基于环境信息的移动搜索个性化查询扩展

王忠民 霍艺伟 邓万宇
(西安邮电大学计算机学院 西安 710121)

摘 要 与传统搜索相比,移动搜索对位置、温度、速度等环境信息更为敏感。为了有效利用环境信息推断用户查询意图,提出了一种基于环境信息的查询扩展方法并应用在移动搜索系统Clever Search Engine(CSE)中。该方法利用专家系统对分词后的查询词和收集到的用户环境信息进行推理和融合,扩展查询词,实现个性化搜索。实验证明,基于环境信息的移动搜索个性化查询扩展方法能有效改善移动用户的搜索体验,比现有的公共搜索引擎(如Google)具有更高的查准率。

关键词 查询扩展,移动搜索,专家系统,个性化,算法
中图法分类号 TP391.3 **文献标识码** A

Personalized Query Expansion Based on Environment Information for Mobile Search

WANG Zhong-min HUO Yi-wei DENG Wan-yu
(School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China)

Abstract Compared with traditional search, mobile search is sensitive to the users' environment information such as location, temperature, speed and so on. In order to use the environment information to infer the user intents, an expansion approach based on environment information was proposed, which has been applied in the Clever Search Engine (CSE)——a mobile search system we developed. Using segmented queries and users' environment information, the approach infers the user intents through the expert system and expands query in order to realize personalized query expansion. Experimental results show that the approach we proposed can significantly improve users' search experiences and has better performance in precision than existing public search engine(e. g. Google).

Keywords Query expansion, Mobile search, Expert system, Personalization, Algorithm

1 引言

随着移动互联网的快速发展以及移动智能终端的日益普及,移动搜索近年来受到了学术界和产业界的广泛关注。和PC机相比,移动设备有屏幕尺寸较小、键盘输入不便等缺陷,使得用户在移动搜索时提交的查询词往往更简短更模糊,但对搜索结果的查准率却有更高的要求。这种情况下普适性的搜索往往无法很好满足用户需求。和传统的桌面搜索不同,移动搜索通常与用户所在的环境信息密切相关,如何利用现有智能移动终端上配备的各种传感器所提供的诸如温度、加速度等环境信息以及GPS提供的位置信息,根据用户输入的查询词结合位置及环境信息更好地理解用户意图,已成为目前研究的热点。

很多时候用户往往对自己想要搜索的话题并不了解,输入的查询词并不能精确表达其搜索意图,难以构造出很好的查询。查询扩展就是根据用户输入的查询词,结合用户所处的环境信息、位置信息以及用户的兴趣模型、历史记录及其所处社交网络等相关信息,生成和用户查询相关的新查询词。

传统的查询扩展方法主要是基于词汇间关系的,以不同方式从语料库中获得和查询词关系密切的词来扩展。如文献[1]利用从用户个人信息库中搜集的术语来扩展查询;文献[2]中提出了一种基于语言模型下相关反馈的查询扩展方法;文献[3]利用伪相关性反馈得到候选扩展词集合。但是,这些方法并不能在移动搜索中充分发挥作用,因为在移动搜索中,还有许多随用户位置改变的环境信息,如何利用这些位置信息及该位置的环境信息,结合用户输入的查询词进行查询扩展,以便更好地理解用户意图是实现移动搜索个性化推荐的关键技术。

近年来,也有一些基于环境信息的移动搜索查询扩展方面的研究。文献[4]采用结果加权和的方法,把用户的位置、个性化信息和个人偏好等在内的上下文信息融合到查询词中,以达到查询扩展的目的。文献[5]中提出了上下文剖面的概念,并将其分为用户剖面、设备剖面、环境剖面和数据剖面4种,然后通过对剖面中的上下文实体赋予不同的权重来扩展查询词。文献[6]中构造概念向量来描述用户的上下文,并以此为依据,利用语义关系选择查询扩展词。这些工作从不

到稿日期:2012-11-17 返修日期:2013-03-14 本文受国家自然科学基金项目(61100166),陕西省工业攻关项目(2011K06),工业和信息化部通信软科学研究计划项目(2012-R-41)资助。

王忠民(1967—),男,博士,教授,主要研究方向为智能信息处理;霍艺伟(1987—),女,硕士生,主要研究方向为移动信息检索;邓万宇(1979—),男,博士,副教授,主要研究方向为个性化推荐。

同的角度利用环境信息对用户查询意图进行推测，取得了一定的进展。本文提出了一种基于环境信息的移动搜索个性化查询扩展方法。该方法首先对环境信息和查询词进行预处理，然后将处理后的环境信息和分词后的查询词送入专家系统，与规则库中的规则进行匹配，最后根据匹配上的规则为查询词追加适当的辅助关键字，从而达到理解用户意图、实现查询扩展的目的。

本文第 2 节详细说明了基于专家系统的移动搜索查询扩展方法涉及的关键问题，并给出了详细的算法描述；第 3 节给出实验描述，并对实验结果进行了分析；实验结果表明，该算法比现有的公共搜索引擎具有更好的查准率，能够有效改善移动用户的搜索体验。

2 基于环境信息的查询扩展

查询扩展主要指在用户的查询词被提交到搜索引擎前，根据移动用户所处的位置、环境信息、历史记录以及兴趣模型等，采用各种算法选出合适的新词对用户查询进行扩展，从而使用户的查询意图更加明确。本文提出一种基于专家系统的查询扩展方法，其具体结构如图 1 所示。

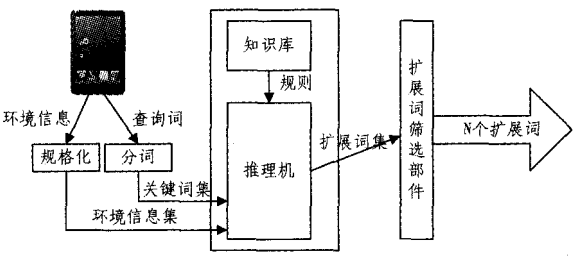

图 1　ES-QEA 结构图

2.1 预处理

环境信息是用户上下文中最多变的部分，它最能反映移动性对用户搜索意图的影响。直接提取的环境信息具有多样性，不利于进行推理运算，所以要将其规格化为方便进行匹配运算的形式。

用户环境信息可以用一个三元组表示：

$e=(ID,Value,Mode)$

其中，ID 是每种环境信息的唯一标识；$Value$ 是该环境信息的取值，根据需要将其量化为数值型，$Mode$ 表示该环境信息的获得方式，1 表示直接获取，0 表示推理得到。环境信息的获取方式决定了该环境信息的精确度，可以将其用作规则匹配时匹配度的一个影响因素。采集到的用户环境信息组成用户环境信息集为：

$E=(e_1,e_2,\cdots,e_n)$

为了完成匹配，专家系统的规则中也需要有相应的环境信息，这里称为条件环境信息。条件环境信息可以用一个四元组表示：

$e'=(ID,Range,Weight,Coefficient)$

这里 ID 是每种环境的唯一标识，和 e 中的相对应；$Range$ 是该环境条件成立时相应环境信息取值的范围；$Weight$ 是该环境信息在规则中所占的权重；$Coefficient$ 为权重系数，$Coefficient\in(0,1]$，当用户环境信息的获取方式为推理得到时，需要乘以该系数，如果 $Coefficient$ 的值为 1，则表示忽略用户环境信息的获取方式。每条规则都有自己的条件环境信息集 $E'=(e_1',e_2',\cdots,e_n')$。

待扩展关键词也用一个三元组表示：

$q=(Key,Field,Type)$

$Type$ 用来区分关键词的来源，true 表示该关键词来自查询词，是查询词分词所得的查询关键词；false 表示该关键词来自专家系统的输出，是扩展所得的辅助关键词。辅助关键词不需要再次被扩展。Key 保存该查询关键词，$Field$ 标识该查询关键词所属的领域。领域标识来源于移动搜索系统 CES(Clever Search Engine)中自建词库的分类。词库中的词按领域分为 12 个一级类，每个一级类包含若干个二级类，部分二级类含有自己的三级类。分词后得到的关键字将会被标识上领域类别号(如果不是一级类，则会同时标识自己的领域类和所属的父领域类)，这样便于在专家系统中快速查找到可能匹配的规则。查询词分词后被表示为一个搜索关键词集 $Q=(q_1,q_2,\cdots,q_m)$。

2.2 知识库

知识库用于存储推理时所用的规则，规则采用产生式表示法。这种表示方法使规则具有相同的格式，并且全局数据库可被所有的规则访问，便于规则统一处理。知识库中的各规则间只能通过全局数据库发生联系，不能直接相互调用，有利于知识的修改和扩充。知识库中规则的具体表示如下：

$$\text{IF } q_i.Field=r_j.Field \text{ AND } Similar>T \text{ AND } ex_j\cap Q=\Phi \text{ Then } q_i=q_i+ex_j \text{ AND } Q=Q\cup ex_j \quad (1)$$

其中，Q 为查询词分词后得到的关键字集，q_i 为 Q 中的第 i 个关键字。r_j 为规则集中的第 j 条规则。$Similar$ 是用户环境信息和规则的条件信息的相似度。T 为相似度的自定义阈值。ex_j 是执行规则 j 后为提交的关键字扩展的辅助关键字，即专家系统的输出。

专家系统知识库中的一条具体规则可以用一个四元组表示：

$r=(ID,Field,E',Ex)$

ID 是规则的唯一标识。$Field$ 是规则适用的领域，它的取值范围与待扩展关键词的 $Field$ 域的一致，但在标识过程中，规则的 $Field$ 域只需要标识它自身所属的领域类，而不需要标识其所属的父领域类。E' 是条件环境信息的集合，其中每个元素都是一个三元组 e'。Ex 是该规则输出的辅助关键字。

以大量移动用户搜索行为模式的研究为基础，综合用户搜索日志的挖掘，我们为知识库设计了 72 条初始规则，之后经过实验筛选优化后得到 23 条规则。这些规则覆盖了与用户搜索意图相关性较大的 6 种环境信息，即位置、日期、速度、温度、季节、交通方式。

2.3 推理机

推理机的核心是规则匹配，这个过程的关键步骤是计算专家系统的输入信息(查询关键字和用户搜索环境信息)和规则的匹配度，即输入信息和规则前件的相似度。为了能够计算确切的相似度，首先需要对环境信息进行抽象和量化，得到用户环境集合 E，其中每个元素都是一个三元组 e。依次比较用户环境集合 E 和规则的条件环境集合 E' 中的对应元素，用浮点型变量 $Similar$ 来记录匹配成功的环境元素的权重和。对于 $e_k.ID=e_k'.ID$ 的一组元素，如果 $e_k.Value$ 的值在 $e_k'.Range$ 表示的范围内，则认为该组环境元素匹配成功，根据 $e_k.Mode$ 的值将 $e_k'.Weight$ 乘以系数 $e_k'.Coefficient$ 后累加

进 *Similar*。*Similar* 的计算公式如下：

$$Similar=\sum_{k=1,e.Mode=1}^{n} e_k'.Weight+\sum_{k=1,e.Mode=0}^{n} e_k'.Coefficient\cdot e_k'Weight \quad (2)$$

2.4 扩展词筛选

由专家系统输出的扩展词可能不止一个，而过多的扩展词往往会影响查询的准确性。引入扩展词筛选部件主要就是为了解决这一问题。它主要利用扩展词和原查询词的共现率，并由扩展过程中计算得到的 *Similar* 值加以辅助，来筛选扩展词。

首先需要构建共现词词典，其中记录两个词的共现频率 F。计算 F 采用的方法是对文献[7]中提出的 FCD 算法的简化，计算公式如下：

$$F_{a,b}=\sum_{d_i\in D,i=1}^{n} f_{ad_i}\cdot f_{bd_i}\cdot \min(r_{db})_{d_i} \quad (3)$$

其中，D 为一个含有 n 个文档的语料库，$d_i\in D$。f_{ad_i} 为词汇 a 在文档 d_i 出现的频率。f_{bd_i} 为词汇 b 在文档 d_i 中出现的频率。$\min(r_{db})_{d_i}$ 为词汇 a 和 b 在文档 d_i 中的最小距离。设置共现频度的阈值 F_T，当 $F>F_T$ 时将词汇对存入共现词词典。

计算每个扩展词 ex，计算它和整个查询的共现频率，公式如下：

$$F_{ex,Q}=\sum_{q_i\in Q,i=1}^{n} F_{ex,q_i} \quad (4)$$

对 $F_{ex,Q}$ 以降序排序，如果出现 $F_{ex,Q}$ 相同的情况，则用该扩展词的 *Similar* 值作为辅助排序条件。取前 N（可以根据实际情况定义）个作为筛选后的扩展词。

2.5 ES-QEA 算法基本工作流程

首先，该算法对采集到的用户环境信息进行分类和量化，得到可用的用户环境信息集 E。然后将 E 和搜索关键词集 Q 一并送入专家系统进行推理。遍历 Q，对其中的每个搜索关键词查找与 q_i 匹配的规则。对特定的 q_i，比较其 *Field* 域和规则的 *Field* 域，如果两个值相等，则开始计算匹配度，否则比较下一条规则。匹配度的计算主要是计算用户环境集 E 和规则的条件环境集 E' 的相似度 *Similar*。如果 *Similar* 大于预先设定的阈值 T，则认为规则匹配成功。比较该规则输出的辅助关键词 Ex，如果 Ex 不存在于 Q 中，则将 Ex 输出，添加到 Q 中，该词辅助关键词的 *Type* 域记为 false，以便在以后的比较中跳过。接着比较下一个关键词 q_{i+1}。重复该过程直到 Q 中所有的搜索关键词都被扩展。

ES-QEA 算法的具体描述如下：

```
1. For each query q_i
2.   IF(q_i. Type == true)
     {
3.   Find rule r whose Field is the same as q_i
4.   Calculate the value of Similar by the E corresponding to q_i and the
     E' contained by r
6.   IF(Similar > T)
7.   Q=QUex_j
   }
```

3 实验结果及评价

本文提出的算法已经在移动搜索系统 CSE 上实现。为了评价该算法的有效性，我们屏蔽了 CSE 的搜索结果重排序功能，所以，在实验中由 CSE 返回的结果是仅经过查询扩展后搜索的结果。为了进一步证明该算法对搜索效果的改善，我们选择 Google 作为一个比较用的搜索系统。

3.1 实验步骤

有 20 位参与者参与了实验。首先参与者同时向 Google 和 CSE(已经屏蔽掉了结果重排序功能)提交查询。然后由查询提交者分别对两个搜索引擎返回的结果的前 10 条进行相关度打分。这里的相关度分为 3 个级别，1 表示高相关，0.5 表示相关，0 表示不相关。最后将分别打过分的来自两个搜索引擎的共计 20 个结果并入实验样本数据集中。实验中我们要求每个参与者至少提交 15 个查询。

最终从这些参与者生成的查询样本数据中选取了环境信息敏感的查询结果共计 5600 个作为评估数据集。这里数据的选择以查询为单位，即如果查询的一个结果被选中，则该查询的其他结果也被选中。事实上，实验只针对环境信息敏感的查询，因为对于非环境信息敏感的查询，CES 将会得到与 Google 搜索引擎返回的相似的结果，而这种查询中收集的数据对比较而言没有价值。

3.2 实验结果

参与评估的 5600 个搜索结果中，相关的有 3104 个，大约占总实验数据集的 56.1%，图 2 和图 3 分别显示了相关记录和无关记录的比较。参与者评价的相关记录中有 65.6%来自于 CSE，34.4%来自于 Google。被参与者认为无关的记录中，Google 返回的占到 69.4%，而 CSE 返回的仅占 30.6%。显然，CSE 比 Google 得到了更多的相关结果。

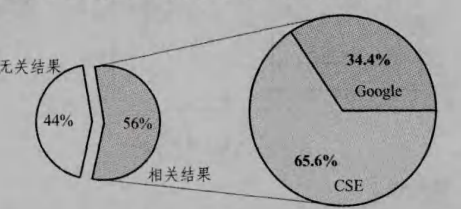

图 2 CSE 与 Google 相关结果对比

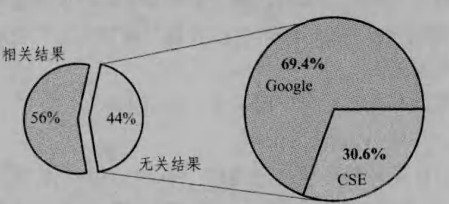

图 3 CSE 与 Google 无关结果对比

在 CSE 的查询扩展中，涉及单一环境信息的结果约占 26.4%，涉及多环境信息的约占 73.6%。各种环境信息的覆盖率如图 4 所示。其中位置信息的查询约占 63.6%，是规则中最常用到的环境信息。

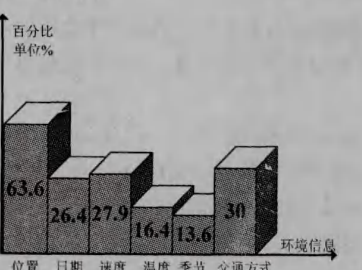

图 4 CSE 查询结果覆盖分布图

结束语 本文详细介绍了基于环境信息的移动搜索个性

(下转第 189 页)

的嵌套 Web Service 事务模型，并给出了其正确性描述。该模型具有很多优点，很适合开放的 Web 环境。

(1)Web Service 事务是长事务，不具有严格的原子性，整个 Web Service 事务不能等到整个事务结束才提交。该模型允许采用分阶段进行提交，在不影响业务逻辑和保证资源约束的前提下，优先提交已完成的子事务，提高了整个事务的执行效率。

(2)采用先替代后补偿的方法，当某阶段中子事务出现失败，首先考虑该子事务的替代事务，若替代事务可以成功执行，那么该阶段子事务可以顺利提交；否则，启动补偿事务对整个 Web Service 事务的失败点之前的各个阶段的子事务进行补偿，保证了整个系统的数据的完整性和一致性。

(3)为高性能地处理 Web Service 事务、灵活地维护 Web 数据的一致性以及并发地控制 Web Service 事务奠定了基础。

参 考 文 献

[1] 岳昆，王晓玲，周傲英. Web 服务核心支撑技术：研究综述[J]. 软件学报，2004，15(3)：428-442

[2] 王剑辉，吴永明. 基于细胞膜模型的 Web Service 事务管理[J]. 计算机应用与软件，2007，24(1)：87-89

[3] 郭玉彬，奚建清. Web service 的事务协调框架研究与实现[J]. 计算机工程与应用，2009，45(36)：22-25

[4] Papazoglou M P. Web 服务：原理和技术[M]. 龚玲，张云涛，译. 北京：机械工业出版社，2009：5-15

[5] Liu Cheng-fei，Zhao Xiao-hui. Towards flexible compensation for business transactions in Web service environment [J]. Service Oriented Computing and Applications，2008，2：79-91

[6] 许峰，徐碧云，黄皓，等. Web 服务事务中的补偿机制研究与实现[J]. 计算机科学，2006，33(7)：242-244

[7] 张晓雯，黄永忠，李占峻. 基于 Web Service 的工作流补偿机制[J]. 计算机工程，2009，35(24)：99-102

[8] 唐飞龙，李明禄，曹健. 一个 Web 服务事务处理模型：结构、算法和事务补偿[J]. 电子学报，2003，31(12A)：2074-2078

[9] Alrifai M，Dolog P，Balke W-T，et al. Distributed Management of Concurrent Web Service Transactions [J]. IEEE Transactions on Services Computing，2009，2(4)：289-302

[10] 申德荣，于戈，张蓉. Web 服务合成中的异构问题[J]. 东北大学学报，2004，25(3)：220-222

(上接第 162 页)

[11] Zhang Yuan-fang，Gill C，Lu Chen-yang. Real-time Performance and Middleware for Multiprocessor and Multicore Linux Platforms[C]//15th IEEE Internaltional Conference on Embedded and Real-time Computing Systems and Applications. Beijing，China，2009：437-446

[12] Baruah S. An improved global EDF schedulability test for uniform multiprocessors[C]//16th IEEE Internaltional Conference on Embedded and Real-Time Computing Systems and Applications. Macau SAR，China，2010：184-192

[13] Li Ning，Kinebuchi Y，Nakajima T. Enhancing Security of Embedded Linux on a Multi-core Processor[C]//17th IEEE Internaltional Conference on Embedded and Real-time Computing Systems and Applications. Toyama，Japan，2011：117-121

[14] Lin T-H，Kinebuchi Y，Shimada H，et al. Hardware-assisted Reliability Enhancement for Embedded Multi-core Virtualization Design[C]//17th IEEE Internaltional Conference on Embedded and Real-time Computing Systems and Applications. Toyama，Japan，2011：101-105

[15] Yoon M-K，Kim J-E，Sha Lui. Optimizing Tunable WCET with Shared Resource Allocation and Arbitrationin Hard Real-time Multicore Systems[C]//32th IEEE Real-Time Systems Symposium. Vienna，Austra，2011：227-238

[16] Nógueira A，Calha M. Predictability and efficiency in contemporary Hard RTOS for multiprocessor systems[C]//17th IEEE Internaltional Conference on Embedded and Real-time Computing Systems and Applications. Toyama，Japan，2011：3-8

(上接第 184 页)

化查询扩展算法，并在 CSE 中实现了该方法。研究的关键在于规则库中规则的形式、匹配的方式和整个扩展算法的设计。最后利用实验证明了该扩展方法比 Google 具有更好的查准率。

这里的扩展仅考虑了环境信息，为了更好地理解用户意图，生成符合用户需求的查询，未来的工作将在基于环境信息的查询扩展研究的基础上，进一步融合用户的其他上下文信息，如历史记录、用户浏览偏好以及用户社交网络等的查询扩展与查询推荐技术。

参 考 文 献

[1] Chirita P-A，Firan C S，Nejdl W. Personalized Query Expansion for the Web[C]//SIGIR 2007 Proceeding. 2007：7-14

[2] 吕碧波，赵军. 基于相关文档池建模的查询扩展[J]. 中文信息学报，2006，20(3)：78-83

[3] 王秉卿，张奇，吴立德，等. 机器学习的查询扩展在博客检索中的应用[J]. 中文信息学报，2008，22(6)：98-102，109

[4] Anderson N. Putting Search in Context：Using Dynamically-Weighted Information Fusion to Improve Search Results [C]// Eighth International Conference on Information Technology. New Generations，2011：66-71

[5] Gui Feng，Adjouadi M，Rishe N. A Contextualized and Personalized Approach for Mobile Search[C]//International Conference on Advanced Information Networking and Applications Workshops. 2009：966-971

[6] Ahmadian N，Nematbakhsh M A，Vahdat-Nejad H. A Context Aware Approach to Semantic Query Expansion[C]// International Conference on Innovations in Information Technology. 2011：57-60

[7] 陈翀，彭波，闫宏飞，等. 一种词汇共现算法及共现词对检索系统排序的影响[J]. 清华大学学报：自然科学版，2005，45(S1)：1857-1860