Slide 1

On the Generalisation and Logical Implementation of Retrieval Models

Barcelona IR Seminar March 2006

Thomas Roelleke

Slide 2

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**

Slide 3

Outline

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

Slide 4

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**

Slide 5

What is IR? Just a matrix?

	sailing	boats	east	coast	$n_T(d,c)$
doc1	**1**	**1**			**2**
doc2		**1**	**1**		**2**
doc3	**1**	**1**			**2**
doc4	**1**				**1**
doc5	**1**		**1**	**1**	**3**
$n_D(t,c)$	**4**	**3**	**2**	**1**	

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

Collection space, content representation, dimensions: D **and** T**.**

Slide 6

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

Slide 7

Notation - Notation - Notation

Replace c by d: document space	
$n_L(t, d)$	Number of locations at which term t occurs in collection d
$N_L(d)$	Number of locations in d

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

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

Slide 8

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.

Slide 9

TF-IDF

Inverse document frequency IDF: Usually this:

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

There we go:

$$RSV(d,q) := \sum_t tf(t,d) \cdot idf(t)$$

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

Slide 10

What tops TF-IDF?

Anchor-text	**Add to document, boosts TF.**
Incoming links	**Who believes in history-biased authorities?**

Slide 11

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?

Slide 12

A bit more of matrix framework

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

Let's multiply each matrix with its transposed matrix.

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

Number of common/shared terms: Document similarity.

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

Number of common/shared documents: Term similarity.

Slide 13

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.

Slide 14

More matrices? Yes!

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

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

Slide 15

TF-IDF in matrix framework?

Let's count documents in which a term occurs:

$$n_D = D^T \times DT$$

$n_D = (4, 3, 2, 1)$ $D^T = (1, 1, ..., 1)$ DT **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:
$n_L(\cdot, d) = L_d^T \times LT_d$

Compare to $n_D(\cdot, c) = D_c^T \times DT_c$

Slide 16

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:**

$$RSV = \left[\begin{array}{c|cccc} & sailing & boats & east & coast \\ \hline d1 & 2 & 1 & 0 & 0 \\ d2 & 0 & 3 & 1 & 0 \end{array}\right] \cdot \vec{q}$$

Slide 17

The trick with the diagonal

Remember? $RSV = DT \cdot G \cdot \vec{q}$

What about? $RSV = DT \cdot IDF \cdot \vec{q}$

What's IDF? A matrix. A diagonal matrix.

$$IDF = diag(idf)$$

$$IDF = \begin{bmatrix} \textbf{\textit{idf}}(sailing) & 0 & 0 & 0 \\ 0 & \textbf{\textit{idf}}(boats) & 0 & 0 \\ 0 & 0 & \textbf{\textit{idf}}(east) & 0 \\ 0 & 0 & 0 & \textbf{\textit{idf}}(coast) \end{bmatrix}$$

Slide 18

Pause

TF-IDF done

We used $L_d^T \times LT_d$ vectors for term frequencies

We used $D_c^T \times DT_c$ vector for document frequencies

We formed a diagonal matrix of idf values

Dual operations? Just to mention one: $P_d^T \times PC_d$

Many more: inverse parent frequency, etc.

Eigenvectors of DD_c, TT_c, PP_c, CC_c: Interesting.

TT_c **Eigenvector: The document/query reflecting term co-occurrence**

Slide 19

More matrices? Yes!!

Document-Assessor matrix DA_q

AA_q: Assessor similarity

Express precision/recall in general matrix framework

Eigenvector of AA_q: ...

Slide 20

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:

$$RSV(d,q) := \sum_{t \in d \cap q} \log \frac{P(t|r) \cdot P(\bar{t}|c)}{P(\bar{t}|r) \cdot P(t|c)}$$

Binary independent retrieval model

Slide 21

Rewrite:

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

So what?

Can be expressed in general matrix framework.

$\boldsymbol{idf}(t,c) - \boldsymbol{idf}(t,r)$**: Relevance information DECREASES the discriminativeness of a term.**

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

Vries/Roelleke:SIGIR:2005

Language modelling

Slide 22

Start with

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

After some arithmetics:

$$RSV(d,q) := \sum_{t \in q} \log\left(\lambda \cdot P(t|c) + (1-\lambda) \cdot P(t|d)\right)$$

Can be expressed in general matrix framework.

Slide 23

Probabilistic Logical Implementation

```
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**

Slide 25

Readings

Rijsbergen:CJ:1986,1989

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

Wong/Yao:TOIS:1995

Interpretations of $P(d \rightarrow q)$

Fuhr/Roelleke:TOIS:1997

Probabilistic relational algebra

Roelleke/Tsikrika/Kazai:IP&M:2006(2004)

General matrix framework

Rijsbergen:2005:

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