EAI Endorsed Transactions
on Ambient Systems
Research Article EAI.EU

A search engine dedicated to the environment around the Congo basin area: Model and Architecture

Mouhamadou Saliou Diallo[1,*], Ousmane Sall[2], Marie N'diaye[3] and Belmond Djomo[4]

[1]Laboratoire d'analyse numérique et d'informatique, Université Gaston Berger de Saint -Louis, ms.diallo@outlook.fr
[2]Université de Thiès, osall@univ-thies.sn
[3]Université de Ziguinchor, marie.ndiaye@univ-zig.sn
[4]MIT University Dakar, djomo.belmond@mit.edu.sn

Considering the environmental field around the Congo Basin area (in Central Africa), many initiatives implement and communicate on their activities via the internet, yet the information they provide remains very little exploited, even inaccessible. The data is scattered on the internet and it is quite difficult to find information in a fairly precise way. Despite the existence of Internet research services (search engines) developed to facilitate the search for information in the vast data network that is the Internet, there are still concerns about quality, and the relevance of information provided in as research results. In this context, we present in this paper a construction approach and the architecture of a search engine dedicated to the environment around the Congo Basin area. This paper develops a theoretical approach that uses scientific analysis and empirical approaches to conceptualize the optimization of the relevance of the results of a thematic search engine by aggregating tools. This approach is a compilation of ideas, methods and tools that, put together, will improve the relevance of the results of a thematic research.

Keywords: search engine, thematic research, search relevance, spiders, web crawlers, results optimization.

Received on 18 July 2018, accepted on 17 September 2018, published on 14 January 2019

doi: 10.4108/eai.23-3-2018.156247

*Corresponding author. Email: ms.diallo@outlook.fr

1. Introduction

Our world tends to be digitized more and more; this phenomenon has been favored by the advent of the Internet and its services. Thanks to the search services offered by search engines, we are able to find what we are looking for on the Internet. However, in the face of this wealth of information that is overflowing with the Internet, finding the right information can sometimes prove to be a long-term exercise. Indeed, the generalist character of the search field of these engines does not always lend to relevant results for the subject of its research. Focusing on the environmental field in Central Africa, it was noted that many initiatives implement and communicate on their activities via the Internet in Central Africa, yet the information they provide remains very little exploited, even accessible. Indeed, this information is disseminated via the Internet in several forms: web articles, documents, video files, audio, images, etc. The data is scattered on the Internet and it is quite difficult to find information in a fairly precise way. Despite the existence of search engines developed to facilitate the search for information in the vast data network that is the Internet, there are always concerns about the quality, richness, and relevance of the information provided as search results.

In this case, it would be desirable to have a research assistance tool for environmental issues in the Congo Basin area, which provides synthetic, rich, relevant and circumscribed results in the target research area. It is in this context that this article finds its justification.

Assuming that the results of a query submitted to a search engine would certainly be more relevant if the search engine takes into account or integrates certain specificities related to the query such as: the thematic district of the research subject, a tool to help the expression of research needs, and some dedicated algorithms. This article proposes conceptual methods that will make it possible to optimize the relevance of the results in a well-defined research context. The thematic constituency to be analyzed is the environmental domain in the Congo Basin in Central Africa. The approach adopted is the aggregation analysis of tools and methods

contributing to the optimization of the relevance of the results. The rest of this article is organized in sections, which will first expose the specificities of a search engine in general. Afterwards, we will analyze the specifications that give a search engine the dedicated character, and then we will discuss the elements to take into account in the design of our search engine dedicated to the environment. Finally, we will present the final architecture proposed for our search engine, and we will conclude this article with the perspectives.

2. How a search engine works

A search engine is a web application for finding documents from a query in the form of words more or less arranged. The documents can be web pages, articles or blogs, images, videos, files, etc. Some websites offer a search engine as the main feature; we then call search engine the site itself.

Although not all search engines work in exactly the same way, they do at least the following basic actions [1]:

(i) Collecting information on the Internet using collector tools called "crawler" or "spider".
(ii) The analysis of the words they find and their indexation;
(iii) The provision of a search service (search server) to help find words or combinations of words in the index database;

The crawler is software for structural, syntactic and semantic analysis of web pages [2]. For each page, it extracts the elements considered significant and relevant, to form a database of keywords related to the page analyzed. In addition, all hyperlinks in the page are extracted and added to a set of URLs to visit: this is the crawl border. As the number of URLs populating the crawl border increases very quickly, a criterion to prioritize the download of certain pages is generally applied. In turn, the top ranked URLs in the crawl border are downloaded and new links are extracted.

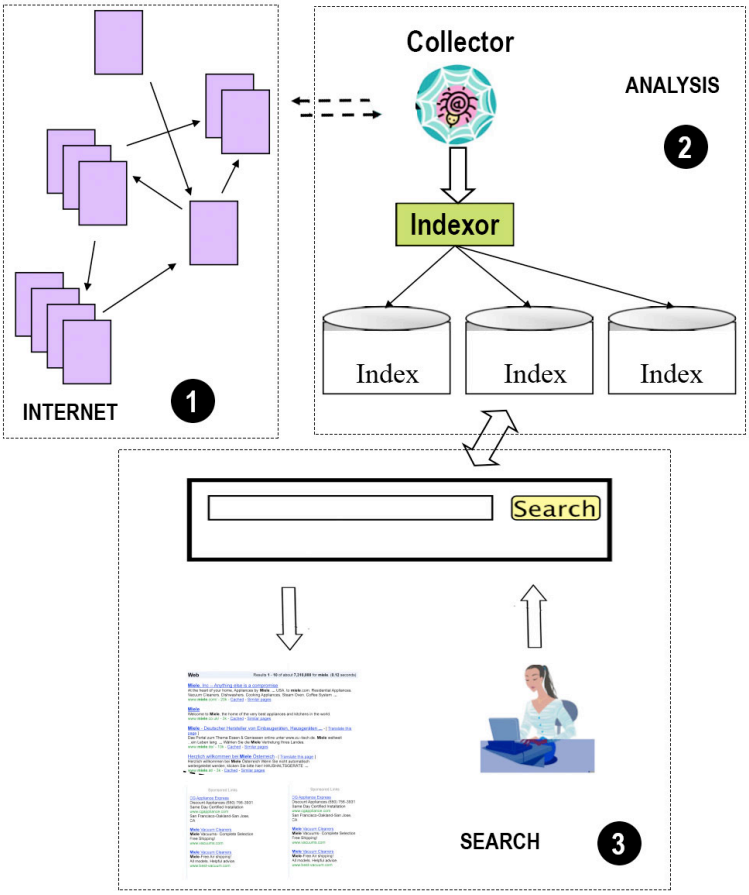

Figure 1. How a search engine works [1]

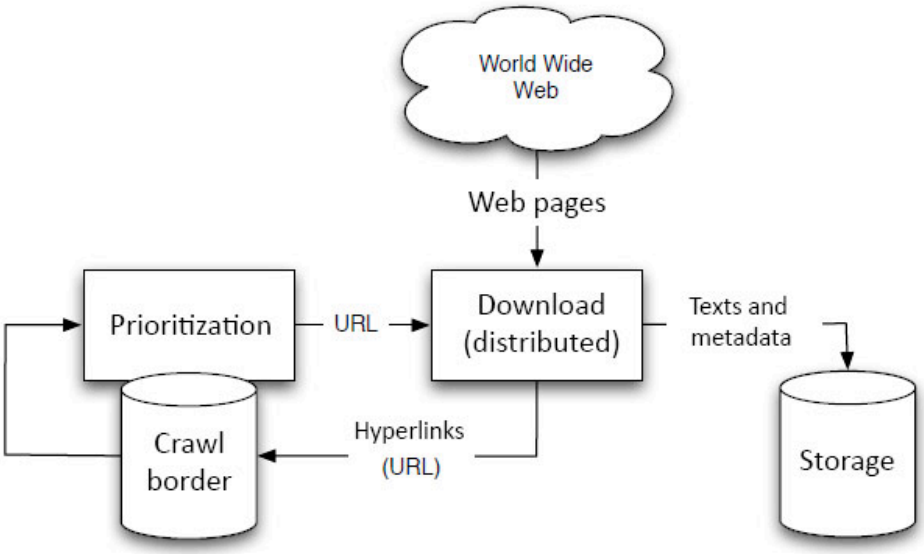

Figure 2. Web Crawling Principle [2]

3. Specificities of a dedicated search engine

A dedicated search engine, also called a vertical search engine, is a search engine specialized in finding specific content.

As we pointed out in the introduction to this document, the obvious main reason why dedicated search engines are developed is that while GPs provide access to countless information, it is not always easy to access in a few clicks to an article of proven relevance, dealing with a given subject. This need for access to even more specific and relevant data is also illustrated by the surveys conducted in [5,6].

Most of the classic search tools for the general public (Google, Bing, Yahoo, etc.); do not rank the results solely by thematic relevance, but also by popularity and many other criteria according to their policies (web referral rate, sponsorship, marketing, etc.) [5]. In addition, the phenomenon of linguistic ambiguity and the fact that users do not always express their queries better contribute to diversify the results provided by these engines.

According to [8], in the context of information retrieval, ambiguity can arise when a query is submitted to a search engine. Queries are artificial formulations, where words appear in a very small linguistic context. The artificial aspect of queries can be explained by the fact that they are built for the specific purpose in order to interact with a technological environment. The average size of requests is 2 to 3 words [5]. The context is therefore absent. For example without context, the query "orange" induces an ambiguity. Indeed, it is a word that can have several interpretations: a fruit,

. a color, a company. The diversity of interpretations can then be reflected in the results of a general search engine: the Orange customer site, the Orange business site, the Wikipedia page devoted to color or various information on the price of fruit in the global market.

This lack of context and the very short length of queries mean that a query can not be disambiguated in a language context. However, the problem of ambiguity affects the performance of the search system [4]. Because in the case of an ambiguous query, and the system will have more difficulty identifying the informational need underlying the query.

Dedicated search engines are characterized by their specialization which may be functional, geographical or thematic [7]. We will focus here on thematic search engines, ie dealing with a particular theme or area.

What makes the specificity of dedicated search engines unique is that they must operate by following the selection principle at all stages of the process [3].

Given their limited perimeter, we can hope to reiterate several advantages of these engines:

(i) First the motor theme reduces ambiguities of homonymic or semantic types. The term "virus", for example, may refer to a computer or biological virus. Faced with this type of query, general search engines have no alternative but to present all the propositions to the users, either implicitly by maximizing the diversity of the results [6], or explicitly [7]. The thematic search engines, on the contrary, carry a context that makes it possible to improve the interpretation of user requests ;

(ii) Reducing the number of documents to be processed makes it possible to improve the freshness of the data collected;
(iii) Engine specialization enables the integration of deep Web, structured data and domain knowledge (ontologies) at different levels of the search engine;
(iv) The human-machine interface can take advantage of the theme to offer unique search tools that perform better than the generic tools of a general search engine. On a cooking recipe site, it is for example interesting to be offered tools to select certain cooking times, types of dishes or families of ingredients;
(v) Users are often domain specialists who will make use of advanced search features.

4. Modeling a search engine dedicated to the environment: Aggregation of tools and methods

For the design of a search engine, a sequential organization of the whole process takes place in four phases: the constitution of the research database, the formulation of a

query, the evaluation of the relevance, the presentation of the results and the ergonomics of the system. Two main approaches exist to create thematic search engines: (I) specialize a general search engine by modifying the query and/or the results returned by the engine; (ii) build a specialized index and build a search engine on it. The second approach allows controlling the freshness and the quality of the data, but also to take better advantage of the field treated to improve the indexing and the research. In this paper, we will try to combine these two approaches to take advantage of each of them.

4.1. Proposed optimization approach for the dedicated engine

To build our search engine, it is necessary to collect a set of thematic documents. To do this, we will rely on the following approaches:

(i) Exploring the Web by orienting its path towards the relevant pages for the theme studied: the focused crawl;

(ii) URL prioritization of sites to scan: crawl border scheduling;

(iii) The use of a motor or several general search engines to find relevant pages for the theme studied: the meta-motor approach;

(iv) The The use of a help wizard for the formulation of user requests.

Regarding the crawl border scheduling, we propose that it be fed by several processes: (i) meta-engine URL submission: URLs from generalist engines; (ii) automatic URL submission: the internal system robot scans the web and retrieves the documents it finds by moving from link to link; (iii) manual URL submission: users submit to an administrator for validation the URLs of the sites deemed interesting.
Meta-engines save a lot of time because they use several other engines to find the information that best fits our expectations. Thus, it would be question of questioning our meta-motor at defined frequencies and to integrate the results obtained at our crawl border. The algorithm below could then be applied for the scheduling of URLs in the crawl border:

Program crawl border (Output)
Let FF be the crawl border (final border considered by the global system crawler)
Let FT be a temporary border for the crawl
Begin
Initialize FF to empty ; Initialize FT to empty ; ***Foreach*** *new URL manually submitted,*
Stack URL in head of FT;
Endforeach Foreach *thematic request transmitted to the metamotor,*
fetch the first 100 URL gathered by he metamotor ; ***Foreach*** *URL,*
If URL is not yet in FT, then stack it in queue of FT ;
Endforeach Endforeach
Foreach *URL gathered by web scanning, If URL is not yet in FT, then stack it in queue*
of FF ;
Endforeach
Stack FT in head of FF ; Return FF ;
end.

4.2. Choice of tools and methods

Since the emphasis is only on the conceptual approach in this document, we will present, in this part, a set of potential tools for the implementation of the approach developed :

Table 1. Choice of tools and methods

Tools	Proposal
Crawler **Search Server** **Metamotor**	Nutch, Websphinx, Scrapy, Php-Crawler, Crawl-Anywhere Apache Solr, Sphinx, Elasticsearch Metamotor based on two external search engines (via their APIs) which, according to the newspaper

	of the net [1], capture the largest parts of search queries on the Internet: Bing API Search and Google API Search.
FOCUSED CRAWL	The "NaiveBayesParseFilter" plugin is a module that classifies the links collected by the crawler as relevant or irrelevant according to the relevance calculated by a text analyzer. [2] Explains how, based on this Naïve BAYES algorithm, he was able to improve the relevance of the search results of a research server focused on articles in the field of physical science. And astronomical in databases of the SAO / NASA Astrophysics Data System (ADS).
Assistant to help with the formulation of requests	We propose that the wizard be structured in an elaborate interface of buttons of choice or elements of orientations constructed on the basis of the database of thematic knowledge of the environmental field of the system. The formulation of the request will consist of a pre-analysis of the key words provided by the user, in order to guide him towards the contextualization of his research in themes and sub- themes until a well-formulated request is obtained. And as explicit as possible.

5. The architecture of the final system

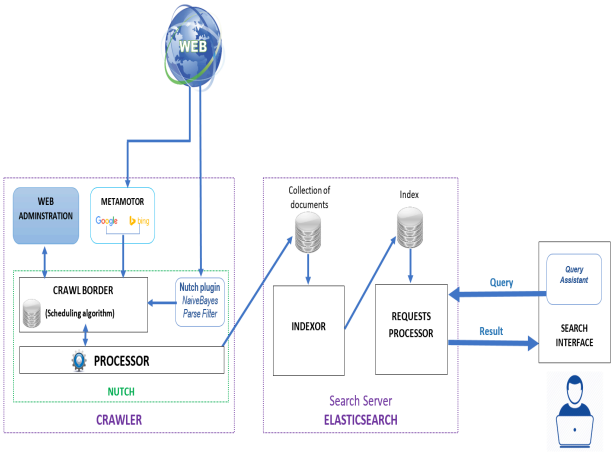

Figure 3. Architecture of the dedicated search engine

This architecture offers the advantage of a better alignment between the data collected and the user request. Indeed, it proposes a system of filter with each entry in the system: entry crawler and entry user. On the one hand, the crawler composed of three modules and information gathering operation in parallel (the manual submission via the administration, the metamotor and the Nutch processor). Each collection module has its own filter system. On the other hand the web search interface has help for the formulation of requests. In the middle, the Elasticsearch search server does the mapping to finish the best results for the user.

6. Conclusion

In this article, we have done several proposals : tools and methods, which put together will improve the performance of a dedicated search engine ; an algorithm applied for the scheduling of URLs in the crawl border ; a data collection approach. We also proposed an architecture allowing the installation of this very important engine for the Congo Basin. In this work, we have emphasized the relevance of the results to be returned by the search engine, which has made the calculation scheme of the results even more complex.

Our future work will be dedicated to the implementation of the architecture followed by a performance test using performance measures. We will also take into account the optimization aspects of calculations. We will also in future work integrate other modules to the set, such as a personalization module of information taking into account the profiles of users. Since this type of engine is intended to be used by people with various profiles and therefore different areas of interest. It will therefore be relevant to present the user with results likely to interest him. For this, we will build on the existing work on the extraction of user profiles [9, 10,11], the personalization of information [13, 14, 15] and the recommendation [12,13].

We also want to integrate an automatic language processing module, to take into account the linguistic context of the user. For the latter, we will use the results obtained as part of our KOCC-intelligent project.

Acknowledgements.

We thank ACE MITIC and Ministry of higher education research and Innovation of Senegal for supporting this work

References

[1] Technical Report : Monica P., Kamyar D. (2005) How search engines work and a web crawler application

[2] Thesis: Sabiha, B. et Hamza, K.(2012) Amélioration de la précision et du temps de réponse d'un moteur de recherche de texte.

[3] Scientific report : Lalleman, F.(2013) Étude de l'ambiguïté des requêtes dans un moteur de recherche spécialisé dans l'actualité : exploitation d'indices contextuels
[4] Conference: Bénédicte, P. et Xavier, L. (1997): Etat de l'art des idées implémentées dans les moteurs de recherche par index sur WWW,in EDF.
[5] Conference: Seymour, Dr. Tom, Frantsvog, D., Kumar, S. (2011). History Of Search Engines. In International Journal of Management & Information Systems
[6] Thesis: Clement De G. (2013) Focused document gathering on the Web for domain-specific information retrieval, Université Paris Sud - Paris XI
[7] Scientific report: Bill, F. et Zillmer, N. (2006). The Emerging Opportunity in Vertical Search, White paper
[8] Conference: Mejdl S., Abdullah A., Dunren C. (2012). Improving Relevance Prediction for Focused Web Crawlers, In proceding of 11th International Conference on Computer and Information Science, (2012), IEEE/ACIS.
[9] Thesis: DIALLO, M. S. (2015). Découverte de règles de préférences: application à la construction de profil utilisateur. PhD thesis Université de Tours, Université Gaston Berger de Saint- Louis, Mars 2015
[10] Conference: DIALLO, M.S, Sall,O. and Badji, I. (2017). Builduing user profiles based on fuzzy preference relation, The 3rd International Conference on Fuzzy Systems and Data Mining *Nov. 24-27, 2017, National Dong Hwa University, Hualien, Taiwan* (accepted)
[11] Technical report: DIALLO,M.S, Sall, O., Cissé M. (2017): Etude des mecanismes de personnalisation d'information, Rapport technique Université de Thies.
[12] Conference: Jelassi M.N., Ben Yahia,S., and Nguifo, E.M. (2013): A personalized recommender system based on users' information in folksonomies.. in proceding of International World Wide Web Conferences Steering Committee / ACM, WWW (Companion Volume), page 1215-1224.
[13] Conference: Jelassi M.N., Ben Yahia,S., and Nguifo, E.M (2015).Towards more targeted recommendations in folksonomies. In Social Network Analysis Mining 5(1): 68:1-68:18
[14] Conference: Bellatreche, L., Giacometti, A., Marcel, P., Mouloudi ,H., and Laurent, D. (2005). A Personalization Framework for OLAP Queries. In Proceedings of DOLAP'2005, pp. 9-18.
[15] Confernce: Bouzeghoub M. and Kostadinov D. (2005), Personnaliation de l'information: aperçu de l'etat de l'art et definition d'un modele flexible de profils. Actes de la 2nde Conférence en Recherche d'Information et Applications CORIA, 2005, p.201-218