基于组合实例的双向优化聚类

李涓子 姬东鸿 黄昌宁
(清华大学计算机系 北京 100084)

摘 要： 本文探讨根据组合实例对不同词性的词同时进行聚类的问题，在聚类过程中，不同词性的词的聚类相互影响。首先将该问题转化为一个基于组合实例的优化聚类问题，以类内平均距离为基础构造目标函数，提出了一个双向优化聚类的算法。我们将此算法应用于基于汉语形容词和名词的组合实例，对形容词和名词同时进行聚类，结果表明算法是有效的。

关键词： 词汇组合 聚类 双向优化聚类 最大距离聚类

Two-dimensional Clustering Based on Compositional Examples

Li Juanzi Ji Donghong Huang Changning
(Intelligent Technology and System Lab. of Tsinghua University)

Abstract: The paper discusses the problem of simultaneous clustering words of different categories according to compositional examples. During clustering, the interactions between diverse categories is considered. The problem is changed to some sort of optimized clustering problem based on compositional examples, the solution of which is two-dimensional clustering algorithm with the objective function of average distance within classes. The further application to clustering some words of adjectives and nouns demonstrates the algorithm is effective.

Key words: words composition, clustering, tow-dimensional optimized clustering, maximal distant clustering.

一、引言

近年来，基于大规模真实文本的语料库语言学研究已经成为计算语言学研究的重点，而从大规模语料库中获取各种对语言分析有用的知识更是许多计算语言学工作者的核心研究课题。在知识获取中获取词汇间的组合关系显得尤为重要。

词汇间的组合关系是指两个或多个词语能否组成一个有意义的短语。从大规模语料库中可以学习到词的组合关系，但组合关系的表示若按具体词来进行，必然存在数据量大、占据空间多的缺点，再者词汇之间的组合关系有一定规律可循，往往表现在一部分词和另一部分词的组合上，所以，词汇组合的研究可归结为词中具有某些性质的类之间的组合关系的研究。

当前关于类的研究大致可归为两种方法，一种是利用现有分类体系，如义类词典。即若词 A 和词 B 之间具有组合关系，则所有与 A 属同一义类的词和所有与 B 属同一义类的词也应具有同种组合关系。但事实表明，具有相同义类的词其修饰或组合能力并不一定相同，可能

只是义类中的一部分具有相同的能力。另一种研究方法是根据词语的分布环境进行的，是基于统计的方法。近些年来，研究者在这方面做了许多卓有成效的工作(Brown et al. 1992，Pereira et al. 1993)。这些聚类都是单向聚类的方法，即根据某一词性词的分布环境，利用信息论中互信息的概念直接对其进行聚类。

实际上，具有组合关系的不同词性之间的聚类是有一定联系的，不以能单纯对某一词性中的词进行聚类，而在其聚类过程中应同时考虑与之相关的另一种词性的聚类。我们针对汉语中的A—N结构，对名词和形容词聚类的方法进行了研究。研究发现，名词的聚类与形容词的聚类有关，反过来形容词的聚类与名词的聚类有关，名词的分布环境应是能与其组合的形容词的类集，而形容词的分布环境是可与之组合的名词的类集。因此根据分布环境对形容词和名词进行聚类，是一个非线性问题，这两个词性间的聚类过程是有相互作用的。

本文基于上述思想，提出了利用组合实例进行双向优化聚类算法的思想，即在聚类过程中考虑到两类词之间的组合关系，同时对这两类词进行聚类。在下文中我们将对这种算法进行描述，并介绍算法应用及其实验结果。

二、基于组合实例的双向优化聚类算法

该问题可一般化表示为：给定一个形容词集合A和一个名词集合N，假设对于A中任一形容词，以可被该形容词修饰的名词为其组合实例，对于N中任一名词，以可以修饰该名词的形容词为其组合实例，算法的目标就是利用这些组合实例先根据已有的较好的聚类算法分别建立形容词和名词的分类，以此作为双向优化聚类的初始条件，再使用双向优化算法对初始分类进行进一步优化调整，得到新的更合理聚类结果，使得每一类的形容词或名词既有相同或相近的意义，又有相同或相近的分布功能，得到更好地表示词间的组合关系的组合框架。

2.1 问题的形式化表示

为了更好地描述算法，先给出一些与算法有关的定义

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，如果满足下述条件：

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。

对于A和N，要首先分别确定满足上述条件的分割〈A_1，A_2，…，A_k〉和〈N_1，N_2，…，N_k〉，然后任何两个形容词和名词的之间的距离可以由其修饰的名词和被修饰的形容词所属的类号确定。

定义2：名词间的距离

设〈A_1，A_2，…，A_k〉是A的一个分割，对于N中任意的n_1和n_2，设w_1和w_2为修饰n_1和n_2的形容词所属的类号集，则这两个名词之间距离为：

$$dis_A(n_1,n_2) = 1 - \frac{|w_1 \cap w_2|}{|w_1 \cup w_2|}$$

同样，我们也可以利用被修饰的名词所在的类号定义 A 中任意的 a_1 和 a_2 之间的距离 $dis_N(a_1, a_2)$。

由上定义可以看出名词间的相似性与形容词的分类有关，而且形容词间的相似性也与名词的分类有关，这恰好反应出名词和形容词类的之间相互作用和影响。

定义 3：双向优化聚类的目标函数Φ(P)

设 P=A ∪ N， A=〈 A_1， A_2，…， A_k〉及 N=〈 N_1， N_2，…， N_k〉，则可定义：

$$\phi(\rho) = \sum_{C \in A} \left[\frac{1}{\left(\frac{|C|}{2} \right)} \sum_{x,y \in C_1} dis_N(x,y) \right] + \sum_{C \in N} \left[\frac{1}{\left(\frac{|C|}{2} \right)} \sum_{x,y \in C} dis_A(x,y) \right]$$

从公式看出，Φ(P)为 A 和 N 类内平均距离的累加和。因此，双向优化聚类思想是从 A 和 N 出发不断调整 A 和 N，使Φ(P)达到最小。即产生出优于 A 和 N 的更适于描述 A 和 N 组合关系新的分割 A'和 N'。

2.2 算法形式化描述

双向优化聚类过程可描述为：

1. 根据 A 和 N，对 N 中任意的 n_1 和 n_2 及 A 中任意的 a_1 和 a_2，由定义 2 计算 $dis_A(n_1, n_2)$ 及 $dis_N(a_1, a_2)$；
2. 由定义 3 计算Φ(P)；
3. 对任意 n_1，计算将它从所在类移入任一其它类时Φ(P)的变化。找出使Φ(P)减少最多的 n，将变化计入$\Phi_1(P)$；
4. 对任意 a_1，计算将它从所在类移入任一其它类时Φ(P)的变化。找出使Φ(P)减少最多的 a，将变化计入$\Phi_2(P)$；
5. 若$\Phi_1(P)$〈 $\Phi_2(P)$ 〈Φ(P)，则确定移 n，否则确定移 a；
6. 移动后重新计算 $dis_A(n_1, n_2)$和 $dis_N(a_1, a_2)$及Φ(P)；
7. 重复 3—6，直到找不出一个移动使Φ(P)减少为止。

注：在考虑移动时，若要移入词与类内其它词的距离大于某一阈值，则不考虑该词的这次移动。

由此最终得到的 A'和 N'分类是计算了 A 与 N 之间相互作用，使其类内平均距离之和为最小的分类。所以，优化结果优于原分类 A 和 N。

三、双向优化聚类应用实例

3.1 应用背景

我们以存入计算机的《现代汉语辞海》这一凝聚了许多语言学家语言知识的语料库为聚类资源，对此算法进行了实验。《现代汉语辞海》中，共收录了近 8000 个常用汉语词汇，近 70 万个组合实例，其中名词近 3000 个，形容词近 2000 个，对于每个词条，都尽可能多地收录了可与之组合的词语。因此《现代汉语辞海》可以作为描述词汇间组合关系的一种资

源。但是这种策略的弱点是其所例举的实例并不能保证其组合的完备性，我们则试图从这些组合
实例出发，对名词和形容词进行聚类以弥补上述缺憾。

作为小规模实验，我们从《现代汉语辞海》中选取了较常用的 30 个名词，并根据修饰该名词的形容词收集了 43 个形容词，分别组成将被聚类的名词集合 N 和形容词集合 A 。其中 N={友谊，友情，田野，感情，原野，爱情，技术，心情，神色，情绪，阳光，春光，春色，情谊，生活，技巧，青春，年华，身体，身子，信心，信念，心境，心灵，心胸，任务，毅力，意志，性格，神情}，

A={真挚,诚挚,宝贵,可贵,美好,美妙,迷人,珍贵,广阔,宽广,美丽,可爱,宽阔,健康,愉快,痛苦,沮丧,苦闷,郁闷,烦闷,疲惫,明媚,欢乐,悲伤,熟练,娴熟.悲痛,不安,懊悔,懊恼,紧张,乐观,疲劳,疲乏,疲倦,明丽,困难,艰巨,坚定,坚强,顽强,艰难}。以 A 和 N 作为被聚类的对象，以 N 的各词条中可以被 A 中修饰的形容词和 A 的各词条中可以修饰 N 中的名词作为聚类时所需要的知识。

3.2 A 和 N 的初始聚类

在进行双向优化聚类之前，要对 A 和 N 进行初始分类，初始分类的好坏对以后的优化聚类有较直接的影响，因此为了保证双向优化聚类的质量，我们使用最大距离聚类法分别对 A 和 N 进行初始聚类。

对 N 中任一 n，以 A 中任一 a_i 与 n 是否具有组合关系为 n 的特征向量。对 A 中任一 a，以 N 中任一 n_i 与 a 是否具有组合关系为 a 的特征向量。

N(A)的聚类过程为：

1. N 中任取一 $n_1(a_1)$作为一类的中心 $n_1(a_1)$；
2. N 中取出到 $n_1(a_1)$距离最大的 $n_2(a_2)$作为第二类中心 $n_2(a_2)$；
3. 对 N 中剩余的每个 $n_i(a_i)$，分别计算到各类中心的距离，令其较小者为 $Dn_i(Da_i)$；
4. 计算 $\max_{n_i \in N}\{Dn_i\}$ $(\max_{a_i \in A}\{Da_i\})$，若其值大于等于某个阈值 t，则取该词为新类中心。
5. 反复进行 3 ～ 4，直到找不到符合上述条件的新类中心。
6. 把剩余的 $n_i(a_i)$作为非中心词，根据它到各类中心的距离，选最近的一个作为其所属的类。

用此算法其分类结果为：N 分为 8 类，A 分为 18 类。其中 N={{友谊 友情 感情 爱情 情谊 年华}，{心情 情绪}，{生活}，{田野 原野 阳光 春光 春色 心灵 心胸}，{信心 信念 毅力 意志 性格}，{神色 心境 神情}，{青春}，{身体 技术 技巧 身体 身子 任务}}；A={{真挚 诚挚}，{迷人}，{沮丧 苦闷 不安 懊悔 懊恼 紧张}，{美好}，{坚定 坚强 顽强}，{美丽}，{健康}，{郁闷 熟练 娴熟 疲倦 艰巨}，{美妙}，{广阔 宽广 宽阔}，{可爱}，{痛苦 悲痛 困难 艰难}{疲惫 疲劳 疲乏}，{宝贵 可贵 珍贵}{愉快 欢乐}，{明媚 明丽}{悲伤}，{烦闷 乐观}}

上述分类已初具雏型，具有一定的可靠性，但有个别词其所在的类有错。

3.3 A 和 N 的双向优化聚类

应用 2.2 双向优化聚类算法，对 3.2 中的结果进行优化，优化后的结果为：A={{友谊 友情 感情 爱情 情谊 年华 }，{心情 情绪}，{生活 青春}，{田野 原野 阳光 春光 春色 心灵 心胸}，{信心 信念 毅力 意志 性格}，{神色 心境 神情}，{技术 技巧}，{身体 身子}，{任务}},A={{真挚 诚挚}，{ 迷人 美好 美妙 美丽}，{愉快 沮丧 苦闷 不安 懊悔 懊恼 紧张}，{坚定 坚强 顽强 }，{ 健康}，{郁闷 疲倦 艰巨}，{熟练 娴熟}，{ 广阔 宽广 宽阔 明媚 明丽}，{可爱 }，{痛苦 悲痛 困难 艰难}，{疲惫 疲劳 疲乏}，{宝贵 可贵 珍贵}，{欢乐}，{ 悲伤 烦闷 乐观}}

从上表中可以看出该算法对 N 和 A 都进行了一定程度上的优化。

3.4 结果评价

从聚类结果上我们发现，第一：属于同一类的词确实在组合能力上具有一定的相似性。由于此聚类方法属于分布聚类，因此在一类中的词在意义上有些是相似的，而有些却截然不同。如“田野”、“心胸”与“春光”等被聚为同一类，虽然它们在意义上不同，但是它们在分布上具有一定的相似性。第二：由聚类的结果，我们可以认为同一类的词具有相似分布功能，因此，可以发现许多新的组合实例，如“田野”的组合实例中没有“宽阔”的组合实例，但因为“田野”与“原野”聚为同一类，则可以学到“宽阔的田野”这一组合实例。

为了对聚类结果进行量纲上的评价，我们初步采用的方法是将算法产生的结果同人进行分类的结果进行比较。为了避免个人的主观因素，可以同时组织多人对 30 个名词和 43 个形容词进行分类，要求分类的个数一定 。以此在估算出及其分类的准确率，我们的实验结果的准确率约为 73.4%。但当词的个数较多时，这种方法难以应用，有关聚类结果的评测方法还有待于进行进一步的研究。

四、结束语

4.1 算法特色

1. 提出一种根据形容词和名词的组合实例，对形容词和名词同时进行聚类的策略。该策略并不局限于 A － N 组合模式，可应用于其它组合模式中，例如 V － N 组合模式。
2. 将聚类问题转化为一个优化问题，并给出一种基于组合实例，以类内平均距离之和为目标函数进行双向优化聚类的算法。

4.2 存在问题及进一步的研究

1. 双向优化聚类对初始分类进行优化时，只可能减少类的个数，并没有考虑在优化过程中类的分裂与合并问题。
2. 以《现代汉语辞海》中形容词和名词的组合实例作为它们聚类的基础，存在数据稀疏问

题，改进设想是以《现代汉语辞海》为基础资源，再从大规模真实语料中获取更多的有关形容词和名词的组合实例，以确保聚类的质量。

3. 研究聚类结果的评测算法。

4.3 聚类结果的应用领域

1. 得到名词和形容词的语义组合框架。以此描述名词所具有的各种可能属性，此结果类似于动词的格框架。
2. 在一定程度上解决形容词和名词在句法或语义上的组合歧义问题。如对 $A+N_1+N_2$，通过 N_1 和 N_2 的组合框架，可以确定该结构为（$A+N_1$）$+N_2$ 还是 $A+$（N_1+N_2）。

参 考 文 献

[1] 姬东鸿，黄昌宁，汉语形容词和名词的语义组合模型，《Communications of COLPS》，1996，1。

[2] 倪文杰等，《现代汉语辞海》，人民中国出版社，1994。

[3] 边肇祺等，《模式识别》，清华大学出版社，1988。

[4] Hermann Ney, Ute Essen & Reinhard Kneser, On sturcturing probabilistic dependences in stochastic language modelling, Computer Speech and Lunguage, 1994, 8.

[5] Peter F. Brown et al., Class-based n-gram Models of Natural Language, Computational Linguistic, 1992, 11.

[6] V. Hatzivassiloglou & K. R. Mckeowen, Towards the Automatic Identification of Adjectival Scales: clustering of Adjectives According to Meaning, In Proceedings of 31st Annual Meeting of ACL, Columbus, Ohio, USA.

[7] F. Pereira, N. Tishby, L. Lillian "Distributional Clustering of English Words", Proceedings of 31st Annual Meeting of ACL, Columbus, Ohio, USA, 183-190, 1993.