Department of Computer Science

CS47300: Web Information Search and Management

Text Clustering

Prof. Chris Clifton

5 October 2020

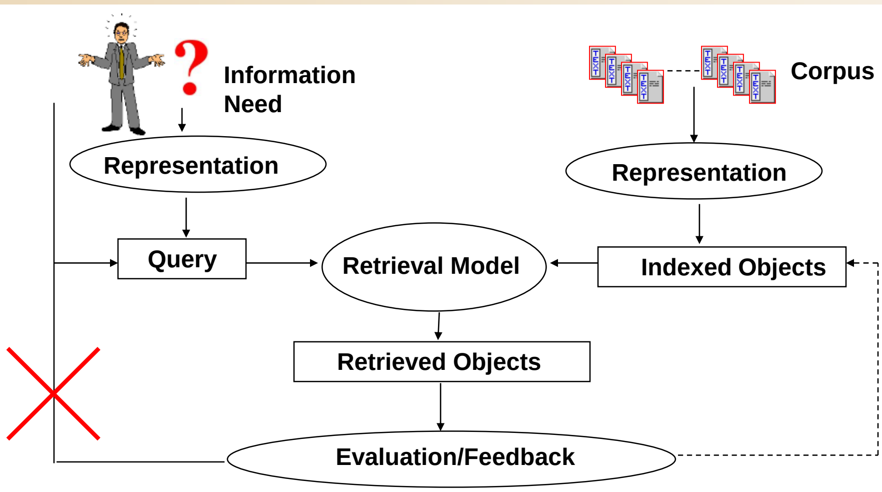
Borrows slides from Chris Manning, Ray Mooney and Soumen Chakrabarti

Indiana Center for Database Systems ™

Department of Computer Science

Midterm 1: All Aspects of AD-hoc IR

October 14 9am – October 15 9am, Gradescope



Pagerank, HITS, and relevance feedback saved for Midterm 2

PURDUE UNIVERSITY
Department of Computer Science

Clustering

- Document clustering
  - Motivations
  - Document representations
  - Success criteria
- Clustering algorithms
  - K-means
  - Model-based clustering (EM clustering)
  - Hierarchical clustering

7

PURDUE UNIVERSITY
Department of Computer Science

What is clustering?

- Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects
  - It is the commonest form of unsupervised learning
    - Unsupervised learning = learning from raw data, as opposed to supervised data where the correct classification of examples is given
  - It is a common and important task that finds many applications in IR and other places

8

PURDUE UNIVERSITY
Department of Computer Science

Why cluster documents?

- Whole corpus analysis/navigation
  - Better user interface
- For improving recall in search applications
  - Better search results
- For better navigation of search results
- For speeding up vector space retrieval
  - Faster search

9

PURDUE UNIVERSITY
Department of Computer Science

Navigating document collections

- Standard IR is like a book index
- Document clusters are like a table of contents
- People find having a table of contents useful

Index
Aardvark, 15
Blueberry, 200
Capricorn, 1, 45-55
Dog, 79-99
Egypt, 65
Falafel, 78-90
Giraffes, 45-59
…

Table of Contents
1. Science of Cognition
   1.a. Motivations
      1.a.i. Intellectual Curiosity
      1.a.ii. Practical Applications
   1.b. History of Cognitive Psychology
2. The Neural Basis of Cognition
   2.a. The Nervous System
   2.b. Organization of the Brain
   2.c. The Visual System
3. Perception and Attention
   3.a. Sensory Memory
   3.b. Attention and Sensory Information Processing

10

PURDUE UNIVERSITY
Department of Computer Science

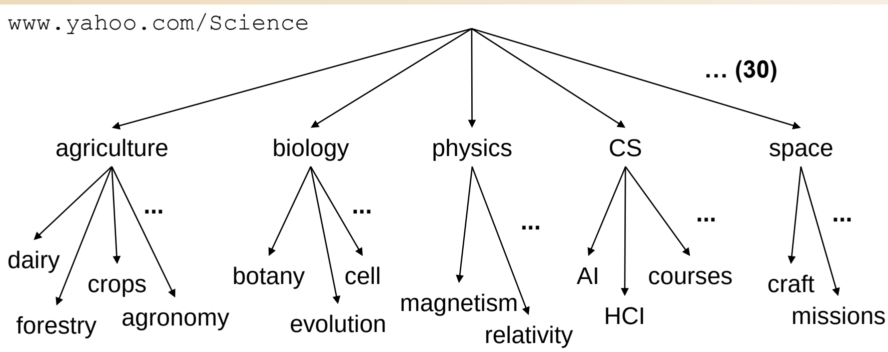
Corpus analysis/navigation

- Given a corpus, partition it into groups of related docs
  - Recursively, can induce a tree of topics
  - Allows user to browse through corpus to find information
  - Crucial need: meaningful labels for topic nodes.
- Yahoo!: manual hierarchy
  - Often not available for new document collection

11

PURDUE UNIVERSITY
Department of Computer Science

Yahoo! Hierarchy



12

© 2020 Christopher W. Clifton

PURDUE UNIVERSITY
Department of Computer Science

For improving search recall

- *Cluster hypothesis* - Documents with similar text are related
- Therefore, to improve search recall:
  - Cluster docs in corpus a priori
  - When a query matches a doc *D*, also return other docs in the cluster containing *D*
- Hope if we do this: The query “car” will also return docs containing *automobile*
  - Because clustering grouped together docs containing car with those containing *automobile*.

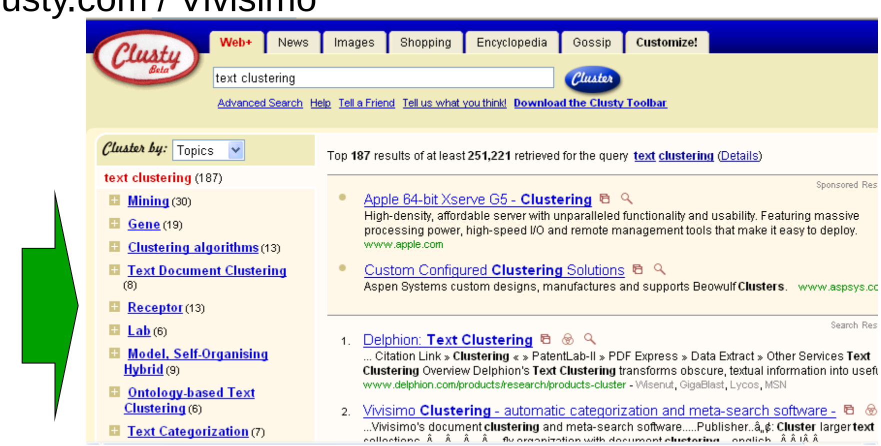
Why might this happen?

13

PURDUE UNIVERSITY
Department of Computer Science

For better navigation of search results

- Grouping search results thematically
  - clusty.com / Vivisimo



14

PURDUE UNIVERSITY
Department of Computer Science

Navigating search results (2)

- One can also view grouping documents with the same sense of a word as clustering
- Given the results of a search (e.g., *jaguar*, ***NLP***), partition into groups of related docs
- Can be viewed as a form of word sense disambiguation
- E.g., *jaguar* may have multiple senses:
  - The car company
  - The animal
  - The football team
  - The video game
- Recall query reformulation/expansion discussion

15

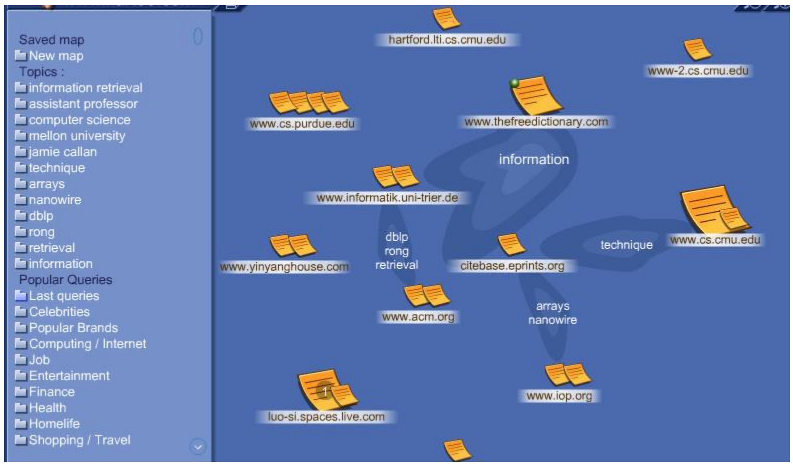
PURDUE UNIVERSITY
Department of Computer Science

Navigating search results (2)

16

For better navigation of search results

- And more visually: Kartoo.com



17

For speeding up vector space retrieval

- In vector space retrieval, we must find nearest doc vectors to query vector
- This entails finding the similarity of the query to every doc – slow (for some applications)
- By clustering docs in corpus a priori
  - find nearest docs in cluster(s) close to query
  - inexact but avoids exhaustive similarity computation

19

Department of Computer Science

What Is A Good Clustering?

- Internal criterion: A good clustering will produce high quality clusters in which:
  - the intra-class (that is, intra-cluster) similarity is high
  - the inter-class similarity is low
  - The measured quality of a clustering depends on both the document representation and the similarity measure used
- External criterion: The quality of a clustering is also measured by its ability to discover some or all of the hidden patterns or latent classes
  - Assessable with gold standard data

20

Department of Computer Science

External Evaluation of Cluster Quality

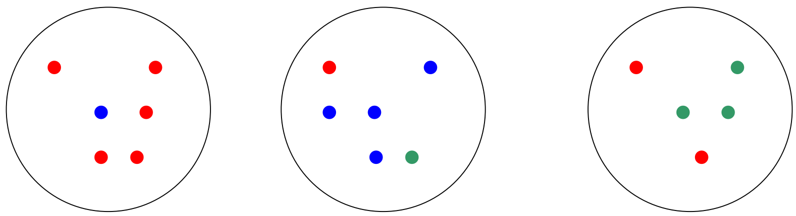
- Assesses clustering with respect to ground truth
- Assume that there are C gold standard classes, while our clustering algorithms produce *k* clusters, $\pi_1, \pi_2, \ldots, \pi_k$ with n_i members.
- Simple measure: purity, the ratio between the dominant class in the cluster π_i and the size of cluster π_i

$$\text{Purity}\,\pi_i = \frac{1}{n_i}\max_j(n_{ij}) \text{ for } j \in C$$

- Others are entropy of classes in clusters (or mutual information between classes and clusters)

21

Purity



Cluster I   Cluster II   Cluster III

Cluster I: Purity = 1/6 (max(5, 1, 0)) = 5/6

Cluster II: Purity = 1/6 (max(1, 4, 1)) = 4/6

Cluster III: Purity = 1/5 (max(2, 0, 3)) = 3/5

22

Issues for clustering

- Representation for clustering
  - Document representation
    - Vector space? Normalization?
  - Need a notion of similarity/distance
- How many clusters?
  - Fixed a priori?
  - Completely data driven?
    - Avoid “trivial” clusters - too large or small
      - In an application, if a cluster's too large, then for navigation purposes you've wasted an extra user click without whittling down the set of documents much.

23

© 2020 Christopher W. Clifton

Department of Computer Science

What makes docs “related”?

- Ideal: semantic similarity.
- Practical: statistical similarity
  – We will use cosine similarity, Docs as vectors
  – We will describe algorithms in terms of cosine similarity:
  Cosine similarity of normalized D_j, D_k:

$$\text{sim}(D_j, D_k) = \sum_{i=1}^{m} w_{ij} \times w_{ik}$$

  Also known as normalized inner product
  – For many algorithms, easier to think in terms of a distance (rather than similarity) between docs.

25

Department of Computer Science

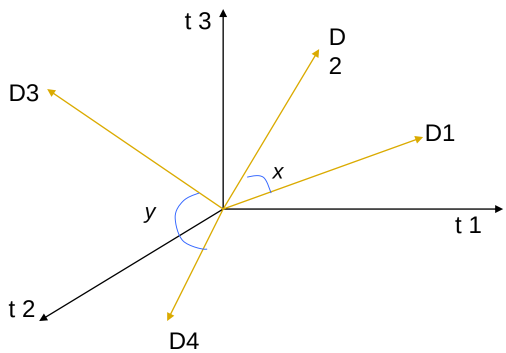
Recall doc as vector

- Each doc *j* is a vector of *tf*×*idf* values, one component for each term.
- Can normalize to unit length.
- So we have a vector space
  – terms are axis - aka *features*
  – *n* docs live in this space
  – even with stemming, may have 20,000+ dimensions
  – do we really want to use all terms?
    - Different from using vector space for search. Why?

26

PURDUE UNIVERSITY
Department of Computer Science

Intuition



Postulate: Documents that are "close together" in vector space talk about the same things.

27

PURDUE UNIVERSITY
Department of Computer Science

Clustering Algorithms

- Partitioning "flat" algorithms
  - Usually start with a random (partial) partitioning
  - Refine it iteratively
    - *k* means/medoids clustering
    - Model based clustering
- Hierarchical algorithms
  - Bottom-up, agglomerative
  - Top-down, divisive

28

© 2020 Christopher W. Clifton

Partitioning Algorithms

- Partitioning method: Construct a partition of n documents into a set of *k* clusters
- Given: a set of documents and the number *k*
- Find: a partition of *k* clusters that optimizes the chosen partitioning criterion
  - Globally optimal: exhaustively enumerate all partitions
  - Effective heuristic methods: *k*-means and *k*-medoids algorithms

How hard is clustering?

- One idea is to consider all possible clusterings, and pick the one that has best inter and intra cluster distance properties
- Suppose we are given n points, and would like to cluster them into k-clusters
  - How many possible clusterings?

$$\frac{k^n}{k!}$$

- Too hard to do it brute force or optimally
- Solution: Iterative optimization algorithms
  - Start with a clustering, iteratively improve it (e.g., K-means)

© 2020 Christopher W. Clifton