

基于查询意图识别与主题建模的文档检索算法

严 锐, 李石君

(武汉大学 计算机学院, 武汉 430072)

摘 要: 传统的搜索引擎仅返回给用户包含查询关键字的文档, 忽略了查询背后用户真正的信息需求。为此, 将文档检索看作个性化推荐问题, 提出一种查询意图识别的主题模型个性化检索算法。对用户检索历史进行潜在狄利克雷分布主题建模, 结合检索历史主题模型识别用户查询的潜在意图, 并按主题相关度进行文档推荐, 计算查询到文档集的 KL 距离对文档集排序, 最终返回给用户个性化检索文档列表。实验结果表明, 与基于协同相似计算和基于用户聚类的推荐算法相比, 该算法能够更准确有效地为用户提供个性化检索。

关键词: 搜索引擎; 查询意图; 文档检索; 个性化推荐; 主题模型; 潜在狄利克雷分布; KL 距离

中文引用格式: 严 锐, 李石君. 基于查询意图识别与主题建模的文档检索算法[J]. 计算机工程, 2018, 44(3): 189-194.

英文引用格式: YAN Rui, LI Shijun. Document Retrieval Algorithm Based on Query Intent Identification and Topic Modeling[J]. Computer Engineering, 2018, 44(3): 189-194.

Document Retrieval Algorithm Based on Query Intent Identification and Topic Modeling

YAN Rui, LI Shijun

(School of Computer, Wuhan University, Wuhan 430072, China)

[Abstract] Conventional search engines collect documents which only contain key words in the query, but not considering the true intent hidden inside its users. Aiming at this problem, taking the document retrieval as a personalized recommendation problem, this paper proposes a personalized retrieval algorithm based on query intent identification and topic model. First, the topic model of Dirichlet Distribution Allocation (LDA) is applied for modeling the historical search data of its user. When a new query comes, latent topic of the query is recognized by the topic model of the historical search of its user, and then appropriate documents are recommended for the correlation of topics. Finally, the KL distance between the query and document sets is calculated, and the documents returning to the user are sorted according to the distance. Experimental results show that the proposed algorithm is better than the method based on collaborative similarity calculation and the method based on user interest clustering on efficiency.

[Key words] search engine; query intent; document retrieval; personalized recommendation; topic model; Latent Dirichlet Allocation (LDA); KL distance

DOI: 10.3969/j.issn.1000-3428.2018.03.032

0 概述

搜索引擎使得人们能够从互联网上快速获取信息^[1], 但用户输入的大部分查询语句都比较短, 一个查询语句中通常包含了多种检索意图, 对于不同用户而言, 一个相同的查询语句所表达的检索意图可能是不同的^[2], 例如搜索关键字为“成都”, 有些人想检索的是成都的地图, 有些人感兴趣的是成都的旅游, 还有一些人可能想要搜索的是赵雷的歌曲《成都》。目前的搜索引擎主要是基于关键字匹配的搜

索模式, 搜索算法如果仅仅只考虑文档中是否包含查询关键字信息, 很有可能返回给用户的检索结果中只包含了用户的部分信息需求, 只覆盖了部分检索意图, 无法充分满足用户需要^[3-4]。与之相比, 个性化检索除了考虑文档与查询语句语义上的相似性, 还结合了用户的兴趣爱好, 不同的人在同样的外部条件下, 可以获得与其本身兴趣爱好、历史记录等相匹配的推荐, 个性化检索能够为用户返回更满意的检索结果^[5]。

个性化检索本质上是一种对查询的个性化推

基金项目: 国家自然科学基金(61272109); 国家自然科学基金青年基金(61502350)。

作者简介: 严 锐(1992—), 男, 硕士研究生, 主研方向为 Web 数据挖掘、自然语言处理; 李石君, 教授、博士生导师。

收稿日期: 2017-02-20 **修回日期:** 2017-03-21 **E-mail:** yanrui1992@qq.com

荐^[6],推荐算法近几年得到了广泛和深入的研究,可以分为协同过滤推荐、基于统计学的推荐、针对内容推荐和混合模型推荐等。其中协同过滤^[7]推荐方法最为常见,其原理是有相同兴趣爱好的用户可能会喜欢相同的物品,因此,协同过滤方法又可以分为对用户、物品的推荐和基于模型的推荐。前者常常需要将用户相关数据读入内存中进行计算,被称为基于内存的协同过滤;后者利用 pLSA 模型、LDA 主题模型和聚类算法等模型,被称为基于模型的协同过滤,需要预先训练数据集。文献[8]将相似用户历史行为数据扩充到单个用户数据集中,应用点击传播产生查询推荐。文献[9]运用改进的 K-means 算法,先形成核心用户聚类,进而形成全部用户聚类,基于用户聚类的算法能够为相似的用户推荐,从一定程度上解决了数据稀疏的问题,但算法复杂性较高,而且随着用户数量逐渐增大,算法扩展性不高。文献[10]和文献[11]都是基于用户标签的微博推荐方法,前者通过关联规则,后者利用 TextRank 算法,算法能够为用户推荐微博,但都是基于关键词的推荐,没有挖掘用户潜在兴趣。文献[12]利用用户浏览日志的上下文进行新闻推荐。文献[13]利用长短期兴趣从搜索日志中获得表达需求的查询模型,这 2 个算法有较高的扩展性,但都需要收集大量的日志和上下文相关数据。

本文提出一种新的文档检索算法。通过对用户查询日志进行主题意图建模,利用用户查询历史中的主题信息扩展查询意图并进行识别,进而推荐文档。同时根据 KL 距离对文档按相关性排序,最终生成推荐列表。

1 本文算法

用户搜索历史能够反映用户对不同主题的兴趣程度,因此,通过对用户查询日志建立合理的兴趣模型,能够扩展用户当前查询主题的意图,并且可以消除查询语句的歧义。将个性化推荐应用到个性化检索中,有利于准确检索出用户想要的信息。

本文提出了一种基于查询意图识别的主题模型检索算法,如图 1 所示,主要分为 3 个步骤:1) 对用户检索历史进行主题建模;2) 对用户查询进行意图识别;3) 针对识别出的意图进行个性化检索。将第 1) 步得到用户兴趣主题模型在检索阶段之前融入到查询语句中,最后对检索的文档进行相关性排序,形成检索推荐列表。当用户的一次检索任务完成时,将查询语句与检索结果列表信息添加到用户检索历史中,并用查询语句和检索文档更新该用户的检索历史主题模型。

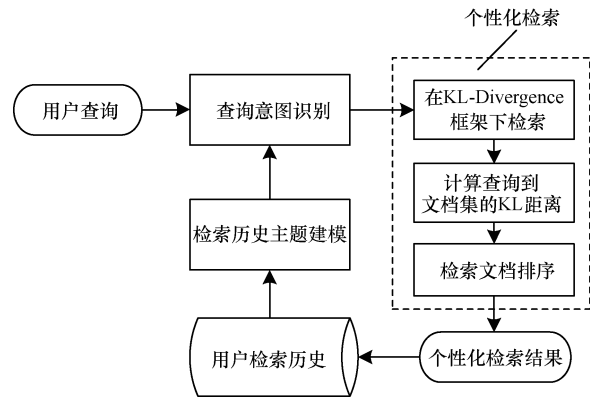


图 1 基于查询意图识别的主题模型检索框架

1.1 用户检索历史主题建模

1.1.1 LDA 主题模型

潜在狄利克雷分布 (Latent Dirichlet Allocation, LDA) 主题模型是一种多层次的贝叶斯概率主题模型。该模型以概率生成作为条件,分为文档、主题和词 3 层结构^[14]。LDA 模型的基本原理为:文档集中共有 K 个主题,每个文档 d 是这 K 个主题上的多项分布,每个主题 z 是词典 V 上的一元语言模型,即词典上的一个多项分布,模型通过对文档建模挖掘出文档中的潜在主题。

LDA 的概率图模型如图 2 所示,其中, W 表示观察到的变量, θ 表示文档在主题上的后验概率分布, φ 表示主题在词语上的后验概率分布, $\alpha \sim Dir(\alpha)$ 是 θ 的先验分布的参数, $\beta \sim Dir(\beta)$ 是 φ 的先验分布的参数, M 为文档集中文档的总数, N 为每篇文档的词语数, K 为主题数^[15]。

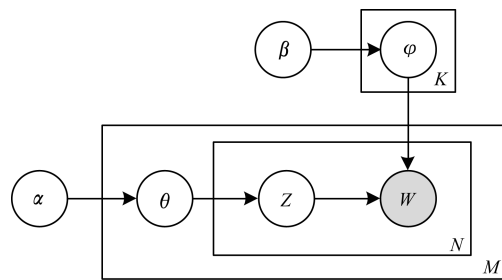


图 2 LDA 概率图模型

用条件概率 $p(z|d)$ 表示一个文档 d 中主题 z 的分布概率,用条件概率 $p(w|z)$ 表示每一个主题 z 中的一个单词 w 的分布概率,则单词 w 在文档 d 中的分布概率公式为:

$$p(w|d) = \sum_{i=1}^K p(w|z_i)p(z_i|d) \quad (1)$$

LDA 模型的参数可以通过吉布斯重复抽样^[16],它是一种快速高效的马尔科夫链蒙特卡洛抽样方法,根据 Markov 链的性质,模型最终会收敛。

1.1.2 检索历史主题模型

用户检索历史反映了用户过去的所有查询和文档的相关信息。用户检索历史可以表示为 $H = \{h_1, h_2, \dots, h_x\}$, $h = [query, documents]$, $query$ 和 $documents$ 分别表示查询关键字及用户点击的网页文档集合, 文档集合表示为 $S = \{d_1, d_2, \dots, d_m\}$, 每个文档 d 都对应着一个检索项, 是由检索项链接指向的网页中标题和正文两部分组成。这样, 整个用户检索历史可以表示为 $C = \{S_1, S_2, \dots, S_n\}$ 。通过对用户检索历史进行 LDA 主题建模, 将文档集 C 聚类成 K 个描述用户兴趣的主题, 用户检索历史主题模型如图 3 所示。

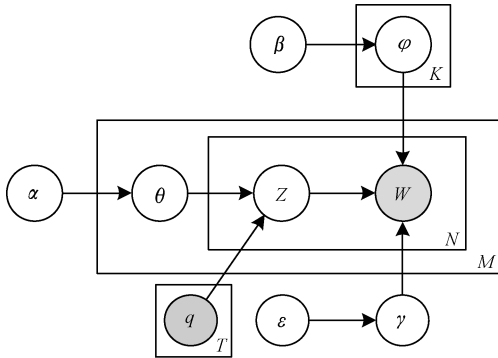


图3 用户检索历史主题模型

在图3中, 灰色圆圈 q 表示查询 q 的主题分布 $p(Z|q)$, T 表示查询 q 中所包含的主题数, 查询 q 对应的搜索文档集中文档 d 的一个词语 w , 可以用一个混合模型表示为:

$$p(w|d, q) = p(d|q)p(w|\gamma_\beta) + (1-\lambda) \sum_{i=1}^K p(z_i|d)p(w|z_i) \quad (2)$$

其中, $p(d|q)$ 表示文档 d 满足查询 q 的概率, γ_β 表示背景模型, 用于检测出现次数较多但与具体主题关系不大的词语, 这些词语出现频率较高但是不具备很强的区分度和信息量。当这类词在背景模型 γ_β 中有较高权重时, 在具体主题模型中的权重就较低, 也即 λ 值越大, 主题模型的区分度越高。用户检索历史文档的生成过程如下:

- 1) 从查询日志中选择查询语句 q 。
- 2) 以概率 $p(d|q)$ 选择文档 d 。
- 3) 从狄利克雷分布 α 中取样生成文档 d 的主题分布 θ 。
- 4) 对于文档 d 中的每个词语重复以下过程:

(1) 判断词语是否由背景模型生成, 如果是, 则从 $p(w|\gamma_\beta)$ 中生成 w ; 否则从 Dirichlet 分布 β 中取样生成主题 k 对应的词语分布 ϕ 。

(2) 需要估计的参数有 $\alpha, \beta, \varepsilon$ 和 $p(d|q)$ 。其中 α, β 的估计采用吉布斯重复抽样, 当抽样次数达到一定值后, 本文用户检索历史的隐含主题会逐渐服从于 Dirichlet 先验分布, 此时参数 α, β 收敛。

(3) γ_β 是背景主题, 用于描述全局经常出现的词语, 对其使用最大似然估计, γ_β 中参数 ε 的分布如式(3)所示。

$$p(w|\varepsilon) = \frac{\sum_d (w, d)}{\sum_{w'} \sum_d c(w', d)} \quad (3)$$

其中, $c(w, d)$ 表示词语 w 在文档 d 中出现的次数。

(4) 估计 $p(d|q)$, $p(d|q)$ 决定了文档 d 中的词在主题模型中的权重, 定义函数 $g(d|q) = \{0, 1, 2, 3\}$ 表示文档 d 与用户查询 q 的相关程度, 因此, $p(d|q) \propto g(d, q)$, $p(d|q)$ 的计算公式如式(4)所示。

$$\sum_{i=1}^M p(d_i|q) = 1 \quad (4)$$

$g(d, q)$ 取值 0、1、2、3 分别表示文档 d 与用户查询 q 不相关、一般相关、基本相关和完全相关 4 个关联程度, 对用户检索历史主题建模, 将用户兴趣分布在 K 个不同的主题上, 可用式(5)表示用户对每个主题的感兴趣程度:

$$p(z|C) \propto \sum_q \sum_{d \in S_q} p(d|q)p(z|d) \quad (5)$$

1.2 查询意图识别

用户兴趣主题可表示为 $Z = \{z_1, z_2, \dots, z_K\}$, 采用贝叶斯的方式: 当用户输出一个新的查询语句, 将用户的历史检索信息融入到当前的查询中, 对当前查询进行主题扩展和意图识别, 用 $p(k|q)$ 表示查询 q 属于主题 k 的概率, $p_H(w|Z_q)$ 为历史检索模型, 其计算公式为:

$$p_H(w|Z_q) = \sum_{k=1}^K p(w|z_k)p(z=k|q) \quad (6)$$

$p(Z=k|q)$ 用于表示查询与用户检索历史主题模型的相似度, 采用余弦相似度来度量^[17], 利用贝叶斯定理, $p(Z=k|q)$ 的计算公式为:

$$p(z=k|q) = \frac{p(q, z=k)}{\sum_{j=1}^K p(q, z=j)} \propto \frac{\text{sim}(z_q, z_j)}{\sum_{j=1}^K \text{sim}(z_q, z_j)} \quad (7)$$

$$\text{sim}(z_q, z_j) = \cos(\mathbf{V}(z_q), \mathbf{V}(\cdot)) \quad (8)$$

其中, $\mathbf{V}(\cdot)$ 表示一个空间向量。用户的检索历史中与当前查询相关的主题在检索模型中有更高的权重。通过计算 $p(z=k|q)$ 的值来对查询意图主题进行排序, 选择前 T 个主题作为意图主题, 可以减少模型的稀疏性, 从而提高算法的检索效率, 也有利于缩小检索空间, 减少检索噪音^[18]。

1.3 个性化检索

个性化检索本质上是一种个性化推荐,即从已有的文档集合中推荐出一个与当前查询符合程度由高到低的文档列表。先在 KL-Divergence 搜索框架^[19]下进行处理得到检索文档,再将用户检索历史主题模型融入当前查询,个性化检索的任务可形式化表示为:计算查询语句 Z_q 与文档 Z_d 的 KL 距离,根据 KL 距离对检索文档进行排序。KL 距离的计算方法如下:

$$D(Z_q, Z_d) = \sum_{w \in V} p(w|Z_q) \ln \frac{p(w|Z_q)}{p(w|Z_d)} \quad (9)$$

一个新的查询可能是一个短语也可能是一句话,查询除了本身包含的信息还可能隐含查询者的意图信息,因此,定义新的查询模型 Z_q 如下:

$$p(w|Z_q) = \mu p_H(w|Z_q) + (1 - \mu) p_{ML}(w|Z_q) \quad (10)$$

$p_{ML}(w|Z_q)$ 是对查询本身的最大似然估计,计算方法为:

$$p_{ML}(w|Z_q) = \frac{c(w, q)}{|q|} \quad (11)$$

其中, $c(w, q)$ 表示词语 w 在查询 q 中出现的次数, $p_H(w|Z_q)$ 表示用户的历史检索模型,参数 μ 用于调节用户历史兴趣当前查询的影响。

1.4 算法描述

根据图 1 和上述推导,本文查询意图识别的主题模型检索算法描述如下:

输入 用户检索日志和对应文档集,当前查询 q , 参数 $\alpha, \beta, \varepsilon, \lambda$ 和 μ

输出 个性化检索列表

- 1) 初始化文档数 M 、主题数 K 和意图主题数 T 。
- 2) For each user U :
- 3) 根据式(4)计算所有的 $p(d|q)$ 。
- 4) 以概率 $p(d|q)$ 选择一篇文档 d :
- 5) 从 Dirichlet 分布中采样生成文档 d 的主题分布 θ 。

6) 对于文档 d 中的每个词语重复以下过程:是否由背景模型生成,是则从 $p(w|\gamma_\varepsilon)$ 中生成 w , 否则从 Dirichlet 分布 β 中取样生成主题 k 对应的词语分布 φ 。

7) 根据式(6)和式(7)计算 q 与检索历史的相似度,根据 $p(Z=k|q)$ 的值降序排序,得到查询 q 的前 T 个潜在相关主题。

8) 根据式(10)中新查询模型对查询 q 进行扩展,在 KL-divergence 检索框架下检索。

9) 根据式(9)计算 10) 中推荐的文档集中的每一个文档到查询 q 的 KL 距离,对文档依据相关性由高到低排序,得到个性化检索列表。

10) 将本次检索的查询和检索列表信息添加到相应用户的检索历史,并用检索列表中的文档更新用户检索历史主题模型。

本文算法分为在线的意图识别检索和离线主题模型构建两部分。耗时较多的是离线的用户检索历史主题建模,吉布斯采样需要对每一个用户建模,但离线建模过程不影响检索性能。在线查询意图识别过程先确定查询的前 T 最大概率主题,需要将查询与所有主题进行相似度计算,时间复杂度为 $O(K)$ 。检索过程利用 KL-Divergence 框架,这是一种用语言模型检索框架,用于衡量 2 个概率分布的距离,本文中的 2 个分布指的是“查询-主题”分布和“文档-主题”分布,KL 散度检索过程的时间复杂度为 $O(T \times K) < O(K^2)$,因此,本文算法的时间复杂度为 $O(K^2)$,其中, K 为主题数。

2 实验结果与分析

2.1 实验数据

本文的实验数据来自于搜狗实验室中的搜索引擎查询日志,共 1 126 374 条,日志时间为 2008 年 6 月。每一条查询日志包括访问时间、用户 ID、查询词、该 URL 在返回结果中的排名、用户点击的序号和用户点击的 URL,对每条查询 URL 对应的网页进行爬取,将每一个网页的标题和正文拼接成单个文档组成实验文档集。选择有 50 条以上查询记录的用户,共 275 091 条查询进行实验,将每个用户的 80% 查询记录作为实验训练数据集,其余的作为测试数据集。算法用 Python 语言编写,分词工具使用哈尔滨工业大学的语言技术平台云 LTP-Cloud 系统。表 1 中列出了本文实验中用到的所有参数及其取值范围。

表 1 实验参数

参数	取值范围
α	[0.5, 1)
β	(0, 0.5)
ε	(0.1, 0.5)
λ	(0.8, 1)
μ	(0, 1)
主题数 K	10 ~ 100
查询意图主题数 T	[1, 5]
迭代次数	100 ~ 200

超参数 $\alpha, \beta, \varepsilon$ 作为伪计数,在取值范围内取一个随机数; λ 取中间值 0.9 使主题模型具有较高的区分度;一个查询所涉及用户兴趣主题总是有限的,本实验中 T 取 5,即选择用户最感兴趣的前 5 个主图作为意图主题。

2.2 实验设计

2.2.1 主题数

文献[13]中提出了 Perplexity 指标来度量主题概率模型的性能,取值越小,模型的性能越好,Perplexity 值计算公式如下:

$$\text{Perplexity}(D) = \exp \left\{ - \frac{\sum_{d=1}^M \text{lb} p(w_d)}{\sum_{d=1}^M N_d} \right\} \quad (12)$$

其中, D 为测试文档集, M 为文档的数量, w_d 为文档集中第 d 篇文档中可以观测到的词语, N_d 表示第 d 篇文档中的词语数。通过设置不同的参数值在迭代次数为 100、150 和 200 次的情况下进行多次实验求平均值,比较不同主题数 K 的 Perplexity 值,实验结果如图 4 所示。

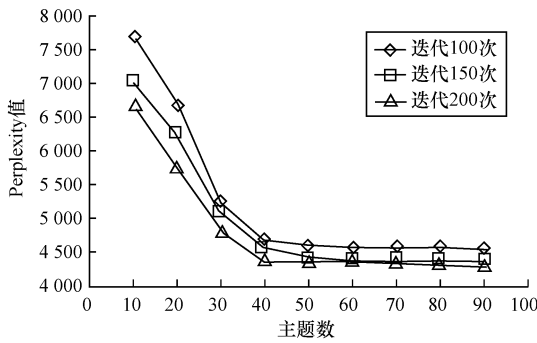


图4 不同主题数下的 Perplexity 值比较

吉布斯采样 LDA 最大的特点是每次迭代内部要经历三重 for 循环,因此每轮迭代经历的循环数是 $M \times N \times K$,算法的时间复杂度为 $O(M \times N \times K)$ 。从图 4 中可以看出,主题数越多,迭代次数越多,模型的 Perplexity 值越低;当主题数为 50 时,再次增加主题数,Perplexity 的减少不再明显。综合考虑算法的时间复杂度和模型的 Perplexity 值,本文选取主题数 $K=50$ 进行实验。

2.2.2 参数

$\mu \in [0, 1]$ 表示用户检索历史信息对当前查询的影响程度,采用信息检索领域常用于查询推荐的评价指标平均准确率^[20] (Mean Average Precision, MAP)是对所有测试查询的平均值,计算公式如下:

$$M_{\text{MAP}} = \frac{1}{M} \sum_{i=1}^M AP_i \quad (13)$$

$$AP_i = \frac{\sum_{j=1}^N P_i(j) \cdot \rho_{ij}}{\sum_{j=1}^N \rho_{ij}} \quad (14)$$

其中, AP_i 是对单个查询在不同召回率点上的正确率进行平均得到的精确度, M 表示测试查询集合中查询的个数,对于每个测试查询,预测一个长度为 N 的推荐列表, $P_i(j)$ 表示第 i 个查询的预测列表在位置 j 的准确率, $\rho_{i,j}$ 是一个示性函数,表示第 i 个查询的预测列表的第 j 个文档是否是正确的推荐,当它是一个正确推荐时, $\rho_{i,j} = 1$, 否则 $\rho_{i,j} = 0$ 。

$\mu = 0$ 时表示只考虑当前检索的内容而不考虑用户的检索历史, $\mu = 1$ 时表示只考虑用户的检索历史而不考虑当前检索的内容。从图 5 中可以看出,一开始不考虑用户检索历史,也即 μ 值很小时,得到的 MAP 也很小;随着 μ 取值的逐渐增大,MAP 的值也逐渐增大,在 $\mu = 0.7$ 时 MAP 达到最大值;之后 μ 值增大,用户检索历史所占比重增大反而使得 MAP 减小。接下来的实验中取 $\mu = 0.7$ 。

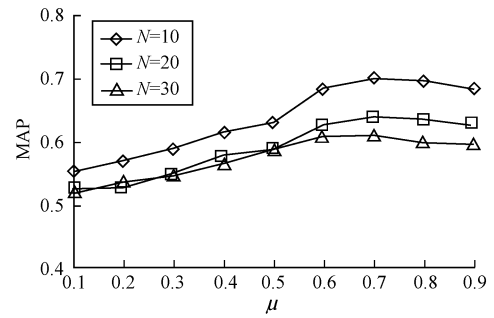


图5 μ 值对检索性能的影响

2.2.3 算法性能比较

用户往往关注检索列表中位置靠前的结果,所以检索列表的性能取决于靠前位置检索结果的相关性,引入归一化折扣增益值^[21] (Normalized Discounted Cumulative Gain, NDCG) 作为预测推荐列表的评价指标,计算公式如下:

$$N_{\text{NDCG}} = \frac{1}{M} \sum_{i=1}^M \frac{DCG_i}{IDCG_i} \quad (15)$$

$$DCG_k = \sum_{j=1}^N \frac{2^{g(d_j, q_k)} - 1}{\text{lb}(j+1)} \quad (16)$$

其中, M 表示测试查询的个数, N 表示推荐列表的大小, $IDCG_i$ 是第 i 个查询推荐列表所有可能取到的最大的 DCG_i 值 (Idea DCG), 即当该列表中所有正确的结果均位于列表最前面时的 DCG 值, $g(d_j, q_k)$ 表示第 k 个查询对应的推荐列表中第 j 个文档与查询 q_k 的相关性,计算方法见式(4)。

为了验证本文基于查询意图识别的主题模型检索算法的有效性及其推荐结果的相关性,将本文算法与基于协同相似计算的查询推荐算法和基于用户聚类的推荐算法做比较,测试结果如表 2 所示, NDCG 的对比如图 6 所示,其中,准确率表示符合需求的文档在推荐列表中所占的比例。

表 2 3 种算法检索性能比较

N	用户聚类推荐		协同相似推荐		意图识别推荐	
	准确率	NDCG	准确率	NDCG	准确率	NDCG
5	0.926	0.489	0.928	0.586	0.931	0.624
10	0.911	0.413	0.909	0.569	0.912	0.608
15	0.887	0.382	0.895	0.518	0.892	0.591
20	0.881	0.365	0.881	0.493	0.885	0.584
25	0.821	0.349	0.823	0.475	0.82	0.558
30	0.769	0.323	0.787	0.45	0.793	0.523

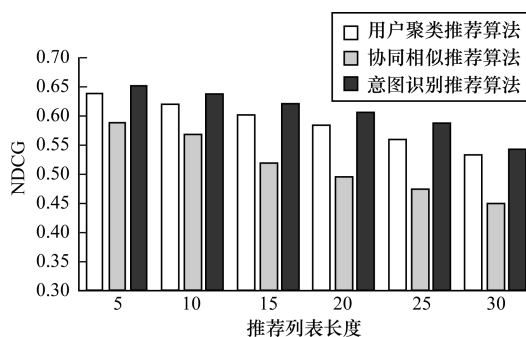


图 6 3 种算法 NDCG 比较

由表 2 可以看出,3 种算法在推荐准确率上相差不多,都能从一定程度上实现个性化推荐,但用户检索时更希望最满意的文档排在检索列表的最前面,结合表 2 中各算法在实验中得到的 NDCG 值和图 6,可以看出本文基于查询意图识别的个性化检索算法取得了较高的 NDCG 值,即能够更好地根据查询推荐相关文档。

3 结束语

本文提出了一种基于查询意图识别与主题建模的文档检索算法。首先通过用户检索历史对用户兴趣进行主题建模,当用户查询检索时,将其检索的历史融入当前查询,结合背景模型和当前查询关键字信息进行主题意图识别。实验数据采用搜狗实验室提供的真实用户检索数据,并在 KL-Divergence 检索框架下进行处理,根据查询与文档集之间的 KL 距离对检索文档排序最终形成推荐。实验结果表明,本文算法能够有效地实现个性化检索。下一步工作是将该算法结合用户兴趣聚类解决数据稀疏及冷启动等问题。

参考文献

- [1] BÜTTCHER S, CLARKE C L A, CORMACK G V. Information Retrieval: Implementing and Evaluating Search Engines[M]. Cambridge, USA: MIT Press, 2016.
- [2] GU J, FENG C, GAO X, et al. Query Intent Detection Based on Clustering of Phrase Embedding [C]//Proceedings of Chinese National Conference on Social Media Processing. Nanchang, China: [s. n.], 2016: 110-122.

- [3] CAI F, de RIJKE M. A Survey of Query Auto Completion in Information Retrieval [J]. Foundations and Trends in Information Retrieval, 2016, 10 (4): 273-363.
- [4] 宋巍. 基于主题的查询意图识别研究[D]. 哈尔滨: 哈尔滨工业大学, 2013.
- [5] DUMAIS S, CUTRELL E, CADIZ J J, et al. Stuff I've Seen: A System for Personal Information Retrieval and Re-use[J]. ACM SIGIR Forum, 2016, 49(2): 28-35.
- [6] 王国霞, 刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用, 2012, 48(7): 66-76.
- [7] 刘知远, 崔安顺. 大数据智能[M]. 北京: 电子工业出版社, 2016.
- [8] 石雁, 李朝锋. 基于协同相似计算的查询推荐[J]. 计算机工程, 2016, 42(8): 188-193.
- [9] 刘璟, 李鹏, 刘欣, 等. 基于用户聚类的推荐算法[J]. 计算机应用与软件, 2015, 32(10): 269-272.
- [10] 马慧芳, 贾美惠子, 李晓红, 等. 一种基于标签关联关系的微博推荐方法[J]. 计算机工程, 2016, 42(4): 197-201, 208.
- [11] 王宁宁, 鲁燃, 王智昊, 等. 基于用户标签的微博推荐算法[J]. 计算机应用研究, 2017, 34(1): 58-61.
- [12] 张骁逸, 苏宇, 晏小辉. 基于用户浏览日志的上下文相关新闻推荐[J]. 计算机工程与应用, 2016, 52(22): 99-104.
- [13] 王晓春, 李生, 杨沐昀, 等. 一种长短期兴趣结合的个性化检索模型[J]. 中文信息学报, 2016, 30(3): 172-177.
- [14] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
- [15] 易军凯, 刘慕凡, 万静. 基于主题与语义的作弊网页检测方法[J]. 计算机工程, 2015, 41(9): 311-316.
- [16] PORTEOUS I, NEWMAN D, IHLER A, et al. Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2008: 569-577.
- [17] ZHAI C, LAFFERTY J. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval[C]//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2001: 334-342.
- [18] HAN S, DU L, SALAZAR E, et al. Dynamic Rank Factor Model for Text Streams[C]//Proceedings of NIPS'14. Berlin, Germany: Springer, 2014: 2663-2671.
- [19] ZHAI C, LAFFERTY J. Model-based Feedback in the Language Modeling Approach to Information Retrieval[C]//Proceedings of the 10th International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2001: 403-410.
- [20] 李亚楠, 王斌, 李锦涛. 搜索引擎查询推荐技术综述[J]. 中文信息学报, 2010, 24(6): 75-85.
- [21] MORENO J G, DIAS G, CLEUZIOU G. Query Log Driven Web Search Results Clustering[C]//Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2014: 777-786.

编辑 金胡考