Notes on Probabilistic Information Retrieval —ProbabilityRanking Principle から BM25 まで—

Yoshihiko Suhara

2012-04-27
Last update: 2012-05-04

概要

ブッチャー本 [1]8 章 Probabilistic retrieval を読んで感動し，そういえば確率的情報検索 (probablistic information retrieval; probabilistic IR) について書かれた日本語資料を見かけなかったのとたので，忘れないようにノートを作成．Probability Ranking Principle から出発し，BM25 の導出をゴールとする．途中で **(ちょっと現実的ではない) 仮定** を置いたり，**(ちょっと無茶な) 近似**をしつつも，1 つの道筋で導出されることを確認し，どういうモチベーションで導出されたかという過程を追いかける．基本的にブッチャー本 8 章のまんまの流れであるが，式変形の行間や原書で語られていないモチベーションなどを勝手に補足してみた．またブッチャー本で typo と思われる部分を適宜修正した．手を加えるうちに解説記事みたいになってきているが，著者は教科書の一章を読んだ程度の素人はだしであることに注意されたし．自分用のノートなので誤りがあるかもしれません．お気づきの点はご指摘頂けると幸いです．

1 Probability Ranking Principle

Probabilistic IR モデルを導入する前に，それらのスタート地点になっている Probability Ranking Principle (PRP) について述べる．文献 [1] (p.259) によると，PRP は

> If an IR system's response to each query is a ranking of the documents in the collection in order to decreasing probability of relevance, the overall effectiveness of the system to its user will be maximized.

というもの．平たくいえば，**ユーザのリクエストに対して，適合確率の降順に文書がランキングされた結果が最適な出力とする原理**である．また，初出と思われる Robertson [2] によると[*1] もう少し詳細に述べられている．

> The probability ranking principle (PRP): If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefullness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the

[*1] 真の初出は未調査．文献 [2] によると本稿での引用部分は W. S. Cooper, "The suboptimality of retrieval rankings based on probability of usefullness. (Private communication)" となっているので，文献に記述したのは文献 [2] が最初ではないかと思う．

> overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data.

ここでは relevance と限定されておらず，probability of usefullness と記述されている点，適合確率について“利用可能なデータを用いて可能な限り正確に推定された”という説明が付与されている点において文献 [1] の表現と異なる．

また Spärck Jones ら [3] では，

> If retrieved documents are ordered by decreasing probability of relevance on the data available, then the system's effectiveness is the best to be gotten for the data.

と記述されている．こちらの方が文献 [1] に比べて初出 [2] に近い表現になっている．細かい違いはあれど，これ以降気にしなくてよいので気にしないことにする．

実際には，検索結果に多様性を求める場合や，冗長な検索結果を不要という観点からは PRP を出発点とするのが必ずしも適切ではない場合もあることに注意する．

2 適合確率のモデル化

さて PRP をスタート地点とするのはよいが，**そもそも適合確率って何?** という疑問が起こる．Spärck Jones ら [3] はこの疑問を “Basic Question” と呼んで以下のように述べている．

> What is the probability that this document is relevant to this query?

これに対する回答はいくつもあるが，これより先は Lafferty ら [4] の方法に従い，(天下り的ではあるが) 適合確率を 3 つの確率変数で表現することにする: 文書 D, クエリ Q, ユーザによる二値の適合性評価 $R \in \{0, 1\}$. これを用いると適合確率を以下のとおり表現することができる[*2].

$$p(R = 1|D = d, Q = q). \tag{1}$$

簡単のため，$D = d$ を D，$Q = q$ を Q と表現すると式 (1) は

$$p(R = 1|D, Q) \tag{2}$$

と書ける．同様に r を $R = 1$，$\bar{r}$ を $R = 0$ とすると，

$$p(r|D, Q) = 1 - p(\bar{r}|D, Q) \tag{3}$$

と書ける．

ベイズの定理

$$p(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{4}$$

を式 (3) に適用すると，

$$p(r|D, Q) = \frac{p(D, Q|r)p(r)}{p(D, Q)} \tag{5}$$

[*2] 離散確率変数に対する確率関数には大文字 P を用いることが多いが，ここでは文献 [1] の記法に従い，p とする

と

$$p(\bar{r}|D,Q) = \frac{p(D,Q|\bar{r})p(\bar{r})}{p(D,Q)} \tag{6}$$

を得る．

ここで式を扱いやすくするためにやや天下り的に対数オッズ (log-odds) またはロジット (logit) を用いて表現する．対数オッズは

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{7}$$

で表現され，p が 0 から 1 に変化する際，対数オッズの値は $-\infty$ から ∞ に変化し，$p = 0.5$ の際，$\mathrm{logit}(p) = 0$ となる．また確率 p と q に対して $p > q$ のとき，かつそのときに限り $\mathrm{logit}(p) > \mathrm{logit}(q)$ が成り立つため，順序不変 (rank-equivalent, rank-preserving, order-preserving) である．

式 (2) に対して対数オッズを取り，式 (5), (6) と同じようにベイズの定理を用いると，

$$l \log \frac{p(r|D,Q)}{1-p(r|D,Q)} = \log \frac{p(r|D,Q)}{p(\bar{r}|D,Q)} \tag{8}$$

$$= \log \frac{p(D,Q|r)p(r)}{P(D,Q)} \cdot \frac{p(D,Q)}{p(D,Q|\bar{r})p(\bar{r})} \tag{9}$$

$$= \log \frac{p(D,Q|r)p(r)}{p(D,Q|\bar{r})p(\bar{r})}$$

このように，対数オッズを取ると $P(D,Q)$ を打ち消すことができる．

ここで $p(D,Q|R) = p(D|Q,R) \cdot p(Q|R)$ という条件付き確率の分解を式 (8) に適用し，再びベイズの定理を用いると，

$$\log \frac{p(D,Q|r)p(r)}{p(D,Q|\bar{r})p(\bar{r})} = \log \frac{p(D|Q,r)p(Q|r)p(r)}{p(D|Q,\bar{r})p(Q|\bar{r})p(\bar{r})} \tag{10}$$

$$= \log \frac{p(D|Q,r)p(r|Q)}{p(D|Q,\bar{r})p(\bar{r}|Q)} \tag{11}$$

$$= \log \frac{p(D|Q,r)}{p(D|Q,\bar{r})} + \log \frac{p(r|Q)}{p(\bar{r}|Q)} \tag{12}$$

を得る．ここで $\log \frac{p(r|Q)}{p(\bar{r}|Q)}$ は D に対して独立なので，文書をランキングする上では順位に影響を与えないので取り除くと，

$$\log \frac{p(D|Q,r)}{p(D|Q,\bar{r})} \tag{13}$$

を得る．この式 (13) は Probabilistic IR 確率的情報検索の心臓部分[*3]に当たる式で，これから先の解説は，ここをスタート地点とする．

3 Binary Independence Model

式 (13) では，クエリと適合性が与えられた下での文書の確率の対数オッズ比をスコアとして用いることを考えた．しかし，文書は多数の単語から成っており，“文書”という単位でモデル化するよりも“単語”という単位でモデル化した方が何かと都合がよい．いくつかの仮定を置くことにより，式 (13) を Binary Independence Model (BIM) と呼ばれるもう少し扱いやすい形に変形することを試みる．

[*3] 余談だが “This formula sits at the heart of the probabilistic retrieval mode.” ([1] p.261) と書かれている

再び天下り的に新たな記法を導入する．文書 D の確率変数を，各次元が語彙 $\mathcal{V}$ の各単語に対応する $D = \langle D_1, D_2, \dots \rangle$ とする．$D_i = 1$ は，i 次元目の単語が文書に出現することを表し，$D_i = 0$ は文書に出現しないことを表す．同様にクエリ Q の確率変数も $Q = \langle Q_1, Q_2, \dots \rangle$ で表現し，$Q_i = 1$ は，i 次元目の単語がクエリに出現することを表し，$Q_i = 0$ はクエリに出現しないことを表す．

これから BIM を導出するために 2 つの強い仮定を置く．1 つ目が independence assumption である:

> Assumption T: Given relevence, terms are statistically independent.

言い換えると，単語は適合しているかどうか R だけに依存しており，**確率変数 R の値が決まれば，単語の出現は互いに独立している**とする仮定である[*4]．もちろん，直観的に理解できるように，この仮定は実際には成り立たない．

この仮定を用いると，式 (13) は個々の単語に対応する確率変数に対する確率の積で表現できる:

$$p(D|Q, r) = \prod_{i=1}^{|\mathcal{V}|} p(D_i|Q, r) \tag{14}$$

$$p(D|Q, \bar{r}) = \prod_{i=1}^{|\mathcal{V}|} p(D_i|Q, \bar{r}). \tag{15}$$

すると式 (13) は,

$$\log \frac{p(D|Q, r)}{p(D|Q, \bar{r})} = \sum_{i=1}^{|\mathcal{V}|} \log \frac{p(D_i|Q, r)}{p(D_i|Q, \bar{r})} \tag{16}$$

と書ける．

2 つ目の仮定は

> Assumption Q: The presence of a term in a document depends on relevance only when that term is present in the query.

という仮定で，これは**クエリに出現しない単語は適合度に影響しない**というこれまた強烈な仮定である．

そのため，この仮定に基づくと $Q_i = 0$ の場合には $p(D_i|Q_i, r) = p(D_i|Q_i, \bar{r})$ となり，

$$\log \frac{p(D_i|Q, r)}{p(D_i|Q, \bar{r})} = 0 \tag{17}$$

となる．

すると式 (16) は，全単語に対する和ではなく，クエリに出現する単語の和に書き換えることができる．

$$\sum_{t \in q} \log \frac{p(D_t|r, Q_t)}{p(D_t|\bar{r}, Q_t)} \tag{18}$$

クエリに出現する単語を対象に和を取っているため，条件に出現する Q_t は必ず 1 となり，冗長であるため

$$\sum_{t \in q} \log \frac{p(D_t|r)}{p(D_t|\bar{r})} \tag{19}$$

[*4] 機械学習の分野でテキスト分類などで用いられるナイーブベイズ仮定 (Naive Bayes Assumption) と同じ仮定である．ナイーブベイズの場合は適合性 R の代わりにクラス C を用いた表現をする．

と書ける．

ここで確率変数 $D = d = \langle d_1, d_2, \dots \rangle$ のうち，単語 t に対応する位置の確率変数を $D_t = d_t$ という表現を用いると，

$$\sum_{t \in q} \log \frac{p(D_t = d_t|r)}{p(D_t = d_t|\bar{r})} \tag{20}$$

と書くことができる．ここで式 (20) において，クエリ全ての単語が出現しない，すなわち $D_t = 0 \ (\forall t \in q)$ の場合，値は定数となるため，この値を式 (20) から引いてもランキングには影響を与えない．そこで

$$\sum_{t \in q} \log \frac{p(D_t = d_t|r)}{p(D_t = d_t|\bar{r})} - \sum_{t \in q} \log \frac{p(D_t = 0|r)}{p(D_t = 0|\bar{r})} \tag{21}$$

を得る．

式 (21) の Σ の中身をクエリと文書両方に出現する単語 $t \in (q \cap d)$ と，クエリに出現して文書に出現しない単語 $t \in (q \backslash d)$ に分けると，確率変数 D_t の値が 1 または 0 のものでまとめることができるため，

$$\sum_{t \in (q \cap d)} \log \frac{p(D_t = 1|r)}{p(D_t = 1|\bar{r})} + \sum_{t \in (q \backslash d)} \log \frac{p(D_t = 0|r)}{p(D_t = 0|\bar{r})} - \sum_{t \in (q \cap d)} \log \frac{p(D_t = 0|r)}{p(D_t = 0|\bar{r})} - \sum_{t \in (q \backslash d)} \log \frac{p(D_t = 0|r)}{p(D_t = 0|\bar{r})} \tag{22}$$

を得る．これを整理して，

$$\sum_{t \in (q \cap d)} \log \frac{p(D_t = 1|r)p(D_t = 0|\bar{r})}{p(D_t = 1|\bar{r})p(D_t = 0|r)} - \sum_{t \in (q \backslash d)} \log \frac{p(D_t = 0|r)p(D_t = 0|\bar{r})}{p(D_t = 0|\bar{r})p(D_t = 0|r)} \tag{23}$$

第二項は打ち消し合って 0 になるため，

$$\sum_{t \in (q \cap d)} \log \frac{p(D_t = 1|r)p(D_t = 0|\bar{r})}{p(D_t = 1|\bar{r})p(D_t = 0|r)} \tag{24}$$

を得る．

式 (24) が仮定 T と仮定 Q の下での式 (13) の変形版で，BIM として知られる．

4 Robertson / Spärck Jones Weighting Formula と IDF との関係

少し寄り道をする．式 (24) を

$$\sum_{t \in (q \cap d)} w_t \tag{25}$$

と記述することにする．なお，これが **BM1** と呼ばれる手法である．

ここで $p_t = p(D_t = 1|r)$，$\bar{p}_t = p(D_t = 1|\bar{r})$ とすると，w_t は，

$$w_t = \log \frac{p_t(1 - \bar{p}_t)}{\bar{p}_t(1 - p_t)} \tag{26}$$

ここで $p(D_t = 0|r) = 1 - p(D_t = 1|r) = 1 - p_t$ と $p(D_t = 0|\bar{r}) = 1 - p(D_t = 1|\bar{r}) = 1 - \bar{p}_t$ を利用した．式 (26) は Robertson/Spärck Jones weighting formula と呼ばれる．

ちょっとまとめると，

- 適合文書に単語 t が出現する確率 (p_t)

- 適合文書に単語 t が出現しない確率 $(1-p_t)$
- 非適合文書に単語 t が出現する確率 $(\bar{p}_t)$
- 非適合文書に単語 t が出現しない確率 $(1-\bar{p}_t)$

という情報を用いてランキングしていることになる．

p_t と $\bar{p}_t$ の推定を考える．ここで N を検索対象群における総文書数，N_t を単語 t を含む文書数，N_r を入力クエリに対する適合文書の真の数，$N_{t,r}$ を単語 t を含む適合文書の真の数とすると，

$$p_t = \frac{N_{t,r}}{N_r} \tag{27}$$

$$\bar{p}_t = \frac{N_{t,\bar{r}}}{N_{\bar{r}}} = \frac{N_t - N_{t,r}}{N - N_r} \tag{28}$$

と推定することができる．これを式 (26) に代入すると，

$$w_t = \log \frac{N_{t,r}(N - N_t - N_r + N_{t,r})}{(N_r - N_{t,r})(N_t - N_{t,r})} \tag{29}$$

を得る．ただし，実際には全てのクエリ，文書に対して適合性評価を付与するのは不可能なので，N_r と $N_{t,r}$ の正確な値はわからないことに注意する．一方 N_t は，単語 t を含む文書を数えればよいので正確な値が求めることができる．

そこでを用いて N_r の代わりに評価済みのデータにおける適合文書数 n_r と，同じく適合文書のうち単語 t を含む数 n_r を用いると

$$w_t = \log \frac{n_{t,r}(N - N_t - n_r + n_{t,r})}{(n_r - n_{t,r})(N_t - n_{t,r})} \tag{30}$$

と書くことができる．ただし，評価済みデータの中に適合文書が含まれていない場合，$p_t = 0$ や $\bar{p}_t = 0$ となり $\log 0$ や $\log \infty$ となるのを避けるため，頻度のスムージングを行い，以下の計算式を得る:

$$w_t = \log \frac{(n_{t,r} + 0.5)(N - N_t - n_r + n_{t,r} + 0.5)}{(n_r - n_{t,r} + 0.5)(N_t - n_{t,r} + 0.5)}. \tag{31}$$

ここで $p_t = \frac{n_{t,r}+0.5}{n_r+1}$，$\bar{p}_t = \frac{N_t - n_{t,r}+0.5}{N-n_r+1}$ という加算スムージングは二項分布のパラメータ推定においてベータ分布を事前分布とした際の MAP 推定をしていることに等しい．

4.1 IDF との関係

ここで式 (29) と式 (31) を眺めると IDF と似ているような気がするので，IDF との関係を考察してみることにする．

まず，式 (26) の対数の中身を 2 つの項に分けて，

$$w_t = \log \frac{p_t}{1-p_t} + \log \frac{1-\bar{p}_t}{\bar{p}_t} \tag{32}$$

$$= \log \frac{N_{t,r}}{N_r - N_{t,r}} + \log \frac{N - N_r - N_t + N_{t,r}}{N_t - N_{t,r}} \tag{33}$$

と書ける．ここで N_r と $N_{t,r}$ が N や N_t に比べて極めて少ないと仮定すると，右辺第二項において $N_r = N_{t,r} = 0$ と近似することができるため，

$$w_t = \mathrm{logit}(p_t) + \log \frac{N - N_t}{N_t} \tag{34}$$

を得る．

Croft と Harper[6] は，$p_t = 0.5$ かつ N_t が N に比べて極めて少ない場合 (つまり $N_t = 0$ という近似を行う)，通常の IDF が式 (34) と一致すると述べている．また，Robertson[7] は

$$p_t = \frac{1}{1 + \frac{N - N_t}{N}} \tag{35}$$

の際に式 (34) が IDF である

$$w_t = \log \frac{N}{N_t} \tag{36}$$

と一致することを示した．

5 単語頻度の考慮

閑話休題．今までは単語が文書に出現するか否かをモデル化してきたが，今度は出現数をモデルに組み込むことを考える．途中 eliteness という概念や 2-poisson モデルというなんでここで出てくるのか一見ワケワカラン手法が出てくるが，BM25 で普段用いられている，k_1 パラメータを用いて単語頻度に対して対数カーブのような飽和曲線を描くような計算式を用いる元ネタとなっている[*5]．

再び天下り的に単語 t の出現数を表す確率変数を $F_t = f_t$ とすると，式 (24) は，

$$\sum_{t \in q} \log \frac{p(F_t = f_t|r)p(F_t = 0|\bar{r})}{p(F_t = f_t|\bar{r})p(F_t = 0|r)} \tag{37}$$

と書ける．これから述べる eliteness の概念を用いてこの式をごにょごにょすることにする．

5.1 eliteness の導入

Bookstein ら [8] は，単語とトピックの関係をモデル化を試み，Robertson ら [9] が *eliteness* と呼ぶ概念を導入した．

文献 [1](p.267) から引用すると，

> A document is said to be *elite* in term t when it is somehow "about" the topic associated with the term.

という，**あるトピックに関連する単語 t を多く含む場合，当該文書は単語 t に対してエリートである**という (正直わかりづらい) 概念である．ざっくり言いかえると，あるトピックについて記述された文書には，そのトピックに関連する単語群に対してエリートであり，それらの文書にはこれらの単語が多数出現すると考える．

エリートという概念で説明するよりも，トピックモデルの潜在変数の代わりに 2 値の隠れ変数 e を挟むとイメージするとかえってわかりやすい．

$$p(F_t = f_t|r) = p(F_t = f_t|e) \cdot p(e|r) + p(F_t = f_t|\bar{e}) \cdot p(\bar{e}|r) \tag{38}$$

$$p(F_t = f_t|\bar{r}) = p(F_t = f_t|e) \cdot p(e|\bar{r}) + p(F_t = f_t|\bar{e}) \cdot p(\bar{e}|\bar{r}) \tag{39}$$

[*5] BM25 は経験的に性能の良さが知られているので，BM25 を使う上ではこの妥当性を知る必要はない．遠回りになるが，背景や理論的裏付け (といっても多くの仮定の上に成り立っているのでなんとも言えないが...) を知りたい人向けの情報

ここでも $\sum_{t\in q}$ の仮定は使われており，e と書かれており，一見単語 t と独立しているように見えるが，実際には**クエリに含まれるある単語 t に対してエリートであるか否か**という確率変数であることに注意する．

これを式 (37) に代入すると，

$$\sum_{t\in q}\log\frac{(p(F_t=f_t|e)p(e|r)+p(F_t=f_t|\bar{e})p(\bar{e}|r))\cdot(p(F_t=0|e)p(e|\bar{r})+p(F_t=0|\bar{e})p(\bar{e}|\bar{r})))}{(p(F_t=f_t|e)p(e|\bar{r})+p(F_t=f_t|\bar{e})p(\bar{e}|\bar{r}))\cdot(p(F_t=0|e)p(e|r)+p(F_t=0|\bar{e})p(\bar{e}|r)))} \tag{40}$$

を得る．少しややこしく見えるが，隠れ変数 $E\in\{e,\bar{e}\}$ に対して全確率の公式を用いて展開しているだけである．

5.2 Bookstein's Two-Poisson Model

次に $p(F_t=f_t|r)$ をどうモデル化するかということを考える．Bookstein は，**ある単語 t に対して文書がエリートである場合の単語の出現頻度と，エリートではない場合の単語の出現頻度をそれぞれふたつのポアソン分布でモデル化する**ことを考えた．

ポアソン分布では単位時間における平均発生回数が μ のイベントにおいて，イベント発生回数を表す確率分布である．ポアソン分布では，各イベントの発生は独立して起こるものという仮定を置いている．単位時間を文書の単位長，発生回数を単語の出現回数と見なすことにより，単語の出現回数をポアソン分布で表現することができる．単位長において単語が平均 μ 回出現する場合，x 回出現する確率はポアソン分布の確率密度関数を用いて，

$$g(x,\mu)=\frac{e^{-\mu}\mu^x}{x!} \tag{41}$$

と求めることができる．

なお一度出現した単語は，文書において再び出現する可能性が高くなると考えられるため，ポアソン分布を用いたモデル化は強い仮定に基づいていることに注意する．また，ポアソン分布では単位あたりの平均出現回数をパラメータとする必要があるため，全ての文書の長さが等しいという仮定を置いていることにも注意する．後者の仮定は，**BM25 における文書長正規化のモチベーション**になっていることも加えて述べておく．

さて，エリートと非エリートに対応するポアソン分布のパラメータを用意する．平均出現数が多い場合にエリートであるため，$\mu_e>\mu_{\bar{e}}$ となるようなパラメータを考える．ここで $q=p(e|r)$，$\bar{q}=p(e|\bar{r})$ とすると，

$$p(F_t=f_t|r)=g(f_t,\mu_e)\cdot q+g(f_t,\mu_e)\cdot(1-q) \tag{42}$$

$$p(F_t=f_t|\bar{r})=g(f_t,\mu_{\bar{e}})\cdot\bar{q}+g(f_t,\mu_{\bar{e}})\cdot(1-\bar{q}) \tag{43}$$

となり，これを式 (37) に代入すると，

$$\sum_{t\in q}\log\frac{(g(f_t,\mu_e)q+g(f_t,\mu_{\bar{e}})(1-q))\cdot(g(0,\mu_e)\bar{q}+g(0,\mu_{\bar{e}})(1-\bar{q}))}{(g(f_t,\mu_e)\bar{q}+g(f_t,\mu_{\bar{e}})(1-\bar{q}))\cdot(g(0,\mu_e)q+g(0,\mu_{\bar{e}})(1-q))} \tag{44}$$

を得る．分子分母を $g(f_t,\mu_e)g(0,\mu_{\bar{e}})$ で割ると，

$$\sum_{t\in q}\log\frac{\left(q+\frac{g(f_t,\mu_{\bar{e}})}{g(f_t,\mu_e)}(1-q)\right)\left(\frac{g(0,\mu_e)}{g(0,\mu_{\bar{e}})}\bar{q}+(1-\bar{q}))\right)}{\left(\bar{q}+\frac{g(f_t,\mu_{\bar{e}})}{g(f_t,\mu_e)}(1-\bar{q})\right)\left(\frac{g(0,\mu_e)}{g(0,\mu_{\bar{e}})}q+(1-q))\right)} \tag{45}$$

となる．さて，ここで

$$\frac{g(f_t,\mu_{\bar{e}})}{g(f_t,\mu_e)}=\frac{e^{-\mu_{\bar{e}}}\mu_{\bar{e}}^{f_t}}{e^{-\mu_e}\mu_e^{f_t}}=e^{\mu_e-\mu_{\bar{e}}}\cdot\left(\frac{\mu_{\bar{e}}}{\mu_e}\right)^{f_t} \tag{46}$$

となる．$\mu_e > \mu_{\bar{e}}$ より，$f_t \to \infty$ のとき，$\left(\frac{\mu_{\bar{e}}}{\mu_e}\right)^{f_t} = 0$ となり，$\frac{g(f_t,\mu_{\bar{e}})}{g(f_t,\mu_e)} = 0$ となる．また，

$$\frac{g(0,\mu_{\bar{e}})}{g(0,\mu_e)} = e^{\mu_{\bar{e}}-\mu_e} \tag{47}$$

より，$f_t \to \infty$ のとき，式 (45) におけるある単語に対する重みは，

$$\log \frac{q(\bar{q}e^{\mu_{\bar{e}}-\mu_e} + (1-\bar{q}))}{\bar{q}(qe^{\mu_{\bar{e}}-\mu_e} + (1-q))} \tag{48}$$

という (単語 t 毎に定められた) 定数に漸近することがわかる．

また，$e^{\mu_{\bar{e}}-\mu_e}$ が小さいと仮定すると，重みは

$$\log \frac{q(1-\bar{q})}{\bar{q}(1-q)} \tag{49}$$

に近似することができる．この形は Robertson/Spärck Jones weighting formula と似ていることに気付く．$q = p(e|r)$ を $p_t = p(D_t|r)$ で近似していると考えるれば，式 (26) は式 (49) の近似と捉えることができる．

5.3 Two-Poisson Model の近似

しかし，e は隠れ変数であり，$p(e|r)$ を直接推定することは難しい．そのため，何かしらの方法で近似することを考える．Robertson と Walker [9] は，two-Poisson モデルの近似として以下の式を提案した:

$$\sum_{t\in q} \frac{f_{t,d}(k_1+1)}{k_1+f_{t,d}} \cdot w_t. \tag{50}$$

ここで $f_{t,d} \to \infty$ のとき，重みは $(k_1+1)w_t$ という定数に漸近するため，式 (45) と同じ性質を保持していることがわかる．経験的には $1 \le k_1 < 2$ の値が用いられ，全ての単語に対して等しい k_1 が設定される．

余談だが，TF-IDF の文脈でも TF に対して対数を取る場合がある (i.e., TF $= \log(f_{t,d}+1)$)．これもある意味で，式 (50) と同様の近似を行っていると見なすことができる．

5.4 クエリ内の単語頻度

今まで文書内における単語出現頻度について論じてきたが，式 (50) はクエリ内の単語頻度に対しても拡張することができる．文書を潜在的なクエリと見なすことにより，クエリと文書を対称的に扱うことができるため，クエリ内の単語頻度を以下の形式で考慮することができる:

$$\frac{q_t(k_3+1)}{k_3+q_t}. \tag{51}$$

ここで k_3 は k_1 と似たようなあらかじめ定められたパラメータである．これを式 (50) に加味すると，

$$\sum_{t\in q} \frac{q_t(k_3+1)}{k_3+q_t} \cdot \frac{f_{t,d}(k_1+1)}{k_1+f_{t,d}} \cdot w_t \tag{52}$$

を得る．

ただし，クエリ長が長くなると，クエリ内に複数出現する単語が，文書内の出現よりも強く効いてしまうため，k_3 の値を k_1 よりもずっと大きくする必要があり，しばしば $k_3 = \infty$ という値も用いられる．$k_3 = \infty$ に

設定すると，

$$\sum_{t \in q} q_t \cdot \frac{f_{t,d}(k_1+1)}{k_1 + f_{t,d}} \cdot w_t \tag{53}$$

という形式を得る．この手法は BM15 とも呼ばれる[*6]．

6 文書長を考慮した手法: BM11 と BM25

ここで，ようやくゴールである BM25 の話にたどり着く．Two-Poisson モデルでは暗黙的に全ての文書が等しい長さであるという現実的ではないな仮定を置いていたので，これを解消することを試みる．

簡単な方法として単語頻度を文書長でスケーリングするという方法が考えられる．

$$f'_{t,d} = f_{t,d} \cdot \frac{l_{avg}}{l_d} \tag{54}$$

$$\sum_{t \in q} q_t \cdot \frac{f'_{t,d}(k_1+1)}{k_1 + f'_{t,d}} \cdot w_t \tag{55}$$

展開すると，

$$\sum_{t \in q} q_t \cdot \frac{f_{t,d}(l_{avg}/l_d)(k_1+1)}{k_1 + f_{t,d}(l_{avg}/l_d)} \cdot w_t = \sum_{t \in q} q_t \cdot \frac{f_{t,d}(k_1+1)}{k_1(l_d/l_{avg}) + f_{t,d}} \cdot w_t \tag{56}$$

を得る．これは BM11 とも呼ばれる．

Robertson[9] は BM11 と BM15 を混ぜ，以下の式を提案した．これが BM25 である:

$$\sum_{t \in q} q_t \cdot \frac{f_{t,d}(k_1+1)}{k_1((1-b) + b(l_d/l_{avg})) + f_{t,d}} \cdot w_t. \tag{57}$$

ここで新たなパラメータ $0 \leq b \leq 1$ を用いる．$b = 0$ のときは式 (53) と等しくなり，$b = 1$ のときは式 (56) と等しくなる．これは分母の $(1-b) + b(l_d/l_{avg})$ を眺めると，全ての文書を平均文書長とする方法と，実際の文書長を平均文書長で正規化する方法の線形和であることからわかる (i.e., $(1-b)(l_{avg}/l_{avg} + b(l_d/l_{avg})$. 検索対象群によっても最適なパラメータは異なるが，経験的には $b = 0.75$ あたりの値が用いられる．

7 まとめ

PRP からスタートして，なんとか BM25 の導出までこぎつけることができた．簡単な流れをおさらいしてみる．

1. PRP をスタート地点とする．
2. 適合確率のモデル化は一意ではないが (Basic Question) ，Lafferty の方法に基づいて確率変数 D, R, Q を用いて定式化する．
3. 適合性が与えられた下で単語の出現は独立，という binary independence assumption と，クエリに出てくる単語だけ考慮する仮定に基づいて BIM の計算式を得る．
4. (寄り道) Robertson/Spärck-Jones weighting formula と IDF との関係についてちょっと解説.

[*6] 文献 [10] を眺めるとどうも BM15 の定義が違う気がするが，未検証．あとでちゃんと調べる．

5. 単語の出現だけでなく，頻度も考慮するモデルを考える．eliteness や two-Poisson モデルが出てくる．
6. eliteness は明示的に扱うことが難しいので，近似することにする．
7. 同じノリでついでにクエリ内の単語頻度も考慮することにする．ただ，クエリ内の単語頻度が効きすぎるので，単純にクエリ内単語頻度をかける計算式にする (BM15).
8. two-Poisson モデルでは全ての文書が同じ長さであるという仮定を置いていたので，文書長正規化を行うことにより，この問題を解消する (BM11).
9. 文書長正規化を完全に効かせるよりかは，それなりに効かせたくらいの方がよさそうなので，新たなパラメータ b を用意して BM11 と BM15 と混ぜた手法を提案する (BM25).

本当は eliteness や two-Poisson モデルのあたりでご飯が 3 杯いける程度の話があるのだろうが，著者のモチベーションと能力の問題でそれはまたいつか.

参考文献

[1] Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack, “Information Retrieval: Implementing and Evaluating Search Engines”, MIT Press, 2010.

[2] S. E. Robertson, “The probability ranking principle in IR”, Journal of Documentation, vol.33, pp.294–304, 1977.

[3] K. Spärck Jones, S. Walker, and S.E. Robertson, “A probabilistic model of information retrieval: Development and comparative experiments – Part 1”, Information Processing & Management, vol.36(6), pp.779–808, 2000.

[4] J. Lafferty and C. Zhai, “Probabilistic relevance models based on document and query generation”, Language Modeling for Information Retrieval, Kluwer International Series on Information Retrieval, Vol.13, 2003.

[5] S. E. Robertson and K. Spärck Jones, “Relevance weighting of search terms”, Journal of the American Society for Information Science, vol.27, pp.129-146, 1976.

[6] W. B. Croft, and D. J. Harper, “Using probabilistic models of documents retrieval without relevance information”, Journal of Documentations, vol.35, pp.285–296, 1979.

[7] S. E. Robertson, and S. Walker, “On relevance weights with little relevance information”, In Proc. SIGIR ’97, pp.16–24, 1997.

[8] A. Bookstein, and D. Kraft, “Probabilistic models for automatic indexing”, Jounal of the American Society for Information Science, vol.25(5), pp.312–319, 1974.

[9] S. E. Robertson, and S. Walker, “Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval”, In Proc. SIGIR ’04, pp.232–241, 1994.

[10] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, “Okapi at TREC-2”, In Proc. TREC-2, pp.21–34, 1993.