ISSN 2616-5775 Vol. 4, Issue 5: 71-75, DOI: 10.25236/AJCIS.2021.040510

An overview of the principle, algorithm improvement and application based on the theory of latent semantic indexing

Dongyu Lan, Anqi Tian, Yingzhou Wang, Yajie Li

Beijing University of Technology, Beijing, 100124, China

Abstract: *Latent Semantic Indexing (LSI) is a latent structural model, aiming to quickly and accurately analyze a large number of texts through statistical calculation methods, and then extract the potential semantic connections between terms, while highlighting the key meanings in the text and weakening the bad influence of polysemy of the words. LSI can simplify the text vector and reduce the dimensionality, with high recall and retrieval speed. This article uses examples of spam filtering to introduce in detail the theoretical basis of latent semantic indexing, that is, singular value decomposition and the construction of multi-dimensional conceptual spaces. And the important link-weight calculation TF-IDF method uses "Sigmoid function" and "location factor" to optimize, which can further emphasize the importance of different words in the text, and is also more conducive to the construction of latent semantic structure space. Then, the paper briefly introduces two applications: research on job description clustering and construction of patent information classification system using LSI. In the end, we elaborated on the of two latent semantic indexes: retrieval and search, parallel examples: research on job description clustering and construction of patent information classification system, and gave a brief introduction.*

Keywords: *information retrieval, singular value decomposition (SVD), latent semantic index (LSI), algorithm improvement, Sigmoid function, location factor.*

1. Introduction

Nowadays, the explosion of massive information has brought huge challenges to technologies such as information retrieval.

Let us consider this scenario: a teacher receives a lot of work and homework emails from colleagues and students every week, but at the same time his mailbox also receives a lot of meaningless spam emails, which makes him miserable.

The latent semantic index can help him solve this problem.

In the traditional Raw Term Space, only the external form of the term participates in the matching, and the existence of uncertain factors such as polysemous and synonymous words in natural language affect the search results and efficiency.

For example: applying the principle of traditional character matching, if the teacher marks the common word "free" in spam emails as a keyword, all emails containing the original word "free" will be directly blocked.

But "free" does not only appear in spam that needs to be filtered out.

If the teacher's students want to discuss academic issues with him, the emails asking if they have free time may also appear "free", then such emails that have nothing to do with spam will also be blocked and filtered.

This is one of the inconveniences of traditional character matching retrieval methods: it cannot solve the problem of ambiguity accurately and efficiently.

In the traditional Raw Term Space, only the external form of the word participates in the matching, and the existence of uncertain factors such as polysemous and synonymous words in natural language will affect the search results and efficiency. For example: if a teacher marks the common word "free" in spam emails as a keyword, all emails containing the original word "free" will be directly blocked. But

ISSN 2616-5775 Vol. 4, Issue 5: 71-75, DOI: 10.25236/AJCIS.2021.040510

"free" does not only appear in spam that needs to be filtered out. If the teacher's students want to discuss academic issues with him, the emails asking if they have free time may also appear "free", then such emails that have nothing to do with spam will also be blocked and filtered. This is one of the inconveniences of traditional character matching retrieval methods: it cannot solve the problem of ambiguity accurately and efficiently. Tests find LSI very promising to such problem.

2. Related theories and operations of LSI

2.1 The basic theory of LSI

Latent semantic indexing adopts the representation method of the Vector Space Model (Proposed by Salton et al. in the 1970s), namely: construct a vocabulary-document matrix from a text set containing n documents:

$$A_{m\times n} = \{a_{ij}\}_{m\times n} \tag{1}$$

Such that each row represents a vocabulary vector, and each column represents a document vector. It successfully represents unstructured text in vector form, making various mathematical processing possible. The work done by latent semantic indexing is to apply semantic structure model to represent vocabulary and term, so as to achieve the ultimate goal of eliminating the correlation between vocabularies and simplifying text vectors.

In order to find the semantic structure in the text, singular value decomposition (SVD) of the matrix is used here. Through the operation of singular value decomposition and intercepting the first k ranks, the key information reflecting the main content of the text set in the vocabulary-document matrix $A_{m\times n} = \{a_{ij}\}_{m\times n}$ is extracted, and at the same time the less critical information is removed. Finally, the text vector is projected into the k-dimensional concept space. The more similar the content of the document, the closer the vector direction, and the closer the cosine value of the vector angle to 1, so that the document similarity can be screened based on the cosine value comparison.

2.2 The operation

2.2.1 The Steps

(1) Collect documents and preprocess

First, split the vocabulary, remove the stop words, and clean the text, and then use the TF-IDF method to convert the cleaned text into the corresponding vocabulary-document matrix.

For the example of spam classification, we denote the vocabulary-document (mail) matrix as $A = [a_{ij}]_{m\times n}$.

Such that the frequency of occurrence of term i in document j is a_{ij}, and A is a sparse high-order matrix. Because different words have different importance in a document, weights should be added to make:

$$a_{ij} = L_{(i,j)} \times G_i$$

Such that $L_{(i,j)}$ is the local weight of word i in document i, and G_i is the global weight of word G_i.

2.2.2 Singular value decomposition of the vocabulary-document matrix (SVD)

Any rectangular matrix $A = [a_{ij}]_{m\times n}$ can be decomposed into the product of three other matrix:

$$A = U_r \Sigma_r V_r^T$$

Such that $U_r \in R^{m\times r}$, $\Sigma_r \in R^{r\times r}$, $V_r \in R^{n\times r}$, $r = rank(A)$, U_r and V_r are the matrices of left and right singular vectors and Σ_r is the diagonal matrix of singular values.

ISSN 2616-5775 Vol. 4, Issue 5: 71-75, DOI: 10.25236/AJCIS.2021.040510

2.2.3 Construct k-dimensional concept space

Sort the r diagonal elements of Σ_r, that is, from large to small, and keep the first k values $(k < r)$. The larger the singular value, the more important the rows of the corresponding matrix U_k and the columns of the matrix V_k^T, and the more important the corresponding terms and documents. The result of this processing constructs a low-rank approximation matrix A_k of the original vocabulary-document (mail) matrix A, where k represents the rank after dimensionality reduction, which is much smaller than the rank of the original matrix.

The SVD form is as follows:

$$A_k = U_k \Sigma_k V_k^T \tag{2}$$

Under the 2-norm, A_k is the matrix closest to A, not only maintaining the internal connection structure (latent semantics) between the terms expressed in the term-document matrix A, but also filtering out the "noise" caused by word usage or the ambiguity of the vocabulary. Therefore, k is much smaller than the total number of words in the document m, and subtle differences in word meaning can be ignored.

2.2.4 Convert documents and terms coordinates (project words and documents into k-dimensional concept space)

Each document vector d in the mail set is sequentially converted into a document coordinate $X = d^T U_k \Sigma_k^{-1}$ in the semantic space, and each word vector t is sequentially converted into a word coordinate $Y = tV_k \Sigma_k^{-1}$ in the semantic space.

In this way, X can be compared with other document vectors for inner product or cosine similarity in the k-dimensional concept space.

2.2.5 Document similarity calculation

The similarity of document vectors X_a, X_b can be calculated by the cosine value formula

$$sim(X_a, X_b) = \frac{X_a X_b}{|X_a| \times |X_b|} = \frac{\sum_{i=1}^{t} d_{ij} \times d_{iq}}{\sqrt{\sum_{i=1}^{t} d_{ij}^2} \times \sqrt{\sum_{i=1}^{t} d_{iq}^2}} \tag{3}$$

The calculation of the similarity of pairwise vocabulary vectors is similar. Mails whose similarity to standard spam exceeds a certain threshold will be judged as spam.

2.3 Improvement of Lexical Weight Algorithm for LSI

The TF-IDF formula $a_{ij} = tf_{ij} \times idf_i$, which is widely used for calculating weights at present. Among them, a_{ij} represents the weight of the vocabulary t_i in the text d_j, tf_{ij} represents the frequency of the vocabulary t_i in the text d_j, and idf_i represents the vocabulary t_i inversely proportional to the text frequency of the vocabulary d_j. The most famous method of weighting TF-IDF is $a_{ij} = tf_{ij} \times \log\left(\frac{N}{n_i}\right)$. N represents the total number of texts in the text set. To a certain extent, it reflects the importance of the vocabulary in the text, but it is only based on linear processing, and does not consider the influencing factors such as the position of the word.

Based on this, the application of Sigmoid function and position factor is mentioned below for further improvement.

2.4 Sigmoid processing and operations considering location factors

In the traditional term-document matrix processing, a linear calculation method is used to calculate the word frequency. However, a word that appears 5 times in the text is obviously not necessarily 5 times

ISSN 2616-5775 Vol. 4, Issue 5: 71-75, DOI: 10.25236/AJCIS.2021.040510

more important than a word that appears once. The linear calculation method is obviously unreasonable.

So, Sigmoid function can be used and process the term-document matrix. Sigmoid function has the character: with the gradual increase of the abscissa, the function converges to 1. When the abscissa increases to a certain extent, the value of the function basically tends to be stable.

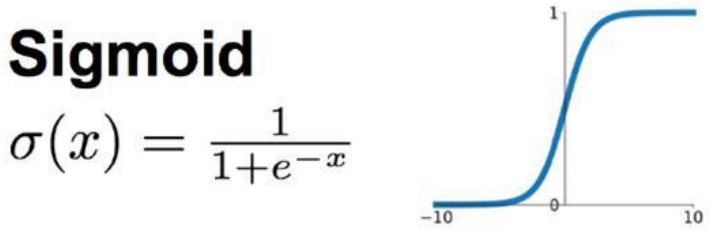

Figure 1: Sigmoid function

In the TF-IDF method, only term frequency factors and inverse term frequency factors are considered, both of which are based on frequency. However, in the text, vocabulary at different positions contributes differently to the importance of the text. Therefore, when weighting, the influence of the position of the word in the text on the word-text matrix should be considered. The position of the word in the text is also taken as a factor. The algorithm improvement of the word text matrix can improve the accuracy of LSI. [1]

2.5 Application scenes

LSI can be used for job description clustering research.

The quality of a company's recruitment of employees will directly affect the future development of the company, so recruitment of employees is very important. For a company with a larger scale and more vacant positions, this research can classify the positions and reduce the types of positions.

Each job description corresponds to a cluster label, that is, the department to which each job description belongs.

Divide all job descriptions into these categories. When a resume matches the position, you can first determine the department of the resume, and then match the resume with the job description of the department to find a suitable one. Compared with before, it can reduce the number of resume matching, improve the efficiency of resume screening, and provide convenience for personnel managers.

LSI can also be applied to the optimization of the index of patent information. The traditional subject search refers to the work of searching for patent information according to the technical subject. The result of the search is to find the relevant patent information containing the technical subject. The technical subject can be searched. The classification number corresponding to the topic and the keywords representing its technical characteristics can be used to construct a structural model for separate searches. [2][3][4]

3. Conclusion

The method LSI is able to afford richness for future extensions. Classification analyses of clustering have been used frequently for structuring documents and terms. It can also be used to solve fundamental problems of statistical NLP such as Search in Matching and Text classification in Classification as we mentioned before. More recently, big data and deep learning provides new opportunity. LSI are used Artificial Intelligence techniques.

4. Appreciation

I would like to thank Mr. Wang Jinru for her guidance on the paper, and also thank The Xinghuo Foundation for its financial support for the project research.

ISSN 2616-5775 Vol. 4, Issue 5: 71-75, DOI: 10.25236/AJCIS.2021.040510

References

[1] LiYuanYuan, Mayongqiang. Weight calculation method of text feature words based on latent Semantic index [J]. Computer application, 2008, 28(6): 1460-1466.
[2] Chenhuahui. A "spam" mail Filtering method based on latent Semantic Index [J]. Computer Application Research, 2000, 10: 17-18.
[3] HuangXinYi, ZhouWeiMin. Research on job description clustering based on latent Semantic index [J]. Network new media technology, 2017, 6(3): 33-37.
[4] BiChen, Jiduo, Caidongfeng. Research on Optimization Technology of latent semantic index based on Patent Information [J]. Journal of Shanxi University (NATURAL SCIENCE EDITION), 2014, 37(1): 26-33.*