

(12) 发明专利申请

(10) 申请公布号 CN 102236719 A

(43) 申请公布日 2011. 11. 09

(21) 申请号 201110207646. 3

(22) 申请日 2011. 07. 25

(71) 申请人 西交利物浦大学

地址 215123 江苏省苏州市工业园区独墅湖
高等教育区仁爱路 111 号

(72) 发明人 史玉回

(74) 专利代理机构 苏州创元专利商标事务所有
限公司 32103

代理人 范晴

(51) Int. Cl.

G06F 17/30(2006. 01)

权利要求书 2 页 说明书 5 页 附图 2 页

(54) 发明名称

基于网页分类的网页搜索引擎及快速查找方法

(57) 摘要

本发明公开了一种具有分类显示的搜索引擎及快速检索方法,该搜索引擎包括处于服务器端的分类模块,用于对每一网页按照中国图书馆图书分类法进行分类,将分类结果索引存入网页索引库;所述结果显示模块通过分栏显示与关键词相匹配的网页索引和与关键词相关的网页分类。该搜索引擎通过网页分类信息更好地帮助用户按网页类别更快速地、更准确地通过搜索引擎寻找到用户所感兴趣的网页。



1. 一种具有分类显示的搜索引擎,包括处于服务器端的:

网页抓取和预处理模块,用于自动从网络上搜集网页,进行预处理将网页信息转化成计算机可读方式的格式化文本信息,并定期实时更新网页信息和新网页信息抓取;

索引模块,用于对网页抓取和预处理模块处理后的格式化文本信息进行分词,并使每一网页与它所包含的分词及分词出现的频率建立具有关联度特征的网页索引库;

查询模块,用于响应用户端的查询请求,并搜索索引模块建立的网页索引库,获得与用户端的查询请求匹配的搜索结果列表;

和处于用户端的:

结果显示模块,用于供用户输入关键词查询请求,并从服务器端的查询模块获得与关键词相匹配的搜索结果列表,并按照关联度由大到小的顺序排列后展示给用户;

其特征在于所述搜索引擎还包括处于服务器端的分类模块,用于对每一网页按照中国图书馆图书分类法进行分类,将分类结果索引存入网页索引库;所述结果显示模块通过分栏显示与关键词相匹配的网页索引和与关键词相关的网页分类。

2. 根据权利要求1所述的具有分类显示的搜索引擎,其特征在于所述搜索引擎服务器端每个网页按照中国图书馆图书分类法进行分类。

3. 根据权利要求2所述的具有分类显示的搜索引擎,其特征在于所述分类采用人工分类或采用粒子群优化算法的机器学习分类。

4. 根据权利要求1所述的具有分类显示的搜索引擎,其特征在于所述结果显示模块包括两栏显示窗口,第一显示窗口用于显示与关键词相匹配的网页列表;第二显示窗口用于显示与关键词相关的网页分类。

5. 根据权利要求4所述的具有分类显示的搜索引擎,其特征在于所述两栏显示窗口呈左右设置,左侧为第二显示窗口,右侧为第一显示窗口;第一显示窗口聚焦当前显示网页列表项时,第二显示窗口相应网页类别加粗或加色。

6.

根据权利要求4所述的具有分类显示的搜索引擎,其特征在于所述第一显示窗口内网页列表项具有与相应网页网站链接的超链接,当用户通过鼠标停留在网页列表项时,相应网页列表项对应的网页类别在第二显示窗口内加粗或变色,相应网页列表项右侧呈现与相应网页列表项相关的网页快照。

7. 根据权利要求6所述的具有分类显示的搜索引擎,其特征在于当用户选择第一显示窗口内的网页列表项时,用户端将直接打开相应网页网站链接供用户浏览;当用户选择第二显示窗口内的网页分类时,第一显示窗口聚类显示用户选择的同一网页分类的网页列表。

8. 根据权利要求4所述的具有分类显示的搜索引擎,其特征在于所述第二显示窗口还包括网页更新时间选项,所述网页更新时间选项设置在网页分类下端供用户选择。

9. 一种基于网页分类的网页搜索结果的快速查找方法,其特征在于所述方法包括以下步骤:

(1) 接收用户端的查询请求;

(2) 响应查询请求,在索引有每一网页与它所包含的分词及分词出现的频率累计的关联度并对每一网页进行分类的网页索引库上执行查询以获得与查询请求匹配的搜索结果

列表；

(3) 将搜索列表按照关联度进行降序排列并将排序后的搜索列表和搜索列表的网页类别分成两栏生成网页呈现给用户。

基于网页分类的网页搜索引擎及快速查找方法

技术领域

[0001] 本发明属于搜索引擎优化技术领域,具体涉及一种基于网页分类的网页搜索引擎及快速查找方法。

背景技术

[0002] 随着互联网技术的快速发展,每天都有大量新的网页出现,互联网网页数量急剧增长。互联网网页中包含了丰富的信息,怎样从大量网页所包含的海量信息中快速地搜索到所感兴趣的信息就显得尤为重要。如果不能从互联网网页中在有限的、可容忍的时间内寻找到有用的信息,互联网的发展就不会那么迅猛,影响不会那么广泛。

[0003] 互联网搜索引擎从互联网网页中提取网页包含的信息,并将这些信息存入数据库。当用户在互联网搜索引擎搜索网页上输入关键词后,搜索引擎从数据库中寻找出与输入关键词相关的网页提供给用户。最早的搜索引擎可追溯到1990年由加拿大McGill大学三名学生设计实现的Archie系统。那时还没有所谓的互联网,Archie系统不是现在意义上的搜索引擎。用户通过输入文件名可用Archie系统搜索到哪一个FTP服务器上拥有可供下载的具有该文件名的文件。随着互联网的出现,陆续出现了许多不同意义上的互联网搜索引擎如WebCrawler,Excite,Infoseek,AltaVista,Yahoo等等。WebCrawler是由美国华盛顿大学Brian Pinkerton于1994年实现的第一个全文搜索引擎。它能搜索任何网页上任何词。而Yahoo实际上更准确地说是属于目录式搜索引擎而不是全文搜索引擎。它的分类目录网页上收集了大量分了类的网站。网页拥有者可自己将拥有的网页按类注册在分类目录网页上。其他用户可根据分类目录从分类目录网页上寻找到感兴趣的网页。其它的网页分类目录网页还有如新浪分类目录网页、DMOZ(www.dmoz.org)等等。大量的商业用户为了更好地利用互联网来推销它们的产品,使他们的产品信息网页更优先地出现在搜索引擎的搜索结果中,开始研究分类目录网页上网站的排名规则,并试图通过一些方法去调整自己网站在分类目录网页上的排名来使自己网站更易被搜索引擎搜索到,并相应地出现搜索结果更靠前的位置,这就出现了搜索引擎优化。另外在搜索引擎网页抓取时也常从这些分类目录网页出发去抓取网页,包含在这些分类目录网页上的网页,更易被搜索引擎抓取,从而显示在搜索引擎的搜索结果中。这些分类目录网页不属于真正的搜索引擎,但这些分类目录网页可被人为地利用来操纵搜索引擎网页搜索结果。不同于目录式搜索引擎,谷歌则采用了PageRank技术,将每一网页被其它网页(特别是一些常被访问的网页)链接的情况结合到搜索引擎搜索到的结果的优先级别里,从而被其它网页链接越多的网页越容易被搜索引擎搜索到。这使得搜索引擎的搜索结果与输入关键词的相关性大大提高。商用的搜索引擎如谷歌、百度等采用了这样的原理。

[0004] 搜索引擎一般包含以下几步:网页抓取、网页分词、网页索引、和网页搜索。每一网页与它所包含的分词及分词出现的频率建立关联度,存入索引数据库供搜索引擎搜索时使用。这样当用搜索引擎搜索时,与输入的关键词相关联的网页一般按照关联度的程度排列来作为搜索结果。最相关的网页则排列在搜索结果的前面,越容易被用户搜索到。用户在

使用搜索引擎搜索时,一般来说输入的关键词有限,二到三个关键词的情况是很普遍的。这样通过网页搜索引擎搜索到的网页结果里常常不一定是用户所真正想要搜索到的网页。经常用户通过一页一页地浏览搜索到的网页而不能找到真正想要搜索到的网页。为了更好地提供给用户真正需要的搜索结果,搜索引擎公司采用了一些方法来提供更好的搜索结果给用户。百度搜索引擎、比应搜索引擎可将搜索结果按图片、视频、新闻等来输出搜索结果,使用户在图片、视频、新闻这些类别里能更好地更准确地搜索到用户所感兴趣的网页。谷歌搜索引擎、有道搜索引擎则不光包含这些,同时也可将搜索结果网页按时间进行分类。如只显示过去一天内、一星期内、一月内或一年内更新过的网页。另外,搜索引擎还通过在每一显示的搜索结果网页旁加上链接指向与该显示结果网页相类似的网页,或者通过给出每一搜索结果网页的摘要或快照,这样用户就不用去每一网页浏览就能知道该网页是否包含用户感兴趣的结果。这些方法提高了提供给用户更好更准确的搜索结果的可能性,但搜索引擎显示的结果网页仍然可能包含很多不感兴趣的搜索结果,需要对这些信息进行过滤从而尽可能地仅提供给用户感兴趣的网页。

发明内容

[0005] 本发明目的在于提供一种具有分类显示的搜索引擎,解决了现有技术中搜索引擎的显示结果常常没有实现用户搜索的目的或者检索的信息太过繁杂使用户难以找到准确的信息等问题。

[0006] 为了解决现有技术中的这些问题,本发明提供的技术方案是:

[0007] 一种具有分类显示的搜索引擎,包括处于服务器端的:

[0008] 网页抓取和预处理模块,用于自动从网络上搜集网页,进行预处理将网页信息转化成计算机可读方式的格式化文本信息,并定期实时更新网页信息和新网页信息抓取;

[0009] 索引模块,用于对网页抓取和预处理模块处理后的格式化文本信息进行分词,并使每一网页与它所包含的分词及分词出现的频率建立具有关联度特征的网页索引库;

[0010] 查询模块,用于响应用户端的查询请求,并搜索索引模块建立的网页索引库,获得与用户端的查询请求匹配的搜索结果列表;

[0011] 和处于用户端的:

[0012] 结果显示模块,用于供用户输入关键词查询请求,并从服务器端的查询模块获得与关键词相匹配的搜索结果列表,并按照关联度由大到小的顺序排列后展示给用户;

[0013] 其特征在于所述搜索引擎还包括处于服务器端的分类模块,用于对每一网页按照中国图书馆图书分类法进行分类,将分类结果索引存入网页索引库;所述结果显示模块通过分栏显示与关键词相匹配的网页索引和与关键词相关的网页分类。

[0014] 优选的,所述搜索引擎服务器端每个网页按照中国图书馆图书分类法进行分类。

[0015] 优选的,所述分类采用人工分类或采用粒子群优化算法的机器学习分类。

[0016] 优选的,所述结果显示模块包括两栏显示窗口,第一显示窗口用于显示与关键词相匹配的网页列表;第二显示窗口用于显示与关键词相关的网页分类。

[0017] 优选的,所述两栏显示窗口呈左右设置,左侧为第二显示窗口,右侧为第一显示窗口;第一显示窗口聚焦当前显示网页列表项时,第二显示窗口相应网页类别加粗或加色。

[0018] 优选的,所述第一显示窗口内网页列表项具有与相应网页网站链接的超链接,当

用户通过鼠标停留在网页列表项时,相应网页列表项对应的网页类别在第二显示窗口内加粗或变色,相应网页列表项右侧呈现与相应网页列表项相关的网页快照。

[0019] 优选的,当用户选择第一显示窗口内的网页列表项时,用户端将直接打开相应网页网站链接供用户浏览;当用户选择第二显示窗口内的网页分类时,第一显示窗口聚类显示用户选择的同一网页分类的网页列表。

[0020] 优选的,所述第二显示窗口还包括网页更新时间选项,所述网页更新时间选项设置在网页分类下端供用户选择。

[0021] 本发明还提供了一种基于网页分类的网页搜索结果的快速查找方法,其特征在于所述方法包括以下步骤:

[0022] (1) 接收用户端的查询请求;

[0023] (2) 响应查询请求,在索引有每一网页与它所包含的分词及分词出现的频率累计的关联度并对每一网页进行分类的网页索引库上执行查询以获得与查询请求匹配的搜索结果列表;

[0024] (3) 将搜索列表按照关联度进行降序排列并将排序后的搜索列表和搜索列表的网页类别分成两栏生成网页呈现给用户。

[0025] 本发明能更好地提供给用户更准确的用户所需的搜索结果。通过将搜索引擎搜索结果通过网页分类信息对搜索结果按网页类别显示搜索结果,从而提供给用户更准确的搜索结果,使用户能更快速地、更好地搜索到用户感兴趣的网页。这种基于网页分类的网页搜索结果的快速查找方法可与现有的商用搜索引擎结合起来提供用户从这些商用搜索引擎提供的网页搜索结果中快速查找到所需要的信息。

[0026] 在对网页抓取、分词、索引、及对每一网页与它所包含的分词及分词出现的频率建立关联度后,对每一网页按照中国图书馆图书分类法进行分类,每一网页可对应于一或几种类别,将所有信息存入数据库中,待网页搜索时用。

[0027] 网页搜索时,搜索引擎显示页面分成左右两格,左边显示网页分类信息,右边显示搜索结果。当用户输入关键词后初始搜索结果显示在网页右边。排列在第一的网页的所属的网页类别在左边显示。左边同时包含返回初始搜索结果及页数的直接链接。选择右边显示的搜索结果里其它的网页则左边将显示该网页所对应的网页类别。双重选择右边的任一网页则会打开这一网页供用户浏览这一网页。选择左边显示的网页分类类别,则右边将会显示搜索引擎的所有搜索结果中属于这一网页类别的搜索结果。这样右边显示的网页链接都属于用户所感兴趣的网页类别的搜索结果,用户可更快速地、更好地、更准确地寻找到用户所需要寻找的信息。当然目前搜索引擎所用的一些方法可被结合起来一起更好地为用户服务,如通过给出每一搜索结果网页的摘要或快照等等。

[0028] 在对网页抓取、分词、索引、及对每一网页与它所包含的分词及分词出现的频率建立关联度的同时,按照图一所示加入对每一网页按照中国图书馆图书分类法进行分类这一步,存入数据库。本专利采用的中国图书馆图书分类法可同样用其它网页分类法替代。网页分类可采用人工的方法,也可采用机器学习的方法如采用粒子群优化算法来对网页进行分类。

[0029] 在用搜索引擎搜索网页时,对搜索引擎给出的网页搜索结果按照图二所示进行显示。搜索引擎搜索结果显示在右边,左边则通过加粗或加色对应网页类别来显示网页类别。

左边可在上部同时显示初始搜索结果及页数的直接链接。显示初始搜索结果及页数的直接链接是为了右边可随时直接回到显示搜索引擎搜索结果,如选择第二页,则右边显示直接来自搜索引擎搜索结果显示的第二页。搜索结果第一次显示时,左边显示排列在第一的网页的所属的网页类别。如在右边显示的网页链接中选择其中的一个网页,则该网页对应的网页类别则显示在左边。选择左边的类别,则右边将会显示搜索引擎的所有搜索结果中属于这一网页类别的所有网页的链接。双重选择右边一网页链接则会直接打开这一网页供用户浏览。如用户知道所要搜索网页对应的类别,则用户可通过左边网页分类类别直接浏览到对应的网页分类类别来选择该网页类别。选择这一网页类别后,右边将仅显示搜索引擎搜索结果中属于该网页类别的网页链接。这样用户可快速直接查找自己感兴趣的网页。

[0030] 目前搜索引擎所用的一些方法可同时被结合起来一起更好地为用户服务,如通过给出每一搜索结果网页的摘要或快照,和只显示过去一天内、一星期内、一月内或一年内更新过的网页等等。

[0031] 相对于现有技术中的方案,本发明的优点是:

[0032] 本发明可通过网页分类信息更好地帮助用户按网页类别更快速地、更准确地通过搜索引擎寻找到用户所感兴趣的网页。

附图说明

[0033] 下面结合附图及实施例对本发明作进一步描述:

[0034] 图 1 为本发明具有分类显示的搜索引擎在服务器端的工作流程图;

[0035] 图 2 为本发明具有分类显示的搜索引擎在用户端的工作流程图。

具体实施方式

[0036] 以下结合具体实施例对上述方案做进一步说明。应理解,这些实施例是用于说明本发明而并不限于限制本发明的范围。实施例中采用的实施条件可以根据具体厂家的条件做进一步调整,未注明的实施条件通常为常规实验中的条件。

[0037] 实施例

[0038] 本发明具有分类显示的搜索引擎,服务器端包括网页抓取和预处理模块,用于自动从网络上搜集网页,进行预处理将网页信息转化成计算机可读方式的格式化文本信息,并定期实时更新网页信息和新建网页信息抓取;索引模块,用于对网页抓取和预处理模块处理后的格式化文本信息进行分词,并使每一网页与它所包含的分词及分词出现的频率建立具有关联度特征的网页索引库;查询模块,用于响应用户端的查询请求,并搜索索引模块建立的网页索引库,获得与用户端的查询请求匹配的搜索结果列表;分类模块,用于对每一网页按照中国图书馆图书分类法进行分类,将分类结果索引存入网页索引库;

[0039] 如图 1 和图 2 所示,网页抓取用新浪分类目录网页 (<http://dir.iask.com/>) 作为起始网页,依据网页上的超链接列表,不断有序抓取新的网页,并将每一抓取过的网页上新的超链接加入超链接列表,供后续抓取新的网页提供网页链接。每一抓取到的网页在超链接列表中标注这次已被抓取过,避免重复抓取已抓取过的网页甚至进入死循环。对抓取的网页信息进行预处理,如去除掉 HTML 文件中标记符号,得到网页文本信息。第一次抓取时,所有抓取到的网页及它对应的网页文本信息存入数据库,供后续步骤使用。没有抓取到的

网页也同样存入数据库但标记网页不存在。第一次抓取后,以后定期(如每晚10时)对网页进行抓取。如网页信息有更新,则抓取该网页的信息处理后存入数据库,如没有更新则不抓取该网页信息,数据库中该网页内容不变。对于网页信息有更新的网页,检查它所含有的超链接有没有包含新的网页,如有则对这些新的网页加入超链接列表,进行新建网页信息抓取。在网页抓取这一过程中,记录每一网页被其它网页链接的次数,并存入数据库。最后对所有存入的网页根据网页信息按照中国图书馆图书分类法进行分类,并将分类类别存入数据库中。

[0040] 在数据库中对新抓取到的网页依据已建立的分词词典(如2词词典、3词词典、专有名词词典等等)进行分词,再根据每一分词在该网页中出现的频率建立该网页与每一出现的分词的关联度,从而可更进一步建立每一分词与包含该分词的所有网页的关联度,供网页搜索时依据关键词搜索网页用。这样当用户输入需要搜索的关键词后,查询模块将找出跟关键词相关的网页并将它们排序。网页排序则依据网页跟输入的关键词的关联度、相关联的关键词的数目及被其它网页链接的次数来确定。与网页关联的关键词越多、关联度越高及被其它网页链接的次数越多,则该网页排序越靠前。

[0041] 用户端包括结果显示模块,用于供用户输入关键词查询请求,并从服务器端的查询模块获得与关键词相匹配的搜索结果列表,并按照关联度由大到小的顺序排列后展示给用户;所述结果显示模块通过分栏显示与关键词相匹配的网页索引和与关键词相关的网页分类。

[0042] 服务器端每个网页按照中国图书馆图书分类法进行分类,分类采用人工分类。用户端结果显示模块包括两栏显示窗口,第一显示窗口用于显示与关键词相匹配的网页列表;第二显示窗口用于显示与关键词相关的网页分类。所述两栏显示窗口呈左右设置,左侧为第二显示窗口,右侧为第一显示窗口;第一显示窗口聚焦当前显示网页列表项时,第二显示窗口相应网页类别加粗或加色。所述第一显示窗口内网页列表项具有与相应网页网站链接的超链接,当用户通过鼠标停留在网页列表项时,相应网页列表项对应的网页类别在第二显示窗口内加粗或变色,相应网页列表项右侧呈现与相应网页列表项相关的网页快照。

[0043] 当用户选择第一显示窗口内的网页列表项时,用户端将直接打开相应网页网站链接供用户浏览;当用户选择第二显示窗口内的网页分类时,第一显示窗口聚类显示用户选择的同一网页分类的网页列表。所述第二显示窗口还包括网页更新时间选项,所述网页更新时间选项设置在网页分类下端供用户选择。

[0044] 进行查询时,服务器端先接收用户端的查询请求;然后响应查询请求,在索引有每一网页与它所包含的分词及分词出现的频率累计的关联度并对每一网页进行分类的网页索引库上执行查询以获得与查询请求匹配的搜索结果列表;最后将搜索列表按照关联度进行降序排列并将排序后的搜索列表和搜索列表的网页类别分成两栏生成网页呈现给用户。

[0045] 上述实例只为说明本发明的技术构思及特点,其目的在于让熟悉此项技术的人是能够了解本发明的内容并据以实施,并不能以此限制本发明的保护范围。凡根据本发明精神实质所做的等效变换或修饰,都应涵盖在本发明的保护范围之内。

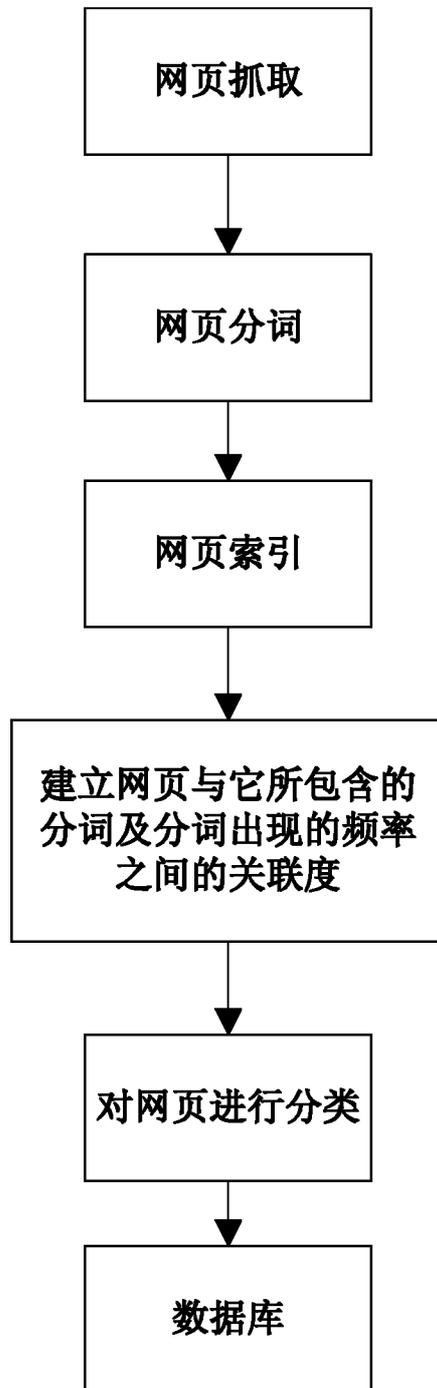


图 1

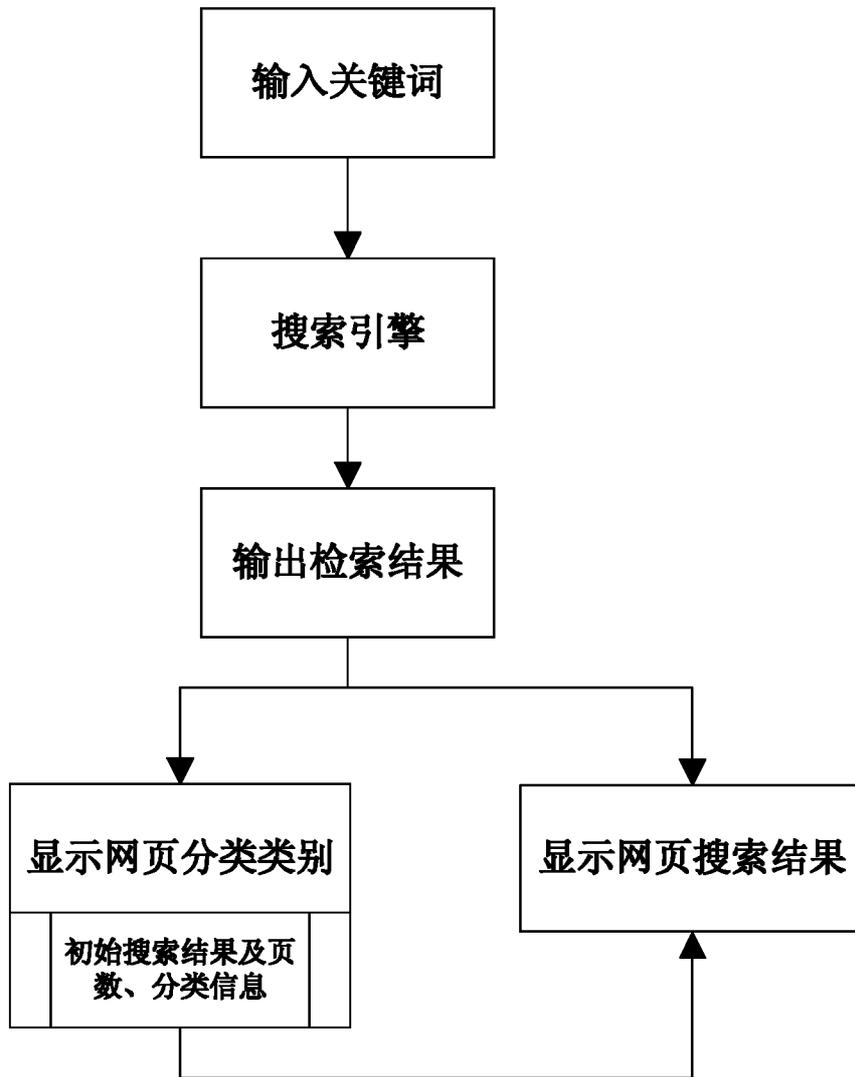


图 2