计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700161

改进的 GHSOM 算法在民航航空法规知识地图构建中的应用

张浩洋　周　良
南京航空航天大学计算机科学与技术学院　南京 211100

摘　要　针对文本聚类过程中簇的数量无法动态改变及文本分类结果不够精确等问题,文中引入并改进了成长型分级自组织映射(Growing Hierarchical Self-Organizing Map,GHSOM)算法,以提高文本聚类的精确度,并尝试使用改进后的 GHSOM 算法构建民航航空法规知识地图。GHSOM 算法为多层分级结构,每一层包含数个独立的成长型 SOM,通过增长规模来在一定程度上更加详细地描述数据集,提高分类效果。在此基础上,以民用航空领域的各项法律、法规条文为样本资料集,结合中文分词、关键词提取、文件向量等技术手段,利用改进的 GHSOM 算法对文本进行聚类分析,并最终完成民航航空法规知识地图的构建。实验结果表明,所提算法具有显著的文本聚类能力,利用该算法构建的民航航空法规知识地图取得了较好的分类效果,其精确度、召回率等评价指标也获得了进一步的提升。

关键词:知识地图;自然语言处理;文本聚类;word2vec;GHSOM
中图法分类号　TP391.1

Application of Improved GHSOM Algorithm in Civil Aviation Regulation Knowledge Map Construction

ZHANG Hao-yang and ZHOU Liang
School of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211100,China

Abstract　Aiming at the problems that the number of clusters cannot be dynamically changed and the text classification results are not accurate enough during the text clustering process,this paper introduces and improves the Growing Hierarchical Self-Organizing Map (GHSOM) algorithm to improve text clustering accuracy,and tries to use the improved GHSOM algorithm to build a knowledge map of civil aviation regulations. The GHSOM algorithm has a multi-level hierarchical structure,and each layer contains several independent growing SOMs. Through the growth of the scale,the data set is described in more detail to a certain extent,and the classification effect is improved. Based on this,taking various laws and regulations in the field of civil aviation as the sample data set,combined with Chinese word segmentation,keyword extraction,file vector and other technical means,the text is clustered and analyzed using the improved GHSOM algorithm,and finally the construction of civil aviation regulation knowledge map is completed. Experimental results show that the proposed algorithm has significant text clustering ability. The civil aviation regulation knowledge map constructed by this algorithm has achieved good classification results,and its evaluation indicators such as accuracy and recall rate have been further improved.

Keywords　Knowledge map,Natural language processing,Text clustering,word2vec,GHSOM

知识经济时代,知识已经成为组织中继物质资源及人力资源后的又一重要资产。组织如何对内部现有知识及外部知识进行有效管理并不断创新已成为组织知识管理的重要课题。知识管理就是组织对知识的获取、传递、共享、传播和创新的过程,其目的是在整个知识管理过程中将最恰当的知识在最恰当的时间传递给最恰当的人,以便其能够利用这些知识作出最恰当的决策。组织知识管理的有效实施离不开信息技术的支持,而知识地图就是利用现代信息技术对组织知识进行管理的有效工具之一。

布鲁克斯提出的"知识地图"主要是指人类的客观知识,他认为人类的知识结构可以绘制成以各个知识单元概念为节点的学科认识地图[1]。而目前对知识地图的定义多是从信息组织和知识管理的角度进行。Davenport 和 Prusak 认为知识地图具有索引的功能,标示出组织中知识的位置,但是无法直接取得知识的内容。当组织成员需要某项专业知识时,可通过知识地图的指引找到所需的知识[2]。EPPLER 指出在组织内推行知识管理时,知识地图具有增加和扩大组织智力资产的功能,他将知识地图定义为:可视化地显示可获得的信息及其相互关系,它促使不同背景的使用者在各个具体层面上进行知识的有效交流和学习。这样的地图包括的知识项目有文本、图表、模型和数字[3]。其他文献中对知识地图的定义大体上与以上定义相似或重叠。

在知识地图的构建技术及构建实例方面,知识地图的研究已由探讨性研究转向实际构建技术的研究。知识地图的构建是一项复杂的系统工程,它需要结合知识管理、社会学、数据挖掘和人工智能等领域的知识。知识地图模型由 3 层构成,如图 1 所示。

通信作者:张浩洋(zhang405002@163.com)

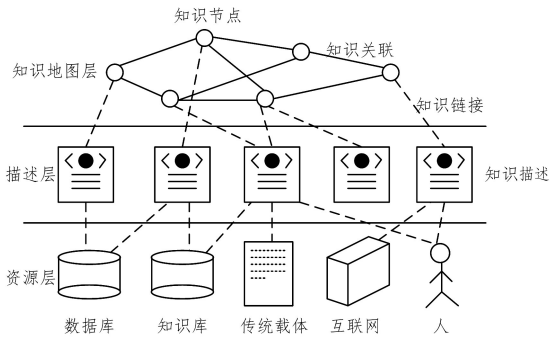

图 1 知识地图模型

Fig. 1 Knowledge map model

在传统的知识建模方面的研究与尝试主要有:叶范波等对产品设计过程的知识进行了定义和描述[4]。潘星等给出了一种基于概念聚类的知识地图模型[5]。此外主题图中的主题建模方法以及本体建模技术可以应用到知识地图的知识建模中,利用主题、本体建模工具可以提高建模的效率及准确性,节省人力和时间。现有常用的本体编辑工具有 Ontosaurus, WebOnto, Protege2000 等,常用的本体及主题图描述语言有 KIF, Ontolingua, CycL 等。利用这些描述语言及辅助构建工具可以统一知识的描述格式,有利于知识在不同平台之间共享,有利于知识地图的及时更新及知识概念的合并。

在大数据不断发展的新阶段,知识地图的构建应该结合数据挖掘及人工智能技术。数据挖掘和人工智能技术可以应用于知识地图模型中资源层与描述层之间。数据挖掘中的关联挖掘可以帮助从数据库、知识库及文本中发现概念,丰富已有知识地图中的概念,关联规则的发现可以帮助提取概念之间的层次性关系及确定适当的抽象层次等。在应用数据挖掘及人工智能技术构建知识地图方面,Li 等针对社区的问答服务数量迅速增加,大量用户需要获得有用的知识这一现象,利用问题-答案对的特征来确定问题-答案对的更精确位置,从而构建了问题和答案档案的知识地图[6]。查文文等针对学科知识特点,研究了学科知识表达方法,提出了多元独立参数的学科知识表示与量化方法,并基于知识地图的生成进行探索,为学科知识表达与展示提供了一种新的思路[7]。范云霞等借助知识地图对学习资源的可视化优势,以《C 程序设计》课程作为系统的资源,实现一个在线个性化学习平台,以提高学生在学习课程时的兴趣和注意力[8]。赵彤等针对数字教育资源服务在资源组织和管理方面的不足,设计了基于知识地图的数字教育资源服务模式,依托 Visual Studio 开发环境,利用 Google Maps API 构建知识地图,实现了数字教育资源服务模式[9]。

本文将首先介绍本次构建民航航空法规知识地图的总体系统架构并给出系统流程图;第 2 节介绍改进的 GHSOM 的算法原理和主要算法流程;第 3 节对实验过程和实验结果做出具体阐述和分析;最后对本文进行总结,并提出未来可行的研究方向。

1 民航航空法规知识地图构建流程

本文提出的知识地图构建流程如图 2 所示,系统整体上由两大模块构成。首先,对所收集的法规文件集进行预处理,将文本信息分解成词并去除无用词组。然后,根据预处理的结果对关键词进行提取并为每份文件生成文件向量。最后,利用改进 GHSOM 算法发展并构建知识地图,依据每个群集内的建议标签,由人工判读给定该分类的主题名称。

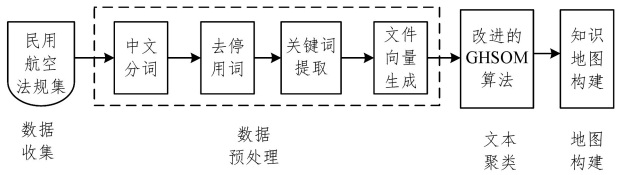

图 2 民航航空法规知识地图构建流程

Fig. 2 Construction process of civil aviation knowledge map

2 改进 GHSOM 算法的设计

2.1 GHSOM 算法的基本原理

由于 SOM 算法存在一些显著的局限性,如网络结构是固定的不能动态改变、网络训练时有些神经元始终不能获胜成为"死神经元"、网络连接权的初始状态和算法中的参数选择对网络的收敛性能有较大影响等,一些学者提出了不同的改进算法,从不同方面不同程度地克服了这些缺点。Dittenbach 等提出 GHSOM 方法,以克服前述问题[10]。

GHSOM 为奠基 SOM 的一种类神经网路模式,它可使众多的资料除了前述 SOM 的单张地图二维呈现外,还可以对多层地图架构进行分群呈现,即依阶层呈现,可更便于资料分析与探索,是一个可用于处理高维度特征空间的稳定且可调整的模式。它可克服前述需事先固定地图大小与非阶层式调整地图架构的问题,可根据资料的结构发展地图大小与阶层架构(见图 3)。

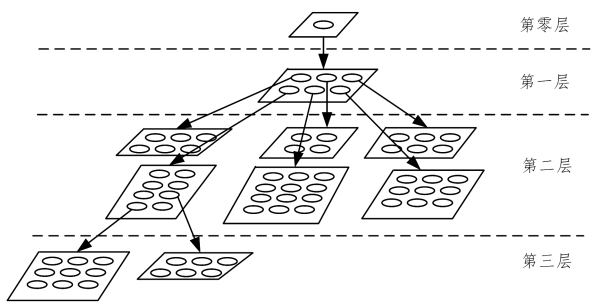

图 3 GHSOM 的典型结构

Fig. 3 Typical structure of GHSOM

GHSOM 是多层分级结构,每一层包含数个独立的成长型 SOM,通过增长规模来在一定程度上更加详细地描述数据集。当其表示一个由较为复杂且相似度较低的数据集组成的单元时,其神经元将被扩展,并在下层形成一个小的成长型 SOM;而当其表示一个相似数据集的单元时,其神经元无需被进一步扩展。因此,GHSOM 通过特有的结构与数据固有的分级结构,表明出良好的适应性。

图 3 中,第 0 层为虚拟映射,为成长过程提供服务。第 1 层映射提供输入数据中主要映射的粗略组织。第 2 层中的两个单独的 SOM 提供数据更详细的表示。同时,数据结构不同,映射的规模也不同。

2.2 改进 GHSOM 算法的流程

根据 GHSOM 的原理,本文算法的主要步骤如下[11]:

(1)计算第 0 层单元的量化误差 qe,计算式如下:

$$qe_0 = \sum_{x_j \in c_0} \| m_0 - x_j \| \tag{1}$$

其中，c_0 为映射到第 0 层单元上的输入向量集（即前述步骤中生成的文件特征向量），为全部向量集；m_0 代表输入向量的平均值。

(2)构建第 1 层映射为 2×2 个单元的 SOM，采用 K-means 方法对向量权值进行初始化，并设置此网络为活动网络，活动网络层级数为 1，训练数据集为全部数据集。

(3)使用 SOM 训练算法训练活动网络，其具体内容如下：

以 t 代表目前训练的回合(iteration)，在每个学习回合 t，输入向量 $x(t)$ 是从所有输入向量中随机选取的并载入地图(map)，选取最高活动层次(activity level)的处理单元 c 为优胜单元 $c(t)$ 来调整，并对其进行调整。通常，处理单元的活动层次是奠基于输入型态与该单元权重向量间的欧氏距离(Euclidean distance)，即最小的欧氏距离为优胜者(winner)。用 i 表示输出空间中的单元，则优胜单元 c 是以式(2)选取：

$$c(t):\|x(t)-m_c(t)\|=\min_i\{\|x(t)-m_i(t)\|\} \tag{2}$$

在每个学习的回合中，会发生前述的调整，为缩短该输入型态与权重向量间的差异，调整的量由学习速率(learning rate)α 所导引，α 调整的原则为在网路初始学习阶段进行较大幅度的调整，随着学习时间愈长，α 将逐渐递减，最后仅进行微幅的调整。在学习时期，优胜者周围的邻居(neighborhood)会被调整朝向该输入型态，这使得属性类似的输入型态可被映射(map)到彼此较靠近的输出单元方格，因此，SOM 的学习过程会产生输入型态(input patterns)的拓朴排列。

在优胜单元周围的邻居可表达为邻近核心函数(neighbor-kernel) h_{ci}，其可视为一种距离的概念，即输出空间中单元 i 与该回合中优胜单元 c 之间的距离，此邻近核心指派一个介于 0 与 1 之间的纯量，以确保附近的单元所调整的强度大于远方的单元，通常以 Guassian 函数(见式(3))表达此邻近核心函数：

$$h_{ci}(t)=\exp\left(-\frac{\|r_c-r_i\|^2}{2\cdot\delta(t)^2}\right) \tag{3}$$

其中，$\|r_c-r_i\|$ 表示输出空间中单元 c 和 i 之间的距离，即 r_i 表示在输出方格中指向处理单元 i 位置的二维向量。通常在学习过程的初期阶段，会选择足够大的邻近核心函数以涵盖整个输出空间，邻近核心函数的空间宽度渐渐随着学习程序缩小，最后仅剩下优胜单元本身被调整，缩小调整范围的速率是由式(3)中随时间调整的参数 δ 来决定的。由以上 SOM 训练的原则，我们用式(4)来表示其学习法则，其中 t 表示现在的学习回合，α 表示随时间变动的学习速率，h_{ci} 表示随时间变动的邻近核心函数，x 表示现在使用的输入向量，m_i 表示指派给单元 i 的权重向量。

$$m_i(t+1)=m_i(t)+\alpha(t)\cdot h_{ci}(t)\cdot[x(t)-m_i(t)] \tag{4}$$

(4)计算活动网络内所有神经元的量化误差 qe_i，并根据平均量化误差 MQE：

$$MQE_m=\frac{1}{n_\mu}\cdot\sum_{i\in\mu}qe_i,n\mu=|\mu| \tag{5}$$

计算当前网络的 MQE_m 值。其中，m 为活动网络所在层级数，qe_i 出自数据投射到的映射单元的子集 μ。

(5)验证级内终止条件：$MQE_m<\tau_1\cdot qe_u$。其中，qe_u 是相应的上层单元的量化误差。条件成立时，转第(7)步。

传统的 GHSOM 算法在该验证操作后，τ_1 值不做改变，并继续按照优胜单元选择法发展下一阶 SOM 子网络，这样往往会使下级 SOM 网络无法精确地表达映射到它们的数据集的结构，降低了聚类效果和文件间相似性比对的准确度。本文针对发展下层 SOM 网络精确度不够理想的问题，调整了验证级内的终止条件，并引入赤迟信息量准则。

赤迟信息量准则由日本统计学家赤迟弘次创立，它是建立在熵的概念基础上，可以权衡所估计模型的复杂度和模型拟合数据的优良性[12]。本文用赤迟信息量准则来衡量独立的 SOM 网络的聚类精度与网络的复杂度。赤迟信息量准则如下：

$$AIC=-2(-k\ln\frac{n}{k}-\sum_{j=1}^{k}\ln(\frac{d_{\max}-d_{\min}}{n_j}))+2u \tag{6}$$

其中，k 为聚类的簇的数目；n_j 为簇 j 的输入向量的数目；n 为输入向量的总数；$d_{\max}$，$d_{\min}$ 分别为类内偏差的最大值和最小值；u 为该 SOM 网络的神经元的总数。当聚类的簇越多，类内的偏差会越小，AIC 值会变小，但是神经元的数目会增加，这又会使 AIC 值变大。只有均衡了两者才能使得 AIC 值整体变小[13]。

本步骤在改进后，当验证级内终止条件的操作结束时，更新水平生长参数。如果网络的神经元个数大于$\sqrt{d}$（d 为映射到该网络的输入向量的总数），则计算网络的 AIC 值并进入下一步，否则降低 τ_1 的值，建立一个新的 2×2 的网络后转向步骤(3)。最后在选择 SOM 网络时，选择具有最小 AIC 值的网络为 GHSOM 网络的子网络。

改进后，SOM 的每次迭代不再使用优胜单元 c 作为其优先处理对象，而是采用 AIC 值作为判据，选择最优网络为 GHSOM 网络的子网络。

(6)选取活动网络中 qe 值最大的单元，标记为错误单元 e。然后按下式得到最相异的邻居 d：

$$d=\arg\max(\|m_e-m_i\|),m_i\in N_e \tag{7}$$

其中，N_e 是 e 的邻居集。在 e 和 d 之间插入一行新的单元[14]，重置 SOM 参数，转第(3)步。

(7)对活动网络单元逐个验证全局终止条件：$qe_i<\tau_2\cdot qe_0$。发现不满足条件的单元时，计算该单元 4 个邻居的模型向量值，然后构建以此 4 个向量值为初始值的 2×2 新映射网络，并设置新建网络为活动网络，层级数加 1。将映射在该单元上的数据作为训练数据，转第(3)步。

(8)完成一个活动网络的验证时，将此网络父亲单元所在网络设置为活动网络，层级数为 1 时结束。否则，层级数减 1，转至步骤(7)。

至此，一个完整的 GHSOM 训练过程执行完毕。

3 实验结果与讨论

3.1 民用航空法规集

本文收集了《中华人民共和国航空法》《中华人民共和国航空器适航管理条例》《中国民用航空规章》《导航设备开放与运行管理规定》《空中交通管理运行单位安全管理规则》《民用航空安全管理规定》《民用航空安全信息管理规定》《民用航空器事故和飞行事故征候调查规定》《气象工作规则》《气象探测环境管理办法》《情报工作规则》《通信导航监视设备飞行校验管理规则》《无线电管理规定》《中国民航航空安全方案》《中国民用航空安全规划纲要》等 80 余篇（不包含航空税法及相关法律法规）最新修订的针对民航领域的法律文件及管理规定。

其中包含行政规则、航空器、航空人员、空中交通管理、机场、航空安全保卫等数十项主要内容。由于短文本长度较短，但是信息描述能力强，因此在预处理环节，每篇法律文献将被按章节拆分，使其变为数个小段文本，随后用这些文本作为文件集用以生成文件向量。

3.2 实验环境与参数设定

在数据预处理环节中，中分语言分词采用混合分词技术，语料库选用维基百科中文语料库，停用词处理环节依照百度停用词表进行处理。

在关键词提取阶段，TF-IDF 技术通常被用来评估一个词汇对于一个文件的重要程度。TF 指的是某一个给定的词语在该文件中出现的频率；IDF 是逆向文件频率，是一个词语普遍重要性的度量。因此，本文采用 TF-IDF 算法按每个词的权重计算结果，在最终结果集中取前 2000 个关键词。

在文件向量的生成过程中，采用已经训练好的 word2vec.tgz 训练模型。

在改进 GHSOM 的参数设定上，起始的学习速率设为 0.5，起始邻近距离设为 3，起始的地图大小设为 2×2 方格（亦即两列两栏的方格）。经过尝试错误法（trial-and-error）后，决定将地图大小设为 0.1，阶层深度设为 0.01。

整个实验过程处于基于 Python 3.6 版本的 Anaconda 实验环境中。

3.3 实验结果

民用航空法规集在经过改进的 GHSOM 聚类后，由人工判读每一类的主要关键词，并依此为每一类设定主题名称。在第一层共产生九大类主题，各部分主题的内容在整个知识地图中的数量占比如图 4 所示。经过人工总结和给定主题词后，文本聚类结果可以被分为安全相关、导航与通信、技能培训、行政检查与规范、气象相关、交通管理、情报工作、违规处理、执照管理与审核九大主题。知识地图的构建结果如图 5 所示。

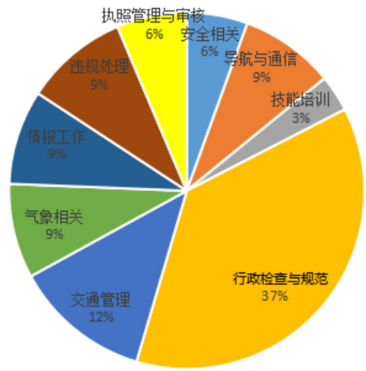

图 4 各类主题内容占比情况

Fig. 4 Proportion of various types of theme content

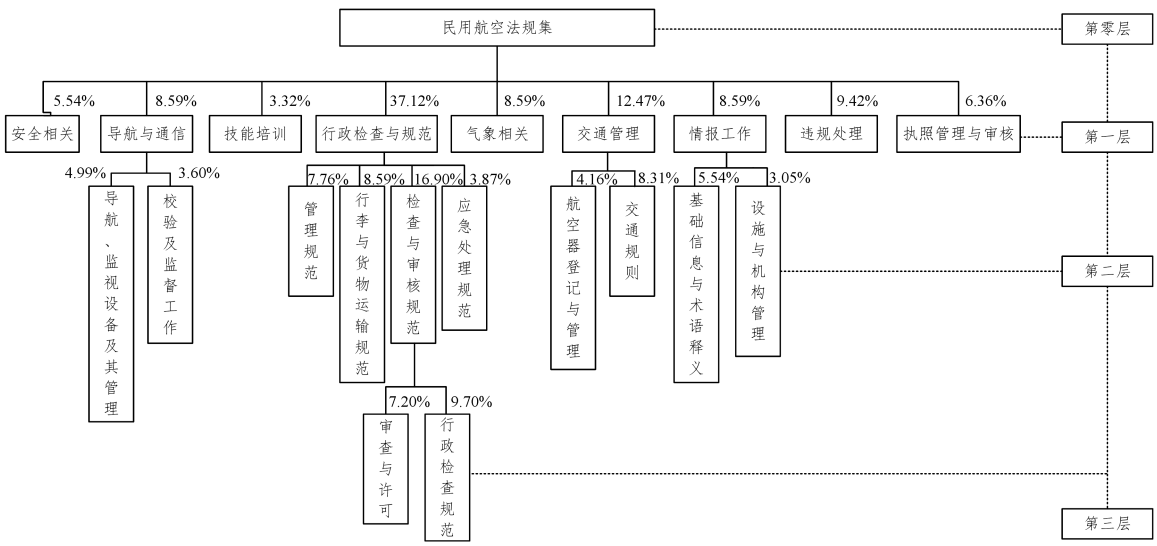

图 5 民航法规知识地图总体结果图

Fig. 5 Overall result of civil aviation laws and regulations knowledge map

3.3.1 安全相关

在第一类安全相关主题内，包含了对民航各运营单位的安全监督及管理规范，本类别中各项法律条文均为保证民用航空安全及正常运行而服务。其中包括《空中交通管理运行单位安全管理规则》第一至六章内容、《民用航空安全管理规定》第一至四章内容、《民用航空安全信息管理规定》第一、六章内容等共计 20 项内容。例如，《民用航空空中交通管理运行单位安全管理规则》在第四章中规定了：发生或者发现民航空管不安全事件，事件发生地的民航空管运行单位应当按照规定及时报告。

上述内容展示了第一类主题中法律内容的特点，此类相关信息均被第一类安全相关主题类别所收纳。

3.3.2 导航与通信

在第二类导航与通信主题内，包含了对通信、导航设备及相关校验机构的一系列规则和要求，规范了通信导航监视及民用航空飞行校验等工作，阐述了相关许可证的获得条件。在这个主题内，本算法将其进一步发展为两类第二层主题，分别为导航、监视设备及其管理主题和校验及监督工作主题。

导航、监视设备及其管理主题中的内容主要以加强对民用航空导航设备的运行管理、保障飞行安全、规范民用航空通信导航监视工作为主，包含了《导航设备开放与运行管理规定》第一章至第四章、《民用航空通信导航监视工作规则》第一至七章等共计 18 项内容。校验及监督工作主题中的内容主要以规范民用航空飞行校验工作、对通信导航监视服务有关的单位实施监督检查等工作为主，包含了《民用航空通信导航监视工作规则》第八章、《通信导航监视设备飞行校验管理规则》第一至六章等共计 13 项内容。此类文本信息均被归纳至第二类主题。

3.3.3 技能培训

在第三类技能培训主题内，包含对不同岗位、不同级别培训的介绍，阐述了参加各项培训需要具备的要求及期望达到的目标，并对培训机构的规范性作出了相关要求。其中包括《民用航空空中交通管制培训管理规则》第一至五章内容、《民用航空情报培训管理规则》第一至五章内容等共计 12 项内容。此类文本信息均被归纳至第三类主题。

3.3.4 行政检查与规范

在第四类行政检查与规范主题中，包含了各项行政检查工作的规则与规范、各类经营许可的获取条件、各类设备设施的运行要求及其他各项具体工作的管理规定。该主题被划分为 4 个次级主题，分别为管理规范主题、行李与货物运输规范主题、检查与审核规范主题与应急处理规范主题。

管理规范主题以约束机场的建设及运营管理为主，阐述了民航机场应该以何种方式建设机场各项设施、以何种方式安排进站安全检查等内容。其主要条文包括《民用航空安全检查规则》第一至七章、《民用机场建设管理规定》第一至八章、《中华人民共和国民用航空法》第六章等共计 28 项内容。

行李与货物运输规范主题收纳了所有对旅客行李和航空货物所作出的要求，以加强航空货物运输的管理、加强对旅客及行李国际航空运输的管理为主，起到保护承运人和旅客的合法权益的作用。其内容包括《中国民用航空货物国内运输规则》第一至七章、《中国民用航空旅客、行李国际运输规则》第一至十二章、《中华人民共和国民用航空法》第九章等共计 31 项内容。

检查与审核规范主题在经过计算后再次被划分，进而形成两个第三层主题，分别为行政检查规章主题和审查与许可主题。行政检查规章主题包含了各类繁杂项目的规章要求与检查要求，介绍了行政检查工作、如何开展行政检查工作、如何开展航空卫生工作、监察员和各类监察机构的责任和权限等非常具体的内容。这个主题包含了《民用航空安全信息管理规定》第四章、《民用航空行政检查工作规则》第一至四章、《中国民用航空航空卫生工作规则》第四至九章等共计 35 项内容。审查与许可主题则收纳了所有与经营许可、使用许可、设备合格审定相关的内容，其中的各项法律条文主要阐述了公共航空运输企业需要获得的经营许可、民用航空产品和零部件的合格审定程序和管理要求及航空器适航管理等内容，具体包括《公共航空运输企业经营许可规定》第一至五章、《民用航空产品和零部件合格审定规定》第一至十一章、《民用机场使用许可规定》第一至四章等共计 26 项内容。

应急处理规范主题主要讲述了如何处理各类突发事件，起到预防、监测与消除突发事件的作用。主要内容包括《民用运输机场突发事件应急救援管理规则》第一至七章、《中国民用航空应急管理规定》第一至六章等共计 14 项内容。

3.3.5 气象相关

在第五类气象相关主题内包含了对气象等环境探测工作的指导和规定，给出了部分监视与管理办法，对例如气象台、气象站等环境探测机构作出了具体工作要求。其中包括《中国民用航空气象工作规则》第一至十二章内容、《民用航空气象人员执照管理规则》第一及第三章内容、《民用航空气象探测环境管理办法》第一及第四章内容等共计 31 项内容。此类相关信息均属于第五类“气象相关”主题。

3.3.6 交通管理

第六类交通管理主题内包含空中及地面交通规则两部分，详细阐述了对空域、跑道、飞行器等的使用和管理规定，给出了交通管制等工作的具体内容和要求。改进 GHSOM 算法在对本类主题的计算中将文献进一步划分为两个次级主题，分别为航空器登记与管理主题和交通规则主题。

航空器登记与管理主题以航空器及加强对民用航空器国籍、机载设备等的管理为主，维护民用航空活动秩序。其主要内容包括《民用航空器国籍登记规定》第一至第五章、《中华人民共和国民用航空法》第二至第七章、《中华人民共和国民用航空法》第十一至第十四章等共计 15 项内容。交通规则主题中的内容则是以规范民用航空活动、保证航空器运行的安全和效率、充分开发和合理使用空域资源为主，对跑道、空域、飞行轨道等交通规则作出规定。其主要内容包括《民用航空使用空域办法》第一至四章、《民用航空使用空域办法》第六章、《中国民用航空空中交通管制工作规则》第一至七章等共计 30 项内容。

3.3.7 情报工作

第七类情报工作主题中对各类情报资料的有效性、各类情报设备的正确使用等作出了具体要求，为各级情报机构工作的开展制定了相应工作规范。该主题衍生出两个次级主题，从两个次级主题所包含的内容来看，较难为这两个主题分配准确的、能够完美表达其内容的主题词。其中，第一个主题主要偏向对情报相关工作的本质做基础性定义及描述，阐述了什么是情报资料、哪些属于情报资料、情报服务的内容、情报人员应该具备的知识和技能等基础性信息，大体可以将其主题命名为基础信息与术语释义。其内容包含《民用航空情报工作规则》第一至九章、《民用机场运行安全管理规定》第十一章等共计 20 项内容。第二个主题主要偏向于对情报机构和各类信息设备进行描述，阐述了情报机构的管理和组织方式，提供了情报机构应当制定的安保方案，规定了民用航空通信、导航、雷达、无线电台站的设置方式等，因此，可以将该主题命名为设施与机构管理。其内容包含《中国民用航空航空卫生工作规则》第一章、《中国民用航空无线电管理规定》第二章、《中国民用航空无线电管理规定》第四章等共计 11 项内容。

3.3.8 违规处理

第八类违规处理主题主要包含了对其他八类主题中各项违规行为的认定及处罚标准，以行政处罚与经济处罚两种方式为主，并且对部分违规行为给出了具体的罚款金额。具体违规行为可以精确到违反了哪一章哪一节甚至哪一条法律规定，并对相应违规违法行为给出不同等级的处罚。其主要内容包括《导航设备开放与运行管理规定》第五章、《公共航空运输企业经营许可规定》第六章、《空中交通管理运行单位安全管理规则》第七章、《民用航空安全管理规定》第五章等共计 34 项内容。例如，在《民用航空情报工作规则》第十章中规定：民用航空情报服务机构违反本规则第十条规定，未按本规则要求建立航空情报质量管理制度的，地区管理局应当责令限期改正，并给予警告。此类文本信息均被归纳至第八类。

3.3.9 执照管理与审核

在第九类执照管理与审核主题内，包含了获取各类执照的学历、年龄、身高等基本要求和必须经过的测试，规定了各类执照的有效期限和挂失、补办等流程。其中包括《航空安全

员合格审定规则》第一至五章内容、《民用航空器领航员飞行机械员飞行通信员合格审定规则》第一至六章内容、《民用航空情报员执照管理规则》第一至二章内容、《民用航空情报员执照管理规则》第四章内容等共计 23 项内容。

3.4 结果评估

在机器学习(ML)、自然语言处理(NLP)、信息检索(IR)等领域,评估(Evaluation)是一个必要的工作,其评价指标有:精确率(Precision, P)、召回率(Recall, R)和 F1-Measure(F)[15]。精确率的定义为分类正确的样本数占样本总数的比例。召回率表示正确被检索的项目占所有应该检索到的与其所属同一类的项目的比例。F1-measure 为综合评价指标,其认为精确率和召回率的权重是一样的,但有些场景下,我们可能认为精确率会更加重要。调整参数,使用 F 值作为评价指标可以帮助我们更好地对结果进行分析和评价。

3.4.1 知识地图分类的结果评估

根据精确率、召回率和 F 值的定义,对基于改进前后的 GHSOM 算法的民航法规知识地图构建效果进行评估,得到的评估结果如表 1—表 3 所列。

表 1 改进算法后知识地图第一层评估结果

Table 1 Evaluation results of the first layer of knowledge map after improved algorithm

(单位:%)

主题名称	精确率	召回率	F 值
安全相关	100.00	71.43	83.33
导航与通信	96.88	91.18	93.94
技能培训	100.00	85.71	92.31
行政检查与规范	94.74	89.26	91.92
气象相关	91.18	88.57	89.86
交通管理	97.83	83.33	90.00
情报工作	96.88	79.48	87.32
违规处理	100.00	85.00	91.89
执照管理与审核	92.00	88.46	90.20

表 2 改进算法后知识地图第二层评估结果

Table 2 Evaluation results of the second layer of knowledge map after improved

(单位:%)

主题名称	精确率	召回率	F 值
导航、监视设备及其管理	94.74	90.00	92.31
校验及监督工作	100.00	92.86	96.30
管理规定	96.55	93.33	94.92
行李与货物运输规范	100.00	96.88	98.42
检查与审核规范	93.85	87.14	90.37
应急处理规范	93.33	93.33	93.33
航空器登记与管理	100.00	88.24	93.75
交通规则	96.77	81.08	88.23
基础信息与术语释义	95.24	74.07	83.33
设施与机构管理	100.00	91.67	95.65

表 3 改进算法后知识地图第三层评估结果

Table 3 Evaluation results of the third layer of knowledge map after improved algorithm

(单位:%)

主题名称	精确率	召回率	F 值
审查与许可	96.30	89.66	92.86
行政检查规章	92.11	85.37	88.61

由计算结果可知,该算法具有较高的分类精确率,每一类主题的精确率普遍高于 90%。相同或相近的法规章节被划分至同一主题中,比较容易为每个类别选定主题词。召回率和 F 值的总体表现也较好,只是在情报工作主题内的表现稍差,这也是由于初始法规集在分解时只精确到章导致的,由于法律文件的特殊特点,每个章节内可能同时包含与安全、情报等多个主题相关的内容,主题的不确定性也较容易导致其在分类过程中的错误。在后续工作中可进一步将其分解至每一小节,从而进一步提高分类效果。

3.4.2 算法改进效果评估

采用传统 GHSOM 算法构建的民航法规知识地图分类结果将民航法规集分为九大类主题,每类的主题词提取结果与目前分类主题有微小偏差,但仍可采用与目前相同的主题词。改进后的算法在精确度上有了更进一步的提升。由于改进后的算法在 SOM 模型迭代过程中动态改变了级内终止条件参数 τ_1,并且在选择 GHSOM 子网络上采用了赤迟信息量准则作为选择判据,知识地图的分类结果的被进一步细化。

主观上,提升的效果体现在主题词的选择上,改进的算法聚类后的结果更容易选定主题词,而改进前每类主题内包含了更多无关的关键词和内容,主题词的选择也更加困难。客观上,提升的效果体现在精确率和召回率上。改进前知识地图第一层的相关数据如表 4 所列。

表 4 改进算法前知识地图第一层评估结果

Table 4 Evaluation results of the first layer of the knowledge map before the improved algorithm

(单位:%)

主题名称	精确率	召回率	F 值
安全相关	98.56	68.95	81.14
导航与通信	72.93	81.18	76.83
技能培训	99.27	83.24	90.55
行政检查与规范	83.19	79.26	81.18
气象相关	92.85	85.54	89.05
交通管理	93.95	88.37	91.07
情报工作	84.81	68.74	75.93
违规处理	100.00	85.00	91.89
执照管理与审核	93.19	82.52	87.53

由于改进前后,安全相关、技能培训、气象相关和违规处理这 4 个主题没有较大幅度的波动,聚类结果比较稳定,因此可将这 4 个主题在算法改进前后的评估结果作对比,如图 6—图 8 所示。

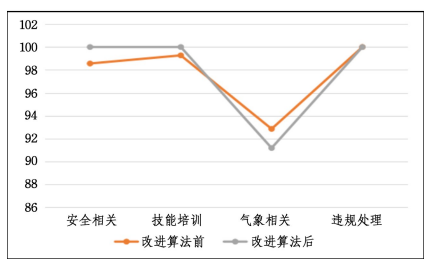

图 6 算法改进前后的 P 值对比

Fig. 6 Comparison of P values before and after algorithm improvement

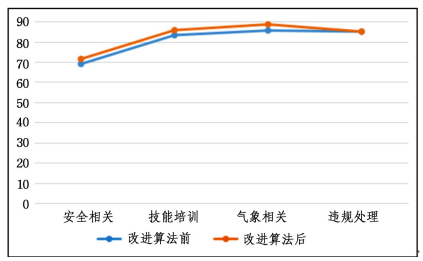

图 7 算法改进前后的 R 值对比

Fig. 7 Comparison of R values before and after algorithm improvement

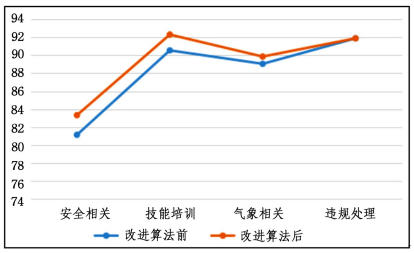

图 8 算法改进前后的 F 值对比

Fig. 8 Comparison of F values before and after algorithm improvement

图 6—图 8 分别表示了算法改进前后这 4 类主题的精确率、召回率和 F 值。可见，改进后算法的精确率和召回率与改进前相比有所提升，主要原因在于聚类结果更精细且主题划分更明确。改进前算法聚类后分为 8 类，这就导致了许多原本应该分为不同类别的法规文献被分在了同一类，进一步导致了类别主题名称的改变。而包含了更多主题词的类别更容易与其他类别中的法规文献产生相似性，拥有更多相似性的法规文献在被分至不同类别中时也导致了召回率的降低。

总体而言，改进的 GHSOM 算法在文本聚类过程中的表现已经十分可观，而改进后的 GHSOM 算法则具有更高的准确度。同时，本文只探讨了利用改进的 GHSOM 算法构建知识地图的可行性，并没有对该算法不同初始参数的值对评估结果的影响做进一步分析。在后续工作中，可尝试对初始参数的值进行不断调整，以探讨其在不同情况下对精确率、召回率及 F 值的影响。

结束语 本研究希望借由改进的 GHSOM 算法来组织并建立民航法规知识地图，通过组织大量法规的内部逻辑内容，为管理人员提供尽可能全面、综合、易查找的知识线索，方便快速寻找知识内容。在对民航法规集的主题分类上收获了较为可观的效果，但仍有其限制和不足。首先，由于民航领域专用语料库及领域专业词表的缺失，在对分词结果进行去停用词处理时，往往得不到较好的处理结果。因此，仅仅靠 TF-IDF 或某一种关键词提取算法来提取关键词是远远不够的，必须要建立民航法规专用词表，以此来使重要的关键词获得更高的权值。此外，虽然该技术提供地图上每个群集数个具有代表性的标签，但是为了提高知识地图的可读性，我们仍须人工选定主题名称。在未来的研究中，可寻找其他办法以达到自动标定主题名称的效果。最后，对于算法中 τ_1 与 τ_2 值的确定，仍然需要更精确的计算方法和流程。

就建立的民航航空法规知识地图而言，未来亦可加入法律检索功能，使用者在搜索某一关键词或新词条时，系统可以重新运行聚类算法，将其投射到某一主题分类下的次级主题中，以便快速检索对应的法律条文，亦可帮助发现和解决法律适用性问题。

参 考 文 献

[1] CHEN Q, LIAO K J, XI J Q. Research status and prospect of knowledge map[J]. Intelligence magazine, 2006, 25(5): 43-46.

[2] DAVENPORT T, PRUSAK L. Working knowledge : how organizations manage what they know[M]. Boston: Harvard Business School Press, 2008.

[3] EPPLER M J. Toward a pragmatic taxonomy of knowledge map-s: classification principles, sample typologies and applicationexamples[C]// Tenth International Conference on InformationVisualization. 2014.

[4] YE F B, TANG R Z. Knowledge Matching Method of Product Design Process Based on Knowledge Map[J]. Journal of Zhejiang University: Engineering Science, 2008, 42(6): 927-932.

[5] PAN C, WANG J, LIU L. A Knowledge Map Model Based on Concept Clustering[J]. Systems Engineering-Theory & Practice, 2007(2): 126-132.

[6] LI M, LU X Z, CHEN L S, et al. Knowledge map construction for question and answer archives[J]. Expert Systems With Applications, 2020, 141: 112923.

[7] CHA W W. Disciplinary knowledge representation and quantification method of multiple independent parameters in knowledge maps [D]. Wuhan: Central China Normal University, 2017.

[8] FAN Y X. Design and implementation of personalized learning system based on knowledge map [D]. Changsha: Hunan Normal University, 2018.

[9] ZHAO T, YU L, ZHAO Q. Research on Digital Education Resource Service Model Based on Knowledge Map [J]. Journal of Southwest China Normal University (Natural Science Edition), 2019, 44(11): 136-141.

[10] BU W X. Based on improved GHSOM intrusion detection technology [D]. Tianjin: Tianjin University, 2016.

[11] CHEN L. Application of Improved GHSOM Algorithm in Text Clustering[J]. Computer and Telecom, 2016(5): 57-61.

[12] TIAN W F. A Method of Feature Selection Based on Word2Vec in Text Categorization[OL]. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CPFD&dbname=CPFDLAST2018&filename=KZLL201807006157&v=MTc5MjNIWXJHNEg5bk1xSTlGWXVvS0N4Tkt1aGRobmo5OFRuanFxeGRFZU1PVUtyaWZaZVp2RUNubFU3Zk5KbG9VTGpm.

[13] XIE Z L, LI N, ZHOU C J. Research on Emotion Classification of Hotel Reviews Based on Word2vec[J]. Journal of Beijing Union University, 2018, 32(4): 34-39.

[14] ZHAO Z B, SHI Y X, LI B Y. Newly-emerging Domain Word Detection Method Based on Syntactic Analysis and Term Vector [J]. Computer Science, 2019, 46(6): 29-34.

[15] LIU W J, LUO J X. Image Retrieval Based on Improved GHSOM Clustering Algorithm[J]. Journal of East China University of Science and Technology(Natural Science), 2015, 41(2): 216-221.

ZHANG Hao-yang, born in 1994, postgraduate. His main research interests include natural language processing and so on.