Distance Weighted Cosine Similarity Measure for Text Classification

Baoli Li and Liping Han

Department of Computer Science
Henan University of Technology
1 Lotus Street, High & New Industrial Development Zone
Zhengzhou, Henan 450001, China
csblli@gmail.com

Abstract. In Vector Space Model, Cosine is widely used to measure the similarity between two vectors. Its calculation is very efficient, especially for sparse vectors, as only the non-zero dimensions need to be considered. As a fundamental component, cosine similarity has been applied in solving different text mining problems, such as text classification, text summarization, information retrieval, question answering, and so on. Although it is popular, the cosine similarity does have some problems. Starting with a few synthetic samples, we demonstrate some problems of cosine similarity: it is overly biased by features of higher values and does not care much about how many features two vectors share. A distance weighted cosine similarity metric is thus proposed. Extensive experiments on text classification exhibit the effectiveness of the proposed metric.

1 Introduction

Similarity calculation is a basic component for many text mining applications. For example, if we have a perfect method to assess how two text segments are similar, we could build an ideal information retrieval system. In the past years, a lot of metrics [1,2], such as Euclidean distance based metric, Cosine, Jaccard, Dice, Jensen-Shannon Divergence based metric, have been proposed to deal with different kinds of information retrieval and natural language processing problems. Among the existing metrics, Cosine, which measures the angle between two vectors, is the most popular one. It is effectively calculated as dot-product of two normalized vectors.

Given two N dimension vectors $\vec{v}$ and $\vec{w}$, the cosine similarity between them is calculated as follows:

$$\text{Cosine}(\vec{v},\vec{w}) = \frac{\vec{v} \bullet \vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

In mathematics perspective, Cosine similarity is perfect. However, if we check it in text mining perspective, it may not always be reasonable. Let's consider a few example vectors shown in figure 1. Suppose these 3-D vectors are derived from five text segments

H. Yin et al. (Eds.): IDEAL 2013, LNCS 8206, pp. 611–618, 2013.

A, B, C, D, and E. The cosine similarities between segment A and the rest are given in the figure. From the values, we can conclude that segment B is the most similar one of A, as it has the highest cosine. However, is it reasonable? Intuitively, text segments C and E, which both have two common terms with segment A, are more relevant to A than B, which contains only one term. Moreover, the segment E has one more term than segment A, but Cosine(A,E) is much lower than Cosine(A,B). If we regard the additional term as a noise and neglect it, E will have the same vector as A.

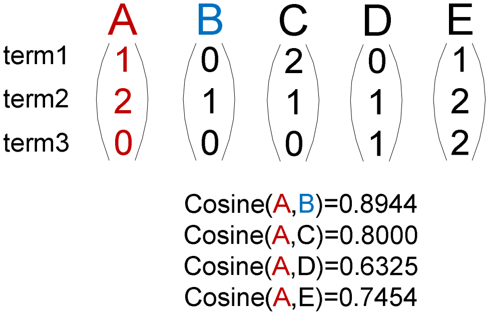

Fig. 1. Cosine similarities between five synthetic vectors

It can thus be derived from the above figure that cosine similarity tends to be overly biased by the features of higher values, but it doesn't care much about how many features two vectors share. In text mining perspective, more features two text segments share, more similar they are. If a part of a text segment is much similar to another segment as whole, the former one is usually thought to be relevant to the latter in Information Retrieval. It is this observation that motivates us to explore more effective similarity metrics than Cosine for text mining.

Because of the proven effectiveness of cosine similarity, we decide to derive new metrics by slightly modifying it. Several distance weighted versions are explored, where distance tends to capture how many features two text segments share. With extensive experiments on a classical text mining problem, i.e. text classification, we obtain a distance weighted cosine metric that performs better than the original cosine metric in most cases. It is also demonstrated with experiments that the similarity metric does have important effects in text mining applications.

The rest of this paper is organized as follows: section 2 introduces the explored distance weighed cosine metrics; section 3 presents extensive experiments on three text classification problems and discussion on the results; Section 4 concludes the paper.

2 Distance Weighted Cosine Similarity Measures

We explore different new similarity metrics with the following evidences or assumptions: 1. cosine similarity is good enough for most text mining applications; 2. more features two text segments share, more similar they are. Therefore, all of the designed metrics have two key components: cosine similarity and distance measure, but they are different in applying different distance measures and assembling strategies.