

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.

G06F 17/30 (2006.01)



## [12] 发明专利申请公开说明书

[21] 申请号 200510119716.4

[43] 公开日 2006 年 7 月 26 日

[11] 公开号 CN 1808430A

[22] 申请日 2005.10.31

[74] 专利代理机构 西安通大专利代理有限责任公司

[21] 申请号 200510119716.4

代理人 李郑建

[30] 优先权

[32] 2004.11.1 [33] US [31] 60/624,249

[71] 申请人 西安迪戈科技有限责任公司

地址 710075 陕西省西安市高新区科技二路  
72 号天泽大厦 5 楼

[72] 发明人 梁平

权利要求书 4 页 说明书 44 页 附图 12 页

### [54] 发明名称

智能、图示和自动化互联网和计算机信息的检索和挖掘方法

### [57] 摘要

本发明公开了一种全新的关于信息的智能检索、挖掘、过滤、图示和检索自动化的方法、软件和系统。能够进行人工智能化信息查询，信息监视，根据概念进行搜索、过滤、图示和跟踪，以协助用户对互联网络和本地计算机的特大数量信息数据进行智能的、高质量地搜索和挖掘。本发明的方法可提取出网页或文件的重要概念和所含的具有创见的概念，对这些概念排序，并图示它们之间的关系。本发明形成的产品将应用于企业管理、规划、消费市场、市场研究、科学研究、技术开发，中高等教育，军事，国家安全，外交等领域。

100

智能搜索引擎

我选择以下组合相关的信息（全选或一个） 101

新闻、娱乐、严肃评论  
 医疗健康  
 教育知识  
 商业财经  
 科学技术  
 学习、研究  
 市场研究

健康花草  
 学校、大学  
 旅游、娱乐、运动  
 研究文献、科技、标准  
 个人、企业、社会网络  
 新闻

用最简单的语言描述您想要找的东西：

104 102 106

120 102

使用下面这个文件的内容描述您想搜索什么： 122 [搜索]

更多设置：  
查找更新时间不超过  
 任意时间  一周  一月  三个月  
 半年  一年  三年 108

选定这个时间段  
今天或从月 到 月日  
(不选择时间限制作为一次性搜索) 110

当发现新的资源或改动时，您将收到一个桌面通知。如果您也想收到邮件通知，请写下您的邮箱地址： 112

选择这些启动概念搜索以扩展搜索  
选择搜索结果到 层。 116

选择这些启动链接搜索以扩展搜索  
选择搜索结果到 层。 118 [搜索]

1. 一种使用用户提供的对搜索的描述产生搜索查询的方法，其特征在于，包括：

从用户提供的对搜索的描述里提取一或多个字、词、短语或句子作为甲集；

把甲集扩展到乙集，乙集含有一或多个和甲集中一或多个字、词、短语或句子概念上相关的字、词、短语或句子；

把乙集作为一个搜索的描述交给一个搜索程序甲去搜索含有乙集中部分或全部的字、词、短语或句子的文件。

2. 如权利要求1所述的方法，其特征在于，进一步包括下列一项或多项：

把甲集扩展到乙集时使用了一或多个知识库；

首先用甲集的一或多个字、词、短语或句子作为一个搜索的描述进行搜索，把甲集扩展到乙集时使用到此搜索的结果；

当甲集含有两个或更多个字、词、短语或句子时，乙集包括甲集、甲集中有其它甲集的字、词、短语或句子的含义支持的字、词、短语或句子的一个或多个含义的同义词；

搜索程序甲在一个网络中搜索信息；

搜索程序甲在用户的个人计算机里搜索信息。

3. 一种信息搜索方法，其特征在于，包括：

提供一个接受用户输入描述甲和描述乙来定义一个搜索的接口；

搜索含有描述甲中部分或全部信息，且不包含或包含描述乙中部分或全部信息的文件或其它信息体。

4. 如权利要求3所述的方法，其特征在于，进一步包括下列一项或多项：

描述甲或描述乙或两者都是有一或多个关键字/词组成；

把一个含有越多的描述乙中信息的文件或其它信息体排序越高。

5. 一种信息搜索方法，其特征在于，包括：

从一个含有一或多个文件或其部分的乙集里提取一或多个信息元，此一或多个信息元形成甲集；

从甲集中选出一或多个信息元形成丙集；

用丙集去获取含有一或多个文件或其部分的丁集。

6. 如权利要求5所述的方法，其特征在于，进一步包括下列一项或多项：

从乙集里提取一或多个信息元形成甲集时使用下列一项或多项来决定提取哪些信息源：字/词或短语的列单、句型的列单、概念或意义的列单、字/词或信息元和上述的一或多个列单里的项的关系、字/词或信息元的位置或格式或上下文、字/词或信息元在文本里的角色、信息元是基于哪些准则鉴别出来的、以及信息元属于哪个类别；

乙集是一个甲搜索的结果，甲搜索是由一或多个描述定义的；

当乙集是一个由一或多个描述定义的甲搜索的结果时，从乙集里提取一或多个信息元形成甲集时使用下列方法之一：

(1). 一或多个搜索引擎利用一或多个信息元和定义甲搜索的一或多个描述的相关性从乙集提取一或多个信息元形成甲集;

(2). 一或多个搜索引擎在甲搜索以前就从存在搜索引擎的部分或全部文件里预先提取一或多个信息元, 当甲搜索时, 用户的计算机从一或多个搜索引擎下载乙集文件所包含的预先提取的一或多个信息元, 用户的计算机利用下载的一或多个信息元和定义甲搜索的一或多个描述的相关性来决定由哪些信息元形成甲集;

(3). 当甲搜索时, 用户的计算机从一或多个搜索引擎下载部分或全部搜索结果, 并从其中提取一或多个信息元形成甲集;

当乙集是一个由一或多个描述定义的甲搜索的结果时, 从甲集中选出一或多个信息元形成丙集包括提供一个用户接口, 让用户选择甲集中一或多个信息元, 以用户的选择形成丙集, 并且用丙集去获取丁集包括把丙集和定义甲搜索的一或多个描述一起当作定义乙搜索的描述交给一或多个搜索程序进行乙搜索, 并由乙搜索的结果或其部分形成丁集;

当乙集是一个由一或多个描述定义的甲搜索的结果时, 从甲集中选出一或多个信息元形成丙集包括提供一个用户接口, 让用户选择甲集中一或多个信息元并可将每个选中的信息元设为存在项或不存在项, 以用户的选择形成丙集, 并且用丙集去获取丁集包括把丙集和定义甲搜索的一或多个描述一起当作定义乙搜索的描述交给一或多个搜索程序进行乙搜索以搜索含有丙集中设为存在项的信息元而且不含有丙集中设为不存在项的信息元的文件或其部分, 并由乙搜索的结果或其部分形成丁集;

从甲集中选出一或多个信息元形成丙集时基于对甲集中一或多个信息元的排序进行的;

甲集中一或多个信息元是概念, 从甲集中选出一或多个信息元形成丙集包括选择以或多个概念, 用丙集去获取丁集包括把丙集中的概念交给一或多个搜索程序进行乙搜索以搜索含有丙集中的概念的文件或其部分, 并由乙搜索的结果或其部分形成丁集;

从丁集中提取一或多个概念, 并多次重复以上方法;

甲集中一或多个信息元是链接, 从甲集中选出一或多个信息元形成丙集包括选择以或多个链接, 用丙集去获取丁集包括把丙集中的链接指向的文件或其部分纳入丁集;

从丁集中提取一或多个链接, 并多次重复以上方法。

## 7. 一种信息搜索方法, 其特征在于, 包括:

从由一或多个文件或其部分形成的甲集里提取的信息元集合中获取一或多个信息元;

对上述获取的一或多个信息元基于下列一或多个排序参数进行排序:

对一个从一组文件中提取的信息元, 基于这组文件的一个链接流行度排序的一个函数; 基于这组文件的一个相关度排序的一个函数; 基于这组文件的一个日期排序的一个函数; 一个信息元可从更多的文件里提取出来则把此信息元的排序提高; 一个信息元可从更少的文件里提取出来则把此信息元的排序提高; 一或多个信息元和一个乙集里的信息元的关系; 一或多个信息元在文体里的位置、格式或角色; 一或多个信息元出现的上下文; 一或多个信息元的含义。

8. 如权利要求 7 所述的方法，其特征在于，进一步包括下列一项或多项：

甲集是一个甲搜索的结果，甲搜索是由一或多个描述定义的；

乙集里的信息元包括一或多个重要字/词和/或短语，句型，概念或含义和论语；

提供一个用户接口让用户调动一或多个排序参数的权重。

9. 一种把文件组织成一个结构或显示此结构的方法，其特征在于，包括：

把两个或更多个文件组织成在一个甲维度上相连接的两个或更多个集，每个集的成员是基于和文件相关的信息元或文件所含的信息元决定的，两个集之间的连结意味着在这两个集之间存在一个甲关系；

把两个或更多个文件组织成在一个乙维度上相连接的两个或更多个集，每个集的成员是基于和文件相关的信息元或文件所含的信息元决定的，两个集之间的连结意味着在这两个集之间存在一个乙关系。

10. 如权利要求 9 所述的方法，其特征在于，进一步包括下列一项或多项：

甲关系和乙关系之一或两者是子集关系，意味着在一个连结一端的集是在连结另一端的集的子集；

甲关系和乙关系之一或两者是一个在一个连结两端的集之间的一个逻辑或语义关系；

在甲维度和乙维度之---或两者上有三个或更多的集连结在一起，且甲关系和乙关系之一或两者是可传递关系；

将文件组织成的结构以图论图或图像的方式显示。

11. 一种计算在搜索结果里的一个文件的排序的方法，其特征在于，包括：

在文件中识别出和用户输入的定义搜索的描述的部分或全部相同或同类或相似的一或多个匹配信息元；

基于在文件中的下列一或多个因素计算一个相关度排序参数：一或多个匹配信息元和它们在定义搜索的描述中的相应部分的相同或同类或相似的程度；两个或更多个匹配信息元出现的顺序和它们在定义搜索的描述中的相应部分出现的顺序的比较；两个或更多个匹配信息元在句子或文体结构里的相对位置；在两个或更多个匹配信息元是否出现标点符号或其它符号；一或多个匹配信息元的格式；一或多个匹配信息元在文件里的角色；一或多个匹配信息元在文件里出现的位置或部分；及是否由和专门针对一个用户的信息相似的信息出现及它们之间的相似程度。

12. 一种信息监视的方法，其特征在于，包括：

在一个浏览应用的窗口提供一个选项，用户可使用此选项选择监视正在此窗口中浏览的 URL 内容的变化或使用此窗口进行的一个搜索的变化；

当用户选择此选项，在一段时间内检查此 URL 或此搜索的内容有无变化；

如此 URL 或此搜索的内容有变化，把探测到的变化通知给用户。

13. 如权利要求 12 所述的方法，其特征在于，进一步包括下列一项或多项：

提供一个选项让用户规定监视的时间段或频率；

检查此 URL 或此搜索的内容有无变化是在用户的计算机上进行；

检查此 URL 或此搜索的内容有无变化包括在一段时间内以某个频率重复访问此 URL 并检查其内容

---

的变化，或在一段时间内以某个频率重复进行此搜索并检查搜索结果内容的变化；

检查此 URL 或此搜索的内容有无变化包括计算并储存此 URL 或此搜索在甲时间的内容的计算校验和或数字摘要，将甲时间储存的计算校验和或数字摘要和在甲时间后的一个时间由此 URL 或此搜索的内容计算的计算校验和或数字摘要进行比较；

14. 一种保护信息的方法，其特征在于，包括：

将一或多个文件或其部分的一或多个特性、信息元或内容的描述保存在甲集里；

含有甲集部分或全部信息的文件或其部分形成乙集，要求用户通过一或多个保护措施才允许用户读或写乙集里的文件或其部分或得到在乙集里的文件或其部分的信息。

15. 如权利要求 14 所述的方法，其特征在于，进一步包括下列一项或多项：

允许用户读或写乙集里的文件或其部分或得到在乙集里的文件或其部分的信息是为用户进行一个搜索，并包括将用户提供的对此搜索的描述和甲集的信息相比较以决定是否要求用户通过一或多个保护措施才进行此搜索；

甲集进一步包括一或多个规则以决定用户可对含有甲集部分或全部信息的文件进行哪些操作；

检查并标记一或多个文件是否含有甲集部分或全部信息，将标记为含有甲集部分或全部信息的文件加入乙集。

## 智能、图示和自动化互联网和计算机信息的检索和挖掘方法

### 技术领域

本发明涉及信息的检索技术领域，更具体的，是关于用于在本地计算机和网络上多台计算机上进行信息的检索、挖掘、过滤、图示和检索自动化的方法及其该方法实现的软件和系统。本发明的优先日是2004年11月1日交到美国专利局的预备申请（申请号 60/624, 249），同时也是2004年12月28日交到中国国家知识产权局的专利申请（申请号 200410073518.4）的部分继续。

### 背景技术

以下列出了当前网络搜索方法的局限性：

1、现有技术的网络搜索方法经常返回大量的结果，例如，一个搜索的条目会有成百上千甚至百万的结果。在实际的应用中，用户不可能在限定的时间内去读取所有的信息。大多数的用户不会读取超过前10到30条搜索结果。结果是用户经常看不到有用的或重要的信息。这就使搜索引擎返回的数千、百万的网页成为无效页。它降低了搜索引擎索引和搜索数十亿网页的有效性。把如此多的搜索结果组织起来的需求已得到广泛的验证。也有以前的搜索引擎使用预先确定的分类或标签或聚类技术。预先确定的网页的分类方法需要一种给定的分类组织。聚类技术比如 Clusty.com 分类搜索结果就是通过从部分搜索结果中提取聚类词来实现的。由于聚类是属于统计性质，它经常会给出没有意义或不相关的聚类。与本发明相比较，以前的聚类技术不仅在提取正确的和重要的词和概念处存在不足，而且它们在多个属性中存在重复文件，不便于用户选择多个属性对搜索结果进行过滤。

2、以前的搜索引擎强迫用户去使用关键字/词或字串去搜索信息。有时，一个用户不知道使用哪些合适的关键字/词进行搜索。更理想的方法是接受用户用自然语言来描述他所寻找的信息。

3、使用以前的搜索方法，用户经常要坐在计算机前，花费数小时去找寻需要的信息。用户需要手动点击和跟踪链接、用已完成的搜索结果中的概念重新描述搜索、和等待大型文件的下载。

4、对于用户来说，以前的搜索技术没有有效的解决方案来监视站点和搜索结果。用户通常需要在一段时间里用多组搜索关键字/词重复地进行搜索，以查看新的信息是否出现或最近访问的网站是否发生变化。

5、在以前的有些搜索技术中，用户必须对互联网和个人计算机分别进行搜索来发现存在这两处的有关信息。在以前对本地计算机上的文件进行索引搜索的解决方案中，在本地计算机上的搜索界面不同于互联网络搜索时用的浏览器界面。另外的一些用相同的界面来显示网站搜索和本地计算机文件搜索的解决方案把这两种搜索被捆绑在一起。即使当一个用户只需要搜索他的计算机硬盘上的文件时，搜索的关键字/词也被发送到网的搜索引擎，泄露了用户的个人行为，这是没有必要的。在有些以前的实现中，当计算机没有接入互联网时，本地的文件搜索不能进行。

6、搜索引擎接收到，通常也记录下用户使用的搜索关键字/词串，这会泄露用户发给搜索引擎的意图或创造性的想法。在有些时候，它涉及到用户的个人隐私或机密。

由前所述可很显然的看出需要发展一种更先进的或智能的方法检索和挖掘互联网和计算机上的信

息，以克服上面提到的各种缺陷。

## 发明内容

本发明的目的在于，提供一种智能、图示和自动化互联网和计算机信息的检索和挖掘方法，该方法包括网搜索的改进，概念搜索，文本挖掘，从搜索结果中提取概念，用户可选择对搜索结果根据概念进行过滤，概念聚类以及统计和逻辑关系的图示化，自动深入和扩展搜索，自动改变探测和跟踪，本地计算机文件搜索，相关的队列或者概念队列，把 META 和用户隐私分开。是一种高级智能搜索，信息挖掘，管理，图示化以及分析工具，它给用户提供了一个空前的能力。

本发明提供了一个非常必要的工具及其方法，它能够帮助用户迅速看到包含在大量的搜索结果中的重要的概念，可作为对搜索结果的概要。它把搜索结果中重要的概念进行提取和排序，计算出它们的统计。这里可能有很多的概念，本发明可让用户选择搜索结果中的概念和其它特征，并以此选择对搜索结果进行过滤、排序、分类。对于其他重要的概念，它也提供了一个基于搜索结果所含的概念对搜索结果的聚类、和搜索结果之间的统计的和逻辑关系的图示法，因而，它使得用户可对大量的搜索结果中所含信息和搜索结果之间的关系尽快达到理解。同时，通过从搜索结果中提取具有特征的重要的概念和他们的统计信息，提供给用户一个更好的信息挖掘的方法。它不仅提取出现频率最高的概念，称为 MPC（最流行概念），而且也提取重要但是出现频率较低的概念，称为 MOC（最新鲜概念）。概念排序可基于和搜索的相关度，搜索结果的统计信息，链接流行度，和新鲜或稀有度。无论是 MPCs 和 MOCs，它可能被排序在前。用户可以选择或排除从搜索结果中所提取的重要概念对搜索结果进行筛选，还能基于从搜索结果中所提取出来的重要概念来细调一个搜索或是改变一个搜索的方向。同时，基于从概念路径图中重要概念的统计以及逻辑关系，本发明也提供了一个图示化的搜索结果的聚类。通过目录和搜索结果的关系，概念路径图给用户提供了一个快捷的形象化和操纵搜索结果的方法。这些都提供了比先前的搜索细化（“Refine Search”）和聚类的技术方法更加灵活和有效的手段。

本发明提供了一个自然语言用户界面，用户能够运用自然语言来描述他所要查找的信息，而无需使用准确的关键字/词。本发明在自然语言的基础上，完成自然语言处理和自动公式化的搜索。本发明通过把搜索关键字/词扩展到由同义词，从属词，关键字/词的子类词，一个概念的首字母缩写形式或完全表达形式等所构成的概念来扩展搜索。同时使用两个或更多的关键字/词之间语义的相互加强来加深理解，通过这种方式从搜索关键字/词的多重语义中适当地消除歧义。

本发明通过自动跟踪链接，使用先前搜索找到的概念再表示搜索以达到加深关键字/词搜索的目的，从而使搜索过程大大地自动化。同时，它也能为用户从搜索结果中自动下载大量的文件。运用这种方法，用户不必再数小时的坐在计算机前用手去点击链接以跟踪一个搜索路径和等待大量文件的下载。实际上，这种搜索是自动的，它或在后台操作，使用户能够做其他的工作，或使用户能够离开计算机去做其他的事情。

本发明提供了一个完整的界面，它允许用户使用相同的、熟悉的浏览界面去搜索互联网和他的计算机，以此获取相关的信息，但对于电脑中机密或和安全相关文件的搜索受用户控制。在此，用户个人电脑的信息搜索就意味着搜索在个人电脑的硬盘上或在一个本地网络的一台计算机上的文件，包括邮件(比如 Microsoft Outlook, Outlook Express, Eudora)和应用文件(比如 Microsoft Word, Excel, Power Point, Adobe pdf, text, Word Perfect, html)和其他包含文本或对文件标题或其属性由文本描述的文件。

本发明提供了一个有效的自动化的方法，它使得用户可以监视选取的网站，监视一个或多个搜索得到的新结果，用户不必再去点击完成搜索或在一段时间内去重复性的浏览。

本发明同时为用户提供了一个方法可以不把所有具有启迪作用的关键字/词都泄漏给任何一个单一的搜索引擎就可进行搜索。运用这种方法，没有任何一个搜索引擎全部接收搜索用户的关键字/词列表，这样，就避免搜索引擎去揣测用户的真正意图或侵犯用户的隐私。它保护了用户的隐私和机密。

#### 附图说明

- 图 1 展示一个智能搜索引擎，它接受一个用户用自然语言描述和自动搜索；  
图 2 展示一个查询发生器的接口；  
图 3 展示一个智能搜索引擎，它接受搜索关键字/词和关键字/词到概念的拓展和自动搜索；  
图 4 展示一个用图表、过滤和图示法的搜索结果表示的一个的用户界面；  
图 5 展示把本发明嵌入到一个网络搜索引擎界面的工具栏中的智能搜索的接口；  
图 6 展示用列表、过滤和图示法的搜索结果表示的一个用来完成图表 5 中一次搜索的一个用户界面接口；  
图 7 展示用分离窗口在本地计算机上显示列表、过滤和图示法的查询结果；  
图 8 展示概念路径图的例子，8 (a) 一个 MPP CPM，8 (b) 一个 MOP CPM，和 8 (c) 一个 MPP CPM 的替换形式；  
图 9 展示一个用户界面窗口的 MPP CPM 的例子，一个结点包含了在 912 中高亮显示的重要概念所包含的网页或文件  
图 10 展示了本发明的索引文件的原理框图和数据库的一个接口；  
图 11 展示了一个可调整的三层工具栏界面使用户可以在其上进行排序项权重的调整；  
图 12 展示了一个将搜索本地计算机硬盘和本发明的新特色进行一体化的改进的一个搜索界面；  
图 13 展示了一个为进行网络搜索建立的本发明的部分实现的高水平的流程图表；  
以下结合附图和发明人给出的具体实施的例子对本发明作更进一步的详细描述。

#### 具体实施方式

本发明的描述将引用图示，在文中的同一数字将代表图示中的同一个部件或部分。下面将描述本专利的实现例子。这些实现例子是用来描述本发明的有关方面，而不应被解释成为限制本发明的范围。当实现例子用到方块图、结构或流程，每一块部件或步骤既代表方法里的一个步骤，也代表实现方法的装置里用于实现一个步骤的一个部件。取决于实现方式，一个装置的部件可由硬件、软件、固件或它们的组合来实现。

为了更清楚的理解本发明，在具体说明本发明之前，发明人给出以下定义：

**概念：**在本发明，当应用于把关键字/词甲或短语甲扩展到它的内涵时，这个词代表和关键字/词甲或短语甲有相同或相近含义的关键字/词或短语的集。这个集可能包括关键字/词甲或短语甲的同义字/词或短语，以及它的母类词和子类词。在本发明中，有些时候，概念、关键字/词或搜索关键字/词或搜索关键字/词串这些术语可交替使用。在这种情况下，它就意味着这些关键字/词、搜索关键字/词或搜索关键字/词串是一个概念的代表。当应用于从文件、网页或搜索结果中抽取有代表性的或被特定规则或标准认为重要的字/词或含义时，概念，或可交替使用的重要概念，就是按照一个或多个规则或标准从一个页或文

---

件中所提取出的关键字/词、关键字/词字串或短语。它可以被扩展到一个具相同或近似含义的关键字/词或短语集。

**文件：**在网络搜索背景下的文件意味着应用一个搜索引擎可以找到的网页或文件。在计算机硬盘进行信息检索的背景下的文件意味着所有文件存储在计算机硬盘或本地网络上的文件。文件的例子包括但不限于微软 Word、Excel Spreadsheet、PowerPoint、PDF、电子邮件、txt、xml、html 和任何含有文字内容的信息体。

**硬盘搜索：**搜索在用户计算机的一个或多个硬盘上或在用户本地网络中计算机上的文件。

**关键字/词，短语：**当术语关键字/词或短语单独使用时，它意味用户用这些关键字/词或字串描述他想要搜索的信息。

**搜索关键字/词，查询关键字/词，搜索关键字/词字串，查询关键字/词字串，搜索短语、查询短语：**搜索时实际使用的关键字/词或关键字/词串。它可从用户提供的关键字/词或短语中生成，但可能不同于这些。在有些情况下，它可能由本发明的查询发生器生成。

**词义：**一个字或短语的含义。一个字或短语可能有多种含义。

**同义词集：**一个字的一组同义词。

**引号内的字/词串表示用于和此字/词串精确匹配的搜索。**为方便，定义一个搜索的**搜索关键字/词或描述**，或有关文件或文件所含的任何信息，如一个字/词、字/词串、短语、句子、句型、论语、概念、链接、文件的 URL、文件类型、日期、标题或作者、等等，都被称为一个信息元。

## 1、智能查询生成器和查询扩展

本发明提供给用户一个如图 1 中所展示的自然语言界面 NLI (Natural Language Interface) 100，而不会强迫用户使用一个关键字/词字串来进行搜索。在这个实现中，在 box 102 里用户可以输入其搜索的一个自然语言描述 NLDS (Natural Language Description)，或就像使用传统搜索引擎那样输入一个关键字/词字串，或使用关键字/词字串和自然语言描述的一个组合。

在一种实现中，在 NLI 的顶部，有一个用户意图表 UIL (User Intentions List) 104 用来让用户明确其搜索意图。在一种实现中，控件“全部选中”101 是默认选中的，并能够搜索和返回所有找到的东西。用户也可以忽略并且不使用 UIL 104。用户的意图可以从 NLDS 102 中提取。同时有按钮 106 可供用户输入关键字/词串进行查询。

运行在用户本地计算机上的 QG (查询生成器) 从 NLDS 中提取出字或字串，并把它们当作搜索关键字/词或字串提交给搜索引擎，或把它们当作搜索关键字/词或字串来执行搜索。个性化搜索可通过两种途径实现：一种是用户对检索的描述和 UIL (如果用户使用了 UIL)；另一种是根据用户偏好以及存储在本机上的搜索结果的历史记录。这种个性化的搜索保护了用户的隐私，因为用户的历史搜索记录或偏好都保存在用户的本地计算机上，而不是搜索引擎。

除了直接从用户对其搜索的描述中提取搜索关键字/词字串外，QG 同样包含了一个自然语言理解模块 202，一个将关键字/词扩展成概念的模块 208，一个安装在用户本地计算机上的知识库 210。这个知识库可用来把用户的自然语言描述解释和翻译成相应的关键字/词，同时将关键字/词扩展成为概念，如表 2 所示。例如，当一个用户输入自然语言描述“我要找一个可以把我所有计算机都连到英特网上的器件” (“I am looking for a device that will be able to connect all my computers wirelessly to the Internet”) 后，自然

语言理解模块 202 利用包含了无线网络知识的知识库 210，将用户描述转换成关键字/词串，如（无线路由器 wireless router），（无线接入点 wireless access point），（WLAN 路由器 WLAN router），（无线宽带路由器 wireless broadband router），等等。另外再举一个例子，当一个用户输入自然语言描述“我想买一台无线路由器把我所有计算机都无线地连到英特网上”（“I want to buy a wireless router that connects all my computers wirelessly to the 互联网”）后，搜索关键字/词串提取模块 204 将利用其中包含有关无线网络的知识库 210，提取关键字/词串（无线路由器 wireless router），（计算机无线地连到英特网 connect computer wirelessly 互联网）；自然语言理解模块 202 和将关键字/词扩展成概念的模块 208 将解释用户的搜索意图为（买 to buy），（购物 to purchase），并把提取的关键字/词串扩展到（无线路由器 wireless router），（无限接入点 wireless access point），（WLAN 路由器 WLAN router），（无线宽带路由器 wireless broadband router），（802.11 路由器 802.11 router），（家庭网络 home networking），等等。

同样，NLI100 给用户提供了更多选项来筛选他的搜索，包括 108 修改日期范围，一个保持一个搜索在一段时间里存活的选项（根据在 110 中指定的日期范围，每隔一定的时间就激活搜索，以检测有无新信息源和已有信息源有无变化），以及当检测到变化时，在本地计算机上通知用户或发送一个邮件给用户的选项。为了这个目的，NLI100 也给用户提供了在 112 输入他的邮件账户的输入框。其它的选项包括下面的 116 的概念跟踪和 118 的链接跟踪选项，在搜索中它们在初始搜索的基础上扩展搜索范围。这些特性将在本发明后面部分将进行详细的说明。

在一种实现里，如果用户点击按钮 106，将出现一个关键字/词用户界面（KUI）300，如图 3 所示。这个 KUI 300 不同于以前的搜索引擎界面之处在于 KUI 300 包含了一个 UIL 302，一个关键字/词扩展成概念选项（按钮 304 和 306），一个“可能用到的词”输入区 308，日期范围过滤器 310，搜索存活日期范围 312 和电子邮件通知用户选项 314。用户输入到 KUI 300 关键字/词串被发送到位于 QG200 中的搜索关键字/词产生模块 206。如果按钮 304 和/或 306 被选中，QG200 将使用一个将关键字/词扩展成概念的模块 208 将用户输入的关键字/词扩展成为概念。然后，QG200 中的搜索关键字/词产生模块 206 将基于用户输入的关键字/词串和关键字/词扩展到概念的结果产生搜索关键字/词串去完成此次搜索，或将其提交给搜索引擎。UIL 302 默认是“全部选中”，这意味着 UIL 中所有意图被选中。这样，本实现将搜索所有能发现的文件并返回。

在另一个实现中，UIL 可能被忽略。本实现可提供给用户按钮 320 来选择 NLDS 界面 100 进行搜索。在另一个实现中，自然语言理解模块 202 和搜索关键字/词串提取模块 204 抽取及产生的关键字/词串被送到将关键字/词扩展成概念的模块 208。208 联合知识库 210 将具有相同或相近含义的字或短语加入关键字/词串。这样，即便用户用不同的词或短语描述他要找的信息，包含这些信息的网页和文件也能被抽取出来。

类似以前的搜索引擎，某些普通词可以不包括在搜索关键字/词当中，比如：是、的、个、而且、也、又、等，除非用户用引号将这些词装入一个句子中，或他们是唯一的查询字。

在以上所有的实现中，关键字/词串的提取和将用户自然语言描述翻译成相应的关键字/词字串都是在用户本地计算机上完成的。在另一个实现中，这些功能都是在搜索引擎服务器中实现的。这样做的好处是关键字/词字串提取模块 204，自然语言理解模块 202 和知识库 210 的维护和更新都能在一个集中的机器上完成。用户的本地计算机直接提交用户的自然语言搜索描述给搜索引擎。在搜索引擎上实现这些功

能的不利条件在于它可能引起搜索引擎的负荷过重。在另外一个实现中，有些功能的实现是利用大量的本地计算机的处理能力来实现的，有些功能的实现是在搜索引擎上实现的，以使用搜索引擎中维护的最新的关键字/词字串抽取方法，自然语言理解方法和知识库来进一步的处理或提高本地计算机的结果抽取和结果翻译的效果。

在一个实现中，当用户的计算机连接到互联网或访问一个搜索引擎或一个服务器时，它将与服务器进行通信，该服务器能够给 QG 的部件提供更新，以使它们保持最新。这些模块包括在用户本地计算机上的搜索关键字/词串提取模块 204，那么将关键字/词扩展成概念的模块 208，自然语言理解模块 202 和知识库 210。更新可以在每次本地计算机连接到互联网，或用户访问一个搜索引擎或服务器时完成，或周期性的去完成。

## 1.1 提取搜索关键字/词串和搜索意图

### 1.1.1 从 NLDS 中提取搜索关键字/词串和搜索意图

在提取关键字/词包含在 NLDS 中的情况下，本发明识别和提取嵌入在 NLDS 中的搜索关键字/词。在某一实现中，它是通过使用已知的句型和线索字获得的。每种语言，比如英语，中文，法语，德语，都有非常频繁使用的特定句型和线索词来描述一个搜索。

在某一实现中，搜索关键字/词字串提取实现 204 扫描寻找 NLDS 中的下列搜索描述：意图，搜索关键字/词，可能词，日期范围，信息源，页类型，被排队在外的其一事物。

在一个 NLDS 中，一个搜索的题目及/或意图很有可能在和下面给出的句型相似的一个或多个句子里给出：

我要找.....的信息	找.....的信息
我要找（或写、理解、学习、调查、研究，等）.....	搜索....
我希望找.....	我对.....感兴趣
（我的）搜索的目的（意图、目标、等）是.....	我的目的是.....
因为.....我搜索.....	.....是要找的.....
..... 是搜索的焦点（目的、动机、等）	.....

在上面所列举的例子句型中，搜索的主题或关键字/词一般是包含在上面句型中的“....”部分。这样，主题或搜索关键字/词及/或意图就能从这样的句型中抽取出来。本发明可建立了一个可以识别这些句型的句型数据库或列表。自然语言处理和理解领域中的自然语言理解算法以及人工智能算法可应用于抽取这样句式中的主题或搜索关键字/词或搜索意图。同样存在一些句子模式，通过这些模式程序可以推断出用户是需要关于这一主题的部分还是全部的信息，例如：

我想得到任何关于.....的信息。	搜索所有关于.....的信息。
查找任何和.....相关信息。	.....

如同在以前的搜索引擎中，用户也可能只是在 NLDS 中键入孤立关键字/词。比如输入（无线网络）、（家庭网络）。这些没有完整句子结构的名词可以用句子部分分析、词类分析和句子结构分析这类自然语言理解算法很容易识别出来。这些算法可用于识别和抽取这一类孤立的搜索关键字/词。

根据一些线索词和短语，也可以确定搜索的目为购物。这些线索词和短语包括：便宜、更便宜、最便宜、低（更低、最低）价、买、购买等等。这些线索词和短语指出用户有很大可能在进行一次关于买

东西的信息的搜索。因此与关键字/词相关的零售商和发明的网站应该在搜索列表中排在前边。这个方法也要包括对于异常情况的处理。例如，买 (buy) 这个词出现在“买或自制 (buy or make)”和“买和自做的对比 (buy vs. make)”中时，它说明搜索是为了做一个是购买还是自己制造的决定，搜索有很大的可能是为了调查竞争和市场的信息，而不是为了搜索零售商和商品从而进行购物。根据这些线索词和短语以及异常可以构建一个数据库或列表并且用它来抽取搜索的目的。

同样的，本发明也可以建立用于指导别的代表或过滤搜索特征或域的抽取的数据库或列表，其中包括：可能用到的词、日期范围、来源、网页类型和排除。

在 NLDS 中，对于“可能用到的词”的搜索以以下句子模式出现的几率非常高。

可能包括..... 可能用到以下的词.....

这些词语可能用到..... 应该包括.....

.....可能包括在内 可能用到的词有.....

.....

“可能用到的词”也可以通过识别包含和“可能”相近义的词的句子来提取。和“可能”相近义的词可以用一个列表来表达，这个列表可包含：(好像、可能、也许、应该、.....)。本实现在搜索是可不包含、包含部分或是包含所有“可能用到的词”，并把搜索结果中包含较多“可能用到的词”的页将要排序在包含较少或是没有包含“可能用到的词”的页的前面。在 NLDS 中，一个搜索的时间范围很可能出现在以下的句子模式中。

网页应该在近期修改过的（写的、发表的、.....） 日期范围.....

返回在过去.....修改的或发表的.....

在 NLDS 中，搜索结果来源的说明很可能出现在以下句子模式中。

我的兴趣是大学（制造商、公司、非盈利机构、等） 仅返回.edu 的搜索结果

只搜索英文（中文、澳大利亚、等）网站 .....

在 NLDS 中，对于搜索页类型的说明很可能出现在以下句子模式中。

只搜索 html（word、pdf、等）网页 仅返回 Word（pdf、html、等）的搜索结果

.....

在 NLDS 中，搜索中需要排除项的说明很可能出现在以下句子模式中。

我不想获得..... 不要搜索.....

不包括..... .....

本实现也可从搜索结果中排除掉包含排除关键字/词的网页和文件。

本发明可建立类似以上的句子模式的数据库或列表，并使用这类数据库或列表来鉴别和抽取表达搜索的各种特征。自然语言处理或理解和人工智能领域中自然语言理解算法可以应用在从这类句子模式中提取出的搜索的各类特征。

这个发明可使用搜索关键字/词抽取排除列表 (Search Word Extraction Exclusion List SWEEL) 来排除一些很常见但是对得到特定信息没有用处的词。在这个列表中的词将不会被认为是搜索关键字/词。SWEEL 列表包括如下的词 (是、这个、那个、我们、她、他、它、然后、而且、不但、但是、..... 等等)。

自然语言理解算法可以从 NLDS 中识别出关键字/词中的或关系。在一个实现中，除非一个关键字/词被识别为“或”或是“可能用到的词”，否则它将被认为与别的关键字/词是与的关系。一个实现可把抽取的关键字/词（和概念的扩展，在下一节中将要描述）按照识别出来的关系“与”或“或”在一起，以及包括“可能用到的词”和排除“可能用到的词”的各种不同情况进行搜索。

在另一种实现中，NLDS 将不会被输入到输入框 102 中，而是将它写入一个文件中，如后缀名为.doc, rtf, pdf, 或.txt 的电脑中的文件里。这个发明将提供选项使用户可以选择这个文件作为 NLDS 并从此文件中产生关键字/词进行搜索。这个功能通过用户向输入框 120 键入文件路径或是点击按钮 122 进行浏览来实现，之后程序将加载用户指定的文件并把它作为 NLDS。

这个发明同样可以从一般的描述和例句或文章中抽出关键字/词，这些一般的描述和例句或文章不一定是为了描述一个搜索而写的 NLDS。例如，用户可以可以在 120 中输入文件路径或在 102 中输入如下文字：“一个无线安全代理使用认证服务器来管理用户认证” (“A wireless security agent uses an authentication server to manage user authentication”)。自然语言理解模块 202 将分析这个句子并可抽取如下搜索关键字/词：(无线安全 wireless security), (安全代理 security agent), (认证 authentication), (认证服务器 authentication server), (用户认证 user authentication)，并且可使用它们来进行搜索。在更高一层，自然语言理解模块 202 不仅可以抽取关键字/词，还可以分析句子中的结构。在此例中，可以提取出：主语（无线安全代理 wireless security agent）、谓语（使用 uses）、直接宾语（认证服务器 authentication server）、副语从句（管理用户认证 manager user authentication，可以进一步分解成动词和宾语）。在这个例子中，一个实现可首先使用抽取的搜索关键字/词串来粗略的搜索，然后可以从粗略搜索的结果中提取出含有和上述一般的描述和例句或文章相似或同义的主语、谓语、直接宾语和副语从句，以及在这些部分间有相似逻辑关系的网页或文件。

在有些情况下用户并不知道使用哪一个词可以描述自己搜索的目标。在这种情况下用户可能会使用描述性语言来描述搜索目标的特征、特性或功能。前面所说的用户向 NLDS 中输入“我要找一个可以把我所有计算机都连到英特网上的器件” (“I am looking for a device that will be able to connect all my computers wirelessly to the Internet”) 就是这样的一个例子。在这种情况下，自然语言理解模块可以使用知识库 210 将用户的描述图示到可能的专业词汇同时生成相应的搜索关键字/词串。在一些专业领域，例如医学、技术、生物学、地理等，本发明可以建立此类领域的领域定义和关系知识库，或把此类领域的领域定义和关系知识库包含在知识库 210 中。

### 1.1.2 从 KUI 中抽取搜索关键字/词串

对于那些习惯于使用以前的搜索引擎和关键字/词串进行搜索的用户，本发明提供了更为实用的功能 KUI 300。在 KUI 300 里，用户可以点击按钮 320 启动 NLI 并使用 NLDS 进行搜索。KUI300 与以前的搜索引擎有如下不同：

- KUI300 提供了 UIL302 让用户可以说明他的搜索意图。例如：购买产品、寻找学习资料、市场研究等等。比起个性化方法试图猜测用户的意图来，KUI 300 允许用户明确地指出自己的意图，这样本发明可以呈现给用户正确信息。当然用户也可以通过点选 301 中的“全部选中”选项跳过这一步。在一种实现中这个选择框在默认下选中的。在别的实现中 UIL 可能被省略。

- 用户可以点击选择按钮 304 或/和 306 使用本发明提供的将用户输入的关键字/词和短语扩展为概念

的功能。将关键字/词扩展成概念的模块 208 与知识库 210 协同工作，将关键字/词和短语扩展使其包含了同义或近义的词和短语，从而保证了搜索可以得到含有用户想得到的信息但是用了和用户的关键字/词和短语不同的文字表达的的页和文件。

- KUI 300 包括了“可能用到的词”输入项 308。308 输入项允许用户输入他也不确定是否会出现自己想要得到页或文件中出现的词和短语。没有以前的搜索引擎可以提供这一功能。

- 与 NLI 100 相似，KUI 300 同样提供对时间过滤 310、保持搜索存活一段时间的选项 312 以获取新的信息源和变化、电子邮件通知选项 314、概念跟踪选项 316 和链接跟踪选项 318。这些将在下面的章节中进行详细的讨论。

用户在输入框 303、305、206 和 309 中输入的关键字/词串将被发送到 QG 200 的搜索关键字/词串生成模块 206。如果按钮 304 和/或 306 被选中，QG 200 将使用将关键字/词扩展成概念的模块 208 将用户输入的关键字/词串扩展为概念集合，例如使其包括与关键字/词同意或近义的词或短语。在此之后根据用户输入的关键字/词和关键字/词扩展到概念的结果，QG 200 中的搜索关键字/词串生成模块 206 将生成搜索用的关键字/词串用来进行搜索或是提交给一个搜索引擎。

本实现可以给用户提供在各个输入项输入内容的例子，以便用户了解如何输入以进行搜索，比如：

输入项 303： 太阳系，火星，生命存在的证据	输入项 308： 红行星，爬行器
-------------------------	------------------

输入项 305： 我相信火星上有生命，热火星	输入项 309： 火星人，外空人
------------------------	------------------

上述可能用到的词的实现提供了一种新的信息搜索方法，该方法包括：

提供一个接受用户输入描述甲和描述乙来定义一个搜索的接口；搜索含有描述甲中部分或全部信息，且不包含或包含描述乙中部分或全部信息的文件或其它信息体。

这个方法还进一步包括下列一项或多项：描述甲或描述乙或两者都是有一或多个关键字/词组成；把一个含有越多的描述乙中信息的文件或其它信息体排序越高。

## 1.2 关键字/词到概念的扩展

这个发明提供了两种关键字/词到概念的扩展方法，说明如下。

### 1.2.1 使用关系字典、领域定义和关系知识库（Domain Ontology）和知识库进行概念扩展

下面给出一种实现的步骤并使用用户输入关键字/词串（上涨的油价 rising cost of oil）作为例子来说明。我们使用 WordNet 作为关系字典的例子来提供词义和同义词集合，同时它还会以指向母类词和子类词（或称为母类词和从属词， hypernyms and hyponyms/troponyms）等的链接形式显示出相关词语之间的概念等级关系。（注：用和英文 WordNet 同样的方法可以建立中文的 WordNet）

1. 首先要获取用户输入的关键字/词的词根和所有变形，删除简单词和联结词（例如：的、了、而且、然后， of, in, at, on, and, is, with 等等），并且生成输入关键字/词的扩展关键字/词列表。举例说在英文里，rising 的根词是 rise。英文的关键字/词扩展列表是 ((rising, rise, rose, risen, rises), cost, (oil, oiled, oiling, oils))。

2. 如果关键字/词甲只有一个意思，那么将这个意思以及和关键字/词甲此意的同义词形成关键字/词甲的查询集合（Query Set QS）。

3. 如果关键字/词甲不只一个意思，那么将它的每一个意思和描述与其它关键字/词的所有意思和描述进行一一比较。如果关键字/词乙的第二个意思的同义词集合中包含了关键字/词甲的第一个意思的同义

词集合中的词, 或关键字/词乙的第二个意思的描述与关键字/词甲的第一个意思的描述相似, 这时关键字/词甲的第一个意思将被选择而且将它的同义词集合加入到关键字/词甲的 QS 中。关键字/词乙的第二个意思也会被选择并且将它的同义词集合加入到关键字/词乙的 QS 中。这种方法叫作互增强 (Mutual Reinforcement MR) 或交叉确认 (Cross Validation CV)。以关键字/词 (上涨 rising, 价 cost) 为例。下面是英文 WordNet 对于 rising 和 cost 返回的结果。

名词 rise 有 10 个意思 (前 6 来自赋加了标记的文本)

1. (9) rise -- (a growth in strength or number or importance)
2. (3) rise, ascent, ascension, ascending -- (the act of changing location in an upward direction)
3. (1) ascent, acclivity, rise, raise, climb, upgrade -- (an upward slope or grade (as in a road); "the car couldn't make it up the rise")
4. (1) rise, rising, ascent, ascension-- (a movement upward; "they cheered the rise of the hot-air balloon")
5. (1) raise, rise, wage hike, hike, wage increase, salary increase -- (the amount a salary is increased; "he got a 3% raise"; "he got a wage hike")
6. (1) upgrade, rise, rising slope -- (the property possessed by a slope or surface that rises)
7. lift, rise -- (a wave that lifts the surface of the water or ground)
8. emanation, rise, procession -- ((theology) the origination of the Holy Spirit at Pentecost; ...)
9. rise, boost, hike, cost increase -- (an increase in cost; "they asked for a 10% rise in rates")
10. advance, rise -- (increase in price or value; "the news caused a general advance on the stock market")

动词 rise 有 17 个意思 (前 16 个来自赋加了标记的文本)

1. (30) rise, lift, arise, move up, go up, come up, uprise -- (move upward; "The fog lifted"; "The smoke arose from the forest fire"; "The mist uprose from the meadows")
  2. (23) rise, go up, climb -- (increase in value or to a higher point; "prices climbed steeply";...)
  3. (20) arise, rise, uprise, get up, stand up-- (rise to one's feet; "The audience got up and applauded")
  4. (8) rise, lift, rear -- (rise up; "The building rose before them")
- .....

名词 cost 有 3 个意思 (前 3 个来自赋加了标记的文本)

1. (379) cost -- (the total spent for goods or services including money and time and labor)
2. (53) monetary value, price, cost -- (the property of having material worth (often indicated by the amount of money something would bring if sold); "the fluctuating monetary value of gold and silver"; "he puts a high price on his services"; "he couldn't calculate the cost of the collection")
3. (17) price, cost, toll -- (value measured by what must be given or done or undergone to obtain something; "the cost in human life was enormous"; "the price of success is hard work"; "what price glory?")

按上述的处理步骤将选择名词 rise 的第 9 个意思、动词 rise 的第 2 个意思以及名词 cost 的第 2、3 个意思。这是因为它们都包括词 value、cost 或是因为与 value 和 cost 的概念相关。因此包括 (rise, rising, rose, risen) 的 QS 现在将包括 (rise, boost, hike, cost increase, rising, rose, risen, go up, went up, gone up, going up, goes up, climb, climbed, climbing, climbs), 而包括 (cost) 的 QS 现

在将包括 (cost, price, monetary value, toll) .

如果没有找到关键字/词甲的互增强或交叉确认,那么可以把关键字/词甲的第一到第三个词义的同义集或所有词义的同义集添加到关键字/词甲的 QS 中。再一个实现中,把多少个词义的同义集添加到 QS 中取决于词义的使用频率或词义在赋加了标记的文本中的使用来决定,使用频率很低的词将会被删除。词义在赋加了标记的文本中的使用由 WordNet 或类似的电子词典提供,在上面的例子里显示在词义序号后的括号 () 中。把上述方法用于中文将产生 (上涨) 的 QS 为 (上涨、上升、上爬、高涨、涨、上增、猛涨、攀高、……), (价) 的 QS 为 (价、价格、费用、单价、要价、批发价、零售价、……)

#### 4. 对所有的关键字/词重复以上操作。

5. 将每个关键字/词的所有被选中意思的母类词和子类词的同义词集合加入到那个关键字/词的 QS 中。在选母类词的同义词集合时,可以在母类词层次结构中向上走一层或两层。在一种实现中,在母类词层次结构中向上走一层词同义词集合将被加入到一个关键字/词的 QS 中,而对于母类词层次结构中向上走的第二层,只有当第二层的母类词和选中的第一层的母类词同义词集或描述、或和关键字/词本身选中的同义词集或描述有很大重合时才会被加入到一个关键字/词的 QS 中。这里所谓的很大一部分可以理解为超过 50% 或是多于两个字/词。我们将以关键字/词 (rise) 为例来说明这个步骤。(rise) 的第 2 个意思和母类词在 WordNet 中是这样的:

Sense 2: rise, go up, climb -- (increase in value or to a higher point; "prices climbed steeply"; "the value of our house rose sharply last year")

=> grow—(become larger, greater, or bigger; expand or gain; "The problem grew too large"; ...)

=> increase – (become bigger or greater in amount; "The amount of work increased")

=> change magnitude -- (change in size or magnitude)

向上第一层母类词是 (grow), 第二层母类词是 (increase)。第一层和第二层母类词的描述都包含 (become, bigger, greater), 所以来自这两层的同义词集合都将被加入到关键字/词 (rising) 的 QS 中去。为简化操作,可以只选择第一层母类词,即在本例中只是加入 (grow)。在中文,用 (油) 做例子,(油) 的母类词向上一层是 (燃料), 向上一层是 (能源)。能源和燃料的描述有很大相同,所以这两层母类词都可以加入 (油) 的 QS 中。

一种方法可向下一层寻求子类词。对于母类词和子类词,只有与已经包含在 QS 中的关键字/词的同义词不同或不包含在 QS 中同义词集合中的字/词或词串时才会被加入到 QS 中。以关键根词 (oil) 的第一个意思为例,它有子类词 (fuel oil, lubricating oil, crude oil, crude, petroleum 等等.)。因为 (fuel oil, lubricating oil, crude oil) 已经包含关键字/词 (oil), 而包含 (fuel oil, lubricating oil, crude oil) 的文件会在查询关键字/词 (oil) 时被检索到,所以只有子类词集中 (crude, petroleum) 才会从被加入到它的 QS 中。相对而言,包含 (crude, petroleum) 的文件将不会在对关键字/词 (oil) 的搜索中被检索到。因此 (crude, petroleum) 将被加入到关键字/词 (oil) 的 QS 中。同样的,在中文里,(油) 的子类词可包括 (石油、原油、汽油、柴油、润滑油、煤油)。因这些词都含 (油) 字,他们就不必被加入 QS 中。

如果是因为关键字/词乙的第二个意思的 MR (互增强) 而选中了关键字/词甲的第一个意思,同时关键字/词甲的第三个意思的子类词集与第一个意思的同义词集合或子类词有交集,那么第三个意思的同义

词集和与第一个意思有交集的第三个意思的子类词的同义词集也将被加入到关键字/词甲的 QS 中。

在一个实现中，只是对名词性和动性的词义进行母类词和子类词的概念扩展。同样，这种概念扩展也可以应用在形容词性和副词性的词义上。

完成上述后，搜索关键字/词串生成模块 206 将使用所有关键字/词的 QS 生成供搜索使用的关键字/词串。搜索关键字/词串生成模块 206 对从每个关键字/词扩展而来的词使用或 (OR) 关系，并且对用户输入的关键字/词使用不同的与关系组合。在 (上涨的油价 rising cost of oil) 的例子中搜索关键字/词串生成模块 206 可以生成以下的搜索。

中文：(上涨 OR 上升 OR 上爬 OR 高涨 OR 涨 OR 上增 OR 猛涨 OR 攀高...) AND (价 OR 价格 OR 费用 OR 单价 OR 要价 OR 批发价 OR 零售价、.....) AND (油 OR 燃料 OR 能源 .....), 英文：(rise OR boost OR hike OR “cost increase” OR “go up” OR climb OR grow OR increase) AND (cost OR price OR value OR toll) AND (oil OR crude OR petroleum)

请注意每个词的不同形式，例如 rise, rising, rose 等等，并不包含在上面的例子中。一种实现也可以包含它们。处理根词不同变化形态的匹配可以在搜索算法阶段或在查询生成算法阶段的得到处理。这个发明的实现可以构建地和任意解决方案接口。

对于用户使用 NLI100 输入的查询描述或关键字/词，如果一个实现不能确定用户对于抽取或生成出的关键字/词之间是想使用与关系还是或关系时，QG200 将使用多种与或组合来进行搜索，而搜索结果排序依赖于以与关系组合在一起的关键字/词个数。包含以与关系组合在一起的所有关键字/词的搜索结果排在最前面。例如，QG200 可以为关键字/词组合产生更多的搜索：(上涨 OR 上升 OR 上爬 OR 高涨 OR 涨 OR 上增 OR 猛涨 OR 攀高...) AND (价 OR 价格 OR 费用 OR 单价 OR 要价 OR 批发价 OR 零售价、.....), (rise OR boost OR ....) AND (cost OR price OR value OR toll), (价 OR 价格 OR 费用 OR 单价 OR 要价 OR 批发价 OR 零售价、.....) AND (油 OR 燃料 OR 能源 .....), (cost OR price OR value OR toll) AND (oil OR crude OR petroleum)。但是搜索：(上涨 OR 上升 OR 上爬 OR 高涨 OR 涨 OR 上增 OR 猛涨 OR 攀高...) AND (价 OR 价格 OR 费用 OR 单价 OR 要价 OR 批发价 OR 零售价、.....) AND (油 OR 燃料 OR 能源 .....), 或(rise OR boost OR hike OR “cost increase” OR “go up” OR climb OR grow OR increase) AND (cost OR price OR value OR toll) AND (oil OR crude OR petroleum) 的结果将会被排在最前面。

自然语言理解模块 202 可以使用句子部分、词性、词类和角色分析算法去分析一个关键字/词是否是一个名词、动词或形容词等等。这可以用来限制在关键字/词到概念扩展时关键字/词的哪些词义将被选择。在作决策时可以使用一些简单的规则。例如，在 (rising cost of oil) 中，如果跟随在 of 后的词是在标点符号前的唯一词或是关键字/词串的末尾，那么自然语言理解模块 202 可以使用“of xxx”模式来确定 xxx 是一个名词。因此在这个例子中 oil 被确定为名词。自然语言理解模块也可以使用“of a/an/the xxx yyy”或“of xxx yyy”模式来确定 xxx 是一个形容词而 yyy 是一个名词，自然它们必须要有相应的词义。自然语言理解模块可以使用诸如辨别在一个句子中的词的词类的简单语言和语法规则来获得很高的正确可能性。这样达到了减少处理的目的，而且 100% 的正确率在这个应用中也是没有必要的。

如果不能确定关键字/词是名词、动词还是形容词等等，那么将关键字/词扩展成概念的模块 208 将使用关键字/词的名词和动词词性或是这个关键字/词的所有词性包括形容词和副词。

### 1.2.2 使用搜索结果的概念扩展

通常来说搜索返回的页和文件都会包含搜索关键字/词的定义、概念扩展、意义和描述。因此这个发明的另一个实现将会解决关键字/词含糊问题。这个发明还会使用与搜索关键字/词匹配的搜索结果文档的上下文和同时出现的词来将关键字/词扩展为概念等同的词的集合。

举个例子来说，一个用户 NLI 100 或 KUI 300 输入关键字/词（QoS）或（WLAN）进行查询。如果知识库包含相关的领域知识，它们就可以被扩展为包括（QoS，服务质量，“quality of service”），（WLAN，“无线局域网”，“wireless LAN”，“wireless local area network”，802.11，802.11a，802.11b，802.11g，WEP，WPA，...）的查询序列。查询可以通过应用概念扩展后的关键字/词进行。然而，如果知识库不包含相关的领域知识，检索只能根据用户输入的关键字/词（QoS）或（WLAN）进行检索。这样的检索结果中很可能包含包括缩写词的定义，本发明就可以使用自然语言处理算法较易的识别和抽取这种信息，比如可以通过搜索如下的句子模式：

QoS=服务质量...	QoS （服务质量） ...
xxx 称为 (or 叫做、缩写为、等) yyy... 服务质量 (QoS) ...	
无线局域网=WLAN...	WLAN 的意思是无线局域网...
.....	
QoS=Quality of Service...	QoS （Quality of Service） ...
Quality of Service (QoS) ...	wireless local area network=WLAN...
xxx is referred to as (or called, abbreviated as, etc) yyy...	
WLAN means wireless LAN...	.....

同样，在检索关键字/词 WLAN 得到的结果中，无线局域网，802.11，802.11a，802.11b，802.11g，WEP，WPA，无线路由器，wireless router，宽带，broadband，宽带，家庭网络，home networking 等词也会以很高的频率出现。这样，本发明可以通过将输入词的查询结果作为知识背景来扩展用户查询，通过这种方式查询的结果比通过一个实体维护的知识库的方式更准确，因为互联网是动态分布式的，它的信息在快速地更新。在上面的例子中，通过应用用户的查询结果，输入关键字/词（QoS）和（WLAN）的查询被扩展成为相当于输入（QoS，服务质量，“quality of service”），（WLAN，无线局域网，“wireless LAN”，“wireless local area network”，802.11，802.11a，802.11b，802.11g，WEP，WPA，无线路由器，wireless router，宽带，broadband，宽带，家庭网络，home networking，...）的查询。

在一个具体的实现中，本发明应用 202 自然语言处理实现，204 输入序列抽取实现和 206 查询词产生实现来分析查询结果，找到定义、等价概念、缩写和查询词相关概念等。应用的方法有句式分析、上下文分析，并发性分析和联想分析等。QG 200 扩展那些有 MR 或可以应用 202 自然语言实现，210 知识基础库和域本体来理解的查询词。获得查询结果后，应用自然语言理解算法在查询结果返回的部分文档中抽取诸如高频词以及与关键字/词高度相关的词等来扩展查询。在另一个具体的实现中，QG200 应用除关键字/词之外的用户输入或抽出的关键字/词来进行概念扩展并进行一次独立的搜索，在搜索结果返回部分文档的基础上应用自然语言理解算法来抽取和查询词同时出现的词语用来扩展查询。

关于这些具体实现的另一些例子如下：

用户输入（软件定义无线接发器 Software Defined Radio），通过使用查询结果上的分析，查询被扩展成为（SDR，软无线，自知无线接发器 cognitive radio）。

用户输入（PSA），通过使用查询结果上的分析，查询被扩展成包括如下关键字/词的一组查询（前列腺特定抗体，Prostate-Specific Antigen，前列腺癌，prostate cancer，自由 PSA，free PSA，fPSA，复 PSA，complex PSA cPSA，pro PSA pPSA，切片化验，biopsy）。

用户输入（无线网络 wireless networks），通过使用查询结果上的分析，查询被扩展为包括如下关键字/词的一组查询（WLAN，无线局域网，wireless local area network，802.11，GSM，3G，蜂窝网，cellular networks，……）。

此类查询扩展方法也可应用于本发明的概念跟踪的实现，这将在以后的章节讨论。

本发明的查询生成和概念扩展的实现提供了一种使用用户提供的对搜索的描述产生搜索查询的新方法，该方法包括：

从用户提供的对搜索的描述里提取一或多个字、词、短语或句子作为甲集；

把甲集扩展到一个含有一或多个和甲集中一或多个字、词、短语或句子概念上相关的字、词、短语或句子的集合，称这个集合为乙集；

把乙集作为一个搜索的描述交给一个搜索程序甲(称为搜索程序甲)去搜索含有乙集中部分或全部的字、词、短语或句子的文件。

上述方法可进一步包括下列一项或多项：把甲集扩展到乙集时使用了一或多个知识库；首先用甲集的一或多个字、词、短语或句子作为一个搜索的描述进行搜索，把甲集扩展到乙集时使用到此搜索的结果；当甲集中含有两个或更多个字、词、短语或句子时，乙集包括甲集、甲集中有其它甲集的字、词、短语或句子的含义支持的字、词、短语或句子的一个或多个含义的同义词；搜索程序甲在一个网络中搜索信息；搜索程序甲在用户的个人计算机里搜索信息。

## 2. 用户概念选择、特征过滤和概念路径图

### 2.1 搜索引擎或本地机的概念过滤和图示

概念过滤和图示的用户界面如图 4 所示，在本发明的这一实现中，概念抽取、过滤和图示（在后面详细讨论）通过一个搜索引擎实现。

如图 1、图 3 所示，用户访问一个预定的搜索引擎的网址时，搜索结果被显示在如图 4 所示的浏览器窗口中。在面板 400 中，如果用户点击了“启动硬盘搜索”选项时，网上搜索得到的结果被显示在中间的面板 408 上，同时用户本地计算机的搜索结果被显示在右侧面板 410 上。在本发明中，硬盘用户本地机硬盘或用户机所在局域网上的硬盘。用户 PC 机或局域网上的计算机都称为本地机或本地计算机。

在具体的实现中，为了明显的区分按钮的选择状态和非选择状态，如“启动硬盘搜索”按钮，当按钮被点击或选中时，它变为高亮度显示，或变化它的颜色或亮度。另外，用户可以通过鼠标拖拽的方式调整面板 408、409 和 410 的宽度。

搜索结果的网页或文件中包含的前 N 个最重要的概念被显示在左侧的 412 面板上，N 是一个正整数，它允许用户设定或采用系统默认值。N 可以通过选择按钮 405 设定，也可以通过输入框 406 设定，N 会被自动限制小于抽出出的概念总数。注：在一个实现中，从结果中抽取出来的概念可能和用户输入的关键字/词相同。

左侧的面板包括以下几个部分：第一部分 412 显示查询结果中取出来的最重要的 N 个概念。在一个具体的实现中，重要概念列表默认显示并且允许用户通过在重要概念列表上进行概念的选择和排除操作来过滤结果。另外的部分 416 允许用户通过其它特性如文件类型、更新时间和域名等来过滤搜索结果。

在 412 部分，紧靠着每一个概念，是一个“选中”选择按钮 420 和“排除”选择按钮 421 来供用户选择和排除概念。当用户使用一个或一组“选中”或“排除”按钮选择时，这个搜索引擎过滤网上返回的结果，将只包含用户输入关键字/词或 NLDS，同时包含用户选择概念并且不包含用户选择排除概念的结果显示在中间面板 408 上。装入用户机的本地搜索程序过滤本地搜索结果，将只包含用户输入或本机搜索引擎抽出的关键字/词，同时包含用户选择概念并且不包含用户选择排除概念的结果显示在右侧面板 410 上。在一个具体的实现中，如果有网页或文件包含的概念被选中的越多，在 410 或 408 面板上它的排序位置越高。

在一个具体的实现中，一旦一个概念（不同于用户原始输入关键字/词）被选择或排除，搜索结果将立即根据这个选择的变化进行过滤。在这个实现中，用户的原始输入检索词被放在了重要概念列表的第一位，这个概念会被自动设为选中状态。用户可以取消选中，当用户取消选中或排除这个概念，并且选择概念列表中的其它概念时，搜索引擎和本地搜索程序认为这是用户通过选择概念、排除概念（如果用户选择了排除概念）的设置进行的新搜索。这样，搜索引擎和本地硬盘搜索程序就会进行一次新的搜索。在另外一个实现中，一个新搜索是这样定义的：用户取消选择或选择排除原始输入关键字/词，在 412 部分选择其它概念，并且/或在选择框 426 种输入新的概念，点击选择按钮 427。以上实现帮助用户根据他对返回结果的理解调整自己的搜索。他可以对原始输入关键字/词串取消选定或排除，选择或排除 412 中的重要概念，也可以在 426 输入框中键入新的关键字/词来重构自己的搜索。

左侧面板底部的输入框 426 被用来添加搜索关键字/词。用户可以选择概念（这些概念可以包括也可以不包括原始输入关键字/词），可以在 426 输入框中输入新的关键字/词，这个关键字/词可以被扩展为概念，点击搜索按钮 427 应用选择和输入的关键字/词和概念来做另外一个搜索。如果用户的原始输入关键字/词被选择，这个搜索将是在原始搜索结果中的一个提炼。如果用户的原始输入关键字/词没有被选择，这个搜索将是一个新搜索。

在另外一种具体实现中，原始输入关键字/词没有被列在面板 412 或 612 的重要概念列表中。提供一个“在结果内搜”和一个“新搜索”按钮。当用户点击“在结果内搜”按钮时，将按照原始输入关键字/词和新输入的关键字/词进行检索。当用户点击“新搜索”按钮时，按照用户新输入的关键字/词进行检索。

在一种实现中，当用户使用最前面的 N 个概念进行概念过滤后，根据匹配的搜索结果中的概念更新重要概念列表。在另外一个实现中，当用户使用概念进行结果过滤时，重要概念列表并不改变而是维持原始的结果，这样用户可以继续在原始结果上进行概念过滤。还有一种实现方式是，用户可以选择使用上面的任何一种方式。

用户界面的 412、416、612 和 616 中显示的“统计”指的是和它同一行的重要概念或过滤特征的统计情况。在一种实现中，这些统计是包含某个重要概念/关键字/词或符合这个过滤特征的网页和文件的数目。在另外一种实现中，“统计”这一项包含更多的统计信息，如一个重要概念在搜索结果中出现的总次数等。

搜索引擎可以预先对网页做概念抽取。在一种实现中，概念抽取是独立于搜索的。这样，在用户构建一个搜索前，一个搜索引擎的网页和文件中的重要概念可以被抽取出来，一个概念-网页/文件索引  $B_{SE}$

可以在搜索引擎上建立。同样的方式，为了支持关键字/词检索，可以建立关键字/词-网页/文件索引  $A_{SE}$ 。这种方式下，当用户应用索引  $A_{SE}$  和用户输入关键字/词检索到一个网页或文件时，这个网页或文件包含的重要概念可以通过索引  $B_{SE}$  立即得到。类似的，一个网页/文件-概念索引  $C_{SE}$  也可以事先在搜索引擎上建立。在一种实现中，本发明针对某个搜索引擎的网页和文件的概念抽取、过滤和图示（在以后的章节详细讨论）可以事先被执行。本地计算机或局域网上的概念抽取、过滤和图示通过一个运行在用户计算机上的程序建立。这种实现的过程如下：

1. 用户通过使用搜索引擎接口 100 或 300 或类似于 Yahoo 和 Google 的常规的搜索引擎接口输入 NLDS 或关键字/词来初始化一个搜索。一个控制程序探测到这个事件，将搜索请求和描述发送到本发明实现的一个搜索引擎实现上去，如果选择了硬盘搜索，同时发送给硬盘搜索程序。
2. 本发明实现的一个搜索引擎识别用户搜索企图并抽取出关键字/词序列。将关键字/词用于概念扩展，并构造关键字/词串来进行搜索。如果用户使用的是类似于 Yahoo 和 Google 的常规的搜索引擎接口，用户输入的关键字/词可以直接用来构建搜索。
3. 如果选择了硬盘搜索，控制程序启动安装在用户机上面的硬盘检索程序抽取出关键字/词，进行概念扩展并生成用来搜索的关键字/词串。如果用户使用的是类似于 Yahoo 和 Google 的常规的搜索引擎接口，用户输入的关键字/词可以直接用来构建搜索。如果没有选择硬盘搜索，跳过这一步。
4. 搜索引擎使用已建立的关键字/词-网页/文件索引 ( $A_{SE}$ ) 来搜索包含查询关键字/词的网页和文件。然后使用已建立的网页/文件-概念索引 ( $C_{SE}$ ) 获得搜索结果中包含的重要概念。搜索引擎将网页/文件以及概念排序，向运行在用户本机上的界面程序返回排序后的搜索结果列表和前 N 个概念。这个用户界面程序在接口 400 中相应的部分进行搜索结果、概念和概念路径图的显示。在一种实现中，搜索引擎应用一个事先建好的网页/文件-概念索引 ( $C_{SE}$ ) 进行检索并且在用户选择搜索结果的网页或文件列表的时候显示网页或文件中的重要概念。
5. 如果选择了硬盘搜索，硬盘检索程序在事先建好的关键字/词-网页/文件索引 ( $A_{PC}$ ) 中查寻包含关键字/词串的文件。硬盘检索程序应用事先建好的一个网页/文件-概念索引 ( $C_{PC}$ ) 检索查询结果中包含的重要概念。硬盘检索程序然后对文件和概念进行排序，向运行在用户本机的界面程序返回查询结果的排序列表和前 N 个重要概念列表，用户界面程序然后在接口 400 中相应的部分进行搜索结果、概念和概念路径图的显示。如果没有选择硬盘搜索，跳过这一步。
6. 在显示概念列表的面板 412 上，当用户把鼠标浮动到概念上方或点击概念的“选中”或“排除”按钮时，或用户选择时间范围、来源、文件类型等过滤属性时，搜索引擎中的过滤程序根据用户选择条件过滤搜索结果并且将过滤结果显示在中间面板 408 上。为了根据用户在面板 412 中选择的概念对搜索结果进行过滤，搜索引擎使用了一个事先定义的概念-网页/文件索引 ( $B_{SE}$ ) 检索网页和文件列表找到含有那些选中概念的结果的交集。搜索引擎同时应用概念-网页/文件索引 ( $B_{SE}$ ) 构建网络搜索结果的概念路径图。
7. 如果选择了硬盘检索，一个本地过滤程序将过滤硬盘检索结果。如果硬盘搜索结果和网络搜索结果同时显示在面板 400 中的浏览器窗口中，将满足过滤条件的结果显示在右侧面板 410 上。如果选择了“Hard Drive Search in New Window”，网络搜索结果过滤和硬盘搜索结果过滤将分别执行和分别进行结果显示。为了根据用户在面板 412 选择的概念进行过滤操作，本地过滤程序应用事先建立的概念-网页/

---

文件索引 ( $B_{PC}$ ) 检索网页和文件列表找到含有那些选中概念的结果的交集。搜索引擎同时应用概念-网页/文件索引 ( $B_{PC}$ ) 构建硬盘搜索结果的概念路径图。

本发明的搜索引擎事先建立索引  $A_{SE}$ ,  $B_{SE}$ , 和  $C_{SE}$ , 也就是, 索引在用户使用搜索引擎进行搜索时就立即可以被使用了。本发明会定期更新这些索引使它们能够和网上的内容及时匹配。本发明的硬盘检索程序也会事先建立索引集  $A_{PC}$ ,  $B_{PC}$ , 和  $C_{PC}$ , 它们的格式和前面提到的类似。在一个实现中, 这些索引在硬盘程序第一次被安装的时候建立, 然后根据默认的时间段更新。为了使索引能够跟得上用户本地计算机文件的更新, 用户可以设置这个默认的时间段的大小。事先建立这些索引可以使本发明提供快速的查询功能。

上面的实现需要一个网络搜索引擎, 用户通过网络访问这个搜索引擎来进行网络搜索。在另外一个实现中, 用户可以自己选择一个网络搜索引擎使用, 如 Yahoo 和 Google, 而本发明中的概念抽取、过滤和图示在用户本地机器上实现。一种办法是使用网浏览器嵌入程序, 如一个 Microsoft 互联网 Explorer 嵌入程序, 把本发明的概念抽取、过滤和图示和搜索引擎结果绑定起来。图 5 展示了一种常规的搜索引擎界面和一个含有工具条的网浏览器接口, 可以用来嵌入本发明。用户点击“Enable DIGGOL”按钮 503, 如图 5 中高亮度显示部分, 来开启本发明的功能。当本发明的功能被开启后, 并且用户在输入框 509 中键入了搜索关键字/词, 点击“Search”按钮 509, 本发明的功能就被启动了。在一种实现中, 一个新的浏览窗口 600 被开启, 如图 6 中所示。如果“启动硬盘搜索”按钮 505 被点击, 这个新浏览窗口将在右侧包含一个面板 623 来显示本地搜索结果, 中间包含一个面板 621 显示网络搜索结果。在这个实现中, 网络搜索结果和本地搜索结果的概念抽取、过滤和图示操作都通过本发明安装在用户机器上的一个程序运行来实现。这个实现的运转过程如下:

1. 用户选择使用一个习惯的搜索引擎, 键入关键字/词串, 如一个类似于 Yahoo 或 Google 的搜索引擎, 然后通过这个搜索引擎进行检索。一个运行在用户本机上面的控制程序探测到这个搜索事件, 打开浏览窗口 600, 如果选择了硬盘检索, 同时把搜索关键字/词串发送给硬盘检索程序。
2. 用户选择的搜索引擎将搜索结果返回给用户本地机器上的搜索引擎接口。用户本机上的控制程序监测到这个事件, 并且初始化一个本地下载程序。下载程序下载搜索引擎返回的结果。它也从搜索引擎下载搜索结果中的网页和文件, 如应用网络服务协议, 或在搜索引擎返回的查询结果中抽取出所有的 URL, 然后根据它们各自的 URL 下载网页或文件。在一个实现中, 下载程序调用病毒扫描程序扫描下载的网页或文件。在一个实现中, 本地排序程序根据原始搜索引擎的排序和本地定义的一组排序规则对搜索结果进行重新排序。
3. 一个本地概念抽取程序在下载的网页和文件中抽取出重要概念, 建立一个概念-网页/文件索引 ( $B_{IP}$ ), 这个索引可以查询包含某个概念的所有网页或文件。在一个实现中, 本地概念抽取程序同时建立一个网页/文件-概念索引 ( $C_{IP}$ ), 这样当用户选择搜索结果中的某个网页或文件时, 用户界面程序可以通过索引  $C_{IP}$  检索并向用户显示这个网页或文件中包含的重要概念。一个排序程序综合应用原始搜索引擎排序和相关度排序对网页和文件重新排序。这个本地排序程序同时对每个文档中抽取出的重要概念进行排序, 然后对所有抽出的概念进行综合排序来取得 612 部分要显示的前 N 个重要概念。排序后的搜索结果和前 N 个概念被发送给运行在用户本地机器的用户界面程序, 这个程序将搜索结果、概念和概念路径图填充在用户界面 600 的面板中显示给用户。

4. 如果选择了硬盘搜索，硬盘检索程序根据用户输入关键字/词串，使用事先建立的关键字/词-网页/文件索引 ( $A_{PC}$ ) 查找包含这个关键字/词串的文件。硬盘检索程序应用一个事先建立的网页/文件-概念索引 ( $C_{PC}$ ) 获取搜索结果中包含的重要概念。硬盘检索程序然后对文件和概念进行排序，向运行在用户本地机器上的界面程序返回搜索结果的排序列表和前 N 个概念，用户界面程序将搜索结果、概念和概念路径图填充在用户界面 600 的面板中显示给用户。如果没有选择硬盘检索，跳过这一步。

5. 在显示概念列表的面板 612 上，当用户把鼠标浮动到概念上方或点击概念的“选中”或“排除”按钮时，或用户在面板 616 中选择时间范围、来源、文件类型等过滤属性时，一个本地过滤程序根据用户选择条件过滤搜索结果并且将过滤结果显示在中间面板 621 上。为了根据用户在面板 612 中选择的概念对搜索结果进行过滤，本地过滤程序使用了上面第 3 步建立的概念-网页/文件索引 ( $B_{IP}$ ) 检索网页和文件列表找到含有那些选中概念的结果的交集。本地过滤程序同时应用概念-网页/文件索引 ( $B_{IE}$ ) 构建网络搜索结果的概念路径图。

6. 如果选择了硬盘检索，一个本地过滤程序过滤硬盘检索结果，如果硬盘搜索结果和网络搜索结果同时显示在面板 600 中的浏览器窗口中，将满足过滤条件的结果显示在右侧面板 623 上。如果选择了“Hard Drive Search in New Window”，网络搜索结果过滤和硬盘搜索结果过滤将分别执行和分别进行结果显示。为了根据用户在面板 612 选择的概念进行过滤操作，本地过滤程序应用事先建立的概念-网页/文件索引 ( $B_{PC}$ ) 检索网页和文件列表找到含有那些选中概念的结果的交集。搜索引擎同时应用概念-网页/文件索引 ( $B_{PC}$ ) 构建硬盘搜索结果的概念路径图。

在一个实现中，网页和文件的下载数目 M 或可以下载的文件大小 K (兆字节 megabytes) 可以默认设置或被用户设置。M 和 K 是正整数，如 M=1,000，表示最初下载 1000 个网页和文件。或 K=100，表示下载的网页和文件的大小不会超过 100MB。当第一批下载的网页和文件集合抵达了 M 或 K 的限制后，下载程序暂时停止下载，并且保存一个 first 指针指向原始结果集中下一个要下载的网页或文件。当第一批下载集合完成了大部分下载以后，如下载了 900 个网页和文件，或 90MB，用户仍然没有停止原始搜索，关闭程序或开启了一个新搜索，则控制程序再次激活下载程序继续下载。下载程序将通过 first 指针从 1001 个网页或文件，或从下载程序没有到达 100MB 以前停止的下一个网页或文件开始下载。

另一个实现是上面两种实现的综合，在搜索引擎上完成概念抽取和索引集  $A_{SE}$ ， $B_{SE}$  以及  $C_{SE}$  的事先定义，但是概念过滤和概念路径图的生成在用户本机上完成。为了做到这一点，搜索时，搜索引擎缩减索引  $B_{SE}$ ，在一些情况下缩减索引  $C_{SE}$ ，使它们只包含搜索结果的网页和文件以及它们的概念。我们把这些索引分别称为  $B'_{SE}$ ，和  $C'_{SE}$ 。一个本地下载程序下载索引集  $B'_{SE}$  和  $C'_{SE}$  到本地客户机。然后，本地过滤程序和概念路径图生成程序可以应用下载的索引集进行概念过滤和生成概念路径图。下载事先建立好的索引集  $B_{SE}$  和  $C_{SE}$  节省了处理时间，这样概念过滤结果和概念路径图可以很快显示给用户。

另一方面，通过下载索引集  $B'_{SE}$  和  $C'_{SE}$  在用户本机上进行搜索结果的概念过滤和概念路径图的生成应用了数以百万计的 PC 机的广大运算信息源。

另一个在本地计算机和搜索服务器之间的任务切分灵活性体现在从 NLDS 中抽取搜索关键字/词串以及在 100 和 300 中将关键字/词扩展为概念。在一种实现中，它们在连接到互联网上的搜索引擎服务器上运行；在另外一种实现中它们在本地计算机上运行，该计算机生成概念化扩展的查询关键字/词串和查询合并，同时发送它们到互联网上的搜索引擎服务器。搜索引擎直接使用提交的搜索词来执行搜索。从

NLDS 中抽取搜索关键字/词串和扩展关键字/词的执行将使用数以百万计的 PC 上大量的可用计算资源。

在用户点击了“新窗口显示硬盘搜索结果”的情况下，硬盘的搜索结果在一个新窗口中显示，如图 7 所示。搜索结果和概念化过滤结果的方法将在第 3 节中说明。

## 2.2 CPM 图

早前的搜索引擎仅仅是把搜索结果显示成一个线性的列表。用户需要拖动卷轴，一页一页地翻看这个列表。聚类搜索引擎提供一个类别列表，如果一个类别有子类别的话，用户需要点击这个类别才能看到它有什么子类别。本发明向用户提供一种简单的图形可视化结果，该可视化结果显示了搜索结果按照其包含的重要概念进行的逻辑和/或统计分布。该可视化结果被称为概念路径图 (CPM) 或简称为概念图。如果一个用户通过点击 400 中 450 或 452, 600 中的 650 或 652, 700 中的 750 以显示概念图，一个概念图生成程序会基于 412 或 612 或 712 区中左边面板显示的概念各自生成一幅概念图，同时用户接口程序会在浏览器窗口 400 或 600 或 700 中分别显示这些概念图。具体实现中会向用户提供两种概念图选项，用户可以挑选其中一个显示：最流行概念图 (MPP) 和最新鲜概念图 (MOP)，其定义在后面说明。对 MPP 来说更具逻辑性的名称是最大交集路径，而 MOP 的名称为最小交集路径。在一个实现中，上面提到的概念或重要概念可能是从结果中抽取出来的查询词和短语。

下面我们用从 100 个查询结果中抽出的 10 个概念来图示说明 CPM。搜索结果可能是互联网或本地计算机或本地网络硬盘中的页或文件。我们称这 10 个概念为 A, B, C, D, E, F, G, H, I, J。其中 A 是查询关键字/词字串。注意在程序中这些概念的任何一个都可能是一个查询词或是查询词的集合或是一个短语。例如，如果用户查询关键字/词串（上涨的油价 rising cost of oil）那么 A=（上涨的油价 rising cost oil），注意“of”“的”没有被用作一个查询词因为它是排除词列表里面的词。而其它概念可能是：B=（OPEC），C=（伊拉克战争 Iraq war），…，I=（俄国 Russia），J=（优科斯 Yukos）。假设这 100 个文件中的概念统计为：A=100, B=70, C=55, D=50, E=41, F=38, G=30, I=10, J=2，这些数字表达有多少页或文件包含这些概念。例如，B=70 意味着有 70 个页或文件包含概念 B（或上面例子中的 OPEC）。

图 8 (a) 显示的 MPP CPM 图中，最流行概念或最大交集概念，即，被最多搜索结果包含的概念，最先被挑选出来，作为通向 CPM 图邻接节点的过渡路径。过渡路径上面的概念的功能就像是一个过滤器，只有包含了标识出的概念的搜索结果才会流动到下一个邻接的节点。在一个具体实现中，从右上角向左下角依次显示最流行概念到最不流行概念。在上面的例子中，在查询词字串 A 之后的第一层中，B 是最流行概念并且作为右上角第一个 1 层过渡路径，称作 1 层路径 B，它指向一个包含了 70 个概念的节点。第一层其它的过渡路径包含有 30 个页或文件，表示为 nB (nB=不包含 B) 路径。假设除了 A，概念 E 是 nB 中最流行的概念且 E=20。这样 E 被用作一层路径 B 下面的第二条 1 层过渡路径，指向一个 20 个搜索结果的节点。在子集 nBnE 中有 10 个概念，假设概念 G 是除了 A 之外的最流行概念，且 G=6，这样 G 是 1 层路径 E 左下方的第三条 1 层过渡路径，指向一个 6 个搜索结果的节点。在子集 nBnEnG 中有 4 个概念，假设两个概念 C 和 I 是除了 A 之外的最流行概念，它们两个有相同的概念数，且 C=2, I=2。这样 C 和 I 被作为一层路径 G 左下方的第四条和第五条 1 层过渡概念路径，分别指向 2 个搜索结果的节点。当两个过渡路径有相同的流行度时，可以按照过渡路径的概念的权值将权值高的排在右上方，同时也可以按照概念的字母顺序排列。在 MPP CPM 的第二层中，相对 B 的子集有 70 个概念，假设概念 C

是除了 A 和 B 之外的最流行概念, C=33。于是 C 被作为一层路径 B 后面右上方的第一个 2 层过渡路径, 指向一个含有 33 个搜索结果的节点。在 BnC 相关子集中有 37 个结果, 假设概念 E 是除了 A 和 B 之外的最流行概念, E=16。那么 E 被用作 B 子集中 2 层路径 C 下面的第 2 条二层过渡路径, 指向一个 16 个搜索结果的节点。在 BnCnE 的子集中有 22 个概念, 假设概念 F 是除了 A 和 B 之外最流行的概念, F=14。于是 F 被用作 B 子集中 2 层路径 E 左面的第 3 条二层过渡路径, 指向一个 14 个搜索结果的节点。概念图将继续扩展下去, 直到一个节点内所有网页和文件所包含的被列出的概念都已经在指向该节点的过渡路径中被使用了, 或在一个节点中只有一个搜索结果了。一个概念路径就是一个过渡路径的序列, 搜索结果按照和过渡路径相关的概念的排列顺序的被过滤, 例如: 图 8(a) 中的概念路径 ABC、ABG、AECD。事实上, ABG 是 AB (nC) G, AECD 是 A (nB) ECD。注意概念在一条路径里面的顺序是十分重要的, 因为搜索结果将根据概念在路径中的顺序被过滤。

在图 8(b) 中显示的一个 MOP CPM 中, 最新鲜概念或最小交集概念, 即被最少搜索结果包含的概念, 是 CPM 中第一个作为通向邻接节点的过渡路径被挑选出来的。事实上, 一个概念被最少的搜索结果包含可能意味着它是一个非常新的, 独特的观点、看法或发现等等。因此它可能非常新颖和有信息量。一个 MOP CPM 图的目标就是在大量的混乱的搜索结果中间挖掘出来这样的网页和文件, 并且清晰明显的显示给用户。在一个 MOP CPM 中, 非常少的过渡途径就可引出包含最不流行概念的网页或文件, 并且可以被显示在显著的位置。类似 MPP, 过渡路径中的概念的功能就像是一个过滤器, 只有包含了过渡路径中标示了的概念的搜索结果才能流到相邻节点。在一个实现中, 按照最稀有或最不流行到最普通或最流行的顺序, 概念从右上方向左下方排列。在上面的例子中, 位于右上方的 J 是最不流行概念, 被作为第一条 1 层过渡路径, 指向含有 2 个结果的节点。第一层剩余的过渡路径表示为 nJ 路径, 包含了 98 个网页和文件。假设概念 I 是 nJ 子集中最不流行的概念, I=9。那么 I 被作为一层路径 J 下面的第二条 1 层过渡路径, 指向一个含有 9 个搜索结果的节点。在子集 nJnI 中有 89 个概念, 假设概念 E 是最不流行概念, E=21。那么 I 被作为一层路径 I 下面的第三条 1 层过渡路径, 指向一个含有 21 个搜索结果的节点。在子集 nJnInE 中有 68 个概念, 假设概念 G 是最不流行概念, G=29。那么 G 被作为第一层路径 E 下面的第四条 1 层过渡路径, 指向一个含有 29 个搜索结果的节点。在子集 nJnInEnG 中有 39 个概念, 假设概念 C 是最不流行概念, E=39。那么 C 被作为一层路径 G 下面的第五条 1 层过渡路径, 指向一个含有 39 个搜索结果的节点。在 MOP CPM 中的第二层中, 假设概念 I 和概念 G 是最不流行概念, I=1 且 G=1。那么 I 和 G 被作为一层路径 J 之后的右上方的第一条和第二条 2 层过渡路径, 各自指向一个包含 1 个搜索结果的节点。当两个过渡路径都最不流行时, 可以按照过渡路径的概念的权值把权值高的排在右上方, 同时也可以按照概念的字母顺序排列。MOP CPM 可以继续扩展下去, 直到一个节点内再没有列出的概念, 或一个节点内只有一个概念。

一般来说, 鉴于受到屏幕大小的限制, 一个概念图只能显示第一层和第二层的过渡路径和节点。其它过渡路径和节点都被收拢起来。被收拢的部分用一个“+”符号它和剩余概念的列表。点击这个“+”号会扩展 CPM 图一层或多层。剩余概念的列表可以是仅显示概念第一个单词的局部列表。当鼠标移动到或点击概念的显示出来的部分时, 一个悬浮窗口将出现并显示完整的概念。用户可以通过点击“+”或“-”符号来展开或收拢 CPM。

在一个实现中, CPM 也可以否定方式显示路径和节点, 例如, 使用上面例子中的 MPP, 第一层的

否定过渡路径是“No B”，它意味着所有不包含 B 的搜索结果可以通过这个节点到相邻节点去。在上面例子的 MPP 的第一层中的一个否定模式中，nB 节点包含了所有不包含概念 B 的搜索结果。如上面图 8(c) 中的 MPP 图，图中示意了否定路径和否定节点的 MPP 图。如图 8(a) 和 (b) 中所示，每一个过渡路径都由一个概念标示。每一个指向第一个节点的过渡路径都像是一个真空管。它将所有包含了指向上述第一个节点的过渡路径中标示出的概念的所有网页和文件吸收进上述的第一个节点。而剩余的网页和文件则继续向下流动。图 8 中 CPM 的各种变化和其它替代图形表示也都可以被用作表示 CPM。

当用户在搜索结果面板中选择了“概念图”，则同时 412 或 612 或 712 或 912 区中左边面板中的一个或更多的概念就被选中。左边面板中被选中的节点将变成高亮度显示或不同颜色和形状显示。这样可以使得用户能通过点击高亮或特殊颜色和形状的节点，快速定位节点或聚类或网页和文件。如图 9 所示，在左边面板 912 区中搜索词（上涨油价 Rising Cost Oil）和两个概念（OPEC）（伊拉克战争 Iraq war）被选中，则 CPM 中的 939 节点变成不同的形状，因为它含有全部被选中的概念。注意图 9 中硬盘搜索没有启用，所以没有显示硬盘搜索的结果。CPM 中的一个节点变成了高亮度或不同的颜色和形状，所以一个概念图产生程序用  $B_{SE}$  或  $B_{IP}$  或  $B_{PC}$  索引来将用户选择的概念和包含这些概念的网页和文件对应起来。对应一个网页可能是一个指向网页的摘要或网页的 URL 的指示器。对应一个文件可能包含一个指向文件的摘要或文件的完整路径的指示器。通过使用  $B_{SE}$  or  $B_{IP}$ , or  $B_{PC}$  索引抽出和各个选中概念相关的网页和文件的集合，概念图生成程序找出所有上述被选中概念相关集合的交集。然后使用上述的交集，它找到含有这个交集的 CPM 节点并且将其高亮化。当用户点击 CPM 中的一个节点的时候，所有属于这个节点的网页和文件的摘要和 URL 都在搜索结果面板中显示。为了实现这一功能，概念图生成程序生成一个索引或列表，该列表可以列出 CPM 中的各个节点中的概念。这一工作将由概念图生成程序在构建概念图的同时完成。

无论 MPP CPM 还是 MOP CPM 都提供搜索结果的统计和逻辑的分布或组织的清晰可视的整体视图。这个功能对之前的搜索引擎技术和界面都是难于完成的。一个用户可以很快地看出通过沿着一条概念路径进行概念过滤的效果，或通过选择左边面板中的高亮节点中的概念进行过滤的效果。一个 MPP 概念图中的概念路径是同一层中最流行概念的搜索结果的连续聚类。流行度可以看作是大众认为什么东西重要的一种投票。这样，一个概念被大量的网页提到说明这些网页的作者认为这个概念是重要的有价值的。在 MPP CPM 图中，各层含有最流行概念的网页和文件在显著的位置向用户显示。在 MPP CPM 中就是同层中最新颖的概念的搜索结果的连续聚类。一个 MOP CPM 图的目标就是挖掘出没有经过广泛认同的新颖的处于发展初期的有潜在价值的观点。

CPM 中的过渡路径可以是基于上面的 MPP 和 MOP 之外的其它关系。在一个实现中，过渡路径是基于两个节点，既两个子集间的逻辑和语义关系。如果两个节点中含有的网页和文件子集所包含的内容和有一个逻辑或语义关系的匹配，那么本发明就可以在两个节点间建立一条过渡路径，这条过渡途径的含义就是这个逻辑或语义关系。在一个实现中，上述的逻辑和语义关系是前提或必要条件关系，如果概念 A 中的网页和文件包含了概念 B 中一些内容的前提或具备条件，那么 AB 间设置一条过渡路径，该过渡路径被称为前提或具备条件过渡路径。

本发明的 CPM, MPP 和 MOP 的实现提供了一种把文件组织成一个结构或显示此结构的新方法，该方法包括：

把两个或更多个文件组织成在一个维度（称为甲维度，如竖轴）上相连接的两个或更多个集，其中一个集的成员是基于和文件相关的信息元或文件所含的信息元决定的，两个集之间的连结意味着在这两个集之间存在一个关系（称为甲关系）；

把两个或更多个文件组织成在另一个维度（称为乙维度，如水平轴）上相连接的两个或更多个集，其中一个集的成员是基于和文件相关的信息元或文件所含的信息元决定的，两个集之间的连结意味着在这两个集之间存在一个关系（称为乙关系）。

上述的方法还可进一步包括下列一项或多项：甲关系和乙关系之一或两者是子集关系，意味着在一个连结一端的集是在连结另一端的集的子集；甲关系和乙关系之一或两者是一个在一个连结两端的集之间的一个逻辑或语义关系；在甲维度和乙维度之一或两者上有三个或更多的集连结在一起，且甲关系和乙关系之一或两者是可传递关系；将文件组织成的结构以图论图或图像的方式显示。

### 2.3 概念显示，概念过滤和概念图的索引结构

在前面描述了3种索引：查询词到网页或文件的索引  $A_{SE}$  和  $A_{PC}$ ，概念到网页或文件的索引  $B_{SE}$ ,  $B_{IP}$ , 和  $B_{PC}$ ，网页或文件到概念的索引  $C_{SE}$ ,  $C_{IP}$ , 和  $C_{PC}$ 。在一个实现中，3个索引的格式为：

$A_{SE}$  和  $A_{PC}$ : {[查询词\_1, (网页\_1, 文件\_2, ..., 网页数量/文件)], [查询词\_2, (文件\_i, 网页\_j, ..., 文件数量)], ...}

$B_{SE}$ ,  $B_{IP}$ , 和  $B_{PC}$ : {[概念\_1, (文件\_1, 网页\_2, ..., 网页的数量/文件)], [概念\_2, (文件\_i, 网页\_j, ..., 网页的数量/文件)], ...}

$C_{SE}$ ,  $C_{IP}$ , 和  $C_{PC}$ : {[网页\_1, (概念\_1, 概念\_2, ..., 抽取出的重要概念的数量)], [文件\_i, (概念\_j, 概念\_k, ..., 抽取出的重要概念的数量)], ...}

在上面，对于一个网页搜索结果，网页\_i 或文件\_j 可以包含网页或文件标题以及 URL，还有对下载和保存在本地硬盘上的网页或文件的版本的指示器。对于用户本地电脑上的文件，文件\_j 可以包含文件的名字和路径。

索引  $A_{SE}$ 、 $A_{PC}$  和  $B_{SE}$ 、 $B_{IP}$ 、 $B_{PC}$  之间的不同在于， $A_{SE}$  和  $A_{PC}$  必须包含除了 SWEEL 之外的用户搜索网页和文件的所有关键字/词，而  $B_{SE}$ 、 $B_{IP}$  和  $B_{PC}$  仅仅包含概念，例如被认为重要并作为重要词抽取出来的词，词组，短语等。 $A_{SE}$  和  $A_{PC}$  中的条目是一个单词或一个常用的短语，而一个  $B_{SE}$ 、 $B_{IP}$  和  $B_{PC}$  中的条目可以是从网页或文件中抽取出的单词串，甚至是一个简单的短语。

在搜索引擎中，为网页搜索准备  $A_{SE}$ ,  $B_{SE}$  和  $C_{SE}$  的功能模块可以提前进行。所有的3个索引都在搜索引擎中维护，如图10所示。图10中椭圆形框内是用户输入和系统的输出。图10中方框是程序的操作。图10中圆柱框内是文件或数据库。相同的模块图也可以应用在本地硬盘的文件搜索中的  $A_{PC}$ ,  $B_{PC}$  和  $C_{PC}$  上，其中所有3个索引也都在本地电脑构建和维护。在另一个混合上面两种实现的方法中，功能模块图类似上面图10，除了它们的维护和使用地点可能变化，例如在搜索引擎服务器上或在用户的PC上，或同时在两者上都有。为了支持快速抽取和快速更新，本发明可用包括哈希表，倒排序，B+树，grid文件，多路B-树在内的合适的数据结构构建索引。

### 2.4 特征过滤

在一个实现中，如416和616区中列出的类似文件类型，修改日期，来源等被提供给用户以过滤搜索结果。一个特征过滤程序抽取特定的源，文件类型和日期范围等等，同时统计这些搜索结果。在一个

实现中，当用户在 104 或 302 中的搜索引擎界面中选择超过一个搜索目标时，区 416 和 616 同样包括可以通过用户选择的（如 400 和 600 所示）搜索目标将研究结果分类的一个域。当用户点击 416 区中列出的搜索目标时，只有匹配了被选中的搜索对象的搜索结果才会显示在 408 的搜索结果面板中。416 和 616 中的特征域可以由用户通过点击+ 或- 符号来展开或收拢。一旦一个新的特征域被选中，之前选中的展开的域会收拢，同时新选中的特征域展开。这使得有限的空间内可以安排多个区。

416 和 616 的来源域中，已知的顶级域名扩展名如，.gov, .edu, .tv, .info 等，国家域名扩展.cn, .us, .ca 等，还有两级域名.edu.cn, .gov.cn, .gov.uk, .ac.uk 都被包含在内。程序中的源聚类模块将计算来自一个网站或域名的网页和文件的数量，例如，cnn.com, ieee.org, irs.gov, ucla.edu 等。在一个实现中，源聚类程序将选出前 S 个包含了最多的网页和文件的网站或域名，其中 S 是由用户指定的或系统默认的一个正整数。这 S 个网站或域名将在来源区域 416 或 616 中列出。这使得用户可以通过选择或排除一个或多个网址或域名来过滤搜索结果。

为 416, 616 或 716 中的各个过滤特性建立了特征到网页/文件索引 (FTFI)，这个索引类似概念到网页/文件的索引  $B_{SE}$ ,  $B_{IP}$  或  $B_{PC}$ 。FTFI 的格式如下

{[过滤特征\_1, (文件\_1, 网页\_2, ..., 网页/文件数目)], [过滤特征\_2, (文件\_i, 网页\_j, ..., 网页/文件数目)], ...}

这样的一个索引可以用在支持选择或排除特性过滤中。当一个过滤特征被选中，特征的 FTFI 可以被用作根据选中的特征过滤网页和文件列表，并且过滤结果将被显示或进一步被其它特征及概念交集过滤。当过滤特征被排除，特征的 FTFI 可被用于根据排除特性抽取网页和文件列表，被抽取出来的网页和文件将从搜索结果中删除。另外概念到网页/文件索引  $B_{SE}$ ,  $B_{IP}$  or  $B_{PC}$  也可扩展包含其它特征。一种扩展格式如下：

{[概念\_1, (文件\_1, 网页\_2, ..., 网页/文件数量)], [概念\_2, (文件\_i, 网页\_j, ..., 网页/文件数量)], ..., [过滤特征\_1, (文件\_k, 网页\_m, ..., 网页/文件数量)], [过滤特征\_2, (文件\_p, 网页\_q, ..., 网页/文件数量)], ...}

网页/文件到概念索引  $C_{SE}$ ,  $C_{IP}$  和  $C_{PC}$  可扩展包含其它过滤特征。一种扩展格式如下：

{[网页\_1, (概念\_1, 概念\_2, 过滤特征\_1, 过滤特征\_2, ..., 抽取出的重要概念数量)], [文件\_i, (概念\_j, 概念\_k, 过滤特征\_1, 过滤特征\_k, ..., 抽取出的重要概念)], ...}

### 3. 在搜索结果或文件中抽取概念并排序

#### 3.1 抽取重要概念

在一个实现中，重要概念是能够表示一个网页或文件特征的名词，短语和首字母缩写词。这将使得一个网页或文件及大量的搜索结果压缩成重要概念列表。

详细的自然语言处理和理解将使概念抽取更加精确。然而，关键问题是如何快速地处理大量的网页和文件。本程序的实现将抽取下面的词组和短语作为重要概念：(1) 在文章的特别位置和特别段落，例如，标题和章节标题；(2) 有特别的统计特性或特征，例如词频最高和最低的词（不包括排除词列表中的普通词），2 或 3 个词的短语，开头大写和全部大写的词，特别要给与大写字母开头或全部是大写字母的两个以上的连续单词以重视，还有高亮度，黑体，斜体，下划线或不同字体颜色修饰的单词，(3) 和查询关键字/词在同一句的单词，和重要词列表 (IW/P) 中的词或词根在同一句中的单词，和 IW/P 列表

在同一个句式集合里面的单词。

每种语言都有一些特定的词和句式用在强调句中。识别出这样的词和句式有助于识别出含有文章重要论点，结论，观点，问题或总结的句子。这样，从这些句子中可抽取出重要的概念。在一个实现中，以英语和中文为例，重要词列表由 3 组词组成。注意每个词都可以扩展成它的不同形态，例如名词，动词，现在时，过去时和将来时态，形容词，副词。由于空间有限，下面仅给出每组的部分子集内容。

IW/P 列表 第一组：根据本组的字/词或短语抽出的概念排序的优先为中等。英文：(better, more, worse, require, outcome, result, important, significant, interesting, true, depend, independent, surprising, oversight, overlook, mistake, investigate, research, study, explore, look into, concept, intriguing, worthwhile, worth, special, specialized, need to, consider, evaluate, improve, enhance, advance, necessary, sufficient, insufficient, standard, new, innovative, overcome, efficient, inefficient, backward, old, outstanding, new, alternative, all -er adjectives or adverbs, etc.); 中文：(较好、更多、重要、依赖于、标准、充分、杰出的、特殊、调查、重大、研究、必要、探索、错误、概念、忽视、考虑、创新、提高、改进、真实、需要、等等)

IW/P 列表 组 2：根据本组的字/词或短语抽出的概念排序的优先为高等。英文：(best, most, worst, referred to as, is/are/was/were called, abbreviated as, critical, crucial, vital, purpose, objective, goal, key, main, major, overwhelming, striking, remarkable, extreme, exceeding, disaster, necessary and sufficient, iff, fundamental, all -est adjectives or adverbs, etc.); 中文：(最好、最坏、最差、称为、关键、目的、主要、必要和充分、等等)

IW/P 列表 组 3：根据本组的字/词或短语抽出的概念排序的优先为最高等。英文：(key idea, main idea, major idea, main purpose, main objective, main goal, main problem, major problem, main difficulty, main obstacle, break through, breakthrough, major development, major innovation, invention, discover, groundbreaking, break new ground, new record, world record, record high, record low, unparalleled, unprecedented, revolutionary, unexpected, never, etc.); 中文：(主要思想、主要目的、关键问题、主要困难、突破、重大发展、发明、开辟新领域、新纪录、空前、革命性、决不、等等)

排除词列表 (ICEEL) 中的普通词可从抽取出的重要概念中排除。注意 ICEEL 可以用作 SWEEL。ICEEL 的一部分内容的例子显示如下：英文：(单个字母或少于 3 位的数字； about after all am among an another any anybody anything anytime are as at be been but by can could did do each everybody find first firstly five for four from had has have he her him his how if in into is it its just little made make many may more much my no not of on one only or other out over people said second secondly see seven shall she should so some somebody something sometimes ten that the their them themselves then there these they thing third thirdly this those three to two up use very via was way we were what when where which who whom will with words would you your, etc.); 中文：(的、关于、以后、是、可以、其它、而且、然而、等等、所以、那里、这里、首先、有关、.....)。

### 3.1.1 使用 IW/P 列表抽取重要概念

在一个实现中，使用 IW/P 列表抽取重要概念可以通过识别包含 IW/P 列表中一个或多个单词的句子来实现。在跨任何标点和定义从句（对于英文，即以 that, those, who, whom, which 开头的子句）的

部分都将切断，删除所有 ICEEL 中的单词，然后将所有的剩余单词作为概念。对于该实现的更细致的说明如下：

1. 从包含至少一个 IW/P 中的单词或短语的句子（不超过句号（。）或分号（;）或引号（“ ”或‘ ’）或冒号（：），但可以跨逗号（，））中抽取出除了排除词列表中的单词之外的所有单词。如果抽出的单词串长度小于 5 那么停止，否则转到第 2 步。
2. 删除上面句子中跨逗号的单词。如果抽出的单词串长度小于 5，停止，否则，到第 3 步。
3. 进一步删除上面句子中跨定语从句或形容词性动词词组的单词。如果抽出的单词串长度小于 5，停止，否则，到第 4 步。
4. 英文：进一步删除上面句子中跨介词（in, on, with, from 等等，但不包括 “of” 和 “to”）的单词。如果抽出的单词串长度小于 5，停止，否则，到第 5 步。中文：进一步删除上面句子中跨联结词、助词的字 / 词。
5. 英文：进一步删除上面句子中跨“of”和“to”的单词。如果抽出的单词串中至少有一个 IW/P 列表之外的单词，停止，否则，用第 4 步抽出的单词串。

保持抽出的单词串和其在原文中出现的顺序完全一样是很重要的。在另外一种实现中，句式和 IW/P 列表中的词被联合用来抽取包含了一个或多个 IW/P 中的单词的句子中的最重要的单词串。同样不要跨越任何标点和从句。这一要求在通过句式，IW/P 或搜索词辨识出的句子中，利用了很多已知的句式，例如：“此研究的目的是.....”，“the goal of this study is to ....”，“结论是.....”，“the conclusion is ....”，等等，并且利用词性来分析识别主语，谓语，宾语，定语从句。利用词类分析识别名词，动词，不定式等。其它可以抽出概念的句式还有“The (形容词) 目的是 ...”，“(名词短语) 提供了 (名词短语)”，“(名词短语) 产生了 (名词短语)”，“(名词短语) 使得 名词短语 )”，以及主语或宾语是大写字母开头的短语的句子等等。

### 3.1.2 重要概念分组

重要概念可能出现在文本的不同部分，并且具有不同的特性和重要性。本发明的一个具体化实现就是将抽取出来的重要概念进行了分组。每组有自己的抽取和排序规则。将从组 A 到 F 抽出的概念作为候选重要概念。重要概念根据预先分配的百分比从六组中选出。从每组选出的重要概念都有不同的排序权值，A 组具有最高的排序权值。

A. (40%) 抽取词在文章的标题和副标题中。一个具有五个或更少的词组成的标题将被作为一个独立的概念抽出。例如，本部分的标题“Grouping of Important Concepts”将被作为一个重要概念抽出。多于五个词组成的标题将根据介词，连接词以及标点截断（如 in, for, with, by, at, on, and, or, 逗号，分号等）成许多部分。例如，标题“Indexing Structure for Concept Display, Conceptual Filtering and Concept Path Maps”将被截为四部分 (Indexing Structure), (Concept Display), (Conceptual Filtering), (Concept Path Maps)。ICEEL 中的词将从每个部分中移除。仅由一个词组成的第一部分将试着与紧随其后的部分组合，如果组合成的词不大于五个，则将其作为一个独立的概念抽出。如果组合成的词大于 5 个，则这两个部分不能组合，并且第一个部分重新试着和下一个部分组合。如果组合得到的词长度不大于 5，将其作为一个独立的概念抽出。如果组合得到的词长度大于 5，这两个部分不能合并。每一个剩余部分将作为一个重要概念抽出。一种具体实现的方法是，抽出的概念将根据概念在文本中出现的次数分配权值，多次

出现或很少出现都赋予较高的权值，由两个到三个词组成概念的权值高于由一个或超过三个词组成的概念，分配权值同时考虑抽出的概念是否包涵关键字/词。出现次数是多还是少可以根据平均出现值或预先设定的值决定。在使用 HTML 或 XML 的结构化文本中，将根据标签确定文章标题或章节标题。在没有标签或非结构化的文本中，文章标题或章节标题根据它是否在一个独立的行，以及是否是紧随一个冒号之后的短语或短行决定。标题中的某些词，如摘要，介绍，背景，讨论，描述，结论，概要等，由于不能传递有关文章内容的重要信息，因此将被排除。

B. (共 12%，其中每组分配 4%) 抽取：(a) 两到四个词的短语中至少包含了关键字/词中的两个词，关键字/词的每一种不同的排列将形成不同的概念；(b) 紧邻一个或多个关键字/词的两到三个词形成的短语；(c) 非关键字/词组成的两到三个词的短语，没有紧邻关键字/词，但它是在有一个或多个关键字/词的句子中。一种具体实现的方法是，抽出的概念根据以下规则排序：抽取自每个子组的概念都被赋予一个介于 0 到 1 之间的子组权值，而子组 (a) 具有最高权值 1，一个抽出的概念将根据关键字/词在短语中或句子中出现的次数，名词的个数以及短语的长度来排序。每个属于本组的排序权值都将被标准化并介于 0 到 10 之间。抽出概念的最终排序权值将由子组的排序权值和本组的排序权值共同决定。

C. (12%) 如果单词或其同义词集在 IW/P 表中或一种指定的句型中，使用上述方法从同一个句子中抽取单词。一种具体实现的方法是，抽出的概念按照如下规则排序：抽出的概念按照介于 0 到 1 的组内权重排序 (IW/P 表中的第三组具有最高的排序权值 1，第二组权值为 0.6，第一组权值为 0.3)。组内权重被标准化为介于 0 到 10 之间，并根据概念在网页或文件中的出现次数决定大小，多次出现或很少出现都将被赋予较高的权值，因此抽取出的概念同时考虑了普遍概念以及特殊概念。具体实现时根据概念出现次数和平均出现值或预先设定值的偏离程度决定。抽出的概念将由子组的排序权值和本组的排序权值共同决定。

D. (共 12%，其中每组分配 4%) 抽取 (a)、首字母大写的两个或更多的词组成的短语，该短语不能被标点分割；(b)、所有字母均大写的一个单词包括缩写词；(c)、除句首大写的由两到三个首字母大写的单词组成的短语，在两个紧连的词中应至少包含一个名词。一种具体实现的方法是，抽出的概念按照如下规则排序：从每个子组抽出的概念都被赋予一个介于 0 到 1 之间的子组权值，(a) 组具有最高的权值 1。组内的排序根据概念在网页或文件中的出现次数决定。一种具体实现的方法是，多次出现或很少出现都将被赋予较高的权值，因此抽取出的概念同时考虑了普遍概念以及特殊概念。具体实现时根据概念出现次数和平均出现值或预先设定值的偏离程度决定。抽出的概念将由子组的排序权值和本组的排序权值共同决定。

E. (12%) 高亮显示，斜体，有下划线或以不同颜色或字体显示的短语将被抽出。如这些词不是名词，则同时抽取出紧随这些词之后的名词或邻近这些词的名词。一种实现方法是，抽出的概念将根据高亮显示，斜体，粗体，有下划线，不同颜色或字体显示这些突出特征出现的次数排序。如果一个网页中超过 10% 的词都是高亮显示，斜体，粗体，有下划线，以不同颜色或字体显示，则该组特征被忽略。

F. (多次出现的关键字/词分配 7%，较少出现的关键字/词分配 5%，抽出的概念应满足两种情况之一) 抽取出现次数最多和出现次数最少的由一个名词或两到三个不是常用词形成的词组，并且抽出的词不能是关键字/词或和关键字/词有相同的含义。如果一个网页或文件中某个名词或短语的出现次数超过 10%，或出现次数最少的词和短语是 ICEEL 中的词，或其中不包含名词，则不抽取这些词。对于出现次

数较多的词或短语，只要出现次数小于 10%，则排序权值随出现次数的增多而上升。对于出现次数较少的词或短语，排序权值随出现次数的减少而上升。

对于上述六组，ICEEL 中的常见词以及被标点分割的短语都不会被抽出。一种具体实现的方法是，一组中权值相等的概念既可以随机选出也可以按照字母顺序选出，以减少处理。每组中标后的百分比代表了从每组中抽取概念数的最高百分比的例子，如果从所有搜索结果包含的网页和文件中抽取的概念总数超过了用户想要列出的概念数，概念总数根据要在 412, 612, 712 或 912 中显示的概念数决定。一种具体实现的情况是，如果一个用户选择列出 N 个概念，从一个网页或文件中抽出的 N 个概念将和搜索结果中其他页或文件中分别抽出的 N 个概念汇总。重复或重叠的概念将被移除。如果一个重要概念已经在一个高权值的组中出现，那么它将从其他较低权值的组中移除。如果两个概念重叠的话，即它们包含相同的词或部分组成它们的词同义，其中的一个概念将被移除。将要移除哪一个概念根据这个概念是否是较高权值组中的概念，是否是由多个词组成的具体概念，是否是一个由较少词组成的普通概念决定。因此，从搜索结果的所有网页和文件中得到的概念将被一起排序，以得到能够显示给用户的前 N 个概念。

如果某组中没有足够的概念来达到分配的百分比，不足部分的百分比将按照比例分配到剩余的组中。一种具体实现的方法是，每类至少保证抽取出一个概念。例如，如果一个用户选择显示 10 个概念，而从 A 到 F 组中共抽出 100 个概念。尽管仅应从 F 组抽取一个概念（10 的 10%），但此时应保留该组出现次数最多的一个概念和出现次数最少的一个概念。在这种情况下，如果 E 组被分配抽到的概念多于一个，F 组将借用 E 组的分配比例。否则，再向上借用。如果  $N < 6$ ，某些组，例如 B, D, E 组的概念抽取都将被忽略。

B 组进行概念抽取是在关键字/词已知的情况下。假定关键字/词是（无线网络 wireless networks），则 B 组 (a) 例子包含（无线局域网络 wireless local area networking），（无线网络接入点 wireless network access point），B 组 (b) 例子包含（无线联结 wireless connectivity），（蜂窝无线 cellular wireless），（网络安全 network security）。很容易看出这些概念更有利得到精确的查询结果。然而，B 组的概念只能在搜索时刻被抽出而不能预先处理，因为当时关键字/词未知。为了减少搜索时的处理时间，每一个网页和文件中的重要概念将被预先抽出。一种具体实现的方法是，A, C, D, E 和 F 中的概念都被预先抽出，而只有 B 中的概念在搜索时抽出。然而另一种具体实现的方法是，B 组的概念没有被使用，分配给 B 组的抽取比例被分给其他组。如给 C, D, E 和 F 组各分配 3%。这样做避免了在搜索时从结果中抽取概念。同样，A 组中的概念权值可以独立于关键字/词预先设定，这样也可以节省搜索时的处理时间。

### 3.2 在本机进行网页结果的概念抽取

就像上述提到的一种具体实现的方法，抽取概念，排序，用户选择概念进行过滤以及 CPM 图都是在搜索引擎端实现。另一种实现的方式是在用户本机做，还有一种实现方式是部分在搜索引擎端，部分在用户本机实现的方式。当在用户本机实现时，需要一个本地下载程序来下载从搜索引擎返回的网页以及文件。这样，用户可以在本机分析下载的网页以及文件来执行概念抽取以及概念的排序操作。由于下载和概念的抽取和排序都需要一定的处理时间，为了在尽可能短的时间给用户一定的结果反馈，一种具体实现的方法是渐进地执行这些任务，也就是在分析部分结果得到的概念以及过滤特征时就显示给用户，同时下载程序继续下载搜索引擎返回的结果网页和文件，并在新的一批结果分析处理完成时定期更新概念列表和相关度排序值。例如，当搜索引擎返回的结果不大于 50 个时，将这些结果页以及文件下载

到本机上，并对这些页以及文件进行概念抽取，排序和过滤特征的提取，将分析结果显示给用户，同时本机上继续进行网页以及文件的下载和分析操作。一种具体实现的方法是，等待时间根据下载和分析最初 50 个结果页的时间进行相应调整。当到达设置的时间点，如 5 秒时，程序应显示给用户相应时间段的部分分析结果。同样，为了避免长时间等待，在进行第一批和第二批的下载时，对于大的网页以及文件（如超过 100KB）不予下载，它们将被安排在后面批次中下载，这样可以快速提供给用户可浏览的分析结果。还要补充的一点是，在对结果进行分析处理来得到概念，文件类型以及其他过滤特征时，为了节约下载时间，网页或文件中的图不下载。然而，图中的文字注释以及其他文字信息都和网页的其他文字部分一样被下载和分析。一种具体实现的方法是，开始时不下载的大于 100k 的网页和文件在已经下载了 M 个网页或文件后开始下载时，对于之后遇到的大型网页和文件也这样处理。

一种具体实现的方法是，当用户选择使用搜索引擎 500 时，点击了按钮 503 “启动 DIGGOL”来启动本发明（当本发明已被默认启动时，这个步骤不是必要的），当用户将搜索字串输入到 507 并点击按钮 509“搜索”时，程序开始进行下载，概念抽取和排序，同时在 5 秒内将部分结果的概念和过滤特征在 612 和 616 中显示给用户。当程序下载了足够多的搜索结果时，从结果中抽取概念，并将新的概念加入到概念池中。重复的概念和子集概念都将被移除，概念池中剩下的将被重新排序，这样概念列表根据概念池中最新的概念及其排序值被更新。

为了用将搜索引擎从用户很少察看的网页和文件中抽取概念，一种具体实现的方法是，本发明中下载和分析网页或文件是从每批结果的两端进行处理，也就是说对于第一批要处理的 50 个结果，下载、概念抽取以及其他过滤特征的提取按照如下顺序进行：1, 50, 2, 49, 3, 48, .....等。在随后的下载中，即使下载的结果数不是 50，也是按照同样的方法处理。这种方法称为“两头烧蜡烛”。该方法既考虑到排序在前的结果的普遍性，同时考虑到排序在后的结果的新颖以及独创性，并且排在后面的结果也可能包含有用信息。本发明中的排序方法将在后面介绍，该方法也遵循上述原则，同时抽取普遍性以及新颖性概念并给予较高的权值。这种“两头烧蜡烛”的处理方法以及排序方法使得搜索结果中排序靠后的网页包含的权值较高的概念能和排序在前的结果中分析得到的概念一起及时显示给用户。以前的搜索引擎不能实现这个功能。

为提示用户程序正在运行中，一种具体实现的方法是，在浏览器窗口的底端显示一个进度栏。这个进度栏显示了所有搜索结果中共有多少个结果被分析过，显示形式如“总共 223, 588 页，1, 250 页已经分析完成”。

为了更进一步缩短概念抽取和排序以及过滤特征提取的处理时间，一种具体实现的方法是，如果网页或文件过大（字数大于 5000），则第一轮仅处理摘要，讨论，结论，概要，文章的开头和结尾，每段的开头一到两句以及结尾一到两句。另一种具体实现的方法是，概念抽取先按照上述原则进行，其余部分的抽取随后继续。后来抽出的任何一个新的概念都将被加入到概念池中。

为了避免用户等待，一种具体实现的方法是，在界面 600 被打开时，由搜索引擎返回的网页搜索结果显示在 650, 612 中显示的概念列表以及 616 中显示的过滤特征都将被激活，搜索结果网页的顺序也将根据结果相关度的排序结果改变。另一方面，由于本地文件已预先被抽取和建立索引，故硬盘搜索结果部分的概念，过滤特征以及相关度计算也将很快被激活。每当部分搜索结果被下载并进行概念提取后，用户才能点击搜索引擎在 408 或 621 返回的 URL 对应的搜索结果来读取网页或文件，或点击按钮 470

或 670“Next”来翻看下一页搜索结果，或通过 412 或 612 选择或排除概念列表中的概念来进行概念过滤。在这种情况下，概念列表一直在更新中，也就是说，对于搜索结果的下载以及下载文件的概念抽取一直在进行，以此来更新概念列表，同时根据用户对列表中概念的选择和排除来进行相应的结果过滤。当用户点击 408 或 621 处搜索引擎返回的链接来观看网页或文件内容时，只要该网页或文件已经被下载或正在下载，保存在本机上的下载版本或正在下载的文件将被直接提供给用户界面，并通过 408 或 621 显示给用户。当用户点击 408 或 621 处搜索引擎返回的链接来观看网页或文件内容时，如果该网页或文件还未被下载，则直接通过搜索引擎返回的 URL 进行下载，并保存至下载队列，同时进行概念抽取以及过滤特征提取。一种具体实现的方法是，当用户点击 408 或 621 处搜索引擎返回的链接来观看网页或文件内容时，该网页或文件将被移至处理队列的最前面来进行概念抽取以及过滤特征提取。另一种实现方法是，当用户点击 408 或 621 处搜索引擎返回的链接来观看网页或文件内容时，如果下载程序仅下载了网页文本部分，则根据搜索引擎返回的 URL 直接重新下载网页所有内容包括图像部分，这样可以显示给用户完整的页。

通常，搜索一个关键字/词会返回大量的搜索结果。在搜索引擎的一种实现中，网页和文件中的重要概念被事先抽取并为其建立索引，这样就可以在概念列表中排序和列举搜索结果的网页和文件中包含的所有重要概念。但是，如果概念抽取和索引建立都在客户机个人 PC 上完成，独立搜索引擎中排序比较靠后的结果将会长时间得不到下载和分析。举例来说，如果下载程序按照原始搜索引擎的顺序下载返回的一百万个搜索结果，则第 999, 901 到一千万页将等待很长的时间才能被下载。在一种实现中，为用户提供一个选择面板，让用户来选择哪一部分搜索结果应该被优先下载和分析。对于最先的 1000 个将要下载和分析的网页和文件，允许用户设置在搜索引擎返回的结果列表的开始、任意中间位置和结尾按照一定比例下载。有些文件由于较新或被链接的较少，使它们处于返回结果的中间或是结尾，但它们可能包含最新的相关信息，如果最先下载和分析这些结果，用户就可以第一时间浏览包含在这些结果中的重要信息。然而这些结果在用户平常使用搜索引擎时是不会被看到的。当用户需要下载搜索结果用来分析和抽取概念时，用户也可以选择下载 M 个网页和文件来节省硬盘空间，如下载 1000 个网页或文件。保存 M 个搜索结果可以使用户在需要时迅速浏览它们而无需等待下载。用户的空余空间越大，他可以下载的页就越多。下载的网页或文件超过指定数量时会自动删除那些已分析和抽取过概念的网页或文件。用户还可以设置一定容量如数个 MBs 来储存下载的结果。当下载的结果超过了容量时，后来下载的结果也会覆盖那些分析和抽取过概念的结果。默认的容量可以设置在 100MB。在一种实现中，可以让用户选择一组规则来决定哪些下载的文件会被保留在分配的存储区中，比如保留所有大于 0.5MB 的文件。这样设定之后大的网页或文件就可以在用户需要查看时被迅速打开而无需等待下载。而小文件因为下载速度快，可以在用户浏览时实时下载而无须保存。当过多的网页和文件需要下载时，不符合给定规则集的网页和文件就将被覆盖以限制空间的使用。

### 3.3 概念的相关度排序及以概念过滤搜索结果

本发明使用自然语言处理，根据搜索结果和搜索关键字/词串的相关度对搜索结果排序。它改进了原有相关度排序方法。一种具体实现的情况是，本项发明将基于内容的相关度排序和搜索引擎本身的排序——如基于投票和流行度的加权平均的 Google 排序算法相组合进行新的排序。

#### 3.3.1 搜索结果的相关度排序

每一个搜索结果都可以根据它的链接情况来排序，或因为使用了其它搜索引擎的结果，那个搜索引擎已经对结果进行了一个排序，如 Google 或 Yahoo。Google 的基于链接的页排序以及其它搜索引擎的排序，都不能很好的表示出结果的相关度。

当用户使用两个或更多的关键字/词进行搜索时，他明显希望返回和关键字/词相关并且文章中含有关键字/词的搜索结果。在原来的搜索引擎中，当用户使用两个或更多的关键字/词进行搜索时，得到的搜索结果网页中关键字/词可能出现在不同的框架中或完全不相关部分中。再举个例子，当用户使用精确短语匹配进行搜索时，如搜索和短语“价格改变”“price change”的精确匹配，以前的搜索引擎经常返回被标点分割的短语，如“...固定价格。改变地址...”，“...fixed price. Change of address ..”，在这个例子中，单词“价格”“price”和“改变”“change”同时出现但是这两个词本身无关并且和用户希望得到的结果无关。

通常页、文件或文章的创建和修改时间也是有用的排序相关因素，这是因为用户往往对最近或是特定日期范围的信息感兴趣。一种实现可以利用基于内容的相关度排序、日期排序和基于链接的排序的加权组合来建立一个新的页排序，如下所示：

$$\text{搜索结果页 } i \text{ 的排序} = PR(i) = a * \text{基于链接排序} + b * \text{相关度排序} + c * \text{日期排序}$$

此处的 a、b 和 c 都是正数且  $a+b+c=1$ ，分别代表基于链接排序、相关度排序和日期排序的权值。举例说明， $a=b=0.4$ ， $c=0.2$ 。假定基于链接的排序最大值是 10。当  $c \neq 0$ ，可有一个默认的日期排序，比如默认日期排序 = {10, 如果  $t \leq$  一周; 8.5 如果  $t \leq$  1 个月, 等等}，此处 t 是页或文件的创建或修改日期。当用户没有在左侧面板 416 或 616 中选择日期范围时，使用默认日期排序。若用户选择了日期范围，则可使用一个选择日期排序，比如默认日期排序 = {10, 如果 t 在选择的日期范围内; 8 如果 t 在选择的日期范围外 1 个月内, 等等}。相关度排序可由以下步骤计算：

1. 每一个由用户输入的关键字/词或其词根变形都带有  $10/N$  的点数。在一个关键字/词扩展为一个概念的情况下，出现在关键字/词的扩展集中的词语的点数为  $9/N$ ，一个出现在关键字/词的上下文中的词语的点数也为  $9/N$ ，关键字/词的子类词的点数为  $9/N$ ，而母类词的点数为  $7/N$ ，此处 N 是用户输入搜索栏的关键字/词个数。

2. 相关度排序 =  $(R1 + R2) / (10N - 1)$ ， $R1 = 10 * P1 * P2$ ，这里  $P1 =$  (两个搜索关键字/词序列按照用户输入顺序在文章中出现的次数)， $P2 =$  这些词语的点数之和，

$R2 = \max\{\max_{\text{所有句子}}[9 * \Sigma (\text{同一句中关键字/词的点数, 不越过逗号和回车})], \max_{\text{所有句子}}[8 * \Sigma (\text{同一句中关键字/词的点数, 不越过句号、分号或换行符})], \max_{\text{所有句子}}[6 * \Sigma (\text{同一段中的关键字/词点数})], \max_{\text{所有句子}}[5 * \Sigma (\text{相邻段中的关键字/词点数})], \max_{\text{所有句子}}[4 * \Sigma (\text{同一区域中的关键字/词点数})], \max_{\text{所有句子}}[3 * \Sigma (\text{同一页中的关键字/词的点数})]\}, (10N - 1)$  是归一化因数。

在计算  $R1$  时，当 M 个关键字/词，且 M 是大于 2 的正整数，按照用户输入的确切顺序依次出现时， $P1=M-1$ 。比如，如果用户输入关键字/词串（无线网络安全 wireless network security）（注：词且分将把无线网络安全且分成 3 个词：无线，网络，安全），然后在页中找到如下的两个词短语（无线网络 wireless networks）（网络安全 network security），此时  $P1=2$ 。如果这个页包含 3 个词的短语（无线网络安全 wireless network security）， $P1=2$  仍然成立。这是因为（无线网络 wireless network）被计为是 2 个在一起的关键字/词，而（网络安全 network security）同样被计为两个在一起的关键字/词。在一种实现下，一个短语，例如（无线网络 wireless networks）和（网络安全 network security），出现次数是不被计数的。每个短

语只记一次。如果用户仅搜索一个单个词语，此时  $P1=0$ ,  $P2=90$ ,  $R2=9*10/(10*1-1) =10$ 。

为了保存计算结果，一旦搜索关键字/词序列中的所有两词短语被找到， $R1$  达到最大值  $R1=10*(N-1)*10$ 。重要概念的抽取和排序程序会停止为计算  $R1$  而进行的文本搜索。相似的，一旦找到一个包含所有关键字/词的句子，程序也将停止为计算  $R2$  进行的文本搜索。举例来说，用户输入（无线网络安全平台的实现 wireless network security platform implementation），如果程序已经找到下面的短语（无线网络安全 wireless network security），（安全平台 security platform）和（平台的实现 platform implementation），它将停止为计算  $R1$  而进行的文本搜索，此时  $P1=4$ ， $R1=10*4*10$  达到可能的最大值。如果所有这些短语都出现在一个句子中，且没有逗号，程序也将停止为计算  $R2$  进行的文本搜索，且  $R2=9*10$  也达到了极值。在这个例子中，相关排序是  $(400+90) / (10*5-1) =10$ 。这个相关度排序的定义使得它在很多情况下很可能只需要扫描一部分文本就可以计算出页或文件的相关度排序。

在一种实现中，页甲基于链接的排序由指向页甲的链接数和类型以及指向页甲的页的基于链接排序。另一种实现是通过以前的搜索引擎实施网络搜索，基于链接排序的条件可以直接使用此搜索引擎的排序结果，例如 Google 或 Yahoo 的排序，或是这些排序的一个函数。对于本地计算机搜索出来的文件，由于没有或只有有限的超链接，可以把所有文件的基于链接的排序值设为 10。或是，可以把所有文件的基于链接的排序值设为 0，同时可将相关度排序项的权值增加到 1。

用户或许希望改变页排序公式中给定三个因素的权值。例如，用户或许对那些相关度排序中高的最近的页更感兴趣，而对那些在基于链接排序高的页不那么关心因为基于链接排序可以被链结场（Link Farms）和链接交换（Link Exchanges）操纵。所以他可能希望选择权向量  $(a, b, c) = (0.2, 0.5, 0.3)$ 。一种实现用可变的三个滑动条界面让用户改变权值，如图 11 所示。一种实现中，用户仅可以改变两栏，比如相关度排序和基于链接排序，因为三者权值相加为 1，而第三项基于文件创建和修改日期的权值可以通过一个权值计算程序自动算出来。在另一种实现中，用户可以调整三个栏目，但是，用户选择的三个向量值将由权值计算程序自动归一化使得其和为 1。

作为对考虑到关键字/词在文章中出现的顺序而计算相关度排序的扩展，在一种实现下，搜索程序支持以“同样的顺序”搜索模式。这个模式获取的搜索结果是包含搜索关键字/词，并且关键字/词出现的顺序和用户原始输入的顺序完全一致的网页或文件。此程序可以进一步支持仅获取关键字/词之间没有标点的网页或文件。正如前面寻找“价格改变”“price change”的例子一样。在另一种实现中，仅考虑关键字/词出现的顺序，而关键字/词之间可以出现词语或文章片断。

本发明的搜索结果相关度的排序的实现提供了一种计算在搜索结果里的一个文件的排序的新方法，该方法包括：

在文件中识别出和用户输入的定义搜索的描述的部分或全部相同或同类或相似的一或多个匹配信息元；

基于在文件中的下列一或多个因素计算一个相关度排序参数：一或多个匹配信息元和它们在定义搜索的描述中的相应部分的相同或同类或相似的程度；两个或更多个匹配信息元出现的顺序和它们在定义搜索的描述中的相应部分出现的顺序的比较；两个或更多个匹配信息元在句子或文体结构里的相对位置；在两个或更多个匹配信息元是否出现标点符号或其它符号；一或多个匹配信息元的格式；一或多个匹配信息元在文件里的角色；一或多个匹配信息元在文件里出现的位置或部分；及是否由和专门针对一个用

户的信息相似的信息出现及它们之间的相似程度。

### 3.3.2 选择从单个页或文件以及搜索结果集合中抽取的概念

对每一个页或文件，抽取出来的重要概念分为 A 到 F 组并在每个组内排序，用户可以根据如前所述的百分比分配选择特定的重要概念。对页和文件的重要概念的抽取、排序和选择在前面已经描述。如果用户选择在重要概念列表 412、612、712 或 912 中显示的 N 个重要概念，本发明的重要概念抽取和排序程序就会对结果集中每一个网页或文件抽取前 N 个最重要的概念。这个称为抽取集的集合，可能是搜索结果中的所有网页和文件，也有可能只是其中的一部分。当重要概念抽取和排序程序只对所有网页和文件中预先确定和预先选择的部分进行抽取时，抽取集是结果集的一个子集。另一种情况是用户在程序完成对所有网页和文件的分析抽取之前结束了程序也会导致抽取集是结果集的子集。另外，当程序仍在运行且未完成对所有文件的抽取时，抽取集也只是搜索结果的子集。这种情况下，随着程序对网页和文件抽取的完成，抽取集也不断增大。如果  $N \geq 6$ ，则页或文件中组 A 到 F 都至少有一个重要概念被选取。如果  $N < 6$ ，其中一些组，例如 B, D, E 可以被忽略。然后，每个来自抽取集中的网页或文件的前 N 个概念可放入抽取概念池。重复或子集概念将从概念池中删除。然后，对概念池中的概念排序。具体实现可用下面的公式计算：

$$\text{概念 } j \text{ 的概念排序} = CR(j) = c * 10 * \max\{Na(j), (Nt - Na(j)) / Nt + d * \{\sum_{\text{所有含概念 } j \text{ 的页}} PR(k)\} / Na(j)\}$$

此处  $c > 0$ ,  $d > 0$ ,  $c+d=1$ ,  $Nt$  是当  $CR(j)$  计算时，抽取集中的网页和文件总数。 $Na(j)$  是抽取集中包含概念  $j$  的网页和文件总数。注意  $Na(j) > 0$ ，因为抽取概念池中的概念至少包含在一个网页或文件中。还要注意到对所有概念而言  $CR(j)$  的最大值是 10。这个排序公式同时对最流行概念 MPC 和最新鲜概念 MOC 排序，这很有意义。因为通常这两类概念比中间概念携带了更多的信息。MPCs 是那些大多数搜索结果认为重要的概念，因此它们很可能是重要。这很像 Google 这类搜索引擎的排序算法。另外，MOCs 则是那些搜索结果中小部分结果认为其重要的概念。因此它们往往与平常的看法有所不同。通常，新发现往往是注意到大众所不关注的，或走一条不是大家都走的路。所以 MOCs 也可能是重要的，所以本发明可以把它们排在前面。相比之下，在先前搜索技术下，稀有概念被掩盖在平凡概念中，使得用户无法看到它们。权因子  $c$  代表一个概念流行和新鲜程度的权重，权因子  $d$  代表包含此概念的网页和文件的平均排序。例如  $c=d=0.5$ 。

在一种实现中，概念抽取和排序算法提供一个接口让户选择两个正值 A 和 B，且  $A+B=N$ ，这样可以选择在重要概念列表 412, 612 or 712 种显示 A 个 MPCs 和 B 个 MOCs，其中 N 是在重要概念列表中显示的概念总数。MPCs 和 MOCs 的排序可以依据下式计算：

$$\text{概念 } j \text{ 的 MPC 排序} = CR(j) = c * 10 * Na(j) / Nt + d * \{\sum_{\text{所有含概念 } j \text{ 的页}} PR(k)\} / Na(j)$$

$$\text{概念 } j \text{ 的 MOC 排序} = CR(j) = c * 10 * (Nt - Na(j)) / Nt + d * \{\sum_{\text{所有含概念 } j \text{ 的页}} PR(k)\} / Na(j)$$

### 3.3.3 在搜索时计算相关度排序和概念排序

计算相关度排序要求知道用户搜索时使用的关键字/词，所以只能在搜索时计算。在重要概念抽取的 6 个组中，组 A、C、D、E 和 F 可以提前取得，但组 B 只能在搜索时取得。这是因为它需要用到搜索时所使用的关键字/词信息。在预处理阶段，可以抽取组 A, C, D, E 和 F 中的重要概念，这些重要概念的索引  $B_{SE}$  和  $C_{SE}$ ，或  $B_{IP}$  和  $C_{IP}$ ，或  $B_{PC}$  和  $C_{PC}$  也可建立。而页排序 PR 和概念排序 CR 则在搜索时计算。

在一次新搜索完成之后，用户在概念列表上选择一个概念后，程序会自动进行对概念的过滤，这等价于把概念作为附加关键字/词的又一次搜索。所以，相关度排序和页排序 PR 需要重新计算。具体实现时，为了减少概念过滤处理的次数以使得过滤结果可以迅速的显示给用户，相关度排序和页排序只在新搜索时计算一次，过滤结果的排序直接应用原始结果相关度排序的结果。具体实现时，概念排序 CR 是由过滤结果计算而来，根据新的排序，概念列表也将更新。另一种实现，为进一步减少概念过滤的处理时间，概念排序 CR 和概念列表不会改变且始终与原始搜索得到的结果一样。在目前的实现下，用户可以选择以上两种方式的任意一种。在一种实现中，只抽取组 A、C、D、E 和 F 中的重要概念，而不抽取概念组 B 的概念。这样，所有的概念抽取可以预处理，排除了在搜索时抽取概念的必要。这进一步减少了搜索时的负担。

如上所述，概念提取、概念过滤和 CPM 既可以在搜索引擎服务器上处理，也可以在用户的 PC 上处理，或两者各自处理一部分。相似地，相关度排序、页排序 PR 和概念排序 CR 也可依照上述方式完成。在个人计算机上处理可以应用网络上千万台个人计算机的处理能力，而不在搜索引擎服务器上集中处理。后者需要同一时间处理高达数以亿计的用户请求，需要使用大批的计算机集群或服务器集群。

在一种实现中，当索引  $C_{SE}$ ，或  $C_{IP}$ ，或  $C_{PC}$  在一个搜索执行前第一次被建立时索引的每一个条目都是网页或文件和搜索结果中抽取出来的所有重要概念列表的一个图示，这些重要概念不包括那些需要知道用户搜索关键字/词后才能抽取出来的概念。列表中的概念数都要除以一个最大值，例如 100，得到的百分比分配到如前所述的每一个组中。分配给组 B 的百分比可以保留到搜索时。组内的概念可以排序。对于组 A，依赖搜索关键字/词的排序部分现在可以忽略。在每个页或文件的索引  $C_{SE}$ ，或  $C_{IP}$ ，或  $C_{PC}$  条目中的排过序的重要概念列表称为预搜索排序列表 (PSRL)。在搜索时，搜索关键字/词是已知的，这样组 B 的概念也可以抽取和排序，而组 A 概念却可以重排序。对每个页或文件的索引条目中的 PSRL 的修改得到了搜索时间排序列表 (STRL)。当选择 N 个重要概念时，每个 STRL 组中的选择依据便是先前分配的百分比。这 N 个来自页或文件的概念被放在一起并且去除了重复的概念和子概念，计算余下的概念得到了概念排序的结果。重要概念列表 412 和 612 上显示的概念就是从重要概念池中选出的排序最高的 N 个概念。在另外一种实现中，为了减少处理时间，每组中排序最高的概念直接从页或文件抽取的索引条目中的 PSRL 得到，此时没有抽取概念组 B 的概念也没有重计算组 A 的概念排序。

本发明的概念或其它信息元的提取和排序的实现提供了一种新的信息搜索方法，该方法包括：

从由一或多个文件或其部分形成的一个集合（称为甲集）里提取的信息元集合中获取一或多个信息元；对上述获取的一或多个信息元基于下列一或多个排序参数进行排序：

对一个从一组文件中提取的信息元，基于这组文件的一个链接流行度排序的一个函数；基于这组文件的一个相关度排序的一个函数；基于这组文件的一个日期排序的一个函数；一个信息元可从更多的文件里提取出来则把此信息元的排序提高；一个信息元可从更少的文件里提取出来则把此信息元的排序提高；一或多个信息元和另一个信息元的集合（称此集合为乙集）里的信息元的关系；一或多个信息元在文体里的位置、格式或角色；一或多个信息元出现的上下文；一或多个信息元的含义。

上述的方法还可进一步包括下列一项或多项：甲集是一个搜索的结果(称此搜索为甲搜索)，甲搜索是由一或多个描述定义的；乙集里的信息元包括一或多个重要字/词和/或短语，句型，概念或含义和论语；提供一个用户接口让用户调动一或多个排序参数的权重。

#### 4. 本地计算机文件搜索

在一种实现下，用户界面提供给用户在本机上对文件进行搜索的选择，如图 1, 3-7 和 9 所示的工具栏选项“启动硬盘搜索”所示。这个整合了网络搜索和本地搜索的界面更加具有亲和力。在具体实现时，网络搜索结果和本地搜索结果都会在同一个窗口中显示，如图 4 和 6 所示。另外一种方式是用户可以选择将本地搜索结果显示在一个独立的窗口中，如图 7 所示，只需点击按钮 430 或 630 “新窗口显示硬盘搜索结果”，这样就有足够的空间显示详细的结果信息。当用户进行网络搜索时，一旦用户选择了“启动硬盘搜索”，PC 的硬盘搜索也将同时进行。另一方面，当用户选择只在本地搜索即点击按钮“仅搜索硬盘”时，搜索关键字/词和其他信息都不会发送给搜索引擎服务器。

硬盘搜索程序会预先建立索引  $A_{PC}$ ,  $B_{PC}$  and  $C_{PC}$ 。这三个索引的使用和关系显示在图 10 中。索引  $A_{PC}$  由关键字/词组成并且图示到包含此关键字/词的文件列表。当查询到关键字/词时，返回的是包含关键字/词的文件名和路径。这个索引的功能即利用关键字/词查找文件。 $A_{PC}$  的关键字/词是从文件名、文件属性文本域（就像通过在本机文件名上点击鼠标右键得到的属性信息一样）和文件文本域中抽取的，搜索程序可以使用文本内容作为文件文本的索引，举例而言，电子邮件文件，图像文件，音/视频文件，程序文件或各种各样应用文件像 Word、PPT、Pdf、html 等等。

索引  $B_{PC}$  由从硬盘上的文件抽取出的重要概念建立，每个索引图示到抽取出重要概念的文件名列表和文件路径。当查询到一个重要的概念时，例如概念过滤时，通过对重要概念列表进行选择来生成 CPM 时，返回的结果是文件名列表和路径名。类似地，FTFI 的建立是为了 716 中列出的每个过滤特性。当查询过滤特性时，返回的也是包括过滤属性的文件名和路径。

索引  $C_{PC}$  由文件名建立且图示文件到从文件中抽取出的重要概念上。当由文件名和路径查询时，比如从搜索结果中检索和选择 N 个重要的概念时，鼠标指到文件名上显示文件所包含的概念，这时返回的结果将是文件中抽取出的已排序的概念列表。这三个索引或许由一个文件组织起来也可能由一批独立的文件组织起来。相似地，在 416 或 616 中的其他过滤条件，如文件类型、日期范围等，也可以从搜索结果中抽取出并组织起索引，使得按属性的过滤可以实施的快一些。

为了提供硬盘搜索的结果和用户选择属性进行过滤的快速图示，硬盘搜索程序预先处理每个文件的重要概念抽取和排序、其他过滤属性的抽取并建立索引。在硬盘搜索程序第一次安装时，它就在后台执行这些任务。为方便告知用户进度，程序会显示一个进度条，例如在窗口工具栏上显示。工具栏将会显示总文件中的多少文件已被处理和索引过。其格式如下所示：“共 923, 588 页 / 文件，925 个已分析检索完成”。在所有文件都被索引后，它会告知用户程序已经准备好可以立刻开始搜索和分析 PC 硬盘上的文件。如果 PC 关机或程序被中断，下次启动时它可以自动从上次中断处继续进行。

如果硬盘上添加了新的文件，索引的建立、概念的抽取和排序以及文件属性的提取都可自动完成，新的结果会添加到索引中。这种更新是阶段性的，而这个阶段的间隔长短可以由用户在浏览器工具栏的选项上自己选择。默认的更新间隔是每天或每周特定时间晚上 10 点，前提是电脑处于开机状态且空闲。

索引建立之后，硬盘搜索结果可以由索引  $A_{PC}$  迅速得到，抽取出的重要概念可以由索引  $C_{PC}$  迅速得到。所以，当用户输入关键字/词后，搜索结果和概念中较高权值得部分可以迅速在 721 和 712 中显示出来。同样，当鼠标悬浮在文件名上时，来自索引  $C_{PC}$  中的重要概念也将显示在一个小窗口中。一旦鼠标离开，小窗口将消失。当双击文件名后，文件将通过相应的应用程序被打开。用户在重要概念列表上选

择或排除概念时，过滤的结果利用索引 C<sub>PC</sub> 和 FTFI 也可以很快得到。

在另一实现中，当用户点击日期、文件名、文件夹或日期域 752，本地控制程序根据用户点击的域相应地以降序或是升序改变硬盘搜索结果的排列方式。这样的界面操作和用户熟悉的 Windows 界面很相似。另外，如果当用户实行搜索时本地计算机没有连接到网络，搜索将自动地被解释为仅硬盘搜索执行而且只执行。

当本地计算机连接了网络时，本发明提供给用户可以只选择搜索硬盘而不实行网络搜索的功能，此时用户只需点击按钮“仅搜索硬盘”。当用户点击了按钮后，本地控制程序调用硬盘搜索程序，并告知它只搜索硬盘而不把用户输入的关键字/词或 NLDS 提交给任何搜索引擎或网络上的计算机。特别是当用户希望进行本地文件的隐私查询而不希望搜索引擎知道时，这种功能就非常有利。仅进行硬盘检索时，硬盘搜索的结果会显示在一个带有左面板的窗口中，左面板显示了重要概念列表和其他的过滤条件，第二个面板显示 PC 硬盘搜索的结果。整个情况如图 7 所示。在一种实现中，当按钮“仅搜索硬盘”被点击后，本地控制程序会在用户计算机上显示一个 IE 页，如图 5 所示，这同早期的搜索引擎界面类似，但输入的关键字/词只被用来搜索本机上的文件。另一种改进的界面如图 12 所示，这种界面提供了新的特性，包括把关键字/词扩展成概念，“可能用到的词”，概念跟踪和链接跟踪。在另一种实现中，本地计算机连上网络，硬盘搜索与本地搜索同时进行，但是两者的结果独立，每个都有自己的文本区用来获取用户的关键字/词输入。

快速硬盘搜索使得任何人都可以方便地获取计算机上的信息。一个未经许可的用户可以从计算机上迅速地找到一些私人信息。他所需要的时间只是不经意的几秒钟。因此，有必要为这种私人信息加以保护以避免进行硬盘检索时此类信息被暴露。

一种实现方法是在进行硬盘搜索前为硬盘搜索程序加上密码或是使用其他的用户认证办法，另一种实现方法是在搜索特定硬盘、分区、文件夹或文件时需要密码或其他认证方法。如果一个用户输入了正确的密码或认证信息，程序就会返回未受密码保护以及特定受密码保护的或认证保护的硬盘、分区、文件夹或文件的搜索结果。否则，硬盘检索程序只返回未受密码保护的硬盘、分区、文件家伙文件的搜索结果。然而在另一种实现中，当用户输入正确的密码和认证信息，硬盘检索程序只返回受特定密码保护或认证保护的信息。还有一种实现方式是，硬盘搜索程序对每个特定受密码保护或认证保护的硬盘、分区、文件夹或文件要求密码或认证信息，但是存在超级密码或认证，一旦输入成功，搜索结果就将返回所有的信息，不论该信息是不是受保护。

在一个实现中，一个保护数据文件或一个保护数据库用来存储所有硬盘、硬盘分区、文件夹或文件。硬盘搜索程序或文件保护程序引用数据库来决定是否需要密码或某种用户的授权来执行搜索、显示搜索结果、打开文件、修改文件、打印文件或执行一个文件操作。硬盘搜索程序或文件保护程序能够提供给用户一个交互界面，用以添加、编辑或删除保护数据文件或保护数据库上的硬盘、硬盘分区、文件夹或文件。一个实现中，在进行过一次硬盘搜索后，硬盘搜索程序询问是否用户想要保护任何硬盘、硬盘分区、文件夹或文件。如果用户选择保护一些硬盘、硬盘分区、文件夹或文件，它们会被添加到保护数据文件或保护数据库。

在某些实现中，用户对于搜索计算机上的某些特定信息的保护感兴趣。在一个实现中，当用户使用确定的词语、短语、句子或概念搜索信息时，或在搜索结果里显示文件，而这个文件的文件名字、文件

类型、属性、作者、文本内容或其他文本特征（指的全部为内容）中包含确定的词语、短语、句子或概念，硬盘搜索程序需要密码或授权。在另一个实现中，这种通过文件内容来保护文件的方法被更进一步地扩展到文件保护程序，它基于其内容来保护文件不受其他文件操作的影响。在这种扩展的实现中，如果文件的文件名、文件类型、属性、文本内容或其他的至少匹配一条规则的文本特征中包含确定的词语、短语、句子或概念，文件保护程序需要密码或用户的一种授权，目的是为了打开文件、修改文件、打印文件或执行一个文件操作。

在一种实现中，保护数据文件或保护数据库被用来存储所有的词语、短语、句子、概念和规则。硬盘搜索程序或文件保护程序查询数据库，决定是否需要密码或用户的某种授权来执行搜索、显示搜索结果、打开文件、修改文件、打印文件或执行一个文件操作。硬盘搜索程序或文件保护程序能够提供给用户一个交互界面，用以添加、编辑或删除保护数据文件或保护数据库上的词语、短语、句子、概念和规则。在一种实现中，硬盘完成搜索后，上述的交互界面询问用户是否需要保护这次搜索。如果用户选择保护这次搜索，这次硬盘搜索所用的关键字/词就会被添加到保护数据文件或保护数据库。在另外的实现中，硬盘搜索程序或文件保护程序能够将保护文件或保护数据库中的词语或短语拓展为概念，例如，将词语或短语拓展，使其包含同义词集合（synsets），母类词（hypernyms），以及子类词（hyponyms/troponyms），在某种意义上，类似于这个发明中1.2节所描述的关键字/词到概念的拓展方法。

在以上的实现中，为了保护信息不被非授权用户执行的硬盘搜索所查找到，硬盘搜索程序可以在它搜索特定硬盘、硬盘分区、文件夹、关键字/词或概念之前，要求用户输入密码或用户的授权。另一个选择，硬盘搜索程序可以搜索所有硬盘，包括受保护的硬盘、硬盘分区或文件夹，或使用受保护的关键字/词或概念搜索，而不需要用户的密码或授权。搜索后，如果从受保护的硬盘、硬盘分区或文件夹中检索到了文件，或，如果通过使用受保护的关键字/词或概念来搜索而检索到了文件，那么，硬盘搜索程序在显示包含受保护的关键字/词或概念的文件之前，需要输入密码或用户的授权。如果用户不输入密码或授权，硬盘搜索程序不会返回受保护的硬盘、硬盘分区或文件夹上的搜索结果，或不返回包含受保护的关键字/词或概念的文件。

本发明的信息保护实现提供了一种保护信息的新方法，该方法包括：

将一或多个文件或其部分的一或多个特性、信息元或内容的描述保存在一个集合里（称为甲集）；

含有甲集部分或全部信息的文件或其部分形成另一个集合（称为乙集），要求用户通过一或多个保护措施才允许用户读或写乙集里的文件或其部分或得到在乙集里的文件或其部分的信息。

上述的方法还可进一步包括下列一项或多项：

允许用户读或写乙集里的文件或其部分或得到在乙集里的文件或其部分的信息是为用户进行一个搜索，并包括将用户提供的对此搜索的描述和甲集的信息相比较以决定是否要求用户通过一或多个保护措施才进行此搜索；甲集进一步包括一或多个规则以决定用户可对含有甲集部分或全部信息的文件进行哪些操作；检查并标记一或多个文件是否含有甲集部分或全部信息，将标记为含有甲集部分或全部信息的文件加入乙集。

## 5. 链接与概念跟踪

使用以前的搜索引擎在互联网上要达到广而精的搜索，用户通常需要在计算机前浪费大量的时间。用户需要跟踪使用原始关键字/词搜索到的结果中的网页上的链接或文件，在其中找到新的关键字/词，等

待下载大的文件。本发明通过自动识别链接和重要的关键字/词或概念并跟踪，自动跟踪并下载大的文件到用户计算机，而不需要用户的参与，从而使这一搜索过程自动化。这扩大了搜索范围，可以检索到那些潜在的有用的信息，而这些信息有可能被以前的搜索引擎技术所忽略掉。使用以前章节描述的本发明的方法，扩展的搜索结果可被分析、提取概念、排序、组织、过滤和图示化。因此，本发明不仅在更大的范围内检索到了更多的信息，扩展了搜索范围，而且为用户提供了分析和图示法工具，用以从海量的信息中提取到有用的信息。同时，许多浏览网页工作是自动的，这样就节约了用户的时间并提高了效率。所有的工作都可以在用户做其他工作或阅读网页时在后台被执行。

在一种实现，一个自动浏览器提供给用户一个交互界面，使用户能够选择概念跟踪的深度和链接跟踪的深度，例如 116 和 118，或 316 和 318，或 1216 和 1218 所示。假设用户输入了原始的搜索关键字/词并选择了概念和链接跟踪的深度 D。自动浏览器需首先利用原始关键字/词检索网络搜索结果。然后，提取到 K 个最重要的概念或来自于每个网页或文件的重要的链接，这些网页或文件以搜索引擎的搜索结果的排列为顺序或以用户选择的格式排序，这样重要的概念或重要的链接都首先是从排序最高的网页或文件上提取出来。参数 K 是一个正整数，它可以被设置为默认值或由用户选择。重要的概念或重要的链接可以在搜索之前由搜索引擎预提取并排序，或通过下载并分析搜索结果页，在用户本地的计算机上提取和排序，或通过预处理和搜索时处理或搜索引擎处理和本地计算机处理的联合来提取和排序。在概念跟踪中，自动搜索程序使用从每个网页或文件提取出的 K 个重要概念来执行额外的网络搜索。这些网络搜索叫做一级或深度 1 概念跟踪。从一级概念跟踪中得到的网络搜索结果被添加到搜索结果中。自动浏览器从每个网页或文件提取到 K 个最重要的概念，在某种程度上类似于用于概念过滤的重要概念抽取，然后把这些抽出的重要概念作为关键字/词，再来进一步执行网络搜索。这些网络搜索叫做第二级或深度 2 概念跟踪。上述过程对于使用原始搜索关键字/词搜索到的网页或文件，对于 D 级或深度 D 来说，每个概念跟踪结果里的网页或文件要重复执行，直到全部重要概念都被跟踪，直到用户终止该过程后停止。D 是一个正整数，它可以被设置为默认值或由用户选择。

在一种实现，一个自动搜索程序使用与概念过滤时的重要概念提取及 CPM 同样的排序来选取用于概念跟踪的 K 个重要概念。描述这些重要概念的关键字/词或短语在概念跟踪过程的搜索中被用作搜索关键字/词。有另外一种实现，组 C 和与组 E 中最少出现的字和短语排在较高的位置上，因为它们更可能扩展原始搜索的结果，这些结果与原始的搜索关键字/词相关，却与原始关键字/词不在同一概念范围。概念跟踪能够成为一种强有力的自动浏览方法，例如，假设用户想要使用原始搜索关键字/词（无线网络安全 wireless network security）调查有关无线网络安全的技术和发明，搜索结果可能包含概念或关键字/词（802.11i），（WPA），（WAPI），（网络接入控制 network access control），（802.1X），（共钥加密 public key encryption），有名气的和新创办的公司的名字等。使用以前的搜索引擎，用户可能需要阅读并手动地点击链接，察看是否有感兴趣的内容，这可能浪费大量时间，并且经常忘记哪些途径已经查看过，哪些途径还没有查看过。更重要地，一些潜在的非常有用的途径可能根本没有被跟踪到。这个发明将能自动地跟踪这些基于重要概念的链接，向用户给出大量的扩展的搜索结果，并且，使用本发明的过滤、排序和 CPM 实现可以将搜索结果过滤、再排序和图示化。本发明可以比基于知识库和领域定义和关系知识库（Domain Ontology）的技术更加高效，因为网络搜索结果能够快速地引入新的进展和正发生的事件，而升级知识库和领域定义和关系知识库（Domain Ontology）需要花费相当长的时间。在上面提到的无线网

络安全例子中，本发明的网络搜索结果能够快速地包括新创办公司的新产品，政府部门制定的新规则，或工业标准的新进展等。在很长一段时间内，这些都可能不会包含到知识库和领域定义和关系知识库（Domain Ontology）中。

有另一种实现，在概念跟踪中，需要知道搜索关键字/词才可进行的重要概念提取和排序以及相关性排序都被忽略掉。跟踪位于 k 级（level-k）的每个重要概念所得到的搜索结果被看作一个 k 级搜索结果池（level-k pool）。这些 k 级池中的搜索结果和提取出的概念在池中排序，在这种情况下，忽略了需要知道搜索关键字/词的重要概念的提取和排序以及相关性排序。然后，多个 k 级池的搜索结果和提取出概念被组合在一起，并且计算这个所有搜索结果组合中每个网页或文件，或重要概念的最终排序。从跟踪一个重要概念得到的在 k 级池中的网页或文件，或重要概念的最终排序可如下计算

$$\text{最终排序} = (\text{产生此池的概念的排序}) * (\text{池中的网页或文件，或重要概念的排序})$$

对于第二级概念跟踪中的网页来说，这个公式意味着这个概念搜索轨迹上的所有重要概念的排序将被链接在一起：

$\text{最终排序} = (\text{原始搜索结果中的重要概念甲的排序}) * (\text{用重要概念甲作为搜索关键字/词串得到的搜索结果中概念乙的排序}) * (\text{用重要概念乙作为搜索关键字/词串得到的搜索结果中的网页或文件，或重要概念的排序})$

本发明可用最终排序选择下一级的链接跟踪中所要跟踪的重要概念，以及选取包含在 412 或 612 的列表中的重要概念。在另外的一种实现中，在概念跟踪中用来做第一搜索关键字/词的第一个重要概念在提取和排序重要概念中被当作搜索关键字/词，这些重要概念依赖于在通过第一搜索关键字/词检索到的搜索结果池中的搜索关键字/词。最终的网页或文件排序，或所有搜索结果组合中的重要概念能够用与上面同样的方式计算出来，除了使用在提取和排序重要概念中的第一搜索关键字/词计算池内排序。

在链接跟踪中，自动搜索程序检索出第一组网页和文件，这些网页和文件通过从利用原始搜索关键字/词搜索到的结果中的网页或文件里提取到的 K 个重要链接所指向，并且将第一组网页和文件以及他们的摘要（如果需要的话）添加到网络搜索结果中。这叫做第一级链接跟踪或深度 1 链接跟踪。然后自动搜索程序从第一组网页和文件中提取出至多 K 个重要链接，检索出第二组网页和文件，这些文件由从第一组中的网页和文件中提取出的重要链接所指向。将第二组网页和文件以及他们的摘要（如果需要的话）添加到网络搜索结果中。这叫做第二级链接跟踪，或深度 2 链接跟踪。上述过程对每个使用原始搜索关键字/词搜索到的结果中的网页或文件重复执行，对于 D 级或深度 D，直到完成每个链接跟踪结果中的网页和文件，或直到跟踪全部重要链接，或直到用户终止了这个过程。

在另外一种实现下，重要概念和需要搜索关键字/词知识的相关性排序的提取和排序规则在链接跟踪中被忽略掉。跟踪位于 k 级（level-k）的链接跟踪的每个重要链接所得到的搜索结果被看作一个搜索结果的 k 级池（level-k pool）。这些 k 级池中的搜索结果和提取出的重要链接在池中排序，在这种情况下，忽略掉重要概念，重要链接和需要搜索关键字/词知识的相关性排序的提取和排序。然后，搜索结果和提取出的重要链接的 k 级池就被组合在一起，并且计算所有 k 级搜索结果组合中的重要链接的最终排序。跟踪重要链接的 k 级池中的重要链接的最终排序为：最终排序 = (产生此池的链接的排序) \* (池中的链接的排序)。

对于第 k 级链接跟踪中的网页来说，这个公式意思是这个概念搜索轨迹上的所有重要链接的排序将

会被链接在一起。最终的排序用来选取在下一级链接跟踪中所要跟踪的重要链接。

为了控制一个搜索使用的处理器信息源总量，除了概念或链接跟踪的深度外，自动浏览程序也可以限制跟踪的重要概念或重要链接的总数，例如，至多 M 个重要概念或重要链接，这里 M 是一个正整数，并且可以设为默认值或由用户选定。这就叫做概念跟踪和链接跟踪的宽度。有一种实现，自动浏览程序首先使用原始搜索关键字/词检索网络搜索结果。然后从每个网页或文件中提取至多 M 个排在最前的重要概念或重要链接。这个提取可以做完搜索结果中所有的网页和文件，或仅仅做完排在搜索结果最前的网页和文件。提取出重要概念或重要链接的网页和文件集合叫做提取集。在另外一种概念跟踪的实现下，自动搜索程序汇集所有从每个页或文件中提取出的重要概念，删除副本和子集概念，并对剩下的重要概念重排序，其形式与引入到重要概念列表（List of Important Concepts LIC）中的最靠前的 N 个重要概念的选择相同。然后，用排序靠前的 M 个重要概念作为搜索关键字/词，执行额外的网络搜索。这些网络搜索叫做第一级或深度 1 概念跟踪。第一级的概念跟踪搜索结果添加到搜索结果中。然后自动浏览程序以类似于上述的方式从每个网页或文件提取至多 M 个最靠前的重要概念，汇集所有从每个页或文件中提取的重要概念，删除副本和自己概念，并以上述相同方式对剩下的重要概念重排序。然后，用排序靠前的 M 个重要概念作为搜索关键字/词，执行额外的网络搜索。这些网络搜索叫做第二级或深度 2 概念跟踪。以上过程重复直到 D 级或深度 D 为止。

链接跟踪的另外一种实现，自动搜索程序从每个原始搜索结果里的网页或文件中提取至多 M 个排序靠前的重要链接。自动浏览程序汇集来自于提取集里每个网页或文件中的重要链接，将其排序，并提取出至多 M 个排序靠前的用于链接跟踪的重要链接。然后自动搜索程序检索出第一组由排在上面提到的 M 个最前面的重要链接所指向的网页和文件，并将第一组网页和文件以及他们的摘要（如果需要的话）加入到网络搜索结果中。这叫做第一级链接跟踪或深度 1 链接跟踪。然后自动搜索程序从第一组网页和文件或第一组的子集中提取出至多 M 个重要链接，每个都作为一个提取集。自动浏览程序汇集来自于提取集里每个网页或文件中的重要链接，将其排序，并提取出至多 M 个排序靠前的用于链接跟踪的重要链接。然后自动搜索程序检索出第二组由排在上面提到的 M 个最前面的重要链接所指向的网页和文件，并将第二组网页和文件以及他们的摘要（如果需要的话）加入到网络搜索结果中。这叫做第二级链接跟踪或深度 2 链接跟踪。以上过程重复到 D 级或深度 D 为止。

有一种实现，自动搜索程序通过排序网页或文件里的链接来决定跟踪什么链接。首先，选取主框架中的链接。一个链接的排位由提取出的与此链接语义上相近的重要概念的排位决定。链接的排位由下面的过程确定：

1. 如果 URL 链接是指向一个字串、或短语、或包含一个提取的重要概念的句子的超链接，链接与重要概念的排位相同，否则，
2. 如果 URL 链接和一重要概念在同一句，链接与重要概念的排位相同，否则，
3. 如果 URL 链接和一重要概念在同一段落，链接的排位是重要概念排位的 0.7 倍，否则，
4. 如果 URL 链接和一重要概念在同一章节，链接的排位是重要概念排位的 0.5 倍，否则，
5. 如果 URL 链接和一重要概念在同一帧，链接的排位是重要概念排位的 0.3 倍。

从用于链接跟踪的网页和文件中提取 K 个重要链接的实现时，K 个链接能被分配给 6 组概念，分别命名为 A 到 F，使用与用于概念过滤得重要概念提取的百分比相同的比率。然后这 K 个链接用于跟踪。

如果  $K < 6$ , 可以忽略提取的与某些重要概念组相关的重要链接。

从位于链接跟踪的每一级或深度的所有网页和文件中提取总计  $M$  个重要链接的实现,  $M$  个排列在最前面的重要链接是从每个网页或文件中提取出来的, 并且添加到一个提取重要链接池中。删除链接副本。剩下的重要链接由下面的公式排序:

$$\text{链接 } j \text{ 的链接排序} = LR(j) = e * 10 * \max\{Na(j), (Nt - Na(j))\} / Nt + f * \{\sum_{\text{所有含链接 } j \text{ 的页}} PR(k)\} / Na(j)$$

其中  $e > 0$ ,  $f > 0$ ,  $e+f=1$ ,  $Nt$  是在提取集中的网页或文件的总数,  $Na(j)$  是包含链接  $j$  的  $Nt$  集合中的页总数。注意到,  $Na(j) > 0$ , 因为至少一个网页或文件必须包含一个被接受的链接。还注意到, 对于所有链接,  $LR(j)$  的最大值是 10。这个排序公式将特别常用的和特别不常用的链接都进行了排序。然后, 选取排在最前的  $M$  个重要链接作为链接跟踪。

为了减少用户等待结果的时间, 概念跟踪和链接跟踪过程都是能改进的, 也就是说, 显示一部分结果给用户, 同时自动浏览程序继续按特定的宽度和深度执行概念跟踪和链接跟踪。一旦得到了新的概念跟踪或链接跟踪结果, 他们就被添加到搜索结果中, 显示给用户。通过其它过滤特征由重要概念过滤, 以及 CPM 可以在部分结果上运行, 并且当获得了新的结果后他们可以不断地升级更新。

重要概念和链接的提取和跟踪能够在搜索引擎服务器上执行, 或在用户本地计算机上执行。在搜索引擎服务器端执行的优点是绝大部分的搜索结果不需要下载到用户的 PC 上, 并且可以预先提取和排序一部分或所有的重要的链接和概念, 因此, 在检索搜索中的网页和文件中, 可以立即获得它们。自动浏览程序仅下载大文件到用户的 PC, 这些大文件排位高, 并且需要额外的下载时间。由于概念跟踪和链接跟踪可能依赖于用户在原始搜索中使用的搜索关键字/词, 一些重要概念和重要链接的提取的排序需要在搜索引擎服务器搜索时执行。这种情况增加了搜索引擎服务器端的工作量。当有上百万用户执行自动概念跟踪和链接跟踪时, 这就需要占用搜索引擎非常高的处理器信息源。在本地计算机运行的优点是它利用了大量可获得的宽带连接, 海量的存储器和几百万 PC 中的快速的处理器。然而它需要下载所有的或大量的搜索结果到用户本地计算机, 并且重要概念和重要链接的提取只能在搜索时执行, 因此, 增加了执行概念跟踪和链接跟踪的时间。有一种综合的实现, 它组合了以上两种实现的优点。在这种情况下, 搜索引擎预先为每个网页和文件提取并排序一部分或所有重要链接和重要概念, 保存它们和一些压缩的上下文, 用来为每个网页和文件提取和排序到一个文件中。在搜索时, 在用户 PC 上运行的自动浏览程序下载这些带有预提取得重要链接和重要概念的文件, 以及他们压缩的每个网页和文件的上下文。基于原始搜索中使用的搜索关键字/词分析他们, 计算依赖于搜索关键字/词的概念排序和链接排序中的组分, 并通过阐述的搜索执行自动浏览, 将它们提交到搜索引擎并检索结果。它仅下载网页和文件, 对于这些网页和文件, 需要额外的重要链接和重要概念提取与排序。

本发明的概念或其它特征或信息元的提取、搜索结果过滤、链接与概念跟踪的实现提供了一种新的信息搜索方法, 该方法包括:

称一个含有一或多个文件或其部分的集合为乙集, 从乙集里提取一或多个信息元, 称以此一或多个信息元形成的集合为甲集; 从甲集中选出一或多个信息元并形成丙集; 用丙集去获取另一个含有一或多个文件或其部分的集合(称这个集合为丁集)。

上述的方法还可进一步包括下列一项或多项:

从乙集里提取一或多个信息元形成甲集时使用下列一项或多项来决定提取哪些信息源: 字/词或短语

的列单、句型的列单、概念或意义的列单、字/词或信息元和上述的一或多个列单里的项的关系、字/词或信息元的位置或格式或上下文、字/词或信息元在文本里的角色、信息元是基于哪些准则鉴别出来的、以及信息元属于哪个类别；

乙集是一个搜索的结果（称此搜索为甲搜索），甲搜索是由一或多个描述定义的；

当乙集是一个由一或多个描述定义的甲搜索的结果时，从乙集里提取一或多个信息元形成甲集时使用下列方法之一：

(1). 一或多个搜索引擎利用一或多个信息元和定义甲搜索的一或多个描述的相关性从乙集提取一或多个信息元形成甲集；

(2). 一或多个搜索引擎在甲搜索以前就从存在搜索引擎的部分或全部文件里预先提取一或多个信息元，当甲搜索时，用户的计算机从一或多个搜索引擎下载乙集文件所包含的预先提取的一或多个信息元，用户的计算机利用下载的一或多个信息元和定义甲搜索的一或多个描述的相关性来决定由哪些信息元形成甲集；

(3). 当甲搜索时，用户的计算机从一或多个搜索引擎下载部分或全部搜索结果，并从其中提取一或多个信息元形成甲集；

当乙集是一个由一或多个描述定义的甲搜索的结果时，从甲集中选出一或多个信息元形成丙集包括提供一个用户接口，让用户选择甲集中一或多个信息元，以用户的选择形成丙集，并且用丙集去获取丁集包括把丙集和定义甲搜索的一或多个描述一起当作定义另一个搜索（称此搜索为乙搜索）的描述交给一或多个搜索程序进行乙搜索，并由乙搜索的结果或其部分形成丁集；

当乙集是一个由一或多个描述定义的甲搜索的结果时，从甲集中选出一或多个信息元形成丙集包括提供一个用户接口，让用户选择甲集中一或多个信息元并可将每个选中的信息元设为存在项或不存在项，以用户的选择形成丙集，并且用丙集去获取丁集包括把丙集和定义甲搜索的一或多个描述一起当作定义乙搜索的描述交给一或多个搜索程序进行乙搜索以搜索含有丙集中设为存在项的信息元而且不含有丙集中设为不存在项的信息元的文件或其部分，并由乙搜索的结果或其部分形成丁集；

从甲集中选出一或多个信息元形成丙集时基于对甲集中一或多个信息元的排序进行的；

甲集中一或多个信息元是概念，从甲集中选出一或多个信息元形成丙集包括选择以或多个概念，用丙集去获取丁集包括把丙集中的概念交给一或多个搜索程序进行乙搜索以搜索含有丙集中的概念的文件或其部分，并由乙搜索的结果或其部分形成丁集；

从丁集中提取一或多个概念，并多次重复以上方法；

甲集中一或多个信息元是链接，从甲集中选出一或多个信息元形成丙集包括选择以或多个链接，用丙集去获取丁集包括把丙集中的链接指向的文件或其部分纳入丁集；

从丁集中提取一或多个链接，并多次重复以上方法。

## 6. 监视站点与监视搜索

本发明也可对选中的网站点或网页进行自动监视，并可在一段时间内对用户定义的搜索题目多次进行搜索（称为保持搜索活跃），从而监视并探测与该题目相关的信息改变和新信息。在一种实现中，本发明的用户界面程序显示了使用搜索关键字/词甲得到的搜索结果后，用户界面为每个搜索结果提供一个选项框“监视此网页”。当用户选择了一个网页的此选项框，用户界面程序显示一个小窗口，请求用户指定

他想监视网页的时间段，以及指定本发明的页/站点监视程序检查页变化的频率。时间段和监视频率可以在下拉菜单或在文本框或选项框中选择。举例来说，用户可以指定监视器在 1 周、1 个月、X 月的时间段内，每隔 2 小时检查一次，一天检查一次，一周检查一次等。可以设置默认值，例如，一个月内每天都检查。它也提供选项选择“把监视扩展到同一目录下的所有网页”，“监视此网页和链接到此网页的所有网页”，“监视此网页和此网页链接到的所有网页”和“把监视扩展到整个网站”等。用户界面程序也可以让用户选择当网页有变化时怎样通知他。例如，小窗口提供给用户一个选项，用户可以输入电子邮件地址，这样页/站点监视程序在检查到变化时可以给用户发送电子邮件。还有一个选择，有一个单选框，可以在桌面通知用户。当选定这个选项后，页/站点监视程序会在用户屏幕上弹出一个提示窗口，通知用户监视的页发生了变化。

针对每个被监视的网页，一个网页/网站监视程序将计算和存储一组校验和、或数字摘要，例如，为它将为每一个网页使用 CRC32, MD5, SHA-1。经过一个特定的时间间隔，控制程序将激活网页/网站监视程序，由监视程序下载那些被监视的网页，再次为每个网页计算校验和、或数字摘要并与存储的旧数据进行比较。如果网页/网站监视程序发现新旧数据不一致，则发送一个提示或电子邮件让用户得知被监视的东西已经改变。监视程序再将新的数据和摘要存储下来。如果没有改变，则网页/网站监视程序不作动作。在下一个时间间隔末尾网页/网站监视程序再次被触发，同样的过程将被重复直到整个监视过程结束。网页/网站监视程序也会询问用户是否延长监视时间。在另一种实现中，网页/网站监视程序也允许用户将需要被监视的网页或网站输入到一个列表。通过这种方法，这个发明可以为用户自动监视网页或网站而不需要用户人工地多次启动搜索。如上述，在同一个用户界面上也可以供用户选择修改监视时间、频率、监视网页的范围。

在另一种实现当中，用户在使用关键字/词乙进行搜索之前，可以使用 110 或 312 选择保持此搜索活跃的开始和终止日期。这样的搜索称为持续搜索。如果没有指定开始日期，则默认到搜索第一次进行的日期。相类似的，用户也可以在这个界面上选择一个长为 X 周或 X 月的时间段。在另一种实现当中，用户界面的工具栏或属性项目中提供一个“保持搜索活跃”的按钮。在本发明的用户界面程序显示了用户使用关键字/词乙进行搜索的结果后，用户可以点击工具栏上的“保持搜索活跃”按钮或属性菜单里的“保持搜索活跃”项。这种情况下，用户界面程序显示一个含有一个属性项“保持搜索活跃 X 天（周或月）”的窗口。用户在框中输入一个数字或在下拉菜单中选择日、周或月。在上述的两种实现中，一个持续搜索程序将会计算和存储搜索引擎返回的搜索结果中的每个网页的校验和、或数字摘要，例如，使用 CRC32, MD5, SHA-1。经过指定的时间间隔，一个控制程序激活持续搜索程序，并由它向搜索引擎提交关键字/词乙并进行新的搜索。持续搜索程序从搜索引擎获得新的搜索结果。它重新计算上述的校验和、或数字摘要，并与原先存储的数据相比较。如果持续搜索程序发现两者间有改变，则发送一个通知或电子邮件让用户得知被监视的信息已经改变。持续搜索程序再将新的校验和和摘要存储下来。如果没有改变，则持续搜索程序不做动作。在下一个时间间隔末尾网页/网站监视程序再次被触发，同样的过程将被重复直到整个监视过程结束。持续搜索程序也会询问用户是否要增长持续搜索的时间。上述的方法监视在搜索结果列表中是否有新的网页或文件，也监视网页或文件的排序是否有改变。在另一种实现中，在每次被激活时持续搜索程序存储列表中的页并对列表进行比较。从而它可以发现新的网页或文件，也可以区分是新加入的，还是排序更改的网页或文件。

在另一种实现中，一个持续搜索程序存储搜索结果列表，并为搜索结果中每个网页或文件计算校验和、或数字摘要。当每次持续搜索程序被激活时，它对比前面的搜索和本次的搜索的结果列表以及它们的校验和、或数字摘要。通过这种方法，持续搜索程序不只发现新的或删除的信息源，也发现网页或文件本身的改变。这方法将持续搜索程序与前面所说的监视程序有效地组合。这就把网页监视程序应用于搜索结果的每个网页和文件。所以这样的过程需要大量的计算资源和一定的时间。

在一种实现中，在上面提到的任意一种实现中的持续搜索程序可以是一个渐进过程，当搜索列表中有一定比例的页或搜索结果中的部分网页或文件被处理完毕后发现有改动，这部分的运行结果将被发送给用户。在另一种实现中，为了限制处理量，持续搜索程序将只处理搜索结果中的前 X 页或前 X 个网页和文件。

在以上的所有实现当中，页/网站监视程序和持续搜索程序可以在一个搜索引擎上实现，也可以在用户的本机上，或在一个搜索引擎上和在用户的本机同时存在、并执行不同的任务分工合作。如果程序是在用户的本机上实现的，则页/网站监视程序和持续搜索程序将在需要的时候调用下载程序去下载相关搜索结果中的网页和文件。页/网站监视程序和持续搜索程序可只在需要的时候存储和计算或对网页和文件的数据作分析。根据搜索的返回结果将生成一个网页列表，持续搜索程序需要计算和存储这个列表中的网页的校验和、或数字摘要。

本发明的信息监视实现提供了一种信息监视的新方法，该方法包括：

在一个浏览应用的窗口提供一个选项，用户可使用此选项选择监视正在此窗口中浏览的 URL 内容的变化或使用此窗口进行的一个搜索的变化；

当用户选择此选项，在一段时间内检查此 URL 或此搜索的内容有无变化；

如此 URL 或此搜索的内容有变化，把探测到的变化通知给用户。

上述方法还可以进一步包括下列一项或多项：

提供一个选项让用户规定监视的时间段或频率；检查此 URL 或此搜索的内容有无变化是在用户的计算机上进行；检查此 URL 或此搜索的内容有无变化包括在一段时间内以某个频率重复访问此 URL 并检查其内容的变化，或在一段时间内以某个频率重复进行此搜索并检查搜索结果内容的变化；检查此 URL 或此搜索的内容有无变化包括计算并储存此 URL 或此搜索在一个时间（称为甲时间）的内容的计算校验和或数字摘要，将甲时间储存的计算校验和或数字摘要和在甲时间后的一个时间由此 URL 或此搜索的内容计算的计算校验和或数字摘要进行比较。

## 7. 分离元搜索

在一个实现中，为了使用户进行的搜索信息私有化，这个发明中的分割搜索程序将被安装到用户的本地计算机上。这个分割搜索程序将字串分解成两个或多个子集，并将每个子集分配给单独的一个搜索引擎。由于每个搜索引擎只使用搜索关键字/词中的一个子集进行搜索，所以这样搜索的结果将以完整关键字/词搜索所得结果的扩展集。分割搜索程序在之后获取或下载每个搜索引擎上的结果，然后在本机运行一个使用在本地计算机上完整的搜索关键字/词串的查询，并将所有这些子查询结果合并起来。这相当于为所有子查询结果寻找交集。通过这样的方法，用户所作的查询信息不会同时完整地被一个搜索引擎使用，这样，就保护了用户搜索的隐私。例如，这样就可以避免搜索引擎或其它的监视者通过猜测用户有创意的意图来监视用户的搜索行为。

在一个实现中，用户界面程序在工具栏中提供一个“分割搜索”按钮或在“选项”菜单中提供一个“将关键字/词串分割交多引擎处理”的选项。用户可以通过选择点击相关的按钮和复选框来选择这样的属性设置。分割程序就将随机地将关键字/词分割并交由某一些搜索引擎处理。在另一种实现中，用户界面程序允许用户去决定关键字/词将要被分割成多少份，由哪些搜索引擎去执行，或选择哪一部分的关键字/词由哪个搜索引擎去执行。

## 8. 系统

在一种实现中，本发明的程序模块化以获得语言独立性的最大化，并提供接入不同的语言的清楚的届扩和语言模块插件。独立于具体语言的模块构成核心系统。此核心系统和语言适应模块，指定语言甲的模块、和指定语言甲的知识库相联接就可达到本发明对指定语言甲的实现和界面，例如，制定语言加可以是英语，法语，中文等。

在一个实现当中，有一个广告模块将搜索关键字/词和用户选择的概念送至选择的服务器甲。这个模块将接受服务器甲的指令，并且令符合服务器返回的相关条件的网页排序提前，并接收从该服务器上传来的广告信息，将其显示在服务器甲在浏览窗口指定的地方。

图 13 是此项发明的部分实现在进行网页搜索时的流程图。图 13 中的“搜索后分析”项显示了重要概念的提取，排序，选择和排列以及其它的过滤细节。这些重要的概念将与其它的过滤细节一起完成过滤，并由 CPM 图整合显示。前面我们曾经讨论过，虚线箭头所指的两个任务可以在搜索引擎服务器上或是在用户本机上执行，也可以将任务分割交由不同主体执行。

虽然前文对本发明的一些优先的实现的陈述已经显示、描述、或举例说明了本发明的基本的创新特征或原理，但是读者应该理解那些对相关技术领域知识的人可以在不离开本发明的精神的情况下，对前面所描述的方法、元素、模块、器件的细节以及他们的应用作出各种不同的省略、替换或改变。因此，本发明的范围不应该被前文的描述所限制。相反地，本发明的原则可适用于在一个很大范围的方法、系统和器件，以取得前文描述的利益或好处，并可取得其他的利益或好处或满足其它的目的。因此，本发明的范围应该被本发明的权利要求定义。

100

工具栏	启动硬盘搜索	在新窗口中显示硬盘搜索结果	仅搜索硬盘	选项	帮助		
<h2>智能搜索引擎</h2>							
<p>我要寻找和以下选中相关的信息 (全部选中 <input checked="" type="checkbox"/> ) <b>101</b></p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; vertical-align: top;"> <input type="radio"/> 销售商, 服务商, 产品评论  <input type="radio"/> 医疗健康  <input type="radio"/> 商业经济  <input type="radio"/> 金融投资  <input type="radio"/> 学习, 研究  <input type="radio"/> 市场研究         </td> <td style="width: 50%; vertical-align: top;"> <input type="radio"/> 政治军事  <input type="radio"/> 学校, 大学  <input type="radio"/> 旅游, 娱乐, 运动  <input type="radio"/> 研究文献, 科技, 标准  <input type="radio"/> 个人, 团组, 社会网络  <input type="radio"/> 新闻         </td> </tr> </table>						<input type="radio"/> 销售商, 服务商, 产品评论 <input type="radio"/> 医疗健康 <input type="radio"/> 商业经济 <input type="radio"/> 金融投资 <input type="radio"/> 学习, 研究 <input type="radio"/> 市场研究	<input type="radio"/> 政治军事 <input type="radio"/> 学校, 大学 <input type="radio"/> 旅游, 娱乐, 运动 <input type="radio"/> 研究文献, 科技, 标准 <input type="radio"/> 个人, 团组, 社会网络 <input type="radio"/> 新闻
<input type="radio"/> 销售商, 服务商, 产品评论 <input type="radio"/> 医疗健康 <input type="radio"/> 商业经济 <input type="radio"/> 金融投资 <input type="radio"/> 学习, 研究 <input type="radio"/> 市场研究	<input type="radio"/> 政治军事 <input type="radio"/> 学校, 大学 <input type="radio"/> 旅游, 娱乐, 运动 <input type="radio"/> 研究文献, 科技, 标准 <input type="radio"/> 个人, 团组, 社会网络 <input type="radio"/> 新闻						
<p>用您的语言简单描述您要找的东西: <input style="width: 80%; height: 40px; margin-bottom: 5px;" type="text"/> <b>104</b> <input style="width: 15%; height: 25px; border: 1px solid black; background-color: #f0f0f0; float: right; margin-bottom: 5px;" type="button" value="用关键词搜索"/></p>							
<p><b>120</b> <b>102</b></p> <p style="text-align: center;">↓</p> <p>使用下面这个文件的内容描述您想搜索什么 <input style="width: 40%; height: 25px; margin-right: 10px;" type="file"/> <b>122</b> <input style="width: 15%; height: 25px; border: 1px solid black; background-color: #f0f0f0; float: right; margin-right: 10px;" type="button" value="搜索"/></p>							
<p><b>更多选项:</b></p> <p>查找更新时间不超过 <input type="radio"/> 任意时间 <input type="radio"/> 一周 <input type="radio"/> 一月 <input type="radio"/> 三个月  <input type="radio"/> 半年 <input type="radio"/> 一年 <input type="radio"/> 三年 <b>108</b></p>							
<p>选定让这个搜索存活的时间段 <input style="width: 20%; height: 25px; margin-right: 10px;" type="text"/> <b>110</b> <input style="width: 15%; height: 25px; border: 1px solid black; background-color: #f0f0f0; float: right; margin-right: 10px;" type="button" value="搜索"/></p> <p>(不选择时间段就作为一次性搜索)</p>							
<p>当发现新的资源或改动的时候, 您将收到一个桌面通知。如果您也想收到邮件通知, 请您写下您的邮箱地址 <input style="width: 40%; height: 25px; margin-right: 10px;" type="text"/> <b>112</b></p>							
<p><input type="checkbox"/> 选择这里启动概念跟踪以扩展搜索 <b>116</b>            选择深度: 跟踪概念到 <input style="width: 20px; height: 25px; margin-right: 10px;" type="text"/> 层.</p>							
<p><input type="checkbox"/> 选择这里启动链接跟踪以扩展搜索 <b>118</b>            选择深度: 跟踪链接到 <input style="width: 20px; height: 25px; margin-right: 10px;" type="text"/> 层. <input style="width: 15%; height: 25px; border: 1px solid black; background-color: #f0f0f0; float: right; margin-right: 10px;" type="button" value="搜索"/></p>							

图 1

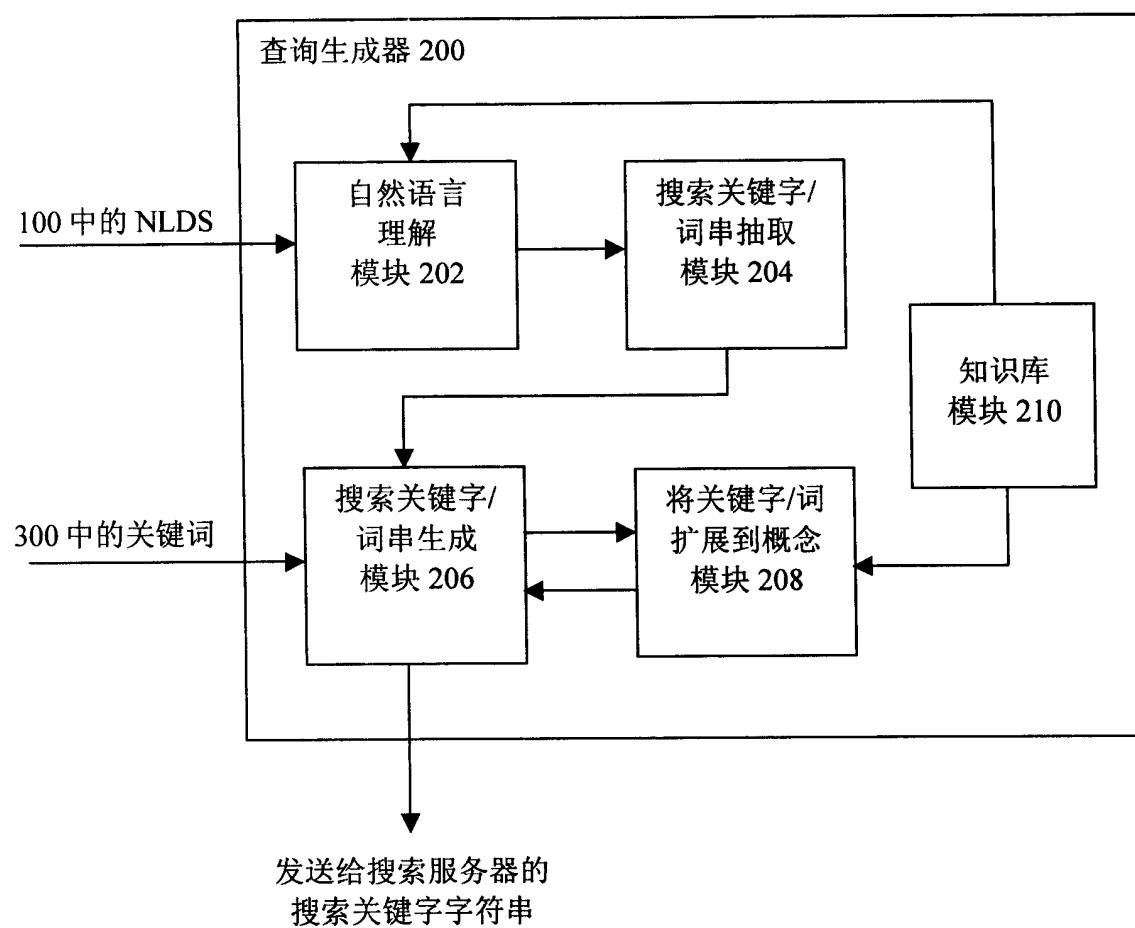


图 2

300

工具栏	启动硬盘搜索	在新窗口中显示硬盘搜索结果	仅搜索硬盘	选项	帮助
<b>智能搜索引擎</b>					
<p>我要寻找和以下选中相关的信息 (全部选中<input checked="" type="checkbox"/> 301)</p> <p><input type="radio"/> 销售商, 服务商, 产品评论      <input type="radio"/> 政治军事  <input type="radio"/> 医疗健康      <input type="radio"/> 学校, 大学  <input type="radio"/> 商业经济      <input type="radio"/> 旅游, 娱乐, 运动  <input type="radio"/> 金融投资      <input type="radio"/> 研究文献, 科技, 标准  <input type="radio"/> 学习, 研究      <input type="radio"/> 个人, 团组, 社会网络  <input type="radio"/> 市场研究      <input type="radio"/> 新闻</p> <p>用您的语言简单描述您要找的东西:</p> <div style="border: 1px solid black; width: 100%; height: 20px; margin-bottom: 5px;"></div> <p style="text-align: right;"><b>点击这里用自然语言搜索</b></p> <p><b>搜索包含以下内容的结果</b></p> <p>包含下面所有的单词和短语 (用逗号分隔, 例如, 太阳系, 火星, 有生命存在的证据) 304</p> <p><input type="radio"/> 点击这里若也要搜索同义词和同义短语</p> <div style="border: 1px solid black; width: 100%; height: 20px; margin-bottom: 5px;"></div> <p>303 并且 精确包含下面的词或短语(用逗号分隔, 例如, 油价上涨)</p> <div style="border: 1px solid black; width: 100%; height: 20px; margin-bottom: 5px;"></div> <p>305 并且 可能包含下面词或短语</p> <div style="border: 1px solid black; width: 100%; height: 20px; margin-bottom: 5px;"></div> <p>306 <input type="radio"/> 点击这里若也要搜索同义词和同义短语,</p> <div style="border: 1px solid black; width: 100%; height: 20px; margin-bottom: 5px;"></div> <p>308 并且 不包含下面的词或短语</p> <div style="border: 1px solid black; width: 100%; height: 20px; margin-bottom: 5px;"></div> <p>307 <input type="radio"/> 点击这里若也要排除同义词和同义短语</p> <div style="border: 1px solid black; width: 100%; height: 20px; margin-bottom: 5px;"></div> <p style="text-align: center;"><b>搜索</b></p> <p><b>更多选项:</b></p> <p>查找更新时间不超过      <input type="radio"/> 任意时间      <input type="radio"/> 一周      <input type="radio"/> 一月      <input type="radio"/> 三个月  <input type="radio"/> 半年      <input type="radio"/> 一年      <input type="radio"/> 三年 310</p> <p>选定让这个搜索存活的时间段      今天 或 年月日 到 年月日  (不选择时间段就作为一次性搜索) 312</p> <p>当发现新的资源或改动的时候, 您将收到一个桌面通知。如果您也想收到邮件通知, 请您写下您的邮箱地址 314</p> <div style="border: 1px solid black; width: 100%; height: 20px; margin-bottom: 5px;"></div> <p><input type="checkbox"/> 选择这里启动概念跟踪以扩展搜索      316  选择深度: 跟踪概念到 <input type="text" value="1"/> 层.</p> <p><input type="checkbox"/> 选择这里启动链接跟踪以扩展搜索      318  选择深度: 跟踪链接到 <input type="text" value="1"/> 层.</p> <div style="text-align: right;"><b>搜索</b></div>					

图 3

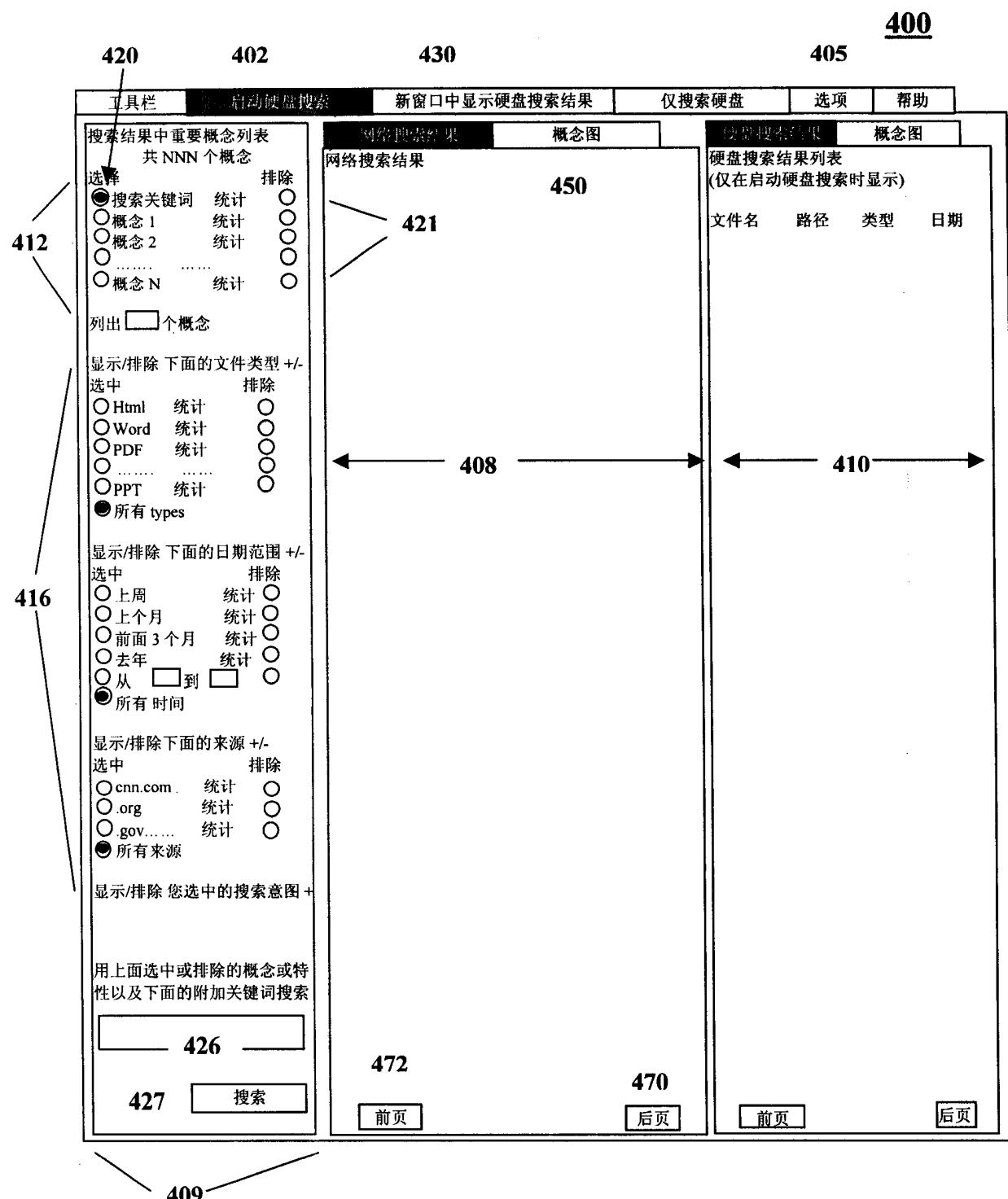


图 4

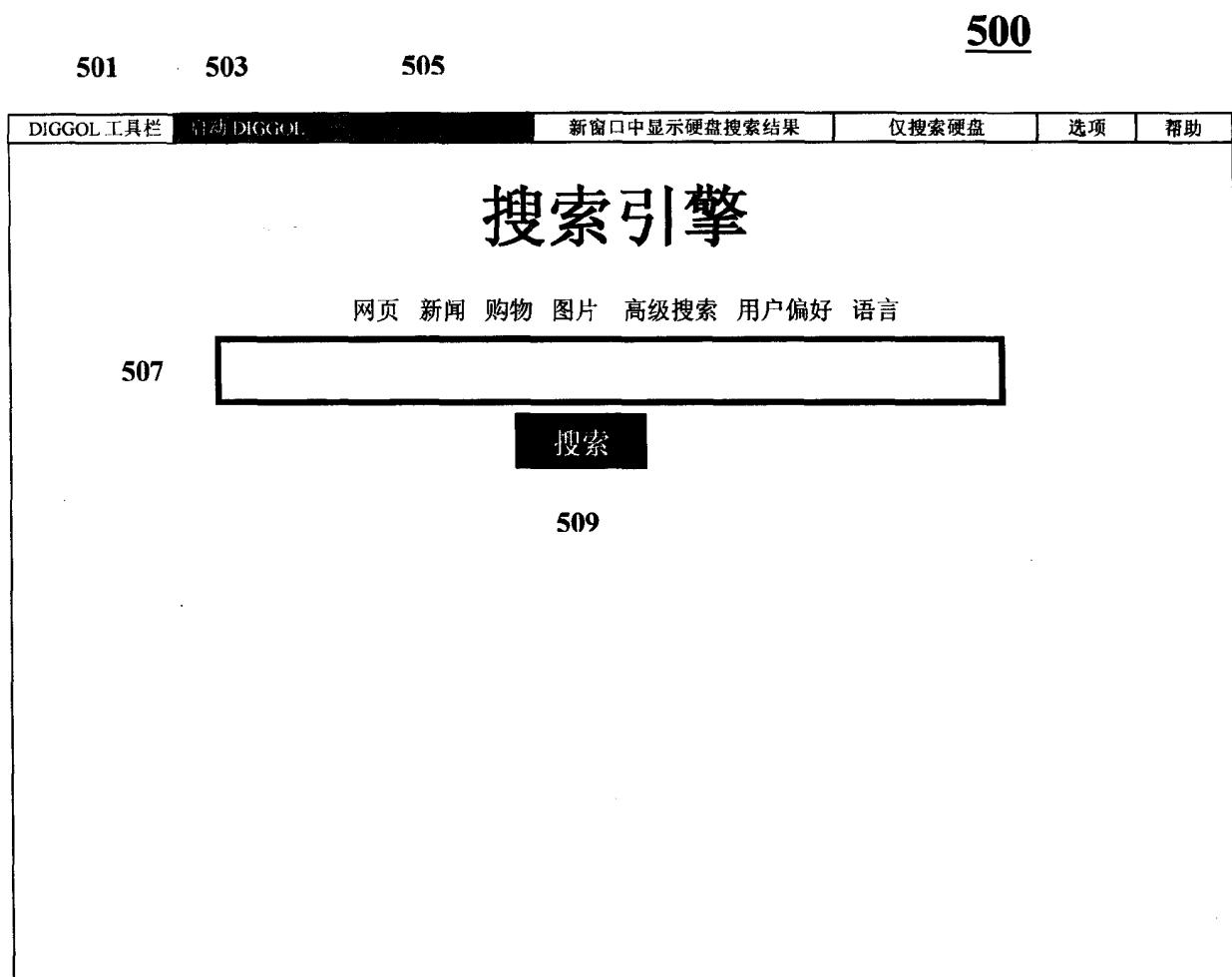


图 5

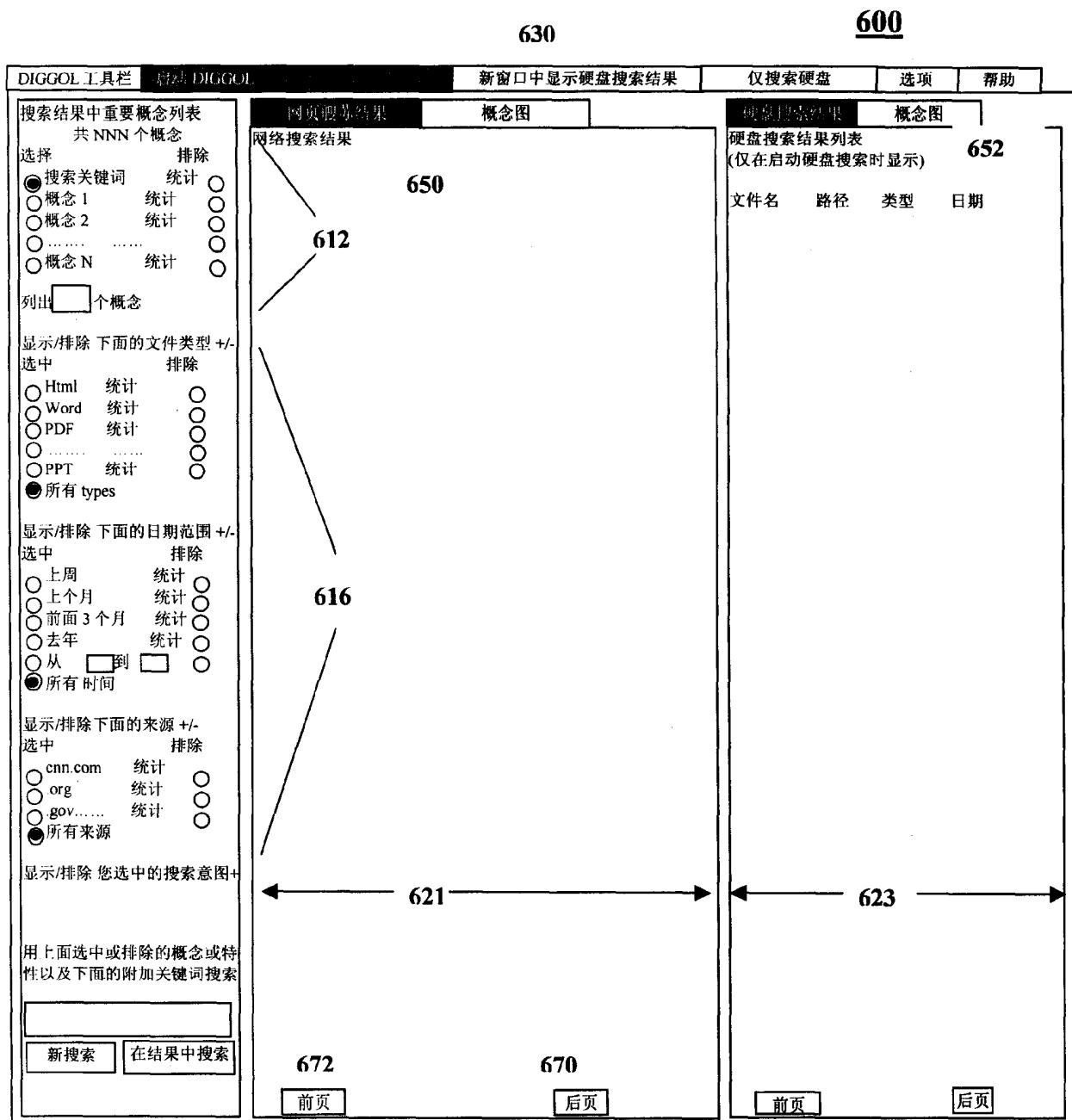


图 6

700

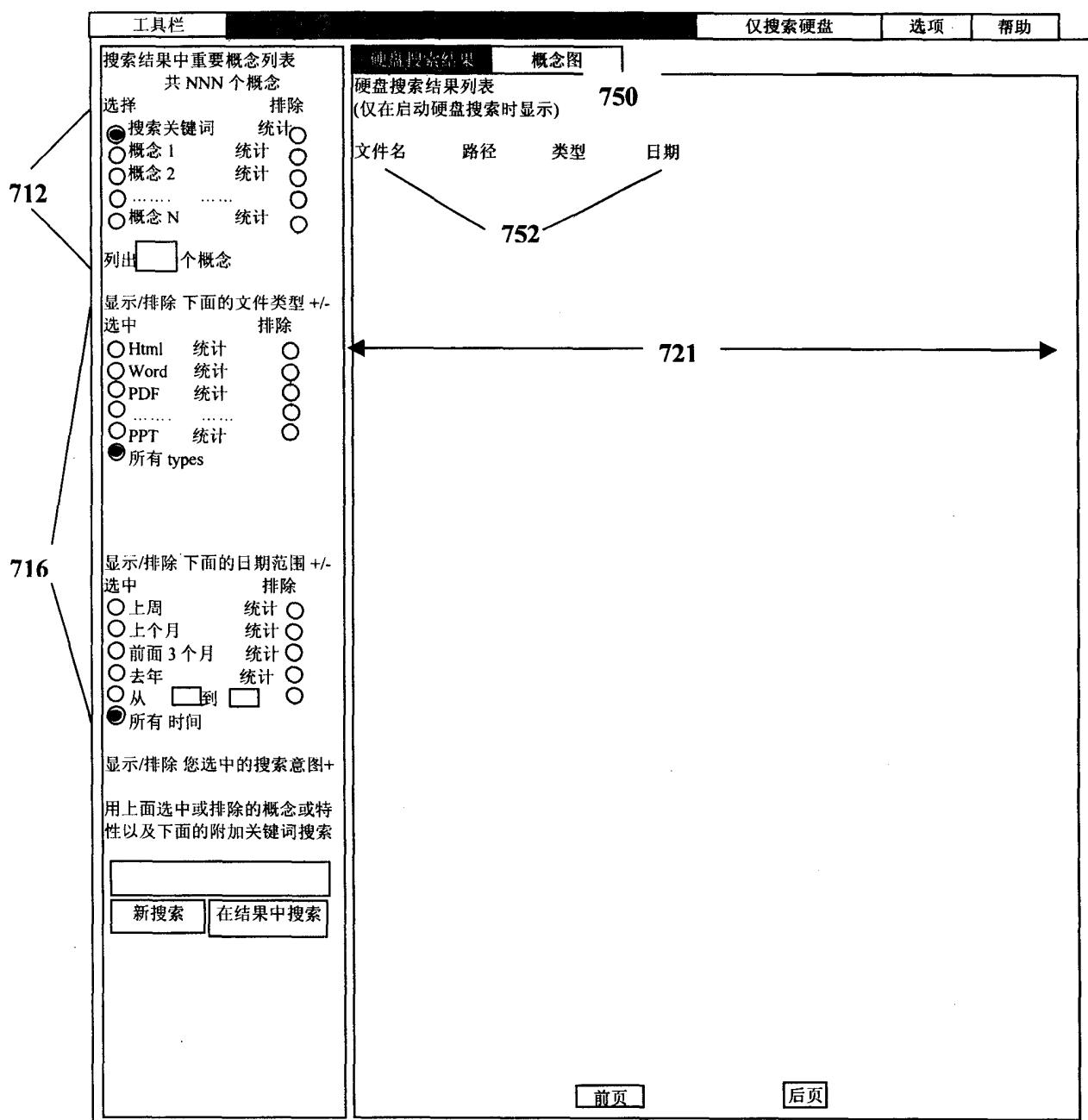


图 7

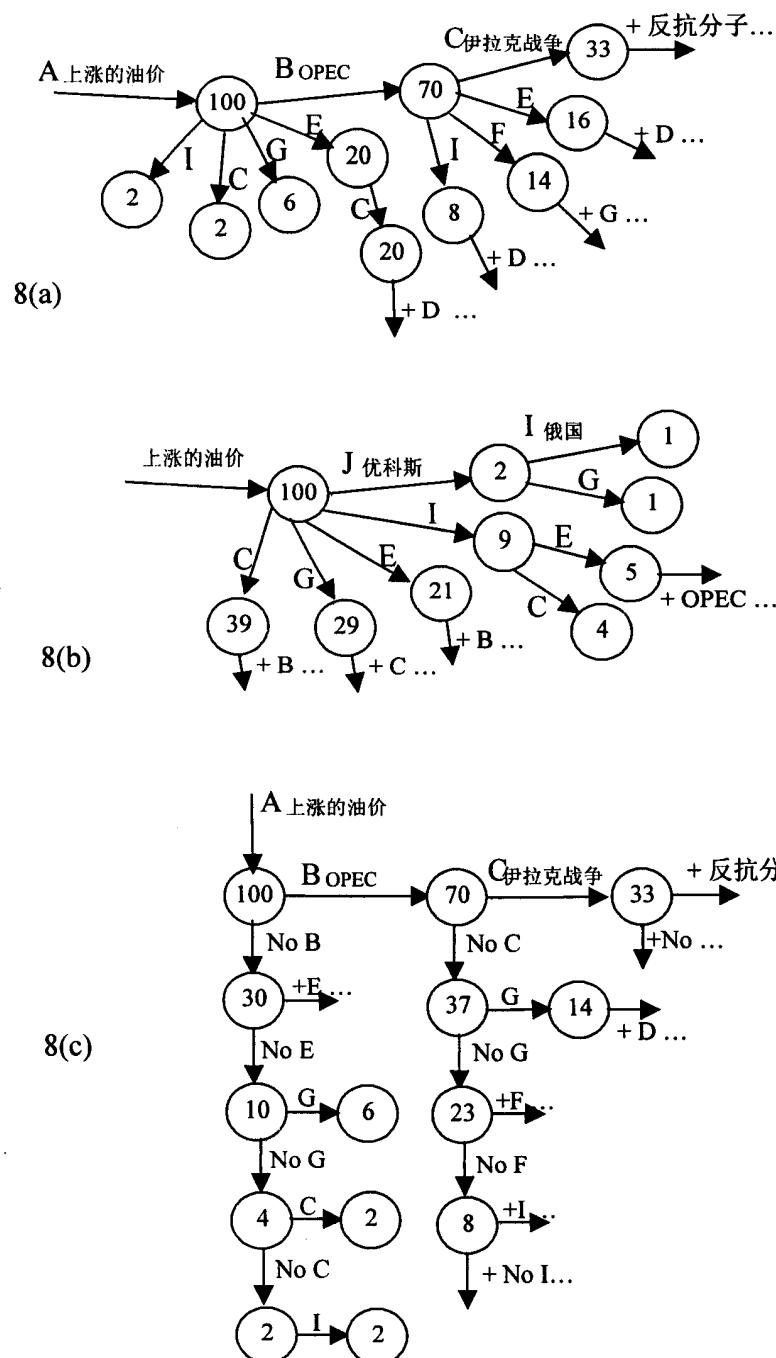


图 8

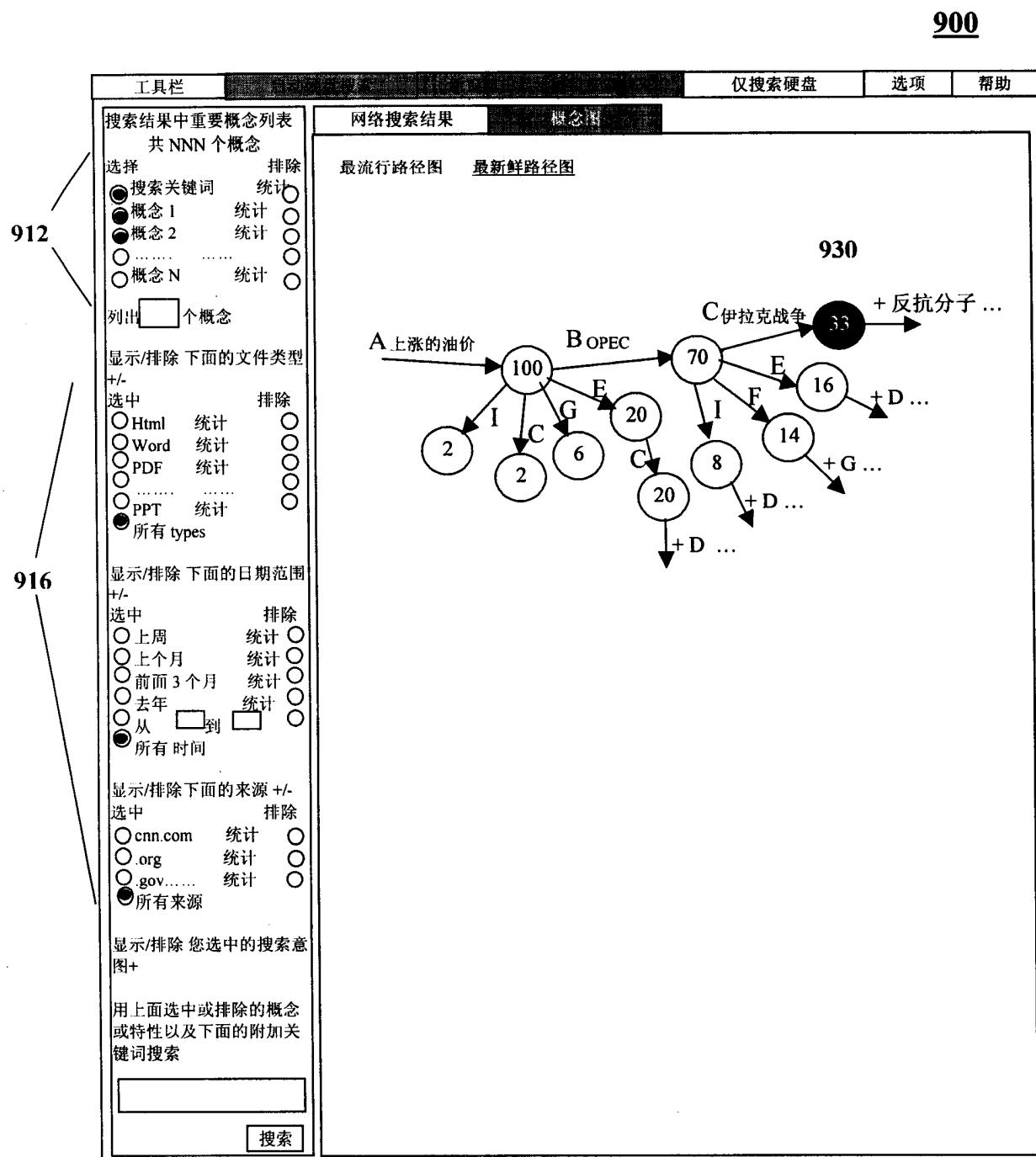


图 9

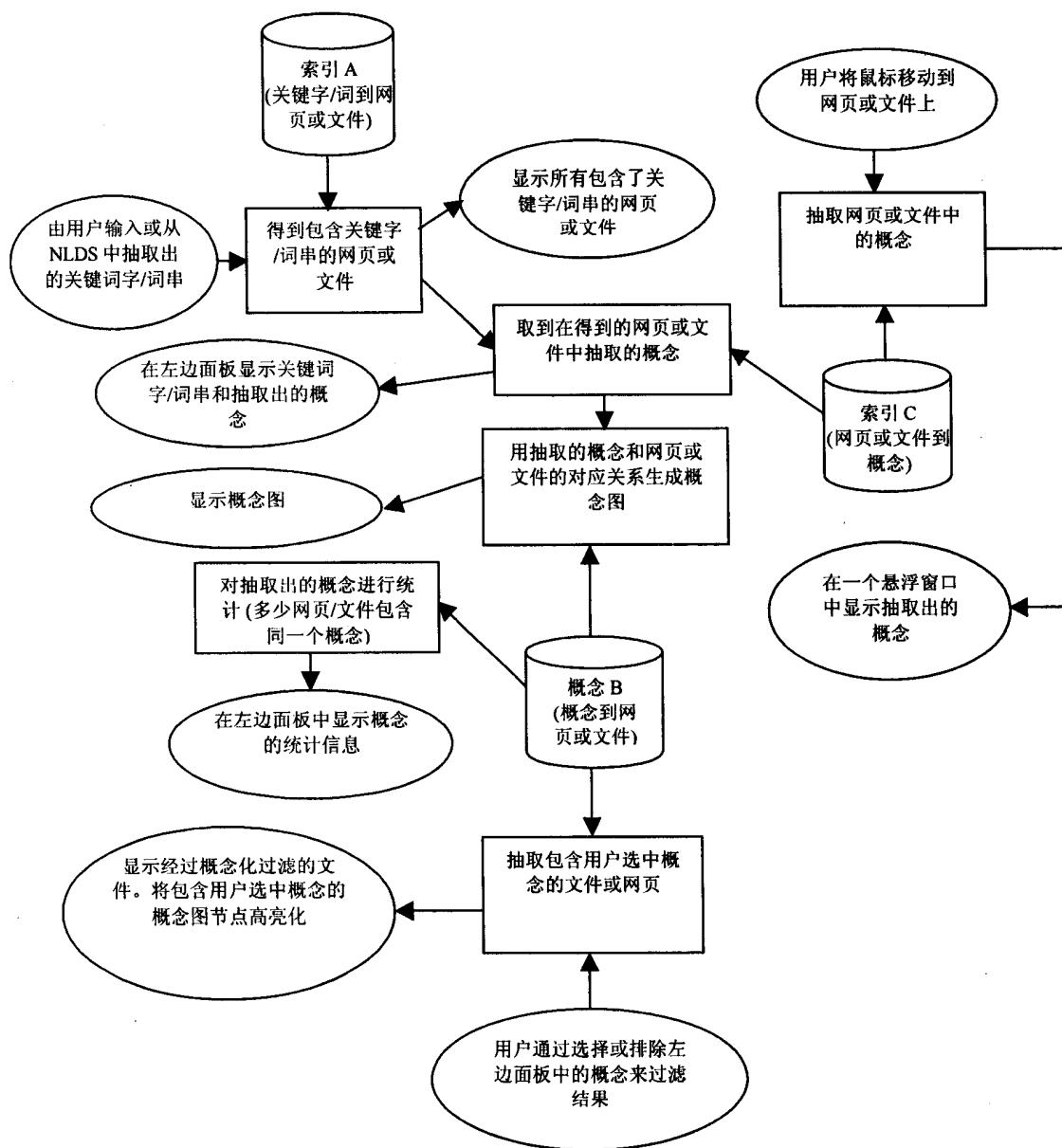


图 10

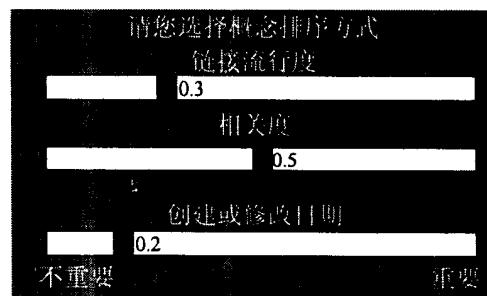


图 11

DIGGOL 工具栏 启动硬盘搜索 新窗口中显示硬盘搜索结果 选项 帮助

## 智能本地搜索

**搜索包含以下内容的结果** 点击这里用自然语言搜索

包含下面所有的单词和短语 (用逗号分隔, 例如, 太阳系, 火星, 有生命存在的证据)  
 点击这里若也要搜索同义词和同义短语

并且 精确包含下面的词或短语(用逗号分隔, 例如, 油价上涨)  
 点击这里搜索同义词和同义短语

并且 可能包含下面词或短语  
 点击这里搜索同义词和同义短语

并且 不包含下面的词或短语  
 点击这里排除同义词和同义短语

**更多选项:** 搜索

查找更新时间不超过  任意时间  一周  一月  三个月  
 半年  一年  三年

选择这里启动概念跟踪以扩展搜索 1216  
 选择深度: 跟踪概念到 1 层.

选择这里启动链接跟踪以扩展搜索 1218 搜索  
 选择深度: 跟踪链接到 1 层.

图 12

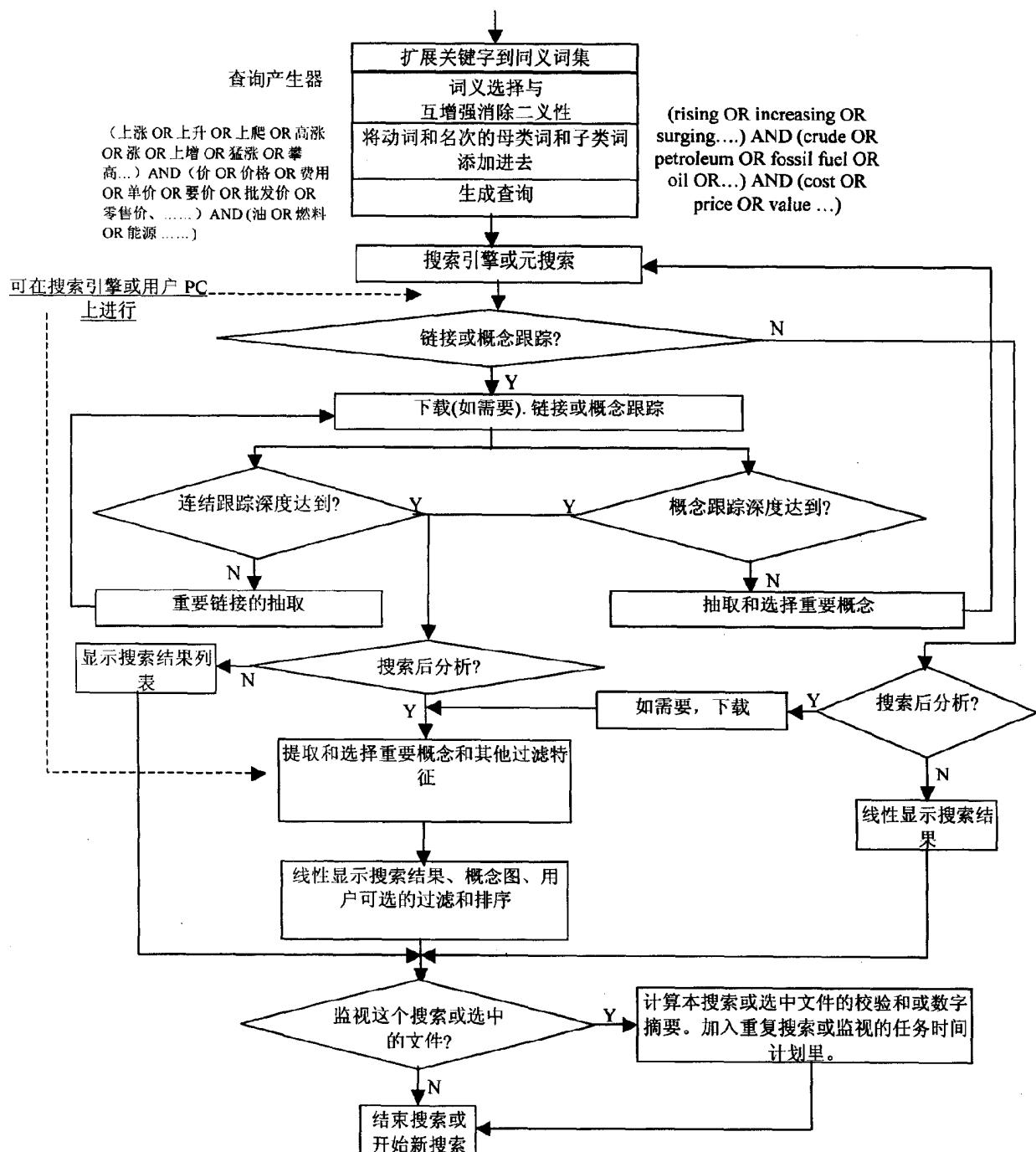


图 13