Categorization and clustering

Hercules Dalianis
DSV-SU-KTH
e-mail:hercules@kth.se
070-568 13 59 / 08-674 75 47

Categorization/Clustering

• Difference between categorization and clustering?
• Categorization is deterministic or predefined by someone.
• Clustering is indeterministic and not decided by anyone

Contents

• Categorization
• Clustering
• Some cluster demos
• Multi text summarizations
• Exam queries

Categorization in search engines

• Automatic language recognition- Swedish, English, Chinese, etc
• Document type HTML, Word, Excel, PDF etc.
• Date
• Categories in form of server, domain or language
Categorization

- Domestic
- Foreign
- Sports
- Economy
- Culture
- etc

Categories not static

- Categories arise continuously
- Categories disappear
  - The predefined categories become quickly obsolete
  - E-commerce category that did not exist 10 years ago
- Automatic categorization through good examples.
- Though most people do not use categories when searching.

Clustering

- Clustering - automatic categorization
- Clustering when searching
- Find new connections between documents.
- Can one group 100,000 news text in some way?

Clustering methods

- Clustering is tricky
- According to what measurements should one cluster?
- Among what methods?
  - Hierarchic clustering
  - K-mean method
Hierarchic clustering

- Take one all-embracing cluster and divide it into smaller clusters.
- Better quality on clusters with hierarchic than with non hierarchic.
- But slow clustering

K-mean algorithm

- Select K initial staring points
- Let all documents group around the K initial starting points. Move the centre of gravity to the center.
- Regroup the documents.
- Move the centre of gravity to the new center of gravity. Regroup the documents until new balance is obtained.

K-mean (cont)

- Predefined number of clusters
- Non hierarchal
- Faster clustering
- Worse quality on clusters.
- Combine hierarchical clustering and K-mean
- Is efficient, is called K-mean bisecting

Similarity measurement for clusters

- Vector space model
- The word vectors product is the similarity measurement
- Word vector - all words in the document,
- Cosine = The angle between the word vectors
- \( d_1 = \{ \text{all words in a document}_1 \} \)
- \( d_2 = \{ \text{all words is a document}_2 \} \)
- \( d_1 \cdot d_2 = \cos(d_1, d_2) = \text{the angle between } d_1 \text{ and } d_2 \)
Refinement

- Remove stop words, *but, with, on, in, maybe* etc.
- Collapse or aggregate all tensed forms to one form (Stemming/lemmatization)
- Use word frequencies
- Compound splitting
- Normalization for length of document.

Other measures

- Number of words that the texts share.
- Number of words that are shared and the word frequencies.

Cluster quality in relation to reference

- Entropy
  - The order in the cluster in relation to expected order. Low entropy - high order.
- F-value
  - $F-value = 2 \times \text{Recall} \times \text{Precision} / \text{Recall} + \text{Precision}$

Precision/Recall

- Recall = Number of found relevant documents / Total number of relevant documents
- Precision = Found relevant documents / Total number of found documents
**Bi-secting K-mean**

1. Take a large cluster
2. Divide it into two clusters (K=2)
3. Use the two clusters as seeds to K-mean. Repeat 1 and 2 to satisfied.

**Vivisimo the clustering search engine**

- The search engine Vivisimo http://vivisimo.com has a clusterer.
- Takes the first 500 best hits and clusters them.

**Categorisation**

- Overlapping or non overlapping categories
- Ordered or non-ordered categories
Clusty news

- Try search words
  - Stockholm
  - Europe
  - Iraq
  - Harbin

Clustering news text

- The clusterer of Rosell (2002, 2003) clustered 5 000 Swedish news text
- Used stemming and compound splitting
- Stemming improved the cluster results with about 5 percent.
- Compound splitting improves the results with 10 percent and a combination gives 13 percent improvement. (Rosell 2003).

Evaluation news text clustering

- Rosell could compare his automatic clusters with the predefined categories from where the news text originated from
- From this he could assess the quality of the clusters.
- Read more in Rosell 2003
The Infomat project

- KTH and the Karolinska Institute
- Twin registry with 45,000 twins
- Query: "Vad arbetar du med?" What do you work with?
- Manual categorization by SCB (Sweden statistics) takes 3 months
- Automatic clustering takes less than an hour. (Show result)

Automatic categorisation

- Twin registry
- "Vad arbetar du med?" What do you work with?
  45,000 answers in short texts.
- K-mean bisecting clustering algorithm (Non-hierarchical algorithm top-down)
- Create at least two initial centers and let then the algorithm find more sub clusters. The center is moved as more sub clusters are created. After a while is equilibrium created.
- http://www.nada.kth.se/~rosell/cluster/result-occupationF-1087930997527.html

Clustering Result

File file: occupationFile.txt
43341 (43375) documents with 6179 different words.
Average number of different words per document: 9.27 (min: 1 max: 51)
Average number of documents each different word occurs in: 65.01 (min: 2, max: 9852)

Some of the most common words:

Some of the most "distinguishing" words:

Clustering Result

- k-means
  - (3532) words: 0.04, bitr: 0.04, skit: 0.03, har: 0.03, harn: 0.03.
  - (3532) words: 0.03, inv: 0.03, art: 0.03, harr: 0.03, skit: 0.03.
  - (3532) words: 0.03, har: 0.03, art: 0.03, sten: 0.03, harn: 0.03.
  - (3532) words: 0.03, inv: 0.03, art: 0.03, harr: 0.03, skit: 0.03.
  - (3532) words: 0.03, har: 0.03, art: 0.03, sten: 0.03, harn: 0.03.
  - (3532) words: 0.03, inv: 0.03, art: 0.03, harr: 0.03, skit: 0.03.
  - (3532) words: 0.03, har: 0.03, art: 0.03, sten: 0.03, harn: 0.03.
  - (3532) words: 0.03, inv: 0.03, art: 0.03, harr: 0.03, skit: 0.03.
  - (3532) words: 0.03, har: 0.03, art: 0.03, sten: 0.03, harn: 0.03.
  - (3532) words: 0.03, inv: 0.03, art: 0.03, harr: 0.03, skit: 0.03.
  - (3532) words: 0.03, har: 0.03, art: 0.03, sten: 0.03, harn: 0.03.

Evaluation interview forms

- The F-value is 86% and 82% respectively of two manual classifications AMSYK & YK80
  - F-value = 2 × P x R/(P+R)
- The clusters have a gain 77% and 71% of the information gain. (Entropy)
- The precision is 85% and 82% respectively of the precision of the classifications.
Semi automatic categorisation

- Give some positive examples on good categories.
- Automatically sort remaining texts depending on similarity into the correct category.

Multi-text summarization

Columbia Newsblaster
Many news articles summarized into one article
http://www1.cs.columbia.edu/nlp/newsblaster/
Multisum-algorithm

There exist many algorithms for this but one can use SweSum text summarization system as a base.
1) Summarize all texts separately
2) Take all summaries and summarize them to one text.

Multi-text sum-algorithm (cont.)

• But some texts might not be relevant, we must remove them.
• Cluster text
• Each cluster should contain only relevant texts.

4-level process

1) Cluster say 100 texts in say 10 clusters. (Each cluster can produce one multi-text summary)
2) Rank each text in each cluster
3) Summarize each text and compile, (concatenate) each text together into one larger text.
4) Summarize each large text into one multi text summary
   • => We obtain 10 multi-text summaries.

Summary

• Categorization is necessary
• Categorization requires lots of manual work
• Clustering maybe to complicated
• Semi automatic categorization
• Clustering to find relations between texts
• Multi-text summarization uses clustering