Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 02
Pages: 100-110
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.2.003

Expedient Information Retrieval System for Web Pages Using the Natural Language Modeling

Dr. P. P. Joby,
Professor and Head,
Department of Computer Science and Engineering,
St. Joseph's College of Engineering and Technology,
Kerala. jobypcse@gmail.com

Abstract: Retrieving of information from the huge set of data flowing due to the day to day development in the technologies has become more popular as it assists in searching for the valuable information in a structured, unstructured or a semi structured data set like text, database, multimedia, documents, and internet etc. The retrieval of information is performed employing any one of the models starting from the simple Boolean model for retrieving information, or using other frame works such as probabilistic, vector space and the natural language modelling. The paper is emphasis on using a natural language model based information retrieval to recover the meaning insights from the enormous amount of data. The method proposed in the paper uses the latent semantic analysis to retrieve significant information's from the question raised by the user or the bulk documents. The carried out method utilizes the fundamentals of semantic factor occurring in the data set to identify the useful insights. The experiment analysis of the proposed method is carried out with few state of art dataset such as TIME, LISA, CACM and the NPL etc. and the results obtained demonstrate the superiority of the method proposed in terms of precision, recall and F-score.

Keywords: Natural Language Modelling, Information Retrieval, LSA- Latent Semantic Analysis, Precision, Recall F-Score

1. Introduction

The recent swift advancements and the emergence of innovative technologies, has led to huge flow of data as a result of the daily routine activities that are taking place all over the world. The huge set of data that are flowing holds useful information that are capable enough to improve the quality of the service. The significant insights hidden in the huge amount of information are recovered by employing the information retrieval process. The information retrieval process is an upcoming procedure popularly used in examining the information from the big amount of data set that is in structured semi-structured or unstructured form. The retrieval of information is performed employing any one of the models starting from the simple Boolean model for retrieving information, or using other frame works such as probabilistic, vector space and the natural language modelling.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 02
Pages: 100-110
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.2.003

The enormous amount of information available in the internet requisites an "information retrieval system" to bring out the information that are relevant to the users. For instance a user may start his search for buying an electronic gadget like lap top or mobile phone that belongs to a particular brand. So he would mention electronic gadget –apple, on this case the information retrieval system is very essential to bring out the information relevant to the search instead of displaying the information about the apple fruit.

There are multitudes of possibilities available to the person undertaking the research to handle the difficulties in the information retrieval. The enormous quantity of information flow and the increase in the data and the webpages heightens the difficulties in the process of information retrieval from recovering the information that are useful and as well as reliable. As almost all the search engines are centered on words rather than concepts. While searching particular information's by employing the "information retrieval system an individual is allowed to only use a certain amount of key words in order to narrow down the search. The search outcomes may be relevant or irrelevant or ranging from tens to hundreds. The "information retrieval" so has become a hot topic of study, but the difficulties in providing an effective information recovering system is still unanswered. So to meet the difficulties and overcome the challenges in the information retrieval system an intellectual interface managing the activities of the "information retrieval system" that straight forwardly communicates with the consumer enabling them to have recover information that are relevant eluding the assistance of an human intercessor.

So to address this problem the paper presents the natural language model based information retrieval to recover the meaning insights from the enormous amount of data available in the internet. The method proposed in the paper uses the latent semantic analysis to retrieve significant information's from the question raised by the user or the bulk documents. The carried out method utilizes the fundamentals of semantic factor occurring in the data set to identify the useful insights

The method utilizes the D-score metric to estimate the documents score and determines the rank by processing the inverted index based on the relevance rate. The retrieval of the information is done setting the verge (Threshold) values for the D-score. The search in the method is carried out using the base of the semantic features that are found in the documents rather than relying on the keywords completely. The method is followed to completely eradicate the information's that are irrelevant and recover only the relevant information.

The proffered paper is laid out with the related work presenting the past works carried out in the portion 2, the proposed semantic analysis based on the latent with the estimation of the D-score in the portion 3, the performance validation of the analysis using the dataset "CACM, LISA, NPL, TIME etc." in portion 4 and the conclusion in portion 5.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 02
Pages: 100-110
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.2.003

2. Related Works

Retrieving information's can be performed using different models starting from the simple Boolean model for retrieving information, or using other frame works such as probabilistic, vector space and the natural language modelling. Miller, et al [1] put forth an HMM-hidden markov model for retrieving the relevant insights from the document. The method included mechanism that produced multiple words within the frame work and proved to provide a better performance compared to the state of art model. The figure.1 below shows the expanded multistate hidden markov model.

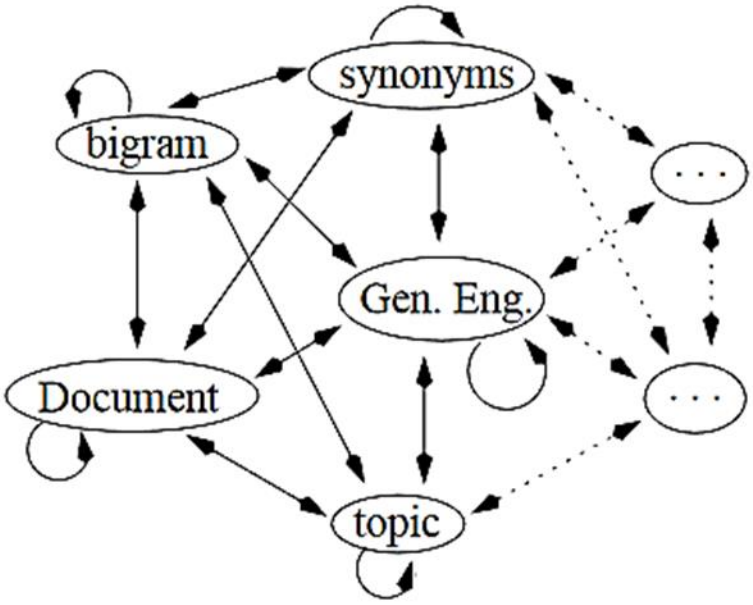

Figure.1 HMM based Multi State Model

Fernandez et al [2] devised the conceptual model with the semantic pattern analysis that essential for small applications in comprehending the low level requirements. The analysis patterns subsidize more to reusability, flexibility and software quality compared to the other varieties. Rosenfeld, et al [3] presents the review on the "statistical language modeling" that determines the variety of natural languages occurrences for the tenacity of speech acknowledgement and extra verbal expertise. Zhai, Jun et al [4] put forth an intelligent transportation using the "semantic information retrieval" centered on the fuzzy ontology.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 02
Pages: 100-110
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.2.003

Thomo et al [5] provides the tutorial on the "latent semantic analysis" Minnie et al [6] presented an "intelligent interface in form of metadata engine for retrieving information in the on various domains. Dhingra et al [7] discussed the demand for the information retrieval that is intellectual and emphasizing on the drawback of the search engine. Dubey et al [8] put forth the enhanced procedure to determine the page rank relying on the optimized normalization techniques. Weston et al [9] provided a "novel learning algorithm for the doing a collaborative retrieval". The mechanism laid put utilized the k-mean algorithm to search and retrieve the information's at a reduced cost. Arora, et al [10] applied the support vector machine to develop the model for effective and intellectual recovery of information's.

Babekr et al [11] the paper exploits the "semantic web technology and the word ontology" in the information retrieval to handle the semantic web technology. Pandian, et al [12] has performed the "Effective Fragmentation Minimization by Cloud Enabled Back up Storage." Weber, et al [13] utilized the "The D-score: a metric for interpreting the early development of infants and toddlers across global settings." Jacob, I. Jeena et al [14] proposed the. "Performance Evaluation of Caps-Net Based Multitask Learning Architecture for Text Classification." Manoharan et al [15] presented a "A Smart Image Processing Algorithm for Text Recognition Information Extraction and Vocalization for the Visually Challenged." Bindhu, V et al [16] has devised the "Biomedical Image Analysis using Semantic Segmentation."

3. Information Retrieval and the Natural language Modelling

The retrieval of information is performed employing any one of the models starting from the simple Boolean model for retrieving information, or using other frame works such as probabilistic, vector space and the natural language modelling. The paper is emphasis on using a natural language model based information retrieval to recover the meaning insights from the enormous amount of data. The method proposed in the paper uses the latent semantic analysis to retrieve significant information's from the question raised by the user or the bulk documents. The paper provides the description of the latent semantic analysis and explains its retrieval process observing the similarity existing between the semantic memory process humans and retrieval procedures. The Latent-SA performs the information retrieval using the query of the users instead of using the keywords that have closer resemblance from the query and the documents.

Latent-SA: Maps the document sets to a "latent semantic space" i.e. a vector space with a minimized dimensionality using the linear-singular value decomposition. The fundamental concept of the L-SA is to represent the document transforming from the word space to topic space. The figure.2 below shows the latent-SA process involved in the converting the original data into topic-encoded data

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 02
Pages: 100-110
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.2.003

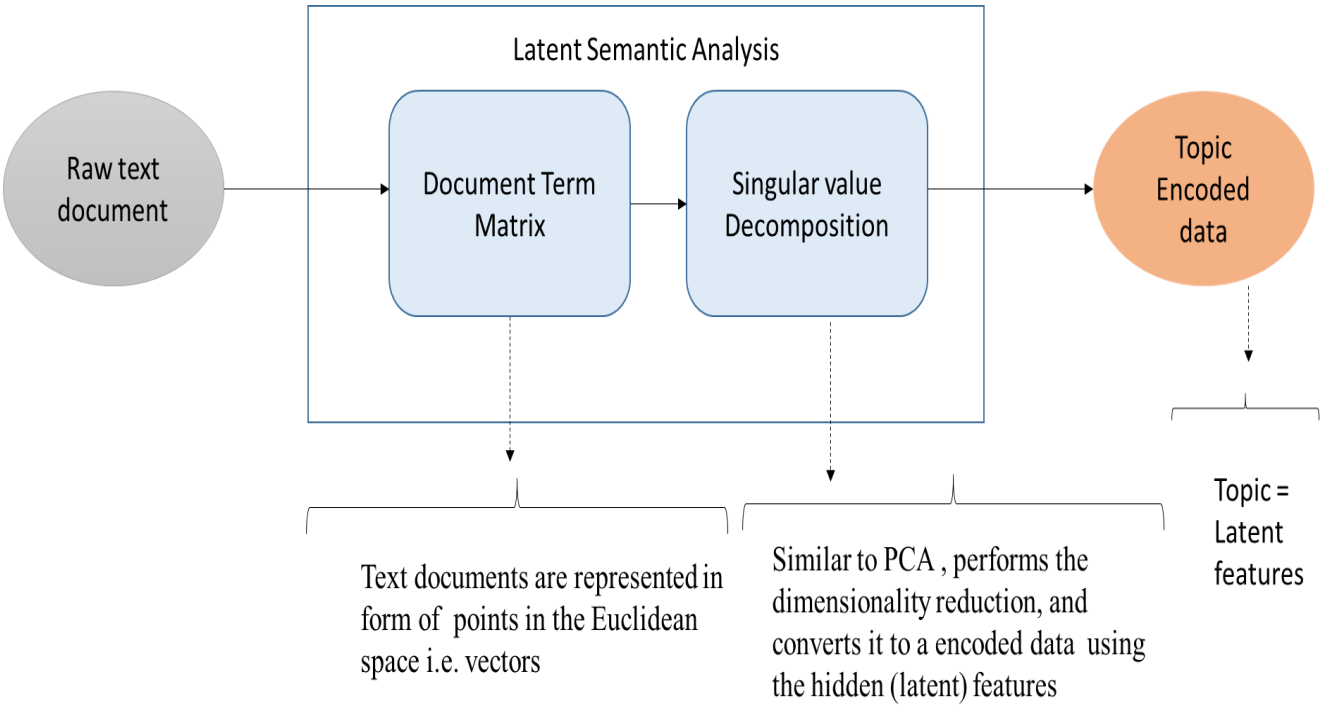

Figure.2 Latent Semantic Analysis

The document or the query is converted into encoded data using the Latent-SA, the singular value decomposition is achieved by the following equation (1)

$$Document_{\ matrix} = OM_1 \sum DM_2^T, \quad (1)$$

Where OM is orthogonal matrix and the DM is the diagonal matrix, 'T' is the text. The query vector is denoted as the $\vec{Q}$ and the document vector is represented as $\vec{D}$. The $\vec{Q}$ is mapped into a latent space indexing using the equation (2)

$$\vec{Q}_{\ l} = \sum_l^{-1} D_l^T \vec{Q} \quad (2)$$

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 02
Pages: 100-110
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.2.003

The mechanism put forth in the paper is encompassed with the following steps to retrieve the information that are relevant from the document. The flow chart in figure.3 shows the step encompassed.

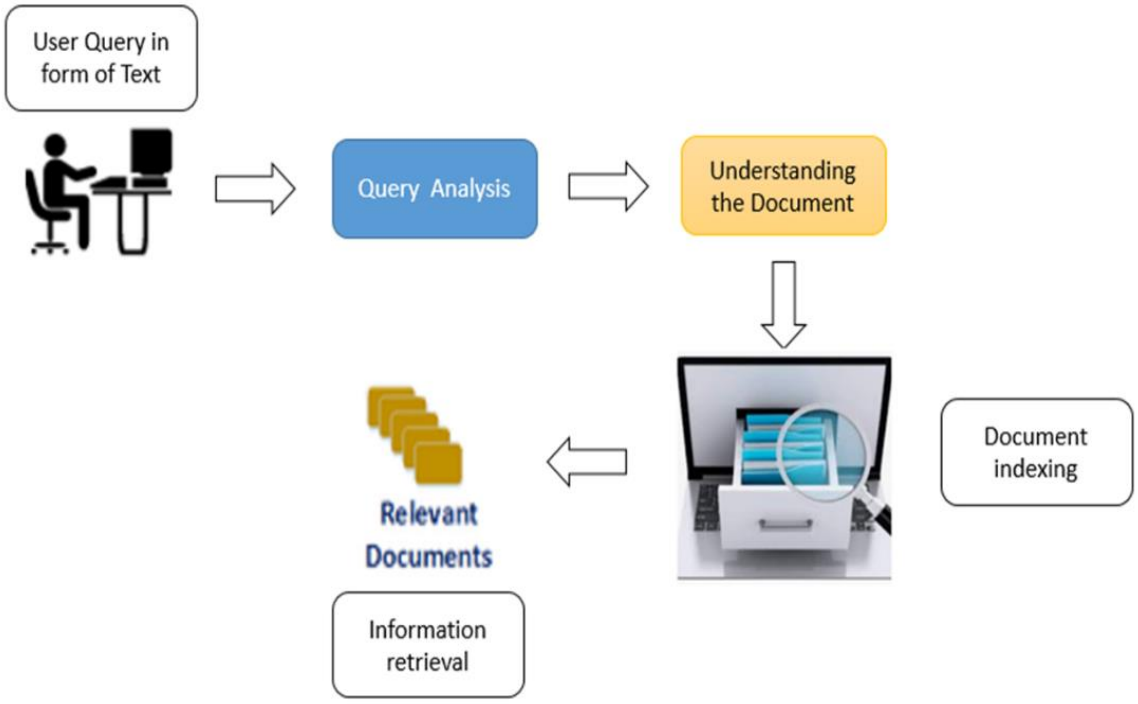

Figure.3 Proposed Information Retrieval Process

User enter the query in the text form using the natural language and the query entered is analyzed to provide the appropriate requirement of the user. In the query analysis the content of the query is estimated and later its intent is predicted. The information about the content and the intent enables in distinguishing the search as (i) information search, (ii) navigation search and (iii) transactional search. The following steps summarizes the query analysis process

- The process starts with the set of query the query are mapped into a $\vec{Q}$.
- For every sentences in the text document the analysis starts with the tokenization where the sentences are broken into pieces or token, every word in a sentence is denoted as a token.
- Then it is verified that the token is not the stop word, if one token is not equal to the stop word the verification proceeds with the next token.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 02
Pages: 100-110
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.2.003

- Once the token are assigned the for every word or the token the stemmer is initialized to remove the unnecessary characters
- Parts of speech tagging is done to determine every tokens parts of speech proceeded by tagging.
- The content and the intent of the text is extracted. (query is processed)

The document is understood by the Latent-semantic analysis [5] and the semantic parsing that performs based on the rules of the grammar and the parts of speech tagging. The synonyms as well as thesaurus are determined by employing the word-net. For understanding the document, the document is read and converted into the document matrix. The document matrix is further compressed into the lower rank matrix and a new matrix with the reduced dimension is obtained. Then the semantic parsing is applied for every sentence and the concept is extracted. The extracted matrix is compared with the new matrix formed to generate the concept matrix. Further the indexing of the document is done employing the "cosine distance and the concept semantic similarity measure" to rank the document based on the query and intensity of the concept is done by determining the average occurrence time of the concept in the document. Further the weight of the semantic is processed which mostly either a 0 or 1. The D-score employed determines the overall rank of every document according to the cosine distance and the weight of semantic. Once the documents is ranked and the score of the each document is estimated using the D-score, the inverted index is calculated based on the level of the information relevance. The threshold value (Th_v) of D-score is set and the retrieval of the document is done. The relevance of the document is estimates based on the equation (3)

$$relevant_{\ document} = \begin{cases} 1 \ if \ Dscore \ \geq \ Th_v \\ 0 \qquad other \ wise \end{cases} \tag{3}$$

The documents are listed in descending order based on the D-score, forming an inverted index and displaying the relevant documents on top.

4. Performance Validation

The mechanism put forth is validated using the standard data sets LISA, TIME, CACM and the NPL, the every data set 50 samples are gathered. So total of 200 queries are gathered the experiments are performed and the offline metric are calculated. The following table.1 gives the particulars of the offline metric evaluated.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 02
Pages: 100-110
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.2.003

Offline metrics	Formulas
Precision	$\frac{\lvert relevant\ document\ \cap retrieved\ document\rvert}{\lvert retrieved\ document\rvert}$
Recall	$\frac{\lvert relevant\ document\ \cap retrieved\ document\rvert}{\lvert relevant\ document\rvert}$
Fall-Out	$\frac{\lvert non-relevant\ document\ \cap retrieved\ document\rvert}{\lvert non-relevant\ document\rvert}$
F –score	$\frac{2.\text{Precision } X\text{Recall}}{\text{Precision } + \text{Recall}}$
Mean Average Precision	$\frac{\sum_{q=1}^{Q} Average\ precision\ (Q)}{Q}$
Normalized Discounted Cumulative Gain	$\frac{Discounted\ Cumulative\ gain\ of\ particular\ ranl\ positon}{ideal\ cumulative\ gain\ of\ rank\ position}$
Discounted Cumulative Gain	$\sum_{i=1}^{position} \frac{relevant\ informattion}{\log_2(i+1)}$

Table.1 Evaluation Metrics

The results in the figure. 4 presents the mean average precision and the normalized discounted cumulative gain for the number of queries tested.

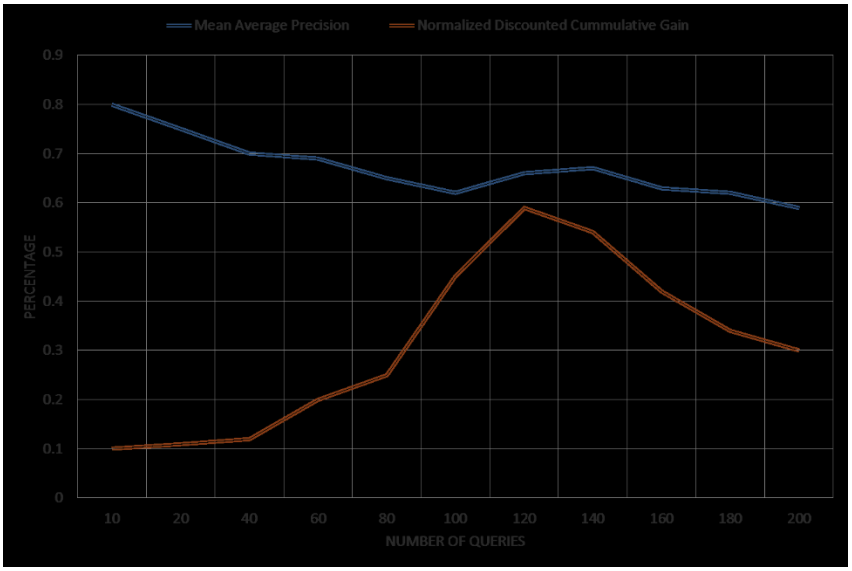

Figure. 3 Mean Average Precision and Normalized Discounted Cumulative Gain

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 02
Pages: 100-110
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.2.003

The Figure .4 below is the precision, recall, F-measure and the fallout observed for the proposed mechanism as well few past mechanism such as the retrieval based on SVM, K-means and the HMM. The results acquired shows that the proposed mechanism out performs the other two with improved precision, recall and minimized fallout and an appropriate F-score.

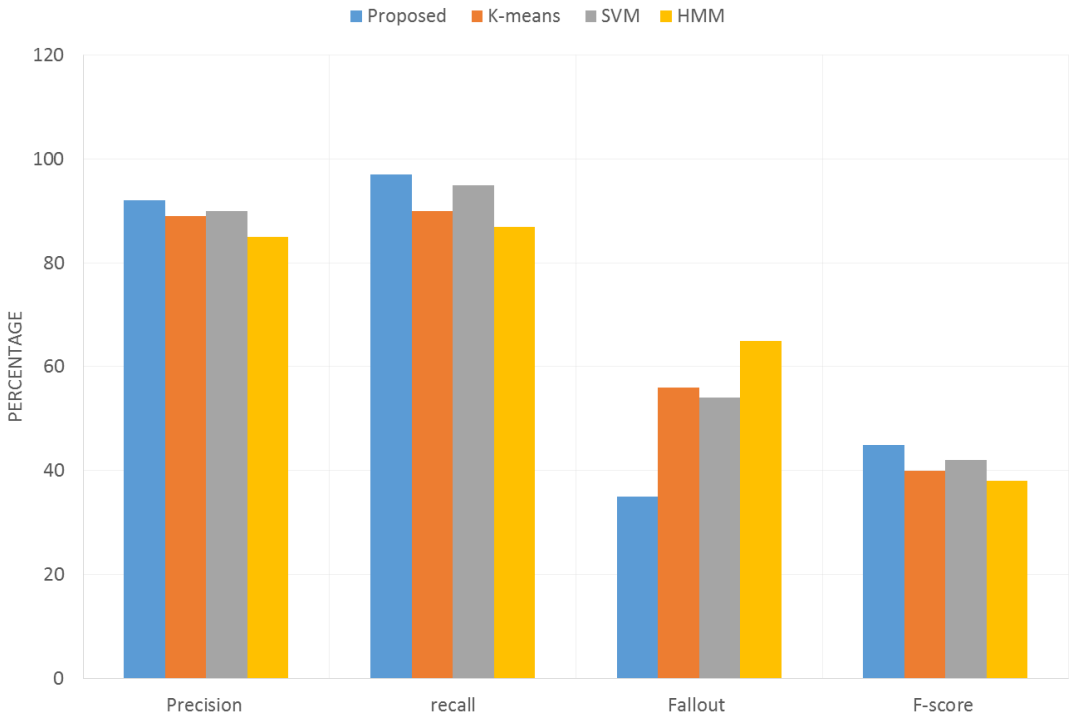

Figure.4 Comparison of Precision, Recall, Fallout and F-score

5. Conclusion

The retrieval of information employing the latent-SA is carried out in the paper, to have a relevant information retrieval from the huge set of data available in the internet. The method utilizes the semantic based analysis instead of retrieving documents based on the keywords. The synonym and the thesaurus are found using the word net in the proposed mechanism. The D-score used in the laid out mechanism ranks the document based on the scores that denote the degree of relevance and are inversely indexed based on the verge level set. The experiments results from the queries gathered from the standard dataset proves the efficiency of the proposed method in extricating the relevant information over the internet.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 02
Pages: 100-110
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.2.003

References

[1] Miller, David RH, Tim Leek, and Richard M. Schwartz. "A hidden Markov model information retrieval system." In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 214-221. 1999.

[2] Fernandez, Eduardo B., and Xiaohong Yuan. "Semantic analysis patterns." In *International Conference on Conceptual Modeling*, pp. 183-195. Springer, Berlin, Heidelberg, 2000.

[3] Rosenfeld, Ronald. "Two decades of statistical language modeling: Where do we go from here?." *Proceedings of the IEEE* 88, no. 8 (2000): 1270-1278.

[4] Zhai, Jun, Yan Cao, and Yan Chen. "Semantic information retrieval based on fuzzy ontology for intelligent transportation systems." In *2008 IEEE International Conference on Systems, Man and Cybernetics*, pp. 2321-2326. IEEE, 2008.

[5] Thomo, Alex. "Latent semantic analysis (Tutorial)." *Victoria, Canda* (2009): 1-7.

[6] Minnie, D., and S. Srinivasan. "Meta search engine with an intelligent interface for information retrieval on multiple domains." *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)* 1, no. 4 (2011): 37-45.

[7] Dhingra, Vandana, and Komal Kumar Bhatia. "Towards Intelligent Information Retrieval on Web." *International Journal on Computer Science and Engineering* 3, no. 4 (2011): 1721-1726.

[8] Dubey, Hema, and B. N. Roy. "An improved page rank algorithm based on optimized normalization technique." (2011).

[9] Weston, Jason, Chong Wang, Ron Weiss, and Adam Berenzweig. "Latent collaborative retrieval." *arXiv preprint arXiv:1206.4603* (2012).

[10] Arora, Monika, Uma Kanjilal, and Dinesh Varshney. "Efficient and intelligent information retrieval using support vector machine (SVM)." *Int. J. Soft Comput. Eng.(IJSCE)* 1, no. 6 (2012): 39-43.

[11] Babekr, Salah T., Khaled M. Fouad, and Naveed Arshad. "Personalized semantic retrieval and summarization of web based documents." *International Journal of Advanced Computer Science and Applications* 4, no. 1 (2013).

[12] Pandian, A. Pasumpon, and S. Smys. "Effective Fragmentation Minimization by Cloud Enabled Back Up Storage." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 01 (2020): 1-9.

[13] Weber, Ann M., Marta Rubio-Codina, Susan P. Walker, Stef van Buuren, Iris Eekhout, Sally M. Grantham-McGregor, Maria Caridad Araujo et al. "The D-score: a metric for interpreting the early development of infants and toddlers across global settings." *BMJ global health* 4, no. 6 (2019).

[14] Jacob, I. Jeena. "Performance Evaluation of Caps-Net Based Multitask Learning Architecture for Text Classification." *Journal of Artificial Intelligence* 2, no. 01 (2020): 1-10.

ISSN: 2582-2012 (online)

[15] Manoharan, Samuel. "A Smart Image Processing Algorithm for Text Recognition Information Extraction and Vocalization for the Visually Challenged." *Journal of Innovative Image Processing (JIIP)* 1, no. 01 (2019): 31-38.

[16] Bindhu, V. "Biomedical Image Analysis using Semantic Segmentation." *Journal of Innovative Image Processing (JIIP)* 1, no. 02 (2019): 91-101.

Author's Biography

Dr. P. P. Joby, is currently the Professor and Head, in Department of Computer Science and Engineering's. St. Joseph's College of Engineering and Technology, Kerala. His research area includes Machine Perception, Robotics, Cyber-Physical Systems, Internet of Things, Complex Networks, Quantitative Network-based modelling, Complex and Intelligent Systems, Networks, Recommendation System, Human-Computer Interface, and Knowledge Representation

ISSN: 2582-2012 (online)