International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501

A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

# A PERSUASIVE SOFTWARE AGENT FOR VALIDITY TEST THROUGH WEB CRAWLERS

Tamanna Jain

M.Tech. (CSE) Student

Maharishi Markandeshwar University, Mullana, Haryana, India

## ABSTRACT

The contemporaryperilousaugmentation of the World Wide Web and other on-line information sources has made captious commitment for some suite of ingeniousreinforcement to a user who is browsing for charismatic information. The voluminous size and changing nature of web make it necessary to continually update web based information extraction systems. Crawlers expedite this process by following hyperlinks in web pages to automatically download new and updated web pages. The main aim of this manuscript is to overview the concept of software agents and the web spiders or crawlers used by these software agents to make the search faster. A new algorithm is also proposed to validate the URLs in the search engine and check for the errors generated for different scenarios like in case when the server is down or in case when the page is not found. The proposed algorithm will check for the errors and take the corresponding steps accordingly. The paper also reviews the HTTP error codes generated whenever the request is made to the server.

## 1. SOFTWARE AGENTS

Software agent is a perpetual, purposeful and go-getter computer program that reciprocates to its environment and runs without continuous absolute oversight to perform some tasks for an end user or another program. Some software agents have user-interfaces. It is basically a computer equivalent to

International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501

A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

a self-governing robot. It can actuate and run itself, not demanding any input form or any other inter-communication with a human user. It can also perform *installation, commanding and abolishing of* other programs or agents including applications and other online intelligent agents. They have specialized purpose and routinely visit one or more web sites in the execution of their assigned tasks. Some of the applications of these agents are listed as below:

a) Administer targeted Internet searches

b) Checking and scheduling incoming E-mail

c) Testing new computer games

d) Controlling online job searches

e) Accumulating tailored new reports

f) Finding good deals in e-commerce

## 1.1 Types of Software Agents

The working of a software agent is generally based on an agreement to act on one's behalf (like some other agents). There are several types of agents that can be used for settling the agreement.

i) *Intelligent Agents*: selectively demonstrating some aspects of artificial intelligence, such as learning and reasoning

ii) *Autonomous Agents*: able to customize the way in which they achieve their objectives

iii) *Distributed Agents*: being executed on physical distinct computers

iv) *Multi-agent Systems*: distributed agents that do not have the capabilities to achieve an objective alone and thus must communicate.

v) *Mobile Agents*: agents that can transpose their execution onto different processors

International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501

A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

1.2 Software Agents Organisation

A software agent is nothing more than a computer program. However, the software acts in an intelligent manner, making assumptions based on preferences defined, or that it has learned by analysing the human behaviour. A goal ('find me pictures of dolphins') is specified, and it searches the same for its users. It might perform these tasks when the screen saver is active, or late at night when the computer is even switched off.

The figure 1.1 shown below illustrates the organisation of a software agent in web mining.

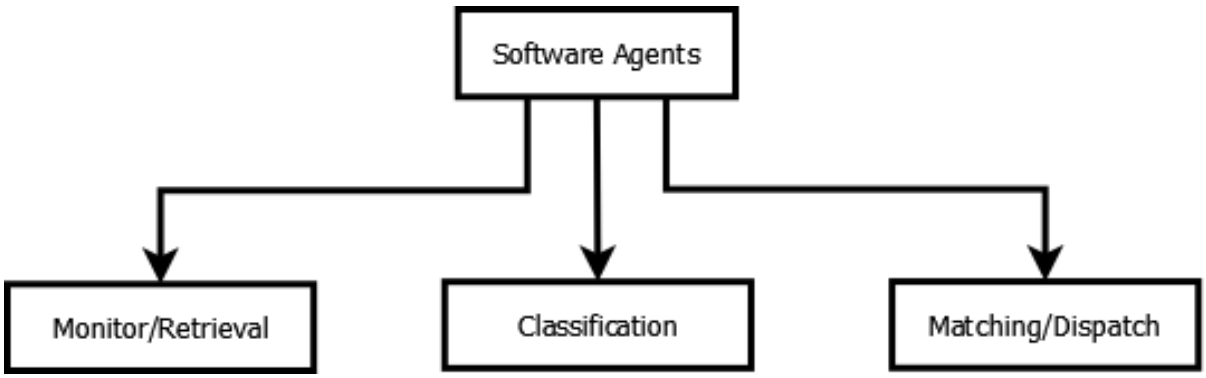


Fig 1.1 Organisation

A) **Monitor/Retrieval**

- Crawling relevant URLs
- Search for required data and links
- HTML parsing and cleaning
- Information extraction
- Transfer raw data to the database

B) **Classification**

International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501

A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

- Data pre processing
- Training data and data cleaning
- Multi class classification of data
- Inserting classified data into the database

C) **Matching/Dispatch**

- Query parsing
- Database transaction
- Data cleaning
- Information retrieval from the database
- Interface and display

## 2. SEARCH ENGINES

A **search engine** is a program accomplished to supportcatch files reserved on a computer, for example a public server on the World Wide Web, or one's own computer. The search engine grants one to ask for media content meeting specific benchmark (typically those containing a given word or phrase) and fetching a list of files that meet those benchmark. A search engine often uses an already made, and frequently rejuvenated index to look for files after the user has enlisted search pattern.

In the ambience of the Internet, search engines usually associate to the World Wide Web. Additionally search engines extract the information available in newsgroups, large databases, or open directories like DMOZ.org. Because the data acquisition is computerized, they are discriminated from Web directories, which are preserved by people.

### 2.1 Challenges faced by search engines

International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501

A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

- The web is flourishing much faster than any contemporary-technology search engine can probably index.
- Many web pages are amended generally, which punches the search engine to return them occasionally.
- The queries one can generate are presently limited to searching for key words, which may results in many false positives.
- Dynamically generated sites, which may be slow or crucial to index, or may return extravagant results from a single site.
- Many dynamically originated sites are not index able by search engines; this paradox is known as the invisible web.
- Some search engines do not order the results by pertinence, but rather according to how much money the sites have paid them.
- Some sites use deception to operate the search engine to display them as the first result returned for some keywords. This can direct to some search results being desecrated, with more relevant links being pushed down in the result catalogue.

### 2.2 **How search engines work**

Web search engines work by accumulating data about a large number of web pages, which they fetch from the WWW itself. These pages are redeemed by a web crawler (sometimes also known as a spider) — a computerized web browser which pursues every link it contemplates. The significance of each page is then scrutinized to impel how it should be indexed. Data about web pages is stored in an index database. Whenever a user makes any query and inserts the corresponding keywords in the search engine, the engine searches the index and provides the best matching web pages according to the query. There is another main type: Real-time search engines which don't use an index. The efficacious of a search engine depends on the applicability of the results it gives back. Web search engines and some other sites use Web crawling or spidering software to update their web content or indexes of others sites' web content.

General purpose search engines use web crawlers or spiders to perpetuatetheir index database, redressing the cost of crawling and indexing over the millions of queries acquired by them. These crawlers are blind and profound in their approach, with extensiveness as their major goal.

## 3. SPIDERS (CRAWLERS)

International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501

A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

Web spiders (also known with some other names like crawlers, web robots, web scutter, automatic indexer, wanderers) is a program or automated script which browses www in an analytical and computerized manner. A number of sites use spidering as a means of providing up-to-date data.

3.1 Uses of Web Spiders

- The web crawler is used to create a duplicate copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.
- They may be used for building focused repositories, automating resource discovery, and facilitating software agents.
- Crawlers can validate hyperlinks and HTML codes.
- They can also be used for automating perpetuation tasks on a web site, such as checking links or validating HTML codes.
- Gathering specific kind of information from web pages, such as harvesting e-mail addresses can also be done with the help of a web crawler.
- They exploit the graph structure of the web to move from page to page

The key motivation for designing web spiders has been to extract web pages and add them to a local repository. Such a repository may then serve particular application needs such as those of a web search engine.

The working of the crawlers starts from a seed page (or a root page) and then uses the external links within it to attend other pages. The process repeats with the new pages adducing more external links to follow, until an acceptable number of pages are diagnosed or some higher-level objective is reached. Behind this simple description, there lays a host of issues related to network connections, spider traps, canonicalizing URLs, parsing HTML pages, and ethics of dealing with remote web servers. Once all the pages had been fetched to a repository, there would be no further need for crawling.

## 4. ERRORS GENERATED

International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501
A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

Whenever a request is made to some server for a page on a web site, then the server returns an HTTP status code in response to the request. This status code provides the information about the status of the request. Some common status codes are:

- 200: server successfully returned the page
- 404: requested page does not exist
- 503: server is temporarily unavailable

Brief Description of the HTTP error codes

- **1xx (provisional response)**: status codes that indicate a provisional response and require the requestor to take action to continue.

| S.No. | Error Code | Description |
|---|---|---|
| 1. | 100 | Continue |
| 2. | 101 | Switching protocols |
| 3. | 102 | Processing |

Fig 4.1 1xx HTTP error codes

- **2xx (successful)**: status codes that indicate that the server successfully processed the request.

| S.No. | Error Code | Description |
|---|---|---|
| 1. | 200 | Successful |
| 2. | 201 | Created |
| 3. | 202 | Accepted |
| 4. | 203 | Non-authoritative information |
| 5. | 204 | No content |
| 6. | 205 | Reset content |
| 7. | 206 | Partial content |
| 8. | 207 | Multi-status |
| 9. | 208 | Already reported |
| 10. | 226 | IM used |
| 11. | 250 | Low on storage space |

Fig 4.2 2xx HTTP error codes

- **3xx (redirected)**: further action is needed to fulfil the request. These codes are generally used for redirection.

International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501
A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

| S.No. | Error Code | Description |
|---|---|---|
| 1. | 300 | Multiple choices |
| 2. | 301 | Moved permanently |
| 3. | 302 | Moved temporarily |
| 4. | 303 | See other location |
| 5. | 304 | Not modified |
| 6. | 305 | Use proxy |
| 7. | 306 | Switch Proxy |
| 8. | 307 | Temporary redirect |
| 9. | 308 | Permanent redirect |

Fig 4.3 3xx HTTP error codes

- **4xx (request error)**: status code indicating that there was likely an error in the request which prevented the server from being able to process it.

| S.No. | Error Code | Description |
|---|---|---|
| 1. | 400 | Bad request |
| 2. | 401 | Not authorized |
| 3. | 402 | Payment required |
| 4. | 403 | Forbidden |
| 5. | 404 | Not found |
| 6. | 405 | Method not allowed |
| 7. | 406 | Not acceptable |
| 8. | 407 | Proxy authentication required |
| 9. | 408 | Request time out |
| 10. | 409 | Conflict |
| 11. | 410 | Gone |
| 12. | 411 | Length required |
| 13. | 412 | Precondition f ailed |
| 14. | 413 | Request entity too large |
| 15. | 414 | Requested URL is too long |
| 16. | 415 | Unsupported media type |
| 17. | 416 | Requested range not satisfiable |
| 18. | 417 | Expectation failed |
| 19. | 418 | I’m a Teapot |
| 20. | 420 | Enhance your calm |
| 21. | 422 | Unprocessable entity |
| 22. | 423 | Locked |
| 23. | 424 | Failed dependency or method failure |
| 24. | 425 | Unordered collection |
| 25. | 426 | Upgrade required |
| 26. | 428 | Precondition required |
| 27. | 429 | Too many requests |
| 28. | 431 | Request header fields are too large |
| 29. | 444 | No response |

International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501
A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

| | | |
|---|---|---|
| 30. | 449 | Retry with (Microsoft) |
| 31. | 450 | Blocked by windows parental controls |
| 32. | 451 | Parameter not understood |
| 33. | 451 | Unavailable for legal reasons |
| 34. | 451 | Redirect (Microsoft) |
| 35. | 452 | Conference not found |
| 36. | 453 | Not enough bandwidth |
| 37. | 454 | Session not found |
| 38. | 455 | Method not valid in this state |
| 39. | 456 | Header field not valid for resource |
| 40. | 457 | Invalid range |
| 41. | 458 | Parameter is read only |
| 42. | 459 | Aggregate operation not allowed |
| 43. | 460 | Only aggregate operation allowed |
| 44. | 461 | Unsupported transport |
| 45. | 462 | Destination unreachable |
| 46. | 494 | Request header too large |
| 47. | 495 | Cert error |
| 48. | 496 | No cert |
| 49. | 497 | HTTP to HTTPS |
| 50. | 499 | Client closed request |

Fig 4.4 4xx HTTP error codes

- **5xx (server error)**: status code indicating that the server had an internal error when trying to process the request. These errors tend to be with the server itself, not with the request.

| S.No. | Error Code | Description |
|---|---|---|
| 1. | 500 | Internal server error |
| 2. | 501 | Not implemented |
| 3. | 502 | Bad gateway |
| 4. | 503 | Service unavailable |
| 5. | 504 | Gateway timeout |
| 6. | 505 | HTTP version not supported |
| 7. | 506 | Variant also negotiates |
| 8. | 507 | Insufficient storage |
| 9. | 508 | Loop detected |
| 10. | 509 | Bandwidth limit exceeded |
| 11. | 510 | Not extended |
| 12. | 511 | Network authentication required |
| 13. | 550 | Permission denied |
| 14. | 551 | Option not supported |
| 15. | 598 | Network read timeout error |
| 16. | 599 | Network connect timeout error |

Fig 4.5 5xx HTTP error codes

International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501

A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

## 5. PROPOSED FRAMEWORK

The key motive behind this manuscript is to design a procedure to check the validity of the URLs in the search engines. The idea behind this is very simple. First of alla database to the search engine is created with the following fields:

| ID | URL | Title | Description | Keywords | Page Relevancy | Error Generated (if any) |
|---|---|---|---|---|---|---|

Then each of the URL will be taken and identified through activated software agents. These agents will check each of the URLs for the corresponding errors generated (if any). If there are no errors then the counter for each URL will be incremented. The URL corresponding to the particular page with the highest counter increment will determine the relevancy of each page.

## 6. PROPOSED ALGORITHM

The algorithm for the above mentioned framework is as:

1. Fetch the database structure

   i←id;

   $u_i$←url;

   p←page;

   c←error code;

   int$x_p$; //counter x to be incremented or decremented for page p

2. Identify each URL, i.e$u_i$←URL
3. Activate the software agent ($S_a$)
4. Repeat step 5 to 8 for i = 1 to n
5. Visit $S_a$ ←→ $u_i$ , while ($u_i$!=NULL)

returnS.code (c) // error code int c

6. If (c==404), print “page not found”

   i++, x--

International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501

A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

Goto step 4

else

7. If(c==503), print “server is temporarily down”

   i++, x--

   Goto step 4

   else

8. If (c==200),

i++, x++ // counter x

Goto step 4

9. If ($u_i$==NULL) //implies no URL left
10. Stop
11. Return x.

## 7. FUTURE WORK

In the present paper, an overview on software agents and their working with spiders is discussed. A detailed study of the errors is shown with the corresponding codes description. A new algorithm is proposed to check the validity of the URLs and the pages associated with these URLs. In the next manuscript this algorithm will be implemented with the results and statistics of sampled database.

## REFERENCES

1. URL: http://en.wikipedia.org/wiki/List_of_HTTP_status_codes

2. URL: http://books.google.co.in/books?hl=en&lr=&id=q0qb5Vi02YAC&oi=fnd&pg=PA153&dq=research+papers+on+software+agents+web+crawlers&ots=-zTwXLO1ec&sig=V69J226lR_LzH_C_oz6e8cPHiWs#v=onepage&q=research%20papers%20on%20software%20agents%20web%20crawlers&f=false

3. URL: http://en.wikipedia.org/wiki/Software_agent

International Journal of Electrical, Electronics and Mechanical Controls ISSN (Online) : 2319 – 7501

A Scientific and Scholarly Resource for Publication of Research Manuscripts

Volume 2 Issue 1 January 2013

http://www.ijeemc.com

4. URL: http://www.sce.carleton.ca/netmanage/docs/AgentsOverview/ao.html

5. URL: http://en.wikipedia.org/wiki/Web_crawler

6. URL: http://www.sciencedaily.com/articles/w/web_crawler.htm