主题网络爬虫研究综述

刘金红，陆余良
（解放军电子工程学院 网络系，合肥 230037）

摘 要：首先给出了主题网络爬虫的定义和研究目标；然后系统分析了近年来国内外主题爬虫的研究方法和技术，包括基于文字内容的方法、基于超链分析的方法、基于分类器预测的方法以及其他主题爬行方法，并比较了各种方法优缺点；最后对未来的研究方向进行了展望。

关键词：主题网络爬虫；信息检索；Web 挖掘
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2007)10-0026-04

Survey on topic-focused Web crawler

LIU Jin-hong, LU Yu-liang
(*Dept. of Network, PLA Electric Engineer Institute, Hefei* 230037, *China*)

Abstract: This paper gave the goal of focused crawling, then comprehensively analyzed the recent advances of the relevant researches and applications about focused-crawler, included focused crawling methods based on text contents, link analyses' methods, classifier-guided methods and other focused methods. Finally pointed out the future direction of focused crawling.
Key words: topic-focused crawler; information retrieval; Web mining

0 引言

随着网络上海量信息的爆炸式增长，通用搜索引擎面临着索引规模、更新速度和个性化需求等多方面的挑战[1,2]。面对这些挑战，适应特定主题和个性化搜索的主题网络爬虫(focused crawler or topical crawler)应运而生[3,4]。基于主题网络爬虫的搜索引擎(即第四代搜索引擎)已经成为当前搜索引擎和 Web 信息挖掘中的一个研究热点和难点。

通用网络爬虫的目标就是尽可能多地采集信息页面，而在这一过程中它并不太在意页面采集的顺序和被采集页面的相关主题。这需要消耗非常多的系统资源和网络带宽，并且对这些资源的消耗并没有换来采集页面的较高利用率。主题网络爬虫则是指尽可能快地爬行、采集尽可能多的与预先定义好的主题相关的网页。主题网络爬虫可以通过对整个 Web 按主题分块采集，并将不同块的采集结果整合到一起，以提高整个 Web 的采集覆盖率和页面利用率。

1 主题爬虫的定义和研究目标

定义 1 网络爬虫是一个自动提取网页的程序，它为搜索引擎从 Web 上下载网页，是搜索引擎的重要组成部分。通用网络爬虫从一个或若干初始网页的 URL 开始，获得初始网页上的 URL 列表；在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入待爬行队列，直到满足系统的停止条件。

定义 2 主题网络爬虫就是根据一定的网页分析算法过滤与主题无关的链接，保留主题相关的链接并将其放入待抓取的 URL 队列中；然后根据一定的搜索策略从队列中选择下一步要抓取的网页 URL，并重复上述过程，直到达到系统的某一条件时停止。所有被网络爬虫抓取的网页将会被系统存储，进行一定的分析、过滤，并建立索引，对于主题网络爬虫来说，这一过程所得到的分析结果还可能对后续的抓取过程进行反馈和指导。

定义 3 如果网页 p 中包含超链接 l，则 p 称为链接 l 的父网页。

定义 4 如果超链接 l 指向网页 t，则网页 t 称为子网页，又称为目标网页。

主题网络爬虫的基本思路就是按照事先给出的主题，分析超链接和已经下载的网页内容，预测下一个待抓取的 URL 以及当前网页的主题相关度，保证尽可能多地爬行、下载与主题相关的网页，尽可能少地下载无关网页。相对于通用网络爬虫，主题网络爬虫需要解决以下四个主要问题：

a) 如何描述或定义感兴趣的主题(即抓取目标)？

b) 怎样决定待爬行 URL 的访问次序？许多主题网络爬虫根据已下载网页的相关度，按照一定原则将相关度进行衰减，分配给该网页中的子网页，而后将其插入到优先级队列中。此时的爬行次序就不是简单地以深度优先或广度优先顺序，而是按照相关度大小排序，优先访问相关度大的 URL。不同主题网络爬虫之间的区别之一就是如何计算 URL 的爬行次序。

c) 如何判断一个网页是否与主题相关？对于待爬行或已下载的网页可以获取它的文本内容，所以可以采用文本挖掘技术来实现。因此不同主题网络爬虫间的区别之二就是如何计算当前爬行网页的主题相关度。

收稿日期：2006-08-16；**修返日期**：2006-12-06
作者简介：刘金红(1978-)，山西永济人，博士研究生，主要研究方向为 Web 挖掘和网络安全(happygold@sina.com)；陆余良(1964-)，江苏宜兴人，教授，主要研究方向为数据挖掘和网络安全。

d)怎样提高主题网络爬虫的覆盖度？如何穿过质量不好(与主题不相关)的网页得到与用户感兴趣主题相关的网页，从而提高主题资源的覆盖度？

对于主题网络爬虫性能的评价，目前主要是基于 harvest rate 来评价。Harvest rate 就是主题相关网页数目占所有抽取网页总数的比率：

$$\text{harvest rate} = \text{numbers of relevant pages}/\text{numbers of all retrival pages} \quad (1)$$

2　主题网络爬虫研究进展

为了高效地抓取与主题相关的网络资源，研究者提出了许多主题定制爬行策略和相关算法，使得网络爬虫尽可能多地爬行主题相关的网页，尽可能少地爬行无关网页，并且确保网页的质量。通过对这些方法进行比较分析，本文将它们分为如下四类。

2.1　基于文字内容的启发式方法

基于文字内容的启发策略主要是利用了 Web 网页文本内容、URL 字符串、锚文字等文字内容信息。不同的分析方法构成了不同的启发式策略和相应的算法。主要包括：

a)Best first search 方法。基本思想是给定一个待爬行 URL 队列，从中挑选最好的 URL 优先爬行。爬行主题采用关键词集合来描述，待爬行 URL 的优先级是根据主题词和已爬行网页 p 的文字内容来计算，用它们的相关度来估计 p 所指向网页的相关度。相关度大的网页，它所指向的网页优先级就高，从而决定了待爬行队列中 URL 的优先级顺序。如果待爬行队列的缓冲区满了，则将优先级最低的 URL 从该队列中移去。它采用式(2)来计算网页与主题间的相关度。

$$\text{sim}(q,p) = \left(\sum_{k\in q\cap p} f_{kq} f_{kp}\right) \Big/ \left(\sqrt{\sum_{k\in p} f_{kp}^2}\ \sqrt{\sum_{k\in q} f_{kq}^2}\right) \quad (2)$$

其中：q 表示主题；p 表示抓取的网页；f_{kq} 表示词 k 在 q 中出现的频次；f_{kp} 表示词 k 在 p 中出现的频次。

该算法有 url_queue 和 crawled_queue 两个堆栈，分别用来存储待爬行的 URL 和已爬行的 URL。在主题网络爬虫研究领域，该算法具有一定的竞争力，所以很多研究者将其作为算法性能的比较基准[6]。J. Cho 等人[7]将待爬行队列分成两个，即 hot_queue 和 url_queue。如果认为是相关的，就将其压入队列 hot_queue；否则就压入 url_queue 中。在爬行时，优先爬行 hot_queue 队列，只有当 hot_queue 队列为空时，才爬行 url_queue。在判断是否相关时，不是采用式(1)，而是作如下判断：如果在 URL 字符串或其对应的锚文字中含有主题词，就认为是 hot_queue，将其压入 hot_queue；否则就是普通的 URL，将其压入 url_queue。这种方式的优点是计算量小，对于单个关键词的主题(宽泛主题)，它的效果还是不错的。但当关键词个数较多时，效果就不是很好。因为 URL 的文字和锚文字并不能很好地反映出多个关键词蕴涵的真实主题。

b)Fish search 方法。1994 年由学者 De Bra 等人[8]提出。它将在网络上遍历的网络爬虫比喻成海里的一群鱼。当它们发现食物(相关信息)时，这些鱼就继续繁殖，寻找新的食物；当没有食物时(没有相关信息)或水被污染(带宽不够)时，它们就死掉。该算法的关键是根据代表用户感兴趣主题的种子站点和主题关键词，动态地维护待爬行的 URL 优先级队列。

当一个网页抓取过来后，抽取它所有的 URL，这些 URL 所对应的网页，称为其孩子网页。如果抓取的网页相关，孩子网页的深度(depth)设成一个预先定义的值；否则孩子网页的深度设置成一个小于父亲网页深度的值。当这个深度为零时，该方向的搜索就停止。

Fish search 算法的入口参数包含种子站点、查询式(主题)、查询宽度 width、深度 depth。深度大于 0 的孩子网页的 URL 按照如下启发策略插入到 url_queue 中：

a) 相关网页的前面 $\alpha \times$ width 个孩子(α 是预定义的大于 1 的常量)加入到 url_queue 的顶部；

b)无关网页的前 width 个孩子 URL 加入到 url_queue 队列中，紧靠着相关网页孩子节点的后面；

c)剩下的孩子 URL 加入到 url_queue 的尾部(即只有在时间允许的情况下，才有可能爬行到)。

上述三种情况可以用一个变量 potential_score 来等价描述它们。第一种，potential_score 设置为 1；第二种设置为 0.5；第三种设置为 0，待爬行 URL 队列就按照 potential_score 来排序。

该算法是一种基于客户端的搜索算法。因为其模式简单、动态搜索，有一定的吸引力，但存在如下缺点：只使用简单的字符串匹配分配 potential_score 的值，并且该值是离散的(只有 1、0.5 和 0 三种)；分配的值并不能完全代表与主题的相关度。在 url_queue 中，优先级值之间的差别太小。当很多 URL(节点)具有相同的优先级并且在爬行时间受到限制时，可能后面更重要的网页被忽略掉了。另外，使用 width 参数来调节删除网页后面 URL 的个数也有点过于武断，可能导致丢掉很多主题相关的重要资源。

c)Shark search 方法[9]。它在 fish search 算法的基础上进行了如下改进：

(a)在 fish 算法中，只有是否相关的二值判断，而在 shark 算法中引入了相似度度量方法，取值为 0～1。

(b)在计算 URL 的 potential_ score 上，不但继承了双亲的值，而且充分利用了锚文字和锚文字的上下文。锚文字的上下文是指围绕在锚文字周围一定距离的文字。

与 fish 算法相比，shark 算法精度更高，能更好地保证爬行器正确的搜索方向，提高相关信息的发现率。

2.2　基于 Web 超链图评价的方法

基于文字内容的算法只是利用网页、URL、锚文字等文字信息，没有考虑到通过超链而形成的 Web 有向图对主题网络爬虫的影响。基于 Web 图的启发策略的基本思想来自于文献计量学的引文分析理论。尽管引文分析理论的应用环境与 Web 并不相同，但到目前为止，网页之间的超链还是比较有价值的一种信息。基于 Web 超链图评价的爬行算法有以下几种：

a)BackLink。一个网页被其他网页所引用的次数越多，就说明越重要。待爬行 URL 队列按照 BackLink 的数量来排序，数量大的优先爬行。

b)PageRank。基于 Web 图，按照式(2)来计算每个网页的 PageRank 值，然后对待爬行 URL 队列按照 PageRank 的值进行

排序。

PageRank 算法是由 Google 的创始人 S. Brin 和 L. Page 提出的,它是一种与查询式无关的算法[10]。在 PageRank 算法中,一个网页的链入网页数量越大,它的重要性就越大,而没有考虑入链的质量问题。实际上,不同质量的网页对网页重要性的贡献是不同的。简单地说,按照 BackLink 算法,要想提高某网页的重要性,只要建立许多网页指向它就可以了。S. Brin 和 L. Page 提出的 PageRank 算法就是为克服 BackLink 的这种不足而设计的。这种计算网页权威度的具体计算方法如下:

$$R(i)=(1-d)+d\times\sum_{j\in B(i)}[R(j)/N(j)] \tag{3}$$

其中:$B(i)$是指向网页 i 的网页集合;$N(i)$表示网页 i 中指向其他网页的超链数目;$R(i)$表示网页 i 的权威度。$d(0<d<1)$是一个衰减因子,表示每个网页本身的权威度有$(1-d)$。它是不用于传递的,所以每个网页实际用于传递的只是该网页权威度的 d 部分,也即该网页权威度的 d 部分被平均传递给该网页的所有链出网页。S. Brin 和 L. Page 通过实验认为 d 的最佳值为 0.85。

每下载一个网页后,抽取其中包含的超链,Web 图都需要作相应的改变,相应的优先级也需要重新计算。基于 Web 图的启发策略计算量一般都很大,所以在实际的网络爬虫设计中,需要进行缓冲,只有下载网页的数量达到设定的阈值后,才重新计算待爬行队列的优先级值。

基于 Web 超链图评价的启发策略有如下缺点:存在很多的导航用超链,顺着这些超链并不能发现更多的主题资源;PageRank 更适合发现权威网页,而不适合发现主题资源;基于图的启发策略的计算量一般都很大,严重影响了爬行器的爬行速度。

2.3 基于分类器预测的方法

为了克服基于文字内容难以精确描述用户感兴趣的主题,以及基于 Web 超链图分析的低效率,研究者提出了基于分类器导引的主题网络爬虫[11],从而可以基于分类模型来描述用户感兴趣的主题和预测网页的主题相关度。通过文本分类模型可以从更深的层次来描述用户感兴趣的主题信息,并可以更加准确地计算网页的主题相关性,而不只停留在基于关键词的匹配上。文本分类技术应用于主题信息搜索中有利于提高主题搜索的正确率和准确率。有关实验结果[12~14]表明,使用主题分类器来指导网络爬虫爬行主题相关网页的效果要好得多。

Chakrabarti 等人[15]叙述了一些有关主题网络爬虫的实验。他们的研究目标包括网页的链接关系和半监督式学习等。该文中网络爬虫使用规范的主题分类,从用户指定的起点(书签)开始。用户将感兴趣的网页做上标记,并将它们归类,如同 Yahoo 的目录层次结构。其主题网络爬虫的主要组成部分包括分类器、过滤器和爬行器。分类器对网页作相关性判断来决定访问哪些链接;过滤器决定访问网页的优先级;爬行器是基于链接的分析。其中使用的评价指标称为获取率,获取率表示的是获取相关网页的频率,以及如何有效过滤不相关网页。

Chakrabarti 等人[16]提出了分别基于两种不同的模型来计算网页主题相关性和 URL 访问次序。计算网页主题相关性的模型可以是任何二值分类器,而计算 URL 访问次序的模型(简称为 apprentice)是通过包含父网页和子网页及其相关度的训练样本集合在线训练得到的。对于每一个抽取的网页,apprentice 模型根据基本分类器(即二值分类器,以此确定父网页属于某一类别的概率)和父网页链接周围的特征进行训练,以此来预测父网页指向网页的主题相关度。然后基于这些相关度预测信息对待爬行队列中的 URL 进行排序。实验结果表明,爬行错误网页的数目极大地减少了(大约减少了 30% ~ 90%)。

傅向华等人[12]将 Web 爬行看做执行序列动作的过程,结合改进的快速 Q 学习和半监督贝叶斯分类器,提出了一种新的具有在线增量自学习能力的聚焦爬行方法。该方法从获取的网页中抽取特征文本,根据特征文本评估网页的主题相关性,预测链接的 Q 值,然后基于 Q 值过滤无关链接。当得到主题相关网页时产生回报,然后将回报沿链接链路反馈,更新链路上所有链接的 Q 值,并选择相应的特征文本作为训练样本,增量地改善主题评估器和 Q 值预测器。实验结果表明,该方法具有很快的自学习能力,获取的网页数目和精度均优于离线聚焦爬行方法,更符合 Web 资源发现的要求。

李盛韬等人[17]在分析主题 Web 信息采集基本问题的基础上,提出了主题网络爬虫的难点以及相关的解决方案,并在此基础上设计实现了"天达"主题 Web 信息采集系统。李卫等人[18]以全信息理论为支撑,吸收传统向量空间模型的思想,采用基于概念的向量空间模型,从词的语义层次对文本进行主题相关性分析,研究并实现了一个基于主题的智能信息采集系统 IFWC。其使用扩展元数据的语义相关性判定算法,对页面内的 URL 进行主题相关性预测。

目前国内外对于基于主题分类器来引导主题 Web 爬虫的研究还非常少[13]。S. Chakrabarti 等人[11]第一次提出基于朴素贝叶斯分类模型引导主题 Web 爬虫。Johnson 等人[19]提出了基于 SVM(support vector machine,支持向量机)分类模型来进行主题爬行。文献[12,14]中通过实验对比表明,基于线性 SVM 分类模型的主题 Web 爬虫要比朴素贝叶斯分类模型的主题爬行效果好很多。

2.4 其他主题爬行方法

J. Cho 等人[7]提出通过先爬行更重要的网页使得爬行更有效,而提出了各种计算网页重要性的方法,如网页与查找项的相关性、指向该页的网页个数(BackLinks)、该网页的 PageRank值和该网页所处的位置。一个网页的 PageRank 被递归地定义为指向该网页所有网页的 PageRank 权值之和。他们使用了基于这些重要性测量的排序机制实现了网络爬虫。该网络爬虫对斯坦福大学的网站(大约 225 000 个网页)进行了测试。他们发现使用 BackLink 重要性测量的网络爬虫类似深度优先查找,在继续访问前先访问特定簇里的网页,并经常受起始点左右。使用 PageRank 的网络爬虫却不受起始点影响,并结合了宽度优先和广度优先,能更好地爬行。同时也发现在较小的网页域中(如网页的子集中),使用 BackLink 计数的网络爬虫执行效果不好。

文献[5]中对不同的主题爬行策略进行了评价,它为每一百个主题建立一个分类器,用于评价已爬行过的网页。作者认

为一个好的主题网络爬虫在向量空间中也应该保证主题相关。他们画了一个随时间变化的抛物线图，根据网络爬虫是否能保持主题相关来评价一个网络爬虫的性能。文中评价了三种不同爬行策略的网络爬虫：

a) BestFirst 网络爬虫。根据主题和网页的相关性排列 URLs 的优先级队列。

b) PageRank 网络爬虫。以 PageRank 的次序排列 URLs，每 25 个网页重新计算网页的 PageRank 值。

c) InfoSpiders。采用后向传播的中枢神经网络，考虑链接周围的文本。

最后实验结果发现，BestFirst 执行效果最好，其次是 Infospiders，最后是 PageRank。研究者认为 PageRank 在主题爬行任务中过于全面，所以效果不好。

M. Diligenti 等人[20]提出了使用上下文图表来引导主题网络爬虫。作者认为该领域主要的问题是在爬行中把任务分配给网页。例如，一些偏离主题的网页指向符合主题的网页，结果形成页层次结构。基于这些问题，提出了基于上下文的网络爬虫。基于上下文的网络爬虫使用一个通用搜索引擎来获取链接到专门文本的网页，并为该文本建立上下文图表。该上下文图表用于训练一组分类器，根据与目标的链接距离的预测值将文本归类。每个种子页都有相应的上下文图和分类器，并能够逐层建立直到特定的等级。这样网络爬虫能够获取与目标主题直接或间接相关的内容，以及通向目标非常简单的路径模型。与普通的宽度优先和传统的主题网络爬虫相比，上下文主题网络爬虫（CFC）的每一层都使用贝叶斯分类器，最先使用最短路径。使用的评价机制是下载的相关网页的比率。实验表明，CFC 维护了更高的相关性。

C. Aggarwal 等人[21]提出了 Web 主题管理系统（WTMS）。他们使用主题爬行技术，只下载相关网页附近的网页（如双亲、孩子和兄弟）。它基于对网页的向量空间表示，网络爬虫下载所有包含主题关键词的 URLs。最后，它为每个特定的类固定了一个相关性域值。如果相关性值下降到某个特定值，将停止下载。该论文的另外一个方面是使用基于 Hubs 和 Authorities 对站点进行逻辑分组的方法，讲述了如何可视化主题。

A. McCallum 等人[22]提出了使用加强学习的方法建立专门域的搜索引擎。贝叶斯分类器根据整个网页的文本和链接文本对超链接进行分类。这样为每个链接计算出了奖励值，用于通过加强学习的方法来训练网络爬虫。通过对计算机系站点的测试表明，与普通的宽度优先网络爬虫相比，主题网络爬虫在更少的时间里发现了更多的研究论文。

M. Ehrig 等人[23]提出了一种基于 ontology 的相关度计算和主题爬行体系框架。首先经过网页进行预处理后，从网页中抽取实体（即 ontology 中出现的关键词）并进行频率统计；然后针对用户选择的感兴趣实体，在 ontology 图上基于多种策略来计算主题相关度，即直接匹配方法、类别关系和复杂关系计算等方法。通过实验与基本的主题爬虫（即简单通过关键词的布尔匹配来计算主题相关度）进行比较。结果证明该方法 harvest rate 有了很大提高，但是它没有与其他类型的主题爬虫进行比较。

3　主题网络爬虫的研究趋势

综上分析可知，未来主题网络爬虫的研究主要是围绕如何提高链接主题预测的准确性，降低计算的时空复杂度，以及增加主题网络爬虫自适应性这几个方面展开。

提高链接价值预测的准确性一直是近年来研究的焦点。将各类评价方法相结合，尤其是基于分类器的主题相关度预测和基于在线训练、反馈的主题爬行方法值得进一步训究。将目前信息检索领域中的概念检索理论应用于链接价值的计算，是一个新的尝试方向。网络爬虫的爬行具有重复性，如何将 Web 动态变化的规律与先前搜索的统计结果相结合，以提高价值计算的准确性，是一个值得研究的问题。降低网络蜘蛛在训练、搜索过程中的计算复杂性，也是有待进一步研究的问题。目前的网络爬虫通常采用固定的搜索策略，缺乏适应性，如何提高网络爬虫的自适应性有待进一步研究。

4　结束语

随着人们对个性化信息服务需要的日益增长，基于主题爬虫的专业搜索引擎的发展将成为搜索引擎发展的主要趋势之一。主题网络爬虫爬行策略的研究，对专业搜索引擎的应用和发展具有重要意义。本文在给出主题网络爬虫的定义和研究目标的基础上，对现有的主题爬行策略进行了分类，系统分析了它们的定制方法，比较了它们的优缺点。最后，给出了若干值得进一步研究的问题。

参考文献：

[1] MURRAY B, MOORE A. Sizing the Internet[M]. [S. l.]: Cyveillance Inc, 2000.

[2] LAWRENCE S, GILES L. Accessibility and distribution of information on the Web[J]. Nature, 1999, 400(8): 107-109.

[3] CHO J, CARCIA M H. The evolution of the Web and implication for an incremental crawler[C]//Proc of the 26th International Conference on Very Large Databases (NVLDB-00). 2000.

[4] BREWINGTON B E, CYBENKO C. How dynamic is the Web[C]// Proc of the 9th International World Wide Web Conference. 2000.

[5] MENCZER F, PANT C, RUIZ M E. Evaluating topic-driven Web crawlers[C]//Proc of SIGIR'01. New Orleans, Louisiana: [s. n.], 2001: 241-249.

[6] MENCZER F, PANT C, SRINIVASAN P. Topic-driven crawlers: machine learning issues[EB/OL]. (2002-05-15). http://dollar.biz.uiowa.edu/~fil/papers.html.

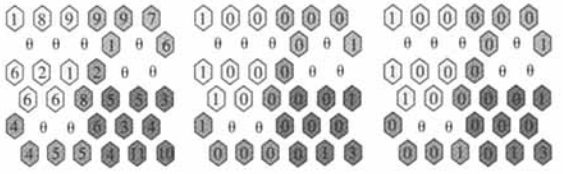
[7] CHO J, GARCIA M H, PAGE L. Efficient crawling through URL ordering[J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 161-172.

[8] DeBRA P, HOUBEN G, KORNATZKY Y, *et al*. Information retrieval in distributed hypertexts[C]//Proc of the 4th RIAO Conference. New York: [s. n.], 1994: 481-491.

[9] HERSOVICI M, JACOVI M, MAAREK Y S, *et al*. The shark-search algorithm: an application: tailored Web site mapping[C]//Proc of the 7th International World Wide Web Conference. Brisbane: [s. n.], 1998: 65-74.

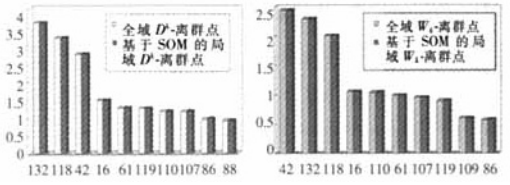
（下转第 47 页）

据分析的基础上，有针对性地选取感兴趣区域进行深入分析，具有交互性的特点。同时，由于可在 SOM 的局部邻域内寻找 k-最近邻居，根据离群数据定义进行算法的设计与实现，使其具有可扩展性、可预测性、简明性等特征。



(a) 数据点分布图　(b) D^k 离群点分布图　(c) w_k 离群点分布图

图 1　Iris 数据集的 SOM 命中标记图



(a) D^k-离群点及 D^k-距离（基于全域及基于 SOM V 集）　(b) w_k-离群点及 w_k-距离（基于全域及基于 SOM V 集）

图 2　Iris 数据集的离群数据及距离

参考文献：

[1] HAN J, KAMBER M. Data mining, concepts and technique[M]. San Francisco: Morgan Kaufmann, 2001.

[2] ESKIN E, AMOLD A, PRERAV M, *et al*. A geometric framework for unsupervised anomaly detection: detecting intrusions in unlabeled data [C]//Applications of Data Mining in Computer Security. Boston: Kluwer Academic Publishers, 2002.

[3] JIN Wen, TUNG A K H, HAN Jia-wei. Mining top-n local outliers in large databases[C]//Proc of ACM SIGKDD Int'l Conf on Knowledge Discovery and Data Mining. San Francisco: [s. n.], 2001.

[4] YU D, SHEIKHOLESLAMI G, ZHANG A. Findout: finding outliers in large datasets[J]. Knowledge and Information Systems, 2002, 4 (4): 387-412.

[5] KNORR E, NG R. Algorithms for mining distance-based outliers in large datasets[C]//Proc of Int'l Conf on Very Large Databases. New York: [s. n.], 1998: 392-403.

[6] RAMASWAMY S, RASATOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[C]//Proc of ACM Int'l Conf Management of Data. Dallas: [s. n.], 2000: 427-438.

[7] ANGIULLI F, PIZZUTI C. Outlier mining in large high-dimensional data sets[J]. IEEE Trans Knowledge and Data Eng, 2005, 17 (2): 203-215.

[8] BREUNIG M M, KRIEGEL H, NG R, *et al*. LOF: identifying density-based local outliers[C]//Proc of ACM Int'l Conf on Managment of Data. Dalles: [s. n.], 2000.

[9] PAPADIMITRIOU S, KITAGAWA, GIBBONS P B. LOCI: fast outlier detection using the local correlation integral[C]//Proc of the 19th International Conference on Data Engineering. Bangalore: [s. n.], 2003: 315-326.

[10] KOHONEN T. Self-organizing maps[M]. Berlin: Springer-Verlag, 1997.

[11] 汪加才，陈奇，赵杰煜，等. VISMiner：一个交互式可视化数据挖掘原型系统[J]. 计算机工程，2003，29(1)：17-19.

[12] ULTYSCH A, SIEMON H P. Kohonen's self-organizing feature maps for exploratory data analysis[C]//Proc of INNC'90, International Neural Network Conference. Dordrecht, Netherlands: [s. n.], 1990: 305-308.

[13] AGGARWAL C, YU P. Outlier detection for high dimensional data [C]//Proc of the SIGMOD. Santa Barbara: ACM Press, 2001: 37-46.

(上接第 29 页)

[10] BRIN S, PAGE L. The anatomy of a large-scale hypertexual Web search engine[C]//Proc of the 7th World Wide Web Conference. Brisbane: [s. n.], 1998.

[11] CHAKRABARTI S, DOM B, INDYK P. Enhanced hypertext categorization using hyperlinks[C]//Proc of the ACM SIGMOD International Conference on Management of Data. Seattle: [s. n.], 1998: 307-318.

[12] 傅向华，冯博琴，马兆丰，等. 可在线增量自学习的聚焦爬行方法[J]. 西安交通大学学报，2004，38(6)：599-602.

[13] PANT G, SRINIVASAN P. Link contexts in classifier-guided topical crawlers[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(1): 107-122.

[14] PANT G, SRINIVASAN P. Learning to crawl: comparing classification schemes[J]. ACM Trans Information Systems, 2005, 23(4): 430-462.

[15] CHAKRABARTI S, BERG M van den, DOM B. Focused crawling: a new approach to topic-specific Web resource discovery[C]//Proc of the 8th International Conference. Toronto: [s. n.], 1999.

[16] CHAKRABARTI S, PUNERA K, SUBRAMANYAM M. Accelerated focused crawling through online relevance feedback[C]//Proc of the 11th International World Wide Web Conference. Hawaii: [s. n.], 2002.

[17] 李盛韬，赵章界，余智华. 基于主题的 Web 信息采集系统的设计与实现[J]. 计算机工程，2003，29(17)：102-104.

[18] 李卫，刘建毅，何华灿，等. 基于主题的智能 Web 信息采集系统的研究与实现[J]. 计算机应用研究，2006，23(2)：163-166.

[19] JOHNSON J, TSIOUTSIOULIKLIS K, GILES C L. Evolving strategies for focused Web crawling[C]//Proc of Int'l Conf Machine Learning. 2003.

[20] DILIGENTI M, COETZEE F, LAWRENCE S, *et al*. Focused crawling using context graphs[C]//Proc of the 26th International Conference on Very Large Databases (VLDB 2000). Cairo: [s. n.], 2000.

[21] AGGARWAL C, AL-GARAWI F, YU P. Intelligent crawling on the world wide Web with arbitrary predicates[C]//Proc of the 10th International World Wide Web Conference. Hong Kong: [s. n.], 2001.

[22] MCALLUM A, NIGAM K, RENNIE J, *et al*. Building domain-specific search engines with machine learning techniques[C]//Proc of 1999 AAAI Spring Symposium on Intelligent Agents in Cyberspace. Stanford, CA: Stanford University, 1999.

[23] EHRIG M, MAEDCHE A. Ontology-focused crawling of Web documents[C]//Proc of ACM Symposium on Applied Computing. 2003.