基于 TF-IDF 和 DeBERTa 混合模型的数据目录语义检索系统研究

谷剑芳
河南省政务大数据中心，河南 郑州 450000
DOI:10.61369/ASDS.2026010005

摘　　要： 本文以政务数据目录智能化搜索为研究目标，针对传统关键词匹配方法存在的语义理解不足、同义词覆盖不全等问题，提出了一种基于 TF-IDF 与 DeBERTa 算法的混合模型的轻量级语义搜索的解决方案。通过整合政务数据目录中的目录名称、申请依据、应用场景、标签等多维度数据构建语料库，利用中文分词、动态权重调整、领域术语扩展等技术，建立融合关键词精确匹配与语义相似度计算的双层检索机制。通过实践，本研究解决了语义识别场景中的三个核心问题：一是利用 TF-IDF 与 DeBERTa 的有机融合，发挥两个算法在关键词匹配和短句语义理解的优势，提高文本搜索的召回率；二是面向政务服务领域应用，建立丰富的知识库，建立专业术语与民间表述之间的映射桥梁，解决语义鸿沟跨越问题；三是通过对搜索结果的二次过滤，解决过度泛化所产生的“语义漂移”，进一步提升搜索的精确度。

关 键 词： 语义分析；TF-IDF；DeBERTa；混合模型；动态权重；数据目录

Research on a Hybrid TF-IDF and DeBERTa Model for Semantic Retrieval in Data Catalog Systems

Gu Jianfang
The Henan Provincial Government Big Data Center, Zhengzhou, Henan 450000

Abstract: This study focuses on intelligent search for government data catalogs. To address the limitations of traditional keyword matching methods—such as inadequate semantic understanding and incomplete synonym coverage—we propose a lightweight semantic search algorithm based on a hybrid model integrating TF-IDF and BERT algorithms.This research aims to build a lightweight intelligent search model for government service scenarios, addressing three core issues. First, it leverages the organic integration of TF-IDF and BERT to utilize the advantages of both algorithms in keyword matching and short sentence semantic understanding, thereby improving the recall rate of text search. Second, it establishes a rich knowledge base for government service applications, creating a mapping bridge between professional terms and folk expressions to overcome the semantic gap. Third, through secondary filtering of search results, it addresses over-generalization and potential "semantic drift" from Word2Vec, further enhancing the precision of government data directory search and providing an efficient intelligent search solution for government data resource sharing scenarios.

Keywords: semantic analysis; TF-IDF; DeBERTa; hybrid model; dynamic weight; data catalog

引言

在数字化转型背景下，政务数据共享已成为提升政府治理能力的关键路径。截至2023年底，全国一体化政务服务平台使用总量超过888亿人次，其中，证照共享服务体系持续优化，已汇聚全国31个地区和26个国务院部门58亿条目录，累计提供电子证照共享服务97亿余次，持续推动“减证便民”[1]。本研究致力于构建轻量级政务数据目录智能搜索系统，实现以下目标：建立跨部门术语知识库，覆盖“社会保险”“行政审批”等12个高频领域的586个同义词对设计动态权重调节机制，使 TF-IDF 与 BERT 的混合权重随查询特征自适应变化开发面向政务场景的查询扩展算法，解决“个人养老金→基本养老保险待遇”等专业映射问题。本研究提出的混合搜索模型，通过融合 TF-IDF 的精准匹配优势与 BERT 的语义泛化能力，经过实验证明，有效提升跨部门目录检索的准确率。特别在社保、卫生、教育、交通、公积金、房地产等高频领域，解决术语标准不统一带来的搜索盲区问题，为构建全国一体化政务大数据体系提供技术支撑。

作者简介：谷剑芳，女，河南省政务大数据中心数据治理部，副部长，研究方向：数字政府建设、政务数据治理、数据要素流通；邮箱：13525568581@sina.com.cn。

一、相关工作

（一）数据目录检索系统的现状

政务数据目录作为政府数据资源的管理核心，具有以下特征：

1. 结构化与非结构化数据并存：目录名称、标签类型等字段具有明确的结构化特征，而应用场景、申请依据等字段多为自然语言描述的非结构化文本。

2. 领域术语密集性：数据目录和应用场景包含大量政策文件特有的专业术语（如“容缺受理”“一网通办”等）和部门专属概念（如“社保基数核定”“税控云监管”等）。

3. 动态更新需求：随着政务服务业务的扩展，跨部门业务协同的丰富，数据目录会持续增加、变更。如：2024年，随着国办推进“高效办成一件事”，大部分一件事都需要跨部门协同办理，新的目录对应新的业务，要求目录搜索算法同步提升对文本的推理能力。

以中部K省为例，政务数据目录超过两千条，目录名称平均长度13.7个字符，包含专有名词占比12.5%。应用场景文本平均长度44.3字符，每句覆盖专用术语：0.11个，专业术语占比81.68%。申请依据文本平均长度75.6字符，文件引用占比53.95%，标签体系包含11类58个。这些特征决定了传统关键词匹配方法难以满足需求，需引入分词、词向量和语义分析等技术。现行目录检索系统主要采用关键词匹配或分词匹配的方法，从语义精准匹配的角度，普遍存在三大痛点：其一，传统关键词匹配对“企业－公司”“养老－退休金”等政务场景高频同义词缺乏映射能力；其二，缺乏跨领域术语差异解析能力，目录名称与应用场景的语义割裂导致跨部门检索效率低下；其三，政策术语快速迭代得静态词库难以适应“一件事一次办”“免申即享“等新型服务模式的搜索需求。

随着政务服务领域“三融五跨”业务的快速增长，数据共享也越来越高频。2025年，国家数据局发布《数字中国发展报告（2024年）》公布[2]：近5年，各地区政务服务平台数据共享累计超过5400亿次。数据目录作为数据共享的驱动，为用数单位找数、供数单位发布服务，提供了基本工作界面和载体。近年来，随着政务数据资源目录和数据共享业务的不断增长，数据目录查询量不断增加，传统关键词检索机制暴露两大核心短板：其一，难以精准识别“生存认证/死亡记录”等业务场景同义词关联；其二，缺乏跨领域术语差异解析能力（如“税控云、成品油监管”等专业表述辨识困难）。本研究是为了促进数据共享，通过算法实现智能化数据目录搜索，为用数单位找到需要的数据提供便利。

（二）TF-IDF模型概述

1.TF-IDF的基本原理。Luhn[3]的TF-IDF加权策略通过逆文本频率强化专业术语表征，TF是词频(Term Frequency)，IDF是逆文本频率指数(Inverse Document Frequency)，其核心思想是如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力[4]。该方法适合用来分类，组成要素包括词频、逆文本频率和TF-IDF值。词频，是词语在单个网页中出现的频率，反映词语对当前网页的重要性。逆文本频率，用于衡量词语在整个语料库中的稀有程度，稀有词具有更高的区分能力。TF-IDF值越高，词语对网页的代表性越强。

2.TF-IDF在信息检索中的应用。TF-IDF在处理中文政务数据时面临两大挑战：一是政务场景特有的缩略语（如“一网通办”“跨省通办”）导致特征稀疏；二是政策文件的多义表述造成语义漂移。Mikolov[5]提出的Word2Vec模型开创了分布式词向量新范式，Lee[6]将其扩展为Doc2Vec实现文档级语义表征。在政务领域，Zhang[7]构建的政务词向量库GOV2VEC包含5.7万个政策术语，但未解决新政策术语的动态更新问题。

（三）DeBERTa模型概述

1.DeBERTa的前身：BERT模型的基本原理。2018年Google AI研究院提出BERT模型。与传统的单向语言模型不同，BERT利用Transformer架构中的多层Encoder结构，实现了对输入序列中每个位置的双向依赖建模，这使得它能够同时考虑输入文本的上下文信息。这种双向性显著增强了模型对复杂语义关系的理解能力，进而提高了在各种自然语言处理任务中的表现。

BERT在自然语言处理中的应用。翁克瑞[8]等结合BERT预训练模型与改进的TextCNN架构，通过分析社交媒体（如抖音）评论数据，识别公众对新能源汽车、碳中和等政策的意愿倾向，分类准确率达82.0%-86.4%，显著优于传统深度学习方法。刘青[9]等利用BERT语义模型+企业知识图谱，从专利摘要中识别劳动节约型技术创新（如自动化设备专利），实现量化劳动力成本上升对企业技术创新的促进作用。采用BERT模型实现语义搜索，其768维向量的计算复杂度进一步提升语义搜索的准确性，同时也会增加内存开销。

2.DeBERTa: 解码增强的BERT与解耦注意力大模型。DeBERTa具有解耦注意力机制和增强的掩码解码器[10]。在解耦注意力机制方面，DeBERTa模型则采用双向量表示法，分别编码词的内容和位置信息，并通过基于内容和相对位置的解耦矩阵计算词间注意力权重。在增强型掩码解码器方面，与BERT类似，DeBERTa采用掩码语言建模（MLM）进行预训练，通过分析掩码词周围的上下文词汇来预测被遮蔽的词语。通过解耦注意力机制捕捉上下文词汇的内容特征和相对位置，通过MLM预训练有效处理词汇的绝对位置信息，两者结合，能够更有效发现词语之间的关联关系。

二、混合模型设计

（一）需求分析

随着政务服务领域“三融五跨”业务的快速增长，数据共享也越来越高频。数据目录作为数据共享的驱动，为用数单位找数、供数单位发布服务，提供了基本工作界面和载体。近年来，随着政务数据资源目录和数据共享业务的不断增长，数据目录的搜索量不断增加，传统关键词检索机制暴露两大核心短板：一是难以精准识别“生存认证/死亡记录”等业务场景同义词关联；二是缺乏跨领域术语差异解析能力（如“税控云、成品油监管”等专

业表述辨识困难）。为了促进数据共享，通过算法实现智能化数据目录搜索，为用数单位找到需要的数据提供便利。确定以下业务逻辑：

1. 明确分析对象。以数据共享订阅历史记录为搜索对象，匹配的字段包括目录名称、应用场景，以及与目录有关联的标签、场景分类、提供部门等字段。

2. 建立权重参数。按照直接和间接表达目录实际用途的价值，为目录名称含义、应用场景、申请依据、场景分类、标签等字段建立相似度计算权重参数。

3. 建立语义识别知识库。语义识别具有场景化的特点，每个应用领域都有行业术语。为了让大模型能够理解行业术语，精准识别文本含义，建立政务服务领域字典和同义词词典，为了过滤语言中的助词、语气词、副词，建立停用字词词典。

（二）系统架构

1. 总体设计。政务数据目录智能化搜索系统的核心架构采用分层设计理念（如图1），通过数据预层、算法层、融合层、增强功能四个核心模块的协同工作，实现从原始数据到精准搜索结果的完整处理流程。构建 TF-IDF 与 DeBERTa 的双通道语义理解体系，突破传统单一算法的局限性。通过 TF-IDF 向量空间（精确匹配目录名称等结构化字段）与 DeBERTa 语义识别的并行计算，实现关键词匹配与语义理解的优势互补。在技术实现上，采用 hstack 矩阵拼接方式融合两种向量表示，并创新性地引入动态权重调整机制：通过将等参数权重的相似度统计特征（均值、标准差、极值）构建状态向量，输入强化学习模型，模型通过预测动作概率分布输出动态权重分配方案，利用加权 MSE 损失函数结合用户反馈奖励进行梯度下降优化，最终实现权重参数的在线自适应调整，显著提升搜索结果中语义相关性与用户意图的匹配精度。

2. 模块划分。主要功能方法包括数据层、算法层、融合层、增强功能、分析模块和接口层，见图1。

图1：整体模型架构

（三）数据预处理

1. 数据清洗与标准化。将对读取的目录名称、申请依据等字段内容去除异常字符，按权重系数拼接。

2. 语料准备。

（1）中文分词：采用结巴分词工具进行细粒度切分。支持 n-gram（1，2），增强分词效果，提高文本特征的表征能力。

（2）停用词过滤：移除“的”“是”等无意义词汇。

（3）领域术语注入：加载政务专属词典。

（四）TF-IDF 模型实现

1. 词频与逆文档频率计算。TF-IDF（Term Frequency-Inverse Document Frequency）通过统计词频衡量词语重要性，其中，$f_{t,d}$为词t在文档d中的出现次数，N为总文档数，n_t为包含词t的文档数。在本系统实现中，通过 TF-IDF 对文本的向量化实现以下优化：

（1）双向 n-gram：设置 ngram_range=(1,2)，捕获“社保_缴费”等复合词。

（2）动态剪枝：min_df=0.001过滤低频噪声词，max_df=0.95去除泛化停用词。

（3）权重增强：目录名称字段在语料构造时重复3次，提升关键字段权重。

在文本处理流程中，通过双向 n-gram 解析设置 ngram_range=(1,2) 构建弹性词窗，精准捕获“社保缴费”类复合语义单元；结合动态剪枝机制，运用 min_df=0.001智能过滤低频噪声词汇并借助 max_df=0.95精准剔除泛化停用词；同时采用语义权重强化策略，在语料构造时对“目录名称”字段内容复制3份，相当于进行了3次语义强化，实现关键字段的战略级权重提升。

2.TF-IDF 向量化。TF-IDF 的优势是对关键词的匹配，但是缺乏语义关联分析能力：

（1）语义盲区：无法识别“企业－公司”等同义词，导致“省直国企名录”无法匹配“省级国有公司登记信息”。

（2）语境缺失：对“养老”一词，难以区分社保场景中的“养老保险办理”与民政场景中的“养老机构监管”。

（3）长尾效应：IDF 逆文档率算法，让政策文件中的一些低频专业术语（如“非税收入划转”）IDF 值过高，从而产生过度加权。

TF-IDF 算法虽在关键词精准匹配层面展现优势，却存在语义关联解析的系统性局限：其机械的字面匹配机制无法辨识“企业－公司”等语义等价关系，导致“省直国企名录”与“省级国有公司登记信息”形成检索断层；对“养老”等多义词汇缺乏上下文感知能力，难以区分社会保障领域的“养老保险办理”与民政管理范畴的“养老机构监管”应用场景；更因 IDF 逆文档频率算法的特性，使文本中“非税收入划转”等低频专业术语产生异常加权，引发算法权重失衡。实证研究表明，基于 TF-IDF 的单一模型在标准测试集上相似度较低，超过50% 的相关目录无法检索出来，语义鸿沟导致的跨领域匹配失效构成主要误差源。

（五）DeBERTa 模型实现

1.DeBERTa 模型选择与微调。采用 deberta-v3-base 预训

练模型（12层 Transformer，768维隐藏层），通过以下策略实现计算效率与语义精度的平衡：

文档向量预计算：系统初始化时离线计算全部目录的 DeBERTa 向量，存储为 numpy 数组。

批处理优化：采用16条 / 批的小批量处理，避免内存溢出。

缓存机制：将预计算结果保存为外部文件，后续直接加载，避免重复计算。

向量提取策略：使用 [CLS] 标记的向量作为文档级语义表征，相比传统词向量平均法更能保留上下文信息。vector = outputs.last_hidden_state[0, 0, :].numpy()

2. 语义嵌入生成

第一步：语料增强处理

（1）停用词过滤算法：建立多级过滤规则，包括163个通用停用词，如“查询”“获取”等操作动词，根据 TF-IDF 权重自动发现低频噪声词。

（2）启用领域术语：构建包含三级结构的领域知识库（如图1），包括核心结构（政务词典基础同义词库、200+ 领域术语库）、重点领域（社保、税务、市场监管等重点政务领域术语）、特色应用（智慧城市、电子证照、区块链存证等），支持“医保→医疗保险→医疗保障局”的递进式扩展。

图2：三级结构的领域知识库

第二步：采取混合向量化算法。建立双通道向量空间模型：建立 TF-IDF 通道，生成稀疏特征矩阵。建立 DeBERTa 通道：训练词向量模型并计算文档向量。

（1）优化设置 TF-IDF 算法参数：包括设置次线性词频缩放参数（Sublinear TF=true），避免高频词因绝对频次过高而主导权重，从而更合理地反映词语在文档中的重要性。设置动态 DF 阈值（min_df=0.001）过滤长尾词。

（2）DeBERTa 模型训练：捕捉语义关联，解决一词多义问题，平均词向量生成文档表示，多线程并行计算显著提升 BERT 训练效率。

（3）动态策略：引入了强化学习（RL）来动态调整，影响算法精准度的参数权重：TF-IDF、应用场景、目录名称、申请依据、标签类型、标签名称、语义。

（4）用户参与：根据用户反馈的无效目录来过滤无关结果，提升搜索质量。

第三步：动态权重调节

基于相似度分布的动态权重算法，根据各字段相似度统计量动态调整权重w_i，实现表现好的字段权重高，波动大的字段权重低的智能分配。

（1）波动抑制：当某字段相似度方差过大时自动降低权重。

（2）重点突出：对稳定高匹配字段给予权重奖励。

（3）归一化处理：保证权重总和为1。

第四步：对检索结果实施二次过滤

为强化关键词匹配效果，采取 TF-IDF 硬过滤，经过多次测算，且判断条件为：目录名称相似度 >0.15或应用场景 >0.15。

（六）系统实现

1. 开发环境与工具

（1）硬件平台：华为笔记本电脑（32GB 内存，INTEL CPU）

（2）软件环境：Python 3.8 、WINDOWS 操作系统

（3）模型：

模型1：TF-IDF 模型（n-gram 分词范围：1-2）

模型2：Word2Vec 模型（向量维度80）

模型3：TF-IDF 混合 Word2Vec 模型（向量维度80、n-gram 分词范围：1-2）

模型4：TF-IDF 混合 BERT 模型（向量维度768、n-gram 分词范围：1-2）

模型5：TF-IDF 混合 DeBERTa 模型（向量维度768、n-gram 分词范围：1-2）

2. 系统功能模块

（1）主要功能

采用 deberta-v3-base 预训练模型（12层 Transformer，768维隐藏层），使用 [CLS] 标记的向量作为文档级语义表征，相比传统词向量平均法更能保留上下文信息。模型要解决三个问题：一是利用 TF-IDF 与 deberta-v3-base 的有机融合，发挥两个算法在关键词匹配和语义理解的优势，提高文本搜索的召回率。二是面向政务服务领域应用，建立丰富的知识库，建立专业术语与民间表述之间的映射桥梁，解决语义鸿沟跨越问题。三是通过对搜索结果的二次过滤，解决过度泛化产生的“语义漂移”，进一步提升政务数据目录搜索的精确度。通过以下策略实现计算

效率与语义精度的平衡。

文档向量预计算：系统初始化时，离线计算全部目录的 DeBERTa 向量，存储为 numpy 数组。

批处理优化：采用16条 / 批的小批量处理，避免内存溢出。

缓存机制：将预计算结果保存为 npy 文件，后续直接加载。

（2）用户输入与查询处理

首先使用自定义的词典对语料进行 jieba 分词，然后对分词结果进行向量化计算，计算词向量的平均值作为文档向量，其中空文档用零向量表示。

（3）检索结果生成

以查询“个人养老金”为例，直接用模型计算，取相似度值前15个。15个结果中，相似度最高的第4项是“机关事业单位工勤岗位登记电子证照”，与养老金没有任务关系，第7项“城乡居民基本养老保险待遇领取信息”属于个人养老金，但是排序靠后。

3. 性能优化

（1）优化检索速度

①领域语料注入：定义政务领域同义词和术语。本案例设置了覆盖教育、医疗、税务、社保、就业、卫生、交通、海关、消防、金融、职称、自然资源等政务行业的128组同义词和83组覆盖智慧城市、数据共享、特色政府服务等领域术语。

②混合相似度计算：通过矩阵拼接实现 TF-IDF 稀疏矩阵与 DeBERTa 稠密向量的协同计算。

（2）提升结果准确性。DeBERTa 算法核心是双向 Transformer 编码器，通过多层自注意力机制和双向上下文学习生成动态词嵌入，基于词向量简单平均的文档级语义表征方法存在明显局限，其聚合过程会损失词汇序列的位置信息，从而产生语义漂移，例如，搜索“个人养老金”，检索出与“个人”相关，但与“养老金”无关的住房贷款、强制执行等数据目录。避免因同义词扩展导致的过度泛化问题，解决算法在语义解读方面产生的“语义漂移”，即避免语义相关性分析关联到与查询词相关性不大的目录，确保结果与查询存在实质性关联，对搜索结果进行停用词二次过滤，并过滤掉目录名称相似度小于0.1或应用场景相似度小于0.1的结果。经过二次过滤，均与个人养老金有直接关系的四条数据目录：离退休人员养老保险退休信息、城镇企业职工基本养老保险待遇领取信息、机关事业单位养老保险待遇领取信息、城乡居民基本养老保险待遇领取信息，排序在1至4，有效数据召回显著提升。

（3）建立动态权重

构建参数权重的深度学习模型，针对影响相似度识别的7个参数 RL_WEIGHT_FIELDS（TF-IDF、应用场景、目录名称、申请依据、标签类型、标签名称、DeBERTa 语义），计算每个字段相似度的均值、标准差、最大值、最小值，组合成状态向量。将 RL_WEIGHT_FIELDS 参数的动态权重状态向量，输入强化学习模型，模型通过预测动作概率分布输出动态权重分配方案，利用加权 MSE 损失函数结合用户反馈奖励（如过滤无关结果惩罚 -1/-2分）进行梯度下降优化，最终实现权重参数的在线自适应调整，显著提升搜索结果中语义相关性与用户意图的匹配精度。工程化流程实现如下：

动态权重生成过程

①状态特征提取：对每个查询，系统实时计算 RL_WEIGHT_FIELDS 中各字段（如 TF-IDF 整体、场景相似度、语义等）的相似度统计特征，包括均值、标准差、最大值、最小值，形成长度为7×4=28维的状态向量（7个参数 ×4种统计量）。

②模型推理：将状态向量输入 PyTorch 神经网络（包含128单元隐藏层和 Softmax 输出层），输出各字段的初始权重概率分布（如 [0.3, 0.2, 0.1, ..., 0.15]）。

③归一化处理：通过 Softmax 归一化确保权重总和为1，例如将 [0.3, 0.2, 0.1, 0.15, 0.05, 0.1, 0.1] 转换为 [0.25, 0.17, 0.08, 0.13, 0.04, 0.08, 0.07]。

④在线调整：根据用户反馈（如标记无关目录）动态更新模型参数，例如若 BERT 权重过高导致语义偏差，后续查询会降低 BERT 的权重分配。

LR 学习过程（强化学习训练）

①损失函数设计：采用加权 MSE 损失函数，公式为：

Loss = MSE(predicted_weights, true_weights) × (1 + Reward)

其中 Reward 根据用户行为动态调整（如过滤无关结果时 Reward=-2，点击相关结果时 Reward=+1）。

②探索与利用策略：通过 ε -greedy 策略平衡探索（随机采样权重）与利用（按模型预测分配权重），初始 ε =0.2，每轮搜索后衰减至0.99倍。

③梯度更新：使用 Adam 优化器（学习率1e-4）进行反向传播，重点优化导致低奖励的动作参数，例如对产生高惩罚权重的神经元增大反向传播梯度。

④经验回放机制：将状态 - 动作 - 奖励三元组存入经验池（rl_training_data），当积累100条数据后进行批量训练，提升样本利用效率。

4. 实验与评估

（1）实验设计

①数据集选择。本实验采用政务数据资源共享平台的实际业务数据，构建包含18,456条目录记录的测试数据集。数据字段包括目录名称（平均长度13.7字符）、应用场景（平均长度44.3字符）、申请依据（平均长度75.6字符）等关键字段。 实验数据集经过严格预处理：统一数字格式和字段内容。使用相同的分词词典、同义词扩展、停用词、弃用词。通过同一组测试输入，取相似度前10个结果，检验不同模型的检索结果与业务的相关性。

②实验指标。选择高频使用的数据目录内容：企业社保缴费，能够体现个人行为和生存状态的信息，包括养老、工伤和医保缴费，死亡，火化，住院和门诊结算等，查询公积金贷款相关目录，文化旅游政务服务和企业开办一件事相关目录。

（2）实验结果分析

①多模型运行效率对比。从1115条检索结果中，针对上述三个模型从评价指标（检索精度、查全率和排序质量）和热词两个方面进行效果评估。从检索结果的热词分布看，WORD2VEC 模型的热词虽然多，但是与检索输入的相关性不高，TF-IDF 与

DeBERTa 混合模型检索结果热词比 TF-IDF 模型更优，不仅与检索输入的相关性高，输出的有效内容更多。以搜索“企业社保缴费”为例，TF-IDF 模型的高频热词是“企业”“养老保险”和“职工”，WORD2VEC 模型的高频热词是“民企”“企业”和“政策”，TF-IDF 与 Word2Vec 混合模型的高频热词是“缴费”“企业”和“养老保险”，TF-IDF 与 DeBERTa 混合模型的高频热词是“缴费”“社会保险”“职工参保”。可以看出，TF-IDF 与 DeBERTa 混合模型的搜索结果更符合业务期望。

②混合模型的性能评估

定义三个评价指标：精确率、召回率、F1和排序质量[11]，计算值见表1，TF-IDF、Word2Vec、TF-IDF 混合 Word2Vec、TF-IDF 混合 BERT 和 TF-IDF 混合 DeBERTa 五个模型的效果比对示意图见图3。定义和计算方法分别如下：

精确率 (Precision)

定义：强调预测的准确性，不受负类样本数量的直接影响，适合不平衡数据集评估。

公式：Precision = TP /(TP + FP)

举例：例如，top_k=10，有3个相关结果，则 precision@10=0.3。

召回率 (Recall)

定义：系统检索到的相关文档占所有相关文档的比例。

公式：recall@k = 检索到的相关数 / 总相关数

举例：总共有5个相关文档，检索到3个，则 recall@k=0.6。

MRR 排序质量 (RankQuality)

定义：首个相关文档在结果中的排名的倒数，未找到则为0。平均倒数排名，取值范围：

0,1，值越大表示排序质量越好，主要衡量模型返回结果的排序合理性。

公式：MRR = 1 / rank_of_first_relevant_doc（如果存在）

举例：第一个相关文档在第3位，则 MRR=1/3 ≈ 0.333。

F1综合性能（精确率 + 召回率调和平均数）

定义：分类模型评估中一个综合性的指标，用于衡量模型在精确率和召回率之间的平衡能力。

公式：F1=2×(Precision×Recall)/(Precision+Recall)，综合反映模型的整体性能。

通过业务专家标注五个模型的搜索结果并建立评估模型，从评估结果看，TF-IDF 混合 DeBERTa 的 MRR 和精确率最优，展现的结果是相同搜索对象，该模型的搜索排序和结果最优，该模型的指标 F1综合性能、召回率和准确率排名第二。

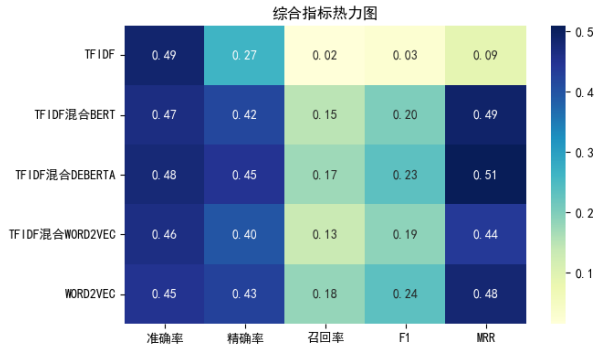

图3：五种模型综合热力图

三、总结

本研究提出 TF-IDF 与 DeBERTa 的动态混合模型，在政务数据目录搜索场景中实现更具有智能化、更贴近用户意图的数据目录。通过向量预计算、缓存等优化技术，使系统响应速度持续提示，支持万级数据目录的文本语义检索。通过语义分析，发现跨部门数据目录的相关性，为挖掘政务服务领域更多具有并联串联办理的“一件事”场景提供线索。未来，需要在三个方面探索政务数据目录智能化搜索的能力：一是构建政务领域预训练语料库。引入主动学习优化同义词库，提升词语推理的理解能力，从而提高语义匹配的准确性。二是持续完善政务语言资源库。包含术语体系、专用词典、同义词库、停用词表及弃用词等核心要素，强化语义解析的领域适应性。三是采取搜索的兜底措施。针对词向量语义分析的不确定性，某些情况下无法正确返回结果甚至结果为空情况，采用分词比对的方法进行兜底，为用户提供有效的结果。

参考文献

[1] 国家互联网信息办公室 . 国家信息化发展报告（2023年）[EB/0L].https://www.cac.gov.cn/2024-09/06/c_1727308607362592.htm.

[2] 国家数据局 . 数字中国发展报告（2024年）[EB/0L].https://www.nda.gov.cn/sjj/zhuanti/sjzgzxd/szzgbg/0605/20240830180401077761745_pc.html.

[3]Luhn, H.P. The automatic creation of literature abstracts[J].1958.《IBM Journal of Research and Development》第2卷，159-165.

[4] 百度百科“tf-idf”词条 [EB/0L].https://baike.baidu.com/item/TF-IDF/8816134.

[5]Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[6]Lee J, Kim S, Song Y. Doc2Vec-based semantic document retrieval in big data environments[J]. Future Generation Computer Systems, 2020, 112: 997-1005.

[7]Zhang Y, Li J, Song Y. GOV2VEC: A domain-specific word embedding model for government documents[C].Proceedings of the 2021 IEEE International Conference on Big Data. IEEE, 2021: 1023-1032.

[8] 翁克瑞，周雅洁，於世为 . 基于 BERT 的多层次特征融合的舆情文本政策意愿识别模型研究 [J]. 中国地质大学学报（社会科学版），2025,25(01):131-140.

[9] 刘青，肖柏高 . 劳动力成本与劳动节约型技术创新——来自 AI 语言模型和专利文本的证据 [J]. 经济研究，2023,58(02):74-90.

[10]Engcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen.《DeBERTa: Decoding-enhanced BERT with Disentangled Attention》[EB/0L].https://arxiv.org/abs/2006.03654.

[11] 王国霞，刘贺平 . 个性化推荐系统综述 [J]. 计算机工程与应用，2012,48(07):66-76.

