International Journal of Soft Computing 8 (2): 72-81, 2013
ISSN: 1816-9503

Affinity Based Nominal Language Model (NLM): A Dynamite Information Retrieval Approach

[1]T. Chellatamilan and [2]R.M. Suresh
[1]Department of CSE, Anna University of Technology, Chennai, India
[2]Jerusalem College of Engineering, Chennai, India

Abstract: In current scenario, retrieving appropriate data from vast data repositories in superlative method is a confrontation. Since, there is a greater competency on the number of pages indexed and the retrieval speed of relevant documents in Information Retrieval (IR) through which the people access information from distinct data stores, disparate methodologies have been developed. In order to reduce noise and to acquire more precise results, researchers proposed an algorithm called Affinity based Nominal Language Model (NLM). The LM basically composes some probability measures which in turn bestows rank to the retrieved documents that afford an essence for the adequate information retrieval process. The methodology comprises the customs of IR such as query expansion, clustering and document ranking and the predominant Nominal Language Model which is working with the description of part of speech of the language that describes the features, combined nouns and adjectives. The NLM in the proposal deals with the affinity rate calculation that is based on the affinity against the document over user concern, combined with probabilistic measurements to find the occurrence rate of a particular term of query within the document. This begets the algorithm to produce optimal results for the applied query with high accuracy rate, less processing time and the minimal use of memory in an adept manner.

Key words: Information retrieval, query expansion, Language Model, WordNet, users

INTRODUCTION

Progression in web mining techniques abets to retrieve the documents facilely with greater precision. As is well known, the main thing has to be noted is how relevant the retrieved document to the user query. Subsequently the enormous growth of content in web, there remains a difficulty in content management and defiance in retrieving in appropriate document. By the consideration of above concerns, researchers developed a dynamite approach called Affinity based Nominal Language Model to provide apt results to the user with efficacy.

The conventional method of information retrieval is completely based on the vector space model which involves in the determination of the direction and distance of the concern to beget the conclusions (Langville and Meyer, 2006). The traditional model for IR abounds with various metamorphoses and in the process it is fed up with a language model to provide lingual informational search. The language modeling approach is consumed in manifold natural language processing operations such as IR, PoS Tagging, Optical Character Recognition, Handwriting Recognition, etc. In the information retrieval process, the language modeling method is consorted with the document amassment with respect to the query input. The annexed documents are ranked regarding probabilistic measures. In the proposal, researchers are highlighting the language model with a unigram model which specifies the probability calculation of hitting an insulated word without the impact from the words pre or post target analysis. The unigram language modeling in IR can be treated as the combination of a bunch of one-state finite automations. The NLM based language modeling accords with the part of speech of a literal language of the given query constitutes the factors with noun and adjectives. The informational query made an attempt to capture the document containing the data which is relevant to the area of analysis. The affinity based NLM proposes a mechanism for adequate information retrieval with various processing mechanisms and similarity based probabilistic calculations. There derived two models for IR using language models called basic retrieval model and the extension of the basic model (Hiemstra, 2001). The basic retrieval model describes the basic matching process of the system whereas the extension model provides query conception methods along with the matching process.

Corresponding Author: T. Chellatamilan, Department of CSE, Anna University of Technology, Chennai, India

The essential operations involved in NLM are Query Expansion, Affinity Rate Calculation, Clustering and Document Ranking. Before researchers begin the process with grabbing queries from user, the approach study with the set of user defined classifiers. Subsequently, the grabbed query will be preprocessed regarding the documents present in every classifier. The above process comes under the training and testing phase of the proposal. The classifier narrows up the searching method to give appropriate results.

Researchers present a framework for proficient information retrieval which provides congruous outcomes. Herewith researchers have used the query expansion that is a process reformulates the user query. While dealing with query expansion, it embeds the recall and precision process with it. The recall process involves in the collection of relevant data from the web document and the precision involves providing accuracy to search mechanism.

The preprocessing of the user concern is followed by the description of the affirmed algorithm NLM. The Nominal Language Model comprises the calculation of stag affinity ratio using the conceits of conditional probability theorem by comparing the user concern with the WordNet. The WordNet is a lingual database for the link language English. WordNet is termed as the abounding lexical database for English that constitutes the group of nouns, verbs, adjectives and adverbs called synsets. Synsets are contrived on conceptual semantic and lingual relations. Consuming the eminence of WordNet in the IR process augments the efficiency and precision rate of the method. Each word of the given query will be analyzed for the affinity ratio calculation using WordNet. The similarity rate will be taken into consideration. There will be an affinity ratio is generated for each word in the user query in similarity basis with the WordNet.

Summation of all the similarity ratios produced by each word is taken as net affinity rate of the concern which makes the search process facile. After all the computations made with the Nominal Language Model, the concept called clustering comes into effect to form clusters over the documents with similar value of affinity ratio. The grouping up of documents is made with the cogitation of the affinity rate values of each document in user query regarding WordNet. By the deliberation of the Net-affinity ratio calculated with the summation value of each clusters similarity rate, the document clusters are ranked. An adept ranking algorithm for a searching mechanism is much substantial for efficient retrieval of data. Probability based ranking methodology which is a traditional ranking mechanism provides a contemporary conviction for the ranking principle. Later, the relevant results of user query will be displayed in the order of the corresponding ranks of the documents. The cogent part of the research is the banding of probability techniques in order to compute the similarity ratio with the user query. The proposal accomplishes a procedure for concrete document retrieval with less processing time. The process elucidates the user with high accuracy rate and utilizes less system memory considerably.

LITERATURE REVIEW

With the aggrandize growth of content volume in the internet, there is an interminable need of methods for the process of information extraction. In order to make the query tracking against the contents stored using language models, a technique called smoothing was used (Chen and Goodman, 1998). The smoothing technique provides statistical modeling for the language modeling procedure. The research comprises the language modeling using N-Gram Model which formulates the probability distribution over the query string. The reearch explored a smoothing methodology for N-Gram Language Model and the future research was assigned as the analysis of the smoothing model with various language models. A research work analyzed the three conventional models on IR namely, Boolean Model, vector space model and the probabilistic model (Hiemstra and De Vries, 2000). There explained a new method for the relevance weighting for information extraction. The above work glossed about the stopword tool and the stemmer tool that uses the Boolean queries for stopword elimination and stemming. In future, the performance of the information retrieval process is further enhanced by relating the relevant weighting of the Language Model with the probability ranking principles. A mechanism with two adaptive language models called a Mixture Based Model and the MAP based model for describing the information retrieval adaptation (Chen *et al*., 2001). The approach was processed with Topic Dependent Language Models. The combination of the mixture based and the MAP based model provided an effective topic tracking mechanism and reducing the error rate up to 2.7%.

The approach can be enhanced with the advanced models in order to reduce the Character Error Rate (CER) considerably. A methodology called Re-Ranking Model regarding document clusters using static cluster and dynamic cluster view was given in (Leea *et al*., 2001). The process bestows the merits of inverted file method and cluster analysis. It provided a significant advancement in similarity search ranking methodology. As future research, plug and play architecture can be embedded with the re-ranking model with a user profile management system in order to improve the performance.

A framework banded the document model and query model with the Bayesian decision Theory of probabilistic functions (Lafferty and Zhai, 2001). The research work focused on Kullback-Leibler Divergence and the

evaluation of Query Language Models. The accuracy rate and the performance can be further amended in future work. Statistical foundations enclosed language model with the enactment of model based feedback approach (Zhai and Lafferty, 2001). The approach calibrated on two distinct methods for the updating query with feedback based Query Language Model. One is Generative Probabilistic Model for feedback document and the next is dependent on the depreciation of the KL-divergence over feedback documents. The pitfalls researchers analyzed over this study are there is a factor called performance decrease when there is the use of large feedback documents. Further augmentation of this research work can be provided with the counter measures for reducing the error and the noise ratio of the retrieved documents.

There proposed a mechanism called Two-Stage Language Model focused on the smoothening of Document Language Model using Dirichlet prior and the second stage worked with comprehend of Query Language Model (Zhai and Lafferty, 2002). The extended work of this study can be traded to impose on the Query Based Language Model with the aspect of predicting the redundancy in the retrieved results. A thesis stated a mechanism for ranking flat queries within the collection of structured documents (Ogilvie and Callan, 2001). The algorithm has the description over storage mechanism and the extracted algorithm which leads to the estimation of unstructured queries. The related research had been planned to the comparison of the retrieved document with the component retrieval. This has been followed by the research work with the exploration of query expansion accomplishing XML document components.

Statistical Language Modeling described in (Croft and Lafferty, 2003) focused on the determination of probabilistic distributions that arrest the statistical regularities in the acquired natural language processing. The study provided distinct methods for IR by the combination of the probabilistic model and the Language Model. There was a description for the integration of Language Models into the retrieval model for better performance and to have better results. An ad hoc information retrieval approach explained by Kurland and Lee (2004) which enhances the information afforded by the document Based Language Model by incorporating the data drawn from the clusters of affined document. Alternative cluster algorithms can be overlapped with the methodology for making a study over distinct results and the observations can be accumulated for reference.

The cluster based retrieval in IR provided significant enhancements in document based retrieval approach which was acquired in an automatic ambience (Liu and Croft, 2004). The approach processed with two Language Models, one for ranking or extracting documents and the next is for smoothing the documents using clusters. There explained the clusters formulated by the static clustering affords optimal results. The future work of this approach is made with the analysis whether the cluster formulated for one collection can be accommodated for some other collections. The further investigation deals with the automatic excerpt of model parameters.

There exists a framework for Dependence Language Model (Gao *et al*., 2004) which is based on the extension of the unigram language modeling approach relieving the independence presumption. The independence presumption termed as the assumption of query in two stages. The first stage involves in generating linkage whereas the second stage focused on the each term formulated is depending on the other related terms with respect linkage. The method provided solution for the problems in classical dependence models such as term dependency estimation and normalization of weight.

A study proposed a mechanism called re-ranking the retrieved document using cluster validation and label propagation by the utilization of intrinsic structure in the vast document data (Yang *et al*., 2006). The cluster validation based on k-means clustering used for the determination of pseudo relevant documents. The ranking of the clustered document is made through the label propagation method. There is a possibility for the evaluation of more approaches for finding the pseudo relevant documents since it was the key factor for the re-ranking mechanism.

There was a mechanism for multilingual indexing for the consideration of distinct languages in IR (Pingali and Varma, 2007). The study composed of the indexing methodology for cross language information retrieval with language modeling techniques. The approach outperformed the restriction over the monolingual conceit of information retrieval. Further enhancement could be studied on the basis of reducing the processing time of the model. An approach defines the query dependent ordering with K-Nearest Neighbor (KNN) with the consideration of significant differences between the queries (Geng *et al*., 2008). The methodology contended to provide different ranking models for different queries which are termed as query dependent ranking. As the future based study, the proposal was enhanced with reducing complexity over online processing of KNN methods. There is no focus on the coverage radius of the k-nearest neighbors to determine the query dependent re-ordering. A rank ensemble approach for optimal searching methodology constitutes proposing cluster attributes and integrating their functionalities through rank aggregation methods (Kurland and Domshalk, 2008). The study can be extended by producing diversity in the list of relevant documents.

The Positional Language Model is another approach of Language Model involved in information retrieval. The key factor of this approach is to construe the Language

Model for each position of the user concern and score the document (Lv and Zhai, 2009). The approach comprises proximity heuristics and the passage retrieval. Position language modeling can be further enhanced with the setting of term specific or query specific evaluation over the user query.

An eminent methodology given by Dai *et al.* (2011) called Freshness and Relevance Ranking (FRR). The ranking methodology provided in this approach is based on the divide and conquer method with the addition of two key elements. The first element used the hybrid labels based on the freshness and relevance of the user concern and the second element determined a query importance factor for document ranking. The approach introduced Criteria-Sensitive relevance ranking based on Divide and Conquer Method (CS-DAC). The modeling based on the occurrence of the hybrid label and tends to ranking mechanism. The document ranking strategies can be made with aggregation of some immense methodologies and calculations. Other ranking architectures can be further implemented to acquire improved results.

The affirmed study outperforms the pitfalls given in the related works. By the devotion of immense computations of probabilistic theorems, researchers acquire effective mechanism for information retrieval. Extracting the relevant data from the large data store against the ample content volume is a great defiance. In such premises, the mechanism provides an adept way for retrieving relevant information.

PROPOSED WORK

In the past decennium, many versions of Language Model have been proposed to prop the process of IR against the enormous growth of content volume in the internet. Henceforth researchers developed a mechanism called Nominal Language Model comprises query expansion, affinity rate calculation, clustering and document ranking. The process commences with query snagging from the user. Subsequently the grabbed query is proceeding with query expansion which is termed as the reformulation of the seed query to improve the performance of the retrieval process. In the context of information retrieval methodologies, query expansion involves in the evaluation of user query. The process of query expansion is followed by the calculations made with nominal language modeling. The NLM involves in the determination of Net-affinity rate by comparing the user query with WordNet data store by the calculations made with the conditional probability conceits. The lingual database named WordNet comprises the data related to English language in which the user query is given. The

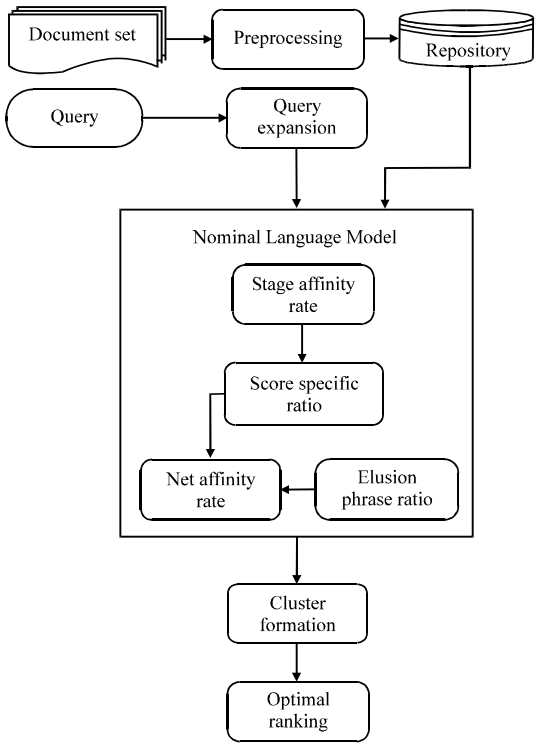

Fig. 1: System architecture

WordNet constitutes the synsets which defines the short, common descriptions and stores the various semantic correlations between these synonym sets. Then, the elusion phrase ratio is calculated for finding the non participation terms of the query against the WordNet data store.

The weight of each term of user query is determined based on its occurrences and the importance and the score ratio is being calculated with the evaluated weight and the Language Model coefficient. The NLM calculations are abided by the document clustering. The clusters are formulated on the basis of the relative affinity rates of each document. Then, the optimal ranking operations come into effect for ordering the retrieved documents in the prioritized manner to make the search facile. The apparent explanation of the methods involved in the proposed affinity based NLM approach.

Figure 1 shows the system architecture of the proposed study. The extraction work starts with the query snagging process from which the process acquires its essence. Then, the acquired query is fed up with the query expansion for analyzing the given query regarding the searching methodology with process such as stopword elimination and the stemming methods. Then, the process is proceeding with the Nominal Language Model descriptions which composed the determination of

stag affinity rate, score specific ratio and Net-affinity rate, the calculations made with probabilistic measurements.

After all the evaluation takes place, the result is given for cluster formation to group up the documents having similar Net-affinity ratio. Then, the modest ranking procedure comes into the act with distinctive ranking methodologies for dashing the results to the used in prioritized manner.

Query snagging: As is well known, the elementary process of the information retrieval is query snagging from the user. The query given by the user is the essence of the process to proceed with. The concern may be a single word or a phrase that will be banded with the results made by the basic initial operations of IR. The process stars with preprocessing of the query regarding the classifiers researchers made earlier for brewing the data extraction process apparent.

Query expansion: In the proposal, researchers emulate the general query expansion process for accomplishing the user concern in accordance with the adequate information retrieval process. The methods of query expansion composed of finding out the synonyms of the given word, the assorted morphological forms of the terms exists in the query, instinctive correction of spelling errors and finally re-write the terms as if in the user concern. The query expansion alsoinvolves in some preprocessing methods such as stopword elimination, stemming, etc. The stopword elimination process involves in dispensing the unwanted terms from the given query and makes the sufficient words for the follow-up procedures. Stemming is the process of cutting up of adorned part of the query in order to make the search brand adroit. The method persuades the user query for effective mining from the large database cogitating the all aspects and features of the query given.

Nominal language model: The Information retrieval process based on NLM is an efficient method for extracting the relevant documents. Language modeling is the conceit processed with natural language processing methodologies. In the proposal, the NLM is assembled with some rate specifications and ratio calculations using probabilistic terms. It involves in comparing the query terms occurrence with the data store with the conditional probability theorem by the occurrence estimation.

Stag affinity ratio: The query expansion process is followed by the calculation of stag affinity ratio. Researchers have applied conditional probability for query term with each term in the document subjected to the number of occurrences. In this proposal, researchers substantially conjugate the semantic methods and the Language Model techniques. This conjugative method of IR audits the distinct meanings of single word given in the query and drifts to the results with greater accuracy. P (Q|W) which is called as LM coefficient for the determination of the occurrence number of each word given in the user concern with WordNet database. The solution of this probabilistic proportion denotes the probability of the presence of query terms and words of each document, respectively. LM coefficient is the verdict of the probability of occurrence of the words in the query against the database regarding the documents. The stag affinity rate is described as the product of the weight of the individual word present in the document with the number of occurrences of the word, here termed as LM Coefficient (LMC):

$$\text{Stag affinity rate} = w \times LMC \quad (1)$$

For each word this affinity rate is calculated and finally iterated for all words present in the document. A document related with the query by the similarity, occurrence and their relationship. In WordNet there exists similarity for some words but not for all terms.

In order to approximate the stag affinity rate, the log value for stag affinity rate will be determined thus made effective evaluation criteria and reduces the fluctuations over the high values determined by the stag affinity rate. Hence, there is a possibility of both the participation and non-participation term present in the user query regarding WordNet, the lexical database for English. In order to improve the consistency of our proposed algorithm according to IR, researchers evaluated the association ratio of participation and non-participation terms of WordNet regarding the user query. The participation terms proceed with the remaining probabilistic measurements and there is also a necessity for determining the ratio of non-participating terms. If researchers can't calculate for each word the accuracy rate is slightly missing. Hence, the non participating terms count is evaluated. The association between the non participation terms and the participating terms are evaluated by applying Kendal coefficient. For non-participating terms the stag affinity rate will be 0. If the result given by the Kendal coefficient equation is negative, it shows there present higher non-participation terms present in the user query. Here, the association between the participation and non-participation word is termed as elusion phrase ratio. The Kendal coefficient is a non-parametric statistic with concordance.

In general, the Kendal coefficient ranges from 0-1 and the value can be made assumption regarding the probabilistic distribution values. The elusion phrase ratio with Kendal coefficient is given by:

$$\text{Elusion phrase ratio}(\tau) = \frac{\left(\frac{P - NP}{P + NP}\right)}{\alpha} \quad (2)$$

Where:

P = The participation terms

NP = The non-participation term regarding WordNet and given here is the arbitrary constant that implies on the range between $0 \le \alpha \le 1$

This calculation is biased for the accuracy determination of the query terms.

Score specific ratio: The dependency between LM coefficient and the conditional probability applied on using the correlation statistics. However, the result was efficient. The score specific ratio is computed for the evaluation of correlation weight and the LM coefficient. The score specific ratio is computed by applying the pearson correlation coefficient to the number of participating items. The determination of score specific ratio is given as:

$$\text{Score specific ratio} = \text{Corr}\left(\text{Weight, LMcoefficient}\right) \quad (3)$$

The correlation calculation defines the dependence statistical relationship between the measured weight and the value of LM coefficient. Correlation coefficient determines the degree of dependency between the given terms against document. The score specific ratio results in producing tedious value. It paves a way for affording unvarying correlation value for further estimations. Thus, monotonousness is evaluated using this correlation calculation. The determination of score specific ratio is a decisive part of the affinity based Nominal Language Model which specifies the correlation between the weight of the words, present in the document set and the Language Model coefficient for excerpting the number of occurrences of the words of user query against the document set.

Net-affinity rate: The Net-affinity ratio is calculated using the stag affinity rate which is calculated for the similarity rate of the query terms regarding WordNet, score specific ratio that specifies the correlation between the weight of the word and the number of occurrences and the elusion phrase ratio for the determination of participation and non-participation terms association. Net-affinity rate can be described further by the ratio of stag affinity rate and the score specific ratio with the destruction of the elusion phrase ratio, mentioned (τ).

$$\text{Net-affinity rate} = \frac{\text{Stage affinity rate}}{\text{Score specific ratio}} - \text{Elusion phrase ratio} \quad (4)$$

Net-affinity rate is for finding the similarity ratio of all the terms given in the user query with all the relevant documents in the classifier. These results in the summation of similarity ratio of all the terms in the query with the entire document set. The terminal calculation of the proposed methodology is the evaluation of Net-affinity rate which summarizes all the similarities and correlates the results. The outcome will be the documents completely relevant to user query with greater accuracy. These documents will be ranked on the basis of the scores for efficient result display.

Cluster formation: Forming clusters describe the alignment of the retrieved documents depending on the similarities predicted over Net-affinity rates. The documents having similar Net-affinity rate will be formed as clusters. The clusters are framed in accordance with the number of retrieved documents based on the affinity rate. The accurate documents will be avail in the cluster having the documents with higher Net-affinity rate. In order to acquire greater accuracy in information retrieval, researchers focus on the cluster with high Net-affinity rate. In the approach, the clustering is done on applying the K-means based on substituting the Net-affinity rate. K-means is a pattern of cluster analysis in data mining which focuses on the partition of n results with k clusters. The K-means clustering process based on the Gaussian distribution factors with the iterative refinement process. The overall score for each cluster is evaluated. The cluster which gains the maximum score is ranked.

Optimal ranking: The documents in the cluster with high Net-affinity rate will be prioritized and ranked. Conferring to the Net-affinity rate calculated using (4) the documents in the cluster are ordered and dashed on the web page. The document with high similarity score will be presented first on the web page. The query has been matched with the web documents on the basis of the methods explained above to give away the results with immense accuracy and efficacy:

```
Algorithm: Affinity based Nominal Language Model
Input: User Query
Output: Relevant documents
Begin: Snagging the user query
    For each document {di|di ε D, 1≤i≤N}
        For each term {tj|tj ε di, 1≤j≤n}
            a. Determine weight and LM coefficient
            b. Calculate stage affinity phrase rate
            c. Evaluete elusion phrase ratio
            d. Calculate scpre specfic ratio
        End for
    e. Determine net affinity rate
    End for
End
```

Algorithm for NLM illustrates the overall procedure of the adduced methodology. The process begins with query snagging from the user. After grabbing the query from the user, the Nominal Language Model calculations are made with document wise analysis and the term wise analysis where the document constitutes. The documents are represented by d_i , ranges from $1 \leq d_i \leq N$ and the term composed inside the documents, represented by t_i that ranges from $1 \leq t_i \leq n$. The determination of weight, LMC, stag affinity rate, elusion phrase ratio, score specific ratio and the Net-affinity ratio is for all the terms and the entire document that comprises the terms. This is an iterative process of evaluation and continues until the determination of last term of the final document mentioned. Then, the process persists with the clustering procedures using K-means followed by the document ranking methodology. The NLM based information retrieval approach with affinity correlations brews the data extraction process more efficient and affords with considerably less processing time than other approaches. The approach outperforms various common language model based information retrieval methods.

EXPERIMENTAL RESULTS

Hence, provide the proof of the affirmed work in consequence of the experimental results. Here, the experiments are carried out with reuters dataset. Researchers initialize our project with training and testing phase. While during the phase of training, preprocessing is being performed with the extracted data files. The training process is processed by using Bayesian approach. There are two classifications will be represented in accordance with the assumption. Let us consider, researchers take two hundred test documents are embraced in the testing phase for precise information retrieval. Those documents are scrutinized into 125 test documents in the training phase, based on the given query terms and the classifiers, determined in the implementation phase. Thus, it reduces the number of test documents up to 37.5% and minimizes the computation time. As is well know from the above descriptions, the files are classified according to the terms present in the provided concern.

The query, researchers are going to acquire should be related to those classifications. The testing process describes the demonstration of extracted data under the defined classifiers. In order to prove the efficiency and accuracy rate of the affirmed work, the produced experimental results are compared with the earlier Freshness and Relevance Ranking (FRR) Method. The conceit of FRR focused on the carrier sensitive ranking

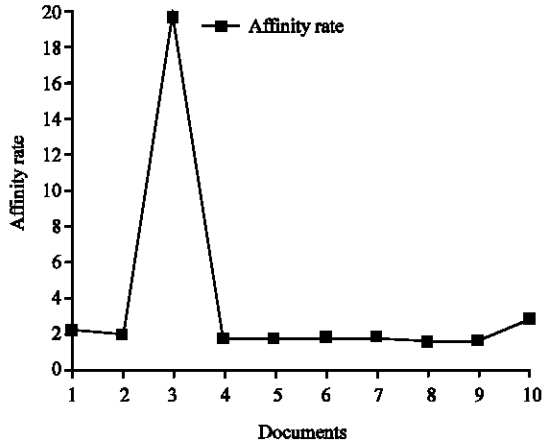

Fig. 2: Documents vs. affinity rate

Table 1: Stipulation of average affinity rate for test samples

Test	Samples average affinity rate
1	3.7133
2	2.5140
3	3.1020
4	1.0250
5	2.6980
6	2.0690
7	3.2670
8	2.7180
9	1.9660
10	3.1370

based on divide and conquer method whereas the predominant work focused on the Nominal Language Model based on the affinity rate of the document set. The process explained later is an immense work to provide more accurate results in an optimal and efficient manner. The documents which are prescribed under classifiers are examined against the affinity rate with the documents related to user concern. Figure 2 demonstrates the relation between the documents over the affinity rate. The affinity rate is predicted by considering the correlation factor over the weight and the LM coefficient of the document. The affinity rate estimation depends on the participation and non-participation terms present in each document. The stag affinity rate is determined by (1) and Net-affinity rate is calculated via (4). Elusion phrase ratio is the term represents the ratio of the association of participation and non-participation terms in a document using Kendal coefficient. The comparison is made with the WordNet database that composed all the information about the prescribed lingual language.

Table 1 examplified the contigency between the average affinity rate and test samples given in the methodology. In the approach, the extraction process starts with the adverse of the documents with WordNet data store. The stag affinity rate is being determined for all

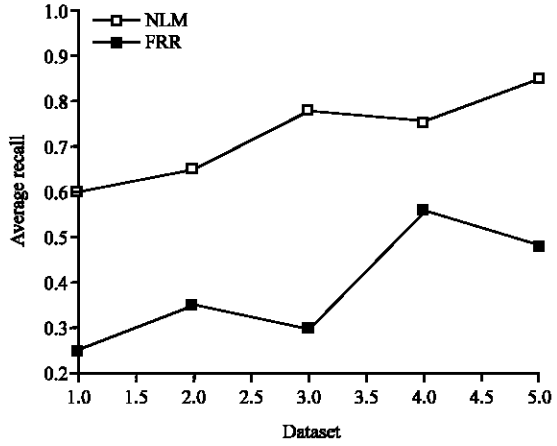

Fig. 3: Dataset vs. average recall

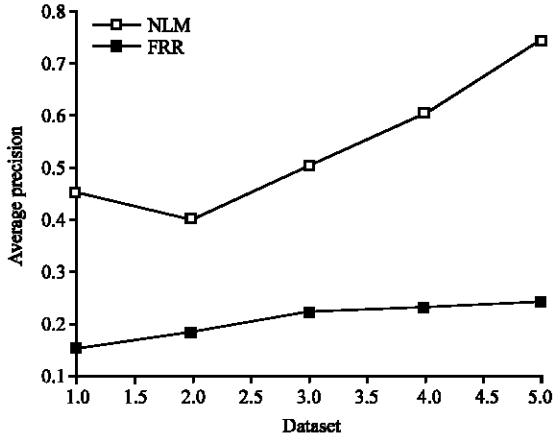

Fig. 4: Dataset vs. average precision

the documents regarding all terms present in the document using conditional probability theorems. For the affinity rate of each terms in a document tends to the estimation of average affinitty rate. The Net-affinity rate evaluated using this stag affinity rate and score specific ratio as in (4), hands us to find documents with high accuracy. The similarity rate is based on the eclusion phrase rate which specifies the association between the participation and non-participation terms of the documents. There is an arbitrary constant mentioned in the eclusion phrase ratio equation α, ranges from $0 \leq \alpha \leq 1$ for approximating the value determined.

Figure 3 elucidates the relationship between dataset and average recall which is evaluated during the IR process using NLM. It is aparent from the above pictorial representation that the proposed methodology NLM produces higher average recall than the existing FRR method. The Query Expansion Method bands the eminence of recall and precision conceit with it for affording accurate search results. Increasing recall conceivably increases the quality of the results with more relevant documents. Recall involves in the retrieval data on the basis of high frequency and close proximity rate. Figure 4 exemplifies that the adduced methodology produce results with high precision rate than FRR approach. Researchers integrate the adroit determinations of conditional probability theorem that provides the affinity rate more accurate by comparing all the terms of entire documents in the data store in an efficient and optimal manner. The efficiency of the dynamite methodology is on the execution speed of the results and the accuracy or precision rate of the acquired result. From various determinations of our methodology, the retrieved data attains high accuracy rate, i.e., the extracted documents are precisely matched with the given user concern.

The above pictorial representation demonstrates the correlation between dataset and execution time for the

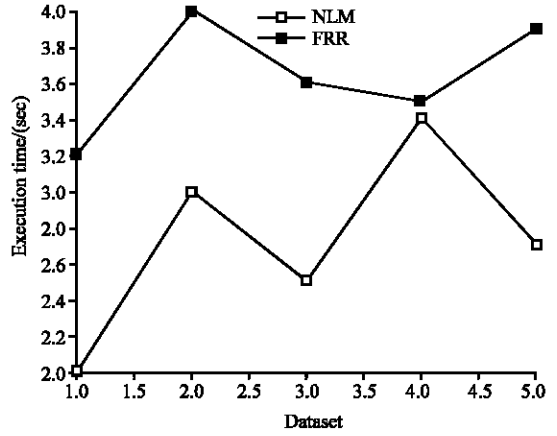

Fig. 5: Dataset vs. execution time

affirmed NLM mechanism and the existing FRR Model (Fig. 5). It's very obvious that the proposal needs less execution time than FRR which uses the carrier sensitive ranking model. The processing time is completely based on the number of datasets present. Researchers described an NLM based approach that specifically reduces the processing time of data extraction from large data store by using proficient method called affinity rate determination in addition to document ranking and clustering. Since, following classification methods and document clustering, the execution time for the retrieval of information is considerably decreased in the work. The probabilistic functions, researchers adopt in the approach provide appropriate results. The above experimental results of the methodology show that the approach is adept in information retrieval and affords accurate results.

CONCLUSION

In the adduced research, researchers have framed a predominant methodology for information retrieval using

the affinity based Nominal Language Model. The NLM based approach comprises the lexical resources of the natural language processing in which the process move along throughout the data extraction process with the given query. Researchers utilized the immense methods of conditional probability theorem for the determination of Affinity rates that makes the approach persuasive. Cluster formulation of the derived results fabricates the retrieval process more facile for the users. The documents in the clusters with high affinity rate are being ranked for arraying the relevant results to the user in an optimal manner by consuming less processing time and high accuracy in retrieving relevant document.

Researchers proposed the approach with the lingual searching methodology where the lexical database comprises the information of single natural language processing. The research can be enhanced for bargaining with multilingual processing on NLM.

Researchers adopted WordNet database for the proposal implementation, perhaps non-participation terms may be higher in referencing WordNet. This can be replaced with some other lexical data stores in the future study.

REFERENCES

Chen, L., J.L. Gauvain, L. Lamel, G. Adda and M. Adda, 2001. Using information retrieval methods for language model adaptation. Eurospeech 2001-Scandinavia. http://perso.telecom-paristech.fr/~chollet/Biblio/Congres/Audio/Eurospeech01/CDROM/papers/page255.pdf.

Chen, S.F. and J. Goodman, 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Harvard University Press, Cambridge, Massachusetts, pp: 359-394.

Croft, W. and J. Lafferty, 2003. Language Modeling for Information Retrieval. Kluwer, Netherlands.

Dai, N., M. Shokouhi and B.D. Davison, 2011. Learning to rank for freshness and relevance. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2011, ACM, New York, pp: 95-104.

Gao, J., J.Y. Nie, G. Wu and G. Cao, 2004. Dependence language model for information retrieval. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 25-29, 2004, ACM, New York, USA., pp: 170-177.

Geng, X., T.Y. Liu, T. Qin, A. Arnold, H. Li and H.Y. Shum, 2008. Query dependent ranking using K-nearest neighborm. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 20-24, 2008, ACM, New York, pp: 115-122.

Hiemstra, D. and A.P. De Vries, 2000. Relating the new language models of information retrieval to the traditional retrieval models. Univsersity of Twente. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.3016.

Hiemstra, D., 2001. Using language models for information retrieval. Ph.D. Thesis, Centre for Telematics and Information Technology, University of Twente, The Netherlands.

Kurland O. and L. Lee, 2004. Corpus structure language models and Ad hoc information retrival. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 25-29, 2004, ACM, New York, pp: 194-201.

Kurland, O. and C. Domshlak, 2008. A rank-aggregation approach to searching for optimal query-specific clusters. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 20-24, 2008, ACM, New York, pp: 547-554.

Lafferty, J. and C. Zhai, 2001. Document language models, query models and risk minimization for information retrieval. Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, LA, USA., pp: 111-119.

Langville, A.N. and C.D. Meyer, 2006. Information Retrieval and Web Search. In: Handbook of Linear Algebra, Hogben, L. (Ed.). Chapman andHall/CRC Press, Boca Raton, FL., pp: 63.1-63.14.

Leea, K.S., Y.C. Park and K.S. Choi, 2001. Re-ranking model based on document clusters. Inf. Proc. Manage., 37: 1-14.

Liu, X. and W.B. Croft, 2004. Cluster-based retrieval using language models. Proceedings of the 27th ACM SIGIR Conference on Research and Development in Information Retrieval, July 25-29, 2004, Sheffield, UK., pp: 186-193.

Lv, Y. and C. Zhai, 2009. Positional language models for information retrieva. Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 19-23, 2009, Boston, MA, USA., pp: 299-306.

Ogilvie, P. and J. Callan, 2001. Using language models for flat text queries in XML retrieval. Language Technologies Institute, Carnegie Mellon University Pittsburgh, PA USA. http://www.cs.cmu.edu/~callan/Papers/inex03-pto.pdf.

Pingali, P. and V. Varma, 2007. Multilingual indexing support for CLIR using language modeling. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. http://sites.computer.org/debull/a07mar/pingali.pdf.

Yang, L., D. Ji, G. Zhou, Y. Nie and G. Xiao, 2006. Document re-ranking using cluster validation and label propagation. Proceedings of the 15th ACM International Conference on Information and Knowledge Management, Arlington, Virginia, November 5-11, 2006, ACM, New York, pp: 690-697.

Zhai, C. and J. Lafferty, 2001. Model-based feedback in the language modeling approach to information retrieval. Proceedings of the 10th International Conference on Information and Knowledge Management, October 5-10, 2001, ACM, New York, USA., pp: 403-410.

Zhai, C. and J. Lafferty, 2002. Two-stage language models for information retrieval. Proceedings of the 25th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland, pp: 49-56.