基于本体概念图的电子课本系统构造技术研究

Research on generating an e – Textbook system Technology based on the Ontology Conceptual Diagram

曾义聪 （长沙民政职业技术学院电子工程系 长沙 410004）

摘 要： 提出一种在 Web 上自动构造电子课本学习系统的方法。学习者通过遍历域本体库的本体概念图，指定主题层次，引导主题爬取 Web 文档，自动构造电子课本，在 Web 上学习就像在读一本书。

关键词： 本体 本体概念图 电子课本 主题爬取

1 引言

最初在 Web 上学习，采用的是搜索引擎，由于 web 文档是为多种用途而制作的，基于无遗漏爬取的搜索引擎，消耗巨大的存储和带宽资源，同时用户很难找到他们特定需要的 web 文档。

一个理想的解决方法是采用主题爬取技术，收集指定主题的 Web 文档，构造电子课本（e – Textbook），让人们在 Web 上学习就像在读一本书[2]。面对信息的海洋，本体（Ontology）作为一种能在语义和知识层次上描述信息系统的概念模型建模工具，是一个日益流行的组织信息的方式。基于本体的信息组织，它是采用语义内容导向的方法，而非传统基于链接分析为导向的方法。与此相适应，我们的主题爬取过程中，更需要以语义内容为导向，使爬取过来的文档与主题语义更相关。

本文介绍一种新颖的方法，学习者通过遍历域本体库的本体概念图，指定主题层次，采用基于本体概念图的主题爬取技术爬取 Web 文档，自动构造电子课本。

2 相关工作

在主题爬取技术方面，文献[1]中的搜索策略采用一个上下文组合图，允许用户查询指向特定文档的网页，为种子文档构造一个合上下文图和对应各层的分类器，不足之处是构造过程须借助搜索引擎的部分功能，并且计算方法复杂。

文献[3]中的搜索策略采用随机冲浪模型，基于网页权值（PageRank）排序爬取，计算方法简单。适用于用户不能访问网上所有可能的网页情况下，对用户来说，优先访问"重要"网页显得尤为重要，但在爬取过程的开始阶段不是很有效，一般与其它方法组合使用。

而文献[4]中提出的智能爬取是采用统计模型，在进行爬取的过程中尽力学习链接结构特征。它利用了入链（inlinking）web 网页的内容信息，候选 URL 记号信息，入链（inlinking）web 网页或兄弟（sublings）网页其它行为，预测候选 URL 对给定爬取有用的概率。但实现中因为每次一个网页被爬取，我们需要分析它入链网页的内容，这是有点麻烦的事。文献[4]的作者利用一个启发式的调整，用侯选者本身的内容代替入链网页的内容，但侯选者并未爬行，它的内容从何知道，这是文献[4]的作者有欠考虑的地方。

所有这些基于链接分析的主题爬取方法共同不足之处是赋予待爬取的 URL 对象，只是链接结构方面的信息，而并非真正的语义信息。

在 Web 学习方面，文献[2]描述了允许用户指定主题层次，系统自动给他们产生完整课本。它的不足之处是首先它假设用户可能知道主题的纲要，而每个人对主题的纲要理解深度不同，从而指定的主题层次准确程度不同。其次是只对搜索引擎的结果按相关性和重要性进行重排，只在扩充结果集时才应用了主题爬取技术，但没有考虑用主题层次引导爬取。

本文提出一种在 Web 上自动构造电子课本学习系统的方法。学习者可通过遍历本体库的本体概念图，加深对概念关系的理解，进而构造准确的主题层次，这个主题层次引导主题爬取，最终实现电子课本自动产生。

3 构造电子课本学习系统

电子课本学习系统构造始于用户指定目标主题，通过遍历域本体库中的本体概念图，获得从根结点至目标主题结点的一条路径，我们称之为“knowledge – path”，构造对应的主题层次，建立相应的分类器，实现 Web 文档的主题爬取。

本体是共享概念模型的明确的形式化规范说明，本体的表示可采用概念图的表示。概念图的所有概念构成一个概念类型的层次结构，概念之间的层次关系为继承关系（kind – of）。将领域本体空间中的元素解释为概念图中的概念节点，领域空间上的概念关系解释为概念关联节点，本体空间元素之间的继承关系对应概念之间的层次关系，这样的概念图，我们称之为本体概念图。以计算机领域为例，本体概念图的顶层元素为“计算机”领域类，像 yahoo 的目录结构把“计算机”分为以下子类：算法、安全、硬件、软件和人工智能等，而软件又分为通信、多媒体和数据库等。“软件”与“计算机”之间即为继承关系，如图 1 所示。

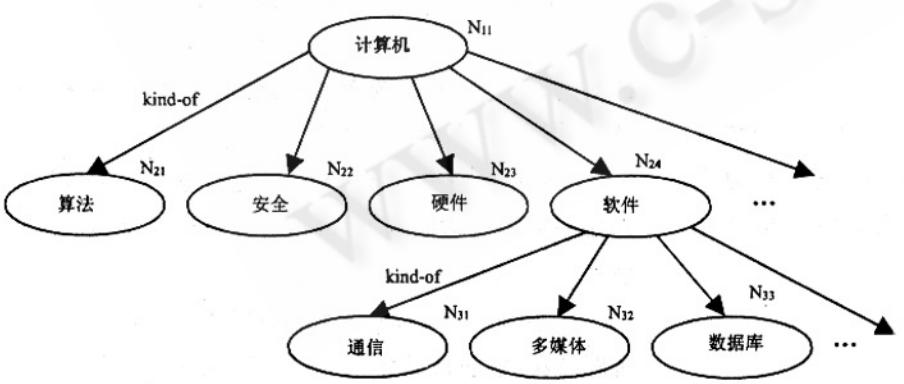

图 1　“计算机”本体概念图

图 1 的描述：

N_{ij} 表示本体概念图的第 i 层，第 j 个结点；

L_{ij} 是 N_{ij} 的标签，它是描述概念结点 N_{ij} 的少数关键字集；

parent（N_{ij}）表示结点 N_{ij} 的父结点；

child（k，N_{ij}）表示结点 N_{ij} 的第 k 个儿女。

假设以图 1 中结点 N33 为目标主题，构成从根结点“计算机”至目标结点“数据库”一条“knowledge – path”和对应的各主题层次，如图 2 所示。

图 2　主题层次图

图 2 的描述（图 1 的描述适用于图 2）：

m 为最大主题层号；

D_i 是主题爬取阶段对于主题层次图中的第 i 层（结点 N_{ij}）产生的描述词集，它包括 Nij 及 N_{ij} 子结点的标签，同时包括概念的别名，同义词，不同语种的译名等，用符号 Lf_i 表示，i = 1 ~ m；

对于主题层次图中的第 i 层（结点 N_{ij}），存在一个称之为“$DescriptList_i$”的 Web 文档表，这些文档是通过主题爬取获得的，它们与第 i 层主题的相似度和它们的重要度的合成权值满足某阀值 Hv；‘DescriptList0’为语义不相关层 0 的 Web 文档表，i = 0 ~ m。

对于主题层次图中的第 i 层(结点 N_{ij}),$AnchList_i$ 是 $DescriptList_i$ 的文档包含的侯选 URL 链接分类表。$AnchList_i$ 中的每项是一个三元组($AnchorText_k$,URL_k,$Weight_k$),并且各项按 $Weight_k$ 的降序排列。对于每一个 $AnchList_i$ 分配一个 URL 队列 $UrlQueue_i$,$i=0\sim m$;

对于某领域本体库的本体概念图,可采用宽度优先顺序遍历。由用户指定目标主题,构造对应的主题层次,对于主题层次上的结点的相关 Web 文档都是通过主题爬取获得。因此电子课本的构造过程可由以下两步完成:

(1) 用户指定的目标主题,遍历本体概念图,获得"knowledge – path"和描述词集;

(2) Web 文档的主题爬取,获取与目标主题相关 Web 文档。

3.1 获得"knowledge – path"和描述词集

由用户给出的目标主题,遍历领域本体库中的本体概念图,获得从根结点至目标主题结点的一条路径,我们称之为"knowledge – path",对应于该路径上的每一个概念结点 N_{ij},产生一个描述词集 D_i:

N_{mj}是叶子结点,则 $D_m=L_{mj}\cup Lf_m$;

N_{ij}是内部结点,当 $0<i<m$ 时,递归求出 $D_i=L_{ij}\cup Lf_i\cup D_{i+1}$,$i=i-1$。

3.2 Web 文档的主题爬取

3.2.1 构造主题层次

"knowledge – path"路径上的每一个概念结点,构造一个主题层,其中包括一个语义不相关层 – 层 0。以"计算机"本体概念图为例,假设用户的目标主题为"数据库",则我们得到如图 2 所示的主题层次图,同时初始化与主题层相对应的 m + 1 个 URL 队列 $UrlQueue_i$($i=0\sim m$),此时 m 等于 3。

3.2.2 基于主题层的 Web 文档分类及文档中 URL 对象分配

主题层次构造完成后,接着建立一个分类器,将网上爬取来的 Web 文档分配到合适的主题层,Web 文档中的侯选 URL 分配到与该文档所在主题层对应的 URL 队列。本文采用检索领域常用的向量空间模型作为页面文档与主题之间的相关性判定方法,同时考虑文档的重要性,将既语义相关又重要的文档分配至相应的 Web 文档表 $DescriptList_i$。

对爬取来的 Web 文档 P,我们用入链 Web 文档 Ti 的文档权值 IR(Ti),递归地定义一个文档 P 的重要度[3]。

$$IR(P)=(1-s)+s(IR(T1)/c1+\cdots+IR(Tn)/cn) \quad (1)$$

其中 Web 文档 P 被文档 T1 …,Tn 链接,ci 为文档 Ti 的出链数,若一个文档没有出链,我们假设它有指向任一 Web 文档的出链;s 为阻尼因子,一般取值为 0.85。

至于 Web 文档 P 与各主题层(层 0 除外)的相关性,可由以下公式计算出文档与各主题层的相似度,其中 d_i 为主题层描述词集 D_i 的特征向量,d_j 为待处理的文档 P 的特征向量,M 为向量的维数,W_k 为向量的第 K 维。

$$S(P)=Sim(d_i,d_j)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{M}W_{ik}^2)(\sum_{k=1}^{M}W_{jk}^2)}} \quad (2)$$

由公式(1)、(2)可得文档 P 的合成权值 W 的公式如下:

$$W=a*IR(P)+b*S(P) \quad (3)$$

其中 a、b 为调节因子。

设主题层次数为 m + 1,第 0 层对应于语义不相关层,第 1 层对应于"knowledge – path"的根结点,…,第 m 层对应于目标主题层,为了加快计算速度,可实例化 m 个分类器线程对应于主题层次 1 ~ m,用公式(3)并行地计算文档 P 与各层的合成权值 W_i($i=1\sim m$),并判断它是否超过阀值 Hv。由于从第 m 层至第 1 层,主题由窄变宽,若有满足情况,将文档 P 分配至层号较大的层次对应的 Web 文档表 $DescriptList_i$,否则将文档 P 分配至语义不相关层(层次 0)对应的 Web 文档表 $DescriptList_0$。同时析取文档 P 中的各 URL 链接,过滤已被爬取的 URL 链接,将所有新的 URL 链接分配至文档 P 所在主题层(层 i)对应的 URL 队列 $UrlQueue_i$。

3.2.3 侯选 URL 的排序

为了让与目标主题相关且重要的文档优先爬取到,我们应对 URL 队列 $UrlQueue_i$ 中的各超链分别排序。对于超链 URL u 考虑两个因素:首先可考虑 URL u 字符串与锚文本中的一些提示信息,其次是析取出 URL 的文档 P 的重要性。考虑前者,一个 URL 相应权

重 f_U 可由下式计算得出[2]：

$$f_U = \alpha * ImpUrlWt + \beta * ExpURlWt + \gamma * ExpAnchorWt \quad (4)$$

其中 ImpUrlWt 是 URL 中内在关键词 IKIs 的数目，ExpURlWt 是 URL 中外在关键词 EKIs 的数目。ExpAnchorWt 代表锚文本中外在关键词 EKIs 的数目。参数 α、β、γ 代表三种类型的相对重要性。内在关键词指那些暗示作者写作目的的词，外在关键词指那些主题层次图中有突出概念的词。

综合两个因素，由公式(3)和公式(4)可得超链 URL u 的合成权值 W_u：

$$W_u = \psi * f_U + (1-\psi) * W \quad (5)$$

其中 ψ 为调节因子。

将保存 W_u 到 $AnchList_i$ 三元组中，按照更新后的 URL 权值，对 URL 队列 $UrlQueue_i$ 中的侯选 URL 链接分别排序。

3.2.4　主题爬取 Web 文档算法

基于本体概念图的主题爬取 Web 文档算法步骤如下：

(1) 输入目标主题，初始 URL，阀值 Hv。

(2) 根据目标主题，从本体概念图中获取"knowledge－path"。

(3) 按照"knowledge－path"构造主题层次，最小层号(层 0)为语义不相关层，...，最大层号 m 为目标主题层(参照第 3.2.1 节)。

(4) 对各主题层(不相关层－层 0 除外)实例化一个分类器线程；

(5) 对各主题层初始化一个对应的 URL 队列($UrlQueue_i$，i = 0 ~ m)。

(6) 初始 URL 进入层 0 对应的 URL 队列 0($UrlQueue_0$)。

(7) 按 URL 队列名的逆序(如 $UrlQueue_m$，$UrlQueue_{m-1}$，…，$UrlQueue_0$)提取队首 URL 元素，爬取 URL 指向的 web 文档 d。

(8) 基于主题层对 Web 文档 d 分类及将文档 d 中的 URL 分配至主题层对应的 URL 队列，赋予了待爬取的 URL 对象层次语义信息(参照第 3.2.2 节)。

(9) 对 URL 队列($UrlQueue_i$，i = 0 ~ m)中的侯选 URL 链接分别排序(参照第 3.2.3 节)。

(10) 若所有 URL 队列不为空，则转(7)，否则结束。

4　结论

本文提出的自动构造电子课本学习系统，采用基于本体概念图的主题爬取技术爬取 Web 文档，基于主题层次对 Web 文档 d 分类及将文档 d 中的 URL 分配至与主题层对应的 URL 队列。赋予了待爬取的 URL 对象层次语义信息。同时对每个 URL 队列中的 URL 对象按合成权值 W_u 排序。实验表明[6]，层次语义引导主题爬取，可加快了主题爬取的速度。主题层次图中由根结点至目标结点，主题由宽变窄，学习者可对窄主题的目标主题层文档的学习，可加深概念含义的理解，同时可对宽主题的目标主题层邻层文档有选择学习，更容易把握概念的整体及相互关系。

参考文献

1 M Diligenti, F M Coetzee, S Lawrence, et al. Focused crawling using context graphs[A]. Proceedings of the 26th International Conference on Very Large Data Bases [C]. Cairo: Morgan Kaufmann Publishers, 2000. 527－534.

2 Jing Cheng, Qing Li, Liping Wang, et al. Automatically Generating an e－Textbook on the Web[A]. Lecture Notes in Computer Science 3143[C]. Berlin: Springer－Verlag Heidelberg, 2004. 35－42.

3 J Cho, H Garcia－Molina, L Page. Efficient crawling through URL ordering[A]. Proceedings of the 7th ACM－WWW International Conference[C]. Brisbane: ACM Press, 1998. 161－172.

4 Aggarwal C, Al－Garawi F, Yu P. Intelligent crawling on the World Wide Web with arbitrary predicates[A]. Proceedings of the Tenth International World Wide Web Conference[C]. Hong Kong: ACM Press, 2001. 96－105.

5 邓志鸿、唐世谓、张谓 等，Ontology 研究综述[J]，北京大学学报(自然科学版)，2002，38(5)：730－738.

6 曾义聪、杨贯中、刘柯，基于概念树的主题爬取技术研究[J]，科学技术与工程，2005，5(12)：785－790.