

基于 XML 和 Mobile Agent 的个性化知识搜索和推荐系统建模与设计

田永鸿 黄铁军 高文

中国科学院计算技术研究所数字化技术室, 北京 100080

{yhtian,tjhuang,wgao,xczhang}@ict.ac.cn

摘要 本文所研究的主要问题是虚拟校园环境下如何根据学生的兴趣和当前的学习活动找到他所需的领域知识。我们基于 RDF 模型和 Bayes 网络对虚拟校园环境下的学生行为和兴趣进行建模; 然后在此基础上, 设计并实现了一个用于虚拟校园环境中使用的个性化知识搜索和推荐 Agent 系统原型。

关键字 虚拟校园、资源描述框架 RDF、学生兴趣模型、个性化知识搜索和推荐

Modeling and Designing Personalized Knowledge Search & Recommendation System Based on XML and Mobile Agent Technology

Y.H. Tian T.J. Huang W. Gao

Digital Multimedia Lab, Institute of Computing Technology

China Academy of Science, China Beijing 100080

Abstract In this paper we focus on how a student in virtual campus is able to find relevant domain expertise according to his interests and current actions. Based on Resource Description Frame model and Bayesian Net, we model student behavior and interests; Moreover, we have designed and implemented a prototype for personalized knowledge search & recommendation agent system used in virtual campus.

Keywords Virtual Campus、Resource Description Frame、Interests Model、Personalized Knowledge Search & Recommendation

1 引言

网络教育是近年来教育界、IT 产业界和学术界讨论和研究的热点。作为网络教育区别于传统远程教育的一个新特点, 虚拟校园的基本特征是支持在线服务人员在网络空间中向学生

提供各种服务; 面向各种层次的教师和学生, 提供丰富的、易用的先进交流手段, 并最终构建面向教育应用的先进的虚拟社群系统[高文等, 2000]。

随着 Internet 中信息量的不断加大, 人们发现越来越难于准确地搜索到自己所需的资料或得到别人的帮助[A.S. Vivacqua,1999]。同样, 在虚拟校

园环境中, 学生往往需要就某个问题向教授或已学过的同学求教, 或当他在虚拟课堂上听课时忽然需要查询与课程知识点有关的一些技术资料或电子文档, 但这时利用常规的搜索引擎输入一个关键词后他常常会发现得到超出想象的数量的 URL 地址, 他必须花大量的时间才能识别出自己感兴趣的東西。

本文所需要解决的主要问题就是如前所述的这类问题, 即是在虚拟校园环境下如何根据学生的兴趣和当前的学习活动找到他所需的知识或相关的领域专家。课题所研究的内容属于中国科学院研究生院网络多媒体教育工程的一部分。我们的解决方案是, 为每一个在虚拟校园环境下学习的学生提供一个“虚拟学习助理”, 它能自动记录学习在虚拟校园中的活动并进行学生行为建模和兴趣建模, 并能根据学生的兴趣和当前的学习活动提供个性化的知识搜索和知识推荐。

本文首先将基于 XML 对虚拟校园环境下的学生行为和兴趣进行建模; 然后在此基础上, 设计并实现了一个用于虚拟校园环境中使用的个性化知识搜索和推荐 Agent 系统原型; 最后将我们的原型系统与相关的工作进行了比较并指出将来需要进行的工作。

2 虚拟校园环境下的学生建模

系统实现个性化服务的一个重要的基础是要有能准确记录学生在虚拟校园环境下所从事的活动, 反映学生兴趣及知识领域的学生模型。虚拟校园环境下的学生模型包括学生行为模型和学生兴趣模型。这些模型随时根据学生的行为进行动态的更新。

2.1 学生行为建模

学生行为建模描述了学生在虚拟校园环境下的活动信息和行为模式。在传统的 WEB 站点中, 并没有对用户

的活动进行建模处理, 而是通过对 Web 服务器的日志数据进行数据挖掘和知识发现来挖掘用户的行为模式 [Fayyad U M,1996]。由于 Web 日志数据的不规则性, 随着 Web 站点数量的不断增加以及对每个 Web 站点的访问信息量的不断加大, 这种方式逐渐显示出局限性, 即在大量的信息中发现相关的信息并挖掘出用户的行为模式将变得越来越困难 [L. Ardissono,1999]。一种可选的解决方案是基于系统的日志文件对学生的行为进行建模, 然后基于学生的行为模式进行挖掘学生的兴趣, 这样不仅可以大大降低数据挖掘的复杂度, 而且还可以向用户提供个性化的服务。在本课题中我们将基于 RDF 模型对虚拟校园环境下的学生行为进行建模。

定义 1 (学生行为模式) 在虚拟校园环境下, 一个典型的学生行为可以描述为: (O, P, ST)

其中: (1) O 是对象标识及其属性;

(2) P 是谓词, 描述了对象的行为、对象与主题间的行为关系等; 在虚拟校园中, 谓词一般包括有动作、时间、地点等;

(3) ST 是主题集, 描述了本次对象行为的对象、所利用的资源等; 主题可以是整个网页, 也可以是网页中的一部分或网页的集合, 还可以是通过 Web 访问的对象、虚拟校园中开展的活动主题等。主题间根据抽象的层次构成语义 Bayes 网络。

资源描述框架 RDF (Resource Description Frame) 提供了一种便于将 WWW 数据转换为机器可以理解的知识库的元数据描述标准。它着眼于自动便捷地处理网络资源, 基于对象模型来描述概念及其相关关系并以 Web 兼容的方式来标注事实和大纲 [S.Staab et al,2000]。因此, 我们用 RDF 来描述虚拟校园中的学生行为模式, 其中对象用 RDF 的申明来加以描述, 谓词用 RDF 的属性来加以描述, 而主题集则

用 RDF 的资源来加以描述。

例：学生 A 正在虚拟教室 M 中上《计算机操作系统原理》的课程，同时又在参加一个有关 LINUX 的讨论组。

则该学生的行为用 RDF 描述如下：

```
<rdf:RDF>
  <rdf:Description
about="Computer Operating
System Principle">
  <s:Property>
    <v:Teacher>Pref.
Wang</v: Teacher>
    <rdf:subClassOf
rdf:"#Computer Science"/>
  </s:Property>
  <s:Object>
    <v:Name>A</v: Name>
  </s:Object>
  <s:Behavior>
    <v:Action>Attend
Lectures</v: Action>
    <v>Date>Am 10:20
10/20/1999</v: Date>
    <v:Place>Virtual
```

```
Classroom</v: Place>
  </s:Behavior>
</rdf:Description>
<rdf:Description
about="Linux Symposium">
  <rdfs:subClassOf
rdf:resource="# Computer
Operating System Course"/>
  <s:Object>
    <v:Name>A</v: Name>
  </s:Object>
  <s:Behavior>
    <v:Action>Converse
</v: Action>
    <v>Date>Am 10:20
10/20/1999</v: Date>
    <v:Place>Virtual
Chat Room</v: Place>
  </s:Behavior>
</rdf:Description>
</rdf:RDF>
```

如图 1 所示：

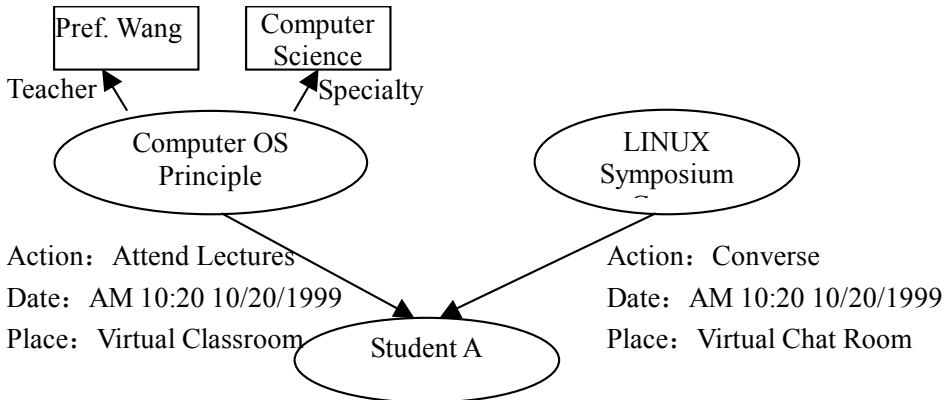


Fig 1. An Example of Student Actions Described By RDF Model

2.2 学生兴趣建模

学生兴趣模型描述了学生的动态特性，是学生模型的一个重要组成部分。这里我们基于传统的用户兴趣建模方法，即通过分析学生的初始信息、学生从事某项活动或主题的频率、阅

读某类文章和关键字列表的时间以及学生对在线调查表的反馈等行为来对学生的兴趣进行建模 [W.R.Picard,1998]，分析结果是由一些带权关键字向量 (Weighted Keyword Vectors, 简称 WKV) 来表示学生兴趣，

其中权值表示学生对关键字相应对象的兴趣程度 [G. Salton and C. Buckley,1987]。

我们将 KWV 应用到虚拟校园领域，得到学生的兴趣向量 (Interests Vector) 定义：

定义 2 (兴趣向量 IV) IV 可表示为： $O(W)$ ，这里： O 表示兴趣对象关键字，而 W 是根据学生操作兴趣对象的频率和时间计算出来的权值，一般简称为兴趣度 (Interests Degree)，一般采用 TFIDF 算法计算 [A.Moukas, 1996]，有：

$$W = \prod_c \times T_f \times idf_k \quad (1)$$

其中， \prod_c 是调节常量， T_f 是兴趣对象在当前访问的 WEB 页面或虚拟场景中出现的频率 (称为对象频率，Object Frequency)， idf_k 是访问页面或场景在学生的全部访问活动中的频率 (称为页面频率或场景频率，Scene Frequency)， idf_k 定义为：

$$idf_k = \log\left(\frac{N}{df_k}\right) \quad (2)$$

其中 N 是学生访问的页面或场景的总数量， df_k 是第 K 个页面或场景在所有访问中出现的频率。

兴趣向量间按兴趣对象的分类层次建立起的继承与引用关系，每一个兴趣对象都与其继承的兴趣对象、引用的兴趣对象以及由它派生的兴趣对象相关，从而可以建立一个学生的兴趣向量间的联系网。由于兴趣向量间的这种相关关系具有不确定性，具有继承与引用关系的兴趣向量间还可以进行因果推理，因此我们用 Bayes 网络来表示学生的兴趣向量间的联系网，建模学生的兴趣。一个典型的学生兴趣模型 (部分) 如下图所示：

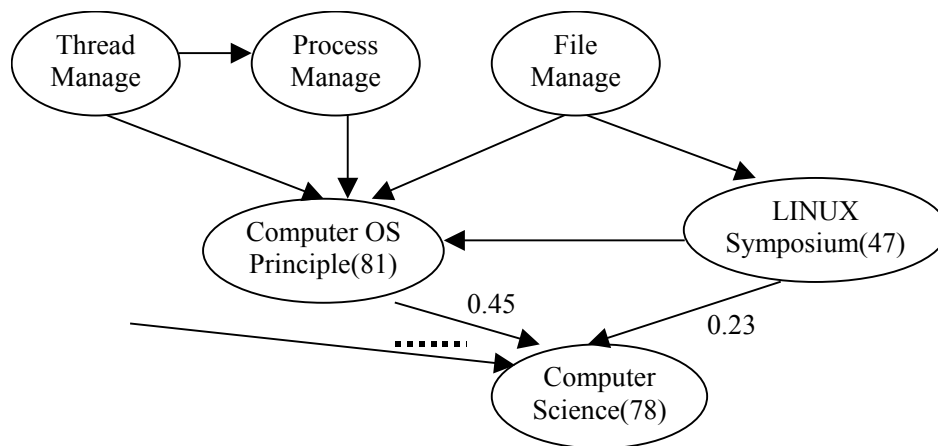


Fig 2 A Bayesian Net Indicating Student Interests

与学生行为建模类似，我们采用 RDF 存储学生的兴趣模型。图 2 所示的学生兴趣模型 (部分) 描述如下：

```

<rdf:RDF>
  <rdf:Description
    ID="Computer Science">
    <rdf:Type
  
```

```

Resource= "http://www.w3.org/TR/1999/PR-rdf-schema-19990303
#Class"/>
  
```

```

<s:InterestsDegree>78
</:InterestsDegree>
  
```

```

<rdfs:subClassOf
  rdf:resource=http://www.w3.org/TR/1999/PR-rdf-schema-19990303#Resource
  v:probability=1/>
</rdf:Description>
<rdf:Description
ID="Computer OS Princople">
  <rdf:Type
    Resource= "http://www.w3.org/TR/1999/PR-rdf-schema-19990303#Class" />

    <s:InterestsDegree>81
  </:InterestsDegree>
<rdfs:subClassOf
  rdf:resource="#Computer Science"
  v:probability=.45/>
</rdf:Description>
<rdf:Description
ID="Linux Symposium">
  <rdf:Type
    Resource= "http://www.w3.org/TR/1999/PR-rdf-schema-19990303#Class" />

    <s:InterestsDegree>47
  </:InterestsDegree>
<rdfs:subClassOf
  rdf:resource="#Computer Science"

```

```

  v:probability=.23/>
<rdfs:subClassOf
  rdf:resource="#Computer OS Princople"
  v:probability=.15/>
</rdf:Description>
.....
<rdf:RDF>

```

学生兴趣模型的使用主要在于分析当前的使用场景以及学生的行为模式，依据使用的经验与学生的行为历史来提供智能服务上，对不同场景的分析和学生的行为模式都使用学生兴趣模型来优化智能服务的内容。学生兴趣的建模过程描述如下：

- 1) 学生行为感知：提取当前场景的特征信息，感知学生在虚拟校园环境内的行为，这一般通过传感器实现；
- 2) 学生行为建模：实时捕获或从日志文件中获得学生在虚拟校园内的活动信息，进行学生行为建模；
- 3) 行为分类：根据行为的主题，对行为进行聚类处理；
- 4) 兴趣对象的收集：收集新增的学生行为模型中的主题信息；
- 5) 兴趣向量的计算：按照 TFIDF 算法计算学生的兴趣向量；
- 6) 兴趣网络的更新：利用新计算出的兴趣向量更新兴趣网络。

3 个性化知识搜索和推荐 Agent 系统

学生建模的目的是为了向虚拟校园环境下的学生提供个性化的推荐与信息服务。在本部分，我们将讨论如何

将学生的兴趣模型嵌入个性化知识搜索和推荐 Agent 系统。

3.1 系统结构

个性化知识搜索和推荐 Agent 系统的基本功能是向虚拟校园环境下的学生提供个性化的知识搜索服务和知识推荐服务。具体地，系统包括以下功

能:

- 1) 记录所有在线的学生;
- 2) 感知学生在虚拟校园中的行为和活动, 并随时记录学生在校园中的活动, 充当学生的数字化学习秘书;
- 3) 建模学生的行为和兴趣, 以便及时掌握学生的学习状况和学习兴趣;
- 4) 根据学生的兴趣以及当前的行为准实时地向学生提供个性化的推荐, 并以生动的形式向学生呈现;
- 5) 根据学生的搜索请求并结合学生的兴趣进行个性化的知识搜索等。

系统由以下两大部分组成, 如图 3 所示:

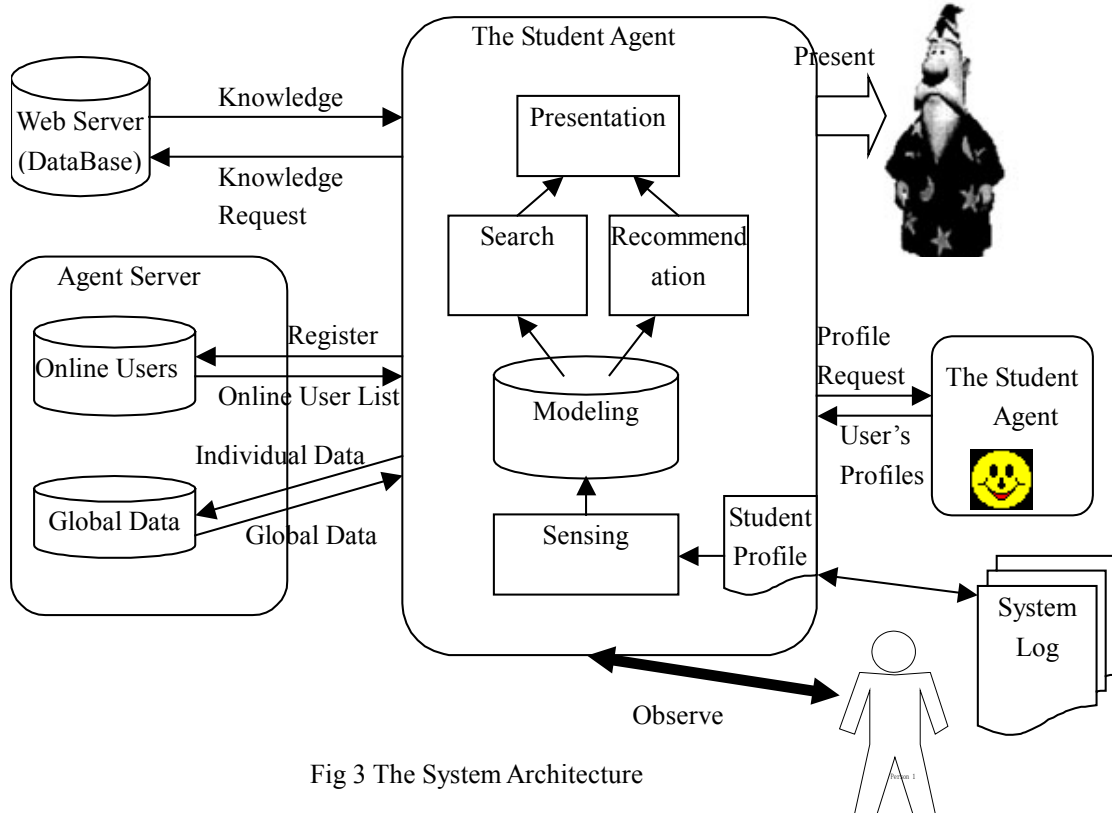


Fig 3 The System Architecture

➤ 学生 Agent (Student Agent): 它是每一个在虚拟校园中的学生的数字化学习秘书和虚拟的学习助理, 其功能可用“感知、建模、推荐、搜索、表示”等十个字概括, 即感知学生在校园中的行为并充当学生的数字化学习助理, 对学生的行为和兴趣进行建模以便提供个性化的服务, 根据学生当前的行为和兴趣进行个性化的知识推荐, 根据学生的搜索请求并结合学生兴趣进行

个性化知识搜索服务, 以生动的形式将信息呈现给学生。

➤ Agent 服务器 (Agent Server): 它提供虚拟校园系统中的所有学生 Agent 的注册、Agent 通信与控制信息的管理等功能。

3.2 学生 Agent

系统的体系结构中的核心部分是学生 Agent, 其结构如图 4 所示。它由下列模块组成:

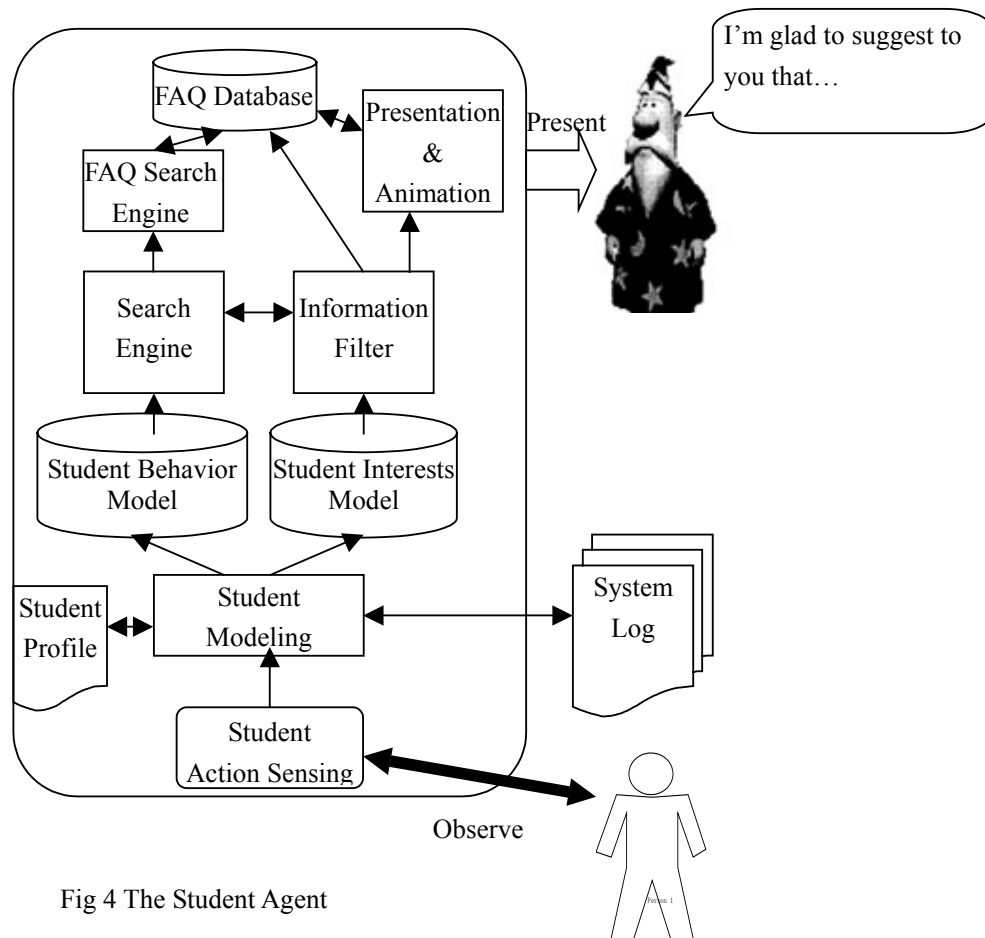


Fig 4 The Student Agent

- 1) 学生行为感知模块 (Student Action Sensing Module): 感知学生在虚拟校园内的活动和行为, 为学生充当数字化学习秘书。
- 2) 学生建模部分 (Student Modeling Component): 动态地建立和更新学生的行为模型和兴趣模型。
- 3) 学生模型 (Student Model): 代表一个特殊学生所需的特征和信息, 包括学生的概要信息(在学生注册时提交)、学生的行为模型以及学生兴趣模型。
- 4) 搜索引擎 (Search Engine): 包括知识搜索和专家搜索两部分, 即根据给定的搜索请求在系统各知识库 (或 Web Server) 中搜索符合要求的知识 (按知识相关程度排序), 或者与其他学生 Agent 进行交互以便获得相关的领域专家信息
- 5) 个性推荐部分 (Personalized Recommendation Component): 包括 FAQ 搜索引擎、FAQ 知识库以及信息过滤三部分。FAQ 知识库主要管理学生在过去一段时间内经常访问的知识、根据学生的主要兴趣提取的个性化推荐的知识等。信息过滤是对搜索引擎所得的知识或信息进行分类和过滤, 并根据学生的兴趣高低向学生推荐。
- 6) 知识表示和 Agent 呈现 (Knowledge Presentation & Animation): 将搜索出的知识或向用户推荐的知识以友好的形式向用户推荐, 包括 Agent 动画的选择和表情、动作合成。

3.3 系统工作流程

对虚拟校园环境下的学生进行建模

的目的是用于提供智能服务上，而系统向学生的智能服务主要包括个性化知识推荐和信息搜索两方面，下面分别讨论其工作流程。

个性化知识推荐的工作流程如下：

- 1) 感知学生行为并对学生行为和兴趣进行建模；
- 2) 搜索 FAQ 知识库，查看是否符合用户当前的行为和兴趣的知识或领域专家信息：如果有，则提交给信息表示模块呈现给用户；
- 3) 如果 FAQ 知识库中没有与学生当前兴趣相关的知识或专家信息，或者学生对推荐给他的知识或专家信息提出不满意，则根据学生新的知识请求启动搜索引擎搜索进行知识搜索与专家搜索；
- 4) 信息过滤器对搜索引擎返回的结果进行分类，并根据学生的当前行为、兴趣以及搜索请求对结果进行相关度计算[N.W.V. Dyke etc al,1999]，然后按相关度大小排列结果并呈现给学生；
- 5) 根据学生对搜索结果的引用或访问情况将相关的知识或信息写入 FAQ 知识库。

个性化知识搜索的工作流程如下：

- 1) 学生提出信息搜索请求，学生 Agent 从中提取请求关键字向量；
- 2) 学生 Agent 启动 FAQ 搜索引擎，提取出匹配学生搜索请求的知识；
- 3) 与此同时，学生 Agent 启动搜索引擎，同时进行知识搜索与领域专家搜索，即一方面通过远程调用或移动方式访问虚拟校园环境所链接的公开知识库（如在线数字图书馆、在线资料库等），另一方面向虚拟校园中的所有其他学生 Agent 发出

请求，提取具有相关兴趣的其他学生 Agent 的 Profile 信息（每一个学生可以决定是否容许自己的 Agent 是否可以向其他学生 Agent 提供信息服务）；

- 4) 信息过滤器对搜索引擎返回的结果进行分类，并进行相关度计算后呈现给学生。

4 与相关工作的比较

当前，许多系统都通过研究用户的兴趣来实现个性化信息服务，但这些系统大多针对网上的信息浏览系统而设计的，对于三维虚拟环境下的用户行为分析以及用户建模则相对较少。比较而言，与本系统相关的工作有两类：第一类是网页信息推荐系统，如美国 MIT 的 Media Lab 的 Letizia 系统 [H.Lieberman, 1997]、Let's Browser 系统 [H.Lieberman, 1999] 以及 Stanford 大学的 "Fab" adaptive Web Page [M. Balabanovic, 1997] 等；第二类是基于文本级或二维图象级的智能搜索与信息推荐系统，如 Butterfly 系统 [N.W.V. Dyke etc al, 1999]、Expert Finder 系统 [A. S. Vivacqua, 1999] 等。其中 Butterfly 系统用于因特网在线聊天系统（Internet Relay Chat, 简称 IRC）中进行个性化推荐；而 Expert Finder 系统立足于用户在进行 JAVA 程序设计时利用 Agent 来发现相关的领域专家。

与上述系统相比，本原型系统的特点是基于三维虚拟环境下对学生的行为和兴趣进行感知和建模，并在此基础上向学生提供实时的数字秘书服务、个性化知识推荐与信息搜索。系统在构筑上针对三维虚拟环境的特点采用了对多场景支持的技术，使学生在校园中的每一个活动都可以得到智能的支持服务；同时系统利用 Microsoft 的 Agent 控件具有表情丰富的动画人物以及语音合成和语音识别功能来进行信息的表示以增加系统的

亲和感与界面的友好性。

5 进一步的工作

基于学生模型的个性化知识搜索和推荐 Agent 系统虽然能够对学生在虚拟校园中的行为和兴趣进行建模以提供一定程度上的个性化服务，但其中几乎都没有提供识别用户的情感的能力。而随着交互式三维虚拟环境的进一步发展，迫切需要面向学生提供多层次的、丰富易用的交流手段，包

括自然语言和体态语言的输入和表达、情感表达和计算等。美国 MIT 媒体实验室 Picard 教授的情感计算 (Affective Computing) 和情感智能 (Emotional Intelligence) 理论 [R.W.Picard,1995, R.W.Picard,1998]向我们展示了在三维虚拟环境下建模用户情感的可能性。因此，如何利用 Agent 来建模学生的情感和兴趣，以此向学生提供人性化的 Agent 服务，将是我们下一步构建高级虚拟校园环境的主要问题。

参考文献

- [高文等, 2000]-高文, 刘峰, 黄铁军等, *数字图书馆——原理与技术实现*, 清华大学出版社, 2000, 10, 416-437
- [A.S.Vivacqua,1999]- Adriana S. Vivacqua, *Agents for Expertise Location*. In Proceedings 1999 AAAI Spring Symposium on Intelligent Agents on Cyberspace, Stanford, CA,USA, March 1999
- [A.Moukas, 1996]- Alexandros Moukas, *Amalthea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem*, in the proceedings of the Conference on Practical Applications of Agents and Multiagent Technology, London, April 1996
- [Fayyad U M,1996]- Fayyad U M, Piatetski-Shapiro G, Smith P. *From data mining to knowledge discovery: An overview*. In: Fayyad U M, Piatetsky-Shapiro G, Smith P, Uthurusamy R eds, *Advances in Knowledge Discovery and Data Mining*. Boston: AAAI/MIT Press, 1996, 1~34
- [G. Salton and C. Buckley,1987]- G. Salton and C. Buckley. *Text Weighting Approaches in Automatic Text Retrieval*. Cornell University Technical Report 87-881, 1987.
- [H.Lieberman, 1997]- H.Lieberman, *Autonomous interface agents*, ACM Conference on Computers and Human Interaction [CHI-97], Atlanta, May, 1997
- [H.Lieberman, 1999]- H.Lieberman, N. Van Dyke, A. Vivacqua, *Let's browser: a collaborative agent*, Knowledge-Based System 12(1999), 427-431
- [L. Ardissono,1999]- Liliana Ardissono, Luca Console, Ilaria Torre, *Exploiting user models for personalizing news presentations*, In Proceedings of the 2nd Workshop on Adaptive Systems and User Modeling on the WWW, CMU,USA,1999
- [M. Balabanovic,1997]- Marko Balabanovic. *Agents'97 Marina del Rey CA USA, An Adaptive Web Page Recommendation Service*, ACM Press, 1997
- [N.W.V. Dyke etc al, 1999]-Neil W. Van Dyke, Henry Lieberman, Pattie Maes, *Butterfly:A Conversation-Finding Agent for Internet Relay Chat*, ACM IUI'99 Redondo Beach CA USA, Jan 1999
- [R.W.Picard,1998]- R.W.Picard, *Toward Agents that Recognize Emotion*, In Actes Proceedings IMAGINA, Monaco, March 1998
- [R.W.Picard,1995]- R.W.Picard, *Affective Computing*. Technical Report 321, MIT Media Lab Perceptual Computing Section, Cambridge, MA, USA

[S. Staab et al,2000]- Steffen Staab, Michael Erdmann, Alexander Maedche,etc, *An Extensible Approach for Modeling Ontologies in RDF(S)*, In Proceedings of ECDL 2000 Workshop on the Semantic Web.