IMPLEMENTATION OF SEARCH ROBOT'S FUNCTION TO COLLECT INFORMATION IN SCIENTOMETRIC SYSTEMS

IMPLEMENTACIÓN DE LA FUNCIÓN DE ROBOT DE BÚSQUEDA PARA RECOPILAR INFORMACIÓN EN SISTEMAS CIENTÍMETRICOS

Anis F. Galimyanov, Dmitriy A. Minullin

[1] Kazan Federal University
e-mail: anis_59@mail.ru, minullin.dima@mail.ru
Tel.: +79050210915

Enviado: 27 de junio de 2019
Aceptado para publicar: 30 de julio de 2019
Publicado: 8 de agosto de 2019

abstract

At present, the World Wide Web is developing rapidly, and every day the problem of automated collection and analysis of information placed on various web resources is becoming increasingly urgent. If in the 90s of the last century, the World Wide Web was a huge amount of poorly structured information, to search in which it was difficult for a person. It was then that the first developments in the field of automated agents began to appear, facilitating the task of finding the necessary information on the web. The main part of such systems is a search robot - a software package that navigates through web resources and collects information for a database. In the Kazan (Volga Region) Federal University, a monthly rating of academic staff is compiled based on data placed in the personal offices of employees in the Electronic University system. Now there is a need to move away from manually filling the Hirsch index in a personal account with KFU staff to avoid incorrect data filing and validation of the entered information by the Prospective Development Center. What was required was the creation of a search robot to automatically collect the Hirsch indices of KFU employees from the Scopus system. This article discusses the search robot: What is it? How does he work? How to write your program to collect information? All these issues were addressed in this article. The possible types of search robots and the whole process of their work were considered. The Scopus scientometric system and scientometric indicator - Hirsch index, its purpose, and calculation were considered. For implementation, the Python programming language was used and the tools for implementing HTTP requests and processing HTML pages were considered.

Keywords: search robot, spider, crawler, bot, parser, crawler, bot, robot, spider, Hirsch index, Scopus, python, requests, Beautiful Soup.

En la actualidad, la World Wide Web se está desarrollando rápidamente, y cada día es cada vez más urgente el problema de la recopilación y el análisis automatizados de la información colocada en diversos recursos web. Si en los años 90 del siglo pasado, la World Wide Web era una gran cantidad de información mal estructurada, para buscar en lo que era difícil para una persona. Fue entonces cuando comenzaron a aparecer los primeros desarrollos en el campo de los agentes automatizados, facilitando la tarea de encontrar la información necesaria en la web. La parte principal de dichos sistemas es un robot de búsqueda, un paquete de software que navega por los recursos web y recopila información para una base de datos. En la Universidad Federal de Kazan (Región del Volga), se compila una calificación mensual del personal académico en función de los datos colocados en las oficinas personales de los empleados en el sistema de la Universidad Electrónica. Ahora es necesario dejar de llenar manualmente el índice Hirsch en una cuenta personal con el personal de KFU para evitar el archivo incorrecto de datos y la validación de la información ingresada por el Centro de Desarrollo Prospectivo. Lo que se requería era la creación de un robot de búsqueda para recopilar automáticamente los índices Hirsch de los empleados de KFU del sistema Scopus. Este artículo aborda el robot de búsqueda: ¿qué es? Como trabaja el ¿Cómo escribir su programa para recopilar información? Todos estos problemas fueron abordados en este artículo. Se consideraron los posibles tipos de robots de búsqueda y todo el proceso de su trabajo. Se consideró el sistema cienciométrico Scopus y el indicador cienciométrico: índice de Hirsch, su propósito y cálculo. Para la implementación, se utilizó el lenguaje de programación Python y se consideraron las herramientas para implementar solicitudes HTTP y procesar páginas HTML.

Palabras clave: robot de búsqueda, araña, rastreador, bot, analizador, rastreador, bot, robot, araña, índice de Hirsch, Scopus, python, solicitudes, sopa hermosa.

Introduction

At present, the World Wide Web is developing rapidly, and every day the problem of automated collection and analysis of information placed on various web resources is becoming increasingly urgent. If in the 90s of the last century, the World Wide Web was a huge amount of poorly structured information, to search in which it was difficult for a person. It was then that the first developments in the field of automated agents began to appear, facilitating the task of finding the necessary information on the web. The main part of such systems is a search robot - a software package that navigates through web resources and collects information for a database.

In the Kazan (Volga Region) Federal University, a monthly rating of academic staff is compiled based on data placed in the personal offices of employees in the Electronic University system. The purpose of rating assessment is to create a system of motivation of the NDP for high-quality and effective activities, the development of initiatives, the achievement of key indicators of the University's Development Program; increasing the level of objectivity in assessing the contribution of each NPR to the educational process and scientific activities.

For the construction of the rating, indicators such as the number and quality of publications indexed in the Web of Science and Scopus database for the reporting period, the period for which the rating is made, are used. The article's belonging to the reporting period is determined by the date it was added to the employee's account.

One of the indicators for building a rating is the Hirsch index, which serves as an assessment of the activity of a scientist in terms of publications.

Now there is a need to move away from manually filling the Hirsch index in a personal account with KFU staff to avoid incorrect data filing and validation of the entered information by the Prospective Development Center. For all teachers, the Hirsch index should be populated using the Scopus database by automatically reading the relevant data from the staff pages.

What is a search robot? Basic concepts

Currently, there are many definitions of the search robot, here are some of them:

A search robot is a special program that is part of a search engine and is designed to crawl Internet pages to enter information about them (keywords) into the search engine database [11]. Also used names: crawler, spider, bot, automatic indexer, ant, web crawler, bot, web robot, web spider.

The search robot is a browser-type program. He constantly scans the network: visits indexed (already known to him) sites, follows links from them and finds new resources [14, 18].

A robot or bot is a special program that performs automatically and/or according to a predetermined schedule any actions through interfaces designed for humans [13].

A search robot is a program that visits hypertext links, extracting all subsequent documents from one resource or another by entering them into a search engine index [12].

Search engine robots are tireless internet workers who constantly browse hundreds of thousands of websites and gigabytes of text in search of the most up-to-date and interesting information [17, 19].

A crawler is a specialized program that uses the graph structure of the Web to navigating the pages of websites to collect the required information [7].

Of all the above definitions, we can distinguish several basic functions that every search robot should have:

- Scan
- Indexing
- Formation of results

Index (search index) is a search engine database in which information about web resources is collected. The purpose of this parameter is to increase the speed of searching for relevant web documents that correspond to specific user requests.

Indexing or indexing is the process of collecting, sorting and adding data to the search engine database. It is carried out according to specially developed algorithms that are available for each search engine.

Consider the types of search robots presented in table №1 [6].

Table 1. Types of search robots

Title	Purpose
National Search Robot (main search robot)	Collection of information from one national domain and web-resources, taken for indexation into the database of the search engine (for example .ru, .su).
Global search robot	Collection of information from national web-resources. Maybe one or more.
Image Indexer	Responsible for indexing graphics.
Audio and Video File Indexer	Responsible for indexing audio and video files.
Mirror mirror	Defines mirrors of web resources.
Reference robot	Responsible for counting the number of links on the resource.

Designer Robot	Responsible for the design of the results issued by the search engine. For example, referring to a web page via the link “Found words” and highlighting the words of the query in its text.
Testing robot	Checks the presence of a web-resource in the database of the search engine and the number of indexed documents.
Snitch robot	One or more robots that determine whether a resource is currently available that is referenced in the corresponding service. If not available for some time, then it is removed from the database.
Spy robot	Searches for links to web resources that are not in the search engine database.
Fast robot	Checks the date of the last update.
Robot explorer	Designed to debug the algorithm of the search engine or research specific web-resources.
Robot Caretaker	Designed to recheck the results.

How to search for robots work

Search robots should be perceived as automated data retrieval programs traveling through the network in search of information and links to information.

To understand how a search robot works, you need to look at the site through the eyes of a robot. To do this, we use the special service "Site through the eyes of a robot" https://pr-cy.ru/simulator/. As an example, take the main page of the KFU website https://kpfu.ru. Before looking at the site through the eyes of a robot look at it through the eyes of a man

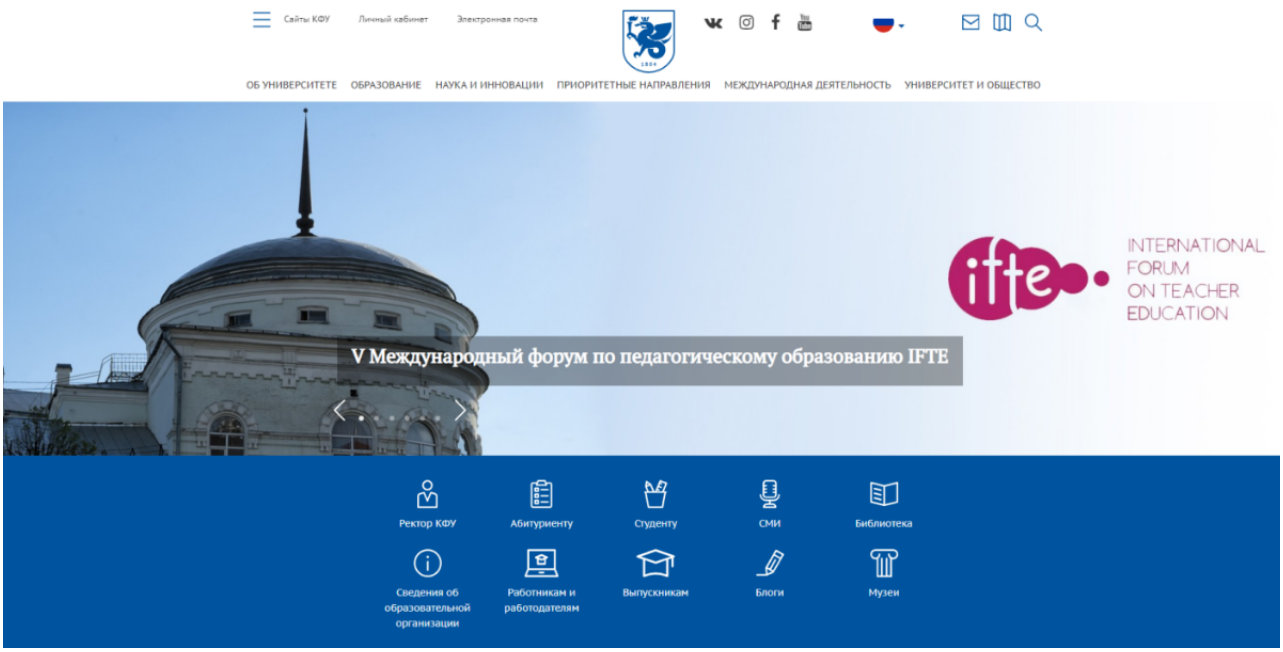

Figure 1. The main page of the KFU site.

Now we can go to the robot. First, it performs an HTTP GET request, which contains the full URL of the page. The server responds to the request in the same way as in the case of browser hits. First, the header of the HTTP Response Header response is presented, as shown in Figure 2, with information about the document [16].

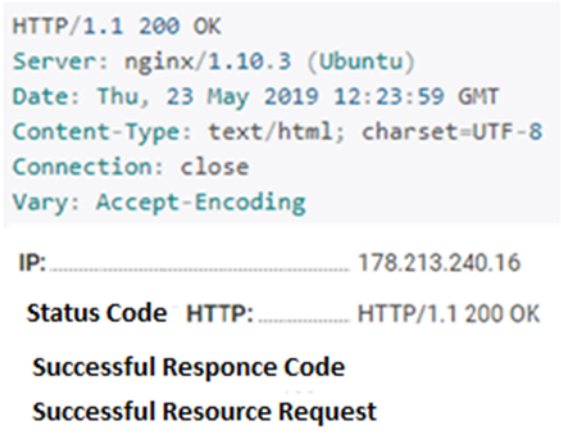

Figure 2. HTTP Response Header

Here we can find out the following information:
• Server response code
• The web server running the site and operating system
• server time
• The MIME type of resource and its encoding
• Connection status
• Content coding on the server
• IP address
There are a large number of headers that can be transferred, but it all depends on the server and site settings.
After the robots receive the HTTP Response Headers, there are two possible developments, depending on the status of the server's response, or the robot continues to work and the person stops and goes to another site. To continue to work or not, the robot decides depending on the status of the server response. There are five classes of answers [10]:
• 1xx - information codes. They are responsible for the data transfer process. These are temporary codes, they inform that the request is accepted and processing will continue.
• 2xx - successful processing. The request was received and successfully processed by the server.
• 3xx - redirection (redirect). These answers state that further action is needed to fulfill the request.
• 4xx - user error. This means that the request cannot be executed due to its fault.
• 5xx - server error. These codes are caused by a server-side error. In this case, the user did everything correctly, but the server cannot complete the request.
If the robot receives a server response belonging to class 1-3, then it continues its work, if to 4 or 5, then skips this site and goes to another. Now that we know the HTTP Response Headers and the robot is ready for further work, the server sends the document itself, usually, this is an HTML page, but there may be other document formats. In Figure 3, we can observe the site through the eyes of a robot [17].

```
<!DOCTYPE html>
<!-- App: main_page -->
<!-- Update: 24/03/2019 Sergey -->
<html lang="ru" dir="ltr">
<head>
    <meta charset="utf-8" />
    <meta name="title" content="Казанский (Приволжский) федеральный университет" />
    <link rel="canonical" href="https://kpfu.ru/" />
    <base href="https://stud-new2.kpfu.ru/" />
    <meta name="MobileOptimized" content="width" />
    <meta name="HandheldFriendly" content="true" />
    <meta name="viewport" content="width=device-width, initial-scale=1.0" />
    <link rel="shortcut icon" href="sites/default/files/favicon.ico" type="image/vnd.microsoft.icon" />
    <meta name="google-site-verification" content="YKj0-jgwp5_8YKVvgocl2-YiFV2n-59bpN5nYpQ0TDo" />
    <meta name="sputnik-verification" content="dGn78dgtlzt2XYa6" />
    <base href="">
    <meta http-equiv="Cache-Control" content="no-cache">
    <title>Казанский (Приволжский) федеральный университет - официальный сайт</title>
    <link rel="stylesheet" href="themes/stable/css/system/components/ajax-progress.module.css">
    <link rel="stylesheet" href="themes/stable/css/system/components/align.module.css">
    <link rel="stylesheet" href="themes/stable/css/system/components/autocomplete-loading.module.css">
    <link rel="stylesheet" href="themes/stable/css/system/components/fieldgroup.module.css">
    <link rel="stylesheet" href="themes/stable/css/system/components/container-inline.module.css">
    <link rel="stylesheet" href="themes/stable/css/system/components/clearfix.module.css">
    <link rel="stylesheet" href="themes/stable/css/system/components/details.module.css">
    <link rel="stylesheet" href="themes/stable/css/system/components/hidden.module.css">
    <link rel="stylesheet" href="themes/stable/css/system/components/item-list.module.css">
    <link rel="stylesheet" href="themes/stable/css/system/components/js.module.css">
    <link rel="stylesheet" href="themes/stable/css/system/components/nowrap.module.css">
```

Figure 3. Part of the HTML page received by the robot.

In Figure 3, we see only a small part of the HTML page that the robot received.
Each page begins with specifying the type of the current document. This is necessary for the browser to understand how to interpret the current web page. The following is a <head> tag which is a container for other elements whose purpose is to help the browser with data. Also, inside this block are meta tags, which are used to store information intended for browsers and search engines. Search engine engines access meta tags for site descriptions, keywords, and other data. After the <head> tag comes the <body> tag which is intended to store the content of the web page (content) displayed in the browser window [15].
The robot runs through the entire HTML page looking for a tag that is responsible for placing the hyperlink, namely the <a> tag. Then it determines which type the hyperlink belongs to: is it external and leads to one of the sections of the site, or else

it is external and a link to this link leads to another site. In our case, as shown in Figure 4, 428 internal links and 521 external links were found on the KFU website.

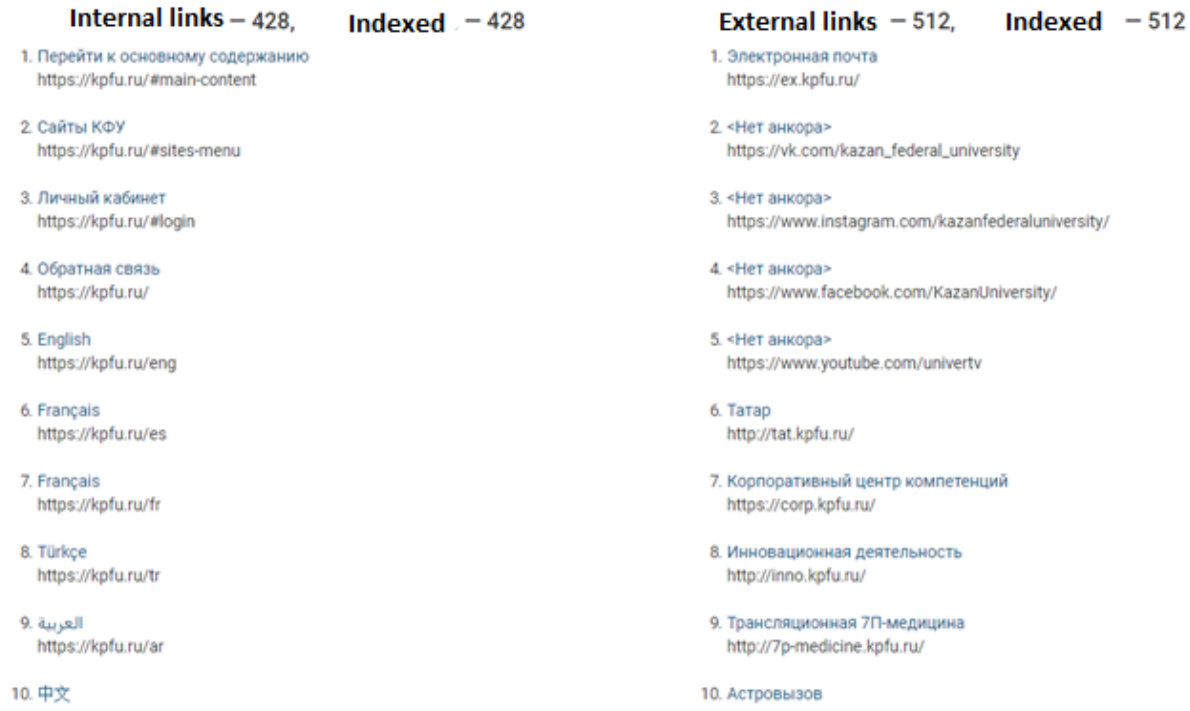

Figure 4. Hyperlinks.

Found the robot adds links to the database. Then, depending on the purpose of the robot, it starts indexing the page: selects keywords, determines the page content, find all images, audio and video files, or much more. After the completion of work on this page, the robot will continue to follow the found links.

1. The Hirsch Index. Scopus.

The Hirsch Index, or h-index, is a scientometric indicator proposed in 2005 by physicist Jorge Hirsch of the University of California, San Diego, initially to assess the scientific productivity of physicists. However, now the Hirsch index is a quantitative characteristic of the productivity of a scientist, a group of scientists, a scientific organization or a country as a whole, based on the number of publications and the number of citations of these publications, serves to assess the activity of a scientist in terms of publications [8]. The index is calculated based on the distribution of citations of the works of this researcher. According to Hirsch:

"A scientist has an index h if h from his h from his N articles is cited at least h times each, while the remaining (N - h) articles are cited no more than h times each."

The citing of an article is called using Scientist 1 in the list of references for his article to refer to the article of Scientist 2. The more different authors used in their research a link to your article, the more quoted it is considered.

The following algorithm can also be used to calculate the Hirsch index [9]:

1. The list of articles is sorted from most to least quoted.
2. Take the first article and look at the number of citations, if the number of citations is more than the serial number of the article, then go to the next one.
3. Repeat step 2 until the article number becomes more than the number of citations.
4. The number of the previous article will be equal to the Hirsch index.

The graphics calculation of the Hirsch index is presented in graph №1.

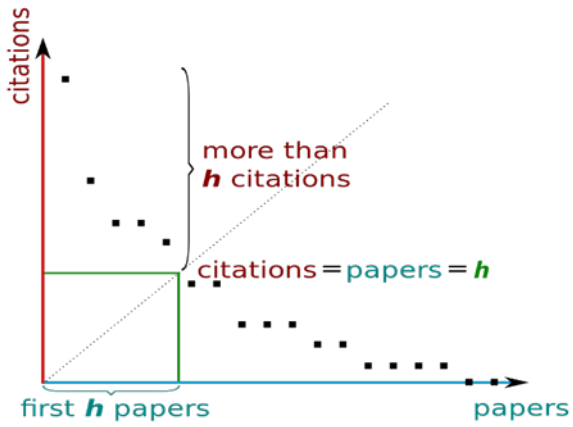

Chart 1. Getting the Hirsch index from the

distribution of articles according to the number of citations

To better understand this definition, consider a few examples:

• The author has published 1 article, which was referred to at least 50 times, his h-index is 1. The same will be the index of the author, who wrote 50 articles that were referred to only 1 time.

• The following example is most common in real life. The author has written 10 articles, of which the first is cited 7 times, the second - 6, the third - 5, the fourth - 4, the rest have been cited less than 4 times. That is, the author has 4 papers, cited at least 4 times, therefore, his Hirsch index will be 4. The remaining 6 papers did not affect the final result since cited less than 4 times, i.e. their serial number is higher than the number of citations.

With the help of the Hirsch index, you can more accurately understand how popular the work of a researcher is. In other words, it highlights the work of the author, which turned out to be the most popular in recent years.

In our case, we will collect the Hirsch index from the pages of the authors-employees of KFU in the Scopus system.

Scopus is a scientometric reference database included in the Elsevier SciVerse database. SciVerse combines materials from the SciVerse Scopus collection of the reviewed literature, the SciVerse ScienceDirect collection of full-text articles, as well as data from the Internet and advanced applications developed by the scientific community that enrich the contents of the database and increase its value [5].

Scientific resources published after 1996 are indexed in the Scopus database along with lists of article bibliographies. Citation in the database is calculated by automated analysis of the contents of these lists. Thus, Scopus counts the number of links to all indexed resources published since 1996 [4].

For authors who have published more than one article, individual accounts are created in Scopus - author profiles with unique author IDs (Author ID).

Using the search function of authors in SciVerse Scopus is easy to find the right author. You just need to enter the name and initials of the author, for greater accuracy, you can enter the organization in which he works, and his ORCID.

Author details page (Author details) contains contextual information about the author, with the help of which you can check whether he is the author who interests you. The following data is displayed here [4]:

• The author's affiliation with the organization, as recorded in the last publication

• Number of documents from this author, presented in the latest publication in the Scopus database

• Number of links to the work of this author in the Scopus database

• The number of documents containing citations from the works of this author.

• Hirsch Index

• Number of collaborators

• Number of web results from the Scirus system

• Subject areas, materials on which this author is published

The Scopus database in many countries is one of the main sources of scientometric data acquisition for conducting state and/or corporate assessment studies.

1. Implementation

To implement the project, we will use the Python language, but before proceeding to work, you need to figure out what the author's page in Scopus is.

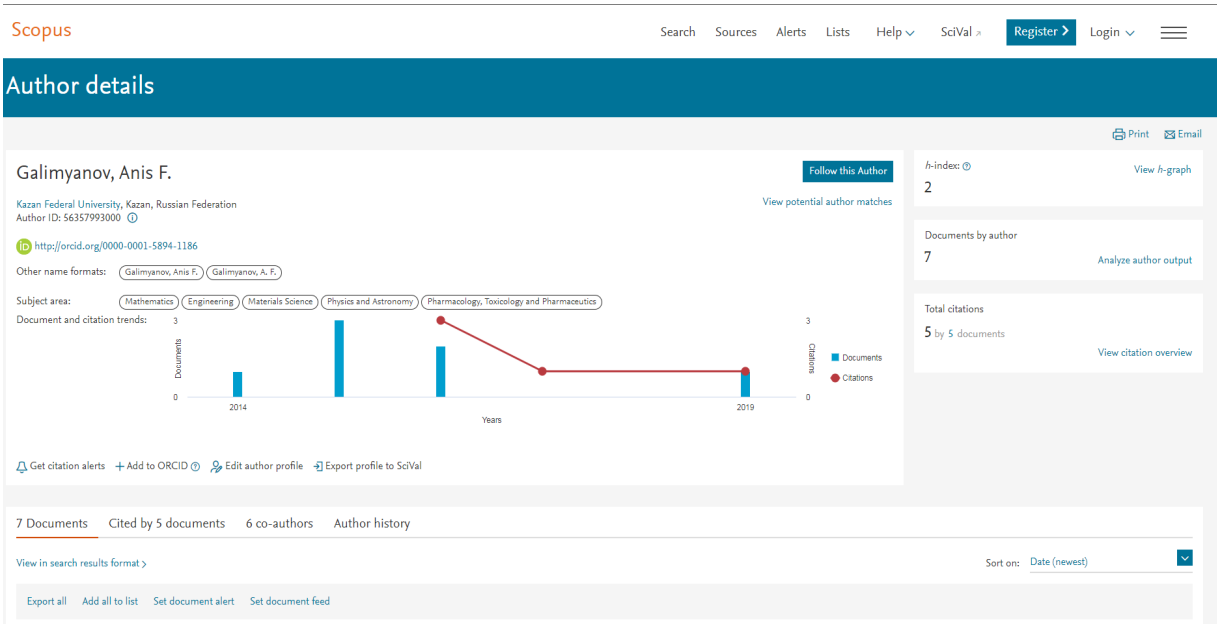

Figure 5. Personal page of the author.

On this page, we can see information about the author: name, place of work, Hirsch index, the number of published documents and the number of citations. Now you need to look into the structure of the HTML page and find out in which block of the code the desired Hirsch index is located. To do this, open the page code inspector in the browser.

Figure 6. Html block of the desired index.

The required element is in the block with the class panel-body. Also, for the full functioning of the program, it is necessary to determine the URL at which we will receive the personal pages of the authors, employees of KFU. The main part of the URL address: https://www.scopus.com/authid/detail.uri?authorId=

Additional part: 5635799300 – author ID - a unique identifier of the author in the Scopus system.

Now we have all the necessary information to proceed with the implementation of our project.

As mentioned earlier, the Python programming language will be used to implement the project. To do this, we need several libraries:

1. The request is a library for making HTTP requests [3].
2. Beautiful Soup is a parser for parsing HTML / XML files, written in the Python programming language, which can even translate incorrect markup into a parse tree. It supports simple and natural ways to navigate, search and modify the parse tree [2].

To get started, we need to get the HTML text of the page. To do this, we use the get method from the requests library.

```
def get_html(url):
    r = get(url)
    if r.status_code == 200:
        return r.text
    else:
        return 'error'
```

Figure 7. HTTP request

The robot sends an HTTP GET request, which contains the full address pointing to the author's page in Scopus. The server responds to the request in the same way as when the browser accesses: it sends the response header (HTTP Response Header) with information about the document, followed by the document itself (usually an HTML page, but there may be other document formats), if the response is not then returned error.

After we got the HTML page, we need to extract the Hirsch index, for this, we use the library Beautiful Soup which provides convenient and intuitive functionality for work that simplifies and accelerates the search for the test, blocks, tags.

```
def get_data(html):
    if html != 'error':
        try:
            soup = BeautifulSoup(html, 'lxml')

            h_index = soup.find('div', class_ = 'panel-body').text
            if not h_index.split():
                h_index = '0'
            return re.sub('\s', "", h_index)
        except Exception:
            return 'error'
    else:
        return html
```

Figure 8. Parsing HTML page.

The HTML content is converted to a BeautifulSoup object, which allows us to work with it as with a normal Python object, i.e. refer to the methods of the module [1].

All functions have been prepared for information, now it remains to launch the robot.

```
def robot(arr, response):
    for i in arr:
        try:
            if i.isdigit():
                url = SCOPUS_URL + i
                h_index = get_data(get_html(url))
                if h_index == 'error':
                    response.append('{0}=error'.format(i))
                    continue
                h_index = h_index.replace('\n', '')
                response.append('{0}={1}'.format(i, h_index))
            else:
                response.append('{0}=error'.format(i))

        except IndexError:
            continue

        except Exception:
            response.append('{0}=responseError'.format(i))
```

Figure 9. Search robot.

The robot receives two arrays at the input: the first, consisting of the authors' identifiers in the Scopus system, the second array is empty, we will write the results into it.

Summary

The search robot is not a scary and mysterious thing as it might seem at first glance. Search robots can now facilitate the work of a person of any profession in finding the necessary information. The job of a search robot is quite simple and for a specialist person to write his program for obtaining information is quite simple.

Conclusion

In this way, we learned what search robots are, how they are classified, how they see a web page, and how they work. Studying the function of a search robot to collect information in scientometric systems, we developed a program for our robot, which collects information from the personal pages of the authors-employees of KFU in the Scopus system and allows us to determine the Hirsch Index.
Automated determination of the Hirsch Index facilitates the work on rating evaluation of the achievements of key indicators of the University staff; increasing the level of objectivity in assessing the contribution of each NPR to the educational process and scientific activities.

Acknowledgments

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

1. Web Scraping with Python. Ryan Mitchell, 2015 https://yanfei.site/docs/dpsa/references/PyWebScrapingBook.pdf
2. Official documentation for Python library Beautiful Soup. https://www.crummy.com/software/BeautifulSoup/bs4/doc/
3. Requests Documentation Release 2.21.0. Kenneth Reitz, 2019 https://buildmedia.readthedocs.org/media/pdf/requests/master/requests.pdf
4. Scopus. 2018 https://ru.wikipedia.org/wiki/Scopus
5. Methods of search in the Scopus database. Dudnikova O.V., Bondarenko S.A., 2011 https://library.sfedu.ru/media/upload/%20Материалы%20ДПО%20/Учебно-методическое%20пособие_Scopus2.pdf
6. Search robots. Markova T.I., Zakharova K.V. 2009 https://cyberleninka.ru/article/v/poiskovye-roboty
7. Adaptive crawler for searching and collecting external hyperlinks A.A. Pechnikov, D.I. Chernobrovkin. 2012 https://cyberleninka.ru/article/v/adaptivnyy-krauler-dlya-poiska-i-sbora-vneshnih-giperssylok
8. Hirsch Index. https://ru.wikipedia.org/wiki/Indeks_Hirsha
9. What is the Hirsch index and how to raise it? Alex Zvansky, 2017 https://wos-scopus.com/chto-takoe-indeks-hirsha/
10. HTTP response codes. 2019 https://developer.mozilla.org/ru/docs/Web/HTTP/Status
11. Search robot. https://ru.wikipedia.org/wiki/Search_robot
12. The search robot is what it is and how it works. http://seo-dnevnik.ru/blogosfera/poiskovyiy-robot-robotyi-poiskovyih-sistem.html
13. Bot (program). https://ru.wikipedia.org/wiki/bot_(program)
14. What is a search robot? https://wiki.rookee.ru/poiskovyj-robot/
15. HTML. 2019, https://ru.wikipedia.org/wiki/HTML
16. Search robots. 2010, http://wiki.webimho.ru/search exploit
17. Search engine robots. 2006, https://www.seonews.ru/masterclasses/robotyi-poiskovyih-sistem/
18. Jahwari, N. A., & Khan, M. F. (2016). ORGANIZATIONAL LEARNING MECHANISMS IN SOHAR UNIVERSITY. *Humanities & Social Sciences Reviews*, *4*(2), 76-87. https://doi.org/10.18510/hssr.2016.423
19. Shirvani, M., Mohammadi, A., & Shirvani, F. (2015). Comparative study of cultural and social factors affecting urban and rural women's Burnout in Shahrekord Township. UCT Journal of

Management and Accounting Studies, 3(1), 1-4.