

# 个性化搜索引擎技术探讨

魏小梅

(华中农业大学 理学院,湖北 武汉 430070)

**摘要:**介绍了搜索引擎的技术基础,并对个性化的搜索引擎技术进行了探讨,同时讨论了在个性化搜索引擎技术中建立用户兴趣模型的方法,给搜索引擎的技术研究提供了参考。

**关键词:**个性化;搜索引擎;兴趣模型;网页分级

中图分类号:TP393.4

文献标识码:A

文章编号:1672-6251(2006)04-0055-02

## Research on personalized search engine

Wei Xiao-mei

(Science college, HuaZhong Agricultureal University, Wuhan 430070, China)

**Abstract:** It introduces the basic technology and personalization methods of search engine in this paper. At the same time, it discusses the methods to get user's interest profile. This paper is useful to study the search engine technology.

**Key words:** Personalization; Search engine; Interest profile; Page rank

## 1 引言

Internet 上网页的爆炸式地增长,虽然通用搜索引擎给人们提供了信息检索的手段,但是随着信息越来越多,搜索引擎返回给用户的信息量也越来越大,其中跟用户无关的垃圾信息也越来越多。如何根据不同的用户兴趣过滤掉不相关的信息,从而返回给用户最有用的信息,就是个性化查询研究的内容。到目前为止,已经提出了许多技术来解决 Web 搜索个性化的问题。

## 2 搜索技术的概述

搜索引擎的基本原理,主要部分可以看作三步:从互联网上抓取网页→建立索引数据库→在索引数据库中搜索排序。一般说来搜索引擎由搜索器、分析器、索引器、索引数据库、检索器和用户接口组成<sup>[1]</sup>。从互联网上抓取网页是由搜索器实现的,搜索器主要就是蜘蛛程序,它以广度优先或深度优先的方法进行 Web 检索并下载相关页面。它的实现方法是沿着任何网页中的所有 URL 爬到其它网页,重复这过程,并把爬过的所有网页收集回来。

分析器对获得的检索结果进行分析,分析收集回来的网页以用于索引,具体包括分词、过滤、转换等工作;

提取相关网页信息,将文档以便于检索的方式存储在索引数据库中,一般采用的方法有矢量空间模型(Vector Space Model)、倒排文档、概率模型等。

当用户输入查询关键词后,搜索系统程序从网页索引数据库中搜索符合该关键词的所有相关网页;将搜索到的网页根据相对查询关键词的网页重要性进行排序,重要性越高,排名越靠前;最后,由页面生成系统将搜索结果的链接地址和页面内容摘要等内容组织起来返回给用户。

## 3 个性化搜索技术

### 3.1 个性化查询服务基本流程

个性化查询服务就是在查询服务的基础上,添加了根据用户偏好,对分级和排序的过程进行影响。个性化查询服务步骤如下:

(1) 用户提交查询条件;

(2) 系统根据检索条件在检索端数据库中进行检索;

(3) 结合用户特性,体现用户的个性。对第(2)步的检索结果考虑进用户兴趣进行重新分级(Re-Ranking),重新排序;重新分级的过程中过滤(Filtering)掉不相关的结果,使相关的结果排在更靠前的位置;

(4) 返回结果,记录用户对结果的反馈信息,为用户兴趣模型的修改提供依据,从而为后面的搜索做准备。

### 3.2 个性化搜索引擎的方法

个性化搜索引擎就是在对搜索结果重新分级的时候,考虑进用户的偏好信息。因此,获得用户的兴趣模型并将其整合进搜索引擎是个性化搜索引擎研究的核心内容。下面讨论几种在搜索引擎中整合用户兴趣的方法:个性化网页权重;查询过滤;个性化元数据搜索系统。

3.2.1 个性化网页权重 除了传统的文本匹配技术,现在的 Web 搜索引擎也按网页的重要性来对结果分类。最著名的例子就是 PageRank 算法,它也是 Google 搜索引擎的基础<sup>[2]</sup>。应用 Web 链接结构,PageRank 给每个网页计算权重,其权重计算公式如下:

$$PR(A)=(1-d)+d(PR(T_1)/C(T_1)+\dots+PR((T_n)/C(T_n)))$$

其中  $PR(A)$ : 页面  $A$  的网页级别;  $PR(T_i)$ : 页面  $T_i$  的网页级别, 页面  $T_i$  链向页面  $A$ ;  $C(T_i)$ : 页面  $T_i$  链出的链接数量;  $d$ : 阻尼系数,取值在 0-1 之间,通常取值为 0.85。

该随机冲浪模型的可用向量表示<sup>[3]</sup>: 设  $A$  为一个方阵,行和列对应网页集的网页。如果网页  $i$  有指向网页  $j$  的一个链接,则  $A_{ij}=1/N_i$ ,  $N_i$  为网页  $i$  的链出网页数,否则  $A_{ij}=0$ 。设  $V$  是对应网页集的一个向量,有  $V=cAV+(1-c)E$ ,  $V$  为  $A$  的特征根为  $c$  的特征向量,  $E$  为引入的阻尼系数,它对应 Rank 的初始值。

在最终的显示给用户的搜索结果中,权重越大的网页越靠前,因此,修正基于用户偏好的权重算式,根据 PageRank 算法能得出个性化的搜索引擎算法。

通常,个性化网页权重积分的计算基于用户定义的一个偏好网页集,那些链接到兴趣网页和被兴趣网页链接的网页应该有更高的网页权重,更可能接近用户的搜索意图。在前面的随机冲浪模型中考虑偏好网页:

用  $P$  表示兴趣网页集,兴趣向量为  $U$ ,  $|U|=1$ ,  $U(p)$  表示对网页  $p$  的偏好度,若  $p \in P$ , 则  $U(p)=\frac{1}{|P|}$ , 否则,  $U(p)=0$ 。对于给定的  $U$ , 个性化网页分级等式可以写成:  $V=cAV+(1-c)U$ 。

3.2.2 查询改进 该方法不修改搜索引擎的算法,而是进行查询改进。通常,该过程包括三步:

#### (1) 根据用户行为建立用户兴趣模型

用户兴趣可以由用户显式输入或者从用户的搜索行为中隐式学习到,在 3.3 中详细讨论。

#### (2) 查询修正

首先系统基于相关用户偏好模型调整输入查询。

然后,修正后的查询交给搜索引擎,然后搜索引擎根据该查询进行搜索。

### (3) 结果修正

从搜索引擎得到查询结果后,系统会对该反馈信息进行提炼。有些系统会根据用户兴趣对搜索再分级,过滤掉不相关的网页,最终返回的网页就是更接近用户意图的少量网页。

3.2.3 个性化元搜索系统 随着 Web 网容量的增大,搜索引擎的覆盖率实际是减少的<sup>[4]</sup>。一个搜索引擎只能获得很低的搜索率,为了解决这个问题,人们提出元数据搜索系统,就是将几个搜索系统整合在一起增加查找的覆盖率。

理想情况是,从多个搜索引擎得到不同网页级别的搜索结果,然后将它们整合,可以得到一个最终的网页级别表,从而通过元数据搜索引擎提高搜索效率。但是,既然元搜索引擎提高了覆盖率(查全率),但是信息量增大了,实际并没有办法提高查询的精确度。为了提高返回结果的精确度,人们又提出了将用户的偏好加到元搜索系统中。

个性化元搜索系统是采用查询改进的方法。通常,这些系统根据相关的用户模型修改输入查询,通过查询统一的用户界面提交给各个独立搜索引擎后,各个搜索引擎根据自己的分级算法返回结果,然后系统根据用户的意图将这些查询结果再做一次重新分级的处理。

### 3.3 个性化用户兴趣模型

3.3.1 用户兴趣模型 在个性化网络搜索引擎的过程中,需要参考用户兴趣模型对搜索结果进行过滤。所以,用户兴趣模型可以用在搜索引擎中提高搜索引擎的执行效率。

用户模型有助于确定查询关键词的意义。例如,不同的用户对相同的查询关键词的查询目标不一定相同,这是要借助用户兴趣来确定重要网页。比如,提交查询关键词“苹果”,用户究竟对“苹果电脑”感兴趣还是对“苹果”这种水果感兴趣,需要用户模型去确定。

用户兴趣模型可以用来进行查询扩展。当模型和用户的查询关键词密切相关时,模型中的关键词可以加到查询中形成更长的查询。众所周知,查询关键词越长,搜索到的结果与查询的匹配精度越精确。

用户模型可以用来过滤初始查询结果。当搜索引擎返回查询结果后,这些结果是基于查询关键词,而不是基于用户兴趣模型,所以将这些结果与用户兴趣模型比较,过滤掉一些网页,得到的结果将更精确。

(下转第 62 页)

### 3.2 自主研发

目前,国内正式运营的140余款网络游戏中,韩国游戏有73款,占据了半壁江山<sup>[1]</sup>。经过几年的经营,中国大陆出品的网络游戏所占份额已经上升到27%,然而,这显然是不够的。由于缺乏自主研发、自有知识产权的网络游戏产品,版权问题制约,在技术支持、升级维护、周边产品的开发营销等方面受制于人,不仅如此,由于游戏开发可以溶入文化传统和价值观,无法掌握开发游戏的优势,还会造成中国传统文化和价值观对青少年一带的影响减弱。

因此,国家鼓励有条件的企业进入游戏开发行业,抓紧自有知识产权网络游戏产品的研发,结合中国的社会体制、意识形态、文化传统和价值观,开发出符合中国国情的产品,以此推动整个民族网络游戏产业的发展。然而,在鼓励运营商进入游戏开发的时候要慎重,如前述,开发与运营的进入壁垒以及成本模式不尽相同,贸然鼓励实力较弱的运营商进入开发行业,容易导致其资金回笼缓慢,进而造成企业发展困难。

### 3.3 行业工会

(上接第56页)

多数人浏览过的网页可能是重要网页,以此为据可以形成推送系统。有些网页如果被许多人浏览过,也就是说,该查询对许多人是有用的,那么对一个新查询来说,这些网页就应该具有高的网页权重,这就是推送系统的基本原理。

3.3.2 获取用户兴趣模型的方法 如何获得用户的兴趣,目前归纳起来有两种方法:一种显式的方法,一种是隐式方法。前一种有些是用户选择一系列的目录来确定其兴趣,或者由用户自己输入一些兴趣词来确定其兴趣,所有的这些方法都需要用户显式地输入或者选择一些信息。但是,有时候并不是每个用户都有足够的耐心和时间来输入相关信息或者选择兴趣方案,所以第二种有其优势,该方法就是从用户搜索的历史记录来学习用户的兴趣,建立兴趣模型,并且每次记录用户的搜索行为,以便对兴趣模型进行更新。该方法具体实现通常是根据用户的浏览历史建立兴趣目录,或者由用户输入搜索关键词,然后在列出的搜索结果中点击需要的网页,搜索引擎根据用户点击网页的行为得到跟兴趣相关的特征向量,然后根据兴趣向量对搜索引擎的搜索结果进行再分级。

用户的搜索行为包括点击某些网页和不点击某些网页,或者在某些网页上停留时间的长短,还有对网页有保存或打印的行为,都可以说明用户对该页感兴趣。机器识别了这些网页并从中提取网页关键词和用户查

尽管网络游戏在中国只有短短的五年的历史,然而,一方面,其面对着巨大的现实和潜在消费市场;另一方面,它还面临着国外同行者以及“私服”、“盗版”等的激烈竞争;其本身又存在诸多不规范因素,因此,建立行业工会势在必行。行业工会一旦建立,就有利于合理的划分市场,便于业者建立行业公约,规范业者行为,也有助于工会成员共同进退,应对外来竞争。

#### 参考文献

- [1] 上海艾瑞市场咨询有限公司.2004年中国网络游戏行业研究报告[R].上海:上海艾瑞市场咨询有限公司,2005.
- [2] IDC中国.2004年度中国游戏产业报告[R].北京:IDC中国,2005.
- [3] 中国互联网络信息中心.中国互联网络发展状况统计报告(2005/1)[R].北京:中国互联网络信息中心,2005.
- [4] 刘瑾.中韩网络游戏产业比较[J].中国电子商务,2005,(8).
- [5] 张瑞良,彭蕾.我国网络游戏产业的现状与对策[J].贵州工业大学学报(社会科学版),2005,(2).
- [6] 苏东水.产业经济学[M].北京:高等教育出版社,1997.
- [7] 迈克尔·波特.竞争战略[M].北京:华夏出版社,2000.

询关键词构成用户兴趣模型。当然不同的时段用户的兴趣会不同,所以兴趣模型会不断更新。

### 4 小结

网络搜索引擎技术经过多年的发展已经给人们提供了快捷的信息搜索方式,成熟的服务模式,它和Web一起改变了人们的生活,改变了人们利用信息的方式。但是随着信息越来越多,搜索引擎返回给用户的信息量也越来越大,其中跟用户无关的垃圾信息也越来越多。如果能根据不同的用户兴趣过滤掉不相关的信息,从而返回给用户最有用的信息,将能大大提高搜索引擎的搜索效率,帮助人们快捷地从庞大的互联网上找到相关的信息,这就是个性化搜索引擎的思想。本文对个性化搜索引擎技术进行了探讨,为进一步进行个性化搜索引擎技术研究提供了良好的基础。

#### 参考文献

- [1] 孟小峰,曹巍.Web查询语言分析与比较[N].计算机世界,2000-04-17.
- [2] OurSearch:Google Technology [URL]. <http://www.google.com/technology/>,2005
- [3] 朱炜,王超,李俊,潘金贵.Web超链分析算法纵览[URL]. <http://www.21ec.net.cn/2005/6-8/133818.html>,2005.12.
- [4] Lawrence,S., Giles,C.L.: Accessibility of Information on the Web[J]. Nature, Vol 400(1999) 107~109.