北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.038

一种基于查询加权的用户建模方法

胡娟　白宇†　蔡东风

沈阳航空航天大学知识工程研究中心，沈阳　110136; † 通信作者, E-mail: baiyu@sau.edu.cn

摘要　通过分析用户的查询日志，模拟用户与搜索引擎之间的交互过程，提出一种基于查询加权的用户建模方法。首先，对查询日志进行会话分割；然后，利用会话中用户查询出现的次数、持续时间及所点击的 URL 排名等行为信息，计算查询权重；最后，采用兴趣投票的方式，完成用户模型的构建。在 AOL 美国在线查询日志数据集上的测试结果表明，基于查询加权的用户建模方法在用户兴趣预测上取得较好的效果。

关键词　用户建模；查询日志；会话分割；查询加权

中图分类号　TP391

A Query Weighted-Based Method for User Modeling

HU Juan, BAI Yu †, CAI Dongfeng

Knowledge Engineering Research Center, Shenyang Aerospace University, Shenyang 110136;
†Corresponding author, E-mail: baiyu@sau.edu.cn

Abstract　A query weighted-based method is proposed for user modeling by simulating the interaction between user and search engine. First, the query log is divided into sessions according to the session division principle. Then, for each session, a group of user behavior information, such as query frequency, duration and the ranks of the clicked URLs, are employed to calculate the weight of queries. Finally, the voting method is used to generate user model. The experiment results show the effectiveness of the method over the AOL query log dataset.

Key words　user modeling; query log; session division; query weighted

互联网规模和覆盖面的增长带来了信息过载(information overload)问题：过量信息同时呈现使得用户很难从中获取对自己有用的部分，信息使用效率反而减低。推荐系统作为一种信息过滤的重要手段，是当前解决信息过载问题的非常有潜力的方法[1]。个性化推荐分为用户兴趣建模、推荐对象建模和推荐算法 3 部分，如图 1 所示。其中用户兴趣建模可以发现用户的偏好，是个性化推荐系统的核心技术之一。在个性化推荐系统中，用户模型能否很好地反映用户的兴趣爱好，直接决定推荐结果的好坏。

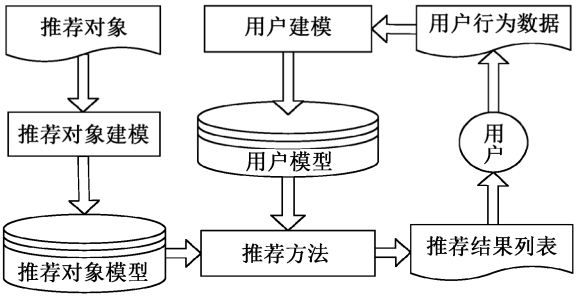

图 1　推荐系统框图

Fig. 1　Framework of recommendation system

用户建模是获取和维护用户兴趣、需求和习惯的过程，最后得到表示用户特有兴趣的用户模型[2]。用户建模一般包括两个方面的内容：一方面，通过记录和分析用户浏览行为、浏览内容及用户反馈等收集用户信息，从中挖掘用户兴趣；另一方面，用合适的方法表示用户兴趣(即用户兴趣模型)，并随用户兴趣变化动态更新用户兴趣模型。目前，用

国家自然科学基金(61403262)和辽宁省教育厅科学技术研究项目(L2013066)资助
收稿日期: 2014–07–26; 修回日期: 2014–10–24; 网络出版时间: 2014-12-01 09:26

户模型表示形式主要有：基于关键词的用户模型、基于类别的用户模型和基于潜在主题的用户模型。

查询日志记录用户和搜索引擎交互的相关行为信息，是研究和分析真实网络用户行为的主要载体[3]。

面对海量的互联网信息资源，搜索引擎利用简单的关键词匹配得到的结果，很难满足用户的需求。因此，利用查询日志对用户的行为进行研究，得到用户模型，变得尤为重要。

会话(session)是查询日志中单个用户在一段时间间隔内所提交的整个查询序列，其中查询串的个数定义为会话长度。用户使用搜索引擎过程中，往往有一个特定的查询意图，当返回的结果满足用户的需求或用户放弃本次查询意图时，用户结束本次会话，开始下一个会话。因此，会话具有以下特点：一个会话是在一定的时间间隔内发生的；会话内部相邻查询的时间间隔较短；一个会话对应特定的查询意图。表 1 为查询日志中某个用户的 3 个会话。

本文通过对用户在搜索引擎中的查询日志进行分析，提出一种基于查询加权的用户建模方法。该方法考虑到会话的特点，将查询日志进行会话分割。模拟用户利用搜索引擎进行会话的过程，提出 3 个假设，充分利用会话内部查询的出现次数、持续时间和所点击的 URL 排名等行为信息，进行查询加权，然后利用兴趣投票的形式，得到用户模型。

1 相关工作

在用户建模的研究中，按照用户模型表示方式不同可分为：基于关键词的用户建模[4−6]、基于类别的用户建模[7−8]和基于潜在主题的用户建模[9−11]。

Chen 等[4]利用用户发送的微博和关注的信息，将每个用户的微博信息看作一篇文档，利用 TF-IDF 计算权重的方法，建立用户关键词集合，得到用户模型，并充分利用用户的微博信息和用户社会信息，提高了用户模型的准确性。Matthijs 等[5]将用户的浏览历史看作用户的相关文档，分别融合 TF-IDF 和 BM25 计算权重的方法，得到用户的关键词及其权重的向量集合，更加准确地计算了用户对某个关键词的权重。张新猛等[6]利用用户的历史评分项目和数据，对二部图进行加权，引入扩散的理论，按照二部图边权占该节点权重和的比例分配资源，实现对用户行为的预测，建立用户项目模型。此方法充分利用了用户与用户之间的关系，用户预测准确率得到有效的提高。Liu 等[7]将用户的查询按照预定义的类别进行分类，得到一个排序的类别结果，作为用户的搜索兴趣模型。Tian 等[8]提出频繁集聚类用户模型，计算用户频繁使用的资源的主题向量，形成多个频繁兴趣簇来表征用户的多个兴趣方向。Sontag 等[9]提出一个概率模型，将用

表 1 查询日志中某用户的会话样例

Table 1 Session sample of an user

UserID	Query	QueryTime	Rank	ClickURL
Session1	midway online literary journal	2006–04–21 11:24:43	3	http://www.mndaily.com
	midway online literary journal	2006–04–21 11:24:44	9	http://www.smallspiralnotebook.com
	meridian literary magazine	2006–04–21 11:38:21	2	http://www.engl.virginia.edu
	meridian literary magazine	2006–04–21 11:38:25	6	http://www.fglaysher.com
Session2	mark twain middle school	2006–04–21 14:38:23	2	http://www.fcps.k12.va.us
	mark twain middle school	2006–04–21 14:38:27	1	http://www.fcps.k12.va.us
Session3	university of massachusettes mfa blog	2006–04–22 07:22:43	3	http://www.thepublishngspot.com
	university of massachusettes mfa blog	2006–04–22 07:22:48	5	http://www.pitt.edu
	university of massachusettes mfa blog	2006–04–22 07:22:49	10	http://snreview.wordpress.com
	university of massachusettes mfa blog	2006–04–22 07:29:42	15	http://www.myspace.com
	university of massachusettes mfa blog	2006–04–22 07:29:45	24	http://maudnewton.com
	babies are fireproof	2006–04–22 07:31:22	1	http://babiesarefireproof.blogspot.com

户和文档表示成潜在主题的形式, 得到主题表示的用户模型。Iwat 等[10]将用户购买的商品看作动态主题模型中构成文档的单词, 建立模型来模拟用户购买商品的过程。在此基础上, 经过模型学习可以得到一系列模型参数, 利用这些参数得到用户兴趣模型。Morgan 等[11]利用查询日志将用户模型表示成基于点击文档的主题空间模型, 将用户的兴趣表示成潜在主题的形式, 挖掘出用户基于语义层次的兴趣。

以上方法考虑了与用户相关的文档或物品信息, 通过分析这些信息挖掘用户的兴趣, 得到用户模型。这些模型并未考虑用户与应用之间的实时交互的行为信息, 而这些信息在很大程度上决定了用户对于某一物品的喜好程度。因此,在以上方法的基础上, 本文考虑用户与搜索引擎交互的行为信息, 完成用户模型的构建。

2 基于查询加权的用户建模

2.1 用户建模方法

本文构建的用户模型表示为<关键词, 兴趣度>向量的形式, 用户模型如式(1)所示, 基于查询加权的用户建模方法的系统框架如图 2 所示。

$$\text{UserInterest}=\{(T_1,W_{T_1})(T_2,W_{T_2})....(T_{T_M},W_{T_M})\}。 \quad (1)$$

首先, 将用户的查询日志按用户的不同进行分割, 筛选出查询日志条数满足某个阈值的用户, 将这部分用户的查询日志作为系统的输入数据。

然后, 对用户的查询日志进行会话分割。一个会话指的是同一个用户在某一小段时间内的连续查询[12]。一般情况下, 用户利用搜索引擎进行检索时, 必定有某个特定用户意图或信息需求, 用户通过一个或多个查询获得用户感兴趣的信息, 或放弃本次意图的搜索过程。这样的一个过程就是用户和搜索引擎之间的一次会话过程, 在整个过程中用户的意图没有发生改变。由此, 可以得到会话的几个特性: 具有特定的时间间隔; 用户在同一会话中的用户意图固定不变; 会话内部查询与查询之间的间隔比较短。因此, 本文利用文献[13]中的会话分割方法对查询日志进行会话分割: 同一会话的时间间隔不超过会话时间阈值; 同一会话中相邻查询之间的时间间隔不超过查询时间阈值; 同一会话中相邻查询之间的余弦相似度不小于查询相似度阈值。

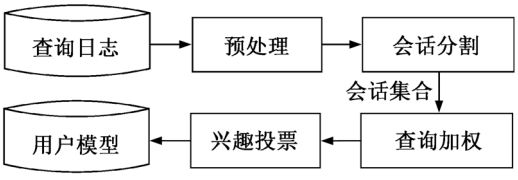

图 2 用户建模系统框图

Fig. 2 Framework of user modeling

最后, 对每个会话中的每个查询进行加权, 并对用户的关键词进行投票, 得到用户模型。模拟用户使用搜索引擎进行某个会话的过程: 当用户多次使用某一特定的查询时, 表明该查询能很好地反映用户的查询意图; 当用户对某一特定查询所持续的时间较长时, 用户很可能找到了感兴趣的信息, 表明该查询能反映用户的查询意图; 当用户输入某个查询, 搜索引擎会按相似度高低进行排序, 将较高相似度的信息排在前面返回给用户。当用户点击的 URL 排名越靠前时, 表明该查询能较好地反映用户的意图。

通过以上分析, 提出以下 3 个假设: 1) 会话中某个查询出现的次数越多, 权重越大; 2) 查询持续的时间越长, 权重越大; 3) 用户提交的某个查询对应的 URL 平均排名越靠前, 查询的权重越大。

得到一个用户的某个会话中每个查询的权重后, 对用户查询日志中出现的关键词进行投票, 最后的得到用户的兴趣模型。

2.2 查询加权及兴趣投票

同一个会话包含若干个查询, 基于本文提出的 3 个假设, 利用每个会话中某个查询出现的次数、持续时间、点击 URL 的平均排名信息对该查询进行加权, 流程如图 3 所示。其中, FreRate 表示查询出现次数的比率, TimeRate 表示查询平均持续时间的比率, AveRank 表示查询点击 URL 平均排名的倒数。

2.2.1 查询出现次数比率

查询在会话中出现的次数表示相同的查询在同一会话中出现的次数。查询 Q_{k_j} 在会话 S_k 中出现次数的比率 $\text{FreRate}_{Q_{kj}}$, 如式(2):

$$\text{FreRate}_{Q_{k_j}}=\frac{\text{Fre}_{Q_{k_j}}}{Q}, \quad (2)$$

其中, Fre_{Q_j} 表示查询 Q_{k_j} 在会话 S_k 中出现次数, Q 表示会话 S_k 中所有查询的总数。

2.2.2 查询平均时间比率

如图 4 所示, 会话中按时间排序的查询串为 $\{Q_1,Q_2,....,Q_i,.....,Q_K\}$, QueryTime_{Q_i} 表示查询 Q_i 的

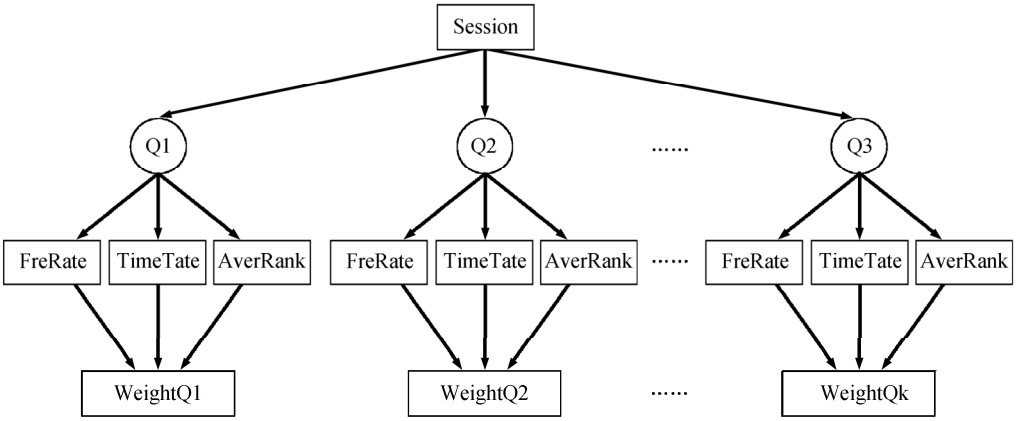

图 3　查询加权框图

Fig. 3　Framework of query weighted

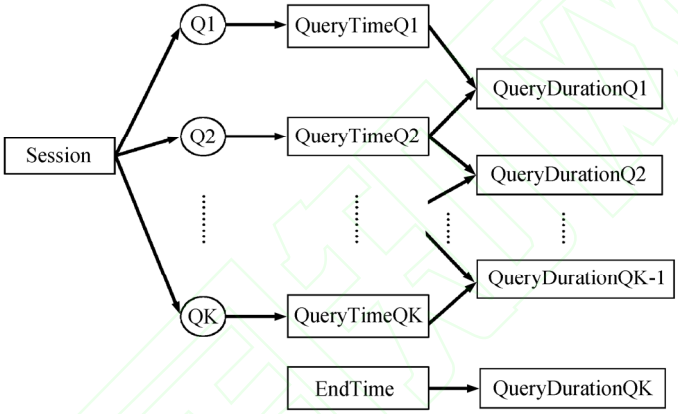

图 4　查询持续时间计算框图

Fig. 4 Framework of the calculation of query duration

查询时间，$\text{QueryDuration}_{Q_i}$ 表示查询 Q_i 的持续时间，计算方法如式(3):

$$\text{QueryDuration}_{Q_i}=\begin{cases}\text{QueryTime}_{Q_i+1}-\text{QueryTime}_{Q_i}, & 1\leqslant i<K,\\ \text{EndTime}, & i=K,\end{cases}\tag{3}$$

$$\text{EndTime}=\begin{cases}10\ (\text{s}) & (Q_K\text{没有点击URL}),\\ 60\ (\text{s}) & (Q_K\text{点击了URL}),\end{cases}$$

查询 Q_{kj} 在会话 S_k 中的平均持续时间比率为 $\text{TimeRate}_{Q_{k_j}}$:

$$\text{TimeRate}_{Q_{k_j}}=\frac{\overline{\text{QueryDuration}_{Q_{k_j}}}}{\text{SessionTime}_{S_k}},\tag{4}$$

$$\overline{\text{QueryDuration}_{Q_{k_j}}}=\frac{\sum^{\text{Fre}_{Q_{k_j}}}\text{QueryDuration}_{Q_{k_j}}}{\text{Fre}_{Q_{k_j}}},\tag{5}$$

其中，$\overline{\text{QueryDuration}_{Q_{k_j}}}$ 表示查询 Q_{kj} 在会话 S_k 中的平均持续时间；$\sum^{\text{Fre}_{Q_{k_j}}}\text{QueryDuration}_{Q_{kj}}$ 表示查询 Q_{k_j} 在会话 S_k 中持续的总时间，SessionTime_{S_k} 表示会话 S_k 的持续总时间，是会话开始时间与结束时间之差。

2.2.3　查询点击 URL 的平均排名倒数

查询 Q_{kj} 在会话 S_k 中点击 URL 的平均排名倒数 $\text{AverRank}_{Q_{k_j}}$ 计算如下：

$$\text{AverRank}_{Q_{k_j}}=\frac{\text{Fre}_{Q_{k_j}}}{\sum^{\text{Fre}_{Q_{kj}}}\text{Rank}_{Q_{k_j}}},\tag{6}$$

其中，$\text{Rank}_{Q_{k_j}}$ 表示查询 Q_{kj} 对应每一次点击的 URL 排名，$\sum^{\text{Fre}_{Q_{k_j}}}\text{Rank}_{Q_{k_j}}$ 表示查询 Q_{k_j} 在会话 S_k 中点击的

URL 的排名总和。所以用户对会话 S_k 中的查询 Q_{k_j} 的兴趣度为

$$W_{Q_{k_j}} = \alpha \cdot \mathrm{FreRate}_{Q_{k_j}} + \beta \cdot \mathrm{TimeRate}_{Q_{k_j}} + \gamma \cdot \mathrm{AverRank}_{Q_{k_j}}, \quad (7)$$

其中，$\alpha+\beta+\gamma=1$，$0 \leqslant \alpha \leqslant 1$，$0 \leqslant \beta \leqslant 1$，$0 \leqslant \gamma \leqslant 1$。

2.2.4 兴趣投票

通过查询加权的方法，得到用户对会话 S_k 中查询 Q_{kj} 的兴趣度。首先，对每个查询进行预处理：按空格分词，去除停用词，去除噪声词，采用 Porter 提取词干的方法提取英文词干。然后，采用权重投票的形式，计算用户对查询日志中关键词 T_i 的兴趣度 W_{T_i}：

$$W_{T_i} = \mathrm{Vote}\,(T_i) = \sum_{k}^{K_i}\sum_{j}^{N_{ki}} (W_{Q_{k_j}} \times F_{ij}), \quad (4)$$

其中，F_{i_j} 表示关键词 T_i 在查询 Q_{k_j} 中出现的次数；K_i 表示在查询日志中有 K_i 个会话包含关键词 T_i；N_{ki} 表示在会话 S_k 的 N_{ki} 个查询中包含关键词 T_i。

3 实验设计及结果分析

3.1 实验语料

本文采用 AOL 美国在线查询日志数据(http://www.datatang.com/data/42724)，该查询日志包括 657426 个匿名用户在 2006–03–01—2006–05–31 共 10154742 条查询记录，每一条记录表示为{UserID, Query, QueryTime, Rank, ClickURL}的形式，其中，UserID 表示匿名用户的 ID；Query 表示用户提交的查询；QueryTime 表示用户提交查询的时间；Rank 表示用户点击返回结果的排名；ClickURL 表示用户点击的结果的 URL。

实验中确保每个用户训练集和测试集的查询条数大于 20，筛选出 376 个用户的日志记录作为实验数据。将每个用户前 2.5 个月(2006–03–01—2006–5–15)的查询日志作为训练数据集，剩余时间(2006–05–16—2006–05–31)的查询日志作为测试数据集。

3.2 评价方法及指标

本文通过训练数据集对用户的查询日志进行分析，采用基于查询加权的用户建模方法，将用户模型表示成关键词集合的形式，完成用户模型的构建。利用测试数据集对用户模型进行测试：首先，对测试数据集的查询日志进行预处理，如分词、去除停用词、去除噪声词以及 Porter 提取英文词干；然后，将测试数据集表示为词向量的形式，每个用户对应一个词向量，将该词向量作为用户的真实兴趣；最后利用用户的真实兴趣评价用户建模得到的用户模型的好坏。

本文采用 MeanP(Mean Precision)和 MAP(Mean Average Precision)来评价用户模型对用户行为的预测能力。MeanP 表示用户预测准确率的平均值，是对所有用户的预测命中率求平均，MeanP 越高表示用户模型预测准确度越高，计算如式(9)；MAP 表示每个用户的平均准确率的平均值，是对所有用户平均准确率(AP，预测兴趣中在每个真实兴趣的位置上的正确率的平均值)求宏平均，MAP 越高表示预测成功的兴趣排名越靠前，计算如式(10)。

$$\mathrm{MeanP} = \frac{1}{|U|}\sum_{u}^{|U|}\frac{\mathrm{Pre\ Num}_u}{M_u}, \quad (9)$$

$$\mathrm{MAP} = \frac{1}{|U|}\sum_{u}^{|U|}\frac{1}{N_u}\sum_{m}^{N_u}\mathrm{Precision}(R_{um}), \quad (10)$$

其中，U 为用户集合；PreNum_u 表示对用户 u 预测正确兴趣的数目；M_u 表示对用户 u 预测兴趣的总数目；N_u 表示用户 u 真实兴趣的数目；用户 $u \in U$ 对应的真实兴趣为 $\{T_1, T_2, ..., T_{N_u}\}$，$R_{um}$ 表示预测兴趣中直到遇见 T_m 后所在位置前(包含 T_m)的所有真实兴趣集合；$\mathrm{Precision}(R_{um})$ 表示集合 R_{um} 的准确率。

3.3 参数估计

在查询加权的过程中，需要估计参数 α，γ，β 的值，保证用户模型最优。本文列举满足 $\alpha+\beta+\gamma=1$ 的所有 α, β, γ 的值，并设置步长为 0.1，得到它们对应的 MeanP 和 MAP 值。图 5 中，横坐标表示总共进行实验的次数，即对应 α, β, γ 值共 65 组，纵坐标表示对应的 MeanP 值。当 $\alpha=0.4, \beta=0.3, \gamma=0.3$ 时，MeanP 取得最大值。

另外，为了验证 3 个特征的作用效果，本文进行了 7 组实验，如表 2 所示。第 4，5，6 和 7 组实验取 MeanP 值最大时的结果。实验 1，2 和 3 是 3 个特征单独的作用效果，实验 4，5 和 6 是 3 个特征两两组合的效果，实验 7 是 3 个特征同时作用的效果。由表 2 可以看出，两两组合的特征选择作用效果比任意特征单独作用的效果好，并且 3 个特征同时作用的效果比两两组合的特征选择作用效果好。通过以上分析，本文取 3 个特征同时作用，即 α =

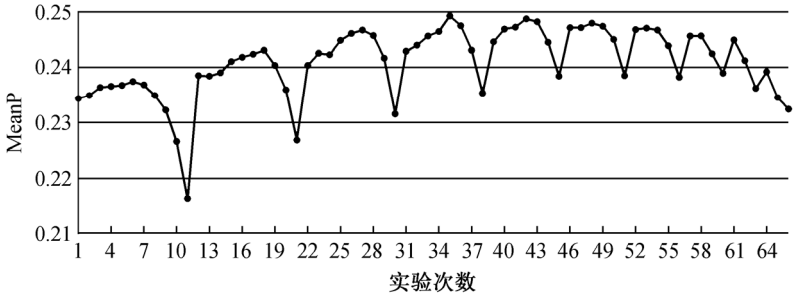

图 5　每组α, β, γ 对应的 MeanP 值

Fig. 5　Value of MeanP corresponding each group of α, β, γ

表 2　特征选择及对应的 MeanP, MAP 值

Table 2　Values of MeanP and MAP corresponding different feature selections

实验编号	特征选择	MeanP	MAP
1	FreRate (α = 1.0, β = 0.0, γ = 0.0)	0.23237	0.26883
2	TimeRate　(α = 0.0, β = 1.0, γ = 0.0)	0.21619	0.25081
3	AveRank (α = 0.0, β = 0.0, γ = 1.0)	0.23432	0.27096
4	TimeRate & AveRank (α = 0.0, β = 0.5, γ = 0.5)	0.23742	0.27569
5	FreRate&AveRank (α = 0.5, β = 0.0, γ = 0.5)	0.24714	0.28657
6	FreRate&TimeRate (α = 0.7, β = 0.3, γ = 0.0)	0.23892	0.27581
7	FreRate&TimeRate&AveRRank (α = 0.4, β = 0.3, γ = 0.3)	**0.24873**	**0.28844**

0.4, β = 0.3, γ = 0.3 时的结果作为用户模型。

3.4　实验结果和分析

通过用户的查询日志对用户进行建模, 本质上是通过用户历史记录来预测用户未来行为的过程。本实验分别采用 3 种方法在 AOL 查询日志数据集上进行对比分析。

1) 采用传统的 TF-IDF 方法对查询日志进行建模, 将训练数据集中一个用户的查询日志看作一篇文档, 对文档内的每个关键词计算 TF-IDF 值, 得到关键词及其权重集合, 作为用户模型(TF-IDF)。

2) 采用文献[6]中的方法, 以项目的评分作为二部图中用户与项目的边权, 采用发散理论, 利用用户的历史行为, 按照用户–项目边权占该节点权重和的比例分配资源, 充分利用用户之间的相关信息, 得到按用户的项目评分高低排序的集合, 作为用户模型(Diffusion_based)。

3) 为本文第 2 节提出的查询加权的方法(Query_weighted)。

如图 6 所示, 基于查询加权的方法在用户平均准确率上比 TF-IDF 的方法提高 5.1%, 比基于扩散的预测方法提高 1.6%, 表明本文提出的方法能更加准确的预测用户的兴趣。比较 3 种方法的 MAP 值, 本文方法比 TF-IDF 方法提高 6.2%, 比基于扩散的预测方法提高 2.1%, 表明考虑用户搜索行为能很大程度地提高预测结果, 并且预测成功的兴趣排名更加靠前。与其他两种方法对比, 本文提出的方法在用户建模问题上更加有效。

以上实验都是在用户模型取前 30 位关键词时得到的结果。本文同时也观察了用户模型取不同的关键词个数对结果的影响。如图 7 所示, 横坐标表示取兴趣度排在前 K 位的关键词作为用户模型, 纵坐标表示对应的 MeanP 值。由图 7 可知, 返回不同数目的兴趣时, 基于查询加权的方法均优于 TF-

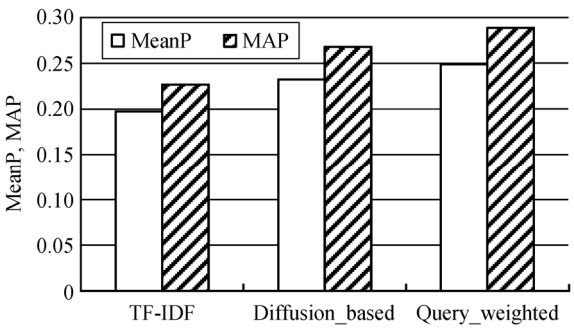

图 6　三种方法的 MeanP 和 MAP 值比较

Fig. 6　Comparison of MeanP and MAP of three methods

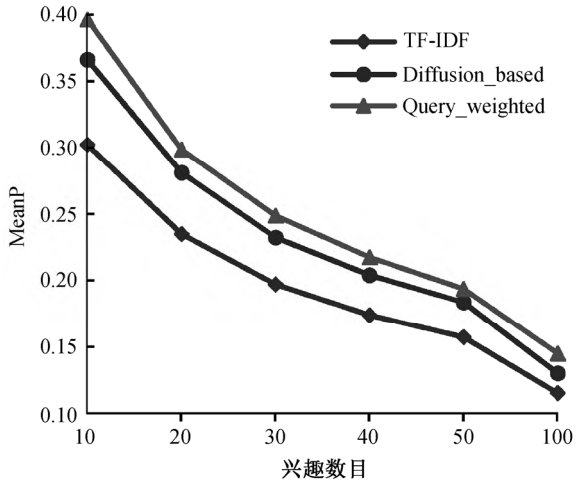

图 7　不同用户模型兴趣数目对应的 MeanP 值

Fig. 7　Value of MeanP corresponding different numbers of user model interests

IDF 方法和基于扩散的预测方法, 说明本文方法对于用户兴趣预测的有效性和稳定性。另外, 对于 3 种方法的运行时间进行了比较, 如表 3 所示, 可以看到基于查询加权的用户建模方法明显比其他两种方法的效率要高。

通过以上分析, 由本文方法构建的用户模型更加满足用户的兴趣爱好, 能够更好反映用户的行为, 并且在算法效率上较其他两种方法也有显著的提高, 取得较好的预测效果。

4　结语

本文提出了一种基于查询加权的用户建模方法, 将用户的查询日志分割成若干会话, 针对每个会话, 考虑其中每个查询的出现次数、持续时间以及点击 URL 的排名等信息, 模拟用户使用搜索引擎进行会话的过程。同时提出了 3 个假设, 对查询进行加权, 最后利用权重投票的方式, 得到用户的兴趣模型。实验结果表明: 该方法在预测准确率和 MAP 值上较 TF-IDF 方法和基于扩散的预测方法均有明显的提升。但是, 由于该方法只是考虑了单个用户的相关信息, 是对单个用户内部进行建模, 并没有将用户之间的关系考虑进去, 因此, 下一步工作将考虑把用户之间的信息融合到该方法中, 以期取得更好的效果。

表 3　特征选择及对应的 MeanP, MAP 值 3 种方法的运行时间比较

Table 3　Running time comparison of three methods

方法	运行时间/s
TF-IDF	115
Diffusion_based	196
Query_weighted	14

参考文献

[1] 许海玲, 吴潇, 李晓东, 等. 互联网推荐系统比较研究. 软件学报, 2009, 20(2): 350–362

[2] 吴丽花,刘鲁.个性化推荐系统用户建模技术综述.情报学报, 2006, 25(1): 55–62

[3] 岑荣伟,刘奕群,张敏,等.基于日志挖掘的搜索引擎用户行为分析.中文信息学报, 2010, 24(3): 49–54

[4] Chen J, Nairn R, Nelson L, et al. Short and tweet: experiments on recommending content from information streams // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Atlanta: ACM, 2010: 1185–1194

[5] Matthijs N, Radlinski F. Personalizing web search using long term browsing history // Proceedings of the Fourth ACM International Conference On Web Search And Data Mining. Sydney: ACM, 2011: 25–34

[6] 张新猛, 蒋盛益. 基于加权二部图的个性化推荐算法.计算机应用, 2012, 32(3): 654–657

[7] Liu F, Yu C, Meng W. Personalized web search by mapping user queries to categories // Proceedings of the Eleventh International Conference on Information and Knowledge Management. New York: ACM, 2002: 558–565

[8] Tian X, Du X, Hu H, et al. Modeling individual cognitive structure in contextual information retrieval. Computers & Mathematics with Applications, 2009, 57(6): 1048–1056

[9] Sontag D, Collins-Thompson K, Bennett P N, et al. Probabilistic models for personalizing web search // Proceedings of the Fifth ACM International Conference on Web Search And Data Mining. Seattle: ACM, 2012: 433–442

[10] Iwata T, Watanabe S, Yamada T, et al. Topic tracking model for analyzing consumer purchase behavior // IJCAI. Pasadena: 2009, 9: 1427–1432

[11] Harvey M, Crestani F, Carman M J. Building user profiles from topic models for personalised search // Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. New York: ACM, 2013:

2309–2314

[12] 余慧佳, 刘奕群, 张敏, 等. 基于大规模日志分析的搜索引擎用户行为分析.中文信息学报, 2007, 21(1): 109–114

[13] Piwowarski B, Dupret G, Jones R. Mining user web search activity with layered bayesian networks or how to capture a click in its context // Proceedings of the Second ACM International Conference on Web Search and Data Mining. New York: ACM, 2009: 162–171