Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7546452

A COMPARATIVE ANALYSIS OF TEXT SIMILARITY MEASURES AND ALGORITHMS IN RESEARCH PAPER RECOMMENDER SYSTEMS

Noore Ilahi[1], Mukka Shirisha[2]
Assistant Professor[1], Assistant Professor [2]
ilahinoore90@gmail.com[1], shirishakirans@gmail.com[2],
Department of CSE[1], Department of CSE(AIML)[2]

Abstract—

Recent advancements in the internet and web technologies are responsible for the surge in online published research articles. Because of the information boom, academics and internet users have a hard time obtaining relevant and reliable data. Finding the optimal mix of algorithms and similarity metrics for research paper recommender systems' article search and recommendation is the goal of this work. In this study, we used text similarity metrics with non-linear classification methods. While several similarity measures are evaluated using existing datasets, an offline assessment method is used to ascertain the correctness and performance of the models. Boosted, Recursive PARTitioning (rpart), and Random Forest are a few machine learning techniques that will be used to datasets that measure the similarity of research papers. With an average accuracy of 80.73 and a time efficiency of 2.354628 seconds, the rpart method outperformed the Boosted and Random Forest algorithms, respectively. When compared to other similarity measures, cosine similarity fared the best. There will be a proposal for new metrics and measurements of similarity. In this study, we show that when trying to build models for research paper similarity assessment and recommendation, there are superior metric and algorithm combinations to apply. We also found several other problems and unanswered questions.

Keywords— recommender systems; documents; data mining; information retrieval; big data; similarity metrics and measures

I. INTRODUCTION

Because of the proliferation of online electronic documents, textual categorisation and document classification in online repositories have become more important. The ever-increasing amount of big data is now being processed by using text mining, machine learning, and natural language processing approaches and techniques. An explosion of research papers and scientific outputs in the form of journals and scientific literary works has also been brought about by widespread research. Because of this, researchers are hell-bent on finding interesting subjects that pertain to their field of study. The desire to learn about recent developments in a certain area of study or the need to include research articles into their citations are two other possible motivating factors.

Researchers are continually finding new information and capturing it in various repositories around the world, which is causing information overload [2]. At the same time, more and more people are using the internet to conduct research and find relevant research papers [1]. Therefore, researchers are using ideas from recommender systems and techniques from information search and retrieval [3] to tackle these issues and meet consumers' informational demands. Finding a related or similar paper can be a challenging task. To overcome this, researchers often combine data mining algorithms with document recommendation techniques and information retrieval methods. These methods are applied to research paper features in order to identify the most relevant and important documents that researchers can use. The researcher is able to automatically find, categorise, and suggest electronic documents thanks to a combination of machine learning algorithms, data mining techniques, and natural language processing [4].

When compared to employing individual predictors, feature engineering—the synthesis of several variables that may serve as model predictors—produces better results. It is extremely probable that using a ratio of two predictors will provide better results than using two or more independent predictors [5]. One way to determine whether two papers are comparable is to compare their phrase similarity. The document's length, the frequency of occurrences of common and uncommon words, and the number of instances of a phrase are other potential metrics. In order to perform tasks such as document retrieval, classification, and summarisation, text mining requires the extraction of textual characteristics from documents [4]. As a result, we want to use the word characteristics of research articles to identify commonalities among them.

First and foremost, this study aims to conduct an exhaustive literature evaluation of scientific articles using document search and retrieval techniques, drawing attention to the similarity metrics used. Second, put models for retrieval and search of information through their paces. Third, we will examine potential future paths and provide suggestions when we have established similarity measures to back up our assertions. An accurate classifier for predicting research paper class labels is built into this article's contribution.

II. LITERATURE REVIEW

There are three distinct types of document similarity, as stated in [6]: to begin with, character-based string- space above might be referred to as the space for research papers. Accordingly, a research-paper vector d may be described as the first two being term-based and corpus-based, and the third being knowledge-based (relatedness, similarity). Utilised in this work include similarity measurements based on terms, such as cosine similarity,

Three measures of similarity: Jaccard similarity, Pearson's coefficient, and the distance between two points on Earth. [7] Conducted a comprehensive comparative research on online document similarity measures. They used a number of clustering methods (k-means, weighted graph partitioning, hyper-graph partitioning, self-organising feature map, and random) in combination with four similarity metrics (Euclidean, cosine, Pearson correlation, and extended

Jaccard). While other studies have focused on these four similarity metrics alone, we've taken a new approach by combining them with classification methods like rpart, boosted, and randomforest. They found that weighted-graph clustering performed the best, and that the cosine similarity measure was the best overall. Similarly, our study evaluates the four similarity metrics by comparing their efficiency with other categorisation methods.

After comparing hierarchical clustering methods with partitional clustering algorithms, A. Huang [8] found that the former produced superior results. More than that, we compared and analysed how well different similarity measures worked for document clustering using similarity measures. Three things were determined by their experiments: $im(b, d)$ is a function that quantifies the significance of which word t within the corpus of research papers. First, we must define the significance function.

We choose to utilise the frequency of terms t in the research paper document d as equal to the significance function $imp(Z, d)$ since there are many other approaches to quantify and estimate the value of words in a corpus. Be advised that the vector d does not include all document information as certain details are omitted while calculating the significance function. To get document d ready for the tasks that decide the importance function, a number of steps must be implemented.

Section B: Data pre-processing

Stop words, such as "the," "is," "an," etc., are ubiquitous in academic writing and contribute to the usage of meaningless, unhelpful terminology that routinely slow down computer systems. Conversely, uncommon words and phrases used in a document or research paper are less common but no less significant. "The rank of an important word is inversely proportional to the frequency with which it appears," says Zipf's law. Here is the format it takes: In a text clustering situation, the objects, distance or similarity measurements, and clustering method utilised all have an impact on the final output. Furthermore, it It has been noted that the current diverse set of data mining distance and similarity metrics does not provide particularly clear results when it comes to text categorisation.

In order to determine how similar two texts are to one another, Aggarwal et al. [9] used a hybrid approach that combined knowledge-based semantic similarity with corpus-based semantic relatedness. The machine learning algorithms for linear regression and bagging were trained with all the collected scores. The researchers found that combining similarity-based and knowledge-based metrics yielded far better results in their studies. Measures of term-based similarity are being used in our investigation. In order to get the words, one may look at the title, abstract, tags, etc. of a research article [1]. This study used the names of the research articles to determine how similar they were to one another.

Step Three: Identify the Issue

In a collection of research papers called corpus c, which includes words t, a research paper

document d may be shown as a vector $d\square$. may be represented as $c\square t$, the symbol for a single unit vector inside a corpus. It is important that all unit vectors within a corpus be perpendicular to one another. group of people. Therefore, a set consisting of all unit vectors $c\square t$ in a corpus

in such case the Term-Frequency will reduce the paramount

ordinary, widely used words while elevating uncommon ones. We used Porter's stemming technique to sort the terms [10]. For the purpose of defining the improved notion of significance, let N represent the total number of research publications in the corpus.

The frequency of the phrase t in research papers is denoted as dft and it represents the total number of documents that include the term t. Consider the term frequency as the number of occurrences of term t in document d in a research article, denoted as bfZ,d. So, in order to explain what the acronym tf-idf stands for, we need to look at the

the above illustration, we will apply the following equation

The function $im(b,\ d)$ is equal to $tfb,d \prod lob2\ (N/df)$ (3).

In this case, the inverse document frequency (IDF) is mitigated by using the log2.

C. Locating papers that are comparable to the inquiry

In order to locate a comparable or applicable article, the metadata of the target papers will be used as a query q, which may be represented as a vector $c\square$. Keep in mind that the query is really a research paper—a document vector representation of the same thing. makes up the formal space of the corpus that will match all the words in the dictionary. The

D. Appraisal of Importance

Selecting an appropriate similarity measure for classification involves calculating the degree of similarity between a query vector $c\square$ and a document d inside the corpus $c\square t$.

The development of a reliable and successful research article recommendation system relies heavily on algorithms.

Part IV: Data Analysis

Given the variety of distance measurements at our disposal, it is imperative that

As shown before, the tfidf values are used as term weights in the set $T = \{t1,\ldots, tm\}$.

The equations $\overline{\overline{qt}}$ and $dfid(Z, c)$ are equivalent.

Coefficient of C. Jaccard

The Tanimoto coefficient is another name for this similarity metric; it calculates the degree of similarity by dividing the intersection of two objects by the sum of their unions.

in order to verify that the measurements being utilised are genuine metrics$\overline{\overline{sq}}\,b\cdot\overline{\overline{d}}ot$

The function $SJM(\overline{c}]\lfloor, \overline{\overline{d}})$ is equal to subject to the following terms: The metric space is defined as the set X with a distance function d that, for any element x and y in X that meets the axioms, assigns a real number d(x,y).

The fact that d(x,y) > 0 means that the distance between any two places can't be negative and must be more than zero.

The equation d(x,y) = 0 indicates that if the two items are equal, then the distance between them might be zero, since x = y.

• The distance between any two locations, d(x,y) = d(y,x), indicates that the measurement's starting point makes no difference.

d(x,y) + d(y,z) > d(x,z), which means that the length of the remaining side is at least as long as the sum of the other two sides.

Section A. Concordance

To be considered a valid metric, the cosine similarity metric must adhere to the aforementioned four axioms. At the time
This metric may take on values between 0 and 1. When $\overline{\overline{ct}}$ equals 1, it is 1. This means that the two papers are identical. In the case when the two papers are completely different, it will be zero if the equality conditions are satisfied. In this case, the similarity distance metric is going to be
The reliability of recommendation systems. The datasets utilised for similarity analysis included some verified information based on the annotations on 220 texts authored by eight (8) AI professionals who have published research papers in the area. Using a combination of the cosine similarity measure and the term-frequency inverse-document frequency, the 30 most comparable papers out of a total of 16597 were identified for each of the 220 articles. As shown in Figure 0-1, the dataset's labels were either comparable (positive) or dissimilar (negative). The evaluation.txt file should include 30 more articles that are similar to the ones in the

testids.txt file. This will allow any research paper similarity approach to be tested. There is a record of every paper that was used in the experiment in the file documens.txt.

Part B: Assessment

We utilised a dataset for assessing research article similarity developed by the Ghent University department of computer science to test our system's efficacy. Offline evaluations, online evaluations, and user studies are the three main assessment approaches in the area of research paper recommender systems [11]. The algorithms used 70% of the dataset for training purposes and 30% for testing purposes, dividing the dataset in half.

The outcomes of the experiments

Our research paper recommender system will be based on data mining methods, thus we ran tests to see how well they work. We conducted testing

Three methods were evaluated for their efficiency and accuracy: Random Forest, Recursive Partitioning, and Boosted Tree. After comparing the three algorithms' performance, we found that the rpart method was the most effective and precise, while the other two performed poorly. Research publications were categorised using the rpart machine learning method, which had the minimum running time for the datasets, and the prediction accuracy was improved. Tables 0-1 display the results of the algorithms.

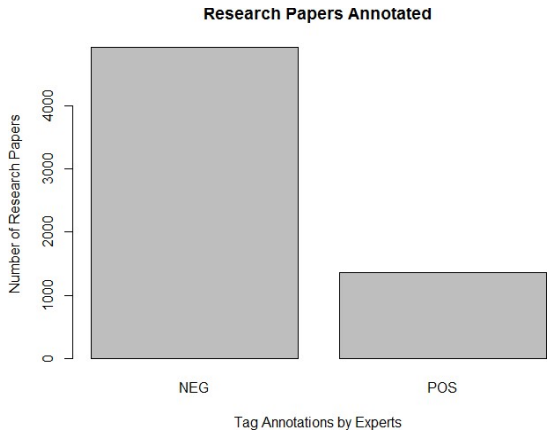

Figure III-1: Proportion of research papers annotated by experts

The similarity between the research papers was accomplished by utilising the cosine similarity. Having measured the cosine similarity, the measures can be taken to collect top-k most similar papers.

Table III-1: Performance of algorithms

	RF	Rpart	Boosted
Time efficiency	39.83342s	2.354628s	41.35908s
Model Accuracy	80.38 %	80.73 %	83.20 %
Area under curve (ROC)	0.6201	0.6201	0.7741

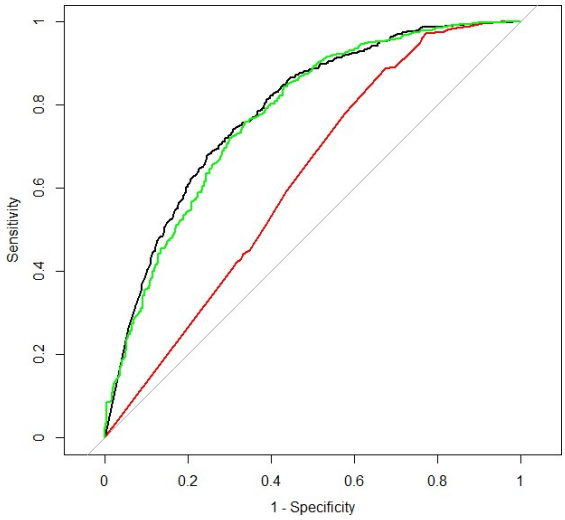

Table III-2: ROC curves of the three algorithms

A. Discussion

The time performance and classification accuracy of the rpart machine learning algorithm were the deciding factors in its selection over the other techniques. With an AUC of 0.7741 and a score of 83.2% for model correctness, the improved performance was remarkable. Unfortunately, processing took 41.35908 seconds, which is very lengthy. Even with an AUC of 0.6201 and a model accuracy of 80.38 percent, the random forest technique took 39.83342 seconds to complete. With a model accuracy score of 80.73 and an area under the curve (AUC) of 0.6201, the rpart method achieved the best time duration of 2.354628 seconds. By calculating the angle θ for each document and returning the N research papers with the shortest angles, the top-N most comparable papers for the query may be rated. One definition of a similarity measure is a function that calculates the degree of similarity in texts. Using similarity metrics, we may reformulate relevance feedback, regulate the number of retrieved documents by enforcing a threshold, and rank documents in order of significance. In text categorisation, using decision trees on a small number of tests usually results in subpar

performance [4]. We present a system that can automatically identify the research paper a researcher is viewing or reading, and then, based on their actions within the paper, extract key features to calculate their document similarity using partitional clustering algorithms, which are better suited to processing large datasets than hierarchical clustering algorithms [8]. In subsequent trials, we will use more sophisticated models, such as latent semantic analysis, which can group texts into the same category despite the lack of common words and phrases.

REFERENCE:

[1] P. Jomsri, S. Sanguansintukul, and W. Choochaiwattana, "A comparison of search engine using “tag title and abstract” with CiteULike —An initial evaluation," in *Internet Technology and Secured Transactions, 2009. ICITST 2009. International Conference for*, 2009, pp. 1-5.

[2] K. Sugiyama and M.-Y. Kan, "A comprehensive evaluation of scholarly paper recommendation using potential citation papers," *International Journal on Digital Libraries,* vol. 16, pp. 91-109, 2015.

[3] A. S. Raamkumar, S. Foo, and N. Pang, "A Framework for Scientific Paper Retrieval and Recommender Systems," *arXiv preprint arXiv:1609.01415,* 2016.

[4] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text- documents classification," *Journal of advances in information technology,* vol. 1, pp. 4-20, 2010.

[5] M. Kuhn and K. Johnson. (2013). *Applied predictive modeling*. Available: http://dx.doi.org/10.1007/978- 1-4614-6849-3

[6] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications,* vol. 68, 2013.

[7] A. Strehl and J. Ghosh, "Impact of similarity measures on web-page clustering," 2000.

[8] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* , 2008, pp. 49-56.

[9] N. Aggarwal, K. Asooja, and P. Buitelaar, "DERI&UPM: Pushing corpus based relatedness to similarity: Shared task system description," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 2012, pp. 643-647.

[10] J. B. Lovins, "Development of a stemming algorithm," 1968.

[11] J. Beel and S. Langer, "A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems," in *International Conference on Theory and Practice of Digital Libraries*, 2015, pp. 153-168.