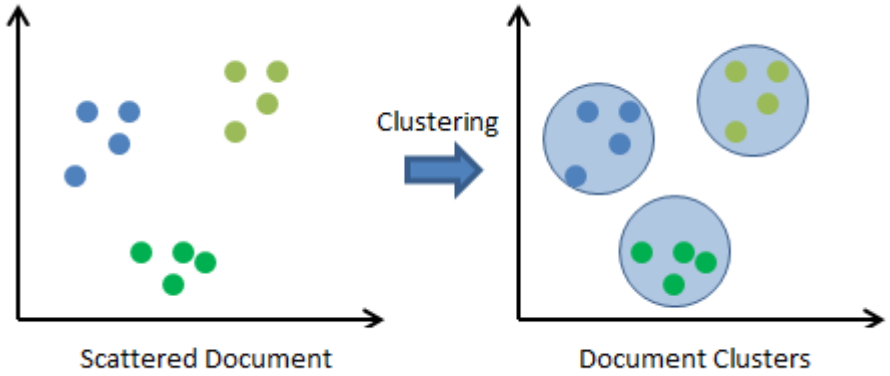
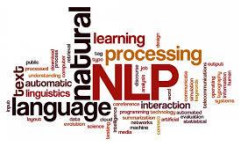


Document Clustering

Natural Language Processing

Agenda

- What is Document Similarity
- Methods to measure Document Similarity
- Cosine Similarity Method



Document clustering

- "I have these millions of documents (unstructured data). Is there a way I can group them into some meaningful categories?"

Clustering

- task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters)

Document clustering

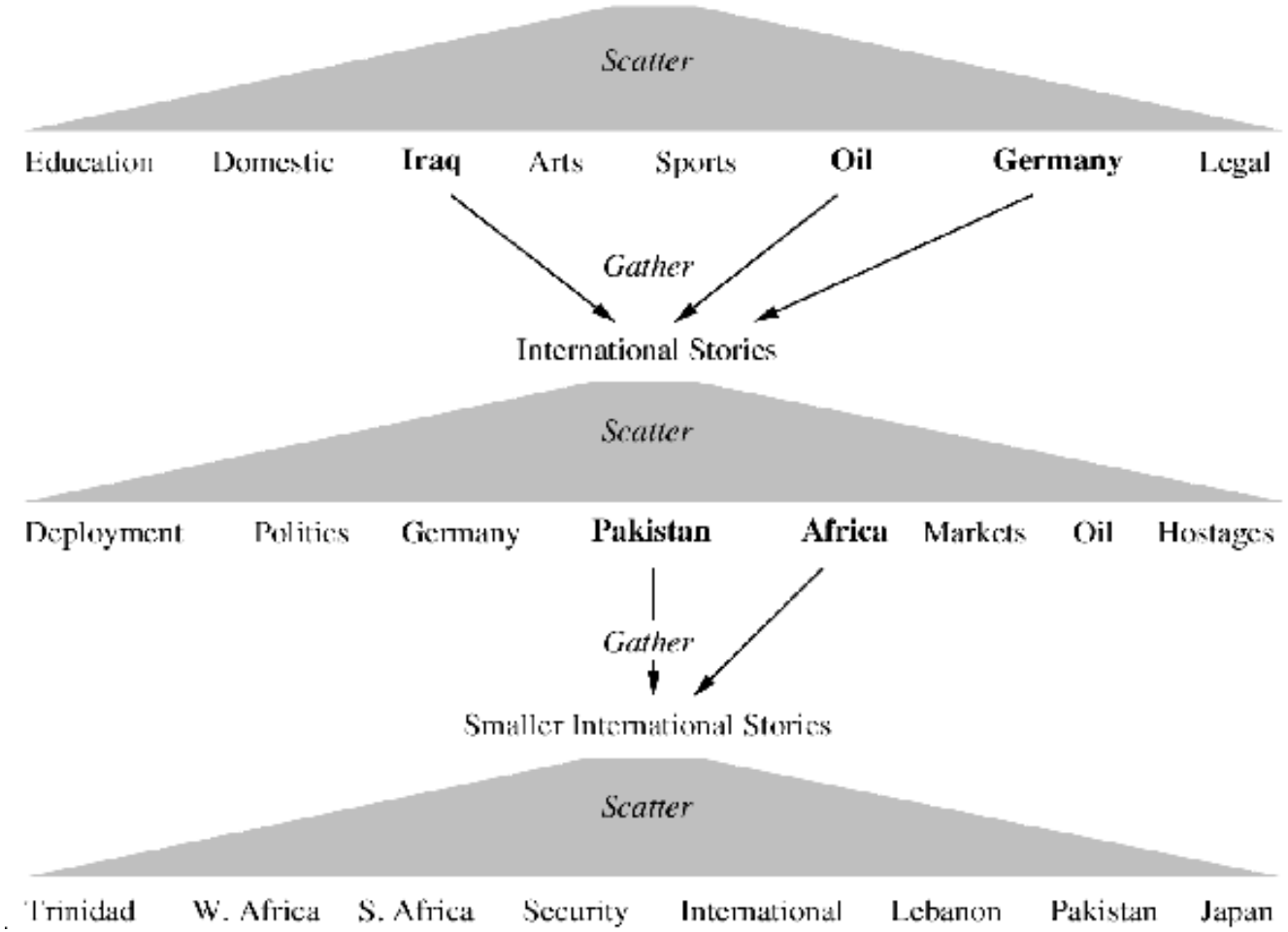
- Called Text clustering.
- application of cluster analysis to textual documents.
- It has applications in automatic document organization, topic extraction and fast information retrieval or filtering.

Applications of clustering in IR

Application	What is clustered?	Benefit
Search result clustering	search results	more effective information presentation to user
Scatter-Gather	(subsets of) collection	alternative user interface: “search without typing”
Collection clustering	collection	effective information presentation for exploratory browsing
Cluster-based retrieval	collection	higher efficiency: faster search


Vivísimo
jaguar
the Web
Search
Advanced Search
Help
Clustered Results
Top 208 results of at least 20,373,974 retrieved for the query jaguar (Details)
Jaguar (203)
Cars (74)
Club (34)
Cat (23)
Animal (13)
Restoration (10)
Mac OS X (8)
Jaguar Model (8)
Request (5)
Mark Webber (5)
Maya (5)
More
Find more clusters
Enter Keywords
Go
1. Jag-lovers - THE source for all Jaguar information [new window] [frame] [cache] [preview] [clusters]
... Internet! Serving Enthusiasts since 1993 The Jag lovers Web Currently with 40661 members The Premier Jaguar Cars web resource for all enthusiasts Lists and Forums Jag lovers originally evolved around its ...
www.jaglovers.org Open Directory 2, Wisenut 8, Ask Jeeves 8, MSN 9, Lookamart 12, MSN Search 18
2. Jaguar Cars [new window] [frame] [cache] [preview] [clusters]
[...] redirect to www.jaguar.com
www.jaguarcars.com - Lookamart 1, MSN 2, Lycos 3, Wisenut 6, MSN Search 9, MSN 29
3. http://www.jaguar.com/ [new window] [frame] [preview] [clusters]
www.jaguar.com - MSN 1, Ask Jeeves 2, MSN Search 3, Lycos 9
4. Apple - Mac OS X [new window] [frame] [preview] [clusters]
Learn about the new OS X Server, designed for the Internet, digital media and workgroup management. Download a technical factsheet.
www.apple.com/macosx - Wisenut 1, MSN 3, Lookamart 26

Applications of clustering in IR



Types

- The application of document clustering can be categorized to two types, online and offline.
- Online applications are usually constrained by efficiency problems when compared to offline applications.
- Text clustering may be used for different tasks, such as grouping similar documents (news, tweets, etc.) and the analysis of customer/employee feedback, discovering meaningful implicit subjects across all documents.

Algorithms

- Hierarchical based algorithm,
- K-means algorithm and its variants.
- Graph based clustering,
- Ontology supported clustering and
- Order sensitive clustering

Hierarchical based algorithm

- includes single link, complete linkage, group average and Ward's method.
- By aggregating or dividing, documents can be clustered into hierarchical structure, which is suitable for browsing.
- Generally hierarchical algorithms produce more in-depth information for detailed analyses
- Usually suffers from efficiency problems.

K-means algorithm

- more efficient and provide sufficient information for most purposes.

K-means algorithm

K-MEANS$(\{\vec{x}_1, \ldots, \vec{x}_N\}, K)$

1 $(\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_K) \leftarrow$ SELECTRANDOMSEEDS$(\{\vec{x}_1, \ldots, \vec{x}_N\}, K)$
2 **for** $k \leftarrow 1$ **to** K
3 **do** $\vec{\mu}_k \leftarrow \vec{s}_k$
4 **while** stopping criterion has not been met
5 **do for** $k \leftarrow 1$ **to** K
6 **do** $\omega_k \leftarrow \{\}$
7 **for** $n \leftarrow 1$ **to** N
8 **do** $j \leftarrow \arg\min_{j'} |\vec{\mu}_{j'} - \vec{x}_n|$
9 $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$ *(reassignment of vectors)*
10 **for** $k \leftarrow 1$ **to** K
11 **do** $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$ *(recomputation of centroids)*
12 **return** $\{\vec{\mu}_1, \ldots, \vec{\mu}_K\}$

Hard / Soft clustering algorithms

- Hard clustering computes a hard assignment – each document is a member of exactly one cluster.
- The assignment of soft clustering algorithms is soft – a document's assignment is a distribution over all clusters.
- In a soft assignment, a document has fractional membership in several clusters

Procedures

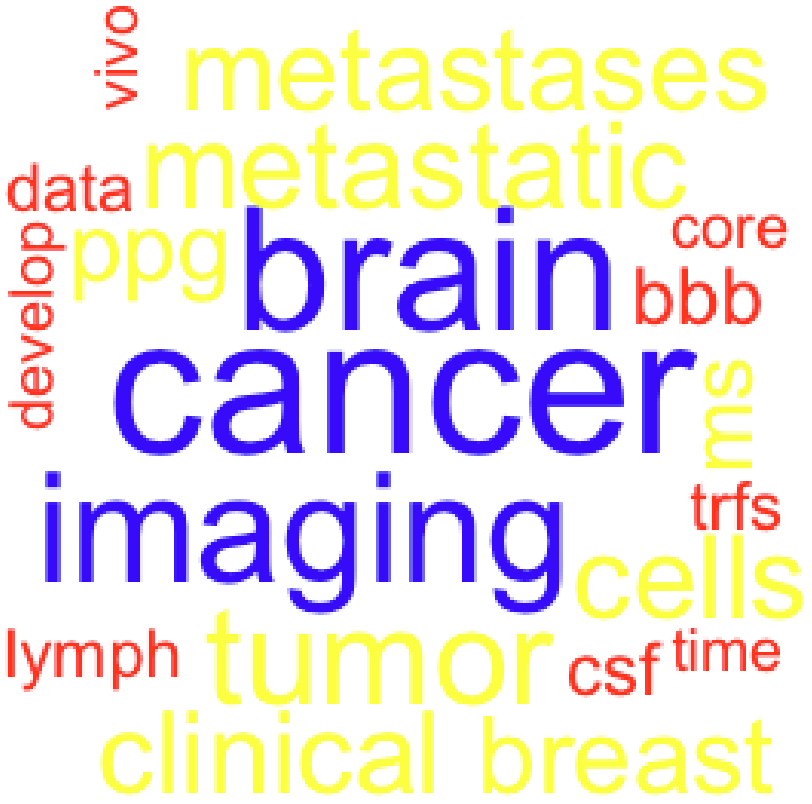
1. Tokenization
2. Stemming and lemmatization
3. Removing stop words and punctuation
4. Computing term frequencies or tf-idf
5. Clustering
6. Evaluation and visualization

Procedures

Cluster summary table

cluster	size	top_words
1	23	risk, health, diabetes, intervention, treatment
2	4	hiv, inflammation, env, testing, study
3	38	cell, infection, determine, cells, function
4	7	research, program, cancer, disparities, students
5	8	cancer, brain, imaging, tumor, metastatic
6	3	microbiome, crc, gut, psoriasis, gut_microbiome
7	3	cdk, nmdar, nmdars, calpain, tefb
8	7	research, core, center, support, translational
9	3	lung, ipf, expression, cells, methylation
10	4	mitochondrial, metabolic, redox, ros, bde

WordCloud


vivo
metastases
data
metastatic
develop
ppg
brain
core
bbb
cancer
ms
imaging
trfs
cells
lymph
tumor
csf
time
clinical
breast