广西科学院学报
Journal of Guangxi Academy of Sciences
2006,22(4):256～259
Vol. 22,No. 4 November 2006

一种基于本体的个性化搜索引擎模型*

A Personalized Searching Model Based on Ontology

罗 伟,李陶深
LUO Wei,LI Tao-shen

(广西大学计算机与电子信息学院,广西南宁 530004)
(School of Computer, Electronics and Information, Guangxi University, Nanning, Guangxi, 530004, China)

摘要:利用信息检索、本体和个性化搜索等相关知识,构建一种基于本体的个性化搜索引擎模型 PSMBO。该模型由用户界面、查询请求处理模块、检索模块、查询结果处理模块、兴趣学习模块以及用户兴趣知识库和本体知识库七个部分组成。该模型在一定程度上提高了搜索引擎在查准率和查全率方面的性能。

关键词:信息检索 本体 个性化 搜索引擎
中图法分类号:TP393.09 **文献标识码**:A **文章编号**:1002-7378(2006)04-0256-04

Abstract: The problems of "different expression", "loyal expression" and "mechanical match" in the current search engines are discussed. In combination with the theory and technology of personalized searching service, a personalized searching model (PSMBO) based on ontology is presented. This model comprises of seven modules. They are user interface, search request disposing module, searching module, result disposing module, interest learning module and user interest deposit and ontology deposit. This model improves the performance of search engine in Recall and Precision in certain degree.

Key words: information retrieval, ontology, personalization, search engine

随着 Internet 的日益普及和 WWW 的迅猛发展,网络与人们的工作、生活联系越来越紧密,通过网络获取信息已经成为人们学习和生活中的一种习惯。但是 Web 上的信息内容广泛、结构松散,而且无时无刻不在变化,各种有用没用甚至有害的信息夹杂在一起,造成了网络信息的检索和利用性差,信息检索的查全率(被找到的信息/全部所需要的信息)和准确率(有用的信息/全部查询结果)难以令人满意。目前网络上的搜索引擎主要使用 2 类检索方法[1,2]:一类是基于内容分类的目录式搜索,另一类是基于关键词匹配的全文搜索。目录式搜索主要通过人工发现信息,依靠编目员的知识进行甄别和分类,存在着成本较高、网站描述十分简略、对网站内部细节的描述能力不够深入等缺陷,最终会造成信息丢失的现象。基于关键词匹配的搜索是最基本也是最常用的方法,这种方式虽然可以保证查全率,但却带来了信息过载的问题。

造成这两种搜索引擎信息丢失与信息过载的原因主要有三个方面[2,3]:"表达差异"问题,"忠实表达"问题和"机械式匹配"问题。从本质上来看,搜索引擎缺乏对所检索关键词在语义上的分析和处理才是造成这些检索问题的根本原因。因此,把信息检索从目前基于关键词的层面提高到基于知识(或概念)层面上,就成了解决上述问题的根本和关键。本体作为一种能在语义和知识层次上描述信息系统的概念模型建模工具,具有良好的概念层次结构和对逻辑推理的支持,它在计算机领域中的应用使得把信息检索从基于关键词的层面提高到基于知识(或概念)层面上成为了可能。近年来,Internet 个性化搜索服务引起了人们的关注,利用用户个人的兴趣爱好,对于同样的检索关键词返回给不同用户不同的检索结

收稿日期:2006-07-17
作者简介:罗伟(1981-),男,河南驻马店人,硕士研究生,主要从事智能搜索引擎研究工作。
*广西自然科学基金(桂科自 0640026)项目资助。

果,使得检索结果更加符合用户的查询需要,也可以进一步提高搜索引擎的性能。因此,将本体和个性化搜索服务结合起来,是解决目前搜索引擎存在问题,提高其搜索性能的一种有效途径。本文利用信息检索、本体和个性化搜索服务等相关知识,构建一种基于本体的个性化搜索引擎模型 PSMBO,以解决搜索引擎中普遍存在的"表达差异"、"忠实表达"和"机械式匹配"问题。

1　本体与个性化 Web 搜索

1.1　本体

本体原本是一个哲学上的概念,用于研究客观世界本质。在计算机领域中,本体是对概念化对象的明确表示和描述[4]。简单的说,本体就是一个由概念组成的知识库,其中包含了概念的定义、概念间的复杂关系以及概念推理的规则。本体的目标是捕获相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇(术语)和词汇之间相互关系的明确定义。

1.2　个性化 Web 搜索

所谓个性化 Web 搜索服务[5],既是一种个性化服务,又是一种信息服务,它通过长期观察用户的搜索行为,从中识别用户的信息需求偏好,并且能够根据用户对搜索结果的评价,自觉调整搜索策略。个性化 Web 搜索服务是个性化 Web 信息服务的一个方面,它可以帮助用户更快、更准确地找到所需信息,还可以避免无关信息的干扰,这其实也是搜索智能化的一个方面。

个性化搜索服务体现在两个方面:(1)用户可以使用比关键字表达方式更为方便灵活、符合用户个性习惯的描述方式来表达自己的信息需求;(2)用户能够从多个信息源中获得最贴近自己需要的信息,即针对同一检索关键词,不同用户能够获得不同的检索结果。

2　基于本体的个性化搜索引擎模型 PSMBO

本文将本体和个性化搜索结合起来,提出的基于本体的个性化搜索引擎模型 PSMBO 如图 1 所示。PSMBO 模型包括:用户界面、查询请求处理模块、检索模块、查询结果处理模块、兴趣学习模块以及用户兴趣知识库和本体知识库七个组成部分。

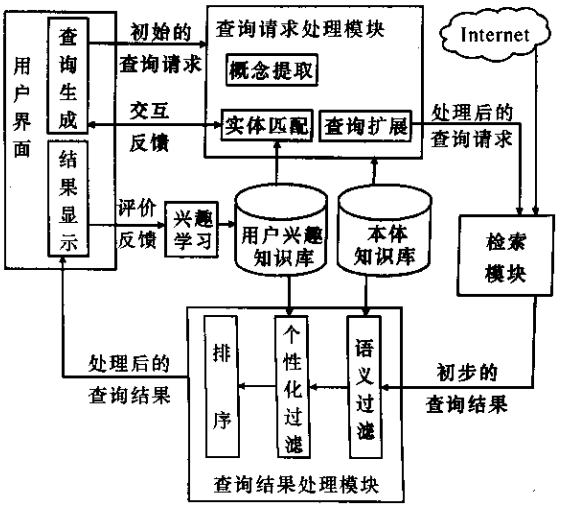

图 1　一种基于本体的个性化搜索引擎模型 PSMBO

2.1　用户界面

在 PSMBO 中,用户界面由查询生成和结果显示两部分组成。

2.1.1　查询生成

查询生成模块给用户提供了一个表达自己查询请求的图形化工具。该工具符合人们的使用习惯,用户可以在图形化的用户界面中输入一系列关键词和布尔操作符(如:and),并选择搜索的一些参数,生成最初的用户查询请求。

2.1.2　结果显示

结果显示模块以统一的格式显示查询结果,从而方便用户浏览。为了使结果清晰明了,该模块采用了两种可供选择的方法:第一种方法是由系统自动完成,按照系统内已设置好的格式去显示查询的结果;第二种方法是允许用户自行定制结果数据的显示格式,这样做会更加符合用户的习惯,且形式上可能会更加简单,但这种方法需要用户具有一定的背景知识,比较适合于专业的用户使用。

2.2　查询请求处理模块

在检索过程中,用户查询请求的准确性直接关系到查询结果的查全率和查准率。查询请求处理模块就是对用户最初的查询请求进行优化,它包括概念提取、实体匹配和查询扩展等三个部分。

2.2.1　概念提取

概念提取模块通过对用户输入的查询语句进行自动分词,经过句法及语义分析,去掉无用的虚词,仅取名词、动词等有实际意义的词和相关词组,提取能正确表达查询语句意义的概念性词或词组,并以此作为用户查询的基本输入概念[6]。例如:用户在利用查询生成模块创建的最初的查询请求是"我想知道该如何使用 word",则通过概念提取之后得到的

查询请求是“使用 word”。

2.2.2 实体匹配

实体匹配模块对概念提取后的查询请求进行语义分析、用户兴趣分析和与用户交互反馈等方式，最终得到用户真正需要查询的实体。在实体匹配过程中，系统首先要对查询请求进行语义分析。例如，用户输入的查询请求为“美洲豹”，通过语义分析可以发现它包含了“动物、汽车和棒球队”等多种含义。此时，实体匹配采用两种方式实现[5,7]：一是系统与用户进行交互，给出显示页面，要求用户自己从排好序的关键词含义中选择最适宜的实体含义。如：系统将关键词“美洲豹”的所有含义返回到用户界面，由用户去选择自己真正感兴趣的实体。如果用户关心的是“美洲豹汽车”，那么用户选择“汽车”。另一种方式是由系统根据用户兴趣知识库自动判别，比如从用户平时使用习惯中了解到用户比较关心汽车方面的网页内容，因而将“美洲豹”自动匹配为“美洲豹汽车”含义。但是应该指出的是，用户的兴趣有时并不能代表自己想要查询内容，因此第一种方式更能准确地进行实体匹配。

2.2.3 查询扩展

查询扩展是在进行实体匹配后，根据本体知识库中关于用户想要查询的实体的概念的定义以及它与相关概念之间的关系，自动地加入新的检索词或短语。扩展的词汇是基于原检索词的同义词以及相关词，也就是说最终的查询请求是基于用户原始查询请求中关键词的一系列同义词及相关词。例如对于“电脑”一词，经过查询扩展后就可将“计算机”也加入到查询关键词中。

2.3 检索模块

检索模块的功能就是根据查询请求处理模块处理后的查询请求，到网站分类或索引数据库中匹配，得到满足要求的网页或 URL。

2.4 查询结果处理模块

查询结果处理模块对检索模块搜索到的结果进行语义过滤、个性化过滤以及排序处理，过滤掉查询结果中的无关信息，并按照信息与用户查询请求及用户兴趣的相关度进行排序，最后将符合用户请求和兴趣的查询结果返回给用户。该模块是实现搜索引擎高查准率的关键部件之一，它由语义过滤、个性化过滤和排序三部分组成。

2.4.1 语义过滤

语义过滤模块根据用户查询请求与查询结果之间的语义相关度，过滤掉检索结果中与用户查询请求不相关的信息。具体的做法是：将用户查询请求和查询的初步结果进行概念提取，然后利用 CWVMA（Concepts-weight vectors matching algorithm）算法[8]计算二者的在概念上匹配度，根据计算过滤不相关结果。

2.4.2 个性化过滤

个性化过滤就是利用存贮在用户兴趣知识库中的知识来评估查询的初步结果，给出结果与用户兴趣的关联度。

2.4.3 排序

排序模块根据语义过滤计算出的语义相关度和个性化过滤得到的用户兴趣关联度，对查询过滤后的结果进行递减顺序的排列，最终把排序的结果返回给用户。

2.5 兴趣学习模块

由于用户的个人兴趣可能会随着工作的需要、环境的改变等条件的影响而改变，处于不断变化的状态当中，这就要求系统能针对用户的兴趣进行学习。兴趣学习模块就是用来完成这个功能的。它接收用户对查询结果的评价和反馈信息，并对它们进行学习，从而建立和修改用户的兴趣知识库。

2.6 用户兴趣知识库

用户兴趣知识库存储的是系统通过对用户个人背景信息、网络使用经验等方面的学习和对用户浏览内容与行为挖掘得到的用户兴趣特征。它是实现个性化搜索的基础，在本搜索引擎的实体匹配、个性化过滤等模块中具有重要的作用。

2.7 本体知识库

在基于语义的搜索机制中，本体知识库是一个捕获语义的重要途径，存放着概念的定义、概念间的复杂关系以及概念推理的规则。它是本搜索引擎的关键组成部分之一，在查询请求处理模块和查询结果的语义过滤中起到非常重要的作用。

3 结束语

与传统的搜索引擎相比，本文提出的基于本体的个性化搜索引擎模型 PSMBO 更具有信息服务的综合性、智能性和个性化等特点，可以在一定程度上提高搜索引擎在查准率和查全率方面的性能。但本模型仍需不断的改进，如本体知识库与用户兴趣知识库的建立和维护、个性化过滤等，这些将是我们今后进一步研究的重点。

参考文献：

[1] NAMBIAR K K. Theory of search engines[J].

Computers and Mathematics with Applications, 2001, 42(12):1523-1526.

[2] 韩婷. 基于本体论的智能搜索引擎模型的研究[D]. 南宁:广西大学,2005.

[3] SUGIURA ATSUSHI,ETZIONI OREN. Query routing for Web search engines: architecture and experiments[J]. Computer Networks,2000,33(1):417-429.

[4] GRUBER CTR. A translation approach to portable ontologies[J]. Knowledge Acquisition,1993,5(2):199-220.

[5] 李雪梅. 基于语义的个性化 Web 搜索[J]. 情报杂志,2003(3):27-31.

[6] 刘维群,李元臣. Web 信息的语义概念检索[J]. 现代情报,2005(7):74-76.

[7] GUHA R,ROB MCCOOL,ERIC MILLER. Semantic search:proceedings of the 12th International Conference on World Wide Web[C]. New York:ACM Press,2003:700-709.

[8] GAO MINGXIA,CHUNNIAN LIU,CHEN FURONG. An ontology search engine based on semantic analysis:proceedings of the Third International Conference on Information Technology and Applications[C]. Washington:IEEE Computer Society,2005:256-259.

(责任编辑:韦廷宗)

(上接第 251 页)

4　结束语

运动性疲劳是运动训练的必然产物,没有疲劳的训练只能是无效训练。疲劳产生的原因,受诸多因素影响,除了生理疲劳外还有心理因素等。本研究对智能技术在竞技项目的应用上作了一些探索性分析,研究结果在空军课题"警卫专业特训学员综合演练系统"中得到了一定的吻合,其长期效果有待进一步观察。

参考文献:

[1] 全国体育学院教材委员会运动生理学教材组. 运动生理学[M]. 北京:高等教育出版社,1990:150-170.

[2] 宋亚军,李晓娟. 运动中枢疲劳的研究现状[J]. 山西体育科技,1999(8):36-38.

[3] 冯炜权. 运动性疲劳和恢复过程运动能力的研究进展[J]. 北京体育大学学报,1993(7):21-23.

[4] WITTEN I,FRANK E. Data Mining:practical machine learning tools and techniques with Java implementation [M]. San Diego CA: Morgan Kaufmann,2000.

[5] JIAWEIHAN. 数据挖掘概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社,2003:90-140.

(责任编辑:韦廷宗)