情报学报 ISSN 1000-0135
第34卷第7期711-716,2015年7月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 34 No. 7, 711-716 July 2015

doi:10.3772/j.issn.1000-0135.2015.007.005

基于VSM和偏好本体的个性化信息检索技术的研究[1)]

张一洲

(中共淮安市委党校,江苏 223003)

摘要 信息检索技术特别注重向量空间模型和偏好本体的结合。为便于找出用户输入的关键词间的关系,利用本体的关联分析及模糊本体值计算关键词间的相似度,本文采用加权算法和排序算法计算每个文档的权重,并根据文档权重进行检索结果的重排。为优化重排模型,本文还对每个检索对象的驻留时间进行合并,并使用衡量搜索引擎质量指标和评价标准F-measure对本文提出的重排机制性能进行测试。实验结果表明,使用本方法进行个性化信息检索的性能优于Google方法。

关键词 信息检索 偏好本体 关联分析 驻留时间

Research of Personalized Information Retrieval Technology Based on VSM and Profile Ontology

Zhang Yizhou

(*Party School of Chinese Communist Party Huai'an City State, Jiangsu* 223003)

Abstract Information Retrieval (IR) techniques specifically focus on combination of Vector Space Model (VSM) with Profile Ontology. In this paper, we propose a novel hybridization of the IR processing to calculate the weight of each document and to find relations between the user entered terms by using the weighting algorithm and the ranking algorithm, and takes advantage of ontology-based correlation analysis which uses the fuzzy ontology value to calculate the similarity score between terms and includes the re-ranking algorithms to display the search results according to the weight of the document. We incorporate the Dwell Time of each retrieval session to optimize re-ranked model, and the performance of our re-ranking mechanism using Discounted Cumulative Gain (DCG) and F-measure was tested. The experimental result shows that the Web retrieval efficiency achieves improvement when our personalized retrieval approach is compared with the Google search.

Keywords information retrieval, profile ontology, correlation analysis, dwell time

1 引 言

搜索引擎指自动从因特网搜集信息,经过一定整理以后,提供给用户进行查询的系统[1]。可是,目前大多数的搜索引擎并没有考虑用户的偏好,无论用户的真正兴趣怎样,只要输入相同的查询都会获得广义关键词的搜索结果。一般情况下,这类搜索引擎只是简单地基于关键词匹配推荐信息,并不分析用户的兴趣,从而产生太多与用户兴趣无关的信息,因为一个特定的用户只对通用检索结果中的一小部分信息感兴趣。由于检索系统需考虑不同背景、不同目的、不同时期、不同用户的查询请求,因此个性化信息检索可有效地提高信息检索的效率[2]。

收稿日期:2015年1月9日

作者简介:张一洲,男,汉,1981年生,硕士,副教授,主要研究方向:信息管理与信息系统、智能化信息处理技术,E-mail:42777278@qq.com。

1) 基金项目:江苏省高校哲学社会科学基金项目(No.2012SJD870001)。

经典的个性化服务研究采用加权算法与排序算法技术相结合的策略确定输入的关键词与文档选择和文档关联 之间是否存在关联。个性化检索是搜索引擎的一个未来发展的重要特征和必然趋势之一。个性化检索通过搜索引擎的社区化产品(即对注册用户提供服务)的方式来组织个人信息[3],并按用户兴趣爱好进行排序,然后在搜索引擎基础信息库的检索中引入个人因素进行分析,获得针对个人不同的检索结果。最新的研究开始关注用户兴趣的自主学习,根据用户兴趣和爱好调整个人需求的检索结果。个性化信息检索技术就是针对这一问题提出的、区别对待用户之间的不同之处,为不同的用户提供不同的服务,以满足不同的需求。

本文提出的基于向量空间模型 VSM(Vector Space Model)和偏好本体相结合的个性化信息检索技术,考虑的应用情景是一个用户在一段时间以来浏览和点击过的文档组成用户模型,并包含用户感兴趣的内容。当用户进行一个新的查询时,新查询与其以前的兴趣相同或关联。通过对用户模型的学习,可得到关联的语义内容,使得获得的结果更接近用户兴趣。

2 个性化信息检索技术

本文提出的信息检索技术包含:创建日志文件、创建偏好本体、创建 VSM 模型、合成 Web 检索及重排检索结果5个主要模块。

2.1 创建日志文件

根据用户创建日志文件推导出构建用户本体和用户偏好模型所需的数据库。日志文件是用户浏览 Web 细节的搜索日志,包括输入的关键词、访问过的网站,访问 Web 的时间和给定的关键词对应的 Google 索引次数,每个用户所对应的 IP 地址也包括在内[4],表1显示的是一个用户日志文件示例。为了收集用户的搜索日志,每个参与者的机器上需安装浏览器扩展软件;为了获得完整的日志数据库,需合并所有用户日志。用 $U_i(0<i\leqslant Q)$ 标识每个用户的日志文件,也就是说,如果 Q 是用户数,那么完整的日志数据库将包含 Q 个用户日志文件细节。

使用 U、P、K、L、T 和 G 分别表示用户 ID、IP 地址、输入的关键词、访问过的 URLs、访问时间和给定的关键词搜索过的 URL[5]。任意用户 i 都可用字段 U_i、P_i、K_i、L_i、T_i 和 G_i 表示,U_i 是用户 i 的 ID,P_i 是分配给 U_i 设备的 IP 地址,K_i 是 U_i 搜索的关键词集,L_i 是 U_i 浏览过的链接集,T_i 是访问超链接的时间,G_i 是 U_i 搜索的次数。K_i 可表示为 $K_i=\{k_{i1},k_{i2},\cdots,k_{iN}\}$,其中 k_{ij} 是 U_i 输入的第 j 个关键词,N 是 U_i 输入的关键词总数。同样地,L_i 可表示为 $L_i=\{l_{i1},l_{i2},\cdots,l_{iM}\}$,其中 l_{ij} 是 U_i 访问过的第 j 个超链接,M 是 U_i 访问过的超链接数。t_{lij} 对应于 l_i 的访问第 j 个超链接的时间,因此时间集 T_i 可表示为 $T_i=\{t_{li1},t_{li2},\cdots,t_{liM}\}$,搜索集 G_i 可表示为 $G_i=\{g_{i1},g_{i2},\cdots,g_{iM}\}$。

2.2 创建偏好本体

根据日志数据库和 URL 超链接关键词间的关系进行本体映射。本体创建时将模糊本体值看作是访问过的文档和输入的关键词间的相似度。对任何 U_i 来说,都可对关键词集 K_i 进行分析,以便本体映射时创建关键词间的关系。本体映射获得时,需对 U_i 使用过的所有关键词进行分析,以便开发完整的偏好本体。图1是建立本体的流程图。

表1 用户日志文件示例(用户 ID:IT0909,IP 地址:192.163.10.29)

关键词	URL 链接	访问时间	Google 索引数
Data Mining	http://www.cedfaculty.com/Technology/datamining.html	8/July/2014:20:18:21 +0790	3
Machine Learning	http://www.laits.tuxesa.edu/~norman/Information-management/Machine-learning.html	8/July/2014:20:18:39 +0790	4
Decision Support System	http://www.data-management1.com/dss/Decision-Support-System.html	8/July/2014:20:18:55 +0790	3

采用模糊本体技术计算任意关键词和任意网址间的相似度。首先进行数据库的创建，然后考虑 U_i 的关键词集 $K_i = \{k_{i1}, k_{i2}, \cdots, k_{iN}\}$，$N$ 是数据库中关键词的总数，关键词集中的关键词按用户输入的顺序排列。每对关键词 k_{ij} 和 $k_{ir}(j \neq r)$ 的相似度初始值都为零。在关键词列表中寻找 U_i 连续输入关键词 k_{ij} 和 k_{ir} 的次数，即 $j-r(1$，并为每对 c_{jr} 增加对应的相似度。当相同的关键词 k_{ij} 和 k_{ir} 连续出现超过 1 次时，每出现一次 c_{jr} 值就加 1，每对关键词最后的相似度就确定了。也就是说，关键词 k_{ij} 和 k_{ir} 间的最后相似度就是它们在关键词列表中连续出现的次数。关键词 k_{ij} 和 k_{ir} 的模糊本体价值 F_{jr} 可由公式(1)定义，其中 X 是数据库中关键词的总数。

$$F_{jr} = c_{jr}/X^2 \tag{1}$$

根据模糊本体值和提供的数据库，可自动获取图 2 所示的本体示意图。假设给定的关键词是 A、B、C 和 D，输入的顺序是 A、B、D、A、D、A、B、C、D、A，那么 $c_{AB}=2$，$c_{AC}=0$，$c_{AD}=4$，$c_{BC}=1$，$c_{BD}=1$，$c_{CD}=1$，且 $F_{AB}=0.125$，$F_{AC}=0$，$F_{AD}=0.25$，$F_{BC}=0.0625$，$F_{BD}=0.0625$，$F_{CD}=0.0625$。

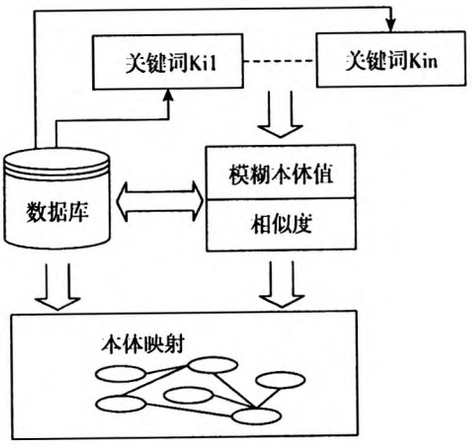

图 1　创建本体流程图

本体创建后，为了获取完整的偏好本体就需对关键词关联的 URL 特征进行分析。数据库创建后需考虑第 i 个用户和给定的关键词 K_T 的 URL 集 $L_i = \{l_{i1}, l_{i2}, \cdots, l_{iY}\}$ 间的关系，Y 是数据库中 URLs 与关键词关联的总数，同时 URLs 需出现在访问的序列中。URLs 和对应的关键词 K_T 连续出现在列表的次数就是 l_{ij} 和 l_{ir} 两 URLs 间的 URL 的相似度 cu_{jr}。模糊本体的 *URL* 值 Fu_{jr} 由公式(2)定义。

$$Fu_{jr} = cu_{jr}/Y^2 \tag{2}$$

设用户输入关键词 K_T 并依次访问 URLs：l_1、l_2、l_3、l_4，观察随后访问的 URLs 并对派生关系进行分析。假设另一用户输入相同的关键词 K_T 并依次访问 URLs：l_1、l_4、l_2，第三个用户也使用相同的关键词 K_T 依次访问 URLs：l_1、l_4、l_3、l_2。然后根据 3 个用户访问 URL 的方式，利用相似度和模糊本体值进行偏好映射，如图 3 所示。$cu_{l1l2}=1$，$cu_{l1l4}=2$，$cu_{l2l3}=2$，$cu_{l2l4}=1$，$cu_{l3l4}=2$，$Fu_{l1l2}=0.0625$，$Fu_{l1l4}=0.125$，$Fu_{l2l3}=0.125$，$Fu_{l2l4}=0.0625$，$Fu_{l3l4}=0.125$。

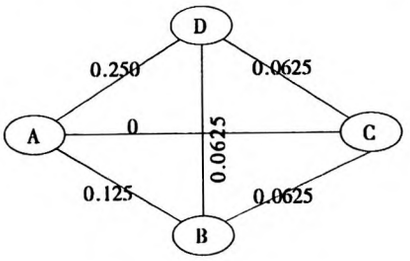

图 2　本体创建示例

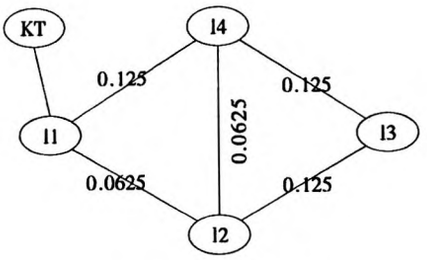

图 3　偏好创建示例

同样的方法对所有的关键词进行相应的偏好映射，最后映射成的偏好本体就是完整的偏好映射。

2.3　创建 VSM 模型

向量空间模型 VSM 是目前表示用户模型最流行的方法，因此本文也采用 VSM 表示网络用户模型。为了跟踪和学习用户兴趣行为，必须为每个用户建立模型。用户模型用于存储用户兴趣、存储管理用户的历史行为、存储学习用户行为知识并进行相关知识功能集的推导[6]。TF-IDF 加权的各种形式常应用于搜索引擎，用作文档与用户查询相关程度的度量或评级。将文档的关键词 TF-IDF 向量表示成 VSM 代数模型。词频 TF(Term Frequency，TF)，反文档频率 IDF(Inverse Document Frequency，IDF)，可用 TF-IDF 判断文档集或语料库中关键词对文档的重要性。关键词的重要性随着关键词在文档中出现的次数成正比增加，但也会随着关键词在语料库中出现的频率成反比下降。

用户首先输入关键词 K_T，返回关联的文档列表，从获得的文档列表中选择排在最前面的 N_d 个

文档 $D(\{d_1,d_2,\cdots,d_{Nd}\})$，对每个文档选择排在最前面的 M_k 个关键词进行处理。在关键词抽取时，首先对每个文档 $d_i(0<i\leqslant N_d)$ 进行处理，然后对停用词进行删除预处理，这样就可抽取出文档中的关键词[7]。

排在最前面的 N_d 个文档可用关键词集 $d_i(\{k_{i1},k_{i2},\cdots,k_{iMk}\})$ 定义，关键词集由 M_k 个最常见的关键词组成。每个关键词 $k_{ij}(0<i\leqslant N_d,0<j\leqslant M_k)$ 的 *TF-IDF* 值由各自关键词确定。基于每个关键词的 *TF-IDF* 值，文档中检索过的所有关键词的权重 W_{ij} 使用公式(3)计算。

$$W_{ij}=tf(k_{ij})\times idf(k_{ij})/\sqrt{\sum(tf(k_{pq})\times idf(k_{pq}))} \tag{3}$$

其中，关键词 k_{ij} 的词频 *TF* 是关键词 k_{ij} 在第 j 个文档中出现的次数，反文档频率 *IDF* 可由公式(4)计算获得，即 N_d 文档总数除以含关键语 N_t 的文档数，然后取商的对数。

$$idf(k_{ij})=\log(N_d/N_t) \tag{4}$$

其中，N_d 是文档的总数，N_t 是关键词 t 在文档 d 中的词频率因子，K_{pq} 是从关联的 URL(即偏好本体示意图)中获得的关键词数。

2.4 合成 Web 检索模型

本文提出的模型是向量空间模型与偏好本体映射的结合，综合使用权重算法和排序算法进行网页重排，并用驻留网页时间进行个性化的性能改进。据了解，在探索个性化检索过程中，几乎还没有人使用驻留网页时间进行导航和事务的相关研究。本文首先进行信息分类，以便计算最终排序时使用。使用公式(5)加权方案获得的文档最终排序有助于提高个性化的程度，即驻留网页时间越长，该文档就越重要。

$$FinalRank=(1-\alpha)Googleindex+\alpha(PersonalizeRank) \tag{5}$$

其中，α 是驻留网页时间，以千秒为单位，范围是 0 到 1。假设时间范围是 3 秒到 15 分钟，20 秒等于 0.020，3 分钟等于(3 * 60)/1000。当 $\alpha=0$ 时，给出的个性化排序就没有权重，但仍使用 Google 索引次数进行排序；相反，如果 $\alpha=1$，个性化的排序就是最终排序。

2.5 重排检索结果

用向量 $\vec{d}=<w_1,w_2,\cdots,w_n>$ 表示文档集中任一文档的关键词，w_i 是关键词 i 的权重，即 $w_i=tf_i*idf_i$，其中 tf_i 是关键词 i 在文档 d_i 出现的频率(索引项频率)，idf_i 是反文档频率，由公式(4)确定。当用户使用特定的查询 q 搜索文档时，可用余弦相似度方法搜索最相似的文档，查询 q 中的关键词应包含在最相似的文档中。基于以下两主要因素，用本文的检索方法得到的检索结果就已被重新排序：①文档集和用户语义文档之间的语义关系；②文档集和用户语义文档最相似簇之间的语义关系。使用算法 1 对用户文档进行映射排序，根据查询 q 的余弦相似度，对每个大类中的所有文档进行重新排序。给用户的语义文档 d_i 分配一优先级($\alpha=5.0$)(第 1 类)、推荐簇的文档 d_i 分配一优先级($\beta=3.0$)(第 2 类)，剩下的文档分配最低的优先级($\gamma=1.0$)(第 3 类)。根据查询 q 的余弦相似度，对每个大类中的所有文档进行重新排序。本文的个性化检索技术(基于 I-Match 方法[8])使用可选等级决定索引文档中每个关键词的重要性，同时累加文档查询关键词之间的余弦相似度。这样，较高优先级的关键词将起到更重要的作用，具体细节见算法 3。修改如下的可选等级：属于第 1 类情况，doc. setBoost( ) = α；属于第 2 类情况，doc. setBoost( ) = β；属于第 3 类情况，doc. setBoost( ) = γ。

算法 1 重排检索结果

输入：q；//关键词搜索
输出：$Rank=\{d_1,d_2,\cdots,d_n\}$； //重排
$Rank=\{d_1,d_2,\cdots,d_n\}$；//查询 q 的默认搜索结果
$UR_i=U_{j=1}^{n}SC_{ji}+U_{k=1}^{l}d_{ki}$；
$RC=U_{c=1}^{l}d_c$；// l 是推荐簇的文档数
For each $d_j\in$ **R***ank*
  If $d_j\in UR_i$ *then*
    $d_j.boost=\alpha$；//用户文档
  End
  Else
If $d_j\in$ **RC** *then*
    $d_j.boost=\beta$；//推荐文档
  End
  Else
  $d_j.boost=\gamma$；
  End
End
根据 $d_j.boost$ 排序

3 实验结果与分析

3.1 实验设置

本文使用不同关键词进行一系列的检索实验，每个检索都会根据 Google 搜索返回结果产生个性化网页重排序列。实验时，既采用 NDCG (Normalized Discounted Cumulative Gain) 测试本文提出的个性化检索质量，又使用精度、召回率及 F-measure 等评价指标对相关的序列和结果进行相关性比照测试。

本实验选择的 60 个用户都在高校 IT 部门工作，且从事 Web 搜索研究。为了获得丰富的日志文件，两周进行一次样本数据收集，收集的样本数据中包含用于选择 URL 的 1356 关键词、Google 搜索次数和每个选定的 URL 驻留时间，其中获得的重要关键词有 708 个(驻留时间在 3 秒到 15 分钟外的关键词被删除掉)，同时将收集到的数据分成两组：每组包含用于测试的 15 个关键词，并采用 Jarvelin 提出的策略进行本文提出的检索技术测试[9]。

进行测试时，选择 8 个志愿者并给测试数据集各自分 15 个关键词。数据选择需符合本文提出的方法：所有志愿者使用最频繁的关键词可为每个用户进行唯一目标查询提供很好的衡量指标，使得用户查询顺序与用户偏好相一致。发给志愿者最频繁的关键词有“数据挖掘”、“机器学习”及“支持决策系统”，其余 166 个关键词用作训练数据库。对数据库剩余的关键词(共 527 个关键词)进行标识以便对本文提出的技术作更多的验证。对每个访问链接，要求志愿者用 0 ~ 3 分值进行等级评价：3 是相关的，2 是部分相关，1 是有些相关，0 是无关紧要。

3.2 结果分析

表 2 和表 3 显示的是 8 个不同志愿者使用 15 个关键词进行检索的实验结果。表 2 显示的是使用本文提出的检索方法和 Google 搜索方法产生检索结果的 NDCG 值；表 3 显示的是使用本文提出的检索方法和 Google 搜索方法计算其精度、召回率及 F-measure 的评价值。

表 2 中的比较结果清楚地表明本文提出的四级个性化检索和传统 Google 搜索计算关键词的 NDCG 值之间存着显著差异。在本文提出的四级个性化检索实验中共有 13 个关键词的 NDCG 值在 0.60 和 0.90 之间，只有两个关键词的 NDCG 值低于 0.60；而 Google 搜索，分值 0.60 以上的关键词只有 6 个，且 NDCG 最高分值是 0.79，更多的关键词(9 个关键词)的 NDCG 分值低于 0.60。

使用本文提出的检索技术和 Google 搜索计算志愿者的反馈和每个链接的精度、召回率及 F-measure 值。表 3 显示的是对前面提到的 3 个查询(“数据挖掘”、“机器学习”、“支持决策系统”)计算得到的精度、召回率及 F-measure 的平均值。从 3 个查询的平均值中可看出本文提出的检索精度和 F-measure 值都比较高，且拥有更多个与 3 个查询相关的检索结果，因而可证得本文提出的检索技术优于 Google 搜索。

表 2 检索 15 个关键词的 NDCG 值

关键词编号	个性化	Google	关键词编号	个性化	Google
1	0.78	0.64	9	0.52	0.39
2	0.65	0.45	10	0.75	0.62
3	0.83	0.65	11	0.68	0.50
4	0.68	0.49	12	0.74	0.56
5	0.69	0.53	13	0.75	0.65
6	0.90	0.79	14	0.81	0.57
7	0.65	0.55	15	0.73	0.60
8	0.56	0.50			

表 3 检索 3 个查询样例的精度、平均召回率和平均 F-Measure 平均值

查询样例		个性化	Google
数据挖掘	精度	0.78	0.57
	召回率	1.00	1.00
	F-Measure	0.89	0.73
机器学习	精度	0.81	0.66
	召回率	1.00	1.00
	F-Measure	0.90	0.82
支持决策系统	精度	0.80	0.62
	召回率	1.00	1.00
	F-Measure	0.89	0.76

从表 3 中可看出使用本文提出的个性化检索技术得到的 3 个查询样例的平均精度分别是 0.78、

0.81 和0.80,而使用 Google 搜索得到的平均精度分别是0.57、0.66 和0.62,平均精度高出19%;同时使用本文提出的个性化检索技术得到的3个查询样例的 F-measure 分别是0.89、0.90 和0.89,而使用 Google 搜索得到的平均 F-measure 分别是0.73、0.82 和0.76,平均 F-measure 高出13%。更好的精度和 F-measure 表明使用本文提出的个性化检索技术取得的平均精度和平均 F-measure 都比 Google 搜索好。

4 结束语

本文提出了一种用于增强个性化服务程度的个性化信息检索新方法。为了给 Web 用户提供高质量的信息,在提出的方法中综合运用向量空间模型和偏好本体技术,根据用户自己兴趣将检索结果按相关性顺序进行检索结果的排序。首先侧重于用户浏览模式的研究,生成单个用户日志文件,并在单个用户偏好的基础之上,将基于用户偏好的本体用作个性化检索输入排序模型,创建所有用户的日志数据库,同时发现源于用户日志的偏好兴趣本体模型具有个性化检索潜力。其次,使用驻留网页时间进行个性化检索的改进,间接确定输入关键词的权重和筛选最相关的文档。在将本文的设计方案与 Google 搜索相比时,发现使用本文的方法进行个性化信息检索的性能优于 Google 搜索方法13%。

参 考 文 献

[1] 搜索引擎[EB/OL]. [2014-06-24]. http://baike.so.com/doc/311390.

[2] 张培颖,李村合. 智能搜索引擎中个性化信息检索技术研究[J]. 科学技术与工程, 2008, 8(17): 5046-5049.

[3] 李树青. 个性化信息检索技术综述[J]. 情报理论与实践,2009,36(5):107-113.

[4] 岑荣伟, 刘奕群, 张敏,等. 网络检索用户行为可靠性分析[J]. 软件学报,2010,21(5):1055-1066.

[5] Radlinski F, Matthijs N. Personalizing web search using long term browsing history[C]//Proceedings of the 4th ACM International Conference on Web Search and Data Mining, Hong Kong, 2011: 25-34.

[6] Vallet D, Fernandez M, Castells P. An adaption of the vector space model for ontology based information retrieval [J]//IEEE Transaction on Knowledge and Data Engineering,2007,19(2):261-272.

[7] Gao J, Yuan W, Li X, et al. Smoothing clickthrough data for web search ranking[C]//Proceedings of the 31nd international ACM SIGIR Conference on Research and Development in information Retrieval, Singapore, 2008:355-362.

[8] 李银松,施水才,张玉杰,等. 用户兴趣分类在个性化搜索引擎中的应用[J]. 情报学报, 2008, 27(4): 535-540.

[9] Kekalainen J, Jarvelin K. IR evaluation methods for retrieving highly relevant documents[C]//Proceedings of the 23rd Annual International SIGIR Conference, Hong Kong,2000:41-48.

(责任编辑 贾 佳)