[摘要 I](#_Toc472003377)

[Abstract II](#_Toc472003378)

[第1章 绪论. 1](#_Toc472003379)

[1.1 选题背景 1](#_Toc472003380)

[1.1.1 文本检索的广泛运用 11](#_Toc472003399)

[1.1.2 文本检索挑战 11](#_Toc472003399)

[1.1.3 选题意义 11](#_Toc472003399)

[1.2 国内外研究现状 1](#_Toc472003381)

[1.2.1 传统文本匹配模型 11](#_Toc472003399)

[1.2.2 深度学习文本匹配模型 11](#_Toc472003399)

[1.2.3 现当今模型存在的问题 11](#_Toc472003399)

[1.3 创新点和技术路线 2](#_Toc472003382)

[1.3.1 模型创新点 11](#_Toc472003399)

[1.3.2 技术路线 11](#_Toc472003399)

[1.4 本章小结 2](#_Toc472003383)

[第2章 深度学习文本索模型基本原理 4](#_Toc472003384)

[2.1 深度学习文本索模型概述 4](#_Toc472003385)

[2.1.1 基于表示的深度模型 11](#_Toc472003399)

[2.1.2 基于交互的深度模型 11](#_Toc472003399)

[2.2 文本检索模型的模块分析 4](#_Toc472003386)

[2.2.1 基于表示的深度模型 11](#_Toc472003399)

2.2.1.1...arci、arcii、dssm、cdssm模型，

[2.2.2 基于交互的深度模型 11](#_Toc472003399)

2.2.1.1...drmm、matchpyramid模型，

[2.3 文本检索模型的性能分析 4](#_Toc472003387)

[2.4 本章小结 4](#_Toc472003388)

[第3章 基于文本检索的深度关联匹配模型算法(DTMM)的研究 6](#_Toc472003389)

[3.1 基于文本检索的深度关联匹配模型算法的改进(DTMM)方案 6](#_Toc472003390)

[3.1.1 交互信息强度对模型的影响 11](#_Toc472003399)

[3.1.2 文档词权重对模型的影响 11](#_Toc472003399)

[3.1.3 根据文档词权重及交互信息强度改进模型 11](#_Toc472003399)

[3.2 基于文本检索的深度关联匹配模型算法的改进模型(DTMM)介绍 6](#_Toc472003391)

[3.1.1 查询和文档词的表示 11](#_Toc472003399)

[3.2.2 交互信息层 11](#_Toc472003399)

[3.2.3 池化层 11](#_Toc472003399)

[3.2.4 全连接层 11](#_Toc472003399)

[3.2.5 损失函数 11](#_Toc472003399)

[3.3 DTMM模型实验与性能分析 6](#_Toc472003391)

[3.2.1 DTMM模型实验数据集 11](#_Toc472003399)

[3.2.2 DTMM模型参数设置 11](#_Toc472003399)

[3.2.3 DTMM模型性能对比 11](#_Toc472003399)

[3.2.4 DTMM模型交互信息模块分析 11](#_Toc472003399)

[3.2.4 DTMM模型池化模块分析 11](#_Toc472003399)

[3.4 本章小结 8](#_Toc472003392)

[第4章 基于知识图谱检索模型的性能研究 9](#_Toc472003393)

[4.1 知识图谱结构与研究 9](#_Toc472003394)

[4.2 知识图谱检索模型 9](#_Toc472003394)

[4.3 模型实验与性能分析 9](#_Toc472003394)

[4.5 本章小结 10](#_Toc472003396)

[第5章 论文总结 11](#_Toc472003397)

[5.1 模型贡献 11](#_Toc472003398)

[5.1.1 DTMM模型 11](#_Toc472003399)

[5.1.2 知识图谱检索模型 11](#_Toc472003399)

[5.2 模型总结 1](#_Toc472003395)2

[5.2.1 DTMM模型 11](#_Toc472003399)

[5.2.2 知识图谱检索模型 11](#_Toc472003399)

[5.3 未来改进 1](#_Toc472003395)2

[5.4 本章小结 13](#_Toc472003403)

[致谢 14](#_Toc472003404)

[参考文献 15](#_Toc472003405)

# [绪论](#_Toc472003379)

## [选题背景](#_Toc472003380)

### [文本检索的广泛运用](#_Toc472003399)

文本检索，即为自然语言检索，不对文献进行任何的标引，通过以自然语言匹配用户所需文本的系统。根据用户查询，文本检索进行匹配的对象包括文章、报告甚整本图书或者它的一部分。随着互联网的发展，文本资源在互联网上数量日益增多，文本的数量级、文本的结构都发生了很大的变化。例如，文本数量大幅增长、文本半结构化，这给文本检索技术提出了更大的挑战和机遇。虽然遍布困难，文本检索任然为当今人们获取知识的一个重要手段。现当今文本检索运用的场景非常广，例如当今国内外最流行的搜索引擎、图书馆检索系统等等都运用的文本检索技术。

### [文本检索挑战](#_Toc472003399)

随着计算机的出现，人们借助计算机可以更加方便的储存更多的文档，现当今很多图书、知识都是以电子文档的储存方式存在的。随着互联网的出现和发展，文本文献在互联网上的数量增加更加迅速，这为数据的存储带来了新一时期的挑战。同时，随着文本文献的数量增多，其质量也表现参差不齐，检索质量与精度也成为新一代的挑战。由此，人们对搜索引擎的性能、准确性要求进一步提高。由于文本数量巨大，搜索引擎往往不能狗一次性快速准确的检索出用户所需求的信息，因此，搜索引擎检索的过程分为初次检索、再次检索的过程。初次检索筛选掉大量不准确的文档，得到小部分比较精确的文档；再次检索（重排）通过将初步检索的文献再次筛选，筛选出用户所需求的小部分文献展示给用户。在初次检索过程中，搜索引擎对筛选模型的性能要求非常高，在再次检索过程中，搜索引擎对模型的精度要求非常高，因为这一过程会带给用户最直接的搜索质量感受。对于再次检索过程中，模型的精度要求非常高，本文主要探讨再次检索模型提升精度的挑战。对于提升精度的挑战，具体又分为以下几个挑战：第一、词语匹配的多元性，不同的词语可能表示的是同一个语义，比如同义词，例如，“荷花” 、“莲花”都表示同一种植物，不同的称呼。同样地，一个相同的词在不同的语境下会有不同的语义，比如“苹果”既可以是一种水果，也可以是一家公司，一个品牌。第二、短语匹配的结构性，多个词语可以按照一定的结构组合成短语，匹配两个短语需要考虑短语的结构信息，例如“深度学习”和“深度学习”是词语顺序匹配的，而“深度学习”和“学习深度”是顺序打乱的，它们的意义是不一样的，它们的匹配程度是不一样的。第三、文本匹配的层次性，文本是以层次化的方式组织起来的，词语组成短语，短语组成句子，句子组成段落，段落组成篇章。这一特性使得我们在做文本匹配的时候需要考虑不同层次的匹配信息，也就是需要考虑不同粒度的信息匹配，模型应该按照层次的方式匹配文本信息。

### [选题意义](#_Toc472003399)

本文撰写的目的是提高搜索引擎重排过程中模型的精度，以此提高搜索引擎质量，给用户提供更加优良的搜索体验，以及优化用户获取知识的手段。文本对研究领域的贡献有以下亮点：第一，本文提出了基于深度文本匹配模型DTMM，该模型发现embedding计算词与词的相似度不尽合理问题；提出将document词以全连接映射，并将其加入交互矩阵以缓解相似度不合理的问题；提出在query维度选出k个最强信号量输入模型，以此去除噪音，从而优化模型。在MQ2007以及Robust04标准数据集上，MAP指标分别提升8%、12.5%；第二，本文提出了基于知识图谱检索的信息检索模型，该模型将额外的知识引入到深度学习信息检索模型中，使模型有着更加直观的可解释性，在性能上面也有大幅度的提升。

## **[国内外研究现状](#_Toc472003381)**

### [传统文本匹配模型](#_Toc472003399)

布尔模型(Boolean)，在模型中，用户查询与文档由其包含的单词集合来表示，两者的相似性通过布尔代数运算（与或非）来进行判定。它的优点是简单直观，但是搜索结果过于粗糙。

向量空间模型(Vector Space Model)，向量空间模型把每个文档表示成一个固定维度的向量，这固定维度是单词、词组、N-gram片段等特征信息。每个特征会计算相应的权重，一般采用TF-IDF框架计算权值，计算公式有多种变体，最简单的就是直接利用词频数作为TF值。其中，IDF衡量不同特征词在数据集中的重要程度，特征词出现过的文档数目越多，IDF值越低，这个词区分不同文档的能力越差。TF为某一文档中该词出现的频率。向量空间模型以查询和文档之间的内容相关性来作为衡量标准，按照文档和查询的相似性得分从高到低排序作为搜索结果。 它的相似度一般采用Cosine计算。

概率检索模型（probability model），它基于概率排序原理，是目前效果最好的模型之一，它已经在商业搜索引擎的网页排序中广泛应用。用户给出一个查询，它计算出每个查询和文档的相关度得分，根据相关度得分将文档从高到低排序，从而得出用户所需文档。

模型首先根据用户的查询将数据集将文档划分为两个集合：相关文档集和不相关文档集，

对某个文档，若其计算出的相关文档集的概率大于不相关文档的概率，认为它与查询相关。

该模型可以考虑IDF因子、文档长度因子、文档词频和查询词频等相关因素，并利用3个超参数对各种因子的权值进行调整。

机器学习排序（Learning to Rank）利用机器学习对搜索结果进行排序近年来取得很好的发展，机器学习是监督学习方法监督学习方法，分为训练和评测两个部分。在训练阶段，最重要的就是训练数据的生成，训练数据可以由两种方式生成。第一种是人工标注数据，这种标注方式比较耗时、耗人力物力。第二种方式是通过用户搜索日志来获取标签，这种方式比前者更加方便，但却存在着很多问题需要解决，例如“position bias”、用户看到的文档数量有限等局限性。机器学习排序方法包括三种，pointwise，pairwise，和listwise，前面两种是将问题转化为分类、回归、顺序分类等问题，但其学习目标和真实目标衡量不一致。后面一种将所有的文档作为一个序列方式进行训练，考虑每个查询下的文档是有序的，但这种方式十分消耗资源。

### [深度学习文本匹配模型](#_Toc472003399)

基于深度学习的文本匹配学习模型。近年来，随着深度学习在计算机视觉[20, 21]，语音识别[22]等领域取得的突破性进展，自然语言处理成为深度学习研究的下一个应用热点。深度学习用于自然语言处理的优势主要体现在：(1) 深度学习模型可以将单词表示为语义空间中的向量，利用向量之间的运算可以更准确地描述两个单词之间的语义关系；(2) 深度学习模型自身的结构是层次化和序列化的，能够比较自然地描述自然语言中的层次结构、序列结构和组合操作；(3) 深度学习模型很好地利用大规模数据的优势和日益发展的高性能计算的能力，将神经网络的灵活结构，匹配上复杂的自然语言的知识表示。直接从大量数据学习既可以模拟人们定义规则（特征）来描述规范的一般的语言规律，又可以刻画例外的、特殊的语言现象，从而大幅提高语言处理的精度。

在传统的自然语言处理领域深度学习已经有了很多突破性的进展，如词性标注[23]、语法分析[24]、情感分析[25]、关系分类[26]等。相关的工具包括卷积神经网络[27]和循环神经网络[28]等。卷积神经网络的卷积核的结构能够建模局部化信息，并有平移不变性的特性[29]，堆叠起来的卷积层可以很方便地模拟语言层次化的特性。而循环神经网络更偏向于序列化建模，类似人类阅读文本的方式每次将历史的信息压缩到一个向量，并作用于后面的计算，符合建模文本的序列性。

**基于Ad-hoc任务的深度关联匹配模型算法研究现状**

机器学习方法已经成功应用于近年来的信息检索（IR）。通常，基于一组人为定义的特征来学习给定查询和文档对的产生相关性分数的排序函数。但是，手工制作特征是耗时的，不完整和过度指定的。另一方面，作为表示学习方法的深度神经网络能够从训练数据中发现对于任务有用的不同抽象级别的隐藏结构和特征。最近，深度模型已被应用于计算机视觉[30]，语音识别[31]和NLP [32,33]等各种应用，并取得了性能上的显著提升。鉴于这些领域的深度学习取得成功，深度学习似乎对IR有重大影响。然而，直到现在，IR任务的深度模型，特别是特殊检索任务能达到指定目标的模型还很少。在一般的情况下，当将深度模型应用于检索时，任务典型地被形式化为两段文本（即，查询和文档）之间的匹配问题。这样的匹配问题通常是广泛的，它可以涵盖检索任务以及其它NLP任务，如释义识别，问答（QA）和自动对话[33,34]。已经提出了各种深度匹配模型来解决这个匹配问题，根据它们的模型结构可以将其分为两类。一种是以表示为中心的模型，它将两个文本用深度神经网络单独的表示，然后再将两个文本的深度表达进行匹配。例如DSSM [35]，C-DSSM [36,37]和ARC-I [34]。另一个是基于交互的模型，首先在两段文本之间建立交互（即局部匹配信号），然后使用深度神经网络学习文本匹配交互信息。例如DeepMatch [33]，ARC-II [34]和MatchPyramid [38]。然而，在这项工作中，我们认为许多NLP任务中的匹配问题和Ad-Hoc检索任务是根本不同的。大多数NLP任务涉及语义匹配，即识别语义并且推断两段文本之间的语义关系，而Ad-Hoc检索任务主要是关于关联匹配，即识别文档是否与给定查询有关。这两个匹配问题之间的三个主要区别，这可能导致深度匹配模型的显着不同的架构设计。大多数现有的深度匹配模型是为语义匹配而非相关匹配而设计的。基于这些差异，提出了一种深度相关匹配模型（DRMM）进行Ad-Hoc检索，通过对相关匹配中的三个主要因素进行明确建模。总的来说，模型是一个以交互为中心的模型，它在查询单词级别采用联合深度架构进行相关性匹配。具体来说，首先建立基于查询和文档的每对词之间的本地交互。对于每个查询单词，我们将可变长度的交互映射到一个固定长度的匹配直方图。基于这个固定长度的匹配直方图，我们使用一个前馈网络来学习分层匹配模式并产生一个匹配分数。最后，通过将来自每个查询单词的分数与词汇门控网络计算聚合权重而生成总体匹配分数。我们展示了我们的主要模型如何设计，包括匹配直方图映射，前馈匹配网络和术语门控网络，解决Ad-Hoc检索的相关匹配中的三个关键因素。实证结果表明，现有的深度匹配模型无法与传统检索模型进行竞争，而所有评估指标方面，该模型在显著优于所有基准模型。

**语意匹配与相关匹配**

许多自然语言处理任务中的匹配，如释义识别，问答和自动对话，主要涉及语义匹配，即识别语义，推理出两段文本之间的语义关系。在这些语义匹配任务中，这两个文本通常是同质的，由几个自然语言句子组成，例如问题/回答句子或对话。为了推断自然语言句子之间的语义关系，语义匹配强调以下几个因素：相似性匹配信号，与精确匹配信号，短语和句子之间的语义相似性/相关性是重要或关键的。例如，在释义识别中，需要识别两个句子是否用不同的表达来表达相同的含义。在自动对话中，我们的目标是找到与之前的对话语义相关的适当的回答，它们之间可能不共享任何常见的单词或短语。由于语义匹配中的文本通常由具有语法结构的自然语言句子组成，因此使用基于这种语法结构的句子的组成含义而不是将它们看作一组词/序列[32]更有利。例如，在回答问题时，大多数问题都有明确的语法结构，可以帮助模型确定问题的组成意义。语义匹配通常把这两个文本作为一个整体来推断它们之间的语义关系，从而进行全局匹配。这部分与语义匹配中的大部分文本长度有关，因此话题范围比较集中。例如，如果整个意思是相同的，那么两个句子就被认为是解释，而一个好的答案完全可以回答这个问题。相反，Ad-Hoc检索中的匹配主要是关于相关性匹配，即识别文档是否与给定查询相关。在这个任务中，查询通常是简短的和基于关键字的，而文档可以在长度上变化很大，从数十个字到数千甚至数万个字。为了估计查询与文档之间的相关性，相关性匹配关注以下三个因素：精确匹配信号：尽管单词不匹配是临时检索中的关键问题，并且已经使用不同的语义相似性信号来处理，但是精确匹配由于现代搜索引擎中的索引和搜索范式，文档中的单词与查询中的单词仍然是临时检索中最重要的信号。例如，Fang和Zhai [42]提出了语义词匹配约束条件，即匹配一个原始查询词应该总是对相关性分数贡献不少于多次匹配一个语义相关词。这也解释了为什么一些传统的检索模型（例如BM25）能够基于精确的匹配信号完全合理地工作。查询词的重要性：由于查询主要是简短的，基于关键词而没有复杂的语法结构，所以重要的是要考虑到词的重要性，而查询词之间的组成关系通常是操作中的简单“和”关系搜索。例如，在查询“比特币新闻”的情况下，预期相关文件将涉及“比特币”和“新闻”，其中术语“比特币”比“新闻”更重要， “比特币”与描述其他事物的“新闻”相关的文档更为相关，文献中对检索模型进行了许多正式的研究，显示了词汇歧义的重要性[43,44]。不同的匹配要求：在文献检索中，相关文献的篇幅可能会很长，文献篇幅有不同的长度[41]，导致了文献的多样性匹配的要求，具体而言，假设假设长文献就像一个短文献，类似的范围，但有更多的话语，在这种情况下，如果我们假设短文档的话题集中，相关性匹配可能是全局的。通过这种方式，相关性匹配可以发生在相关文档的任何部分，并且我们不要求整个文档与查询相关。正如我们所看到的，在许多NLP任务中，临时检索中的相关匹配和语义匹配之间存在显着差异。这些差异会影响深层模型架构的设计，可能会很困难。

**DRMM模型**

基于以上分析，提出了一个新的深度匹配模型，模型为深度相关匹配模型（DRMM）。总体而言，模型类似于以交互为中心的模型，而不是表示表达为中心的模型，因为后者将不可避免地丢失对Ad-Hoc检索中的关联匹配至关重要的详细匹配信号。具体而言，我们的模型在查询单词级别采用联合深层架构，在查询和文档单词之间进行相关匹配的交互。我们首先在查询和基于词嵌入的文档中的每对词之间建立局部交互。对于每个查询单词，我们然后将可变长度的局部交互变换成固定长度的匹配直方图。基于定长匹配直方图，我们使用前馈匹配网络来学习分层匹配模式，并为每个查询词产生匹配分数。最后，通过将来自每个单个查询单词的分数与单词门控网络计算聚合权重而生成总体匹配分数。以下为该模型的计算公式：其中每个查询表示为：q={w1(q), . . . , wM(q)}，wi(q)为查询中的每个词；每个文档表示为：d = {w1(d), . . . , wN(d)}；其中wi(d)为文档中的每个词,h()表示将查询与文档的匹配值映射为匹配直方图；W (l)为需要训练的权重；b(l)为偏置量；tanh()为激活函数；gi代表softmax函数。

### [现当今模型存在的问题](#_Toc472003399)

根据现有的文献[35]，特定检索中的核心问题，即给定特定查询的文档的相关性的计算，可以被形式化为如下的文本匹配问题。给定两个文本T1和T2，匹配度通常被定义为基于每个文本的表示，然后由评分函数产生分数：其中Φ是将每个文本映射到表示向量的函数，F是基于它们之间的交互的评分函数。这样的文本匹配问题被认为是通用的，因为它也描述了许多NLP任务，如释义识别，问题回答和自动对话[33,34]。已经提出了各种深度匹配模型用于特定的特别检索任务或一般匹配问题。根据你如何选择这两个函数，现有的深度匹配模型可以分为两种类型。第一个是以表示为中心的模型，试图将具有深度神经网络的单个文本建立为一个良好的表示，然后在两个组合和抽象文本表示之间进行匹配。在这种方法中，Φ是一个复数表示映射函数，而F是一个相对简单的匹配函数。例如，在DSSM [35]中，Φ是一个前馈神经网络，而F是余弦相似函数。在C-DSSM [36,37]中，Φ是一个卷积神经网络（CNN）[30]，而F是余弦相似函数。在ARC-I [34]中，Φ是一个CNN，而F是一个多层感知器（MLP）。一般地，所有关于表示集中模型的模型体系结构都可以被看作是对文本输入的对称体系结构。第二个是以交互为中心的模型，首先基于一些基本表征建立两个文本之间的局部交互，然后利用深度神经网络学习匹配的分层交互模式。在这种方法中，Φ通常是一个简单的映射函数，而F是一个复杂的深度模型。例如，在DeepMatch [33]中，Φ简单地将每个文本映射到一个单词序列，而F是一个前馈神经网络，在单词交互矩阵上由主题模型驱动。在ARC-II [34]和MatchPyramid [38]中，Φ将每个文本映射到一个单词向量序列，而F是两个文本单词向量之间交互矩阵的CNN。在一般性的情况下，所有关于交互模型的模型体系结构都可以被看作是一个深层的体系结构。尽管在这样的匹配模式下已经提出了各种深度匹配模型，但是其中大部分仅在一些NLP任务（如释义识别和QA）方面被证明是有效的[34,39]。在Ad-Hoc检索任务方面几乎没有积极的结果。即使是专门为Web搜索而设计的深层模型，例如DSSM和C-DSSM，也只是在<query，doc title>对上进行评估，而这些对不是典型的Ad-Hoc检索设置。如果我们直接将这些深度匹配模型应用于某些基准检索集合，例如TREC集合，与传统的排序模型（如语言模型[40]和BM25 [41]）相比，发现性能相对较差。

## **[创新点和技术路线](#_Toc472003382)**

### [模型创新点](#_Toc472003399)

基于深度学习的文本检索排序模型，该模型发现embedding计算词与词的相似度不尽合理问题；提出将document词以全连接映射，并将其加入交互矩阵以缓解相似度不合理的问题；提出在query维度选出k个最强信号量输入模型，以此去除噪音，从而优化模型。在MQ2007以及Robust04标准数据集上，MAP指标分别提升8%、12.5%；

基于知识图谱检索的信息检索模型，该模型将额外的知识引入到深度学习信息检索模型中，使模型有着更加直观的可解释性，在性能上面也有大幅度的提升。

### [技术路线](#_Toc472003399)

实验数据收集。MQ2007数据集、Robust04数据集、WikiPidea数据。

数据预处理。包括文字规范化、分词、去标点、去停用词。WikiPidea数据集建立知识图谱。

安装图数据SqlGraph，建立知识图谱。

安装环境。包括Python、tensorflow、GPU、Matchzoo、Anaconda环境。

在Matchzoo中实现模型

数据评测、观测模型性能。观测MAP、NDCG、Percision指标。

分析实验结果，修改模型

根据实验结果撰写论文

本节阐述了技术路线，从安装环境到数据预处理、模型的搭建、模型的评测与修改等方面进行了拟定。

## **[本章小结](#_Toc472003383)**

本章主要阐述了文本检索在现实世界中的应用，以及当下面临的挑战。在搜索引擎广泛被运用以及人们日益增长的获取高精度知识需求的前提下，对文本检索模型进行改进提升是十分有必要的。本章同时也说明了国内外文本检索面临的现状，传统模型与深度学习模型的对比，在此大前提下，本文拟定两个方面的研究点，即基于深度学习的文本检索排序模型和基于知识图谱的检索模型。