Intelligent Indexing and Sorting Management System – Automated Search Indexing and Sorting of Various Topics

Apoorva Ganapathy[1*], **Takudzwa Fadziso**[2]

[1]Senior Developer, Adobe Systems, San Jose, California, **USA**
[2]Institute of Lifelong Learning and Development Studies, Chinhoyi University of Technology, **ZIMBABWE**

[*]Corresponding Contact:
Email: apganapa@adobe.com

ABSTRACT

An issue that the majority of the databases face is the static and manual character of indexing activities. This traditional method of indexing and sorting different topics is confirmed to shake the dataset performance somewhat, making downtime and a potential effect in the presentation that is normally addressed by manually indexing operations. Numerous data mining methods can accelerate this process by using proper indexing structures. Choosing the appropriate index generally relies upon the kind of operation that the algorithm performs against the dataset.

Topic indexing is the operation of recognizing the principal topics covered by a document. These are helpful for some reasons: as subject headings in libraries, as keywords in scholarly articles, and as hashtags on social media platforms. Knowing a document's topic assists individuals with deciding its importance quickly. In any case, assigning topics manually is a tedious and redundant task. This paper shows the best way to create them automatically in a way that contends with manual indexing done by humans.

This paper also talks about the issues and the techniques for identifying applicable data in a huge variety of documents. The contribution of this thesis to this issue is to foster better content analysis techniques that can be utilized to describe document content with automated index terms. Index terms can be used as meta-data that defines documents and is utilized for seeking various topics. The main point of this paper is to show the way toward creating an automatic indexer which analyzes the topic of documents by integrating proof from word frequencies and proof from the linguistic analysis given by a syntactic parser. The indexer weighs the expressions of a document as per their assessed significance for depicting the topic of a given document based on the content analysis.

Key words:
Automated indexing, manual indexing, artificial intelligence, content management system, sorting

12/20/2020 **Source of Support:** None, **No Conflict of Interest:** Declared

INTRODUCTION

The empowering thought that one-day computers will understand human language stays a slippery dream, a long way from the real world. Tasks that require understanding normal language, for example, intelligent machine translation and text summarization, have been researched for a long time, however, once in a while settled at a level similar to human performance. Researchers should deal not just with the inconceivable complexity of our language, yet also with the deeply abstract nature of the tasks. Deciding whose interpretation is more exact or whose indexing is preferable is in some cases significantly harder than creating one.

Automatic indexing is extraordinary, especially, in terms of topic indexing due to the dynamic nature of the text. It is a simpler task. But the real question "What is this document about?" expects some comprehension of a regular language, however, confines the answer to a fewer bunch of expressions that depict the document's fundamental topics. Thorough interpretation isn't really needed: even human indexers will in general skim the content instead of completely go through it and understand it (Ganapathy, 2018). Topic indexing is additionally simpler to assess. Those topics on which most of the people concur should certainly be the right ones. Various votes change an emotional view into an ideal answer.

In the significant sections of this paper, the methodology, types, and approaches of existing indexing and sorting frameworks have been introduced. Of all the operations needed in information retrieval, the most significant and likely the most difficult one comprises assigning suitable terms and identifiers appointed for addressing the content of the collection things. This task, known as indexing, is ordinarily performed manually via skilled human indexers. In current conditions, the indexing tasks can be performed automatically. This paper is centered on the procedures utilized for automatic indexing, and the impact and performance of these methods.

The fundamental indexing task is first mentioned, followed by an analysis of manual and automatic indexing. Essential procedures are then inspected for choosing suitable index terms and for assigning weights to the terms as per their assumed value for content identification and overall analysis. A simple automatic indexing methodology is then proposed, as well as refinements comprising of the utilization of term expressions. The usage of different methodologies and strategies in automatic indexing is likewise presented. At last, every section is summarized in a detailed conclusion that examines the overall output to show the adequacy of the proposed indexing and sorting procedures for various topics.

LITERATURE REVIEW

Progressively, various sorts of data resources are being made accessible on the internet. Current search engines yield great results for specific searching, indexing, and sorting responsibilities however are unacceptable for the applied or domain-based tasks that ask for expanded accuracy and quality research, majorly related to academic domain or real-world demands (Keyser, 2012).

Studies conducted on intelligent indexing started with the access of digital content during the 1950s (Luhn, 1957) and keeps on being a difficult subject to solve for various business needs and purposes. For the academic outline of automated indexing, Stevens (1965) and Jones (1972) covered the early time of automated indexing, and Lancaster (Lingle, 2005) covered the later developments. A similar term to this concept is computer-aided indexing, sometimes called computer-supported indexing, in which a human indexer makes decisions based on an idea given by the computer (Paruchuri & Asadullah, 2018). A similar system (Martinez-Alvarez, et al., 2012) is additionally developer that alludes to an approach in which simply those suppositions obligated to be found right are trained normally, while human indexers take the more unpredictable choices. But, such topic index terms need extensive research sources for providing a quality content analysis. Due to the steadily expanding amount of documents, perceived objectives of bibliographic systems, like indexing every document on a particular topic would be tedious and might left behind in the process. For instance, a research study (Golub, 2016) shows that topic access isn't talked about efficiently, meaning that advanced information systems are applied to an extremely restricted degree, consequently preventing quality search across them.

MANUAL INDEXING VS. AUTOMATED INDEXING

Before going to a depiction of automatic indexing strategies, it could be helpful, to sum up, some traditional manual indexing practices. As a rule, a controlled indexing language is utilized in which a single standard term or expression addresses a wide range of related terms.

Arguments w.r.t Manual Indexing

Backlogs in libraries not generally due to the increasing number of fresh arrivals in the catalog area; they might be because of numerous different variables: restricted budget, which can cause understaffing, slower classifying practices, and so on. Indexing can for sure assume a part in this as well. It requires a couple of minutes for a cataloger to discover what the specific topic of a document is and which terms are the best interpretation of that topic. If a cataloger needs five minutes to index a document and another 15 to make another inventory record, he could save a fourth of his time if indexing terms were added automatically. These sorts of computations might be appealing, particularly to managers, when it comes to cutting down budget plans.

As a result of the amount of time it takes to find reasonable indexing terms, indexing will cost a great deal of money as well. The best way to make it less expensive – or even moderate– is to outsource it to countries where skilled workers accomplish this task for considerably lower wages. Be that as it may, it is additionally costly to create, update and gain proficiency with a controlled vocabulary. At any rate, one colleague should commit part of their time to update the thesaurus, and then he/she should consult different colleagues or experts every once in a while to seek their recommendation.

Arguments w.r.t Automated Indexing

Controlled vocabularies make a differentiation among preferred and non-preferred terms. Among every single possible synonym, a decision is made for that term which is the best interpretation of the thought behind it: the preferred term.

Any remaining synonyms identify with that term. Automatic indexing brings about a heap of words with no sort of connection between them. Morphological and semantic variations are overlooked (Ganapathy, 2017). Also, in case of words are extracted from a book via automatic indexing, they will be in the language of the source text. This implies that we are obliged to use search terms in however many languages as we probably are aware to find enough texts that were written in them.

It isn't really evident that automatic indexing separates all words from the writings in the specific morphological structure wherein they appear in the content. Various methods are developed to change them into more controlled terms. Normal language indexing is against indexing with controlled vocabularies, for example, thesauri, or grouping codes. Automatic indexing projects can think about file terms with controlled vocabularies or classifications and recover file terms from them.

METHODOLOGY OF AUTOMATED INDEXING AND SORTING

Usually, automatic topic indexing commonly goes through a few significant steps. The initial step is to plan what documents are to be trained to make an appropriate representation for file processing. The mentioned process is identical to the processing and training of documents for information retrieval.

Pre-processing

In the first step, a sequence of words occurring in the document is made according to the tokenization. Likewise, the full stops, commas, and any other sort of punctuation is removed altogether. Next, the words that will in general convey less importance are completely removed such as, pronouns – commonly referred to as stop-words. This type of reports is otherwise called a bag-of-words approach.

Another developed model named as n-gram model is also used for similar purposes, for instance, when phrases should be removed in the process of indexing or while marching certain strings compared to the terms with more than one word. Word in this model might be unigrams, bigrams, and trigrams, and so on. Additionally, NLP (Natural Language Processing) methods might be used that implies removing the affixes of a word – for instance, unlawfully might be decreased to its affix (lawful) whereby its prefix and its suffix are removed (Ganapathy, 2015). The thought behind is that those words having a similar pattern have similar importance. Likewise, grammatical form taggers and syntactic parsers can likewise be applied.

Term Weighting

The next significant step is deciding the significance every term has in order to understand the relevance of a particular document. The selected term can either be a single word or a collection of phrases - depending upon the task at hand. A weight is assigned to every term as a quantifier. Various approaches can be used here in the domains of statistics and heuristics. For statistical principles, the words occurring a lot of times both in the current text and in any other documents in the file, are not likely to be characteristic of the topic, and the other way around. This is also known as TF-IDF commonly. It integrates term frequency, where the weight of the term is viewed as relative to the number of times it occurs in the text, with TD-IDF, where the weight of a certain term is a reverse part of a document that actually contain that particular word.

Various Representations

After the completion of above two steps, every document is changed in a sequence of terms and respective weights as described in the 2nd step. There are potentially two approaches – one is Vector Representation and second one is known as String Matching.

Vector Representation is a methodology where the aftereffect of the initial steps is changed into vector space. Every term connected to its respective weight is addressed as one particular dimension (Ganapathy & Neogy, 2017). Vector Representation leverages advanced numerical algorithms to make it easier and simplify the arrangement of text.

String Matching is a relatively less commonly used approach that defines the methodology followed with respect to the terms from a text and defining ideas from a particular language. In intelligent automated indexing approach, a similar idea is followed to cater to list terms. Every idea can be referred by the terms removed from the document that has been manually indexed. This representation should also be changed into the form of vectors when the files are indexed to enable a detailed analysis.

Allocation of Index Terms

This final step leverages either estimations with the help of vectors or matching relevant strings of the different terms having target list are performed. Normally, a bunch of terms is the main outcome, out of that the best performing terms are chosen with the help of statistical and heuristic approaches. The main approach is always to select the best term on the off chance that the term is ranked amongst the best 5 terms at the top and also happens to occurs in the document title. Another approach is to choose the top 3 candidate terms having the highest weight.

DIFFERENT TECHNIQUES OF INDEXING

An essential meaning of indexing was given in 1988 (Salton, 1988) as the support for precision by collecting and storing information. The precision assistance is carried out by the utilization of different strategies and techniques of indexing. As mentioned earlier, users now look for precision in the indexing to overshadow human performance. The indexing process generally has a consolidation mechanism that allows using ideas from different domains. It’s been defined that there are numerous indexing techniques used in the research and academic disciplines (Vadlamudi, 2020). In this section, we will talk about every strategy in detail and also describe the way they work.

Inverted Files

This technique is characterized as the main part of an indexing algorithm that looks through data has the objective of optimizing the speed of querying. It means discovering files if a specific word is found. Afterwards, at that point, the following approach is creating an index. The created file assumes a part of indexing arrangements of particular terms in each file. Then it is rearranged, resulting in a transformed file. Consecutive iteration is generally needed to access the index.

A study (Belew, 2006) proposed that a process of repetition needs to be executed for every term to enable the verification of a file that relates to a certain query. Actually, a certain time and memory is needed in the presentation of such case does not have a part of being practical. Nonetheless, the structure of files that are created sequence the documents word by word. It is done sequencing the document and each word, where the terms must be inverted on each file. To execute a successive iteration of the whole document, we take an arrangement of files. Later on, we assume that each file is allocated a list that includes some keywords. The set of such keywords is also called attributes (Vadlamudi, 2019). We additionally believe that there are relevance weights for each keyword. Having covered all these points, the arranged list of catchphrases will be a rearranged document. Each keyword will be connected to the files that contain the particular term.

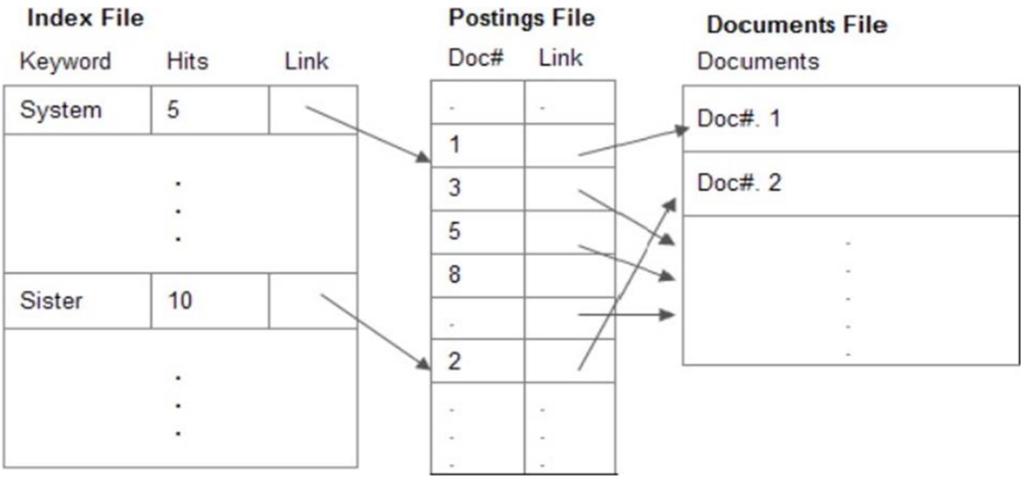

Figure 1: Methodology and Structure of Inverted Files

Suffix Trees

It is a technique that is based on the suffixes of a text that have a connection with them as well as their attributes as the positions present in the content. Basically, compressed tries are somewhat related to a data structure called TRIE. Thought behind this technique had been presented by Weiner (1973). The application of this technique is applied while tackling different issues related to strings - happening in free content search and word processing. These trees are normally used in computer science and other different areas related to word processing.

In the case of string S and length N, the definition of a suffix tree should meet a few parameters (Paruchuri & Asadullah, 2018). Initially, there should be actually n leaves indexed within the suffix tree. Each parent-node should have at least two children. The labels assigned to the edges are mapped with substrings (Ganapathy, 2016). Any two edges starting from a parent-node must include string names that begin with an alternate character. Such conditional parameter implies that it isn't feasible for a suffix parent-node being a prefix parent-node (Ganapathy, 2016). Ultimately, a suffix is formed by the set of string extracted from the connection of substrings.

Signature Files

It is an indexing technique that typically makes a filter. Bloom filter is a well-known example that keeps each one of these existing files connected to the problem faced by a user and wants to store the files that don't relate with the rules. This process is accomplished through the formation of a signature for each file that is usually an adaptation of a hash coding. In this way, any signature is a replica of a file that has already been mapped. These files are produced through 2 fundamental techniques: Word Signatures and Layered Coding.

First approach includes the technique of hashing parameters that are fundamentally expressions of signature file. These patterns then form the signature file through connection. Then again, Superimposed Coding of mapped signatures includes the process of hashing every term to a certain bits, assume a signature file (S) is mapped with a number of bits (B). Layered Coding is performed through OR operator on the most recent files for the formation of the signature files. Figure 2 demonstrates a visualization of a file being prepared through making a sequence of exceptional words. A sequence of usual words should also be made to remove any kind of uncommon words to avoid any extra training of documents. Common words as such have no impact on characterizing the whole document in general.

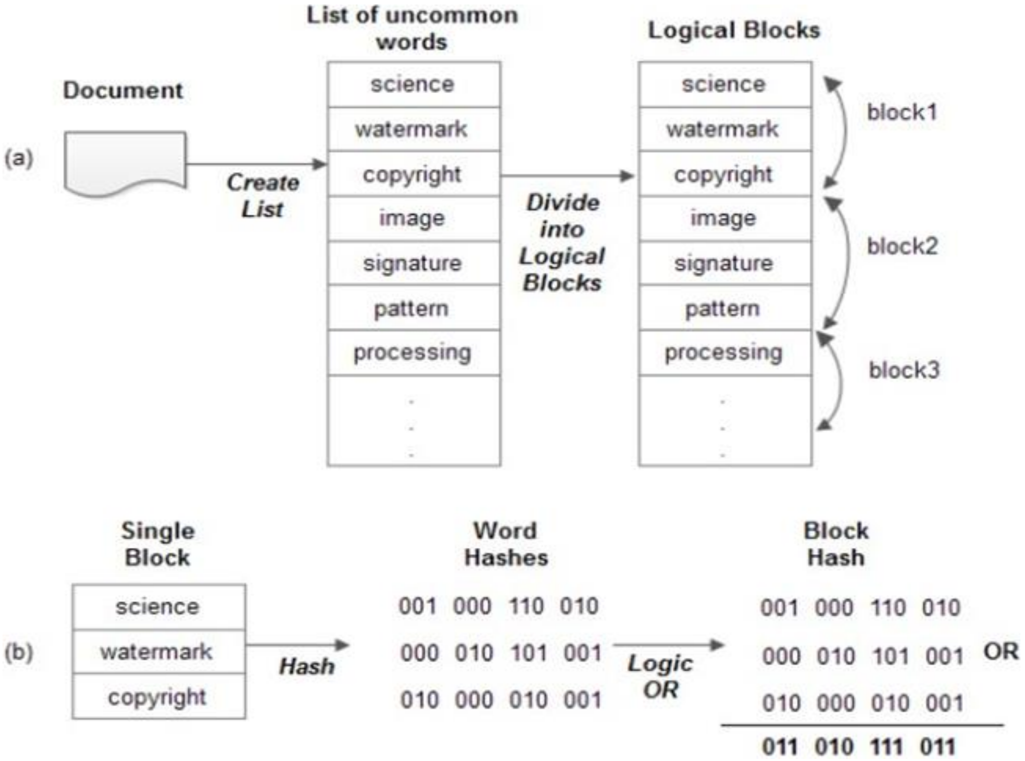

Figure 2: Representation of Signature Files Framework

Types of Topic Indexing Sorting

Term Assignment

Term assignment refers to the primary topics in a document using terms from a pre-characterized controlled vocabulary, for example, a domain-specific thesaurus. These terms don't really occur inside the document. Subject indexing is a more normal term for a similar task with regards to library science (Paruchuri, 2018). A controlled vocabulary records ideas applicable to a given domain with the help of two types of terms: descriptors and non-descriptors. Descriptors are the preferred terms for referring to the ideas. Non-descriptors, likewise called entry terms, are generally synonyms for the comparing descriptors. Leveraging descriptors in indexing improves consistency.

Most documents belong to a similar domain, similarly, as paper articles are classified into distinct domains like Politics, Fashion, Sport, Ads. Articles that have a connection with at least two classifications are unusual. A class can be allocated based on the document's content as well as on its style. For instance, articles in the class Ads are presented by short sentences, abbreviations, and numbers. Automatic content classification lies outside of the extent of this task since it requires a different methodology in comparison to automatic term assignment (Paruchuri, 2019).

Keyphrase Extraction

Key phrase extraction normally means an assignment performed by an algorithm that chooses noticeable expressions occurring in a document. Keyphrase indexing refers to a wider scope of tasks than keyphrase extraction and term assignment, where the source of phrasing isn't limited. For instance, scholar distributers urge their authors to allocate use chosen keyphrases. These phrases could or could not occur in the content.

Two tasks that are like keyphrase extraction are terminology extraction where each domain-specific expression should be separated from a document, and full-text indexing, where the whole vocabulary of the document is changed into an index. Be that as it may, terminology extraction can be addressed with keyphrase extraction techniques.

Tagging

Like key phrases, tags can be chosen uninhibitedly. There are no proper rules. Users choose which terms and the number of terms should be assigned and perform the task mainly for their own advantage. Tagging is leveraged mostly on social media platforms that have user-oriented content, for example, writing and posting to a blog platform, internet-based bookmarking services, and data-sharing websites. Frequently a few users tag similar item and their tags are converted into a single set (Vadlamudi, 2019).

As of late, researchers started creating techniques to assign tags automatically. Such techniques can give ideas that can work with tagging and urge more users to use tags to expand the reach of content to a greater extent. Automatic tagging, or auto-tagging, gets tags from various sources, including the wording of the actual document, different documents owned by the users (stored on their computer), and tags recently used in similar documents.

Table 1: Various Tasks of Automated Indexing and Sorting

Task Name	**Description**
Text Categorization	Not many general classifications, similar to Politics or News, are assigned from normally a minimal vocabulary.
Term Assignment	Fundamental topics are communicated using terms from a huge set of vocabulary, for example, a thesaurus.
Keyphrase Extraction	Main topics are communicated utilizing the most prominent words and expressions in a document.
Terminology Extraction	All domain-specific words and expressions are removed from a file.
Full-text Indexing	All words and expressions, sometimes excluding stop words are removed from a document.
Keyphrase Indexing	An overall term, which means both term assignment and key phrase extraction.
Tagging	The user characterizes however many topics as required. Any word or expression can fill in as a tag. Applies basically to websites and social media platforms.

Conclusion

Automated indexing and sorting, or topic indexing (identify the primary topics in a record) are usually performed by people. Libraries use skilled and proficient indexers. On the internet, where a lot of documents are created every day, topics are given by volunteer taggers, who put together pieces of the internet that are important to them. This paper has explored how to perform topic indexing automatically. The fundamental theory is that with access to domain knowledge and general semantic information, computers can index just like people. It likewise sums up the methodology and types used to test this theory, leading to its approval through the development and assessment of the various methods.

Study discussed in this thesis has obviously accomplished a basic understanding and exploration of techniques related to automated indexing. It also has provided insights into the development of different strategies. Also, this paper presented thoroughly described patterns that make up these techniques and how they work. Because of the complexities related to precision, existing frameworks and approaches have not been enough tried, and hence information about their value for operational frameworks is by all means defective. Nonetheless, extensive research and refined methodologies are required to experimentally test the hypothesis and derive the most suitable analysis approaches for various topic indexing tasks.

References

Belew, R. K. (2006). Adaptive information retrieval, in Machine Learning in Associative Networks. Michigan: University of Michigan Press, pp. 78-83.

Ganapathy, A. (2015). AI Fitness Checks, Maintenance and Monitoring on Systems Managing Content & Data: A Study on CMS World. *Malaysian Journal of Medical and Biological Research*, *2*(2), 113-118. https://doi.org/10.18034/mjmbr.v2i2.553

Ganapathy, A. (2016). Blockchain Technology Use on Transactions of Crypto Currency with Machinery & Electronic Goods. *American Journal of Trade and Policy*, *3*(3), 115-120. https://doi.org/10.18034/ajtp.v3i3.552

Ganapathy, A. (2017). Friendly URLs in the CMS and Power of Global Ranking with Crawlers with Added Security. *Engineering International*, *5*(2), 87-96. https://doi.org/10.18034/ei.v5i2.541

Ganapathy, A. (2018). Cascading Cache Layer in Content Management System. *Asian Business Review*, *8*(3), 177-182. https://doi.org/10.18034/abr.v8i3.542

Ganapathy, A., & Neogy, T. K. (2017). Artificial Intelligence Price Emulator: A Study on Cryptocurrency. *Global Disclosure of Economics and Business*, *6*(2), 115-122. https://doi.org/10.18034/gdeb.v6i2.558

Golub, K. (2016). Potential and Challenges of Subject Access in Libraries Today on the Example of Swedish Libraries, *International Information & Library Review*, *48*(3), 204-210, https://doi.org/10.1080/10572317.2016.1205406

Jones, K. S. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1), 11-21. https://doi.org/10.1108/eb026526

Keyser, P.d. (2012). Indexing: From Thesauri to the Semantic Web. 1st ed. Chandos Publishing. https://2lib.org/book/2337706/ba39da?id=2337706

Lingle, V. A. (2005). Indexing and Abstracting in Theory and Practice. *Journal of the Medical Library Association*, *93*(1), 133. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC545136/

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, 1(4), 309-317. https://doi.org/10.1147/rd.14.0309

Martinez-Alvarez, M., Yahyaei, S., and Roelleke, T. (2012). Semi-automatic Document Classification: Exploiting Document Difficulty. *Lecture Notes in Computer Science: Advances in Information Retrieval*, 7224, 468-471.

Paruchuri, H. (2018). AI Health Check Monitoring and Managing Content Up and Data in CMS World. *Malaysian Journal of Medical and Biological Research*, *5*(2), 141-146. https://doi.org/10.18034/mjmbr.v5i2.554

Paruchuri, H. (2019). Market Segmentation, Targeting, and Positioning Using Machine Learning. *Asian Journal of Applied Science and Engineering*, *8*(1), 7-14. Retrieved from https://journals.abc.us.org/index.php/ajase/article/view/1193

Paruchuri, H., & Asadullah, A. (2018). The Effect of Emotional Intelligence on the Diversity Climate and Innovation Capabilities. *Asia Pacific Journal of Energy and Environment*, *5*(2), 91-96. https://doi.org/10.18034/apjee.v5i2.561

Salton, G. (1988). Automatic Text Processing, in the Translation Analysis and Retrieval of Information by Computer. Washington: Cambridge, Addison-Wesley Publishers, 3(2), pp. 45-70.

Stevens, M. E. (1965). Automatic Indexing: A State of the Art Report, Monograph 91. Washington, D.C.: National Bureau of Standards. https://digital.library.unt.edu/ark:/67531/metadc70462/

Vadlamudi, S. (2019). How Artificial Intelligence Improves Agricultural Productivity and Sustainability: A Global Thematic Analysis. *Asia Pacific Journal of Energy and Environment*, *6*(2), 91-100. https://doi.org/10.18034/apjee.v6i2.542

Vadlamudi, S. (2019). How Artificial Intelligence Improves Agricultural Productivity and Sustainability: A Global Thematic Analysis. *Asia Pacific Journal of Energy and Environment*, *6*(2), 91-100. https://doi.org/10.18034/apjee.v6i2.542

Vadlamudi, S. (2020). The Impacts of Machine Learning in Financial Crisis Prediction. *Asian Business Review*, *10*(3), 171-176. https://doi.org/10.18034/abr.v10i3.528

Weiner, P. (1973). Linear pattern matching algorithms. 14th Annual Symposium on Switching and Automata Theory (swat 1973), 1-11, https://doi.org/10.1109/SWAT.1973.13

--0--

How to Cite:

Ganapathy, A., & Fadziso, T. (2020). Intelligent Indexing and Sorting Management System – Automated Search Indexing and Sorting of Various Topics. *Engineering International*, *8*(2), 101-110. https://doi.org/10.18034/ei.v8i2.554