
09    SortByPriority(CandidateURLs);
10    (parent_url,current_url) ← PickTop (CandidateURLs);
11    Page ← FetchPage (current_url);
12    If cos(page,ExamplesSet) >= δ
13      save(PagesCollection,Page);
14      PageRelevancy ← true;
15    Else
16      PageRelevancy ← false;
17    UpdateCrawledURLs(CrawledURLs,(current_url,parent_
url,relevancy));
18    Foreach url in Page
19      If url not in CrawledURLs
20        infos ← ExtractInfos(url,page,CrawledURLs)
21        relevantResultSets←AnalyzeFeatures(url,infos)
```

```
22          URLRelevancy← CalculateByStatistics(relevant-
ResultSets)
23          If URLRelevancy >= γ
24          Update (CandidateURLs,(url,current_url,URLRele-
vancy))
25      End;
26  End
```

从上述的伪代码中可以看出，主题爬虫使用两个URL列表，一个是候选URL列表(candidate URL list, CUL)，一个是已抓取的URL列表(fetched URL list, FUL)。

在初始时，利用一个种子URL列表对CUL进行初始化，设置每个URL的优先级为γ，并将自身设为父页面。接着，主题爬虫的网页抓取过程如下：

(1)首先，在10行，从CUL中取出优先级最高的URL，即为current_url，抓取其对应的网页，进行分析，若网页是主题相关的，则保存当前网页到网页集合中，并将当前的页面主题相关性pr(page relevancy)设置为True，表示页面是主题相关的，反之，则表示页面是主题无关的，pr设置为False。同时，保存current_url和其相关信息保存到已抓取URL列表(FUL)中。

(2)其次，从18行开始，循环地从网页中提取出每个URL，设其为u。若u未被抓取过，则提取出其相关特征信息，提交给特征信息分析模块进行分析，然后计算出每个特征信息的主题相关性结果，再对结果运用统计模型，计算出u的主题相关度$P(C)$，即u的优先级值。若$P(C)\geq\gamma$，则将u与其相关信息保存到候选URL列表(CUL)中，反之，若$P(C)<\gamma$，则过滤掉u；

(3)重复(1)，(2)直至候选URL列表(CUL)中的所有链接都被抓取。

5 实 验

主题爬虫的主要目标是抓取Web中和某个特定主题相关的网页，而Web中网页的主题是多样性的，且网页数量又是巨大的，其中和某个特定主题相关的网页只是整个Web空间的极小一部分。因此，对于主题爬虫而言，能够在检索过程中，遍历最小的Web空间，抓取最多的主题相关网页是较为关键的。所以，为了评估主题爬虫的效率，我们定义一个准确度*precision*，其计算方式如公式(13)所示。

$$precision=\frac{RP}{TP} \tag{13}$$

式中：*RP*——检索过程中所抓取到所有主题相关的网页，*TP*——检索过程中抓取到的所有网页。

为了避免网络相关因素对爬虫抓取性能的影响，我们通过一个简单的基于链接分析的主题爬虫，抓取了10000多张和数据库主题相关的网页，利用LSI计算出每个页面的主题相关度，按相关度从中选择出5000张网页，并对这些页面中的URL链接进行修改，建立成一个本地网页空间。为了评估基于统计模型的主题爬虫的性能，我们选择了基于链接分析的和基于页面内容分析的主题爬虫来进行本地网页空间抓取实验对比。我们用LC表示基于链接分析的主题爬虫，PC表示基于页面分析的主题爬虫，SC表示基于统计模型的主题爬虫。3种主题爬虫在检索结束后所抓取的网页数和其中主题相关的网页数如图2所示。

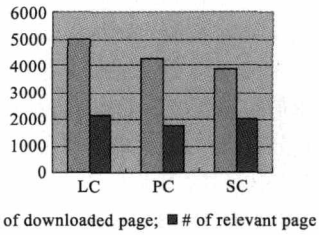

图2 主题爬虫抓取的网页数对比

从图2显示的最终结果，我们可以看出基于统计模型的主题爬虫，在检索了较少的网络空间的同时抓取到了相对较多的主题相关的网页。为了更好的比较不同的主题爬虫的效率，我们根据图2的结果，设置3种主题爬虫抓取网页的上限为4000，进一步对主题爬虫遍历过程中的数据进行统计、分析和对比。

3种主题爬虫在检索过程中所抓取的主题相关的网页数的曲线图如图3所示，主题爬虫在抓取过程中的准确度如图4所示。

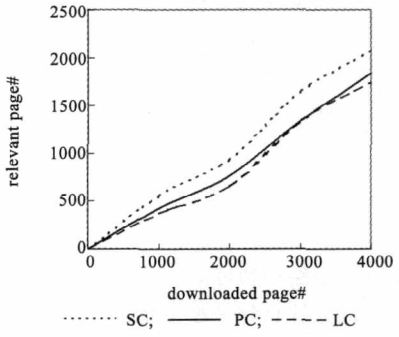

图3 主题相关的网页数的曲线

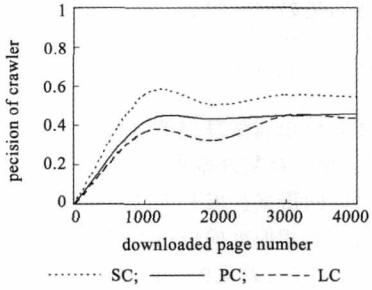

图4 准确度曲线

通过上述实验，我们可以看出，在抓取相同网页数量的过程中，相对于其它两种主题爬虫，基于统计模型的主题爬虫能够检索更少的网页，同时能够抓取到更多主题相关的网页，并能保持相对稳定的抓取精度。

6 结束语

本文通过对不同主题爬虫的研究，针对已有的主题爬虫所存在的缺陷，提出了一种基于统计模型的主题爬虫系统，它全面地分析了检索过程中所能获得的对URL主题相关度有

影响的特征信息，利用统计模型计算出 URL 的主题相关度，对 URL 进行过滤。通过实验对比得出，基于统计模型的主题爬虫能够更少的检索网络空间而抓取更多的主题相关网页，提高了主题爬虫的检索精度。但是，在对 URL 过滤时，需要对多种信息进行分析，相对其它的主题爬虫，它降低了爬虫的运行速度。因此，在以后的工作中，我们将改进特征信息分析算法，在提高检索精度的同时也加快其运行速度。

参考文献:

[1] Pant G.,Srinivasan P.Learning to crawl:Comparing classification schemes[J].ACM Transactions on Information Systems,2005,23(4):430-462.

[2] Menczer F, Pant G, Srinivasan P, et al. Evaluating topic-driven web crawlers[C].Proc 24th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval,2001: 241-249.

[3] Assis G T,Laender AHF,Silva ASd,et al.The impact of term selection in genre-aware focused crawling[C].Proceedings of the 23rd ACM Symposium on Applied Computing,2008:1158-1163.

[4] Liu H,Janssen JCM,Milios EE.Using HMM to learn user browsing patterns for focused web crawling[J].Data and Knowledge Engineering,2006,59(2):270-291.

[5] Eda Baykan, Monika Rauch Henzinger, Ludmila Marian, et al. Purely URL-based topic classification [C]. WWW, 2009: 1109-1110.

[6] Li Jun,Kazutaka Furuse,Kazunori Yamaguchi.Focused crawling by exploiting anchor text using decision tree [C]. ACM, 2005: 1190-1191.

[7] Chakrabarti S, Van Den Berg M, Dom B. Focussed crawling: A new approach to topic specific resource discovery[C]. Proceedings of the WWW Conference,1999:545-562.

[8] Deerwester S C,Dumais S T,Landauer T K,et al.Indexing by latent semantic analysis[J].Journal of the American Society of Information Science,1990,41(6):391-407.

[9] Berry M W. LSI: Latent semantic indexing web site [EB/OL]. http://www.cs.utk.edu/~lsi/,2006.

(上接第 3699 页)

(1)分词词典中缺少领域词汇；

(2)复合句导致计算句子相关度误差增大；

(3)句子逻辑结构复杂；

总之，本文所提出的方法，去除人为因素，自动批改的结果与人工批改结果趋向一致。

3 结束语

本文利用自然语言处理中的语句相似度计算方法，把词形、词序和词义相结合起来做相似度计算，利用同义词词林和知网来处理词语的语义扩展分析，提高了正确率。按照词形、词序和词义对句子相似度的作用不同，分别对其加以不同的权值，使相似度计算达到最优。在研究中，主要考虑通过学生答案与标准答案之间的特征作为相似度识别重点，首先将句子分割为一系列词串，通过同义词扩展及知网的语义扩展分析，得到词与词之间的相似度 ;然后对句子通过词与词的结构形式来计算相似度。由于汉语自身的复杂性，主观题自动阅卷技术还有很多需要处理的问题，研究工作还有待不断的深入和扩展。如果能更准确的对词语进行同义词与近义词的扩展，增强对复杂句式的分析能力，更好利用知网等资源的语义分析处理，会达到更好的效果。

参考文献:

[1] 高思丹. 基于自然语言理解的主观试题自动批改技术的研究与初步实现[D].南京:南京大学,2003:31-38.

[2] 周振波.考试系统基于中文分词技术的主观题评分尝试[J].科技信息,2009,28:609-610.

[3] 孟爱国,卜胜贤,李鹰,等.一种网络考试系统中主观题自动评分的算法设计与实现[J].计算机与数字工程,2005,33(7):147-150.

[4] 刘培奇.基于模糊含权概念图的主观题自动阅卷方法研究[J].计算机应用研究,2009,26(12):4565-4567.

[5] 贾电如. 基于语句结构及语义相似度计算主观题评分算法的研究[J].信息化纵横,2009(5):5-7.

[6] 南铉国.基于多层次融合的语句相似度计算模型[J].延边大学学报(自然科学版),2007,33(3):192-194.

[7] 吕学强,任飞亮,黄志丹,等.句子相似模型和最相似句子查找算法[J].东北大学学报(自然科学版),2003,24(6):531-534.

[8] 车万翔,刘挺,秦兵,等.基于改进编辑距离的中文相似句子检索[J].高技术通讯,2004(7):15-19.

[9] 秦兵,刘挺,王洋,等.基于常问问题集的中文问答系统研究[J].哈尔滨工业大学学报,2003,35(10):1179-1182.