REAL-TIME CONTEXT BASED INFORMATION RETRIEVAL USING SEMANTIC AND FEATURES ON STREAMING CONTENT

N.Karthikeyan[1], R.Dhanapal[2]

[1]*Research Scholar, Research and Development centre, Bharathiar University Name, TamilNadu, India*
[2]*Research Supervisor, Research and Development centre, Bharathiar University Name, TamilNadu, India*

Abstract

Information retrieval has become one of the major necessities in the current information age. Appropriate information retrieval is still complex due to data heterogeneity and unconventional modes of data dissemination. This paper presents a context based information retrieval architecture to perform real time retrieval of data. This model operates on the semantic similarity of the data, rather than the content similarity. Hence this technique exhibits better and efficient retrieval levels providing high reliability. The retrieved documents are pre-processed and feature vectors are created from the tokens. Polarity matching is used to filter semantically correlated results and magnitude based result ranking is performed to perform further elimination of inappropriate results. Experiments conducted using the proposed model exhibits very high true retrieval rates, along with high precision and recall levels exhibiting the competence of the proposed work.

Keywords: *Information Retrieval, Context Based Data Analysis, Semantic Correlation, Polarity Identification, Sentiment Analysis*

----------***----------

1. INTRODUCTION

Information retrieval has become the major component of the current systems, however retrieving the appropriate information is becoming more and more complex due to the availability of a huge amount of data. This has made the content based retrieval systems obsolete and has led to the raise of context aware systems [1]. Context based information retrieval systems operate based on the internal meaning corresponded by the data, rather than absolute text matching. Hence, context based information retrieval techniques are considered to provide better and more meaningful filters compared to the content based retrieval systems [2]. The basic operations to be carried out by any information retrieval system is to initially identify the context of the input query and relate it to the result set to identify a subset of results that matches the context of the query. This semantic matching is carried out in several ways such as; identifying content based similarity of components, identifying sentiment polarity and the sentiment level of the data and finally by matching the context. This is performed by identifying the significant terms in the text and analyzing their semantic similarity with the query. Conventional models identified significant terms by identifying their frequency of occurrence [3]. However, term frequency is not always considered to be the best measure in identifying the significance of a text. Hence the current systems are moving towards identifying the contextual significance of the documents [4]. This paper proposes a context based information retrieval system operating based on the semantic features of the text. The proposed architecture is parallelized to provide real-time efficiency.

2. RELATED WORKS

Context based information retrieval systems are on the raise due to the increase in the amount of data available online. Several contributions in this domain has been witnessed and this section discusses some of the most prominent and the most recent researches in this domain.

A semantic based information retrieval system based on graph theory was proposed by Kalloubi et al. in [5]. This technique operates based on graph-of-concepts, rather than the conventional bag-of-concepts. A graph is created by considering the relationship among concepts and named entities. This also presents a similarity identification module to correlate graphs and identify appropriate information. Another context based information retrieval technique specifically designed for retrieving information from the web was proposed by Zakos et al. in [6]. This technique operates by identifying the significance of the term during the retrieval process, making the process rely towards the contextual nature of the results rather than the term frequency. A context based document relevance assessment system for effective information retrieval was proposed by Lv et al. in [7]. This technique proposes an enhanced technique for concept representation and document relevance recognition. The vector space model and statistical language model were used as the major components for analysis. The model was constructed to enhance the environmental decision making process. A collaborative learning approach for effective geographical information retrieval was presented by Mata-Rivera et al. in [8]. This technique operates on query contextualization to address the problem of data heterogeneity. A similar technique operating on heterogeneous data based

information retrieval was proposed by Moreno-Schneider et al. in [9]. Other similar techniques incorporating multimedia based information retrieval includes an image retrieval mechanism [10], medical data retrieval [11], multiple textual unstructured data retrieval [12], pervasive computing based information retrieval [13] etc. Concept based retrieval systems are currently in the raise, that are similar to the content based retrieval system. The current contributions in this area includes educational resource identification [14], a portal retrieval engine [15] and a generic system using bag of words retrieval technique [16].

3. REAL-TIME CONTEXT BASED INFORMATION RETRIEVAL USING SEMANTIC AND FEATURES ON STREAMING CONTENT

Context based retrieval of information has become the major need for the current information retrieval systems. The major reason for this requirement is that analyzing the content might not necessarily provide the required results. This leads to inappropriate content presented to the user. However, analyzing the text based on the context rather than the content can provide effective correlations and hence better semantically close content [17]. The proposed information retrieval architecture is presented in figure 1.
Information Retrieval Models in the Context of Retrieval Tasks

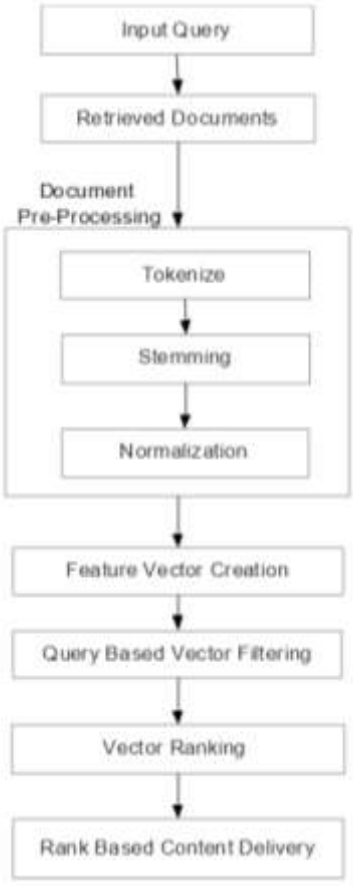

Fig -1: Context based Information Retrieval Architecture

Input query is presented to the query creator, which creates queries based on the data sources. The queries are applied and the corresponding documents are retrieved from the data sources. These contain text, images, multimedia content, XML and HTML content. All textual contents are considered directly and the multimedia data are analyzed with the help of the metadata associated with them. This corresponds to the retrieved documents.

3.1 Document Preprocessing

The retrieved documents contain raw textual data. The data is considered to be in its raw format, because it is laden with several tokens and characters that are not required for the current processing architecture. The data pre-processing phase operates on these impurities to provide cleaned usable data that can be used in the actual processing module. The data pre-processing phase is divided into three sections; tokenization, stemming and normalization.

Tokenization is the process of dividing the data into distinct entities called tokens. This is a heuristic process, performed by setting the splitting constraints, which includes symbols and white spaces. Output from this phase corresponds to individual entities representing tokens.

Stemming is the process of identifying affixed words and eliminating the affixes to provide the root word. The initial stemming algorithm was proposed by Lovins in [18]. Next predecessor is the Porter Stemmer algorithm [19]. This is one of the most widely used stemming techniques. The second version of Porter Stemmer is updated and maintained [20]. This work parallelizes Porter Stemmer 2 and is used as the stemmer. The tokens are first analyzed individually. If the token is identified as a stop-word, it is eliminated from the list, otherwise it is passed to the stemmer for identifying the root word. Stop word identification is carried out using the WordNet 3.0 repository [21].

Normalization is the final phase of the pre-processing module. Stemmer operates on the basis of regular expressions. Hence it strips off any token that matches the patterns, leading to ripping off several proper words. Further, data input by human users tend to contain missed spellings and inflections. Relating semantics to such words is not possible. Hence it necessitates normalization. Normalization compares the tokens with a word lexicon to provide its corrected or synonymous form of the word.

3.2 Feature Vector Creation

A feature set is created using all the shortlisted tokens as the components. Each document is composed of a feature vector. The feature vectors are combined to form a feature matrix. This creates an $n \times m$ matrix, where n refers to the number of documents and m refers to the components of the document.

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

$$a_{ij} = \begin{cases} 0 \dots if\ word_j \notin review_i \\ count(word_j) \dots if\ word_j \in review_i \end{cases}$$

The feature matrix is created with all the reviews components/ tokens. If the token is contained in the document, then the number of occurrence of the component is added to the matrix, otherwise 0 is added and the feature matrix is created. The feature matrix is passed to the next phase for filtering.

3.3 Query Based Vector Filtering

Semantic based analysis and filtering is carried out in this phase on the basis of the input query. Semantic based analysis is carried out by first identifying the polarity of the input query and then matching it with the polarity of the results to eliminate semantically uncorrelated results. Semantic analysis is carried out using the SentiWordNet 3.0 repository [22]. The human annotated nature of the dataset makes it more reliable for processing. Polarity of documents are calculated by identifying individual polarity of the components and aggregating them to identify the final polarity.

$$Polarity_d = \sum_{i=1}^{n} a_{id}\left(Polarity_{(pos,i)} - Polarity_{(neg,i)}\right)$$

Where n is the number of tokens in the document $Polarity_{(pos,i)}$refers to the positive polarity associated with the term i and $Polarity_{(neg,i)}$ refers to the negative polarity associated with the term i and a_{id}refers to the number of occurrence of the token i in the document d.

Polarity identification is carried out for the query as well as the retrieved documents. All the documents matching the polarity of the query are shortlisted for the next phase. This phase is carried out to match the semantic nature of the query with the result sets. Documents are retrieved on the basis of content similarity, providing a first level filtration, this phase acts as a second level filtration leading to results with high semantic correlation.

3.4 Vector Ranking and Content Delivery

Polarity identified in the previous module not only corresponds to the magnitude, but also the intensity of the polarity level in the document. Rank of a document is identified by finding the absolute difference of the polarity of the vector with the polarity level of the query.

$$Rank_d = abs(Polarity_d - Polarity_{Query})$$

Vectors that correspond to the lowest difference are considered to have higher correlation with the input query. Vectors are ranked on the basis of their difference and presented to the user.

4. RESULTS AND DISCUSSION

Experiments were conducted on data obtained from the Twitter corpus [23]. Implementation is done using PySpark and the base data is obtained from HDFS. The input query is applied on the Twitter corpus to shortlist data in the first level. This data is then preprocessed and passed to the feature vector creation phase and finally the polarity identification and ranking phase. Operational efficiency of the vector filtering phase is analyzed and the confusion matrix [24] is created. True Positive Rates (TPR), False Positive Rates (FPR), True Negative Rates (TNR), False Negative Rates (FNR), Precision, Recall, Accuracy and F-Measure are used as the analysis metrics.

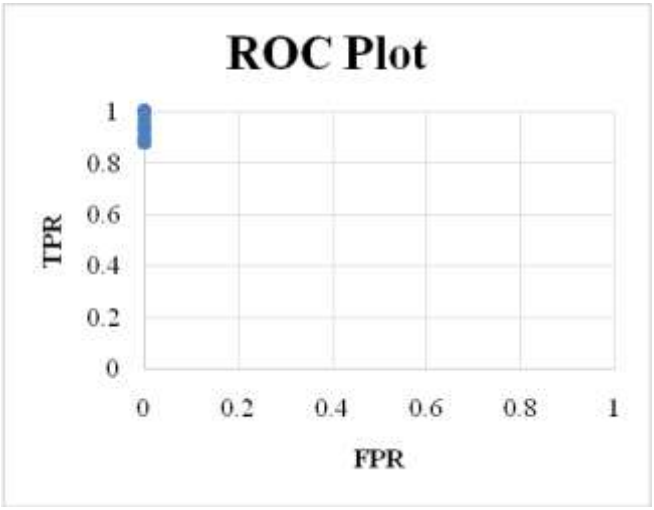

Fig -2: ROC Plot

Receiver Operating Characteristic (ROC) plot constructed using TPR and FPR levels is presented in figure 2. It could be observed that the proposed algorithm exhibits very high TPR levels >0.9 and very low FPR levels (~0). This exhibits the efficiency of the processing algorithm being used.

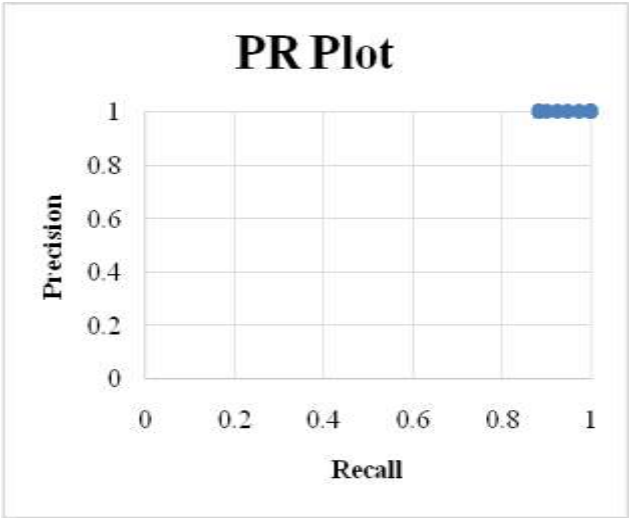

Fig -3: PR Plot

The Precision Recall (PR) plot helps identify the retrieval rates of the algorithm. It could be observed that the proposed technique exhibits very high precision and recall levels (>0.9). This shows the efficiency of the vector filtering phase. The other performance metrics related to the proposed algorithm is presented in table 1. It could be observed from the table that the proposed technique exhibits high accuracy of 0.96 and F-measure of 0.97, hence making the approach affective to be used in applications requiring high performance. However, the true negative rate, i.e. the prediction rate of negative classes is very low (0.37), leading to inappropriate classification of negative classes.

Table -1: Performance Metrics

Metric	Value
Accuracy	0.959966
F-Measure	0.973501
TPR	0.950069
TNR	0.373134
Recall	0.950069
Precision	1
FNR	0.049931
FPR	0

5. CONCLUSION

This paper presents an effective real-time processing technique that can be used to retrieve information effectively from heterogeneous data. Spark based implementations are used to guarantee faster processing. Efficiency of the proposed technique can be observed from the plots. However, the proposed technique exhibits low TNR levels. Future extensions of this technique will be extended to address this issue. Further, it was also observed that the proposed technique does not handle sarcasm. This is attributed to the low TNR levels exhibited by the architecture. Future extensions of this architecture will be programmed to identify sarcasm to make the sentiment identification module more robust.

REFERENCES

[1] C. Manning, P. Raghavan, H. Shutze, "Introduction to Information Retrieval," Cambridge University Press, Cambridge, England, 2009.

[2] P.J. Brown, and G. J. Jones, "Context-aware retrieval: Exploring a new environment for information retrieval and information filtering." Personal and Ubiquitous Computing, 5(4), pp.253-263, 2001.

[3] H. Luhn, "A statistical approach to mechanized encoding and searching of literary information." IBM J. Res. Develop. 1(4):309–317, 1957.

[4] L. Tamine-Lechani, M. Boughanem, and M. Daoud, "Evaluation of contextual information retrieval effectiveness: overview of issues and research." Knowledge and Information Systems, 24(1), pp.1-34, 2010.

[5] F. Kalloubi, and E.H. Nfaoui, "Microblog semantic context retrieval system based on linked open data and graph-based theory." Expert Systems with Applications, 53, pp.138-148, 2016.

[6] J. Zakos, and B. Verma, "A novel context-based technique for web information retrieval." *World Wide Web*, *9*(4), pp.485-503, 2006.

[7] X. Lv, and N.M. El-Gohary, N.M., "Enhanced context-based document relevance assessment and ranking for improved information retrieval to support environmental decision making." *Advanced Engineering Informatics*, *30*(4), pp.737-750, 2016.

[8] F. Mata-Rivera, M. Torres-Ruiz, G. Guzmán, M. Moreno-Ibarra, and R. Quintero, "A collaborative learning approach for geographic information retrieval based on social networks." Computers in Human Behavior, 51, pp.829-842, 2015.

[9] J. Moreno-Schneider, P. Martínez, and J.L. Martínez-Fernández, "Combining heterogeneous sources in an interactive multimedia content retrieval model." *Expert Systems with Applications*, *69*, pp.201-213, 2017.

[10] J.E. Camargo, and F.A. González, "Multimodal latent topic analysis for image collection summarization." *Information Sciences*, *328*, pp.270-287, 2016.

[11] J. Kalpathy-Cramer, H. Müller, S. Bedrick, I. Eggel, A. García Seco de Herrera, & T. Tsikrika,"The CLEF 2011medicalimage retrieval and classification tasks." Working notes of clef 2011 (cross language evaluation forum), 2011.

[12] T. Demeester, D. Trieschnigg, D. Nguyen, & D. Hiemstra, (2013). Overview of the TREC 2013 federated web search track.

[13] J.P. Loyall, and R.E. Schantz, "Using context awareness to improve quality of information retrieval in pervasive computing." In *IFIP International Workshop on Software Technolgies for Embedded and Ubiquitous Systems* (pp. 320-331). Springer Berlin Heidelberg, 2009.

[14] R. Pérez-Rodríguez, L. Anido-Rifón, M. Gómez-Carballa, and M. Mouriño-García, "Architecture of a concept-based information retrieval system for educational resources." *Science of Computer Programming*, 2016.

[15] E. Negm, S. AbdelRahman, and R. Bahgat, "PREFCA: A portal retrieval engine based on formal concept analysis." Information Processing & Management, 53(1), pp.203-222, 2017.

[16] M. Carrillo, and A. López-López, "Concept based representations as complement of bag of words in information retrieval." In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 154-161). Springer Berlin Heidelberg, 2010.

[17] O.L. Golitsyna, and N. V. Maksimov, 2011. Information retrieval models in the context of retrieval tasks. *Automatic Documentation and Mathematical Linguistics*, *45*(1), pp.20-32, 2011.

[18] J. B. Lovins, "Development of a stemming algorithm." MIT Information Processing Group, Electronic Systems Laboratory, 1968.

[19] M.F. Porter, "An algorithm for suffix stripping," Program, 14(3) pp 130−137, 1980.

[20] https://tartarus.org/martin/PorterStemmer/
[21] https://wordnet.princeton.edu/
[22] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In LREC, Vol. 10, pp. 2200-2204, 2010.
[23] http://www.sananalytics.com/lab/twitter-sentiment/
[24] https://en.wikipedia.org/wiki/Confusion_matrix

BIOGRAPHIES

N. Karthikeyan, Msc., Mphil, Research Scholar in Bharthiar University, and Working as HOD (UG) in Computer Science, Srimad Andavan Arts &Science College, T. V. Kovil, Trichy-5.

Prof. Dr. R. Dhanapal. Research supervisor in Bharathiar university. He has got 78 papers on his credit in international and national journals and 13 scholars obtained Ph. D under his guidance and supervision. He has been serving as Editor In Chief for the International Journal of Research and Reviews in Artificial Intelligence (IJRRAI) United Kingdom and serving as reviewer and member of editorial in accredited peer reviewed national and international journals including Elsevier Journals. He has Visited USA, Japan, Malaysia, and Singapore for presenting papers in the International conferences and to demonstrate the software developed by him. He is the recipient of the prestigious „Life-time Achievement" and „Excellence" Awards instituted by Government of India, „Best Professor Award" Instituted by ASDF and Government of Puducherry.