文章编号: 1003-0077(2010)03-0055-07

基于检索历史上下文的个性化查询重构技术研究

宋 巍,张 宇,刘 挺,李 生

(哈尔滨工业大学 信息检索研究中心,黑龙江 哈尔滨 150001)

摘 要: 基于检索历史隐式地学习用户偏好是个性化检索研究的热点,而根据用户检索历史重构新的查询输入是其中主要的研究内容。已有的研究在利用检索历史进行查询重构时,通常不区分检索历史中的内容是否与当前查询相关,而是将全部检索历史视为整体,因而使重构后的查询含有较多噪声。该文基于相关词语在上下文中大量共现的特征,将用户历史检索结果的网页摘要作为上下文语境,结合用户点击,选择检索历史中与当前查询共现程度最高的词语重构查询模型。对初始检索结果重排序的实验表明,该方法可以有效地选择相关词语,减少噪声。用 p@5 和 NDCG 两种指标评价,比最好的基准系统分别相对提高 12.8%和 7.2%,比初始排序结果相对提高 26.0%和 11.4%。

关键词: 计算机应用;中文信息处理;个性化检索;隐式反馈;查询重构

中图分类号: TP391 **文献标识码**: A

Personalized Query Reformulation Based on Search Context

SONG Wei, ZHANG Yu, LIU Ting, LI Sheng

(Center for Information Retrieval, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Learning user preference implicitly is a hot research topic for personalized search, and query model reformulation based on user search history is a key issue. Existing work considers the search history as a whole without distinguishing whether it is relevant to current query, resulting in much noise. In this paper, assuming that the relevant terms tend to co-occurrence in context, we treat each past snippet as a context and reformulate the query by selecting the most relevant terms to the whole query from the user clicks. The experiment results show that the algorithm can select relevant terms and reduce noise. With the evaluation metrics of p@5 and NDCG, the system achieves a relative improvement against the best baseline system by 12.8% and 7.2% respectively, 26.0% and 11.4% against the original ranking.

Key words: computer application; Chinese information processing; personalized web search; implicit feedback; query reformulation

1 引言

当前,通用搜索引擎主要基于关键词匹配的方法进行检索。存在的一个问题是:用户查询时输入的有限词语并不能完全准确表达其检索的真正意图,查询本身存在的歧义性导致搜索引擎返回大量与用户需求无关的文档。另一方面,具有不同应用背景、偏好的用户,在输入相同的查询词时可能也有着各自不同的信息需求。因此,系统对输入相同查询关键词的所有用户返回同样的结果不能使单个用户满意度达到最大。鉴于以上原因,结合用户反馈

收稿日期: 2009-09-11 **定稿日期**: 2010-03-14

基金项目: 国家自然科学基金重点资助项目(60736044);国家自然科学基金面上资助项目(60675034);国家 863 计划探索类专题资助项目(2008AA01Z144);语言语音教育部—微软重点实验室开放基金资助(HTT.KLOF.2009020)

作者简介: 宋巍(1983—),男,博士生,主要研究方向为个性化信息检索,个性化推荐;张宇(1972—),男,博士,副教授,CCF 高级会员,主要研究方向为信息过滤,自动问答,自然语言处理;刘挺(1972—),男,博士,教授,博导,CCF 高级会员,主要研究方向为自然语言处理,信息检索。

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的个性化检索成为近年来学术界[1-2]研究的热点。

用户的反馈信息通常用来重构当前的查询,以使新的查询模型与用户检索意图更为接近。按照反馈方式的不同,用户反馈可分为显式反馈和隐式反馈两种。显式反馈指用户主动向系统提供自己的兴趣偏好或对系统返回的结果进行相关性评价。隐式反馈指通过分析用户与系统正常的交互行为来推测用户检索意图,不需用户做额外的相关性评价。由于用户通常不愿花费精力进行主动反馈,隐式反馈成为个性化信息检索研究的重点。

用户的检索历史是隐式反馈信息最主要的来源之一,通常包括查询输入、结果集、用户点击等,含有用户多方面的偏好信息,同时也存在大量噪声。以往的方法将用户检索历史当成一个整体考虑,或者将其视为历史单元的融合,利用其中的所有词来重构查询模型。但是,检索历史中并非所有词语都与当前查询相关,这种方法自然会导致新的查询模型中包含与当前查询无关的噪声词,影响最终的检索性能。

本文认为与当前查询相关的词语和查询中的词语在检索历史中经常共现。词语的上下文指在它周围一定大小的文本窗口内出现的所有文本。两个词语共现指它们出现在同一上下文内。我们以检索结果的网页摘要(snippet)作为上下文,结合用户点击对词语的共现关系进行建模,考虑检索历史中的候选词语与当前查询中的所有词语的共现关系,进而选择与整个查询相似度最高的候选词作为扩展词语。通过对检索结果重排序的实验证明,该方法可以有效地从用户历史检索中挖掘出与查询相关的词语,减少噪声词,提高排序质量。

本文第2节介绍基于用户检索历史的个性化研究相关工作;第3节叙述基于检索历史上下文的查询重构方法;第4节和第5节分别介绍实验的设置及实验结果分析;第6节做出结论,并对下一步的研究工作进行展望。

2 相关工作

挖掘检索历史的方法可分为基于短期历史和长期历史两种。短期历史针对单个的查询会话(query session)中用户的反馈来修正查询模型[3-5]。这类方法优点是反馈直接针对当前查询,噪声较少,缺点是可获得的信息有限。与之相比,基于长期历史的方法则以用户为中心,收集从不同来源获取的用户信息,建立长期用户模型对当前查询模型进行重构。Sugiyama K等[6]以时间为主轴划分用户的浏览信息为长期反馈、近期反馈和当前反馈进行线性融合作为查询模型。类似的还有基于用户桌面索引[7]和领域本体[8]的方法。这类方法优点是无需进行查询会话划分,能够全面刻画用户兴趣。缺点是长期历史包含多个主题,存在大量噪声。从中发现与当前查询相关的信息,利用这些信息预测用户的检索意图是高效利用长期历史进行个性化检索的关键。

Bin Tan等[9]将用户检索历史以查询为单位构建若干历史单元,计算当前查询与历史单元的相似度作为权值对历史单元进行线性插值形成用户历史模型。其目的是赋予与当前查询相似度高的历史单元中的词语更高的权值,降低了不相关单元对最终查询模型的影响。然而计算查询之间的相似度本身是很难的任务,因为查询较短反映的信息有限,若使用返回的结果集对查询进行扩充则依赖于初始返回的结果,若其中包含很多不相关文档,计算出的相似度与实际情况会有较大偏差。此外该方法使用检索历史中出现的所有词语建模,用户检索历史中的一些词语与查询并不相关,但由于出现次数较多也会获得较高的权值形成噪声。Jing Bai等[3]利用查询的上下文及查询内的词语间关系从相关的兴趣领域中挖掘与当前查询相关的词语重构查询模型。本文与其思想类似,不同的是本文针对用户的检索历史进行挖掘并对用户反馈进行建模,在选择相关词语的过程中考虑了用户的个性化信息。

与本文相关工作还包括基于伪相关反馈的查询扩展[10-11]。这类方法假设初次检索排序靠前的文档或段落与查询相关,并从中选择词语扩展查询。其问题是过于依赖于系统初次检索的质量。已有学者尝试个性化查询扩展方法。Paul Alexandru Chirita等[12]从用户个人桌面文档中选择词语扩展到查询模型中。梅翔等[13]将用户对网页的偏好转化为对知识库中词语的偏好,建立用户兴趣模型挑选出与用户偏好关联最紧密的关键词加入原查询。

3 基于检索历史上下文的查询重构

检索历史包含不同的主题,其中多数与当前查询无关。同时,一篇文章中仅有部分词语能够反映其主题,其余词语起辅助作用。检索历史中与当前查询无关的词语形成噪声,其来源可分为两类:一类为不相关主题中的词语,另一类为在各种主题中广

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

泛存在的起辅助作用的词语,随着检索历史不断增加,此类噪声不断累积。

本文以用户当前查询为中心,基于相关词语在检索历史上下文中的共现及用户点击信息,选择检索历史中与当前查询最相关的词语重构查询模型。设当前查询为 $Q = \{q_i\}$,其中 q_i 是查询关键词。在用户检索历史中,每个历史查询可对应一组信息,这些信息可用一个元组 < 查询输入,结果集,点击页面 > 来表示,结果集包括返回结果中所有网页的标题、摘要以及正文链接。查询模型重构过程如下:

1) 将用户检索历史中的网页摘要进行索引,用当前查询从中检索,得到相关的历史查询网页摘要并提取其中的词语形成候选词语集。

2) 选取候选词语的一个子集来重构查询模型,称该子集中的词语为扩展词语。以网页摘要作为上下文语境,计算每个候选词语与当前整个查询的相似度并依此对候选词语进行排序,选取前 k 个候选词语作为扩展词语。

3) 利用得到的扩展词语重构查询模型。

最后,利用新的查询模型对初始的检索结果进行重排序。具体处理过程如图 1 所示。

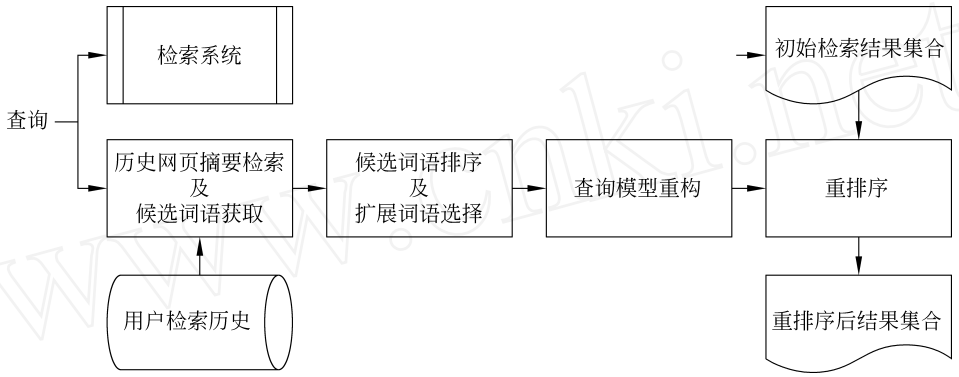

图 1　查询模型重构过程

3.1　候选词语获取

候选词语应与当前查询中的词语在检索历史中经常共现。我们将用户的检索历史以网页摘要为单位(即将一个网页摘要作为一个独立文档)进行索引,以当前查询作为关键词集合(去除其中的停用词),检索索引的历史查询网页摘要,然后选用排序靠前的 n 个网页摘要,从中提取所包含的词语形成候选词语集。索引时网页摘要按照 TF-IDF 进行建模。当前查询按照公式(1)估计权值。

$$w(t) = (0.5 + \frac{0.5tf(t,Q)}{\max tf(*,Q)}) \cdot \log \frac{N}{df(t) + 1} \tag{1}$$

其中,t 为查询中的一个词语,$tf(t,Q)$ 表示 t 在查询 Q 中出现的次数。$df(t)$ 为包含 t 的网页摘要数,N 为网页摘要总数。

3.2　扩展词语选择

本文利用词语在检索历史上下文内的共现并结合用户点击作为隐式反馈,来选取扩展词语。首先,考察候选词语与查询中单个词语的共现关系,在此基础之上计算其与整个查询的相似度,最终选取与整个查询相似度最大的 k 个候选词语作为扩展词语重构查询模型。

在用户的检索历史中,对应某个查询,系统返回的结果中可能包含多个主题,其中只有部分主题的网页文档真正满足用户的信息需求。利用词语在上下文共现来提取相关词语时,选择的文本窗口过大会引入很多与当前查询并无关系的噪声词语。因此,我们将检索历史中搜索引擎给出的网页摘要作为度量词语共现的上下文。一个网页摘要是相对较小的文本窗口而且主题一致。以网页摘要作为上下文能够更好地估计词语间的共现关系,同时易于结合用户反馈。

3.2.1　词语间共现度

首先,考察候选词语与查询中单个词语的共现关系。采取共现度[11]来度量两个词语共现程度,其基本计算公式如(2)所示。

$$co\_degree(c,q_i) = \log(co(c,q_i))\, idf(c) / \log(n) \tag{2}$$

其中,

$$co(c,q_i) = \sum_{j}^{n} tf(c,s_j)\, tf(q_i,s_j) \tag{3}$$

s_j 为排序在前 n 位的网页摘要。$tf(c,s_j)$ 表示词语 c 在 s_j 中出现的次数,$tf(q_i,s_j)$ 表示词语 q_i 在

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

s_j 中出现的次数。$idf(c)$ 为词语 c 在整个数据集上的逆文档数,以降低高频词语的权值。$co(c,q_i)$ 用来度量词语 c 与 q_i 在前 n 个网页摘要中的共现次数。显然,两个词语共同出现的网页摘要数越多 $co(c,q_i)$ 的值越大。

3.2.2 结合用户点击的共现度

公式(3)仅对词语共现进行统计而没有考虑网页摘要的质量,也没有结合用户的反馈。在个性化检索中,用户的反馈信息对于预测其真正的检索意图起着重要的作用。我们将公式(3)进行修改以实现对用户隐式反馈与词语共现的统一建模,如公式(4)所示。

$$co(c,q_i) = \sum_{j=1}^{n} tf(c,s_j)\, tf(q_i,s_j)\, quality(s_j) \tag{4}$$

其中,

$$quality(s_j) = \begin{cases} 1 & if\ s_j\ is\ clicked \\ \mu \in [0,1) & else \end{cases} \tag{5}$$

$quality(s_j)$ 用来衡量网页摘要 s_j 的质量,本文利用用户点击来估计。如果用户曾经点击过 s_j,$quality(s_j)$ 设为1,否则设为可调系数 μ。μ 约束了词语权值估计时用户反馈的重要性。当 $\mu=1$ 时未点击的网页摘要与点击的网页摘要重要性相同,此时等同于没有考虑用户点击信息,公式(3)即为 $\mu=1$ 时的特殊情况。$\mu=0$ 时表示仅统计候选词语与查询中词语在用户点击过的网页摘要中的共现情况。$0<\mu<1$ 时表示考虑所有网页摘要,但未被点击的网页摘要的重要性按比例衰减,从而突出用户点击过的网页摘要的重要性。因此,结合用户反馈的共现度在衡量词语间共现关系的同时结合用户反馈,估计候选词语的权值时既考虑其与查询在统计上的相关性,又考虑了其与用户偏好的相关性。

3.2.3 扩展词语选择

扩展的词语应与用户输入的整个查询相关而不是仅与其中的某个词语相关。我们在计算候选词语与查询中单个词语的共现度的基础上,度量其与整个查询的相似度,进而根据相似度的大小对候选词语进行排序,从中选择扩展词语。通过公式(6)计算候选词语 c 与查询 Q 的相似性。

$$w(c,Q) = \prod_{q_i \in Q} (\delta + co\_degree(c,q_i) \cdot imp(q_i)) \tag{6}$$

$co\_degree(c,q_i)$ 为 c 与查询中词语 q_i 的共现度。$w(c,Q)$ 通过连乘的方式考察了候选词语 c 与查询中所有词语的共现度进而表现其与整个查询的相似度。其中 δ 设为0.1以避免乘积为零。

查询中可能包含多个词语,这些词语不应等同对待,区分性强的词语应当具有更高的重要性。相应地,与查询中重要的词语共现度大的候选词语也应当被赋予较高的权值,从而进一步减少噪声。这里的 $imp(q_i)$ 代表查询中词语 q_i 的相对重要性。采取类似于逆文档数(IDF)的思想,将 Google 索引的网页视为大规模语料库,利用词语在 Google 的返回结果数(Google Hits)来评价查询中词语的相对重要性,这里假设 Google Hits 小的词语具有更大的相对重要性。设 $ghs(q_i)$ 为词语 q_i 的 Google Hits,$imp(q_i)$ 的计算公式如公式(7)所示。

$$imp(q_i) = \frac{1/ghs(q_i)}{\sum_{q_j \in Q} (1/ghs(q_j))} \tag{7}$$

按照 $w(c,Q)$ 对候选词语进行排序,最终选择前 k 个词语作为扩展词语重构查询模型。

3.3 查询模型重构与重排序

查询模型重构包括两个步骤:确定用来重构查询的词语集合和估计词语权值。经候选词语排序,得到 k 个扩展词语,将其表示为 $C_k=\{c_1,c_2,\cdots,c_k\}$。重构查询模型使用的词语集合为 $Q\cup C_k$。对于 q_i,按照公式(1)进行权值估计。使用 $w(c_i,Q)$ 作为 c_i 的权值。

我们对初始检索结果集合中每个文档的网页摘要(包括标题)按照空间向量模型进行建模,计算每个网页摘要模型与重构的新查询模型间的余弦相似度,并依此对初始检索结果实现重排序。

4 实验设置

4.1 数据集

针对个性化信息检索,开发了基于天网100G语料的个性化评测语料标注辅助系统[14]。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,包括查询、检索结果、用户查看的网页等。针对每个查询,标注者对系统返回的前二十个网页判断是否符合其检索意图,符合标注为相关,否则标记为不相关。

收集了5名用户的标注结果。平均每人进行了230余次检索。从每名用户的历史查询中,按照检索时间由后向前的顺序,选择了一系列查询用于测

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

试。这些查询要求同时满足以下 2 个约束:

1) 至少有 2 个或 2 个以上相关文档。

2) 至少符合用户之前提交的查询表达的兴趣,如:科技、电影等,或者属于某个查询片段(用户为了达到查询目的提交的一系列查询)。

对每一用户,测试查询与查询总数的比例在 10 %～20 %之间。数据集的统计如表 1 所示。

表 1　数据集统计

	用户 1	用户 2	用户 3	用户 4	用户 5
查询总数	218	256	177	206	311
查询平均长度	2.6	2.8	2.6	2.7	2.9
平均相关文档数	4.17	4.22	3.89	4.12	3.24
平均点击次数	2.02	2.83	2.15	2.41	2.02
训练查询数	174	226	151	171	251
测试查询数	43	36	26	35	60
测试所占比例	19.7 %	14.1 %	17.2 %	17.0 %	19.3 %

4.2　评价方法

采取 p @5 和 Normalized Discounted Cumulative Gain (NDCG)作为评价方法,对所有测试查询取平均值来评价系统的表现。其中 p @5 方法表示结果集的前 5 篇文档中相关文档比例。DCG[15]赋予排序高的文档以更高权值并且结合不同的反馈级别(高度相关、相关和不相关),如公式(8)所示。

$$DCG(i) = \begin{cases} G(1) & if\ i = 1 \\ DCG(i-1) + \dfrac{G(i)}{\log(i)} & otherwise \end{cases} \tag{8}$$

本文对于相关文档令 $G(i) = 1$,对于不相关文档令 $G(i) = 0$。NDCG 是通过将 DCG 与理想状况下(所有相关文档排在结果集合的最前面)的 DCG 值(IDCG)做比值获得,其值处于 0、1 之间,越高说明系统表现越好。

4.3　实验结果及分析

实验部分着重比较与分析以下 3 个方面:1) 选取的扩展词语数对系统的影响。2) 用户点击对系统的作用。3) 本文的查询重构方法与以往基于检索历史重构查询的方法的比较。选择的基准系统包括:基于 Lucene 实现的默认检索系统,记为 Default;将用户的检索历史中所有点击过的网页摘要基于 TF-IDF 建模并取质心构建用户模型,记为 Whole;采取 Bin Tan 等[9]的思想实现的对比系统 Bin Tan。Bin-Tan 将每个查询形成历史单元模型,通过计算当前查询与所有历史单元模型的相似度作为插值系数对所有历史单元进行线性插值形成历史模型。最终由当前查询模型和历史模型融合得到重构的查询模型,并利用该模型对结果集合进行重排序。本文基于检索历史上下文分析的个性化查询重构方法记为 PQR。

PQR 需要设置的参数为选取网页摘要的数目 n、扩展词语数目 k 及系数 μ。这里设定 $n = 30$,以保证返回的网页摘要的相关性。分别考察 $k = 10, 20, \cdots, 100$ 及 μ 在 0、1 间不同取值时系统的表现。参数 k 反映了扩展词语的精度,参数 μ 体现了用户点击对系统的影响。

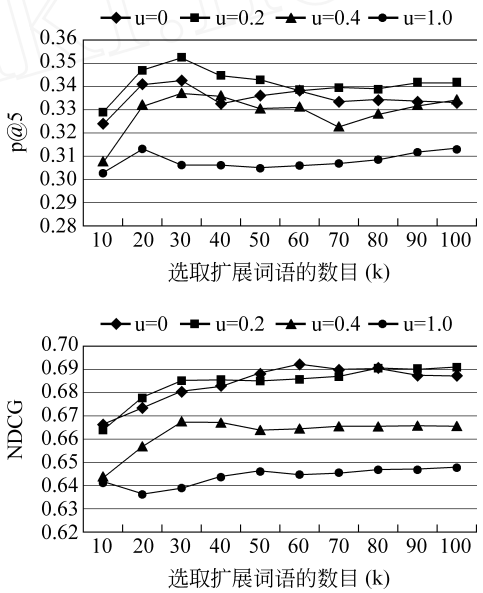

图 2　在不同 k 值及 μ 值下 PQR 的表现

图 2 给出了 μ 取不同值时,PQR 随扩展词语数 k 变化的趋势。当使用 p @5 评价时,系统的表现对 k 的变化较为敏感。在开始阶段随 k 的增大表现提高,当 k 达到 30 左右时系统表现最优,而后随着 k 值增大而下降直至趋于稳定,说明需要小心地设置参数 k 以使相关文档尽可能排到结果集合前列。使用 NDCG 评价时,系统表现随 k 变化相对平稳,说明重构的查询模型能够可靠地提高整个结果集中相关文档的排名。

我们发现当 μ 值较小时,系统的表现更优。如 3.2.2 所述,μ 反映了用户反馈的重要性,当其较小时系统赋予用户点击过的网页摘要相对更高的权重。可见用户点击是可靠且有效的隐式反馈,利用用户点击过的网页摘要可以更好地获得符合用户需

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

求的相关词语。

表2给出了针对2个查询样例,μ分别取0和1时PQR得到的扩展词语中及BinTan历史模型中权值最高的前10个词语。“自然语言发展”是一个具有歧义的查询,既可能指人类语言的演化也可能指计算机科学中的自然语言处理技术。当$\mu=1$(所有网络摘要的重要性相同)时,得到的扩展词语涉及多方面的内容,不能完全准确预测用户意图。当$\mu=0$(仅考虑用户点击过的网页摘要中的词语)时,扩展的词语可以判断用户关注的是自然语言处理、信息检索相关的内容。但对于一些情况,扩展词语过于具体会使用户局限在已有的偏好内。

表2 排序前十位的扩展词语举例

查询:自然语言发展			查询:推荐科幻电影		
μ=0	μ=1	BinTan	μ=0	μ=1	BinTan
计算机	起源	网	Dvd	Dvd	电影
技术	人类	中国	在线	生化	网
搜索	声音	运动	第五元素	科幻片	运动
搜索引擎	问题	技术	主演	旋风	下载
科学	达尔文	健康	侏罗纪	影片	中国
检索	改变	下载	上映	导演	软件
信息	模仿	软件	公园	主演	俱乐部
模型	计算机	能	该部	在线	音乐
智能	科学	电脑	引爆	第五元素	介绍

如查询“推荐科幻电影”,当$\mu=0$时,得到的扩展词语会使系统倾向于将用户过去检索过的对象,如《第五元素》和《侏罗纪公园》等影片相关的网页排在前面。$\mu=1$时,扩展词语中包含了一般化的相关词语,如“科幻片”、“主演”等,这有助于用户找到新的潜在感兴趣的对象。因此设置合适的系数μ,既强调用户点击的重要性又增加一般的相关词语来重构查询是更好的选择。由图2可见,$\mu=0.2$时系统表现较好。BinTan得到的历史模型中,权值最高的词多为在多个主题下经常出现的词,如:“网”,“中国”,“下载”等。说明针对在各种主题中广泛存在的噪声词语,BinTan去噪能力有限。随着用户检索历史增加,此类噪声不断积累,使用户历史模型准确描述用户兴趣的能力逐渐降低。

表3给出了$k=30$,$\mu=0.2$时PQR与基准系统的比较。从中可以发现,将用户的检索历史视为一个整体建模(Whole)时系统的性能反而下降。这说明对于个性化检索任务来说,将检索历史视为整体而不区分其中信息是否与当前查询相关不能有效地利用用户历史信息,提高检索系统的性能。BinTan与Default相比则有明显提升,因为它考虑了历史查询与当前查询的相似性,增加了检索历史中与当前查询相关的历史单元中的词语的权值,对不相关主题中的噪声具有抑制作用。使用p@5进行评价时,PQR相对BinTan提高12.8%,相对于Default提高26%;使用NDCG进行评价时,相对于BinTan提高7.2%,相对于Default提高11.4%。结合图2,在大多数参数条件下,PQR的表现都好于3个基准系统。可见,PQR能够较好处理检索历史中的两类噪声词语,有效地选择相关词语重构查询。适当设置参数时,可以大幅提高满足用户需求的网页的排序,改善用户体验。

表3 PQR与基准系统的比较

	p@5	NDCG
Default	0.28	0.614 4
Whole	0.260 6	0.585 0
BinTan	0.312 8	0.638 3
PQR	0.352 8	0.684 5

5 结论与未来工作

本文针对用户检索历史包含大量与当前查询无关的噪声的问题,将用户的检索历史中的网页摘要视为上下文语境,结合用户点击考察词语在上下文中的共现,选取与整个查询最相关的词语重构查询模型。检索结果重排序的实验表明,在词语选择过程中,用户点击是有效的隐式反馈,对相关词语的选择作用明显。选择与当前查询相关性最高的若干词语重构查询模型比将检索历史视为整体考虑更为合理,可以有效地减少噪声。

本文的方法对用户检索历史规模有一定依赖,利用当前查询检索网页摘要时可能会面临数据稀疏问题,对参数设置也有一定要求。在今后工作中,将局部反馈与用户检索历史相结合以及自适应地确定参数等方面内容是我们需要进一步研究的课题。

参考文献

[1] 曾春,邢春晓,周立柱.个性化服务技术综述[J].软件学报,2002,13(10):1952-1961.

[2] Nicholas J. Belkin. Some (what) challenges and grand

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

challenges for information retrieval[J]. ACM SIGIR Forum, 2008,42(1): 47-54.
[3] Jing Bai, Jian-Yun Nie, Guihong Cao, Hugues Bouchard. Using query contexts in information retrieval [C]// Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007: 15-22.
[4] Xuehua Shen, Bin Tan, ChengXiang, Zhai. Implicit user modeling for personalized search [C]// Proceedings of the 14th ACM international conference on Information and knowledge management. 2005: 824-831.
[5] Yuanhua Lv, Le Sun, Junlin Zhang, Jian-Yun Nie Wan Chen, Wei Zhang. An iterative implicit feedback approach to personalized search [C]// Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. 2006: 585-592.
[6] Sugiyama K, Hatano K, K Yoshikawa M. Adaptive web search based on user profile constructed without any effort from users[C]// Proceedings of the 13th international conference on World Wide Web. 2003: 675-684
[7] Susan Gauch, Jason Chaffee, Alaxander Pretschner. Ontology-based personalized search and browsing[J]. Web Intelligence and Agent Systems. 2003, 1(3-4): 219-234
[8] Teevan, J., Dumais, S. T., & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activites[C]// Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005: 449-456.
[9] Bin Tan, Xuehua Shen, ChengXiang Zhai. Mining long-term search history to improve search accuracy [C]// Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006: 718-723
[10] Lavrenko, V. and Croft, W.B. Relevance-based language models[C]// Proc. 24th ACM SIGIR Conf. On Research and Development in Information Retrieval. 2001: 120-127.
[11] Jinxi Xu, W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis[J]. ACM Transactions on Information Systems (TOIS). 2000, 18(1): 79-112.
[12] Paul Alexandru Chirita, Claudiu S. Firan, Wolfgang Nejdl. Personalized query expansion for the web [C]// Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007: 7-14.
[13] 梅翔, 陈俊亮, 徐萌. 一种基于偏好的查询扩展方法[J]. 高技术通讯, 2007,17: 1142-1146.
[14] 张宇, 范基礼, 郑伟, 邹博伟, 刘挺. 基于人工标注的个性化检索系统评测的研究[J]. 中文信息学报, 2009, 23(2): 62-53.
[15] Kalervo Järvelin, Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents [C]// Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000: 41-48.

(上接第 48 页)

于单 IP 多用户的点击,可以统计是否此时连续点击的比例也较高,是否这两种情况叠加出现,比单独一种情况出现时用户点击的异常特征更加明显。此外,目前的实验只是以一天为周期,然后平均十天的数据得到的结论,文献[8]以一个更长的时间段为周期,这可能成为未来工作的方向。

参考文献

[1] 中国互联网络信息中心. 第 25 次中国互联网络发展状况统计报告[OL]. 2010 年 1 月. 在线资源位置: http://news.sciencenet.cn/upload/news/file/2010/1/2010115164254332.pdf.
[2] 余慧佳,刘奕群,张敏,等. 基于大规模日志分析的网络搜索引擎用户行为研究[C]// 第三届学生计算机语言学研讨会论文集,2006,202-207.
[3] Rongwei Cen, Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma. Study on the Click Context of Web Search Users for Reliability Analysis[C]// The Fifth Asia Information Retrieval Symposium (AIRS) ,2009 年.
[4] 刘奕群,岑荣伟,张敏,等. 基于用户行为分析的搜索引擎自动性能评价[J]. 软件学报, 2008, 19 (11): 3023-3032.
[5] 陈磊,刘奕群,茹立云,等. 基于用户日志挖掘的搜索引擎广告效果分析[J]. 中文信息学报, 2008, 22(6): 92-97.
[6] 余慧佳,刘奕群,张敏,茹立云,等. 基于目的分析的作弊页面分类[J]. 中文信息学报,2009,23(2): 95-101.
[7] 网站分析在中国-从基础到前沿. 那些虚无的网事-人工流量的八大怪现状[OL]. 2008 年 12 月 8 日. 网页地址: http://www.chinawebanalytics.cn/?p=242.
[8] 王继民,彭波. 搜索引擎用户点击行为分析[J]. 情报学报,2006,25(2): 154-162.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net