网页模糊归类算法的应用与实现

曹乐乐　　东南大学软件学院
韩正忠　　东南大学应用数学系
（南京 东南大学, 210096　E-mail: cjnxcll2001@163.com）

摘要：本文运用以模糊综合评判为核心的理论实现对网页的模糊自动归类，详细阐述了网页模糊归类算法(FWCA)，并且通过一个实例阐明了实现过程。作者利用此算法亲自设计实现了一个“网页模糊归类测试系统”，通过分析大量实验数据证明了此算法的归类效果非常稳定和准确。
关键词：FWCA　模糊综合评判　网页归类　分类浏览　搜索引擎

Application and Implementation of Fuzzy Web Categorization Algorithm

Han Zhengzhong　Department of Applied Mathematics in Southeast University
Cao Lele　College of Software Engineering in Southeast University
(Southeast University Nanjing, 210096　E-mail: cjnxcll2001@163.com)

Abstract: This paper implements mainly the Fuzzy Comprehensive Evaluation in the Fuzzy Web Categorization Algorithm (FWCA), and introduces this algorithm in detail. In addition to the presentation of a practical example, the author also realized a “Fuzzy Web Categorization Test System” with FWCA. It proves that the result of implementing FWCA is quite satisfactory through the analysis of data based on plenty of experiments on that system.
Keywords: FWCA　Fuzzy Comprehensive Evaluation　Web Categorization　Categorized Browse　Search Engine

自有文字和书籍以来，人类就开始注意文章的分门别类和编撰目录。那些目录事实上就将文章按照内容的类别进行了分类。九十年代以来，Internet 以惊人的速度发展起来，Web 的容量增长迅速，平均每天增加 100 万个页面。计算技术发展到今天，靠人来阅读互联网上信息和对网上信息做分门别类和总结已经不可能。

搜索引擎的分类浏览模式由此应运而生。它的目录分类的质量较高，检索效果好；但是需要人工维护，因此存在成本高、信息更新慢、维护的工作量大的缺点。而基于模糊技术的网页自动归类能依据网页中所包含的文本的语义将大量的网页自动分门别类，从而更好地帮助人们把握网络信息。

网页模糊归类的步骤与算法

简单地说，网页自动归类所要完成的任务就是在给定的分类体系下，根据网页的内容自动地确定网页关联的类别。如果从纯数学角度来看，网页分类的过程实际上就是一个多对多的映射过程。依据“贝叶斯假设”的内容，可以假定组成网页的元素在确定网页类别的作用上相互独立。这样，可以使用网页中出现的字或词的集合来代替网页，即用一个向量来表示文本：D（W1,W2,W3……Wn），其中 Wi 为第 i 个元素（以下均称为“特征项”）的数值。当然，这将丢失大量关于网页内容的信息，但是这种假设可以使网页的表示和处理形式化，从而让计算机可以处理网页。

构成网页中的文本的词汇，数量是相当大的，因此，表示网页的向量空间的维数也相当大，可以达到几万维，所有几万个词汇对网页分类的意义是不同的。首先，需要考虑词语的性质。一些通用的、各个类别都普遍存在的词汇对分类的贡献是很小的，因此特征提取过程需要去掉对表达网页类别不太重要的词汇。例如“的”、“地”、“得”、“着”、“了”等等。其次，在某特定类中出现比重大而在其他类中出现比重小的词汇对文本分类的贡献大，为了提高分类精度，可以利用词语的互信息量筛选出针对该类的特征项集合。具体操作方法是算出每个词语的互信息量并排序，然后抽取前 n 个词语作为该类别的特征项，抽取的原则是反复试验使得网页归类效果最优。互信息量计算公式由下式给出：

$$\text{词语互信息量}=\log\left(\frac{P(W\mid C_j)}{P(W)}\right)\qquad P(W\mid C_j)=\frac{1+\sum_{i=1}^{|D|}N(W,d_i)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|}N(W_s,d_i)}$$

$P(W\mid C_j)$ 为 W 在 C_j 中出现的比重
$|D|$ （分子）为该类的训练网页样本数
$N(W,d_i)$ 为词 W 在 d_i 中的词频
$|V|$ 为总词数
$\sum_{s=1}^{|V|}\sum_{i=1}^{|D|}N(W_s,d_i)$ 为该类所有词的词频和
$P(W)$ 与分子上的计算公式相同，只是计算词在所有训练网页样本中的比重
$|D|$ （分母）为全体训练网页样本数。

为了让计算机为我们进行网页的自动归类，必须先对计算机进行训练。只要训练网页足够多，那么由计算机进行的归类活动也将是准确的。所有的训练样本都需表示为向量 $X=(x_1,x_2,\Lambda\ ,x_n)$。并使用每个词的相对词频（TF-IDF 公式）对网页样本的特征项进行量化。然后，将每个类别中的所有训练样本数据合成为一个平均参照样本，计算方法就是将每个特征项的值求算术平均。相对词频计算公式由下式给出：

$$W(t,\vec{d})=\frac{tf(t,\vec{d})\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in\vec{d}}\left[tf(t,\vec{d})\times\log(N/n_t+0.01)\right]^2}}$$

$W(t,\vec{d})$ 为词 t 在网页样本 $\vec{d}$ 中的相对词频
$tf(t,\vec{d})$ 为词 t 在网页样本 $\vec{d}$ 中的词频
N 为训练网页样本的总数
n_t 为训练网页样本集中出现 t 的网页数目

在归类过程中，采用三级模糊综合评判。一级指标因素集（网页中出现位置）包括：*网页题名*、*文章标题*、*第一段首句*、*第一段尾句*、*第二段首句*、*第二段尾句*、*第三段首句*、*第三段尾句*、*首段*、*尾段*、*HTML 标记*。二级指标因素集（词性）包括：*名词, 动词, 形容词, 副词, 介词, 连词, 助词, 数字, 符号*。三级指标因素集：待分类网页中所包含的全部词语的频数。评价集确定为 V={V1(不属于 0), V2(不太可能属于 0.25), V3(可能属于 0.5),

V4(很可能属于 0.75), V5(属于 1)}。

专家随机抽取了 300 篇网页，对这些网页进行人工自由标引、人工打分、词频统计，并进行统计数据的分析、研究，将一级指标因素权重集确定为 A={0.128, 0.128, 0.128, 0.104, 0.104, 0.104, 0.06, 0.06, 0.06, 0.06, 0.05, 0.05}；根据语言学专家对各类别中不同词性的词语对标志一个类别（以中图分类法为标准）重要性程度统计和评分，将二级指标因素权重集确定为 An={0.28, 0.18, 0.24, 0.06, 0.05, 0.04, 0.04, 0.06, 0.05}；根据词语的互信息量确定出三级指标因素权重为 $A_{nm}=\{A_{nm1}, A_{nm2} \cdots A_{nmx}\}$　其中，A_{nmx} 即为对应词语的互信息量

隶属函数采用卡夫曼教授提出的隶属函数确定方法（正态分布模型）确定如下：

① 词频针对“不属于”的隶属函数

$$U_{V1}(x)=1-e^{-0.196(x-a_{xyz})^2}$$

② 词频针对“不太可能属于”的隶属函数

$$U_{V2}(x)=\begin{cases} e^{-0.689(x-a_{xyz}+4.268)^2} & x \le a_{xyz} \\ e^{-0.689(x-a_{xyz}-4.268)^2} & x > a_{xyz} \end{cases}$$

③ 词频针对“可能属于”的隶属函数

$$U_{V3}(x)=\begin{cases} e^{-0.892(x-a_{xyz}+2.162)^2} & x \le a_{xyz} \\ e^{-0.892(x-a_{xyz}-2.162)^2} & x > a_{xyz} \end{cases}$$

④ 词频针对“很可能属于”的隶属函数

$$U_{V4}(x)=\begin{cases} e^{-1.236(x-a_{xyz}+0.845)^2} & x \le a_{xyz} \\ e^{-1.236(x-a_{xyz}-0.845)^2} & x > a_{xyz} \end{cases}$$

④ 词频针对“属于”的隶属函数

$$U_{V5}(x)=e^{-0.196(x-a_{xyz})^2}$$

其中，a_{xyz} 是训练样本中词语的相对词频；x 为样本网页中对应词的统计词频；式中的系数是通过人工评判得到一些特殊点，由待定系数法求出的。

下面只需根据多级模糊综合评判的计算方法与步骤将待归类网页与所有类别的平均参照样本进行一遍计算，得出一组表示该网页与各个类别贴近度的数值（$V_1,V_2\cdots,V_n$）。然后按照“最大隶属原则”，将网页划到 V_n 值最大的对应的类别中；或者用“域值法”，事先确定一个不大于 1 的域值 λ，若 $V_n \ge \lambda$ 则认为网页属于此类别，因此，一个网页可能同时属于多个类别。

网页模糊归类实例

（1）. 前期工作

- 简化的分类的标准：经济类，体育类，科教类
- 训练样本数目：48 篇（三类各 16 篇）
- 待归类网页：

```
<html><head><title>经济快讯</title></head><body>
<p>经济：我国经济水平连续三年实现翻番，人民生活水平日益提高</p>
</body></html>
```

- 一级指标因素及权重：U={U_1=0.50, U_2=0.50}
- 二级指标因素及权重：U_1={U_{11}=1.00}
  U_2={U_{21}=0.40, U_{22}=0.26, U_{23}=0.34}
- 三级指标因素及权重：
  U_{11}={U_{111}=0.86, U_{112}=0.14)}
  U_{21}={U_{211}=0.11, U_{212}=0.35, U_{213}=0.21, U_{214}=0.06, U_{215}=0.10, U_{216}=0.17}
  U_{22}={U_{221}=0.26, U_{222}=0.38, U_{223}=0.36}
  U_{23}={U_{231}=0.46, U_{232}=0.54}
- 经济类训练网页样本相对词频（其余两类略）：
  a_{11}={a_{111}(经济 1.1923), a_{112}(快讯 1.1912)}
  a_{21}={a_{211}(我国 1.1017), a_{212}(经济 2.2100), a_{213}(水平 1.7930), a_{214}(三年 0.5029), a_{215}(人民 0.8996), a_{216}(生活 1.1891)}
  a_{22}={a_{221}(实现 1.1988), a_{222}(翻番 1.8002), a_{223}(提高 1.6966)}
  a_{23}={a_{231}(连续 1.5891), a_{232}(日益 1.7293)}

（2）. 模糊综合评判

首先统计待分类网页的各个词语的绝对词频如下：

a'_{11}={ a'_{111}(经济 1), a'_{112}(快讯 1)}
a'_{21}={ a'_{211}(我国 1), a'_{212}(经济 2), a'_{213}(水平 1), a'_{214}(三年 1), a'_{215}(人民 1), a'_{216}(生活 1)}
a'_{22}={ a'_{221}(实现 1), a'_{222}(翻番 1), a'_{223}(提高 1)}
a'_{23}={ a'_{231}(连续 1), a'_{232}(日益 1)}

总共可以得到 4 个一级模糊综合评判矩阵如下：

$$R_{11}=\begin{bmatrix} 0.0072 & 0.0000 & 0.0314 & 0.5906 & 0.9928 \\ 0.0071 & 0.0000 & 0.0314 & 0.5896 & 0.9929 \end{bmatrix}$$

$$R_{21}=\begin{bmatrix} 0.0020 & 0.0001 & 0.0227 & 0.5052 & 0.9980 \\ 0.0086 & 0.0000 & 0.0334 & 0.6075 & 0.9914 \\ 0.1160 & 0.0002 & 0.1879 & 0.9967 & 0.8840 \\ 0.0473 & 0.0001 & 0.0844 & 0.8611 & 0.9527 \\ 0.0020 & 0.0001 & 0.0226 & 0.5040 & 0.9980 \\ 0.0070 & 0.0000 & 0.0311 & 0.5876 & 0.9930 \end{bmatrix}$$

$$R_{22}=\begin{bmatrix} 0.0077 & 0.0000 & 0.0321 & 0.5968 & 0.9923 \\ 0.1179 & 0.0003 & 0.1912 & 0.9975 & 0.8821 \\ 0.0907 & 0.0002 & 0.1473 & 0.9731 & 0.9093 \end{bmatrix}$$

$$R_{23}=\begin{bmatrix} 0.0658 & 0.0001 & 0.1100 & 0.9222 & 0.9342 \\ 0.0990 & 0.0002 & 0.1603 & 0.9836 & 0.9010 \end{bmatrix}$$

构造二级模糊综合评判矩阵：

①采用 M(∧,∨)算子的运算结果

$$R_1=[U_{11}\bullet R_{11}]=[0.0072 \quad 0.0000 \quad 0.0314 \quad 0.5906 \quad 0.8600]$$

$$R_2=\begin{bmatrix} U_{21}\bullet R_{21} \\ U_{22}\bullet R_{22} \\ U_{23}\bullet R_{23} \end{bmatrix}=\begin{bmatrix} 0.1160 & 0.0002 & 0.1879 & 0.3500 & 0.3500 \\ 0.1179 & 0.0003 & 0.1912 & 0.3800 & 0.3800 \\ 0.0990 & 0.0002 & 0.1603 & 0.5400 & 0.5400 \end{bmatrix}$$

②采用 M(·,⊕)算子的运算结果

$$R_1=[U_{11}\bullet R_{11}]=[0.0072 \quad 0.0000 \quad 0.0314 \quad 0.5905 \quad 0.9928]$$

$$R_2=\begin{bmatrix} U_{21}\bullet R_{21} \\ U_{22}\bullet R_{22} \\ U_{23}\bullet R_{23} \end{bmatrix}=\begin{bmatrix} 0.0318 & 0.0001 & 0.0663 & 0.6795 & 0.9682 \\ 0.0795 & 0.0002 & 0.1340 & 0.8845 & 0.9205 \\ 0.0837 & 0.0002 & 0.1372 & 0.9554 & 0.9163 \end{bmatrix}$$

构造三级模糊综合评判矩阵：

①采用 M(∧,∨)算子的运算结果

$$R=\begin{bmatrix} U_1\bullet R_1 \\ U_2\bullet R_2 \end{bmatrix}=\begin{bmatrix} 0.0072 & 0.0000 & 0.0314 & 0.5906 & 0.8600 \\ 0.1179 & 0.0003 & 0.1912 & 0.3500 & 0.3500 \end{bmatrix}$$

②采用 M(·,⊕)算子的运算结果

$$R=\begin{bmatrix} U_1\bullet R_1 \\ U_2\bullet R_2 \end{bmatrix}=\begin{bmatrix} 0.0072 & 0.0000 & 0.0314 & 0.5905 & 0.9928 \\ 0.0618 & 0.0002 & 0.1080 & 0.8266 & 0.9382 \end{bmatrix}$$

多因素综合评判：

①采用 M(∧,∨)算子的运算结果

$$\underset{\sim}{B}=\underset{\sim}{U}\bullet\underset{\sim}{R}=[0.5 \quad 0.5]\bullet\begin{bmatrix} 0.0072 & 0.0000 & 0.0314 & 0.5906 & 0.8600 \\ 0.1179 & 0.0003 & 0.1912 & 0.3500 & 0.3500 \end{bmatrix}$$

$$\underset{\sim}{B}=[0.1179 \quad 0.0003 \quad 0.1912 \quad 0.5000 \quad 0.5000]$$

$$V=\frac{0.1179\times 0+0.0003\times 0.25+0.1912\times 0.50+0.5000\times 0.75+0.5000\times 1}{0.1179+0.0003+0.1912+0.5000+0.5000}=0.74$$

②采用 M(•,⊕)算子的运算结果

$$\underset{\sim}{B}=\underset{\sim}{U}\bullet\underset{\sim}{R}=[0.5\quad 0.5]\bullet\begin{bmatrix}0.0072 & 0.0000 & 0.0314 & 0.5905 & 0.9928\\ 0.0618 & 0.0002 & 0.1080 & 0.8266 & 0.9382\end{bmatrix}$$

$$\underset{\sim}{B}=[0.0345\quad 0.0001\quad 0.0697\quad 0.7086\quad 0.9655]$$

$$V=\frac{0.0345\times 0+0.0001\times 0.25+0.0697\times 0.50+0.7086\times 0.75+0.9655\times 1}{0.0345+0.0001+0.0697+0.7086+0.9655}=0.86$$

网页归类决策：

通过三轮计算得出下表：

样本与类别贴近度	经济类	体育类	科教类
采用 M(∧,∨)算子	0．74	0．16	0．27
采用 M(•,⊕)算子	0．86	0．01	0．14

由上表对照评价集可知，如果用“M(∧,∨)算子”计算，此网页很可能属于经济类，不太可能属于体育类和科教类；如果用“M(•,⊕)算子”计算，则可以认为属于经济类，不太可能属于科教类，不属于体育类。

不管采用哪一种算子，如果用“最大隶属原则”判断，显然都应该属于“经济类”；如果用“域值法”（λ≤0.74）判断，也应该都属于“经济类”。

结果分析

由上述算例可以看出，若用“最大隶属原则”判断，取 λ >0.74，采用 M(∧,∨)算子的算法就无法对此网页归类了，而采用 M(•,⊕)算子却可以对网页正确归类。另外，采用 M(•,⊕)算子的结果区分效果比较明显，与人工归类的结果比较接近。由此可见，采用 M(•,⊕)算子的算法明显优于采用 M(∧,∨)算子的算法。

本文的实例网页最后得出的与“经济类”网页的贴近值仅 0.86，比理想值（人工估计为 0.9）偏低了了一些，与其他类别的贴近值也存在一些偏差。这是因为本文中举的例子为了简单起见，训练文本才 48 篇，导致计算机训练不足；另外，待归类网页过于简单。这些都导致了归类结果与理想值的偏差,在实际情况下，这些问题都可以避免。

作者在自行开发的“网页模糊归类测试系统”平台上作了大量对于网页的归类测试工作 (详见附录) ，测试文档与训练网页都是取自“中国新闻网”新闻网页。在训练网页达到 1200 篇的时候，归类准确率封闭测试为 85.73%，开放测试为 78.82%。虽然这种以模糊综合评判为核心的算法实现的系统初始化工作比较繁琐，但是归类的结果准确率很高，因此还是非常具有实际应用价值的。

参考文献

[1] 卜东波. 聚类/分类理论研究及其在大规模文本挖掘中的应用, 北京:中国科学院计算技术研究所, 2000.

[2] 边肇祺, 张学工. 模式识别(第二版), 北京:清华大学出版社, 2000, 83-159, 284-300.

[3] 韩正忠, 方宁生. 模糊数学应用, 南京:东南大学出版社 2003.2

[4] 刘智颖. 自然语言理解与机器翻译, 清华大学出版社 2001.7

[5] 刘祖根. 基于 WordNet 的文本分类技术研究和实现, 长江大学 2002

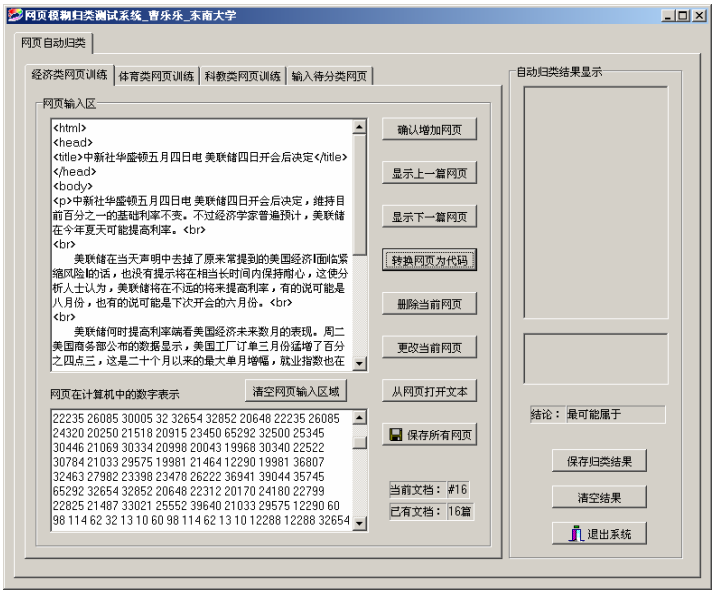
[6] 庞剑锋, 卜东波, 白硕. 基于向量空间模型的文本自动分类系统的研究与实现, 计算机应用研究, 2001, 9(9): 23-26.

[7] 刘增良. 模糊技术与应用选编, 北京航空航天大学出版社, 1997.2(1) ISBN 7-81012-691-1

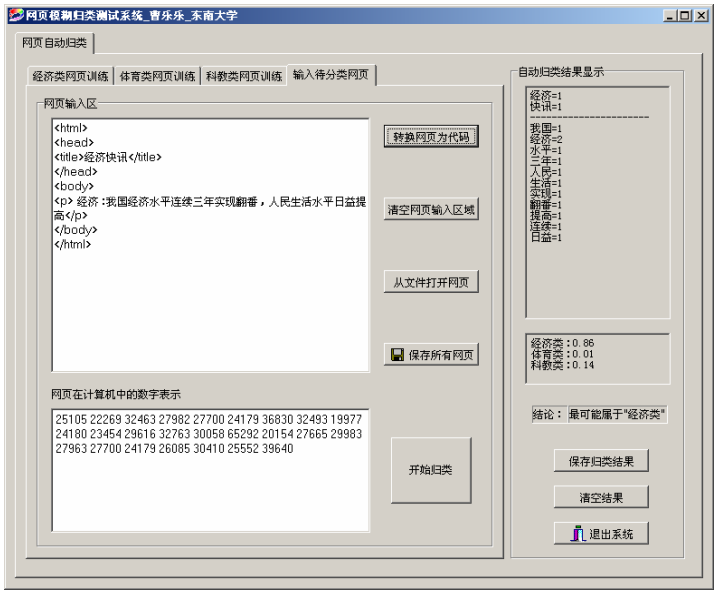
[8] 孙贻源. 模糊数学, 华中工学院出版社, 1984

[9] 张俊福. 应用模糊数学, 地质出版社, 1988.11

附录（“网页模糊归类测试系统”使用案例）



（网页训练）



（网页归类）