个性化搜索引擎
原理与技术

李树青　韩忠愿　编著

科学出版社
www.sciencep.com

个性化搜索引擎原理与技术

李树青　韩忠愿　编著

科学出版社
北京

内 容 简 介

本书通过对基于个性化信息推荐技术的搜索引擎框架和基本技术的探讨，介绍了利用搜索引擎服务器日志中所具有的关键词序列得到用户模式，并按照事务模式聚类的方法实现用户个性化特征的表达，最后在搜索引擎的网页索引中，利用得到的用户个性化特征改进传统的 PageRank 算法。通过上述工作，本书构建了一个完整的基于 Web 个性化信息推荐技术的搜索引擎框架结构。

本书可作为计算机专业的本科生和研究生的参考用书，也可供相关技术人员参考。

图书在版编目(CIP)数据

个性化搜索引擎原理与技术/李树青，韩忠愿编著. —北京：科学出版社，2008

ISBN 978-7-03-022255-8

Ⅰ. 个… Ⅱ. ①李…②韩… Ⅲ. 互联网络—情报检索 Ⅳ. G354.4

中国版本图书馆 CIP 数据核字(2008)第 081972 号

责任编辑：任 静 王志欣 杨 然 / 责任校对：郭瑞芝

责任印制：刘士平 / 封面设计：耕者设计工作室

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

骏走印刷厂印刷

科学出版社发行 各地新华书店经销

*

2008 年 6 月第 一 版 开本：B5(720×1000)

2008 年 6 月第一次印刷 印张：8 1/4

印数：1—4 000 字数：158 000

定价：30.00 元

(如有印装质量问题，我社负责调换(环伟))

前　言

快速发展的现代互联网在带给人们大量信息的同时，也不可避免地产生了难以让用户快速获取有效信息的问题。作为一种常见的 Web 信息资源检索工具，搜索引擎日益受到人们的关注并得到广泛的使用。它面向任何 Web 用户，无需用户具有较高的专业检索知识，使用方式也较为简单。搜索引擎已经成为人们获取 Web 资源的一种主要方式。

然而，现代搜索引擎也存在很多不足。其中，最为主要的一个问题就是由于采用了全文检索的匹配方法，用户往往会得到相当多的查询结果网页，而用户一般只会访问其中感兴趣的网页，但是很多搜索引擎往往缺乏对用户个性化信息的利用，从而不能实现有针对性的个性化信息服务。实际的情况就表现为即使是具有不同个性化信息需求的用户，在输入相同检索词语的时候也会得到相同的结果，甚至是相同的网页排列次序。这些问题显然需要得到解决。

借鉴在电子商务网站中广泛使用的 Web 个性化信息推荐技术，本书提出了一个较为可行的解决方案，即在搜索引擎中使用 Web 个性化信息推荐技术，以实现个性化搜索引擎。然而，传统的 Web 个性化信息推荐技术具有很多并不适合搜索引擎的特点。只有结合搜索引擎工作的原理和特点，在现有的 Web 个性化信息推荐技术基础上加以改进，才能设计出具有个性化信息推荐能力的智能搜索引擎。

通过对基于个性化信息推荐技术的搜索引擎框架和基本技术的探讨，本书构建了一个完整的基于 Web 个性化信息推荐技术的搜索引擎框架结构。这种框架结构的设计思想主要考虑了两点内容：一是尽量减少用户使用的复杂度，能够让用户在完全无需关注个性化过程的情况下，来表达自己的个性化信息需求和得到所需的个性化信息；二是尽量在现有搜索引擎技术基础上进行优化和完善，无需对现有技术和平台环境做过大的调整。

本书的基本结构如下所示：

第 1 章对相关技术和概念做了介绍，包括 Web 信息检索、Web 挖掘和 Web 个性化信息推荐服务等。

第 2 章对网页权重分析技术从网页质量和网页相关度两个角度做出了说明。其中，前者根据网页质量的定义方法，介绍了不同的网页权重设计模型，而后者则

根据网页相关度的定义方法，介绍了 PageRank 算法及其优化计算等问题。

第 3 章对目前个性化搜索引擎的研究现状和相关概念进行了分析，在给出各种常见搜索引擎个性化形式的同时，也分析了这些技术所存在的主要问题。通过各种方法的比较，本书认为利用用户个性化信息以完善现阶段的个性化网页权重的方法是个不错的搜索引擎个性化方案。

第 4 章主要比较了各种常见的用户模式识别方法，指出适用于当前 Web 环境的用户模式识别方法所应具有的特点，并提出一种利用搜索引擎服务器日志信息得到关键词访问序列的思路，据此设计了基于关键词序列的用户模式识别方法。

同时，本章还对基于关键词序列的用户事务模式的相似度匹配方法做了深入的分析和研究，探索了利用基于关键词外在特征的传统事务模式相似度计算方式、基于用户兴趣度的事务模式相似度计算方式、基于关键词语义信息的事务模式相似度计算公式和基于查询文档语义信息的事务模式相似度计算方式。

第 5 章提出了基于修改网页权重值的个性化 PageRank 算法和基于添加修正参数的个性化 PageRank 方法。针对传统的个性化 PageRank 算法，利用基于关键词序列的用户事务聚类模式和主题化事务聚类模式，改进了个性化 PageRank 算法中用户个性化信息特征的表达方法，并给出相应的计算方法。

第 6 章给出了一个较为完整的个性化搜索引擎系统原型，并对基于关键词序列的用户模式识别方法和相关聚类方法，以及个性化 PageRank 方法的运行效果给出了必要的测试。

目　录

第1章 绪 论

本章主要介绍了与个性化搜索引擎相关的基本技术内容，主要包括Web信息检索、Web挖掘和Web个性化信息推荐服务等。

1.1 Web信息检索

随着互联网技术的快速发展，网络信息资源呈现一种爆炸式的增长态势，比如在2005年，Google搜索引擎能够遍历到的Web网页数量就已经达到近60亿[①]。这些网络信息资源在给人们带来丰富知识和极大便利的同时，也暴露出一些亟待解决的问题。其中，最主要的问题表现在这种信息资源的增长速度远远超出了人们能够处理它们的能力，动辄千万吉的信息量让用户实际上难以获取所需的有效信息，更难以对收集来的海量信息进行分析和获取知识。奈斯比特在《大趋势》一书中准确形容了人们目前所处的困境，即信息是丰富的，而知识是贫乏的[1]。

针对上述特点，人们在传统信息检索系统的基础上开发出新的Web信息检索系统，典型的系统就是搜索引擎。自从1994年问世以来，搜索引擎逐渐成为人们获取网络信息资源的主要方式，相关搜索引擎网站也是Web用户使用网络时首选的访问站点。另外，它和免费电子邮箱、网络实时通信软件构成了当今门户网站用来吸引用户访问的三种主要方式。现在，相关搜索引擎厂商日益成为促进互联网产业发展的重要力量。

1.1.1 Web信息检索模型

广义的信息检索是指信息用户为处理和解决各种问题而查找、识别、获取相关的事实、数据、文献的活动及过程，而狭义的信息检索主要是指信息用户在计算机信息检索系统上进行的信息查询行为。具体的计算机检索行为包含脱机批处理检索、联机检索、光盘检索和网络化联机检索。Web信息检索是一种网络化的联机检索，它的检索对象就是互联网上的Web资源。

从逻辑上看，信息检索涉及三个重要的处理过程：文档集的逻辑表示、用户查询的表示和相似度匹配及排序算法。所以，信息检索模型一般可以表示为一个三

① http://www.google.com

元体的框架，即 D 为文档集中的一组文档逻辑视图（即文档的表示），Q 为一组用户信息需求的逻辑视图表示（即用户查询），$R(Q,D)$ 表示文档与用户查询之间联系的相关度函数，它根据文档信息和用户查询信息输出一个与特定查询 q_i 和文档表示 d_i 有关的实数，以此来表示文档和用户查询之间的相关度[2]。

据此，可以看出传统信息检索模型主要关注于查询和文档相似度的匹配。然而，在现代 Web 网络检索系统中，面向用户的个性化服务逐渐成为主流方式，在这种情况下，可以进一步得到个性化的信息检索模型，即为一个四元体的框架，较前者多个 U，也就是包含个性化信息的用户特征逻辑视图。此时的相关度函数可以概括为 $R(Q,D,U)$，在这种情形下，即使是相同的用户查询和相同的检索集合，但对于具有不同个性化要求的检索用户而言，其命中的结果也是不同的。

按照信息检索模型表达方式的不同，可以将信息检索模型分为布尔模型、向量空间模型、概率模型和逻辑模型等。其中，向量空间模型更加适合网络 Web 文本信息检索的要求，具有简单易行的特点，所以在现代 Web 信息检索服务中被广泛采用。本书主要对向量空间信息检索模型在网络检索系统中的使用方式做出介绍。

1.1.2 向量空间模型

现代 Web 网页信息大多是以 HTML 文档形式存在的，用户与服务器之间的信息传递主要依赖于超文本传输协议 HTTP，然而这种协议仍是基于 HTML 语言来表达网页信息的。所以，HTML 语言组成的信息资源在 Web 网络中广泛存在。

在具体的 HTML 网页文档中，所有的信息都是直接面向显示的。例如，用于定义标题的“<head>”标记、定义元数据信息的“<meta>”标记和负责网页文档具体显示的格式化标记等。这些标记的具体显示效果是由浏览器来负责进行的，而信息内容的理解工作本身则由用户自己来完成。所以，要想构造界面良好的网页文档，除了在形式上美观易用外，在内容上也要符合用户的信息需求。

从结构上看，Web 网页文档通常缺乏一个统一明确的格式，包含的信息内容也是相当庞大，非结构化特征表现得相当明显；然而将 Web 网页文档中存在的每个词语取出来组成一个向量序列，就可以较为简单地实现 Web 网页信息的结构化表达。当然，在抽取词语的时候，一般是将 HTML 语言标记去除，必要的时候还可以结合停用词表来过滤无关词语，以及通过剔除词尾保留词根等方法，来尽可能地消除网页中无关词语对 Web 网页信息分析过程的干扰。

对每篇 Web 网页文档都进行相应的向量化处理，就可以利用得到的不同词语项集来表达不同的 Web 网页文档。如把每个词语看成是一个维度，则所有词语的集合就构成了一个 N 维的文档空间，N 为所有 Web 网页文档中出现的有效词语

数量总和。Web 文档集合中的任何一篇文档都可以表示为这个多维空间中的一个向量，因此称之为文档向量。文档向量在每个词语维度上的取值能反映该词语在该文档中的权重[3,4]。

具体地说，在 Web 向量空间模型中，可以将网页文档看成是一组词语(T_1，T_2，…，T_n)构成的序列，对于每一词语 T_i，都根据其在网页文档中的重要程度赋予一定的权值 W_i。如果将 n 个词语集合看成一个 n 维坐标系，W_i就为某网页文档对应的坐标值，因此每一篇网页文档都可以映射为由一组词语组成的网页文档特征向量，具体表示为((T_1,W_1)，(T_2,W_2)，…，(T_n,W_n))。这种方法的优点在于将非结构化的网页表示为向量形式，使得各种基于结构化的信息处理技术得以使用。

在传统的向量空间模型中，对词语权重的设计较为简单。具体说明如下：首先是利用词频，通常某词语在某文档中出现的频率越高，则对该文档来说，此词语较其他出现频率不高的词语更为重要，词频通常用 TF(term frequency)表示，所以词语权重与词频成正比关系；其次是文档频率，通常越多的文档含有某词语，说明该词语的专指度越差，它对文档的区分度效果越不明显，相对来说，重要程度就越低，所以词语权重与文档频率成反比关系。为了计算方便，在设计权重时，一般都采用逆文档频率(文档频率的倒数)来计算，通常用 IDF(inverse document frequency)表示，此时词语权重与逆文档频率成正比关系。

当然，由于词语的绝对词频直接受到文档本身含有词语数量的影响，所以有必要在具体计算之前进行权重的规范化处理，如最大 TF 规范、对数 TF 规范、轴规范和余弦规范等。其中，常见的方式是余弦规范法，它分别用每个向量的词语权重值除以该文档向量的欧几里得长度，如文档向量为(w_1，…，w_n)，其中，w_i = TF · IDF，则得到的规范化权重公式为

$$w = \frac{w_i}{\sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}} \quad (i \in [1,n]) \tag{1-1}$$

对于两个文档向量的相似度匹配函数，可以采用的方法较多，普遍采用的是余弦夹角值，即对于文档 d_i 和 d_j，它们的向量分别表示为 $V(d_i)$和 $V(d_j)$，$V(d_i) \cdot V(d_j)$表示两个文档向量的点积，$|V(d_i)|$和$|V(d_j)|$分别表示两个文档向量的欧几里得长度，计算文档向量相似度的公式为

$$\mathrm{SIM}(d_i,d_j) = \frac{V(d_i) \cdot V(d_j)}{|V(d_i)||V(d_j)|} \tag{1-2}$$

在现代向量空间模型中，对于词语权重的设计可以从更多的方面来进行考查。除了上述诸如利用布尔值、词频和 TF/IDF 等内容的方法外，还有一些改进的方法，诸如 TFC(term frequency collection)、LTC(length term collection)和信息熵等。另外，一些结合词语与文档的主题相关度方法也取得了不错的效果。这

些都使得词语的权重更能真实地反映文档的内容。然而，这些方法都具有计算复杂的特点。由于在 Web 网络环境下数据量是相当庞大的，所以对网页文档的相似度计算通常也应该考虑性能和效率问题。近几年来，一些不直接利用文档语义分析而利用间接方式获取文档词语与文档主题相关程度的思路逐渐产生。例如，有些方法是利用 Web 文档被命中时所键入的查询词语是否相同来判定文档主题是否相关[5]。

从总体上看，向量空间模型的优点是比较明显的，它能以简单的形式将非结构化的 Web 文档转换成易于进行数学运算的结构化形式。但是它也存在很多问题：首先，它不能有效地处理用户提交的结构化查询，如布尔查询表达式。其次，它只能抽取出文档中的词，但不能反映这些词之间的关联。研究者提出的改进方案有很多，其中 N 元语法信息表示模型不仅能够抽取文档中出现的词，而且可以有效反映词之间的相互关联[6]。最后，它在处理向量相似度的时候，并没有考虑词语之间内在的语义联系，而是单纯地假设词语之间没有任何联系，即认为构成文档空间的词语维度向量是彼此两两正交的，所以在非正交关系的实际情况下，会使得检索效果产生误差。后来提出的广义向量空间检索模型和潜在语义标引(latent semantic indexing)检索模型等方法在这些方面进行了极大的改进。但是从本质上看，由于信息检索只是以文档中自然语言作为基本的处理对象，因此目前这些传统的方式全部建立在一种基于“索引假设”的信息检索模型之上，只能处理非常简单的语言成分。通过引入更深层次的自然语言处理理论和方法来改进网络文本信息的组织和管理模式，有效地对词义加以最大限度地利用，以提高信息检索模型的性能，这被认为是未来最重要的发展方向。

近年来，人们提出了一个同时利用词语和义项进行信息检索的模型，称为“义项矩阵模型”(sense matrix model，SMM)。它利用自然语言中词和义项的复杂关联关系设计出一种新的文档表示，即把文档表示成为一个 term×sense 矩阵，由此也可以把 SMM 看成是向量空间模型(vector space model，VSM)的一个“矩阵式权重”扩充，取得了不错的效果[7]。

虽然说现代互联网上的大量网页文档信息给信息检索带来了巨大的挑战，但也为信息检索研究带来了新的发展机遇。根据 2003 年 SIGIR(Special Interest Group on Information Retrieva)论坛提出的一份报告显示，未来 5～10 年的研究框架主要应围绕一个全球范围的多语言、分布式、个性化的信息访问平台所需的基础理论和应用技术①。其中，个性化的要求在现阶段显得尤为重要，可以通过与 Web 个性化推荐服务进行有效结合来完善 Web 信息检索的要求；同时，这也可能是极为可行的一个解决方案。

① http://www.sigir2003.org

1.1.3 搜索引擎

较一般的信息检索而言，Web 网页和 Web 网络自身的复杂结构使得 Web 信息检索难度更大，这具体表现在三个方面：一是如何快速全面地获取海量 Web 网页数据；二是如何将异构化更加明显的 Web 网页进行信息整序以实现结构化存储；三是用户如何准确地表达自己的查询请求，以便和 Web 信息检索系统进行有效交互。其中个性化信息的表达逐渐成为现代 Web 信息检索技术的研究热点。

经过十来年的发展，现代 Web 信息检索系统已经发展出多种主要形式，其中常见的形式有搜索引擎、Web 资源目录、多元搜索引擎和信息检索代理等。其中，搜索引擎是最为用户广泛使用的一种 Web 信息检索工具，很有代表性。

1. Web 网络及其资源特点

当然，要想准确和深入地了解搜索引擎为什么在现代信息社会扮演着如此重要的角色，那就必须从理解搜索引擎所赖以生存的环境——Web 网络开始。

在现代社会中，人类的信息生产能力已经远远超出了人类对信息本身的处理、组织和吸收能力。英国科学家 Martin 认为，人类的科学知识在 19 世纪是每 50 年增加一倍，20 世纪中期是每 10 年增加一倍，到 20 世纪 70 年代就已经缩短到每 5 年增加一倍。随着 20 世纪 90 年代互联网的出现和发展，极大地促进了人类信息的增长速度。同时，由于 Web 信息的分散、交叉引用的频繁，又使得信息的无序化相当严重，这一点在互联网上表现得很明显。所以，研究和设计有效的 Web 信息检索系统就成为在现阶段有效利用互联网信息来获取所需知识的必要前提。另外，在诸如商业站点、推荐服务系统等很多领域中，信息检索技术也有广泛的应用。它一方面能够有效地增强用户使用的方便程度；另一方面也为其他一些新技术提供必要的技术支持。

应该说，Web 网络是现代搜索引擎得以产生的基础环境，这个网络的特点可以从以下几个方面来看：

首先，它具有动态性。互联网的信息资源是直接面向广大普通用户的，信息发布机制简单易学，不需要用户有过多的专业领域知识，Web 用户就可以很方便地发布信息，所以 Web 网页的更新相当频繁。据发表在《科学》杂志 1999 年 7 月的文章《Web 信息的可访问性》估计，平均每月有 40%的 Web 网页发生变化，Web 网页总数大概以每 4～6 个月翻一番的速度在增长，而且这个数字本身还在提高[8]。这些都说明，Web 信息资源的数量是极其庞大的，它的增长和变化也是相当迅速的，使得人们不可能按照传统的方式构建信息检索系统数据库，以进行完全的网络资源复制，同时造成网页间的链接关系经常改变，以及空链接、错误链接

的出现概率增大。

其次，它具有异构性。随着网页技术的发展，Web 网页可以支持的信息格式越来越多，除了简单的文本信息外，诸如动画、视频等多媒体信息也得到广泛的支持。但是从总体上看，这些多媒体信息缺乏统一明确的结构定义，而且随着 Web 网页本身的结构日渐复杂，大量的嵌套结构（如网格、框架等）增大了对网页的解析处理难度。从总体上看，Web 网页也没有按照任何有关排列次序加以组织，更没有分类索引和其他按标题或者作者等信息的索引。这些都使得 Web 网络呈现出相当明显的非结构化特点，也对基于结构化信息的传统信息检索技术提出了更高的技术要求。

最后，它具有广域分布性。互联网中的服务器分布于世界各地，Web 信息被存放在不同的站点和平台上，而且计算机之间的通信链接变化多端，加之这些数据组织没有统一的整体结构，所以 Web 信息资源具有一种前所未有的广域分布特点，同时，不可避免地产生大量重复信息。据统计，大约有 30%的网页信息是重复的，这一点也增加了信息检索的开销，使得即时查询难以有效实现[9]。

通过以上分析可以看出，Web 资源具有很多和传统信息资源不同的特点，这些特点都使得对 Web 网络信息进行加工处理的难度显著增大。

2. 搜索引擎的发展历史

搜索引擎作为一种伴随着互联网产生而产生、发展而发展的重要技术，它的历史与互联网有密切的联系。一般而言，可以按照搜索引擎对网页索引的技术不同，将其划分为若干阶段。

第一阶段的代表系统是 1994 年出现的 Yahoo 搜索引擎，当时处于搜索引擎的早期阶段，网页资源总量较低，爬虫技术尚不成熟，所以类似于 Yahoo 等搜索引擎广泛地采用人工标引等辅助技术来对网页内容进行标注，同时在检索时，也主要借助传统的信息检索算法，如词语文档频率等。

第二阶段的代表系统是 1998 年出现的 Google 搜索引擎，此时的互联网已经进入快速发展的新阶段，网页数量激增，所以 Google 推出了以网页超链分析为基础的 PageRank 算法，以此来完成网页权重的表示和检索结果网页的相关度排序。这些类似的技术在国内的 Baidu 等中文搜索引擎中也得到广泛应用。

现在正处于搜索引擎发展的第三阶段，代表性系统还没有出现，但是这一时期的搜索引擎广泛地采用了现代信息检索技术的发展成果，在用户接口、网页权重表示和结果排序技术方面都取得了较大的完善和改进。其中，广受重视的就是个性化搜索引擎技术的出现和推广。

1.1.4 搜索引擎工作原理

不同的Web信息检索系统在工作原理上各不一样，下面就结合搜索引擎来具体说明一下一般Web信息检索系统的特点。虽然各个搜索引擎的具体实现不尽相同，但一般包含爬虫程序、分析程序、索引程序、检索程序和用户接口界面5个基本部分，而且大致的工作原理是相同的。Web搜索引擎的基本结构如图1-1所示。

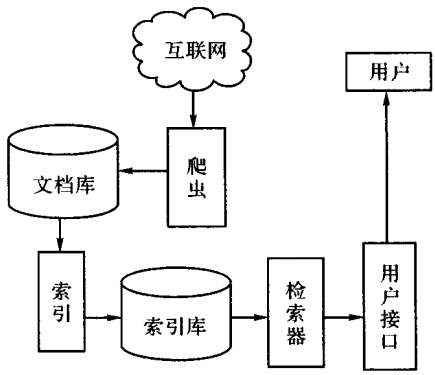

图1-1 Web搜索引擎的基本结构示意图

Web搜索引擎主要是通过爬虫程序定期遍历互联网，将网页的统一资源定位符(URL)、内容和采集时间等相关信息收集到Web服务器上，然后通过必要的信息索引和存储优化处理，利用特定的检索界面对Web用户直接提供服务。这种处理方式在很多方面适应了Web信息的特点。例如，爬虫程序的定期遍历可以将不断动态变化的Web网页信息采集过来，既能有效反映最新的网页信息，又能将分布于各地的Web信息统一存储在搜索引擎的本地服务器上，实现信息资源的本地化，以实现对用户查询的快速响应；同时，搜索引擎提供了基于关键词的全文检索方式，避免了不必要的词语分析和语义处理，适应半结构化网页信息的处理特点，而且还能提高信息的查全率。

它的具体工作流程包括以下几步：

第一步，由爬虫程序采用一定的搜索策略对Web网络进行遍历并下载网页，系统中维护一个超链队列或者堆栈，其中包含一些起始URL；爬虫程序从这些URL出发，下载相应的页面，并从中抽取出新的超链加入到队列或者堆栈中。上述过程不断重复直到堆栈为空。为提高效率，搜索引擎中可能会有多个爬虫程序进程同时遍历不同的Web子空间。为了便于将来扩展服务，爬虫程序应能改变搜索范围和搜索策略，一般采用以宽度优先搜索策略为主、深度优先搜索策略为辅的搜索策略。

第二步，由分析程序对爬虫程序下载的网页进行分析以用于索引，网页分析技术一般包括分词(有些仅从文档某些部分抽词，如Altavista)或者使用停用词表(stop list)来过滤网页信息，同时还提供诸如单复数转换、词缀去除和同义词替换等词语转换，这些技术的具体实现往往与处理方式以及系统的索引模型密切相关。

第三步，索引程序将网页信息表示为一种便于检索的方式并存储在索引数据库中。索引的质量是Web信息检索系统成功的关键因素之一。一个好的索引模

型应该易于实现和维护、检索速度更快、空间需求更低。搜索引擎普遍借鉴了传统信息检索中的索引模型，包括倒排文档、向量空间模型和概率模型等。

第四步，检索程序从索引中找出与用户查询请求相关的网页信息，采用与分析网页文档相似的方法来处理用户查询请求，最后将相关度大于阈值的所有网页按照相关度递减的顺序排列并返还给用户，当然搜索引擎的相关度判断并不一定与用户的需求完全吻合。传统的方式只是利用词频和词语出现的标记和位置来设置权重，新的设置权重方式主要利用基于超链分析的方法，然而只有在系统中引入用户特征模式信息后才能最终为用户提供个性化的信息检索服务。

最后，用户接口为用户提供可视化的查询输入和结果输出界面。在查询界面中，用户按照搜索引擎的查询语法使用检索词语及各种简单、高级的检索条件，构造自己的查询表达式。在输出界面中，搜索引擎将检索结果展现为一个线性的网页列表，其中包含了网页的标题、摘要和相关超链等信息。由于 Web 信息是动态变化的，因此，Robot 分析器和索引器模块要定期更新数据库，时间通常约为一个月。索引数据库越大，更新也越困难。这就使得这种传统的被动服务方式其实不是非常及时有效，借鉴信息推送服务的思想可以极大地提高 Web 信息检索系统的及时性。

1.1.5 相关度排序技术

相关度排序技术的产生主要是由搜索引擎的特点决定的。首先，现代搜索引擎能够访问的 Web 网页数量已经达到上十亿的规模，哪怕用户只是搜索其中很少的一部分内容，基于全文搜索技术的搜索引擎也能返回成千上万的页面。即便这些结果网页都是用户所需要的，用户也没有可能对所有的网页浏览一遍，所以能够将用户最感兴趣的结果网页放于前面，势必可以增强搜索引擎用户的满意度。其次，搜索引擎用户自身的检索专业能力通常很有限，在最为普遍的关键词检索行为中，用户一般只是键入几个词语。例如，Spink 等曾对 Excite 等搜索引擎的近 300 位用户做过实验调查，发现人均输入的检索词为 3.34 个。国内部分学者也有相似的结论，发现 90%左右的用户输入的中文检索单字为 2～6 个，而且 2 字词居多，约占 58%，其次为 4 字词(约占 18%)和 3 字词(约占 14%)。过少的检索词事实上无法真正表达用户的检索需求，而且用户通常也不去进行复杂的逻辑构造，只有相当少的用户进行布尔逻辑检索、限制性检索和高级检索等方法，仅有 5.24%的检索式中包含有布尔逻辑算符。国内的部分学者的研究结果也表明，约 40%的用户不能正确运用字段检索或二次检索，80%左右的用户不能正确运用高级检索功能，甚至还发现用户缺乏动力去学习复杂的检索技能，多数用户都寄希望于搜索引擎能够自动地为他们构造有效的检索式[9]。由于缺乏过去联机检索中常常具备的检索人员，因此，用户实际的检索行为与用户理想的检索行为存在

事实上的差距，检索结果的不满意也是不奇怪的。正是由于这个特点，搜索引擎就必须设法将用户最想要的网页结果尽可能地放到网页结果的前面，这就是网页相关度排序算法在搜索引擎中为什么非常重要的原因。

现阶段的相关度排序技术主要有以下几种：一是基于传统信息检索技术的方式，它主要利用关键词本身在文档中的重要程度来对文档与用户查询要求的相关度做出测量，如利用网页中关键词出现的频率和位置。一般而言，检索出的网页文档中含有的查询关键词个数越多，相关性越大，并且此关键词的区分度越高；同时，查询关键词如果出现在诸如题名字段等重要位置上，则比出现在正文的相关度要大。二是超链分析技术，使用此技术的代表性搜索引擎有 Google 和 Baidu 等。和前者相比，它以网页被认可的重要程度作为检索结果的相关度排序依据。从设计思想上看，它更注重第三方对该网页的认可，如具有较大链入网页数的网页才是得到广泛认可的重要网页，而根据关键词位置和频率的传统方法只是一种网页自我认可的形式，缺乏客观性。最后还有一些其他方式，如由用户自由定义排序规则的自定义方式。北京大学的天网 FTP 搜索引擎就采用这种排序方式，它可以让用户选择诸如时间、大小、稳定性和距离等具体排序指标来对结果网页进行相关度排序①。再如收费排名模式，它作为搜索引擎的一种主要赢利手段，在具有网络门户特点的大型搜索引擎中广为使用，但由于担心影响搜索结果的客观性，这种方式不是它们的主流排序方式，而仅仅作为一个补充显示在付费搜索栏目中。

相关度排序技术主要依赖于超链分析技术实现。超链分析技术可以提供多种功能，其中的主要功能就是解决结果网页的相关度排序问题。它主要是利用网页间存在的各种超链指向，对网页之间的引用关系进行分析，依据网页链入数的多少计算该网页的重要度权值。一般认为，如果 A 网页有超链指向 B 网页，相当于 A 网页投了 B 网页一票，即 A 认可了 B 网页的重要性。深入理解超链分析算法，可以根据链接结构把整个 Web 网页文档集看成一个有向的拓扑图，其中每个网页都构成图中的一个结点，网页之间的链接就构成了结点间的有向边，按照这个思想，可以根据每个结点的出度和入度来评价网页的重要性。

对于超链分析技术，有代表性的算法主要是 Page 等设计的 PageRank 算法和 Kleinberg 创造的 HITS 算法。其中，PageRank 算法在实际使用中的效果要好于 HITS 算法，这主要是由于以下原因：首先，PageRank 算法可以一次性、脱机且独立于查询的对网页进行预计算以得到网页重要度的估计值，然后在具体的用户查询中，结合其他查询指标值，一起对查询结果进行相关性排序，从而节省了系统查询时的运算开销；其次，PageRank 算法是利用整个网页集合进行计算的，不像

① http://e.pku.edu.cn

HITS算法易受到局部连接陷阱的影响而产生“主题漂移”现象，所以现在这种技术广泛地应用在许多信息检索系统和网络搜索引擎中，Google搜索引擎的广获成功也表明了以超链分析为特征的网页相关度排序算法日益成熟。

PageRank技术基于一种假设，即对于Web中的一个网页 u，如果存在指向网页 v 的超链，则可以将 v 看成是一个重要的网页。PageRank认为网页的链入超链数可以反映该网页的重要程度，但是由于现实中的人们在设计网页的各种超链时往往并不严格，有很多网页的超链纯粹是为了诸如网页导航、商业推荐等目的而制作的，显然这类网页对于它所指向网页的重要度贡献程度并不高。但是，由于算法的复杂性，PageRank没有过多考虑网页超链内容对网页重要度的影响，只是使用了两个相对简单的方法：其一，如果一个网页的链出网页数太多，则它对每个链出网页重要度的认可能力相应降低；其二，如果一个网页由于本身链入网页数很低造成它的重要度降低，则它对链出网页重要度的影响也相应降低。所以，在实际计算中，网页 v 的重要性权值正比于链入网页 u 的重要性权值，并且和链入网页 u 的链出网页数量成反比。由于开始无法知道网页 u 自身的重要性权值，所以决定每个网页的重要权值需要反复迭代地进行运算才能得到。也就是说，一个网页的重要性决定着同时也依赖于其他网页的重要性。

这种情况也可以从Web网页图结构 G 的随机漫步理论来解释。设网页 u 指向网页 v 表示图 G 中在节点 u 和 v 之间存在一个边，并且设 $\deg(u)$ 表示网页 u 的链出网页数量。对于任意时间 k 上被随机冲浪者选中的网页 u，在下一个冲浪步骤中，冲浪者从网页 u 的链出网页集合中平均地随机选择了一个节点 v_i，也就是说，在时间 $k+1$ 时，冲浪者访问某个网页 u 链出网页的可能性为 $1/\deg(u)$。

对于用户的偏好网页集 P，可以概括为一个偏好向量 U，其中 $|U|$ 等于1，$U(p)$ 代表了对其中的网页 p 的偏好程度。比如，如果用户想平均地设定偏好网页集合中每个网页的偏好程度，可以在网页 p 属于 P 集合的时候，令 $U(p)=1/|P|$，否则为0。接下来，对于个性化PageRank值的计算可以使用矩阵向量方程式来进行。设 A 为整个网页网络 G 的对应矩阵，其中的 A_{ij} 代表着网页 p_i 和网页 p_j 的链接关系，如果这个链接存在，则 A_{ij} 为 $1/|O(j)|$，否则为0。这里的 $|O(j)|$ 表示网页 P_i 链出节点的个数。为了简化问题，可以假设所有的网页至少都有一个链出网页，即对于没有链出网页的网页可以设置一个指向网页本身的链接。利用具体的 u 向量，个性化的PageRank值可以计算为：$v=(1-c)Av+cu$，其中 c 就是一个跳转因子，一般约为0.15，试验也证明这个因子的微小变化其实并不会对最终的结论产生很大的影响[10]。其中，最初的 v 可以利用一个常量向量来表达，最终收敛的 v 向量可以代表一个较为稳定的值，代表着随机冲浪产生的访问网页分布情况。按照马尔可夫算法理论，只要满足所需条件，最终收敛的 v 向量总是存在并唯一的，其实质是一个非周期性的遍历马尔可夫链，并且这个 v 向量就是