企业搜索引擎个性化排序方法①

文必龙, 张 璇, 赵晶浩, 赵 满

(东北石油大学 计算机与信息技术学院, 大庆 163318)

摘 要: 针对传统搜索引擎“面向检索”而非“面向用户”的缺点, 将个性化服务思想引入到企业搜索引擎排序中, 对其关键技术即用户兴趣建模进行了研究, 将模型用于查询扩展及排序中, 并为企业搜索引擎设计基于用户兴趣的个性化排序方法, 能为不同用户的同一检索请求提供不同的检索结果列表. 通过将研究用于油田企业搜索引擎的实验证明, 本研究能有效地提高企业搜索引擎检索精确度及满足用户的个性化检索需求, 并具有较好的自适应能力.

关键词: 个性化排序; 用户兴趣模型; 企业搜索引擎; 查询扩展; VSM

Personalized Ranking in Enterprise Search Engine

WEN Bi-Long, ZHANG Xuan, ZHAO Jing-Hao, ZHAO Man

(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: Traditional search engine is retrieval-oriented rather than user-oriented, in view of this weakness, the paper introduces the idea of personalized service to the enterprise search engine ranking. The user interest modeling technology which is the core of personalized service was researched in the paper. We apply the user interest model to query expansion and ranking and designed the personalized ranking method for the enterprise search engine. So the engine can display different result of the same retrieval for different users. We apply the personalized ranking to enterprise search engine in oilfield enterprise search and the tests show that, the research not only can effectively enhances the retrieval accuracy of the enterprise search engine and satisfies the user's personalized retrieval demand, but also has good adaptability.

Key words: personalized ranking; user interest model; enterprise search engine;query expansion; VSM

21 世纪的人类社会是一个信息化的社会, 无论是网络上、各政府机关, 还是各企事业单位内部都保存着大量的资料, 各种信息内容和数量都呈现爆炸式增长. 通用搜索引擎和一般的企业搜索引擎, 例如 Baidu、Google、Excite、Oracle SES 等已比较成功的解决了信息检索问题. 但是, 这些传统搜索引擎是“面向检索”的, 其局限性主要体现在:

1) 企业信息数量庞大、数据类型众多, 而传统搜索引擎对所有用户提供相同的界面和服务, 面对动辄返回的成千上万、良莠不齐的检索结果, 用户越来越难找到最需要的信息, 面临着“信息过载”和“资源迷向”问题[1].

2) 传统搜索引擎对用户不具有自适应能力, 不同用户的同一检索请求得到的是完全相同的检索结果. 而企业用户使用搜索引擎的目的性比一般用户更强, 对排序的期望值更高, 对同一检索请求希望根据用户的职业背景、兴趣爱好得到不同的检索结果, 即需要搜索引擎具有“面向用户”的性质.

3) 传统搜索引擎不具备良好的交互性和即时性, 用户在不同时期或阶段的同一检索请求只能得到相同的检索结果[2]. 它不能跟踪、分析用户浏览行为, 不能感知用户兴趣的变化, 也就不能针对不同用户提供个

① 基金项目:国家科技重大专项(2011ZX05023-005-012);黑龙江省教育厅科学技术研究项目(11551018)

收稿时间:2012-09-19;收到修改稿时间:2012-10-16

性化重排序的检索结果.

正是在这种需求驱动下, 搜索引擎排序技术与个性化服务技术得到了长足发展, 成为当前信息服务领域的研究热点之一. 本文针对传统搜索引擎“面向检索”而非“面向用户”的缺点, 提出一种基于用户兴趣的个性化排序的企业搜索引擎系统架构; 通过对个性化服务的关键技术即用户兴趣建模技术进行研究, 获取用户兴趣、建立用户兴趣模型, 并采用机器学习方式更新模型; 将模型用于查询扩展和排序中, 并为企业搜索引擎设计基于用户兴趣的个性化排序方法, 为用户提供更为满意的个性化检索结果, 这是本研究的目的、内容和意义.

1 个性化排序的搜索引擎体系架构

本文提出的个性化排序的企业搜索引擎体系架构图如图 1 所示.

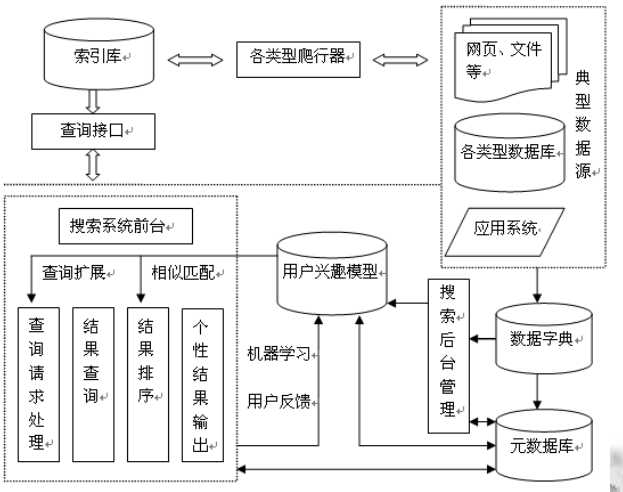

图 1 个性化排序的搜索引擎体系架构

本系统特色是具有基于用户兴趣的个性化排序功能, 可为不同用户的同一检索请求提供不同的检索结果列表, 并支持企业具有代表性的、典型的数据源类型的数据检索, 包括门户网站、内容管理系统、数据库、项目文件服务器、商务办公系统等应用系统.

1) 与传统架构的异同: 与传统搜索引擎结构相同的是包括爬行器、索引器、检索器、后台管理等模块不同之处是包含用户兴趣模型、各种类型数据源对应的数据字典、存储搜索数据源信息的元数据库.

2) 用户兴趣模型: 是本系统个性化排序的基础. 系统根据 VSM 理论建立用户兴趣模型. 在用户使用过程中, 通过机器自学习机制分析用户的浏览行为、搜索日志以及反馈信息, 自动更新用户兴趣模型[3].

3) 查询扩展: 系统根据用户兴趣模型, 自动调整用户检索表达式, 优化扩展查询请求.

4) 结果查询: 本文研究的是基于全文检索技术实现结果搜索的引擎, 并且是基于索引的搜索算法实现技术. 从全文检索的原理可知, 任何不基于索引的搜索效率都非常低. 因此, 基于索引的搜索算法可以保证搜索的效率.

5) 基于用户兴趣的结果排序: 这是区别于其它搜索引擎排序的关键部分. 企业信息搜索排序要比传统搜索引擎一般的页面排序更为复杂, 需要考虑网页结果、数据库结果和其它文档结果如何进行排序. 系统按照特定算法计算搜索结果的排序得分值, 调整融合各结果集得到个性化的排序结果.

6) 个性化结果列表输出: 输出排序好的结果列表, 尽量满足用户的个性化需求.

2 用户兴趣模型研究

用户兴趣模型又称用户模型或个性化模型, 常被理解为对用户在某段时间内相对稳定的信息需求的规范化的数学描述. 它具有特定的数据结构, 是一种对用户兴趣偏好形式化的、面向算法的数学描述. 用户兴趣建模是指从能够体现用户兴趣偏好的信息如浏览行为、浏览内容、职业背景等信息中提取可计算的用户兴趣模型的过程[4]. 值得指出的是企业信息搜索用户的个性化的信息需求是相对稳定的、时间相对长久的信息需求, 因此可以对其建立兴趣模型. 用户兴趣模型是搜索引擎实现个性化排序的基础和关键, 只有建立了高质量的用户兴趣模型, 才能保证搜索引擎个性化排序工作更好的进行.

2.1 用户兴趣的获取方式

通常用户兴趣的获取方式分两种. 第一种是通过用户的显式反馈收集相关信息. 这种方式主要是让用户主动提交兴趣信息, 如自动提交个人工作内容关键词、兴趣偏好等信息; 它的优点是可靠性高, 缺点是占用用户时间可能导致用户反感. 第二种是通过隐式反馈方式收集用户信息. 这种方式主要是通过跟踪记录用户的搜索信息、浏览行为和内容来挖掘用户的兴趣; 其优点是无需用户主动参与, 缺点是可靠性不够高.

因此, 本文采用将两种方式结合, 通过显示方式获取用户的静态信息, 通过隐式方式获取动态信息. 用户兴趣建模的主要信息来源包括:

1) 用户使用搜索引擎时, 输入过的查询词;
2) 用户浏览过的页面;
3) 用户的浏览行为, 如点击、翻页等;
4) 用户保存或下载的网页及其它文档资料;
5) 记录在服务器日志中的信息;
6) 用户填写的兴趣信息及其他手工输入信息;

本系统采用机器学习的隐式方式获取动态用户兴趣信息. 智能系统不断的积累经验改善系统性能的过程就是机器学习. 通过机器学习的方式获取用户兴趣模型可以减少用户干预并增强用户适应性, 同时适合定期动态更新[5].

2.2 用户兴趣模型的表示与建立

用户兴趣模型的表示决定了用户模型反映用户真实信息的能力和可计算能力. 常见的表示方法有: 主题表示法、用户 Bookmark 表示法、关键词列表表示法、向量空间模型表示法、基于本体的表示法、兴趣粒度表示法. 其中, 向量空间模型(Vector Space Model, 简称 VSM)表示法的基本原理是把每个文档和查询各自映射为高维空间中的一个向量, 两个向量间的夹角越小则文档和查询的相似度越高, 文档越接近查询要求[6]. 它是目前改进搜索引擎执行效率、提高个性化排序质量最有效的数学工具之一. 因此, 本文采用 VSM 表示法建立企业用户兴趣模型.

在VSM中选取词作为特征项, 则当文档或查询可被看作是由它所包含的字、词、短语等构成的集合时, 这些基本语言单位统称为项[7]. 假定存在一个项的集合, 于是文档和查询均可用由项构成的向量来表示. 则文档 D 可表示为: $D=(t_1, t_2, \cdots, t_n)$. 其中项 t_i 常被赋予一个数值 w_i, 表示它在文档中的重要程度, 称为项 t_i 的权重. 因此, 我们一般用 $D=(w_1, w_2, \cdots, w_n)$的形式表示文档, 并称其为文档 D 的向量表示. 则用户模型 U 的向量表示为 $U=(D_1, D_2, \cdots, D_n)$. 同理 $Q=(w_1, w_2, \cdots, w_k)$为查询Q的向量表示. 其中, 特征项的权重采用 TF-IDF 公式[8]计算, 公式如下:

$$Wd_i= TF_i * IDF_i = \frac{Frq_i}{Max_1 Frq_1} * \log \frac{N}{n_i} \quad (1)$$

其中, Frq_i 表示项 t_i 在文档中出现的次数, n_i 表示 t_i 出现的文档的次数, N 表示文档集合中项的总数. 而查询 Q 的权重表示为:

$$Wq_i=\begin{cases}1, ti \in 查询条件 \\ 0, ti \notin 查询条件\end{cases} \quad 1 \le i \le n \quad (2)$$

则文档与查询之间的相似度可以用其对应的向量间的夹角余弦来表示, 即

$$SIM(D,Q)=\cos\theta = \frac{\sum_{i=1}^{n} Wd_i * Wq_i}{\sqrt{\sum_{i=1}^{n} Wd_i^2 * \sum_{i=1}^{n} Wq_i^2}} \quad (3)$$

2.3 用户兴趣模型智能更新

一般来说, 用户的兴趣在一定时间内具有相对稳定性, 但不是一成不变, 系统应根据需求更新用户模型. 本系统为提高用户自适应能力采用机器学习方式对用户模型进行智能更新, 主要分以下三种:

1) 若通过机器学习获得用户模型之外的新关键词, 则计算相应权值, 扩展用户模型.

2) 若通过机器学习获得用户模型中已有的关键词, 则调整模型中对应关键词的权值.

3) 若用户模型中存在关键词条超过规定的最大容量, 则删除低权值词条. 由于用户模型空间有限, 所以当词汇量超过规定的最大容量值时, 应当删除一些低权值词条, 使词汇量固定在某一范围内, 以实现对用户兴趣的动态追踪.

3 基于用户兴趣的个性化排序方法研究

3.1 查询优化扩展

查询优化扩展是在原查询词的基础上加入与当前用户相关的词或者词组, 组成新的、更准确的查询词序列. 它可以在一定程度上弥补用户查询信息不足的缺陷, 是查询处理的一部分, 为个性化搜索结果排序做好早期准备工作, 步骤如下:

1) 系统提取用户的查询请求 Q , 从 Q 中提取出各个查询关键词项 k_i 组成的关键词集合 K_i;

2) 在用户模型中, 查找出与 K_i 集合中某项相关的关键词集合 T_i;

3) 将 T_i 作为扩展的查询关键字(当 T_i 为空集时, 说明查询未得到有效扩展);

4) 用户输入的初始查询关键字一般最能反映查询要求, 因此将初始查询关键字的权重均设为 1, 而扩

展的查询关键字的权重均设为 0.8, 防止查询偏移;

5) 经优化后的查询 Q′ =Q∪T_i, 权重向量为:

$$Q' =(W_{1,q},W_{2,q},\cdots,W_{m,q},W_{1,t},W_{2,t},\cdots,W_{m,t}) \quad (4)$$

其中, $W_{1,q}=W_{2,q}=\cdots=W_{m,q}=1$; $W_{1,t}=W_{2,t}=\cdots=W_{m,t}=0.8$.

3.2 基于用户兴趣的个性化排序方法

综合以上研究所述, 本文提出基于用户兴趣模型的个性化排序方法思想如下:

1)企业信息搜索结果中存在特殊的数据库类型结果, 本研究规定将数据库中属于同一张表的单条记录结果使用企业数据字典进行语义解释后合为一条表结果. 考虑到企业用户若选择的查询范围包含数据库, 则证明其得到数据库结果意图明显. 因此决定对其进行优先排序, 即将此类型数据源优先提交给搜索接口, 并将搜索结果集 R_1 放在最终结果集 R 的最前面.

2)对于其它数据类型的结果, 将经过搜索引擎常规初步排序后的结果集 R_2 结合用户兴趣进行重排序, 计算公式如下:

$$RScore=Score+\alpha\ UScore \quad (5)$$

其中, RScore 为最终排序得分值, Score 为搜索引擎常规初步排序得分值, UScore 为当前某条结果经相似度评分后的得分值, 所以 UScore= SIM(R,U) *Score; 而 α 表示 UScore 所占比重值. SIM(R,U) 表示查询结果文档 D 与相关的用户兴趣文档 U 之间的相似度. 公式(5)演变为:

$$RScore=Score+\alpha\ SIM(R,U)Score \quad (6)$$

重排序过程如下:

①系统中用户兴趣模型与系统初步检索出的文档均采用 VSM 表示. 系统用查询 Q′ 得到初步检索结果集 P.

②在用户模型 U 中取权值高的 n 个文档, 然后与 P 中的每个结果逐个求相似度, 参考公式(3)得到:

$$SIM(P,U)=\frac{\sum_{i=1}^{n} W_{i,j} * W_{i,p}}{\sqrt{\sum_{i=1}^{n} W_{i,j}^2 * \sum_{i=1}^{n} W_{i,p}^2}} \quad (7)$$

其中, $W_{i,j}$ 表示文档 D_j 中词条 t_i 的权重, $W_{i,p}$ 表示词条 t_i 在用户兴趣向量中的权重.

③求相似度的平均值作为最终相似度:

$$SIM(R,U)=\sum_{i=1}^{n} SIM(P,U) \Big/ n \quad (8)$$

然而, 检索结果文档集合数量庞大, 若将结果集合中的每个文档都与用户模型文档求相似度则会严重影响系统检索效率. 由于用户一般只访问检索结果中排在前几页的结果, 所以我们只对结果集中一定数量的结果计算相似度, 例如取结果集中前 500 个文档与用户模型计算相似度.

④按照公式(7)计算结果集 R_2 中每个结果的得分值, 然后按最终得分值 RScore 由大到小进行排序得到结果集 R_2' , 即是按照用户兴趣排序的结果.

3) 将 R_2' 放在 R_1 的后面形成最终排序好的结果集 R. 至此, 可输出按用户兴趣输出个性化的排序的结果列表, 从而提高检索结果与用户需求之间的相关性, 达到面向用户的检索目的.

4 实验与分析

油田企业数据类型比较丰富, 具有代表性的典型数据源类型有 SPS 门户网站、CMS 内容管理系统、Oracle 勘探开发等数据库、FTP 项目文件系统和 Lotus Domino 办公系统, 正对应本系统支持的数据类型, 适于将本研究用于油田企业信息搜索进行验证实验.

笔者参与开发的油田信息搜索引擎(Oilfield Information Search, 简称 OIS)是一个企业级的搜索引擎应用研究实例项目, 在本研究之前 OIS 只使用基本的排序方法而未采用基于用户兴趣的个性化排序方法. 油田企业用户通过大量实验对 OIS 采用本研究改进前后的检索效果进行对比, 验证本研究的正确性和可适应能力. 我们以油田企业用户“贾承业”在存在兴趣偏好的情况下的检索结果举例说明. 用户“贾承业”要查找自己写过的一篇开发文档, 使用同一检索请求在 OIS 采用个性化排序方法之前和之后的检索结果片段分别如图 2 和图 3. 由于用户“贾承业”与用户模型中“贾承业”的兴趣相似度更高, 于是图 2 第 2 条结果经个性化排序之后的位置得到提升如图 3, 更加满足用户“贾承业”的个人兴趣需求.

通过油田企业用户大量实验验证, 本文中公式(6)中的取值在 0.50-0.83 范围内对基于用户兴趣排序的结果得分值影响适当, 排序结果更能满足用户的个性化需求.

图 2 采用个性化排序之前的检索结果

图 3 采用个性化排序之后的检索结果

由实验结果对比举例可知，在用户存在兴趣偏好、采用本个性化排序方法的情况下的检索效果优于其未采用之前，达到了“面向用户”的检索目的，能够更好地为用户提供个性化的检索结果.

5 结语

本文针对传统搜索引擎不能满足用户的个性化检索需求的缺点，从个性化信息检索服务的角度出发，对个性化服务的关键技术即用户兴趣建模技术进行了研究，并为企业搜索引擎设计基于用户兴趣的个性化排序方法. 经过实验验证，本文研究能有效提高企业搜索引擎检索精确度及满足用户的个性化检索需求，并对用户有较好的自适应能力.

本文下一步将致力于从语法、语义的角度进一步探讨个性化服务的应用，提高个性化信息服务的准确率和效率.

参考文献

1 Liu F, Yu C, Meng WY. Personalized web search for improving retrieval effectiveness. IEEE Trans. on Knowledge and Data Engineering, 2004,16(1):28–40.

2 陆安江,董旭晖.个性化搜索引擎模型的研究与设计.计算机与现代化,2011,(1):139–141.

3 袁薇,高淼.搜索引擎系统个性化机制的研究.微电子学与计算机,2006,23(2):68–72.

4 冯子威.用户兴趣建模的研究.哈尔滨:哈尔滨工业大学,2010.

5 崔顷顷.基于个性化搜索的系统研究与设计.北京:北京交通大学,2011.

6 申艳光,王敏,范永健.面向隐私保护的个性化搜索结果排序方法研究.数学的实践与认识,2011,41(19):97–100.

7 杨光伟.基于 Lucene 的个性化搜索引擎的研究与实现.呼和浩特市:内蒙古大学,2009.

8 肖卓程,荆金华.基于用户兴趣的搜索引擎.计算机应用与软件,2007,24(9):134–136.