用户信息保护下的学习资源知识点自动提取

谢 铭[1,2] 吴产乐[1]

(武汉大学计算机学院 武汉 430072)[1] (广西经济管理干部学院 南宁 530007)[2]

摘 要 提出一种用户信息保护下的网络学习资源知识点内容自动提取方法,即在信息保护层中加入用户信息保护状态HMM模型,一旦判断保护状态无效,自动退出知识点内容提取流程,防止用户信息受到侵犯。使用用户信息保护HMM模型,对4家学习网站227名用户的查询浏览行为、用户链接、用户配置信息的真实数据集进行了实验,结果表明,进行500次随机消息测试时,模型对用户信息保护状态的判断正确率为94%,对虚假安全消息的误判率为0.04。根据4家学习网站在120天中的用户评分数据,系统使用后的平均分数较系统使用前平均增幅达23.23%。

关键词 用户信息保护,隐马尔可夫模型,动态聚合,语义级共享

中图法分类号 TP391.1 **文献标识码** A

Topic Extracting with User Information Protection on Web

XIE Ming[1,2] WU Chan-le[1]

(Computer School of Wuhan University, Wuhan 430072, China)[1]

(Guangxi Economic Management Cadre College, Nanning 530007, China)[2]

Abstract This paper proposed a method of topic extracting with user information protection in Web learning resources. We added the HMM model of user information protect stated into user information protection layer. The model can exit topic extracting procedure automatically once judging the invalid state of user information protection to prevent violations of user information. This paper evaluated HMM model of user information protection with the real data sets of 227 user's query browsing behavior, user link, the user configuration information in four study sites. The experimental data show that for 500 random massages test, the correct ration of the model judgment on the user information protection states is 94%, the incorrect ration of that on false security messages is 0.04. According to the user ratings data of four learning sites in the 120 days, the average rate of increase of user rating reaches to 23.23% after the system is used.

Keywords User information protection, Hidden Markov model, Dynamic aggregation, Semantic level sharing

1 前言

在信息挖掘的过程中用户隐私易受侵犯,主要表现为用户查询浏览行为、用户链接、用户配置信息3方面保护缺失。在用户信息得到保护的前提下依赖人工方式对海量以非结构化的文本或多媒体形式存在的课程信息进行知识点内容提取将耗费大量时间和人力。如何在非结构化学习资源中实现知识点自动提取,在保护用户信息的同时,快速、准确地获得自己想要的知识点信息是下一代电子学习系统理论和技术研究的核心问题。

用户信息保护下的学习资源知识点内容自动提取研究对于用户信息保护、海量离散、异构无序的网络学习资源的组织管理、动态聚合、语义级共享、精确查找有较大的理论意义和实用价值。

本文重点研究用户信息保护下的知识点内容自动提取技术,引入了用户信息保护模块,通过知识点内容将大量的课程信息用简洁、准确的形式呈现给用户。为了消除课程特有的冗余信息,本文提出了一种基于聚类算法的知识点自动提取模型,其通过信息融合技术合并提取出知识点内容。

本文设计了一种基于OPTICS改进算法NOP的用户信息保护下多文档自动文摘系统,该系统对3个方面进行了改进:(1)通过用户信息保护模块对整个流程实施监控,在信息保护层中加入用户信息保护状态HMM模型,预测当前用户保护是否处于有效状态,当保护状态失效时,自动退出知识点内容提取流程。(2)对句子相似度计算方法进行了改进,根据句子中的词与词之间的语义关系计算句子之间的语义相似度,并在此基础上进行句子聚类,完成对子主题的划分,最后分别在各个子主题中以一定策略抽取出若干句子作为摘要句。再分别在各个子主题中以一定策略抽取出若干句子作为知识点内容描述句。(3)对原有的OPTICS算法进行了改进,将原算法的数据空间对应到句子的向量空间中去,并给出了一种重新定位稀疏节点的方法。

2 国内外相关研究

Wu等使用了一种特别的用户信息保护技术[1]。他们从单个可信来源开始,将标准可信来源中所需子集的属性加以

到稿日期:2010-04-25 返修日期:2010-08-02 本文受国家自然科学基金(60672051),武汉大学博士生科研自主基金(20102110101000110),广西教育厅科研项目(200911LX541),广西教育科学"十一五"规划课题(2010C186),新世纪广西高等教育教改工程项目(2010JGB125)资助。

谢 铭(1978-),男,博士,主要研究方向为语义挖掘、电子学习,E-mail:wolfgangtse@gmail.com;吴产乐 教授,博士生导师。

联系，构造用户信任关系。但是他们没有提出处理非直接信任关系的方案。而实现用户信息保护下的知识点提取，需要给出扩展可信来源、可信组元素的方法。Liu 等则从相反的角度来消除查询中的歧义[2]，他们处理较长的多词查询，利用查询内的上下文信息为查询中的每个词找到精确的词义。消歧的大体步骤是：首先识别出查询中的短语，并利用 WordNet 中的信息为短语中的词消歧，然后仍利用 WordNet 为查询中的其他词消歧。对于没有成功消歧的词，如果它在 WordNet 中出现频率最高的词义大于其他所有词义的频率和，则赋以它这个词义，否则借助一次 Web 检索来帮助消歧。消歧实验在 TREC13Robust 评测的 250 个查询上进行，结果表明，Liu 的消歧方法可以为查询中的 333 个歧义词全部成功消歧，精度为 90%，并在 5 个 TREC 数据集上将检索实验结果提高了 10%到 25%[3]。但实验结果同时表明，如果词义消歧存在 20%到 30%的错误率，那么成功消歧所带来的检索效果的改进都会被消歧失败导致的负面影响抵消，因此需要进一步优化 WordNet 在词义消歧、概念特征、概念距离向量中的应用。

Mihael Ankerst 提出的 OPTICS 算法具有基于密度聚类、跨数据集、可挖掘数据自身的语料结构等特性[4]。但该算法原来用于对数据空间的点聚类，为了将 OPTICS 算法应用在文档处理领域中，必须将数据空间模型与文档空间模型对应起来，使之适合知识点内容自动提取系统。因此，需要对 OPTICS 算法进行一些修改，可以将文档集合中的句子看作 OPTICS 算法中的数据点，通过对数据点进行聚类来对多文档集合中的句子进行聚类，从而挖掘出多文档集合的不同主题。传统的基于词形的相似度计算不能发现词汇之间的深层信息。与 OPTICS 使用的欧氏距离相比，文本处理领域更常用的相似度计算方法为余弦相似度，因为余弦相似度更关注向量方向上的一致性，即特征共现一致性，所以在词特征向量方面，余弦相似度更适合作为句子相似度计算。可以选择其中的词特征向量与词义特征两方面来分析和计算，将原方法中的词向量和词汇距离分别替换为概念向量和概念距离。

Newsblaster 是哥伦比亚大学在多文档文摘方面的一个科研成果[5]，它是一个新闻跟踪的工具，可以为每天的主要新闻做出相关的文摘。NeATS[6] 是南加州大学在多文档文摘技术上的成果，该系统通过可靠的统计信息来抽取重要的概念，根据句子的开始词信息和句子的位置信息来过滤句子。因为有的句子是以转折连词或动词开始，如果这些句子进入文摘句的话，会导致文摘的不连贯，所以得把这些句子过滤掉，利用 MMR 方法来去除冗余信息，对文摘句进行合理的排序（比如按时间顺序），形成文摘。与多文档自动摘要技术相比，知识点内容自动提取在主题概念聚类上有更高的要求。

3 方法

本文通过对用户信息保护信任方案及基于密度的算法 OPTICS 的研究，设计一个用户信息保护下基于句子聚类的知识点内容自动提取系统。该系统在用户信息保护监控模式下，采用隐马尔可夫模型，观察用户查询浏览行为、用户链接、用户配置信息的趋势，预测用户信息是否受到侵犯。通过计算句子之间的相似度进行聚类，把相同主题的句子聚集到一起，使得每个类集合都表示为学习资源文档集合的一个子主题，然后采用一定策略从每个子主题中抽取一定数量的句子生成知识点内容描述句，最后在充分保护用户信息的前提下将知识点内容向学习者呈现。本文的总体架构如图 1 所示。

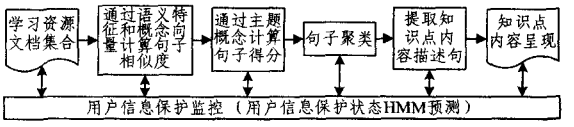

图 1 用户信息保护下知识点内容自动提取总体架构

总体技术路线图如图 2 所示。

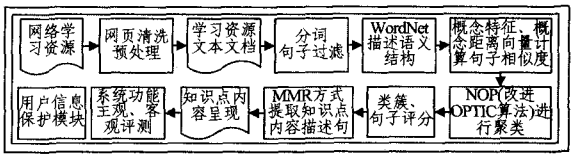

图 2 总体流程

4 详细算法设计

本文认为，用户信息保护下网络学习资源知识点自动挖掘过程中，设用户保护状态为离散马尔可夫链，用户信息保护可能处于两种不同状态：有效保护状态、无效保护状态。而影响用户保护状态的主要有用户查询浏览行为、用户链接、用户配置信息 3 个因素。虽然无法直接获知当前用户信息保护状态，但可以根据用户查询浏览行为、用户链接、用户配置信息的趋势采用隐马尔可夫模型（HMM）进行状态判断。根据总的用户信息保护趋势及各影响因素，可以得到隐马尔可夫模型的参数。

算法 用户信息保护状态判断

输入：状态数目；每个状态下可能的观察值

输出：观察值概率分布，发射概率

参数说明：Valid 为有效保护状态、Invalid 为无效保护状态、Behavior 为用户查询浏览行为保护、Link′为用户链接保护、Config 为用户配置信息保护

```
1. states=('Valid','Invalid')
  //状态数目，两个状态：有效保护状态、无效保护状态
2. observations=('Behavior','Link','Config')
  //每个状态下可能的观察值
3. start_probability={'Valid':0.6,'Invalid':0.4}
  //初始状态空间的概率分布
4. transition_probability={
  //与时间无关的状态转移概率矩阵
5. 'Valid':{'Valid':0.7,'Invalid':0.3},
6. 'Invalid':{'Valid':0.4,'Invalid':0.6},
7. }
8. emission_probability={
9. 'Valid':{'Behavior':0.1,'Link':0.4,'Config':0.5},
10. 'Invalid':{'Behavior':0.6,'Link':0.3,Config':0.1},
11. }
//给定状态下，观察值概率分布，发生概率
```

算法举例：start_probability 代表了初始判断时用户信息保护状态的不确定性（假设用户信息更多地处于有效保护状态下）。这个特定的概率分布并非平衡的，平衡概率应该接近（在给定变迁概率的情况下）{'Valid':0.571,'Invalid':0.429}。transition_probability 表示马尔可夫链下的保护状态变迁情况，本例中，如果当前用户信息保护处于有效状态，那

么下个监控时段用户信息保护处于无效状态的概率只有30%。emission_probability 表示保护状态的概率。如果保护状态有效，有50%的概率是用户配置信息被保护；如果保护状态无效，则有60%的概率用户查询浏览行为受保护。设用户状态保护观察结果如下：时段1用户查询浏览行为被保护、时段2用户链接被保护、时段3用户配置信息被保护。根据Viterbi算法可知在此时间序列下的用户信息保护状态。

在用户信息保护状态有效的前提下，为了挖掘词汇之间的深层联系，引入提取主题概念。通过两个句子中的主题概念构成概念向量空间，利用这两个向量夹角的余弦值计算相似度。两个句子 S_1 和 S_2 的主题概念构成的向量空间为 $V=\{C_1,C_2,\cdots,C_n\}$，其中 C_i 为句子的主题概念。句子 S_1 的向量 $V_1=\{\omega_1,\omega_2,\cdots,\omega_n\}$，其中 ω_i 为主题概念 C_i 在 S_1 中出现的次数。句子 S_2 的向量 $V_2=\{\phi_1,\phi_2,\cdots,\phi_n\}$，其中 ϕ_i 为主题概念 C_i 在 S_2 中出现的次数。则两个句子的概念特征向量相似度计算公式为：

$$\text{Similarity}(S_1,S_2)=\vec{V}_1\cdot\vec{V}_2=\frac{\sum_{i=1}^{n}\omega_i\phi_i}{\sqrt{\sum_{i=1}^{n}\omega_i^2}*\sqrt{\sum_{i=1}^{n}\phi_i^2}} \quad (1)$$

在词义特征方面主要采用词义距离来计算相似度。通过WordNet语义资源来描述多文档文摘的语义结构，在概念树中可以求出两个概念的距离，概念的距离定义为两个概念在概念树中的公共祖先节点到两个概念的最大深度。由此可得两个句子 S_1 和 S_2 的概念距离向量相似度 Similarity_d 为：

$$\text{Similarity}_d(S_1,S_2)=\frac{1}{\sum_{i=1}^{m}\sum_{j=1}^{n}\text{Distance}(X_i,Y_j)} \quad (2)$$

式中，句子 S_1 的概念集合为 $\{X_1,X_2,\cdots,X_i\}$，$(1\leqslant i\leqslant m)$；$S_2$ 的概念集合为 $\{Y_1,Y_2,\cdots,Y_j\}$，$(1\leqslant j\leqslant n)$。$\text{Distance}(X_i,Y_j)$ 为概念间的距离。最后得到句子相似度的计算公式为：

$$\text{SIM}(S_1,S_2)=\alpha*\text{Similarity}(S_1,S_2)+\beta*\text{Similarity}_d(S_1,S_2) \quad (3)$$

根据句子相似度计算方法，当词特征向量和词义特征信息项系数分别取0.699378和0.397034时结果最优。因此本文的概念特征向量和概念距离项的系数在实际计算中分别取 α 为0.7，β 为0.3。

计算权值之后对每个类簇中的句子进行排序，将权值较高的句子排在前面；并计算类簇中有效句子的数量（有效句子指的是权值不为0的句子）。为了去除多文档中的不重要信息（或称为噪声），本文将类簇分为有效类（valid cluster）和无效类（invalid cluster）。一般来说，类簇中的有效句子数量越小，说明这个类簇包含越少的信息，也就越远离多文档集合的中心主题。所以本文设定一个阈值 V，当类簇中的句子大于等于 V 时，为有效类；否则为无效类，可以忽略。本系统中 V 取 $0.8*\text{Avg}_{\text{stNum}}(C)$。

$$0.8*\text{Avg}_{\text{stNum}}(C_i)=\frac{C_i\text{ 中的有效句子数量}}{\text{所有类簇中的有效句子之和}} \quad (4)$$

当有效类提取出来之后，应该对类簇进行打分，并根据分值高低进行排序，决定抽取的顺序。本文通过类簇中包含的句子数及其包含的句子在各文档中的分布来确定该类簇表示的子主题与中心主题的相关度。如果一个类簇中包含的句子个数越多，并且句子分布在不同文档的数量越大，则认为该子主题越重要。类簇的得分计算公式如下：

$$SCORE(C_i)=\frac{NUM_i(Sen)*NUM_i(Doc)}{NUM(Sen)*NUM(Doc)} \quad (5)$$

式中，C_i 表示某个子主题，$Num_i(Sen)$ 表示该类的句子数，$Num_i(Doc)$ 表示该子主题中的句子分别所属原始文档的个数，$Num(Sen)$ 表示原始文档句子总数，$Num(Doc)$ 表示该原始文档中包含文档的个数。根据上述公式可以确定子主题与中心主题的相关程度，实际上也确定了类簇的抽取顺序。确定了类簇抽取的顺序之后可按顺序抽取每个类簇中的句子，若一轮抽取过后还未达到知识点描述文摘要求的最小长度，则返回头一个类簇继续抽取，直到满足要求的长度为止。为了减小知识点描述文摘的冗余，抽取句子时可参照MMR方法进行抽取。

5 实验及分析

用户信息保护实验

本文使用用户信息保护HMM模型对4家学习网站、227名用户的查询浏览行为、用户链接、用户配置信息的真实数据集（http://grid.whu.edu.cn/rainbow/datasets 提供下载）进行测试，以评价模型对防止用户信息被侵犯的计算有效性。测试时，随机引入虚假侵犯消息FI（当信息保护处于有效状态时宣称数据遭侵犯）、虚假安全消息FS（当信息保护处于无效状态时宣称数据安全）。对于产生的消息，启用用户信息保护HMM模型根据状态时序和每个状态下可能的观察值、初始状态空间的概率分布进行甄别。将观察序列['Behavior','Link','Config']输入模型中，得到Valid有效保护状态或Invalid无效保护状态。

如表1所列，随着消息测试次数的增加，用户信息保护HMM模型预测的正确性得到提升。进行100次消息测试时，用户保护状态正确判断的百分比为87%，而进行500次消息测试时，正确率提高到94%。其中，虚假安全消息FS的误判率从0.17降至0.04，说明模型在较大规模用户信息保护方面具有优势。

表1 不同消息测试下模型的错误矩阵（%correct是当消息发出时用户保护状态得到正确判断的百分比）

	100次消息测试		200次消息测试		500次消息测试	
	Valid	Invalid	Valid	Invalid	Valid	Invalid
FI	0.79	0.21	0.82	0.18	0.92	0.08
FS	0.17	0.83	0.13	0.87	0.04	0.96
	%correct=87%		%correct=92%		%correct=94%	

知识点内容自动提取效果实验

拟参照Edmundson评测法，设计两类实验以评测本文提出的知识点内容自动提取系统的效果。

主观评测：以1至2门课程为例，分别采用人工提取（由领域专家和教学名师共同确定）、自动提取两种方式，提取1至2门课程的知识点内容，由专家比较知识点内容自动提取的结果与知识点内容人工提取的结果所含的信息，然后给出一个等级分，等级为：完全不相似、基本相似、很相似、完全相似等。

客观评测：实验采用DUC2005数据集进行评测，并利用ROUGE 1.5.5进行评分。选取ROUGE-N（其中 N 取1到4）、ROUGE-L和ROUGE-W-1.2等6个任务进行测试。

本文在用户信息保护下的网络学习资源知识点内容自动

（下转第230页）

00). Boston, Massachusetts, 2000: 355-359

[5] Pei J, Han J, Mortazavi-Asl B, et al. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth[C]// Proc. of 17th International Conf. on Data Engineering (ICDE'01). Heidelberg, Germany, 2001: 215-224

[6] Yun U. A New Framework for Detecting Weighted Sequential Patterns in Large Sequence Databases[J]. Knowledge-based systems, 2008, 21(2): 110-122

[7] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts[J]. Ordered Sets, 1982: 445-470

[8] 刘宗田，强宇，周文，等. 一种模糊概念格模型及其渐进式构造算法[J]. 计算机学报，2007，30(2)：184-188

[9] Bordat J P. Calcul pratique du treillis de galois d'unc correspondence[J]. Math. Et Sci. Humaines, 1986, 96: 31-47

[10] Bettini C, Wang X S, Jajodia S. Mining Temporal Relationships with Multiple Granularities in Time Sequences[J]. Data Eng. Bull, 1998, 21: 32-38

[11] Godin R, Missaoui R, Alaoui H. Incremental concept formation algorithms based on Galois lattices[J]. Computational Intelligence, 1995, 11(2): 246-267

[12] 谢志鹏，刘宗田. 概念格的快速渐进式构造算法[J]. 计算机学报，2002，25(5)：490-495

(上接第 205 页)

提取系统中加入了投票功能，对用户反馈的信息加以收集。用户根据用户信息保护下的网络学习资源知识点内容自动提取平台的表现进行评分。评分标准：1～100 分。

本文收集了 4 个学习网站在 120 天中的用户评分信息，对比了用户信息保护下的网络学习资源知识点内容自动提取系统使用前后的平均分数，结果如表 2 所列。

表 2　4 个学习网站在系统使用前后的用户评分比较

	site1	site2	site3	site4
Before	63	67	71	61
After	76	79	84	83

分析

如图 3 所示，使用用户信息保护下的网络学习资源知识点内容自动提取系统后，4 家学习网站的用户评分较之前的均有提升，site4 增幅最大，达 36%；site2 增幅最小，为 17.91%；4 家网站用户评分增幅平均提高达 23.23%。真实数据集显示的实验结果证明了本文提出的方法的有效性。但是本文也注意到实验中用户评分的提高可能会受其他因素的影响。例如，site4 在使用本系统前，用户评价最低，但之后评分的迅速提高可能是受到其在实验过程中的广告促销的带动。实验中，至少 3 家网站的数据清晰地显示了用户信息保护下的网络学习资源知识点内容自动提取系统对加强用户信息保护及知识点自动提取方面具有显著的效果。本文用真实数据集支持了实验结论。

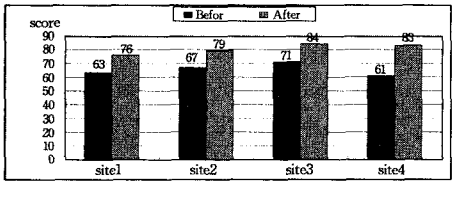

图 3　4 家学习网站用户评分比较图

结束语　本文重点研究用户信息保护下的知识点内容自动提取技术，通过计算用户信息保护状态、影响因素观察项、初始概率、转移概率、发射概率，构造了 HMM 模型来预测用户信息是否被侵犯，实现了保护状态失效时数据处理程序的自动退出。在用户信息得到保护的前提下，通过知识点内容将大量的课程信息用简洁、准确的形式呈现给用户。为了消除课程特有的冗余信息，本文提出了一种基于聚类算法的知识点自动提取模型，其通过信息融合技术合并提取出的知识点内容。

今后的研究主要围绕以下方面展开：

在协同式学习中，建立用户之间单向、双向以及多向等复杂信任关系的评价图。在保护用户隐私的条件下，通过获取用户社会网络联系信息来构建基于信任链的协同式知识点自动提取系统。

在原 OPTICS 算法中，有序队列总是按照可达距离升序排列，所以，算法总选择可达距离最小的点进行处理。这样会导致算法总是沿着稠密的区域进行扩张，只有处理完当前的稠密区域才会跳到下一个稠密区域。而跳入新的稠密区后，原先未处理的稀疏节点将无法得到处理，一直到所有稠密区域全部找出后才会处理稀疏节点。这种处理节点的方式会导致聚类结果的偏差，不能完全反映数据空间的真实结构。为了提高聚类的准确性，需要给出一种处理稀疏节点的方法：拟先将所有节点使用 OPTICS 算法进行聚类，找出基本的聚类簇；然后再对稀疏节点进行处理，计算其到每个簇的相似度，并将其加入相似度最大的簇进行聚类。将稀疏结点恰当地分到每个类簇中去，使聚类结果比原 OPTICS 算法能更加准确地反映学习资源文档集合空间的结构。

参 考 文 献

[1] Wu Z. Dynamic Trust Establishment with Privacy Protection for Web Services [A]// Proceedings of the IEEE International Conference on Web Services[C]. ACM Press, 2005: 811-812

[2] Liu Shuang, Liu Fang, Yu C, et al. An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases[A] // Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. ACM Press, 2004: 266-227

[3] Liu Shuang, Yu C, Meng Wei-yi. WordSense Disambiguation in Queries [A]// Proceedings of the 14th ACM International Conference on Information and Knowledge Management[C]. ACM-Press, 2005: 525-532

[4] Ankerst M, Breunig M M, Kriegel H-P, et al. OPTICS: ordering points to identify the clustering structure. International Conference on Management of Data [A]// Proceedings of the 1999 ACM SIGMOD international conference on Management of data [C]. ACM Press, 1999: 49-60

[5] Barzilay R, Elhadad N, McKeown K. Inferring Strategies for Sentence Ordering in Multidocument News Summarization[J]. Journal of Artifical Intelligence Research (JAIR), 2002, 17: 35-55

[6] Lin Chinyew, Hovy E. Automated Multi-document Summarization in NeATS[A]// Proceedings of the Human Language Technology Conference[C]. 2002