An Insight Into Vector Space Modeling and Language Modeling

Kun Lu
School of Information Management
Wuhan University
kunlu_whu@126.com

Abstract

Vector Space Modeling (VSM) and Language Modeling (LM) are the two most influential retrieval models currently. They appear to have different perspectives and use different mathematical tools. However, they are actually closely related. The current study analyzed their relationship, compared their weighting schemes and revealed their connections. Our findings suggest that although the VSM and the LM originated from different perspectives, they are closely related. The backbone of the LM weighting is still a TF-IDF like weighting scheme.

Keywords: vector space modeling, language modeling, relationships

Introduction

Vector Space Modeling (VSM) and Language Modeling (LM) are the most popular information retrieval models currently. They provide different ways to represent documents and queries, as well as different means to evaluate documents against queries. The VSM was first proposed in Salton and McGill (1983). It uses linear algebra tools to model the documents and terms. A document is represented as a vector and the terms are its elements. On the other hand, the LM was first brought to information retrieval by Ponte and Croft (1998). It is a branch of probabilistic models. A document is viewed as a language model, which is essentially a probability distribution over its terms. At the first glance, the two models take quite different perspectives and use very different mathematical tools. However, it has been speculated that the two models are related. A search on the literature suggests few studies have thoroughly examined the relationship between the two models. Zhai and Lafferty (2001) briefly discussed the connections when studying the smoothing in the LM. Robertson (2004) also pointed out that the weighting in the LM achieves a similar effect as the classical TF-IDF weighting in the VSM by somewhat different means. However, no deep analysis and further discussion was provided with respect to how they are related. The purpose of this study is to uncover the connection between the two models and provide an in-depth insight into their relationship. A good understanding of their relationship will help us to better interpret our search results.

Brief Description of the Two Models

Vector Space Modeling

The VSM uses vectors to represent documents and the elements of a vector consist of words appearing in the collection. The mathematical representation is given as follows:

$$V_{nm} = \begin{pmatrix} v_{11} & & \cdots & v_{1m} \\ v_{21} & & \cdots & v_{2m} \\ \cdots & \cdots & v_{ij} & \cdots \\ v_{n1} & & \cdots & v_{nm} \end{pmatrix} \qquad (1)$$

Lu, K. (2013), An Insight Into Vector Space Modeling and Language Modeling. *iConference 2013 Proceedings* (pp. 717-721). doi:10.9776/13332

The rows of the matrix are defined as documents in the vector space while the columns of the matrix are defined as the terms which are used to describe or index the documents in the vector space. This matrix is commonly referred to as the document-term matrix. An element v_{ij} *(1 ≤ i ≤ n, 1 ≤ j ≤ m)* in the document-term matrix reflects the normalized weight of the indexing term t_j assigned to the document d_i. Here *n* and *m* are the number of documents and indexing terms in the vector space respectively. The prominent TF-IDF defines the term weight to be proportional to the term frequency in the document and inversely proportional to the number of documents that contain the term (i.e. document frequency):

$$v_{ij} = tf_{ij} * log\frac{|C|}{n(t_j)} \qquad (2)$$

where tf_{ij} is the term frequency weighting of the *j*th term in *i*th document, $n(t_j)$ denotes the document frequency , and |*C*| denotes the number of documents in the collection. To control the effect of the document length, the document length normalization is usually applied to the TF component (Harman & Voorhees, 2006).

Language Modeling

Language modeling was first used in natural language processing to model the probability of a sequence of words. Ponte and Croft (1998) introduced LM to information retrieval by considering retrieval as a generative process. The LM ranks the documents according to their probabilities to generate the query terms. To estimate the probability of seeing a term in a document language model, the maximum likelihood estimation is usually adopted:

$$\hat{P}(t|D) = \frac{f(t,D)}{|D|} \qquad (3)$$

where *f(t ,D)* is the frequency of a term in the document, and |*D*| is the document length. Because a document language model is estimated from a limited sample (i.e. one document), it is likely to have a data sparseness problem. So, smoothing becomes apparent for the LM. A commonly used smoothing method is the Jelinek-Mercer smoothing. It mixes the probability estimated from the document with the one from the collection:

$$\hat{P}_{smooth}(t|D) = \lambda\, \hat{P}(t|D) + (1-\lambda)\hat{P}(t|C) \qquad (4)$$

where the maximum likelihood estimation of the $\hat{P}(t|C)$ is the collection frequency of the term divided by the total term count in the collection.

Model Comparison

Both models naively assume the independence of terms in documents. On the surface, the two models provide similar functions (represent documents and weight terms) from quite different perspectives. However, a detailed investigation uncovers their connections. First, a vector in the VSM and a probability distribution in the LM are similar in containing the term weights although they have very different mathematical intuitions (i.e. geometry vs. probability). The only difference is that a probability distribution is normalized to sum to one while a vector does not have such requirement. In terms of their term weighting schemes, The TF-IDF is actually closely related with the probability weighting method in the LM. The standard LM weighting can be decomposed into the following equation (Zhai & Lafferty, 2001):

$$logp(Q|D) = \sum_{i:c(q_i,D)>0} log\frac{p(q_i|D)}{(1-\lambda)\,p(q_i|C)} + m\log(1-\lambda) + \sum_{i=1}^{m} \log p(q_i|C) \qquad (5)$$

where $p(q_i|D)$ denotes the probability of a query term in a document, $p(q_i|C)$ is the probability of a query term in the collection, λ is a parameter to control the amount of smoothing. In equation 5, the only component that influences the rankings is the first addend, $\sum_{i:c(q_i;D)>0} log \frac{p(q_i|D)}{(1-\lambda)\, p(q_i|C)}$, which indicates that the LM weighting is actually proportional to the term frequency in the document and inversely proportional to the collection frequency. This is very close to what is described in the TF-IDF except that the collection frequency is used in the LM instead of the document frequency. Therefore, the relationship between the VSM and the LM is rooted in the relationship between the collection frequency and the document frequency. However, in terms of how the collection frequency is related with the document frequency, the authors did not provide any further evidence. To investigate this relationship, we selected a number of different document collections from TREC, and examined the collection frequencies and the document frequencies of the terms. The following section will report our results.

Collection Frequency versus Document Frequency

A number of representative document collections from TREC were selected. This includes the collections for the Genomics Track 2006, the WT10G, the TREC-6, and the Medline collection. The Genomics Track is a collection of full text academic articles in the field of biology linked with genomics information. The WT10G collects a large number of English web pages. The TREC-6 contains newspaper and government records. The Medline collection is a small collection of documents from Medline plus database. For each collection, TREC provides a number of test topics and their relevance judgments. The data collections were indexed by the Indri search engine (www.lemurproject.org). Stop words were removed and stemming was applied. The descriptive statistics of the document collections are listed in Table 1.

Table 1
Descriptive Statistics of the Document Collections

Corpus	**Corpus size**	**# of queries**	**Avg. doc length**	**# of unique tokens**
Genomics	162,259	64	6,595	2,075,859
WT10G	1,692,096	100	617	5,256,472
TREC-6	556,077	50	526	767,503
Medline	1,033	30	155	9,537

To understand the connection between the LM weighting and the TF-IDF, we examined the correlations between the collection frequencies and the document frequencies of the query terms. The query terms are obtained from the title field of the TREC retrieval topics. Those query terms that do not appear in the collection are dropped as they are not affecting the retrieval. Spearman's ρ is reported since the data could be highly skewed. The results are provided in Table 2.

Table 2
Correlations between the Collection Frequencies and the Document Frequencies of the Query Terms in the Collections

	Genomics	**WT10G**	**TREC-6**	**Medline**
Spearman's ρ	0.966**	0.986**	0.989**	0.952**

**indicates significant correlation at 0.01 level (2-tailed)

From Table 2, we can tell that the collection frequencies of the query terms are both strongly and significantly correlated with their document frequencies in all the collections. This indicates the two measurements are highly accordant in weighting the query terms.

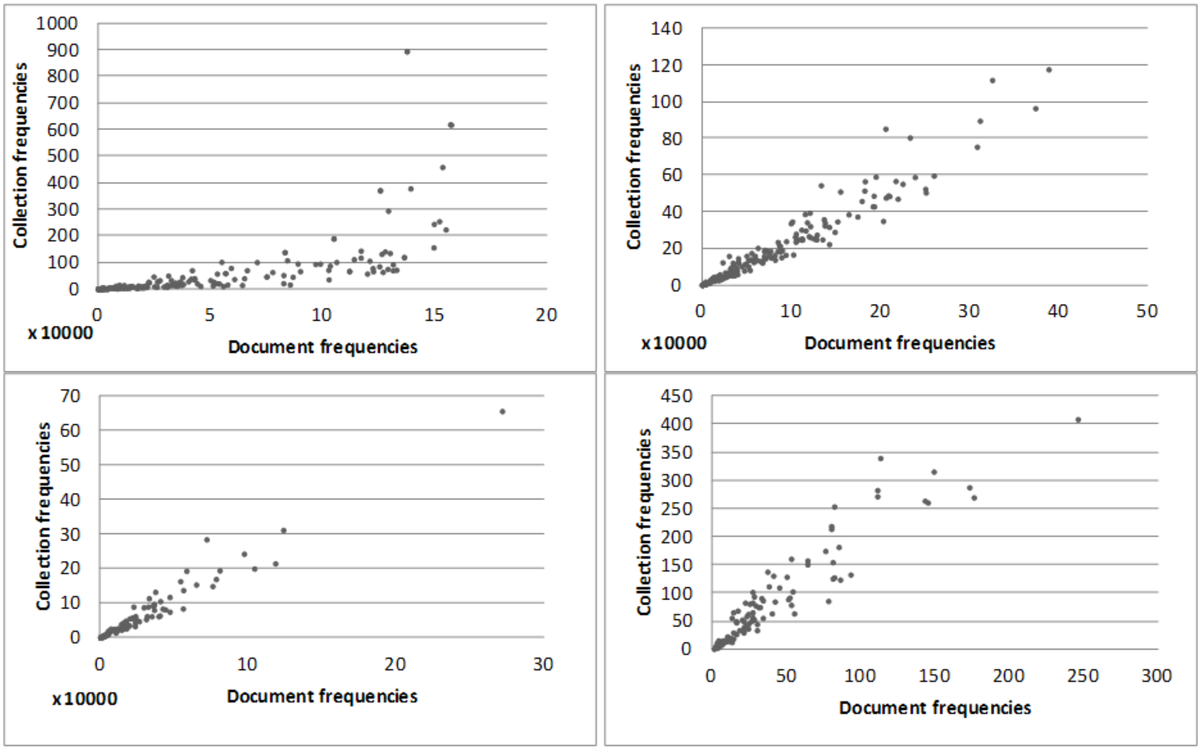

Figure 1. Scatter plots of the term collection frequencies and document frequencies in four collections (upper left is Genomics, upper right is WT10G, lower left is TREC-6, and lower right is Medline)

Figure 1 gives the scatter plots of the document frequencies and the collection frequencies in the four collections. All the charts showed the strong positive correlations between the two measurements. Given that the TF-component is exactly the same as the term probability in the document, we can conclude the TF-IDF weighting and the LM weighting are comparable. However, it should be noted that the two weighting are not exactly the same. First, there are some differences in their calculations. For example, TF-IDF applies logarithm on the IDF part before multiplying the TF part, while according to equation 5 the LM does the multiplying first and then apply the logarithm. Second, the strong and significant correlations do not imply the sameness. Some terms can have similar collection frequencies but different document frequencies or vice versa.

Conclusion

The VSM and the LM are the two most influential retrieval models currently. They appear to have different perspectives and use different mathematical tools. However, they are actually closely related. The current study analyzed their relationship, compared their weighting schemes and revealed their connections. The TF component (with document length normalization) in the TF-IDF weighting is exactly same as the probability of seeing a term in a document language model. The IDF component is implicitly related to the smoothing methods in the LM. After decomposing the LM weighting, we have found that the difference is that the LM uses the collection frequency instead of the document frequency in the VSM. An examination on the relationship between the collection frequencies and the document frequencies of the terms in several representative TREC collections indicates that they have both strong and significant correlations. Therefore, we conclude that although the VSM and the LM originated from different perspectives, they are closely related. The backbone of the LM weighting is still a TF-IDF like weighting scheme.

References

Harman, D. K., & Voorhees, E. M. (2006). TREC: An overview. *Annual Review of Information Science and Technology, 40*(1), 113-155.

Ponte, J., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st ACM SIGIR Annual International Conference on Research and Development in Information Retrieval* (pp. 275-281). Melbourne, Australia.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation, 60*(5), 503-520.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 334-342). New Orleans, Louisiana, USA.