The Influence of Document Indexing on the Bilinear Property of Vector Space Model

Peng Qu
Department of Information Management, Peking University
Beijing 100871, China
pqu@pku.edu.cn

Abstract
The paper discusses on the influence of TFIDF indexing method on inner product and bilinear function which are fundamental for Vector Space Model; and comes to the conclusion that indexing process can change the bilinear property of inner product space. It also comes to the finding that the maximum term frequency influences vector space greatly; and that the normalization factor has both retrieval and geometrical values.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms
Algorithms, Theory.

Keywords
Indexing, Vector Space Model, Inner Product

1. INTRODUCTION
Cater discusses the norm p’s influences on information retrieval space. He proves that information retrieval systems with different weights and norm values are homeomorphous in topology [1]. Researchers using topological methods have made many valuable attempts since then. Representatives of such researches come from Egghe and Rousseau [5, 6] and Dominich’s research on Information Retrieval Framework. Dominich proves that the similarity information retrieval on E is equivalent to the Hausdorff Information Retrieval Framework. Starting from the difference, we define the indexing function α to explicate indexing process formerly hidden in information retrieval models. We differentiate document d from document vector d; and bridge the two with indexing function α. We will examine the influence of indexing on VSM’s inner space in the paper.

2. INDEXING PROCESS INFLUENCE ON DOCUMENT REPRESENTATION
Given document set D, indexing term set T, documents d₁, d₂, ⋯, dₘ, in which, m = |D|. The indexing function α is defined as

\[ α(dᵢ) = dᵢ = (w₁, w₂, ⋯, wₙ)’ \]

in which, n = |T|, wᵢⱼ is the weight of indexing term tᵢ ∈ T in document dᵢ. According to standard TFIDF weighting formula, we have

\[ wᵢⱼ = tfᵢⱼ · idfᵢ = \frac{Freqᵢⱼ}{max Freqᵢⱼ} \log \frac{n}{nᵢ} \]

Freqᵢⱼ is the raw frequency of indexing term tᵢ in document dᵢ, nᵢ is the number of documents in which indexing term tᵢ appears. Define the mapping from document set D and query set Q to real number field as match function

\[ ρ: D × Q → R \]

It is thus accordant with the similar match in Vector Space Model. Meanwhile, for the present information retrieval systems, we can regard the document set D and query set Q are approximately identical. It thus satisfies the basic requirement of inner product formally. For d₁ ∈ D, q ∈ Q, we have

\[ ρ(d₁, q) = α(d₁)q = d₁q = (w₁₁, w₁₂, ⋯, w₁n)q \]

We examine the linearity on document d₁ ∈ D now. For ∀d₁, d₂ ∈ D, we have

\[ ρ(d₁ + d₂, q) = (w₁₁, w₁₂, ⋯, w₁n)q \]

Noticing that we cannot simply add the document vectors d₁, d₂ together, we take d₁ + d₂ as one document. For any member w₁₂ ∈ D in d₁ + d₂ = α(d₁ + d₂), we have
the equal sign stands only when the equation
\[ \max_1 \text{Freq}_{ij} = \max_2 \text{Freq}_{2j} = \max_{1+2} \text{Freq}_{(1+2)} \]
is satisfied.

Although the equality above needs to be met only to keep the bilinearity of VSM, it is very strict in fact. It requires that on the dimension of the most frequent terms, the two documents in consideration must be orthogonal. This means the content of the two documents must be totally different on the two dimensions. This is very strict in real situation. \( \max_1 \text{Freq}_{ij} \) is the cause of variation of vector space geometrically. From information retrieval aspect, the variation is caused by indexing and weighting process. More precisely, the loss of similarity is large for similar topics, while it is small for different topics.

Secondly, we will discuss on \( \max_1 \text{Freq}_{ij} \). It is usually regarded as normalization factor. Discussions on normalization factor have been fruitful in past researches. Sparck Jones, Walker, and Robertson deemed that normalization factor was related to document length and average document length \([8, 9]\). Singhal, Buckley, and Mitra experiment explained that there was a cross between the probability curves of relevance and retrieval \([7]\).

When using document length to normalize, the linear combination of previous normalization method and pivot. The paper explicates the normalization factor’s effect on the vector space. If we do not take the normalization factors into consideration, i.e., taking raw frequencies as TF part in term indexing, indexing will not change the linear property of VSM’s inner product space. This means the normalization factors cause the variation of linear characters of vector spaces. The influence of different indexing method is heterogeneous and needs discussions separately.

4. CONCLUSION AND FUTURE WORK

The discussion in the paper explicates that vector space model is a formal framework and that it is affected greatly by indexing method. Strictly speaking from algebraic deduction, it will not fulfill the requirements for bilinear functions under TFIDF indexing scheme selected in the paper. The main factor that accounts is normalization factor. In TFIDF indexing scheme, the maximum frequency of one document is of great meanings geometrically and for information retrieval.

“Bag of words” is the assumption of the paper, which guarantees the expansion of \( \alpha(d_1 + d_2) \). With the development of n-gram indexing, the “bag of words” assumption might be replaced. While n-gram indexing hitherto is not good enough, the “bag of words” assumption is still widely accepted in the present IR research. The paper thus contributes to Vector Space Model under “bag of words” assumption only.

These results are still very elementary. We have discussed on the basic situation of Vector Space Model. Related issues such as clustering and relevance feedback are not covered in the paper, which leaves many problems to be resolved. Discussion on maximum term frequency and normalization factors need further study on its role in information retrieval and geometry.

5. ACKNOWLEDGMENTS

The author is partially supported by China Scholarship Council (No.: 2009601175) and Graduate School of Peking University.

6. REFERENCES


