第 30 卷 第 3 期
2015 年 6 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 30 No. 3
Jun. 2015

DOI:10.13364/j.issn.1672-6510.20140135

一种融合本体和最小二乘支持向量机的主题爬行方法

马永军，杨海波
（天津科技大学计算机科学与信息工程学院，天津 300222）

摘 要：针对现有的主题爬行方法存在收益率不高和不稳定的问题，融合本体和最小二乘支持向量机理论，提出一种主题爬行方法 Ontology-LSSVM. 该方法将本体作为抓取主题的背景知识表示，得到一组主题相关概念的集合，再将其在网页文本中出现的词频作为输入，构造一个主题相关度 LS-SVM 分类器，用于网页的分类. 使用舆论热点的食品安全问题作为爬行主题领域，建立问题食品本体，进行实验. 结果表明：在本实验条件下，本文方法相比基于 LS-SVM、基于本体和基于关键字的主题爬行，能够维持更高的收益率.

关键词：本体；支持向量机；主题爬行；收益率；食品安全
中图分类号：TP391.3 **文献标志码**：A **文章编号**：1672-6510(2015)03-0072-06

A Focused Crawling Approach Combining Ontology and LS-SVM

MA Yongjun，YANG Haibo
(College of Computer Science and Information Engineering, Tianjin University of Science & Technology, Tianjin 300222, China)

Abstract: There are many implementation approaches to focused crawler. However, it is difficult to maintain a high and stable crawling process. To solve this problem, a focused crawling approach based on ontology and LS-SVM (least squares support vector machine) was proposed. An LS-SVM classifier was created using the problematic food ontology and applied to the classification of web pages. Experimental results show that the proposed approach can get a higher harvest rate than other focused crawling approaches.

Key words: ontology; LS-SVM; focused crawling; harvest rate; food safety

随着互联网信息数量迅猛增长，信息获取过程变得不顺畅. 在大数据背景之下，如何快速、精准地获取行业领域的数据成为亟待解决的问题[1]. 在互联网的海量数据中抽取主题相关数据就必须依赖于主题爬行，因此，寻找高效的主题爬行方法是研究的重点.

主题爬行[2]是指根据预设的主题在海量网页中抓取主题相关页面，即面向特定领域的爬行. 目前，已有众多主题爬行方法：Chakrabarti 等[3]首次提出了主题爬行，并给出包括分类器、蒸馏器和收集器的通用主题爬行架构；在众多主题爬行方法中基于文本内容的方法先被提出，而后出现基于超链接层级结构分析的方法，如基于上下文图模型的主题爬行算法的出现，也催生了该算法的多种改进算法[4–5]，基于文本与链接分析结合方法[6]也相继被提出；为了使爬虫能够智能遍历链接，机器学习算法也被应用到主题爬行中，譬如基于隐马尔可夫模型 HMM 的主题相关链接预测方法[7]、基于 SVM[8]自主学习的方法等；此外，主题爬行在行业领域的应用中也得到广泛关注，例如 Can 等[9]创建的 MedicoPort，专注于收集医疗相关文档.

但是，上述主题爬行方法均未考虑到应用背景知识，如用户喜好或者领域本体等. 背景知识对理解问题和分析具体情况非常有用. 由于本体是一种结构清晰的知识表示方案，使用概念和关系来呈现高层语义化背景知识. 为了提高信息管理系统的性能，现已被一些系统采用. 如 Rosaci[10]利用本体和 ANN 表示

收稿日期：2014–10–23；**修回日期**：2015–03–04
基金项目：天津市科技支撑计划重点资助项目(12ZCZDGX02400)
作者简介：马永军（1970—），男，吉林长春人，教授，yjma@tust.edu.cn.

用户的喜好和行为，提升了系统的推荐质量. 在主题爬行领域，也有一些方法使用本体作为相关领域的背景知识[11−12]，不足之处在于其使用本体概念的权重计算相关度. 由于本体概念的权重是在计算网页主题相关度之前预先试探性设置，所以存在爬行过程中难以获得最优概念权重的问题，导致得到相关度不能最佳地反映主题相关度.

本文提出一种融合本体和最小二乘支持向量机的主题爬行方法，使用本体作为主题爬行的背景知识，同时采用最小二乘支持向量机作为分类器，通过训练样本量化得到概念权重. 支持向量机（support vector machine，SVM）是一种使被广泛应用的机器学习方法，但是存在计算复杂性高的问题. 最小二乘支持向量机（least squares support vector machine，LS-SVM）是对标准支持向量机的扩展，使用最小二乘线性系统替换了标准支持向量机中凸二次规划，将求解过程变成了求解方程组，有效地避免了二次规划耗时问题. 最小二乘支持向量机仍然可以采用核函数，作原始特征空间向高维空间的映射，解决在原始输入空间中线性不可分的问题，具有更好的泛化能力. 本文针对食品安全领域，选择 40 个相关网站，收集我国 2005—2012 年间发生的 3 300 例食品安全事件作为研究样本[13]，建立了问题食品本体，实现领域背景知识在语义层面扩展主题关键字，构建了一个 LS-SVM 分类器，用于网页的分类，以提高主题爬行的收益率.

1　本体与 SVM 简介

1.1　本体

本体概念源于哲学领域，是由古希腊哲学家亚里士多德提出的，用以解决事物分类问题. 目前本体已经被应用到了诸多领域，并于 1993 年由 Gruber 引入计算机科学与信息科学领域，并定义其为一种“形式化、语义化的共享概念体系的明确规范”[14]. 主要用于表述特定领域知识，并从不同层面对概念与概念之间的相互关系进行明确定义.

定义 1　特定领域的本体可以表示为

$$O=\{C,R,I,F\} \tag{1}$$

其中：C 为一组概念的集合，c 是代表某一具体事物的概念（$c\in C$）；R 为一组关系的集合，代表特定领域中概念间相互作用；F 为一种关系映射，例如关系 $is-a(C_1)=C_2$ 表示 C_1 是 C_2 的 1 个子类，关系 $instanceof(C)=I$ 表示 I 是 C 的一个实例，I 则为一组实例的集合，某个概念的具体实现.

1.2　支持向量机

支持向量机是一种基于统计学和机构风险最小化理论的分类算法[15]，其核心思想是使用预先选定的核函数（kernel method）作为输入特征向量到高维空间的映射，在高维空间内获得最优的分类超平面，使得最大化分类器间隔. 令分类超平面函数表达式

$$f(x)=\boldsymbol{w}^{\mathrm{T}}x+b \tag{2}$$

在此最优化问题中，将原始问题巧妙地转化为对偶问题，利用 Lagrange 对偶性，引入 Lagrange 对偶变量 α，采用 Lagrange 乘数法得到目标函数：

$$\begin{aligned}&\max_{\alpha_i\geqslant 0} L(\boldsymbol{w},b,\boldsymbol{\alpha})=\frac{1}{2}\|\boldsymbol{w}\|^2-\sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w}^{\mathrm{T}}x+b)-1]\\&\quad\text{s.t. }\alpha_i\geqslant 0,i=1,2,\cdots,n\end{aligned} \tag{3}$$

原生支持向量机用于处理线性可分的情况，而对于非线性的情况，则需要借助核函数完成. 通过特征数据映射到高维空间，解决在原空间中线性不可分的问题. 分类函数的内积形式表达式为

$$f(x)=\sum_{i=1}^{n}\alpha_i y_i\kappa(x_i,x)+b \tag{4}$$

把用于计算两个向量在映射后高维空间中内积的函数称之为核函数 κ. 相应地目标函数变为

$$\begin{aligned}&\max_{\alpha_i\geqslant 0}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\alpha_i\alpha_j y_i y_j\kappa(x_i,y_i)\\&\quad\text{s.t. }\alpha_i\geqslant 0,i=1,2,\cdots,n\\&\quad\sum_{i=1}^{n}\alpha_i y_i=0\end{aligned} \tag{5}$$

其中 α 的求解一般采用序列最小优化算法 SMO.

1.3　最小二乘支持向量机

最小二乘支持向量机将最小二乘的思想引入支持向量机，目的是构造一个间隔最大的超分类平面，是标准支持向量机的扩展[16]. 较之标准的支持向量机，其增加了松弛变量 e_i，同时将约束条件变成了等式约束，把问题转换为线性方程组的求解问题，算法的计算量锐减，也有效避免了惩罚因子 C 选择问题.

求解最优化超平面问题，可以表述为式（6）的二次凸优化问题：

$$\begin{aligned}&\min_{w,b,e} J_{LS}(\boldsymbol{w},b,e)=\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{\gamma}{2}\sum_{k=1}^{N}e_k^2\\&\quad\text{s.t. }y_i(\boldsymbol{w}^{\mathrm{T}}\varphi(x_i)+b)=1-e_i,i=1,\cdots,l\end{aligned} \tag{6}$$

则其 Lagrange 函数为

$$\begin{aligned}L(\boldsymbol{w},b,e;\boldsymbol{\alpha})=&J_{LS}(\boldsymbol{w},b,e)-\\&\sum_{k=1}^{N}\alpha_k\{y_k[\boldsymbol{w}^{\mathrm{T}}\varphi(x_k)+b]-1+e_k\}\end{aligned} \tag{7}$$

根据 KKT 条件求解，可以得到矩阵：

$$\begin{bmatrix} 0 & -\boldsymbol{Y}^{\mathrm{T}} \\ \boldsymbol{Y} & \boldsymbol{\Omega}+\boldsymbol{\gamma}^{-1}\boldsymbol{I} \end{bmatrix}\begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix}=\begin{bmatrix} 0 \\ \vec{1} \end{bmatrix} \tag{8}$$

其中有 $\boldsymbol{Y}=[y_1,\cdots,y_N]^{\mathrm{T}}$，$\vec{1}=[1,\cdots,1]^{\mathrm{T}}$，根据 Mercer 条件，$\boldsymbol{\Omega}$ 可以用式(9)表示.

$$\Omega_{kl}=y_k y_l\varphi(x_k)^{\mathrm{T}}\varphi(x_l)=y_k y_l\kappa(x_k,x_l) \tag{9}$$

根据式(8)和式(9)，选择合适的核函数就可以求解此线性方程组，得到最小二乘支持向量机分类器.

2　Ontology-LSSVM 主题爬行方法

2.1　爬行系统工作流程

Ontology-LSSVM 主题爬行方法由 3 个工作独立的阶段构成，如图 1 所示，主要包含数据准备阶段、训练阶段以及抓取阶段. 数据准备阶段负责准备训练样本；在训练阶段，使用训练样本来训练 LS-SVM 分类器；在抓取阶段，抓取网页并由 LS-SVM 分类器来决定是否与爬行主题相关，相关则下载到本地，反之丢弃；最终，将下载的网页存储在特定领域的网页资料库.

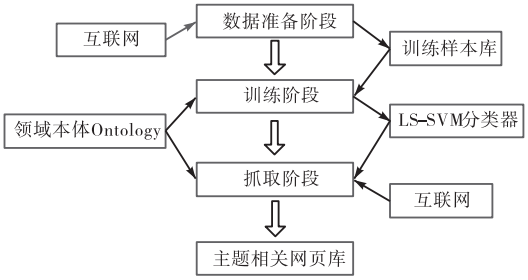

图 1　Ontology-LSSVM 的主题爬行工作流程

Fig. 1　Framework of focused crawling based on Ontology-LSSVM

2.1.1　数据准备阶段

在数据准备阶段，使用 Scrapy 爬行框架编写一个简易的爬虫，抓取网页作为训练样本. Scrapy 是一个高效稳定的爬行框架，能够模拟标准浏览器的行为. 给定一个种子 URL，爬虫将会准确地抓取指定的网页. 在抓取到一组网页集合后，人工决定具体网页与爬行主题的相关度. 这样就获得了实验必需的正训练样本和负训练样本.

2.1.2　训练阶段

在训练阶段中，使用第一阶段获得的训练样本来训练 LS-SVM 分类器，根据表示背景知识的特定领域本体来选定一些相关概念. 这些概念必须与爬行主题相似或相近，即主题相关概念. 训练样本中的每个网页通过预处理得到一组主题相关概念的集合.

一般预处理模块包括：网页从 HTML 到文本的转换、中文分词处理、去除停用词. HTML Parser 被用于完成 HTML 文本转换，去除所有的 HTML 标签. 调用中国科学院计算技术研究所研发的中文分词系统 ICTCLAS 做中文分词处理. 在对抓取网页预处理之后，抽取出其中主题相关概念，再统计出主题相关概念的词频. 最后把它们作为网页主题的特征来训练 LS-SVM 分类器.

2.1.3　抓取阶段

抓取阶段主要关注的是实际抓取和 LS-SVM 分类器的使用. 首先，爬虫会遍历主机的 robot.txt 文件信息来决定该网页是否允许抓取. 若被抓取网页通过检查，则爬虫将会抓取该网页并计算其主题的相关度. 利用建立的爬行主题领域的本体，即一组关系与概念集，来获得爬行主题相关的背景知识. 在爬行过程中，通过预处理得到网页的主题相关概念的词频，再将词频作为 LS-SVM 分类器的输入，计算出该网页的主题相关度. 若网页的主题相关度被纳入相关，则下载该网页并存入特定主题资料库；反之，则放弃本次爬行.

2.2　主题相关度的计算机制

主题相关度的计算在此爬行方法中是至关重要的一步. 从本体定义可以知道，本体就是使用一组概念及概念间的相互关系表示特定领域知识. 给定某个特定的领域，可以根据背景知识选择出相关概念来计算网页的主题相关度. 此外，选择相关概念的范围是由爬行主题与本体中概念之间的距离决定的.

定义 2　本体概念间的距离

$$d(t,c_i)=l,\quad l\in N \tag{10}$$

式中：l 表示在本体 O 中 t 到 c_i 结点之间的连接数；t 表示在本体 O 中与爬行主题对应的概念；c_i 表示本体 O 中概念；$i\in\{1,\cdots,n\}$，n 为本体 O 中概念的个数. 连接数是指在本体的层级结构中，概念结点连通时经过的最小连接线的数量.

概念之间的距离反应的是本体中概念与爬行主题的相关性，若 $d(t,c_i)$ 大，则说明此概念与爬行主题的相关度较低. 同时随着本体中需要进行距离运算的概念个数增加，又因为本体的树状层级结构，相关概念距离计算量将会指数级增长，导致爬行时间复杂度的快速增长. 为了能够保持爬行的高效性，必须通

过实验选择一个合适的主题与概念间距离 $d(t,c_i)$.

根据爬行主题与概念间距离的定义，给定一个爬行主题，就能够在本体中找到一组相关概念 c_i. 对爬虫抓取到的网页，通过预处理模块就能够统计出相关概念在网页文本中出现次数，即词频. 主题相关概念在网页文本中出现的词频能够反映出该概念对此网页文本语义的重要程度. 将相关概念的词频作为 LS-SVM 分类器的输入，输出网页与爬行主题的相关度，根据相关度决定是否下载这个网页.

图 2 呈现了一个主题相关度计算的过程实例. 主题相关度计算需要两步：首先通过预处理完成中文分词及词频统计；再利用 LS-SVM 分类器计算主题相关度. 在图 2 中，C_1 到 C_n 分别代表主题相关概念 c_1 到 c_n 的词频，将其输入到 LS-SVM 分类器，便能获得网页的主题相关度. 使用“牛奶”作为爬行主题，借鉴文献[17]的研究成果，可建立如图 3 所示的问题食品本体. 将概念间距离不大于阈值的概念作为主题相关概念，如 $d(t,c_i)\leqslant 2$. 再统计主题相关概念 c_i 的词频，输入 LS-SVM 分类器，可得到该网页的主题相关度，即该网页与“牛奶”的相关度.

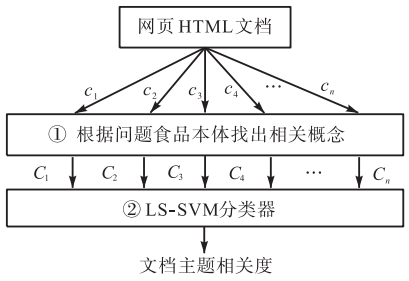

图 2　相关度计算过程实例

Fig. 2　An example of relevance computation

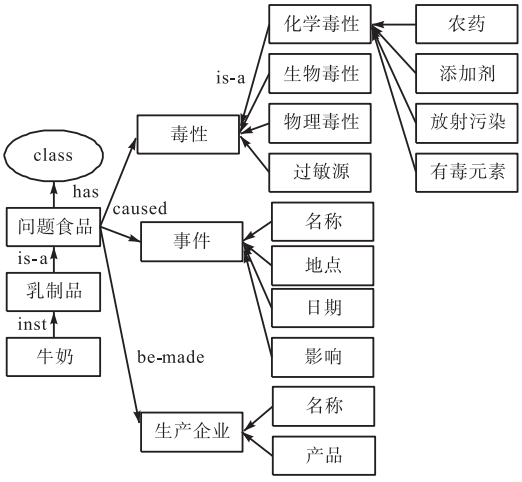

图 3　问题食品本体

Fig. 3　Problematic food ontology

3　实　验

将本文方法与基于关键字、基于特定领域本体及基于 LS-SVM 的主题爬行方法进行比较. 基于 LS-SVM 的主题爬行方法只是使用 LS-SVM 分类器，并未使用特定领域的本体，将网页中所有的词都进行词频统计，将其作为网页的特征用于训练 LS-SVM 分类器，得到网页的主题相关度，决定是否相关. 实验中采用与本文方法相同的训练样本训练.

主题相关网页的过滤效率对于主题爬行是一个至关重要的评价标准，即收益率. 文献[3]等定义的收益率为

$$h_r=\frac{r}{p},h_r\in[0,1] \tag{11}$$

收益率是抓取到满足主题的网页 r 在所有抓取网页 p 中占的比重. 高收益率意味着主题爬虫能够高效抓取主题相关网页，反之，则爬虫需要花费大量时间处理主题无关网页. 因此，高收益率是高效爬虫的重要标志.

实验环境采用 Intel Core2 Duo 2.80 GHz 处理器，2.0 GB 内存，操作系统为 Microsoft Windows 7 Ultimate.

在数据准备阶段，针对爬行主题“牛奶”，收集到了 150 个训练样本，其中包括 98 个正样本和 52 个负样本. 在训练阶段，首先确定最适合的本体概念距离. 对于不同的概念距离 $d(t,c_i)$，选择出的主题相关概念. 再将这些概念的词频作为输入，构建 LS-SVM 分类器.

图 4 为在 20 min 内爬虫以“牛奶”为爬行主题，分别在不同概念距离下抓取到的主题相关网页数量. 可以看出：在以“牛奶”为爬行主题，距离 $d(t,c_i)\leqslant 3$ 时，爬虫具有最好的性能. 当概念距离过小时，得到的主题相关概念很少，爬虫抓取网页的命中主题数量变少. 反之，取过大的概念距离，主题相关概念就会过多，致使每个网页的相似判断耗时越长，将引起爬虫抓取网页的命中主题数量下降.

各种方法的爬行收益率见图 5 和表 1. 所有爬行过程在刚开始具有较高的收益率，但随着抓取网页数量的增长有所下降. 这是因为实验选择与主题相关度较高的种子 URL，随着爬行的进行与主题不相关的网页数量增多导致收益率降低. 本文方法的收益率高于基于本体的主题爬行，主要原因在于基于本体

的方法中主题相关度依赖于预设的权重值，有较大局限性，而本文方法的主题相关度是通过 LS-SVM 学习得到的.

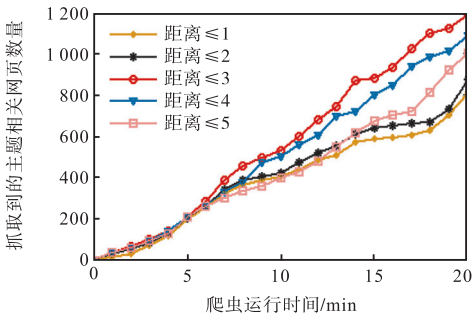

图 4　爬行主题不同距离的比较

Fig. 4　Comparison of different distances

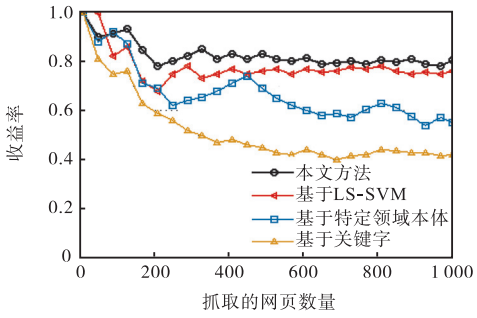

图 5　各种方法的爬行收益率

Fig. 5　Harvest rate of each approach

表 1　各种主题爬行方法的平均收益率

Tab. 1　Average harvest rate of each approach

主题爬行方法	平均收益率
本文方法	0.831 0
基于 LS-SVM	0.724 6
基于特定领域本体	0.685 1
基于关键字	0.437 2

同时，实验表明基于 LS-SVM 的爬行方法也有较高的收益率，是因为其主题相关度也是通过训练好的 LS-SVM 分类器得到的. 但其总收益率低于本文方法，原因是：一方面，通过特定领域本体可以有效地应用背景知识扩展主题，提高了爬行的精准性；另一方面，基于 LS-SVM 的爬行方法是将所有页面中词的词频作为输入计算相关度，在计算复杂性上高于本文方法.

从性价比的角度对几种爬行方法进行比较，实验统计了爬行 20 min 各爬虫的主题命中数[18]，实验结果见图 6. 可以看出：基于特定领域本体的方法和本文方法都有较高的命中数，且基于特定领域本体的方法高于本文方法；基于关键字的方法由于未考虑主题相关扩展，所以在整个过程中主题命中数最低；基于 LS-SVM 的方法将网页中所有的词都作为特征输入，所以计算复杂度高，主题命中数也不高. 基于特定领域本体方法的主题命中数高于本文方法，原因在于本文方法采用 LS-SVM 计算网页主题相关度，较之前者需要更多计算时间. 尽管基于特定领域本体的爬行方法具有更高的主题命中数，但是本文方法能够维持高收益率，这就意味着本文方法能够更高效地抓取主题相关网页. 同时，与基于 LS-SVM 的爬行方法相比，本文方法网页文本特征向量的维度更低，性能更好.

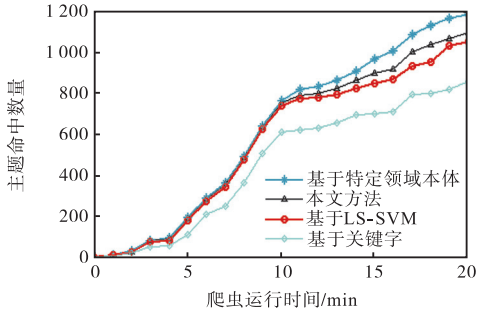

图 6　20 分钟内的主题命中情况

Fig. 6　Number of topic hits during 20 min

4　结　语

本文提出了一种融合本体与支持向量机的爬行方法 Ontology-LSSVM，用以解决主题爬行收益率不高的问题. 此方法应用领域本体作为背景知识扩展，得到主题相关概念集，将其在网页文本中的词频输入 LS-SVM 分类器，计算出该网页的主题相关度，指导后续网页的下载、存储、索引等工作. 实验结果表明，本文提出的主题爬行方法优于简单基于本体、LS-SVM 等的实现方法，主题爬行的效率得到提高，并且能够维持高收益率.

当然本文方法受限于采用本体的质量，构建的本体的完善度直接影响到爬行结果. 因此，在建立大型完备的系统之前，应该先建立一个能够完全体现特定领域背景知识的本体.

参考文献：

[1] Shestakov D. Current challenges in web crawling[M]// Lecture Notes in Computer Science. Heidelberg: Springer Berlin Heidelberg, 2013: 518–521.

[2] Olston C, Najork M. Web crawling[J]. Foundations and

Trends in Information Retrieval, 2010, 4(3): 175–246.

[3] Chakrabarti S, Van den Berg M, Dom B. Focused crawling: a new approach to topic-specific Web resource discovery[J]. Computer Networks, 1999, 31(11): 1623–1640.

[4] Hu Qiang, Du Yajun, Yang Jiaying, et al. An optimized relevancy Context Graph based on social network[J]. Journal of Information and Computational Science, 2014, 11(6): 2029–2038.

[5] 张永，吴崇正. 基于词频差异特征选取的 Context Graph 算法改进[J]. 计算机工程与应用，2014，50(10): 141–146.

[6] Almpanidis G, Kotropoulos C, Pitas I. Combining text and link analysis for focused crawling: An application for vertical search engines[J]. Information Systems, 2007, 32(6): 886–908.

[7] Liu H, Janssen J, Milios E. Using HMM to learn user browsing patterns for focused web crawling[J]. Data & Knowledge Engineering, 2006, 59(2): 270–291.

[8] Shein K P P, Nyunt T T S. Sentiment classification based on ontology and SVM classifier[C]//Proceedings of the 2nd International Conference on Communication Software and Networks. Piscataway: IEEE, 2010: 169–172.

[9] Can A B, Baykal N. MedicoPort: A medical search engine for all[J]. Computer Methods and Programs in Biomedicine, 2007, 86(1): 73–86.

[10] Rosaci D. CILIOS: Connectionist inductive learning and inter-ontology similarities for recommending information agents[J]. Information Systems, 2007, 32(6): 793–825.

[11] Su C, Gao Y, Yang J, et al. An efficient adaptive focused crawler based on ontology learning[C]//Proceedings of the 5th International Conference on Hybrid Intelligent Systems. Piscataway: IEEE, 2005: 73–78.

[12] Luong H P, Gauch S, Wang Q. Ontology-based focused crawling[C]//Proceedings of IEEE International Conference on Information, Process, and Knowledge Management. Piscataway: IEEE, 2009: 123–128.

[13] 张红霞，安玉发，张文胜. 我国食品安全风险识别、评估与管理：基于食品安全事件的实证分析[J]. 经济问题探索，2013(6): 135–141.

[14] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition, 1993, 5(2): 199–220.

[15] 顾亚祥，丁世飞. 支持向量机研究进展[J]. 计算机科学，2011，38(2): 14–17.

[16] 顾燕萍，赵文杰，吴占松. 最小二乘支持向量机的算法研究[J]. 清华大学学报：自然科学版，2010，50(7): 1063–1066.

[17] Yang Y, Du J, He B. A novel ontology-based semantic retrieval model for food safety domain[J]. Chinese Journal of Electronics, 2013, 22(2): 247–252.

[18] 陈能成，陈泽强，王伟. 一种基于能力匹配和本体推理的高精度 Web 地图服务发现方法[J]. 武汉大学学报：信息科学版，2009，34(12): 1471–1475.

责任编辑：常涛