Computer Science and Application 计算机科学与应用, 2024, 14(2), 360-370
Published Online February 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.142037

Hans汉斯

智慧医疗资讯个性化服务平台医学感知

刘彦宏，郑琳那，王　榕，曲金鑫，毛　云，崔永瑞

大连民族大学计算机科学与工程学院，辽宁 大连

收稿日期：2024年1月23日；录用日期：2024年2月22日；发布日期：2024年2月29日

摘　要

本文旨在研究一款智慧医疗资讯个性化服务平台——医学感知。传统的医疗资讯平台并不具备人工智能的相关算法帮助医生或者病人更准确、高效地获取医疗建议。本平台融合了三种不同的自然语言处理技术TF-IDF、Word2Vec、BERT等机器学习算法，通过实验比较出三者对于医疗信息匹配的不同特点从而为用户打造一个更加智能、更具人性化的医疗资讯服务平台。该平台基于Python语言及Django框架和MySQL数据库进行搭建，通过Requests和Beautiful Soup库实现对医疗数据的采集。

关键词

医学感知，自然语言处理，TF-IDF，Word2Vec，BERT，Django框架，数据挖掘

Medical Perception of Intelligent Medical Information Personalized Service Platform

Yanhong Liu, Linna Zheng, Rong Wang, Jinxin Qu, Yun Mao, Yongrui Cui

School of Computer Science and Engineering, Dalian Minzu University, Dalian Liaoning

Received: Jan. 23rd, 2024; accepted: Feb. 22nd, 2024; published: Feb. 29th, 2024

Abstract

The purpose of this paper is to introduce a smart medical information personalized service platform—medical perception. The previous medical information platforms did not have artificial intelligence algorithms to help doctors or patients obtain medical advice more accurately and efficiently. The platform integrates three different natural language processing technologies, including TF-IDF, Word2Vec, and BERT machine learning algorithms. Through experiments, the differences in medical information matching characteristics of the three are compared in order to create a more intelligent and more personalized medical information service platform for users. This platform is built based on the Python language, Django framework, and MySQL database. It collects medical data through Requests and Beautiful Soup libraries.

文章引用: 刘彦宏, 郑琳那, 王榕, 曲金鑫, 毛云, 崔永瑞. 智慧医疗资讯个性化服务平台医学感知[J]. 计算机科学与应用, 2024, 14(2): 360-370. DOI: 10.12677/csa.2024.142037

Keywords

Medical Perception, Natural Language Processing (NLP), TF-IDF, Word2Vec, BERT, Django Framework, Data Mining

1. 引言

近年来，随着人工智能技术的不断发展，机器学习和深度学习算法在医疗领域的应用[1] [2]日益广泛，它们被用于医学图像和信号处理、计算机辅助检测与诊断、临床决策支持、医疗信息挖掘和检索等多个方面。通过结合自然语言处理技术可以提供精准的医疗资讯[3]，能够帮助医疗从业者和患者更准确、高效地获取医疗建议。

在过去的几年中，文本表示学习方法在自然语言处理领域取得了巨大的进展。其中，词频－逆文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)词频统计[4]、单词映射向量[5] (word to vector, Word2Vec)和双向编码器转换器[6] (Bidirectional Encoder Representations from Transformers, BERT)是三种经典的文本表示学习方法，它们在文本相似度计算和语义理解方面表现出色。

本文旨在通过研究 TF-IDF 词频统计、Word2Vec 和 BERT 算法，比较它们在医疗方案推荐中的应用效果。通过分析和对比这三种方法在文本表示和相似度计算方面的性能，旨在为医疗资讯平台提供更准确和个性化的解决方案。同时，将结合了这三种算法的智慧医疗资讯平台和传统的医疗资讯平台的资讯结果作比较，从而为医生和患者打造出更好的医疗体验和服务。

2. 相关工作

2.1. 系统介绍

医疗方案推荐系统，旨在帮助医疗从业者和患者更准确、高效地获取医疗建议。采用了三种不同的算法进行实验和比较，包括 TF-IDF 词频统计分析、Word2Vec 和 Bert。这些算法在文本表示和相似度计算方面表现出色，可以为医疗决策支持系统提供更可靠的技术支持。使用 Django 框架搭建了一个测试页面，以便进行算法性能测试和用户反馈收集。核心算法使用了 NPL 技术，基于 Python 语言及 Django 框架和 MySQL 数据库搭建。我们还使用 Requests 和 BeautifulSoup 库对 39 健康网的病例信息进行爬取，以获得更准确的数据。通过本系统，患者可以更轻松地获取到符合其需求的准确医疗方案，从而提高医疗服务的质量和效率。同时，医疗从业者也可以更快速地制定出符合患者需求的医疗方案，提高工作效率和医疗质量。我们将继续优化和完善这个系统，为医疗领域的发展做出贡献。

2.2. TF-IDF 词频统计分析

TF-IDF 是一种常用于信息检索与文本挖掘的统计算法，用于衡量一个词对于一个文档集合的重要程度。TF (Term Frequency, TF)指的是一个词在文档中出现的频率，可以简单地计算为该词在文档中出现的次数除以文档的总词数，即词频，衡量了一个词在文档中的相对重要程度，认为在文档中频繁出现的词更为重要。IDF (Inverse Document Frequency, IDF)指的是一个词在整个文档集合中的稀有程度即逆文档频率，它通过计算文档集合中包含该词的文档数，并将总文档数除以该值取对数，然后取倒数得到。IDF

衡量了一个词的普遍重要程度，认为在整个文档集合中较少出现的词更为重要。

TF-IDF 算法是经典的关键词提取方法，算法可分为文本预处理模块、权重计算模块以及提取关键词模块三大模块。首先把输入的文本 c 进行分词等预处理操作，然后得到由特征项组成的集合 $c_i=\left(x_1,\cdots,x_j,\cdots,x_C\right)$，根据每个特正项在文本中的重要性赋予其权重 ω_j，再通过 TF 和 IDF 来计算权重。最后再按照去找你红大小对特征词 x_j 排序，选择前 n 个词作为 u 文本最终关键词。TF-IDF 计算公式如下：

$$\text{TF-IDF}\left(c_i,x_j\right)=\text{TF}\left(c_i,x_j\right)\times\text{IDF}\left(x_j\right)=\text{TF}\left(c_i,x_j\right)\times\log\left(\frac{K}{\text{DF}\left(x_j\right)}\right) \tag{1}$$

其中 $\text{TF}\left(c_i,x_j\right)$ 表示 x_j 在当前文本 c_i 出现的次数，$\text{DF}\left(x_j\right)$ 表示数据集中出现 x_j 的个数，K 为总文本数。

TF-IDF 值越大，表示该词在文档集合中的重要性越高，优点是简单有效，能够帮助识别和提取关键词，能够突出文档中重要的词语，并过滤掉常见的词语，从而提高信息检索和文本挖掘的准确性。

TF-IDF 算法的基本流程包括文本预处理、构建文档－词项矩阵、计算词项的 TF 值、计算逆文档频率 IDF 值、文档相似度计算以及排序和检索。首先对文本进行预处理，去除标点符号、停用词和数字，并将文本转为小写字母。然后构建文档－词项矩阵，其中每行代表一个文档，每列代表一个词项，矩阵元素表示词项在文档中的出现频率。接下来计算词项的 TF 值，即词项在文档中的出现频率。然后计算逆文档频率 IDF 值，反映词项在整个文集中的重要性。使用计算得到的 TF-IDF 值，可以计算文档之间的相似度，常用的相似度度量是余弦相似度。最后，根据文档相似度进行排序和检索，选择相似度最高的文档作为检索结果。TF-IDF 算法在信息检索中具有广泛应用，能够提供准确的文本相似度度量和排序功能。

TF-IDF 算法是一种常用的文本信息检索方法，可用于搜索引擎、文档分类、文档聚类等各种自然语言处理任务。它通过考虑词项在文档中的频率和在整个文集中的重要性来识别相关性高的文档。

2.3. BERT

BERT 是一种基于 Transformer 模型的预训练语言表示模型，由 Google 在 2018 年提出。它通过在大规模无标签语料上进行预训练，学习出通用的文本表示，然后在各种下游自然语言处理任务上进行微调，以提高任务性能。

BERT 模型的输入由字向量、句向量和位置向量三部分组成，其中包含特殊的分类标记([CLS])作为起始 Token 和特殊令牌([SEP])作为结束 Token。[CLS]表示用于分类模型的特征，对于非分类模型可以省略；[SEP]表示输入语料中用于分隔两个句子的符号。当输入由两个句子组成时，我们在第一个句子的开头加入[CLS]符号，并在第一个句子的末尾和第二个句子的末尾分别添加[SEP]符号。BERT 模型的输入表示如图 1 所示。

BERT 的主要创新之处在于采用了双向训练机制，即同时利用上下文的左侧和右侧信息来预测当前词语。这种双向训练方式使得 BERT 能够更好地理解词语之间的关系和上下文语境，从而提取出更丰富的语义信息。

BERT 模型结构由多层 Transformer 编码器组成，其中包含自注意力机制和前馈神经网络。自注意力机制能够在不同位置的单词之间建立关联，并捕捉它们之间的依赖关系，而前馈神经网络则能够对每个位置的词语进行非线性变换。

BERT 模型的预训练任务包括掩码语言建模(Masked Language Modeling, MLM)和下一句预测(Next Sentence Prediction, NSP)。在 MLM 任务中，部分输入词语被随机掩盖，模型需要根据上下文预测被掩盖的词语是什么。在 NSP 任务中，模型需要判断两个句子是否是连续的。

BERT 算法用于计算文本相似度的基本流程包括准备数据、文本编码、预训练 BERT 模型、微调模型和计算相似度。首先，准备文本数据，包括需要比较相似度的文本对。然后，将文本转化为数字表示，

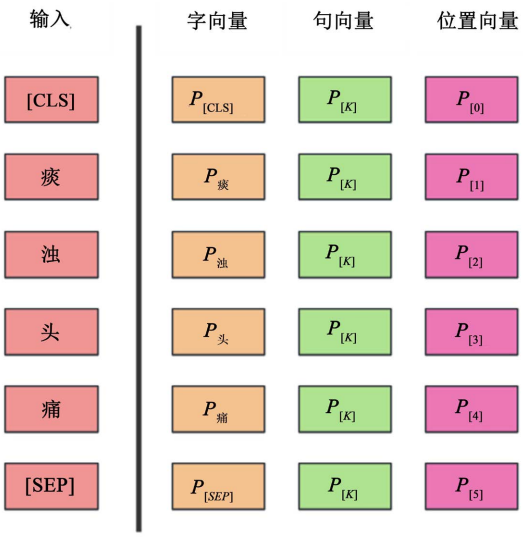

Figure 1. Input representation of BERT model
图 1. BERT 模型的输入表示

使用 WordPiece 嵌套单词表示进行编码。接下来，使用在大规模文本数据上预训练的 BERT 模型，学习文本之间的语义关系。对于文本相似度任务，需要微调预训练的 BERT 模型，将文本对的编码作为输入，通过一个分类或回归层预测相似度得分。最后，使用微调后的 BERT 模型计算文本对之间的相似度得分，得分范围通常为 0 到 1，较高的得分表示更相似，较低的得分表示更不相似。BERT 算法在自然语言处理中广泛应用，能够提供准确的文本相似度计算能力。

BERT 算法在问答系统、语义搜索、文档相似性比较等文本相似度相关任务中具有广泛应用。它通过理解文本的上下文信息，能够更好地捕捉语义含义，从而在文本相似度计算中表现出色。通过预训练和微调，BERT 模型适应各种下游任务，提升了自然语言处理的性能。

2.4. Word2Vec

Word2Vec 是一种基于神经网络的词嵌入方法，用于将词语映射到低维向量空间中。它是自然语言处理领域中广泛应用的模型之一，旨在捕捉词语之间的语义关系。Word2Vec 模型有 CBOW (Continuous Bag of Words)和 Skip-gram 两种主要的实现方式。

CBOW 模型的目标是通过上下文词来预测中心词，即用 $x_{n-2}x_{n-2}, x_{n-1}, x_{n+1}, x_{n+2}$ 去预测 x_n。它通过将上下文词向量求和并输入到一个神经网络中来预测中心词，CBOW 模型图如图 2 所示。CBOW 模型适用于较小的数据集和常见的词汇，其训练输入是某一个特征词的上下文相关的词对应的词向量，而输出就是这特定的一个词的词向量。

Skip-gram 模型和 CBOW 的思路是反着来的，即输入是特定的一个词的词向量，而输出是特定词对应的上下文词向量，Skip-gram 模型图如图 3 所示。它在训练过程中通过最大化中心词和上下文词之间的相似性来学习词向量。Skip-gram 模型适用于较大的数据集和较少出现的词汇。Skip-gram 以当前词预测其上下文词汇，即用 x_n 去预测 $x_{n-2}x_{n-2}, x_{n-1}, x_{n+1}, x_{n+2}$。

Word2Vec 算法用于计算文本相似度。首先，需要准备大规模的文本语料库，这些文本可以是单词、句子或文档的集合。这些文本数据将用于训练 Word2Vec 模型。接下来对文本进行预处理，包括分词、去除停用词、标点符号和数字，将文本转换为小写字母等，以减小噪音和统一文本格式。然后使用预处理的文本语料库训练 Word2Vec 模型。Word2Vec 模型有两个主要变种：Skip-gram 和 CBOW。这些模型

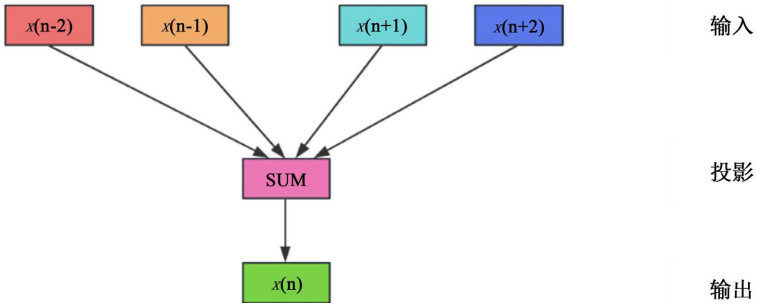

Figure 2. CBOW model diagram
图 2. CBOW 模型图

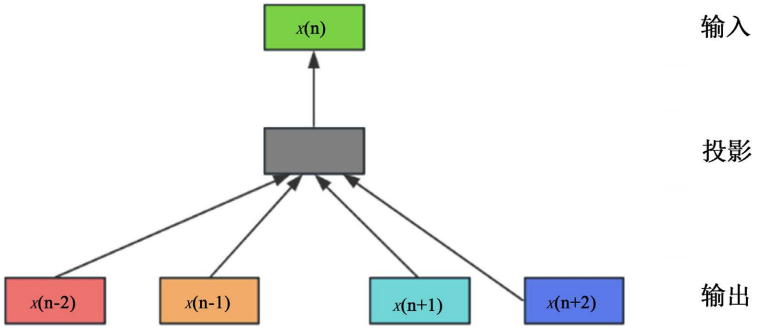

Figure 3. Skip-gram model diagram
图 3. Skip-gram 模型图

通过学习词向量来捕获单词之间的语义关系。在训练过程中，模型将尝试预测上下文词汇或目标词汇，以更新词向量的权重。一旦模型训练完成，每个单词都会有一个对应的词向量，表示单词的语义信息。要计算文本相似度，首先需要将文本转换为向量表示。这可以通过将文本中的所有单词的词向量进行平均、求和或通过其他方法来实现。一旦文本被表示为向量，可以使用不同的相似度度量方法，例如余弦相似度，来计算文本之间的相似度。

Word2Vec 模型是基于浅层神经网络结构的词嵌入模型，通过训练大规模的文本语料库来学习词向量。它能够捕捉词语之间的语义关系，特别适用于计算文本相似度的任务，如信息检索、文档聚类、情感分析和推荐系统等。Word2Vec 在自然语言处理领域被广泛应用，其优势在于能够提供准确的语义表示和较好的相似度计算性能。

3. 实验过程

3.1. 数据采集和预处理

如表 1 所示，数据采集：使用 Requests 和 BeautifulSoup 库对 39 健康网的病例信息进行爬取。通过爬虫程序自动化地获取病例名称、病症描述和治疗方法等相关信息，并保存到本地文件或数据库中。

数据清洗：对采集到的数据进行清洗，去除不必要的 HTML 标签、特殊字符和噪声数据，以确保数据的干净和一致性。

数据预处理：对清洗后的数据进行预处理，包括以下步骤：

分词：将病例描述进行分词操作，将长句子划分为短语或单词，以便后续的文本表示和相似度计算。

去停用词：去除常见的停用词，如“的”、“是”、“在”等，这些词对文本表示和相似度计算没有实质性的贡献。

Table 1. Medical information collection data table
表 1. 医疗信息采集数据表

序号	病例名称	…	详细症状
1	冠心病	…	冠心病症状典型症状：胸骨后的压榨感……
2	慢性胃炎	…	慢性胃炎症状典型症状：上腹疼痛和饱胀……
3	肩周炎	…	肩周炎症状早期症状：肩部呈阵发性疼痛……
…	……	…	……
1192	慢性咽炎	…	慢性咽炎症状典型症状：一般无明显……

词干提取或词形还原：对词汇进行词干提取或词形还原，将单词还原为其原始形式，以减少词汇的冗余和多样性。

构建语料库：将预处理后的数据构建成一个语料库，用于训练和比较不同的算法模型。

3.2. 实验结果分析

3.2.1. 症状输入

如图 4 所示，症状输入模块是医疗方案推荐系统的核心功能之一。用户可以通过该模块输入症状信息，系统将根据用户输入的症状，推荐相应的医疗方案。在症状输入模块中，用户可以输入一个或多个症状，例如头痛、发热、咳嗽等。用户还可以选择症状的严重程度和持续时间等信息，以提供更准确的推荐结果。该模块采用了先进的自然语言处理技术，结合 TF-IDF 词频统计分析、Word2Vec 和 Bert 等算法，对用户输入的症状进行文本表示和相似度计算。系统会将用户输入的症状与已有的医疗方案进行比对和匹配，找出与用户输入最相近的医疗方案，并将其推荐给用户。用户可以通过简单直观的界面输入症状信息，系统将迅速生成相关的医疗建议，帮助用户更准确地理解和应对症状。这将为患者提供方便快捷的就医指南，同时也能帮助医疗从业者更高效地制定治疗方案，提升医疗服务的质量和效率。通过不断优化和完善症状输入模块，我们将为用户提供更精准、个性化的医疗方案推荐，为医疗领域的发展做出积极贡献。

Figure 4. Symptoms entry display
图 4. 症状输入显示图

3.2.2. 疾病诊断

如图 5 所示，疾病诊断模块是医疗方案推荐系统的重要组成部分，旨在帮助医疗从业者和患者准确诊断疾病。该模块采用了先进的自然语言处理技术和数据挖掘算法，结合已有的症状信息和医疗知识库，

实现自动化的疾病诊断和推荐。在疾病诊断模块中，用户可以输入自己或他人的症状信息，包括具体症状、症状的严重程度和持续时间等。系统将根据用户输入的症状，进行文本匹配和相似度计算，找出与用户输入最相近的医疗知识库中的疾病信息。为了提高诊断准确性，系统会综合考虑多个因素，如症状的权重、症状之间的关联性以及疾病的流行病学特征等。通过对大量的病例数据进行分析和学习，系统可以不断优化和更新模型，提高疾病诊断的准确性和精确度。在诊断完成后，系统将向用户推荐相关的医疗方案，如药物治疗、手术建议、康复计划等。同时，系统还可以提供相关疾病的基本信息、预防措施和常见的并发症等，以帮助用户更好地了解和管理自己的健康状况。通过疾病诊断模块，医疗从业者可以快速准确地诊断疾病，为患者提供个性化的治疗方案和咨询服务。患者也可以在家中或移动设备上方便地获取到准确的疾病诊断结果和医疗建议，提高就医效率和医疗质量。

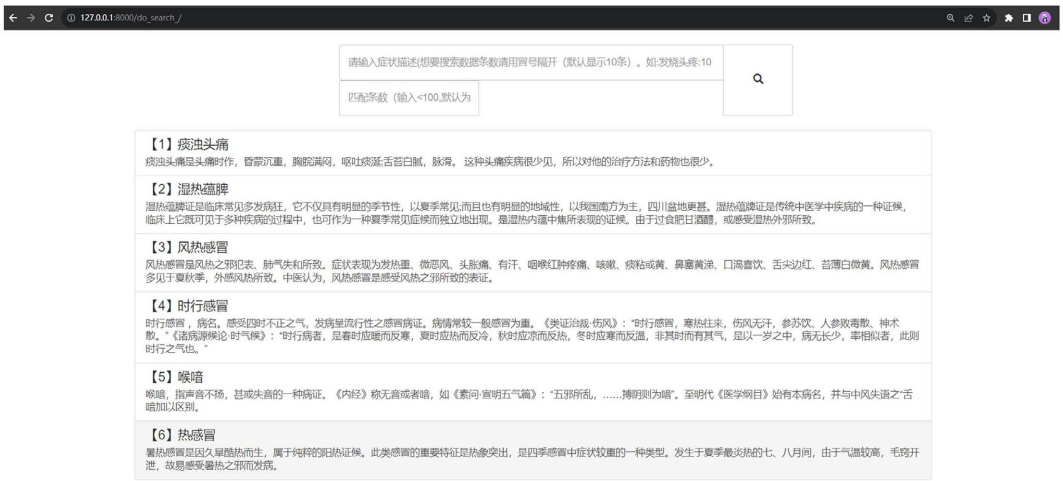

Figure 5. Disease diagnosis display diagram
图 5. 疾病诊断显示图

3.2.3. 治疗建议

如图 6 所示，治疗建议模块是医疗方案推荐系统的重要组成部分，旨在为患者提供个性化的治疗建议。根据用户输入的症状和诊断结果，系统可以为患者提供以下治疗建议：

药物治疗：系统会根据疾病的特点和严重程度，推荐适合的药物治疗方案。包括药物名称、用法、用量、注意事项等信息，以便患者正确使用药物。

手术建议：对于需要手术治疗的疾病，系统会提供手术建议。包括手术的适应症、手术方法、手术风险、术后护理等信息，以帮助患者做出明智的决策。

康复计划：对于康复治疗的疾病，系统会提供相应的康复计划。包括康复目标、康复措施、康复期限等信息，以帮助患者恢复身体功能和提高生活质量。

生活方式建议：除了药物治疗和手术建议，系统还会提供相关的生活方式建议。如饮食调整、运动锻炼、心理疏导等，以促进患者的康复和健康。

预防措施：针对某些疾病，系统会提供相应的预防措施。包括疫苗接种、注意事项、避免诱因等，以帮助患者预防疾病的发生和复发。

3.2.4. 算法评估

本文实现了三种不同的算法，包括 TF-IDF 词频统计分析算法、BERT 算法和 Word2Vec 算法。根据每个算法已经在实验数据上进行了训练和性能评估。此外，成功搭建了一个基于 Django 框架的测试页面，

可以接受用户输入病症描述，并根据不同的算法进行医疗方案的推荐。本文对三个算法输入普通感冒、咽炎、急性病毒性喉炎和疱疹性咽峡炎，来比较输出的医疗方案结果。

Figure 6. Treatment suggestion display diagram
图 6. 治疗建议显示图

下图是使用词频统计对大量医疗病例数据集进行词频统计，相似词频出现最高的五个统计结果如图 7 所示。

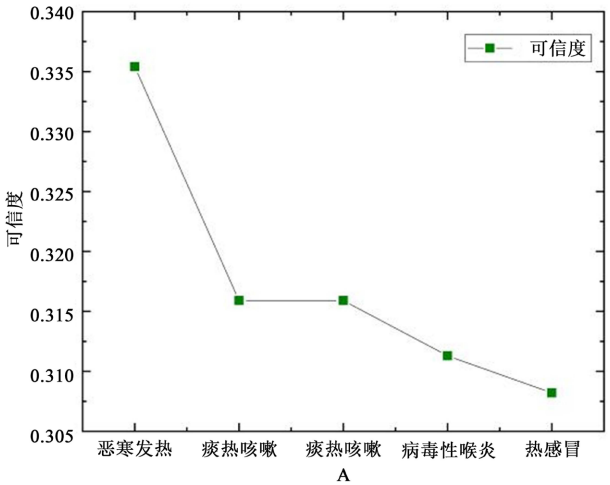

Figure 7. Word frequency statistics result dot-line diagram
图 7. 词频统计结果点线图

由表 2 结果可知，得到了相似词频出现最高的五个词：① 恶寒发热(可信度：0.3354)，② 痰热咳嗽(可信度：0.3159)，③ 痰热咳嗽(可信度：0.3159)，④ 病毒性喉炎(可信度：0.3113)和⑤ 热感冒(可信度：0.3082)，还有每个病症对应的症状表现。这些在病症词的可信度分数表示了它们与输入病例的相似程度，数值越大表示相似度越高。

Table 2. Word frequency statistics result
表 2. 词频统计结果

病症词	恶寒发热	痰热咳嗽	痰热咳嗽	病毒性喉炎	热感冒
可信度	0.3354	0.3159	0.3159	0.3113	0.3082

使用 BERT 算法将病例描述文本转化为向量表示，并计算不同病例之间的相似度。根据计算结果，得到相似度最高的五个词：① 眼睑热性疱疹，② 结膜炎，③ 鼻咽炎，④ 急性甲状腺炎和⑤ 心肌炎，以及每个病症对应的症状表现。

使用 Word2Vec 算法将病例描述文本转化为向量表示，并计算不同病例之间的相似度。根据计算结果，得到相似度最高的五个词：① 儿童顿咳，② 时行感冒，③ 热感冒，④ 病毒性支气管炎和⑤ 急性甲状腺炎，以及每个病症对应的症状表现。

根据 Word2Vec 和 BERT 算法得到的词的排名越高，通常意味着病例之间的相似度越高。

本文对 TF-IDF 词频统计分析算法、Word2Vec 算法和 Bert 算法三个算法模型输入普通感冒、咽炎、急性病毒性喉炎和疱疹性咽峡炎这些症状，来比较输出的医疗方案结果，三个算法实验结果对比如下表 3 所示。

Table 3. Comparison of experimental results of three algorithms
表 3. 三个算法实验结果对比

输入	TF-IDF 算法	BERT 算法	Word2Vec 算法
普通感冒；咽炎；急性病毒性喉炎；疱疹性咽峡炎	病毒性喉炎；热感冒	鼻咽炎	时行感冒；热感冒；病毒性支气管炎

由表 3 可以看出，TF-IDF 算法输出的结果是病毒性喉炎和热感冒。BERT 算法输出的结果是鼻咽炎。Word2Vec 算法输出的结果是时行感冒、热感冒和病毒性支气管炎。对比输入的普通感冒、咽炎、急性病毒性喉炎和疱疹性咽峡炎这些症状，可以看出 Word2Vec 算法具有较高的准确性，说明 Word2Vec 算法比其它两个算法更优。

3.2.5. 实验结果对比

如图 8 所示，通过在传统的医疗资讯平台如 39 健康网上输入患者症状：“我的头很痛”。根据搜索内容网站返回链接显示，根据传统的匹配算法，将头痛拆分为了“头”字和“痛”字，再根据关键字匹配原理在数据源中找到包含“头”字和“痛”字的内容并返回相关链接。由结果图 9 所示：第一条返回内容为“俯卧撑肱三头肌肘关节痛”和输入内容“头痛”的含义大相径庭。而其他的搜索内容也与患者头痛的语义差别较大。由此可见传统的医疗资讯平台的搜索算法并不具有文本语义匹配的功能，导致返回给用户的结果并不理想。

如图 8 所示，在本文研究的智慧医疗资讯平台输入相同症状：“我的头很痛”。根据搜索内容返回的结果中显示大部分病例都是包含“头痛”或者由头痛导致的相关症状。说明结合了三种文本语义匹配算法的医疗资讯平台在资讯结果上会给用户带来更准确且更具人性化的方案。

4. 实验结论

本文通过对智慧医疗资讯个性化服务平台的研究和对传统医疗资讯平台的实验对比证明了该平台在

Figure 8. 39 Health Net matching result diagram
图 8. 39 健康网匹配结果图

Figure 9. Matching result diagram of intelligent medical information platform
图 9. 智慧医疗资讯平台匹配结果图

医疗建议的准确性方面更具优势。未来的研究可以进一步改进算法、扩充数据集、引入用户反馈与评估机制，并加强隐私和安全保护，以推动智慧医疗领域的发展和创新。

基金项目

大连民族大学创新创业训练计划(202312026027)资助。

参考文献

[1] Abubakar, H.D., Umar, M., and Bakale, M.A. (2022) Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec. *SLU Journal of Science and Technology*, **4**, 27-33.

[2] 李龙, 金铄, 黄霞. 基于改进 TF-IDF 算法的毕业生就业推荐算法研究[J]. 计算机与数字工程, 2023, 51(9): 1985-1989+2118.

[3] Cahyani, S.N. and Saraswati, G.W. (2023) Implementation of Support Vector Machine Method in Classifying School Library Books with Combination of TF-IDF and Word2Vec. *Jurnal Teknik Informatika* (*Jutif*), **4**, 1555-1566. https://doi.org/10.52436/1.jutif.2023.4.6.1536

[4] Xu, S.T., Leng, Y.H., Feng, G.F., *et al*. (2023) A Gene Pathway Enrichment Method Based on Improved TF-IDF Algorithm. *Biochemistry and Biophysics Reports*, **34**, Article ID: 101421. https://doi.org/10.1016/j.bbrep.2023.101421

[5] 彭俊利, 王少泫, 陆正球, 等. 基于 LDA-TF-IDF 和 Word2vec 文档表示[J]. 浙江纺织服装职业技术学院学报, 2023, 22(2): 91-96.

[6] Popova, E. and Spitsyn, V. (2021) Sentiment Analysis of Short Russian Texts Using BERT and Word2Vec Embeddings. *GraphiCon* 2021: 31*st International Conference on Computer Graphics and Vision*, Nizhny Novgorod, 27-30 September 2021, 1011-1016. https://doi.org/10.20948/graphicon-2021-3027-1011-1016