TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 4, August 2020, pp. 1874~1883
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i4.15001

Semantics-based clustering approach for similar research area detection

Marion Oluwabunmi Adebiyi[1], Emmanuel B. Adigun[2], Roseline Oluwaseun Ogundokun[3], Abidemi Emmanuel Adeniyi[4], Peace Ayegba[5], Olufunke O. Oladipupo[6]
[1,3,4,5]Department of Computer Science, College of Pure and Applied Sciences, Landmark University, Nigeria
[1,2,6]Department of Computer Science, Covenant University Ota, Nigeria

Article Info

Article history:

Received Dec 6, 2019
Revised Jan 22, 2020
Accepted Mar 18, 2020

Keywords:

K-means clustering
Latent semantic indexing
Nigeria University
Ontology-based preprocessing
Semantics-based clustering

ABSTRACT

The manual process of searching out individuals in an already existing research field is cumbersome and time-consuming. Prominent and rookie researchers alike are predisposed to seek existing research publications in a research field of interest before coming up with a thesis. From extant literature, automated similar research area detection systems have been developed to solve this problem. However, most of them use keyword-matching techniques, which do not sufficiently capture the implicit semantics of keywords thereby leaving out some research articles. In this study, we propose the use of ontology-based pre-processing, Latent Semantic Indexing and K-Means Clustering to develop a prototype similar research area detection system, that can be used to determine similar research domain publications. Our proposed system solves the challenge of high dimensionality and data sparsity faced by the traditional document clustering technique. Our system is evaluated with randomly selected publications from faculties in Nigerian universities and results show that the integration of ontologies in preprocessing provides more accurate clustering results.

Corresponding Author:

Roseline Oluwaseun Ogundokun,
Department of Computer Science,
Landmark University,
Kwara State, Nigeria.
+2347036261504
Email: ogundokun.roseline@lmu.edu.ng

1. INTRODUCTION

Background of study

For the purpose of research, topics and relevant articles from digital libraries and online databases are sought for in order to gain a better understanding. These sources are useful in the retrieval of relevant articles by renowned researchers [1]. Citations (a document cites another), bibliographic coupling (documents sharing a reference in their bibliography) and co-word linkages (documents share certain words) are some of the existing methods that have been used to identify textual documents that are necessary for research. Traditional document clustering approaches do not sufficiently capture semantic relations between keywords leading to ambiguity and high dimensionality thereby reducing the accuracy of clustering results [2, 3]. Traditional clustering methods often ignore semantic relationships or connections between words and there produce inaccurate representations of such articles. Text data carry high level semantic information and diverse vocabulary, hence the need for text clustering techniques to improve quality.

When a prospective researcher embarks on a research endeavor, the researcher must conduct a review of related works already documented in literature in that field for a robust research publication. The researcher would This study utilizes a dataset containing bibliographic information of Nigerian researchers as a case study and applies co-word linkages to determine documents in the same field. The goal of this study is to develop a semantics-based clustering framework for detecting similar research areas. The system determines the suitability of an author to a research field based on the semantic similarity between an author's work and previous works. This would be particularly useful for researchers who intend to locate researchers within their research field and similar research field. The detection of similar research areas based on keywords could prove beneficial to tertiary institutions of learning and research center. The research networks created is intended to enhance the prospect of research collaborations in the continent. According to [4-6], the existing methods that have been used to identify textual documents that address a particular subject matter include:

- Citations (a document cites another)
- Bibliographic coupling (documents sharing a reference in their bibliography)
- Co-word linkages (documents share certain words)

The problem

Traditional document clustering approaches do not sufficiently capture semantic relations between keywords leading to ambiguity and high dimensionality thereby reducing the accuracy of clustering results [7]. Existing keyword matching techniques can be significantly improved by integrating semantics in document similarity computation. The detection of similar research areas based on keywords could prove beneficial to tertiary institutions of learning and research centres. This study intends to utilize a dataset containing bibliographic information of Nigerian researchers as a case study. We expect the research networks created to enhance the prospect of research collaborations in the continent

The proposed solution

In this study, an automated similar research area detection system is proposed that generates similar research areas and publications to that of a prospective researcher. Based on the expertise of the prospective researcher, the system automatically assigns prospective researchers to already existing researchers in the same or similar research field. This is done using the similarity score obtained using the LSA model and cosine similarity to calculate the semantic similarity between them. The proposed system is built for researchers to detect similar research areas depending on their research preferences. The system is accessible as a web application.

Therefore, the aim of the research was to develop a semantics-based clustering method for detecting similar research areas using Nigerian publications as a case study and to achieve this aim, the following objectives were carried out;

- Creation of a dataset
- Develop a framework for similar research area detection
- Implement a prototype for the proposed framework
- Validation and evaluation of the proposed approach

2. LITERATURE REVIEW

A review of existing semantic clustering techniques already document in literature is outlined below [8] presented a deep hypergraph model for sentiment classification and online reviews. The model used a hierarchical clustering algorithm to discover semantic cliques. The model, tested with movie reviews and product reviews (books, DVD, electronic and kitchen) was compared with seven other methods of sentiment classification and results showed the model outperformed all other methods in all cases. Also, [9] used semantic clustering to locate and access web documents. The text corpus is pre-processed, stemming is performed using the WordNet ontology. The term frequency-inverse document frequency algorithm was used to construct a feature matrix. Hierarchical agglomerative clustering was used to perform clustering on the feature matrix. The approach used, improved the accuracy of the clusters generated. The drawback is that Hierarchical agglomerative clustering is not suitable for large datasets. Similarly, [10] performed enhanced semantic clustering with the WordNet ontology. The text corpus was pre-processed with WordNet ontology to perform word sense disambiguation. The term frequency-inverse document frequency technique is used to derive a feature representation of the words in the text corpus. The K-Means clustering algorithm is applied to cluster the feature vectors [11]. The pre-processing method used eliminated the dimensionality problem encountered in traditional document clustering. The limitation is that the K-Means clustering algorithm suffers from the local optima problem. In [12] as well used Semantic clustering to solve the topic drift problem in information retrieval systems. Search snippets are preprocessed and extract the longest common subsequence between two snippets by GST. Evaluate Word similarity using HowNet ontology and construct a lexical chain to select features of snippets. A feature vector is constructed and evaluate snippet similarities.

The Improved Chameleon algorithm used for clustering the feature vectors improved the within-class density and between-class variation of the cluster labels. However, the Chameleon algorithm has a high time and space complexity which does not make it suitable for high dimensional datasets.

In [13] used semantic clustering to cluster search result documents based on the semantics of retrieved documents. A search engine was queried to retrieve information, the search engine results were then pre-processed, and extraction of features was carried out. The features were enhanced using concepts from an ontology and semantic network. A dissimilarity matrix of the documents was created using the Floyd-Warshall algorithm [14]. The Hierarchical agglomerative clustering algorithm is used to cluster the feature vectors in the similarity matrix. High precision results were obtained as the approach outperformed existing approaches for web search result clustering. Hierarchical agglomerative clustering can be computationally expensive to use for large datasets. In [15] also proposed the use of semantic clustering to determine similar text documents. The text corpus is extracted and pre-processed. The term frequency-inverse document frequency algorithm is used to identify frequently occurring terms and construct a document matrix. A domain ontology is constructed from the text corpus to provide a vocabulary for filtering relevant terms. A Fuzzy equivalence relation is used to determine the level of membership of terms in the text corpus. Singular value decomposition is used to transform the document matrix into a concept space. Bisecting K-means algorithm is used to perform clustering of the concept space. The use of a domain ontology in the pre-processing stage improves clustering results. The limitation of the method is the performance of the method is entirely dependent on the quality and comprehensiveness of the ontology used. In [16] as well used semantic clustering to classify customer reviews. The text corpus is extracted by crawling customer review websites and then it is pre-processed. Ontology is used to generate a concept mapping in the text corpus. Euclidean distance metrics are used to calculate the similarity of sentences in the bag of words vector space model. The modified K-Means algorithm is used to cluster the bag of words. Experimental results revealed that the accuracy of the clusters generated is increased with the use of ontology in the pre-processing stage.

Further methods for identifying existing collaborations between various researchers from various publication databases are presented below [17] developed a co-authorship network to reveal the interactions between researchers. A system for selecting a collaborator with similar research interests for joint research was modeled as a link prediction problem. Authors with similar known features were determined using Cosine similarity computed on a vector constructed to model the descriptive statistics of various research activities. The co-authorship network was determined using the hierarchical clustering of research interests in various co-occurrence networks. They also used logistic regression with lasso regularization on normalized feature vectors. The disadvantage of this approach is that information retrieval was used to obtain data from the bibliography database and the semantic meaning of the terms was not taken into consideration.

A novel architecture was proposed by [18] for joining multiple bibliographic sources to identify common research areas and relationships between authors and their publications. The scientific publications were retrieved from various bibliographic sources using APIs and Linked data practices, the data is analyzed to provide a structure and if there is no explicit structure, the data model is produced using web scraping. An ontology mapping model is used to unify the data from different bibliographic sources, and data disambiguation is performed to eliminate data inconsistencies and duplications. A vector space model of the retrieved information is generated using the TF-IDF algorithm. The K-means clustering algorithm using the Cosine Similarity measure was used to automatically discover similarities and group the authors into their research areas. The information retrieval method used did not utilize semantics in retrieving the information. In [19] proposed a method for determining collaborations between university research and industry research. A heterogeneous social network [20] was constructed to describe the relationship between researchers and companies, a company and a researcher are deemed to have a relationship if they have co-authored academic articles, co-participated in projects or co-invented patents. A dataset [21, 22] is created for researchers who have directly collaborated with companies before, it is assumed that researchers within the same domain as the researchers in this dataset can collaborate with them. Company similarity is also used to determine potential collaborations, companies that have worked on similar patents, articles, and projects. Keywords are extracted from company technological documents, pre-processing is then performed using tokenization, stop words removal, normalization and stemming, the vector space model is used to index the extracted keyword frequencies. The cosine similarity measure is used to determine the similarity of the different companies using their keywords, the top K most similar companies are stored as neighbor companies and researchers with connections to a neighboring company are potential collaborators of its neighboring companies. The approach used by [23] to predict potential research collaborations involves the use of the online social network [20] to determine the co-authorship network. The latent dirichlet allocation (LDA) algorithm is used to model a set of topics from a document corpus consisting of authored papers, these are then represented in a K-dimensional vector. LDA is also used to determine the content similarity of

the papers. The weighted support vector machine was used to perform link prediction to determine potential collaborators. The drawback of this method is that the use of LDA makes it computationally expensive.

2.1. Document clustering

Document clustering is the task of grouping a set of text documents into groups or clusters. Documents belonging to the same cluster share the same features while documents in another cluster do not share similar features. Traditional document clustering techniques use a bag of words representation; which do not take semantics into consideration. In [2] described the typical process of document clustering in Figure 1.

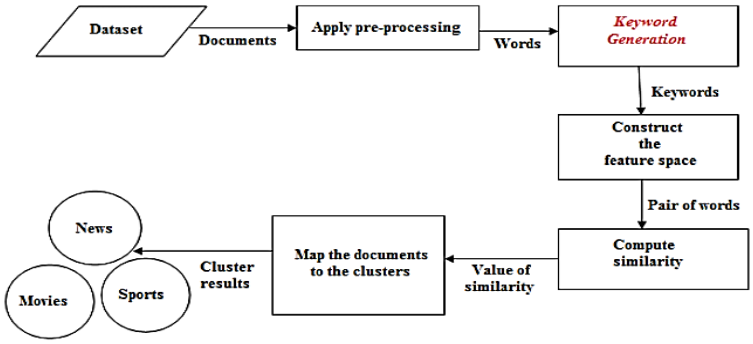

Figure 1. Document clustering process [2]

2.2. Feature representation

2.2.1. Vector space model

The vector space model (VSM) is largely considered the basic model for feature representation and has been modified severely to cater for its inadequacies. In the generic VSM model, each text document is represented as follows:

$$\phi: d \longmapsto \phi(d) = [w_{t_1,d}, w_{t_2,d,\ldots,w_{t_m,d}}]^T \in \Re^D \quad (1)$$

The term frequency-inverse document frequency technique has been widely used for feature representation of text documents, where is the TF-IDF weight of term t in document d. T denotes the transpose operator, denotes the document d as a weighted term vector in the m-dimensional space of terms. This function could also represent the mapping of a document to its vector space representation. The documents are then weighted by their inverse document frequency (IDF). This weighting is done to determine terms that appear frequently across the set of text documents. The TF-IDF matrix is represented thus:

$$(tf - idf)_{ij} = tf_{ij} \times idf_i \quad (2)$$

The drawback of the VSM model is that it suffers from the sparsity problem and does not perform well with large documents [24].

2.2.2. N-gram model

The N-gram model traditionally focuses on bi-grams, which are pairs of words but recently, the use of character N-grams and byte N-grams is commonplace. Character N-gram is a language autonomous text representation method. Text documents are transformed into high-dimensional feature vectors where features represent substrings. N-grams are typically N adjacent characters from the alphabet. The dimensionality of N-gram features can be as high as |A|N even for mid-range values of N. Generally, only a sizeable portion of N-grams are available in a given set of text documents. The N-gram model has the further advantage of being robust and tolerant of grammatical and typographical errors [25]. The limitation of the n-gram model is that the semantics of words and word order is not taken into consideration.

2.2.3. Latent semantic indexing

Latent semantic indexing is an algebraic based algorithm that is used for feature representation. It works based on a primary or latent structure to the word pattern usage in a text document and utilizes statistical techniques in determining this structure. It considers latent higher-order structures in the relationship between terms and documents. This technique can be applied to synonymy and polysemy problems. Latent semantic indexing is also used for dimensionality reduction using singular value decomposition (SVD).

The drawback of LSI is that it uses a bag-of-words approach which can lead to unstructured information and it only works on singular terms [26].

2.2.4. Probabilistic latent semantic analysis (PLSA)

Probabilistic LSA is a statistical technique for co-occurrence data. PLSA is derived from LSA by making it a probabilistic model. Probabilistic LSA is based on combination decomposition derived from a latent class model, unlike the standard latent semantic analysis which is from linear algebra. Documents are modeled as a multinomial combination of topics which gives room for document-document comparison. This makes PLSA a more popular technique for analysis of co-occurrence data. PLSA is used to model, the probability of each co-occurrence as a combination of independent multinomial distributions.

$$\mathrm{P(w,d)} = \sum_c P(c)P(d|c)P(w|c) = P(d)\sum_c P(c|d)P(w|c) \quad (3)$$

P(w,d) is the symmetric formulation where w and d are computed from the latent class in similar ways using the conditional probabilities P(d|c) and P(w|c), for each document [27].

3. MATERIAL AND METHODS

The 2010-2018 Publication data from randomly selected Nigerian institutions was retrieved through the Scopus API listed in the Scopus database. Each publication retrieved with a unique Elsevier ID is used to retrieve its abstract which builds a database of 9800 publication data. The abstracts are concatenated and stored in a database and the file format of the abstracts retrieved is in JavaScript Object Notation (JSON) file format.

3.1. Data preprocessing

The datasets were obtained in their raw form and required text processing and formatting to make them intelligible. Several operations were carried out on the datasets to extract the required text are discussed below:

- Non-printable characters: Regular expressions were used to remove characters that did not conform to the Unicode text encoding format.
- Tokenization: The natural language toolkit (NLTK) class was used to perform tokenization and the conversion of each word to lowercase characters.
- Number removal: The built-in python module was used to eliminate numbers but not words representing numbers.
- Stop words removal: Using the NLTK toolkit, stop words list was used to remove frequently occurring English words such as conjunctions and prepositions which do not reflect the content of the text corpus.
- Lemmatization: The NLTK wordnet and the WordNet Lemmatizer was used to obtain the stem versions of each word in the text corpus.

4. THE PROPOSED SYSTEM

The main components include dataset collation, pre-processing & database, document representation module, and the Pattern detection module. The system architecture is laid out in the Figure 2. The proposed framework consists of the presentation layer, business logic layer and the data layer, all having their roles in the total functionality of the system. Figure 3 presents the three-tier showing the different layers and their interactions.

- Presentation layer

From this layer, users can input a query and receive a response to their query. This layer cannot carry out computations on its own, but it interacts with the business logic layer through the Django web framework to provide more functionalities.

- Business logic layer

This layer consists of the Python application, which provides the functionalities to the presentation layer. It also interacts with the data layer through python SQLite connector to process necessary data useful for the working of the system. The Gensim library is a python library that is used for document representation implementation while the natural language toolkit handles natural language computations. The DBpedia and WordNet ontology is used for semantically annotating documents.

- Data layer

This is the layer where all the information needs of the system are stored. SQLite is used as the database management system platform for storing and managing records of individual reviewers. The data layer communicates with the business logic layer through python SQLite connector.

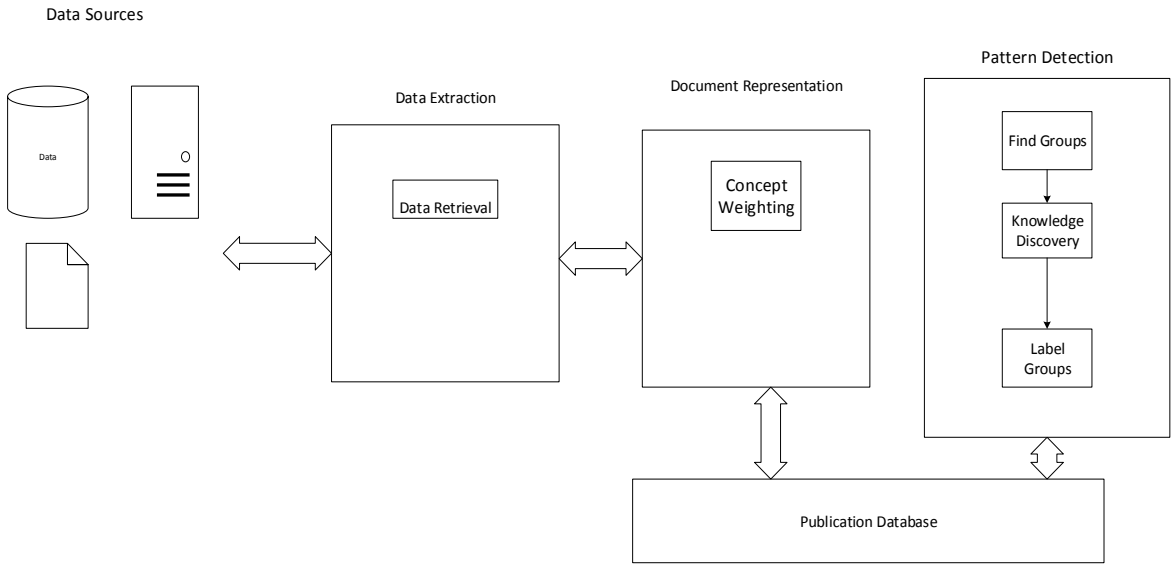

Figure 2. Proposed framework

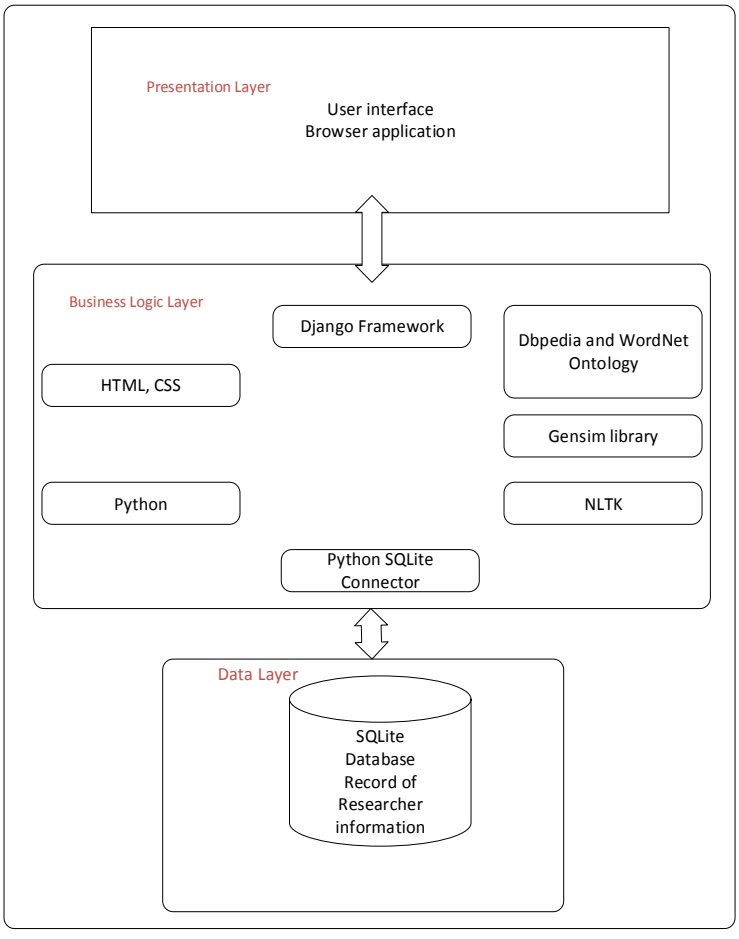

Figure 3. Three-tier architecture

4.1. Activity diagram

In Figure 4, the flow of activities in similar research area detection is shown. The diagram displays the flow of activities between the user and the system. The flow of activities includes: registration, login,

uploading an existing publication, the system recommends similar publications, the user checks the research area, the result is displayed, if the user is not pleased with the result, the user will be redirected to the similar research area page to adjust the uploaded publication.

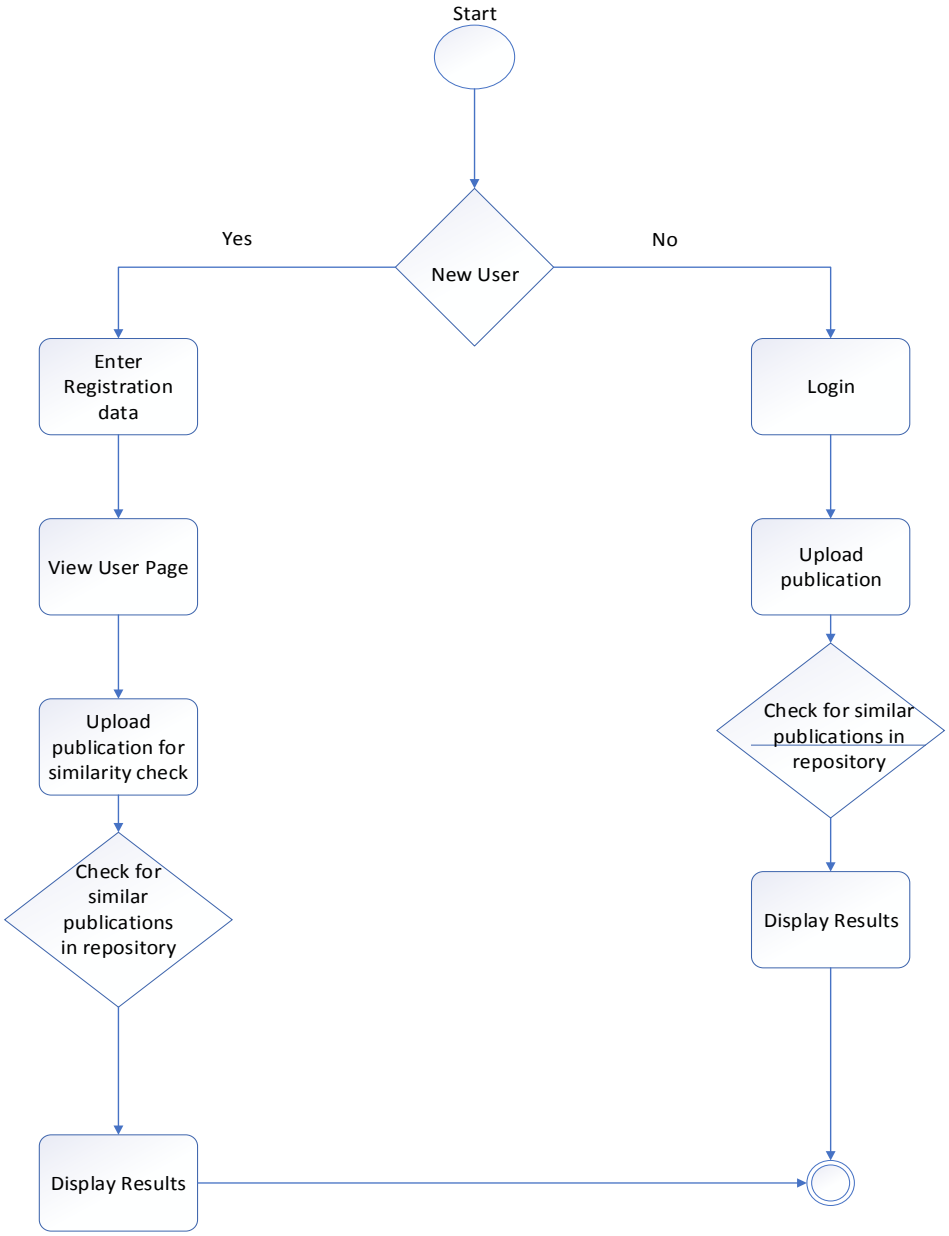

Figure 4. Activity diagram

4.2. Sequence diagram

Sequence diagrams model the dynamic behaviour of a system; they are interactive diagrams that depict the passing of messages between objects in a system. The messages are passed from the user to the registration module, and from the registration module to the database and back to the user to display a successful registration. Figure 5 shows the sequence diagram of the similar research publication discovery process and how messages are exchanged between the different modules from the user to the publication repository and back to the user with the result of the message passed across.

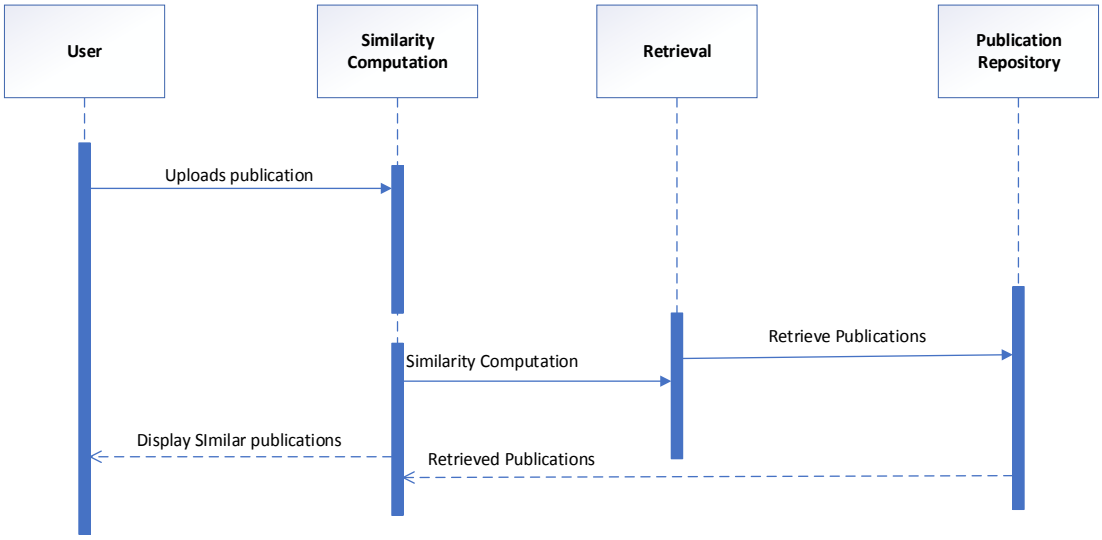

Figure 5. Sequence diagram for similar publication discovery

5. RESULTS AND DISCUSSION

5.1. Experimental results

The elbow curve was used to validate the optimal number of clusters beforehand, we used a range of values of K and selected the optimal value based on the elbow curve of the clusters. The evaluation of the clusters generated was performed using Silhouette analysis, which revealed a high level of accuracy and consistency as evidenced by an average 80% silhouette score for all the data points as shown in Figure 6. The figure shows that the number of clusters increases with the scores.

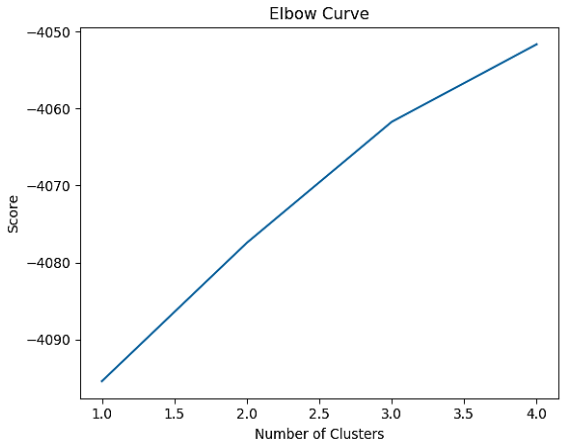

Figure 6. Evaluation results

6. CONCLUSION

In this work, a prototype system was developed to provide a platform for enhancing collaborations between researchers in a research field by enabling a researcher to identify other researchers in a given research field. A rigorous review of the literature was conducted to examine existing approaches that have been utilized in discovering similar research areas, this study utilized an approach involving the use of publication documents sharing certain keywords. Randomly selected Nigerian University publication documents from 2010 to 2018 were retrieved and used as a case study. This approach was modeled as a clustering problem. In order to improve the accuracy of the clustering results, the DBpedia and WordNet ontology was used to capture domain terms and semantically related terms in the publication dataset used. LSI and TF-IDF were used to model the text documents and generate feature vectors. The K-Means clustering algorithm was used to cluster

the feature vectors. The clustering results were evaluated using the silhouette analysis technique, which revealed a high intra-cluster similarity of 0.8 on the average across all the data points. A prototype system was developed using Python programming language, while SQLite Database Management System was used to manage the database. Our results show that the prototype system showed high clustering accuracy and could be deployed for large scale utilization.

7. CONTRIBUTION TO KNOWLEDGE

This work contributes to the existing body of knowledge by developing a prototype system that integrates ontology-based pre-processing, Latent Semantic Indexing and K-Means clustering to discover similar research domain publications. The system determines the similarity between publication documents using semantic similarity techniques. The clustering results show an improvement through the use of ontology semantics in pre-processing the documents. It is also believed that this approach will be useful for unearthing hidden and implicit patterns in the document dataset.

8. RECOMMENDATIONS AND FUTURE WORK

For the system to be continuously relevant, the publication repository will have to undergo regular updates to improve the robustness of the system. Hence, for future improvements in this research, the following recommendations are proposed:

- The system can be integrated with locally available conference and journal publication repositories to provide similar research area detection services.
- To improve the pre-processing stage, an ontology learning model can be incorporated into the concept of weighting and concept-document construction stage to produce domain ontologies pertinent to the publication text corpus.
- Other document clustering algorithms could be utilized, such as Bisecting K-means and K-medoids algorithm.
- The use of some other document representation techniques such as Word2Vec or Recurrent Neural Network Language can be used to improve the semantic similarity computation and further capture the implicit semantics of the text corpus.

ACKNOWLEDGMENTS

This research is fully sponsored by Landmark University Centre for Research and Development, Landmark University, Omu-Aran, Nigeria and Covenant University for Research and Discovery (CUCRID) for their immense support in this research.

REFERENCES

[1] Cagliero L., Chiusano S., Garza P., Bruno G., "Pattern set mining with schema-based constraint," *Knowledge-Based Systems*, vol. 84, pp. 224-38, 2015.
[2] Naik M. P., Prajapati H. B., Dabhi V. K., "A survey on semantic document clustering," *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2015.
[3] Szczuka M., Janusz A., Herba K., "Clustering of rough set related documents with use of knowledge from DB pedia," *International Conference on Rough Sets and Knowledge Technology*, 2011.
[4] Wang Y., Song S., Zhou F., Zheng X., "Chinese WeChat and Blog Hot Words Detection Method Based on Chinese Semantic Clustering," *Intelligent Automation and Soft Computing*, vol. 23, no. 4, pp. 613–618, 2017.
[5] Velden, Theresa, Boyack, Kevin, Glaser, Jochen, Koopman, Rob, Scharnhorst, Andrea, Wang, Shenghui, "Comparison of topic extraction approaches and their results," *Scientometrics*, vol. 111, pp. 1169-1221, 2017.
[6] Wang, Shenghui, Koopman, Rob., "Clustering articles based on semantic similarity," *Scientometrics*, vol. 111, pp. 1017-31, 2017.
[7] Naik M. P., Prajapati H. B., Dabhi V. K., "A survey on semantic document clustering," *Proceedings of 2015 IEEE International Conference on Electrical, Computer and Communication Technologies*, 2015.
[8] Yuan X, Sun M, Chen Z, Gao J, Li P., "Semantic clustering-based deep hypergraph model for online reviews semantic classification in cyber-physical-social systems," *IEEE Access*, vol. 6, pp. 17942-51, 2018.
[9] Fahad S. A., Yafooz W. M., "Design and develop semantic textual document clustering model," *Journal of Computer Science*, vol. 3. no. 2, pp. 26-39, 2017.
[10] Mahajan S., Shah N., "Efficient pre-processing for enhanced semantics based distributed document clustering," *2016 3rd International Conference on Computing for Sustainable Global Development*, 2016.

[11] Oladele T. O., Aro T. O., Adegun A. A., Ogundokun R. O., "Prediction of Student's Academic Performance using K-menas Clustering and Multiple Linear Regressions," *Journal of Engineering and Applied Sciences*, vol. 14, no. 22, pp. 8254-8260, 2019.
[12] Zhang H., Wang D., Wang L., Zhuming Bi., Yong Chen, "A semantics-based method for clustering of Chinese web search results," *Enterprise Information Systems*, vol. 8, no. 1, pp. 147-65, 2014.
[13] Soliman S. S., El-Sayed M. F., Hassan Y. F., "Semantic clustering of search engine results," *The Scientific World Journal*, vol. 2015, pp. 1-9, 2015.
[14] Oladele T., Adegun A., Ogundokun R. O., keyinka A., Ayeni L., "Application of Floyd-Warshall's Algorithm in Air Freight Service in Nigeria," *International Journal of Engineering Research and Technology*, vol. 12, no. 12, pp. 2529-2535, 2019.
[15] Yue L., Zuo W., Peng T., Wang Y., Han X., "A fuzzy document clustering approach based on domain-specified ontology," *Data & Knowledge Engineering*, vol. 100, pp. 48-66, 2015.
[16] Sulthana A. R., Subburaj R., "An improvised ontology-based K-means clustering approach for classification of customer reviews," *Indian Journal of Science and Technology*, vol. 9, no. 15, pp. 1-6, 2016.
[17] Makarov I., Bulanov O., Zhukov L. E., "Co-author recommender system," *International Conference on Network Analysis NET 2016: Models, Algorithms, and Technologies for Network Analysis*, pp. 251-257, 2016.
[18] Sumba X., Sumba F., Tello A., Baculima F., Espinoza M., Saquicela V., "Detecting similar areas of knowledge using semantic and data mining technologies," *Electronic Notes in Theoretical Computer Science*, vol. 329, pp. 149-67, 2016.
[19] Arumawadu H. I., Rathnayaka R. M. K. T., Illangarathne S. K., "Mining profitability of telecommunication customers using k-means clustering," *Journal of Data Analysis and Information Processing*, vol. 3, no. 3, pp. 63–71, 2015.
[20] Awotunde J. B., Ogundokun R. O., Ayo F., Ajamu G. J., Adeniyi E., Ogundokun E. O., "Social Media Acceptance and Use Among University Students for Learning Purpose Using UTAUT Model," *International Conference on Information Systems Architecture and Technology*, 2019
[21] Ogundokun R. O., Adebiyi, M. O., Abikoye, O. C., Oladele, T. O., Lukman A. F., Adeniyi A. E., Adegun A. A., Gbadamosi B., Akande N. O., "Evaluation of the scholastic performance of students in 12 programs from a private university in the south-west geopolitical zone in Nigeria," pp. 154, 2019. F1000Research 8 [version 1].
[22] Ogundokun R. O., Adebiyi M. O., Abikoye O. C., Oladele T. O., Lukman A. F., Adeniyi A. E., Adegun A. A., Gbadamosi B., Akande N. O., "Evaluation of the scholastic performance of students in 12 programs from a private university in the south-west geopolitical zone in Nigeria," 2019. F1000Research 8 [version 2].
[23] Oladele T. O., Ogundokun R. O., Kayode A. A., Adegun A. A., Adebiyi M. O., "Application of Data Mining Algorithms for Feature Selection and Prediction of Diabetic Retinopathy," *International Conference on Computational Science and Its Applications*, 2019.
[24] Wang Y., Song S., Zhou F., Zheng X., "Chinese WeChat and Blog Hot Words Detection Method Based on Chinese Semantic Clustering," *Intelligent Automation & Soft Computing*, vol. 23, no. 4, pp. 613-8, 2017.
[25] Chuan P. M., Ali M., Khang T. D., Dey N., "Link prediction in co-authorship networks based on hybrid content similarity metric," *Applied Intelligence*, vol. 48, no. 8, pp. 2470-86, 2018.
[26] Deshmukh A., Hegde G., Lathi R., Govikarn S., "A literature survey on latent semantic indexing," *International Journal of Engineering Inventions,"* vol. 1, no. 4, pp. 2278-7461, 2012.
[27] Shafiei M., Wang S., Zhang R., Milios E., Tang B., Tougas J., Spiteri R., "Document representation and dimension reduction for text clustering," *2007 IEEE 23rd international conference on data engineering workshop*, 2007.