利用单词超团的二分图文本聚类算法

朱 君[1,2]，曲 超[1]，汤 庸[2]

(1. 东莞理工学院计算机科学与技术系 广东 东莞 523000; 2. 中山大学计算机科学系 广州 510275)

【摘要】鉴于目前传统文本聚类方法中利用文档间的相似度进行聚类存在的问题，在传统的文本挖掘基础上提出了一种新的文本聚类算法—— 利用单词超团的二分图文本聚类算法。该算法用文档中单词的关联模式来评估文档间的相似度及主题类别预测，并利用图划分策略来大大降低文档相似度比较算法的复杂度，同时将超团作为特征结构的扩展，可以在一定范围内减少语言信息的丢失，提高聚类效果。经实验证明该算法具有较高的有效性。

关 键 词 二分图划分; 文本聚类; 单词超团

中图分类号 TP309. 4 文献标识码 A

Clustering Algorithm of Bipartite Graph Partition Based on Word Hyperclique

ZHU Jun [1, 2], QU Chao[1] , and TANG Yong [2]

(1. Department of Computer Science, DongGuan University of Technology Dongguan Guangdong 523000;

2. Department of Computer Science, Sun Yat-sen University Guangzhou 510275)

Abstract This paper proposes a new algorithm for document-word co-clustering. After mining semantics with word hyperclique patterns, the document dataset with a bipartite graph is described. Then, the efficient graph partitioning algorithm is employed to partition this graph, so that the high computational overhead of traditional clustering algorithms over huge document datasests can be avoided. During clustering, word hyperclique patterns that are full of document semantics are preserved. In this way, our algorithm partially circumvents the problem of loosing document semantics, which happens a lot in traditional clustering algorithms based on document pairwise similarity alone. Finally, the extensive experimental results demonstrated the effectiveness of this algorithm in document clustering accuracy and cluster topic detection.

Key words bipartite partition; documents clustering; word hyperclique

文本挖掘是信息领域当前的一个研究热点[1]。作为一种无监督的机器学习方法，聚类技术已成为对文本信息进行有效组织、摘要和导航的重要手段。传统聚类方法大多利用文档间相似度进行聚类，然后对各个类产生描述。但传统聚类方法的主要问题有：(1) 对于文档向量两两之间相似度的比较随着文档数量的增加其复杂度呈指数函数增加；(2) 对于文档的聚类，只以其中的线性部分作为特征，不能正确表述实际的语义。(3) 在计算文档之间相似度时，不同的特征结构对聚类的结果会产生不同的影响[2-3]。基于上述问题,本文提出一种新的文本聚类算法，即利用单词超团的二分图文本聚类算法。该算法用文档中单词的关联模式来评估文档间的相似度及主题类别预测，并利用图划分策略降低文档相似度比较算法的复杂度。另外，将超团作为特征结构的扩展可以在一定范围内减少语言信息的丢失，提高聚类效果。

1 文本聚类评测标准

文本聚类是无教师的机器学习，它没有预先定义好的主题类别，目标是将文档集合分成若干个簇，要求同一簇内文档内容的相似度尽可能大，而不同簇间的相似度尽可能小。对于聚类结果的评测，除了时间和空间等运行效率外，更重要的是对信息检索系统的检索性能进行评测。现在通用的评测标准主要包括以下6类[4-5]：

收稿日期：2008 – 04 – 11; 修回日期：2008 – 04 – 20
基金项目：国家自然科学基金(60673135; 60373081; 60736020)
作者简介：朱 君(1976 –)，女，博士生，讲师，主要从事CSCW、群体感知和信息检索方面的研究.

(1) 查准率：查准率是指检出的相关文献量与检出文献总量的比率，是衡量信息检索系统检出文献准确度的尺度，定义为 $\mathrm{Precision}=\frac{N_{\mathrm{correct}}}{N_{\mathrm{correct}}+N_{\mathrm{Incorrect}}}$。

(2) 召回度：召回度反映检索系统对某个查询返回结果中正确结果占全部正确结果的比例，定义为 $\mathrm{Recall}=\frac{N_{\mathrm{correct}}}{N_{\mathrm{q}}}$。

(3) 正规化互信息：正规化互信息是对查准率和召回度统计信息的一种均衡性度量，用来避免真实聚类和算法结果因共享信息所导致的噪音，定义为 $\mathrm{NMI}=\frac{I(T,C)}{\sqrt{H(T)H(C)}}$，其中$H(X)$表示$X$中的平均信息量；$I(X,Y)$表示$X$与$Y$的平均互信息量，定义为 $I(X,Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\log_2\frac{p(x_iy_j)}{p(x_i)p(y_j)}$。由此可以得到 $\mathrm{NMI}=\frac{H(T)+H(C)-H(T,C)}{\sqrt{H(T)H(C)}}$。

(4) 条件熵：条件熵表示在已知C的条件下在T中保存的有效信息，定义为 $\mathrm{CE}=H(T\mid C)=H(T,C)-(C)$。通常，CE的取值范围依赖于真实类别的个数$T$，为了达到测量尺度的正规化，通常采用正规化的条件熵 $\mathrm{NCE}=\frac{H(T\mid C)}{H(T)}=\frac{H(T,C)-H(C)}{H(T)}$。

(5) 错误率：错误率表示划分块中与占比率最大的实际类别样本不同的样本所占的比例，即划分到错误聚类中的样本比率，可以看作是$H(T|C)$的一个简化，定义为 $\mathrm{ERR}=\sum_{i=1}^{n}\frac{\left\|C_i-\max 1,2,\cdots,m(C_i\cap T_j)\right\|}{\left\|C_i\right\|}$。

(6) F-measure：F-measure包含了查准率和召回度的内部信息关联[8]。将每个簇看作是一个查询结果，同时将每个类别看作希望得到的文档集，可以按照公式 $P_{ij}=C_{ij}/C_{ij}$ 及 $R_{ij}=C_{ij}/C_{+j}$ 计算每个给定类别的查准率和召回度。公式中 C_{+j} 表示第j个真实类别对应第i个划分块的查准率；R_{ij} 表示第j个真实类别对应第i个划分块的召回度；C_{ij} 表示第j个真实类别划分到第i个划分块中的文档个数；C_{i+} 和 C_{+j} 分别表示第i个划分块中文档的个数和第j个真实类别中文档的个数。由于同一篇文档在聚类过程中可能被划分在多个划分块中，因此 C_{+j} 通常会大于第j个真实类别中文档的数目。聚类划分和真实类别之间的F-measure可以定义为 $F_{ij}=(2P_{ij}R_{ij})/(P_{ij}+R_{ij})$。

2 单词超团特征结构的设计

2.1 超团的定义

“超团”(hyperclique)是在频繁项集(frequent item set)的基础上提出的一个较新的概念，是一种特殊的频繁项集[5-6]。

定义 1 超团信任度：项目集 H_c 的超团信任(h-confidence)度 hconf(P) 反映项目集P中全部项目的整体关联程度，定义为：

$$\begin{aligned}\mathrm{hconf}(P)=\min\{&\mathrm{conf}\{i_1\to i_2,\cdots,i_m\},\\&\mathrm{conf}\{i_2\to i_1,i_3,\cdots,i_m\},\cdots,\\&\mathrm{conf}\{i_m\to i_1,\cdots i_{m-1}\}\}\end{aligned}$$

式中 conf为普通意义上的信任度。

定义 2 超团：对于一个项目集P，如果称之为一个超团，当且仅当 $\mathrm{support}(P)\geqslant\theta$，并且 $\mathrm{hconf}(P)\geqslant H_c$ 其中 θ 和 H_c 分别是用户自定义的支持度阈值和超团信任度阈值。作为一个项目集P，如果满足超团的定义，则该项目集称为一个超团，其中每个项目的支持度不低于θ，且任意两个项目之间的信任度均不低于 H_c；对于P的子集$\boldsymbol{D}$中的项目，同样满足超团的定义，因此也可以称为超团。但实际挖掘中，真正感兴趣的是项目集P本身而不是它的子集。一个超团子集并不能反映整体的关联程度，因此需要给出一个最大超团的定义。

定义 3 最大超团(maximal hyperclique pattern，MHP)[7]：对于一个超团HP，如果任意一个包含它的集合都不能够满足超团的定义则称HP是一个最大超团，即 $\forall P'\supset P,P'\notin\mathrm{MHP}$，$P\in\mathrm{HP}$ $P\in\mathrm{MPH}\Leftrightarrow$，并且 $\forall P'\supset P,P'\notin\mathrm{MHP}$。

最大超团依赖于给定的支持度和超团信任度，对于不同的支持度和超团信任度，一个项目集所包含的超团的数目和各个超团的大小是不相同的。

2.2 单词超团特征结构

给定单词集合 $I=\{i_1,i_2,\cdots,i_k\}$，单词集合的子集集合文档向量集合 $\boldsymbol{D}=\{d_1,d_2,\cdots,d_m\}$，其中，$d_i\subset I$。给定最小支持度$s$和最小超团信任度hc对$\boldsymbol{D}$进行最大超团集合挖掘，得到超团集合 $\mathrm{HP}=\{p_1,p_2,\cdots,p_n\}$，其中，$p_i\subset I$。

定义 4 对于$\boldsymbol{D}$中任意一个文档向量 $\boldsymbol{d}_k$，若 $p_i\subset\boldsymbol{d}_k$，则称超团 P_i 是文档 $\boldsymbol{d}_k$ 的一个扩展特征分量。

定义 5 $\boldsymbol{d}_k$ 所有扩展特征分量构成的集合称为 $\boldsymbol{d}_k$ 的特征分量集，记为 $P\boldsymbol{d}_k$。

将单词集合I和超团集合HP的并集作为新的特

征分量集合IH， $IH=\{ih_1,ih_2,\cdots,ih_n\}$，其中$n=k+m$，且 $\forall ih_\partial(\partial\leqslant k)\Rightarrow(ih_\partial=i_\partial)$，$\forall ih\partial(k<\partial\leqslant k+m)\Rightarrow(ih_\partial=P_{\partial-K})$。

对文档向量集合可重新构造如下：

$$MD=\{md_1,md_2,\cdots,md_m\}$$

式中　$md_i=d_i\cup Pd_i$。则集合MD即为新构造的以超团作为扩展特征结构的文档集合，对原始文档集合进行的聚类挖掘即可转化为在MD上进行相关的聚类操作。

2.3　向量分量权值的设定

本文选用Tf-idf作为文档向量的权值算法，然后对各个向量进行正规化。对超团分量的Tf-*i*df权值做如下定义。

定义 6　超团分量的权值 $\overline{\omega}_i$ 定义为 $\overline{\mathrm{tf}}*i\mathrm{df}$，其中 $\overline{tf}$ 定义为超团中包含的所有单词的tf之和的平均值；*i*df=log(N/n_i)，其中N为文档向量的数量，n_i为出现第i个超团的文档数量。

经过如上定义，可以对包含权值的文档向量的结构作出新的定义。

定义 7　带有权值的以超团作为扩展特征向量的文档向量结构为：

$$\omega md_i=\{i_{a_1},\omega_1,i_{a_2},\omega_2,\cdots,i_{a_k},\omega_k,p_{b_1},\overline{\omega}_1,p_{b_2},\overline{\omega}_2,\cdots,p_{b_1},\overline{\omega}_1,\}$$

式中　$i_{a_x}\in md_i$；$\omega_i=\mathrm{tf}_i*i\mathrm{df}_i$；$P_{b_i}\in HP$ 且 $P_{b_i}\in md_i$。

3　二分图划分聚类算法定义

图划分算法能够完成数据集中相似元素的分类工作，可将图划分法引入作为聚类的方法。通常的用于文档聚类的图划分方法分为两类，第一类是将每个文档作为图中的节点，通过计算节点之间的两两相似度来进行划分；第二类是将文档和单词都作为图中的节点，构成二分图(bipartite graph)，对二分图进行划分[8]。给定一个具有n个文档的集合S，相似度图 C_s 是通过将每个文档作为一个结点，任意两个结点之间存在一条边，边的权值表示该边连接的两个结点之间的相似度。对于这样的相似度图来说，可以定义很多种不同内部目标函数、外部目标函数，或联合函数来衡量整体的聚类划分效果。

本文定义“最小剪切的最大化内部相似度”(MinMaxCut)作为衡量聚类划分的依据。MinMaxCut函数是对内部目标函数和外部目标函数的综合体现，定义为：

$$\min\ \mathrm{imize}\sum_{r=1}^{k}\frac{\mathrm{cut}(S_s,S-S)}{\sum_{d_id_j\in S_r}\mathrm{sim}(d_i,d_j)}\qquad(1)$$

式中　$\mathrm{cut}(S_r,S-S_r)$ 表示对 S_r 子图和它的补图 $S-S_r$ 之间边的剪切，$\sum_{d_id_j\in S_r}\mathrm{sim}(d_i,d_j)$ 表示子图中所有顶点之间的相似度之和。对于结点集合A和B之间边的剪切可以定义为A和B之间任意节点之间的连接数之和。该函数的实际目标是要保证在划分过程中使被剪切的边的数量最小，同时保证被划分成的各个子图中的所有节点的相似度之和最大。这样可以保证划分的平衡性，避免划分由于单纯考虑内部目标函数导致出现较小的划分块，也可以避免由于只考虑外部目标函数而导致的划分不均衡。如果使用余弦值来表示文档之间的相似度，则式(1)可表示为：

$$\sum_{r=1}^{k}\frac{\sum_{d_i\in S_r,d_j\in S-S_r}\cos(d_i,d_j)}{\sum_{d_i,d_j\in Sr}\cos(d_i,d_j)}$$

二分图划分是将单词和文档都作为节点来构成图 $C_b=(V,E)$，图中的节点集合V由V_d和V_t两部分组成，其中，V_d由文档集合组成，每个文档作为其中一个节点；V_t由单词集合组成，每个单词作为其中一个节点。对于V_d中的任意一个节点i和V_t中的节点j，如果文档i包含单词j，则i和j之间存在一条边；V_d中任意两个节点之间没有边相连，V_t亦是如此。节点之间边的权值由tf-*i*df来计算。对于这样的二分图来说，聚类算法可以看作是对文档节点和单词节点的同时划分。定义“规范化剪切”(normalized cut)来作为衡量划分效果的目标函数，其定义为：

$$\min\mathrm{imize}\left(g_2=\sum_{r=1}^{k}\frac{\mathrm{cut}(V_r,V-V_r)}{W(V_r)}\right)$$

式中　V_r是第r个划分块中的节点集合；$W(V_r)$是和V_r邻接的划分块与V_r之间节点连接的权值之和，V_r中既包含文档节点也包含单词节点。该目标函数的意义与式(1)类似，所不同的是在进行划分的过程中消去的是单词与文档之间的边。

4　实验及结果比较

4.1　实验数据集

为了验证所定义算法的聚类效果，从不同领域借鉴6个数据集来进行测试。其中数据集k1a、k1b选自WebACE；数据集re0、re1来自于路透社-21578文本分类测试集1.0；数据集la1、la2是从TREC-5 中挑选出来的。每个数据集去除其中的公共单词，用Porter’ssuffix-stripping算法对保留下来的单词加以

处理。各数据集的特征如表1所示。

表1 数据集特征

	k1a	k1b	re0	re1	la1	la2
文档数	2 340	2 340	1 504	1 657	3 204	3 075
单词数	21 839	21 839	2 886	3 758	31 472	31 472
文档类别数	20	6	13	25	6	6

4.2 实验结果及比较

4.2.1 原始数据划分结果

为了对理论设计进行验证，首先对原始数据集采用二分图进行划分，得到相应的测评指标值如表2所示。

表2 原始图划分结果

	NMI	CE	ERR	F-measure
k1a	0.514 642	1.721 594	0.523 932	0.441 122
k1b	0.577 733	0.539 664	0.370 940	0.644 877
la1	0.246 216	0.872 771	0.616 729	0.420 352
la2	0.321 217	0.840 732	0.569 106	0.466 708
re0	0.315 199	2.527 995	0.402 926	0.363 055
re1	0.374 787	2.976 249	0.494 267	0.322 628

4.2.2 实验内容

为了按算法完成实验数据的聚类分析，得到对应的评测数据，实验共设计了6个功能模块，包括超团挖掘预处理、超团挖掘、建立扩展特征集、权值正规化、图划分文件构造、二分图划分和结果统计[9-10]。由于本文篇幅有限，具体的实验步骤将另文描述。

4.2.3 利用超团扩展特征划分结果

依据经验，对不同的数据集选择不同的参数，分别尝试使用原始类别数和原始类别的两倍作为划分块的数目，得到的较好的划分结果如表3所示。

表3 实验结果

		NMI	CE	ERR	F-measure
k1a	STD	0.514 642	1.721 594	0.523 932	0.441 122
	WHP	0.590 880	1.389 075	0.497 009	0.501 794
k1b	STD	0.577 733	0.539 664	0.370 940	0.644 877
	WHP	0.658 099	0.573 262	0.132 051	0.844 936
la1	STD	0.246 216	0.872 771	0.616 729	0.420 352
	WHP	0.472 987	1.278 142	0.312 734	0.646 572
la2	STD	0.321 217	0.840 732	0.569 106	0.466 708
	WHP	0.419 227	1.419 978	0.378 211	0.576 292
re0	STD	0.315 199	2.527 995	0.402 926	0.363 055
	WHP	0.329 434	1.791 696	0.428 856	0.361 028
re1	STD	0.374 787	2.976 249	0.494 267	0.322 628
	WHP	0.401 863	2.129 910	0.653 591	0.324 963

表3中带下画线的数值表示比较指标中较好的结果。由实验结果可以看出，单词超团作为文档向量扩展特征的二分图划分算法相对于传统的二分图划分算法具有一定的优越性。

5 总 结

传统的聚类方法考虑文本之间的两两相似度，采取K-means等算法来进行聚类挖掘。本文在传统的文本挖掘基础上提出了一种利用单词超团的二分图文本聚类算法，该算法用文档中单词的关联模式来评估文档间的相似度及主题类别预测；用图划分策略来大大降低文档相似度比较算法的复杂度；将超团作为特征结构的扩展可以在一定范围内减少语言信息的丢失。实验证明，该算法具较高的有效性，可有效提高聚类效果。

参 考 文 献

[1] 张俊林. 基于语言模型的信息检索系统研究[D]. 北京: 中国科学院, 2004.
[2] 熊云波. 文本信息处理的若干关键技术研究[D]. 上海: 复旦大学, 2006.
[3] 袁军鹏, 朱东华, 李 毅, 等. 文本挖掘技术研究进展[J]. 计算机应用研究, 2006, 02.
[4] HU Tian-ming, SAM Y S. Finding centroid clusterings with entropy-based criteria[J]. Knowledge and Information Systems, 2006,10(4): 505-514.
[5] HU Tian-ming, QU Chao, CHEW L T, et al. Preserving patterns in bipartite graph partitioning. Proc[C]//18th IEEE Int Conf Tools with Artificial Intelligence (ICTAI'06). Washington D C: IEEE Press, 2006: 489-496.
[6] HU Tian-ming, XIONG Jin-zhi, ZHENG Geng-zhong. Similarity-based combination of multiple clusterings[J]. Int J Computational Intelligence and Applications, 2005, 5(3): 351-369.
[7] XIONG Hui, TAN Pang-ning, VIPIn K. Hyperclique pattern discovery[J]. Data Mining and Knowledge Discovery Journal, 2006, 13(2): 219-242.
[8] PAK K C, MARTINE D F, SCHLAG, et al. Spectral K-way ratio-cut partitioning and clustering[J]. IEEE Trans on CAD of Integrated Circuits and Systems ,1994,13(9): 1088-1096.
[9] DHILLON I, GUAN Y, KULIS B. A fast kernel-based multilevel algorithm for graph clustering[C]//In:. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago: ACM, 2005.
[10] HUANG Yao-chun; XIONG Hui; WU Wei-li, et al. A hybrid approach for mining maximal hyperclique patterns[C]//In: Proceedings of ICTAI. Boca Roton: IEEE Press, 2004: 354-361.

编 辑 熊思亮