Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

# 自组织映射在专利文本聚类中的应用研究*

□ 曲军伟 乔晓东 桂婕 / 中国科学技术信息研究所 北京100038

摘要：自组织映射（SOM）是一种基于人工神经网络的聚类方法，通过将相似的输入数据映射到相同或者相近神经元达到相似相聚的目的，有着不需要先验知识、保持拓扑结构不变、无监督自我学习和易于可视化的优点。由于专利文献有着数量大、文字晦涩冗长、专业性强等特点，分析难度较大，自动聚类分析能挖掘专利文献内在相似性，作为基础性处理用于后期应用，例如专利数据清洗、专利检索、主题分析和专利地图生成等众多领域。基于SOM的专利文本聚类与传统聚类方法相比效率和准确率较高，并且易于可视化展示。本文使用了SOM、k—means和TwoStep算法分别在专利文本聚类中作了对比，得出SOM较优的结论。

关键词：自组织映射，专利聚类，文本挖掘，可视化
DOI：10.3772/j.issn.1673—2286.2010.09.004

## 1 引言

专利制度在技术革命中起到了非常重要的作用，它分享了人类的智慧财富，缩短了研发时间和开发成本，减少了重复劳动。但是，由于技术创新周期更短、创新过程更加复杂，决策管理者对专利信息的需求越来越快速和精确；同时，专利分析面临着专利数量巨大、篇幅庞大、技术与法律术语并存、用词晦涩、人工阅读专利文献耗费时间长，及人工成本高等问题。以上原因使得专利分析的自动化、规模化、可视化的需求越来越迫切[1]。聚类分析将分析的粒度从单个专利提升到相同类簇上，在保证聚类结果可靠的基础上能有效降低工作强度、提高分析效率、实现自动监测，同时也可以从较宏观角度解析不同类簇的联系，将定量分析和定性分析结合，挖掘数据内涵的意义。此外，专利文本聚类作为文本挖掘的典型应用，能够为专利文档聚合确定相应类别，方便用户准确定位所需信息和分流信息，便于专利数据清洗、专利检索、主题分析和专利地图生成等众多领域，已经成为一项具有较高实用价值的关键技术。

专利信息虽然包含说明书附图，但是主要存储形式是文本，附图也通常需要配加文字说明，而专利的技术信息80%包含于专利文本中[2]。因此，本文聚类研究的对象为专利文献的文本信息。

本文尝试将基于神经网络自组织映射（Self-Organizing Map，SOM）的聚类方法应用到专利分析中，以支持专利聚类效率的提高，增强深层挖掘内涵效果，使分析结果更形象直观。

## 2 自组织映射（SOM）

自组织映射（SOM）是一种基于人工神经网络的聚类算法，在1981年由芬兰人卡汉（T. Kohonen）提出，也被称为Kohonen自组织映射[3-5]。Kohonen认为人脑的特定区域只对特定的行为反应，该行为会正向刺激临近区域，同时反向抑制远离区域，大脑通过“聚类”过程针对性地响应和识别外界信号。自组织映射正是模仿这一生物学特征，无需监督和先验知识自动分析样本的内在特征，揭示样本的相似性，将高维数据映射到低维，实现模式识别和聚类映射。SOM的特性使其在模式识别[6]、优化问题求解[7]和智能控制[8]等自然科学学科得到了一定研究和应用。

* 本文得到国家科技部“十一五”科技支撑计划（项目编号：2006BAH03B03）、中国科学技术信息研究所重点工作项目（项目编号：2009KP01—7—1）、中国科学技术信息研究所2009年度预研基金项目（项目编号：YY—200906）等项目的资助。

## 2.1 拓扑结构

典型的SOM拓扑结构如图 1所示，分为输入层和输出层。输入层节点数等于要输入向量个数，每一个节点接收一个输入向量X的分量$x_i$；输出层一般是一维或者二维的的阵列。对于二维输出层，其拓扑结构以六边形为常见，也可以是矩形或者任意形状。对于每一个输入向量都会在输出层通过竞争的方式得到一个最佳匹配单元（Best Match Unit，BMU），同时模拟生物刺激，通过临域函数（图 2）控制对临近的神经元正强化，远离的神经元负强化，达到相似相聚的目的。

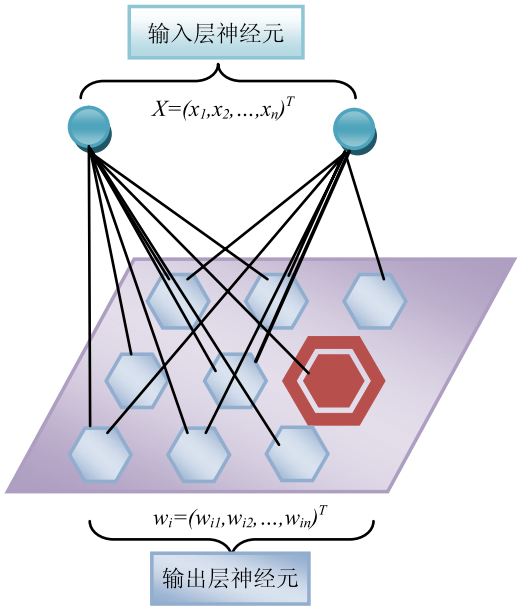


图1 SOM拓扑结构图

## 2.2 算法流程

（1）连接权值初始化 对初始权值向量w(0)取(0,1)区间上的随机值，j=1,2,…,m，其中m是网络输出层神经元数目。

（2）输入 一次将预处理过后向量化的文档$x_i$作为输入向量输入，其中$x_i$是l维向量，l是网络输入层神经元的数目。

（3）训练 步数n（n=0,1,2,…），输入向量为x(n)，权值向量为$w_j(n)$，则进行如下操作：

① 竞争：计算对应x(n)的获胜节点i(x)，即使得x(n)和wj(n)之间的欧氏距离最小的那个神经元节点为BMU。

$$i(x)=\arg\min_j \|x(n)-w_j(n)\|, j=1,2,\cdots,n \quad \text{公式 1}$$

通过这种方式就实现了从高维向量空间映射到二维的点，并保持了其拓扑有序性。

② 合作：计算获胜节点i(x)的临域大小。如果采用正方形临域，临域宽度σ(n)是动态变化的，随时间n增加而逐渐下降。

$$\sigma(n)=\sigma_0 e^{-n/\tau_1} \quad \text{公式 2}$$

其中$\sigma_0$为初始临域宽度，初始临域宽度一般为平面的一半；$\tau_1$为时间常数。

③ 调整：更新②中求出的所有临域内神经元的权重向量

$$W_j(n+1)=w_j(n)+\eta(n)h_{i(x),j}(n)(x(n)-w_j(n)) \quad \text{公式 3}$$

其中学习率 随着时间n增加而逐渐下降，并且：

$$\eta(n)=\eta_0 e^{-n/\tau_2}，\ \tau_2\text{为时间常数} \quad \text{公式 4}$$

神经元i(x)的临域函数$h_{i(x),j}(n)=e^{-d_{i,j}^2/2\sigma^2(n)}$，d(i,j)为i和j之间的距离

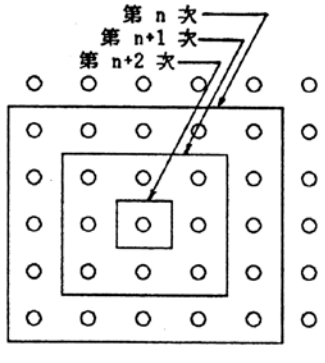


图2 临域函数调整示意图

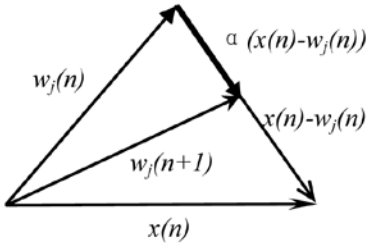


图3 权值调整示意图

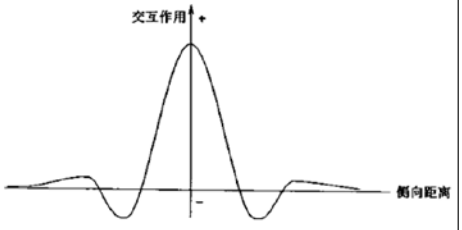


图4 临域函数调整示意图

（4）余下的每一个输入向量重复上述过程(3)，n=n+1，直到W收敛或者达到既定的训练次数。

当所有输入都完成的时候，网络就进行了一次迭代。要得到稳定状态通常需要迭代若干次。在排序阶段，一般迭代1000次或者更多；学习率η(n)初始值较大，接近1.0，然后逐渐减小到大于0.1；临域函数$h_{i(x),j}(n)$初始几乎包括所有神经元，然后随步数逐渐收缩。在收敛阶段，迭代次数一般为网络输出神经元的500倍；学习率η(n)应该在0.01的数量级上，但是不能为0；临域函数$h_{i(x),j}(n)$应该仅包括获胜神经元的最近临域，最终减到一个。

## 3 SOM可视化

SOM网络竞争层的每个神经元都拥有一个二维坐标，分别计算每个神经元与相邻神经元的距离取均值（或者最大值）作为第三维坐标，即高度坐标其平面拓扑如图 5所示。U-matrix的本质是三维空间中用高度展示神经元的相似度，距离越远相似度越低。

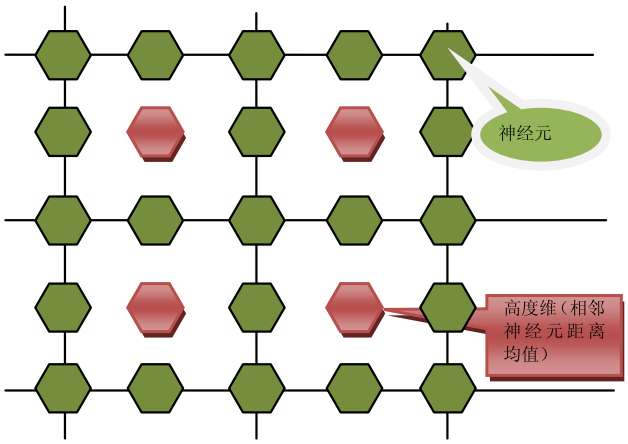


图5 U−matrix拓扑示意图

假设SOM网络输入层位n维，竞争层结构为$n_x \times n_y$，每个神经元的坐标为b(x,y)，其权值向量为wi(x,y)，则神经元(x,y)与其相邻神经元（x+1,y+1）的距离为：

$$d_{xy}(x,y)=\frac{1}{2}\left[\frac{\|b(x,y)-b(x+1,y+1)\|}{\sqrt{2}}+\frac{\|b(x,y+1)-b(x+1,y)\|}{\sqrt{2}}\right]$$

公式 5

U-matrix结构有$(2n_x-1)\times(2n_y-1)$个节点，其中神经元的高度值有两种表示方式，一种是RGB颜色，一种是灰度值。在颜色和距离之间对应一个映射，距离越远，颜色越深，反之，距离越近颜色越淡。通过颜色标注，可以非常形象地展示神经元之间的距离，这个距离反映了不同类簇的差别。

## 4 基于SOM的再聚类

整体看来，基于SOM的文本处理是一种对大规模文档进行聚类特别适合的方法。SOM训练后，生成了一个二维的权值矩阵，每一个列向量对应一个类别。当文本数不多，而且文本集合的选择非常精细时可能进行一次聚类就可得到比较理想的分类结果。但是，由于文本实际类别数少于我们设定的输出层神经元的个数，所以还需要进行再次聚类，将不同的神经元合并为一个类（图 6所示）。有实验证明，如果神经元个数非常少，甚至为一维的情况下，SOM的算法和K-means算法几乎是相似的，体现不出其优势来，这样一来要求SOM应该有较多（适度）的神经元以便更好地体现其特性。这使得单独一个神经元作为一类更加不合理，因此需要将内容相似的神经元合并到一起得到较少的聚类簇（图 6所示）。对神经元手工归类，虽然能提高准确率，但是工作强度非常大，所以Juha Vesanto提出了基于SOM的再分类的方法[9]，将SOM输出层神经元权重作为输入数据使用第二种聚类方法（例如k-means）进行显式地再聚类。

图6 基于SOM的再聚类

关于再聚类的类别数确定方法，一种理想状态是从SOM可视化结果中明显地看到类别数；另一种方法是通过类别数判定指数（例如Davies-Bouldin Index[10]）来确定。

## 5 实验和结果

### 5.1 实验设计和操作

在文本聚类领域，一直缺乏公认的非常优秀标准

测试集，很多学者在聚类测试中通常要自己用类别已知的数据构造，并在此基础上引用信息检索领域的召回率（R）、准确率（P）来衡量算法的性能。因此，本文从专利文献库中随机抽取A-H八大部类（精确到大组）的专利文献各50篇，取专利名称、摘要、主权利要求为文本采集字段。每篇专利文献处理为一个文本文档，共包括三个段落，分别是名称、摘要、主权利要求，保存为“部类+序号.txt”（A01.txt,A02.txt,…,B01.txt,B02.txt,…,H50.txt）存放在一个文件夹中，共计400篇。

摘要是对专利的综合说明，权利要求书是专利申请人请求专利保护的范围，记载了解决技术问题的必要技术特征，是保护专利权利的重要依据。它们的区别度大，适合作用于专利文献聚类。

对既得的测试集进行分词处理、向量化表示后作为SOM的输入层神经元，使用SOM算法排序、收敛。待训练结束后，统计每个神经元映射进来的向量，将每个神经元内的文档都作为一个单独的聚类簇。这样在聚类簇数目多于原来的类别个数时候，对于一个聚类簇i和类别j，$F_1(i，j)$表示i作为j类时候的F1值，对于类别j而言哪一个聚类簇i的F1值最高，就认为聚类簇i是类别j的映射。

$$F_1(i，j)=\frac{2\times P(i,j)\times R(i,j)}{P(i,j)+R(i,j)} \quad \text{公式 6}$$

聚类过程使用K-means和TwoStep算法结果作为对比实验，步骤如下：

（1）分词和词性标注：使用ICTCLAS30工具对专利文本分词并标注一级词性。

（2）去停用词：通过TF和TF/IDF结合的方式自动构建停用词表，并进行人工修改。

（3）特征选择：特征选择不但能有效降低维度，并且可以提高聚类的精度。通过对分词后的专利文本进行分析发现，大多数有意义的词都是名词或者动词，它们或者是对对象的描述或者对技术的改进，区分度比较大。本试验通过词性筛选的方式选择实词构造词空间向量。

（4）文档向量化：文档向量化就是把文本内容数据转换为便于计算机处理的结构化数据的形式。目前，在信息处理领域，向量空间模型（VSM）是应用较多且效果较好的表示方法之一[11]。最基本的思想就是用词袋法（bag of words）表示文本，所有单词不区分文档间的结构关系和先后关系等。文本D用特征项$t_1,t_2,t_3,\cdots\cdots,t_n$来表示，其中，特征项$t_k(1\le k\le n)$指的是文档中的字、词、短语。然后对每一项$t_k$赋予一定的权重$w_k$表示它们在文本D中的重要程度，即$D=D(t_1:w_1,t_2:w_2,\cdots,t_n:w_n)$。在此$w_k$是使用多重因子加权的TF/IDF的特征项权重表示VSM。

TF-IDF是一种用于信息检索与文本挖掘的常用加权技术。TF-IDF是一种统计方法的经验公式，其基本思想是：一个词在特定的文档中出现的频率越高，说明它区分该文档内容属性方面的能力就越强（TF）；一个词在文档中出现的范围越广，说明它区分文档内容的属性就越低（IDF）。公式7是TF/IDF的经典计算方式

$$\text{TF-IDF}=tf_{i,j}\times idf_i=\frac{n_{i,j}}{\sum_k n_{k,j}}\times\log\frac{|D|}{|\{d:t_i\in d\}|} \quad \text{公式 7}$$

本文对TF/IDF进行了补充完善，使其更符合当前语料，加入了位置权重和词长权重：

① 位置权重

国内有人抽样统计，国内中文期刊自然科学论文的标题与文本的基本符合率为98%，新闻文本的标题与主题的基本符合率为95%[11]。所以单词d位置权重location(d)设置为

$$\text{location}(d)=\begin{cases}1.2 & \text{term在标题中出现}\\ 1.0 & \text{term在普通原文出现}\end{cases} \quad \text{公式 8}$$

② 词长权重

一般来说，一个词语的词长越短，它出现的频率就越高，多义现象就更严重，其专指性低，区分度就差；而一个词语的词长较长，频率低，则指代明确，区分度强。因此长词更能精确地反映特征词在文章中的重要程度，应该具有较高的权重，设置词长权重表达式为：

$$W_i(d)=\ln(\text{len}(\text{词})+1) \quad \text{公式 9}$$

得到改进的TF/IDF公式为：

$$TFIDF'=tf_i(d)*location_i(d)*w_i(d)*idf_i(d) \quad \text{公式 10}$$

经过处理后，得到400行3671列的改进TF/IDF矩阵作为输入向量。

（5）聚类操作

① SOM聚类 SOM的输入向量400×3671，设置参数如下：

1, 神经元个数为9行11列；2, 神经元形状为六角

形；3, 神经元权重初始化函数为随机函数；4, 临域函数为高斯函数；5, 排序训练1000步，初始学习率0.8，初始半径9；6, 收敛训练1000步，初始学习率0.02，初始半径4

② 使用k-means聚类 指定聚类数为8，迭代1000次

③ 使用TwoStep聚类TwoStep是spss公司在BIRCH层次聚类算法基础上做了改进，取类簇数为8的聚类结果

④ SOM训练结果k-means聚类 将训练结束后的SOM输出层神经元权值作为k-means的输入向量，指定分类数8进行再聚类

### 5.2 实验结果及分析

① SOM训练结束后，落入各个神经元中的文档向量如表 1所示。

从表 1可以发现，同一个神经元中映射的文档纯

表1 SOM神经元映射结果（部分截取）

| | |
|---|---|
| 1 | B04,B09,B10,B15,B19,B20,B22,B31,B32,B33, |
| 2 | B07,B23,B43, |
| 3 | B02,B11,B16,B17,B44,B48, |
| 4 | B29, |
| 5 | B14,B26,B34, |
| 6 | A50,D28, |
| …… | |
| 16 | B12,B49,D49,F21, |
| …… | |
| 41 | D06,D08,D10,D13,D16,D18,D20,D22,D23,D25,D32,D34,D37,D38,D40,D43,F28, |
| …… | |
| 90 | E03,E04,E08,E17,E21,E22,E23,E24,E25,E26,E28,E32,E33,E34,E36,E38,E39,E41,E43,E44,E46,E48, |
| 91 | E06,E07,E14,E19,E20,E30,E31,E37,E47, |
| 92 | E02,E05,E09,E10,E11,E12,E13,E15,E18,E27,E40, |
| 93 | |
| 94 | D07,D11,D17,D21,D33,D35,D42,D46,D47,D50, |
| 95 | |
| 96 | D04,D15,D41, |
| 97 | H42,H43,H47, |
| 98 | H12,H24,H25,H26,H30,H38,H44, |
| 99 | H04,H08,H09,H13,H15,H17,H18,H19,H20,H22,H32,H35,H45,H49,H50, |

度非常高，几乎都是同类别的，说明SOM在自组织映射方面非常优秀，无监督的自我学习效果非常好。

根据公式6将每个神经元中隶属于某个类F值最大的类作为该神经元的分类，经过整理合并为8大类，统计召回率、准确率结果为表 2。

② K-means算法的聚类结果不十分理想，其中有393篇聚为一类，其余7篇各自为类。究其原因，有学者指出，K-means在进行高维稀疏矩阵聚类的时候由于相似性和差异性表现不明显，容易导致聚类不准确，聚类效果较差。

③ TwoStep算法指定聚类簇为8的时候，得到的聚类结果如图 7所示。

表2 SOM聚类效果评估

| 类别 | 召回率R | 准确率P | F值 |
|---|---|---|---|
| A | 98.00% | 94.23% | 96.08% |
| B | 100.00% | 96.15% | 98.04% |
| C | 100.00% | 96.15% | 98.04% |
| D | 88.00% | 95.65% | 91.67% |
| E | 100.00% | 100.00% | 100.00% |
| F | 86.00% | 91.49% | 88.66% |
| G | 100.00% | 100.00% | 100.00% |
| H | 96.00% | 94.12% | 95.05% |

表 3 TwoStep聚类效果评估

| 类别 | 召回率R | 准确率P | F值 |
|---|---|---|---|
| A | 48.00% | 55.81% | 51.61% |
| B | 56.00% | 48.28% | 51.85% |
| C | 58.00% | 51.79% | 54.72% |
| D | 28.00% | 48.28% | 35.44% |
| E | 60.00% | 43.48% | 50.42% |
| F | 54.00% | 60.00% | 56.84% |
| G | 32.00% | 31.37% | 31.68% |
| H | 70.00% | 71.43% | 70.71% |

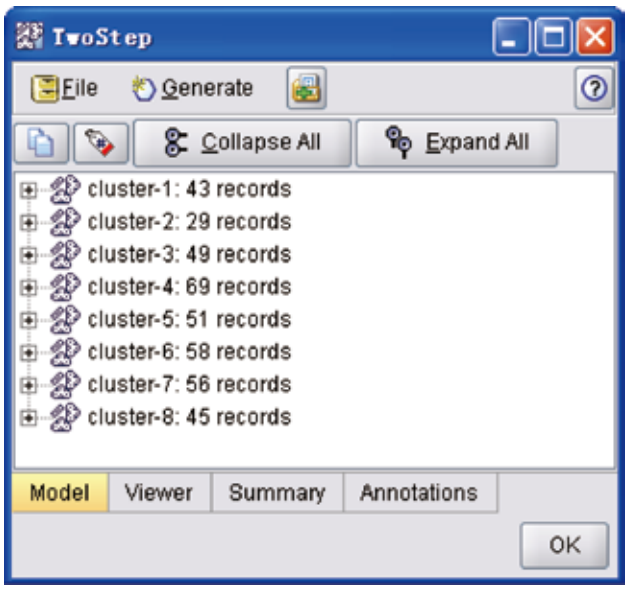


图7 TwoStep聚类结果

表3是对聚类结果进一步处理，统计召回率R，准确率P和F值。

④ SOM训练结果k-means聚类可视化如图 8所示，左侧是神经元之间距离的u-matrix视图，理想状态是类簇间隔明显，表现为图 9中的①②等，山脊位置作为类簇的界定分割线。如果视图明显则可以人工识别类簇数目，解决k-means需要指定聚类数目的缺陷。当分割线不明显时，可以借助很多界定类间距离的指数来确定最优类簇数目，本文使用Davies-Bouldin Index。图8右侧是使用该指数计算最佳聚类簇为15类时候的分簇颜色标记结果。

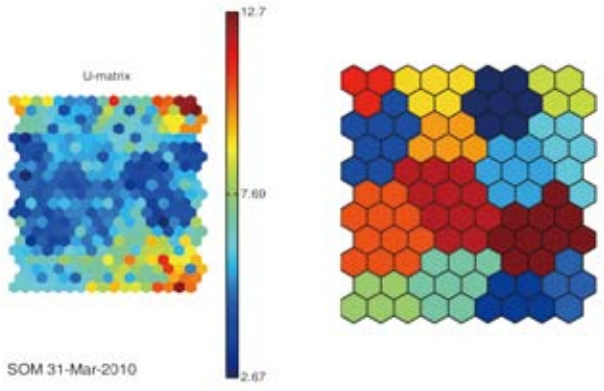


图8 SOM可视化展示

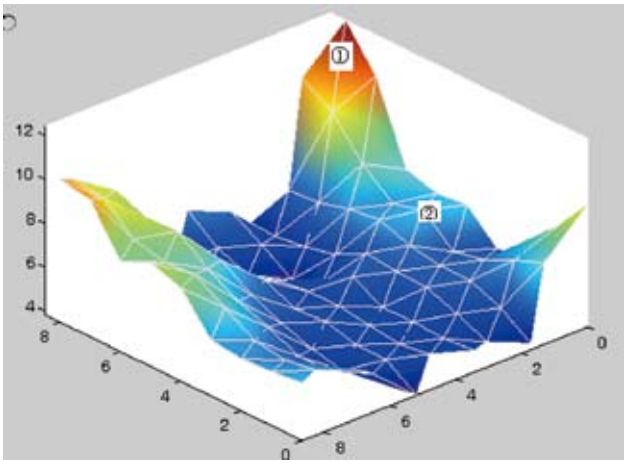


图9 距离3D视图

## 6 结论

综合比较上述3种方法，特别是表 2和表 3的对比

中发现，SOM的聚类效果非常好，优于层次聚类算法BIRCH（TwoStep）和K-means算法。

本次的测试文档仅仅是用词性对词空间进行了筛选，可能会产生高维稀疏现象，同时也会含有噪声词；而专利文献选择的随机性又增添了异常数据出现的可能性，对实验数据的质量都一定影响。在这种实验条件下，基于平面划分的k-means算法夸大了异常点的影响，且忽略了高维向量的相似度，表现较差；SOM算法有效地克服上述问题，聚类效果佳；TwoStep（BIRCH）算法效果居中。

有关本试验的问题，由于SOM只是界定了输入向量的相似性，其实是一个隐性的分类，并没有给出一个明显的边界界定，想要得到显而易见的结果通常还需要借助其他的算法辅助。本文SOM的较高的评测结果是建立在每个神经元都会在界定边界的时候分配到极佳的、正确的簇里，这在实际应用中是比较有难度的。

另外，对SOM输出层再聚类后，可以通过提取类簇中TF/IDF值较大的词（或其他抽取算法）作为类簇的标签，实现聚类结果自动标引，并生成专利主题地图。

综上所述，基于神经网络理论基础的SOM算法，在专利文献的聚类中表现了良好的拓扑保序性，不需要先验知识无监督进行自组织自学习和反馈，拥有良好的高维数据处理能力，将其引入到专利分析领域中是一个成功的尝试，值得进一步研究和推广使用。

**参考文献**

[1] 姜彩红. 基于本体的摘要知识抽取[D]. 北京:中国科学技术信息研究所,2009.
[2] 钟晓,马少平,俞瑞钊. 数据挖掘综述[J]. 模式识别与人工智能,2001,14(1):513-520.
[3] KOHONEN T. Self-Organization and Associative Memory[M]. New York: Spring-Verlag,1989.
[4] KOHONEN T. The Self-Organizing map[J]. Proceedings of IEEE,1990,78(9):1464-1480.
[5] KOHONEN T. Self-Organizing Maps[J]. Springer Series in Information Sciences,1995,v30.
[6] MA Q. Emergence of Chinese Semantic Maps from Self-Organization[C]// Shanghai: 8th International Conference on Neural Information Processing,2001.
[7] 王学敏,程君实,等. 四足步行机器人模糊规则优化算法[J]. 机器人,1997,19(5):360-364.
[8] 余金宝,谷立臣,孙颖宏. 利用SOM网络可视化方法诊断液压系统故障[J]. 工程机械,2007,36:54-57.
[9] VESANTO J, ALHONIEMI E. Clustering of the Self-Organizing Map[J]. IEEE Transaction on Neural Networks,2000,22(3):586-600.
[10] DAVIES D L, BOULDIN D W. A cluster separation measure[C]// IEEE Trans. Pattern Anal. Mach. Intelligence 1979,1:224–227.
[11] 邹娟,周经野. 特征值提取中同义词处理方法[J]. 中文信息学报,2005,19(6):44-49.

**作者简介**

曲军伟(1985- )，中国科学技术信息研究所 硕士研究生在读，研究方向：信息技术应用，特别是数据挖掘技术在专利信息分析中的算法研究和具体实现。通讯地址：北京市复兴路15号信息技术支持中心 100038。E-mail: qujunwei@126.com
乔晓东(1965- )，硕士，研究员，研究方向：信息服务和信息资源管理。通讯地址同上。E-mail: qiaox@istic.ac.cn
桂婕(1976- )，博士，助理研究员。研究方向：专利分析和科技创新管理。通讯地址同上。E-mail: guij@istic.ac.cn

## A Research on Patent Document Clustering-analysis Using Self-Organizing Map

Qu Junwei, Qiao Xiaodong, Gui Jie / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Self-Organizing Map (SOM) is a method for clustering based on ANN, which maps the similar data to the same or close neurons in order to categorize the inputs. SOM has several advantages: requiring no prior knowledge, topological structure mapping, unsupervised, visualization, etc. Using SOM, the patent document clustering is easier, more automatic and efficient.

Keywords: SOM, Patent clustering, Text mining, Visualization

（收稿日期：2010-08-15）