新闻的特征抽取方法

October 17, 2013

1 整篇新闻的特征

1.1 TF * IDF

将文章里的词都提取出来，计算每个新闻的tf*idf 来得到一个向量

2 句子的特征

2.1 Centroid-based summarization of multiple documents

[1]

Centroid value

The centroid value Ci for sentence Si is computed as the sum of the centroid values Cw;i of all words in the sentence. For example, the sentence “President Clinton met with Vernon Jordan in January” would get a score of 243.34 which is the sum of the individual centroid values of the words (clinton = 36.39; vernon = 47.54; jordan = 75.81; january = 83.60).

$$C_i = \sum_{w} C_{w,i} \tag{1}$$

Positional value

The positional value is computed as follows: the first sentence in a document gets the same score Cmax as the highest-ranking sentence in the document according to the centroid value. The score for all sentences within a document is computed according to the following formula:

$$P_i = \frac{n - i + 1}{n} * C_{max} \tag{2}$$

First-sentence overlap

The overlap value is computed as the inner product of the sentence vectors for the current sentence i and the first sentence of the document. The sentence vectors are the n-dimensional representations of the words in each sentence, whereby the value at position i of a sentence vector indicates the number of occurrences of that word in the sentence.

$$F_i = \overrightarrow{S_1}\overrightarrow{S_i} \tag{3}$$

References

[1] D. R. Radev, H. Jing, M. Styś, and D. Tam, “Centroid-based summarization of multiple documents,” *Information Processing & Management*, vol. 40, no. 6, pp. 919–938, 2004.