Investigation of Similarity Paradigms for Electronic Document Query and Retrieval

[1]Samuel S. Udoh*, [2]Uduak D. George, [3]Okure U. Obot [4]Isaiah S. Tom

[1,2,3,4] Department of Computer Science
Faculty of Science
University of Uyo
Nigeria

*Corresponding author's email: samueludoh@uniuyo.edu.ng

Abstract

Many similarity models for electronic document retrieval have difficulties in retrieving and ranking relevant information from document repositories in response to queries. This stems from the fact that the natural language employed in queries sometimes contain ambiguous elements which sponsor the retrieval of irrelevant information. This research is aimed at investigating similarity models with a view to selecting appropriate model for deployment in document query and retrieval system. Models of Cosine, Okapi, Jaccard, Dice and Fuzzy logic-based similarity algorithms were designed and implemented using Java programming tools. My Structured Query Language (MySQL) database was designed for data repository. Course materials totaling 5025 in the Department of Computer Science University of Uyo, Nigeria were collected and stored as documents in the data repository. Queries were presented to the repository via the similarity models in the program interface. Mean score of similarity assessment obtained from three (3) human experts served as the parameter for evaluating the scores derived from the similarity models in the Java-based program. Results of similarity analysis showed a strong correlation value of 0.999 between the Expert score and the Fuzzy model followed by Okapi (0.958), Dice (0.940), Cosine (0.936) and Jaccard (0.757). In the document ranking analysis, Cosine model had the least correlation value of 0.767, while Fuzzy model had the highest correlation value of 0.978 and the least error value of 0.01. Fuzzy model is therefore considered the closest model to human Expert result in the domain of document query and retrieval.

Keywords: Similarity Models, Fuzzy Algorithm, Document Query, Retrieval

1. Introduction

The main task of an information retrieval (IR) system is to provide a list of relevant documents for a user query. This task is generally formulated as a ranking problem. The ranking function is obtained by computing similarity scores between queries and documents in the repository, Baeza-Yates & Ribeiro-Neto [7]; Obot *et al*.[10]. The higher the score the greater the importance of document to query term, Swain *et al.* [4]. Documents are retrieved from the repository when they contain index terms specified in queries. However, this approach neglects other relevant documents that do not contain index terms specified in queries. When working with specific domain, this problem could be solved by incorporating a knowledge-base of the domain such as ontology which depicts relationships in index terms, Leite [3].

In IR, one of the challenges is the inability of search engines or databases to precisely understand user's needs. Sometimes, users do not know the precise vocabulary of the topic to be searched to get the best results, Delgado *et al*. [8]. He & Ounis [2], proposed an entropy, measure that estimated the spread of query terms over returned documents. It was shown that entropy in the top 5 returned documents is normally very high and that the entropy decreases rapidly in the remaining documents. Having only top 5 web documents returned as relevant documents to the user's query implies that the documents are not properly ranked. Hence, the need to incorporate effective document ranking algorithm.

In order to deal with the vagueness typical of human knowledge, the fuzzy set theory could be used to manipulate the knowledge-base for optimal results Obot *et al.* [10]. The expectation is that the indexed terms could improve the quality of retrieved documents bringing the most relevant documents to the initial query. One possible solution to confront the uncertain and vague information is inserting the fuzzy logic into the construction process of Information Retrieval System, Tho [6]. Fuzzy set theory, and computational intelligence techniques improve the effectiveness of indexing, classification and clustering in information retrieval systems, Leite [3], Nagpal, [9]; Qiu *et al*.[11], Samuel *et al*. [12].

The boolean model, vector space model (VSM), and the probabilistic model have been proposed to implement document ranking and selection in IR systems. Similarity measures such as (Cosine, Okapi Jaccard and Dice) use the weighting scheme in the similarity computation, Obot *et al.* [10], while fuzzy logic simulates human-like intelligence in computation of document similarity, Mohammad & Al-Ibrahim [1], Qiu *et al*. [11]. This study seeks to investigate similarity models in quest of exploiting uncertainty and imprecision in finding solutions to documents query, ranking and retrieval. The remainder of the work is organized as follows. Reviews of related works on document similarity models are carried out in Section 2. The system design and components interaction are conceptualized in Section 3. System implementation and results are discussed in Section 4 while Section 5 presents the conclusion of the research and recommendations for further work.

2. Related Works

A comparison of Cosine, Jaccard and Dice similarities coefficient is presented in Vikas & Vivek [14], to find the best fitness value for retrieval of web documents. An experiment was conducted using keywords from 10 pages of the google-searched document. The best fitness values were obtained using the Cosine similarity coefficients followed by Dice and Jaccard. The superiority of the Cosine similarity measure over Dice and Jaccard could not be substantiated due to insufficient data employed for empirical confirmation. In Ramzy, [5], the effects of similarity measures were assessed on genetic algorithm-based information retrieval system. Cosine similarity measure was more effective than others in detecting the similarity of documents that were vastly different in their sizes. 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. Jaccard similarity measure was effective for calculating the similarity between the queries and documents in which the index starts with a minimum value of 0 (completely dissimilar) and goes to a maximum value of 1 (completely similar), Wael & Aly [15].

In Mohammad & Al-Ibrahim [1], fuzzy logic (FL) system is applied for information retrieval in electronic libraries. Index algorithms were applied in FL system to accomplish indexing operations as well as computation of document ranking for retrieval. The system addressed the challenges of document ranking in web and digital libraries. Qiu *et al*. [11], proposed FL retrieval system on continuous bag-of-words (CBOW) model. A CBOW algorithm was developed to generate vector representations of a language vocabulary to enable word encoding in vector space structure. The FL approach combined the techniques of deep learning and fuzzy set theory to capture the relationships between words and query item. Similarity models in information retrieval were used to determine the resemblance between the texts selected for document clustering. Weights were assigned to the terms of a query to increase their relative importance in generating better results. Nagpal [9], applied soft computing techniques in information retrieval. There was significant improvement in efficiency in acquiring knowledge related to a user's query compared to traditional retrieval methods. Soft computing paradigm exploits uncertainty, imprecision and approximate reasoning to achieve low-cost solutions that are robust and tractable. Soft computing deals with the implementation of optimization techniques to find probable solutions to hardcore problems.

Sharma & Mittal [13], presented three approaches namely: N-grams, word sense disambiguation and K-nearest neighbor (KNN) algorithm to improve query-based document ranking presented for information retrieval. The ranking of documents was improved by reformulating the query and performing a spell check using n-grams. Contextual meaning of the ambiguous terms (polysemy words) were identified using WordNet and the relevant terms were added to the original user's query. The similarity score was assigned to the extended query and the KNN technique was applied to find the most relevant documents along with their ranking. Since the user query does not index and rank all the relevant terms properly, it could lead to less accurate results. In this regard,

Sharma & Mittal [13] enhanced retrieval of documents by finding suitable terms that are similar to the original query terms and including those terms in the original query to facilitate retrieval of relevant documents. Guo & Gomes [16] proposed an approach for automatically searching and ranking structured documents. The model employed support vector machine patent ranking (SVMPR) which incorporated margin constraints that directly capture the specificities of patent citation ranking. The approach combined patent domain knowledge features with meta-score features from different general IR methods. The training algorithm extended the Pegasos algorithm. Experiments on a homogeneous essential wireless patent dataset showed that SVMPR performs on average, 30%-40% better than many other state-of-the-art general-purpose IR methods in terms of the normalized discounted cumulative gain measure at different cut-off positions.

Obot *et al.* [10] compared the effectiveness of Jaccard, Cosine, Jaro and Dice similarity measures for grading short answers to examination questions. The similarity measures were tested with the aim of ascertaining the measure that rank closest to the average scores provided by 3 human examiners. Results showed that Jaro similarity measure ranked closest to the mean score of the examiners. Soft computing paradigms were not incorporated to facilitate human-like reasoning and decision making.

3. System Design

The main components of the system architecture are document repository, user query, similarity algorithm, ranking module, evaluation/selection module and query response as shown in Figure 1.

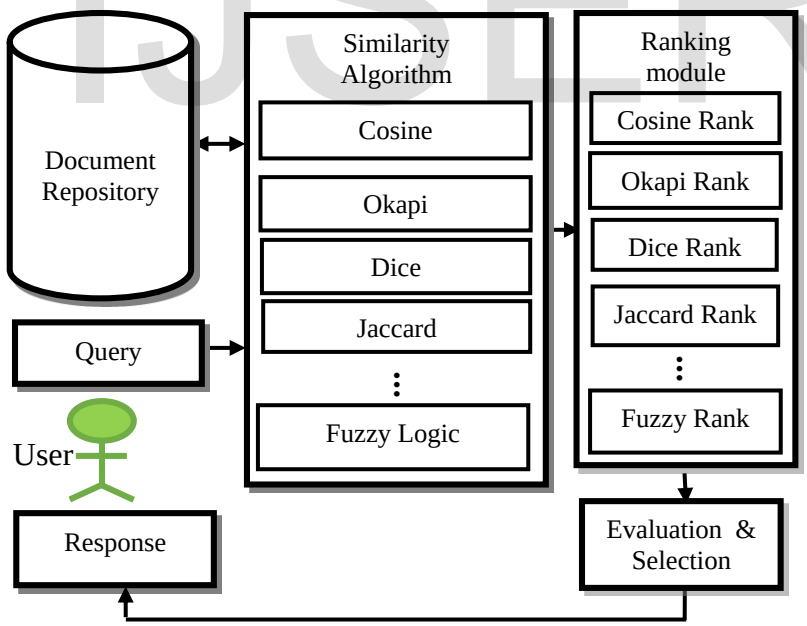


Figure 1: Document Query and Retrieval System Architecture

The document repository is a compendium of databases and electronic files of an organization or items accumulated over time. In this research, the document repository is

the databases of course materials offered by students from the Department of Computer Science University of Uyo, Nigeria at undergraduate and postgraduate levels. The course materials were accumulated over the period of 11 years from year 2010 to year 2021. The user query is the search term provided by the user to the system and sent to the similarity algorithms such as cosine, okapi, dice and fuzzy logic models. Similarity algorithms compute the resemblance between a query term and a list of documents in the document repository. Each similarity algorithm computes a similarity score (how similar a search term is to each document in the document database) in a one-to-many fashion. Ranking module positions the document based on descending order of their similarity scores. Evaluation/Selection module applies the mean score of human-experts to evaluate and determine the algorithm that should be incorporated in the final implementation of the electronic document retrieval system. Equations 1, 2, 3, 4 and 5 represent the Cosine, Okapi, Dice, Jaccard and Fuzzy logic similarity measures respectively as adopted in the system; while Equation 6 represents the human expert similarity assessment guide.

$$\mathrm{Cos}(\mathrm{T1},\mathrm{T2}) = \frac{\sum_{j=1}^{m}(w1_j * w2_j)}{\sqrt{\sum_{j=1}^{m}(w1_j)^2} * \sqrt{\sum_{j=1}^{m}(w2_j)^2}} \quad (1)$$

where: T1 is the correct answer to a question
T2 is the answer given by a student
$w1_j$ is the term frequency of words in T1
$w2_j$ is the term frequency of words in T2

The cosine similarity measure uses the term frequency that is the number of times a term or word occurs in a document. Each term or word in the document is a dimension in Euclidean space and the frequency of each word corresponds to the value in the dimension.

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot (1-b+b \cdot \frac{|D|}{avgdl})} \quad (2)$$

The Okapi model is a bag-of-word retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of their proximity within the document. Given a query Q, containing keywords $q_1, \ldots, q_{n,}$ the Okapi score of a document D is shown in Equation 2.

where:
$f(q_i, D)$ is $q_i's$ term frequency in the document D
$IDF(q_i)$ is the inverse document frequency of the query term q_i
$|D|$ is the length of the document D in words
$avgdl$ is the average document length in the text collection from which documents are drawn k_1 and b are free parameters, usually chosen as $k_1 \in [1.2, 2.0]$ and $b = 0.75$

$$\text{Dice (d,q)} = \frac{2*|d \cap q|}{|d|+|q|} \quad (3)$$

where:

d = document in the repository
q = query supplied by the user
The Dice similarity measure is the ratio of twice the intersection of the strings under consideration to the union of the strings.

Jaccard similarity is computed using the following formula

$$J(D,Q) = \frac{|D \cap Q|}{|D \cup Q|} = \frac{|D \cap Q|}{|D|+|Q|-|D \cap Q|} \quad (4)$$

where:
D and Q are the sets of documents and query items respectively
$|D|$ and $|Q|$ are the cardinality of D and Q respectively representing the count of the number of elements in set D and set Q respectively.
∪ is the union of two sets
∩ is the intersection of two sets

Jaccard measure is considered as the size of the intersection divided by the size of the union of the document and query vectors. It is also known as the Intersection over Union (IoU) measure.

Fuzzy logic similarity measure is given in Equation 5.

$$f(x;a,b,c) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ \frac{c-x}{c-b}, & b \leq x \leq c \\ 0, & c \leq x \end{cases} \quad (5)$$

where:
a - the left leg of the membership function
b - the center of the function
c - the right leg of the function
x - the crisp input
f - a mapping function

Equation 6 shows the human expert similarity assessment guide.

$$Expert(Sim) = \begin{cases} \text{Very Low} & \text{if } Sim < 0.1 \\ \text{Low} & \text{if } 0.1 \le Sim < 0.3 \\ \text{Moderate} & \text{if } 0.3 \le Sim < 0.6 \\ \text{High} & \text{if } 0.6 \le Sim < 0.8 \\ \text{Very High} & \text{if } 0.8 \le Sim \le 1.0 \end{cases} \quad (6)$$

4. Results and Discussion

Course materials totaling 5025 in 84 courses comprising 66 courses offered at undergraduate level and 18 courses offered at postgraduate level in the Department of Computer Science University of Uyo, Nigeria were collected and stored as document repository. Queries were presented to the repository via the search term interface. Document similarity and selection process employed Cosine, Jaccard, Okapi, Dice and Fuzzy logic, similarity algorithms to compute the similarity between the search term and documents in the repository. The document query and retrieval interface depicted in Figure 2 accepts selection threshold, which determines the number of documents filtered from the repository in response to user's query. The dynamic threshold value determines the number of documents that are presented to the user. The documents presented to the user are ranked using the similarity score and arranged in descending order of importance to the user's query.

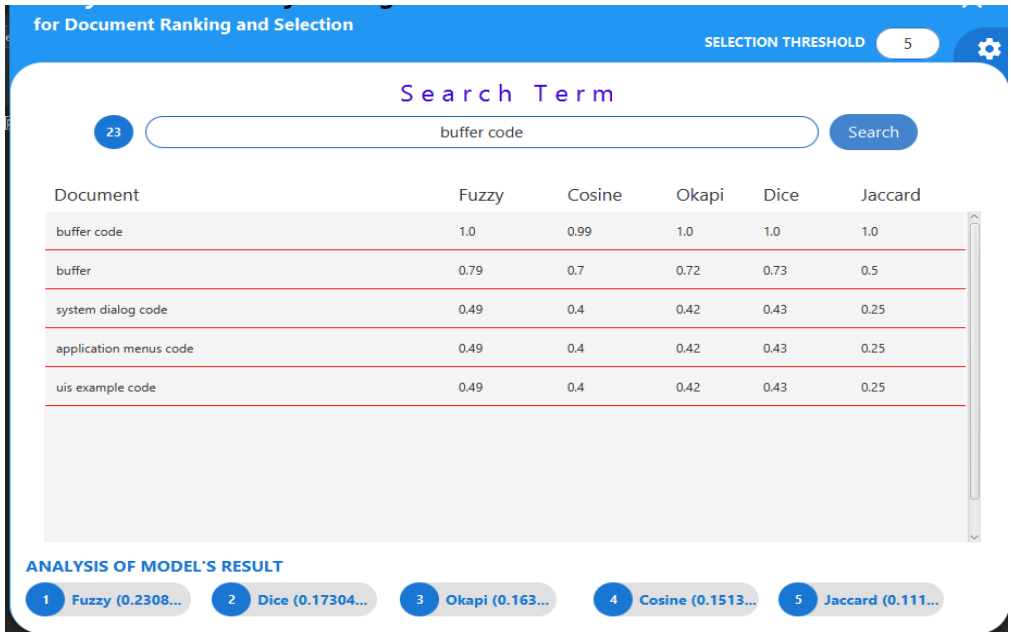


Figure 2: Document Query and Retrieval Interface

Table 1 presents the similarity scores of the algorithms in response to user's query as well as the mean similarity score obtained from three (3) human experts. Table 2, displays excerpts of 25 documents stored as info1, info2, … , info25, retrieved by similarity algorithms in response to the user's query: "buffer code". The ranking of query term similarity with retrieved document by human experts is included to facilitate evaluation of the similarity algorithms. Tables 3 and 4 present similarity correlation and ranking correlation respectively between the query term and the retrieved document.

Table 1: Similarity Scores of Document to Query

SN	Fuzzy	Cosine	Okapi	Dice	Jaccard	Mean Expert Score
1	1.00	0.98	0.95	0.87	0.86	0.99
2	0.79	0.70	0.72	0.72	0.50	0.80
3	0.82	0.75	0.78	0.74	0.65	0.81
4	0.54	0.54	0.56	0.65	0.57	0.54
5	0.49	0.40	0.42	0.43	0.25	0.52
6	0.61	0.69	0.53	0.57	0.65	0.62
7	1.00	0.99	1.00	1.00	1.00	0.99
8	0.79	0.70	0.72	0.72	0.50	0.80
9	0.49	0.40	0.42	0.43	0.25	0.50
10	0.73	0.71	0.76	0.65	0.64	0.71
11	0.69	0.70	0.62	0.73	0.75	0.68
12	0.61	0.69	0.53	0.57	0.65	0.62
13	1.00	0.99	1.00	0.99	1.00	1.00
14	0.79	0.70	0.72	0.72	0.50	0.80
15	0.59	0.60	0.72	0.73	0.25	0.60
16	0.49	0.40	0.42	0.43	0.25	0.51
17	0.51	0.45	0.47	0.43	0.75	0.52
18	0.61	0.69	0.53	0.57	0.65	0.62

19	1.00	0.99	0.93	0.97	1.00	1.00
20	0.79	0.70	0.72	0.72	0.50	0.80
21	0.49	0.40	0.42	0.43	0.25	0.50
22	0.49	0.40	0.42	0.43	0.25	0.51
23	0.49	0.40	0.42	0.43	0.25	0.52
24	0.61	0.69	0.53	0.57	0.65	0.62
25	0.71	0.79	0.73	0.67	0.65	0.71

Table 2: Document Ranking in Response to Query

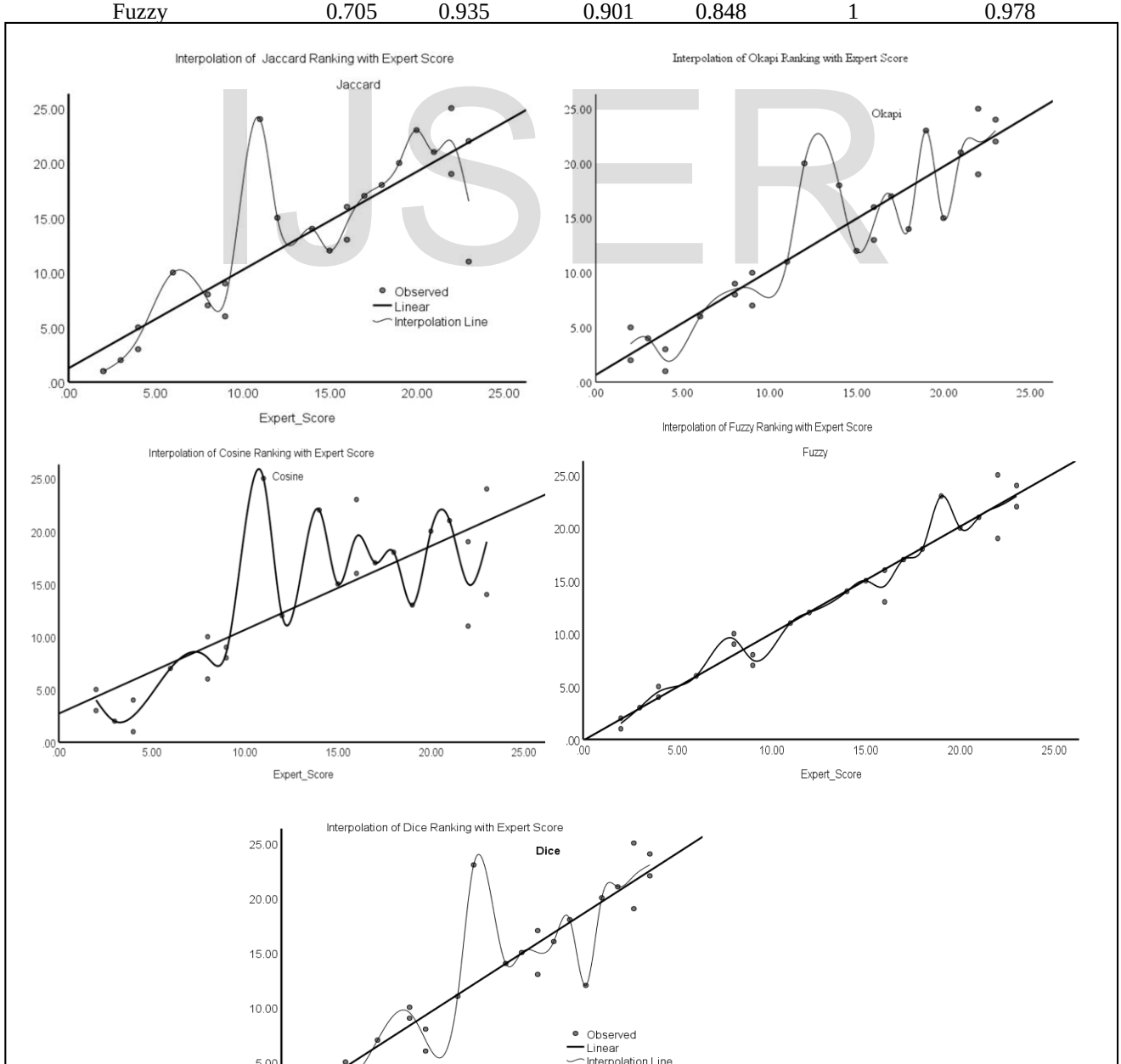
SN	Fuzzy	Cosine	Okapi	Dice	Jaccard	Expert1	Expert2	Expert3
1	Info2	Info3	Info2	Info3	Info1	Info2	Info2	Info3
2	Info3	Info2	Info4	Info4	Info2	Info3	Info3	Info2
3	Info4	Info4	Info3	Info2	Info3	Info4	Info4	Info4
4	Info5	Info1	Info1	Info5	Info5	Info5	Info1	Info5
5	Info1	Info5	Info5	Info1	Info1	Info1	Info5	Info1
6	Info6	Info7	Info6	Info7	Info10	Info7	Info6	Info6
7	Info7	Info8	Info7	Info8	Info6	Info10	Info10	Info7
8	Info10	Info10	Info9	Info9	Info7	Info6	Info7	Info10
9	Info8	Info9	Info10	Info6	Info9	Info8	Info9	Info9
10	Info9	Info6	Info8	Info10	Info8	Info9	Info8	Info8
11	Info25	Info11	Info25	Info25	Info25	Info24	Info19	Info24
12	Info24	Info24	Info24	Info24	Info11	Info25	Info25	Info19
13	Info11	Info25	Info11	Info11	Info24	Info11	Info11	Info11
14	Info22	Info14	Info22	Info22	Info22	Info19	Info24	Info25
15	Info21	Info21	Info21	Info21	Info21	Info21	Info21	Info21
16	Info14	Info22	Info18	Info14	Info14	Info14	Info14	Info14
17	Info19	Info19	Info19	Info19	Info19	Info22	Info22	Info22
18	Info18	Info18	Info14	Info18	Info18	Info18	Info18	Info18
19	Info17	Info17	Info17	Info16	Info17	Info17	Info17	Info17
20	Info16	Info16	Info16	Info17	Info16	Info16	Info16	Info16
21	Info15	Info15	Info12	Info15	Info12	Info15	Info15	Info15
22	Info20	Info20	Info15	Info20	Info23	Info20	Info20	Info20
23	Info13	Info23	Info13	Info13	Info13	Info13	Info23	Info13
24	Info12	Info12	Info20	Info23	Info15	Info12	Info13	Info12
25	Info23	Info13	Info23	Info12	Info20	Info23	Info12	Info23

Table 3: Similarity correlation of query term and retrieved document

Model	Cosine	Okapi	Dice	Jaccard	Fuzzy	Expert_Score
Cosine	1	0.927	0.932	0.868	0.943	0.936
Okapi	0.927	1	0.970	0.739	0.965	0.958
Dice	0.932	0.970	1	0.761	0.947	0.940
Jaccard	0.868	0.739	0.761	1	0.774	0.757
Fuzzy	0.943	0.965	0.947	0.774	1	0.999
Expert_Score	0.936	0.958	0.940	0.757	0.999	1

Table 4: Ranking correlation of query term and retrieved document

Model	Cosine	Okapi	Dice	Jaccard	Fuzzy	Expert_Score
Cosine	1	0.692	0.688	0.725	0.705	0.767
Okapi	0.692	1	0.902	0.802	0.935	0.921
Dice	0.688	0.902	1	0.803	0.901	0.911
Jaccard	0.725	0.802	0.803	1	0.848	0.846
Fuzzy	0.705	0.935	0.901	0.848	1	0.978



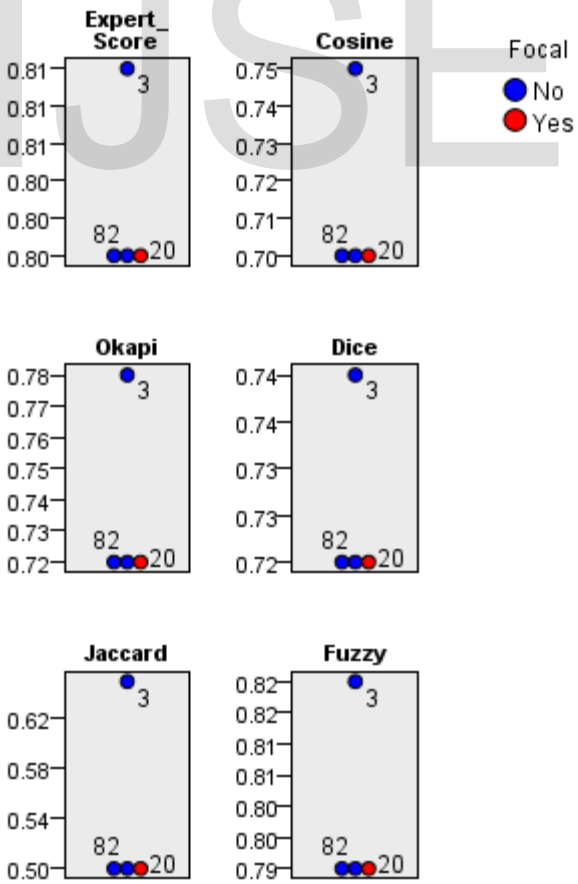


Figure 4: Document Retrieval Similarity Pears Chart

Statistical Package for Social Sciences (SPSS version 26 developed by International Business Machines was used in the similarity and ranking correlation analysis. Summary of results for similarity correlations are presented in Table1 while summary of results for ranking correlations are captured in Table 2. A strong similarity correlation value of 0.999 was obtained between the Expert score and the Fuzzy algorithm score followed by Okapi (0.958), Dice (0.940), Cosine (0.936) and Jaccard (0.757). In the document ranking analysis, cosine measure had the least correlation value of 0.767, followed by Jaccard (0.846), Dice (0.911), Okapi (0.921) while Fuzzy had the highest correlation of 0.978.

The interpolation graphs in Figure 3, illustrate the response of the similarity algorithms to Expert score value. The observed values connected by interpolation curve showed significant variation with the linear line for Jaccard, Okapi, Dice and Cosine measures. Insignificant deviation is noticed between the Fuzzy interpolation curve and the linear line, thereby attesting to the inference that Fuzzy similarity measurement values are very close to human Expert measurement.

Retrieval similarity peers chart shown in Figure 4 compares the patterns of similarity algorithms with the Expert's score at selected focal point. At many focal points the fuzzy algorithm retrieval was observed as the closest neighbour to expert score than any other algorithm. For instance, at focal point 20, the document similarity measurement obtained from Expert, Fuzzy, Cosine, Okapi, Dice and Jaccard scores were 0.80, 0.79, 0.70, 0.72, 0.72 and 0.50 respectively. The difference between the Expert score and other scores: (Fuzzy, Cosine, Okapi, Dice, Jaccard) derived from the similarity pear chart were 0.01, 0.10, 0.08, 0.08,0.30 respectively. The least error value of 0.01 was observed between the Expert and Fuzzy Score, followed by Okapi, Dice and Cosine. While the greatest error of 0.30 was observed in Jaccard measurement. It is inferred that Fuzzy score gives the least error in measurement compared to other similarity algorithms at many focal points and therefore considered the closest neighbour to human Expert measurement.

5. Conclusion

In this paper, models of Cosine, Okapi, Jaccard, Dice and Fuzzy logic-based similarity algorithms were designed and implemented using Java programming tools and MySQL database. Course materials from undergraduate and postgraduate programmes in the Department of Computer Science, University of Uyo, Nigeria were collected and stored as documents in the database. Queries were presented to the database via the search term box and the similarity models in the program interface. Mean score from three (3) human experts served as assessor to the scores derived from the similarity models. Analysis of similarities showed a strong correlation value of 0.999 between the expert score and the Fuzzy model while Jaccard model had the least similarity value of 0.757. In the document ranking analysis, Cosine model had the least correlation value of 0.767, while Fuzzy model had the highest correlation value of 0.978. Fuzzy model is therefore considered the closest model to human Expert result and hence, most appropriate for incorporation in the development of document

retrieval system to handle uncertainties, vagueness and imprecision in user queries. In further work, adaptive neuro-fuzzy paradigm as well as lexical and semantic models would be incorporated for more optimal results.

REFERENCE

[1] Mohammad Ali & Al-Ibrahim H. (2019) Fuzzy Logic System for Retrieval of Information in Electronic Libraries, Modern Applied Science; Vol. 13, No. 11; ISSN 1913-1844 E-ISSN 1913-1852

[2] He B. & Ounis, I. (2009) "Studying query expansion effectiveness," in The 31th European conference on IR research on advances in IR, Toulouse, France, , pp. 611-619.

[3] Leite M. A. A. & Ricarte,I. L. M (2008.) "Fuzzy information retrieval model based on multiple related ontologies," in Proceedings of The 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '08), pp. 309–316,

[4] Swain M., Andersen J. A. & Korrapati R (2005) "Study of information retrieval using fuzzy queries," in IEEE Southeastcon 2005: Excellence in Engineering, Science and Technology, pp. 527–535.

[5] Ramzy, M. Girgis, A. A. & Azzam F. M. (2014): "The Effect of Similarity Measures on Genetic Algorithm-Based Information Retrieval", International Journal of Computer Science Engineering and Information Technology Research, Vol. 4, Issue 5, pp. 91-100.

[6] Tho Q. T., Hui S. C., Fong A. C. M., & Cao T. H. (2006)., "Automatic Fuzzy ontology generation for semantic Web," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 6, pp. 842–856,

[7] Baeza-Yates R. & Ribeiro-Neto B. (2011): Modern Information Retrieval: The Concept and Technology behind Search, England: Pearson Education Limited,

[8] Delgado, S. M. Martin-Bautista M., D. Sanchez, J. Serrano & M. Vila, (2009) "Association rules and fuzzy association rules to find new query terms," in EUSFLAT, Lisbon, Portugal, pp. 49-53.

[9] Nagpal, Namrata. (2018)."Applying Soft Computing Techniques in Information Retrieval." *International Journal of Advanced Engineering, Management and Science*, vol. 4, no. 5,

[10] Obot, O.U., Udoh, S. S. & Attai, K. F. (2021): The suitability of similarity measures to the grading of short answers in examination. Inderscience Enterprises Ltd. International Journal of Quantitative Research in Education, 5 (3), 207–222.g3

[11] Qiu D., Jiang H. & Chen S. (2020) Fuzzy Information Retrieval Based on Continuous Bag-of-Words Model. Journal of Computer Science Symmetry. Vol.12 pp225.

[12] Samuel S. Udoh, Francis B. Osang, Olutola O. Fagbolu & Michael E. Isang**, (**2021**):** Intelligent Vehicular Traffic Control System using Priority Longest Queue First Model. Global Journal of Computer Science and Technology (GJCST): Neural and Artificial Intelligence, 21(1), 9-18.

[13] Tanuj Sharma & Kanika Mittal (2017), Query Based Document Ranking For Enhanced Information Retrieval, International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Vol-4, Issue-7

[14] Vikas T. & Vivek J. (2013). Comparison of Jaccard, Dice, Cosine Similarity Coefficient for Web Retrieved Documents Using Genetic Algorithm. International Journal of Innovations in Engineering and Technology (IJIET). 2(4), 202-205

[15] Wael H. G. & Aly A. F. (2013). A Survey of Text Similarity Approaches, International Journal of Computer Applications 68(13),75 – 87.

[16] Yunsong Guo & Carla Gomes (2017), Ranking Structured Documents: A Large Margin Based Approach for Patent Prior Art Search, Department of Computer ScienceCornell University.