

clustering for next generation search engines

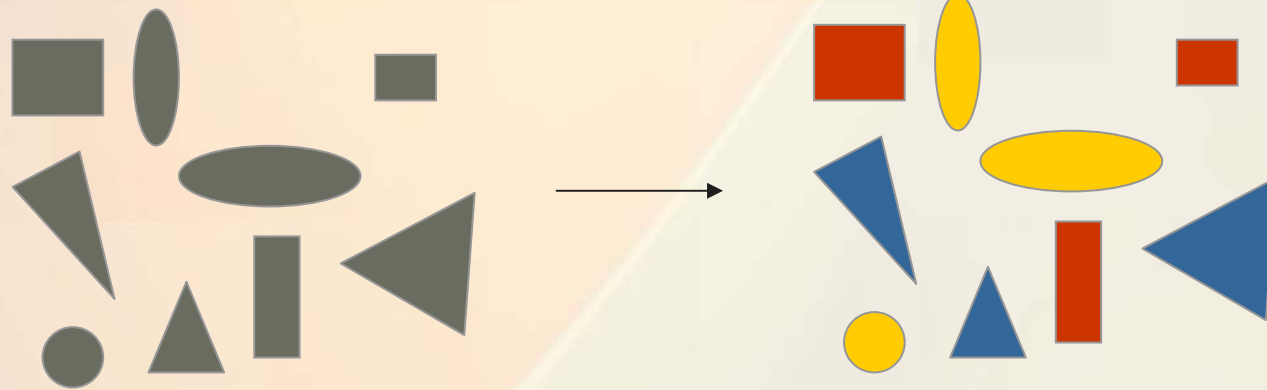
COALA : seeking alternate clusterings

Eric Bae

CSSE, The University of Melbourne

Brief Introduction to Clustering

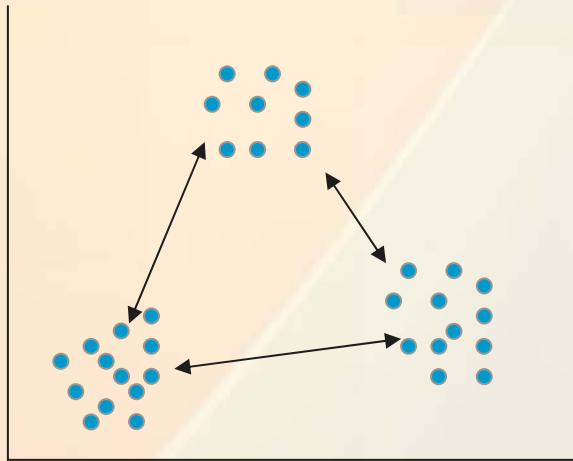
- **An unsupervised task of finding 'similar' objects and grouping them into 'clusters'**



Group similar objects together.....

Brief Introduction to Clustering

- **the objects are clustered to maximize inter-cluster similarity and minimize intra-cluster similarity**



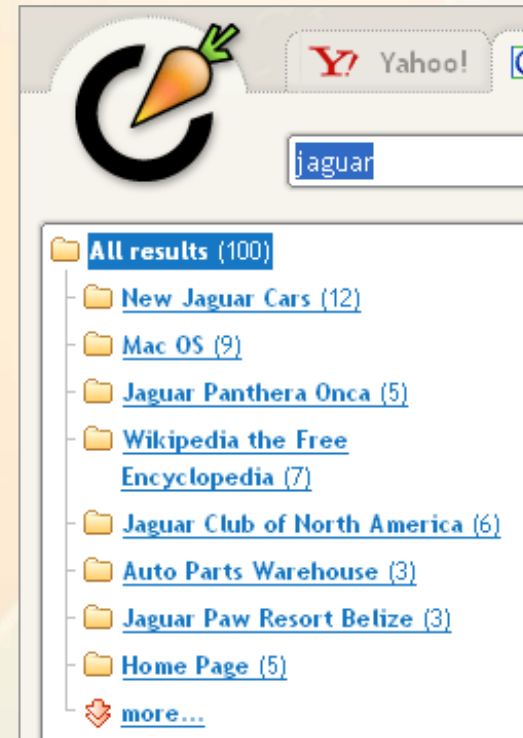
- **numerous algorithms and techniques available**

Clustering Documents (WWW)

- **Automatically categorizing documents into topical similarities (i.e. Yahoo!)**
- **Semantic disambiguation (i.e. Java, Jaguar, Storm)**
- **Provides more effective and efficient browsing**

Currently Available Tools

- **Clusty, Carrot2**



- **Also from Microsoft, Ask.com (Teoma)**

Digging Deeper : Alternate Clustering

- **Multiple clusterings are present in the dataset and current techniques retrieve only a *single clustering* as a solution**
- **Applying different algorithms is not appropriate**
- **Applying different initial parameters is not appropriate**

COALA

- **Constraint Orthogonal Average Link Algorithm**
- **“Given a pre-defined clustering, find another clustering which is dissimilar to the given clustering, and also of high quality”**
- **Two important requirements – Dissimilarity & Quality**

Experimental Results

Cluster 1

- Chinese restaurant
- Chinese cuisine
- Chinese new year
- Lunar new year celebration

Cluster 1

- Chinese restaurant
- Chinese cuisine
- Italian restaurant
- Papa Gino's pizza & Italian food

Cluster 2

- Italian restaurant
- Papa Gino's pizza & Italian food
- Liberation day for Italy
- Liberazione public holiday Italy

Cluster 2

- Chinese new year
- Lunar new year celebration
- Liberation day for Italy
- Liberazione public holiday Italy

Conclusion

- **Clustering can aid users in finding answers more effectively**
- **Alternate clustering can re-organize the groups in different perspectives**
- **Combining original and alternate clusterings together**