Robertson-Spärck-Jones Probabilistic Model Tutorial

Abstract – This is a tutorial on the Robertson-Spärck-Jones Probabilistic Model. The model computes global weights, known as RSJ weights, based on Independence Assumptions and Ordering Principles for probable relevance.

Keywords: probabilistic model, independence assumptions, ordering principles, idf, inverse document frequency
Published: 03-30-2009; Updated: 09-26-2016
© E. Garcia, PhD; admin@minerazzi.com

Note: This article is part of a legacy series that the author published circa 2009 at http://www.miislita.com, now a search engine site. It is now republished in pdf format here at http://www.minerazzi.com, with its content edited and updated. The original articles can be found referenced in online research publications on IR and elsewhere.

Introduction

In 1976, Stephen Robertson and Karen Spärck-Jones proposed a probabilistic model for information retrieval under the following assumptions and principles:

Independence Assumptions:

- I1 – The distribution of terms in relevant documents is independent and their distribution in all documents is independent.
- I2 – The distribution of terms in relevant documents is independent and their distribution in non-relevant documents is independent.

Ordering Principles:

- O1 – Probable relevance is based *only* on the presence of query terms in the documents.
- O2 – Probable relevance is based on *both* the presence and absence of query terms in the documents.

In the next section of this tutorial we explain these assumptions and principles.

Discussion

I1 states that the presence of a term in a relevant document does not impact the presence of other terms in the same document or its presence in other relevant documents. I1 says nothing about the distribution of terms in non-relevant documents.

I2 extends I1 to non-relevant documents by stating that the presence of a term in a non-relevant document does not impact the presence of other terms in the same document or its presence in other non-relevant documents. Since documents are either relevant or non-relevant to a query, this is why I2 is more realistic than I1.

O1 indicates that documents should be ranked only if they contain all of the terms specified in a query. It is an AND approach. It says nothing about the absence of query terms in the documents and is therefore incorrect.

O2 takes O1 a little further and states that we should consider both the presence and absence of query terms. It is an OR approach. So for a query consisting of two terms t1 and t2, documents mentioning both terms should rank higher than those mentioning one or none of these terms.

To implement O2, a system using an inverted index has to identify all terms present and not present in a document. To avoid exhaustively tracking the inverted index, we can assign zero probability of relevance to documents lacking of all query terms. Adopting this strategy implies that we have some evidence of non-relevance. It also has the effect of artificially converting O2-based weights to presence-only O1 weights. This makes O2 more practical than O1.

Table 1 summarizes the above assumptions and principles.

Table 1. Assumptions-Principles Contingency Table.

		Independence Assumptions	
		I1	I2
Ordering	O1	F1	F2
Principles	O2	F3	F4

In Table 1, F1-F4 are weighting functions. According to Robertson and Spärck-Jones (1976), I2 is more realistic than I1 while O2 is correct and O1 is incorrect. The model then predicts that F4 is likely to yield the best results and is therefore the best match.

In the next section, we explain how these functions rank documents in the presence and absence of relevance information.

Derivation

Given a query consisting of a term and a collection of documents, the Robertson-Spärck-Jones Probabilistic Model (RSJ-PM) addresses two cardinal questions:

- Is the term present in the documents?
  Answer: 1 = Yes (Present), 0 = No (Absent).
- Are the documents relevant to the term?
  Answer: 1 = Yes (Relevant), 0 = No (Non-Relevant).

Said binary treatment can be extended to queries consisting of several terms. To simplify, in this tutorial we limit the discussion to one-term queries. We begin by stating the following definitions:

r = number of relevant documents that contain the term.
n – r = number of non-relevant documents that contain the term.
n = number of documents that contain the term.
R – r = number of relevant documents that do not contain the term.
N – n – R + r = number of non-relevant documents that do not contain the term.
N - n = number of documents that do not contain the term.
R = number of relevant documents.
N – R = number of non-relevant documents.
N = number of documents in the collection.

Next, we construct the following contingency table.

Table 2. Contingency Table for a collection and a one-term query.

		Are the documents relevant to the term?		
		1 = Yes (Relevant)	0 = No (Non-Relevant)	Collection-wide incidence
Is the term present in the documents?	1 = Yes (Present)	r	n - r	n
	0 = No (Absent)	R - r	N – n – R + r	N - n
Total number of documents		R	N - R	N

Normalizing column elements, a table of probabilities is generated.

Table 3. Array of Probabilities.

Relevant	Non-relevant	Collection-wide incidence
r/R	(n – r)/(N-R)	n/N
(R – r)/R	(N – n – R + r)/(N – R)	(N – n)/N

A table of odds is then obtained by taking ratios.

Table 4. Array of Odds.

Relevant	Non-relevant	Collection-wide incidence
r/(R – r)	(n – r)/(N – n – R + r)	n/(N – n)

Now let *t* be a term. Reading from left to right Tables 3 and 4:

r/R = probability that a relevant document contains *t*.

(n – r)/(N-R) = probability that a non-relevant document contains *t*.

n/N = probability that a document contains *t*.

(R – r)/R = probability that a relevant document does not contain *t*.

(N – n – R + r)/(N – R) = probability that a non-relevant document does not contain *t*.

(N – n)/N = probability that a document does not contain *t*.

r/(R – r) = odds that a relevant document contains *t*.

(n – r)/(N – n – R + r) = odds that a non-relevant document contains *t*.

n/(N – n) = odds that a document contains *t*.

We now do some collection-wide and distribution-specific comparisons. The fraction of relevant documents containing the term (r/R) is compared in two different ways:

- Against the fraction of documents in the collection containing the term; i.e., (n/N).
- Against the fraction of non-relevant documents containing the term; i.e., (n – r)/(N-R).

Likewise, the odds that relevant documents contain the term (r/(R – r)) is compared in two different ways:

- Against the odds that documents from the collection contain the term; i.e., n/(N – n).
- Against the odds that non-relevant documents contain the term; i.e., (n – r)/(N – n – R + r).

To account for the fact that term weights are additive we take logarithms. This yields explicit expressions for the four weighting functions given in Table 1. The functions score global weights which we now call RSJ weights, after Robertson and Spärck-Jones. These weights are summarized in Table 5.

Table 5. Scoring Functions.

Weighting Function	Remarks
$\mathbf{F1} = \mathbf{log}\left(\frac{\mathbf{r/R}}{\mathbf{n/N}}\right)$	F1 evaluates the ratio of the proportion of relevant documents in which the term occurs to the proportion of the entire collection in which it occurs.
$\mathbf{F2} = \mathbf{log}\left(\frac{\mathbf{r/R}}{(\mathbf{n}-\mathbf{r})/(\mathbf{N}-\mathbf{R})}\right)$	F2 evaluates the ratio of the proportion of relevant documents to that of non-relevant documents.
$\mathbf{F3} = \mathbf{log}\left(\frac{\mathbf{r}/(\mathbf{R}-\mathbf{r})}{\mathbf{n}/(\mathbf{N}-\mathbf{n})}\right)$	F3 evaluates the ratio between the “relevance odds” for the term (i.e., the ratio between the number of relevant documents in which it does occur and the number in which it does not occur) and the “collection odds” for the term.
$\mathbf{F4} = \mathbf{log}\left(\frac{\mathbf{r}/(\mathbf{R}-\mathbf{r})}{(\mathbf{n}-\mathbf{r})/(\mathbf{N}-\mathbf{n}-\mathbf{R}+\mathbf{r})}\right)$	F4 evaluates the ratio between the term relevance odds and its “non-relevance odds”.

In Table 5, F1 through F4 are scoring functions that evaluate the weight of term i, $w(t_i)$, as log transformations. These comparisons and transformations are not arbitrary. Let see why.

In the absence of relevance information the only information available is collection-wide incidence: the n/N and n/(N – n) ratios. It seems intuitively correct to propose scoring functions that use the n/N and n/(N – n) ratios as reference points. By doing so, we are effectively comparing against collection-wide proportions.

If we recall, log(N/n) is called Inverse Document Frequency (IDF) and log ((N – n)/n)) is its "odds version" also known as IDF Probabilistic (IDFP). Considering these as weighting functions we can write

$$F0 = \log\left(\frac{N}{n}\right) = IDF \quad (1)$$

$$F00 = \log\left(\frac{N-n}{n}\right) = IDFP \quad (2)$$

As weighting functions, F0 and F00 evaluate the weight of term i, $w(t_i)$, but with one caveat: without incorporating relevance information. Thus (1) and (2) are collection-wide estimators of the discriminatory power of a term (*term specificity*).

Thus, an IDF value is an RSJ weight in the absence of relevance information. This is also true for IDFP. It is now evident that

$$F1 = \log\left(\frac{r}{R}\right) + IDF \quad (3)$$

$$F3 = \log\left(\frac{r}{(R-r)}\right) + IDFP \quad (4)$$

That is, F1 and F3 compensate for the lack of relevance information in IDF and IDFP weights by incorporating a relevance component. Note that F1 and F3 are related by comparing the relevant document distribution of a term to its entire collection distribution.

In the case of F2 and F4, these functions are related by comparing relevant and non-relevant distributions. It is possible to derive IDF and IDFP from F2 and F4 by making specific assumptions about the degree of relevance information available. For instance, IDFP can be obtained by setting R and r equal to zero in F4 (Robertson, 2004).

Regarding the use of logarithms, we must remember that these are additive: the log of a product is a sum of logs. This additive property is frequently assumed in IR with term matching coefficients (Robertson, 2004, Robertson & Spärck-Jones, 1976). In the next section, we show that RSJ model can be used in two different ways:

- retrospectively
- predictively

Using the Model Predictively

According to Robertson and Spärck-Jones, if the model is used retrospectively, the use of proportions as estimates is recommended.

However if used predictively, the model breakdown when n, r, N, R, n - r, N – n, or N – R = 0. This can be avoided by adding a correction factor k to the entries of Table 2. See Table 6.

Table 6. Contingency Table with Correction Factor k.

		Are the documents relevant to the term?		
		1 = Yes (Relevant)	0 = No (Non-Relevant)	Collection-wide Incidence
Is the term present in the documents?	1 = Yes (Present)	r + k	n – r + k	n + 2k
	0 = No (Absent)	R – r + k	N – n – R + r + k	N – n + 2k
Total number of documents		R + 2k	N – R + 2k	N + 4k

Using the model predictively means making inferences about the probabilities on the basis of sample information available. This is problematic for small samples. In their original paper, the authors used k = 0.5 (a *point-5 correction*) and obtained the scoring functions depicted in Table 7.

Table 7. RSJ Model with correction factor k.

RSJ Predictive Functions	k = 0.5
$\mathbf{F1 = \log\left(\frac{(r+k)/(R+2k)}{(n+2k)/(N+4k)}\right)}$	$F1 = \log\left(\frac{(r+0.5)/(R+1)}{(n+1)/(N+2)}\right)$
$\mathbf{F2 = \log\left(\frac{(r+k)/(R+2k)}{(n-r+k)/(N-R+2k)}\right)}$	$F2 = \log\left(\frac{(r+0.5)/(R+1)}{(n-r+0.5)/(N-R+1)}\right)$
$\mathbf{F3 = \log\left(\frac{(r+k)/(R-r+k)}{(n+2k)/(N-n+2k)}\right)}$	$F3 = \log\left(\frac{(r+0.5)/(R-r+0.5)}{(n+1)/(N-n+1)}\right)$
$\mathbf{F4 = \log\left(\frac{(r+k)/(R-r+k)}{(n-r+k)/(N-n-R+r+k)}\right)}$	$F4 = \log\left(\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)}\right)$

A Working Example

Robertson and Spärck-Jones (1976) applied their model to a collection of 200 documents of which 5 were relevant to terms *a*, *b*, *c*, *d*, and *e*.

In Table 8 we have reproduced their results. For comparison purposes we computed results for F0 and F00. Results were computed using both versions of the model.

Table 8. RSJ weights for five terms (*a*, *b*, *c*, *d*, and *e*) with k = 0 and k = 0.5.

		Retrospective Mode (k = 0)						Predictive Mode (k = 0.5)			
a	N	R	n	r		*a*	N	R	n	r	
	200	5	5	1			200	5	5	1	
F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
1.60	1.59	0.90	0.99	0.99	1.08	1.53	1.51	0.93	1.04	1.04	1.15
b	N	R	n	r		*b*	N	R	n	r	
	200	5	5	4			200	5	5	4	
F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
1.60	1.59	1.51	2.19	2.19	2.89	1.53	1.51	1.40	1.99	1.99	2.59
c	N	R	n	r		*c*	N	R	n	r	
	200	5	100	1			200	5	100	1	
F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
0.30	0.00	-0.40	-0.40	-0.60	-0.62	0.30	0.00	-0.30	-0.31	-0.48	-0.49
d	N	R	n	r		*d*	N	R	n	r	
	200	5	100	4			200	5	100	4	
F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
0.30	0.00	0.20	0.21	0.60	0.62	0.30	0.00	0.18	0.18	0.48	0.49
e	N	R	n	r		*e*	N	R	n	r	
	200	5	20	3			200	5	20	3	
F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
1.00	0.95	0.78	0.84	1.13	1.20	0.98	0.94	0.75	0.82	1.08	1.15

In Table 9 we have reordered Table 8 results with respect to the r/n ratio (the probability that documents containing a term are relevant).

Table 9. Reordering of Table 8 results with respect to the r/n ratio.

R/N (%) = 2.5						Results with k = 0						Results with k = 0.5					
	r	n	r/N (%)	n/N (%)	r/n (%)	F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
c	1	100	0.5	50	1	0.30	0.00	-0.40	-0.40	-0.60	-0.62	0.30	0.00	-0.30	-0.31	-0.48	-0.49
d	4	100	2	50	4	0.30	0.00	0.20	0.21	0.60	0.62	0.30	0.00	0.18	0.18	0.48	0.49
e	3	20	1.5	10	15	1.00	0.95	0.78	0.84	1.13	1.20	0.98	0.94	0.75	0.82	1.08	1.15
a	1	5	0.5	2.5	20	1.60	1.59	0.90	0.99	0.99	1.08	1.53	1.51	0.93	1.04	1.04	1.15
b	4	5	2	2.5	80	1.60	1.59	1.51	2.19	2.19	2.89	1.53	1.51	1.40	1.99	1.99	2.59

To understand better these results, we have also computed the R/N and n/N ratios. It is clear that:

- A weight of zero is obtained when R/N = r/n. In addition, the theory predicts that F00 should give a zero IDFP weight when n/N = 0.5 and a negative weight when n/N > 0.5.
- For r/n < R/N all four functions (F1 through F4) give a negative weight to *c* since documents chosen at random from those containing the term *c* is less likely to be relevant than one chosen at random from the whole collection.
- As r/n increases and n/N decreases, F1-F4 separate terms as expected: *b* > *a* and *d* > *c*.
- The relationship of *a* and *e* shows that the four functions do not necessarily rank terms in the same order.
- With k = 0, F2 and F3 assign the same weight to *a* (0.99) and to *b* (2.19). This is also observed when k = 0.5.
- F4 assigns higher weights than F1, F2, and F3.
- F0 and F00 assign higher weights than F4 when both r/N and r/n are small (for *a*, these respectively are 0.5 % and 2.5 %).

Revisiting k

Robertson and Spärck-Jones adopted the idea of using the correction factor k from Cox (1970).

From the above discussion, it is still unclear the role of k in their information retrieval model. In particular,

- What is the effect of varying k for a given weighting function across terms?
- What is the effect of varying k for a given term across weighting functions?

To address these questions, in Figure 1 we inspected the effect of varying r/n for each of the scoring functions. Notice that varying k from 0 to 0.5 dampens down the curves. These results confirm the generalized perception that k is a smoothing correction.

In Figure 2, we examined the effect of varying k for each of the scoring functions. In the figure, w(a) stands for the weight assigned to *a*, w(b) for the weight assigned to *b*, and so forth. Again, k acts as a smoothing correction.

It should be underscored that setting k does impact the scoring functions in a non-trivial way. Figure 2 shows that curve slopes differ across terms and scoring functions. This is a reflection of the several combinations of relevant/non-relevant documents.

The absolute values of the slopes are indicative of how sensitive the scoring functions are to k. When the curves overlap, selecting one scoring function over the other for a particular k does not matter that much.

By contrast, when function curves are orthogonal to the y axis, using these predictively or retrospectively should return identical results. This is the case of F0 and F00 with terms *c* and *d*. Notice that F4 gives higher weights to *b*, *d*, and *e*; i.e. to terms with a high relevant document incidence, r.

We can extend on this subject and argue that varying k does provide some insight as to when and why some functions assign lower or higher weights. Such a discussion is a great homework and complementary research work for this tutorial.

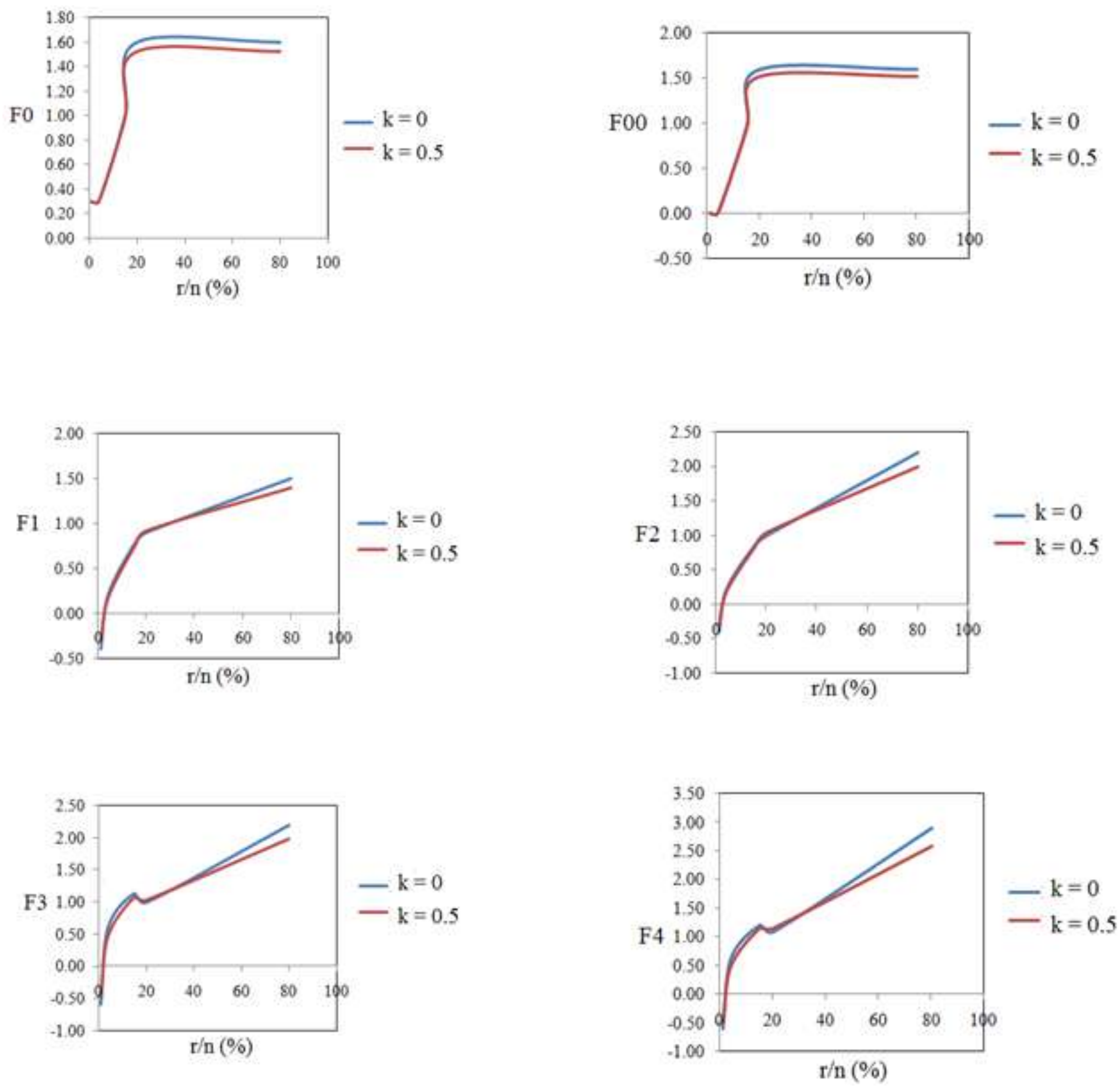

Figure 1. Profile curves of term weights (w) vs. r/n ratios at k = 0 and k = 0.5 for *a*, *b*, *c*, *d*, and *e* terms.

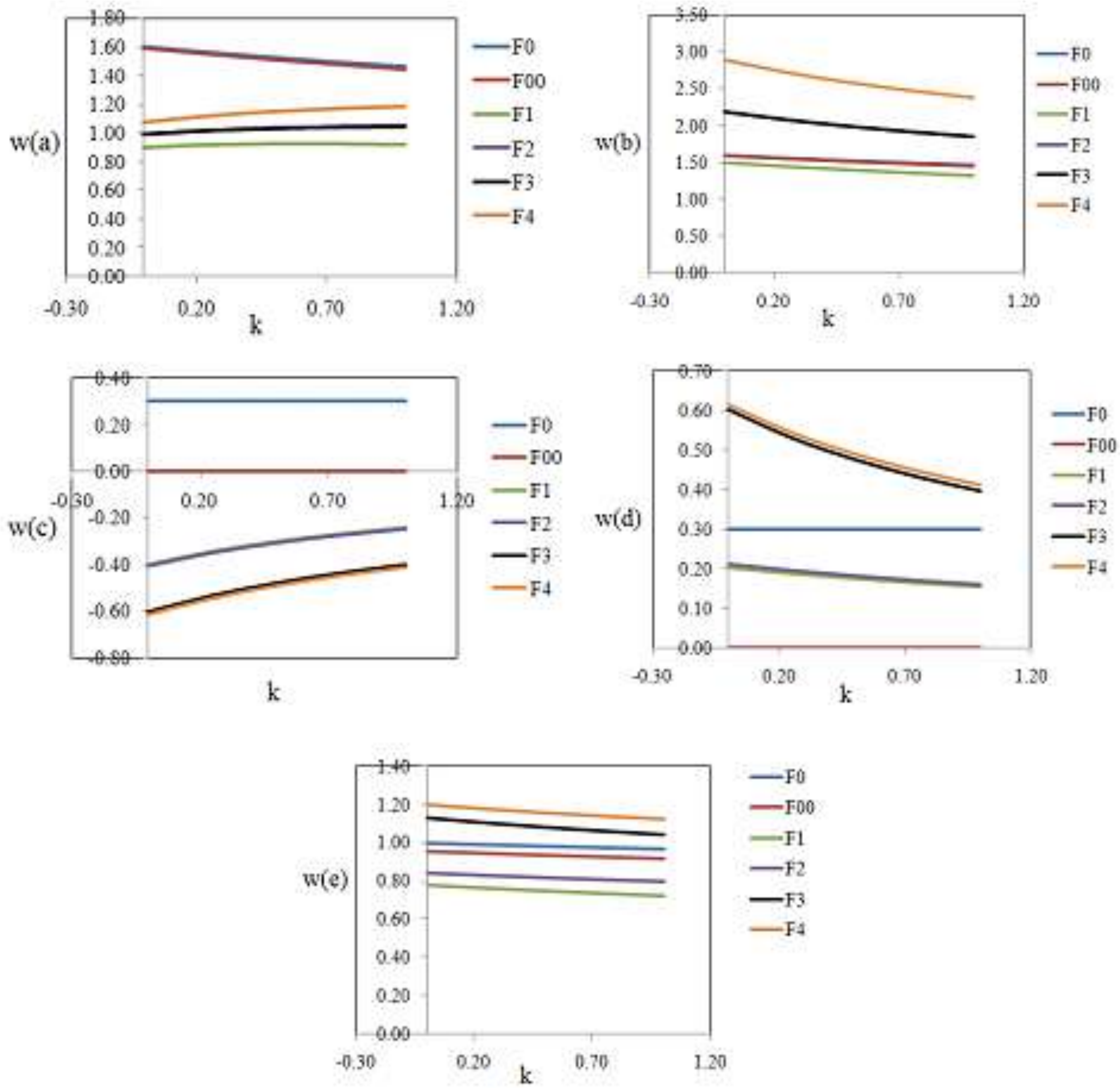

Figure 2. Profile curves for *a*, *b*, *c*, *d*, and *e* terms showing term weights (w) for several values of k.

An Illustrative Example without Relevance Information

The examples given in Figures 1 and 2 illustrate several scenarios wherein documents are marked as either relevant or non-relevant. However, early in this tutorial we mentioned that in the absence of relevance information the only information available is collection-wide incidence.

Revisiting Table 7, for R = 0, r = 0, and k = 0.5 the best match function F4 reduces to

$$F4 = \log\left(\frac{N-n+0.5}{n+0.5}\right) \tag{5}$$

corresponding to an IDFP with a 0.5-smoothing correction ($IDFP_{0.5}$). The correction essentially resets the scale of weights to avoid a mathematical error when N – n = 0. The error can also be avoided by simply adding a trailing term to IDFP before taking logs, like this

$$F4 = \log\left(\frac{N-n}{n} + l\right) \tag{6}$$

which we may refer to as IDFP *plus* ($IDFP_{+}$) where l works as a "lift" for the IDFP curve. (6) is easier to compute than (5) as it merely shifts the scale of RSJ weights. The technique is frequently used in power transformations problems. Furthermore, the use of logs for transforming data is just one type of Box-Cox Power Transformations (Garcia, 2016).

Table 10 and Figure 3 compare the corresponding IDF, IDFP, $IDFP_{0.5}$, and IDFP+ curves for a collection of N = 100 documents and several n values. Since a small k is used IDFP and $IDFP_{0.5}$ curves are superimposed, but can be resolved as we increase k.

Table 10. Calculation of RSJ weights without relevance information.

n	N	IDF	IDFP	$IDFP_{0.5}$	$IDFP_{+}$
10	100	1.00	0.95	0.94	0.98
20	100	0.70	0.60	0.59	0.65
30	100	0.52	0.37	0.36	0.45
40	100	0.40	0.18	0.17	0.30
50	100	0.30	0.00	0.00	0.18
60	100	0.22	-0.18	-0.17	0.07
70	100	0.15	-0.37	-0.36	-0.03
80	100	0.10	-0.60	-0.59	-0.12
90	100	0.05	-0.95	-0.94	-0.21

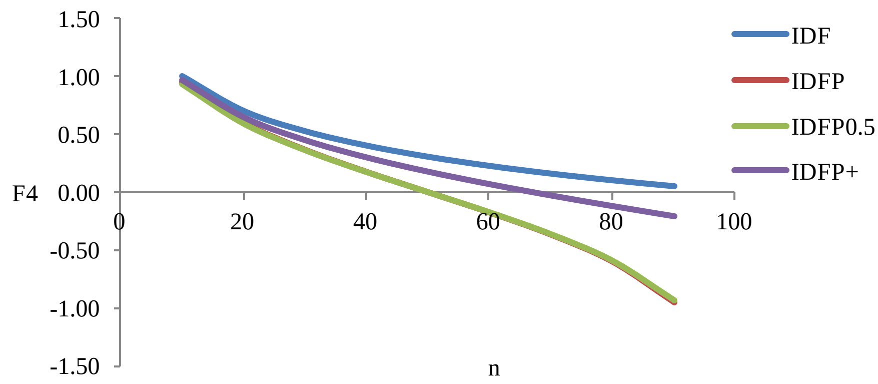

Figure 3. RSJ weight curves in the absence of relevance information.

What is really important from Figure 3 is that with (5) zero or negative weights are obtained for terms mentioned in half or more than half of a collection of documents; i.e., when $n \geq N/2$. These weights introduce unwanted complications.

Figure 3 shows that setting l = k = 0.5 yields an $IDFP_+$ curve more tolerant to negative weights. As expected, l = 1 reduces $IDFP_+$ to IDF, l > 1 avoids negative weights altogether, and setting l = 2 reduces (6) to $\log\left(1+\frac{N}{n}\right)$, an expression derived by Lee (2007).

Last but not least, we should mention that negative weights can be avoided by simply resetting them to zero, effectively treating the corresponding terms as stopwords. This can be justified on practical grounds. Consider a collection of documents about jobs. Assume that more than half of the documents mention the terms *job, résumé, career, education*, or *curriculum*. Searching with these terms is an inefficient way of discriminating between relevant and non-relevant documents from said collection. The terms actually behave as stopwords.

Conclusion

We have presented a tutorial on the Robertson-Spärck-Jones Probabilistic Model. The model computes global weights, known as RSJ weights, based on Independence Assumptions and Ordering Principles for probable relevance. The model subsumes IDF and IDFP as RSJ weights in the absence of relevance information.

Exercises

1. The following example is taken from *Information Retrieval:Algorithms and Heuristics* (Grossman & Frieder, 2004). Let Q be a query and d_1, d_2, and d_3 be documents of a collection. Thus, N = 3.

   Q: gold silver truck

   d_1: Shipment of gold damaged in a fire.

   d_2: Delivery of silver arrived in a silver truck.

   d_3: Shipment of gold arrived in a truck.

   Assuming term independence, rank documents in decreasing order of weights using F0 through F4 with (a) k = 0 and (b) k = 0.5. Compare weighting function results for each k and between k values.
2. Show that RSJ weights scored with function F4 can be expressed as follows

   $$w(t_i) = \log\left[\left(\frac{p_i}{(1-p_i)}\right)\left(\frac{(1-q_i)}{q_i}\right)\right]$$

   where p_i = P(document contains t_i|document is relevant) and q_i = P(document contains t_i|document is not relevant). That is, p_i is the probability that the document contains t_i provided that it is relevant and q_i is the probability that the document contains t_i provided that it is not relevant.
3. Figure 1 suggests the dampening power of k is almost insignificant at lower values of r/n . Why?
4. Why the slopes of some of the curves shown in Figure 2 are either positive or negative? You need to evaluate the weighting function derivatives respect to k (i.e., dF/dk) using the data given in Table 9.

References

Cox, D. R. (1970). *Analysis of Binary Data*. Methuen, 1970, London.

Garcia, E (2016). Local Term Weight Models from Power Transformations Development of BM25IR: A Best Match Model based on Inverse Regression. Retrieved from http://www.minerazzi.com/tutorials/bm25ir.pdf See also https://arxiv.org/abs/1608.01573

Grossman, D. A. & Frieder, O. (2004). *Information Retrieval: Algorithms and Heuristics*. Springer, Netherlands.

Lee, L. (2007). IDF Revisited: A Simple New Derivation within the Robertson-Spärck Jones Probabilistic Model. SIGIR 07, July 23-27,2007. Amsterdam, The Netherlands. ACM 978-1-59593-597-7/07/0007. Retrieved from http://www.cs.cornell.edu/home/llee/papers/idf.pdf

Robertson, S. E., & Spärck-Jones, K. (1976). Relevance weighting of search terms, Journal of the American Society for Information Science, Volume 27, 1976 pp. 129–146. Retrieved from http://www.staff.city.ac.uk/~sb317/papers/RSJ76.pdf

Robertson, S. E. (2004). Understanding Inverse Document Frequency: On theoretical arguments for IDF. Journal of Documentation, 60, 5, 503-520. Retrieved from http://nlp.cs.swarthmore.edu/~richardw/papers/robertson2004-understanding.pdf

Spärck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28, 1, 11-21. Retrieved from https://www.cl.cam.ac.uk/archive/ksj21/ksjdigipapers/jdoc72.pdf