基于 FCM 聚类算法的单词型术语识别方法

周浪[1,2] 史树敏[2] 冯冲[2] 黄河燕[2,3]

1. 南京理工大学计算机科学与技术学院，江苏 南京 210094

2. 中国科学院计算机语言信息工程研究中心，北京 100097

3. 北京理工大学计算机学院，北京 100081

E-mail: yzzhoulang@126.com

摘 要： 大部分的术语抽取工作都将重点放在词组型术语的识别上，忽略了单词型术语。虽然在整个术语系统中，单词型术语的数量要比词组型术语少得多，但它却是构成词组型术语的重要元素。由于单词型术语具有语法边界清晰的特点，引入模糊 C-均值聚类算法，将术语识别工作转化为两类聚类任务，从而实现无监督自动标注的目的，并获得了令人满意的结果。

关键词： 自然语言处理；模糊聚类；FCM 算法；单词型术语

A Single-Word Term Recognition Approach Based on FCM Clustering Algorithm

ZHOU Lang[1,2], SHI Shu-Min[2], FENG Chong[2], HUANG He-Yan[2,3]

1. School of Computer Science and Technology, University, Nanjing University of Science and Technology, Nanjing 210094;

2. Research Center of Computer & Language Information Engineering, CAS, Beijing 100097

3. School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081

E-mail: yzzhoulang@126.com

Abstract: Since multi-word terms occur much more frequently than the single-word terms, most term extraction systems focus on the former. But many foundational terms in specific domain are in the form of single word, and they're the important components in the multi-word terms. In the single-word extraction process, fuzzy C-means clustering algorithm is introduced to transform the extraction task into clustering task. Without manual assistant and additional resource, this approach could get a satisfying result.

Keywords: Natural Language Processing; Fuzzy Clustering; FCM Algorithm; Single-word Term;

1 引言

随着网络的高速发展以及各类资讯的日益公开化，人们越来越容易获得所需的各种信息，其中以文档出现的资源占据了绝大多数。为了能从这些资源中高效地获取感兴趣的内容，同时保证交流的顺畅，自动术语抽取技术的发展越来越受到人们的重视。

由于在整个术语系统中，有 80%以上的术语都以短语的形式出现[1]，词组型术语在数量上占据了绝对的优势，因此很多术语抽取方法都将研究目标锁定在了词组型术语上[2-3]。但是从某种意义上讲，是否为术语就是一个词条的固有属性，是否具有该属性不应该受到词数的限制[4]。在各专业领域中，很多重要的术语都是以单个词汇的形式出现，如物理领域的“力”、“场”等等。而

且，单词型术语也是词组型术语的重要组成元素[1]。

由于单词型术语自身包含的信息量较少，所以目前使用的单词型术语抽取方法都采用了多种辅助资源。Lemay & L'Homme[5]提出了两种抽取策略：一种在拥有面向专业领域的语料库的同时，还需要一个通用语料库，通过对比单词在不同语料库中的词频来获取术语；另一种方法则是将专业领域的语料库划分为若干子集，通过比较单词在子集和整体中的词频来判断是否为术语。Zan[6]则是使用了北大的双语语义词典 CCD 来辅助抽取法律专业的术语。

以上的方法都需要加入额外的辅助措施，但是构造这些资源需要耗费大量的人力物力。本文的目的就是尝试使用统计的方法，以较少的资源来实现单词型术语的抽取。因此，引入了模糊聚类机制，将抽取工作转化为聚类任务。

2 FCM 聚类算法

在众多模糊聚类算法中，模糊 C-均值[7]（Fuzzy C-Means, FCM）聚类算法是理论发展最为完善、应用领域最广泛的一种模糊聚类算法。FCM 算法是一种基于目标函数的模糊聚类算法，可以将聚类归结为一个带约束的非线性规划问题，并通过优化求解获得数据集的模糊划分和聚类。这种方法设计简单，解决问题的范围广，还可以转化为优化问题而借助经典数学的非线性规划理论求解，并易于在计算机上实现。FCM 算法利用均方逼近理论构造了带约束的非线性规划函数，采用类内加权平方误差和作为聚类目标函数。

设有样本集合 $X=\{x_1,x_2,\mathrm{L}\ ,x_n\}$，其中 n 表示样本的个数。FCM 算法在对 X 中的数据进行划分的过程中，通过更新各模糊簇的类中心，使得目标函数达到最小或满足其他收敛条件。FCM 算法的目标函数为：

$$J_m(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}\left(u_{ij}\right)^m\left(d_{ij}\right)^2$$

其中 $U=\{u_{ij}\}$ 是隶属度矩阵，u_{ij} 表示样本 x_i 对第 j 个类别的隶属程度，满足 $u_{ij}\in(0,1)$ 并且有 $\sum_{j=1}^{c}u_{ij}=1$；$V=\{v_j\}$ 表示类中心矩阵；m 为模糊加权指数，且 $1\le m<\infty$；c 表示聚类后的类别数，$c\ge 2$；$d_{ij}=\|x_i-v_j\|$ 表示样本 x_i 到类中心 v_j 的距离。

FCM 算法是对自变量 (U,V) 的一个约束优化处理，通过初始化类中心或者隶属度矩阵、方程迭代，直到使得目标函数最小化。类中心和隶属度在迭代过程中的更新方程如下所示：

$$v_j=\frac{\sum_{i=1}^{n}\left(u_{ij}\right)^m\cdot x_i}{\sum_{i=1}^{n}\left(u_{ij}\right)^m},\ j=1,2,\mathrm{L}\ ,c$$

$$u_{ij}=\left[\sum_{k=1}^{c}\left(\frac{\|x_i-v_j\|}{\|x_i-v_k\|}\right)^{2/(m-1)}\right]^{-1},i=1,2,\mathrm{L}\ ,n$$

3 基于 FCM 聚类算法的单词型术语抽取方法

由于单词型术语的语法边界较为清晰，只需使用分词处理文本，即可获得满意的结果。因此，术语的分辨任务可以转化为对候选集合的一个两类划分过程。FCM 聚类过程中存在两个主要难题：一是需处理的数据太多时，时间开销很大；二是无法预知初始类中心的位置或属性，只能随机选择初始类中心。针对这两个问题，下文中结合自然语言处理措施，使其术语抽取过程中得到了妥善的解决。

3.1 候选术语集的生成

在《中国大百科全书》中，将术语定义为“各门学科中的专门用语。术语可以是词，也可以是词组，用来正确标记生产技术、科学艺术、社会生活等各个专门领域中的事务、现象、特性、关系和过程”。术语能够表达特定领域中一个专门的概念[8]，因此虚词不可能成为术语。冯志伟[1]对汉语单词型术语的词类进行分析后，总结出只有名词、形容词、动词、数词和量词可以在句子中以术语的形式独立出现。

在生成候选术语集时，只接纳名词、形容词、动词、数词和量词进行下一步的处理。这不仅可以提高术语抽取结果的正确率，还可以大规模减少需处理的数据，提升聚类性能。

3.2 初始类中心的选择

在 FCM 算法中，各类中心的初始位置和属性是随机选取的，如果初始类中心与实际的类中心非常接近，则迭代次数很小，快速收敛于实际类中心。反之，则会耗费大量的聚类时间。因此，我们并不采用随机的方式来选择初始类中心，而是使用 TFIDF[9]准则度量每个候选样本，分别选择 TFIDF 值最大的样本和值最小的样本作为两类的类中心。

$$TFIDF(t)=tf(t)\cdot\log\frac{N}{df(t)}$$

其中，$tf(t)$表示词t在语料中出现的词频；$df(t)$表示词t的文档频率；N表示语料库中包含的所有文档数。

TFIDF 方法已经被成功嵌套于很多术语抽取方法[10-11]中，辅助术语抽取工作。如果词t在少量的文本中频繁出现，则极有可能是专业术语，相应地也能获得较高的 TFIDF 值。因此 TFIDF 值最高的样本很可能就是真正的术语且具备典型的术语分布特征，选择该样本作为初始正例类中心；同理，选择 TFIDF 值最低的样本作为初始负例类中心。相较于随机选择的类中心点，使用 TFIDF 度量后的设置更加接近实际的类中心，可以减少迭代计算次数，加快聚类速度。

3.3 特征表示及距离计算

为了简化问题，采用向量空间模型（Vector Space Model，VSM）作为特征表示方法，来表示每个词语在语料中的分布特性。设待聚类的词语特征向量为$x_i(x_{i1},x_{i2},\mathrm{L}\ x_{ik},\mathrm{L}\ ,x_{iN})$，$N$表示特征维数。用词语在每篇文档中出现的比重来衡量词语在文档中的权重，并按权重的大小进行排序。

$$x_{ik}=\sqrt{\frac{tf_{ik}}{Nr_k}}$$

其中，tf_{ik} 表示词语 x_i 在文档 k 中出现的次数；Nr_k 表示文档 k 中包含的实词数。

本文使用欧氏距离作为样本点到类中心的距离度量。

$$d_{ij}=\left\|x_i-v_j\right\|=\sqrt{\frac{1}{M}\sum_{k=1}^{M}\left(x_{ik}-v_{jk}\right)^2}$$

其中，M 取值的约束条件为：$M=\min\left(k,l\right)$，k 和 l 的取值满足 $x_{ik}\neq 0,\forall x_{ik}^{-}=0$，$v_{jl}\neq 0,\forall v_{jl}^{-}=0$。

4 实验及结果分析

4.1 实验语料及参数设置

实验中所用的测试语料为 200 篇计算机领域的技术论文，使用中科院计算所的词法分析工具 ICTCLAS[12]进行分词处理后，共包含了 605,730 词次。去除停用词后，则有 13,754 个单词，选择其中出现 2 次以上的名词、动词、形容词、数词和量词作为候选项，共有 4,964 个单词。

聚类的类别数设置为 2，即 $c=2$；模糊加权指数 m 的值置为 2，收敛阈值 ε 设为 $1e^{-3}$，即当第 k 次迭代和第 $k+1$ 次迭代类中心向量的误差 $\left\|V^k-V^{k-1}\right\|\leq\varepsilon$ 时，则停止迭代计算。

4.2 实验结果及分析

对候选集中的近 5,000 条单词进行人工判断是否为术语。其中有 1,568 个单词属于真正的术语。使用 FCM 聚类算法对候选集进行划分后，结果如表 1 所示。

表 1 FCM 聚类算法划分结果

	人工判断结果	FCM 聚类划分结果	FCM 划分正确结果
正例数	1,568	925	752
负例数	3,369	4,039	3,223

由表 1 中的结果数据，可以看出使用 FCM 聚类算法来识别语料库中的单词型术语时，能够获得 81.30%的正确率，但是召回率偏低，仅为 47.96%。

由于可用于单词型术语抽取的统计方法较少，我们选择 TFIDF 方法做为对比，来比较 FCM 算法应用于抽取术语中的效果。TFIDF 的实验结果如图 1 中所示。

通过图 1 可以看出，当正确率维持在 70%以上的时候，召回率仅为 19.64%；而召回率大于 30%的时候，TFIDF 方法所获得的正确率下降速度较快；当召回率为 50%的时候，正确率仅为 49.06%。

通过以上的实验可以发现，FCM 聚类算法能够成功应用于单词型术语的识别任务中，而且能够获得较满意的结果。

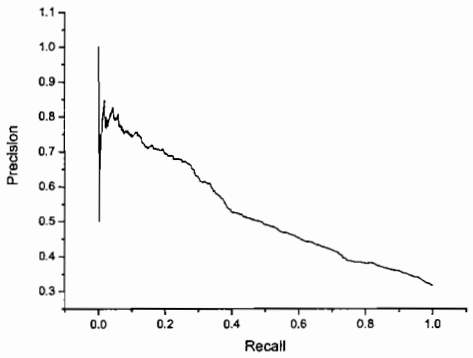

图 1 TFIDF 实验结果

5 总结及下一步工作

由于使用现有的语言处理工具，可以快速高效识别单词的边界，因此在识别单词型术语的任务中，引入了模糊聚类机制，成功地将抽取任务转化为分类任务。而 FCM 聚类算法是一种经典的模糊聚类方法，具有理论基础完善，应用范围广的优点。同时结合自然语言处理中常用的统计度量策略辅助处理 FCM 算法中的数据压缩和初始类中心选择问题，在减少需计算的数据量的同时，能够减少聚类迭代次数，加快收敛速度。在无需人为辅助及其他资源的情况下，可以获得较为理想的术语识别结果。

在下一步的工作中，可以结合短语边界识别方法，将模糊聚类方法应用于词组型术语的抽取工作中。

参考文献

[1] 冯志伟. 现代术语学引论[M]. 语文出版社. 1997.

[2] Justeson J., Katz S.. Technical Term: Some Linguistic Properties and an Algorithm for Identification in Text[J]. Natural Language Engineering, 1995, 1(1):9-27.

[3] Frantzi K.T., Sophia Ananiadou, Hideki Mima. Automatic Recognition of Multi-word terms: the C-value/NC-value Method[J]. International Journal on Digital Libraries, 2000, 3(2):115-130.

[4] Le An Hua. Advances in Automatic Terminology Processing: Methodlogy and Application in Focus[D]. PhD Thesis of University of Wolverhampton.

[5] Chantal Lemay, Marie-Claude L'Homme, P. D. Two Methods for Extracting "Specific" Single-word Terms from Specialized Corpora: Experimention and Evaluation[J]. International journal of corpus linguistics, 2005,10(2): 227-256.

[6] Hongying Zan, Guocheng Duan, M. F.. Single Word Term Extraction using a Bilingual Semantic Lexicon-based Approach[C]. Proceedings of the Third International Conference on Natural Computation (ICNC 2007).

[7] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. New York: Plenum Press, 1981.

[8] 陈原. 陈原语言学论著[M]. 辽宁教育出版社. 1988.

[9] Manning, C. D., and H. Schütze. 1999. Foundations of Statistical Natural Language Processing[M]. Cambridge, Massachusetts: MIT Press.

[10] Nakagawa. Experimental evaluation of ranking and selection methods in term extraction[C]. Recent Advances in Computational Terminology, Amsterdam: John Benjamins. 2001:303-326.

[11] Patry, A., and P. Langlais. 2005. Corpus-Based Terminology Extraction[C]. In Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, Copenhagen, Danemark. 2005:313-321.

[12] http://ictclas.org