Information Retrieval

Dhruv Gupta

dhruv.gupta@ntnu.no

20-September-2022

NTNU | Norwegian University of Science and Technology

1. Administrative
   - Announcements
   - References
   - Q&A
2. Evaluation
   - Single Value Summaries
     - Precision at K
     - Mean Average Precision
     - R-Precision
     - Mean Reciprocal Rank
   - User-Oriented Measures
3. Query Expansion
   - Introduction
   - Framework
   - Explicit Relevance Feedback
     - The Rocchio Method
     - The Probabilistic Model
   - Implicit Relevance Feedback
     - Local Analysis
     - Global Analysis

Announcements

- Assignment 1: due (22.September.2021).
- Assignment 2: will be published on BlackBoard this week.
- Reference Group: volunteers needed for feedback regarding course.
    - Interested? Please contact me by email!

References

- 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.

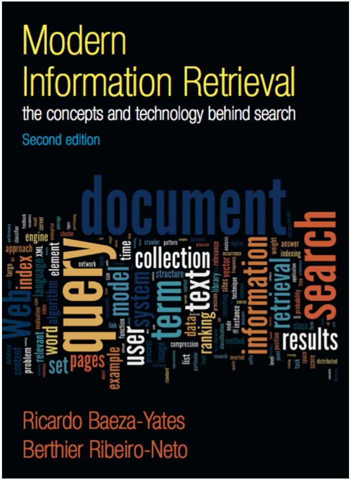


Image Credit: `http://grupoweb.upf.es/mir2ed/`

Precision and Recall

- Consider a reference collection and a set of test queries.
- Let R_{q_1} be the set of relevant docs for a query q_1:

$$R_{q_1} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}.$$

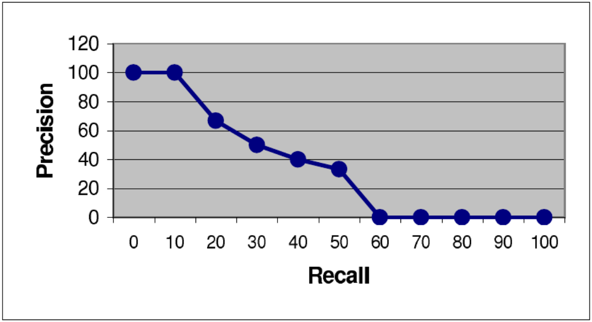
- Consider a new IR algorithm that yields the following answer to q_1 (relevant docs are marked with a bullet):

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.

Precision and Recall

- If we proceed with our examination of the ranking generated, we can plot a curve of precision versus recall as follows:
- Note that precision at recall levels greater 50% is zero because not all the relevant documents are retrieved.



Recall	Precision
0	100
10	100
20	66.6
30	50
40	40
50	33.3
60	0
70	0
80	0
90	0
100	0

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

1. Administrative
   - Announcements
   - References
   - Q&A

2. Evaluation
   - Single Value Summaries
     - Precision at K
     - Mean Average Precision
     - R-Precision
     - Mean Reciprocal Rank
   - User-Oriented Measures

3. Query Expansion
   - Introduction
   - Framework
   - Explicit Relevance Feedback
     - The Rocchio Method
     - The Probabilistic Model
   - Implicit Relevance Feedback
     - Local Analysis
     - Global Analysis

Single Value Summaries

- Average precision-recall curves constitute standard evaluation metrics for information retrieval systems.
- However, there are situations in which we would like to evaluate retrieval performance over individual queries.
- The reasons are two-fold:
  - First, averaging precision over many queries might disguise important anomalies in the retrieval algorithms under study.
  - Second, we might be interested in investigating whether a algorithm outperforms the other for each query.
- In these situations, a single precision value can be used.

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

Precision at k: $P@K$

- In the case of Web search engines, the majority of searches does not require high recall.
- Higher the number of relevant documents at the top of the ranking, more positive is the impression of the users.
- Precision at 5 ($P@5$) and at 10 ($P@10$) measure the precision when 5 or 10 documents have been seen.
- These metrics assess whether the users are getting relevant documents at the top of the ranking or not.

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:

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

- where, N_q is the total number of queries.

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

Mean Average Precision: MAP

- To illustrate, consider again the ranked list of documents returned for the example query q_1.

$$R_{q_1} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}.$$

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 •

$$\mathrm{MAP}_1 = \frac{1 + 0.66 + 0.5 + 0.4 + 0.33 + 0 + 0 + 0 + 0 + 0}{10} = 0.28.$$

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

Mean Average Precision: MAP

- To illustrate, consider again the ranked list of documents returned for the example query q_2.

$$R_{q_2} = \{d_3, d_{56}, d_{129}\}.$$

01. d_{425}	06. d_{615}	11. d_{193}
02. d_{87}	07. d_{512}	12. d_{715}
03. d_{56} •	08. d_{129} •	13. d_{810}
04. d_{32}	09. d_4	14. d_5
05. d_{124}	10. d_{130}	15. d_3 •

$$\mathrm{MAP}_2 = \frac{0.33 + 0.25 + 0.20}{3} = 0.26,$$

$$\mathrm{MAP} = \frac{\mathrm{MAP}_1 + \mathrm{MAP}_2}{2} = \frac{0.28 + 0.26}{2} = 0.27.$$

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.

R-Precision

- The R-precision measure is a useful for observing the behavior of an algorithm for individual queries.
- Additionally, one can also compute an average R-precision figure over a set of queries.
    - However, using a single number to evaluate a algorithm over several queries might be quite imprecise.

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

Mean Reciprocal Rank: MRR

- MRR is a good metric for those cases in which we are interested in the first correct answer such as:
  - Question-Answering (QA) systems.
  - Search engine queries that look for specific sites:
    - URL Queries.
    - Homepage queries.

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.

Mean Reciprocal Rank: MRR

- To illustrate, consider again the ranked list of documents returned for the example query q_1.

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 •

$$\mathrm{RR}_1 = \frac{1}{1} = 1.$$

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

Mean Reciprocal Rank: MRR

- To illustrate, consider again the ranked list of documents returned for the example query q_2.

01. d_{425}	06. d_{615}	11. d_{193}
02. d_{87}	07. d_{512}	12. d_{715}
03. d_{56} •	08. d_{129} •	13. d_{810}
04. d_{32}	09. d_{4}	14. d_{5}
05. d_{124}	10. d_{130}	15. d_{3} •

$$\mathrm{RR}_2 = \frac{1}{3} = 0.\overline{3},$$

$$\mathrm{MRR} = \frac{\mathrm{RR}_1 + \mathrm{RR}_2}{2} = \frac{1 + \frac{1}{3}}{2} = \frac{2}{3} = 0.\overline{6}.$$

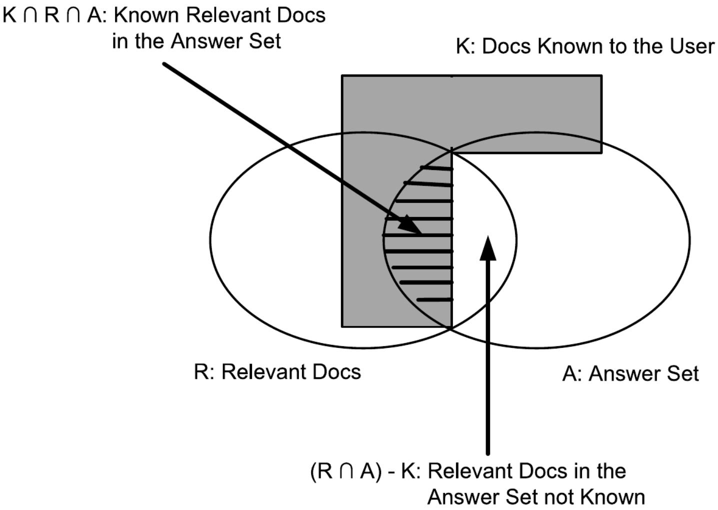
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

User-Oriented Measures

- Recall and precision assume that the set of relevant docs for a query is independent of the users.
- However, different users might have different relevance interpretations.
- To cope with this problem, user-oriented measures have been proposed.
- As before,
  - Consider a reference collection, an information request I, and a retrieval algorithm to be evaluated.
  - with regard to I, let R be the set of relevant documents and A be the set of answers retrieved.

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

User-Oriented Measures



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

User-Oriented Measures

- 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.

User-Oriented Measures: Additional Measures

- Relative Recall: ratio between the number of relevant docs found and the number of relevant docs the user expected to find.
- Recall Effort: ratio between the number of relevant docs the user expected to find and the number of documents examined in an attempt to find the expected relevant documents.

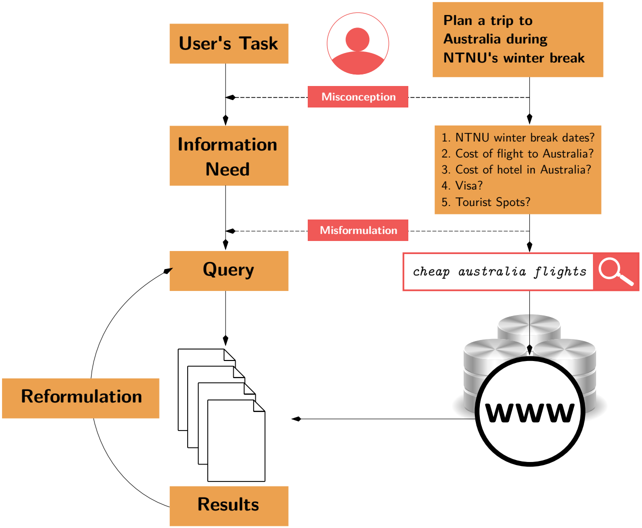
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

1. Administrative
   - Announcements
   - References
   - Q&A

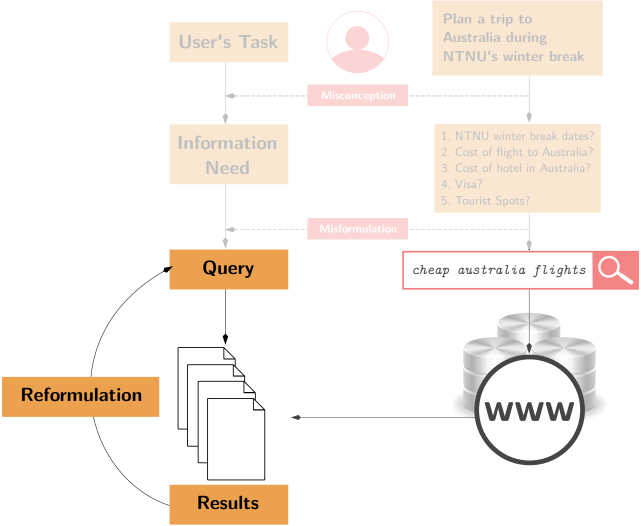
2. Evaluation
   - Single Value Summaries
     - Precision at K
     - Mean Average Precision
     - R-Precision
     - Mean Reciprocal Rank
   - User-Oriented Measures

3. Query Expansion
   - Introduction
   - Framework
   - Explicit Relevance Feedback
     - The Rocchio Method
     - The Probabilistic Model
   - Implicit Relevance Feedback
     - Local Analysis
     - Global Analysis

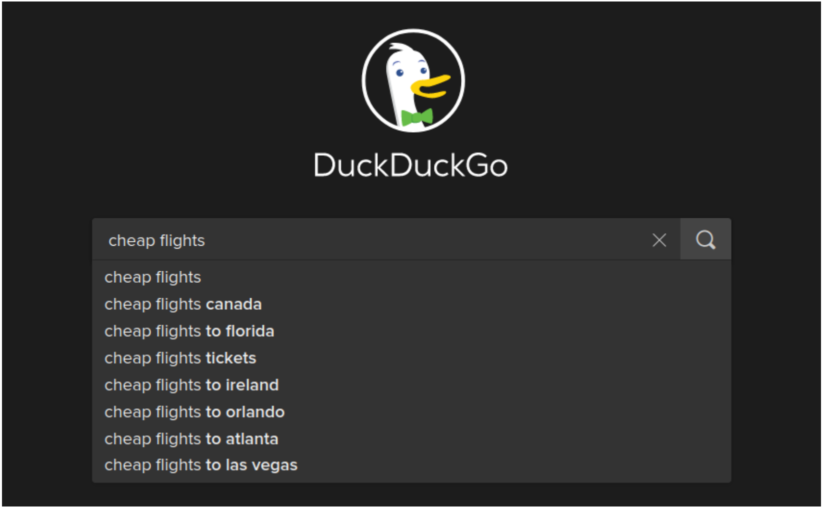
Introduction



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.

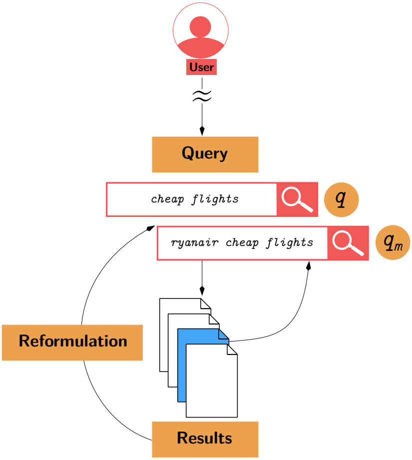
All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

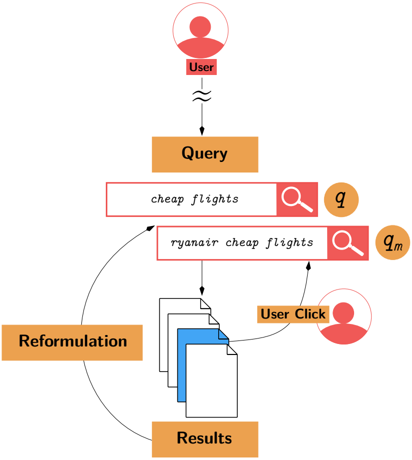
Framework

- 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





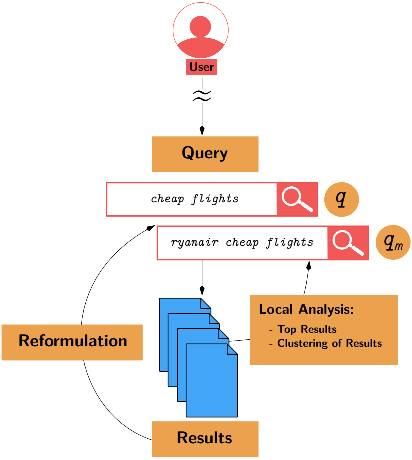
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

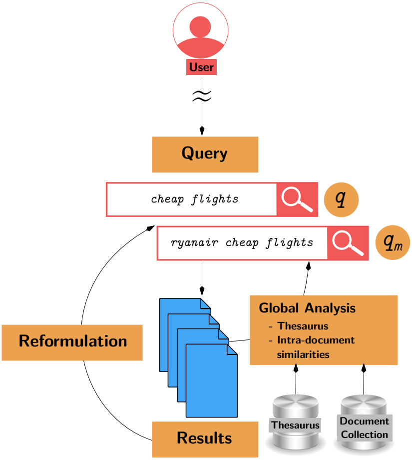
Framework — Implicit Feedback Information

- 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:

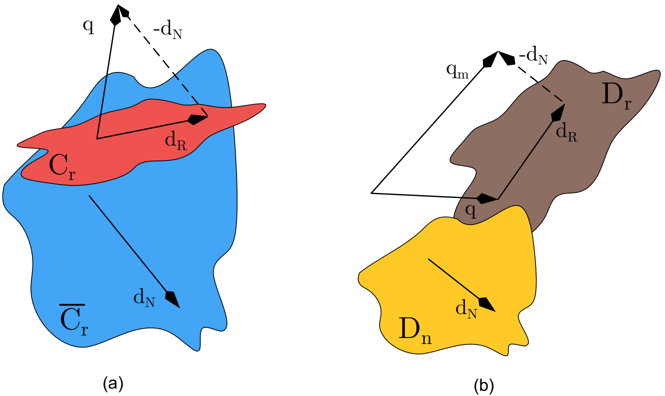
$$\vec{q}_{\text{opt}} = \frac{1}{|C_r|} \cdot \sum_{\forall \vec{d}_i \in C_r} \vec{d}_i - \frac{1}{N - |C_r|} \cdot \sum_{\forall \vec{d}_i \notin C_r} \vec{d}_i$$

  - $|C_r|$ 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.



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,

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

- 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 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:

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

- 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.

1. Administrative
   - Announcements
   - References
   - Q&A
2. Evaluation
   - Single Value Summaries
     - Precision at K
     - Mean Average Precision
     - R-Precision
     - Mean Reciprocal Rank
   - User-Oriented Measures
3. **Query Expansion**
   - Introduction
   - Framework
   - Explicit Relevance Feedback
     - The Rocchio Method
     - The Probabilistic Model
   - **Implicit Relevance Feedback**
     - Local Analysis
     - Global Analysis

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.

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:

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

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.

Association Clusters

- 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.

- 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.

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 $\mathrm{sim}(q, c)$ 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.

Local Context Analysis

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

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

- 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 $\text{sim}(q, c)$ 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.

Local 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.

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.

Similarity Thesaurus

- 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,

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

- 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:

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

- 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 $\text{sim}(q, w_v)$ 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.

Similarity Thesaurus

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

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

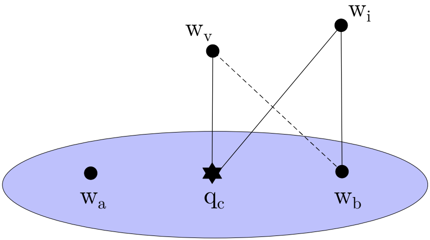
- where, $m_{i,q}$ is a term-query weight computed using the equation for $m_{i,j}$ but with $\vec{q}$ in place of $\vec{d}_j$.
- For the second step, the similarity $\mathrm{sim}(q, w_v)$ is computed as:

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

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 global analysis 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:

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

- 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.

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.

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.

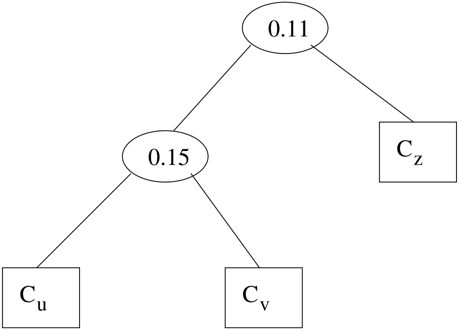
All from: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Statistical Thesaurus

- 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 tight.

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

- 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.

Statistical Thesaurus

- 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.

- 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.