Journal of Computer Applications
计算机应用,2015,35(7):1975 - 1978,1983
ISSN 1001-9081
CODEN JYIIDU
2015-07-10
http://www.joca.cn

文章编号:1001-9081(2015)07-1975-04
doi:10.11772/j.issn.1001-9081.2015.07.1975

面向文献搜索系统的用户实时需求发现方法

徐　浩,陈　雪*,胡晓峰
(上海大学 计算机工程与科学学院,上海 200444)
(*通信作者电子邮箱 xuechen@shu.edu.cn)

摘　要:针对当前文献搜索系统不能理解用户实时需求的问题,提出了一种面向文献搜索系统的用户实时需求发现方法。首先,分析用户浏览、下载等个性化搜索行为;其次,根据用户搜索行为与用户需求的关系构建用户实时需求文档(RD);然后,从用户需求文档中提取用户需求关键词网络;最后,运用随机游走的方法提取出关键词网络的核心节点构成用户需求图。实验结果表明:在模拟用户需求的环境下,提取需求图的方法比*K*-medoids算法在检索指标*F*值上平均高2.5%;在用户搜索文献真实情况下,提取需求图的方法比DBSCAN算法在检索指标*F*值上平均高5.3%,因此,在用户需求比较稳定的文献搜索中,该方法能够获取用户需求从而提升用户体验。

关键词:用户行为分析;实时需求;文献搜索系统;个性化;关键词网络
中图分类号:TP391.3　　**文献标志码**:A

Finding method of users' real-time demands for literature search systems

XU Hao, CHEN Xue*, HU Xiaofeng
(*School of Computer Engineering and Science, Shanghai University, Shanghai* 200444, *China*)

Abstract: Because of the literature search system failing to comprehend users' real-time demands, a method to find users' real-time demands for literature search systems was proposed. Firstly, this method analyzed the users' personalized search behaviors such as browsing and downloading. Secondly, it established users' real-time Requirement Documents (RD) based on the relations between users' search behaviors and users' requirements. And then it extracted keyword network from requirement documents. Finally, it gained users' demand graphs which were formed by core nodes extracted from keyword network by means of random walk. The experimental results show that the method by extracting demand graphs increases the *F*-measure by 2.5%, in the comparison of the *K*-medoids algorithm on average, under the condition that users' demands are emulated in the experiment. And it also increases the *F*-measure by 5.3%, in the comparison with the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm on average, under the condition that users really searches for papers. So, when the method is used in literature search systems where users' requirements are stable, it will be able to gain users' demands to enhance users' search experiences.

Key words: user behavior analysis; real-time demand; literature search system; personalization; keyword network

0　引言

随着互联网技术的发展,在线文献数量急剧增长,学术研究者可以通过使用文献搜索系统来减少获取文献的时间和成本。然而当前文献搜索系统呈现出搜索不准确、返回结果有效信息稀疏等缺点,而用户潜意识地希望只付出较少的检索代价就能准确获得想要的结果,因此,用户希望搜索系统能够提供从有效信息稀疏的返回结果中挑选出用户所需要文献的个性化检索服务。为用户不同的检索需求定制个性化检索结果来提高搜索质量的个性化的搜索将是一种有发展前景的方法[1]。

造成文献搜索系统呈现出上述缺点的主要原因在于搜索系统不能理解用户的实时需求。目前大多数文献搜索系统的入口都是条件搜索框,而用户在向系统输入需求时由于他们的表述能力的不足和搜索框的功能限制,通常只以少数检索词来表示自己的需求。而目前搜索系统的重点仍然停留在关键词的匹配方面,因此对于不同用户不同场景搜索,搜索系统仍然提供相同的结果。特别地,在学术搜索系统中,用户大部分是学术研究者,他们有着各自的研究领域,每个用户熟悉的专业知识也不一样,对相同的查询词汇,不同用户有着不同的理解,因此要提高用户的搜索体验为用户提供个性化服务,首先就要对用户的查询意图进行识别和提取出用户的需求。

1　研究现状

目前很多研究者都对学术文献搜索方面进行了研究:Kaya等[2]着重从个性化、可伸缩性和探索性搜索的三个重要属性方面改进学术推荐服务来提供更好的学术搜索。文献[3]通过采用能够区别文档相关性和位置偏好的检验假设来建立用户点击模型,观察用户点击行为是否能被相关性和位置偏好完全解释。Griffiths等[4]在作者主题模型[5]的基础上提出作者会议主题连接(Author Conference Topic Connection, ACTC)模型来研究学术网络,该模型在作者会议主题(Author Conference Topic, ACT)模型上增加了会议主题和主题之间的

收稿日期:2015-01-27;**修回日期**:2015-03-31。　**基金项目**:上海市教委创新项目(B.10-0108-14-202)。
作者简介:徐浩(1990 -),男,江苏南通人,硕士研究生,主要研究方向:海量Web信息挖掘;　陈雪(1981 -),女,河南信阳人,副教授,博士,CCF会员,主要研究方向:语义Web、对等网络、并行体系结构;　胡晓峰(1991 -),男,湖南湘潭人,硕士研究生,主要研究方向:海量Web信息挖掘。

潜在映射信息。文献[6]在处理个性化信息检索时进行了用户兴趣建模,分析真实可以代表用户兴趣的浏览、保存、收集或打印网页的用户访问行为;文章阐述了用户兴趣和访问行为的关系,通过对文档量化探讨了文档结构和遗忘因子对用户兴趣度的影响,通过兴趣模型来对结构过滤,最终提高用户的满意度。

上述的工作大都是对文献或用户的研究,而本文研究的目的是增强文献搜索系统对用户需求的理解,从而为用户提供个性化服务,提高用户的搜索体验,因此本文的研究重点是用户实时需求的获取。用户搜索文献是一种有较强目的的行为,因此用户的搜索过程存在着一定的规律,而目前搜索系统在理解用户需求时通常忽略了用户检索文献的个性化行为。用户的搜索行为包括浏览、下载、略过、翻页、收藏、分享等操作;这些行为准确地反映了用户的需求,而记录这些行为只会增加很小负担,分析这些行为有助于更好地理解用户的偏好[7]。另一方面由于文献文本有着半结构化的特殊的形式,文献搜索系统对文献呈现方式也存在着特点,通常搜索系统将文献的标题、摘要、关键词、引用、正文等都分开显示,因此学术搜索系统的返回对象的格式十分统一。

由于用户搜索文献的行为和搜索系统返回文献的格式都比较统一,因此可以在搜索系统的基础上增加用户需求分析模块,在用户搜索文献时对用户进行需求分析并为用户提供个性化服务。整个需求获取流程如图 1 所示,在文献列表呈现后增加用户需求分析模块,该模块首先对用户浏览、下载的行为以及相应的文献进行分析,并根据用户搜索行为建立需求文档并提取其关键词网络,再提取出用户关键词网络的核心语义以图的形式呈现给用户并对用户需求的表达进行提示。从用户搜索过程可以看出,在搜索系统呈现出文献列表之后,不管用户进行了哪些操作,搜索系统都不会对用户有任何响应。而在增加需求分析模块后,充分利用了用户与搜索系统的交互信息,对用户的每种检索行为都能给出响应和提示和当前页面文献列表更新。

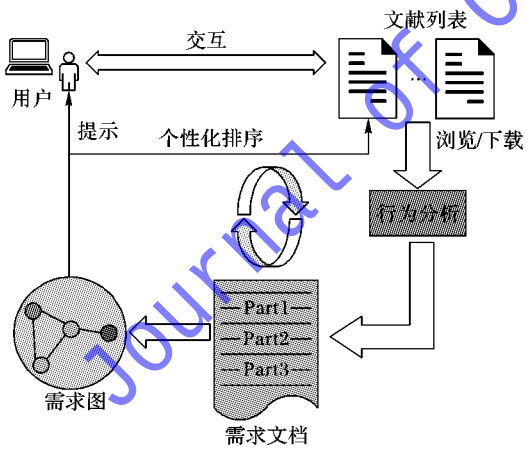

图 1 实时需求获取流程

2 用户需求提取方法

用户需求的提取过程包含 3 个步骤:首先,是能够形式化的表示出文献文本;其次,对用户的搜索行为和与行为相关的文献进行分析;最后,提取并形式化地表示出用户需求。

2.1 文献文本的表示

2.1.1 文献文本的特点

文献文本不同于一般的长文本:1)文献文本中关键词会随所处领域的不同而表现出不同的含义,比如“LDA”在文本表示中就是潜在狄利克雷分布(Latent Dirichlet Allocation, LDA)的主题模型,而在模式识别领域一般为线性判别式分析(Linear Discriminant Analysis, LDA)。2)文献文本由标题、摘要、正文等多部分组成,每个部分的意义和作用都不一样,通常标题部分的信息比摘要部分的信息重要,而摘要部分的信息又比正文中的信息重要。

为了能够合理地表示文献文本,针对文献文本的第一个特点本文使用图表示模型,以关键词作为图的节点,关键词与关键词的关系作为图的边。关键词的具体意义通过该关键词与周边词汇的关系来体现。为了体现出文献文本的半结构化的特性,本文提出了基于结构权重的文本表示方法,该方法把一篇文献分为几部分,每个部分赋予不同的权重,越重要的部分结构权重越大。对标题、摘要和正文处要赋予合理的权值将是一个难点,文献[7]中的方法是对不同位置赋予人为的权重,那种方法太主观,本文结合文献文本的自身特点最终使用“关键词密度”作为权值的衡量标准,某部分“关键词密度”反映了该部分关键词的密集程度,越重要的部分关键词越密集。最终密度公式为:

$$Density = n/N$$

其中:$Density$ 为该部分关键词的密度,n 为该部分所有关键词的个数,N 为该部分的所有词的个数。结合以上两种因素最后通过建立关键词网络来表示文献文本。

2.1.2 关键词网络的形式化表示

在获得一篇文献的标题、摘要、关键词、正文等信息后就很容易建立该篇文献的关键词网络,最终关键词网络用四元组表示:

$$KWN = \{V, WV, E, WE\}$$

其中:V 表示节点集合,$V = \{v_1, v_2, \cdots, v_i, \cdots, v_n\}$,每个节点 v_i 表示一个关键词。WV 表示节点的权重集合,$WV = \{wv_1, wv_2, \cdots, wv_i, \cdots, wv_n\}$,$wv_i$ 为节点 v_i 的权重。E 表示边的集合,$E = \{e_1, e_2, \cdots, e_i, \cdots, e_m\}$,每一条边 e_i 表示两个节点之间的相连。关键词与另一个关键词共现于同一个句子,则认为这两个关键词的节点间存在边。WE 表示边的权重集合,$WE = \{we_1, we_2, \cdots, we_i, \cdots, we_m\}$,$we_i$ 为边 e_i 的权重。

节点权重计算方法为:

$$Weight(A) = Density_t \times t_A + Density_a \times a_A + Density_m \times m_A$$

其中:$Density_t$、$Density_a$、$Density_m$ 分别为标题、摘要和正文部分的关键词密度,t_A、a_A、m_A 为关键词 A 在标题、摘要和正文中出现的次数。

边的权重计算方法为:

$$Weight(e_{AB}) = Density_t \times t_{AB} + Density_a \times a_{AB} + Density_m \times m_{AB}$$

其中:t_{AB}、a_{AB}、m_{AB} 分别为关键词 A 和 B 在标题、摘要和正文中共现的句子数。

最终节点和边的权重进行归一化处理得:

$$wv_i = Weight(v_i) \Big/ \left(\sum_{i=1}^{n} Weight(v_i) \right)$$

$$we_i = Weight(e_i) \Big/ \left(\sum_{i=1}^{m} Weight(e_i) \right)$$

2.1.3 关键词网络相似度

传统的图文本表示模型的相似度计算方法大都是将图的节点与边分开来考虑的,这样很难表现出节点与边对相似度的共同影响,本文在文献[8]的基础上提出了基于最大公共子图的相似度计算方法,通过动态地改变节点与边的权衡因

子,关键词网络相似度较小时主要考虑节点之间的关系,相似度较大时边的影响就会较大,这样使之在相似度较低的文献中和相似度较高的文献中都有很好的分辨能力,最终得到相似度计算公式:

$$Sim(KWN_1, KWN_2) = \beta \sum_{v \in V(KWN_C)} wv + (1-\beta) \sum_{e \in E(KWN_C)} we$$

其中:KWN_C 为关键词网络 KWN_1 和 KWN_2 的最大公共子图;$\sum_{v \in V(KWN_C)} wv$ 为最大公共子图所有节点的权重之和;$\sum_{e \in E(KWN_C)} we$ 为最大公共子图所有边的权重之和。

参数 β 为调节因子,动态变化的值公式如下:

$$\beta = \exp\left(-\left(\sum_{v \in V(C)} wv\right)/n_C\right)$$

其中 n_C 为最大公共子图 C 节点的个数。

2.2　用户需求文档构建方法

用户在搜索文献时总是按照自己的需求对搜索系统返回的结果进行挑选,略过自己不感兴趣的论文,浏览自己感兴趣的论文,下载自己所需要的论文。用户与搜索系统交互所产生的文献序列就是用户需求的一种体现,因此可以通过用户的浏览、下载的文献来表达出用户的实时需求。然而用户的需求是一种很模糊的抽象概念,面临的问题就是如何通过用户浏览、下载的文献序列来合理地构建出用户需求。用户搜索文献的行为通常包含略过、浏览、下载、收藏、分享、保存等操作。文献[9]通过提出了一种基于访问时间的方法来区分这些行为,访问时间越长则认为越重要;而文献[10]在访问时间的基础上结合页面大小和历史记录来计算用户行为的重要程度。

用户在搜索文献时下载一篇文献,说明该篇文献内容符合用户的需求;用户浏览一篇文献时说明该篇文献 Title 部分内容用户比较感兴趣,而用户放弃下载该篇文献则说明 Abstract 部分内容不是用户的主要需求。用户的收藏、分享、保存行为都是文献内容符合用户的需求的表现归属于下载操作。最终为了能够充分体现用户的需求,本文根据用户检索流程采用文档构造法来建立用户需求文档(Requirement Document,RD),再提取出用户的需求。需求文档可用四元组表示如下:

$$RD = \{K, P_1, P_2, P_3\}$$

其中:$K = \bigcup_{i=1}^{n} K_{d_i}, P_1 = \bigcup_{i=1}^{n} T_{d_i}, P_2 = \bigcup_{i=1}^{n} A_{d_i}, P_3 = \bigcup_{i=1}^{m} T_{v_i}, K_{d_i}$ 为下载的第 i 篇文献的关键词集合,T_{d_i} 为下载的第 i 篇文献的标题中句子集合,A_{d_i} 为下载的第 i 篇文献的摘要中句子集合,T_{v_i} 为第 i 只浏览未下载文献的标题中句子集合。

构建过程如图 2 所示,首先把用户下载的所有文献的 Title 合成为 Part1;把用户下载的所有文献的 Abstract 部分合成 Part2;把用户只浏览未下载的所有文献的 Title 部分合成 Part3。这 3 部分最终构成一个新的文档为 D。该文档 D 的关键词为所有下载的文献关键词的集合。最后提取出需求文档关键词网络 KWN_D。

2.3　需求图的提取方法

需求文档的关键词网络虽然能够反映出用户的需求,但是节点比较多并且包含很多冗余信息,如果直接用来作为用户查询意图的上下文计算,不但计算量很大而且不能突出用户的主要需求点,因此本文从需求文档的关键词网络中提取核心节点来构成需求图作为用户的需求上下文。提取思想:把关键词网络节点按重要程度进行排序,选取前 K 个节点作为需求图的节点。

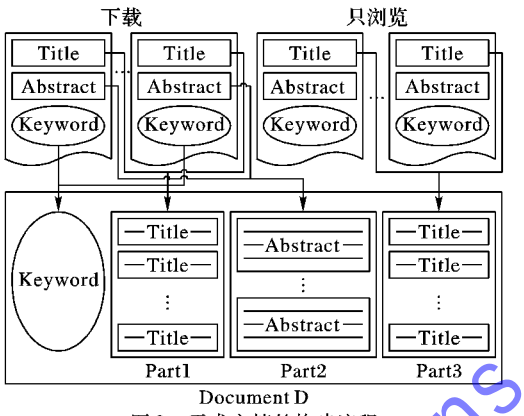

图 2　需求文档的构建流程

关键词网络节点重要程度取决于两个因素:1)节点的权重,节点权重越高则越重要。2)网络拓扑结构中节点的关键程度。目前传统的网络节点排序方法很多,例如局部中心性排序方法[11]、k-壳分解法[12]以及 PageRank[13]等方法。这些方法大都只考虑网络的某一指标,而且关键词网络边和节点都含有权重,很难直接运用上述方法。本文为了综合考虑节点权重和拓扑结构关键程度两种因素,最终使用随机游走方法对网络节点进行排序,方法描述如下:

假设一个 Agent 在关键词网络节点间随机游走,Agent 只能选择跳转和游走。当 Agent 处于节点 v_i,则 Agent 下一步选择跳转的概率为 $p_{v_i \to \text{jump}} = 1/(Degree(v_i)+1)$;如果 Agent 选择跳转,则跳到节点 v_j 的概率为 $p_{\text{jump} \to v_j} = wv_j$;如果选择游走则只能游走到相邻节点,游走到节点 v_j 的概率为 $p_{\text{walk} \to v_j} = we_{ij}/\left(\sum we_{i\cdot}\right)$,其中 $\sum we_{i\cdot}$ 为节点 v_i 组成的所有的边权重之和。

节点重要度求取步骤如下:

1) Agent 以概率 $p_{\text{jump} \to v_i}$ 跳转到某一节点 v_i,Agent 经过节点 v_i 次数加 1。

2) Agent 以概率 $p_{v_i \to \text{jump}}$ 选择跳转或以 $1 - p_{v_i \to \text{jump}}$ 选择游走,如果选择跳转则转到步骤 1),选择游走则转到步骤 3)。

3) Agent 以概率 $p_{\text{walk} \to v_j}$ 从当前节点 v_i 游走到邻节点 v_j,Agent 经过节点 v_j 次数加 1,转到步骤 2)。

一直循环上述步骤并记录 Agent 经过每个节点 v_i 次数为 $n(v_i)$,估算出 Agent 经过该节点概率 $p(v_i)$,在 Agent 经过每个节点的概率波动小于阈值时停止循环。最后 $p(v_i)$ 就为该节点的 Rank 值。

该随机游走方法在循环过程直接运用关键词网络节点的权重作为跳转概率,不存在跳转概率的更新因此计算规模只和关键词网络的规模相关,相比传统的随机游走方法[14],该方法收敛速度更快。

将所有的节点按照 Rank 值从大到小排序,提取出前 K 个节点以及这些节点之间的边来构成需求图,然后可以将需求图直接展示给用户,对用户下一个查询词进行提示;计算当前页面文献列表的文献与需求图的相似度,对文献列表进行排序。

3　实验分析

3.1　实验设置

需求图的提取是一种半监督学习。为了验证需求图代表用户需求的能力,实验采用一种比较需求图的分类效果和半监督聚类效果的方法。由于实验验证涉及到用户的需求因此实验采用了两种方案。

第 1 种方案　描述如下。

实验数据　选取了 7 个领域的论文，分别是 JiaWei Han 在 Data Mining 领域的论文；Ling Liu 在 Peer to Peer 领域的论文；Francisco Herrera 在 Fuzzy Theory 领域的论文；Edward Fox 在 Digital Library 领域的论文；W. Bruce Croft 在 Information Retrieval 领域的论文；Hermann Ney 在 Machine Translation 领域的论文和 Amit Sheth 在 Semantic Web 领域的论文，每类各 50 篇总共 350 篇，另外加上 50 篇作为实验数据杂质的其余类别的论文，共计 400 篇论文，每领域的论文都是同一作者在同一研究课题所写的文献；每一类论文作为用户的同一需求。

实验方案　采用需求图分类方法。每类随机选择 n 篇论文作为用户下载的论文来模拟用户行为，构建需求文档提取出需求图；得到所有类的需求图后，计算剩余论文与各需求图的相似度，把论文归到与之相似度最大的需求图类别。半监督聚类：每类也随机选择 n 篇论文，将 400 篇论文聚成 7 类，采用 K-medoids 聚类方法，要求每类选择的 n 篇论文必须聚到同一类，不同类论文不能聚到一类。最后，比较两种方法在 n 分别取 2，3，5，10 的情况下检索的准确率、召回率和 F 值。

第 2 种方案　描述如下。

实验数据　获取多名用户在 IEEE Xplore Digital Library 搜索文献的日志，选取 20 个下载量较高的查询的前 50 篇文献，其中包含用户略过、浏览和下载的文献。平均每个查询 50 篇文献中有 7.4 篇被下载。

实验方案　采用需求图分类方法。在用户下载了 n 篇论文时，从日志中获取用户在该次搜索中前面下载和浏览的所有文献，结合用户下载和浏览的行为与相关文献提取出需求图；计算需求图与列表中文献的相似度，将相似度大于阈值的论文归为用户需求类。半监督聚类：用 DBSCAN（Density-Based Spatial Clustering of Applications with Noise）[15] 聚类方法将文献列表中 50 篇文献聚成若干个类。聚类时增加约束，保证下载的前 n 篇论文必须聚到同一类作为需求类；在下载的第 n 篇论文前的所有浏览和略过的论文都不能聚到需求类中。两种方法都是把日志中用户下载的所有论文作为用户的需求标准论文集，比较需求类和需求标准集中的文献，计算检索的准确率、召回率和 F 值。最后在 n 分别取 1，2，3，4，5 的情况下对这两种方法进行对比。

3.2　实验结果分析

第 1 种方案实验效果如图 3 所示，从实验结果可以看出，在总体趋势上随着选取的论文数目 n 的增加，需求图和半监督聚类的效果均越来越好。在论文篇数 $n = 2$ 时，半监督聚类方法比需求图分类效果要好。当选择论文数量大于 3 篇时，需求图的分类效果就要好于半监督聚类，最终在 n 取 3，5 时可计算得到需求图的效果比半监督聚类效果在 F 值上平均高 2.5%。

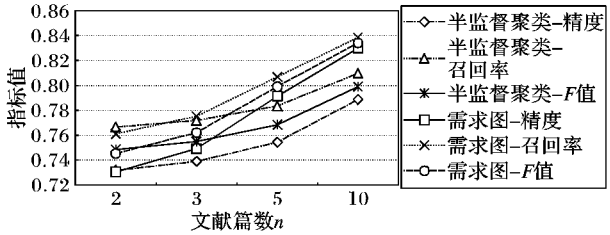

图 3　第 1 种实验方案的实验结果

第 2 种方案实验效果如图 4 所示，当用户下载了一篇论文时，半监督聚类和需求图两种方法效果都普遍偏低，但半监督聚类要比需求图方法好很多。当 $n = 2$ 时需求图方法的结果 F 值就已经好于半监督聚类；在 $n = 3$ 时需求图检索 F 值比半监督聚类高 5.3%。随着 n 的增大，需求图方法的精度、召回率和 F 值均有所增加，而半监督聚类下载了几篇论文后精度呈现下降趋势，最终 F 值不会有很大的提升。

两种方案实验的趋势表明，提取需求图的方法比半监督聚类方法效果更好，在用户进行文献搜索时，随着用户交互的增加，需求图将越来越接近用户的实时需求。

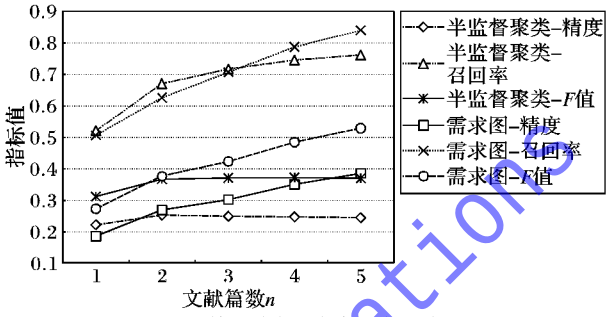

图 4　第 2 种实验方案的实验结果

4　结语

本文在文献搜索系统的基础上，结合文献文本的特点提出了基于结构权重的文本表示模型；通过对用户搜索文献时的交互行为进行深入研究，分析并提取用户与搜索系统交互所反映的与用户需求相关的信息，构建需求文档来进一步明确用户检索意图。最后提取出用户搜索的核心需求，以图的形式对用户的下一步搜索进行提示，从而节约用户搜索文献的成本。

参考文献：

[1] XU Y, WANG K, ZHANG B, *et al*. Privacy-enhancing personalized Web search [C]// Proceedings of the 16th International Conference on World Wide Web. New York: ACM, 2007: 591 – 600.

[2] KUCUKTUNC O, SAULE E, KAYA K, *et al*. Towards a personalized, scalable, and exploratory academic recommendation service [C]// Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York: ACM, 2013: 636 – 641.

[3] HU B, ZHANG Y, CHEN W, *et al*. Characterizing search intent diversity into click models [C]// Proceedings of the 20th International Conference on World Wide Web. New York: ACM, 2011: 17 – 26.

[4] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, *et al*. The author-topic model for authors and documents [C]// Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Arlington: AUAI Press, 2004: 487 – 494.

[5] WANG J, HU X, TU X, *et al*. Author-conference topic-connection model for academic network search [C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012: 2179 – 2183.

[6] ZHU Z, WANG J, CHEN M, *et al*. User interest modeling based on access behavior and its application in personalized information retrieval [C]// Proceedings of the 2010 International Conference on Information Management, Innovation Management and Industrial Engineering. Piscataway: IEEE, 2010: 266 – 270.

[7] NANDA A, OMANWAR R, DESHPANDA B. Implicitly learning a user interest profile for personalization of Web search using collaborative filtering [C]// Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies. Piscataway: IEEE, 2014: 54 – 62.

（下转第 1983 页）

[3] TSO-SUTTER K H L, MARINHO L B, SCHMIDT-THIEME L. Tag-aware recommender systems by fusion of collaborative filtering algorithms [C]// Proceedings of the 2008 ACM Symposium on Applied Computing. New York: ACM, 2008: 1995 – 1999.

[4] ZHOU T C, MA H, KING I, *et al*. TagRec: leveraging tagging wisdom for recommendation [C]// CSE'09: Proceedings of the 2009 International Conference on Computational Science and Engineering. Washington, DC: IEEE Computer Society, 2009: 194 – 199.

[5] SUN L, LI S. Social tagging recommendation system based on *K*-means cluster and tensor decomposition [J]. Journal of Jiangsu University of Science and Technology: Natural Science Edition, 2012, 26(6): 597 – 601. (孙玲芳, 李烁朋. 基于 *K*-means 聚类与张量分解的社会化标签推荐系统研究[J]. 江苏科技大学学报: 自然科学版, 2012, 26(6): 597 – 601.)

[6] SYMEONIDIS P, NANOPOULOS A, MANOLOPOULOS Y. A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(2): 179 – 192.

[7] LIAO Z, WANG C, LI X. *et al*. Tag recommendation and new user tag recommendation algorithms based on tensor decomposition [J]. Journal of Chinese Computer Systems, 2013, 34(11): 2472 – 2476. (廖志芳, 王超群, 李小庆, 等. 张量分解的标签推荐及新用户标签推荐算法[J]. 小型微型计算机系统, 2013, 34(11): 2472 – 2476.)

[8] de LATHAUWER L, de MOOR B, VANDEWALLE J. A multilinear singular value decomposition [J]. SIAM Journal on Matrix Analysis and Applications, 2000, 21(4): 1253 – 1278.

[9] SYMEONIDIS P, NANOPOULOS A, PAPADOPOULOS A, *et al*. Scalable collaborative filtering based on latent semantic indexing [C]// ITWP 2006: Proceedings of the 2006 IJCAI Workshop on Intelligent Techniques for Web Personalization. Boston: [s. n.], 2006: 1 – 9.

[10] MA H, YANG H, LYU M R, *et al*. SoRec: social recommendation using probabilistic matrix factorization [C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York: ACM, 2008: 931 – 940.

[11] HUANG W, MENG X, WANG L. A collaborative filtering algorithm based on users' social relationship mining in mobile communication network [J]. Journal of Electronics and Information Technology, 2011, 33(12): 3002 – 3007. (黄武汉, 孟祥武, 王立才. 移动通信网中基于用户社会化关系挖掘的协同过滤算法[J]. 电子与信息学报, 2011, 33(12): 3002 – 3007.)

[12] YU H, LI J. Collaborative filtering recommendation algorithm using social and tag information [J]. Journal of Chinese Computer Systems, 2013, 34(11): 2467 – 2471. (于洪, 李俊华. 结合社交与标签信息的协同过滤推荐算法[J]. 小型微型计算机系统, 2013, 34(11): 2467 – 2471.)

[13] MA H, KING I, LYU M R. Learning to recommend with social trust ensemble [C]// Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2009: 203 – 210.

[14] YAN Z, ZHOU J. User recommendation with tensor factorization in social networks [C]// Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington, DC: IEEE Computer Society, 2012: 3853 – 3856.

[15] WU L, CHEN E, LIU Q, *et al*. Leveraging tagging for neighborhood-aware probabilistic matrix factorization [C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012: 1854 – 1858.

[16] LIU H, HU Z, MIAN A, *et al*. A new user similarity model to improve the accuracy of collaborative filtering [J]. Knowledge-Based Systems, 2014, 56: 156 – 166.

[17] HetRec 2011. Datasets [EB/OL]. [2014-06-08]. http://grouplens.org/datasets/hetrec-2011/.

(上接第 1978 页)

[8] TOMITA J, NAKAWATASE H, ISHII M. Calculating similarity between texts using graph-based text representation model [C]// Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. New York: ACM, 2004: 248 – 249.

[9] LIANG T, LAI H. Discovering user interests from Web browsing behavior: an application to Internet news services [C]// HICSS 2002: Proceedings of the 35th Annual Hawaii International Conference on System Sciences. Piscataway: IEEE, 2002: 2718 – 2727.

[10] LI Y, FENG B, WANG F. Page interest estimation based on the user's browsing behavior [C]// ICIC'09: Proceedings of the Second International Conference on Information and Computing Science. Piscataway: IEEE, 2009: 258 – 261.

[11] CHEN D, LYU L, SHANG M, *et al*. Identifying influential nodes in complex networks [J]. Physica A: Statistical Mechanics and Its Applications, 2012, 391(4): 1777 – 1787.

[12] KITSAK M, GALLOS L K, HAVLIN S, *et al*. Identification of influential spreaders in complex networks [J]. Nature Physics, 2010, 6(11): 888 – 893.

[13] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks and ISDN systems, 1998, 30(1): 107 – 117.

[14] JIN D, YANG B, LIU J, *et al*. Ant colony optimization based on random walk for community detection in complex networks [J]. Journal of Software, 2012, 23(3): 451 – 464. (金弟, 杨博, 刘杰, 等. 复杂网络簇结构探测——基于随机游走的蚁群算法[J]. 软件学报, 2012, 23(3): 451 – 464.)

[15] ESTER M, KRIEGEL H P, SANDER J, *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise [C]// KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996: 226 – 231.

[16] WANG H, SONG Y, CHANG M W, *et al*. Learning to extract cross-session search tasks [C]// Proceedings of the 22nd International Conference on World Wide Web. New York: ACM, 2013: 1353 – 1364.

[17] YIN X, TAN W. Semi-supervised truth discovery [C]// Proceedings of the 20th International Conference on World Wide Web. New York: ACM, 2011: 217 – 226.

[18] YANG Q, HAO H, NENG X. The research on user interest model based on quantization browsing behavior [C]// ICCSE 2012: Proceedings of the 7th International Conference on Computer Science & Education. Piscataway: IEEE, 2012: 50 – 54.

[19] ALHARBI A, SMITH D, MAYHEW P. Web searching behavior for academic resources [C]// Proceedings of the 2013 Science and Information Conference. Piscataway: IEEE, 2013: 104 – 113.

[20] BASHIR M B, ABD LATIFF M S, ABDULHAMID S M, *et al*. Grid-based search technique for massive academic publications [C]// ICT-ISPC 2014: Proceedings of the Third ICT International Student Project Conference. Piscataway: IEEE, 2014: 173 – 176.