时间敏感查询词补全关键技术研究综述

田 萱,张 骁,孟祥光,陈志泊
(北京林业大学信息学院,北京 100083)

摘 要: 搜索引擎的查询词补全技术给搜索用户提供了较好的用户体验.针对用户检索需求随时间变化而不同这一问题,时间敏感查询词自动补全成为研究热点.时间敏感查询词补全在生成查询词补全候选列表时拟合多种时间因素,呈现出与传统查询词补全不同的特点.本文首先介绍了时间敏感查询词补全的定义和分类,然后从查询词时间敏感类型判断、补全候选词权值计算、候选词排序计算三个步骤分析了关键技术,最后对技术评价方法和技术未来发展难点与热点进行了总结和展望.

关键词: 时间敏感;查询词补全;信息检索;候选词权值计算
中图分类号: TP391.1 **文献标识码:** A **文章编号:** 0372-2112 (2015)06-1160-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.06.018

Research Review of Time-Sensitive Query Auto-completion Technique

TIAN Xuan, ZHANG Xiao, MENG Xiang-guang, CHEN Zhi-bo
(*School of Information Science & Technology, Beijing Forestry University, Beijing* 100083, *China*)

Abstract: Query auto-completion of search engines provides good experience for the users. With the user's search intention changing over time, time-sensitive query auto-completion (TSQC) comes to be a research focus. Different from traditional query auto-completion, recommendation list of TSQC is made according to the attaching time features of the query words. First, the definition and classification of TSQC are introduced. Then the key steps of TSQC are presented and analyzed, which include type judgment of time-sensitive query, weight calculation of candidates and recommendation list ranking. Finally, technique evaluation and future development of TSQC are analyzed and summarized.

Key words: time-sensitive; query auto-completion; information retrieval; candidates' weights calculation

1 引言

当前,查询词补全(Query auto-Completion, QC)技术广泛应用于信息检索系统中,使用户花费较少的时间得到符合用户查询期望的搜索词,为用户节省搜索时间,提高用户搜索满意度.目前的查询词补全技术主要依赖群体智慧策略(Wisdom of Crowd),根据大多数用户的需求推测某个用户的查询需求.但这种只考虑点击量、相关度的方法已经不能满足用户对补全准确率、实时性的要求.为了更准确的满足用户个性化查询词补全需求,个性化查询词补全应运而生.个性化查询词补全可以依赖很多不同的个性化上下文因素,如地点、兴趣、时间、桌面行为等,其中时间因素因其能够轻而易举由系统获得并且具有延伸性而成为一个备受关注的重要因素.近年来研究人员并提出了"时间敏感[1~4]查询词补全(Time-Sensitive Query Completion, TSQC)"的概念,利用查询词的时间属性预测用户感兴趣的话题.

在近几年举办的国际信息检索大会(Special Interest Group on Information Retrieval, SIGIR)上,时间敏感的查询词补全广受关注.很多学校和机构,如伊利诺伊大学[5]、佐治亚理工学院[6]、谷歌[1]、以色列理工大学[7]、微软研究院[3]等都强调了时间因素在查询词排序方面的重要性,并提出了一些解决问题的思路.

当前,国内外针对查询词补全技术的综述大都针对传统的查询词补全算法,重点关注词语的补全和构造,极少涉及到查询词的时间敏感性.考虑到时间因素在查询词补全方面具有重要研究意义和实用价值,我们总结该领域现阶段的研究成果,并分析和预测其发展过程中的难点和创新点,期望能够为未来研究工作提供指导和帮助.

收稿日期:2014-01-17;修回日期:2014-09-15;责任编辑:梅志强
基金项目:中央高校基本科研业务费专项资金资助(No.YX2014-19)

2　时间敏感查询词补全概述

2.1　传统查询词补全介绍

查询词是检索过程的开端,直接决定检索结果的内容和质量.但很多时候用户并不能准确输入反映自己检索意图的查询词,或者在输入查询词时对词中表达的人名、地名等专有名词不能确定,或想通过较少操作得到较好结果,此时查询词补全[3,8~10]、模糊匹配[11,12]、查询推荐[13~19]等技术就发挥了重要作用.

查询词补全是指根据用户在搜索框中输入的词语,遵循扩展前缀匹配原则[20],通过一定算法,将用户想要查询的词语补全完整.查询词补全技术可分为两种[3],第一种是随用户输入动态生成并排序候选词.这种类型的补全随用户输入字符进行动态计算,并将通过计算得到的排名首位候选词作为补全词,其余词语以列表形式推荐给用户.另外一种是随用户输入匹配索引中词语,将存储好的候选词列表补全给用户.该类算法需要提前计算好每个词语的候选词列表并存储在索引中,当用户输入字符时,在索引中调用相应词语的候选词列表.

对比以上两种补全类型,第一种方法随用户输入实时计算候选词列表,速度较慢,但有相对较高的准确性、灵活性;第二种补全类型需要较大索引存储空间,实时性不够强,但节省了计算环节,速度较快.两种查询词补全类型根据特点不同,有不同应用.前者较多应用在搜索引擎中,搜索引擎对灵活性实时性要求较高,而且用户可能输入的字符情况也多种多样.后者较多应用在网站的搜索中,例如购物网站,用户在某一网站中可以搜到的词语有限,提前存储并排序所有可能情况可以较快地满足用户需求.

传统判断用户对一个词语的感兴趣程度通常使用词语的点击或查询频率等因素,将词语的访问次数作为补全排序的主要依据[10,20].例如在亚马逊等购物网站,当用户在搜索框中输入商品名称的一个或几个字时,系统就会自动补全匹配该前缀的最受欢迎的(即点击率或查询率最高的)商品.该方法为"最流行匹配算法"(Most Popular Completion[3,8],MPC),直观有效,是目前主流的查询词补全匹配方法,形式化如公式(1)所示.

$$\begin{aligned} \mathrm{MPC}(P) &= \arg\max_{q\in C(p)} w(q), \\ w(q) &= \frac{f(q)}{\sum_{i\in Q} f(i)} \end{aligned} \tag{1}$$

其中,$C(p)$表示候选词列表,$f(q)$表示查询 q 出现在搜索日志 Q 中的次数.

2.2　时间敏感查询词补全(TSQC)

2.2.1　时间敏感查询词补全定义

用户的检索行为随时间发生变化,不同用户在搜索中关注的焦点也不尽相同.即在不同时间,用户的查询倾向不同;在相同时间,不同用户的查询倾向也不同.分析时间因素对用户搜索行为的影响,为用户补全符合时间趋势、季节性、周期性的查询词,将大大提升用户搜索效率和用户搜索满意度.

时间敏感查询词补全不仅考虑词语出现的频率,还要充分分析时间因素对候选词排序的影响.例如,在仅根据访问频率的条件下,"国家统计局"的访问频率高于"国家公务员",并作为输入"国家"的首位补全词.但是在国家公务员考试前期,输入"国家"后,"国家公务员"应该作为首位补全词排在"国家统计局"之前.

考虑到时间因素对查询词补全的影响,在式(1)的基础上,Milad Shokouhi 和 Kira Radinsky[3]拟合时间因素,对时间敏感查询词补全(Time-Sensitive,TS)进行了形式化定义,如式(2)所示.

$$\mathrm{TS}(P,t) = \arg\max_{q\in C(p)} w(q\mid t) \tag{2}$$

其中,$w(q\mid t) = \dfrac{\hat{y}_t(q)}{\sum_{i\in Q}\hat{y}_t(i)}$

$C(p)$、Q 等变量含义与式(1)相同,$\hat{y}_t(q)$表示在时间因素影响下,查询词 q 出现在搜索日志中次数的预测拟合值.该公式准确描述了查询词在时间条件影响下的权值,对时间敏感查询相关研究有重要借鉴意义.

2.2.2　时间敏感查询词分类

根据查询词附带的时间特征不同,本文将查询词分为显式时间敏感查询词(Explicit Time-Sensitive Query,ETSQ)和隐式时间敏感查询词(Implicit Time-Sensitive Query,ITSQ).其中,ETSQ 包括与时间词语语义上紧密相关的查询词,例如词语"SIGIR",它的当前首位补全词为"SIGIR 2015";也包括在特定时间内被密集访问的查询词,例如词语"国家",在国家公务员考试期间,首位补全词为"国家公务员考试".ITSQ 是指除了 ETSQ 以外那些和时间因素关联不明显的词语.

2.3　时间敏感查询词补全整体流程

虽然上述两类时间敏感查询词的时间特征不同,但查询词补全整体思想和流程是相同的:得到查询词后,首先分析查询词,找到满足前缀匹配原则的补全候选词,再应用时间敏感技术计算每个候选词权值,根据权值补全首位候选词,其余形成补全列表,最后依据用户选择结果重新调整候选词权值,总结如图 1 所示.本文第二、三章中会对算法涉及的关键技术做详细的说明和介绍.

3　时间敏感查询词排序算法研究

上述两类时间敏感词语算法思想流程一致,但由于查询词时间特征类型不同,采用的权值计算方法也

就有所不同,技术路线对比如图 2 所示.本章将根据图 2 流程对时间敏感查询词排序算法过程进行分析.

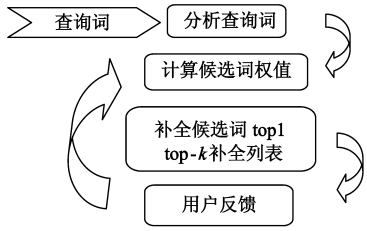

图1 时间敏感查询词补全整体流程

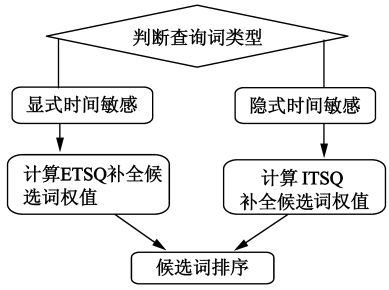

图2 候选词排序技术路线

3.1 判断查询词类型

Wisam Dakka[1],Luis Gravano 等人将 ETSQ 表述为,查询词不是随着时间一直存在,而是被限定在某个时间段里.Donald Metzler 和 Rosie Jones 等人也对 ETSQ 的特点进行了分析[21],词语与不同时间有关并且与时间有关的可能性比与时间无关的可能性大.若某词语在某一个时间段内的查询频率远远高于其他时间,或者该词语经常与某些时间相关词语一同出现在网页文档或用户上下文中,则表示该词语是显式时间敏感,形式化表示如式(3)所示.

$$\text{isETSQ}=\begin{cases}1, & |(y:w(q,t)>0)|\geq 2\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

其中,$w(q,t)=\#(q\cdot t)+\#(t\cdot q)$,isETSQ 表示是否为 ETSQ 的判定.$\#(q\cdot t)$ 表示时间后限定时 q 的访问次数.同理,$\#(t\cdot q)$表示时间前限定时 q 的访问次数.

综合以上分析,本文将 ETSQ 定义为经常与时间词语共同出现或在某一段时间内经常出现的查询词,形式化为式(4).

$$\text{isETSQ}=\begin{cases}1, & P(q,t)\geq\alpha, P(q\mid \text{period}_{t_1\cdots t_n})\geq\beta\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

其中,$P(q,t)=\dfrac{w(q,t)}{\#(q)}$,$P(q\mid \text{period}_{t_1\cdots t_n})=\dfrac{\sum_{i=1}^{n}\#(q\mid t_i)}{\#(q)}$,$w(q,t)=\#(q\cdot t)+\#(t\cdot q)$;$\#(q)$表示 q 出现的总次数,$\text{period}_{t_1\cdots t_n}$表示从时间 t_1到时间 t_n 的时间段,$\#(q\mid t_i)$表示在 t_i时刻下,q 的访问次数.另外,根据 ETSQ 性质,其与时间相关的概率大于与时间无关的概率(即与时间相关的概率大于 0.5),所以 α 和 β 为大于 0.5 的参数.

3.2 ETSQ 候选词权值计算

由于 ETSQ 具有时间属性特征,所以在计算 ETSQ 候选词权值时,既可以采用研究一般词语的方法,也可根据查询词附带的时间属性进行计算.

(1)ETSQ 作为一般词语

若把 ETSQ 作为一般的包含时间相关词的词语,则可根据计算一般词语间相关度的方法得到候选词权值.但考虑到通过计算词汇相关度得到候选词权值的方法已广泛应用于传统查询词匹配算法中,与时间敏感关联较小,所以本文只做简单表述.

在词汇相关度研究中,Heasoo Hwang、Hady WLauw 等人提出了几种基于不同考虑的计算两词语相关度的方法[14],包括基于访问时间、基于检索结果、基于出现频率等等,分别从词语访问时间间隔程度、检索结果相同程度及共同出现概率等不同角度对词语相关度进行分析并进行了实验.实验中,Heasoo 等人采用了三类基线方法,包括基于时间、文本相似度、和查询词点击情况.实验发现,考虑单一方法时无法看到对比强烈的实验结果,所以仅采用一种方法计算词汇相关度并不具备充分的说服力和突出的代表性,因此有待综合考虑多种因素从多种不同角度综合分析拟合,以得到具有较强可信度的词汇相关度结果.

另外计算两词语相似性也是相关度研究的一个方面,这里相似度是指基于词语字符组成的相似程度.其中,Elias Losif 等人提出了词语间的无监督语义相似度计算方法[13](使用基于 web 的度量标准 Web-based metrics),实验中与基于词语、基于页数计算、基于上下文的度量标准进行了对比,实验证明该方法优于最先进的基于监督的语义相似度算法.Shengyue Ji 等人[11]提出树状存储结构,从词语字符的存储角度分析得到效率更高的相似度计算结果,并从单词、多词不同方面进行了模糊匹配的算法分析.Berry MW 等人[22]提出基于 Jaccard 相似性系数的算法,并有效应用于数据挖掘中.以上方法中,Elias 和 Berry 等人关注得到更高准确率的相似性判断,其中 Elias 等人算法特点在于利用网络文档实现无监督的词汇相似度计算;而 Shengyue 等人提出树状结构存储字符,重点关注提高相似度计算方法的效率.判断词语相似度主要从两词语字符的一致性角度分析,在推荐系统中有较大优势.但由于在查询词补全中需遵循最大前缀匹配原则,所以在查询词补全方面的应用有较大局限性.以上算法实验结果对比如表 1 所示.

表 1　ETSQ 作为一般词语时候选词权值算法比较

算法名称	算法特点	实验结果
无监督语义相似度算法[13]	基于 web 的无监督语义相似性度量标准	数据集: Charles-Miller & MeSH 结论:该算法准确率是 Jaccard 算法准确率的 2.15 倍
模糊匹配算法[11]	树状索引存储单词多词快速模糊匹配	数据集: DBLP 和 MEDLINE 大量公开数据 结论:随着字符长度增加,匹配准确率越高于一般算法
基于 Jaccard 相似性系数算法[22]	经典算法,利用词语交集与合集的比值计算相关概率	该算法应用广泛,在无监督语义相似度算法中也有对该算法的验证,而无监督语义相似度算法在相关度比较准确率上优于该算法

(2)ETSQ 作为特殊词语

ETSQ 作为特殊词语,是指将查询词的时间属性加入到相关度计算中,这种时间属性包括两类,第一类是上下文共现的时间属性,上下文共现是指查询词与表示时间的词语同时出现在某个文档局部上下文中,这个时间词语可以用来描述 ETSQ 的时间属性;第二类是与 ETSQ 所在文档访问相关的时间属性,即用文档的访问时间戳来描述 ETSQ 的时间属性.

对于第一种类型,Wisam Dong, Rq Zhang 等人提出了响应一般时间敏感查询方法[1],方法中引入了贝叶斯规则,用文档相关的查询 q 与时间 t 的条件概率 $p(t|q)$来计算词语相关度,如式(5)所示.

$$p(t \mid q) = \frac{p(q \mid t) \cdot p(t)}{p(q)} = \frac{p(q \mid t) \cdot p(t)}{\sum_{\hat{t} \in \text{dates}(D)} p(q \mid \hat{t}) \cdot p(\hat{t})} \tag{5}$$

其中,$p(q|t) = \frac{\#(R_q, t)}{\#(D, t)}$, $\#(R_q, t)$表示与 q 相关文档 R_q 中出现时间 t 的次数, $\#(D, t)$表示在所有文档中出现时间 t 的次数,dates(D)表示文档集 D 的时间跨度.方法引用了计算概率的经典算法贝叶斯规则,并引入了词语的时间属性,最后得到了普遍的、易于理解的基于时间敏感的词语相关度计算方法.

对于第二类时间属性,Jones 和 Diaz 提出了一种语言模型框架[23],利用基础检索模型得到若干个与查询词 q 匹配的文档 d,再进一步计算 q 与访问时间戳 t 有关的概率,从而用文档的访问时间戳来预测 ETSQ 的时间属性.

以上两类时间属性都考虑到了将文档的时间属性与查询词相关联,使查询词得到更加丰富多样的时间属性.对未来研究查询词时间敏感性有重要的参考价值.以上算法实验结果对比如表 2 所示.

在实际应用中,大多应用到 ETSQ 时间属性中的最近访问时间.最近访问时间越接近当前时间,则权值越大,排序越靠前.该算法思想已广泛应用于当前流行搜索引擎中,用户最近搜索过的词会出现在查询词补全的第一位,实现了针对单个用户的简单个性化,是目前较流行的补全算法之一.

表 2　ETSQ 作为特殊词语时候选词权值算法比较

算法名称	算法特点	实验结果
时间敏感查询语言模型[1]	利用文档相关的查询词与时间的条件概率(贝叶斯规则)计算候选词权值	数据集:TREC 和 Newsblaster 多年新闻数据 结论:时间敏感的算法平均优于当前突出的基础语言模型 20%
语言模型框架[23]	将与查询词匹配的文档的时间戳引入到查询词的时间属性中	数据集:TREC 三组新闻数据 结论:75%情况下,利用时间属性对数据进行分类并得到权值的方法平均优于人工方法 6 个百分点

3.3　ITSQ 候选词权值计算

ITSQ 候选词权值计算可采用上述计算 ETSQ 候选词权值的计算方法,但由于 ITSQ 不具备 ETSQ 的明显时间属性,算法效果不佳.针对这个问题,研究人员发现时间序列预测法[3,5,7,24~26]可应用到查询词补全预测中,并取得了一些研究成果.

时间序列(Time-Series)是利用变量本身的历史数据进行预测的方法,由以下四个因素构成:长期趋势(T),季节变动(S),循环变动(C),不规则变动(I).在实际应用中,限于数据范围局限性,可只考虑其中两三个因素进行预测.但随着考虑因素逐渐增多,预测的准确率也逐渐提高.预测步骤如下:搜集数据→分析数据模式→按照模式进行预测.

应用时间序列进行预测的方法有很多种,文献[3]中发现了时间敏感查询词具有趋势和季节性特征,例如用户输入"d"字母时,根据群体策略趋势可得到"dictionary"作为补全词;但考虑季节性因素后,若用户在节假日输入"d"字母,最佳补全词可能为"disney";若加入周期性考虑,用户输入"d"字母的时间为寒暑假,则"disney"的权值要相应增加.针对这个问题,文献[3]采用了时间序列中的指数平滑预测法,如式(6).

$$\bar{y}_t = \lambda y_t + (1 - \lambda)\bar{y}_{t-1} \tag{6}$$

其中 $\bar{y}_t$ 表示在 t 时刻,查询词权值平滑值. λ 为取值 0 到 1 的平滑参数.但单指数平滑法不能充分体现线性趋势在时间序列中的作用,所以加入在时间 t 的线性趋势 F_t 考虑,形成二指数平滑法,如式(7)所示.

$$\begin{aligned} \bar{y}_t &= \lambda_1 y_t + (1 - \lambda_1)(\bar{y}_{t-1} + F_{t-1}) \\ F_t &= \lambda_2(\bar{y}_t - \bar{y}_{t-1}) + (1 - \lambda_2) F_{t-1} \end{aligned} \tag{7}$$

在二指数平滑法的基础上加入季节性考虑 S_t 会让

预测时效性更强,于是产生三指数平滑法,如式(8)所示.

$$\begin{aligned}\bar{y}_t &= \lambda_1(y_t - S_{t-\tau}) + (1-\lambda_1)(\bar{y}_{t-1} + F_{t-1})\\ F_t &= \lambda_2(\bar{y}_t - \bar{y}_{t-1}) + (1-\lambda_2)F_{t-1}\\ S_t &= \lambda_3(y_t - \bar{y}_t) + (1-\lambda_3)S_{t-\tau}\\ \lambda_1 &+ \lambda_2 + \lambda_3 = 1\end{aligned} \tag{8}$$

使用以上时间序列预测算法可以得到较高准确率的预测值.当在原来预测值的基础上再加入新的考虑因素时,预测值的考虑将更全面,预测准确率将得到提升,如表 3[3]所示.

表 3 显示了 4 个不同词语利用时间序列预测算法(TS)、基线算法(P1)预测的实验对比结果,其中 TS(S)表示"考虑到季节性因素的时间序列预测算法",TS(C)表示"考虑到周期性因素的时间序列预测算法".SMAPE(Symmetric Mean Absolute Percentage Error)和 MAE(Mean Absolute Error)的值越小,预测结果越好.

时间序列预测算法中,根据实际情况,可挑选时间序列因素中的一种或几种进行预测.当考虑多种因素时,应用过程可以如图 3 所示.

表 3 ITSQ 候选词权值计算实验结果对比

	"irs"		"australian open"		"nascar"		"irish lottery"	
	TS(S)	P1	TS(S)	P1	TS(C)	P1	TS(C)	P1
SMAPE	0.18	0.26			0.06	0.19		
MAE			0.49	0.52			0.09	0.52
TS 与 P1 对比	TS 比 P1 优出 31%		TS 比 P1 优出 6%		TS 比 P1 优出 68%		TS 比 P1 优出 83%	
TS(S)与 TS(C)对比	SMAPE 结果 TS(C)比 TS(S)优出 67%,MAE 结果 TS(C)比 TS(S)优出 82%							

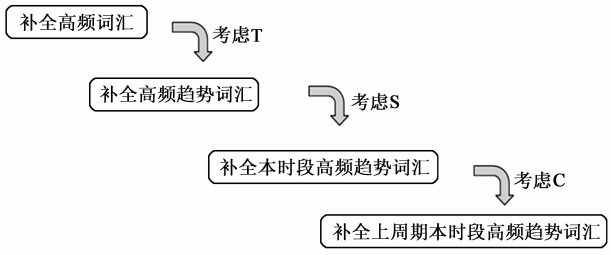

图3 时间序列在查询词补全中应用过程

3.4 候选词排序技术研究

利用时间敏感技术得到候选词权值后,即可根据权值生成候选词列表.但在生成候选词列表时,既可将权值计算应用于首次列表生成过程中,也可将权值计算应用于首次排序(传统算法得到的排序列表)后的二次排序过程中.两种排序类型分别为时间中排序和时间后排序.

(1)时间中排序

时间中排序是指将时间因素始终应用于权值计算过程.传统查询词补全排序多用机器学习两类分类器分析数据集为正样本的概率来预估点击率,例如逻辑回归、支持向量机、决策树/促进树、神经网络等[15].

时间敏感的候选词排序算法在原算法基础上加入含有时间上下文因素的用户偏好分析,结合预估点击率从未评分项目集中选出偏好值最高的若干个项目作为补全集列表.针对偏好值的计算,文献[16]基于贝叶斯网络模型;文献[18]采用朴素贝叶斯分类器计算时间等上下文信息与项目的关联概率;另外,文献[27,28]虽然没有明确提及时间因素的结合分析,但提出利用马尔科夫链蒙特卡洛方法获取用户属性、项目属性和上下文等各种关联信息,继而得到各种可结合的模型参数,该方法也对时间敏感数据采集有启发意义.

另外,随着社交网络的逐渐发展,社交网络实时信息也逐渐应用到查询词补全中.文献[6]提出一种利用 Twitter 的及时性、实时性,挖掘最新敏感检索数据信息进行结果项排序的方法来提升效率.文献[29,30]选择性地监视新闻文档来得到最近敏感查询词汇,并将敏感内容应用到信息检索结果排序算法中.文献[31]提出一种使用高精度分类器自动监视并响应近期敏感查询的检索系统,该系统使用机器学习算法训练数据集得到排序模型,并响应最近敏感查询词汇.由于社交媒体、网络,新闻,门户网站等网络社区具有较高实时性、及时性,所以在时间敏感查询词补全中具有较高利用价值.

(2)时间后排序

时间后排序是一种重排序(Re-ranking)[4,32],首先采用传统信息检索算法生成一次补全结果,再拟合时间因素生成最终补全列表.重排序的关键技术在于将时间因素结合到首次补全列表排序中.文献[32]提出了一种利用时间语言模型,结合查询时间或潜在时间趋势生成重排序补全列表的有效算法.文献[4,33]提出在时间敏感查询中利用反馈控制调节思想,通过挖掘查询日志和文档时间戳标识网页标题和链接,并利用排序误差调节排序行为.文献[34]通过关注首次排序行为的不同点理解用户意图并根据首次排序结果的数量和时间特点对查询分类,以采用时间序列和周期性分析

达到重排序的目的.上述方法思想大都一致,但均存在不可控因素,经过首次补全排序得到的候选词列表中并没有与时间相关性较大的候选词,所以该思想有待进一步深入研究.

4　时间敏感的查询自动补全技术效用评价

4.1　时间敏感的查询词补全相关数据集

在 TSQC 中,时间因素往往依托于用户个性化操作和用户私人信息,所以涉及到可行性、隐私性等问题,都给研究带来很大挑战.但研究人员还是结合具体研究情况使用了多种形式的数据集,包括真实数据集(nature data set)、模拟数据集(synthetic data set)和混合数据集(mixed data set).

(1)真实数据集.真实数据集在检索结果效用评价中应用广泛,如文献[1,3,4,6,7,21,25,32,35,36].真实数据集具有较强说服力,能较好地反映客观事实,但也具有一定局限性.例如,适应性不强、样本容量有限等.

真实数据集具备较强的普遍适用性,在信息检索领域效用评价中常用到的真实数据集包括来自 Bing、TREC、WIKI、UKGOV 等.如文献[7]采用了 Bing 中从 2010 年 12 月 15 日到 2011 年 4 月 25 日的用户访问行为数据,数据包含每个查询对应的 URL 及其中每个 URL 的点击情况;文献[1,5]使用 TREC 数据集,其中文献[1]采用了 TREC 卷四、卷五的数据,包括金融时报中从 1991 到 1994 年和洛杉矶时报中 1989 到 1990 年的新闻文章,但不包括国会议事录和联邦登记中的文件档案;文献[5]包括 TREC 中来自 LA Times、aquaint、genomics 和 wt10g 的近三百万篇文档;文献[35]使用两组数据集分别来自 WIKI 和 UKGOV,其中 WIKI 的数据集包含从 2011 年 1 月到 2005 年 12 月的用户编辑历史并去除编辑量较少的记录,UKGOV 数据集由网络存储基础提供,包括 2004 到 2005 年 11 个英国政府网站每周爬虫记录.

(2)模拟数据集.此类数据根据一定规则自动生成,适应性较强,可用于评价大规模复杂信息排序效果,如文献[37],根据时间序列算法特点,将 500 个时间序列点模拟成正弦和余弦两种类型数据集.但与真实数据集相比,模拟数据集说服力不强,适用算法类型少,评价效果不突出.

(3)混合数据集.此种数据类型中既包含真实数据集,也包含模拟数据集,采用两种(或多种)数据集混合的方式全面对算法效果进行评价,如文献[5,14,38],均在实验时涉及了真实和模拟两类数据集,例如文献[14],既使用了某商业搜索引擎的真实用户搜索日志,同时为了更好地拟合论文算法,又加入了 200 组用户在论文搜索引擎上的搜索行为数据.混合数据集综合以上两类数据集的优点,具有较好的评价效果.

4.2　时间敏感查询词补全效用评价指标

TSQC 效用评价指标与传统信息检索对结果的评价指标一致,需根据具体研究数据集情况选择合适的评价指标.评价系统的主要评价指标是查全率和查准率,查全率是系统检出相关网页结果的能力,而查准率是系统拒绝不相关文献的能力,评价指标包括 nDCG[4,6,21], MRR[3], Spearman (ρ)[3], MAP[1,5,32], P@N[1,5,32], RMSE[5], Precision[7,25,32,37], Recall[7,25,37]等等.根据不同数据集情况,可采用其中某一种评价指标作为补全结果评价标准考量,也可选择多种标准对结果进行综合比较分析.评价的目的是为了改进以得到更好的操作结果,所以得到评价效果后需对效果进行分析.若效果不理想,不能得到用户满意的排序结果,需要进一步分析计算过程中出现的问题,修正算法,逐渐优化系统.

5　时间敏感的查询自动补全技术研究难点与热点

随着用户对搜索服务要求的不断提高,TSQC 领域研究取得了一定的进展,但作为新兴技术研究还存在许多问题和挑战.

(1)显/隐式时间敏感定义

根据查询词的显隐性采用不同的查询词补全策略是 TSQC 关键步骤.但从定义发现,词语的显隐性定义并不清晰.对于一个词语与时间相关的概率大小,无法给出确切的计算方法.所以,找到区分显隐式词语更明确的界限,或找到更好的从不同角度对词语分类的方法,或研究准确的对两类词语均适用的补全算法将成为 TSQC 算法重要研究方向.

(2)时间因素局限性

时间因素虽性质固定,但形式多样.在得到复杂的时间因素后,选择哪些时间因素建模使之更有效地参与 TSQC 形式规约,如何有效利用全部或部分信息分析建模并没有统一明确的定义或模型.另外对于精确化候选词排序结果,仅考虑时间因素对计算过程的影响是远远不够的.在考虑时间敏感的前提下,结合用户信息、上下文信息等将有效提高 TSQC 准确率.另外,TSQC 最终是为用户补全用户满意的信息,所以综合交叉研究用户心理、用户情感等学科知识将是一个值得深入研究的问题.

(3)冷启动问题

TSQC 存在冷启动问题,在没有任何数据反馈的条件下,检索出符合用户时间要求的信息存在挑战.但在用户信息不外泄、用户使用个性化补全系统(如单机信息检索)等情况下,冷启动问题终究无法避免.解决实际应用中 TSQC 系统冷启动问题值得相关研究人员关

注.目前的实验证实,该问题可从两方面入手:优化分析多用户横向信息[39]和分析单用户多属性纵向信息[40].协同过滤是一种应用最广的"优化分析多用户横向信息"方法,从其他相似用户信息中获取目标用户信息.例如,文献[41]采用了协同过滤系统的分类方法,结合相似技术和预测机制,利用人口行为数据确定具有相似行为的用户,利用相似用户产生推荐.文献[42]从Twitter应用中获取目标用户的好友信息,应用LDA(Latent Dirichlet Allocation)[43]模型产生潜在的用户群体,从潜在的用户群体中学习目标用户信息."分析单用户多属性纵向信息"是从已知的目标用户信息学习未知的目标用户信息.例如,文献[44]提出一种上下文感知半监督协同训练方法.使用因素分解模型来捕捉细粒度的用户内容,使用不同的上下文实例构造的不同预测模型,然后采用合作培训策略,让弱预测模型学习其他的预测模型.文献[45]采用功能矩阵分解方法,构造决策树,决策树的每个节点是用户的已知一个信息,根据决策树上用户的信息可适性的推荐.以上两类方法应用条件不同,要根据用户个人信息是否丰富,当前网络环境是否允许获得多用户横向信息等来选择.通过以上两类信息丰富用户初始化信息,是目前解决冷启动问题的主流方法.

(4)TSQC效率

在对时间敏感查询词进行补全或生成补全列表的过程中存在计算效率问题.用户在输入查询词时,往往希望补全列表可以随用户输入实时出现,而无需等待,这给补全列表的生成效率带来巨大挑战.目前很多搜索引擎都较好地解决了这个问题,例如文献[12,41]指出使用关联模型(Relevance models)可将反应速度提高.但在实际应用中,较少研究涉及到TSQC的效率分析,于是如何更快、更及时地得到时间敏感补全列表也是值得关注的研究方向.

(5)补全结果评价

TSQC的算法评价仍然采用传统信息检索评价方法,使用国际通用指标.但TSQC一切从用户的角度出发,在结果评价上也应注重用户体验,而传统评价方法指标单一,无法从人性化角度对结果做出评价,尤其缺少对时间因素影响力的评价,在结果评价使用的数据集方面,覆盖面窄、代表性不强等问题都为准确的结果评价带来障碍.设计更加人性化、精准、灵活的结果评价机制对TSQC具有重要意义.

(6)用户隐私问题

用户隐私和安全问题是所有个性化系统共同的问题,只有全面了解用户才能对用户需求把握到位,进行个性化的补全和检索.TSQC中,在提取时间信息时,群体智慧策略和个性化策略都将涉及用户个人隐私信息、操作信息,而且要根据一定规则提取用户操作的时间因素,这必将为用户带来不安全感.由于目前并没有完善解决该问题的有效措施方案,所以在实际应用中有重重障碍.因此,对TSQC的隐私安全保护和策略研究将成为研究热点.

6 结束语

在信息化的今天,信息检索中的技术一直不断更新,而更好、更快、更准的补全查询词不仅会推进信息检索学科的发展也会改变人们的工作生活.人们的思想、行为会随着时间的变化而剧烈变化,所以在信息检索中考虑时间敏感的查询补全必将大大提高信息检索的效率.鉴于目前国内外还没有针对该领域的综述文章,所以本文对时间敏感的查询词补全的关键技术做了介绍说明,希望对该领域的发展有促进作用.

参考文献

[1] Dakka W, Gravano L, Ipeirotis PG. Answering general time-sensitive queries[J]. Knowledge and Data Engineering, IEEE Transactions on, 2012;24(2):220 – 235.

[2] Fu C-L, Silver D. Time-sensitive Sampling for Spam Filtering [M]. Ontario, Canda: Springer, 2004. 551 – 553.

[3] Shokouhi M, Radinsky K. Time-sensitive query auto-completion [A]. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. Portland: ACM, 2012. 601 – 610.

[4] Zhang R, Chang Y, Zheng Z, Metzler D, Nie J-y. Search result re-ranking by feedback control adjustment for time-sensitive query[A]. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume [C]. Boulder: Association for Computational Linguistics, 2009. 165 – 168.

[5] Efron M. Linear time series models for term weighting in information retrieval[J]. Journal of the American Society for Information Science and Technology, 2010, 61(7): 1299 – 1312.

[6] Dong A, Zhang R, Kolari P, Bai J, Diaz F, Chang Y, et al. Time is of the essence: Improving recency ranking using twitter data[A]. Proceedings of the 19th International Conference on World Wide Web[C]. Raleigh: ACM, 2010. 331 – 340.

[7] Radinsky K, Svore K, Dumais S, Teevan J, Bocharov A, Horvitz E. Modeling and predicting behavioral dynamics on the web[A]. Proceedings of the 21st International Conference on World Wide Web[C]. Portland: ACM, 2012. 599 – 608.

[8] Bar-Yossef Z, Kraus N. Context-sensitive query auto-completion[A]. Proceedings of the 20th International Conference on World Wide Web[C]. Hyderabad: ACM, 2011. 107 – 116.

[9] Bast H, Weber I. Type less, find more: fast autocompletion

search with a succinct index[A]. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Seattle: ACM, 2006.364 – 371.

[10] Chaudhuri S, Kaushik R. Extending autocompletion to tolerate errors[A]. Proceedings of the 35th SIGMOD International Conference on Management of Data[C]. Providence: ACM, 2009.707 – 718.

[11] Ji S, Li G, Li C, Feng J. Efficient interactive fuzzy keyword search[A]. Proceedings of the 18th International Conference on World Wide Web[C]. New York: ACM, 2009.371 – 380.

[12] Li G, Wang J, Li C, Feng J. Supporting efficient top-k queries in type-ahead search[A]. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Portland: ACM, 2012.355 – 364.

[13] Iosif E, Potamianos A. Unsupervised semantic similarity computation between terms using web documents[J]. Knowledge and Data Engineering, IEEE Transactions on, 2010, 22(11): 1637 – 1647.

[14] Hwang H, Lauw HW, Getoor L, Ntoulas A. Organizing user search histories[J]. Knowledge and Data Engineering, IEEE Transactions on, 2012, 24(5): 912 – 925.

[15] Khribi MK, Jemni M, Nasraoui O. Automatic recommendations forelearning personalization based on web usage mining techniques and information retrieval[A]. Advanced Learning Technologies, 2008 ICALT'08 Eighth IEEE International Conference on[C]. Piscataway: IEEE, 2008.241 – 245.

[16] Ono C, Kurokawa M, Motomura Y, Asoh H. A Context-Aware Movie Preference Model Using a Bayesian Network for Recommendation Andpromotion[M]. Berlin Heidelberg: Springer, 2007.247 – 257.

[17] Xu H-L, Wu X, Li X, Yan B. Comparison study of Internet recommendation system[J]. Journal of Software, 2009, 20(2): 350 – 362.

[18] Yu Z, Zhou X, Zhang D, Chin C-Y, Wang X. Supporting context-aware media recommendations for smart phones[J]. Pervasive Computing, IEEE, 2006, 5(3): 68 – 75.

[19] 付博,赵世奇,刘挺. Web 查询日志研究综述[J]. 电子学报,2013,41(9):1800 – 1808.
Fu Bo, Zhao Shiqi, Liu Ting. Research on analysis and mining of web query logs[J]. Acta Electronica Sinica, 2013, 41(9): 1800 – 1808.

[20] Alfonseca E, Ciaramita M, Hall K. Gazpacho and summer rash: Lexical relationships from temporal patterns of Web search queries[A]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing[C]. Philadelphia: Association for Computational Linguistics, 2009. 1046 – 1055.

[21] Metzler D, Jones R, Peng F, Zhang R. Improving search relevance for implicitly temporal queries[A]. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Boston: ACM, 2009.700 – 701.

[22] Berry MW, Browne M. Lecture Notes in Data Mining[M]. Singapore: World Scientific, 2006.27 – 38

[23] Jones R, Diaz F. Temporal profiles of queries[J]. ACM Transactions on Information Systems (TOIS), 2007, 25(3): 14.

[24] Kim HD, Nikitin D, Zhai C, Castellanos M, Hsu M. Informationretrieval with time series query[A]. Proceedings of the 2013 Conference on the Theory of Information Retrieval[C]. New York: ACM, 2013.14.

[25] Shokouhi M. Detecting seasonal queries bytime-series analysis [A]. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. Beijing: ACM, 2011.1171 – 1712.

[26] Hamilton JD. TimeSeries Analysis[M]. Cambridge: Cambridge University Press.1994.

[27] Adomavicius G, Sankaranarayanan R, Sen S, Tuzhilin A. Incorporating contextual information in recommender systems using a multidimensional approach[J]. ACM Transactions on Information Systems (TOIS), 2005, 23(1): 103 – 145.

[28] Adomavicius G, Tuzhilin A. Context-Aware Recommender Systems[M]. New York: Springer, 2011.217 – 253.

[29] Diaz F. Integration of news content into web results[A]. Proceedings of the Second ACM International Conference on Web Search and Data Mining[C]. Barcelona: ACM, 2009. 182 – 191.

[30] König AC, Gamon M, Wu Q. Click-through prediction for news queries[A]. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information retrieval[C]. Boston: ACM, 2009.347 – 354.

[31] Dong A, Chang Y, Zheng Z, Mishne G, Bai J, Zhang R, et al. Towards recency ranking in web search[A]. Proceedings of the Third ACM International Conference on Web Search and Data Mining[C]. New York: ACM, 2010.11 – 20.

[32] Kanhabua N, Nørvåg K. Determining Time of Queries for Re-Ranking Search Results[M]. Berlin Heidelberg: Research and Advanced Technology for Digital Libraries, 2010.261 – 72.

[33] Agichtein E, Brill E, Dumais S. Improving web search ranking by incorporating user behavior information[A]. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Seattle: ACM, 2006.19 – 26.

[34] Murata M, Toda H, Matsuura Y, Kataoka R, Mochizuki T. Detecting periodic changes in search intentions in a search engine[A]. Proceedings of the 19th ACM International Conference on Information and Knowledge Management[C]. Toron-

to: ACM, 2010. 1525 – 1528.
[35] Anand A, Bedathur S, Berberich K, Schenkel R. Index maintenance for time-travel text search[A]. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Portland: ACM, 2012. 235 – 244.
[36] Vlachos M, Meek C, Vagena Z, Gunopulos D. Identifying similarities, periodicities and bursts for online search queries [A]. Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data[C]. New York: ACM, 2004. 131 – 142.
[37] Keogh EJ, Pazzani MJ. Relevance feedback retrieval of time series data[A]. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Berkeley: ACM, 1999. 183 – 190.
[38] Wang TD, Deshpande A, Shneiderman B. A temporal pattern search algorithm for personal history event visualization[J]. Knowledge and Data Engineering, IEEE Transactions on, 2012, 24(5): 799 – 812.
[39] Zhang Z-K, Liu C, Zhang Y-C, Zhou T. Solving the cold-start problem in recommender systems with social tags[J]. EPL (Europhysics Letters), 2010, 92(2): 28002.
[40] Gantner Z, Drumond L, Freudenthaler C, Rendle S, Schmidt-Thieme L. Learning attribute-to-feature mappings for cold-start recommendations[A]. Data Mining (ICDM), 2010 IEEE 10th International Conference on[C]. Shenzhen: IEEE, 2010. 176 – 185.
[41] Blerina Lika, Kostas Kolomvatsos, Stathes Hadjiefthymiades. Facing the cold start problem in recommender systems[J]. Expert Syst, 2014, 41(4): 2065 – 2073.
[42] JLin, KSugiyama, M-Y Kan, T-S Chua. Addressing cold-start in app recommendation: Latent user models constructed from twitter followers[A]. 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013)[C]. Dublin: ACM, 2013. 283 – 292.
[43] Heping Li, Feng Zhang, Shuwu Zhang. Multi-feature hierarchical topic models for human behavior recognition[J]. Science China Information Sciences, 2014, 57(9): 1 – 15.
[44] Mi Zhang, Jie Tang, Xuchen Zhang, Xiangyang Xue. Addressing cold start in recommender systems: a semi-supervised co-training algorithm[A]. 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'14)[C]. Gold Coast: ACM, 2014. 73 – 82.
[45] Ke Zhou, Shuang-Hong Yang, Hongyuan Zha. Functional matrix factorizations for cold-start recommendation[A]. 34th International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. New York: ACM, 2011. 315 – 324.

作者简介

田 萱 女,1976 年生于山东济宁.北京林业大学信息学院副教授.研究方向为智能信息处理、智能信息检索.
E-mail: tianxuan@bjfu.edu.cn

张 骁 女,1989 年生于辽宁抚顺.北京林业大学计算机软件与理论专业硕士研究生.研究方向为智能信息处理.
E-mail: zhangxiao0818@163.com

陈志泊(通信作者) 男,1967 年出生于山东日照,北京林业大学毕业,现为北京林业大学信息学院教授、博士生导师,研究方向为数据库技术.
E-mail: zhibo@bjfu.edu.cn