DOI:10.20033/j.1003-7241.(2025)05-0090-05.

基于感知哈希算法的能源数据安全快速检索

萧展辉，张世良，宋云奎
（南方电网数字电网研究院有限公司，广东 广州 510000）

摘 要：检索能源数据时极易泄露用户隐私信息，导致检索安全性下降，设计基于感知哈希算法的能源数据安全快速检索方法。利用云环境创建能源数据库，运用已知数据模型和背景数据模型保证数据安全性。采用感知哈希算法把数据映射为较小长度的比特位，通过比特位相等位数判断数据相似度。结合数据相似度计算结果进行多粒度文本匹配，加权求和检索语句和文档的相关性分数，将相关性分数最高的数据提供给用户，实现高效率能源数据检索。实验结果分析表明，该方法的能源数据检索执行时间短，检索索引空间代价小，F1值较高，数据检索安全性好。

关键词：感知哈希算法；能源数据；数据检索；云环境；多粒度文本匹配
中图分类号：TP391 文献标识码：A 文章编号：1003-7241(2025)05-0090-05

Fast Retrieval of Energy Data Security Based on Perceptual Hash Algorithm

XIAO Zhan-hui, ZHANG Shi-liang, SONG Yun-kui
(Southern Power Grid Digital Power Grid Research Institute Co., Ltd., Guangzhou 510000 China)

Abstract: When retrieving energy data, it is very easy to disclose users' privacy information, which leads to the decline of retrieval security. A fast and secure energy data retrieval method based on perceptual hash algorithm is designed. The cloud environment is used to create an energy database, and the known data model and background data model are used to ensure data security. The perceptual hash algorithm is used to map the data into bits of smaller length, and the data similarity is judged by the equal number of bits. Combine the data similarity calculation results to carry out multi granularity text matching, weighted sum the relevance score of retrieval statements and documents, and provide the data with the highest relevance score to users to achieve efficient energy data retrieval. The experimental results show that this method has the advantages of short execution time, low index space cost, high F1 value and good data retrieval security.

Keywords: perceptual hash algorithm; energy data; data retrieval; cloud environment; multi granularity text matching

0 引言

近年来，我国经济增长速度加快，但与此同时，能源供应危机越发严峻[1]。在能源资源开发过程中，能源管理工作是十分重要的。在我国工业能源消耗量是最大的，需要提升能源利用质量，以此降低能源的消耗量。因此，能源综合利用的问题越来越受到人们的关注，而增强能源数据分析，提升能源利用率是减少能耗的重要策略[2]。因此，如何精准查找能源数据是所有能源工作开展的必要条件。

面向能源数据搜索问题，国内外学者从多角度给出不同的解决方案。国外学者们广泛研究雾计算下数据检索模式，但该方法存在一定误判，检索结果不尽如人意。国内对于该问题的研究已经取得了一定的进展。吴飞[3]等人利用全局对抗网络完善自编码器重构流程，采用极值博弈方法获取原始特征，并对特征进行了重构处理。根据特征重构结果使用隐含层对抗网络划分不同的模态数据，降低多模态数据搜索过程中的数据分布差异，通过搭建数据检索架构实现数据检索的目标。冯姣[4]等人在残差神经网络引入注意力机制，通过约束网络映射提高数据检索关键词和检索结果之间的匹配程度，并将检索结果反馈给用户。但上述两种方法在设计过程中，没有考虑网络环境安全性，导致检索时极易泄露用户隐私信息，不能够满足数据安全检索需求。

综上，提出一种基于感知哈希算法的能源数据安全快速检索方法，以期为能源研究领域提供安全可靠的检索工具。

1 云环境下能源数据库构建

为了保证能源数据存储安全，利用云环境创建能源

*基金项目：南方电网数字电网研究院有限公司科研项目（00022000000076144）
收稿日期：2024-01-25

数据库，为后续数据检索提供良好的云安全环境。在系统中包含数据持有者、数据应用者与云服务器三种实体。能源数据库系统结构如图1所示。

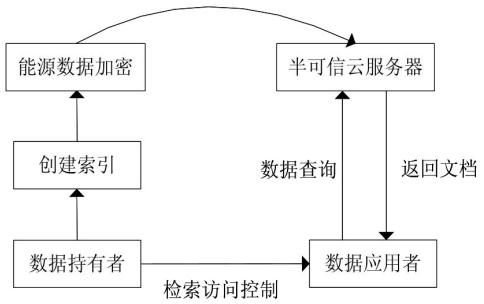


图1 能源数据库系统结构示意图

数据持有者能够将数据与索引加密传输至云服务器内。不仅如此，数据持有者还可以成为数据发起者与接收者[5]。云服务器供应了数据检索所需的超高容量储存空间与计算资源，在接收到数据使用者的合理请求后，云服务器要查找最相近的前个文档，数据持有者可以自行设置值。能源数据库的设计目标即不泄露数据内容基础上，提升数据检索速率与精准度。

假设数据应用者是可信的，云服务器半可信，也就是云服务器能够按照指令完成相关操作，通过分析服务器的数据与索引架构明确用户的数据搜索意图。当前云环境下能源数据库具有已知数据模型与背景数据模型两类安全模型。对于已知数据模型，云服务器能够获取用户保存的信息，也就是加密后的数据集；对于背景数据模型中，云服务器不但能够获取模型的全部数据，还能计算与分析云服务器的所有信息，推测能源检索关键词属性，以此最大程度上保证数据索引安全性。

2 基于感知哈希算法的数据相似度计算

为了准确划分能源数据属性，本文使用感知哈希算法计算数据相似度，进一步锁定检索范围，提升检索精度。邻近数据属性相似度计算是安全检索的核心内容，其性能优劣直接影响了能源数据检索正确率。感知哈希是一种可靠的数据认证处理方案[6]，一般用来计算检索过程中不同数据之间的相似度。和其他相似度计算方法相比，感知哈希算法无需考虑数据类别，计算快捷方便，效率更高。

通常情况下，感知哈希算法有平均哈希与离散余弦变换哈希两种模式[7-8]。平均哈希表示一个数据类型区域内，全部数据值和数据均值之间的比例，以明确数据哈希值；若数据值高于均值，将数据哈希值设定为1，反之是0。但此方法容易受到外部扰动影响，使得能源数据检索速率变慢。离散余弦变换哈希算法使用离散余弦变换替换数据属性均值，让相似度计算更具通用性与鲁棒性。数据属性利用离散余弦变换处理后能获得一个二维系数矩阵，使用该矩阵就能有效处理数据属性。

根据感知哈希算法自身特点，把数据映射为长度很小的比特位，将其运用在数据属性匹配中，利用比特位相等位数评估数据相似度，相等或相近的数据会获得近似的感知哈希值。假设感知哈希函数为 $A(x)$，通过检索目标 C 生成的函数值 d 被称作感知哈希值[9]，计算公式如下：

$$d = A(x)C \tag{1}$$

两个数据的感知哈希间距 b 可利用多距离度量获得[10]：

$$b = d \times E\left(d_i, d_j\right) \tag{2}$$

式中，d_i、d_j 表示两个不同数据的感知哈希值，$E(\cdot)$ 代表距离度量函数。

推算不同数据属性哈希值之间的相似度[11-12]，计算公式如下：

$$h\left(J_1, J_2\right) = \sum_{i=1}^{e}\left(d_i \oplus d_j\right) \tag{3}$$

式中，J_1、J_2 分别表示两个不同的 n 维哈希值，e 表示相似度计算迭代次数，$\oplus$ 表示异或运算。

3 数据安全快速检索

能源数据检索多数使用具有海量标注数据功能的有监督学习方法，但这种方法人工成本较高，为此本文设计一种能源数据安全快速检索模型，结合数据相似度计算结果，从能源数据概念、词语级别、词组级别及三个方面实施多粒度文本匹配，多粒度语义匹配的输入值是用户检索语句，输出为数据检索结果。

3.1 词语级别匹配

针对用户输入的检索内容，计算相似度后利用分词和去停用词把检索内容变换为关键字序列，采用词嵌入策略把关键字序列替换成矩阵 $G_{n\times k}$。将能源数据库内的能源数据集也进行同样操作，变换为矩阵 $B_{m\times k}$。二者用下述公式表示：

$$G_{n\times k} = \left[g_1, g_2, \cdots, g_{n-1}, g_n\right] \tag{4}$$

$$B_{m\times k} = \left[b_1, b_2, \cdots, b_{m-1}, b_m\right] \tag{5}$$

式中，n 表示检索语句分词处理后的关键词数量，m 表示分词处理后的关键词数量，k 表示关键词变换为词嵌入后的矢量维数，g_i，i=1，2，…，n 是检索内容预处理后序列内第 i 个词矢量，b_j，j=1，2，…，m 是能源数据集预处理后序列内第 j 个词矢量[13]。

结合注意力机制对 g_i、b_j 进行处理，获得二者的注意力信号 g_i^{att}、b_j^{att}，记作：

$$g_i^{att} = \sum_{i=1}^{n}\frac{l^{O_{i,j}}}{\sum l^{O_{i,j}}}u_i \tag{6}$$

$$b_j^{att}=\sum_{j=1}^{m}\frac{l^{O_{i,j}}}{\sum l^{O_{i,j}}}s_j \tag{7}$$

式中，$l^{O_{i,j}}$表示注意力因子参数，u_i表示检索内容中原始词矢量均值，s_j表示能源数据集中原始词矢量均值。

假设词语匹配的关联性矩阵为$P_{n\times m}$，对$P_{n\times m}$进行最大池化操作获得矢量r_{word}，令矢量r_{word}中各元素r_{word_i}乘以该词的文档频率实施加权求和，获得检索词语与能源数据集中词语级别的相关性分数$score_{word}$，记作：

$$score_{word}=\sum_{i=1}^{n}idf_i^{query}\cdot r_{word_i} \tag{8}$$

式中，idf_i^{query}表示元素r_{word_i}的逆文档频率。

3.2 词组级别匹配

能源领域文档涉及诸多专业术语与固定句式，不同词语搭配会产生不同含义。如果此类词语记忆被划分成独立词句，就丧失了原本的内涵，使得表达错误。为处理以上问题，对词组级别匹配的关联性矩阵$Q_{n\times m}$实施滑动窗口为3×3的平均池化操作，获得窗口中词语的相关数据，得到池化操作后的关联性矩阵：

$$Q_{i,j}^{ph}=\frac{\left(Q_{i,j}^{word}+Q_{i+1,j}^{word}+Q_{i,j+1}^{word}+Q_{i+1,j+1}^{word}\right)}{9} \tag{9}$$

式中，$Q_{i,j}^{word}$为矩阵中第i行第j列元素，$Q_{i+1,j}^{word}$为矩阵中第$i+1$行第j列元素，$Q_{i,j+1}^{word}$为矩阵中第i行第$j+1$列元素，$Q_{i+1,j+1}^{word}$为矩阵中第$i+1$行第$j+1$列元素。

以行为单位，最大池化操作式(9)的关联性矩阵，获得矢量r_{phrase}，令矢量r_{phrase}内各元素r_{phrase_i}乘以该词组的逆文档频率并加权求和，获得检索内容与能源词组的相关性分数$score_{phrase}$：

$$score_{phrase}=\sum_{i=1}^{n}idf_i^{query}\cdot r_{phrase_i}\cdot Q_{i,j}^{ph} \tag{10}$$

3.3 基于概念匹配的数据检索

对能源数据库中现存的全部能源数据实施分词与去停用词处理后，采用TransE算法把知识词典内的实体数据、实体之间的耦合关联映射至低维矢量空间，记作三元组(v,w,z)，v、z表示不同的实体矢量信息，表示关系矢量，也是两个实体矢量的相似性系数。经过训练调整三个矢量，则下述公式成立：

$$v+w\approx z \tag{11}$$

式中，z与$v+w$的间距足够远。

提取分词与去停用词处理后，通过检索语句出现于知识词典内的信息，即可获取检索语句的实体矢量描述矩阵，记作：

$$G_{\delta\times k}^{ent}=z\times\left[g_1^{ent},g_2^{ent},\cdots,g_{n-1}^{ent},g_n^{ent}\right] \tag{12}$$

式中，δ表示检索语句在知识词典内的实体词语数量，g_i^{ent}，$i=1,2,\cdots,n$表示TransE算法下，序列内第i个实体词矢量的输出值。

假设能源数据概念匹配关联性矩阵为$R_{n\times m}^{ent}$，其中第i行第j列元素为$R_{i,j}^{ent}$，将其定义为：

$$R_{i,j}^{ent}=\cos\left(g_i^{ent}\right) \tag{13}$$

注意力机制下，检索语句实体与文档实体的注意力信号g_i^{entatt}、b_j^{entatt}分别为：

$$g_i^{entatt}=\sum_{i=1}^{n}\frac{R_{i,j}^{ent}}{\sum R_{i,j}^{ent}}u_i \tag{14}$$

$$b_j^{entatt}=\sum_{j=1}^{m}\frac{R_{i,j}^{ent}}{\sum R_{i,j}^{ent}}s_j \tag{15}$$

最大池化操作关联性矩阵$R_{n\times m}^{ent}$，获得矢量r_e，令r_e内各元素r_{e_i}乘以该词的逆文档频率，再进行加权求和，获取检索语句与能源数据概念的相关性分数$score_e$：

$$score_e=\sum_{i=1}^{n}idf_i^{query}\cdot r_{e_i} \tag{16}$$

$$r_{e_i}=\max\left(\tilde{R}_i\right) \tag{17}$$

式中，$\tilde{R}_i$代表相似度指数。

把词语匹配相关性分数$score_{word}$、词组匹配相关性分数$score_{phrase}$、能源数据概念匹配相关性分数$score_e$实施加权求和，输出能源数据检索语句和文档之间的相关性分数$score_f$：

$$score_f=\eta\cdot score_{word}+\beta\cdot score_{phrase}+\lambda\cdot score_e \tag{18}$$

式中，η、β、λ分别表示$score_{word}$、$score_{phrase}$、$score_e$的权重。

结合$score_f$计算结果，将相关性分数最高的数据提供给用户，完成高效率能源数据安全快速检索，可以有效提升用户满意度。

4 实验分析

4.1 实验环境与数据集

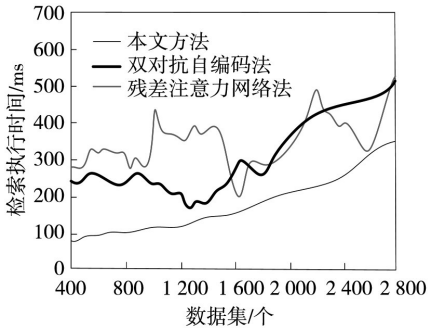
实验过程中使用开源Java语言与分词系统内的Jna-Test_NLPIR开发包，实验平台为MATLAB。实验数据集来自某化工厂能源数据库，其中涵盖11 215个文档数据集，每篇文档内平均具备135个关键词。为验证方法的高效性与精准性，将文献[3]方法、文献[4]方法作为对比方法。实验指标为检索效率、空间开销代价、F1值、检索功能安全性。

4.2 实验结果

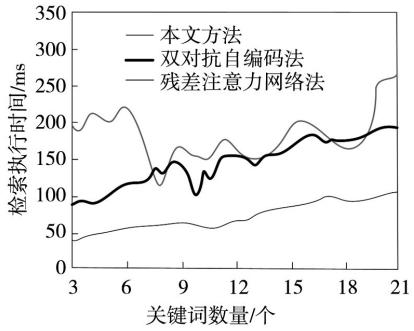
检索效率对比实验中，将数据规模与关键词数量拟作可变参数，实验结果如图2所示。

分析图2可知，无论是不同数据规模还是不同关键词数量下，文中方法的检索执行时间都是最短的，说明该方法的检索效率更优。两种文献方法检索执行时间曲线都

出现大幅度波动，计算稳定性较差。出现此种现象的原因在于，文献[3]方法使用的双对抗自编码法检索机制会伴随数据规模的改变调节预设字典，计算代价也随之提升；文献[4]方法使用的残差注意力网络法搜索能源数据时要自适应调节位数结构，也造成了极高的时间损失。而文中方法使用感知哈希算法可有效分析检索信息相似度，使检索效率得到质的提升。



(a) 不同数据规模下检索时间对比



(b) 不同关键词数量下检索时间对比

图2 能源数据检索执行时间对比

检索方法的空间开销可体现在数据在网络内的传输与储存代价方面，索引空间开销代价越小，证明方法的使用寿命越长。设定不同的数据规模进行实验探究，随机抽取文档数据集，对比不同方法生成的能源数据索引空间代价，结果如图3所示。

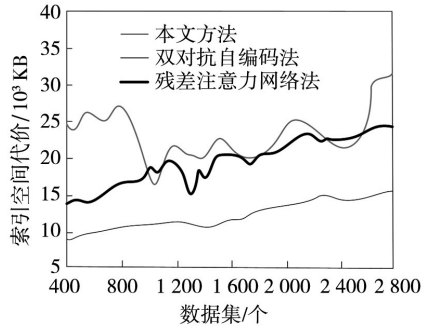


图3 检索索引空间代价对比

从图3看出，同一实验环境下，文中方法的索引空间开销代价最小，而两个文献方法伴随数据规模的增加开销会大幅度提升，用户端无法承受多个数据拥有者使用选择性能源数据检索，形成网络拥堵，实际应用效果差。

为验证三种方法的检索准确性，采用F1值指标综合评估其可靠性。F1值为精确率与查全率的调和均值，最大值是1，最小值是0。具体的计算公式如下。

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{19}$$

式中，P表示精确率，R表示查全率。

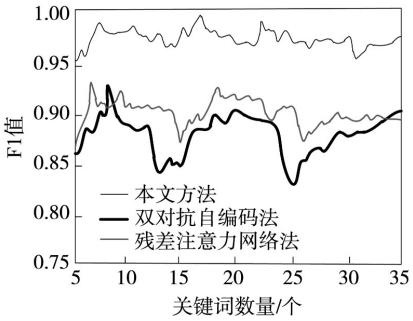


图4 能源数据检索F1值对比

设定检索关键词数量为35个，实验结果如图4所示。

从图4中能够看出，随着关键词数量的增加，两个文献方法的F1值均小于文中方法。文中方法的F1值曲线较为稳定，且始终保持在0.95以上。这也表明了文中方法检索精度要显著优于两个对照组，能源数据检索指向性更强，让用户得到满意的检索内容。

下面从安全信道、权限控制、攻击抵御等方面检验数据检索安全性，检验结果如表1所示。

表1 检索方法功能比较

方法功能	无需安全信道	权限控制	抵御在线关键词猜测攻击	抵御离线关键词猜测攻击
本文方法	是	是	是	是
双对抗自编码	是	否	否	是
残差注意力网络	否	否	是	否

综合表1的安全功能分析结果看出，文中方法不需要浪费额外时间创建安全信道，在离线与在线状态下均能很好地抵抗恶意攻击，支持多用户访问控制，能够满足数据检索对于安全性的要求。

5 结束语

随着能源数据数量的日益增多，数据检索逐步成为一项具有挑战性的工作。针对能源数据类型多样化特征，设计基于感知哈希算法的能源数据安全快速检索方法。运用云计算技术搭建能源数据库，能够最大程度上保证数据安全性。采用感知哈希算法计算检索信息与能源数据库信息的相似度，把能源数据划分为能源数据概念、词语级别与词组级别三个层面，利用多粒度文本匹配获得检索文本相关性分数，将相关性数最高的结果提供

(下转第98页)