一种基于互联网智能元搜索引擎的研究*)

包骏杰[1] 马 燕[2]
(重庆教育学院计算机与现代教育技术系 重庆400067)[1]
(重庆师范大学信息技术系 重庆400047)[2]

摘 要 WWW的迅速发展,使得开发新型的搜索引擎成为Web发展过程中亟待解决的问题之一。结合信息检索领域和人工智能领域最新的发展状况,本文提出了一种全新的解决方案——互联网智能元搜索引擎。文章在回顾已有的搜索引擎技术的基础上,首先提出了智能元搜索引擎的基本框架,之后详细介绍了实现的关键技术。这些技术包括:机器学习中的文本学习技术,中文文本表示技术,信息过滤技术,个性化用户建模技术等。

关键词 智能元搜索,文本学习,信息过滤,个性化用户建模

Research of an Intelligent Meta Searching Engine Based on Internet

BAO Jun-Jie MA Yan
(Depertment of Computer and Modern Technology, Chongqing Education College, Chongqing 400067)[1]
(Department of Information Technology, Chongqing Normal University, Chongqing 400047)[2]

Abstract The rapid development of the WWW has made the exploration of the new tape of Search Engine become the urgent question in the process of the Web development. Combining the latest development of information retrieve with the AI, the author in this thesis puts forward a new solution-Web Intelligence Element Search. On the basis of reviewing the available technology of Search Engine nowadays, the author sets forth firstly, the basic framework of the Intelligence Element Search, and introduces the key technologies to realize it in detail, including Textual Study, Indicative Technology of Chinese Words, Information Filtration, Individual Customer Modeling etc in the machine learning.

Keywords Intelligence element search, Textual study, Information filtration, Individual customer modeling

1 搜索引擎

互联网的迅速发展,正成为知识经济时代不可缺少的信息载体。为了帮助人们搜索网络信息,一大批网络信息搜索工具应运而生。网络搜索引擎的研究开发成为当今互联网应用开发的热点,其发展历程由主题指南发展到独立型搜索引擎,再由独立型搜索引擎发展为集成型搜索引擎,直到目前的元搜索引擎。

目前在互联网上大部分搜索引擎的工作原理如图1。通过搜索机器人(搜索蜘蛛)程序,在全球范围内的Web服务器上定期搜集网页;将搜集到的网页进行分析、整理,提取关键词后,放入搜索引擎的索引数据库中;用户向搜索引擎的界面输入关键词进行查询。但是,搜索引擎还存在以下缺陷:查全率不高,查准率不稳定,不能很好地表达用户需求,提交的搜索结果中包含大量用户无关信息,即缺乏个性化和智能化。针对这些问题,目前提出了以下解决办法。

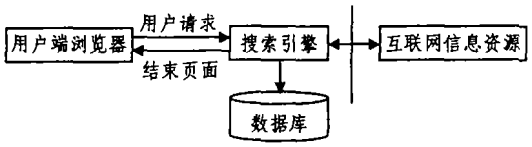

图1 搜索引擎工作原理示意图

1.1 Agent技术

Agent技术是当前人工智能领域发展最快的技术之一,其核心思想在于自适应、学习、协作。将Agent技术与搜索引擎技术相结合,主要是希望Web信息检索更加智能,具有学习的能力,能够捕获用户不断变化的兴趣,过滤掉用户无关的信息。Letizia、Syskill和Webert、Personal Web Watcher等都是比较成功的信息Agent系统。Letizia可以跟踪用户的浏览行为,发现用户的兴趣,建立用户的个性化模型,根据用户模型对用户下一步将要选择浏览的超链进行推荐。Harvest、FAQFinder、Information Manifold等是面向某个特定领域的信息Agent系统。另外一类智能信息Agent,它们能够在所采集到的信息资源中提取概念层次,因此可以更加容易得到满足人类语义查询的需求。比如有:ILA(Internet Learning Agent)、HyPurSuit。

1.2 元搜索引擎

从数据库提取信息的单个搜索引擎具有局限性,因此有必要发展性能更优越的新型引擎搜索技术,元搜索引擎应运而生。元搜索引擎(Meta Search Engine),也称集成搜索引擎、多搜索引擎、索引搜索引擎等,用户只需递交一次检索请求,由元搜索引擎负责转换处理后提交给多个预先选定的独立搜索引擎,并将所有查询结果集中起来以整体统一的格式呈现到用户面前。它是将整个因特网作为一超大型的动态数据库。由于采用了一系列优化运行机制,能在尽可能短的时间内提供相对全面、准确的信息,即使不能完全满足用户需求,仍可

*)重庆市教委应用基础研究项目(编号020805)资助,"重庆市高等学校优秀中青年骨干教师资助计划"资助。包骏杰 讲师,主要研究方向:计算机网络与数据挖掘。马 燕 教授,研究方向:计算机网络新技术、多媒体数据检索。

以作为相对可靠的参考源进行扩展搜索，因此成为备受推崇的信息搜索首选入口。元搜索引擎主要分为基于服务器端和基于用户端两大类。基于用户端的元搜索引擎根据用户应用模式又可分为基于万维网的免费搜索引擎、可供免费下载的客户端桌面应用型、可共享或授权使用的桌面应用型等。

2 IMSE 系统的体系结构

智能元搜索引擎 IMSE(Intelligent Meta Search engine)的核心思想就是要将多个独立的搜索引擎的检索能力集成于统一的用户界面之下。在向用户反馈查询结果的时候，需要删除多个搜索引擎中提交的重复 URL。Metacrawler 是由华盛顿大学研制的元搜索引擎，它能够同时调用6个搜索引擎并行检索，包括 Lycos、Infoseek、WebCrawler、Excite、AltaVista。SavvySearch 是一个并行的元搜索引擎，可以同时调用21个独立的搜索引擎。除此而外，还有大量的桌面元搜索引擎，如 EchoSearch、Webcompass、Webseeker 等。桌面搜索引擎的一个突出特点在于，能够跟踪用户在客户端的浏览行为，从而方便地建立用户模型。

针对搜索引擎的研究现状，本文主要采用 Agent 技术，将机器学习、信息检索、中文处理等方面的多种技术应用到搜索引擎领域，设计和实现了一种新型的互联网智能元搜索引擎 IMSE。下面对 IMSE 系统的体系结构、实现中的关键技术进行论述。

2.1 搜索引擎选择功能

智能元搜索引擎是对多个独立搜索引擎的集成，每个搜索引擎对特定问题的查询效率是不同的。在 IMSE 中对于用户提交的查询请求，可以自动选择对此问题查询效率高的部分引擎完成检索，其余搜索引擎同时就可以处理别的查询请求，即在元搜索引擎中加入异步处理的概念，可以提高整个系统的吞吐量。关于查询问题和搜索引擎的匹配程度，在 IMSE 中有专门的评价数据库和评价算法支撑。

2.2 检索指令转换功能

元搜索引擎中的每个独立搜索引擎都拥有各自的查询语言，因此需要把元搜索引擎界面下输入的查询指令，转换成为各个独立搜索引擎自己的查询指令，即在元搜索引擎中还要集成“全局/局部指令辞典”。

2.3 搜索结果的合并

通过多个独立搜索引擎得到的检索结果，最终要合并成为逻辑上统一、格式一致的结果提交给用户。合并的任务有两个：第一，剔除重复的 URL；第二，对查询得到的结果项进行排序。在合并的过程中还应该考虑到每个用户的兴趣差异，即在检索结果的合并过程中可以加入个性化的设计。

另外，本系统具有智能化的功能，因此还应该在搜索引擎的设计中加入对访问用户个性化建模的功能。由此得到系统体系结构图如图2。该图中各个模块功能如下：

界面模块的主要功能是搜集用户在搜索引擎中输入的关键词，以及记录用户输入此项关键词之后访问过的部分网页(即那些在搜索引擎数据库中存在的网页)。界面模块将采集到的关键词存入用户关键词的数据库中；而将用户访问的网页也直接从其它独立搜索引擎的数据库中转存过来，放入 HTML 文档数据库中。学习 Agent 采用文本学习等机器学习的方法，对用户关键词数据库和 HTML 文档数据库进行学习，从而得到用户的个性化模型。推荐引擎的主要工作有三个：第一，整合独立搜索引擎返回的搜索结果，删除其中重复的 URL；第二，根据各个检索项与用户模型的相关程度，对它们进行评分和排序，从中过滤掉分数很低的检索项；第三，评价各个独立搜索引擎对此类问题进行查询时的表现，结果保存到 SECR 数据库(Search Engine Credit Rate)中。

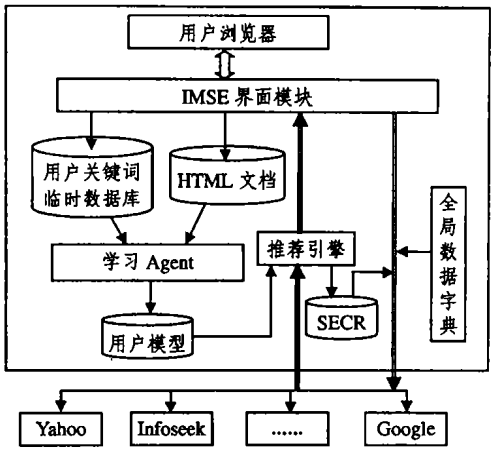

图2 IMSE 体系结构

3 IMSE 实现的关键技术

本文所设计的 IMSE 系统与其它普通元搜索引擎的区别在于它引入了大量人工智能方面的技术，使得该系统具有个性化和智能性。这些技术包括基于 Web 的挖掘、用户的个性化建模、文本分类和学习等。下面对其中的关键技术予以介绍：HTML 文档表示法、用户兴趣模型的建立和更新、在搜索结果合并时所采用的信息过滤技术。

3.1 HTML 文档的表示

当前计算机中对“本文”信息的表示，主要采用“文本特征提取”以及矢量空间模型(VSM)的方法，其中 TFIDF 文本向量表示法是最具代表性的方法之一。由于 HTML 页面是半结构化的文档，大量的文本信息包含于标记(tag)之中。某些 tag 中的文本信息对于提示文档主题非常有用，比如：题头标记〈TITLE〉、标题〈H1〉、…、〈H6〉、格式〈B〉、〈U〉、〈I〉、超链〈a href=“*.htm”〉等，对于这些 tag 中的文本信息应该赋予较重的权值。由此，提出以下解决方法：

定义 tag 集 $S=\{TITLE,H1,H2,H3,H4,H5,H6,B,U,I,URL,Meta,Ordinary\}$，相应的权值集为 $W=\{W_i|s\in S\}$(其中 *Ordinary* 表示普通文本信息，即 $W_{Ordinary}=1$)。则有：

$$w_i=\frac{\sum_{i\in S}(W_i*tf_i)*\log(N/n_i)}{\sqrt{\sum_j(\sum_i(W_i*tf_j)*\log(N/n_j))^2}} \quad (1)$$

w_i 表示第 i 个特征的权值。其中 tf_i 表示第 i 个特征在标记 s 中出现的频率，N 表示文档集中的文档总数，n_i 表示特征至少出现一次的文档数量。

3.2 用户模型的创建和更新

IMSE 中的用户模型，是通过对界面模块捕获用户输入的关键词和用户访问过的 Web 页面，进行学习的基础上而得到的。首先通过用户在界面中输入关键词来建立用户模型的初值。若最初设定初始类集合为 $C=\{C_1,C_2,\cdots,C_t\}$，类中心向量集合为 $O=\{\vec{o}_1,\vec{o}_2,\cdots,\vec{o}_t\}$，则用户模型的初值即为，$O=$ {用户输入的关键词 }。

之后，将用户访问过的 HTML 文档 $D=\{\vec{d}_1,\vec{d}_2,\cdots,\vec{d}_m\}$ 作为训练事例集，其中 $\vec{d}_i$ 是采用式(1)表示的某个 HTML 文

档的特征向量。通过机器学习中文本学习的方法，可以对初始的用户模型进行修正，从而获得动态的用户模型。具体算法描述如下：

输入：聚类算法的最大半径 R；类集合 $C=\{C1,C2,\cdots,Ct\}$；类中心向量集 $O=\{\vec{o}_1,\vec{o}_2,\cdots,\vec{o}_t\}$；待分类的文本特征向量集为 $D=\{\vec{d}_1,\cdots,\vec{d}_n\}$；

输出：已经更新的类集合 C 和类中心向量集合 O；

算法：

(1) $k=1$；

(2) $\forall \vec{d}_k\in D$，$\min=DIS(\vec{d}_k,\vec{o}_i)$；$x=\min$；classNo$=i$；

(3) 若 $x<=R$，则 $C_{classNo}=\vec{d}_k\cup C_{classNo}$；

$$\vec{o}_{classNo}=\frac{Num(C_{classNo})*\vec{o}_{classNo}+\vec{d}_k}{Num(C_{classNo})+1};$$

(4) 若 $x>$ R，则构造一个新类 C_j；

$C_j=C_j\cup\vec{d}_k$；$\vec{o}_i=\vec{d}_k$；

(5) $k=k+1$；

(6) 如果 $k<=n$，则跳转至(2)，否则结束。

该算法可以得到不断更新的类集合 C 和类中心向量集合 O，其中类中心向量集合 O 就是用户的兴趣模型。

3.3 搜索结果的合并和过滤

当每个独立搜索引擎都得到检索结果的时候，还不能把这些结果直接推荐给用户，因为这当中包含有大量的重复索引项，即大量重复的 URL；还包含有大量用户无关的索引项。因此，集成独立搜索引擎的检索结果时，至少需要两个步骤的工作：第一步，去掉重复 URL；第二步，根据用户模型过滤掉无关 URL，只保留用户感兴趣的索引项，并按照它们与用户模型的相关度从大到小排序，再提交给用户。关于如何去掉重复的 URL，请参见相关文献；在此我们仅就 URL 的过滤，或者说 URL 和用户兴趣的相关度排序的算法进行讨论。

各个独立的搜索引擎提交的查询结果都有两项评估数据：第一，查询结果项在当前搜索引擎中的排序；第二，查询结果项和用户模型的相似度。若假设 t_{ij} 表示在第 j 个搜索引擎中的第 i 个检索结果项，则有：

(1)用 $Rank(t_{ij})$ 来表示检索结果项 t_{ij} 在搜索引擎 j 中的分值，则 $Rank(t_{ij})=(m_j-i+1)/m_j(i\in 1,2,\cdots,m_j)$。其中 m_j 表示第 j 个搜索引擎中所返回的索引项的总数。

(2)用 $Sim(t_{ij})$ 表示检索结果项 t_{ij} 和用户模型的相似度，则将 t_{ij} 对应的文本向量 $\vec{t}_{ij}$ 与最终的用户模型对应类中心向量集 O，进行向量之间相似度的计算：

$$Sim(t_{ij})=Max(Dis(\vec{t}_{ij},\vec{o}_1),\cdots,Dis(\vec{t}_{ij},\vec{o}_t))$$

(3) 最后得到检索结果项 t_{ij} 的评估值为 $R(t_{ij})=Rank(t_{ij})+Sim(t_{ij})$。

在 IMSE 中，针对某个输入的检索关键词，如何选择性能较优的独立搜索引擎进行检索，这方面的问题很多文献中均已提到，在此不再详述。

结束语 Web 信息检索系统作为用户层和 Web 信息层之间的中间层，包含了搜索引擎与目录、元搜索引擎、信息检索 Agent 等多个层次。这些系统在传统信息检索的基础上采用了针对 Web 特点的各种技术以提高检索效果。我们开发的智能元搜索引擎 IMSE 兼备了若干个层次的特点，将多种技术融合起来。比如 IMSE 既采用了元搜索引擎的技术，又采用了人工智能中机器学习的技术。它能够自动建立用户兴趣模型，并且随着用户兴趣的变化而不断变化；同时还采用了文档自动分类，信息过滤等技术。随着 Web 上用户群体、信息量的迅猛发展，现有 Web 信息检索系统在信息收集、分类、可视化等方面仍缺乏良好的可扩充性。如何利用自然语言处理、数据挖掘和分布式处理等技术来提高 Web 信息检索能力，将是 Web 信息检索在未来几年中的一个重要发展趋势。

参 考 文 献

1 Buraga S C. Web Technologies(in Romanian). Matrix Rom Publishing House, Bucharest, 2001

2 Buraga S C, Rusu T. Search Semi-Structural Data on Web. In: The Seventh Intl. Symposium on Automatic Control and Computer Science-SACCS 2001 CD-ROM Proc. Iasi 2001

3 Cooley R, Mobasher B, Srivastava J. Web Mining: Information and Pattern Discovery on the World Wide Web, [DB/OL]http://www.google.com

4 Bhrart K. Web SearchPad: Explicit Capture of Search Context to Support Web Search. In: Proc. of the Eight Intl. World Wide Web Conf. WWW9, ACM Press, 2000

5 Lieberman H. Letizia: An Agent that Assists Web Browsing. In: Intl. Joint Conf. on Artificial Intelligence, 1995

6 Braden B, et al. Kann: Introduction to the ASP execution environment: [Technical report]. USC/Information Science Institute, Feb. 2000

7 李晓黎，刘继敏，史忠植. 基于支持向量机与无监督聚类相结合的中文网页分类器. 计算机学报，2001，21(1)