## Key Points

By assuming independence between query terms, Robertson and Sparck-Jones proposed for the probability $p(Rel|d,q)$ the following model (the RSJ model [2]):

$$\log(p(Rel|d,q)) \propto \sum_{t \in q} \log \frac{p(t|Rel) \cdot p(\overline{t}|\overline{Rel})}{p(t|\overline{Rel}) \cdot p(\overline{t}|Rel)} \quad (1)$$

where $\overline{Rel}$ indicates the event of non-relevance; $t$ and $\overline{t}$ indicate the events that the term $t$ occurs in document $d$ or does not, respectively. For each query term $t$, the probability $p(Rel|d,t)$ is given by the sum of two log-odds, $\log \frac{p(t|Rel)}{p(t|\overline{Rel})}$ and $\log \frac{p(\overline{t}|\overline{Rel})}{p(\overline{t}|Rel)}$.

If $N$ is the number of documents in the whole collection, $R$ is the number of relevant documents, $r$ is the number of relevant documents containing $t$, $N_t$ is the document frequency, i.e., the number of documents containing $t$, [3] instantiated the RSJ model as follows:

$$w^{(1)} = \log \frac{(r+0.5)(N - N_t - R + r + 0.5)}{(R - r + 0.5)(N_t - r + 0.5)} \quad (2)$$

where $w^{(1)}$ is the raw weight of a term $t$ in a document $d$. The number 0.5 is used to avoid assigning negative weights. The formula is called the "point-5" formula.

If relevance information is not available, i.e., $R = r = 0$, the point-5 formula can be written as:

$$w^{(1)} = \log \frac{N - N_t + 0.5}{N_t + 0.5} \quad (3)$$

As one of the most well-established IR systems, Okapi uses a weighting model that is based on the RSJ model introduced above, and takes also term frequency ($tf$) and query term frequency ($qtf$) into consideration.

## Cross-references

- ▶ Information Retrieval
- ▶ Information Retrieval Models
- ▶ Term weighting

## Recommended Reading


1. Robertson S.E. The probability ranking principle in IR. J. Doc., 33:294–304, 1977.
2. Robertson S.E. and Sparck-Jones K. Relevance weighting of search terms. J. Am. Soc. Inf. Sci., 27:129–146, 1977.
3. Robertson S.E. and Walker S. On relevance weights with little relevance information. In Proc. 20th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1997, pp. 16–24.


# Probability Smoothing

Djoerd Hiemstra
University of Twente, AE Enschede, The Netherlands

## Definition

Probability smoothing is a language modeling technique that assigns some non-zero probability to events that were unseen in the training data. This has the effect that the probability mass is divided over more events, hence the probability distribution becomes more *smooth*.

## Key Points

Smoothing overcomes the so-called *sparse data problem*, that is, many events that are plausible in reality are not found in the data used to estimate probabilities. When using maximum likelihood estimates, unseen events are assigned zero probability. In case of information retrieval, most events are unseen in the data, even if simple unigram language models are used documents are relatively short (say on average several hundreds of words), whereas the vocabulary is typically big (maybe millions of words), so the vast majority of words does not occur in the document. A small document about "information retrieval" might not mention the word "search," but that does not mean it is not relevant to the query "text search." The sparse data problem is the reason that it is hard for information retrieval systems to obtain high recall values without degrading values for precision, and smoothing is a means to increase recall (possibly degrading precision in the process). Many approaches to smoothing are proposed in the field of automatic speech recognition [1]. A smoothing method may be as simple so-called Laplace smoothing, which adds an extra count to every possible word. The following equations show respectively (1) the unsmoothed, or maximum likelihood estimate, (2) Laplace smoothing, (3) Linear interpolation smoothing, and (4) Dirichlet smoothing [3]:

$$P_{ML}(T = t|D = d) = tf(t,d) / \sum_{t'} tf(t',d) \quad (1)$$

$$P_{LP}(T = t|D = d) = (tf(t,d) + 1) / \sum_{t'} (tf(t',d) + 1) \quad (2)$$

$$P_{LI}(T = t|D = d) = \lambda P_{ML}(T = t|D = d) + (1 - \lambda)P_{ML}(T = t|C) \quad (3)$$

$$P_{Di}(T = t|D = d) = (tf(t, d) + \mu P_{ML}(T = t|C)) / ((\sum_{t'} tf(t', d)) + \mu) \quad (4)$$

Here, $tf(t,d)$ is the frequency of occurrence of the term $t$ in the document $d$, and $P_{ML}(T|C)$ is the probability of a term occurring in the entire collection $C$. Both linear interpolation smoothing and Dirichlet smoothing assign a probability proportional to the term occurrence in the collection to unseen terms. Here, $\lambda$ ($0<\lambda<1$) and $\mu$ ($\mu>0$) are unknown parameters that should be tuned to optimize retrieval effectiveness. Linear interpolation smoothing has the same effect on all documents, whereas Dirichlet smoothing has a relatively big effect on small documents, but a relatively small effect on bigger documents. Many smoothed estimators used for language models in information retrieval (including Laplace and Dirichlet smoothing) are approximations to the *Bayesian predictive distribution* [2].

### Cross-references

► Language Models
► N-Gram Models

### Recommended Reading

1. Chen S.F. and Goodman J. An empirical study of smoothing techniques for language modeling. Technical report TR-10-98, Center for Research in Computing Technology, Harvard University, August 1998.
2. Zaragoza H., Hiemstra D., Tipping M., and Robertson S. Bayesian extension to the language model for ad hoc information retrieval. In Proc. 26th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2003, pp. 4–9.
3. Zhai C. and Lafferty J. A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst., 22(2):179–214, 2004.

## Procedure Order

► Clinical Order

## Procedure Request

► Clinical Order

## Process Composition

► Composition

## Process Definition

► Workflow Model

## Process Evolution

► Workflow Evolution

## Process Life Cycle

Nathaniel Palmer
Workflow Management Coalition, Hingham, MA, USA

### Synonyms

Workflow lifecycle; Thread lifecycle; Process state model

### Definition

The stages of life from the start to the end of a process instance within the context of workflow management.

### Key Points

The Process Life Cycle represents the stages of a process instance as it evolves from instantiation to termination. This life cycle is most closely related to the life cycle of a thread, and is distinct from the life cycle approach to Business Process Management initiatives, involving an iterative or recursive evolution through the five stages of design, modeling, execution, monitoring, and optimization.

The latter notion of Business Process Management Life cycle is associated with the discipline of continuous process improvement, whereby processes are