(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101719145 B
(45)授权公告日 2011.08.10

(21)申请号 200910238155.8
(22)申请日 2009.11.17
(73)专利权人 北京大学
地址 100871 北京市海淀区中关村颐和园路5号
(72)发明人 张铭 孙韬
(74)专利代理机构 北京市商泰律师事务所 11255
代理人 毛燕生
(51)Int.Cl.
G06F 17/30(2006.01)
(56)对比文件
US 2002/0123987 A1,2002.09.05, 全文.
EP 1538838 A1,2005.06.08, 全文.
CN 101540874 A,2009.09.23, 全文.
CN 101556603 A,2009.10.14, 全文.
审查员 王艳臣

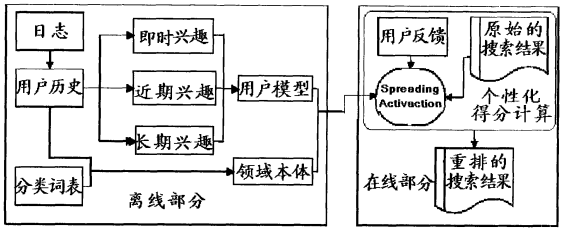
权利要求书 2 页 说明书 7 页 附图 2 页

(54)发明名称

基于图书领域本体的个性化搜索方法

(57)摘要

本发明提供基于图书领域本体的个性化搜索方法,属于个性化网络搜索服务。该方法包括:建立领域本体,引入协同过滤思想,加入体现用户之间影响的语义联系;对日志进行分析处理,建立基于用户兴趣偏好的用户模型;个性化得分计算,基于用户模型和领域本体,通过图挖掘算法(SpreadingActivation,简称SA)来计算;以及,重排搜索结果,按照个性化得分从高到低的顺序对原搜索引擎返回的结果进行重排并返回给用户。本发明方法将协同过滤思想引入领域本体,并建立及时体现用户兴趣变化的用户模型,通过图挖掘算法准确分析用户需求,有效消除关键词歧义,大幅提高用户对搜索结果的满意度。



CN 101719145 B

1. 基于图书领域本体的个性化搜索方法，其特征在于，包括在离线部分用户模型和领域本体的建立，以及在线部分个性化特征计算和重排搜索结果，具体步骤如下：

步骤 1，建立领域本体：基于一定的图书分类法，在图书两两之间建立加权有向不对称 borrowIntent 联系，其中 borrowIntent 的具体定义为：若有 n_1 个读者借阅了图书 b_1，n_2 个读者借阅了图书 b_2，$b_1 \rightarrow b_2$ 的边权重 (link weight) 为：$borrowIntent(b_1, b_2) = |n_1 \cap n_2| / n_1$，同理，有 $b_2 \rightarrow b_1$ 的边权重：$borrowIntent(b_2, b_1) = |n_1 \cap n_2| / n_2$；进而建立图书领域本体，提供该特定领域的概念定义和概念之间的语义关系；

步骤 2，建立用户模型：对日志进行分析处理，分析用户历史记录，根据时间顺序建立用户模型；

步骤 3，个性化得分计算：根据已经建立的领域本体和用户模型，通过图挖掘算法 SA 来计算该个性化得分；所述的图挖掘算法 SA 的具体计算步骤如下：首先，把领域本体看作图，在领域本体上运用图挖掘算法 SA，并将用户模型中的被赋予了权值的图书为传播扩散的初始点；然后，设置 SA 算法的循环次数限制、传播路径限制以及传播终点限制，以提高算法的效率；最后，通过得分更新公式，不断迭代更新每个点的分数值，直到整个算法结束；

步骤 4，重排搜索结果：根据所述 SA 得到的个性化得分，按照从高到低的顺序对原搜索引擎返回的结果进行重排，然后返回给用户。

2. 根据权利要求 1 所述的基于图书领域本体的个性化搜索方法，其特征是：步骤 1 中所述领域本体是本搜索算法的传播网络，通过提供丰富的语义信息和强化实体之间的语义关系，克服了使用原始搜索结果中的多义词、同义词和单词依赖现象，从而起到消除语义歧义的作用；并引入协同过滤思想，能体现用户之间的互相影响。

3. 根据权利要求 1 所述的基于图书领域本体的个性化搜索方法，其特征是：步骤 2 中，由于读者的兴趣随着时间的推移不断发生变化，按照时间顺序对用户兴趣进行分类和加权，从而建立起体现用户兴趣偏好的用户模型。

4. 根据权利要求 3 所述的基于图书领域本体的个性化搜索方法，其特征是，所述按照时间顺序对用户兴趣进行分类和加权是指，按照借阅时间段将所述时间顺序分为即时兴趣、近期兴趣和长期兴趣三类，并赋予这三类兴趣从高到低的权值。

5. 根据权利要求 1 或 4 所述的基于图书领域本体的个性化搜索方法，其特征是：步骤 2 中，所述用户模型通过分析用户历史记录，根据用户的行为习惯判别用户在检索过程中的行为偏好，并体现用户兴趣的更新和迁移，从而利用用户兴趣偏好实现个性化服务。

6. 根据权利要求 1 所述的基于图书领域本体的个性化搜索方法，其特征是：在所述图挖掘算法 SA 的每个循环结束之后，收集用户反馈信息，来更新下一次传播的初始点激活值，所收集的用户反馈信息包括：用户新点击链接的图书和用户新借阅的图书，这两部分图书都作为即时兴趣，并赋予权值。

7. 基于图书领域本体的个性化搜索系统，其特征是，包括：

领域本体模块，系统基于一定的图书分类法，在图书两两之间建立加权有向不对称 borrowIntent 联系，其中 borrowIntent 的具体定义为：若有 n_1 个读者借阅了图书 b_1，n_2 个读者借阅了图书 b_2，$b_1 \rightarrow b_2$ 的边权重 (link weight) 为：$borrowIntent(b_1, b_2) = |n_1 \cap n_2| / n_1$，同理，有 $b_2 \rightarrow b_1$ 的边权重：$borrowIntent(b_2, b_1) = |n_1 \cap n_2| / n_2$；进而建立图书领域本体，提供该特定领域的概念定义和概念之间的语义关系；

用户模型，其对日志进行分析处理，分析用户历史记录，由于用户兴趣随着时间的推移不断发生变化，按照时间顺序对所述用户兴趣进行分类和加权，从而建立起基于用户兴趣偏好的用户模型；

个性化得分计算模块，根据已经建立的领域本体和用户模型，通过图挖掘算法 SA 来计算该得分；所述的图挖掘算法 SA 包括如下单元：初始值确定单元：把领域本体看作图，将用户模型中的被赋予了权值的图书作为传播扩散的初始点；SA 算法的循环设置单元：设置 SA 算法的循环次数限制、传播路径限制以及传播终点限制，以提高算法的效率；迭代单元：通过得分更新公式，不断迭代地更新每个点的激活值，直到整个算法结束；

重排搜索结果模块，其根据所述 SA 得到的个性化得分，按照从高到低的顺序对原搜索引擎返回的结果进行重排，然后返回给用户。

基于图书领域本体的个性化搜索方法

技术领域

[0001] 本发明涉及图书馆的个性化服务，尤其涉及为图书馆提供个性化搜索的方法，属于计算机应用技术信息管理技术领域。

背景技术

[0002] 信息时代，随着信息量的爆炸式增加，“信息过载”逐渐成为了一个不可忽视的问题。通用搜索引擎比如 Google 和百度等，能够返回成千上万的搜索结果，但这给用户带来了信息筛选的困难。而且用户经常提交一些有歧义的关键词，比如，喜欢惊险小说的用户 A 使用关键词“达芬奇”来搜索丹·布朗的代表作《达·芬奇的密码》，而热衷文艺复兴艺术的用户 B 也同样选择关键词“达芬奇”来搜索达·芬奇的画作。显然他们有着不同的信息需求，但Google、百度会返回给他们同样的搜索结果。数字图书馆也同样面临着这个严峻的问题，由于数字文件数量的不断增长和关键词的不明确性，用户不得不花费越来越长的时间从返回结果中选择真正所需。个性化搜索，能够通过分析用户历史记录、建立用户模型，针对用户真正需求返回更为精准的搜索结果，从而解决“信息过载”问题。

[0003] 目前，在个性化搜索技术领域，国内外学者已经展开了大量而深入的研究工作，主要的个性化搜索方法有：基于个性化 PageRank 算法 (T.H.Haveliwala.，Topic-sensitive PageRank.Proceedings of the 11th internationalconference on World Wide Web. New York，USA,2002.)，基于该算法的个性化搜索方法根据用户浏览历史分析用户兴趣，然后在随机游走 (Random Walk) 时偏向特定类别的文档；基于聚类算法 (Ferragina Gulli,A Personalized SearchEngine Based On Web Snippet Hierarchical Clustering,Software Practiceand Experi ence，Volume 38,2008.)，该方法对文档进行聚类，然后高亮用户感兴趣的特定类别等等。以上方法虽然能够根据用户兴趣实现一定程度的个性化搜索，但是并不能有效消除关键词的歧义，也没有考虑被搜索领域的语义知识，造成了文档之间的语义联系的缺失。

[0004] Yolanda Blanco-Fernández 等人提出了利用语义推理来实现个性化服务的方法 (Yolanda Blanco-Fernández，JoséJ.Pazos Arias，etc.，SemanticReasoning：A Path to New Possibilities of Personalization，the 5th AnnualEuropean Semantic Web Conference，Tenerife，Spain,2008.)。该方法首先根据领域知识建立领域本体，然后运用语义推理技术的 ρ-path、ρ-join、ρ-cp 方法得到实例 (Instance) 之间潜在的语义联系。利用推理技术扩充本体之后，再基于领域本体计算各实例与用户偏好之间的相似度。该方法虽然能利用领域本体有效消除关键词歧义，但存在如下缺点：

[0005] 1. 没有考虑用户之间的相互影响，仅仅从实例的角度来计算相似性；

[0006] 2. 没有考虑用户兴趣的变化性，无法跟踪用户最近期的需求。

发明内容

[0007] 为了克服现有技术的不足，本发明提供一种基于图书领域本体的个性化搜索方法

及系统。该方法首先建立图书领域本体，考虑用户之间的影响，加入新的语义联系；然后建立用户模型 (User Profile)，对兴趣按照时间顺序进行了分类和加权；再运用通过图挖掘算法 SA 计算个性化得分，并据此重排搜索引擎返回的结果，实现个性化搜索。本发明方法通过利用领域本体，有效消除关键词歧义，即时体现用户兴趣变化，从而准确分析用户需求，大幅提高用户对搜索结果的满意度。本发明解决其技术问题所采用的技术方案是：

[0008] 基于图书领域本体的个性化搜索方法，其包括离线部分用户模型的建立和领域本体的建立，以及在线部分个性化得分的计算和搜索结果的重排，具体步骤如下：

[0009] 步骤 1，建立领域本体：根据用户历史，以一个特定领域的分类词表作为描述对象的本体，并加入协同过滤的思想，建立领域本体，从而提供该特定领域的概念定义和概念之间的语义关系。

[0010] 所述领域本体提供了丰富的语义信息，强化了实体之间的语义关系，从而克服了使用原始搜索结果中的多义词、同义词和单词依赖等现象，起到消除歧义的作用，并且能反映用户之间兴趣的互相影响。该领域本体将作为本发明的搜索算法的传播网络。

[0011] 步骤 2 建立用户模型，根据用户历史对日志进行分析处理，分析用户历史记录。读者的兴趣会随着时间的推移不断发生变化，因此按照时间顺序对用户兴趣进行分类和加权。按照借阅时间段分为即时兴趣、近期兴趣和长期兴趣三类，并赋予这三类兴趣从高到低的权值，从而建立起基于用户兴趣偏好的用户模型。

[0012] 所述用户模型能够及时体现用户兴趣的更新和迁移。

[0013] 步骤 3 个性化得分计算，根据已经建立的领域本体和用户模型，通过图挖掘算法 SA 来计算该个性化得分，具体计算步骤如下：

[0014] 首先，把领域本体看作图，在领域本体上运用 SA 算法，并将用户模型中的被赋予了权值的图书为传播扩散的初始点 (Initial Nodes)；然后，设置 SA 算法的循环次数限制、传播路径限制以及传播终点限制，以提高算法的效率；最后通过得分更新公式，不断迭代更新每个点的激活值，直到整个算法结束。

[0015] 在所述图挖掘算法 SA 的每个循环结束之后，收集用户反馈信息，来更新下一次传播的初始点激活值，所收集的用户反馈信息包括：用户新点击链接的图书和用户新借阅的图书，这两部分图书都作为即时兴趣，并赋予权值。

[0016] 步骤 4 重排搜索结果，根据所述 SA 得到的个性化得分，按照从高到低的顺序对原搜索引擎返回的结果进行重排，然后返回给用户。

[0017] 基于图书领域本体的个性化搜索系统，其包括：

[0018] 领域本体模块，系统通过以一个特定领域作为描述对象的本体，建立领域本体，从而提供该特定领域的概念定义和概念之间的语义关系；

[0019] 用户模型，其对日志进行分析处理，分析用户历史记录，由于用户兴趣随着时间的推移不断发生变化，按照时间顺序对所述用户兴趣进行分类和加权，从而建立起基于用户兴趣偏好的用户模型；

[0020] 个性化得分计算模块，根据已经建立的领域本体和用户模型，通过图挖掘算法 SA 来计算该得分；以及，

[0021] 重排搜索结果模块，其根据所述 SA 得到的个性化得分，按照从高到低的顺序对原搜索引擎返回的结果进行重排，然后返回给用户。

[0022] 所述个性化得分计算模块中的所述图挖掘算法 SA 包括如下单元 :

[0023] 初始值确定单元 :把领域本体看作图,将用户模型中的被赋予了权值的图书作为传播扩散的初始点 ;

[0024] SA 算法的循环设置单元 :设置 SA 算法的循环次数限制、传播路径限制以及传播终点限制,以提高算法的效率 ;以及,

[0025] 迭代单元 :通过得分更新公式,不断迭代地更新每个点的激活值,直到整个算法结束。

[0026] 本发明的有益效果 :

[0027] 1. 本发明方法将协同过滤思想和 SA 算法相结合,在建立领域本体的时候引入新的语义联系 -borrowIntent,从而从用户兴趣的角度反映两个实体之间的相似性,大大丰富和提高了领域本体的表达能力,同时保证了 SA 传播网络的信息完整性。

[0028] 2. 更精确细致地分析用户兴趣以建立用户模型。在建立用户模型的时候,区分了用户不同时间段的兴趣,通过赋予不同的权值,客观、全面表达用户兴趣知识,体现和跟踪用户兴趣的变化,并且保证了 SA 算法初始点权值的合理性。

[0029] 本发明的方法使用北京大学图书馆 (http://www.lib.pku.edu.cn) 的真实日志数据进行了评测,实验数据表明,通过本发明方法的个性化搜索重排结果,能够在返回的搜索结果中,有效消除关键词歧义,并且大幅提高用户感兴趣的图书排名,从而节省用户浏览时间,提高用户满意度。同时,本发明的方法并不只限于图书馆领域,可以扩展运用到其他领域,具有较高的实验价值。

附图说明

[0030] 图 1 为本发明方法的领域本体示意图 ;

[0031] 图 2 为根据本发明为用户提供个性化搜索服务的过程示意图 ; ;

[0032] 图 3 为采用本发明方法与其他三种方法 Top N 结果的 Norm DCG 平均值比较图 ;

[0033] 图 4 为采用本发明方法与其他三种方法 Top N 中用户感兴趣结果个数的比较图。

具体实施方式

[0034] 下面结合附图和具体实施方式对本发明作进一步详细描述 :

[0035] 实施例 1 :以北京大学图书馆 2008 年 1 月到 2008 年 6 月的真实日志数据为例,结合图 1 的领域本体示意图详细描述本发明的具体实施方式。

[0036] 本具体实施方式描述的是为图书馆用户提供个性化搜索服务的方法。目标是对于同一关键词的检索请求,不同用户能够得到最贴近的自己需要的信息,从而为用户带来更好的用户体验。在本具体实施方式中,基于图书领域本体的个性化搜索方法的系统结构如图 2 所示。

[0037] 具体描述如下 :

[0038] 第一部分,离线部分。离线部分的工作在线下完成,即在用户提交关键词搜索之前完成。具体的步骤如下 :

[0039] 步骤 1,建立领域本体 :以一个特定领域作为描述对象的本体,建立领域本体,从而提供该特定领域的概念定义和概念之间的语义关系。具体地,领域本体建议过程如下 :

[0040] 领域本体示意图如图 1 所示。建立本体时，引入协同过滤 (Badrul Sarwar, George Karypis, Joseph Konstan, John Riedl., Item-Based CollaborativeFiltering Recommendation Algorithms, Proceedings of the 10thinternational conference on World Wide Web, Hong Kong, 2001.) 的思想，考虑用户之间的影响，加入新的语义联系。然后建立用户模型，对兴趣按照时间顺序进行了分类和加权。再运用 Spreading Activation(SA)Model(AM Collins, EF Loftus. A spreading-activation theory of semantic processing. Psychological review. V. 82 p. 407-428, 1975.) 算法，重排搜索引擎返回的结果，实现个性化搜索。本发明建立的领域本题中，概念 (concept) 包括图书类别 (class) 和图书实体 (instance)，概念之间的联系包括 W3C 推荐的 rdfs:subClassOf, rdf:type, dc:creator, dc:subject 以及本发明新提出的 borrowIntent。

[0041] 具体来说，对于图书领域，基于中国图书馆分类法 (CLC)，建立顶层类别“F 经济”，“I 文学”，“J 艺术”等，以及子类别“J2 绘画”，“J22 中国绘画作品”等。类别与子类别之间用 rdfs:subClassOf 联系。对于每本图书，按照其对应的中图分类号，分类至 CLC 的最底层类别。比如《达·芬奇密码》的分类号为 I712.45/598，则分类至“I7”。图书与类别之间用 rdf:type 联系。根据图书的作者和主题信息，继续在领域本体引入 dc:creator 和 dc:subject 联系，比如《达·芬奇密码》与其作者“丹·布朗”之间以 dc:creator 联系，《达·芬奇画作》与其主题“文艺复兴艺术”之间以 dc:subject 联系。

[0042] 之后，借鉴协同过滤的思想，从读者的阅读兴趣角度出发，引入图书两两之间的加权有向不对称联系 borrowIntent。

[0043] borrowIntent 的具体定义为：若有 n_1 个读者借阅了图书 b_1，n_2 个读者借阅了图书 b_2，$b_1 \rightarrow b_2$ 的边权重 (link weight) 为：$borrowIntent(b_1, b_2) = |n_1 \cap n_2|/n_1$，同理，有 $b_2 \rightarrow b_1$ 的边权重：$borrowIntent(b_2, b_1) = |n_1 \cap n_2|/n_2$

[0044] 建立领域本体后，进行日志整理：

[0045] 通常情况下，日志记录的格式比较混乱而且会含有大量无用信息。因此，首先需要整理日志，去除不合法或者错误的记录，比如带有“MISSING”或者“??”。然后，将所有日志信息整理成表 1 的格式并存入关系数据库。其中，entry_id 表示记录编号，book_id 表示该图书的中图分类法编号，user_id 为用户编号（用户编号只为系统区分用户所用，不能推断出任何用户信息，不涉及用户隐私），timestamp 为该条记录的日期。

[0046] 表 1 日志信息表

[0047]

entry_id	book_id	user_id	timestamp
1	B516.47/9.2	00000001	2008-01-02
...	...	...	...
389,138	C37/2	00010009	2008-06-30

[0048] 同时，需要维护另一张图书信息的表，其中，book_title 表示该本图书的完整书名。

[0049] 表 2 图书信息表

[0050]

book_id	book_title
I712.45/598	达·芬奇密码
...	...
K835.4657/6e	达·芬奇画传＝Da vinci

[0051] 步骤 2,建立用户模型 :对日志进行分析处理,分析用户历史记录,由于读者的兴趣随着时间的推移不断发生变化,按照时间顺序对用户借阅的图书进行分类和加权,从而建立起基于用户兴趣偏好的用户模型。具体过程如下 :

[0052] 通过分析日志信息表,可以得到特定用户的借阅历史,从而分析用户兴趣,建立用户兴趣模型。本发明将用户兴趣根据时间段分为三类 - 即时兴趣 (本次借阅的其他图书),近期兴趣 (一个月之内借阅) 和长期兴趣 (其他)。对于用户兴趣模型中的每一本图书 i,权值 A[i] 赋予公式如下所示,

[0053]

$$A[i]=\begin{cases}\dfrac{\alpha}{|\text{即时兴趣}|} & \text{如果 i}\in\{\text{即时兴趣}\}\\ \dfrac{\beta}{|\text{近期兴趣}|} & \text{如果 i}\in\{\text{近期兴趣}\}\\ \dfrac{\gamma}{|\text{长期兴趣}|} & \text{如果 i}\in\{\text{长期兴趣}\}\end{cases}$$

[0054] 上式中,本发明经实验对比最终选用的参数为 α ＝ 4, β ＝ 2, γ ＝ 1。直观地讲,表示即时兴趣的重要程度是近期兴趣的两倍,近期兴趣的重要程度是长期兴趣的两倍。

[0055] 由于用户的兴趣并非一成不变,本发明的用户兴趣模型随着时间推移不断更新。

[0056] 第二部分在线部分。在线部分的工作在线上完成,即在用户提交关键词搜索之后完成。在线部分的具体工作步骤如下 :

[0057] 步骤 3,个性化得分计算 :根据已经建立的领域本体和用户模型,通过图挖掘算法 SA 来计算该个性化得分。用 SA 计算个性化得分的具体方法如下 :

[0058] 个性化得分即通过 SA 计算得到的激活值 (Activation Score)。离线部分建立的领域本体是 SA 的传播网络,在该网络中,结点 (node) 表示图书、类别、作者和主题 ;Link(i, j) 表示连接结点 i 与结点 j 的边。SA 传播中,除 borrowIntent,其他边都看作无向边,这样保证了激活值能从书 b_1 传播到 b_1 对应的类别 / 作者 / 主题,再传播到书 b_2。也就是说,所有边之中,只有 borrowIntent 是有向加权边。同时,离线部分得到的用户兴趣模型成为 SA 的加权初始点,其他所有点初始激活值为 0。SA 传播过程中,结点 j 的激活值 A[j] 按以下公式更新。

[0059] $$A[j]=A[j]+\sum_{i\in\{i|link(i,j)\}}A[i]*DecayFactor$$

[0060] 上式中,DecayFactor 为衰减系数,表示由结点 i 传播到邻居结点 j 的衰减后激

活值。本发明按照 Ming-Hung Hsu(Ming-Hung Hsu，Hsin-Hsi Chen.A methodto predict social annotations.CIKM，Napa Valley，CA，USA，2008.）使用的参数设定，衰减系数默认值设为 0.8。与 Ming-Hung Hsu 不同的是，当边为 borrowIntent 时，衰减系数为该边的边权重。

[0061]

$$DecayFactor=\begin{cases}\text{边权重}, & \text{如果} link\in\{borrowIntent\}\\ 0.8, & \text{其他}\end{cases}$$

[0062] SA 传播过程中，为了提高效率，本发明为 SA 设置了以下限制：

[0063] (1) 循环次数限制。在本具体实施方式中，限制 SA 的循环次数为 3。

[0064] (2) 传播路径限制。控制传播的距离，在本具体实施方式中，限制最远传播距离为 2。

[0065] (3) 传播终点限制。在本具体实施方式中，传播终点限制在于，当传播遇到特定点之后，传播停止。

[0066] 需要指出的是，在 SA 的每个循环结束之后，本发明会收集用户反馈信息，来更新下一次传播的初始点激活值。可以收集的用户反馈信息包括：用户新点击链接的图书，用户新借阅的图书。这些图书都将作为即时兴趣，并赋予权值。

[0067] 步骤 4，重排搜索结果：根据所述 SA 得到的个性化得分，按照从高到低的顺序对原搜索引擎返回的结果进行重排，然后返回给用户。具体实施方案如下：

[0068] SA 结束后，个性化重排可以看作根据 SA 得到的最终个性化得分，对原始搜索结果进行重排的过程。

[0069] 根据本发明方法的实际评测结果如下：

[0070] 确定评测指标。评测指标为 Discounted Cumulative Gain(DCG)。由麻省理工大学的 Jaime Teevan 在 J J Teevan，ST Dumais，E Horvitz.PersonalizingSearch via Automated Analysis of Interests and Activities.Proceedingsof the 28th annual internat ional ACM SIGIR，Salvador，2005.New York，ACM Press，2005：178~185 中提出，利用人工对查询结果打分的方式结合 DCG 公式来评测个性化检索系统的方法。此方法依据不同网页在检索结果中排序的不同给其赋予不同的重要度，排序越高的检索结果重要度越大，用户对其的打分对系统性能的影响也越大。因此利用 DCG 公式将用户对检索结果的打分与结果的排序位置结合，计算出的值作为系统性能的评测指标。

[0071] 本发明的测评中，给出的搜索结果里，用户确实借阅的图书 G(i) ＝ 3，其他结果 G(i) ＝ 1，DCG 迭代的计算公式如下

[0072] $$DCG(i)=\begin{cases}G(1) & i=1\\ DCG(i-1)+G(i)/\log(i) & i>1\end{cases}$$

[0073] 由于每次搜索返回的结果数量都不一致，还需要做归一化。越相关结果排序越高的搜索为理想搜索，此时的 DCG(i) 作为 ideal DCG(i)，则最终的评测公式如下所示。

[0074] $$normalized\ DCG(i)=\frac{DCG(i)}{ideal\ DCG(i)}$$

[0075] 显然，标准 DCC(normalized DCG，简称 Norm DCG) 越高，说明搜索结果越与用户兴

趣吻合，个性化搜索的效果越好。

[0076] 评测结果测试数据为北京大学图书馆2008年1月到2008年6月的真实日志数据。本发明的方法于三种其他方法进行了对比，各方法的具体描述如下：

[0077] “Lucene/VSM”方法：开源搜索引擎Lucene的Lucene Score API利用VectorSpace Model(VSM)的原始搜索结果。之所以与Lucene的结果相比较是由于Lucene已经被世界范围内多家数字图书馆采用为索引和搜索引擎，比如佛罗伦萨国家图书馆(http://www.planetware.com/florence/national-library-i-to-fbc.htm)，纽约公共图书馆(http://www.nypl.org/)等等。Lucene的结果能最大限度地模拟现实中各个图书馆的搜索结果。

[0078] “SA”方法：与Ahu Sieg在Ahu Sieg,Bamshad Mobasher,Robin Burke.Websearch personalization with ontological user profiles.CIKM,Lisbon,Portugal,2007.中采用的方法相似：领域本体中没有borrowIntent，也不对用户兴趣进行分类。

[0079] “SA+B”方法：领域本体中加入borrowIntent，但是不对用户兴趣进行分类。

[0080] 本发明的方法“SA+B+S”：领域本体中加入borrowIntent，同时对用户兴趣进行分类。各方法的性能比较结果如下表所示：

[0081] 表3各方法性能比较

[0082]

	未经重排(Lucene/VSM)	个性化重排		
		无 *borrowIntent*(SA)	加入 *borrowIntent*	
			用户兴趣不分类(SA+B)	用户兴趣分类(SA+B+S)
Norm DCG 平均值	0.642	0.725	0.813	0.848

[0083] 基于上表可以看出，本发明的方法性能最优。经过重排SA方法与Lucene得原始结果相比，Norm DCG平均值提高了12.9%，通过引入borrowIntent并在建立用户兴趣模型时对用户兴趣进行分类和加权，本发明方法的Norm DCG平均值达到了0.848。

[0084] 在实际搜索中，用户往往只浏览列在前两页的排名最高的Top N结果。基于此，在设计实验比较本发明的方法与前述三种方法在取Top N结果时的性能，结果如图3，图4所示。从图3，图4的曲线可以明显看出，本发明的方法效果好于其他方法，大大提高了用户感兴趣的结果的排名。

[0085] 本发明并不以上述实施例为限，本发明方法同样适用于电子产品、电子书籍、手机等用户关联度的扩大销售。此外，上述仅为本发明的较佳实施例，并不用来限定本发明的实施范围。也就是说，任何依照本发明的权利要求范围所做的同等变化与修改，皆为本发明的权利要求范围所涵盖。

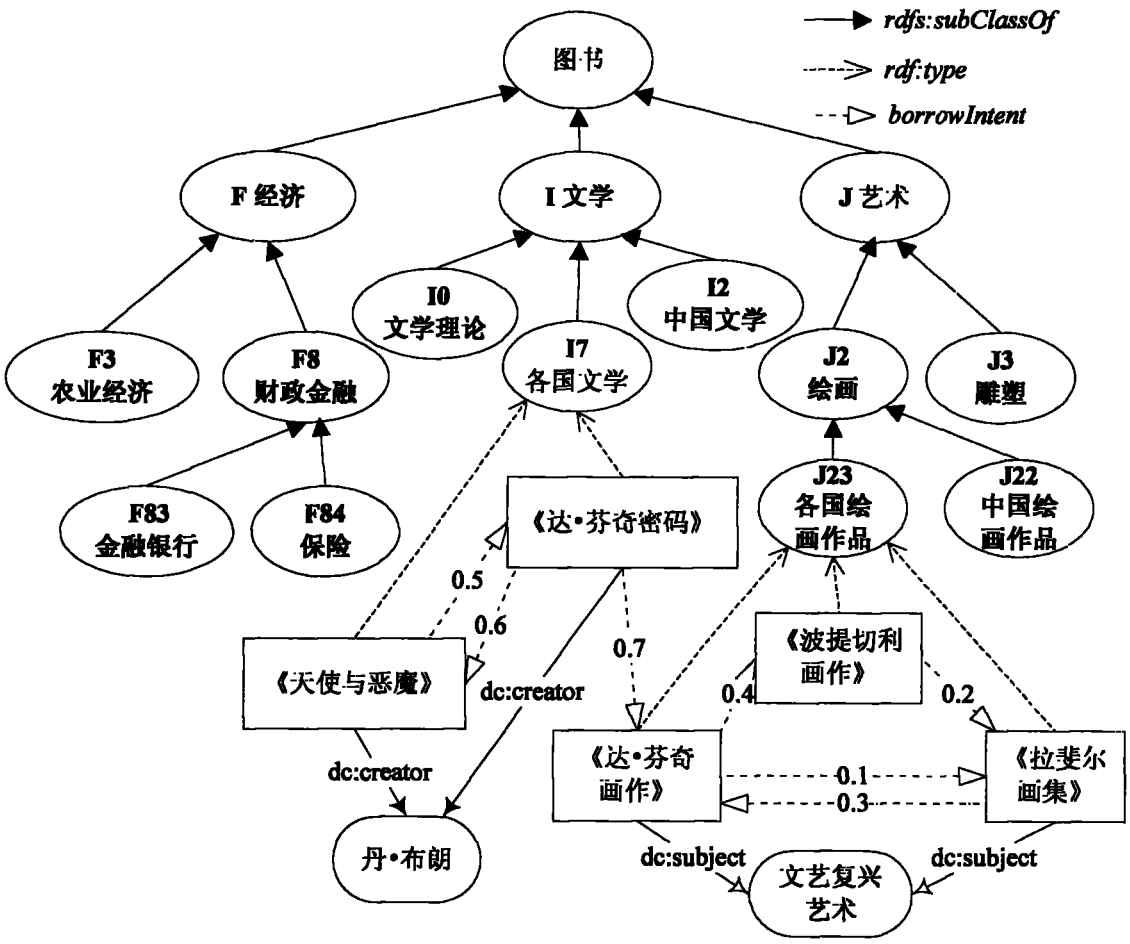

rdfs:subClassOf
rdf:type
borrowIntent
图书
F 经济
I 文学
J 艺术
I0 文学理论
I2 中国文学
F3 农业经济
F8 财政金融
I7 各国文学
J2 绘画
J3 雕塑
F83 金融银行
F84 保险
《达·芬奇密码》
J23 各国绘画作品
J22 中国绘画作品
0.5
0.6
0.7
《天使与恶魔》
dc:creator
《波提切利画作》
0.4
0.2
dc:creator
《达·芬奇画作》
0.1
0.3
《拉斐尔画集》
丹·布朗
dc:subject
文艺复兴艺术
dc:subject

图 1

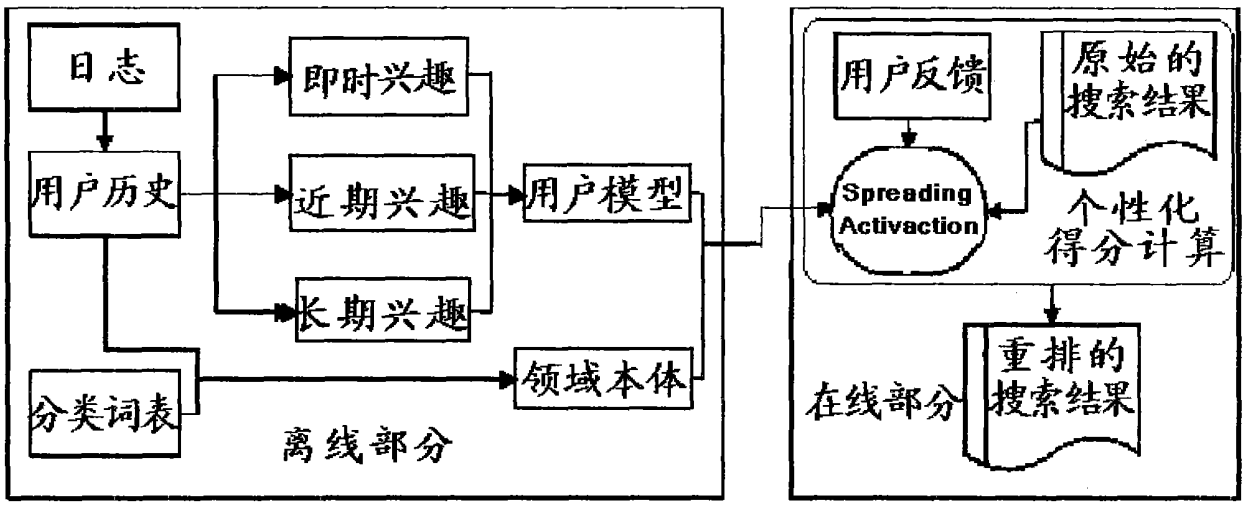

日志
用户历史
即时兴趣
近期兴趣
长期兴趣
用户模型
分类词表
领域本体
离线部分
用户反馈
原始的搜索结果
Spreading Activaction
个性化得分计算
在线部分
重排的搜索结果

图 2

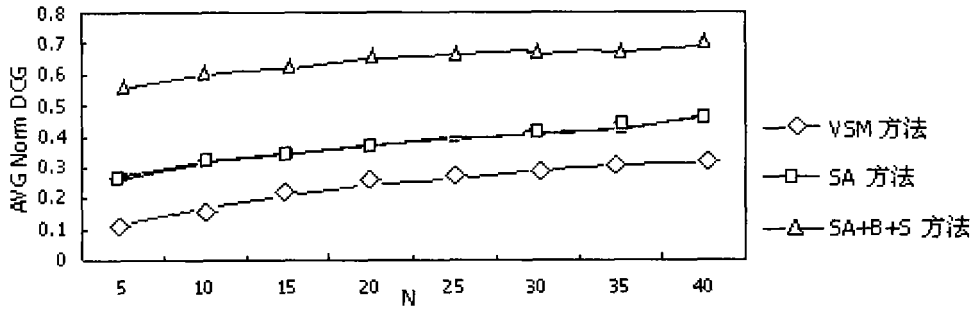

AVG Norm DCG
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30
35
40
N
VSM 方法
SA 方法
SA+B+S 方法

图 3

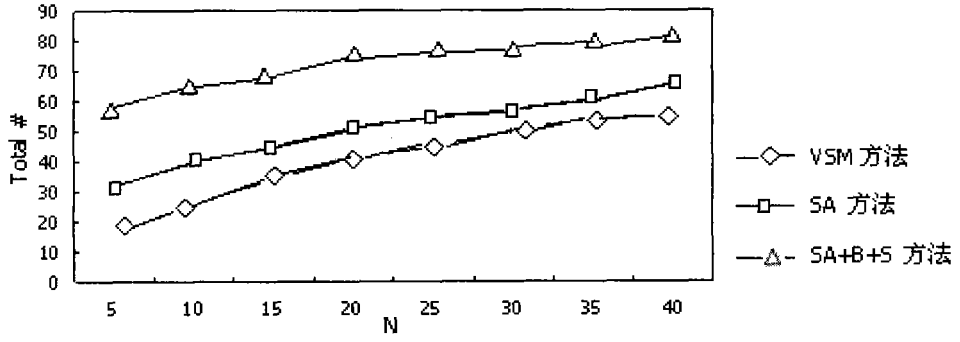

Total #
90
80
70
60
50
40
30
20
10
0
5
10
15
20
25
30
35
40
N
VSM 方法
SA 方法
SA+B+S 方法

图 4