# Use case Document

## Search Engine & Crawler

**Project Groep Thywin**

**9-5-2014**

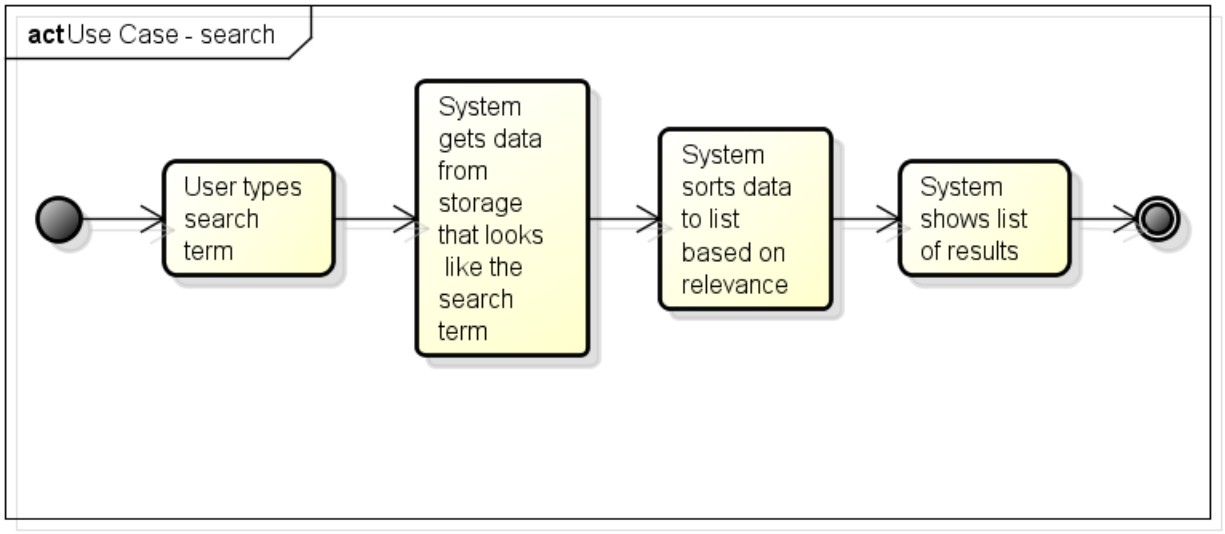
# Inhoud

# Use Case 1: Search

The user goes to http://thywin.com. The browser then shows a page. The user types a search term in the search field. The server then shows all the results based on the search term.

| **Primary Actor: User** | |
|---|---|
| **Stakeholders: Site owners** | |
| **Preconditions:** | |
| **Post conditions:** | |
| **Main success scenario:** | |
| 1. User goes to **http://thywin.com** | 2. Webserver sends page back |
| 3. User types search term in the search field | 4. Webserver shows results. |
| **Extensions: ( or Alternative flow)** | |
| | [If no results]<br>4a. Webserver shows message “no results” |

## Activity diagram




## Sequence diagrams

# Use Case 2: Crawler

## Korte omschrijving (brief vorm)

De Crawler vraagt het systeem on een url die gecrawled moet worden. Het systeem geeft crawler daarna een url. Hierna zal de crawler kijken of het een http / https url is.

Vervolgens zal de crawler de header opvragen van de pagina en de status code en het type van de reply controleren (STATUS: 200 OK – TYPE: HTML). Hierna wordt het gehele document op gehaald en opgeslagen op het systeem, wannaar het wordt doorgegeven aan de parser.

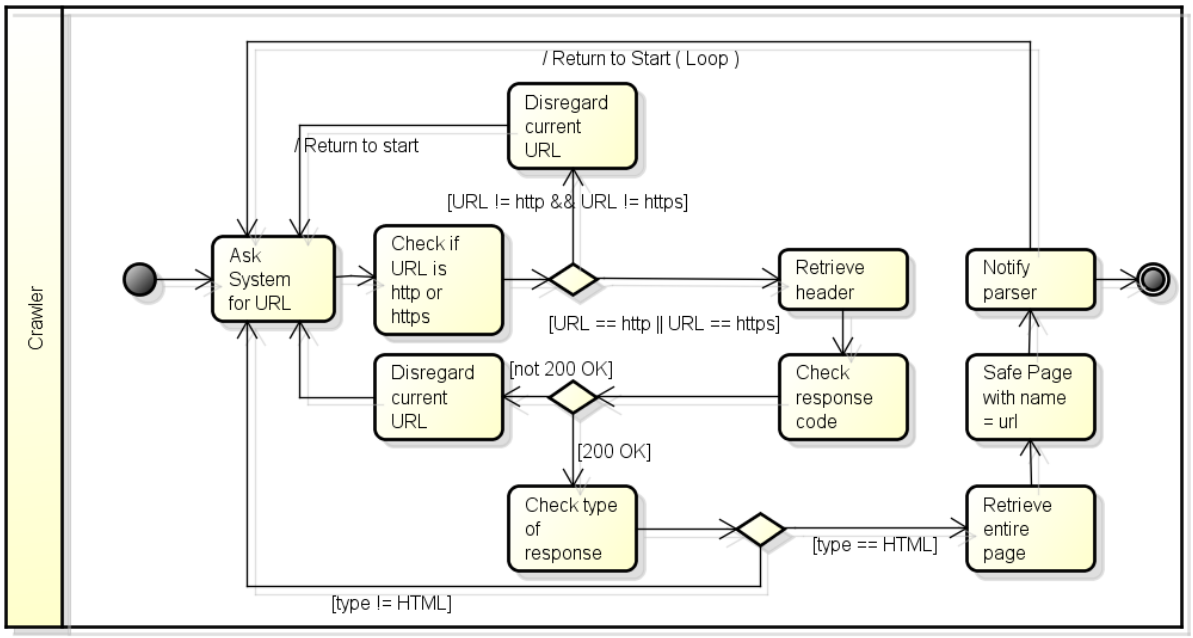
| **Fully Dressed** | |
|---|---|
| **Primary actor**: Crawler | |
| **Stakeholders & interest**: Scheduler, Parser | |
| **Preconditions**: Connection to the internet, list of urls to crawl | |
| **Main success Scenario** | |
| | |
| 1. Crawler vraagt systeem om een url die gecrawled moet worden | 2. Systeem pakt een url uit de nog te crawlen lijst |
| | 3. Systeem kijkt of de terug gekomen url een http of https protocol is |
| | 4. Systeem vraagt header van pagina |
| | 5. Systeem kijkt of response code in header 200 OK is |
| | 6. Systeem controleert de type van de response: 'type/html' |
| | 7. Systeem haalt gehele pagina op |
| | 8. Systeem slaat pagina op met de url als naam |
| 9. kijkt of de terug gekomen url een http of https protocol is | 10. Systeem laat de parser weten dat er een nieuwe pagina is gecrawled, welke naam die heeft en waar deze te vinden is |
| | |
| *5. Response code is geen 200 OK* | |
| | 6a. Systeem stopt<br>END OF USE CASE |
| *6. Type is geen HTML* | |
| | 7a. Systeem stopt<br>END OF USE CASE |

# Kenmerken

| Kenmerk | Omschrijving |
|---|---|
| **Aanleiding** | |
| **Actors** | Crawler |
| **Wijze van uitvoering** | |
| **Samenhang met andere use cases / requirements** | Use Case – parsen, Use case – Scheduler.<br>Requirement: Ophalen van web pagina's. |
| **Frequentie van uitvoering** | Herhalend |

# Volgorde van gebeurtenissen

## Activiteitendiagram

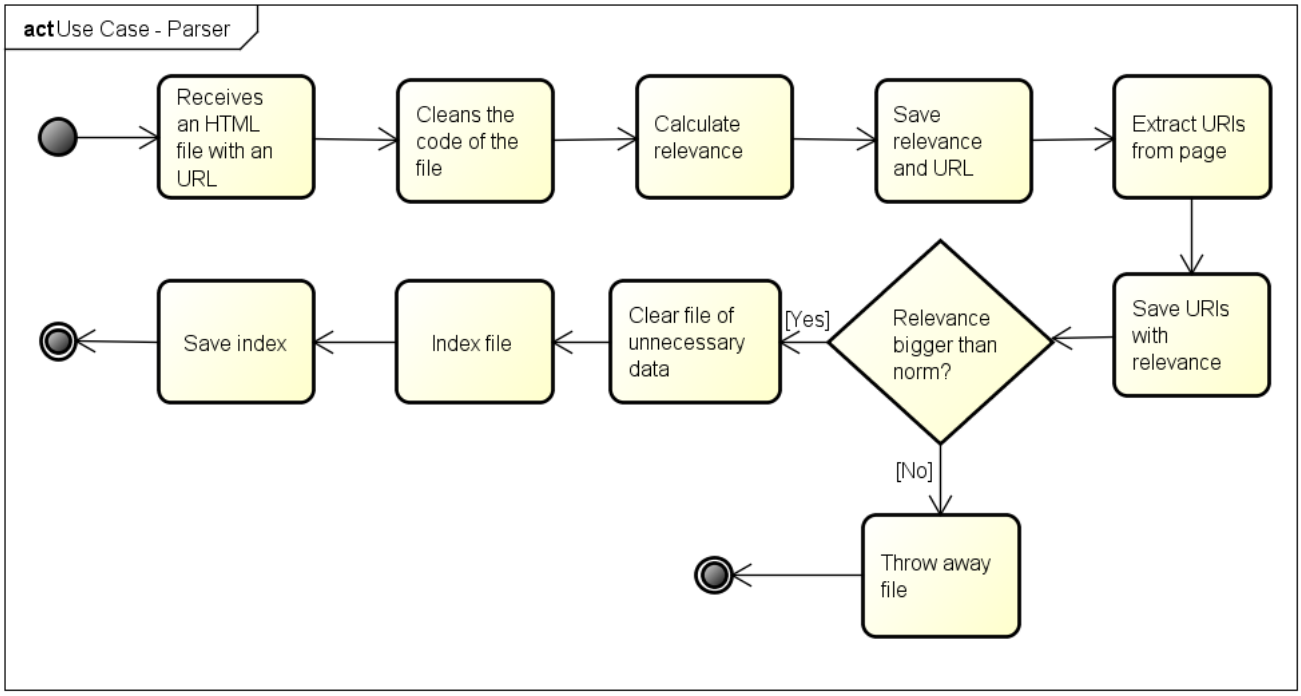


# Use Case 3: Parsen

The parser receives an HTML file and the related URL from the crawler. The first thing it does is clean the code of the file. After the code is cleaned, it will determine the relevance to the given source document. The URL, in combination with the relevance score, are stored in the database. The parser will then extract the URIs from the file. The URIs, in combination with the relevance score, are stored in the database. If the relevance is bigger than 0, the parser will clear the code so that only plain text remains. The file will then be indexed according to its words. The index is then stored in the database.

| **Primary Actor: Crawler** | |
|---|---|
| **Stakeholders:** Crawler, Parser, Database | |
| **Preconditions:** Connection to the database and the crawler, file is HTML format | |
| **Post conditions:** File is indexed and the index, URIs and relevance is stored | |
| **Main success scenario:** | |
| **Actor action** | **System action** |
| | 1. System receives an HTML file and URL |
| | 2. System cleans the code of the file |
| | 3. System determines relevance and saves it together with the URL |
| | 4. System extracts URIs from page and save them with the relevance |
| | 5. If the relevance is bigger than the minimum norm, the system clears the file |
| | 6. System indexes file and saves the index |
| **Extensions (Alternative flow)** | |
| 5. Relevance is lower than the minimum norm | |
| | 5. System deletes the file<br>END OF USE CASE |

# Activity diagram




# Sequence diagrams

# Use case 4: Master

TBA