基于感兴趣度的 WWW 个性化信息发现*)

WWW Personalized Information Retrieval Based on Measures of Interestingness

卢 超 梅卫峰 陈俊良 徐永森

(南京大学网络信息中心 软件新技术国家重点实验室 南京210093)

Abstract Popular use of the Internet as a global information system has flooded us with a tremendous amount of data and information. The traditional methods of information retrieval are unable to meet users' expectation . To solve this problem, the personalized IR has made good progress. In this paper, we apply the model of interestingness to guide the IR on Web, and support the automatic studying of users' search habit by means of relevant feedback. Finally, we present the implementation of a completed PSA (Personalized Searching Assistant)system.

Keywords Information retrieval, Measures of interestingness, Relevant feedback, Personalize

1 引言

随着 Internet 的飞速发展,网络信息量爆炸式的增长,信息的更新速率也成倍加快。再加上 WWW 本身的分布性和动态性,使得发现特定的信息变得越来越困难。传统的 WWW 信息发现方法[1],使用称之为 spiders 或 robots 的自动网站信息发现程序,在 Internet 上进行漫游,将它们所发现的 Web 文档下载到本地。在本地机上,通过离线数据库的建立,对搜集到的文档进行特征抽取、聚类、索引。用户的搜索请求,由信息发现系统转化为对数据库的查询,而得到最后的搜索结果。当今一个著名的 robot 的实现系统就是由科罗拉多大学开发的 WWW Worm 系统[2]。但是,这种基于离线数据库的信息发现策略有它先天的缺陷:首先,网络信息量的急剧增加,使得数据库的更新越来越困难,更新所耗费的时间越来越长,从而导致数据库中大量无效文档的出现,使得系统查询精度大大下降。其次,数据库在对下载到本地的文档进行特征抽取的过程中,不可避免地会将某些关于文档结构和内容的重要信息抽取掉,从而使得数据库中的文档特征不能精确地表示其对应的文档信息。最后,由于该策略中信息搜集和查询过程相互独立,这样使得查询过程中的反馈信息不能有效地用来改进信息收集的效率,更不能通过对用户反馈信息的学习,实现个性化的搜索。

为了解决传统搜索策略的这些弱点,基于客户端的主动式个性化信息发现技术得到了蓬勃的发展。它的主要思想是:模拟用户进行网络浏览的过程,自动地在 Web 页面中进行导航,在线地进行信息发现[3]。相比传统搜索策略而言,个性化信息发现不需要专门的数据库系统支撑,提供了对瞬时信息的发现能力,每次处理的对象是一个完整的 Web 文档,保证了信息的完整性,最主要的是:提供了学习用户反馈信息的能力,可以非常好地支持用户对信息发现系统的个性化定制。

目前,比较成熟的个性化信息发现系统有由艾恩德霍芬大学开发的绑定于 Mosaic 的 Search-Tool[4],它采用了一种称为 fish-search 的信息发现算法,该算法的主要思想是对以起始节点为根的 Web 结构树进行深度优先的遍历,在遍历过程中使用关键字的匹配算法获得相关页面。

显然,缺乏启发性的遍历算法,不可能实现对网络信息空间的自适应,达不到搜索效率的最优化。同时,由于在 Search-Tool 中缺乏对用户反馈信息量化的有力模型,因而,对用户搜索习惯学习的能力也不强。

为了解决个性化信息发现系统中的以上问题,本文通过引入当前数据挖掘技术中的感兴趣度模型,提出了基于感兴趣度导引的 WWW 信息发现算法,提高了系统的效率,并且给出相应的相关反馈度量公式,以优化系统的学习能力。在文章的最后,简要地介绍了一个采用以上思想建立的 WWW 个性化搜索系统——PSA。

2 基于感兴趣度的 WWW 个性化信息发现

2.1 感兴趣度模型简介

在知识发现过程中的感兴趣度分为客观感兴趣度和主观感兴趣度。客观感兴趣度主要根据模式或规则的形式和数据集中的数据进行定义,属于数据驱动;而主观感兴趣度还要考虑用户的参与等人为因素的影响,属于用户驱动。单纯地考虑客观感兴趣度是不够的,它不能考虑模式和规则的所有方面,同时感兴趣度的问题从本质上来说是一个主观的问题,不同的用户感兴趣的规则和模式是不一样的,因而需要用户的参与。所以需要综合地使用客观感兴趣度和主观感兴趣度这两种标准[5]。

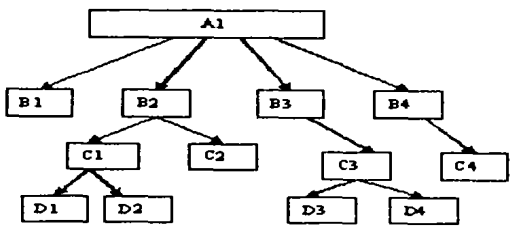

图1 Web 页面链接树

作为主动式的网上信息发现技术,个性化搜索引擎完全模仿用户平时上网浏览时的信息发现模式。即用户以某一特定的 Web 页面为信息发现的起点,以该页面上的超链接(hyper link)为导引,进行下一层页面的浏览,如此反复,直到找到感兴趣的内容。在该过程中,特定的用户查询和搜索页面的匹配度就表现了客观的感兴趣度,而用户对查询结果的反馈信息则体现了主观感兴趣度。通过客观感兴趣度的启发,我们可以方便地实现对搜索空间的智能剪枝,提高系统的搜索效率。而在得到用户主观感兴趣度的基础上,我们就能学习用户

*)本课题为江苏省应用基础计划资助项目。卢 超 硕士研究生,研究方向为 Internet 信息发现。梅卫锋 硕士研究生,研究方向为信息发现,形式方法。陈俊良 教授,研究方向为网络技术、网上信息发现。徐永森 教授,研究方向为软件理论。

的搜索习惯，从而实现搜索的个性化。

2.2 基于客观感兴趣度启发的 Web 信息发现算法

主动式的个性化信息发现系统以图1所示的 Web 页面链接树为基础进行信息发现。在图1中，每一个节点代表一个 Web 页面，由节点 i 指向节点 j 的有向边 D_{ij} 表示在页面 i 中链接为 j 页面的超链接，搜索以根节点为起点，在超链接的导引下不断进行到下一层的页面搜索。

目前，对 Web 页面链接树 T 的搜索算法主要有深度优先算法和广度优先算法，但这两种算法都有致命的弱点。

对一个网上信息发现系统而言，准确性和效率是两个至关重要的要求。由于信息发现的过程中需要将相关页面下载到本地上处理，故而盲目的页面采集会极大地浪费网络带宽，加重网络传输负担，降低系统效率。同时也由于大量不相关页面的出现，系统的准确率也会大大降低。

而正是由于缺乏相关的启发策略，深度优先和广度优先算法只能盲目对信息空间进行遍历，因此效果并不好。不过，在信息发现过程中，这两种算法也有各自的特点，深度优先算法可以对特定主题进行深度搜索，从而可得到较高的精度。而广度优先算法在主题不明确的情况下，可以对指定层数的所有页面进行穷举搜索，从而得到较高的查全率。

在以上研究的基础上，我们提出了基于客观兴趣度导引，综合深度优先和广度优先的二阶段搜索策略：

首先我们给出以下概念：兴趣阈值 IT，兴趣基准 IB，客观感兴趣度函数 I，匹配度函数 M，遍历层数 m。

在介绍最终的算法之前，我们先给出匹配度函数 M 的定义：为了提高信息检索的精度，我们采用基于矢量空间模型(VSM)的匹配度算法[6]。在 VSM 中，文档 D 和查询 Q 均被表示为 t 维空间中的数字矢量，即 $Q=(q_1,q_2,\cdots,q_t)$，$D=(d_1,d_2,\cdots,d_t)$，其中 q_i 和 d_i 对应于关键字 i 分别在 Q，D 中的权值大小。则 Q，D 的匹配度为：

$$M(Q,D)=\frac{\sum_{i=1}^{t}q_i d_i}{\left(\sum_{i=1}^{t}(q_i)^2\cdot\sum_{i=1}^{t}(d_i)^2\right)^{\frac{1}{2}}}$$

相应的信息发现算法描述为：

输入：起始搜索节点 $A1$，用户查询关键字 Q

①预先给定一组基于经验值的 IT，IB，m。

②计算起始节点 $A1$ 处的匹配度 $M(A1)$(注：本算法中的匹配度函数由上文介绍的 VSM 算法给出)，并使 $A1$ 处的兴趣度 $I(A_1)=M(A_1,Q)$，作为初始的客观感兴趣度。

以下，首先进行第一阶段即基于广度搜索的兴奋点发现过程。

③使用广度搜索算法遍历 $A1$ 节点下 m 层内的所有节点，并且计算全部第 m 层节点 $node_i$ 的兴趣度

$I(node_i)=M(node_i,Q)$，**if**

$I(node_i)\geq IT$ **then** {$node_i$ 标记为兴奋点} **else** {在 $node_i$ 处结束向下遍历}。

然后，进行第二阶段的基于深度优先算法的专题性搜索。

④以第3步得到的兴奋点集合中的元素为第二阶段搜索的起点，进行基于兴趣度启发的深度优先搜索阶段：

以图1为例，不妨设 B_2 为一兴奋点，则按深度优先算法，首先游历 C_1 节点，给出该点的兴趣度为

$I(C_1)=I(B_2)+M(C_1)-IB$，**if**

$I(C_1)\geq IT$ **then**{继续游历后继节点 D}**else**{结束 C_1 点的游历，返回父节点 B_2，取分枝 C_2 继续遍历}

⑤重复第四步，直到遍历所有节点。

本算法综合了深度优先和广度优先的优点，在保证了较高精度的前提下，也得到了很高的查全率。并且，由于引入了基于兴趣度的启发式算法，能对搜索树进行有效的剪枝，大大减小了搜索空间，从而提高了搜索的效率。

2.3 相关反馈实现阈值的自适应

在2.2节中我们对搜索策略所采用的参数 IT，IB，m 赋予了一组经验值，如何能够学习用户的搜索习惯，实现阈值的自适应，是一个个性化系统所必须解决的问题。

相关反馈是信息检索系统中的一种指导性学习的技术，用以提高系统的检索能力。我们可以通过接受用户反馈的哪些是符合用户要求的正反馈结果，那些是不符合要求的负反馈结果，来实现阈值的微调[7]。这里，主观感兴趣度作为阈值调整的评价度量无疑是一个非常好的选择。以下，我们给出在给定阈值 IT，IB，m 之下，用户对搜索结果主观感兴趣度的度量公式：

设 S 为大小为 n 的试验结果样本空间$\{S_1,S_2,\cdots,S_n\}$，fit_i 为试验 i 中用户满意的结果数，$total_i$ 为试验 i 中系统返回的结果数，则 $S_i\equiv\frac{fit_i}{total_i}$。令 X 为取值为0.2，0.4，0.6，0.8，1的离散型随机变量，函数 num(X)定义为样本中所有大于 X-0.2，小于 X 的元素个数，则 X 的概率为 $P(X)=\frac{num(X)}{n}$，对应的 X 的分布率为：

X	20%	40%	60%	80%	100%
P(X)	P1	P2	P3	P4	P5

则用户的主观感兴趣度为

$I=E(X)=0.2P_1+0.4P_2+0.6P_3+0.8P_4+P_5$

我们可方便地根据该公式，给出用户对当前阈值的满意度，自动地进行阈值微调，直到主观感兴趣度达到峰值。

3 实例系统的设计与实现

3.1 PSA 系统总体设计及工作流程

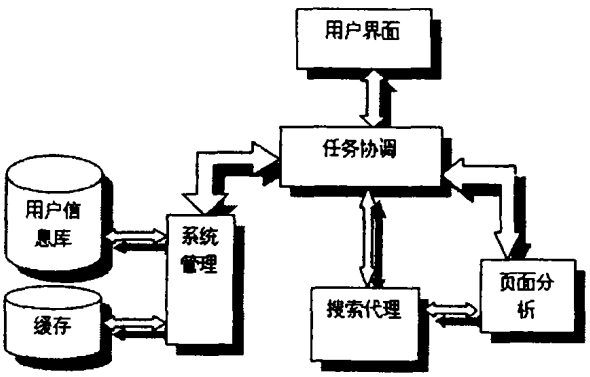

图2 PSA 系统结构图

PSA 系统(Personalize Search Assistant)是为了提供网络个性化信息发现服务而设计开发的，它采用了 VSM 匹配算法，基于客观兴趣度导引的源信息搜集策略，以及相关反馈技术。主要由用户界面、任务协调、系统管理、搜索代理、页面分析5部分组成(见图2)。

3.2 主要模块功能简介

用户界面：接受用户查询，返回搜索结果，并且得到用户

反馈信息。

任务协调：负责系统各个模块之间的协同、通信。

系统管理：提供对用户信息库和缓存的管理，并负责用户搜索习惯的学习。

搜索代理：采用2.4节所论述的策略，进行网上源信息采集。

页面分析：对搜索代理取得的页面，进行特征提取，并使用 VSM 算法给出匹配度。

4 实验结果及评估

4.1 IR 系统的两大评价标准

1)查全率(recall)：挖掘到的文档资料数与实际相关文档资料数之比。

2)查准率(precision)：结果集合中的相关文档数与结果集文档数之比。

作为一个好的 IR 系统，在保证查准率的同时还要能有较高的查全率[6]。

4.2 兴趣度导引算法与 fish-search 算法及广度优先算法的比较结果

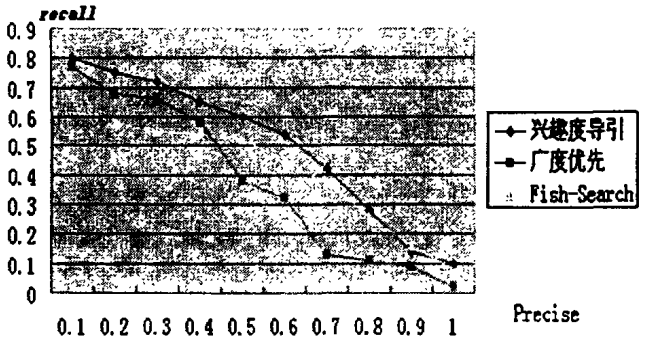

图3 评估结果 r-p 图

我们以 http://WWW.edu.cn/为搜索空间，使用 PSA 系统分别使用基于兴趣度导引的算法(取经验阈值 IT 为 0.34，IB 为0.1，m 为3)、fish-search 算法、广度优先算法，对20个查询进行试验，得到结果如图3。

测试结果表明，广度优先算法具有较高的查全率，但在查准率方面衰减非常快。Fish-Search 算法精度高，但是相应的查全率在低查准率区间表现很差。而兴趣度算法具有极好的稳定性，表现明显优于其它两种算法。

结论 本文引入数据挖掘中的感兴趣度模型，提出了基于客观感兴趣度导引的 Web 页面采集算法，及基于主观感兴趣度的相关反馈度量公式，并且给出了相应的实现系统 PSA。实验证明，以上两种策略的实施大大提高了 WWW 个性化搜索系统的精度和效率，并加强了系统对用户搜索习惯学习的能力。

参 考 文 献

1 邹涛，王继成，等. WWW 上的信息挖掘技术及实现. 计算机研究与发展，1999，36(8)

2 McBryan OA 1994. GENVL AND WWW: Tools for taming the Web Proc. Intl. World Wide Web Conference, ed. By Nierstrasz O. Genneva: CERN

3 Menczer F, Belew R K. Artificial Life Applied to Adaptive Information Agents. Communication Technology Lab, Image Science Group ETH- Zertrum, ETZ F86

4 De Bra, Post. Searching for Arbitratry Information in the WWW: the Fish-Search for Mosaic

5 杨炳儒，等. 感兴趣度的研究综述. 计算机科学，2001，28(10)：43～45

6 Salton G. The Smart Retrieval System-Experiment in Automatic Document Processing. Prentice-Hall, Englewood Cliffs, New Jersey 1971

7 Harper D, van Rijsbercen C J. An evaluation of feedback in document retrieval using co-occurrence data. Journal of Documentation, 1978, 34: 189～216

8 van Rijsbergen C J, Retrieval. London: Butterworths, 1979

(上接第60页)

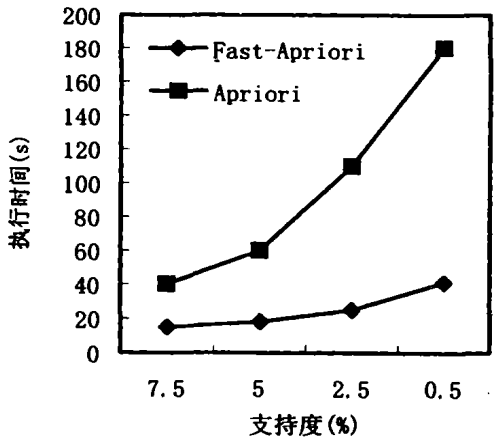

图1 算法执行时间(|T|=20 000)

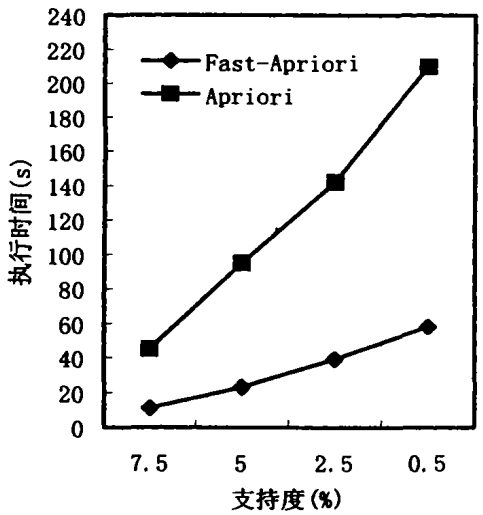

图2 算法执行时间(|T|=40 000)

4 Agrawal R, Srikant R. Fast algorithm for mining association rules. In: Proc. 20th Int. Conf. on VLDB, Santiago, Chile, 1994. 487～499

5 Houtsma M, Swami A. Set-oriented mining for association rules in relational databases. In: Yu P, Chen A, eds. Proc. of the Int. Conf. on Data Engineering, Los Alamitos, CA: IEEE Computer Society Press, 1995. 25～33

6 Savasere A, Omiecinski E, Navathe S. An efficient algorithm for mining association rules. In: Proc 21th Int. Conf. on VLDB, Zurich, Switzerland, 1995. 432～444