A Probabilistic Approach on Document Retrieval

By Acar ERDİNÇ

Outline

- Description
- Motivation
- Methodology
- Expected Results

Description

1. Query the Search Engine
2. ??
3. Retrieve the Ranked List of Relevant Documents

Description

How can we obtain the ranked list without using the traditional IR methodologies,

BUT

rather using a probabilistic approach?

Description

With enough data, everything can be learned.

There are many search engine on the web, providing this data.

Motivation

- Learning from the relevance information already provided.
- Avoiding similarity measurements, clustering…
- Personal research field & interest.

Methodology

1. Obtain the data
2. Learn the model
3. Estimate the probability of a document being relevant to a given query. And list the probabilities.

$$P(R \mid D, Q)$$

Methodology

Two possible approach,

- Considering the terms in the documents.
- Ignoring the terms in the documents.

Methodology

- Naïve Bayesian Classification

  Fi is conditionally independent, given the class C,

$$p(F_i|C,F_j) = p(F_i|C), \; p(F_i|C,F_j,F_k) = p(F_i|C), \; p(F_i|C,F_j,F_k,F_l) = p(F_i|C),$$

  join model is,

$$\begin{aligned} p(C|F_1,\dots,F_n) &\propto p(C)\; p(F_1|C)\; p(F_2|C)\; p(F_3|C)\; \cdots \\ &\propto p(C) \prod_{i=1}^{n} p(F_i|C). \end{aligned}$$

Expected Results

Similarity between the acquired ranked list and the ground truth, for a given query.

Ground Truth?

Thanks

FOR

LISTENING