

面向探索式搜索过程的查询推荐

马超, 张引, 张斌

(东北大学 信息科学与工程学院, 辽宁 沈阳 110819)

摘 要: 查询推荐是一种提高用户搜索效率的重要工具,但是传统的推荐方法对于探索式搜索的推荐效果不好. 针对此问题提出了一种新的面向探索式搜索过程的查询推荐方法,即根据用户搜索的行为模式,通过试探性查询重构和确认性查询重构两个过程,对探索式搜索过程进行建模,并根据影响探索式搜索过程的三种因素提出了一种排序算法,将确认性子查询中的查询推荐给用户. 通过与传统推荐方法的对比实验验证了本模型及其推荐方法的有效性.

关 键 词: 查询推荐;探索式搜索;查询链;试探性查询重构;确认性查询重构

中图分类号: TP 311.13

文献标志码: A

文章编号: 1005-3026(2015)06-0777-04

Query Recommendation for Exploratory Search Process

MA Chao, ZHANG Yin, ZHANG Bin

(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: MA Chao, E-mail: macmacmac@yeah.net)

Abstract: Query recommendation is an important tool for improving searching efficiency. But traditional recommendation method cannot provide good recommendations for exploratory search. A novel model based on search behaviors for the above problem was proposed, which successfully build the exploratory search process by tentative query refinement and deterministic query refinement. Then a query recommendation algorithm based on the three factors affecting exploratory search process was proposed. Finally the results of comparative experiments showed that this model and algorithm achieved a good effect.

Key words: query recommendation; exploratory search; query chain; tentative query refinement; deterministic query refinement

随着搜索引擎的日益普及,搜索领域的主要研究方向已经从简单的查询-应答(query-response)模式中的信息查找(look-up)转到更为复杂的信息探寻(seeking)^[1]. 从2006年开始一种新的搜索模式“探索式搜索”被提出^[2]. 该模式从用户为中心的视角出发,认为用户的搜索方向和期望的结果会随着搜索过程的变化而不断变化. 学习和调查是该模式中最为重要的行为特征^[3-5]. 随着这一概念的提出,探索式搜索逐渐成为信息检索、信息科学、人机交互以及认知科学等领域的研究热点.

目前探索式搜索面临很多挑战,其中面向探索式搜索的查询推荐就是一个亟待解决的问题.

现有的查询推荐方法对下面具有明显探索性的搜索任务其推荐效果不明显,无法帮助用户缩短探索过程:

1) 当用户对于搜索目标的相关背景知识了解不足; 2) 有多种搜索目标可供选择时,用户提交了一个含义非常宽泛的查询.

面对这种需求时,目前的推荐方法都是基于以往搜索结果的文本相关性进行推荐的,因此当用户提交了一个含义宽泛的查询之后,其得到的推荐结果往往也是一个很宽泛的查询,对整个探索过程缺乏指导性. 典型的例子:一名中国大学生要参加一位法国留学生举办的生日宴会,想选择合适的生日礼物. 当用户输入“生日礼物”时,目

收稿日期: 2014-04-24

基金项目: 国家自然科学基金资助项目(61100090); 中央高校基本科研业务费专项资金资助项目(N110204006).

作者简介: 马超(1978-),男,辽宁抚顺人,东北大学博士研究生; 张斌(1964-),男,辽宁沈阳人,东北大学教授,博士生导师.

前的推荐方法给出的推荐结果是“生日礼物送什么好”、“创意生日礼物”、“女生生日礼物”、“男生生日礼物”等含义很宽泛的查询建议,无法帮助用户确定下一步的探索方向。

针对上述问题,本文提出了一种利用概率方法通过查询链图对探索式搜索过程进行行为建模,并利用该模型提出一种 rank 算法将确认性子查询中的查询推荐给当前用户,并对本方法与基准方法进行了对比试验。

1 探索式搜索过程的分析和定义

1.1 探索式搜索过程的分析

本文组织多位志愿者针对前面提到的例子展开实际的搜索,并收集全部搜索日志进行搜索行为分析,通过分析发现每次的搜索行为都具有以下特征(图 1):

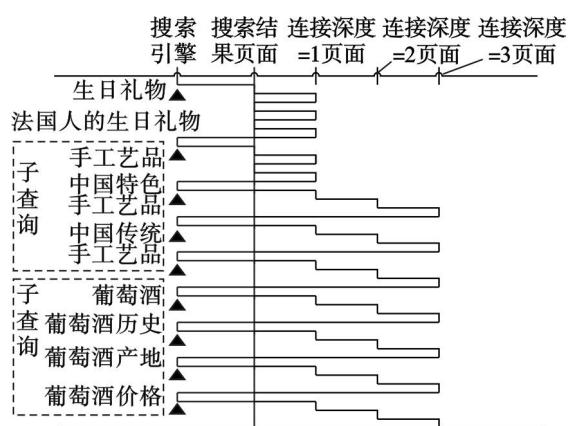


图 1 探索式搜索过程的行为特征

Fig. 1 The behavior characteristics of exploratory search

从图中可以看出探索式搜索过程主要由 3 种行为构成:

1) 快速浏览. 提交一个宽泛的查询词, 然后从搜索结果中, 点击一些链接深度为 1 的页面进行短时间浏览。

2) 深入浏览. 提交一个相对具体的查询词, 从链接深度为 1 的页面开始通过页面间的跳转, 逐步地对一些链接深度大于 1 的页面进行长时间的阅读, 通常这种浏览行为的链接深度在 1 ~ 3 之间。

3) 集中子查询. 通过对基本查询词的增删改提交一组查询目标基本相同的查询, 然后进行一些链接深度大于 1 的深入浏览。

1.2 探索式搜索过程的定义

根据探索式搜索的行为特征, 探索式搜索过

程可以明确地定义为: 如果在一个搜索过程中存在一个连续的试探性查询重构过程和确认性查询重构过程, 并且重构发生之前的查询对应的浏览行为依次为快速浏览、深入浏览, 那么该搜索过程可以称为探索式搜索过程。

2 探索式搜索过程的构建

2.1 构建查询链

以单个用户日志为基本分析单元, 依据查询提交时间的先后顺序, 组成一个查询链, 如图 2 所示。



图 2 查询链图

Fig. 2 The query chain

2.2 子查询划分

断开查询链中所有节点间的链接, 计算相邻节点间的查询相似度, 判断节点间是否发生确认性查询重构, 如果发生, 那么节点间重新建立一个链接(图 3)。

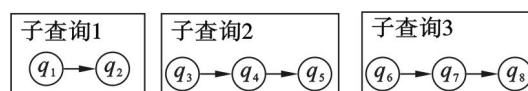


图 3 子查询链图

Fig. 3 The subquery chain

查询相似度 $\text{Sim}(q', q)$ 可以通过一种编辑距离的方法来计算^[6]。

2.3 浏览方式的判定

计算平均链接深度和平均浏览时间, 从而判定每个子查询对应的浏览行为是快速浏览还是深入浏览(图 4)。

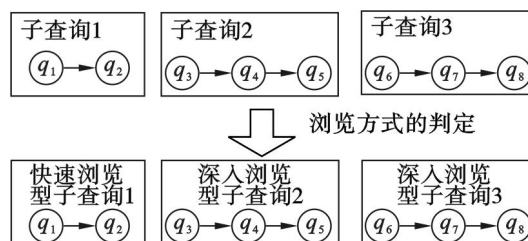


图 4 浏览方式判定图

Fig. 4 The identification for browse type

对于平均链接深度 $\text{Lnd}(q)$, 可以通过用户点击文档的数量来计算。对于平均浏览时间 $\text{Ret}(D)$, 可以利用文档点击时间差来计算。

2.4 探索式搜索过程构建

快速浏览型子查询与深入浏览型子查询进行

交叉组合,通过主题概率模型判定子查询间是否发生试探性查询重构(图5)。

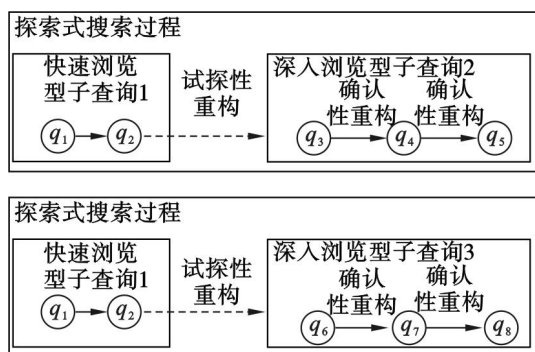


图5 探索式搜索过程构建

Fig. 5 The building process of exploratory search

对于查询词 q' 的出现概率的计算,本文采用一种主题预测模型来进行计算。具体的形式化描述如下:

对于一个 session 内用户所有点击文档进行 LDA 主题分析,得到主题集合 $T = \{t_1, t_2, \dots, t_n\}$,假定 $c = \{q_1, q_2, \dots, q\}$ 是用户提交 q' 之前提交的查询序列,那么对于重构后查询节点 q' 的出现概率 $P(q' | c)$ 可以表示为

$$P(q' | c) = \sum_{t_i \in T} \sum_{d_j \in D_c} P(q' | c; t_i, d_j) \times P(t_i, d_j | c). \quad (1)$$

式中的 $P(q' | c; t_i, d_j)$ 可以通过 Jelinek - Mercer 平滑模型^[7]进行估计。

3 探索式搜索的查询推荐方法

本文设计的推荐方法可分为两个部分:1) 线下部分。利用前面建立的模型挖掘出用户搜索日志中的探索式搜索过程。2) 在线部分。当用户提交一个查询时,利用那些包含该查询的探索式搜索过程构成一个被推荐集合,然后利用 rank 函数对被选集合进行排序并从中选出 top - k 个确认性子查询作为推荐结果。

对于 rank 函数的设计,主要依据影响推荐效果的几个因素。通过对建模过程的分析,发现影响推荐效果主要因素为

- 1) 发生试探性查询重构的链接权重(概率) w ,权重越大,推荐结果与搜索目标的相关性越强。
- 2) 确认性子查询间的平均相似度 s ,相似度越大探索的价值越大。
- 3) 探索深度 d ,即从用户当前所提交的查询节点到发生试探性查询重构节点间的链接个数,

深度越大,离搜索目标越远。

根据上述三种因素,本文设计的 rank 函数为

$$\text{Score}(q) = \beta \times \frac{w}{d+1} + (1 - \beta) \times s. \quad (2)$$

其中 β 为调节参数。

4 实验结果和讨论

4.1 实验数据

本文的实验数据来源于 100 位在校大学生及研究生在近 6 个月的 Firefox 的浏览历史记录,在实验数据收集期间,根据文献[8-9]设计了 10 个探索式搜索任务。

4.2 评价推荐效果

本文将文献[10]中提到的基于 session 的聚类方法作为基准方法,根据不同查询条件将通过基准方法得到的查询推荐结果与本文设计的面向探索式搜索过程的推荐结果进行对比。

表 1 列出了以“法国人的生日礼物”为查询条件,通过基准方法和本文方法得到的推荐结果。

表 1 “法国人的生日礼物”的推荐结果对比
Table 1 The comparative results for “the French birthday gift” recommendation

基准方法	本文方法
“给妈妈的生日礼物”	“手工艺品”
“特别的生日礼物”	“中国特色手工艺品”
“给爸爸的生日礼物”	“中国传统手工艺品”
“送给男生的生日礼物”	“葡萄酒”
“给女朋友的生日礼物”	“葡萄酒的产地”
“送给闺蜜的生日礼物”	“葡萄酒的价格”

通过表 1 可以发现,当查询词的查询目标比较宽泛时,本文提出的推荐方法相对于基准方法,对用户来说更具有启发性。

此外,本文还组织了 10 位志愿者对推荐结果的相关性进行人工打分。利用 LDA 模型对查询结果进行主题提取,可以得到每个用户查询对应的主题个数。随机选取 100 个用户查询,然后按照其对应的主题个数进行排序(降序),最后选出 top - 1 至 top - 5 五个查询,并对它们的推荐结果进行打分。

从相关性得分对比(图 6)可以得到与表 1 相同的结论,即当查询词的查询目标比较宽泛时(对应的主题数比较多),本文提出的推荐方法相对于基准方法更具有启发性。

(下转第 785 页)