Probabilistic Model

(Teaching Inspired by Research)

Prof. Dr. Marcin Grzegorzek and
the Medical Data Science Team

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MEDIZINISCHE INFORMATIK

Contents of the Course

Week	Lecture	Practical Exercises
1	**(05/04)** Introduction to Medical Information Retrieval (MIR)	**(05/04)** Introduction to Python
2	**(12/04)** Main Components and Classification of MIR Systems	**(12/04)** Introduction to Python
3	**(19/04)** Metadata in Medical Information Retrieval Systems	**(19/04)** CBIR in Medical Applications
4	***(26/04)*** *No Lecture due to a Business Trip*	**(26/04)** CBIR in Medical Applications
5	**(03/05)** Set Theoretic Model: Boolean Retrieval	**(03/05)** CBIR in Medical Applications
6	**(10/05)** Set Theoretic Model: Fuzzy Retrieval	**(10/05)** Flask Tutorial
7	**(17/05)** Vector Space Model: Similarity Measures	**(17/05)** Flask Tutorial

Contents of the Course

8	**(24/05)** Vector Space Model: Distance Functions	**(24/05)** HTML
9	**(31/05)** Vector Space Model: Latent Semantic Indexing	**(31/05)** HTML
10	**(07/06)** Probabilistic Model	**(07/06)** HTML
11	**(14/06)** Text-based Retrieval of Medical Information	**(14/06)** Deep Learning
12	**(21/06)** Audio-based Retrieval of Medical Information	**(21/06)** Deep Learning
13	**(28/06)** Image-based Retrieval of Medical Information	**(28/06)** Relevance Feedback
14	**(05/07)** Demonstrators from Current Research Projects	**(05/07)** Relevance Feedback
15	**(12/07)** Summary and Conclusions	**(12/07)** Evaluation

Probabilistic Model – Optimum Effectiveness

“If a reference retrieval system’s response to each request is a ranking of the documents in the collection in order of ***decreasing probability of relevance*** *to the user who submitted the request, where the* ***probabilities are estimated as accurately as possible*** *on the basis of whatever data have been made available to the system for this purpose,* ***the overall effectiveness of the system*** *to its user* ***will be the best*** *that is obtainable on the basis of those data.”*

[van Rijsbergen 1979]

Probability of Relevance Based on Historical Data

The probability of relevance can be modelled using historical data, if available:

$$P(R = 1|\boldsymbol{d}_i, \boldsymbol{q}) = \frac{N_{\boldsymbol{q},\boldsymbol{d}_i,R=1}}{N_{\boldsymbol{q},\boldsymbol{d}_i,R\in\{0,1\}}} \quad .$$

Probability of Relevance for Unseen Queries and Documents

For unseen queries and/or documents, we can assume the following approximation:

$$P(R = 1|\boldsymbol{d}_i, \boldsymbol{q}) \approx p(\boldsymbol{q}|\boldsymbol{d}_i, R = 1) \quad .$$

A Priori and A Posteriori Probability

- A priori probability: $P(\boldsymbol{d}_i)$.
- A posteriori probability: $P(\boldsymbol{d}_i|\boldsymbol{q})$.

Likelihood Density Function

- The likelihood density function $p(\boldsymbol{q}|\boldsymbol{d}_i)$ describes how vectors $\boldsymbol{q}$ are distributed within $\boldsymbol{d}_i$.
- It is usually trained from examples.

Bayes Decision Theory for the Retrieval Problem

Known:

- Documents: $\boldsymbol{d}_i$
- A priori probabilities: $P(\boldsymbol{d}_i)$
- Likelihood density functions: $p(\boldsymbol{q}|\boldsymbol{d}_i)$
- A query to be processed: $\boldsymbol{q} = (q_1, q_2, \ldots, q_I)^{\mathrm{T}}$

Unknown:

- A posteriori probabilities: $P(\boldsymbol{d}_i|\boldsymbol{q})$

Computation of the A Posteriori Probability

- Using the **Bayes rule** we obtain:

$$P(\boldsymbol{d}_i|\boldsymbol{q}) = \frac{p(\boldsymbol{q}|\boldsymbol{d}_i)P(\boldsymbol{d}_i)}{p(\boldsymbol{q})} \quad .$$

Final Retrieval Result Based on Bayes Decision Theory

Final Retrieval Result Based on Bayes Decision Theory

Statistical Modelling of Text Documents

What is the probability of the next word:

$$p(\text{house}|\text{this is the}) = ? \qquad p(\text{did}|\text{this is the}) = ?$$

Statistical Modelling of One-Dimensional Time Signals

Statistical Modelling of Images

Final Statements

- If documents can be represented as probability density functions over possible queries, the statistical approaches used for supervised classification can also be applied for retrieval.
- Multimedia documents of different kind (text, audio, image, etc.) usually require different techniques for the statistical modelling of their contents.