Information Retrieval

Dhruv Gupta

dhruv.gupta@ntnu.no

27-September-2022

NTNU | Norwegian University of Science and Technology

Announcements

- Assignment 2: due next Wednesday (06.October.2022).
- Reference Group: volunteers needed for feedback regarding course.
  - Interested? Please contact me by email!

Lecture Plan (Tentative)

Week/ Day	Topic	Lecturer	Chapter Number
34 / Tuesday	Welcome, General Information, and Introduction	Dhruv Gupta	1
35 / Tuesday	Classical Similarity Models	Dhruv Gupta	3
36 / Tuesday	Classical Similarity Models (continued), BM25, and Language Model	Dhruv Gupta	3
37 / Tuesday	Classical Similarity Models (continued), BM25, and Language Model	Dhruv Gupta	3
38 / Tuesday	Evaluation in Information Retrieval	Dhruv Gupta	4
39 / Tuesday	User Relevance Feedback and Query Expansion	Dhruv Gupta	5
-	Text Operation	Heri Ramampiaro (recorded video)	6
-	Indexing and Searching	Heri Ramampiaro (recorded video)	9
-	Guest Lectures (Tentative)	Tentative	-
-	Web Search and Search Engines	Heri Ramampiaro (recorded video)	11
-	Introduction to Multimedia Retrieval and Image Retrieval	Heri Ramampiaro (recorded video)	-
-	Audio Retrieval	Heri Ramampiaro (recorded video)	-
-	Audio Retrieval	Heri Ramampiaro (recorded video)	-

- Text and diagrams of some slides are based on the material from the book: Baeza-Yates and Ribeiro-Neto, “Modern Information Retrieval", Second Edition. Pearson Education Limited, 2011.

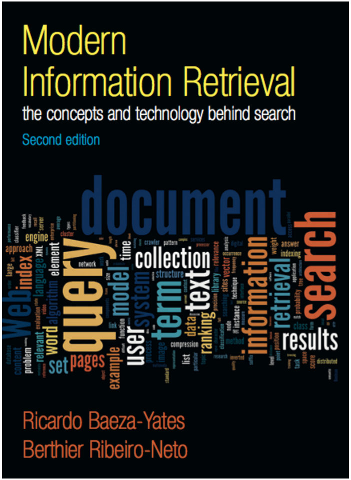
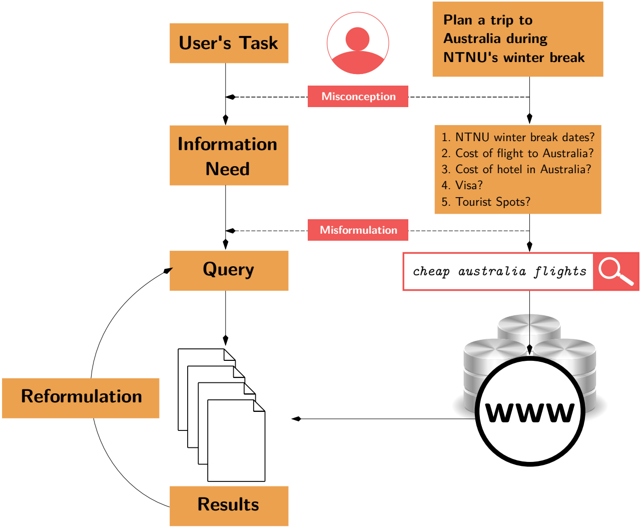
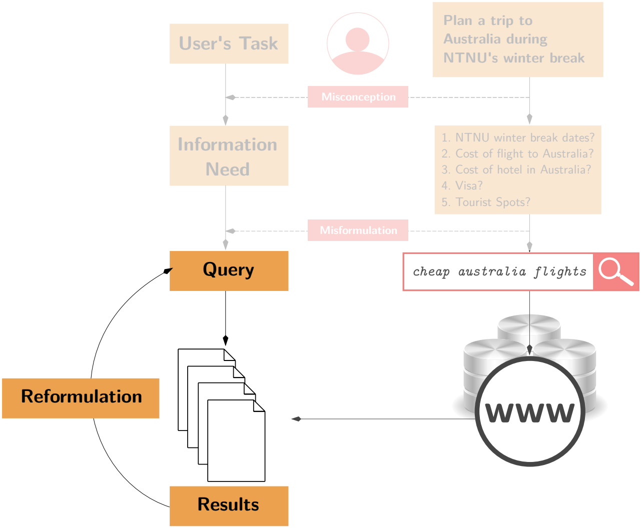


Image Credit: http://grupoweb.upf.es/mir2ed/

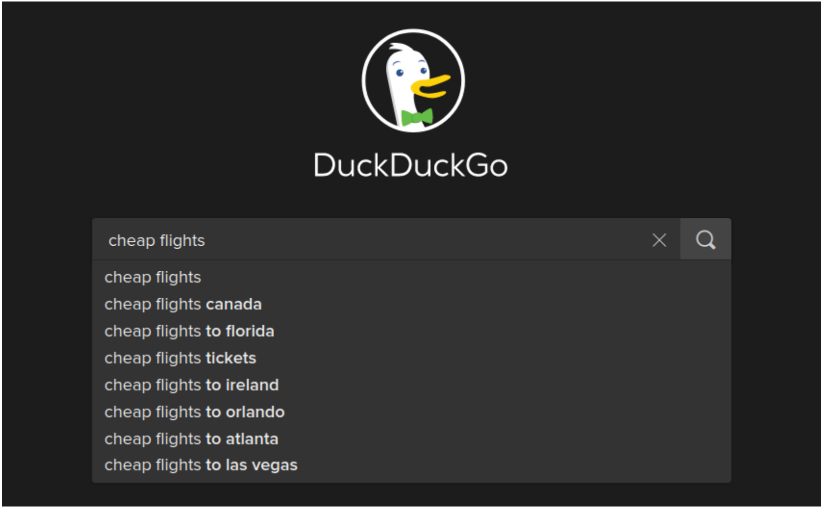
Introduction



Introduction



Introduction



Introduction

- Most users find it difficult to formulate queries that are well designed for retrieval purposes.
- Yet, most users often need to reformulate their queries to obtain the results of their interest.
  - Thus, the first query formulation should be treated as an initial attempt to retrieve relevant information.
  - Documents initially retrieved could be analyzed for relevance and used to improve initial query.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Introduction

- The process of query modification is commonly referred as:
  - relevance feedback, when the user provides information on relevant documents to a query.
  - query expansion, when information related to the query is used to expand it.
- We refer to both of them as feedback methods.
- Two basic approaches of feedback methods:
  - Explicit Feedback: information for query reformulation is provided directly by the users.
  - Implicit Feedback: information for query reformulation is implicitly derived by the system.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Framework

- Consider a set of documents D_r that are known to be relevant to the current query q.
- In relevance feedback, the documents in D_r are used to transform q into a modified query q_m.
- However, obtaining information on documents relevant to a query requires the direct interference of the user.
  - Most users are unwilling to provide this information, particularly on the Web.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Framework

- Because of this high cost, the idea of relevance feedback has been relaxed over the years.
- Instead of asking the users for the relevant documents, we could:
  - Look at documents they have clicked on.
  - Look at terms belonging to the top documents in the result set.
- In both cases, it is expect that the feedback cycle will produce results of higher quality.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Framework

- A feedback cycle is composed of two basic steps:
  - Determine feedback information that is either related or expected to be related to the original query q.
  - Determine how to transform query q to take this information effectively into account.
- The first step can be accomplished in two distinct ways:
  - Obtain the feedback information explicitly from the users.
  - Obtain the feedback information implicitly from the query results or from external sources such as a thesaurus.

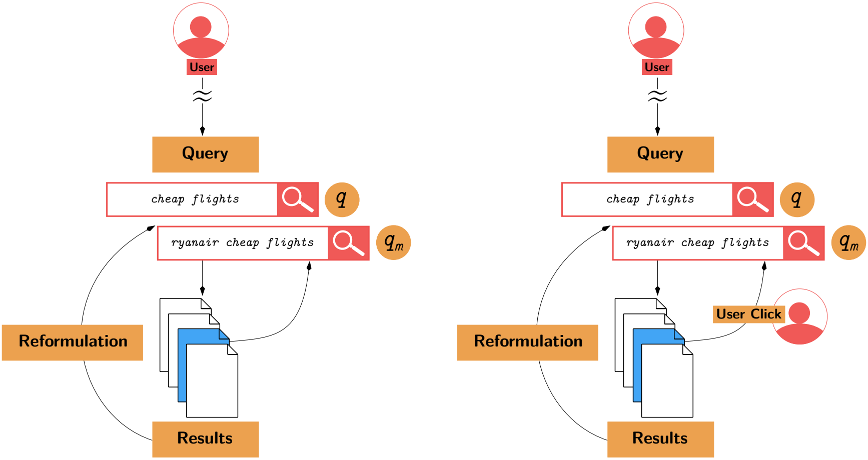
All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Framework

- In an explicit relevance feedback cycle, the feedback information is provided directly by the users.
- However, collecting feedback information is expensive and time consuming.
- In the Web, user clicks on search results constitute a new source of feedback information.
- A click indicates a document is of interest to the user in the context of the current query.
  - Notice that a click does not necessarily indicate a document that is relevant to the query.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Framework — Explicit Feedback Information



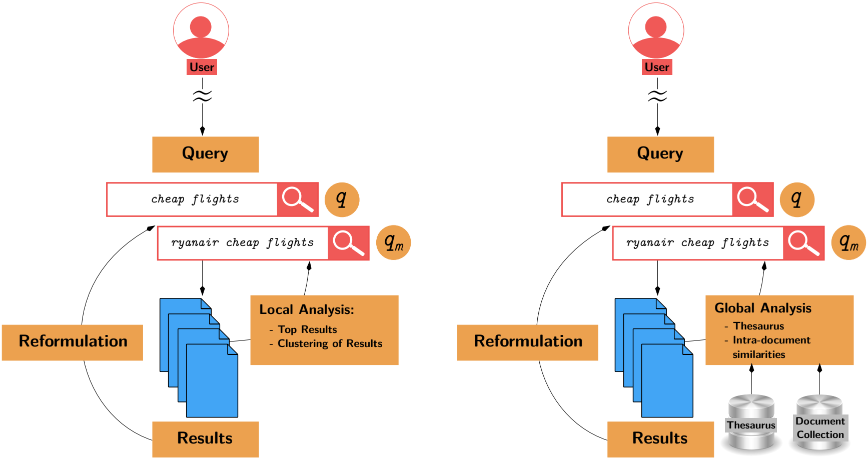
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Framework — Implicit Feedback Information

- In an implicit relevance feedback cycle, the feedback information is derived implicitly by the system.
- There are two basic approaches for compiling implicit feedback information:
  - Local Analysis: which derives the feedback information from the top ranked documents in the result set.
  - Global Analysis: which derives the feedback information from external sources such as a thesaurus.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Framework — Implicit Feedback Information



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Explicit Relevance Feedback

- In a classic relevance feedback cycle, the user is presented with a list of the retrieved documents.
- Then, the user examines them and marks those that are relevant.
- In practice, only the top 10 (or 20) ranked documents need to be examined.
- The main idea consists of:
  - Selecting important terms from the documents that have been identified as relevant.
  - Enhancing the importance of these terms in a new query formulation.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Explicit Relevance Feedback

- Expected effect: the new query will be moved towards the relevant documents and away from the non-relevant ones.
- Early experiments have shown good improvements in precision for small test collections.
- Relevance feedback presents the following characteristics:
  - It shields the user from the details of the query reformulation process (all the user has to provide is a relevance judgement).
  - It breaks down the whole searching task into a sequence of small steps which are easier to grasp.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Rocchio Method

- Documents identified as relevant (to a given query) have similarities among themselves.
- Further, non-relevant documents have term-weight vectors which are dissimilar from the relevant documents.
- The basic idea of the Rocchio Method is to reformulate the query such that it gets:
  - Closer to the neighborhood of the relevant documents in the vector space, and
  - Away from the neighborhood of the non-relevant documents.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Rocchio Method

- Let us define terminology regarding the processing of a given query q, as follows:
  - D_r: set of relevant documents among the documents retrieved.
  - N_r: number of documents in set D_r.
  - D_n: set of non-relevant documents among the documents retrieved.
  - N_n: number of documents in set D_n.
  - C_r: set of relevant documents among all documents in the collection.
  - N : number of documents in the collection.
  - $\alpha, \beta,$ and γ: tuning constants.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Rocchio Method

- Assume that the set C_r is known in advance:
- Then, the best query vector for distinguishing the relevant from the non-relevant documents is given by:

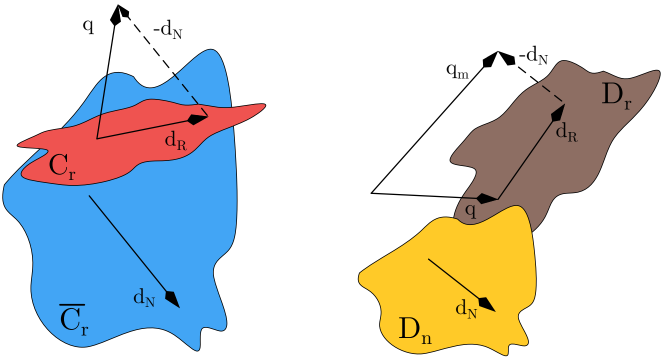
$$\vec{q}_{\text{opt}} = \frac{1}{|C_r|} \cdot \sum_{\forall \vec{d}_i \in C_r} \vec{d}_i - \frac{1}{N - |C_r|} \cdot \sum_{\forall \vec{d}_i \notin C_r} \vec{d}_i$$

  - $|C_r|$ refers to the cardinality of the set C_r.
  - $\vec{d}_i$ is a weighted term vector associated with document d_i, and
  - $\vec{q}_{\text{opt}}$ is the optimal weighted term vector for query q.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Rocchio Method

- However, the set C_r is not known a priori.
- To solve this problem, we can formulate an initial query and to incrementally change the initial query vector.



(a) (b)

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Rocchio Method

- There are three classic and similar ways to calculate the modified query $\vec{q}_m$ as follows,

$$\begin{aligned}
\text{Standard Rocchio} &: \vec{q}_m = \alpha \cdot \vec{q} + \frac{\beta}{N_r} \cdot \sum_{\forall \vec{d}_i \in D_r} \vec{d}_i - \frac{\gamma}{|N_n|} \cdot \sum_{\forall \vec{d}_i \notin D_n} \vec{d}_i \\
\text{Ide Regular} &: \vec{q}_m = \alpha \cdot \vec{q} + \beta \cdot \sum_{\forall \vec{d}_i \in D_r} \vec{d}_i - \gamma \cdot \sum_{\forall \vec{d}_i \notin D_n} \vec{d}_i \\
\text{Ide Dec Hi} &: \vec{q}_m = \alpha \cdot \vec{q} + \beta \cdot \sum_{\forall \vec{d}_i \in D_r} \vec{d}_i - \gamma \cdot \mathrm{maxrank}(D_n)
\end{aligned}$$

  - where, $\mathrm{maxrank}(D_n)$ is the highest ranked non-relevant document.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Rocchio Method

- Three different setups of the parameters in the Rocchio formula are as follows:
  - $\alpha = 1$, proposed by Rocchio.
  - $\alpha = \beta = \gamma = 1$, proposed by Ide.
  - $\gamma = 0$, which yields a positive feedback strategy.
  - The current understanding is that the three techniques yield similar results.
- The main advantages of the above relevance feedback techniques are simplicity and good results.
  - Simplicity: modified term weights are computed directly from the set of retrieved documents.
  - Good results: the modified query vector does reflect a portion of the intended query semantics (observed experimentally).

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Probabilistic Model

- The probabilistic model ranks documents for a query q according to the probabilistic ranking principle.
- The similarity of a document d_j to a query q in the probabilistic model can be expressed as:

$$\mathrm{sim}(d_i, q) \propto \sum_{w_j \in q \wedge w_j \in d_i} \left[\log\left[\frac{P(w_j|R)}{1 - P(w_j|R)}\right] + \log\left[\frac{1 - P(w_j|\bar{R})}{P(w_j|\bar{R})}\right]\right]$$

- where,
  - $P(w_j|R)$ stands for the probability of observing the term w_j in the set R of relevant documents.
  - $P(w_j|\bar{R})$ stands for the probability of observing the term w_j in the set $\bar{R}$ of non-relevant documents.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Probabilistic Model

- Initially, the equation above cannot be used because $P(w_j|R)$ and $P(w_j|\bar{R})$ are unknown.
- Different methods for estimating these probabilities automatically were discussed earlier.
- With user feedback information, these probabilities are estimated in a slightly different way.
- For the initial search (when there are no retrieved documents yet), assumptions often made include:
  - $P(w_j|R)$ is constant for all terms w_j (typically 0.5).
  - The term probability distribution $P(w_j|\bar{R})$ can be approximated by the distribution in the whole collection.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Probabilistic Model

- These two assumptions yield:

$$P(w_j|R) = 0.5 \qquad\qquad P(w_j|\bar{R}) = \frac{n_j}{N}$$

- Substituting into similarity equation, we obtain:

$$\text{sim}_{\text{initial}}(d_i, q) = \sum_{w_j \in q \wedge w_j \in d_i} \log\left[\frac{N - n_j}{n_j}\right]$$

- For the feedback searches, the accumulated statistics on relevance are used to evaluate $P(w_j|R)$ and $P(w_j|\bar{R})$.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Probabilistic Model

- Let $n_{r,j}$ be the number of documents in set D_r that contain the term w_j.
- Then, the probabilities $P(w_j|R)$ and $P(w_j|\bar{R})$ can be approximated by:

$$P(w_j|R) = \frac{n_{r,j}}{N_r} \qquad\qquad P(w_j|\bar{R}) = \frac{n_j - n_{r,j}}{N - N_r}$$

- Using these approximations, the similarity equation can be rewritten as:

$$\mathrm{sim}(d_i, q) = \sum_{w_j \in q \wedge w_j \in d_i} \left[\log\left[\frac{n_{r,j}}{N_r - n_{r,j}}\right] + \log\left[\frac{N - N_r - (n_j - n_{r,j})}{n_j - n_{r,j}}\right] \right].$$

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Probabilistic Model

- Notice that here, contrary to the Rocchio Method, no query expansion occurs.
- The same query terms are re-weighted using feedback information provided by the user.
- The formula above poses problems for certain small values of N_r and $n_{r,j}$.
- For this reason, a 0.5 adjustment factor is often added to the estimation of $P(w_j|R)$ and $P(w_j|\bar{R})$:

$$P(w_j|R) = \frac{n_{r,j} + 0.5}{N_r + 1} \qquad P(w_j|\bar{R}) = \frac{n_j - n_{r,j} + 0.5}{N - N_r + 1}.$$

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Probabilistic Model

- The main advantage of this feedback method is the derivation of new weights for the query terms.
- The disadvantages include:
  - Document term weights are not taken into account during the feedback loop.
  - Weights of terms in the previous query formulations are disregarded.
  - No query expansion is used (the same set of index terms in the original query is re-weighted over and over again).
- Thus, this method does not in general operate as effectively as the vector modification methods.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Local Analysis

- Local analysis consists of deriving feedback information from the documents retrieved for a given query q.
- This is similar to a relevance feedback cycle but done without assistance from the user.
- Two local strategies are discussed here:
  1. Local Clustering.
  2. Local Context Analysis.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Implicit Relevance Feedback

Local Analysis – Local Clustering

Local Analysis using Local Clustering

- Adoption of clustering techniques for query expansion has been a basic approach in information retrieval.
- The standard procedure is to quantify term correlations and then use the correlated terms for query expansion.
- Term correlations can be quantified by using global structures, such as association matrices.
- However, global structures might not adapt well to the local context defined by the current query.
- To deal with this problem, local clustering can be used, as we now discuss.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Modeling Documents — Term Document Matrix

	w_1	w_2	w_3	w_4	w_5	$\cdots$	$w_{\lvert\mathcal{V}\rvert}$
d_1	1	0	1	1	0	$\cdots$	1
d_2	1	1	0	0	1	$\cdots$	0
d_3	0	0	0	1	0	$\cdots$	0
d_4	0	1	0	0	1	$\cdots$	1
d_5	0	0	1	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	1	1	0	1	1	$\cdots$	0

Term-Term Correlation Matrix

- For classic information retrieval models, the index term weights are assumed to be mutually independent.
  - This means that $m_{i,j}$ tells us nothing about $m_{i+1,j}$
- This is clearly a simplification because occurrences of index terms in a document are not uncorrelated.
- For instance, the terms *computer* and *network* tend to appear together in a document about *computer networks*.
  - In this document, the appearance of one of these terms attracts the appearance of the other.
  - Thus, they are correlated and their weights should reflect this correlation.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Term-Term Correlation Matrix

- To take into account term-term correlations, we can compute a correlation matrix.
- For the correlation matrix, we reverse our convention explained earlier: now rows correspond to words w_i in the vocabulary $\mathcal{V}$ and columns correspond to documents d_j in the collection $\mathcal{D}$.
- Let $\mathbf{M}$ be a term-document matrix $|\mathcal{V}| \times |\mathcal{D}|$.
- The matrix $\mathbf{C} = \mathbf{M} \times \mathbf{M}^T$ is a term-term correlation matrix.
- Each element $c_{u,v} \in \mathbf{C}$ expresses a correlation between terms w_u and w_v given by:

$$c_{u,v} = \sum_{d_j} m_{u,j} \cdot m_{v,j}$$

- Higher the number of documents in which the terms w_u and w_v co-occur, stronger is this correlation.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Term-Term Correlation Matrix

- Term-Term correlation matrix for a sample collection.

$$\mathbf{M} \times \mathbf{M}^T = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{bmatrix} m_{1,1} & m_{1,2} \\ m_{2,1} & m_{2,2} \\ m_{3,1} & m_{3,2} \end{bmatrix} \end{array} \times \begin{array}{c} \\ d_1 \\ d_2 \end{array} \begin{array}{c} \begin{array}{ccc} w_1 & w_2 & w_3 \end{array} \\ \begin{bmatrix} m_{1,1} & m_{2,1} & m_{3,1} \\ m_{1,2} & m_{2,2} & m_{3,2} \end{bmatrix} \end{array}$$

$$\mathbf{M} \times \mathbf{M}^T = \begin{array}{c} \\ w_1 \\ w_2 \\ w_2 \end{array} \begin{array}{c} \begin{array}{ccc} w_1 & w_2 & w_3 \end{array} \\ \begin{bmatrix} m_{1,1}m_{1,1} + m_{1,2}m_{1,2} & m_{1,1}m_{2,1} + m_{1,2}m_{2,2} & m_{1,1}m_{3,1} + m_{1,2}m_{3,2} \\ m_{2,1}m_{1,1} + m_{2,2}m_{1,2} & m_{2,1}m_{2,1} + m_{2,2}m_{2,2} & m_{2,1}m_{3,1} + m_{2,2}m_{3,2} \\ m_{3,1}m_{1,1} + m_{3,2}m_{1,2} & m_{3,1}m_{2,1} + m_{3,2}m_{2,2} & m_{3,1}m_{3,1} + m_{3,2}m_{3,2} \end{bmatrix} \end{array}$$

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Association Clusters

- For a given query q, let:
  - D_ℓ: local document set, i.e., set of documents retrieved by q.
  - N_ℓ: number of documents in D_ℓ.
  - V_ℓ: local vocabulary, i.e., set of all distinct words in D_ℓ.
  - $\mathrm{f}_{i,j}$: frequency of occurrence of a term w_i in a document $d_j \in D_\ell$.
  - $\mathbf{M}_\ell = [m_{ij}]$: term-document matrix with V_ℓ rows and N_ℓ columns.
  - $m_{ij} = f_{i,j}$: an element of matrix $\mathbf{M}_\ell$.
  - $\mathbf{M}_\ell^T$: transpose of $\mathbf{M}_\ell$.
- The matrix $\mathbf{C}_\ell$ is a local term-term correlation matrix, given by:

$$\mathbf{C}_\ell = \mathbf{M}_\ell \times \mathbf{M}_\ell^T.$$

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Association Clusters

- Each element $c_{u,v} \in \mathbf{C}_\ell$ expresses a correlation between terms w_u and w_v.
- This relationship between the terms is based on their joint co-occurrences inside documents of the collection.
- Higher the number of documents in which the two terms co-occur, stronger is this correlation.
- Correlation strengths can be used to define local clusters of neighbor terms.
- Terms in a same cluster can then be used for query expansion.
- We consider three types of clusters here:
  1. Association Clusters.
  2. Metric Clusters.
  3. Scalar Clusters.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Association Clusters

- An association cluster is computed from a local correlation matrix $\mathbf{C}_\ell$.
- For that, we re-define the correlation factors $c_{u,v}$ between any pair of terms w_u and w_v, as follows:

$$c_{u,v} = \sum_{d_j \in D_\ell} f_{u,j} \cdot f_{v.j}.$$

- In this case the correlation matrix is referred to as a local association matrix.
- The motivation is that terms that co-occur frequently inside documents have a synonymity association.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

- The correlation factors $c_{u,v}$ and the association matrix C_ℓ are said to be unnormalized.
- An alternative is to normalize the correlation factors:

$$c'_{u,v} = \frac{c_{u,v}}{c_{u,u} + c_{v,v} - c_{u,v}}.$$

- In this case the association matrix $\mathbf{C}_\ell$ is said to be normalized.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Association Clusters

- Given a local association matrix $\mathbf{C}_\ell$, we can use it to build local association clusters as follows.
- Let $C_u(n)$ be a function that returns the n largest factors $c_{u,v} \in \mathbf{C}_\ell$, where v varies over the set of local terms and $v \neq u$.
- Then, $C_u(n)$ defines a local association cluster, a neighborhood, around the term w_u.
- Given a query q, we are normally interested in finding clusters only for the $|q|$ query terms.
- This means that such clusters can be computed efficiently at query time.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Metric Clusters

- Association clusters do not take into account where the terms occur in a document.
- However, two terms that occur in a same sentence tend to be more correlated.
- A metric cluster re-defines the correlation factors $c_{u,v}$ as a function of their distances in documents.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Metric Clusters

- Let $w_u(n,j)$ be a function that returns the n$^{\text{th}}$ occurrence of term w_u in document d_j.
- Further, let $r(w_u(n,j), w_v(m,j))$ be a function that computes the distance between:
    - The n$^{\text{th}}$ occurrence of term w_u in document d_j.
    - The m$^{\text{th}}$ occurrence of term w_v in document d_j.
- We define,

$$c_{u,v} = \sum_{d_j \in D_\ell} \sum_n \sum_m \frac{1}{r(w_u(n,j), w_v(m,j))}.$$

- In this case the correlation matrix is referred to as a local metric matrix.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Metric Clusters

- Notice that if w_u and w_v are in distinct documents we take their distance to be infinity.
- Variations of the above expression for $c_{u,v}$ have been reported in the literature, such as $\frac{1}{r^2(w_u(n,j),w_v(m,j))}$.
- The metric correlation factor $c_{u,v}$ quantifies absolute inverse distances and is said to be unnormalized.
- Thus, the local metric matrix $\mathbf{C}_\ell$ is said to be unnormalized.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Metric Clusters

- An alternative is to normalize the correlation factor.
- For instance,

$$c'_{u,v} = \frac{c_{u,v}}{\text{total number of } [w_u, w_v] \text{ pairs considered}}.$$

- In this case the local metric matrix $\mathbf{C}_\ell$ is said to be normalized.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Scalar Clusters

- The correlation between two local terms can also be defined by comparing the neighborhoods of the two terms.
- The idea is that two terms with similar neighborhoods have some synonymity relationship:
  - In this case we say that the relationship is indirect or induced by the neighborhood.
  - We can quantify this relationship comparing the neighborhoods of the terms through a scalar measure.
  - For instance, the cosine of the angle between the two vectors is a popular scalar similarity measure.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Scalar Clusters

- Let,
  - $\vec{s}_u = \langle c_{u,x_1}, c_{u,x_2}, \ldots, c_{u,x_n} \rangle$: vector of neighborhood correlation values for the term w_u.
  - $\vec{s}_v = \langle c_{v,x_1}, c_{v,x_2}, \ldots, c_{v,x_n} \rangle$: vector of neighborhood correlation values for the term w_v.
- Define,

$$c_{u,v} = \frac{\vec{s}_u \cdot \vec{s}_v}{|\vec{s}_u| \cdot |\vec{s}_v|}.$$

- In this case the correlation matrix $\mathbf{C}_\ell$ is referred to as a local scalar matrix.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Scalar Clusters

- The local scalar matrix $\mathbf{C}_\ell$ is said to be induced by the neighborhood.
- Let $C_u(n)$ be a function that returns the n largest $c_{u,v}$ values in a local scalar matrix $\mathbf{C}_\ell$, $v \neq u$.
- Then, $C_u(n)$ defines a scalar cluster around term w_u.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Neighbor Terms

- Terms that belong to clusters associated to the query terms can be used to expand the original query.
- Such terms are called neighbors of the query terms and are characterized as follows.
- A term w_v that belongs to a cluster $C_u(n)$, associated with another term w_u, is said to be a neighbor of w_u.
- Often, neighbor terms represent distinct keywords that are correlated by the current query context.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Neighbor Terms

- Consider the problem of expanding a given user query q with neighbor terms.
- One possibility is to expand the query as follows.
- For each term $w_u \in q$, select m neighbor terms from the cluster $C_u(n)$ and add them to the query.
- This can be expressed as follows:

$$q_m = q \cup \left\{ w_v | w_v \in C_u(n), w_u \in q \right\}.$$

- Hopefully, the additional neighbor terms w_v, will retrieve new relevant documents.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Neighbor Terms

- The set $C_u(n)$ might be composed of terms obtained using correlation factors normalized and unnormalized.
- Query expansion is important because it tends to improve recall.
- However, the larger number of documents to rank also tends to lower precision.
- Thus, query expansion needs to be exercised with great care and fine tuned for the collection at hand.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Implicit Relevance Feedback

Local Analysis – Local Context Analysis

Local Context Analysis

- The local clustering techniques are based on the set of documents retrieved for a query.
- A distinct approach is to search for term correlations in the whole collection.
- Global techniques usually involve the building of a thesaurus that encodes term relationships in the whole collection.
- The terms are treated as concepts and the thesaurus is viewed as a concept relationship structure.
- The building of a thesaurus usually considers the use of small contexts and phrase structures.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Local Context Analysis

- Local context analysis is an approach that combines global and local analysis.
- It is based on the use of noun groups, i.e., a single noun, two nouns, or three adjacent nouns in the text.
- Noun groups selected from the top ranked documents are treated as document concepts.
- However, instead of documents, passages are used for determining term co-occurrences.
    - Passages are text windows of fixed size.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Local Context Analysis

- Local context analysis procedure operates in three steps:
  - First, retrieve the top n ranked passages using the original query.
  - Second, for each concept c in the passages compute the similarity $\text{sim}(q, c)$ between the whole query q and the concept c.
  - Third, the top m ranked concepts, according to $\text{sim}(q, c)$, are added to the original query q.
- A weight computed as $\left[1 - 0.9 \cdot \frac{i}{m}\right]$ is assigned to each concept c, where:
  - i: position of c in the concept ranking.
  - m: number of concepts to add to q.
- The terms in the original query q might be stressed by assigning a weight equal to 2 to each of them.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Local Context Analysis

- Of these three steps, the second one is the most complex and the one which we now discuss.
- The similarity $\mathrm{sim}(q, c)$ between each concept c and the original query q is computed as follows:

$$\mathrm{sim}(q, c) = \prod_{w_i \in q} \left[\delta + \frac{\log\left[f(c, w_i) \cdot \mathrm{idf}_c \right]}{\log(n)} \right]^{\mathrm{idf}_i}$$

- where, n is the number of top ranked passages considered.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Local Context Analysis

- The function $f(c, w_i)$ quantifies the correlation between the concept c and the query term w_i and is given by:

$$f(c, w_i) = \sum_{j=1}^{n} \mathrm{pf}_{i,j} \cdot \mathrm{pf}_{c,j}.$$

- where,
  - $\mathrm{pf}_{i,j}$ is the frequency of term w_i in the j^{th} passage.
  - $\mathrm{pf}_{c,j}$ is the frequency of the concept c in the j^{th} passage.
- Notice that this is the correlation measure defined for association clusters, but adapted for passages.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Local Context Analysis

- The inverse document frequency factors are computed as:

$$\text{idf}_i = \max\left[1, \frac{\log_{10}\left[\frac{N}{\text{np}_i}\right]}{5}\right]$$

$$\text{idf}_c = \max\left[1, \frac{\log_{10}\left[\frac{N}{\text{np}_c}\right]}{5}\right]$$

- where,
  - N is the number of passages in the collection.
  - np_i is the number of passages containing the term w_i.
  - np_c is the number of passages containing the concept c.
- The idf_i factor in the exponent is introduced to emphasize infrequent query terms.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Local Context Analysis

- The procedure above for computing $\mathrm{sim}(q, c)$ is a non-trivial variant of tf-idf ranking.
- It has been adjusted for operation with TREC data and did not work so well with a different collection.
- Thus, it is important to have in mind that tuning might be required for operation with a different collection.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Global Analysis

- Local analysis methods extract information from the local set of documents retrieved to expand the query.
- An alternative approach is to expand the query using information from the whole set of documents — a strategy usually referred to as global analysis procedures.
- We distinguish two global analysis procedures:
  1. Query expansion based on a similarity thesaurus.
  2. Query expansion based on a statistical thesaurus.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Implicit Relevance Feedback

Global Analysis – Similarity Thesaurus

Query Expansion based on a Similarity Thesaurus

- We now discuss a query expansion model based on a global similarity thesaurus constructed automatically.
- The similarity thesaurus is based on term to term relationships rather than on a matrix of co-occurrence.
- Special attention is paid to the selection of terms for expansion and to the re-weighting of these terms.
- Terms for expansion are selected based on their similarity to the whole query.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Similarity Thesaurus

- A similarity thesaurus is built using term to term relationships.
- These relationships are derived by considering that the terms are concepts in a concept space.
- In this concept space, each term is indexed by the documents in which it appears.
- Thus, terms assume the original role of documents while documents are interpreted as indexing elements.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

- Let,
  - t: number of terms in the collection.
  - N: number of documents in the collection.
  - $f_{i,j}$: frequency of term w_i in document d_j.
  - t_j: number of distinct index terms in document d_j.
- The,

$$\text{itf}_j = \log\left[\frac{t}{t_j}\right]$$

- where, itf_j is the inverse term frequency for document d_j (analogous to inverse document frequency).

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Similarity Thesaurus

- Within this framework, with each term w_i is associated a vector $\vec{w}_i$ given by:

$$\vec{w}_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,N} \rangle$$

- These weights are computed as follows:

$$m_{i,j} = \frac{\left[0.5 + 0.5 \cdot \frac{f_{i,j}}{\max_j(f_{i,j})}\right] \cdot \mathrm{itf}_j}{\sqrt{\sum_{\ell=1}^{N} \left[0.5 + 0.5 \cdot \frac{f_{i,\ell}}{\max_\ell(f_{i,\ell})}\right]^2 \cdot \mathrm{itf}_\ell^2}}.$$

- where, $\max_j(f_{i,j})$ computes the maximum of all $f_{i,j}$ factors for the i^{th} term.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Similarity Thesaurus

- The relationship between two terms w_u and w_v is computed as a correlation factor $c_{u,v}$ given by:

$$c_{u,v} = \vec{w}_u \cdot \vec{w}_v = \sum_{\forall d_j} m_{u,j} \cdot m_{v,j}.$$

- The global similarity thesaurus is given by the scalar term-term matrix composed of correlation factors $c_{u,v}$.
- This global similarity thesaurus has to be computed only once and can be updated incrementally.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Similarity Thesaurus

- Given the global similarity thesaurus, query expansion is done in three steps as follows:
  1. First, represent the query in the same vector space used for representing the index terms.
  2. Second, compute a similarity $\mathrm{sim}(q, w_v)$ between each term w_v correlated to the query terms and the whole query q.
  3. Third, expand the query with the top r ranked terms according to $\mathrm{sim}(q, w_v)$.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Similarity Thesaurus

- For the first step, the query is represented by a vector $\vec{q}$ given by:

$$\vec{q} = \sum_{w_i \in q} m_{i,q} \cdot \vec{w}_i$$

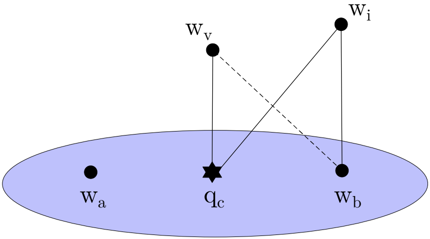
- where, $m_{i,q}$ is a term-query weight computed using the equation for $m_{i,j}$ but with $\vec{q}$ in place of $\vec{d}_j$.
- For the second step, the similarity $\text{sim}(q, w_v)$ is computed as:

$$\text{sim}(q, w_v) = \vec{q} \cdot \vec{w}_v = \sum_{w_i \in q} m_{i,q} \cdot c_{i,v}.$$

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Similarity Thesaurus

- A term w_v might be closer to the whole query centroid q_C than to the individual query terms.
- Thus, terms selected here might be distinct from those selected by previous methods.



All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Similarity Thesaurus

- For the third step, the top r ranked terms are added to the query q to form the expanded query q_m.
- To each expansion term w_v in query q_m is assigned a weight m_{v,q_m} given by:

$$m_{v,q_m} = \frac{\mathrm{sim}(q, w_v)}{\sum_{w_v \in q} m_{i,q}}.$$

- The expanded query q_m is then used to retrieve new documents.
- This technique has yielded improved retrieval performance (in the range of 20%) with three different collection.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Implicit Relevance Feedback

Global Analysis – Statistical Thesaurus

Query Expansion based on a Statistical Thesaurus

- We now discuss a query expansion technique based on a global statistical thesaurus.
- The approach is quite distinct from the one based on a similarity thesaurus.
- The global thesaurus is composed of classes that group correlated terms in the context of the whole collection.
- Such correlated terms can then be used to expand the original user query.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Statistical Thesaurus

- To be effective, the terms selected for expansion must have high term discrimination values.
  - This implies that they must be low frequency terms.
- However, it is difficult to cluster low frequency terms due to the small amount of information about them.
- To circumvent this problem, documents are clustered into classes.
- The low frequency terms in these documents are then used to define thesaurus classes.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

- A document clustering algorithm that produces small and cohesive clusters is the complete link algorithm:
  1. Initially, place each document in a distinct cluster.
  2. Compute the similarity between all pairs of clusters.
  3. Determine the pair of clusters $[C_u, C_v]$ with the highest inter-cluster similarity.
  4. Merge the clusters C_u and C_v.
  5. Verify a stop criterion (if this criterion is not met then go back to step 2).
  6. Return a hierarchy of clusters.

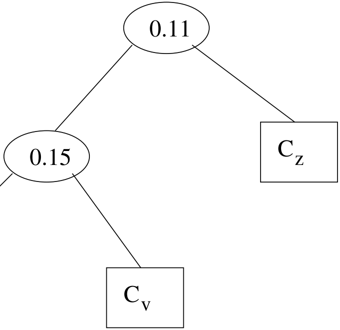
All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

- The similarity between two clusters is defined as the minimum of the similarities between two documents not in the same cluster.
- To compute the similarity between documents in a pair, the cosine formula of the vector model is used.
- As a result of this minimality criterion, the resultant clusters tend to be small and cohesive.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Statistical Thesaurus

- Consider that the whole document collection has been clustered using the complete link algorithm.
- Figure below illustrates a portion of the whole cluster hierarchy generated by the complete link algorithm where the inter-cluster similarities are shown in the ovals.



All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Statistical Thesaurus

- The terms that compose each class of the global thesaurus are selected as follows.
- Obtain from the user three parameters:
  - TC: threshold class.
  - NDC: number of documents in a class.
  - MIDF: minimum inverse document frequency.
- Parameter TC determines the document clusters that will be used to generate thesaurus classes:
  - Two clusters C_u and C_v are selected, when TC is surpassed by $\text{sim}(C_u, C_v)$.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

- Use NDC as a limit on the number of documents of the clusters:
  - For instance, if both C_{u+v} and C_{u+v+z} are selected then the parameter NDC might be used to decide between the two.
- MIDF defines the minimum value of IDF for any term which is selected to participate in a thesaurus class.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Statistical Thesaurus

- Given that the thesaurus classes have been built, they can be used for query expansion.
- For this, an average term weight wt_C for each thesaurus class C is computed as follows:

$$\mathrm{wt}_C = \frac{\sum_{i=1}^{|C|} m_{i,C}}{|C|}.$$

- where,
  - $|C|$ is the number of terms in the thesaurus class C.
  - $m_{i,C}$ is a weight associated with term-class pair $[w_i, C]$.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Statistical Thesaurus

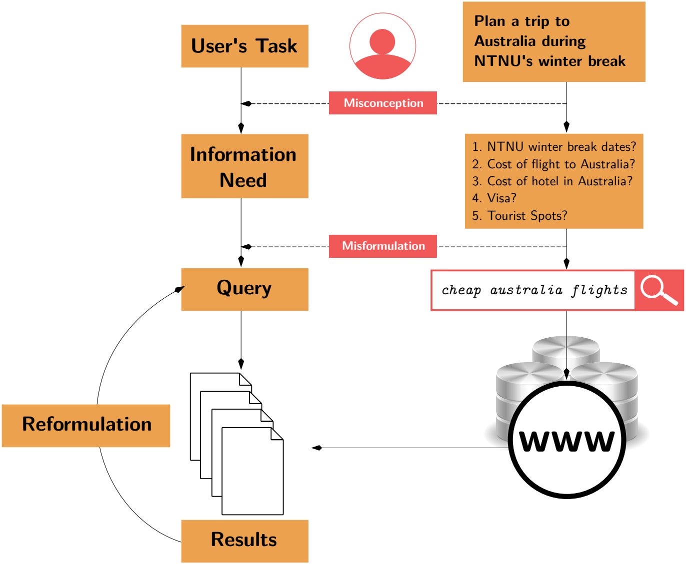
- This average term weight can then be used to compute a thesaurus class weight m_C as:

$$m_C = \frac{\mathrm{wt}_C}{|C|} \cdot 0.5.$$

- The above weight formulations have been verified through experimentation and have yielded good results.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Information Retrieval — Formalization



Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Modeling Documents — Term Document Matrix

	w_1	w_2	w_3	w_4	w_5	$\cdots$	$w_{\|\mathcal{V}\|}$
d_1	1	0	1	1	0	$\cdots$	1
d_2	1	1	0	0	1	$\cdots$	0
d_3	0	0	0	1	0	$\cdots$	0
d_4	0	1	0	0	1	$\cdots$	1
d_5	0	0	1	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	1	1	0	1	1	$\cdots$	0

Boolean Retrieval

- Consider a Boolean query:
  $q = w_1 \wedge (w_2 \vee \neg w_3)$.
- Term vector for $w_1 = \langle 1, 1, 0, 0, 0 \rangle$.
- Term vector for $w_2 = \langle 0, 1, 0, 1, 0 \rangle$.
- Term vector for $w_3 = \langle 1, 0, 0, 0, 1 \rangle$.

	1	2	3	4	5
w_3	1	0	0	0	1
$\neg w_3$	0	1	1	1	0

	1	2	3	4	5
$\neg w_3$	0	1	1	1	0
w_2	0	1	0	1	0
OR	0	1	1	1	0

	1	2	3	4	5
$w_2 \vee \neg w_3$	0	1	1	1	0
w_1	1	1	0	0	0
AND	0	1	0	0	0

	w_1	w_2	w_3	w_4	w_5
d_1	1	0	1	1	0
d_2	1	1	0	0	1
d_3	0	0	0	1	0
d_4	0	1	0	0	1
d_5	0	0	1	0	0

Weighted Term Vectors — TF-IDF

- The best known term weighting schemes use weights that combine idf factors with term frequencies.
- TF-IDF weight of term t in document d:

$$m_{d,t} = \left[1 + \log\left[tf_{d,t}\right]\right] \cdot \left[\log\left[\frac{N}{\mathrm{df}_t}\right]\right].$$

- Relevance of a document to a query:

$$\mathrm{score}(q, d) = \sum_{t \in q \cap d} \left[1 + \log\left[tf_{d,t}\right]\right] \cdot \left[\log\left[\frac{N}{\mathrm{df}_t}\right]\right].$$

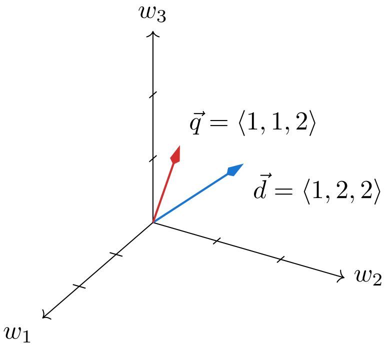
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

The Vector Space Model

- They are represented as unit vectors of a $|\mathcal{V}|$-dimensionsal space.
- The representations of document d and query q are $|\mathcal{V}|$-dimensional vectors given by:

$$\vec{d}_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,|\mathcal{V}|} \rangle$$
$$\vec{q} = \langle m_{q,1}, m_{q,2}, \ldots, m_{q,|\mathcal{V}|} \rangle$$



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

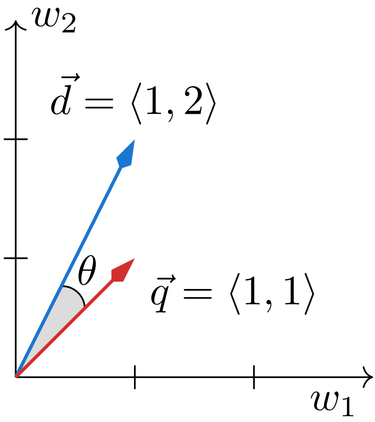
The Vector Space Model

- Similarity between a document and query $\text{sim}(\vec{d}, \vec{q})$ is equated to its cosine-similarity:

$$\vec{d}_i \cdot \vec{q} = |\vec{d}_i| \cdot |\vec{q}| \cdot \cos(\theta)$$

$$\cos(\theta) = \frac{\vec{d}_i \cdot \vec{q}}{|\vec{d}_i| \cdot |\vec{q}|}$$

$$\cos(\theta) = \frac{\sum_{j=1}^{|\mathcal{V}|} m_{i,j} \cdot m_{q,j}}{\sqrt{\sum_{j=1}^{|\mathcal{V}|} m_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{|\mathcal{V}|} m_{i,q}^2}}$$



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Probability Ranking Principle

"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."

— van Rijsbergen, 1979.

All from: Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

The Binary Independence Model

- Binary Independence Model Assumption 2: $\forall w_j \notin q, p_{iR} = q_{iR}$.
- Converting log products into sums of logs, we have

$$\text{sim}(d_i, q) \sim \sum_{w_j \in q \wedge w_j \in d_i} \log\left[\frac{p_{iR}}{1 - p_{iR}}\right] + \log\left[\frac{1 - q_{iR}}{q_{iR}}\right].$$

- The above formula is a key expression for ranking computation in the probabilistic model.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Ranking Formula

- In the previous formula, we are still dependent on an estimation of the relevant docs for the query.
- For handling small values of r_j, we add 0.5 to each of the terms in the formula above, which changes $\mathrm{sim}(d_i, q)$ into

$$\mathrm{sim}(d_i, q) \sim \sum_{w_j \in q \wedge w_j \in d_i} \log\left(\frac{r_j + 0.5}{R - r_j + 0.5} \cdot \frac{N - n_j - R + r_j + 0.5}{n_j - r_j + 0.5} \right).$$

- This formula is considered as the classic ranking equation for the probabilistic model and is known as the Robertson-Sparck Jones Equation.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Probabilistic Model

- The probabilistic ranking formula:

$$\mathrm{sim}(d_i, q) \sim \sum_{w_j \in q \wedge w_j \in d_i} \log\left[\frac{N - n_j}{n_j}\right].$$

- To avoid problems with $D = 1$ and $D_j = 0$:

$$p_{jR} = \frac{D_j + 0.5}{D + 1} \text{ and } q_{jR} = \frac{n_j - D_j + 0.5}{N - D + 1}.$$

- Also,

$$p_{jR} = \frac{D_j + \frac{n_j}{N}}{D + 1} \text{ and } q_{jR} = \frac{n_j - D_j + \frac{n_j}{N}}{N - D + 1}.$$

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

BM25 Ranking Formula

- BM25: combination of the BM11 and BM15.
- The motivation was to combine the BM11 and BM15 term frequency factors as follows.

$$\mathcal{B}_{i,j} = \frac{(K_1 + 1) \cdot \mathrm{tf}_{i,j}}{K_1 \cdot \left[(1 - b) + b \cdot \frac{\mathrm{len}(d_i)}{\mathrm{avg\_doclen}} \right] + \mathrm{tf}_{i,j}}.$$

- where, b is a constant with values in the interval $[0, 1]$.
  - If $b = 0$, it reduces to the BM15 term frequency factor.
  - If $b = 1$, it reduces to the BM11 term frequency factor.
  - For values of $b \in (0, 1)$, the equation provides a combination of BM11 and BM15.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

BM25 Ranking Formula

- The ranking equation for the BM25 model can then be written as:

$$\text{sim}_{BM25}(d_i, q) \sim \sum_{w_j \in q \wedge w_j \in d_i} \mathcal{B}_{i,j} \cdot \log\left[\frac{N - n_j + 0.5}{n_j + 0.5}\right]$$

- where, K_1 and b are empirical constants.
  - $K_1 = 1$ works well with real collections.
  - b should be kept closer to 1 to emphasize the document length normalization effect present in the BM11 formula.
  - For instance, $b = 0.75$ is a reasonable assumption.
  - Constants values can be fine tuned for particular collections through proper experimentation.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Statistical Language Models in IR

- Each document is treated as (the basis for) a language model.
- Given a query q.
- Rank documents based on $P(d|q)$.

$$P(d|q) = \frac{P(q|d) \cdot P(d)}{P(q)} \propto P(q|d) \cdot P(d)$$

- $P(q)$ is the same for all documents, so ignore.
- $P(d)$ is the prior – often treated as the same for all d.
  - But we can give a higher prior to "high-quality" documents, e.g., those with high PageRank.
- $P(q|d)$ is the probability of q given d.
- Under the assumptions we made, ranking documents according to $P(q|d) \cdot P(d)$ or $P(d|q)$ is considered equivalent.

All from: Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Jelinek-Mercer Smoothing

- Jelinek-Mercer Smoothing:

$$P(q|d) \propto \prod_{1 \leq k \leq |q|} \lambda \cdot P(t_k|M_d) + (1 - \lambda) \cdot P(t_k|M_c)$$

- What we model: the user has a document in mind and generates the query from this document.
- $P(q|d)$ is the probability that the document that the user had in mind was in fact this one.

All from: Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Dirichlet Smoothing

- Dirichlet Smoothing:

$$P(t|d) = \frac{\mathrm{tf}_{d,t} + \mu \cdot P(t|M_c)}{|d| + \mu}$$

- The background distribution $P(t|M_c)$ is the prior for $P(t|d)$.
- Intuition: before having seen any part of the document we start with the background distribution as our estimate.
- As we read the document and count terms we update the background distribution.
- The weighting factor μ determines how strong an effect the prior has.

All from: Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Relevance — Query vs. Information Need

- User satisfaction is equated with the relevance of search results to the query.
- Information Need i: *I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
- This is an information need, not a query.
- Query q: *red wine white wine heart attack*.
- Consider document d': *At the heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.*
- d' is an good match for query q.
- d' is not relevant to the information need i.

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

The Cranfield Paradigm

- The next step was to devise a set of experiments that would allow evaluating each indexing system in isolation more thoroughly.
- The result was a test reference collection (Cranfield-2 collection) composed of:
  1. Documents,
  2. Queries, and
  3. Relevance Judgements.
- The reference collection allows using the same set of documents and queries to evaluate different ranking systems.
- The uniformity of this setup allows quick evaluation of new ranking functions.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Reference Collections

- Reference collections, which are based on the foundations established by the Cranfield experiments, constitute the most used evaluation method in IR.
- A reference collection is composed of:
  1. A set $\mathcal{D}$ of pre-selected documents.
  2. A set $\mathcal{I}$ of information need descriptions used for testing.
  3. A set of relevance judgements associated with each pair $[i, d]$, where, $i \in \mathcal{I}$ and $d \in \mathcal{D}$.
- The relevance judgement has a value of:
  - 0 if document d is non-relevant to i.
  - 1 if document d is relevant to i.
- These judgements are produced by human specialists.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision and Recall

	Relevant	**Non-Relevant**
Retrieved	TP (True Positive)	FP (False Positive)
Not Retrieved	FN (False Negative)	TN (True Negative)

$$\text{Precision} = P = \frac{TP}{TP + FP}$$

$$\text{Recall} = R = \frac{TP}{TP + FN}$$

Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

Combining Precision and Recall: F-measure

- F-measure allows us to trade off precision against recall.

$$F = \frac{1}{\alpha \cdot \frac{1}{P} + (1-\alpha) \cdot \frac{1}{R}} = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}, \qquad \text{where } \beta^2 = \frac{1-\alpha}{\alpha}.$$

- $\alpha \in [0, 1]$ and therefore $\beta^2 \in [0, \infty]$.
- Usually used: balanced F-measure with $\alpha = 0.5$ or $\beta = 1$. This correspond to the harmonic mean of precision and recall.

$$\frac{1}{F} = \frac{1}{2} \cdot \left(\frac{1}{P} + \frac{1}{R}\right) = \frac{2 \cdot P \cdot R}{P + R}.$$

- The F-measure is derived from the E-measure with the following equivalence:

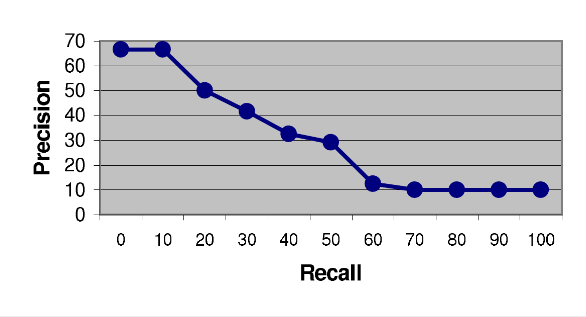
$$E = 1 - F = 1 - \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}.$$

Manning et al., “Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

Schütze et al.: `https://www.cis.lmu.de/~hs/teach/14s/ir/`.

Precision and Recall

- To illustrate, the figure below illustrates precision-recall figures averaged over queries q_1 and q_2:



Recall	Precision
0	66.6
10	66.6
20	49.9
30	41.6
40	32.5
50	29.1
60	12.5
70	10
80	10
90	10
100	10

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Precision at k: $P@K$

- To exemplify, consider again the ranking for the example query q_1 we have been using.
- For this query, we have $P@5 = 40\%$ and $P@10 = 40\%$.
- Further, we can compute $P@5$ and $P@10$ averaged over a sample of 100 queries, for instance.
- These metrics provide an early assessment of which algorithm might be preferable in the eyes of the users.

01. d_{123} •	06. d_9 •	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} •	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} •	15. d_3 •

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Mean Average Precision: MAP

- The idea here is to average the precision figures obtained after each new relevant document is observed.
- For relevant documents not retrieved, the precision is set to 0.
- MAP_i: the mean value precision for query q_i is:

$$\text{MAP}_i = \frac{1}{|R_i|} \cdot \sum_{k=1}^{|R_i|} P(R_i[k]).$$

- where, R_i is the set of relevant documents for query q_i.
- where, $P(R_i[k])$ is the precision when the $R_i[k]$ document is observed in the ranking of q_i.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Mean Average Precision: MAP

- MAP: the mean average precision over a set of queries, is defined as:

$$\text{MAP} = \frac{1}{N_q} \cdot \sum_{i=1}^{N_q} \text{MAP}_i.$$

- where, N_q is the total number of queries.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

R-Precision

- Let R be the total number of relevant docs for a given query.
- The idea here is to compute the precision at the R–th position in the ranking.
- Example: consider query q_1,
  - The R value is 10 and there are 4 relevant documents among the top-10 documents in the ranking.
  - Thus, the R-Precision value for q_1 is $\frac{4}{10} = 0.4$.
- Example: consider query q_2,
  - The R value is 3 and there is 1 relevant document among the top-3 documents in the ranking.
  - Thus, the R-Precision value for q_2 is $\frac{1}{3} = 0.\overline{3}$.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Mean Reciprocal Rank: MRR

- Let,
  - $\mathcal{R}_i$: ranking relative to a query q_i.
  - $S_{\text{correct}}(\mathcal{R}_i)$: position of the first correct answer in $\mathcal{R}_i$.
  - S_h: threshold for ranking position.
- Then, the reciprocal rank $\text{RR}(\mathcal{R}_i)$ for query q_i is given by:

$$\text{RR}(\mathcal{R}_i) = \begin{cases} \frac{1}{S_{\text{correct}}(\mathcal{R}_i)}, & \text{if } S_{\text{correct}}(\mathcal{R}_i) \leq S_h \\ 0, & \text{otherwise} \end{cases}$$

- The mean reciprocal rank (MRR) for a set Q of N_q queries is given by:

$$\text{MRR}(Q) = \frac{1}{N_q} \cdot \sum_{i}^{N_q} \text{RR}(\mathcal{R}_i).$$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

- The coverage ratio is defined as the fraction of the documents known and relevant that are in the answer set, that is:

$$\text{coverage} = \frac{|K \cap R \cap A|}{|K \cap R|}.$$

- A high coverage indicates that the system has found most of the relevant docs the user expected to see.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

User-Oriented Measures

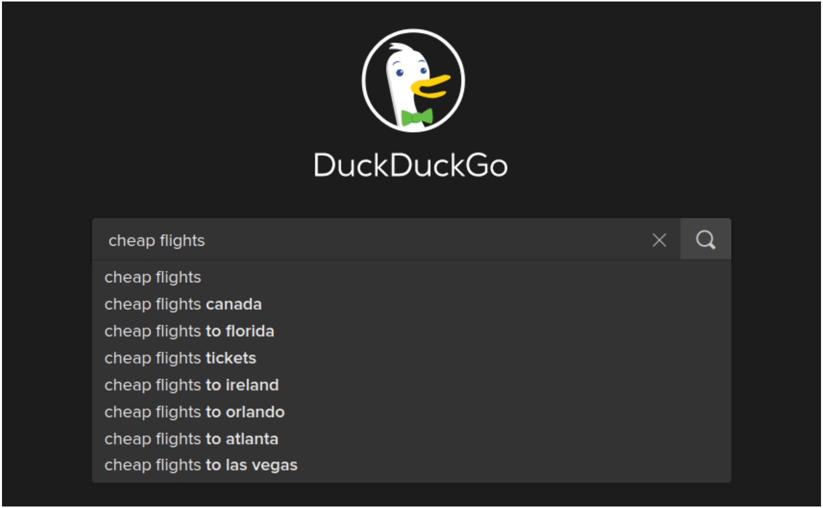
- The novelty ratio is defined as the fraction of the relevant documents in the answer set that are not known to the user, that is:

$$\text{novelty} = \frac{|(R \cap K) - A|}{|R \cap A|}.$$

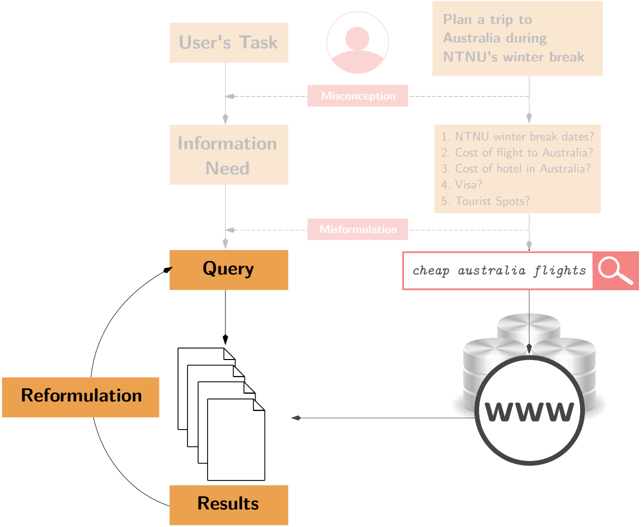
- A high novelty indicates that the system is revealing many new relevant docs which were unknown.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

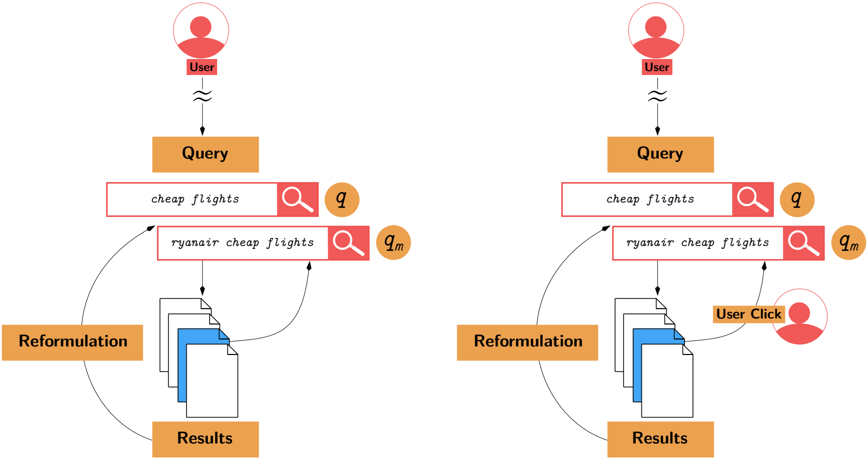
Introduction



Introduction

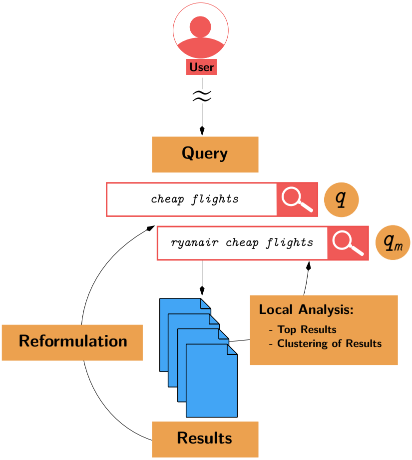


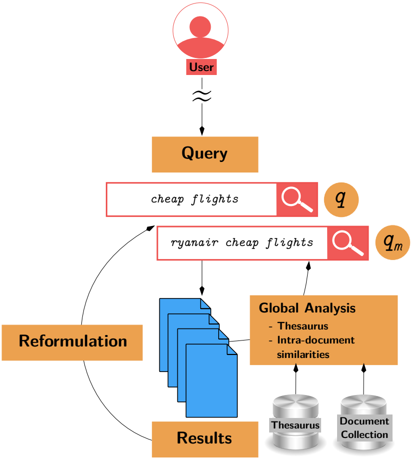
Framework — Explicit Feedback Information



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Framework — Implicit Feedback Information





Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Rocchio Method

- There are three classic and similar ways to calculate the modified query $\vec{q}_m$ as follows,

$$
\begin{aligned}
\text{Standard Rocchio} &: \vec{q}_m = \alpha \cdot \vec{q} + \frac{\beta}{N_r} \cdot \sum_{\forall \vec{d}_i \in D_r} \vec{d}_i - \frac{\gamma}{|N_n|} \cdot \sum_{\forall \vec{d}_i \notin D_n} \vec{d}_i \\
\text{Ide Regular} &: \vec{q}_m = \alpha \cdot \vec{q} + \beta \cdot \sum_{\forall \vec{d}_i \in D_r} \vec{d}_i - \gamma \cdot \sum_{\forall \vec{d}_i \notin D_n} \vec{d}_i \\
\text{Ide Dec Hi} &: \vec{q}_m = \alpha \cdot \vec{q} + \beta \cdot \sum_{\forall \vec{d}_i \in D_r} \vec{d}_i - \gamma \cdot \text{maxrank}(D_n)
\end{aligned}
$$

  - where, $\text{maxrank}(D_n)$ is the highest ranked non-relevant document.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Probabilistic Model

- Let $n_{r,j}$ be the number of documents in set D_r that contain the term w_j.
- Then, the probabilities $P(w_j|R)$ and $P(w_j|\bar{R})$ can be approximated by:

$$P(w_j|R) = \frac{n_{r,j}}{N_r} \qquad\qquad P(w_j|\bar{R}) = \frac{n_j - n_{r,j}}{N - N_r}$$

- Using these approximations, the similarity equation can be rewritten as:

$$\mathrm{sim}(d_i, q) = \sum_{w_j \in q \wedge w_j \in d_i} \left[\log\left[\frac{n_{r,j}}{N_r - n_{r,j}} \right] + \log\left[\frac{N - N_r - (n_j - n_{r,j})}{n_j - n_{r,j}} \right] \right].$$

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Term-Term Correlation Matrix

- Term-Term correlation matrix for a sample collection.

$$\mathbf{M} \times \mathbf{M}^T = \begin{array}{cc} & \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{array}{c} w_1 \\ w_2 \\ w_3 \end{array} & \begin{bmatrix} m_{1,1} & m_{1,2} \\ m_{2,1} & m_{2,2} \\ m_{3,1} & m_{3,2} \end{bmatrix} \end{array} \times \begin{array}{cc} & \begin{array}{ccc} w_1 & w_2 & w_3 \end{array} \\ \begin{array}{c} d_1 \\ d_2 \end{array} & \begin{bmatrix} m_{1,1} & m_{2,1} & m_{3,1} \\ m_{1,2} & m_{2,2} & m_{3,2} \end{bmatrix} \end{array}$$

$$\mathbf{M} \times \mathbf{M}^T = \begin{array}{cc} & \begin{array}{ccc} w_1 & w_2 & w_3 \end{array} \\ \begin{array}{c} w_1 \\ w_2 \\ w_2 \end{array} & \begin{bmatrix} m_{1,1}m_{1,1} + m_{1,2}m_{1,2} & m_{1,1}m_{2,1} + m_{1,2}m_{2,2} & m_{1,1}m_{3,1} + m_{1,2}m_{3,2} \\ m_{2,1}m_{1,1} + m_{2,2}m_{1,2} & m_{2,1}m_{2,1} + m_{2,2}m_{2,2} & m_{2,1}m_{3,1} + m_{2,2}m_{3,2} \\ m_{3,1}m_{1,1} + m_{3,2}m_{1,2} & m_{3,1}m_{2,1} + m_{3,2}m_{2,2} & m_{3,1}m_{3,1} + m_{3,2}m_{3,2} \end{bmatrix} \end{array}$$

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Association Clusters

- Each element $c_{u,v} \in \mathbf{C}_\ell$ expresses a correlation between terms w_u and w_v.
- This relationship between the terms is based on their joint co-occurrences inside documents of the collection.
- Higher the number of documents in which the two terms co-occur, stronger is this correlation.
- Correlation strengths can be used to define local clusters of neighbor terms.
- Terms in a same cluster can then be used for query expansion.
- We consider three types of clusters here:
  1. Association Clusters.
  2. Metric Clusters.
  3. Scalar Clusters.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Local Context Analysis

- Local context analysis procedure operates in three steps:
  - First, retrieve the top n ranked passages using the original query.
  - Second, for each concept c in the passages compute the similarity $\mathrm{sim}(q, c)$ between the whole query q and the concept c.
  - Third, the top m ranked concepts, according to $\mathrm{sim}(q, c)$, are added to the original query q.
- A weight computed as $\left[1 - 0.9 \cdot \frac{i}{m}\right]$ is assigned to each concept c, where:
  - i: position of c in the concept ranking.
  - m: number of concepts to add to q.
- The terms in the original query q might be stressed by assigning a weight equal to 2 to each of them.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Similarity Thesaurus

- Given the global similarity thesaurus, query expansion is done in three steps as follows:
  1. First, represent the query in the same vector space used for representing the index terms.
  2. Second, compute a similarity $\text{sim}(q, w_v)$ between each term w_v correlated to the query terms and the whole query q.
  3. Third, expand the query with the top r ranked terms according to $\text{sim}(q, w_v)$.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Statistical Thesaurus

- A document clustering algorithm that produces small and tight clusters is the complete link algorithm:
  1. Initially, place each document in a distinct cluster.
  2. Compute the similarity between all pairs of clusters.
  3. Determine the pair of clusters $[C_u, C_v]$ with the highest inter-cluster similarity.
  4. Merge the clusters C_u and C_v.
  5. Verify a stop criterion (if this criterion is not met then go back to step 2).
  6. Return a hierarchy of clusters.

All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.