doi:10.3969/j.issn.1672 - 4348.2015.04.014

基于词性标注的文本聚类算法

王金水，唐郑熠，薛醒思

（福建工程学院 信息科学与工程学院，福建 福州 350118）

摘要：针对传统的文本聚类容易受到噪声影响的问题，提出一个基于词性标注的文本聚类算法。该算法利用词性标注从文本中识别并抽取最能体现文本特征的关键词，再基于所抽取的关键词进行聚类操作。实验发现，相对传统的聚类算法，基于词性标注的文本聚类算法能够有效地提高聚类结果的质量。

关键词：文本聚类；词性标注；自然语言处理；聚类分析

中图分类号：TP393.08　　**文献标志码**：A　　**文章编号**：1672 - 4348(2015)04 - 0372 - 04

A text clustering algorithm based on part-of-speech tagging

Wang Jinshui, Tang Zhengyi, Xue Xingsi

(College of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, China)

Abstract: To tackle the problem that traditional text clustering methods are susceptible to the effects of noises, a text clustering algorithm based on part-of-speech tagging was proposed. Firstly, the part-of-speech tagging was utilized to recognize the keywords that well characterize the text features. A text clustering based on the recognized keywords was performed via the proposed algorithm. The experimental results show that comparing with the clustering results generated by the traditional clustering algorithm, our proposal was able to effectively improve the quality of clustering results.

Keywords: text clustering; part-of-speech tagging; natural language process; cluster analysis

随着互联网的高速发展，网络数据量呈现出爆炸式的增长态势。然而，这些海量数据大多以文本形式存在，且这些数据与存放在数据库中的高度结构化的数据不同，它们往往是以半结构化、甚至是无结构化的形式存在。因此，如何有效地分析和处理庞大的文本数据便成为一个亟待解决的问题。文本聚类作为文本信息处理领域一个重要分支，其目标就是研究如何更有效地组织和管理文本信息，并快速、准确、全面地从海量文本数据中找到用户所需要的信息。因此，文本聚类被视为解决信息杂乱和信息爆炸的一个有效技术，并得到了越来越多的关注[1]。

文本聚类技术作为一种无监督的机器学习方法，它能够在给定的某种相似性度量下完成对文本集合的分组，从而将彼此相近或相似的文本分到同一个组内。文本聚类用于辅助用户信息获取可表现在以下四个方面：1）合理组织检索结果，它能够自动地根据语义信息将不同文本划分到多个类别，使用户能够快速定位到所关注的信息；2）个性化信息推送，文本聚类能够识别和划分对不同事务感兴趣的用户小组，从而帮助系统根据用户偏好进行个性化信息的推送；3）提供可视化

收稿日期：2015 - 06 - 10
基金项目：福建省中青年教师教育科研项目（JA15348，JB14069，JB12146）；福建省科技厅高校项目（JK2012033）；福建工程学院科研启动基金项目（GY - Z13113，GY - Z14068，GY - Z13112）
第一作者简介：王金水（1981 - ），男，福建漳州人，讲师，博士，研究方向：软件工程。

的多文本摘要，可通过聚类结果和文本间关联关系将不同文本组成可视化的文本关联图，以提供更加友好的用户查询接口；4）加速检索过程，经过文本聚类处理之后，用户进行检索时只需将检索项与各组的类中心进行相似度比较，而不用计算检索项与所有文本的相似度，从而能够有效地提高检索效率[2]。

在传统的文本聚类技术中，文本数据通常采用向量空间模型（vector space model，VSM）进行描述，每个单词都被视为特征空间坐标系的一维，而每个文本则是特征空间的一个向量。但是，文本特别是 Web 文本中包含着大量与文本特征无关的单词（例如停用词、虚词等），它们作为“噪声”不仅影响着文本聚类的准确率和效率[3]，而且往往会导致文本向量空间变得异常高维且稀疏。为了减少文本噪声对聚类过程及结果所造成的不利影响，本文提出了一个基于词性标注的聚类算法，该算法利用词性标注从文本中识别并抽取最能体现文本特征的关键词，再基于所抽取的关键词进行文本聚类操作。为了验证方法的有效性，本文在一个开放的测试文档集合上比较了基于传统聚类算法与基于词性标注的聚类算法所得到的聚类结果。实验发现，相对传统的聚类算法，基于词性标注的文本聚类算法能够有效地提高聚类结果的质量。

1　相关技术

1.1　文本聚类

聚类分析是在无类别标记信息的情况下，根据某种相似性计算公式对物理或抽象的待分析对象集合进行分组，使每个分组能够自我识别并且区分于其他分组。聚类的过程可以描述为：对对象集合 $X = \{x_1, x_2, \cdots, x_n\}$ 进行分组，进而得到分组集合 $C = \{c_i \mid c_i \subset X, i = 1, 2, \cdots, m, \bigcup_{i=1,\cdots,m} c_i = X\}$。其中，$C_i$ 也被称为一个簇或聚类簇。因此，聚类可视为一种通过对对象集合按照某种规则进行划分，进而从中发现隐含的有用信息的知识发现方法。

文本聚类主要是基于以下的聚类假设而产生的：同类的文档相似度较大，而不同类的文档相似度较小。与聚类算法类似，它可以被描述为：对一个给定的文本集合 $D = \{d_1, d_1, \cdots, d_n\}$，经过聚类后可得到一个聚类簇集合 $C = \{c_i \mid c_1, c_1, \cdots, c_m\}$，其中，$c_i \subset D(i = 1, 2, \cdots, m)$，使得 $\forall d_i (d_i \in D)$，$\exists c_j (c_j \in C)$ 都有 $d_i \in c_j$ 成立，且使得某个代价函数 $f(C)$ 达到最小。

作为一种无监督的机器学习方法，文本聚类具有较高的自动化处理能力，被视为对文本信息进行组织、摘要和管理的重要方法[4]。但是，由于传统的文本聚类算法往往受到文本噪声的影响，使得生成的向量空间高维且稀疏，从而不仅影响了算法的效率，并且聚类结果也很难令人满意[5]。研究发现，通过在原始的特征集合中选择一小部分最有效的特征可以非常有效地解决高维稀疏的问题[6]。但由于缺乏类信息的指导，文本聚类难于从大量的单词中准确地识别最能体现文本特征的单词。

1.2　词性标注

现有的文本处理技术大多采用词袋模型（bag of words）作为文本的表示模型，然而词袋模型忽略了单词之间的词义关系和潜在的概念结构，所以难于识别最能体现文本特征的关键词。一般而言，名词和动词是句子中最重要的成分。研究发现，动词的无标记匹配是事件或命题，在句法结构中充当谓语核心；而名词的无标记语义匹配是事物，在句法结构充当论元。表述完整的句子语义需要名词和动词同时发挥作用[7]。在文本处理时，选择具有重要词性角色的单词作为特征不仅有助于提高效率，还可以提高处理结果的质量[8]。

在自然语言中，表达意义的单词无论在句法层面还是语义层面都可能存在歧义，即词性兼类问题。在句法层面上，一个单词可能同时拥有多个词性。例如，“评论”可在句子“张三写了一个评论”中作为名词，而在句子中“张三评论了这一事件”则作为动词。在语义层面上，一个单词可能有多个义项。例如，“苹果”在句子“张三买了一斤苹果”中表示一种水果，而在句子“张三买了一个苹果手机”中表示一个品牌。然而，各种兼类词在特定的上下文中总是具有确定的词性[9]。词性标注（part-of-speech tagging，POS tagging）的作用就是通过采取适当的方法，根据上下文的语境关系消除句子中单词的语法兼类，使得每个单词在特定的场合下只保留一种词性。

综合以上分析，为抽取句子中的关键词，应综合考虑到不同单词在文本中所发挥的作用及其被

使用的方式。动词和名词作为句子的核心最能体现文本特征,因此本文利用词性标注从文本中抽取出动词和名词作为关键词,并通过关键词改进文本聚类的质量。

2 方法框架

基于词性标注的文本聚类算法的主要处理步骤如图 1 所示。其中,矩形框表示制品,圆角矩形表示操作。本文在实现过程中采用 OpenNLP[①] 提供的组件完成句子切分和词性标注的任务。

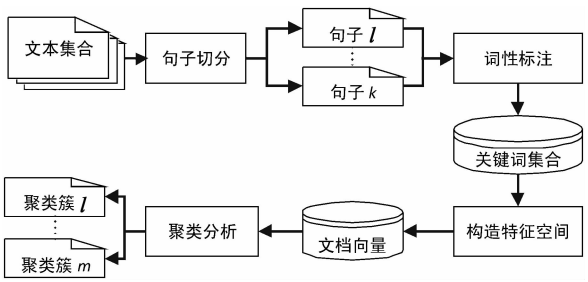

图 1 基于词性标的文本聚类算法过程

Fig. 1 Flowchart of text clustering algorithm based on part-of-speech tagging

完整且独立的语句是词性标注的必要输入项,因此需要先将文本集合切分为一组不相交的句子。在此之后,可通过词性标注技术将文本中的名词和动词识别出来,并将其作为关键词集合以参与后续的分析。得到文档中的关键词之后,可通过构造特征空间得到文档向量,并通过聚类分析得到最终的聚类结果。考虑参与聚类的文档数量较为庞大,因此需要采用较高运算效率的聚类算法。对此,本文采用 Carrot 所提供的 Lingo 算法来完成最终的聚类分析。[②] Lingo 算法是基于奇异值分解的索引结果聚类算法。相对其他聚类算法,Lingo 算法有以下 2 个优点:首先,它生成的聚类簇尺寸较小;再者,它的运算速度更快,适合需要进行大量聚类运算的场景。具体算法描述如下:

输入:文本集合 $D = \{d_1, d_1, \cdots, d_n\}$

输出:已标注的 m 个聚类簇,聚类簇集合 $C = \{c_i \mid c_1, c_1, \cdots, c_m\}$

(1) 对文本集合 D 中的每个文档 d_i 进行句子切分,得到句子集合 $S = \{s_1, s_2, \cdots, s_k\}$。

(2) 通过词性标注识别所有句子 s_i 中的名词和动词,将其保存为关键词集合。

(3) 将文本中的每个关键词视为特征空间坐标系的一维,而每个文本则作为特征空间的一个向量,构造文档向量。

(4) 基于所构建的文档向量,通过聚类算法生成聚类簇集合 C。

3 实验分析

通过一个案例研究分析词性标注对文本聚类所产生的影响。以开源自然语言处理工具 LingPipe 提供的 178 个文本文件[③] 作为实验对象,通过比较不同词项集合的文本聚类结果的质量验证基于词性标注的文本聚类算法的有效性。

3.1 实验项目

在案例研究中使用如表 1 所示的归属于 4 个不同主题的 178 个文本文件。其中,最大的文本文件包含 23 485 个字符(不计空格)和4 767个文字,最小的文本文件包含 168 个字符(不计空格)和 32 个文字。

表 1 实验文本文件主题及编号

Tab. 1 The subject and number of text files

主题名	文件数/个	文件编号	
		最小编号	最大编号
soc. religion. christian	45	21412	21499
alt. atheism	46	53302	53399
misc. forsale	44	74756	76199
talk. religion. misc	43	82796	83992

3.2 实验算法

由于已知文本的总主题数,因此采用了层次聚类算法中的 complete-link 完成聚类分析,并设定聚类簇数量为 4。在实验开始之前,需标注所有文件所属的主题,以作为标准答案来评判聚类结果的质量。为了验证基于词性标注的文本聚类算法的有效性,对比了以下几种基于不同词项集合的文本聚类算法得到的结果。

① http://opennlp. apache. org/。

② http://project. carrot2. org。

③ http://alias-i. com/lingpipe/demos/tutorial/cluster/read-me. html。

1）Normal：不进行词性标注，以文件中所有词为相似度计算依据进行文本聚类。

2）POS：通过词性标注获取文件中的名词和动词，并以此为相似度计算依据进行文本聚类。

3.3　评测指标和评测结果

聚类算法类型的不同导致了聚类结果质量评测标准的不同，到目前为止缺乏一个统一的评测方法来比较不同聚类算法所得到的结果[10]。本文采用了在信息检索领域中常见的查全率（recall）、查准率（precision）和 F-measure。对于聚类结果的查全率和查准率而言，需要与人工标注的类别相结合，具体定义如下：

$$\text{recall} = n_i^j/n_i,\ \text{precision} = n_i^j/n_j,$$

$$F\text{-measure} = \frac{1 + b^2}{1/\text{precision} + b^2/\text{recall}}$$

式中 n_i 为人工标注的类别 i 的文档数量，n_j 为聚类结果中类别 i 的文档数量，n_j^i 为属于人工标注类别 i 且被分配到聚类类别 j 中的文档数量。在实验中，直接采用 LingPipe 提供的聚类结果验证类①完成聚类结果的评测，并将 F-measure 中的 b 设置为 1。

从表 2 可看出，相对于 Normal 方法，通过 POS 方法得到的聚类结果在查全率、查准率和 F-measure 的评测结果都更加理想（分别提高了 5%、6% 和 6%）。文件中存在的噪声对聚类过程及结果带来的不利影响得到较好的抑制。

表 2　聚类结果

Tab. 2　POS clustering results

%

方法	查全率	查准率	F-measure
Normal	50	33	40
POS	55	39	46

4　结论

本文提出了一个基于词性标注的聚类算法，该算法利用词性标注从文本中识别并抽取最能体现文本特征的关键词，再基于所抽取的关键词进行文本聚类操作。为了验证方法的有效性，本文在一个开放的文档集合上比较了基于传统聚类算法与基于词性标注的聚类算法所得到聚类结果。实验发现，相对传统聚类算法，基于词性标注的文本聚类算法能够有效地提高聚类结果的质量。

本文的工作为通过自然语言处理技术优化数据挖掘提供了一个新思路。由于在实验分析中使用的数据集较小，无法应用统计分析技术验证方法的有效性。因此，采用更大的数据集和更多的数据类型，并从理论上对本文所提出的方法进行有效性分析将是下一步研究的内容。

参考文献：

[1] 杨震. 文本分类和聚类中若干问题的研究[D]. 北京：北京邮电大学，2007.

[2] 王春龙. 文本聚类关键技术研究[D]. 北京：华北电力大学，2014.

[3] 叶宇飞，安世全，代劲. 一种新的 Web 中文文本聚类方法研究[J]. 计算机应用与软件，2013(12)：222－225.

[4] 姚清耘，刘功申，李翔. 基于向量空间模型的文本聚类算法[J]. 计算机工程，2008，34(18)：39－41.

[5] Aggarwal C C, Yu P S. Finding generalized projected clusters in high dimensional spaces[J]. Sigmod，2002，29(2)：70－81.

[6] Dash M, Koot P W. Feature Selection for Clustering[M]. Berlin：Springer，2000：110－121.

[7] 刘丹青. 汉语是一种动词型语言——试说动词型语言和名词型语言的类型差异[J]. 世界汉语教学，2010(1)：3－17.

[8] 韩普，王东波，刘艳云，等. 词性对中英文文本聚类的影响研究[J]. 中文信息学报，2013，27(2)：65－73.

[9] 郭永辉，吴保民，王炳锡. 一种用于词性标注的相关投票融合策略[J]. 中文信息学报，2007，21(2)：9－13.

[10] 苏冲. 基于最大频繁项集的搜索引擎查询结果聚类方法[D]. 哈尔滨：哈尔滨工业大学，2009.

（责任编辑：肖锡湘）

① http://alias-i.com/lingpipe/demos/tutorial/cluster/read-me.html。