

需求分析器中搜索工具研究

刘志雄, 陈松乔, 孙莹

(中南大学 信息科学与工程学院, 湖南 长沙 410083)

摘要:需求分析器是异构服务组装平台的重要组成部分, 需求分析工具的成功开发是进行异构服务组装和工作流调用的前提。根据需求分析应当与服务搜索交互进行的特点, 对搜索机制进行了研究。探讨了服务搜索引擎框架的设计方式, 然后讨论了在此基础上确定需求分解粒度的方法。该引擎能够进行简单的个性化服务搜索, 可以帮助用户更快、更准确地找到所需的服务信息, 还可以避免无关服务信息的干扰。

关键词:服务组装; 需求分析; 分解粒度; 搜索引擎; 用户兴趣库

中图分类号: TP311.5

文献标识码: A

文章编号: 1673-629X(2006)10-0011-03

Research on Searching Tools for Requirement Analyser

LIU Zhi-xiong, CHEN Song-qiao, SUN Ying

(Information Science and Engineering College, Central South University, Changsha 410083, China)

Abstract: Requirement analyser is an important part of isomeric platform for service - composition, the development of its tools is the premise for isomeric service - composition and workflow - calling. Discuss the service - searching mechanism according to the characteristic of the relationship between requirement analysis and service - searching. Moreover, an approach to deciding the granularity of requirement decomposition and a method for designing the frame of the service - searching engine are to be recommended. This engine presents users with simple specific - searching service, to help deriving information more quickly and more accurately. It also helps to stop unnecessary information from entering.

Key words: service - composition; requirement analysis; decomposition granularity; searching engine; user - interest library

0 前言

需求分析工具的研究是异构服务组装平台重要的组成部分。在这个研究领域, 以下两个问题一直是关注的焦点:

1) 如何设计一种灵活的机制, 帮助开发者确定需求分解的粒度。

2) 如何合理地设计一个服务搜索引擎, 帮助用户准确、迅速地找到所需服务, 并避免无关信息的干扰^[1]。

而这两个问题之间, 又存在着直接的关联。这是因为, 一个合理的服务搜索引擎的出现, 能够快速地搜索到异构服务与需求功能进行对应, 势必会帮助用户更好地确定功能分解的粒度。

随着对异构网络服务以及组件(服务)组装技术研究的日趋深入, 出现了以分布式搜索引擎占主导地位的如下4类搜索引擎:

(1) 分布式元搜索引擎。分布式搜索引擎设计简单、快速, 并且任何一个单元可以随时地摘掉, 且影响不太大,

但是对于大规模的并发搜索并非好的解决办法。

(2) 散列分布搜索引擎。该引擎抗压, 但是对于单个索引服务器或者文档服务器的容量等动态的调整较困难。

(3) P2P 分布搜索引擎^[2]。Peer 2 Peers 搜索引擎可以很大, 而且基本上不需要有维护成本, 但是中心服务器的更新效率很低, 信息源不稳定。

(4) 局部遍历型搜索引擎。容易解决抗压, 搜索精度高、搜索效率高, 但是设计复杂, 调整索引所在节点的位置不容易。

总体来说, 搜索引擎的设计方法很多, 但是并没有明确的针对需求分析的搜索机制, 也没有较好的支撑确定功能分解粒度的算法, 用户在进行功能分解的时候往往会陷入困境。

文中根据需求分析应当与服务搜索交互进行的特点, 借鉴个性化网络搜索引擎设计理念, 对异构服务搜索工具进行了研究, 探讨了个性化服务搜索引擎的设计方式, 并在此基础上讨论了功能分解粒度的控制问题。

1 服务搜索引擎

异构服务组装平台的用户在进行需求分析的时候, 提出的查询请求往往很模糊, 仅仅根据检索词条来判定, 常常在返回结果中包含大量不相关文档, 接下来还要对搜索

收稿日期: 2006-01-11

作者简介: 刘志雄(1982-), 男, 湖南长沙人, 硕士研究生, 研究方向为软件复用技术; 陈松乔, 教授, 博士生导师, 研究方向为计算机网络、软件工程。

结果进行过滤。这个过程很烦琐,而且经常达不到预期的效果。文中提出了如下构想:用户在进行异构服务组装前很长一段时间内,要从事查询并浏览与该软件项目有关的网页和文档的工作,因此,可以给开发者提供异构服务搜索引擎的服务,通过长期观察、监视用户的搜索行为,引擎能够识别出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,自我调整搜索策略,返回最贴近用户需要的信息。

1.1 服务搜索引擎框架

搜索引擎通用框架主要包括下面几个部分:用户服务接口、需求分析器、检索器、索引知识库、索引器、分析器、spider(简记为 sp)、查询过滤器、用户兴趣库。如图 1 所示^[3]。

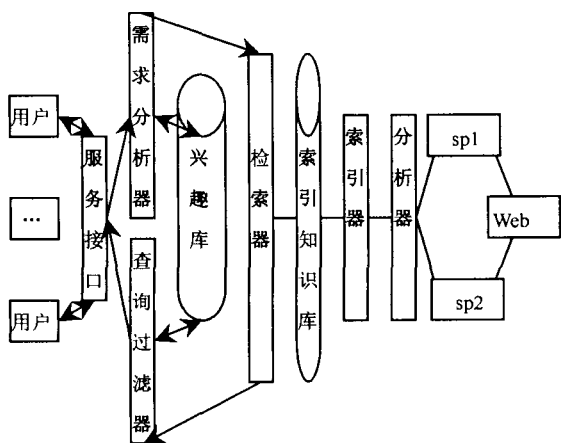


图 1 服务搜索引擎框架

与普通搜索引擎相比,服务搜索引擎多了 3 个部分:信息需求分析器、信息查询过滤器和用户兴趣库。其中,用户兴趣库起着非常重要的作用,是实现精确搜索的关键。下面将这几部分的功能简述如下:

(1)信息需求分析器。用户提出的查询请求往往很模糊,仅仅根据检索词条来判定,常常在返回结果中包含大量的不相关文档,根据用户兴趣知识库中信息对用户查询请求进行特定分析则可以帮助服务搜索引擎确定用户确切所指。

(2)信息查询过滤器。信息查询过滤器主要指对检索器返回的查询结果进行特定处理。比如根据用户兴趣词条来对返回文档打分,然后将结果排序,并设定阈值,最后输出用户真正感兴趣的文档。

(3)用户兴趣库。根据一定的用户兴趣模型,存放用户兴趣知识。好的用户兴趣模型不仅要求客观、全面表达用户兴趣知识,而且还要具备良好的后期兴趣评估可操作性。

1.2 用户兴趣知识提取以及用户兴趣模型的建立

用户兴趣知识必须建立在一定的兴趣提取方法和兴趣模型建立规则基础上。

1.2.1 兴趣知识的提取

主要通过如下两种方法提取用户兴趣知识:

(1)依据用户的查询行为进行提取。即通过跟踪用户的浏览行为来推测用户兴趣以建立用户兴趣模型。规则有很多,比如,如果用户保存某个页面,则推测出用户对该页面的内容感兴趣。反之,如果用户没有保存链宿页面就马上返回,或者进入更深一级链接,则表明用户对该链宿页面不感兴趣。又比如,如果用户反复地回到某一页面,则说明用户对该页面感兴趣。根据用户浏览行为提取特征中存在大量的“噪音”。例如,一些网页可能不能提供任何信息给用户,却会被经常访问,因为它们有大量的超级链接。因此,可以设定一个阈值,如果一个文档中包含的超级链接超过这个阈值,就可以被看成是一个参考链接的目录页。

(2)依据用户查看的内容进行提取。即根据用户的浏览内容建立用户兴趣模型,而用户浏览内容通常是指用户浏览过的文档,使用最多的是词频法,根据文件中词条的频率特性进行目标特征的提取。现在基于内容的用户兴趣提取的对象一般都是文档,文档中存在很多标记信息,这些标记信息往往对文档的内容有很高的概括性,因此可以利用这些标记信息提高用户兴趣特征提取精度。比如出现在标题中的词条或以粗体、大字体显示的词条,显然比较重要,因此在词频法中可以增加它的权重。此外,词条在文档中的不同位置也体现了词条的不同重要程度。比如出现在文档起始区域和终止区域的词条,往往也被认为十分重要。在实际兴趣提取中,也可以通过赋予不同的加权参数,以提高这些词条的权重。

1.2.2 兴趣模型的建立

兴趣模型是指对于用户感兴趣的信息的可计算描述,这里介绍一种用三元组表示的兴趣模型(兴趣词条、兴趣权重、词条新鲜度)。兴趣结点用三元组 (p_i, w_i, x_i) 表示,简记为 $\text{Node}(p_i)$,其中 $p_i \in P$, P 为词条集合, $P = \{p_1, p_2, \dots, p_m\}$, p_1, p_2, \dots, p_m 分别表示兴趣(词条), m 为字典的大小, w_i 为兴趣词条 p_i 的权重, x_i 为兴趣词条 p_i 的新鲜度。所有兴趣的集合构成兴趣字典,记为 U 。 D 为 WWW 缓存中的文本集合。

在词频中,考虑到各个词条在文档中的不同位置体现其不同的重要性,对词条兴趣加权重,即位置词频 spf_{ij} 。为了计算词条新鲜度,对于文档 dn ,使用一个文档新鲜度函数 $dtx(n)$ 。这是一个单调非递减函数,用来保证越是最近访问的页面,对用户当前的兴趣作用越大。其中 n 指缓冲中的第 n 个时间页面^[4]。

兴趣结点 $\text{Node}(p_i)$ 的词条权重公式如下:

$$\text{Node}(p_i) \cdot w_i = \sum_{j=1}^n (spf_{ij} \times E_j) \quad (1)$$

兴趣结点 $\text{Node}(p_i)$ 的词条新鲜度公式如下:

$$\text{Node}(p_i) \cdot x_i = \sum_{j=1}^n \frac{spf_{ij} \times E_j}{\text{Node}(p_i) \cdot w_i} \times dtx(j) \quad (2)$$

式中 spf_{ij} 为词条 p_i 在文本中 d_j 的位置词频, n 为 D 中文本的个数, E_j 为文本兴趣系数, $dtx(n)$ 为文档新鲜度函数。

得到兴趣词条 p_i 的权重和新鲜度后,可以根据公式 $t_i = w_i \times f(x_i)$ 计算词条 p_i 的兴趣度,式中 $f(x)$ 为词条新鲜度对权重的影响函数。

词条兴趣度是网络服务搜索特性分析的最终依据。

1.3 异构服务搜索的特性分析

现举一个例子来说明异构服务搜索的特性分析处理过程。

假设已经得到了用户的兴趣字典,它由若干兴趣词条与相应词条的权重以及新鲜度组成,根据这些词条与相应词条权重、新鲜度构造用户个人兴趣树。兴趣树的根节点是总类,叶节点是兴趣字典中的词条,其他节点可能是兴趣字典中的词条也可能不是(这依赖于兴趣字典中词条是否存在蕴涵关系)。从根节点到叶节点概念上越分越细。每一个节点由词、权重、新鲜度三部分组成。兴趣生成树中每一中间节点词的权重按公式(3)计算,每一中间节点词的新鲜度按公式(4)计算^[5]。

$$\text{Node}(p_j) \cdot w_j = \sum_{i=1}^k w_i \cdot w_i \quad (3)$$

$$\text{Node}(p_j) \cdot x_j = \sum_{i=1}^k \frac{w_i x_i}{w_i} \quad (4)$$

式中 w_j 为中间节点 p_j 的权重, x_j 为中间节点 p_j 的词条新鲜度, k 为节点 p_j 的子节点个数, w_i 为子节点兴趣词条 p_i 的权重, x_i 为子节点兴趣词条 p_i 的新鲜度。

在获得中间节点的权重与新鲜度后,可以根据公式 $t_i = w_i \times f(x_i)$ 计算每一节点的兴趣度。如果一个节点的兴趣度越大,则认为用户对此节点中包含的内容兴趣越大。因此当用户输入检索词存在一词多义情况下,可以通过比较几种含义在用户个人兴趣树中兴趣度来选择用户真正感兴趣的含义。

以 interest 关键词的查询为例,假设已建立兴趣搜索树,如图 2 所示(图中所标注的兴趣度已经被单位化)。interest 一词有两种含义:一是经济术语“利润”;二是文化上的“兴趣”。那么用户是想搜索“利润”的含义还是“兴趣”的含义,可以通过分析用户个人兴趣树来选择。在用户个人兴趣树中,用户对经济的兴趣度为 0.15,对文化的兴趣度为 0.08。由于经济的兴趣度比文化的兴趣度大,因此认为用户对经济类型的东西更感兴趣,因而系统认为用户实际查询的是“利润”含义的。实际上,这种方法是按兴趣度预先将用户划归为不同兴趣类型,然后再根据用户兴趣类型选择相应检索词含义,以达到特性分析的目的^[6]。

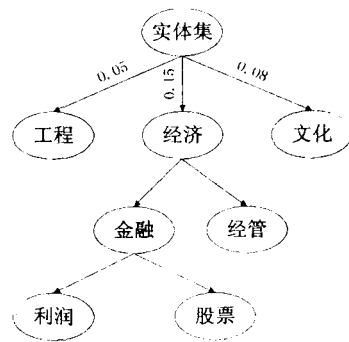


图 2 兴趣搜索树

2 结论及相关研究

综上所述,与其他需求分析工具相比,文中所叙的需求分析器具有以下优点:由于备有个性化的服务搜索引擎,可以协助用户更加恰当地表达需求,从而精确、迅速地从 Internet 上搜索到所需要的信息,并且有效地避免了无关信息的干扰;同时,在需求分析与服务搜索交互的基础上,能够为需求分解粒度的控制提供参考。

异构服务组平台需求分析工具的研究是一个涉及面相当广泛的课题,它包括应用需求的表示工具研究、应用需求的功能分解以及分解建模后的运行流程的表示工具研究等。后续研究包括:需求分析建模工具的集成、需求分解粒度的精确控制研究、需求分析的成果—运行流程的形式化推导以及运行流程的文档化表示等。

参考文献:

- [1] 李雪梅. 网络搜索的个性化服务[J]. 中国信息导报, 2003(3): 26-27.
- [2] 胡 坤. 等待第三代搜索引擎[J]. 电子商务世界, 2005(8): 40-44.
- [3] 戴建中. GnetFtp 搜索引擎的算法设计与实现[J]. 汕头大学学报(自然科学版), 2005(3): 69-74.
- [4] Ozan E. Virtual reality in requirement analysis for CIM system development suitable for SMEs[J]. International Journal of Production Research, 2002(7): 3693-3708.
- [5] Merunka V. Object-oriented approach in requirement engineering for the analysis of information systems[J]. Journal of Forest Science, 2005(3): 13-18.
- [6] Drake J M. Approach and case study of requirement analysis based on acquisition ontology[J]. International Journal of Intelligent Systems, 2000(4): 1125-1155.

(上接第 10 页)

参考文献:

- [1] Artiges M. BEA Weblogic Server 8.1 Unleashed[M]. 北京: 机械工业出版社, 2005.
- [2] 王爱冬, 阳国贵, 张 涛. J2EE 架构下的数据库访问技术分析与研究[J]. 齐齐哈尔大学学报, 2005, 21(3): 39-42.

- [3] 杨 瑞, 蔡 虹. 连接池技术及其 java 实现[M]. 应用技术, 2003(6): 26-29.
- [4] 黄 文, 谢寄石. 基于 J2EE 的数据库连接服务[J]. 电子科技大学学报, 2002(2): 68-71.
- [5] 陈梅容, 郭 俊, 朱兵章. 在 JSP 中采用连接池技术优化数据库连接[J]. 机电工程技术, 2004, 33(4): 59-60.