基于半监督图聚类的项目主题模型构建方法

石林宾 余正涛 严 馨 宋海霞 洪旭东
(昆明理工大学信息工程与自动化学院 昆明 650500)

摘 要 项目文档主题表征的好坏直接影响后续评审专家的推荐效果。为有效利用项目文档片段之间的关联关系进行项目主题分析,提出一种基于半监督图聚类的项目主题模型构建方法。该方法首先分析项目文档的结构特点,提取项目名称、项目关键字等能表征主题的结构信息,结合专家证据文档、专家主题关系网等能表征专家主题的外部资源,定义及提取项目文档片段之间的关联关系特征;然后,利用不同类型的关联关系计算项目文档片段之间的相关性,构建项目文档片段间的无向图模型;最后,利用已标记关联关系特征作为聚类的监督信息,采用半监督图聚类算法对项目文档片段进行聚类,从而实现项目主题的提取。项目主题提取对比实验结果验证了所提方法的有效性,项目文档结构化特征、专家证据文档以及专家主题关系网对项目主题模型的构建具有一定的指导作用。

关键词 主题模型,半监督图聚类,关联关系特征,评审专家推荐
中图法分类号 TP391.2 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.5.024

Project Topic Model Construction Based on Semi-supervised Graph Clustering

SHI Lin-bin YU Zheng-tao YAN Xin SONG Hai-xia HONG Xu-dong
(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract The quality of project topic model has a direct impact on recommended effect of the follow-up evaluation experts. In order to effectively exploit the association relationships among project document fragments to analyze project topics, we proposed a project topic model construction method based on semi-supervised graph clustering. We first analyzed structural characteristics of project documents to extract project name, project keywords and other structural information that responds project topics. Combined with expert evidence documents, expert topic relationship networks and other external resources which can indicate expert topics, we defined and extracted the association relationship features among project document fragments. Then, we used different association relationships to calculate correlation among project document fragments and built undirected graph model for project document fragments. Finally, using the marked association relationship features as supervised information for clustering, we applied semi-supervised graph clustering algorithm to cluster for project document fragments to realize the construction of the project topic model. The comparative experimental results of project topic extraction verify the effectiveness of the proposed method. Structural features of the project documents, expert evidence documents and expert topic relationship networks have certain guidance function for the construction of the project topic model.

Keywords Topic model, Semi-supervised graph clustering, Association relationship features, Evaluation experts recommendation

1 引言

随着各类项目的立项、申报等活动日益增多,主管机构人工遴选项目评审专家面临着严峻的挑战。项目评审专家的遴选需要主管机构对项目申请的领域以及评审专家的研究领域有比较深的了解,同时需要考虑评审专家的其他因素,工作量很大,因此亟需一种为申报项目自动推荐合适评审专家的解决方案,以克服人工遴选评审专家带来的弊端[1]。为项目推荐评审专家首先需要构建项目主题模型,提取项目的主题信息。主题模型是对项目文档的主题表征,相当于个性化推荐过程中的用户模型,其模型对主题表征的好坏直接影响后续项目评审专家的推荐效果,所以构建项目主题模型具有非常重要的意义。

项目主题模型的构建过程主要是对项目文档进行分析,挖掘能表征项目文档的主题信息。主题模型[2]的构建过程通常是对所提供的项目文档的挖掘过程,目前多数主题的获取

到稿日期:2014-01-20 返修日期:2014-04-01 本文受国家自然科学基金(61175068),国家中小企业创新基金(11C26215305905),云南省教育厅基金重大专项项目资助。

石林宾(1989—),男,硕士生,主要研究方向为信息检索、专家推荐;余正涛(1970—),男,博士,教授,博士生导师,CCF高级会员,主要研究方向为自然语言处理、信息检索,E-mail:ztyu@hotmail.com;严 馨(1969—),女,副教授,主要研究方向为自然语言处理、数据挖掘;宋海霞(1986—),女,硕士生,主要研究方向为自然语言处理、数据挖掘;洪旭东(1989—),男,博士生,主要研究方向为自然语言处理。

采用无监督学习方法[3-5]。其中，Blei 等人在 PLSA 的基础上，利用贝叶斯理论使其参数具备了概率分布，将其转化为随机变量，提出了基于隐藏的狄利克雷分析（Latent Dirichlet Allocation，LDA）的无监督主题模型[3]。Chong 等人基于高斯马尔科夫随机场对不同主题间存在的相关性进行估计，在 LDA 的基础上提出了马尔科夫主题模型，并将该模型应用到专业会议的主题发现上，取得了一定的效果[4]。孙艳等人通过在 LDA 模型中融入情感模型，提出一种无监督的主题情感混合模型，并将该模型用于情感分类任务[5]。由于无监督学习方法在主题聚类方面存在聚类结果分散的问题，最终导致提取的主题效果不理想。于是，很多学者考虑在模型中加入一定的监督信息，期望达到更好的主题聚类效果，其中，相关学者基于有监督学习的思想对主题模型的构建开展了相关研究[6-8]。例如，Blei 等人将文档类别信息作为监督信息，提出有监督 LDA 模型并将其用于文本分类任务，该模型在 LDA 模型的基础上加入文档类别标记作为监督信息，在文档主题模型的获取上取得了很好的效果[6]。李文波等人在传统 LDA 模型基础上融入文本类别信息，将参数按照类别细化，提出了 Labeled-LDA 模型并用于文本分类任务[7]。江雨燕等人通过研究文档标记与 LDA 模型中主题的映射关系，将文档标记映射为局部主题和共享主题的组合，提出一种新的 LabeledLDA 模型，该模型可以有效地利用文档的标记信息提升文档聚类效果[8]。另外，相关学者基于半监督学习的思想对主题模型的构建开展了相关研究[9,10]，例如 Ville 等人以聊天者之间的链接关系作为主题聚类的指导资源，提高了聊天数据主题挖掘的准确性[9]。TanXu 等人利用微博中的关键词与 Wikipedia 之间的关联关系作为指导资源，对微博中的信息进行主题扩展，完成了微博的主题聚类排序[10]。

这些前期工作表明，为了提高主题模型的准确性，需要深度挖掘可用的监督信息和指导关系参与主题分析。项目文档结构化特征明显，其中存在着项目名称、项目关键词、所属领域等能表征主题的结构化信息，这些结构化信息能够有效地指导项目文档的主题聚类，例如包含同一个项目关键词的项目文档片段往往会聚成一类。针对面向项目主题的专家推荐任务而言，项目文档片段信息与专家主页、专家微博页、专家百度百科页等专家证据文档之间存在着主题关联关系；另外，可以利用专家之间的关系构建专家主题关系网，项目文档片段主题信息与专家主题关系网在一定程度上也表现出很多关联，这些专家关联关系对项目主题提取具有一定的约束作用。因此，本文借助于半监督图聚类思想，利用专家证据文档及专家关系网等外部资源以及项目文档本身的结构化特征，定义和挖掘项目文档片段之间的关联关系特征，利用这些关联关系作为约束，研究基于半监督图聚类的项目主模型构建方法。

本文第 2 节分析并提取项目文档结构化特征以及文档片段关联关系特征；第 3 节提出基于半监督图聚类的项目主题模型构建方法；第 4 节给出实验结果及分析；最后给出结论，并对下一步工作进行展望。

2 特征分析

2.1 项目文档结构化特征

项目文档包含很多能表征项目主题的结构化信息，例如项目名称、所属领域、项目关键词等信息。这些结构化信息对项目主题模型的构建具有指导和约束作用，并且对于分析项目文档片段之间的关联关系具有很大的作用。经过分析，选取项目名称、项目关键词等 6 种能表征项目主题的结构化特征，如表 1 所列。

表 1 项目文档结构化特征

序号	结构化特征	特征表示
1	项目名称	Name(pn)
2	所属领域	Domain(pd)
3	所属学科	Subject(ps)
4	项目关键词	Key(pk)
5	项目成员研究方向	Orientation(po)
6	项目成员在本课题中承担的任务	Task(pt)

本文采用实验室开发的 CRF-tf 系列工具对上述定义的 6 种结构化特征进行标注，并将其处理成 CRFs 单行模型训练的格式，然后采用 CRFs 工具包对上述已标注的项目结构化特征训练识别模型，通过识别模型对测试语料中项目文档的结构化信息进行识别。

2.2 项目文档片段关联关系特征

对项目文档进行划分。项目文档是项目文档片段的集合，项目文档片段之间可能存在着一些显示或者隐含的关联关系特征，如项目文档片段包含同一个项目关键词，项目文档片段包含项目成员在本课题中承担的同一个任务，项目文段片段与项目名称以及所属学科等存在的隐含主题关系等。同时针对面向项目主题的专家推荐任务，专家的主题对于项目主题的获取具有一定的约束作用，项目与专家之间在一定程度上存在着主题关联关系，例如项目文档片段与专家证据文档、专家主题关系网存在着主题关联关系等。所以本文根据项目文档中能表征主题的结构化信息定义关联关系特征 f_1 —f_6，根据项目文档片段与专家证据文档之间的主题关联关系定义关联关系特征 f_7，根据项目文档片段与专家主题关系网之间的主题关联关系定义关联关系特征 f_8。项目文档片段关联关系特征如表 2 所列。

表 2 项目文档片段关联关系特征

序号	特征名称	特征值
1	与项目名称关联	[0,1]
2	与所属学科关联	[0,1]
3	与所属领域关联	[0,1]
4	包含同一项目关键字	{0,1}
5	与项目成员研究方向关联	[0,1]
6	包含项目成员承担的同一任务	{0,1}
7	与专家证据文档关联	[0,1]
8	与专家主题关系网关联	[0,1]

对于布尔类型关联特征通过正则匹配方式确定其特征值；对于非布尔类型关联特征，利用《知网》计算相似度[15]确定其特征值。

3 基于半监督图聚类的项目主题模型构建

3.1 半监督图聚类主题模型构建

将项目文档片段看作是待聚类的节点集合 $V=\{v_1, v_2, \cdots, v_n\}$，$v_i$ 之间的相关性通过以上定义项目文档片段之间的关联关系进行相似度计算获得。建模时为不同的关联关系和资源定义不同的权重，并通过期望最大化算法确定权重，并将

权重进行归一化加权求和确定节点间的相似性，节点间的相似度计算公式为 $A_{ij}=\sum_{m=1}^{8} w_m f_m$，最终得到相似度矩阵 A。主题模型可以表示为无向图 $G=(V,E,A)$，其中 V 为项目文档片段集合，A 表示相似度矩阵，E 为节点间边集合，其权重通过 A 体现。至此，主题模型的构建问题转换为图聚类过程。

为了能够让聚类的主题不发散，需要定义约束规则。Wagstaff 等人最早在文献[11]中引入两种类型的成对点约束即 must-link 和 connot-link 来作为聚类约束，must-link 限制两个项目文档片段必须在同一聚类中，connot-link 限制两个项目文档片段必须在不同的聚类中。通过启发式规则定义项目文档片段之间的约束规则，如当项目文档片段包含同一项目关键词时，认为对应的两个节点间存在 must-link 约束；而当项目文档片段包含不同项目关键词时，对应的两个节点间存在 connot-link 约束。本文在分析项目文档片段之间的关联关系的基础上，定义“must-link”和“connot-link”约束规则，如表 3 所列。

表 3　约束规则定义

序号	“must-link”	“connot-link”
1	包含相同的项目关键词	包含不同的项目关键词
2	包含相同的任务词	包含不同的任务词
3	与同一专家主题关系网存在主题关联	与不同专家主题关系存在主题关联
4	与同一专家证据文档存在主题关联	与不同专家证据文档存在主题关联，且主题不相似

以上定义了 4 种类型的约束规则，通过分析每两个项目文档片段之间存在的约束，构建约束关联矩阵 W。当两个项目文档片段之间满足“must-link”约束时，$w_{ij}=1$，当两个项目文档片段之间满足“connot-link”约束时，$w_{ij}=-1$，否则 $w_{ij}=0$。

将项目主题模型的构建过程转换为项目文档片段的图聚类过程，并且该过程可以看作是对目标函数的求解过程。基于图的半监督聚类的主题模型的目标函数表示为如下求解最优化的过程[12]。

$$D(\{\pi_c\}_{c=1}^{k})=\sum_{c=1}^{k}\frac{links(V_c,V\backslash V_c)}{\deg(V_c)}-\sum_{\substack{x_i,x_j\in M\\ l_i=l_j}}\frac{w_{ij}}{\deg(V_{l_i})}+\sum_{\substack{x_i,x_j\in C\\ l_i=l_j}}\frac{w_{ij}}{\deg(V_{l_i})} \quad (1)$$

其中，$links(V_c,V\backslash V_c)=\sum_{i\in V_c,j\in V\backslash V_c}A_{ij}$ 表示第 c 个聚类与其它节点的相似度，π_c 表示第 c 个聚类，$\deg(V_c)=\sum_{\substack{i\in V_c,j\in V_c\\ i\neq j}}A_{ij}$ 表示聚类中节点间的相似度。M 和 C 表示节点的先验约束条件，M 表示“must-link”约束，C 表示“connot-link”约束，w_{ij} 表示聚类违反约束的惩罚值，表 3 已经给出了约束规则的定义。这些约束条件作为聚类过程中的监督信息，为聚类过程提供指导。至此，融合关联关系的半监督图聚类主题模型的构建已完成。

3.2　主题聚类过程

项目文档片段的图聚类过程可以看作是对目标函数的求解过程，也即是在聚类过程中使目标函数 D 最小化。同时，针对主题模型的构建而言，聚类的数目是未知的，其不仅与提供的项目文档相关，而且与模型自身的粒度有关，所以必须事先确定聚类数 k。聚类数 k 通过统计模型的贝叶斯信息准则(Bayesian Information Criterion，BIC)来估计。

BIC 起源于贝叶斯学习理论[13]，是确定最优聚类数目的常用方法。文献[14]假设聚类数为 K 时得到一系列聚类中心点，其集合为 $M_k=\{c_1,c_2,\cdots,c_k\}$，通过相应的贝叶斯信息值公式 $BIC_k=\ln(P(DB|K))-p_k\times\ln|DB|$ 计算贝叶斯信息测度值，并以贝叶斯信息测度值作为寻找聚类过程中的最优聚类数目的判别标准。其中，$\ln(P(DB|K))$ 是数据集 DB 和集合 M_k 中的中心点相匹配的概率；$p_k=K+K\times d$ 是聚类数为 K 时的参数个数；$p_k\times\ln|DB|$ 是惩罚项，对于同一数据集，p_k 值越小相应的惩罚也越小。BIC 方法通过平衡上述两项，选择最优类数目，BIC_k 值越大，则聚类数为 K 时得到的聚类结果越能准确反映真实数据的分布情况。

聚类的主要过程是：首次聚类时置 $k=1$，在使得目标函数 D 最小时完成项目文档片段的聚类，在每次聚类完成后计算其 BIC；然后将某个聚类中心分裂为两个后重新聚类，如果 BIC 的值显著增大则接受新的聚类，反之拒绝；最后，通过不断迭代确定聚类数 k 的值，同时完成对项目文档片段的聚类。具体算法描述如表 4 所列。

表 4　基于半监督图聚类的项目主题聚类算法

Input A: input affinity matrix, obj: clustering objective, k: number of clusters, W: constraint penalty matrix
t_{max}: optional maximum number of iterations.
Output $\{\pi_c\}_{c=1}^{k}$ final portioning of the points
Process:
step 1　set k=1, t=0
//初始化聚类中心
step 2　Get initial Clusters $\{\pi_c^{(t)}\}_{c=1}^{k}$, Initialize the k clusters center M_k^t.
//对每个对象聚类
step 3　For each clustering objective obj_i, computer $D(\{\pi_c\}_{c=1}^{k})$ according to Eq. (1).
step 4　Minimum $D(\{\pi_c\}_{c=1}^{k})$, for each clustering objective obj_i to clusters c, updated the k clusters center M_k^t.
//计算贝叶斯信息值来确定是否接受当前的聚类结果
step 5　For every cluster $\{\pi_c^{(t)}\}_{c=1}^{k}$ and M_k^t, computer $BIC_k^{t+1}=\ln(P(DB|K))-p_k\times\ln|DB|$ if $BIC_k^{t+1}>BIC_k^t$ updated clusters $\{\pi_c^{(t)}\}_{c=1}^{k}$, otherwise, stop and output final clusters $\{\pi_c\}_{c=1}^{k}$
//如果算法收敛输出聚类结果，否则跳转到第 2 步
step 6　If not converged or $t_{max}>t$, set t=t+1, k=k+1 and goto step 2; otherwise, stop and output final clusters $\{\pi_c\}_{c=1}^{k}$.

通过半监督图聚类主题模型聚类后，主题数即为聚类数 k，每一个项目文档片段聚类簇表示一个主题，这些主题需要一些主题词来表示。本文通过计算词频，取词频最高的前 N 个词作为当前主题的主题词，实现了项目主题模型的构建。

4　实验以及结果分析

4.1　实验数据

由于没有开放的权威语料资源，本文项目申报的文档数据来自昆明理工大学科技项目管理系统。从系统文档库中按项目文档的领域整理出了内容较完整的 4 个类别的申报项目文档共 400 篇，人工收集了 4 个领域共 200 个专家的证据文档，其中包括专家的主页、专家博客页、百度百科页和维基百科页。此外实验中还用到已构建的 20 个主题的专家关系网。

4.2 评价指标

为了综合衡量所提方法的有效性，在统计正确率(Precision)和召回率(Recall)两个指标的基础上，采用F值作为衡量聚类效果最终的评测指标。准确率、召回率以及F值的公式表示如下：

$$正确率(P)=\frac{T_p}{T_p+F_p}\times 100\% \quad (2)$$

$$召回率(R)=\frac{T_p}{T_p+F_n}\times 100\% \quad (3)$$

$$F值=\frac{2\times P\times R}{P+R}\times 100\% \quad (4)$$

其中，T_p 表示应该放在一类的项目文档被放在了一类的文档数目，F_p 表示不应该放在一类的项目文档被错误地放在了一类的文档数目，F_n 表示不应该分开的项目文档被错误地分开的文档数目。

4.3 实验及结果分析

为了验证关联约束数目对聚类产生的影响，通过实验找到最优的关联约束数目。本实验选取面向冶金、生物、化工、信息技术4个领域的项目申报文档，每类随机选100篇，共400篇。把4个领域的项目文档分别分成10份，每次选取2份项目文档集合作为测试集，将其余8份作为训练集。实验中页面关联特征约束数目(M和C的总对数)取自0－300之间，由于关联约束数目是连续的，因此使约束数目每次新增50来反映聚类的整体情况。图1给出了不同的关联约束数目对4个领域测试集的F值评测指标变化情况。从图1可以看出，基于半监督图聚类的项目主题模型构建方法得到的F值随着关联约束数量的增加而增大，在关联约束数目为200时，F值达到最大，而当约束数目大于200时聚类性能呈现下降趋势，所以当关联约束数目为200时得到的聚类数是最优的。

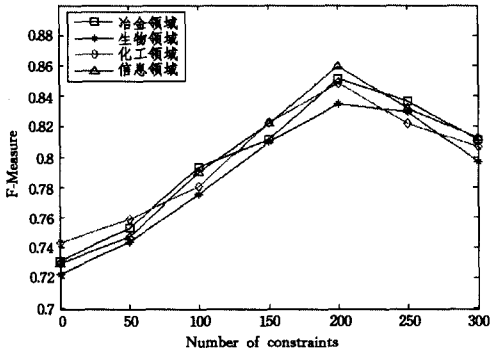

图1 关联约束数目对测试集F值的影响

为了分别验证项目的结构化特征、专家证据文档以及专家主题关系网对项目主题模型构建的有效性，本文主要比较了无监督LDA主题模型与不同关联特征的半监督图聚类主题模型的主题聚类效果。比较的算法包括：

Baseline，使用无监督的主题聚类模型，目前使用得最多的无监督聚类模型是LDA主题模型，所以本文使用LDA作为基准算法。

SSGC-T1，使用的是半监督图聚类算法，加入项目的结构化特征关联约束，也即使用的关联特征约束为 f_1-f_6。

SSGC-T2，使用的是半监督图聚类算法，该算法在SSGC-T1的基础上加入与专家证据文档关联约束，也即使用的关联特征约束为 f_1-f_7。

SSGC-T3，使用的是半监督图聚类算法，该算法在SSGC-T2的基础上加入与专家主题关系网关联约束，也即使用的关联特征约束为 f_1-f_8。

对于每个领域数据集，30%作为测试数据，70%作为训练数据。在本实验中标注的关联约束数目为200。实验采用十折交叉验证的方法，表5比较了不同关联特征下的主题聚类效果。从表5可以看出，加入了关联约束后，获取项目主题的效果得到了很大的改善，这也说明了项目文档中能表征主题的结构化特征对项目主题模型的构建具有约束和指导作用，专家证据文档及专家主题关系网等能表征专家主题的外部资源对项目主题模型的构建具有重要的支撑作用。

表5 不同关联特征下主题聚类效果的比较

		Baseline	SSGC-T1	SSGC-T2	SSGC-T3
冶金领域	P(%)	78.52	83.10	83.94	84.27
	R(%)	72.90	76.65	77.13	77.80
	F(%)	73.60	79.74	80.44	80.90
生物领域	P(%)	76.30	82.81	83.21	83.90
	R(%)	69.53	74.70	76.02	76.86
	F(%)	72.75	78.55	79.45	80.23
化工领域	P(%)	79.25	83.85	84.63	85.01
	R(%)	73.37	76.82	77.68	78.05
	F(%)	74.90	80.18	81.48	81.38
信息领域	P(%)	76.70	81.52	82.04	82.94
	R(%)	70.80	75.90	76.51	77.35
	F(%)	73.63	78.61	79.18	80.04
平均值	P(%)	77.69	82.82	83.46	84.03
	R(%)	71.65	76.02	76.84	77.52
	F(%)	73.72	79.27	80.14	80.64

结束语 为有效利用项目文档片段之间及与专家关系网、专家证据文档等外部资源的关联关系进行项目主题分析，本文提出一种基于半监督图聚类的项目主题模型构建方法。通过利用项目本身的结构信息以及专家证据文档、专家关系网等外部资源，定义提取项目文档片段之间的关联关系特征，采用半监督图聚类算法对项目主题进行抽取。实验验证了所提方法的有效性。进一步工作将探索基于主题模型的专家推荐方法。

参 考 文 献

[1] 许云红. 基于网络方法的专家知识推荐[D]. 安徽：中国科学技术大学，2010

[2] 徐戈，王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报，2011，34(8)：1423-1436

[3] Blei D M，Lafferty J D. Dynamic topic models[C]// Proceedings of the 23rd International Conference on Machine Learning. New York，USA：ACM，2006：113-120

[4] Chong Wang，Bo T，Christopher M，et al. Markov Topic Models [C]// Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Clearwater Beach，USA，2009：583-590

[5] 孙艳，周学广，付伟. 基于主题情感混合模型的无监督文本情感分析[J]. 北京大学学报：自然科学版，2013，49(1)：102-108

[6] Blei D，McAuliffe J. Supervised topic models[C]// Advances in Neural Information Processing Systems (NIPS). Vancouver，

Canada,2008

[7] Li Wen-bo,Sun Le,Zhang Da-kun. Text classification based on labeled-LDA model[J]. Chinese Journal of Computers,2008,31(4):620-627

[8] 江雨燕,李平,王清. 基于共享背景主题的 LabeledLDA 模型[J]. 电子学报,2013,41(9):1794-1799

[9] Ville H T,Henry T. Combining Topic Models and Social Networks for Chat Data Mining[C]// IEEE/WIC/ACM International Conference on Web Intelligence. Los Alamitos, USA: IEEE Computer Society Press,2004:206-213

[10] Tan Xu, Douglas W O. Wikipedia-based Topic Clustering for Microblogs[J]. American Society for Information Science and Technology,2011,48(1):1-10

[11] Wagstaff K,Cardie C. Clustering with instance-level constraints [C]// Proceedings of the 17rd international conference on Machine learning. Morgan Kaufmann,2000:1103-1110

[12] Brian K, Sugato B, Inderjit S D, et al. Semi-supervised graph clustering: a kernel approach[J]. Machine Learning, 2009,74(1):1-22

[13] Kass R,Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarzcriterion[J]. Journal of the American Statistical Association,1995(10):928-934

[14] 郑苗苗,吉根林. 一种基于密度的分布式聚类算法[J]. 南京大学学报,2008,44(5):536-543

[15] 刘群,李素建. 基于《知网》的词汇语义相似度计算[C]//第三届汉语词汇语义学研讨会. 台北,2002

(上接第 87 页)

结束语 本文中,我们将以前独立的两种衡量方式综合考虑,希望找到那些高质量即既频繁又高效用的项集。另外,我们将问题转化为 top-k 的频繁和高效用项集的挖掘。由于效用值和质量值都既不具有单调性也不具有反单调性,因此提出了利用相对效用值和质量值上界来剪枝的 FHIMA 算法。同时,FHIMA 利用了 Prefixspan 算法的思想,避免了非频繁候选项集的产生。实验结果证明,FHIMA 算法通过利用上界值剪枝能大大提高算法的效率,同时紧的上界值在效率上要优于松的上界值。

参 考 文 献

[1] Agrawl R, Srikant R. Fast algorithms for mining association rules[C]// VLDB. 1994:487-499

[2] Ahmed C F,Tanbeer S K,Jeong B-S,et al. Efficient tree structures for high utility pattern mining in incremental databases[C]// TKDE. 2009:1708-1721

[3] Ahmed C F,Tanbeer S K,Jeong B-S,et al. Efficient tree structures for high utility pattern mining in incremental databases [J]. TKDE,2009,21(12):1708-1721

[4] Chan R, Yang Q, Shen Y. Ming high-utility itemsets[C]// ICDM. 2003:19-26

[5] Cheung Y L,Fu A W. Mining frequent itemsets without support threshold:with and without item constraints[C]// TKDE. 2004: 1052-1069

[6] Chuang K,Huang J,Chen M. Mining top-k frequent patterns in the presence of the memory constraint[J]. VLDB Journal,2008, 17:1321-1488

[7] Fu A W,Kwong R W,Tang J. Mining n-most interesting itemsets[C]// ISMIS. 2000:59-67

[8] Gade K, Wang J, Karypis G. Efficient closed pattern mining in the presence of tough block constraints[C]// KDD. ACM,2004: 138-147

[9] Han J,Pei J,Yin Y. Mining frequent patterns without candidate generation[C]// SIGMOD. 2000:1-12

[10] Han J, Wang J, Liu Y, et al. Mining top-k frequent closed patterns without minimum support[C]// ICDM. 2002:211-218

[11] Li H F,Huang H Y,Chen Y C,et al. Fast and memory efficient mining of high utility itemsets in data streams[C]// ICDM. 2008:881-886

[12] Liu M,Qu J. Mining high utility itemsets without candidate generation[C]// CIKM. ACM,2012:55-64

[13] Liu Y,Liao W K,Choudhary A. A fast high utility itemsets mining algorithm[C]// Workshop on WUDM. 2005:90-99

[14] Lucchese C,Orlando S,Palmerini P,et al. KDCI:A multi-strategy algorithm for mining frequent itemsets[C]// ICDM Workshop FIMI. 2003:372-390

[15] Pei J,Han J,Mortazaviasl B,et al. Prefixspan:Mining sequential patterns efficiently by prefix-projected pattern growth[C]// ICDE. 2001:215-224

[16] Savasere A,Omiecinski E R,Navathe S B. An efficient algorithm for mining association rules in large databases[C]// VLDB. 1995:432-443

[17] Seno M,Karypis G. Lpminer:An algorithm for finding frequent itemsets using length-decreasing support constraint[C]// ICDM. 2011:505-512

[18] Tseng V S,Chu C J,Liang T. Efficient mining of temporal high-utility itemsets from data streams[C]// KDDWorkshop on UBDM. 2006:18

[19] Tseng V S,Wu C W,Shie B E,et al. Up-growth:an efficient algorithm for high utility itemset mining[C]// KDD. 2010:253-262

[20] Vo B,Nguyen H,Ho T B,et al. Parallel method for mining high utility itemsets from vertically partitioned distributed databases [C]// KES 2009. 2009:251-260

[21] Wu C,Shie B,Tseng V S,et al. Mining top-k high utility itemsets[C]// KDD. 2012:78-86

[22] Yao H, Hamilton H J, Buts C J. A foundational approach to mining itemset utilities from databases[C]// SDM. 2004:251-260

[23] Zaki M. Scalable algorithms for association mining [J]. Knowledge and Data Engineering,2000,12(2):372-390

[24] Frequent itemset mining dataset repository,2012[OL]. http://fimi.ua.ac.be