Hans Journal of Data Mining 数据挖掘, 2013, 3, 33-39
http://dx.doi.org/10.12677/hjdm.2013.33007 Published Online July 2013 (http://www.hanspub.org/journal/hjdm.html)

Research of Chinese News Classification Based on Titles*

Haitao Wang[1], Yanqiong Zhao[2], Bang Yue[1]

[1]College of Computer Science & Software Engineering, Shenzhen University, Shenzhen
[2]Network Department, China Mobile Limited (Anhui), Hefei
Email: htwang@szu.edu.cn

Received: May 17th, 2013; revised: Jun. 9th, 2013; accepted: Jun. 19th, 2013

Abstract: Retrieving online information efficiently becomes a crucial issue in nowadays online experience. Compared with traditional news in paper form, online news are faster, more convenient and more flexible. It is a trend that online news are replacing their traditional counterpart and becoming the most common way for people to obtain daily information. However, the volume of frequent updated news becomes so large that the traditional manual news classification cannot meet the needs of online users. One of the solutions for this will be applying automatic text classification technologies to classify online news. Many IT companies are developing automatic news classification systems. There are different forms of network news. Some of the news are composed mostly by graphics or videos instead of text and therefore not able to be coped with by classic text classification. A new approach of news classifier based on news titles is proposed to dealing with such news. In this paper, the title based classification model was created. The model was evaluated by a built corpus and compared with contents based classification. A two-phase news classification system is constructed and category key feature is proposed.

Keywords: Text Classification; Title Classification; News Classification; Semantic Similarity

基于标题的中文新闻分类研究*

王海涛[1]，赵艳琼[2]，岳　磅[1]

[1]深圳大学计算机与软件学院，深圳
[2]安徽移动网络部，合肥
Email: htwang@szu.edu.cn

收稿日期：2013 年 5 月 17 日；修回日期：2013 年 6 月 9 日；录用日期：2013 年 6 月 19 日

摘　要：如何快捷、准确、全面地检索互联网信息是互联网时代的重要问题。网络新闻比传统纸质媒体新闻速度更快、内容更丰富、形式更灵活生动，正逐渐取代传统新闻媒体成为很多人获取新闻信息的主要途径。然而，面对快速更新的大量新闻信息，传统的手工分类方式无法满足用户的需求。新闻的主要内容一般都是以文本的方式呈现，因此，利用文本自动分类技术对网络新闻进行自动分类是解决手工新闻分类问题的一个有效途径。由于网络新闻信息形式多样，很多新闻内容完全是由图片或者视频组成，不包含文本内容。本文提出通过新闻标题对网络新闻进行分类的方法，比通过内容进行分类的方法分类速度更快，并且有更强的适应性，可对无文本内容的新闻(如图片新闻、标题新闻等)进行分类。本文创建了基于标题的文本分类模型；从网络上获取新闻语料，验证模型的工作情况；并通过与基于内容的文本分类方法比较，验证基于标题的文本分类模型的优劣。本文构建了基于标题的两步分类系统，所提出的类别唯一特征，对于可分样本可以实现高分类准确率。

关键词：文本分类；标题分类；新闻分类；语义相似度

*资助信息：国家自然科学基金面上项目，编号 61170076；2010 年深圳市基础研究项目，编号 JC201005280408A。

1. 引言

随着信息技术的发展，特别是互联网技术的发展和普及，网络已经成为人们发布、交流和获取信息的主要途径。然而，网络上的信息正在爆炸性地增长。Google 官方博客[1]指出，Google 检索的独立 ULR 数量已经达到万亿级别，并且 Google 工程师发现，互联网上每天新增网页数量达到数十亿个。

以网络新闻为例，它以更新速度快、内容丰富、形式多样的特点逐渐替代报纸、广播或者电视成为很多人获取新闻的主要来源。然而网络新闻更新快、内容多的优点同时也成为不利于人们阅读的缺点，人们为了找到自己关心的新闻往往要费一番功夫。为了满足阅读者的需求，各新闻网站都在对自己的新闻进行越来越详细多样的分类。然而这些分类基本上都是手工完成的，对于迅速更新的大量新闻需要耗费大量的人力。同时，由于个人的分类标准具有很大的主观性，导致分类结果存在差异。

目前，很多新闻门户网站都在发展自动分类技术的应用，例如谷歌(Google)的“谷歌资讯(Google News)[2]”，将超过 1000 个中文网站的新闻进行汇集，整合相似报道，其网站内容完全是由计算机自动生成的，其中大量使用了文本分类和聚类技术。

由于网络新闻信息形式多样，很多新闻内容完全是由图片或者视频组成，不包含文本内容。本文提出通过新闻标题对网络新闻进行分类的方法，比通过内容进行分类的方法有更强的适应性，可对无内容的新闻(如图片新闻、标题新闻等)进行分类，而且在 RSS 精确阅读等方面可以提供有效的帮助。

本文以网络新闻为例，只通过新闻标题对新闻进行分类，实验语料库从网络新闻中获取，选自 QQ 新闻[3]的 6 类新闻内容：财经、房产、科技、汽车、体育、游戏，总数为 8200 多条。其中 70%作为训练样本，30%作为测试样本。

使用 N 元模型和中文分词两种方式提取新闻标题中的特征，通过训练样本中的特征建立类别的特征表示，实验验证两种方式的分类效果；提取特征中对相应类别具有代表性的特征，定义为唯一特征，通过唯一特征提高分类的准确率；使用新闻内容文本，利用基于 VSM 的余弦距离和基于机器学习的 KNN 文本分类模型对新闻进行分类，通过实验对比基于标题和基于内容的分类速度和分类准确率。

2. 基于 N 元模型的特征选择及实验

N 元模型是一个简单但是非常实用的统计语言模型，它是对统计语言模型的简化。假设一个文本序列为 $W = w_1 w_2 \cdots w_n$，那么想要计算 W 在文本中出现的概率 $P(w)$，需要计算 w_1 到 w_n 的所有词的出现概率，而每一个 w_i 的出现概率都与它前面的 $i-1$ 个词的概率有关，这样计算起来太复杂，如果只与前面的 $N-1$ 个词有关那么就可以大大简化计算，这样简化之后的模型就是 N 元模型。

$$P(w_i) = P(w_i | w_{i-N+1} \cdots w_{i-1}|) \tag{1}$$

其中最常使用的是 2 元和 3 元模型，当 $N=3$，公式可简化为：

$$P(w_i) = P(w_i | w_{i-2} w_{i-1}) \tag{2}$$

尽管这个模型非常简单，但其效果却相当好，远远超过单独使用统计和语法模型，科学家曾试图用别的方法来代替此模型，但都没有获得成功，这本身也是语言模型的一个困惑，即如此简单的一个模型，效果却为何远远超过许多复杂模型[4]。N 元模型只是简单地利用了字和词的同现信息，但是在自然语言处理的很多领域的应用中起着有效的作用，问题在于目前国内外还没有哪一种语言的句法语义分析系统可以胜任大规模真实文本处理的重任。因此，对于世界各国的语言来说，当前的主流技术仍是语料库方法和统计语言模型[5]。

即使经过大大简化后的 N 元模型的时间复杂度依然还是较高，在实际应用中大多不使用单词作为基本的单位，而是使用基于英文短语的 N 元模型来简化计算。David Lewis 认为，英文文本分类中使用优化合并后的词组比较合适[6]。

在本文研究中首先使用 N 元模型作为特征表示，提取语料中的特征。然后通过实验测试 N 元模型对于基于标题的分类效果。

2.1. 特征空间的建立

在使用 N 元模型时，N 值的选择对于效率有着决定性的影响。因为只是对新闻的标题进行实验，鉴于标题的长度都很短，为了更好地提取出标题中的特

征，将 N 值定在 2~5 之间，也就初步建立的特征空间将包含长度在 2 到 5 之间的所有特征。

对于英文的 N 元模型一般在切分的同时进行剪枝的操作，以便除掉切分过程中产生的停用词，减少计算和存储的开销。对于中文来说，由于中文的词语之间没有明确的分隔，所以无法确定在切分过程中产生的汉字序列是否对以后的分析处理有作用，在切分过程中不能进行剪枝的操作。这样就导致在切分结束后产生大量的特征，因此在利用 N 元模型进行切分完成之后，需要使用特征降维方法进行降维。

在本文的实验中，初步建立的特征空间的容量是 262,922 个特征，这其中绝大部分特征的出现频率为 1。在实验中，首先使用 TF 对特征进行降维，降维后的特征数量为 38,606。

2.2. 实验及结果

在实验中，特征权重分别使用 TF 和 TFIDF。两个长度不同的特征，长的特征对于主题的表达作用明显要大于短的特征，因此，对于 TF 权重，不同长度的特征词赋予不同的权重，对应长度为 2、3、4、5 的特征，其特征权重分别为 1、2、4、8。

对于新闻标题与类别的相似度，通过测试样本中包含的特征在类别中的共现频率来确定。同时，因为特征的提取是通过 N 元模型的方式实现的，因此就会出现长特征中包含短特征的情况，在计算过程中，对于同一个标题中的短特征的贡献度要除去包含该短特征的长特征的贡献。例如在一个标题中包含“电脑”和“电脑城”两个特征，它们在某个类别中的共现频率分别为 25 和 20，那么“电脑”的共现频率就应该修改为 5。

在计算 TFIDF 权重时，由于标题长度很短，把一个标题作为一个文档来处理不合适，因此在计算过程中，把整个类别包含的所有特征作为一个文档，来计算特征相对于每个类别的 TFIDF 权重。

从表 1 的实验结果可以看出，使用 N 元模型进行特征切分的分类准确率不高。在使用 TFIDF 权重之后，对于 N 元模型的影响效果不大，在本身分类结果精度不高的情况下，只提高了 1 个百分点。这说明，使用 N 元模型对特征进行切分从而产生的特征空间对于基于标题的新闻标题不适合，下面将采用基于中文分词工具的特征选择。

Table 1. The classification result based N-gram
表 1. 使用 N 元模型的分类实验结果

	样本数	使用 TF 权重	使用 TFIDF 权重
财经	450	77.33%	76.67%
房产	240	65.42%	64.17%
科技	540	72.78%	76.30%
汽车	744	75.27%	76.34%
体育	216	49.54%	50.00%
游戏	121	72.73%	76.86%
微平均	2311	71.53%	72.70%

3. 基于 ICTCLAS 中文分词的特征选择及实验

中文自动分词是中文信息处理的基础，在中文信息检索、中文自动翻译等领域被广泛使用。与英文词语之间有空格不同，中文词语之间没有明显的分界，而中文词语比中文汉字拥有更多的信息。因此，为了更有效地处理中文文本，首先需要对中文文本进行自动分词，将由汉字组成的字串正确切分为中文词语序列。

3.1. ICTCLAS 中文分词

本文选择开源项目 ICTCLAS[7]作为分词组件。ICTCLAS 是中国科学计算技术研究所在多年研究积累的基础上研制的汉语词法分析系统。主要功能包括中文分词、词性标注、命名实体名、新词识别等。该系统分词速度快、精度高，其最新版本 ICTCLAS 3.0 的分词速度单机 996 KB/s，分词精度达到 98.45%，是当前世界上最好的汉语词法分析器[7]。

3.2. 特征权重的确定

训练样本经过分词处理后去除停用词，包括虚词(如连词、叹词、拟声词、助词、标点、语气词等)和表示媒体类别的名词(如组图、视频等)，形成特征词表 $V=(t_1,t_2,t_3,\cdots,t_n)$。

通过对特征词表中的词汇进行加权处理后形成类别的一般特征。类别 C_i 的一般特征可以表示为 $P_i=(\langle t_1, w_{i1}\rangle,\langle t_2, w_{i2}\rangle,\langle t_3,w_{i3}\rangle,\cdots,\langle t_n,w_{in}\rangle)$。要确定特

征词的权重应该考虑以下因素：1) 特征词在一个类别中的出现次数，出现的次数越多说明该特征词对该类别的影响越大；2) 包含一个特征词的类别个数，类别个数越多，特征词的影响越小；3) 特征词的长度，长度越大特征词的影响越大；4) 训练样本的个数对特征词的出现次数也有决定性的影响，应予以考虑。

定义一个特征词 t_i 在类别 C_j 中的权重 w_{ij} 为：

$$w_{ij} = tf_{ij} \times icf(t_i) \times lw_i \times cw_j \quad (3)$$

其中，tf_{ij} 为特征词在类别 C_j 中的出现次数，$icf(t_i)$为特征词 t_i 的逆类别频率值，lw_i 为特征词 t_i 的长度权重，cw_j 为类别 C_j 的调整参数。下面分别介绍各个部分的计算方法：

1) 逆类别频率 $\mathrm{icf}(w_i)$。借用逆文档频率 idf 的计算方法来计算逆类别频率 icf：

$$icf(w_i) = \log(N/cf_{ij}) \quad (4)$$

其中，N 为类别总数，cf_{ij} 为出现特征词 t_i 的类别数。

2) 长度权重 lw_i。不同长度的特征词对于分类的影响显然是不同的，尤其是标题本身的长度就很短，一个长度为 5 的特征词比一个长度为 2 的特征词对决定该标题所属类别的重要性要大的多。规定特征词的长度 l 和长度权重 lw_i 的关系如表 2 所示。

3) 类别调整参数 cw_j。三个彼此相关的参数可以影响 cw_j，分别为：该类别的训练样本数 CoS，出现在该类别中的特征词数 CoF，出现在该类别中的特征词词频之和 SoF。为了确定类别调整参数，分别使用上述三个参数的指数函数作为类别调整参数进行了实验。通过图 1 的实验结果可以看出，当 $cw = \mathrm{CoS} - 0.5$ 和 $\mathrm{cw} = \mathrm{SoF} - 0.6$ 时，取得最佳的分类准确率。

4. 类别唯一特征

在特定的情况下，有时根据标题中的一个特征词就能基本确定标题所属的类别，例如：包含奇瑞、雅阁、凯美瑞等词语的标题一般可以确定为汽车类新闻。这样的特征词本文定义为类别的唯一特征。

选择特征词表中的人名、地名、团体名、其他专有名词、英文名词、简称、习用语作为备选的唯一特征，然后根据各备选唯一特征在各个类别中的出现次数和出现次数所占所有类别中出现次数之和的比率，来确定其是否成为唯一特征。设类别的一般特征向量

Table 2. The relation between the length of features and weights
表 2. 特征词的长度 *l* 和长度权重 lw_i 的关系

l	2	3	4	5	6	7	≥8
lw_i	1	2	4	8	12	14	$L+7$

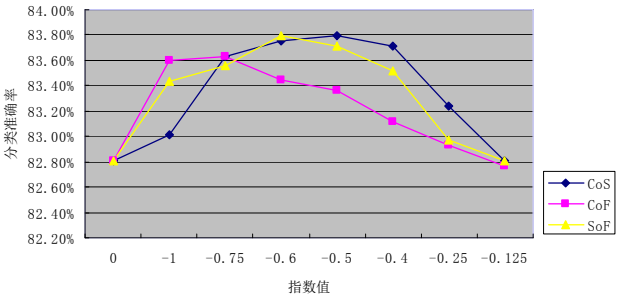

Figure 1. Experiment result about class adjustment parameter
图 1. 类别调整参数实验结果

为：

$$P = (\langle t_1, w_1 \rangle, \langle t_2, w_2 \rangle, \langle t_3, w_3 \rangle, \cdots, \langle t_n, w_n \rangle)$$

对于特征词 t_i，设其词频为 tf_{ij}，如果 tf_{ij} 满足条件：

$$tf_{ij} \geq \Theta, \left(tf_{ij} \Big/ \sum_{k=1}^{N} tf_{ik} \right) \geq \theta$$

则特征词 t_i 称为唯一特征，其中 Θ 和 θ 为预先确定的阈值，N 为类别数。一般特征向量 P 的唯一特征向量为：

$$U = (\langle t_1, f_1 \rangle, \langle t_2, f_2 \rangle, \langle t_3, f_3 \rangle, \cdots, \langle t_n, f_n \rangle)$$

其中 f_i 是特征词 t_i 的唯一性权重，如果 t_i 是唯一特征，那么

$$f_i = \left(tf_{ij} \Big/ \sum_{k=1}^{N} tf_{ik} \right) \times w_i \quad (5)$$

否则 $f_i = 0$。

选择不同的频率阈值 Θ 和比率阈值 θ 进行实验。测试样本数量为 2311。

从表 3 的实验结果可以看出，在 θ 等于 0.90 时，单独使用类别唯一特征进行分类可以取得非常高的平均准确率；随着阈值 Θ 和 θ 的增大，分类平均准确率增高，同时可分样本数相对于测试样本 2311 的总数量逐渐降低。

5. 基于标题的两步新闻分类系统

从上一节的实验结果可以看出，单独使用唯一特征对于可分样本可以达到很好的分类效果，但是很多

的样本无法通过唯一特征进行分类。通过两步分类可以充分利用唯一特征分类准确性的优势，同时对所有的样本进行分类。

对于某一新闻标题，因其长度很短，所含词汇很少，容易得到其一般特征向量 $TP=\left(\langle t_1,h_1\rangle,\langle t_2,h_2\rangle,\langle t_3,h_3\rangle,\cdots,\langle t_n,h_n\rangle\right)$。其中 h_i 一般为 1 或者 0，如果新闻标题包含特征 t_i，h_i 值为 1，否则 h_i 值为 0。

如图 2 所示，分类过程由两步组成：

1) 计算标题一般特征向量 *TP* 与类别唯一特征向量 *U* 的相似度。如果标题中包含类别唯一特征，标题就可分，则分类完成，否则通过下一步对标题进行分类；

$$Sim1(T,C)=Sim(T_P,U)=\sum_{i=1}^{n}h_i\times f_i$$

2) 计算标题一般特征向量 *TP* 与类别一般特征向量 *PN* 的相似度，确定标题所属的类别；

$$Sim2(T,C)=Sim(T_P,P)=\sum_{i=1}^{n}h_i\times w_i$$

下面的实验首先只使用 TFIDF 的一般特征进行分类，然后在唯一特征频率阈值 $\Theta=10$ 和比率阈值 $\theta=1.00$ 时，通过两步分类对测试样本进行分类。从表 4 的结果中可以看出，使用两步分类方法对分类结果有一定的改进，平均准确率为 89%，在分率精度上可以满足实际应用的需求。

6. 基于标题与基于内容的实验对比与分析

基于内容的文本分类实验采用与基于标题的分类实验对应的新闻内容语料。实验分别采用基于 VSM 的余弦距离分类算法和 KNN 分类器。Salon 等人提出的向量空间模型(VSM)[8,9]被广泛地应用在了信息检索和文本分类领域，使一串离散的文本能够以一个向量的方式来表示，现在已经成为最简便高效的文本特征表示模型之一。以选择 KNN 分类器进行基于内容的分类实验，是因为 KNN 分类器虽然是简单易行的文本分类器，但是分类效果良好，对于不同数据集都有很好的可操作性，被广泛地应用于基于统计的机器学习中[10]。

6.1. 特征权重和特征选择

在进行文本分类的过程中，因为出现在文本不同

Table 3. The classification result based unique features
表 3. 使用唯一特征的分类结果

Θ	θ	可分样本数	正确样本数	平均准确率
3	0.50	1289	1123	87.12%
	0.60	1127	1040	92.28%
	0.70	1073	1013	94.40%
	0.80	971	934	96.19%
	0.90	840	820	97.62%
	1.00	747	734	98.26%
5	0.50	1106	971	87.79%
	0.60	954	887	92.98%
	0.70	914	867	94.86%
	0.80	842	811	96.32%
	0.90	728	715	98.21%
	1.00	633	627	99.05%
10	0.50	872	697	89.13%
	0.60	697	658	94.40%
	0.70	677	646	95.42%
	0.80	612	597	97.55%
	0.90	565	556	98.41%
	1.00	466	464	99.57%

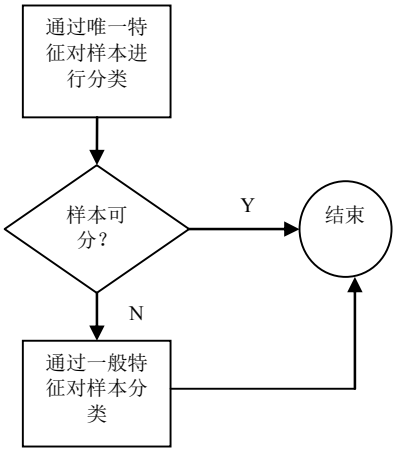

Figure 2. The processing of two steps classification algorithm
图 2. 两步分类算法流程图

Table 4. The classification result based normal and unique features
表 4. 综合使用一般特征和唯一特征的分类结果

	样本数	只使用 TFIDF	一般特征和唯一特征
财经	450	74.67%	78.67%
房产	240	65.41%	88.75%
科技	540	86.11%	89.26%
汽车	744	93.10%	93.28%
体育	216	84.26%	95.37%
游戏	121	77.69%	90.91%
微平均	2311	83.47%	89.10%

地方的特征词具有不同的影响力，对于不同位置的特征词应该赋予不同的特征权重。例如出现在标题中的特征词因为更能代表该文本的主题，应该赋予最大的特征权重；而出现在首段的特征词相对于其他段落的权重要大；而对于一个段落中，出现在段首、段中、段尾的特征词，也将赋予不同的权重；一些科技类论文还拥有摘要和关键词，在处理这类文本的权重时对这些位置出现的特征词，其权重都应该相应地加大。

在本文应用中，对于一般特征词的权重设定为 1；如果出现在段首和段尾将其特征权重加 1；如果出现在首段和末段，其特征权重加 1；如果出现在标题中其权重将为最大值 5。因此对于出现在不同地方的特征词，它的特征权重根据重要性的大小分别被赋予 1、2、3 和 5。

在特征降维阶段，采用互信息(MI)的方法。由于互信息对于低频特征词的过渡拟合现象会导致低频特征词评价过高问题，因此在实际使用中首先通过特征词频 *TF* 过滤掉低频特征词，再计算各个特征词的互信息大小。选择互信息大于指定阈值的词作为特征词汇表中的特征词。

6.2. 实验及结果分析

与基于内容的文本分类结果从两个方面来进行比较：1) 分类准确率；2) 分类速度。

从表 5 的实验结果中可以看出，基于标题的分类结果好于基于内容的 KNN 分类结果。基于内容的 KNN 分类结果各个类别的分类准确率比较平均，而基于标题的分类结果对某些类别的分类效果明显要好于其他类别，这是由于基于标题的分类方法对于某些类别的标题敏感度比较高，对另一些类别要差一些；而基于内容的 KNN 分类由于使用了新闻的全部文本内容，能够更好地表现样本的主题。

下面将进行分类速度与分类准确率的综合实验对比。其中基于内容的分类实验使用两种分类方法，一种使用上面的 KNN 分类器，另一种使用基于 VSM 余弦距离的分类算法，以下简称 VSM 分类算法。

在进行 VSM 分类实验时，使用所有训练样本的权重之和来表示类别的特征权重。类别与测试样本的相似度通过类别向量与测试样本向量的余弦值来计算。

从表 6 的实验结果可以看出，对于分类精度，基于标题的分类结果最好，KNN 次之，VSM 最差；对于分类所用的时间，基于标题的时间最少，VSM 次之，KNN 所用时间最长。

Table 5. The comparison of classification result between basing content and basing titles
表 5. 基于内容 KNN 分类与基于标题分类结果对比

	测试样本数	基于内容(KNN)		基于标题	
		正确样本	准确率	正确样本	准确率
财经	450	385	79.56%	358	79.56%
房产	240	207	86.25%	213	88.75%
科技	540	465	86.11%	482	89.26%
汽车	744	675	90.73%	693	93.15%
体育	216	189	87.50%	207	95.83%
游戏	121	103	85.12%	112	92.56%
微平均	2311	2024	87.58%	2065	89.36%

Table 6. The comparison of time-consuming between basing content and basing titles
表 6. 基于内容与基于标题分类的用时对比

基于内容				基于标题	
VSM		KNN			
平均准确率	分类用时	平均准确率	分类用时	平均准确率	分类用时
83.75%	285s	87.58%	435s	89.36%	100s

KNN 与基于标题的分类精度在上面已经作了分析，对于 VSM 和 KNN 的分类效果差别是因为 VSM 只是简单地计算了类别特征向量与测试样本的特征向量的余弦值，而类别特征向量的设置完全通过所有训练样本的特征向量来确定，这样就将个别训练样本的噪声全部累加到类别特征向量之中，导致类别特征向量的噪声增大。

对于分类所用时间的差异，可以从对样本的处理(包括中文分词和特征向量的生成)时间和分类算法分类所用的时间两方面来考虑。对比 VSM 与基于标题的分类实验，两者在第二步所作的操作所用的时间差别不大，所以主要的用时差异在第一步。对于基于标题的分类，两者处理样本的用时主要是由样本文本长度决定的。由于标题的长度明显短于文本内容的长度，最终导致 VSM 的分类用时是基于标题的用时的 2.85 倍。

因为 VSM 与 KNN 在处理样本的时候所做的操作基本相同，因此对于 VSM 与 KNN 的分类用时的差异

主要是由分类阶段用时差异决定的。在分类阶段，VSM 只需要计算样本特征向量与类别特征向量的相似度，而 KNN 要计算样本特征向量与所有训练样本特征向量的相似度，从而导致 KNN 用时比 VSM 的用时有了巨大的增加。

对于本文中 KNN 分类所用时间过长是由于在分类阶段计算了测试样本与所有训练样本的相似度，最后选择相似度最小的 K 个训练样本，通过它们所属的类别来确定测试样本的类别。可以通过一些改进算法来加快这一步分类过程，这方面的研究比较多。文献[11]中在 KNN 算法中加入了训练阶段，将训练结果保存到特殊设计的数据库中，在分类阶段通过检索引擎快速返回 K 个最邻近的结果；通过优化检索引擎可以在百万级训练样本的情况下用几百毫秒返回搜索结果。文献[12]中提出了一种基于核的 KNN 思想，在训练阶段将每个类别的训练样本聚类为多个小的类别，用这些小的类别的中心代表全部训练样本，而这些中心就是类别的核，在分类阶段只需要计算测试样本与这些类别的核的相似度，从而加快分类速度。

由于本文的研究重点不在这里，所以使用了最简单的办法来进行 KNN 分类。如果使用上面提到的第一种改进算法，KNN 分类所用的时间将会与 VSM 分类所用的时间接近。

参考文献 (References)

[1] The Official Google Blog. http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

[2] 谷歌资讯(Google News)[URL]. http://news.google.cn

[3] QQ 新闻[Z]. http//news.qq.com

[4] E. I. Sicilia-Garcia and F. J. Smith. Statistical language modeling. Encyclopedia of Library and Information Science, 2002, 71(34): 309-338.

[5] 黄昌宁. 统计语言模型能做什么?[J]. 语言文字应用, 2002, 1(2): 77-84.

[6] D. D. Lewis. Representation and learning in information retrieval. University of Massachusetts, Amherst, 1992.

[7] ICTCLAS 中文分词工具[URL]. http://ictclas.org

[8] S. Chakrabarti. Hypertext databases and data mining. *Proceedings of the* 1999 *ACM SIGMOD International Conference on Management of Data*, 1999, 28(2): 508.

[9] G. Salton, M. J. McGill. Introduction to modern information retrieval. New York: Mc Graw Hill, 1983

[10] Y. Yang, J. O. Pedersen. A comparative study on feature selection in text categorization. Morgan Kaufmann Publishers, Burlington, 1997: 412-420.

[11] 张庆国, 张宏伟, 张君玉. 一种基于 k 最近邻的快速文本分类方法[J]. 中国科学院研究生院学报, 2005, 22(5): 554-559.

[12] 刘斌, 黄铁军, 程军等. 一种新的基于统计的自动文本分类方法[J]. 中文信息学报, 2002, 16(6): 18-24.

