

WWW 站点的自动信息提取与分类

程学旗* 余智华* 白硕* 夏威**

*国家智能计算机研究开发中心, 北京 100080

**中国人民解放军总后研究所

Email: cxq@ncic.ac.cn

摘要: 网络信息资源在迅速增长, 要提高网上信息服务的性能和效率, 简单的信息导航或检索已经很难满足实际需求, 而必须加入对信息源本身结构和语义的自动分析。通过对 WWW 站点进行分类, 并分析站点组织结构, 可为浏览提供智能导航。将其与搜索引擎相结合, 可提高信息检索的有效性。本文讨论了站点信息的自动提取、自动分类的方法及相关技术, 实现了站点结构分析、站点主题获取和领域分类。同时, 当站点涉及多个领域时将其适当分解成多个子站点并确定每个子站点的领域。相关的技术已经应用到曙光“天罗”网络信息服务平台中。

关键词: 站点分类 站点结构分析 智能导航 搜索引擎

Automatic Information Extraction and Classification to the Web Sites

Cheng Xueqi Yu Zhihua Bai Shuo Xia Wei

National Research Center for Intelligent Computing Systems(NCIC), Beijing 100080

Email: cxq@ncic.ac.cn

Abstract: Although there are many popular search engines and many Web Portals help people gain information from the Internet, we need some more intelligent methods which be able to direct and navigate our exploring in Internet by semantic comprehension. So we developed methods and some other relevant techniques to automatically analyse and classify Web sites. We also put out some methods which extract important information from the web sites. Integrating these with the search engine, we can provide more extensive services than earlier.

KeyWords: Classify to the Web Sites, Information Extraction, Intelligent Information Navigating, Search Engine

1 引言

随着网络上信息的高速增长和快速变化, 人们在运用网络获取信息时遇到了一些无法避免的困难。如何在 Internet 这么一个动态变化的环境下对各种信息进行收集、分析及评价并提供高效的检索和导航服务成为目前计算机研究领域的一个热点。通过对 WWW 站点进行自动分类, 并分析站点组织结构, 抽取相关的信息, 可为浏览提供智能导航, 同时将其与搜索引擎相结合, 可提高信息检索的有效性。

Internet 信息随时可能处于变化之中, 搜索引擎必须不停地刷新数据, 但仍无法避免无

效的检索结果。1995年的调查表明，通过 Internet 中最常用的一些搜索引擎查询到的结果 URL 中，14.9%的目标页面已经失效了[Selberg and Etzioni 1995]。通过对网上 WWW 站点的分析我们发现，虽然站点的内容处于不断更新之中，但站点的结构通常比较稳定。比如一个新闻站点，其中的新闻报道虽然更新频繁，但仍保持固定的结构，经济栏始终都是经济新闻，而科技板块仍然是有关科技的文章。如果能分析出这些结构并据此进行检索，就能较好地保证检索结果的正确性。

WWW 站点中的页面并不是孤立存在的，很多页面通过超链接结合起来构成比较完整的内容。仅仅查询出单个页面往往不能满足用户的需要。因而若能将这些页面组合起来进行检索并一次提交给用户，就可以在提高检索效率的同时满足用户的需要。这样做还可以避免将同一组内容（如电子书）的各个页面分别提交给用户而产生重复的查询结果。

由上面分析的情况可以看出，要提高搜索引擎的性能和效率，已不能仅从检索上来考虑，而必须加入对信息源（尤其是 WWW）结构和语义的分析。本文讨论了站点分析、分类的方法及相关技术，实现了站点结构分析、站点主题获取和领域分类，以及当站点涉及多个领域时将其适当分解成多个子站点并确定每个子站点的领域。

2 站点信息提取与站点分类模型

站点信息的自动内容提取与站点的分类与分解的主要目的是为了更好的信息导航和信息发布。其中，从页面内部提取的信息主要是需求驱动的，包括对页面的文摘、LINK、内容的类别、内部的结构特征等方面进行综合提取，提取出来的信息统一存放到页面信息库内。而站点的自动分类与子类的自动分解则主要是基于站点内部页面提取出来的信息和站点本身的拓扑结构分析获得的。

站点信息分析提取和站点分类的系统模型如下图所示：

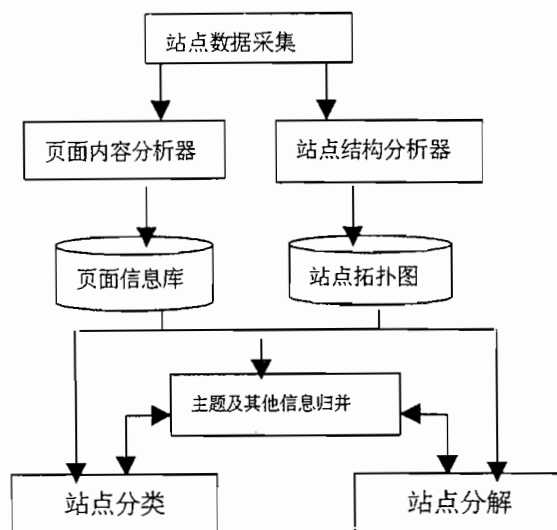


图 1 系统模型示意图

系统主要包括六大功能模块：数据采集、页面分析、站点结构分析、综合信息归结、站点分类、站点分解等。

一个完整的站点的原始数据由一个网络信息采集器完成，随后对其中的 HTML 页面进行分析。页面分析过程中提取出页面的正文和链接等信息，并将这些内容分析、分类。根据页面及链接中的结构信息构造出站点的组织结构，然后基于站点结构从每个页面的内容归纳出整个站点的主题领域。如果站点主题不明确，则需按照站点的结构从主页开始自顶向下地逐层分解，最终将站点适当地划分成一系列子站点。

3 站点分析与分类的相关算法描述

3.1 页面内容分析与信息提取

对采集到的页面首先进行 HTML 语法分析，剔除各种标记，提取出正文、链接及标题等其它结构数据。通过对页面中的正文进行文本分类得到页面的主题领域信息，这些主题信息是判定站点主题的基础。而页面链接表达了页面间的联系，是构造站点结构和分析站点主题的重要依据，因此需要将这些链接全部记录下来并加以分类。

在对页面正文分类时，我们不仅是判别该页面主要论述哪个领域，而且需要获得该页面所涉及的各个领域的更精确的信息。在实验系统中我们采用了基于关键词的向量空间模型（VSM）。首先按照我们设定的领域分类体系从大量训练语料中统计出每个领域的特征向量，在分析页面时通过计算待分类文本与这些特征向量的空间距离获得该文本与每个领域的相似度。这些相似度表示了该页面与哪些领域有关，并定量地确定了相关的程度。

站点的内部结构和联系主要体现在页面链接中，而不同的链接所体现的头尾两个页面间的关系是有差别的。我们把 Spertus 的分类方法[Spertus 1997]加以扩充，基于所表示的页面关系把页面中包含的链接分成以下 6 类，判别的标准主要是页面的 URL：

- DOWNWARD——下行链。目标页面是当前页面的下级页面，通常是当前页面提供更详细的信息，可以看作是当前页面的附属子页面。典型运用在主页、索引页面中。
- UPWARD——上行链。与 DOWNWARD 链相反，目标页面是当前页面上级，也可能是跨了好几级。许多页面都提供这样的链接让浏览者方便地返回上一级索引或直接返回主页。
- HORIZONTAL——水平链。同一目录下的页面间的链接被定义为水平链。
- CROSSWISE——交叉链。前面三种链接两个页面都处在站点目录中同一条路径上，当两者分别位于两个不同的分枝上时，就无法比较它们的上下级，我们称之为交叉链。
- OUTWARD——外向链。链接指向其它站点中的页面。此类链接所表示的页面关系更加模糊，它指向的页面内容可能与当前页面毫无关联。而且要分析相关的所有站点往往超出了系统的能力，不可能搜集到链接有关的完整信息。
- FRAME——框架链。此类链接比较特殊，从外观上看目标页面实际上嵌入到当前页面中，完全可以看作当前页面的一部分，因而它所表示的关系是最紧密的。它另一个特殊的方面在于该类别的判定不是根据 URL，而是在语法分析构成中抽取链接时就

已经确定了。

3.2 站点结构分析

站点的结构信息由两个部分组成：页面树和链接表，分别记录页面和链接相关的属性，两部分通过多种指针紧密结合在一起。

页面树是将站点中所有页面和目录依照 URL 中表示的目录层次组织成树型结构。树中的每个节点对应一个页面或/和目录，叶子节点对应一页面，中间节点对应一级目录。页面节点中包含指入链和指出链的头指针，分别指向链接表中的链接节点，由此可以获得指向本页面的链接以及本页面中包含的链接。此外页面节点还包含页面标题、正文长度、相似度向量等信息。

链接表用十字链表的方式组织，一行中的节点组成的链表包含此页面中的所有链接，即该页面的指出链；一列中的节点组成的链表包含指向此页面的所有链接，即该页面的指入链。链接节点中包含链接的类型和重数。指向同一页面内任何具体位置的链接都看作目标相同而合并，忽略链接 URL 中的页面内链接点标识。

基于站点和结构信息，我们可以对页面进一步加以分析并分类。这里所说的分类与页面分析中对页面正文的分类不同，是判断每一个页面在站点结构中的地位和作用，并依此分类，因此我们称之为按功能分类。

在站点中，并不是每个页面都有相同的重要性。显而易见，WWW 站点的主页是很重要的，它是大多数浏览者访问该站点的出发点，应能使浏览者迅速了解该站点的主要内容和结构，并有效地引导浏览者寻找他所感兴趣的页面。其实不仅仅是站点的主页，一般说来组织良好的站点其内部也建立了一系列的中心页面，每个中心页面在其所代表的页面集中起着类似主页的作用，概括了该部分页面的内容。这些页面当然具有比其它页面更大的重要性。在普通页面中，由于内容的不同也有主次之分。比如说在一部用超文本组织的电子书中，标题和摘要等通常放在主页中，以便读者一开始就能对全书内容有大致了解。在书的其它部分中除了包含正常章节内容的页面外，还有些页面记载的是诸如附录、参考书目等附加信息。显然从内容上比较，这些记录附加信息的页面就不如包含章节内容的页面重要。

Pirolli 提出了页面的一种分类方法，将页面分成五类及一些子类，并试图通过分析页面的特征找出可用于分类的规则 [Pirolli 1996]。分析他所提出的这个分类体系我们认为他在定义页面类别时倾向于从页面的内容和含义上考虑，从人的认识角度出发力求精细，而没有考虑这些类别在应用上是否有益以及是否便于自动分类。从他的分类结果看，分类的准确度基本在 50%到 70%之间。而且他的结果是基于对“www.xerox.com”这一个站点的统计分析得出的。在推广到所有的 Internet 站点时页面的某些特性会发生变化，从而更加影响分类方法的有效性。

我们定义页面类别的出发点是从应用的角度考虑，根据页面的功能和作用加以区分，而不考虑其内容。为提高自动分类的效率和准确度，我们尽可能使不同类别的页面间有较明显的区别，并且都有一定的实际意义和应用价值，而不追求分类体系的细致。我们将页面划分为以下 4 个类别：

- 主页。页面集的中心页面，概要介绍整个页面集的内容或主题，并提供链接指向页

面集的具体内容或下级主页。

- 索引页面。引导用户浏览其它站点或本站点中其它板块的页面。页面中包含大量指向其它站点或其它页面集中的链接，所指向的站点或页面与本页面集的主题可能相关也可能无关。
- 内容页面。不具有特殊功能的普通页面，其对站点的贡献就是页面里的内容，在浏览时没有附加的作用和意义。
- 参考页面。提供说明或引用等辅助信息的页面，与内容页面的差别在于会反复被多个页面所引用，而不固定包含在某个页面集中。

对于我们站点分析分类系统来说，上述 4 类页面中最重要的是主页。主页不仅概括了站点的内容，而且是站点结构的重要标志，因此准确地识别站点及其中各页面子集的主页是页面分类的首要目标。内容页面是站点内容的主要载体，也是站点主题信息的主要来源。参考页面的内容对站点主题也有一定贡献，虽然与普通页面相比参考页面与站点主题的联系不是那么紧密，但很难精确衡量其中的差别。事实上参考页面的另一个重要价值体现在站点结构的分析上，由于参考页面往往难以归入某个页面子集，有效地区分出参考页面可以大大减少它们对站点层次结构分析的干扰。索引页面通常没有实质内容，而且其中链接所指的页面或站点并不一定与本站点的主题相近，因此在站点分析中没有多大价值。索引页面与站点中其它页面的联系并不紧密，把它们忽略掉对站点的结构并没有太大影响。

我们总结了各类页面的结构特征，包括链接数、链接的类型分布、页面长度、文件名、页面在站点中位置等。通过统计页面的这些特征量来自动判别页面的功能类型。

有些站点并不按页面间的关系来组织目录。处理这些站点时，用上面所说页面树就无法很好地分析页面间的关系和站点的结构，尤其是当站点主题不唯一而划分比较困难。要解决这个问题，就要从目录层次以外发掘页面间的相互关系，构造出一个层次结构来。把在目录中处于同等位置的、看似互不所属的页面用另外一个结构汇集成一个一个的页面集。这个结构我们称之为站点的逻辑结构。

系统中的逻辑结构中仍保留链接表，只是页面的组织方式有所不同。页面也用一棵树来表示，如图 2。树中的节点代表页面，树的边用来表示节点的上下级关系。站点有一个中心节点作为整棵树的根，剩下的页面被划分成多个页面集，这些页面集都是根节点的后代。对各个页面集重复这个划分过程，最终形成整个站点的逻辑树。

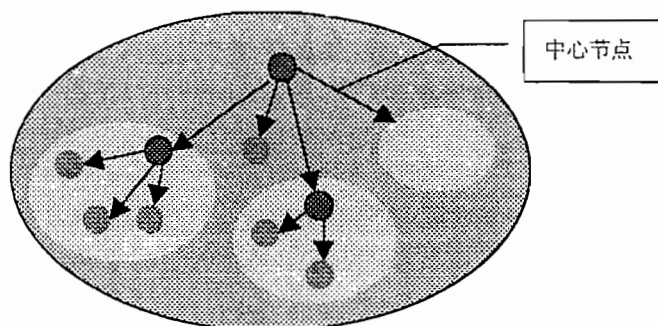


图 2 站点逻辑树示意图

由此可见，站点逻辑结构的核心是各级中心页面，它们构成了逻辑树的骨架，其它页面

都附属在中心页面下。与目录树不同的是，页面间的上下级关系不是由 URL 确定，而是由页面间的联系决定，页面间的这种联系综合页面类型、链接及页面的其它结构特征得出。在逻辑树中也不存在目录节点，每个节点都对应站点中的一个页面。

逻辑树从原理上看是对站点中页面集的不断划分，而其构造过程却相反，是一个自底向上的过程。由逻辑树的构成可以看出，每个页面都直接附属于一个中心页面。我们的构造算法就是：依序处理站点中的每个页面，从它的相邻页面中找出与它关系最紧密而又具有中心页面特征的页面作为它的上级页面，在这两个页面间加入一条边从而形成了一棵子树。对每个页面重复这一过程确定其父节点。在处理过程中肯定会出现不止一棵子树，但随着页面逐个被处理，子树会被合并。当站点中除了主页外所有页面都确定了父节点后，整个站点的逻辑树就已经建立起来，站点的主页就是这棵树的根节点。

构造逻辑树的关键是上级页面的判定。判定过程分为两步，首先从链接相邻页面中确定候选页面，然后再从中评比出唯一的上级页面。由于链接的类型代表了一定的页面联系，我们根据链接类型来确认可能的上级页面。在选择时按照 FRAME、DOWNWARD、HORIZONTAL、CROSSWISE 的优先次序。FRAME 和 DOWNWARD 都表示明显的上下级，而 HORIZONTAL 和 CROSSWISE 就不那么确定了。中心页面应该能有效地组织页面集并作为页面集的入口，即包含大量的指向下级页面的链接并且有链接从上级页面指入。利用链接数和链接类型分布可以从候选页面中挑选出最符合此标准的页面，在判别时考虑“index”等特殊文件名的使用可以提高判别的效率。在有多个中心页面都是当前页面的上级时，选择最近的作为直接上级页面，以使整个逻辑结构尽量精确。

在构造站点逻辑结构的过程中，索引页面和参考页面可能会带来一些麻烦，往往可以有多种各种类型的链接从各个位置的页面中指向它们。考虑到它们对站点主题来说并不是最主要的，在分析逻辑结构时可以做特殊处理甚至将它们撇开，这样可以获得一个比较清晰的站点结构。

3.3 站点分类及分解

站点的主题通过统计其中页面的主题信息获得，简单地说，哪个领域讨论的最多哪个就是站点的主题。链接表明了页面之间一定的关系，给页面提供了一些附加的主题信息。有些页面如索引页面本身内容不多，无法确定其主题，但通过分析其中的链接可能会发现指向的这些页面都是同一个领域的，由此可知该页面也应属于此领域。

站点主题统计过程分为三个步骤：扩散、合并以及累计。前两个步骤计算页面链接对页面主题的影响，累计则将主题信息逐层向上统计直到站点的主页。

扩散过程就是页面将自身的主题信息由链接传递给相邻页面的过程。在传递时根据链接类型的不同施加的影响也不一样。扩散算法建立在对不同类型链接所表示的页面间关系的分析基础上。我们发现，页面链接从统计意义上讲表示了一定程度的相关性。因此我们每种链接指定一个权值，由这个权值决定了其对页面主题的影响，最后综合所有链接的影响作为页面的附加主题信息。对每个链接计算了对相关页面的影响后，页面节点中累计了所有指出/指入链的附加主题信息。需要将这些附加信息与页面的原始主题按一定比例合并。

站点主题信息累计就是将站点中所有页面的主题信息逐层累加到上层节点，最终在站点的主页也就是根节点上统计出整个站点的主题信息。累计过程基于站点结构实现，利用页面树或逻辑树的深度优先遍历实现对所有页面的统计。页面的主题信息是通过表示上下级关系的链接累加到上层节点的，累计过程中使用了 DOWNWARD 和 FRAME 两类链接。在统计时用页面的正文长度为主题信息加权。

站点主题统计过程最终在站点的根节点即主页上获得了用领域相似度向量表示的站点整体主题信息。由这个向量我们来判定站点属于那个领域。一个站点归于某个领域必须具备两个特点，首先是站点内容与该领域密切相关，反映在主题信息中就是对应的相似度足够高。我们通过设定阈值来判断。另外一个特点是要保证领域的唯一性，就是要找出最相关的领域，而将其它领域排除掉。在排除其它领域时需要考虑这些领域是否真是次要的，我们通过计算各领域相似度的级差来进行过滤。具体计算方法是首先找出相似度最高的两个领域，计算两者相似度相差的倍数，它代表了两个领域与站点相关程度的差别。若级差值大于设定的阈值说明对该站点来说第二个领域的重要性要差很多，其它相似度更小的领域就更不用说了。由此可以断定第一个领域已经包含了站点的主要内容，忽略掉其它领域对站点主题信息造成的损失不大。

当无法判定站点的主题时，就需要从站点内部分解出能确定主题的子站点。一个站点在组织其内容时，总是把相近的页面组织成一个页面集，再把一些小的页面集成较大的板块。因此如果能按照这种结构反过来分解下去，肯定能得到内容紧凑、主题单一的板块。把这些板块独立出来作为子站点正是我们站点划分的目标。划分时自顶向下将站点结构树分解成一系列子树，判断每棵子树根节点的主题，若不能确定则继续分解此子树。划分过程中限制了站点划分的深度，以免将站点分解得过细而产生许多小站点。并根据节点的权值（即节点所代表页面集内容的总长度）、节点子树的深度等将内容不多的节点或零散的页面过滤掉。

站点划分要求站点结构与页面集的层次关系相吻合，因此采用站点逻辑树可以获得准确、高效的结果。若基于页面树划分只能按站点的目录结构分解，有时不能真正反映站点内容的组织结构，可能出现将同一主题的内容分解成多个子站点的情况。但通过巧妙地设置划分和过滤策略仍可以对多数站点获得比较满意的划分结果。

4 实验与分析

我们选取了 4 个专业站点分析站点分类的效果，分类结果如表 1。表中每个站点列出了自动分类后相似度最高的两个领域及其相似度，站点“中国传统医药网”除了“医疗卫生”外其它领域相似度为 0，故只列一个。由表中的数据可以看到，专业站点的主题非常明显，其它领域的相似度还不到其 1/10，因而不会出现混淆。

另一个分析实例是站点“serve.cei.gov.cn”（“中经网为您服务”）。它是一个较大的综合性站点，我们从该站点采集下来的共有 10300 多个页面，涉及教育、服务与旅游业、交通运输、医疗卫生、通信行业等多个领域，内容很复杂。表 2 列出了该站点分类后各领域的相似度（最高的 6 个），按相似度从高往低排列。

表 1 专业站点测试结果

站点 URL	站点内容	主题	相似度
Www.dalu.online.sh.cn	大陆旅游信息网	服务与旅游业	0.029953
		材料	0.000495
Www.medicinchina.com	中国传统医药信息网	医疗卫生	0.054194
Www.sport.gov.cn	中国体育信息网	体育	0.090095
		法律	0.004879
Www.zgl69.ncl/~qinglong	发烧军事	军事	0.066681
		文学理论	0.001973

表 2 站点“serve.cei.gov.cn”分类结果

主题	相似度
教育	0.010838
服务与旅游业	0.010618
通信行业	0.009074
交通运输	0.009065
艺术	0.006757
计算机	0.006012

除了表 2 中列出的 6 个领域，还有另外 8 个领域的相似度在 0.001 以上。可见该站点主题分布很广，而且由于内容分散于各个领域，必然地导致每个领域的相似度都很低。对比表 1 中主次领域相似度差别在 10 倍以上的专业站点，此站点最大相似度 0.010838 到它的 1/10 即 0.001 之间包含了 14 个领域。

表 3 站点“serve.cei.gov.cn”划分结果

子站点 URL	子站点内容	主题	相似度
serve.cei.gov.cn/index01.htm	政府机关	法律	0.017898
		电子行业	0.008812
serve.cei.gov.cn/index02.htm	文教机构	教育	0.037113
		电子行业	0.005028
serve.cei.gov.cn/index03.htm	医疗机构	医疗卫生	0.045384
		建筑与城市建设	0.008766
serve.cei.gov.cn/index04.htm	交通服务	交通运输	0.032772
		通信行业	0.012997
serve.cei.gov.cn/index05.htm	电话邮编	通信行业	0.031583
		服务与旅游业	0.010531

对此表 3 中提到的站点进行划分后得出的结果如表 3 所示。采用的是基于物理结构的划分算法，在划分时限制深度为 1，即只分解到站点主页下一层。在自动过滤了几个零散页面后，系统提交了 5 个子站点。

由表 3 中数据来看，划分后的子站点主题已经比较明确了，主次领域的相似度级差在 2

倍以上。对照划分前站点主页上的分类结果，不仅主题突出了，排除了不同领域间的干扰，而且主题的相似度也大大提高了。

5 结束语

从网络信息服务的实际需求出发，我们认为有必要对网站信息进行一定程度的自动信息提取和自动化的站点分类与子站点分解。我们在这个方面进行了一些尝试，提出了一些实用模型并对相关的技术与算法进行了实验，结果表明有关的策略比较可行。目前，相关的技术已经应用到曙光“天罗”网络信息服务平台上。

然而，由于网络资源本身的开放性、多变性，并且相关的协议一直在发展，不可能存在一个一劳永逸的方法使得对网络信息资源的自动处理能够完全满足实际的信息服务的需要。目前，我们正在对新型的关于元信息的知识描述方法和知识检索的相关理论与算法进行探讨研究，希望能够在此基础上提供一个真正的柔性化、一体化、智能化的网络信息服务平台。

参 考 文 献

- [Frei 1995] H.P. Frei and Stieger D, "The Use of Semantic Links in Hypertext Information Retrieval", *Inform. Proc. and Management*, Vol. 31, No.1, pages.1-13, 1995;
- [Pirolli 1996] Peter Pirolli, James Pitkow and Ramana Rao, "Silk from a Sow's Ear: Extracting Usable Structures from the Web", *CHI '96 Proceedings: Conference on Human Factors in Computing Systems: Common Ground*, pages 118-125, 1996;
- [Selberg and Etzioni 1995] Erik Selberg and Oren Etzioni, "Multi-Service Search and Comparison using the MetaCrawler", *Proceedings of the 4th International World Wide Web Conference*, 1995;
- [Spertus 1997] Ellen Spertus, "ParaSite: Mining Structural Information on the Web", *The Sixth International World Wide Web Conference*, April 1997