



学术学位硕士研究生学位论文开题报告

及课题研究计划

所在学院 计算机科学与工程

学科专业 计算机系统结构

研究方向　　 信息检索

年　　级　　 2016级

学　　号 52160329101

研 究 生 杨州

指导教师 王越(教授)、朱小飞(教授)

入学年月 2016年9月

重庆理工大学研究生处制

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 论文题目 | 基于文本检索的深度关联匹配模型算法的研究与改进 | | | | | | |
| 题目来源 | 国　家 | 部 委 | | 省、市、自治区 | | 横　向  联　系 | 自　选 |
|  |  | |  | |  | √ |
| 题目类型 | 基础研究 | | 应用基础 | | 工程技术 | | 其　他 |
| √ | |  | |  | |  |
| 经　费 |  | | | | | | |
| 一、论文选题的意义  文本匹配是自然语言理解中的一个核心问题。对文本匹配的研究可以应用到大量已知的自然语言处理任务中，例如信息检索[1]、自动问答[2]、机器翻译[3]、对话系统[4]、复述问题[5]等等。这些自然语言处理的任务都可以在一定程度上抽象成文本匹配问题，比如信息检索可以归结为查询项和文档的匹配，自动回答可以归结为问题和候选答案的匹配，机器翻译可以归结为两种语言间的匹配，对话系统可以归结为前一句对话和回复的匹配，复述问题则可以归结为两个同义词句的匹配。这些匹配需要关注的特性具有很大不同，如何利用一个较好的文本匹配模型，针对不同任务找到最适合的匹配方式，成为研究文本匹配这个核心问题最大的挑战。  文本匹配面临的挑战主要来源于以下几个方面：  （1）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，“荷花”、“莲花”、“水芙蓉”、“芙蕖”，它们表示的都是同一种植物；同理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，亦可以是一个品牌。  （2）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。比如“机器学习”和“机器学习”是两个词顺序匹配的，而“机器学习”和“学习机器”只有词语是匹配的，而顺序是打乱的。这两种情况的匹配程度是不一样的。  （3）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。这样一种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，按照层次的方式组织我们的文本匹配信息。  最近文本匹配问题的研究，渐渐从传统文本匹配模型向深度文本匹配模型转移。由于传统的文本匹配模型需要基于大量的人工定义和抽取的特征而且可以学习调整的参数相对较少，所以这些特征总是根据特定的任务（信息检索，或者自动问答）人工设计的，很大程度上限制了模型的泛化能力。在信息检索方面很多工作是基于传统的检索模型的改进，例如将文档模型和LDA模型融合[6]，将词向量的结果作为特征[7]；自动问答的大量工作是基于知识库检索也既是对结构化的数据的检索问题；而对话系统的研究还处于起步阶段，部分利用传统模型的工作是基于句型模式识别和语义提取[8, 9]。传统模型在一个任务上表现很好的特征很难用到其他文本匹配任务上。而利用深度学习方法[10]，可以自动从原始数据中抽取特征，免去了大量人工设计特征的开销。首先特征的抽取过程是模型的一部分，根据训练数据的不同，可以方便适配到各种文本匹配的任务当中。以此同时，深度文本匹配模型结合上词向量（Word2Vec[11, 12]）的技术，更好地解决了词语匹配的多元性问题。最后得益于神经网络的层次化设计原理，深度文本匹配模型也能较好地符合短语匹配的结构性和文本匹配的层次性的特性。  考虑以上问题，在阅读了大量的文献和资料后，确定将改进基于Ad-Hoc任务的深度关联匹配模型算法作为研究对象。一定程度上解决了需要人工选取特征的问题，在构建模型与人力资源等方面节约了很大的成本。 | | | | | | | |
| 1. 国内外研究现状   首先是基于传统文本匹配学习模型。传统的文本匹配研究主要基于人工提取的特征，因此问题的焦点在于如何设置合适的文本匹配学习算法来学习到最优的匹配模型。以互联网搜索为例，查询项与网页被认为是两个异质空间中的对象，多种匹配学习模型被提出来去计算查询与网页的相关度。Berger和Lafferty [13]提出使用统计机器翻译模型计算网页词和查询词间的“翻译”概率，从而实现了同义或者近义词之间的匹配映射；Gao 等人[14]在词组一级训练统计机器翻译模型并利用用户点击数据进行模型训练，获得了很好的效果。进一步地说，典型相关分析 (CCA, canonical correlation analysis) [15]和偏最小二乘 (PLS, partial least square) [16]等隐空间模型试图为两种对象建立一个公共的隐空间，任意给定的查询和文档都可以被映射到此隐空间中，且在隐空间中查询和文档有一致的表达方式和特征维度，从而可以方便地计算两者的相似度或者距离，进而对其是否具有相同的“语义”做出判断。例如，Wu等人[17]提出正则化隐空间映射(regularized mapping to latent space, RMLS)把查询项和网页映射到同一隐空间中，并在模型训练中引入了正则化因子以避免奇异解，Bai等人[18]提出有监督学习语义索引模型(supervised semantic indexing, SSI)，Gao等人[19]扩展了话题模型提出双语话题模型(bilingual topic model, BLTM)，对隐空间模型进行概率化建模。  尽管这些模型已经在诸如网络搜索，推荐和问答等应用中取得了良好的效果，然而还是存在许多问题。（1）人工提取特征的代价很大。需要花费大量人力物力才能提取到少量的比较有效的特征，这其中不仅需要有经验的工程师来设计，还需要大规模的特征选择过程。（2）基于主题模型的隐空间模型还比较粗糙，难以克服文本匹配中的语义鸿沟问题。（3）传统模型很难发掘一些隐含在大量数据中，含义不明显的特征，然而往往有些特殊情况需要这样的特征才能提高性能。  其次是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。  **2.1 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。  **2.2 关于Ad-hoc任务的问题**  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。  **2.3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。  **2.4 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。  IMG_256  IMG_256  **参考文献：**  [1] Li H, Xu J. Semantic matching in search [J]. Foundations and Trends in Information Retrieval, 2014, 7(5): 343-469.  [2] Xue X, Jeon J, Croft W B. Retrieval models for question and answer archives; proceedings of the Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, F, 2008 [C].  [3] Brown P F, Pietra V J D, Pietra S a D, et al. The mathematics of statistical machine translation: Parameter estimation [J]. Computational linguistics, 1993, 19(2): 263-311.  [4] Serban I V, Sordoni A, Bengio Y, et al. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models [J]. arXiv preprint arXiv:150704808, 2015.  [5] Dolan W B, Brockett C. Automatically constructing a corpus of sentential paraphrases; proceedings of the Proc of IWP, F, 2005 [C].  [6] 卜质琼, 郑波尽. 基于 LDA 模型的 Ad hoc 信息检索方法研究 [J]. 计算机应用研究, 2015, 32(5): 1369-72.  [7] Ganguly D, Roy D, Mitra M, et al. Word embedding based generalized language model for information retrieval; proceedings of the Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, F, 2015 [C]. ACM.  [8] 黄沛杰, 黄强, 吴秀鹏, et al. 语法和语义相结合的中文对话系统问题理解研究 [J]. 中文信息学报, 2014, 28(6): 70-8.  [9] 俞凯, 陈露, 陈博, et al. 任务型人机对话系统中的认知技术———概念, 进展及其未来 [J]. 计算机学报, 2015, 38(12): 2333-48.  [10] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-44.  [11] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality; proceedings of the Advances in neural information processing systems, F, 2013 [C].  [12] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [J]. CoRR, 2013, abs/1301.3781  [13] Berger A, Lafferty J. Information retrieval as statistical translation; proceedings of the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, F, 1999 [C].  [14] Gao J, He X, Nie J-Y. Clickthrough-based translation models for web search: from word models to phrase models; proceedings of the Proceedings of the 19th ACM international conference on Information and knowledge management, F, 2010 [C].  [15] Hardoon D R, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods [J]. Neural computation, 2004, 16(12): 2639-64.  [16] Rosipal R, Kr\"A M, Nicole. Overview and recent advances in partial least squares [M]. Subspace, latent structure and feature selection. Springer. 2006: 34-51.  [17] Wu W, Lu Z, Li H. Learning bilinear model for matching queries and documents [J]. The Journal of Machine Learning Research, 2013, 14(1): 2519-48.  [18] Bai B, Weston J, Grangier D, et al. Supervised semantic indexing; proceedings of the Proceedings of the 18th ACM conference on Information and knowledge management, F, 2009 [C].  [19] Gao J, Toutanova K, Yih W-T. Clickthrough-based latent semantic models for web search; proceedings of the Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, F, 2011 [C].  [20] Lecun Y, Bengio Y. Convolutional networks for images, speech, and time series [J]. The handbook of brain theory and neural networks, 1995, 3361(10): 1995.  [21] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions; proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, F, 2015 [C].  [22] Abdel-Hamid O, Mohamed A-R, Jiang H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition; proceedings of the Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, F, 2012 [C].  [23] Collobert R, Weston J, Bottou L E, On, et al. Natural language processing (almost) from scratch [J]. The Journal of Machine Learning Research, 2011, 12(2493-537.  [24] Vinyals O, Kaiser L, Ukasz, Koo T, et al. Grammar as a foreign language; proceedings of the Advances in Neural Information Processing Systems, F, 2015 [C].  [25] Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank; proceedings of the Proceedings of the conference on empirical methods in natural language processing (EMNLP), F, 2013 [C].  [26] Zeng D, Liu K, Lai S, et al. Relation Classification via Convolutional Deep Neural Network; proceedings of the COLING, F, 2014 [C].  [27] Lecun Y, Bottou L E, On, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-324.  [28] Levin E. A recurrent neural network: limitations and training [J]. Neural Networks, 1990, 3(6): 641-50.  [29] Lecun Y, Others. Generalization and network design strategies [J]. Connections in Perspective North-Holland, Amsterdam, 1989, 143-55.  [30] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10):1995, 1995.  [31] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, 29(6):82–97, 2012.  [32] R. Socher, E. H. Huang, J. Pennin, C. D. Manning, and A. Y. Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In NIPS, pages 801–809, 2011.  [33] Z. Lu and H. Li. A deep architecture for matching short texts. In NIPS, pages 1367–1375, 2013.  [34] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In NIPS, pages 2042–2050, 2014.  [35] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In CIKM, pages 2333–2338. ACM, 2013.  [36] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. Learning semantic representations using convolutional neural networks for web search. In WWW, pages 373–374, 2014.  [37] J. Gao, P. Pantel, M. Gamon, X. He, L. Deng, and Y. Shen. Modeling interestingness with deep neural networks. EMNLP, October 2014.  [38] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng. Text matching as image recognition. 2016.  [39] S. Wan, Y. Lan, J. Guo, J. Xu, L. Pang, and X. Cheng. A deep architecture for semantic matching with multiple positional sentence representations. arXiv preprint arXiv:1511.08277, 2015.  [40] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In SIGIR, pages 334–342. ACM, 2001.  [41] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In SIGIR, pages 232–241. ACM, 1994.  [42] H. Fang and C. Zhai. Semantic term matching in axiomatic approaches to information retrieval. In SIGIR, pages 115–122. ACM, 2006.  [43] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In SIGIR, pages 49–56. ACM, 2004.  [44] H. Fang, T. Tao, and C. Zhai. Diagnostic evaluatio | | | | | | | |
| 三、研究目标、内容、拟解决的关键问题、研究方法、技术路线等   * 1. **研究目标** * 研究原始的深度相关匹配模型(DRMM)模型，并理解模型的意义以及方法。 * 研究深度相关匹配模型(DRMM)模型中关于查询与文档之间位置关系对查询精度的影响。 * 研究深度相关匹配模型(DRMM)模型中多粒度对查询结果的关系和影响。   1. **研究内容** * 关于查询与文档之间位置关系对查询精度的影响   给出一个查询时，查询词语在文档之间的位置关系对查询时有影响的，例如某个查询词汇在文档的开始与文档的结尾频繁出现的重要程度是不一样的，因此对于该模型加上位置信息是十分有必要的。   * 多粒度对查询的影响   在文本检索Ad-Hoc任务中，许多模型都是单粒度的，而当人类进行查询的时候，往往是根据词或者多个词的表达来查询的，因此，多粒度的查询对于模型的改进是有很大的帮助的，在此基础之上，对模型进行研究，在多粒度的角度之下来探究粒度对于查询的影响。   * 将查询与文档之间的位置信息、多粒度对查询的影响分别进行不同的实现   在前两个研究内容的基础之上，分别对深度相关匹配模型(DRMM)模型进行基于位置信息的改进与多粒度的改进，并做大量的实验，对两种不同的研究思想加以实现。   * 1. **拟解决的关键问题** * 查询与文档之间位置关系对查询精度的影响。通过统计词汇在文本中的位置信息统计，然后将位置信息组合成向量，对向量进行卷积操作或者全连接操作，进行信息的整合，之后再对其进行原始模型同样的操作得到匹配分数。 * 多粒度对查询的影响。因为不同词的embedding相加是有意义的，所以讲两个词相加之后得到的embedding信息可以作为查询的拓展，以及可以作为文档信息的更高维度的拓展，从而增加查询与文档之间的交互信息。 * 将查询与文档之间的位置信息、多粒度对查询的影响分别进行不同的实现。对模型两个方面的改进进行实践，在一定程度上提升模型的性能。   1. **研究方法**   研究已有的深度关联匹配模型，并做大量的前期调研，对位置信息是否在文本检索中有贡献以及有多大的贡献做一个调研，对多粒度信息的加入是否对文本检索有大的贡献做一个详细的调研，之后再基于此基础之上，对模型进行改进。在python语言中，基于开源项目MatchZoo对模型进行开发与拓展。具体方案如下：   * 对数据进行收集与整理，标注出document中每个词在文档中的位置，计算出多个粒度在去除禁用词的情况下的逆文档频率(IDF值)； * 对模型进行改进，包括两个方向的改进，即基于位置信息的改进与基于多粒度的改进，其中关于位置信息的改进需要对文档的输入矩阵变换为二维矩阵，即第一维度为每个单词的embedding,第二个维度为每个单词的位置信息；此外，在基于多粒度的改进中，将不同的embedding进行相加之后再对其均值，计算出新的embedding,再将其输入到模型之中，特别地，由于多个粒度形成了某个新的词的embedding,所以需要用到统计的多个粒度的逆文档频率(IDF值)； * 最后在MatchZoo的环境之下对新模型进行评测，评测指标有MAP,NDCG,Precesion等。   1. **技术路线**   针对基于以上问题。具体技术路线如图1所示。  图1 技术路线图 | | | | | | | |
| 四、研究计划进度  2017年12月—2018年02月 深度关联匹配模型(DRMM)的研究  2018年03月—2018年04月 位置信息对文本匹配的影响调研  2018年05月—2018年07月 多粒度对文本匹配的影响调研  2018年08月—2018年11月 对模型进行改进以及调整参数  2018年12月—2019年03月 论文撰写，完成终稿  2019年03月—2019年06月 毕业答辩 | | | | | | | |
| 导师意见    导师（签名）：  年 月 日 | | | | | | | |

注：可自行加页。本表一式两份，分别交研究生处及所在学院。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 开题报告会记录  时 间：  地 点：  参加人：  主持人：  记录人：  提问及回答要点：  主持人（签名）：  记录人（签名）：  年 月 日 | | | | | | |
| 专家组对开题报告的评议  1.对开题报告的整体评价  2.存在的主要问题和建议  3.开题报告结论：  开题报告专家组共　　人，其中　　票评定该生的开题报告为优秀；　　票评定该生的开题报告为合格；　　票评定该生的开题报告为不合格。  因此，专家组认为该生的开题报告 □优秀 □合格 □不合格    专家组组长签名：  年 月 日 | | | | | |
| 参加学位论文开题报告的专家名单 | | | | |
| 姓名 | 职称 | 学科、专业 | 工作单位 | 专家签名 |
| 李 梁 | 教授 | 计算机科学与技术 | 重庆理工大学 |  |
| 闫 河 | 教授 | 计算机科学与技术 | 重庆理工大学 |  |
| 刘万平 | 副教授 | 计算机科学与技术 | 重庆理工大学 |  |
|  |  |  |  |  |
|  |  |  |  |  |
| 所在学院学位评定分委员会意见  该生的开题报告 □优秀 □合格 □不合格  学位评定分委员会主席（签名）：  年 月 日 | | | | | |

注：可自行加页。本表一式两份，分别交研究生处及所在学院。