

融合知识需求的扩展主题图在 E-learning 中的应用

谭文安^{1,2}, 崔林山¹, 唐安琼²

(1 南京航空航天大学 计算机科学与技术学院, 江苏 南京 210016;

2 上海第二工业大学 计算机与信息学院, 上海 201209)

摘 要: 本文针对传统教学中知识滞后学非所用的不足, 结合 E-learning 环境可以便捷更新和获取学习资源的优势, 提出一种融合知识需求的扩展主题图在 E-learning 中的应用方法, 其采用 Web 信息抽取方法从招聘类网站获取知识需求并整合于传统主题图, 形成一种扩展主题图模型用于 E-learning 的学习资源组织与导航, 以热点的知识需求引导教育者更新知识体系有针对性的开发组织学习资源, 引导学习者开展切实所需目的明确的在线学习。

关键词: 主题图; 在线学习; web 信息抽取; 知识管理

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)11-0001-04

Integrating Knowledge Requirements into Topic Map for E-learning

TAN Wen-an^{1,2}, CUI Lin-shan¹, TANG An-qiong²

(1 School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China;

2 School of Computer and Information Science, Shanghai Second Polytechnic University, Shanghai 201209, China)

Abstract: The paper addressed the problem of people are learning hysteric knowledge in traditional learning environment by an extended topic model which integrate knowledge requirements of society into E-learning environment. We can use web information extraction methods to grab knowledge requirements from job-list website and integrate it to the traditional topic map, to form an extended topic map model for the E-learning. The proposed model can be used to guide educators to update their outdated knowledge system based on requirements and guide learners to learn the valuable and required knowledge.

Key words: topic maps; e-learning; web information extraction; knowledge management

1 引言

E-learning 作为教育与信息技术结合的应用, 从最初的教材电子化到 web2.0 支持的交互式学习, 再到支持个性化在线学习, 其融入了越来越多的教育理念与计算机科学^[1]. 近年来, 在 E-learning 学习资源的组织导航机制研究上整合了较多计算机学科的研究成果, 例如结合社会化 web, 组建的虚拟学习社区形成了以内容为核心, 通过博客, wikes, 资源分享等社会化交互方式驱动的学习资源的创建, 获取^[2-3]; 整合以主题图为代表的知识组织机制, 弥补

了传统“资源, 索引和元数据目录”三要素为核心的资源组织管理机制很难体现知识内部关联的不足; 采用人工智能与数据挖掘技术, 用于分析学习者学习兴趣及行为, 实现学习资源的个性化导航等. 不难看出, 现有的研究关注于在学习主题明确的前提下更好的组织导航学习资源以改善学习体验, 提高学习效率^[1]. 本文在现有研究的基础上针对学习主题滞后引起的学非所用问题, 提出融合知识需求的扩展主题图模型, 以当前社会的知识需求为引导, 帮助教育者更新旧的知识体系开发与组织社会所需的知识服务, 帮助学习者开展主题明确的学习。

2 知识需求的获取与处理

通过实地的调研获取社会对人才应掌握知识的需求,既费时又费力,而且不能实时的查看和更新需求的变化.为此,我们采用 web 信息抽取方法从招聘类网站发布的企业招聘信息中抽取知识需求,主流招聘类网站招聘信息一般包括如下几项信息:1. 工作地点 2. 招聘人数 3. 工作年限 4. 学历要求 5. 职位名称 6. 知识技能需求描述. 这些信息从招聘类网站抽取后进行结构化存储,方便进一步的统计与挖掘,发现我们需要的热点知识需求.

2.1 基于 web 信息抽取的知识需求获取

web 信息抽取技术,用于从海量的半结构化的数据中发现有价值的信息.当前主要形成了两类主流的 web 信息抽取技术,第一种称为基于包装器的信息抽取,第二种称为基于概念模型的信息抽取^[4].前者侧重分析 web 页面结构,通过对样本页面内容与组织结构的归纳学习获取可以高效描述信息特征的模板,使用获取的模板作为后续信息抽取的规则.而后者基于概念模型的信息抽取,多用于纯文本类型,将文本信息与定义的概念关系匹配后把文本描述为概念明确的对象.

以招聘类网站一段典型的招聘信息 html 片段为例见图 1.将该片段转化为一棵 DOM 树结构后,我们发现从 1,2,3 节点可以直接获取到工作地点,需求人数,职位名称信息,而 4 节点中包含的知识技能需求描述有较多的冗余数据.所以使用单一基于包装器的抽取获得的数据有较多冗余信息,而基于概念模型的信息抽取获得的数据又不完备.所以本文采用基于包装器与基于概念模型相结合的方式完成招聘信息的获取.具体采用基于包装器方法获取招聘信息中的工作地点,招聘人数,工作年限,学历要求,职位名称等简单信息,再将基于概念模型的方法用于包含冗余数据的知识技能描述信息的抽取.

```

<table>
<tr>
1  <td class="txt_1" width="12%">工作地点: </td>
2  <td class="txt_2" width="12%">招聘人数: </td>
3  <td class="txt_3" width="12%">招聘人数: </td>
4  <td class="job_detail">系统架构设计师<div style="padding-bottom:30px;">
    <div style="width:100%; height:100%; border:1px solid black; padding:5px;">
    <div style="padding-bottom:30px;">
    <br>1、软件项目的总体设计师,负责软件新产品的开发与集成
    <br>2、熟悉软件开发工程规范:掌握C++, Java等一种以上的开发工具
    <br>3、精通SQL server,Oracle等一种大型数据库系统!
    </div>
    </div>
    </div>
</td>
</tr>
</table>

```

图 1 招聘信息的 html 片段

融合两种信息抽取技术的招聘信息获取过程描述为图 2.首先,定义行业知识概念本体,采用本体方式定义知识概念一方面定义行业知识技能名的语料库,另一方面区分了同一知识技能在不同行业的不同含义.之后针对 Html 格式的招聘信息,将其转换为 DOM 树结构对于不包含冗余数据的信息,可直接抽取节点信息,对于包含冗余数据的知识技能需求描述信息再结合定义的行业知识概念本体进一步抽取,识别出其中的知识技能实体.

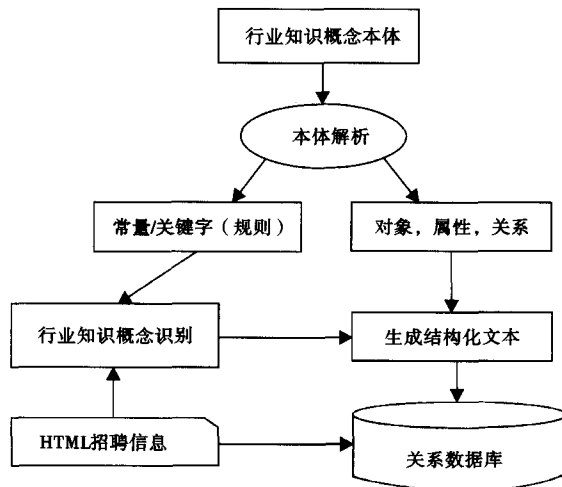


图 2 招聘信息的抽取过程

2.2 招聘信息处理

对于抽取并结构化存储的招聘信息进一步处理才能发现对我们更有价值的信息,主要采用统计和数据挖掘技术.首先,按行业、按地域显示 top N 知识技能的需求;其次,使用 Apriori 算法^[5],挖掘知识技能需求间的关联规则.例如,知识主题 A 和 B 总是共现与企业的知识技能需求中,那么教育者可以考虑将 A 与 B 知识主题的学习安排在一起.总之,根据统计挖掘结果,教育者可以针对真实的知识需求重构知识体系,开发组织新的知识产品;学习者可以根据当前社会的 top N 知识技能需求再结合学习兴趣开展目的明确的学习.

根据上述方法开发的知识需求抽取工具已经在计算机行业的知识需求抽取中得到验证,发现了上海,北京地区自 2010 年以来对 android,IOS,java, Objective-C 知识的需求剧增,符合移动互联网到来后对移动应用开发人才需求的增长.

3 融合知识需求的扩展主题图模型

3.1 主题图在 E-learning 中应用

主题图作为一种描述信息资源知识结构的元数

据格式,由描述事物的主题(topic),构成主题间关系的关联(association),以及主题与信息资源关系的事件(occurrence)三个核心要素构成^[6].它可以表示概念体系,可以定位知识概念所在的资源位置,可以表示知识概念间的相互联系.主题图(Topic Map)的这种TAO结构应用于E-learning中可以克服学习资源缺乏有效的信息组织及语义描述的缺点.如图3描述了E-learning中基于主题图的学习资源导航.其上层为学习资源的逻辑组织,是一个结构化的语义网.其中 $T_1, T_2, \dots, T_n$ 表示了不同的知识主题, β 表示了这些知识主题间的语义关系.下层为学习资源的物理组织,即具体的学习资源, $R_1, R_2, \dots, R_n$ 可以是ppt,文本,视频等各种类型的教材,关系 α 表示某知识主题的学习会被导航至与其有 α 关系的学习资源处.

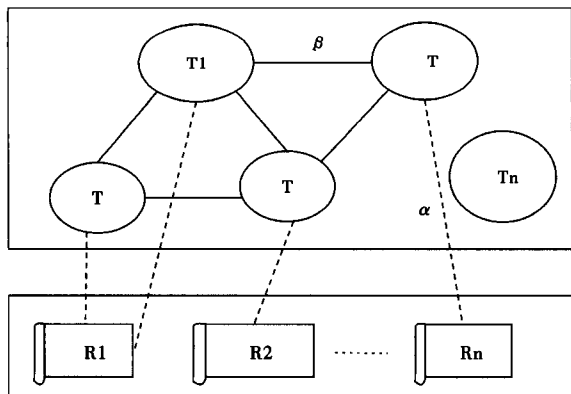


图3 E-learning 中基于主题图的学习资源导航

3.2 融合知识需求的扩展主题图

结合第二部分获取的知识需求扩展E-learning中的传统主题图,首先从层次上除了学习资源的物理组织层与逻辑组织层外添加知识需求层,同时增加了层次内部及层次之间的语义关系,融合知识需求后的扩展主题图描述如图4所示.

该模型形式化描述为一个七元组

$(T, RT, R, \alpha, \beta, \theta, \gamma)$

其中:

$R = \{R_1, R_2, \dots, R_n\}$, 表示具体的学习资源;

$T = \{T_1, T_2, \dots, T_n\}$, 表示特定行业的知识主题;

$RT = \{RT_1, RT_2, \dots, RT_n\}$, 表示从招聘类网站抽取出的知识需求;

$\alpha = \{\text{resourceRef}\}$, 表示 T 与 R 之间的导航关系,例如 T_m 与 R_n 存在该关系则 T_m 知识主题的学习将被导航至 R_n 学习资源;

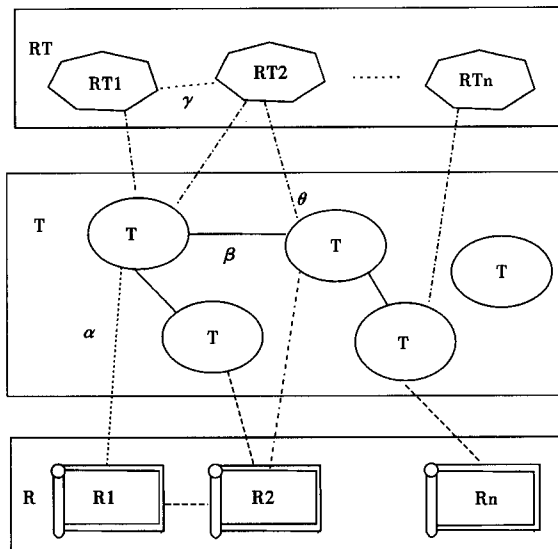


图4 E-learning 中融合知识需求的扩展主题图

$\beta = \{\text{partof}, \text{preLearn} \dots\}$, 表示 T 内部各知识主题之间一组关系集. 例如, reLearn 表示知识主题间的学习顺序, partof 表示知识主题间的隶属关系等;

$\theta = \{\text{knowledgeRef}\}$, 表示知识需求主题 RK 与知识主题 T 之间的关系. 例如, $RT_m \text{ knowledgeRef } T_n$ 表示学习者为满足 RT_m 的需求需要学习 T_n 知识主题;

$\gamma = \{\text{co-occur}\}$, 表示 RT 内部抽取的知识需求主题间的关系集. 例如, RT_m, RT_n 间存在 co-occur 关系, 表示 RT_m 和 RT_n 以大于阈值的概率共现与社会对人才的知识需求中.

模型的实现上采用XML机制描述七元组的定义以支持计算机的处理,参考ISO/IEC3250定义的基于XML的主题图表示XTM^[7],除XTM中的19个基本标签<topic>, <basename>, <basename-String>, <association>, <instanceOf>, <occurrence>...外添加扩展模型特有的<knowledgeRef>, <co-occur>等,这样可以兼容过去没有知识需求层描述的学习资源.

4 结束语

本文针对传统教育知识滞后导致的学非所用问题^[8-9],提出了E-learning环境下面向需求的学习资源导航方法.具体的,采用结构识别与概念识别相结合的Web信息抽取方法获取社会热点知识需求,并将该需求整合于传统的基于主题图的学习资源导航中,方便教育者针对社会热点的知识需求更新知识体系开发组织学习资源,引导学习者开展切实所需

目的明确的在线学习^[10-11]。目前,结合各行业专家定义的行业知识概念本体开展知识需求抽取工具的进一步测试,及扩展主题图模型的解析是下一步主要的工作。

参考文献:

- [1] Sergio Martin, Gabriel Diaz. New technology trends in education: seven years of forecasts and convergence [J]. Computer & Education, 2011(57): 1893-1906.
- [2] Zongkai Yang, Qingtang Liu. Research and development of web-based virtual online classroom[J]. Computers & Education, 2007(48): 171-184.
- [3] Hsi-mei Chen, Cris Yu, Cheng-sian Chang. E-home-book system: a web-based interactive education interface[J]. Computers & Education, 2007 (49) 160-175.
- [4] Alberto H F Laender, Sergio Martin. A brief survey of web data extraction tools[J]. SIGMOD Record, 2002,31(2):84-93.
- [5] Rakesh Agrawal, Ramakrishnan Srikant. Fast algorithm for mining association rules in large database[C] // VLDB,1994. Chille, Santiago, 1994: 487-499.
- [6] Lu Jiang, Jun Liu, Qinghua Zheng. ETM toolkit: a development tool based on extended topic map[C] // International Conference on Computer Supported Cooperative Work in Design. Santiago, 2009.
- [7] Mihaela M Brut, Florence Sedes. A semantic-oriented approach for organizing and developing annotation for E-learning[C] // IEEE Transactions On Learning Technologies. USA: LOS Alamitos, 2011: 239-249.
- [8] Peter A C Smith, Moira McLaughlin. Knowledge management: people are important [J]. Journal of Knowledge Management Practice, 2004:2-7.
- [9] Lijia Lin, Robert K Atkinson. Using animations and visual cueing to support learning of scientific concepts and processes [J]. Computers & Education, 2011 (56): 650-658.
- [10] Maja Cukusic, Niksa Alfirevic, Andrina Granic, et al. e-Learning process management and the e-learning performance: results of a European empirical study[J]. Computers & Education, 2010 (55): 554-65.
- [11] Lida Xu, WenAn Tan. An approach to enterprise process dynamic modeling supporting enterprise process evolution[J]. Information System Frontiers, 2008 (10):611-624.

作者简介:

谭文安 男,博士,教授.研究方向为知识管理、智能信息系统、协同计算等.

崔林山 男,硕士.研究方向为 E-learning、服务计算等.