基于特征关联度的汉语文本自动分类系统的设计与实现

张玥杰　　姚天顺
东北大学计算机系（110006）

摘要：本文提出一种基于预定义类别与文本特征之间关联度的自动分类算法，并详细阐述系统的设计与实现过程。为测试分类系统的实现性能，建立具有12类别的分类体系，构造包含近500篇汉语新闻语料的测试集。实验结果表明，评价自动分类算法的两个重要指标：查全率和查准率，都比较令人满意。

关键词：信息检索、自动分类、特征向量、关联度。

The Designation and Implementation of Chinese Text Automatic Classification System based on Feature Relativity

Zhang,Yuejie and Yao,Tianshun
Dept. of Computer Science, Northeastern University

Abstract: The paper presents a kind of automatic classification algorithm based on the relativity between the predefined categories and the text.The processes of designation and implementation about the system are elaborated in detail.For evaluating the accomplished performance on classification model,we build a classifying system which includes 12 categories.Meanwhile,we construct a test set which includes nearly 500 pieces of Chinese news.The results of the experiment show that the two important norms which are used for evaluating the automatic classification algorithm,precision and recall,are quite satisfying.

Key words: infomation retrieval,automatic classification,feature vector,relativity.

一、引言

随着科学技术的高度发展，信息情报激增，有大量的科技文献、新闻语料等各种文本需要管理，这就要求人们花费大量的时间和金钱，以有效地保留大的文本集合。对文本进行有效管理的方法之一，就是将它们进行系统地分类，而不浪费人类资源。

文本自动分类，就是基于内容将自然语言文本自动分配给预定义的类别。在文本自动分类的方法学中，是通过使用信息检索方法提出的一些自动文本分类方法。传统的信息检索技术主要是采用关键词查找和统计技术来检索相关的文本[1]。对于基于词的技术来说，

具有很大的局限性[2]，即同义词、多义性、短语、局部上下文和全局上下文等问题。

本文中的汉语文本自动分类系统，通过对部分汉语文本的统计学习，描述分类项和分类文本作为规范化特征向量之间的内部联系，从而有效地为文本分配类别编码。在使用基于词的技术基础上，将分类项特征与关键词综合考虑，更精确地对文本进行自动分类处理。

二、基于特征关联度的自动分类算法

基于特征关联度的分类算法为系统实现的核心，首先描述以下定义及准则:

[定义 1] 规范化特征向量

所谓规范化特征向量，就是指以反映对象特征或对象间联系的因素作为元素，描述对象自身特征或对象之间的联系，并遵循规范化准则，经过规范化处理的特征向量。

[定义 2] n 重加权

对于规范化特征向量，遵循加权准则，实施 n 次的加权处理。

[定义 3] 关联度

遵循关联度测量准则，利用关联度描述对象之间的相关程度。

[准则 1] 规范化准则

对特征向量有必要进行规范化处理，使其具有相同的长度。

设特征向量 $Vector(v_1, v_2, v_3, \cdots\cdots, v_i)$（1≤i≤N），对于 $Vector$ 中的每个 v_i

$$v_i \leftarrow \frac{v_i}{v_{S_v}}, \qquad \text{其中} \quad v_{S_v} = \sqrt{\sum_{i=1}^{N} v_i^2}$$

[准则2] 加权准则

确定与特征向量中各元素对应的权值，建立加权因子向量，进行加权处理。

设特征向量 $Vector(v_1, v_2, v_3, \cdots\cdots, v_i)$（1≤i≤N），加权因子向量 $Weight(w_1, w_2, w_3, \cdots\cdots, w_i)$，按如下公式将特征向量加权：$v_i \leftarrow v_i \times w_i$。

[准则3] 关联度测量准则

设规范化特征向量 $Vector1(v_{11}, v_{12}, v_{13}, \cdots\cdots, v_{1i})$ 与 $Vector2(v_{21}, v_{22}, v_{23}, \cdots\cdots, v_{2i})$（1≤i≤N），二者关联度按如下公式计算：$\mathrm{Re}l(Vector1, Vector2) = \sum^{N}(v_{1i} \times v_{2i})$

2.1 构造规范化类别特征向量

设 $C_1, C_2, C_3, \ldots, C_i, \ldots, C_M$ 是预定义类别，M是类别数目，N是义类词典中类别的数目。

（1）初始类别特征向量—设 $C_i = (t_{i1}, t_{i2}, t_{i3}, \ldots, t_{ij}, \ldots, t_{iN})$（1≤i≤M，1≤j≤N），$C_i$ 中的项 $t_{i1}, t_{i2}, t_{i3}, \ldots, t_{ij}, \ldots, t_{iN}$ 分别是义类词典中每一类别的词项中属于第i类的词的总数。

（2）一重加权因子向量—设 $WO_{Ci}(wo_{i1}, wo_{i2}, wo_{i3}, \cdots, wo_{ij}, \cdots, wo_{iN})$，$wo_{ij}$ 为对于类j的每个 $t_{ij} > 0$ 中预定义类别的个数（$0 \leq wo_{ij} \leq M$）。存在特殊情形：对于类j，当所有预定义类别的 t_{ij} 均为0或大于0时，即每个 $t_{ij}=0$ 或 $t_{ij} > 0$ 时，wo_{ij} 均为0。在所有预定义类别中从未出现或都出现的义类词典的类别，其重要性为最小。

（3）二重加权因子向量—设 $WSc_i(ws_{i1}, ws_{i2}, ws_{i3}, \cdots, ws_{ij}, \cdots, ws_{iN})$，又设 $K_x = (e_1, e_2, e_3, \ldots, e_i, \ldots, e_M)$，其中 $e_1, e_2, e_3, \ldots, e_i, \ldots, e_M$ 为关键词分属于主题词表中相应类别的数目。ws_{ij} 按如下公式计算：

$$ws_{ij} = \frac{e_i}{\sqrt{\sum_{i=1}^{M} e_i^2}}$$

遵照加权准则，将初始类别特征向量由一重和二重加权因子向量进行多重加权处理，最后按照规范化准则，获取最终的规范化类别特征向量，形成规范化特征向量表，如表1。

	抽象事物	心理活动	……	时间与空间	关联
C1 文化	t_{11}	t_{12}		$t_{1,N-1}$	t_{1N}
……					
C12 农业	$t_{12,1}$	$t_{12,2}$		$t_{12,N-1}$	$t_{12,N}$

表1. 规范化类别特征向量表

2.2 构造规范化文本特征向量

（1）初始文本特征向量—设 $F_x = (x_1, x_2, x_3, \ldots, x_i, \ldots, x_N)$（1 ≤ i ≤ N），其中 $x_1, x_2, x_3, \ldots, x_i, \ldots, x_N$ 为文本中属于义类词典所分各类加权后关键词项的总和，最初设为相同出现，即为0。

（2）一重加权因子向量—设 $WO_x(wo_1, wo_2, wo_3, \cdots, wo_i, \cdots, wo_N)$，$wo_i$ 按如下公式计算：

$$wo_i = \sum_{j=1}^{W_{Num}} W_{word_j}$$

其中：W_{Num} 为属于义类词典中类i的关键词数目，W_{word_j} 为关键词 $word_j$ 的总权值。

（3）关键词的总权值—该权值是通过该关键词 $word_j$ 在文本中出现的位置来决定，以下是三种出现情形：1. 文本标题；2. 标题的组成部分；3. 正文的组成部分。设 fre_{ij}（1 ≤ j ≤ 3）为 $word_j$ 在情形j中的频率，W_j 为 $word_j$ 在情形j中所对应的权值。总权值按如下公式获得：

$$W_{word_i} = \sum_{j=1}^{3} (fre_{ij} \times w_j)$$

遵照加权准则，将初始文本特征向量由一重加权因子向量进行单重加权处理，最后按照规范化准则，获取最终的规范化文本特征向量。

2.3 测量分类文本与每一预定义类别的关联度

遵循关联度测量准则，测量每一预定义类别和分类文本的两种规范化特征向量之间的关联度，获得关联度向量 $REL(rel(C_1, x), rel(C_2, x), \cdots\cdots, rel(C_i, x))$（1 ≤ i ≤ M）。关联度越大，文本内容和预定义类别越关联。文本最终所属类别即为具有最大关联度的类别。

三、自动分类系统的设计与实现

3.1 关键词抽取模块

从分类文本中有意义地抽取关键词项，同时保存词项相关信息，是非常重要的技术，也是汉语语言处理的基本要求。模块的实现依靠名词和动词，它们是文本中具有代表性的关键词项。模块设计基于具有 78,000 个词条的汉语基本词典。《图 1 》说明模块的构成。

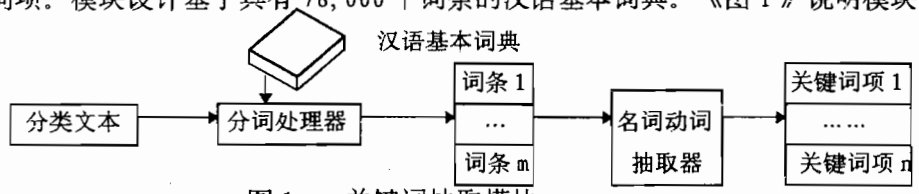

图 1.　关键词抽取模块

3.2 规范化类别特征向量表生成模块

为生成规范化类别特征向量表，需要预定义类别及所对应的编码，义类词典，分类主题词表。《图 2 》说明向量表生成模块的组成。在该图中，输入数据为类别编码，由近 70,000 个词条组成的义类词典，及由 6,780 个各预定义类别词条构成的汉语分类主题词表。

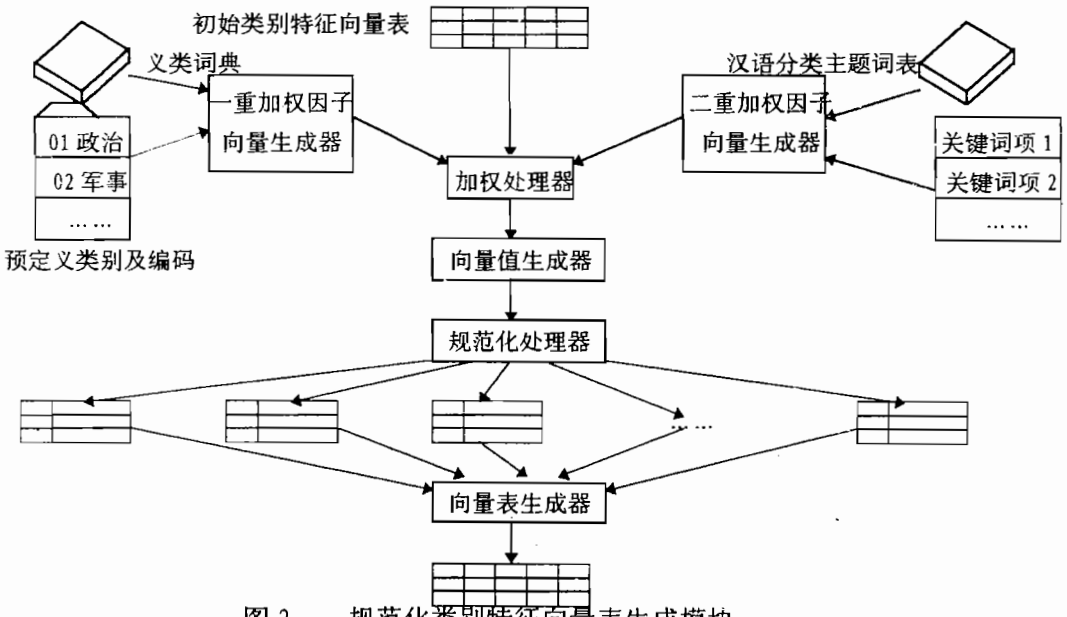

图 2.　规范化类别特征向量表生成模块

3.3 规范化文本特征向量生成模块

《图 3 》简略地描述了规范化文本特征向量生成模块。在实现过程中，可将输入分类文本转换为多维特征向量，进行各项处理，为之后的关联度测量作准备。

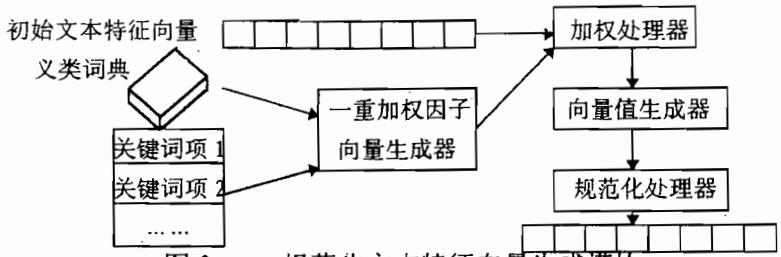

图 3.　规范化文本特征向量生成模块

3.4 自动分类模块

《图 4 》中说明的模块可从早先生成的规范化类别特征向量表，及规范化文本特征向量表中，测量二者之间关联度，为输入的分类文本，分配相关类编码。

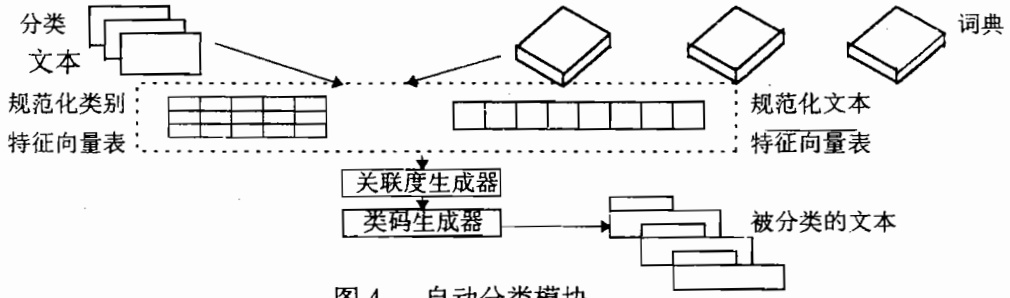

图 4.　自动分类模块

3.5 反馈模块

通过自动分类模块被分配类编码的文本，被用作为输入数据，以提高已构规范化类别特征向量表的性能。反馈模块，通过汉语基本词典与义类词典，可对分类主题词表进行扩充，因此改变各预定义类别与义类词典各类之间的内部联系，有必要重新构造向量表。

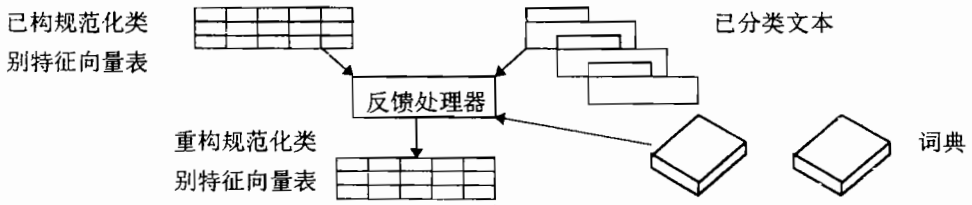

图5.　反馈模块

四、自动分类系统的实验结果

评价与测试文本自动分类算法需要两个重要指标：查全率(recall)和查准率(precision)。查全率是指通过分类算法被正确分类的文本(Correct)占未分类之前属于该类的文本(Alltext)的百分比，查准率是指通过分类算法被正确分类的文本(Correct)占被分类为该类的文本(Classify)的百分比。

首先，通过查阅《同义词词林》、《中国分类主题词表》和《人民日报》新闻语料的已分类情况，将分类体系确定为十二类别。其次，构造规模为近500篇新闻语料的测试集。利用所实现的汉语文本自动分类系统进行实验，《表2》说明了实验结果。

类别	政治	军事	经济	法律	农业	体育	卫生	工业	文化	交通	生活	宗教	合计
Correct	12	26	48	16	43	119	39	43	31	5	15	0	397
Clssify	17	32	53	17	45	119	39	47	43	6	21	2	441
AllText	15	34	58	24	52	138	49	47	33	6	19	0	475
recall	80%	76.4%	82.8%	66.7%	82.7%	86.2%	79.6%	91.5%	93.9%	83.3%	78.9%	0.0%	83.6%
precisi	70.6%	81.2%	90.6%	94.1%	95.6%	100%	100%	91.5%	72.1%	83.3%	71.4%	0.0%	90.2%

表2.　实验结果

其中存在一些文本，由于某种原因而不能被分类处理，但数量不多，可通过人工方式解决。从实验结果可看出，测试指标查全率和查准率数值较高。这与通常信息检索系统中的情况略有差异，其造成与测试集的规模及分类信息的抽取相关。针对目前所建立的预定义类别而言，系统实现性能较好。可见，这种文本自动分类方法，还是比较可行的。

五、结束语

基于人类专家所建立的汉语语言学知识，我们提出一种基于类型与文本特征之间关联度的汉语文本自动分类方法。首先，寻求每一预定义类别及分类文本的特征，将它们作为规范化向量，并通过相应不同的加权处理，说明二者之间的关联度，实现自动确定文本的所属类别。其次，使用分类正确的文本作为更新预定义类别特征向量表，扩充知识库的输入数据。实践证明，该方法基于预定义大类，对于汉语新闻语料的分类，还是切实可行的。

实验之后，我们发现关键词抽取精度的提高，在相当程度上可提高系统性能。所以，提高关键词抽取的正确率，并将抽取领域扩展为名词、动词、名词短语及动词短语的综合，成为我们将来的研究。

参考文献

[1]. Salton.G， Automatic Text Processing： The Transformation： Analysis and Retrieval of Information by Computer， Addison-Wesley， Reading， Mass， 1989.

[2]. Ellen Riloff and Wendy Lehnert， Information Extraction as Basis for High-precision Text Classification， ACM Tramsactions on Information System， vol. 12， No 3， July 1994.

[3]. Mauldin.M， Retrivel performance in Ferret： A conceptual infomation retrieval system， In Proceding of SIGIR 1991， ACM， New York， 1991.

[4]. 刘湘生、候汉清等，中国分类主题词表，华艺出版社，1994。

[5]. 梅家驹、竺一鸣等，同义词词林，上海辞书出版社，1983。

[6]. 吴军、王作英、禹锋、王侠，汉语语料的自动分类，中文信息学报，vol.9，No.4，1995。

[7]. 姚天顺等，自然语言理解，清华大学出版社，1995。