Interconnected text documents search and clustering

A. Kozitsyn, S. Afonin

Moscow State University

Search engines problems

- Low search relevance

Search improvement techniques

- Logical structure usage (semistructured data)
- Document clustering and classification
- Metadata usage
- Semantic (linguistic) analysis
- Link structure analysis
- Automated learning (user feedback)

Link structure analysis

"The network structure of a hyperlinked environment can be a rich source of information about the content of the environment, provided we have effective means for understanding it." (Kleinberg, 1998)

Algorithms for "authorities" search:

- Kleinberg's algorithm - hyperlink is the resource authority confirmation
- Google PageRank - random walks
- SALSA, HITS, ...

Data clustering

Clustering is the unsupervised classification of objects (data items, feature vectors) into groups (clusters).

The main clustering approaches are:

- Hierarchical
- Partitional

- Agglomerative *vs.* divisive
- Hard *vs.* fuzzy
- Deterministic *vs.* stochastic

Graph-theoretic clustering

Similarity matrix W may be considered as incident matrix of an undirected graph G=(V,E) with weighted edges.
A set of edges that separates the graph into two disconnected parts is called an ***edge-cut***.

Clustering is a recursive graph partitioning in accordance with the criteria of the goodness of the partition (i.e. sum of edge-

Text document clustering

- Document representation
  - bag of words, tf.idf
  - Latent Semantic Index (LSI)
  - Tesaurus concepts (WordNet)
- Similarity measure
  - cosine normalisation (scalar product)
  - distance in vector space
- Clustering algorithm
  - a relevant data clustering algorithm

Hypertext clustering (He)

Similarity metric (He, Zha, Ding, Simon 2002):

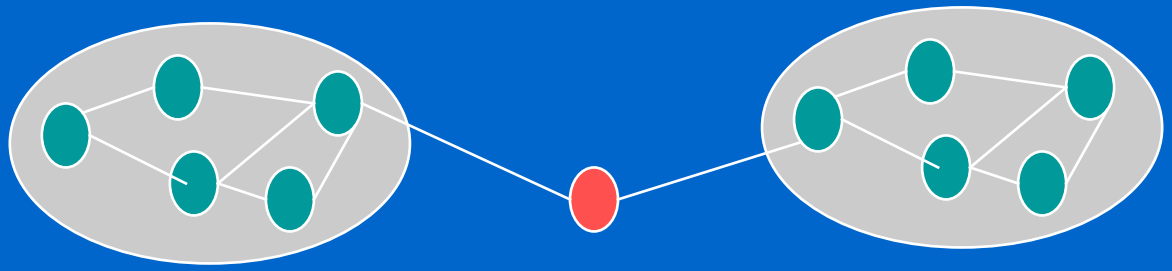
- Hyperlink structure **A**
- Textual similarity **S**
- Co-citations **C**

$$W = \alpha \frac{A \otimes S}{\|A \otimes S\|_2} + (1 - \alpha) \frac{C}{\|C\|_2}$$

Clustering by graph partitioning using *normalised-cut*; authority search in clusters

He’s approach properties

- Only interconnected documents form a cluster (by definition of graph partitioning)
- Similar documents that are divided by a non-relevant one could be assigned to different clusters



$$W = \alpha \frac{A \otimes S}{\|A \otimes S\|_2} + (1-\alpha)\frac{C}{\|C\|_2}$$

Possible scenarios

1. Classification learning (by clustering or test set)
2. Documents classification
3. Classification refinement by link structure analysis
4. Initial query (keywords) processing
5. User feedback, query refinement (similar documents)

Clustering vs. Classification

- Classification is much faster then clustering. Model learning process only a small fraction of documents.
- Classes are fixed at the beginning.

Probabilistic classification

- Cluster definition:
  For each term
  - P_1 = **P**{t *in* d | d *in* C}
  - P_2 = **P**{t *in* d | d *not in* C}
- Document weight $\mathbf{P_0}$ is calculated iteratively in accordance with the following formula :

$$\frac{P_1^k \cdot P_0^{k-1}}{P_1^k \cdot P_0^{k-1} + P_2^k \cdot (1 - P_0^{k-1})}$$

Classification refinement

- A clusters weight (probability) vector is assigned to each document
- For any cluster we can construct a vertex-weighted graph G=(V,E) and then apply “Kleinberg’s algorithm”
- The result is a vector of corrected documents weights.

Query processing

Given a number of words as a query the set of relevant clusters can be constructed and their content may be considered as an answer.

For relevant cluster search the “Kleinberg’s algorithm” can be applied to clusters similarity matrix.

Conclusion

- Link structure might be useful for clustering or classification refinement.
- The described search method require large internet coverage.
- Surrogated data can not be used for testing purposes.