计算机研究与发展 ISSN 1000-1239/CN 11-1777/TP
Journal of Computer Research and Development 46(9): 1594-1600, 2009

基于改进 TextTiling 方法的用户新兴趣发现的研究

邹博伟 张 宇 范基礼 郑 伟 刘 挺
(哈尔滨工业大学信息检索研究室 哈尔滨 150001)
(bwzou@ir.hit.edu.cn)

Research on Personalized Information Retrieval Based on User's New Interest Detection

Zou Bowei, Zhang Yu, Fan Jili, Zheng Wei, and Liu Ting
(Information Retrieval Laboratory, Harbin Institute of Technology, Harbin 150001)

Abstract An important characteristic of next generation search engine is personalization. Personalized information retrieval (PIR) focuses on users. It captures users' interest in different kinds (explicit, implicit interest and interest of similar users). These information of users are integrated and used to improve the result of information retrieval system. Personalized information retrieval can grasp the users' retrieval intention and find personalized results. The authors propose the new interest detection task, which identifies the queries containing users' new retrieval interest by the change of retrieval object. Simultaneously, by using and improving the TextTiling algorithm, the retrieval system is enabled to automatically choose the appropriate dynamic threshold and detect the change of users' interest. The retrieval information and labeled answers of users are used to establish the experimental dataset. The evaluation matrix includes false alarm rate, miss alarm rate, and cost of detection. In the experiment of personalized information retrieval system, the improved TextTiling algorithm improves the new interest detection system by 16.4%. What's more, the new interest detection task improves the performance of the personalized information retrieval system is by 3.8%. The experiment shows that mining users' interest with this method can decrease the false information in users' models and improve the result of precision of users' interest detection.

Key words personalized information retrieval; users' new interest detection; TextTiling algorithm; dynamic threshold; search interest change

摘 要 个性化信息检索可以根据用户的检索兴趣返回个性化的检索结果.提出了用户新兴趣发现子任务,根据用户检索对象的变化识别包含新检索兴趣的查询.同时,引入 TextTiling 方法并对其进行改进,使系统可以自动选择合适的动态阈值并准确发现用户检索兴趣的转移.在构建的标准评测集上的实验结果表明,改进的 TextTiling 方法使得用户新兴趣发现系统性能提高了 16.4%,而且此子任务使得最终的个性化检索系统的性能提高了 3.8%.

关键词 个性化信息检索; 用户新兴趣发现; TextTiling 算法; 动态阈值; 检索兴趣转移

中图法分类号 TP391

收稿日期: 2009-01-06; 修回日期: 2009-05-08
基金项目: 国家自然科学基金项目(60736044, 60675034); 国家"八六三"高技术研究发展计划基金项目(2008AA01Z144)

现有的通用搜索引擎主要依赖语言学、符号匹配和网络链接关系反馈用户需求, 忽视了用户个体的背景信息和兴趣偏好, 从而使反馈的结果包含了大量与用户个性需求不相关的信息, 限制了用户获取相关信息的精确性和快捷性[1]. 因此, 如何提高搜索引擎检索结果的精度并向用户提供个性化服务已成为搜索引擎技术的一个新的发展方向和研究热点. 个性化检索系统提供了更好的方式来确定用户检索意图, 收集和分析用户的个性化信息和查询的上下文, 判断用户的真实需求, 并根据不同的用户返回不同的检索结果.

在检索过程中, 往往会出现用户连续输入的查询词(query) 从不同侧面描述同一检索目标的情况, 如果利用之前描述相同检索目标的 query 对当前的查询进行扩充, 有利于更准确地把握用户的检索意图, 从而利用用户查询过程中的"上下文信息" 优化个性化检索系统的性能[2]. 由此, 本文提出了用户新兴趣发现任务. 该任务将用户的兴趣按检索对象的不同划分为若干相对独立的兴趣表达段落. 在个性化检索中, 可以利用与当前 query 处于同一段落内的 query、以及用户检索的隐式信息(包括用户的点击历史, 浏览网页的时间等) 对用户模型进行扩充, 从而定位用户的检索需求. 比如, 用户查询"苹果" 之前, 在同一兴趣表达段落里查询过有关苹果牌电脑的信息, 则利用该用户在查询和浏览中的行为作为判断当前用户检索兴趣的依据, 从而更加准确地把握用户的检索意图, 给出更加符合用户兴趣的检索结果. 针对用户新兴趣发现任务, 本文还提出了改进的 TextTiling 方法, 该方法可以实现实时自动地对 query 段落进行划分, 不需要训练阈值, 划分结果也明显好于依赖阈值训练的相似度划分方法.

1 相关研究

个性化检索的重要研究任务是如何建立个性化的用户兴趣模型. 针对这一课题, 现有的相关研究主要分为两种: 1) 建立用户的个性化需求模型, 利用该模型对原检索结果进行重排序; 2) 根据检索历史对用户输入的查询内容进行扩充从而进行重新检索. 在第 1 类研究中, 林鸿飞等人提出了利用用户提供的文档[3], 将表示用户兴趣的段落作为识别用户兴趣的基本单位, 利用将段落聚类的方法获得用户兴趣类别, 建立用户兴趣模型, 其优点在于可以最直接地获取用户对当前检索结果的评价, 因此可以提供最准确的信息更新用户模型, 但此方法需要用户花费较多精力对结果进行评价, 所以在实际应用中的价值并不突出. Teevan 等人针对每个查询建立向量空间模型, 并通过将检索结果的标题和摘要与查询内容相融合的方式建立用户的兴趣模型[4]. 此方法考虑到用户在浏览检索结果时所点击网页与用户的检索目的有较大的关联性, 因此使用标题和网页摘要信息对用户兴趣模型进行更新, 可以更精确地描述用户兴趣. Liu 等人采用对 query 进行分类的方法构建用户兴趣模型[5], 此方法根据用户点击过的网页将 query 划分到若干类别中, 每个类别下有相关的网页文档, 从而计算用户检索关键词与表现用户兴趣的类别之间的相似度, 按相似度将类别进行降序排序, 最后选择前 k 个类别辅助搜索引擎为用户反馈结果. Qiu 等人提出了利用基于话题的 PageRank 算法进行个性化检索的方法[6]. 该方法事先将用户的兴趣划分为若干类(topic), 利用用户对结果的点击行为建立用户对各话题的兴趣矩阵, 最后利用此兴趣矩阵结合主题相关的 PageRank 公式计算每个结果网页的 Rank 值.

由于用户输入的 query 无法很好表达用户检索意图, 因此另一种研究方法尝试利用用户相关的查询历史对当前 query 内容进行扩充并对网页重新检索. Shen 等人[2] 提出一个假设: 当用户检索一种信息时, 如果第 1 遍搜索的结果中不包含相关反馈, 则用户会变换检索关键词重新搜索此信息, 因此用户连续输入的检索关键词可能是对同一信息的描述. 基于此假设, 先利用 query 内容及返回的检索结果的标题和摘要建立每个 query 的模型, 然后根据相邻 query 之间的相似度判断前一个 query 与当前 query 是否相关, 如果相关, 则利用前一个 query 的信息对当前 query 进行扩充, 从而把握用户的检索意图. 此方法优点在于实现比较简单, 不足是需要训练判断相似度的阈值, 从而对语料的依赖比较大.

2 基于用户新兴趣发现任务的个性化信息检索

2.1 个性化信息检索任务

个性化检索任务主要是挖掘用户的兴趣, 并利用用户兴趣对检索结果进行优化, 使得检索结果更加符合用户的个性化需求. 例如, 用户想观看一些关于计算机智能方面的电影, 利用搜索引擎进行检索, 如图 1 所示:

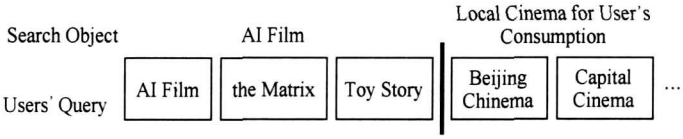

Fig. 1 Process of users' search.

图 1 用户检索过程

由于刚开始不知道具体的电影名, 用户首先输入检索关键词"计算机智能 电影", 然后, 根据从结果中获得的信息, 用户找到自己可能感兴趣的电影(例如"黑客帝国"), 再输入精确的电影名以获得更加详细的相关信息. 在这个过程中, 用户需要不断优化自己的 query, 从而最终搜索到自己想要的结果. 个性化检索任务是在用户查询的过程中, 系统通过用户的检索行为, 例如点击网页、浏览时间、跳转信息等用户隐式信息, 建立用户兴趣模型, 获得用户的兴趣, 从而对检索结果进行处理, 将更准确更多的结果返回给用户, 优化了用户的查询过程.

2.2 用户新兴趣发现任务

如果检索系统可以判断用户的检索对象, 并自动识别出用户新的检索兴趣的出现, 则可以找到检索对象相同的 query, 利用与当前 query 检索对象相同的其他 query 中用户的行为判断用户对哪些信息感兴趣, 哪些网页是用户真正需要的, 这样就可以使得对用户隐式信息的利用更加准确.

本文提出将用户新兴趣发现任务作为个性化检索任务中的一个独立子任务, 此子任务的目的是对用户的检索过程进行分析, 识别用户检索对象的变化, 发现包含用户新检索兴趣的 query, 将检索对象相同的 query 划分为同一段落. 用户新兴趣发现任务需要利用每个用户的 query 内容、系统返回的检索结果、用户查看过的结果网页、对网页的浏览时间、用户对结果的翻页信息等进行处理, 由系统生成 query 段落划分结果, 然后将这个结果与标准答案进行对比, 对系统的性能进行评价. 评价指标将采用话题检测与跟踪中的误检率和漏检率方法, 具体的评测方法和评测指标将在第 5 节介绍.

3 基于改进 TextTiling 方法的用户新兴趣发现

3.1 TextTiling 方法

TextTiling 方法主要应用于新闻报道中文章段落划分. Hearst[7] 的研究表明, 在一篇新闻报道中, 通常是相邻的若干句子表达同一个子主题, 这些表达同一子主题的句子中相邻两句的相似度都比较高, 而子主题有转折的两个相邻句子的相似度相对会有大幅度下降, TextTiling 方法将表达同一个子主题的句子划分为同一段落.

TextTiling 方法把每两个相邻句子的相似度作为一个点, 将每个点与前后两点的相似度下降值之和作为深度值(depth score), 如果出现当前点的相似度高于前面点或后面点的情况, 对应的相似度下降值取 0, 深度值计算公式为

$$depth_i = \max\{(sim_{i-1,i} - sim_{i,i+1}), 0\} + \max\{(sim_{i+1,i+2} - sim_{i,i+1}), 0\}, \quad (1)$$

其中, $sim_{i,i+1}$ 是句子 i 和句子 $i+1$ 的相似度, 相似度采用计算向量空间模型夹角的余弦值[7] 的方法获得. 动态阈值定义为

$$threshold_D = S - \sigma/2, \quad (2)$$

其中, S 为深度值的平均值, σ 为深度值的标准偏差.

如果深度值大于动态阈值, 则将认为这两个句子表达不同的子主题. 如图 2 所示, 每个点代表相邻两个句子的相似度, 点旁边的数值表示相邻两个句子的深度值. 句子 3, 4 的深度值为 0.9, 大于动态阈值 $threshold_D = 0.22$, TextTiling 将判定句 3 和句 4 表达不同子主题, 并将句 3 和句 4 之间作为不同子主题的边界进行切分. 在用户新兴趣发现任务中, 相邻的若干 query 描述同一个检索对象, 这些描述同一检索对象的 query 内容比较相似, 而用户变换检索对象时的 query 内容会有较大转折, 因此用户新兴趣任务中 query 段落的划分类似于新闻报道中句子的划分, 所以本文将 TextTiling 方法应用到用户

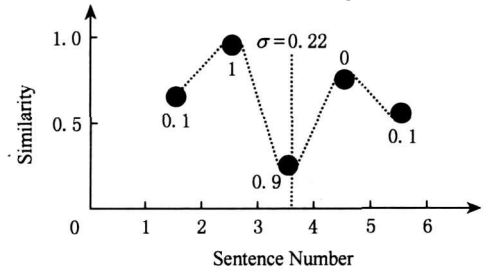

Fig. 2 TextTiling algorithm.

图 2 TextTiling 方法

新兴趣发现任务中. 但由于 TextTiling 方法无法将相似度较低的连续 query 划分开来, 而且在使用 TextTiling 方法时无法实现在线划分, 本文对 TextTiling 方法进行了改进.

3.2 改进的 TextTiling 方法

由于用户输入的 query 都比较短, 词的个数少, 为克服 query 数据稀疏而使相似度计算不准确的问题, 系统利用两个 query 检索时用户点击过的网页摘要计算相似度作为其相似度, 采用向量空间模型 (vector space model, VSM) 表示网页摘要, 基于 TF-IDF 权重公式描述每个网页摘要的词权重向量 $\boldsymbol{s}_i$, 然后在向量空间模型中将原始 query 向量与网页摘要的词权重向量融合, 计算质心向量:

$$\boldsymbol{x} = \alpha\boldsymbol{q} + (1-\alpha)\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{s}_i, \qquad (3)$$

其中, $\boldsymbol{q}$ 是 query 向量, k 表示在查询为 q 时用户点击网页的数目, 参数 α 用于分配 query 和用户点击过的网页对质心向量的影响, 在本文方法中, α 取 0.5. 每个质心向量表示用户的一个 query 模型, 文章段落在大多数情况下一个话题中包含许多句子, 而用户输入的一系列 query 通常会出现针对同一检索对象只使用一个 query 的情况, 这样就造成连续几个 query 之间的相似度都比较低的现象, TextTiling 方法无法将句子划分开来, 这种情况下 TextTiling 方法不能很好地识别 query 段落. 此外, 针对 query 流的切分与文章段落切分的区别在于前者是在线划分的, 即对于当前 query 来说只存在与前面 query 的相似度. 因此, 本文对 TextTiling 方法做了两点改进: 一是用相对坡度下降值代替传统方法中的绝对坡度下降值(式(4)), 这样可以有效地解决连续 query 之间的相似度低造成的段落无法正确划分的问题. 改进的相对深度计算公式如下:

$$depth_i = \max\{(sim_{i-1,i} - sim_{i,i+1})/sim_{i-1,i}, 0\} + \max\{(sim_{i+1,i+2} - sim_{i,i+1})/sim_{i+1,i+2}, 0\}. \qquad (4)$$

第 2 个改进是用前面 $i-1$ 个 query 模型相似度的平均值代替当前 query 与下一 query 的相似度. 当系统判断当前 query 是否为用户新兴趣时, 因为系统是在线的, 即每当新来一个 query 就要判断其是否与前一个 query 处于相同段落中, 没有后一个 query 的信息可以利用, 所以无法计算与下一个 query 的相似度, 因此从期望的角度来考虑, 当前 query 与下一 query 的相似度的数学期望值可以近似地用前 n 个相邻 query 相似度的平均值代替, query 数量越多这个值越接近期望值, 这也从一定程度上反映了用户的检索行为习惯. 使用改进的 TextTiling 方法对 query 进行分段的优点在于, 系统可以根据相邻 query 相似度的相对值自动对 query 进行段落划分, 而不需要设定一个静态的相似度阈值.

3.3 基于用户新兴趣发现的个性化检索

采用用户新兴趣发现任务中的方法判断当前查询的 query 与前一个 query 是否属于同一段落. 若属于同一段落, 则利用相同段落中的 query 信息对当前 query 进行扩充, 扩充公式如下:

$$\boldsymbol{q}_f = \beta\boldsymbol{q}_c + (1-\beta)\boldsymbol{q}_p, \qquad (5)$$

其中, $\boldsymbol{q}_f$, $\boldsymbol{q}_c$, $\boldsymbol{q}_p$ 分别表示扩充后的 query 模型、同一段落内其他 query 的模型、当前 query 的模型, β 是用于扩充的信息在扩充后的 query 模型中所占权重, 同一 query 段落描述同一个用户检索对象, 各 query 的信息同样重要地描述了该检索对象, 因此本文中 β 设置为 0.5.

利用扩充后的 query 模型与评测集中要求评价的网页计算相似度, 相似度计算采用向量空间模型中计算向量夹角余弦值的方法, 再通过系统阈值判断其中哪些网页与当前 query 相关, 最后将系统的结果与正确答案比较, 得到个性化检索系统的性能.

4 实验及结果分析

4.1 实验语料

针对个性化信息检索我们建立了标准评测集, 开发了基于天网 100G 语料的个性化评测语料标注辅助系统[8], 标注者利用此系统模拟正常的检索行为, 系统记录下用户在检索过程中的各种隐式信息, 并由标注者判断结果网页是否为检索目标的正确结果.

利用个性化检索标注辅助系统我们收集了 9 名标注者的标注结果. 其中每个人对 100 个检索问题进行检索和标注, 平均每个人进行了 230 次检索. 在用户兴趣发现任务的标准答案中, 本文将其中前 50% 的标注结果作为训练集, 后 50% 作为测试集.

4.2 评测方法

用户新兴趣发现任务的评价指标借鉴话题跟踪与检测(topic detection and tracking, TDT) 中的评价指标[9]. 因为在 TDT 评测中, P_{target} 描述了正确答案在语料总数中的比例, 可以更好地反映不同语料上的实验效果, 而且 DET 曲线和 $(C_{Det})_{Norm}$ 值能够更准确地描述取不同相似度阈值时系统性能的好坏.

基于TDT 2003 的评测方法[8], 通过误检率和漏检率对系统性能进行评测. 其计算公式如下:

$$P_{FA} = \frac{A}{A + C},\ P_{Miss} = \frac{B}{B + D}, \qquad (6)$$

其中 A, B, C, D 如表 1 所示, A 为系统判定是用户新兴趣的 query 且在标准答案也是新兴趣 query 的个数, B, C, D 同理. P_{FA}, P_{Miss} 是系统误检率和漏检率, 值越小则系统性能越好.

Table1 Parameters of evaluation

表 1 评测的参数

Item	Relevance by System	No-Relevance by System
Relevance by Answer	A	B
No-Relevance by Answer	C	D

之后, 通过误检率和漏检率计算总的评价指标 $(C_{Det})_{Norm}$, 公式如下:

$$(C_{Det})_{Norm} = (C_{Miss}P_{Miss}P_{target} + C_{FA}P_{FA}P_{non\text{-}target})/\min(C_{Miss}P_{target}, C_{FA}P_{non\text{-}target}), \qquad (7)$$

其中, C_{Miss} 是系统进行一次漏检的代价; C_{FA} 是系统进行一次误检的代价, 由于实际中, 错误的段落划分和漏掉正确段落划分对个性化检索任务的影响基本等价, 因此将 C_{Miss} 和 C_{FA} 都设为 1; P_{target} 是每个 query 为用户新兴趣的概率; $P_{non\text{-}targe}$ 是非新兴趣的概率, 针对语料中的正确答案, 将 P_{target} 和 $P_{non\text{-}target}$ 分别设为 0.435 与 0.565; $(C_{Det})_{Norm}$ 是系统性能损耗代价, 此值越小则系统性能越好.

为了使系统性能得到更直观的体现, 本文引入 TDT 的中的决策错误权衡曲线(decision error tradeoff curve, DET 曲线)评测系统性能. 横坐标是误检率, 纵坐标是漏检率, 曲线越靠近图的左下角则性能越好, 在图中还标出了最小性能损耗代价, 此值越小则系统综合性能越好. 个性化检索任务的评测采用相同的方法. 个性化信息检索研究通常采用基于人工参与的传统评价方式, 对评价者要求较高, 不同的评价者可能会产生不同的评价结果, 同时增加了评价成本; 而本文采用标准评价方式, 减少了评价中的人工参与, 方便跨系统评价, 使评价具有更强的客观性.

4.3 实验设计

针对用户新兴趣发现任务, 本文采用 Shen 等人[2]提出的基于向量空间模型, 通过计算相似度对 query 进行段落划分的方法构建用户新兴趣发现任务的 baseline 系统(System1); 采用改进的 TextTiling 方法实现自动地对描述相同检索兴趣的 query 进行段落划分, 形成用户新兴趣发现系统(System2), 评价改进的 TextTiling 方法对用户新兴趣发现任务的影响.

针对个性化检索任务, 本文采用 Shen 等人提出的个性化检索方法作为 baseline 系统(System3), 该系统只利用了与当前 query 相关的前一个 query 信息进行扩充; 采用基于改进的 TextTiling 方法的用户新兴趣发现任务对当前 query 进行扩充, 形成个性化检索系统(System4), 评价用户新兴趣发现任务对个性化检索系统性能的影响.

4.4 实验结果及分析

在训练集上, 考察了 TextTiling 方法与改进 TextTiling 方法在用户新兴趣发现任务中的效果, 基于 TextTiling 方法的 $(C_{Det})_{Norm}$ 值为 0.241549, 而经过改进的 TextTiling 方法的 $(C_{Det})_{Norm}$ 值达到了 0.197033, 性能提高了 18.4%.

图 3 是在训练集上考察改进的 TextTiling 方法对用户新兴趣发现任务的影响. 对于 System1, 需要在训练集上训练判断相邻两个 query 是否属于同一段落的相关性阈值 θ. 对应图中使训练集上 $(C_{Det})_{Norm}$ 能够取得最小值的 θ 值为 0.103. 其中的曲线是 baseline 系统的性能, 曲线上黑色的点是 baseline 方法在训练集上的最小 $(C_{Det})_{Norm}$ 值. 曲线下方的"+"是 System2 在训练集上的 $(C_{Det})_{Norm}$ 值, 由于 TextTiling 方法不需要训练参数, 所以只得到一个点, 这个点就是系统的最优效果. 从图中可以看出, System1 要好于 System2, $(C_{Det})_{Norm}$ 由 0.210484 降到 0.197033, 系统性能提高了 6.4%. 图 4 是在测试集上考察改进的 TextTiling 方法对用户新兴趣发现任务的影响. 其中, System1 的 $(C_{Det})_{Norm}$ 为

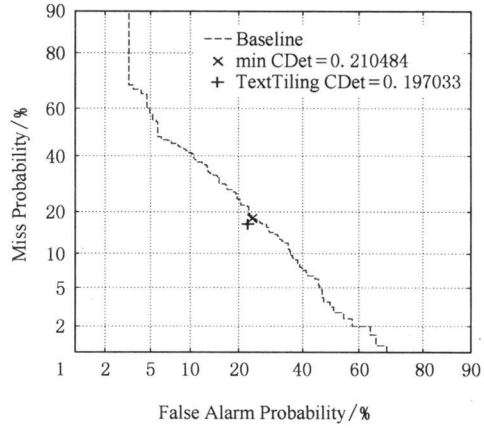

Fig. 3 Effect of TextTiling algorithm to New Interest Detection task on training set.

图 3 训练集上 TextTiling 方法对用户新兴趣发现任务的影响

0.17257, System2 的 $(C_{Det})_{Norm}$ 为 0.14419, 性能提高了 16.4%. 在训练集上和测试集上相比, 改进的 TextTiling 方法和 baseline 方法对用户新兴趣发现任务效果提高有比较明显的差别, 主要原因是在训练集上(如图 3 所示), baseline 方法可以对判断相邻 query 之间相似度的阈值 θ 进行充分的训练, 而测试集上(如图 4 所示)无法对阈值 θ 进行训练. 而改进的 TextTiling 方法利用相似度之差, 即下降坡度来描述用户兴趣的变化, 可以更好地体现用户检索兴趣的变化, 使划分 query 段落达到较好的效果, 并且改进的 TextTiling 方法不需训练相似度阈值, 不同语料上的可移植性好.

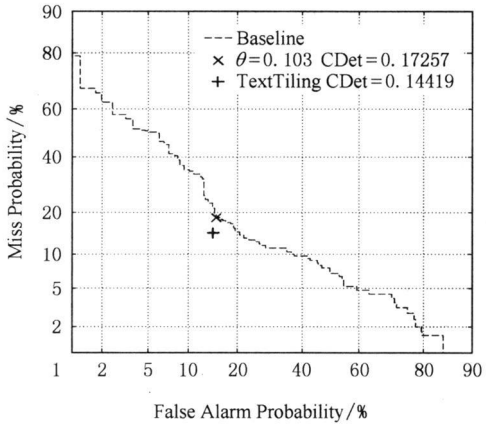

Fig. 4 Effect of textTiling algorithm on new interest detection task in testing set.

图 4 测试集上 TextTiling 方法对用户新兴趣发现任务的影响

图 5 给出了测试集上分别基于 baseline 方法和改进的 TextTiling 方法的用户新兴趣发现任务, 应用于个性化信息检索系统后获得的 DET 性能曲线图. System4 的效果好于 System3, 将系统性能提高了 3.8%. 在个性化检索中, 采用基于改进 TextTiling 方法的用户新兴趣发现任务, 利用整个兴趣段落中的 query 信息对当前 query 进行扩充, 克服了用于扩充的 query 包含信息较少的缺点, 并提高了对 query 扩充内容的准确性, 使得用于扩充的 query 能够更准确地描述用户的检索意图, 从而使个性化检索效果得到提高.

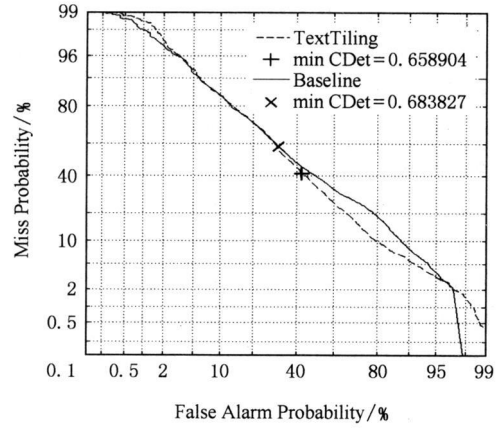

Fig. 5 DET curve of personalized information retrieval system on testing set.

图 5 测试集上个性化检索系统的 DET 曲线

5 结论与未来工作

针对现有工作对挖掘检索兴趣准确性较低的问题, 本文提出了用户新兴趣发现任务, 该任务是将用户已输入过的 query 按照检索对象的不同进行段落划分, 以便获得更准确的用户检索兴趣, 此任务使得最终的个性化检索系统的性能提高了 3.8%. 针对用户新兴趣发现任务, 本文还提出了采用改进的 TextTiling 方法实现用户新兴趣发现任务, 此方法较为准确地刻画了用户兴趣变化, 而且不需要训练参数, 其结果比 baseline 方法的效果提高了 16.4%. 用户新兴趣发现任务增强了对当前 query 检索意图的把握, 但同时也会引入少量噪声, 影响个性化检索系统的性能. 因此在今后的研究工作中, 还需要对在用户新兴趣发现的基础上增加对用户兴趣的判断, 即剔除掉不属于用户兴趣的信息, 进一步提高个性化检索的性能.

参 考 文 献

[1] Li Xiaoming, Yan Hongfei, Wang Jimin. Theory, Technology and System of Search Engine [M]. Beijing: Science Press. 2005: 212-215 (in Chinese)
(李晓明, 闫宏飞, 王继民. 搜索引擎原理、技术与系统[M]. 北京: 科学出版社. 2005: 212-215)

[2] Shen Xuehua, Tan Bin, Zhai Chengxiang. Implicit user modeling for personalized search [C] //Proc of CIKM 2005. New York: ACM, 2005: 824-831

[3] Lin Hongfei, Yang Yuansheng. The representation and update mechanism for user profile [J]. Journal of Computer Research and Development, 2002, 39(7): 843-847 (in Chinese)
(林鸿飞, 杨元生. 用户兴趣模型的表示和更新机制[J]. 计算机研究与发展, 2002, 39(7): 843-847)

[4] Teevan J, Dumais S T, Horvitz E. Personalizing search via automated analysis of interests and activities [C] //Proc of SIGIR 2005. New York: ACM, 2005: 449-456

[5] Liu Fang, Yu Clement, Meng Weiyi. Personalized Web search by mapping user queries to categories [C] //Proc of CIKM 2002. New York: ACM, 2002: 558-565

[6] Qiu Feng, Cho Junghoo. Automatic identification of user interest for personalized search [C] // Proc of WWW 2006. New York: ACM, 2006: 727-736
[7] Hearst M A. Multi-paragraph segmentation of expository text [C] // Proc of ACL 1994. Morristown, NJ, USA: ACL, 1994: 9-16
[8] IR-Lab, HIT. Personalized Search Engine [CP/OL]. [2008-06-20]. http://ir.hit.edu.cn/demo/bwzou/index.jsp
[9] NIST. Topic Detection and Tracking Evaluation [EB/OL]. [2008-06-20]. http://www.itl.nist.gov/iad/mig/tests/tdt/

Zou Bowei, born in 1984. Master candidate in the School of Computer Science and Technology. Harbin Institute of Technology. His research direction: information retrieval.
邹博伟, 1984 年生, 硕士研究生, 主要研究方向为信息检索.

Zhang Yu, born in 1972. Senior member of China Computer Federation. Associate professor and PhD in the School of Computer Science and Technology, Harbin Institute of Technology. His research direction: natural language processing and information retrieval.
张 宇, 1972 年生, 博士, 副教授, 中国计算机学会高级会员, 主要研究方向为自然语言处理、信息检索(zhangyu@ir.hit.edu.cn).

Fan Jili, born in 1986. Graduated from the School of Computer Science and Technology, Harbin Institute of Technology. His research direction: information retrieval.
范基礼, 1986 年生, 本科, 主要研究方向为信息检索(jlfan@ir.hit.edu.cn).

Zheng Wei, born in 1984. Master in the School of Computer Science and Technology, Harbin Institute of Technology. His main research direction: information retrieval.
郑 伟, 1984 年生, 硕士, 主要研究方向为信息检索(zw@ir.hit.edu.cn).

Liu Ting, born in 1972. Senior member of China Computer Federation. Professor and PhD supervisor in the School of Computer Science and Technology, Harbin Institute of Technology. His research direction: natural language processing and information retrieval.
刘 挺, 1972 年生, 教授, 博士生导师, 中国计算机学会高级会员, 主要研究方向为自然语言处理、信息检索(tliu@ir.hit.edu.cn).

Research Background

This paper is supported by the National Natural Science Foundation of China (grant No. 60736044). The name of fund is "the Next Generation Search Engine-Personalized information retrieval". An important characteristic of next generation search engine is personalization. Personalized Information Retrieval (PIR) focuses on users. It captures users' interest in different kinds (explicit, implicit interest and interest of similar users). These information of users are integrated and used to improve the result of information retrieval system. Personalized information retrieval can grasp the users' retrieval intention and find personalized results. This paper proposes the new interest detection task, which identifies the queries containing users' new retrieval interest by the change of retrieval object. Simultaneously, this paper uses and improves the TextTiling algorithm to automatically choose the appropriate dynamic threshold and detect the change of users' interest. The retrieval information and labeled answers of users are used to establish the experimental dataset. The evaluation matrix includes false alarm rate, miss alarm rate and cost of detection. In the experiment of personalized information retrieval system, the improved TextTiling algorithm improves the new interest detection system by 16.4%. What's more, the new interest detection task improves the performance of the personalized information retrieval system by 3.8%. Experiment shows that mining users' interest with this method can decrease the false information in users' models and improve the result of precision of users' interest detection.