文章编号：2096-1472(2018)-10-14-03
DOI:10.19644/j.cnki.issn2096-1472.2018.10.004

基于社交网络的小说聚类

楼锴毅，霸元婕，李绍昂
(吉林大学计算机科学与技术系，吉林 长春 130012)

摘　要：目前小说的受众群体越来越大，其中蕴含着巨大的商业价值。文本聚类的研究领域也在突飞猛进，但对于其中的现实领域：小说聚类，相关的研究却较少。本文研究了一种基于小说中的社交网络对其进行聚类的方法。该方法首先提取出小说中的社交网络，在得到网络的特征向量后，基于其进行聚类，并将结果与依据小说作者的划分进行对比。实验结果表明，该方法可以在一定程度上反映出不同作者写作风格的不同，效果可以接受，并拥有进一步提升的可能。

关键词：小说；社交网络；聚类算法
中图分类号：TP391.1　**文献标识码**：A

Clustering of Novels Based on Social Network

LOU Kaiyi,BA Yuanjie,LI Shaoang
(*School of Computer Science*,*Jilin University*,*Changchun* 130012,*China*)

Abstract: At present,more and more people are reading novels,which contains great commercial value.The research field of text clustering is also advancing by leaps and bounds,but for the real practice—novel clustering,there are few related researches.This paper uses a method based on social network in the novel to cluster it.The method first extracts the social network in the novel.After obtaining the feature vector of the network,it clusters based on it and compares the result with the division according to the author of the novel.The experimental result shows that the method can reflect the different writing styles of different authors to a certain extent,the effect is acceptable,and further improvement is possible.

Keywords:novels;social network;clustering algorithm

1 引言(Introduction)

随着第三产业的发展，移动互联网时代的到来，文娱产业对人们日常生活的影响越来越大，特别是近几年小说的受众群体越来越大，因此基于小说的各种文学定量分析越来越成为重要的课题。与此同时，以机器学习和统计方法为基础，各种各样的文本分类技术也在飞速发展。特别是在近几年来，基于CNN、RNN等深度神经网络的相关方法取得了很好的结果，因此人们对文学分析定量方法的研究兴趣也日益增加[1,2]。

小说的关键维度包括形式、结构、人物、情节等。目前来讲，人们对其的定量研究大多集中在形式和内容上。而对于小说中的情节、结构、人物关系等，由于其量化较为复杂，而少有研究[3]。在本文中，我们实现了基于小说中的社交网络的聚类。我们首先提取出小说中的社交网络，之后得到其特征向量并根据其进行聚类。因此，聚类的结果也是根据小说结构进行的分组，通过与小说作者的对比，我们也能得到小说社交网络与小说的风格流派和作者风格特征的联系程度。

2 相关工作(Related work)

2.1 文本分类

文本分类的相关研究可以追溯到20世纪50年代，而到目前它已经成为了NLP领域的经典问题，一直到现在都是人们研究的热点。而其算法的发展，也伴随着人工智能研究领域的发展而不断地更新。在20世纪，文本分类往往基于规则和语料库，其虽有准确率高等优点，但是耗费资源过多、可移植性很差。到了20世纪90年代的时候，人工智能的研究领域开始向基于统计和数据驱动的方法过度，与此同时基于特征工程和各种分类器的文本分类方法也开始逐渐兴起。

但是传统分类方法依然存在着诸多不足，比如特征表达能力较弱，成本较高，等等。近年来，随着深度学习的发展，基于其的一些方法也开始被应用到了文本分类的领域。

深度学习解决文本分类问题，一般都是先解决文本表示，之后利用CNN、RNN等自动获取特征表达能力，从而端到端的解决问题[4]。

2.2 文学计算分析

自从计算机诞生之后，人们便一直尝试将其算法应用到文学分析的领域，即文学的计算分析。这种方法往往用定量的方式，基于文本的语言结构特征对文章的风格进行刻画。因此，这种研究方法最重要的就是两个问题：语言特征的选择和研究方法的选择。不过一般来讲，大多数方法利用的都是基于主题和内容的特性。然而对于一部小说来讲，我们不应该只从标点、词法、句法、语义的维度进行分析。这种文学形式还有情节、人物、叙事结构，等等。可以说每一部小说都是一个社会的缩影[5]。

因此，人们也逐渐开始关注量化情节的方法，以及人物对情节的影响。特别是可以将小说刻画成社交网络，并通过其研究小说中的情节结构。目前，通过提取复杂网络并基于其分析文本已经成为了一个十分重要的学术流派。人们的研究表明，通过提取小说中的人物关系网络来分析小说中社会结构、意义和作者观点是完全可行的。

3 网络的构建(Network construction)

3.1 人物的自动识别

社交网络起源于网络社交，目前可以理解为一种形容人际关系的网络结构，其本身作为一种复杂网络，可以反映出网络中点与点之间的联系。而在小说中，每个人物正是社交网络中的结点，人物与人物之间的关系为社交网络的边。因此，一般将其分为四个步骤：人物角色标记、角色指代消解、人物关系识别与网络关系表示[6]。在人物角色标记中，需要识别出所有表示人的单词；在角色指代消解中，需要将代词或者非人名的词替换为其指代的人名；在人物关系识别中，需要识别并提取人与人之间的关系；关系网络表示则是将网络用数学模型表示出来。

在人物自动识别这一步骤中，需要解决的问题是人物关系识别和指代消解，其也被称为命名实体识别。而对于这类问题，笼统地可以分为三种解决的方法：基于规则的方法、基于统计的方法和近年来兴起的基于深度学习的方法。基于规则的方法一般由语言学专家手工构造规则模版，因此存在代价大、移植性差等缺点，目前只有在数据量小或者非常特殊的场合才会使用。基于统计的方法有：隐马尔科夫模型、较大熵模型、支持向量机、条件随机场等，这类方法一般对语料库的依赖较大。近年来随着深度学习的发展，人们也将其应用到了命名实体识别中，一般方法为将NN、CNN、RNN与条件随机场结合[7,8]。本文采用的方法是条件随机场，采用开源工具CRF++。

条件随机场，一般简称为CRF，由于其具备长距离依赖性和交叠性能力，是目前一种非常常用的用于命名实体识别的，判别式的概率图模型[9]。定义无向图G=(V,E)，单词序列$x=(x_1, x_2, \cdots, x_n)$，每个单词x_i有对应的实体类型标记$y_i$，标记序列集合$y=\{y_i\}$。则节点集合V为单词或其对应的实体标记类型，边集合E表示单词对应节点与该单词实体标记类型对应节点间的连线，于是(x，y)构成一个条件随机场。由于链式结构为最简单的结构和建模方式，因此人们一般采用的是线性链条件随机场，如图1所示。

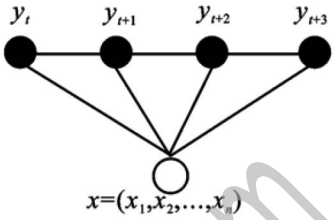

图1 线性链条件随机场

Fig.1 Linear-chain CRF

在线性链条件随机场中，给定观察序列x，特定序列y概率为P(y|x)，则有

$$P(y \mid x) = \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right)$$

其中，t_k为转移函数，s_l为状态函数，λ_k与μ_l分别表示两个函数的权重。

转移函数和状态函数均为特征函数，一般取0或1，即满足特征函数的为1，否则为0。若将它们统一用特征函数的形式来表示，再加上归一化的过程，则可以得到最终条件随机场的条件概率公式为

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{k=1}^{K} \lambda_k F_k(y, x)\right)$$

其中，

$$F_k(y, x) = \exp\left(\sum_{i=1}^{n} f_k(y_{i-1}, y_i, x, i)\right)$$

$$Z(x) = \sum_{y} \exp\left(\sum_{k=1}^{K} \lambda_k F_k(y, x)\right)$$

对于其中的参数，一般采用极大似然法进行估计，并采用迭代技术来确定参数。

3.2 网络的构建

对于小说中的人物关系识别，一般有两种方法：基于人物对话的方法和基于人物共现的方法[6]。第一种方法为只考虑小说中的对话，即如果两个角色有语言或者对话的交互，就将两个角色进行关联，得到的网络为人物对话网络。这种方法为目前大多数文献所采用，尤其是对于剧本这种只通过对话来进行角色间互动的文本，该方法十分有效。但是，对于大多数小说，人物间的许多互动都是通过叙述者的描述或者间接的互动来完成的。这时我们就应该考虑第二种方法，即通过人物间的共现关系来构建网络，每当两个角色出现在同一个文本窗口或者语境下时，将二者进行关联，得到的网络为人物共现网络。在本文中，我们使用Python库Networkx来构建网络，并将其存储在表示人物关系的邻接矩阵中。

4 计算与聚类(Calculation and clustering)

4.1 特征选择

我们可以把特征分为两种。第一种为网络的拓扑特征，其指标有度分布、集聚系数、网络特征路径长度、直径、主节点的相关性等。但是对于小说而言，其更像一个小型的社会，所以我们还应该考虑社会指标。一些常考虑的社会指标包括男性角色比例、视角的比例，等等。通过这些特征，我们可以分析出小说中社交网络的结构特性，并根据其进行聚类[5]。

4.2 聚类

本文采用k-means算法进行聚类，它是目前最简单的聚类算法之一，也是应用最广泛的一种聚类算法。其具体过程可以分为四步：选择k个初始聚类中心，根据对象与中心的距离对其重新划分，计算更新后的均值，迭代至测度函数收敛。在算法中，k值即为数据集中作者的个数，初始聚类中心为数据集中随机选择的k个值，距离将采用余弦距离，即通过向量空间中两个向量夹角的余弦值来衡量个体间差异的大小，公式为

$$\cos(\theta)=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}}$$

5 实验(Experiment)

5.1 实验语料

本文共选取了65篇小说作为语料，分别来自作家卡夫卡、张爱玲、老舍、狄更斯，数据集见表1。

表1 数据集

Tab.1 Data set

类型	数量
卡夫卡	20
老舍	20
张爱玲	15
狄更斯	10

5.2 评测指标

本文主要采用常见的三种指标：准确率、召回率与F_1值。

准确率：$P=\frac{A}{B}$

召回率：$R=\frac{A}{C}$

F_1值：$F-measure=\frac{2PR}{P+R}$

其中，A表示正确识别的相关小说数，B表示识别的小说数，C表示相关的小说总数。

5.3 实验结果

由于我们的语料分别来自四位作家，因此在我们的聚类算法中，将k设为4。最终算法将会把所有的小说分为四类，我们以每类含有的最多的小说作者作为该类的标签，并以此作为评价的基准。我们将计算准确率、召回率、F_1值，并将其作为评价我们算法的依据。实验结果如表2所示。

表2 实验结果

Tab.2 Experimental results

类型	准确率	召回率	F_1值
卡夫卡	0.71	0.60	0.65
老舍	0.50	0.45	0.47
张爱玲	0.48	0.50	0.49
狄更斯	0.57	0.80	0.67

6 结论(Conclusion)

目前的文学定量分析方法大多是基于文本的形式和内容，对于结构、情节、人物关系等的量化与分析方法较少。在本文中，我们基于小说本身就是一个小型社会的特点，研究了基于社交网络对小说进行聚类的方法。在实验中，我们发现小说的社交网络能够在一定程度上反映出小说的风格流派及作者的风格特征。该方法具备一定的实用性，并且有进一步提升的可能。

参考文献(References)

[1] Abualigah L M,Khader A T,Al-Betar M A.Unsupervised feature selection technique based on harmony search algorithm for improving the text clustering[C].International Conference on Computer Science and Information Technology,IEEE,2016:1-6.

[2] Scrivner O,Davis J.Interactive Text Mining Suite: Data Visualization for Literary Studies[C].Corpora in the Digital Humanities,2017.

[3] Jarynowski A,Boland S.Social Networks Analysis in Discovering the Narrative Structure of Literary Fiction[J].Biuletyn Instytutu Systemow Informatycznych,2013,12(2):35-42.

[4]Ji Y L,Dernoncourt F.Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks[C]. North American Chapter of the Association for Computational Linguistics,2016:515-520.

[5] Ardanuy M C,Sporleder C.Structure-based Clustering of Novels[C].The Workshop on Computational Linguistics for Literature,2014:31-39.

[6] 刘海燕,尹晓虎.文学作品中的“小世界”——菲茨杰拉德小说人物关系网络的实证分析[J].统计与信息论坛,2015,30(12):102-107.

[7] Chen L C,Papandreou G,Kokkinos I,et al.Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J].Computer Science,2015(4):357-361.

[8] Ritter A,Clark S,Etzioni O.Named entity recognition in tweets:an experimental study[J].Emnlp,2011,61(3):1524-1534.

[9] Lafferty J D,Mccallum A,Pereira F C N.Conditional Random Fields:Probabilistic Models for Segmenting and Labeling Sequence Data[C].Eighteenth International Conference on Machine Learning.Morgan Kaufmann Publishers Inc.2001:282-289.

作者简介：

楼锴毅(1996-)，女，本科生.研究领域：数据挖掘.

霸元婕(1997-)，女，本科生.研究领域：数据挖掘.

李绍昂(1997-)，男，本科生.研究领域：数据挖掘.