基于加权概念网络的用户兴趣建模*

许欢庆[1] 王永成[1] 孙强[1]

[1]（上海交通大学计算机系 上海 200030）

E-mail: xuhuanqing@sjtu.edu.cn

摘要：用户兴趣建模是互联网个性化信息服务的关键技术。本文中，提出一种基于加权概念网络的用户兴趣建模方法。该方法利用动态学习算法，挖掘蕴含在用户反馈文档中的概念及其概念关系，建立加权概念网络的用户模型，从而捕捉和表述用户兴趣偏好。基于加权概念网络用户兴趣模型，提出了检索提问个性化理解，以及文档个性化重评价的实现方法。为了检验提出方法的建模性能，设计了信息过滤仿真试验。测试结果表明：加权概念网络有较好的用户建模性能。

关键词： 加权概念网络; 用户建模; 概念映射

引言

个性化信息服务已成互联网技术的研究热点。其通过分析用户浏览行为以及反馈的评价信息，建立反映用户信息需求的兴趣模型，依据用户模型提供信息服务，提高了互联网服务质量。用户兴趣模型是个性化信息服务的基础，模型的准确性、时效性直接决定服务质量的优劣。用户兴趣建模的方法很多[1][2][3][4]，多采用抽取文档中特征词作为用户兴趣的特征项，通过学习算法，动态调整特征项的权值适应用户兴趣的变化。由于，特征词在语义上的多义性，一定程度上影响了用户模型的准确度。本文中，我们提出将概念作为用户兴趣的特征项，利用动态学习算法，挖掘蕴含在用户反馈文档中的概念及其概念关系，建立加权概念网络的用户模型，从而捕捉和表述用户兴趣偏好。基于加权概念网络用户兴趣模型，实现了检索提问的个性化理解，以及文档个性化重评价。

1 加权概念网络

加权概念网络从结构体系而言，类似语义网络。60 年代，美国学者 R. Quillian 对人类联想记忆进行研究，他认为词汇在人的头脑中是以概念网络的方式存在，是一种语义网络，语义的内容即为概念的内容。我们在语义网络的基础上进行相应的改变，设计加权体系，建立加权概念网络。

1.1 概念

概念是对象本质在人脑中的映象。词或词组是概念的一种表述形式。区别于词与词组，概念具有语义唯一性，并且，在语义空间上，概念间存在丰富的语义关联关系。概念词典 HowNet[5]对概念以及概念关系的类别进行总结。根据信息服务的需求出发，我们选择其中几种用于加权概念网络的建设，

* 本项目受到国家自然科学基金资助(60082003)。

其中包括：上下位关系(Hyponym)、同义关系(Synonym)、部件-整体关系(Part)、施事/经验者/关系主体-事件关系(Agent-Event)、受事/内容/领属物等-事件关系(Patient-Event)，同现关系（Concept Co-occurrence）。

1.2 加权概念网络的结构

加权概念网络继承语义网络的结构，同时，引入了加权体系。网络由节点和弧构成。网络节点包含概念信息以及权重，概念信息包括表示同一概念的特征词。节点间弧表示概念间语义关联关系，其权值表示关系受用户重视程度。加权概念网络 $WCN=(N,E)$，N为节点集，E为弧集。节点集N，$N=\{n \mid n \in NodeSet\}$。$NodeSet=\{(c,w^N) \mid c \in C, 0 \le w^N \le 1\}$。$w^N$ 为概念节点权重，C为概念集。弧集 E，$E=\{e \mid e \in EdgeSet\}$

$EdgeSet=\{(\langle n,n'\rangle, w^E) \mid n,n' \in N, rel(n.c,n'.c) \in Concept\_Relation, 0 \le w^E \le 1\}$。

其中，$Concept\_Relation=\{R_{Hyponym}, R_{Synonym}, R_{Part}, R_{Agent-Event}, R_{Patient-Event}, R_{co}\}$，$w^E$ 为弧的权重，$rel(x,y)$ 表示概念x, y间关系类别。

1.3 加权概念网络的扩展

加权概念网络的建立是以网络扩展的方式进行。网络扩充包括二种情况，分别进行如下处理。

- 已有概念节点进入

已存在的概念进入概念网络，对网络中已有概念以及概念间关系有所加强。设概念c_{old}，权重为w_{old}。在概念网络中，找到对应的概念节点n。调整节点对应的权值调整如下：

$$n.w^{N'} = \alpha \cdot n.w^N + \beta \cdot w_{old} \quad (1)$$

其中，$\alpha = \dfrac{n.w^N}{n.w^N + \delta \cdot w_{old}}$，$\beta = \dfrac{\delta \cdot w_{old}}{n.w^N + \delta \cdot w_{old}}$，$\delta$ 为调整系数。

概念网络中，与节点n相连的弧构成集合$E'=\{\langle n,n'\rangle \mid \langle n,n'\rangle \in E\}$。对弧$e \in E'$，分别调整$e$的权值，

$$e.w^{E'} = e.w^E + \alpha_{old} w_{old} \cdot w_{rel(n,n')} \quad (2)$$

α_{old}为加权系数。$w_{rel(n,n')}$为对应节点n和n'间关系类型的权重。概念关系一定程度上体现了概念间语义距离，我们对不同类型的概念关系设置权重：

$w_{Hyponym}$=0.4，$w_{Synonym}$=0.8，w_{Part}=0.3，w_{Agent_Event}=0.6，w_{CO}=0.7 $w_{Patient\_Event}$=0.6。

- 新概念节点加入

新概念加入概念网络，应考虑与网络中其他概念节点的关系。设新概念c_{new}，权重为w_{new}。将概

念 c_{new} 作为新节点插入概念网络中，新节点权重如下公式计算：

$$w^N = \delta_{new} w_{new} \tag{3}$$

δ_{new} 为新节点的加权系数。借助概念词典，查询与概念 c_{new} 存在关联关系的概念，构成关系集 R'。设概念 c_i，与概念 c_{new} 的关系为 $rel(c_{new}, c_i)$。在概念网络中找寻 c_i 对应的概念节点，如果概念节点存在，添加新弧，并对弧进行赋权。概念 c_i 对应的节点为 n'，

$$w_e^E = \alpha_{new} \cdot \left(\frac{w_{new}^{\ 2} + n'.w^{N2}}{w_{rel(c_{new}, c_i)}} \right) \tag{4}$$

其中，α_{new} 为加权系数，

1.4 相似度计算

基于加权概念网络的用户兴趣建模体系中，文档与用户兴趣间的相关度可以通过计算文档特征概念向量与兴趣概念网络的相似度获得。加权概念网络 WCN_1，文档 d 与之的相似度计算步骤如下：

- 分析 WCN_1 的主题概念

对概念网络中节点权值进行如下归一操作：

$$\overline{w}_{n_i}^N = \frac{w_{n_i}^N}{\sum_{n_j \in WCN_1.N} w_{n_j}^N} \tag{5}$$

其中，$w_{n_i}^N$ 为节点 n_i 的权值，$\overline{w}_{n_i}^N$ 是归一后的权值。同样，对弧权值进行规范化处理。概念节点 n_i 在网络中的主题重要性 W_{n_i} 可以采用如下公式计算：

$$W_{n_i} = w_{n_i}^N + \eta \sum_{\langle n_i, n_j \rangle \in WCN.E} \overline{w}_{n_i n_j}^E \cdot w_{rel(n_i.c, n_j.c)} \cdot \overline{w}_{n_j}^N \tag{6}$$

其中，η 为系数。对计算值进行如下规范化，

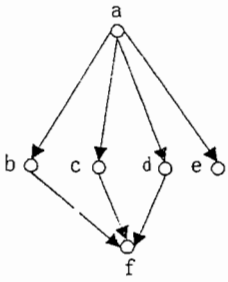
$$\overline{W_{n_i}} = \frac{W_{n_i}}{\sum_{n_j \in WCN_1.N} W_{n_j}} \tag{7}$$

- 相似度计算

文档 d_0 的特征概念向量表示为 $d_0 = (c_1, c_2, ..., c_m)$，$c_i$ 为特征概念。w_i 为概念 c_i 在文档中的权重，$1 \le i \le m$。相似度计算公式如下：

$$Sim(d_0, WCN_i) = \frac{\sum_{k=1}^{m} w_k \cdot \overline{W_{n_k}}}{\sqrt{\sum_{k=1}^{m} w_k^{\,2}} \cdot \sqrt{\sum_{k=1}^{m} \overline{W_{n_k}}^{\,2}}} \tag{8}$$

下面给出了一个计算相似度的简单示例，设概念网络 WCN_1 如图 1 所示；



节点权重

w_a=5, w_b=2, w_c=2, w_d=1, w_e=4, w_f=1

弧权重

w_{ab}=1, w_{ac}=2, w_{ad}=3, w_{ae}=3, w_{bf}=2, w_{cf}=2, w_{df}=1

弧代表的关系

a→b:Hyponym, a→c:Synonym,
a→d:Part, c→f:CO, a→e:Patient,
b→f:Hyponym, d→f:Agent

图1 相似度计算示例

文档 d 的概念向量为 $d=(a,b,c)$，对应权重为 w_a =0.7, w_b =0.2, w_c =0.1。概念网络的主题概念计算结果为：

表 1 概念节点的主题重要度

a	B	C	d	e	F
0.403	0.096	0.175	0.096	0.139	0.091

计算文档和概念网络相似度

$$Sim(d, WCN_i) = \frac{403 * 0.7 + 0.096 * 0.2 + 0.175 * 0.1}{\sqrt{0.7^2 + 0.2^2 + 0.1^2} \cdot \sqrt{0.403^2 + 0.096^2 + 0.175^2}}$$，结果值为 0.960。

2 用户兴趣建模

用户的信息需求跟用户所处的社会、文化、职业背景相关。通常，用户存在多个信息需求领域，不同兴趣领域又存在不同的兴趣偏好。例如，专业背景为心脏病病理研究的医生关注心脏病治疗最新研究动态，而对其他病症的信息可能仅关心简单护理知识。因此，用户兴趣建模时，应该根据用户的不同需求领域建立相应的加权概念网络。同时，捕捉需求领域内用户兴趣的变化，动态调整对应的概念网络进行适应。

2.1 文档特征概念向量

词是概念的一种表述形式，经过抽象化后形成概念。我们在提取文档主题概念时，首先抽取出反映文档主题的特征词，建立特征词向量，通过概念映射获得特征概念向量。

2.1.1 特征词抽取

用户需求领域确定，反馈文档分为相关，不相关二种。设需求领域 I_i，判定为相关的文档组成集合

S_1，不相关文档构成集合S_2，$S_0 = S_1 + S_2$。文档d 进行分词和剔除禁用词处理，获得特征词集$T = \{t_1, t_2, ..., t_m\}$，$w_i (1 \le i \le m)$是特征项$t_i$ 在文档d 中的权重。计算特征词t_i 在文档集中区分文档相关与否的能力 *DV*，计算公式如下：

$$DV_i = \frac{freq_{1i}}{n_1} \sum_{d_j \in S_1} f_{ji} - \frac{freq_{2i}}{n_2} \sum_{d_j \in S_2} f_{ji} \tag{9}$$

其中，$freq_{1u}$ 表示特征项t_i 在集合S_1 中出现的文档频次，$freq_{2i}$ 是特征项t_i 在集合S_2 中出现的文档频次，n_1 是集合S_1 的文档数，n_2是集合S_2 的文档数，f_{jk} 为特征项t_k 在文档d_j 中的词频。特征词权重如下计算：

$$w_i = f_i * DV_i \tag{10}$$

f_i 为特征项在文档中的词频。权重超过阈值的特征词，组成文档特征词向量。规范化向量中特征词权重。

2.1.2 概念映射

查询概念词典 HowNet，可以获得特征词对应的概念，完成概念映射。文档中一些特征词存在多重语义，可能对应多个概念；此外，存在一些未在概念词典中标注的新词，具有较强的主题提示作用，比如：人名，特定的事件名等，需要特殊处理。

文档特征词相互之间存在语义关联关系，这种关系可以用于确定特征词的语义。从形式上而言，特征词间语义关联关系一定程度上表现为特征词间共现频率。我们给出共现的定义：设特征词x 和y 出现在文档d_i 中的同一个句子认为二者共现，词间共现率 *CO* 如下计算：

$$CO_d(x, y) = \frac{f_{xy}^d}{f_x^d + f_y^d} \tag{11}$$

f_{xy}^d 为特征词 x 和 y 共现句子数，f_x^d 为特征词 x 的词频。文档特征词集$T = \{t_1, t_2, ..., t_m\}$，特征词$t_i \in T$，对应的概念有 c_1，c_2，…，c_k。根据共现率，选取特征词，构成词集$T_{CO}^{t_i} = \{t_j \mid CO_d(t_i, t_j) \ge \varphi\}$，$\varphi$ 是预设阈值。特征词t_i 隶属于概念c_i 可能性如下计算：

$$p_{c_i}^{t_i} = \sum_{t_k \in T_{CO}^{t_i}} \lambda_p CO_d(t_i, t_k) \delta_{rel(c_i, Concept(t_k))} \tag{12}$$

其中，λ_p 为系数，$Concept(x)$ 表示特征词 x 的概念。$\delta_{rel(c_i, c_j)}$ 的取值跟概念间关系有关，如果概念c_i 和c_j 之间的关系为 *Co* 时，$\delta_{rel(c_i, c_j)}$=0.7；关系为 *Agent* 或 *Patient* 时，$\delta_{rel(c_i, c_j)}$=0.5；关系为 *Synonym* 时，$\delta_{rel(c_i, c_j)}$=0.2。选择隶属度最大的概念作为其特征词的概念。一些未在概念词典中标注的新词，通常具有

很强的提示作用，我们直接保留其作为特定概念，加入特征概念向量。

2.2 用户模型建立

对应于需求领域的用户反馈文档集作为兴趣建模的训练样本集。用户需求领域I_i，对应加权概念网络为WCN_{I_i}。初始状态下，WCN_{I_i}为空。训练样本集中的相关文档d_i，经过预处理后转化成特征概念向量形式。将特征概念向量表示的文档以概念网络扩展的方法，加入到概念网络WCN_{I_i}中。可见，用户兴趣建模是动态增量学习过程。随着需求领域的相关文档增加，概念网络将逐渐膨胀，噪音随之增加。为了将概念网络约束在一定范围内，一定时间内，应对概念网络进行精简。将主题重要性低的概念节点以及与之相连的关系删除。

2.3 用户模型的激活

互联网个性化信息服务主要有信息推荐、信息过滤和个性化信息检索三种。用户模型被激活的方式主要如下：

- 检索提问个性化理解

信息检索质量提高的前提是正确理解用户提交的检索提问。检索提问的正确理解包括两个方面：检索提问所属的需求领域和所属需求领域的用户偏好。用户的信息需求通过用户提交的检索提问式表达，通过聚类分析检索提问式可以获得用户的信息需求领域集。同时，借助信息需求领域的用户兴趣概念网络模型，理解用户提问的真实信息需求。首先，用户检索提问q作为文档进行处理，计算q对应的特征概念向量与各用户需求领域对应概念网络的相似度。根据计算出来的相似度，确定用户提问的需求领域归属：相似度最大的概念网络对应的需求领域是提问所属的领域。加权概念网络中概念节点间关系表示概念间的语义关联关系，所带权值表示用户对概念关系的重视程度。我们利用提问所属的加权概念网络进行用户提问扩展。设WCN_i为提问q归属领域的概念网络，概念特征向量$q = (c_1,c_2,\ldots,c_m)$。对于概念c_i，在概念网络中寻找到与之对应的节点，如果节点存在，与之相连弧的权重超过给定的阈值，则将关联节点包含的概念c'对应的特征词按如下规则加入提问中，实现提问扩展：

规则1．如果概念c'与c_i存在同义或共现关系，其对应的特征词加入概念c_i对应特征词的“或集合”中。

规则2．如果概念c'与c_i存在施事或受事关系，其对应的特征词加入概念c_i对应特征词的“与集合”中。

- 文档个性化重评价

用户的信息需求不同，导致对文档的相关评价不尽相同。文档的评价应以用户为中心。个性化重评价过程就是计算文档主题与用户兴趣模型的相似度，对文档集进行重排序。

3 性能评价

3.1 试验设计

3.1.1 数据集

为了检验加权概念网络用于用户兴趣建模的性能，我们设计了信息过滤仿真试验，并对测试结果进

行分析。选取 25 个实际用户的评价文档集作为试验数据集。数据集共包含 500 篇文档，涉及 5 个话题(体育、科技、经济、医疗、军事)，每个话题对应 100 篇文档，我们将话题作为用户的需求领域，分别进行测试。用户对文档的评价分为相关和不相关两种，按评价时间顺序排列话题对应文档集。话题对应的评价文档集被分割为两个集合：前 60 篇组成训练集，用于建立话题对应的用户兴趣加权概念网络模型，剩余的文档组成测试集。

3.1.2 对比方案

为了进行性能对比，我们选择向量空间模型(VSM)作为用户兴趣建模对比方案。VSM 用户建模时，采用同样方法抽取特征词项，构成用户兴趣向量。兴趣向量与文档间的相似度用于文档相关度预测。

3.1.3 评价方法

11 Point Precision Average 被选作为性能评价方法。基本步骤是：

- 计算测试集文档的用户兴趣相关度，并按照相关度计算值递减排列文档。假设测试集 D，文档 $d_i \in D(1 \le i \le n)$。如果 $i < j$，则 d_j 的相关度小于 d_i。
- 计算测试集中被用户评价为相关的文档的召回率(*Recall*)和准确率(*Precision*)。相关文档 d_i，召回率和准确率的计算公式如下：

$a=$ 位于该文档 d_i 之前的所有相关文档的个数；$b=$ 位于该文档 d_i 之前的所有文档的个数；$c=$ 整个文档集中相关文档的个数；

$$\mathrm{Re}call = \frac{a}{c} \quad , \qquad \mathrm{Pr}ecision = \frac{a}{b} \tag{13}$$

- 将区间[0,1]分割成十等分。11 个边界点值作为召回率，计算每个召回率下的准确率，即在召回率区间范围内最大的准确率值。将准确率的计算值表示成 11 维向量形式。试验测试阶段，两种方法分别对应了 125 个向量。
- 平均处理向量，绘制 Precision vs Recall 曲线图。

3.2 测试结果与分析

图 3, 4, 5, 6, 7 绘制了两种建模方法对于五个话题领域的 *Precision* vs *Recall* 曲线图。曲线表明：采用加权概念网络用户兴趣模型的信息过滤系统的准确率和查全率较向量空间模型有较大程度的提高。可见，加权概念网络提高了用户兴趣建模的性能。

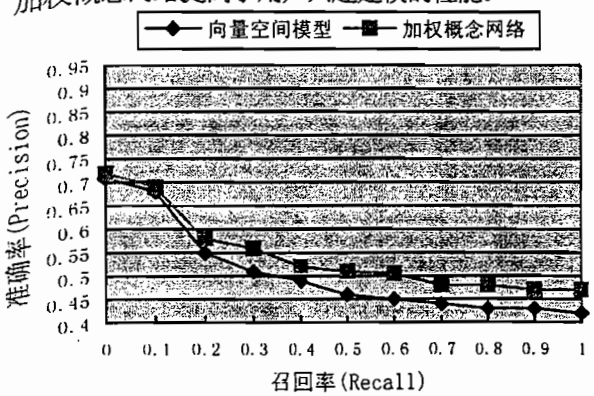


图3 性能比较图(军事)

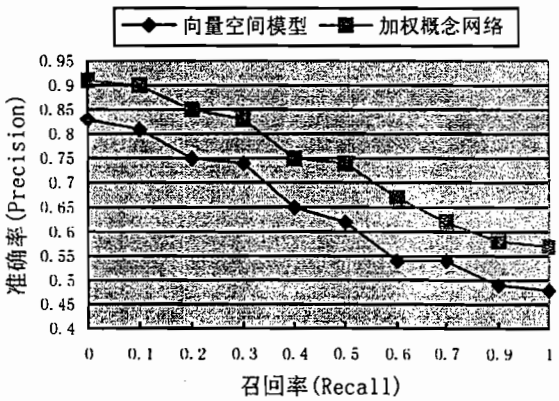


图4 性能比较图(体育)

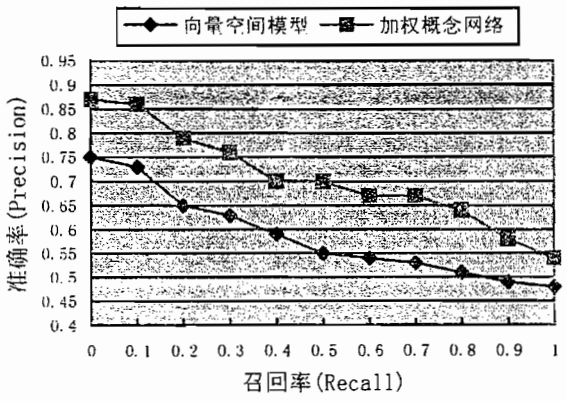


图5 性能比较图(科技)

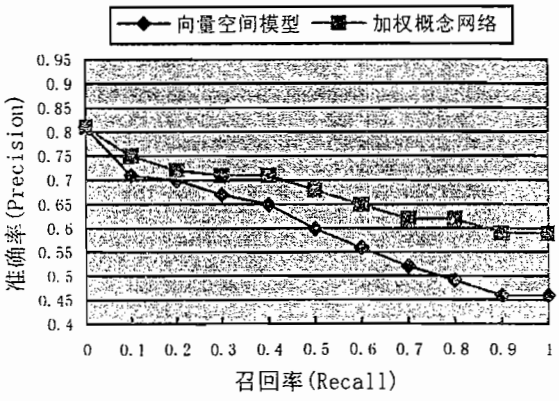


图6 性能比较图(经济)

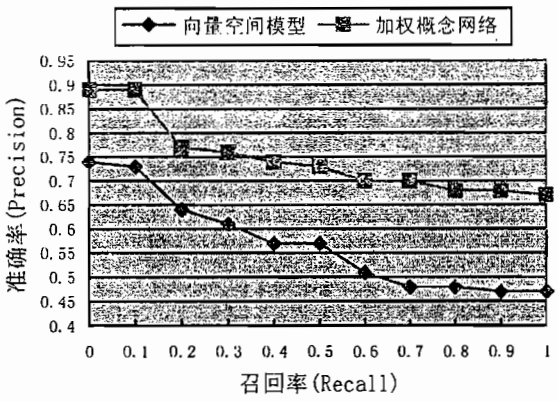


图7. 性能比较图(医疗)

4 结束语

随着互联网信息的快速膨胀，用户对信息服务的个性化需求愈加迫切。利用加权概念网络对用户兴趣建模，能够较准确地捕捉和表述用户信息需求和偏好，保证了个性化信息服务的质量。通过相应的改进，加权概念网络也可用于互联网信息分类领域。我们接下来的研究重点是概念节点和弧的权值调整方法，以使用户模型能够动态适应兴趣的迁移。

参考文献:

[1] T. Yan. H. Garcia-Molina. SIFT- a tool for wide-area information dissemination. In Proc. 1995 USENIX Technical Conf. 177-180.

[2] M. Pazzani. J. Muramatsu. D. Billsus. Syskill&Weber: identifying interesting web sites. In Proc. 13th Natl. Conf. on AI, 1996.

[3] Wiener. E., Pederson. A.S., A Neural Network Approach to Topic Spotting, Proc. of the 4 th Annual Symposium on Documents Analysis and Information Retrieval, 1995 .

[4] Min-Huang Ho, Ming-Chun Cheng, Yue-Shan Chang, Shyan-Ming Yuan, A GA-based dynamic personalized filtering for Internet

search service on multi-search engine. Electrical and Computer Engineering, 2001. Canadian Conference on , Volume: 1 , 2001 age(s): 271 -276 vol.1.

[5] How-Net, http://www.keenage.com

作者简介： 许欢庆，1973 年生，博士研究生，主要研究领域为智能信息检索，Web 挖掘，信息过滤与推荐。 王永成，1939 年生，博士生导师，主要研究领域为信息检索，自动摘要，自然语言理解。 孙强，1974 年生，博士研究生，主要研究领域为自然语言理解，信息家电。

User Modeling Based on Weighted Concept Network

XU Huan-qing[1], WANG Yong-cheng[1], SUN Qiang[1]

[1](Department of Computer Science, Shanghai Jiao Tong University, Shanghai 200030, China)

E-mail: xuhuanqing@sjtu.edu.cn

Abstract: User modeling is one of the crucial techniques of personalized information service. In this paper, we propose a new approach for user modeling based on Weighted Concept Network. This approach presents user's preference using concepts and concept relations implied in documents proposed by the user's relevance feedback. Incremental learning mechanism was applied to structure the Weighted Concept Network for user's interests. The personalized query expansion and re_ranking algorithm based on Weighted Concept Network were implemented. The simulated experiment about information filtering was designed. Experimental results indicate that this approach has better performance in user modeling.

Keywords: Weighted Concept Network; User Modeling; Concept Mapping