Geometric Perspectives of the BM25

Giorgio Maria Di Nunzio
Dept. of Information Engineering
University of Padua
Outline

• Probabilistic Retrieval Models
• Binary Independence Model (BIM)
• BestMatch 25 (BM25)
• Likelihood Spaces
• Conclusions
Probability Ranking Principle (PRP)

- The objective of a retrieval function is to rank the documents of a collection according to a specific information need or query.

- Optimal retrieval can be achieved by ranking documents in order of decreasing probability of relevance (PRP).

\[ P(R = r | D) \]

Probability Ranking Principle (PRP)

- The probability of a document being relevant is rank equivalent (given a query) to the odds

\[
P(R = r|D) \propto_{\text{rank}} \frac{P(R=r|D)}{P(R=\bar{r}|D)}
\]

\[
= \frac{P(D|R=r)P(R=r)}{P(D|R=\bar{r})P(R=\bar{r})}
\]

\[
\propto_{\text{rank}} \frac{P(D|R=r)}{P(D|R=\bar{r})}
\]
Bayesian Decision Theory 1/2

• The identification of relevant documents can be achieved by selecting the decision that minimizes the conditional risk (under zero-one loss):

\[ P(R = r|D) > P(R = \bar{r}|D) \]

\[ \frac{P(R = r|D)}{P(R = \bar{r}|D)} > 1 \]
Bayesian Decision Theory 2/2

- The identification of relevant documents can be achieved by selecting the decision that minimizes the conditional risk (under zero-one loss):

\[ P(D|R=r)P(R=r) > P(D|R=\bar{r})P(R=\bar{r}) \]

\[ \frac{P(D|R=r)}{P(D|R=\bar{r})} > \frac{P(R=r)}{P(R=\bar{r})} \]

Binary Independence Model (BIM)

- Documents as binary vectors. Each word is distributed as a Bernoulli function.

\[
\frac{P(D = d_i|R = r)}{P(D = d_i|R = \bar{r})} = \prod_{w_t \in \mathcal{V}} \frac{\theta_{w_t|r}^{x_{it}} (1 - \theta_{w_t|r})^{1 - x_{it}}}{\theta_{w_t|\bar{r}}^{x_{it}} (1 - \theta_{w_t|\bar{r}})^{1 - x_{it}}} 
\]

\[
\log \left( \frac{P(D = d_i|R = r)}{P(D = d_i|R = \bar{r})} \right) = \sum_{w_t \in \mathcal{V}} x_{it} \log \left( \frac{\theta_{w_t|r}}{\theta_{w_t|\bar{r}}} \frac{(1 - \theta_{w_t|r})}{(1 - \theta_{w_t|\bar{r}})} \right) + \sum_{w_t \in \mathcal{V}} \log \left( \frac{(1 - \theta_{w_t|r})}{(1 - \theta_{w_t|\bar{r}})} \right)
\]

Binary Independence Model (BIM)

- Documents ranked according to the sum of relevance weights RW.

\[
\log \left( \frac{P(D|R = r)}{P(D|R = \bar{r})} \right) \propto \text{rank} \sum_{w_t \in d_t} \log \left( \frac{\theta_{w_t|r}(1 - \theta_{w_t|r})}{\theta_{w_t|\bar{r}}(1 - \theta_{w_t|\bar{r}})} \right)
\]

\[
RW(w_t) = \log \left( \frac{p(1 - q)}{q(1 - p)} \right)
\]
BM25 (exact formula)

• Each word is a mixture of 2 Poisson distributions
  • $tf = 0$, $w_t = 0$
  • monotonic in $tf$
  • asymptote approximates BIM

$$w_t = \log \left( \frac{(p' \lambda^{tf} e^{-\lambda} + (1 - p') \mu^{tf} e^{-\mu}) (q' e^{\lambda} + (1 - q') e^{\mu})}{(q' \lambda^{tf} e^{-\lambda} + (1 - q') \mu^{tf} e^{-\mu}) (p' e^{\lambda} + (1 - p') e^{\mu})} \right)$$

BM25 (approximated)

- TF and document length normalisation
- Pivoted normalisation

\[ w(t, d_i) = \frac{TF(t, d_i)}{TF(t, d_i) + K} \log \left( \frac{p(1 - q)}{q(1 - p)} \right) \]

\[ K = k_1 \left( (1 - b) + b \frac{dl(d_i)}{\Delta} \right) \]

Likelihood Spaces

- Coordinates of a two-dimensional space

\[ \log(P(D|R = r)) > \log(P(D|R = \bar{r})) + \log \left( \frac{P(R = \bar{r})}{P(R = r)} \right) \]
Ranking in Likelihood spaces

Tipster 4-5

Select topic: international organized crime

Validation on fold 1:
- Num of objects: 2011
- Num of positive examples: 19
- P5, P10, P20, P50, P100, P500, P1000
  - R5: 0.600
  - R10: 0.800
  - R20: 0.800
  - R50: 0.380
  - R100: 0.190
  - R500: 0.038
  - R1000: 0.019

Test:
- Num of objects: 15940
- Num of positive examples: 448
- P5, P10, P20, P50, P100, P500, P1000
  - R5: 0.800
  - R10: 0.800
  - R20: 0.750
  - R50: 0.740
  - R100: 0.640
  - R500: 0.338
  - R1000: 0.210

G. M. Di Nunzio

Geometric Perspectives of the BM25
Ranking in Likelihood Spaces

- Change slope $M$ of the decision line

\[
\log(P(D|R = r) \mid X) > \log(P(D|R = \bar{r}) \mid Y)
\]

\[
M \log(P(D|R = r) \mid X) + Q > \log(P(D|R = \bar{r}) \mid Y)
\]

Ranking in Likelihood Spaces
Conclusions

- Study probabilistic models on two dimensions
- BM25 vs BIM
- BM25 and Bayesian Decision Theory
- Ranking Line and BDT