# 基于维基百科类别的文本特征表示*

王锦[1,2] 王会珍[1,2] 张俐[1,2]

1. 东北大学自然语言处理实验室，辽宁沈阳 110004

2. 医学影像计算教育部重点实验室（东北大学），辽宁沈阳，110819

E-mail: wangjin@ics.neu.edu.cn

**摘　要**：本文提出了基于维基百科类别体系的文本特征表示方法，该方法将文本中的词映射到维基百科的类别体系中，使用类别作为特征来对文本进行表示。基于维基类别的文本特征表示方法可以增强文本特征表示能力，降低文本特征空间维数。针对维基百科条目在语料中覆盖度不足的问题，本文提出了一种基于全局信息自学习维基百科类别的方法。本文构造基于维基百科类别为文本表示的分类系统，实验结果证明，基于维基百科类别作为文本表示特征，相对于词袋模型，具有明显的降维效果，在特征数为700个时，分类的F1值提高了5. 14%。

**关键词**：文本分类，维基百科类别，文本表示

# Text Representation Using the Wikipedia Category

Wang Jin[1,2], Wang Huizhen[1,2], Zhang Li[1,2]

1.Natural Language Processing Lab, Northeastern University Shenyang,Liaoning, P.R.China, 110004

2. Key Laboratory of Medical Image Computing (Northeastern University), Ministry of Education, Shenyang, Liaoning, P.R.China, 110819

E-mail: wangjin@ics.neu.edu.cn

**Abstract:** In this paper, we present a text representation method by using wikipedia categories as text features. This method can map each word of text to one of wikipedia categories. It can enhance the representation ability of features and reduce the dimensions of a text vector. An approach is presented by using clustering techniques to resolve the limited coverage of wikipedia categories by mapping unknown words into predefined categories Then a text category system is developed that uses these learned wikipedia categories as text features. The experimental results show that text representation based on wikipedia categories has the obvious effect of dimension reduction, and achieves 5.14% F1 improvement over the BOW-based method when 700 features are used for text classification.

**Key words:** text classification, text representation, wikipedia category.

## 1. 引言

文本分类一直是自然语言处理领域研究的一个重要课题。近年来，国内外许多研究人员对文本分类任务做了深入研究，包括在文本表示、特征选取、分类模型等方面的探索。在传统的文本表示中，文本被表示成一个文本特征向量，文本特征用词来表示，即文本表示采用BOW（Bag of Words）模型。这种方法简单、易行，目前大多数文本分类系统都是使用这种文本特征表示方法。

但是，词作为文本特征存在特征空间维数过高、表达能力有限[1]等问题。该方法仅仅用词作为特征，并没有使用人们掌握的知识[2]。针对这些问题，国内外研究人员对知识库在文本分类中

*本课题研究工作部分得到了国家自然科学基金项目（60873091）、中央高校基本科研业务费专项资金资助。

的应用进行了研究。Scott[3]等人利用WordNet的语义关系Hypernym来表示文本特征，但是这些知识库都存在覆盖度不足的问题。研究人员还对词簇作为文本特征做了很多研究。Baker和McCallum[4]提出一种基于词的类别分布来进行词聚类，然后用这些词簇表示文本。Chen[5]等提出了基于全局信息词聚类作为文本表示的方法，该方法将类别分布相似的词归为一簇，用簇作为特征表示文本。

本文在词聚类作为文本表示的基础上，引入了维基百科的类别体系，将词进行有指导的聚类，即将文本中所有词映射到维基百科类别上，采用维基百科的类别作为文本表示的特征。目前，维基百科是世界上最大的开放式百科全书，由人工标注而成，具有较快的更新速度。维基百科的类别能把表达不明确的维基百科条目映射为理解能力更强的信息，如："狮子王"、"美女与野兽"、"米老鼠"都被映射为"迪士尼动画"这个维基类别，而人们很容易把"迪士尼动画"和文化、艺术等主题类别联系起来。针对维基百科类别对语料的覆盖度不足的问题，本文提出了一种全局信息自学习维基类别的词聚类方法，用维基百科的类别来表示词聚类得到的簇，并使用簇的信息表示文本，构造了基于簇的文本分类系统。

## 2. 基于维基百科类别的文本特征表示

在传统的文本分类中，文本特征用词来表示，存在表达能力有限的问题[8]。所以，本文试图寻找一种准确描述文本内容的表示方法来表示文本。维基百科是目前最大的在线知识库之一，而且，维基百科中提供了一个由大众来进行编辑的格状分类体系。每一个条目都能映射到分类体系中的某些类别，这个信息是人工标注的，具有很高的准确度。因此，本文选用维基百科的类别对文本进行表示。与本文工作最相似的前期工作Chen等[5]曾利用人工构建的领域知识库将文本中所有词映射到预定义的领域特征改善文本表示。本文与前人工作的最大区别在于没有采用人工构建的领域知识库，而是从维基百科中自动获取部分词与维基百科类别的对应关系，然后进行自动扩展，用于改善文本特征表示，提高文本文类的性能。整个过程没有涉及到额外的人工标注代价，方法的基本动机与Chen等[5]的工作相似，但技术的处理角度和方法不同。

### 2.1 维基百科的类别体系

维基百科是目前世界上最大的多语种的面向互联网的开放式的百科全书。它的基本组成单元叫"概念"或"条目"，每个条目都有一篇文章来解释[6]。维基百科的每个条目都对应一组维基百科类别，维基百科类别体系是基于层次结构的网状类别体系[9]。表1是维基百科类别的部分实例。当然，维基百科的类别体系和中图法[7]的类别体系有所差异，并且在一个条目对应的所有类别中，很多类别不能准确的表达分类信息，只是有助于查找在这个类别下的其他条目，这个类

表1 维基百科类别的部分实例

| 条目 | 类别 |
|---|---|
| 阿伏伽德罗定律 | 气体定律 \| 物理定律 \| 热学 \| |
| 阿道夫·希特勒 | 德国总理 \| 德国总统 \| 德国第二次世界大战人物 |
| 狮子王 | 迪士尼动画 \| 1994年電影 \| 美国电影作品 \| 百老汇音乐剧 |
| 美女与野兽 | 迪士尼动画 \| |
| 米老鼠 | 迪士尼动画 \| |

别体系有待进一步研究。本文用的维基语料是，从维基百科网站[10]上下载的2010-3-3版的XML格式的语料，它包含有553709个页面，其中有类别的页面数为149272个，类别体系中的类别数为135214个，本文将词映射到维基百科的类别中总共覆盖到类别体系中14052个类别。

### 2.2 基于维基百科类别的文本特征表示

在本文中，维基百科的类别作为文本特征，表示成一个文本特征集合，也就是维基类别的集合，这里用M表示维基类别。具体过程如下：

(1) 建立维基百科的每个条目和其对应一组类别的映射关系。维基百科的条目集合 $T=\{t_1,t_2,\ldots,t_n\}$，第i个条目对应的维基类别集合$M(t_i)=\{m_j|t_i$条目的类别标签为$m_j\}$。

(2) 构建863语料中出现的维基条目的词的集合T。使用东北大学自然语言处理实验室的分词和专名标注系统（为了保证分词的一致性，可以事先将维基百科条目作为临时辞典参与分词过程）对文本进行分词，本文称这里分词得到的普通词为W，将普通词W中是维基百科条目的词放入T集合中。

(3) 利用T和维基类别M的映射关系，最终将语料中每篇文档映射成只有维基百科类别的特征集合$M_k$，用tf表示特征的权重。

在863文本分类评测语料上进行统计，863语料中共有107469个词，维基百科中覆盖了其中的17570个词，大部分词在维基百科中没有类别信息，仅仅使用现有维基百科条目对文本的覆盖度明显不足。为了解决这个问题，本文提出了基于全局信息自学习维基类别的方法（本质上是词聚类技术）来对没有维基类别信息的其它词自动赋予维基类别标记。

## 3. 基于全局信息自学习维基类别方法

语料中没有维基百科类别的词（也就是不是维基条目的词），这些词用UW表示：$UW=\{uw|uw\in W \text{ and } uw\notin T\}$。本文提出一个基于聚类技术的自动学习维基类别的方法，将UW中的词与维基百科的类别M建立映射关系。基本步骤是，利用词在文本类别中的分布，把所有的词表示成向量的形式，将每个词簇m中的所有元素（也就是维基百科条目）的均值作为词簇的中心点，通过计算uw和每个中心点的距离，来获得与uw相似度最大的词簇$m_i$，建立UW和M的映射关系。

### 3.1 定义

(1) T：维基百科条目集合，$T=\{t_1, t_2, t_3, \ldots, t_n\}$；

(2) M：维基类别的集合$M=\{m_1, m_2, \ldots, m_n\}$，第i个类别对应维基条目集合$T(m_i)=\{t_j|\ t_j$条目的类别标签为$m_i\}$；

(3) C：是863评测语料中的类别集合$C=\{c_1, c_2, \ldots, c_{36}\}$。

(4) $p(C|w)$：表示词w在整个类别间的分布，也就是词w在每个类别c中的频数$N(c_j|w)$。

(5) $p(C|m_i)$：表示簇（维基类别）$m_i$在整个类别C的分布，也就是36维的向量。其计算方法就是计算簇中的元素（维基条目）t在整个类别间的分布的均值，计算公式如公式（1）所示。

$$p(C|m)=\sum_{i=1}^{n}p(C|ti)/n \quad (1)$$

其中，n表示簇m中的元素个数。

### 3.2 基于全局信息的自学习算法

首先将训练语料进行预处理，将训练语料分词后得到的普通词W中不是维基百科条目的词放入UW集合，然后将UW中的每一个词划分到维基类别M中。具体过程如下：

算法1. 自学习算法

---

输入：待划分维基类别的词集UW={uw | uw∈W and uw∉T}，维基百科类别集合M={ $m_1$，$m_2$，…，$m_n$ }。

输出：UW中的词$uw_i$对应的类别$m_k$。

步骤：

(1) 用公式(1)计算簇的中心点，得到每个簇在整个文本类别C中的分布p(C| $m_j$)。

(2) Loop，直到所有UW都加入到M集中{

① 从待划分维基类别的词集合UW中取出一个词$uw_i$；

② 用公式(2)计算待划分词$uw_i$和每个簇m的距离：D={ D($uw_i$, $m_1$)，D($uw_i$, $m_2$)，…，D($uw_i$, $m_n$)}；

$$D(uw,m)=D\left(p(C|uw),p(C|m)\right)=\sum_{t=1}^{36}\left(N(c_t|uw)-N(c_t|m)\right)^2 \quad (2)$$

③ 求$m_k$，k=arg min（D($uw_i$,$m_j$)）；/*1≤j≤n*/

④ $uw_i \rightarrow m_k$；

}

---

通过全局信息自学习维基类别的方法，使得语料中没有维基类别信息的词UW和维基类别M建立一一对应的映射关系。利用T→M和UW→M的映射关系，重新构造文本特征，将语料中每篇文档映射成只有维基百科类别的特征集合$M_k$，用tf表示特征的权重。

## 4. 实验与分析

### 4.1 实验语料

863中文评测语料，该语料来源于2004年国家863 中文文本分类评测的语料，其中采用中图法进行构建分类体系，共36类，每类包含100篇中文文本。在语料预处理过程中，分词工具采用东北大学自然语言处理实验室开发的分词工具NEUCSP，去掉禁用词后，剩下的词汇个数为107469。

在分类实验过程中，采用十次交叉检验的方法，90%语料作为训练语料，剩下的10%语料作

为测试语料，将十次交叉检验的分类性能指标取平均值作为最后分类性能评价。

## 4.2 分类模型选择

本文实验选用最大熵分类器（ME）、朴素贝叶斯分类器（NB）、支持向量机分类器（SVM）三种分类器进行对比实验。最大熵使用张乐开发的工具包，支持向量机采用了$SVM^{light}$作为SVM的实现，使用$SVM^{light}$的默认参数。

## 4.3 评价方法

在本文实验中，以文本分类的性能来衡量文本表示方法的性能。本文使用Macro F1来评价分类性能。计算公式如下：

$$MacroP = \frac{1}{n}\sum_{j=1}^{n} P_j \qquad MacroR = \frac{1}{n}\sum_{j=1}^{n} R_j$$

$$MacroF1 = \frac{MacroP \times MacroR \times 2}{MacroP + MacroR}$$

其中，n是类别总数，$P_j$为第$j$类的准确率，$R_j$为第$j$类的召回率。

## 4.4 实验设置

1）以词作为文本特征表示的分类系统

共构建三个分类器：BOW-NB表示采用词为特征，使用朴素贝叶斯分类模型的分类系统；BOW-ME表示采用词为特征，使用最大熵分类模型的分类系统；BOW-SVM表示采用词为特征，使用支持向量机分类模型的分类系统。

2）以维基类别作为文本特征表示的分类系统

共构建三个分类器：Wiki-NB表示采用维基类别为特征，使用朴素贝叶斯分类模型的分类系统；Wiki-ME表示采用维基类别为特征，使用最大熵分类模型的分类系统；Wiki-SVM表示采用维基类别为特征，使用支持向量机分类模型的分类系统。

3）基于全局信息自学习维基类别的分类系统

共构建三个分类器：Global-Wiki-NB表示采用维基类别为特征，使用朴素贝叶斯分类模型的分类系统；Global-Wiki-ME表示采用维基类别为特征，使用最大熵分类模型的分类系统；Global-Wiki-SVM表示采用维基类别为特征，使用支持向量机分类模型的分类系统。

## 4.5 实验结果

本实验对3个分类系统进行了比较。图1是使用朴素贝叶斯（NB）分类器的3个分类系统的分类结果，y轴是各分类系统的F1值，x轴是表示该系统使用的文本特征数目。从整体上看，基于Wiki-NB方法的F1值并没有比BOW-NB的F1值高，说明维基类别存在明显的的覆盖度不足的问题，然而，Global-Wiki-NB的分类性能高于BOW-NB，尤其是在特征数少的时候。进一步考察基于Global-Wiki-NB的方法，在特征数为200~2000之间明显优于BOW-NB，特征数为700时，基于

Global-Wiki-NB方法的F1值达到72.56%，比相同特征数的BOW-NB方法高5.14%，这与基于BOW-NB方法特征数为2000时的性能，达到相当的效果。

图2是最大熵（ME）分类器的3个分类系统的分类结果，在特征数为200~2000之间时，Global-Wiki-ME的分类性能也明显优于BOW-ME，特征数为700时，基于Global-Wiki-ME方法的F1值达到72.53%，比相同特征数的BOW-NB方法高3.25%。图3是支持向量机（SVM）分类器的3个分类系统的分类结果，特征数为800时，基于Global-Wiki-SVM方法的F1值达到73.31%，比相同特征数的BOW-SVM方法高3.89%。

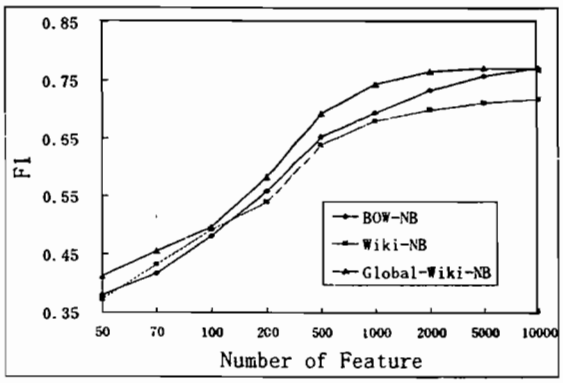


图1 NB分类器的3个分类系统的实验结果

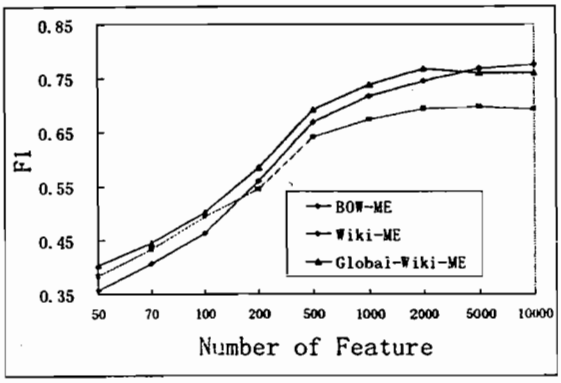


图2 ME分类器的3个分类系统的实验结果

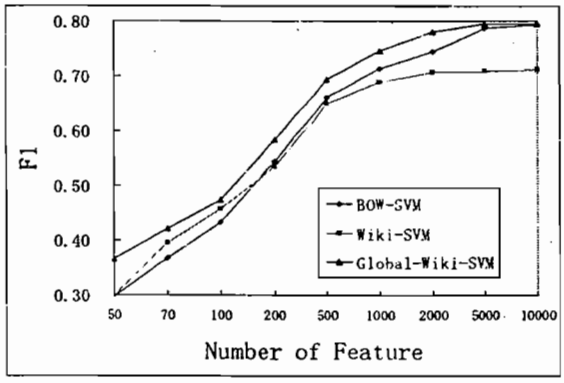


图3　SVM分类器的3个分类系统的实验结果

## 4.6 讨论

本文提出了基于维基百科类别的文本特征表示方法。该方法优于前人的工作，因为维基类别是从维基百科中自动获取的，并且可以进行自动扩展，无需人工构建知识库。同时，从实验结果可以看出，在特征数很少的情况下，基于Global-Wiki的方法已经达到很好的效果。因为在自学习维基类别的过程中，将大量的词映射到了少量的维基类别中，这不仅能起到了降维作用，有效的降低时间复杂度，减少了系统的计算开销，而且能增强特征的表达能力。本文用词频tf作为特征的权重，然而很多词频低的信息都是表达能力强的信息，比如“姚明”，当选择一定数量的特征时，这些信息很可能被过滤掉。Global-Wiki的方法会把这些信息聚到少量的维基类别上，使得

在特征数很少时，这些信息也可以被利用上，这就使得在特征数很少时，本方法能达到很好的性能。从图中我们同样可以看出，在特征数增加到5000以上时，Global-Wiki的分类性能与基于BOW的分类性能趋于相同甚至下降，这表明，再增加特征，也只是引入了噪音，对文本分类没有起到作用。

## 5. 结论

本文提出了一种新的文本特征表示方法，用维基百科的类别作为文本的特征，并且结合了全局信息自学习维基类别的方法，来解决维基类别对文本的覆盖度不足的问题。这种方法，克服了传统的词作为文本特征的空间维数过高和表达能力有限等问题。实验结果表明：

(1) 用维基百科的类别作为文本特征，有助于增强文本特征的表达能力；

(2) 基于自学习方法的维基类别作为文本特征可以很好的改善文本分类的性能，特别是在特征数目少的情况下表现出更优的效果；

下一步的工作的研究重点是：一是，如何过滤掉更多无用的维基类别，用更少的特征来表示文本进行文本分类；二是，探索维基百科知识库在自然语言处理领域的其他应用。

## 参考文献


[1] Sangkon Lee , Masami Shishibori. Passage segmentation based on topic matter. Computer Processing of Oriental Languages , 2002 ,15 (3) : 305～340

[2] 陈文亮, 朱靖波. 基于领域词典的文本特征表示. 计算机研究与发展. 2004.

[3] Scott, Sam, and Stan Matwin. Text classification using wordnet hypernyms. In The COLING ACL Workshop on Usage of WordNet in Natural Language Processing Systems, 1998.

[4] L. D. Baker, A. K. MCallum. Distributional clustering of words for text classification. In: Proc. 21st Annual Int'l ACM SIGIR Conf. Research and Development in Information Retrieval. New York : ACM Press , 1998. 96～103

[5] Chen Wenliang , Chang Xingzhi , Wang Huizhen , et all Automatic word clustering for text categorization using global information. AIRS2004, Beijing , 2004.

[6] P.Wang, J.Hu, H.-J.Zeng, L.Chen. Improving text classification by using encyclopedia knowledge. In Internation Conference on Data Mining, pages 332-341, Omaha, NE, 2007.IEEE.

[7] China Library Categorization Editorial Board China Library Categorization. The 4th ed. Beijing: Beijing Library Press ,1999.

[8] 陈文亮. 面向文本分类的文本特征学习技术研究. 东北大学博士学位论文,2005.

[9] Xiaohua Hu, Xiaodan Zhang. Exploiting Wikipedia as External Knowledge for Document Clustering. ACM, 2009.

[10] http://zh.wikipedia.org/zh-cn/Wikipedia:%E9%A6%96%E9%A1%B5