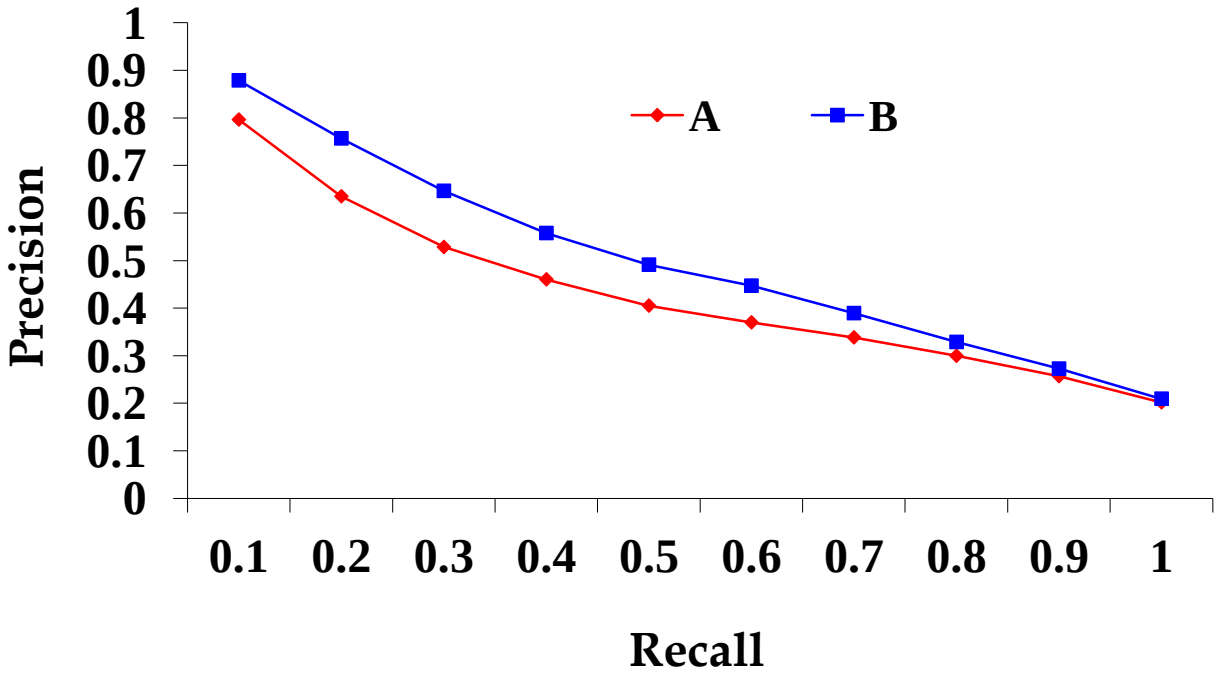
Information Retrieval Systems Class Notes (16.02.2014 – 21.02.2014)

Emre Nevayeshirazi

1. Evaluation of Information Retrieval Systems

Last time we have seen,

- Recall (R)
- Precision (P)
- F Measure 2PR / (P + R)



Using precision-recall graph we can compare two information retrieval systems. The curve closest to upper right corner of graph indicates better results.

To compare whether the difference is significant or not, we can use **Paired T Tests**. This test is also called **Student's T Test**.

Rank	1	2	3	4	5	6	7	8	9	10

To use Student's T Tests we need **Information Retrieval Test Collection**

IR Test Collection is:

- Set of Documents
- Set of Queries
- Relevant documents from query

We need to use statistical tests to show the results are not by chance.

R – Precision

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	x
13	772	x
14	990	

R precision is one of the information retrieval evaluation methods.

Precision at the R^{th} position in the ranking of results for a query that has R relevant documents.

R-Precision = 4/6

Recall Precision Example

- Assume we have a test collection.
- It means for given query, we know all the relevant documents.
- Assume no. of relevant documents is 10 for given query.

Relevance	0	1	0	1	1	1	1	0	0	0
Precision	0/1	1/2	1/3	2/4	3/5	4/6	5/7	5/8	5/9	5/10
Recall	0/10	1/10	1/10	2/10	3/10	4/10	5/10	5/10	5/10	5/10

2. Clustering

Clustering is a process that puts similar items into the same group and dissimilar items into different groups.

Classification is a similar to clustering. In classification we assign objects to one of the existing groups.

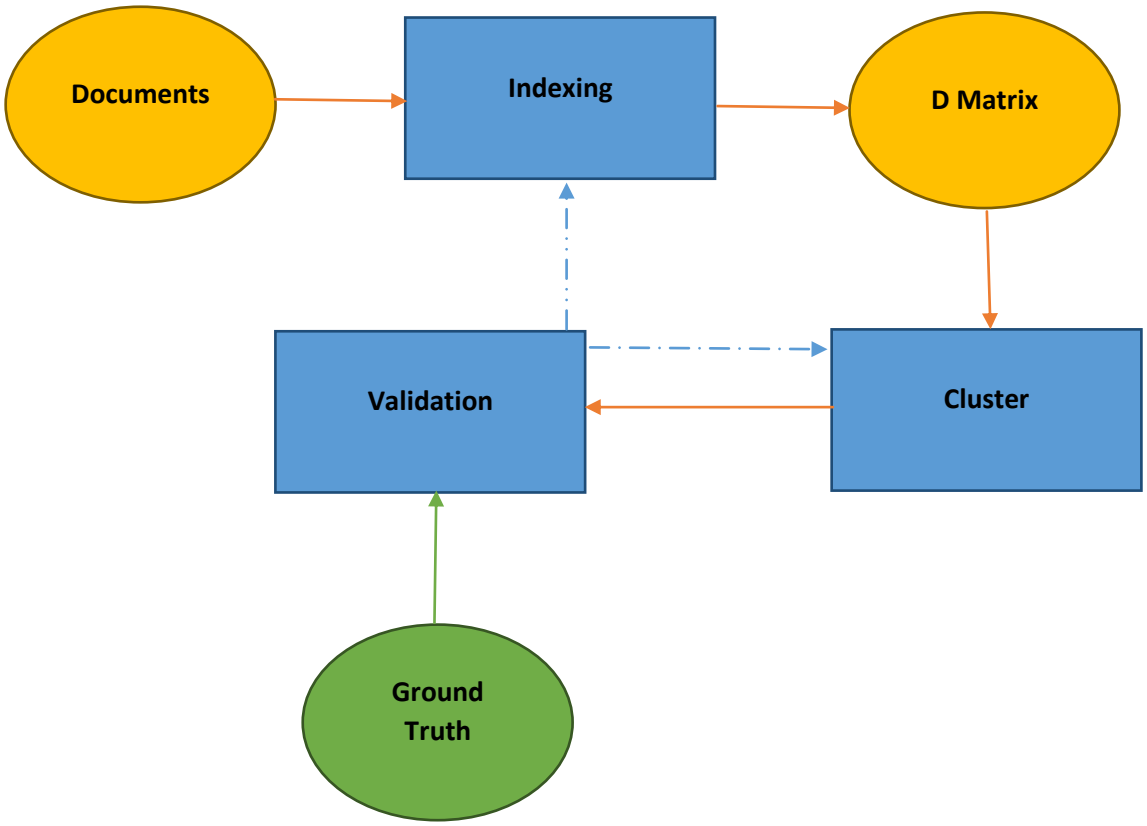
In clustering there is no predefined groups. Classification is supervised, clustering is not.

Motivation for Clustering

- Cluster IR search results
- Find similar documents
- Cluster terms so that we can use additional terms during query

Clustering Process

If validation is not satisfactory, we can go back to indexing or cluster phase to tweak our approach. If validation passes, we make a test with random values not controlled values.



Cluster Based Retrieval

- Choose one representative from cluster.
- Or we can use average document vector for that cluster.

CBR - 2

n_c = total number of clusters

1. Compute the query with all cluster representations.
2. Choose top number of most similar clusters. ($n_s = n_c \times 0.1$)
3. Consider the members of the selected cluster and rank them according to their similarity to the query.

Classification of Clustering Algorithms According to Clustering Structure

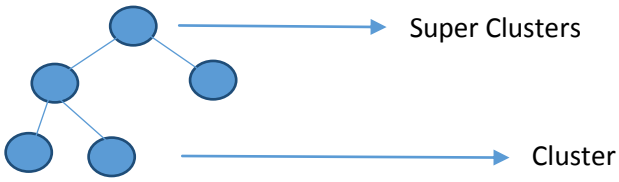
1. Partitioning

Intersection of two clusters do not have common members when these two clusters are not equal.

2. Overlapping

Intersection of two clusters may not be null set when these two clusters are not equal to each other.

3. Hierarchical



It can be top down or bottom up.

It is especially good for terms.

4. Fuzzy Structured

Object can be members of multiple clusters by some chance.

Classification of Clustering Algorithms In terms of Implementation

- Single Pass
- Multi Pass
- Graph Theoretical
- Query Based

1. Single Pass

A. Seed Oriented

Choose clusters seeds and assign non seeds to seed objects.

Problems

- No of clusters
- Similarity measure to use for member assignment
- Selection of seeds

Seed oriented clustering algorithm can produce both portioning and overlapping structures.

Singleton : Cluster with single object.

Ragbag : Cluster that may contain anything.

Seed Selection

1. They must be distinct. (decoupled)
2. They must be connected to other objects. (cohesion)
3. Can be selected randomly.
4. c^3m selection
5. Using inverted lists.

B. Heuristic

1. The first document becomes cluster.
2. Take the next document
    a. If it is not similar to one of the existing clusters
        i. It becomes its own cluster
    b. Else
        i. Assign it to one of the existing clusters that it is most similar to
3. Repeat step 2

This is an order dependent algorithm.

2. Multi Pass

1. Obtain initial clusters (quick and dirty)
2. Repeat
    a. Create a centroid vector for each existing clusters
    b. Assign each object to most similar clusters

   Until no. of iterations is equal to max no. of iterations or stability of clusters is greater than or equal to >= threshold value

Characteristics of Good Clustering Algorithms

1. Efficiency and effectiveness (meaningful)
2. Order independence
3. Stable (addition of new clusters do not destroy original structure)
4. Maintainable (we should not recluster every time, it should be incremental)
5. Robustness (mistakes in selection of parameters or mistakes in D matrix should not affect the clustering structure)

Questions

1. For given document set, query and relevant documents calculate the precision and recall. Total number of relevant documents is 4.

Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Relevance	0	0	1	1	0	1	0	0	1	0
Precision										
Recall										

2. What is the R precision for above set of documents?

3. What is the main difference between clustering and classification ?

4. What is the IR Test Collection and why do we need it?

5. In clustering process which phases should be repeated if validation fails ?

Answers

1.

Rank	1	2	3	4	5	6	7	8	9	10
Relevance	0	0	1	1	0	1	0	0	1	0
Precision	0/1	0/1	1/3	2/4	2/5	3/6	3/7	3/8	4/9	4/10
Recall	0/4	0/4	1/4	2/4	2/4	3/4	3/4	3/4	4/4	4/4

2. R precision is the precision at R^{th} position in the ranking of results for a query that has R relevant documents.

   In our case we have 4 number of relevant documents. So we need to find precision at 4^{th} point.

   The precision at 4^{th} point is **2/4.**

3. In classification we assign objects to one of the existing groups. In clustering there is no predefined group. In short, classification is supervised whereas clustering is not.

4. IR Test collection is a set of documents and set of queries and relevant documents for each query.

   We need IR Test collections so that we can easily compare the performance of different retrieval techniques with each other.

5. If validation fails during clustering process, we can either go back to indexing (generate D matrix in a different way, e/g/. use a different stemming algorithm) or clustering (change the number of clusters, etc.) phases to tweak our approach.