ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 10, December 2012

Comparison of Extracting Content with Minimization of Lexeme in a Text Corpus by Using Different Dimension Reduction Techniques

Ms.I.Kirubaraji[1], Ms.R.Jothilakshmi[2]

Assistant Prof, Dept of I.T, RMD Engineering college, Chennai, India[1]

Associate Prof, Dept of I.T, RMD Engineering college, Chennai ,India[2]

ABSTRACT: Document retrieval is a member of information retrieval in which information are extracted or gaining appropriate knowledge from unstructured text.i.e Unstructured text is in the form of NLP,HTML, AML format. Each document symbolized in the form of term vector model. Term vector model represented by an identifiers of objects as index terms. A single document contains more than ten thousand index terms, Seeking information from this archive is not easy. Dimension of tem vector models are high, So pertaining information from this large space is painful. Scaling of data is rigid. For the sake of effective information retrieval dimension of each document feature should be reduced. This is achieved by different dimension reduction techniques. This paper focuses on populous dimension reduction techniques such as LLE, t-SNE, Isomap and LDA and its advantages and disadvantages.

Keywords*:* documents, vectors, terms, dimension reduction

I.INTRODUCTION

Information retrieval is the exertion of gaining appropriate knowledge from a group of proper assets. It is the system of inquiring information in documents, determining for documents themselves, seeking for metadata, which describe documents or searching within documents.

Document retrieval is a branch of information retrieval, it illustrates the resembling of some form of problematic query opposed to a set of free text records. These manuscript could be any type of unstructured text.

Several textual informations are accessible electronically. For powerful recovery and tunneling, pursuing informations are more difficult without standardization of data .So each archive represented by a set of significant items. Each archive comprises ten thousand words. That can be represented in the form of vectors. Vectors are high dimensional in nature, that can be reduced by dimensionality degradation [1].

This study figure out the conventional illustration of documents, scope of dimension reduction , popular techniques of dimension reduction , together with comparison of different reduction methods.

II.ILLUSTRATIONS OF DOCUMENT

2.1 Bag of words

Let W be the dictionary – the set of all terms that occur at least once in a aggregation of document D.The bag of words achievement of document dn is a vector of weights w.

W represents (w_{1n},w_{2n} …….w_{in}).In transparent, the weights $w_{in} \in \{0,1\}$and it implies presence or absence of selective terms in a document.win represent the frequency of ith item in nth document, it comes from the term frequency representation.[2].

2.2 Terms

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 10, December 2012

Terms may represent unique word or mixed word units. The terms are represented in terms of frequency in a document (tf) and inverse document frequency (idf).

$$W_{tf}(tj, di) = tf(tj, di) * idf(tj) \qquad (2.2.1)$$

$$idf(tj) = log_2\left(\frac{N}{d_f(tj)}\right) \qquad 2.2.2$$

N=Number of documents.

df=Number of document in which term tj is occurred.

2.3 N-gram frequency

In the field of computational linguistics and probability ngram is a neighboring sequence of n-items related to be a source of grouping text. ngram model predicts x_i based on $x_{i-(n-1)}$……x_{i-1}.In probability terms this is $P_{(xi|xi-(n-1))}$.

III.DIMENSION REDUCTION

3.1.Reason

The fundamental desire of dimensionality reduction is to reduce the number of features .For the sake of too many features deriving the information from a text or document is difficult along with scaling of inputs are rigid.

3.2 Definition

Given a sample

$\{t_n\}$ C τ find :

A sample of X dimension L

A dimensionality reduction mapping F:

$$F: \tau \rightarrow \varphi$$

$$t \rightarrow x = F(t)$$

We call x the reduced dimension representative of t.

$$(dim1 \quad dim2 \ldots . dimd) \underset{\Longrightarrow}{k<D} (dim1 \quad dimk) \qquad 3.2.1$$

3.3 Classes of dimension reduction

Visualization: High dimensional data onto 2D or 3D.

Data Compression: Efficient storage and retrieval

Noise removal: Positive effective on query processing.

IV.TECHNIQUES

Dimensionality reduction can be done in two ways. Feature selection and feature extraction. Feature selection is mechanism that chooses an excellent subset of features equivalent to un preprocessed function. Feature reduction refers to the mapping of the original high dimensional data on to a lower lower dimensional space.[3].

4.1.Feature reduction Techniques

4.1.1. Locally Linear Embedding(LLE)

It is an unsupervised nonlinear learning algorithm, that estimate low dimensional adjacency embeddings of high dimensional data.It plans to observe non linear structure in high dimensional data by employing the local symmetries of linear reconstruction.[4]

LLE- Algorithm:

1.Enumerate the neighbors of each data point xi.

2.Compute the weights wij that best reconstruct each data point xi from its neighbours by minimizing the cost

$\varepsilon(w) = \sum_i \left|xi - \sum_j wij\ xj\right|^2$ 4.1.1.1$a_a$wij=j th data point to ith reconstruction.

3.Compute the vectors yi best reconstructed by the weights wij, minimizing the quadratic equation form

$$\emptyset(y) = \sum_i \left|y_{i-\sum_j wij\ y_j}\right|^2 \qquad 4.1.1.b$$

By its bottom non-zero eigenvectors.

Y is given by eigenvectors of the lowest d non –zero eigenvalues of the matrix.

$$M = (I - W)^T (I - W) \qquad 4.1.1.c$$

4.1.1.1 Advantages

- Only one parameter is needed.
- Smoother manifold is needed.
- Neighbourhood value should be larger.
- Faster Optimization.

4.1.1.2 Disadvantage:

- Local Geometry of data.

4.1.2.Isomap

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 10, December 2012

Isomap explores to conserve pairwise distance between input points.It was the first algorithm introduced for manifold learning. In mathematics, a manifold of dimension n is a topological space that near each point resembles in n- dimensional Euclidean space.[5].

Algorithm:

1.Estimate the geodeisc distances(distances along a manifold) between points in the input using shortest path distances on data set k.

2.Use Multi dimensional scaling to find points in low dimensional Euclidean space whose interpoint distances match the distances found in step 1.

Distance Matrix can be viewed as a kernel matrix.

$$k = \frac{1}{2} HD^2H \qquad 4.1.2.a$$

Where $D^2 = D_{ij}^2 = \left(D_{ij}\right)^2$ is the elementwise square of geodesic distance matrix D=[D_{ij}].H is the entering matrix given by ,

$$H = I_n - \frac{1}{N} e_M e_N^T \qquad 4.1.2.b$$

4.1.2.1.Advantages

- Nonlinear
- Globally optimal
- Guarantee asymptotically to recover the true dimensionality.

4.1.2.2 Disadvantages

- May not be stable, dependant on topology of data.
- Computation cost is more.
- If dimension N is small, it will be more inaccurate.

4.1.3 Topic Modeling:

It is a effective probablistic model for accumulating of discrete data such as text corpus.It is a three level hierarchial bayesian model,in which each item of a cumulation is modeled as afinite mixture over an primitive set ot topic probabilities.[6].

Algorithm:

1.Choose N ~ Poission($\in$).

2.Choose $\theta \sim dir(\alpha)$

3.For each of the N words wn

a) Choose a topic Z_n~Multidimensional(θ)

b)Choose Word wn from $P(W_n|Z_n, \beta)$, a multinominal probability conditioned on the topic Z_n.

4.1.3.1Advantages:

- Compute classes automatically.
- Able to assign different topics to same phase in *different contents.*

4.1.3.2 Disadvantages:

- Excercsie the network.

4.1.4 T-Distributed Stochastic Neighbour Embedding:

T-SNE visualizes high dimensional data by giving each data point a location in better visualisation.[7][8]

Algorithm:

Data: Data set $X = \{X_1, X_2, \ldots \ldots X_n\}$

Cost function parameters: perplexity Perp

Optimization Parameters: number of iteration T, Learning rate η,Momentum $\alpha(t)$.

Result: low dimensional data representation

Y(T)={y_1,y_2,……y_n}.

Advantages:

- It computes small pair wise distance with in a local structure Map

V.COMPARISION OF DIFFERENT REDUCTION TECHNIQUES

Algorithm	**Applications**	**Learning Type**
LLE	Face Recognition	UnSupervised learning
Isomap	Explore video sequences	Unsupervised Learning
LDA	Document Classification	Supervised learning algorithm
T-SNE	Document classification with 2d data	Supervised Learning

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 10, December 2012

V1.CONCLUSION

The ultimate aim of this analysis is discussing distinctive techniques of dimension reduction advantages and applications of various methods.The most favourable method for document reterival is t-SNE.Because it represents each object by a point of two dimensional scatter plot and arranges points in near by points so it computes small pairwise distances in local structure itself.

REFERENCES

[1].David Gering,"Linear and nonlinear data dimension reduction",Apr 2002.

[2].Lei Yu,Jieping Ye,"Dimensionality reduction for data mining techniques and applications and trends,"Arizona state University.

[3].L.K.Saul and S.T.Roweis,"An introduction to locally linear embedding technical report",At& T Labs research and Gatsby ,Computational neuro science unit.UCL 2001.

[4].M.Balasubramanian and E.L.Schwartz,"The Isomap learning algorithm and topological stability,Science 295:7a ,Jan2002.

[5].Yui,K.Balasubramanian,Lebanon,"Dimensionality reduction for text using domain knowledge",Georgia University.

[6].Jieping Ye,"Manifold learning MDS and Isomap",Department of Computer Science and engineering.

[7].Blei, David M,Ng.Andrew Y "Latent Drichlet Allocation".

[8].Lauren Vander Marten,"Visualizing data using t-SNE",Journal of Machine learning research(2008)pg.No:2579-2589.