Information Retrieval Methods

Helena Ahonen-Myka
Spring 2006, part 6
Clustering (ryvästäminen, klustring)
Translation: Greger Lindén

In this part

- Clustering of documents
  - Methods based on the document-document similarity matrix
  - Heuristic methods
- Using clustering in information retrieval

2

Clustering

- Clustering:
  - A set of documents is divided into groups, where documents within each group are more similar than documents that belong to different groups
- Clustering hypothesis: similar documents are usually relevant for the same retrieval task
- → If similar documents are grouped, the retrieval can be made more effective
  - Support for browsing
  - Another way of accessing documents, in addition to inverted files
  - Similar documents can be physically stored together→ faster search

3

Similarity between documents

- The similarity between two documents i and j can be described with the cosine measure $\cos(D_i,D_j)$
- We compute similarities between all documents from the document-term matrix
  - → we obtain a document-document matrix (of size n x n if there are n documents)
- The document-document matrix is symmetric, so it is enough to represent only the values on either side of the matrix diagonal

4

Document-term matrix

document			ter	ms				
vector	a	b	c	d	e	f	g	h
D1	1	0	1	1	0	1	0	1
D2	0	1	1	0	0	1	1	0
D3	0	0	0	1	1	0	1	0
D4	1	1	0	0	1	0	0	1

5

Document-document matrix

docu-		docu-	ments	
ments	D1	D2	D3	D4
D1	1	0.45	0.26	0.45
D2		1	0.29	0.22
D3			1	0.26
D4				1

6

Clustering

- Two subproblems related to retrieval
  - – Constructing clusters
    - usually not performed often, so it can take time and space
    - when the collection is updated, the clustering must usually be redone
  - – Using clusters in retrieval
    - must be fast

Desired properties of clustering methods

- Efficiency, mostly concerning time
- Properties of the clustering result
  - – The clusters should not change very much when new documents are added
  - – If there are small errors in the descriptions of the documents, the influence of these errors on the clustering should also be small
  - – The clustering should be independent of the order in which the documents are processed (clustered)

Clustering methods

- Methods that are based on the document-document similarity matrix, e.g. hierarchical methods
  - – Satisfy the constraints related to the clustering result
  - – slow: O(n^2)
- Heuristic methods, e.g. one-pass method, k-means
  - – Do not satisfy the constraints, but the result is usually fairly good
  - – fast: O(n log n)

A simple method

- The similarity between two documents is measured by the cosine function cos(d_i, d_j)
- We choose some threshold
- If the similarity value of documents d_i and d_j exceeds the threshold, then d_i and d_j belong to the same cluster
- E.g. in Slide 6
  - – threshold 0.4: clusters {D1, D2, D4} and {D3}
  - – threshold 0.5: each document forms its own cluster

Hierarchical methods

- For information retrieval, a hierarchical clustering is more suitable
- The simple method above could be repeated for different thresholds
  - – But to choose suitable threshold values is hard
- There are also special methods that are based on constructing a clustering hierarchy (=nested clusters)
  - – We can use the whole hierarchy or define a threshold after which the hierarchy is cut → we can have several separate clusters

Hierarchical methods

- We can start from the point where each document forms its own cluster
  - – We combine clusters until there is only one cluster left → agglomerative clustering
- Or all documents can be in the same cluster at the start
  - – We divide clusters until each document forms its own cluster → divisive clustering

Hierarchical, agglomerative clustering

1. Construct a document-document matrix
2. Assign each document to its own cluster
3.
   - Construct a new cluster by combining clusters i and j, the similarity value of which is the highest
   - Update the similarity matrix removing all rows and columns corresponding to the clusters i and j
   - Add a new row ij and compute the similarities of the new cluster regarding all other clusters → the elements of the row
4. Repeat step 3 until there is only one cluster left

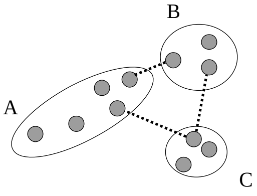
Hierarchical, agglomerative clustering

- The similarity between two clusters
  - If there is only one document in each cluster, the similarity is the similarity between the documents
  - If there are more documents in the clusters, we will have to define how to compute the similarity value between the clusters; there are several alternatives
    - single link
    - complete link
    - group average

Single-link clustering

- The similarity between a pair of clusters:
  - the similarity between the most similar pair of documents, one of which appears in each cluster
- Each cluster member will be more similar to at least one member in that same cluster than to any member of another cluster
- Single-link clustering tends to produce a small number of large, poorly linked clusters

Single-link clustering

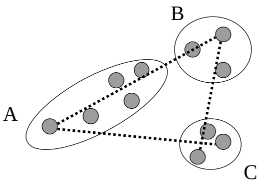


We combine the two clusters whose shortest distance is the smallest: A and B

Complete-link clustering

- The similarity between a pair of clusters:
  - the similarity between the least similar pair of documents from the two clusters
- Each cluster member is more similar to the most dissimilar member of that cluster than to the most dissimilar member of any other cluster
- Complete-link clustering produces a larger number of small, tightly linked clusters

Complete-link clustering



We combine the two clusters whose longest distance is the smallest: B and C

Group-average clustering

- A compromise
- Each cluster member has a greater average similarity to the other members of its cluster than it does to all members of any other cluster

Heuristic methods

- Heuristic methods are based on comparing document vectors only when needed
  - There is no need to construct a document-document matrix
- Usually we need parameters that have been experimentally determined (and that might not be easy), e.g.
  - The number of clusters desired
  - A minimum and maximum size of each cluster (i.e. number of documents)
  - A threshold on the document-to-cluster similarity measure, below which a document will not be included in the cluster; the control of overlap between clusters

Heuristic methods: one-pass method

- One-pass method: each document is processed only once
  - The first document forms its own cluster
  - Each of the other documents is compared to each existing cluster (e.g. to the centroid)
    - If the similarity value exceeds a threshold, the document is added to the cluster (and the centroids are updated)
    - Otherwise the document forms a new cluster
- Note: document can be added to several clusters

The centroid vector

- A cluster can be “summarised” by its centroid
- centroid = the average vector of the vectors of the cluster
- cen ({D_1,...,D_m}) = (w_1*, ..., w_t*), where each w_k^* (k = 1,...,t) is the average of the corresponding components:

$$w_k^* = (1/m)\sum_{i=1}^{m} w_{ik}$$

Computing the centroid vector

D1	1	0	1	1	0	1	0	1
D2	0	1	1	0	0	1	1	0
D4	1	1	0	0	1	0	0	1
sum	2	2	2	1	1	2	1	2
centroid	0.67	0.67	0.67	0.33	0.33	0.67	0.33	0.67

Heuristic methods: one-pass method

- The one-pass method usually forms clusters of very different sizes
- Balancing the clustering: we can change the size of the clusters, the number of the clusters, and how much the clusters overlap (number of joint documents)
- We can e.g. specify the average size of a cluster
  - If the cluster grows over the given size, it is split into two clusters

Heuristic methods: k-means

- We choose k to be the number of clusters
- We divide the documents into k clusters in some way
- Each document is compared at a time to the centroids of the clusters and added to the cluster whose centroid is the most similar to the document
- After the addition of each document, the centroids are recomputed and we repeat adding documents until there are no significant changes in the clusters any more

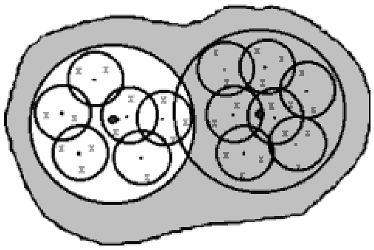
Hybrid methods

- Example I
  - We form a "coarse" clustering using some heuristic method
  - We use a method based on the document-document matrix to divide the clusters into subclusters
- Example II
  - We select a subset of documents and cluster them using a method based on the document-document matrix
  - The remaining documents are clustered into the existing clusters using some heuristic method

Cluster hierarchy for search

- For each document cluster, we compute a centroid
  - The centroid summarises the cluster
- We can compare the similarity of the centroids and group clusters to higher level clusters (with their own centroids)
- → we build a cluster hierarchy
  - At the top level, one or a few clusters
  - The term vectors of the documents are on the lowest level in the hierarchy
  - The documents in a cluster are also stored as a cluster (physically in the same place) → retrieval is fast

Cluster hierarchy for search



Retrieval from a clustered collection

- The query vector is compared to the centroids of the top-level clusters
- Clusters with centroids that are similar (exceeds a threshold) are studied further
- We compare further the centroids of the subclusters of these clusters to the query vector
- We continue until we have reached the lowest level in the hierarchy
- The documents of these clusters are returned as answer

Using clustering in retrieval

- In the previous slides we assumed that all documents were clustered
  - After an update, the clustering must be redone
  - When performing a search, the clusters already exist
  - The user's query is compared to the clusters and the best clusters are returned as answers
- We could also restrict ourselves to clustering dynamically only the results
  - The document space is smaller and does not change
  - The clustering must be fast
  - The clustering hypothesis works better because the documents are an answer to some query

Clustering search results: the Scatter/Gather method

1. We first do a clustering
2. We show the user sample text summaries (e.g. Titles) for each cluster
3. The user selects the interesting clusters
4. Selected clusters are combined and reclustered
5. We repeat from step 2 until the user is satisfied

Clustering search results: the Scatter/Gather method

- The user can browse documents at any level
- The clusters are on ever narrower topics, but not necessarily on subtopics of topics chosen in the beginning
  - The user may also change the focus of his interests

In this part

- Clustering methods based on the document-document similarity matrix
  - A simple method
  - Hierarchical, agglomerative clustering
- Heuristic methods
  - one-pass method
  - k-means
- Hybrid methods
- Using clustering in information retrieval; clustering answers