Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data

Stanisław Osiński   Dawid Weiss

Institute of Computing Science
Poznań University of Technology

May 20th, 2004

About various goals of evaluation...

- Reconstruction of a predefined structure
  Test data: merge-then-cluster, manual labeling
  Measures: precision-recall, entropy measures...

- Labels "quality", descriptiveness
  User surveys, click-distance methods

What types of "errors" can an algorithm make structure-wise?

- Misassignment errors (document→cluster)
- Missing documents in a cluster
- Incorrect clusters (unexplainable)
- Missing clusters (undetected)
- Granularity level confusion (subcluster domination problems)

We tried to answer the following questions:

- Clusters' structure:
  1. Is Lingo able to cluster similar documents?
  2. Is Lingo able to highlight outliers and "minorities"?
  3. Is Lingo able to capture generalizations of closely-related subjects?
  4. How does Lingo compare to Suffix Tree Clustering?
- Quality of cluster labels
  - Are clusters labelled appropriately? Are they informative?

- Data set: a subset of the Open Directory Project
- Rationale:
    - Human-created and maintained structure
    - Human-created and maintained labels
    - Descriptions resemble search results (snippets)
    - Free availability
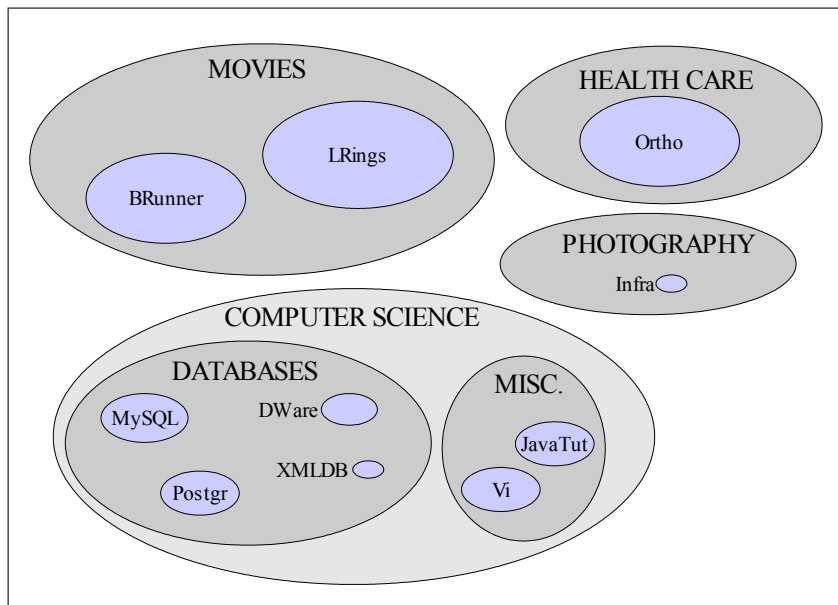
Test sets

Test sets were combinations of categories designed to help in answering the set of questions.

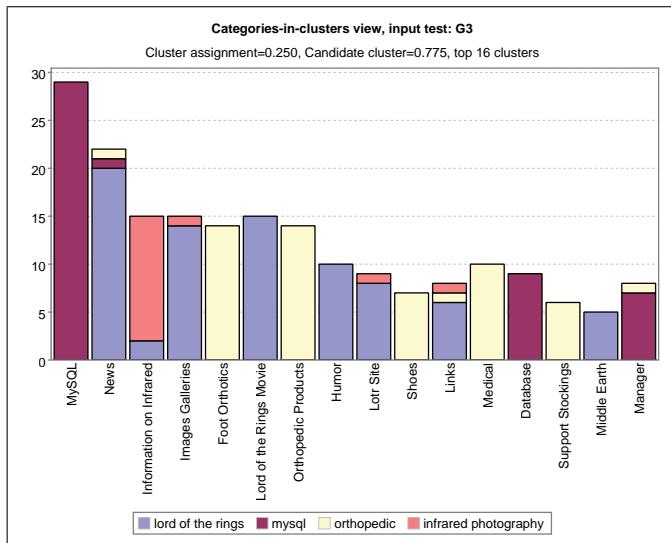| Identifier	Merged categories	Test set rationale
G1	*LRings*, *MySQL*	Separation of two unrelated categories.
G2	*LRings*, *MySQL*, *Ortho*	Separation of three unrelated categories.
G3	*LRings*, *MySQL*, *Ortho*, *Infra*	Separation of four unrelated categories, highlighting small topics (*Infra*).
G4	*MySQL*, *XMLDB*, *DWare*, *Postgr*	Separation of four conceptually close categories, all connected to database.
G5	*MySQL*, *XMLDB*, *DWare*, *Postgr*, *JavaTut*, *Vi*	Four conceptually very close categories (database) plus two distinct, but within the same abstract topic (computer science).
G6	*MySQL*, *XMLDB*, *DWare*, *Postgr*, *Ortho*	Outlier highlight test – four dominating conceptually close categories (databases) and one outlier (*Ortho*)
G7	All categories	All categories mixed together. Cross-topic cluster detection test (movies, databases).

- Lingo's implementation $\rightarrow$ Carrot[2] framework
- The algorithm's thresholds:
  - Fixed at "good guess" values (same as those used in the on-line demo)
  - Stemming and stop-word detection applied to the input data

Method of analysis

Manual investigation of document-to-cluster assignment charts.

- Helps understand the internal structure of results
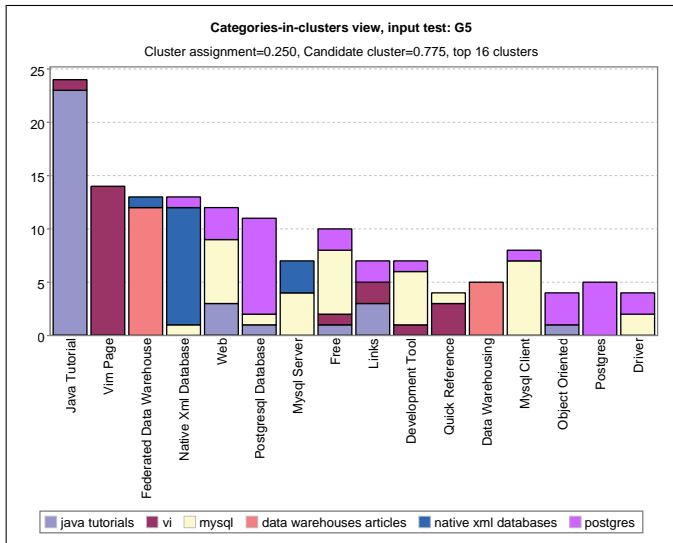- Prevents compensations inherent in aggregative measures

Categories-in-clusters view, input test: G3

Cluster assignment=0.250, Candidate cluster=0.775, top 16 clusters

Legend: lord of the rings | mysql | orthopedic | infrared photography

G1–G3: clear separation of topics, but with some extra clusters

G1: granularity problem

Categories-in-clusters view, input test: G5

Cluster assignment=0.250, Candidate cluster=0.775, top 16 clusters

Legend: java tutorials, vi, mysql, data warehouses articles, native xml databases, postgres

G5: misassignment problem
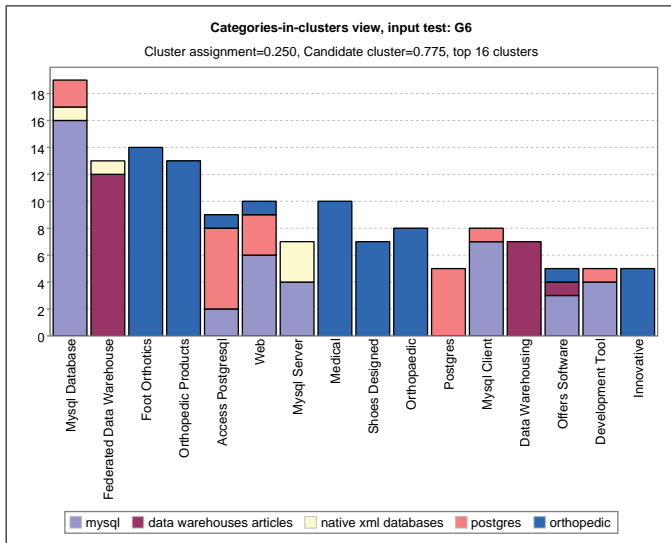
Ortho category (outlier), *XMLDB* consumed by MySQL!

Infra category (outlier)
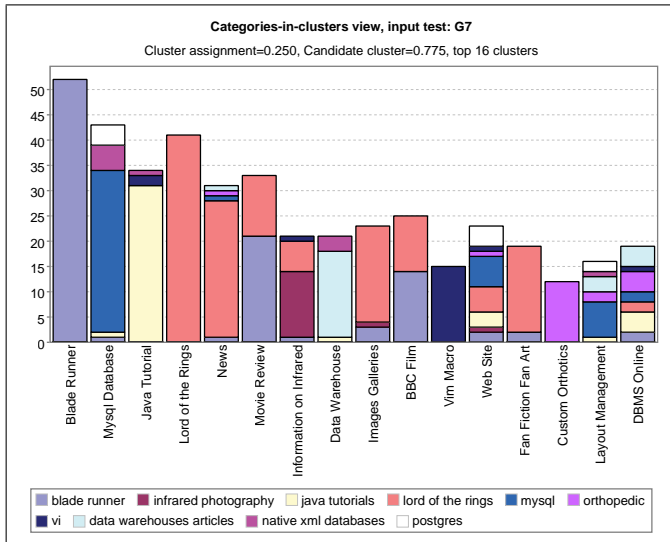
XMLDB category (outlier)

Categories-in-clusters view, input test: G7

Cluster assignment=0.250, Candidate cluster=0.775, top 16 clusters

Legend:
- blade runner
- infrared photography
- java tutorials
- lord of the rings
- mysql
- orthopedic
- vi
- data warehouses articles
- native xml databases
- postgres

Categories: Blade Runner, Mysql Database, Java Tutorial, Lord of the Rings, News, Movie Review, Information on Infrared, Data Warehouse, Images Galleries, BBC Film, Vim Macro, Web Site, Fan Fiction Fan Art, Custom Orthotics, Layout Management, DBMS Online
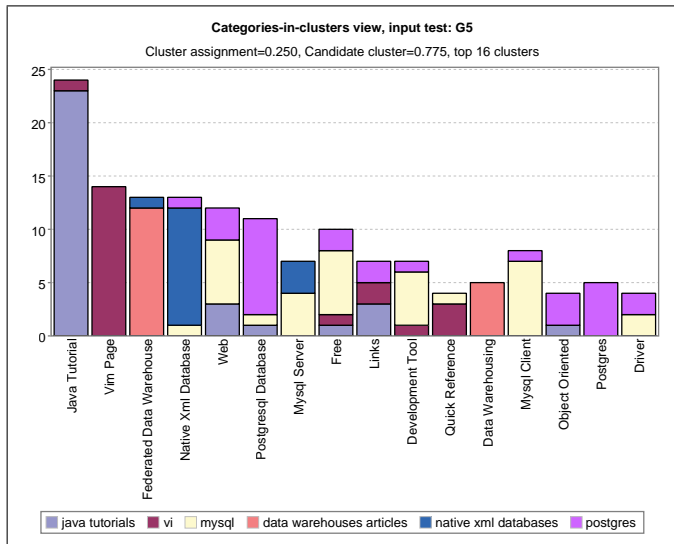
"movie review" cluster is a generalization, but. . .

Categories-in-clusters view, input test: G7
Cluster assignment=0.250, Candidate cluster=0.775, top 16 clusters

Legend: blade runner, infrared photography, java tutorials, lord of the rings, mysql, orthopedic, vi, data warehouses articles, native xml databases, postgres
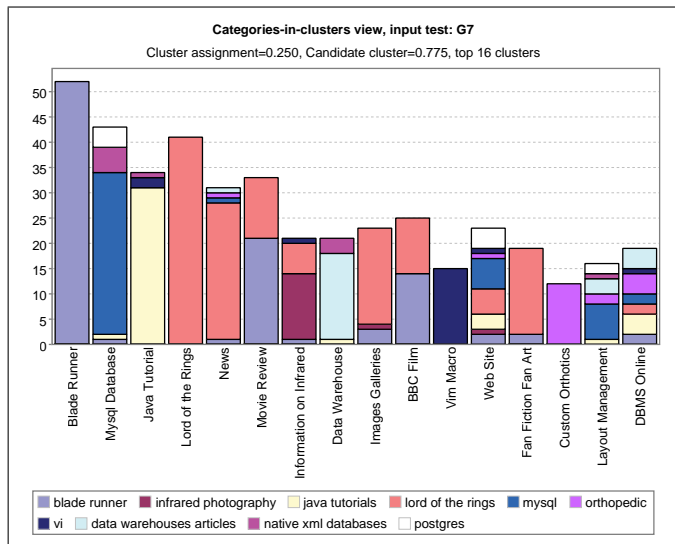
Clusters are usually orthogonal with SVD, so no good results should be expected in this area.
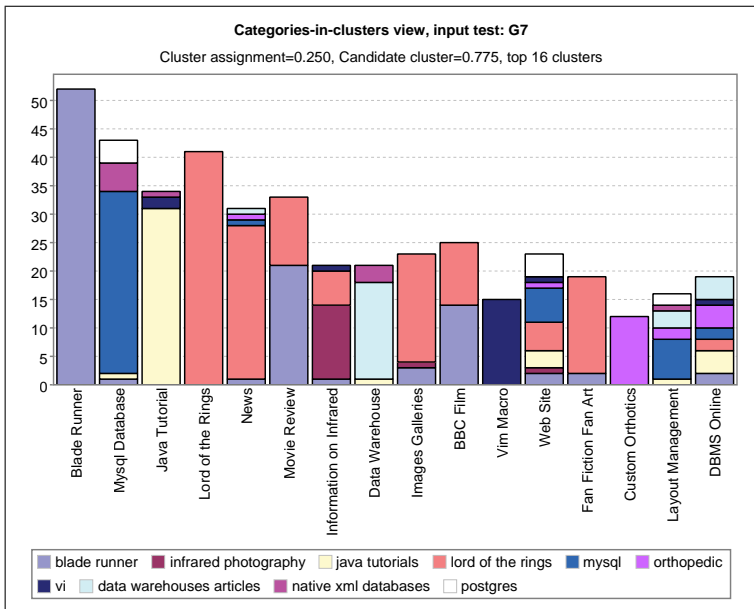
Categories-in-clusters view, input test: G7

Cluster assignment=0.250, Candidate cluster=0.775, top 16 clusters

Categories (legend): blade runner, infrared photography, java tutorials, lord of the rings, mysql, orthopedic, vi, data warehouses articles, native xml databases, postgres

Categories-in-clusters view, input test: G7

STC algorithm, Top 16 clusters
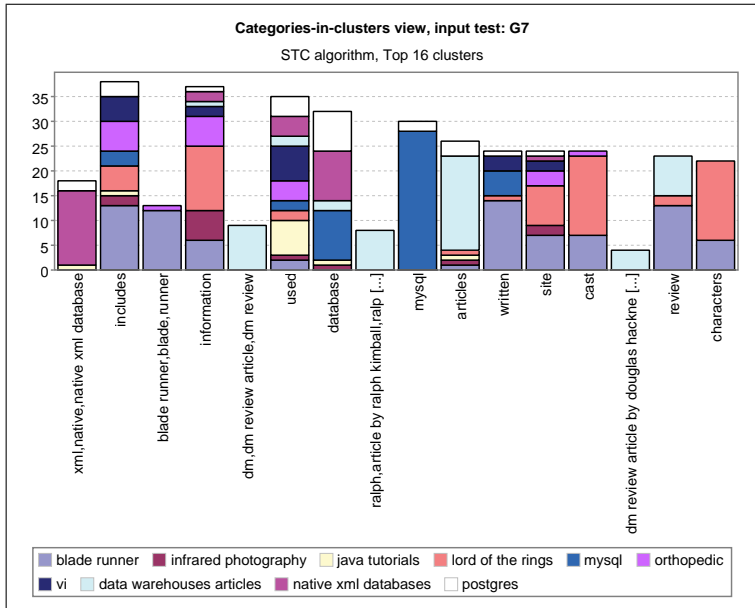
- Size-dominated clusters in STC
- Cluster labels much less informative
- Common-term clusters in STC

- Performed manually
- Problems:
  - Single term labels usually ambiguous or too broad ("news", "free")
  - Level of granularity usually unclear (need for hierarchical methods?)

Can these conclusions be derived using formulas?

We think so: cluster contamination measures might help.

Online demo

A nice form of evaluation (although scientifically doubtful), is the online demo's popularity and feedback we get from users.

carrot²

Proces: | Google, English stemming, STC, Dynamic Tree | ▼ | Pobrać

Sort: [flat] [group] [score]

All groups (74)
- 庶正然 (5)
- 網趣翻譯 (9)
- 薔薇膝知識庫 表示 (5)
- 薔薇華 (4)
- 街出 (4)
- 小薔華 圍終 小薔華 小薔華大事紀 (2)
- 新資雜誌 商業周刊838 Mon Dec (2)
- 網站內容分融之嫌疑概寫 (2)
- 網路100強 中國領先 台灣勢第的100強 (2)
- PC Home Online 網路家庭·重要新聞 (2)
- 薔薇膝樂位科技知識庫 (3)
- 國拓文教基金會 (3)
- Yahoo (2)
- (Other) (55)

sub topics

---

3   **一月主題** 🗗

... 二月，紀念二二八，吳俊興、蕭景燧 . 三月，三八婦女 節，庶正然、蕭景燧 . 四月，慶祝兒童
http://home.yam.org.tw/jan/

5   **P080** 🗗

... 照顧僑胞，當然有正面的意義，」薔薇膝數位科技公司的創辦人之一，也是淡江大學資訊系教
http://www.new7.com.tw/weekly/old/605/605-080.html

5   **新新聞510期:網路國發會的推動背景與前景** 🗗

... 也針對柯林頓遺任之，後的使命窟追殺打。據番薔 薇網站成員 蕭景燧 指出，最近的美. 國網路
http://www.new7.com.tw/weekly/old/510/article089.html

7   **資訊圖書館** 🗗

... 名人專訪- 梁小萍（Accenture台灣區總裁）. 名人專訪- 薔薇膝知識長 蕭景燧 (下). 名人專
http://infolib.ncl.edu.tw/info/in6.asp

9   **國拓文教基金會/關於我們** 🗗

... 董事袁螃嬈，淡水河衡接電台台長. 董事蘇魚智，立法委員. 董事 蕭景燧，薔薇膝數位科技知
http://www.frontier.org.tw/about.htm

10   **e天下網站-從八里到墾丁，玩出全台「行動力」** 🗗

再搭豪華遊艇 你愛 週即堆也有無處上網的批戰 . 薔薇膝知識長 蕭景燧 . 也變淡水河岸

http://carrot.cs.put.poznan.pl

Thank you. Questions?