基于 VSM 算法的集中性运维信息智能检索方法

张海涛
（深圳供电局有限公司，广东 深圳 518000）

摘 要 由于网络页数的覆盖量急剧增加，从海量运维信息中获取真正的所需信息非常困难，检索后的数据结果与实际需求不符，难以保证较高的准确率。为了提高运维信息检索的准确性，本研究提出基于 VSM（Vector Space Model）算法的集中性运维信息智能检索方法。通过模糊聚类法对文本进行空间向量转换设定，对集中性运维信息进行分类；基于 VSM 算法计算信息相似度，对语句进行分词权重统计；定义集中性运维信息概念空间，扩展信息的语义能力，完成信息的智能匹配检索。实验结果表明：在指标为 P@K500 组的检索结果中，本文方法的准确率为 0.95，且本文方法在单项指标 MAP 的准确率也为 0.95，提高了运维信息检索的准确性，具有实际应用效果。

关键词 VSM 算法 集中性运维信息 检索方法 模糊聚类
中图分类号：TP311 **文献标识码**：A **文章编号**：1007-0745（2022）10-0019-03

在互联网技术不断发展的进程中，网络信息和网络用户的数据不断激增，互联网也从信息发送和接收端口，正逐渐转换为信息流的传输渠道。在大数据网络时间潮流中，仅通过传统的信息检索方法，难以支撑运维数据这种多源异构超文本数据的搜索和应用，因此需要研究一种更加快速且智能的信息检索方式 [1-3]。集中性运维信息管理方法解决了这一问题，但随着网络页数的覆盖量急剧增加，用户发现越来越难以应用该检索方法，帮助自己找到所需要的运维信息数据。随着计算机技术和互联网的进一步普及，为更好地满足用户检索要求，该领域的研究者改进了集中性运维信息管理检索方法，提出了面向特定运维主体的信息检索技术，即在给定的信息内容中，有选择性地从网络中搜索出边缘信息，提高了检索效率。但这种方法只在所要求的范围内进行针对性检索，虽然在一定程度上满足了区域内的检索，但若不事先给定预设运维数据库，则仍然难以真实地描述用户检索需求。另外该方法在大量的数据信息中，也不能精准有效地检索出所有相关信息，而放弃不相关信息，存在数据信息判断不准确的问题。现有信息检索方法想要从海量运维信息中准确获取真正的所需信息，依旧非常困难。向量空间模型 VSM 可以考虑词频之间的有效关系，以权值计算的方法，对具有相似性的文本进行聚类。为了提高集中性运维信息智能检索的准确性，本文以 VSM 算法为基础，研究基于 VSM 算法的集中性运维信息智能检索方法，为信息的同步获取提供理论支持。

1 基于模糊聚类法分类集中性运维信息

对集中性运维信息进行检索，主要是对其特征内容进行选择和设定。其中，信息选择是以特征相似模糊聚类方式对集中性运维信息进行检索需求的特征提取，在聚类组成后进行分类处理。

对文本信息进行分类处理主要分为预处理和聚类两个部分。在预处理过程中，主要是将即将分类的信息，以中文分词的形式进行特征选择，并将其映射至空间向量模型中。通过文本信息预处理，将待分类的本文信息按照不同的向量形式，进行初始文本集合的若干分类。

将文本表示为以特征权值项的维度形式，对其进行简化选择，经过模糊聚类的方式，对文本进行空间向量转换设定 [4-5]。

根据内容所示，对文本信息中的任意一个文本进行设定，将 V_I 对应在选择的向量中，表示为：

$$B(V_I)=(N_1(V_I),N_2(V_I),...,N_M(V_I)) \quad (1)$$

公式中：特征向量权值表示为 $B(V_I)$。向量个数表示为 M，其中 M=1,2,...，主要为文本集合中进行特征向量值计算时，所有特征项的总数。$N_M(V_I)$ 表示文本 V_I 在 I 维度中的数值，也是在第 I 个特征项中，文本所计算出的权值。

由于模糊聚类算法属于无监督学习形式，即可以不用进行预先的样本训练，直接以模糊相似聚类的形式对预处理后的数据进行规则分类，并按照一定的规则进行类和簇的组合。正常分类情况下，每个类中的相似度需大于类间的相似度。数据分类完成后，要在其具备准确性的前提下，对其相似度进行计算，以准确率和召回率为指标，表示为：

$$q=\frac{w_1}{w_1+w_2} \tag{2}$$

$$r=\frac{w_1}{w_1+w_3} \tag{3}$$

式中：准确率用 q 来表示。召回率用 r 来表示。在聚类结果为该类的数据中，w_1 表示真正属于该类的文本数量，w_2 表示不属于该类的文本数量。而当聚类完成后，其结果为非该类的数据集合时，则 w_3 表示真实属于非类的数据量，而不在其类型的文本数量为 w_4，可不计入计算内。在利用准确性和召回率完成数据对比分类后，采用 VSM 算法计算信息的相似度。

2 基于 VSM 算法计算信息相似度

基于分类后的数据利用 VSM 算法进行集中性运维信息的相似度计算，对语句进行分词权重统计，以扩展语义使其自身具有匹配能力，为信息的智能检索奠定基础。

对文本信息进行统计，若在两组文本中出现相同的词汇较少，或者从未出现较为相似的词汇，则其相似值可能会很低，甚至为 0。将余弦系数计算与 VSM 算法进行融合，计算所有语句中所有词汇的相似度，并利用概念对应的距离形式，设定检索信息与需求信息之间的关系。

通过 VSM 模型计算向量空间中的内容，用以描述信息中的具体内容，将词转变为词向量，从而进行余弦相似度的计算，当两个向量的余弦夹角值越小，说明两个文本之间更为相似，反之则存在很大的不同之处[6-7]。假设需要检索的运维信息中，含有 A_1 和 A_2 两组语句，利用 VSM 计算方式，具体步骤如下：

对 A_1 和 A_2 两组语句进行分词处理，其中 $A_1=\{S_1,S_2,...,S_D\}$、$A_2=\{F_1,F_2,...,F_G\}$。当 A_1 和 A_2 语句分别完成分词后，共同建立为一个数据集合 H。将 A_1 和 A_2 中出现的所有词汇，进行合并处理，即 $H=\{S_1,S_2,...,S_D,F_1,F_2,...,F_G\}$。统计 A_1 和 A_2 两个语句中，每个词汇在集合 H 中，出现的次数，即可作为每组词汇的权重，能够完成本文数据的特征向量值。

将 A_1 和 A_2 中每个词汇的权重进行汇总，定义 A_1 中的文本特征向量为 $J_{KS}=(Z_{S,1},Z_{S,2},...,Z_{S,D})$ 和 $J_{KF}=(Z_{F,1},Z_{F,2},...,Z_{F,D})$，两个特征向量空间夹角为 β，则：

$$X=J_{KS}\times J_{KF}=\sum_{C=1}^{D}Z_{S,C}Z_{F,C} \tag{4}$$

公式中：两组向量的内积为 X。向量个数为 $C=(1,2,...,G)$。利用余弦系数进行相似度求解，如下：

$$SIM_{IJVSM}=A\left(K_S,K_F\right)=COS\beta=\frac{\sum_{C=1}^{D}Z_{S,C}\times Z_{F,C}}{\sqrt{\left(\sum_{C=1}^{D}Z_{S,C}{}^2\right)\left(\sum_{C=1}^{D}Z_{F,C}{}^2\right)}} \tag{5}$$

公式中：对两个文本之间的相似度，用 SIM_{IJVSM} 来表示。在 VSM 算法中会出现高频词汇和低频词汇，因此对本文中词汇权重的计算尤为重要，通过上述方法获得权重汇总，将词汇中的奇异值进行剔除，寻找到集中性信息的相似度。通过集中性运维信息的相似度计算，以扩展语义方法，进行文本信息的内容扩充，使其自身具有匹配能力，完成信息的智能检索。

3 实现智能匹配信息检索

用户进行集中性运维信息的检索，需要通过自然语言检索进行表达。在自然语言检索下，直接以分词和语义进行分析，完成概念之间的逻辑关系转换，形成新的逻辑关系概念集合，即用户检索概念空间集。一般情况下，对信息检索的整个过程，即是在概念空间里，对运维信息进行检索匹配的过程。而检索中难免会出现失败现象，为避免用户信息检索中出现失误，需要优化和拓展信息所处的语言空间集合，对用户需求充分表达，拓展语义能力，处理运维信息检索过程中的缺陷问题，实现智能匹配信息检索。

以扩展语义能力为基础，利用 ONTOLOGY 的关联关系，对信息所处的空间集合进行优化和拓展。在原始空间为 $\{Q,W\}$ 的前提下，其中 Q 为用户查询过程中的检索项目集合，W 为概念逻辑关系的集合。对其进行语义扩展优化，主要分为两个部分。首先是将 Q 中关于用户的概念，以 ONTOLOGY 中的概念定义，映射为新的概念集合 E。其次，在语义关系和原始逻辑中，利用 W 对 E 进行规则转换，确定新空间中概念之间的逻辑属性，形成新的一个隶属概念空间。

对于第一步中的概念假设问题，即在 Q 中设置为 $(Q_1...Q_N)$ 种概念项，对于每一组项目进行 ONTOLOGY 内部的逻辑匹配，包括同类型词汇以及词条的变化形式。在每次转换成功后，均可产生一组匹配记录 $(Q_I...E_I)$，其中 Q_I 为 Q 中的某一个概念项目检索，E_I 是 ONTOLOGY 中能够与 Q_I 相匹配的概念。而由于 Q_I 可能会匹配出多个 E_I，因此 Q_I 可以拥有多条运维信息记录，以此在所有的 E_I 总计中生成新概念集合 E。至此完成用户检索概念空间集优化拓展，实现智能检索方法设计。

4 实验测试分析

4.1 实验数据准备

采用 DBLP 数据集中的一个子集代表海量运维信息，其中包含有 AUTHOR 数据表、PAPER 数据表、WRIRE 数据表和 CITE 表。每种数据表中的信息记录分别为 290000 条、450000 条、900000 条、120000 条。通过对 DBLP 数据集中抽取，构造其检索对象的级别关系模式。

在数据子集中的数据表，所属关系为互通形式，符合运维数据关系特征。在处理后对数据中的检索对象进行统计，其中论文对象共计 440000 组、作者对象共计 290000 组，最终形成的检索对象模式图的节点数为 740000 个。基于以上数据，对测试的检索方法进行效果论证。

4.2 选择评估指标

信息检索的目的是通过一系列相关操作，找到所需要的数据信息。为验证本文方法的有效性，对设计的检索方法进行评估。由于检索的目的主要是尽可能多地检索出所需信息，并且排除掉不相关信息。选择 P@K 指标和 MAP 指标进行评价：

1.P@K 指标：表示准确率的变形，是指在检索结果中占据前 K 个结果的准确率。

2.MAP 指标：反映检索方法在全部数据检索过程中的单项指标，为平均准确率。

通过选择的两组指标，验证本文方法与传统方法的检索效果。

4.3 对比检索效果

按照选择的两组指标，首先进行准确率的变形测试，设定指标为 P@K100、P@K200、P@K300、P@K400、P@K500。每个指标共进行 10 组测试，对检索记录的结果均进行登记后，统计其准确率平均值。

本文的检索方法准确率指数，均在传统方法之上。当指标为 P@K500 时，本文检索方法的准确度为 0.95，较比传统方法高出 0.35。

在此基础上，针对 P@K 指标测试情况，分别对比 P@K100、P@K200、P@K300、P@K400、P@K500 的 10 组查询 MAP 值。

传统方法在初始阶段的准确率与本文方法较为一致，但随着测试指标的增加，本文检索方法更具有优势，其中仍以 P@K500 时作为参考，本文方法的 MAP 值为 0.95，传统方法为 0.65，说明本文方法更加有效。

5 结语

信息检索在数据应用中具有重要作用，随着互联网信息的快速融合，为保证用户能够完成所需信息的准确检索，本文以 VSM 算法为基础，设计了集中性运维信息的智能检索方法。在实验论证下，本文方法取得了一定优势，无论是 MAP 指标和 P@K 指标均可以保证较高的准确率。但由于此次时间有限，在研究过程中没有对数据的吞吐情况和丢失情况进行分析，存在不足之处。后续研究中会进一步进行分析，为实现高效能的信息检索提供理论支持。

参考文献：

[1] 容秀婵，邹湘军，李承恩，等．基于数据驱动的虚拟场景搭建及模型检索优化方法 [J]. 中国农机化学报，2022,43(08):128-135.

[2] 潘华峰，王春玲，毋涛．结合哈希网络和敏感散列的图像检索推荐研究 [J]. 计算机技术与发展，2022,32(07):173-178.

[3] 孔建．跨库文献检索方法应用于科研文献系统中的研究与分析 [J]. 黑龙江科学，2022,13(12):38-40.

[4] 郑庆荣，赵建立，盛明，等．基于知识图谱的全链路数据自动检索方法 [J]. 自动化与仪器仪表，2022(05):170-173.

[5] 童林，官铮，杨文韬，等．潜在低秩表示下 VSM 联合 PCNN 的红外与可见光图像融合 [J]. 国外电子测量技术，2021,40(10):84-90.

[6] 任锦标．基于数据仓库及决策树算法的电网事故事件信息智能检索方法研究 [J]. 集成电路应用，2019,36(12):86-87.

[7] 袁仁进，陈刚，李锋，等．基于 VSM 和 Bisecting K-means 聚类的新闻推荐方法 [J]. 北京邮电大学学报，2019,42(01):114-119.