Rocchio algorithm

From Wikipedia, the free encyclopedia

The **Rocchio algorithm** is based on a method of relevance feedback found in information retrieval systems which stemmed from the SMART Information Retrieval System around the year 1970. Like many other retrieval systems, the Rocchio feedback approach was developed using the Vector Space Model. The algorithm is based on the assumption that most users have a general conception of which documents should be denoted as relevant or non-relevant.[1] Therefore, the user's search query is revised to include an arbitrary percentage of relevant and non-relevant documents as a means of increasing the search engine's recall, and possibly the precision as well. The number of relevant and non-relevant documents allowed to enter a query is dictated by the weights of the a, b, c variables listed below in the Algorithm section.[1]

Contents

Algorithm

The formula and variable definitions for Rocchio relevance feedback is as follows:[1]

$$\overrightarrow{Q_m} = \left(a \cdot \overrightarrow{Q_o}\right) + \left(b \cdot \frac{1}{|D_r|} \cdot \sum_{\overrightarrow{D_j} \in D_r} \overrightarrow{D_j}\right) - \left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\overrightarrow{D_k} \in D_{nr}} \overrightarrow{D_k}\right)$$

As demonstrated in the Rocchio formula, the associated weights (**a**, **b**, **c**) are responsible for shaping the modified vector in a direction closer, or farther away, from the original query, related documents, and non-related documents. In particular, the values for **b** and **c** should be incremented or decremented proportionally to the set of documents classified by the user. If the user decides that the modified query should not contain terms from either the original query, related documents, or non-related documents, then the corresponding weight (**a**, **b**, **c**) value for the category should be set to 0.

In the later part of the algorithm, the variables **Dr**, and **Dnr** are presented to be sets of vectors containing the coordinates of related documents and non-related documents. Though **Dr** and **Dnr** are not vectors themselves, $\overrightarrow{Dj}$ and $\overrightarrow{Dk}$ are the vectors used to iterate through the two sets and form vector summations. These sums are normalized (divided) by the size of their respective document set (**Dr**, **Dnr**).

In order to visualize the changes taking place on the modified vector, please refer to the image below.[1] As the weights are increased or decreased for a particular category of documents, the coordinates for the modified vector begin to move either closer, or farther away, from the centroid of the document collection. Thus if the weight is increased for related documents, then the modified vectors coordinates will reflect being closer to the centroid of related documents.

Variable	Value
$\overrightarrow{Q_m}$	Modified Query Vector
$\overrightarrow{Q_o}$	Original Query Vector
$\overrightarrow{D_j}$	Related Document Vector
$\overrightarrow{D_k}$	Non-Related Document Vector
a	Original Query Weight
b	Related Documents Weight
c	Non-Related Documents Weight
D_r	Set of Related Documents
D_{nr}	Set of Non-Related Documents

Time complexity

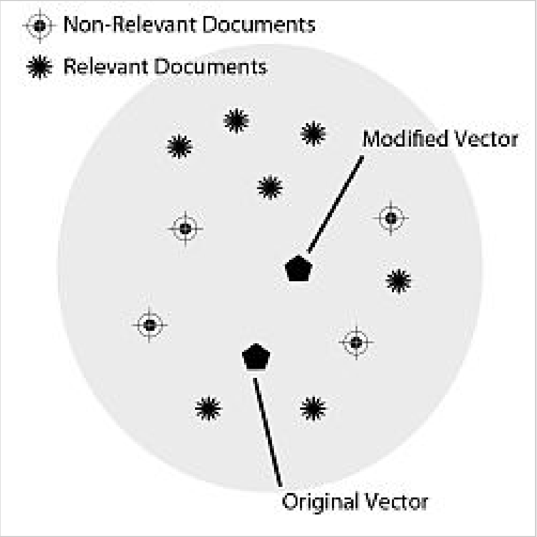
Variable	Value
$\mathbb{D}$	Labeled Document Set
L_{ave}	Average Tokens Per Document
$\mathbb{C}$	Class Set
V	Vocabulary/Term Set
L_a	Number of Tokens in Document
M_a	Number of Types in Document

The time complexity for training and testing the algorithm are listed below and followed by the definition of each variable. Note that when in testing phase, the time complexity can be reduced to that of calculating the euclidean distance between a class centroid and the respective document. As shown by: $\Theta(|\mathbb{C}|M_a)$.

Training = $\Theta(|\mathbb{D}|L_{ave} + |\mathbb{C}||V|)$
Testing = $\Theta(L_a + |\mathbb{C}|M_a) = \Theta(|\mathbb{C}|M_a)$[1]

Usage

Though there are benefits to ranking documents as not-relevant, a relevant document ranking will result in more precise documents being made available to the user. Therefore, traditional values for the algorithm's weights (**a**, **b**, **c**) in Rocchio Classification are typically around **a = 1**, **b = 0.8**, and **c = 0.1**. Modern information retrieval systems have moved towards eliminating the non-related documents by setting **c = 0** and thus only accounting for related documents. Although not all retrieval systems have eliminated the need for non-related documents, most have limited the effects on modified query by only accounting for strongest non-related documents in the **Dnr** set.



Rocchio Classification

Limitations

The Rocchio algorithm often fails to classify multimodal classes and relationships. For instance, the country of Burma was renamed to Myanmar in 1989. Therefore the two queries of "Burma" and "Myanmar" will appear much farther apart in the vector space model, though they both contain similar origins.[1]

See also

- Nearest centroid classifier, aka Rocchio classifier

References

1. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: *An Introduction to Information Retrieval*, page 181. Cambridge University Press, 2009.

- Relevance Feedback and Query Expansion (http://nlp.stanford.edu/IR-book/pdf/09expand.pdf)
- Vector Space Classification (http://nlp.stanford.edu/IR-book/pdf/14vcat.pdf)
- Data Classification (http://cs.nyu.edu/courses/fall07/G22.2580-001/lec7.html)

Retrieved from "http://en.wikipedia.org/w/index.php?title=Rocchio_algorithm&oldid=653512526"

Categories: Information retrieval