文本自动分类的模糊方法

王小华　　陆　蓓　　张国煊
杭州电子工业学院，杭州市文一路65号，310037
wxhhie@yahoo.com.cn

摘要：本文提出基于统计的文本分类特征词的自动提取方法，使特征词反映文本分类的类别特征，系统能通过自学习完善分类特征关键词；同时本文探讨模糊文本自动分类模型，提出一种模糊文本自动分类模型用的隶属度与语义关联度的计算方法，使分类系统具有较高的精度。

关键词：　模糊　文本分类

An Automatic Fuzzy Text Categorizing Model

Wang Xiaohua　　Lu Bei　　Zhang Guoxuan
Hangzhou Institute of Electronics Engineering, Hangzhou, 310037, China.P.R
wxhhie@yahoo.com.cn

Abstract: In this thesis, an automatic extract method of text-categorizing words is put forward on statistical basis. Thus, the extracted word can embody the characteristics of each categorized text. The system will perfect by self-learning the key word to categorizing. Meanwhile, an automatic fuzzy text Categorizing is also discussed in this thesis. A calculation method is suggested about the membership and semantic association for the automatic fuzzy text categorizing so that the categorizing system can attain comparatively high precision.

Keywords: Fuzzy, Text Categorizing

1．文本分类技术的发展

随着 Internet 的迅猛发展，网上信息量急剧增加，信息分布全球化，信息结构更趋复杂；能否快速、准确地检索到所需要的信息资料，是人们普遍关心的问题。衡量文献检索系统优劣的性能指标，通常由检索召回率（recall）和检索精度（precision）[1] 来度量。检索召回率是指检索得到的相关文献占实际符合用户要求的总文献数的比例。检索精度是指检索得到的文献中符合用户要求的相关文献所占的比例。在实际应用中，召回率和检索精度通常是较难兼顾的。要想同时达到高召回率和高精度是困难的。许多优秀的检索系统建立在良好的文本分类上，如 Yahoo 的 WWW 索引系统在对下载的 Web 文档进行索引前，都对文档分类处理；分类后检索，可降低非用户要求的相关文献所占的比例，为信息检索打好基础，提高检索系统的性能。

文本自动分类技术产生和发展的原因之一，是为了提高文本检索的精度和速度。文本自动分类本身也是一种文献检索的手段，与普通的文本检索不同的是文本自动分类技术预

先设定了一个类别集合，对检索的文本，根据一定的判别法则，判断其是否属于这个集合中的某个类[3]。这种文本分类技术对 Internet 网上搜索有很大的实用价值。显然，如果能够使用文本自动分类技术对检索结果进行过滤，剔除掉无关的文献，无疑将有效提高网上文献检索的精度。此外，文献分类对于语料库语言学的发展也将有很大的推动作用。随着语料库语言学的发展,要求语料库的规模越来越大;而电子出版业的迅速发展，使获取大量的电子文本建立大规模语料库已成为可能。但语料处理的速度却相对落后于语料收集的速度。因为收集来的粗语料是杂乱无章的，在加工整理前必须进行分类处理，目前对粗语料的分类处理过程仍然是以手工为主，不但效率低，而且对从事分类的工作人员水平要求较高，如果能够代之以自动分类，无疑将大大加快语料处理的速度。此外，对图书馆的电子文本资料管理的研究等问题，都促使对文本的自动分类技术进行研究。

文本自动分类技术尚处研究阶段，国外有几个实验系统，如美国马萨诸塞大学曾使用信息提取技术从样本文献内提取特征信息，然后根据一定的算法将待处理的文献进行分类。中文文献自动分类研究目前还处发展阶段， 96 年，吴军等在《中文信息学报》上讨论了有关“汉语语料的自动分类”[11]。98 年，刘开瑛在 “中文文本中抽取特征信息的区域与技术”[4]一文中，讨论了如何从各种文本抽取特征信息。99 年，何新贵等发表了“中文文本的关键词的自动抽取和模糊分类方法”[5]，并将文本分类方法实际应用于“全国政协提案处理”课题。 南京大学软件新技术国家重点实验室的邹涛专题讨论了“Web 信息的采集、文档的识别与分类”[6]。

2. 文本自动分类方法

文本自动分类技术预先设定了一个类别集合，每个子集具有类似的内容特征，在词级表现为相同词条特征；文本自动分类技术根据文本特征，依据一定的判别法则，判断其是否属于这个集合中的某个类。目前的文本分类方法有模糊分类法、向量空间法、距离分类法等。

1. 模糊分类方法[5]

基于语义关联度的模糊分类法是利用模糊集间的语义距离或语义关联度来进行文本的分类。任一文本或文本类都可通过它的特征关键词来描述它的内容特征的。因此可用一个定义在特征关键词类上的模糊集来描述它们。

设 $L = \{ l_1 , l_2 , \cdots , l_n \}$ 为由 n 个特征关键词组成的论域，则任一文本或文本类可用定义在特征关键词论域 L 上的一个模糊集来描述：

$F = \{ u_1/l_1 , u_2/l_2 , \cdots , u_n/l_n \}$

其中 $l_1 , l_2 , \cdots , l_n$ 为特征关键词，$u_1 , u_2 , \cdots , u_n$ 为每个特征关键词的隶属度，隶属度可以是重要性、频度和代表性等。

设文本类别集合 $S = \{ C_1 , C_2 , \cdots , C_m \}$，其中 C_i 表示第 i 类上的文本集。定义在第 k 类上的模糊集记为：

$F_k = \{ u_{k,1}/l_1 , u_{k,2}/l_2 , \cdots , u_{k,n}/l_n \}$

其中 k=1，2，…m，　某些 $u_{k,i}$ 可以为 0，表示特征关键词 l_i 对第 k 类分类无贡献。

设待分类文本 T 的模糊集记为：

$F_T = \{ u_{T,1}/l_1 , u_{T,2}/l_2 , \cdots , u_{T,n}/l_n \}$

则判分类文本T 所属的类别可以通过计算文本T的模糊集 F_T 分别与这 m 个文本类的模糊集 F_k（k=1，2，…m）的语义关联度 SR 获得，关联度越大，则说明语义关系越密切。满足

$SR(F_T, F_j) = MAX\ SR(F_T, F_k)$　　（k=1，2，…m）

的 j 表示文本 T 应被分在第 j 类中。

模糊分类法关键在于对隶属度的确定和语义关联度 SR 的计算。

2．向量空间模型[2]

向量空间模型是效果较好的一种文本检索模型。用于文本分类时，向量空间模型可以看作是模糊分类法的特例。特征关键词 l_1，l_2，…，l_m要求相互独立，并被看成一个 n 维坐标系中的坐标轴；模糊分类法中的隶属度在此称为权值，各词对应的权值则理解为相应的坐标值。这样由（l_1，l_2，…，l_m）分解而得的正交词条矢量组就构成了一个向量空间，文本与文本类则映射成为空间中的点。文本与文本类的匹配问题转化为向量空间中的矢量匹配问题，两者的相似程度可用向量之间的夹角来度量，夹角越小，说明相似度越高。待分类文本 T 与文本类 C_k 的相似度计算公式如下：

$$Sim(T,C_k)=\frac{\sum_{j=1}^{n}u_{T,j}u_{k,j}}{\sqrt{\sum_{j=1}^{n}(u_{T,j})^2\sum_{j=1}^{n}(u_{k,j})^2}}$$

3．距离分类方法[8]

类似于向量空间模型，距离分类方法中的文本映射成为 n 维空间中的点，m 个类别看成是 n 维空间的 m 个聚类，每个聚类有一个聚类中心，它也是 n 维空间中的点。利用多元统计分析中的距离判别方法判定某个待分类文本 T 与 k 个聚类中心的距离，距离为近者，则 T 属于该类别。距离计算可采用欧氏距离或马氏距离。在概率意义下，马氏距离能保证获得最小的误判概率；实验表明，采用马氏距离的分类方法的精度接近向量空间模型，但计算量巨大；欧氏距离的分类方法计算简单，但精度明显下降。

除上述方法之外，还有利用概率计算的 Nvaive Bvayes 分类法和利用文本向量到文本类向量的映射关系的矩阵变换分类法等。

3．分类特征词的统计计算

从上一段可知，目前应用的分类方法一般都采用特征关键词描述文本的内容特征，比较待分类文本与训练语料在特征关键词上的相似或差异程度进行文本分类，多数可以归结为特征关键词词频的比较。本文认为：特征关键词最好具有如下特征：

特征关键词的定义最好简单明确，具有可操作性。

特征关键词的选取应由程序自动确定，避免人工干预。

特征关键词应能反映文本的类别特征。

根据第一项，我们的实验系统没有严格按照国标 GB13715《信息处理用现代汉语分词规范》[9]的规定分词，而是直接采用词表分词加歧义校正的方法，分词算法简单快速，词频统计亦很方便。事实上对文本分类而言，待分类文本的分词算法可以更简单，即直接对“特征关键词”分词。

文献[5]介绍了采用预过滤系词、代词等，结合专家选词意见和词频统计等方法的抽取分类特征关键词的过程。本文认为若具有第二项特征（即特征关键词由机器自动选择确定）就能使系统通过分类过程不断学习，自我完善分类特征关键词，并在加入新的分类类别时自动选择分类特征关键词，但如何选择分类特征关键词使其具有第二项特征是问题所在。

设文本分为 m 个类别，$x_1(l)$，$x_2(l)$，… $x_m(l)$表示词 l 在这 m 个类别中的词频，定义：

$$\bar{x}(l)=\frac{1}{m}\sum_{i=1}^{m}x_i(l)$$ 表示词频的平均数

$$s(l)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}x_i^2(l)-\bar{x}^2(l)}$$ 表示词频的类间标准差

$$Max\_x(l)=\underset{i=1}{\overset{m}{MAX}}x_i(l)$$ 表示数值最大的那个词频值

$$v(l)=\frac{s(l)}{\bar{x}(l)}$$ 表示词频的变异系数

以 Max_x(l)和 v(l)作为选择分类特征关键词的两项指标。前者主要是为了保证词频高于一定值的词才能选作特征关键词，虽然在一定程度上影响封闭性测试的结果，但却可以确保开放性测试结果的稳定性，而不采用词频的平均值是考虑到有个别词一般仅出现在少数类别中，而它们对分类有较重要的贡献，应将其选作特征关键词。指标 v(l)（词频变异系数）实质上是单位词频的词频标准差，它与词频的类间标准差相比，更能反映词在文本上的类别特征。

在实验中，用于小类测试的十个类别均属于“计算机技术”类，当选择最高词频大于 1.38‰，变异系数大于 0.8 时，自动提取 789 个特征关键词，基本反映这十个小类的类别特征。下面是按汉语拼音顺序排列的前 3 行自动提取的特征关键词：

安全　安装　按钮　板子　办公　帮助　包含　保护　报表　报告　背景
奔腾　本地　彼此　弊病　边框　编辑　编码　编写　编译　变换　标点
标签　标题　标准　表达式　表格　并行　病毒　波特　波形　播放　补救

4. 基于统计特征词模糊文本自动分类模型

前面提到，模糊分类法关键在于对隶属度的确定和语义关联度 SR 的计算，下面讨论本文提出的方法。

对于由 m 个类别的文本构成的训练语料，设 $L=\{l_1, l_2, \cdots, l_n\}$为选中的 n 个特征关键词，在论域 L 上定义文本类别 k 的一个模糊集：

$Fk=\{u_{k,1}/l_1, u_{k,2}/l_2, \cdots, u_{k,n}/l_n\}$

其中 $u_{k,1}$，$u_{k,2}$，$\cdots u_{k,n}$为在文本类别 C_k 上对应特征关键词的隶属度。计算如下：

$$u'_{k,j}=\frac{x_k(l_j)}{\bar{x}(l_j)}$$ 表示文本类别k上的关键词l_j的隶属度。

事实上，选用

$$u'_{k,j}=\frac{x_k(l_j)}{Max\_x(l_j)}$$ 或

$$u'_{k,j}=\frac{x_k(l_j)}{Other\_x_k(l_j)}$$ 当$Other_x_k(l_j)$为0时，用一较小数代替。

作为文本类别 k 上的特征关键词 lj 的隶属度，效果也相当好。这里：

$$Max\_x(l_j)=\overset{m}{MAX}\,x_i(l_j)$$

$$Other\_x_k(l_j)=\sum_{\substack{i=1\\k\neq i}}^{m}x_i(l_j)$$

最后，对上述求得的隶属度进行规格化处理，即：

$$u_{k,j}=\frac{u'_{k,j}}{\sum_{i=1}^{n}u'_{k,i}*x_k(l_i)}$$

设待判文本 T 的模糊集记为

$F_T = \{ u_{T,1}/l_1 , u_{T,2}/l_2 , \cdots , u_{T,n}/l_n \}$

其中 $u_{t,j}$ （j=1, 2, …n,）取待分类文本 T 中对应的特征关键词的词频。则待分类文本 T 与类别 C_k 的语义关联度：

$$SR(F_T,F_k)=\sum_{i=1}^{n}u_i*u_{k,i}$$

关联度越大，则说明语义关系越密切，计算文本 T 的模糊集与全部 m 个文本类上的模糊集的语义关联度即知文本 T 应被分在何类中。

显然，若待判文本中的特征关键词的词频与某类的训练文本中的对应词词频一致，则计算得到的语义关联度为 1。

据文献[6][7]介绍和我们的实验，向量空间模型是使用效果较好的分类模型，它通过计算多维空间的向量间夹角来度量待分类文本与各类的相似度。而距离分类法则是以待分类文本与各聚类中心的距离度量它们的相似度。模糊分类方法则通过计算语义关联度来度量相似性，隶属度的确定是模糊分类方法的关键，本文提出的利用特征关键词词频的差异计算隶属度的方法，实验表明有很好的分类精度。对于选中的某一特征关键词，在距离判别法中，待判文本中该词的词频高于或低于训练语料中的相同词的词频，都会增加待判文本与该类的距离。而在本文设计的模糊分类法中，待判文本中特征关键词的词频越高，则归属于该类的可能性越大。

5. 实验结果

我们的实验分封闭性测试和开放性测试，封闭性测试是指训练语料库中的文本作为被测试文本，对其进行分类测试。开放性测试是指被测试文本不包含在训练语料库中，对其

进行分类测试。与封闭性测试相比，开放性测试的结果更具有实际意义。当训练语料库的规模达到相当大的程度，封闭性测试结果与开放性测试结果应趋于吻合。

在进行分类测试时，我们使用语料均来自于公开发表的带分类号的科技期刊论文。文本分类规范按照1999年版的《中国图书馆分类法》[10]，这是目前国内图书分类的专业规范。对分类号精确到字母后一位数字的类别（简称大类），无论是在封闭性测试还是在开放性测试中，我们选择的三个大类分类的精度已达到95%以上。故我们主要对文本在小类上的自动分类正确率进行测试，即分类号精确到字母后数字的第二位的类别。

小类自动分类实验中共计有325篇语料，分布在同一大类的10个小类中，其中有11篇语料兼两小类。测试中凡兼类文献只要被分入任一兼类，即认为分类正确。在封闭性测试中，所有语料同时作为训练语料（样本）和封闭性测试语料。在开放性测试中，按语料小类，分别抽取20%作为开放性测试语料，剩余的80%作为系统的训练语料（样本）。

在封闭性测试中，基于统计特征词模糊文本自动分类的平均精度为81%，相对照的向量空间模型的平均精度为60%，距离分类方法的平均精度为58%。在开放性测试中，基于统计特征词模糊文本自动分类的平均精度为72%，相对照的向量空间模型的平均精度为53%，距离分类方法的平均精度为52%。

从封闭性测试和开放性测试结果来看，二者的正确率差距不大；在训练语料库不变的条件下，它们比以往的实验结果更为接近，本文认为这与特征关键词的选取方法有关。

由于本文分类正确性的统计是由待分类文本T的模糊集与全部的文本类别上的模糊集的最大语义关联度决定，若考虑次最大的语义关联度，分类正确度可提高10个百分点以上，因此系统基本上满足实用要求。

本文的研究受军事电子预研基金项目“DJ8.4.1 基于分布式人工智能技术的军事电子文本检索方法的研究”的资助。

参考文献

[1] GERARD.S, Another Look at Automatic Text-Retrieval Systems. Communications of the ACM. 29,7, 1986. p648-656.

[2] GERARD.S, Developments in Automatic Text Retrieval. Science . 253 ,30, 1991. p974-1012

[3] ELLEN.R AND WENDY.L, Information Extraction as a Basis for High-Precision Text Classification. ACM Transactions on Information Systems. 12,3, 1994. p296-333

[4] 刘开瑛等，中文文本中抽取特征信息的区域与技术。中文信息学报，12，2，1998. p1-7

[5] 何新贵等，中文文本的关键词的自动抽取和模糊分类方法。中文信息学报，13,1,1999. p9-15

[6] 邹涛， Web信息的采集、文档的识别与分类。计算机世界报，1999.4.19

[7] 邹涛，检索模型，计算机世界报，1999.4.19

[8] 王小华等， 基于多元统计分析的电子文本自动分类，杭州电子工业学报，19,3,1999

[9] 刘源等，信息处理用现代汉语分词规范，清华大学出版社，1994

[10] 中国图书馆分类法编辑委员会，中国图书馆分类法（第四版），中国图书馆出版社，1999

[11] 吴军等，汉语语料的自动分类. 中文信息学报. 9,4, 1996. p25-32.