搜索引擎输入方式的研究*)

Research on the Query Method of Search Engine

周 晋 路海明 李衍达

(清华大学自动化系 北京 100084)

Abstract Search engines has become the most important tools to help people find information among the rapidly increasing World Wide Web pages, but low quality of results of search engines can not satisfy users. To improve the quality, the query method of search engine should be improved firstly. In this paper, we analyse and compare recent query methods of search engines, and lastly point out several ways to improve the query method.

Keywords Query method, Search engine, Keywords, Information retrieval

1 引言

随着 Internet 的飞速发展,网络上信息量呈几何级数增长,截至 2001 年以前,网上约有数十亿网页。因特网上的信息是极其无序的,因此,信息量越大,越难被利用。如何获取和利用因特网上的信息就成了一个重要问题。目前解决这一问题的最佳途径便是利用搜索引擎。

2001 年,Roper Starch 的调查指出,36%的互联网用户一个星期花了超过 2 个小时时间在网上搜索;71%的用户在使用搜索引擎的时候遇到过麻烦;86%的互联网用户感到应当出现更有效的、准确的信息搜索技术。

从调查数据中不难看出,搜索引擎的检索质量亟待提高。提高搜索引擎的检索质量要从多方面入手,其中,用户需求的输入是搜索引擎工作开始的第一步,是提高检索质量的重要突破口。

在以前的搜索引擎的研究中,搜索服务提供者在搜索技术方面已经花费了大量的时间和精力,由于人工智能技术水平有限,往往收效甚微。而被人们忽视了的输入环节,还只停留在最简单的关键词输入方式上。

实际上,输入方式往往决定了整个搜索引擎系统的索引方式、检索方式,甚至体系结构。

输入环节所要完成的任务是:将用户需求准确无误地传递给计算机。但由于自然语言理解和人机交互技术不发达、用户对搜索引擎不熟悉、用户文化水平参差不齐以及个性化等问题的存在,导致搜索引擎不能准确地理解用户需求,而建立在偏颇或错误的理解之上的搜索结果自然也不尽如人意。

抛弃自然语言理解技术不发达、用户对搜索引擎不熟悉等客观因素,我们依然可以在挖掘用户认知能力、设置个性化环节等方面改进输入方式,从而提高检索质量(例如:应用了关键词调整输入方式的 Inquirus 2[1]),所以,改进输入方式是提高搜索引擎检索质量的重要手段之一。

本文以提高搜索引擎的检索质量为基本点,具体介绍了目前存在的几种主要输入方式,通过分析和比较各自的优、缺点,给出了改进输入方式的几种可行途径。

2 输入方式的介绍

目前,搜索引擎采取的输入方式主要有以下几种:关键词、分类目录、浏览模型[2]、关键词重构、关键词调整、文章集输入方式,详见表 1。

表 1 已有的搜索引擎输入方式

输入方式	英文名称	典型系统
关键词	Keywords	Infoseek 大多数搜索引擎采用的输入方式
分类目录	Directory	Yahoo!
浏览模型	Browsing Modal	
关键词重构	Query Reformulation	HiB[3]
关键词调整	Query Modification 或 Query Expansion	Inquirus 2
文章集输入方式	Documents Set	

2.1 关键词输入方式

关键词输入方式是目前最常用的输入方式,绝大多数搜索引擎采用了这种输入方式。在关键词输入方式中,用户只需输入描述自己需求的一个或多个关键词,系统便可根据这些关键词检索出结果。这种输入方式简单、直观,便于数据库的检索。

但随着用户对信息检索质量需求的日益提高,关键词输入方式暴露出越来越多的弊端,具体表现如下:

①关键词语义空间过大,容易发生歧义,检索质量不高;

②某些需求难以表达;

③常常导致搜不到结果或结果过多的现象;

④难以体现用户个性化特征;

⑤无法检索视、音频等多媒体元素。

这些弊端严重地阻碍了搜索引擎进一步提高检索质量,设计出更先进的输入方式,已成为迫在眉睫的事情。

2.2 分类目录输入方式

由于机器检索的不准确性,人们考虑将人的分类能力融入到搜索引擎中,分类目录就是基于这种思想工作的。在分类目录搜索引擎中,首先,信息资源按类别、层次进行人工的分类,然后用户在分类完毕的目录中,逐级浏览,找出满足自己

*)国家自然科学基金(60003004)资助。周 晋 博士生,研究方向为网络信息服务,人工智能,MAS。路海明 博士,讲师,研究方向为网络信息服务,人工智能,MAS。李衍达 教授,博士生导师,中国科学院院士,研究方向为网络信息服务,生物信息学,智能信号处理等。

需求的结果。

1)分类目录的优点 ①搜索结果准确性高;②分类目录预先设定好,故检索速度快;③分类固定,便于用户熟悉搜索引擎。

2)分类目录的缺点: ①无法浏览已有类别以外的内容,仍需使用标准关键词输入方式进行搜索;②类别划分主要依靠人为分类,主观成分多;③由于自动分类技术还不成熟,目前目录的维护以手工为主、自动为辅,维护工作量大;④某些类别要经过多级点击才能浏览到,输入过程时间较长。

2.3 浏览模型输入方式

许多时候,人们的需求往往难以用语言描述清楚,或者人们不情愿自己把问题阐述得十分透彻,而更乐于在交互的过程中,让系统帮助自己理清并最终理解自己的目的。针对这个问题,Kevin Cox[2]提出了浏览模型输入方式。

在浏览模型检索系统提供的对象选择集中,用户通过有倾向、有选择的浏览行为,最终找出满足自己需求的网页。这里的对象可以是:网页、索引或关键词向量。浏览模型可以充分挖掘用户本身认知、识别能力,根据用户对有用资源所在位置的判断,把握搜索方向。Kevin Cox 认为,这比计算机按照固定算法去查找东西要准确、快捷得多。

1)浏览模型的执行过程

a. 用户在一个分类目录里选择初始选择集。

b. 用户在选择集中选择最接近自己目标的对象(可以多个)。

c. 系统根据选择对象,按照最近邻网络的匹配原则,返回新一轮的选择集。

d. 用户继续在选择集中选择最接近自己目标的对象。

e. 重复上述过程,直到用户找到满意结果。

2)浏览模型的优点 ①用户无需给出关键词表达式;②充分挖掘用户的认知、识别能力,自行把握搜索方向与速度;③由于数据库结构事先已经设计完成,且数据库操作简单,因此操作起来速度快;④系统适用性广,数据库中的对象可以是:网页、索引或关键词向量;⑤相似度度量适用性广,可以应用任何一种相似度算法。

3)浏览模型的缺点 ①按照最近邻原则连接的数据库可能突现异常现象,导致搜索方向偏离;②无法根据用户的个性特点,适应性调整系统;③对于用户完全未知的事物,系统很难搜索出满意的结果。

2.4 关键词重构输入方式

调查表明,用户在使用搜索引擎时,在输入关键词后,如果检索不到满意的结果,往往在原有关键词基础上进行修改,然后再次查询,这样的过程可能重复多次。针对这种行为,人们设计了关键词重构输入方式,它将原有关键词的多种演化形式自动提供出来,用户只需在这些重构关键词中进行选择,便可以完成进一步的搜索。关键词重构简化了用户的输入过程,并可以帮助用户理清思路、确定需求。

P. D. Bruza[3]等人通过对用户行为的分析,发现用户修改关键词时,有多种重构方式,如表 2 所示。

统计分析表明,REP、ADD、DEL 为主要修改方式(其中,REP 没有讨论的意义,不再考虑),ADD 占统计总数的 14%,DEL 占 21%,所以搜索引擎在设计关键词重构时,主要考虑 ADD、DEL 两种情况即可。

1)关键词重构输入方式的执行过程

a. 假定用户输入关键词为:Internet;

b. 系统给出调整后的关键词的集合:Internet Direct、Guide for Internet、……、Internet Security、Internet Solution;

c. 从中,用户选择了 Internet Security;

d. 系统再给出调整后的关键词集合:Internet Security Firewall、Internet Security Software、……、Mid-range Internet Security;

e. 用户选择 Internet Security Software;

f. 重复上述过程,直到用户找出满意的关键词,然后进行查询。

表 2 关键词重构方式

重构方式	说明	举例
SPL	分离或连接关键词	rockclimb→rockclimb center point→centerpoint
DEL	删除某些关键词	malaysia electricity→malaysia
ADD	加入某些关键词	windows95→windows95 help
REP	重复原有关键词	soccer→soccer
SUB	替代某些关键词	electronic commerce → electronic contact
DER	原关键词的派生词	jobs→job tourism→tour
SPE	修改拼写错误	
ABR	缩写或扩展缩写	jpl→jet propulsion laboratories
PUN	修改符号	hitch-hikers guide → hitchhikers guide
CAS	修改大小写	food→FOOD
MIS	多种重构方式混合使用	

2)关键词重构的优点 ①克服用户不知道如何用关键词语法表达信息需求的问题,帮助用户理清思路;②帮助用户尽可能地精炼信息需求,同时剔除不相关的需求,避免了歧义。例如:若 surfing 精炼为 wave surfing,则剔除了 internet surfing 的含义;③满足用户不断升级信息需求的情况。例如:quilting store 升级为 quilting store in their location;④简化了用户的输入过程。

3)关键词重构的缺点 ①检索范围有限;②难以评价系统的检索性能;③缺乏个性化设置。

2.5 关键词调整输入方式

在以前的搜索引擎中,用户输入关键词后,系统原封不动地将关键词送往数据库进行检索。由于用户输入的关键词往往简单、模糊、有歧义,导致检索质量不高。

关键词调整输入方式允许用户在输入关键词的同时,同时选择本次检索的个性偏好。然后,系统在不改变用户原意的前提下,根据个性偏好,适当调整或修改关键词,使之更准确地描述用户需求,从而提高检索质量。关键词调整应用举例如下:

例 1 用户初衷:找出最近的体育新闻。

用户输入的关键词:sports news

用户选择的个性偏好:current events

系统调整后的关键词:sports news + in the last two weeks

说明:在限定了"sports news"的具体时限"in the last two weeks"后,就可以更符合用户本意的新闻了,否则许多过期的新闻也会被一并检索出来。

例 2 用户初衷:找出有关"information filtering"的科技

文献

用户输入的关键词:information filtering

用户选择的个性偏好:research papers

系统调整后的关键词:information filtering + abstract + keywords + introdution

说明:如果仅检索"information filtering"会找出很多不属于科技文献的结果,根据科技文献的特征,系统加入"abstract"等词加以限定,大大提高了结果为科技文献的可能性。

关键词调整有如下优点:①提高了检索精度、检索质量;②缩小了检索范围;③加快了检索速度;④对于 meta search 类的搜索引擎,当主搜索引擎将关键词发往元搜索引擎时,可以根据元搜索引擎各自的特点,进行关键词调整,充分利用了各个元搜索引擎的优势。

2.6 文章集输入方式

文章集输入方式的主要思想是:把输入内容由关键词变成了一篇或多篇文章,用文章集的含义来表达用户的信息需求,通过检索与文章集相似的信息,来得到用户想要的结果。

1)文章集输入方式的优点

·用文章集来表达用户需求,避免了用户选择关键词或选择类别的困难。因为,有时候人们很难说出自己的信息需求,但是能说出所看到的文章是否是自己需要的;

·文章集包含的语义空间比关键词更具体、更精确,缩小了检索范围,提高了检索质量,加快了检索速度;

·有助于多语种信息检索的处理。多语种文章检索,往往采用自动翻译、关键词翻译或向量翻译等方法,这些方法都有很大的信息损失。而使用一个有限的文章集表示用户的信息需求,可以将该文章集从一种语言 A 的表达人工地翻译成另一种语言 B,在 B 中检索相似信息。这样,用户就可以输入自己母语文章集,对不同语言的信息进行检索。

2)文章集输入方式的缺点

·输入过程较为复杂;

·选取到合适的文章集有一定的困难。

3)输入方式的比较 通过对几种输入方式的了解,我们可以总结出一些衡量输入方式的指标,输入方式的指标比较如表 3 所示。下面先对这些指标做一下说明。

·反映输入性能方面的指标有:覆盖面(指输入方式的表达范围)、无歧义性、易表达性、输入速度(衡量整个输入过程快慢的指标)、个性化(指根据用户自身特点调整输入内容的能力);

·反映检索性能方面的指标有:检索速度、查全率、查准率;

·反映系统维护方面的指标有:占用硬盘空间(指检索系统占用的硬盘空间)、工作量(维护系统所花费的手工工作量)。

表 3 输入方式的比较

输入方式	输入性能					检索性能			系统维护	
	覆盖面	无歧义性	易表达性	输入速度	个性化	检索速度	查全率	查准率	占用硬盘空间	工作量
关键词	优	差	良	优	差	差	优	差	大	小
分类目录	差	优	优	差	差	优	差	优	小	大
浏览模型	良	良	优	差	差	优	良	良	大	中
关键词重构	良	良	优	差	差	良	良	良	大	中
关键词调整	优	优	良	优	良	良	优	优	大	中
文章集	差	良	良	良	差	差	差	良	大	小

在比较的基础上,针对输入方式的发展情况,我们有以下分析:

·关键词输入方式:各项性能不高,是导致系统检索质量不高的重要原因之一;

·分类目录输入方式:利用人的智慧进行分类,在一定程度上提高了检索质量,这也正是 Yahoo! 成为互联网上最受欢迎的搜索引擎的重要原因。但是分类目录输入方式的检索质量难以提高,并且由于分类固定,用户的个性化特点难以体现,这些都阻碍了分类目录搜索引擎有更大的发展。

·浏览模型输入方式:能解决一些关键词输入方式难以克服的问题,但其技术发展尚不充分,在短期内,浏览模型输入方式还只是其他输入方式的有益补充。但它的挖掘用户认知、识别能力的思想,为研究输入方式提供了一个崭新的思路。

·关键词重构输入方式:它简化了用户修改关键词的过程,但对提高检索质量,没有本质上的改进。

·关键词调整输入方式:从表 3 中可以看出,它的综合性能最好,实践中,NEC 研究机构的 Inquirus 2 系统的良好表现也验证了这一点。关键词调整技术容易嫁接到其他输入方式上,并且对原有输入方式几乎无负面影响。今后,关键词调整技术将会受到越来越多搜索引擎开发者的青睐。

·文章集输入方式:综合性能差。虽然将关键词换成文章集可以提高需求描述的准确性,但最终结果的检索依赖于文章间相似度的计算,而目前相似度技术不成熟,所以文章技术输入方式的优越性也不能得到充分体现。

结论 提高搜索引擎检索质量应从改进输入方式入手,在分析和比较了几种已有的输入方式后,我们可以大胆地展望改进搜索引擎输入方式的几种可行途径。

1)多种输入方式的有机融合 关键词输入方式和关键词调整输入方式在结构上十分相似,二者的融合是很自然的事情,可以说,融入关键词调整技术对关键词输入方式是有百利而无一害的。可以应用关键词输入方式的地方,几乎都可以融入关键词调整方式。

分类目录、浏览模型、关键词重构等浏览式输入方式由于分类数据预先固定,导致了覆盖面窄、查全率低等缺点。但如果借用 meta-search 的思想,当用户的关键词在本地数据库中无法检索到时,可以向支持关键词检索的搜索引擎发出检索请求,获取检索结果,显示给用户。这便是浏览式输入方式和关键词输入方式的一种合理融合。

文章集输入方式和关键词输入方式也可以融合在一起,即:网页+关键词作为输入内容一同递交给搜索引擎,其中,关键词作为网页的一个有益补充,这样可以解决网页不能完全表达用户需求的缺点。

总之,在没有出现一种完美的输入方式之前,各种输入方式有机地融合在一起、取长补短,是一种改进输入方式的有效解决方案。

2)个性化设计 从表 3 中看出,只有关键词调整加入了个性化设计,其余的都没有考虑。

越来越多的需求表明,搜索引擎不仅能给出准确的结果,还要能提供个性化搜索。个性化设计是提高搜索引擎检索质量的一个重要方面。也就是说,搜索引擎通过对用户的不断了解、分析,给出的搜索结果要符合每个用户的个性需求。这就要求输入方式也要能适应用户的个人特点,准确无误地将用户需求传递进来。

在输入方式中加入个性化设计，可以从以下几条思路入手：

·输入方式的个性化选择：不同的用户可能喜欢不同的输入方式，这就要求搜索引擎能提供多种输入方式供用户选择，当然相应的检索系统应该能接受多种形式的输入。

·输入内容的个性化调整：通过用户个人描述或总结用户以往检索历史，适当调整用户的输入内容。例如：从用户以往的检索历史中得知，该用户的关键词"网络安全"实际上是指"局域网网络安全"，那么用户再次检索"网络安全"时，输入应调整为"局域网网络安全 OR 网络安全"，且第一个关键词的权重更大。

·输入界面的个性化调整：个性化调整输入界面，可以改善输入的友好程度、简化用户的输入过程、提高用户的输入速度。例如：Yahoo！分类目录可以根据用户的国籍调整输入界面的语言，还可以根据用户的地理位置、年龄段，给出不同的目录内容。

3）专用化设计也称垂直化设计　是随着用户检索领域的变化，输入方式也相应变化的一种技术，专用化输入方式可以让用户在一个专用领域内，详细、准确地表达自己的需求。例如：检索领域选择为"学术文献"，则输入方式允许用户输入题名、作者、出处等项目进行检索。又如：检索领域选择为"汽车销售"，则输入方式允许用户输入型号、厂家、功率、颜色等项目进行检索。

随着可扩展标记语言 XML 的普及，网页数据可以分门别类地表达出来，这样，专用化输入方式可以更方便地检索出准确的结果。

此外，改进输入方式还可以考虑融合人工智能、多代理等新技术，以及挖掘用户认知能力等方面。总之，提高检索质量必须建立在搜索引擎充分理解用户需求的基础上，所以在输入方式领域，我们还应进行更深入的研究。

参 考 文 献

1 Glover E J，et al. Recommending Web Documents Based on User Preferences. In：ACM SGIR 99 Workshop on Recommender Systems，Berkeley，CA，Aug. 1999

2 Cox K. Information Retrieval by Browsing. In：Proc. of The 5th Intl. Conf. on New Information Technology，1992

3 Bruza P D，Dennis，S. Query reformulation on the Internet：Empirical data and the hyperindex search engine. In：Proc. of the RIAO97 Conf. -Computer-Assisted Information Searching on the Internet，Centre de Hautes Etudes Internationales d'Informatique Documentaires，June，1997. 488～499

4 Lawrence S，Giles C L. Searching the World Wide Web. Science，1998，280(5360)：98

5 Lawrence S，Giles C L. Context and page analysis for improved web search. IEEE Internet Computing，July-August，1998. 38～46

6 Arens Y，Knoblock C A，Shen W. Query reformulation for dynamic information integration. Journal of Intelligent Information Systems，1996

7 Nastar C，Mitschke M，Meihac C. Efficient Query Refinement for Image Retrieval. IEEE Conf. on Comp. Vis. and Pattern Recognition，Santa Barbara，California，1998. 547～552

8 Jansen B J，Spink A，Bateman J，Saracevic T. Real life information retrieval：A study of user queries on the Web. ACM SIGIR Forum，1998，32(1)：5～17

（上接第 14 页）

$[\![s]\!]$），记作 $B[\![t]\!] \rightarrow B[\![s]\!]$。

定义 10（闭包与范式）　$\stackrel{\cdot}{\rightarrow}$是$\rightarrow$的自反传递闭包。对于项 s，若不存在满足 $s \rightarrow u$ 的项 u，则称 s 是一个范式。对于项 t 和 s，若 $t \stackrel{\cdot}{\rightarrow} s$ 且 s 是范式，则称 s 是 t 的一个范式。

例 7　ADD→INSTANTIATE（RECURSIVE（rec（s（0），recursive（0，s（0））））。同样，ADD→INSTANTIATE

```
([ recursive;x,y:
   recursive(0,y) →base(y),
   recursive(s(x),y)→s(recursive(x,y))]
(recursive(s(0),s(0))))
```

下面的项是 ADD 的（唯一的）范式：

```
[ base,rec,0,s;x,y:
  base(x) →x,
  rec(x,y) →s(y),
  recursive(x,y)→redundant]
([ recursive;x,y:
   recursive(0,y) →y,
   recursive(s(x),y)→s(recursive(x,y))]
 (s(s(0))))
```

4. 动态项重写计算的特征

从上述语法和操作语义的定义和示例可以看出，DTRC 具有以下主要特征：

1. 高度层次化的结构：DTRC 项中的系统嵌套呈树形结构。

2. 灵活的动态重写：可以使用上层系统中的规则重写下层系统中规则，并通过函数声明和变量声明控制这样的动态重写。

3. 项重写系统的扩展：DTRC 的框架是项重写系统的扩展。这使我们得以使用 DTRC 项形式地描述项重写系统乃至于 DTRC 框架的元计算，并对形式描述进行形式验证[2-6]；同时，由于语法和语义上的相似性，可以使用和扩展项重写系统的研究手法来研究 DTRC 的合流性、终止性等基本性质[7]。

4. 元变量的实现：可以在 DRTC 项中简单地实现在形式证明中起重要作用的元变量的概念[2-6]。这些特征及对 DTRC 的研究和应用声明，DTRC 是项重写系统的强有力的元计算模型。

参 考 文 献

1 Dershowitz N. Termination of Rewriting. J. Symb. Comput.，1987，3：69～115

2 Feng S，et al. Mechanizing Weak Termination Proving of Term Rewriting Systems by Induction. In：Proc. ICYCS'2001. 2001. 15～19

3 Huet G. Confluent Reductions：Abstract Properties and Applications to Term Rewriting Systems. J. ACM，1980，27：797～821

4 Knuth D E，Bendix P. Simple Word Problems in Universal Algebra. In：J. Leech，eds. Computational Problems in Abstract Algebra：Oxford，Pergamon Press，1970. 263～297

5 冯速．项重写系统弱基终止性的归纳证明．计算机科学，2001，28(7)：105～108

6 Feng S，et al. Mechanizing Explicit Inductive Equational Reasoning by DTRC. IEICE Trans. Inf. & System.，1995，E78-D(2)：113～121

7 Feng S，et al. Confluence Property of Simple Frames in Dynamic Term Rewriting Calculus. IEICE Trans. Inf. & Syst.，1997，E80-D(6)：625～645