智能 Web 浏览器及其关键技术

Intelligent Web Browser and Related Key Techniques

应晓敏　窦文华

（国防科技大学自动化研究所　长沙 410073）

Abstract　With the exponential growth of the Wold Wide Web, there is also a growing demand in intelligent Web browser (IWB), which can provide users with personal services like guiding users while searching the Web, filtering the information that users aren't interested in, notifying users when there are valuable changes in the Web sites or pages that users care, and so on. In this paper, we introduce some former researches, present the architecture of the IWB, describe its main functions, and discuss key techniques in the resarch of IWB.

Keywords　Intelligent web browser, Feature selection, Text categorization, User interests model, Information filtering

1　引言

互联网的飞速发展正在改变着人们的生活，WWW 已经成为人们交流和获取信息的重要媒介。1989 年起源于 CERN 的 WWW 已经发展成为拥有 8 亿页面的分布式信息空间，而且这个数字仍在以每 4 至 6 个月翻一倍的速度增加[1,2]。然而资源的极大丰富也带来了资源使用的困难，人们发现，在浩瀚的 Web 信息资源中查找和发现用户感兴趣的信息成为一件非常耗时耗力的事情。传统 Web 浏览器已经不能适应网络信息资源的迅速增长，主要表现在以下几个方面：

1）*没有考虑用户兴趣的差异*　尽管每个用户兴趣各异，但各 Web 站点提供的内容对任何用户都相同。传统的 Web 浏览器只是简单地接收用户的访问请求，与服务器交互后将传来的页面显示给用户，不能根据用户的兴趣提供有针对性的信息，用户不得不花大量的时间和精力从中找出自己感兴趣的信息。

2）*深度优先搜索*　传统 Web 浏览器的界面结构易于导致用户进行深度优先搜索，每点击一个超链接，该超链接所指向的链宿页面就马上显示出来。用户必须退回到上一级页面，才能搜索链源页面的兄弟链接。如果用户毫无目的和导引地随意浏览，则很可能"迷失在超空间"中[3]。

3）*缺乏信息过滤机制*　用户在网上搜索信息时，经常面对一个或多个超链接。在用户进入该超链接之前，无法确知该超链接所指向的链宿页面是否包含用户感兴趣的信息。这种情况尤其发生在使用搜索引擎搜索信息时，每当用户输入关键字，搜索引擎返回的结果往往成百上千，其中不乏大量重复、业已不存在和内容不相关的 URL，逐个浏览每一个 URL 指向的页面十分耗时耗力。围绕提高搜索引擎匹配精度和基于内容的检索问题已经开展了很多研究，综合集成此类技术的智能搜索引擎是目前研究的热点，但由于智能搜索引擎仍是一个主要运行于服务器端的通用系统，对用户信息知之甚少，其效果仍然不能令人满意。

4）*没有自动查新功能*　对于用户特别感兴趣的站点或页面，用户不得不经常访问以获得该站点或页面的最新信息。目前已经有少数几家网络公司和 Infomant、KARNAK 向用户提供站点查新服务，定期查找用户定义的站点，若发现新页面，则记录其 URL，将查新结果发到用户的电子信箱中。然而这种查新功能只是考察页面的增加，而不是页面内容的修改，而且用户仍需要到相应的站点去访问新页面。

为了提高浏览效率并满足个性化服务的需要，一些著名的学术机构先后开展了智能 Web 浏览器的研究，其目的在于使 Web 浏览器具有一定智能，即能够对不同兴趣类别的用户提供不同的服务和内容，实现智能导游、信息过滤、自动查新和主动服务等功能。

本文在国内外研究的基础上，通过剖析几种具有代表性的智能 Web 浏览器原型系统，提出智能 Web 浏览器的一般体系结构，对其中的关键技术进行了讨论，最后对智能 Web 浏览器的发展前景做了展望。

2　智能 Web 浏览器的体系结构与功能描述

智能 Web 浏览器作为用户与 Web 交互的中间媒介，较之传统 Web 浏览器在功能上有很大改进，能够

给用户提供个性化、智能的服务。其体系结构如图1所示。

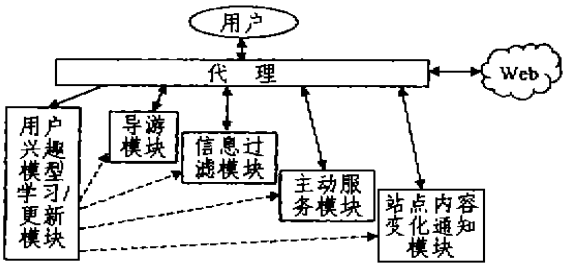

图1 智能Web浏览器的体系结构

1)*代理* 是用户、各功能模块和Web交互的中间媒介。它接收用户的访问请求，与Web交互，获取所请求的页面，并将URL、页面以及用户浏览行为交给用户兴趣模型学习/更新模块，最后将各功能模块输出的个性化页面提交给用户。

2)*用户兴趣模型学习/更新模块* 接收代理转来的URL、页面以及用户浏览行为，根据用户的浏览内容和浏览行为学习用户的兴趣，建立或更新用户兴趣模型，作为其它功能模块的基础。

3)*导游模块* 实现的功能类似于旅游团的导游，根据用户的兴趣和不同页面之间的相关性，指导用户沿着恰当的路径浏览Web。

4)*信息过滤模块* 接收用户定义的需求或原始页面，根据用户兴趣模型或用户定义的需求，找出用户不感兴趣的内容，通过修改HTML页面源码，将之从页面中过滤，并将过滤后的页面提交给代理。信息过滤功能可以将用户从大量不感兴趣的内容中解脱出来，直接浏览感兴趣的信息。

5)*主动服务模块* 接收用户定义的需求，或根据用户兴趣模型自动访问Web页面，将满足要求的页面保存在本地，供用户浏览。主动服务功能可以代替用户自动地从Web中搜索用户感兴趣的信息。

6)*站点内容自动查新模块* 代替用户定时访问用户关注的站点或页面，并与访问记录中该站点或页面的内容进行比较，若发现用户感兴趣的变化，如增加或修改了内容，则及时通知用户。

我们举例说明智能Web浏览器的工作过程。

假设某个用户经常上网看新闻栏目，但不喜欢体育栏目，而且该用户希望某几个站点的新闻能在他每天上网的时候已经下载到本地，以节省等待时间。此外，该用户在Web中搜索关于机器学习的论文，已经搜集了http://www.ics.uci.edu/MLpapers.html页面中的所有论文，并希望一旦该页面有新论文上传，就通知他访问。

当用户通过智能Web浏览器浏览Web时，浏览器通过分析用户的浏览内容和浏览行为，学习并记录用户兴趣，以构建和调整用户兴趣模型，及时反应并形式化地描述用户兴趣的变化。每当用户访问Web页面时，智能Web浏览器对所获取的页面进行兴趣匹配，若发现体育栏目，则通过修改页面的源码将体育栏目屏蔽。在用户浏览的同时，智能Web浏览器主动获取该页面中所有超链接指向的链宿页面，对页面进行兴趣分析，若发现新闻栏目或关于机器学习的内容，则在链源页面中该超链接上做上醒目标记，提醒用户浏览。对于用户预先设置要下载新闻栏目的Web站点，智能Web浏览器会自动代替用户访问，并把新闻栏目下载到本地保存，供用户浏览。对于用户设置需要自动查新的页面http://www.ics.uci.edu/MLpapers.html，智能Web浏览器会定期访问，并与保存的用户浏览记录比较，若发现页面内容有用户感兴趣的变化，则通知用户访问。

由此可见，使用传统浏览器时用户不得不花大量时间反复从事的操作可以由智能Web浏览器代为执行，从而节省浏览时间，提高浏览效率。

3 典型的智能Web浏览器原型系统

从1995年开始，CMU、MIT、UCI和Stanford大学先后开展了智能Web浏览器的研究，至今已研究开发了几种分别实现导游、主动服务和站点内容自动查新功能的原型系统。

1)**Personal Web Watcher(PWW)**[4] 是CMU开发的具有导游功能的智能Web浏览器原型系统。在用户浏览Web的整个过程中，PWW根据超链接的标记文本预测链宿页面的兴趣度，若属于用户感兴趣的页面，则在当前页面中的该超链接做上标记，建议用户访问。PWW同时记录其访问的页面和地址。在用户离线后，PWW分析用户访问的页面内容，学习、更新用户兴趣模型，指导用户的下一次浏览。PWW可以帮助用户从大量链接中找到自己感兴趣的链接，而不必像使用传统Web浏览器那样逐个浏览。此外，由于对用户浏览内容定期进行学习，PWW的用户兴趣模型可以随着用户兴趣的变化相应地调整。

2)**Letizia**[5] 是MIT开发的具有导游功能的智能Web浏览器原型系统。当用户浏览Web时Letizia跟踪用户的浏览行为，主动获取当前页面中所有超链接指向的链宿页面(即宽度优先搜索)，在分析页面内容后与用户兴趣模型比较，找出用户可能感兴趣的页面，在单独的窗口中显示推荐给用户的URL列表。

PWW采用的是基于内容的用户兴趣建模，即根据用户的浏览内容建立用户兴趣模型。与PWW不同

的是，Letizia 采用了一种基于行为的用户兴趣建模方法，即通过跟踪用户的浏览行为推测用户兴趣，建立用户兴趣模型。如用户保存某个页面，则推测用户对该页面的内容感兴趣；如用户跳过某一超链接，则推测用户对该超链接的标记文本不感兴趣。另外，Letizia 能够自动从用户当前页面开始进行宽度优先搜索，获取链宿页面，与用户兴趣模型比较后向用户推荐下一步的浏览目标，而 PWW 则只是根据超链接的标记文本预测链宿页面的兴趣度。

3）**Syskill&Webert**[5~7] 是 UCI 开发的具有导游和主动服务功能的智能 Web 浏览器原型系统。Syskill&Webert 要求用户为每一个兴趣构建一个目录页面，该页面包含指向同一兴趣、不同页面的超链接，用户从目录页面开始浏览 Web。当用户浏览 Web 页面时，Syskill&Webert 要求用户对页面做出"喜欢"或"不喜欢"的评价，而后保存页面和相应的评价。Syskill&Webert 分析评价后的页面内容，学习用户兴趣模型，并根据用户兴趣模型推荐用户可能感兴趣的超链接，具体的做法与 Letizia 相似，即主动获取当前页面中所有超链接指向的链宿页面（即宽度优先搜索），分析其内容后与用户兴趣模型比较，估算出用户可能感兴趣的概率。所不同的是，Syskill&Webert 直接将概率标示在链源页面的超链接前，而 Letizia 在单独的窗口显示推荐的链接。

此外，Syskill&Webert 还能依据学习到的用户兴趣模型构建查询语句，到搜索引擎 LYCOS 中查找满足用户兴趣的页面，并对返回的查询结果依据用户兴趣模型计算用户可能感兴趣的概率，标示在相应的 URL前，作为用户决定浏览与否的依据。Syskill&Webert 的突出特点是能够辅助用户查找长期感兴趣的信息，如跟踪某一领域的研究。

4）**LIRA**[1,2] 是 Stanford 大学开发的具有主动服务功能的智能 Web 浏览器原型系统。LIRA 代替用户搜索 Web，选择与用户兴趣模型相似度高的页面提交给用户，要求用户给出明确的评估值（－5 和＋5 之间的一个值），根据用户提供的相关反馈修改搜索和选择启发值，调整用户兴趣模型。LIRA 的突出特点是采用了人工智能中经典的启发式搜索算法搜索 Web，考虑到执行效率的问题，对搜索规模进行了限制，以使其能在规定的时间内中止。

5）**DICA**[3] 是 California 大学 Irvine 分校研究开发的具有页面内容自动查新功能的智能 Web 浏览器原型系统。DICA 从用户提供的样本中学习用户感兴趣的页面内容变化，而后定期访问用户指定的目标页面，确定自上次访问后发生的变化，若符合用户兴趣，则给用户发电子邮件通知用户访问。

4 智能 Web 浏览器的关键技术

智能 Web 浏览器的研究涉及计算机网络、人工智能、模式识别、机器学习等领域，主要关键技术包括特征提取、文档分类、用户兴趣模型的学习与更新、信息过滤等。

4.1 特征提取

Web 中的文档主要是 HTML 页面，去掉标记部分的 HTML 页面就是普通文档。（后续论文中的"文档"即指普通文档。）文档特征提取和文档表示主要针对普通文档。由于 HTML 文档的标记部分提供了有关文档结构的信息，如 title、head、字体等，有人提出利用标记确定文档特征[1,4,10]，但尚未取得明显效果。目前的智能 Web 浏览器在进行后续阶段处理前均将 HTML 文档转化为普通文档。

特征提取是文档表示、文档分类、构建用户兴趣模型和信息过滤的基础，是影响智能 Web 浏览器性能的重要因素。特征提取在信息检索领域被广泛使用并不断得到改进，较为成熟的方法主要有根据词频（包括通过词频计算权重）提取特征、根据互信息量（Mutual Information）提取特征、根据期望信息增益（Expected Information Gain）提取特征、根据期望交叉熵（Expected Cross Entropy）提取特征、根据几率比（Odds Ratio）提取特征等等。在智能 Web 浏览器的研究过程中，有些研究者提出一种与用户浏览行为相结合的特征提取方法——根据用户浏览行为提取特征。

目前典型的智能 Web 浏览器原型系统所采用的特征提取方法主要有四种：一是根据词频提取特征，如 LIRA 和 DICA；二是根据互信息量提取特征，如 PWW；三是根据期望信息增益提取特征，如 Syskill&Webert；四是根据用户浏览行为提取特征，如 Letizia。

1）*根据词频提取特征* LIRA 采用的特征提取方法是：对所有收集到的 Web 页面，去掉标记部分和停止字（stop list）（频繁出现且无实际意义的词，如英文中的 the）采用 porter Suffix-stripping 算法将所有同根词归为一个词根（如 computer，computers，computability 归为 comput），统计所有的词根，最后选出 27,000 个词根构成一个字典集 $D=\{d_i|_i=1,2,\cdots,27000\}$，$d_i$ 为字典集的词根；同时统计包含词根 d_i 的文档数目 $DF(i)$。

文档 doc 中每个词根 d_i 的权重 v_i 为

$$v_i=\frac{(0.5+0.5\frac{TF(i)}{TF_{max}})(\log\frac{n}{DF(i)})}{\sqrt{\sum_{d_j\in doc}((0.5+0.5\frac{TF(j)}{TF_{max}})^2(\log\frac{n}{DF(j)})^2)}} \quad (1)$$

其中，$TF(i)$是指词根 d_i 在文档 doc 中出现的次数，n 是文档总数，TF_{max}是指 $TF(i)$中的最大值，即 $TF_{max}=$

$\max\{TF(t), t=1,2,\cdots,27000\}$。出于对计算量的考虑，LIRA 选取 10 个权值最高的词根作为特征 f_t，$t=1,2,\cdots,10$。这样，每个文档都可以表示成特征矢量空间 $F=f_1\times f_2\times\cdots\times f_{10}$ 的一个矢量 $\vec{V}=(v_1,v_2,\cdots,v_{10})$，$v_t$ 为对应词根 d_t 在该文档中的权重。如果文档不包含 d_t，则 $v_t=0$。

2)根据互信息量提取特征　在文档特征提取中，可以通过计算词的互信息量确定该词能否作为文档的特征。PWW 先将 HTML 页面转化为普通文档并删除停止字，而后考虑文档中出现的词和类值，定义词的互信息量[11]为

$$MutualInfo(w)=\sum_i P(c_i)\log\frac{P(w/c_i)}{P(w)} \qquad (2)$$

其中，$P(c_i)$是第 i 类文档出现的概率，$P(w)$是词 w 出现的先验概率，$P(w/c_i)$是词 w 在 c_i 类文档中出现的后验概率。互信息量高意味着区分文档类别的能力强，由于 PWW 将文档只分为两类，即用户感兴趣类与用户不感兴趣类，所以词的互信息量高就意味着该词区分用户感兴趣文档和不感兴趣文档的能力强。毫无疑问，选取互信息量高的词作为特征可以比较准确地反映文档的含义。

3)根据期望信息增益提取特征　Syskill&Webert 的研究者希望提取在用户感兴趣的文档中出现频率高而在用户不感兴趣的文档中出现频率低的词作为文档特征。与 PWW 相同的是 Syskill&Webert 先将 HTML 文档转化为普通文档并删除停止字，所不同的是它采用计算期望信息增益的方法提取文档特征。期望信息增益定义为

$$ExpInfoGain(w,S)=I(S)-[P(w)I(S_w)+P(\overline{w})I(S_{\overline{w}})] \qquad (3)$$

其中，$I(S)=-\sum_i P(S_{c_i})\log_2(P(S_{c_i}))$，$c_i$ 为类别，S 是文档集，$P(w)$是词 w 出现的概率，$P(\overline{w})$是词 w 不出现的概率，S_w 是包含词 w 的所有文档，$S_{\overline{w}}$是不包含词 w 的所有文档，Syskill&Webert 根据计算的结果最后选取了 128 个词作为文档特征。

4)根据浏览行为提取特征　用户的浏览行为是用户兴趣的暗示，如果用户保存一个页面，则说明用户认为该页面的内容重要，用户感兴趣；如用户跳过某一超链接，则推测用户对该超链接的标记文本不感兴趣；如果用户在某一页面停留的时间长，说明用户对该页面内容感兴趣；等等。Letizia 根据用户的浏览行为推测用户感兴趣的文档，选取用户感兴趣的文档中的关键词为文档特征。

根据浏览行为提取特征的缺点是必须先对用户浏览行为建立准确的模型，用户浏览行为模型一旦出现偏差，提取的文档特征就不能正确地反映用户的兴趣，系统的性能会受到较大的影响。而且很难对用户浏览行为表现出的用户兴趣进行量化，即很难通过浏览行为确定用户对哪些内容更感兴趣，而对哪些内容兴趣一般。

4.2　用户兴趣模型的学习与更新

用户兴趣模型是指对于用户感兴趣的信息的可计算描述，是所有其它智能化功能的基础，用户兴趣模型的学习与更新是智能 Web 浏览器研究的核心内容。现有的原型系统采用兴趣相关反馈、Naive 贝叶斯分类器以及根据用户浏览行为学习用户兴趣模型。

LIRA 的用户兴趣模型表示为特征矢量空间下的矢量 $\vec{M}$，在系统初始时被置为 $\vec{0}$。LIRA 采用兴趣相关反馈更新用户兴趣模型，用户不断地浏览新的页面，并给出页面明确的评估值 e_i（[−5，+5]间的整数），$\vec{M}$ 随之调整。

$$\vec{M}\leftarrow\vec{M}+\sum_{i=1}^{10}e_i\vec{V}_i \qquad (4)$$

PWW、Syskill&Webert 和 DICA 均采用 Naive 贝叶斯分类器学习用户兴趣模型。所有用户浏览过的 HTML 页面和未浏览的超链接均表示成特征空间中的矢量作为学习用户兴趣的正负样本（Syskill&Webert 要求用户对页面做“喜欢”或“不喜欢”的评估），而后采用 Naive 贝叶斯分类器进行学习。用户兴趣模型的更新必须保留所有样本，以便重新学习。

Letizia 通过用户的浏览行为学习和更新用户兴趣模型，其规则包括：

1)如果用户保存某个文档，则表明用户对该文档感兴趣。

2)如果用户进入某一超链接，则表明用户对超链接的主题感兴趣。由于用户在进入链宿页面之前无法知道页面的内容，因此这种表示是试验性的。

3)如果用户没有保存链宿页面就马上返回，或者就进入更深一级链接，则表明用户对链宿页面不感兴趣。

4)如果用户反复地回到某一页面，则说明用户对该页面感兴趣。

5)假设用户的浏览习惯是从上到下，从左至右，若用户跳过某一链接，则说明用户对该链接不感兴趣。

4.3　文档分类

文档分类是研究文档的兴趣归属问题。如果用户兴趣模型只区分感兴趣和不感兴趣，则文档分类就是一个二类问题。目前所有原型系统均只区分用户感兴趣类和不感兴趣类，采用最小距离判别和 Naive 贝叶斯分类算法进行分类。

最小距离判别算法[9]的思想很简单：对于任意一个文档 $\vec{V}$，比较它与类别矢量 $\vec{C}_i$ 之间的相似度 $\cos(\vec{V},\vec{C}_i)$，文档属于相似度最高的那一类文档。LIRA 对最小距离判别算法做了进一步的简化，只将文档 $\vec{V}$ 与用户兴趣模型 $\vec{M}$ 比较，高值意味着属于用户感兴趣类，

否则为用户不感兴趣类。

采用 Naive 贝叶斯分类算法要考虑文档的表示方法。每一个需要预测的文档都表示成特征空间中的矢量，通过计算文档属于各类的概率确定文档的类别。假设文档所有的特征与特征之间不相关，如果文档表示成频率矢量(即每个分量是对应特征在文档中出现的频率)则给定文档 doc 属于类 c 的概率为

$$P(c/doc)=\frac{P(c)\prod_{f_j\in F}P(f_j/c)^{TF(f_j,doc)}}{\sum_i P(c_i)\prod_{f_j\in F}P(f_j/c_i)^{TF(f_j,doc)}} \quad (5)$$

其中，$P(f_j/c)=\frac{1+TF(f_j,c)}{|F|+\sum_i TF(f_i,c)}$，$TF(f_i,c)$是特征 f_j 在类 c 文档中出现的频率，$TF(f_j,doc)$是特征 f_j 在文档 doc 中出现的频率，$|F|$是文档表示中不同特征的总数目。PWW 和 DICA 均采用此法进行文档分类。

如果文档表示成布尔矢量(即每个分量为布尔值，1 表示对应特征在文档中存在，否则，表示对应特征不存在)则给定文档 doc 属于类 c 的概率为[31]

$$P(c/doc)=\frac{P(c)\prod_{f_j\in F}P(f_j/c)}{\sum_i P(c_i)\prod_{f_i\in F}P(f_i/c_i)} \quad (6)$$

其中 $P(f_j/c)=\frac{1+DF(f_j,c)}{2+DF(c)}$，$DF(f_i,c)$是 c 类文档中包含特征 f_j 的文档数，$DF(c)$是 c 类文档包含的文档总数。Syskill 和 Webert 采用此法进行文档分类。

Letizia 根据文档中是否包含用户兴趣模型中的关键字来判断该文档是否属于用户感兴趣的文档。

4.4 信息过滤

随着网络信息的指数增长，越来越多的用户希望能在浏览 Web 时过滤掉大量不相关的信息，信息过滤在 90 年代初开始受到关注。根据过滤文档方式的不同，过滤系统可以分为认知系统、社会系统和经济系统[8]。认知系统根据文档的内容进行过滤，社会系统根据他人的推荐过滤文档，经济系统通过某些价值衡量机制计算用户的代价和利益进行过滤。

智能 Web 浏览器所需要的个性化过滤应该满足三个要求：

1)*根据用户兴趣过滤* 不同的用户有着不同的兴趣，智能 Web 浏览器应该能学习不同用户的兴趣，并根据用户的兴趣进行过滤。

2)*能随着用户兴趣的变化相应地调整* 根据用户兴趣维持的时间，我们可以将用户兴趣划分为长期兴趣和短期兴趣。长期兴趣是指用户在一段相对长的时间里一直感兴趣的主题和内容；短期兴趣则指的是用户的"突发兴趣"。但需要指出的是，无论是长期兴趣还是短期兴趣，都会随着时间的推移而改变。因此，智能 Web 浏览器应该能察觉用户兴趣的变化并作相应的调整，以提供准确的服务。

3)*能发现满足用户潜在兴趣的新信息*[9] 如果仅仅根据学习到的用户兴趣对用户浏览内容进行过滤，则可能出现用户浏览的内容范围越来越集中的现象。智能 Web 浏览器应该能从学习到的用户兴趣中推测用户的潜在兴趣，在过滤时保留相关的内容以供用户浏览。

结束语 经过几年的研究，已经出现了一些在某一方面具有较好功能的智能 Web 浏览器原型系统，但距离实用化的要求还有相当的距离。智能 Web 浏览器是人工智能、心理学、网络技术等多学科相互交叉的产物，要使之真正符合个性化的要求，还需要深入研究人的认知心理和行为心理，并充分借鉴智能科学已有的大量研究成果，其最终实现必须依靠各学科研究人员的共同努力。

如果说智能搜索引擎是信息海洋中的导航灯的话，那么智能 Web 浏览器就是在 WWW 中冲浪的飞舟，用智能信息处理技术改造传统 Web 浏览器不仅是一种技术上的尝试，更是 Web 信息资源急速膨胀的必然要求。随着广大网络用户对个性化服务需求的日益增长，相信会有更多的有识之士投入到这项研究中来。

参 考 文 献

1 Balabanovic M, Shoham Y. Learning information retrieval agents: experiments with automated Web browsing. In: Proc of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Resource. Stanford, CA, March 1995

2 Balabanovic M, Shoham Y. An adaptive Agent for automated Web browsing. Journal of Visual Communication and Image Representation, 1995, 6(4)

3 Lieberman H. Letizia, an Agent that assists Web browsing. In: Int Joint Conf. on Artificial Intelligence. Montreal, August 1995

4 Mladenic D. Personal WebWatcher: design and implementation: [Technical Report IJS-DP-7472]. Department for Intelligent System, J. Stefan Institute

5 Pazzani M, Muramatsu J, Billsus D. Syskill&Webert: identifying interesting web sites. In: AAAI Conf. Portland, August 1996

6 Ackerman M, et al. Learning probabilistic user profile. AI Magazine, summer 1997. 47～56

7 Pazzani M, Billsus D. Learning and revising user profiles: The identification of interesting Web sites. Machine Learning, 1997, 27: 313～331

8 Sheth B D. A learning approach to personalized information filtering: [MS thesis]. MIT. Feb. 1994

9 Joachims T. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In: Int Conference on Machine Learning97, 1997. 143～151

10 Culter M, Shih Y, Meng W. Using the structure of HTML documents to improve retrieval. USENIX Symposium on Internet Technolgies and Systems (NSITS'97). Monterey, California, Dec. 1997. 241～251

11 王伟强，高文，段立娟. Internet 上的文本数据挖掘. 计算机科学，2000，27(4)：32～36

12 王继成，潘金贵，张福炎. Web 文本挖掘技术研究. 计算机研究与发展，2000，37(5)，513～520