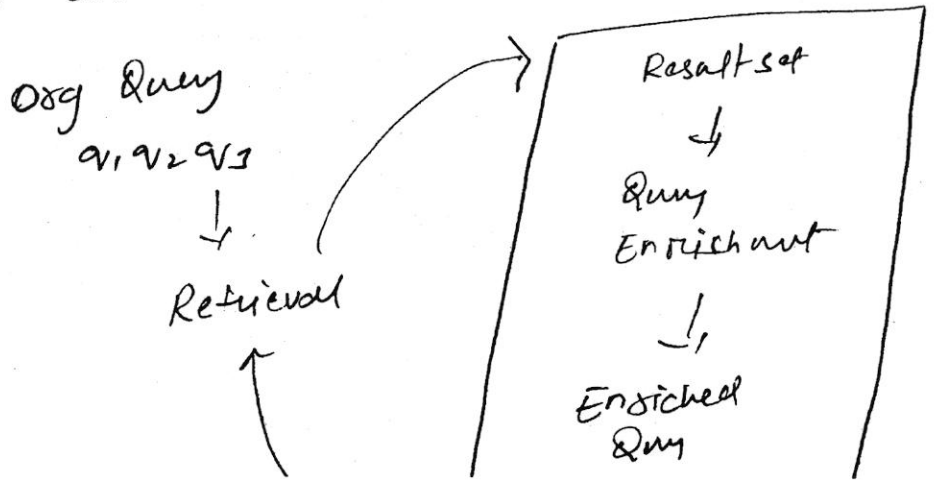
UNIT-2

Retrieval Utilities:

Many different utilities improve the results of a retrieval strategy. Most utilities add or remove terms from initial Query in an attempt to refine the Query.

Relavance feedback: The top doc found by an initial Query are identified as relevant. These doc are then examined. They might be deemed relevant either by manual intervention or by an assumption that the top n doc are relevant.

The basic idea is to implement retrieval in multiple passes. The user retrial the Query in each pass based on the results of earlier Query. The user indicates which of the doc presented in response to an initial Query are relevant & new terms are added to the Query based on this selection. Additionaly existing terms in the Query can be reweighted based on user feedback.



An alternative approach is to avoid asking the user anything & to simply assume the top ranked doc are relevant. Using either manual or automatic, the initial Query is modified & the new Query is used.

Let find info surrounding the various conspiracy theories about the assassination of John f. Kennedy. has both useful keywords & Noise.

keywords are Assasi & John

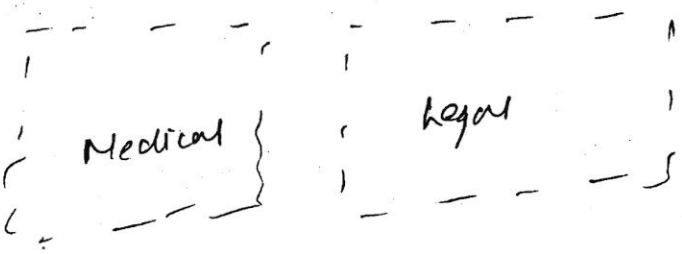
various, info are not stop words.

they used words a, an, and, the are ignored

the idea is to use all the terms in the Query & ask the user if top ranked doc are relevant.

Assume a highly Ranked document contains the term Oswald. It is reasonable to expect that adding the term oswald to the Query will improve both precision & recall. And if it may contain many occurances of assination, the wt used in initial Query for this term should be increased.

Clustering: doc clustering attempts to group doc by content to reduce the search space required to respond to a Query. for eg a document collection that contains both medical & legal doc might be clustered such that all med doc are placed into one cluster & all legal doc are assigned to legal cluster. A query over legal material might be directed to the legal doc cluster



Several clustering algorithms have been proposed. In many cases for evaluation has been challenging because it is difficult to automatically point a query to a doc cluster.

① Result set clustering:

clustering was used as a utility to assist relevance feedback. In those cases ontop the result of a query were clustered & in relevance feedback process, by one new term from large cluster were selected.

Recently, web search results were clustered based on significant phrases in the result set. first documents in a result set are parsed & two term phrases are identified. characteristics about these phrases are then used as input- algorithm. (eg

are used in this work. A cluster is initially identified as the set of documents that contain one of the most significant phrases.

Eg "New York" all doc that contain this phrase would be placed into a cluster. finally these initial clusters are merged based on doc-doc similarity.

(2) Hierarchical Agglomerative clustering.

First the NxN doc similarity matrix is found. Each doc is placed into its own cluster. The following two steps are repeated until one cluster exists.

The two clusters that have the highest similarity are found.

These two clusters are combined & the similarity between the newly formed cluster & the remaining clusters recomputed

As the larger cluster is formed the clusters that merged together are tracked & form a hierarchy.

Assume doc A B C D E exist & a doc-doc sim matrix exists. At this pt each doc is a cluster by itself.

{ {A} {B} {C} {D} {E} }

{ {A B} {C} {D} {E} }

After certain steps

{ {A B C D E} }

However the history of the formation of this