【专委会成果推荐-第十四期】SIGIR 2020 |Personal Embed for Personalized Search

原创 姚菁 中国中文信息学会信息检索专委会 2020-08-26

点击蓝字关注我吧

为促进信息检索相关领域的学术交流，中国中文信息学会信息检索专委会将定期推送国内学者在相关领域顶级国际会议上发表的优秀论文。本期介绍的论文《Employing Personal Word Embeddings for Personalized Search》发表于刚刚落下帷幕的SIGIR 2020会议。

Employing Personal Word Embeddings for Personalized Search

Jing Yao[2], Zhicheng Dou[1], Ji-Rong Wen[3,4]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]School of Information, Renmin University of China
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods
[4]Key Laboratory of Data Engineering and Knowledge Engineering, MOE
{jing_yao,dou}@ruc.edu.cn,jirong.wen@gmail.com

研究动机

个性化搜索任务是根据用户个人的兴趣爱好对查询的文档列表进行排序以更好地满足用户的信息需求。大多数现有的个性化搜索模型遵循一个通用的模式：首先基于用户的搜索历史建立用户兴趣画像，然后通过计算候选文档与用户兴趣画像的匹配得分来对文档列表重排。在本文中，我们尝试从另一个角度来解决个性化搜索问题。自然语言中存在很多具有多种不同含义的词，比如"苹果"，而知识背景和兴趣爱好不同的用户往往对这些词的具体含义会有不同的理解。那么，对于不同的用户来说，同一个词就应该具有不同的语义和表示。从这个想法出发，我们提出了基于个人词向量的个性化搜索模型（PEPS）。在模型中，我们利用用户个人的查询日志为语料集为其训练个性化词向量；然后分别获得查询和文档的个性化词表示向量和通过自注意力机制计算的上下文表示向量；最后利用一个匹配模型来计算查询和文档个性化表示之间的匹配得分。在AOL查询日志和一个商业查询日志上的实验证明我们设计的模型在效果上显著优于现有模型。

相关工作

基于深度学习的个性化模型

随着深度学习的发展，很多利用深度模型获得用户兴趣画像表示向量的工作被提出。Ge et al. 设计了一个带注意力机制的分层RNN（HRNN）模型来学习用户动态的长期和短期兴趣画像。Lu et al. 利用对抗神经网络来增强个性化模型的训练数据，优化用户画像。这些工作都是基于用户的兴趣画像来对文档进行排序，从而实现搜索结果个性化。本文从另一个全新的角度来解决个性化搜索问题，不需要建立用户兴趣画像。

基于词向量的个性化搜索模型

近年来，也有一些工作尝试使用词向量来实现个性化搜索。他们首先基于用户个人的查询日志训练词向量，然后根据词向量之间的相似度获得查询关键词的同义词来扩充查询或者作为当前查询下用户的兴趣画像，进而根据文档与扩充查询或兴趣画像的匹配程度来实现个性化排序。然而，在本文的模型中，我们直接将用户兴趣嵌入到用户个人词向量中，文档和查询的个性化表示向量匹配得分即为该文档的个性化得分。

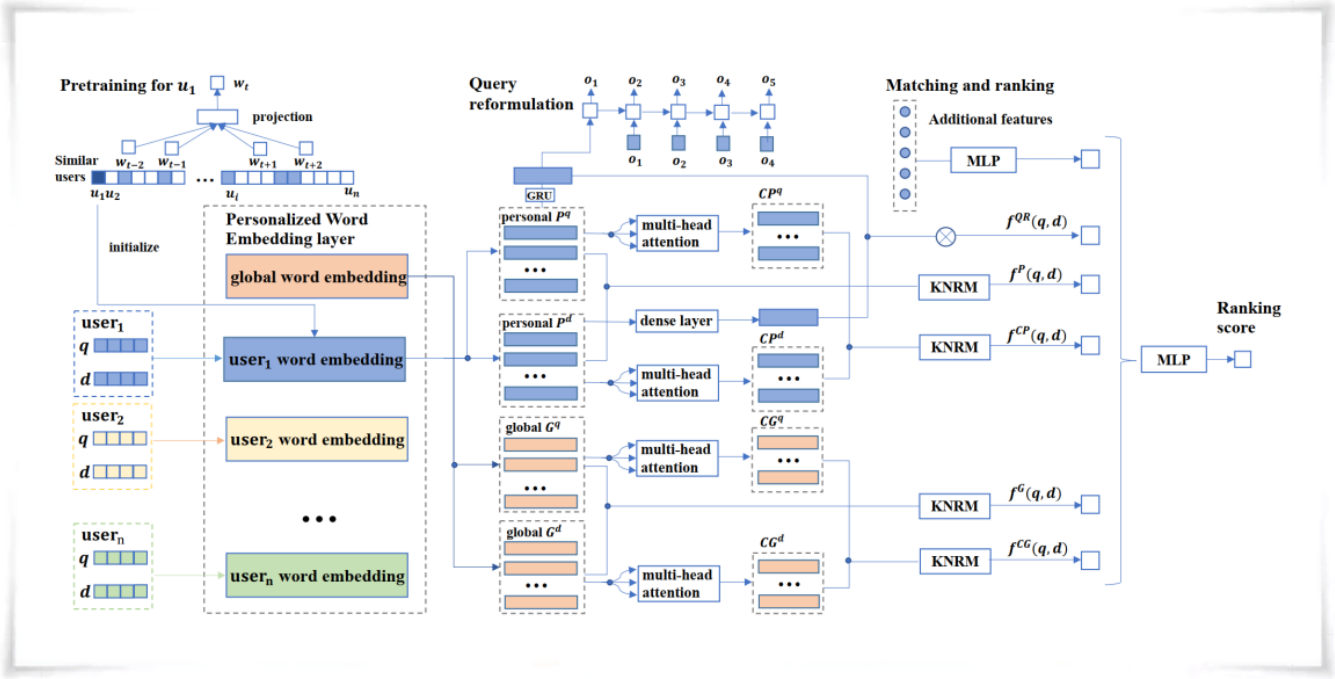

图1 PEPS模型的架构

PEPS模型：基于个人词向量的个性化搜索模型

我们提出从个人词向量的角度来解决个性化搜索问题。通过为每个用户训练个人词向量，得到的词向量主要包含用户个人对于该词所了解和感兴趣的词义，以此来表示用户的查询能够明确用户当前所表达的个性化查询意图。PEPS模型结构如图1，主要包括四个组成部分：首先，我们设置了一个个性化词向量层，包括共享的全局词向量和用户的个人词向量；然后我们获得查询和文档的表示向量；利用KNRM匹配模型来计算文档和查询的匹配得分并采用LambdaRank算法来训练整个排序模型；最后，我们加上一个查询重构模块来辅助个人词向量的学习和个性化排序模型的训练。

个性化词向量层

我们利用两个标识（单词，对应的用户编号）来标志每个单词，因此不同用户词向量中的同一个单词会被标志为不同的单词，比如用户i的单词“Apple”被表示为“Apple+i”，而“Apple+j”则表示用户j的词“Apple”。每个用户的个人词向量都只用该用户的查询日志进行训练，训练得到的个人词向量不包括该词多种不同的含义，而主要只包括该用户查询过或者感兴趣的语义。

我们还需要确定用户个人词向量中包括哪些单词，直接使用全局词汇表会浪费大量的内存空间。因此我们按照如下规则筛选全局词汇表中的单词来构建用户个人词汇表：

- 去掉全局词汇表中的停用词；
- 去掉在用户个人查询日志中出现次数小于c次的单词
- 去掉WordEntropy值小于一定阈值的单词。

WordEntropy为单词的交叉熵，由所有包含该单词的查询点击熵的平均值计算得到。

设置好个性化词向量层后，需要对其进行初始化。一般来说，词向量的预训练需要依赖一定规模的语料集合，仅基于用户个人的查询日志不太足够。因此，我们利用全局的word2vec模型来初始化用户个人词向

量，或者用若干个相似用户的查询日志训练word2vec模型来进行初始化。

文本表示模块

通过个性化词向量层，我们可以将查询和文档映射到高维空间并获得它们的表示向量，一共包括四种不同粒度和角度的文本表示向量。

(1) Personalized Word Representation：输入用户的查询和文档，通过对应用户的个人词向量矩阵会得到查询和文档的个性化词向量表示 $P^q \in R^{dim \times |q|}, P^d \in R^{dim \times |d|}$ ，实现了词级别的个性化。

(2) Personalized Contextual Representation：为了捕捉上下文之间的交互信息并获得查询级别的个性化向量表示以便明确用户的查询意图，我们在个性化词向量表示上加一个多头自注意力层对其进行处理，得到查询和文档的个性化上下文表示 $CP^q \in R^{dim \times |q|}, CP^d \in R^{dim \times |d|}$ 。查询的个性化上下文表示计算过程如下：

$$CP^q = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^A,$$

$$\text{head}_i = \text{softmax}(\frac{P^q W_i^Q (P^q W_i^K)^T}{\sqrt{d_k}})(P^q W_i^V),$$

(3) Global Word Representation：在实际搜索中，用户的兴趣是多变的，知识也是在不断增长的。因此，除了个性化的文本表示之外，我们同时也关注查询和文档的全局表示 $G^q \in R^{dim \times |q|}, G^d \in R^{dim \times |d|}$ 。

(4) Global Contextual Representation：与个性化上下文表示的计算过程相同，我们利用多头注意力层对全局词向量表示进行处理，得到查询和文档的全局上下文表示向量 $CG^q \in R^{dim \times |q|}, CG^d \in R^{dim \times |d|}$ 。

文本匹配排序模块

获得查询和文档的个性化及全局表示后，我们计算查询和文档的个性化匹配得分来对候选文档集实现重排序，使用匹配模型KNRM来计算得分。对于文档和查询的个性化词向量表示,我们首先构建一个相似度矩阵；然后利用多个RBF核函数从相似度矩阵中抽取查询和文档之间的多层级匹配特征，最后使用多层感知机计算匹配得分。

$$\phi(S^P) = \sum_{i=1}^{|q|} \log(\vec{K}(S_i^P)),$$

$$\vec{K}(S_i^P) = \{K_1(S_i^P), \ldots, K_K(S_i^P)\},$$

$$K_k(S_i^P) = \sum_j \exp(-\frac{(S_{ij}^P - \mu_k)^2}{2\sigma_k^2}).$$

$$f^P(q,d) = \tanh(W_P^T \phi(S^P) + b^P)$$

与以上个性化词向量之间匹配得分的计算相同，我们计算其他文本表示之间的匹配得分，我们还计算了查询重构模块中重构得到的查询表示向量和文档之间的匹配得分。同时，我们也引入了一些点击特征和相关性特征来帮助文档排序，将这些特征输入多层感知机中计算相关性得分。最后，组合以上所有得分得到该文档的个性化得分，并选择基于文档对的学习排序算法LambdaRank来训练个性化排序模型。

查询重构模块

日常使用搜索引擎的大多数用户很难直接用准确的查询关键词来表达自己的查询意图。在PEPS模型中，我们能够得到包含用户兴趣的个性化查询表示向量，因此我们能够基于此来推断用户真实的查询意图，并对用户的查询进行重构来提升个性化排序的效果，反过来也能够促进用户个人词向量的学习。从这个动机出发，我们设置了查询重构模块并搭建多任务框架将其与个性化排序任务同时进行训练，使用"编码器—解码器"结构并用查询序列中的后一个查询作为监督信息。

$$p(o) = \prod_{t=1}^{T} p(o_t | \{o_1, \ldots, o_{t-1}\}, h_{|q|}).$$

我们通过最小化目标序列的负对数生成概率训练查询重构模块。整个多任务模型通过最小化负对数概率和文档对损失来联合进行训练。

在线更新词向量

在实际应用场景中，用户会不断地输入新的查询，这些查询可能体现出新的用户兴趣。为了确保模型中的个人词向量蕴含最新的用户兴趣，我们应该根据用户新输入的查询来时刻调整用户个人词向量，此时可以保持排序模型的其他参数固定不变。为此，我们设计了三种不同的在线调整方式：

按阶段更新：第一步，在线下用当前的查询历史训练好一个模型；第二步，规定一个阶段时长，在这个时间内我们只收集用户点击行为，不调整词向量；第三步，在本阶段结束后，根据收集的用户点击行为来调整

对应用户的个人词向量，同时保持排序模型中的其他参数不变。不断重复第二步和第三步操作，使得词向量始终包含最近的查询历史中体现的用户兴趣。

按会话更新：在检索中，我们通常认为查询会话是反映用户查询意图和短期兴趣的单位。因此，与上一方法的操作步骤相同，我们提出以查询会话为间隔来更新词向量。

按查询更新：为了捕捉查询会话内部的短期用户兴趣，我们也设计了一种以查询为间隔来更新词向量的方式。

Table 1: Statistics of the datasets.

Dataset	AOL Dataset			Commercial Dataset		
	Train	Valid	Test	Train	Valid	Test
#session	187,615	26,386	23,040	71,731	13,919	12,208
#query	814,129	65,654	59,082	188,267	37,951	41,261
avg query len	2.845	2.832	2.895	3.208	3.263	3.281
avg #click	1.249	1.118	1.115	1.194	1.182	1.202

实验设置

数据集

我们在两个非个性化的搜索日志上进行实验，即公开的AOL查询日志和一个商业搜索引擎的查询日志。AOL数据集记录了2006年3月1日至2006年5月31日的查询日志。我们依据连续查询的相似度大小将用户的查询序列划分为若干个查询会话，前5周的日志作为用户历史，之后的日志按照6:1:1划分为训练集、验证集和测试集。由于AOL数据集中只记录了每个查询下的点击文档，我们参考Ahmad et al.利用BM25算法来构建同时包含点击和未点击文档的候选文档列表。商业数据集记录了2013年1月1日至2013年2月28日两个月的查询日志。以30分钟的不活动时间为间隔划分查询会话，前6周的日志设为用户历史，之后的日志按照4:1:1划分训练集、验证集和测试集。

评价指标

我们选取最常用的排序评价指标MAP、MRR以及P@1来评估模型。此外，考虑到用户是否点击某个文档不仅仅取决于该文档的相关性，同时也会受排序位置的影响，我们还采用了Lu et al. 提出的P-Improve评价指标。

基线模型

除了与查询日志记录的原始文档排序进行对比，我们还选取了目前表现最好的排序模型和个性化搜索模型作为基线模型。排序模型包括KNRM和Conv-KNRM模型，个性化搜索模型包括P-Click、SLTB、HRNN、PSGAN、PPWE和PWEBA。

实验结果

Table 2: Overall performances of models. Relative performances compared with PSGAN are in percentages."†" indicates significant improvements over all baselines with paired t-test at $p < 0.05$ level, and ‡ for t-test at $p < 0.01$ level. The best results are shown in bold. PEPS(fix) means the personalized word embedding layer is fixed during training.

Model	AOL Dataset						Commercial Dataset							
	MAP		MRR		P@1		MAP		MRR		P@1		P-Imp.	
Adhoc search model														
Ori.	.2504	-54.3%	.2596	-53.6%	.1534	-68.6%	.7399	-9.1%	.7506	-8.8%	.6162	-14.1%	-	-
KNRM	.4291	-21.7%	.4391	-21.6%	.2704	-44.7%	.4916	-39.6%	.5001	-39.3%	.2849	-60.3%	.0655	-73.7%
ConvK	.4738	-13.5%	.4849	-13.4%	.3266	-33.2%	.5872	-27.8%	.5977	-27.4%	.4188	-41.6%	.1422	-42.9%
User profile based personalized search model														
PClick	.4224	-22.9%	.4298	-23.3%	.3788	-22.6%	.7509	-7.7%	.7634	-7.3%	.6260	-12.7%	.0611	-75.5%
SLTB	.5072	-7.5%	.5194	-7.3%	.4657	-4.8%	.7921	-2.6%	.7998	-2.9%	.6901	-3.8%	.1177	-52.7%
HRNN	.5423	-1.0%	.5545	-1.0%	.4854	-0.8%	.8065	-0.9%	.8191	-0.5%	.7127	-0.7 %	.2404	-3.4%
PSGAN	.5480	–	.5601	–	.4892	–	.8135	–	.8234	–	.7174	–	.2489	–
Embedding based personalized search model														
PWEBA	.4284	-21.8%	.4368	-22.0%	.2687	-45.1%	.7415	-8.9%	.7529	-8.6%	.6201	-13.6%	.0433	-82.6%
PPWE	.6542‡	19.4%	.6668‡	19.1%	.5613‡	14.7%	.8138	0.1%	.8249	0.2%	.7187	0.2%	.2338	-6.1%
PEPS(fix)	.6971‡	27.2%	.7107‡	26.9%	.6153‡	25.8%	.8209†	0.9%	.8310†	0.9%	.7232†	0.8%	0.2516	1.1%
PEPS	**.7127**‡	30.1%	**.7258**‡	29.6%	**.6279**‡	28.4%	**.8221**†	1.1%	**.8321**†	1.1%	**.7251**†	1.1%	**.2545**†	2.3%

模型整体表现

表1记录了所有基线模型和PEPS模型在整体数据集上的实验结果。我们发现：(1) 相比所有的基线模型，PEPS模型在两个数据集的所有评价指标上都有显著提升，在AOL数据集上通过了$p<0.01$的T检验，在商业数据集上能通过$p<0.05$的T检验。(2) 所有的个性化搜索模型都能够大幅度提升原始排序的结果，表明搜索结果个性化能够有效提升用户的搜索体验。

Table 3: Results of ablation experiments. Relative performances compared with complete PEPS are in percentages. PWE/GWE means personal/global word embeddings.

PEPS Variant	AOL Dataset						Commercial Dataset							
	MAP		MRR		P@1		MAP		MRR		P@1		P-Imp.	
PEPS	.7127	-	.7258	-	.6279	-	.8221	-	.8321	-	.7251	-	.2545	-
w/o Attn.	.6869	-3.62%	.7008	-3.44%	.6021	-4.11%	.8145	-0.84%	.8254	-0.83%	.7196	-1.18%	.2446	-4.08%
w/o Attn, PWE	.6693	-6.09%	.6823	-5.99%	.5771	-8.09%	.8126	-1.07%	.8242	-0.97%	.7181	-1.39%	.2388	-6.35%
w/o Attn, GWE	.6686	-6.19%	.6822	-6.01%	.5796	-7.69%	.8139	-0.91%	.8249	-0.89%	.7191	-1.25%	.2418	-5.18%
Ablation on query reformulation														
w/o Multi-task	.7113	-0.20%	.7246	-0.17%	.6266	-0.21%	.8186	-0.34%	.8295	-0.34%	.7256	-0.36%	.2513	-1.45%
w/o Query Ref	.7101	-0.36%	.7232	-0.36%	.6247	-0.51%	.8202	-0.15%	.8306	-0.20%	.7253	-0.40%	.2392	-6.20%

消融实验结果

PEPS模型包括几个主要的组成结构：个人词向量层、文本表示模块和查询重构模块。为了进一步分析每个模块对于个性化的作用，我们进行了消融实验。表2的结果表明几个模块对于搜索结果个性化都有一定的作用，尤其是词向量层中的个人词向量和全局词向量，以及文本表示模块的多头自注意力机制。

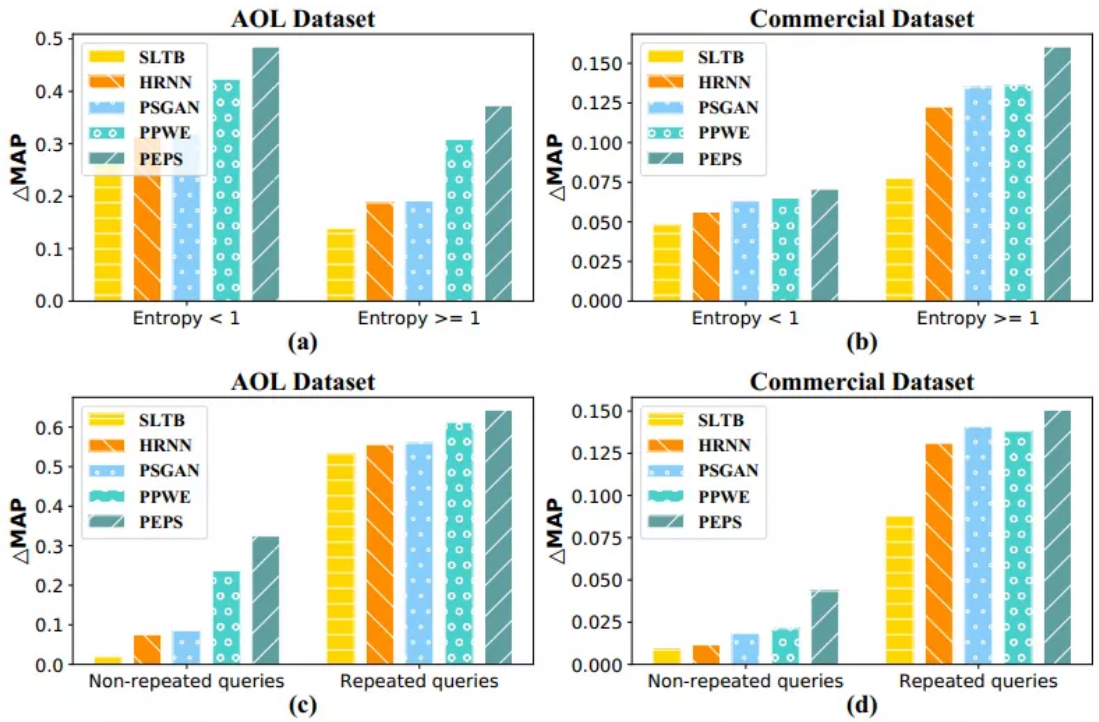

Figure 2: Experimental results on different query sets. (a) and (b) are results about queries with different entropies, (c) and (d) are results on repeated/non-repeated queries.

不同查询集合分析

为了分析PEPS模型在不同查询、不同场景下的效果，我们将所有查询按照不同的标准划分为不同的集合，并观察PEPS模型在不同查询集合上的表现，结果如图。(1) 根据查询的点击熵是否大于1将其划分为两个集合informational queries和navigational queries。PEPS模型在两个集合上都能取得最好的效果，在navigational queries集合上表现更突出。(2) 根据查询是否在用户历史中出现过将其划分为重复查询和非重复查询集合。所有的个性化模型都倾向于在重复查询集合上表现出更好的效果，但是相比之下，PEPS模型在非重复查询上的提升更显著。

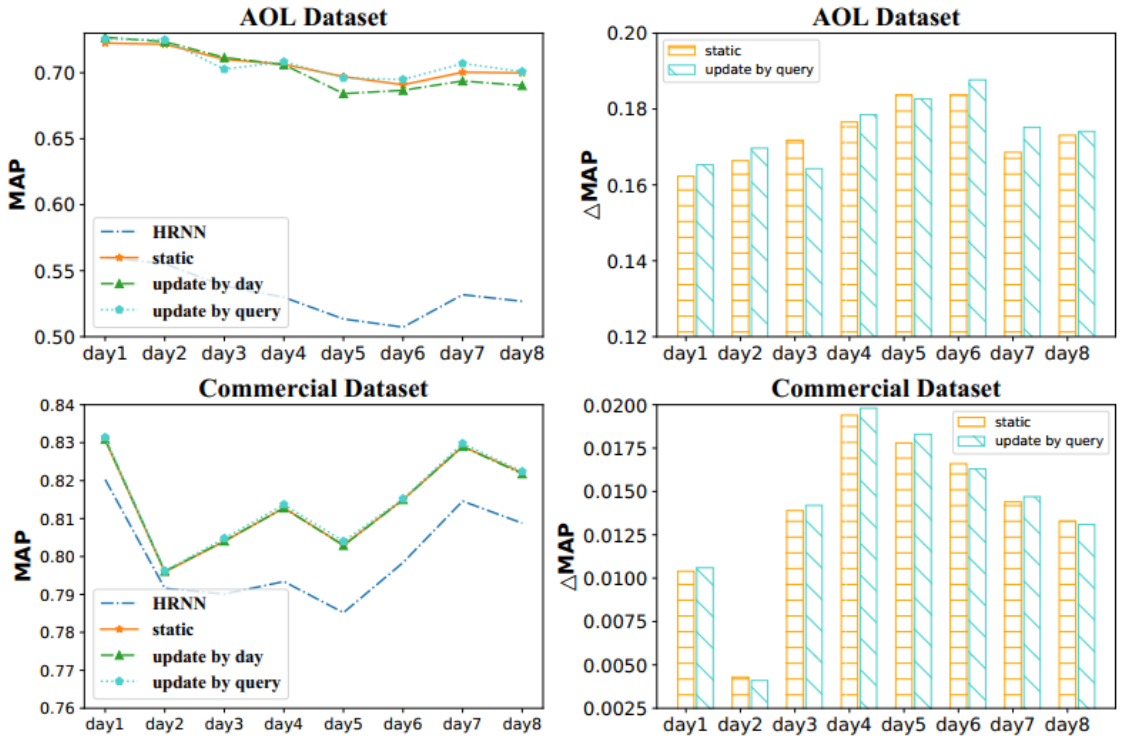

Figure 3: Performance of different online update methods.

在线更新词向量

为了使个人词向量包含该用户在新输入的查询中体现的兴趣，我们设计了三种不同的词向量在线更新方式，并且在测试集上进行了模拟实验，结果如图。以查询为单位进行更新的方式表现最好，但是总体来看，在线词向量更新在前期表现较好而在后期可能会比静态词向量更差。我们认为可能是增量调节的方式难以使模型达到最优状态，因此，我们建议采用短期的在线更新，一段时间后重新训练模型。

总结

大部分现有的个性化搜索模型基于用户历史构建用户兴趣画像，然后根据用户兴趣画像来对文档列表实现重排。在本文中，我们从明确用户利用当前查询关键词所表达的查询意图出发来实现查询结果个性化。我们认为不同的用户对同样的单词会有各自不同的理解，并根据这个想法提出了一个基于用户个人词向量的个性化搜索模型，结果表明我们的模型能够有效提高个性化搜索的效果。

编辑：庞 亮 毛佳昕