US 20100287152A1

(19) United States

(12) Patent Application Publication

Hauser

(10) Pub. No.: US 2010/0287152 A1

(43) Pub. Date: Nov. 11, 2010

(54) SYSTEM, METHOD AND COMPUTER READABLE MEDIUM FOR WEB CRAWLING

(75) Inventor: **Robert R. Hauser**, Frisco, TX (US)

Correspondence Address:
RG & ASSOCIATES
1103 TWIN CREEKS, STE. 120
ALLEN, TX 75013 (US)

(73) Assignees: **Paul A. Lipari**, Frisco, TX (US); **SUBOTI, LLC**, Frisco, TX (US)

(21) Appl. No.: 12/435,774

(22) Filed: May 5, 2009

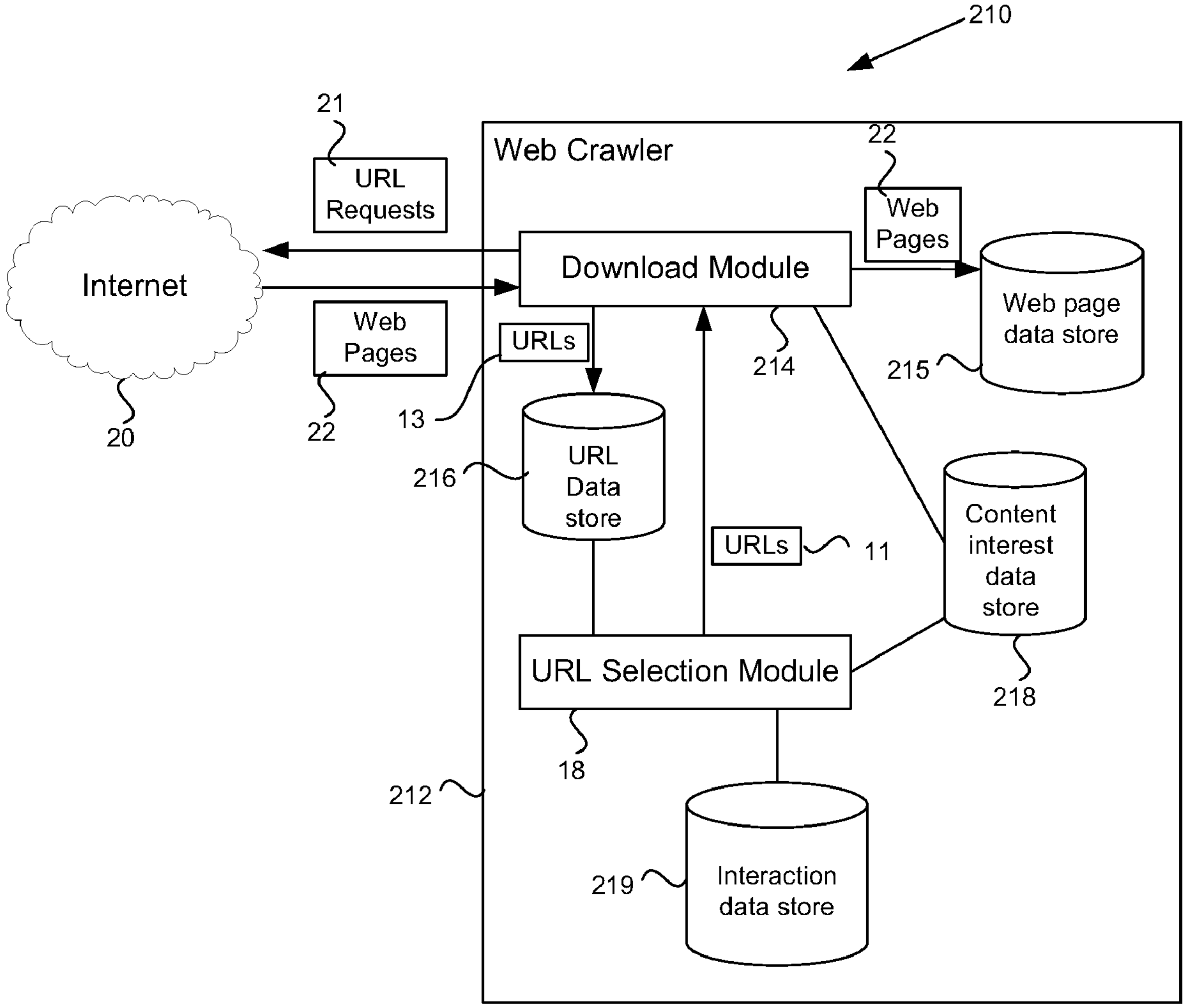
Publication Classification

(51) Int. Cl.
G06F 17/30 (2006.01)

(52) U.S. Cl. **707/707**; 707/706; 707/726

(57) ABSTRACT

In a web crawler, a URL selection module selects URLs for pages to be downloaded. The URL selection module accesses an interaction data store that stores interaction data for web pages, including interaction data that indicates human interactions with the pages. To reduce the effects of link farms, the URL selection module filters the URLs to select only those URLs that have human interaction histories and provides the selected URLs to a download module for web page downloading.



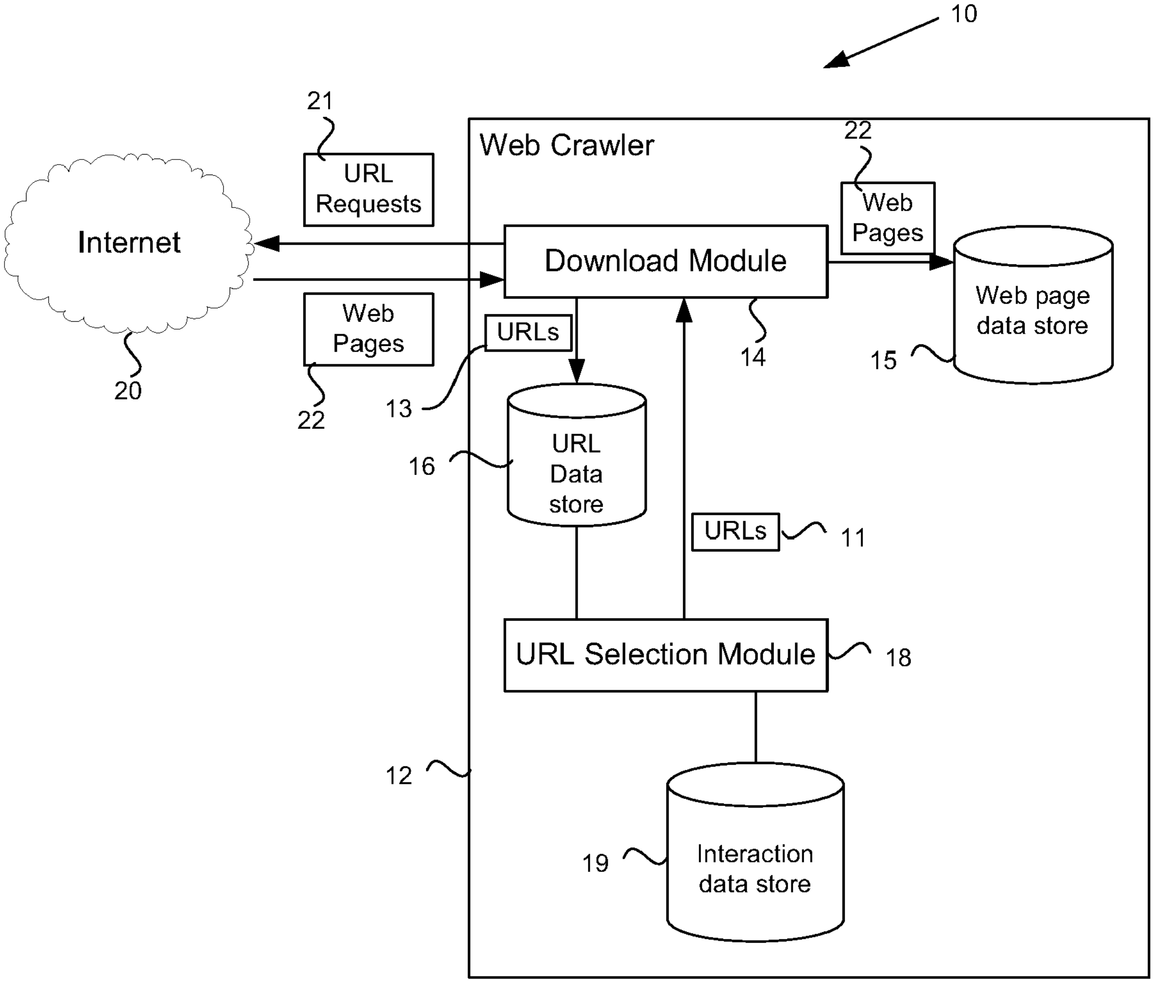

10
21
URL Requests
Internet
20
Web Pages
22
Web Crawler
22
Web Pages
Download Module
14
Web page data store
15
URLs
13
URL Data store
16
URLs
11
URL Selection Module
18
12
Interaction data store
19

Figure 1

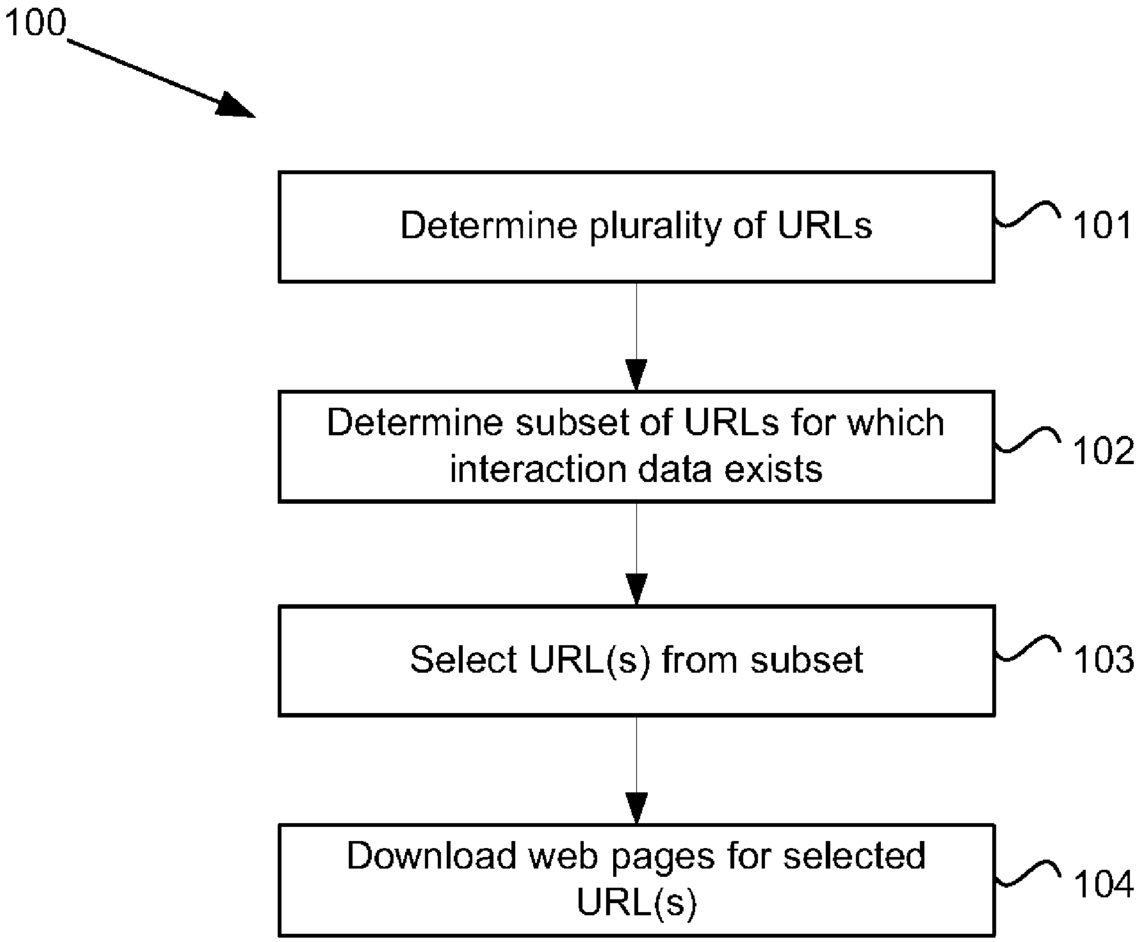

100
Determine plurality of URLs
101
Determine subset of URLs for which interaction data exists
102
Select URL(s) from subset
103
Download web pages for selected URL(s)
104

Figure 2

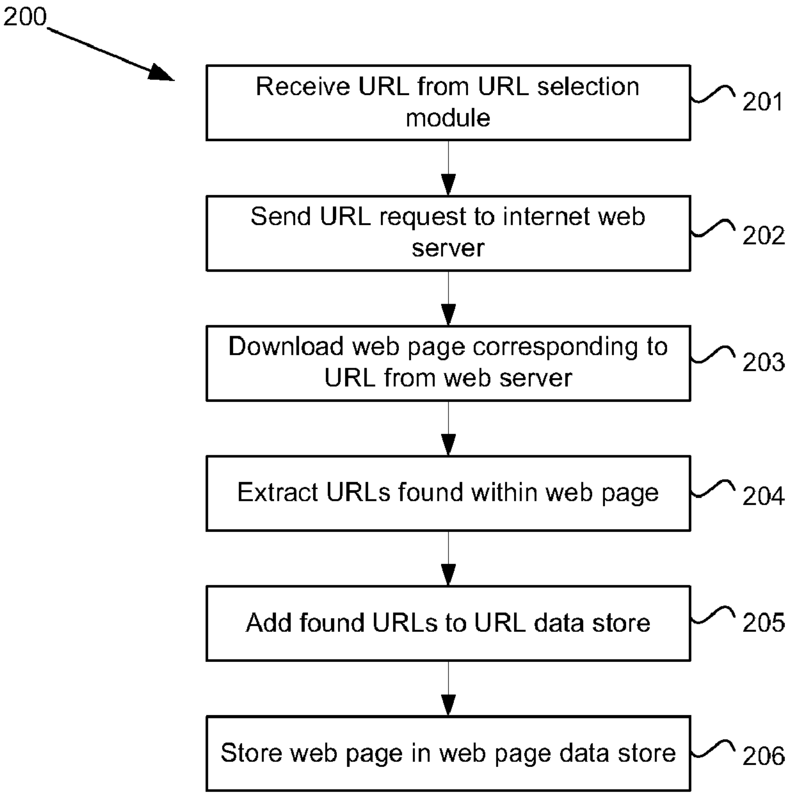

200
Receive URL from URL selection module
201
Send URL request to internet web server
202
Download web page corresponding to URL from web server
203
Extract URLs found within web page
204
Add found URLs to URL data store
205
Store web page in web page data store
206

Figure 3

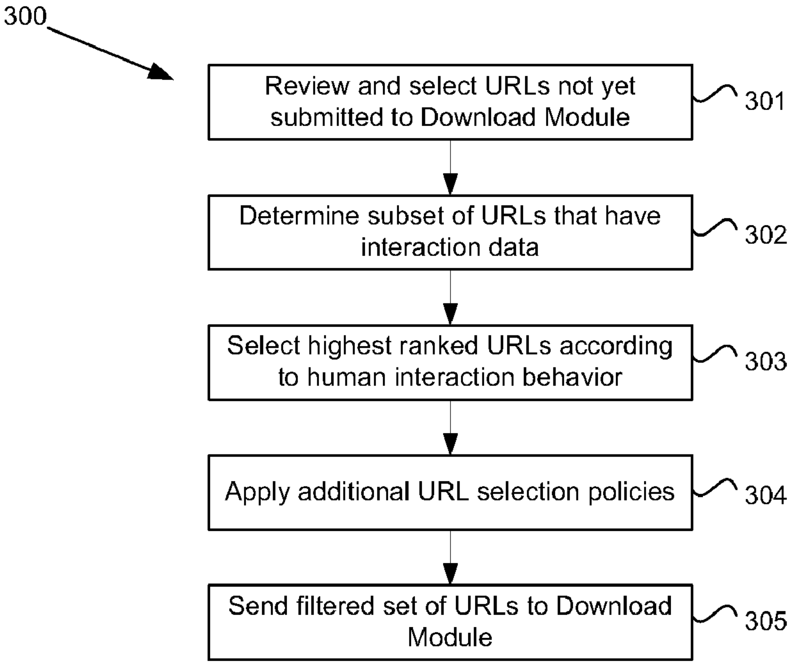

300
Review and select URLs not yet submitted to Download Module
301
Determine subset of URLs that have interaction data
302
Select highest ranked URLs according to human interaction behavior
303
Apply additional URL selection policies
304
Send filtered set of URLs to Download Module
305

Figure 4

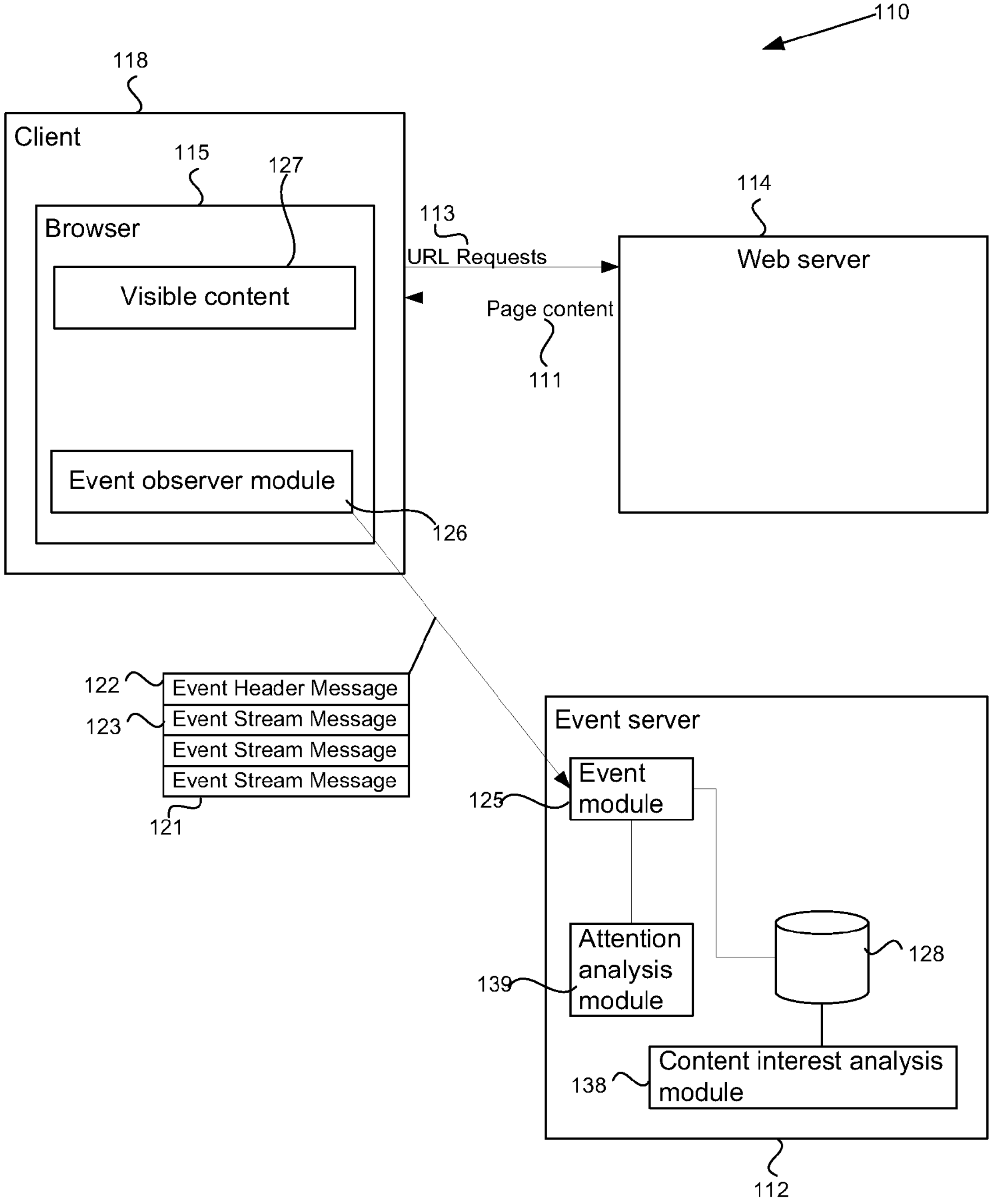

110
118
Client
115
127
Browser
Visible content
Event observer module
126
113
URL Requests
Page content
111
114
Web server
122
Event Header Message
123
Event Stream Message
Event Stream Message
Event Stream Message
121
Event server
Event module
125
Attention analysis module
139
128
Content interest analysis module
138
112

Figure 5

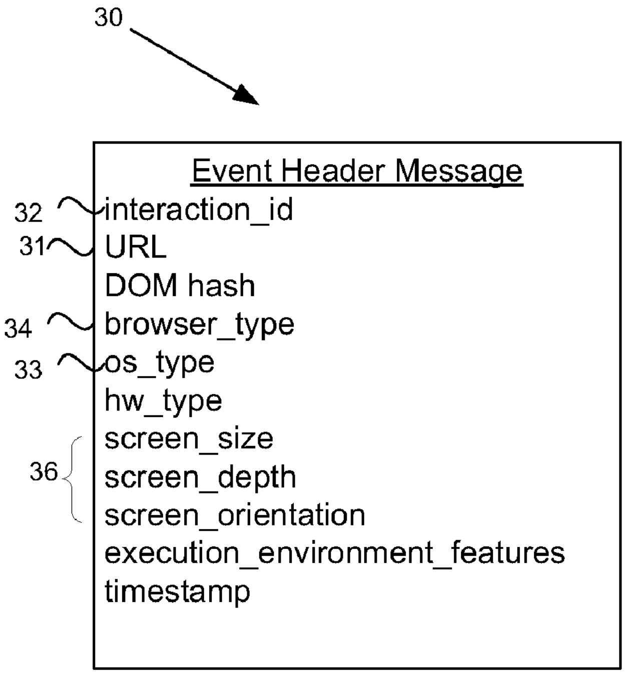


Figure 6

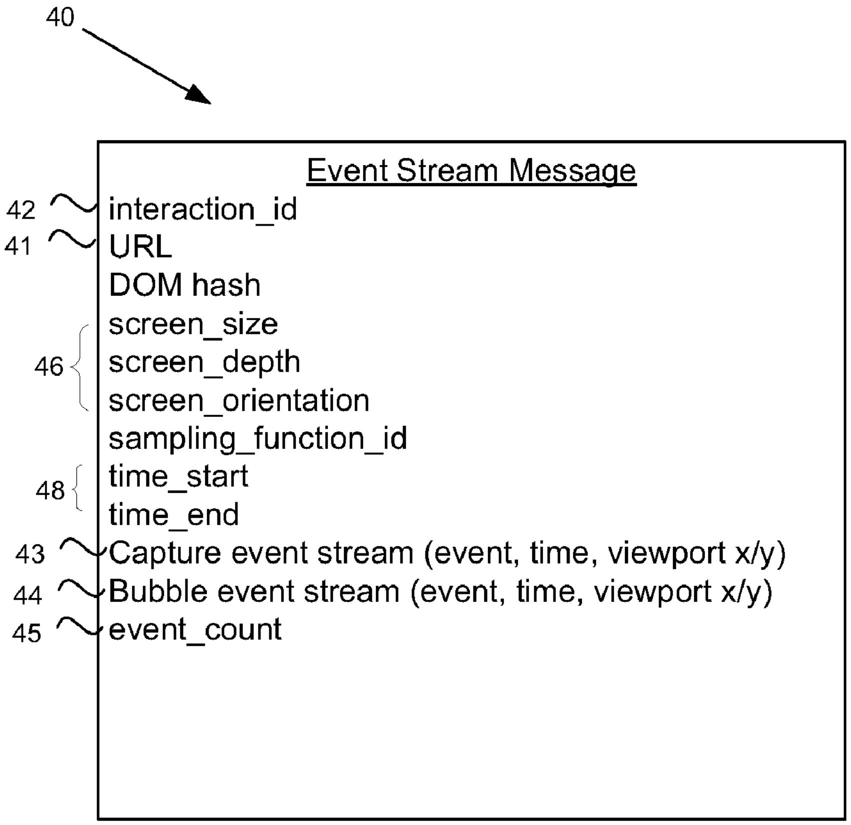


Figure 7

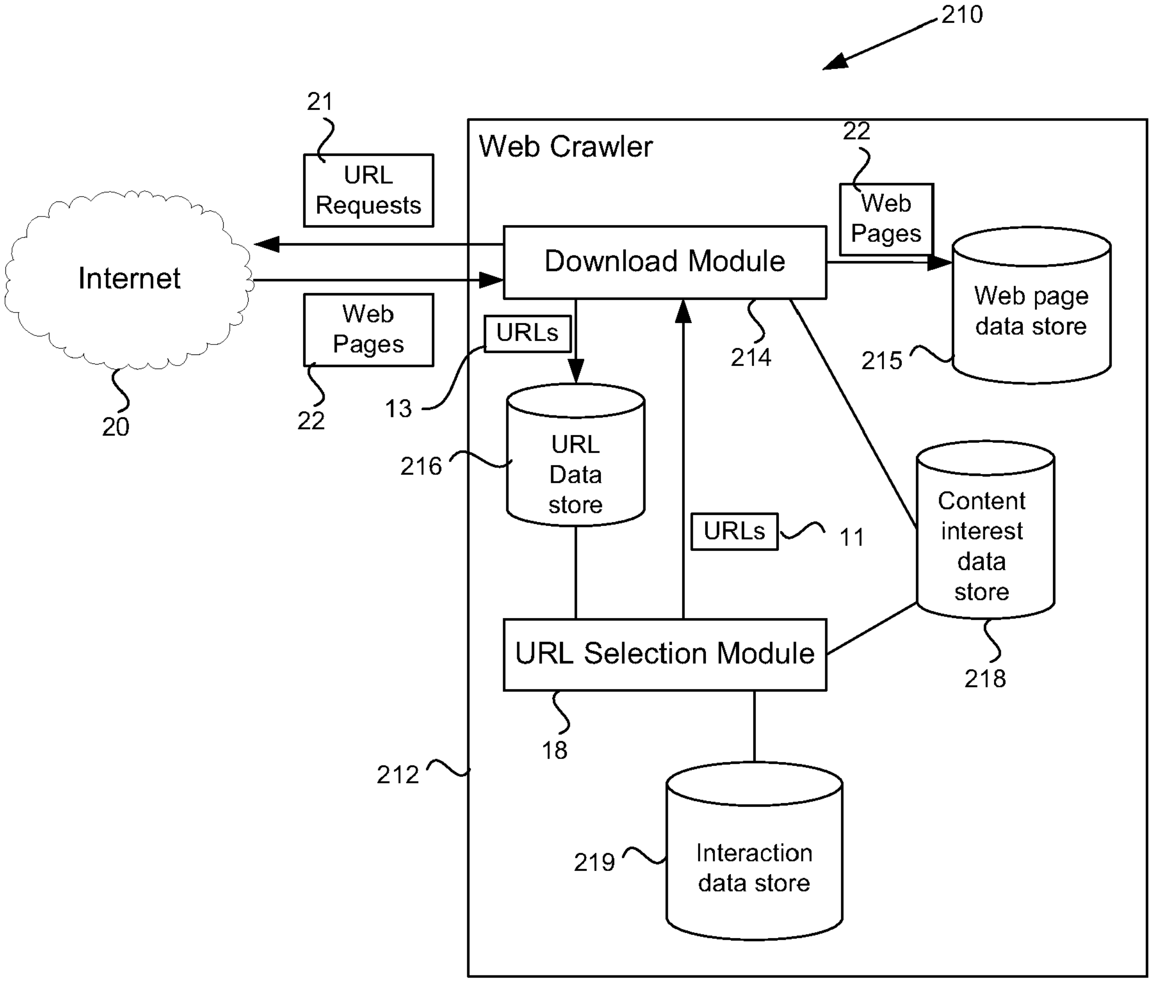

210
21
URL Requests
Internet
20
Web Pages
22
Web Crawler
22
Web Pages
Download Module
Web page data store
215
214
URLs
13
URL Data store
216
URLs
11
Content interest data store
218
URL Selection Module
18
212
219
Interaction data store

Figure 8

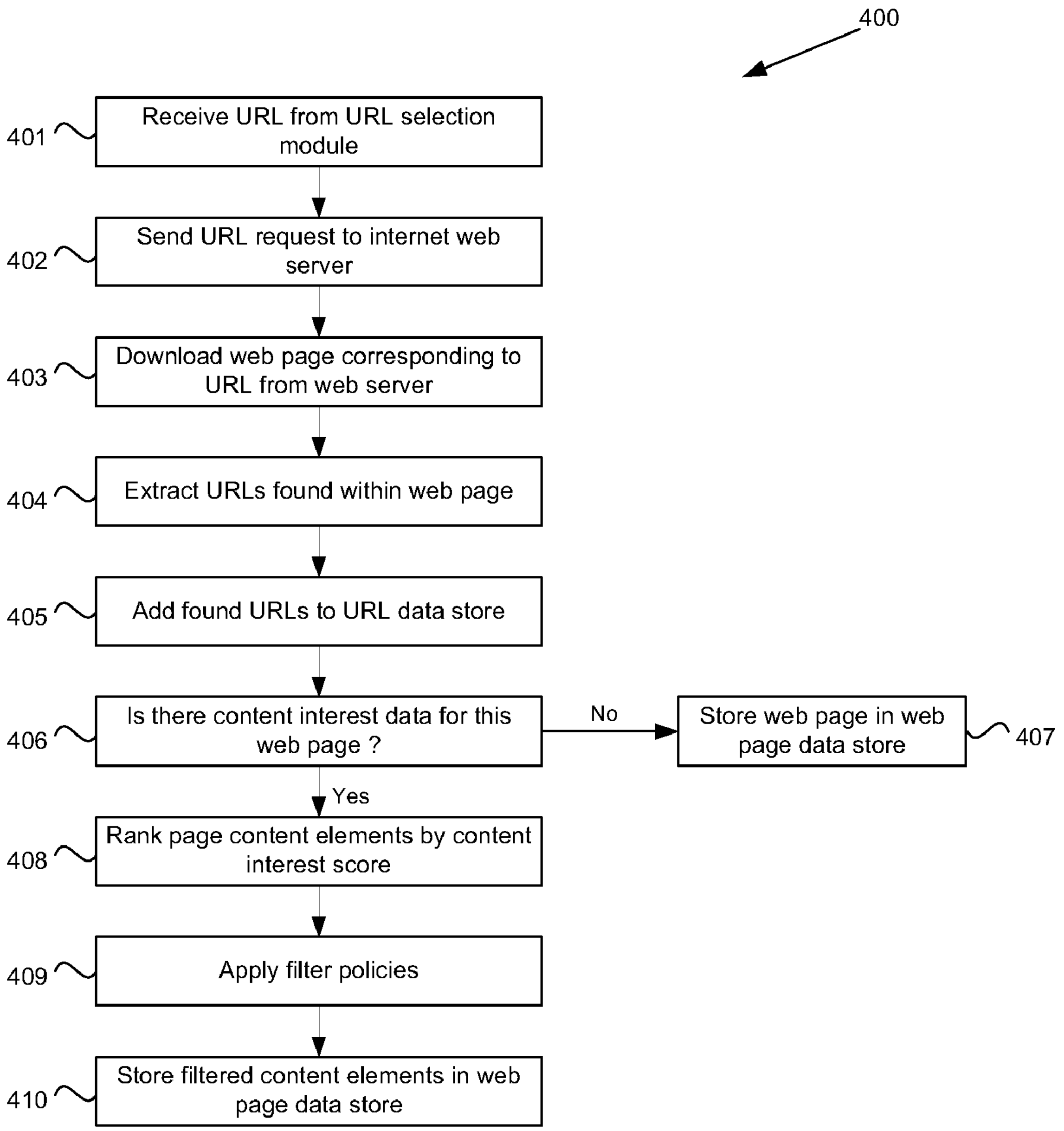

400
401
Receive URL from URL selection module
402
Send URL request to internet web server
403
Download web page corresponding to URL from web server
404
Extract URLs found within web page
405
Add found URLs to URL data store
406
Is there content interest data for this web page ?
No
Store web page in web page data store
407
Yes
408
Rank page content elements by content interest score
409
Apply filter policies
410
Store filtered content elements in web page data store

Figure 9

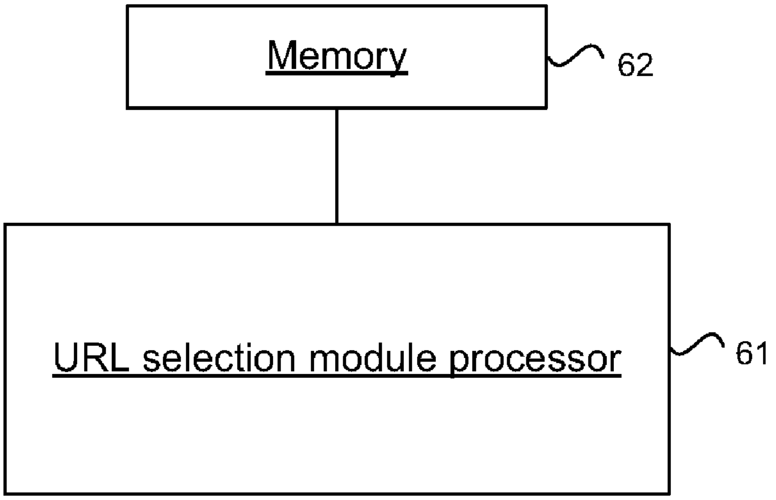


Figure 10

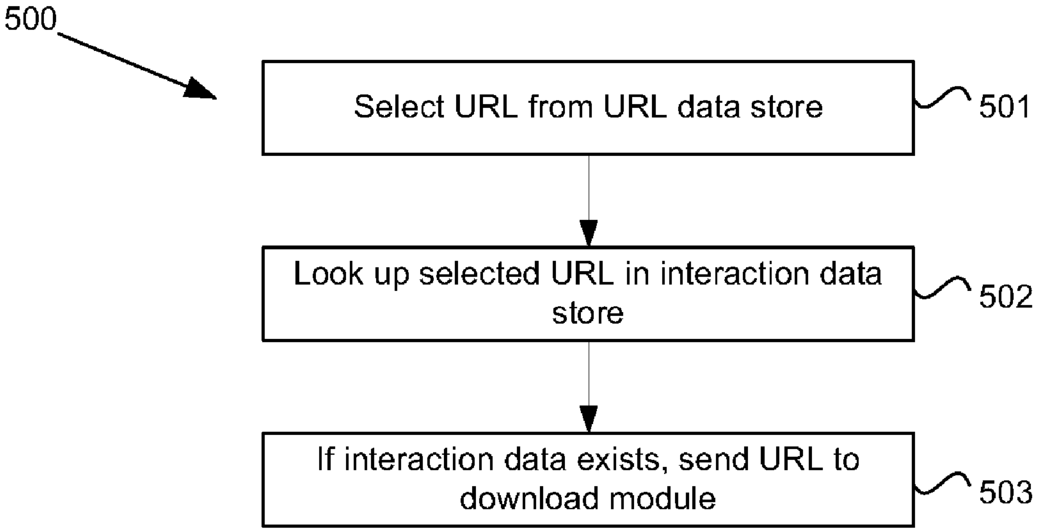

500
Select URL from URL data store
501
Look up selected URL in interaction data store
502
If interaction data exists, send URL to download module
503

Figure 11

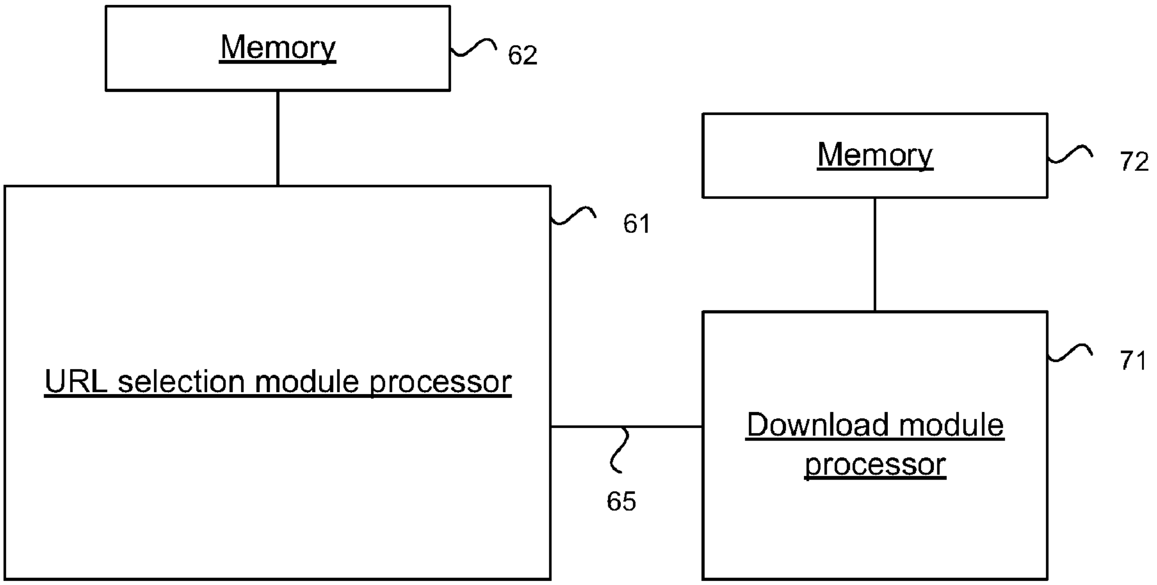

Memory
62
Memory
72
61
URL selection module processor
71
Download module processor
65

Figure 12

SYSTEM, METHOD AND COMPUTER READABLE MEDIUM FOR WEB CRAWLING

FIELD OF THE INVENTION

[0001] This invention relates to a system, method and computer medium for crawling the web to find relevant internet content.

BACKGROUND OF THE INVENTION

[0002] In internet technology, web crawlers are used to find new web pages by collecting and following URLs (Uniform Resource Locators). By following an URL and downloading the corresponding web page the links within that web page can be added to the web crawler's URL collection. The web pages are stored for indexing and ranking by internet search engines. Internet search engines use web page ranking algorithms that relate the links within a web page to the relevance of the web page.

[0003] The use of link popularity algorithms to rank web pages has lead to the problem of "link farms". In order to manipulate a web page's ranking, a large sub-web of interlinked web pages is created and linked to a web page so that the page receives a high search engine ranking. In addition to distortion of web page rankings, a problem with link farms is that a web crawler spends a lot of resources following links and collecting web pages for eventual indexing into a search engine, even though many of these pages are created only for page ranking and are not otherwise used by, nor useful for humans.

[0004] What is required is a system, method and computer readable medium that provides enhanced web crawling.

SUMMARY OF THE INVENTION

[0005] In one aspect of the disclosure, there is provided a method for web crawling comprising determining a plurality of Uniform Resource Locators (URL)s, determining a subset of the plurality of URLs that have associated interaction data, selecting at least one URL of the subset, and downloading a web page corresponding to the at least one selected URL.

[0006] In one aspect of the disclosure, there is provided a web crawler comprising at least one Uniform Resource Locator (URL) data store that stores a plurality of URLs, at least one interaction data store that stores interaction data for a plurality of web pages, at least one download module that downloads web page content corresponding to a URL, and at least one URL selection module in communication with the at least one URL data store and the at least one interaction data store. The interaction data indicates an interaction between a human and a web page corresponding to a URL. The at least one URL selection module selects at least one URL from the at least one URL data store that has interaction data in the at least one interaction data store. The at least one URL selection module provides the at least one selected URL to the at least one download module.

[0007] In one aspect of the disclosure, there is provided a computer-readable medium comprising computer-executable instructions for execution by a processor, that, when executed, cause the processor to select a Uniform Resource Locator (URL) from a URL data store, look up the selected URL in an interaction data store to determine if interaction data exists for the selected URL in the interaction data store, and if interaction data exists for the selected URL, provide the selected URL to a download module.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] Reference will now be made, by way of example only, to specific embodiments and to the accompanying drawings in which:

[0009] FIG. **1** illustrates a web crawler in accordance with an embodiment of the disclosure;

[0010] FIG. **2** illustrates a method for web crawling;

[0011] FIG. **3** illustrates a method for downloading web pages in the web crawler of FIG. **1**;

[0012] FIG. **4** illustrates a method of a URL selection module;

[0013] FIG. **5** illustrates a system for recording and analyzing event data;

[0014] FIG. **6** illustrates an event header message;

[0015] FIG. **7** illustrates an event stream message;

[0016] FIG. **8** illustrates an alternative web crawler embodiment;

[0017] FIG. **9** illustrates an alternative method of the URL selection module;

[0018] FIG. **10** illustrates a processor and memory of a URL selection module;

[0019] FIG. **11** illustrates an instruction set that may be executed on the processor and memory of FIG. **10**; and

[0020] FIG. **12** illustrates the processor and memory of FIG. **10** in association with a processor and memory of an download module.

DETAILED DESCRIPTION OF THE EMBODIMENTS

[0021] A system **10** for providing web crawling in accordance with an embodiment of the disclosure is illustrated in FIG. **1**. In the system **10**, a web crawler **12** provides URL requests **21** to the internet **20** and downloads web pages **22** corresponding to the URL requests. The web crawler **12** includes a download module **14** that downloads the web pages **22** and provides the web pages **22** to a web page data store **15**. The web crawler **12** also includes a URL data store **16**, a URL selection module **18** and an interaction data store **19**. While the components of the web crawler are shown within the web crawler **12**, a person skilled in the art will readily understand that these components may be provided in a distributed form. For example, the various data stores may be co-located with processing modules such as the download module **14** or URL selection module **18**. Alternatively, the various data stores may be located offsite with data retrieval occurring through appropriate communication links.

[0022] A web crawling method using the system **10** of FIG. **1** is illustrated in the flowchart **100** of FIG. **2**. At step **101**, the URL selection module **18** determines from the URL data store **16** a plurality of URLs for which web page content is to be downloaded. The URLs may represent new web pages or web pages for which web page content, ranking statistics etc have not been updated for a period of time. At step **102**, the URL selection module **18** accesses the interaction data store **19** to determine which URLs have interaction data associated with them. URLs for which interaction data exist are formed into a subset. At least one URL is selected from the subset (step **103**) and provided to the download module **14** so that the web page corresponding to the URL can be downloaded (step **104**).

[0023] The download module **14** downloads web pages **22** from the internet **20** and extracts linked URLs **13** from the download pages. The operation of the download module in accordance with an embodiment of the disclosure is illustrated in the flowchart **200** of FIG. **3**. The download module receives a URL **11** to fetch from the URL selection module **18** at step **201**. The download module **14** sends a URL request **21** to an appropriate web server within the internet **20** (step **202**) and downloads the web page **22** corresponding to the URL request **21** from the internet **20** (step **203**). At step **204**, the download module **14** extracts any URLs **13** found within the web page **22**. These URLs are added to the URL data store **16** (step **205**). As is known, duplicate URLs are not stored multiple times and links extracted from web pages may need to be normalized into their respective URLs. The web page **22** is also loaded into the web page data store **15** (step **206**).

[0024] The operation of the URL selection module **18** in accordance with an embodiment of the disclosure is shown in the flowchart **300** of FIG. **4**. At step **301**, the URL selection module **18** reviews the URL data store **16** for URLs that have not yet been submitted to the download module **14**. At step **302**, the URL selection module **18** accesses the interaction data store **19** to determine the subset of the selected URLs that have interaction data records. The URL selection module then selects the highest ranked URLs according to human interaction behavior (step **303**). In general, there are at least two types of human interaction behaviour that may be considered. There is a first type of human interaction in selecting a link (URL) that corresponds to a web page, even though the selected linked URL/web page may not have been downloaded and may not have any interaction data. This “human interaction” uses the analysis of the source element ranking and attention shift that happens in the various web pages that link to the URL/web-page-of-interest. Another case of “human interaction” utilizes the content of interest ranking within the URL/web-page-of-interest. This helps rank the importance of this URL in the link-graph. That is, how much, if any, content on the web-page/URL gets human attention time (independent of any links within the page). In ranking the URLs, a specific human behavior is an out-click of the URL on web pages that have the URL as a link. One ranking measure for human out-clicks may be the ratio of human out-clicks to total interaction exits per web page that display the URL as a link. Another ranking could rely on the attention ranking of the content area containing the URL/link (the location of the URL/link within human attentive areas of the web pages containing the URL as a link). Within the web page corresponding to a URL, the most preferred URLs have corresponding web pages that have highly ranked content areas, ranked by the amount of human attention, e.g time, that those content areas receive. Other ranking and selection policies may be applied to filter the URLs at step **304** and a filtered set of the URLs is sent to the download module at step **305**, returning the cycle to step **201** of FIG. **3**.

[0025] The interaction data in the interaction data store **19** may be derived from interactions between users and the web page at client browsers, for example as described in any of the Applicant’s co-pending patent applications Attorney Docket Nos. HAUSER001, HAUSER002, HAUSER006, HAUSER007, HAUSER007B, HAUSER008, HAUSER009, HAUSER010, the entire contents of each of which are explicitly incorporated herein by reference. In particular, event recorders provided within the web pages may record event data during these interactions and provide event streams to an event server. An example of an event data processing system is illustrated in FIG. **5**. In the system **110**, a client **118** provides web page requests **113** to a web server **114**, in response to which, the web server **114** provides page content **111** for display in a browser **115** of the client **118**. Typically, the web page **111** will include visible content **127** as well as javascript applications.

[0026] The web server **114** may be modified such that the web page content provided to the client **118** includes an event observer module **126** which may be provided as appropriate code or scripts that run in the background of the client’s browser **115**. In one embodiment, code for providing the event observer module **126** is provided to the web server **114** by a third party service, such as provided from an event server **112**, described in greater detail below.

[0027] The event observer module **126** observes events generated in a user interaction with the web page **111** at the client **118**. The event observer module **126** records events generated within the web browser **115**, such as mouse clicks, mouse moves, text entries etc., and generates event streams **121** including an event header message **122** and one or more event stream messages **123**. It will be apparent to a person skilled in the art that terms used to describe mouse movements are to be considered broadly and to encompass all such cursor manipulation devices and will include a plug-in mouse, on board mouse, touch pad, pixel pen, eye-tracker, etc.

[0028] The event observer module **126** provides the event streams **121** to the event server **112**. An example of an event header message **30** is illustrated in FIG. **6** and an example of an event stream message **40** is illustrated in FIG. **7**. The messages **30**, **40** show a number of components that can be included, though in various embodiments, not all of these components may be required and additional components may be added. Primarily, an Interaction_ID **32**, **42** uniquely identifies an interaction between the client **18** and the web server **14** and aids to identify the particular event stream **121**. The event header message **30** and the event stream message **40** may also identify the Uniform Resource Locator (URL) **31**, **41**. Fixed parameters such as the operating system **33** and browser type **34** may form part of the event header message **30**. Screen parameters **36**, **46** such as the screen size, depth and orientation may be included in either or both of the event header message **30** or the event stream message **40**. A capture event stream **43** and a bubble event stream **44** specifies the events recorded in respective event capture and bubbling phases during the web page interaction. Each event may be indicated by the event type, time and x/y location relative to the viewport. Not all web browser types support event capture, and thus the capture event stream **43** may be empty. Where required, events missing from the event bubble stream may be inferred, for example as described in the Applicant’s co-pending application Attorney Docket No. HAUSER002, referenced above. An event_count field **45** may indicate the total number of unique events observed by the event observer module including those events not included in the event stream message **40** due to a current sampling function excluding them. Timing parameters **48** may indicate the relevant period over which the event stream message **40** is current.

[0029] During an interaction with the web page **111**, a user navigates the web page **111** and may enter content where appropriate, such as in the HTML form elements. During this interaction events are generated and recorded by the event observer module **126**. Periodically, the event observer module **126** formulates an event stream message **123** preceded by an event header message **122** if one has not yet been sent. The event observer module **126** passes the event stream messages

123 to an event module **125** of the event server **112**. In the embodiment illustrated in FIG. **5**, the event stream **121** is provided directly to the event module **125**. However, the event stream **121** may also be provided indirectly, e.g. via the web server **114**.

[0030] The event server **112** processes the event stream **121** in the event module **125** or an equivalent component, to analyze the event stream data. Analyzed data may be stored with the raw event stream messages in a content data store **128**. Additional modules of the event server may include an attention analysis module **139** as described in the Applicant's co-pending application HAUSER008 reference above, and a content interest processing module **138** as described in the Applicant's co-pending application HAUSER009 referenced above. In one embodiment, the event stream data can be analyzed to determine the probability that the interaction that created the event stream at the client is a human dependent interaction, for example as described in the Applicant's co-pending patent application Attorney Docket No. HAUSER001 referenced above. In the present embodiment, the existence of any human interaction within the content areas of the web page, such as hints, lingers or clicks within the content areas, may be used to indicate the validity of a URL, and such statistics may be loaded into the interaction data store **19**. In one embodiment, the web crawler **12** may include the event server **112** such that the web crawler is self contained. In an alternative embodiment, human interaction data may be provided to the interaction data store as a third party service by an event server operator. Alternatively, the event server **112** may maintain its own interaction data store and provide access to the interaction data store as a service.

[0031] The interaction data store **19** may store raw event streams with processing of the event streams being performed by the URL selection module **18**, for example to rank the URLs according. Alternatively, the interaction data store may have an associated processing module (not shown) that preprocesses the interaction data so that the interaction data store stores the URLs in a ranked form. For example, a processing module may process the event streams to determine an event generator type (e.g. human, non-human, computer assisted human, etc) as described in the Applicant's co-pending patent application HAUSER001 and HAUSER006 referenced above. Once an interaction with a webpage has been classified as a human interaction, the data may be further processed to rank the particular behavior of the interactions. For example, the event streams may be processed to select those events streams containing out-click events, i.e. events that a user produces to exit a web page. The event streams and/or the page content may also be analyzed to determine additional preferred behavior, such as a breadth-first traversal of the web site, backlink count, partial page-rank calculations, page-rank calculations using a link graph with URLs only if those URLs have sufficient human interaction, etc. In one embodiment, the interaction pattern for parked pages, link farms, auto generated "spam" pages (that use random snippets from a variety of authentic web pages just to get high search engine ranking based on the keywords in the snippets) may be identified and used to remove these URLs from the crawl graph (not pursue the links) and/or remove such URLs from page-rank calculations.

[0032] A summary of the event statistics including any data used to rank the web pages may be stored in the interaction data store **19**.

[0033] An alternative embodiment is illustrated in FIG. **8**. In this embodiment, the web crawler **212** is modified to include a content interest data store **218**, which may be the same as the content interest data store described in the Applicant's co-pending application Attorney Docket No. HAUSER009 referenced above. The content interest data store **218** may store content interest data including a content interest score that ranks the various elements of a web page by the interest they receive during interactions between users and the respective web page. In one particular embodiment, the elements are document object model (DOM) elements of the web page. Content interest may be derived from an attention analysis of the event streams to determine where a user's attention focus was directed during an interaction, as described in the Applicant's co-pending application Attorney Docket No. HAUSER008 referenced above.

[0034] An operation of the download module **214** is illustrated in the flowchart **400** of FIG. **9**. Steps **401**-**405** are equivalent to steps **201**-**205** of the download module **14** described with reference to FIG. **3** above. Once the URLs found within a web page are added to the URL data store **16** at step **405**, the content interest data store **217** may be queried (step **406**) to determine if the web page being downloaded has content interest data. If no content interest data is available, the process proceeds as before by storing the web page content into the web page data store **15** (step **407**). If content interest data is available, then the page elements may be ranked by their respective content interest score (step **408**). At step **409**, a filter policy may be applied to the content elements. For example, content elements may be filtered for the most interesting content or for content above a certain threshold content interest score. Alternatively, content elements, such as DOM elements, may be grouped by similar content interest scores for an indexer or search engine. After filtering, the content elements of the web page, or at least those elements that pass the filter, may be stored in the web page data store **15** (step **410**).

[0035] In a further embodiment, the modified web crawler **212** of FIG. **8** may be used to apply alternative URL selection policies in the URL selection module **18**. In one embodiment, the URL selection module may select only URLs having human out-clicks, where the source element, i.e. the content element where a user's attention was directed prior to the out-click event, is a content element with a high content interest score. This selection policy requires a data correlation between the last content element to have the focus of the user's attention and the link (URL) of a human out-click. Such a data correlation may be built from the event stream and attention analysis data.

[0036] An alternative URL selection policy may specify that URLs (or human out-click URLs) will only be followed if there is some form of human area of interest within the page where the URL was found, e.g. a content element with a high enough content interest score. A further alternative URL selection policy may specify that URLs (or human out-click URLs) will only be followed if they are found within a content element with high enough content interest.

[0037] The URL selection policies followed by the URL selection module focus the web crawlers resources towards those web pages that are actively used by humans and thus generate particular attention events. Using the selection policies may significantly increase the efficiency of the web crawler and assist in providing higher quality page ranking statistics. Furthermore, as described above, common human browsing patterns, can be recognized via attention analysis for link farm pages, parked pages where the most interesting content is advertisements, and auto-generated "spam" pages. Human outclicks on pages that have no content of interest other than ads can be ignored by the URL selection module.

[0038] The embodiments described herein provide an enhanced system and method for web crawling that avoids spending resources collecting web pages that are not useful to humans. The effect of these embodiments is to reduce or eliminate the advantages of a link farm and to remove search engine spam. At current internet growth rates, the requirement to crawl less of the internet can provide large resource savings as well as making page ranking of web pages more efficient and useful for humans. By focusing crawling to the web pages relevant to and used by humans, the ability of artificially manipulate search engine rankings is reduced.

[0039] The web crawler **12** may be embodied in hardware, software, firmware or a combination of hardware, software and/or firmware. In a hardware embodiment, components of the web crawler **12** may be embodied in a device, such as server hardware, computer, etc. For example, the URL selection module **18** may include a processor **61** operatively associated with a memory **62** as shown in FIG. **10**. The memory **62** may store instructions that are executable on the processor **61**. In addition, the memory **62** may provide elements of the URL data store **16** and/or interaction data store **19**. An instruction set **500** that may be executed on the URL selection module processor **61** is depicted in the flowchart of FIG. **11**. Specifically, when executed, the instruction set **500** allows the processor to select (step **501**) a URL from the URL data store and look up the URL in the interaction data store to determine if interaction data for the URL exists (step **502**). If interaction data does exist, the processor **61** provides the selected URL to the download module. The download module may also be embodied in hardware and have a processor **71** and operatively associated memory **72** as shown in FIG. **12**. The download module processor **71** may communicate with the URL selection module processor **61** by an appropriate communication link. If appropriate, aspects of the URL selection module may be performed by the download module. For example, the download module may perform ranking of the URLs to be downloaded or may choose to ignore URLs provided by the URL selection module if content interest data for a URL is unavailable or indicates an insufficient content interest score.

[0040] Although embodiments of the present invention have been illustrated in the accompanied drawings and described in the foregoing description, it will be understood that the invention is not limited to the embodiments disclosed, but is capable of numerous rearrangements, modifications, and substitutions without departing from the spirit of the invention as set forth and defined by the following claims. For example, the capabilities of the invention can be performed fully and/or partially by one or more of the blocks, modules, processors or memories. Also, these capabilities may be performed in the current manner or in a distributed manner and on, or via, any device able to provide and/or receive information. Further, although depicted in a particular manner, various modules or blocks may be repositioned without departing from the scope of the current invention. Still further, although depicted in a particular manner, a greater or lesser number of modules and connections can be utilized with the present invention in order to accomplish the present invention, to provide additional known features to the present invention, and/or to make the present invention more efficient. Also, the information sent between various modules can be sent between the modules via at least one of a data network, the Internet, an Internet Protocol network, a wireless source, and a wired source and via plurality of protocols.

What is claimed is:

1. A method for web crawling comprising:

determining a plurality of Uniform Resource Locators (URL)s;

determining a subset of the plurality of URLs that have associated interaction data;

selecting at least one URL of the subset; and

downloading a web page corresponding to the at least one selected URL.

2. The method according to claim **1** wherein determining the subset comprises:

accessing an interaction data store that stores interaction data that associates a URL with an interaction with a web page corresponding to the respective URL; and

selecting a URL into the subset if a web page corresponding to the URL has interaction data.

3. The method according to claim **2** comprising ranking the subset of URLs using the interaction data.

4. The method according to claim **3** wherein the interaction data indicates one or more out-click events from the corresponding web page of an associated URL and wherein ranking the subset of URLs comprises ranking the URLs dependent on the one or more out-click events.

5. The method according to claim **4** wherein ranking a URL is dependent on a source content element of the web page prior to an out-click event.

6. The method according to claim **5** wherein ranking a URL is dependent on an attention analysis ranking of the source content element.

7. The method according to claim **3** wherein selecting at least one URL comprises selecting a highest ranked URL.

8. The method according to claim **2** comprising selecting a URL into the subset if a web page corresponding to the URL has interaction data that indicates at least one human dependent interaction with a web page associated with the URL.

9. The method according to claim **1** wherein downloading a web page comprises determining content interest of one or more content elements of the web page.

10. The method according to claim **9** comprising storing content elements that satisfy a threshold content interest requirement.

11. The method according to claim **1** wherein said downloading comprises providing the subset of URLs to a download module.

12. The method according to claim **1** further comprising storing the interaction data comprising:

receiving an event stream from an interaction between a user and a web page;

analyzing the event stream; and

storing the analyzed event stream in association with a URL for the respective web page.

13. A web crawler comprising:

at least one Uniform Resource Locator (URL) data store that stores a plurality of URLs;

at least one interaction data store that stores interaction data for a plurality of web pages, the interaction data indicating an interaction between a human and a web page corresponding to a URL;

at least one download module that downloads web page content corresponding to a URL; and at least one URL selection module in communication with the at least one URL data store and the at least one interaction data store;

wherein the at least one URL selection module selects at least one URL from the at least one URL data store that has interaction data in the at least one interaction data store; and

wherein the at least one URL selection module provides the at least one selected URL to the at least one download module.

14. The web crawler according to claim **13** further comprising:

at least one content interest data store that stores an attention ranking of one or more content elements of a web page; and

at least one web page data store;

wherein the download module is configured to:

utilize the at least one content interest data store to determine a content interest score of one or more content elements of a downloaded web page; and

store the one or more content elements of the downloaded web page in the at least one web page data store dependent on the respective content interest score.

15. The web crawler according to claim **14** wherein the at least one web page data store is configured to group content elements of a plurality of web pages according to their content interest score.

16. The web crawler according to claim **13** comprising an event server that:

receives at least one event stream generated during an interaction with a web page on a client browser;

analyzes the event stream; and

stores the analyzed event stream in the interaction data store.

17. The web crawler according to claim **16** wherein the event server analyzes the at least one event stream to determine an event generator type of the event stream.

18. The web crawler according to claim **13** wherein the at least one interaction data store receives interaction data from an event server.

19. A computer-readable medium comprising computer-executable instructions for execution by a processor, that, when executed, cause the processor to:

select a Uniform Resource Locator (URL) from a URL data store;

look up the selected URL in an interaction data store to determine if interaction data exists for the selected URL in the interaction data store; and

if interaction data exists for the selected URL, provide the selected URL to a download module.

20. The computer readable medium according to claim **19** comprising computer-executable instructions for execution by the processor, that, when executed, cause the processor to:

select a plurality of URLs from the URL data store that have corresponding interaction data in the interaction data store;

rank the plurality of URLs according to out-click event data of the interaction data;

provide at least one of the plurality of URLs to the download module;

wherein the at least one URL provided to the download module is provided depending on the rank.

* * * * *