

文章编号:1001-9081(2012)05-1332-03

doi:10.3724/SP.J.1087.2012.01332

基于改进 k -means 算法的中文词义归纳

张宜浩^{1,2*}, 金 澎^{1,2}, 孙 锐^{1,2}

(1. 乐山师范学院 计算机科学学院, 四川 乐山 614004;
2. 乐山师范学院 智能信息处理与应用实验室, 四川 乐山 614004)
(*通信作者电子邮箱 yhaozhang@163.com)

摘要:汉语中一词多义现象普遍存在,词义归纳就是对在不同语境中具有相同语义的词进行归类,本质上是一聚类问题。目前广泛采用无指导的聚类方法对词义归纳进行研究,提出一种改进的 k -means 算法,该算法主要从初始簇中心的选取以及簇均值的计算两个方面进行改进,在一定程度上克服了其对“噪声”和孤立点数据的敏感。在特征表示上用同义词词林中词的分类编号来降低特征维度。实验表明改进 k -means 算法在性能上有较大的提升,F-Score 达到了 75.8%。

关键词:词义归纳; k -means 算法; 聚类; 同义词词林

中图分类号: TP391 文献标志码:A

Chinese word sense induction based on improved k -means algorithm

ZHANG Yi-hao^{1,2*}, JIN Peng^{1,2}, SUN Rui^{1,2}

(1. School of Computer Science, Leshan Teachers' College, Leshan Sichuan 614004, China;
2. Laboratory of Intelligent Information Processing and Application, Leshan Teachers' College, Leshan Sichuan 614004, China)

Abstract: Polysemy is an important and pervasive semantic phenomenon in Chinese; the task of word sense induction is to classify words with the same semantics in different contexts, which is a clustering problem essentially. Currently, unsupervised clustering algorithm has been widely used in its research. In this paper, an improved method of k -means was proposed, which mainly improved the selection of initial cluster centers and the calculation of cluster centroid and overcame the “noise” and the sensitivity of isolated point in data to some extent. Another idea was to use the classification coding of word in Tongyici Cilin to reduce the feature dimension. The experimental results show that the performance has great improvement with the improved k -means, of which the F-Score reached 75.8%.

Key words: word sense induction; k -means algorithm; clustering; Tongyici Cilin

0 引言

在自然语言处理中,面向信息处理的词义分析一直是自然语言处理的焦点。词义知识的获取问题已经成为知识库构建、词义消歧等诸多领域的瓶颈问题^[1]。研究表明使用词义比单纯地使用词形能够改善信息检索^[2]、信息抽取和机器翻译^[3]的结果。而词义知识获取的重要手段之一就是词义归纳,它根据同形异义词语在不同上下文环境中的义项区分出词义,特别是多义词的词义内容,确定词语有多少义项,以及各个义项以何种形式表征等问题^[4]。国内外已经展开了许多相关的研究,但主要集中在英语上,而中文的研究较少。

国外关于英文词义归纳的研究包括:词义聚类方法、词义特征的选取、词义归纳应用方法等。2002 年 Patrick 等提出了一种基于词聚类的词义归纳方法^[5]。该方法选用了词语在文本中的具有句法关系的上下文环境作为特征,利用 CBC(Clustering By Committee)聚类算法从文本中抽取词义。1998 年 Schütze^[6]首次提出了基于上下文分组的词义归纳思想,他指出可以把基于上下文分组的词义归纳方法看成是词义消歧的第一个步骤。2003 年 Ji 等^[7]利用“上下文关系词”完成词义

归纳。他们首先对每一个目标词,在语料中找出所有与它搭配的词语,形成该目标词的候选上下文关系词。在汉语方面,词义归纳也有研究,但是由于汉语在词语切分上的错误、特征的提取与发现困难等方面的原因,其效果不甚理想。为了选取有效的特征,减少噪声,很多词义归纳的研究都是建立在句法分析的基础上,但目前在汉语中也很难找到一个性能很好的句法分析器。基于此本文提出了利用同义词词林改进特征向量的表示,并用聚类的方法研究词义归纳。

1 特征获取与表示

1.1 特征的获取

一般来说,影响词义的因素可以归纳为四类:词法、句法、语义以及语用。其中句法功能需要高准确率的句法分析器的支持,现在虽然在句法分析上取得了积极的进展,但人们对语境类特征的研究则刚起步;限于目前的技术水平,深层语义分析仍然很难实现,人们也仅仅集中研究一些浅层语义分析;而语用 = 知识目前就更是难以获取。因此本文选用搭配词作为词义归纳的句法知识,并通过调整搭配窗口的大小来获得较好的系统性能^[8]。

收稿日期:2011-11-16;修回日期:2012-01-09。

基金项目:四川省教育厅科研项目(10ZB025);国家自然科学基金资助项目(61003206)。

作者简介:张宜浩(1982-),男,河南信阳人,讲师,博士研究生,主要研究方向:自然语言处理; 金澎(1977-),男,河南开封人,副教授,博士,主要研究方向:自然语言处理; 孙锐(1977-),男,四川眉山人,讲师,硕士,主要研究方向:自然语言处理。

1.2 特征的表示

特征本文的形式化表示在文本处理中非常重要。通常的方法是采用词袋模型(Bags of words)^[9],这样会出现严重的数据稀疏问题。鉴于此本文采用《哈尔滨工业大学同义词词林扩展版》^[10]将向量空间模型中基于词的特征项进行语义分析,使用同义词或相关词集合的类代替单个词条,将传统向量空间模型中的特征项由词映射为代表深层次语义的概念。这种表示方法能降低特征的维数,实验表明其在词义归纳研究中极其有效。以“暗淡”为例,提取出其一搭配词为“经济”,在同义词词林中可以找到相应的五级编码“Dj01A01”,然后分别对五级编码赋予从高到低的权重,来表示特征词(在实际实验中只利用高层的两级编码取得了较好的实验结果)。

1.3 数据的规范化

数据的规范化就是将属性值按比例缩放,使之落入一个小小的特定区间,如0.0~1.0。数据的规范化处理对距离度量的分类算法和聚类特别有用,它可以防止具有较大初始值域的属性和具有较小初始值域的属性相对权重过大。本文采用最小-最大规范化对原始数据进行线性变换。假定 A_{\min} 和 A_{\max} 分别为属性A的最小值和最大值,最小-最大规范化通过计算 $v' = \frac{v - A_{\min}}{A_{\max} - A_{\min}}(A_{\text{new_max}} - A_{\text{new_min}}) + A_{\text{new_min}}$ 将A的值映射到区间 $[A_{\text{new_min}}, A_{\text{new_max}}]$ 中的 v' 。

2 k-means 算法及其改进

2.1 k-means 算法

k-means 算法^[11]是以k为输入参数,把n个对象的集合分为k个簇,使得结果簇内的相似度高,而簇间的相似度低。k-means 算法的处理流程如下。首先,随机地选择k个对象,每个对象代表一个簇的初始均值或中心;对剩余的每个对象,根据其与各个簇均值的距离将其指派到最相似的簇;然后计算每个簇的新均值。这个过程不断重复,直到准则函数收敛。通常采用平方误差准则,其定义如下:

$$E = \sum_{i=1}^k \sum_{p \in c_i} \|p - m_i\|^2$$

其中:E是数据集中所有对象的平方误差和,P是空间中的点, m_i 是簇 c_i 的均值。k-means 聚类算法描述如下。

算法:k-means 用于划分的k均值算法,每个簇的中心用簇中对象的均值表示。

输入:簇的数目k,包含n对象的数据集D。

输出:k个簇的集合。

- 1) 从D中任意选择k个对象作为初始簇的中心;
- 2) Repeat
- 3) 根据簇中对象的均值,将每个对象指派到最相似的簇;
- 4) 更新簇均值,即计算每个簇中对象的均值;
- 5) Until 不再发生变化。

2.2 改进的k-means 算法

利用上述的k-means 算法进行词义聚类分析时,发现实验的结果稳定性方面存在问题。因此本文提出了一种簇初始中心选择和簇均值计算的改进方法。

2.2.1 簇初始中心选择的改进方法

由于k-means 算法在初始化簇的种子对象时的随机性,不同的初始中心的选择对聚类结果有较大的影响。本文利用

高频搭配词预判簇的初始中心,进行约束指导聚类,以期得到较好的聚类结果。

假定 X_{text} 是篇章中的所有词(候选目标词除外),而后逐一统计这些词的频率 $P(X_{\text{text}})$,并提取出频率最大的若干个词 X_n 。在分析目标词的词义时,假定 X_s 为该句子中的所有词(候选目标词除外),逐一取 X_s 集合中的单个词 $X_{s(i)}$,如果 $X_{s(i)} \in X_n$,则断定词义相同,选取其中之一作为簇中心,否则为两个簇的中心,直至找到所有簇中心后终止比较。这种方法看起来虽然比较简单,但是在实验中却十分有效。

例如:在对“吃饭”这个目标词进行词义归纳分析时,发现很多句子中有“靠文凭吃饭”、“靠力气吃饭的劳动者”、“靠工资吃饭”、“靠外表吃饭”、“靠总结经验吃饭”、“靠本事吃饭”这些短语。而这些短语中又无一例外地有“靠”这个动词(统计表明其是高频词),通过这个词可以实现对“吃饭”词义的较好聚类,判断每个句子中是否包含这一关键词(如有聚为一族,否则为另一簇)来选取簇初始中心。

2.2.2 簇均值计算的改进方法

由于k-means 算法对孤立点很敏感,少量这样的点对聚类结果有较大的影响^[12]。而在词义归纳实验数据中恰恰存在着少量的数据远离高密度的数据集。针对这一问题,改进算法在迭代计算新的簇均值时过滤掉了少量的离群点。k-means 算法中簇均值计算方法的改进如下:

- 1) 对于第 $k-1$ 轮聚类,计算簇中的每个对象 $x_{i(k-1)}$ 与簇聚类中心 $c_{n(k-1)}$ 的距离,并将距离度量 $s_{i(k-1)}$ 以及对应的对象 $x_{i(k-1)}$ 存储在相应的集合 M_{k-1} 中;
- 2) 假定 M_{k-1} 的数量为 s_{num} ,提取出 βs_{num} 个最小距离($\beta \in (0,1)$),计算这 βs_{num} 距离的均值为 $c_{\beta n(k-1)}$;
- 3) 将 $c_{\beta n(k-1)}$ 作为第 k 轮聚类的种子,重复1)、2)两步进行迭代聚类。

通过上述改进可以保证在更新簇均值时过滤掉少量的离群点,避免了离群点致簇中心的较大偏移。上述对k-means 算法的改进在小样本数量上取得了较好的实验性能,在后面的实验中得以验证。

3 实验分析

3.1 实验数据及评测指标

本次词义归纳评测采用的是中国科学院软件研究所基础软件国家工程研究中心信息检索实验室开发的中文词义归纳语料。该语料包括50个目标词,2500个句子以及人工标注的答案。

对于词义归纳系统的评测,本文把具有相同词义标签的gold standard当作一个类,然后将输出结果中具有最大权重的相同词义标签的例子当作一个类进行比较来计算F-Score^[13]。设 c_r 是gold standard的一个类, s_i 是系统输出的一个类。

$$F\text{-Score}(c_r, s_i) = 2 \times P \times R / (P + R)$$

$$P = \frac{\text{结果类中标注正确的例子个数}}{\text{对应类的例子总数}}$$

$$R = \frac{\text{结果类中标注正确的例子个数}}{\text{标准答案中对应类的例子总数}}$$

$$\text{对于一个给定的类 } c_r, \text{ 有 } F\text{-Score}(c_r) = \max_{s_i} F\text{-Score}(c_r,$$

$$s_i)). \text{ 而总体的 } F\text{-Score} = \sum_{r=1}^c \frac{n_r}{n} F\text{-Score}(c_r), \text{ 其中: } c \text{ 代表总}$$

的类数, n_c 是类 c 中元素的数据, n 是总的元素数目。

3.2 实验结果分析

为了验证本文所提的特征表示方法以及改进 k -means 算法的有效性, 做如下实验。实验中用到的 Simple k -means 算法和 EM 聚类算法是来自 weka 中的集成算法, 而改进 k -means 算法是本文实现的算法。实验选择特征窗口大小为 5 时, 三种不同的聚类算法获得的实验结果如表 1。

表 1 三种不同的聚类算法获得的结果

特征表示 方法	聚类算法	<i>F-Score</i>		
		最好	最坏	平均
词袋模型 ^[9]	EM 聚类算法	0.806	0.403	0.626
	简单 k -means	0.779	0.415	0.619
	改进 k -means	0.860	0.502	0.658
同义词词林 ^[10]	EM 聚类算法	0.826	0.506	0.702
	简单 k -means	0.845	0.486	0.696
	改进 k -means	0.875	0.656	0.758

从表 1 中不难看出, 相比采用词袋模型表示特征的方法, 用同义词词林表示特征的方法具有较好的实验效果。在三种聚类算法上其 *F-Score* 分别提高了 6%、7%、10%。

从表 1 中也可以看出, 在两种特征表示方法下, EM 聚类算法和简单 k -means 之间的性能差距很小, 而改进 k -means 算法则分别有 4% 和 6% 的提高, 这表明了本文提出的改进 k -means 算法在词义归纳实验中是有效的。

为了最大可能地提高系统的性能, 本实验测试了几组不同窗口大小的实验结果如下(实验中的聚类算法为改进的 k -means 算法, 特征表示利用同义词词林)。

表 2 不同窗口大小的实验结果

窗口大小	<i>F-Score</i>		
	最好	最坏	平均
3	0.823	0.605	0.725
5	0.875	0.656	0.758
10	0.812	0.632	0.731

从表 2 不同窗口大小的实验结果可以看出, 选取窗口大小为 5 的特征窗口时, 考虑当前词的前后各 5 个词, 将这 10 个词作为特征, 其 *F-Score* 取得了最佳的性能, 因此在最终的系统中选取窗口大小为 5 作为特征窗口。

4 结语

本文提出了一种改进的 k -means 算法对词义归纳进行研究。统计分析发现, 待聚类词语所在句子中的某些高频词可以实现词义的初步区分, 正是利用这一特点 k -means 算法初始聚类中心的改进方法。而由于 k -means 算法对于“噪声”和

孤立点数据极为敏感, 提出了选取离簇均值中心较近点优先聚类, 这样就避免孤立点数据对其他正常数据的影响。实验表明, 本文所采取的对 k -means 算法的改进是有效的。下一步的工作将重点考虑从语义学的角度对句子进行分析, 提取出更准确的搭配词, 这对词义归纳性能的提升是十分有效的。

参考文献:

- [1] 朱虹, 刘扬. 词汇语义知识库的研究现状和发展趋势[J]. 情报学报, 2008, 27(6): 870–877.
- [2] OZLEM U, BORIS K, DENIZ Y. Word sense disambiguation for information retrieval[C]// Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Conference on Innovative Applications of Artificial Intelligence. [S. l.]: American Association for Artificial Intelligence, 1999: 985–986.
- [3] VICKREY D, BIENALD L, TEYSSLER M, et al. Word-sense disambiguation for machine translation[C]// Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2005: 771–778.
- [4] ENEKO A, AITOR S. Evaluating word sense induction and discrimination systems[C]// SemEval-2007: Proceedings of the 4th International Workshop on Semantic Evaluations. Stroudsburg: Association for Computational Linguistics, 2007: 7–12.
- [5] PANTEL P, LIN DEKANG . Discovering word senses from text[C]// Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002: 613–619.
- [6] SCHUTZE H. Automatic word sense discrimination[J]. Computational Linguistics, 1998, 24(1): 97–124.
- [7] JI H, PLOUX S, WEHRLI E. Lexical knowledge representation with contextonyms[EB/OL].[2011-10-22]. <http://www.cs.toronto.edu/~gh/Courses/2528/Readings/Ji-et-al-Contextonyms.pdf>.
- [8] 何径舟, 王厚峰. 基于特征选择和最大熵模型的汉语词义消歧[J]. 软件学报, 2010, 21(6): 1287–1295.
- [9] 王锦, 王会珍, 张俐. 基于维基百科类别的文本特征表示[J]. 中文信息学报, 2011, 25(2): 27–31.
- [10] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报: 信息科学版, 2010, 28(6): 602–608.
- [11] de AMORIM R C, MIRKIN B. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering[J]. Pattern Recognition, 2012, 45(3): 1061–1075.
- [12] RONG HUIGUI, LI MINGWEI, CAI LIJUN. An early recognition algorithm for BitTorrent traffic based on improved K-means[J]. Journal of Central South University of Technology, 2011, 18(6): 2061–2067.
- [13] ZHAO YING, KARYPIS G. Hierarchical clustering algorithms for document datasets[J]. Data Mining and Knowledge Discovery, 2005, 10(2): 141–168.
- [15] 冈萨雷斯. 数字图像处理[M]2 版. 阮宇智, 阮秋琦, 译. 北京: 电子工业出版社, 2007.
- [16] TAKAGI T, SUGENO M. Fuzzy identification of systems and its applications to modeling and control[J]. IEEE Transactions on Systems, Man and Cybernetics, 1985, 15(1): 116–130.
- [17] ABREU E, LIGHTSTONE M. A new efficient approach for the removal of impulse noise from highly corrupted images[J]. IEEE Transactions on Image Processing, 1996, 5(6): 1012–1025.

(上接第 1295 页)

- [12] NIKOLOVA M. A variational approach to remove outliers and impulse noise[J]. Journal of Mathematical Imaging and Vision, 2004, 20(1): 99–120.
- [13] RUSSO F. Noise removal from image data using recursive neurofuzzy filter[J]. IEEE Transactions on Instrumentation and Measurement, 2000, 49(2): 307–314.
- [14] 王利朋, 刘东权. 基于粒子群算法的柔性形态学滤波器[J]. 计算机应用, 2010, 30(10): 2811–2814.

- [15] 冈萨雷斯. 数字图像处理[M]2 版. 阮宇智, 阮秋琦, 译. 北京: 电子工业出版社, 2007.
- [16] TAKAGI T, SUGENO M. Fuzzy identification of systems and its applications to modeling and control[J]. IEEE Transactions on Systems, Man and Cybernetics, 1985, 15(1): 116–130.
- [17] ABREU E, LIGHTSTONE M. A new efficient approach for the removal of impulse noise from highly corrupted images[J]. IEEE Transactions on Image Processing, 1996, 5(6): 1012–1025.