PRO-SER: Project Search Engine & Repository Using Pattern Matching & NLP

Kiran Kumar A V[1], Abhilash D R[2], Debanjan Poddar[3], Biplav Karki[4], Dr. Vikas Reddy S[5]

[1,2,3,4] Department of Computer Science Engineering, S J C Institute of Technology, Chikkaballapur,
[5] Associate Professor, Department of Computer Science & Engineering, SJC Institute of Technology, Chikkaballapur.

---***---

Abstract - This brief synopsis on the title PRO-SER: Project Search Engine and Repository using Pattern Matching and NLP, introduces and examines the idea of academic project search engine predicated on the few recently done industry implementation and research studies, directions are supplied on the best way to enhance the search techniques for search engine and search motors with several model search engines like Google Scholar Search engine, Bing Scholar Search Engine, IEEE Xplore Search engine, Quora Search Engine, Stack Overflow Search Engine, Amazon Search Engine, Facebook Search Engines. This project report deliberates on use of python as a programming language along with Flask Library with the use of advanced Machine Learning tools and techniques along with Search techniques and algorithms like Cosine Similarity. Advanced features like Auto suggestion, and pattern matching along with Python File Handling concepts are also being incorporated to make search experience better. The design towards search engine is like the engine iterates over the data, by finding and counting words that are similar and for the recommendation after finding similar ones, the engine checks the current project title of the source and compares with the corresponding project title of the similar targets and then finds the project title that is next.

Key Words: Search Engine, Machine Learning, Cosine Similarity, Pattern Matching.

1. INTRODUCTION

There are several open-source search engines available to download and use. In this study it is presented a list of the available search engines and an initial evaluation of them that permits to have a general overview of the alternatives. In search engine market Archie was the initial research engine, that has been used to look for FTP (File Move Protocol) documents and in the other area the initial text-based search engine is called Veronica. Because large research engines include thousands and occasionally billions of pages, many research engines aren't only just looking the pages but in addition exhibit the outcome dependent on their importance. This significance is generally determined by applying various algorithms. There are now different types of research engines accessible like Bing, Google, Ask.com, Google, Alta vista etc.

For the students are very difficult to choose project because of old concepts. For the project guide and coordinator, it's very difficult to search all the previous projects. We have to open and read all the records manually that causes time wastage also. Prior to this period most Students projects are store in the Archive Room in various University, this project is then sorted out depending on the area of research, in time this work is sorted by student who are looking for a project topic. The order in which these files are kept after used can lead to a project hidden in the wrong compartment.

2. LITERATURE SURVEY

Web impact factors and search engine coverage [1]: There is an increasing amount of academic and other information on the web. There is also an increasing number of online journals as well as online versions and indices of traditional journals. A survey was conducted in order test the coverage of search engines and to decide whether their partial coverage is indeed an obstacle to using them to calculate Web Impact Factors. The results indicate that search engine coverage, even of large national domains is extremely uneven and would be likely to lead to misleading calculations.

Trends in web Based Search Engine [2]: The use of web search engines is an essential part of the ordinary life. It is difficult to underestimate the tremendous role they have for internet users. Over the years several useful web-based search engines such as Lycos, Excite, AltaVista, Google, Yahoo, Bing and the likes emerge. This work gives an insight into the trend of web-based search engine, diverse ways in which it works, and its future.

A New methodology for search engine optimization without getting sandboxed [3]: This research work implies a new methodology of Search Engine Optimization (SEO) without getting sandboxed by search engines like Google, Bing and other. In the past, the algorithm was based on the quantity of back links that a site has. This process involves in implementing safe link building techniques with link velocity as its key without compromising the on-page optimization. The latest algorithmic updates are taken in to consideration and the strategy is developed to rank for a keyword. By implementing this method, any organization can take advantage of the traffic from the search engines and have a good online presence. This paper is done based on the basic guidelines recommended by all the search engines for proper indexing without sandboxing. Hence even in the future; this method will not hinder the online progress of any business.

The Anatomy of a Large-Scale Hypertextual Web Search Engine [4]: In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http: llgoogle.stanford.edu/ To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of Web pages involving a comparable number of distinct terms.

Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co [5]: This article introduces and discusses the concept of academic search engine optimization (ASEO). Based on three recently conducted studies, guidelines are provided on how to optimize scholarly literature for academic search engines in general and for Google Scholar in particular. In addition, we briefly discuss the risk of researchers' illegitimately 'overoptimizing' their articles.

3. PROPOSED SYSTEM

This informative article introduces and examines the idea of PRO-SER: Project Search Engine & Repository. Predicated on three lately done studies, directions are supplied on the best way to enhance scholarly literature for academic search motors generally and for Bing Scholar in particular. The ability to search a specific topic for the page you are looking for is a very useful feature. However, searching can be complicated and providing a good search experience can require knowledge of multiple programming languages.

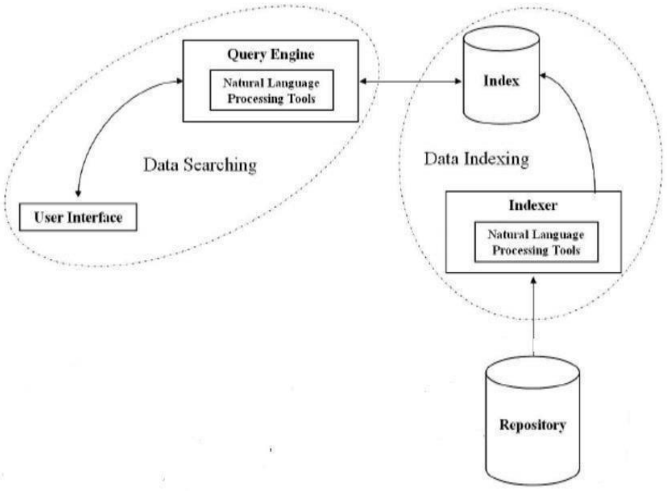


Fig – 2: System Architecture

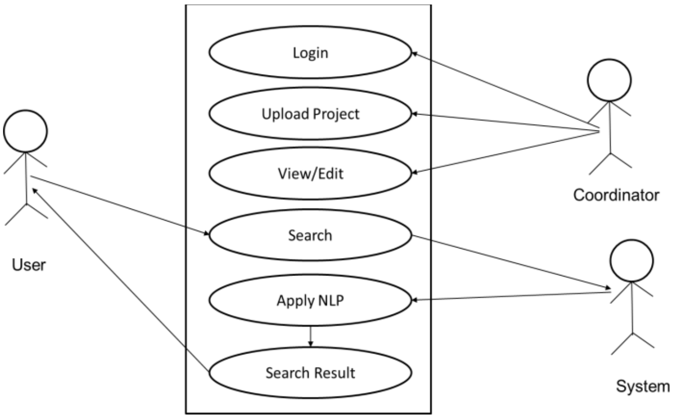


Fig – 3: Use case Diagram

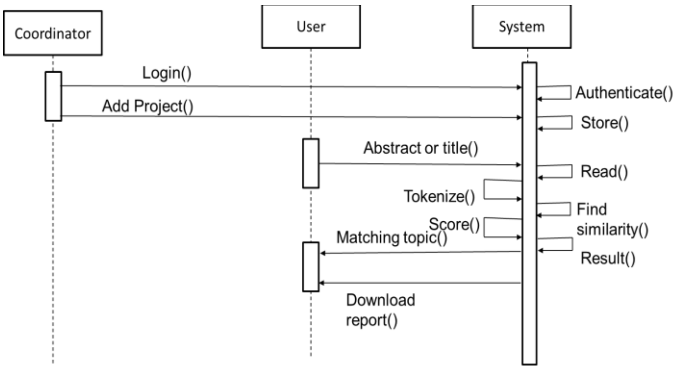


Fig – 4: Sequence Diagram

4. METHODOLOGY

In this project we are using the cosine similarity algorithm is used to calculate the similarity between the new topic with old topic. Cosine similarity calculates similarity by measuring the cosine of angle between two vectors.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Fig -1: Figure

Mathematically speaking, Cosine similarity is a measure of similarity between two nonzero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any angle in the interval (0, π] radians. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors oriented at 90° relative to each other have a similarity of 0, & two vectors diametrically opposed have a similarity of -1, independent of their magnitude.

The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

5. SIGNIFICANCE AND RELEVANCE OF WORK

The significance of this study is to move from manual documentation of projects to Computerized documentation of projects for easy retrieval, storage, accuracy and security. This research work will offer the following advantages to the various departments in the university;

1. **Reduced Storage:** The cost of commercial property and the need to store documentation for e.g., retrieval, regulatory compliance means that paper-based project storage competes with people for space within an organization. Scanning projects and integrating them into a project management system can greatly reduce the amount of prime storage space required by paper.
2. **Flexible Indexing:** Indexing paper in more than one way can be done, but it is awkward, costly and time-consuming. Images of projects stored within a project management system can be indexed in several different ways simultaneously.
3. **Improved, faster and more flexible search:** Project Management Systems can retrieve files by any word or phrase in the document – known as full text search – a capability that is impossible with paper.
4. **Improved Security:** A project management system can provide better, more flexible control over sensitive projects. Many project management system solutions allow access to projects to be controlled at the folder and/or document level for different groups and individuals. Paper projects stored in a traditional filing cabinet or filing room does not have the same level of security i.e., if you have access to the cabinet, you have access to all items in it.
5. **Disaster Recovery:** A project management system provides an easy way to back-up projects for offsite storage

and disaster recovery providing failsafe archives and an effective disaster recovery strategy. Paper is a bulky and expensive way to back-up records and is vulnerable to fire, flood, vandalism and theft.

6. **No Lost Files:** Lost projects can be expensive and time-consuming to replace. Within a Project Management System, imaged projects remain centrally stored when being viewed, so none are lost or misplaced. New documents are less likely to be incorrectly filed and even if incorrectly stored can be quickly and easily found and moved via the full-text searching mechanisms.

7. **Digital Archiving:** Keeping archival versions of projects in a project management system helps protect paper documents that still have to be retained, from over- handling.

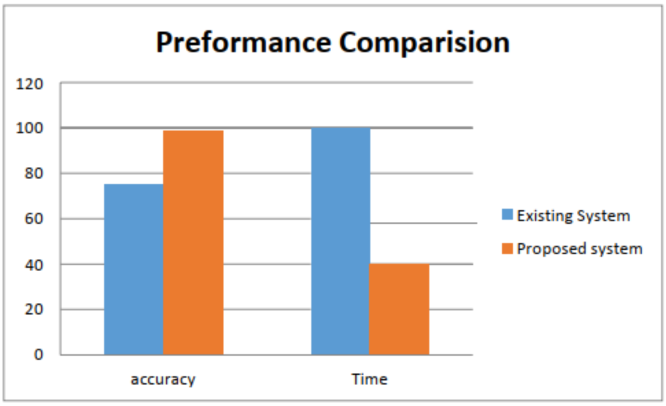
6. ADVANTAGES

▪ Our application is very effective search engine for academic project.
▪ Our application helps students to select their academic projects and also avoid the already done projects.
▪ Our application reduces the time of project selection and submission.
▪ Our application keeps large amount of database of projects which helps the students.
▪ Our application helps project coordinators to digitally handle their many tasks regarding the project work.

7. APPLICATIONS:

In Educational Institutions and research centers (across departments): The scope and application of this project pertains to the education field, academia and higher educational institutions like universities and colleges ranging from undergraduate till PhD. Sometimes there will be confusion among the students as well as project coordinators while choosing and allotting the projects for their final year project or research work, as to whether this is a repetition or a copy of an already existing project. Therefore, this project solves that problem by digitizing the entire process, and acts as a search engine where students can check their topic with the already existing topic and the percentage of match. This also digitizes the working of project coordinator and makes his/her work seamless and smooth in this pandemic and future in which working safely & going digital is an utmost priority.

8. PERFORMACE EVALUATION



In this section, we compare the performance of system with against time and accuracy of the System against all the search cases. From figure above, first of all, it can be seen from the figure that the average time consumption of proposed system is better than of existing systems. It is because we use Machine learning which will improve the software over the time as it learns from the interaction of the user and the system.

In summary, compared to the other systems, proposed system has the advantages of low time consumption and more accurate for results.

9. CONCLUSIONS

The “PRO-SER” search engine will give the best of the support to the students and project coordinator by giving them the result about the percentage of match with the existing projects done by the seniors and an overview of whether they can go ahead with their project or is it a redundant topic. It also acts as a repository of executed projects for juniors to find the gaps in existing projects and improve and work upon them by downloading the reports and resources. It also gives a redirecting link option to get the new paper on the project to get checked on plagiarism. The project coordinator will have the full control to manage the search engine database and repository and the project guides will have certain control than students for few features that are required.

REFERENCES

[1] M. Thelwall, web impact factors and search engine coverage, Journal of documentation 56(2) (2000) 185-189.
[2] Koyoro Shadeo, Trends in web Based Search Engine ‘Journal of emerging trends in computing and information Sciences’ Vol 3, No-6, June 2012, ISSN – 2079-8407.
[3] Dr. S. Sarvankumar, A New methodology for search engine optimization without geeting sandboxed ‘International journal of Advanced research in computer and communication Engineering Vol 1, issues, Sept 2012, PP- 472-475.
[4] Sergey Brin and Lawrence Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,”. [Online].Available: http://infolab.stanford.edu/~backrub/google.html
[5] Meng Cui, Songyun Hu, “Search Engine Optimization Research For Website Promotion, Transport Management College, Dalian Maritime University, Dalian, 116026, China”. [Online]. http://www.ftsm.ukm.my:8080/kt/attachments/article/69/06113701.pdf
[6] Joeran Beel, Bela Gipp, Erik Wilde. UC Berkeley School of Information.” Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co.”[Online].Available: http://www.beel.org/files/papers/2010-ASEO--preprint.pdf.
[7] "10 SEO Techniques All Top Web Sites Should Use”. [Online]. Available: http://freelancefolder.com/10-top-seo-techniques/.
[8] “Search Engine Optimization”. [Online]. Available: http://en.wikipedia.org/wiki/Search_engine_optimization.
[9] “Ultimate Benefits of SEO”.[Online]. Available: http://www.nextsbd.com/seo/benefits-of-seo.php.