Atlantis Highlights in Computer Sciences, volume 4
Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021)

A Study of Prototyping in Information Retrieval Process Utilizing Context Requirements and Feature Sample Sizes

Sridevi K N[1,*] Prakasha S[2]

[1] *R&D Centre, RNSIT, Bengaluru, India*
[2] *RNSIT, Bengaluru, India*
[*] *Corresponding author. Email: sridevi.kn23@gmail.com*

ABSTRACT
Information retrieval can provide organizations with immediate value, Information retrieval has become a core of the current information managing systems, this has given profound importance on how we gather effective information in a huge pile of data. Information retrieval is a highly debated issue in a variety of fields. Today right from digital library to media searching information retrieval has taken up the place in the industry. Effective simulation is the foundation of effective information retrieval. These frameworks were divided into two groups based on their statistical estimation methods: models where the terminology dependency is being checked and also the models where the weighting features of the terms are being taken into consideration. Most models found an understanding term freedom, whereas others take this relationship into account. Term weighting, but in the other end, is a technique for effectively indexing a text. The intention of this research article is to compare and contrast various information retrieval structures that use term dependence/independence and term weighting.

Keywords: *Information retrieval system, Mathematical modeling formatting, Term dependency modeling, Term weighting.*

1. INTRODUCTION

A complete process that really can collect, restore, and preserve information is given as an information retrieval methodology [18]. Information (including quantitative and timeline info), voice, videos, and other digital are all examples of knowledge. Information retrieval modelling is critical in assisting scientists in the layout and implementation of an effective information system.

Numerical simulation can be extended to a number of disciplines, including geography, medicine, and so on. The information retrieval framework assists users in forecasting and describing what they can discover applicable to a given issue. Older models which can be of either boolean type or it can also be vector space model and models having probabilistic feature are the three types of numerical retrieval simulation. to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceeding.

2. RELATED WORKS

Classical Process models

There are two defining characteristics throughout this group: Boolean and region models. This prototype has identical comparison in [22]. Logical operators are used in these systems (and, or, not). "Intersection, union, and difference" are the terms of these operations.

Figure 1 shows the mathematical models for retrieving and Figure 2 illustrates the various Boolean system procedures.

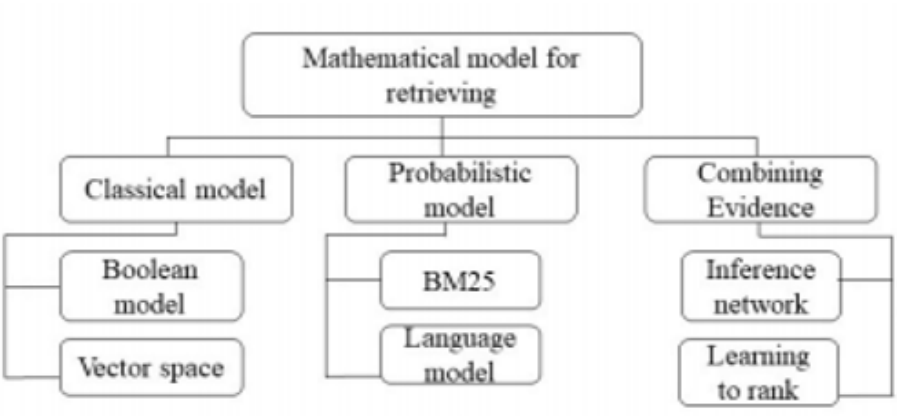

Figure 1 Classification of Mathematical Modeling

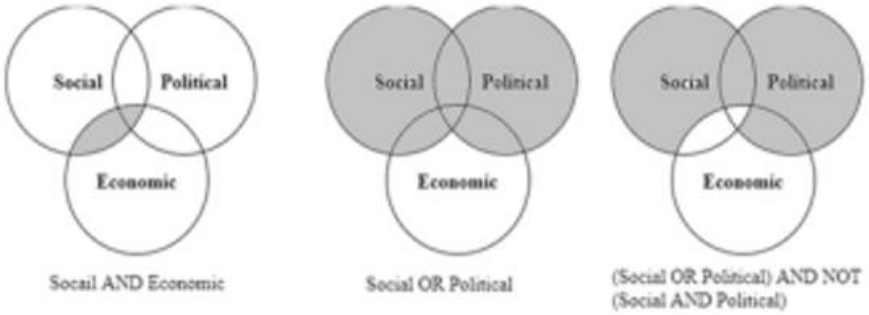

Figure 2 Boolean operation using Venn Diagram

Whose benefits where granted an accurate matching result, but perhaps the documentation was not ranked. There is indeed a distinction amongst these in that the region model is meant to locate semi-structured information.

Using the "Euclidean, Manhattan, and cosine correlation," the graph based study was to compare the similarity "length" among documents and requests. If the cosine angle is negative, the variable is orthogonal; otherwise, the magnitude is zero. The feature space of the applicable and unrelated records is calculated by placing the demand in a vector space model. It improves retrieval efficiency by transferring the query to the clustering point of important instead of just irrelevant data. The three typical drawbacks are the perception of word measurement and the freedom of definitions.

Probabilistic type of Models

Probability distribution is used in statistical methods like probabilistic models, which include two criteria, [22], a related paper, and large requests. The related text is found by calculating the likelihood that now it includes documentation words.

To discern here between appearance and lack of terms in texts, long queries are being used. The fact because both important and non-relevant records are usable is a positive thing. There is indeed a distinction in between stochastic and an experience and understanding methodology. With attribute values and delivery, the probability scheme does not hereby require any supplementary data, although the knowledge-based methodology might.

Utilizing only a combination of compounds Poisson, the Double Poisson Method is constructing a series of numerical rules to classify aggregation terms [7, 22]. The paper became generated at randomly as a source of term instances, according to this template. A combination of two Poisson distributions shown in Equation (1) can also be used to represent a variety of incidents word frequency (TF) of features in files:

$$p(x = tf) = \lambda \frac{e^{-\mu_1 - (\mu_1)^{tf}}}{tf!} + (1 + \lambda) \frac{e^{-\mu_2 - (\mu_2)^{tf}}}{tf!} \quad (1)$$

Here, the term x is presented as a random variable, Also, u1, u2 are regarded as the Mean number of frequencies of a particular term.

These have the merit of not requiring the implementation of an extra term weighting technique. It's one of the most effective terminology weighting techniques on the market.

A stochastic guided path is used in Bayesian clustering algorithms, which are probabilistic concepts. Since there is no direct path between A and Z, a direct graph is cyclic in nature. Through introducing a probabilistic model as a directed graph, graph theory can be used to test abstract conditional probability assumptions.

Pattern recognition has been using the linguistic technique [22]. The auditory template and the language model are two deterministic models used in speech processing. The inference engine could generate a reducing possibility order example. "Good evening," "mood evening," are examples. The proposed approach that decides phrases like "good evening" is even more likely. In English, it appears more often than in other languages. It creates a new paper for each one. The term "retrieval" has a high likelihood in the learning algorithm [22]. If the question includes this term, this means that it is indeed a successful candidate for retrieval. It's useful in situations where prototypes of resemblance expression or material priors are needed.

The famous algorithm i.e Page rank on Google is based on an optimization method. It was used to monitor the effectiveness of web sites [17,22]. The goal of this model is to identify several issues with initial google analytics' information ranking algorithms, which is using word docs for internet pages to obtain intelligence with no clear correlation connection among themselves. A static rating function is what it's labeled.

That's been used because only consistent and dedicated relationships between entities need to be designed. The strengths of PageRank would include possibility that it is indeed a methodology where it is no longer need of queries global metric, although the disadvantages have included the assumption that it would be very simple to improve your own PageRank and that it is always 'buying' a connection on a high-PageRank website.

Combining the Necessary Evidence

Wide text extraction includes a learning to rank algorithm. It's an ensemble learning exercise. Requests and records make up results. Upon this training package, it was also the first to be qualified [26]. It's divided into three main parts those are pointwise, pairwise and list wise.

Consider the method of ranking as if it were a standard grouping. It produces a class as a result. Its aim is to reduce the quantity of incorrect classifications [27]. It converts a rating into a pointwise grouping. Its aim is to reduce the number of graded pairs that are out of order. List wise is a variant of pair - wise. It decreases the loss function.

Ultimately, two tests are also used to evaluate the retrieval method's efficacy. This one is known as the precision rate, which again is defined as the number of valid documents that are actually obtained. Second, there's the recall rate. It's the same as retrieving important records that are currently relevant. We must limit queries if we're to improve accuracy. We extend the questionnaire if we want to maximize the recall. As a result, the relationship between accuracy and recall is opposite [28-30]. The third one, F-measure, blends accuracy and recall.

3. SCHEMATIC ANALYSIS AND DISCUSSIONS

The assumption that definitions are independently and that term calculation is natural but formal seem to be the two main concerns in the vector space model, as shown in Figure 3.

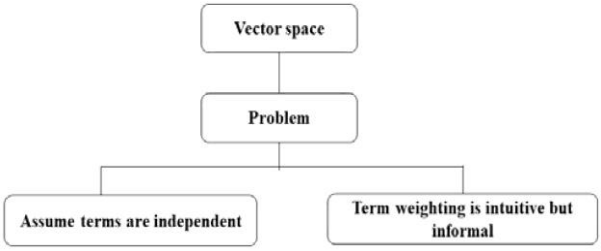

Figure 3 Problems of vector space

3.1 Term Dependency Modeling

Many frameworks, including such Conditional independence vocabulary, ontology, and others, presume how words are independent. The topic of dependency modelling is now being debated [31]. In the given below Figure 4, the modelling independency is shown.

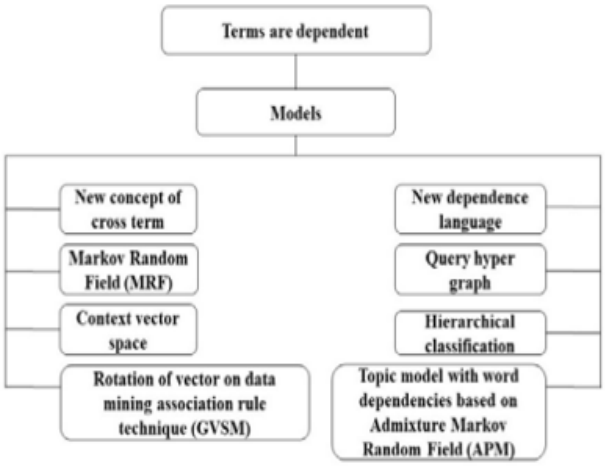

Figure 4 Dependency Modeling

In the paper [1] propose a different definition of the crossover word based on bigram, n-gram, and kernel based process. Cross Phrase with bigram is being used as a starting point, and then n-gram was extended for n queries such that (n> =3). The influence of the search query is defined using form functions [32]. Properties ought to be happy with the purposes (non-negative, continues, symmetric, monotonic and identified). Form has several kernel functions that fulfill the query term's effect, including such (Gaussian kernel, triangle function, circle function, cosine kernel computation using Equation (2), Quadratic kernel computation using Equation (3), Epanechnikov kernel computation using Equation (4) and

triweight kernel computation using Equation (5)). In neural networks, the Gaussian kernel was commonly used. In genomic graphics, the triangle, circle, and Cartesian coordinates were used. The length was calculated using these attributes. As the gap is expanded, the impact is reduced [33-36].

$$kernel\ (u) = \frac{1}{2}\left[1 + \cos\left(\frac{u\pi}{\sigma}\right)\right] \cdot 1_{\{u \le \sigma\}}, \quad (2)$$

Quadratic Kernel:

$$kernel\ (u) = \left(1 - \left(\frac{u}{\sigma}\right)^2\right)^2 \cdot 1_{\{u \le \sigma\}}, \quad (3)$$

Epanechnikov Kernel:

$$kernel\ (u) = \left(1 - \left(\frac{u}{\sigma}\right)^2\right) \cdot 1_{\{u \le \sigma\}}, \quad (4)$$

Triweight Kernel:

$$kernel\ (u) = \left(1 - \left(\frac{u}{\sigma}\right)^2\right)^3 \cdot 1_{\{u \le \sigma\}}, \quad (5)$$

This added a package of terms [2], a bi-term, and a slew of constraints. There are several bi-term variants like Divergence extracted from Randomness, BM5 model as well as several term dependences models (BM25-Span Model, Proportional language model (PLM)). They both presume the words are interdependent. Each interval is evaluated using the BM25-Span model, which distorts the complete width present and count of query terms in each phase. PLM calculates the likelihood of each question in the document occurring in nearby locations. The occurrence of each word in Report is then determined using kernel functions. Gaussian kernel attributes were being used as kernel.

In paper [3], authors were using a modern dependency vocabulary that applies to unigram-based languages. The majority of requirements do not result in an increase in retrieval efficacy when recovering large items. This really is due to two factors.

To begin with, estimating constraints on a broad scale is challenging. Second, and in weighting schema, the incorporation of all single terms and references. The bi-gram linguistic blueprint outperforms the unigram language model. Term constraints are represented as a cyclic, planar, and undirected graph in the new version. In two points, the request is produced from documents: -

The word "linkage created" was coined. This research uses various models including such binary autonomy retrieval (BIR), unigram (UG), dependency model (DM), bi-gram language model (BG), and biterm language model on six different datasets (BT1 and BT2). Specificity improves as a result of these contrasts

Response hyper graphs are being used to catch dynamic correlations amongst query concepts declarations in [4]. Requests serve as vertices and edges in the graph. "Dependencies" represents the distance among edges shown in Figure 4 [23,24]. The question refers to the vertex. A ranking function is extracted from the query hyper graph shown in Figure 5 that deals with definition and dependence concepts. Three key characteristics are integrated into the model presented in this study.

1) As either a definition, design the subjective terms dependencies.
2) It models connections between any of these definitions using passage-level information.
3) Both frameworks are given a weight: definition specifications and key aspects are assigned a weight.

Newswire and online corporate array was used in trials. So many extraction frameworks are improved by using this system.

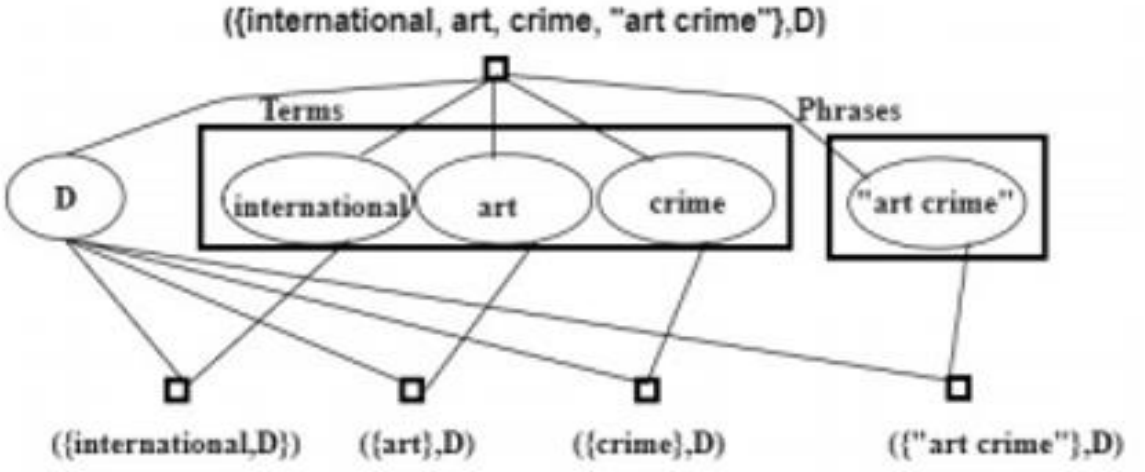

Figure 5 An example of a hyper graph representation for the query international art crime

Markov Chains: To explore complete randomness, simultaneous requirements, and full dependencies, the Random Field "MRF" has been used in [5, 6], with a recent variant in [21]. MRF is built from of the undirected

in [5.] Edges describe autonomy amongst parameters, while nodes represent statistical properties. Create a graph for query term constraints, describe a set of possible functions, and rank the document are the steps of MRF.

Complete independencies (FI) are words that exist without being influenced by other terms.

Markov Chains to explore complete randomness, simultaneous requirements, and full dependencies, the Random Field "MRF" has been used in [5-9], with a recent variant in [21]. MRF shown in Figure 6 is built from of the undirected in [5]. Edges describe autonomy amongst parameters, while nodes represent statistical properties. Create a graph for query term constraints, describe a set of possible functions, and rank the document are the steps of MRF.

Full independencies (FI) are words that exist without being influenced by certain definitions.

SD stands for sequential dependencies, which are constraints between neighbors.

Full dependencies (FD) are words that are interdependent and present a total graph, capturing relatively long connections [37-41].

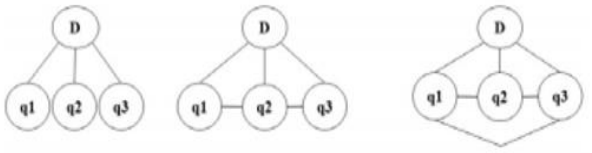

Figure 6 An example of MRF for three query term dependencies (left “F1”, middle “SD”, right “FD”)

To achieve full productivity, the possible plan is responsible. The most consistent with each other under specified set are given the highest value in the good potential function shown in Equation (6). Model - based interactions improves efficacy over a wide spectrum of TREC selection, according to our findings. On the smaller set, concurrent dependencies utilizing structured functionality are more efficient, whereas the complete set is more productive. Greater versions are indeed the finest.

$$\log \varphi_t(c) = \lambda_t \log p(q_i \backslash D) = \lambda_t \log \left[(1 - \alpha_D) \frac{tf_{qi,D}}{|D|} + \alpha_D \frac{cf_{qi}}{|C|} \right] \quad (6)$$

Here IDI is the cumulative number of indicators in articles, ICI is the set's width, tf(qj,D) is the amount of instances a term appears in a report, cfq1 is the amount of instances a term appears in the entire library, and also an is really the toning factor.

A novel Markov Random Field variability based on BM-25 was presented in [21]. Several of the reliance frameworks is MRF. In current history, it has gotten much more coverage. Because with the computational complexity costs, it presents a realistic usefulness rather than efficiency. TREC8, GOV2, and Clueweb09-Category-B samples are all using the new model. It saves up to 60% on costs while maintaining the same level of usefulness.

A novel subject process focuses on a Poisson Markov Random Field (APM) admixture was presented in Equation (7). In comparison to prior individual statistical methods such as PLSA, APM models interactions among sentences. Over nonparametric method. the results showed, Poisson MRFs (PMRFs) provide a JOINT distribution. The PMRF (0, 0) process is characterized as continues to follow:

$$pr_{PMRF}(x \backslash \theta, \Theta) \propto exp\left\{ \theta^T x + x^T \Theta x - \sum_{s=1}^{p} ln(x_s!) \right\} \quad (7)$$

If theta is low i.e. in negative form, the dependence is seldom co-occurred, due to the dependency of parameter theta. Otherwise, they occasionally co-occur. Grolier encyclopedias are used in the tests, and the results are visually pleasing and easy to understand.

The framework in [8] is an expansion of the graph based application that includes data analysis approaches' association rule to find a series of words that co-occur in the information array. Partial matching, strong rating, simplicity, speed, it's the most commonly used concept of term weight are all benefits of the vector space model. The incorporation of correlation details between features in the set of vector space model management to make in this study [28]. A further variation of its feature space is the generalized vector space model (GVSM). Non-orthogonal definitions can indeed be expressed by either a minor variable named "minterms" on binary weight in GVSM. In VSM evaluates the degree of similarity between query and documents based on distance. The similarity between two vectors is calculated using Equation (8).

$$sim(d_j, q) = \frac{d_j \cdot q}{|d_j| \times |q|} = \frac{\sum_{i=1}^{t} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{t} {w_{i,j}}^2} \sqrt{\sum_{i=1}^{t} {w_{i,q}}^2}} \quad (8)$$

And the confidence of two vectors is computed using Equation (9).

$$confidence(A \Rightarrow B) = p(B/A) = \frac{freq(A \cap B)}{freq(A)} \quad (9)$$

For feature selection, apriori defines a fundamental observed data. The co-occurrence of words is detected using this technique. The trust of terms is a vector space with an interaction norm [25]. Trust defines the current angle between the word vectors. Laws are used to put term vectors closer together. If an angle here between vectors Ki and Kj is 90 degrees, so only Ki and Kj are rotated. The angle is calculated using Equation (10).

$$\theta_{i,j} = 90(1 - c_{i,j}) \quad (10)$$

Where, Θ_{ij} is a new vector between (K_i, K_j) and c_{ij} is a confidence of association rule K_i -> K_j.

In [9] used a creative query retrieval system called "hierarchical question grouping." People should ask questions in natural language rather than using section words and respond to questions that have been posted online or offline. The importance of question retrieval has piqued the attention of researchers and the general public. Figure 7 illustrates a high-level description of query retrieval.

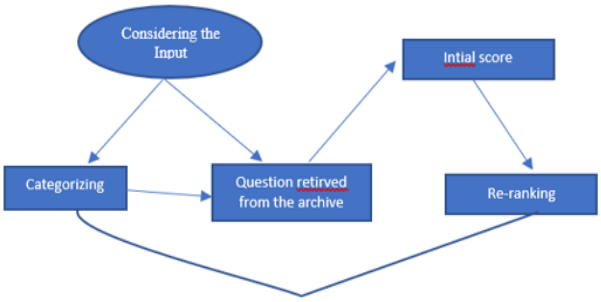

Figure 7 Question Retrieval Process

The usefulness of this approach as opposed to yet another state-of-the-art process is shown by an experiment results on the Yahoo dataset.

Paper [10] used a phrase referring vector model based on term co-occurrence in the same paper. The meaning vector for a document is measured using vectors. Terms that appear around an object in a document are used to create a background vector. The sense of a single event is represented by vectors. The following nxn matrix can be used to describe a collection of word meaning vectors:

$$\begin{pmatrix} c_{11} & \cdots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{nn} \end{pmatrix}$$

It was tested on a validation sample and discovered whether TF*IDF has a strong degree of importance. The word weighting variation is mentioned in [10-13]. Word which appear in all documentation with a higher wavelength than terms that appear in all documents only are more significant. Different instances should boost retrieval efficiency, according to a type of deviation. Table 1 shows various local weight formulae.

Table 1. Local Weight Formula

Formula	Name	Abbr
$1 \quad if \quad f_{ij} > 0$ $0 \quad if \quad f_{ij} = 0$	Binary	BNRY
f_{ij}	Within document frequency	FREQ
$1 + \log f_{ij} > 0 \quad if \; f_{ij} > 0$ $0 \quad if \; f_{ij} = 0$	log	LOGA
$\frac{1+\log f_{ij}}{1+\log a_{ij}} \quad if \; f_{ij} > 0$ $0 \quad if \; f_{ij} = 0$	Normalized log	LOGA
$0.5 + 0.5\left(\frac{f_{ij}}{x_{ij}}\right) \quad if \; f_{ij} > 0$ $0 \quad if \; f_{ij} = 0$	Augmented normalized term frequency	ATF1

When measuring the intensity of terms pairs, the concept calculation is based on a query retrieval system that reveals the affinity between terms pairs. To address this problem, [14] proposes a novel term weighting scheme that incorporates the dependence relation between different pairs. To start, create a dependency graph shown in Figure 8 (a) and measure the frequency of each term's relationship. Second, the original term intensity should be fine-tuned. Any term pair has a dependence link direction thanks to the undirected graph. If the route is shorter, it indicates a closer relationship, as seen in Figure 8 (b).

Table 2 shows global weight formulae which represents various things like inverse document frequency, entropy, global frequency IDF and Table 3 shows normalization factors like pivoted normalization, and another one which is pivoted unique normalization.

Table 2. Global Weight Formula

Formula	Name	Abbr
$\log\left(\frac{N}{n_i}\right)$	Inverse document frequency	IDFB
$\log\left(\frac{N-n_i}{n_i}\right)$	Probabilistic inverse	IDFP
$1 + \sum_{j=1}^{N} \frac{\frac{f_{ij}}{F_i}\log\frac{f_{ij}}{F_i}}{\log N}$	Entropy	ENPY
$\frac{F_i}{n_i}$	Global frequency IDF	IGFF
1	No global weight	NONE

Table 3. Normalization Factors

Formula	Name	Abbr
$\frac{1}{\sqrt{\sum_{i=0}^{n} (G_i L_{ij})^2}}$	Cosine normalization	COSN
$\frac{1}{(1-slope)+slope\, l_j}$	Pivoted unique normalization	PUQN
1	None	NONE

This method is applied using large real-world data set from Yahoo and the result is to improve retrieval effectiveness.

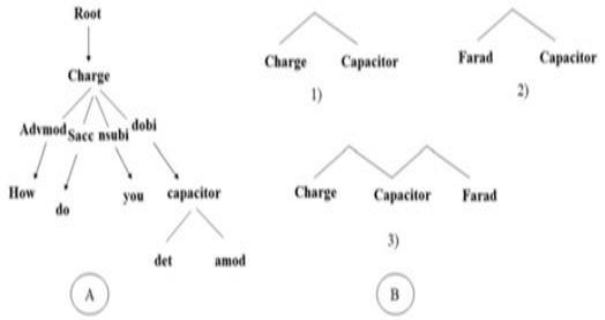

Figure 8 (a) Dependency Paring Tree (b) Dependency Relation Path

In [15], a new system of word weighting is implemented. Trust intensity is the name of this form. It is defined as a mathematical estimate of the word's significance. It also has the advantage of facilitating feature discovery. As an alternative to TF*IDF, trust weight is used. It is capable of performing well even though no functions are included. If a function is meaningless, the weight it receives from trust weight is minimal. Table 4 shows three data sets used to test, like reuters, which consists of groups relating to business news reports, ohsumed, which is derived from vast text collections and is seldom used for all available categories, and documents of term weight.

Table 4. Comparison of Methods

Method	Stable	Sensitive	Accuracy
Gain information ratio	More	Less	Less
Confidence weight	More	More	More
TF*IDF	Less	More	Less

A new word measurement approach was introduced in [16]. This approach doesn't use the query information, but rather information similarity information. We use the information benefit ratio to map cluster similarities to weight. As the quantity of knowledge in a cluster varies after it is annexed into sub-clusters, the volume of material in the cluster is used to calculate the configuration of the sub-clusters. Two sub-clusters make up the cluster. A sub-cluster is a set of recovered reports that would be split into smaller classes. The remainder of the database seems to be another choice. The term "structured similarity" refers to the relationship between texts. The retrieval paper for clustering is based on IGR, as seen in Figure 9.

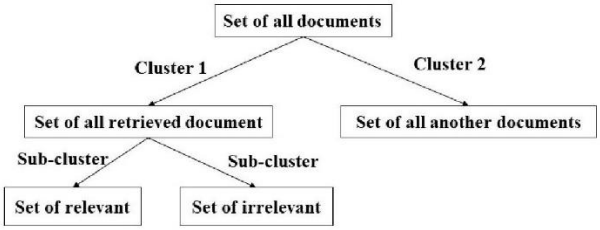

Figure 9 Clustering Retrieval Documents

Term weighting can be classified into three categories (TF, IDF, and IGR). The transmission of each word in records affects TF*IDF. IGR is built around the concept of information clusters and uses the decision variable algorithm to analyze correlation among retrieved documents. As shown below, the distance is determined using the Euclidean distance formula shown in the Equation (11), weight is computed using Equation (12), tf (D_i, w_k) is computed using Equation (13) and idf (w_k) is computed using Equation (14).

$$d(D_i, D_j) = \sqrt{\sum_k \left(weight_{ik} - weight_{jk}\right)^2} \quad (11)$$

$$weight_{ik} = tf(D_i, w_k) idf(w_k) \quad (12)$$

$$tf(D_i, w_k) = \frac{freq(D_i, w_k)}{|D_i|} \quad (13)$$

$$idf(w_k) = \log_2 \frac{N}{df(w_k)} \quad (14)$$

Here N represents total count of documents, ID1I is the number of morphemes in D1, freq (D1, $_{Wk)}$ is the frequency of the word Wk in D• and df $_{(Wk)}$ is the document frequency of the word.

Clustering is a form of unsupervised learning. It aids in the classification of results. A sample is a collection of entities that are identical within one cluster but dissimilar in another. The k-mean algorithm is one of the most widely used algorithms. [17] is where you'll find it.

Text classification is one of the most rapidly emerging fields of research. It is an unsupervised learning strategy that aids in the classification of related documents in order to increase retrieval. Preprocessing in [18, 19] should be completed before implementing

this technique, such as (tokenization, preprocessing and feature extraction).

$$E(c_i) = -\frac{1}{log\ c}\sum_{h-1}^{k}\frac{{n_i}^h}{n_i} log\left(\frac{{n_i}^h}{n_i}\right) \quad (15)$$

Where, C_i is the cluster and n_i is the size

Thirdly, purity: It is computed using Equation (16) and it evaluates the coherence of cluster. To achieve high quality, maximize F-measure, purity and minimize entropy defined in Equation (15).

$$p(c_i) = \frac{1}{n_i} max_h({n_i}^h) \quad (16)$$

The intensity of phrase frequency and reciprocal text intensity is taken into account in traditional TF-IDF, but the importance of several other word features is ignored. The intensity of the word's location is determined by a new approach called TF-IDF adaptation position [20]. On Chinese expression, TF-IDF-AP is used. As compared to traditional TF-IDF, the F-measure of TF-IDF-AP has increased by 12.9 percent. The computation of first occurrence's location evaluation is done using Equation (17).

$$first\ position(word) = \frac{FPBeforecount\ (word)+1}{\sum_{i=0}^{n} count\ (word_i) - FPBeforecount\ (word)} \quad (17)$$

The postion of last occurrence is computed using Equation (18).

$$last\ position(word) = \frac{LPAftercount\ (word)+1}{\sum_{i=0}^{n} count\ (word_i) - LPAftercount\ (word)} \quad (18)$$

Equation (19) shows the formula to compute Adaptive weight.

$$position\ weight = \frac{1}{First\ position\ (word)+Last\ position\ (word)} \quad (19)$$

The weight is computed using Equation (20).

$$weight\ (word, Doc) = \frac{TF*IDF*position\ weight}{\sqrt{\sum_{word,Doc}[TF*IDF*position\ weight]^2}} \quad (20)$$

4. CONCLUSION

It is occurred to perform a survey of requirements structures and weighting methods by reviewing different dependencies modelling and term weighting methods. On six separate sets of TREC dataset, unigram, Bigram, and n-gram models, as well as a new dependency language, are built on unigram to increase accuracy and effectiveness. All six TREC standards are improved in Bigram Cross Term (CRTER2). The efficiency of data analysis retrieval using hierarchical sorting, hyper graphs, and correlation rules has improved, but on various data sets, such as the yahoo response data set, newswire, and site corporation. In addition to association rules, context vector model improves average precision. Also, Context vector improves performance. As the time passes by with this technology growing day by day we can find latest technologies that replace our current system, as time passes by we can still enhance the performance of this system using the help of data analytics approach where we can reform the methodology of fetching the output in much faster desired manner. TF*IDF and Information Gain Ratio are two criteria for determining term measurement that increase efficiency, precision, and time. "Confidence weight" is the latest form, which is a substitute for TF*IDF. Compared to the other techniques, it is more reliable, adaptive, and precise.

REFERENCES

[1] Jıashu Zhao, Jımmy Xıangjı Huang, and Zheng Ye, Modeling Term Associations for Probabilistic Information Retrieval, ACM Trans. Inf. Syst. 32, 2, Article 7 (April 2014), 47 pages, (2017).

[2] Samuel Huston and W. Bruce Croft, A Comparison of Retrieval Models using Term Dependencies, ACM; November 3-7, 2014, Shanghai, China, (2018).

[3] JianfengGao, Guangyuan Wu, Dependence Language Model for Information Retrieval, Special İnterest Group of İnformation Retrieval "SIGIR"; ACM, (2004).

[4] Michael Bendersky, W. Bruce Croft, Modeling higher order term dependencies in information retrieval using query hyper graph, Special İnterest Group of İnformation Retrieval "SIGIR", ACM; USA, (2015).

[5] Donald Metzler, W. Bruce Croft, Modeling Query Term Dependencies in Information Retrieval with Markov Random Fields, Special İnterest Group of İnformation Retrieval "SIGIR" ;(2016).

[6] Michael Bendersky, Donald Metzler, W. Bruce Croft, Learning Concept Importance Using a Weighted Dependence Model, ACM; (2010).

[7] David I. Inouye, PradeepRavikumar, Inderjit S. Dhillon, Admixture of Poisson MRFs: A Topic Model with Word Dependencies, International Conference on Machine Learning; JMLR; volume 32; Beijing, China, (2017).

[8] Silva, I.R., Univ. Fed. de Uberlandia, Brazil, Dependence Among Terms in Vector Space Model, Database Engineering and Applications Symposium; IEEE, (2004).

[9] Wen Chan, Jintao Du, Weidong Yang, Jinhui Tang, Xiangdong Thou, Term Selection and Result Re ranking for Question Retrieval by Exploiting Hierarchical Classification, ACM, November 03-07 2014, Shanghai, China, (2017).

[10] HolgerBillhardt, Daniel Borrajo, Victor Maojo, A Context Vector Model for Information Retrieval, American Society for Information Science and Technology; vol. 53, n. 3; p. 236-249, (2002).

[11] Xiaolu Lu, Alistair Moffat, J. Shane Culpepper; How Effective are Proximity Scores in Term Dependency Models, ACM, November 27- 28 2014, Melbourne, Victoria, Australia, (2014).

[12] DeepikaMatta, ManojVerma, Evaluating Relevancy Of

Words In Document Queries Using Vector Space Model, Journal of Engineering, Computers & Applied Sciences (JEC&AS); Volume 2, No.6, ISSN No: 2319- 5606, (2013).

[13] Erica Chisholm and Tamara G. Kolda, New term weight formulas for the vector space method in information retrieval, Computer Science and Mathematics Division, (1999).

[14] Weinan Zhang, Zhao- yan Ming, The Use of Dependency Relation Graph to Enhance the Term Weighting in Question Retrieval, Computational Linguistics "COLING"; pages 3105-3120; Mumbai, (2018).

[15] Pascal Soucy, Guy W. Mineau, Beyond TFIDF Weighting for Text Categorization in the Vector Space Model, İnternational joint conference on artificial intelligence - IJCAI (2014).

[16] Tatsunori Mori, Miwa Kikuchi, Kazufumi Yoshida; Term Weighting Method based on Information Gain Ratio for Summarizing Documents retrieved by IR systems, Journal of Natural Language Processing.

[17] Taylor & Francis Group, The top ten algorithms in data mining, LLC, (2009).

[18] Gerald J. Kowalski, Mark T. Maybury, Information storage and retrieval systems theory and İmplementation, Second edition.

[19] Monika Gupta, Kanwal Garg, Attribute Weighted K-means For Document Clustering, International Research Journal of Engineering and Technology (IRJET), June-2016, (2018).

[20] Jie Chen, Cai Chen and Yi Liang, Optimized TF-IDF Algorithm with the Adaptive Weight of Position of Word, 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE), (2017).

[21] Xiaolu Lu, Alistair Moffat, J. Shane Culpepper, Efficient and Effective Higher Order Proximity Modeling, ICTIR, 16 , September 12-16, 2016, Newark, DE, USA, ACM, (2018).

[22] Goker, A., and Davies, J., Information Retrieval Models, November 2009, (2009).

[23] Sridevi kN, Dr. Prakasha S. (2021). Analytic comparision of Various Classification CiiT International Journal of Biometrics and Bioinformatics. 10. 74-77.

[24] S. Priyadarsini Panda and J. Prasad Mohanty, "A Domain Classification-based Information Retrieval System," *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, 2020, pp. 122-125, doi: 10.1109/WIECON-ECE52138.2020.9398018.

[25] N. Z. Tawfeeq, W. S. Abed and O. G. Ghazal, "A semantic model of morphological information retrieval: A comparative accumulative analysis," *2020 2nd Annual International Conference on Information and Sciences (AiCIS)*, 2020, pp. 1-6, doi: 10.1109/AiCIS51645.2020.00011.

[26] C. Tian, "On the Storage Cost of Private Information Retrieval," in *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7539-7549, Dec. 2020, doi: 10.1109/TIT.2020.3015818.

[27] M. Kulkarni and S. Kale, "Information Retrieval based Improvising Search using Automatic Query Expansion," *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 2021, pp. 1226-1230, doi: 10.1109/ICICV50876.2021.9388573.

[28] Y. Sun, J. Liu, K. Yu, M. Alazab, K. Lin, "PMRSS: Privacy-preserving Medical Record Searching Scheme for Intelligent Diagnosis in IoT Healthcare", IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2021.3070544.

[29] Z. Guo, L. Tang, T. Guo, K. Yu, M. Alazab, A. Shalaginov, "Deep Graph Neural Network-based Spammer Detection Under the Perspective of Heterogeneous Cyberspace", Future Generation Computer Systems,https://doi.org/10.1016/j.future.2020.11.028.

[30] Puttamadappa, C., and B. D. Parameshachari. "Demand side management of small scale loads in a smart grid using glow-worm swarm optimization technique." *Microprocessors and Microsystems* 71 (2019): 102886.

[31] Parameshachari, B. D., H. T. Panduranga, and Silvia liberata Ullo. "Analysis and computation of encryption technique to enhance security of medical images." In *IOP Conference Series: Materials Science and Engineering*, vol. 925, no. 1, p. 012028. IOP Publishing, 2020.

[32] L. Tan, H. Xiao, K. Yu, M. Aloqaily, Y. Jararweh, "A Blockchain-empowered Crowdsourcing System for 5G-enabled Smart Cities", Computer Standards & Interfaces, https://doi.org/10.1016/j.csi.2021.103517

[33] C. Feng et al., "Efficient and Secure Data Sharing for 5G Flying Drones: A Blockchain-Enabled Approach," IEEE Network, vol. 35, no. 1, pp. 130-137, January/February 2021, doi: 10.1109/MNET.011.2000223.

[34] N. Shi, L. Tan, W. Li, X. Qi, K. Yu, "A Blockchain-Empowered AAA Scheme in the Large-Scale HetNet", Digital Communications and Networks, https://doi.org/10.1016/j.dcan.2020.10.002.

[35] Y. Sun, J. Liu, K. Yu, M. Alazab, K. Lin, "PMRSS: Privacy-preserving Medical Record Searching Scheme for Intelligent Diagnosis in IoT Healthcare", IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2021.3070544.

[36] Rajendran, Ganesh B., Uma M. Kumarasamy, Chiara Zarro, Parameshachari B. Divakarachari, and Silvia L. Ullo. "Land-use and land-cover classification using a human group-based particle swarm optimization algorithm with an LSTM Classifier on hybrid pre-processing remote-sensing images." *Remote Sensing* 12, no. 24 (2020): 4135.

[37] Hu, Liwen, Ngoc-Tu Nguyen, Wenjin Tao, Ming C. Leu, Xiaoqing Frank Liu, Md Rakib Shahriar, and SM Nahian Al Sunny. "Modeling of cloud-based digital twins for smart manufacturing with MT connect." *Procedia manufacturing* 26 (2018): 1193-1203.

[38] Seyhan, Kübra, Tu N. Nguyen, Sedat Akleylek, Korhan Cengiz, and SK Hafizul Islam. "Bi-GISIS KE: Modified key exchange protocol with reusable keys for IoT security." *Journal of Information Security and Applications* 58 (2021): 102788.

[39] Bhuvaneswary, N., S. Prabu, S. Karthikeyan, R. Kathirvel, and T. Saraswathi. "Low Power Reversible Parallel and

Serial Binary Adder/Subtractor." *Further Advances in Internet of Things in Biomedical and Cyber Physical Systems* (2021): 151.

[40] Prabu, S., Balamurugan Velan, F. V. Jayasudha, P. Visu, and K. Janarthanan. "Mobile technologies for contact tracing and prevention of COVID-19 positive cases: a cross-sectional study." *International Journal of Pervasive Computing and Communications* (2020).

[41] Nguyen, Tu N., Bing-Hong Liu, Nam P. Nguyen, and Jung-Te Chou. "Cyber security of smart grid: attacks and defenses." In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pp. 1-6. IEEE, 2020.