# 基于认知广度和深度的个性化信息检索模型*

邹 海[1,2],郇秀花[1,2]

(1.教育部智能计算与信号处理重点实验室,安徽 合肥 230039;
2.安徽大学 计算机科学与技术学院,安徽 合肥 230039)

**摘 要**:受心理学激活-扩散模型的启发,提出了在领域本体基础上的用户认知结构模型。该模型依据用户提供的认知中心,一方面,根据领域本体中概念之间的语义相关性推导出用户的认知范围;另一方面,根据概念之间的语义相关度刻画出用户的认知深度。从认知范围和认知深度两方面,描述用户对某领域知识的认知结构。实验结果表明,该模型与通用本体模型相比,具有较高的查准率。

**关键词**:认知结构;激活扩散模型;认知广度;认知深度;个性化检索

**中图分类号**:TP391 **文献标识码**:A **文章编号**:1674-7720(2011)13-0088-04

# Personalized information retrieval model based on cognitive range and depth

Zou Hai[1,2], Huan Xiuhua[1,2]

(1.Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Hefei 230039, China;
2.The School of Computer Science and Technology, Anhui University, Hefei 230039, China)

**Abstract**: It proposes a model to get user's individual cognitive structure. It is inspired by the spreading-activation model of psychology and is built on the domain ontology. According to cognitive center which is provided by the user and semantic association between some concepts, this model can not only infer to user cognitive range, but also inscribe user cognitive depth. The experimental results show that this model is more accurate than the normal Ontology model.

**Key words**: cognitive structure; spreading-activation model; cognitive range; cognitive depth,; personalized information retrieval

随着信息的急剧膨胀,人们希望借助信息检索工具如搜索引擎来获取自己需要的信息显得尤为迫切。然而,传统的基于关键词匹配的信息检索技术往往只是得到“千人一面”的检索结果,难以理解用户检索目的和区别用户的需求。造成这种情况的主要原因有两方面:一是当前的互联网不能恰当的处理语义;二是缺乏对用户的理解。针对第一个原因,Tim Berners-Lee 提出了语义 Web(Semantic Web)的概念[1]。其引入了以本体(Ontology)来表示概念和语义关联信息这一思想,来实现不同系统之间的信息共享,提高网络服务的智能化与自动化。语义 Web 通过把当前 Web 上无序的信息变为有序的知识,为解决数据管理有序性与 Web 上信息无序性相矛盾,搜索引擎的查全查准要求与数据缺乏语义相矛盾等问题指明了方向[2]。针对第二个原因,许多学者引入了用户上下文信息,如用户工作内容、专业背景、兴趣、爱好、生活习惯、经验、点击反馈、用户认知(Cognition)、理解水平等因素都属于用户上下文信息。这些上下文信息都是理解用户个性化需求的关键信息。

随着语义 Web 的研究,人们纷纷在本体的基础上对上下文信息进行分析和描述[3],这些研究具有以下特点:

(1)研究对象仅仅只是用户的兴趣,缺乏从多角度对用户个性化需求,如理解水平、认知结构等的理解和挖掘。

(2)分析只是集中于利用上下语义关系,缺乏精确的分析和表示。这些研究工作大都基于 WordNet、dmoz ODP(Open Directory Project)之类的通用本体,只在概念间的父子关系基础上进行分析,而不能从细粒度上对用户的兴趣进行精确分析和表示。

*基金项目:国家科技重大专项资助项目(2008ZX05039-004)

(3)研究方法多集中在定性的分析,缺乏定量分析和描述。这些研究大部分从父子语义关系入手来描述用户兴趣范围,缺乏对用户兴趣深度的描述和表示。如文献[4]的正例/反例扩展向量和文献[5]中的个性化层次树,只要描述的关键词相同,那么用户的个性化模型也必然相同。

心理学上认为,人们的兴趣、认识和情感密切联系。认识越深刻,情感就越丰富,兴趣也就越浓厚。用户的爱好、理解水平、表达等都和用户认知结构紧密相关[6]。因此,从用户的认知结构入手可以更好地理解用户的个性化需求。尤其在专业领域范围内,用户的检索目标往往和自身在该领域的认知结构相适应。

受认知心理学上激活-扩散模型(spreading-activation model)的启发,本文提出了一种基于领域本体来描述用户认知结构的模型 ObSAM (Ontology based Spreading-Activation Model)。激活-扩散模型是认知心理学领域里一种表征个体知识的模型,它认为个体内部知识不是按照层次组织的,而是根据概念间的语义关系或者语义之间的距离来组织和表示的。当概念在用户大脑里出现时,用户语义记忆中相对应的概念节点会被激活,被激活了的概念节点就开始扩散到其他的概念上,尤其会扩散到那些在语义上有紧密联系的概念。根据这个模型,本文提出了用户认知结构模型,依据用户给出的认知中心概念,一方面,根据领域本体中概念之间的语义相关性推导出用户认知范围;另一方面,通过概念之间的语义相关度刻画出用户认知深度,从这两个方面描述用户对某领域知识的认知结构。

## 1 激活-扩散模型

1968 年 Quillian 提出了最早的语义记忆模型。在这个模型中,他用 type 来描述概念,用 token 描述词语,用带有标签说明的激活扩散行为来描述两个节点之间关联时涉及到的中间节点。1975 年 Collins 和 Loftus 最早提出了激活-扩散模型。他们认为个体内部知识不是按层次组织的,而是根据语义关系或语义之间的距离来组织和表示的,并提出了描述人类认知的激活-扩散模型。

激活-扩散模型认为,个体头脑里所存储的知识是一种组织巨大的概念网络,概念之间是通过语义关系相关联。激活-扩散模型有两个关于知识结构的假设:(1)连接节点的线段表示概念之间的联系,连线越短,表明两个概念之间的联系越紧密;(2) 语义的距离是知识组织的基本原则,即概念的内涵是由它相关联的其他概念,特别是联系密切的概念来确定的。它认为,当概念出现时,认知中相应的概念节点会被激活,被激活了的概念节点就开始扩散到其他的概念,特别是那些在语义上有紧密联系的概念。而激活-扩散的远近主要由以下因素决定:最初被激活节点的激活强度、从最初被激活的节点到目前节点的语义距离、扩散时间等。

20 世纪 80 年代,激活-扩散模型已经被应用到信息检索领域,主要运用在文档和词汇查询过程中用以扩展词汇和文档集。F.Crestani 曾经综述了激活-扩散模型在信息检索领域中的应用,指出了激活-扩散模型中典型的四点约束:扇出约束、路径约束、距离约束以及激活约束。本文试图在信息检索领域直接按照激活-扩散模型的本意来描述用户的认知结构,并把它应用到个性化信息检索中。

## 2 基于领域本体的认知模型

### 2.1 基本定义

定义 1 领域本体:一个领域本体是关于领域知识的概念以及概念之间的关系集合,用二元组定义 $O=\{C,S\}$,$C$ 表示概念的集合,$S$ 表示概念之间的语义关系集合。

要构建用户的认知结构,需要用户先给出若干个描述其认知结构的中心概念。

定义 2 认知中心概念:由用户 $u$ 指定的,描述在该领域内比较关注和掌握的领域本体概念,称为用户 $u$ 认知中心概念。由用户的认知中心概念构成的集合被称为用户的认知中心 $V_u$。

定义 3 概念认知深度 DOC(Degree Of Cognition):用户 $u$ 对概念 $C_j$ 赋予一个数值 $DOCu(C_j)$,描述对该概念的掌握程度,$0<DOCu(C_j)\leqslant 1$,称为用户 $u$ 对概念 $C_j$ 的概念认知深度。

定义 4 基于领域本体的认知结构模型 ObSAM (Ontology based Spreading-Activation Model):给定一个领域本体 $O=\{C,S\}$,$\theta$ 为用户认知结构扩展的阈值,$V_u$ 是用户给定的认知中心,用户 $u$ 在领域本体上的认知结构模型 $ObSAM\ \ O_u=\{C',S\}$ 定义如下:

(1) $C'=\{Ci|DOC_u(Ci)\geqslant\theta\}$

(2) $S'=\{(C_i,C_j)|(C_i,C_j)\in S,C_i\in C',C_j\in C'\}$

### 2.2 语义相关度

由于 ObSAM 模型中,需要根据概念之间的语义相关度刻画用户的认知深度,下面引出关于本体中语义关系和语义相关度的形式化定义。

(1) 语义等价关系:如果 $x$ 被定义为 $y$ 的 owl:equivalentClass,则称 $x$ 和 $y$ 语义等价,表示为 $x\equiv y$ 或 $y\equiv x$。

owl:equivalentClass 意味着两个概念有相同的概念外延(即它们包括同样的实例集合)。

(2) 语义父子关系:若 $x$ 被定义为 $y$ 的 rdfs:subClassOf,则表示 $x$ 被 $y$ 语义包含,忽略概念包含它自身的情况,表示为 $x\subset y$。

rdfs:subClassOf 意味着属于 $x$ 概念外延实例的集合是 $y$ 概念外延的实例集合的子集。

(3) 若 $x$ 被定义为 $y$ 的 owl:ObjectProperty 或 rdf:Property,则称 $x$ 和 $y$ 语义关联,表示为 $y\propto x$。

Owl:ObjectProperty 或 rdf:Property 表示 $x$ 和 $y$ 通过属性关联，其中 $x$ 是域概念，$y$ 是范围概念。

(4)语义相关度(DSA)：如果领域本体中从概念 $x$ 到概念 $y$ 存在一种语义关系 $r$，则存在一条从概念 $x$ 到概念 $y$ 的有向边，并且定义 $w_x(y;r)$ 为这条边上的权值，它表示概念 $x$ 经 $r$ 语义关系到概念 $y$ 的语义关联程度。

根据本体上两个相邻概念之间的语义关系，给出 MDSA(Macro Degree of Semantic Association)：领域本体中任意概念之间的语义相关度。定义如下：

$$MDSA(C_i,C_j)=\begin{cases}1.0,\text{if } C_i\equiv C_j\vee C_i==C_j\\ W(C_i,C_j,r),\text{if } C_j\propto C_i\vee Cj\subset Ci\\ 0.0,\text{if}\neg(C_j\propto C_i\vee C_j\subset C_i\vee C_i\equiv C_j\vee C_i==C_j)\\ \text{Max}\{W(C_k,C_j,r)\times MDSA(C_i,C_k)\},\text{otherwise}\\ r\in(C_j\propto C_k,C_j\subset C_k,C_j\equiv C_k)\end{cases}\quad(1)$$

其中，$C_i==C_j$ 表示 $C_i$ 和 $C_j$ 是同一个概念。根据式(1)，对间接相邻的概念 $C_j$ 和 $C_i$，若 $C_j$ 到 $C_i$ 只有一条同向可达的路径，则路径上的语义相关度乘积便为从 $C_j$ 到 $C_i$ 的语义相关度；若 $C_j$ 到 $C_i$ 有多条同向可达的路径，则路径上的最大 MDSA 便为从 $C_j$ 到 $C_i$ 的语义相关度；若 $C_j$ 到 $C_i$ 不存在同向可达的路径，则 $C_j$ 到 $C_i$ 的语义相关度为 0。

### 2.3 基本思想

由于用户指定的认知中心概念数目不会太多，所以用户可以给出每个认知中心概念的认知深度。设用户为认知中心概念指定的概念认知深度为 $\lambda_i$，$0\leqslant\lambda_i\leqslant 1$。但是用户不能给出所有概念的认知深度，下面给出用户 $u$ 对任意概念 $C_i$ 的概念认知深度：

$$DOC_u(C_i)=\begin{cases}\lambda_i,\text{if } C_i\in V_u\\ \text{Max}\{DOC_u(C_j)\times(MDSA(C_i,C_j)+MDSA(C_j,C_i))\},\text{if } C_i\in(C-V_u)\\ C_j\in V_u\end{cases}\quad(2)$$

认知中心概念是由用户指定的，它相对应的概念认知深度也是由用户给定的。对领域本体中其他的概念，通过式(2)推导出用户对这些概念的认知深度，即概念认知深度是随着它们和认知中心概念关联强度的变化而变化。这种推导方式来源于认知心理学中的激活-扩散模型，即激活扩散的远近一般由最初被激活节点的激活强度、从最初被激活的节点到目的节点的相关程度等因素影响。

对用户给定一个深度阈值 $\theta$，$0\leqslant\theta\leqslant 1$，并且 $0\leqslant\theta\leqslant\min(\lambda_i)$（其中 $\lambda_i$ 为用户对认知中心概念 $C_i$ 给定的概念认知深度）。以用户的认知中心 $V_u$ 为中心，可以依据概念相关度在领域本体内进行概念扩展，形成用户认知结构模型 ObSAM，ObSAM 从广度和深度两个方面描述出用户在对领域知识的认知程度。

例如，假设用户 $u$ 给出认知中心概念为(经济危机，金融危机)，给定相应的认知深度为(1，0.9)，指定的深度阈值为 0.5。结果在生成的 ObSAM 中，共有概念为 15 个(包括 2 个认知中心概念)。图 1 显示了该用户关于经济方面的认知结构。

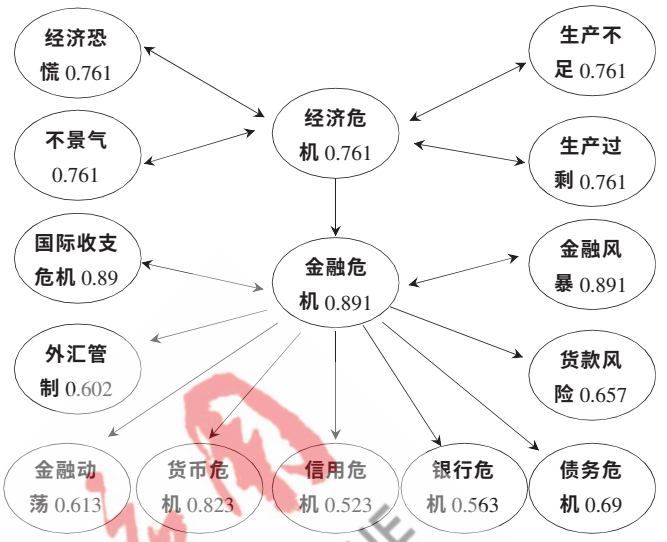


图 1 用户的认知结构

根据认知心理学上的激活-扩散模型，基于对领域本体精确丰富的语义关系的分析和利用，ObSAM 从深度和广度入手描述了用户对领域知识的认知结构。传统关键词列表在描述用户个性化需求时，缺乏从完整的体系中考虑并利用关键词之间的语义相关性，因此不能准确地定性分析；而近年来发展的基于本体用户个性化表示方式，往往是基于大型的概念层次结构如 WordNet、Yahoo！等，一方面缺乏对语义关系的精细分析与利用，另一方面由于过于庞大而很难从定量的角度分析利用。利用 ObSAM 描述用户个性化需求，一方面从领域知识定性的角度分析用户对领域知识的认知范围，另一方面从定量角度分析用户对领域知识的概念认知深度。表 1 中列出了 ObSAM 和其他表示方式的异同。

表 1 ObSAM 和其他模型的比较

| | 来源 | 表现形式 | 特点 | 分析方法 |
|---|---|---|---|---|
| 传统的关键词列表 | 同义词典等 | 向量结构 | 同义关系利用 | 简单定性分析 |
| 其他基于本体表现形式 | 通用本体如 wordnet 等 | 层次结构 | 父子语义关系的定性分析与应用 | 定性分析 |
| ObSAM | 领域本体 | 网状结构 | 基于父子、关联等语义关系描述 | 定性分析和定量描述 |

## 3 实验

### 3.1 实验设置

为了表现出在领域本体上构建模型 ObSAM 比通用本体有优势，实验中采用了 2 个本体进行对比，一个是通用本体 WordNet，另一个是经济学领域本体 EO(economic ontology)(假设该领域本体包含所有的经济领域词汇)。WordNet 的读取采用了 SourceForge 开放源码社区提供的 JWNL 接口(http://sourceforge.net/projects/

jwordnet)；EO 是 NSFC 资助项目“通用网上知识编辑器及示范主题语义网研究”的一部分成果，基本包含了经济学领域的重要概念和关系。

对应于两种不同的本体，相应采用的测试数据集是：一个是美国国家标准技术局 NIST (National Institute of Standards and Technology) 与 2004 年公开发布的 TREC2001 Filtering Track 中使用的 REuters 数据集(http://www.jmlr.org/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm)，另一个是中国人民大学数字图书馆个性化服务系统 DLPers V2.0 中的数字资源作为测试数据集。

### 3.2 实验评测标准和实验结果分析

实验主要从查询准确率方面进行评价，查准率采用 *Precision@n* 和 *AP@k* 来衡量。*Precision@n* 是前 $n$ 个结果文档中查询准确率，用来衡量大多数用户关注的前 $n$ 个结果文档的准确率。*AP@k* 用来衡量前 $n$ 个结果文档中相关文档的排序情况。*Precision@n* 和 *AP@k* 在一起能更全面对 top-k 检索结果进行评价，因为大多数用户习惯在检索过程中主要关注 top-k 检索结果[7]。

*Precision@n* 的计算方式是：*Precision@n*=#of relevant docs in top-n retrieved/$n$，其中 $n$ 表示前 $n$ 个结果文档；*AP* 的计算方式是：$AP@k=\frac{1}{r}\sum_{rank_i\leq jk}\frac{j}{rank_j}$ 其中，$r$ 表示前 $k$ 个结果文档中相关文档的个数，$j$ 表示前 $k$ 个结果文档中第 $j$ 个文档；$rank_j$ 表示第 $j$ 个相关文档在结果文档中的排序。通常用户只关注前 20 个检索结果，这里取 $n=k=20$。实验结果如表2 所示。

表 2 两种本体上的查询准确率对比

| | *Precision@20* | *AP@20* |
|---|---|---|
| 通用查询 | 31.46% | 22.45% |
| ObSAW 查询 | 70.41% | 37.62% |

本文以认知心理学上的“激活-扩散模型”为基础，提出了一种基于用户认知结构的 ObSAM 模型。它具有以下优点：(1)它是基于领域本体而不是通用本体。由于人类知识的构建本身是分领域进行的，所以基于领域本体更有利于表达用户的认知结构，可以提供更精确和细致的分析。(2)基于概念之间的概念相关度来合理刻画出用户的认知深度，对用户的个性化需求增加了定量分析，从认知广度和认知深度两个方面，加深对用户个性化需求的理解。

### 参考文献

[1] Berners-Lee T, Hendler J, Lassila O.The Semantic Web-A New Form Of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities [J]. Scientific American, 2001, 284(5):34-43.

[2] Berners-Lee T, Hendler J. Publishing On The Semantic Web-the Coming Internet Revolution Will Profoundly Affect Scientific Information[J]. Nature 2001,410(6832):1023-1024.

[3] Middleton S, Shadbolt N, De Roure D.Ontological user profiling in recommender systems [J]. ACM Transactions on Information Systems 2004,22(1):54-88.

[4] Sieg A, Mobasher B, Burke R, et al. Representing User Information Context with Ontologies [C]. In: Proceedings of 11thInternational Conference on Human-Computer Interaction (HCII2005): Las Vegas, Nevada, USA,2005.

[5] Chaffee J, Gauch S. Personal Ontologies for Web Navigation [C]. In: Proceedings of the ninth international conference on Information and knowledge management; McLean,Va.,USA,2000, P.227-234.

[6] 梁宁建.当代认知心理学[M].上海：上海教育出版社，2003.

[7] 田萱，杜小勇，李海华.语义查询扩展中词语-概念相关度的计算[J]，软件学报，2008，19(8)：2043-2053.

(收稿日期：2011-02-14)

作者简介：

邹海，男，1969 年生，博士，副教授，硕士研究生导师，主要研究方向：数据挖掘与信息检索，中间件理论与技术，工作流理论与技术。

郇秀花，女，1985 年生，硕士研究生，主要研究方向：数据挖掘与信息检索。