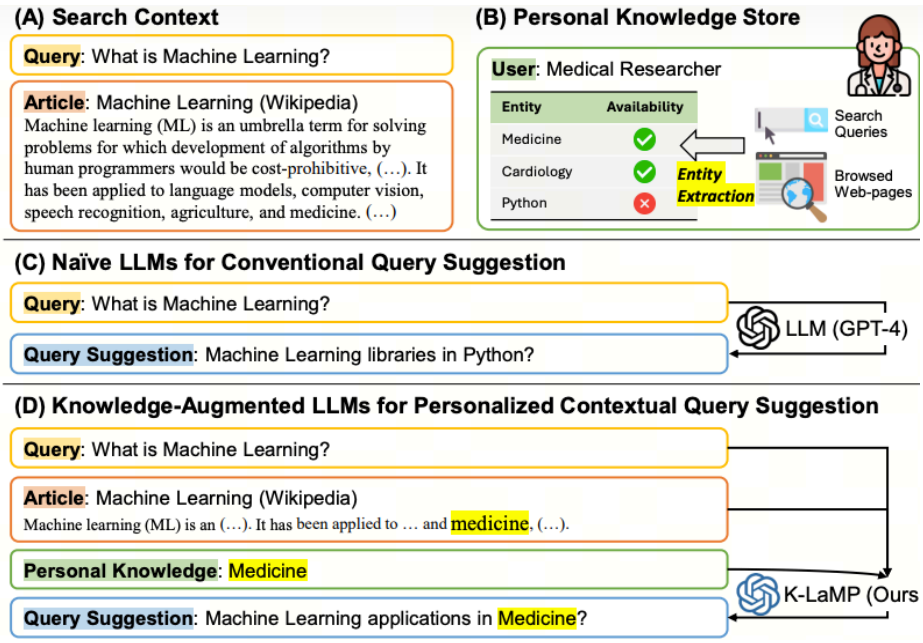
知乎



微软2024：大模型（LLM）如何赋能个性化搜索

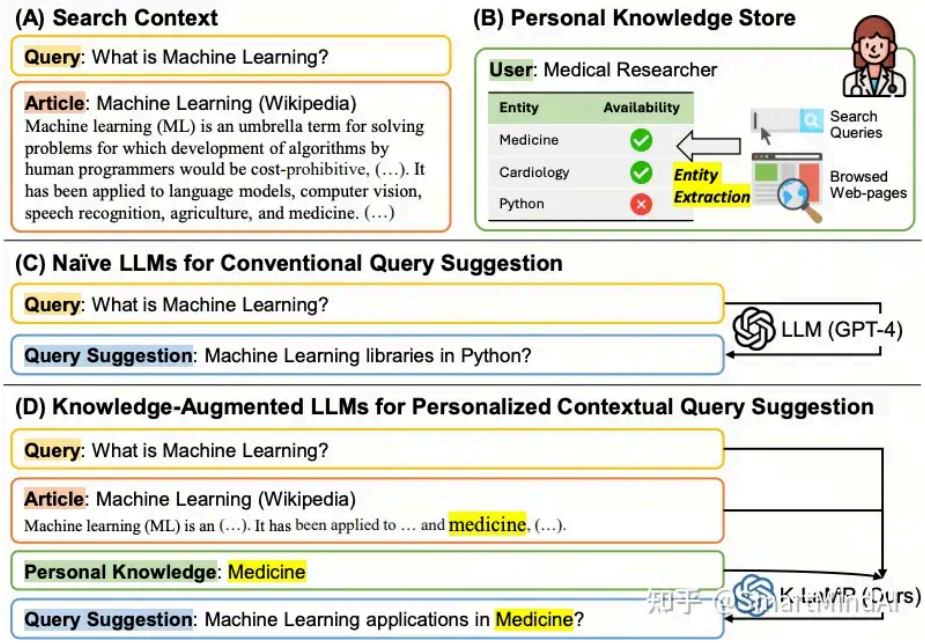
SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

原文《Knowledge-Augmented Large Language Models for Personalized Contextual Query Suggestion》

Introduction

LLMs如GPT-4，通过海量参数训练，掌握广泛领域知识，能对非特定任务的用户输入生成连贯、合理且有益的回答。它们已在问答和对话生成等多任务中显示出出色的性能。

尽管LLMs如GPT-4在广泛任务中表现出众，但个性化到满足个人偏好、需求的响应仍面临难题，主要由于单个用户大规模重训成本高。为克服这一挑战，研究人员尝试通过"在上下文中的学习"方式，如利用用户历史数据，改进模型。然而，处理大量用户数据可能引发隐私保护和系统扩展问题。一些研究还通过创建深度用户档案，但这也带来隐私风险。因此，优化大型语言模型的个性化策略，特别是在考虑隐私和扩展性的同时深度理解用户，仍是当前研究的重点。



知乎

LLMs的知识。这个个人知识库是基于用户与现代搜索引擎交互的搜索日志构建的，随着时间的推移，它通过聚合用户查询中出现的实体和浏览过的网页来不断更新。

1. 提供更深入的个性化内容，基于用户已知和兴趣。
2. 遵循隐私保护原则，利用现有日志基础设施，降低成本和复杂性。
3. 易于集成并适用于不同个性化任务，不局限于搜索。
4. 成本效益⁺高，因为只需添加少量实体相关提示，而非对模型进行全面重调。

K-LaMP: Knowledge-Augmented LLMs for Personalized Query Suggestion

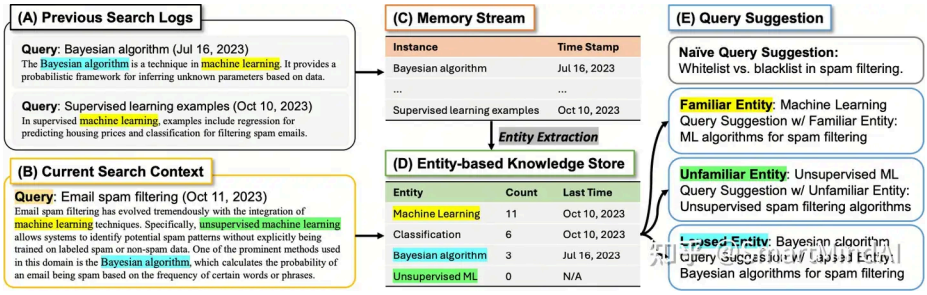
我们通过提出和实施K-LaMP方法，结合了先进的知识增强技术和实体中心化的个人知识库，以生成针对用户独特需求的个性化输出。这种方法特别针对处理情境化查询建议，即在搜索建议的基础上，考虑用户当前浏览网页⁺的内容。这与传统的基于表面联系的查询建议不同，我们的方法能深度理解和利用用户的专业领域和知识。我们在实验中利用Bing搜索引擎的搜索日志来建立这样的知识库，并通过与基于LLM的基线对比，证明了K-LaMP在生成更相关、有用且随用户交互历史增长性能提升的方面显著优于其他方法。

Problem Statement

我们从大型语言模型⁺（LLMs）的基础入手，它们是通过大量无标签文本训练的Transformer架构模型，擅长从文本中学习并应用于问答和对话生成等任务。然而，个性化上下文查询建议这一特定场景提出了挑战，即如何让LLMs针对每个用户生成与他们专业领域相关且个性化的建议，而不只是通用响应。为此，我们引入了知识增强和实体中心化的个人知识库策略，以提升模型的理解能力和生成能力。

我们首先定义了大语言模型⁺（LLMs），这是一种参数化模型$\boldsymbol{\theta}$，通过学习大量无标签文本训练，能处理文本任务如问答和对话。针对个性化上下文查询建议，我们面临的问题是如何让LLM理解用户的专业领域并生成个性化的建议，而非通用答案。为此，我们引入了实体中心化的个人知识库策略来增强模型。接下来，论文将详述如何利用LLM在给定的上下文查询建议任务中，利用搜索日志中的实体信息，以用户已知知识为导向生成更精确和动态的建议。具体涉及上下文序列$\boldsymbol{c}$的设计及其如何促进个性化输出。

Knowledge Store for Output Personalization



在本研究中，我们提出了一种针对LLM的个性化增强框架。首先，我们强调LLM的补充上下文$\boldsymbol{c}$应当包含反映用户个性化信息，如偏好、兴趣和专业知识的元素。为此，我们构建了一个用户特定的知识库$\boldsymbol{\mathcal{K}}$，其中$\boldsymbol{k}$是线性化用户数据的文本片段，它嵌入到用户当前上下文的环境中，形成

$$\boldsymbol{c} = [\boldsymbol{q_h} \cdot \boldsymbol{w} \cdot \boldsymbol{k}]$$

这里$\boldsymbol{q_h}$代表用户历史交互的特征$\boldsymbol{w}$是用户当前浏览的网页内容。 1. 设计和构建$\boldsymbol{\mathcal{K}}$的关键在于收集和处理用户数据。这可能涉及数据清理、隐私保护和用户同意的过程。我们需要确保数据的准确性和有效性，同时遵守相关法规，以合法且透明地整合用户信息。

1. 为了捕获用户兴趣和知识$\boldsymbol{k}$的选择和利用需要基于对用户行为和查询模式的深入理解。例如，我们可以通过关联用户的搜索历史、购买记录、浏览行为等多维度数据来创建一个更全面的用户画像⁺，从而选择最相关的$\boldsymbol{k}$。

知乎

函数，或者调整模型的架构以处理更复杂的关系和上下文信息。通过这样的方式，我们期望模型能够在接收实体增强的上下文后，生成更具针对性的建议。

Constructing Personal Knowledge Stores

我们有两个主要的$\mathcal{K}$集：$\mathcal{K}_s$，它通过简单线性方式记录用户过去的查询习惯和浏览行为，用来捕捉全局的查询模式；而$\mathcal{K}_e$是一个实体为中心的版本，它整合了用户独特的兴趣和专业知识，代表了个体层面的深入偏好。这两种策略分别反映了用户行为的整体趋势和个体差异，旨在通过这些定制化的知识源，提升模型生成个性化查询建议的精准度。

对于搜索历史的个人知识库$\mathcal{K}_s$，我们基于用户频繁提出和点击与特定主题相关的查询来建立。例如，用户频繁询问和点击与"机器学习"相关的关键词，表明他们对该领域有显著兴趣。这个记忆流模型通过时间戳记录用户的行为，展示了一个用户逐步获取知识的过程。

为了构建这样的库，我们从搜索引擎日志中提取数据，包括用户提问和点击的页面。重要的是，这个数据集相对小巧，反映了用户行为的直接证据，同时保持了隐私（因为主要依赖公开数据），且易于管理，避免了大规模存储和隐私保护的挑战。

个人实体中心知识库 ($\mathcal{K}_e$)是为克服基于搜索历史内存流的局限性而设计的。它不直接基于线性存储和检索，而是通过实体作为焦点，这些实体代表用户感兴趣的主题和领域。实体识别技术简化了操作，使得聚合过程更为高效。

实体链接技术用于标记和标准化用户查询及访问网页中的实体，确保信息的一致性和准确性。尽管每个实体都有时间戳，但我们通过分析用户全面互动历史中的实体频率，而非仅依赖时间，来更深入地聚合和理解用户的专业兴趣。这样做的目的是创建一个结构清晰、易于理解和处理的个性化知识库，既能处理单一主题，又能防止混淆，同时减少了处理复杂网页文本的负担。

尽管$\mathcal{K}_e$相对于更深度的个人档案可能不那么精简，但它作为一种轻量级的个性化方法，具备显著优势。它依赖于广泛可用的实体链接工具，这些工具能快速处理大量数据，适应性强。隐私保护方面，实体链接过程通常在公共知识库子图上进行，遵循k匿名化原则，保护了用户的个人信息。若需要，只需删除与实体关联，就能实现数据的匿名化。这种设计兼顾了实用性和隐私保护，使得我们的实体中心知识库在提供个性化服务的同时，保持了数据处理的便捷性和合规性。

Contextual Retrieval from Personal Knowledge Stores

在利用搜索历史知识库$\mathcal{K}_s$进行检索时，我们聚焦于查询q_j与用户当前互动文档w相关的上下文内容。考虑到查询本身的简洁性不足，我们通常忽视直接查询，而是侧重于网页上的详细内容。检索过程如下：

1. 将所有记录以嵌入形式表示，这通常涉及将用户过去的查询和网页信息转化为数值向量。
2. 使用 Contriever（比如基于检索的模型）计算查询q_j的嵌入，以评估其与记录间的相似性。
3. 根据相似性，选取与输入最匹配的上下文项k，作为个性化建议的基础。

对于基于实体中心的$\mathcal{K}_e$，检索同样基于嵌入表示，但实体链接器负责将文本中的实体链接到公共知识库。检索步骤包括：

1. 使用实体链接器将输入x中的实体映射到公共知识库的小型子图。
2. 计算这些实体在知识库中的相关性。
3. 返回与实体相关性最高的上下文项k，结合实体的上下文信息，生成个性化建议。

这两种方法都旨在通过挖掘用户过去的交互数据，利用嵌入技术，从知识库中有效地检索与当前情境相关的上下文，以增强LLM生成个性化查询的能力。

对于实体中心的$\mathcal{K}_e$知识库，检索过程基于实体和它们的时间关联。实体通过链接器映射到知识库的小型子图，然后根据实体的活跃度进行匹配。我们采用三种策略：

1. 熟悉实体：依据在$\mathcal{K}_e$出现频率排序，选取出现最频繁的前5个。
2. 不熟悉实体：与熟悉实体相反，按照出现频率反向排序，同样选取5个。
3. 过时实体：首先筛选出时间戳在最近两周内的实体，然后从中随机抽取，保持频率分布。

知乎

Experimental Setup

在评估环节，我们使用了特定的数据集和模型来验证实体中心知识库$\mathcal{K}_e$的效果。具体来说，我们利用实际的搜索历史数据以及相应的实体链接信息作为输入，构建了实验模型。实施细节包括如何从搜索引擎日志中提取有效信息，如何根据实体的出现频率、时间和活动状态定义匹配策略，以及如何运用这些策略进行检索以生成个性化建议。通过这些步骤，我们旨在量化和比较不同实体匹配策略在提升查询建议相关性和准确性方面的表现。

Data

在评估阶段，我们利用Bing搜索引擎2023年5月至7月的搜索日志数据，进行了数据预处理。首先，我们筛选出包含点击搜索结果的会话，然后专注于链接到维基百科和热门新闻站点的页面，因为这是我们的实体链接器适用范围。为保证隐私，我们排除了访问频次少于100次的用户，且遵循企业级隐私标准，仅保留至少50人请求的查询。数据集总量巨大，我们随机选择了1,000个用户作为基准，平均有493个查询、109个会话、177篇点击文章和3,053个实体。为了验证效果，我们将数据划分为两部分：近期10个会话作为预测目标，分别构建基于搜索历史（$\mathcal{K}_s$）和实体（$\mathcal{K}_e$）的个性化知识库，用于生成上下文建议。

Baselines and Our Model

通过对比，评估了我们的K-LaMP（基于LLM的个性化知识增强模型）与几种基线，这些基线仅利用搜索上下文生成建议。为了保证公正性，所有模型都基于LLMs。然而，由于RNNs和bart等传统方法在处理长上下文和复杂数据上的局限，特别是缺乏特定训练数据时，我们在主要实验中并未包括它们。另外，由于它们不擅长处理用户当前浏览网页的长期情境，直接对比可能不够准确。

模型对比包括： 1. K-LaMP: 专为处理上下文关联查询的我们提出的方法。 2. GPT-4: 作为通用模型的代表。 3. Search Context Baselines: 仅用搜索历史的基线。 4. Entity-centric Baselines: 不考虑搜索，依赖实体信息的建议系统。

这些对比展示了K-LaMP在个性化建议生成上的优越性。

Evaluation Setup

我们采用3点量表来评估模型的性能，具体指标如下： 1. 相关性（Context Relevance）：评价模型建议查询与用户当前搜索内容的实际相关性，满分2分，高分表示高度相关。

1. 用户兴趣匹配（Interest Match）：衡量建议是否能准确捕捉到用户潜在的兴趣，满分为2分，高分表示强烈兴趣相关。
2. 知识一致性（Knowledge Alignment）：考察建议查询与用户已知知识的契合度，满分2分，高分表示信息吻合度高。

通过这三个维度，我们比较了K-LaMP（基于LLM的个性化知识增强模型）与基线模型，以揭示其在生成个性化上下文关联查询方面的优势。

对于模型的评估，我们引入了第四个维度------**排名**，它衡量建议的吸引力，根据其对用户兴趣、知识和搜索上下文相关性的预期点击率进行排序。这项指标关注的是建议的实际吸引力，而非单纯的质量。我们在印度通过第三方服务商雇佣了12位评估专家，每人每小时支付11.98美元，他们接受了详细的手册培训，包括任务指导、指标解释和示例标注。总共收集了5,236条来自不同模型在Section中给出的1,309组上下文建议的主观评分，涵盖了各个模型在质量和交互吸引力上的表现。

知乎

Agreements	Metrics	Scores (↑)
Exact match	Validness	0.963
	Relatedness	0.850
	Usefulness	0.819
Cohen's kappa coefficient	Validness	0.606
	Relatedness	0.652
	Usefulness	0.622
Spearman's correlation coefficient	Ranking	0.654

除了对模型性能的直接评估，我们还采用了双重标注来保证标注质量，并利用Cohen's κ来衡量评估一致性。对于排名的评估，我们采用了Spearman等级相关系数，来分析两组排名的关联性。结果显示，评估者间存在适中至高的共识，这支持了他们对个性化上下文查询建议的有效判断。在实现上，我们选择了GPT-4⁺（版本日期为2023年6月13日）作为基准，并设置了特定的超参数⁺ $\textbf{temperature} = 0.7$和$\text{top}_p = 0.95$。

Experimental Results

Table 1: Main results on our contextual query suggestion task. The best results are highlighted in bold.

Types	Models	Validness (↑)	Relatedness (↑)	Usefulness (↑)	Ranking (↓)
Baselines	Query Suggestion	1.769	0.962	0.948	2.736
	Contextual Query Suggestion	**1.966**	1.267	1.245	2.415
	Contextual Query Suggestion w/ $\mathcal{K}_s$	1.822	1.192	1.166	2.654
Ours	**K-LaMP (Ours)**	**1.966**	**1.482**	**1.455**	**2.160**

实验结果显示，K-LaMP模型在相关性、实用性和排名方面明显领先于所有基线。尽管在有效性上与Contextual Query suggestion不相上下，这虽符合预期（因个人上下文不一定提高检索效率），但也显示出一些有趣现象。使用$\mathcal{K}_s$进行的建议并未超越常规上下文，可能是因为从记忆库检索的关联性不足，反而产生了误导。附录提供了GPT-4的超参数设置和获取响应的提示。所有数据和分析均基于严谨的评估方法，包括双重标注和Cohen'sκ以保证质量。

Table 3: Results of different retrieval strategies on Retrieval Relevance to the user's current search context.

Retrieval	Types	Retrieval Relevance (↑)
History-based Retrieval ($\mathcal{K}_s$)	Past Documents	0.299
Entity-centric Retrieval ($\mathcal{K}_e$)	Familiar Entities	**0.936**
	Unfamiliar Entities	**0.810**
	Lapsed Entities	**0.849**

发布于 2024-04-11 10:44 · IP 属地北京

LLM 微软（Microsoft） 个性化搜索