融合主题模型及多时间节点函数的用户兴趣预测研究*

桂思思[1]　陆　伟[1,2]　黄诗豪[1]　周鹏程[1]
[1](武汉大学信息管理学院　武汉 430072)
[2](武汉大学信息资源研究中心　武汉 430072)

摘要: 【目的】针对用户兴趣随时间推移不断变化的问题, 利用主题模型及时间节点函数预测用户兴趣。【方法】使用主题模型生成用户兴趣, 针对用户的所有兴趣, 分别利用多时间节点函数对每个兴趣的每次出现进行加权, 用以预测用户兴趣在下一个时间节点的分布情况。【结果】在 Sogou 搜索日志上, 与基于记忆的用户兴趣模型、基于遗忘曲线的用户兴趣度多阶段量化模型进行对比实验, 余弦相似度及 KL(Kullback-Leibler)距离均表明本文方法能较准确地预测用户兴趣。【局限】仅在 Sogou 搜索日志上进行实验测试, 还需在其他数据集上进一步检验。【结论】充分考虑用户历史数据中每一个时间点可更准确地对用户兴趣进行预测。

关键词: 主题模型　时间函数　用户兴趣　预测
分类号: TP393

1　引　言

用户兴趣建模是从能够体现用户兴趣偏好的信息(如浏览行为、浏览内容、知识背景等)中归纳出可计算的用户兴趣模型的过程[1]。用户兴趣建模能为个性化信息服务提供可信赖的用户信息, 有效改善个性化信息服务的服务质量, 是个性化信息服务有效开展的基础与核心。

用户兴趣可根据用户行为进行预测, 例如协同过滤(Collaborative Filtering)利用用户自身或相似用户的历史行为预测用户兴趣。但是用户兴趣并非一成不变, 而是随着时间的推移不断变化。协同过滤模型更偏重用户间或资源间的相似性, 却忽略了用户兴趣随时间变化的过程。时间窗口模型和遗忘模型能够反映用户的兴趣变化[2]。但是时间窗口模型不仅不能反映用户兴趣衰减, 还容易忽略窗口之外的数据; 遗忘模型虽能合理利用历史数据, 但只考虑了历史记录中的初始时间点、最近时间点数据。

本文认为用户的某类兴趣在某个时间点出现后, 会对当前时间点以及下一个时间节点的该类兴趣权重的计算产生影响, 因此预测用户兴趣时需考虑历史记录中每一个时间点的用户兴趣数据。在这个假设前提下, 本文提出一个新的用户兴趣预测模型, 该模型使用主题模型生成用户兴趣, 随后针对用户所有兴趣, 分别利用多时间节点函数对每个兴趣的每次出现进行加权, 用于预测下一个时间节点的用户兴趣分布情况。利用该模型在 Sogou 实验室发布的用户查询需求日志上进行实验, 结果表明本文提出的模型具有较好的预测效果, 优于其他模型。

2　相关研究

2.1　基于主题模型的用户兴趣表示

主题模型在用户兴趣表示上已有部分研究成果。Ahmed 等[3]认为用户兴趣是主题的集合, 用户检索时

通讯作者: 陆伟, ORCID: 0000-0002-0929-7416, E-mail: weilu@whu.edu.cn。

*本文系教育部人文社会科学基地重大项目“面向细粒度的网络信息检索模型及框架构建研究”(项目编号:10JJD630014)和国家自然科学基金面上项目“面向词汇功能的学术文本语义识别与知识图谱构建”(项目编号: 71473183)的研究成果之一。

先确定主题，再选用能表示该主题的查询词进行检索，并在此基础上提出基于主题模型的用户模型构建框架。Veningston 等[4]在研究个性化检索问题时，认为用户兴趣可以表示为用户 u 提交查询 q 时检索主题 T 的概率分布，并认为主题模型是一个较好的实现工具。在研究基于用户兴趣的推荐系统时，Sakamoto 等[5]、Pennacchiotti 等[6]、Liu 等[7]及 Mao 等[8]采用<用户–项目–兴趣>用户兴趣三层表示模型，并将其与主题模型中的<文档–词–主题>相对应，使用主题模型抽取用户兴趣。

与传统聚类方式相比，主题模型有较好的聚类效果。Ding 等[9]从主题识别及主题演化两个维度将主题模型法与基于词共现、基于引文共现等传统聚类算法相比较，综合比较后发现主题模型法在主题识别及主题演化方面优于其他两种聚类方法。

前人的研究成果表明主题模型可用于用户兴趣的表示、生成，且在划分用户兴趣时，优于一般聚类方法。

2.2 用户兴趣预测方法研究

用户兴趣预测可采用协同过滤的思想，但是基于传统协同过滤的用户兴趣预测方法更偏重用户间或资源间的相似性，容易忽略用户兴趣随时间动态变化的过程。为了准确地预测用户兴趣，必须考虑时间因素的影响。Lee 等[10-11]以移动电子商务推荐系统为例，考虑用户购买时间、评论时间、商品上线时间、上述时间的时间差以及各种时间组合，并证明考虑时间因素能有效提高推荐准确度。

基于协同过滤的个人兴趣预测改进法是利用时间因素，描述用户兴趣动态变化的过程，对协同过滤得到的用户兴趣结果集合进行加权排序。常见的方法有时间窗口法[12-13]、遗忘模型法以及混合模型法[14-15]。时间窗口法容易忽略时间窗口之外的历史数据，这些窗口外的数据也可能反映出用户的一些常规需求，不应随意剔除。Maloof 等[16]针对该问题，专门探讨了历史数据的选择问题。遗忘模型法认为用户兴趣衰减与自然遗忘规律相似，提出一个用于模拟用户兴趣遗忘规律的时间函数[17-18]。Chen 等[19]虽未在文中明确提出一个时间函数，但采用遗忘模型法思想，根据用户的评分时间在[0,1]取值并分段赋值；其他研究者常选用一个单调递减函数作为时间函数，例如指数函数[20-22]、逻辑函数[23]、线性函数[24]、幂函数[25]、复合函数[26]等。利用时间函数加权重主要有以下方式:

(1) Zhang 等[21]、于洪等[25]利用项目初始评分时间点与最后一次评分时间点之间的时间差；

(2) Chen 等[19]、邢春晓等[24]、Wu 等[26]利用项目初始评分时间点与整个时间段的最后时间点之间的时间差；

(3) Karahodza 等[22]、Wang 等[23]在上述基础上还区分了不同用户对同一个项目评分的差异性，利用某项目最后一次被评分时间点与该项目被某单个用户最后一次评分时间点之间的时间差。

不考虑协同过滤，单纯利用时间因素描述用户兴趣变化过程的研究成果相对较少。例如，Liu 等[27]、Cheng 等[28]选用指数函数作为时间函数，利用遗忘模型描述博客上用户兴趣演化过程。Rybak 等[29]选用线性函数作为时间函数，描述一段时间内专家专长的变化情况。Wu 等[30]提出基于记忆的用户模型(Memory-based User Profile)，简记为 Memory-UP，该模型同时考虑了用户学习、遗忘的过程，并以在线新闻网站用户日志中的点击数据为例，对用户兴趣进行预测[31]。于洪涛等[32]提出基于遗忘曲线的用户兴趣度多阶段量化模型(简记为 Multi-Step-UP)，该模型把整个时间段分成多个阶段，认为每一个阶段都是一个新的遗忘过程，并在腾讯微博数据上验证模型的预测效果。

由上述分析可知，基于协同过滤法的用户兴趣预测在考虑时间因素时，仅考虑了项目的初始评分时间点、最后评分时间点以及整个时间段的最后时间点，忽略了项目初始评分时间点与整个时间段的最后时间点之间的其他时间点；不考虑协同过滤的用户兴趣预测法研究成果较少。本文在上述研究成果的基础上，利用遗忘模型思想，提出一个多时间节点函数对兴趣的每次出现进行加权并预测，并在用户日志中的查询数据集上与 Memory-UP 模型和 Multi-Step-UP 模型效果进行对比。

3 研究方法

本文使用主题模型生成用户兴趣，针对用户所有兴趣，分别利用多时间节点函数对每个兴趣的每次出现进行加权，用于预测下一个时间节点的用户兴趣分布情况。

3.1　基于主题模型的用户兴趣模型构建

网络日志(Web Log)记载了用户访问某网站的完整记录，包括大量用户行为以及用户 IP、访问时间等数据，这些数据可潜在反映用户兴趣。

为方便实验，本文不区分单个用户的兴趣，将全体用户的兴趣作为研究对象，因此查询日志中的记录可简化为如下形式，表示在 $time_i$ 时，用户向搜索引擎提交查询 $query_i$。

$$Log=<time_1:query_1,\ time_2:query_2,\ \cdots,\ time_n:query_n>$$

本文认为用户向搜索引擎提交的查询词是用户兴趣的表现，查询词构成的潜在主题集合 Z 是用户真正的兴趣，该主题需要使用主题模型生成。主题模型是一种用来发现文档集合中隐含主题的统计模型，常见的有 PLSI[33]与 LDA[34]，它认为文档集合中的每篇文档是由多个主题按照一定比例组合而成的，且每个主题可以表示为词表中词的分布。

对用户查询记录而言，每一个查询(query)均由不同的查询词(word)构成。可以将本文中查询、查询词、兴趣映射成主题模型中文档、词、主题，即：$P(w|q)=P(w|z)P(z|q)$，其中$P(w|q)$为查询 q 关于查询词 w 的分布、$P(w|z)$为用户兴趣 z 关于查询词 w 的分布、$P(z|q)$为查询 q 关于用户兴趣 z 的分布。

由于 LDA 相比 LSI 与 PLSI 而言，具有较好的建模能力及相对较低的计算复杂度[35]，因此使用 LDA 求得$P(w|z)$。

3.2　基于多时间节点函数的兴趣预测

(1) 时间函数

由 2.2 节可知，指数函数、逻辑函数、线性函数、幂函数、复合函数等均可作为时间函数，但是以函数本身的变化趋势而言，指数函数优于逻辑函数[36]，所以本文的多时间节点函数如下：

$$f_{word_i}(\tau,Z_j)=e^{-\lambda_{Z_j}\tau}\qquad \tau=t_n-t_{n-1} \tag{1}$$

公式(1)表示属于主题 Z_j 的查询词 $word_i$ 的时间函数，它随τ的增大而逐渐降低，其取值范围为(0,1]。τ为时间差，是$word_i$第 n 次与第n–1次出现的时间差。λ是遗忘因子，表示用户对某主题失去兴趣的快慢程度，在此处，λ_{Z_j}表示用户对主题 Z_j 失去兴趣的快慢程度：λ_{Z_j}越大，函数图像变化得越快，表示用户对主题Z_j遗忘得越快，即失去兴趣的速率越快；否则反之。

(2) 遗忘因子

遗忘因子λ对刻画用户兴趣度尤为关键。本文以全体用户为研究对象，探讨整体用户的兴趣变化情况，涉及资源范围领域广，确定半衰期较为困难，故在 Zhang 等[21]提出的遗忘因子计算方法的基础上做出相应改进：对于主题Z_j，遗忘因子λ_{Z_j}计算公式如下：

$$\lambda_{Z_j}=a^{\left(\frac{N_{Z_j}}{N}\right)^{\frac{1}{m}}}\qquad 0<a<1 \tag{2}$$

其中，N_{Z_j}表示用户在某个时间段内查询属于主题 Z_j 的查询词个数，N 表示用户在该时间段内查询所有查询词的个数；a, m 均为参数。λ_{Z_j}是一个关于N_{Z_j}的减函数，当N_{Z_j}增大时(用户对主题Z_j查询次数增多)，λ_{Z_j}减小，因此时间函数变化较平缓(用户对其兴趣保持相对水平，不会很快减弱)。为了保证通过时间函数计算后多数查询词权重不为 0，且分布相对散开，在粗略尝试后，a, m 初始取值如下：a=0.38, m=100。

(3) 查询词权重计算

研究一段时间中某个时间点状态时，需综合考虑该时间点之前的情况以及该时间点新增的情况。该思想在 Rybak 等[29]以及 Wu 等[30]的论文中采用过。

本文认为用户每向搜索引擎提交一次查询，都会改变该查询词在当前时间点的权重。因此，在考虑某个时间点查询词权重时，需要综合考虑该时间点之前该查询词的权重，以及该时间点因查询操作而新产生的权重。某个时间点的查询词权重计算公式如下：

$$w^{(n)}_{word_i}=f^{(n)}_{word_i}(\tau,Z_j)+g^{(n-1)}_{word_i} \tag{3}$$

$$f^{(n)}_{word_i}(\tau,Z_j)=e^{-\lambda_{Z_j}\tau}\qquad \tau=t_n-t_{n-1} \tag{4}$$

$$g^{(n-1)}_{word_i}=\begin{cases}\dfrac{1}{n-1}\displaystyle\sum_{k=1}^{n-1}w^{(k)}_{word_i} & n\geq 2\\ 0 & n=1\end{cases} \tag{5}$$

w_{word_i}表示某个时间点查询词$word_i$的权重，$w^{(n)}_{word_i}$表示查询词$word_i$第n次查询时在该时间点的权重，$f^{(n)}_{word_i}(\tau,Z_j)$是因第 n 次查询操作而新产生的权重，$g^{(n-1)}_{word_i}$是前n–1次查询操作对本查询词$word_i$的累计权重。

当$word_i$第一次出现时(n=1)，$w^{(1)}_{word_i}=1$，即认为在此时间点的权重为 1；当$word_i$第二次出现时(n=2)，相当于用户第二次查询该词，此时权重应在第一次查询权重的基础上加上本次查询产生的新权重，即：

$$g_{word_i}^{(1)}=\frac{1}{2-1}\sum_{k=1}^{1} w_{word_i}^{(k)}=w_{word_i}^{(1)}=1$$

$$w_{word_i}^{(2)}=f_{word_i}^{(2)}(\tau,Z_j)+1 \qquad \tau=t_n-t_{n-1}$$

(4) 主题权重计算

本文认为文本主题可用加权树表示，查询词为叶子节点，主题为非叶子节点，每一个主题(非叶子节点)可划分出多个查询词(叶子节点)。对于主题权重的计算方法可以借鉴专家专长研究[29]的计算方法:

若主题Z_j内共有 m 个查询词，则兴趣主题Z_j的得分为:

$$Score_{Z_j}=\sum_{i=1}^{m} w_{word_i}^{(n)} \tag{6}$$

即兴趣主题Z_j的最终预测得分等于属于该主题的所有查询词的权重之和。为了保证评测的可比性，主题Z_j最终得分为依据所有主题的得分和归一化后的值，即:

$$Score_{Z_j}^{Nor}=\frac{Score_{Z_j}}{\sum Score_{Z_j}} \tag{7}$$

4 实验以及结果分析

4.1 数据获取与预处理

为了验证本文模型预测的精准性，从 Sogou 实验室①获取 2008 年 6 月 1 日至 2008 年 6 月 29 日(无 6 月 10 日)共 28 天的搜索日志，并从原始数据集非空记录中抽取“访问时间 \t 用户 ID \t [查询词]”三项信息，共计 51 537 394 条。

利用 ICTCLAS②(2014 年版)对 Sogou 日志中的用户查询词进行分词。为了保证分词质量，笔者根据该工具的分词结果，结合人工判断，新增 657 个新词至用户词典。重新对 Sogou 日志中用户查询词进行分词。

4.2 构建主题模型

本实验利用主题模型工具 MALLET(MAchine Learning for LanguagE Toolkit)③生成主题模型。使用主题模型必须提前确定主题数，虽然常使用困惑度(Perplexity)评判主题数的最佳取值[37]，但是本实验的关注点在于划分出用户兴趣的类别，而不在于兴趣类别划分的精确性，因此不检验困惑度。主题数常设置为 100[37]，故实验中主题数设定为 100。

实验中将 Sogou 分词后的文本作为输入文件，利用 MALLET 自带的 LDA 算法构建主题模型，最后的输出文件格式为: “doc \t source \t pos \t typeindex \t type \t topic”，即记录了每一个词的原始位置以及所属主题的编号。利用主题模型时，可能出现同一个词属于不同主题。对于该问题，MALLET 在生成主题模型时，已经计算过同一个词属于不同主题的概率P(w|z)，并在这个概率的基础上，将相同的查询词划分到不同的主题中。

4.3 实验评价方式

实验原始数据总时长为 28 天: 取前 21 天数据作为训练数据(Training Data)，用以预测后 7 天(测试数据, Test Data)每一天的用户兴趣分布。

对于测试数据，采用词频统计方法计算每一天用户兴趣的分布情况，并将其作为真实用户兴趣分布: 统计属于某一个主题所有查询词的词频，除以所有主题的查询词的词频，进行归一化处理，从而求得真实的用户兴趣分布情况。测试数据共有 7 天，因此可计算主题分布相似性 7 次。

余弦相似度(Cosine Similarity)及KL距离(Kullback-Leibler Divergence)[38]是用来计算两个主题分布相似度的常用方法[39-40]。本文同时使用这两种方法计算预测的兴趣分布与真实兴趣分布之间的相似度: KL 距离的值恒不为负，值越小，表示两个分布越接近，即预测的结果越准确; 余弦相似度取值范围为[0,1]，值越大，表示两个分布越接近，即预测结果越准确。

为了体现本模型(Multi-Time-UP)的有效性，选取 Memory-UP、Multi-Step-UP 进行对比，并利用双尾 T 检验对不同模型的预测结果之间是否存在显著差异进行检验。

(1) Memory-UP[30]: 该模型较好地模拟了用户学习、遗忘等过程，利用用户日志中的点击数据预测用户兴趣;

(2) Multi-Step-UP[32]: 该模型考虑时间因素，把

① http://www.sogou.com/labs/dl/q.html/.
② http://ictclas.nlpir.org/.
③ http://mallet.cs.umass. edu/.

整个时间段分成多个阶段，认为每一个阶段都是一个新的遗忘过程，与本文思路有类似之处。相关参数取文献[32]中默认值。

4.4　实验结果与分析

Multi-Time-UP、Memory-UP、Multi-Step-UP 三个模型预测的用户兴趣分布与真实用户兴趣分布的 KL 距离与余弦相似度如表 1 所示。其中，组号为月和日组成的 4 位数字，如 0623 表示 6 月 23 日。表 2 为三个模型 KL 距离差异的显著性检验结果。表 3 为三个模型余弦相似性差异的显著性检验结果。

表 1　三个模型预测的用户兴趣分布与真实用户兴趣分布的 KL 距离与余弦相似度

组号	Multi-Time-UP		Memory-UP		Multi-Step-UP	
	KL 距离	余弦相似度	KL 距离	余弦相似度	KL 距离	余弦相似度
0623	0.1400	0.8727	0.7215	0.5751	0.9243	0.6833
0624	0.2268	0.7919	0.7482	0.5622	1.1499	0.6012
0625	0.3417	0.6867	0.8691	0.4885	1.1221	0.5739
0626	0.2936	0.7335	0.8235	0.5135	1.0488	0.6137
0627	0.1978	0.8238	0.6977	0.5952	1.0293	0.6459
0628	0.1693	0.8532	0.6259	0.6599	1.0455	0.6466
0629	0.1526	0.8583	0.6403	0.6170	1.0183	0.6290
平均值	0.2174	0.8029	0.7323	0.5731	1.0483	0.6277

表 2　三个模型 KL 距离差异的显著性检验

	Multi-Time-UP	Memory-UP	Multi-Step-UP
Multi-Time-UP	–	0.0000	0.0000
Memory-UP	–	–	0.0001
Multi-Step-UP	–	–	–

表 3　三个模型余弦相似性差异的显著性检验

	Multi-Time-UP	Memory-UP	Multi-Step-UP
Multi-Time-UP	–	0.0000	0.0000
Memory-UP	–	–	0.0197
Multi-Step-UP	–	–	–

根据表 2 和表 3 可知，三个模型的实验结果均有显著差异(<0.05)。由表 1 可知，在三个模型中，Multi-Time-UP 的 KL 距离(平均值为 0.2174)普遍小于其他两个模型的结果，余弦值相似性(平均值为 0.8029)普遍大于其他两个模型的结果，用户兴趣预测效果最优。

就 KL 距离的结果而言，Multi-Step-UP 的 KL 距离(平均值为 1.0483)普遍大于其他两个模型的结果，用户兴趣预测效果最差；Memory-UP 的预测效果居于 Multi-Time-UP 与 Multi-UP 之间。Multi-Time-UP 预测的兴趣主题分布与真实兴趣分布的 KL 距离最低可达 0.1400，KL 距离最高值 0.3417 也比其他两个模型的 KL 最低值小。就余弦相似性的结果而言，Multi-Step-UP 的预测效果次之(平均值为 0.6277)；Memory-UP 预测效果相对而言不太理想(平均值为 0.5731)。

总结来看，以KL距离评估模型准确性，预测准确性排序为：Multi-Time-UP、Memory-UP、Multi-Step-UP；以余弦相似性评估模型准确性，预测准确性排序为：Multi-Time-UP、Multi-Step-UP、Memory-UP。因此，本文模型较其他模型具有更好的预测效果。

5　结　语

用户兴趣随着时间推移不断改变，本文提出一种新的用户兴趣动态预测模型，该模型利用多时间节点函数充分考虑了用户历史数据中每一个时间点的历史数据。实验结果表明，与基于记忆的用户兴趣模型、基于遗忘曲线的用户兴趣度多阶段量化模型相比，本文模型能较准确地实现用户兴趣的动态预测，说明预测用户兴趣时需考虑历史记录中每一个时间点的用户兴趣数据。然而本文研究也有一定的局限性：研究对象为集体用户兴趣，而非个体用户兴趣；数据时间跨度较小。今后的研究方向包括：将该方法应用于个体用户兴趣研究；尝试将该方法应用于分析用户兴趣周期上的可行性；尝试用户兴趣动态预测的相关应用，结合实际问题探讨模型的适用性。

参考文献：

[1]　冯子威. 用户兴趣建模的研究[D]. 哈尔滨：哈尔滨工业大

学, 2010. (Feng Ziwei. Research on User Interests Modeling [D]. Harbin: Harbin Institute of Technology, 2010.)

[2] 杨杰, 陈恩红. 面向个性化服务的用户兴趣偏移检测及处理方法[J]. 电子技术, 2009(11): 72-76, 63. (Yang Jie, Chen Enhong. Personalized Service Oriented User Interest Shift Detection and Processing [J]. Electronic Technology, 2009(11): 72-76, 63.)

[3] Ahmed A, Low Y, Aly M, et al. Scalable Distributed Inference of Dynamic User Interests for Behavioral Targeting [C]. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011: 114-122.

[4] Veningston K, Shanmugalakshmi R. Combining User Interested Topic and Document Topic for Personalized Information Retrieval [A]. //Big Data Analytics [M]. Springer International Publishing, 2014: 60-79.

[5] Sakamoto S, Mikawa K, Goto M. A Study on Recommender System Based on Latent Class Model for High Dimensional and Sparse Data [C]. In: Proceedings of the 14th Asia Pacific Industrial Engineering and Management Society Conference, Cebu, Philippines. 2013.

[6] Pennacchiotti M, Gurumurthy S. Investigating Topic Models for Social Media User Recommendation [C]. In: Proceedings of the 20th International Conference Companion on World Wide Web. ACM, 2011: 101-102.

[7] Liu Q, Chen E H, Xiong H, et al. Enhancing Collaborative Filtering by User Interest Expansion via Personalized Ranking [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012, 42(1): 218-233.

[8] Mao Q, Feng B, Pan S. Modeling User Interests Using Topic Model [J]. Journal of Theoretical and Applied Information Technology, 2013, 48(1): 600-606.

[9] Ding W, Chen C. Dynamic Topic Detection and Tracking: A Comparison of HDP, C-word, and Cocitation Methods [J]. Journal of the Association for Information Science and Technology, 2014, 65(10): 2084-2097.

[10] Lee T Q, Park Y, Park Y T. A Time-Based Approach to Effective Recommender Systems Using Implicit Feedback [J]. Expert Systems with Applications, 2008, 34(4): 3055-3062.

[11] Lee T Q, Park Y, Park Y T. An Empirical Study on Effectiveness of Temporal Information as Implicit Ratings [J]. Expert Systems with Applications, 2009, 36(2): 1315-1321.

[12] Widmer G, Kubat M. Learning in the Presence of Concept Drift and Hidden Contexts [J]. Machine Learning, 1996, 23(1): 69-101.

[13] 郝水龙, 吴共庆, 胡学钢. 基于层次向量空间模型的用户兴趣表示及更新[J]. 南京大学学报:自然科学版, 2012, 48(2): 190-197. (Hao Shuilong, Wu Gongqing, Hu Xuegang. Presentation and Updation for User Profile Based on Hierarchical Vector Space Model [J]. Journal of Nanjing University: Natural Sciences, 2012, 48(2):190-197.)

[14] 宋丽哲, 牛振东, 余正涛, 等. 一种基于混合模型的用户兴趣漂移方法[J]. 计算机工程, 2006, 32(1): 4-6,89. (Song Lizhe, Niu Zhendong, Yu Zhengtao. A Method of Drifting User's Interests Based on Hybrid Model [J]. Computer Engineering, 2006, 32(1): 4-6,89.)

[15] 布红艳, 王国胤, 董振兴. 邮件系统中的兴趣漂移混合模型[J]. 计算机工程与设计, 2011, 32(12): 4026-4029. (Bu Hongyan, Wang Guoyin, Dong Zhenxing. Hybrid Interest Drifting Model of E-mail Systems [J]. Computer Engineering and Design, 2011,32(12): 4026-4029.)

[16] Maloof M A, Michalski R S. Selecting Examples for Partial Memory Learning [J]. Machine Learning, 2000, 41(1): 27-52.

[17] Koychev I. Gradual Forgetting for Adaptation to Concept Drift [C]. In: Proceedings of ECAI 2000 Workshop on Current Issues in Spatio-Temporal Reasoning, Berlin, Germany. 2000.

[18] Koychev I, Schwab I. Adaptation to Drifting User's Interests [C]. In: Proceedings of ECML2000 Workshop: Machine Learning in New Information Age. 2000: 39-46.

[19] Chen Z, Jiang Y, Zhao Y. A Collaborative Filtering Recommendation Algorithm Based on User Interest Change and Trust Evaluation [J]. International Journal of Digital Content Technology and Its Applications, 2010, 4(9): 106-113.

[20] Zheng N, Li Q. A Recommender System Based on Tag and Time Information for Social Tagging Systems [J]. Expert Systems with Applications, 2011, 38(4): 4575-4587.

[21] Zhang Y, Liu Y. A Collaborative Filtering Algorithm Based on Time Period Partition [C]. In: Proceedings of the 3rd International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, China. IEEE, 2010: 777-780.

[22] Karahodza B, Supic H, Donko D. An Approach to Design of Time-Aware Recommender System Based on Changes in Group User's Preferences [C]. In: Proceedings of the 2014 X International Symposium on Telecommunications. IEEE, 2014: 1-4.

[23] Wang Q, Sun M, Xu C. An Improved User-Model-Based Collaborative Filtering Algorithm [J]. Journal of Information and Computational Science, 2011, 8(10): 1837-1846.

[24] 邢春晓, 高凤荣, 战思南, 等. 适应用户兴趣变化的协同过滤推荐算法[J]. 计算机研究与发展, 2007, 44(2): 296-301. (Xing Chunxiao, Gao Fengrong, Zhan Sinan, et al. A Collaborative Filtering Recommendation Algorithm Incorporated with User Interest Change [J]. Journal of Computer Research and Development, 2007, 44(2): 296-301.)

[25] 于洪, 李转运. 基于遗忘曲线的协同过滤推荐算法[J]. 南京大学学报:自然科学版, 2010, 46(5): 520-527. (Yu Hong, Li Zhuanyun. A Collaborative Filtering Recommendation Algorithm Based on Forgetting Curve [J]. Journal of Nanjing University: Natural Sciences, 2010, 46(5): 520-527.)

[26] Wu Y K, Wang Y, Tang Z H. A Collaborative Filtering Recommendation Algorithm Based on Interest Forgetting Curve [J]. International Journal of Advancements in Computing Technology, 2012, 4(10): 148-157.

[27] Liu K, Chen W, Bu J, et al. User Modeling for Recommendation in Blogspace [C]. In: Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops. IEEE, 2007: 79-82.

[28] Cheng Y, Qiu G, Bu J, et al. Model Bloggers' Interests Based on Forgetting Mechanism [C]. In: Proceedings of the 17th International Conference on World Wide Web. ACM, 2008: 1129-1130.

[29] Rybak J, Balog K, Nørvåg K. Temporal Expertise Profiling [C]. In: Proceedings of the 36th European Conference on IR Research, Amsterdam, Netherlands. 2014: 540-546.

[30] Wu D, Zhao D, Zhang X. An Adaptive User Profile Based on Memory Model [C]. In: Proceedings of the 9th International Conference on Web-Age Information Management. IEEE, 2008: 461-468.

[31] Wang W, Zhao D, Luo H, et al. Mining User Interests in Web Logs of an Online News Service Based on Memory Model [C]. In: Proceedings of the 8th International Conference on Networking, Architecture and Storage. IEEE, 2013: 151-155.

[32] 于洪涛, 崔瑞飞, 董芹芹. 基于遗忘曲线的微博用户兴趣模型[J]. 计算机工程与设计, 2014, 35(10): 3367-3372, 3379. (Yu Hongtao, Cui Ruifei, Dong Qinqin. Micro-Blog User Interest Model Based on Forgetting Curve [J]. Computer Engineering and Design, 2014, 35(10): 3367-3372, 3379.)

[33] Hofmann T. Probabilistic Latent Semantic Indexing [C]. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1999: 50-57.

[34] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[35] 崔凯. 基于 LDA 的主题演化研究与实现[D]. 长沙: 国防科学技术大学, 2010. (Cui Kai. The Research and Implementation of Topic Evolution on LDA [D]. Changsha: National University of Defense Technology, 2010.)

[36] Ding Y, Li X. Time Weight Collaborative Filtering [C]. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management. ACM, 2005: 485-492.

[37] Cao J, Xia T, Li J, et al. A Density-Based Method for Adaptive LDA Model Selection [J]. Neurocomputing, 2009, 72(7-9): 1775-1781.

[38] Kullback S, Leibler R A. On Information and Sufficiency [J]. The Annals of Mathematical Statistics, 1951,22(1): 79-86.

[39] Jeong D H, Song M. Time Gap Analysis by the Topic Model-Based Temporal Technique [J]. Journal of Informetrics, 2014, 8(3): 776-790.

[40] Newman D, Asuncion A U, Smyth P, et al. Distributed Algorithms for Topic Models [J]. Journal of Machine Learning Research, 2009, 10: 1801-1828.

作者贡献声明：

桂思思：提出研究命题，设计实施方案，数据分析处理，论文起草与修订；

陆伟：设计研究方案，论文最终版本修订；

黄诗豪：Sogou 数据集预处理，使用主题模型生成用户兴趣；

周鹏程：在 Sogou 数据集上实现基于记忆的用户兴趣模型、基于遗忘曲线的用户兴趣度多阶段量化模型。

收稿日期：2015-04-03
收修改稿日期：2015-05-03

User Interest Prediction Combing Topic Model and Multi-time Function

Gui Sisi[1] Lu Wei[1,2] Huang Shihao[1] Zhou Pengcheng[1]
[1] (School of Information Management, Wuhan University, Wuhan 430072, China)
[2] (Center for the Studies of Information Resources, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] User interest is not static and it changes dynamically as time goes by, this paper proposes a user interest prediction model based on topic model and multi-time function. [**Methods**] Generate user interests by topic model, and calculate the weights of each user interest at every time point by applying multi-time function in order to predict user interest at next time point. [**Results**] Compared with memory-based user profile model and multi-step user profile model, cosine similarity and Kullback-Leibler divergence of the experimental results on search engine log data provided by Sogou Lab show that this model can predict user interests more effectively. [**Limitations**] The proposed method is only tested on search engine log data provided by Sogou Lab, and it need further examination on other data sets. [**Conclusions**] It is more effective to take every time point of user history data into consideration for user interest prediction.

Keywords: Topic model Time function User interest Prediction

新的隐私准则鼓励图书馆和内容供应商一起保护读者隐私

2015 年 6 月 29 日, 美国图书馆协会(American Library Association, ALA)的知识自由委员会批准了“图书馆隐私指南: 给电子书借阅和数字内容供应商”。该指南给出了厂商保护图书馆用户隐私的最佳做法, 旨在鼓励内容供应商和图书馆一起努力, 为图书馆读者进行电子书借阅和数字内容交付制定有效的隐私保护政策和程序。该文件由 ALA 知识自由委员会隐私权小组联合 ALA 的其他委员会, 以及相关的利益集团等共同撰写。

“伴随着图书馆提供数字内容、结合现代互联网提供个性化服务而发展起来的共同数据管理实践, 和图书馆一直以来的隐私保护之间逐渐形成鸿沟。” ALA 知识自由委员会隐私小组组长 Michael Robinson 说: “图书馆隐私指南中提供的这些准则试图平衡读者隐私保护的需求和图书馆收集用户数据并提供个性化服务的需求, 同时还能尊重和保护个人权利, 允许用户根据自己的需求决定与他们相关的数据的私密性。”

“即使图书馆转变为利用新技术和交付系统提供内容, 图书馆仍然是读者隐私坚定的保护者。” ALA 总裁 Sari Feldman 表示: “这些指导方针对帮助图书馆与供应商制定必要的读者隐私保护政策非常有帮助。”

该指南已可在 ALA 网站上访问, 详见: http://www.ala.org/advocacy/library-privacy-guidelines-e-book-lending-and-digital-content-vendors。

(编译自: http://www.oif.ala.org/oif/?p=5522)

(本刊讯)