基于多粒度树模型的 Web 站点描述及挖掘算法*

田永鸿[1+], 黄铁军[1,2], 高 文[1,2,3]

[1](中国科学院 计算技术研究所,北京 100080)

[2](中国科学院 研究生院,北京 100039)

[3](哈尔滨工业大学 计算机科学与工程系,黑龙江 哈尔滨 150001)

A Web Site Representation and Mining Algorithm Using the Multiscale Tree Model

TIAN Yong-Hong[1+], HUANG Tie-Jun[1,2], GAO Wen[1,2,3]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

[2](Graduate School, The Chinese Academy of Sciences, Beijing 100039, China)

[3](Department of Computer Science, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: Phn: +86-10-82649529, Fax: +86-10-82649298, E-mail: yhtian@jdl.ac.cn, http://www.jdl.ac.cn

Received 2003-06-02; Accepted 2003-07-08

Tian YH, Huang TJ, Gao W. A Web site representation and mining algorithm using the multiscale tree model. *Journal of Software*, 2004,15(9):1393~1404.

http://www.jos.org.cn/1000-9825/15/1393.htm

Abstract: With the exponential growth of both the amount and the diversity of the web information, web site mining is highly desirable for automatically discovering and classifying topic-specific web resources from the World Wide Web. Nevertheless, existing web site mining methods have not yet handled adequately how to make use of all the correlative contextual semantic clues and how to denoise the content of web sites effectually so as to obtain a better classification accuracy. This paper circumstantiates three issues to be solved for designing an effective and efficient web site mining algorithm, i.e., the sampling size, the analysis granularity, and the representation structure of web sites. On the basis, this paper proposes a novel multiscale tree representation model of web sites, and presents a multiscale web site mining approach that contains an HMT-based two-phase classification algorithm, a context-based interscale fusion algorithm, a two-stage text-based denoising procedure, and an entropy-base pruning strategy. The proposed model and algorithms may be used as a starting-point for further investigating some related issues of web sites, such as query optimization of multiple sites and web usage mining. Experiments also show that the approach achieves in average 16% improvement in classification accuracy and 34.5% reduction in processing time over the baseline system.

Key words: algorithm; Web site mining; multiscale site tree; context model; hidden Markov tree (HMT); multiscale

* Supported by the "Knowledge Innovation Initiative" of the Chinese Academy of Sciences under Grant No.Kgcxz-103 (中国科学院知识创新工程)

作者简介: **田永鸿**(1975－),男,四川德阳人,博士生,主要研究领域为数据挖掘,多媒体分析与检索;**黄铁军**(1970－),男,博士,研究员,主要研究领域为数字媒体,数字图书馆,智能交互,模式识别,图像处理;**高文**(1956－),男,博士,教授,博士生导师,主要研究领域为多模式用户接口,多媒体技术,人工智能.

classification; entropy-based pruning

摘 要: 随着 Web 所拥有的信息量和信息种类的急剧增长,Web 站点挖掘对于自动实现特定主题的 Web 资源发现和分类具有重要的意义.然而现有的 Web 站点分类或挖掘算法在利用上下文语义信息、去除噪声信息以进一步提高分类准确率等方面还缺乏深入研究.从站点的采样尺寸、分析粒度和描述结构 3 个方面分析了设计高效的 Web 站点挖掘算法所需要解决的问题.在此基础上,提出了一种新的 Web 站点多粒度树描述模型,并描述了包括基于隐 Markov 树的两阶段分类算法、粒度间上下文融合算法、两阶段去噪程序以及基于熵的动态剪枝策略在内的多粒度 Web 站点挖掘算法.站点的多粒度描述方法及挖掘算法为多站点查询优化、Web 效用挖掘等的深入研究奠定了基础.实验表明,该算法相对于基线系统平均可以提高 16%的分类准确率,并减少了 34.5%的处理时间.

关键词: 算法; Web 站点挖掘;多粒度站点树;上下文模型;隐 Markov 树;多粒度分类;基于熵的剪枝

中图法分类号: TP393 **文献标识码**: A

Web 已经成为科学研究、教育学习等领域最重要的信息源和知识库.随着 Web 所拥有的信息量和信息种类的急剧增长,Web 站点挖掘对于自动实现特定主题的 Web 资源发现、收集、分类和组织具有重要的意义[1].一般地,Web 站点挖掘包括如下两个过程:

- 搜索过程:给定初始主题集 T 或种子站点集,从一个无限大的信息集合 S^i 中,通过采用某种搜索策略 f_1,得到候选目标集 S^c,即: $f_1 : S^i \to S^c$,其中 $|S^i| >> |S^c|$.
- 分类过程:将未分类的候选目标集 S^i 映射到分类体系 C 中(该映射既可以是一对一映射,也可以是一对多映射),即: $f_2 : S^c \to S^l$,其中 S^l 为已分类的目标集,且 $|S^c| = |S^l|$.

由上可见,影响分类目标集 S^l 大小的主要因素为搜索过程,而影响 S^l 类别正确性的主要因素为分类过程.因此,为设计高效的 Web 站点挖掘算法,必须考虑如下 3 个问题:

1) 站点的采样尺寸问题:站点的采样尺寸对挖掘算法的效率有着重要影响.大部分分类算法都必须离线运行,同时下载一个远程站点网页的时间远远大于本地内存分类操作的时间[1],因此,在 Web 站点挖掘过程中必须采用某种采样算法来只下载尽可能少的页面却能达到相同甚至更高的分类准确率.

2) 站点的分析粒度问题:大部分已有的 Web 分类算法常以页面作为原子分析结点[2,3].但 Web 页面通常包含有多个主题的内容,并且很多页面都存在着噪声信息(如广告、导航条等),从而导致了页面的分类准确度较低.因此,Web 站点挖掘算法必须使用更小的分析粒度以便能够有效地去除噪声信息的影响,即需要将页面进一步分解为只包含单一主题的逻辑块[2,3],并以页内逻辑块作为挖掘的基本分析结点.最常用的逻辑块是 DOM(document object model)结点.在文献[2,3]中已探索了如何基于 DOM 结点来进行 Web 页的主题提炼(topic distillation)和信息抽取,本文将从站点、页面以及 DOM 结点 3 个层次来提出一种站点多粒度描述方法,以便于既能在多个粒度上对站点内容做去噪处理,又可以建模不同粒度分析结点间的主题相关关系.

3) 站点的描述结构问题:在 Web 站点分类算法中,通常有超页(super-page)[1,4]、主题向量(topic vector)[1]以及树[1]或图[5]描述等几种站点描述结构.相比较而言,树描述方式更适合刻画站点层次化的语义结构.因此本文用多粒度树(multiscale site tree)来作为站点的描述模型,其中页面之间以及页内 DOM 结点间都分别以树来描述其链接结构(分别称为页面树和 DOM 树),而不同层次的结点则用一棵多粒度树联系起来.同时,我们提出 4 种上下文模型来刻画多粒度站点树中结点间的主题相关关系.

基于上述站点多粒度树描述模型,本文提出一种新的多粒度 Web 站点挖掘算法.算法包括基于隐 Markov 树(hidden Markov tree,简称 HMT)的两阶段分类算法、粒度间上下文融合算法、两阶段去噪程序和基于熵的页面树动态剪枝策略 4 个部分.我们利用在信号处理领域提出的 HMT[6]来刻画页面树和 DOM 树中结点间的粒度内上下文相关关系(intrascale contextual dependencies).因此,在不考虑结点之间的粒度间上下文相关关系(interscale contextual dependencies)的情况下,可以直接将基于 HMT 的分类器分别应用到 DOM 树和页面树上,从而构成两阶段站点分类(two-phase site classification).为解决细粒度的 DOM 结点预分类过程中存在的过局部

化(over-localization)问题[7],我们提出粒度间上下文融合算法以便沿着粒度由粗到细的方向(coarse-to-fine),对两阶段站点分类所得的分类结果进行优化.这样,整个多粒度站点分类过程就包括了初始分类、粒度间融合以及为获得最终结果的再分类过程等 3 个步骤.同时,为进一步提高分类准确率并减少网页下载开销,我们还将利用两阶段文本去噪程序来去除站点内属于噪声信息的 DOM 结点和页面,以及利用基于熵的页面树动态剪枝策略来实现对站点的有效采样.算法的具体内容将在后面详细加以阐述.

我们通过在不同数量的种子站点集上执行实际的物理学科 Web 站点挖掘任务来评价算法的性能.作为性能比对的基线系统(baseline)只下载站点中小于固定深度的页面来代表站点的内容,同时采用基于超页描述方式的统计文本分类方法来进行站点分类.实验结果表明,本文所提算法相对于基线系统平均可以提高 16%的分类准确率,并减少了 34.5%的处理时间.

与文献[1,4,5]中描述的Web站点分类算法相比,本文的主要贡献是提出了站点的多粒度树描述模型和基于上下文的多粒度站点挖掘算法.已有的研究表明,文献[4]采用的超页方法由于噪声问题而性能较差[1],但由于其分类算法简单,因而通常用作构造基线系统.文献[5]采用基于图的 Web 站点描述以及基于超链的分类方法来对站点进行分类.实验表明,这种分类方法只适合于 Focused Crawling[8]等辅助任务.文献[1]的研究工作给了我们很大的启迪,但本文的研究与其有如下 4 个重要区别:(1) 本文采用多粒度树作为站点描述模型,并引入多粒度分类方法来求解Web站点分类问题;(2) 本文利用更广泛的上下文模型以使站点分类过程能够充分利用各种相关的语义信息;(3) 本文使用相对熵(relative entropy 或 Kullback Leibler Distance[9])而不是条件概率方差来构造站点剪枝策略;(4) 本文提出两阶段文本去噪程序从不同层次去除站点内的噪声信息,以便达到更高的分类准确率.

本文第 1 节提出了基于多粒度树的 Web 站点描述模型.第 2 节描述了包括基于 HMT 的两阶段分类和粒度间上下文融合的多粒度站点分类算法.第 3 节描述了两阶段去噪程序以及基于熵的动态剪枝策略等辅助挖掘算法.第 4 节是实验及其结果.本文最后是关于结论和进一步工作的讨论.

1 Web 站点的多粒度树描述模型

Web 站点的描述模型直接影响到分类或挖掘算法的有效性.超页方法[4]仅仅是简单地将 Web 站点看做是一些关键字或术语的集合,并直接将文本分类方法应用到站点和页面分类上,因而分类性能较差.类似地,主题向量方法[1]将站点描述为主题向量(其中每个主题各对应一个关键字向量),其实质仍是分两阶段应用基于关键字的文本分类方法来解决 Web 站点分类问题.另一方面,基于树的站点描述方法能够在分类过程中有效地利用站点的语义结构,并直观地将站点的采样尺寸问题转化为页面树的动态剪枝问题.但是在现存的基于树描述的站点分类方法中,页面的预分类仍然采用基于关键字的文本分类方法,因而同样无法消除页面内噪声信息的影响,从而影响了最终的站点分类准确度.基于此,本文将现存的树描述方法扩展到 DOM 结点层,提出一种新的Web 站点多粒度树描述模型.模型包括站点结构描述和上下文模型两部分.

1.1 站点结构描述

本文使用树作为站点描述中的基本数据结构.基于如下假设提出站点结构模型:

假设 1(树结构假设). 大多数 Web 站点是层次式而非网络式的结构[1];每个 HTML/XML 页面可以表示为一棵 DOM 树[2,3,10].

根据上述假设,Web 站点的结构定义如下:

定义 1(站点结构模型). 一个 Web 站点可以表示为一棵树 $T_P = T(P, E)$.其中:结点集 $P = \{p_1, ..., p_n\}$,树根 p_1 表示站点首页,任意结点 $p_i \in P$ 表示站点内的一个 HTML/XML 页面,它可以进一步表示为一棵 DOM 树,即 $p_i = DOM_i(DN, DE)$;页面 p_i 和 p_j 之间的链接用边 $(p_i, p_j) \in E$ 表示,称 p_i 为 p_j 的双亲结点, p_j 为 p_i 的孩子结点.因此,Web 站点可进一步表示为一棵多粒度树 $T_M = T(\{DOM_i(DN, DE)\}, E)$.

图 1 描述了这种多粒度树站点结构模型.整个 Web 站点由站点、页面以及 DOM 结点 3 种描述粒度构成,页面之间以及页内 DOM 结点间分别以页面树和 DOM 树来描述其链接结构,而联接不同粒度的结点则构成一

棵多粒度树.

本文采用广度优先搜索(breadth-first search)策略来构建站点的页面树,并且只采样站点内URL以域名开始的页面,而不考虑具有多域名的分布式站点采样问题[5].例如,若首页地址为http://phys.cts.nthu.edu.tw/en/index.php,则只下载拥有前缀http://phys.cts.nthu.edu.tw/en/的网页.

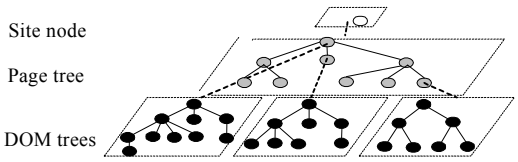


Fig.1 Represent the structure of Web sites as a multi-scale tree

图 1 基于多粒度树的站点结构描述

1.2 上下文模型

在超链环境下,链接包含了大量指示页面主题的语义信息,而这些语义信息能够有助于使分类过程达到更高的分类准确率[11].本文称这些语义信息为页面的上下文(context),并将其推广到站点和 DOM 结点的分类中.因此,根据 Web 链接拓扑结构(link topology)[5,12],我们有:

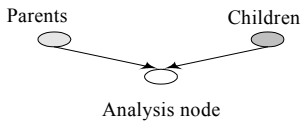


Fig.2 Two kinds of context nodes

图 2 两种上下文结点类型

假设 2(上下文假设). 在 Web 挖掘过程中,上下文是指与分析结点具有主题相关关系的结点和环境信息.上下文信息有助于改善对分析结点的分类准确性[7,12,13].

根据一阶 Markov 随机场论(Markov random field,简称 MRF)[13,14],在树结构中,结点的双亲和孩子结点都可以建模为其上下文(如图 2 所示).因此考虑上述多粒度树中不同层次结点可得的上下文信息和相应的主题相关关系,我们有如下 3 种上下文模型:

定义 2(站点层上下文模型 SC). 每个站点和与它紧密链接的站点间具有主题相关关系[5,14].令 N_i 为站点 S_i "紧密链接"的站点集,则 S_i 的站点层上下文 $SC(S_i)=N_i\approx N_i^K=N_i^I\cup N_i^O$,其中 N_i^K 表示 N_i 中已知分类站点集, N_i^I , N_i^O 分别为 N_i^K 中 S_i 的链入(in-neighbors)和链出(out-neighbors)邻居站点集[14],

$$P(C_{S_i}\mid\{C_{S'}\mid S'\neq S_i\})=P(C_{S_i}\mid\{C_{S'}\mid S'\in SC(S_i)\}) \tag{1}$$

其中 C 为隐藏的类变量.

特殊地,若 S_i 为一个孤立站点或 $N^K=\varnothing$,则 $P(C_{S_i}\mid\{C_{S'}\mid S'\neq S_i\})=P(C_{S_i})$,即无法根据站点的链接邻居信息来近似估计站点的类信息.

定义 3(页面层上下文模型 PC). 每个页面与其链入、链出页面之间具有主题相关关系[10,12,14].在页面树 T_P 中, p_i 的父结点 ρ_i 为 p_i 的链入页面, p_i 的子结点 $(p_{c_1},...,p_{c_{n_i}})$ 为 p_i 的链出页面,则 p_i 的页面层上下文 $PC(p_i)=\{\rho_i,p_{c_1},...,p_{c_{n_i}}\}$,其中

$$P(C_{p_i}\mid\{C_{p'}\mid p'\neq p_i\})=P\{C_{p_i}\mid C_{\rho_i},C_{p_{c_1}},...,C_{p_{c_{n_i}}}\} \tag{2}$$

定义 4(DOM 层上下文模型 DC). 每个 DOM 结点的主题和 DOM 树中的双亲、孩子具有主题相关关系[2,3].即 DOM 结点 DN_i 的 DOM 层上下文 $DC(DN_i)=\{\rho(DN_i),DN_{c_1},...,DN_{c_{m_i}}\}$,其中

$$P(C_{DN_i}\mid\{C_{DN'}\mid DN'\neq DN_i\})=P\{C_{DN_i}\mid C_{\rho(DN_i)},C_{DN_{c_1}},...,C_{DN_{c_{m_i}}}\} \tag{3}$$

另一方面,DOM 结点的预分类通常采用基于关键字的文本分类方法.由于 DOM 结点中所能提取的特征信息较少,同时缺乏必要的局部上下文信息,从而导致其分类准确率较低,并最终影响到站点的分类准确度.因此为了解决由于分析粒度过细而带来的过局部化问题[7],本文引入多粒度上下文模型,使得在分类过程中既能利用同一粒度中不同结点间的语义相关关系,又能利用不同粒度间的结点语义相关关系,从而使站点分类过程达到更高的准确率.因此,有如下定义:

定义 5(多粒度上下文模型 MC). 相邻分析粒度上对应结点间具有主题相关关系[7,13,15].例如,页面的主题类可以为其 DOM 结点的主题分类提供先验知识.在多粒度树 T_M 中,令 W_i^l 为第 l 层(l=2,3 分别为页面层和 DOM 结点层)的第 i 个结点, ρ_i^{l-1} 为其在 l–1 层的对应父结点,则 W_i^l 的多粒度上下文 $MC(W_i^l)=\{\rho_i^{l-1}\}$,其中

$$P(C_{W_i^l}\mid\{C_{W^{l'}}\mid l'<l\})=P\{C_{W_i^l}\mid C_{\rho_i^{l-1}}\} \tag{4}$$

不失一般性,我们把站点 S_i 的邻集 N_i 看做是第 0 层,则站点层上下文可以看做是多粒度上下文的一个特例.因此本文将站点层和多粒度上下文模型统称为粒度间上下文模型,相应地,称页面层和 DOM 层上下文模型为粒度内上下文模型.

综合上述站点结构模型和上下文模型,就构成了 Web 站点的多粒度树描述模型,图 3 给出了其图形描述.与站点结构模型不同的是,站点多粒度树描述模型使用隐藏树(hidden tree)[10]来建模结点间的主题相关关系,而不是直接建模结点间的链接关系,其定义如下:

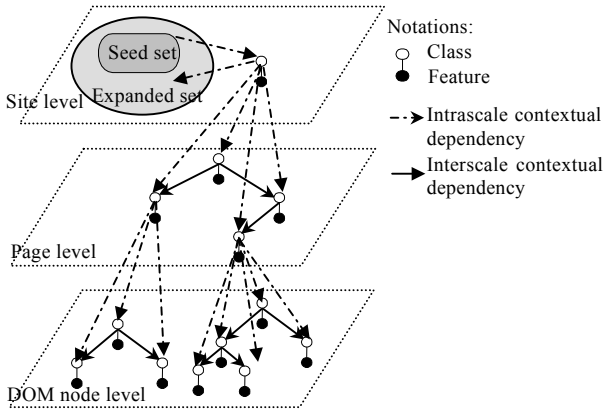


Fig.3 The multiscale tree representation model of sites

图 3 站点的多粒度树描述模型

定义 6(站点多粒度树描述模型 MST). 站点多粒度树描述模型表示为 $MST=(T_M,\{SC,PC,DC,MC\})$, 其中 $T_M=T(\{DOM_i(DN,DE)\},E)$为其结构模型,粒度内上下文模型 $\{PC,DC\}$ 建模了同层结点间的主题相关关系,粒度间上下文模型 $\{SC,MC\}$ 刻画不同层结点间的主题相关关系.

通过引入多粒度数据描述和多种上下文模型,站点的多粒度树描述模型为深入分析和挖掘 Web 站点奠定了基础.在 Web 站点挖掘中,采用站点多粒度树描述模型的好处有:(1) 能引入多粒度分类体系来求解 Web 站点分类问题,通过有效地利用粒度内和粒度间的结点主题相关关系来获得比使用单一粒度描述模型更高的分类准确率[7];(2) 能从 DOM 结点和页面两个层次来进行去噪处理,以进一步提高站点分类的准确率;(3) 可以通过页面树的剪枝算法来求解站点的采样尺寸问题.下面我们分别进行详细阐述.

2 多粒度 Web 站点分类算法

多粒度分类(multiscale classification)是信号处理和图像分割领域常用的一种分类方法[13,15,16].本文将其扩展到 Web 站点挖掘中来,以充分利用结点间的主题相关关系来进行站点分类.多粒度 Web 站点挖掘算法包括基于 HMT 的两阶段分类和粒度间上下文融合两个部分.

2.1 基于HMT的两阶段分类

在不考虑多粒度站点树中结点之间的粒度间上下文相关关系时,可以用一个基于树分类器的两阶段分类过程来实现 Web 站点分类,如图 4 所示.因此,两阶段分类算法的关键问题在于怎样选择页面树和 DOM 树的统计模型并设计相应的树分类器.

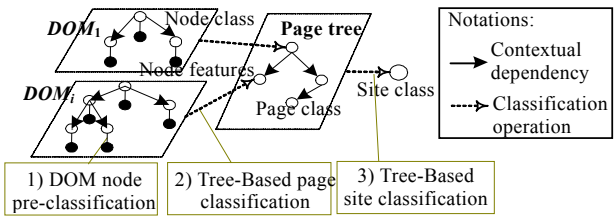


Fig.4 The two-phase site classification procedure

图 4 两阶段站点分类过程

根据上述 Web 站点描述模型,页面树或 DOM 树中双亲和孩子结点共同构成了分析结点的粒度内上下文.因此在为页面树和 DOM 树选择统计模型时,不能采用文献[1]中的一阶 Markov 树,而需要采用信号处理领域提出的隐 Markov 树模型(HMT)[6,16,17].图 5 给出了 Markov 树和 HMT 在建模结点间上下文相关关系上的区别:Markov 树仅以结点的双亲结点作为其上下文,结点间为因果相关关系;而 HMT 则建模双亲结点和孩子结点同时作为结点的上下文,结点间为非因果相关关系.即在 HMT 中,给定父状态 $C_{\rho(u)}$,结点对$\{C_u,W_u\}$独立于树中除 C_u 的孩子结点 $\{C_{c_1}^u,...,C_{c_{n_u}}^u\}$ 外的所有结点(称为 Markov 树特性[6]):

$$P(C_u \mid C_{u'} \neq C_u) = P(C_u \mid C_{\rho(u)}, C_{c_1}^u, ..., C_{c_{n_u}}^u) \tag{5}$$

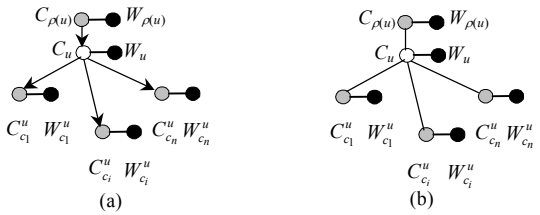


Fig.5 Dependency among class labels in the Markov tree and HMT models

图5 Markov树和HMT模型中的结点类相关关系

一个 HMT 模型λ可以由如下参数描述[6]:根状态结点 C_1 的初始分布 $\pi = (\pi_k)_{k\in\{1,...,K\}}$,状态转移矩阵 $A = (a_{\rho(u),u}^{rm})$(其中 $a_{\rho(u),u}^{rm} = P(C_u = m \mid C_{\rho(u)} = r)$),观察概率 $B = \{b_j(k)\}$.

类似于 HMM 模型,可以采用标准 EM 算法[6]来进行 HMT 的模型参数训练.但在本文的实验中,不同主题类别的训练数据其分布极不均匀,因此本文采用增量 EM 算法[18]来在每次挖掘过程结束后利用新获得的数据重新训练模型参数.

在此基础上,本文基于最大后验概率的原则(maximum a posteriori,简称 MAP)来构造基于 HMT 的分类器.令 W 为树中的全部观察数据,则分类器可以表示为

$$C_i = \arg\max_i P(C_i \mid W, \lambda) \tag{6}$$

利用 downward-upward 算法[6],可以很容易得到:

$$P(C_i = m \mid W, \lambda) = \frac{P(C_i = m, W \mid \lambda)}{P(W \mid \lambda)} = \frac{\beta_i(m)\alpha_i(m)}{\sum_{k=1}^{K}\beta_i(k)\alpha_i(k)} \tag{7}$$

其中 $\beta_i(k) = P(T_i \mid C_i = k)$ 为上向变量,$\alpha_i(k) = P(C_i = k, T_{1\setminus i})$ 为下向变量.

由于本文建模 HMT 树中结点的隐藏状态变量为结点对应的主题类标签,因此只需为页面树和 DOM 树分别训练一个 HMT 分类器,从而极大地简化了分类器的训练.另一种方式则是为页面树和 DOM 树的各类分别训练一个 HMT 模型,采用最大相似度(maximum likelihood,简称 ML)原则来构造分类器[16],在此不再赘述.

我们也可以直接将 HMT 分类器应用到以页面作为基本分析单位的站点分类中,此时,站点分类过程包括页面预分类和基于 HMT 的页面树分类两个步骤.

2.2 粒度间上下文融合算法

在使用 HMT 分类器来进行两阶段站点分类时,只利用了结点间的粒度内上下文.因此,下面将利用粒度间上下文模型来对两阶段站点分类的各层分类结果进行优化.

令 W_i^l 为第 l 层(l=1,2,3)的第 i 个结点,$MC_i^1 = SC(W_i^1)$(l=1)和 $MC_i^l = MC(W_i^l)$(l=2,3)为其粒度间上下文模型,$P(C_i^l \mid W_i^l)$ 为采用 HMT 分类器时结点的分类结果(当 l=3 时,则为 DOM 结点预分类结果,采用基于关键字的文本分类方法),则

$$P(C_i^l \mid W_i^l, MC_i^l) = \frac{P(W_i^l \mid C_i^l, MC_i^l)P(C_i^l \mid MC_i^l)}{P(W_i^l \mid MC_i^l)}$$

$$= \frac{P(W_i^l \mid C_i^l)P(C_i^l \mid MC_i^l)}{P(W_i^l \mid MC_i^l)} \text{(给定 } C_i^l, W_i^l \text{ 不依赖于 } MC_i^l \text{)}$$

$$= \frac{P(W_i^l)}{P(W_i^l \mid MC_i^l)} \cdot \frac{P(C_i^l \mid W_i^l)P(C_i^l \mid MC_i^l)}{P(C_i^l)}$$

$$= \frac{\alpha P(C_i^l \mid W_i^l)P(C_i^l \mid MC_i^l)}{P(C_i^l)} \text{(其中 } \frac{P(W_i^l)}{P(W_i^l \mid MC_i^l)} \text{ 可视为比例常数 } \alpha \text{ [14])} \tag{8}$$

其中, $P(C_i^l)$ 为第 l 层中类 C_i 的先验概率, $P(C_i^l \mid MC_i^l)$ 称为上下文概率(contextual probability).为简化参数估计问题,同一层中使用同样的参数 $P(C_i^l)$ 和 $P(C_i^l \mid MC_i^l)$.因此,若 $|C|$ 表示主题类的个数,则共需估计 $2\times3\times|C|$ 个参数.先验概率 $P(C_i^l)$ 可用训练集中第 l 层结点(l=1,2,3)中 C_i^l 类结点的相对频率来估计.因此,关键问题是需要进行上下文概率 $P(C_i^l \mid MC_i^l)$ 的估计.

在图像分割领域,上下文概率的估计方法有两种.最常见的是采用上下文标记树(contextual labeling tree,简称 CLT)来计算[16,19].CLT 是一个在观察结点上增加了一个上下文结点的树状结构.多个上下文(包括粒度间上下文和粒度内上下文)通过 CLT 可以依次与先前的分类结果相融合,最后获得更优的分类结果[19].另一种方法是类概率树(class probability tree,简称 CPT)[13].这种方法通过一系列测试或决策来得到上下文概率.

本文分如下两种情况来计算上下文概率:

1) 站点层的上下文概率 $P(C_i^1 \mid MC_i^1)$:根据站点层上下文, $MC_i^1 = SC(S_i) = N_i \approx N_i^K = N_i^I \cup N_i^O$,则

$$P(C_i^1 \mid MC_i^1) = P(C_i^1 \mid N_i^K) = \frac{P(N_i^K \mid C_i^1)P(C_i^1)}{P(N_i^K)} \tag{9}$$

其中, $P(N_i^K)$ 为站点的邻集 N_i 中已分类邻集 N_i^K 的相对频率,而 $P(N_i^K \mid C_i^1)$ 可以根据下式计算[14]:

$$P(N_i^K \mid C_i^1) = \prod_{\delta_j \in N_i^I} P(C_j \mid C_i^1, j \to i) \prod_{\delta_k \in N_i^O} P(C_k \mid C_i^1, i \to k) \tag{10}$$

2) 页面层和 DOM 层的上下文概率 $P(C_i^l \mid MC_i^l)$ (l=2,3):根据多粒度上下文, $MC_i^l = MC(W_i^l) = \{\rho_i^{l-1}\}$.我们可以采用文献[16,19]所描述的 CLT 树来计算页面层和 DOM 层的上下文概率.但为了减少计算的复杂性,本文直接使用 $P(C_i^{l-1} \mid \rho_i^{l-1})$ 来近似,即

$$P(C_i^l \mid MC_i^l) = P(C_i^{l-1} \mid \rho_i^{l-1}) \ (l=2,3) \tag{11}$$

当计算得到 $P(C_i^l)$ 和 $P(C_i^l \mid MC_i^l)$ 之后,融合了多粒度上下文模型的结点 W_i^l 的分类规则为

$$\hat{C}_i^l = \arg\max_i P(C_i^l \mid W_i^l, MC_i^l) \tag{12}$$

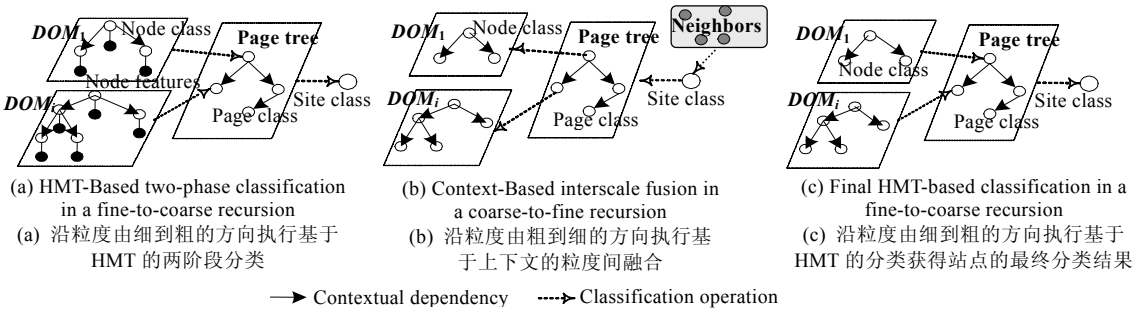
至此,我们已经完成了多粒度 Web 站点分类算法的讨论,算法分如下 4 步(如图 6 所示):

第 1 步.使用基于关键字的文本分类方法对站点的所有 DOM 结点进行预分类;

第 2 步.使用 HMT 分类器沿 Fine-to-Coarse 方向依次对页面和站点进行第 1 次粗分类(raw classification);

第 3 步.利用多粒度上下文模型对 HMT 分类器的各层分类结果沿 Coarse-to-Fine 方向依次进行优化,得到第 2 次分类结果;

第 4 步:再次使用 HMT 分类器沿 Fine-to-Coarse 方向依次对页面和站点进行第 3 次分类,得到最终结果.



(a) HMT-Based two-phase classification in a fine-to-coarse recursion
(a) 沿粒度由细到粗的方向执行基于HMT的两阶段分类

(b) Context-Based interscale fusion in a coarse-to-fine recursion
(b) 沿粒度由粗到细的方向执行基于上下文的粒度间融合

(c) Final HMT-based classification in a fine-to-coarse recursion
(c) 沿粒度由细到粗的方向执行基于HMT的分类获得站点的最终分类结果

Contextual dependency Classification operation

Fig.6 The multiscale Web site classification procedure

图 6 多粒度 Web 站点分类算法过程图解

3 去噪和剪枝

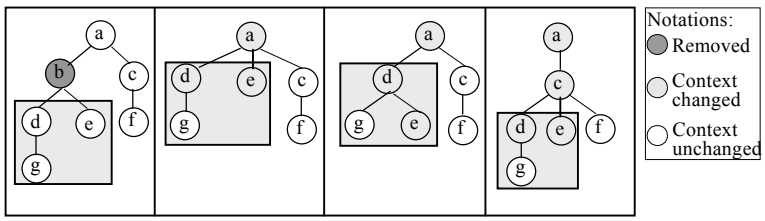
如引言所述,完整的 Web 站点挖掘算法除多粒度 Web 站点分类算法以外,还需考虑去噪和剪枝问题.

3.1 两阶段去噪

去噪的任务是去除与挖掘的主题无关或不能被文本分类器识别的 DOM 结点或网页,例如网站的动画前导页、框架(frames)、导航条以及广告[2,3]等.因此,很自然地可以利用基于文本的去噪方法.然而,当采用文献[2,3]所使用的中心向量文本去噪方法时,Web 内容的多样化不仅导致中心向量的维数很高,而且使 DOM 结点的关键字向量中的大多数元素为 0,因此本文采用基于主题词表的文本去噪方法,即只要给定主题的关键词或术语在 DOM 结点中出现的频率低于一定值,我们即认定该结点属于“噪声”.在本文的实验中,主题词表由一个具有 9181 个术语的物理学科主题词表(physics subject thesaurus)构成,同时我们从《物理文摘》杂志中提取大量专业关键词来充实主题词表.实验显示,在我们的应用环境下,主题词表法的性能要优于中心向量法.

在此基础上,本文采用基于多粒度站点树模型的两阶段去噪程序,即先用基于主题词表的文本去噪方法在 DOM 层删除属于“噪声”的 DOM 结点;如果一个页面的大部分 DOM 结点都被标记为“噪声”,则整个页面都将被删除.

如何重构去噪后的 DOM 树或页面树,使其尽可能地保持结点间的语义上下文关系,是我们在去噪处理中需要考虑的问题.图 7 例示了 3 种树重构方法.图中 b 为将被删除的“噪声”结点,则 3 种树重构方法分别为:将 b 的子树作为其父结点 a 的子树;以 b 的左孩子 d 代替 b;将 b 的子树作为其右兄弟结点 c 的子树.根据加权树编辑距离(edit distance of trees)[20]的思想,本文选择方法 1 作为去噪处理后 DOM 树和页面树的树重构方法.



(a) Original tree (a) 初始树 (b) 1st method (b) 方法 1 (c) 2nd method (c) 方法 2 (d) 3rd method (d) 方法 3

Fig.7 Three approaches for reconstructing the tree after denosing

图 7 3 种去噪后树重构方法

3.2 基于熵的剪枝

为进行站点采样而执行的页面树剪枝任务显得更为复杂.文献[1]利用一条路径上各站点类条件分布的方差来度量该路径对站点分类的重要程度,并提出了基于该方差值和路径长度的剪枝算法.为捕获多于二阶统计量的数据结构,本文提出基于相对熵或 KL 距离[9]的剪枝方法.方法基于如下假设:

假设 3(剪枝必要性假设). 在 Web 挖掘过程中,下载一个远程站点网页的时间远远大于本地内存分类操作的时间[1].

该假设可被如图 8 所示的实验结果所证实.假设不仅表明在 Web 挖掘过程中剪枝是极其必要的,更重要的是,为大量降低网页下载时间,我们可以通过适当增加本地内存操作以使总的挖掘时间有所优化.

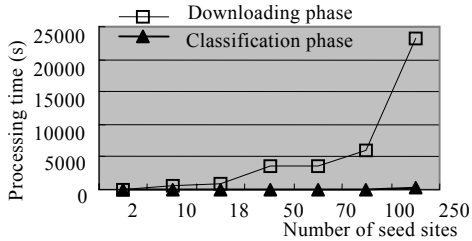


Fig.8 Compare processing time of the baseline system between downloading and classification phases

图 8 基线系统下载和分类阶段的处理时间比较

假设 4(样本尺寸假设). Web站点的分类需要下载描述其网页来提取分类特征.但当下载的网页数据超过一定量时,再下载更多的网页并不能显著地增加站点分类的准确度.

目前还没有更好的方法来衡量多少网页数就能"足够"对站点进行分类操作.直观上说,有时候站点页面树上的某个路径或子树上的网页并不显示任何清晰的类属性[1],因此,从它们中提取的特征不仅不能增加反而会降低站点分类的准确度.本文用 KL 距离来度量初始页面树和剪枝后页面树之间在概率分布上的差异:

令 u 为当前下载的网页,$T_{1:u}^{-}$ 和 $T_{1:u}^{+}$ 分别为增加结点 u 前后的页面树,$P(C_i \mid T_{1:u}^{-}, \lambda_P)$ 和 $P(C_i \mid T_{1:u}^{+}, \lambda_P)$ 分别为增加结点 u 前后的页面树相似度函数(λ_P 为页面层 HMT 树模型参数),则 $T_{1:u}^{-}$ 和 $T_{1:u}^{+}$ 之间的 KL 距离可以按下式计算[9]:

$$KL(T_{1:u}^{-};T_{1:u}^{+}) = \sum_i P(C_i \mid T_{1:u}^{-}, \lambda_P) \log \frac{P(C_i \mid T_{1:u}^{-}, \lambda_P)}{P(C_i \mid T_{1:u}^{+}, \lambda_P)} \tag{13}$$

因此,站点的页面树剪枝策略(如图 9 所示)为:检查在页面树中增加结点 u 是否会减少分类结果的不确定性.若是,则将结点 u 加入页面树中;不是,则把结点 u 及以它为根的子树都从页面树中剪去,即不再下载结点 u 的所有子孙结点.重复这个过程直到没有新的结点.即:

If ($KL(T_{1:u}^{-};T_{1:u}^{+}) \geq depth(T_{1:u}^{+}) \cdot \Delta$) then add u to the page tree, else suppress the growth of the tree to node u.

其中 $depth(T_{1:u}^{+})$ 为树 $T_{1:u}^{+}$ 的深度,Δ为页面层 HMT 模型训练的收敛参数.类似文献[1]描述的方法,随着 $depth(T_{1:u}^{+})$ 的增加,剪枝阈值越来越大.实验表明,这种基于熵的动态剪枝策略能有效地达到只下载一小部分站点内容即能获得相同甚至更高分类准确率的目的.

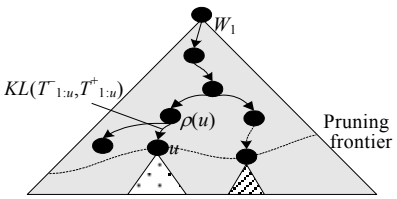


Fig.9 Pruning the page tree of a Web site

图 9 站点的页面树剪枝示意图

4 实验及结果分析

本文研究的目的是为国家科学数字图书馆开发特定学科的网络资源智能分析器 iExpert.当前,我们通过在不同数量的种子站点集上执行实际的物理学科 Web 站点挖掘任务来评价算法的性能.种子站点集包括 528 个

物理学科网站,其中每个网站的所有网页都已经被全部下载到本地,同时,这些网站也全部由领域专家按照物理主题分类法(physics subject classification)进行了分类标定.物理主题分类法包括 10 个大类、71 个小类,这些类别之间内容相互交叠、可区分度小,因而加大了分类任务的难度.

作为性能比对的基线系统采用超页作为站点描述模型,即将整个站点描述为其所有下载页面合成的一个“虚拟页”(称为超页).在此基础上,采用基于核加权 KNN 文本分类器来对站点的“超页”进行统计文本分类.基线系统没有剪枝功能,为减少下载数据量,只简单下载站点中小于 3 层的所有网页.同时,基线系统和本文的多粒度分类算法(简称 MSC)都采用同样的超链分析程序和网页抓取器.本文的所有实验都在如下环境内运行:800MHz CPU,256MB RAM,整个局域网共享 2MB 带宽.

Web 站点挖掘系统的主要评价指标包括系统的扩展能力(spotting capability)、分类准确率(classification accuracy)和时间开销.系统的扩展能力定义为新发现的该学科网站数 N_{new} 与种子站点数 N_{seed} 的比值,即

$$spotting = \frac{N_{new}}{N_{seed}} \times 100\% \tag{14}$$

在本文所有的实验中,种子站点数分别设定为 2,10,18,50,70,100 以及 250(种子站点同时用作该挖掘过程的训练样本).实验表明,基于 MSC 的 iExpert 的平均扩展能力为 187%.同时,本文使用分类准确率来作为 MSC 和其他站点分类方法的主要比较指标.在本文的实验中,分类准确率定义如下:

$$accuracy = \frac{1}{10}\sum_{i=0}^{9}\frac{M^{(i)}}{N_{new}^{(i)}} \times 100\% \tag{15}$$

其中, $N_{new}^{(i)}$ 是新发现的站点中属于第 i 类的站点个数,而 $M^{(i)}$ 是被正确分类的第 i 类的站点个数.

图 10 比较了给定不同种子站点数条件下各分类器的分类准确率,包括基线系统、MSC 算法、页面树 HMT 分类器、两阶段 HMT 分类器以及文献[1]描绘的 0 阶 Markov 树分类器.实验结论包括:

(1) 虽然 MSC 算法在小训练样本集时的分类性能还不尽如人意,但其平均分类准确率仍比基线系统高 16.7%.同时,基线系统采用的超页方法的平均分类准确率只达到 59%,因此只适合于辅助验证系统之用.

(2) 不同去噪方法对性能的影响:由于 MSC 算法和两阶段 HMT 分类器在 DOM 层和页面层上都进行了去噪处理,因此其平均分类准确率高于只在页面层进行去噪处理的页面树 HMT 分类器(约分别高 11%和 7%).这再一次证明了噪声信息是妨碍网站和网页分类准确率无法进一步提高的重要因素.

(3) 粒度间上下文融合对性能的影响:MSC 算法对 DOM 结点的预分类结果进行了修正,相对于两阶段 HMT 分类器,其分类准确率有约 3.9%的提高.

(4) 与零阶 Markov 树分类器比较:零阶 Markov 树分类器的平均分类准确率为 69.3%,分别比 MSC 算法和两阶段 HMT 分类器低 6.4%和 2.6%,但比页面树 HMT 分类器高 4.4%.零阶 Markov 树分类器采用页面树为其站点描述模型,只在页面层进行了去噪处理,因此页面内的噪声信息降低了其分类准确度;同时,它比页面树 HMT 分类器的训练参数要少,因此二者虽然采用同样的站点描述模型,但其模型训练误差较小.由于本文采用的分类体系中某些类别间的区分度较小,因此极容易造成误分现象,故在本文所述的实验中,零阶 Markvo 树分类器的分类性能要远远低于文献[1]中所描述的 87%.

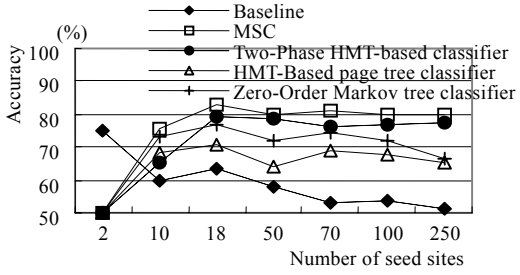


Fig.10 Compare the accuracy of several classifiers given different numbers of seed sites

图 10 不同种子站点数条件下各分类器分类准确率的实验比较结果

我们也比较了基线系统和基于 MSC 的系统在处理时间方面的开销.相对于基线系统,基于 MSC 的系统平均减少了 46.7%的下载时间(如图 11 所示),增加了 66.2%的分类时间,但在总的挖掘过程处理时间上降低了 34.5%.这也验证了假设 3,即可以通过增加本地内存操作的代价来换取总的处理时间有所优化.同时,通过比较采用基于熵的动态剪枝策略和固定下载 3 层网页时的挖掘算法性能(如图 12 所示),我们还发现,采用基于熵的剪枝策略的算法不仅没有显著增加总的下载数据量,还能在分类性能上提高约 5%.这是因为动态剪枝策略有目的性地对站点的特征提取过程进行了一定程度的控制,剔除了不能显著提高分类确定性的网页.同时也证明了基于熵的动态页面树剪枝策略是有效的.

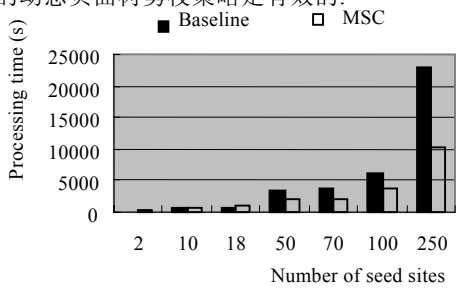


Fig.11 Compare the processing time of the baseline and the MSC-based systems on downloading phase
图 11 基线系统和 MSC 在下载阶段处理时间的比较

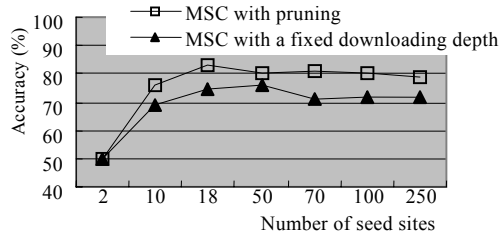


Fig.12 Compare the classification accuracy of the MSC algorithms with and without pruning
图 12 是否采用基于熵的剪枝策略时的分类准确率比较

总之,多粒度 Web 站点挖掘算法能在有效的处理性能下达到较高的分类准确率.不足之处在于,算法需要进一步提高在有限训练样本情况下的分类性能.

5 结 论

本文从站点的采样尺寸、分析粒度和描述结构 3 个方面引出问题,将现有的单一粒度的站点树描述方式扩展到更细粒度的 DOM 结点层,提出了一种新的 Web 站点多粒度树描述模型,并提出 4 种上下文模型来建模和刻画树中不同层次的结点在粒度内和粒度间的上下文相关关系.站点的多粒度描述方法,对于应用上下文信息、提高分类准确率以及针对 DOM 结点的去噪等奠定了基础,并对多站点查询优化、Web 效用挖掘等相关研究具有指导作用和应用价值.基于此模型,我们提出了多粒度 Web 站点挖掘算法,算法的核心包括基于 HMT 的两阶段分类、粒度间上下文融合、两阶段去噪程序以及基于熵的页面树动态剪枝策略.大量的实验表明,多粒度 Web 站点挖掘算法能在有效的处理性能下达到较高的分类准确率.

本文的一个重要特点是,将在信号处理和图像分割领域内提出的一些概念和方法,例如多粒度数据模型、上下文模型、去噪处理以及多粒度分类等,扩展到 Web 挖掘领域.在接下来的工作中,我们将进一步研究融合文本和多媒体特征的 Web 挖掘算法,以便能够更有效地从 Web 中发掘更有用的知识.

References:

[1] Ester M, Kriegel HP, Schubert M. Web site mining: A new way to spot competitors, customers and suppliers in the world wide web. In: Hand D, ed. Proc. of the SIGKDD 2002. Edmonton: ACM Press, 2002. 249~258.

[2] Chakrabarti S, Joshi M, Tawde V. Enhanced topic distillation using text, markup tags, and hyperlinks. In: Kraft DH, ed. Proc. of the 24th ACM-SIGIR Conf. on Research and Development in Information Retrieval. New Orleans: ACM Press, 2001. 208~216.

[3] Chakrabarti S. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In: Shen VY, ed. Proc. of the WWW 2001. Hong Kong: ACM Press, 2001. 211~220.

[4] Pierre JM. On the automated classification of web sites. Computer and Information Science, 2001,6(001).

[5] Terveen L, Hill W, Amento B. Constructing, organizing, and visualizing collections of topically related web resources. ACM Trans. on Computer-Human Interaction, 1999,6(1):67~94.

[6] Crouse MS, Nowak RD, Baraniuk RG. Wavelet-Based statistical signal processing using hidden Markov models. IEEE Trans. on Signal Processing, 1998,46(4):886~902.

[7] Li J, Gray RM. Context-Based multiscale classification of document images using wavelet coefficient distributions. IEEE Trans. on Image Processing, 2000,9(9):1604~1616.

[8] Chakrabarti S, Berg M, van den Dom B. Focused crawling: A new approach to topic-specific web resource discovery. Computer Networks, 1999,31(11-16):1623~1640.

[9] Minh ND. Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov model. IEEE Signal Processing Letters, 2003,10(4):115~118.

[10] Diligenti M, Gori M, Maggini M, Scarselli F. Classification of HTML documents by hidden tree-Markov models. In: Tombre K, *et al*, eds. Proc. of the Int'l Conf. on Document Analysis and Recognition (ICDAR 2001). Los Vaqueros: IEEE Computer Society Press, 2001. 849~853.

[11] Han JW, Chang KC. Data mining for web intelligence. IEEE Computer, 2002,35(11):64~70.

[12] Attardi G, Gulli A, Sebastiani F. Automatic Web page categorization by link and context analysis. In: Hutchison C, Lanzarone G, eds. Proc. of the THAI'99, European Symp. on Telematics, Hypermedia and Artificial Intelligence. Varese, 1999. 105~119. http://faure.isti.cnr.it/~fabrizio/Publications/THAI99.pdf

[13] Cheng H, Bouman CA. Multiscale Bayesian segmentation using a trainable context model. IEEE Trans. on Image Processing, 2001,10(4):511~525.

[14] Chakrabarti S, Dom B, Indyk P. Enhanced hypertext categorization using hyperlinks. In: Tiwary A, Franklin M, eds. Proc. of the SIGMOD'98. Seattle: ACMP Press, 1998. 307~318.

[15] Ye Z, Lu CC. Unsupervised multiscale classification using wavelet-domain hidden Markov tree model. In: Proc. of the 2002 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002). 2002. 4188~4191. http://www.cs.kent.edu/~zye/research.htm

[16] Choi H, Baraniuk R. Multiscale image segmentation using wavelet-domain hidden Markov models. IEEE Trans. on Image Processing. 2001,10(9):1309~1321.

[17] Romberg J, Choi H, Baraniuk R. Bayesian tree-structured image modeling using wavelet domain hidden Markov models. IEEE Trans. on Image Processing, 2001,10(7):1056~1068.

[18] Gotoh Y. Hochberg M, Silverman H. Efficient training algorithms for HMMs using incremental estimation. IEEE Trans. on Speech and Audio Processing, 1998,6(6):539~548.

[19] Fan G, Xia XG. Multiscale texture segmentation using hybrid contextual labeling tree. In: Proc. of the IEEE Int'l Conf. on Image Processing (ICIP 2000). Vancouver: IEEE Computer Society, 2000. 576~579. http://www.vcipl.okstate.edu/Publications/icip2000b.pdf

[20] Bertino E, Guerrini G, Mesiti M. Measuring the structural similarity among XML documents and DTDs. Technical Report, DISI-TR-02-02, Department of Computer and Information Sciences, University of Genova, 2001. http://www.disi.unige.it/person/MesitiM