站点主题结构与导航归纳技术*

于 龙,尹 浩
(解放军理工大学 通信工程学院,江苏 南京 210007)

摘 要:站点主题描述了互联网站点中信息的聚合与分类,体现着信息逻辑结构,是分析站点信息的关键。分析站点逻辑结构是站点设计的逆向过程,为了准确分析站点中的主题,提出了站点主题结构的理论模型,以形式化的方式描述了站点中不同主题的组织形式、逻辑关系及相关性质,为面向主题的网络信息抽取提供必要的理论基础。在此基础上,进一步研究自动构建站点主题结构的技术,提出基于导航的主题结构归纳方法,并进行了算法描述和实验分析。实验结果证明,站点主题结构的理论模型概括了目前大多数站点的主题结构特征,基于导航的主题结构归纳方法能正确地建立站点的主题结构,并具有较快的运行时间。

关键词:站点;主题结构;导航
中图分类号: TP391 **文献标志码**: A **文章编号**:1001-2486(2012)05-0090-06

Website topic structure and navigation induction

YU Long, *YIN Hao*
(Institute of Communications Engineering, PLA University of Science and Technology, Nanjing 210007, China)

Abstract: Website topics, describing aggregation and classification of website information, embodying information logic structure, is crucial for website information analysis. Analysis of logical structure is the reverse process of website design. In order to accurately analyze the site topics, the research proposed a topic structure model describing the organizational forms, logic relations and related properties of different website's topics in a formal way, providing the necessary theoretical basis for the topic oriented web information extraction. On this basis, navigation-based topic structure induction was proposed with algorithm and experimental analysis to automatically construct topic structure of websites. Experimental results show that topic structure model generalizes most of the site's topic structural characteristics, while the navigation based topic structure induction can correctly establish the site's topic structure, and has a faster running time.

Key words: website; topic hierarchy; navigation

站点的主题结构描述了其中信息的组织形式与逻辑关系,是面向主题的信息抽取的基础。分析站点的主题结构是站点设计的逆向过程,要根据已有的信息去推测其中的逻辑结构。这样的逻辑结构能够更加准确地指导面向主题的网络信息抽取。

自动构建站点主题结构的技术被越来越多的研究者关注,目前自动获取站点主题结构的典型方法可以归纳为三类:基于页面间的链接特征获取,基于目录结构与URL特征获取和基于站点地图获取。

(1) 基于页面间的链接特征的获取方法

文献[1]将页面集建模为加权有向图 $G(V, E, \omega)$,其中 V 是页面,E 是链接,w 是权值。在定义了页面文本距离 $\mathrm{dcontent}(u,v)$ 和页面路径距离 $\mathrm{dpath}(u,v)$ 之后,将 w 定义为 $\mathrm{dcontent}(u,v)$ 与 $\mathrm{dpath}(u,v)$ 的加权和。之后采用相关的算法,在图中寻找最短路径,并由这些最短路径得到主题结构。在此基础上,文献[2]采用机器学习的方法对链接特征分类,并提出了一系列分类特征:URL特征、文件名称特征、文本相似矢量、导航条包含、在文本中的出现位置、链接文本长度、链接文本字体。以这些特征作为加权有向图的权值,进而得到站点主题结构。在文献[3]中对这一方法进一步深入,采用了三种分类器来学习链接的八个特征,得到加权值,用三种不同的算法来实现图结构的关键页面提取,进而生成站点主题树。通过实验分析总结出,采用判定树的机器学习方法,配合最小扩展树的图归纳算法,能够取得最高的站点主题获取精度。

基于页面间的链接特征的获取方法将站点主题建立为一个树形结构模型,这样的建模仅适用

* **收稿日期**:2012-04-03
基金项目:国家自然科学基金资助项目(60903042);国家863高技术资助项目(2010AA)
作者简介:于龙(1976—),男,北京人,博士研究生,E-mail:blade_fish@163.com;
尹浩(通信作者),男,研究员,硕士,博士生导师,E-mail:yinhao@263.net

于单一主题树的站点;这种方法虽然能够得到站点中的一组关键页面,但是链接层面的关键页面不足以描述逻辑层面的主题结构,该模型仅对那些关键页面恰好是主题的站点是可行的;需要收集站点中的全部页面并进行复杂度为 $O(|V|^2)$ 的图迭代算法,对页面规模较大的站点难以实现。

(2) 基于目录结构与 URL 特征的获取方法

文献[4]提出了链接结构和内容结构的分离,并在目录结构的基础上,将链接分为 5 类。通过对目录结构的分析,得到站点主题结构。采用同样技术路线的文献还包括文献[5-8]。

基于目录结构与 URL 特征的获取方法存在一个必要的前提:站点主题结构与目录结构或 URL 结构密切相关。虽然许多站点的主题和其目录结构存在关联,但这样的关联程度各不相同,关联的方式也存在很大差异,一些采用查询词动态构建 URL 的站点,难以通过机器学习的方式去发现其规律;基于目录结构与 URL 特征的获取方法因为目录结构与主题的复杂关系对机器学习算法提出了很高的要求,人工参与程度也相应增加,因此难以自动实现大规模站点的主题结构获取。

(3) 基于站点地图的获取方法

为了使爬虫能更有目的去搜索站点中的内容,搜索引擎需要了解站点的主题结构,站点地图也应运而生[9]。站点地图描述了站点的主题结构,可以作为快速获取站点主题的一种方式。然而,并不是所有站点都有站点地图,目前的站点地图大部分是手工完成的[10],对于大规模站点难以进行全面而完整的描述。基于站点地图的获取方法具有极高的效率,它仅需要解析一个文件就可以完成任务,但是这种方法对站点地图有着非常苛刻的要求。既要完整翔实到每一个主题,又要随着站点结构的变化而更新。目前具有这样性质的站点地图几乎不存在。

综上所述,为了逆向获取站点的主题结构,现有的方法都存在着各自的局限性。目前并未存在成熟的理论模型和成熟的算法来解决这类问题。

随着 HTML 语言的发展[11],导航信息成为页面中的重要元素。因为导航信息源于站点自身,又包含着对主题的描述与主题关系,因此可被用来构建站点的主题结构。在下文中将详述构建站点主题结构的理论模型和基于导航的主题结构归纳算法及相关实验。

1　主题结构的定义

描述站点的主题结构要形式化定义主题及主题之间的关系。为此引入如下定义:

定义 1　主题可被形式化定义为二元组:

$t = \{ \text{topicName}, \text{topicPageUrl} \}$

其中,topicName 是阐述该主题的词汇或短语,称为主题名称;topicPageUrl 是该主题所对应的广义页面地址。

定义 2　设有主题 t_1, t_2,若在逻辑上 t_2.topicName 是 t_1.topicName 的子主题,称 t_1,t_2 具有父子关系,以有序的二元关系对形式定义为:(t_1, t_2)。

定义 3　主题包含关系可定义为:

(1)若 (t_1, t_2),则 $t_1 \supset t_2$

(2)若 $t_1 \supset t_2$ 并且 $t_2 \supset t_3$,则 $t_1 \supset t_3$

定义 4 主题结构用于描述多个主题及相互关系,可被定义为二元组:$T = \{ \text{Topics}, \text{fsRelations} \}$,其中,Topics 是主题集合,fsRelations 是定义在 Topics 上的父子关系集;

若 T_1.Topics $\supset$ T_2.Topics,并且 T_1.fsRelations $\supset T_2$.fsRelations,则称主题结构 T_2 是主题结构 T_1 的子结构,记为:$T_1 \supset T_2$。

两个主题结构相加将对应的主题集与关系集分别取并集,形式化定义为

$T_1 + T_2 = \{T_1.\text{Topics} \cup T_2.\text{Topics}, T_1.\text{fsRelations} \cup T_2.\text{fsRelations}\}$

定义 5　设主题结构 $T = \{ \text{Topics}, \text{fsRelations}\}$,若 fsRelations 满足:

(1) $\forall t \in \text{Topics}$,$t$ 的父主题唯一;

(2) $\forall t_1, t_2 \in \text{Topics}$,若存在 $t_1 \supset t_2$,则不存在 $t_2 \supset t_1$;

称主题结构 T 逻辑一致。

定义 6　主题路径是特殊的主题结构,可被定义为

$\text{TopicPath} = \{t_1, t_2, \cdots, t_n\}, \forall i, (t_i, t_{i+1})$

其中,$\{t_1, t_2, \cdots, t_n\}$ 是对应自然数列的有序集。

定义 7　站点主题结构是逻辑一致的主题结构,记为

$\text{S.TopicStructure} = \{ \text{Topics}, \text{fsRelations} \}$

其中,Topics 是站点中所有主题的集合,fsRelations 是定义在 Topics 上的所有父子关系集合。

2　主题结构的性质

利用导航信息归纳主题结构,需要归纳逻辑一致的主题路径,因此需要研究主题结构的相关性质。

定理 1　设有主题结构 T_1,T_2,那么 $T_1 + T_2$

$= T_2 + T_1$。

证明 设 $T_a = T_1 + T_2, T_b = T_2 + T_1$,由定义 4, T_a. Topics $= T_1$. Topics $\cup T_2$. Topics, T_b. Topics $= T_2$. Topics $\cup T_1$. Topics,故 T_a. Topics $= T_b$. Topics;同理可证 T_a. fsRelations $= T_b$. fsRelations。由定义 4 中主题结构相等的充要条件可得, $T_a = T_b$,即 $T_1 + T_2 = T_2 + T_1$。

主题路径聚合是通过导航归纳站点主题的主要操作,由定理 1 可以看出,主题路径聚合与顺序无关。

定理 2 设有主题结构 T_1, T_2,其中 T_2 是 T_1 的子结构,若 T_1 逻辑一致,那么 T_2 逻辑一致。

证明 采用反证法,假设 T_2 逻辑不一致,那么由定义 5 分为两种情况:(1) $\exists t_a, t_b, t \in T_2$. Topics, $(t_a, t), (t_b, t) \in T_2$. fsRelations,并且 $t_a \neq t_b$;由于 $T_1 \supset T_2$,可知, $t_a, t_b, t \in T_1$. Topics, $(t_a, t), (t_b, t) \in T_1$. fsRelations;因 T_1 逻辑一致,所以推出 $t_a = t_b$,与假设情况(1)矛盾;(2) $\exists t_a, t_b \in T_2$. Topics,并且 $t_1 \supset t_2$,同时成立 $t_2 \supset t_1$;由于 T_1 逻辑一致,同理可证假设情况(2)不成立;综合两种情况可以推出, T_2 逻辑不一致这个假设不能成立,即 T_2 逻辑一致。

基于导航信息的主题结构归纳有时候会得到站点主题结构的子结构,由定理 2 可以看出,若站点主题结构逻辑一致,那么归纳得到的子结构也逻辑一致。

推论 设有主题结构 T, T_1, T_2,其中 T_1, T_2 都是 T 的子结构,若 T 逻辑一致,那么主题结构 $T_1 + T_2$ 逻辑一致。

证明 假设 $T_a = T_1 + T_2, \forall t \in T_a$. Topics,可得到 $t \in T_1$. Topics $\cup T_2$. Topics,若 $t \in T_1$. Topics,因 T_1 是 T 的子结构, T_1. Topics $\subset T$. Topics,故 $t \in T$. Topics,若 $t \in T_2$. Topics,因 T_2 是 T 的子结构, T_2. Topics $\subset T$. Topics,故 $t \in T$. Topics,归纳得到 $\forall t \in T_a$. Topics, $t \in T$. Topics,因此 T_a. Topics $\subset T$. Topics,同理可证 T_a. fsRelations $\subset T$. fsRelations;由定义 4 中子结构定义可以推出 $T_a \subset T$,又因 T 逻辑一致,根据定理 2 可知, T_a 逻辑一致,即 $T_1 + T_2$ 逻辑一致。

基于导航的主题结构归纳需要将若干主题结构相加,由推论可知,若站点主题结构逻辑一致,那么相加之后的主题结构也是逻辑一致的。

定理 3 逻辑一致的主题结构等效多棵树构成的森林,且此森林结构唯一。

证明 设 $T = \{\text{Topics}, \text{fsRelations}\}$,以集合 Topics 中的主题作为结点,集合 fsRelations 中的父子关系作为结点联系,可以构造多个图结构 $\{G_1, G_2, \cdots, G_k\}$。因为 T 是逻辑一致的,由定义 5 可以推出 $\forall i, G_i$ 无链,并且 $\forall t \in G_i, t$ 的父结点唯一,因此可推出 $\forall i, G_i$ 具有树形结构,进而 $\{G_1, G_2, \cdots, G_k\}$ 成为多棵树构成的森林结构 Structure $= \{\text{Tree}_1, \text{Tree}_2, \cdots, \text{Tree}_k\}$。

假设由主题结构 T 归纳的森林结构不唯一,并假设 Structure_1、Structure_2 是由 T 归纳的不同的森林结构,则必然存在某个主题 t,在 Structure_1 中具有关系 (t_1, t),同时在 Structure_2 中具有关系 (t_2, t),并且 $t_1 \neq t_2$,由定义 5 可以推出主题结构 T 不是逻辑一致的。这个结论与命题条件矛盾,可知原假设不成立,因此由逻辑一致的多元主题结构归纳的森林结构唯一。

与主题结构等效的森林结构如图 1 所示。

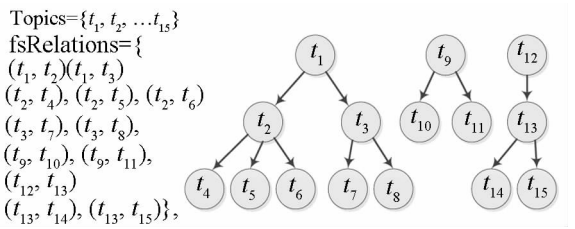

图 1 与主题结构等效的森林

Fig. 1 The forest equal with topic structure

由定理 3 可知,S. TopicStructure 可归纳为多棵树构成的森林,因此站点 S 的主题结构也可表示为多棵树的集合:S. TopicStructure $= \{\text{Tree}_i\}$,其中独立的树结构 Tree_i 称为站点 S 的主题树。某新闻站点主题树如图 2 所示。

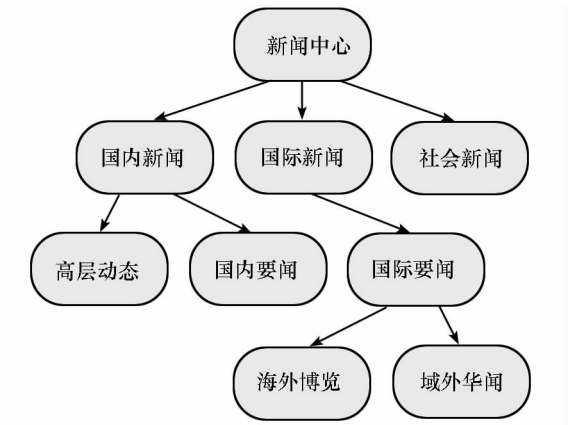

图 2 某新闻站点主题树

Fig. 2 Topic tree of a news website

3 基于导航的主题结构归纳

通过手工方式归纳一个站点的主题结构是非常繁琐的,为此需要寻找高效的方法,对站点主题结构自动归纳。目前大量的站点都采用主题导航条对其中的页面进行定位,这类普遍存在的页面

元素也简称为导航条。

导航条中蕴含着站点主题结构的部分定义。由于越来越多的站点采用高自动化的生成模板及动态页面的生成技术,同一个站点内的导航条具有高度的逻辑一致性。因为导航条的定义源自站点自身,它是站点的页面对其内部主题的简称及引用,又因为这样的信息具有高度的逻辑一致性,所以为自动构建站点主题结构提供了重要线索。一个典型的导航条如图 3 所示。

图 3　新闻站点导航条及其 HTML 文本

Fig. 3　Navigation bar and its HTML

导航条是分散于站点的诸多页面中的,每个页面中的导航条都是站点主题结构的一部分,由于这种分散与不完整的特点,要对大量页面的不同导航条进行抽取,并由此归纳出完整的站点主题结构,将这个过程称为基于导航信息的主题结构归纳。

在主题模型中,导航条可以看作主题路径,它是站点主题结构的子结构,构建站点主题结构的问题转化为通过若干主题路径来归纳站点主题结构的问题。基于导航信息归纳站点主题结构的算法如表 1 所示。

表 1　站点主题结构归纳算法

Tab. 1　Algorithm of website topic structure induction

算法:站点主题结构归纳算法
输入:广义页面集 P,主题导航元素分类 nav
输出:主题集 Topics 与父子关系集 fsRelations

```
1  Topics = null;  //主题集
2  fsRelations = null;  //父子关系集
3  foreach(page p in P)
4  {
5     navElement = GetElement(p , nav);
6     topicsInNav = GetTopicsFormNavElement (navElement);
7     Topics. addTopics(topicsInNav);
8     for(i = 1; i < topicsInNav. Length; i + + )
9  {
10     topicPair = null;
11     topicPair. father = topicsInNav [i - 1];
12     topicPair. son = topicsInNav [i];
13     fsRelations. add(topicPair);
14  }
15 }
16  return Topics ,  fsRelations;
```

算法通过对广义页面集的遍历,不断收集每个页面中的主题及父子关系,并汇集到预备好的数据结构中。函数 GetElement 通过导航元素分类 nav,在页面中获取导航元素;函数 getTopicsFormNavElement 用于将导航元素中的主题提取并返回主题集;topicPair 对象通过 father,son 属性来描述主题父子关系;在输出之前,算法遍历每个主题对应的页面,若其中存在分页元素,则将页面地址转化为超级页面地址的形式。在运行结束时,算法将主题集与父子关系集输出。

在归纳了站点主题结构的基础上,可以进一步通过主题结构生成主题森林,如表 2 所示。

表 2　站点主题森林归纳算法

Tab. 2　Algorithm of website topic forest induction

算法:站点主题森林生成算法
输入:主题集 Topics 与父子关系集 fsRelations
输出:根结点集 Rootlist 与结点集 Nodelist

```
1  Nodelist = null
2  foreach (topic in Topics)
3     {
4     treeNode = createTreeNode( );
5     treeNode. content = topic;
6     treeNode. father = null;
7     treeNode. subNodes = null;
8     Nodelist. add(treeNode);
9     }
10    foreach (fsRelation in fsRelations)
11    {
12    //以主题为索引,填充树结点的子结点集,并记录父结点
13    fatherTopic = fsRelation. father;
14    sonTopic = fsRelation. son;
15    tempNode = treeNode[sonTopic]
16    treeNode[fatherTopic]. subNodes. Add(tempNode);
17    treeNode[sonTopic]. father = treeNode[fatherTopic];
18    }
19    RootList = null;   //根结点集
20    foreach (treeNode in Nodelist)
21    {
22       //寻找根结点
23    if ( treeNode. father = = null)
24       {
25    RootList. Add(treeNode);   //添加根结点对象
26       }
27    }
28    return Rootlist, Nodelist;
```

算法构造了树结点对象 treeNode,与数据结构 Nodelist,遍历主题集,以每个主题做为 treeNode 对象的 content 属性,之后通过遍历父子

关系集,以主题为索引,将父子关系中的子主题对应的 treeNode 对象添加到父主题对应的 treeNode 对象的 subNodes 属性中,算法最后遍历所有的 treeNode,将父结点为空的所有结点存放于集合 Rootlist 中,再将 Rootlist 与 Nodelist 输出。

4 实验

本文实验环境和数据准备:主机 CPU 为 AMD PhenomII X4 925 2.8GHz;内存为 4G X 2 双通道 DDR3;网络环境为 20Mb/s ADSL;页面解析使用 HtmlParser 工具集;实验中采用的站点来自文献[12]。

4.1 模型的正确性验证

将不同规模的站点划分为四类,如表 3 所示。

表 3 四类不同规模站点

Tab. 3 4 classes of websites by page scale

	Ⅰ	Ⅱ	Ⅲ	Ⅳ
页面规模	500 个以下	500 ~ 2000 个	2000 ~ 10000 个	10000 个以上

在每个分类中选取 30 ~ 50 个站点进行完整爬取,用导航路径集构造每个站点的主题树,将所得到的主题集合与站点实际主题集合进行对比,计算每个站点主题获取精度(F 值)[13] 并在该类站点中求得统计平均,实验结果如图 4 所示。

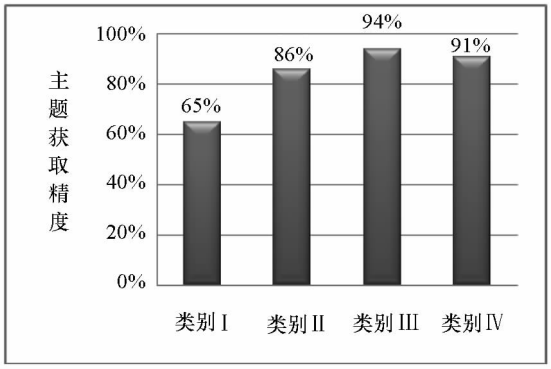

图 4 不同规模站点的主题构建准确度

Fig. 4 Construction accuracy in different website scales

从实验结果可以看出,第Ⅰ类站点,开发站点的工作多是手工完成,其中的部分站点及页面缺少导航,或者导航不具备逻辑一致性,导致一部分主题无法准确获取,平均精度为 65%;第Ⅱ类站点,开发过程的自动化程度增高,大部分页面都具备良好的导航结构,对主题的获取准确性相比Ⅰ类站高,平均精度为 86%;第Ⅲ类站点,由于页面规模不可能手工完成,开发过程高度自动化,未能正确获取到的导航多是因为 JavaScript 的出现,在爬取时并未对 JavaScript 作出相应解析,平均精度为 94%;第Ⅳ类站点,由于主题繁多,一些较深层次的主题树未能准确获取,平均准确度比第Ⅲ类站点略低,平均精度为 91%。

综合以上的分析可以发现,站点主题结构的理论模型是准确的,基于导航路径自动构建站点主题结构的方法能够对目标站点正确地构建主题结构。并且,构建准确程度与站点开发自动化程度密切关联。

4.2 不同爬取过程的构建效率

选取 100 个第 III 类站点,分别以首页和站点地图页面进行完整爬取作为基准主题结构。在不同的相对采样成本点记录其平均构建效率。其中的爬取代价定义为:已爬取页面数/站点中所有页面数;构建效果定义为:已爬取页面中获取的主题数/站点全部主题数。实验结果如图 5 所示。

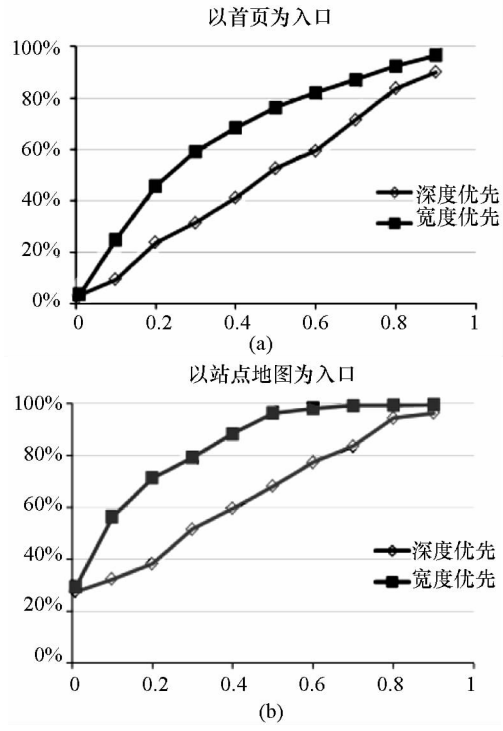

图 5 不同爬取过程的构建效率

Fig. 5 Efficiency of diffenent crawl mathod

从实验结果可以看出,构建站点结构的收敛速度与爬取算法密切相关。在相同的采样成本下,宽度优先算法比深度优先算法超出 5% ~ 38%;图(a)以首页为入口,图(b)以站点地图为入口,图(b)的构建有着更高的起点;以站点地图为入口的宽度优先爬取,可以在 30% 的代价下,达到 80% 的构建效果。

4.3 当前典型算法对比

在 4 类不同的页面规模下,每类规模选取 10 个站点做平均,分别用四种不同的方法获取站点

主题结构,并与正确的站点主题集合对比,计算不同方法的主题构建精度,实验结果如图 6 所示。

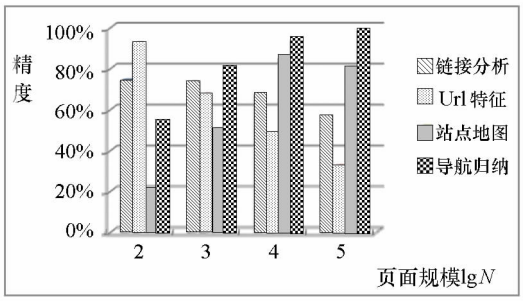

图 6　与当前典型算法对比

Fig. 6　Accuracy comparison with classic method

从实验结果可以看出,在站点规模较小、自动化程度不高时,导航归纳的精确度不如典型方法,平均精度为 54%;随着站点规模增大,页面自动化增加时,导航归纳的精度逐渐超过了典型方法,在页面规模在 10^5 个时,精度达到了 98%。

5　小　结

伴随互联网技术的迅猛发展,开发站点的技术也在不断变革。高效开发工具的涌现使开发者能够快速构建大规模站点,这也给主题结构逆向分析带来了巨大的挑战。为此,需要不断研究能描述符合当前网络特点的分析模型,探索效率更高、适应性更广泛的分析方法。

本文提出了站点主题结构的理论模型,并在此基础上提出了基于导航的主题结构归纳的方法。实验结果证明该模型符合目前大多数站点的特征,相应的归纳方法能正确地建立站点的主题结构,并具有较快的运行时间。

从实验结果看,虽然构建方法具有比较理想的精度,但仍然存在改进空间。增加对 JavaScript 的解析能进一步提高识别导航条的精度;对逻辑不一致的导航路径纠错也可以提高主题的获取精度。

本文提出的模型与方法仅适应导航结构完整的站点,对缺少导航结构的站点则无用武之地。HTML5 的出现为研究者带来了新的希望,其中专门添加的标签 < Nav > 恰恰证明了导航条的价值,在未来的一段时间里,随着遵循 HTML5 的站点增多,可以期待算法得到更精确的结果。

参考文献(References)

[1] Liu N, Yang C. Mining web site's topic hierarchy[C]// Proceedings of International World Wide Web Conference, Tokyo, Japan, 2005: 980 - 981.

[2] Liu N, Yang C. A link classification based approach to website topic hierarchy generation[C]// Proceedings of International World Wide Web Conference, May 8 - 12, 2007, Banff, Alberta, Canada, 2007: 1127 - 1128.

[3] Yang C, Liu N. Web site topic-hierarchy generation based on link structure [J]. Journal of the American Society for Information Science and Technology2009, 60(3): 495 - 508.

[4] Chen Z, LiuS, et al. Building a web thesaurus from web link structure[C]// Proceeding of the ACM SIGIR July 28 - August 1, Toronto, Canada, 2003: 48 - 55.

[5] 冯雁,王申康. Web 站点层次结构抽取算法的分析和实现[J]. 浙江大学学报:工学版, 2005, 39(10): 1507 - 1511.
FENG Yan, WANG Shenkang. Analysis and implementation of extraction algorithm of Web hierarchy structure[J]. Journal of Zhejiang University (Engineering Science) 2005, 39(10): 1507 - 1511. (in Chinese)

[6] 刘继红,吴军华,任明鑫. 基于改进的网络蜘蛛算法抽取 Web 站点结构的方法[J]. 江南大学学报:自然科学版, 2009, 8(5): 555 - 559.
LIU Jihong, WU Junhua, REN Mingxin. Method of the web structure recovery basedon the improved spider algorithm[J]. Journal of Jiangnan University(Natural Science Edition) 2009, 8(5): 555 - 559. (in Chinese)

[7] Baykan E, Marian L, Weber I. Purely url - based topic classification[C]// Proceedings of International World Wide Web Conference, April 20 - 24, 2009, Madrid, Spain, 2009: 1109 - 1110.

[8] Yang Q, Jiang P, Zhang C. Reconstruct logical hierarchical sitemap for related entity finding[C/OL]// Proceedings of 19th Text REtrieval Conference, Gaithersburg, Maryland, November 16 - 19, 2010[2012 - 5 - 27]. http://trec.nist.gov/pubs/trec19/papers/beijing.inst.tech.blog.entity.rev.pdf.

[9] Schonfeld U, Shivakumar N. Sitemaps: above and beyond the crawl of duty[C]// Proceedings of 18th International World Wide Web Conference, 2009: 991 - 1000.

[10] Liu N, Yang C. Keyphrase extraction for labeling a website topic hierarchy[C]// Proceedings of 11th International Conference on Electronic Commerce, Taipei, Taiwan, 2009: 81 - 88.

[11] HTML5 - A vocabulary and associated APIs for HTML and XHTML[OL]. W3C Working Draft [2012 - 5 - 27] http://www.w3.org/TR/html5/.

[12] Top 500 Sites of China in Alexa[EB/OL]. [2012 - 5 - 27] http://www.alexa.com/topsites/countries/CN.

[13] 程显毅,朱倩,王进. 中文信息抽取原理及应用[M]. 北京:科学出版社, 2010: 19 - 20.
CHENG Xianyi, ZHU Qian, WANG Jin. Principle and application of Chinese information extraction[M]. Beijing: Science Press, 2010: 19 - 20. (in Chinese)