Lecture 11: Classical Probabilistic IR: 2-Poisson model

William Webber (william@williamwebber.com)

COMP90042, 2014, Semester 1, Lecture 10
What we’ll learn in this lecture

Non-binary probabilistic models for IR
- Two-Poisson model
- BM25
Binary independence model

- Binary independence uses term occurrence 0, 1
- Models $p_t^{\{1\}} = P(d_t = 1|R, q)$ as Bernoulli RV, with param $p$
- $p$ estimated as prop of rel docs that $t$ occurs in.
- Similarly $u_t^{\{1\}} = P(d_t = 1|\bar{R}, q)$, param $u$
- $u$ estimated as prop of irrel docs that $t$ occurs in.

Weight $w_t$ of query term $t$ occurring in document $d$ is then:

$$w_t^{\{1\}} = \log \frac{p_t^{\{1\}}(1 - u_t^{\{1\}})}{u_t^{\{1\}}(1 - p_t^{\{1\}})}$$

(1)

Note that $1 - p_t^{\{1\}}$, $1 - u_t^{\{1\}}$ terms are for documents where query terms do not occur (see working from last lecture)
\textit{n-ary frequency}

Represent document as vector of term frequencies:

\[
\vec{d} = \langle d_1, \ldots, d_{|T|} \rangle, \quad d_i \in \{0, 1, 2, \ldots\}
\]

Then an equivalent \textit{n-ary} expression for Equation 1 is\footnote{Robertson and Walker, “Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval”, \textit{SIGIR}, 1994.}

\[
w_{tf} = \log \frac{p_{tf} u_0}{u_{tf} p_0}
\]

where

\[
\begin{align*}
    p_{tf} &= P(f_{d,t} = f|R,q) ;
    u_{tf} = P(f_{d,t} = f|R,q) , \quad f \in \{1, 2, \ldots\} \\
    p_0 &= P(f_{d,t} = 0|R,q) ;
    u_0 = P(f_{d,t} = 0|R,q)
\end{align*}
\]

\textbf{NOTE:} \( p_0 \neq (1 - p_{tf}) \); \( p_0 \) models non-occurrence, not complement of \( p_{tf} \)
Modelling $f_{d,t}$

- We need some model of:

$$p_{tf} = P(f_{d,t} = f|R, q) \quad (3)$$

and $u_{tf}$, $p_0$, $u_0$ as probability distributions

- that is, of $f_{d,t}$ as a random variable over $\{0, 1, 2, \ldots\}$

- Simplest suitable distribution is Poisson
  - Simple because it only requires us to estimate one parameter (like Bernoulli)
The Poisson process

Poisson process

A process in which events occur over time(-like dimension) independently and at random, e.g.:

- arrival of radioactive particles at Geiger counters
- emails to mail server
- failure of electronic components

More formally:

- Rate of arrivals $\lambda$ is constant over time
The Poisson process

Poisson process

A process in which events occur over time(-like dimension) independently and at random, e.g.:

- arrival of radioactive particles at Geiger counters
- emails to mail server
- failure of electronic components

More formally:

- Rate of arrivals $\lambda$ is constant over time
- Expected arrivals in interval $u$ is $\lambda u$
- Number of arrivals in disjoint intervals independent
A random variable $X$ has Poisson distribution with param $\lambda$ if:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{for } k = 0, 1, 2, \ldots$$

- $X$ is number of arrivals in unit interval of a Poisson process.
- $\lambda$ estimated as observed average arrivals
The Poisson Model

- Term frequency can be modelled as a Poisson process
- Assumes that terms occur “randomly” in documents
- ... around some common rate

One-Poisson Model

\[ P(f_{d,t}) \sim \frac{\lambda^k}{k!} e^{-\lambda} \]

\[ \hat{\lambda} = \frac{c_t}{N} \]

where \( c_t \) is collection frequency of \( t \) (i.e. total occurrences of \( t \), not just number of documents occurring in; \( c_t \geq f_t \)).

- In practice:
  - One-Poisson model reasonable fit for content-less words
  - But poor fit for content-bearing words (higher \( f_{d,t} \) more likely than Poisson model predicts)
The One-Poisson model

Empirically, one-Poisson fits content-less words ok
But poor fit for content-ful words
  More frequent high $f_{d,t}$ than expected$^2$

---

---

Table 1. Frequency Distributions for 19 Word Types and Expected Frequencies Assuming a Poisson Distribution with $\lambda = 53/650$

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Word Type</th>
<th>Number of Documents Containing $k$ Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>51</td>
<td>act</td>
<td>608</td>
</tr>
<tr>
<td>51</td>
<td>actions</td>
<td>617</td>
</tr>
<tr>
<td>54</td>
<td>attitude</td>
<td>610</td>
</tr>
<tr>
<td>52</td>
<td>based</td>
<td>600</td>
</tr>
<tr>
<td>53</td>
<td>body</td>
<td>605</td>
</tr>
<tr>
<td>52</td>
<td>castration</td>
<td>617</td>
</tr>
<tr>
<td>55</td>
<td>cathexis</td>
<td>619</td>
</tr>
<tr>
<td>51</td>
<td>comic</td>
<td>642</td>
</tr>
<tr>
<td>53</td>
<td>concerned</td>
<td>601</td>
</tr>
<tr>
<td>53</td>
<td>conditions</td>
<td>604</td>
</tr>
<tr>
<td>55</td>
<td>consists</td>
<td>602</td>
</tr>
<tr>
<td>53</td>
<td>factor</td>
<td>609</td>
</tr>
<tr>
<td>52</td>
<td>factors</td>
<td>611</td>
</tr>
<tr>
<td>55</td>
<td>feeling</td>
<td>613</td>
</tr>
<tr>
<td>52</td>
<td>find</td>
<td>602</td>
</tr>
<tr>
<td>54</td>
<td>following</td>
<td>604</td>
</tr>
<tr>
<td>51</td>
<td>force</td>
<td>603</td>
</tr>
<tr>
<td>51</td>
<td>forces</td>
<td>609</td>
</tr>
<tr>
<td>52</td>
<td>forgetting</td>
<td>629</td>
</tr>
<tr>
<td>53</td>
<td>expected, assuming Poisson distribution</td>
<td>599</td>
</tr>
</tbody>
</table>

---

Two-Poisson Model

Suggests fitting with two Poisson distributions:

Elite dist $a_{tf}$ for docs “about” concept represented by term.

Non-elite dist $n_{tf}$ for docs not “about” concept

Model $a_{tf} = P(f_{d,t}|E)$, $n_{tf} = P(f_{d,t}|\bar{E})$ as Poisson distributions with different rates:

\begin{align*}
    a_{tf} & \sim \frac{\lambda^k}{k!} e^{-\lambda} \quad (5) \\
    n_{tf} & \sim \frac{\mu^k}{k!} e^{-\mu} \quad (6)
\end{align*}

($\lambda > \mu$). Then distribution of $f_{d,t}$ given by:

\begin{equation}
P(f_{d,t} = f) = \pi \frac{\lambda^k}{k!} e^{-\lambda} + (1 - \pi) \frac{\mu^k}{k!} e^{-\mu} \quad (7)
\end{equation}

where $\pi$ is probability that document is elite. This can be made to fit data ok.
Eliteness and relevance

- Eliteness is not the same thing as relevance
- Document can be elite but not relevant, relevant but not elite
- But term frequency, conditioned on eliteness, is independent of relevance
- Therefore:

\begin{align*}
P(f_{d,t} = f | R) &= P(f | E)P(E | R) + P(f | \neg E)P(\neg E | R) \quad (8) \\
P(f_{d,t} = f | \neg R) &= P(f | E)P(E | \neg R) + P(f | \neg E)P(\neg E | \neg R) \quad (9)
\end{align*}
Expanding the Two-Poisson Model

Writing:

\[ p' = P(E|R) ; \quad q' = P(E|\bar{R}) \tag{10} \]

we can then expand Equation 2:

\[ w_{tf} = \log \frac{p_{tf} u_0}{u_{tf} p_0} \tag{11} \]

with Equations 8 and 9 as\(^3\):

\[ w_{tf} = \log \frac{(p' \lambda^f e^{-\lambda} + (1 - p') \mu^f e^{-\mu})(q' e^{-\lambda} + (1 - q') e^{-\mu})}{(q' \lambda^f e^{-\lambda} + (1 - q') \mu^f e^{-\mu})(p' e^{-\lambda} + (1 - p') e^{-\mu})} \]

\(^3\)Robertson and Walker, 1994
Estimating the Two-Poisson

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

(12)

Apparently going backwards:

- Now have four or five parameters to estimate per term
- \( p' = P(E|R) \) can’t be estimated, even with rel judgments
  - Would have to also judge “eliteness”
Approximating the Two-Poisson

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

(13)

At this point, Robertson and Walker (1994) throw up their hands and suggest approximating the “shape” of Equation 13:

1. Zero for \( tf = 0 \)
2. Increases monotonically with \( tf \)
3. To asymptotic maximum
4. Of Equation 1-like form \( \log \frac{p'(1-q')}{q'(1-p')} \)

From this, they suggest:

\[ w_{tf} = \frac{tf}{k_1 + tf} \cdot w_t^{\{1\}} \]  

(14)

for some tunable constant \( k_1 \), and recalling that \( w_t^{\{1\}} \) simplifies to IDF if we set \( p_t \) to 0.5.
Robertson and collaborators developed series weight functions:

\[
w = 1 \quad \text{(BM0)}
\]

\[
w_t^{1} = \log \frac{N - f_t + 0.5}{f_t + 0.5} \times \frac{f_{q,t}}{k_3 + f_{q,t}} \quad \text{(BM1)}
\]

If \( k_3 = 0 \), a slight variant on IDF. Behaves strangely if \( f_t > N/2 \).

\[
w_{15} = \frac{f_{d,t}}{k_1 + f_{d,t}} \times w_t^{1} + k_2 \times |q| \frac{\overline{|d|} - |d|}{|d| + |d|} \quad \text{(BM15)}
\]

Robertson and Walker (1994), with doc length and qry freq.

\[
w_{11} = \frac{f_{d,t}}{\frac{k_1 \times |d|}{|d|} + f_{d,t}} \times w_t^{1} + k_2 \times |q| \frac{\overline{|d|} - |d|}{|d| + |d|} \quad \text{(BM11)}
\]

Same as BM15 except \( f_{d,t} \) downweighted by document length.
BM25

\[ w_{25} = \log \frac{N - f_t + 0.5}{f_t + 0.5} \times \frac{(k_1 + 1)f_{d,t}}{k_1((1 - b) + \frac{b|d|}{|d|}) + f_{d,t}} \times \frac{(k_3 + 1)f_{q,t}}{k_3 + f_{q,t}} \]  

(BM25)

- BM25 combines aspects of B11 and B15
- \( k_1, b, \) and \( k_3 \) need to be tuned (\( k_3 \) only for very long queries).
  - \( k_1 \approx 1.5 \) and \( b \approx 0.75 \) common defaults.
- BM25 highly effective, most widely used weighting in IR
- Has TF, IDF, and document length components
- But only loosely inspired by probabilistic model
What have we achieved?

Pros
- Started from plausible probabilistic model of term distribution
- Shown how it can be made to fit something like TF*IDF
- Providing a probabilistic justification TF*IDF-like approaches

Cons
- Directly trying to estimate $P(f_{dt}|R)$ not practicable in retrieval (too many parameters, not enough evidence)
- Such approaches end up as ad-hoc as geometric model
- Progress requires letting query tell us what relevance looks like
- This the approach of language models
Probabilistic models promise to directly estimate (monotonic function of) $P(R|d, q)$

Classical models attempt to build upon collection statistics (e.g. $P(d_t|R, q) = $ proportion of relevant documents containing $t$.)

But lack of evidence at retrieval time forces very rough approximations.

Effective weighting schemes like BM25 are at best “inspired” by probabilistic ideas.
Looking back and forward

Forward

- Braver steps are required to make probabilistic models practical
- In particular, query must tell us more about relevance
- Language models attempt to implement this
Further reading


- Robertson and Waller, “Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval”, \textit{SIGIR}, 1994 (how to go from 2-Poisson model to something implementable like BM25).

