 ISSN: 2454-4248
17 – 24

Information Retrieval and Query Ranking of Unstructured Data in Dataspace using Vector Space Model

Niranjan Lal
Computer Science & Engineering
Research Scholar, Suresh Gyan Vihar University, Jaipur, Rajasthan, India"
niranjan_verma51@yahoo.com

Shamimul Qamar
Computer Science & Engineering
Computer Network Engineering Department,
King Khalid University, Kingdom of Saudi Arabia
jsqamar@gmail.com

Savita Shiwani
Suresh Gyan Vihar University, Jaipur, Rajasthan, India
savitashiwani@gmail.com

***Abstract*—** There is a vast amount of data is available on the web in the form of WebPages, on the clouds or in the repositories of any organization. All data are stored digitally by any companies, enterprises or any organization, these data may be text data, streamed data, images, Facebook data, Twitter data, Videos and other documents available digitally on the Internet related any areas like manufacturing, engineering, medical, etc. collectively called Dataspace. The data available over the internet may be structured data, unstructured or without any format. The storing mechanism is different for each organization but searching and retrieval of data should be easy from the user's point, they are able to find the relevant information efficiently and accurate information that should be satisfied them, so there should be a proper model, search engine or interface for finding the information. Retrieving information from the Internet and large databases are quite difficult and time-consuming especially if such information is unstructured. Several algorithms and techniques have been developed in the area of data mining and information retrieval yet retrieving data from large databases continue to be problematic.

In this paper, the Vector Space Model (VSM) technique of information retrieval is used, by using VSM model documents and queries can be represented as a vector, whose dimension is considered as terms to build the index represent the unstructured data. VSM is widely used for retrieving the documents and data due to its simplicity and efficiency work on a large number of datasets. VSM is based on term weighting on document vectors using three steps 1) First step is used to create indexes of the documents to retrieve the relevant data, 2) In the second step weighting of the indexed terms is used to retrieve the appropriate document for the end user, and (3) In the Finally steps the similarity measures is between documents to rank the documents relevant to the end user query using.

The cosine measure is often used. We then found out that it is easier to retrieve data or information based on their similarity measures and produces a better and more efficient technique or model for information retrieval.

Keywords- *Information Retreival, Ranking Indexing, Vector Space Model, Untsructuctured Data, Dataspace*

*****

I. INTRODUCTION

Information Retrieval from heterogeneous information systems is required but challenging at the same as data is stored and represented in different data models in different information systems. Information integrated from heterogeneous data sources into single data source are faced upon by major challenge of information transformation were in different formats and constraints in data transformation are used in data integration for the purpose of integrating information systems, at the same is not cost effective.

Information retrieval from heterogeneous data sources remains a challenging issue, as the number of data sources increases more intelligent retrieval techniques, focusing on information content and semantics, are required.

Dataspace has recently gained much attention in machine learning. In many large/small organization or enterprises, managing the heterogeneity among data at various levels has made a challenging task for its management community. In an organization, data may vary from fully structured to completely unstructured. The existing data management systems fail to manage such data in an efficient manner. Now, Dataspace technology addresses the problem of heterogeneity present in data and solving various shortcomings of the existing systems.

Retrieving information [1,[2,3] from the Internet and from a large database is quite difficult and time consuming especially unstructured information [4,5]. A lot of algorithms and techniques have been developed in the area of data mining and information retrieval yet retrieving data from large databases continue to be problematic. In this research work, we used the vector space model to rank information based on computed cosine. First, we computed the similarity scores

 ISSN: 2454-4248
17 – 24

using the weighted average of each item. The cosine measure is then used to compute the similarity measure and to determine the angle between document's vector and the query vector since Vector Space Models [6, 7] are based on geometry whereby each term has its own dimension in a multi-dimensional space, queries and documents are points or vectors in this space. The cosine measure is often used.

Information retrieval, Basically Information retrieval is the process of revelation of stored data in the organization, enterprises, universities, that extract the relevant information from this storage.

Storage and management of information should be in the proper way, so the end user can access relevant and accurate information from Dataspace, within stimulate time period efficiently.

Structured Data [8, 9]: It concerns with all the data which is stored in the tabular format, it contains any information that consists of a fixed field placed called structured data like details stored using spreadsheets and relational databases. Structured data is increasing slower that unstructured data.

Semi-Structured Data [8,9]: with structured and unstructured data, there is no predefined format or schema in semi-structured data it contains some format for storing information. , this is not residing in a relational database but has some structural nature that makes them relaxed to analyze. There are many semi-structured models have introduced for this types of data like; XML documents, NoSQL database, and some object-oriented based databases.

Unstructured Data [8, 9]: Unstructured data is everywhere that is generated by humans or storing in the organizations. It refers to the fact there is no proper format available for this type of Data. Unstructured data cannot be stored in tabular formats. It includes text and multimedia data. Examples are e-mail, MS word documents, videos, photos, audio files, presentations, Web Pages and many other kinds of business documents.

Heterogeneous Data [8, 9]: Traditional relational database systems work on simple queries with structured data, those are in the proper format, while information retrieval systems should work for more ranked keyword search queries over Dataspace.

Dataspace [8, 9, 10, 11, 12]: Dataspace is the collection of various types of data stored in different formats in organizations, those data are heterogeneous in nature, that includes text documents, LATEX documents, XML repository, RDB, code collection, online data available on web, a different packages used, repositories used for emails, those are stored in Dataspace in heterogeneous format and on distributed systems with set of relationship of them .

Ranking Unstructured Data [9, 13]: Ranking gives an appropriate and efficient a result of documents searched on the web, for given information is obtained by their relativity is an intrinsic part of each search engine. The ranking function evaluates the power of searching by giving a good chance of relevant results is found on Dataspace.

In information retrieval system, a ranking of documents is done according to matched query for phrase keywords. Ranking problems arise in Dataspace due to the heterogeneity of data; in this case end user can retrieve any type of information's from the Dataspace according to their preference like, Reports, Research papers, Titles, E-books, PDF, PPTS, Email etc., from University Dataspace. Songs, Lyrics, Movies etc. from Multimedia Dataspace and medical data or biological data from the health Dataspace. There are many applications where it is desirable to rank rather than to classify instances, for instance: ranking the importance of web pages, evaluating the financial credit rating of a person, and ranking the risk of investments. When the end user entered a query, the index of information is used to get the documents most relevant to the end user query and resulted documents are then ranked according to the importance and their degree of relevance, different information retrieval techniques are shown in Fig.1[9].

II. REVIEW OF LITERATURE

We have studied several papers and articles on search engines, information retrieval, categories of data, unstructured data, clustering techniques and clustering algorithms for unstructured data, vector space model, indexing techniques, Dataspace, and Indexing of Dataspace, some researchers work related to our area described here. In this section, we have discussed the related work to our research area.

Gordon M. et. al [14] discussed the effectiveness of search engines for the users for retrieval of information from the World Wide Web. For internet access there are many search engines developed, these search engines provide the surprising information to users, As the search engine development market is growing day by day since its inception, the most prominent search engines.

Nielsen Netrating [16] and **the economist Survey [15]** described some important search engines like; GOOGLE, YAHOO!, OPEN TEXT, ALTAVISTA, EXCITE, INFOSEEK, HOTBOTARE.

Baeza-Yates et. al [1] have covered the Information retrieval work, they have also explained information retrieval mechanism for schema designing, storing the information,

 ISSN: 2454-4248
17 – 24

mapping of information, and access to information items. Information retrieval represents, store and organized the information in an easy way, so users can find the relevant information accurately and efficiently.

BELKIN, N. J . et. al[17] discussed the main tasks to querying and searching techniques for the user's information retrieval system. They have discussed the different Information retrieval component. They have also pointed out the research challenge and requirement to support querying and searching for all participants regardless of their different data model.

Niranjan Lal et. al[9] and **Rolf Saint et. al[18]** covered the different types of data available in organizations or any enterprises locally or globally. They have also explained the categories of data with proper examples.

The storing and management of every day growing amount data, changing the format of data is the challenging task for the organizations. Not long ago, datasets contained thousands of data items. Currently, different technologies can store, manage, and process data with increasing volumes of unstructured data

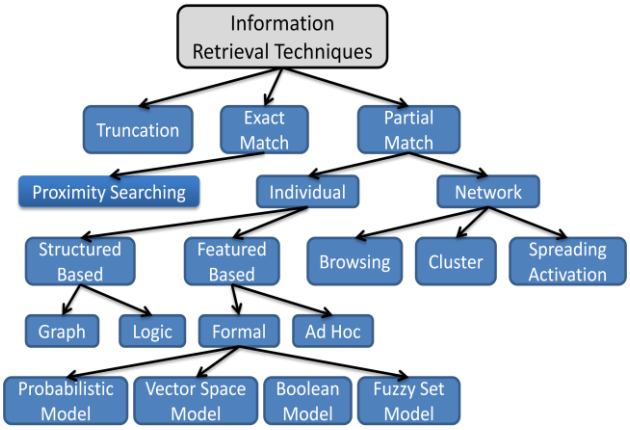

Fig.1 Information Retrieval Techniques

Berry et al. [19] propose a vector space model called orthogonal factorization matrix for retrieving information in a large database. The technique uses mathematical concepts from linear algebra for determining the weighted value of terms in documents using matrix approach whereby terms are represented in rows and documents are represented in columns each document is then checked against the query to determine the frequency of occurrence of each term. These are then represented in a matrix form and the cosine score of each ranked document is then computed. The resulting matrix of the term-by-document is then normalized and then transposed by ranked reduction by using QR factorization followed by SVD to identify and remove redundant information in the matrix representation of the database. By using the ranking of the term-by-document matrix, a geometric interpretation of the vector space model is formed.

Singh and Dwivedi [20] discuss the various approaches of vector space model to compute similarity score of hits in information retrieval. These approaches are Term Count Model, TF-IDF model and the vector space model based on normalization. Based on the similarity function between vector document and query term, the similarity function is computed using database collection of retrieved documents, query and index term. The term frequency-inverse document frequency (Tf-idf) is used to determine how important a word is in a document based on weighting factor in information retrieval and converts the textual representation of information into a vector space model, the term-count model gives better results for long documents when compared with small documents. Thus long documents have a score. The Tf-IDF method uses weight to show the importance of words in the document especially for stop-words (e.g., a, an, the etc.) filtering that is common to give weights to meaningful terms. This is because stop-words are known to have a low weight. The three approaches of vector space all perform well for long documents where the frequency term in documents is high.

Hiemstra and De Vries [21] propose the language model by the retrieval algorithms to the widely accepted traditional algorithms for information retrieval: the Boolean model, the vector space model, and the probabilistic model. The proposed algorithm is used to match terms when computed for efficient information retrieval. The language models for information retrieval are somehow similar to both the tf.idf term weighting in vector space model and relevance weighting in a probabilistic model. Thus the work proposed a strong theoretical approach to the language modeling approach by showing that the approach performs better than the weighting algorithms developed in traditional models. Therefore, the language modeling approach results in tf.idf term weighting, the tf component and the IDF component are both logarithmic thus making it a tf + IDF algorithm and not tf.idf algorithm as formerly claimed. Furthermore, they use collection frequency instead of document frequencies.

III. MOTIVATION OF RESEARCH

The main motivation of this paper is to interpretation considering the importance of large and complex unstructured data and presenting an improved indexing and ranking algorithms for Dataspace that may applicable in all field and analyzing the knowledge and information from wide data for the improvement in the business world, also enhancement in the efficiency of information extraction methods. If we want to take advantage of this huge data; the retrieval of data is simply not enough it requires a gadget for the customized layout of data, extraction of the core of information set away, and the disclosure of cases in unrefined data. With the massive measure of dataset away in records, databases, and diverse storage facilities, it is logically basic, to develop software for

 ISSN: 2454-4248
17 – 24

examination and illustration of such data and for the extraction of fascinating data that could help in decision making. The primary reaction to all above is information retrieval from Dataspace.

Baeza-Yates et. al [17] discussed some Traditional information retrieval models, they have described the problems and solution of information retrieval, they have covered the retrieval of information using keyword and phrase query based on the type of data or information. For the formulation of information retrieval models main 4-tuples [D_{set}, Q_{set}, F_{dq}, $R_f(q_i,d_j)$] used for denotation;

Where

D_{set} – Collection of Documents, Q_{set} - Collection of Query,

F_{dq} – Framework for Modeling, Representation, and relationship of Documents and Query.

$R_f(q_i,d_j)$-Ranking Function for pair q_i and d_j where $q_i \in Q_{set}$ and $d_j \in D_{set}$, here score is used to ranked the documents, the value of score can be real Number.

IV. Research Problem Statement

Whole world taking the advantages of increasing technologies in each field but it is found that the information is increasing data by day for each and every organization throughout the world so there will be problem with data storage and meaningful information from large data, which needs analysis so that knowledgeable information can be extracted and use for further development in either companies or businesses, for this reason, data mining is popular in current days which is most important technology to mine data from the large sets of data. For storing this large amount of data, computer's memory (supercomputers) and internet (Dropbox, emails etc.), that are good for the maintenance of records which is quite easy to access and flexible.

V. Research Objectives

By using data mining techniques mainly clustering, we want to propose search engine system (query-answer system) by applying indexing and ranking on heterogeneous data sets with the help of R programming in Rstudio 3.3.2 environment and Apache Lucene, Lucene-core-2.4.1 version[22].

Ranking mechanisms are one of the most necessary mechanisms of every search engine and also require high attention during the development of search engine. An efficient Indexing and Ranking algorithms have a significant role in any information retrieval system. In a web search engine, there are different need of users (Casual, Naïve, Parametric, Sophisticated, and Standalone), and the role of the search engines come first in the picture for the retrieval of information that becomes critical and considerable.

Nowadays everyone is speaking about unstructured data but actually, what does this unstructured data means? Where the unstructured data comes from, how they are analyzed, processes and also how the results are really used. More than 90% of Big data is unstructured data. In 2003, we are created almost 5 Exabyte of data that in the present time is believable in only two days. Going from 2003 to the following years, until 2012 the data expanded to almost 2.72 Zettabytes, and by 2017 they are expanded to 8 Zettabytes, which are doubled every two years. Every day size of the web pages and heterogeneity of data increasing data and whenever we fire a query using a search engine it will display thousands of results, and cannot wait for more time on visiting all pages, everyone needs results on the first page of the browsers or hardly on second page results, there should be good search engines with better indexing and ranking mechanism that can satisfy the users.

VI. A Vector Space Representation of Information

By using VSM model documents and queries can be represented as a vector, whose dimension is considered as terms to build the index represent the unstructured data. VSM is widely used for retrieving the documents and data due to its simplicity and efficiency work on a large number of datasets. VSM is based on term weighting on document vectors using three steps *1) First step is used to create indexes of the documents to retrieve the relevant data, 2) In the second step weighting of the indexed terms is used to retrieve the appropriate document for the end user, and (3) In the Finally steps the similarity measures is between documents to rank the documents relevant to the end user query using.*

The Vector Space model for three documents, three terms, and a query is shown in Fig.2. The vector is represented as follows in equations.

$$V_j = [W_{1j}, W_{2j}, W_{3j}, W_{4j}..., W_{nj}]$$

Where, W_{ij}- Weight of i^{th} keyword and j^{th} document.

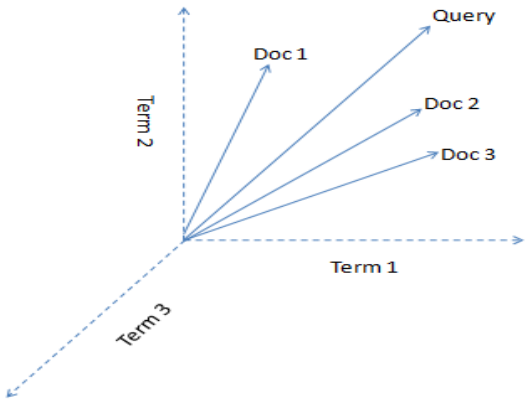

Fig.2 Vector Space Model of 3 Documents, 3 Terms, Query

 ISSN: 2454-4248
17 – 24

Calculations in the VSM used the concept of geometry with each term as a dimension, individual query and where each document is considered as points or vectors in the multidimensional space.

Coding Execution steps involved in vector space model goes as follow. Construct a matrix D_D

- D_D is Term-Document Matrix if rows of D_D - terms, columns of D_D - documents/passages
- D_D is Document-Term Matrix if rows of D_D - documents/passages, and columns of D_D - terms
- each cell - typically a frequency with which a word occurs in a doc
- also apply weighting for Indexing text: TF or TF-IDF
- finally applying Cosine similarity: for Ranking of documents

A. PREPROCESSING OF TEXT

Steps:

1) Modeling the unstructured data into a vector space is to create Metadata information as a dictionary of terms.
2) In a second step all simple terms are select from the unstructured data, and then convert these documents to a dimension in the vector space.
3) Remove stop words from unstructured data, and ignore them.
4) The prepossessing makes the text uniform.

   - Solution on Heterogeneity
   - Text cleaning- removing HTML/XML tags...etc
   - Text conversion/transformation
   - Text reduction- applying stemming
   - Discretization (numerical evaluation and implementation on digital computers.) and creating concept hierarchies.
   - The ranking algorithm, in this we have used the Page ranking Algorithm for unstructured data with Dataspace.

B. TERM WEIGHTING TECHNIQUE

Term frequency and Inverse Document frequency formed the Vector space model and tf-IDF, First find the important word from document, and second it convert the textual information into a Vector Space Model, Third if any term is frequently available in document the weight is assigned to these frequent terms, and finally **tf** is multiplied by **idf** to calculate the weight.

1) TERM FREQUENCY (TF)

Each term in vector space is representing by the **term-frequency**, that represents the frequency of the term in an unstructured document, as the documents, in reality, are of different size, so for the lager documents, the term frequency is normalized according to the size of the document. For the calculation purpose, the term frequency is divided by the total number of terms, higher the term frequency in any documents shows the importance of that document.

$$TF\ (Term) = c\ (Term, Doc)$$

C(Term, Doc) – Counting of term t in document d

The term frequency of documents evaluated as follow:

TermFreq (t, d) =

$$\frac{\text{Number of times the keyword appeared in that particular document}}{\text{Total number of keywords in the document}}$$

Inverse Document Frequency

The inverse document frequency calculates the terms available in the documents. This is calculated as follows;

$$IDF_i\ (Term, Doc) = \log_{10} \frac{N}{N_i}$$

Where, N-Number of Documents, N_i-Number of Doc. Contains term t_i

2) TF-IDF

When the calculation of weight, we find out :

- Find specified and Related word from the unstructured data
- Check the frequency of word present in the unstructured data by combining TF and IDF
- IDF to reduce weights of terms that occur more frequently to ensure that document matching is done with more discriminative words as the result, terms appearing too rarely or too frequently are ranked low

The most frequently used metric for computing term weights in a VSM. The underlying idea: value those terms that are not so common in the corpus (relatively high IDF), but still have some reasonable level of frequency (relatively high TF).

General formula for computing TF-IDF:

$$TF\text{-}IDF\ (Term, Doc) = TF\ (Term, Doc) * IDF_i\ (Term, Doc)$$

- One popular ‘instantiation’ of this formula:

 ISSN: 2454-4248
17 – 24

TF-IDF (Term, Doc) = TF (Term, Doc) * $\log_{10} \frac{N}{N_i}$

Steps for construction of Index Terms; First step is the Tokenization, in tokenization the punctuations are removed from the document and after removing punctuations from the documents the text is converted into lowercase, In Second step filtering process done to check the stop-words removed or not, and in Final step steaming process performed for the document.

Example- Find the relevant document for the following query in twelve documents stored in the local directory;

Query=" Beijing duck recipe cat food" and collection of "documents" consists of twelve (D=12).

The number of documents taken for analysis in this paper is twelve(12); these documents are different in their formats to each other so putting them together in one directory folder and called as heterogeneous database collection from which the searching of the document for our query matching.

C. Similarity Measure of the Documents using Vector Space Model

In this step first, we calculate the degree between two vectors than a degree will be used to find the similarity between query and document that will rank the document to retrieve the relevant result of the end user's query from Dataspace that is also controlled the size of the retrieved document.

Similarity between query Q and vectors for the document D

$$sim(D, Q) = D \text{ o } Q = W_{xy} \cdot W_{xz}$$

Where

W_{xy}- Weight of term x in document y

W_{xz} - Weight of term x in document z

D. Cosine Similarity Measure

Cosine similarity measure is used to find the angle between two vectors. In this case, the cosine of the angles between vectors is equal to the document vector's similarity of a query vector.

In cosine similarity, the user preference is considered as a point in n-dimensional space. Now create a line between origin point (0,0,0,….,0) two created points. Whenever two users are similar in documents, the rating of these two users will similar type of rating, means they closely related to each other in the space, they will be in the same direction from the origin point that we created. The angles will small between these two lines if they are approximately same in two documents but if points are not similar there angle will be wide in size and these points will be in different directions from the origin. By using the angle a similarity metric can be calculated for Euclidean distance, this based on the cosine of the angels for similarity value that will lies between –1 and 1, if the value of cosine angel is around 1, it will show the small angel that shows the high similarity in the document, there may be another cosine of larger angel near 180 degrees that is close to –1, that shows the less importance of the document.

All set of unstructured documents represents a set of vectors in a vector space and each term that will consider for matching in the document have their own axis. For the calculation of similarity between any two documents is given in equation 3.9. I exiting approach similarity is deal only with numbers, here we are also dealing with text, for this reason first we have to convert the text into numbers for the calculation of cosine similarity between any two documents, for the conversion of text into numbers we have used the term frequency and inverse document frequency that will be represented in vectors, in this case, users query can also be represent as vector then we will calculate the TF*IDF for the query.

The cosine similarity for query and document is calculated by the equation 3.9.

$$\text{Cos\_Sim (Q, Doc1)} = \frac{\text{Dot product (Query ,Document 1)}}{\|Query\| * \|Document\ 1\|}$$

⋮

$$\text{Cos\_Sim(Q, Doc12)} = \frac{\text{Dot product (Query ,Document 1)}}{\|Query\| * \|Document\ 12\|}$$

After calculating the cosine similarity values between query and documents the document ranked. Example for finding similarity values using entire Vector Space model is shown below.

Example: Calculate the cosine similarity between two texts:

Text 1: Hindbala loves me more than Priya loves me
Text 2: Priyanka likes me more than Hindbala loves me

STEPS:

Step-1: In Fist step we have found out all distinct words in Text 1 and Text 2.

Step-2: In a second step we have found out the frequency of occurrences of words identified in step 1 in Text 1 and Text 2, and created a vector of these words.

Step-3: In the third step, we have applied the cosine similarity function given below;

$$\cos \frac{a.b}{\sqrt{a^2}\ \ \sqrt{b^2}}$$

In a simple form

 ISSN: 2454-4248
17 – 24

$$\cos(a, b) = \frac{a \cdot b}{\|a\| * \|b\|}$$

NOTE: if a = {a1, a2…a_n} and b = {b1, b2…b_n}
Then
a.b = Sum (a1*b1 + a2*b2 + … + a_n*b_n)
||a|| = sqrt (a1^2 + a2^2 + … + a_n^2) and
||b|| = sqrt (b1^2 + b2^2 + … + b_n^2).

Example: The vector representation for the above two text;

Table 1 Vector representation

Distinct Words from both texts, #word	Frequency in Text-1 # Vector A	Frequency in Text-2 # Vector B
Hindbala	1	1
Loves	2	1
Me	2	2
More	1	1
Than	1	1
Priya	1	0
Priyanka	0	1
Likes	0	1

A_{Vector} = [a1=1, a2=2,a3=2, a4=1,a5=1,a6=1,a7=0,a8=0]
B_{Vector} = [b1=1, b2=1,b3=2, b4=1,b5=1,b6=0,b7=1,b8=1]
Evaluation using above formulas:
VectAB=9.0
VectA_Sq=12
VectB_Sq=10.0
Cosine similarity=0.82158383

In this way, we can find the similarity values between every document to the Query we are using for information retrieval from the collection of heterogeneous data.

E. STEPS OF VSM APPROACH

In order to retrieve information, the existing vector space model is designed by the following scenario:

1. Input: Group of unstructured documents and user Query(applying vector space model)
2. Preprocessing text of unstructured documents
3. Find term-document matrix for the calculation of Indexing
4. Find term frequency and inverse document frequency
5. Apply **"Cosine similarity"** between user's query and unstructured documents
6. Cluster the matching documents in one and unmatched in other
7. Ranking by sorting them according to higher to lower similarity values.

Currently, the data/information on the internet is growing extremely and it is getting uneasy for users to get the significant documents. While surfing, it might take a lot of time to discover the applicable information. Actually, it is a challenge in the area of information retrieval. Hence a fair amount of research papers are approaching in this area.

VII. CONCLUSIONS

Retrieving information from the Internet and from a large database is quite difficult and time-consuming especially if such information is unstructured. A lot of algorithms and techniques have been developed in the area of data mining and information retrieval yet retrieving data from large databases continue to be problematic. In this research work, we used the vector space model for retrieving information on the Internet. First, we computed the similarity scores using the weighted average of each item.

The cosine measure is then to compute the similarity measure and to determine the angle between documents vector and the query vector since VSMs are based on geometry whereby each term has its own dimension in a multi-dimensional space, queries and documents are points or vectors in this space. The cosine measure is often used. We then found out that it is easier to retrieve data or information based on their similarity measures and produces a better and more efficient technique or model for information retrieval. This research work is very significant in that it aims to design a tool that will enable users to retrieve information from the Internet more efficiently and effectively. Another feature that Information Retrieval Systems share with DBMS is Database Volatility. A typical large Information Retrieval application, such as a Book Library System or Commercial Document Retrieval Service, will change constantly as documents are Added, Changed, and Deleted. This constrains the kinds of data structures and algorithms that can be used for Information Retrieval.

Nowadays, managing unstructured data is main problems in the technology and in the industry; the main cause is that the techniques and tools that have proven so successful transform unstructured data into structured data but not in Heterogeneous Data. Unstructured data produced largely from email conversations, social networking sites as graphics and text. Heterogeneity and security problems with large Data hamper the progress at each stage of the process that can create huge value from data. Most of today's data not in a structured layout, as blogs and tweets are weakly structured text fragments, while images and videos are unstructured for storage and visualization, but not for semantic content and search: to transform that content into a structured format for further analysis is a great challenge.

Based on the literature reviewed we have studied for this, many works have been done by many researchers related our work. In this chapter, we have studied the search engines for information retrieval, indexing, and ranking approaches, clustering and clustering algorithms for different data, vector space model used for creating a vector for documents for information retrieval, Dataspace, and indexing of Dataspace.

Based on literature review, we have to do more work in this area for heterogeneous data on Dataspace. We have to create better indexing and ranking model for Dataspace for efficient retrieval of data.

According to above researchers work and literature review, there should be an improved framework that supports all types of queries in Dataspace for heterogeneous data, that can also create the better index and efficient and appropriate ranking method to retrieve the highest ranked result of the query from the Dataspace.

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, " Modern Information Retrieval," ACM Press.

[2] T. Y. Liu, J. Xu, T. Qin, W. Xiong and H. Li, "LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval," In Proceedings of the Learning to Rank workshop in the 30th annual International ACM SIGIR Conference (SIGIR'07) on Research and Development in Information Retrieval.

[3] J. Xu and H. Li. , "Adarank: A Boosting Algorithm for Information Retrieval," In Proceedings of the 30th Annual International ACM SIGIR (SIGIR'07) Conference on Research and Development in Information Retrieval, pp. 391–398, New York, NY, USA, 2007.

[4] P. Castells, M. Fernandez, and D. Vallet," An Adaptation of the Vector Space Model for Ontology-Based Information Retrieval," Knowledge and Data Engineering. IEEE Transactions, Vol. 19, No. 2, pp. 261-272, 2007.

[5] A. B. Manwar, H. S. Mahalle, K. D. Chinchkhede and V. Chavan, " A Vector Space Model for Information Retrieval: A MATLAB Approach," Indian Journal of Computer Science and Engineering (IJCSE), Vol. 3, No. 2, pp. 222-229, 2012.

[6] C. Zeng, Z. Lu and J. Gu, "A New Approach to Email Classification Using Computer Vector Space Model," In Proceedings of the Future Generation Communication and Networking Symposia, 2008, FGCNS'08, pp. 162-166.

[7] I. R. Silva, J. N. Souza, and K. S. Santos, " Dependence Among Terms in Vector Space Model," Proceedings of the Database Engineering and Applications Symposium, pp. 97-102, 2004.

[8] Podolecheva M, Prof T, Scholl M, Holupirek E. Principles of Dataspaces. Seminar From Databases to Dataspaces Summer Term; 2007.

[9] Niranjan Lal, Samimul Qamar, "Comparison of Ranking Algorithm with Dataspace", International Conference On Advances in Computer Engineering and Application(ICACEA), pp. 565-572, March 2015.

[10] Mrityunjay Singh and S.K. Jain: A Survey on Dataspace, National Institute of Technology, Kurukshetra-136119, India CNSA 2011, CCIS 196, pp. 608–621, 2011.

[11] Dong, X., Halevy, .: Indexing Dataspaces. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, 2007.

[12] Rocha C, Schwabe D, Aragao MP. A hybrid approach for searching in the semantic web. In Proceedings of the 13th International Conference on World Wide Web; p. 374– 83, 2004.

[13] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 2. Information Processing & Management, doi: 10.1016/S0306-4573(00) 00016-9. http://www.sciencedirect.com/science/article/B6VC8-40V4CH5-2/2/99b7ca71c12a8845a56ae46c78497f80, 36(6):809{840, November 2000

[14] Gordon, M., & Pathak, P. Finding information on the World Wide Web: the retrieval effectiveness of search engines. Information Processing and Management, 35(2), 141–180, (1999).

[15] The Economist Survey, Virtual fun. The Economist, 1–5. Available from http://www.economist.com/surveys/ display story.cfm?Story_id=2646152, (2004b, 13 May).

[16] Nielsen Netrating, "Top Web properties March 2002 [website]. Nielsen/Netrating. Retrieved, Available from http://www.nielsen-netratings.com. 27 September 2002.

[17] BELKIN, N. J., and W. B. CROFT. "Retrieval Techniques," in Williams, M. (Ed.), Annual Review of Information Science and Technology, ed. M. Williams, New York: Elsevier Science Publishers.pp. 109-45, 1987.

[18] Rolf Saint, Sebastian Schaffert, Stephanie Stroka and Roland Ferst "Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis", 4th Workshop on Semantic Wikis (SemWiki2009) at ESWC09, Heraklion, Greece, June 2009‖.

[19] M. W. Berry, Z. Drmač and E. R. Jessup, " Matrices, Vector Spaces, and Information Retrieval," Society for Industrial and Applied Mathematics, Vol. 41, No. 2, pp. 335-362, 1999.

[20] J. N. Singh and S. K. Dwivedi, "Analysis of Vector Space Model Information Retrieval. In Proceedings of the National Conference on Communication Technologies and its Impact on Next Generation computing (CTNGC'12), Int'l Journal of Computer Applications (IJCA), 2012.

[21] D. Hiemstra and A. P. De Vries, "Relating the New Language Models of Information Retrieval to the Traditional Retrieval Models," CTIT Technical Report TR-CTIT-00-00, http://www.ctit.utwente.nl, pp. 1-14, 2000.

[22] Apache Lucene Core tool , https://lucene.apache.org/core/