高技术通讯 2010 年 第 20 卷 第 6 期:596 ~ 601

doi:10.3772/j.issn.1002-0470.2010.06.008

基于线性回归的相关查询推荐①

翟海军②* ** 张 刚** 张 瑾**

(* 中国科学技术大学计算机科学与技术系 合肥 230027)

(** 中国科学院计算技术研究所信息智能与信息安全研究中心 北京 100080)

摘 要 在分析搜索引擎查询日志的基础上,提出了一种基于线性回归的相关查询推荐方法。该方法考虑了查询串之间的多种关联关系,包括查询串会话共现、点击页面共享和查询串文本相似性,以避免因直接应用查询串之间的部分关联关系造成易受查询日志中噪音的影响。相比以往的方法,采用线性回归的方法来识别相关查询推荐的有效特征,能更好地解决噪音问题和进行有效的相关查询推荐。实验证实,采用线性回归挖掘的识别中文相关查询串的特征非常有效,且所提出的回归模型的预测准确率优于现有的方法。

关键词 查询日志,查询会话,相关查询推荐,线性回归

0 引 言

已有的 Web 搜索技术在一定程度上能够满足用户的信息需求[1]。但用户的查询请求通常都很简短,不能准确地描述用户的信息需求[2]。仅使用用户提交的初始查询,返回的相关文档通常很难满足用户的信息需求。如何根据用户的初始查询更有效地满足用户信息需求,成为信息检索研究的难点和热点[3]。现有的解决方案有查询扩展[4]、相关反馈[3]、相关查询推荐等技术[5,6]。

构造一个好的查询并非易事,特别是对搜索引擎不熟悉的用户。很多用户在搜索过程中会通过不断优化其查询串来满足其信息需求。本文通过分析搜索引擎查询日志进行相关查询推荐。当有相同信息需求的用户提交类似的查询串时,推荐以往用户优化的相关查询串给当前用户,以提高 Web 搜索的效率和用户体验。尽管该方法相对简单,但是这种简单的交互被证明在 Web 检索中非常有效。然而,查询日志中存在大量的噪音,如点击数据稀疏、查询会话切分错误和点击错误等。点击错误是指用户点击的结果链接(uniform resource locator, URL)与所提交的查询串不相关。本文通过线性回归(linear regression,LR)的方法来挖掘相关查询推荐的有效特征。所有的特征基于查询串之间的关联关系抽取得到。这些关联关系包括查询串会话共现、点击页面共享和查询串文本相似性。查询串会话共现是指两个查询串在同一会话中出现。点击页面共享是指两个查询串有一个或多个相同的点击结果 URL。以往的研究直接应用查询串之间的部分关联关系,这样很容易受上述噪音的影响。相比以往的方法,本文采用线性回归的方法来识别相关查询推荐的有效特征,这样可以更好地解决噪音问题和进行有效的相关查询推荐。最后的实验表明,采用线性回归挖掘的识别中文相关查询串的特征非常有效,并且本文所提出的回归模型的预测准确率优于现有的方法。

相对于英文相关查询推荐的广泛研究,当前中文相关查询推荐研究工作相对较少,尤其是缺乏中文相关查询推荐有效特性的研究。本文采用中文搜索日志作为语料,并采用线性回归来挖掘识别中文相关查询串的有效特征。

对相关查询推荐技术的性能评价,目前还没有一种客观、一致的评价方法。此前相关查询推荐技术的评价,都是基于人工评价或在线点击率。这些评价方法的客观性和合理性值得怀疑,另外,人工评价需要大量的人力,代价太大。评价方法的研究仍需进一步开展。本文基于文献[7]的思想,给出了预测实验,以此来评价相关查询推荐技术的性能。该方法更加客观,测试集容易获取而且无需大量的人力。

① 863 计划(2006AA010105,2007AA01Z416)资助项目。

② 男,1982 年生,博士生;研究方向:信息检索,数据挖掘;联系人,E-mail: zhaihaijun@software.ict.ac.cn

(收稿日期:2008-06-17)

1 相关工作

查询扩展[4]是指在用户所给查询用词的基础上,按照一定的扩展策略构建与该查询用词相关的词表,从而在检索时能够返回更多的相关文档。查询扩展的目的是为了提升检索结果的召回率,使用户需要的内容尽可能地包含在返回结果中。查询扩展可能会导致扩展后的查询主题漂移。

相关反馈技术[3]是指根据用户对检索结果的反馈来调整查询处理结果,从而提高检索的性能。但通常额外的操作影响用户体验[3],明确的相关反馈很难得到。伪相关反馈技术[3]作为相关反馈技术的变种,它假定返回结果列表的前 n 个文档为相关文档。伪相关反馈技术结果的好坏依赖于初始查询返回的前 n 个文档的相关性。同时,Web 检索中,隐式相关反馈成为一种重要的技术[8],它通过分析搜索引擎查询日志来获取相关性信息。然而由于查询日志中存在大量噪音,很难得到准确的相关性信息。

相关查询推荐[5,6,9,10]通过分析以往用户提交的查询串与当前查询串之间的关系,推荐最相关的候选查询串给当前用户,以提高 Web 搜索的效率。文献[5]通过分析查询日志中的点击数据,根据查询串与用户点击结果 URL 之间的关联关系,采用迭代聚类的方法挖掘相关查询串。用户查询日志中存在大量的噪音,并且点击数据非常稀疏。对于聚类算法,如何更好地解决噪音和点击数据稀疏问题仍需进一步研究。文献[6]基于查询日志中的查询串会话共现关系来挖掘查询词的关联查询词。对用户提交的查询串,通过选取会话共现查询串和替换部分关联查询词来获取候选推荐查询串集合。然后采用学习算法,排序候选查询串集合。文献[9]中进一步考虑了点击相同页面的查询串之间的关联,扩展了候选关联查询串范围。文献[10]给出了线性模型来排序候选查询串。

本文采用线性回归来挖掘相关查询串。文中明确定义了相关性类别,以此来消除训练样本标注可能带来的噪音。同时,通过组合不同特征来挖掘识别中文相关查询串的有效特征,这些特征有别于以往的研究,最后的性能稳定性实验和预测实验都证明了这些特征的有效性。最后的预测实验说明我们从查询会话中生成训练样本,能更有效地学习用户的查询行为。

2 基于线性回归的相关查询推荐方法

2.1 基本概念

首先定义文中用到的基本概念,如查询记录、查询事务、查询会话、候选推荐查询串集合和查询串对。查询记录是指搜索引擎查询日志所记录的用户单个查询行为。查询事务是指用户提交查询串到搜索引擎,并点击结果列表的一个或多个 URL。查询会话是用户为某信息需求而进行的多个查询事务。形式化定义如下:

定义 1 查询记录 R 对应查询日志中的一个条目:$(t, userID, q, click, url)$。其中,$t$ 是查询行为的时间标签,$userID$ 标识进行查询行为的用户,q 为用户提交的查询串,$click$ 标识查询事务中用户的点击序号,url 为点击的链接。

定义 2 查询事务 T 是一组查询记录的序列集合:$\{(t_1, userID_1, q_1, click_1, url_1), \cdots, (t_n, userID_n, q_n, click_n, url_n)\}$,对所有的正整数 i,有 $t_i < t_{i+1}$,$click_i = i$,$userID_1 = \cdots = userID_n$,$q_1 = \cdots = q_n$。查询事务 T 的用户标签记为 $userID(T)$,开始时间记为 $Start(T)$,结束时间记为 $End(T)$,查询串记为 $q(T)$。

定义 3 查询会话 S 是一组查询事务的序列:$\{T_1, T_2, \cdots, T_n\}$,满足以下条件:$userID(T_1) = userID(T_2) = \cdots = userID(T_n)$,$Start(T_{i+1}) - End(T_i) <$ 时间阈值,对所有的正整数 i。并且不能加入其他的查询事务,得到更大的查询会话。

查询会话的开始时间和结束时间,分别为查询事务 T_1 的开始时间 $Start(T_1)$ 和查询事务 T_n 的结束时间 $End(T_n)$。查询会话切分就是确定这两个时间。

定义 4 查询串 q 的候选推荐查询串集合 $C(q)$ 是一组查询串集合:任意 $q_c \in C(q)$,存在会话 S,$T_i \in S$,$T_j \in S$,$q_c = q(T_i)$,$q = q(T_j)$ 或存在记录 $R_i = (t_i, userID_i, q_i, url_i)$,$R_j = (t_j, userID_j, q_j, url_j)$,其中:$q_i = q$,$q_j = q_c$,$url_i = url_j$。

定义 5 查询串对 $P = (q_I, q_T)$,其中 $q_T \in C(q_I)$;q_I 称为初始查询串,q_T 称为目标查询串。

2.2 流程简介

首先介绍相关查询推荐方法的基本操作流程。查询日志记录了所有提交到搜索引擎的查询事务,我们先将其切分成查询会话。然后基于查询会话

集,进行查询串关联关系分析,获取查询串对的特征。基于这些特征,采用线性回归的方法训练排序模型。对用户提交的查询串,先根据其与以往查询串的关联关系,取得候选推荐查询串集合,然后用训练好的排序模型,排序候选推荐查询串集合得到相关查询串列表,返回结果列表给当前用户。

2.3 查询会话切分

在进行查询会话切分前,先介绍我们采用的查询日志数据。查询日志是用户向搜索引擎提交的查询以及相关数据的记录。比如查询时间、查询内容、点击的链接等。这些数据往往能够反映出用户的兴趣、查询用词的特点、查询内容与用户点击结果链接之间的关系等。本文采用搜狗提供的2007年3月1日至2007年3月7日查询日志作为实验语料,该语料的总记录数为10041593,用户数为2201847,不同的查询串数为1308500。

查询会话切分就是要确定查询会话的开始时间和结束时间。这里采用文献[2]提出的时间阈值方法来进行查询会话切分。尽管该方法存在少量的切分错误,但并不会影响我们的算法的性能,最后的实验证明了这一点。图1给出了含多个查询事务的查询会话数目随切分时间阈值变化的曲线。从图1可以看出,在时间阈值为200s时,曲线逐渐平稳。因此设定切分时间阈值为200s,这个值小于文献[2]中的300s。这与查询日志的记录方式有关,这里查询日志时间标签更加精细。采用200s作为时间阈值进行查询会话切分,结果是:总查询会话数为353965,包含多个查询事务的查询会话数为968380,占总会话数的27.36%。该结果显示含多个查询事务的查询会话的比率与文献[2]中给出的非常相近。

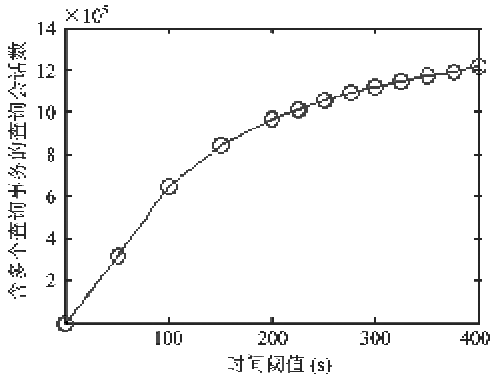

图1 含多个查询事务的查询会话数目随切分时间阈值变化的曲线

得到切分好的查询会话集合后,选择训练样本,进行查询会话分析和特征抽取。

2.4 查询会话分析

在进行查询会话分析和特征抽取前,先选择训练样本,并进行样本标注。表1给出了查询串之间相关性类别标签。基于文献[11]的思想,查询串由多个话题部分组成。对查询串对 (q_I, q_T) 考虑了3种相关性类别,它们各自的意义如下:

表1 训练集合分布表

	相关性描述	打分	样本数	比率
相关性类别1	很相关	3	426	47.33%
相关性类别2	相关	2	157	17.45%
相关性类别3	不相关	1	317	35.22%
总样本数	–	–	900	100%

- 相关性类别1:q_T 与 q_I 具有相同的主话题,如查询串对:(老友记,老友记在线);(幼儿教育,幼儿识字);(北京户口最新政策,北京市户口)。
- 相关性类别2:q_T 与 q_I 话题上相互关联。如查询串对:(分数线,复试);(理财,和迅(一个理财网));(天空(一个软件网),软件网)。
- 相关性类别3:q_T 与 q_I 不存在话题和语义上的关联。

尽管这里只标识了3个类别,但是这3个类别易于区分,后面的实验也证明了其有效性。同时更进一步细化类别,很难区分,可能会导致更多的噪音。相关工作有待于进一步开展。

训练集合的获取分为两个阶段。第一阶段,考虑到时序上的关系,我们首先从切分好的查询会话集中,随机选取2007年3月7日的查询会话,要求选取的查询会话都包含多个查询事务。对采样所得查询会话,去除有切分错误的查询会话,得到训练样本查询会话。实验中我们发现,每天的查询热点都具有一定的突发性。因此这里采用随机采样,而不是按查询频率选择样本。接着从训练样本查询会话构建查询串对,即将查询会话中初始事务查询串 $q(T_1)$ 和所有后续事务的查询串 $q(T_i)$(其中 $i>1$)组合得到查询串对 $(q(T_1), q(T_i))$。对所有的查询串对,过滤含新查询串的查询串对,新查询串是指未在2007.03.01—2007.03.06的查询记录中出现的查询串,从而得到训练集合的第一部分。

第二阶段中,对上面得到的第一部分训练集合中的每个查询串对 (q_I, q_T),从 q_I 的候选查询串集合 $C(q_I)$ 中随机选取查询串 q_C,从而得到新的查询

串对(q_I, q_C),这些新查询串对构成训练集合的第二部分。最后,对训练集合中所有的查询串对进行相关性类别标注,训练集合的类别分布如表1所示。

对训练集合中的所有查询串对,本文按照查询串对间的关联关系抽取了20个特征,这些特征可分为3类,详细的特征描述如表2所示。其中,查询会话相关和点击结果URL相关特征是从2007.03.01—2007.03.06查询会话集合中抽取得到的。

• 查询会话相关:查询会话相关的统计特征,包括会话共现频率、查询频率、初始查询串置信率、目标查询串置信率等特征。

• 查询串文本相关:查询串文本的相关特征,包括查询串长度、查询串词数、查询串对文本串相似度、查询串对编辑距离[12]等特征。

• 点击结果URL相关:点击结果URL相关的统计特征,包括查询串点击的URL数、查询串对URL关联率、目标查询串URL置信率、初始查询串URL置信率等特征。

表2 本文所考虑的查询串对的所有特征(用*标记显著性水平为0.001的特征)

特征	描述
初始查询串长度(Length(q_I))	初始查询串 q_I 的中文字符数
目标查询串长度(Length(q_T))	目标查询串 q_T 的中文字符数
初始查询串词数(WordNO(q_I))	初始查询串 q_I 所包含的的中文词数
目标查询串词数(WordNO(q_T))	目标查询串 q_T 所包含的的中文词数
相同字符数(Common_Character)*	查询串对中的相同中文字符数
相同词数(Common_Word)*	查询串对中的相同中文词数
查询串对文本串相似度(Text_Similarity)*	计算公式如(1)
查询串对文本词相似度(Word_Similarity)*	计算公式如(2)
查询串对文本串编辑距离(Edit_Distance)*	计算公式如(3)
查询串对相同前缀(Prefix_Overlap)*	查询串对最大的相同前缀中文字符数
查询串对会话共现频率(QCO-OC)*	查询串对出现在同一查询会话中的频率
初始查询串查询频率(QCount(q_I))	初始查询串 q_I 出现的在不同查询会话的次数
目标查询串查询频率(QCount(q_T))	目标查询串 q_T 出现的在不同查询会话的次数
初始查询串置信率(Init_Confidence)*	会话共现频率比上初始查询串查询频率
目标查询串置信率(Target_Confidence)*	会话共现频率比上目标查询串查询频率
初始查询串点击的URL数(URLNO(q_I))	查询日志所有记录中,初始查询串 q_I 所对应的不同URL数
目标查询串点击的URL数(URLNO(q_T))	查询日志所有记录中,目标查询串 q_T 所对应的不同URL数
查询串对URL关联率(URL_Support)*	初始查询串与目标查询串的点击的相同的URL数
初始查询串URL置信率(URL_Init_Confidence)*	初始查询串与目标查询串的共同点击的URL数比上初始查询串点击的URL数
目标查询串URL置信率(URL_Target_Confidence)*	初始查询串与目标查询串的共同点击的URL数比上目标查询串点击的URL数

$$Text\_Similarity(q_I,q_T)=\frac{Common\_Character(q_I,q_T)^2}{Length(q_I)\times Length(q_T)} \quad (1)$$

$$Word\_Similarity(q_I,q_T)=\frac{Common\_Word(q_I,q_T)^2}{WordNO(q_I)\times WordNO(q_T)} \quad (2)$$

$$Edit\_Distance(q_I,q_T)=1-\frac{Levenshtein\_Distance(q_I,q_T)}{\max(Length(q_I),Length(q_T))} \quad (3)$$

3 学习算法

对前面得到的训练样本,考虑采用不同的学习算法进行学习。下面分别采用了线性回归(LR),以及支持向量机(support vector machine, SVM)。

3.1 线性回归学习算法

考虑到线性回归模型优越的运算性能,这里采用线性回归分析方法。线性回归模型只要求几个简单的特征就能快速地计算出排序结果。同时实验结果表明线性回归模型非常有效。这里采用标准的线性回归分析方法。

我们首先进行特征选择,从表2所描述的20个特征中选择出能够很好地识别相关查询串的有效特征。选择有效的特征进行学习,不仅在样本不多的条件下可以改善学习算法的性能,而且可以简化特征获取的过程,以提高算法的性能。特征选择的任务是从生成的20个特征中选择出数量为 $d(d<20)$ 的一组最优特征来。为此有两个问题要解决:一是选择的标准,这里采用线性回归统计量拟合优度 R,即要选出使 R 值达到最大的特征组来;另一个问题是特征选择算法,这里采用逐步引入删除法。

通过显著性分析,我们保留显著性水平为0.001的12个特征。然后采用逐步引入删除法,组合不同特征,学习得到最简单有效的回归函数:

$$f(q_I,q_T) = 1.313 + 2.368 \times Text\_Similarity(q_I,q_T) + 0.174 \times Common\_Word(q_I,q_T) + 0.001 \times QCO\_OC(q_I,q_T) + 0.534 \times Init\_Confidance(q_I,q_T) - 0.195 \times Target\_Confidance(q_I,q_T) \quad (4)$$

由于点击数据的稀疏性,点击结果 URL 相关的统计特征未出现在上述公式当中。中文存在分词的问题,由于查询串的简短,查询串分词存在一定的错误。此时,未分词的文本串相似度特征更加有效。同时查询串对中相同的词有助于识别相关的中文查询串对。上述公式可以直观地理解为,对给定的初始查询串,偏向于选择文本上更相似,查询串会话共现更加频繁的候选查询串。公式中的查询会话置信率有助于消除查询会话切分所导致的噪音,识别相关的中文查询串对。

3.2 支持向量机学习算法

支持向量机学习[13]算法具有优越的分类性能。本文采用支持向量机学习算法作为参照对象来考察线性回归方法的性能。首先将所有未归一化的特征用该特征的最大值进行归一化,然后应用支持向量机学习算法进行学习,学习时让支持向量机自动地进行参数修正。

4 结果及其分析

4.1 性能稳定性实验

首先考察学习算法的性能。对上面给定的两个学习算法,我们采用标注的数据进行 100 次随机实验,其中 90%的数据用于训练,10%用于测试。统计平均准确率(precision)和召回率(recall)分布情况。

图 2 给出了 LR 和 SVM 随机实验的平均准确率-召回率散点分布。LR 的平均准确率-召回率点都集中在准确率为 80%,召回率为 80%的位置附近,这说明了 LR 的稳定性。同时,从图中可以看出,LR 的性能与 SVM 的性能基本相当。考虑到线性回归快速有效,后面只采用公式(4)所描述的线性回归模型做预测实验。下面,我们将学习得到的线性回归模型应用于预测实验,进一步考察线性回归模型的预测准确率。

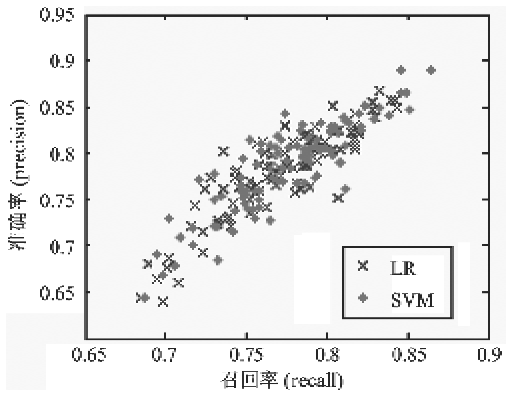

图 2 LR 和 SVM 随机实验的平均准确率-召回率散点图

4.2 预测实验

首先我们从 2007 年 3 月 7 日的查询会话集合中,随机选择了 100 个不同于训练集合的查询会话,使得测试集合不交于训练集合。同时要求每个查询会话含有多个不同查询串,并且每个查询会话的初始查询串和至少一个目标查询串出现在 2007 年 3 月 1 日至 2007 年 3 月 6 日的查询会话集合中。这里将查询会话的目标查询串作为预测的目标。然后对所有测试查询会话的初始查询串,生成其候选推荐查询集合。测试查询会话的候选查询串分布见表 3。

表 3 测试查询会话的候选查询集合大小分布表

候选查询集合大小(size)	查询会话数	所占比率
size < = 10	19	19%
10 < size < = 50	32	32%
50 < size < = 200	30	30%
200 < size < = 500	11	11%
500 < size	8	8%
总查询会话数	100	100%

这里我们将测试查询会话中,目标查询串在结果列表前 n 个中所占的比例作为一种新的度量-预测准确率指标(P@N)。实验中,将我们组合不同特征学习得到的线性回归模型 LR,与文献[5]的聚类算法(clustering, CLU)以及文献[10]的线性模型(linear model, LM)进行了对比分析。预测实验结果见表 4。

表 4 预测结果表(P@N)

	LR	LM	CLU
P@5	0.31	0.15	0.22
P@10	0.45	0.24	0.34
P@20	0.57	0.36	0.42
P@50	0.61	0.51	0.53
P@100	0.65	0.54	0.58
P@500	0.67	0.67	0.67
P@N	0.67	0.67	0.67

由于数据稀疏,有 33%的测试查询会话的目标查询串没有出现在初始查询串的候选查询串集合中。尽管数据中存在大量的模糊查询和查询会话切分错误,但在上面预测实验中,取排序列表的前 10 个结果时,LR

的预测准确率为67.16%(对所有67可以预测的查询会话),优于CLU(50.75%)和LM(35.82%)的预测准确率。由于查询日志中存在大量的噪音,更好地解决噪音问题,挖掘识别相关查询串的有效特征,对相关查询推荐至关重要。实验表明,我们采用回归分析挖掘的识别中文相关查询串的特征非常有效,与以往的方法相比,本文所采用的方法能更好地解决噪音问题和进行有效的相关查询推荐。

5 结 论

本文在分析搜索引擎查询日志的基础上,提出了一种基于线性回归的相关查询推荐方法。该方法考虑了查询串之间的多种关联关系,采用随机采样来选择查询会话,以避免样本选择带来的偏好,同时明确定义了相关性类别,消除样本标注可能导致的噪音,最后采用线性回归分析,挖掘中文相关查询推荐的有效特性。本文最后的实验表明,我们采用回归分析挖掘的识别中文相关查询串的特征非常有效,相比以往的方法,本文所采用的方法能更好地解决噪音问题,能有效地进行相关查询推荐。尽管本文取得了一定的成功,但仍有许多工作有待深入开展,如可以考虑上下文信息,尝试更好的学习算法,构建更加完备的测试方法和更加细致的类别标注等。期望通过上述改进工作能够取得更好相关查询推荐效果。

参考文献

[1] Broder A. Taxonomy of web search. *SIGIR Forum*, 2002,36(2): 3-10

[2] Silverstein C, Marais H, Henzinger M, et al. Analysis of a very large web search engine query log. *SIGIR Forum*, 1998, 33(1): 6-12

[3] Ruthven I, Lalmas M. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 2003,18(2): 95-145

[4] 崔航,文继荣,李敏强.基于用户日志的查询扩展统计模型.软件学报,2003,14(9):1593-1599

[5] Beeferman D, Berger A L. Agglomerative clustering of a search engine query log. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, USA, 2000.407-416

[6] Jones R. Generating query substitutions. In: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland, 2006. 387-396

[7] Zhang Z Y, Nasraoui O. Mining search engine query logs for query recommendation. In: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland, 2006. 1039-1040

[8] Shen X, Tan B, Zhai C. Context-sensitive information retrieval using implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, 2005. 43-50

[9] Cucerzan S, White R W. Query suggestion based on user landing pages. *SIGIR Forum*, 2007, 875-876

[10] 王继民,彭波,孟涛.基于搜索引擎日志发现相近Web查询.北京邮电大学学报,2005,28(1):44-48

[11] He X F, Yan J, Ma J W, et al. Query topic detection for reformulation. Proceedings of the 16th International Conference on World Wide Web, Banff, Canada, 2007. 1187-1188

[12] Chang C C, Lin C J. LIBSVM: A Library for Support Vector Machines, 2001. Software available at http://www.csie.ntu.edu.tw/_cjlin/libsvm

[13] Gilleland M. Levenshtein Distance in Three Flavors. http://www.merriampark.com/ld.htm

Related query recommendation based on linear regressio

Zhai HaiJun[*,**], Zhang Gang[**], Zhang Jin[**]

([*]Department of Computer Science and Technology, University of Science & Technology of China, Hefei 230027)

([**]Research Center of Information Intelligent and Information Security, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

Abstract

In this paper linear regression is applied to related query recommendation. A novel linear regression model is proposed to recommend related queries from query logs. In this model many types of association relationships, including query session co-occurence, URL-clicked sharing and text similarity, are identified and leveraged, to avoid the circumstance that part of these relationships which are directly applied may be largely affected by the noise in query logs. In this work the linear regression is proposed to identify effective features, and this method can effectively deal with the noise problem. The experiments demonstrate that the features identified with linear regression are very effective. Moreover, the prediction precision of the proposed linear regression model outperforms the existing methods.

Key words: query log, query session, related query recommendation, linear regression