支持向量机在化学主题爬虫中的应用

祝宇[1,2]，夏诏杰[1,2]，聂峰光[1]，郭力[1]
（1. 中国科学院过程工程研究所多相反应实验室，北京，100080；2. 中国科学院研究生院，北京，100049）

摘要：爬虫是搜索引擎的重要组成部分，它沿着网页中的超链接自动爬行，搜集各种资源。为了提高对特定主题资源的采集效率，文本分类技术被用来指导爬虫的爬行。本文把基于支持向量机的文本自动分类技术应用到化学主题爬虫中，通过SVM分类器对爬行的网页进行打分，用于指导它爬行化学相关网页。通过与基于广度优先算法的非主题爬虫和基于关键词匹配算法的主题爬虫的比较，表明基于SVM分类器的主题爬虫能有效地提高针对化学Web资源的采集效率。

关键词：支持向量机(SVM)；化学主题爬虫；文本分类；搜索引擎
中图分类号：TP 393
文献标识码：A
文章编号：1001-4160(2006)04-329-332

Research on chemistry focused crawler with support vector machine classifier

Zhu Yu[1,2], Xia Zhaojie[1,2], Nie Fengguang[1] and Guo Li[1]
(1. Lab of Multi-Phase Reaction, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, 100080, China; 2. Graduate University of Chinese Academy of Sciences, Beijing, 100049, China)

Abstract: Crawler is an important component of search engine, which collects Web pages through hyperlink between the pages. In order to enhance the performance of topic-specific search engines, text categorization techniques can be used to direct the crawling of focused crawlers. Based on Support Vector Machine, a new chemistry focused crawler is proposed in this paper. It can guide the focused crawler to collect the chemistry Web pages, and ignore the irrelevant information. The experiment results show that the focused crawler with SVM classifier is more effective to collect chemistry relevant pages, compared to the crawlers based on breadth first and keyword matching.

Key words: support vector machine, chemistry focused crawler, text categorization, search engine

Zhu Y, Xia ZJ, Nie FG and Guo L. Research on chemistry focused crawler with support vector machine classifier. Computers and Applied Chemistry, 2006, 23(4):329-332.

1 引言

随着因特网的快速发展，网络资源成为巨大的知识库，搜索引擎已经成为Web用户获取各种信息的一种重要手段。通用搜索引擎检索返回的结果过多、主题相关性不强。随着人们对提供的各项信息服务的要求越来越高，它们只能部分地解决资源发现问题[1,2]。面向特定领域的专业搜索引擎采用信息处理技术对检索进行指导，过滤掉大量不相关的资源，能帮助用户方便、快速、有效地获取所需信息。

爬虫(Crawler)[3,4]是搜索引擎的重要组成部分，也被称为Splider、Robot或Wander。它根据一定的网络协议，沿着网页中的超链接在Web上爬行，搜集各种资源并下载到本地服务器。通常搜索引擎会有多个爬虫并行搜索以提高效率。为了提高针对特定领域资源的采集效率，文本分类技术被用来指导爬虫的爬行，这些增加了爬行控制模块的爬虫称作主题爬虫，它能够对网页进行识别，过滤掉与主题不相关的资源，提高专业搜索引擎的服务质量。

支持向量机(support vector machine，SVM)方法建立在统计学习理论的VC维理论和结构风险最小化原理基础上[5]，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以期获得最好的推广能力。SVM是一种有监督学习的算法，是目前分类方法中较好的一种方法，已经被广泛应用到许多模式识别问题中，如数字识别、人脸识别、指纹识别等[6]。

本文将SVM技术应用于化学主题爬虫中，指导主题爬虫采集化学相关信息，提高针对化学Web资源的采集效率。

收稿日期：2005-08-16；**修回日期**：2005-12-28
基金资助：国家自然科学基金资助项目（20273076）

2 主题爬虫的设计和实现

广度优先算法(Breadth First)是通用的指导爬行算法,爬虫是按照网页的先进先出(FIFO)顺序爬行,采集相关资源。基于广度优先算法的爬虫是一种非主题爬虫。本文设计了两种主题爬虫,分别采用关键词匹配算法(Keyword Matching)和SVM监督算法(SVM Supervise)。

在关键词匹配算法中,我们建立了一个化学领域的关键词典,该词典收录的692个关键词是从ChIN[7]关键词库中提取出的化学领域基本术语,如chemistry、chemical等。当爬行到某网页时,如果该网页出现了词典中的关键词,则爬虫认为此网页是与主题相关,赋予该网页的分值(Score)为1,否则分值为0,并对该网页所链接的URLs赋予同该网页一样的分值,下次爬行将从分值最大的URL开始。基于关键词匹配算法的主题爬虫结构简单、易实现、效率高。

SVM监督算法与关键词匹配算法类似,不同之处在于使用的判别方法是基于SVM的文本分类器。主题爬虫的流程图见图1。

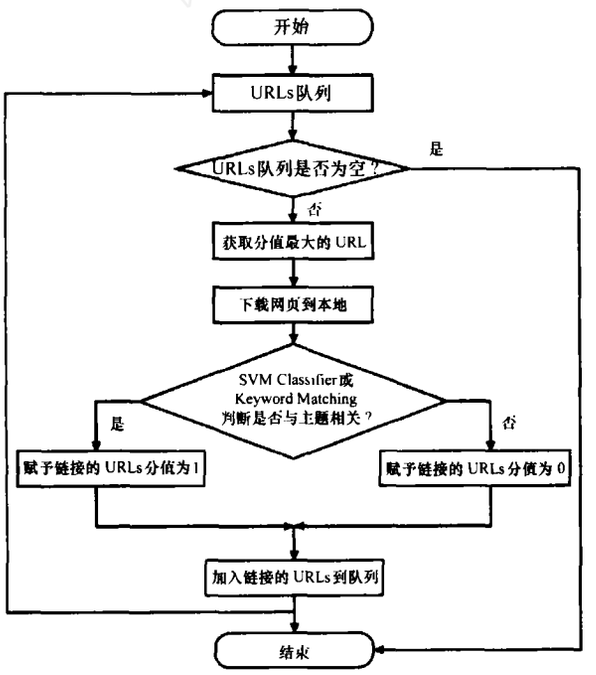

Fig. 1 Flow chart of chemistry focused crawler.

图1 主题爬虫的流程图

3 基于SVM的网页分类器的设计

3.1 对网页的处理

60年代末Salton等人提出的向量空间模型(vector space model,VSM)[8],将文档表示成计算机处理的所需形式,其基本思想是以向量来表示文本。利用此模型,一篇网页可以表示成$(d_i, w_1, w_2, \cdots, w_j, \cdots, w_n)$,其中$d_i$表示网页的类别信息,$w_j$表示网页第$j$个特征项的权重。

本文利用向量空间模型,对网页进行如下的处理,将网页表示成向量。

Step1:一个HTML文件中包含所有将显示在网页上的文字信息,其中也包括对浏览器的一些指示,如哪些文字应放置在何处,显示模式是什么样的等。本文去除这些指示,只提取文字信息。

Step2:进行分词,统计词频(term frequency,TF)和文档频率(document frequency,DF)。

Step3:去除停用词(Stopword)和还原词根(Stemming)。

Step4:特征提取(Feature Selection),去除区分度较小的噪音特征项可以提高分类准确率,去除重要性较低的低频特征项可以加快运行速度。常用的特征选择方法有文档频次、互信息、信息增益、χ^2统计量[9]。本文采用效果较好的信息增益方法,选择5000个特征项。

Step5:计算特征项权重。本系统采用了一种改进的TF-IDF公式:

$$W(i,j) = \frac{tf(i,j) \times \log(N/n_i + 0.01)}{\sqrt{\sum_{i \in j} [tf(i,j) \times \log(N/n_i + 0.01)]^2}}$$

其中,$W(i,j)$为词i在文本j中的权重,而$tf(i,j)$为词i在文本j中的词频,N为网页集合的总数,n_i为网页集中出现i的网页数,分母为归一化因子。

经过上述处理后,每篇网页表示成空间中的一个向量。

3.2 基于SVM的网页分类器

获取网页空间向量后,就建立好了一个SVM训练样本集$\{d_i, x_i\}_{i=1}^{l}$,其中$d_i = \{1, -1\}$是网页x_i的类别信息,l是训练样本集的大小。

Vapnik等人基于统计学习理论提出的支持向量机(SVM)方法是用于解决分类问题的一种新的学习方法[10-12],是近年来机器学习研究的一项重大成果。其基本思想是通过非线性变换将输入空间变换到一个高维空间,使样本线性可分,然后求取最优分类面(见图2),即分类间隔(Margin)最大的超平面。所谓间隔是指该超平面到最近样本的距离,即$2/\|\omega\|$。

文本具有高维的输入空间、特征项相关性小、文

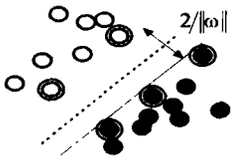

Fig. 2　Optimal separating hyperplane.

图 2　最优分类面

档向量稀疏、分类问题线性可分等特性[8],SVM 适合于文本(网页)分类。

在网页训练集合经过 SVM 分类器训练后,得到一个分类器的模型(Model)。一个新的网页到来之后,要经过上述相似的处理过程,得到空间向量,不同之处在于类别 d 是未知的,需要经过分类器预测 d 值,从而得到该网页的类别信息。

4　实验与分析

4.1　化学主题模型的建立

我们从 DMOZ[13]下载化学领域(生化、化工、高分子化学等)相关网页 1 949 篇和非化学领域(新闻、经济、体育和其它科学领域等)网页 2 231 篇。利用这些网页建立分类训练集,经过 SVM 分类器训练后,得到关于化学和非化学的分类主题模型。

4.2　初始 URLs 的选择

初始 URLs 的选择必须十分慎重[14],因为它将影响采集的效率,尤其是爬虫初始爬行的准确率。为了测试初始 URLs 对采集效率的影响,我们选取化学相关和无关的 URL 各 150 个,每 50 个 URL 作为一组,共得到六组初始种子。为了检验本文所提方法的有效性,对每组初始种子均采用了 SVM 监督、广度优先和关键词匹配三种爬行算法。

4.3　测试结果及分析

对爬行下来的大量网页进行与主题相关性的判断,采用人工浏览的方法是不切实际的,本文采用已经建立好的 SVM 分类器判断网页是否与主题相关。

使用同样的分类器来指导爬虫爬行和判断爬行下来的网页是否与主题相关,似乎在方法上存在缺陷,实际上并非如此[15],首先训练集和测试集使用的是不同的网页,其次也不是在用分类器本身来检验早期的工作,我们只是在评价这些高相关性的链接资源,而不是在精确分类。本文将三种算法爬行下来的网页,均使用 SVM 分类器进行与主题相关性的判断。

对于每组初始 URLs,我们均爬行 30 000 篇网页,按照爬行的先后顺序,每隔 500 篇网页统计一次局部爬准率和全局爬准率。局部爬准率是指 500 篇网页的爬准率,全局爬准率是指已经爬行的全部网页的爬准率,结果如图所示。

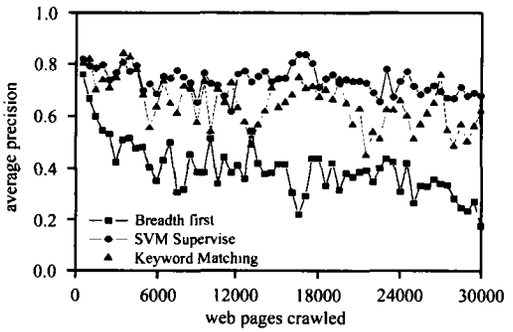

Fig. 3　Local average precision of web pages crawled with seed urls of chemistry.

图 3　初始种子与化学相关时局部爬准率曲线

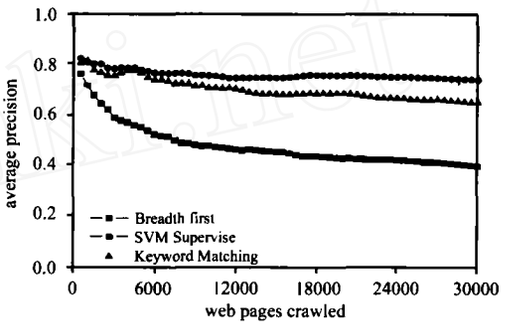

Fig. 4　Whole average precision of web pages crawled with seed urls of chemistry.

图 4　初始种子与化学相关时全局爬准率曲线

图 3 图 4 展示的是初始种子与化学主题相关时的局部和全局爬准率曲线,可以看出,主题爬虫采集资源的效果明显优于非主题爬虫。SVM 监督算法的采集效果最好,波动小,保持了较高的爬准率;关键词匹配算法局部波动较大,导致了全局爬准率的下降;广度优先算法效果较差,随着爬行网页的增多,局部和全局的爬准率越来越接近于零。

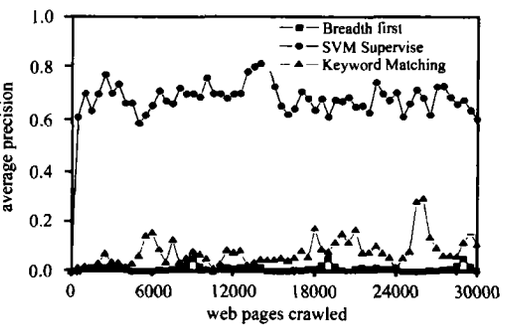

Fig. 5　Local average precision of web pages crawled with seed urls of non-chemistry.

图 5　初始种子与化学无关时局部爬准率曲线

图 5 图 6 展示的是初始种子与化学主题无关时

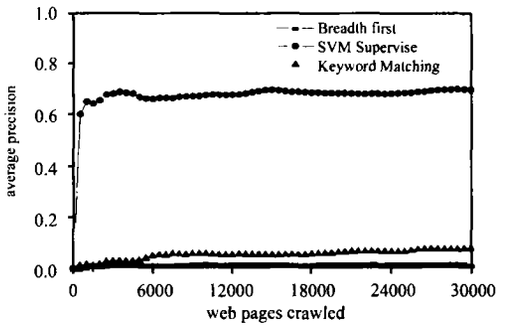

Fig. 6 Whole average precision of web pages crawled with seed urls of non-chemistry.

图 6 初始种子与化学无关时全局爬准率曲线

的局部和全局爬准率曲线,可以看出,主题爬虫采集效果优势更加突出,非主题爬虫几乎找不到与化学主题相关的资源。SVM 监督算法较好地消除了初始 URLs 对采集效率的影响,仍然保持了较高的爬准率;广度优先和关键词匹配算法受初始 URLs 的影响较大,爬准率较低,采集效果很差。

5 结论

通过与基于广度优先算法的非主题爬虫和基于关键词匹配算法的主题爬虫的比较,运用基于 SVM 的文本自动分类技术的 SVM 监督算法,消除了初始 URLs 对采集效率的影响,并且对主题爬虫的爬行指导效果好,保持了较高的爬准率,波动较小,能有效地过滤掉非化学的信息,对化学主题的 Web 资源采集优势明显。

References

1 Du AN, Fang BX and Hu MZ. A decision tree-based web mining system for chinese pages. Advances of Search Engine and Web Mining in China. Beijing: Higher Education Press, 2003, 3.

2 Li XX, Yang ZY and Xu ZH. Today and tomorrow's chemical resources on internet. Computers and Applied Chemistry, 1999, 16(5):325 - 326.

3 *Heydon A and Mercator NM. A scalable, extensible web crawler.* World Wide Web, 1999, 2(4):219 - 229.

4 Wang JC, Xiao R, Sun ZX and Zhang FY. State of the art of information retrieve on the web. Journal of Computer Research and Development, 2001, 38(2):187 - 193.

5 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer, 1995.

6 Burges CJC. A tutorial on support vector machines for pattern recognition. Knowledge Discovery and Data Mining, 1998.

7 Institute of Process Engineering. Chinese Academy of Sciences. The Chemical Information Network. http://www.chinweb.com, 2005, 4,19.

8 Salton G and McGill MJ. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

9 Yang Y and Pedersen J. A comparative study on feature selection in text categorization. International Conference on Machine Learning (ICML), 1997.

10 Joachims T. Text categorization with support vector machines: Learning with Many Relevant Feature. Proceedings of European Conference on Machine Learning(ECML), 1998.

11 Cortes C and Vapnik V. Support vector networks. Machine Learning, 1995, 20:273 - 297.

12 Statistical Learning Theory. New York: Wiley, 1998.

13 DMOZ. http://www.dmoz.net, 2005,5,20.

14 Li ST, Yu ZH, Cheng XQ and Bai S. A survey on web crawling. Computer Science, 2003, 30(2):151 - 157,171.

15 Chakrabarti S Vandenberg M and Dom B. Foused crawling: a new approach to topic-specific web resource discovery. In Proceedings of the Eighth International World Wide Web Conference, 1999.

附中文参考文献

1 杜阿宁, 方滨兴, 胡铭曾. 一个基于决策树的中文 Web 文本挖掘系统. 搜索引擎与 Web 挖掘进展. 北京:高等教育出版社, 2003, 3.

2 李晓霞, 杨章远, 许志宏. Internet 化学资源的发展状况与展望. 计算机与应用化学, 1999, 16(5):325 - 326.

4 王继成, 萧嵘, 孙正兴, 等. Web 信息检索研究进展. 计算机研究与发展, 2001, 38(2):187 - 193.

7 中国科学院过程工程研究所. 化学信息门户, http://www.chinweb.com, 2005,4,19.

14 李盛韬, 余智华, 程学旗, 白硕. Web 信息采集研究进展. 计算机科学, 2003, 30(2):151 - 157,171.