



[12] 发明专利申请公开说明书

[21] 申请号 02100486.2

[43] 公开日 2003 年 1 月 8 日

[11] 公开号 CN 1389811A

[22] 申请日 2002.2.6 [21] 申请号 02100486.2

[71] 申请人 北京造极人工智能技术有限公司

地址 100086 北京市海淀区中关村大街 49 号
大华科技商厦 A408

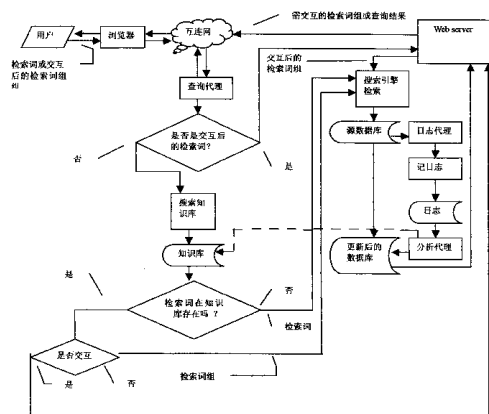
[72] 发明人 韩立岩 王 强 张寅生 周 密

权利要求书 3 页 说明书 14 页 附图 4 页

[54] 发明名称 搜索引擎的智能化搜索方法

[57] 摘要

本发明由 3 项子发明构成,涉及到计算机网络技术中对具有自然语言特征的文本和文献目录的搜索方法,特别涉及到搜索引擎的智能化搜索方法。发明 1 是按照用户点击次数对网页及网站地址和文献目录进行排序的方法,通过服务器日志或计数器统计用户输入检索词后用户对对应检索词的网址和目录的点击率,按点击率大小对网址和目录排序。发明 2 是与文本内容匹配的基于检索词的知识库的建立和应用的方法,它按照建立数据库的方法把实词的关联词集合起来,使得用户输入一个检索词时系统给出其关联词。发明 3 是搜索引擎检出文献目录的方法,它使得搜索引擎不只能搜索文本的地址,还能搜索文献目录。



本发明由 3 项子发明构成,发明 1 是按照用户点击次数对网页及网站地址和文献目录进行排序的方法,发明 2 是与文本内容匹配的基于检索词的知识库的建立和应用的方法,发明 3 是搜索引擎检出文献目录的方法。

发明 1 权利要求书:

1 一种按照用户点击次数对网页及网站地址和文献目录进行排序的方法,其特征在于,通过计算机程序记录用户对输入检索词后搜索引擎给出的对应该检索词的网址和目录的点击次数和下载次数,按照一定时间内记录的多个用户输入的同一检索词对应的网址和目录的用户点击次数进行排序,排序的规则是:同一检索词对应的网址或目录击次数多,则排前列,为以后输入该词的用户优先给出;某检索词对应的某些网页或网站下载记录次数作为排序的权值。

2 如权利 1 所述的方法,记录对输入检索词后搜索引擎给出的对应该检索词的网址和目录的点击次数或下载次数,

3 如权利 1 所述的方法,记录对输入检索词后搜索引擎给出的对应该检索词的网址和目录的点击次数或下载次数系统是服务器日志,

4 如权利 1 所述的方法,记录对输入检索词后搜索引擎给出的对应该检索词的网址和目录的点击次数和下载次数系统是用户点击计数器统计分析系统软件,

5 如权利 1 所述的方法,按照对输入检索词后搜索引擎给出的对应该检索词的网址和目录的用户的点击次数对网址和目录进行排序,

6 如权利 1 所述的方法,按照对输入检索词后搜索引擎给出的对应该检索词的网址和目录的用户的点击次数对网址和目录进行排序,当对应某检索词的网址有下载记录时,下载记录作为点击次数的权重进行排序,

7 如权利 1 所述的方法,对现存的搜索引擎的存放检索词和网址的数据库进行更新,更新后的数据库中检索词、网址和点击次数是对应的,多个用户输入的同一检索词对应被记录的输入该检索词后由现存的搜索引擎搜索到的网址,形成检索词和网址的关系,该网址对应用户输入该检索词后由现存的搜索引擎搜索到的网址的点击次数,形成检索词和网址和点击率的关系,

8 如权利 1 所述的方法,对现存的搜索引擎的存放检索词和文献目录的数据库进行更新,更新后的数据库中检索词、目录和点击次数是对应的,多个用户输入的同一检索词对应被记录的输入该检索词后由搜索引擎搜索到的目录,

形成检索词和目录的关系，该目录对应用户输入该检索词后目录的点击次数，形成检索词和目录和点击率的关系，

9 如权利 1 所述的方法，对现存的搜索引擎的存放检索词和网址的数据库进行更新，更新后的数据结构的主元是网址，对应的关系是检索词和点击次数，

10 如权利 1 所述的方法，对现存的搜索引擎的存放检索词和网址的数据库进行更新，更新后的数据结构的主元是检索词，对应的关系是网址和点击次数，

11 如权利 1 所述的方法，对现存的计算机文献检索系统中的存放检索词或目录的数据库进行更新，更新后的数据结构的主元是目录，对应的关系是检索词和点击次数，

12 如权利 1 所述的方法，对现存的计算机文献检索系统中的存放检索词或目录的数据库进行更新，更新后的数据结构的主元是检索词，对应的关系是目录和点击次数，

13 如权利 8 所述的方法，对文献目录的数据库进行更新，文献目录包括专利文献检索系统、国家标准检索系统、图书检索系统，但不对这些检索系统本身提出权利要求。

14 如权利 7 所述的方法，更新后的数据库中检索词、网址和点击次数被编号，每一个检索词的编号都是唯一的，每一个网址的编号都是唯一的。

15 如权利 8 所述的方法，更新后的数据库中检索词、目录和点击次数被编号，每一个检索词的编号都是唯一的，每一个目录的编号都是唯一的。

16 如权利 1 所述的方法，网址是 IP 地址或 ARP 转换的地址。

发明 2 权力要求书:

17 一种与文本内容匹配的基于检索词的知识库的建立和应用的方法，其特征在于，对实词按照内容的关联性建立联系，对每一个词汇进行唯一的编号，按照数据库格式和语言将实词及其关联性关系进行存储和管理，成为知识库，嵌入搜索引擎，当用户输入检索词后，搜索引擎首先将该检索词在知识库元组中搜索，如果搜索到该检索词，则知识库给出与检索词对应的一组关联词，再将检索词与关联词都输入搜索引擎进行搜索；知识库给出的检索词及与之对应的关联词，可以与用户交互，由用户选择其中的词汇，也可以不与用户交互；如果在知识库中没有搜索到该检索词，则搜索引擎将该检索词进行搜索。

18 如权利 17 所述的方法，对自然语言的实词按照内容的关联性建立联系，分为 6 级：(1)第一上属关系，(2)第二上属关系，(3)同义词关系，(4)相关关系（因果关系、原料-产品关系、服务-被服务关系），(5)第一下属关系，(6)第二下属关系。

19 如权利 18 所述的方法，知识库的格式和语言是数据库的，检索词与关联性的词汇构成数据库的对应的关系。

20 如权利 18 所述的方法，知识库的格式和语言是数据库的，其主元是检索词。

21 如权利 17 所述的方法，检索词及其关系可以增设用相关度衡量的功能，相关度为可以转化为 0~1 之间的数值，该相关度可以作为该组词的各个词汇对应搜索结果（即不同词汇对应的网址或目录序列）排序的权值。

22 如权利 21 所述的方法，相关度可以被用户所交互，按照用户对关联词的点击率折算成权值进行修改。

发明 3 权利要求书:

23 一种用搜索引擎检出文献目录的方法，其特征在于搜索引擎系统内建立并使用文献目录数据库和文献目录检索系统，通过与用户交互由用户决定是否需要专利、标准、图书馆文献目录，哪一个图书馆的目录，当用户输入检索词时，增设的知识库将给出与检索词意义相关的一组词，即：将一个检索词变成检索词组，转变方法见发明 2，检索词组进入文献目录检索系统后检索系统通过扫描文献目录的类别的名称的词汇或摘要与检索词组的匹配程度（频率和概率），确定目录的类别，再根据目录类别对应的文献目录的号码，检出文献目录，该目录以被用户点击的记录的大小作为排序的依据。

搜索引擎的智能化搜索方法

本发明由 3 项子发明构成,发明 1 是按照用户点击次数对网页及网站地址和文献目录进行排序的方法,发明 2 是与文本内容匹配的基于检索词的知识库的建立和应用的方法,发明 3 是搜索引擎检出文献目录的方法,三种方法均涉及计算机网络技术中对具有自然语言特征的文本和文献目录的搜索方法,具体地说,涉及到搜索引擎的智能化搜索方法。

计算机网络中的搜索引擎是搜索网页及网站的工具。现存的搜索引擎的基本原理是:通过网页及网站的搜集软件,从互联网上自动收集网页地址及其文本,将搜集所得的网页文本交给索引和检索系统,由计算机程序通过扫描文本中的每一个词,建立以词为单位的倒排文件,检索程序根据用户给出的检索词在文本中出现的频率和概率,对包含这些检索词的文本进行排序,最后输出网页及网站的排序的结果。这种搜索方法存在以下缺陷:第一,无用的(对于检索词不相关的或相关度低的)搜索结果过多。这是因为单一的检索词在文本中出现的频率和概率并不完全代表检索词与文本内容的相关性的值。即:目前的搜索方法不能识别与检索词相关的文本内容。第二,它只能搜

索可链接于因特网的网页及网站地址，不能以文献目录为搜索单位，即不能检出文献目录。例如，现有的搜索引擎不能找到某一检索词对应下的专利号、国家标准号以及某一图书馆中的书号。

本发明能够弥补这两个缺陷。其中，发明1和发明2能够弥补第一个缺陷，发明3能够弥补第二个缺陷，这就是本发明的目的和动机所在。

发明1的要点是：

对于搜索网址的工作，在现有的搜索引擎系统中加入服务器日志或用户点击计数器统计分析系统，该日志或系统完成以下工作：1。记录不同用户对输入同一检索词后搜索引擎按照检索词在网页或网站的文本中出现的频率和概率的多少进行排序的网页的点击次数和下载次数。2。对一定时间内记录的多个用户输入的同一检索词中对应的网页或网站的用户选择结果（不同网页或网站的点击次数）进行排序，排序的规则是：同一检索词对应的网页或网站点击次数多，则排前列作为以后输入该词的用户优先给出的搜索结果；某检索词对应的网页或网站下载记录次数作为排序的权值。这两步工作的目的是记录按照人（用户）的智力对目前的搜索引擎对网页或网站的搜索结果与索引词的相关性的判断。亦即：在加入本发明以前，搜索引擎对网页或网站的搜索是计算机系统依据检索词在网页或网站的文本中出现的频率和概率排序（现有的搜索方法）；而加入本发明后，将在此现存的搜索方法搜索的结果后增补一个步骤，该步骤由新增的计算机

软件系统记录了人（用户）的智力对现存的搜索方法条件下的搜索结果在检索词与网页或网站的内容的相关程度上的判断和选择，以此作为对下一个输入同一词的用户提供搜索结果的排序依据。

实现上述要点，主要通过以下步骤实现，每个步骤均由一个软件模块按顺序执行：

1. 对现存的搜索引擎系统中网页或网站的数据库进行更新，其方法是定期抽取服务器日志或其它网页点击计数器中与用户输入检索词对应的网址的点击次数，形成新的数据库，本过程所发挥的功能定义为建立与检索词对应情况下网页或网站的点击次数的数据库模块。该数据库的基本特征三个参数（属性值）的组合，即：（1）用户输入的检索词与（2）现存的搜索系统中通过检索词文本扫描后匹配的对应的网址以及（3）用户在输入此检索词后对该词对应的网址的点击次数（访问次数）的组合，从而反映出在一定时间内不同用户输入同一检索词后用户在现存的搜索引擎给出的搜索结果（多个网页或网站）中选择了哪（些）个。对现有的搜索引擎系统中网页或网站的数据库进行更新后建立的数据库其结构可有多种设计，可以选择IP 地址或检索词作主元。图 1 是与检索词对应情况下网页或网站的点击次数的数据库（更新后的数据库）结构示意图。

2. 通过查询与检索词对应情况下网页或网站的点击次数的数据库对与检索词对应的网页或网站按照点击次数进行排序，对于有下载

记录的网址，将每一次下载和每一次点击赋予不同权重进行排序。即本模块定义为与检索词对应的网页或网站按照点击次数和下载次数排序模块。在排序后与检索词对应的网页或网站的网址序列作为下一个用户输入同一词时的网址搜索结果序列。当用户输入的查询词被外部系统给出内容相关的一组词时，本模块可按该组词的给定的相关度对该组词的每一个词对应的搜索到的网址序列进行加权，从而给出不同权重的多组对应的的网址序列，再在此多组网址序列中按不同权重重新排为一个序列提供给用户。

上述步骤与现存的搜索引擎相结合，实现了将现存的搜索引擎的搜索结果由检索词在文本中出现频率和概率大小作为排序依据转变为由用户（人的智力）对检索词与文本在内容的关联性的大小的判别（对应检索词的网页或网站的访问次数，即点击率）为排序依据。

对于搜索文献目录的工作，发明 1 在现有的搜索引擎系统中加入服务器日志或用户点击计数器统计分析系统，将对检索词（组）对应的文献目录数据库的目录的点击数和点击次序进行记录。此后对目录按点击数和点击次序进行排序，点击次序被赋予权重加入点击数排序之中，形成一个排序序列。

搜索引擎搜索目录的方法见发明 3。

搜索引擎在加入发明 1 之后与计算机目录检索系统（如现存的图书馆的图书计算机检索系统、专利、国家标准计算机检索系统）的差别在于：1. 在发明 1 应用以前，计算机目录检索系统在输入检索词

之后由系统对目录的搜索是计算机系统依据检索词在目录中出现的频率和概率排序；而加入本发明后，将在此传统的计算机检索的结果后增补一个步骤，该步骤由新增的计算机系统记录了人（用户）的智力对传统的计算机检索的结果在检索词与网页或网站的内容的一致程度（相关性）上的判断和选择，以此作为下一个输入同一词的用户提供搜索结果的排序依据。2. 在发明 1 应用以前，搜索引擎的数据库不包含文献目录检索系统而只包含文本所在的网页或网址的检索系统，当用户需要检索文献目录时，现存的搜索引擎只能链接文献目录的网址，再由用户使用文献目录的检索系统检索。加入发明 1 后，搜索引擎的数据库将包含文献目录检索系统，用户对检索词对应的搜索单位将含盖网址和文献目录，使用户不必访问文献目录所在网页，而由搜索引擎一次完成网址和文献目录的搜索。

实现上述要点，主要通过以下步骤实现，每个步骤均由一个软件模块按顺序执行：

1. 对现有的被检索的目录数据库进行更新，其方法是定期抽取服务器日志或点击计数器中与用户输入检索词对应的目录的点击次数，形成新的数据库。此模块定义为记录与检索词对应的目录的点击次数模块。该数据库的基本特征三个参数（属性值）的组合，即：

（1）用户输入的检索词与（2）现存的搜索系统中通过检索词匹配的目录以及（3）用户在输入此检索词后对该词对应的目录的点击次数（访问次数）的组合，从而反映出在一定时间内不同用户输入同一检

索词后用户在现存的搜索引擎给出的搜索结果（目录）中选择了哪（些）个。对现有的用于计算机目录检索的目录数据库进行更新后建立的数据库其结构可有多种设计，可以选择目录或检索词作主元。图2是与检索词对应情况下目录的点击次数的数据库（更新后的数据库）结构示意图。

2. 通过查询与检索词对应情况下目录的点击次数的数据库，对与检索词对应的目录按照点击次序和次数进行排序（将次序折算成权值加入点击次数排序）。本模块定义为与检索词对应的目录按照点击次数排序模块。在排序后与检索词对应的目录序列作为下一个用户输入同一词时的目录搜索结果序列。当用户输入的查询词被外部系统给出内容相关的一组词时，本模块可按该组词的给定的相关度对该组词的每一个词对应的按照点击次数进行排序的目录序列进行加权，从而给出不同权重的多组对应的的目录序列，再在此多组网址序列中按不同权重重新排为一个序列提供给用户。

上述步骤与现存的搜索引擎相结合，实现了将现存的搜索引擎的搜索结果由检索词在目录中出现频率和概率大小作为排序依据转变为由用户（人的智力）对检索词与文本在内容的关联性的的大小的判别（对应检索词的目录的访问次数，即点击率）为排序依据。

使用本发明流程见图3。

图3的用语说明如下：

知识库：记录词汇之间的意义（内容）相关性的软件系统，其功能在于在用户输入检索词后按该检索词的意义的相关性给出一组检索词。

源数据库：现存的搜索引擎的数据库，用于存储网址或文献目录的和按照检索词在文本或目录中出现频率和概率大小作为排序依据排序的与检索词对应的网址或文献目录数据库。

更新后的数据库：对源数据库按照被搜索对象（网址或文献目录）在对应检索词的情况下用户的点击率多少进行重新排序的数据库。

检索词：用户输入的符合自然语言的语法规则的词汇、词组，以及非句子的连续的词汇的集合，即在用户界面中要求用户输入的搜索的“关键词”。

从图3的左端开始，在用户提出了检索词的查询请求之后，通过浏览器界面，网络服务器端得到用户的请求，启动查询代理进行响应。查询代理搜索知识库，从中找出是否存在与用户查询的检索词匹配或是相关的检索词组。判断是否相关的依据是专家事先对知识的分类的知识点间的相关性或定量化的相关系数。如果存在，则生成一个相关的“检索词组”。如果系统设计成与用户交互式，即由用户对他们感兴趣的由知识库给出的与检索词相关的检索词组进行选择，则通过网络服务器提供给用户。通过提交表单，得到用户的选择。用户选择后由查询代理经由网络服务器交给搜索引擎检索。如果系统不设计成与用户交互式，则知识库生成的相关的“检索词组”直接交给搜索引擎

检索。如果用户输入的索引词在知识库不存在,该检索词没有相关词匹配,直接(单一地)交给搜索引擎检索。

这样,搜索引擎检索系统可能接受三种中的一种要求检索的检索词的集合:检索词(用户最初输入的单一的检索词),或由知识库按相关性对检索词进行匹配的一组检索词,该词组不经过用户选择(检索词组),或由知识库按相关性对检索词进行匹配的一组检索词,该词组经过用户选择(交互后的检索词组)。无论三种中的哪一种检索词的集合,只要检索词的集合分别交给现存的搜索引擎,现存的搜索引擎都会在源数据库的文本或目录中扫描该检索词并按该词在文本或目录中出现的频率和概率的大小将文本的网址或目录排序。本发明加入后,将记录用户输入的检索词(无论是单一的还是一组,无论该组词是与用户交互过的还是未交互过的)和用户对按该词在文本或目录中出现的频率和概率的大小将文本的网址或目录进行排序的搜索结果(对现存的搜索引擎而言的搜索结果)的选择,即将记录用户输入的检索词和对应该检索词的网页或网站的点击率或目录的点击率。此后,发明1再根据点击率大小重新排序,当检索词为多个时,则对检索词对应的网页或网站或目录进行加权排序。排序后的结果作为应用了本发明的搜索引擎的对应该词的查询结果。

发明2的背景是:现存的计算机网络搜索引擎和计算机文献目录检索系统用用户输入的检索词搜索或检索网址或目录的基本原理是

由计算机程序通过扫描文本或目录中的每一个词,检索程序根据用户给出的检索词在文本中出现的频率和概率,对包含这些检索词的文本或目录进行排序,最后输出网页及网站的排序的结果。这种方法的缺陷之一是搜索或检索结果不很全面、不很准确。其进一步的原因之一是因为检索词在文本或目录中出现的频率和概率并不代表检索词与文本或目录内容的相关性的值。即:目前的搜索方法不能识别与检索词相关的文本内容。

例如,当用户输入“高等数学”检索词时,现存的计算机网络搜索引擎和计算机文献目录检索系统会把含有“高等数学”这一词汇出现频率高的文本网页或网站的地址或目录搜索或检索到,但是,许多以“微积分”作为主题词或标引词但没有或较少出现“高等数学”词汇的文本网页或网站的地址或目录则被丢弃或派在后位,尽管它可能比含有“高等数学”这一词汇出现频率高的文本网页或网站的地址或目录在内容上更接近高等数学。

发明 2 的目的和动机是弥补现存的计算机网络搜索引擎和计算机文献目录检索系统用用户输入的检索词搜索或检索网址或目录的缺乏内容的关联性的缺陷。

它的基本要点是:

按照词汇的语义的内容关联关系,建立词汇间的关联关系,并储存于计算机。在计算机内存中驻留一个小的进程,侦听用户的请求,用户输入的检索词被截获后转给搜索知识库。知识库把与检索词相关的一组词提供给用户选择,或直接提供给搜索引擎检索系统,进入现存

的搜索引擎工作程序。这样，搜索引擎就由现存的对用户的一次的一个检索词的搜索变为一组词的搜索，该组词上是由专家确定为内容相关的，可以增设用相关度衡量的功能，相关度为 0 - 1 之间，该相关度作为该组词的不同相关词汇对应搜索结果（即不同词汇对应的网址或目录序列）的权值，当需要将权值进行调整时，可以依据专家进行人为设定，也可通过与用户的交互进行，如果通过用户的交互进行调整，则通过设置服务器日志或用户点击计数器统计分析系统统计用户对检索词的关联词的点击率，再将点击率作为权值调整相关度。

实现的步骤是：

1. 建立知识库。按照知识分类，将词汇建立内容关联关系。关联关系分为 6 级：(1)第一上属关系，(2)第二上属关系，(3)同义词关系，(4)相关关系（如因果关系、原料-产品关系、服务-被服务关系），(5)第一下属关系，(6)第二下属关系。

例如：

酒（检索词）

酿造工业（第二上属关系）

酿酒工业（第一上属关系）

酿酒微生物（相关关系），粮食（相关关系），酒精（相关关系）.....

白酒（第一下属关系）

低度白酒（第二下属关系）

例如：

“脑血管意外”检索词，可以得到图 4 的关系结构。从图中可以

看到：“脑血管疾病”和“神经疾病”是它的上属级别，“中风”、“高血压”和“偏瘫”是同级的（“中风”是同义词，“高血压”和“偏瘫”是相关关系），其余的四个词则是它的下属关系的词。

图 4 是知识库关系的逻辑示意图。

2. 建立的知识库的结构如图 5、6 所示。

在图 5 中对于知识体系建立大的几个类，在大类的底下可以再分成一些子类，对于每一个知识大类的关键词进行统一编号。例如，用 10 位十进制的阿拉伯数字表示一个关键词。如“0101000001”。其中，前两位“01”表示的是“大类编号”，紧跟着的两位是大类下的子类的编号，后面 6 位数字表示的是“类内编号”。用这种方法就可以唯一确定一个关键词，并且通过分级的方法可以提高检索的效率。对以后的查询只要对关键词进行检索。在统一定义了关键词的编号后，以后的数据库中凡是出现关键词的地方都是用编号进行记录，从而减少了存储量，提高了效率。

在图 6 中，紧接在每个检索词之后的 c 是该检索词与该行首列检索词之间的相关系数。在专家打分给出初始的相关系数值之后，通过用户的使用，应用发明 1 中的用户对检索词输入后给出的一组相关的检索词的选择。用户对同一检索词的关联词的对应的网址或目录的点击次数可作为对同一检索词的关联词的选择次数，根据此次对检索词之间的系数进行更新。这样，就得到了一张如表 6 所示的“检索词关系表”。

3. 知识库的关联度的调整可以依据专家进行人为设定，也可通

过与用户的交互进行,如果通过用户的交互进行调整,则通过设置服务器日志或用户点击计数器统计分析系统统计用户对检索词的关联词的点击率,再将点击率作为权值调整相关度。这一过程见图3中的虚线。

因此,加入发明2之后,搜索引擎对搜索对象的排序标准就由按照单一检索词在文本或目录中出现的概率大小次序排序变为按照内容相关的一组词在文本或目录中出现的概率大小次序排序,从而使用户输入的检索词在内容上更加接近文本或目录。

发明3的背景是:现存的搜索引擎的存在以下缺陷:它只能搜索可链接于因特网的网页及网站地址,不能直接搜索链接于因特网上的网页及网站中的文献目录,即不能按照检索词检出文献目录,而只能检出网址(网址目录)。如果用户要求在网络环境下搜索网页内的文献目录,如某一检索词对应下的专利号、国家标准号以及某一图书馆中的书号,则用户只能按照搜索引擎提供的或自己知道的网址分别访问不同具有文献检索系统的网站,分别使用该文献检索系统检索文献目录,在此例中,用户需分别访问专利、国家标准以及某一图书馆的专门网站,使用该网站的专业检索系统,分别获得同一检索词对应的文献目录。这一过程的进一步的缺陷在于:用户的多次搜索和检索会增加许多造成失败的不确定性,如用户可能不知道文献目录所在的网址,或者增加网络搜索和文献目录检索时间网络的运行可能出现拥塞,用户操作复杂,等等。

发明 3 的应用将克服只能搜索可链接于因特网的网页及网站地址,不能直接搜索链接于因特网上的网页及网站中的文献目录的这一缺陷及其随之而来的进一步的缺陷。这是发明 3 的目的和动机。

发明 3 的要点和实现过程是:在搜索引擎系统内建立文献目录数据库和文献目录检索系统。通过与用户交互选择文献目录种类或文献目录系统,即由用户决定是否需要专利、标准、图书馆文献目录,哪一个图书馆。当用户输入检索词时,增设的知识库将给出与检索词意义相关的一组词,即:将一个检索词变成检索词组,转变方法方法见发明 2。检索词组进入文献目录检索系统。检索系统通过扫描文献目录的类别的名称(描述)中的词汇与检索词组的匹配程度(频率和概率),确定目录的类别,再根据目录类别对应文献目录的号码,检出文献目录。文献目录检索系统分为专利、标准、图书馆的图书检索系统以及专业数据库等多个种类。

专利检索系统中建立《国际专利分类表》,将检索词组在《国际专利分类表》的部、大类、小类、组的目录中进行扫描和匹配,根据检索词组部、大类、小类、组的目录中出现的频率和概率检出部、大类、小类、组的目录;通过部、大类、小类、组的目录的类号检出文献,再将检索词组在检出的文献文本中进行全文扫描和匹配,根据检索词组在检出的文献文本中出现的频率和概率将文献目录排序,该序列作为搜索引擎在专利文献检索系统中的搜索结果。

标准检索系统中建立《中国标准文献分类法》,将检索词组在《中国标准文献分类法》的类目(两级类目)中进行扫描和匹配,根据检

索引词组类目中出现的频率和概率检出类目；通过类目标识号检出文献，再将检索词组在检出的文献文本中进行全文扫描和匹配，根据检索词组在检出的文献文本中出现的频率和概率将文献目录排序，该序列作为搜索引擎在标准文献检索系统中的搜索结果。

图书馆的图书检索系统中，将检索词组在《中国图书馆图书分类法》的类目中进行扫描和匹配，根据检索词组类目中出现的频率和概率检出类目；通过类目分类标引号检出文献标题，再将检索词组在检出的文献标题中进行扫描和匹配，根据检索词组在检出的文献题目中出现的频率和概率将图书目录排序，该序列作为搜索引擎在标准文献检索系统中的搜索结果。如果建立的图书检索系统中附于图书题目还有关键词、主题词、标引词、内容简介等涉及图书内容等文字，则检索词组在检出的文献标题中进行扫描和匹配时可同时扫描这些文字，并以检索词组在这些文字中出现的频率和概率作为图书目录排序的依据。

这种过程的应用，用户在使用计算机搜索引擎时直接搜索到专利、标准、图书馆的图书以及专业数据库等多个种类文献目录，而不必先搜索到该检索系统的网站，再用该网站的搜索系统检索，检索的方法也由将索引词直接扫描文献目录变为由检索词生成意义相关的检索词组，再确定目录类别，然后在该类别下进行扫描文献目录，提高了文献目录检索的效率和准确性。

图 1:

IP 地 址	检索词 集	相关词 编号 1	访问 次数 1	相关词 编号 2	访问 次数 2	相关词 编号 n	访问 次数 n
地址 1	M1	K1	N2	K2	N2		Kn	Nn

图 2:

目 录 条	检索词 集	相关词 编号 1	访问 次数 1	相关词 编号 2	访问 次数 2	相关词 编号 n	访问 次数 n
目录 条 1	M1	K1	N2	K2	N2		Kn	Nn

图 3:

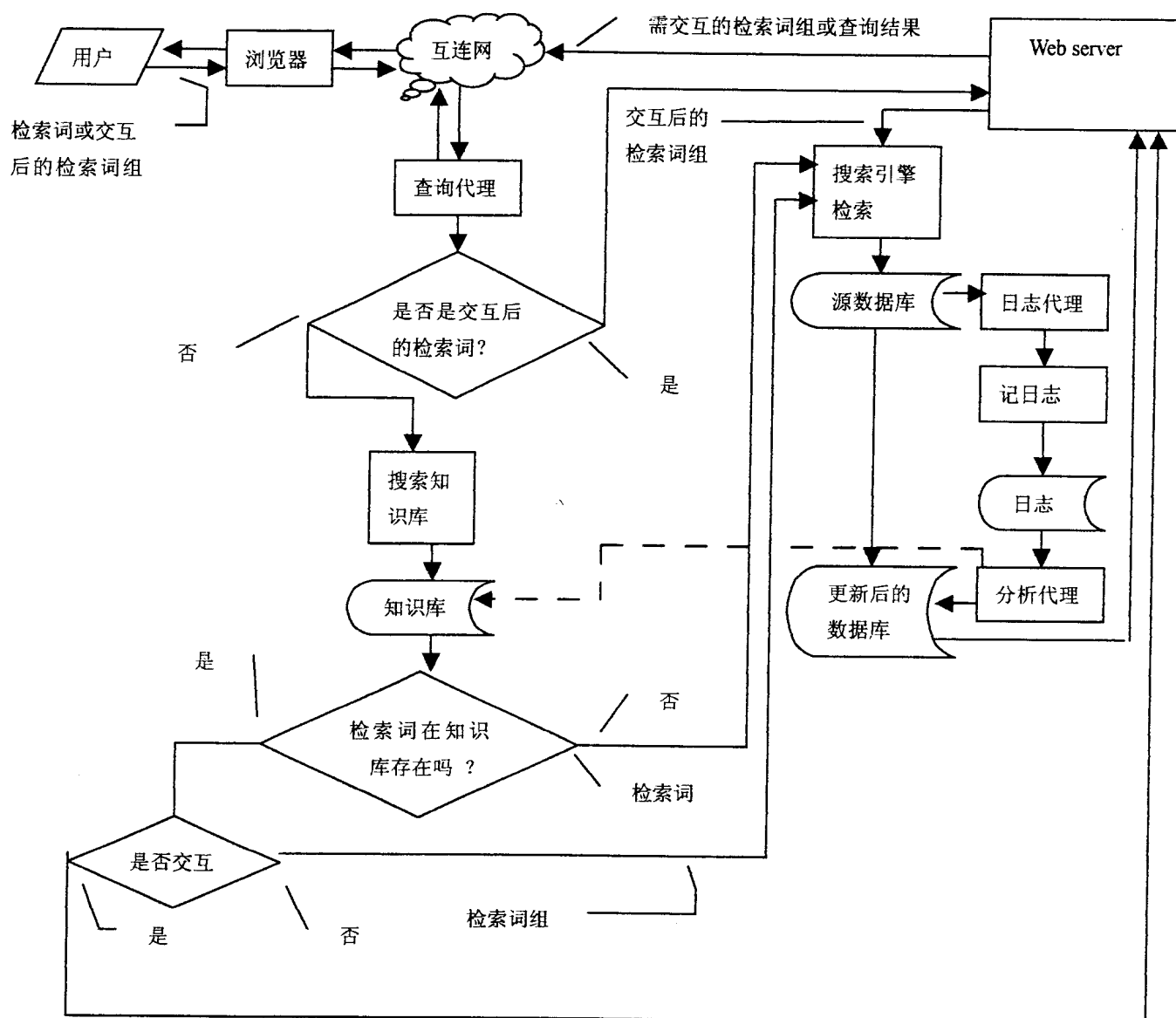


图 4:

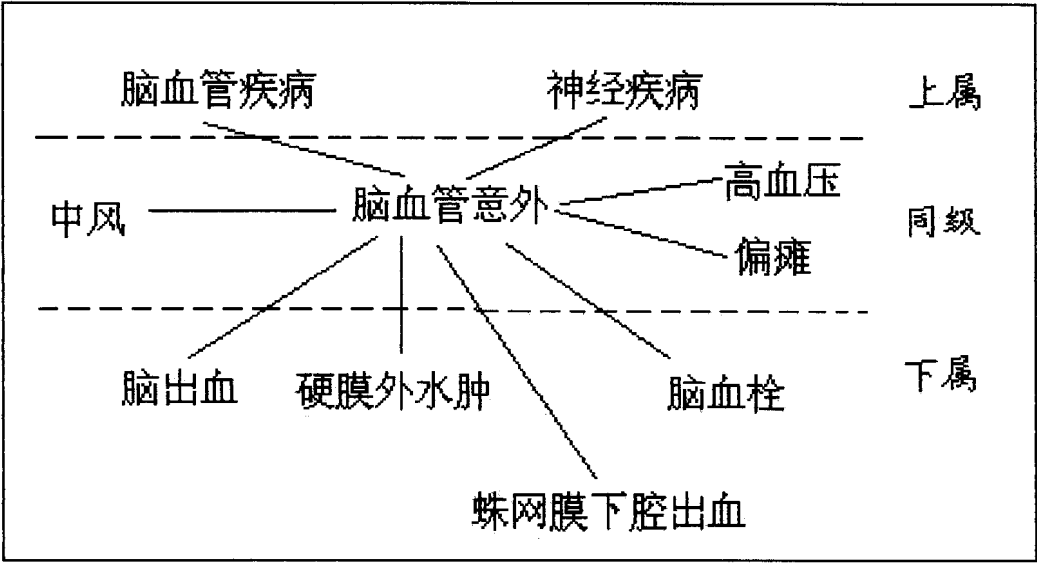


图 5:

检索词	编号
检索词 1	编号 1
.....

图 6 :

检索词	上属 1	C1	上属 2	C2	同级 1	C3	同级 2	C4	同级 3	C5	下属 1	C6	下属 2	C7
编号 1