DEVELOPING SMART WEB-SEARCH USING REGEX

Ikechukwu Onyenwe[1], Stanley Ogbonna[1], Ebele Onyedimma[1],
Onyedikachukwu Ikechukwu-Onyenwe[1], Chidinma Nwafor[2]

[1]Nnamdi Azikiwe University Awka, Anambra State, Nigeria
[2]Nigerian Army College of Environmental Science and Technology,
Makurdi, Benue State, Nigeria

ABSTRACT

Due to the increasing storage data on the Web Applications it becomes very difficult to use only keyword-based searches to provide comprehensive search results, thus increasing the difficulty for web users to search information on the web. In this paper, we proposed using a combined method of keyword-based and Regular expressions (RegEx) searches to perform search using strings of targeted items for optimal results even as the volume of data around the world on the Internet continues to explode. The idea is to embed RegEx patterns as part of search engine's algorithm in a web application project to provide strings related to the targeted items for more comprehensive coverage of search results. The user's search query is a string of characters guided by search boundaries selected from the entry point. The results returned from the search operation are different results within a category determined by the search boundaries. This is designed to be beneficial to a user who has an obscure idea about the information he/she wanted to search but knows the boundries within which to get the information. This technique can be applied to data processing tasks such as information extraction and search refinements

KEYWORDS

Web, RegEx, User Input, Information Extraction.

1. INTRODUCTION

Web-search involves searching for information on the Web. The search results are generally presented in a line of results often referred to as search engine results pages (SEROs). The information may be a mix of web pages, images, and other types of files. An entry point of most web-search is a search box/bar.

A search box is a graphical control element used in applications such as file managers, web browsers and websites. A search box is usually a single-line textbox with the dedicated function of accepting user input to be searched for in a database. On web pages users are allowed to enter a query to be submitted to a web-search engine server-side script, where an index database is queried for entries that contain one or more of the user's keyword research. it is an integral part of the site search functionality, which is an important element of website design for content-rich websites. On some websites, site search is more prominent than on others. E-commerce websites typically use search boxes, and thus site search, as a primary navigation tool.

Regular expression (RegEx) is based on the concept of a state machine, which is a process that will sequentially read in the symbols of an input word and decide whether the current state of the machine is one of acceptance or non-acceptance [**3**]. Using a RegEx requires defining all of the criteria that need to reach an accepted state in order for a valid pattern match to occur. Illustrating RegEx concept, we used example from [**3**] where *XYZ* is a RegEx to perform a pattern match on

the string *ZXYXYZ*. The RegEx solution starts by reading *Z* symbol into the state machine which failed to match the first condition of the RegEx pattern thus causing a state of non-acceptance and hence failure of the matching operation. The next characters X and Y in the string matched the first and second condition of the RegEx but failed at the character *Y* to reach an accepted state. This string portion starting with second character *X* failed to reach an accepted state. The third character of the string *Y* is used as a start symbol for the state machine which failed to match the first condition of the start machine and results in a failure as well. Finally, the last three characters X, Y and X of the string read into the state machine successfully match the XYZ RegEx conditions.

Getting a comprehensive search result on the Web even when the user is not sure of the right search keywords to use is an important module required for developing a user-friendly web application. The explosive rate of information growth and availability often makes it increasingly difficult to locate information pertinent to users' needs. Use of only keyword based search methodologies are not adequate for describing the information users seek. According to [**2**], The Web search engines started to take form with the use of regular expressions to search through their indexes. Regular expressions were chosen for these early search engines because of both their power and easy implementation. It is a fairly trivial task to convert search strings into regular expressions that accept only strings that have some relevance to the query. In the case of a search engine, the strings input to the regular expression would be either whole web pages or a pre-computed index of a web page that holds only the most important information from that web page.

$$(\Sigma^*\texttt{regular}\Sigma^*\texttt{expression}\Sigma^*)^* \cup (\Sigma^*\texttt{expression}\Sigma^*\texttt{regular}\Sigma^*)^*$$

Figure 1. RegEx query. Source [**Carter**]

A query such as regular expression in *Fig. 1* could be translated into the following regular expression, then, of course, would be the set of all characters in the character encoding used with this search engine. The results returned to the user would be the set of web pages that were accepted by this regular expression. [**4**] focused on text preprocessing of automotive advertisements domains to configure a structured database. The structured database was created by extract the information over unstructured automotive advertisements, which is an area of natural language processing. Information extraction deals with finding factual information in text using learning regular expressions. [**5**] presents the automation of a Web advertising recognition algorithm, using regular expressions. The tests were carried out in three Web browsers. As a result, the detection of advertisements in Spanish, that distract attention and that above all extract information from users was achieved. Extraction of information from the Web is a well-known but unsolved and critical problem when it comes to accessing complex information systems. These problems are related to the extraction, management and reuse of the huge amount of Web data available. These data usually has a high heterogeneity, volatility and low quality (i.e. format and content mistakes), so it is quite hard to build reliable systems. [**1**] proposed an Evolutionary Computation approach to the problem of automatically learn software entities based on Genetic Algorithms and regular expressions. One use of regular expressions that used to be very common was in web search engines. Archie, one of the first search engines, used regular expressions exclusively to search through a database of filenames on public FTP servers. [**6**] presented a novel approach for generating string test data for string validation routines, by harnessing the Internet. The technique uses program identifiers to construct web search queries for regular expressions that validate the format of a string type (such as an email address). It then performs further web searches for strings that match the regular expressions, producing examples of test cases that are both valid and realistic. Following this, our technique mutates the regular

expressions to drive the search for invalid strings, and the production of test inputs that should be rejected by the validation routine.

Recently, Properties Finders Web Apps (PFW) is on the increase in Nigeria. **PFW** helps users to find properties for sale and **rent** in the nation. This approach of finding properties is believe to be effective than the traditional approach where users will employ physical strength to find/locate property of their choices. This is a non-trivial task. PFW saves lots of time and energy in property finding, especially when the users know exactly what they are looking for in a rental home. For example, if a user's budget requires him/her to look for cheap houses for rent, he/she can do a preliminary search on PFW to get an overview of the available rental homes within his/her price range.

In this paper, we propose the use of a search method that combined keyword-based and RegEx searches for an efficient but comprehensive search output on PFW. We used keywords to serve as search boundaries and strings to match and filter all items pertinent to what the users seek within a confined boundary. This technique can be applied to data processing tasks such as information extraction and search refinement. RegEx is a tool in natural language processing (NLP) for information extraction and used at initial stage of NLP preprocessing pipeline. This research aims at using smart search tools to enhance PFW responses to clients' needs in a more rapid fashion, thereby building a positive customer experience. As an added benefit, PFW users will discover and buy relevant properties according to their preferences.

2. REVIEW OF RELATED LITERATURE

In modern websites, techniques such as RegEx, a web scraping and pattern matching methods, has been used to enhance user web searches experience. In this section, we present some previous works related to the use of RegEx in web applications. In [7], authors applied web crawling and scraping methods on an e-commerce website to get HTML data for identifying products updates based on the current time. [8] uses a feature extraction method in order to get more fine-grained structured data as an input for entity linking tools, and thus improve the matching precision. Their goal is to extract product attributes from product offers, by means of regular expressions, in order to build well-structured product specifications. E-Commerce websites must provide trust to attract consumers, [9] develop tools that can search the trust attributes to assists customers assess the trustworthiness of ane-Commerce website. These trust attributes in e-Commerce websites have been identified and located usually in 'Homepage' and 'Contact Us' pages. Since the trust attributes are usually placed in unstructured text, they authors used information extraction to extract the data. They used data from eCommerce websites in United Kingdom (UK), United States (US) and Malaysia to create patterns using regular expression. The patterns are tested against the e-Commerce websites from UK, US and Malaysia. Based on the results, which show that they can be used as a tool to enhance trust search attributes assisting the users to place trust in an e-Commerce website as quickly as possible. RegEx, according to [10] are used to help digital marketers to eliminate and automate the most boring and time-consuming parts of data analysis. For example, a digital marketer can perform custom extractions using Regex. Custom extractions allow users to extract tons of useful information from a website such as Email addresses, tracking IDs, Schema Markup, Page Titles, URLs, and tons more.

3. EXPERIMENTAL TOOLS

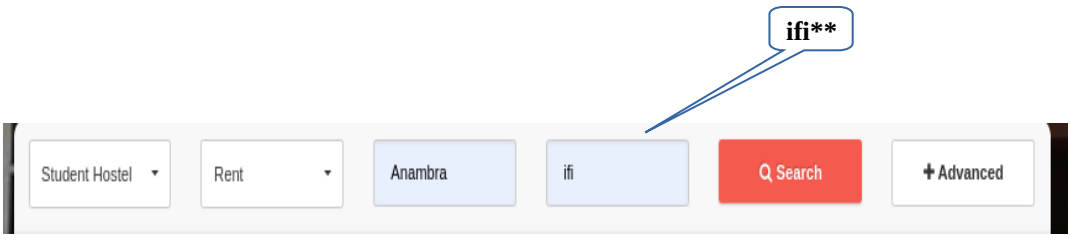
The followings describe the tools we used in this paper. Basically, we used Django, Python and Regular Expression commonly known as RegEx. We have introduced and extensively discuss RegEx in the introduction section.

Django is a Python-based free and open-source web framework that follows the model–template–views (MTV) architectural pattern. It is maintained by the Django Software Foundation (DSF), an American independent organization established as a non-profit. Django's primary goal is to ease the creation of complex, database-driven websites. The framework emphasizes reusability and "pluggability" of components, less code, low coupling, rapid development, and the principle of do not repeat yourself (DRY). Python is used throughout, even for settings, files, and data models. Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models.

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

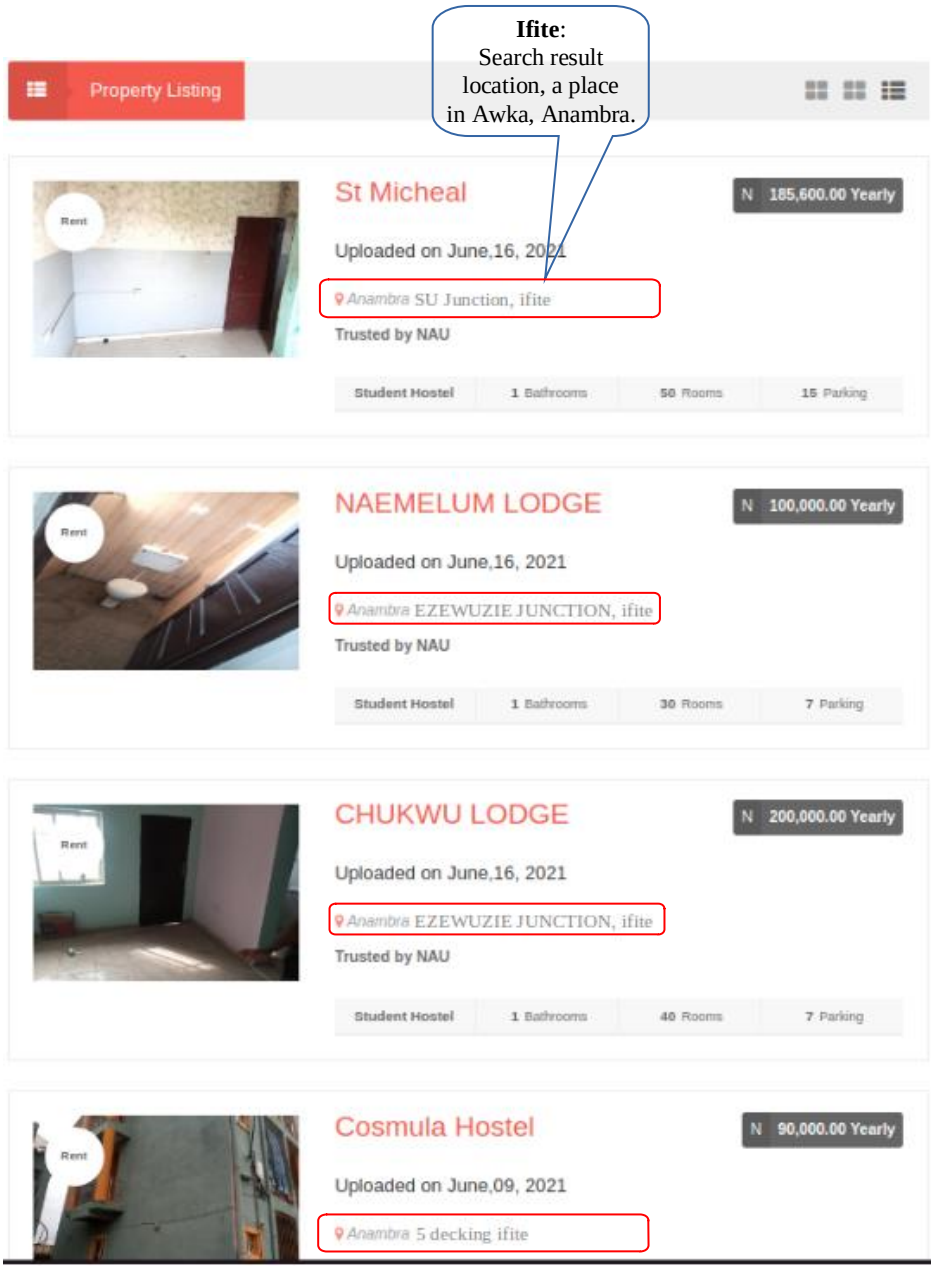
4. RESULTS AND DISCUSSION

For this paper, we used a property network website known as **propertywithin.com.ng** and integrate our combo search engine component in it. P**ropertywithin** is a friendly accommodation hub (search engine) aimed to facilitate users ability to find the property most suited to their needs regardless of location and position swiftly. There are subscribed trusted agents that bring onboard their properties for easy identification by possible occupants and create a customer-friendly atmosphere to choose property without much stress of physical sighting across Nigeria (including accommodation needs around campuses). It is an African real estate management agency that facilitates ease of identifying properties for both leasing, sales (buying), management, and securing of property' lives.



(a)

Fig. 2 (a) is the search box of the property within website. The search engine here combined the use of keyword-based and RegEx-based searches. From this Figure, observe the keywords **{Student Hostel, Rent, Anambra}** and string **{ifi}**. We used the **keywords** to set our search boundaries to be *only student hostels for rent within a location in Awka Anambra* and the **string** to be *a RegEx pattern to match and filter all locations that have available accommodations.*



(b)

Figure 2. Search box and search results using propertywithin.com.ng.

The search results of Fig. 2 (a) is displayed in Fig. 2 (b). Take note of the last words in the red marked boxes in Fig. 2 (b). The words indicate that the found accommodations are in Ifite, a place in Awka of Anambra most populated by the students of Nnamdi Azikiwe University. Also note that there are streets/closes we don't hve an idea about them but they are populated alonside their accommodations.

5. CONCLUSION

In this paper, we have shown how combination of the keyword-based and regular expression (RegEx)-based searches can be used to improve search efficacy in a comprehensive way. Illustrations in Fig.1 and 2 have shown how RegEx based pattern matching is a highly useful tool for augmenting keyword based search methodologies in that it can be used to dramatically reduce the number of irrelevant search results the user will be presented with. Again, the search results are comprehensive enough that even the locations the user has an obscure idea about the information he/she wanted to search are also populated. It works in such a way that once a search pattern is created such as **ifi** in Fig. 2 (a), a search system is initiated and the search results must be certainly located within the associated search attributes such as *student Hostel, Rent, etc.* as seen in Fig. 2 (a) and (b). This study is important since the task of searching the trust attributes is somehow hard especially for the beginner computer user and tools to help them in searching the trust attributes should be developed.

The future work that will be carried out shall consider testing the effectiveness and the accuracy of thetechnique using regular expression as regards to search results that are correct and finally testing the practicability of the system that has been developed.

REFERENCES

[1] Barrero, D.F., Camacho, D. and R-moreno, M.D., 2009. Automatic web data extraction based on genetic algorithms and regular expressions. In *Data Mining and Multi-agent Integration* (pp. 143-154). Springer, Boston, MA.

[2] Carter, B.A., Hubert, L.A. and Walton, A.C., 2007. Applications of regular expressions.

[3] Frenz, C.M., 2008, "Introduction to Searching with Regular Expressions", Proceedings of the 2008 Trenton Computer Festival.

[4] Jalal, A.A., 2020. Text Mining: Design of Interactive Search Engine Based Regular Expressions of Online Automobile Advertisements. *International Journal of Engineering Pedagogy*, *10*(3).

[5] Riaño, D., Piñon, R., Molero-Castillo, G., Bárcenas, E. and Velázquez-Mena, A., 2020. Regular Expressions for Web Advertising Detection based on an Automatic Sliding Algorithm. *Programming and Computer Software*, *46*(8), pp.652-660.

[6] Shahbaz, M., McMinn, P. and Stevenson, M., 2015. Automatic generation of valid and invalid test data for string validation routines using web searches and regular expressions. *Science of Computer Programming*, *97*, pp.405-425.

[7] Ikechukwu Onyenwe, Ebele Onyedinma, Chidinma Nwafor, Obinna Agbata, 2021. Developing Products Update-Alert System for e-Commerce Websites Users Using HTML Data and Web Scraping Technique. International Journal on Natural Language Computing, October 2021, Volume 10, number 5.

[8] Petrovski, P., Bryl, V. and Bizer, C., 2014, October. Learning Regular Expressions for the Extraction of Product Attributes from E-commerce Microdata. In *LD4IE@ ISWC* (pp. 43-54).

[9] Rusli, M.R., Che-Hussin, A.R. and Dahlan, H.M., 2010, February. Regular expression patterns for searching trust attributes in e-commerce website. In *Proceedings of the 9th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases* (pp. 162-167).

[10] Using Regex (Regular Expressions) in Digital Marketing. https://www.codefixer.com/blog/using-regex-regular-expressions-in-digital-marketing/