基于量词的名词概念获取研究*

王萌[1]，俞士汶[2]

（1.江南大学人文学院，江苏 无锡 214122；2. 北京大学计算语言学教育部重点实验室，北京 100871）

摘要：概念获取是自然语言理解领域中重要的研究课题。本文提出了一种基于汉语量词的名词概念描述方法，设计并实现了一个权重计算方案。通过聚类实验探索了量词对名词语义区分的作用和贡献，实验结果表明基于量词的名词概念表达方式是有效的，可以区分大部分名词概念。

关键词：概念获取；量名搭配；量词；聚类

中图分类号：TP391　　**文献标识码：**A

Concept Acquisition Based on Chinese Measure Words

Meng Wang[1]，Shiwen Yu[2]

（1.School of Humanity, Jiangnan University , Wuxi, Jiangsu, 214122, China；2. Key Laboratory of Computational Linguistics, Peking University, Beijing, 100871, China）

Abstract： Concept acquisition from corpora has become increasingly important in NLP. This paper presents a new concept representation based on measure words. Concepts are modeled as vectors with one component corresponding to each measure word. We propose a weighting scheme that assigns each measure word a weight in a concept. Then we conduct experiments to identify concept similarities via clustering, and the results show measure words can categorize most concept classes.

Key words： Concept acquisition; classifier-noun collocation; measure words; cluster

1 引言

概念获取（Concept Acquisition）又称概念学习（Concept Learning），是自然语言理解领域中重要的研究课题，获取的结果可以直接应用于词汇知识库或者领域本体的建设，也可服务于信息检索（Information Retrieval）、词义消歧(Word Sense Disambiguation)等应用研究。目前大多数研究工作，是基于语料库或者Web自动获取概念的相关属性信息或语法关系，然后通过聚类的方法来区分概念。Grefenstette提出用语料中自动获取的语法关系（grammatical relations）来表达概念[1]，如动词-宾语（verb-object）、主语-动词（subject-verb）、名词-名词（noun-noun）和形容词-名词（adjective-noun）等。Lin将名词概念“狗 dog”表示为<语法关系, 值>对[2]，如<adj-mod brown>（意思为，“狗”的形容词修饰语为“brown”），获取这些语法关系要求语料是经过句法分析的。在基于Web的方法中，研究者定义若干模板以获取概念的各种属性值。例如，Almuhareb提出用模板“the * of the C [is|was]”在Web上收集概念C的属性信息[3]，如“the price of the car was”，则得到price是car的一个属性，在此基础上，通过获取的属性来定义概念。上述这些概念表示方法，向量维数通常比较大。

本文提出了一种基于汉语量词的名词概念表示方法。该方法涉及的信息容易从语料中获得，而且向量维数较低易于计算。实验结果表明基于量词的名词概念表达方式是有效的，可以区分大部分名词概念。

2 量名搭配的特点

早有研究者指出汉语量词的主要功能是对名词的语义做分类[4]，Huang 利用 Shannon 熵

*收稿日期：2014-5-28　定稿日期：2014-7-22

基金项目：国家自然科学基金“面向语义检索的汉语名名组合自动释义研究”(No.61300152)

计算名词量词搭配的信息含量[5]，推导出一个大致可行的汉语名词语义分类架构。考察量名搭配的特点，量词所修饰的名词通常是有限的（“个”是使用最为广泛的量词，几乎所有的个体名词都能论“个”），只能与某些类名词搭配，而这一类中的名词通常会有一些相近的特点。例如，量词“条”，可以说“一条鱼”、“一条蛇”及“一条河”，“条”隐含了其修饰的名词是具备“长”、“柔软”以及“像绳子一样”等特征的物体；量词“张”则隐含了物体有延展的平面，如“一张桌子”、“一张床”等。此外，还有一些约定俗成的搭配，例如“一把扇子”、“一把锁”等。

名词也可有不同的量词和它搭配，不同量词反映了名词不同方面的特征，例如，名词“布”，既可以用“块”修饰，表现了布的局部特征，又可用“匹”修饰，表现的是较大整体的卷状形态。

量名搭配的特点启发我们，既然量词可以指示名词的某种语义特征，那么是否与某个名词搭配的所有的量词的组合能够在一定程度上描述该名词的语义？换句话说，量词能否作为名词概念的一种描述方法为名词的概念区分提供有效的信息？为此，本文设计以下实验，首先，从《中文概念词典》[1]（CCD，Chinese Concept Dictionary）中抽取了 11 个不同语义类的名词，共 875 个，用向量空间模型（VSM，Vector Space Model）将每个名词表示为向量，向量中的每一维对应一个量词；其次，提出了一个权重计算方案，为每个名词的向量赋值；最后，通过聚类的方法来计算名词之间的语义相似度，并对聚类结果进行评测。

3 基于量词的名词概念描述

向量空间模型的提出源自信息检索领域，它与布尔模型（Boolean Model）和概率模型（Probability Model）同属于该领域的代表性文本检索模型，而向量空间模型是最有效的文本表示模型之一。一个向量空间是由一组线性无关的基本向量组成，向量维数与向量空间维数一致。向量空间模型具有表示简洁和计算简便的特点，可以利用空间相似性来逼近语义相似性。本文采用该模型来表示名词概念,每个名词被表示为由所有量词构成的向量空间中的一个点，下面介绍权重计算方法，即如何给每个向量赋值。

假设 C 是名词概念集合，M 是与 C 构成搭配的所有量词集合。通过以下三步来计算向量每一维的权重。

1. 量词频次（Measure word Frequency）

量词频次是指量词 m 与名词概念 c 之间的共现次数，记为 $\mathbf{mf}_{m,c}$。根据从语料中获取的量名搭配对，计算每一对量名搭配的出现次数，即可以得到量词频次。

2. 信息容量（Information Load）

对集合 M 中的每一个量词 m，如果它可以和集合 C 中的 n 个名词搭配，按照公式（1）计算其信息容量，记为 $\mathbf{il}_m$。

$$\mathrm{il_m} = \log \frac{\|C\|}{n} \qquad \text{（公式 1）}$$

信息容量表示了一个量词的区分能力[5]，反映了量词的特异性。一个量词的信息容量的值越高，说明该与该量词搭配的名词越少，更具有区别性。反之，信息容量小说明该量词更为通用，能与更多的名词搭配。如果一个量词可以和集合 C 中的所有名词搭配，那么 n=||C||，它的信息容量为 0。

3. $\mathtt{mf\text{-}il}_{m,c}$权重方案（$\mathtt{mf\text{-}il}_{m,c}$ weighting scheme）

量词频次是一个局部信息，它反映了一个量词在某个名词概念中的“密度”，而信息容量是一个全局信息，它反映了一个量词在整个名词概念集合中的“稀有”程度。综合考虑二

---

1 CCD 是基于 WordNet 构建的，它根据汉语的特点，继承并优化了 WordNet 的语义分类体系，为中文选择倾向的研究提供了基础，本文采用的是 2006 年的版本。

者，将局部信息和全局信息结合起来，构成最终的权重计算公式，如公式（2）所示：

$$\mathbf{mf\text{-}il}_{m,c} = \mathbf{mf}_{m,c} \times \mathbf{il}_{m} \qquad \text{（公式 2）}$$

对公式（2）中的 $mf_{m,c}$ 进行归一化，如公式（3）所示，其中 mf_{max} 是与名词概念 c 搭配的量词中的频率最大值。

$$mf_{m,c} = \begin{cases} 0.5 + 0.5\dfrac{mf}{mf_{max}} & (mf \neq 0) \\ 0 & (mf = 0) \end{cases} \qquad \text{（公式 3）}$$

值得注意的是，由于数据稀疏问题，一些合理的量名搭配并没有出现在真实语料中。例如，名词“教官”可以和量词“位、名、个”搭配，但是在语料中，“教官”只和量词“位”共现，如果仅基于语料获取的量名搭配计算向量的权值，就会造成信息缺失。因此，本文用《现代汉语语法信息词典》（GKB，Grammatical Knowledge Base）进行加 1 平滑[6]，对于那些合理但是没有出现在语料中的量名搭配，将它们的 mf 值赋为 1。

基于上述步骤，可以为每个名词概念构造一个向量，这些向量组合起来便构成了一个矩阵，其中行代表名词，列代表量词。该矩阵将作为下一步聚类实验的输入。

4 聚类方法及评价指标

4.1 聚类方法

本文使用的聚类工具是 CLUTO 2.1.2[2]，该工具实现了三种聚类算法：划分法、凝聚法和图分割法[7]。CLUTO 使用的是硬聚类算法，即每个概念只能被分配到一个类中。实验中的参数设置如下：clustering method = Repeated Bisection, similarity function = cosine, criterion function = I2，No.of classes=11。本文采用了多组参数进行实验，结果表明，在该组参数设置下得到的结果最好。

4.2 评价指标

聚类结果用以下五个指标来评价：准确率（Accuracy，A）、精确率（Precision，P）、召回率（Recall，R）、F 值和漏识率（Fallout，F）。其中，准确率（A）是聚类结果正确的概念个数占整个集合概念个数的百分比。对于系统产生的聚类结果，我们需要将它对应到已有的语义类上，对应的办法是：对于结果中的每一类，找出该类中每一个词语实际的语义类，个数最多的语义类将被分配为该结果的语义类。那么，在该类中如果一个词语的实际语义类与它所分配的语义类相同，则被看做是正确，不相同的则视为错误。在此结果上，就可以得到整个聚类结果的准确率，准确率在总体上反映了词语是否被正确地分配到所属的类中。

其余四个评价指标是从联立矩阵（Contingency Table）中计算得来的。这里需要将聚类结果转化为联立矩阵，为此本文引入“共现问题”以实现转换[3]。具体方法如下：

对集合中任何一对名词概念，提问“它们是否出现在同一类中？”，回答“是”或“否”。

那么，对于同一个“共现问题”，真实情况和系统聚类结果将会分别给出两组不同的“是-否”答案。例如，有四个概念 A、B、C、D，真实的情况是（A，B）和（C，D），即 A 和 B 同类，C 和 D 同类，系统的聚类结果是（A）和（B，C，D），那么回答共现问题，可以得到如表 1 所示的结果。通过对“是”和“否”的组合计数，就可以得到联立矩阵，见表 2

---

2 该软件网址：http://glaros.dtc.umn.edu/gkhome/views/cluto/

所示。根据联立矩阵，精确率（P）、召回率（R）、F 值和漏识率（F）的计算方式如公式（4）-（7）所示：

$$P = \frac{a}{a+b} \quad \text{(公式 4)} \qquad R = \frac{a}{a+c} \quad \text{（公式 5）}$$

$$\text{Fallout} = \frac{b}{b+d} \quad \text{（公式 6）} \qquad F\text{-measure} = \frac{2\times P\times R}{P+R} \quad \text{（公式 7）}$$

表 1 共现问题的系统答案及正确答案

问题	系统答案	正确答案
(A, B) 出现在同一类中吗？	否	是
(A, C) 出现在同一类中吗？	否	否
(A, D) 出现在同一类中吗？	否	否
(B, C) 出现在同一类中吗？	是	否
(B, D) 出现在同一类中吗？	是	否
(C, D) 出现在同一类中吗？	是	是

表 2 联立矩阵

系统答案	正确答案	
	是	否
是	a（1）	b（2）
否	c（1）	d（2）

5 实验

5.1 实验数据

首先，从中文概念词典 CCD 中抽取 875 个名词，这些名词来自 11 个不同的语义类，抽取过程是随机的。在 CCD 的上下位关系中，同一类中的名词都拥有共同的上位概念，表 3 给出了词语语义类分布及部分样例。可以看出，大多数语义类中的名词个数都大于 80，只有三个语义类：树（65）、出版物（66）和事件（43），包含名词个数少于 80 个，主要原因是属于这三个语义类的词语在语料中出现不多，观察不到足够的样例。如果语料规模扩大，可以得到更为平衡的实验数据集。

表 3 实验名词样例

语义类	个数	名词样例		语义类	个数	名词样例
动物	92	大象 老虎 骆驼 乌鸦		交通工具	90	汽车 巴士 吉普 计程车 摩托车
建筑	93	古堡 校舍 别墅 教堂		出版物	66	参考书 词典 报刊 期刊 杂志
衣物	86	牛仔裤 马甲 毛衣 棉袄 裤子		机构	82	监狱 法院 学校 医院 议院
食物	86	粥 面粉 米饭 面条 羊肉		事件	43	博览会 会议 宴会 晚宴 舞会
树	65	白杨 柏树 松树 枫树 桃树		资产	80	报酬 补贴 退休金 贷款 奖金
人	92	秘书 教授 学生 乘务员 医生				

其次，基于在 1998 年上半年《人民日报》语料中获取的真实量名搭配，共 152,430 对[8]。可以得到与这 875 个名词构成搭配的所有量词，共计 194 个，这些量词作为向量空间的每

一维。对于每个量词，计算其信息容量，表 4 给出了部分量词的信息容量值。

表 4 部分量词的信息容量

量词	信息容量 （information load）
枝	6.7742
幅	6.0811
段	4.9825
册	4.1352
所	3.6387
辆	2.8824
只	2.2203

对每一个名词，按照第三节中介绍的权重计算方法，从量名搭配表中获取量词频次，结合量词的信息容量，生成其向量表示。

5.2 实验结果

本文用 CLUTO *vcluster* 命令以及第四节中的参数设置进行实验，得到聚类结果后，通过回答“共现问题”将其转化为联立矩阵，结果如表 5 所示。五个评价指标的详细情况见表 6。准确率显示，在基于量词的名词概念描述方式下，大约 86%的名词概念都可以正确聚类。

表 5 实验结果的联立矩阵

系统答案	**真实结果**	
	是	否
是	27401	9698
否	8173	337103

表 6 聚类结果评价

评价指标				
准确率（A）	**精确度（P）**	**召回率（R）**	**漏识率（F）**	**F 值**
0.8617	0.7386	0.7703	0.0280	0.7541

5.3 分析

为了对聚类结果做详细的分析，本文给出了混淆矩阵（confusion matrix），见表 7。斜体字所对应的语义类就是分配给系统结果的语义类，例如表中的第一行序号为 0 的聚类结果的语义类为“资产”。从聚类结果中可以看出，“人、资产、事件”三类是最均质的，因为它们各自所包含的名词概念都聚集在同一类中。而“交通工具、衣物”这两类最不均匀，内部差异较大，它们包含的名词概念分散在很多类中。以“交通工具”为例，该类的名词被聚在不同的 5 个类中，其中 50 个属于“交通工具”类，37 个属于“动物”类。

对上述现象给出解释和分析。首先，拥有独特量词的词语通常能够被正确聚类。例如，“人”这一类名词概念，通常和“位、名”等量词搭配，如“一名商人、一位运动员”，这些量词通常不会修饰别的语义类的名词概念。再如“资产”类名词概念，也有比较固定的

量词与之搭配，如“笔、元”等，构成“一笔报酬、50 元奖金”，这些量词通常也较少修饰其它语义类的名词。同理，与“事件”类名词搭配的量词如“次、场”也非常固定，如“一场婚礼、一次宴会”等。因此，这三类名词概念在基于量词的描述方式下都非常容易地与其它类的名词区分开来，聚类准确率相对较高。

其次，对于那些内部差异较大、非均质的类，与不同下位名词概念搭配的量词会发生变化。换句话说，这一类的名词概念不共享相同的量词集合。以“交通工具”类为例，包含两个子类，一是“机动车”类，如“汽车、巴士、吉普、计程车、摩托车、救护车”等，二是“船”类，如“油轮、小船、游艇、渔轮”等。对于“机动车”类名词概念，通常与量词“辆”搭配，对于“船”类名词概念，通常与“艘、只”搭配。而量词“只”同时是“动物”类名词概念的常用量词，因此，这是为什么聚类结果中“船”类名词与“动物”类名词被聚在同一类中的原因。

表 7 混淆矩阵

系统答案	正确答案										
	动物	建筑	衣物	食物	树	交通工具	出版物	机构	事件	资产	人
0	0	0	1	0	0	0	0	0	0	***80***	0
1	0	7	0	0	0	0	1	***62***	0	0	0
2	0	0	0	0	0	***50***	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	***92***
4	0	2	0	1	***63***	0	0	0	0	0	0
5	0	0	0	3	0	1	0	6	***43***	0	0
6	0	***80***	0	0	0	0	0	14	0	0	0
7	0	1	1	3	0	1	***62***	0	0	0	0
8	4	1	***65***	2	0	0	1	0	0	0	0
9	***81***	2	15	0	0	37	1	0	0	0	0
10	7	0	4	***76***	2	1	1	0	0	0	0

6 结语

本文提出了一种基于量词的名词概念描述方法，通过聚类实验探索了量词对名词语义区分的作用和贡献，实验结果表明量词可以为大部分名词语义类的区分提供有效信息。

本方法的一个局限性在于，它只能描述那些能够与量词搭配的名词概念，对于那些不能受量词修饰的名词，如“心胸、长短”等，这种描述方法就不适用了。对《语法信息词典》名词库中的所有名词进行考察，发现大于 80%的名词都可以受量词修饰。因此，该方法适用于大部分名词。此外，另外一个因素不可忽视的因素是，量词与名词之间的语义关系有时是不明确的，如一些约定俗称的搭配，这也一定程度上造成了该方法在某些语义类上表现较差。

下一步的工作可以从两个方面入手，一是名词概念的选择，需要考虑概念的出现频次及多义性等因素，选择更为平衡的数据。二是尝试用层次聚类的方法，以反映细粒度的概念之间的同源关系。

参 考 文 献

[1] Grefenstette, Gregory. SEXTANT: Extracting Semantics from Raw Text Implementation Details [J]. Heuristics: The Journal of Knowledge Engineering, 1993.

[2] D. Lin, Automatic Retrieval and Clustering of Similar Words [C], In Proc. of COLING-ACL, 1998, pp.768-774.

[3] Almuhareb, A. and Poesio, M. Attribute-based and value-based clustering: an evaluation [C], In Proc. of EMNLP, 2004

[4] Tai, James H. Y. Chinese Classifier Systems and Human Categorization [M]. "*In Honor of Professor William S-Y. Wang: Interdisciplinary Studies on Language and Language Change*, Matthew Chen and Ovid Tseng, eds. Pyramid Publishing Company, 1994, pp. 479-494.

[5] Huang chu-ren, CHEN keh-jiann and GAO Zhao-ming, Noun Class Extraction from a Corpus-based Collocation Dictionary: An Integration of Computational and Qualitative Approaches [J], Quantitative and Computational Studies of Chinese Linguistics, 1998, pp: 339-352.

[6]俞士汶、朱学锋、王惠等. 现代汉语语法信息词典详解（第二版）[M]. 北京：清华大学出版社，2003 年 2 月

[7] Karypis, G, CLUTO: A Clustering toolkit [R], Technical Report 02-017, University of Minnesota, 2002.

[8]王萌、俞士汶、段慧明、孙薇薇，现代汉语名词语法属性的计量研究初探[J]，中文信息学报，Vol.22（5），2008.

[9]Dongdong Zhang, Mu Li and Nan Duan, Measure Word Generation for English-Chinese SMT System [C], In Proc. of ACL, 2008, pp: 89-96.

[10]Dominic Widdows and Beate Dorow, A Graph Model for Unsupervised Lexical Acquisition [C], In Proc. of COLING, 2002, pp.1093-1099.

[11]Hong Zhang, Numeral Classifiers in Mandarin Chinese [J], East Asian Linguist, 2007 (16), pp.43-59.

作者简介：王萌(1977——)，女，江南大学文学院讲师，主要研究方向为计算语言学，Email：wangmengly@163.com；俞士汶（1938——），男，北京大学计算语言学研究所教授，博导，主要研究方向为计算语言学，Email：yusw@pku.edu.cn。