Information Retrieval

Miriam Butt
December 2002

Information Retrieval

What is it good for: Storage and Retrieval of all kinds of media. Main application so far is with *text documents* (also known as **Data Mining**).

Document: indexed unit of text indexed (e.g, a Webpage).

Collection: set of documents (e.g, the WWW).

Term: lexical item in a collection (e.g., bass).

Query: user's informational need expressed as a set of terms (e.g., where can I catch bass?).

Information Retrieval

Level of Sophistication: No information beyond the word. **Bag of Words** approach is common: *I see what I eat* and *I eat what I see* are treated as equivalent.

Other Tasks: 1) Document Categorization

2) Document Clustering

3) Text Segmentation

4) Text Summarization

Document Categorization

Classify a Document: Figure out which of an existing class of documents a given document should be identified as.

Most Common Method: Supervised Machine Learning

Good For: 1) Routing, e.g, getting e-mails to the right person to answer them.

2) Filtering, e.g., spam mails

3) Identifying the Language of a Document, e.g., to retrieve only those

Document Clustering

Discover a Cluster of Documents: Maximize within-cluster document similarity, minimize between-cluster similarity.

Efficiency: Clustering Documents allows for more efficient overal information retrieval.

Cluster Hypothesis (Jardine and van Rijsbergen 1971)**:** Identifying clusters should allows for greater precision/recall. But, no real empirical support so far. (More interesting recent work seems to be coming out of a study of how **Networks** work: comparing the WWW and human networks).

Precision/Recall

These measures are used generally to test the performance of a system. In terms of information retrieval, one can calculate the following:

Recall = $\frac{\text{\# of relevant documents returned}}{\text{total \# of relevant documents in collection}}$

Precision = $\frac{\text{\# of relevant documents returned}}{\text{\# of documents returned}}$

Text Segmentation

Document Organization: break down larger documents into smaller, more coherent units, e.g. if one document covers a wide variety of topics.

Text Summarization

Produce a shorter summary version of an existing document.

Knowledge Based: Detailed syntactic/semantic analysis which produces a meaning representation of the text. This representation is then passed on to a **generator**, which produces a new piece of text summarizing the original, longer text (this is the ideal world).

Selection Based: word frequency and discourse structure heuristics are used to identify the "important" sentences. A predetermined number of such important sentences are pulled out and included in the summary document.

Ad Hoc Retrieval

Ad Hoc Retrieval: An unaided user poses a question to a retrieval system. The system returns a set of ordered and hopefully useful documents.

There are several possible methods to achieve this, the one most popularly used is the **Vector Space Method.**

The Vector Space Model

Documents and queries are represented as **vectors of features.** The value of the feature indicates the presence or absence of a term (this could also be a weighted value).

Document: $\vec{d_j} = (t_{1,j}, t_{2,j,}, t_{3,j,}, \ldots t_{N,j,})$

Query: $\vec{q_k} = (t_{1,k}, t_{2,k,}, t_{3,k,}, \ldots t_{N,k,})$

The Vector Space Model--- An Example

Document1: This is Miriam Butt's Web Page.
Vector of Features: [1, 1, 1, 1, 0, 0]

Document2: This is Tracy King's Web Page.
Vector of Features: [0, 0, 1, 1, 1, 1]

Query: Miriam Butt
Vector of Features: [1, 1, 0, 0, 0, 0]

Comparison: Figure out the number of terms two vectors have in common (via a similarity metric, J&M p. 648, (17.17)).

Calculating Similarity

In the previous example, vectors were compared by simply summing the number of terms they share (function words such as *this* and *is* or *the* and *and* are generally left out because they are not useful similarity indicators, see notion of "stop list").

Terms are given a **binary** value: either they are found, or they are not found.

However, some terms tend to be more important than others, so it is generally better to assign **weighted** values instead.

Term Weighting

Term Frequency: Simple check to see how frequent a given term is in a document. The assumption is that a frequently occurring term will be more important.

Inverse Document Frequency: Check for a term across a collection of documents. The fewer documents a term occurs in, the higher its weight (i.e, it is a very important term in the context of that document).

Stop List

List of functional high-frequency words which are eliminated from a document for the purposes of information retrieval (determiners, conjunctions, auxiliaries, etc.)

Problem: *To be or not to be* could end up being looked up simply under "not".

Vexed Morphology

In a simple, term by term treatment, the following words will all be treated as completely unrelated terms:

process, processing, processed

This is clearly not desirable. One possible quick fix: integrate a stemmer (such as the Porter stemmer) to preprocess terms.

Problem: Throw away "too much" information. Example, not being able to distinguish *stockings* (stock) from *stocks* (stock) can prove to be extremely embarassing.