文章编号：1003-0077（2011）00-0000-00

基于主题网络的伪主题分析*

闫蓉[1,2]，高光来[1,2]

[1]（内蒙古大学 计算机学院，内蒙古 呼和浩特 010021）

[2]（内蒙古自治区蒙古文信息处理技术重点实验室，内蒙古 呼和浩特 010021）

摘要：传统无监督的主题建模方法利用相互独立的主题变量抽象描述文本语义，忽略了各主题内部隐含的结构和联系，粗粒化的文本主题分析加剧了“强制主题”问题对文本建模的影响。本文通过研究主题网络社区内部结构，结合主题内部语义耦合关系与网络拓扑结构，提出伪主题分析方法来识别和解释主题，实现从网络结构角度描述文本语义特征，弥补统计主题分析方法对文本语义结构刻画的不足。

关键词：伪主题分析；主题网络；文本理解

中图分类号：TP391　　　　　**文献标识码：**A

Pseudo Topic Analysis Based on Topical Networks

YAN Rong[1,2] , GAO Guanglai[1,2]

[1] (College of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia 010021, China)

[2] (Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Hohhot, Inner Mongolia 010021, China)

Abstract: Traditional unsupervised topic models usually represented the document semantic by using a set of topics where no relationships between the topics, which would result in the imperfection to the text topic modeling in a coarse-grained manner, and intensify the effect of the ‘forced topic’ problem for the text topic modeling because ignoring the internal structure and relationships within each topic. Based on the study of the community inner structure, and combined with the internal coupling relationships and network topology, this paper proposed a novel pseudo topic analysis approach. It achieved identify and explain the topic, and accomplished the description the textual semantic features from the network structure point of view so as to remedy the deficiency of the textual semantic structure of the statistical topic modeling methods.

Key words: pseudo topic analysis; topical network; text understanding

1 引言

概率主题模型，如 LDA（Latent Dirichlet Allocation）[1]和 PLSA（Probabilistic Latent Semantic Analysis）[2]为用户在海量信息中筛选和挖掘有效信息发挥了重要的作用[3]。已经有很多工作致力于构建新的主题模型和改进算法来捕获主题结构[4-6]和实现主题模型的可视化[7-9]。但是，该类文本主题分析技术多数是利用统计方法实现文本主题获取，通常考虑词频较大的词项对于文本内容的贡献，核心假设是利用文本集中包含特定数目的潜在主题变量，来构建文本语义描述空间。这些数目的潜在主题变量在表达文本集固有抽象的同时，也利用多个不同主题变量抽象地表示文本的不同语义，实现了文本间的区别。但是这种方法由于受到其概率主题建模机理的限制，文本主题分析结果并不理想，原因有三点，分别是：第一，利用统计方法获取这些潜在主题变量的同时，假设各潜在主题变量之间是相互独立的，尽管各潜在主题变量之间有结构，但是潜在主题变量内部描述却无结构和无联系。而实际情况是，

* **收稿日期：**　　　　　**定稿日期：**

基金项目：国家自然科学基金项目（61662053）；内蒙古自然科学基金项目（2018MS06025）；内蒙古大学高层次人才项目（21500-5175128）

作者简介：闫蓉（1979—），女，讲师，博士，主要研究领域为自然语言处理和信息检索；高光来（1964—），男，教授，博士生导师，主要研究领域为智能信息处理。

各潜在主题变量在表达文本时，它们之间并不是孤立的，同一词项会同时出现在多个不同潜在主题变量中，使得利用潜在主题变量实现文本内容表达效用降低。第二，文本主题建模所抽象表达的语义，是通过描述各潜在主题变量中排名靠前的那部分词项的分布来实现的，但是这些词项间并无明显关联关系，人工界定主题解释非常困难。第三，各文本语义由于被“强制”利用特定数目的潜在主题变量表达，会由于“强制主题”问题（forced topic problem）[10]，有可能造成对不同文本的主题表达结果一致，无法有效辨识文本语义。尤其对于短文本的主题分析，进而影响到与之相关的诸多文本处理任务，如文本检索和文本分类等。

到目前为止，有诸多研究工作都致力于改善这种状况。在这其中值得关注地是，在过去的几十年间，大量的数据分析表明“无标度”特性广泛存在于各种网络中。近年来对语言的社会网络分析成果[11-13]，使得我们可以实现文本的复杂网络结构表达，并利用现有社会网络分析技术对其进行分析和研究，重新审视和实现文本理解。

本文致力于结合主题内部语义耦合关系与网络拓扑结构分析，识别和解释文本主题语义，梳理和获取更加细化的主题分析结果，提出一种基于主题网络的伪主题分析方法（Pseudo Topic Analysis, PTA），通过构造文本主题网络图，诣在通过对各主题网络的社区内部结构分析和解释，获取描述各主题词项之间更加细化的语义关联关系，调整主题网络中各词项重要度，突显描述主题语义的词项，实现丰富和补充主题内容表达，有助于更好地解释主题表达内涵。

2 相关工作

复杂网络显著的动力学特征之一就是具有社区结构[14]，即社区内各节点连接紧密，但两个社区之间节点连接稀疏。知晓复杂网络社区结构，对实现更准确地理解并分析复杂系统的拓扑结构及动力学特性起着重要的作用。关于复杂网络社区结构的研究主要包括两种：社区结构及关联关系的研究和社区结构识别的研究。

关于文本网络的社区结构研究，大体包括与文本处理相关具体任务实现和文本主题内容分析两种。其中，相关任务实现包括词义消歧[15]、文本分类[16]和信息推荐[17]等。文本的主题内容分析主要集中对文本主题识别研究[9,18-22]。Smith 等[9]通过获取主题内各词项间关联关系构建各主题内部词项间的网络关系图和主题间的网络关系图。但是，该文所构建的词项间的网络关系图仅考虑了主题内部各词项间的局部关联关系，未充分考虑各词项在文本数据集中的全局关联关系。Zhou 等[18]利用社会网络社区发现方法，提出一种自动文本主题生成方法 HLSM。Lancichinetti 等[19]利用社区发现方法，优化概率主题建模结果。Arruda 等[20]提出新的文本社会网络表示方法，同时兼顾文本内容和主题结构，获取词项间的语义关联关系。Akimushkin 等[21]研究了文本中不同部分的词共现网络的拓扑演化。Chen 等[22]利用社区识别算法实现文本主题发现，其工作本质上构建的是一种基于知识源的主题网络图，通过模块度计算划分社区获取主题分布，并利用各主题节点的紧度值评估其对于文本内容贡献的重要程度。

但是，这些方法并没有从主题内部各词项间所具备的潜在语义耦合关系与网络拓扑结构相结合，实现对文本各主题的理解。从某种角度而言，其分析结果仍是一种粒度较粗的文本语义分析。但事实是，出现在不同主题中的相同词项对于主题内容贡献程度是不一样的，其不同的语义贡献程度不仅仅体现在词项-主题概率分布中的概率值大小的不同，还在于词项间语义关联关系的强度程度不同所体现的语义表达的不一致。

近几年，付京成等[23,24]的研究致力于通过研究社区内部结构，从而获取更加合理的网络中各节点在社区结构中的作用，即在社区结果内部识别两种不同的社区组织结构，分别是领导者社区和自组织社区。其中，在领导者社区内部存在一个或者多个具有较大度数的节点，其地位要高于自组织社区中各节点。各领导结点不仅连接了社区中其余节点，还保证了社区

的稠密和维护社区之间的通信，体现的是网络拓扑结构中的中心性原则。自组织社区内各节点度数基本一致，各节点在社区中的地位等同，体现的是网络拓扑结构中的自组织性原则。

综上，我们可以在文本的主题建模的复杂网络结构中，通过社区划分识别其内部的领导社区和自组织社区，可以实现从复杂网络社区内部结构，来审视主题变量在抽象表达文本语义过程中的生成机制，从而细化明确各主题变量所隐含的内部语义，将有助于文本的主题语义分析，减少“强制主题”问题对文本分析影响，获取更加精细的文本间语义相似度。

3 基于主题网络的伪主题分析

基于主题网络的伪主题分析过程，本质上是在各主题的网络拓扑结构中，分析和识别其隐含的社区结构，并将表达主题内涵的词项通过社区内部结构分析，实现主题内部语义耦合关系与网络拓扑结构相结合，获取新的主题特征来描述原主题分析结果，即不断的修正主题网络中各词项节点的重要程度及词项节点对之间的关联程度，将其作为新的主题分析结果。图 1 所示为伪主题分析获取构架图。

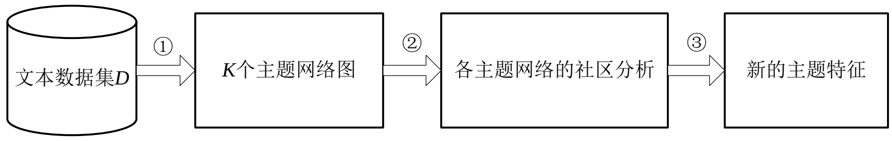


图 1 伪主题分析获取构架图

3.1 主题网络图的构建

本文采用标准的 LDA 对文本数据集进行主题建模。设文本数据集 D 有 K 个主题 $T=\{T_1,T_2,\ldots,T_K\}$，就有 K 个主题网络图，表示为 $G=\{G_1,G_2,\ldots,G_K\}$。其中，每一个主题网络可以表示为无向图 $G_i=(V_i,\ E_\text{i})$，$i\in[1,K]$。每个主题网络的节点集，表示为 $V=\{v_1,v_2,\ldots,v_n\}$，节点总数记为 $n=|V|$，节点 v 的度记为 k_v；每个网络的边集 E 中每条边 $e_{i,j}$ 对应 V 集中节点对 (v_i,v_j)之间的连接关系，边总数记为 $m=|E|$。如图 2 所示为构建的主题网络图。

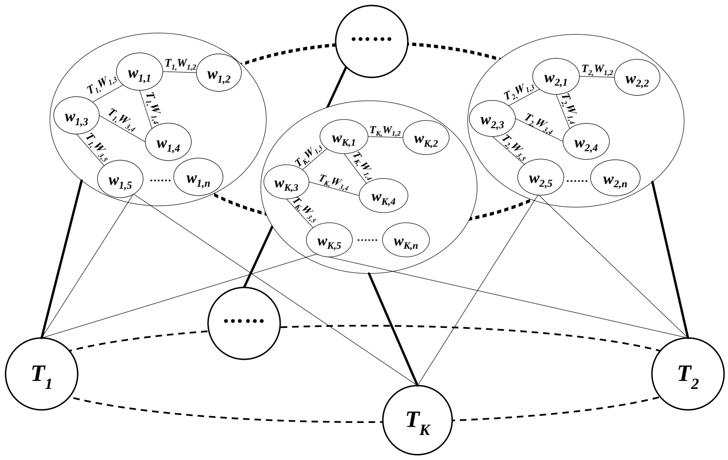


图 2 构建的主题网络图

其中，在每个主题网络图中，各节点是描述该主题的各词项节点，各节点的权重体现的是该词项节点描述主题内容的重要程度。节点对之间的连边权重体现的是，各词项在描述主题语义时所体现的语义关联关系。具体的定义如下所述。

3.1.1 节点的权重定义

本文把数据集主题建模后，将描述各主题排名靠前的 n 个词项作为各主题网络图的 n 个词项节点。节点的权重即为该节点在主题网络中的重要度。实质上，本文的伪主题分析就是从各主题网络中，抽取出更能抽象各主题内容表达的节点描述特征，并利用这些新的描述特征来构造数据集的伪主题分析结果。这就要求这些新的描述特征，不仅能够抽象各词项节点在各个主题网络中的重要程度，同时也要增加不同主题网络之间的区别。基于以上原则，

将主题网络 G_i 中各节点的权重定义如公式（1）所示：

$$Node_{weight}\left(v_j\right)=\sum_{w\in N\left(v_j\right)}k_{v_j}\cdot\phi_{i,v_j}\cdot w\left(v_j,w\right) \tag{1}$$

其中，$N(v_j)$表示节点 v_j 邻接节点的集合，ϕ_{i,v_j} 表示在主题 i（即主题网络 G_i）中第 j 个词项 v_j 的概率值，k_{v_j} 表示节点 v_j 的度数。$w(v,w)$表示节点对(v,w)之间的边权重。

3.1.2 边及边权重定义

判断每个主题网络中每个词项节点对之间是否存在连边，可以通过计算该节点对之间是否存在某种语义联系来获取。本文将利用工具 work2Vec[1]，将每个词项节点用词向量来抽象表示，通过计算两个词项节点向量之间的相似度值的大小，判断该节点对之间是否存在连边。若节点对相似度大于 0，则该节点对存在连边。反之，该节点对不存在连边。

为了能够更加准确地度量描述主题的各词项节点对的关联强度，需要对主题网络中节点连边的权重进行定义。通常，各种不同类型的复杂网络中边权重往往具有一定的实际意义，有助于社区的识别。因此，本文在定义主题网络图中节点连边权重的时候，不仅要考虑网络的拓扑结构，还要考虑节点之间连边的实际意义。这里，我们的工作主要是想通过对主题的网络结构描述，实现从网络结构角度描述文本特征，弥补统计方法对文本语义结构刻画的不足，所以在主题网络的边权重定义时，要从整个数据集层面来考虑，本文的边权重定义如公式（2）所示。

$$w\left(v,w\right)=\frac{w'\left(v,w\right)}{w'_{avg}} \tag{2}$$

其中，

$$w'_{avg}=\frac{\sum_{v,w\in E(G)}w'\left(v,w\right)}{\left|E\left(G\right)\right|} \tag{3}$$

$$w'\left(v,w\right)=\lambda_1 sim\_con\left(v,w\right)+\lambda_2 sim\_word\left(v,w\right) \tag{4}$$

$$sim\_con\left(v,w\right)=\frac{\left|N\left(v\right)\cap N\left(w\right)\right|^2}{\min\{\left|N\left(v\right)\right|,\left|N\left(w\right)\right|\}^2} \tag{5}$$

$|E(G)|$表示图 G 的边总数。sim_con 和 sim_word 分别表示节点对之间的网络拓扑结相似度和词向量相似度。$N(v)\cap N(w)$表示节点 v 和节点 w 的公共邻接节点集合。

3.2 主题网络图社区结构分析

描述主题的各词项，在共同抽象地表达主题语义时，对主题语义的贡献程度是不一样的。首先体现在词项-主题概率分布中的概率值大小的不同。通常，概率值较大的词项认为贡献程度较大。另外，还体现在这些词项间语义关联关系的强度不同所体现的语义表达的不一致。通常，主题所表达语义是由其中少数词项通过协调和语义关联其它词项实现的，且其所表达语义描述较强。同时，其它词项对这部分词项所表达语义起补充作用，且彼此间关联关系较弱。这些均为主题内部的耦合关系。

传统基于统计的概率主题建模方法，由于受其建模机理限制，无法获取主题内部耦合关系。值得注意的是，这种耦合关系与社区内部结构非常相似。我们可以利用社区内部结构分析方法应用到主题网络内部耦合关系的获取。其中，社区内部结构分为领导者社区和自组织

[1] http://code.google.com/p/word2vec

社区[23,24]。在领导者社区内部存在少数几个领导节点高度关联其余节点。同时，其余节点必须通过这几个少数节点的支配才能相互联系。在自组织社区内部各点，不存在任意节点具有支配其它节点的功能，且社区内部各节点地位等同。

在付京成等 2017 年的工作中，通过计算社区内各节点度数的方差，与相同节点数的随机零模型的节点度数的方差比值作为社区划分依据[24]。但是，在实际的网络中，节点度数仅是节点属性描述特征之一，还包括具体网络中节点的实际含义，即节点点强度。在本文所描述的主题网络中，网络中各节点点强度即为其描述主题内涵的强度大小。所以我们对划分依据进行了部分调整，如公式（6）所示：

$$\rho=\frac{VAR_{\mathrm{real}}}{VAR_{rand}} \tag{6}$$

其中，VAR_{real} 的 VAR_{rand} 分别表示主题网络中社区的节点度数及点强度的方差和对应随机社区的节点度数及点强度的方差的期望。随机社区节点的点强度就是节点的点度数。这里，我们采用和文献[24]相同的阈值标准，将 1 作为阈值。当 $\rho>1$ 时，识别为领导者社区；当 $\rho<1$ 时，识别为自组织社区；当 $\rho=1$ 时，既不是领导者社区也不是自组织社区。

除此之外，在实际的主题建模过程中，一定会有一部分词项同时出现在多个不同主题描述中的情况发生，即有部分词项节点在社区识别过程中，会出现在多个不同社区中，存在重叠社区现象。通常，处于重叠社区的那些节点，对完成网络间语义信息流动和不同网络间意义的关联起到关键作用。所以，在实际的主题网络社区识别结果中，对于处理重叠社区的那部分词项节点，本文将适当增加其节点属性重要度。

3.3 新的主题特征的获取

在整个伪主题分析获取构架中，最关键的部分就是识别主题网络中最能体现主题语义内涵的词项节点信息。直观地讲，重要程度大且能够最大语义关联其它节点的那些节点，是最有可能体现主题语义内涵的。这与社区内部结构中的领导者节点特点是一致的。本文将各主题网络图结构中，处于领导者社区且权重较大的节点，作为体现主题语义内容新的主题词项特征集。

4 实验与结果分析

4.1 实验数据

本文将对中、英两种不同语料进行实验。其中，中文采用 NTCIR8[2]提供的新华社简体中文四年的新闻语料 XINHUA（2002 年-2005 年），包括 308,845 个文档，涉及多种主题新闻语料。英文采用 MEDLINE[3]提供的五年的医疗文档语料 OHSUMED（1987 年-1991 年），包括 348,566 个文档，涵盖 270 种医学杂志发表的医疗文献。表 1 列出了中、英两个不同数据集的基本情况。

表 1 实验数据集描述

数据集	文档数	词项数	数据集	文档数	词项数
XINHUA2002	64,251	12,721,776	OHSUMED1987	36,890	2,819,114
XINHUA2003	73,431	15,444,634	OHSUMED1988	47,054	3,673,655
XINHUA2004	84,287	18,268,710	OHSUMED1989	49,805	3,976,735
XINHUA2005	86,858	18,647,309	OHSUMED1990	49,480	4,095,198
			OHSUMED1991	50,216	4,288,755

4.2 社区划分方法及评价指标

本文采用基于模块度最大化最好的社区划分算法之一 BGLL 算法[25]作为主题网络社区划分方法。

由于本文所构建的主题网络是无社区划分标签，所以评价标准采用模块性 EQ（Extended Modularity）[26]来度量社区发现质量。

[2] http://research.nii.ac.jp/ntcir/index-en.html
[3] http://medline.cos.com

设社区划分结果为 $C=\{C_1,C_2,\ldots,C_M\}$，EQ 值的计算如公式（7）所示：

$$EQ=\frac{1}{2}\sum_{i=1}^{M}\left(\sum_{v\in C_i}\sum_{w(\neq v)\in C_i}\frac{1}{O_vO_w}\left[A_{vw}-\frac{k_vk_w}{2m}\right]\right) \tag{7}$$

其中，M 为社区划分数，O_v 表示在最终社区划分结果中节点 v 属于的社区数，$\mathbf{A}$ 是原始网络的邻接矩阵，m 是社区划分前原始网络的总边数。

4. 3 主题获取

本文采用开源的 JGibbLDA[4]工具实现对文本数据集的主题建模。设置初始主题数目 K=10，超参数设定 α=50/K、β=0.01；Gibbs 采样的估计迭代次数设定为 100 次，返回主题描述词项个数 *word_number*=20。主题数目依次取 K=10、20，直至 100，分别对数据集进行主题建模。为了降低少数低频词对文本建模结果的影响，实验预先去除了数据集中词频低于 5 的部分词项，其中包括 XINHUA 中 130,363 个词项和 OHSUMED 中 77,322 个词项。本文利用困惑度 *Perplexity*[6]度量建立的主题模型的生成性能，取困惑度取值最低值对应的主题数目作为数据集的最佳主题数目 K。

模型困惑度值采用公式（8）计算：

$$Perplexity\left(R_{test}\right)=exp\left(\frac{-\sum_{j=1}^{J}log\left(P\left(d_j\right)\right)}{\sum_{j=1}^{J}N_j}\right) \tag{8}$$

其中，R_{test} 表示有 J 个文档的测试集，N_j 表示第 j 篇文档 d_j 包含的词项数；$P(d_j)$表示模型产生文档 d_j 的概率。由如图 3 所示中、英数据集 *Perplexity* 值变化曲线，可知中、英文数据集最佳主题数目分别为 60 和 70。

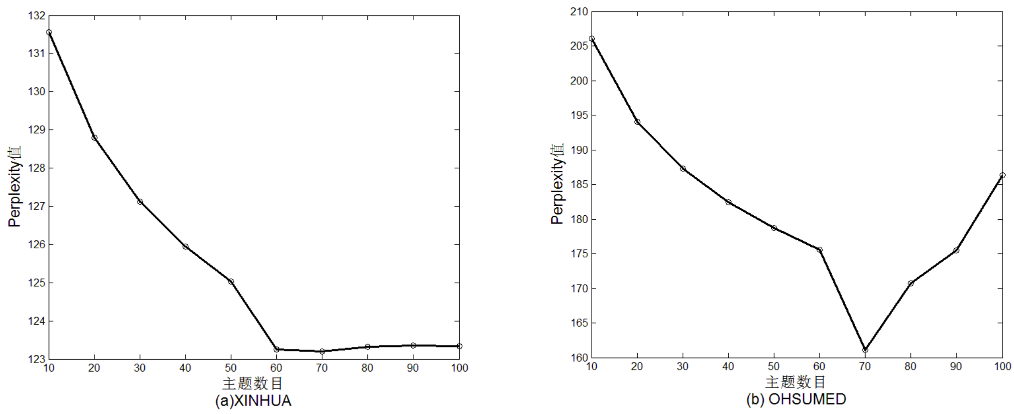


图 3 中、英数据集 *Perplexity* 值变化曲线

4. 4 实验结果和分析

表 2 所示为中、英两种不同文本数据集原始的主题分析结果（*top*-20）和经过伪主题分析的样例结果比较。图 4 所示为相应样例的伪主题图结果描述。

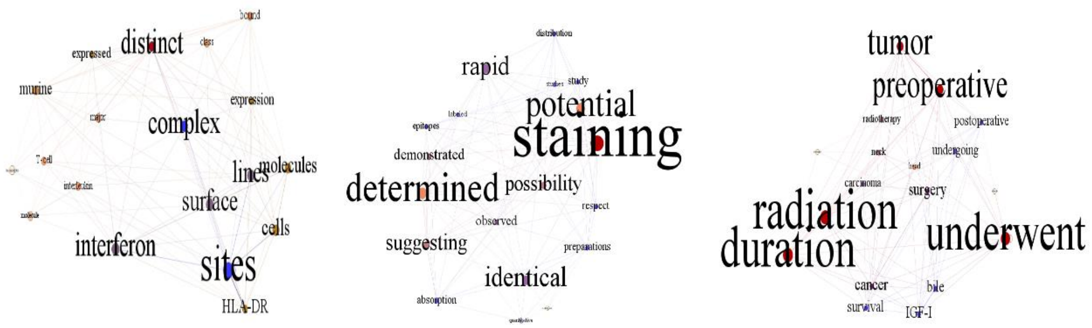
表 2 中、英数据集原始主题分析和伪主题分析结果样例比较

XINHUA 数据集			OHSUMED 数据集		
主题	原始主题分析结果	伪主题分析结果	主题	原始主题分析结果	伪主题分析结果
Topic2	增长，去年，今年，美元，经济，出口，下降，增加，统计，同期，达到，消费，占，显示，季度，上升，同比，达，报告，减少	增长，经济，美元，消费，同期，同比，出口，达到，上升，统计	***Topic4***	expression, cells, class, surface, lines, complex, expressed, T-cell, molecules, cells., major, bound, interferon, HLA-DR, murine, molecule, sites, interleukin, distinct, transcripts	sites, complex, distinct, interferon, lines, HLA-DR, molecules, cells

[4] http://sourceforge.net/progjects/jgibblda/

Topic4	教育，大学，学生，学校，儿童，人才，培训，妇女，专业，学习，就业，青年，培养，学院，青少年，工作，务，职业，高校，社会	培养，儿童，职业，工作，妇女，服务，就业，青少年，青年，教育，培训，人才	***Topic16***	observed, study, studies, distribution, suggesting, rapid, investigated, epitopes, demonstrated, potential, determined, absorption, staining, possibility, labeled, quantitative, respect, identical, preparations, investigated	staining, determined, potential, identical, rapid, suggesting, possibility
Topic6	上海，国际，举办，世界，城市，中心，协会，来自，展览，举行，活动，主办，上海市，成功，申办，博览会，今天，中国，世博会，浦东	世界，博览会，展览，城市，上海，举办，活动，中心，主办，协会，举行	***Topic18***	hospital, patient, support, study, time, program, costs, nursing, programs, status, admitted, elderly, community, care, patients, survey, recommended, systems, improve, benefits	underwent, duration, radiation, preoperative, tumor, surgery

(a) Topic2 (XINHUA) (b) Topic4 (XINHUA) (c) Topic6 (XINHUA)



(d) Topic4 (OHSUMED) (e) Topic16 (OHSUMED) (f) Topic18 (OHSUMED)

图 4 样例的伪主题图结果描述

从表 2 和图 4 的结果可以看出，对各主题网络的伪主题分析结果不仅可以更能体现主题表达内涵，还进一步体现了这些词项间的关联关系。

图 5 所示为中、英数据集各主题网络图模块性结果。

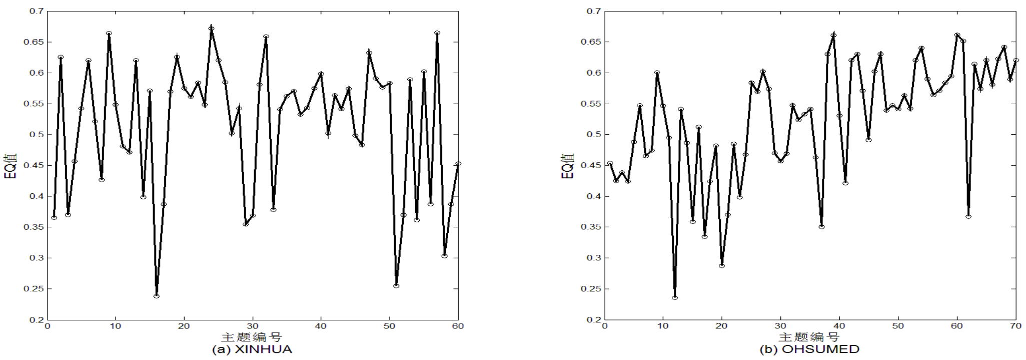


图 5 中、英文本集各主题网络图 EQ 值结果

从图 5 结果来看，本文所提方法对各主题网络模块性整体表现良好。图 5 中存在个别主题的模块性值较低，分析其主要原因是由于该主题描述中组成词项关联关系缺乏影响社区划分结果。

总体而言，本文所提方法，在主题内容发现过程中，综合考虑了网络的拓扑特征和原始描述主题词项的权重信息，能够给出更符合主题所表达语义的伪表达结果。

5 总结

本文提出了一种基于主题网络的伪主题分析方法，该方法综合考虑网络拓扑结构和主题网络社区内部结构，从全局数据集角度考虑，评估主题网络各社区节点重要度，实现从网络结构角度抽象描述文本语义特征，弥补统计方法对文本语义结构刻画的不足。对实际文本数据集的主题网络的伪主题分析实验中，模块性表现良好。本文所提方法可以帮助用户更好地分析和理解大规模数据，进一步可用于文本主题内容可视化分析应用中。

参考文献

[1] Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3: 993-1022.

[2] Hofmann T. Probabilistic latent semantic indexing[C]. In Proceedings of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 1999), 1999, pp. 50-57.

[3] Zhai C. X. Probabilistic topic models for text data retrieval and analysis[C]. In Proceedings of the 40th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 2017), ACM Press, New York, NY, 2017, pp. 1399-1401.

[4] Lancichinetti A., Sirer M. I., Wang J. X., et al. High-reproducibility and high-accuracy method for automated topic classification[J]. Physical Review X, 2014, 5(1), No. 11007: 1-11.

[5] Blei D. M., Griffiths T. L., Jordan M. I. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies[J]. Journal of the ACM, 2010, 57(2):17-24.

[6] Blei D. M., Lafferty J. D. Correlated topic models[C]. In Proceedings of the 18th International Conference on Neural Information Processing Systems, 2005, pp. 147-154.

[7] Allison June-Barlow C., Blei D. M. Visualizing topic models[C]. In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, 2012, pp. 419-422.

[8] Wei F. R., Liu S. X., Song Y. Q., et al. TIARA: a visual exploratory text analytic system[C]. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 153-162.

[9] Smith A., Chuang J., Hu Y., et al. Concurrent visualization of relationships between words and topics in topic models[C]. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Baltimore: ACL, 2014, pp. 79-82.

[10] Li X., Ouyang J., Lu Y., et al. Group topic model: organizing topics into groups[J]. Information Retrieval, 2015, 18(1):1-25.

[11] 刘知远, 孙茂松. 汉语词同现网络的小世界效应和无标度特性[J]. 中文信息学报, 2007, 21(6):52-58.

[12] Cong J., Liu H. Approaching human language with complex networks.[J]. Physics of Life Reviews, 2014, 11(4):598-618.

[13] Kulig A., Drożdż S., Kwapień J., et al. Modeling the average shortest-path length in growth of word-adjacency networks.[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2015, 91(3):032810.

[14] Girvan M., Newman M. E. J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America (PNAS), 2001, 99(12):7821-7826.

[15] Edilson A. C. J., Alneu A. L., Diego R. A. Word sense disambiguation: a complex network approach[J]. Information Sciences, 2018, 442-443: 103-113.

[16] De A. H. F., Costa L. D. F., Amancio D. R. Using complex networks for text classification: Discriminating

informative and imaginative documents[J]. Epl, 2016, 113(2):28007.
[17] Yu L., Huang J., Zhou G., et al. TIIREC: a tensor approach for tag-driven item recommendation with sparse user generated content[J]. Information Sciences, 2017, 411: 122-135.
[18] Zhou G. R., Chen G.. Hierarchical latent semantic mapping for automated topic generation[J]. International Journal of networked and distributed computing, 2016, 4(2):127-136.
[19] Lancichinetti A., Sirer M. I., Wang J. X., et al. High-reproducibility and high-accuracy method for automated topic classification[J]. Physical Review X, 2015, 5(1): (011007)1-11.
[20] Arruda H. F. D., Costa L. da F., Amancio D. R. Topic segmentation via community detection in complex networks[J]. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2016, 26(6):163-222.
[21] Akimushkin C., Amancio D. R., Jr O. O. Text authorship identified using the dynamics of word co-occurrence networks[J]. PLoS ONE, 2017, 12(1):1-15.
[22] Chen Q., Guo X., Bai H. Semantic-based topic detection using markov decision processes[J]. Neurocomputing, 2017, 242:40-50.
[23] Fu J. C., Wu J. L., Liu C. J., Xu J. Leaders in communities of real-world networks[J]. Physica A, 2016, 444:428-441.
[24] Fu J. C., Zhang W. X., Wu J. L. Identification of leader and self-organizing communities in complex networks[J]. Scientific Reports, 2017, 7(704):1-10.
[25] Blondel V. D., Guillaume J. L., Lambiotte R., et al. Fast unfolding of community hierarchies in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10:1-12.
[26] Shen H. W., Cheng X. Q., Cai K., et al. Detect overlapping and hierarchical community structure in networks[J]. Physica A Statistical Mechanics & Its Applications, 2009, 388(8):1706-1712.

闫蓉（1979—），女，讲师，博士，主要研究领域为自然语言处理和信息检索。Email: csyanr@imu.edu.cn；

高光来（1964—），男，教授，博士生导师，主要研究领域为智能信息处理。 Email:csggl@imu.edu.cn。