///

Information Retrieval and Machine Learning

Massimo Melucci

University of Padua
Department of Information Engineering
massimo.melucci@unipd.it

CIMI School in Machine Learning 2015

///

///

Information Retrieval and Machine Learning

Massimo Melucci

University of Padua
Department of Information Engineering
massimo.melucci@unipd.it

CIMI School in Machine Learning 2015

Summary

Two general approaches to Information Retrieval (IR)

- First, theory, then experimentation.
- First, experimentation, then theory.
- Importance of observation.
- Circular path.

$$
\begin{tikzcd}
\text{Real world} \arrow[r] & \text{Theory} \arrow[r] & \text{Model} \arrow[r] & \text{Experiment} \arrow[ll, bend right] \arrow[lll, bend right]
\end{tikzcd}
$$

Theory and model

Theory	Model
Probability Statistics	Probabilistic (e.g. BIR)
	Language Model
	Authority Model
	Quantum Model
Geometry	Vector Space Model
	Quantum Model
Logic	Classical Logic (Boolean)
	Non-classical (Quantum) Logic

Model of IR

- A set of abstract structures (algebraic structures?) that describes documents and queries.
- It defines an operation called retrieval function that maps structures to the real field.
- It does not provide implementation details, but the retrieval function should be like

$$\sum_{\substack{\text{query and document} \\ \text{content descriptor } t}} \text{weight}(t)$$

  for efficiency reasons.
- It is based on a metaphor, that is, a figure of speech in which a word or phrase is applied to an object or action to which it is not literally applicable.

Summary

Introduction

- Most used model in early IR System (IRS).
- Little used in search engines.
- Little used by most end users.
- Requires interaction and expertise.
- It may be very effective.
- See [1], [2], [3], [7], [11], [30], [32].

Metaphor

- A content descriptor is a set of documents.
- Documents are set elements.
- Queries are Boolean expressions.
- Complex descriptors are Boolean expressions.
- The retrieval function maps a document to a real number.
- In most cases, the image is $\{0, 1\}$ (i.e. true, false).

Cognitive overload

- Users are expected to know the Boolean logic.
- And the domain of the document collection.
- Otherwise, this model is cause of cognitive overload.
- That is, frustation. In particular:
- Confusion about which operator (i.e. AND, OR, NOT) should be used.
- No unique Boolean expression of a natural language expression.
- Some alleviation from (graphical) user interfaces.

Retrieved document set dimension

- The retrieved document set may be very large or very small. Two extreme cases:
- Null output.
- Output overload.

DNF

- Every Boolean expression can be translated into an equivalent Disjunctive Normal Form (DNF).
- Example: “(apple OR orange) AND NOT juice” becomes “(apple AND NOT juice) OR (orange AND NOT juice)”.

Coordination level

- Term weights and heuristic weight functions.
- Term weight:

word x	weight $w(x)$
apple	2
orange	1
juice	3

- Weight function for AND:

$$w(x_1 \text{AND} w_2) = w(x_1) + w(x_2)$$

- Weight function for NOT:

$$w(\text{NOT} x) = -w(x)$$

- Weight function for OR:

$$w(x_1 \text{OR} w_2) = \max\{w(x_1), w(x_2)\}$$

Coordination level

- Suppose a document is indexed by apple, orange and juice.

$$
\begin{aligned}
& w((\text{apple AND NOT juice) OR (orange AND NOT juice)} \\
&= \max\{w(\text{apple AND NOT juice}), w(\text{orange AND NOT juice})\} \\
&= \max\{w(\text{apple}) + w(\text{NOT juice}), w(\text{orange}) + w(\text{NOT juice})\} \\
&= \max\{w(\text{apple}) - w(\text{juice}), w(\text{orange}) - w(\text{juice})\} \\
&= \max\{2 - 3, 1 - 3\} \\
&= -1
\end{aligned}
$$

Coordination level

- Suppose a document is indexed by orange, juice but not apple.

$$
\begin{aligned}
& w((\text{apple AND NOT juice}) \text{ OR } (\text{orange AND NOT juice}) \\
&= \max\{w(\text{apple AND NOT juice}), w(\text{orange AND NOT juice})\} \\
&= \max\{w(\text{apple}) + w(\text{NOT juice}), w(\text{orange}) + w(\text{NOT juice})\} \\
&= \max\{w(\text{apple}) - w(\text{juice}), w(\text{orange}) - w(\text{juice})\} \\
&= \max\{0 - 3, 1 - 3\} \\
&= -2
\end{aligned}
$$

Coordination level

- Suppose a document is indexed by orange and but not juice.

$$
\begin{aligned}
& w((\text{apple AND NOT juice}) \text{ OR } (\text{orange AND NOT juice}) \\
&= \max\{w(\text{apple AND NOT juice}), w(\text{orange AND NOT juice})\} \\
&= \max\{w(\text{apple}) + w(\text{NOT juice}), w(\text{orange}) + w(\text{NOT juice})\} \\
&= \max\{w(\text{apple}) - w(\text{juice}), w(\text{orange}) - w(\text{juice})\} \\
&= \max\{2 - 0, 1 - 0\} \\
&= 2
\end{aligned}
$$

Summary

Vector-space modeling

- The early formulation by Gerald Salton was in the 1960s.
- It became well known in the 1970s.
- It was applied to several tasks in the 1980s
- and industrialized in the 1990s.
- Its name is Vector Space Model (VSM).
- See [12], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27].

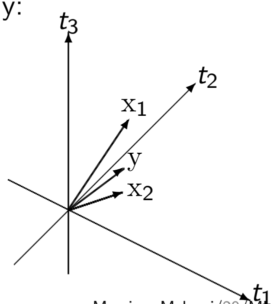
VSM made simple

- Documents and queries are vectors.
- Documents are ranked by inner product.
- Example: two document vectors $\mathrm{x}_1, \mathrm{x}_2$ and one query vector y:

$$\mathrm{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \qquad \mathrm{x}_1 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \qquad \mathrm{y} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\mathrm{x}_1^* \mathrm{y} = 3 \qquad \mathrm{x}_2^* \mathrm{y} = 0$$

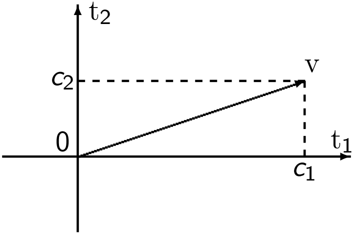
- Graphically:



What are vectors?

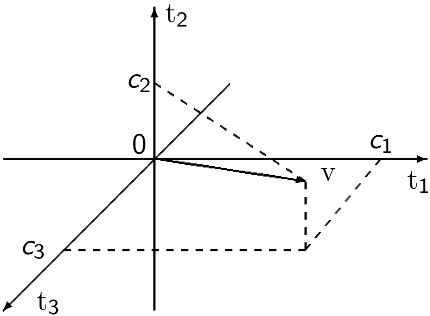
- A content descriptor is a basis vector.
- An index is a basis of a real vector space.
- The number of distinct descriptors is the dimension of the space.
- Documents are vectors.
- Queries are vectors.
- Complex descriptors are vectors.
- Passages are vectors.
- ...
- The retrieval function maps a document-query to a real number.

Searching in the two-dimensional space



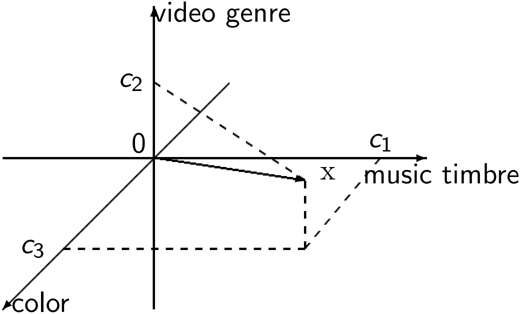
- start from $(0, 0)$ without terms in mind
- choose t_1 with weight c_1
- the query is $c_1\mathrm{t}_1$
- choose t_2 with weight c_2
- the query/document is $c_1\mathrm{t}_1 + c_2\mathrm{t}_2$

Searching in the three-dimensional space



- start from $(0, 0)$ without terms in mind
- choose t_1 with weight c_1
- choose t_2 with weight c_2
- choose t_3 with weight c_3
- the query/document is $c_1\mathrm{t}_1 + c_2\mathrm{t}_2 + c_3\mathrm{t}_3$

Searching in the three-dimensional *multimedia* space



- start from $(0, 0)$ without descriptors in mind
- choose video genre with weight c_{video}
- choose music timbre$_2$ with weight c_{timbre}
- choose color with weight c_{color}
- the query/document is
  c_{video}video genre $+$ c_{timbre}music timbre $+$ c_{color}color

Vector space concepts

- Linear independence.
- Vector basis.
- Inner product.
- Orthogonality.
- Orthonormality.

Linear independence

- Let
$$T = \{\mathrm{t}_1, \ldots, \mathrm{t}_k\}$$
be a set of vectors of $\mathbb{R}^n$.
- T is linearly independent when any t cannot be linear combination of the other t's.
- For any vector x of the k-dimensional space spanned by T
$$\mathrm{x} = \sum_{i=1}^{k} c_i \mathrm{t}_i$$
- T represents an index.
- T includes one t_i for each index term.
- Linear independence means that no index term can be expressed as "linear combination" of other index terms.
- A basis vector of T is often a canonical vector and T becomes
$$\{(1, 0, \ldots, 0), (0, 1, \ldots, 0) \ldots, (0, 0, \ldots, 1)\}$$

Inner product

- Given two vectors x, y of the same vector space.
- The inner product is the real number

$$\mathrm{x}^* \mathrm{y} = \sum_{j=1}^{n} x_j y_j$$

where x^* is a row vector.

Relevant concepts

- Weighting schemes.
- Normalization schemes.
- Correlation.
- Cluster hypothesis.

Weighting schemes

- A set of rules that compute the c_{ij}'s for each document i and term j.
- Binary.

$$c_{ij} = \begin{cases} 1 & \text{if } t_j \text{ occurs in } i \\ 0 & \text{otherwise} \end{cases}$$

- Term Frequency (TF).

$$c_{ij} = f_{ij} \qquad f_{ij} \text{ is the frequency of term } j \text{ in document } i$$

- Inverse Document Frequency (IDF).

$$c_{ij} = \log N/n_j \qquad n_j \text{ is the number of documents indexed by term } j$$

- TF $\times$ IDF (TFIDF).

$$c_{ij} = f_{ij} \log \frac{N}{n_j}$$

Normalization schemes

- Short documents contain little data.
- Long documents contain much data.
- Short documents might containt little both relevant and non-relevant information.
- Long documents might containt much both relevant and non-relevant information.
- Normalization keeps control of document length.
- Three methods:
- Cosine: normalize by $\sqrt{\mathrm{x}^*\mathrm{x}\mathrm{y}^*\mathrm{y}}$ (the result is the cosine of the angle between x, y.
- Maximum weight: normalize by $\max_i x_i$.
- Pivot: normalize by smoothed document length when this length is relatively large.

Correlation

- Let x, y be two vectors
- Inner product

$$\mathrm{x}^*\mathrm{y} = \sum_{i=1}^{k} \sum_{j=1}^{k} c_i b_j \mathrm{t}_i \mathrm{t}_j$$

- Correlation matrix $\mathrm{T}^*\mathrm{T} = \mathrm{R} = (\mathrm{t}_i \mathrm{t}_j)$
- Suppose that

$$\mathrm{R} = \begin{pmatrix} 1 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathrm{c}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} \qquad \mathrm{c}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

- We have that

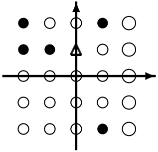
$$\mathrm{c}_1^*\mathrm{R}\mathrm{c}_2 = \frac{3}{2}$$

while

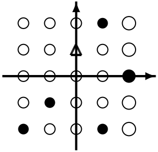
$$\mathrm{c}_1^*\mathrm{c}_2 = 1$$

Cluster Hypothesis

- Relevant documents tend to resemble relevant documents more than non-relevant documents.
- Cluster Hypothesis "holds":



- Cluster Hypothesis "does not hold":



- Why is the cluster hypothesis important?
- Efficiency reasons.
- Effectiveness reasons.

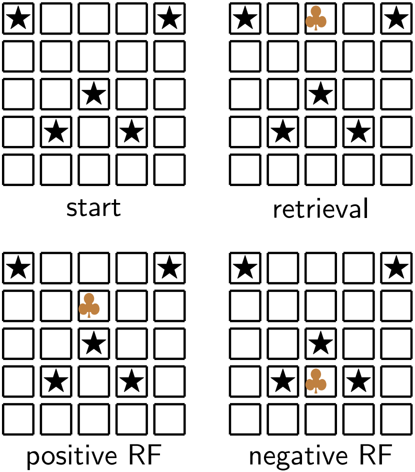
VSM-Relevance Feedback (RF)

- Query y.
- r relevant documents $x_1, \ldots, x_r$.
- $n-r$ non-relevant documents $x_{r+1}, \ldots, x_n$.
- Modified query:

$$\mathrm{y}' = \mathrm{y} + \sum_{j=1}^{r} \alpha_j \mathrm{x_j} + \sum_{h=r+1}^{n} \beta_h \mathrm{x_h} \qquad \alpha_j \geq 0 \qquad \beta_h \leq 0$$

- Positive RF: $\sum_{j=1}^{r} \alpha_j \mathrm{x_j}, \alpha_j \geq 0$.
- Negative RF: $\sum_{h=r+1}^{n} \beta_h \mathrm{x_h}, \beta_h \leq 0$.
- The α's and the β's are *free parameters*.

Positive and negative RF



Summary

Relevance probabilistic modeling

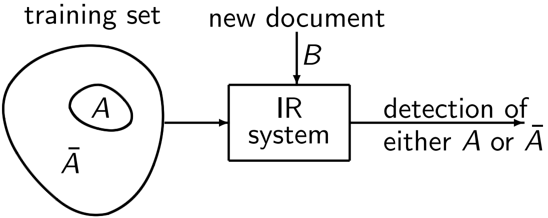
- One of the most successful approaches to IR.
- Early studies and results date back to Sixties.
- Currently, it is the foundation of IR systems.
- Integrated with Machine Learning (ML) approaches.
- See [4], [5], [10], [16], [15], [14], [29], [28], [31].

A few definitions of probability

- Elementary event: a single occurrence of a process or phenomenon.
- Cannot be decomposed into simpler occurrences.
- Event: a set of elementary events.
- Probability measure: a function that maps an elementary event to $[0, 1] \subset \mathbb{R}$.
- Degree of belief that the elementary event occurs.

Metaphor of relevance probabilistic modeling

- Document collection as elementary event space.
- Terms (or descriptors) are document sets.
- Relevance is a document set A.
- A changes for each information need.
- Retrieval is decision.



Retrieval decision

- Retrieval decision is affected by uncertainty.
- Statistical decision.
- Perfect retrieval: all relevant documents and no non-relevant documents.
- Two errors.
- Retrieve non-relevant documents.
- Miss relevant documents.
- Two costs.
- False alarm.
- Loss of recall.
- Optimal retrieval: the largest number of relevant documents provided the maximum number of non-relevant documents.

Decision costs

True Relevance	Decision	
	Relevant	Non-relevant
Relevant	$c(A, A)$	$c(A, \bar{A})$
Non-relevant	$c(\bar{A}, A)$	$c(\bar{A}, \bar{A})$

Decision risks

- Risk:

$$R(A|B) = c(A, A)P(A \mid B) + c(\bar{A}, A)P(\bar{A} \mid B)$$

$$R(\bar{A}|B) = c(A, \bar{A})P(A \mid B) + c(\bar{A}, \bar{A})P(\bar{A} \mid B)$$

- Decision for retrieval:

$$R(A \mid B) < R(\bar{A} \mid B)$$

- If and only if:

$$P(A \mid B) > c \qquad c = \frac{c(\bar{A}, A) - c(\bar{A}, \bar{A})}{c(\bar{A}, A) - c(\bar{A}, \bar{A}) + c(A, \bar{A}) - c(A, A)}$$

Probability Ranking Principle (PRP)/1

If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.

Implementation of $P(B|A)$ and $P(B|\bar{A})$

- For each $b \in B$ $P(\{b\}|A)$ and $P(\{b\}|\bar{A})$ is needed.
- Documents in B are described by k properties.
- Properties are described by a random variable X.
- Simplest approach is binary:

$$X_i(\omega) = 1 \quad \text{term occurs in } \omega \quad \omega \in B$$

Curse of dimensionality

- Let B mapped to X.
- Then $P(B|A) = P(\mathrm{X} = \mathrm{x}|X_A = 1)$
- and $P(B|\bar{A}) = P(\mathrm{X} = \mathrm{x}|X_A = 0)$.
- As $\mathrm{X} = \mathrm{x}$ is $X_1 = x_1, \ldots, X_k = x_k$,

$$P(\mathrm{X} = \mathrm{x}|X_A = 1) = P(X_1 = x_1, \ldots, X_k = x_k|X_A = 1)$$

- However, we need $O(2^k)$ estimators.

Conditional stochastic independence

- Assumption:

$$P(\mathrm{X} = \mathrm{x} | X_A = 1) = P(X_1 = x_1 | X_A = 1) \cdots P(X_k = x_k | X_A = 1)$$

$$P(\mathrm{X} = \mathrm{x} | X_A = 0) = P(X_1 = x_1 | X_A = 0) \cdots P(X_k = x_k | X_A = 0)$$

- Let

$$p_i = P(X_i = 1 \mid X_A = 1) \qquad q_i = P(X_i = 1 \mid X_A = 0)$$

- Then we have the following two likelihoods:

$$P(\mathrm{X} = \mathrm{x} \mid X_A = 1) = \prod_{i=1}^{k} p_i^{x_i} (1 - p_i)^{1 - x_i}$$

$$P(\mathrm{X} = \mathrm{x} \mid X_A = 0) = \prod_{i=1}^{k} q_i^{x_i} (1 - q_i)^{1 - x_i}$$

Retrieval decision

- The retrieval decision between relevance and non-relevance implies a hypothesis test.
- Likelihood ratio:

$$L(\mathrm{x}) = \frac{P(\mathrm{X} = \mathrm{x}|X_A = 1)}{P(\mathrm{X} = \mathrm{x}|X_A = 0)}$$

Theorem (Neyman-Pearson's Lemma (NPL))

When performing a hypothesis test between two hypotheses (e.g. relevance vs non-relevance) then the likelihood ratio which rejects relevance in favour of non-relevance when

$$L(\mathrm{x}) \leq \lambda$$

where

$$P(\mathrm{X} = \mathrm{x}|X_A = 0) = \alpha$$

is the most powerful test of size α for a given threshold λ.

Application to IR of NPL

- Application to IR gives the likelihood ratio of the Binary Independence Retrieval (BIR) model:

$$L(\mathrm{x}) = \frac{P(\mathrm{X} = \mathrm{x} \mid X_A = 1)}{P(\mathrm{X} = \mathrm{x} \mid X_A = 0)} = \frac{\prod_{i=1}^{k} p_i^{x_i}(1-p_i)^{1-x_i}}{\prod_{i=1}^{k} q_i^{x_i}(1-q_i)^{1-x_i}}$$

- Logarithmic transformation:

$$\ell(\mathrm{x}) = \sum_{i=1}^{k} x_i w_i + \sum_{i=1}^{k} \log \frac{1-p_i}{1-q_i}$$

- Term Relevance Weight (TRW):

$$w_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

What about query?

- The query is not modeled.
- Relevance is modeled.
- However, efficiency reasons requires query modeling.
- Let $\mathrm{Y} = (Y_1, \ldots, Y_k)$ where $Y_i = 1$ term i occurs in the query, 0 otherwise.
- Let $\mathrm{Z} = (Z_1, \ldots, Z_k)$ where $Z_i = X_i Y_i$.
- Then, rank documents by $\ell(\mathrm{z})$.

Parameter estimation

- Contingency table:

	A	$\bar{A}$	
$X_i = 1$	r_i	$n_i - r_i$	n_i
$X_i = 0$	$R - r_i$	$N - n_i - R + r_i$	$N - n_i$
	R	$N - R$	N

- Maximum likelihood estimators:

$$\hat{p}_i = \frac{r_i}{R} \qquad \hat{q}_i = \frac{n_i - r_i}{N - R}$$

- Laplace smoothing:

$$\hat{p}_i = \frac{r_i + \frac{1}{2}}{R + 1} \qquad \hat{q}_i = \frac{n_i - r_i + \frac{1}{2}}{N - R + 1}$$

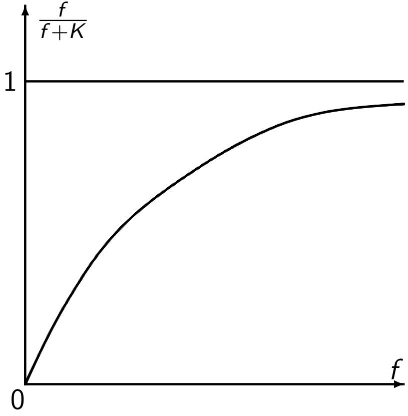
$$\mathrm{TRW} = \log \frac{r_i + 0.5}{R - r_i + 0.5} - \log \frac{n_i - r_i + 0.5}{N - n_i - R + r_i + 0.5}$$

Best Match N. 25 (BM25)

$$w_{ij} = \underbrace{\frac{(k_1+1)f_{ij}(k_3+1)f_{qj}}{(k+f_{ij})(k_3+f_{qj})}}_{\text{saturation term}} \mathrm{TRW}_{ij}$$

- l average sample document length.
- l_i length of document i.
- $k = k_1((1-b) + b\frac{l_i}{l})$.
- b is a *free parameter* (usually 0.75).
- k_1 e k_3 are *free parameters* (usually, 1.2 and something betwee 7 and 1000).
- f_{ij} is frequency of j in document i.
- f_{qj} is frequency of j in the query.

Saturation



Relevance Feedback (RF)

- Start with No relevance data and rank by:

$$g^{(0)}(\mathbf{z}) = \sum_i z_i w_i^{(0)} \qquad w_i^{(0)} = \log \frac{N - n_i + \frac{1}{2}}{n_i + \frac{1}{2}}$$

- Collect some relevance data and rank at step t by:

$$g^{(t-1)}(\mathbf{z}) = \sum_i z_i w_i^{(t-1)}$$

where

$$w_i^{(t-1)} = \log \hat{p}_i^{(t-1)} + \log 1 - \hat{q}_i^{(t-1)} - \log \hat{q}_i^{(t-1)} - \log 1 - \hat{p}_i^{(t-1)}$$

$$\hat{p}_i^{(t-1)} = \frac{r_i^{(t-1)} + a^{(t-1)}}{R^{(t-1)} + b^{(t-1)}} \qquad \hat{q}_i^{(t-1)} = \frac{n_i - r_i^{(t-1)} + c^{(t-1)}}{N - R^{(t-1)} + d^{(t-1)}}$$

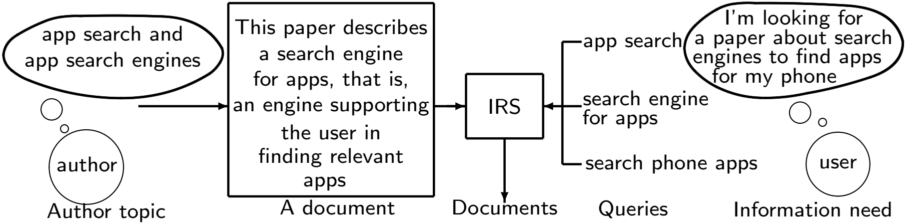
$$a^{(t-1)} = c^{(t-1)} = \frac{1}{2} \qquad b^{(t-1)} = d^{(t-1)} = 1$$

usually.

Summary

Metaphor

- Author thinks about queries for his document.
- He writes the document using queries and variations of them.
- Pictorially,



- See [13], [34], [35], [9], [6], [8], [33].

Assumptions

- The user is assumed to have a good idea of what he is searching.
- The author is assumed to have a good idea of the user's need.
- Both are assumed to use an effective and the same language.
- Under these assumptions, the documents generated by the authors are likely relevant to the user's information need.

Language Model

- Let w be a symbol.
- A language L is defined as a set of symbol

$$L = \{w_1, \ldots, w_N\}$$

- A Language Model (LM) is a language L provided with a probability function
- Example:

upon the bench the goat lives
under the bench the goat dies

- Remove stopwords and stem words to obtain

bench goat live bench goat dies

- Language is $L = \{\text{bench}, \text{goat}, \text{live}, \text{die}\}$ such that

$$P(\text{bench}) = \frac{2}{6} \quad P(\text{goat}) = \frac{2}{6} \quad P(\text{live}) = \frac{1}{6} \quad P(\text{die}) = \frac{1}{6}$$

QLM

- Mostly used in IR.
- Queries are LMs.
- Documents are samples.
- The IRS looks for the most likely document given a query:

  $$B^* = \arg_B \max P(B \mid Q)$$

  where Q is the Query Language Model (QLM) and B is a document event.
- Documents are ranked by $P(B \mid Q)$.

How to estimate a QLM

- However, Q is not completely known: the language is known but the probability is unknown.
- Bayes' theorem:

$$P(B \mid Q) = \frac{P(Q \mid B)P(B)}{P(Q)}$$

- $P(Q)$ is constant.
- $P(B)$ uniform or estimated by external sources.
- Estimation: B is an n-gram $w_{(1)} \ldots w_{(n)}$

  $$P(Q|B) = p_B(w_{(1)})p_B(w_{(2)}|w_{(1)}) \cdots p_B(w_{(n)} \mid w_{(n-1)} \cdots w_{(1)})$$

- Stochastic independence:

  $$P(Q|B) = p_B(w_{(1)}) \cdots p_B(w_{(n)})$$

  where

  $$p_B(w_{(i)}) = \frac{f(w_{(i)}, B)}{\sum_{i=1}^{n} f(w_{(i)}, B)}$$

  and $f(w, B)$ is the frequency of $w_{(i)}$ in B.

Mixture and smoothing

- Problem: $f(w,B)$ might be 0.
- Solution: mixture.

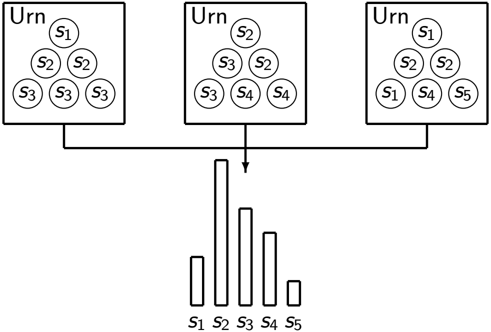
$$\hat{p}_B(w_{(i)}) = (1-\lambda)\frac{f(w_{(i)},B)}{\sum_{i=1}^{n} f(w_{(i)},B)} + \lambda\frac{f(w_{(i)},\mathcal{V})}{\sum_{i=1}^{n} f(w_{(i)},\mathcal{V})}$$

where $\mathcal{V}$ is the collection language.

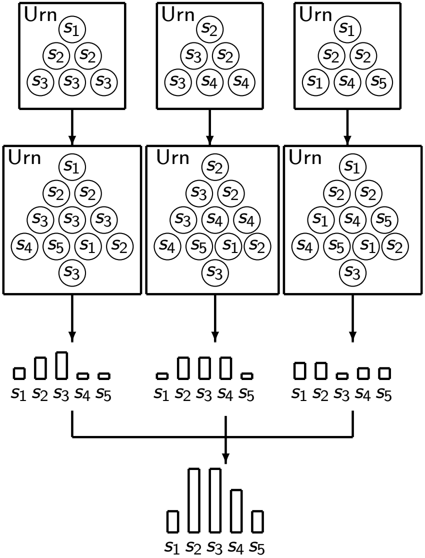
- Alternatively, smoothing.

$$\hat{p}_B(w_{(i)}) = \frac{f(w_{(i)},B)+a}{\sum_{w\in B} f(w_{(i)},B)+a+b}$$

Mixture


Urn
s_1
s_2 s_2
s_3 s_3 s_3
Urn
s_2
s_3 s_2
s_3 s_4 s_4
Urn
s_1
s_2 s_2
s_1 s_4 s_5
s_1 s_2 s_3 s_4 s_5

Smoothing



Urn
s_1
s_2 s_2
s_3 s_3 s_3
Urn
s_2
s_3 s_2
s_3 s_4 s_4
Urn
s_1
s_2 s_2
s_1 s_4 s_5
Urn
s_1
s_2 s_2
s_3 s_3 s_3
s_4 s_5 s_1 s_2
s_3
Urn
s_2
s_3 s_2
s_3 s_4 s_4
s_4 s_5 s_1 s_2
s_3
Urn
s_1
s_2 s_2
s_1 s_4 s_5
s_4 s_5 s_1 s_2
s_3
s_1 s_2 s_3 s_4 s_5
s_1 s_2 s_3 s_4 s_5
s_1 s_2 s_3 s_4 s_5
s_1 s_2 s_3 s_4 s_5

Summary

Setting

- Topic set: TREC-6, TREC-7, TREC-8.
- Mean Average Precision (AP) (MAP) presented by:
- Query type: topic title-only, topic title and description.
- Model: LM, BM25, VSM-TFIDF.
- RF: without Pseudo Relevance Feedback (PRF), with PRF (i.e. no explicit RF).
- When with PRF: N. PRF documents, N. PRF terms $= \{5,10\} \times \{5,10\}$.
- Note that the number of PRF terms and the number of PRF documents are free parameters.

Some general comments

- Long queries are not worse (usually better) than short queries.
- RF improves TFIDF and BM25 and does not improve LM.
- RF improves effectiveness with a few documents and terms ($n = 5, k = 5$); larger numbers do not provide further increment.
- LM seems slightly superior to TFIDF and BM25 when RF is not applied, but...
- The experiments have been performed using Lemur, which is the IRS developed by a LM research group.

Detailed results follow

Evaluation Results

- Topic title-only queries.
- N. PRF documents: 5.
- N. PRF terms: 5.
- TREC-6 topic set.

LM		0.1402
LM	PRF	0.1424
BM25		0.1129
BM25	PRF	0.1424
VSM-TFIDF		0.1302
VSM-TFIDF	PRF	0.1424

Evaluation Results

- Topic title-only queries.
- N. PRF documents: 5.
- N. PRF terms: 5.
- TREC-7 topic set.

LM		0.1807
LM	PRF	0.1800
BM25		0.1549
BM25	PRF	0.1800
VSM-TFIDF		0.1687
VSM-TFIDF	PRF	0.1800

Evaluation Results

- Topic title-only queries.
- N. PRF documents: 5.
- N. PRF terms: 5.
- TREC-8 topic set.

LM		0.1708
LM	PRF	0.1682
BM25		0.1582
BM25	PRF	0.1751
VSM-TFIDF		0.1588
VSM-TFIDF	PRF	0.1747

Evaluation Results

- Topic title and description queries.
- N. PRF documents: 5.
- N. PRF terms: 5.
- TREC-6 topic set.

LM		0.1582
LM	PRF	0.1516
BM25		0.1377
BM25	PRF	0.1516
VSM-TFIDF		0.1743
VSM-TFIDF	PRF	0.1516

Evaluation Results

- Topic title and description queries.
- N. PRF documents: 5.
- N. PRF terms: 5.
- TREC-7 topic set.

LM		0.1773
LM	PRF	0.1759
BM25		0.1427
BM25	PRF	0.1759
VSM-TFIDF		0.1818
VSM-TFIDF	PRF	0.1759

Evaluation Results

- Topic title and description queries.
- N. PRF documents: 5.
- N. PRF terms: 5.
- TREC-8 topic set.

LM		0.1498
LM	PRF	0.1499
BM25		0.1351
BM25	PRF	0.1418
VSM-TFIDF		0.1594
VSM-TFIDF	PRF	0.1602

Evaluation Results

- Topic title-only queries.
- N. PRF documents: 10.
- N. PRF terms: 5.
- TREC-6 topic set.

LM		0.1402
LM	PRF	0.1403
BM25		0.1129
BM25	PRF	0.1204
VSM-TFIDF		0.1302
VSM-TFIDF	PRF	0.1276

Evaluation Results

- Topic title-only queries.
- N. PRF documents: 10.
- N. PRF terms: 5.
- TREC-7 topic set.

LM		0.1807
LM	PRF	0.1804
BM25		0.1549
BM25	PRF	0.2032
VSM-TFIDF		0.1687
VSM-TFIDF	PRF	0.1940

Evaluation Results

- Topic title-only queries.
- N. PRF documents: 10.
- N. PRF terms: 5.
- TREC-8 topic set.

LM		0.1708
LM	PRF	0.1680
BM25		0.1582
BM25	PRF	0.1773
VSM-TFIDF		0.1588
VSM-TFIDF	PRF	0.1715

Evaluation Results

- Topic title and description queries.
- N. PRF documents: 10.
- N. PRF terms: 5.
- TREC-6 topic set.

LM		0.1582
LM	PRF	0.1454
BM25		0.1377
BM25	PRF	0.1727
VSM-TFIDF		0.1743
VSM-TFIDF	PRF	0.1812

Evaluation Results

- Topic title and description queries.
- N. PRF documents: 10.
- N. PRF terms: 5.
- TREC-7 topic set.

LM		0.1773
LM	PRF	0.1751
BM25		0.1427
BM25	PRF	0.1955
VSM-TFIDF		0.1818
VSM-TFIDF	PRF	0.2005

Evaluation Results

- Topic title and description queries.
- N. PRF documents: 10.
- N. PRF terms: 5.
- TREC-8 topic set.

LM		0.1498
LM	PRF	0.1503
BM25		0.1351
BM25	PRF	0.1407
VSM-TFIDF		0.1594
VSM-TFIDF	PRF	0.1594

Evaluation Results

- Topic title-only queries.
- N. PRF documents: 10.
- N. PRF terms: 10.
- TREC-6 topic set.

LM		0.1402
LM	PRF	0.1426
BM25		0.1129
BM25	PRF	0.1205
VSM-TFIDF		0.1302
VSM-TFIDF	PRF	0.1297

Evaluation Results

- Topic title-only queries.
- N. PRF documents: 10.
- N. PRF terms: 10.
- TREC-7 topic set.

LM		0.1807
LM	PRF	0.1854
BM25		0.1549
BM25	PRF	0.2150
VSM-TFIDF		0.1687
VSM-TFIDF	PRF	0.2028

Evaluation Results

- Topic title-only queries.
- N. PRF documents: 10.
- N. PRF terms: 10.
- TREC-8 topic set.

LM		0.1708
LM	PRF	0.1743
BM25		0.1582
BM25	PRF	0.1880
VSM-TFIDF		0.1588
VSM-TFIDF	PRF	0.1843

Evaluation Results

- Topic title and description queries.
- N. PRF documents: 10.
- N. PRF terms: 10.
- TREC-6 topic set.

LM		0.1582
LM	PRF	0.1517
BM25		0.1377
BM25	PRF	0.1773
VSM-TFIDF		0.1743
VSM-TFIDF	PRF	0.1866

Evaluation Results

- Topic title and description queries.
- N. PRF documents: 10.
- N. PRF terms: 10.
- TREC-7 topic set.

LM		0.1773
LM	PRF	0.1915
BM25		0.1427
BM25	PRF	0.1997
VSM-TFIDF		0.1818
VSM-TFIDF	PRF	0.2077

Evaluation Results

- Topic title and description queries.
- N. PRF documents: 10.
- N. PRF terms: 10.
- TREC-8 topic set.

LM		0.1498
LM	PRF	0.1436
BM25		0.1351
BM25	PRF	0.1405
VSM-TFIDF		0.1594
VSM-TFIDF	PRF	0.1603

[1] A. V. Aho and J. D. Ullman.
Foundations of computer science.
http://i.stanford.edu/~ullman/focs.html.

[2] G. Boole.
An Investigation of the laws of Thought.
Walton and Maberly, 1854.

[3] W. Cooper.
Getting beyond Boole.
Information Processing & Management, 24:243–248, 1988.

[4] W. Cooper.
Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval.
ACM Transactions on Information Systems, 13(1):100–111, Jan. 1995.

[5] W. Croft and D. Harper.
Using probabilistic models of document retrieval without relevance information.
Journal of Documentation, 35:285–295, 1979.

[6] W. Croft and J. Lafferty, editors.

Language Modeling for Information Retrieval, volume 13 of *Kluwer International Series on Information Retrieval*.

Kluwer Academic Publishers, 2002.

[7] D. E. Knuth.

The Art of Computer Programming, volume 1-4.

Addison-Wesley Professional, III edition, 2011.

[8] J. Lafferty and C. Zhai.

Probabilistic relevance models based on document and query generation, chapter 1.

Volume 13 of Croft and Lafferty [6], 2002.

[9] V. Lavrenko and W. Croft.

Relevance-based language models.

In *Proceedings of SIGIR*, pages 120–127, New Orleans, LO, USA, Sept. 2001.

[10] M. Maron and J. Kuhns.

On relevance, probabilistic indexing and retrieval.

Journal of the ACM, 7:216–244, 1960.

[11] M. Melucci.

The boolean model.
Encyclopedia on Database Systems, 2009.

[12] M. Melucci.
The vector space model.
Encyclopedia on Database Systems, 2009.

[13] J. Ponte and W. Croft.
A language modeling approach to information retrieval.
In *Proceedings of SIGIR*, pages 275–281. ACM Press, Aug. 1998.

[14] S. Robertson.
The probability ranking principle in information retrieval.
Journal of Documentation, 33(4):294–304, 1977.

[15] S. Robertson and S. Walker.
Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval.
In *Proceedings of SIGIR*, pages 232–241, 1994.

[16] S. Robertson and H. Zaragoza.
The probabilistic relevance framework: BM25 and beyond.
Foundations and Trends in Information Retrieval, 3(4):333–389, 2009.

[17] J. Rocchio.
Relevance feedback in information retrieval.
In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter 14, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[18] G. Salton.
Automatic Information Organization and Retrieval.
Mc Graw Hill, New York, NY, 1968.

[19] G. Salton.
A theory of indexing.
Philadelphia : Society for Industrial and Applied Mathematics, 1975.

[20] G. Salton.
Mathematics and information retrieval.
Journal of Documentation, 35(1):1–29, 1979.

[21] G. Salton.
Automatic Text Processing.
Addison-Wesley, 1989.

[22] G. Salton and C. Buckley.

Term weighting approaches in automatic text retrieval.
Information Processing & Management, 24(5):513–523, 1988.

[23] G. Salton and C. Buckley.
Improving retrieval performance by relevance feedback.
Journal of the American Society for Information Science, 41(4):288–297, 1990.

[24] G. Salton, E. A. Fox, C. Buckley, and E. M. Voorhees.
Boolean query formulation with relevance feedback.
Technical Report TR83-539, Cornell University, Computer Science Department, Jan. 1983.

[25] G. Salton and M. McGill.
Introduction to Modern Information Retrieval.
McGraw-Hill, New York, NY, 1983.

[26] G. Salton, A. Wong, and C. Yang.
A vector space model for automatic indexing.
Communications of the ACM, 18(11):613–620, 1975.

[27] A. Singhal, C. Buckley, and M. Mitra.
Pivoted document length normalization.
In *Proceedings of SIGIR*, pages 21–29, 1996.

[28] K. Sparck Jones.
Search term relevance weighting given little relevance information.
Journal of Documentation, 35:30–48, 1979.

[29] K. Sparck Jones and P. Willett.
Readings in Information Retrieval.
Morgan Kaufmann, San Francisco, CA, 1997.

[30] H. Turtle and J. Flood.
Query evaluation: strategies and optimizations.
Inf. Process. Manage., 31(6):831–850, Nov. 1995.

[31] C. J. Van Rijsbergen.
Information Retrieval.
Butterworths, London, second edition, 1979.

[32] C. J. van Rijsbergen.
A non-classical logic for Information Retrieval.
The Computer Journal, 29(6):481–485, 1986.

[33] C. Zhai.
Statistical Language Models for Information Retrieval: A Critical Review.

Foundations and Trends in Information Retrieval. Now Publishers Inc., 2008.

[34] C. Zhai and J. Lafferty.
Model-based feedback in the language modeling approach to information retrieval.
In *Proceedings of CIKM*, pages 403–410, Atlanta, GA, USA, Nov. 2001.

[35] C. Zhai and J. Lafferty.
A study of smoothing methods for language models applied to ad-hoc information retrieval.
In *Proceedings of SIGIR*, pages 334–342, New Orleans, LA, USA, Sept. 2001.