DOI: 10.13657/j.cnki.gxkxyxb.2005.s1.020

基于分词和基于 N-Gram的网页分类系统比较研究

A Comparative Study of Word-Segment and N-Gram Categorization System

高伟锋,刘连芳

Gao Weifeng, Liu Lianfang

(南宁市平方软件新技术有限责任公司,广西南宁 530003)

(Nanning Pingsoft New Technology Co. Ltd., Nanning, Guangxi, 530003, China)

摘要: 设计并实现一个网页分类系统,采用相同的特征权值计算方法,特征选择算法以及分类算法,进行基于分词的网页分类系统和基于 N-Gram的网页分类系统的对比实验,分析两者的分类效果。结果表明,基于 N-Gram的网页分类系统能达到并在一定程度上高于基于分词的网页分类系统的效果。

关键词: 中文网页 分类 N-Gram 分词 KNN

中图法分类号: TP393.092 文献标识码: A 文章编号: 1002-7378(2005)S0-0058-03

Abstract This page designs a Chinese web categorization system, with the same feature weight, feature selection and categorizing algorithm, based on Word-Segment categorization system and N-Gram categorization system. The experiment demonstrates that being based on N-Gram categorization system has the same effect as being based on Word-Segment categorization system, which is more effective in some aspects.

Key words chinese web, categorization, N-Gram, word-segment, KNN

随着 Internet的快速发展,Web上的网页数目正在飞速发展。为了能从中提取出有效的信息,我们需要对网页进行分类。虽然从现在的技术来看,人工分类的效果优于自动分类,但是自动分类的效率要大大优于人工分类。针对 Web上的海量信息,网页自动分类便成了快速有效的组织网络上海量信息的一个重要技术。

网页分类是按照一定的类别体系,对文本进行自动标识,给一个待识别的文本给予一个或多个类别标识。目前文本分类的研究主要采用的向量空间模型。在向量空间模型中,特征值的选择一般可以选择字、词、词组等。通常情况下选择词作为文档特征进行文档分类,这是因为词是表征语义的最小语言单元。由于中文文字的特征,词与词之间没有很明显的截断,需要用其它技术进行区分。基于字典的匹配方法是当前最为常用和成熟的中文自动分词技术,它的目的是通过对文本进行分词处理,从而得到该文本的词条集合,即可以成为向量空间模型中的向量维。由于分词方式建立在词典完备的理论假设之下,所以它会带来许多问题,其中一个比较重要的问题是对于未登陆词的处理上。一本词典不可能包括所有的词,而且随着 Internet的发展,新词会层出不穷,这样基于词典的方法就没有办法处理新词。另外,还存在词的切分歧义问题。为此我们考虑利用文档的 N-Gram作为文档特征,和基于词文档特征的分类系统进行比较。为了体现两者最终的效果上的真实区别,我们设计并实现一个网页分类系统,采用相同的特征权值计算方法,特征选择算法以及分类算法,进行基于分词和基于 N-Gram的网页分类系统的对比实验,对比分析两者的分类效果,希望能找出解决现有的基于分词的分类系统问题的方法。

收稿日期: 2005-09-07

作者简介: 高伟锋(1976-),男,广西横县人,助理研究员,主要从事中文信息处理研究和软件开发工作。

1 N-Gram信息简介

对于长度为 L 的中文文档 d,如果不考虑标点符号和其它各种字符,也就是说,文档是长度为 L 的汉字序列,那么,这个文档中包含的N-Gram信息项总共为 $L(L+1)/2$ 由此可见,文档中包含的 N-Gram信息项非常丰富。所以在用 N-Gram信息进行文档分类时,必须有所选择。

不是所有出现在文档中的 N-Gram信息项对分

类有用。一个 N-Gram信息项对分类的有用性可以从频度、分散度和集中度来衡量。下面分别给出它们的定义。

定义 **1**(频度) 在文档 d中, N-Gram信息项 t的频度用它在 d中出现的次数 tf表示。

定义 **2**(分散度) 在文档类 c中, N-Gram信息项 t的分散度用 c中包含 t的文档数目 df表示。df越大,则 t在 c中越分散;反之,越不分散。

定义 **3**(集中度) 在文档集 D中, N-Gram信息项 t的集中度用 D中包含 t的文档数目 cf表示。cf越小,则 t在 D中越集中;反之,越不集中。

直观地,对于 N-Gram信息项 t,其频度越高、分散度越大、集中度越强,则对分类越有用。由于现在没有很好的数学方法能证明如何选取频度、分散度和集中度这三个因素,使得选出的文档特征能获得最优的分类效果。我们只能根据自己的经验,确定如下两个条件:

(1)在文档 d中某个 N-Gram信息项 t被提取的条件是它在 d中的 $tf \geqslant$ 预先给定的最小频率值 min-tf。

(2)在文档类 c中某个 N-Gram信息项 t被提取的条件是它在 c中的 $df \geqslant$ 预先给定的最小分散度值 min-df。

在我们的实验中,我们选取 min-tf 和 min-df 为 2。N-Gram一个比较重要的问题是 N的取值,在中文文档中,我们知道大部分词的长度一般不会超过 4个字。根据现代汉语频率词表的统计,在最常用的 9000个词中,单字词占 26.7%,二字词占 69.8%,三字词占 2.7%。这个结果显示,在汉语中二字词占绝大部分,而且三字词,四字词也可以由二字词构成。所以我们考虑可以选择 $N = 2$。

2 网页分类系统设计

为了进行对比实验,我们设计并实现了一个网页分类系统。网页分类由训练器和分类器构成。在训练器中,首先将训练文本集进行特征分解(基于分词的系统通过分词进行分词处理,将词作为特征;基于 N-Gram的系统则选取 2-Gram作为特征),并计算特征的权重,得到文档特征向量;通过特征选择算法从文档特征向量中抽取一个最优的文档特征向量。在分类器中,首先将测试文本用最优文档特征向量表示,在经过分类器分类,得到测试文本的所属的类别。

2.1 向量空间模型

采用由 Salton提出的向量空间模型(VSM)[1],研究文本分类。该模型以特征项作为文档表示的基本单位,特征项可由字,词和短语组成。在此基础上,建立文档向量空间模型,以基坐标轴,把文档表示为 M维向量,再用两个向量的点积或夹角的余弦来表示不同文档之间的相关程度。

2.2 文档特征权重

根据文献[2]比较,我们只要选择相同的权重计算方法就可以比较基于分词的分类系统和基于 N-Gram的分类系统,所以我们选择 TF-IDF权重计算函数作为权重计算函数来研究文档的特征权重,以提高查询的查全率和查准率。选择的特征权重函数见表 1。

表 **1** 特征权重函数

名称	权重函数
Boolean Weighting	1,如果 $f_{ik} > 0$ 0,其他
NNN	f_{ik}
TD-IDF	$f_{ik} \times \log(\frac{N}{n_i})$
TFC	$\frac{f_{ik} \times \log(\frac{N}{n_i})}{\sqrt{\sum_{j=1}^{M}[f_{ik} \times \log(\frac{N}{n_i})]^2}}$
ITC	$\frac{\log(f_{ik}+1.0) \times \log(\frac{N}{n_i})}{\sqrt{\sum_{j=1}^{M}[\log(f_{ik}+1.0) \times \log(\frac{N}{n_i})]^2}}$
ENTROPY	$\log(f_{ik}+1.0) \times (1+\frac{1}{\log(N)}\sum_{j=1}^{N}[\frac{f_{ij}}{n_i}\log(\frac{f_{ij}}{n_i})])$
Okapi	$\frac{tf}{0.5+1.5\frac{dl}{avg\_dl}}\log(\frac{N-df+0.5}{df+0.5})$

2.3 特征选择

一般向量空间模型所构造出来的向量的维数通常都很高。高维的特征向量不但会影响系统的效率,而且各个特征对分类的贡献也不一样,例如“的”,“是”等词在每一个类别都会出现很多,而且这类词对分类没有任何贡献,应当取掉。为了提高分类算法的效率,需要在不牺牲分类准确性的前提下降低特征空间的纬度。我们在测试中选取 CHI方法[3]为基本的特征项选取算法。

$$x^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (C+D)},$$

其中: A为 t和 c同时出现的次数; B为 t出现而 c没有出现的次数; C为 c出现而 t没有出现的次数, D为

t和 c同时没有出现的次数;N 为所有训练文档数。

2.4 分类算法

文档自动分类算法采用 KNN分类算法[4],设置 $k = 0$ KNN 算法的定义如下:

$$f(D, c_j) = \sum_{d_i \in kNN} sim(D, d_i) y(d_i, c_j) - b_j,$$

其中,D为一篇待分类网页的向量表示;d为训练集中的一篇实例网页向量表示;c_j 为一类别;$y(d, c_j) \in \{0, 1\}$(当 d属于 c_j时取 1;当 d不属于 c_j时取 0);b_j 为预先计算得到 c_j的的最优截尾阈值;$sim(D, d_i)$ 为待分类的网页与网页实例之间的相似度计算,我们用的是夹角余弦。

KNN算法的分类器不需要使用训练集进行训练,训练时间复杂度为 0 KNN分类的复杂度和训练集中文档数目成正比,也就是说,如果训练集中的文档总数为 n,那么 KNN 分类时间复杂度为 $O(n)$。

3 实验结果与分析

实验采用的是北大天网提供的数据集,它包括 10558个训练网页实例和 2784个测试网页实例,详见表 2

表 **2** 训练集和测试集的分布情况

类别编号	类别名称	训练集	测试集
01	人文与艺术	395	101
02	新闻与媒体	120	18
03	商业与经济	813	211
04	娱乐与休闲	1479	367
05	政府与政治	286	82
06	社会与文化	1063	290
07	教育	282	82
08	自然科学	1551	412
09	社会科学	1501	403
10	计算机与因特网	828	217
11	医疗与健康	2240	601
合计		10558	2784

从表3的分类结果来看,基于N-Gram的分类系统除了新闻与媒体之外其分类正确率和召回率均等同,甚至高于基于分词的分类系统,特别对于政府与政治以及社会与文化两个类别,基于 N-Gram的系统比基于分词的系统要高出 10多个百分点。由此可知,基于 N-Gram的系统能达到基于分词的分类系统的分类效果,对一些类别的分类还高于基于分词的分类系统的分类效果。

表 **3** 分类结果

类别名称	类别文章数	分词		N-Gram	
		正确率*	召回率*	正确率	召回率
人文与艺术	101	0.6129	0.5644	0.7262	0.6040
新闻与媒体	18	0.6111	0.6111	0.5294	0.5000
商业与经济	211	0.6447	0.6019	0.7326	0.6493
娱乐与休闲	367	0.8045	0.7847	0.8798	0.8174
政府与政治	82	0.5195	0.4878	0.6986	0.6220
社会与文化	290	0.4424	0.4103	0.6932	0.6310
教育	82	0.6282	0.5976	0.6974	0.6463
自然科学	412	0.7778	0.7476	0.8226	0.7767
社会科学	403	0.7870	0.7519	0.8032	0.7395
计算机与因特网	217	0.8505	0.8387	0.8900	0.8571
医疗与健康	601	0.8733	0.8486	0.8794	0.8369

* 正确率= 分类正确文本数 /实际分类的文本数;召回率= 分类正确的文本数 /应有所有文本数

4 结束语

本文应用相同的中文网页训练集和测试集,比较研究了基于 N-Gram的分类系统和基于分词的分类系统。得出的结论是基于 N-Gram的系统能达到基于分词的分类系统的分类效果,在某些类别还高于基于分词的分类系统的分类效果。从我们实验的结果,采用基于 N-Gram的分类系统能取代基于分词的分类系统,它能有效的解决中文网页中不断出现新词和词切分异常的问题。在中文网页分类领域内能取得比较好的效果。我们下一步工作是进一步提高现有系统的分类正确率和基于 N-Gram分类系统的效率。

参考文献:

[1] 刘少辉,董明楷,等.一种基于向量空间模型的多层次文本分类方法[J].中文信息学报, 2001, 16(3): 8-26.

[2] 吴 科,石 冰,卢 军,等.基于文本集密度的特征选择与权重计算方案[J].中文信息学报, 2004, 18(1): 42-47.

[3] Yang Yiming, Pedersen J O. A comparative study on feature selection in text categorization [D]. Proceedings of the Fourteenth International Conference On Machine Learning(ICML 97), 1997.

[4] Yang Yiming, Liu Xin. A re-examination of text categorization methods [D]. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR 99), 1999. 42-49.

(责任编辑: 邓大玉)