情报科学
第37卷第11期 2019年11月

面向用户生成内容的本体构建方法

郑姝雅[1],黄　奇[2],张　戈[3],李雨轩[1],陈　雪[1]

(1.南京大学 信息管理学院,江苏 南京 210046; 2.南京大学 国家信息资源管理南京研究基地,江苏 南京 210046;
3.南京大学 工程管理学院,江苏 南京 210046)

摘　要:【目的/意义】本文希望通过自动化本体构建将非结构化的用户生成内容组织成为语义丰富的本体。【方法/过程】综合运用机器学习、自然语言处理等技术,从用户生成内容中抽取本体概念、同义关系及分类关系,形成领域本体,并且通过京东商城用户评论进行实证。【结果/结论】本文实现了手机本体的自动构建,发现该模型能够达到较高准确率,消除了大量冗余,更符合用户需求。

关键字:本体构建;用户生成内容;层次聚类;主题模型

中图分类号:G254　　DOI:10.13833/j.issn.1007-7634.2019.11.007

A Ontology Construction Method for User Generated Content

ZHENG Shu-ya[1],*HUANG Qi*[2],*ZHANG Ge*[3],*LI Yu-xuan*[1],*CHEN Xue*[1]

(1. *School of Information Management,Nanjing University,Nanjing* 210046,*China*;
2. *Nanjing Research Base of National Information Management,Nanjing University,Nanjing* 210046,*China*;
3. *School of Management & Engineering, Nanjing University,Nanjing* 210046,*China*)

Abstract: 【Purpose/significance】This paper is aimed at organizing unstructured user-generated content into rich-semantic ontology through automated ontology construction. 【Method/process】This paper comprehensively uses machine learning, natural language processing and other techniques to extract ontology concepts, synonymous relationships and classification relationships from user-generated content, in order to form domain ontology, and conduct empirical research by Jingdong Mall user reviews. 【Result/conclusion】This paper implements the automatic construction of the mobile phone ontology, and finds that the model can achieve high accuracy, eliminate a lot of redundancy, and more meet user needs.

Keywords: ontology construction;user generated content;hierarchical clustering;topic model

1 引　言

随着技术和观念的革新,互联网服务也不断的更新迭代,由最初的网站向用户单向传输的模式转变为如今的用户参与的大众互联网时代,并且正在向以Berners-Lee提出的语义网为核心的组织模式过渡。

用户主导的互联网服务模式产生了大量的用户生成内容,这一类特殊的组织形式成为很多学者、企业研究利用的重要资源。比如,利用股评文章可以研究它对于股票市场的影响,利用消费者生成内容可以挖掘消费者行为、社交媒体、销售情况和公司策略四者的有机联系。但是,由于人人可以参与、人人可以创作的特性,网络上用户生成内容的数据量是巨大的,质量参差不齐,并且多为难以处理的非结构化数据。有学者提出采用主题模型从语义层面来分辨高质量的用户生成内容,从众多书评中提炼出有价值的信息[1]。

本体是一种形式化的知识表示方法,并且可以通过机器推理查询,在检索、推荐等研究中起到了非常重要的作用,研究还发现利用领域本体可以减少信息不对称性[2]。由于本体的特性,很多学者将本体引入到相关领域的知识组织中[3]。但传统方法构建的本体对于用户来说含有大量冗余概念,且词典、词表等工具对于流通知识的组织通常存在滞后的情况,无法满足社会化媒体用户的需要[4],这要求我们对高度专业化、规范化的传统信息组织进行改进和创新。

收稿日期:2019-04-21

基金项目:国家社会科学基金重点项目"基于知识组织的产品分类本体研究"(13ATQ005)

作者简介:郑姝雅(1995-),女,研究生,主要从事网络信息资源的语义化、电子商务研究;通讯作者:黄　奇.

本文借鉴传统的本体构建方法，面向用户生成内容提出了一种中文本体自动构建方法。通过LDA（潜在狄利克雷分布）主题模型、Word2vec、层次聚类等技术自动化地抽取概念及概念间关系，以期将社交媒体平台中用户发布的文本数据组织成为结构化的本体，为商业生产和科学研究提供更多支持和新的思路。

2 研究基础

目前的本体构建方法主要面向三大类信息源，包括结构化、半结构化以及非结构化的信息源。

结构化的信息源主要包括词表和关系数据库。郭朝敏[5]提出从数据库中抽取关系模式并写入XML（可扩展标记语言）格式的关系模式文件中，他们在研究中将本体构建分为语义信息的发现和本体映射两部分自动构建初始本体，然后补充领域规则来最终实现目标本体的构建。陆佳莹[6]等人基于eCl@ss产品分类利用形式概念格方法进行本体构建。

半结构化信息主要指的是XML、HTML（超文本标记语言）网页等具有一定的结构性，但缺乏固定或严格结构的信息。如Rung-Ching Chen[7]根据领域相关的网页中词汇的TF-IDF值得到领域关键词，并且利用自适应谐振网络对关键词进行聚类，通过贝叶斯网络来挖掘概念间的层次关系，自动化的构建领域本体。Xu[8]等人提出将XML文档首先映射为实体联系模型，然后将实体联系模型映射为OWL本体，实现本体构建。

面向非结构化信息源的本体构建方法大多基于自然语言处理技术。李志义[9]基于信息熵、互信息以及TF-IDF（term frequency-inverse document frequency）等统计信息完成领域概念的自动抽取，然后基于向量空间模型和关联规则的方法分别实现分类关系和非分类关系的提取，实现了电子商务领域本体的自动构建。Azevedo[10]等人利用斯坦福大学的NLP（Natural Language Processing，自然语言处理）工具对文本进行句法分析，根据得到的依存关系对文本中的所有名词及形容词进行合并，并且基于语义角色标注完成本体工程中的概念抽取和关系抽取任务，实现半自动化的本体构建。

网络用户的信息需求呈现出多元化的特征[11]，面向用户需求越来越多学者将研究的目光放到UGC（用户生成内容）上来。张云中[12]等人利用社会化标签对电影资源进行标注，通过建立电影标签集和资源集向本体关系进行映射，实现电影资源的本体构建。但是目前面向UGC信息源的本体构建方法大多仍基于社会化标签这类半结构化信息上，对于非结构化的UGC信息的研究还较少。本文将探讨面向UGC文本的领域本体自动化构建。

3 模型提出

用户生成内容指的是Web 2.0环境下产生的一种网络信息资源创作与组织模式，泛指任何形式在网络上发表的由用户创作的文字、图片、音频和视频等内容[13]。本文对用户生成内容中的文本类数据进行处理，基于此类数据进行中文本体的自动化构建。

前人的研究将本体构建任务划分为预处理模块、术语抽取模块、概念抽取模块、关系抽取模块以及形成本体模块。借鉴前人的研究，本文综合运用自然语言处理技术、机器学习技术，结合统计学、语言学相关知识，从用户生成内容中自动地抽取领域概念和概念间的关系，设计了一种领域本体自动构建方法（见图1）。具体过程如下：

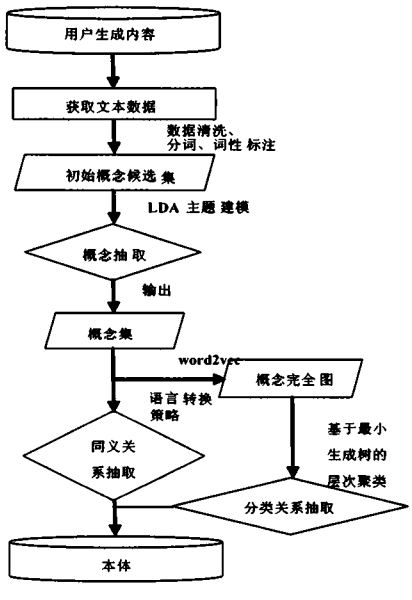

图1 面向用户生成内容的本体构建工作流程图

①收集某领域用户生成内容，对数据进行清洗，利用分词工具对文本数据进行分词以及词性标注；

②利用LDA主题模型从文本中抽取领域相关概念；

③基于语言转换策略实现同义关系的抽取；

④基于word2vec和最小生成树算法对概念进行层次聚类，挖掘分类关系，并基于句法模式识别其中的部分整体关系；

⑤将抽取出的概念和关系进行整合，形成领域本体。

3.1 数据预处理

用户生成内容的文本多是大众即时产生的想法、心情、生活场景的分享，含大量口语化表达，因此数据预处理是一项重要工作。

和英文文本最大的不同，中文文本的内容是连贯的，因此利用中文文本时，首先需要进行分词处理。Jieba中文分词工具包功能强大，完美支持中文分词，能够在Python平台调用该模块实现分词、词性标注等功能。Jieba中文分词模块提供了三种不同的分词模式：精确模式、全模式以及搜索引擎模式。其中，精确模式适合用于文本分析，提供最精准的分词；全模式能够识别出整句中所有可以成词的词语，例如“手机运行速度很高”，精确模式下得到的结果是“手机 运行 速度 很高”，全模式下输出的结果是“手机 运行 速度 运行速度 很 高 很高”，可以发现全模式下能够保留所有的词

语和词组，但是不能解决歧义问题；搜索引擎模式对精准模式的分词结果中所有的长词再次切分。根据三种分词模式的特点，本文选择精确模式分词，以得到完整准确的领域概念。

本文最终目的是构建领域本体，本体中的概念一般为名词或者名词性词组，因此可以通过词性标注筛选文本中的名词。同时，并不是所有的名词都可以作为本体中的概念，例如利用产品评论构建商品领域本体时，人名、地名等均不属于领域相关概念，这一类名词则可以利用二级词性标注进行剔除。此外，还可以自定义停用词表对名词集中的非领域概念进一步剔除。

对文本进行清洗之后，再利用Jieba工具包的分词功能和词性标记功能，即可得到初始候选概念集。

3.2 概念抽取

领域概念是用于表达领域主题的核心概念。对于领域概念的抽取，一部分学者直接将基于TF-IDF、句法分析等方法提取的领域术语作为领域概念，或者在领域术语的基础上进行聚类形成领域概念。还有一部分研究基于WordNet等现有资源直接从文本中抽取领域概念。

一般来说，领域概念是那些在领域相关数据中分布范围较广，使用频次较高，并且在非领域相关数据中使用较少的术语[14]。LDA主题模型是一个概率主题模型，用于识别大型文档或语料库中的潜在主题信息，模型的基本假设如下：同一主题下相同主题的词汇出现的概率往往比较相近；不同主题分布下，词汇的概率分布特征也不同。可以发现从用户生成内容中抽取领域概念符合LDA训练文本的特点，并且研究证明LDA模型算法性能良好，具有较好的领域移植性[15]。因此，本文选择LDA模型完成用户生成内容的概念抽取任务。

笔者对经过清洗、分词、词性标注后得到的初始名词集合进行LDA主题模型训练，并通过词频统计来进一步筛选得到最终的概念集合。概念抽取任务的详细步骤如下：

Step1:将文本所对应的候选概念集合进行TF(词频，term frequency)转换，得到每个词在初始候选概念集合中的权重；

Step2:训练LDA主题模型，现有的工具如Java、Python等均带有LDA的库或者包，可以直接调用。在应用LDA主题模型训练语料时，需要确定四个参数：LDA模型超参数α和β，主题数K，特征值的候选阈值ε。其中，α和β根据经验值分别以50/K和0.01为最佳[16]，通过实验证明这种取值效果最好，对于主题数根据样本量划定一定的范围，特征值的候选阈值取值范围为0-1之间，然后根据模型困惑度的大小，确定最优参数；

Step3:利用训练好的模型进行求解，得到用户生成内容的主题分布，将所有的主题词构成候选特征集合文档；

Step4:将初始候选集进行TF-IDF转换，对上一步骤所得到的候选特征集合文档按TF-IDF排序，筛选高频率的主题词，得到最终的本体概念集。

通过LDA主题模型将用户生成内容中最核心的主题概念抽取出来，接下来需要识别概念之间的关系，以最终形成层次结构。

3.3 关系抽取

概念间关系的抽取是本体构建任务中最重要的一个步骤，它决定了最终的本体结构。一般来说，概念间的关系分为分类关系和非分类关系两类。对于概念间分类关系的抽取目前的主流技术有基于模板的方法、基于概念聚类的方法、基于词典的方法以及混合方法。针对非分类关系的研究相对较少，因为概念间的非分类关系太过复杂，目前研究主要依赖于句法结构和依存关系分析。由于非分类关系类别相对分类关系更为隐蔽，并且难以明确关系的具体类别，目前的技术还不能够做到完全自动化的准确识别所有的非分类关系，因此本文只关注于分类关系以及同义关系的抽取。

3.3.1 同义关系抽取

在WordNet中，同义关系描述为同义词关系，因此同义关系的抽取可以等价为同义词的识别。

学者提出基于双语词典来识别词语的同义关系是可行方法[17]，因此本文基于语言转换策略，以英汉词典作为知识库，实现同义关系抽取。详细来说，对于中文概念C1，对其进行英文转换得到单词或者短语集合W1={w1,w2,…,wn}，同时对概念C2也进行英译得到单词或者短语集合W2={w'1,w'2,w'm}，如果概念$W1 \cap W2 \neq \varnothing$，则概念C1和C2是同义词，在本文表示为同义关系sy。

3.3.2 分类关系抽取

本节所探讨的分类关系包括上下位类关系和部分整体关系。对于概念间的分类关系的抽取，可以利用语言学规则或模式构建出匹配模式，然后通过模式匹配实现，但由于语言的复杂性，这种方法的可扩展性不高。也可以由专家或者根据背景知识预定义一些关系模板来进行分类关系的抽取，但准确性取决于专家的知识背景。

为了弥补上述两种方法的不足，本文结合基于word2vec的层次聚类方法实现分类关系的抽取，并采用基于句法模式对层次聚类的结果进行进一步的细分。层次聚类方法是一种无监督学习，应用更灵活，同时增加一部分简单的句法模式匹配对文本进行进一步的深层次分析，能够获得更高性能的本体知识[18]。word2vec模型能够将每个词映射为一个向量，并且将词与词之间的语义关系通过向量间的距离来反映。因此以word2vec方法训练的距离作为测度标准进行层次聚类能够很好的将概念集按语义相似度进行划分，得到准确的分类结果。但是由于层次聚类方法在每次合并时都需要从更新的距离矩阵中找出最小值，算法复杂度很高，因而笔者引入最小生成树算法，则可以大大降低算法的时间复杂度。

综合上述，本文将分类关系抽取分为两大步，第一步首先利用基于最小生成树改进的层次聚类算法识别出包括上

下位类关系和部分整体关系在内的所有分类关系，详细步骤如下：

Step1:将2.2节得到的本体概念作为完全图G(V,E)的顶点V，其中如果概念C1和概念C2已经确定为同义关系，那么两个概念标记为同一个顶点vi；

Step2:利用gensim工具包的word2vec模块计算概念对应的词向量之间的相似度，将其倒数表示为概念顶点之间边的权值，概念间的相似度越大，权值越小；

Step3:基于Kruskal算法得到图G的最小生成树T(V,E)；

Step4:基于最小生成树对概念进行层次聚类，将语义相近的概念聚集到一个簇中，即可得到概念间的分类层次关系。

依靠上述算法得到概念间分类关系之后，第二步基于句法模式进一步识别出其中的部分整体关系。

部分整体关系的句法模式有很多，笔者对其进行梳理，整理如表1所示。其中，A表示整体，B表示部分，部分词指代“配件”、“部件”等表示部分整体关系的词，匹配时由(配件|部件|零件|元件|组件|构件|器件)替换。

表1 分整体关系的句法模式总结

模式	示例
A由B等(构成|组成)	电脑由主机、显示屏、键盘等构成。
A的B[部分词]等	这对相机的镜头要求很高。
B等A[部分词]	耳机、充电器、电池等手机配件。

利用给定的句法模式，笔者从已获取的分类关系中得到部分整体关系，具体步骤如下：

Step1:提取出分类关系中所有的概念对，表示为<A,B>，A为高层次概念，B为下层概念，给定概念集合C，由所有的概念A组成；

Step2:在语料库中对C中的概念A根据句法模式进行模式匹配，查找所有符合句法模式的概念D，得到候选概念对<A,D>；

Step3:对Step1得到的所有概念对和Step2得到的所有概念对取交集，即得到所有有部分整体关系的概念对。

基于上述步骤笔者得到存在部分整体关系的概念对，根据前文对分类关系的定义，则第一步中得到的分类关系中的剩余概念对之间为上下位类关系。

4 实验及结果分析

商品评论是一种领域性很强的用户生成内容，所有的商品评论都是围绕某一特定产品展开的描述。因此笔者选择京东商城平台的用户评论作为语料对模型进行实证。

4.1 实验数据

笔者选择京东商城上荣耀10、iPhone x、小米6三款手机的评论数据作为实验数据，利用八爪鱼爬虫软件共爬取10 744条评论。对评论数据进行清洗，将诸如“好评！”“用过一段时间再来评价”等无用文本以及重复评论数据删除后，共有评论9289条。

使用Jieba分词工具包进行数据预处理时，需要特别注意文本文件的编码格式，混乱的编码模式会导致模块运行出错。因此，笔者将文本数据全部转换为utf-8编码的文本文档，再进行后续操作。

利用上述评论数据，本实验拟利用所提出的模型构建一个手机产品特征本体，以证明方法的可行性。

4.2 实验结果

将实验数据输入模型，由于产品评论文本口语化表达严重，在分词时会遇到一些问题从而影响后续环节的效果，因此在分词时，笔者加载了自定义用户词典以提高结果的精度。

利用LDA主题模型进行概念抽取时，主题数是非常重要的超参数。笔者根据困惑度的值，选择了最小困惑度时的主题数20来训练模型。表2展示了提取出的部分概念。

表2 概念抽取结果(部分)

概念	概念	概念	概念
手机	手感	信号	指纹识别
屏幕	性价比	电量	品牌
耳机	摄像头	处理器	价格
速度	画面	边框	功能
外观	声音	人脸识别	发货速度
系统	内存	尺寸	颜色
电池	音质	运行速度	质量

接下来对提取出的领域概念基于语言转换策略识别出其中的同义词，以同义关系sy标记概念词对。本文利用有道翻译词典，对所有的概念进行英文翻译，最终得到概念间同义关系词对如表3所示。

表3 概念间同义关系(部分)

概念间同义关系
价钱 价格 sy
价钱 价位 sy
质量 品质 sy
色彩 颜色 sy
外观 外形 sy
外观 样子 sy

在识别出所有同义词的基础上，将语料中同义概念进行替换统一，利用gensim工具包的word2vec算法训练模型，将所有概念对之间的相似度的倒数作为权值，构建概念完全图。然后，基于克鲁斯卡尔最小生成树算法进行层次聚类抽取概念间分类关系。在得到分类层次结构的基础上，笔者利用正则表达式根据句法模式进一步识别部分整体关系。提取出的分类关系结果如表4所示。其中，准确率是指校验后的分类关系数和抽取的所有分类关系的比率。

表4 分类结果统计

分类关系总数	校验后分类关系总数	准确率
101	66	65.35%

最终利用protégé工具基于自动抽取的概念及概念间关

系形成手机产品本体。笔者在protégé中添加了对象属性isPartOf表示概念间的部分整体关系，添加对象属性Sy表示概念间的同义关系。得到最终的手机产品本体在图2(部分)展示。

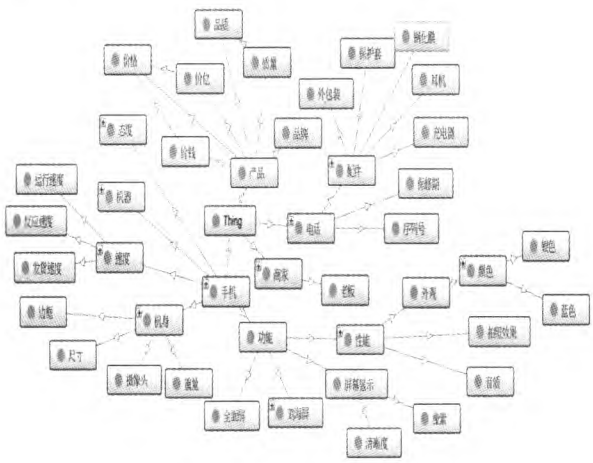

图2 手机产品本体(部分)

通过实验发现，利用本文模型能够有效构建领域本体，并且完全自动化的过程省去了大量的人工工作。实验得到的手机产品本体可以应用于手机产品推荐服务，基于知识本体的推荐方法能够很好的解决传统推荐技术会遇到的"冷启动"、数据稀疏性等问题。本文模型还可以应用于其他社交媒体平台，如利用简书特定专题用户分享笔记可以构建该专题领域知识本体，利用领域知识本体可以设计在线问答系统，构建学术知识地图[19]，也可提高搜索引擎检索准确率。

5 结 语

本文面向用户生成内容，研究了以文本作为数据源的本体构建中概念自动抽取和关系抽取两个任务，提出了本体自动化构建的方法。研究涉及了自然语言处理、主题模型、层次聚类等方法。基于模型笔者以消费者生成内容为实验数据进行了实证，实验验证了本文模型大大节省了本体开发过程中的人工参与，缩短了本体构建时间，并且关系抽取精度能够达到较高水平。利用本文模型，能够将丰富的用户生成内容自动化构建为领域本体，降低冗余，为相关领域的研究提供帮助。

未来研究的方向可以针对不同领域不同来源的中文文本进行对比研究，探索模型对于不同场景下的文本内容构建效果是否一致，进而更好的完善模型。

参考文献

1 阮光册, 夏 磊. 高质量用户生成内容主题分布特征研究[J]. 图书馆杂志, 2018,(4):95-101.
2 李雨轩,黄 奇,陈 雪,郑姝雅,张 戈.利用领域本体提高信息对称性的研究[J].情报学报,2018,37(7):678-685.
3 朱 光, 杨嘉韵, 吴先华,等. 基于FCA的气象灾害领域术语层次关系分析和本体构建研究[J]. 现代情报, 2017, 37(5):79-88.
4 胡 华. 基于中文UGC信息源的半自动应用本体构建研究[D]. 武汉:武汉大学, 2014.
5 郭朝敏,姜丽红,蔡鸿明.一种关系数据库到本体的自动构建方法[J].计算机工程与应用,2012,48(7):115-120,248.
6 陆佳莹,袁勤俭,黄 奇,钱韵洁.基于概念格理论的产品领域本体构建研究[J].现代图书情报技术,2016,(5):38-46.
7 Chen R C, Liang J Y, Pan R H. Using recursive ART network to construction domain ontology based on term frequency and inverse document frequency[J]. Expert Systems with Applications, 2008, 34(1):488-501.
8 Xu J, Li W. Using Relational Database to Build OWL Ontology from XML Data Sources[C]// International Conference on Computational Intelligence and Security Workshops. IEEE, 2007:124-127.
9 李志义,李德惠,赵鹏武.电子商务领域本体概念及概念间关系的自动抽取研究[J].情报科学,2018,36(7):85-90.
10 Azevedo R R D, Freitas F, Rocha R G C, et al. An Approach for Learning and Construction of Expressive Ontology from Text in Natural Language[C]// Ieee/wic/acm International Joint Conferences on Web Intelligence. IEEE, 2014:149-156.
11 魏 敏.信息组织4.0:变革历程和未来图景[J].国家图书馆学刊,2018,27(1):78-85.
12 张云中, 李佳佳. 基于社会化标签的电影资源本体构建研究[J]. 图书情报工作, 2016, 60(12):130-138.
13 赵宇翔,范 哲,朱庆华.用户生成内容(UGC)概念解析及研究进展[J].中国图书馆学报,2012,38(5):68-81.
14 刘柏嵩. 中文领域本体自动构建理论与应用研究[M]. 杭州:浙江大学出版社, 2014.
15 马柏樟,颜志军.基于潜在狄利特雷分布模型的网络评论产品特征抽取方法[J].计算机集成制造系统,2014,20(1):96-103.
16 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J].Journal of Machine Learning Research,2012,(3):993-1022.
17 宋培彦,李静静,赵 星.跨语言术语同义关系推荐方法及其实证[J].现代图书情报技术,2013,(5):40-45.
18 任飞亮,沈继坤,孙宾宾,朱靖波.从文本中构建领域本体技术综述[EB/OL].http://kns.cnki.net/kcms/detail/11.1826.TP.20170506.1214.010.html,2018-10-11.
19 刘晓燕,王 晶,单晓红.基于本体的学术知识地图构建——以国内动态能力研究为例[J].情报理论与实践,2017,40(7):122-126.

(责任编辑:徐 波)