Scatter/Gather: A Cluster-based Approach

by Cutting et al.

Handan Kulan – Troya Çağıl Köylü

Outline

- Introduction
- Scatter/Gather Browsing
- Document Clustering (Previous + Purposed)
- Application to Scatter/Gather
- Conclusion

Introduction

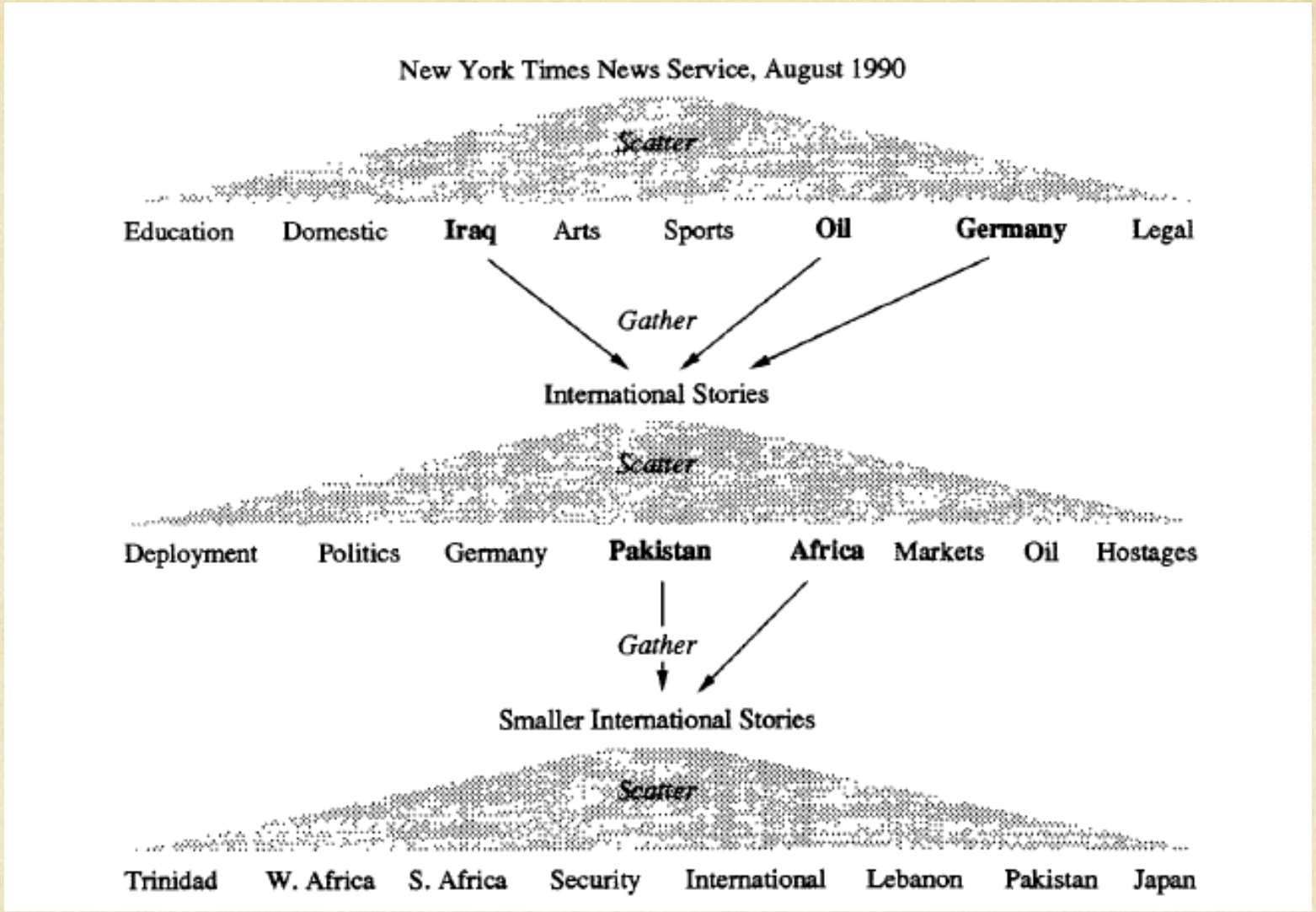
- Browsing vs. Searching

Textbook analogy: "If one has a specific question in mind, and specific terms which define that question, one consults the index [...]. However, if one is simply interested in gaining an overview, or has a general general question, one peruses the table of contents, which lays out the logical structure of the text." [1]

Scatter/Gather Browsing

1. A large document collection (after the general query)
2. Scattering → Gathering

3. Smaller document collection (that the user is interested in)


New York Times News Service, August 1990
Scatter
Education
Domestic
Iraq
Arts
Sports
Oil
Germany
Legal
Gather
International Stories
Scatter
Deployment
Politics
Germany
Pakistan
Africa
Markets
Oil
Hostages
Gather
Smaller International Stories
Scatter
Trinidad
W. Africa
S. Africa
Security
International
Lebanon
Pakistan
Japan

[1]

Previous Work on Clustering

- Document similarity measures
- Different clustering types
  - Flat partition

  Can be iterated for better results

  Some have rectangular ($O(kn)$) running times

  Computational benefits disappear with increasing size
  - Hierarchical clustering

  They are agglomerative & global ➡ Running time: $\Omega(n^2)$

Purposed Document Clustering

- Present 2 partitioning algorithms : Buckshot, Fractionation
  - Find k cluster centers
  - Assign each document in collection to a center
  - Refine the partition so constructed
- The result is a set P of k disjoint document
- Assume the existence of some algorithm which clusters well, but run slowly. Each of algorithms uses this cluster subroutine over small sets, and builds on its results to find the k centers

Finding Initial Centers

Buckshot

- Applies the cluster subroutine to a random sample to find centers
- Rectangular time clustering algorithmO(kn)

Fractionation

- Successive application of the cluster subroutine over fixed sized groups to find centers (more accurate)

Assigning Documents to Centers

- Iteratively, assign- to- Nearest algorithm assigns each document to the nearest center
  - Divide collection into k groups and in each group find nearest document looking similarity btw document and centers
- Split algorithms seperates poorly defined clusters into two well seperated parts
- Join merges clusters which are two similar

Application to Scatter/Gather

- *Offline* corpus

  Slower but more accurate clustering algorithm (very slow)

  OR Fractionation + Great deal of refinement (with Split, Join, Iterated Assign-to-Nearest)

- *Online* corpus

  Buckshot + Bare minimum of refinement (Not deterministic)

Conclusion

- Areas of Application
- Development