DOI:10.13656/j.cnki.gxkx.1995.03.007
广西科学 Guangxi Sciences 1995 2 (3): 24~27

神经模型在超文档检索中的应用
Application of Neural Model in Hyperdocument Retrieval

覃健文
Qin Jianwen

(广西计算中心广西软件新技术实验室 南宁市星湖路 32 号 530022)
(Guangxi New Software Technology Lab., Guangxi Computing Center, 32 Xinghu Road, Nanning, Guangxi, 530022)

摘要 介绍“轴 K -平均算法”在文档检索中的应用。利用数值分析的方法，先把文档分为各主题，然后对主题里的文档和关键词进行排序形成半轴，由此可得到全局的主题轴和局部主题轴，通过它们，读者就可以查到所需的信息。

关键词 术语 关键词 文档 轴K－平均算法

Abstract Introduces “axial K- means” algorithm using in documents retrieval. By using the methods of numeric analysis, topics are separated from documents, then ranking both documents and keywords in each topic. So the global and local topic axes appeal. In this way, the user can get the information he needs.

Key words term, keyword, document, axial K-means algolithm

1 引言

超文档（Hyperdocument）由超文本界面和数据库组成，用户通过超文本的界面来访问相关的数据库。由于超文本系统结构是高度自由化的，用户从中获得有关信息会遇到极大的困难，因而希望能得到导航的帮助。通常超文本系统非常庞大，而且用户又是根据自己的需要将节点链接起来，加上信息内部的联系，形成纵横交错的网络。如果能用神经网络的方法来处理，是一种好方法。

从用户的角度来看，在浏览超文本系统时，自然希望计算机能够充分理解用户的想法，根据用户的意图和所提供的信息，动态地生成检索的链，根据这些链，用户能很快地获得所需的信息。在形成检索链路的过程中，这些链最好以简单的图形的形式显示出来，这样更能使用户一目了然。

超文本系统是由节点和链组成，而节点可以被看成是独立的语义单元，这些节点用链连接，每条链嵌入某种联系，用户沿着这些链，能自由地浏览文档库。路径显然是有用的，但不适合于多维联合的机制，也限制用户自己任意地存取文档。

本文内容有：查询模式；数据的表达；神经算法；神经浏览器。希望通过这些内容能描述神经机制在超文档中的应用。

2 查询模式

当前流行的模式是交互式的，即查询是通过用户和机器的对话进行。多数有窗口式的图形界面。当前的研究集中在具有用户表达式的查询系统。但基本观点仍是：如何表达系统里数据库的内容。如果没有考虑这一点，那么就不能较好地进行人机对话，而且在大多数超媒体系统里，比较注重于人机界面的设计，较少地考虑到数据库内容的表达。

近年来，对导航的研究也比较重视。导航在浏览文档的过程中是很有用的，用户可以利用手中的导航工具来访问整个数据库，导航的一般策略如下：

- 有效的导航必须在局部进行；
- 导航必须基于对象之间的亲缘（affinity）关系来进行；
- 必须给用户提供直观的超图（Hypergraphic）（节点和链的概图）；

在超文档(超文本技术与文档库组成的超文本系

1995-07-01 收稿。

统）里，对文档库一般比较关心它的文档名、主题（topic）和术语（term）（关键词）。下面给出的“轴K——平均”算法[1]，对超文档的数据表达分两层：

·用2维图来表示全局的对象（即“主题”），而不是基本对象（术语）。

·对每个主题，用1维的“映象图”表示，根据主题里的对象（关键词）与主题的关系密切程度排列成轴，用清单的形式列在窗口上。

这样就把文档分析转化为数值分析的形式，也比较有利于导航，这是因为把以术语描述的文档集转化为由0和1组成的矩阵（在那里0比1多得多）。这个矩阵由文档集里的文档和关键词集里的关键词所确定，利用数值分析的方法就可以对文档集进行分析，因而很容易地实施上面所提到的两层的数据表达。

另外，全局的2维映象图对用户也是基本的定向工具，因为每个主题都把权重（weight）加到术语里，大致地把它刻划出来。

3 数据的表达

前面我们实际上已提到，用矩阵来表达文档集，分两层来表示。对整体而言，我们介绍“轴K－平均”算法，对局部而言，我们介绍“局部成份分析”，“轴K－平均算法”是K－平均聚合（aggragation）算法的变形，它获得全局内部轴的极大化惯性评判的半轴，而不是获得轴中心（原点）。在这些半轴上排列簇的描述符和文档。用这种方法，虽然限制了簇化技术，但能获得对轴的大体解释。它能快速地处理大量的数据。

局部成份分析的方法大致是这样：先给出“粗略”的参数，定义局部全惯性索引或密度测量值，为每条轴遍历所有的数据空间，经过多次遍历后，通过“接合权重向量”，汇合局部的极大值，取得最大点，因而获得主题。

4 神经算法

4.1 轴K－平均算法

神经模型轴K－平均算法如下：

·K个神经先被随机地初始化，它们的权重向量M（k）在描述符空间I里被表示为每簇的引力中心；

·为确定最接近的权重描述，每个数据x和K个权重m(k)之间的距离被计算出和比较。这就等于在两个“扩张”向量：$(1, x_1, x_2, \cdots x_I)^T$ 和 $(\| m(k) \|^2/2, m_1(k), m_2(k), \cdots m_k(k))^T$ 之间寻找极大点η；

·这种神经算法结果对仅有的权重向量进行修改。新的已积累的t-1个元素簇k的引力中心的位置如下计算：

$$m_t(k) = m_{t-1}(k) + (1/t)(x - m_{t-1}(k))$$

这个公式可以理解为简单的学习规则。在数据集里经过多次遍历后，权重向量汇合，而且簇的内容趋于稳定。为使输出的是在原描述符空间而不是在扩张空间里生成的$<x, m(k)>$，目标函数用：

$$\Sigma \mathrm{Max} <x, m(k)>^2$$

$$k=1, K \; x \in \text{簇 } K$$

而不是 $\mathrm{Min}\Sigma\Sigma d^2(x, m(k))$ 结果是：

$$k=1, K \; x \in \text{簇 } K$$

·簇被经过原点O的超平面小方格所分。

·每条簇为由O开始的半轴表示，簇和其他簇的描述符和描述对象在半轴上排列。

4.2 局部成分分析

当聚合（aggragation）时，必须注意数据空间里高度密集的地方，在这里“结构函数”比密度函数更能清确地表达。“结构函数”在数据空间的任意点定义一个密度常数，它依赖于“粗略”参数在两条边界之间的变化。一方面，很多点被集结，另一方面找到单条簇。

单个神经算法模型如下：

转移函数：$\eta'_t = f^+ \{\eta_t (1 - \mathrm{ctg}\theta_0 \mathrm{tg}(m_t, x_t))\}$

这里θ_0是结构函数的参数 $(0 < \theta_0 < \pi)$

X_t是第t个数据向量，m_t是在时间t时刻的权重向量；

$$\eta_t = <m_t, x_t>;$$

$f^+: R \to R;$

$$f^+(x) = \begin{cases} x & x>0 \\ 0 & \text{其它} \end{cases}$$

学习规则：

$$m_{t+1} = m_t + \alpha\eta'_t (x_t - \eta_t x_t)(1 + \mathrm{ctg}\theta_0 \mathrm{tg}(m_t, x_t))$$

这里α为正常数

这个“倾斜上升”规律导出客观对象的极大化

$$\Omega = \sum_{t=1,T} \eta'^2_t$$

给出θ_0，所有极大值构造了绝对最优化。给出处理最少的极大神经元，而且假设每个神经元都必须获得最大值时，本模型提供轴的聚合过程，它是绝对最优化的，比前面提到的“轴K－平均”算法前进了一

步。

5 神经浏览器原理

5.1 全局映射图

提供给用户的全局映象图应是图形的，并能反映出主题的重要性并和主题所在的文档子集里文档的数量（轴K－平均算法）或其它指标如惯性索引值（局部成份分析）相称，比如可用面积不等的园表示。

每个主题标上标签（tag），标签在主题的清单里是最突出的项，但须专家确认才有效。如果该条款不合适，专家可以在清单里选择别的条款，或建议更加适当的条款。

每个文档都连上一些有价值的内容集，如主题轴的投射值、所有的术语、术语的布尔表达式、作者、原标题、作者简历等描述。每选择一次，都高亮地显示一个或几个主题，主题的投射值越大，越灰暗。在用户的显示器里，这样处理强调了给出的术语、术语表达式、作者、原标题、作者简历等相关的上下文关系。

5.2 局部主题

当用户选中一个给定的主题，所选的窗口接着显示连到这个主题上的所有文档、术语、作者、原标题、作者简历等等清单，再选择这些清单的一个元素就得出进一步更详细的信息：

——详细的文档参考资料

——术语、作者、索引文档的数量、第一次建立和最后一次索引文档的日期等。

5.3 导航问题

用户可以通过查询初始化来处理导航问题。他可以对某一主题进行直接探索。在主题里，某些术语和参考文献似乎与他的研究使用特别有关系，他可以选择它们的进一步信息，而且也可以考虑在全局映射图里所选择高亮显示的其它的上下文关系，如果这些在上下文关系里看上去有与他关系更紧密的，他可以继续探索它，等等，如此重复下去。

在这里，用户可以做如下事情：

· 查看相关上下文里与某个术语相连或与某个综合啶 文档密切相关的文档。注意，在这些文档里，可能没有通常的初始查询的文档的术语。查询过程是与上下文相关的，这个性质对多义术语特别有用。

· 确定在给出的上下文关系里与所列的作者相似的作者。

· 找出有哪些实验室也在对给出的主题进行研究。

· 知道关于所列的概念或研究方向里哪个术语的变种正被索引者采用。

L. Alain 利用上述算法。已开发出在 Mac 机器的 HyperCard 环境里运行的神经浏览模型，他是这样做的：

主题簇有下列两个特点：

· 这些簇是交叉的，作为文档或关键词可以同时属于几个簇。

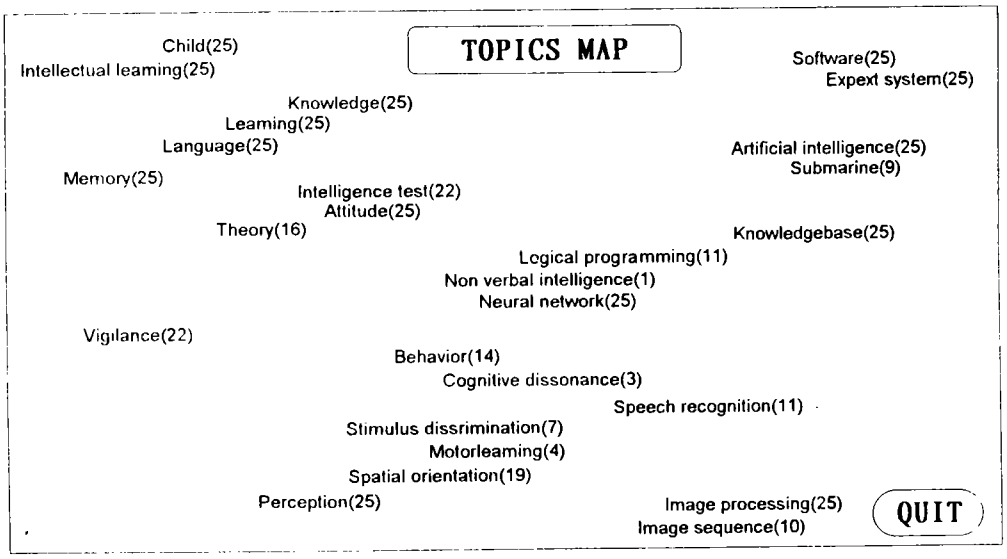

图 1 全局主题映象图 概述认知科学领域的 529 篇文档，括号里的数字是互不交叉的簇的文档的数量（至多 25 篇文档）

Fig. 1 A global topic map, summarizing 529 documents in the field of cognitive science Numbers in brackets indicate the size of documents non－overlapping clusters (up to 25 documents, due to implementation constraints of the prototype)

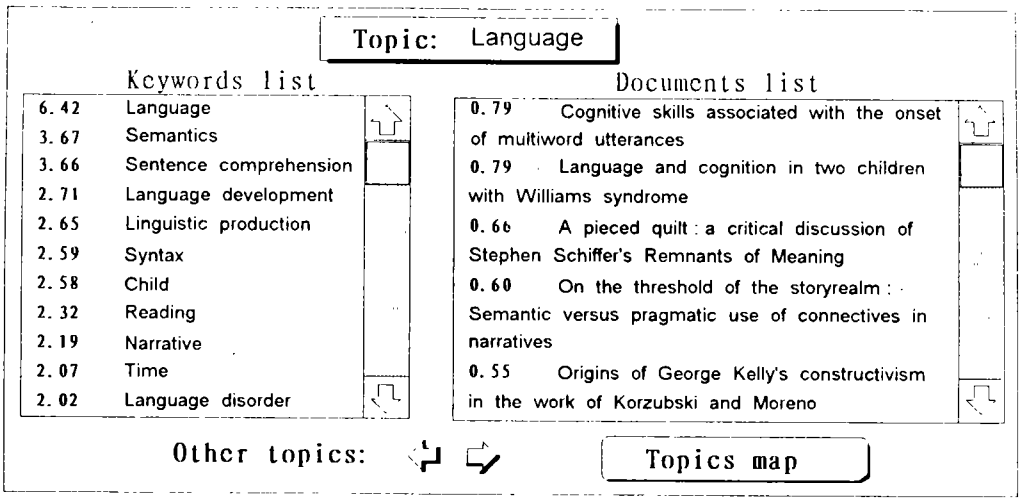

图 2　主题卡片　显示关键词清单和文档清单，清单左边的数字分别是关键词和文档在主题轴的坐标

Fig. 2　A topic card, displaying the kerwords windows and the documents windows. The decimal numbers indicate the coodinates of kerwords and documents along the topic axis.

·每条簇的元素（即文档或关键词）根据它们的“意思－类型”相似的程度来排列。

用户拿到的浏览器是包含在 HyperCard 卡堆里的。高端卡是全局映射图（见图 1），它显示有关主题的位置。这个图是建立在整个关键词向量空间主题范围的基本成份分析上的（布局由标准的事务图形软件解决）。每个主题表达成轴，在轴里，文档和关键词被分组排列（见图 2）。通过选择一个主题，用户可以访问它的内容。主题由关键词和相关的文档清单组成，根据关键词和文档在主题轴里的投射的值来排列，排列后的关键词和文档清单显示在窗口上。

卡堆由三种不同类型的卡片组成：主题的全局映射图，连到主题的关键词和排好序的文档清单，文档标题、作者和简历、原标题、关键词和摘要见图 2。

6　结论

超文档结构过于自由，链的连接由于作者的观点而各异。对这样自由的系统，如果一切都依靠读者去判断，那么读者的负担必然不轻。因此开发出高效、智能的超文本系统是一条出路，在超文本系统中应用神经机制是一种好办法。

参考文献

1　Alain L. Chaire F.. Hypertext paradigm in the field of information retrieval: a neural approach. Proceeding of Hypertext, 1992, 112～120.

2　Biennier F., Guivarch M., Pinon J M. Browsing in hyperdocument with the assistance of a neural network. Hypertext: concepts, systems and applications, Cambridge university press, 1990, 288～297.