QUERY LANGUAGE

- A query is the formulation of a user information need.
- Most query languages try to use the content (semantics) and the structure of the text (syntax), to find relevant documents.
- The retrieved unit is the basic element which can be retrieved as an answer to a query.
- The retrieval unit can be a file, a document, a webpage, etc.
  - Keyword Based Queries
    - Single Word Queries
    - Context Queries
    - Boolean Queries
    - Natural Languages
  - Pattern Matching
  - Structural Queries

1. KEYWORD BASED QUERY

- A query is composed of keywords and the documents containing such keyword are searched for.
  - Single Word Queries
    - The most elementary query that can be formulated in text retrieval is a word.
    - Documents are assumed to be long sequences of words.
    - Word is a sequence of letters surrounded by separators.
    - Some characters are not letters but do not split a word. For eg: hyphen (co-education).
    - The result of word queries is the set of documents containing at least one of the words of the query.
    - Further, the set of resulting documents are ranked according to a degree of similarity to the query, i.e. tf, tdf.
  - Context Queries
    - Many systems complement single word queries with the ability to search words in a given context, i.e. hear other words.

- Words which appear near each other may signal a higher likelihood of relevance than if they appear apart.
- Two types of queries:

1. *Phrasal Queries*
   - Phrase is a sequence of single word queries.
   - An occurrence of the phrase is a sequence of words.
   - Relevance documents are those that contain a specific phrase, i.e. ordered list of contiguous word. For example: "buy camera" matches "buy a camera", "buying the cameras", etc.
   - Must have an inverted index that also stores positions of each keyword in a document.
   - Retrieving a document and position for each individual word, intersect documents and then finally checks for ordered contiguity of keyword positions.

2. *Proximity Queries*
   - A more relaxed version of the phrase query.
   - In this case, a sequence of single words or phrase is given together with a maximum allowed distance between them.
   - List of words with specific maximal distance constraints between terms.
   - For example: "dogs" and "race" within 4 words match.

     "……dog will begin the race……"
   - May also perform stemming and/or/not count stop words.

- Boolean Queries
  - A Boolean query has a syntax composed of basic queries that retrieve documents and of Boolean operators which work on their operands and deliver set of documents.
  - Since this schema is general compositional, a query syntax tree is naturally defined, where the leaves corresponds to the basic queries and the internal nodes to the operators.

- For example:

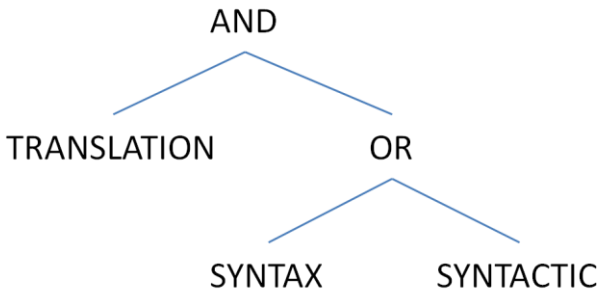


Fig.: it will retrieve all the documents which contain the word "translate" as well as either the word "syntax" or the word "syntactic".

- Operators most commonly used in Boolean queries are:
  1. *OR (e1 OR e2)*
  2. *AND (e1 AND e2)*
  3. *BUT (e1 BUT e2)* → *satisfy e1 but not e2*

- Natural Language
  - Full text queries as arbitrary strings.
  - All the documents matching a portion of the user query are retrieved.
  - Higher ranking is assigned to those documents matching more part of the query.
  - Typically, such process is used by vector space model.

2. PATTERN MATCHING

- A pattern is a set of syntactic features that must occur in a text segment.
- Those segments satisfying the pattern specifications are said to match the pattern.
- Examples:
  1. ***Prefixes:*** Pattern that matches start of the word. For example: "anti" means "antiquity", antibody", etc.

  2. ***Suffixes:*** Pattern that matches end of the word. For example: "ix" matches "fix", "matrix", etc.

3. ***Substrings:*** Pattern that matches arbitrary sub-sequence of characters. For example: "rapt" matches "enrapture", "velociraptor", etc.

4. ***Ranges:*** Pair of strings that matches any word alphabetically between them. For example: "tin" to "tix" matches "tip", "tire", "title", etc.

ALLOWING ERRORS

- What if query or document contains misspellings?
- Judge the similarity of words using edit distance.

EDIT (LEVENSTEIN) DISTANCE

- Minimum number of character deletions, additions or replacements needed to make two strings equivalent.
- For example: "misspell" to "misspell" is distance 1.
  "misspell" to "mistell" is distance 2.
  "misspell" to "misspelling" is distance 3.

REGULAR EXPRESSIONS

- Some text retrieval systems allow searching for regular expression.
- Examples:
  1. (u/e) nabl (e/ing) matches
     - unable, unabling, enable, enabling
  2. (un/en) *able matches
     - able, unable, untenable, enununenable

RELEVANCE FEEDBACK & QUERY EXPANSION

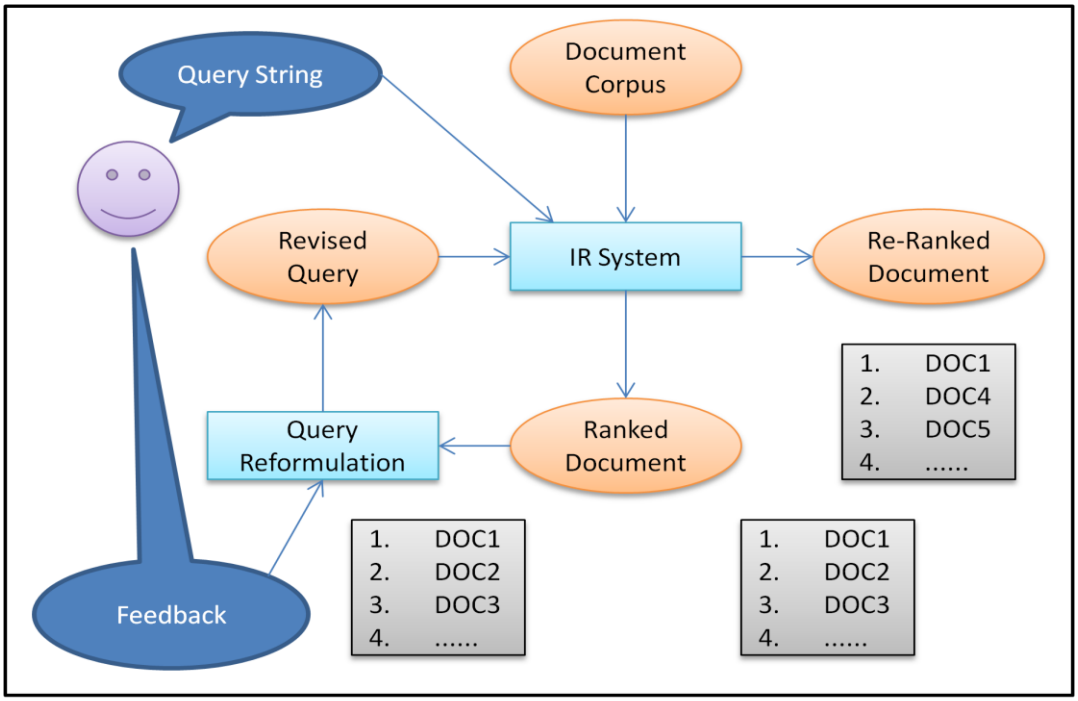
- In most collections, the same concept may be referred to using different words (synonyms).
- For example: a search for 'restaurant" to match "café".

- Such problem can be addressed by user manually.
- Also the system can help with query refinement.
- Such methods for tackling this problem by system are classified into two classes.

1. *Global Methods*
   - Query expansion/reformulation with a thesaurus or word net.
   - Techniques like spelling correction.
2. *Local Methods*
   - Relevance feedback
   - Pseudo-relevance feedback
   - Indirect relevance feedback

RELEVANCE FEEDBACK

- The idea of relevance feedback is to involve the user in the retrieval process so as to improve the final result set.
- In particular, the user gives feedback on the relevance of documents in an initial set of results.



RELEVANCE FEEDBACK ARCHITECTURE

BASIC PROCEDURE

- The user issues a query.
- The system returns an initial set of retrieval results.
- The user marks some returned documents as relevant or non-relevant.
- The system computes a better representation of the information need based on the user feedback.
- The system displays a revised set of retrieval results.
- Seeing some documents may lead users to refine their understanding of the information they are seeking.
- Image search provides a good example or relevance feedback.

WHEN DOES RELEVANCE FEEDBACK WORK?

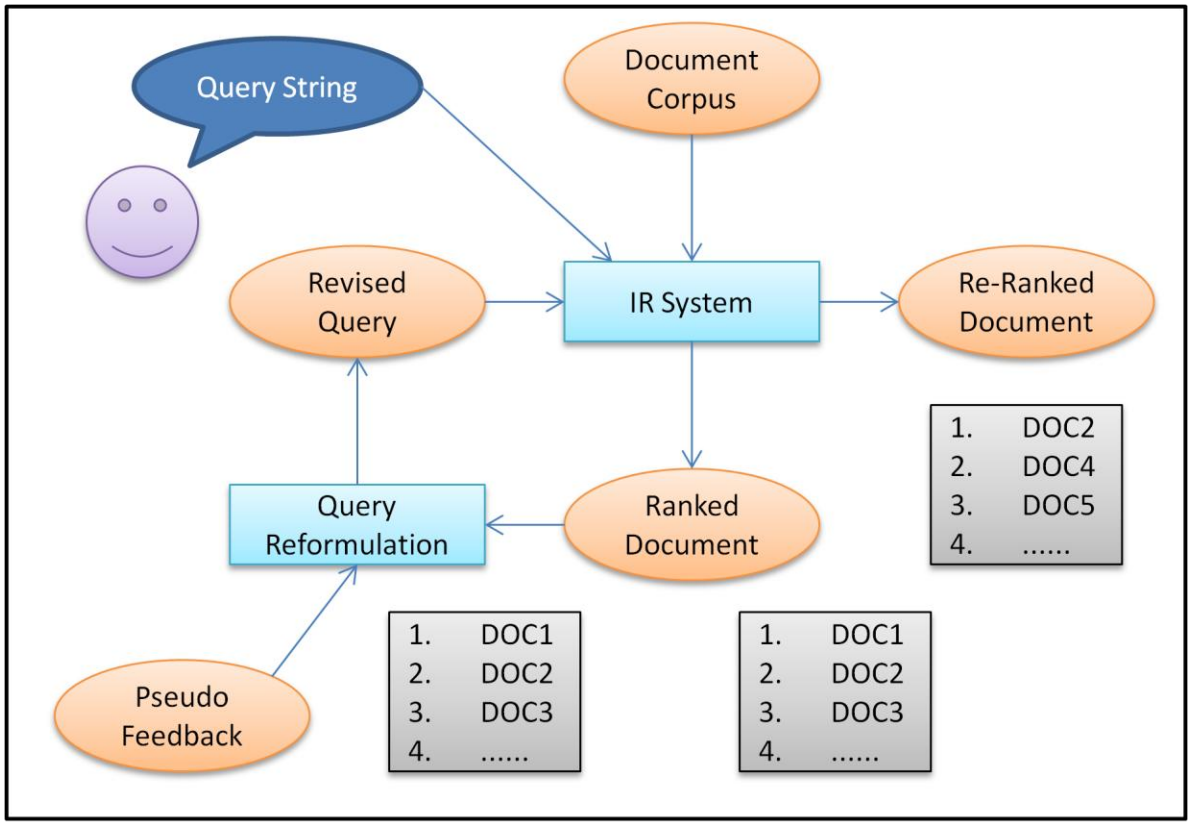
- The success of relevance feedback depends on certain assumptions.
- Firstly, the user has to have sufficient knowledge to be able to make an initial query which is at least somewhere close to the documents they desire.
- Secondly, the relevance feedback approach requires relevant documents to be similar to each other, i.e. they should cluster.

CASES WHERE RELEVANCE FEEDBACK ALONE IS NOT SUFFICIENT

1. Misspellings
   - If the user spells a term in different way to the way it is spelled in any document in the collection then relevance feedback is unlikely to be effective.
2. Cross language information retrieval (CLIR)
   - It is difficult to cluster the same relevance documents in different language rather than some language.
3. Mismatch of searcher's vocabulary versus collection vocabulary
   - For example: if the user searches for "laptop" but all the documents use the term "notebook computer" then the query fails.

PSEUDO RELEVANCE FEEDBACK

- Also called blind relevance feedback.
- Provides a method for automatic local analysis.
- Use relevance feedback methods without explicit user input.
- Just assume the top m retrieved documents are relevant and use them to reformulate the query.
- Allows for query expansion that includes terms that are correlated with the query terms.



PSEUDO FEEDBACK ARCHITECTURE

INDIRECT RELEVANCE FEEDBACK

- On the web, direct hit introduced the idea of ranking more highly documents that users choose at more often.
- Clicks on the links were assumed to indicate that the page was likely relevant to the query.
- Click stream mining.

THESAURUS

- A thesaurus provides information on synonyms and semantically related words and phrases.
- The IR system might also suggest search terms by means of a thesaurus.
- A user can also be allowed to browse lists of the terms that are in the inverted index and thus find good terms that appear in the collection.
- For example: Physician

  Syn (Synonyms): doc, doctor, MD, medical, medicines, medico

  Rel (Related): medic, general practitioner, surgeon

THESAURUS BASED QUERY EXPANSION

- For each term t in a query, expand the query with synonyms and related words of t from the thesaurus.
- May weight added terms less than original query terms.
- Generally increases recall.
- May significantly decrease precision, particularly with ambiguous terms.
- For example: "interest rate" → "interest rate fascinate evaluate".

WORD NET

- Word net is a lexical database for the English language.
- It groups English words into sets of synonyms called synsets providing various semantic relations between these synonym sets.
- Word net is more detailed database of semantic relationships between English words.
- Developed by famous cognitive psychologist George Miller and a team at Princeton University.
- About 144000 English words.

WORD NET SYNSET RELATIONSHIP

1. Antonym: front → back
2. Attribute: benelovelence → good (noun to adjective)

3. Pertainym: alphabetical → alphabet (adjective to noun)
4. Similar: unquestioning → absolute
5. Cause: kill → die
6. Entailment: breathe → inhale
7. Holonym: chapter → text (part of)
8. Meronym: computer → CPU (whole of)
9. Hyponym: tree → plant (specialization)
10. Hypernym: fruit → apple (generalization)

WORD NET QUERY EXPANSION

- Add synonyms in the same synset.
- Add hyponyms to add specialized term.
- Add hypernyms to generalize a query.
- *Note:* *Y is a holonym of X, if X is a part of Y*
  Y is a meronym of X, if Y is a part of X

SPELLING CORRECTION

- Correcting spelling errors in queries.
- For instance, we may wish to retrieve documents containing the term "carrot" when the user types the query "carot".
- Two steps to solve this problem:
  i. *Edit distance*
  ii. *K-gram overlap*

IMPLEMENTING SPELLING CORRECTION

- Of various alternative correct spellings for a misspelled query, choose the nearest one (i.e. the smallest edit distance).

- When two correctly spelled queries are tied, select the one that is more common. For example: "grunt" and "grant" both seem equally plausible as correction for "grnt". Correction is done then by examining which term (grunt or grant) is typed by the user in the query.

FORMS OF SPELLING CORRECTION

- Two forms:
  i. *Isolated term correction*
  ii. *Context sensitive correction*
- In isolated term correction, correct a single query term at a time, even when we have multiple term queries.
- But sometimes, such isolated term correction fails to detect.
- For example: "flew form Nepal" → contains the misspelling of the term "from" but not detected by isolated term correction. In such case we need context sensitive correction.

EDIT DISTANCE

- Given two character strings S_1 and S_2, the edit distance between them is the minimum number of edit operations required to transform S_1 into S_2.
- Most commonly edit operations include the following operations:
  i. Insert a character into a string.
  ii. Delete a character from a string.
  iii. Replace a character of string by another character.
- Edit distance is also called Levenstein distance.
- **Algorithm:**

```
EDIT DISTANCE (S1, S2)
int M[i, j] = 0
for i = 1 to |S1|
        do M[i, 0] = i
for j = 1 to |S2|
```

do M[0, j] = j

for i = 1 to $|S_1|$

do for j = 1 to $|S_2|$

do M[i, j] = min { M[i−1, j−1] + if (S_1[i] = S_2[j]) then 0 else 1, M[i−1, j] + 1, M[i, j−1] + 1}

return M[$|S_1|$, $|S_2|$]

- The [i, j] entry of the matrix (after execution of algorithm) will hold the edit distance between the strings consisting of the first i characters of S_1 and first j characters of S_2.

K-GRAM INDEXES FOR SPELLING CORRECTION

- A k-gram is a sequence of k characters.
- Example: "cas", "ast", "stl" are 3 grams occurring in term "castle".
- Use the k-gram index to retrieve vocabulary terms that have many k-grams in common with the query.
- Example:

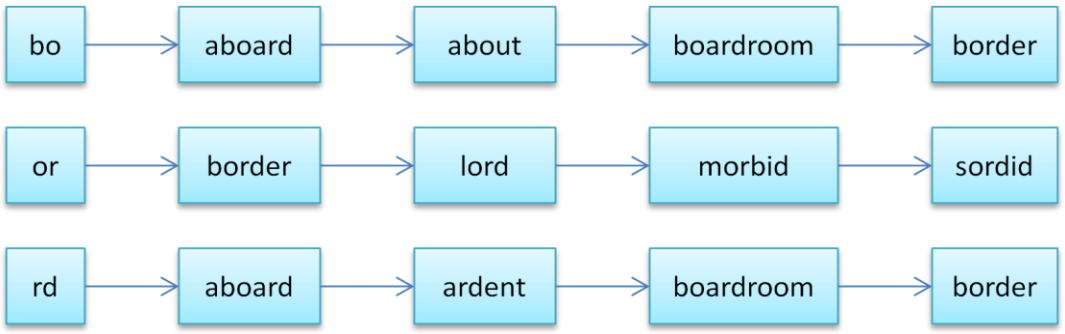


Fig.: Matching at least two of the three 2 gram in the query "bord"

- Suppose we want to retrieve vocabulary terms that contained at least two of these bigrams. We would enumerate aboard, boardroom and border.