# 基于语言网络的文本表示模型研究[1]

吴思竹[1]，张智雄[2]

（1 中国医学科学院医学信息研究所 北京 100020，

2 中国科学院国家科学图书馆 北京 100190）

**摘要：**本文对基于语言网络的文本表示模型及其构建的主要思路和方法进行研究。对模型的特点进行分析，总结和归纳构建基于语言网络的表示模型的主要流程及模型的主要应用领域。最后，对模型研究的相关问题进行初步探讨。

**关键词：**文本表示模型，语言网络，网络图，语义网络

# Research on Text Representation Model based on Language Network

WU Si-zhu[1]，ZHANG Zhi-xiong[2]

（1 Institute of Medical Information，Chinese Academy of Medical Sciences, Beijing 100020, China

2 National Science Library, Chinese Academy of Sciences, Beijing 100190, China）

**Abstract：**This paper mainly researches on research ideas and methods of text representation model based on language network and its construction. It analyses the characteristics of the model and summarizes the procedures of the model construction and the main application fields of the text representation model application。Finally，it preliminary studies on the relevant problems of text representation model based on language network.

**Keywords：**text representation model，language network，network graph，semantic network

## 1 引言

文本表示模型能够将结构化和非结构化的自然语言文本映射为计算机能够识别、处理、分析的结构化特征表示形式，将文本抽象为描述及替代文本的数学模型，它是文本挖掘、信息检索、文本分类等研究的基础问题。常用的文本表示模型有 N-gram、向量空间模型（Vector Space Model，VSM）、布尔逻辑模型、概率模型等。N-gram 是一种统计语言模型，第 N 个词出现的概率由前 N-1 个词决定，其保留了文本中的词序特征。VSM 将文本信息映射为数值信息。每篇文本用一行向量表示，向量中的数值表示对应的特征词是否在文本中出现或表示通过其他计算方式(如 tf*idf)得到的权重。通常 VSM 被用于对多文本建模。这几种表示模型在自然语言处理中应用广泛，但是其在文本特征揭示上存在一定的局限性，如 N-gram 虽然保留了文本的词序信息，但是文本中的词间结构关系、上下文关系无法体现。VSM、概率模型等忽略了文本中的上下文关联，词序关系、结构特点。一些研究者对二维模型进行了拓展，Liu 提出将文本表示为线性代

[1]本文系国家自然科学基金项目“基于语言网络的文本主题中心度计算方法研究”（项目编号：61075047）

数中的三维张量模型，建立三个维度坐标，用“a”到“z”的 26 个字母表示，而其他字符都用“-”代替。结合 N-gram 将文本映射为三维张量 A=$\{a_{i1},a_{i2},a_{i3}\}\in R^{I_1\times I_2\times I_3}$[1]。Hu 通过扩展词性特征生成三维空间的文本表示模型[2]。第一维表示特征词的词频，第二维表示特征词的词性，第三维表示文本数。单篇文本被表示为特征词和词性的二维向量。虽然三维空间表示模型在一定程度上补充了文本的某些特征属性，但是仍不能充分、灵活的对文本中的词序、词间关系、词性、词的角色功能等类似信息予以揭示。

在语言学领域，研究者将语言（语音或文本）构建为语言网络，研究同种语言或多种语言间的复杂网络结构特征，并将语言网络分为静态网络和动态网络，认为通过词表等语言资源构建的网络是静态语言网络，而根据真实文本构造的网络为动态语言网络[3]。自然语言处理领域中，研究者把文本构建为语言网络表示模型（Language Network Text Representation Model，LNTRM），即网络或图的表示形式，其提供的是一种机器可读的形式，能够相对灵活的表示文本结构、语序，揭示语法、语义信息。语言网络具有较好的可扩展性和可理解性，与图挖掘的相关算法结合，为文本表示和文本挖掘提供了新的视角。

本文主要对基于语言网络的文本表示模型的组成、构建方法进行分析和归纳，总结模型特点和当前的主要的应用情况。

# 2 基于语言网络的文本表示模型及其构建

1968 年 J.R.Quillian 在研究人类联想记忆时曾提出语义网络，将其作为知识表示方法。1972 年，西蒙在自然语言理解系统中将语义网络表示方法用于推理。语义网络被定义为一种有向标识图，图中节点表示各种事物、概念、属性、动作、状态等，有向弧表示它所连接的节点间的某种语义联系。语义网络中所表示的语义完全依赖于系统预定义和自定义的语义关系。而后，随着对基于复杂网络或图的方法研究的逐渐增多，研究者以更多灵活的形式将自然语言文本映射为语言网络，并将其广泛应用于文本处理、挖掘等诸多领域。

语言网络模型是通过对自然语言文本特征的识别和抽取，将其映射为一种基于网络结构的特征图，它是对文本的抽象表征。实际应用中，基于语言网络的文本表示模型的定义不似传统的语义网络的定义那般严格，而是较为灵活。一般来说，节点和边是构建语言网络的重要组成部分，文本集或文本可以简化地被表示为 G={V，E}，其中 V 表示网络中的节点集合，E 表示边集合，此时，G 表示边无权重的语言网络，或 G={V，E，W}，W 表示边的权重集合，此时，G 表示边有权重的语言网络。

复杂的自然语言网络模型还包括节点属性和边属性，更多的文本特征可以通过节点属性和边属性进行进一步的补充和拓展。

## 2.1 节点类型及构建方法

研究者根据文本处理的最终应用需求，按不同粒度划分文本单元作为网络节点，并结合单

元间的不同关系构建文本语言网络模型。本文对现有研究和应用中的语言网络表示模型的构建方法进行归纳和总结，网络节点可以分别表示句子、实体、概念、事件、短语、词等。边可以通过文本中的词序关系、共现关系、语法关系（依赖关系、角色功能）、语义关系等构建。研究和应用中，最常选用的是以句子或单个词作为节点的网络构建方式。以句为单位进行文本分割或以空格为切分再将句子细化为单词，作为节点。Erkan 等将文本映射为以句子作为节点的网络，节点间的边表示成对句子间的相似关系，计算基于逆文档频改进的余弦相似度作为边的权重，构建文本表示模型[4]。Ohsawa 抽取文本中的词，将其作为文本表示网络的节点，计算词的同句共现值作为边的权重[5]。

短语、实体或事件的抽取方法复杂，构建以这些文本单元作为节点的语言网络模型需要更复杂的词性识别、句法解析、语义消歧等自然语言处理技术。虽然抽取难度大，但与以单个词作为节点的网络相比，短语、实体及事件在语义表达上更加明确和具有可理解性。Rusu 采用文本深层句法解析，利用 TreePenn 工具抽取句子的逻辑三元组：主语、谓语、宾语。将主语和宾语作为节点，并生成由主语指向宾语的边，谓语作为边的标签，将文本映射为基于句子三元组的有向网络[6]。Xie 构建基于短语网络的文本表示模型，其对文本进词切分和词性标注，通过定义 35 个解析规则将解析出的短语作为网络节点，节点间的边通过句间或文本间层次关系生成[7]。Qu 构建了一种基于属性事件依赖图的文本表示模型，节点表示属性事件，边表示事件间的因果关系，事件间的关系强度表示边的权重，边的方向是由前提事件指向后继事件[8]。其中，边被分为两类，一种是内容关系，其权重由两个属性事件在同一前提条件或结果事件中出现的共现频率表示；另一种是依赖关系，事件相互依赖的概率值作为边的权值。网络中的每个节点可以同时具有这两种关系。

此外，概念也被应用于构建语言网络节点。结合外在知识源（如本体或词表），将文本中的词映射为本体或词表中的规范概念，依据概念在本体或词典中的层级类属关系或语义相似度计算构建网络表示模型，这种网络可以看作为一种语义网络或概念网络。

基于语言网络的文本表示模型的节点并不局限于一种类型的文本单元，也可扩展为多种类型单元的组合。Grobelinik 等通过对新闻故事进行命名实体抽取和主题识别，将新闻集映射为实体和主题关系的网络图[9]。抽取时间、人物、地点和组织四种实体，并通过主题建模选取新闻主题作为节点，计算实体与主题的关系、实体间的关系、主题间的关系，构建新闻主题间的网络模型。

## 2.2 边类型及构建方法

Liu[10]将语言网络根据词汇间关系归纳为三种：共现网络、语法网络和语义网络。笔者认为其主要是根据边的构建方法来划分的，可以进一步归纳为边的生成是通过文本中的共现关系、语法关系（依赖关系）、语义关系（语义相似或非相似，本体或词表中的上下位关系、相关关系）。其中，词序关系没有单独列出，作为简单的表示方式被融合在其他关系中。（1）共现关系：利用文本单元共同出现在同一窗体（同文本、同句或特定间隔词的长度）的关系构建网络，将文本的结构化特征抽象为数值表示形式，并且可以根据文本中的词序构建有向网络模型。

Mialeea[11]对文本进行词性标注，提取名词或动词，选择 1-gram 作为网络节点。边的构建利用共现关系，如果两个词出现在最大 N 个词的窗口中，则它们之间存在联系，N 一般设置为 2-10 个词。边的方向为文本中的词序顺序。Palshikar 将文本构建为无向网络，通过过滤标点符号、数字、停用词及词根提取将文本划分为词，通过词频控制选择出现频率大于特定阈值的词作为网络节点[12]。节点间的边通过共现关系建立，边的权重表示词间的不相似度。（2）语法关系：通过语法解析，生成节点间的语法依赖关系网络，这种网络是有向的，边的方向表示语法结构，施动关系、因果关系等。Marneffe 对句子进行句法解析，通过依赖关系构建语法树[13]。Huang 将每篇文本建模为网络，节点表示词，节点属性表示词频信息，边利用语法信息及同句相邻的共现关系构建，其通过词性标签区别不同语法关系，并赋予边的权重[14]。（3）语义关系：利用外在词表、本体(如 WikiPedia、WordNet）等，将节点映射为本体或词表中的概念或术语，通过节点间属性、类属、泛化、所属等多种关系进行语义相似度、相关度计算构建语义关系。Grineva 通过文本抽取 N-gram，对每个 N-gram 构建它的变形词，找出其在 WikiPedia 中对应的标题，利用 Wikipeida 进行消歧作为网络节点。通过计算节点在 WikiPedia 中的语义相似度作为边的权重，构建了表示文本的加权网络图[14]。

不同层面构建的网络的结构特征、表达含义的深度具有差别。同一段文本经过不同的构造方式产生的语言网络具有差别，如将下面一段来自 PubMed 数据库中的摘要文本构建为共现、语法及语义网络：

Foot problems in patients with diabetes cause substantial morbidity and may lead to lower extremity amputations. These risks may be reduced by appropriate screening and intervention measures. Effective screening assigns the patient to a risk category and dictates both the type and frequency of appropriate foot interventions. Less than half of diabetic patients in tertiary care hospital in Thailand received annual foot examination and there are limited data available on the nature of foot problems in such setting. This study reported a cross-sectional data of 438 diabetic patients attend tertiary diabetic clinic in the university hospital in Northern Thailand. Neuropathy manifestations as skin dryness, limitation of joint mobility and insensate to monofilament was the most common manifestation of diabetic foot problems in this setting. Most patients were not protected by proper footwear. More effort is needed to educate diabetic patients about foot care and improve their choice and selection of footwear.

语言学研究侧重于研究语言网络的结构、组成、语言特点等，将虚词作为句法组成的重要部分，认为其在语言表达中起到重要作用。因此，构建语言网络会保留大量虚词。但是在文本处理中，文本表示模型主要用于计算机理解，侧重获取文本的实质内容，因此，多保留有意义的实词作为有效知识单元，而虚词被作为停用词滤除。在构建的三种不同层面的语言网络中，共现网络利用同句共现关系构建，节点为该段文本中的名词短语和动词。共现关系建立的是文本整体关联，是潜在的关系，较易实现。语法网络主要通过句子语法依赖关系构建，利用语法解析工具解析句子，通过相同节点合并网络。虽然共现和语法网络均反映文本的原始结构，但语法网络不等同于共现网络。语法依赖关系是对句子结构的进一步解析，词在句子中位置不同，其语法功能作用也有所区别。语义网络相对前两种，是一种更深层的语言网络，是在文本结构之上的语义层面的关系网络。语义网络构建以概念作为节点，示例中是将能够映射到一体化医学语言系统 UMLS 中的概念作为节点，边是基于语义相关度计算而获得的，简单示例见图 1、图 2、图 3。三种网络从不同角度和层次反应文本的结构和内容关联。共现网络较为密集，体现

上下文的潜在关联；语法网络中，动词占据网络中心节点；语义网络虽然节点数量和关联有限，但基于更高层次，表达上层语义关联。

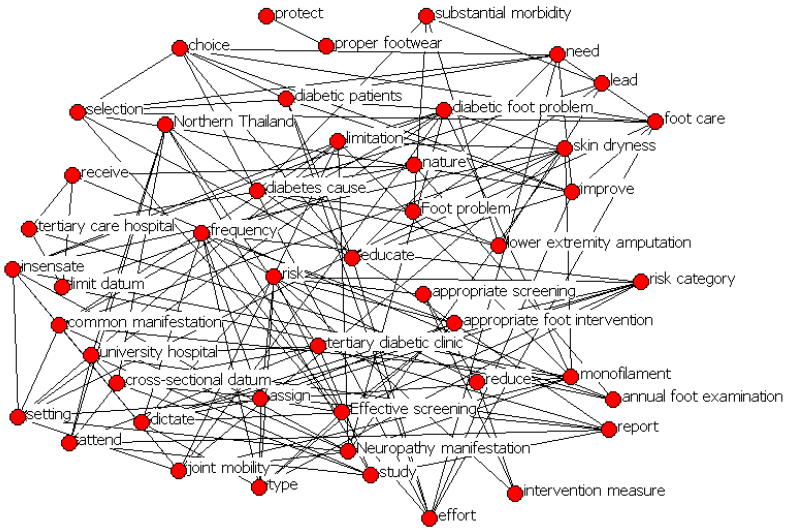


图 1 共现网络

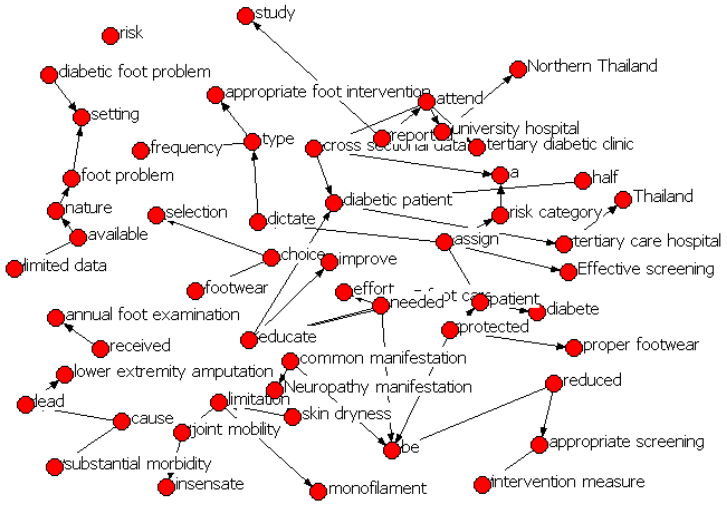


图 2 语法网络

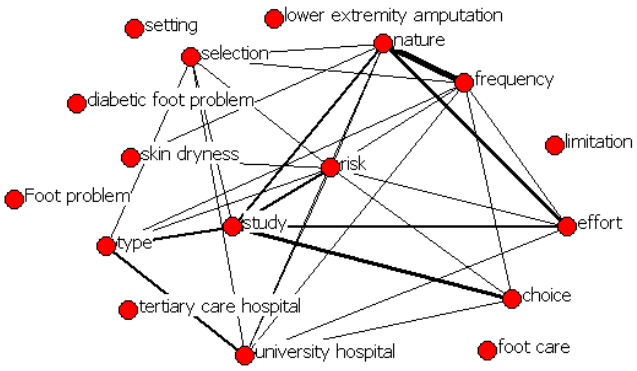


图 3 语义网络

## 2.3 模型构建的主要步骤

通过上述研究及文本语言网络节点和边的类型选择及构建方式可归纳出，基于语言网络的文本表示模型是由节点、边、节点属性及边属性四部分构成。根据文本的多种划分粒度和类型，现有研究中，基于语言网络的文本表示模型的节点可以是句子、实体、概念、事件、主题、短语、词，也可以是一种类型或多种类型的组合。节点属性是对节点的特征描述，包括词性、词频、角色功能及节点单元在文本中不同位置出现的重要性权值等。网络的边通过文本共现关系、语法依赖关系、语义关系构建，边的属性是节点间联系的特征描述，包括边强度、边的权重、方向、具体关系类型等。其中，边的方向通过文本中节点单元出现的次序、语法依赖方向、语义本体定义的上下位等关系来确定，边的权重可以通过共现、各种语义相似、相关度计算（如余弦、Jaccard 等）、权值定义等获得。因此，构建的语言网络可以包括有向、有权、无向、无权四种网络的组合。

基于语言网络的文本表示模型的构建（如图 4），输入为一个文本集或一篇文本，输出为一个具有文本特征的网络图模型。主要步骤可以概括为：

（1）对文本进行预处理，如以句为网络节点，将文本集或文本划分为句。若对节点进行更细划分，在分句后对句子进行词切分，去停用词、取词根、词性标注、句法解析等。

（2）节点识别：根据选择节点类型不同，进行词、短语识别、三元组识别、实体抽取或概念映射等，获取相应节点，并将计算词频、词性、实体类型等存储为节点属性。

（3）边关系构建：根据选择节点类型，生成节点间的边关系，可以通过在特定文本单元长度（如以句为单位长度）进行共现关系计算、进行语法关系解析、实体关系抽取或利用外部词典进行语义关系的识别和语义相似度计算。边的关系类型、方向和权重作为边的属性。

（4）网络合并集成：得到的大量节点和边的关系，如直接构建网络，则获得的密集网络，通常选取特定范围节点，如利用词频排序，选取大于一定阈值的词作为网络最终节点，通过合并相同词根，及词义相同、词形不同的节点，进行语义消歧后的节点合并；边的关系通过合并

相同节点可以得到消减，并通过控制边权重阈值控制，大于某阈值范围的边被除去来构建最终的基于语言网络的文本表示模型。

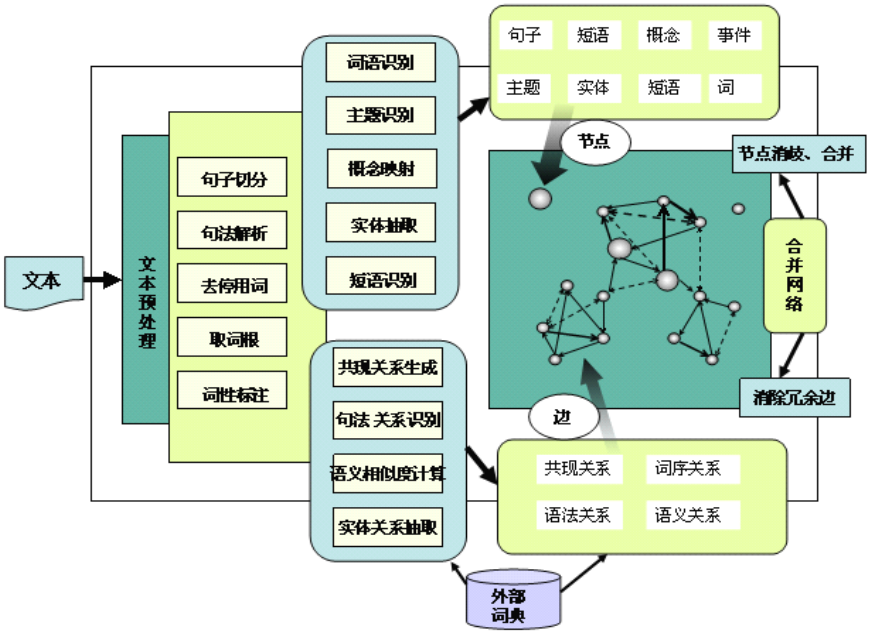


图 4 基于语言网络的文本表示模型的构建

# 3 基于语言网络的文本表示模型的应用研究

文本的网络结构表示比 VSM 等模型能表达更多的语法、语义维度，表达更丰富的文本特征。自然语言网络具有小世界特征，随着复杂网络、图挖掘等大量算法的借鉴和引入，促使自然语言处理领域中的文本表示模型研究和应用逐渐深入。截止到 2012 年，ACL-HLT 已举办 6 届基于网络或图模型的自然语言处理会议，主要宗旨是将网络或图模型用于自然语言处理或计算机语言的研究，包括 NLP 的动态图表示；词典、语义、语法和语音图的属性表示；利用图的算法，基于语言网络进行信息检索、语义消歧、文本摘要、关键词抽取研究等。下面本文对当前基于语言网络的文本表示模型的应用情况进行梳理。

## 3.1 文本分类及聚类

基于语言网络的文本表示模型从多角度揭示文本特征，在文本分类和聚类研究中，主要是利用网络的多特征生成多维特征向量，进行训练或相似特征识别，进行类别划分。

Jiang 等将文本集建模为网络图集合，应用基于文本结构和内容的权重图挖掘算法来抽取最高频子图，用于生成分类向量[16]。网络模型捕捉的文本特征包括(1)词根(2)词性(3)词序(4)上位词(5)句子结构(6)句子切分(7)句子顺序。模型中包含四种节点：①表示句子组成的内在结构的结构节点。②词性节点③token 节点④表示词的附加信息，如它的词根和上位概念的语义节点。包含五

种类型的边：①记录文本结构的边；②连接 token 和其自身的边。③记录词和句子顺序的边。④连接到词根的边。⑤连接到上位概念的边。结果表明，通过对加权子图的挖掘能够提高分类的效果。Yoo 等提出新的文本聚类方法，通过本体将文本映射为网络图。其中，节点表示映射的本体概念，边表示本体中概念层级间的语义关系[17]。方法首先构建每篇文本的图结构而后将其合并为文本集层次的网络图，然后利用无标度聚类算法聚类，通过识别 k 个最高密度的子图捕捉每个类的核心语义关系，最后将每篇文本分配到适合的类下。实验采用来自 MEDLINE 的文本，结果表明该方法能够提高文本聚类的质量，并且能通过每个文本类提高文本的可理解性。Zhang 认为现有用于文本挖掘的文本表示方法缺少集成外部/领域知识，而图模型是有效的表示方式。他结合语义知识、分类知识等进行文本聚类和主题分析，提出新的基于图的表示形式和基于图的聚类方法，应用于生物医学的文本聚类，取得较好效果[18]。

## 3.2 文本摘要

基于语言网络文本表示模型进行文本摘要生成，通常是以句子或对句子逻辑三元组的抽取作为节点，句子间共现关系构建文本表示网络，而后利用分类算法或排序算法选取网络中的中心句作为最终的摘要主题句。

Rusu 提出通过对文本进行深层语法分析的文本摘要方法[19]。他将文本划分为句，从句子中抽取主-谓-宾逻辑三元组，并通过共指消解、语义标准化对三元组进行优化。最后将其合并到创建的语义图模型中。而后使用支持向量机对语义子图进行学习生成文本摘要，分类向量的由语义图模型中节点的属性特征构成，包括节点的出/入度、PageRank 权重、Hub 和 Authority 权重、强弱连接性等。Chen 等提出多文本摘要的新方法[20]。表示模型中，每个节点表示一个句子，边表示句子间的非对称关系。他提出通过向原始图添加一个超级节点（super-vertex）来测量节点子集的重要性。理论上，添加的这个节点不能明显影响原来图中节点的中心度分布。而后应用抽取和近似算法计算超级节点的中心度，最后通过启发规则找到最重要的超级节点作为摘要句。方法基于 DUC 数据进行测试取得较好的效果。

## 3.3 文本关键词抽取

大量研究者将语言网络的文本表示模型应用于文本关键词抽取，在应用中结合了文本网络的小世界特性，利用聚集度、中心度算法或社区探测等算法。

Litvak 将文本表示为网络图模型，利用监督和无监督方法进行关键词识别 [21]。无监督方法使用 Hits 算法计算节点的重要度；监督方法利用提取网络特征包括出度、入度、度、词频、Tfidf、局部分值、高频词分布、标题分值生成统计向量，分别利用 SVM 等分类方法构建二元分类模型，计算节点的重要性。Grineva[15]利用 wiki 语义相似度计算生成文本语义网络，采用 Grivan-Newman 算法进行网络社区探测获得表达文本中多个主题的关键词[22]。生物医学文本呈指数增长，自动标引成为重点，然而现存方法不能很好的识别核心概念。Herskovic 等其将生物医学文本映射为

概念图，而后对概念进行排序识别最重要的概念。他将 MEDRank 结合到医学文本标引系统（Medical Text Indexer，MTI)中，并对集成 MEDRank 方法的 MTI 和未结合该方法的 MTI 进行了对比，结果表明基于图的改进方法具有更好的提高检索效果的作用[21]。

# 4 结语

基于网络的文本表示形式为自然语言处理提供了丰富的语法、语义特征，与 N-gram、VSM 等模型相比，扩充了模型所包含的信息量，具有很多优点：（1）结构化：它是一种结构化的表示方法，通过网络化形式揭示文本单元及单元间的各种特征和联系。（2）特征化：将文本映射为抽象网络结构，将文本内容特征化表示。（3）可视性：语言网络模型便于可视转换，与可视化技术结合，将文本单元的特征和联系显性地表示出来，利于文本理解。（4）灵活性：网络模型定义灵活，节点和边都包含多种类型，能够灵活组合。（5）可扩展性。通过节点、节点属性和边、边的属性可以扩展表达丰富的文本特征及属性。但该模型也具有一定局限性，在构建前需要对节点、边及其属性预先进行明确的定义。一些自然语言特征识别算法，如语法解析、实体识别、语义计算等相对复杂，消耗运算时间并且这些计算的精度直接影响网络模型对文本特征的表达效果。现有研究的语言网络多针对全局和文本的浅层结构关系构建，缺乏局部考虑，缺乏与语言学知识的结合。而且，在特定学科领域当中，语言具有自身特点，构建文本表示模型用于专业领域文本研究，需要结合该领域文本的语言特点，结合大量语言学知识，才能提高语言网络的文本表示模型的可用性和降低模型构建过程中操作的复杂度及提高其表达文本特征的准确程度。

综上，基于语言网络的文本表示模型研究能够有助于自然语言文本理解，并更加快速、准确、全面地获取文本中的重要信息和特征，其在文本处理和挖掘方面具有极其重要的研究和应用意义。今后的研究将进一步结合实际数据对不同层次的语言网络模型的特点、联系与区别进行深入分析；结合特定领域，如在医学领域中结合医学文本特征进行构建方法研究和探索。

**参考文献**

[1]Liu N, Zhang B Y, Yan J et.al. Text Representation:from vector to Tensor[EB/OL]. [2012-8-10]. http://www-connex.lip6.fr/~gallinar/Enseignement/2009-Papiers-ARI/icdm2005-Liu-Tensors.pdf.

[2]Hu J Z, Xiong Z X, Shu J B et.al. A novel method of three dimensional text Representation[C]. Liu F, Reid,R D et.al. eds. International Conference on Management and Service Science (MASS2009). IEEE,2009:1-4.

[3]刘海涛.语言网络：隐喻，还是利器[J].浙江大学学报. 2011,41(2):169-180.

[4]Erkan G, Radev D. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization[J]. Journal of Artificial Intelligence Research .2004 (22):457-479.

[5]Ohsawction Y, Benson N E, Yachida M. KeyGraph:Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor [EB/OL]. [2012-8-10]. http://sclab.yonsei.ac.kr/courses/06mobile/4-1.pdf.

[6]Rusu D, Fortuna B, Mladenic D etal. Visual analysis of documents with semantic graphs[EB/OL]. [2012-8-10].http://www.hiit.fi/vakd09/papers.html.

[7] Xie Z L. Centrality measures in text minging: prediction of noun phrases that appear in abstracts[C]. Proceedings of

the ACL Student Research Workshop 2005, Ann Arbor, Michigan June 27, 2005. ACL2005.

[8] Qu Q, Qiu J G, Sun C Y et.al.Graph-based knowledge representation model and pattern retrieval[C]. FSKD '08. Jinan Shandong,China,October18-20, 2008. IEEE2008.

[9]Grobelinik M, Mladenic.D. Visualization of news articles[EB/OL]. [2012-7-13]. http://kt.ijs.si/dunja/SiKDD2004/Papers/GrobelnikMladenic-Contexter.pdf.

[10] Liu JG, Wang JH.Keyword extraction using language network[C]. NLP-KE 2007. Beijing, China, Aug.30 -Sept.1, 2007. IEEE2007:129-134.

[11] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[EB/OL]. [2012-7-13]. http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf.

[12]Palshikar G. Pattern Recognition and Machine Intelligence[M]. Heidelberg :Springer Berlin. 2007: 503-510

[13] Marne M, D. M C.Stanford typed dependencies manual[EB/OL]. [2012-8-10]. http://nlp.stanford.edu/software/dependencies_manual.pdf.

[14]Huang C, Tian Y H, Zhou Z et.al.Keyphrase Extraction using semantic networks structure analysis[EB/OL]. [2012-8-10]. http://www.jdl.ac.cn/user/chuang/paper/icdm_full.pdf.

[15]Grineva M, Grinev M, Lizorkin D. Extraction key terms from noisy and multi-theme documents[EB/OL]. [2012-7-13].http://www2009.eprints.org/67/1/p661.pdf.

[16]Jiang C T, Coenen F, Sanderson R etal. Text Classification using graph mining-based feature extraction [EB/OL]. [2012-7-13]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.158.2002&rep=rep1&type=pdf.

[17] Yoo I, Hu X H.Scale-Free Network based Clustering using Knowledge-enriched Graph Representation of Biomedical Documents[EB/OL]. [2012-7-13]. http://www.ischool.drexel.edu/faculty/thu/research-papers/Yoo_IS06.pdf.

[18] Zhang X D, Jing L P, Hu X H.et.al. Exploiting External/Domain Knowledge to Enhance Traditional Text Mining Using Graph-based Methods[EB/OL]. [2012-7-13]. http://idea.library.drexel.edu/bitstream/1860/3076/1/Zhang_Xiaodan.pdf.

[20] Chen S Y, Minlie Huang M L, Hu Z Y. Summarizing Documents by Measuring the Importance of a Subset of Vertices within a Graph[C]. Yates R B,Purdue E B .et.al.eds.WI-IAT '09 Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology.Washington,DC: IEEE Computer Society. 2009:269-272.

[21]Litvak M, Last M.Graph-Based Keyword Extraction for Single-Document Summarization[C]. Bandyopadhyay S, Poibeau T .et.al. eds.MMIES '08 Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization. Stroudsburg, PA: ssociation for Computational Linguistics. 2008:17-24.

[22] Herskovic J R, Cohen T, Subramanian D, et.al. MEDRank: using graph-based concept ranking to index biomedical texts[J].International Journal of Medical Informatics.2011:80(6):431-41.

**作者简介**：吴思竹，女，1981 年生，博士，助理研究员，主要研究领域：文本挖掘、信息抽取和领域分析相关研究，现发表论文 9 篇。张智雄，男，1971 年生，研究馆员，主要研究领域：信息系统与智能信息处理，现发表论文 70 多篇。