Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

# 词聚类技术研究综述*

郭怀思 朱礼军 徐硕 / 中国科学技术信息研究所 北京 100038

摘要：词聚类是一种面向词语的聚类技术，广泛应用于自然语言处理的各个方向。文章将部分已有的词聚类方法分为基于语法特征、基于语义特征和基于语用特征三类，并对各类方法进行了归纳整理。

关键词：词聚类，语法特征，语义特征，语用特征
DOI：10.3772/j.issn.1673—2286.2010.05.004

## 1 引言

词汇作为最小的能够独立运用的语言单位，它们经常在语法、语义或语用上表现出一定的共性。词聚类就是根据词汇的语法、语义或语用特性，对特定的词汇集合进行聚类，使得每个聚簇中词汇间的共性尽量大，不同聚簇中词汇间的共性尽可能小。但部分研究者将词汇在文档集中的分布特征（如：某个词汇在整个文档集中的出现频率不高，但在单个文档中出现的频率很高的现象）也称为词聚类，如Thom and Zobel[1]，不过本文综述的词聚类技术指的是前者。

众所周知，词汇是一个开放的系统，很难说清楚一种语言究竟包含多少词汇，从而导致构建统计语言模型中经常出现数据稀疏的问题[2]。具体来说，这一问题是指无论收集到的语料库有多大，都难以包含所有的语言现象，因此经常难以准确估计语言模型中的许多参数。然而，词类是一个相对封闭的系统，而且词类的数量远小于词汇的数量。如果首先对词汇进行聚类处理，然后将类条件概率引入语言模型之中，这样可大大缓解数据稀疏的问题，词聚类最初也正是为了解决该问题而被提出的[3-7]。当然，因为词类的数量比较少，使得构建高阶语言模型也成为可能。随着词聚类技术的发展，它还在语义消歧[8-9]、主题抽取[10]、文本分类[11-13]、信息检索[14-15]等应用中发挥着重要的作用。

词聚类过程通常分为两个步骤：（1）特征提取或相似度/相关度计算，这是词聚类的前提；（2）选择合适的聚类方法完成聚类分析，其中划分聚类法及层次聚类法常见于各种词聚类的文献报道中。参照自然语言分析处理的层面，本文从提取词语特征的角度把词聚类方法分为基于语法特征、基于语义特征和基于语用特征三大类，分别见本文的第2节、第3节和第4节，最后本文从多个角度比较了这三类方法，并指出了当前词聚类技术存在的问题，以及未来可能的研究方法。

## 2 基于语法特征的词聚类

基于语法特征的词聚类，也被称为基于语料库的词聚类。因为语料库是自然语言运用的实例，基本符合语言的语法规律，因此这类方法的基本思想是从目标词的上下文中提取目标词的语法特征，使得语料库中语法特征相近的词汇通过聚类分析能够很好地聚在一起。容易看出，这类方法是建立在统计基础之上的，属于基于统计的方法。根据词汇特征维度的确定方式，又可进一步将其分为事先确定特征维度法以及

* 国家“十一五”科技支撑计划课题“知识组织系统的集成及服务体系研究与实现”（2006BAH03B03）和“科技文献信息服务系统关键技术研究及应用示范”（2006BAH03B06）资助项目。
中国科学技术信息研究所重点工作项目“汉语科技词系统建设与应用工程——新能源汽车领域完善及领域扩展”（2009KP01—3—2）资金项目。

实时确定特征维度法。

### 2.1 事先确定特征维度法

该类方法需要事先选择一些有代表意义的词汇作为目标词的特征维度，统计语料库中目标词与作为特征维度的词共现的频次，从而将目标词表示为一种特征向量的形式，这样许多成熟的聚类方法都可以被采用。为了捕获有效而且敏感的特征，在统计目标词与作为特征维度的词的共现频次时，通常不是在整篇文档范围内，而是在目标词周围设置一个窗口，针对这个窗口进行统计。

关于特征维度的选取问题，有研究表明[16]，可以选取和目标词构成相关关系的词作为特征维度，这与所聚类词的语言特点也有联系。比如，汉语中的虚词作为特征维可以对实词进行有效表示[17]。而日语中常选取动词-名词组合，然后互为特征维构建词空间[38]。

### 2.2 实时确定特征维度法

事先确定特征维度法通常用于对不同词性的词汇集合进行聚类分析，它对特征维度的选取特别敏感，从一定程度上限制了它的适用范围。而实时确定特征维度法只在目标词的附近选择一些词汇作特征揭示，即利用目标词的上下文表示目标词。目前，目标词的上下文尚没有明确的定义，不同的研究者采用不同的形式，可以是紧随目标词其后的一个词，也可以是目标词前后的若干词。

Farhat等人[18]给出了目标词上下文的形式化表示，假设目标词T和它所在文本环境为$...w_nw_{n-1}...w_1Tw_{-1}w_{-2}...w_{-n}...$，则目标词T的长度为n的上下文可定义为n + 1种形式：$C_1=(w_n, w_{n-1}, ..., w_1)$，$C_2=(w_{n-1}, ..., w_1, w_{-1})$，...，$C_{n+1}=(w_{-1}, w_{-2}, ..., w_{-n})$。Farhat等人通过实验发现，上下文长度为2时聚类效果最佳，这样目标词就可被表示为二元随机变量，并且KL（Kullback-Leibler）距离被引入，用于计算两个目标词间的相似度。实际上，上下文长度究竟应该多大，通常与语料库的规模有关，而且经常存在长程效应现象，比如Gal等[19]发现距离目标词1000个词以外还存在与之相关的词汇。

另外还有一类方法（如最大互信息法[37]），它将特征维度的确定与聚类分析同时进行，采用的聚类标准是最小化语料库的困惑度（Perplexity）。具体来说，对于语料库$L=w_1w_2...w_n$，基于类的二元语言模型的困惑度定义为：$H(L)=-\frac{1}{n}\log p(w_1w_2\cdots w_n)\approx-\frac{1}{n-1}\sum_{i=1}^{n-1}\#(w_iw_{i+1})\log(p(c_i|c_{i+1})p(w_{i+1}|c_{i+1}))$

其中，$c_i$和$c_{i+1}$分别为wi和$w_{i+1}$所属的聚簇，$\#(w_iw_{i+1})$表示词对$w_iw_{i+1}$出现的频次。对上式进行简单变换，容易发现当词类间的互信息最大时，语言模型的困惑度最小。于是聚类准则就转换为词类间互信息总和为最大，从而可采用迭代划分的方式对语料库中的词语进行聚类。

### 2.3 改善聚类效果的方法

#### （1）对语料库进行语法分析

基于语法特征的词聚类属于基于统计的方法，它的效果依赖于语料库资源的质量。如果直接统计共现频率，可能会出现偏差，比如英文中插入语部分的词汇可能和目标词相关性不大，但也被统计在内。为解决这一问题，Habert等人[20]首先对语料库中的句子进行语法分析，然后精简语法树得到基本句子，最后从中提取句子中的主题词汇，提高了语料库的质量，因而改善了词聚类效果。

#### （2）采用多语言平行语料库资源

Wang等人[21]尝试利用双语平行语料库进行词聚类。他们将一种统计式的翻译模型和单语言词聚类中使用的互信息聚类算法结合在一起，利用英-德双语平行对齐语料库进行了小规模的词聚类实验。实验表明，这种方法从两种语言中吸取了有用信息，得到的结果会更准确，并且很适合应用在机器翻译中。缺点在于大规模的双语平行语料库资源不容易获得。

## 3 基于语义特征的词聚类

该类方法通常依赖某种语义知识库，如同义词词林、Hownet、Wordnet等，这些语义资源是语言专家对语义规律的直接编码，因此属于基于规则的方法。这一类方法常以词汇的语义相似度/相关度矩阵的形式表示词语，然后使用KNN、逐对K-Means[22]、层次聚类或刊登于*Science*杂志的AP（Affinity Propagation）聚类[23]得到聚类结果。

聚类过程中经常会涉及词集间的语义相似度/相关度计算的问题，目前常用的计算方法为single linkage、complete linkage以及average linkage等[24]。因为这些聚类

方法相对已经比较成熟，目前研究重点主要集中在如何有效地计算词汇间的语义相似度/相关度。

Rada等人[25]于1989年以及Lee等人[26]于1993年分别利用上下位关系词典来计算词语的语义相似度。其原理是：在上下位分类体系中两个结点词之间的路径越短，语义相似度越大。如果这两个词之间有多条通路，则选择最短路径作为它们的相似度度量的依据。这种方法的一个假设是，在分类体系中节点间的链接代表的长度是相同的。而事实上，结点疏密不同的子类中的链接代表的距离也不同。Agirre 和Rigau[27]于1996年定义的语义相似度指标不仅与词语结点在分类体系中的深度有关，还对所在的子类中的结点密度敏感，在较密的子结构中的结点间的距离要更近。除了使用上下位词典，研究者们还尝试了多种语义词典。Nagao同时使用上位词典和同义词词典计算语义相似度[28]，还有研究者利用词语的形态学信息，甚至反义词信息作为语义相似度计算的依据[29]。

Resnik[30]打破了基于路径长度的语义相似度的计算方法。他结合使用了分类体系和语料库来表示语义相似度，先计算出两个词语$W_{1p}$和$W_{2p}$的共同的上位词$c_p$的熵值，然后用最大的熵值来表示这两个词的语义相似度，即$sim(W_{1p},W_{2p})=\max_{c_p\in\{x|sub(x,W_{1p})\wedge sub(x,W_{2p})\}}\{-logP(c_p)\}$其中，函数sub(a, b)表示a是b的上位词，$P(c_p)$表示$c_p$在语料库中出现的频率。

以上方法只能对分类体系中的词进行词聚类，对于未登录词（比如由分类体系中的多个词语构成的复合词）就不能适用了。我国相关研究人员李峰和李芳[31]针对这种情况，提出了汉语复合词相似度计算的方案。先将两个汉语复合词切分为原子词，再利用汉语语义词典计算两个复合词中原子词的语义相似度的大小，确定原子词之间的配对关系，最后构建模型计算得到复合词的相似度。该方案的缺点是原子词匹配时只考虑了语义相似的因素，而忽略了语序，这样最明显的问题是无法区分原子词相同而语序不同的复合词。徐硕等人[32]针对这个问题，引入了生物信息学中的全局双序列比对算法对原子词匹配，得到了更为合理的匹配结果，汉语复合词语义相似度的计算结果也更为准确。

## 4 基于语用特征的词聚类

基于语用特征的词聚类针对某种具体的应用提取词汇特征，是面向应用的词汇聚类方法。语用特征是指面向具体应用时词汇表现出的特征，词与词之间的相似度不再局限于语义接近程度，而在于共同指向某一特定对象的程度。下面简要介绍几种具体的基于语用特征的词聚类方法的思想：

### （1）基于搜索引擎的搜索关键词聚类

网络搜索引擎用户如果对搜索结果不满意，会调整搜索策略，采用别的搜索关键词，或者和其他关键词合并检索。而前后使用的搜索关键词由于指向了同样的搜索目标，所以在语用上可能存在相似关系。搜索引擎通过cookie或者session来跟踪用户的搜索行为，就可以收集到这种面向搜索对象的数据资源，进而用于词聚类。得到的聚类结果反馈应用于搜索关键词推荐。

### （2） 利用文档分类资源做词聚类

Jerome R. BeUegarda等人[33]利用文档分类资源做词聚类。这里的文档分类资源是指事先整理好的文档分类结果，词语共现的范围扩大到了文档类，也就是利用文档类作为词语的特征维对目标词进行向量表示。这样做一方面可以缓解数据稀疏问题，另一方面由于以文档类为特征维，得到的聚类结果是面向文档集的，反映了目标词集在特定文档集中的分布情况。当然这种方法要依赖于在文档层级标注的语料资源，Jerome R. BeUegarda使用的是ARPA北美商业新闻语料库（the ARPA North American Business News Corpus）。

### （3） 通过搜索引擎利用网页资源做词聚类

词聚类技术最初是为了解决统计语言模型中的数据稀疏问题的，于是数据稀疏在基于语料库的词聚类过程中依然是不可避免的。使用大规模的语料库是克服数据稀疏的一个方法。Yutaka Matsuo等人[34]通过搜索引擎利用互联网网页这一海量数据资源进行词聚类。他们先统计在网页范围内目标词两两之间的共现次数，在此基础上计算目标词两两相似度，最后使用纽曼聚类（Newman Clustering）算法得到聚类结果。利用搜索引擎所做的聚类，词语共现的范围是网页，也是一种面向具体应用的词聚类。

## 5 词聚类方法归纳

### （1）基于统计和基于规则的两个方向

基于语义特征的词聚类属于基于规则的方法，基于语法特征和基于语用特征词聚类都属于基于统计的

方法。基于语法特征的词聚类其实是基于语用特征的词聚类的一个特例，只不过前者的应用比较特殊，它应用在了语言模型构建。这三类方法在目的、使用资源、词语的特征表达以及评价方法均各有差异，详见表1。

（2）基于语料库的方法提取的是语法特征相近的词

基于语料库的方法提取的是语法特征相近的词，而不能直接提取出语义相近的词，这是由词语特征提取所采用的资源和算法决定的。语法特征的相似程度可以用替换相似度来衡量。替换相似是指在特定的语言上下文中，两个词可以互换而不影响上下文的结构。文献[35]把替换相似度（Substitutional Similarity）也称为语义相似度（Semantic Similarity）是不准确的。语义相似和替换相似不等价，即语义相似的词语互换而不影响上下文的结构，而替换相似的词语语义上未必相近，可以是相反，或者只是相关。这可以从一些使用基于语法特征的词聚类方法得到的词簇有时候是语法相近，有时候是语义相近得到印证。

（3）在词聚类中采用模糊聚类

传统的词聚类算法，每一个词最终都归入一个且仅归入一个词类，属于明确聚类（Crisp Clustering）。但是现实中，一个词语有多个义项，明确聚类很明显忽略了词语的多义性，在应用中词类的质量会降低。Guihong Cao[36]等人将模糊聚类（Fuzzy Clustering）引入到词聚类算法中，并做了小规模数据的词聚类实验，取得了良好的效果。这种模糊聚类模型给每一个词附以多个度数，表示这个词归属于生成的各个词类的程度，以此来体现词语的多义性。这种方法应当广泛应用于词聚类中。

## 6 结语

词聚类广泛应用于自然语言处理的各个方向，是一种基础性技术。词聚类技术的提高对于自然语言处理领域有着重要意义。本文将部分已存在的词聚类方法分为基于语法特征、基于语义特征和基于语用特征三类，并对各类方法进行了归纳整理，并指出了当前词聚类技术存在的问题，以及未来可能的研究方法。

表1 词聚类方法对比分析

| | 基于语法特征的词聚类 | 基于语义特征的词聚类 | 基于语用特征的词聚类 |
|---|---|---|---|
| 目标或应用方向 | 改善语言模型 | 建立语义词典 | 满足具体应用 |
| 使用的资源 | 语料库 | 语义知识库 | 从应用中获取的数据资源 |
| 词的表示方式 | 相似度矩阵/词空间向量 | 相似度矩阵 | 相似度矩阵/词空间向量 |
| 评价方法 | 最大互信息标准等 | 人工判定 | 实践检验 |

**参考文献**

[1] JAMES A T, JUSTIN Z. A Model for Word Clustering[J]. Journal of the American Society for Information Science and Technology, 1992.
[2] PETER F B, VINCENT J D P, PETER V D, JENIFER C L,ROBERT L M. Class-Based n-gram Models of Natural Language[J]. Computational Linguistics, 1992.
[3] 陈浪舟,黄泰翼. 一种新颖的词聚类算法和可变长统计语言模型[J]. 计算机学报, 1999,22(9).
[4] SHINSUKE M, MAKOTO N. A Stochastic language model using dependency and its improvement by word clustering[C]// Université de Montréal, Government of Canada. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 1998: 898-904.
[5] JOHN G M, FRANCIS J S. Improving Statistical Language Model Performance with Automitically Generated Word Hierarchies[J]. Computational Linguistics, 1996,22(2):217-247.
[6] BASSIOU N K, KOTROPOULOS C L. Interpolated distanced bigram language models for robust word clustering[C]// Nonlinear Signal and Image Processing.[出版者不详], 2005.
[7] SHINSUKE M, NISHIMURA M, NOBUYASU I. Language Model Adaptation using Word Clustering[J]. Joho Shori Gakkai Kenkyu Hokoku, 2003,2003(14):89-94.
[8] DAN T, RADU I, NANCY I. Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets[C]// Proceedings of the 20th international conference on Computational Linguistics, 2004, Geneva, Switzerland. Association for Computational Linguistics, 2004:1312-1318.
[9] JIN P, SUN X, WU Y, YU S. Word Clustering for Collocation-based Word Sense Disambiguation[C]// Proceedings of the 8th International Conference on Computational Linguistics, Geneva, Switzerland. Association for Computational Linguistics, 2004.

[10] 陈炯,张永奎. 一种基于词聚类的中文文本主题抽取方法[J]. 计算机应用,2005.
[11] CHEN W, CHANG X, WANG H, ZHU J, YAO T. Automatic Word Clustering for Text Categorization Using Global Information[C]// First Asia Information Retrieval Symposium (AIRS 2004), 2004.10, Beijing. [出版者不详], c2004:1-6.
[12] 朱慕华,陈文亮,朱靖波. 词聚类在文本分类中的应用[C]// 中国中文信息学会. 第二届全国学生计算语言学研讨会论文集. [出版者不详],2004:399-405.
[13] INDERJIT S D, SUBRAMANYAM M, RAHUL K. Enhanced word clustering for hierarchical text classification[C]// SIGKDD,SIGMOD, AAAI. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2002:191-200.
[14] SAEEDEH M, DIETRICH K. A word clustering approach for language model-based sentence retrieval in question answering systems[C]// Conference on Information and Knowledge Management Proceeding of the 18th ACM conference on Information and Knowledge Management. ACM, 2009:1911-1914.
[15] MICHIKO Y, HIDETOSHI Y. Term Clustering based on Lengths and Co-occurrences of Terms[C]// Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia. [出版者不详],2009.
[16] 闻扬，苑春法，黄昌宁. 基于搭配对的汉语形容词一名词聚类[J].中文信息学报,2000,14(6).
[17] WANG B, Wang H F. A Comparative Study on Chinese Word Clustering[J]. Lecture Notes in Computer Science, 2006, 4285/2006:157-164.
[18] FARHAT A, ISABELLE J, O'SHAUGHNESSY D. Clustering words for statistical language models based on contextual word similarity[C]// Acoustics, Speech and Signal Processing. IEEE Computer Society, 1996:180-183.
[19] GALE, CHURCH, YAROWSKY. A Method for Disambiguating Word Senses in a Corpus[J]. Computers and the Humanities, 1992:415-439.
[20] BENOÍT H, ELIE N, ADELINE N. Symbolic word clustering for medium-size corpora[C]// Proceedings of the 16th conference on Computational linguistics. ACM, 1996:490-495.
[21] WANG Y Y, LAFFERTY J, WAIBEL A. Word Clustering with Parallel Spoken Language Corpora[C]// Proceedings of the Fourth International Conference on Spoken Language. [出版者不详],1996:2364-2367.
[22] RICHARD O D, PETER E H, DAVID G S. Pattern Classification[M]. John Wiley & Sons, Inc, 2001.
[23] BRENDAN J F, DELBERT D. Clustering by Passing Messages between Data Points[J]. Science, 2007(315):972-976.
[24] ANIL K. J, RICHARD C D. Algorithms for Clustering Data[M]. Prentice-Hall, Inc, 1988.
[25] RADA R, HAFEDH M, BICKNELL E, BLETTNER M. Development and application of a metric on semantic nets[J]. IEEE Transactions on System, Man, and Cybernetics, 1989,19(1):17-30.
[26] LEE J H, KIM M H, LEE Y J. Information Retrieval based on conceptual distance in IS-A hierarchies[J]. Journal of Documentation, 1993,49(2):188-207.
[27] AGIRRE E, RIGAU G. Word sense disambiguation using conceptual density[C]// Proceedings of COLING'96. [出版者不详], 1996.
[28] NAGAO M. Some Rationales and Methodologies for Example-Based Approach[C]// Proceedings of International Workshop on Fundamental Research for the Future Generation of Natural Language Processing, 30-31 July 1992, Manchester, UK. [出版者不详], 1992:82-94.
[29] NIRENBURG S, DOMASHNEV C, GRANNES D I. Two Approaches to Matching in Example-Based Machine Translation[C]// Proceedings of TMI-93. [出版者不详], 1993:47-57.
[30] RESNIK P. Using information content to evaluate semantic similarity in a taxonomy[C]// Proceedings of IJCAI. [出版者不详],1995.
[31] 李峰,李芳. 中文词语语义相似度计算—基于《知网》2000[J]. 中文信息学报, 2007,21(3):99-105.
[32] XU S, ZHU L J, QIAO X D, XUE C X. A Novel Approach for Measuring Chinese Terms Semantic Similarity based on Pairwise Sequence Alignmen[C]// The Fifth International Conference on Semantics, Knowledge and Grid , [出版者不详],2009.
[33] JEROME R B, JOHN W B, Y L CHOW, NOAH B C, DEVANG N. A Novel Word Clustering Algorithm based on Latent Semantic Analysis[C]// Proceedings of the Acoustics, Speech, and Signal Processing. IEEE Computer Society, 1996:172-175.
[34] YUTAKA M, TAKESHI S, KOKI U, MITSURU I. Graph-based Word Clustering using a Web Search Engine[C]// Proceedings of Empirical Methods on Natural Language Processing. Association for Computational Linguistics, 2006:542-550.
[35] ANTONIO S,. Word Clustering[R/OL].The EAGLES Lexicon Interest Group, 1998[2010-03-22].http://www.ilc.cnr.it/EAGLES96/rep2/node37.html.
[36] CAO G, SONG D, BRUZA P. Fuzzy K-Means Clustering on a High Dimensional Semantic Space[J]. Lecture notes in computer science, 2004.
[37] 同17，158-159.
[38] Li H, ABE N. Clustering Words with the MDL Principle[C]// Proceedings of the 16th conference on Computational linguistics. ACL, 1996:4-9.

**作者简介**

郭怀思，中国科学技术信息研究所硕士生，研究方向为词聚类和聚类可视化。通讯地址：北京市复兴路15号 中国科学技术信息研究所研究生部 100038。E-mail: job-green@163.com
徐硕，男，博士，目前于中国科学技术信息研究所从事博士后科研工作。研究方向：数据挖掘，信息抽取，生物信息等。通讯地址：北京市复兴路15号 中国科学技术信息研究所信息技术支持中心 100038。E-mail: xush@istic.ac.cn
朱礼军，男，副研究员，硕士生导师。2005年起，在中国科学技术信息研究所信息技术支持中心从事一线科研工作。目前主要研究Semantic Web、Web service和知识技术在科技信息服务、电子政务/商务中的应用以及知识组织系统的集成与服务体系。通讯地址：北京市复兴路15号 中国科学技术信息研究所信息技术支持中心 100038。E-mail: zhulj@istic.ac.cn

## A Survey on Word Clustering Technique

Guo Huai'en, Zhu Lijun, Xu Shuo / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Word clustering is a word-oriented clustering technique, which is widely applied in a number of NLP tasks. This survey paper provides a categorization of some of the existing word clustering methods .

Keywords: Word clustering, Grammatical feature, Semantic feature, Pragmatic feature

（收稿日期：2010-02-28）