

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.

G06F 17/30 (2006.01)



[12] 发明专利申请公布说明书

[21] 申请号 200610007905.7

[43] 公开日 2007 年 8 月 29 日

[11] 公开号 CN 101025737A

[22] 申请日 2006.2.22

[21] 申请号 200610007905.7

[71] 申请人 王东

地址 100089 北京市海淀区板井路远大园 6
区 3 楼 3F 室

[72] 发明人 王东

[74] 专利代理机构 北京乾诚五洲知识产权代理有限公司

代理人 付晓青 刘毅

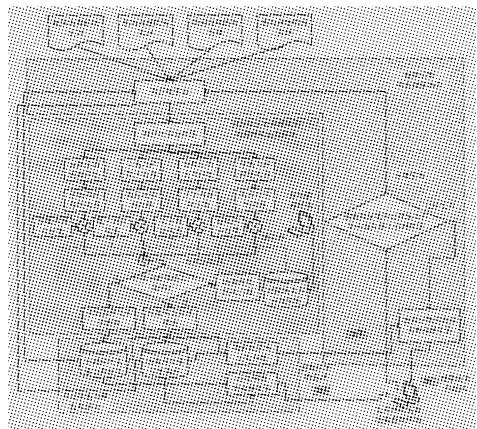
权利要求书 11 页 说明书 52 页 附图 12 页

[54] 发明名称

基于关注度的同源信息搜索引擎聚合显示方法及相关系统

[57] 摘要

本发明涉及一种基于关注度的同源信息搜索引擎聚合显示方法及系统。搜索引擎找到全部符合条
件的目标站点作为原始搜索结果；依据内容质量、
显示加权权力采购者的账户信息和服务质量等要
素，将原始搜索结果聚合成一条标题搜索结果；只
将标题搜索结果作为搜索结果展示给查询者，当其
需要时才为其展开查看全部搜索结果。本系统采用
统计服务器配合网络浏览器，将用户的全部操作行
为换算成对该网页的关注程度评分值 PageFocus，并
发送回统计服务器来代表其内容质量，从而可以作
为搜索引擎的选择“标题搜索结果”和进行结果显
示排名的方法。本发明还涉及一种能够自动判断用
户状态并提供恰当的网页风格与内容的方法。



1、一种同源信息站点搜索引擎聚合显示方法，其包括下列步骤：

- (1) 查询者通过 Web 浏览器或应用软件访问搜索引擎，并输入需要查询的关键词；
- (2) 由搜索引擎找到全部符合条件的目标站点作为原始搜索结果；
- (3) 由“同源信息处理模块”查询标题搜索结果权力采购者的账户信息，并结合其它判断规则在原始搜索结果中选取用来作为标题搜索结果的对象；
- (4) 由搜索引擎 Web 服务器或应用服务器只将选中的标题搜索结果作为搜索结果展示给查询者，并为其提供一个带有展开查看细节或其它信息含义的按钮；
- (5) 查询者还可按动与之对应的按钮，搜索引擎再向其展示在(2)中找到的原始搜索结果。

2、根据权利要求 1 所述的同源信息站点搜索引擎聚合显示方法，其特征在于，

所述“同源信息处理模块”包括如下步骤：

- (1) 由信息种类判断模块对网络搜索器收到的信息进行种类判断；
- (2) 将相同种类的信息集中发送到相应的“同源信息处理模块”；
- (3) 将由信息处理模块处理后的搜索信息返回 Web 服务器。

3、根据权利要求 2 所述的同源信息站点搜索引擎聚合显示方法，其特征在于，

所述由“同源信息处理模块”处理网页信息的步骤如下所示：

- (1) 在搜索引擎搜索部分接收需要查询的关键词的时候，首先由搜索结果已经发布在 Web 服务器上的判决器判断该关键词是否近期已经被其它人查询过，如果被查询过，并且结果已经在搜索引擎搜索结果 Web 服务器上发布，则直接返回搜索结果，该结果中已经将具有相同来源的网页聚合成一条搜索结果，点击“同源网页”按钮后，可以在搜索引擎搜索结果 Web 服务器上看到另一个

包括全部搜索结果的搜索结果网页，完成整个查询过程；

(2) 如果在搜索引擎搜索部分接收需要查询的关键词的时候，由搜索结果已经发布在 Web 服务器上的判决器判断该关键词近期没有被其它人查询过，并且也没有相应的查询结果在搜索引擎搜索结果 Web 服务器上发布则：

A. 启动“网页搜索器”搜索“非同源网页结果数据库”和“同源网页结果数据库”找到符合搜索关键词的网页地址，并获取这些网页的内容；

B. 如果“网页搜索器”在“非同源网页结果数据库”和“同源网页结果数据库”中没有找到符合搜索关键词的网页地址，则返回查询者“没有符合条件网页”的结果，并且将该搜索关键词加入到下一轮更新“非同源网页结果数据库”和“同源网页结果数据库”的任务中，如果在更新过程中找到了符合条件的网页地址则根据其是否具有同源网页而选择入“非同源网页结果数据库”或“同源网页结果数据库”，这样如果再有人搜索同样的关键词是就可以找到结果；

(3) 由“网页内容分离器”将找到的网页内容及超级链接目标分解成：多媒体、图片、文字、超级链接等种类；

(4) 分别由各种内容判决器产生判决结果：

A. 由“多媒体内容判决器”产生目标网页所含“相同多媒体文件程度 SMS (Same Media Score)”；

B. 由“图片内容判决器”产生目标网页所含“相同图片的程度 SPS (Same Photo Score)”；

C. 由“文字内容判决器”产生目标网页所含“相同文字的程度 STS (Same Text Score)”；

D. 由“链接内容判决器”产生目标网页所含“相同超级连接的程度 SHS (Same Hyperlinks Score)”；

(5) 从“同源网页判决规则库”分别获取“多媒体判决权重 SMP”、“图片判决权重 SPP”、“文字判决权重 STP”、“链接判决权重 SHP”并分别与第(4)步生成的“相同多媒体文件程度 SMS”、“相同图片的程度 SPS”、“相同文字的程度 STS”、“相同超级连接的程度 SHS”做数学乘法；

(6) 将第(5)步获得的数学乘法结果做加法，获得网页的“同源程度 SSS

(Same Sourc Score)", 同源程度 $SSS = (SMS*SMP) + (SPS*SPP) + (STS*STP) + (SHS*SHP)$;

(7) 判断该网页的“同源程度 SSS”是否超出门限，如果超出门限则判定为与其它网页的“同源网页”，如果没有超出门限则判定为“非同源网页”；

(8) 将第(7)步产生的“非同源网页”由“非同源网页处理模块”入“非同源网页结果数据库”；将第(7)步产生的“同源网页”由“同源网页处理模块”入“同源网页结果数据库”；

(9) 由“搜索结果网页发布器”根据“同源网页结果数据库”和“非同源网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果 Web 服务器”，再通过浏览器呈现给查询用户；

(10) 作为第(9)步的另一种实现方法，也可以通过“动态网页 Web 服务器”直接通过浏览器呈现给查询用户。

4、根据权利要求 2 所述的同源信息站点搜索引擎聚合显示方法，其特征在于，

所述“同源信息处理模块”包括如下步骤：

(1) 在收到查询者的搜索关键词，并通过软件根据关键词内容和关键语法判断需要找的文件或服务；

(2) 判断“要搜索的内容已经发布在 Web 服务器上吗？”，如果搜索的目标已经发布在“搜索引擎搜索结果 Web 服务器”上则直接返回搜索结果，该结果中已经将符合搜索条件具有相同来源的多媒体的获取接口聚合成一条“标题搜索结果”，点击“同源文件”按钮后，可以在“搜索引擎搜索结果 Web 服务器”上看到另一个包括全部搜索结果的网页，使查询者可以看到符合查询条件的全部搜索结果，完成搜索过程；如果搜索的目标没有发布在“搜索引擎搜索结果 Web 服务器”上则从第(3)步开始；

(3) 返回查询者没有符合条件的结果；

(4) 将该搜索关键词加入到下一轮更新“同源信息索引数据库”和“非同源信息索引数据库”的任务中，并定期启动两个数据库的更新过程；

(5) “同源信息索引数据库”和“非同源信息索引数据库”的更新过程：

A. 由搜索器搜索网页新出现的目标文件或服务入口，通过软件进入该入口获取该文件或服务；

B. 由“内容判决器”判断新找到的信息“与当前“同源信息索引数据库”的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同源信息索引数据库”的该类别；如果“否”则由“内容判决器”判断它“与当前非同源信息索引数据库”的内容属于同一内容吗？”；

C. 如果“是”则：“为当前的信息和与之同源的并已经存贮在‘非同源信息索引数据库’中的信息，新建一个类别并全部转移到‘同源信息索引数据库’”；如果“否”则“为当前的信息新建一个类别，并存入‘非同源信息索引数据库’”；

(6) 由“搜索结果网页发布器”根据“同源网页结果数据库”和“非同源网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果 Web 服务器”，再通过浏览器呈现给前来搜索的查询者；

(7) 作为第(6)步的另一种实现方法，也可以通过“动态网页 Web 服务器”直接通过浏览器呈现给查询用户。

5、根据权利要求4所述的同源信息站点搜索引擎聚合显示方法，其特征在于，所述由“同源信息处理模块”处理文档时，“同源信息索引数据库”和“非同源信息索引数据库”的更新过程为：

(1) 由“文档搜索器”搜索网页新出现的文档文件或链接入口，通过软件进入该入口获取该文件或服务；

(2) 由“文字内容判决器”和“图片内容判决器”判断新找到的文档内容“与当前‘同源文档索引数据库’的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同源文档索引数据库”的该类别；如果“否”则由“文档内容判决器”判断它“与当前非同源文档索引数据库”的内容属于同一内容吗？”；

(3) 如果“是”则：“为当前的文档和与之同源的并已经存贮在‘非同源文档索引数据库’中的文档，新建一个类别并全部转移到‘同源文档索引数据库’”；如果“否”则“为当前的文档新建一个类别，并存入‘非同源文档索引数据库’”。

6、根据权利要求3、4或5任一项权利要求所述的同源信息站点搜索引擎聚

合显示方法，其特征在于，所述相关内容判决器模块包括如下步骤：

(1) 接收“被判断对象”：可以接收多个来源的多媒体，并记录被判断对象的数量 **InputQuantity**；

(2) 查找“被判断对象”既定的可参与比对的属性，记录当前属性具有相同值的“被判断对象”的数量 **SameQuantity**；

(3) 输入当前属性在判断过程中的“权重”值 **Power**；

(4) 计算被全部“被判断对象”在当前属性上的吻合度：**PSame**=
SameQuantity* **Power**；

(5) 返回(1)对下一个“属性”执行(1)~(4)，得到该属性的 **PSame**，直至获得全部属性的 **PSame** 值；

(6) 计算并返回“被判断对象”的相同内容程度值：**SameMediaPower**=
(全部 **Psame** 值的数学累加值) / **InputQuantity**。

7、根据权利要求3、4或5任一项权利要求所述的一项权利要求所述的同源信息站点搜索引擎聚合显示方法，其特征在于，内容判决器模块为文字内容判决器时，其包括如下步骤：

(1) 找出文字内容中具有相同的单词或句子的部分的总计长度值 **SameLenth**；

(2) 找出输入的多个文字内容中，长度最短的输入文字的长度值 **MinLenth**；

(3) 返回文字相似程度值 **SameTextPower**=**SameLenth** / **MinLenth**。

8、根据权利要求3、4任一项权利要求所述的同源信息站点搜索引擎聚合显示方法，其特征在于，内容判决器模块为链接内容判决器时，其包括如下步骤：

(1) 接收“被判断对象”：多组超级链接的 URL 地址；

(2) 统计“被判断对象”相似程度：**SameURLPower**=在每组超级链接均出现过的 URL 地址数量；

(3) 返回 **SameURLPower**。

9、根据权利要求4所述的同源信息站点搜索引擎聚合显示方法，其特征在于，内容判决器模块为商业信息内容判决器时，其包括如下步骤：

- (1) 比对参与比对的商业信息是否是相同的产品或服务，如果“不是”返回“不一致”，如果“是”进入第(2)步；
- (2) 判断参与比对的商业信息是否具有地理位置敏感性，如果“不是”返回判断结果“一致”，如果“是”则进行第(3)步；
- (3) 判断参与比对的商业信息的提供者是否处于相同的城市或区域，如果“不是”返回判断结果“不一致”，如果是返回判断结果“一致”。

10、根据权利要求1所述的同源信息站点搜索引擎聚合显示方法，其特征在于，

“标题搜索结果”选择的具体实现方法如下：

(1) 计算每个“同源搜索结果”成为“标题搜索结果”的概率权值 PW_n ：

$$PW_n = TP * PageFocus / (RespDelay - K)$$

n : 该搜索结果为第 n 条

当 $(RespDelay - K)$ 小于等于零时， $(RespDelay - K)$ 应取值为 1

$PageFocus$: 网页关注度值

$RespDelay$: 网页服务响应延迟

K : 服务响应常数：小于本值的服务延误将不被察觉，例如： K 设置为 50 毫秒

TP : 标题搜索结果权力

(2) 统计求和全部原始“同源搜索结果”的概率权值 PW_n 的总和： $PWall$ 全部概率权值；

(3) 计算每条“同源搜索结果”成为“标题搜索结果”的概率： $P_n = PW_n / Pwall$ ；

(4) 按照 P_n 值的概率，随着搜索者的访问动作，动态地随机选择“标题搜索结果”，呈现给搜索者。

11、根据权利要求10所述的同源信息站点搜索引擎聚合显示方法，其特征在于，

所述“标题搜索结果”的概率权值 PWn 的计算方法还可以是：

- a. PWn= (TP+PageFocus) / (RespDelay-K) 或,
- b. PWn= (TP+PageFocus) / RespDelay/K 或,
- c. PWn=TP*PageFocus/RespDelay/K。

12、根据权利要求 1 所述的同源信息站点搜索引擎聚合显示方法，其特征在于，所述“同源信息处理模块”：

- (1) 可以内嵌在搜索引擎中；
- (2) 可以放置在“搜索引擎”和“搜索引擎搜索结果 Web 服务器”之间；
- (3) 也可以作为预处理模块放置在“搜索引擎”和被搜索站点之间。

13、根据权利要求 1 所述的同源信息站点搜索引擎聚合显示方法，所述展开查看细节或其它信息含义的按钮可为超级连接或各种软件界面控件。

14、一种获取权利要求 1 所述搜索方法生成结果而必须具备的网页用户关注度 PageFocus 的系统，包括 PageFocus 网络服务器、PageFocus 网络浏览器，其特征在于：

- (1) PageFocus 网络服务器包括 PageFocus 浏览器 ID 注册服务器、PageFocusAccServer 网页关注统计服务器、PageFocus 浏览器在线升级服务器及数据加解密模块；
- (2) PageFocus 网络浏览器包括 PageFocus 浏览器 ID 注册模块、关注分值 PageFocus 计算模块。

15、根据权利要求 14 所述的获取网页用户关注度的系统，其工作步骤如下：

(1)“PageFocus 网络浏览器”，每个浏览器均在安装时具备全球唯一的 ID 标识号，或在使用时主动寻找网络上的“PageFocus 浏览器 ID 注册服务器”以获得全球唯一的 ID 标识号；

(2)“PageFocus 网络浏览器”具备具有常规网络浏览器，并将用户对浏览器的操作和对网页的操作，以及网页内容特征按照权重转换成网页的“关注分值 PageFocus”并形成“PageFocus 数据包”，以加密方式通过网络协议传递至本搜索引擎的“PageFocusAccServer 网页关注统计服务器”；

(3) “**PageFocusAccServer** 网页关注统计服务器”在收到全球的每一个“**PageFocus** 网络浏览器”发来的“**PageFocus** 数据包”后将其内部包含的“关注分值 **PageFocus**”累加到相应的网页上；

(4) “**PageFocusAccServer** 网页关注统计服务器”上包含的全球每一个网页的“关注分值 **PageFocus**”，这些信息可以通过各种处理方法形成：搜索引擎对网页排行依据、搜索引擎在具有相同内容搜索结果中选择可以作为“标题搜索结果”的依据、也可以直接公布出来作为“网页热门程度排行榜”的服务。

16、根据权利要求 15 所述的获取网页用户关注度的系统，其特征在于，**PageFocus** 数据包可以在浏览器彻底关闭该网页时形成，也可以定时形成，也可以累计到某个分值时再形成，以减少 **PageFocusAccServer** 网页关注统计服务器的计算压力。

17、根据权利要求 15 所述的获取网页用户关注度的系统，其特征在于，所述关注分值 **PageFocus** 按照下表所列权重形成：

浏览器操作行为	权重	说明
打开网页	1	用户有兴趣打开对网页
在网页停留小于 10 秒	-10	另人讨厌的网页
在网页停留小于 30 秒	-3	不兴趣或内容非常简单
在网页停留小于 1 分钟	-1	没什么兴趣或内容非常简单
在网页停留超过 1 分钟	3	普通的浏览
在网页停留超过 5 分钟	5	比较认真的浏览
文字阅读速度	100/ 文字阅读速度	1. “文字阅读速度”：可以通过判断用户使用键盘、鼠标滚轮、“滚动条拖动”的操作幅度、频率和网页内容字体大小来计算； 2. “文字阅读速度” = 被滚动的文字数量 /

浏览器操作行为	权重	说明
		滚动时间间隔； 3.“文字阅读速度”的单位为：字/每秒 4. 说明：1秒看100字说明很不认真，得1分；1秒看一个字说明很认真，得100分
网页内容中的图片被点击，而图片上的超级链接指向一个图片	3	说明该图片可能提供了“放大图片”，代表该图片可能是网页的主体内容，并且用户点击了它，说明用户想仔细看看放大的图片
网页内容中的图片被点击，而图片上的超级链接指向网站内一个网页	1	很可能是个“广告”图片
网页内容中的图片被点击，而图片上的超级链接指向网站外一个网页	0	很可能是个“广告”图片
网页内容中的超级链接被点击，而目标与当前网页处于同一目录下	2	很可能是同一篇文章的网页连载，本网页和被点击超级链接指向的网页均可获得本“权重”的得分，例如：“ http://www.yoogol.com/news/01.htm ”和“ http://www.yoogol.com/news/02.htm ”
网页内容中的超级链接被点击，而目标站外网页	0	和本网页没什么关系
网页内容中含有标题为“1234”等带有序号含义的文字，而目标均指向同一个URL目录的网页的超级链接	N	本网页和标题为“1234”等带有序号含义，并且目标均指向同一个 URL 目录的网页通常为同一篇文章的分页显示；文章的任何页面的得分（包括负数得分），其它页面即使没有被打开也可得到相同的得分

浏览器操作行为	权重	说明
用户使用浏览器右键菜单投票:“投票-10分”	PageFocus*0%	1 浏览器对网页的任何部分均可使用鼠标右键菜单，菜单中含有“投票 10 分”、“投票 5 分”、“投票 1 分”、“投票-1 分”、“投票-5 分”、“投票-10 分”的菜单选项；
用户使用浏览器右键菜单投票:“投票-5分”	PageFocus*10%	2 用户使用右键菜单投票当前网页背景时，其 PageFocus 得分 =“当前 PageFocus 得分” *权重；
用户使用浏览器右键菜单投票:“投票-1分”	PageFocus*50%	3 用户使用右键菜单投票当前网页各种网页元素时，该元素超级链接所指向的网页得到 PageFocus 得分 = 当前网页“当前 PageFocus 得分” *权重。
用户使用浏览器右键菜单投票:“投票10分”	PageFocus*10	
用户使用浏览器右键菜单投票:“投票5分”	PageFocus*5	
用户使用浏览器右键菜单投票:“投票1分”	PageFocus*1	

18、根据权利要求 15 所述的获取网页用户关注度的系统，其特征在于，所述 PageFocus 数据包包含 PageFocus 浏览器 ID、网页 URL 及网页 PageFocus 得分值字段。

19. 根据权利要求 15 或 17 所述的获取网页用户关注度的系统，其特征在于，具备“同源网页”的每一个网页在参与搜索引擎提供的网页排名过程中，可以使用每一个“同源网页”获得的用户关注度 PageFocus 分值的总和作为排名的依据，即：**A** 在“同源网页”的“标题搜索结果”在参与搜索引擎结果排名时可以采用每一个“同源网页”获得的用户关注度 PageFocus 的总和作为排名依据；
B “同源网页”内的每一个网页在参与搜索引擎结果排名时也可以采用其从属的“同源网页”的每一个网页获得的用户关注度 PageFocus 的总和作为排名依据。

20、一种自动判断用户状态并提供恰当的网页风格与内容的方法，其包括
如下步骤：

(1)：首先将服务器集群划分成“工作风格”和“个人和休闲风格”两大类，无论是静态页面还是动态页面，在向这两类服务器更新相同的内容的时候，自动产生两类风格，以便不同场合的用户在访问相同页面 URL 地址时看到不同的风格；

(2)：在“网站服务器集群入口”收到用户首次访问本网站网页的请求后，首先在访问协议里（或 IP 层协议里）获取其 IP 地址；

(3)：根据 IP 地址在“IP 地址属性数据库”中查询其 IP 地址是“工作场合 IP 地址”还是“私人或休闲场合的 IP 地址”，若是“工作场合 IP 地址”则进行第(4)步，若是“私人或休闲场合的 IP 地址”则进行第(5)步；

(4)：获取“工作场合 IP 地址”所处的地理位置，并得到该地理区域的行政时间，若是该 IP 地址所属区域正处于工作时间（周 1~5 的 8: 00~20: 00）则将其访问分配到服务器集群中的“工作风格服务器”上向其提供适合工作场合使用的页面服务，否则进行第(5)步；

(5)：则将其访问分配到服务器集群中的“个人和休闲风格服务器”上向其提供适合个人和休闲状态使用的页面服务。

基于关注度的同源信息搜索引擎聚合显示方法及相关系统

技术领域

本发明涉及计算机网络技术,特别是利用计算机在互联网或企业内部网上提供搜索服务的搜索引擎技术。本发明还涉及一种获取网页用户关注度的系统及网站内容风格自适应装置与方法。

背景技术

目前在 Internet 上存在着大量的“相同(或类似)来源的网页或网络服务”,例如:1由同一个人或组织写作的被大量复制的文章、观点、信息网页;2由同一个人或组织采访(或发布)的被大量复制的新闻报道网页;3由同一个人或组织在 BBS 论坛发言帖子的转贴;5由同一个人或组织产生的不同数据格式、压缩比例的多媒体文件;6由同一个人或组织产生的可执行程序、数据、设计文件;7其他方式产生的并被广泛复制的信息内容。这些“相同(或类似)来源的网页或网络服务”在目前的搜索引擎搜索结果中被一一列举,占据大量篇幅,内容却雷同,不便查询者的浏览。

目前各种搜索引擎和网页排行服务系统,均仅仅采用了点击流量和网页停留时间的方式来衡量网页的热门程度,而采取的方法主要:1) 搜索引擎类:依靠查询者对搜索结果的点击来计算网页的热门程度,例如 google、百度。2) ALEXA 网站排行类:依靠内嵌在浏览器上的工具条软件,把用户对超级链接的点击和网页停留时间发送回服务器(参数包括当前网页地址、页面打开时间),但是不包含其他评估方法。 Alexa 工作原理可参见:

http://www.singtaonet.com/it/it_sp/t20051110_43674.html,

<http://www.people.com.cn/GB/it/8219/41552/41597/3109586.html>。

目前各种网站可以划分为如下类别:

类别一:全部网站内容对任何用户在同一时刻均具有同样风格与内容(例如:新闻网站)。

类别二:可以根据用户的设定显示不同的风格与内容(例如: google 的新闻

网站)。

但是这些网站，不能在实时根据用户的不同状态给出不同的显示风格与内容。

发明内容

为了改进上述问题的不足，本发明提供这样一种搜索方法，其能把因内容相同而对搜索者具有相同使用价值的搜索结果聚合成一条记录，即标题搜索结果，根据需要再展开查看其他结果的装置和方法，从而避免“标题搜索结果”由于频繁被点击而导致目标服务器访问量过大而瘫痪的，把“标题搜索结果”点击自动分散到其他搜索结果目标上的装置与方法。本发明还提供了这样一种系统，其利用能够和网络上的统计服务器配合的网络浏览器，将用户的全部操作行为换算成对该网页的评分，并发送回统计服务器，作为对网页的关注程度的评分，从而可以作为搜索引擎的排名方法和工具。本发明还提供这样一种方法：利用各种可能获得的、有助于判断用户所处环境和状态的信息，在同一时刻、同一个网站内、甚至时同一个页面内，向不同状态的用户提供不同的显示风格和内容。

为了实现上述目的，一种将同源信息站点搜索引擎聚合显示的搜索方法，其包括下列步骤：

- (1) 查询者通过 Web 浏览器或应用软件访问搜索引擎，并输入需要查询的关键词；
- (2) 由搜索引擎找到全部符合条件的目标站点作为原始搜索结果；
- (3) 由“同源信息处理模块”查询“成为标题搜索结果”的权力采购者的账户信息，并结合其他判断规则在原始搜索结果中选取用来作为“标题搜索结果”的对象；
- (4) 由搜索引擎 Web 服务器或应用服务器只将选中的“标题搜索结果”作为搜索结果展示给查询者，并为其提供一个带有“展开查看细节或其他信息”含义的“按钮”；
- (5) 查询者还可按动与之对应的“按钮”，搜索引擎再向其展示在(2)中找到的原始搜索结果。

“同源信息处理模块”有多个“(相应信息种类的)同源信息处理模块”组成，例如：“同源网页处理模块”、“同源多媒体处理模块”、“同源图片处理模块”、“同源文档处理模块”、“同源软件处理模块”、“同源数据或数据库处理模块”、“同源 GIS 信息处理模块”、“同价值网络服务处理模块”、“同价值商业信息处理模块”

等。

所述“同源信息处理模块”包括如下步骤：

- (1) 首先由“信息种类判断模块”对网络搜索器收到的信息进行种类判断；
- (2) 将相同种类的信息集中发送到“(相应信息种类的)同源信息处理模块”；
- (3) 将由“(相应信息种类的)同源信息处理模块”处理后的搜索信息归档进入“非同源(相应信息种类的)结果信息库”或“同源(相应信息种类的)结果信息库”。
- (4) 由系统把“非同源(相应信息种类的)结果信息库”和“同源(相应信息种类的)结果信息库”发布到 Web 服务器上，供查询者查询。作为另一中实现方法，也可以根据这两个数据库直接向查询者提供基于动态网页的查询服务。

所述由“同源网页处理模块”处理网页信息的步骤如下所示：

- (1) 在“搜索引擎 搜索部分”接收需要查询的关键词的时候，首先由“搜索结果已经发布在 Web 服务器上的判决器”判断该关键词是否近期已经被其他人查询过，如果被查询过，并且结果已经在“搜索引擎搜索结果 Web 服务器”上发布，则直接返回搜索结果，该结果中已经将具有相同来源的网页聚合成一条搜索结果，点击“同源网页”按钮后，可以在“搜索引擎搜索结果 Web 服务器”上看到另一个包括全部搜索结果的搜索结果网页，完成整个查询过程；
- (2) 如果在“搜索引擎 搜索部分”接收需要查询的关键词的时候，由“搜索结果已经发布在 Web 服务器上的判决器”判断该关键词近期没有被其他人查询过，并且也没有相应的查询结果在“搜索引擎搜索结果 Web 服务器”上发布则：

- A.** 启动“网页搜索器”搜索“非同源网页结果数据库”和“同源网页结果数据库”找到符合搜索关键词的网页地址，并获取这些网页的内容；
- B.** 如果“网页搜索器”在“非同源网页结果数据库”和“同源网页结果数据库”中没有找到符合搜索关键词的网页地址，则返回查询者“没有符合条件网页”的结果，并且将该搜索关键词加入到下一轮更新“非同源网页结果数据库”和“同源网页结果数据库”的任务中，如果在更新过程中找到了符合条件的网页地址则根据其是否具有同源网页而选择入“非同源网页结果数据库”或“同源网页结果数据库”，这样如果再有人搜索同样的关键词是就可以找到结果；

(3) 由“网页内容分离器”将找到的网页内容及超级链接目标分解成：多媒体、图片、文字、超级链接等种类；

(4) 分别由各种内容判决器产生判决结果：

A. 由“多媒体内容判决器”产生目标网页所含“相同多媒体文件程度 SMS (Same Media Score)”；

B. 由“图片内容判决器”产生目标网页所含“相同图片的程度 SPS (Same Photo Score)”；

C. 由“文字内容判决器”产生目标网页所含“相同文字的程度 STS (Same Text Score)”；

D. 由“链接内容判决器”产生目标网页所含“相同超级连接的程度 SHS (Same Hyperlinks Score)”；

(5) 从“同源网页判决规则库”分别获取“多媒体判决权重 SMP”、“图片判决权重 SPP”、“文字判决权重 STP”、“链接判决权重 SHP”并分别与第(4)步生成的“相同多媒体文件程度 SMS”、“相同图片的程度 SPS”、“相同文字的程度 STS”、“相同超级连接的程度 SHS”做数学乘法；

(6) 将第(5)步获得的数学乘法结果做加法，获得网页的“同源程度 SSS (Same

Sourc Score)”， $SSS = (SMS * SMP) + (SPS * SPP) + (STS * STP) + (SHS * SHP)$ ；

(7) 判断该网页的“同源程度 SSS”是否超出门限，如果超出门限则判定为与其它网页的“同源网页”，如果没有超出门限则判定为“非同源网页”；

(8) 将第(7)步产生的“非同源网页”由“非同源网页处理模块”入“非同源网页结果数据库”；将第(7)步产生的“同源网页”由“同源网页处理模块”入“同源网页结果数据库”；

(9) 由“搜索结果网页发布器”根据“同源网页结果数据库”和“非同源网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果 Web 服务器”，再通过浏览器呈现给查询用户；

(10) 作为第(9)步的另一种实现方法，也可以通过“动态网页 Web 服务器”直接通过浏览器呈现给查询用户。

所述由“同源信息处理模块”也可包括如下步骤：

- (1) 在收到查询者的搜索关键词，并通过软件根据关键词内容和关键词语法判断需要查找的文件或网络服务；
- (2) 判断“要搜索的内容已经发布在 Web 服务器上吗？”，如果搜索的目标已经发布在“搜索引擎搜索结果 Web 服务器”上则直接返回搜索结果，该结果中已经将符合搜索条件并具有相同来源的文件或网络服务的获取入口聚合成一条“标题搜索结果”，点击“同源文件”按钮后，可以在“搜索引擎搜索结果 Web 服务器”上看到另一个包括全部搜索结果的网页，使查询者可以看到符合查询条件的全部搜索结果，完成搜索过程。如果搜索的目标没有发布在“搜索引擎搜索结果 Web 服务器”上则从第（3）步开始；
- (3) 返回查询者“没有符合条件的结果”的提示；
- (4) 将该搜索关键词加入到下一轮更新“同源信息索引数据库”和“非同源信息索引数据库”的任务中，并定期启动两个数据库的更新过程；
- (5) “同源信息索引数据库”和“非同源信息索引数据库”的更新过程：
 - A. 由搜索器搜索网页新出现的目标文件或服务入口，通过软件进入该入口获取该文件或网络服务；
 - B. 由“内容判决器”判断新找到的信息“与当前“同源信息索引数据库”的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同源信息索引数据库”的该类别；如果“否”则由“内容判决器”判断它“与当前非同源信息索引数据库”的内容属于同一内容吗？”
 - C. 如果“是”则：“为当前的信息和与之同源的并已经存贮在‘非同源信息索引数据库’中的信息，新建一个类别并全部转移到‘同源信息索引数据库’”；
 - D. 如果“否”则：“为当前的信息新建一个类别，并存入‘非同源信息索引数据库’”；
- (6) 由“搜索结果网页发布器”根据“同源网页结果数据库”和“非同源网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果 Web 服务器”，再通过浏览器呈现给前来搜索的查询者；
- (7) 作为第（6）步的另一种实现方法，也可以通过“动态网页 Web 服务器”直接通过浏览器呈现给查询用户。

所述由同源信息处理模块处理文档时，“同源信息索引数据库”和“非同源信息索引数据库”的更新过程为：

- A. 由“文档搜索器”搜索网页新出现的文档文件或链接入口，通过软件进入该入口获取该文件或服务；
- B. 由“文字内容判决器”和“图片内容判决器”判断新找到的文档内容“与当前‘同源文档索引数据库’的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同源文档索引数据库”的该类别；如果“否”则由“文档内容判决器”判断它“与当前非同源文档索引数据库”的内容属于同一内容吗？”；
- C. 如果“是”则：“为当前的文档和与之同源的并已经存贮在‘非同源文档索引数据库’中的文档，新建一个类别并全部转移到‘同源文档索引数据库’”；如果“否”则“为当前的文档新建一个类别，并存入‘非同源文档索引数据库’”；

所述相关内容判决器模块包括如下步骤：

- (1) 接收“被判断对象”：可以接收多个来源的多媒体，并记录被判断对象的数量 InputQuantity；
- (2) 查找“被判断对象”既定的可参与比对的属性，记录当前属性具有相同值的“被判断对象”的数量 SameQuantity；
- (3) 输入当前属性在判断过程中的“权重”值 Power；
- (4) 计算被全部“被判断对象”在当前属性上的吻合度： $PSame = SameQuantity * Power$ ；
- (5) 返回(1)对下一个“属性”执行(1)~(4)，得到该属性的 PSame，直至获得全部属性的 PSame 值；
- (6) 计算并返回“被判断对象”的相同内容程度值： $SameMediaPower = (\text{全部 } PSame \text{ 值的数学累加值}) / InputQuantity$ 。

内容判决器模块为文字内容判决器时，其包括如下步骤：

- (1) 找出文字内容中具有相同的单词或句子的部分的总计长度值

SameLenth:

(2) 找出输入的多个文字内容中，长度最短的输入文字的长度值 MinLenth;

(3) 返回文字相似程度值 $\text{SameTextPower} = \text{SameLenth} / \text{MinLenth}$ 。

内容判决器模块为链接内容判决器时，其包括如下步骤：

(1) 接收“被判断对象”：多个超级链接的 URL 地址；

(2) 统计“被判断对象”相似程度： $\text{SameURLPower} = \text{在被判断的每个超级链接所指向的页面上均出现过的} \text{ URL 地址数量} / \text{被判断的 URL 地址数量}$ ；

(3) 返回 SameURLPower 。

内容判决器模块为商业信息内容判决器时，其包括如下步骤：

(1) 比对参与比对的商业信息是否是相同的产品或服务，如果“不是”返回“不一致”，如果“是”进入第（2）步。

(2) 判断参与比对的商业信息是否具有地理位置敏感性，如果“不是”返回判断结果“一致”，如果“是”则进行第（3）步。

(3) 判断参与比对的商业信息的提供者是否处于相同的城市或区域，如果“不是”返回判断结果“不一致”，如果是返回判断结果“一致”。

“标题搜索结果”选择的具体实现方法如下：

(1) 计算每个“同源搜索结果”成为“标题搜索结果”的概率权值 PWn :

$$PWn = TP * PageFocus / (RespDelay - K)$$

n: 该搜索结果为第 n 条

当 (RespDelay-K) 小于等于零时，(RespDelay-K) 应取值为 1

PageFocus: 网页关注度值

RespDelay: 网页服务响应延迟

K: 服务响应常数，建议 K 设置为 50 毫秒 (ms)。

TP: 标题搜索结果权力

(2) 统计求和全部原始“同源搜索结果”的概率权值 PWn 的总和：PWall 全部概率权值；

(3) 计算每条“同源搜索结果”成为“标题搜索结果”的概率：Pn

=PWn/Pwall;

(4) 按照 Pn 值的概率，随着搜索者的访问动作，动态地随机选择“标题搜索结果”，呈现给搜索者。

所述“标题搜索结果”的概率权值 PWn 的计算方法还可以是：

- a. PWn= (TP+PageFocus) / (RespDelay-K) 或，
- b. PWn= (TP+PageFocus) /RespDelay/K 或，
- c. PWn=TP*PageFocus/RespDelay/K。

所述“同源信息处理模块”：

- A. 可以内嵌在搜索引擎中；
- B. 可以放置在“搜索引擎”和“搜索引擎搜索结果 Web 服务器”之间；
- C. 也可以作为预处理模块放置在“搜索引擎”和被搜索站点之间。

所述展开查看细节或其他信息含义的按钮可为超级连接或各种软件界面控件。

一种获取网页用户搜索结果关注度的系统，包括 PageFocus 网络服务器、PageFocus 网络浏览器及网页计分服务器，

PageFocus 网络服务器包括 PageFocus 浏览器 ID 注册服务器、PageFocusAccServer 网页关注统计服务器、PageFocus 浏览器在线升级服务器及数据加解密模块；

PageFocus 网络浏览器包括 PageFocus 浏览器 ID 注册模块、关注分值 PageFocus
计算模块。

其工作步骤如下：

(1) “PageFocus 网络浏览器”，每个浏览器均在安装时具备全球唯一的 ID 标识号，或在使用时主动寻找网络上的“PageFocus 浏览器 ID 注册服务器”以获得全球唯一的 ID 标识号；

(2) “PageFocus 网络浏览器”具备具有常规网络浏览器，并将用户对浏览器的操作和对网页的操作按照权重转换成网页的“关注分值 PageFocus”并形成

“PageFocus 数据包”，以加密方式通过网络协议传递至本搜索引擎的“PageFocusAccServer 网页关注统计服务器”；

(3) “PageFocusAccServer 网页关注统计服务器”在收到全球的每一个“PageFocus 网络浏览器”发来的“PageFocus 数据包”后将其内部包含的“关注分值 PageFocus”累加到相应的网页上；

(4) “PageFocusAccServer 网页关注统计服务器”上包含的全球每一个网页的“关注分值 PageFocus”，这些信息可以通过各种处理方法形成：搜索引擎对网页排行依据、搜索引擎在具有相同内容搜索结果中选择可以作为“标题搜索结果”的依据、也可以直接公布出来作为“网页热门程度排行榜”的服务。

所述 PageFocusAccServer 网页关注统计服务器可以采用数学对数或科学计数法记录得分。

所述 PageFocus 数据包可以在浏览器彻底关闭该网页时形成，也可以定时形成，也可以累计到某个分值时再形成。

所述关注分值 PageFocus 按照下表所列权重形成：

浏览器操作行为	权重	说明
打开网页	1 (或 1.1、1.3、1.5)	用户有兴趣打开对网页
在网页停留小于 10 秒	-10 (或 7、8、9)	另人讨厌的网页
在网页停留小于 30 秒	-3 (或 -5、-4、-2)	不兴趣或内容非常简单
在网页停留小于 1 分钟	-1 (或 -3、-2、-1.5)	没什么兴趣或内容非常简单
在网页停留超过 1 分钟	3 (或 1、2、5)	普通的浏览
在网页停留超过 5 分钟	5 (或 3、4、5)	比较认真的浏览
文字阅读速度	100/ 文字阅读速度	1. “文字阅读速度”：可以通过判断用户使用键盘、鼠标滚轮、“滚动条拖动”的操作幅度、频率和网页内容字体大小来计算。

浏览器操作行为	权重	说明
		<p>2. “文字阅读速度” = 被滚动的文字数量 / 滚动时间间隔。</p> <p>3. “文字阅读速度”的单位为：字/每秒</p> <p>4. 说明：1 秒看 100 字说明很不认真，得 1 分；1 秒看一个字说明很认真，得 100 分。</p>
网页内容中的图片被点击，而图片上的超级链接指向一个图片	3 (或 1、2、4)	说明该图片可能提供了“放大图片”，代表该图片可能是网页的主体内容，并且用户点击了它，说明用户想仔细看看放大的图片。
网页内容中的图片被点击，而图片上的超级链接指向网站内一个网页	1 (或 0.1、0.2、0.9)	很可能是个“广告”图片。
网页内容中的图片被点击，而图片上的超级链接指向网站外一个网页	0 (或 0.1、0.2、0.3)	很可能是个“广告”图片。
网页内容中的超级链接被点击，而目标与当前网页处于同一目录下。	2 (或 0.5、1、2.5)	<p>很可能是同一篇文章的网页连载，本网页和被点击超级链接指向的网页均可获得本“权重”的得分。</p> <p>例如：“http://www.yoogol.com/news/01.htm”和“http://www.yoogol.com/news/02.htm”</p>
网页内容中的超级链接被点击，而目标站外网页。	0 (或 -0.5、-0.3、-0.1)	和本网页没什么关系

浏览器操作行为	权重	说明
网页内容中含有标题为“1 2 3 4”等带有序号含义的文字，而目标均指向同一个 URL 目录的网页的超级链接	N	本网页和标题为“1 2 3 4”等带有序号含义，并且目标均指向同一个 URL 目录的网页通常为同一篇文章的分页显示。文章的任何页面的得分（包括负数得分），其他页面即使没有被打开也可得到相同的得分。
用户使用浏览器右键菜单投票：“投票-10 分”	PageFocus*0%	1 浏览器对网页的任何部分均可使用鼠标右键菜单，菜单中含有“投票 10 分”、“投票 5 分”、“投票 1 分”、“投票-1 分”、“投票-5 分”、“投票-10 分”的菜单选项。
用户使用浏览器右键菜单投票：“投票-5 分”	PageFocus*10%	
用户使用浏览器右键菜单投票：“投票-1 分”	PageFocus*50%	2 用户使用右键菜单投票当前网页背景时，其 PageFocus 得分 =“当前 PageFocus 得分” *权重。
用户使用浏览器右键菜单投票：“投票 10 分”	PageFocus*10	3 用户使用右键菜单投票当前网页各种网页元素时，该元素超级链接所指向的网页得到 PageFocus 得分 = 当前网页“当前 PageFocus 得分” *权重
用户使用浏览器右键菜单投票：“投票 5 分”	PageFocus*5	
用户使用浏览器右键菜单投票：“投票 1 分”	PageFocus*1	

注释：

1 表格中的权重值是实施例，其它数值也可采用，均为本发明的范围。

所述文字阅读速度的计算步骤如下：

- A. 鼠标滚轮滚动：文字阅读速度 = (显示区宽度/字体宽度) * 每次滚动的文字行数/滚动时间间隔；
- B. 键盘翻页：文字阅读速度 = (显示区宽度/字体宽度) * 每次翻页的文字行数/翻页时间间隔；
- C. 窗体滚动条滚动：文字阅读速度 = (显示区宽度/字体宽度) * 每次滚动的文字行数/滚动时间间隔。

所述 PageFocus 数据包包含 PageFocus 浏览器 ID、网页 URL 及网页 PageFocus 得分值字段。

具备“同源网页”的每一个网页在参与搜索引擎提供的网页排名过程中，可以使用每一个“同源网页”获得的用户关注度 PageFocus 分值的总和作为排名的依据，即：A 在“同源网页”的“标题搜索结果”在参与搜索引擎结果排名时可以采用每一个“同源网页”获得的用户关注度 PageFocus 的总和作为排名依据；B “同源网页”内的每一个网页在参与搜索引擎结果排名时也可以采用其从属的“同源网页”的每一个网页获得的用户关注度 PageFocus 的总和作为排名依据。

一种自动判断用户状态并提供恰当的网页风格与内容的方法，其包括如下步骤：

- (1) 在“网站服务器集群入口”收到用户首次访问本网站网页的请求后，首先在访问协议里在或 IP 层协议里获取其 IP 地址；
- (2) 根据 IP 地址在“IP 地址属性数据库”中查询其 IP 地址是“工作场合 IP 地址”还是“私人或休闲场合的 IP 地址”，若是“工作场合 IP 地址”则进行第（3）步，若是“私人或休闲场合的 IP 地址”则进行第（4）步；
- (3) 获取“工作场合 IP 地址”所处的地理位置，并得到该地理区域的行政时间，若是该 IP 地址所属区域正处于工作时间，则将其访问分配到“工作风格服务器”上向其提供适合工作场合使用的页面服务，否则进行第（4）步；
- (4) 则将其访问分配到“个人和休闲风格服务器”上向其提供适合个人和休闲状态使用的页面服务。

通过上述方案，可把内容相同而对搜索者具有相同使用价值的搜索结果聚合成一条记录，即标题搜索结果，根据需要再展开查看其他结果的装置和方法。设

计了避免“标题搜索结果”由于频繁被点击而导致目标服务器访问量过大而瘫痪，把“标题搜索结果”点击自动分散到其他搜索结果目标上的装置。本发明除了具备现有搜索引擎外，还具备搜索各种“多媒体”、“文档”、“软件”、“软件硬件源代码或设计文件”、“数据或数据库”、“信息”的各种网络服务，例如文件共享、FTP 服务、P2P 服务等的功能。

利用能够和网络上的统计服务器配合的网络浏览器，将用户的全部操作行为换算成对该网页的评分，并发送回统计服务器，作为对网页的关注程度的评分，从而可以作为搜索引擎的排名工具。

通过网站内容风格自适应方法，用户可以：

- 1. 周 1~5 的早上 9: 00~18: 00 属于工作时间，处于工作状态的人需要看到简洁、相对严谨的风格和尽量与工作状态相关内容。**
- 2. 周 1~5 的晚上 18: 00~早上 9: 00 和周 6~7 的全天，属于休闲时间，处于休闲状态的人需要看到活波、热闹、休闲的风格和内容。**
- 3. 处于工作场所人需要看到简洁、相对严谨的风格和尽量与工作状态相关内容。**
- 4. 处于家庭和休闲场所的人需要看到活波、热闹、休闲的风格和内容。**
- 5. 处于其它环境或状态的人需要看到与当时的环境和状态相适应的的风格和内容。**

附图简要说明

图 1 为同源信息站点搜索引擎聚合显示方法的系统工作结构图；

图 2 为同源信息处理模块内部结构图；

图 3 为同源网页处理模块流程图；

图 4 为同源多媒体处理模块流程图；

图 5 为同源图片处理模块流程图；

图 6 为同源文档处理模块流程图；

图 7 为同源软件处理模块流程图；

图 8 为同源数据或数据库处理模块流程图；

图 9 为同源 GIS 信息处理模块流程图；

图 10 为同价值网络服务处理模块流程图；

图 11 为同价值商业信息处理模块流程图；

图 12 为获取网页用户关注度系统结构图；

图 13 为不具备内容和风格自适应技术的现有常规搜索引擎网站系统；

图 14 为本发明具备内容和风格自适应技术的搜索引擎网站系统。

具体实施方式

现结合附图对本发明做进一步的说明。

图 1 为同源信息站点搜索引擎聚合显示方法的系统工作结构图。第 1 步：由查询者通过 Web 浏览器或应用软件访问搜索引擎，并输入需要查询的关键词。第 2 步：由搜索引擎找到全部符合目标站点作为“原始搜索结果”。第 3 步：由“同源信息处理模块”查询“成为标题搜索结果”权力采购者的账户信息，并结合其他判断规则在“原始搜索结果”中选取用来作为“标题搜索结果”的对象：A “同源信息处理模块”可以内嵌在搜索引擎中；“同源信息处理模块”可以放置在“搜索引擎”和“搜索引擎搜索结果 Web 服务器”之间；C “同源信息处理模块”也可以作为预处理模块放置在“搜索引擎”和被搜索站点之间。第 4 步：由搜索引擎 Web 服务器或应用服务器只将选中的“标题搜索结果”作为搜索结果展示给查询者，并为其提供一个带有“展开查看细节或其他信息”含义的“按钮（包括超级连接或各种软件界面控件）”。第 5 步：只有查询者还希望进一步展开某条“标题搜索结果”，并按动与之对应的“按钮”时，搜索引擎再向其展示在“第 2 步”中找到的“原始搜索结果”。

图 2 为同源信息处理模块内部结构图。“同源信息处理模块”定义为：1) 主要用来判断按照搜索关键词找到的一组信息节点中是否有多个节点只是一个或多个同信息源的重复站点（这些站点对查询者具有相同搜索价值或使用价值，通常不必全部直接展现给查询者），并且将这些重复站点聚合成一条搜索结果发给查询者，只有查询者需要其他同等价值的站点时才将这些搜索结果呈现出来。2) 和现有搜索引擎主要集中于网页的搜索不同，“同源信息处理模块”除了需要处理“Html 网页”外还能够处理各种“多媒体”、“文档”、“软件”、“软件硬件源代码或设计文件”、“数据或数据库”、“信息”的各种网络服务，例如：文件共享、FTP 服务、P2P 服务等。

“同源信息处理模块”采用模块化结构，可以根据需要逐步开发和实施其中

的每一个模块，并且具备扩展能力，同时每一个模块也可以进一步加强其自动判断的准确性，其中包括：

1 “信息种类判断模块”：判断信息的种类，并将同类型信息集中发送到相应类型信息的处理模块，如下列模块。

2 “同源网页处理模块”：用来判断并处理找到的属于同一来源并对查询者具有相同价值的网页，例如：Html, ASP, JSP, PHP, BBS 论坛的内容等。

3 “同源多媒体处理模块”：用来判断并处理找到的属于同一来源，并对查询者具有相同价值的多媒体文件或网络服务，例如：.MP3,.AVI,.WMV,.MPEG,.WAV,.RM 等各种视频文件,以及各种基于流媒体技术的视频服务接入端口。

4 “同源图片处理模块”：用来判断并处理找到的属于同一来源或具有相同内容的，并对查询者具有相同价值的图片，例如：.GIF, .JPG, .BMP, .PNG 等。

5 “同源文档处理模块”：用来判断并处理找到的属于同一来源、具有相同或相关内容，并对查询者具有相同价值的各种格式文档文件或网络服务，例如：“.Doc”, “.Txt”, “.Pdf”, “.XLS”, “.PPT” 等。

6 “同源软件处理模块”：能够判断并处理找到的计算机应用软件安装程序属于同一作者的同一软件，它们可以是适应不同或相同操作系统的，相同或不同版本的软件安装程序。

7 “同源数据或数据库处理模块”：用来判断并处理找到的属于同一来源或具有相同内容的，并对查询者具有相同价值的，已知格式的数据文件或数据库文件，例如：.DAT, .XLS, .MDF, .DBF 等。

8 “同源 GIS 信息处理模块”：用来判断并处理找到的属于同一来源或具有相同内容的，并对查询者具有相同价值的数字地图文件或服务。

9 “同价值网络服务处理模块”：用来判断并处理找到的属于同一来源或具有相同内容的，并对查询者具有相同价值的网络服务，例如：相同文件的 FTP 下载服务，同时转播一个电视台的 IPTV 服务，同时提供 1GB 容量的邮件服务等。

10 “同价值商业信息处理模块”：用来判断并处理找到的属于同一来源或具有相同内容的，处于相同地理或行政区域的，并对查询者具有相同价值的，通过网络发布自己的商业产品或服务的广告内容，例如：在同一个街区提供的鸡蛋出售信息，在同一个街区提供的理发服务出售信息，在同一个城市可以使用的电话

通讯服务等。“信息种类判断模块”

“信息种类判断模块”主要用于在搜集到的信息中，分类出其类型，并送至相应的信息处理模块。

“信息种类判断模块”处理的信息来源主要有3种形式：

(1) 网页形式：信息来自于网站的网页内容，同时网页中还含有指向特定文件类型的超级链接，例如：“http://www.008.org.cn/up/the_quiet_american.mp3”

(2) 网络服务形式：包括各种网络服务器提供的网络服务入口，例如：FTP文件下载服务、各种P2P(Pear To Peer)软件(例如：BT下载、eMule下载)的种子服务，新闻服务器服务等。对于网络服务入口的获知可以有两种途径：

A. 网页上可以查到的网络服务：通过解析网页内容可以获知的网络服务入口。

B. 直接由网络服务提供者向本搜索引擎提交其网络服务入口或内容。

(3) 数据或数据库形式：由搜索引擎直接向网络提供信息录入服务，由网络用户提交自己的信息，最终形成数据文件或数据库形式的信息，在本搜索引擎被查询时，从中提取信息来满足查询者的要求。

“网页形式”信息的种类判断方法如下：

网页本身就可以直接作为“网页”输出给“同源网页处理模块”进行处理，另外，“信息种类判断模块”按照网页语法(例如：Html、Java、JSP、ASP、ASPX、PHP等等语言)针对“超级链接”的语法，可直接解析出其指向的文件类型，根据不同的文件类型可以区分出其信息类型，详见下表：

信息类型	超级链接内的文件类型
多媒体	“.avi”：视频文件、“.swf”：Flash文件、“.wav”：声音文件、“.mp3”：声音文件等等
图片	图片格式文件：“.gif”、“.bmp”、“.jpg”、“.png”等。
文档	文档格式文件：“.txt”、“.doc”、“.xls”、“.ppt”等。
软件	软件安装盘“.exe”、“.msi”等。
GIS 信息	GIS 文件：Mapinfo 公司的“.tab”等
压缩文件	常见压缩文件格式：“.rar”、“.zip”、

举例：

1. 网页中含有：“http://xxx/xxx/song.mp3”超级链接，即可判断其目标为“多媒体”类型信息。
2. 网页中含有：“http://xxx/xxx/song.rar”超级链接，找到该目标文件后解压缩，发现里面只含有“song.mp3”仍可判断目标为“多媒体”类型信息。
3. 网页中含有：“http://xxx/xxx/song.rar”超级链接，找到该目标文件后解压缩，发现里面含有的文件和目录的文件个数、每个文件的名称和大小均与某种已知软件的安装盘相同，即可判断其为“软件”类型信息。

“网络服务形式”信息的种类判断方法如下：

第1步：作为普通用户访问该服务，以获取其内容。

第2步：将获得的内容按照下表进行分类。

信息类型	超级链接内的文件类型
多媒体	“.avi”: 视频文件、“.swf”: Flash 文件、“.wav”: 声音文件、“.mp3”: 声音文件等等
图片	图片格式文件: “.gif”、“.bmp”、“.jpg”、“.png”等。
文档	文档格式文件: “.txt”、“.doc”、“.xls”、“.ppt”等。
软件	软件安装盘 “.exe”、“.msi” 等。
GIS 信息	GIS 文件: Mapinfo 公司的 “.tab” 等
压缩文件	常见压缩文件格式: “.rar”、“.zip”、

第3步：如果获得的是压缩格式文件，则需要展开其内容后在按照第2步进行分类。

“数据或数据库形式”信息的种类判断方法如下：

第1步：访问数据文件或数据库，以获取其内容。

第2步：如果从数据文件或数据库中获得的信息是文件则直接进行“第4步”。

第3步：如果从数据文件或数据库中获得的信息是存放文件的位置，则需要

访问该位置以获得目标文件。

第4步：将获得的内容按照下表进行分类。

信息类型	超级链接内的文件类型
多媒体	“.avi”: 视频文件、“.swf”: Flash 文件、“.wav”: 声音文件、“.mp3”: 声音文件等等
图片	图片格式文件: “.gif”、“.bmp”、“.jpg”、“.png” 等。
文档	文档格式文件: “.txt”、“.doc”、“.xls”、“.ppt” 等。
软件	软件安装盘 “.exe”、“.msi” 等。
GIS 信息	GIS 文件: Mapinfo 公司的 “.tab” 等
压缩文件	常见压缩文件格式: “.rar”、“.zip”、

第5步：如果获得的是压缩格式文件，则需要展开其内容后在按照4步进行分类。“同源网页处理模块”

图3为“同源网页处理模块”流程图。“同源网页处理模块”主要功能：将根据搜索关键词找到的，具有相同主要内容的网页，以一条“标题搜索结果”形式展现给查询者，并且通过“展开”含义按钮可以看到全部查询到的具有相同主要内容的网页的查询结果。为最大化地提高本系统的工作性能，我们采用了如下技术：

采用了网页发布技术，使用“搜索结果网页发布器”将搜索结果提前发布到“搜索引擎搜索结果 Web 服务器”，直接响应已经被查询过的搜索要求，避免根据请求动态从数据库生成动态网页的大量计算。

“同源信息处理模块”将处理结果分类别放置在“非同源网页结果数据库”和“同源网页结果数据库”中，并定期由“搜索结果网页发布器”发布到“搜索引擎搜索结果 Web 服务器”，避免了重复计算和减少了计算等待时间。

“同源信息处理模块”处理流程如下：

第1步：在“搜索引擎 搜索部分”接收需要查询的关键词的时候，首先由“搜索结果已经发布在 Web 服务器上的判决器”判断该关键词是否近期已经被其他人查询过，如果被查询过，并且结果已经在“搜索引擎搜索结果 Web 服务

器”上发布，则直接返回搜索结果（见图“M1”标记），该结果中已经将具有相同来源的网页聚合成一条搜索结果，点击“同源网页”按钮后，可以在“搜索引擎搜索结果 Web 服务器”上看到另一个包括全部搜索结果的搜索结果网页，完成整个查询过程。

第 2 步：如果在“搜索引擎 搜索部分”接收需要查询的关键词的时候，由“搜索结果已经发布在 Web 服务器上的判决器”判断该关键词近期没有被其他人查询过，并且也没有相应的查询结果在“搜索引擎搜索结果 Web 服务器”上发布则：

启动“网页搜索器”搜索“非同源网页结果数据库”和“同源网页结果数据库”找到符合搜索关键词的网页地址，并获取这些网页的内容。

如果“网页搜索器”在“非同源网页结果数据库”和“同源网页结果数据库”中没有找到符合搜索关键词的网页地址，则返回查询者“没有符合条件网页”的结果，并且将该搜索关键词加入到下一轮更新“非同源网页结果数据库”和“同源网页结果数据库”的任务中，如果在更新过程中找到了符合条件的网页地址则根据其是否具有同源网页而选择入“非同源网页结果数据库”或“同源网页结果数据库”，这样如果再有人搜索同样的关键词是就可以找到结果。

第 3 步：由“网页内容分离器”将找到的网页内容及超级链接目标分解成：多媒体、图片、文字、超级链接等种类。

第 4 步：分别由各种内容判决器产生判决结果

- A. 由“多媒体内容判决器”产生目标网页所含“相同多媒体文件程度 SMS”(Same Media Score) (多媒体的定义包括：Flash 类、视频/音频文件的播放服务或文件服务、IPTV/卫星直播/音视频监控/表演/人工应答等实时信息的播放服务或文件服务，其他多媒体服务)。
- B. 由“图片内容判决器”产生目标网页所含“相同图片的程度 SPS”(Same Photo Score)。
- C. 由“文字内容判决器”产生目标网页所含“相同文字的程度 STS”(Same Text Score)。
- D. 由“链接内容判决器”产生目标网页所含“相同超级连接的程度 SHS”(Same Hyperlinks Score)。

第 5 步：从“同源网页判决规则库”分别获取“多媒体判决权重 SMP”、“图片判决权重 SPP”、“文字判决权重 STP”、“链接判决权重 SHP”并分别与第 4 步

生成的“相同多媒体文件程度 SMS”、“相同图片的程度 SPS”、“相同文字的程度 STS”、“相同超级连接的程度 SHS”做数学乘法。

第 6 步：将“第 5 步”获得的数学乘法结果做加法，获得网页的“同源程度 SSS(Same Sourc Score)”，同源程度 $SSS = (SMS * SMP) + (SPS * SPP) + (STS * STP) + (SHS * SHP)$

第 7 步：判断该网页的“同源程度 SSS”是否超出门限，如果超出门限则判定为与其它网页的“同源网页”，如果没有超出门限则判定为“非同源网页”。

第 8 步：将“第 7 步”产生的“非同源网页”由“非同源网页处理模块”入“非同源网页结果数据库”；将“第 7 步”产生的“同源网页”由“同源网页处理模块”入“同源网页结果数据库”。

第 9 步：由“搜索结果网页发布器”根据“同源网页结果数据库”和“非同源网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果 Web 服务器”，再通过浏览器呈现给查询用户。(见图“M2”标记)。

作为第 9 步的另一种实现方法，也可以通过“动态网页 Web 服务器”直接通过浏览器呈现给查询用户。(见图“M3”标记)。

“网页内容分类器”可通过软件实现，直接根据“Html 语法”，“ASP/ASPX 语法”，“PHP”，“JSP”等各种网页上使用的语法解析出每一个内容的类型。

“同源多媒体处理模块”

图 4 为“同源多媒体处理模块”流程图。对于符合搜索条件的多媒体文件或服务，“同源多媒体处理模块”均采用 Html 网页中的超级链接方式提供给被查询者。为最大化地提高本系统的工作性能，我们采用了如下技术：

采用了网页发布技术，使用“搜索结果网页发布器”将搜索结果提前发布到“搜索引擎搜索结果 Web 服务器”，直接响应已经被查询过的搜索要求，避免根据请求动态从数据库生成动态网页的大量计算。

“同源信息处理模块”将处理结果分类别放置在“非同源多媒体索引数据库”和“同源多媒体索引数据库”中，并定期由“搜索结果网页发布器”发布到“搜索引擎搜索结果 Web 服务器”，避免了重复计算和减少了计算等待时间。

“同源多媒体处理模块”处理流程如下：

第 1 步：在收到查询者的搜索关键词，并通过软件根据关键词内容和关键词语法判断需要找的是多媒体文件或服务（例如，关键词中含有“.MP3”表示需要寻找的是.MP3 文件而不是含有该文字的网页）。

第2步：判断“要搜索的内容已经发布在Web服务器上吗？”，如果搜索的目标已经发布在“搜索引擎搜索结果Web服务器”上则直接返回搜索结果（见图“M1”标记），该结果中已经将符合搜索条件具有相同来源的多媒体的获取接口聚合成一条“标题搜索结果”，点击“同源文件”按钮后，可以在“搜索引擎搜索结果Web服务器”上看到另一个包括全部搜索结果的网页，使查询者可以看到符合查询条件的全部搜索结果，完成搜索过程。如果搜索的目标没有发布在“搜索引擎搜索结果Web服务器”上则从第3步开始。

第3步：返回查询者“没有符合条件多媒体”的结果。

第4步：将该搜索关键词加入到下一轮更新“同源多媒体索引数据库”和“非同源多媒体索引数据库”的任务中，并定期启动两个数据库的更新过程。

第5步：“同源多媒体索引数据库”和“非同源多媒体索引数据库”的更新过程：

- A. 由“多媒体搜索器”搜索网页新出现的多媒体文件或服务入口，通过软件进入该入口获取该文件或服务。
- B. 由“多媒体内容判决器”判断新找到的多媒体内容“与当前“同源多媒体索引数据库”的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同源多媒体索引数据库”的该类别；如果“否”则由“多媒体内容判决器”判断它“与当前非同源多媒体索引数据库”的内容属于同一内容吗？”
- C. 如果“是”则：“为当前的多媒体和与之同源的并已经存贮在‘非同源多媒体索引数据库’中的多媒体，新建一个类别并全部转移到‘同源多媒体索引数据库’”；如果“否”则“为当前的多媒体新建一个类别，并存入‘非同源多媒体索引数据库’”；

第6步：由“搜索结果网页发布器”根据“同源网页结果数据库”和“非同源网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果Web服务器”，再通过浏览器呈现给前来搜索的查询者（见图“M2”标记）。

作为第6步的另一种实现方法，也可以通过“动态网页Web服务器”直接通过浏览器呈现给查询用户。（见图“M3”标记）。

“同源图片处理模块”

图5为同源图片处理模块流程图。对于符合搜索条件的图片文件或链接，“同

源图片处理模块”均采用 Html 网页中的超级链接方式提供给被查询者。为最大化地提高本系统的工作性能，我们采用了如下技术：

采用了网页发布技术，使用“搜索结果网页发布器”将搜索结果提前发布到“搜索引擎搜索结果 Web 服务器”，直接响应已经被查询过的搜索要求，避免根据请求动态从数据库生成动态网页的大量计算。

“同源信息处理模块”将处理结果分类别放置在“非同源图片索引数据库”和“同源图片索引数据库”中，并定期由“搜索结果网页发布器”发布到“搜索引擎搜索结果 Web 服务器”，避免了重复计算和减少了计算等待时间。“同源图片处理模块”处理流程如下：

第 1 步：在收到查询者的搜索关键词，并通过软件根据关键词内容和关键词语法判断需要找的是图片文件或链接（例如，关键词中含有“.JPG”表示需要寻找的是.JPG 文件而不是含有该文字的网页）。

第 2 步：判断“要搜索的内容已经发布在 Web 服务器上吗？”，如果搜索的目标已经发布在“搜索引擎搜索结果 Web 服务器”上则直接返回搜索结果（见图“M1”标记），该结果中已经将符合搜索条件具有相同来源的图片的获取接口聚合成一条“标题搜索结果”，点击“同源文件”按钮后，可以在“搜索引擎搜索结果 Web 服务器”上看到另一个包括全部搜索结果的网页，使查询者可以看到符合查询条件的全部搜索结果，完成搜索过程。如果搜索的目标没有发布在“搜索引擎搜索结果 Web 服务器”上则从第 3 步开始。

第 3 步：返回查询者“没有符合条件图片”的结果。

第 4 步：将该搜索关键词加入到下一轮更新“同源图片索引数据库”和“非同源图片索引数据库”的任务中，并定期启动两个数据库的更新过程。

第 5 步：“同源图片索引数据库”和“非同源图片索引数据库”的更新过程：

- A. 由“图片搜索器”搜索网页新出现的图片文件或链接入口，通过软件进入该入口获取该文件或服务。
- B. 由“图片内容判决器”判断新找到的图片内容“与当前“同源图片索引数据库”的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同源图片索引数据库”的该类别；如果“否”则由“图片内容判决器”判断它“与当前非同源图片索引数据库”的内容属于同一内容吗？”

C. 如果“是”则：“为当前的图片和与之同源的并已经存贮在‘非同源图片索引数据库’中的图片，新建一个类别并全部转移到‘同源图片索引数据库’；如果“否”则“为当前的图片新建一个类别，并存入‘非同源图片索引数据库’”。

第6步：由“搜索结果网页发布器”根据“同源网页结果数据库”和“非同源网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果 Web 服务器”，再通过浏览器呈现给前来搜索的查询者（见图“M2”标记）。

作为第6步的另一种实现方法，也可以通过“动态网页 Web 服务器”直接通过浏览器呈现给查询用户。（见图“M3”标记）。

“同源文档处理模块”

图6为同源文档处理模块流程图。同源文档处理模块”支持常见文档格式：“.Txt”, “.Doc”, “.PPT”, “.PDF”, “.XLS”等等。对于符合搜索条件的文档文件或链接，“同源文档处理模块”均采用 Html 网页中的超级链接方式提供给被查询者。为最大化地提高本系统的工作性能，我们采用了如下技术：

采用了网页发布技术，使用“搜索结果网页发布器”将搜索结果提前发布到“搜索引擎搜索结果 Web 服务器”，直接响应已经被查询过的搜索要求，避免根据请求动态从数据库生成动态网页的大量计算。

“同源信息处理模块”将处理结果分类别放置在“非同源文档索引数据库”和“同源文档索引数据库”中，并定期由“搜索结果网页发布器”发布到“搜索引擎搜索结果 Web 服务器”，避免了重复计算和减少了计算等待时间。“同源文档处理模块”处理流程如下：

第1步：在收到查询者的搜索关键词，并通过软件根据关键词内容和关键词语法判断需要找的是文档文件或链接（例如，关键词中含有“.PDF”表示需要寻找的是. PDF 文件而不是含有该文字的网页）。

第2步：判断“要搜索的内容已经发布在 Web 服务器上吗？”，如果搜索的目标已经发布在“搜索引擎搜索结果 Web 服务器”上则直接返回搜索结果（见图“M1”标记），该结果中已经将符合搜索条件具有相同来源的文档的获取接口聚合为一条“标题搜索结果”，点击“同源文件”按钮后，可以在“搜索引擎搜索结果 Web 服务器”上看到另一个包括全部搜索结果的网页，使查询者可以看到符合查询条件的全部搜索结果，完成搜索过程。如果搜索的目标没有发布在“搜索引擎搜索结果 Web 服务器”上则从第3步开始。

第3步：返回查询者“没有符合条件文档”的结果。

第4步：将该搜索关键词加入到下一轮更新“同源文档索引数据库”和“非同源文档索引数据库”的任务中，并定期启动两个数据库的更新过程。

第5步：“同源文档索引数据库”和“非同源文档索引数据库”的更新过程：

- A. 由“文档搜索器”搜索网页新出现的文档文件或链接入口，通过软件进入该入口获取该文件或服务。
- B. 由“文字内容判决器”和“图片内容判决器”判断新找到的文档内容“与当前‘同源文档索引数据库’的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同源文档索引数据库”的该类别；如果“否”则由“文档内容判决器”判断它“与当前非同源文档索引数据库”的内容属于同一内容吗？”
- C. 如果“是”则：“为当前的文档和与之同源的并已经存贮在‘非同源文档索引数据库’中的文档，新建一个类别并全部转移到‘同源文档索引数据库’”；如果“否”则“为当前的文档新建一个类别，并存入‘非同源文档索引数据库’”。

第6步：由“搜索结果网页发布器”根据“同源网页结果数据库”和“非同源网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果 Web 服务器”，再通过浏览器呈现给前来搜索的查询者（见图“M2”标记）。

作为第6步的另一种实现方法，也可以通过“动态网页 Web 服务器”直接通过浏览器呈现给查询用户。（见图“M3”标记）。

“同源软件处理模块”

图7为同源软件处理模块流程图。对于符合搜索条件的软件文件或链接，“同源软件处理模块”均采用Html网页中的超级链接方式提供给被查询者。为最大化地提高本系统的工作性能，我们采用了如下技术：

采用了网页发布技术，使用“搜索结果网页发布器”将搜索结果提前发布到“搜索引擎搜索结果 Web 服务器”，直接响应已经被查询过的搜索要求，避免根据请求动态从数据库生成动态网页的大量计算。

“同源信息处理模块”将处理结果分类别放置在“非同源软件索引数据库”和“同源软件索引数据库”中，并定期由“搜索结果网页发布器”发布到“搜索引擎搜索结果 Web 服务器”，避免了重复计算和减少了计算等待时间。“同源软

件处理模块”处理流程如下：

第1步：在收到查询者的搜索关键词，并通过软件根据关键词内容和关键词语法判断需要找的是软件文件或链接（例如，关键词中含有“.EXE”表示需要寻找的是.EXE文件而不是含有该文字的网页）。

第2步：判断“要搜索的内容已经发布在Web服务器上吗？”，如果搜索的目标已经发布在“搜索引擎搜索结果Web服务器”上则直接返回搜索结果（见图“M1”标记），该结果中已经将符合搜索条件具有相同来源的软件的获取接口聚合成一条“标题搜索结果”，点击“同源文件”按钮后，可以在“搜索引擎搜索结果Web服务器”上看到另一个包括全部搜索结果的网页，使查询者可以看到符合查询条件的全部搜索结果，完成搜索过程。如果搜索的目标没有发布在“搜索引擎搜索结果Web服务器”上则从第3步开始。

第3步：返回查询者“没有符合条件软件”的结果。

第4步：将该搜索关键词加入到下一轮更新“同源软件索引数据库”和“非同源软件索引数据库”的任务中，并定期启动两个数据库的更新过程。

第5步：“同源软件索引数据库”和“非同源软件索引数据库”的更新过程：

- A. 由“软件搜索器”搜索网页新出现的软件文件或链接入口，通过软件进入该入口获取该文件或服务。
- B. 由“软件内容判决器”判断新找到的软件内容“与当前“同源软件索引数据库”的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同源软件索引数据库”的该类别；如果“否”则由“软件内容判决器”判断它“与当前非同源软件索引数据库”的内容属于同一内容吗？”
- C. 如果“是”则：“为当前的软件和与之同源的并已经存贮在‘非同源软件索引数据库’中的软件，新建一个类别并全部转移到‘同源软件索引数据库’”；如果“否”则“为当前的软件新建一个类别，并存入‘非同源软件索引数据库’”；。

第6步：由“搜索结果网页发布器”根据“同源网页结果数据库”和“非同源网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果Web服务器”，再通过浏览器呈现给前来搜索的查询者（见图“M2”标记）。

作为第6步的另一种实现方法，也可以通过“动态网页Web服务器”直接

通过浏览器呈现给查询用户。(见图“M3”标记)。

“同源数据或数据库处理模块”

图 8 为同源数据或数据库处理模块流程图。对于符合搜索条件的软件文件或链接，“同源数据处理模块”均采用 Html 网页中的超级链接方式提供给被查询者。为最大化地提高本系统的工作性能，我们采用了如下技术：

采用了网页发布技术，使用“搜索结果网页发布器”将搜索结果提前发布到“搜索引擎搜索结果 Web 服务器”，直接响应已经被查询过的搜索要求，避免根据请求动态从数据库生成动态网页的大量计算。

“同源信息处理模块”将处理结果分类别放置在“非同源数据索引数据库”和“同源数据索引数据库”中，并定期由“搜索结果网页发布器”发布到“搜索引擎搜索结果 Web 服务器”，避免了重复计算和减少了计算等待时间。“同源数据处理模块”处理流程如下：

第 1 步：在收到查询者的搜索关键词，并通过数据根据关键词内容和关键词语法判断需要找的是数据文件或链接（例如，关键词中含有“.DBF”表示需要寻找的是 .DBF 文件而不是含有该文字的网页）。

第 2 步：判断“要搜索的内容已经发布在 Web 服务器上吗？”，如果搜索的目标已经发布在“搜索引擎搜索结果 Web 服务器”上则直接返回搜索结果（见图“M1”标记），该结果中已经将符合搜索条件具有相同来源的数据的获取接口聚合成一条“标题搜索结果”，点击“同源文件”按钮后，可以在“搜索引擎搜索结果 Web 服务器”上看到另一个包括全部搜索结果的网页，使查询者可以看到符合查询条件的全部搜索结果，完成搜索过程。如果搜索的目标没有发布在“搜索引擎搜索结果 Web 服务器”上则从第 3 步开始。

第 3 步：返回查询者“没有符合条件数据”的结果。

第 4 步：将该搜索关键词加入到下一轮更新“同源数据索引数据库”和“非同源数据索引数据库”的任务中，并定期启动两个数据库的更新过程。

第 5 步：“同源数据索引数据库”和“非同源数据索引数据库”的更新过程：

- A. 由“数据搜索器”搜索网页新出现的数据文件或链接入口，通过数据进入该入口获取该文件或服务。
- B. 由“数据内容判决器”判断新找到的数据内容“与当前“同源数据索引数据库”的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同源数据索引数据库”的该类别；如果“否”则由“数

据内容判决器”判断它“与当前非同源数据索引数据库”的内容属于同一内容吗？”

- C. 如果“是”则：“为当前的数据和与之同源的并已经存贮在‘非同源数据索引数据库’中的数据，新建一个类别并全部转移到‘同源数据索引数据库’”；如果“否”则“为当前的数据新建一个类别，并存入‘非同源数据索引数据库’”；。

第6步：由“搜索结果网页发布器”根据“同源网页结果数据库”和“非同源网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果 Web 服务器”，再通过浏览器呈现给前来搜索的查询者（见图“M2”标记）。

作为第6步的另一种实现方法，也可以通过“动态网页 Web 服务器”直接通过浏览器呈现给查询用户。（见图“M3”标记）。

“同源GIS信息处理模块”

图9为“同源GIS信息处理模块”流程图。对于符合搜索条件的GIS信息文件或链接，“同源GIS信息处理模块”均采用Html网页中的超级链接方式提供给被查询者。为最大化地提高本系统的工作性能，我们采用了如下技术：

采用了网页发布技术，使用“搜索结果网页发布器”将搜索结果提前发布到“搜索引擎搜索结果 Web 服务器”，直接响应已经被查询过的搜索要求，避免根据请求动态从数据库生成动态网页的大量计算。

“同源信息处理模块”将处理结果分类别放置在“非同源 GIS 信息索引数据库”和“同源 GIS 信息索引数据库”中，并定期由“搜索结果网页发布器”发布到“搜索引擎搜索结果 Web 服务器”，避免了重复计算和减少了计算等待时间。“同源 GIS 信息处理模块”处理流程如下：

第1步：在收到查询者的搜索关键词，并通过软件根据关键词内容和关键词语法判断需要找的是GIS信息文件或链接（例如，关键词中含有“.JPG”表示需要寻找的是.JPG文件而不是含有该文字的网页）。

第2步：判断“要搜索的内容已经发布在 Web 服务器上吗？”，如果搜索的目标已经发布在“搜索引擎搜索结果 Web 服务器”上则直接返回搜索结果（见图“M1”标记），该结果中已经将符合搜索条件具有相同来源的GIS信息的获取接口聚合成一条“标题搜索结果”，点击“同源文件”按钮后，可以在“搜索引擎搜索结果 Web 服务器”上看到另一个包括全部搜索结果的网页，使查询者可

以看到符合查询条件的全部搜索结果，完成搜索过程。如果搜索的目标没有发布在“搜索引擎搜索结果 Web 服务器”上则从第 3 步开始。

第 3 步：返回查询者“没有符合条件 GIS 信息”的结果。

第 4 步：将该搜索关键词加入到下一轮更新“同源 GIS 信息索引数据库”和“非同源 GIS 信息索引数据库”的任务中，并定期启动两个数据库的更新过程。

第 5 步：“同源 GIS 信息索引数据库”和“非同源 GIS 信息索引数据库”的更新过程：

- A. 由“GIS 信息搜索器”搜索网页新出现的 GIS 信息文件或链接入口，通过软件进入该入口获取该文件或服务。
- B. 由“GIS 信息内容判决器”判断新找到的 GIS 信息内容“与当前“同源 GIS 信息索引数据库”的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同源 GIS 信息索引数据库”的该类别；如果“否”则由“GIS 信息内容判决器”判断它“与当前非同源 GIS 信息索引数据库”的内容属于同一内容吗？”
- C. 如果“是”则：“为当前的 GIS 信息和与之同源的并已经存贮在‘非同源 GIS 信息索引数据库’中的 GIS 信息，新建一个类别并全部转移到‘同源 GIS 信息索引数据库’”；如果“否”则“为当前的 GIS 信息新建一个类别，并存入‘非同源 GIS 信息索引数据库’”；。

第 6 步：由“搜索结果网页发布器”根据“同源网页结果数据库”和“非同源网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果 Web 服务器”，再通过浏览器呈现给前来搜索的查询者（见图“M2”标记）。

作为第 6 步的另一种实现方法，也可以通过“动态网页 Web 服务器”直接通过浏览器呈现给查询用户。（见图“M3”标记）。

“同价值网络服务处理模块”

图 10 为“同价值网络服务处理模块”流程图。对于符合搜索条件的网络服务，“同价值网络服务处理模块”均采用 Html 网页中的超级链接方式提供给被查询者。为最大化地提高本系统的工作性能，我们采用了如下技术：

采用了网页发布技术，使用“搜索结果网页发布器”将搜索结果提前发布到“搜索引擎搜索结果 Web 服务器”，直接响应已经被查询过的搜索要求，避免根

据请求动态从数据库生成动态网页的大量计算。

“同价值信息处理模块”将处理结果分类别放置在“非同价值网络服务索引数据库”和“同价值网络服务索引数据库”中，并定期由“搜索结果网页发布器”发布到“搜索引擎搜索结果 Web 服务器”，避免了重复计算和减少了计算等待时间。“同价值网络服务处理模块”处理流程如下：

第 1 步：在收到查询者的搜索关键词，并通过软件根据关键词内容和关键词语法判断需要找的是网络服务文件或链接（例如，关键词中含有“.JPG”表示需要寻找的是.JPG 文件而不是含有该文字的网页）。

第 2 步：判断“要搜索的内容已经发布在 Web 服务器上吗？”，如果搜索的目标已经发布在“搜索引擎搜索结果 Web 服务器”上则直接返回搜索结果（见图“M1”标记），该结果中已经将符合搜索条件具有相同来源的网络服务的获取接口聚合成一条“标题搜索结果”，点击“同价值文件”按钮后，可以在“搜索引擎搜索结果 Web 服务器”上看到另一个包括全部搜索结果的网页，使查询者可以看到符合查询条件的全部搜索结果，完成搜索过程。如果搜索的目标没有发布在“搜索引擎搜索结果 Web 服务器”上则从第 3 步开始。

第 3 步：返回查询者“没有符合条件网络服务”的结果。

第 4 步：将该搜索关键词加入到下一轮更新“同价值网络服务索引数据库”和“非同价值网络服务索引数据库”的任务中，并定期启动两个数据库的更新过程。

第 5 步：“同价值网络服务索引数据库”和“非同价值网络服务索引数据库”的更新过程：

- A. 由“网络服务搜索器”搜索网页新出现的网络服务文件或链接入口，通过软件进入该入口获取该文件或服务。
- B. 由“网络服务内容判决器”判断新找到的网络服务内容“与当前“同价值网络服务索引数据库”的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同价值网络服务索引数据库”的该类别；如果“否”则由“网络服务内容判决器”判断它“与当前非同价值网络服务索引数据库”的内容属于同一内容吗？”
- C. 如果“是”则：“为当前的网络服务和与之同价值的并已经存贮在‘非同价值网络服务索引数据库’中的网络服务，新建一个类别并全部转移到‘同价值网络服务索引数据库’”；如果“否”则“为当前的网络服务新建一个类别，并存入‘非同价值网络服务索引数据库’”；

第 6 步：由“搜索结果网页发布器”根据“同价值网页结果数据库”和“非同价值网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果 Web 服务器”，再通过浏览器呈现给前来搜索的查询者（见图“M2”标记）。

作为第 6 步的另一种实现方法，也可以通过“动态网页 Web 服务器”直接通过浏览器呈现给查询用户。（见图“M3”标记）。

“同价值商业信息处理模块”

图 11 为“同价值商业信息处理模块”流程图。对于符合搜索条件的商业信息，“同价值商业信息处理模块”均采用 Html 网页中的超级链接方式提供给被查询者。为最大化地提高本系统的工作性能，我们采用了如下技术：

采用了网页发布技术，使用“搜索结果网页发布器”将搜索结果提前发布到“搜索引擎搜索结果 Web 服务器”，直接响应已经被查询过的搜索要求，避免根据请求动态从数据库生成动态网页的大量计算。

“同价值信息处理模块”将处理结果分类别放置在“非同价值商业信息索引数据库”和“同价值商业信息索引数据库”中，并定期由“搜索结果网页发布器”发布到“搜索引擎搜索结果 Web 服务器”，避免了重复计算和减少了计算等待时间。“同价值商业信息处理模块”处理流程如下：

第 1 步：在收到查询者的搜索关键词，并通过软件根据关键词内容和关键词语法判断需要找的是商业信息文件或链接（例如，关键词中含有“.JPG”表示需要寻找的是 JPG 文件而不是含有该文字的网页）。

第 2 步：判断“要搜索的内容已经发布在 Web 服务器上吗？”，如果搜索的目标已经发布在“搜索引擎搜索结果 Web 服务器”上则直接返回搜索结果（见图“M1”标记），该结果中已经将符合搜索条件具有相同来源的商业信息的获取接口聚合成一条“标题搜索结果”，点击“同价值文件”按钮后，可以在“搜索引擎搜索结果 Web 服务器”上看到另一个包括全部搜索结果的网页，使查询者可以看到符合查询条件的全部搜索结果，完成搜索过程。如果搜索的目标没有发布在“搜索引擎搜索结果 Web 服务器”上则从第 3 步开始。

第 3 步：返回查询者“没有符合条件商业信息”的结果。

第 4 步：将该搜索关键词加入到下一轮更新“同价值商业信息索引数据库”和“非同价值商业信息索引数据库”的任务中，并定期启动两个数据库的更新过程。

第 5 步：“同价值商业信息索引数据库”和“非同价值商业信息索引数据库”的更新过程：

- A. 由“商业信息搜索器”搜索网页新出现的商业信息文件或链接入口，通过软件进入该入口获取该文件或服务。
- B. 由“商业信息内容判决器”判断新找到的商业信息内容“与当前“同价值商业信息索引数据库”的内容属于同一内容吗？”如果“是”则将它作为一个新的元素归入“同价值商业信息索引数据库”的该类别；如果“否”则由“商业信息内容判决器”判断它“与当前非同价值商业信息索引数据库”的内容属于同一内容吗？”
- C. 如果“是”则：“为当前的商业信息和与之同价值的并已经存贮在‘非同价值商业信息索引数据库’中的商业信息，新建一个类别并全部转移到‘同价值商业信息索引数据库’”；如果“否”则“为当前的商业信息新建一个类别，并存入‘非同价值商业信息索引数据库’”；。

第 6 步：由“搜索结果网页发布器”根据“同价值网页结果数据库”和“非同价值网页结果数据库”的内容动态生成搜索结果的静态网页，发布到“搜索引擎搜索结果 Web 服务器”，再通过浏览器呈现给前来搜索的查询者（见图“M2”标记）。

作为第 6 步的另一种实现方法，也可以通过“动态网页 Web 服务器”直接通过浏览器呈现给查询用户。（见图“M3”标记）。

“同价值商业信息处理模块”的特点在于能够根据商品或服务特点、供应与查询者的分布自动判断多个商业信息目标是否对查询者具有相同的使用价值，从而作为将其聚合成一条“标题搜索结果”的依据，以及查询结果排序的依据。

内容判决器可以在各种“同源（同价值）信息处理模块”中通用。

“内容判决器”具体实现方案

“多媒体内容判决器”具体实现方案：

- 1 输入：可以接收多个来源的多媒文件（如果是播放服务就将起录制成文件，或从播放服务器上获取媒体文件信息）。
- 2 处理：进行多媒体内容吻合度比对。
- 3 返回：计算输入多媒体中具有的相同内容程度值：SameMediaPower。

具体实现方法：

第1步：接收“被判断对象”：可以接收多个来源的多媒体。并记录被判断对象的数量：InputQuantity。

第2步：查找“被判断对象”在下表中可参与比对的属性，记录当前属性具有相同值的“被判断对象”的数量：SameQuantity（例如，5个被判断对象中，有3个对象的属性具有相同的值，则该属性的SameQuantity=3）

第3步：输入当前属性在判断过程中的“权重”值（从下表中查到）：Power

第4步：计算被全部“被判断对象”在当前属性上的吻合度：PSame = SameQuantity * Power

第5步：返回“第1步”对下一个“属性”执行“第1步”～“第4步”，得到该属性的PSame。直至获得全部属性的PSame值。

第6步：计算并返回“被判断对象”的相同内容程度值：SameMediaPower = (全部Psame值的数学累加值) / InputQuantity。

视频文件或播放服务判断内容：

视频文件属性	权 重	WMV	AVI	MPEG	其他 格式	说明
文件名	5	是	是	是		仅适于视频文件
文件长度	5	是	是	是		
文件校验值	5	是	是	是		
文件类型	5	是	是	是		仅适于视频播放服务
播放服务器 URL 地址	200	是	是	是		
播放服务器 IP 地址	5	是	是	是		
播放服务器端口	5	是	是	是		包括文件头内的全部属性和值
文件头信息	5	是	是	是		
标题	2	无	无	是		
主题	2	无	无	是		

视频文件属性	权重	WMV	AVI	MPEG	其他格式	说明
作者	2	无	无	是		
来源	2	无	无	是		
视频像素高度	10	是	是	是		
视频像素宽度	10	是	是	是		
受保护	2	是	无	是		
播放持续时间	10	是	是	是		
音频流数据速率	5	是	无	是		
音频采样宽度	5	是	无	是		
音频频道	5	是	无	是		
音频采样率	5	是	无	是		
视频流数据速率	5	是	是	是		
视频采样宽度	5	是	无	是		
视频流名称	5	是	无	是		
视频帧速率	5	是	是	是		
视频编码标准	5	是	是	是		

注释：

1. 本发明在于采用“权重”值计算每种属性的比对重要性的方法，而不仅仅是表中所列具体数值，表中“权重”具体数值仅为典型值，根据实际需要改变其具体数值仍然属于本发明范畴。

2. 根据实际情况，某些属性值可能为“空（Null）”，计算过程中属性值为“空”时不应被视作属性相等。

音频文件判断内容：

音频文件属性	权重	Wav	MP3	MP4	其他格式	说明
文件名	5	是	是	是		仅适于音频文件
文件长度	5	是	是	是		
文件校验值	5	是	是	是		
播放服务器 URL 地址	200	是	是	是		仅适于音频播放服务
播放服务器 IP 地址	5	是	是	是		
播放服务器端口	5	是	是	是		
文件头信息	5	是	是	是		包括文件头内的全部属性和值
艺术家	5		是	是		
唱片标题	10		是	是		
发行年	10		是	是		
曲目编号	10		是	是		
流派	10		是	是		
歌词	200		是	是		
标题	10		是	是		
备注	1		是	是		
来源受保护	1		是	是		
播放持续时间	10		是	是		
播放数据速率	10		是	是		
频道	10		是	是		
音频采样速率	10		是	是		
音频格式	10	是	是	是		

注释:

1 本发明在于采用“权重”值计算每种属性的比对重要性的方法，而不仅仅是表中所列具体数值，表中“权重”具体数值仅为典型值，根据实际需要改变其具体数值仍然属于本发明范畴。

2 根据实际情况，某些属性值可能为“空（Null）”，计算过程中属性值为“空”时不应被视作属性相等。

Flash 文件判断内容：

音频文件属性	权重	Flash			其他格式	说明
文件名	5	是				仅适于 Flash 文件
文件长度	5	是				
文件校验值	5	是				
播放服务器 URL 地址	200	是				仅适于 Flash 播放服务
播放服务器 IP 地址	20	是				
播放服务器端口	20	是				
文件头信息	20	是				包括文件头内的全部属性和值

注释：

1. 本发明在于采用“权重”值计算每种属性的比对重要性的方法，而不仅仅是表中所列具体数值，表中“权重”具体数值仅为典型值，根据实际需要改变其具体数值仍然属于本发明范畴。

2. 根据实际情况，某些属性值可能为“空（Null）”，计算过程中属性值为“空”时不应被视作属性相等。

“图片内容判决器”具体实现方案

1 输入：可以接收多个来源的图片。

2 处理：进行图片内容吻合度比对。

3 返回：计算输入图片中具有的相同内容程度值：SamePicPower。

具体实现方法：

第1步：接收“被判断对象”：可以接收多个来源的图片。并记录被判断对象的数量：InputQuantity。

第2步：查找“被判断对象”在下表中可参与比对的属性，记录当前属性具有相同值的“被判断对象”的数量：SameQuantity（例如，5个被判断对象中，有3个对象的属性具有相同的值，则该属性的SameQuantity=3）

第3步：输入当前属性在判断过程中的“权重”值（从下表中查到）：Power

第4步：计算被全部“被判断对象”在当前属性上的吻合度：PSame=SameQuantity* Power

第5步：返回“第1步”对下一个“属性”执行“第1步”～“第4步”，得到该属性的PSame。直至获得全部属性的PSame值。

第6步：计算并返回“被判断对象”的相同内容程度值：SamePicPower=(全部Psame值的数学累加值)/InputQuantity。

根据图片各种属性以及图像识别软件对于相似程度的判断。

图片文件属性	权 重	JPG	BMP	GIF	其他格式	说明
文件名	5	是	是	是		对同类型文件起作用
文件长度	5	是	是	是		
文件类型	5	是	是	是		
文件校验值	5	是	是	是		
文件头内容属性	5	是	是	是		
图片像素高度	5	是	是	是		
图片像素宽度	5	是	是	是		
水平分辨率	5	是	是	是		
垂直分辨率	5	是	是	是		
位深度	5	是	是	是		
帧数	5	是	是	是		

图片文件属性	权重	JPG	BMP	GIF	其他格式	说明
标题	2	是	否	否		
主题	2	是	否	否		
关键字	2	是	否	否		
备注	2	是	否	否		
作者	2	是	否	否		
图像相似程度识别	20	是	是	是		可跨文件类型 起作用

注释：

1. 本发明在于采用“权重”值计算每种属性的比对重要性的方法，而不仅是表中所列具体数值，表中“权重”具体数值仅为典型值，根据实际需要改变其具体数值仍然属于本发明范畴。
2. 根据实际情况，某些属性值可能为“空（Null）”，计算过程中属性值为“空”时不应被视作属性相等。

“文字内容判决器”具体实现方案

“文字内容判决器”，可通过软件实现：

- 1 输入：可以接收多个来源的文字，作为“被判断对象”。
- 2 处理：进行图片内容吻合度比对。
- 3 返回：“被判断对象”之间的一致程度值 SameTextPower。

实现方法：

第1步：找出输入的多个图片中

文字内容中，具有相同的单词或句子的部分的总计长度值：SameLenth。

第2步：找出输入的多个文字内容中，长度最短的输入文字的长度值，MinLenth。

第3步：返回文字相似程度值：SameTextPower=SameLenth /MinLenth

按照此方法找到的文字中：长度最长文字的通常是同一篇文章分页数量少或含有大量广告和外部超级链接的，而长度最短的文字通常是同一篇文章被分成更多页数或含有最少广告和外部超级链接的。

“链接内容判决器”具体实现方案

“链接内容判决器”，可通过软件实现：用来比对多个网页上所含有的超级链接是否具有共同特征。

1 输入：多组超级链接的 Url 地址（每组超级链接通常是从一个网页上获得的全部超级链接）。

2 处理：进行各组间超级链接 Url 地址吻合度计算

3 返回：各组间具有相同超级链接数。

实现方法：

第 1 步：接收“被判断对象”：多组超级链接的 URL 地址。

第 2 步：统计“被判断对象”相似程度：SameURLPower=在每组超级链接均出现过的 URL 地址数量。

第 3 步：返回 SameURLPower。

“软件内容判决器”具体实现方案

“软件内容判决器”，用来比对输入的多个软件是否是同种软件。

1 输入：可以接收多个来源的软件。

2 处理：进行软件内容吻合度比对。

3 返回：软件内容吻合度数值。

具体实现方法：

第 1 步：接收“被判断对象”：多个输入的文件或目录。并记录被判断对象的数量：InputQuantity。

第 2 步：查找“被判断对象”在下表中可比对的属性，记录当前属性具有相同值的“被判断对象”的数量：SameQuantity（例如，5 个被判断对象中，有 3 个对象的属性具有相同的值，则该属性的 SameQuantity=3）

第3步：输入当前属性在判断过程中的“权重”值（从下表中查到）：Power

第4步：计算被全部“被判断对象”在当前属性上的吻合度：PSame =
SameQuantity * Power。

第5步：返回“第1步”对下一个“属性”执行“第1步”～“第4步”，
得到该属性的PSame。直至获得全部属性的PSame值。

第6步：计算并返回“被判断对象”的吻合值：SameSoftPower = (全部 Psame
值的数学累加值) / InputQuantity。

文件及属性	权 重	未 压缩文 件	压 缩文 件	安 装目 录	某 属性	同 类 型 属性
软件包文件总长度	5	无	是	是		
软件包文件总数量	5	无	是	是		
软件包各文件名称	5	无	是	是		
文件名	5	是	是	是		对同类型文件 起作用
文件长度	5	是	是	是		
文件校验值	5	是	是	是		
文件头内容属性	5	是	是	是		
版本信息：文件版本	1	是	是	是		对于单个文件 有效 对于文件包中 的每个文件都 有效。
版本信息：描述	1	是	是	是		
版本信息：版权	1	是	是	是		
版本信息：产品版本	2	是	是	是		
版本信息：产品名称	2	是	是	是		
版本信息：公司	2	是	是	是		
版本信息：商标	2	是	是	是		
版本信息：内部名称	1	是	是	是		

文件文件属性	权重	安装文件	压缩文件	安装目录	其他格式	说明
版本信息：文件版本	1	是	是	是		
版本信息：语言	1	是	是	是		
版本信息：源文件名	1	是	是	是		
兼容性信息：适用操作系统	1	是	是	是		
摘要信息：作者	1	是	是	是		
摘要信息：来源	1	是	是	是		
摘要信息：修订版本	1	是	是	是		

注释：

1. 本发明在于采用“权重”值计算每种属性的比对重要性的方法，而不仅仅是表中所列具体数值，表中“权重”具体数值仅为典型值，根据实际需要改变其具体数值仍然属于本发明范畴。

2. 根据实际情况，某些属性值可能为“空（Null）”，计算过程中属性值为“空”时不应被视作属性相等。

“数据或数据库内容判决器”具体实现方案

逐一比对不同数据库文件内的每条数据记录内容是否相等，返回参与比对的数据库一致程度值 SameDBPower 是否超过门限。

SameDBPower=字段名称相同并且数值相等的记录数/参与比对的数据库拥有该字段的最少记录个数。

SameDBPower 反映了相同内容记录数相对拥有最少记录数的数据库的比例， SameDBPower 取值为：0~1。

“数据或数据库内容判决器”具体实现方案

对于数据文件可采用如下实现步骤：

第1步：在参与比对的多个数据文件中，随机选取一个文件作为“比对标准”。

第2步：进行其它文件与“比对标准”的一致性的粗略比较：文件长度、文件校验和、标题、主题、版本、作者、类别、关键字、备注等文件属性信息。

第3步：如果一致则判定为“粗略一致”，这样的判断结果可以直接作为“数据或数据库内容判决器”的输出。

第4步：如需进一步的比对，在获得“粗略一致”的输入文件中，进行第5步。

第5步：精细比较：文件属性信息和文件中每一个字节的逐一比对。全部特征均吻合的文件可以判定为“完全一致”，作为“数据或数据库内容判决器”的输出。

对于数据库文件可采用如下实现步骤：

第1步：对输入的数据库文件根据文件名后缀和文件属性判断是否符合同种数据库格式。

第2步：对于同种数据库格式进行第3步，对于不同种数据库格式直接第4步

第3步：同种格式数据库粗略比较：文件长度、文件校验和、标题、主题、版本、作者、类别、关键字、备注等文件属性信息。上述特征不完全符合作为“不一致”判断结果输出，对于完全符合的数据库文件进行第4步。

第4步：数据库精细比较：（本步骤适应各种不同的数据库文件参与内容比对）。按照每种数据库文件的格式逐一提取其“数据库表”，判断其“数据库表”结构是否一致：不一致作为“不一致”输出，一致的数据库文件进行第5步。

第5步：逐一比对参与比对的数据库文件的每条记录的内容：遇到记录内容相同的情况：为计数器“SameRecNum”字段名称相同并且数值相等的记录数”加1。

第6步：计算“SameDBPower 数据库一致程度值” = “SameRecNum”字段名称相同并且数值相等的记录数” / “参与比对的数据库拥有该字段的最少记录个数”。（SameDBPower 反映了相同内容记录数相对拥有最少记录数的数据库的比例， SameDBPower 取值为：0~1）。

第7步：判断“SameDBPower 数据库一致程度值”是否超过门限，超过门

限则输出“一致”作为判断结果，否则输出“不一致”作为判断结果。

“GIS信息内容判决器”

“GIS 信息内容判决器”，可通过软件实现：

- 1 输入：可以接收多个来源的数字地图，作为“被判断对象”。
- 2 处理：进行数字地图的覆盖范围的吻合度比对。
- 3 返回：“被判断对象”之间的一致程度值 SameMapPower（取值 0~1）。

实现方法：

第 1 步：按照数字地图的格式打开参加比对的数字地图文件。

第 2 步：找到数字地图的西北角和东南角（也可以是其它形式的地图对角）的经纬度。

第 3 步：比对参加比对的数字地图的西北角和东南角的经度、纬度误差，计算地图覆盖区域的一致性值 SameMapPower：

假设“地图 1”和“地图 2”参与比对：

则：

SameMapPower=两副地图重叠区域的面积/两副地图中最小地图的面积。

第 4 步：返回 SameMapPower 值。

第 5 步：判断 SameMapPower 是否超过门限（例如：门限值 =0.8），是则判定为相同的地图，不是则判定为不相同的地图。

“网络服务内容判决器”

“网络服务内容判决器”的 FTP 服务内容判决：

第 1 步：采用相应的 FTP 协议登陆参与比对的服务，并获取其内部的文件。

第 2 步：在获取 FTP 服务的文件后，首先根据文件名后缀判断文件类型是否一致，若不一致返回“不一致”作为输出，若文件类型一致，进行第 3 步。

第 3 步：根据文件类型采用“多媒体内容判决器”、“图片内容判决器”、“文字内容判决器”、“软件内容判决器”、“数据或数据库内容判决器”或“GIS 信息内容判决器”判决其文件内容是否一致，并返回其判断结果。

Email 网站提供的邮箱服务内容判决：

Email 网站提供的邮箱服务信息主要是通过软件搜索各个网站的网页，并从网页标签中解析出邮箱的大小、收费情况、是否支持 POP 协议等信息。

第 1 步：将邮箱尺寸划分成相应的等级，（例如：10MB~25MB、25MB~100MB、100MB~300MB、300MB~1GB、1GB~100GB 等），然后判断参与比对的邮箱是否在同一个级别里，如果“不是”则返回“不一致”，如果“是”则进行第 2 步。

第 2 步：比对“收费情况”是否一致，如果“不是”则返回“不一致”，如果“是”则进行第 3 步。

第 3 步：比对支持 POP 协议条件是否一致，如果“不是”则返回“不一致”，如果“是”则返回“一致”。

“商业信息内容判决器”

在网页上发布的产品或服务出售信息是否相同，并在相同的自然地理范围内、相同的行政地理范围、相同距离范围内。

第 1 步：比对参与比对的商业信息是否是相同的产品或服务，如果“不是”返回“不一致”，如果“是”进入第 2 步。

第 2 步：判断参与比对的商业信息是否具有地理位置敏感性（例如：生活消费类商品、需要到现场服务的服务具有地理位置敏感性，例如冰激凌、家教服务等），如果“不是”返回判断结果“一致”，如果“是”则进行第 3 步。

第 3 步：判断参与比对的商业信息的提供者是否处于相同的城市或区域，如果“不是”返回判断结果“不一致”，如果是返回判断结果“一致”。

“获取网页用户关注度子系统”

图 12 为获取网页用户关注度子系统结构图。该搜索引擎能够和与之配套的网络浏览器（或兼容该搜索引擎能够和与之配套的网页浏览器之间通讯协议的其他第三方浏览器）的协同工作方式，由网络浏览器采集用户对每个网页的关注程度，并上报给搜索引擎，作为搜索引擎进行搜索结果排名或选择“标题搜索结果”的依据。本方法和装置还可以单独于搜索引擎之外，独立形成能够提供“网页热

门程度排行榜”的 Web 查询系统，并可以进行收费业务或作为交换条件换取其他利益。

本系统主要包含两大部分：“PageFocus 网络服务器”和“PageFocus 网络浏览器”。

“PageFocus 网络服务器”构造

“PageFocus 网络服务器”通过“PageFocus 网络浏览器”获取全球用户对每一个网页的关注程度，并形成该网页的“关注分值 PageFocus”数据库，做为网页的热门程度的衡量值。

“PageFocus 网络服务器”由下列组成：

(1) “PageFocus 浏览器 ID 注册服务器”：为正在网络上使用的“PageFocus 网络浏览器”分配全球唯一的 ID 标识号。

(2) “PageFocusAccServer 网页关注统计服务器”：接收全球正在运行的“PageFocus 网络浏览器”发来的“PageFocus 数据包”内包含的对于一个或多个网页的“关注分值 PageFocus”。ID 号用来区分不同的浏览用户。

(3) “PageFocus 浏览器在线升级服务器”：用来向全球“PageFocus 网络浏览器”提供在线升级服务。

(4) “数据加解密模块”：用来在“PageFocus 网络服务器”和“PageFocus 网络浏览器”之间传递加密数据，放置被攻击或窃取信息。

“PageFocus 网络浏览器”构造

“PageFocus 网络浏览器”通过网络向“PageFocus 网络服务器”汇报当前用户对于某个网页的关注程度。

“PageFocus 网络浏览器”由下列组成：

(1) “关注分值 PageFocus 计算模块”：根据用户对“PageFocus 网络浏览器”的操作，计算用户对某网页的关注程度，并形成“PageFocus 数据包”向“PageFocusAccServer 网页关注统计服务器”汇报。

(2) “PageFocus 浏览器 ID 注册模块”：与“PageFocus 浏览器 ID 注册服务器”通讯以获得全球唯一的标识 ID，作为区分不同用户的依据。

(3) “PageFocus 浏览器在线升级模块”：与“PageFocus 浏览器在线升级服

务器”通讯，以保持当前用户计算机上的“PageFocus 浏览器”是最新版本。

本装置包含：本发明创造的“PageFocus 网络浏览器”、“PageFocus 浏览器 ID 注册服务器”和“网页计分服务器”，具体实现方法如下：

第 1 步：开发一个特殊的“PageFocus 网络浏览器”，每个浏览器均在安装时具备全球唯一的 ID 标识号，或在使用时主动寻找网络上的“PageFocus 浏览器 ID 注册服务器”以获得全球唯一的 ID 标识号。

第 2 步：“PageFocus 网络浏览器”具备具有常规网络浏览器（例如：微软公司的 IE 浏览器）的全部功能。

第 3 步：“PageFocus 网络浏览器”还具备将用户对浏览器的操作和对网页的操作按照下表所列权重转换成网页的“关注分值 PageFocus”并形成“PageFocus 数据包”，以加密方式通过网络协议传递至本搜索引擎的“PageFocusAccServer 网页关注统计服务器”。

第 4 步：“PageFocusAccServer 网页关注统计服务器”在收到全球的每一个“PageFocus 网络浏览器”发来的“PageFocus 数据包”后将其内部包含的“关注分值 PageFocus”累加到相应的网页上。

第 5 步：“PageFocusAccServer 网页关注统计服务器”上包含的全球每一个网页的“关注分值 PageFocus”，这些信息可以通过各种处理方法形成：搜索引擎对网页排行依据、搜索引擎在具有相同内容搜索结果中选择可以作为“标题搜索结果”的依据、也可以直接公布出来作为“网页热门程度排行榜”的服务。

“PageFocus 网络浏览器”计算“关注分值 PageFocus”的方法：

由于“PageFocus 网络浏览器”具有普通浏览器的全部功能，故此可以在用户使用浏览器的时候，按照下表采集其操作行为，并按照每种行为的“权重”对该网页进行“关注分值 PageFocus”计分，并在浏览器彻底关闭该网页的时候形成一条关于该网页“关注分值 PageFocus”的分值记录，以“PageFocus 数据包”的形式发给

“PageFocusAccServer 网页关注统计服务器”。

浏览器操作行为	权重	说明
打开网页	1	用户有兴趣打开对网页
在网页停留小于 10 秒	-10	别人讨厌的网页

浏览器操作行为	权重	说明
在网页停留小于30秒	-3	不兴趣或内容非常简单
在网页停留小于1分钟	-1	没什么兴趣或内容非常简单
在网页停留超过1分钟	3	普通的浏览
在网页停留超过5分钟	5	比较认真的浏览
文字阅读速度	100/ 文字阅读速度	<p>1. “文字阅读速度”：可以通过判断用户使用键盘、鼠标滚轮、“滚动条拖动”的操作幅度、频率和网页内容字体大小来计算。</p> <p>2. “文字阅读速度” = 被滚动的文字数量 / 滚动时间间隔。</p> <p>3. “文字阅读速度”的单位为：字/每秒</p> <p>4. 说明：1秒看100字说明很不认真，得1分；1秒看一个字说明很认真，得100分。</p>
网页内容中的图片被点击，而图片上的超级链接指向一个图片	3	说明该图片可能提供了“放大图片”，代表该图片可能是网页的主体内容，并且用户点击了它，说明用户想仔细看看放大的图片。
网页内容中的图片被点击，而图片上的超级链接指向网站内一个网页	1	很可能是个“广告”图片。
网页内容中的图片被点击，而图片上的超级链接指向网站外一个	0	很可能是个“广告”图片。

浏览器操作行为	权重	说明
网页		
网页内容中的超级链接被点击，而目标与当前网页处于同一目录下。	2	<p>很可能是一篇文章的网页连载，本网页和被点击超级链接指向的网页均可获得本“权重”的得分。</p> <p>例如：“http://www.yoogol.com/news/01.htm”和“http://www.yoogol.com/news/02.htm”</p>
网页内容中的超级链接被点击，而目标站外网页。	0	和本网页没什么关系
网页内容中含有标题为“ <u>1 2 3 4</u> ”等带有序号含义的文字，而目标均指向同一个 URL 目录的网页的超级链接	N	本网页和标题为“ <u>1 2 3 4</u> ”等带有序号含义，并且目标均指向同一个 URL 目录的网页通常为同一篇文章的分页显示。文章的任何页面的得分（包括负数得分），其他页面即使没有被打开也可得到相同的得分。
用户使用浏览器右键菜单投票： “投票-10 分”	PageFocus*0%	1 浏览器对网页的任何部分均可使用鼠标右键菜单，菜单中含有“投票 10 分”、“投票 5 分”、“投票 1 分”、 “投票-1 分”、“投票-5 分”、“投票-10 分”的菜单选项。
用户使用浏览器右键菜单投票： “投票-5 分”	PageFocus*10%	
用户使用浏览器右键菜单投票： “投票-1 分”	PageFocus*50%	2 用户使用右键菜单投票当前网页背景时，其 PageFocus 得分 = “当前 PageFocus 得分” * 权重。
用户使用浏览器右键菜单投票： “投票 10 分”	PageFocus*10	3 用户使用右键菜单投票当前网页各种网

浏览器操作行为	权重	说明
用户使用浏览器右键菜单投票：“投票 5 分”	PageFocus*5	该元素超级链接所指向的网页得到 PageFocus 得分 = 当前网页“当前 PageFocus 得分” * 权重
用户使用浏览器右键菜单投票：“投票 1 分”	PageFocus*1	

注释：

1. 用本评分标准虽然可能有误判，但是通过网络上的大量操作可以获得统计上的准确性。
2. 表中所列“权重”具体数值，仅为典型值，本发明在于通过浏览器为页面打分，任何其他“权重项目”和“权重”的改变，均属本发明范畴。
3. 采用用户对网页投票的方式是基于对于网民公德的充分信任，故此其“权重”对整体得分的数学乘法，而不是数学加法。
4. 由于每个网页均可能得到大量的 PageFocus 得分，可能导致软件变量的溢出，故此在“PageFocusAccServer 网页关注统计服务器”可以采用“数学对数”或“科学计数法”记录得分。
5. 为本方法的其他途径，除了在浏览器彻底关闭该网页时可以形成“PageFocus 数据包”外，还可以用其他任意规则来确定“PageFocus 数据包”的时机，例如：定时，累计到某个分值等等，这些方法均属于本发明范畴。
6. 表中“每行文字阅读速度”的详细计算方法：
 - A. 鼠标滚轮滚动：文字阅读速度 = (显示区宽度/字体宽度) * 每次滚动的文字行数/滚动时间间隔。
 - B. 键盘翻页：文字阅读速度 = (显示区宽度/字体宽度) * 每次翻页的文字行数/翻页时间间隔。
 - C. 窗体滚动条滚动：文字阅读速度 = (显示区宽度/字体宽度) * 每次滚动的文字行数/滚动时间间隔。

“PageFocus 数据包”的形成方法

“PageFocus 数据包”的内容：

记录编号	字段名称	说明
记录 1	PageFocus 浏览器 ID	发送 “PageFocus 数据包”的 PageFocus 浏览器 ID
	网页 URL	被打分的网页 URL 地址
	网页 PageFocus 得分值	被打分的网页 PageFocus 得分值
记录 2		
记录 N		

注释：每个“PageFocus 数据包”可以包含多个网页的得分记录。每条网页得分记录还可以添加其他的属性，但是为了提高效率，表中仅仅列出最重要的内容，在表中添加其他属性也属本发明范畴。“PageFocus 数据包”发送时机的选择：

为了减少发送“PageFocus 数据包”占用的带宽和给服务器端带来的压力，可以采取如下几个方案之一：

当某个网页被彻底从浏览器关闭时发送“PageFocus 数据包”。

当浏览器被彻底关闭时发送“PageFocus 数据包”。

浏览器将“PageFocus 数据包”以文件形式保留在本地计算机，积累到特定数量、或特定长度、或特定时间周期时再发送。

“标题搜索结果”选择算法

本算法主要用于如何在原始搜索结果中选择可以用来作为“标题搜索结果”的“同源搜索结果”。本算法需要解决如下问题：

1. 通过网络使用者行为和网页内容判断网页的内容质量，质量高的优先显示。
2. 避免某一搜索结果因成为“标题搜索结果”而承受过多的点击流量，导致网站处理速度变慢甚至崩溃。
3. 避免某一搜索结果因成为“标题搜索结果”而承受过多的点击流量导致服务响应速度变慢，而降低访问者的使用体验好感。
4. 使成为“标题搜索结果”作为一种权力，可以提供给需要的网站，这些网站可以购买这种权力。

5. 每个“同源搜索结果”的原始结果均有机会按照某种概率成为“标题搜索结果”。

“标题搜索结果”选择方法在于，在“同源搜索结果”中选择“标题搜索结果”时，同时考虑了“搜索结果内容质量”、“加权值”和“服务响应延迟”三个要素，即：内容质量高的优先显示、有加权的优先显示、网络服务好的优先显示；而在排列全部“同源搜索结果”时则仍然按照这种原则，而“加权值”可以向本发明的系统运营者购买。“标题搜索结果”选择的具体实现方法如下：

第1步：计算每个“同源搜索结果”成为“标题搜索结果”的概率权值 PWn（该搜索结果为第 n 条）：

$$PWn = TP * PageFocus / (RespDelay - K)$$

注释 1：当 (RespDelay-K) 小于等于零时，(RespDelay-K) 应取值为 1。

注释 2：公式中变量含义如下

A. PageFocus 网页关注度值：是该搜索结果根据本发明中“获取网页用户关注度的方法与装置”所获得的“PageFocus 值”。

B. RespDelay 网页服务响应延迟：是该搜索结果在向搜索者提供服务访问时的响应延迟。（由于访问体验取决于网站的响应延迟，反应越慢，体验越差）。

C. K 服务响应常数：是可以定义的常数，建议使用 50 毫秒 (ms)，低于 K 值的服务响应延迟将不被察觉，不影响使用体验，从而可以忽略。

D. TP 标题搜索结果权力：作为一种加权，任何人都可以和本发明系统的运营者通过各种交换条件而获取“TP 标题搜索结果权力”。

E. 作为本公式的其它实现算法，还可以有如下其它形式：

a. $PWn = (TP + PageFocus) / (RespDelay - K)$

b. $PWn = (TP + PageFocus) / RespDelay / K$

c. $PWn = TP * PageFocus / RespDelay / K$

第2步：统计求和全部原始“同源搜索结果”的概率权值 PWn 的总和：PWall 全部概率权值。

第3步：计算每条“同源搜索结果”成为“标题搜索结果”的概率：Pn = PWn/PWall。

第4步：按照Pn值的概率，随着搜索者的访问动作，动态地随机选择“标题搜索结果”，呈现给搜索者。

网站内容风格自适应的装置与方法

本发明的内容在于：利用各种可能获得的、有助于判断用户所处环境和状态的信息，使处于不同工作或生活休闲状态的用户在无需任何操作、注册、设定或Cookie设定的前提下，访问相同页面URL地址时看到不同的风格，其中包括：

- 1.利用用户的IP地址判断其所处的国家或区域，再结合通过本网站时间便可以计算出访问者的当地行政区域时间，通过他的时间可以判断他处于工作状态还是休闲状态。
- 2.通过用户的IP地址，可以查询到该IP地址的属性：家庭、工作场所。根据其所处场所提供适合其所处环境的风格和内容。
- 3.通过用户的IP地址可以获知其所处的地理位置，在查询商业信息时，可以自动将距离他最近的供应商排列在最前面。

举例如下：

同一时刻，不同的用户访问本网站内相同URL的一个网页时看到的内容是不同的：

- A.工作状态和环境中的用户看到的是庄重、简介、不含休闲消遣娱乐信息的页面。
- B.休闲状态和环境中的用户看到的是热闹、可含有休闲消遣娱乐信息、可含有个人消费广告信息的页面。

本发明可以部分或全部地被应用到搜索引擎以外的网站系统，均属本发明范畴。

目前各个大型网站，为了满足大流量的访问，均采用了服务器集群，甚至在区域建立本地服务分系统，来分流用户访问。但是目前的服务器集群的重要特征就是每个集群成员均提供完全相同的内容。如图13：前来访问的用户被“网站服务器集群入口”设备，部分任何特征地，直接分配到某个具有相同内容的服务器集群成员服务器上。

如图14，而本发明的装置对上述结构做了部分改动，在“网站服务器集群入口”收到访问用户后，根据其访问网站时发过来的IP地址等各种用户属性信

息判断其是否处于工作状态，并根据其是否处于工作状态向其提供不同风格和内容的信息服务。

自动判断用户状态并提供恰当的网页风格与内容的方法

第1步：首先将服务器集群划分成“工作风格”和“个人和休闲风格”两大类，无论是静态页面还是动态页面，在向这两类服务器更新相同的内容的时候，自动产生两类风格，以便不同工作或生活休闲状态的用户在访问相同页面 URL 地址时看到不同的风格。

第2步：在“网站服务器集群入口”收到用户首次访问本网站网页的请求后，首先在访问协议里（或 IP 层协议里）获取其 IP 地址。

第3步：根据 IP 地址在“IP 地址属性数据库”中查询其 IP 地址是“工作场合 IP 地址”还是“私人或休闲场合的 IP 地址”，若是“工作场合 IP 地址”则进行第4步，若是“私人或休闲场合的 IP 地址”则进行第5步。

第4步：获取“工作场合 IP 地址”所处的地理位置，并得到该地理区域的行政时间，若是该 IP 地址所属区域正处于工作时间（周 1~5 的 8:00~20:00）则将其访问分配到服务器集群中的“工作风格服务器”上向其提供适合工作场合使用的页面服务，否则进行第5步。

第5步：则将其访问分配到服务器集群中的“个人和休闲风格服务器”上向其提供适合个人和休闲状态使用的页面服务。

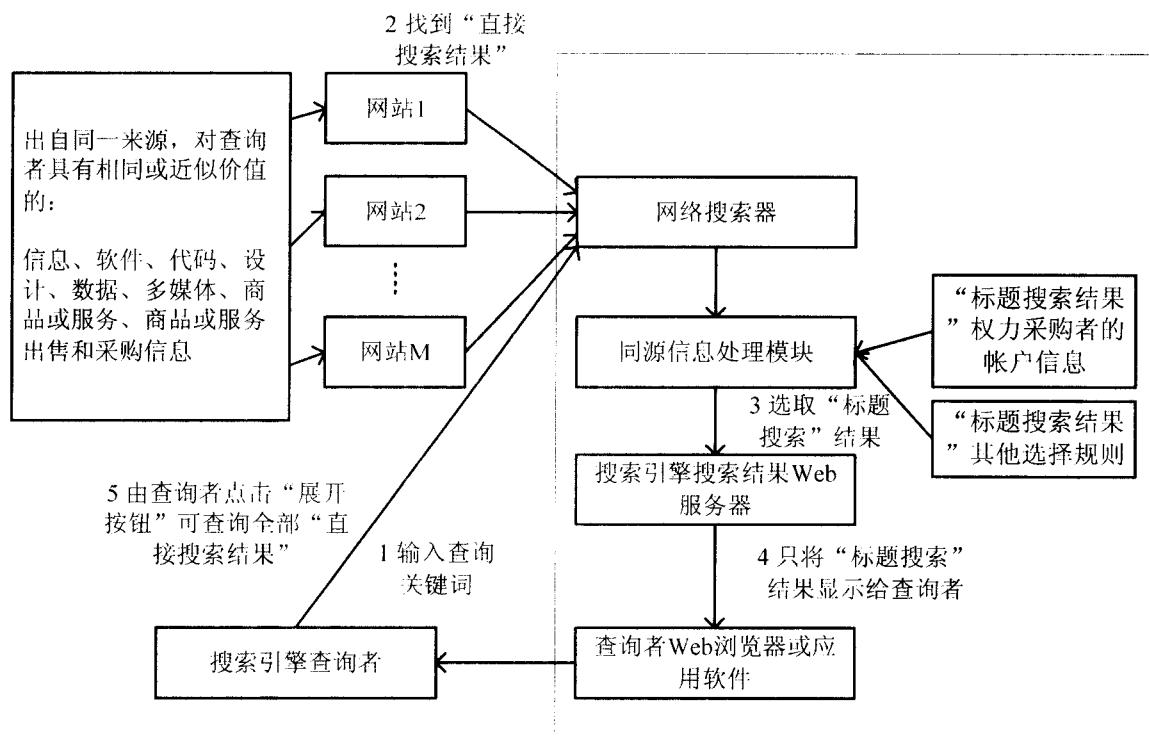


图 1

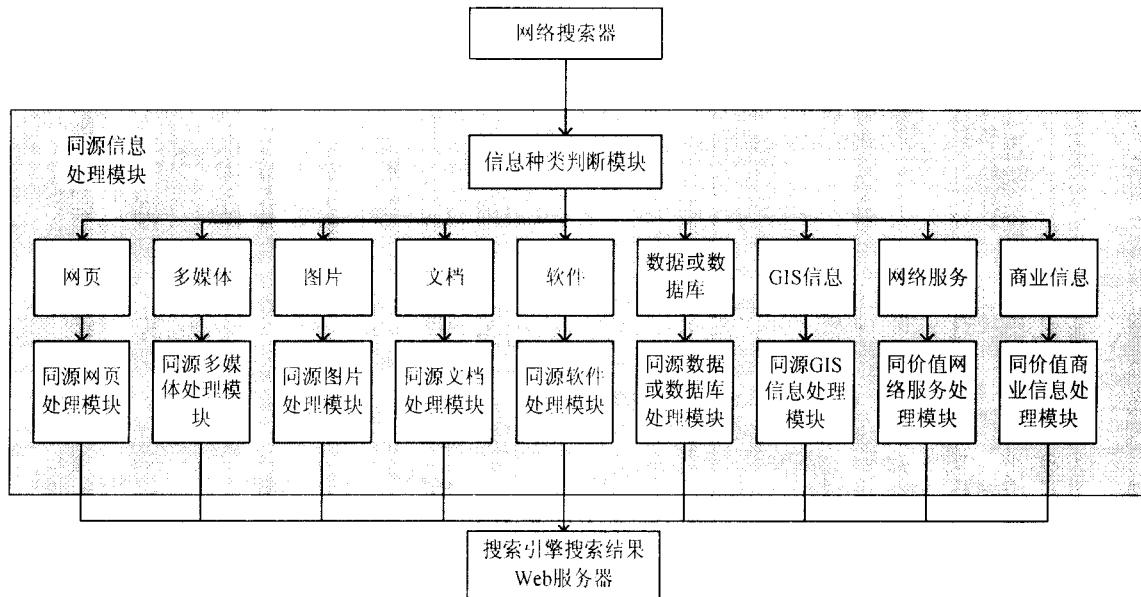


图 2

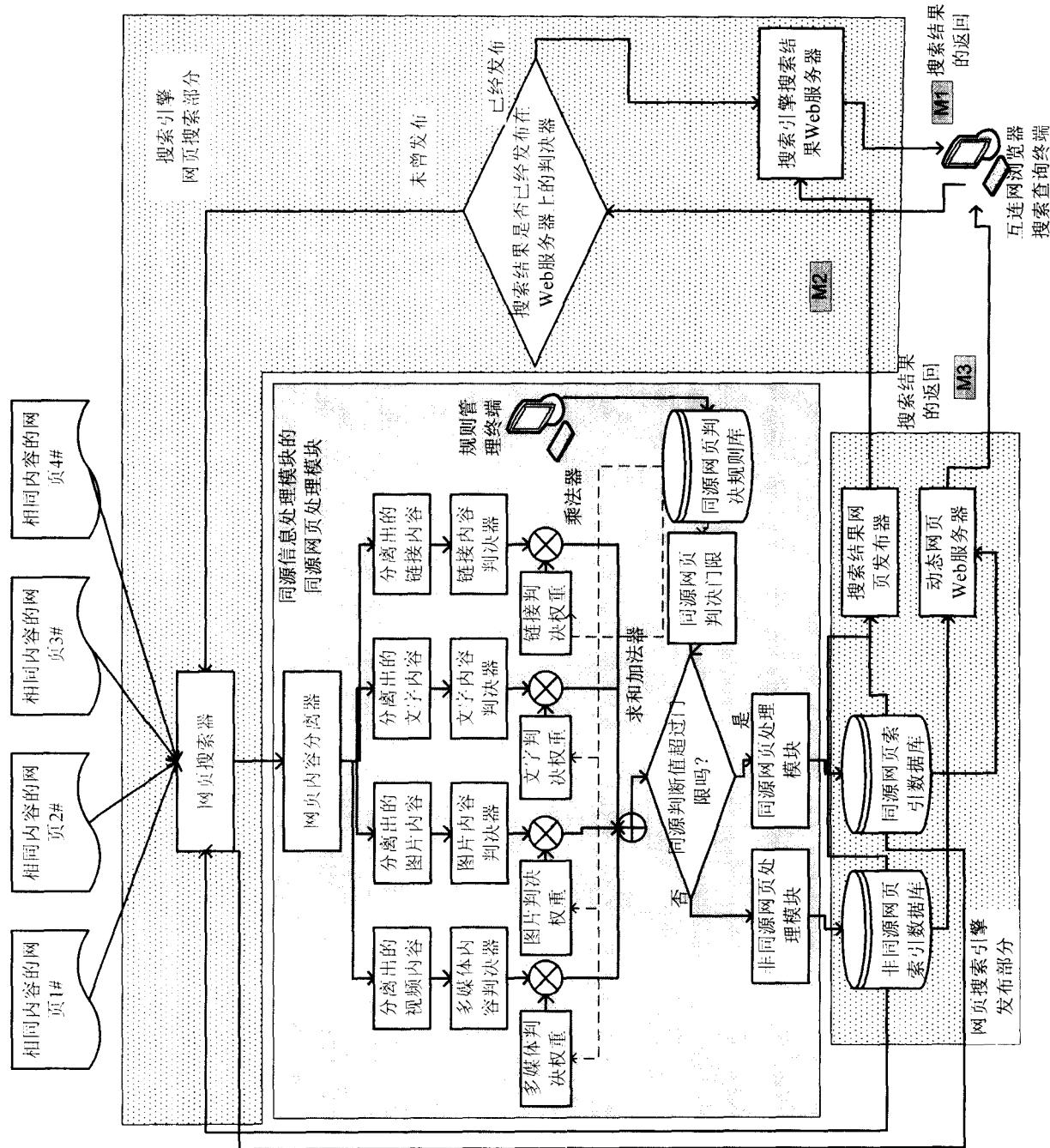


图3

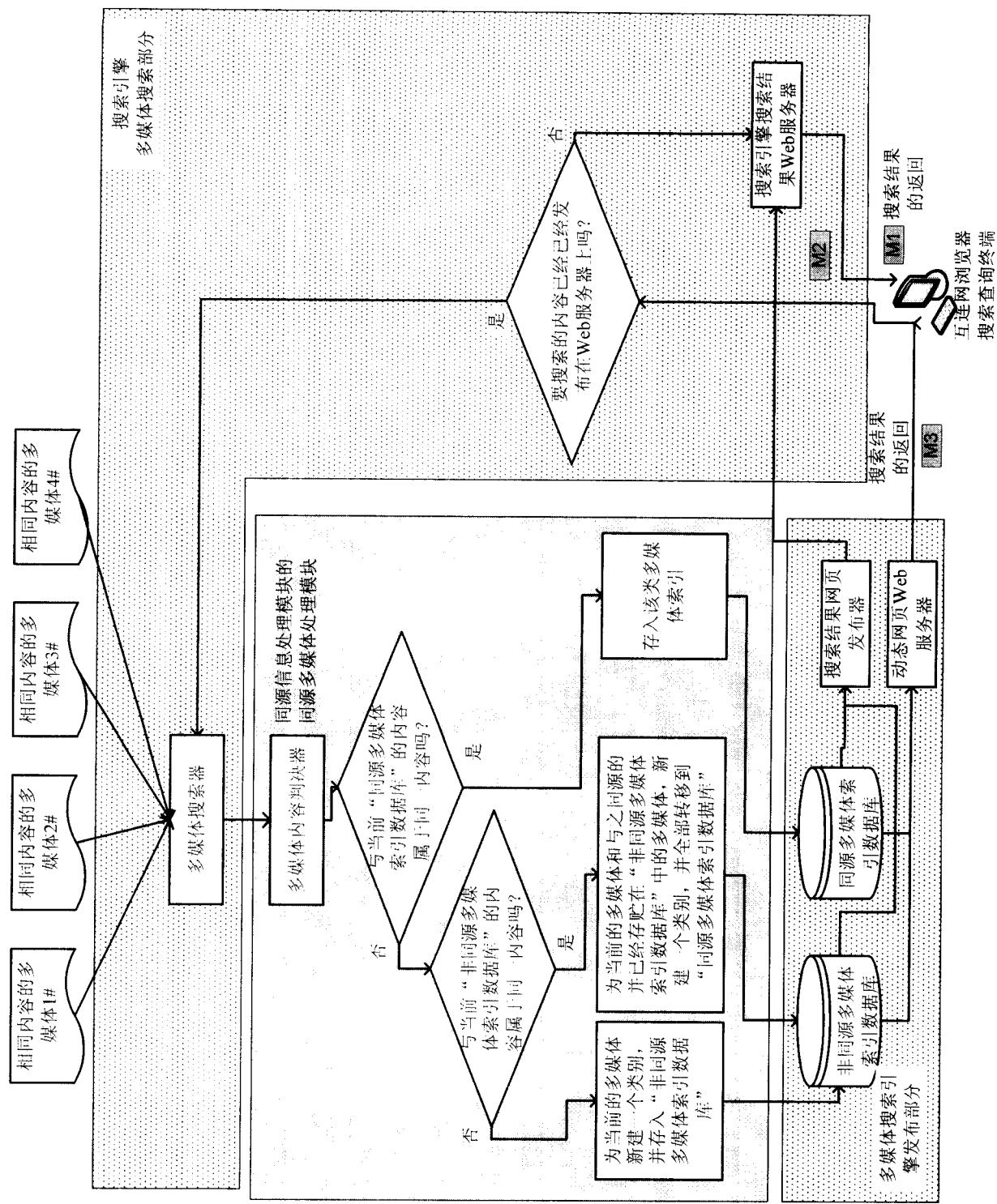


图4

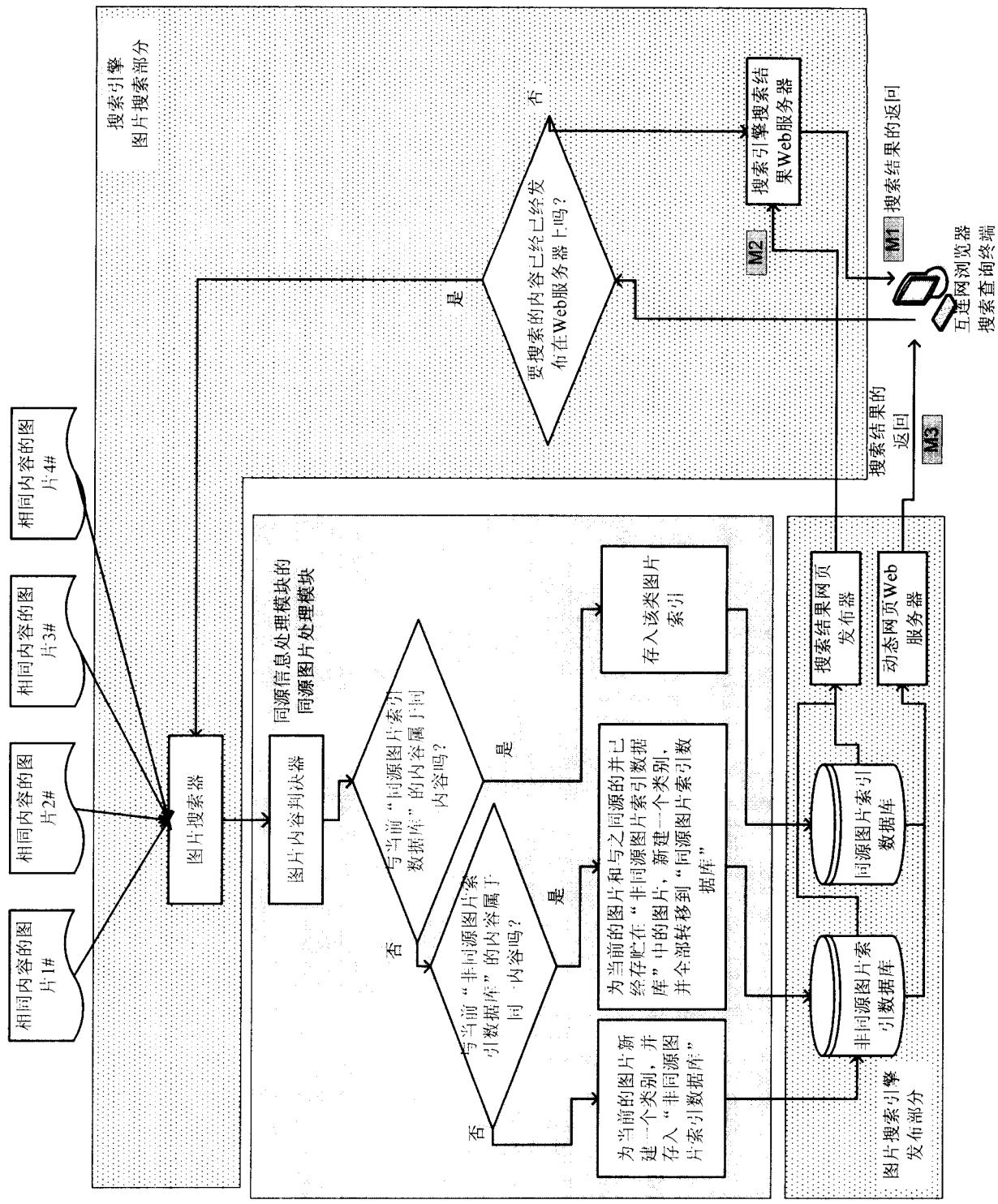


图5

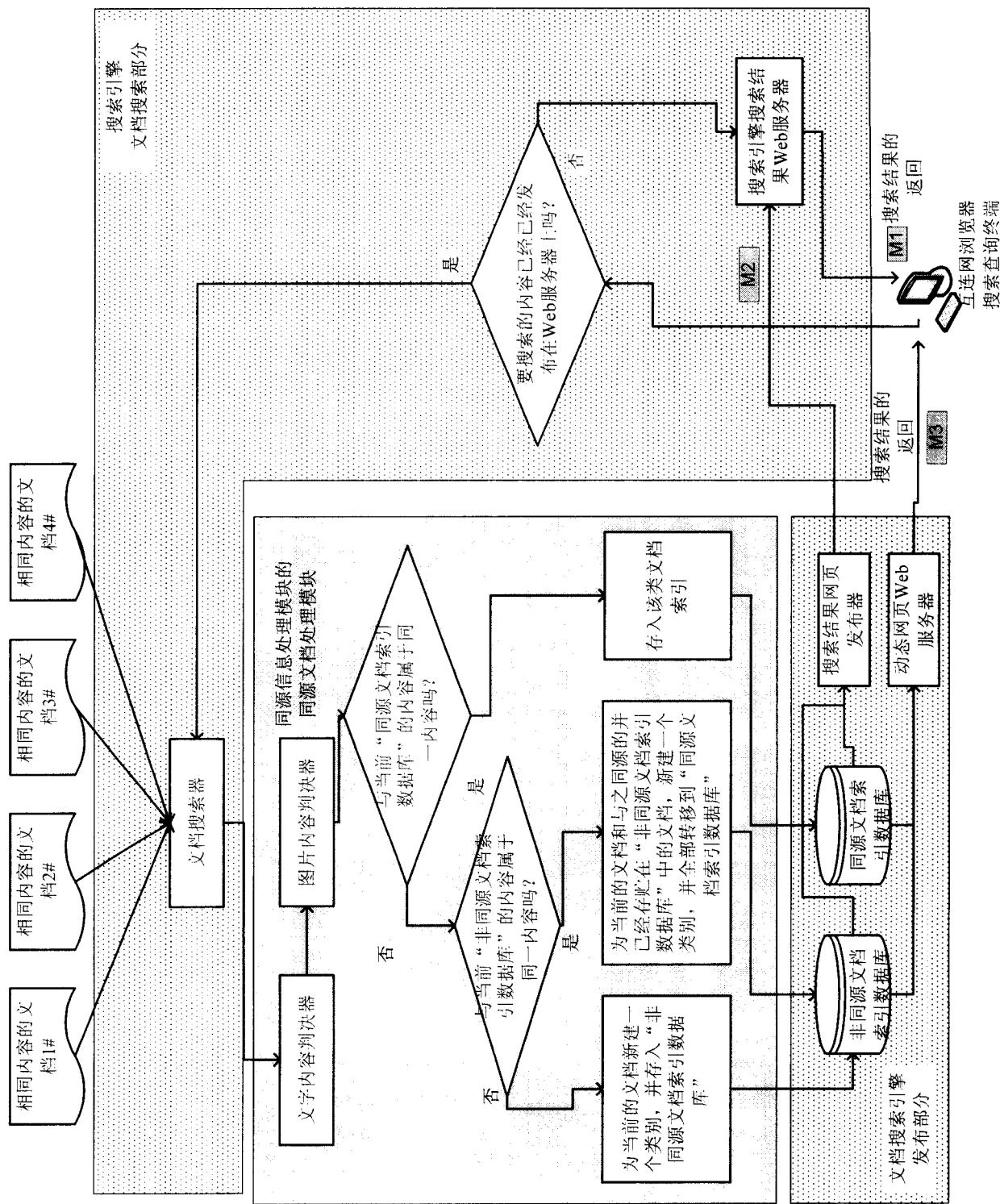


图6

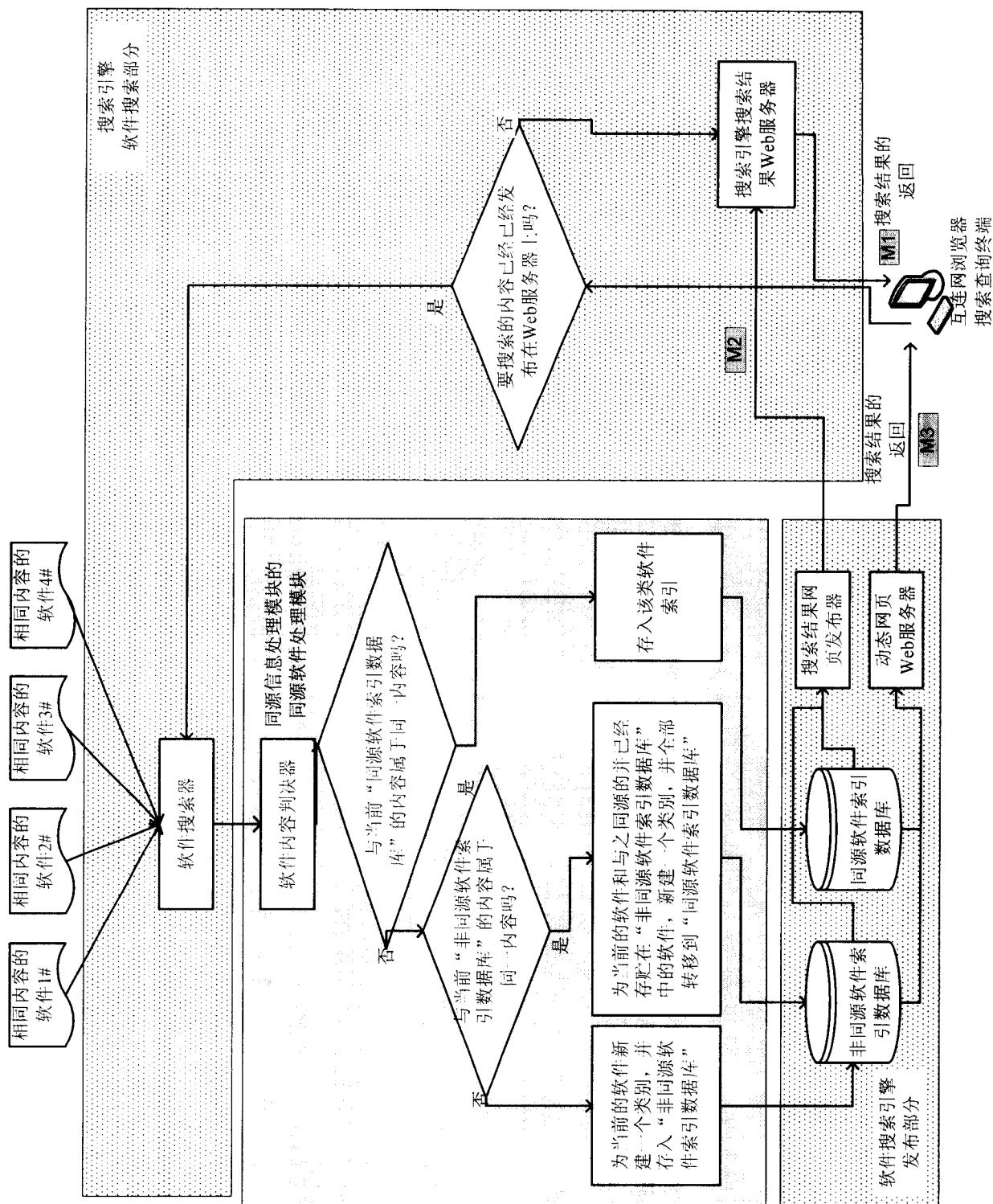


图 7

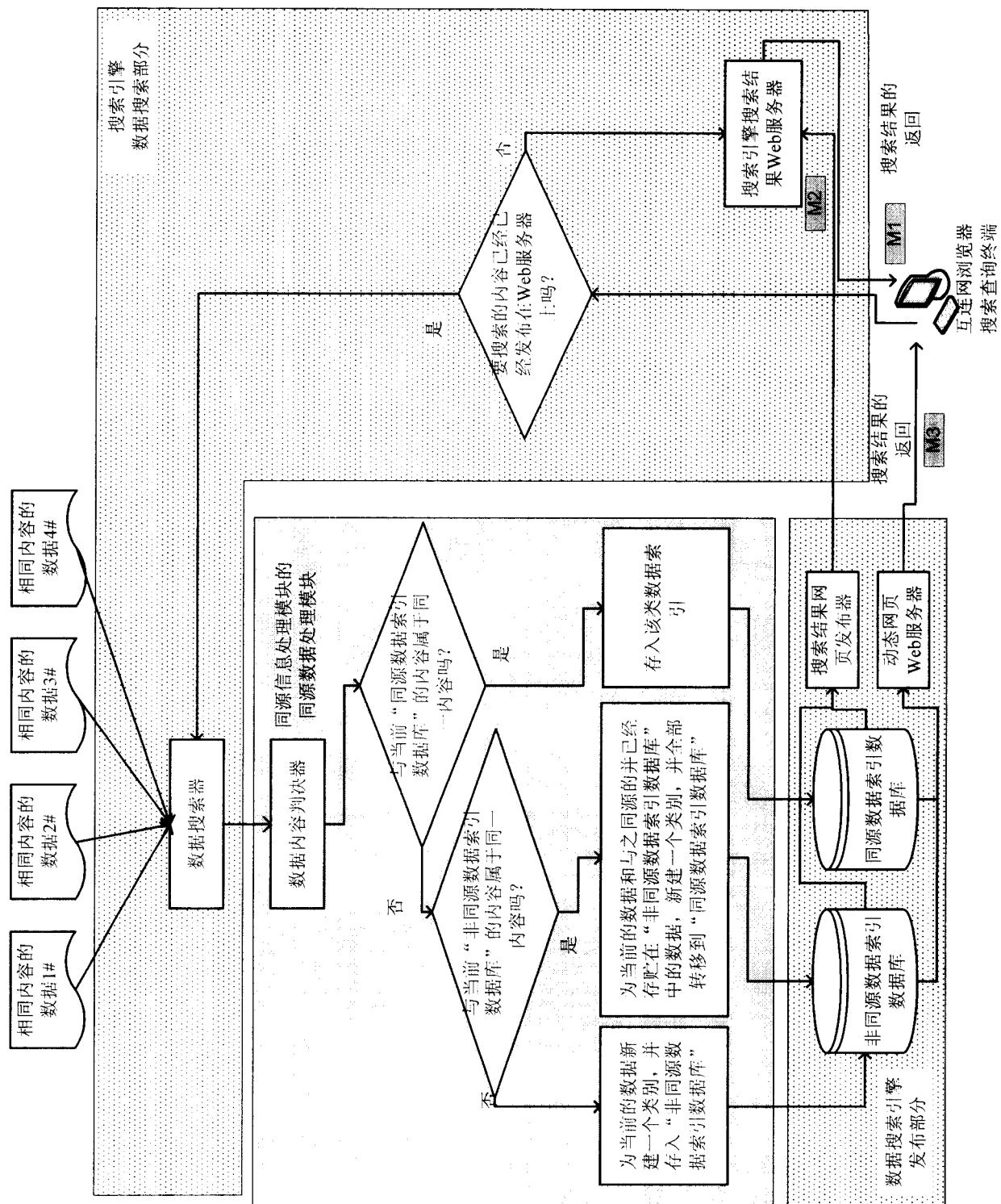


图8

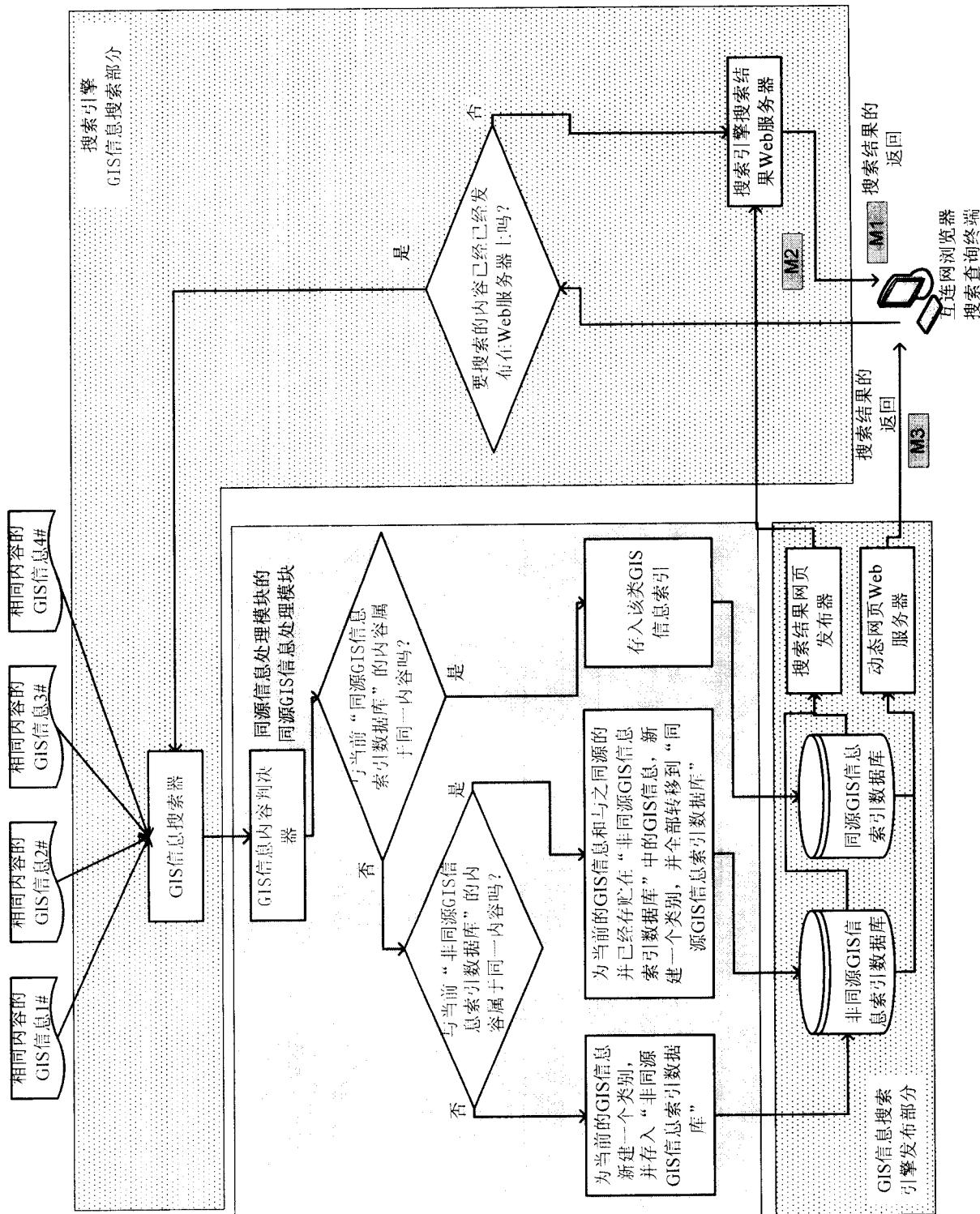


图9

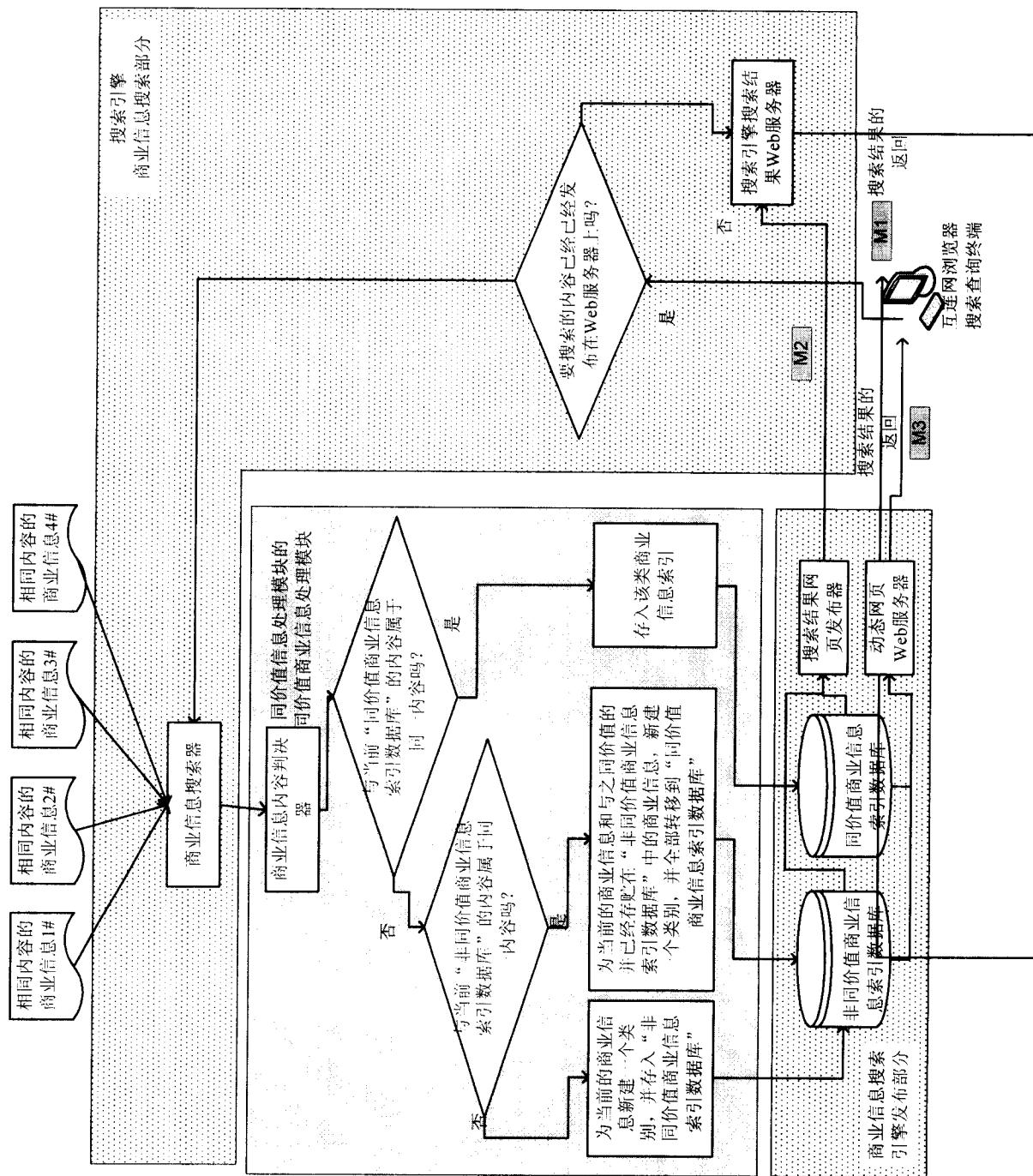


图 10

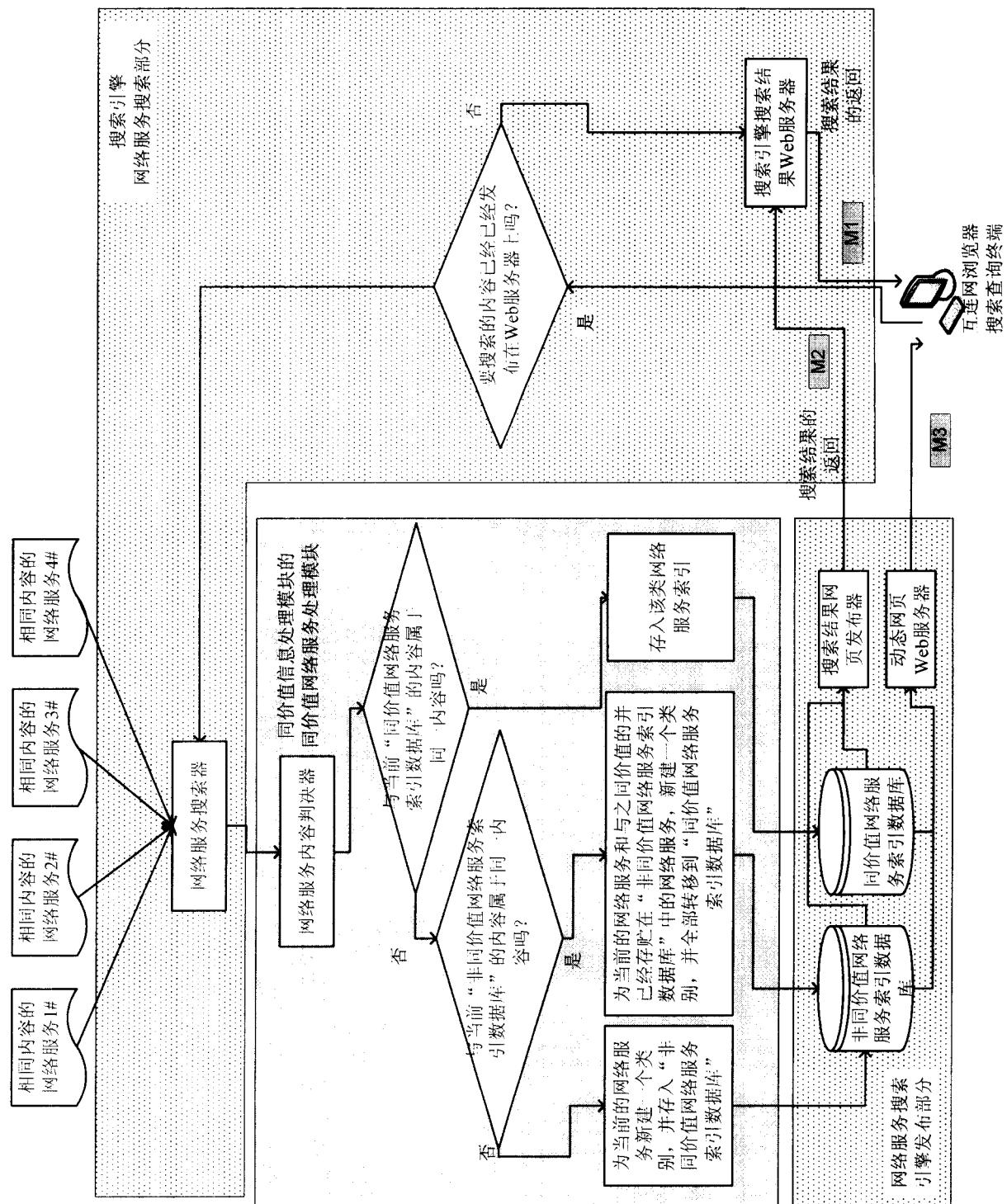


图11

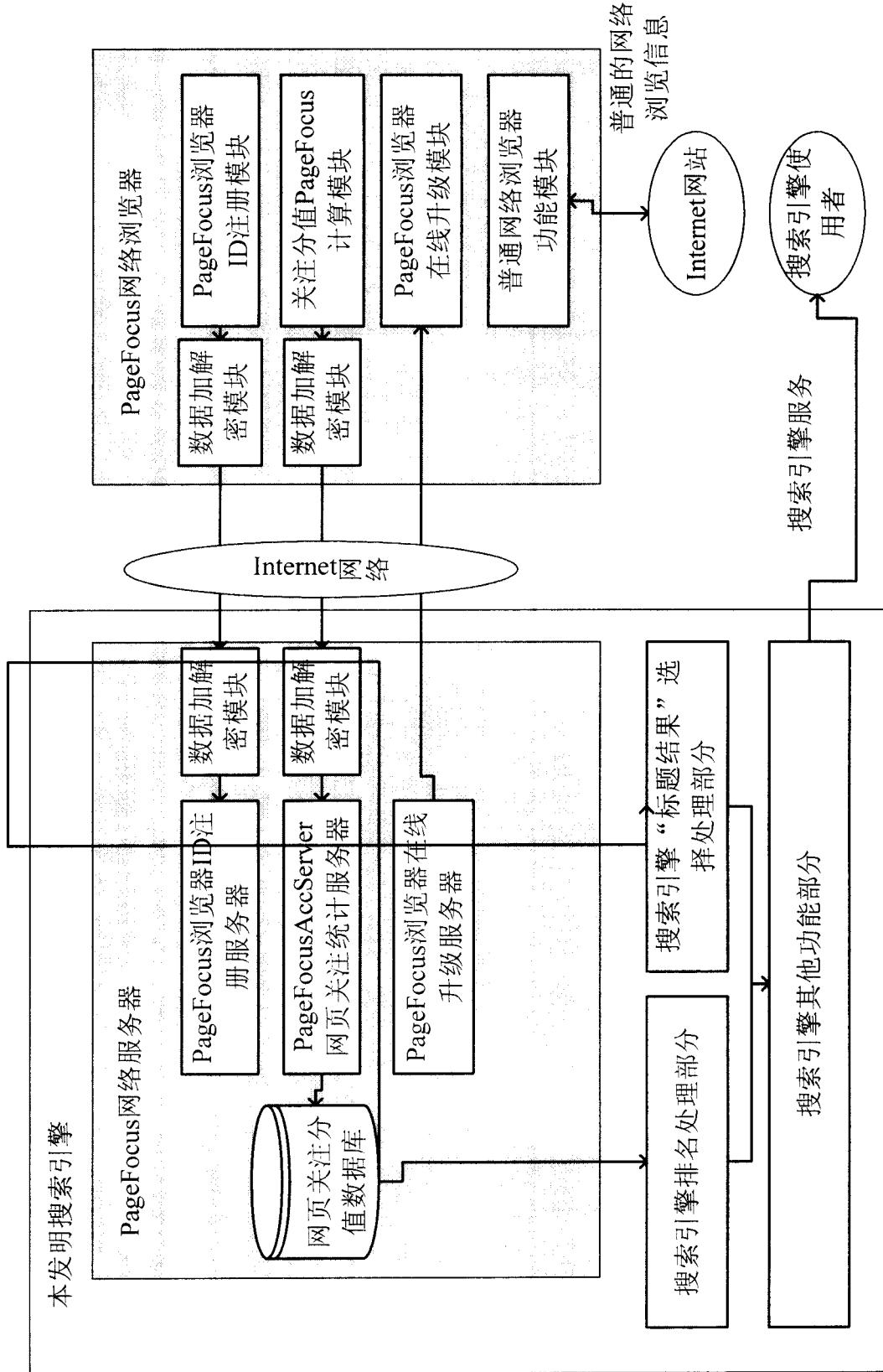


图12

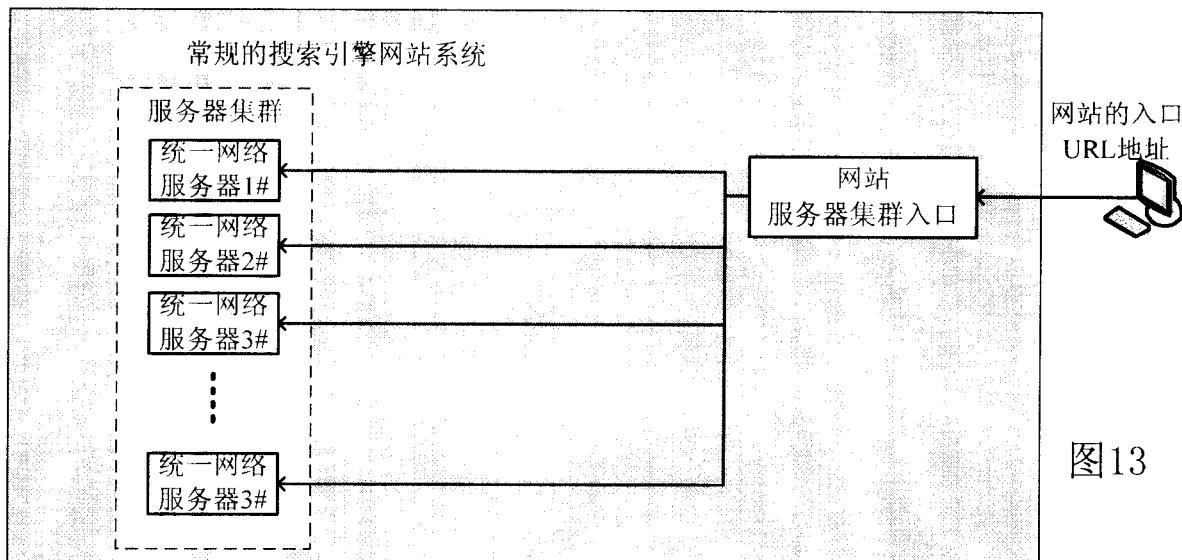


图13

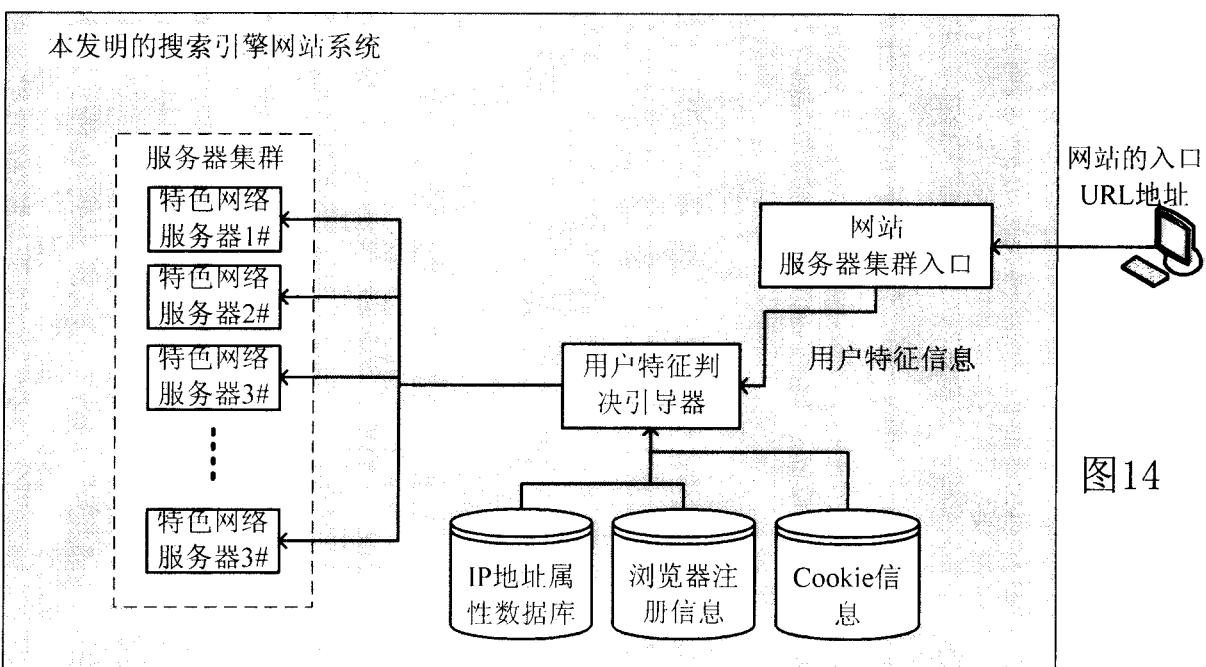


图14