图书情报工作
第57卷 第18期 2013年9月

基于学术产出挖掘的用户兴趣建模研究*

姚晓娜 祝忠明 王思丽

[摘 要] 为了解决用户兴趣建模初期存在的冷启动问题,以科研用户公开发表的学术产出作为用户兴趣建模的数据源,采用文本挖掘和基于本体的模型表示技术进行用户建模,并提出一种通过实体关系表示用户兴趣的方法。该方法与使用单个关键词或实体的表示方法相比,语义信息更为丰富,能更好地描述用户兴趣。最后,将生成的用户兴趣本体实例存储到 Sesame 本体数据库中,支持通过 SeRQL 和 SPARQL 语言进行查询,实现了用户兴趣信息的语义化存储和检索。

[关键词] 个性化服务 用户兴趣建模 文本挖掘 本体 实体关系对

[分类号] G250.76

DOI: 10.7536/j.issn.0252-3116.2013.18.021

数字图书馆个性化服务是基于用户的信息使用行为、习惯、偏好、特点及用户特定的需求,来向用户提供满足其个性化需求的信息内容和系统功能的一种服务[1]。个性化服务实现的核心在于用户建模——建立用户兴趣、目标和行为的数据结构。用户模型所包含信息的丰富程度,决定个性化服务的可靠准确程度和水平[2]。目前,用户建模的主要方法有两种:一种是从用户提供的兴趣描述或样本文档中提取用户兴趣,另一种是对用户对网页的浏览行为进行分析,挖掘用户兴趣和行为模式。第一种方法需要用户主动提供资料,在实际应用中可行性较差。第二种方法是由系统自动地发现用户潜在的兴趣,但前提是系统已有用户一段时间内的行为数据,存在"冷启动"[3]的问题,即在用户使用个性化服务初期由于缺少兴趣信息而无法推测用户的需求。

数字图书馆的用户大多为科研人员,他们的学术产出也是当前数字图书馆的重要资源。学术产出不仅包含了科研人员的领域背景和研究方向,也隐含了关于科研人员研究兴趣的信息。如果能将学术产出作为用户兴趣的数据源,从中挖掘用户兴趣,那么不仅能够丰富个性化服务中的用户模型信息,而且能够避免用户在使用个性化服务系统初期存在的冷启动问题。

1 相关工作概述

目前,基于学术产出挖掘的研究主要集中在文献计量、知识图谱以及社会网络分析等方面,其中也包含一些用户兴趣分析的研究。文献[4]针对计算机领域科学文献数据库 DBLP 的个性化服务问题,提出使用社会网络分析法(social network analysis,SNA)从历史学术产出中挖掘研究社群,主要通过分析作者-会议、作者-会议-主题等关系计算作者之间的相关度,并生成研究社群,最终根据研究社群相关的作者、会议以及主题等信息向作者推荐相关文献。但其中的主题采用频繁 N-grams 表示,缺乏语义信息。在网络个性化推荐和搜索领域,许多研究者将本体技术应用到用户兴趣建模中,以建立语义化、可共享和重用的模型。文献[5]针对个性化推荐系统中的相似度计算问题,提出了一种基于本体的语义相似度计算模型,通过概念属性间的树状关系计算两个实体间的相似度,该方法与传统的 VSM 等方法相比,能识别出同义或近义的实体,正确率更高,但缺点是只考虑了实体在本体中的层次关系,且文档的表示仍是实体的简单集合,没有考虑集合中实体间的关系。文献[6]对目前国内的用户兴趣建模研究现状进行了总结,指出现有研究多集中在

* 本文系中国科学院国家科学图书馆青年人才前沿领域基金项目"基于学术产出挖掘的用户兴趣建模研究"(项目编号:Y200081001)研究成果之一。

[作者简介] 姚晓娜,中国科学院国家科学图书馆兰州分馆/中国科学院资源环境科学信息中心馆员,E-mail:yaoxn@llas.ac.cn;祝忠明,中国科学院国家科学图书馆兰州分馆/中国科学院资源环境科学信息中心研究员,信息系统部主任;王思丽,中国科学院国家科学图书馆兰州分馆/中国科学院资源环境科学信息中心馆员。

收稿日期:2013-07-24 **修回日期**:2013-09-02 **本文起止页码**:122-126 **本文责任编辑**:刘远颖

模型表示上，其中基于关键词和基于概念、层次概念的模型表示较多，基于语义的模型表示较少，同时，用户兴趣模型进化、评价的研究略显不足。

本研究以科研用户公开发表的学术产出为数据源，首先避免了用户兴趣建模初期存在的冷启动问题。然后通过本体技术进行用户兴趣建模，在表示用户兴趣主题时，采用了句法分析技术从文本中抽取实体关系对，映射到本体模型中，并计算用户对实体关系对的兴趣。该方法与单个关键词或实体相比，实体关系对显然包含了更多的语义信息，能更好地描述用户兴趣。

2 用户兴趣模型的表示

2.1 学术产出分析

本研究首先对学术产出中可能包含的兴趣信息进行分析，学术产出有期刊论文、会议论文、学位论文、专著、专利等多种形式，其中期刊论文和会议论文是相对比较容易获取的，本研究的学术产出主要是指这两种。期刊论文和会议论文的格式大致相同，一般都包括作者、发表时间、标题、关键词、摘要、正文和参考文献等信息。其中作者信息可用于描述用户的基本信息，发表时间可用于描述兴趣的时间属性，标题、关键词、摘要以及正文包含了用户在研究领域的兴趣主题，参考文献中包含了用户可能感兴趣的作者、会议、期刊、机构等信息，通过对作者和机构的分析，还能够得到用户兴趣的地理位置分布。

2.2 基于本体的用户兴趣模型

本研究采用本体技术进行模型表示，模型中的所有信息都表示为本体形式，为了使模型可共享和重用，在建模过程中尽量复用已有的本体。用户兴趣模型与学术产出以及复用本体的映射关系如图 1 所示：

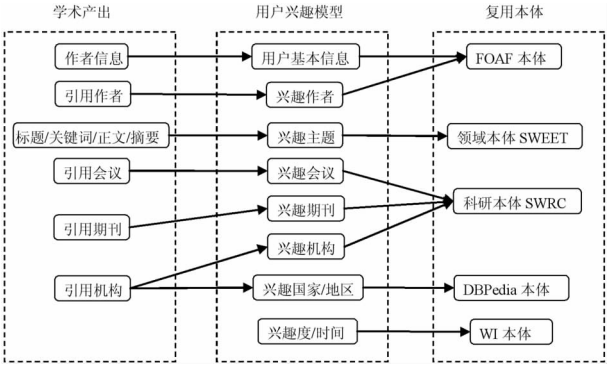

图 1 用户兴趣模型映射关系

本文采用 FOAF(friend of a friend) [7] 表示用户以及用户感兴趣的作者，FOAF 是目前国际上使用最多的描述个人信息的本体。用户在学术领域的兴趣主题需要相应的领域知识库支持，本研究选取 SWEET (Semantic Web for earth and environmental terminology) [8] 作为领域本体知识库。SWEET 是由美国 NASA 开发的一个地球科学本体项目，目标在于实现地球科学数据的互操作。SWRC(Semantic Web for research communities) [9] 是一个描述科研活动的本体，包含了机构、会议、期刊、文献等相关概念，本研究采用该本体表示用户感兴趣的会议、期刊以及机构等信息。DBPedia [10] 是一个基于维基百科的关联数据知识库项目，基本包含了全球所有的国家、地区及城市的名称，因此使用 DBPedia 本体来表示用户感兴趣的国家或地区。WI(weighted interest) 本体 [11] 是一个基于 FOAF 的用户兴趣本体，定义的信息包括兴趣主题、兴趣发生时间、兴趣持续时间、当前兴趣度、累计兴趣度等。WI 本体包含的相关类与属性见图 2，在 WI 本体中，兴趣的内容采用 wi: topic 属性来表示，可以是任意一个本体概念，而感兴趣的程度采用 wo: Weight 类来表示。

图 2 WI 本体的相关类和属性

3 用户兴趣模型的构建

3.1 用户兴趣的数据采集

完成学术产出数据的自动采集是所有研究工作的基础。Web of Science 数据库是国际上使用非常广泛的科学引文数据库，通过 Web of Science 数据库提供的 Web Service 接口 [12]，用户可以通过程序进行数据检索和获取，除了文献自身的元数据之外，还可以获取到参考文献的元数据信息。本文基于 Web Service 接口，以目标用户信息为查询条件，获取该用户发表的论文以及参考文献的元数据信息，存储到本地数据库中。由于 Web of Science 数据库不包含全文，因此，在建模时忽略了此部分信息。

3.2 用户兴趣项的识别

在原始数据中，作者、发表时间、机构、期刊、会议等元数据可以直接作为兴趣项，进行兴趣度的计算。

而对于标题、摘要等非结构化的数据，需要采用自然语言处理技术进行文本挖掘，从中抽取兴趣项。本研究利用开源软件 GATE（general architecture for text engineering，文本工程通用框架）完成对文本中的地名、时间以及领域本体概念等实体的识别，其中本体概念匹配使用的是 GATE 的本体语义词典工具 Onto Root Gazetteer。GATE[13] 是由英国谢菲尔德大学开发的一个开源文本挖掘软件，提供了一系列可重用的自然语言处理的组件和类库，从而能够广泛地应用于各种文本处理任务。除了数据采集和兴趣度计算，本研究的大部分工作都是基于 GATE 的 JAVA API 开发完成的。由于关键词一般都是专业术语，因此将关键词也作为实体，并表示成相关概念类的实例。

在实体识别的基础上，本研究采用依存句法分析技术，对文本中的句子进行分析，从中提取存在依存关系的实体关系对。由于数据源是英文文本，因此下面所述的依存关系特指英文的语法概念。本研究使用 GATE 的 Stanford Parser[14] 句法分析插件完成实体关系对的抽取，其中的依存关系为折叠模式[15]，抽取的依存关系主要包括以下几种：

- 定语关系。对应于 Stanford Parser 中的 amod 或 nn，一般是多个领域本体概念组成的复合短语，如 climatic warming 和 rain distribution，本研究将这些复合短语也作为命名实体，在完成定语关系的抽取后，重新进行句法分析，并进行其他关系的抽取。
- 介词关系。对应为 Stanford Parser 中的 prep_in、prep_of 等以 prep 开头的关系，如“carbon cycle in China”，carbon cycle 依存于 China，关系为 prep_in。
- 嵌套介词关系。上述介词关系中某些依存对象不是实体概念，如“release of 553 Tg of CO2”，release 依存于 Tg，但 Tg 不是实体概念，而 Tg 与 CO2 之间存在介词关系，针对这种情况，通过 Tg 确定 release 和 CO2 的关系对。
- 主谓宾关系。对应为 Stanford Parser 中的 nsubj 加 dobj 关系，如“human disturbance has greatly impacted the floristic composition”，谓语 impacted 同时依存于 human disturbance 和 floristic composition，关系分别为 nsubj 和 dobj。
- 分词修饰 + 介词关系。对应为 Stanford Parser 中的 partmod 加 prep 开头的关系，如“organic carbon released through land desertification”，organic carbon 依存于 released，released 依存于 land desertification，通过 released 确定 organic carbon 和 land desertification 的关系对。
- 并且关系。对应于 Stanford Parser 中的 conj_and 关系，表示两个实体概念常一起出现，如“soil moisture and soil thickness”。
- 同位语关系。对应于 Stanford Parser 中的 appos 关系，如“middle reach of the Heihe River，China”，Heihe River 依存于 China。
- 被动关系。对应于 nsubjpass 加 prep 开头的关系，如“organic carbon was released from desert lands”，released 同时依存 organic carbon 和 desert lands，关系分别为 nsubjpass 和 prep_from。

本研究自定义本体类 user: Relation、属性 user: subject、属性 user: predicate 以及属性 user: object 进行实体关系对的表示，user: subject 对应关系对中依存的一方，user: object 对应被依存的一方，user: predicate 表示关系的类型，“carbon cycle in China”在本体中的表示见图 3，其中 China 是一个地名实体，表示为 spac2: Location 的实例。

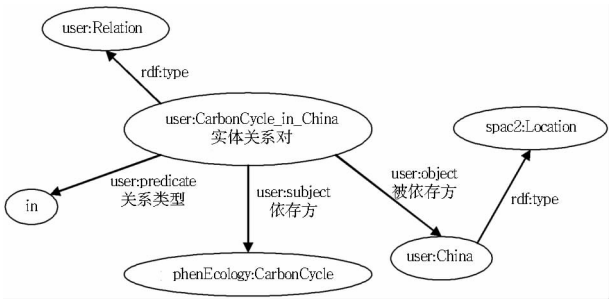

图 3　实体关系对的本体表示

3.3　用户兴趣度的计算

兴趣度用于表示用户感兴趣的程度，对应于 0 到 1 之间的一个实数，0 和 1 分别表示无兴趣和最大兴趣。本研究首先对学术产出中可能对兴趣度产生影响的因素进行了分析，用户感兴趣的作者、会议、期刊、机构、国家/地区这些兴趣项，来源于参考文献，因此只受出现次数和出现时间的影响。而用户在研究领域的兴趣主题，还受到出现文档数和出现位置（标题、关键词、摘要）的影响。

假设兴趣项集合 C（集合中的兴趣项为同一类型，如兴趣作者），兴趣项 c 的影响因素及函数表示如下：①出现次数 freq（c）；②出现文档数 doc（c）；③在题名中出现次数 title（c）；④在关键词中出现次数 keyword（c）；⑤在摘要中出现次数 abstract（c）；⑥在参考文献中出现次数 reference（c）；⑦最近出现时间 time（c）。

兴趣度 Interest（c）的计算公式如下：

$$Interest(c) = \frac{\sum_{i \in F} w_i * f_i(c)}{\sqrt{\sum_{c \in C}[\sum_{i \in F} w_i * f_i(c)]^2}} \cdot time(c) \quad (1)$$

其中，F 表示影响因素的函数集合，w_i 表示第 i 个影响因子的权重，权重之和为 1。f_i 表示第 i 个影响因素的计算函数，time(c) 为兴趣度衰减函数[16]，最近出现时间越早，兴趣度越小，计算公式如下：

$$time(c) = e^{-\lambda t} \quad (2)$$

其中，t 对应于最近出现时间的分段函数，5 年为一个时间段，如 1 到 5 年取值为 1，6 到 10 年取值为 5，本研究设定兴趣度的半衰期为 10 年，因此 λ 取值为 0.17。其他影响因素的计算公式如下，其中 count(c) 为兴趣项对应影响因素的出现次数：

$$f(c) = \frac{count(c)}{\sqrt{\sum_{c \in C}[count(c)]^2}} \quad (3)$$

4 实验结果

本研究从 Web of Science 数据库中采集了 10 名科研人员发表的 269 篇文献以及 3 411 篇参考文献的元数据，作为实验的数据源。由于引用作者数量较多，因此只选择第一作者为兴趣项。在计算兴趣度时，为了使得兴趣不过于分散，排除了只出现一次的兴趣项。在进行实体关系抽取的实验时，共计抽取了 4 780 个实体关系对，出现次数超过一次的为 701 个。通过与人工标注的结果进行对比，实体关系抽取的查准率为 35.8%，查全率为 44.7%。经分析，主要是因为摘要中包含的复杂句、长句较多，导致抽取效率较低。

抽取完成后，通过 GATE 的 Sesame[17] 接口，将构建好的用户兴趣实例写入 Sesame 本体数据库。Sesame 支持通过查询语言 SeRQL 和 SPARQL 进行基于图模式的检索，下面以查询用户 User1 最感兴趣的前 10 个实体关系对为例，SPARQL 查询语句如下：

```
PREFIX user: <http://semantic.llas.ac.cn/user-interest/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX wi: <http://purl.org/ontology/wi/core#>
PREFIX wo: <http://purl.org/ontology/wo/core#>
SELECT ? t ? v
WHERE
{
user:User1 wi:preference ? p.
? pwi:topic ? t.
? t rdf:type user:Relation.
? pwi:overall_weight ? w.
? w wo:weight_value ? v
}
ORDER BY DESC(? v) LIMIT 10
```

查询结果见图 4，其中 T 列为关系对实例，V 列代表兴趣度。

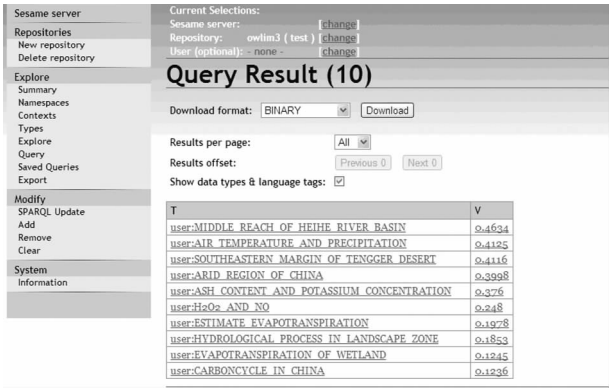

图 4 SPARQL 查询结果：用户 1 最感兴趣的前 10 个实体关系对

5 结 语

实体关系对不仅包含单个的实体概念，还包含了实体间的语义关系，与单个实体或关键词组成的兴趣集合相比，由实体关系对组成的兴趣模型显然能更准确地描述用户兴趣。本研究采用 Sesame 本体数据库进行用户兴趣模型的存储和检索，也为以后实现基于本体的个性化服务做好准备。笔者下一步拟将用户兴趣模型应用到个性化服务系统当中，研究相应的模型更新算法，并通过系统性能的改变来实现对模型的评价。

参考文献：

[1] 郭海明，刘昆雄. 数字图书馆个性化服务方式综述[J]. 津图学刊，2003，21(6)：33－36.

[2] 应晓敏. 面向 Internet 个性化服务的用户建模技术研究[D]. 长沙：国防科学技术大学，2003.

[3] Cold start [EB/OL]. [2013－08－14]. http://en.wikipedia.org/wiki/Cold_start.

[4] Zaiane O，Chen Jiyang，Goebel R. Mining research communities in bibliographical data [C]//Advances in Web Mining and Web Usage Analysis. Berlin：Springer-Verlag，2009：59－76.

[5] Hajian B，White T. Measuring semantic similarity using a multi-tree model [C]//Proceedings of the 9th Workshop on Intelligent Techniques for Web Personalization & Recommender Systems. Barcelona. CEUP Workshop Proceedings，2011：7－14.

[6] 孙雨生，刘伟，仇蓉蓉，等. 国内用户兴趣建模研究进展[J]. 情报杂志，2013，32(5)：145－149，165.

[7] FOAF vocabulary specification [EB/OL]. [2013 - 07 - 06]. http://xmlns.com/foaf/spec.
[8] SWEET Ontologies [EB/OL]. [2013 - 07 - 06]. http://sweet.jpl.nasa.gov/sweet.
[9] SWRC Ontology [EB/OL]. [2013 - 07 - 06]. http://ontoware.org/swrc.
[10] DBPedia Ontology [EB/OL]. [2013 - 07 - 06]. http://wiki.dbpedia.org/Ontology.
[11] The weighted interests vocabulary [EB/OL]. [2013 - 07 - 06]. http://smiy.sourceforge.net/wi/spec/weightedinterests.html.
[12] Web of science Web services [EB/OL]. [2013 - 07 - 06]. http://wokinfo.com/products_tools/products/related/webservices.
[13] GATE, a general architecture for text engineering [EB/OL]. [2013 - 07 - 06]. http://gate.ac.uk.
[14] The Stanford Parser [EB/OL]. [2013 - 07 - 06]. http://nlp.stanford.edu/software/lex-parser.shtml.
[15] Stanford dependencies [EB/OL]. [2013 - 07 - 06]. http://nlp.stanford.edu/software/stanford - dependencies.shtml.
[16] 潘建国. 基于语义的用户建模技术与应用研究 [D]. 上海: 上海大学, 2008.
[17] OpenRDF.org [EB/OL]. [2013 - 07 - 06]. http://www.openrdf.org.

User Interest Modeling Based on Literature Mining

Yao Xiaona　Zhu Zhongming　Wang Sili
Lanzhou Branch of the National Science Library/Scientific Information Center for Resources and Environment, CAS, Lanzhou 730000

[Abstract] In order to solve the cold start problem existing at the beginning of user interest modeling, this paper puts the literature published by researchers as the data source of user interest modeling, and applies the technology of text mining and ontology modeling in user modeling, and proposes a method which describes user interest by using entity relationship pair. Compared to the method by using single keyword or entity, this method has more semantic information and can describe user interest better. Finally, the individuals of user interest ontology are stored into the Sesame ontology database and can be queried through SeRQL and SPARQL, and the semantic storage and retrieval of user interest information is implemented.

[Keywords] personalized service　user interest modeling　text mining　ontology　entity relationship pair

(上接第 97 页)

Study On In-depth Integration Services of Internet Resources of Discipline Book Reviews in University Libraries

Li Ming
School of Information Management, Nanjing University, Nanjing 210093

[Abstract] Discipline book reviews about teaching and research in universities are the running-up and most valuable reference internet resources. Based on three layers of internet resource navigation, database construction, deeply mining resources and optimizing services of discipline book reviews, the practicable ways is analyzed to step-by-step develop in-depth integration services of internet resources of discipline book reviews. It hopes to provide some new ideas for university libraries to develop deep discipline-oriented services.

[Keywords] discipline book review　discipline-oriented service　integration of Internet resources　university library