网页去重的改进算法

王 静[1]，刘观宁[2]，张钰辉[1]
(1. 西安电子科技大学 计算机学院，陕西 西安 710071；
2. 安徽省技术创新服务中心，安徽 合肥 230001)

摘 要：针对网页内容相似重复的特点，提出了一种改进算法对网页进行去重处理。该方法能够有效地对网页进行去重，并能对网页信息进行冗余识别处理。实验结果表明，与原有网页去重算法相比，该算法的执行效果提高了 14.3%，对网页去重有了很明显的改善。

关键词：网页去重；特征提取；特征表示
中图分类号：TP391 **文献标识码**：A **文章编号**：1674-7720(2011)12-0016-03

Theory of blind signal separation and Markov random fields

Wang Jing[1], Liu Guanning[2], Zhang Juehui[1]
(1.School of Computer Science and Engineering, Xidian University, Xi'an 710071, China)
2. Anhui Province Technology Innovation Service Center, Hefei 230001, China)

Abstract: In this paper we introduce the theory of blind signal separation and two methods: in dependent component analysis(I-CA) and Markov random field (MRF).In the end ,we give the developing direction of blind signal separation.

Key words: blind signal separation (BSS); independent component analysis (ICA); Markov random field (MRF)

随着互联网的高速发展，Web 已经成为最大的信息来源。但是如何获取这些 Web 信息为我所用则是大家面临的共同问题。网页去重是 Web 网页信息处理的重要环节，只有在对网页的去重基础上才可以准确处理网页中的信息。本文介绍网页的去重算法。

提取出来的网页，有些内容可能很相似，对于这些内容相似的网页没必要保存。针对系统中的人才招聘网页更是必要：一个公司的招聘信息很可能会在数十家招聘网站以及自己公司主页同时发布，所以有必要对这些网页去重。

1 网页的特征表示

词、词组和短语是组成文档的基本元素，在不同内容的文档中各词条出现频率有一定的规律性，不同的特征词条可以区分不同内容的文本。因此可以抽取一些特征词条构成特征矢量，在 VSM[1]模型中把 $t_1,t_2,\cdots,t_n$ 看成一个 N 维的坐标系 $w_1(d),w_2(d),\cdots,w_n(d)$ 为相应的坐标值，因而文本 d 被看成是 N 维空间中的一个规范化特征矢量：$\boldsymbol{V}(d)=(t_1,w_1(d);\cdots;t_i,w_i(d);\cdots;t_n,w_n(d);)$

对于网页，t_i 就表示特征词条，$w_i(d)$ 就是文本 d 中 t_i 的权值。用这个特征矢量来表示网页文本。在网页表示中，对任一特征而言有两个因素影响特征的权值。一是词在 HTML 文档中出现的词频，另一个是该词在该文档中出现的位置。词频指的是某一词条在文档中出现的频率，频率越高（当然不包括那些停用词）则说明该词越重要，越能代表该网页的内容。对于网页的主题包含在<title>和</title>之间的词组比在<body>和</body>之间的词组更具有代表性。因此本文提出了一种把该词出现的频率以及该词出现的位置相结合的权重计算方法，能够更有效地表示网页。公式如下：

$$w_i(d)=\left[\frac{1}{2}(CW_j(t,\bar{d}))+\frac{1}{2}(SW(t_j,d_i))\right] \tag{1}$$

式(1)中的 $CW_j(t,\bar{d})$ 是针对词在 HTML 文档中出现的词频的权重计算方法，公式如下：

$$CW(t,\bar{d})=\frac{tf(t,\bar{d})\times\log(N/n_t+0.01)}{\sqrt{\sum_{N\in\bar{d}}[tf(t,\bar{d})\times\log(N/n_t+0.01)]^2}} \tag{2}$$

其中 $CW_j(t,\bar{d})$ 一般被定义为 t_i 在 d 中的出现频率 $tf_i(d)$ 的函数，$tf(t,\bar{d})$ 为词 t 在文本 d 中的词频。N 是整个文本的数量，n_i 是文本中出现第 i 个关键词的文本数量。式(1)中的 $SW(t_j,d_i)$ 则是针对该词在该 HTML 文档中出现的位置来计算权重。公式如下：

$$SW(t_j,d_i)=\sum_{e_k}(w(e_k)\cdot TF(t_j,e_k,d_i)) \tag{3}$$

这里 e_k 是一种 HTML 元素，它指的是 TITLE 标签。$w(e_k)$ 表示分给 e_k 的权重。$TF(t_j,e_k,d_i)$ 表示了词组 t_j 出现在 HTML 网页 d_i 的元素 e_k 中的次数。定义了 $w(e)$ 如下：

$$w(e)=\begin{cases}\alpha, & \text{if } e \text{ is META or TITLE}\\ 1, & \text{其他}\end{cases} \tag{4}$$

这里 $\alpha=2$，$\alpha=3$，$\alpha=4$ 和 $\alpha=6$ 都是经过实验得到的。实验结果也证明了此改进算法对网页分类正确率的有效性。

2 网页的特征提取

使用 VSM 模型表示法时，表示文档的特征向量的维数会达到成百上千。同时，具有代表性的特征以及词汇特征也会很大，并且是冗余的。这种未经处理的文本矢量会给后继的处理工作带来巨大的计算开销。特征提取主要用于排除那些被认为无关或关联性不大的特征。基于 VSM 常用的特征项提取算法有：词频、信息增益、互信息量[2]及 X^2 统计量[3]等。在中文文本分类中使用较多的是互信息量和 X^2 统计量。

(1) 互信息量

互信息是信息论中的概念，它用于度量一个消息中两个信号之间的相互依赖程度。在特征选择领域中人们经常利用它来计算特征 t 与类别 c 之间的依赖程度，将特征 t 与各个类的互信息融合起来作为特征的权重。特征 t 与第 i 类的互信息计算公式如下(两个公式等价)：

$$I(t,c)=\log\frac{P_r(t,c)}{P_r(t|c)\times P_r(t)} \tag{5}$$

$$I(t,c)=\log P_r(t|c)-\log P_r(t) \tag{6}$$

(2) X^2 统计量

基于 X^2 统计量的特征提取方法又被称作开方拟合检验(CHI，X^2-test)，其计算公式如下：

$$x^2(t_k,c_i)=\frac{g[P(t_k,c_i)P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)} \tag{7}$$

其中：t_k 表示任意特征项(特征词)；c_i 表示任意类别；g 为训练集中所有文本数；$P(t_k,c_i)$ 为 t_k 和 c_i 同时出现的概率(即对于任意一篇文章 X，含有特征项 t_k 且文章 X 属于类别 c_i 的概率)；$P(t_k)$ 为文章中出现特征项 t_k 的概率；$P(c_i)$ 为文章属于类别 c_i 的概率，类似地不难理解 $P(t_k,\bar{c}_i)$、$P(\bar{t}_k,c_i)$、$P(\bar{t}_k,\bar{c}_i)$、$P(\bar{t}_k)$ 和 $P(\bar{c}_i)$。

(3) 联合特征提取方法

虽然 X^2 统计量法是目前常用的特征提取方法之一，但该方法仍存在一些缺点，如它提高了在指定类中出现少而在其他类中出现较高的特征的权重以及降低了低频词的权重等。根据公式(3)~(5)，对于指定类中出现频率低而其他类中出现频率高的词语，当 $P(t,c_i)\to 0$，而 $P(t)$ 和 $P(c_i)$ 均不趋向于零，则 $P(t,c_i)/(P(t)P(c_i))\to 0$，于是 $I(t,c)$ 将趋向于负无穷，故这些词语会被过滤掉。根据式(6)，对于有相同 $\log P_r(t|c)$ 的词语来说，低频词的权重将更高，即在多类中普遍出现的高频词的权重将比只在特定类中出现的低频词的权重低。这样就很好地解决了上述问题，所以本文提出一种联合特征提取的方法，该方法综合了 X^2 统计量法和互信息量法，可以获得较好的结果。该方法可以描述为：

$$E(t,c)=\alpha E_1(t,c)+\beta E_2(t,c) \tag{8}$$

$$0<\alpha,\ \beta<1,\ \alpha+\beta=1$$

其中 $E_1(t,c)$ 是使用 X^2 统计量法得到的特征权重；$E_2(t,c)$ 为使用互信息量法得到的特征权重。

3 SOM 神经网络算法

3.1 向量归一化

向量的归一化是对输入向量进行预处理的第一步。其目的是把所有不同长短和方向的向量变成方向不变、长度为 1 的单位向量。设：

$$X=(x_1,x_2,x_3\cdots x_n) \tag{9}$$

X 的模：

$$\|X\|=sqrt(x_1,x_2\cdots x_n)(x_1,x_2\cdots)^{\mathrm{T}}) \tag{10}$$

X 的单位化向量：

$$X=(x_1/\|X\|,x_2/\|X\|,\cdots,x_n/\|X\|) \tag{11}$$

在网络训练过程开始时，定义获胜节点的邻域节点是为了能使二维输出平面上相邻输出节点对相近的输入模式类做出特别反应。假设本次获胜节点为 N_j，它在 t 时刻的邻域节点用 NE_j 表示，$NE_j(t)$ 是包含以 N_j 中心而距离不超过某一半径的所有节点。随着训练过程的进行，$NE_j(t)$ 的半径逐渐减小，最后只包含获胜节点 N_j 本身，也就是说在训练的起始阶段不仅对获胜节点做权值调整，而且也对其较大范围内的几何邻节点做相应的调整，随着训练过程的继续进行，与输出节点相连的权向量也越来越接近其代表的模式类。这时，在对获胜节点的权值进行比较细微的调整时，只对其几何邻节点比较近的节点进行相应的调整，直到最后只对获胜节点本身做细微的调整。在训练过程结束后，几何上相近的输出节点所连接的权向量既有联系又有区别，这样，保证了对某一类输入模式获胜节点能够做出最大“响应”，而相邻节点做出“较大”响应。几何上相邻节点代表特征上相近的模式类别。

自组织特征映射学习过程包括描述最佳匹配神经元的选择和描述权矢量的自适应变化过程两部分。SOM 输出层通常由两维 $m\times m$ 的网格节点组成，从输入向量到网络输出层的每个节点 j 的权值向量定义为 w，w 和 x_i 的维数是相同的，设为 d，影射节点的数量从数十个到

数千个决定 SOM 正确性和概化能力。

3.2 Kohonen 网络训练算法[4~5]

其算法步聚如下：

(1)权连接初始化：初始化输出层节点 j 的权值矢量 w_{ij} 时可选随机值，初始值通常要选择小一点。初始化学习率和领域函数时要尽量大一些，对连接输入神经元和输出神经元之间的权系数设定为小的随机数 a，一般有 $0<a<1$，同时，设定邻近区域的初始半径。

(2)网络输入模式为：

$$X^k=(x_1,x_2\cdots x_n) \tag{12}$$

(3)在 SOM 迭代训练的每一步，从输入数据集中随机地选择文本向量 x_i 属于实数集，计算 x_i 和 som 输出层所有节点 j 的权值向量 w_{ij} 的距离，最匹配的点用 d 表示，权值向量用 w_{ij} 表示，它是输出层节点中最接近 x_i 的。

$$d_j=\sum_{i=1}^{n}(x_i^k-\omega_{ij})^2,i\in\{1,2,\cdots,n\},j\in\{1,2,\cdots,m\} \tag{13}$$

(4)具有最小距离的节点 $j\times N$ 竞争获胜：

$$d_j=\min_{j\in\{1,2,\cdots,m\}}(d_j\} \tag{14}$$

(5) 在每一步学习中 CN 的神经元自适应变化而 CN 外的神经元保持不变，调整输出节点所连接的权值以及几何邻域内节点所连权值为：

$$\Delta\omega_{ij}=\eta(t)(x_i^k-\omega_{ij}),N_j\in NE_j(t),i\in\{1,2,\cdots,n\} \tag{15}$$

式中 $\eta(t)$ 为标量自适应增益，$0<\eta(t)<1$，$\eta(t)$ 是单调降函数，它可以是线性指数的或者是与其成反比的形式等，通常选择 $\eta(t)=0.9(1-t/1\,000)$，它与 $N(t)$ 都是经验函数。

(6) 若还有输入样本数据则 $t=t+1$ 转到步骤(2)。

网络输出与权值调整竞争学习算法规定，获胜神经元输出为 1，其余输出为零。只有获胜神经元才有权调整其权向量 $j\times w$，调整后权向量为：

$$W_{j^*}(t+1)=W_{j^*}(t)+\alpha(X-W_{j^*}(t)) \tag{16}$$

$$W_j(t+1)=W_j(t),\ j\neq j^* \tag{17}$$

其中，$\alpha\in(0,1]$ 为学习率，一般其值随着学习的进展而减小。可以看出，当 $j\neq j^*$ 时，对应神经元的权值得不到调整，其实质是“胜者”对它们进行了抑制，不允许它们兴奋。另外，调整后得到的新向量不再是单位向量，因此需要对调整后的向量重新归一化。步骤(3)完成后回到步骤(1)继续训练，直到学习率 α 衰减到 0。

4 实验结果

采用以上介绍的算法，对一批数量在 50~100 之间的网页集合进行去重处理，集合中包含了一与此内容完全相同或部分相同的网页，将实验结果与人工判别的结果进了比较，发现重复网页的正确率达到 95%以上，出现错误的判断的是由于网页转载时出现错码等现象，有的是两个重复网页的段落排列差异太大。测试结果如图 1 所示。

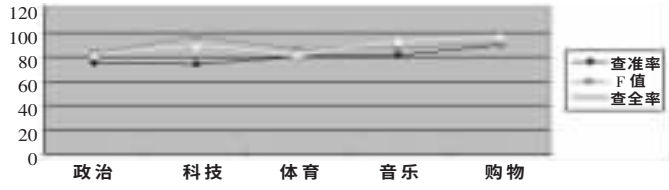

图 1 选择 SOM 聚类结果比较

本文将 SOM 的思想和方法引入中文 Web 文档的聚类问题.探索向用户提供高质量的网页信息具有很强的理论意义和实际价值。但是，这种方法的不足之处是当网络的连接过多、节点数目庞大时其计算量大，需要较长的学习时间。所以对于上述问题，笔者正在研究通过网络剪枝技术，在不增加聚类错误的前提下，剪去多余的连接和节点，降低特征向量空间的维数从而减少计算工作量。

参考文献

[1] LINSKER R. An application of the principle of maximum information preservation to linear systems[Z]. Adv. Neural Inform. Process Systems, 1989,1.

[2] JUTTEN C, HERAULT J. Blind separation of sources,Part 1:An adaptive algorithm based on neuromimetic architecture [J]. Signal Processing, 1991,24:10.

[3] COMMON P. Independent component analysis,a new concept[J]. Signal Processing, 1994,36:287-314.

[4] TONAZZINI A, BEDINI L, KURUOGLU E E. Blind separation of auto-correlated images from noisy images using mrf models,. in 4th Int. Symp. on ICA and Blind Source Separation, Nara, Japan, 2003.

[5] SHULMAN D, HERVE J Y. Regularization of discontinuous flow fields. in Proc. Workshop on Visual Motion, 1989:81.86.

[6] BOUMAN C, SAUER K. A generalised gaussian image model for edge-preserving MAP estimation,. IEEE Trans. Image Processing, vol. 2, pp. 296-310,1993.2704.

(收稿日期：2011-03-16)

作者简介：

王 静，女，1981 年生，博士士研究生，主要研究方向：数据挖掘。

刘观宁，男，1954 年生，工程师，主要研究方向：自然语言理解。

张钰辉，男，1989 年生，本科生，主要研究方向：文本挖掘。