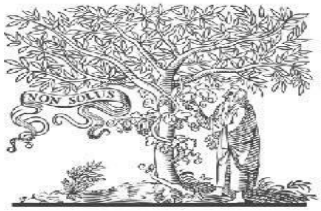
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 05th Apr 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 04

10.48047/IJIEMR/V12/ISSUE 04/12

Title **MINING THE BLOGS VIA LATENT SEMANTIC ANALYSIS**

Volume 12, ISSUE 04, Pages: 84-90

Paper Authors

D.Vamsi, Ch.Kusuma Kumari, E.Vyshnavi, Ch.Suchitra, Ch.Madan Mohan

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

MINING THE BLOGS VIA LATENT SEMANTIC ANALYSIS

D.Vamsi[1], Associate Professor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

Ch.Kusuma Kumari[2], **E.Vyshnavi[3]**, **Ch.Suchitra[4]**, **Ch.Madan Mohan[5]**
[2,3,4,5] UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
[1,2 3,4,5] d.vamsi1@gmail.com, kusumachinni0202@gmail.com,
vyshnavie567@gmail.com,
suchitra.challa17@gmail.com,madanmohanchunduri12@gmail.com

Abstract

Blogs are a popular tool for internet users to share information and express their thoughts on a variety of topics. As a result, blogs have become an increasingly important source of information for users seeking new ideas. To facilitate this exchange of information, we have created a platform for both bloggers and readers. Bloggers can regularly update their content, while readers can visit the website to access and gain knowledge from these posts. The subjective nature of blog content allows readers to gain insight into a blogger's perspectives and observations on various topics. Our website provides bloggers with separate accounts for them to log in and write or update their posts, while readers can easily navigate to the homepage to view the blogs. Our focus is on creating a Techie Blog website.

Keywords: Latent Semantic analysis, Flask, NLTK (Natural language Toolkit), Document term matrix, Singular value decomposition.

Introduction

An online diary or informational website that displays content in reverse chronological order, with the most recent posts appearing first, is known as a blog (a contraction of "weblog").

It is a website that provides information, frequently in the form of casual diary-style text entries, and serves as a forum for writers to express their opinions on specific topics. It is comparable to an online blog where a person, a team, or an organization keeps track of their activities, ideas, or views. A blog is defined as a frequently updated website or web page that is normally managed by a person or small group, written in an informal or conversational tone, and composed of a series of entries. Posts are archived and are typically arranged into categories. It is comparable to a newspaper in that it consistently produces fresh content and maintains the previous article's current.

In order to comprehend and portray the public's viewpoints in-depth, bloggers identify the sentiments, including both good and negative comments about the

subject. To read earlier postings, readers can browse these categories on the blog. In order to gain further insights, it frequently entails searching for and analysing blogs, serving as a source of information for the user's thoughts. The capacity to process vast amounts of text material effectively is provided by mining the blogs via latent semantic analysis. This study of blog mining is used to analyze and find the relevant content of online blog posts in a simpler fashion. With its ability to handle vast amounts of text data efficiently, mining blogs via lsa can be a useful tool for learning more about a particular subject. Due to its ability to handle massive amounts of text data effectively, the voluminous content is condensed, enabling the user to effortlessly sift through a flood of data in a shorter amount of time and producing faster search results.

The procedures here are:

1. Data pre-processing[1]
2. As a result, blogger updates pre-processed data
3. A summary of the modified information (involves text summarization)
4. Readers browse the content and access the essential blogs where the torrent of content is summarised, users gain knowledge about a blog in a shorter amount of time, and search results are returned quickly.

Literature survey

Literature employs a variety of text summary techniques. Text summarization Techniques are two types. One is extractive summarization and the one is abstractive summarization. Research has been done on latent semantic analysis under extractive summarization.

In paper [1] the Author OM Foong describes about the text summarization based on taking the input documents downloaded from the Document Understanding Conference 2002 dataset in 2015.

In paper [2] the Author RA Rofiq describes the text summarization based on taking the input documents as new articles and summarizing the article as the output in 2021.

In paper [3] the Author Leo Kim assesses the utility of both personal blogs and mass media coverage of future technology and forecasting the prospect of industrial technology, published in the year 2019.

In paper [4] the Authors I. A. Adigun, M.O. Adigun, and A. O. Ajayi describe text summarization with Latent Semantic Analysis, published in the year 2013.

Problem Identification

Content saturation is indeed a significant challenge that bloggers are currently facing. With the sheer volume of information available online, readers are becoming overwhelmed and finding it increasingly difficult to process all the content that they come across[2].

To address this issue, bloggers need to focus on creating high-quality, unique, and valuable content that stands out from the crowd. This can involve taking a more strategic approach to content creation, including conducting in-depth research

on topics that are relevant to their audience, and investing time and resources into creating high-quality visuals and multimedia content.

Additionally, bloggers can use technology to their advantage by implementing tools and software that help to organize and categorize their content, making it easier for readers to find what they are looking for. This might include using keywords and tags to improve searchability, as well as employing machine learning algorithms and other advanced techniques to help identify relevant content and recommend it to users.

Methodology

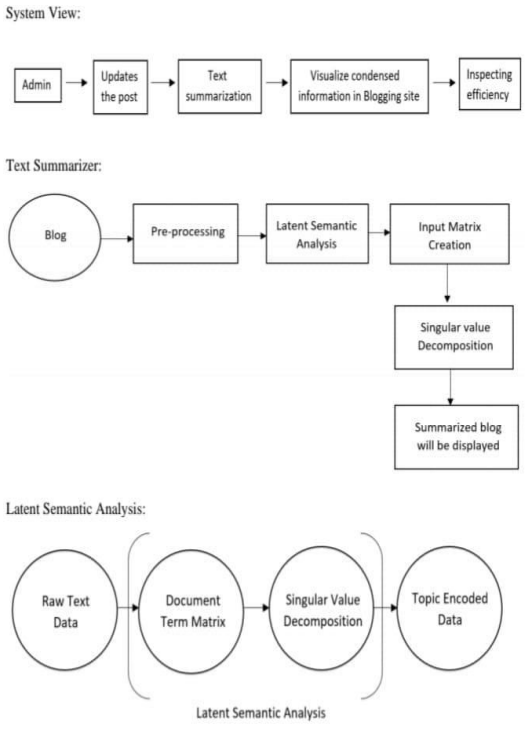


Fig.1. System Architecture

The conceptual model is what describes a system's structure, behaviour, and other aspects. A formal description and representation of a system, arranged to facilitate reasoning about the structures and behaviours of the system, is known as an architecture description. System components and created sub-systems that will cooperate to accomplish the overall system might make up a system architecture.Our system describes about the system view which contains the text summarizer contains the lsa algorithm. When an administrator updates a new blog, the modified content is briefly summarised and stored in the database. Users can get summary content here as they would information about a blog. This section of the blog is where the distilled information is displayed. The next step is to assess effectiveness by gathering information on information both before and after summary. As a result, the pre-processed data may be provided for a graphical representation with a visual impact for a before-and-after comparison of summarization. During this process, blog content is pre-processed using the Latent Semantic Analysis algorithm[3], resulting in the output of the blog's summary content. Typically, there are two steps in the latent semantic analysis process: the document term matrix[11], where the content gathered is represented in a matrix, and the singular value decompose, which allows for the extraction of the ranked text from the matrix and the extraction of the extracted data, which is in encoded matrix format, and the extraction of the extracted data as sentences.

Implementation

In this paper, the latent semantic analysis technique is used for text summarization for the content which we will present in the blogs.

Pre-processing:

Making the text acceptable and simple to handle for the phrases of the text document is known as text pre-processing[12]. To increase the effectiveness of subsequent calculations, the system employs stemming and stop-word removal processes. One of the most crucial phases in text mining is preprocessing. We need a suitable data set with appropriate attributes in order to retrieve information accurately. There are numerous phases involved in text preparation, such as eliminating stop words, erasing punctuation and URLs, stemming, and ultimately formatting the text uniformly.

Basic functions involve in pre-processing[4] are:

a)Word tokenizing: Converting the document into paragraphs or sentences into words.

b)Removal of punctuations and tags: Next step is to remove the punctuation is does not explain any meaning.

c)Removal of stop words: Removing the connectivity words like is, an, the, a, then, that, become,..etc.

Latent semantic analysis:

To represent text data in terms of features and latent features is the primary goal of latent semantic analysis [5]. A mathematical technique called latent semantic analysis (LSA) uses sample corpora of natural literature to analyze the meaning of words and passages and simulate it on a computer. Several facets of learning and interpreting human languages can be accurately approximated by LSA. It enables a range of information retrieval applications. There are two stages in the latent sematic analysis[10]:

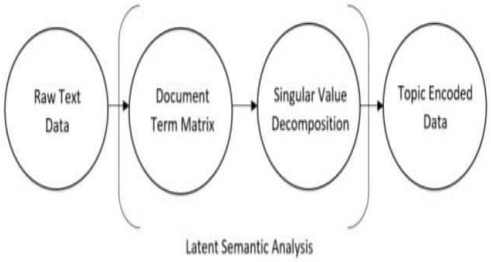


Fig.2. Latent Semantic Analysis

First stage is Document term matrix and another one is SVD.

A. Document term matrix:

The frequency of terms used in a collection of documents can be determined using a mathematical matrix called a document-term matrix [1],[2],[6]. In a document-term matrix, the rows correspond to the collection's documents, and the columns to its terms. This matrix serves as an illustration of a document-feature matrix, in which "features" refer to more than just document terms. Both computational text analysis and natural

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

language processing benefit from them.

	brown	dog	fox	lazy	quick	red	slow	the	yellow
"the quick brown fox"	1	0	1	0	1	0	0	1	0
"the slow brown dog"	1	1	0	0	0	0	1	1	0
"the quick red fox"	0	1	0	0	1	1	0	1	0
"the lazy yellow fox"	0	0	1	1	0	0	0	1	1

Fig.3.Example for Document term matrix

B.Singular value decomposition:

It resembles principal component analysis in many ways. By encoding the data set with latent characteristics, it makes it possible to provide the dimensionality of the data set. The Singular Value Decomposition (SVD) of a matrix is an operation that factors it into three distinct matrices[8]. It conveys important geometrical and theoretical understandings of linear transformations and possesses various fascinating algebraic features. It has a few important applications in data science as well.

The SVD interpretation can be obtained by determining the mapping between m-dimensional space and r-dimensional singular vector space (rank of input matrix = r). separates the original document into r linearly independent basis vectors or concepts.

By identifying salient and vectors, SVD can semantically cluster words and sentences. The phrase with the highest index value will best describe this pattern[9]. We must choose the sentences to serve as summaries after running SVD to identify the key ideas in the text.

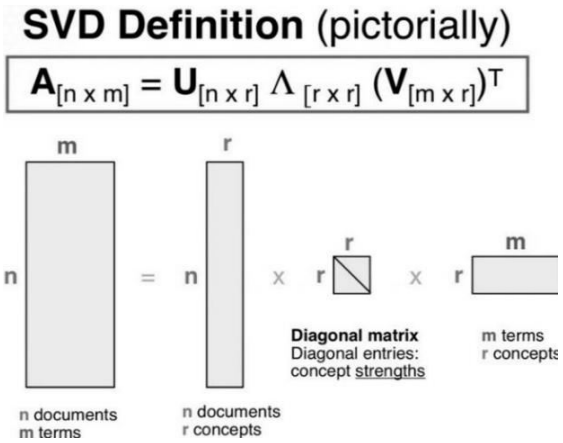


Fig.4.SVD Definition

In fig 4,Where U is an m × m orthogonal matrix. V is an n × n orthogonal matrix. Σ is an m × n matrix whose ith diagonal entry equals the ith singular value σi for i = 1,...,r. All other entries of Σ are zero.

Results & Conclusion

The study of mining the blogs is used to more quickly assess and search for pertinent blog posts online. The supply of summarised content is one of the key goals. Because the material is presented in a concise manner, one may quickly understand vast amounts of information, and surfing produces quicker results. By removing search optimization, blog mining may be understood as a view that also enables users to recognise the abundance of material on blogging websites.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

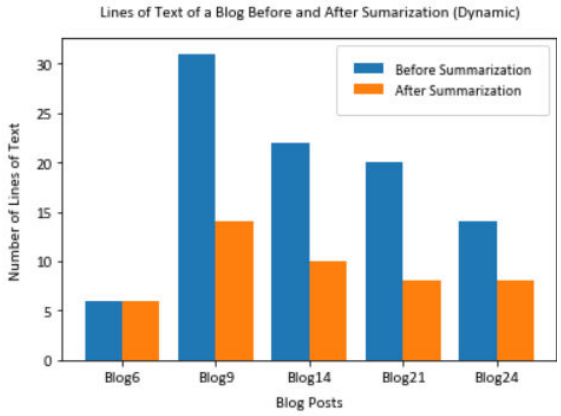


Fig.5. Lines of text in a blog

In fig 5, the graph describes how much content is reduced before and after summarization.

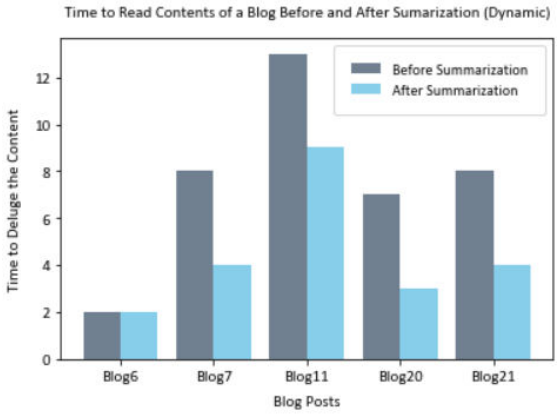


Fig.6.Time taken to read contents

In fig 6, the graph describes the how much time was taken read to the content in the blog before and after summarization.

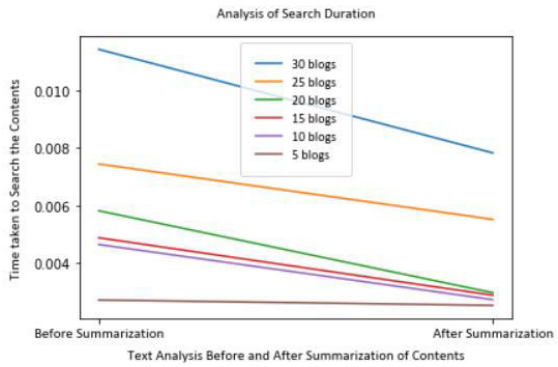


Fig.7.Analysis of Search duration

In fig 7,the graph describes how long does it take to retrieve results from a blog search before and after summarization.

Future Scope

In the future, we can use different languages, like Hindi. This enables users who prefer their native tongue to be more enriched; the fact that the source speaks their native tongue encourages many users to take an interest in reading the blogs' engaging material.

References

[1] Oi-Mean-Foong, suet peng yong ,farha am jaid "Text summarization using latent semantic analysis Model in mobile android platform.", IEEE Access, Vol. 11, Issue 3, September 2015.

[2] Rizka aniur rofiq, suyanto "Indonesian news extractive text summarization using latent semantic analysis", IEEE Access, Vol. 11, Issue 3, November 2021.

[3] Leo Kim, Jaewook Ju "Text-mining approach to the on-line newspaper and blog's representation of prospective industrial technologies", ,Science Direct ,Vol.56,Issue 4,July 2019.

[4] I.A.Adigun, M.O.Adigun, A.O.Ajayi "Text Summarization with altent semantic analysis" in International journal of computer Applications ,Science Direct,2013.

[5] Igor Mashechkin, Dmitry tsarve ,Mikhail petrovskiy "Automatic Text Summarization using Latent

Semantic Analysis", Research gate, Vol. 11, Issue 3, November 2011.

[6] Makblue Gulcin ozsoy,Ferda Nur Alpaslan,llyas cicekil "Text summarization using latent semantic analysis", Research gate, Vol. 11, Issue 3,August 2011.

[7] T.E.Ramya, N.Maghesh "Text summarization using latent semantic analysis", IJSRCSE, Vol. 8, Issue 1, feb 2020.

[8] Josef steinberger.Karel jezek "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", Research gate, Vol. 11, Issue 3, jan 2004.

[9] Song, Y., Gu, H., Zhang, L., & Wang, X. "Text summarization based on latent semantic analysis and improved ant colony algorithm".IEEE access in 2017.

[10] El-Kassas, S., & Abulkhair. "Extractive text summarization using Latent Semantic Analysis." In International Conference on Computer Science and Information Techonology in 2016.

[11] Singh, J., & Jain, "A Review of automatic text summarization techniques", IEEE Access,Vol 11,Issue 3,2016.

[12] Wang, Y., Liu, "Text summarization Based on Latent Semantic Analysis with Support Vector Regression", IEEE Access,2015.