一种构建个性化网络购物搜索引擎模型研究 *

李世威[a], 钱晓东[b]
(兰州交通大学 a. 交通运输学院; b. 经济管理学院, 兰州 730070)

摘 要: 通过分析在电子商务环境下购物搜索引擎所面临的问题,提出了一种跨网站式的模糊识别多媒体信息购物搜索引擎的模型架构方案,并结合用户个性化的需求进行学习和调整来提高用户的搜索满意度,以提升其购物意愿,进而促进电子商务的发展。运用相关检索指标对该模型进行效能评估,以证明模型的可行性和有效性,并通过分析模型的局限性,提出未来的改进方向。

关键词: 网络购物搜索引擎; 模糊识别; 个性化; 信息检索; 模型架构; 评估
中图分类号: TP302 **文献标志码**: A **文章编号**: 1001-3695(2010)06-2176-05
doi:10.3969/j.issn.1001-3695.2010.06.052

Model of personalized online shopping search engine

LI Shi-wei[a], QIAN Xiao-dong[b]
(*a. School of Traffic & Transportation, b. School of Economic & Management, Lanzhou Jiaotong University, Lanzhou* 730070, *China*)

Abstract: This paper analyzed the problems faced by shopping search engine in e-commerce environment, presented a cross-site type model architecture program of shopping search engine based on fuzzy recognition multimedia information, and combined with the user personality demand to learning and adjusting the model, in order to improve the user's search satisfaction and enhance their shopping wishes, thus contributing to the development of e-commerce. And the study used the retrieval norms to evaluate performance of the model, to demonstrate its feasibility and effectiveness, and proposed improvement direction of the model for future by analyzing its limitations.

Key words: online shopping search engine; fuzzy recognition; personality; information retrieval; model architecture; evaluation

随着电子商务的蓬勃发展,网络上充斥着海量的多媒体信息,用户如何全面有效地查询到所需信息已成为当前信息检索的重要课题之一。本文提出了一种跨网站式的模糊识别多媒体信息购物搜索引擎模型,通过构建相关主题的知识库与主题词汇库,并对 HTML、XML、VRML、ASP、JSP、CGI 等语言描述的相关网页进行特征萃取,建立了相对应主题的网页资料库,然后通过与用户交互式的学习,进行相关特征比对和主题分类分配,模糊识别用户个性化的查询需求,建立相应的初始化隶属度函数,并制定规则来约束搜索的例外情况,以建立能较好地满足用户个性化需求的多媒体信息搜索引擎。

1 个性化信息检索背景

1.1 信息检索

信息检索(information retrieval)从 1950 年开始发展至今,产生了许多查询方式,大致可分为以下三类:

a)布尔逻辑检索(Boolean model)[1]。在面对用户明确的检索需求时,可以处理不同层次的数据或相同层次的多个关键字。布尔逻辑模式可以迅速缩小检索范围,但这种模式最大的问题在于无法判断不同文件对于检索条件的适应度和重要程度。因此,通常会出现检索结果满足检索条件,但与用户实际需求不符的状况。

b)信息过滤检索(information filtering)[1]。这种检索模式是由用户事先向系统提供个人的信息需求,再由系统主动搜集符合需求的相关信息,以定期或不定期的方式呈现给用户。在信息过滤检索模式中,用户的查询需求较为固定,其检索结果也较为符合用户的个性化查询需求。

c)对话反馈检索[1]。这种检索模式结合了人机交互式以及渐进式查询模式,用户通过界面与查询系统进行对话,系统通过语法分析用户的信息需求意图,然后利用解析器进行实际检索,并将结果反馈给用户,当用户再对反馈结果作出相关响应时,系统将根据这些响应信息作出进一步检索,并调整检索策略以符合用户需求。

随着互联网上信息量的指数倍激增,用户对于网络搜索引擎的依赖也越来越深。目前,一般搜索引擎(Google、Yahoo、Alta Vista 等)都会提供三种基本功能:具有利用关键字搜寻所需信息的功能;具有数据更新的功能;当搜索引擎检索到数据后,具有提供相关网页基本信息的功能。

1.2 多媒体信息检索发展趋势

随着网络信息的多元化发展,互联网上充斥着大量的多媒体信息,不再局限于传统的文字信息。如何有效地检索到用户所需的多媒体信息已成为信息检索发展的主要趋势。Web 查询语言就此孕育而生,在这个领域中著名的研究成果有 WISE[2]、WEBSQL[3]等。这些研究是将互联网视为一个海量的数据库,将网页内文件的特征(如关键字、卷标、锚点等)作

收稿日期: 2009-11-16; **修回日期**: 2009-12-20 **基金项目**: 国家社科基金资助项目(08XTQ010)

作者简介:李世威(1981-),男,甘肃白银人,讲师,硕士,主要研究方向为数据挖掘、决策分析(lst9647@126.com);钱晓东(1973-),男,副教授,博士后,主要研究方向为数据挖掘、信息处理.

为网页查询时的基础,以开发出类似 T-SQL 语言来查询网页数据库。这种方法将搜索引擎的适用范围延伸到世界范围的任何领域,因此该模式的检索规则在于所有网页的共同特性而不再局限于个别领域所规定的特殊处理流程。但是互联网上存在着大量的非结构化和半结构化的信息,且不同领域、不同类型的数据格式存在着巨大差异,这就使得在检索过程中很难设计一种标准的查询语言进行检索;此外,如何为缺乏专业知识背景的用户设计出友好的交互系统也是信息检索发展亟待解决的问题。

1.3　个性化搜索引擎

随着互联网的日益普及,用户对于信息的要求也逐渐由大众化需求转换为个性化需求[4],Magedanz 等学者提出了信息检索的研究将朝着个性化方向发展,而智能代理器(intelligent agent)的功能正好能够满足用户的个性化检索要求。Chang 等学者在智能代理器的定义中提到:设计智能代理器的目的是要让系统具备了解用户的工作范围并学习用户处理工作的习惯,达到主动分担用户工作的目的[2]。用户将自己的信息需求提交给智能代理器,代理器代替用户到各相关数据库检索数据,然后将结果集反馈给用户,这种检索模式最大的优势在于:检索无时空限制、系统能够自组织自学习、信息处理过程完全自动化。

目前,绝大多数搜索引擎提供的个性化服务大致可以分为两类[6]:

a)直接定义法。通过直接对用户信息需求的规定来获取用户的个性化需求信息,如要求用户填写相关表格和问卷。直接定义法的优点在于可以收集到用户的基本信息,但缺点在于需要用户的高度配合。

b)隐藏式数据萃取法。由系统特定功能模块担当感应器的角色,学习用户对于事物的反应及处理流程,进而模拟用户隐藏的特征。通常可以根据以下规则来识别用户的隐藏特征:(a)用户对于该网页所停留的时间;(b)用户是否使用滚动条来浏览该网页信息;(c)用户是否点击该网页中的超链接来浏览其他相关信息。

2　模糊逻辑理论

自 1965 年美国控制论专家 Zadeh 提出了模糊集思想起,模糊集合(fuzzy sets)、模糊法则、推论机制与推论模式[7]被广泛地应用于许多领域。1995 年我国学者首次提出公理模糊集理论(axiomatic fuzzy sets,AFS),成为模糊集理论一个新的研究方法[8]。公理模糊集理论应用 AFS 代数和 AFS 结构来描述自然语言语义的不确定性和原始数据随机分布的不确定性,为模糊度隶属函数及其逻辑运算提供了客观统一的确定方法,克服了传统研究方法中隶属函数确定的主观性和模糊逻辑算子选择的随意性[9]。近年来,AFS 理论与概率理论相结合使用,将人类主观的模糊性和客观的不确定性有机地统一起来,成为新的应用方向。

2.1　*AFS* 代数

定义 1[10,11]　设 $X_1,X_2,\cdots,X_n,M$ 是 $n+1$ 个非空集合,则将集合 $EX_1\cdots X_nM$ 定义为

$$EX_1\cdots X_nM=\{\sum_{i\in I}(u_{1i}u_{2i}\cdots u_{ni}A_i)\mid A_i\in 2^M,u_{ri}\in 2^{X_r}\}$$

其中:$r=1,2,\cdots,n$,I 是一个非空指标集,当 $n=0$ 时,上述公式可变换为

$$EM=\{\sum_{i\in I}A_i\mid A_i\in 2^M,i\in I\}$$

其中:I 是一个非空指标集。

定义 2[10,11]　设 $X_1,X_2,\cdots,X_n,M$ 是 $n+1$ 个非空集合,在 $EX_1\cdots X_nM$ 上的一个二元关系 R 定义为

$$\forall\sum_{i\in I}(u_{1i}u_{2i}\cdots u_{ni}A_i),\sum_{j\in J}(v_{1j}v_{2j}\cdots v_{nj}B_j)\in EX_1\cdots X_nM$$

则有　$[\sum_{i\in I}(u_{1i}u_{2i}\cdots u_{ni}A_i)]R[\sum_{j\in J}(v_{1j}v_{2j}\cdots v_{nj}B_j)]\Leftrightarrow$

(1) $\forall(u_{1i}u_{2i}\cdots u_{ni})A_i$,$\exists(v_{1h}v_{2h}\cdots v_{nh})B_h$ 使得 $A_i\supseteq B_h$,$u_{ri}\subseteq v_{rh}$。其中 $i\in I,h\in J,1\leqslant r\leqslant n$。

(2) $\forall(v_{1j},v_{2j},\cdots,v_{nj})B_j$,$\exists(u_{1k}u_{2k}\cdots u_{nk})A_k$ 使得 $B_j\supseteq A_k$,$v_{rj}\subseteq u_{rk}$。其中 $j\in J,k\in I,1\leqslant r\leqslant n$。

$EX_1\cdots X_nM/R$ 记为

$$EX_1\cdots X_nM\cdot\sum_{i\in I}(u_{1i}\cdots u_{ni}A_i)=\sum_{j\in J}(v_{1j}\cdots v_{nj}B_j)$$

该等式表示$\sum_{i\in I}(u_{1i}u_{2i}\cdots u_{ni}A_i)$和$\sum_{j\in J}(v_{1j}v_{2j}\cdots v_{nj}B_j)$有关系 R 下等价。

定理 1[10,11]　设 $X_1,X_2,\cdots,X_n$,M 是 $n+1$ 个非空集合,$(EX_1\cdots X_nM,\wedge,\vee)$在如下定义的 $\wedge$,$\vee$ 二元运算下,成为一个完全分配格:

$\forall\sum_{i\in I}(u_{1i}u_{2i}\cdots u_{ni}A_i),\sum_{j\in J}(v_{1j}v_{2j}\cdots v_{nj}B_j)\in EX_1\cdots X_nM$ 有 $\sum_{i\in I}(u_{1i}u_{2i}\cdots u_{ni}A_i)\wedge\sum_{j\in J}(v_{1j}v_{2j}\cdots v_{nj}B_j)=\sum_{i\in I,j\in J}[(u_{1i}\cap v_{1j})(u_{2i}\cap v_{2j})\cdots(u_{ni}\cap v_{nj})(A_i\cup B_j)]$和$\sum_{i\in I}(u_{1i}u_{2i}\cdots u_{mi}A_i)\vee\sum_{j\in J}(v_{1j}v_{2j}\cdots v_{nj}B_j)=\sum_{k\in I\cup J}(\omega_{1k}\omega_{2k}\cdots\omega_{nk}C_k)$其中:$\forall k\in(I\cup J)$,$(I\cup J)$是 I 与 J 的不交并,如果 $k\in I$ 则$C_k=A_k$,$\omega_{rk}=u_{rk}$;如果 $k\in J$ 则 $C_k=B_k$,$\omega_{rk}=v_{rk}$,$1\leqslant r\leqslant n$。此时,$(EX_1\cdots X_nM,\wedge,\vee)$被称为 $X_1,X_2,\cdots,X_n$ 和 M 上的 EI^{n+1}代数,$X_1\cdots X_n\Phi$ 是 $EX_1\cdots X_nM$ 的最大元,$\Phi\cdots\Phi M$ 是 $EX_1\cdots X_nM$ 的最小元,当 $n=0$ 时,上述 EI^{n+1}代数就成为 EI 代数$(EM,\wedge,\vee)$。

此方法可以将少数的几个模糊概念生成用 EM 表示的非常多的概念,$\wedge$和$\vee$是这些模糊概念的交、并运算,且 EM 中每个元素都有确切的语意[11]。因此,本文将 EI^{n+1}代数引入搜索引擎的模式识别,可以从用户所提交的较少信息中生成新的概念模式,从而进一步充分识别用户的个性化需求。

2.2　*AFS* 模糊逻辑系统

定义 3[11,12]　设 ξ 是论域 X 上的一个属性或概念,ξ 与 X 上的一个二元关系 R_ξ(即 $R_\xi\subseteq X\times X$)相对应。其中$(x,y)\in R_\xi$,说明 x 以某种程度属于 ξ 且 x 属于 ξ 的程度要强于或等于 y 属于 ξ 的程度。

定义 4[11,12]　设 X 为集合,R 是集合 X 上的二元关系。如果对于 $x,y\in X$ 且 $x\neq y$,若 R 满足:

a)如果$(x,y)\in R$,则$(x,x)\in R$;

b)如果$(x,x)\in R$,$(y,y)\notin R$,则$(x,y)\in R$;

c)如果$(x,y)\in R$,$(y,z)\in R$,则$(x,z)\in R$;

d)如果$(x,x)\in R$,$(y,y)\in R$,则或$(x,y)\in R$,或$(y,x)\in R$。

则称 R 为弱偏好关系(sub-preference relation),与弱偏好关系对应的概念称为简单概念,反之称为复杂概念。

定义 5[10~12]　设 X、M 为两个集合,2^M 是 M 的幂集,$\tau=X\times X\to 2^M$,如果对于任意的 $x_1,x_2,x_3\in X$,τ 满足下面的公理:

$AX_1:\tau(x_1,x_2)\subseteq\tau(x_1,x_1)$;

$AX_2:\tau(x_1,x_2)\cap\tau(x_2,x_3)\subseteq\tau(x_1,x_3)$,

则将(M,τ,X)称为一个 AFS 结构,X 称为论域,M 称为属性域,τ 称为结构。在实际应用中,通常构造的 M 是论域 X 上的

简单属性组合,所以定义如下公式来构造 AFS 结构[11]:

$$\tau(x,y)=\{m\mid m\in M,(x,y)\in R_m\}$$

定义 6[11~14]　对逻辑非(′)运算作如下定义:

$$\forall \sum_{i\in J}A_i\in EM,\text{有}(\sum_{i\in I}A_i)'=\wedge_{i\in I}\{\vee_{a\in A_i}(a')\}$$

其中:$a\in M$,a'是简单概念 a 的非。

根据以上定义,将代数系统($EX_1\cdots X_nM$, $\wedge$, $\vee$, ′)称为 AFS 模糊逻辑系统。在构建搜索引擎用户个性化学习模块中,采用 AFS 模糊逻辑系统可以从用户较少的自然语言描述中获得最大的信息需求特征,进而最大限度地匹配用户所需要的网页资料。

2.3　AFS 模糊集隶属函数

定理 2[11,12]　设 X、M 为两个集合,(M,τ,X)是一个 AFS 结构,$B\subseteq X$,$A\subseteq M$,定义符号:

$$\underline{A}(B)=\{y\mid y\in X,\tau(x,y)\supseteq A,\forall x\in B\}$$

对于给定的 $x\in X$,如果存在映射关系 $\phi_x:EM\to EXM$, $\forall \sum_{i\in I}A_i\in EM$,使得:

$\phi_x(\sum_{i\in I}A_i)=\sum_{i\in I}\underline{A}_i(\{x\})A_i\in EXM$ 成立,则称 ϕ_x 是从(EM, $\wedge$, $\vee$)到(EXM, $\wedge$, $\vee$)上的代数同态。

定义 7[11,12]　设 X 是一个集合,S 是 X 上的 σ 代数,存在 $\rho:X\to R^+=[0,\infty)$,$0<\sum_{x\in X}\rho(x)<\infty$,对任意的 $A\in S$,使得:

$m(A)=\dfrac{\sum_{x\in A}\rho(x)}{\sum_{x\in X}\rho(x)}$成立,则称 m 为 S 上的由 ρ 导出的测度。

由于搜索引擎在模糊识别用户所提交的自然语言过程中,绝大多数情况下属性是离散的,本文只讨论离散情况下的模糊测度的导出。文献[12]对连续情况下的模糊测度导出进行了详细地阐述。

定义 8[8,11,12]　设 ξ 是 X 上的一个简单概念,且存在 ρ_ξ: $X\to R^+=[0,\infty)$,如果 ρ_ξ 满足下列条件:

a)$\rho_\xi(x)=0\Leftrightarrow(x,x)\notin R_\xi,x\in X$;

b)$(x,y)\in R_\xi\Rightarrow\rho_\xi(x)\geqslant\rho_\xi(y),(x,y)\in X$,

则将 ρ_ξ 称为简单概念 ξ 的隶属度函数。

定义 9[8,11,12]　设 X 为论域,M 是 X 上的一些简单概念构成的一个集合,S 是 X 上的 σ 代数,对于 $\forall\alpha\in M$,m_α 是由 α 的隶属度函数 ρ_α 导出的 S 上的测度,对于 $\sum_{i\in I}a_iA_i\in EXM$,如果满足 $a_i\in S$, $\forall i\in I$,则定义 $\sum_{i\in I}a_iA_i$ 的范数为

$$M(\sum_{i\in I}a_iA_i)=\sup_{i\in I}(\prod_{\alpha\in A_i}m_\alpha(a_i))\in[0,1]$$

对于在半认知空间(M,τ,X,S)中,可测的模糊概念 $\sum_{i\in I}A_i\in EM$,定义其表示模糊概念的 Zadeh 模糊集隶属函数为 $\forall x\in X$。

$$\mu_{\sum_{i\in I}A_i}(x)=M((\sum_{i\in I}A_i)(x))=M(\sum_{i\in I}\underline{A}_i(x)A_i)\in[0,1]$$

3　网络购物搜索引擎模型构建

3.1　模型框架

用户在商品信息搜索过程中,存在很大的不确定性,而且以往简单的文本信息不再满足用户的需求。如何从用户简单的自然语言中识别用户的需求,进而将丰富的多媒体信息呈现给用户,成为当今购物搜索引擎发展的主要趋势。

因此基于上述理论,本文构建了一种能够较好地满足用户个性化查询需求的网络购物搜索引擎模型,该模型框架如图 1 所示。

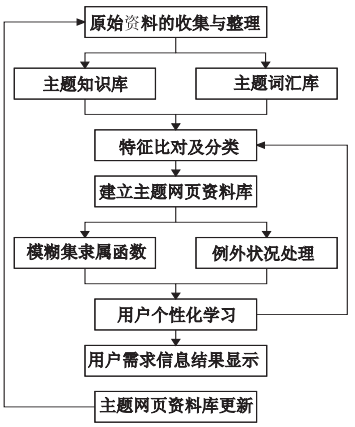

图1　一种个性化网络购物搜索引擎模型

1)原始网页资料的收集与整理　由于互联网上的信息量是极其庞大的,如果用户搜集相关网页信息时,系统实时初始化原始网页资料库,则会导致检索服务器的传输带宽严重不足,致使检索系统无法顺畅运行。为了避免这种情况,借鉴大多数搜索引擎的运作方式,采用离线运作模式进行原始网页资料的收集与整理。根据不同主题,先通过全文搜索引擎搜集相关主题的网页资料,如果收集到该主题信息的 URL 时,通过网络下载软件 LAN Spider(或 Teleport Pro、LAN Search Pro),进行该主题资料的收集,统一集中到本系统的服务器上,并将收集到的资料按相关主题进行分类汇总。

2)建立主题知识库和主题词汇库　为了使系统具有对特定领域内的信息进行处理和相关推理的能力,就需要根据相关主题领域的特征和规则构建主题知识库。本文针对数码产品领域内的特征、行业报告,以及根据主要的生产商和零售商对相关产品的使用说明和促销宣传文件进行分词处理,建立了以数码产品型号为单位的主题知识库。由于数码产品种类繁多,本研究只针对行业内核心企业的产品,构建相关主要产品的主题知识库。

当用户输入自然语言进行检索时,检索系统需要分析出其中对于查询结果具有影响的词与字,为了实现这种语言解析的功能,进而辅助模糊集隶属函数的构建,就必须建立相关领域内的完整主题词汇库。建立主题词汇方法很多,本文依照主题知识库构建基本的主题词汇库,然后采用问卷调查和用户访谈的形式来扩充数码产品的词汇库。当系统运行后,根据与用户的交互式学习、例外状况的判断以及相关主题网页资料的更新来进一步丰富该主题的词汇库。主题知识库与主题词汇库构建过程如图 2 所示。

3)特征比对及分类　因为收集来的网页资料含有大量的信息,且是不同的 Web 语言(如 HTML、XML、VRML、ASP、JSP、CGI 等)编译而成,因此就需要根据所建立的主题知识库和主题词汇库对每个网页资料进行特征识别,并将其核心信息分类存放在服务器的数据库里。例如,对于 HTML 形式的网页,可以通过〈TITLE〉来识别网页标题、IMG SRC = 来识别图片名称、ALT = 来识别相关图片说明、〈P〉〈BR〉〈TR〉〈TD〉来识别网页内容等等。通过特征比对与分类,可以建立相关主题词汇与网页资料的关联,进而提高系统的检索效率。

4)建立主题网页资料库　通过特征比对与分类模块,将相关主题的网页按照与主题知识和主题词汇相关联的规则,存放在本系统的数据库服务器上,形成相关主题的网页资料库。

5)建立用户的模糊集隶属函数　设 X 为相关主题的网页

资料集合,M 是 X 上的简单属性集合(即相关主题词汇集合),τ 是用户相关的检索结构,A 是用户相关的检索要求,(M,τ,X)是一个 AFS 结构,因此可以根据上述的 AFS 理论建立用户相关的模糊集隶属函数[8,11,12]:

$$\mu_{\sum_{i\in I}A_i}(x)=M((\sum_{i\in I}A_i)(x))=M(\sum_{i\in I}\underline{A}_i(x)A_i)\in[0,1]$$

其中:1 表示该网页真正符合用户的检索要求;0 表示该网页不符合用户的检索要求。

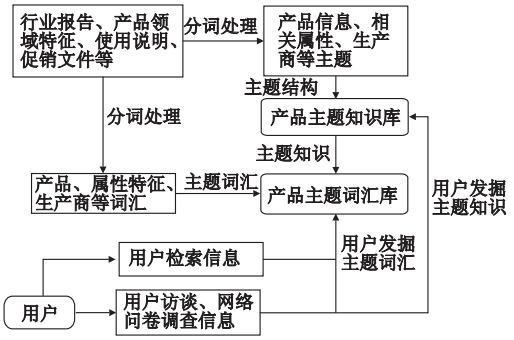

图2　主题知识库与主题词汇库构建流程

6)例外状况处理　若主题词汇库从用户的自然语言中无法解析任何关键字,那么系统将存储该语句,并调用该主题的通用查询语句来显示查询信息。例如,数码相机 + 佳能 + 索尼 + 奥林巴斯 + … + 最新 + …。用户可以修改这个查询语句得到最终查询结果,然后系统将所保存的未解析的语句,与修改后的查询语句和查询结果关联,进而扩充该主题的词汇库。

7)用户个性化学习　不同用户在搜索商品时,所提出的需求因个人偏好会有很大的差异,但不同用户又有自身的消费定式,如何识别用户的消费定式,就成为个性化购物搜索引擎构建的关键。在此,模型采用隐藏式数据萃取法[6],通过制定规则、监测用户的网上行为来识别用户潜在的消费偏好,并结合所定义的模糊集隶属函数,让用户对于检索结果进行判定,进而修正该用户的隶属函数,以达到对其消费定式较好的识别。

8)用户需求信息结果显示　通过样本集的训练,如果达到用户检索的满意度,那么系统基本上识别了该用户的消费定式,并确定了相关的模糊集隶属函数,从而系统就可以为用户提供个性化的检索功能,并按照用户的偏好显示检索结果。

9)主题网页资料库更新　由于本系统采用的是离线运作方式,而且互联网的信息是不断动态变化的,为了有效更新主题资料库,而又不影响用户检索效率,系统采用分布式服务器的方式,单独构建一个更新主题网页资料的服务器,运用网络下载软件 LAN Spider(或 Teleport Pro、LAN Search Pro)进行相关主题网页收集,然后在用户访问量较少的时间段更新原始主题资料库,通过主体知识库和主题词汇库对新的网页资料进行特征比对和分类,最后根据用户的模糊集隶属函数与用户检索偏好建立关联。

3.2　模型效能评估

为了对该购物搜索引擎模型的效能进行评估,作如下定义来计算系统信息检索的正确率、错误率和漏检率。

设 Attain_$H(L)$为该购物搜索引擎检索到的与用户需求相关度高(低)的产品信息数;Correct_$H(L)$为该搜索引擎检索到的与用户需求相关度高(低)且符合用户需要的产品信息数;Error_$H(L)$为该搜索引擎检索到的与用户需求相关度高(低)且不符合用户需要的产品信息数,因此有 Error_$H(L)$ = Attain_$H(L)$ − Correct_$H(L)$;A_Page_$H(L)$为该购物搜索引擎检索到的与用户需求相关度高(低)的网页数;C_Page_$H(L)$为该购物搜索引擎检索到的与用户需求相关度高(低)且符合用户需要的网页数;E_Page_$H(L)$为购物该搜索引擎检索到的与用户需求相关度高(低)且不符合用户需要的网页数,因此有 E_Page_$H(L)$ = A_Page_$H(L)$ − C_Page_$H(L)$;Total_Page 为用户从原始资料库中手工检索到相关产品的网页总数;Miss_Page 为购物该搜索引擎漏检到的网页数,有 Miss_Page = Total_Page − C_Page_$H(L)$;$C_{H(L)}$为检索信息相关度高(低)的正确率;$E_{H(L)}$为检索信息相关度高(低)的错误率;$CP_{H(L)}$为检索网页相关度高(低)的正确率;$EP_{H(L)}$为检索网页相关度高(低)的错误率;$MP_{H(L)}$为检索网页相关度高(低)的漏检率。则定义如下式:

$$C_{H(L)}=\frac{\text{Correct\_}H(L)}{\text{Attain\_}H(L)} \tag{1}$$

$$E_{H(L)}=\frac{\text{Error\_}H(L)}{\text{Attain\_}H(L)} \tag{2}$$

$$CP_{H(L)}=\frac{C\text{\_Page\_}H(L)}{A\text{\_Page\_}H(L)} \tag{3}$$

$$EP_{H(L)}=\frac{E\text{\_Page\_}H(L)}{A\text{\_Page\_}H(L)} \tag{4}$$

$$MP_{H(L)}=\frac{\text{Miss\_Page}}{\text{Total\_Page}} \tag{5}$$

为了检验系统检索的准确性,随机抽取若干个用户样本进行系统学习,形成样本各自偏好的模糊集隶属函数,然后用户通过系统检索自身偏好的商品,并通过人工判断来识别检索结果的正确性,进而来分析该检索系统的效能。表 1 和 2 是经过系统反复学习后生成的所有样本平均效能指标值。从表 1 和 2 中可以看出,在高相关的检索条件下,系统检索网页的正确率为 0.986,在低相关的检索条件下,系统检索网页的正确率为 0.904,说明系统能够很好地满足用户的检索要求。

表 1　相关产品信息数检索效能表

C_H	E_H	C_L	E_L
0.829	0.171	0.794	0.206

表 2　相关产品网页数检索效能表

CP_H	EP_H	MP_H	CP_L	EP_L	MP_L
0.986	0.014	0.295	0.904	0.096	0.059

但是,在高相关的检索条件下,系统的漏检率 MP_H 同样也是比较高(0.295),进一步分析研究发现,在进行相关主题网页特征比对和分类时,对自然语言的解析过于片断化,这样做的目的是过滤不符合要求的错误网页(容错率),但同时也导致了漏检网页数量的增加。如何有效地过滤掉不符合主题词汇的网页信息并降低网页的漏检率,是该模型改进的方向。

4　结束语

根据上述分析,针对不同用户的个性化需求,建立了一种基于 AFS 模糊识别方法的购物搜索引擎模型,能够通过与用户的交互式学习发现其消费定式,生成符合用户需求偏好的模糊集隶属函数,能够很好地为用户检索提供个性化需求信息。但是,模型也存在一些不足,如何在容错率和漏检率之间寻求一个平衡并结合用户的需求偏好建立一个多目标函数是该模型尝试改进的方向。

参考文献:

[1] BELKIN N J, CROFT W B. Information filtering and information re-

trieval: two sides of the same coin[J]. Communication of the ACM, 1992, 35(6): 29-38.
[2] MOHAGEG M F, GRAPHICS S. The influence of hype rtext linking structures on the efficiency of information retrieval[J]. Human Factors, 1992, 34(3): 351-367.
[3] DOUG R. An architecture of integrated agents[J]. Communication of the ACM, 1994, 37(7): 106-116.
[4] BUDI Y, LEE D L. A world wide web resource database system[J]. IEEE Trans on Knowledge and Data Engineering, 1996, 8(4): 548-554.
[5] ZUMBACH J. Enhancing learning from hypertext by inducing a goal orientation: comparing different approaches[J]. Instructional Science, 2002, 30(4): 243-267.
[6] CALISIR F, GUREL Z. Influence of text structure and prior knowledge of the learner on reading comprehension, browsing and perceived control[J]. Computer in Human Behavior, 2003, 19(2): 135-145.
[7] ZADEH L A. Fuzzy sets[J]. Information and Control, 1965(8): 338-353.
[8] LIU Xiao-dong. A new fuzzy model of pattern recognition and hitch diagnoses of complex systems[J]. Fuzzy Sets and Systems, 1999, 104: 289-297.
[9] LIU Xiao-dong, WANG Wei, CHAI T Y. The fuzzy clustering analysis based on AFS theory[J]. IEEE Trans on Systems, Man and Cybernetics Part B, 2005, 35(5): 1013-1027.
[10] LIU Xiao-dong. The fuzzy theory based on AFS algebras and AFS structure[J]. Journal of Mathematical Analysis and Applications, 1998, 217: 459-478.
[11] REN Yan, SONG M L, LIU X D. New approaches to the fuzzy clustering via AFS theory[J]. Internal Journal of Information and Systems Sciences, 2007, 3(2): 307-325.
[12] LIU Xiao-dong. The fuzzy sets and systems based on AFS structure[J]. Fuzzy Sets and Systems, 1998,95(2): 179-188.
[13] LIU Xiao-dong, CHAI Tian-you, WANG Wei. Approaches to the representations and logic operations for fuzzy concepts in the frame-work of axiomatic fuzzy set theory II[J]. Information Sciences: an International Journal, 2007, 177(4): 1027-1045.
[14] LIU Xiao-dong, CHAI Tian-you, WANG Wei. AFS fuzzy logic system and its applications to model and control[J]. International Journal of Information And Systems Sciences, 2006, 2(3): 1-21.
[15] LIU Xiao-dong. The development of AFS theory under probability theory[J]. International Journal of Information And Systems Sciences, 2007, 3(2): 326-348.
[16] FERNANDEZ E, LEYVA J C. A method based on multi-objective optimization for deriving a ranking from a fuzzy preference relation[J]. European Journal of Operational Research, 2004, 154(1): 110-124.
[17] PENEVA V, POPCHEV I. Properties of the aggregation operators related with fuzzy relations[J]. Fuzzy Sets and Systems, 2003, 139(3): 615-633.
[18] HERRERA F, HERRERA-VIEDMA E, CHICLANA F. Militiperson decision-making based on multiplicative preference relations[J]. European Journal of Operational Research, 2001, 129(2): 372-385.
[19] HERRERA-VIEDMA E. Modeling the retrieval process of an information retrieval system using an ordinal fuzzy linguistic approach[J]. Journal of the American Society for Information Science and Technology, 2001, 52(6): 460-475.
[20] HERRERA-VIEDMA E, HERRERA F, CHICLANA F, *et al*. Some issues on consistency of fuzzy preference relations[J]. European Journal of Operational Research, 2004, 154(1): 98-109.
[21] GAO J Q. A personalized WWW image-text shopping engine: a case study on cellular phones[D]. Taiwan: Da-Yeh University, 2000.
[22] PASI G. Modeling users' preferences in systems for information access[J]. International Journal of Intelligent Systems, 2003, 18(7): 793-808.
[23] Van De WALLE B. A relational analysis of decision makers' preference[J]. International Journal of Intelligent Systems, 2003, 187: 775-791.

(上接第 2172 页)

由以上测试和分析结果表明,采用扩展的 WKT 格式的矢量地理数据表示模型是可行的,能够很好地利用 P2P 网络的共享机制,减少数据传输量,实现数据的快速分片与合并。

5 结束语

P2P 网络为各领域的研究带来了广阔的前景, P2PGIS 更是 P2P 网络中一个很好的例子。在此基础上,本文着重研究了矢量地图的分片技术,并分析 P2P 网络的特性,结合用户对数据请求的快速响应需求,最终构造了一个基于分布式拓扑的 P2P 矢量地理数据的表示模型。这种表示模型能够根据请求数据的大小,动态地调整网络中的数据量,尽可能地减少网络负载,提高传输效率,保证数据合并的正确性。该模型比较简单,可以进一步在此基础上改进分片和合并算法,提高原型系统的性能。

参考文献:

[1] 刘兴权,严米. 基于 J2EE 和 XML 的分布式 GIS 研究[J]. 地理空间信息,2007,5(3):12-14.
[2] 刘德刚,向金海,周刚. 基于 P2P 的 Web GIS 系统架构设计[J]. 计算机与现代化,2007(11):129-131.
[3] MONDAL A, YI Li-fu, KITSUREGAWA M. P2PR-tree: an R-tree based spatial index for peer-to-peer environments[C]//Proc of International Workshop on Peer-to-Peer Computing and Databases, 2004. Greece: Heraklion, 2004:516 -525.
[4] TANIN E, HARWOOD A, SAMET H. Using a distributed quad tree index in peer-to-peer networks[J]. The VLDB Journal, 2007, 16(2): 165- 178.
[5] BERGAMINI J A, Dr HAUNGS M. Enabling P2P cooperative WMS proxy caching and prefetching in an educational environment[C]// Lecture Notes in Geoinformation and Cartography. Berlin: Springer, 2007:1-14.
[6] 高波,郭朝珍,丁善镜. 基于 GML 矢量图层分割的空间数据分布式协同处理[J]. 计算机应用,2009,29(1):297-300,303.
[7] 骆炎民,涂超. 基于 GML 的 WebGIS 地理信息建模[J]. 计算机工程与应用,2004,40(15):218-221.
[8] 杨建宗,杨崇俊,明冬萍,等. WebGIS 系统中矢量数据的压缩与化简方法综述[J]. 计算机工程与应用,2004,40(32):36-38,92.
[9] VIVID Solutions. JTS Topology Suite Technical Specifica- tions v1.4 [EB/OL]. (2003-10-17) [2009-11-8]. http://www. vividsolutions. com/jts/jtshome. htm.