

基于本体的元搜索引擎 4 级结果处理算法

秦 杰, 王春云, 谢 蕙, 朱海丰

(河南工业大学 信息科学与工程学院, 河南 郑州 450001)

摘 要: 提出一种基于本体的元搜索引擎 4 级结果处理算法. 该算法结合用户的兴趣本体库, 计算网页特征与本体库概念的语义相关度, 把无相关性或相关性低的网页信息进行过滤. 通过验证表明, 该算法可以有效地加大无关信息的过滤量, 提高了搜索引擎查询结果的相关性.

关键词: 元搜索引擎; 本体; 用户模型; 结果处理

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2010)06-0039-04

Ontology-based Four Degree Results-processing Algorithm for Meta Search Engine

QIN Jie, WANG Chun-yun, XIE Hui, ZHU Hai-feng

(College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China)

Abstract: An Ontology-based four degree results-processing algorithm using semantic relevance to filter the irrelevant information from the results set is designed and implemented in this paper. And the algorithm increased the page numbers of irrelevant information which be filtered. Especially, when the expression of the keywords is not clear, the algorithm in improving the relevance of query results is particularly significant.

Key words: meta search engine; ontology; user model; results-processing

1 引言

随着互联网上信息资源的急剧增加, 搜索引擎作为一种用户检索信息的重要工具, 得到了越来越广泛的应用. 然而, 现有的搜索引擎不能完全满足用户的查询需求, 主要表现在缺乏对用户的查询意图分析处理, 导致结果存在大量的无关信息. 为满足用户精确查询的需求, 建立在本体技术基础上的搜索引擎技术已成为研究的热点^[1].

本体是概念模型的明确的形式化规范说明, 具有精确的语义^[2]. 将本体应用到元搜索引擎中, 能够准确地了解用户的需求. 文中设计一种新的基于本体的用户信息模型, 它将本体论的思想与用户兴趣相结合来描述用户的信息和行为, 实现用户查询需求的清晰语义表达. 同时结合基于本体的用户信

息, 提出一种基于本体的元搜索引擎 4 级结果处理算法, 该算法引入语义相关度对与用户查询需求无关的信息进行过滤, 提高了结果的查询相关性.

2 基于本体的用户信息模型

2.1 用户模型设计

目前利用本体知识表达用户信息在信息搜索和个人兴趣管理方面得到了应用^[3]. 文中设计了一种基于本体的用户模型, 定义如下:

$$\text{UserModel} = (\text{Personall}, \text{PersonalO})$$

其中, Personall 代表用户的信息, 包括用户的基本信息、兴趣描述背景知识和常用的关键词, 用概念类表示; PersonalO 是包含了用户信息的兴趣领域本体库; 在该模型中, 用户的基本信息和兴趣表达都以本体概念也即是本体类的形式出现, 兴趣领域则

收稿日期: 2009-05-08; 修回日期: 2009-07-08

基金项目: 河南工业大学校基金项目(2006BS009); 郑州市科技攻关项目(083SYG25121-7)

以对应的领域本体库 OWL 文件名来表现,OWL 文件对应的是一个完整的可以实现重用、共享的领域本体库。

用户模型中定义用户的一些信息,比如:教育程度、当前动态、兴趣爱好等,这些信息可以被用来当作整体模块类的信息,如计算机专家可能会查看与计算机技术相关的信息.定义用户的常用信息,列出最上层的一些本体概念如表 1 所示.其中,兴趣领域如表 2 表示.

表 1 用户信息的上层本体概念

类名	类的描述
人(Person)	用户的基本信息,如姓名,生日等.
居住地	有关用户的居住信息.
联系群(Contact)	与该用户有联系的人.朋友,同事等.
偏爱(Preference)	表示用户的偏爱,如喜欢猫,蓝色等.
兴趣(Interest)	用户的爱好或相关工作.计算机等.
教育(Education)	用户受教育情况.如大学学位语言等.
职业(Profession)	用户所从事的职业.
关键词(KeyWords)	用户常用的搜索关键词.

表 2 用户兴趣概念层次表

兴趣概念层次	兴趣实例(“InterestInstances”)
商业	(<Root>,Business)
投资	(Business,Investing)
股票	(Investing,Stocks & Bonds)
体育	(<Root>,Sports)
篮球	(Sports,Basketball)
职业赛	(Basketball,Professional)

领域本体库,领域本体(Domain ontology 即领域特异性本体)所建模的是某个特定领域,或者现实世界的一部分.领域本体所表达的是那些适合于该领域的那些术语的特殊含义.这样就可以清晰地表达用户兴趣领域内的概念信息,如计算机系统的简单结构图如图 1 所示.

2.2 用户信息静态数据表示

文中在定义好上面的用户模型和领域本体库后,把他们结合起来定义基于本体的用户模型,其中包括用户的信息本体库、兴趣、兴趣对应的本体库名称和本体库的 OWL 文件.基于本体的用户模型的静态数据列表,如表 3 所示.

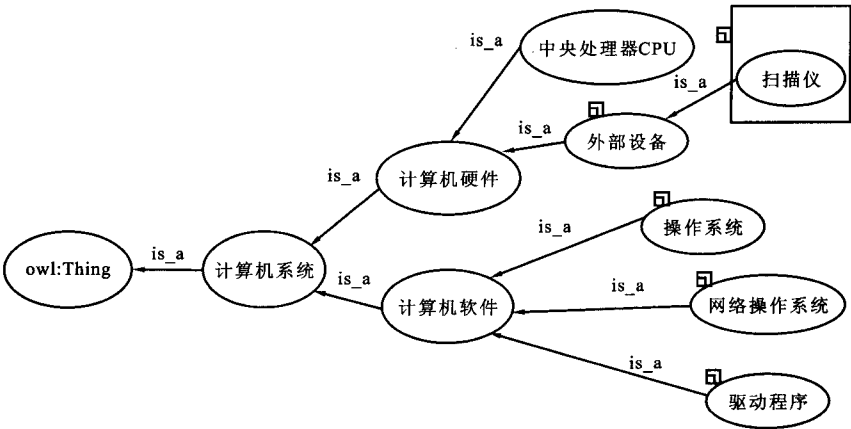


图 1 计算机系统的简单结构图

表 3 静态用户模型数据

用户	兴趣	搜索词	主要概念	OWL 文件
1	玩游戏	网络游戏...	网络游戏,传奇,仙剑,红警...	Game.owl
2	看小说	网络小说...	网络小说,小说名,主人公...	Novel.owl
3	计算机	网络,计算机...	软件,硬件,网络...	Computer.owl

该用户模型的一个突出优点就是可以减少领域本体概念的重复定义,实现了本体库的重用和共享.用户的基本信息和兴趣表达都以本体概念也即是本体类的形式出现,兴趣领域则以对应的领域本体库 OWL 文件名来表现,OWL 文件对应的是一个完整的可以实现重用、共享的领域本体库.例如:如果用户 5 和用户 6 的兴趣都是“计算机”,在以往的定义中,在用户 5 和用户 6 对应的本体库中都定义计算

机领域的相关领域概念,这样就造成了概念的重复定义.

3 基于本体的 4 级结果处理算法

在深入研究多种算法^[4,6]基础上,文中提出了一种基于本体的 4 级结果处理算法.其主要思想是:由于网页的标题和摘要表达了网页的关键信息,对标题和摘要进行基于本体的相关度计算,当相关度

值小于某个阈值时,即认为该网页和用户的查询需求不相关,从而被过滤;最后按相关度大小进行排序,相关度大的排在前面。

该算法详细描述如下所示。

Step1:收集各个元搜索引擎中参于检索任务的源搜索引擎(目标搜索引擎)的查询结果,以此为1级结果集,记作 $C = \{P_1, P_2, \dots, P_n\}$. 每个命中网页包括网页标题、网页摘要、URL、成员引擎中的序号这4项信息。

Step2:对1级结果集进行网页去重得到2级结果集. 采用元搜索去重算法,思想如下:(1)提取记录的网页元数据,判断网页地址是否重复. 如果地址重复,转(5),否则,转(2);(2)判断网页标题是否重复. 如果相同,则转(3),否则,转(4);(3)依次提出每个网页摘要的特征码信息,对特征码与平衡搜索树进行比较,若相似度大于系统阈值,则两条记录重复,否则,转(4);(4)将提取出的特征码插入到平衡搜索树中,并转(1);(5)进行下一个网页处理。

Step3:对2级结果集的结果进行基于目标搜索引擎的权重进行排序,取其排位在前的一部分网页,得到3级结果集;排序算法主要思想如下:

设关键词 keyword 在 Google 搜索引擎的搜索结果结果为 m_1 个,分别是 $page_id(1), page_id(2), \dots, page_id(m_1)$, Google 的权重指数为 K_1 ; 在 Baidu 搜索引擎的搜索结果结果为 m_2 个,分别为 $page_id(1), page_id(2), \dots, page_id(m_2)$, Baidu 的权重指数为 K_2 , 则它们结合后的相关性排序办法就是 $m_1 + m_2$ 个网页的搜索位置和搜索引擎权重结合后的倒排序结果,即

$$\begin{aligned} & \text{Relevance}(page_id(i), K(i)) \\ &= (\text{select } K * page_id \text{ from search_list order by } K \\ & \quad * page_id \text{ desc}) \\ &= \text{sort collect-A and collect-B by } k * page_id \text{ decs} \end{aligned} \quad (1)$$

其中式(1)中,集合 A:

$$\text{Collect-A} = \{K_1 * page_id(1), k_1 * page_id(2), \dots, k_1 * page_id(m_1)\},$$

集合 B:

$$\text{Collect-B} = \{K_2 * page_id(1), k_2 * page_id(2), \dots, k_2 * page_id(m_2)\},$$

搜索引擎权重:

$$K = K(i) = \begin{cases} K = K_1, & \text{Page}(i) \in \text{Page}(\text{google}) \\ K = K_2, & \text{Page}(i) \in \text{Page}(\text{baidu}) \end{cases}$$

Step4:对3级结果集,将网页标题及摘要内容

进行分词处理,保留词性为动词(/v)和名词(/n)的词作为网页特征词,并根据其位置赋上相应权值,相同的词,权值相加,形成网页特征二维数组{(网页特征词,权值 $i(w_i)$)}。

进行领域本体解析,得到本体中的类、实例或属性,并设定这些词的权值,得到特征词集{(本体特征词,权值 $j(w_j)$)},其中与用户兴趣紧密度高的特征词所对应的权值大. 将这两个特征词集看作两个向量,计算夹角的余弦值,作为该网页与该本体概念的相关度 Sim. 计算公式如下:

$$\text{Sim} = \frac{\sum_{k=1}^n (w_{ik} \cdot w_{jk})}{\sqrt{(\sum_{k=1}^n w_{ik}^2) \cdot (\sum_{k=1}^n w_{jk}^2)}} \quad (2)$$

Step5:设定一个相关度阈值,通过阈值策略来判断网页 k 是否是用户查询相关的网页. 如设定阈值 Sim0, 如果得到的相关度值大于 Sim0, 则网页 k 作为4级结果输出,反之则过滤掉,然后按相关度大小进行排序,相关度大的排在前面。

4 实现与分析

4.1 实现

本文实现的系统环境:操作系统为 Windows XP Professional; PC 设备配置为 Pentium D 2.8GHz 处理器, 1GB 内存, 通过局域网连接到 Internet. 本文实现的元搜索引擎模型是基于 Java 语言建立的. 开发工具是 MyEclipse 6.0, 本体技术方面利用 Protégé3.2 构建本体库, 存为 OWL 文件的形式, 在 MyEclipse 6.0 中加载 Jena, 然后通过 Jena API 加载构建的本体模型, 使用 Jena 工具操作本体库. 在关键词输入框中输入“计算机”, 点击“查询”按钮, 结果显示如图2所示。

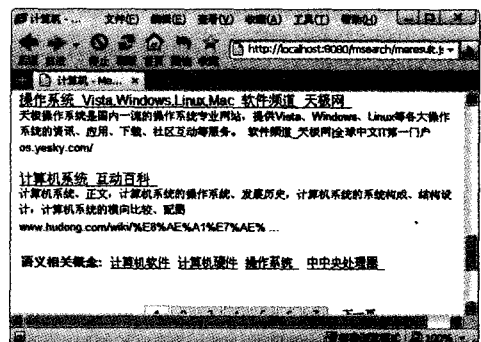


图2 查询效果图

4.2 验证分析

在实验中选取了“计算机”, “操作系统”, “计算

机软件”三个关键词分别进行检索,跟踪每次检索的 100 个网页的处理结果,将处理结果与人工处理结果相比较,统计结果如表 4 所示。

表 4 相关度实验结果

关键词	总过滤 条目数	误过滤 条目数	过滤比 例/%	误过滤 率/%
计算机	27	7	20	7
操作系统	30	6	24	6
计算机软件	28	3	25	3

从表 4 中可以看出,平均 23% 的过滤比例,比元搜索引擎有道(www. youdao. com)的平均去重(过滤)比例 7.7%(10%,7%,6%),高出了十几个百分点,有效实现了无关信息的过滤.同时排序正确率可以高达 90%,实现了相关结果靠前显示,方便了用户查询^[7-8].

同样,以相同的硬件环境,分别采用该结果处理算法和不采用结果处理算法进行测试,共进行 10 次测试,得到的结果如图 3 所示。

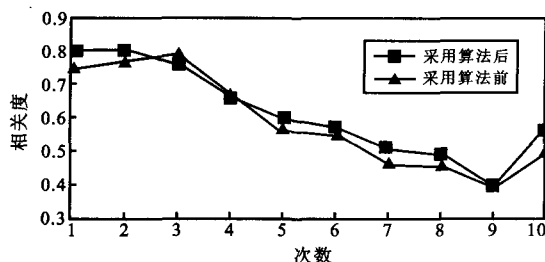


图 3 采用算法前后相关度对比图

在实验中发现,网页特征权值、本体概念权值和相关度阈值这几个参数的设置,影响显示结果的过滤和排序,这几个参数的设置是由用户的兴趣,本体概念的相关度决定的.结果的排序前后与用户的需求是成正比的,说明此算法可以满足基于用户查询意图的查询请求。

再次以查询“网络”为例,在 Baidu 和 Google 中用关键词“网络”查询可以看到,结果信息中包括计算机网络、网络电视、网络小说、网络游戏、网络广告、公司名字含有“网络”二字的公司信息,更有甚者是在网络上传播的不健康信息.在文中设计的系统以用户查询关键词“网络”进行查询,假设用户模型中用户数据完整,根据用户模型兴趣库匹配得到所属领域为计算机领域.点击查询,系统返回 192 条记录.由于时间精力和相关性的关系,一般人们只会查看搜索引擎不超过前面 10 页的查询结果.本文选取各个搜索引擎的前 5 页的显示结果来进行分析,结果中含有网络设备、网络论文等都看作是和计算机相关的.结果分析如表 5 所示。

表 5 系统查询结果分析表

搜索引擎	页数	结果数	无关结果数	相关结果数	查询相关度
Google	5	50	35	15	30%
baidu	5	50	38	12	24%
比比猫	5	50	34	16	32%
有道	5	50	39	11	22%
本系统	5	50	9	41	82%

查询结果中有些网页虽然含有与查询网页相匹配的概念,却不属于用户确定的领域范围,这些网页的特征词与本体中的概念术语特征词不相匹配或者相关度不高,通过相关度的阈值策略,这些网页将从结果集中剔除,为用户消除了信息噪音.表 5 中的查询相关度,利用柱状图表示如图 4 所示。

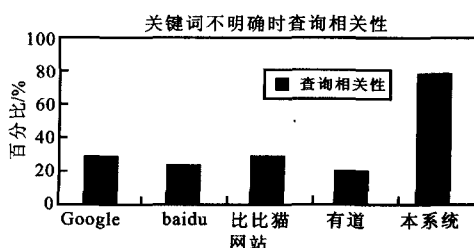


图 4 查询相关性表示图

从上面的分析可以看出,像 Google, Baidu 这样的常用搜索引擎,虽然技术方面相对较完备,但是由于缺乏对用户的查询意图进行处理,所以当用户关键词表达不清时,检索到的结果并不理想.文中通过基于兴趣匹配和基于本体的相关度计算,有效地减少了查询结果中的无关信息,尤其是当用户的关键词表达不明确时,在提高系统的查询结果相关度方面,效果尤为显著。

5 结束语

文中提出了基于本体的元搜索引擎 4 级结果处理算法,详细说明了算法的步骤,同时设计了基于本体的用户模型.该算法增大了无关信息的过滤量,从而提高了信息的查询相关性,尤其当用户的查询关键词不规范时,效果更显著.但文中的本体库还不太完善,将在以后的工作中做进一步的研究。

参考文献:

- [1] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo. Query reformulation over ontology - based peers[C] // Proceedings of the Twelfth Italian Symposium on Advanced Database Systems. Italy, 2004:418 - 425.
- [2] 张体首,蔡明.语义搜索引擎概念模型[J].微电子学与计算机,2007,24(3):171 - 173.

(下转第 46 页)

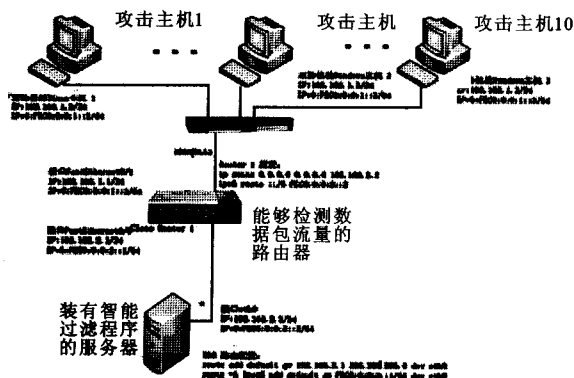


图4 PDIF 攻防模型测试拓扑图

小包长(IPv4 64 Byte, IPv6 80 Byte)的 IPv6/IPv4 混合 75Mb/s 和 100Mb/s 的背景流,捕获率分别为 97.3% 和 72.9%,系统 CPU 使用率 < 80%。攻击检测的指标包括误报率和漏报率,根据以下公式计算得出:

漏报率 = (攻击数 - 报警数) / 攻击数 × 100%,

误报率 = 误报数 / 报警数 × 100%。

在 100 次试验中,分别采用 SYN Flood 以及 IPv6 ND 攻击工具,测试结果漏报率低于 5%,误报率低于 8%。可见,当发生攻击后启动智能过滤器,可以有效地过滤攻击数据包。

6 结束语

文中结合数据包流量检测技术和改进的 TCP Cookie 技术,建立了一种数据包检测——智能过滤 PDIF 防御 SYN Flood 攻击的模型。该算法采用快速算法,能够在很短的时间内对网络攻击作出反应。在检测出现 SYN Flood 的 DDoS 攻击之后,采用了一种验证远程客户端 TCP 连接有效性的智能算法来辨别到达的分组是否是攻击分组,从而智能地将

DDoS 攻击分组拦截,而让正常的网络流量通过,提供一个智能化、高效、实时的抗 SYN Flood 的 DDoS 攻击的方法。以后我们将针对 IPv6 环境下的 IPv6 分片攻击以及 IPv6/IPv4 Land 攻击等新的问题,做进一步的研究。

参考文献:

- [1] 沈清,金心宇,周绮敏. 基于 SYN Cookie 下防分布式拒绝服务攻击算法的分析与实现[J]. 计算机应用,2005,25(12):2745-2747.
- [2] 李磊,赵永祥,陈常嘉. SYN Flood 攻击原理及其应对策略[J]. 中国数据通信,2003(3):22-25.
- [3] Sebastian Kiesel. On the use of cryptographic cookies for transport layer connection establishment[R]. Norway: Trondheim, Summer School,2002.
- [4] 曲同胤,赵敬中. SYN Flood 网络攻击的原理及其防御方法[EB/OL]. [2006-10-10]. <http://www.yesky.com>.
- [5] 刘董,郑宁,徐海涛. 网络异常检测中一种隶属度函数优化的新方法[J]. 微电子学与计算机,2008,25(6):186-189.
- [6] 见晓春,吴振强,霍成义. 同源 SYN 报文两次接收法防御 SYN Flood 攻击[J]. 计算机工程与设计,2008,29(6):1440-1442.
- [7] 蒋凌云,王汝传. 基于流量自相似模型的 SYN Flood DDoS 攻击防范[J]. 南京邮电大学学报,2007,27(2):90-94.
- [8] 王斐,陆建德. 交换网 ARP 欺骗自动检测系统的设计与实现[J]. 微电子学与计算机,2009,26(1):193-196.

作者简介:

连红女, (1951-), 教授. 研究方向为计算机网络安全。

(上接第 42 页)

- [3] 陈沈焰,吴军华. 基于本体的概念语义相似度计算及其应用[J]. 微电子学与计算机,2008,25(12):96-99.
- [4] Wikipedia. Semantic web[EB/OL]. (2008-12-01) [2008-12-19]. <http://en.wikipedia.org/wiki/Semantic>[CD#*2]Web.
- [5] Trajkova J, Gauch S. Improving ontology-based user Profiles[C] // Proceedings of RIAO 2004. France, 2004: 380-389.
- [6] 陈琳娜. 基于本体的个性化用户模型研究[D]. 秦皇岛: 燕山大学信息科学与工程学院,2006.

- [7] 张强弓,喻国宝. 一种元搜索引擎结果处理模型[J]. 华南理工大学学报:自然科学版,2004(11):47-51.
- [8] 王梅. 基于本体进行自动分类的元搜索引擎的设计与实现[J]. 现代图书情报技术,2008(9):58-62.

作者简介:

秦杰男, (1968-), 博士, 副教授, 硕士生导师. 研究方向为网络信息处理。

王春云女, (1981-), 硕士研究生. 研究方向为信息处理。

谢蕙 硕士研究生. 研究方向为信息处理。

朱海丰 硕士研究生. 研究方向为信息处理。