Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (46-56), Month: May - August 2015, Available at: www.noveltyjournals.com

Review on Capturing Context-Based Closeness between Two Text Documents

[1]Pallavi Bharambe, [2]Prof. Sanjivani Deokar

[1, 2] Department of Computer Engineering, Lokmanya Tilak college of Engineering, Navi-Mumbai, India

Abstract: **Information present on the internet is in the form of text. Even other sources of information like images, audios, videos also must be described in the form of text to convey their value to the users. So Information retrieval (IR) and Text mining applications have become very important. These applications require an proper way to represent text document in the form of mathematical objects so that it can be processed as required. Also an intelligent system which calculates closeness between two text documents is integral part of such applications. There are many ways to represent text document in mathematical form like Vector Space Model (VSM), Distance Graph, Concept Link Graph etc. VSM is most efficient and is widely used in many applications requiring text processing rather than other techniques. But VSM has major limitation that it can't maintain semantics of document. E-VSM: Enhanced-Vector Space Model is proposed to overcome original VSM's limitation of not able to maintain ordering of terms in text document and novel approach to find context-based closeness between two text documents which uses E-VSM to represent text document.**

Keywords: **Text mining, Information extraction, summarization, clustering, Text analysis.**

I. INTRODUCTION

Text Mining:

Text Mining is the located by computer of new, previously unknown information, by extracting information from different resources.

Text Mining is the process of applying automatic methods to analyze and structure textual data in order to create useable knowledge from previously unstructured information.

This paper Describes text mining with its techniques as well as its role and applications in various areas

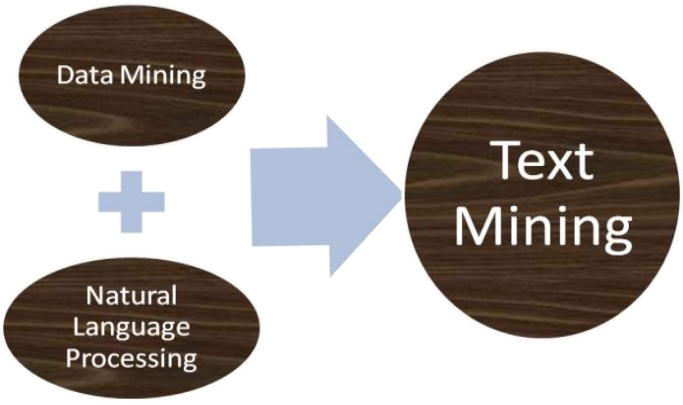


Fig 1: Text mining

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (41-48), Month: May - August 2015, Available at: www.noveltyjournals.com

A key element is the linking together of the extracted information together to form new factor new hypotheses to be explored further by more conventional means of experimentation.

Characteristics of Text:

1. Several input modes Text is intended for different consumers, i.e. different languages (human consumers) and different formats (automated consumers)

2. Dependency Words and phrases create context for each other.

3. Ambiguity Word ambiguity. Sentence ambiguity

4. Noisy data Erroneous data. Misleading data

5. Unstructured text Formal and, Informal speech

6. High dimensionality (sparse input) Tens of thousands of words (attributes).

Only a very small percentage is used in a typical document.

Purpose of text mining:

1. To discover and use knowledge that is contained in a document collection as a whole, extracting essential information from document collections and from a variety of different sources.
2. Text mining lets executives ask questions of their text-based resources quickly extract information and find answers they never imagined.
3. "Preprocessing" the text to distill the documents into a structured format.
4. Reducing the results into a more practical size.
5. Text mining the reduced data with traditional data mining techniques.
6. Text preprocessing transforms text into an information-rich, term-by-document matrix. This large grid indicates the frequency of every term within the document collection. During this stage, feature extraction is also used to locate specific bits of information, such as customer names, organizations and addresses. Next, a mathematical technique called singular value decomposition (SVD) is used to replace the original term-by-document matrix with a much smaller matrix. As part of this process, unimportant words get discarded or ignored, and more important or highly relevant words are
7. singled out. The new matrix can be used to place associated terms and documents into categories.
8. Lastly, clustering, classification and predictive methods are applied to the reduced data using traditional data mining techniques.

Text Representation:

Text document is represented by the words (features) it contains and their occurrences.

Two main approaches of document representation

1. “Bag of words”.
2. “Vector Space”.

“Bag of words” Document Representation:

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text is not represented in sequences of words. Now a days the bag-of-words model has also been used for computer vision.[1]

The bag-of-words model is mostly used in methods of document classification, where the occurrence of each word is used as a feature for training a classifier.

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (41-48), Month: May - August 2015, Available at: www.noveltyjournals.com

An early reference to "bag of words" in a linguistic context can be found in Zellig Harris's 1954 article on Distributional Structure

The following models a text document using bag-of-words.

Here are two simple text documents:

Rahul likes to watch movies. Mahesh likes too.

Rahul also likes to watch football games.

Based on these two text documents, a dictionary is constructed as:

{“Rahul”:1,“Likes”:2,“to”:3,“watch”:4,“movies”:5,“also”:6,“football”:7“games”:8,“Mahesh”:9,“too”:10}

Which has 10 distinct words. And using the indexes of the dictionary, each document is represented by a 10-entry vector:

[1, 1, 1, 1, 1, 0, 0, 0, 1, 1]

[1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

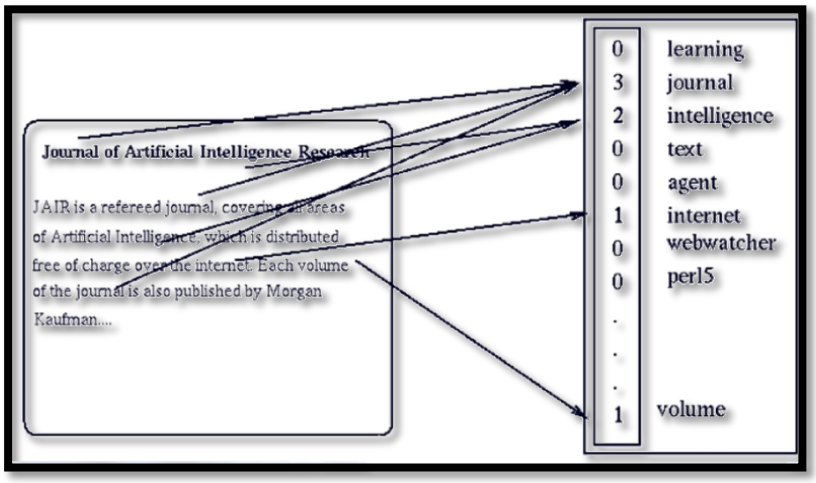


Fig.2 “Bag of words”

Approaches of Text Mining:

The following diagram shows the different approaches of text mining;

Fig.3 Approaches of Text Mining

Text Mining vs.

1. Data Mining (DM)

In Text Mining, patterns are extracted from natural language text rather than databases.

2. Web Mining

In Text Mining, the input is free unstructured text, whilst web sources are structured.

3. Information Retrieval

No genuinely new information is found.

The desired information merely coexists with other valid pieces of information.

4. Computation Linguistics (CPL) & Natural Language Processing (NLP)

CPL computes statistics over large text collections in order to find useful patterns which are used to inform algorithms for various sub-problems within NLP, e.g. Parts Of Speech tagging.

Text Mining Process:

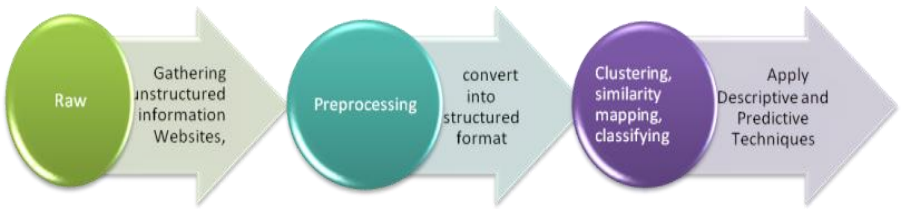


Fig.4 Text Mining Process

Information Retrieval (IR)

IR Systems identify the documents in a collection which match a user's query. The most commonly used IR systems are search engines such as Google, which identify those documents on the internet that are relevant to a set of given words. IR systems are also used in libraries, where the documents are typically not the books themselves but digital records containing information about the books.IR systems allow us to narrow down the set of documents that are relevant to a particular problem.

As text mining involves applying very computationally-intensive algorithms to large document Collections, IR can speed up the analysis n considerably by reducing the number of documents for analysis..

Natural Language Processing (NLP)

NLP one of the oldest and most difficult problems in the field of artificial intelligence It is the analysis of human language so that computers can understand natural languages as humans do. Although this goal is still some way off, NLP can perform some types of analysis with a high degree of success. Shallow parsers identify only the main grammatical elements in a sentence, such as noun phrases and verb phrases, whereas deep parsers generate a complete representation of the grammatical structure of a sentence. The role of NLP in text mining is to provide the systems in the information extraction phase with syntactical data that they need to perform their task. This is done by explanting documents with information like sentence boundaries, part-of-speech tags, parsing results, which can then be refer by the information extraction tools.

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (41-48), Month: May - August 2015, Available at: www.noveltyjournals.com

Data Mining (DM)

DM is the process of identifying patterns in large sets of data. The aim is to uncover previously unknown, useful knowledge. When use in text mining, DM is applied to the facts generated by the information extraction phase. We put the results of our DM process into another database that can be queried by the end-user via a suitable graphical interface. The data generated by such queries can also be represented visually.

Information Extraction (IE)

IE is the process of automatically extracting structured data from an unstructured natural language document. Often this involves defining the general form of the information that we are interested in as one or more templates, which are then used to guide the extraction process. IE systems rely heavily on the data generate by NLP systems.

Different methods to compare two text documents, recent happenings in clustering techniques and various 'Statistical semantics hypothesis' [1]

To calculate closeness between two text documents, there are many methods like Cosine similarity [4] Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle.

Manhattan Distance [1], The Manhattan distance function computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components. The formula for this distance between a point X=(X1, X2, etc.) and a point Y=(Y1, Y2, etc.) is:

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Where n is the number of variables, and Xi and Yi are the values of the ith variable, at points X and Y respectively.

Euclidean Distance [5] is the distance between two points (p, q) in any dimension of space and is the most common use of distance. When data is dense or continuous, this is the best proximity measure. All methods use VSM to represent text document in vector form. Since VSM itself is lossy and can't preserve semantics of text document, therefore none of the above methods can actually calculate context-based closeness.

'Statistical semantics hypothesis' [1]

Statistical patterns of human word usage can be used to figure out what people mean. If units of text have similar vectors in a text frequency matrix, then they tend to have similar meaning. This general hypothesis can be divided into several more specific hypothesis such as 'bag-of-Words hypothesis', the 'distributional hypothesis', the 'extended distributional hypothesis', and 'latent relation hypothesis'[I]. This work considers first two of above.

'Bag-of Words hypothesis' [1]: In mathematics, bag is just like a set, except that duplicates are allowed. For example, { p, q, q, q, r, r } is a bag containing elements p, q and r. In case of bags and sets, order of elements doesn't matter. Therefore bags { m, m, n, p, q }and { n, p, m, m, q}, both are equivalent. Bag { m, m, n, n, n,0} can be represented with vector x=<2,3,I> where, elements of x shows the frequency of respective elements in the bag. This is nothing but the Vector Space Model where text document is considered to be a bag of words. But in case of text documents ordering of words is important and it decides the meaning of text in document. This ordering of words is lost in VSM.

'Distributional hypothesis' [1]

States that 'Words that occur in similar context tend to have similar meaning' That means, words in proximity generally highlights single context. Individually a word can be very ambiguous but the context in which it is used decides its meaning. This gives the justification to use clustering to find out the context in text document. Various clustering methods like.

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (41-48), Month: May - August 2015, Available at: www.noveltyjournals.com

Centroid-based clustering [7]

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as a optimization problem: find the cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized. Are studied.

'k-means' [7] algorithm is widely used centroid-based clustering method. It partitions the data space in Voronoi cells and assigns all elements in the cell to its centroid.

Drawback of this method is, it requires the 'number of cluster-k' to be specified in advance so can't be used to find the context of document.

Density- based clustering [6]

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points.

Density-based clustering algorithms overcome the limitation of centroid-based clustering algorithms. Also it tries to find out only dense regions above threshold and considers sparse elements as noise. Various existing density-based clustering algorithms are DBSCAN [8], QIDBSCAN [9], U-DBSCAN [10], OPTICS [11] etc which can be used depending on application.

II. EXISTING SYSTEM

2.1.1 Vector Space Model (VSM)

Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. Its first use was in the SMART Information Retrieval System .In Vector Space Model text document is represented as vectors. When VSM is used for large number of documents, vocabulary of terms is created and appearance frequency of term in document is used as the value of respective dimension in document vector.

Dj = {fl,f2,....,ft}

Where, Dj is vector representation of document j,

fi is the frequency of the ith term in document j

This representation loses the ordering of term in which they appeared in document. It loses the context of the word; it's proximity with other words in document. It is a unidirectional representation that means vector form can be created from document but document can't be regenerated from its vector form.

3.1.2Limitations of VSM

1. In order in which the term appears in document is lose in vector space representation.

2. Long documents are poorly represented because they have poor similarity values.

3. Search keywords must precisely match document terms; word substrings might result in a "false positive match"

4. Semantic sensitivity; documents with similar context but different term vocabulary won't be associated, resulting in a "false negative match".

5. The order in which the terms appear in the document is lost in the vector space representation.

6. Theoretically assumes terms are statistically independent.

7. Weighting is intuitive but not very formal.

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (41-48), Month: May - August 2015, Available at: www.noveltyjournals.com

III. PRPOESED SYSTEM

3.1.1Enhanced Vector Space Model (E-VSM)

To overcome the limitations of VSM, E-VSM is introduced Following is the proposed enhancement to VSM:

Dj={{f1,(p1,p2,....pfl)},{f2,(p1,p2,.....pf2)},.{ft,(p1,p2,...pft)}}

Where, Dj is vector representation of document j,

fi is the frequency of the ith term in document j and

(P1…..Pfi) represents the positions at which ith term appears in document j.

Above representation of text document is lossless and bidirectional. It is lossless since ordering of terms in document is preserved and it is bidirectional since original text document can be regenerated from its vector form. This representation preserves the context of the document as well. E-VSM form of a text document can be best stored in memory as Generalized Linked List (GLL). Head node stores frequency of a term and remaining nodes in chain store the positions of same term in text document. Since text document is always read from top, positions of term will always be stored in ascending order. This representation simplifies the regeneration of text document from its E- VSM form. Normally while comparing two text documents, stop words are not considered. E-VSM considers stop word to regenerate text document more accurately. But depending on the system requirement stop words can be detected and may not be considered at the time of generating E-VSM from text document.

System Overview:

System includes all components as shown in .It takes text document to be compared and set of target text documents as input. Text document to be compared and target text document are converted into Boolean Vector and E-VSM respectively. Then proposed context-based approach is used to calculate closeness score and finally all target text documents are ranked in the descending order of closeness score. That means giving top rank to the most close target document.

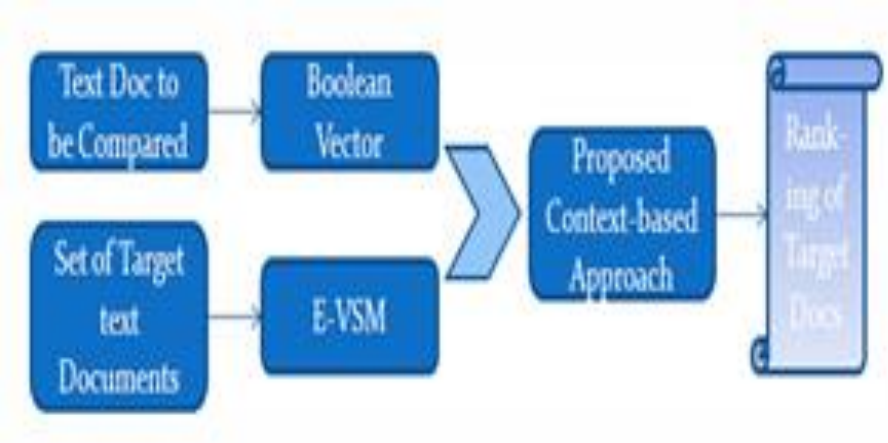


Fig. 3.1

CONTEXT-BASED CLOSENESS USING DENSITY- BASED CLUSTERING

Clustering means to 'Group (cluster) the elements having similar attributes'.

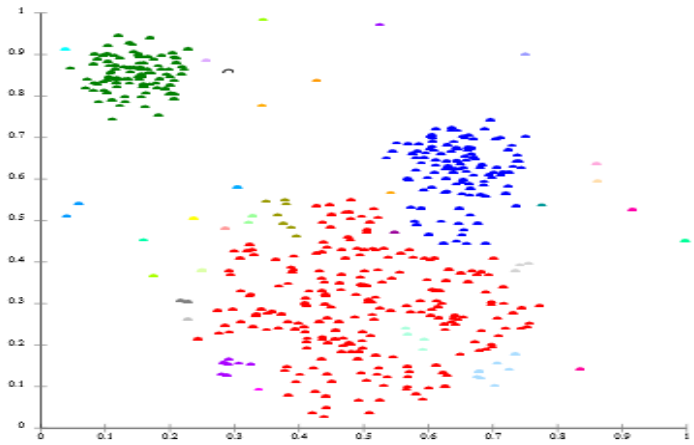
In density-based clustering clusters are defined as area of higher density than remainder of data set. Sparse elements which don't fit in any of density regions are considered as noise and are not included in any of clusters. Number of clusters are dynamically decided, depending on the nature of data set, need not know in advance like in 'k-means'(KM) clustering.

DBSCAN (density based spatial clustering of application with noise) is a popular example of density-based clustering.

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (41-48), Month: May - August 2015, Available at: www.noveltyjournals.com



(DBSCAN-DENSITY DATA)

Fig. 3.2: DENSITY-BASED CLUSTERING

DBSCAN requires two parameters: ε(eps) and the minimum number of points required to form a cluster (minPts). It starts with an arbitrary starting point that has not been visited. This point's ε -neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in sufficiently sized ε-environment of a different point and hence be made part of a cluster.

If a point is found to be a dense part of a cluster, its ε -neighborhood is also part of that cluster. Hence, all points that are found within the ε -neighborhood are added, as is their own ε -neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise

Algoritam

```
DBSCAN (SetOfPoints, Eps, MinPts)
                                SetOfPoints is UNCLASSIFIED
 ClusterId = nextId(NOISE);
             FOR i FROM 1 TO SetOfPoints.size DO
              Point = SetOfPoints.get(i);
  IF Point.ClId = UNCLASSIFIED
          THEN
    IF ExpandCluster (SetOfPoints, Point,ClusterId,  Eps, MinPts)
             THEN
      ClusterId = nextId(ClusterId)
                  END IF
               END IF
        END FOR

                           END;
```

Different techniques to find closeness Cosine similarity:

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of-1,independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (41-48), Month: May - August 2015, Available at: www.noveltyjournals.com

Note that these bounds apply for any number of dimensions, and Cosine similarity is most commonly used in high-dimensional positive spaces. For example, in Information Retrieval and text mining, each term is notionally assigned a different dimension and a document is characterized by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter. The technique is also used to measure cohesion within clusters in the field of data mining.

e.g

a: [1, 1, 0]

b: [1, 0, 1]

$$\theta = \frac{(1*1 + 1*0 + 0*1)}{\sqrt{(1-1)^2+(1-0)^2+(0-1)^2}}$$

$$\theta = \frac{1}{\sqrt{(1)^2+(-1)^2}} = \frac{1}{\sqrt{2}}$$

$\theta = 0.707$

Cos-θ= cos-(0.707)

$\theta = 45^0$

Euclidean Distance

Euclidean Distance is the distance between two points (p, q) in any dimension of space and is the most common use of distance. When data is dense or continuous, this is the best proximity measure The Euclidean distance between points p and q is the length of the line segment connecting them (pq).In Cartesian coordinates, if p = (p1, p2,..., pn) and q = (q1, q2,..., qn) are two points in Euclidean n-space, then the distance from p to q, or from q to p is given by:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

e.g

	Var 1	Var 2
P1	20	80
P2	30	44
P3	90	40

Distance between P1 & P2

$d= \sqrt{(20-30)^2+(80-44)^2} = 37.36$

Distance between P1 & P3

$d= \sqrt{(20-90)^2+(80-40)^2} = 80.62$

Distance between P2 & P3

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (41-48), Month: May - August 2015, Available at: www.noveltyjournals.com

$d= \sqrt{(30-90)^2+(44-40)^2} = 60.13$

Also Distance between P1,P2& P3

$d= \sqrt{(20-80)^2+(30-44)^2+(90-40)^2} = 79.34$

Euclidean distance is weak as compare to Cosine similarity, when multidimensional and sparse data is present.

Manhattan Distance:

The Manhattan distance function computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components. The formula for this distance between a point X=(X1, X2, etc.) and a point Y=(Y1, Y2, etc.) is:

$$d= \sum_{i=1}^{n} |x_i - y_i|$$

Where n is the number of variables, and Xi and Yi are the values of the ith variable, at points X and Y respectively.

The following figure illustrates the difference between Manhattan distance and Euclidean distance:

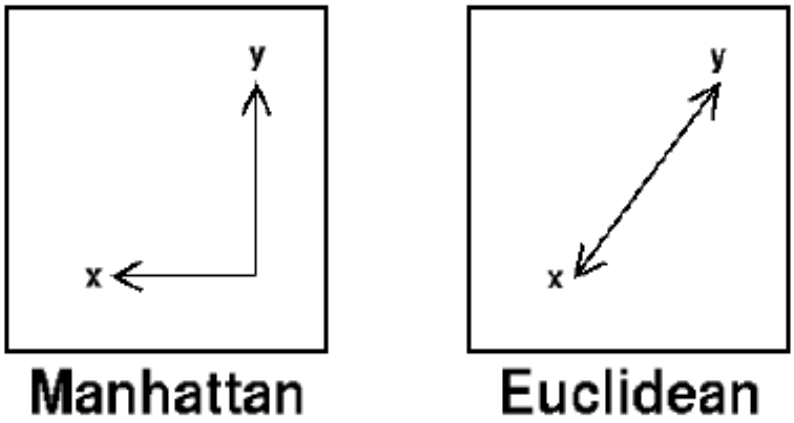


Fig 3.3 Manhattan distance and Euclidean distance

IV. CONCLUSION

Experimentation show that E-VSM gives much richer representation of text document as compared to original VSM as follows:

Stores positions of words in text document in addition with their frequencies It is a bidirectional method that means original text document can be partially regenerated from its vector representation

Doesn't lose the ordering of terms in text document in turn the context of it Maintain the semantics of text document An Enhancement to Vector Space Model (VSM) is proposed and new approach to find context-based closeness between two text documents using density-based clustering. Proposed E-VSM also includes positions of matched terms in addition to v their frequencies as in original VSM. Experimental results show that it is much richer representation of text document as compared to original VSM and has many applications. Using density-based clustering to find context-based closeness between two text documents is completely new approach. Results show that it gives very good results especially when document to be compared is very much close to particular region of target document. Currently proposed approach considers the proximity of matched terms in target text document only; it doesn't take into account the proximity of same terms in document to be compared. We would like to suggest, this consideration about the proximity of cluster terms in document to be compared may help us decide quality of that particular cluster and its contribution in calculating Closeness Score.

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (41-48), Month: May - August 2015, Available at: www.noveltyjournals.com

REFERENCES

[1] P. D Turney and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics", Journal of Artificial Intelligence Research, Volume 37,pages 141-188,2010.

[2] Charu C. Aggarwal and Peixiang for Text: "A New Paradigm for Text Representation and Processing" SIGIR'10, July 19-23 2010 Geneva Switzerland. ACM 978-1-60558-896-4/10/07.

[3] Rowena Chau, Ah Chung Tsoi, Markus Hagenbuchner and Vincent C.S Lee "A ConceptLink Graph for Text Structure Mining"' ACSC '09, Proceeding of the Thirty-Second Australasian Conference on Computer Society, Inc.Darlinghurst, Australia, 2009.

[4] P.-N. Tan, M. Steinbach & V. Kumar,"Introduction to Data Mining",Addison-Wesley (2005), ISBN 0-321-32136-7, chapter 8; page 500.

[5] http://inside.mines.edu/ckarlsso/miningortfolio/similarity. html#euclid

[6] http://en.wikipedia.orglwiki/Cluster_analysis#Density-based_clustering

[7] http://en.wikipedia.orglwiki/Cluster_analysis#Centroid-based_clusterig

[8] http://en.wikipedia.orglwikilDBSCAN

[9] Cheng-Fa Tsai, Tang-Wei Huang, "QIDBSCAN: A Quick Density-Based Clustering Technique", International Symposium on Computer, Consumer and Control (IS3C), Pages 638-641,4-6 June 2012.

[10] Tepwankul A., Maneewongwattana S., "U-DBSCAN : A density-based clustering algorithm for uncertain objects", IEEE 26th International Conference on Data Engineering Workshops(ICDEW),Pages 136-143,1-6 March 2010.

[11] http://en.wikipedia.orglwiki/OPTICSalgorithm

[12] http://upload.wikimedia.orglwikipediaicommons/0105/DBSCAN- density-data.svg