《现代图书情报技术》 1998 年 第 1 期

Internet 搜索引擎 A ltaV ista 的研究

郑 菲

(中国科学院文献情报中心 北京 100080)

【摘要】 介绍了一个 Internet 信息资源的搜索引擎 A ltaV ista 的产生背景和技术特点, 并研究了它的查询方法与技巧。

【关键词】 Internet A ltaV ista 信息查询 搜索引擎

Ths Study of Search Engine on the Internet——AltaVista

Zheng Fei

(*The Documentation & Information Center of the Chinese Academy of Sciences*)

【Abstract】 The article introduces the growth background and technical characters of A lta V ista, which is a powerful search engine to search information on the Internet. And it studies A ltaV ista's searching methods and skills.

A ltaV ista 是 Internet 上WWW 的重要导航工具之一, 是由DEC 公司于 1995 年正式推出的, 目前它收集了 3000 万个网址(W eb)、14000 个新闻组(new sgroups)的全文数据, 每天访问它的次数超过 1800 万次。A ltaV ista 已成为 Internet 上最热门的搜索引擎。

1 A ltaV ista 的产生背景

众所周知, Internet 网上的信息浩如烟海, 要想从中快速、准确、全面地查找出所需信息是很难实现的。1995 年, DEC 公司在美国加州 Palo A lto 的研究室中开发出了 A ltaV ista——能够对整个 Internet 资源进行索引的工具, 它以难以致信的查询速度解决了世界上最大信息资源 Internet 网的检索难题。利用 A ltaV ista 可以查到任何一个在WWW 或U senet(新闻组)中出现的词, A ltaV ista 已成为一个非常有效的查找信息的工具, 并且具有广阔的商业应用前景。

A ltaV ista 全部W eb 的索引总容量有 60GB, 在强大的软件和 Internet 网及DEC 的A lpha 技术支持下, 可以在任何地方的任何一台微机上快速地查询 Internet 网上的信息。

2 A ltaV ista 的特点

A ltaV ista 被誉为 Internet 网上综合性的、功能最强的查询引擎, 它具有如下特点:

1) 搜索速度快, 提交检索式后可在几秒钟之内得到结果;

2) 用自然语言进行查询, 操作简单;

3) 查询方式灵活, 既可采用简单查询方式, 又可采用高级查询方式, 用布尔逻辑关系将词和短句进行组配;

4) 可查询 25 种语言的信息;

5) 查询结果全面、丰富, 不受学科限制;

6) 有多种检索点, 方便查询, 如 title, image, link 等;

7) 对检索结果进行排序, 将最为相关的信息放在前面;

8) 有五个国际站点, 分别位于亚洲、澳洲、拉美、南欧和北欧, 任何一个站点的查询功能均与在美国的主站点相同, 并且在本地站点还可提供当地语言的帮助和信息。

3 联接方法

在任何一台联入 Internet 网的计算机上, 用 Netscapes 导航软件键入地址 http: //www. altavista. digital. com /, 即可进入A ltaV ista 的主页。

4 查询方法

A ltaV ista 的查询方式分为两种: 简单查询方式和

收稿日期: 1997- 08- 22

高级查询方式。简单查询通过输一个或几个关键词后提交查询，如图 1；高级查询方式是用布尔逻辑算符将两个以上的词组成检索式，检索式中可用括弧，如图 2。

图 1　AltaVista 简单查询方式界面

图 2　AltaVista 高级查询方式界面

4.1　基本检索方法

4.1.1　逻辑算符

简单查询：前一项和后一项为“与”关系，用+；前一项和后一项为“非”关系，用-。

高级查询：是一种结构式的操作和公式表达的方式，用布尔逻辑算符将两个以上的词组成检索式，逻辑算符如下：

逻辑算符	符号表达	举　例	内　容　说　明
AND	&	red AND blue	表示命中结果中同时含有 red 和 blue 两个词
OR	|	red OR blue	表示命中结果中至少含有其中一个词
AND NOT	!	red AND NOT blue	表示命中结果中含有前一个词 red 而除去第二词 blue
NEAR	～	red NEAR blue	表示命中结果中两个词之间可间隔 10 个以内的词，且词序可变

逻辑算符执行顺序是：NEAR、ANDNOT、AND、OR。

查询时逻辑算符用大、小写表示均可，也可用符号表示。如需把逻辑算符作为检索词时，可将它用引号标引起来，如：Portland AND (Oregon OR“OR”)。

4.1.2　关键词

在 AltaVista 的简单查询和高级查询中均支持关键词检索，并可将其限定在 Web 页或新闻组的特殊字段中。如简单查询方式，可直接敲入提问式，title: king；高级查询方式通常用逻辑表达式。在 AltaVista 中，可以把任意一个单词、短语或一句话作为检索条件，但字符(如 &，%，$，/，#，_，～)或空格等均不能作为检索对象，因为它们在词典中没有明确的拼写。当单词间夹有各种符号、数字时，检索时将被认为是两个词的结构，如：don't，x-y，AT&T，U.S.。

4.1.3　截词

用 * 记号作为截词形式，可以查到一组类似的词。比如，查 sing，singers，singing 等一系列词时，提问式可记为 sing *。但有时也要注意，AltaVista 同时还会查到一些不相关的词，如用以上提问式，还会查出 singe，single，singular，和外来词如法语 singulier。

使用 * 记号时要注意以下几点：

(1) 只能在有三个字母的词干后使用 * 号；

(2) 当在小写字母后加 * 进行查询时，* 可查出附加零至五个字母相应的检索词；

(3) 当用大写字母或数字作为检索词时，查询结果为零，即用大写字母或数字进行查询不能用 * 记号。

(4) * 号插入的位置可以在词尾，也可在词的中间，要根据检索词的结构加以分析来确定。例如，查找 color 和 colour 时，提问式 col * r 就不是最理想的，查到的结果可能会有 collector 和原子钴 collider。最有效的提问式是 colo * r，这样就可同时查到 color 和 colour。

4.1.4　大写字母

当查询一个词时，通常是用其小写形式，因为小写字母可表示任何匹配形式。如果用大写字母，那么只能查出与查询条件完全匹配的词。

因此，当用小写形式查询 turkey 一词，可能查到任何一种在文献中出现的形式：turkey，Turkey，或 TURKEY。但用大写形式 Turkey 查询，只能查出含有 Turkey 的文献。

4.1.5　短语

短语是一个文件中连接词的串，它们可能会被一些空格或标点分开。例如：

□President of the U.S.A.　(6 个词串)

□http://www.election.digital.com (5 个词串)

由于标点和空格在AltaVista中不作为检索对象(它们是非限定词),因此以上词组与下面的写法是无区别的:

□President of the USA

□http www election digital com

在查短语或固定配搭时,最好去掉短语中所有的标点符号,并用引号将词标明出来,如短语"a sequence of words separated by spaces and surrounded by double quotes"。

通常标点符号&|!和~在高级查询方式中有自己的含义,而*号可同时用于简单和高级两种查询方式。

4.2 检索式的限定

4.2.1 字段的限定:

将关键词限定在link, title, image,等字段中,后面紧跟冒号,且必须用小写。查询的内容可分别限定在Web或USENET中。

检索Web页时:

title:"The Wall Street Journal"

查到的匹配页题目中含有The WallStreet Journal。

其它类似限定: **url, text, anchor, applet, link, image, host, domain**

在Usenet新闻组中构造检索式:

from: napoleon@elba.com

匹配新闻的**From**:项中含有napoleong@elba.com。

其它类似检索项: **subject, newsgroups, summary, keyword**

4.2.2 语言限定

AltaVista可以查询25种不同语言的信息,其中包括英语、日语、汉语、德语等,但是查询时不能用含有双字节的字符(如汉语、日语、朝鲜语等)进行查找。如查找汉语的信息,可以用英语查找,并将结果限定在汉语中。

4.2.3 时间限定

在高级查询方式中可以查询特定时间段内的信息。时间限制的方式为dd/mmm/yy。dd表示日期;mmm表示月份的英文缩写;yy表示年份的后两位数字。如:09/Jan/96。如果省略年份,AltaVista即确定为当年。

4.3 修改策略指令(Refine)

在AltaVista的查询主页中有一个Refine键。当查看检索结果时,可能会发现有一些不相关的内容,只有修改提问式,再按Search键重新开始查询。但此时若使用Refine重新查询,查到的结果可能会更接近你所需的内容。

4.4 结果排序(Ranking)

在简单查询方式中,查到的结果即为已经排序的结果。比如查询结果为200条,一般来说,很可能在结果的第一或第二页就能发现你最需要的信息。如果在前几页没有发现,修改提问式,然后执行Refine。

在高级查询方式中,有两种结果排序。一般查询后的结果是未经排序的,只有在Ranking一栏中按顺序依次键入关键词时,查询的结果就按照关键词出现次数进行排序。

5 简单查询与高级查询方式比较

当访问AltaVista查询主页时,将直接进入简单查询方式,90%以上的用户首先采用此方式进行查询。采用简单查询方式的优点是:操作简单;用自然语言进行查询,如查询"what is the weather";查询结果进行自动排序,最相关的结果排在前面。

选择高级查询方式的优点是:可利用逻辑算符进行组配;进行时间限定;可根据自己的意途进行结果排序;可以用Near将两个词限定在一定范围内(10个单词以内)。

一般情况下,先采用简单查询方式进行搜索,因为其操作简单易学;在需要进行较复杂的逻辑组配时,再采用高级查询方式,并可查看帮助信息。

6 结 语

经过本人的多方面实践认为,在众多的搜索引擎中,AltaVista是一个较出色的查询工具,它查询速度快,响应时间短,且内容丰富,涉及主题广泛,特别是当你不知道所查的信息属于哪一类时,使用AltaVista更为方便。但是它也存在着缺点,明显一点就是查询的结果有时令人感到价值不大。这是由于它是一个网上公共的搜索引擎,任何人都可将任何信息提供给它作为索引,因此它搜索的信息特别的繁杂。总之,当你需要系统或详细的查找某方面信息时,可先利用AltaVista查得一些线索,比如得到数据库主页的网址或某个公司的网址,进而再进入具体的数据库或公司的主页中进行查询。

参考文献

1 http://www.altavista.digital.com/