# 对基于本体的搜索中用户偏好库的算法研究

韩岳松,李宝敏
(西安工业大学 计算机科学与工程学院,陕西 西安 710032)

**摘 要**:介绍了本体的基本概念及其基本的元素。探讨了在基于本体的智能搜索中用户偏好库的作用和其类型,特别是客观世界中某一特定领域或主题的搜索中用户偏好库的研究。研究了用户偏好库中用户兴趣剖像生成的提取算法,即TF * IDF算法和TF * IWF * IWF算法和基于本体的查询扩展算法。并且讨论了各个算法之间的联系,论述了用户偏好库在基于本体的搜索系统中的独特作用。

**关键词**:本体;用户偏好库;兴趣剖像生成算法;TF * IDF算法;TF * IWF * IWF算法;扩展查询算法
**中图分类号**:TP18 **文献标识码**:A **文章编号**:1673-629X(2007)12-0064-04

# According to User's Preference Storage of Algorithm Research on Intelligent Search Based Ontology

HAN Yue-song, LI Bao-min
(Institute of Computer Science and Engineering, Xi'an Technology University, Xi'an 710032, China)

**Abstract**: Introduced the basic concept and the basic elements of Ontology. Discussed the user's preference storage's function and type in based on the Ontology's intelligent search, specially, in some particular fields or themes user's preference storage' research in the objective world. Research user's preference's information and its extraction algorithm, those are TF * IDF algorithm, TF * IWF * IWF algorithm and inquiry expansion algorithm based on Ontology. And discussed the relations between those algorithms' ties, elaborated the unique function of user's preference in the search system based on Ontology.

**Key words**: Ontology; user's preference storage; interest splits algorithm; TF * IDF algorithm; TF * IWF * IWF algorithm; expansion inquiry algorithm

## 1 本体的相关概念内容

本体这一概念源自于形而上学的哲学分支,原意是指对客观世界的事务进行分解,发现基本的组成部分,进而发现其抽象本质的研究方法。近年来,人们将本体的概念引入人工智能、知识工程和数据处理等领域,用以解决知识概念表示和知识组织体系方面的有关问题。学者们对本体的定义都有着各自不同的看法。综合起来本体可定义为:它是一个关于某些主题的、层次清晰的规范说明,是一个已经得到公认的形式化的知识表示体系,它包含词表或名称表、术语表,词表中的本体全是与某一专业领域相关的,词表中的逻辑声明全部是用来描述那些术语的含义和术语间关系的,即它们是怎样和其它术语相关联的[1]。

本体可以被看作是一座桥,桥的一端是具体的语法形式,另一端是这种表达形式的概念模型,而桥的下面则是难以逾越的语义大河。本体是对领域实体存在本质的抽象,它强调实体间的联系,它通过某些元素来表示知识,这些元素包括:

(1)概念:在某个领域中,包括一般意义上的概念以及任务、功能、策略、行为和过程等,概念在实现过程中通常用类来定义,而且通常具有一定的分类层次关系。

(2)属性:描述概念的性质,是一个概念的性质,是一个区别于其他概念的特性。

(3)关系:表示概念之间的关联,例如,一些常用的关联:父关系、子关系和相等关系等。

(4)函数:表示一类特殊的关系,即由前 $n-1$ 个要素来唯一决定第 $n$ 个要素。

(5)公理:表示永真式,在本体中,对于属性、关系和函数都具有一定的关联和约束,这些约束就是公理。

(6)实例:表示属于某个概念类的具体实体[2,3]。

**收稿日期**:2007-04-12
**基金项目**:国家"星火计划"资助项目(2004EA850069)
**作者简介**:韩岳松(1981-),男,河南安阳人,硕士研究生,研究方向为智能搜索与语义网;李宝敏,教授,硕士生导师,研究方向为计算机系统结构、计算机网络及语义网。

## 2　用户偏好库的作用和类型

根据语义网的发展需要,首先要解决各个知识领域的语义知识的构架。所以,在此主要针对客观世界中某一特定领域或主题的搜索引擎中用户偏好库的研究。

用户偏好库是用于开放引擎运行过程中用户每次输入的信息,经常选择的其感兴趣的信息、用户的知识背景、专业词汇等,是一种作为用户个性特征的记录信息库,为用户快速准确查询提供参考,可以通过反馈来更新。用户的检索请求及反馈结果经机器学习后生成用户兴趣偏好模型,随着用户的不断使用,用户兴趣偏好模型越来越准确。用户偏好库可把偏好分为两种类型。

第一种类型描述了用户的基本背景知识,比如用户的年龄、学历、专业、爱好等。这种类型的特点是结构稳定,容易操作。在基于本体的用户偏好库中,用户可以用类来定义。一个类代表一种类型的用户,每一个具体的用户都是类的一个实例。这种模式的层次结构的优势体现在两个方面:一是在无法获取用户信息或用户信息不全的情况下可以通过类继承关系推导出其基本的特征属性;二是系统可以针对不同的用户类型建立不同的术语系统和检索策略。

第二种类型用于表示用户的兴趣,兴趣信息反映了单个用户的特定的搜索需要,比如说用户习惯于自己专业的那些术语名称,在选择 a,b,c,d 等信息时最多的是选择了哪类信息。因为在很多时候用户的兴趣偏好信息属于一种隐性信息,用户很难对其直接或明确的说明。还有,用户的兴趣知识随着对系统使用的深入而变化,就好像用户偏好不是一成不变的而是动态的。这样以来就要求系统对用户不断深入了解,适应用户变化而做出相应的调整[4]。

所以说在智能搜索系统就要求用户偏好库对用户的兴趣爱好进行深入地提取和使用。

## 3　个人兴趣剖像生成的方法和算法

再来看用户偏好信息的提取。用户偏好库的提取,即剖像文件的生成,主要用两种方法。

1)用户主动把自己的兴趣爱好等个性化的东西通过在线文档形式提供给系统,而系统则进行整理、记录。也可以让用户回答一些问题,然后根据用户选择的答案,启发式地转到下一个问题。当用户搜索的结果经过用户兴趣剖像过滤返回后,让用户对结果给出一个评价,用该评价及用户输入的检索词对用户个人兴趣剖像进行调整。

2)系统可以通过记录用户的历史信息来判断用户的偏好。因为用户所访问的页面和内容大部分是自己的喜好。历史信息的来源多样,可以是用户目录下的文件,也可以是浏览器、缓冲器中的文件等等。

利用第一种方法生成用户兴趣剖像时,要求用户回答一系列问题可以构成一棵兴趣树。从兴趣生成树的根节点往下,提供给用户选择的兴趣越来越具体。从兴趣树的根节点通往叶节点的过程是发现用户兴趣的过程。用户个性化信息提取的过程就是从兴趣树中找出一棵用户兴趣子树的过程,如图 1 所示。

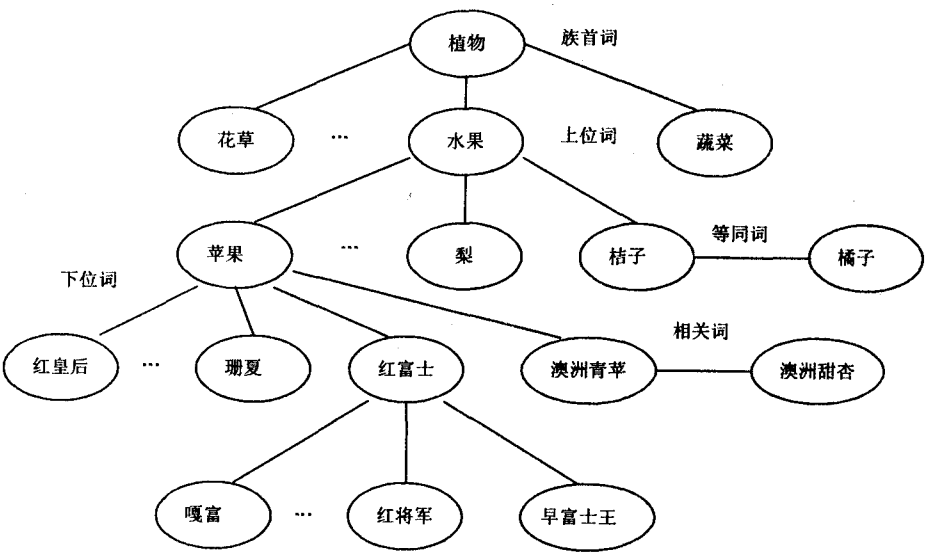


图 1　兴趣生成树

在用户个人兴趣模型中,特别是在其中某一领域或某一主题的用户偏好库中,为了更好地描述用户剖像的生成过程,把用户个人兴趣剖像定义为由二元组组成的集合,即

$$U = \{(L_1, W_1), (L_2, W_2), \cdots, (L_n, W_n)\}$$

其中,$L_K \in T$,$T = \{t_1, t_2, t_3, \cdots, t_m\}$ 为词条集(词典),$W_K \in \{0, \cdots, 1\}$ 为 $L_K$ 词条在用户兴趣剖像中的权重,$\sum_{k=1}^{n} w_k = 1$。

利用此式可得到用户个人兴趣剖像生成算法,根据用户在兴趣生成树上的选择结果得到用户兴趣子树,从用户兴趣子树可以得到用户个人兴趣剖像 $U$。

(1) 置兴趣树中叶节点权 sub $w_i = 1$;

(2) 中间节点权 mid $w_k = \sum_{i=1}^{x} \text{sub } w_i$，其中 $x$ 为该中间节点的下一级节点数目；

(3) 将所有节点按照权值从大到小排序，取前 $n$ 个节点为 $(L_1, V_1), (L_2, V_2), \cdots, (L_i, V_i), \cdots, (L_n, V_n)$，其中 $L_i$ 为词条，$V_i$ 为对应词条在兴趣树中的权重；

(4) 令 $U = \{(L_1, W_1), (L_2, W_2), \cdots, (L_k, W_k), \cdots, (L_n, W_n)\}$，其中 $(L_k, W_k) = (L_k, V_k / \sum_{k=1}^{n} v_k)$，$1 \leqslant k \leqslant n$，$U$ 为用户个人兴趣剖像即用户个性化信息的数学表达。

## 4 TF * IDF 算法和 TF * IWF * IWF 算法

在基于本体的智能搜索中，搜索器把从网络上搜集到的文本或网页信息放到了元数据库中，当用户提出搜索请求后，系统便结合用户偏好库从元数据库中提取用户所感兴趣的信息。而此时元数据库中的信息是繁杂的。如何才能让用户偏好库帮助检索器从这些繁杂的信息中找出用户感兴趣的信息呢？这就要求对元数据库中的文本或网页内容进行特征提取，进而提取利用价值高的文本或网页。使用重要特征词条来表示文本的主要内容，而这些特征词条往往与用户的偏好有着重要的联系。不同的特征词条权重是不一样的，权重由特征词条在文档中的位置和出现的频率所决定。比如特征词条出现在标题中那么这个特征词条的权重一定是较大的，与用户偏好的联系也是紧密的。

在此问题上有两个重要的算法。启发式权重 TF * IDF 算法(Salton,1993)。

计算公式为：

$$W(f_i, d) = \text{TF}(f_i, d) * \text{IDF}(f_i) = N(f_{id}) * \log(N/(f_i/N))$$

其中，$W(f_i, d)$ 是特征词条，$f_i$ 是在文本或网页中的权重，$N(f_{id})$ 是文本 $d$ 中出现 $f_i$ 的次数。

6 年后，在此算法的基础上 Roberto Basili(1999) 又提出了 TF * IWF * IWF 算法(Robert et al,1999)，计算公式如下：

$$W(w_i, d) = \text{TF}(w_i, d) * \text{IDF}(w_i) = N(w_{id}) * (\log(N(w_i/N))^2$$

其中 $N(w_{id})$ 是训练集中出现 $w_i$ 的次数，$N$ 是训练集中所有词出现的次数之和，$N(w_{id})$ 是文本 $d$ 中出现 $w_i$ 的次数[5]。

对于这两种关于特征词权重的算法，尤其是后一种，保证了用户偏好库能够帮助检索器比较准确地从元数据库中找出相关信息。

## 5 扩展查询算法

但在得到了用户偏好后从元数据库中进行查询时，由于各种原因很多时候得不到用户真正需要的信息，这样就必须回过头来利用本体库中的概念信息对用户的兴趣爱好进行“放大”。于是便有了扩展查询算法。

虽然扩展查询算法不能与用户原始需求一致，但可以做到尽可能相关，使得用户在搜索时能够比较准确地找到自己感兴趣的信息。利用领域本体中的概念、属性、关系、函数等信息对用户查询条件进行扩展，可从两方面入手：

1)利用类层次结构的关系，这里的层次关系包括本体知识库和推理器推理得到的信息。领域本体中的概念类层次结构所体现的父子关系可以作为过滤信息时的依据，利用“族首词”或“上位词”的通用概念代替用户的检索概念，或者用抽象的属性代替具体的属性值都可以减少对查询的限制，获得更多的结果。利用“等同词”的专指概念代替用户的检索概念可以获得更深的语义内容和更多的语义表达形式，也可能产生更多的搜索结果。

2)知识库中的概念间除了“族首词”、“上位词”、“下位词”的关系之外还有“等同词”、“替代词”或“用代词”，比如：电子计算机是电脑的“等同词”，电脑、计算机是电子计算机的“替代词”，另外还有“相关词”，比如计算机应用和自动控制是“相关词”。所以说可以利用这些关系划分成不同的类来有效控制，灵活应用。比如利用相似的概念来替代给定的要搜索的概念。

在智能搜索中特别是基于本体的关于某一主题的或某一领域的属性、关系，使用函数、公理、实例可以形成一个节点交错的网络(见图 1)。所以，可以利用本体中的概念关系知识作为搜索系统的启发式知识，将有向网络中的节点作为输入，在搜索的过程中每激发一个链接都将产生一个中间节点，作为下一次激活操作的起始集合，用于记录的兴趣偏好，使下次搜索的查全率和查准率可针对不同兴趣爱好的用户得以有效提高。

如果要搜索关于某个概念的信息，依据的启发知识是领域本体中的父类和子类的关系，则系统进行启发式扩展查询算法可描述为：

$$\text{self} * \Phi_1 \cup (\text{self} \rightarrow \text{ZSub}) * \Phi_2 \cup (\text{self} \rightarrow \text{TSub}) * \Phi_3 \cup (\text{self} \rightarrow \text{ZSub}) * \Phi_4 \cup (\text{self} \rightarrow \text{TSuper}) * \Phi_5 \cup (\text{self} \rightarrow (\text{ExtendBrother})) * \Phi_6 \cup (\text{self} \rightarrow (\text{ExtendTBrother})) * \Phi_7 \cdots$$

其中 self 是本类，ZSub 是直接子类，TSub 是经推理的

得到的子类,ZSuper 是直接父类,TSuper 是经推理得到的父类,Extend Brother 是扩展的兄弟类,ExtendT-Brother 是扩展后推理得到的兄弟类。$\Phi_1,\Phi_2,\Phi_3,\Phi_4,\Phi_5,\Phi_6,\Phi_7$ 表示各节点的权值。一般来说 $\Phi_1$ 的权值最高,往后依次降低。∪ 表示集合并集操作。这样对用户的偏好进行了"放大",依照权值的大小,顺序选择所要搜索的信息,才能提高查准率和查全率。

最后,用户偏好库提取了用户的兴趣爱好,知道它应该选择一些什么样的信息。然后对搜索器从 Internet 抓取回来的信息进行过滤,步骤如下:

(1)对于取回来的文本信息或页面,根据用户个人兴趣剖像中的词条及权重,对该文本信息或页面进行分值计算[2]。

$$\mathrm{Score}(\mathrm{URL}_i) = \sum_{i=1}^{n}\mathrm{Count}(l_i\ \mathrm{URL}_i) * W_i$$

$\mathrm{Count}(l_i,\mathrm{URL}_i)$ 为用户个人兴趣剖像中第 $i$ 个词条在 $\mathrm{URL}_i$ 所对应的页面中出现的次数,$L_i \in T, T = \{t_1,t_2,\cdots,t_m\}$ 为词条集(词典),$W_i \in [0,1]$ 为 $L_i$ 词条在用户兴趣剖像中的权重,$\sum_{i=1}^{n} w_i = 1$。

(2)当对搜索器搜索集的 URL 过滤完全后,按照 $\mathrm{Score}(\mathrm{URL}_i)$ 的分值由高到低依次排序,这样排序高的最有可能就是用户所需要的。

(3)最后用户偏好库要进行自我更新,以便搜索器再搜索时提高查全率。

## 6　结束语

本体作为一种论点新颖、起点较高,并且哲学渊源悠久的知识组织体系,在理论上具备很多优越性和潜在功能。将它应用在智能搜索中,必然有其独到之处。在智能搜索的用户偏好库模型中,在本体论的支持下,采用了用户个人兴趣剖像算法、扩展查询算法,使得在基于本体的智能搜索中更能体现出用户的兴趣爱好,能够大大提高检索的速度和查准率。

**参考文献:**


[1] 李　景.本体理论在文献检索中的应用研究[M].北京:北京图书馆出版社,2005.

[2] 凌　云,王　勋,费玉莲.智能技术与信息处理[M].北京:科学出版社,2003.

[3] 周　宁,张玉峰,张李义.信息可视化与信息检索[M].北京:科学出版社,2005.

[4] 徐宝文,张卫丰.搜索引擎与信息获取技术[M].北京:清华大学出版社,2003.

[5] 秦玄铮.基于本体的个性化信息检索系统的设计与实现[D].北京:北京邮电大学,2006.


(上接第 63 页)

尽可能多的安全数据,经过 Rough 集的数据约简算法,降低了安全数据的冗余,减轻了特征码构造的负担,提取出有效的最简规则集和信息可信度,将特征属性的取值离散划为 4 个区间,进行二进制编码构造特征码,得出最初 Self 集;同时利用上近似 Self 集和下近似 Self 集的构造实现了 Self 集的演化,使得 Self 集具有自动演化更新的能力,为整个免疫系统的构造打下了良好的基础。

当然,在 Self 集的构造和演化中,仍有一些关键的问题没有解决,需要做进一步的工作[6,7]。

**参考文献:**


[1] 李　涛.计算机免疫学[M].北京.电子工业出版社,2004:39-43.

[2] Hofmeyr S,Forrest S. Architecture For An Artificial Immune System[J]. Evolutionary Computation, 2000, 8(4): 443-473.

[3] Forrest S, Hofmeyr S A. Immunology as information processing. In: Segel, Cohen eds. Design Principles for the Immune System and Other Distributed Autonomous Systems. USA: Oxford University Press, 2000: 224-227.

[4] Jin X Y, Du H F, He W H, et al. Optimizing the weights of neural Networks based on antibody colonel simulated annealing algorithm[C]//The International Symposium Neural Networks (ISNN2004). Heidelberg: Springer-Verlag, 2004: 299-304.

[5] 梁意文,李俊涛,郭学理.一种基于用户行为的 Self 集构造和演化方法[J].计算机应用研究,2001(9):7-9.

[6] Garrett S. From Natural to Artificial Immune Systems[M/OL]. users. aber. ac. uk /smg/Modules/CX211_2001-2002/immunity. ppt, 2004: 110-115.

[7] Harmer P K, Williams P D, Gunsch G H, et al. An Artificial Immune System Architecture for Computer Security Applications[J]. IEEE Transaction on Evolutionary Computation, 2002(3): 229-335.

[8] 刘　清. Rough 集及 Rough 推理[M].北京.科学出版社,2001:194-201.

[9] 杨孔雨,王秀峰.入侵检测免疫模型中抗体基因库的生成和进化[J].计算机应用,2003(7):26-28.