Language Models and Smoothing Methods for Collections with Large Variation in Document Length

Najeeb Abdulmutalib, Norbert Fuhr

University of Duisburg-Essen
najeeb@uni-due.de

TIR-08
5th International Workshop on Text-based Information Retrieval
Turin, Italy
1 September 2008
Motivation

- Document length effect on the retrieval effectiveness
- Smoothing and the retrieval performance
An Outline

- Models
- Smoothing methods
- Experiments
- Results
Basic model

\[ P(q|d) = \prod_{t_i \in q^T} P(t_i|d) \]

\[ = \prod_{t_i \in q^T \cap d^T} P_s(t_i|d) \prod_{t_i \in q^T - d^T} P_u(t_i|d) \]

\[ = \prod_{t_i \in q^T \cap d^T} \frac{P_s(t_i|d)}{P_u(t_i|d)} \prod_{t_i \in q^T} P_u(t_i|d) \]  

(1)
An Odds model

As an alternative to the basic prob. model, we propose an odds-like model

\[
\frac{P(d|q)}{P(\bar{d}|q)} = \frac{P(q|d)}{P(q|\bar{d})} \cdot \frac{P(d)}{P(\bar{d})}
\]

\[
\approx \prod_{t_i \in q^T \cap d^T} \frac{P(t_i|d) P(d)}{P(t_i|\bar{d}) P(\bar{d})}
\]

\[
= \prod_{t_i \in q^T \cap d^T} \frac{P_s(t_i|d)}{P_s(t_i|\bar{d})} \prod_{t_i \in q^{T-d^T}} \frac{P_u(t_i|d)}{P_u(t_i|\bar{d})} \cdot \frac{P(d)}{P(\bar{d})}
\]
Some known smoothing methods

- The Jelinek-Mercer method involves a linear interpolation

\[ P_{s,\lambda}(t_i|d) = (1 - \lambda) \cdot P_{ML}(t_i|d) + \lambda \cdot P_{avg}(t_i|C) \]

- Bayesian parameter estimation with Dirichlet distribution is a document length-dependent smoothing factor

\[ P_{s,\mu}(t_i|d) = \frac{c(t_i; d) + \mu P_{avg}(t_i|C)}{\sum_{t_i \in d^T} c(t_i; d) + \mu} \]
Exponential formula

- Our alternative way of smoothing, combining \( P_{ML}(t_i|d) \) and \( P_{avg}(t_i|C) \) as an estimate of \( P_s(t_i,d) \)
- And we estimate \( P_u(t_i|d) \) as a function of \( P_{avg}(t_i|C) \)

\[
\begin{align*}
P_s,e(t_i|d) &= P_{ML}(t_i|d)^{\alpha_d} \cdot P_{avg}(t_i|C)^{1-\alpha_d} \\
P_u,e(t_i|d) &= P_{avg}(t_i|C)^{\beta_d}
\end{align*}
\]
Retrieval Functions

- We have combined the models with the exponential smoothing method
- The Odds model

\[ \rho_{o,e} = \prod_{t_i \in q^T \cap d^T} \left( \frac{P_{ML}(t_i | d)}{P_{avg}(t_i | C)} \right)^{\omega_d} \cdot \prod_{t_i \in q^T - d^T} P_{avg}(t_i | C)^{\gamma_d} \cdot \frac{P(d)}{P(\bar{d})} \]

- The Prob model

\[ \rho_{p,e} = \prod_{t_i \in q^T \cap d^T} \frac{P_{ML}(t_i | d)^{\alpha_d}}{P_{avg}(t_i | C)^{\beta_d + \alpha_d - 1}} \prod_{t_i \in q^T} P_{avg}(t_i | C)^{\beta_d} \]
Collection

- INEX 2005 IEEE collection, version 1.9
- 16,819 journal articles in XML format, comprising 764 MB of data
- We regarded each XML element as an independent document
- 21.6 million documents with a collection size of more than 253 million words
- We have a good test case for investigating the influence of document length variation on the retrieval quality of language models.
Distribution of document length in our test collection
Experiments

For the retrieval part

- we considered the CO queries from INEX 2005 along with the official adhoc assessments
The effect of considering document length

- We assume that the probabilities $P(d)$ and $P(\bar{d})$ are proportional to document length.
- For the Odds model, the factor $\frac{p(d)}{p(\bar{d})}$ was omitted from the retrieval formula $\rho_{o,e}$ when document length was ignored.
- In the case of the Probability model, the functions for $\rho_{p,e}^d$ and $\rho_{p,e}$ were compared.
Models results with and without using dl

![Bar chart showing MAP values for different models with and without deep learning.]
Influence of smoothing parameters on MAP when using Odds model
Influence of smoothing parameters on MAP when using Prob. model
Best results for models and smoothing methods

MAP

0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0

Expo. J-M Baysian

Odds Prob Z-L
P@k relevant documents

![Precision-Rank Curve for the Best Runs](attachment:image.png)

#### Abbildung:
Precision-rank curve for the best runs

Najeeb Abdulmutalib
Language Models and Smoothing Methods
University of Duisburg-Essen
Conclusions and Outlook

- New language model based on an odds formula
- New smoothing method called exponential smoothing
- Our new model along with the new smoothing method give very good results
- Document length is an important factor for language models, so models ignoring this parameter lead to very poor results