文章编号:1001-9081(2006)11-2635-04

用户 ontology的构建及其在个性化检索中的应用

卢林兰,李 明
(兰州理工大学 计算机与通信学院,甘肃 兰州 730050)
(linlanlu@sohu.com)

摘 要:为使用户能够从信息庞杂的网络中方便准确地找到自己所需的信息,在传统检索技术的基础上,结合 ontology及个性化检索技术的研究,提出个性化检索中用户 ontology的概念,讨论了基于行为的用户描述文件的建立,并制定了个性化检索中语义关系的提取规则。最后在此基础上给出了用户 ontology的构建方法,并通过比较实验说明了用户 ontology在个性化检索中的有效性。

关键词:个性化检索;用户描述文件;语义关系;用户 ontology
中图分类号: TP18; TP391 **文献标识码**: A

Construction of user ontology and its application in personalized retrieval

LU Lin-lan, LI Ming
(School of Computer and Communication, Lanzhou University of Technology, Lanzhou Gansu 730050, China)

Abstract: To let users easily and correctly find the correct information they want from the Internet where the information is numerous and jumbled, on the basis of traditional retrieval technique, the concept of user ontology in personalized retrieval was proposed in combination with the research of ontology and personalized retrieval technique. The construction and maintenance of user profile based on behavior was introduced. And the extraction rules of semantic relations were discussed in detail. The method of constructing the user ontology was given based on the discussions. Finally, the availability of using user ontology in personalized retrieval was shown by a comparative experiment.

Key words: personalized retrieval; user profile; semantic relation; user ontology

0 引言

目前,基于关键字匹配的网络检索技术已经远远不能满足用户的需求,用户在用关键字进行检索后,得到的仍然是一堆杂乱的信息,要在这些信息中找到自己感兴趣的信息需要花费大量的时间和精力。

个性化、专业化正逐渐成为网络检索中重要的组成部分。目前,各大搜索网站都提供一定程度的专业化查询服务。如 Yahoo、Google、百度、中搜等,将网络资源按主题进行分类,用户在查询时先选择相应主题,然后再输入关键字进行查询。另外,中搜还在关键字查询之后提供智能导航[1](例如,当我们输入"计算机"进行查询后,会出现以下导航信息:软件下载/公司/教育/学校教务教学系统),检索结果按用户选择的导航信息重新排序。这在一定程度上减少了信息冗余,提高了检索效率。但是,这离我们所希望的个性化检索还有很大的差距。为用户建立用户描述文件(user profile)并利用描述文件对检索结果重排序[2,3]是近年来个性化检索的热门话题。

本文结合 ontology[4~6]和个性化检索研究成果,提出用户 ontology的概念,讨论了用户描述文件的建立,制定了语义关系提取规则,在此基础上给出了用户 ontology的构建过程,并用实验说明其在个性化检索中的有效性。

1 用户描述文件(User Profile)

用户描述文件既是构建用户 ontology的基础,又是过滤检索结果的依据。本文提出的用户描述文件是基于行为的,即对用户 Web日志进行分析,从而了解用户浏览行为。通过对用户访问历史页面分析得知用户偏好并建立相应的用户描述文件。考虑到用户的兴趣可能涉及到几个不同的领域,应对用户检索过程中涉及到的每一个领域都建立一个用户描述文件。以下参照文献[3]中用户描述文件的构建方法,给出本文中用户描述文件的构建及维护方法。

1.1 相关假设和计算

本文中的用户描述文件是基于以下假设的:

假设 1 用户本次从查询页面列表中点击了 L个页面,构成本次点击页面集 T。

假设 2 通常情况下,当用户打开一个页面后,发现并非所需页面,则立即关闭该页面,这段时间不超过 10s。

在用户描述文件的建立和维护过程中,涉及到以下两种计算:

将 HTML的不同标签 P分为 6类,每类给定对应的位置权值 T_p,标签类分别为:标题($T_1=6$)、一级标题和链接锚文字($T_2=5$)、二级标题($T_3=4$)、三级标题($T_4=3$)、正文体(加重字、黑体字、斜体字)($T_5=2$)、正文体(其他)($T_6=1$)。通过公式(1)来计算页面中每个关键术语 t_i的权值 w_i:

$$w_i=\sum_{p=1}^{i} N_p \times C_p \tag{1}$$

其中,N_p表示关键词 t_i在 P部分出现的次数;C_p为 P类标签的归一化权值,$C_p=T_p/\sum_{j=1}^{6} T_j$。

通过对用户浏览行为进行分析,得到用户对页面 p_i的反

收稿日期:2006-05-19;**修订日期**:2006-08-04 **基金项目**:甘肃省自然科学基金资助项目(3ZS042-B25-007)

作者简介:卢林兰(1978-),女,甘肃景泰人,助理工程师,硕士研究生,主要研究方向:ontology及其在信息检索中的应用; 李明(1959-),男,河北辛集人,教授,主要研究方向:智能信息处理。

馈度 feedback(p_i)。

用户浏览页面 p_i的次数为 Fre(p_i),第 j次浏览页面 p_i的时间为 Time(p_i, j),浏览页面 p_i的总浏览时间为 $\sum_{j=1}^{Fre(p_i)} Time(p_i, j)$。T中所有页面的总浏览时间可以看成一个数列,计算此数列的平均值 μ和标准方差 σ,然后用高斯归一化公式处理。根据 3-σ 规则,数列中每个数归入[-1, 1]区间内的概率约为 99%,再通过平移操作使 feedback(p_i)最终落在[0, 1]上。综上所述,页面反馈度计算如下:

$$feedback(p_i) = \frac{\sum_{j=1}^{Fre(p_i)} Time(p_i, j) - \mu + 3\sigma}{6\sigma} \quad (2)$$

1.2 用户描述文件的建立及维护

本文中用户描述文件是通过机器学习的方法来建立和维护的,主要是对关键字的处理。关键字分为机器新学习到的关键字和文件中已经存在的关键字。其中,已存在的关键字又分为经常使用的和不经常使用的两种。以下将以新学到的关键字为主,介绍用户描述文件的建立及维护过程。

对未包含在描述文件中的关键字的处理主要是计算其权值,并将其添加到用户描述文件中,具体操作如下:

1)对于每个被用户点击的页面,筛选出相对新增关键字集 K_{p_i},按公式(1)计算相应关键字权值。形成页面集 T的新增关键字集 $K_{new} = \bigcup_{i=1}^{L} K_{p_i}$。

2)根据公式(2)页面反馈度算法,计算每个页面 $p_i \in T$ 的反馈度 feedback(p_i)。

3)计算 T(含 L个页面)中关键字集 K_{new} 的权值均值向量: $W = \frac{\sum_{i=1}^{L} \{(w_{i1}, w_{i2}, \cdots, w_{in}) \times Feedback(p_i)\}}{L}$。

4)调整用户描述文件 U为: $U = (UK \cup K_{new}, UW \cup W)$,其中,UW为用户描述文件 UK的关键字权值向量。

如果某个关键字经常出现在用户浏览页面中,则认为用户对该关键字有兴趣,应提高其权值。按照公式(1)计算该关键字的权值,将计算所得权值与原关键字的权值相加,结果作为该关键字的新权值存入用户描述文件中。

如果某个关键字长期未被使用(30天之内未被使用),则认为用户对该关键字已经不感兴趣,应该降低其权值。按如下公式调整关键字的权值:

$$w_i = \begin{cases} w_i \times \left(1 - \frac{D_{now} - D_{visited}}{D_{now} - D_{created}}\right), & D_{now} - D_{visited} \geqslant 30 \\ w_i, & \text{否则} \end{cases}$$

其中: D_{now} 表示当前日期; $D_{visited}$ 表示该关键字最后一次被访问的日期; $D_{created}$ 表示该关键字在用户描述文件中被创建的日期。若 $w_i < 1$,则删除关键字 k_i。

2 语义关系提取规则

语义关系是某一句子中表达相关词之间语义的一种关系。构建 ontology的目的就是让计算机理解自然语言语义。因此,语义关系提取是建立 ontology的核心,语义关系提取的准确与否直接决定着 ontology中所含信息的准确程度。

2.1 语义关系及其表示[7]

事物之间的关系是复杂的,但是这种复杂的语义关系可以通过把许多基本的语义关系关联到一起实现。常用的基本事实语义关系主要有:类属关系(ISA, AKO)、属性关系(Have, Can, Is等)、包含关系(Part-of, Composed-Of)、位置关系(Location-at, Location-under等)以及相似关系(Similar-To, Near-to等)。

基本事实语义关系一般用二元谓词或语义网来表示。例如:"张三是一个学生"可以用二元谓词表示为: ISA(张三,学生),也可以用语义网络表示如图 1。

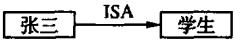

图 1 语义网络表示

复杂的语义关系可以用多元谓词表示,也可以用一组二元谓词来表示,还可以表示为复杂的语义网。例如:"John gives Mary a book"可以用三元谓词表示为: Give(John, Mary, Book)。用二元谓词组表示就是: ISA(G1, Giving-Event), Giver(G1, John), Receptor(G1, Mary), Object(G1, Book)。将以上一组二元谓词分别转换成语义网并互相连接起来就是该三元语义关系的语义网表示形式。由于网络相对复杂,这里不再给出。

2.2 语义关系提取规则的建立

靠手工提取语义关系建立 ontology不仅费时费力,而且由于提取者知识水平的限制和主观因素的影响,还可能造成语义提取不准确,对 ontology表达信息的准确性造成影响。因此,在本文中,我们尝试制定一些语义提取规则,使系统能够根据这些规则自动从网页中提取语义关系。表 1给出了基本事实语义关系提取规则。

在表 1中,S表示句子, NP表示名词短语, SNP表示个体名词, CNP表示类名词, VP表示动词短语, VPbase表示动词原型, PP表示介词断语。规则中所有动词形式都包含其动词的各种变形。

表 1 基本事实语义关系提取规则

序号	语义关系	提取规则
1	ISA	If (S→NP1 be NP2)∧(NP1∈ SNP, NP2∈ CNP) Then ISA(NP1, NP2)
2	AKD	If (S→NP1 be NP2)∧(NP1, NP2∈ CNP) Then AKO(NP1, NP2)
3	Have	If S→NP1 have/has NP2 Then Have(NP1, NP2)
4	Can	If S→NP can VP Then Can(NP, VP)
5	Is	If S→NP be ADJ Then Is(NP, ADJ)
6	Part-Of	If S→NP1 be[a] part/parts of NP2 Then Part-Of(NP1, NP2)
7	Composed-Of	If S→((NP1 be[a] part/parts of NP2) ∨(NP2 be composed of NP1)) Then Composed-Of(NP2, NP1)
8	Belong-To	If S→((NP1 include NP2)∨(NP2 belong to NP1)) Then Belong-To(NP2, NP1)
9	时间	If (Time1-Time2)>0 Then Before(time2, Time1) or After(Time1, Time2)
10	位置	If S→NP1 be/locate PP NP2 Then Location-PP(NP1, NP2)
11	其他	If (S→NP1 VP NP2)∧¬(ISA, AKO, Have, Part-Of, Composed-Of, Belong-To) Then VPbase(NP1, NP2)

规则 1表示的意义为:如果句子 S能够表示为 NP1 be NP2的形式,并且 NP1是个体名词, NP2为类名词,则提取语义关系 ISA(NP1, NP2)。其他规则参照规则 1理解。

3　用户 ontology的构建

用户描述文件只是用户感兴趣的关键字的集合,各关键字之间没有任何内在关系,仅依靠用户描述文件无法指导检索系统找到用户真正感兴趣的信息。本文采用用户 ontology来描述关键字之间的语义关系,以便对用户输入的关键字进行语义扩展,从而使信息的查全率得到提高。

本文中的用户 ontology是根据用户描述文件建立的。我们用语义网络[7]来表示用户 ontology。用户 ontology采用一种基于中心词的半自动 ontology构建方法[8]来构建。

3.1　基础 ontology的建立

建立用户 ontology的第一步是构建中心词网络。在本文中,我们将用户描述文件中权值较高的关键字作为中心词来构建中心词网络。举例说明:假设某用户描述文件涉及经济领域,并且在该描述文件中,关键字“Company”,“Stock”,“Trade”,“Staff”和“Chairman”的权值最高,我们将它们手工连接起来,构成中心词网络,如图 2所示。

图 2并不是真正意义上的 ontology。因为图 2只是简单的关键字网络,并不能表达用户语义,需要找到这些词之间的语义关系。这就需要我们找到建立描述文件时用到的源网页,从中找出这些中心词之间的语义关系(语义关系按照表 1进行规则提取),将提取出的语义关系添加到中心词网络中,就构成了我们需要的基础 ontology。图 3为在中心词网络中添加语义关系之后形成的基础 ontology。

基础用户 ontology是用户 ontology的中心部分,表达了用户感兴趣领域的基本组成部分及它们之间的关系。但是要让计算机对该领域有全面的理解,只靠基础 ontology是远远不够的。因此需要对基础 ontology进行扩展,形成语义关系完整的用户 ontology。

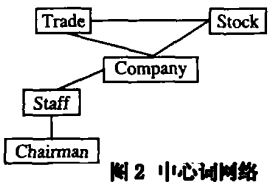

图 2 中心词网络

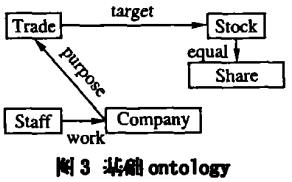

图 3 基础 ontology

3.2　用户 ontology的扩展

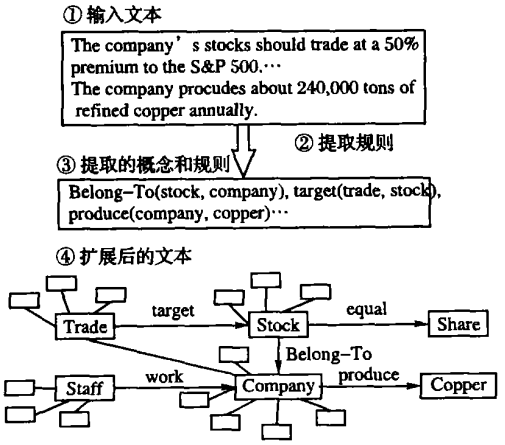

图 4　围绕中心词建立的概念与关系

本文选取 Protégé 3.1[9,10](可从相关网站下载)作为 ontology扩展工具,Protégé采用图形化界面,界面上以多个 tab分别支持 Classes Forms Slots Instances Queries的编辑,利用其 FAQ插件功能完成基础 ontology的扩展[11]。

将手工构建的基础用户 ontology、语义提取规则和经过标注的网页输入 Protégé, Protégé将自动完成用户 ontology的扩展。图 4是中心词“company”扩展后的结果。

对 ontology进行自动扩展的方法不仅可以在较短的时间内建立起语义关系较完整的 ontology,而且可以随着领域知识的变化不断地对 ontology进行动态更新,不断添加新的语义关系。如果在使用过程中发现某些语义表述不准确,只需修改相应规则,而不必作较大改动,就可以得到更加准确的 ontology。

4　试验及结果分析

我们分析了某用户的查询日志和点击日志,发现该用户兴趣主要在计算机领域,按本文方法为其建立了用户描述文件和用户 ontology。

为进行实验,用 Wget[12]从互联网中抓取计算机领域相关网页 5234篇,并建立了相应的索引库和链接库以及网页的倒排索引。

我们设计了一个小型搜索引擎系统,分别用三种检索方法进行实验:

方法一　采用现有天网搜索引擎[13]系统结构,对检索结果不作任何处理(目前搜索引擎采用的方法)。

方法二　在方法一基础上利用用户描述文件对检索结果进行重排序[3]。

方法三　在方法二基础上加入用户 ontology模块,供用户查询时选择,从而对用户查询关键字进行有效语义扩展(本文方法)。

表 2　三种方法查全率比较

序号	检索词	相关网页(篇)	方法一		方法二		方法三	
			返回网页	查全率	返回网页	查全率	返回网页	查全率
1	人工智能	1256	895	0.71	895	0.71	1135	0.90
2	信息检索	2258	1256	0.56	1256	0.56	2115	0.94
3	人工智能 信息检索	2978	1896	0.64	1996	0.64	2755	0.93
4	搜索引擎	1159	684	0.59	684	0.59	956	0.82
5	自动分词	895	523	0.58	523	0.58	746	0.83
6	特征提取	1466	984	0.67	984	0.67	1356	0.92
7	模式匹配	2356	2032	0.86	2032	0.86	2145	0.91
8	CBIR	1358	746	0.55	746	0.55	1196	0.88
9	机器人	779	635	0.82	635	0.82	760	0.98
10	信息管理系统	3526	3125	0.89	3125	0.89	3452	0.98

表 3　三种方法查准率比较

序号	检索词	方法一		方法二		方法三	
		用户满意网页	查准率	用户满意网页	查准率	用户满意网页	查准率
1	人工智能	3	0.20	9	0.60	10	0.67
2	信息检索	6	0.40	11	0.73	10	0.67
3	人工智能 信息检索	7	0.47	13	0.87	14	0.93
4	搜索引擎	10	0.67	12	0.80	13	0.87
5	自动分词	6	0.40	14	0.93	11	0.73
6	特征提取	5	0.33	13	0.87	13	0.87
7	模式匹配	6	0.40	11	0.73	9	0.60
8	CBIR	4	0.27	9	0.60	12	0.80
9	机器人	2	0.13	10	0.67	14	0.93
10	信息管理系统	9	0.60	12	0.80	11	0.73

我们从查准率和查全率两个方面对三种方法作比较。在实验过程中,该用户分别输入 10种不同的检索词进行检索,查全率计算如表 2所示。我们取根据每一种检索词检索得到的前 15个页面来计算查准率,实验数据如表 3。

通过表 2可以看出,利用 ontology对用户查询进行语义扩展后查全率明显提高。由表 3可以看出,利用用户描述文件对检索结果重排序后查准率明显提高。综上所述,本文方法查全率与查准率均明显高于方法一,与方法二比较,查全率有显著提高,但查准率略有下降,这主要是由于分析数据过少造成误差。

5 结语

本文在语义分析和个性化检索研究的基础上提出了利用用户 ontology实现个性化检索的方法,并对用户 ontology的构建过程作了详细说明,最后用实验说明了本文方法的有效性。

虽然本文从语义角度对个性化检索作了有益的探索,但是仍存在很多不足之处,有待于进一步探索研究:1)由于用户行为是随机的,2.1中的假设 2并不总是成立,这就有可能将用户不感兴趣的网页也作为感兴趣的网页,从而使构建的用户描述文件和用户 ontology不够准确,进而造成系统查准率降低,检索结果不符合用户要求。2)文章只是讨论了基本语义关系的提取。对于复杂的语义关系,如何制定相应的提取规则,才能使提取出的语义关系完整而准确未作讨论。3)由于技术水平限制,实验设计不太完善,导致实验结果存在较大误差。

参考文献:

[1] 中国搜索[EB/OL]. http://www.zhongsou.com, 2006.

[2] WU Y-H, CHEN Y-C, CHEN ALP. Enabling personalized recommendation on the web based on user interests and behaviors[A]. KLASW, ed. Proceedings of the 11th International Workshop on Research Issues in Data Engineering[C]. Los Alamitos, CA: IEEE CS Press, 2001. 17-24.

[3] 赵仲孟,袁薇,何世丽,等.个性化搜索引擎中用户模型智能调整算法的研究[J].计算机工程与应用,2005,41(24):184-187.

[4] CHANDRASEKARAN B, JOSEPHSON JR, BENJAMINS VR. What are ontologies and why do we need them?[J]. IEEE Intelligent Systems, 1999, 14(1): 20-26.

[5] KAYED A, COLOMB RM. Extracting ontological concepts for tendering conceptual structures[J]. Data and Knowledge Engineering, 2002, 40(1): 71-89.

[6] STOJANOVIC N. On the query refinement in the onto logy-based searching for information[J]. Information Systems, 2005, 30(7): 543-563.

[7] 王文杰,叶世伟.人工智能原理与应用[M].北京:人民邮电出版社,2004.226-257.

[8] KOO SO, YEON LIM S, LEE SJ. Building an ontology based on hub words for information retrieval[A]. Proceedings of the IEEE/WIC International Conference on Web Intelligence[C]. Halifax, Canada: IEEE Computer Society, 2003. 466-469.

[9] NOY NF, SINTEK M, DECKER S, et al. Creating semantic web contents with protege 2000[J]. IEEE Intelligent Systems, 2001, 16(2): 60-71.

[10] CRUBEZY M, O'CONNOR M, PINCUS Z, et al. Ontology-centered syndromic surveillance for bioterrorism[J]. IEEE Intelligent Systems, 2005, 20(5): 26-35.

[11] SWRL Editor FAQ[EB/OL]. http://protege.stanford.edu/plugins/owl/swrl/, 2006.

[12] GNU Wget[EB/OL]. http://www.gnu.org/software/wget/wget.html, 2006.

[13] 天网 2006[EB/OL]. http://e.pku.edu.cn, 2006.

(上接第 2634页)

期限制情况下优于本文算法,此时所有订单只要在 90个单位时间内交货即可,对于此算例而言相当于没有严格的交货期限制。此时两种算法的完成时间相同,而专家系统在成本方面略优于本文算法。随着交货期限制条件变严,专家系统距离最优解(此类调度问题的最优解很难精确获得,这里指的最优解是实验过程中采用不同算法,通过多次运算获得的相对最优解)的差距越来越明显,而本文算法则表现出较好的寻优和适应能力。多组订单数据的比较结果均与上述情况类似,且订单数量越多,本文算法所表现的优势就越大。

4 结语

大规模定制 CODP之后的供应链调度是整个供应链效率的关键。本文在分析此类供应链调度特点的基础上,建立了符合实际的优化调度数学模型,并设计了用于模型求解的蚁群算法。试验数据证明算法模型不仅调度优化效果良好,同时具有很好的适用性和稳定性。

参考文献:

[1] JIAO J, TSENG MM, DUFFY VG, et al. Product family modeling for mass customization[J]. Computers & Industrial Engineering, 1998, 35(3/4): 495-498.

[2] SUA JCP, CHANGB Y-L, FERGUSON M. Evaluation of postponement structures to accommodate mass customization[J]. Journal of Operations Management, 2005, 23(3/4): 305-318.

[3] GHIASSIA M, SPERAB C. Defining the Internet-based supply chain system for mass customized markets[J]. Computers & Industrial Engineering, 2003, 45(1): 17-41.

[4] GU XJ, QIGN, YANG ZX, et al. Research of the optimization methods for mass customization[J]. Journal of Materials Processing Technology, 2002, 129(1-3): 507-512.

[5] 姚建明,周国华.大规模定制模式下供应链计划调度优化分析[J].管理科学学报,2003,6(5):58-64.

[6] 姚建明,蒲云.基于动态生产能力约束的 MC模式下供应链调度优化[J].系统工程,2005,23(2):25-30.

[7] MOON C, KIM J, HUR S. Integrated process planning and scheduling with minimizing total tardiness in multi-plants supply chain[J]. Computers and Industrial Engineering, 2002, 43(1): 331-349.

[8] PEREA-LOPEZ E, YDSTIE BE, GROSSMANN IE. A model predictive control strategy for supply chain optimization[J]. Computers and Chemical Engineering, 2003, 27(8/9): 1201-1218.

[9] 姜桦,李莉,乔非,等.蚁群算法在生产调度中的应用[J].计算机工程,2005,31(5):76-78,101.

[10] 胡燕海,马登哲,叶飞帆.制造系统通用作业计划与蚁群算法优化[J].计算机集成制造系统,2005,11(1):104-108.

[11] 吴启迪,汪雷.智能蚁群算法及应用[M].上海:上海科技教育出版社,2004.57-93,101-118.

[12] 叶志伟,郑肇葆.蚁群算法中参数 α、β、r设置的研究——以 TSP问题为例[J].武汉大学学报(信息科学版),2004,29(7):597-601.