Information Retrieval - Lecture 4 - Weighting schemes

colm.oriordan

October 2020

1 Recap

As we have seen in previous classes, one of the big challenges is how to best assign a weight for a term in a given document. We assign to each term in a document a weight for that term that depends on the number of occurrences of the term in the document. We would like to compute a score between a query term t and a document d, based on the weight of t in d. A standard approach is to assign the weight to be a function of the number of occurrences of term t in document d. We typically adopt a "bag of words" model where the actual ordering of the terms in a document is ignored but the number of occurrences of each term is important.

Hence, the document "Mary is quicker than John" is viewed as being identical to the document "John is quicker than Mary".

Using a term's raw frequency suffers from a serious problem: all terms are considered equally important when it comes to assessing relevancy of a document to a query. In reality, certain terms have little or no discriminating power. One approach is to use some global feature (e.g. collection frequency or document frequency) to scale down the weight based on the term frequency.

2 Variants of tf

A common variation is to use the logarithm of the term frequency, which assigns a weight to a term in a document as

$$1 + log(tf_{i,j}), \text{ if } tf_{t,d} > 0$$

The intuition behind this is that repeated occurrences should increase the weight of the term; however, that increase should not necessarily be linear.

Another approach to modifying the raw term score is maximum term normalisation. The approach is to normalize the tf weights of all terms occurring in a document by the maximum tf in that document.

For each document d, let $tfmax(d)$ be the maximally occurring term in the document. The normalized term frequency can then be calculated as:

$$ntf_{i,j} = a + (1-a)\frac{tft_{i,j}}{tfmax(d)}$$

where a is a value between 0 and 1. The term a is a smoothing term to weaken the contribution of the second term.

This is a good method to penalise long documents but does suffer from some problems

3 Weighting Schemes

We can view most weighting schemes as comprising local factors, global factors (collection wide) and query related features. Many, if not all, of the developed or learned weighting schemes can be represented in the following format:

$$sim(q,d) = \sum_{t \in q \cup d} (ntf(D) \times gw_t(C) \times qw_t(Q))$$

where $ntf(D)$ is the normalised term frequency in a document $gw_t(C)$ is the global weight of a term across a collection $qw_t(Q)$ is the query weight of a term in Q, a query

There have been many approach to term weighting that fits this template. Many term weighting schemes, normalisation schemes, document normalisation and global (idf-like) approaches have been developed and empirically investigated on collections.

4 Axiomatic approaches

Beginning in 2004, there have been several bodies of work that have attempted an axiomatic approach to defining term weight functions. These approaches attempt to define axioms (or constraints) that all weighting schemes should obey or adhere to. The advantages of this approach include providing a means to identify where certain weighting schemes may be flawed and a means to guide the search for weighting schemes.

- Adding a query term to a document must always increase the score of a document. The intuition here is that by having that extra term in the document increases the likelihood that the document is 'on topic' or relevant to the query.

- Adding a non-query to a document must always decrease the score of a document. The intuition is that by adding a term not in the query, we are increasing the length of the document but not adding any terms that would indicate the document is more 'on topic'.
- Adding successive occurrences of a term to a document must increase the score of a document less with successive occurrences. Essentially any term-frequency factor should be sub-linear. The intuition is that repeated occurrences of a term increase the likelihood that the document is 'on topic', but that that likelihood does not grow linearly with occurrence. Encountering a term 30 times does not increase the likelihood of relevance by a factor of thirty.
- Ensuring that the document length factor is used in a sub-linear function will ensure that repeated appearances of non-query terms are weighted less.