汉语名词和形容词的聚类算法研究*

王宁[1]　苑春法[1]　黄昌宁[2]

[1]智能技术与系统国家重点实验室

清华大学计算机科学与技术系，北京 100084

[2]微软中国研究院

Email: cfyuan@tsinghua.edu.cn

摘　要：以《现代汉语辞海》中的搭配对为原始数据，根据词之间的组合搭配关系计算词之间的距离，本文提出了对汉语名词和形容词聚类的多种算法并进行了实验。为了评价各种聚类算法的优劣，对实验结果进行了认真的分析比较，并构造了聚类的评价函数。

关键词：聚类　搭配对　搭配度

A Study on Clustering Algorithms of Chinese Adjectives & Nouns

WANG Ning[1]　YUAN Chunfa[1]　HUANG Changning[2]

[1]State Key Laboratory of Intelligent Technology and System

Dept. of Computer Science & Technology, Tsinghua University, Beijing 100084

[2]Microsoft Research China

Email: cfyuan@tsinghua.edu.cn

ABSTRACT: Based on the collocations of a thesaurus <Xiandai Hanyu Cihai> and the distance between words, some different clustering algorithms of Chinese adjectives & nouns are proposed and tested. In order to comment all of these clustering algorithms and recognize the best one finally, the experiential results of the different algorithms are analyzed and compared seriously, and an evaluation function is also constructed.

Keywords: clustering; compositional pairs; collocation degree.

1 引言

词汇间的组合搭配关系是指两个或者多个词语能否组成一个有意义的词组或短语。词汇之间的组合搭配关系有着一定的规律，往往表现为一部分词和另一部分词之间存在着搭配关系。因此，对词汇间的组合搭配关系的研究可归结为寻找具有共同搭配属性的类，然后研究

*自然科学基金(69773031) 和 973（1998030507）资助项目

类和类之间存在的搭配关系。

当前关于基于组合的类的研究大致可以归为两种方法，一种是利用现有分类体系，如利用义类词典，对词进行分类。但是事实表明，具有相同义类的词的组合能力并不一定相同，可能只是义类中的某一部分具有相同的组合能力。而且，现有的分类体系并不完备。另一种研究方法是根据词语的分布环境进行聚类，是基于统计的方法。近年来，研究者在这方面做了不少卓有成效的工作（Brown et al. 1992，Pereira et al. 1993，李涓子 等 1997，闻扬 等 2000）。

本文的研究工作也是基于词语的分布环境的。以《现代汉语辞海》中的搭配对为原始数据设计和使用多种聚类算法，对名词和形容词进行了聚类，并对聚类结果进行分析评价。

2 问题的描述

当我们研究形容词的聚类时，可以把问题归结为寻找形容词集合的一个分割，这个分割的每一个子集都具有相同的和名词组合的能力。此时名词是形容词的分布环境。反之亦然。

[定义 1] 分割

假设 U 是一个非空有限集，U_1，U_2，…，U_k 是 U 的子集，如果满足：

1） 对于任意的 i 和 j，$i \neq j$；$U_i \cap U_j = \Phi$

2） $U = \bigcup_{1 \le i \le k} U_i$

则称<U_1，U_2，…，U_k>是 U 的一个分割。

[定义 2] 距离

- 距离 1：

记 A 为所有形容词的集合，N 为所有名词的集合。对于 N 中任意的两个名词 n1 和 n2，设 W1 和 W2 是分别和 n1 和 n2 有搭配关系的形容词集合，则这两个名词之间的距离为：

$$DIS\_W(n1,n2)=1-|W1 \cap W2|/|W1 \cup W2|$$

- 距离 2：

记 A 为所有形容词的集合，N 为所有名词的集合。对于 N 中任意的两个名词 n1 和 n2，设 C1 和 C2 是分别和 n1 和 n2 有搭配关系的形容词的语义代码的集合，则这两个名词之间的距离为：

$$DIS\_C(n1,n2)=1-|C1 \cap C2|/|C1 \cup C2|$$

同样，我们可以给出有关形容词距离的定义。

[定义 3] 加权距离

由于《现代汉语辞海》存在着数据稀疏问题，不少情况下如果只依据搭配词集合，计算出的两个词之间的距离较真实值偏远，导致本来应归为同一类的词没有被归入同一类。因此在计算距离时同时考虑搭配词的语义代码集合是很有必要的。

因此，考虑了距离 2 的影响，修改后的加权距离公式为：

DIS(n1,n2)=plus×DIS_W(n1,n2)+(1-plus)×DIS_C(n1,n2)

其中 plus 是加权系数，取值范围(0,1)，需要在实验过程中进行调整。

[定义 4] 搭配度

记 CA 为聚类后所形成的一个形容词类，类成员包括<a1,a2,…>，共 p 个；CN 为聚类所形成的一个名词类，类成员包括<n1,n2,…>，共 q 个；C 表示所有实际存在的搭配对的集合；M(CA,CN)表示 CA 和 CN 的搭配度；则：

$$M(CA,CN)=\frac{\left|\{(ab)\mid a\in CA,b\in CN,ab\in C\}\right|}{p\times q}$$

[定义 5] 评价函数

为了从总体上评价聚类结果，本文初步提出如下的评价函数：

$$P=\frac{\sum M_i^2}{\sqrt{Num_a\times Num_n}}$$

其中 P 表示评价函数值，该值越高意味着聚类效果越好；Num_a和 Num_n分别表示聚类结果中形容词的类数和名词的类数；M_i表示各搭配度的值。

用搭配度的平方和而不是直接求和，是为了在搭配度之和相等的情况下区分聚类效果仍存在着的优劣。假设有两种聚类结果，每种结果都得到两个搭配度，第一种结果的搭配度为 a 和 b，第二种结果的搭配度为 c 和 d，并且有 a+b=c+d 和|a-b|>|c-d|。很容易可以推出 $a^2+b^2>c^2+d^2$。前者聚类效果优于后者，相应评价函数值也高于后者。因此该评价函数在搭配度之和相等的情况下是具有区分能力的。举一个实际例子，假设名词被聚成了两类，记为 N1 和 N2；形容词也被聚成了两类，记为 A1 和 A2。一个理想的结果是：A1 和 N1，A2 和 N2，A1 和 N2，A2 和 N1 的搭配度分别为 1,1,0,0。而一个不够理想的结果是没有得到明确的划分，比如上述四个搭配度都是 0.5。这样则有 1+1+0+0=2；0.25+0.25+0.25+0.25=1；2>1。由此可见本文提出的评价函数是实际有效的。

3 聚类算法设计

对于同一种词性的词，知道了其中任意两个词之间的距离，也就相当于知道了它们在空间里的相对位置分布。基于分布情况，本文设计了多种算法对词进行聚类。其中第一种算法是别人提出的算法：（李涓子 等 1997）。

[算法一] 基于点之间的最大距离

1. 从某种词性的所有词的空间点集 X 中取一个点 x1 作为第一类的中心；
2. 取 X 中距离 x1 最远的点 x2 作为第二类的中心；
3. 对 X 中剩余的每个点 xi，分别计算到各类中心的距离，令其小者为 Dxi；
4. 找出 Dxi 中最大的，如果大于某个阈值，则取该点为新的中心；
5. 反复进行 3-4，直到找不到新的中心点；

6. 其余的点作为非中心点，根据到各中心的距离，归入最近的一个类中。

[算法二] 用到类的平均距离代替到点的距离

1. 从某种词性的所有词的空间点集 X 中取一个点 x1 作为第一类的中心；
2. 取 X 中距离 x1 最远的点 x2 作为第二类的中心；
3. 对于其余的每个点 xi，计算它到各个类 Sj 内所有点的平均距离 Dij，求出 Dij 的最小值 Dik，若大于阈值则把它作为新的中心点，否则把它归入和它平均距离最近的类 Sk；
4. 重复进行 3，直到所有点都被归入了某一类。

[算法三] 基于分布密度

1. 计算每个点周围一定半径内的点的密度，将密度大于阈值的点定为中心点；
2. 以各个中心点为球心，将一定半径内的点聚成一类；
3. 对于还没有被归入任何一类的点 xi，计算它到各个类 Sj 内所有点的平均距离 Dij，求出 Dij 的最小值 Dik，若大于阈值则把它作为新的中心点，否则把它归入和它平均距离最近的类 Sk；
4. 重复进行 3，直到所有点都被归入了某一类。

[算法四] 逐层扩展聚类

1. 计算每个点周围一系列半径(r1,r2,r3,…)内的点的密度(m1,m2,m3,…)，将指定密度 mk 大于阈值的点定为中心点；
2. 以各个中心点为球心，从一个最小的半径开始以环的形式向外扩展，半径逐渐扩大，观察每环的密度，当密度出现突然下降时停止扩展，以当前半径聚成一类；
3. 对于还没有被归入任何一类的点 xi，计算它到各个类 Sj 内所有点的平均距离 Dij，求出 Dij 的最小值 Dik，若大于阈值则把它作为新的中心点，否则把它归入和它平均距离最近的类 Sk；
4. 重复进行 3，直到所有点都被归入了某一类。

4 实验结果与分析

- 实验结果

本文所使用的初始数据是《现代汉语辞海》，它包涵了五十多万条多种词性的词语之间的搭配对。我们从中将名词与形容词的搭配对(包含 4536 个名词和 2569 个形容词，共 37346 个搭配对)抽取出来，建立 A-N 数据库，在这个数据库上进行聚类。

聚类算法	基于最大距离	用到类的平均距离代替到点的距离	基于分布密度	逐层扩展聚类
名词聚类数目	608	578	310	377
形容词聚类数目	421	422	349	482
搭配度最大值	88.9%	89.5%	89.5%	89.5%
评价函数值	0.0498	0.0541	0.0510	0.0568

由于篇幅关系，这里仅列出算法四（逐层扩展聚类）的部分结果如下：

形容词：

〈超群　出众　非凡　过人　赫赫　卓越〉

〈笨拙　规范　机械　灵巧　容易　生硬〉

〈暗淡　斑驳　灰暗　昏暗　明朗　柔和　柔媚〉

〈黛绿　红彤彤　黄澄澄　绛紫　桔黄　水绿　杏红〉

名词：

〈哀乐　教训　泪水　伤　痛苦　危亡　罪孽〉

〈匕首　笔锋　刺刀　钉子　锋芒　剪刀　剑锋　箭头　镰刀　杂文　针头　锥子〉

〈刀枪　步法　技法　技艺　枪法　手艺　译笔　章法〉

〈彩霞　朝霞　晨曦　春光　春天　风光　风景　光景　景色　景物　色泽　夜景〉

- 评价与分析

第一种算法（基于最大距离）由于第一个中心点的选取是随机的，因此不能保证选取的第一个中心点在实际分布中是一个中心点，而其它中心点的选取都是建立在第一个中心点的基础上的，这样如果第一个中心点不准确的话，可能会导致聚类结果的不准确。

第二种算法（用到类的平均距离代替到点的距离）是第一种算法的改进型。通过对算法一的结果进行分析，发现经常出现下面的情况：某个非中心词 X 由于和类 P（P 由聚类形成）的中心词 Y 距离不够接近，没有被归入类 P；而实际上 X 和类 P 中其他词距离都很近，应该被归入类 P。因此在算法二中当判断一个非中心点是否应该归入某一类时，考察它到该类内的所有点的平均距离，也就是说，以它到该类的距离而不是到该类的中心点的距离作为它是否归入该类的依据。因此，算法二的聚类效果要优于第一种算法，但是它仍然不能解决第一个中心点的选取问题。

第三种算法（基于分布密度）根据点在空间的相对位置分布，将密度大的那些点定为中心点，然后以各中心点为球心以一定半径进行聚类。未被归入任何一个球体的散点最后处理，若它们距离各类都很远则单独成类，否则归入最近的类中。这样的中心点选取策略在直观上是优于基于距离的中心点选取方案的，在实际中通过考察聚类结果发现优势并不明显。研究发现原因是这些点在空间中的分布并不是等体积的多维球体，也就是说，聚类形成的类的大小相互间可能有很大的差异，比如有些类可能包含上百个词，有些类可能只有几个词。

第四种算法（逐层扩展聚类）是在第三种算法的基础上进行改进而设计的，对每一个点以多半径逐层扩展的办法来考察其周围点的密度并进行聚类，效果比第三种算法略好一些，但是调试和运行的空间和时间开销都非常大。

5 问题与讨论

- 关于原始数据

原始数据存在着严重的数据稀疏。对于收录了 4536 个名词和 2569 个形容词的 A-N 库，仅存在搭配对 37346 条，这相当于只有不到百分之一的搭配关系。即使考虑到一些名词和形容词之间没有搭配关系，这个数据仍然很不完备。

在本文的研究过程中，主要采用的是调整加权系数的办法。在计算距离时，搭配词集合

的权重过大就使得词之间的距离过远，聚类的结果显得过于分散；搭配词的语义代码集合的权重过大就使得词之间的距离过近，聚类的结果显得过于密集。因此在计算距离时，调整两个距离分量的权重，以寻求一个更接近实际情况的值。

另外还可以根据聚类结果来对原始数据进行增补。当聚类完成之后，名词将被划分为一些类<N1, N2, N3, ….>，形容词也将被划分为一些类<A1, A2, A3, ….>。假设 N2 和 A3 的搭配度比较高，那么很有可能 N2 内的词和 A3 内的词所构成的在原始数据库中不存在的搭配对也是实际存在的，因此可以考虑在辅助人工判断的情况下，向原始数据库中添加搭配对。这种方法可以解决一部分原始数据的数据稀疏问题，但是由于聚类结果本身是基于一个数据稀疏的原始数据得到的，因此效果有限。

不论采用什么办法来弥补原始数据的不足，数据稀疏的问题都是无法完全解决的，其不良影响在聚类结果中有着直接的体现。原始数据是研究工作的基础，因此期望以后能够得到一个更加完备和准确的数据源，以便更好的进行研究。

- 关于算法

聚类的基础是词和词之间的距离，目前使用的距离公式并不见得是最优的，在今后的工作中可以考虑改进或者改变。聚类算法是多种多样的，本文只提出了比较有代表性的几种算法，在今后的工作中可以尝试构造其它的聚类算法。评价函数目前还只是处于初步研究阶段，有待于在今后的工作中继续完善和改进。

6 结束语

汉语中词的聚类问题的研究有着重要的意义，是自然语言处理研究范围里的一个基础研究课题。本文基于词之间的搭配实例，设计了多种聚类算法对汉语中名词和形容词的聚类问题进行了大量的实验和研究，并依据聚类结果对所使用的算法进行了比较和分析，并构造了一个评价函数对几种算法进行了评价。结果表明，在我们的聚类环境下，第四种聚类算法明显优于前三种算法。

参考文献

[1]倪文杰等：《现代汉语辞海》，中国人民出版社，1994。

[2]姬东鸿、黄昌宁："汉语形容词和名词的语义组合模型"，Communications of COLPS, 6(1), P25-33, 1996。

[3]李涓子、姬东鸿、黄昌宁："基于组合实例的双向优化聚类"，《语言工程》，清华大学出版社，P164-169, 1997。

[4]闻扬、苑春法、黄昌宁："基于搭配对的汉语形容词-名词聚类"，中文信息学报，第十四卷第六期，P45-50, 2000。

[5]Peter F.Brown et al.："Class-based n-gram Models of Natural Language"，Computational Linguistic，1992。

[6]F.Pereira，N.Tishby，L.Lillian："Distributional Clustering of English Words"，Proceedings of the 31st Annual Meeting of ACL，P183-190，1993。