On the Generalisation and Logical Implementation of Retrieval Models

Barcelona IR Seminar March 2006
Thomas Roelleke

Motivation

What is an IR model?

Try a mathematical definition: An IR model is a function

\[ RSV : D \times Q \rightarrow R \]

\( D \): Set of documents
\( Q \): Set of queries
\( R \): Set of real numbers
Outline

- Part 1: General matrix framework and TF-IDF
- Part 2: Probabilistic models

Outline

- What is IR? Just some matrix/vector algebra?
- Notation - Notation - Notation
- TF-IDF
- Binary independent retrieval model
- Language modelling
- Implementation of IR models in probabilistic data models
- Conclusion
What is IR? Just a matrix?

<table>
<thead>
<tr>
<th></th>
<th>sailing</th>
<th>boats</th>
<th>east</th>
<th>coast</th>
<th>$n_T(d, c)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>doc2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>doc3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>doc4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>doc5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>(n_D(t, c))</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**DT matrix:** \(N_D(c) \times N_T(c)\) matrix.

Collection space, content representation, dimensions: \(D\) and \(T\).

Notation - Notation - Notation

Motivation: A consistent and dual notation:

- \(n_D(t, c)\): Number of documents in which term \(t\) occurs in collection \(c\)
- \(N_D(c)\): Number of documents in \(c\)
- Replace set \(D\) by set \(L\)

- \(n_L(t, c)\): Number of locations at which term \(t\) occurs in collection \(c\)
- \(N_L(c)\): Number of locations in \(c\)
**Notation**

<table>
<thead>
<tr>
<th>Replace $c$ by $d$: document space</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_L(t, d)$</td>
</tr>
<tr>
<td>$N_L(d)$</td>
</tr>
</tbody>
</table>

See "A general matrix framework for IR", IPM 2006

TF-IDF, binary independent retrieval, language modelling, Poisson model, divergence from randomness: we are ready.

---

**TF-IDF**

Term frequency TF: $tf(t, d)$: How "representative" is $t$ for $d$?

What about:

$$tf(t, d) := \frac{n_L(t, d)}{N_L(d)}$$

Ok, we know there is better:

$$tf(t, d) := \frac{n_L(t, d)}{K + n_L(t, d)}$$

$K$: A constant for all $t$, might depend on $d$ and $c$.

Hm, term frequency is actually location (token) frequency.
**TF-IDF**

Inverse document frequency IDF: Usually this:

\[ idf(t) := - \log \frac{n_D(t,c)}{N_D(c)} \]

There we go:

\[ RSV(d,q) := \sum_t t f(t,d) \cdot idf(t) \]

Still a competitive baseline, after so many years. Add a document length normalisation, and you are close to the top-performing BM25.

---

**What tops TF-IDF?**

- **Anchor-text**: Add to document, boosts TF.
- **Incoming links**: Who believes in history-biased authorities?
IS THERE A FRAMEWORK THAT ALLOWS FOR
MODELLING, EXPLAINING, INVESTIGATING AND COMPARING
TF-IDF AND ADD-ONS AND ALTERNATIVES?
Candidates: Matrix/vector algebra. Logic. Probability theory. Information theory?

A bit more of matrix framework

The starting point for CONTENT representation: Document-Term matrix: $D_T$.

Let’s multiply each matrix with its transposed matrix.

$DD = D_T \times D_T^T$: What is in $DD$?

Number of common/shared terms: Document similarity.

$TT = D_T^T \times D_T$: What is in $TT$?

Number of common/shared documents: Term similarity.
More matrices? Yes.

The starting point for STRUCTURE representation: Parent-Child matrix: $PC$.

Let’s play with dualities: Use $PC$ rather than $DT$, and phrase accordingly:

$PP = PC \times PC^T$: What is in $PP$?

Number of common children: Parent similarity.

$CC = PC^T \times PC$: What is in $CC$?

Number of common parents: Children similarity.

More matrices? Yes!

The starting point for CONTENT representation: Location-Term matrix: $LT$

Slide 14

There is a $LT$ for each document. $LT_d$

There is a $DT$ for each collection: $DT_c$

Dito for Parent-Child matrix: $PC_c$

One more: $PC_d$

MORE MATRICES? Yes, but let’s do TF-IDF next
TF-IDF in matrix framework?

Let’s count documents in which a term occurs:

\[ n_D = D^T \times DT \]

\[ n_D = (4, 3, 2, 1) \quad D^T = (1, 1, ..., 1) \quad DT \text{ as before} \]

Let’s divide by \( N_D(c) = 5 \). Then apply \(-\log\).

\[ idf = (\log \frac{4}{5}, ..., \log \frac{1}{5}) \]

Good start. We get term frequency with the same deal:

\[ n_L(\cdot, d) = L^T_d \times LT_d \]

Compare to \( n_D(\cdot, c) = D^T_c \times DT_c \)

How to get IDF in?

- \( n_L(\cdot, d) \): Term frequencies for each document
- \( idf(\cdot) \): inverse document frequency

Remember the vector space model?

\[ DT \cdot \vec{q} \]

Ok, that’s coordination level match.

Better use \( n_L(\cdot, d) \) vectors, the term frequencies:

\[
RSV = \begin{bmatrix}
\text{sailing boats east coast} \\
\text{d1} & 2 & 1 & 0 & 0 \\
\text{d2} & 0 & 3 & 1 & 0
\end{bmatrix} \cdot \vec{q}
\]
The trick with the diagonal

Remember?  \[ RSV = DT \cdot G \cdot \vec{q} \]
What about? \[ RSV = DT \cdot IDF \cdot \vec{q} \]


\[ IDF = diag(idf) \]

\[
IDF = \begin{bmatrix}
idf(sailing) & 0 & 0 & 0 \\
0 & idf(boats) & 0 & 0 \\
0 & 0 & idf(east) & 0 \\
0 & 0 & 0 & idf(coast)
\end{bmatrix}
\]

Pause

TF-IDF done
We used \( L_d^T \times L_d \) vectors for term frequencies
We used \( D_c^T \times D_c \) vector for document frequencies
We formed a diagonal matrix of idf values

Dual operations? Just to mention one: \( P_d^T \times P_C_d \)
Many more: inverse parent frequency, etc.

Eigenvalues of \( DD_C, TT_C, PP_C, CC_C \): Interesting.
\( TT_C \) Eigenvector: The document/query reflecting term co-occurrence
More matrices? Yes!!

Document-Assessor matrix $DA_q$

$AA_q$: Assessor similarity

Express precision/recall in general matrix framework

Eigenvector of $AA_q$: ...

Binary independent retrieval model

Start with ranking criteria:

$O(r|d, q) = \frac{P(r|d, q)}{P(\bar{r}|d, q)}$

Brings you to

$P(r, d, q) = P(d|q, r) \cdot P(q, r)$

After some arithmetic exercise and assumptions:

$RSV(d, q) := \sum_{t \in d \cap q} \log \frac{P(t|r) \cdot P(t|c)}{P(t|r) \cdot P(t|c)}$
**Binary independent retrieval model**

Rewrite:
\[
RSV(d, q) := \sum_{t \in d \cap q} -idf(t|r) - idf(\bar{t}|c) + idf(\bar{t}|r) + idf(t|c)
\]

So what?
Can be expressed in general matrix framework.

\(idf(t, c) - idf(t, r)\): Relevance information DECREASES the discriminativeness of a term.

For a term that occurs in many relevant docs: \(idf(t, r) \approx 0\).

Vries/Roelleke:SIGIR:2005

---

**Language modelling**

Start with
\[
P(r, d, q) = P(q|d, r) \cdot P(d, r)
\]

After some arithmetics:
\[
RSV(d, q) := \sum_{t \in q} \log (\lambda \cdot P(t|c) + (1 - \lambda) \cdot P(t|d))
\]

Can be expressed in general matrix framework.
Slide 23

**Probabilistic Logical Implementation**

```prolog
tf(T,D) :- ...
idf(T) :- ...
retrieve(D) :-
  tf(T,D) & idf(T) & query(T)
```

**HySpirit/Apriorie framework**: components for describing the required probability estimations.

`SELECT * FROM ...`

`ASSUMPTION IDF`

`EVIDENCE KEY (...)`

Needs notion of "probability of being informative": SIGIR:2003

Slide 24

**Summary and Conclusion**

- IR models "modelled" in general matrix framework: collection, document, and query spaces
- Supports the application and investigation of dualities
- Supports the aggregation of parameters: content + structure
- Related to the probabilistic logical implementation of IR models
- What for? Make system development more productive
Readings


$P(d \rightarrow q)$ as general framework

Wong/Yao: TOIS: 1995

Interpretations of $P(d \rightarrow q)$

Fuhr/Roelleke: TOIS: 1997

Probabilistic relational algebra


General matrix framework

Rijsbergen: 2005:

Vector algebra/spaces: That’s IR’s geometry