MMR: Maximal Marginal Relevance

Devrim Şahin & Selçuk Emre Solmaz

$$MMR = \arg\max_{D_i \in R \setminus S} \left[\lambda\, Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right]$$

D_i: Documents in the collection C
Q: Query,
R: Relevant documents in C,
S: Current result set

- Constructs the result set incrementally
- User-tunable diversity through λ parameter
- High λ = Higher **accuracy**
- Low λ = Higher **diversity**

EXAMPLE:

Assume that we are given a database of 5 documents d_i and a query q, and we calculated, given a symmetrical similarity measure, the similarity values as below. Further assume that λ is given by the user to be 0.5:

$$S = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ q \end{array} \begin{array}{c} \begin{array}{cccccc} d_1 & d_2 & d_3 & d_4 & d_5 & q \end{array} \\ \left[\begin{array}{cccccc} 1 & 0.11 & 0.23 & 0.76 & 0.25 & 0.91 \\ & 1 & 0.29 & 0.57 & 0.51 & 0.90 \\ & & 1 & 0.02 & 0.20 & 0.50 \\ & & & 1 & 0.33 & 0.06 \\ & & & & 1 & 0.63 \\ & & & & & 1 \end{array}\right] \end{array}$$

1^ST ITERATION

Currently our result set S is empty. Therefore the second half of the equation, which is the max pairwise similarity within S, will be zero. For the first iteration, MMR equation reduces to:

MMR = arg max (Sim (d_i, q))

d_1 has the maximum similarity with q, therefore we pick it and add it to S. Now, S = {d_1}.

2^ND ITERATION

Since S = {d_1}, finding the maximum distance to an element in S to a given d_i is simply sim(d_1,d_i).

For d_2:

sim(d_1, d_2) = 0.11
sim (d_2, q) = 0.90
Then MMR = λ0.90 – (1-λ)0.11 = 0.395

Similarly MMR values for $d_{3,\,4,\,5}$ are 0.135, -0.35 and 0.19 respectively. Since d_2 has the maximum MMR, we add it to S. Now S = {d_1, d_2}.

3^RD ITERATION

This time S = {d_1, d_2}. We should find max. of sim (d_i, d_1) and sim (d_i, d_2) for the second part of the equation.

For d_3:

max{sim (d_1, d_3), sim (d_2, d_3)} =
max {0.23, 0.29} = 0.29
sim (d_3, q) = 0.50
MMR = 0.5*0.5 - 0.5*0.29 = 0.105

Similarly, other MMRs are calculated as:
d_4: -0.35, d_5: 0.06

d_3 has the maximum MMR, therefore

S = {d_1, d_2, d_3}.

If we didn't have diversity at all (λ = 1), then our S would have been {d_1, d_2, d_5}. Notice that the total pairwise similarity of the diverse case is:

sim (d_1, d_2) + sim(d_1, d_3) + sim (d_2, d_3) = **0.63**

whereas the non-diverse version has a total pairwise similarity of **0.87**. We have effectively made the items in the result set more dissimilar to each other. Also note that total similarity to the query has reduced from **2.44** to **2.31**. We traded off some accuracy for the sake of diversity.