学号 2017302590132

密级

武汉大学本科毕业论文

顾及制图特征和语义属性的 WMS 图层检索意图识别

院（系）名 称：遥感信息工程学院

专 业 名 称 ：遥感科学与技术

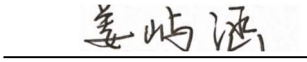
学 生 姓 名 ：姜屿涵

指 导 教 师 ：桂志鹏 副教授

二〇二一年五月

郑 重 声 明

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名：  日期：2021/4/27

摘　要

随着地球科学系统数据采集与共享技术日益完善，互联网上各式地图大量涌现，为地学研究和应用提供丰富数据资源。但从海量数据中搜索特定数据如同“大海捞针”，准确高效的资源检索成为了实现数据价值的关键。而地图数据丰富的视觉特征、文本属性，用户检索意图的模糊性、多样性等，给检索意图的提取提出了巨大的挑战，如何建立综合地图多模态信息的检索意图识别方法，具有很大的意义。为此，本文顾及地图的制图方法和语义属性，提出一种基于 DBSCAN 聚类算法的地图检索意图提取模型。具体研究内容如下：

（1）基于意图维度空间建立意图的形式化表达模型。定义意图的维度空间，实现对子意图的表达，由存在联系的子意图构成复合意图，实现意图的具象化。

（2）基于 DBSCAN 聚类实现地图检索意图识别。利用 SWEET 本体库的结构，构建语义相似度度量模型；其次通过构图完成复合意图的提取，并在复合意图的基础上，综合考虑聚类稳定性、正样本利用率和平均簇凝聚度确定聚类参数，使用 DBSCAN 算法提取子意图；最后使用 kruskal 算法生成子意图的最小生成树，利用维度空间完成子意图的表达。

（3）基于置信度实现迭代式相关反馈。构建包括复合意图质量、复合意图突出度、子意图维度权重、子意图突出度的置信度评价机制。借用“肘部法则”的思想，确定置信度阈值，完成迭代式相关反馈。

（4）基于典型检索场景实现模型的有效性分析。构建典型检索场景生成合理的样本集合，对模型提出的动态半径、复合意图综合评价指标、子意图维度表达和迭代式相关反馈的有效性进行验证和分析。

综上，本文基于聚类算法，建立了一种顾及制图方法和语义属性的地图多模态信息的检索方法，以意图识别的形式实现地图精准检索。为地图检索意图的形式化表达和提取方法提供一种新思路，为实现海量地图数据的有效利用提供强有力的支撑。

关键词：地图检索意图；意图形式化表达；DBSCAN 聚类；迭代式相关反馈；置信度评价机制

ABSTRACT

As the technology of data collection and sharing in the earth science system is becoming more and more perfect, a large number of maps have emerged on the Internet, providing rich data resources for the research and application of geosciences. However, searching for specific data from massive data is like "finding a needle in a haystack." Accurate and efficient resource retrieval has become the key to realizing the value of data. The rich visual features and text attributes of map data, and the multi-dimensionality, abstraction, and ambiguity of user retrieval intentions pose a huge challenge to the extraction of retrieval intentions. How to build a comprehensive map multi-modal information retrieval intention recognition method is of great significance. To this end, this article taking into account the mapping method and semantic properties of the map, proposes a map retrieval intention extraction model based on the DBSCAN clustering algorithm. The specific research content is as follows:

(1) Establish a formal expression model of intention based on the dimensional space of intention. Define the dimensional space of intentions composed of tags, realize the expression of sub-intentions by the dimensional space of intentions, and form composite intentions by the sub-intents with connections.

(2) Realize map retrieval intention recognition based on DBSCAN clustering. Using the structure of the SWEET ontology library, construct a semantic similarity model to achieve the measurement of semantic distance. Secondly, complete the extraction of compound intentions through composition. On the basis of compound intentions, after comprehensively considering clustering stability, positive sample utilization and average cluster aggregation to determine the clustering parameters, we use DBSCAN algorithm to extract sub-intents. Finally, the kruskal algorithm is used to generate the minimum spanning tree of the sub-intent, and the dimensional space is used to complete the expression of the sub-intent.

(3) Realize iterative correlation feedback based on confidence. Construct a confidence evaluation mechanism including the quality of the composite intention, the prominence of the composite intention, the weight of the sub-intent dimensions, and the

prominence of the sub-intent. Borrow the idea of "elbow rule" to determine the confidence threshold and complete iterative correlation feedback.

(4) Analysis of the effectiveness of the model based on typical retrieval scenarios. Construct a generation method of typical retrieval scenarios to generate a reasonable sample set. Verify and analyze the validity of the dynamic radius, composite intention comprehensive evaluation index, sub-intention dimension expression and iterative related feedback proposed by the model.

In summary, based on the clustering algorithm, this paper establishes a map multi-modal information retrieval method that takes into account mapping methods and semantic attributes, and achieves accurate map retrieval in the form of intention recognition. It provides a new idea for the formal expression and extraction method of map retrieval intent, and provides strong support for the effective use of massive map data.

Keywords: Map retrieval intention; Formal expression of intention; DBSCAN clustering; Iterative correlation feedback; Confidence evaluation mechanism

目　录

1 绪论

1.1 研究背景与意义

地图是空间数据的重要载体和表现形式，具有广泛的应用场景。随着地球科学系统数据采集与共享技术日益完善，互联网上各式地图大量涌现，为地学研究和应用提供丰富数据资源的同时，也给准确高效的资源检索带来巨大挑战[1]。由于地图丰富的视觉及文本特征，用户地图检索意图的抽象性及模糊性，如何建立综合地图多模态信息的检索意图识别方法，成为实现地图精准检索的关键。

由开放地理空间联盟(OGC)制定的网络地图服务(WMS)，可通过提供接口实现地理数据的实时绘制等操作，被广泛运用于各学科和工业领域[2]。据报告[3]，来自全球各领域的 WMS 数据量已超过 4 万个，而一个 WMS 地图由多个图层构成，WMS 图层的总数目已超 30 万张，覆盖地质、农业、能源等多个主题，从海量数据中搜寻特定数据如同“大海捞针”，如何辅助用户实现快速准确的地理资源检索成为了研究热点方向。

搜索引擎是帮助用户快速获取信息的重要工具，但由于用户输入的查询语句较短，对一个用户而言，一个查询可能包含多种意图，对不同用户而言，一个相同的查询可能表达不同的意图[4]。总体而言，精准检索给用户和检索系统带来了两个重要的挑战[5]：从用户的角度，挑战在于如何准确地表达自己的需求；从系统的角度，挑战在于如何根据用户提出的查询来了解他们想要什么。建立意图的形式化表达机制，既是实现用户检索意图提取的基础，能够有效的描述用户的意图偏好；也是实现多模态检索的基础，将用户多样的意图表达归纳为统一的形式，成为系统与用户之间沟通的通用语言。而相关反馈则是“双向交流”的过程，通过系统与用户的反复沟通确认，明确检索意图，降低噪声等不稳定因素的影响，实现个性化检索，使系统更加准确的捕捉到用户的检索意图[6]。

相较于自然图像与遥感影像，地图图像具有强烈的制图学、地理学和社会人文学科特性。地理要素的符号化表达和制图综合使得地图内容高度抽象化，进一步导致地图的尺度依赖性，用户对于地图的检索需求也更加综合复杂[7]。“语义鸿沟”和“意图鸿沟”进一步扩大，地图资源的精准检索需要兼顾图层的图像数据和元数据文本所蕴含的视觉特征和文本特征，并融合反馈机制挖掘用户的检索意图。对视

觉特征而言，如何针对地图的特性设计图像特征以实现对地图关键信息的全面描述至关重要。对文本特征而言，地图具有语义、空间、时间等多方面信息，相对于时间和空间这类可量化的信息而言，如何检索语义意图是实现用户地图检索意图识别的关键。

本研究将以 WMS 图层数据作为检索对象，结合制图特征和语义属性，挖掘地图语义上多模特征的共性与组合关系，建立意图及偏好的描述机制，结合相关反馈，构建识别用户意图的方法，实现地图的精准检索。为地图检索意图的形式化表达和提取方法提供一种新思路，为实现海量地图数据的有效利用提供强有力的支撑。

1.2 国内外研究现状

1.2.1 地图图像的检索

地图图像的检索大致分为基于语义的检索（SBIR）、基于内容的检索（CBIR）和基于用户相关反馈的检索。

基于语义的检索是利用地图元数据完成检索[8]。用户通过输入文本来表达检索需求，使用语义工具比较用户输入文本与候选集中图像的属性文本的相似性实现检索，地学本体库是最常用的语义工具之一[9]。该方法相较于基于关键词匹配检索方法，解决了语义歧义的问题，但对用户和地图元数据的要求较高，一方面要求用户具有较高的专业素养，能够用专业语言准确表达检索需求，另一方面要求地图元数据具有较高的质量。

基于内容的检索是由用户输入一个与检索需求相对应的可视化查询，例如地图图像，通过对比用户输入图像集与候选集中图像的低层次特征，如颜色、纹理等的相似性来实现[10]。随着卷积神经网络的重大突破，通过对图像高层特征的建模，并提取深度全连接特征作为图像表示向量，CBIR 的精度得到了显著的提高[11,12]。该方法较好的解决了用户意图表达不准确的问题，改善地图元数据缺失或错误的问题，但地图图像具有内容抽象概括、形式丰富多样等特点[13]，以图识图的结果差强人意，容易导致视觉特征“过拟合”与“欠匹配”，最终“只识其形未得其神”，低层次图像特征与高层语义之间始终存在较大的“语义鸿沟”。

此外还有一种结合 CBIR 和 SBIR 的方法，该方法要求用户以 CBIR 的方式输入检索序列，然后利用与输入图像关联的属性文本而非低层图像特征来实现检索[14]，这一方法试图克服前两类的缺点，避免用户语言表达的不准确，让图像自己说

话，同时考虑视觉特征和文本特征，有助于消除底层图像特征与地图高层语义之间的“语义鸿沟”，避免误检漏检现象，改善地理信息检索质量。但该方法要求用户拥有能表达意图的图像素材，与现实检索场景不完全一致；另外，大量的 WMS 还存在元数据缺失或图文不符的情况，使得该方法依然存在较大的局限性。

基于用户相关反馈的方法，通过记录用户的鼠标标记、轨迹追踪等相关反馈数据，通过分类器训练，获得检索结果呈现给用户后，再次捕捉用户的相关反馈数据，直到用户获得理想的检索结果。该方法通过引入用户相关反馈的机制[15]，降低用户检索意图表达的成本，但缺少对显示意图的表达，用户难以判断系统是否正确的理解其意图，无法引导用户指明意图，容易导致多次反馈未果，反而增大了用户检索操作的负担。

1.2.2 检索意图的识别与形式化表达

检索意图识别，本质上是将用户输入的检索序列具象化，“翻译”为系统可理解的对象，即意图。目前检索意图识别直接相关的研究较少，但检索意图识别与可解释性推荐具有较大的相关性，意图可以作为可解释性推荐中对推荐结果的解释，意图识别是实现个性化推荐的一种有效策略，因此检索意图识别是可解释性推荐的一种尝试，也是实现个性化推荐的一种方式。

个性化推荐算法目前可以大致分为基于内容的推荐[16]、基于协同过滤的推荐[17]和混合推荐[18]。基于内容的推荐是通过搜集用户的背景信息[19]，如性别、年龄、偏好等来构建用户画像，同时根据物品的特征信息，如种类、口味、功能等构建物品画像，两个画像通过匹配完成推荐，但是用户画像和物品画像的构建都需要大量的时间，且无法保证画像的准确性。基于协同过滤的算法应用最为广泛，其原理是有相似特征的用户可能具有相近的喜好，根据模型的作用对象，又可以分为基于用户、物品和模型的协同过滤，其中前两者需要用户相关的数据，后者包括 PLSA 模型、LDA 主题模型和聚类算法[20]等。但是由于用户在开始使用系统时，很少甚至没有历史行为记录，该算法最大的问题之一就是冷启动问题[21]。

可解释性推荐的目标是在给用户展示推荐结果的同时给出推荐理由，使得推荐结果更具有说服力。目前已有的系统从多种角度给出了推荐理由[22]，包括以用户或商品为推荐理由的，如找相似商品；以商品特征作为推荐理由的，即展示商品的某项特征可能符合用户的检索需求；以对商品的意见作为推荐理由的，即结合评

论数据，给用户展示评论中的相关意见；文本解释，直接以文本语句解释推荐理由；可视化解释[23]，提供分析图表等可视化的推荐理由。而可解释性推荐中使用的具体方法也多种多样，包括矩阵分解、知识图、主题建模、深度学习等。

在检索意图的表达方面，现有的检索系统有使用如兴趣树、画像等形式。阿里曾提出一类基于兴趣树的深度学习推荐模型[24]，通过利用从粗到精的方式从上到下检索兴趣树的节点为用户生成推荐候选集，该方法适合从海量商品中进行快速检索，是一种较为高效的方法。基于用户画像的方式[25]，是以标签的形式将用户的偏好、个人信息、习惯等表现出来，再结合用户的性别、年龄等基础信息计算出用户的兴趣模型，即用户的偏好商品列表，再将兴趣模型标签化，这些标签即为检索意图，该方式主要根据用户的背景和历史信息完成意图的表达，缺乏灵活性，并且在地图检索系统中，更多考虑的是用户的专业学术背景和当下的研究热点，个人信息如年龄可能对于地图检索的帮助相对较小。

本研究拟选取对用户地图检索有参考意义的地图视觉特征和语义特征作为意图提取的目标，设计意图的形式化表达机制来实现对意图的解析，同时采用显示反馈的机制引导用户进行相关反馈。

1.3 研究内容与技术路线

1.3.1 研究目标

本研究旨在实现用户检索意图的形式化表达，并结合迭代式相关反馈，设计能够有效提取用户检索意图的模型，在有效平衡用户友好性和反馈准确性的基础上，完整且精确的概况地图信息和用户检索意图，解决检索意图多维性、多样性、抽象性、模糊性、多模性的问题。

1.3.2 研究内容

本研究主要包括意图的形式化表达机制、用户检索意图识别模型和迭代式相关反馈机制，如图 1.1 所示。

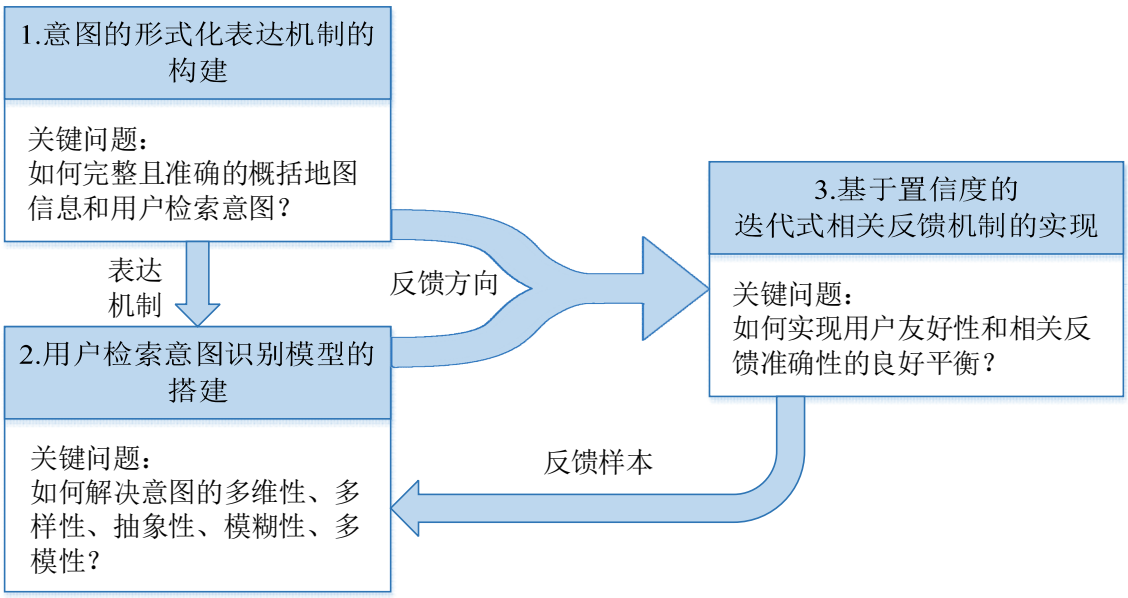


图 1.1 研究内容及其之间的关系与关键问题

（1）意图的形式化表达机制的构建：解决地图信息具有抽象性、复杂综合性和尺度依赖性等问题，形成意图的描述机制，建立系统与用户之间交流的语言。

（2）用户检索意图识别模型的搭建：结合制图特征与语义属性实现适用于多种典型场景的地图检索意图的识别方法，利用形式化表达机制对意图进行解析，实现意图对用户的可见性和可理解性。并在不同场景下探究模型对意图多维性、多样性、抽象性、模糊性等问题的适用性。

（3）迭代式相关反馈机制的实现：构建基于置信度的迭代式相关反馈机制，在保证检索意图提取精度的同时，最大可能的降低用户检索操作的负担。

1.3.3 技术路线

本研究的技术路线包括样本生成、意图的形式化表达模型、意图提取和相关反馈共四个步骤，如图 1.2 所示。

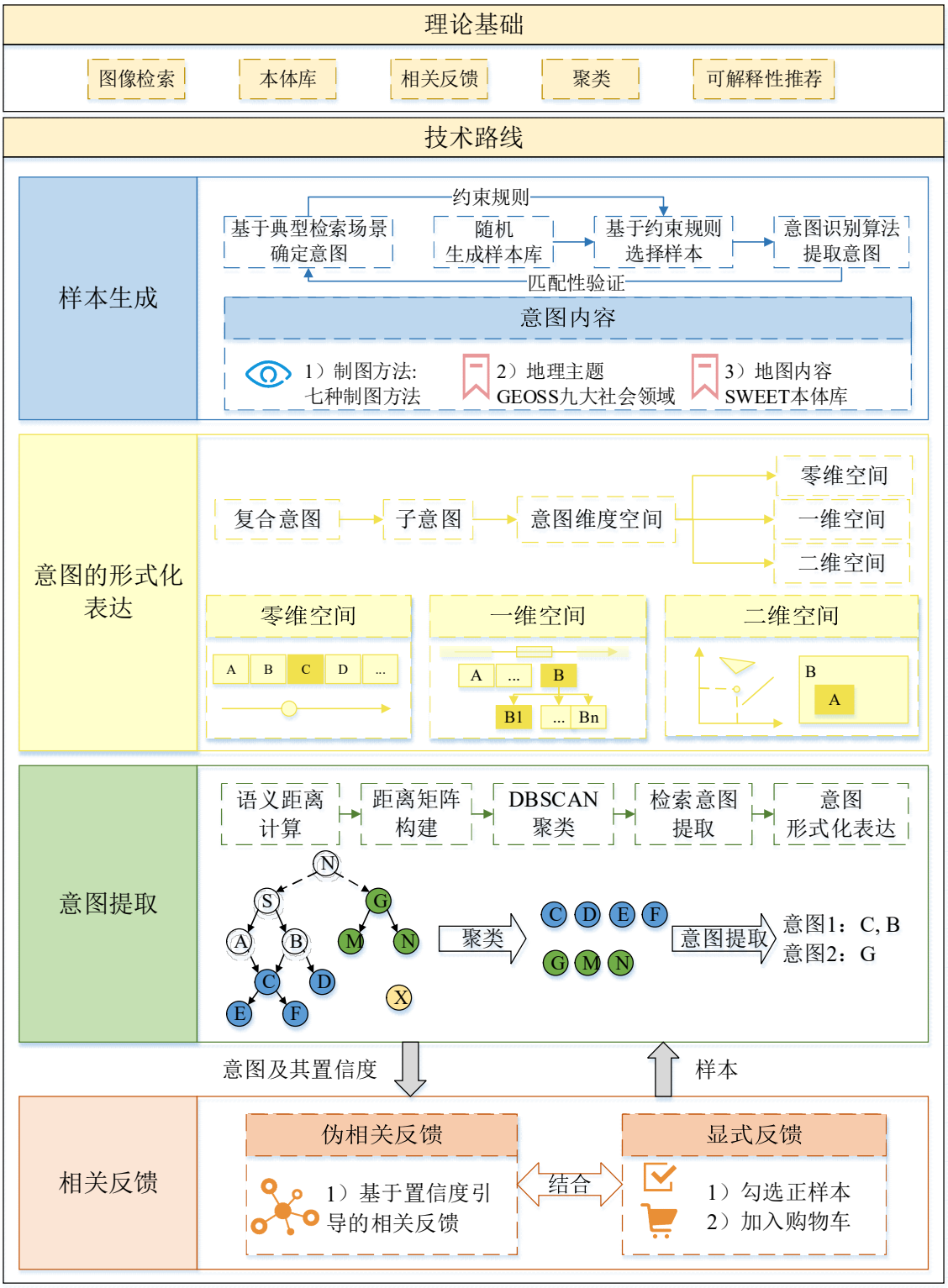


图 1.2 论文技术路线图

（1）样本生成：根据典型检索场景和样本生成的规则，保证有足够样本对应意图的前提下，生成具有制图方法、地理主题和地图内容标签的样本，用于后续的实验和验证。

（2）意图的形式化表达：本研究将用户的意图划分为复合意图和子意图，复

合意图由具有关联性的子意图组成，类比于真实空间，使用意图维度空间对子意图进行表达。

（3）意图提取：根据 SWEET 本体库的结构，实现语义距离的度量并生成距离矩阵，基于 DBSCAN 聚类实现意图提取，然后使用意图的形式化表达模型完成意图表达和置信度评估。

（4）相关反馈：基于置信度的引导，生成下一次相关反馈的样本集，由用户勾选正样本后，再次完成意图提取，直到用户满意。

1.4 论文章节安排

本文主要研究顾及制图特征和语义属性的 WMS 图层检索意图识别算法的设计与实现，共分为六章进行论述。

第一章 绪论。主要介绍地图检索意图识别的背景和意义，目前意图识别与形式化表达、地图图像检索的研究现状，本文研究的目标、内容和技术。

第二章 地图检索意图的形式化表达。首先从用户的角度总结地图检索意图的特点，明确形式化表达模型需要解决的问题；其次从地图的角度，构建意图的分类体系，以及本研究所关注的制图特征和语义属性的提取，确定检索意图的值域；最后基于空间的维度概念，构建意图的维度空间，形成意图的形式化表达机制。

第三章 基于 DBSCAN 聚类的地图检索意图提取。首先基于本体库建立语义相似度模型，度量样本间的语义相似度，生成距离矩阵；其次使用 DBSCAN 聚类，提取意图；最后建立意图的形式化表达机制，完成意图的表达。

第四章 基于置信度的迭代式相关反馈。首介绍相关反馈的特点，为本文的相关反馈方法奠定基础；其次从聚类效果和意图质量两个方面，提出聚类稳定度、簇凝聚度、样本利用率、意图突出度等多个评价指标，构建意图的置信度评价机制；最后根据置信度，引导生成样本集，完成下一次的相关反馈。

第五章 基于典型检索场景的实验分析。设计典型检索场景样本生成方法，对模型涉及的多个参数和指标进行有效性检验。

第六章 总结和展望。阐述顾及制图特征和语义属性的 WMS 图层检索意图识别算法的优势和不足，从聚类算法、置信度评价机制、匹配性验证、样本利用的角度对未来的工作进行展望。

2 地图检索意图的形式化表达

地图检索意图的形式化表达旨在从用户和地图两个角度出发，填补“语义鸿沟”。用户检索意图具有多维性、多样性、抽象性、模糊性和多模性的特点，通过建立意图的形式化表达机制，将意图具象化。地图信息具有抽象性、复杂综合性和尺度依赖性的特点，用户的检索意图囊括地图所有的信息，需要对杂糅的地图信息建立科学的分类体系，便于更好的整理和发现用户的意图。地图检索意图的形式化表达是准确的理解用户意图的基础，也是检索系统与用户沟通的语言。

2.1 地图检索意图的分类体系和特点

2.1.1 意图的分类体系

本文根据相关背景调查，建立了意图的分类体系，将检索意图分为 5 个方面：语义、空间、时间、制图和其它，如图 2.1 所示。分别用于描述用户检索地图时对于地图内容主题、地图空间范围与空间关系、地图相关的时间、地图制图与地图资源本身。

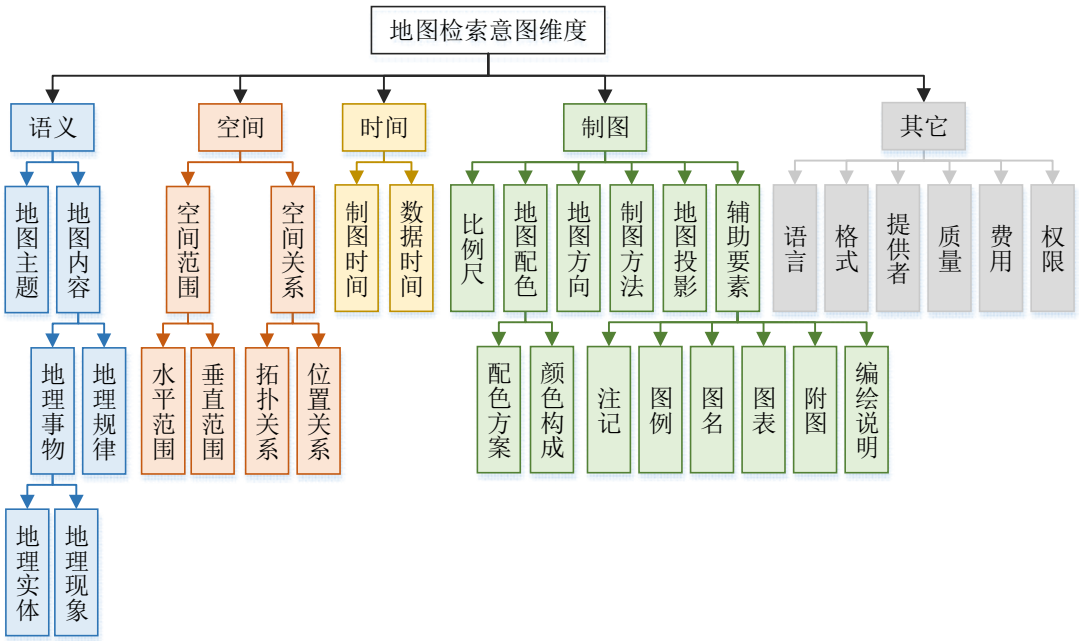


图 2.1 检索意图类别

（1）语义：由地图主题和地图内容构成。其中地图内容又分为地理事物和地理规律，地理事物是客观存在的，包括地理实体如：河流、湖泊，地理现象如：大气压、风速，地理规律是体现地理事物变化规律的地图，如描述气压沿赤道向两极变化的地图。

（2）空间：分为空间范围和空间关系。空间范围指地理对象所处的空间范围，水平范围可以使用如行政区划、边界矩阵经纬度的方式描述，垂直范围可以使用海拔范围、大气层划分的方式描述。空间关系指地理对象间的相互关系，拓扑关系包括邻接、相离等，位置关系包括方位、距离等。

（3）时间：分为制图时间和数据时间。制图时间一定程度上可以反应地图制作时的标准和质量，成为检索意图的标准之一。

（4）制图：由于地图制图的复杂性和综合性，该类意图较为杂糅，包括比例尺、地图配色、地图方向、制图方法、地图投影和辅助要素。在制图中，颜色具有较强的象征性，与地图语义形成对照，如蓝色代表海洋、绿色代表植物，深色代表密、多、强，浅色代表稀、少、弱，因此将颜色分为配色方案和颜色构成，配色方案侧重描述颜色种类如红、绿，颜色构成侧重描述色彩整体的特征如深浅构成、空间分布等。制图方法是制图中最具代表性的意图，整体概况了制图的手段、形式，决定成图的部分特点。

（5）其它：包括语言、格式、提供者、质量、费用、权限共 6 个叶子节点。

2.1.2 意图的特点

根据 WMS 图层的特点以及用户检索地图的特点，本研究总结了以下 5 个意图的特点。

（1）多维性：在意图的维度空间中，检索意图不止存在于单一维度，而是在每个维度均有不同权重的意图。

（2）多样性：如 2.1.1 节的介绍，用户检索意图的范围可以覆盖 5 个方面共 29 个叶子节点，每个意图可以由多个叶子节点组合而成。

（3）抽象性：用户的检索意图为高层语义，与地图信息并非完全对等，如中国北部土壤分布图，北部是对水平空间范围的抽象概括。

（4）模糊性：用户在开始检索时，若无法明确个人的检索意图，意图识别系统需要根据意图方向，对当前检索出的意图进行泛化或细化，引导用户具象化意图。

（5）多模性：地图具有丰富的视觉特征和文本特征，因此地图检索意图也呈多模特性。视觉特征包括形状、颜色等，文本特征包括内容、主题等。

根据意图分类体系和意图的特点，评估不同意图在使用频率、视觉相关性、抽象性、模糊性、值域范围四个方面的表现，如表 2.1 所示。通过背景调查，最终选取具有代表性的三者作为本研究的意图提取对象，分别是在视觉表现方面突出的

制图方法，在使用频率表现突出的地图主题，在抽象模糊性和值域范围上表现突出的地图内容。其中选取制图方法作为视觉特征的代表，还因为制图方法是地图区别于自然地图的一种独有的具有概括性的图像特征，一方面它能综合表达出地图的颜色、形状、制图目的等多方面的特征；另一方面，制图特征能够将视觉特征转化为文本语言，与文本特征相结合，实现多模检索的融合。

表 2.1 三类意图的评价（★~★★★）

指标	地图主题	地图内容	制图方法
使用频率	★★★	★★	★
视觉相关性	★	★	★★★
抽象模糊性	★★	★★★	★
值域范围	★	★★★	★

2.2 地图检索意图的取值和表达

2.2.1 意图的取值

获取地图标签值的过程分为两步，首先从 WMS 图层的图像和元数据提取出两大地图属性，分别是视觉特征和文本特征[26,27]；其次是从地图属性中抽象出地图标签[28]，如图 2.2 所示。

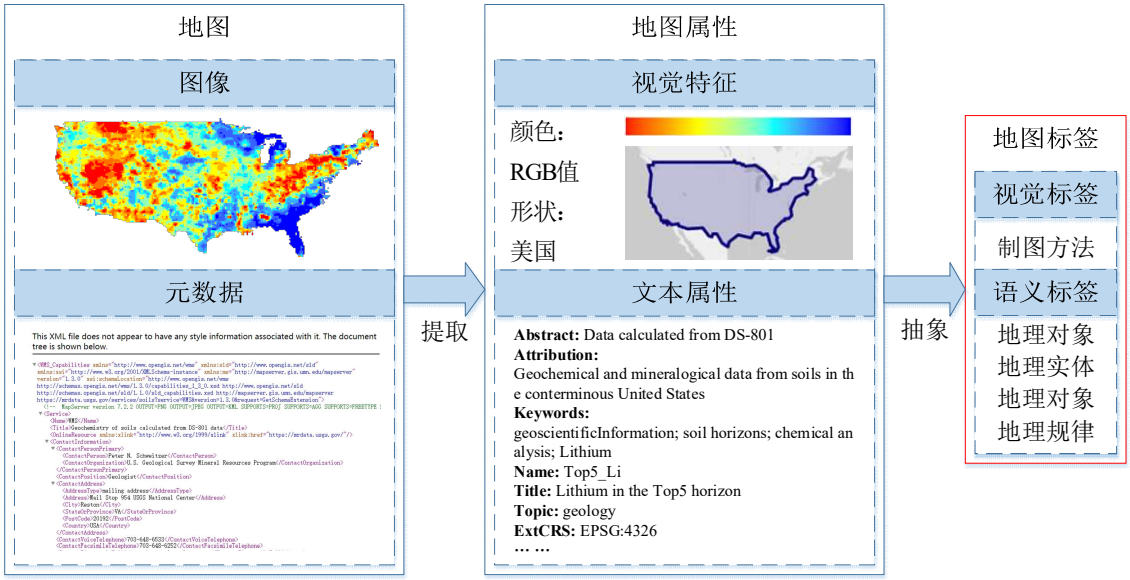


图 2.2 地图标签的获取

本研究的关注重点在于如何从已经有各类标签的地图中完成检索意图的提取，因此本研究在已经有各类标签的地图基础上，完成检索意图的提取，检索意图的值

域即为标签值。

制图方法共 7 个特征值，参考专题地图制图教材和使用频率确定。地理主题共 10 个特征值，为地质和 9 大社会效益领域，其中社会效益领域是全球系统中的地球观测系统项目围绕的九个环境领域，每个领域关注的重点主要是环境问题与人类活动和健康之间的关系。地图内容为地球与环境术语语义网（SWEET）中的所有节点，SWEET 本体库是由美国航天局地球办公室构建的，它通过网络平台实现信息的共享，是迄今为止规模最大的地球科学数据与术语研究项目，是实现地学知识的规范化表达和地学数据语义共享的基石[29,30]。三类标签的具体值域如表 2.2 所示。

表 2.2 地图标签的值域

标签	值域
制图方法	无、点状符号法、线状符号法、范围法、质底法、分级统计图法、其他
地理主题	地质、农业、生物多样性、气候、灾难、生态系统、能源、水、天气、健康
地图内容	SWEET 本体库中的类

如图 2.3 所示为本体库简单的图结构，根节点为顶层概念，顶层概念之间不存在语义关系且相互独立，如 Root1 节点和 Root2 节点。中间节点存在上下位节点，如节点 D 的上位节点为 A 表示 A 为 D 的父类，下位节点为 F 表示 F 是 D 的子类，上下位之间存在语义关系。由于制图方法与地理制图均为枚举值，值与值之间也相互独立，可视为本体库的顶层概念，因此在后文的模型中将制图方法与地图内容的取值均作为顶层节点处理。

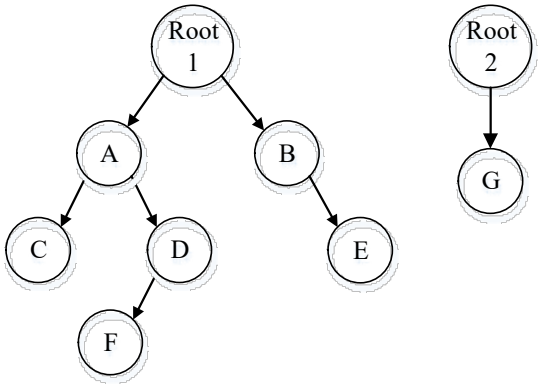


图 2.3 本体库结构

2.2.2 意图的表达

本研究建立了由四个部分组成的意图形式化表达机制，如图 2.4 所示。首先是提取标签值，即确定每个样本在地理主题、地图内容和制图方法上的取值；其次是由标签值形成意图的维度空间；第三是使用意图维度空间和维度权重向量完成子意图的表达；最后将具有关联性的子意图组合成复合意图，根据子意图间的关系，可以在复合意图内部进行语义推理和意图延展。

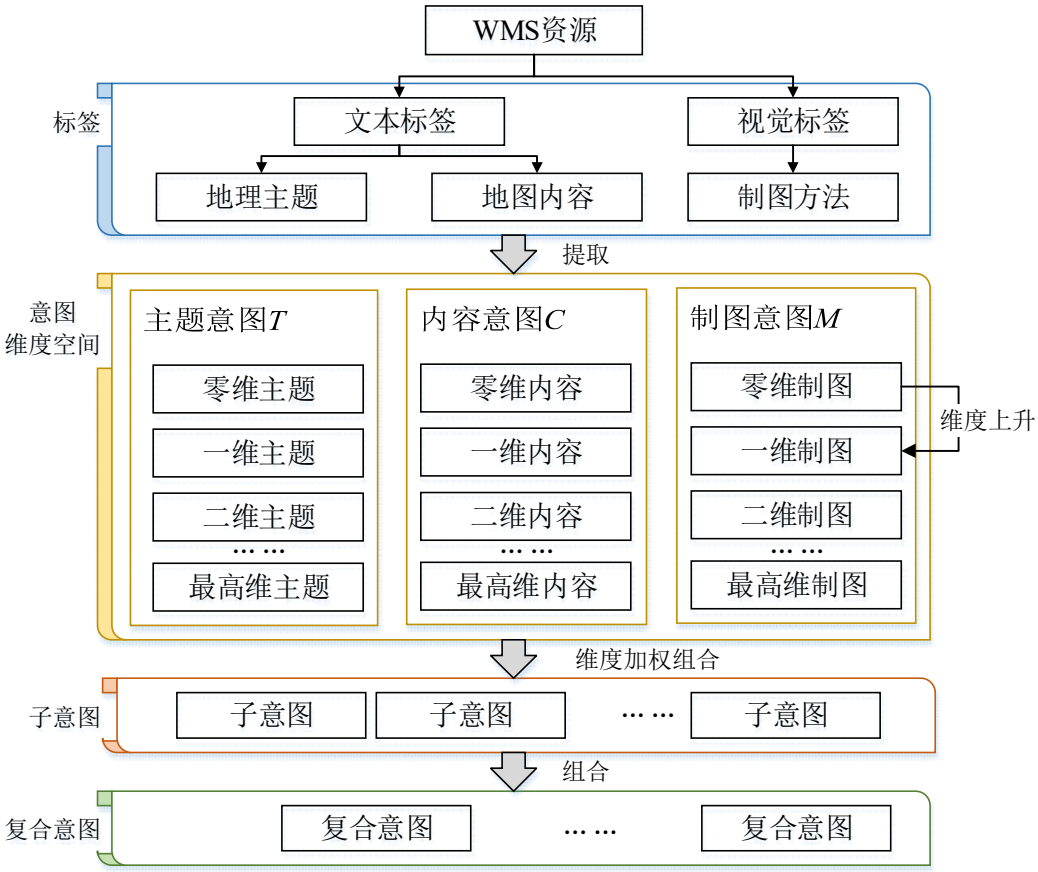


图 2.4 意图形式化表达

其中意图维度空间是类比于显示世界中对于空间的维度定义，由本研究创新性提出的。如公式（2.1）、（2.2）、（2.3）所示，主题意图 T、内容意图 C 和制图意图 M 分别形成 3 个意图维度空间。在每个维度空间中，以维度权重向量 P，量化用户在每个维度的意图需求，实现对意图的形式化表达，通过求解每一个意图维度空间的维度权重向量，识别出用户检索意图指向的维度，如公式（2.4）所示，同时满足 3 个维度空间的形态即可完整的表达子意图 I，如公式（2.5）所示，具有关联性的子意图组合即为复合意图 CI。

$$T = [T_1, T_2, \dots, T_n]^t \tag{2.1}$$

$$C = [C_1, C_2, \dots, C_n]^t \tag{2.2}$$

$$M = [M_1, M_2, \dots, M_n]^t \tag{2.3}$$

$$I = P_t T \cdot P_c C \cdot P_m M \tag{2.4}$$

$$CI = \sum_{k=1}^{n} I_k \tag{2.5}$$

图 2.5 为维度空间的示意图。目前本研究依据本体库的特点，构造了意图的零维空间、一维空间和二维空间，具体定义如下所述：

（1）零维空间

即为点，没有长度和方向，呈现最大聚集度的意图即为零维意图。标签值均为离散值，如图 2.5 中的零维空间，样本集的标签为 A、B、C、D，若用户勾选的正样本均有 C 标签，即在 C 上呈现最大聚集，则零维意图为 C。

（2）一维空间

即为线，有长度和单方向，呈现为取值范围的意图即为一维意图。标签值在数量上的意图为连续值，比如意图是制图方法超过 2 种类型的地图，标签值在取值范围上的意图为一维意图。地图内容的一维空间是由根据本体库生成的节点连接图构成，图结构中上层概念包含下层概念，顶层概念深度最小，底层概念深度最大。定义意图线为从某节点开始，单方向以深度优先的向下或向上寻找连接点终点，此时从起点到终点穿过的所有节点相连，即为意图线，节点数为意图线的长度。如图 2.5 中的一维空间，样本集的标签为 A、B、B1、Bn，若用户勾选的正样本标签为 B、B1，则地图内容取值上的一维意图是以 B 和 B1 为线段端点的“线段”或按深度增加向下层概念延伸的“线”或按深度降低向上层概念延伸的“线”。

（3）二维空间

即为面，有长度、多个方向和面积，呈现为取值面的意图即为二维意图，二维意图由多条意图线组成。如图 2.5 中的二维空间，样本集为 A、B、C、A1、An、B1、Bn，用户勾选正样本为 C、A、A1、B、B1，形成两条线，此时地图内容上的二维意图为以 C 为起点，向叶子节点以广度优先的延展。

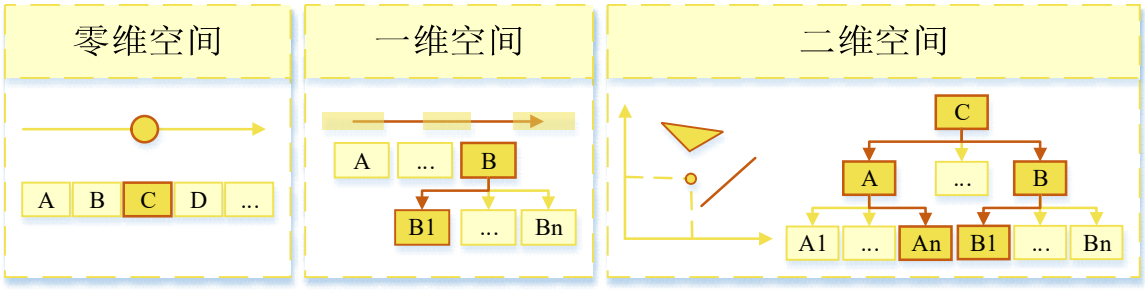


图 2.5 维度空间

2.3 本章小结

本章主要介绍了地图检索意图的分类体系和特点、地图检索意图的取值和表达。从语义、空间、时间、制图和其它共5个方面29个叶子节点，构建意图的分类体系，并根据意图的多维性、多样性、抽象性、模糊性和多模性的特点，以及在检索中的使用频率、视觉相关性、值域范围，最终确定以地图主题、地图内容、制图方法作为本研究的意图内容，参考相关资料确定3类意图内容的具体取值。最后类比于现实空间，提出意图维度空间，依次以标签、意图维度空间、子意图、复合意图，层层递进的建立意图的形式化表达机制，形成系统与用户之间交流的语言，实现对意图的具象化表达。

3 基于 DBSCAN 聚类的地图检索意图提取

地图内容由于抽象模糊性高、值域范围广，是意图提取上的重点和难点。本研究以 DBSCAN 为核心，首先基于 SWEET 本体库建立语义相似度模型，根据节点之间的距离矩阵，初步切分距离过大的节点集合，形成复合意图；其次在复合意图的基础上提出动态参数的 DBSCAN 聚类，以聚类稳定性、正样本利用率、平均簇凝聚度三个指标综合确定最佳聚类参数，完成子意图提取；最后利用意图的形式化表达机制对子意图进行解析，具象化用户意图。

3.1 模型概述

本文综合地图内容维度上不同取值之间存在复杂语义关系，设计一种基于聚类的提取方法，主要包括七个步骤，如图 3.1 所示。

（1）生成节点距离序列：每个样本的内容标签均对应本体库的一个节点，根据节点在本体库中的位置，单向向上遍历样本节点的上位节点和单向向下遍历样本节点的下位节点，按照一定的规则计算该样本节点到对应上下位节点的距离，得到每一个样本节点的（上/下位节点：距离）距离序列。

（2）寻找纽带节点：对比每个节点的距离序列，寻找节点与节点之间的最深共同祖先节点或最浅共同根节点，称这类连接起两个节点的共同节点为纽带节点。

（3）计算语义相似度：通过纽带节点，用最短距离建立节点之间的连接，若两个节点之间没有纽带节点，节点间的距离记为最大距离。

（4）构图提取复合意图：计算包括纽带节点在内，任意两节点之间的最短距离，建立大小为 $(C_{\text{样本节点数+纽带节点数}} \times C_{\text{样本节点数+纽带节点数}})$的对称阵，即距离矩阵。切断距离较大的节点间的连接，初步将节点分为不同节点组，每个节点组即为一组复合意图。

（5）DBCSCAN 聚类提取子意图：以复合意图的节点组为对象，确定半径 EPSILON 和阈值 MINPTS，使用 DBCSCAN 聚类提取子意图。

（6）kruskal 算法生成最小生成树：类比于人的连续性思维，本研究认为检索需求与需求之间的逻辑关系也是最短距离的连接关系，因此地图检索意图与意图之间的逻辑连接为最小生成树，使用 kruskal 算法生成最小生成树。

（7）分维度表达意图：使用第二章中的意图形式化表达机制对子意图完成子

意图的分维度表达。

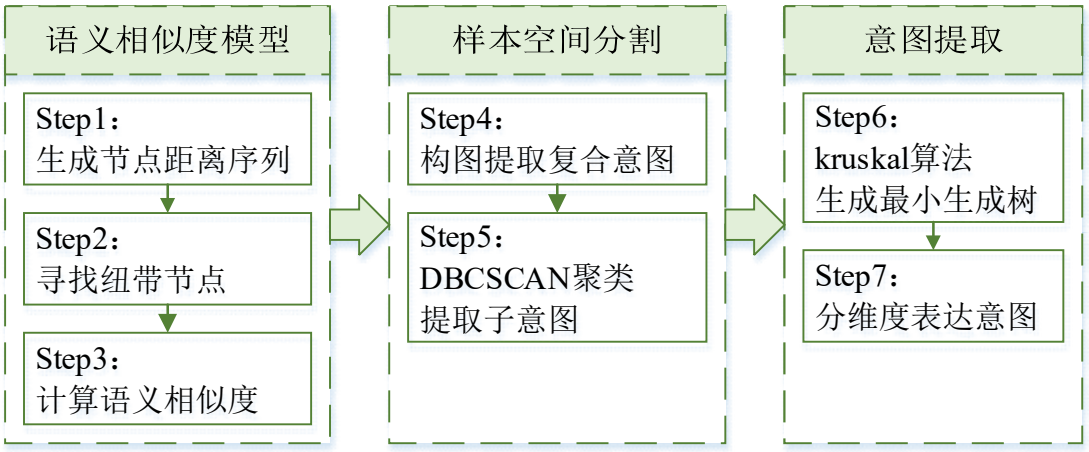


图 3.1 地图内容意图提取步骤

3.2 语义相似度模型的建立

（1）生成节点距离序列

本体库具有以下两个特点：

a. 节点所在深度越深，则语义相似度越高，反之则反。

b. 本体库的顶层概念，即根节点 Root，根节点之间没有语义关系，同一根节点下的所有叶子节点均有语义关系。

基于以上特点，针对每一个节点所处的位置，以自然常数为底数，计算其上下位节点到该节点的距离，建立语义相似度模型。如图 3.2 所示，灰色节点为样本点 X 及其在本体库中的位置，无填充色的节点为样本点 X 在本体库中的上下位节点。每个节点到根节点的距离计算如公式（3.1）所示，节点与节点间的距离计算如公式（3.2）所示。根据样本节点在本体库中的位置及距离根节点的距离，计算得到每一个样本节点的（上/下位节点：距离）距离序列。

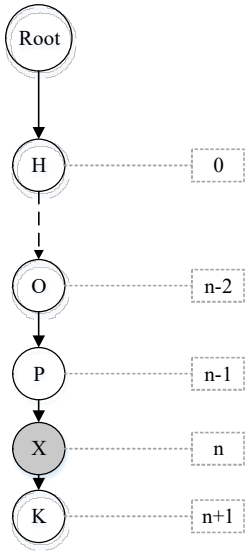


图 3.2 节点序列

$$S(X \to root) = \sum_{i=0}^{n} e^{-i} \tag{3.1}$$

$$S(O \to X) = S(X \to O) = |S(O \to root) - S(X \to root)| \tag{3.2}$$

（2）寻找纽带节点

图 3.3 所示为一组样本点及其在本体库中的位置，灰色节点均为样本节点，无填充色的节点即为样本节点在本体库中的上下位节点。红色边框的节点即为纽带节点，可见除了 X 节点，其余任意两节点之间均有连接，X 节点与任意节点都没有连接。

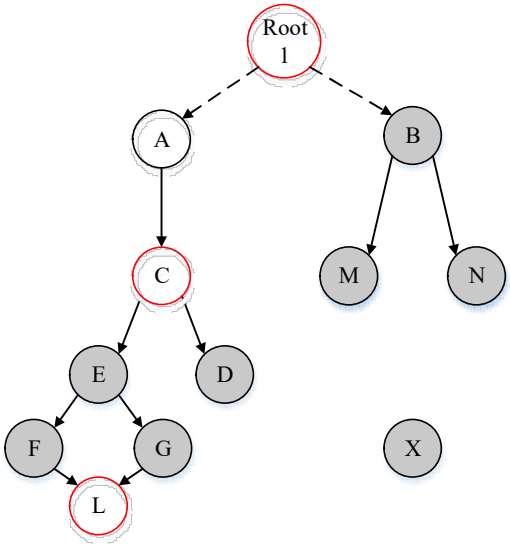


图 3.3 样本节点示例

（3）计算语义相似度

当两节点间存在连接时，由公式（3.3）完成语义相似度的度量，若两个节点间有多个纽带节点，取连接后距离最小的作为纽带节点，比如图 3.3 中的 F 与 G，因为 L 和 E 均是共同节点，L 作为下位概念距离 FG 更近，因此选择 L 作为纽带节点。当节点间不存在连接的节点，如图 3.3 中 X，由公式（3.4）计算可知，属于同一根节点下的节点间的最大距离一定小于MAX_DIS，因此将不存在连接的节点间的语义相似度记为MAX_DIS。通过对本体库的观察，本体库中最深的根节点，即向下的最大步长为$MAX_STEP = 8$,因此如公式（3.4）所示，可得不同节点间的最小距离MIN_DIS。

$$S(Node_1 \rightarrow Node_2) = min[S(Node_1 \rightarrow Link) + S(Link \rightarrow Node_2)] \tag{3.3}$$

$$MAX\_DIS = 2 * \lim_{n\to\infty} \sum_{i=0}^{n} e^{-i} = 2 * \frac{e}{e-1} \tag{3.4}$$

$$MIN\_DIS = e^{-MAX\_STEP}, \quad MAX\_STEP = 8 \tag{3.5}$$

3.3 样本空间的分割

（1）构图提取复合意图

计算包括纽带节点在内，任意两节点之间的语义相似度，建立大小为$(C_{\text{样本节点数+纽带节点数}} \times C_{\text{样本节点数+纽带节点数}})$的对称阵。表 3.1 为图 3.3 中较有代表性的 5 个节点间的距离矩阵，其中 X 与 Root1、B、F、G 的距离均为MAX_DIS；B 与其余节点的距离，均以根节点 Root1 作为纽带节点，距离均大于 2 小于MAX_DIS；FG 以 L 作为纽带节点获得最短距离。

表 3.1 部分距离矩阵

	Root1	B	F	G	X
Root1	0	$S(Root1 \rightarrow B)$	$S(Root1 \rightarrow F)$	$S(Root1 \rightarrow G)$	MAX_DIS
B	$S(B \rightarrow Root1)$	0	$S(B \rightarrow Root1 \rightarrow F)$	$S(B \rightarrow Root1 \rightarrow G)$	MAX_DIS
F	$S(F \rightarrow Root1)$	$S(F \rightarrow Root1 \rightarrow B)$	0	$S(F \rightarrow L \rightarrow G)$	MAX_DIS
G	$S(G \rightarrow Root1)$	$S(G \rightarrow Root1 \rightarrow B)$	$S(F \rightarrow L \rightarrow G)$	0	MAX_DIS
X	MAX_DIS	MAX_DIS	MAX_DIS	MAX_DIS	0

切断距离($S \geq MAX\_DIS$)的节点间的连接，将节点分为节点组，每个节点组即为一组复合意图，组内任意两节点间语义距离($S < MAX\_DIS$)。由于图 3.3 中 X 与其余节点的距离均为MAX_DIS，图 3.3 所示的样本组被分为了两组，如图 3.4 所示。设置复合意图的数量阈值，当复合意图中的样本数目大于 2 则暂且保留复合意图，图 3.4 所示的两个复合意图，复合意图 1 成立，复合意图 2 不成立。仅复合意图 1 进入下一步子意图提取。

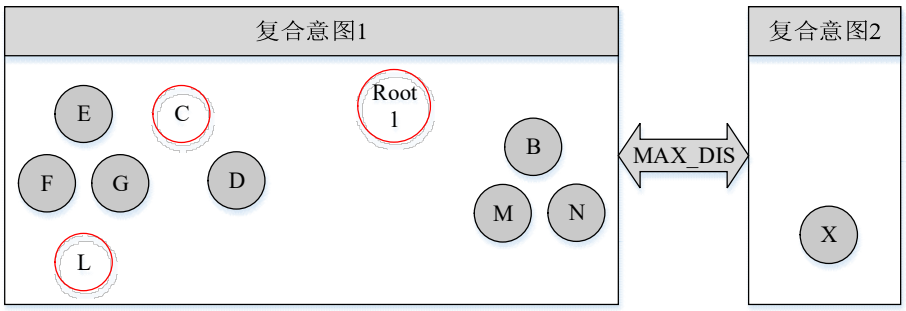


图 3.4 复合意图

（2）DBSCAN 聚类提取子意图

DBSCAN 算法是一种基于密度的聚类方法，由密度可达关系找到密度相连的样本集合即为簇，该算法可以在有噪声的数据集中发现任意形状的簇，此处不再赘述 DBSCAN 的具体聚类步骤。如图 3.5 所示即为对图 3.4 中复合意图 1 进行聚类后得到的蓝色和绿色两组子意图。

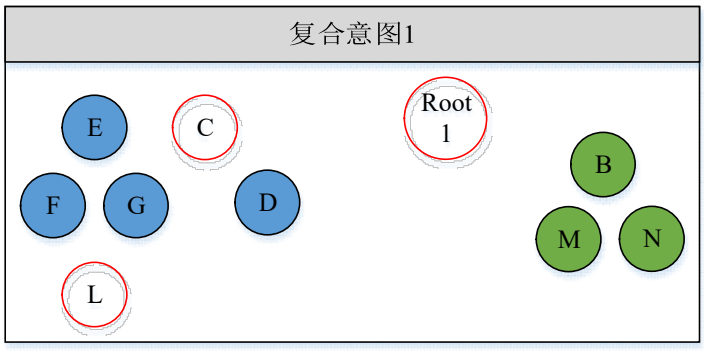


图 3.5 子意图

DBSCAN 聚类中最重要的两个参数为阈值 MINPTS 和半径 EPSILON，当到某节点距离小于半径的节点数目大于阈值，该节点为核点。本文在复合意图样本组的基础上完成聚类，当存在距离相近的两个样本点，则有意图存在的可能，因此阈值参数不变且为 2。由于不同的样本组所在深度不同，半径对聚类结果会产生较大的影响，本研究采取动态半径的聚类法，从最小半径形成一个簇，至最大半径将所有复合意图中的样本均聚集为一个簇停止。在有效利用正样本的情况下，聚类簇的数

量保持相对稳定则子意图相对稳定可靠，节点间语义距离越小越紧密则子意图越突出，因此本研究结合平均簇凝聚度、正样本利用率和聚类稳定性，共三个指标综合选择最佳聚类参数。

如公式（3.6）和（3.7）为第l种聚类结果下，簇凝聚度的计算。公式（3.6）所示为均方根$RMSE$的计算，其中S_{ij}为属于同一聚类簇的任意两节点间的语义距离，$\overline{S_k}$为第 k 个聚类簇的平均语义聚类，n_k为第 k 个聚类簇中的样本数量，n为总样本数量。平均簇凝聚度的计算如公式（3.7）所示，$Cohesion_k$为单个簇的簇凝聚度，$Cohesion_k$越大，则子意图越突出，$Cohesion_l$ 为第l种聚类结果下的平均簇凝聚度，平均簇凝聚度越大，则复合意图的聚类结果越紧密。

$$RMSE_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(S_{ij} - \overline{S_k})^2} \quad , \tag{3.6}$$

$$\overline{S_k} = \frac{1}{n_k}\sum_{j=1}^{n_k} S_{ij}\,, i = 1,2,\dots,n$$

$$Cohesion_k = \frac{MIN\_DIS}{MIN\_DIS + RMSE_k} \tag{3.7}$$

$$Cohesion_l = \frac{1}{n}\sum_{k=1}^{n} Cohesion_k$$

如公式（3.8）所示为第l种聚类结果下的正样本利用率U_l，$Count_l$为第l种聚类结果下形成子意图的正样本总数，$Count_{all}$为复合意图的中的样本总数，利用率越高则形成子意图的正样本越多，正样本利用越有效，当所有样本均形成簇，正样本利用率为 1，当没有形成聚类簇时，正样本利用率为 0。

$$U_l = \ Count_l/Count_{all} \tag{3.8}$$

如公式（3.9）表示在第l种聚类结果下的聚类稳定性S_l，$Class_length_l$表示在第l种聚类结果下，半径可变的最大长度，$max(Class_length)$表示在当前复合意图中，聚类数保持不变的最大半径可变长度，S_l即当类别数保持不变时，半径可变的相对最大长度，聚类稳定性最大为 1 表示最稳定的聚类结果，当没有形成聚类簇时，为 0。

$$S_l = Class\_length_l/max(Class\_length) \tag{3.9}$$

公式（3.10）表示在第l种聚类结果下的相对簇凝聚度C_l， $Cohesion_l$为第l种聚

类结果的簇凝聚度，$max(Cohension)$为当前复合意图中的最大平均簇凝聚度，平均簇凝聚度最大的聚类结果，相对簇凝聚度为最大值 1。

$$C_l = Cohesion_l / max(Cohension) \tag{3.10}$$

如公式（3.11）所示为第l种聚类结果下。综合簇凝聚度、正样本利用率、聚类稳定性的复合意图聚类结果评分，三个衡量指标的最大值均为 1，当总评分最高时，为最佳聚类结果，即取得最佳子意图。

$$ConScore_l = S_l \times U_l \times C_l \tag{3.11}$$

3.4 地图内容检索意图的提取

（1）kruskal 算法生成最小生成树

带权图各边的权值总和称为该树的权，权值最小的生成树即为最小生成树，生成最小生成树需解决以下两个问题：

a. 尽量选取权值小的边，但不能构成回路。

b. 共选取$n-1$条边，连通n个顶点，使得当前生成树的权值最小。

kruskal 算法的思想是以边为处理对象，每次选取最小边权的边，并实时判断该边链接的两个节点有没有间接联通，若未联通则保留该边，若已联通则放弃该边，直到所有节点均被联通。如图 3.6 所示为示例样本中复合意图的最小生成树示例。

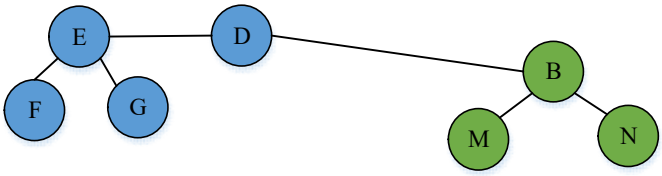


图 3.6 最小生成树

（2）分维度表达意图

根据不同维度的特点，本研究提出点聚集度、线性系数和面性系数三个指标分别作为意图维度空间中零维意图、一维意图、二维意图的权重。

零维意图：以聚集度最高的点作为点状意图的中心，如公式（3.12）所示，$S_{i \to j}$为以 i 节点作为中心，j 节点到 i 节点的语义距离，当以 i 节点作为中心时，平均距离最小，$Concentration_k$最大时，第 k 个子意图的零维意图是 i 节点。$Concentration_k$越大，则零维意图越突出，当形成该子意图的样本节点内容相同时，点聚集度取最大值为 1。

$$Concentration_k = \frac{MIN\_DIS}{\left(MIN\_DIS + min\left(\frac{1}{n}\sum_{j=1}^{n} S_{i\to j}\right)\right)} \quad (3.12)$$
$$i = 1,2,\ldots,n$$

一维意图：在最小生成树的基础上，一维意图是从根节点出发，长度最长的意图线。以蓝色子意图为例，如图 3.7 所示，以深度最浅的 E 节点为起点，深度最深的 F 节点为终点，形成红色边框线的“E-F”意图线即为一维意图。意图线至少由深度不同的两个节点构成。如公式（3.13）所示，$LineK_{longest}$表式第 k 个子意图中的最长意图线上样本节点的数量，$Count_k$表示第 k 个子意图的样本总数，$LineIndex_k$越大则表示该子意图中形成一维意图的样本越多，则一维意图越突出，当所有样本点都在意图线上时，线性系数取最大值为 1。

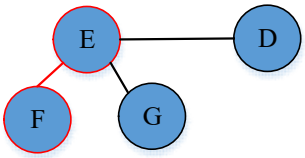


图 3.7 一维意图

$$LineIndex_k = \frac{LineK_{longest}}{Count_k} \quad (3.13)$$

二维意图：在最小生成树的基础上，首先寻找所有意图线，各意图线分布越均匀，说明意图向多个方向发展，则意图形状越接近“面”状。如公式（3.14）所示，以线性系数的均方差的倒数作为第 k 个子意图的面状的评价指标，$AreaIndex_k$越大，则说明二维意图越突出，当意图线均等长时，面性系数取最大值为 1。公式（3.15）表示综合交叉度，即多条意图线的交点，$X_{frequncy}$表示经过 X 节点的意图线的数量，$Line_{总}$为意图线的总数，$Crossover$最大为 1，表示所有意图线均通过 X 节点，X 节点为面状意图的中心。公式（3.16）表示连接度，X_{degree}表示当前节点在最小生成树中的度，$\max(XDegree)$表示 X 节点度的最大值，$Degree$最大为 1，说明面状意图以 X 节点为中心，向所有方向延展。如公式（3.17）所示，若点 X 的$Score_x$为最大，即综合交叉度和连接度均较高时，该二维意图是以 X 节点作为起点的，按广度优先向叶子节点扩展的“面”，$Score_x$越大，面状意图越突出。

$$AreaIndex_k = 1/\left(1 + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(LineIndex_i - \overline{Lineindex})^2}\right) \quad (3.14)$$

$$\overline{Lineindex} = \frac{1}{n}\sum_{i=1}^{n} LineIndex_i, \qquad i = 1,2,\dots,n$$

$$Crossover = \frac{X_{frequncy}}{Line_{总}} \tag{3.15}$$

$$Degree = \frac{X_{degree}}{max(XDegree)} \tag{3.16}$$

$$Score_x = Crossover * Degree \tag{3.17}$$

3.5 本章小结

本章主要介绍了地图检索意图的提取方法。基于本体库建立了语义相似度模型，实现节点间语义相似度的度量并生成距离矩阵，通过构图完成复合意图的提取，在复合意图的基础上使用动态半径的 DBSCAN 聚类提取子意图，采用平均簇凝聚度C_l、正样本利用率U_l和聚类稳定性S_l，共三个指标对每个聚类结果综合评价，保留最佳聚类结果，完成子意图的提取。最后利用意图的形式化表达模型，在 kruskal 算法生成的节点最小生成树的基础上，完成对子意图的维度表达，对零维意图提出点聚集度$Concentration_k$，聚集度最大的节点即为零维意图；对一维意图提出线性系数$LineIndex_k$，$LineIndex_k$最大的意图线为以深度最浅的节点为起点，深度最深的节点为终点的线，该意图线即为一维意图；对二维意图提出面性系数$AreaIndex_k$，各意图线分布越均匀面性系数越大，面状中心为综合交叉度$Crossover$和连接度$Degree$最大的节点。以上各类指标是构建置信度评价机制的基础。

4 基于置信度的迭代式相关反馈

采用迭代式相关反馈的方式是为了提高在大量地图中检索的效率和解决用户的意图可能产生漂移的问题，用户意图的漂移是指用户在检索过程中，意图发生改变，与最初的检索意图产生出入，此时相关反馈的正负样本集合和权重应当发生改变。本研究结合两种相关反馈的形式，首先是用户勾选正样本或加入购物车的显式反馈；其次是基于置信度的伪相关反馈，在下次反馈时引导用户完成相关反馈，明确意图的方向。当用户对此次检索结果不满意时，再次勾选正样本，系统按置信度进行迭代反馈，直到取得令用户满意的检索结果。

4.1 相关反馈的特点

从如何获取反馈信息方式的角度[31]，相关反馈行为可分为显示反馈、隐式反馈和伪相关反馈。显示反馈[32]要求用户直接对检索结果做出相关与否的标记，是相对简单高效的反馈方式，但会复杂化用户的操作。隐式反馈通过分析用户的行为来发现用户的兴趣和爱好，从而提高检索效率，如[33,34]鼠标点击、停留等，但是该方法存在冷启动的问题，同时搜集用户行为存在一定的隐私侵犯。伪相关反馈则可以解决以上两种方式存在的部分问题，该方法不需要用户的参与，常用的方式是对初次检索结果排名靠前的 N 项做分析扩展[35]，对改善查询质量有一定的辅助作用。早期的伪相关反馈主要是基于文本的伪相关反馈[36]，但结果却差强人意，尤其是当前 N 项初次查询结果中存在噪声时，扩展的查询甚至会降低检索质量[37,38]。随着社交网络的普及，出现了基于图的伪反馈方法[39,40]，该方法不需要分析文档的具体内容，主要是基于拓扑结构，将检索的前 N 项结果作为相关节点，研究对象间的关系来寻找用户的检索目标，能够有效改善系统的检索结果，但改善程度依赖于本体库的可靠性。随着信息技术的进一步发展，出现自然语言处理及深度学习相结合的方法[41]。

综合考虑到地图检索的复杂性和相关反馈算法的效果，深度学习的反馈方法缺乏可解释性，且算法复杂性太高。其次，由于缺乏历史数据，为了避免冷启动问题，本研究暂不考虑隐式反馈。显示反馈能够更好的体现用户的检索偏好，实现个性化检索，伪相关反馈能够减少用户的参与，提高系统的用户友好性。本研究以结合显示反馈和基于图的伪相关反馈的方式，并设置置信度检查机制，综合反馈达到

提高检索精度的目的。

如图 4.1 所示，在显式反馈中，用户存在两种正反馈行为：加入购物车和勾选相关样本。加入购物车行为说明该地图在较大程度上符合用户的需求，勾选正样本说明该地图与用户需求存在较大的关联性或相似性。因此加入购物车的正反馈行为可以作为未来对用户检索偏好的研究，为隐式反馈打下基础。目前本研究仅关心在当前检索情况下，从用户勾选的正样本中提取检索意图。在用户显式反馈的基础上，基于图将用户勾选的正样本作为相关节点，在第三章的基础上，求取意图维度空间的权重，将权重作为置信度，完成伪相关反馈，有效降低显式反馈对用户勾选正样本的准确性和完备性要求。

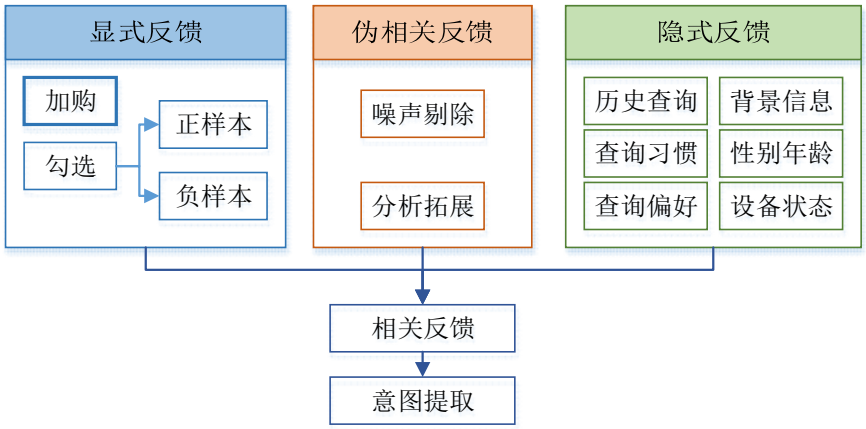


图 4.1 相关反馈的一般工作流程

4.2 置信度评价机制的构建

结合第三章的内容，本研究从整体到局部，首先确定复合意图的置信度，如图 4.2 中的蓝色部分所示，从复合意图质量和复合意图突出度两个方面对复合意图进行置信度评估；如图 4.2 中的绿色所示，在复合意图置信度的基础上，结合子意图维度权重和子意图突出度，共三个方面对子意图进行置信度评估。

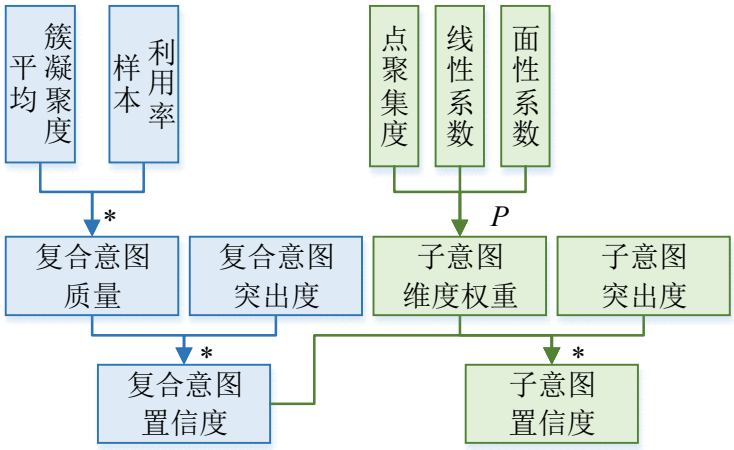


图 4.2 置信度机制

（1）复合意图置信度

复合意图质量Q_{con}：如公式（4.1）所示，复合意图质量由复合意图的平均簇凝聚度和样本利用率决定。其中平均簇凝聚度$Cohension$用于衡量复合意图聚类结果的质量，为直接可比量，当所有聚类簇中的样本节点间距离为 0 时，簇凝聚度为最大值 1。样本利用率 U用于衡量复合意图中的有效正样本数，为直接可比量，当所有正样本均参与聚类时，样本利用率为最大值 1。复合意图的质量最佳时，Q_{con}为最大值 1。

$$Q_{con} = Cohension \times U \tag{4.1}$$

复合意图突出度Sig_{con}：如公式（4.2）所示，以当前复合意图中的正样本总数相对于正样本总数最多的复合意图的比例作为突出度的衡量指标，当所有正样本形成一个符合意图时，突出度取最大值为 1。

$$Sig_{con} = Count_{con}/max(Count_{con}) \tag{4.2}$$

复合意图置信度$Conf_{con}$：如公式（4.3）所示，当复合意图的质量和突出度均较大时，复合意图的的置信度取得较大值，最大为 1。

$$Conf_{con} = Q_{con} \times Sig_{con} \tag{4.3}$$

（2）子意图置信度

子意图维度权重向量P：如公式（4.4）所示，P是由点聚集度$Concentration$、线性系数$LineIndex$、面性系数$AreaIndex$，三者归一化后构成的向量。其中点聚集度衡量一维意图，当形成该子意图的样本节点内容相同时，点聚集度取最大值为 1。线性系数$LineIndex$衡量一维意图，当所有样本点都在意图线上时，线性系数取最大值为 1。面性系数$AreaIndex$衡量二维意图，当所有意图线均等长时，面性系

数取最大值为 1。

$$P = \left[\frac{Concentration}{sum}, \frac{LineIndex}{sum}, \frac{AreaIndex}{sum}\right]^t \tag{4.4}$$
$$sum = Concentration + LineIndex + AreaIndex$$

子意图突出度Sig：如公式（4.5）所示，以当前子意图中的正样本总数相对于复合意图中正样本数最多的子意图的比例作为突出度的衡量指标，当所有正样本形成一个子意图时，即复合意图与子意图大小相等，子意图突出度取最大值为 1。

$$Sig = \ Count/max(Count) \tag{4.5}$$

子意图置信度$Conf$：$Conf$为(3×1)的向量，如公式（4.6）所示，当复合意图置信度、子意图突出度、和某个维度的权重均最大时，则子意图在该维度的置信度取得最大值为 1。

$$Conf = Conf_{con} \times \ Sig \times P \tag{4.6}$$

此时意图集合的表达式如表 4.1 中的示例，其中 Dimension 标识意图的维度，0 代表零维，1 代表一维，2 代表二维；Node 标识关键点，零维意图关键点为中心节点，一维意图关键点为意图线上的节点，二维意图关键点为中心节点；Confidence 标识置信度。

表 4.1 意图集合示例

Dimension	Node	Confidence
0	A	$Conf_1$
1	[A,B,C]	$Conf_2$
2	A	$Conf_3$

4.3 迭代式相关反馈的实现

本节的目标是根据上一小节的置信度，合理的设置阈值，舍弃置信度低的意图，保留置信度高的阈值，然后以置信度作为比例，生成用户下一次相关反馈的样本集合。由于每次相关反馈的情况差异较大，本研究参考“肘部法则”的思想，寻找置信度变化曲线上发生明显转折的点，即为“肘部点”，肘部点所在处即为置信度的阈值。

（1）拟合置信度曲线

如公式（4.7）所示，将各个意图的置信度按从大到小的顺序排列后，如公式

（4.8）所示，以X作为自变量，置信度$ConfList$作为因变量，用三次曲线完成拟合。

$$ConfList = \ sort(Conf_1, Conf_2, \dots, Conf_n) \tag{4.7}$$

$$f(x) = \ fit(X, ConfList, 3), \qquad X = 1,2, \dots n \tag{4.8}$$

（2）置信度变化率曲线

根据模型特点，一定存在置信度较高的意图和置信度较低的意图，本模型以置信度变化率达到 50%时的置信度作为置信度的阈值。如公式（4.9）所示是求置信度变化率达到 50%时的横坐标$CutX$。

$$f(x)' = 0.5, \qquad CutX = x \tag{4.9}$$

（3）置信度分割点

公式（4.10）为当变化率达到 50%时的置信度，$CutY$即为置信度阈值，置信度大于$CutY$的意图保留，置信度小于$CutY$的意图丢弃。

$$f(CutX) = CutY \tag{4.10}$$

（4）下次反馈样本集

公式（4.11）为按照置信度，将各意图按比例分配，生成下一组样本供用户进行勾选。其中零维意图的样本生成为带零维意图标签的样本，一维意图的样本生成为意图线穿过的所有节点并向线的两端以深度优先的延伸，二维意图的样本生成为以中心节点作为面状中心向所有意图线以广度优先的等长延展。

$$proportion_i = \frac{Conf_i}{sumConf}, \qquad sumConf = \sum_{i=1}^{n} Conf_i \tag{4.11}$$

4.4 本章小结

本章首先介绍了相关反馈的特点，结合模型背景，本研究以结合显示反馈和基于图的伪相关反馈的方式，并设置置信度检查机制，综合反馈达到提高检索精度的目的。因此，本研究利用各个指标建立置信度评价机制，以复合意图质量和复合意图突出度确定符合意图置信度，在此基础上，再结合子意图维度权重、子意图突出度，共同确定子意图置信度。最后参考“肘部法则”的思想，确定置信度的阈值，保留高置信度意图，丢弃低置信度意图，按比例生成下次反馈的样本集合，迭代回意图提取模型中，再次完成检索意图的提取，直到取得另用户满意的检索结果。

5 基于典型检索场景的实验分析

由于本研究缺少真实的用户数据，因此本研究根据相关背景，设置典型检索场景，并根据合理的规则生成样本集合，对本研究在建立检索意图提取模型时提出的动态半径、复合意图综合评价指标、子意图维度表达和迭代式相关反馈的置信度机制进行有效性分析和验证。

5.1 典型检索场景样本的生成

样本生成如图 5.1 所示，首先参考相关检索系统，构造典型检索场景，设计地图检索意图。第二步以七种制图方法、九大地理主题和 SWEET 本体库中的类作为标签的取值，按照一定的规则随机生成总样本库。第三步则根据第一步中设计的意图，自动筛选得到意图所对应的样本组，并人工验证以保证样本空间覆盖以及有足够样本对应意图。最后利用意图识别算法提取意图，与设计的意图进行匹配性验证，以检验算法的有效性、准确性。如表 5.1 所示为用户在单方面有单意图典型检索场景的样本示例，意图取值由意图维度和关键节点标识，以地图内容为例，表中的目标意图为面状意图，因此其中一个正样本的地图内容为二维意图中心节点的子类。

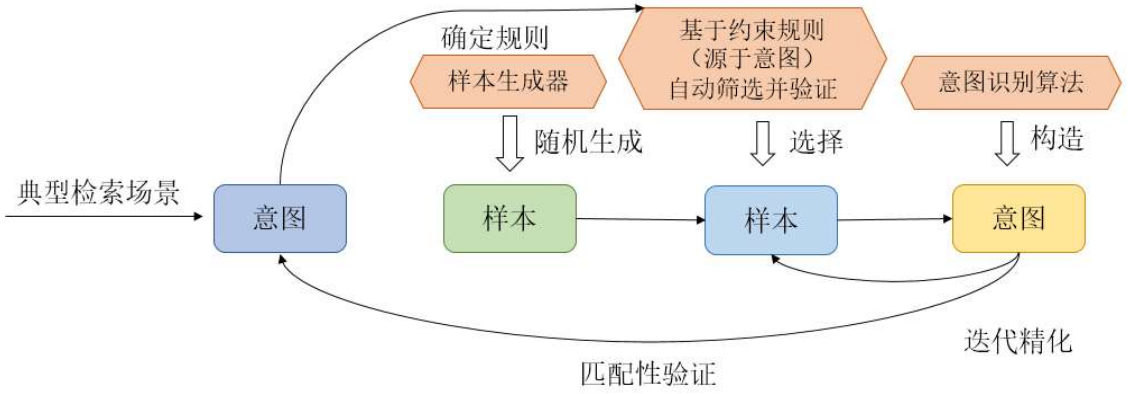


图 5.1 样本生成与验证流程

表 5.1 单意图单类型检索场景

	T(Theme)	C(Content)	M(Mapping)
Intention	"0","None"	"2","http://sweetontology.net/propTemperature/Temperature"	"0","None"
Positive samples:			
"T": ["Disaster"], "M": ["None"], "C": ["http://sweetontology.net/propTemperature/WindChill"]			
"T": ["None"], "M": ["Others"], "C": [http://sweetontology.net/propTemperature/PotentialTemperature]			
…			
Negative samples:			
"T": ["Climate"], "M": ["None"], "C": ["http://sweetontology.net/stateTimeFrequency/Erratic"]			
…			

5.2 示例分析

5.2.1 动态半径

为了说明半径对聚类结果的影响，本研究选取了一组子意图较多的样本集，如图 5.2 所示为该样本集在不同半径下完成聚类后的类别数和簇凝聚度变化情况。随着半径的增大，聚类簇的数量先增大，至最大类别数后，距离相近的簇合并，类别数减小，簇凝聚度逐渐减小。

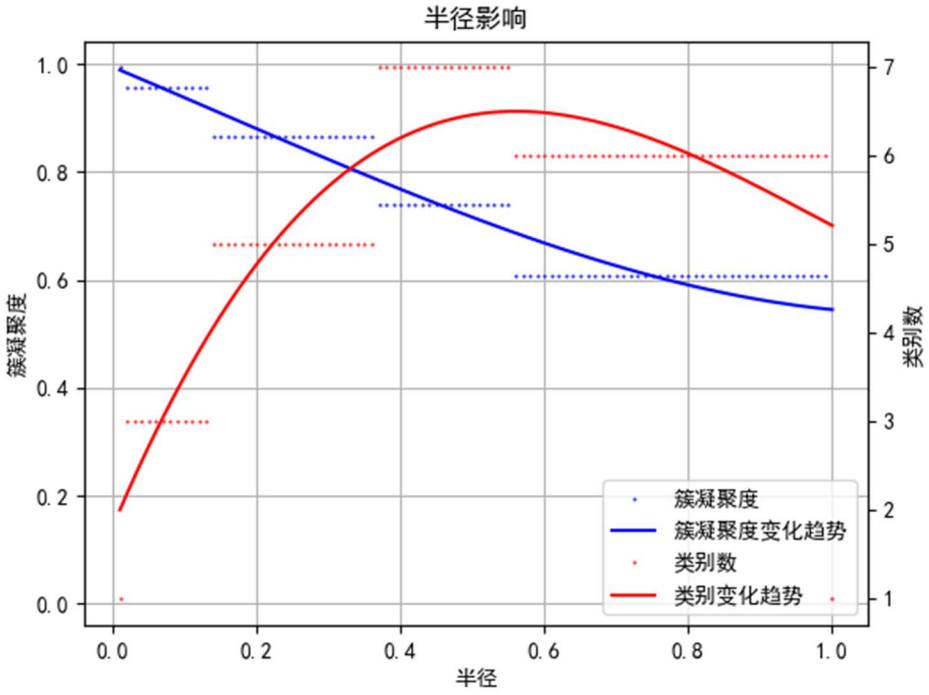


图 5.2 半径影响

5.2.2 复合意图综合评价指标

如图 5.3 所示为同一组样本在不同聚类结果下的综合评价指标得分，横坐标表示第l种聚类结果及当前聚类数“$l-\ Count_l$”，纵坐标为不同评价指标的分值，在该样本情况下，随着半径的增大，共形成 6 个类别的聚类结果，簇凝聚度逐渐降低，样本利用率逐渐上升，聚类稳定性在形成 6 个簇时最稳定，综合第 5 种聚类结果，形成 6 个簇，即 6 个子意图时，聚类结果最好。

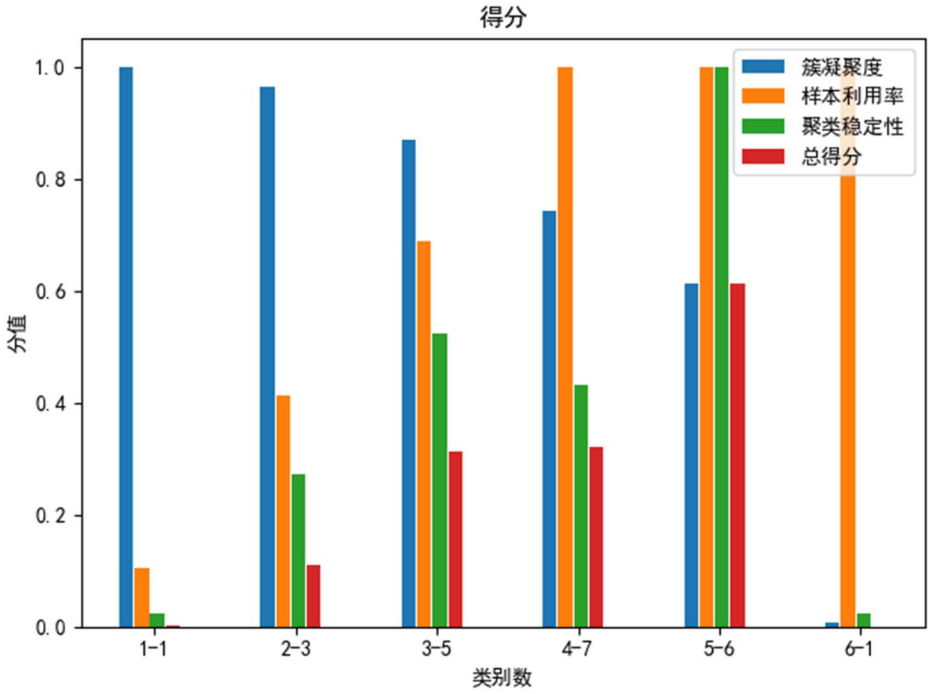


图 5.3 复合意图综合评价指标

5.2.3 子意图维度表达

为了更好的说明面状系数的有效性，以图 5.4 所示的样本组为例，该样本组是从上文所使用的复合意图样本组，在最佳聚类情况下，提取到的子意图之一，图 5.4 为该子意图的最小生成树，其中蓝色的为样本节点，无填充色的为纽带节点。

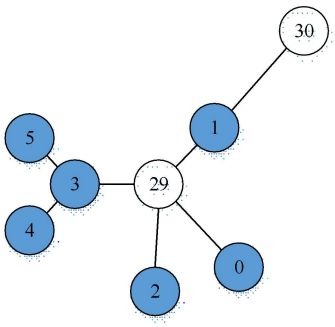


图 5.4 样本组示例

从该子意图中提取到的零维意图、一维意图、二维意图结果如表 5.2 所示。其

中零维意图是 29 号节点，聚集度达到 0.96；一维意图包括两条长度相等的意图线，线性系数均为 0.5；二维意图是以 29 号节点为中心，向四周延展的面，面状系数为 0.89。在面状系数中心节点的选择时，分别以各点作为面状中心的得分如图 5.5 所示，可见 29 号节点在连接度和交叉率上都是最符合二维意图中心的。通过对子意图样本组的观察，可见本研究所建立子意图表达模型具有一定的可靠性。

表 5.2 意图集合示例

Dimension	Node	Confidence
0	29	0.2012
1	[30，1，29，3，4]	0.5000
1	[30，1，29，3，5]	0.5000
2	29	0.8891

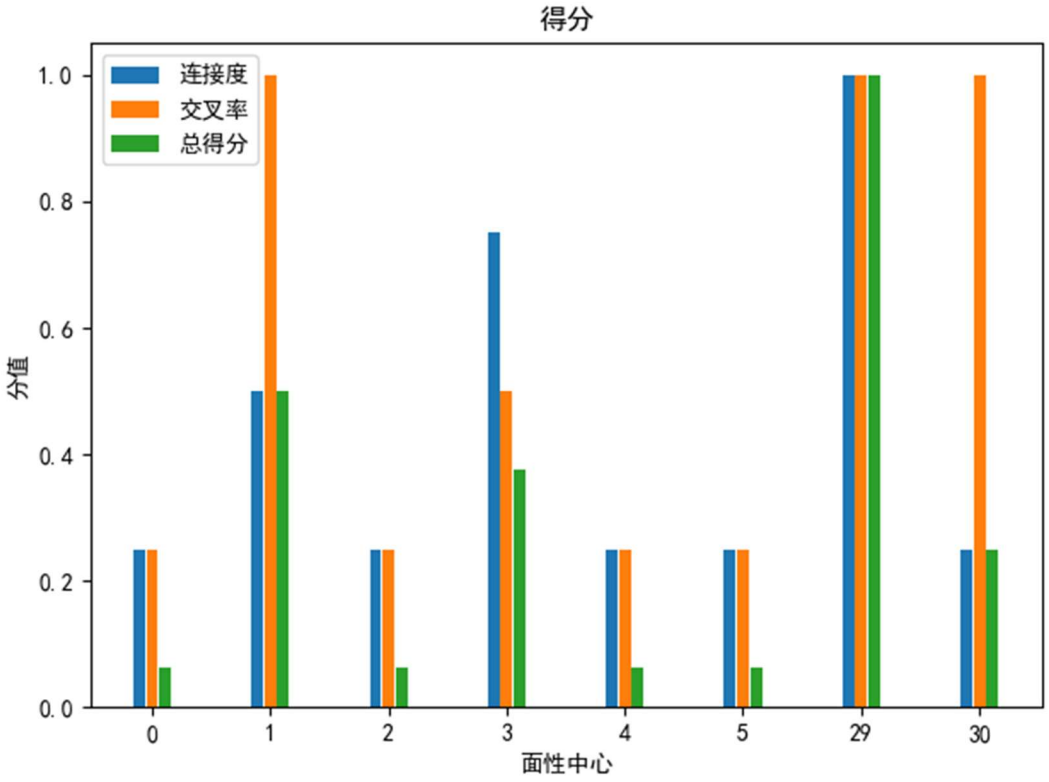


图 5.5 面状中心得分

5.2.4 迭代式相关反馈

从上文的同一样本集中，共提取到 17 组子意图，置信度变化如图 5.6 所示，蓝色为置信度曲线，绿色为置信度曲线的斜率变化线，绿色节点为置信度变化率达到 50%时的分割点，垂直方向的红色虚线为$CutX$，与蓝色的置信度曲线相交，得

到水平方向的红色虚线为$CutY$，可见水平方向红色虚线以上即为保留下来的子意图，此次相关反馈共提取到 6 个较为突出的子意图，将置信度换算为比例后，即可据根据相关原则，完成下一次相关反馈样本集的生成，具体如表 5.3 所示。

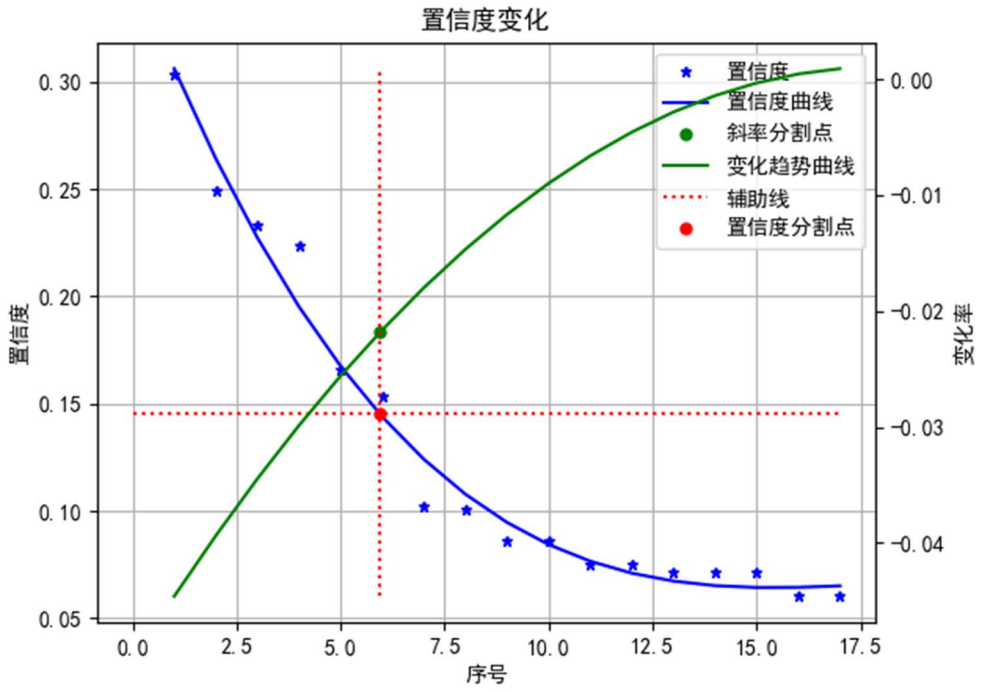


图 5.6 置信度变化

表 5.3 意图集合示例

Dimension	Node	Confidence	Proportion
0	29	0.1659	0.1225
2	29	0.1533	0.1182
0	36	0.2234	0.1657
2	36	0.2495	0.1899
0	37	0.2333	0.1754
2	37	0.3037	0.2283

5.3 匹配性验证

在单意图单维度的场景下，为零维意图、一维意图、二维意图分别设计 3 组单意图单维度的样本，以 3 组样本的准确率均值代表维度的准确率。利用本体库生成每个意图对应的正样本全集，以随机抽取正样本的形式模拟用户在相关反馈中

勾选正样本的行为，以正样本个数的增加来模拟迭代反馈数量的增加。如图 5.7 所示为匹配性验证的流程图，以随机抽取 10 次样本数为 N 时的平均准确率的作为最终准确率。零维意图和二维意图准确率以关键节点间的语义距离来度量，一维意图的准确率以提取的意图线与真实意图线的重叠度来度量。

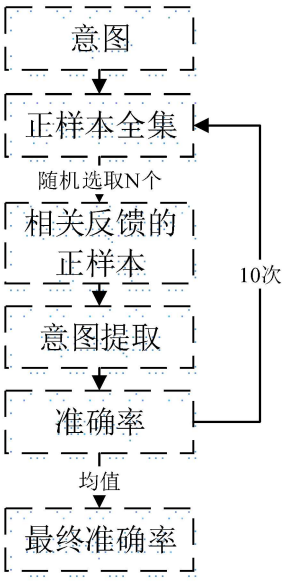


图 5.7 匹配性验证

零维意图、一维意图、二维意图提取的准确性随着样本数增加的变化如图 5.8 所示，三个维度的准确率均逐渐提高，其中零维意图和二维意图均可达到 100%的准确率，一维意图相对于二维意图和零维意图的准确率较低，一维意图扩展的方式还有待完善。

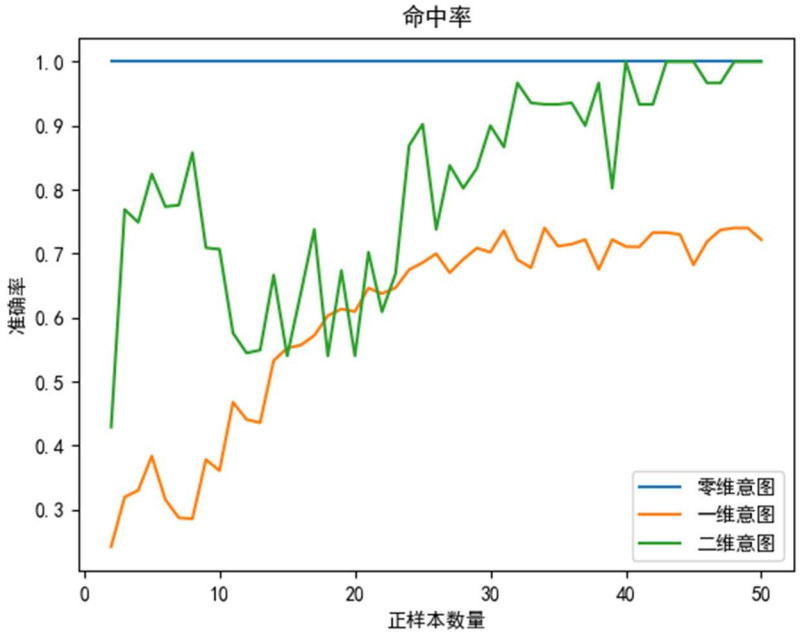


图 5.8 准确率

5.4 本章小结

本章设计了典型检索场景样本的生成原则，生成合适的样本集。首先以一组样本示例，从 4 个方面对模型的有效性进行了检验。在动态半径的设置上，模型能够提取到较好的半径阈值完成子意图的提取，对复合意图的综合评价符合真实情况，子意图的表达也较为合理，迭代式相关反馈能够完成 Top-N 意图的提取。其次针对单意图单维度的情况进行了匹配性验证，随着正样本数量的迭代增加，其中零维意图和二维意图的准确率均可达到 100%，一维意图的准确率相对较低，意图提取算法的设计还有待完善。本研究对模型的实验还需加以补充，讨论模型在不同场景下的意图提取能力和优缺点。

6 总结和展望

6.1 总结

本研究基于 DBSCAN 聚类，在顾及地图制图特征和语义属性的同时，设计了地图检索意图提取的模型，并对模型的有效性进行了验证。主要内容如下：

（1）根据相关背景调查，建立意图的分类体系，并根据意图的多维性、多样性、抽象性、模糊性、多模性的特点，确定以地图主题、地图内容和制图方法作为本研究意图提取的目标。根据专题地图制图教材确定制图方法的 7 个特征值，根据 9 大社会效益领域确定地图主题的 10 个特征值，根据 SWEET 本体库确定地图内容的值域。类比与现实世界的维度定义，创新性的构造意图的维度空间，实现对意图的形式化表达。

（2）基于 DBSCAN 聚类算法完成地图检索意图提取。首先通过对本体库的特点进行总结，基于 SWEET 本体库建立语义相似度模型，根据节点之间的距离矩阵，初步切分距离过大的节点集合，形成复合意图；其次在复合意图的基础上提出动态参数的 DBSCAN 聚类，以聚类稳定性、正样本利用率、平均簇凝聚度三个指标综合确定最佳聚类参数，完成子意图提取；最后利用意图的形式化表达机制对子意图解析，明确用户意图。

（3）结合显示反馈和基于图的伪相关反馈的方式，并设置置信度检查机制，综合实现迭代式相关反馈，达到提高检索精度的目的。以复合意图质量和复合意图突出度确定符合意图置信度，在此基础上，再结合子意图维度权重、子意图突出度，共同确定子意图置信度。参考“肘部法则”的思想，确定置信度的阈值，保留高置信度意图，丢弃低置信度意图，按比例生成下次反馈的样本集合。

（4）设计典型检索场景样本生成方法，从动态半径、复合意图综合评价指标、子意图维度表达、迭代式相关反馈共 4 个方面对模型的有效性进行了检验。

6.1.1 创新与特色

（1）建立基于本体库的语义相似度模型：不同于基于步长的语义相似度度量模型，本文考虑到不同深度的节点语义相似度的差异性，建立更加合理的语义相似度度量模型。

（2）构造意图的维度空间：类比于现实世界对空间的定义，以维度的形式定

义检索意图，为意图的形式化表达提供一种新的思路。

（3）基于置信度的迭代式相关反馈：结合显示反馈和基于图的伪相关反馈的方式，并设置置信度检查机制，在保证检索精度的同时，最大化提高系统的可用性和用户友好性。

6.1.2 不足

（1）负样本利用不足：目前本模型只考虑从正样本中提取检索意图，未考虑负样本的聚集模式和对正样本的干扰性等信息，与正样本形成参照，提高意图提取的准确性。

（2）未考虑维度间的联系：制图方法、地理主题和地图内容，三大意图取值方面之间，存在具有较大参考价值的信息，通过对它们之间相关信息的挖掘能够有效的提高意图提取的有效性。

（3）对模型的讨论不足：缺少真实的用户数据，典型检索场景的合理性无法验证。缺少对模型鲁棒性、意图提取效率等多方面的考察。

6.2 展望

本文研究属于利用聚类算法完成地图检索意图提取的初步探索，在意图的形式化表达、聚类算法和迭代式相关反馈上都还有很大的发挥空间。

（1）意图的形式化表达：本文对意图维度空间的定义目前只实现到二维空间，对高维空间还可以进行更多的探索，扩展意图的形式，挖掘不同维度空间的联系。

（2）聚类算法：本研究采用较为简单的 DBSCAN 算法完成对意图的提取，但本文的意图提取实际是基于本体库的图结构，采用相关的图聚类算法或超图分割算法可能更加有效。

（3）迭代式相关反馈：根据上述不足，在迭代式相关反馈中，提取负样本中的有效信息，提高相关反馈信息的利用率，从而提高意图提取的准确性。

参考文献

[1] Gui, Z., Yang, C., Xia, J., Liu, K., Xu, C., Li, J., Lostritto, P., 2013. A performance, semantic and service quality enhanced distributed search engine for improving geospatial resource discovery. International Journal of Geographic Information Science, 2013, 27(6), 1109-1132.

[2] Li Z., Yang C., Wu H., et al. An optimized framework for seamlessly integrating OGC Web Services to support geospatial sciences [J]. International Journal of Geographical Information Systems, 2011, 25(4): 595-613.

[3] Gui Z., Cao J., Liu X., et al. Global-Scale Resource Survey and Performance Monitoring of Public OGC Web Map Services [J]. ISPRS International Journal of Geo-Information, 2016, 5(6): 88.

[4] Gu J., Feng C., Gao X., et al. Query Intent Detection Based on Clustering of Phrase Embedding [J]. 2016: 110-122.

[5] Lamine M. Review of Human-Computer Interaction Issues in Image Retrieval [M]. Advances in Human Computer Interaction. 2008.

[6] Spink A, Losee R.M. Feedback in information retrieval [J]. Annual Review of Information Science Technology. 1996, 31: 112-118

[7] 李牧闲. 顾及制图方法和主体内容的 WMS 图层检索方法研究[D]. 武汉大学, 2020.

[8] Liu, K., Yang, C., Li, W., Gui, Z., Xu, C., Xia, J., 2014. Using semantic search and knowledge reasoning to improve the discovery of Earth science records: an example with the ESIP Semantic Testbed. International Journal of Applied Geospatial Research, 2014, 5(2), 44-58.

[9] Liaoat M., Khan S., Majid M. Image retrieval based on fuzzy ontology [J]. Multimedia Tools and Applications, 2017, 76(21): 22623-22645.

[10] Lowe D.G. Object recognition from local scale-invariant features; proceedings of the Proceedings of the Seventh IEEE International Conference on Computer Vision[C], F 20-27 Sept. 1999.

[11] Hu, K., Gui, Z., Cheng, X., Qi, K., Zheng, J., You, L., Wu, H., 2016. Content-based Discovery for Web Map Service using Support Vector Machine and User Relevance Feedback. PLoS ONE, 2016, 11(11): e0166098. DOI: 10.1371/journal.pone.0166098

[12] Gong Y., Wang L., Guo R., et al. Multi-scale Orderless Pooling of Deep Convolutional Activation Features; proceedings of the Computer Vision – ECCV 2014, Cham [C]. European Conference on Computer Vision. Springer International Publishing. 2014: 392-407.

[13] 闫浩文, 王家耀. 基于 Voronoi 图的点群目标普适综合算法[J]. 中国图象图形学报, 2005, 10(05):633-636.

[14] Rasiwasia N., Moreno P., Vasconcelos N. Bridging the Gap: Query by Semantic Example [J]. Multimedia, IEEE Transactions on, 2007, 9: 923-938.

[15] 李牧闲, 桂志鹏, 成晓强, 吴华意, 秦 昆. 多核学习和用户反馈结合的 WMS 图层检索方法. 测绘学报, 2019, 48(10), 1320-1330.

[16] Lops P., De Gemmis M., Semeraro G. Content-based Recommender Systems: State of the Art and Trends [M]. Recommender Systems Handbook. Boston, MA; Springer US.

2011: 73-105.
[17]Schafer J.B., Frankowski D., Herlocker J., et al. Collaborative Filtering Recommender Systems [M]. BRUSILOVSKY P, KOBSA A, NEJDL W. The Adaptive Web: Methods and Strategies of Web Personalization. Berlin, Heidelberg; Springer Berlin Heidelberg. 2007: 291-324.
[18]Balabanovic M., Shoham Y. Fab: Content-Based, Collaborative Recommendation [J]. Communications of the ACM, 1997, 40: 66-72.
[19]Martinez L., Pérez L., Barranco M. A multigranular linguistic content-based recommendation model [J]. International Journal of Intelligent Systems, 2007, 22: 419-434.
[20]刘睨, 李鹏, 刘欣, et al. 基于用户聚类的推荐算法 [J]. 2015, 000(010): 269-272.
[21]Schein A., Popescul A., Ungar L., et al. Methods and Metrics for Cold-Start Recommendations [M]. Sigir: International Acm Sigir Conference on Research & Development in Information Retrieval. ACM. 2002: 253-260.
[22]Zhang Y., Chen X. Explainable Recommendation: A Survey and New Perspectives [M]. 2018.
[23]Gui, Z., Yang, C., Xia, J., Li, J., Abdelmounaam, R., Sun, M., Xu, Y., Fay, D., 2013. A visualization-enhanced graphical user interface for geospatial resource discovery. Annals of GIS, 2013, 19(2), 109-121.
[24]Zhu H., Zhang P., Li G., et al. Learning Tree-based Deep Model for Recommender Systems [J]. 2018,
[25]王顺箐. 以用户画像构建智慧阅读推荐系统 %J 图书馆学研究 [J]. 2018(04): 92-96.
[26]Yang, Z., Gui, Z., Wu, H.*, Li, W., 2020. A Latent Feature-based Multimodality Fusion Method for Theme Classification on Web Map Service. IEEE Access, 2020, 8, 25299-25309.
[27]张敏, 桂志鹏*, 成晓强, 曹军, 吴华意. 一种 WMS 领域主题文本提取及元数据扩展方法. 武汉大学学报信息科学版, 2019, 44(11), 1730-1738.
[28]Wei, Z., Gui, Z.*, Zhang, M., Yang, Z., Mei, Y., Wu, H., Liu, H., Yu, J., 2021. Text GCN-SW-KNN: A Novel Collaborative Training Multi-Label Classification Method for WMS Application Themes by Considering Geographic Semantics. Big Earth Data, 5(1), 66-89.
[29]Tomás Pariente, José María Fuentes, María Angeles Sanguino, et al. A Model for Semantic Annotation of Environmental Resources: The TaToo Semantic Framework[J]. 2017.
[30]Ramachandran R, Graves S, Raskin R. Ontology Re-engineering Use Case: Extending SWEET to map Climate and Forecasting Vocabulary Terms[J]. Agu Spring Meeting Abstracts, 2006.
[31]于莹莹, 陈燕, 张金松. 相关反馈在信息检索中的研究综述 [J]. 情报理论与实践, 2016, 39(12): 135-139.
[32]Rocchio J. Relevance Feedback in Information Retrieval, in the SMART Retrieval System[M]. The SMART Retrieval System: Experiments in Automatic Document Processing. 1971: 313-323.

[33]JOACHIMS T. Optimizing search engines using clickthrough data [M]. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. Edmonton, Alberta, Canada; Association for Computing Machinery. 2002: 133–142.
[34] Joachims T., Granka L., Pan B., et al. Accurately Interpreting Clickthrough Data as Implicit Feedback [J]. ACM SIGIR Forum, 2017, 51: 4-11.
[35]Robertson S.E., Jones K.S. Relevance weighting of search terms [M]. Document retrieval systems. Taylor Graham Publishing. 1988: 143–160.
[36]Ruthven I.A.N., Lalmas M. A survey on the use of relevance feedback for information access systems [J]. The Knowledge Engineering Review, 2003, 18(2): 95-145.
[37]Sakai T., Manabe T., Koyama M. Flexible pseudo-relevance feedback via selective sampling [J]. ACM Trans Asian Lang Inf Process, 2005, 4: 111-135.
[38]Vassilvitskii S., Brill E. Using web-graph distance for relevance feedback in web search [M]. International Acm Sigir Conference on Research & Development in Information Retrieval. ACM. 2006: 147-153.
[39]Su Y., Yang S., Sun H., et al. Exploiting Relevance Feedback in Knowledge Graph Search [M]. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 1135-1144.
[40]Kundu M.K., Chowdhury M., Rota Bulò S. A graph-based relevance feedback mechanism in content-based image retrieval [J]. Knowledge-Based Systems, 2015, 73: 254-264.
[41]ALMasri M., Berrut C., Chevallet J-P. A Comparison of Deep Learning Based Query Expansion with Pseudo-Relevance Feedback and Mutual Information [C]. European Conference on Information Retrieval. 2016: 709-715.

致　谢

在此我首先要感谢桂志鹏老师，从进入小组开始，桂老师一直耐心的指导我，介绍组内的研究方向。前期由于没有接触过该课题，入门很慢，是桂老师的坚持不懈坚定了我的信心，确定此次毕业设计的研究方向。在毕业设计阶段，从开题、中期到现在，桂老师都很认真负责，前期两周一次小组会讨论，后期由于进度紧张，一周一次讨论，给我提供了莫大的帮助和很多宝贵的建议，是桂老师的一丝不苟、积极的态度驱使着我不断的前进。

其次我要感谢同组的师兄、师姐，前期确定研究方向时，他们都十分耐心的回答我的问题，帮助我尽快了解研究内容。在后期实现阶段，他们也不遗余力的给予我帮助，无论是技术上的问题，还是思路上的探讨，都为我提供了很多宝贵的经验。除开此次毕业设计，在与师兄师姐的交流和学习中，我对将来的就业趋势也有了更多的了解，对个人的未来也有了更多的思考。

最后我要感谢家人和朋友，四年的本科生涯即将顺利结束，一路走来离不开大家的陪伴、支持和包容。我会不忘初心，砥砺前行。