# 一种适合动态文本推荐的K_means算法最佳聚类数确定方法

边鹏[1,2]　赵妍[3]　苏玉召[1,2]

(1. 中国科学院文献情报中心，北京，100190; 2. 中国科学院研究生院，北京，100049; 3. 郑州航空工业管理学院，郑州，450015 )

[摘要] 最佳聚类数确定方法是K-means算法的一个研究热点，本文从嵌入式NSTL文本推荐系统的需求入手，分析了原有方法的不足，引入共词分析方法和分化理论，提出了一种新的最佳聚类数确定方法，改进最小类间距离和平均类内距离的计算方法，强化了聚类结果的推荐效果，同时使推荐效果可以随着样本数据的变化而动态调整。实验结果验证了该方法的有效性。

[关键词] K-means聚类；聚类数；文本聚类；个性化推荐

[分类号] TP18, G350

# An improved method of determining optimal number of clusters in K-means clustering algorithm for dynamic text recommending

Bian Peng[1,2]　Zhao Yan[3]　Su Yuzhao[1,2]

1 (National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

2 (Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

3 (Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou, 450015, China)

[Abstract] The method for determining optimal number of clusters has been a focus of K-means algorithm researches, and the authors of this paper analysis the shortage of the original method, based on the text recommending requirement from the embedded NSTL Recommending System. Then the authors raise a novel algorithm to determine optimal number of clusters, and optimize the calculation of the minimal distance between clusters and the average distance within one cluster, introducing the co-word analysis method and the differentiation theory. The improved algorithm enforces the effect of the text recommending in the cluster result. Moreover it can adjust the recommending effect according to the change of the amount of sampling data. At last, the lab result shows it is effective.

[Keywords] K-means Cluster; Cluster number; Text Clustering; Personalized Recommending

## 0. 引言

机器学习技术在个性化文本推荐上的大量应用，使后者获得了长足进步。K-means 聚类算法作为一种典型机器学习技术，得到了广泛使用。K-Means 算法容易实现对大规模数据集的聚类[1]，且其简单直观和易于理解，因而得到广泛应用，但是其缺点也比较明显，除了抗干扰性较差、仅能处理数字属性等问题外，还需要预知类数目[2]，虽然，X-Means 算法[3]能进行最佳类数目估计，但是目前关于 k 值的准确界定依然还是一个难题。为此，不断有人提出自动确定最佳聚类数 k 的方法[4]-[9]，这些方法的一般过程是设计一个函数指标，然后使用聚类算法划分样本数据集合，聚类数在一定范围内变化，计算得到不同的函数指标值，最后，找出最符合条件的指标值对应的聚类数作为最佳聚类数。其中，较为典型的方法有 Calinski-Harabasz（CH）指标[4]、Davies-Bouldin（DB）指标[5]、Krzanowski-Lai(KL)指标[6]、Weighted inter-intra(Wint)指标[7]、In-Group Proportion(IGP)指标[8] 、Between-Within Proportion（BWP) [9]指标。

为了克服 BWP 方法在处理单一样本类时的不足，笔者提出了 BWP4 指标和相应的算法。但该方法在文本聚类时效果有限，无法获得足够的推荐数据，而且它的推荐效果也无法随着样本数据的变化而动态调整，适应不了文本推荐系统数据动态增长的实际情况。例如，在嵌入式 NSTL 个性化检索词推荐系统实验初期，样本量较小，存在一些本该被分到一类中的用户被划分到不同的两类中。本文针对 BWP4 方法的局限性，提出了改进的方法。实验表明，该方法具有更好的效果。

## 1. K-means 聚类算法

K-means 算法是由 Mac Queen[10]提出，该算法是一种基于划分的聚类算法，它通过不断的迭代来实现聚类，当算法收敛到结束条件时就终止迭代过程，得出聚类结果[1]。传统的 K-means 算法[11]涉及到以下的公式及步骤。

设聚类的样本集为 ：$X = \{x_1, x_2, \cdots, x_n\}$ ，n 为样本总数， 得到 C 个聚类中心为$\{z_1, z_2, \cdots, z_c\}$。令 $a_j$ （j= 1，2，… ，C)表示聚类的 C 个类别，$n_j$表示第 j 类的样本数，则 ：

$$z_j = \frac{1}{n_j} \sum_{x \in a_j} x \qquad (\text{I})$$

定义目标函数：

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n_i} d_{ij}(x_j, x_i) \qquad (\text{II})$$

目标函数 J 为每个样本数据点到相应聚类中心的距离平方和，即聚类的最小均方误差。

传统的 K-means 算法步骤[11]如下：

(1) 随机指定 C 个样本点 $z_1$（1），$z_2$（1），…，$z_c$（1）为初始聚类中心；

(2) 按照距离最近的原则，对样本集合聚类，确定每个样本的类属关系；

(3) 使用公式（Ⅰ），计算新的聚类中心 $z_1$（k），$z_2$（k），…，$z_c$（k）(k 表示迭代次数)；

(4) 重复执行(2)-(4)，直到聚类中心稳定为止 。

**2. BWP4 指标含义、不足及改进思路**

为了估算 K-means 算法的最佳类数目，周世兵等人[9]在借鉴了 Peter J. Rousseeuw[12]的指标说明方法，提出了类间类内划分(Between-Within Proportion，简称 BWP) 指标，在寻找最佳聚类数方面进行了尝试，并取得了一定成效，但该方法在处理单一样本类时，存在一定问题，进而 BWP4 方法[16]被提出。

**2.1 BWP4 指标含义**

BWP 指标[9]的定义如下：

**定义 1** 令 K = {X, R} 为聚类空间，其中 X = {$x_1$，$x_2$，…，$x_n$}，假设 n 个样本对象被聚类为 c 类，定义第 j 类的第 i 个样本的最小类间距离 b(j，i) 为该样本到其他每个类中样本平均距离的最小值，即：

$$b(j,\ i) = \min_{1\le k\le c,\ k\neq j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\| x_p^{(k)} - x_i^{(j)} \|^2\right) \qquad (1)$$

其中：k 和 j 表示类标，$x_i^{(j)}$ 表示第 j 类的第 i 个样本，$x_p^{(k)}$ 表示第 k 类的第 p 个样本，$n_k$ 表示第 k 类中的样本个数，$\| \cdot \|^2$ 表示平方欧氏距离。

**定义 2** 令 K = {X, R} 为聚类空间，其中 X = {$x_1$，$x_2$，…，$x_n$}，假设 n 个样本对象被聚类为 c 类，定义第 j 类的第 i 个样本的类内距离 w(j,i) 为该样本到第 j 类中其他所有样本的平均距离，即：

$$w(j,\ i) = \frac{1}{n_j - 1}\sum_{q=1,q\neq i}^{n_j}\| x_q^{(j)} - x_j^{(i)} \|^2 \qquad (2)$$

其中：$x_q^{(j)}$ 表示第 j 类的第 q 个样本，并且 $q\neq i$，$n_j$ 表示第 j 类中的样本个数。

**定义 3** 令 K = {X, R} 为聚类空间, 其中 X = {$x_1$, $x_2$, …, $x_n$}, 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的聚类距离 baw(j, i) 为该样本的最小类间距离和类内距离之和, 即:

baw(j, i) = b (j, i) + w (j, i) =

$$\min_{1\le k\le c,\,k\neq j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\| x_p^{(k)} - x_i^{(j)} \|^2\right) + \frac{1}{n_j - 1}\sum_{q=1,q\neq i}^{n_j}\| x_q^{(j)} - x_j^{(i)} \|^2 \qquad (3)$$

**定义 4** 令 K = {X, R} 为聚类空间, 其中 X ={$x_1$, $x_2$, …, $x_n$}, 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的聚类离差距离 bsw(j, i) 为该样本的最小类间距离和类内距离之差, 即:

bsw (j, i) = b (j, i) — w (j, i) =

$$\min_{1\le k\le c,\,k\neq j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\| x_p^{(k)} - x_i^{(j)} \|^2\right) - \frac{1}{n_j - 1}\sum_{q=1,q\neq i}^{n_j}\| x_q^{(j)} - x_j^{(i)} \|^2 \qquad (4)$$

**定义 5** 令 K = {X, R} 为聚类空间, 其中 X = {$x_1$, $x_2$, …, $x_n$}, 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的类间类内划分(Between-Within Proportion, BWP) 指标 BWP(j, i) 为该样本的聚类离差距离和聚类距离的比值, 即:

$$\text{BWP (j, i)} = \frac{\text{bsw(j,i)}}{\text{baw(j,i)}} = \frac{\text{b (j,i) - w (j,i)}}{\text{b (j,i) + w (j,i)}} \qquad (5)$$

$\text{avg}_{\text{BWP}}$ (k) 表示数据集聚成 k 类时的平均 BWP 指标值, $k_{opt}$ 表示最佳聚类数。

$$\text{avg}_{\text{BWP}}\ (k) = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\ \text{BWP}\ \ (j,\ i) \qquad (6)$$

$$k_{opt} = \frac{\text{argmax}}{2\le k<n}\{\text{avg}_{\text{BWP}}(K)\} \qquad (7)$$

最后，通过上面两个公式找出使 $avg_{BWP}$ (k)值（为简单起见，以下简称 BWP 值）最大的 k，即为最佳聚类数 $k_{opt}$。

笔者针对 BWP 在处理单一样本类时的局限性，提出用类内距离最大值集合中的最小值来表示类内距离 w(j，i)，命名为 BWP4 指标[16]。

**定义 6** 令 K = {X， R} 为聚类空间， 其中，$X = \{x_1, x_2, \cdots, x_n\}$， 假设 n 个样本对象被聚类为 c 类，则

$$w_{single4}(j,i) = \min_{1\le k\le c}\left(\max_{1\le h\le n_k} w(h,k)\right)$$

且 $n_j=1$。

进一步表示如下：

$$w_{single4}(j,i) = \min_{1\le k\le c}\left(\max_{1\le h\le n_k} \frac{1}{n_k-1}\sum_{q=1,q\neq h}^{n_k} \| x_q^{(k)} - x_h^{(k)} \|^2\right)$$

且 $n_j=1$。 （8）

这里，$x_q^{(k)}$表示第 k 类中的第 q 个样本， $n_k$ 表示第 k 类样本的个数。

## 2.2 BWP4 的算法

1) 选择聚类数的搜索范围 [2，n)。

2) 从 2 循环至 n-1

① 调用 K-means 算法；

② 利用式(5)计算单个样本的 BWP 指标值，若类内仅有一个样本，则调用式（8）计算单个样本的 BWP 指标值；

③ 利用式(6) 计算平均 BWP 指标值。

3) 利用式(7) 计算最佳聚类数。

4) 输出最佳聚类数、有效性指标值和聚类结果。

## 2.3 BWP4 在文本推荐方面存在的不足

笔者将 BWP4 方法应用于 NSTL 嵌入式资源服务的推荐系统原型，对注册用户的日志

数据进行相似兴趣建模，使用 K-means 的聚类算法得到用户聚类的结果，见表 1。共有 24 名注册用户的检索信息参与聚类，为保护隐私，隐去真实用户 ID，以用户序号代替。易见，虽然聚类结果的准确性较高——同类中很少有不适合推荐的词，而且聚类结果里也有不同的检索词可供推荐使用，如对使用“web crawler”检索的 10 号用户就会推荐“web”，但同类别的用户几乎都是使用相同的期刊类检索词，真正可供推荐的检索词太少。

| 用户序号 | 用户使用过的检索词 | 类别 |
|---|---|---|
| 1 | web document document library web search web search web web search web retrivel web search web search web search library | 0 |
| 2 | Library | 1 |
| 3 | Library | 1 |
| 4 | Library | 1 |
| 5 | Library | 1 |
| 6 | air crawler web2.0 web web web library library | 10 |
| 7 | computer document 机构 库 development [ Ljava.lang.String;@49481c web search web2 web search web2 web library | 11 |
| 8 | web web crawler | 13 |
| 9 | web crawler | 13 |
| 10 | search crawler crawler crawler crawler crawler crawler web web web web web service web service | 14 |
| 11 | 文献 管理 | 12 |
| 12 | web crawling | 2 |
| 13 | web crawling | 2 |
| 14 | computer computer | 3 |
| 15 | Computer | 3 |
| 16 | Computer | 3 |
| 17 | Computer | 3 |
| 18 | Web | 4 |
| 19 | Web | 4 |
| 20 | 清华 | 5 |
| 21 | Eclipse | 6 |
| 22 | ROUTING AND WAVELENGTH ASSIGNMENT OF SCHEDULED LIGHTPATH DEMANDS IN A WDM OPTICAL TRANSPORT NETWORK | 7 |
| 23 | data mining e gov e gov 2010 education computer data intergration data intergration 2010 data intergration 2009 data intergration 2009 ALL | 8 |
| 24 | web search | 9 |

表 1：同类用户聚类结果

可供推荐的检索词少固然与注册用户数少有一定关系，但是，确实存在着一些更适合推荐的情况存在，如 2-5 号用户都检索了“library”，而 6 号用户检索过“library”两次，在推荐资源较少的情况下，可以作为 2-5 号用户的同类用户。

造成上述文本聚类效果不佳的原因主要是由于将 BWP 方法直接应用到文本聚类中产生的两个问题。一个问题是 BWP 方法的距离计算方法应用到文本聚类时产生的局限性，将用户作为向量，其使用的检索词映射到向量空间里后，样本数据非常稀疏，各个样本之间的空间

距离相距甚远，往往导致样本里的无关属性值发挥了决定作用，聚类结果反映出方法一定的盲目性。另一个问题是，BWP 聚类方法没有考虑到聚类样本集合发展的过程，无论样本集合是新建还是已经稳定运行，均采取一样的策略，不符合信息系统数据动态增长的实际情况，反映其存在一定的机械性。

**2.4 BWP4 指标的改进思路**

要弥补 BWP4 指标的不足，首先让我们看看人类的思维是如何逐渐认知世界的。根据 Gibson 夫妇提出的分化理论[13]，人类从出生开始就在一个变化的世界里主动搜索环境的不变特征（Invariant features），即保持稳定的特征。例如：在模式知觉中，婴儿起初面对的是大量令人摸不到头脑的刺激，但是，不久以后，他们开始搜索刺激的边界突出特征，并且面向一张具有代表性的面孔图像。随着时间的推移，婴儿越来越细致地检查刺激中的不变特征。这也就是说，人类在婴儿期，掌握的信息有限，在识别新事物时，首先是着重考查事物之间的共性，这反映出其聚类的特征；随着人类的发展成长，其掌握信息逐渐增多，这种对强调共性的聚类特征也相应减弱。人类的这种认知过程符合一个信息系统从无到有、从小到大的客观过程。如对于一个新建的推荐信息系统，用户的使用信息较少，可供推荐的资源也较少，这时为了获得更多的推荐可能，可以对推荐的准确度要求适当放宽。

为了模拟人类的认知过程，引入两个方法——共词奖惩机制和发展调节方法。

共词分析方法[14]是由法国文献计量学家提出，该方法利用文献集中词汇对或名词短语共同出现的情况，来确定该文献集所代表学科中各主题之间的关系。一般认为[15]词汇对在同一篇文献中出现的次数越多，则代表这两个主题的关系越紧密。本文在文本推荐过程中，将共词作为“不变特征”，一旦发现两个用户使用的检索词之间存在共词，则尽量将这两个用户聚在同一类中。在 BWP 方法中，理想的情况是同类中的样本相距越近越好，不同类的样本相距越远越好，即平均类内距离 w(j，i)越小越好，最小类间距离 b(j，i)越大越好。笔者将共词引入到 BWP 方法，建立共词奖惩机制，即：

若同类内出现相同属性——共词，则减小平均类内距离 w(j，i)以达到奖励的效果；若不同类间出现相同属性——共词，则减小最小类间距离 b(j，i)以达到惩罚的效果。

为此，笔者构造了共词奖惩变量$\Gamma$：

$$\Gamma = \prod_{p=1}^{k} \mathrm{Share}(x_p^{(k)}, x_i^{(j)}) \qquad (1 \le k \le c，k \ne j，\ \mathrm{Share}(x_p^{(k)}, x_i^{(j)}) \ne 0)$$

$\mathrm{Share}(x_p^{(k)}, x_i^{(j)})$ 表示样本 $x_p^{(k)}$ 与 $x_i^{(j)}$ 在属性比较时相同属性（共词）数量的占比。这样，$\Gamma$ 就随着样本 $x_p^{(k)}$ 与 $x_i^{(j)}$ 有相同属性（共词）的情况增加而迅速减小。将 $\Gamma$ 与 w(j，i)和 b(j，i)分别相乘，就可以模拟出共词奖惩机制。

发展调节系数$\sigma$是用来控制上述共词奖惩变量$\Gamma$随处理信息量的增加而减小的。

$$\sigma = \frac{\mu}{n}$$

$\mu$ 为临界样本规模，n 为样本总数。将$\sigma$作为$\Gamma$幂，即$\Gamma^{\sigma}$。这样，当样本数量 n 小

于 $\mu$ 时，$\Gamma^{\sigma}<\Gamma$；当样本数量 n 大于 $\mu$ 时，$\Gamma^{\sigma}>\Gamma$；随着 n 的增长，变量$\Gamma$奖惩效果不断减弱。从$\sigma$的计算方法易知，临界样本规模$\mu$值成为影响发展调节系数的关键，$\mu$的作用及确定将在下文中讨论。

**3．对 BWP 聚类有效性指标的改进**

量化阴影的方法可能有很多，为了能够与原有 BWP 指标衔接，笔者选取 BWP 的类内距离 w(j,i)表示方块部分的阴影，选取 BWP 指标的最小类间距离 b(j,i)和类内距离 w(j.i)为参考，对单一样本类提出四种阴影表示指标 $w_{single1}$、$w_{single2}$、$w_{single3}$、$w_{single4}$ 替代原来的类内距离 w（j，i）。

**3.1 最小类间距离 b（j,i）的修正**

**修正后的定义 1** 令 K = {X， R} 为聚类空间， 其中 X = {$x_1$，$x_2$，…，$x_n$}， 假设 n 个样本对象被聚类为 c 类， 定义第 j 类的第 i 个样本的最小类间距离 b(j，i) 为该样本到其他每个类中样本平均距离的最小值， 即：

$$b(j,\ i)=\min_{1\le k\le c,\,k\neq j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\|x_p^{(k)}-x_i^{(j)}\|^2\times\prod_{p=1}^{n_k}\mathrm{Share}(x_p^{(k)},x_i^{(j)})^{\frac{\mu}{n}}\right)\qquad(9)$$

其中，$\mathrm{Share}(x_p^{(k)},x_i^{(j)})\neq 0$，$\mathrm{Share}(x_p^{(k)},x_i^{(j)})$ 表示样本 $x_p^{(k)}$ 与 $x_i^{(j)}$ 在属性比较时相同属性（共词）数量的占比，$\mu$ 为临界样本规模，用以衡量在什么样的样本规模下惩罚效果更加明显。

当样本数量 n 远大于临界样本规模 $\mu$ 时，文本样本的分布将更加均匀，容易聚成相似类，此时，惩罚系数接近为 1，惩罚效果可以忽略；当样本数量 n 远小于临界样本规模 $\mu$ 时，如文本推荐系统建立初期，用户少，使用的文本样本数量也少，且分布很难均匀，导致文本样本距离较远，经常形成单一样本类，难以聚成相似类，不利于向用户推荐，此时，惩罚系数很大，惩罚的效果也非常明显。

其中：k 和 j 表示类标，$x_i^{(j)}$ 表示第 j 类的第 i 个样本，$x_p^{(k)}$ 表示第 k 类的第 p 个样本，$n_k$ 表示第 k 类中的样本个数，$\|\cdot\|^2$ 表示平方欧氏距离。

**3.2 类内平均距离 w（j,i）的修正**

**修正后的定义 2** 令 K = {X， R} 为聚类空间， 其中 X = {$x_1$，$x_2$，…，$x_n$}， 假设 n 个样本对象被聚类为 c 类， 定义第 j 类的第 i 个样本的类内距离 w(j，i) 为该样本到第 j 类中其他所有样本的平均距离， 即：

$$w(j,\ i)\ =\ \frac{1}{n_j-1}\sum_{q=1,q\neq i}^{n_j}\| x_q^{(j)}-x_j^{(i)}\|^2\times\prod_{p=1}^{k}\mathrm{Share}(x_p^{(k)},x_i^{(j)})^{\frac{\mu}{n}} \qquad (10)$$

$\mathrm{Share}(x_p^{(k)},x_i^{(j)})$ 表示样本 $x_p^{(k)}$ 与 $x_i^{(j)}$ 在属性比较时相同属性（共词）数量的占比，$\mu$ 为临界样本规模，用以衡量在什么样的样本规模下奖励效果更加明显。

当样本数量 n 远大于临界样本规模 $\mu$ 时，文本样本的分布将更加均匀，容易聚成相似类，此时，奖励系数接近为 1，奖励效果可以忽略；当样本数量 n 远小于临界样本规模 $\mu$ 时，如文本推荐系统建立初期，用户少，使用的文本样本数量也少，且分布很难均匀，导致文本样本距离较远，经常形成单一样本类，难以聚成相似类，不利于向用户推荐，此时，奖励系数很大，奖励的效果也非常明显。

其中：$x_q^{(j)}$ 表示第 j 类的第 q 个样本，并且 $q\neq i$，$n_j$ 表示第 j 类中的样本个数。

## 5. 实验与分析

本文通过将 BWP4 方法和适合动态推荐的 BWP 方法进行数据对比实验，分析改进的效果。首先，通过实验确定关键系数——临界样本规模值 $\mu$，然后通过与 BWP4 方法比较准确性和推荐效果。实验均采用 10 作为 K-means 聚类的随机种子。本文在方法实验中使用了 JAVA、MySQL Server5.1 和 Weka3.6 等工具，硬件平台环境是 Intel(R)Core(TM)2 Duo CPU 2.4GHz 和 2G 内存，操作系统是 WindowXP。

本文的实验使用了 NSTL 嵌入式资源服务的推荐系统原型中的日志数据，事先邀请中国科学院国家科学图书馆的研究生参与使用该原型系统进行文献检索，历时 2 周。笔者运用 K-means 聚类分析用户使用的检索词进而判断用户的相关性，将与用户检索词偏好相似的其他用户的检索词推荐给用户。根据这些数据进行了实验研究，但由于时间较短，参与用户较少，用户的使用数据样本也相对较少，用户之间的相关性并不显著，这种情况适用本文确定最佳聚类数的方法。

日志数据集合通过清洗后得到 24 个注册用户在嵌入式 NSTL 系统中使用过的检索词，经过分类、去除停用词、去重和 Weka 的字符串-词向量转换函数 StringToWordVector（样本对于不存在的属性维度赋值为空），形成 24 个样本 43 维的数据集合，样本分布较为稀疏，只有少部分样本比较相似。

### 5.1 临界样本规模 $\mu$ 的确定

笔者令 $\mu$ 取不同值获得不同的聚类结果，该聚类结果里存在着适合推荐的类数，详见表 2。限于篇幅，本文不再列举每种情况下详细的聚类结果。从表 2 中易见，$\mu$=107 时就可以得到较好的推荐效果，下面的实验中，$\mu$ 就取此值。需要指出的是，临界样本规模 $\mu$ 随着不同样本集合会发生变化，这里实验得到的具体数值仅适合本原型系统得到的样本集合，但这种方法可以应用在更多的样本集合上。

| 取值范围 | $\mu$ <=7 | 8<= $\mu$ <=14 | 15<= $\mu$ <=106 | $\mu$ >=107 |
|---|---|---|---|---|
| 适合推荐的类数 | 2 | 2 | 2 | 3 |

表 2：临界样本规模 $\mu$ 取不同值时得到的适合推荐的类数

### 5.2 对比分析

将改进后的方法与 BWP4 方法分别应用到文本推荐系统上，进行对比实验，详见表 3。

表 3 中列出了人工分类，但是，人工分类的可能性也不只一种，这个表里列出的是为了推荐而进行的分类，同时，两种方法的聚类结果如果与人工分类不符，则用红色字体标出。易见，虽然改进后的聚类结果存在一个错误分类——1 号类（底色为绿色），但是，原有聚为一类的样本在改进后依然聚在一类中——保证了推荐效果未降低，而且，准确率从 70.8%提高到 83.3%，更为重要的是，在原有基础上新增了 6 个同类用户（底色为蓝色），推荐效果更加显著，如对使用“web crawler”检索的 3 号用户原来只会推荐 4 号用户的检索词“web”，改进后可以推荐 2 号用户的“web search”和 1 号用户的“web service”。

| 用户序号 | 用户使用的检索词集合 | 改进后的分类 | BWP4 的分类 | 基于推荐的人工分类 |
|---|---|---|---|---|
| 1 | search crawler crawler crawler crawler crawler crawler web web web web web service web service | 0 | 14 | 0 |
| 2 | web search | 0 | 9 | 0 |
| 3 | web crawler | 0 | 13 | 0 |
| 4 | web web crawler | 0 | 13 | 0 |
| 5 | 文献 管理 | 1 | 12 | 1 |
| 6 | Computer | 1 | 3 | 2 |
| 7 | Computer | 1 | 3 | 2 |
| 8 | Computer | 1 | 3 | 2 |
| 9 | Computer computer | 1 | 3 | 2 |
| 10 | 清华 | 1 | 5 | 3 |
| 11 | Eclipse | 1 | 6 | 4 |
| 12 | data mining e gov e gov 2010 education computer data intergration data intergration 2010 data intergration 2009 data intergration 2009 ALL | 2 | 8 | 5 |
| 13 | Computer document 机构 库 development [ Ljava.lang.String;@49481c web search web2 web search web2 web library | 3 | 11 | 0 |
| 14 | ROUTING AND WAVELENGTH ASSIGNMENT OF SCHEDULED LIGHTPATH DEMANDS IN A WDM OPTICAL TRANSPORT NETWORK | 4 | 7 | 6 |
| 15 | web crawling | 5 | 2 | 7 |
| 16 | web crawling | 5 | 2 | 7 |
| 17 | Web | 5 | 4 | 7 |
| 18 | web | 5 | 4 | 7 |
| 19 | web document document library web search web search web web search web retrivel web search web search web search library | 6 | 0 | 8 |
| 20 | library | 6 | 1 | 8 |
| 21 | library | 6 | 1 | 8 |
| 22 | library | 6 | 1 | 8 |
| 23 | library | 6 | 1 | 8 |
| 24 | air crawler web2.0 web web web library library | 6 | 10 | 8 |

**表 3：改进后的方法与 BWP4 方法在文本推荐系统上的对比**

上述实验验证了新方法的有效性，改进后的 BWP 方法在未降低原有 BWP4 方法有效性的前提下，进一步提高了准确度，获得了更加显著的推荐效果，特别是其可以在小样本数据集合到大样本数据集合的过程中，动态调整推荐结果。

## 6．结语

BWP4 方法应用在文本推荐时遇到两个问题，一个是推荐效果不明显，另一个是无法随着样本数量的变化而动态调整推荐策略，不符合信息系统数据动态增长的实际情况。本文对 BWP4 方法进行了改进，借鉴分化理论的不变特征在人类分类方面所起到的作用，向 BWP4 指标中引入共词分析方法模拟人的认知行为，优化了最小类间距离和平均类内距离的计算方法，使其在强化推荐效果的同时，还可以根据样本数量的变化动态地调整这种推荐效果。将改进的 BWP 方法应用在嵌入式 NSTL 推荐原型系统中，BWP4 已经具有的推荐结果均得到保留，并且获得了更多的推荐数据，准确率也得到提升，取得良好效果。本文的不足之处是：虽然改进的方法可以随着样本数量变化而动态调整聚类结果的推荐效果，但是本文并未在样本增加的情况下实验。未来可以在更为复杂的系统、或更大的样本集合上实践本文的方法。

参考文献：
[1] 谢娟英，蒋帅等，一种改进的全局 K-均值聚类算法[J]，陕西师范大学学报（自然科学版），2010,38(2):18-22.
[2] 姜园，张朝阳等，用于数据挖掘的聚类算法，电子与信息学报[J]，2005,27(4):655-662.
[3] Pelleg D, Moore A. X-means: Extending K-means with efficient estimation of the number of clusters[C]. In: Proc. 17th ICML. Stanford University. 2000: 727-734.
[4] CALINSKI R, HARABASZ J. A dendrite method for cluster analysis[J]. Communications in Statistics, 1974,3(1):1-27.
[5] DAVIES D L, BOULDIN D W. A cluster separation measure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979,1(2):224-227.
[6] DUDOIT S, FRIDLYAND J. A prediction-based resampling method for estimating the number of clusters in a dataset[J]. Genome Biology, 2002,3(7):1-21.
[7] DIMITRIADOU E, DOLNICAR S, WEINGESSEL A. An examination of indexes for determining the number of cluster in binary datasets[J]. Psychometrika, 2002,67(1):137-160.
[8] KAPP A V, TIBSHIRANI R. Are clusters found in one dataset present in another dataset?[J]. Biostatistics, 2007,8(1):9-31.
[9] 周世兵，徐振源，唐旭清，K-means 算法最佳聚类数确定方法[J]，计算机应用，2010，30（8）：1995-1998.
[10] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C] Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, 1967：281-297.
[11] 李飞，薛彬，黄亚楼，初始中心优化的 K-Means 聚类算法[J]，计算机科学，2002,29（7）:94-96.
[12] PJ Rousseeuw，a graphical aid to the interpretation and validation of cluster analysis[J]，Journal of computational and applied mathematics, 1987，20（1）：53-65.
[13] Gibson, E. J. The development of perception as an adaptive process[J]. American Scientist, 1970, 58(1):98-107.
[14] Callon M， Law J， Rip A. Mapping the Dynamics of Science and Technology：Sociology of Science in the Real world[M]. Macmillan, 1986.
[15] 钟伟金，李佳，共词分析法研究（一）——共词分析的过程与方法[J]，情报杂志，2008(5):70-72.
[16] 边鹏，赵妍，苏玉召，一种改进的 K-means 算法最佳聚类数确定方法，现代图书情报技术，2011(9):34-40.

作者简介：
边鹏，男，（1978--），中国科学院文献情报中心，中国科学院研究生院博士研究生，已发文 4 篇。
主要研究方向：网络信息管理技术与信息系统、Web 日志挖掘。
地址：北京中关村北四环西路 33 号
邮政编码：100190
联系邮箱：bianpeng@mail.las.ac.cn
电话：13641395632

赵妍，女，（1979--），郑州航空工业管理学院计算机科学与应用系，讲师，已发文 10 余篇。

主要研究方向：数据挖掘，计算机网络。

苏玉召，男，（1975--），中国科学院文献情报中心，中国科学院研究生院博士研究生，已发文 5 篇。
主要研究方向：网络信息管理技术与信息系统、Web 日志挖掘。