

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

Some background: how to evaluate an SRC algorithm?

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

Some background: how to evaluate an SRC algorithm?

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)

Evaluation of Lingo's performance

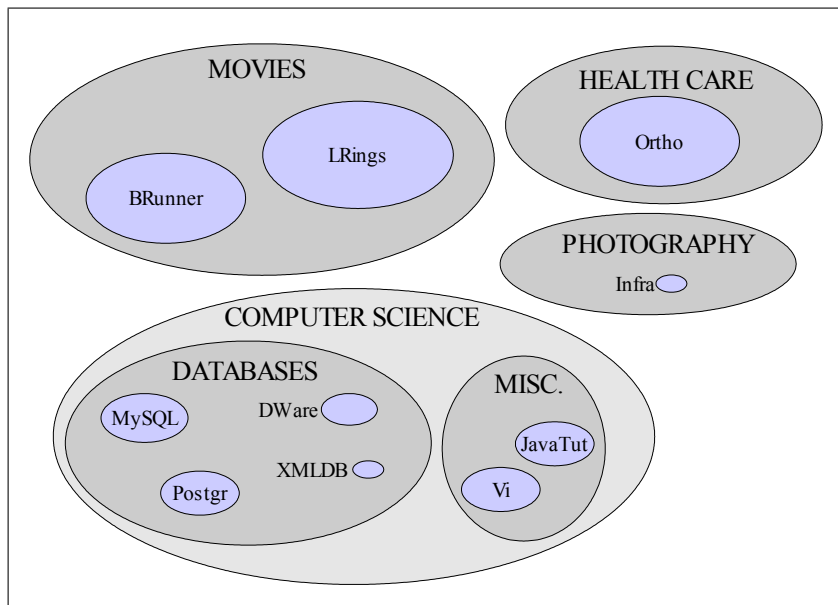
We tried to answer the following questions:

- 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 for the experiment

- 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

ODP Categories chosen for the experiment



Test sets for the experiment

Test sets

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

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

The experiment

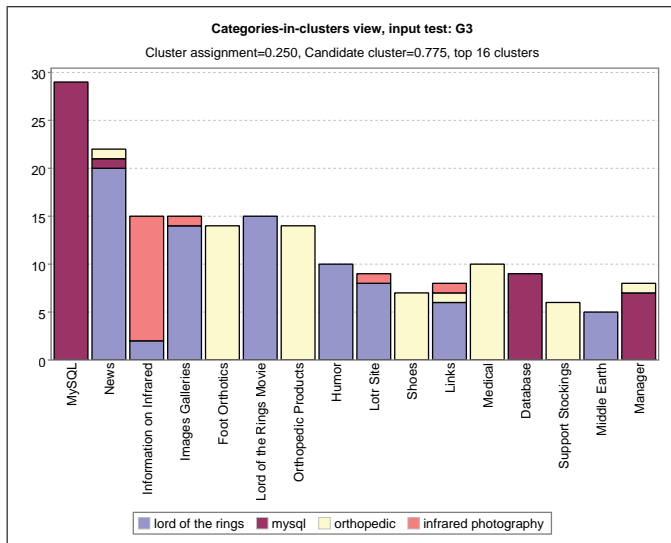
- Lingo's implementation → Carrot² framework
- The algorithm's thresholds:
 - Fixed at “good guess” values (same as those used in the on-line demo)
 - Stemming and stop-word detection applied to the input data

Method of analysis

Manual investigation of document-to-cluster assignment charts.

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

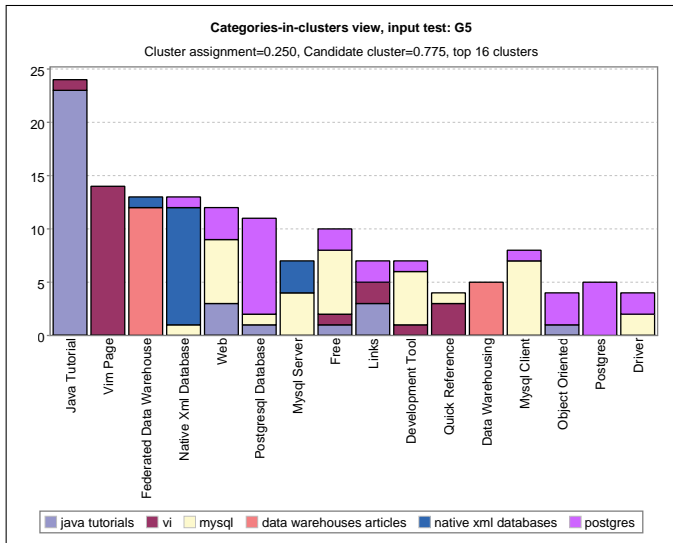
→ Is Lingo able to cluster similar documents?



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

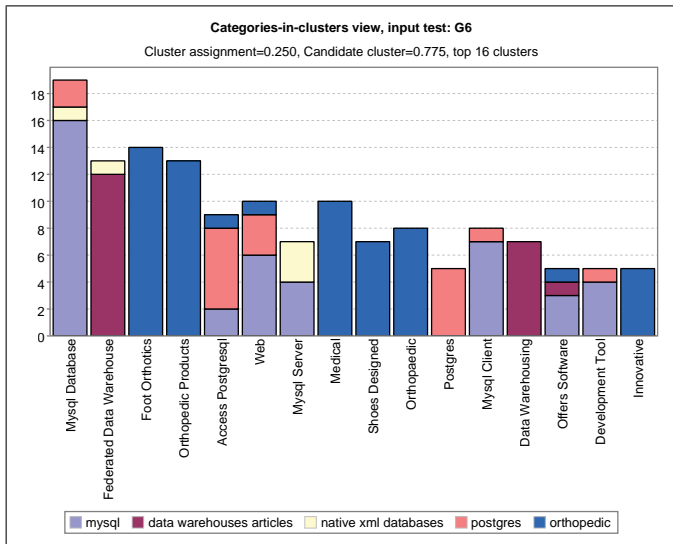
G1: granularity problem

→ Is Lingo able to cluster similar documents?



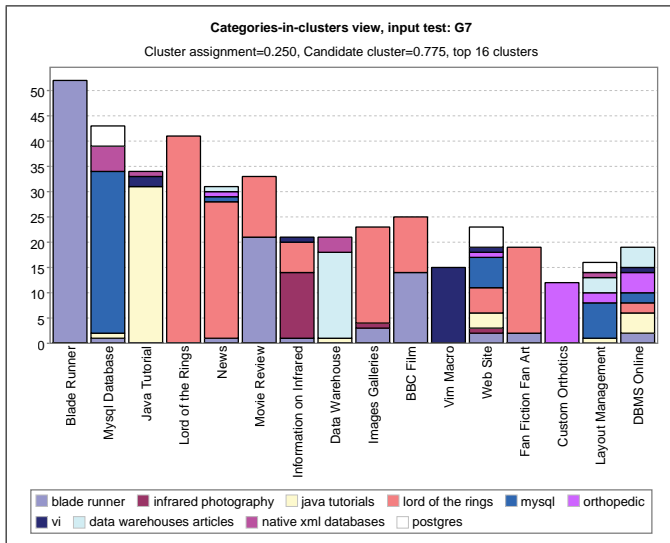
G5: misassignment problem

→ Is Lingo able to highlight outliers and “minorities”?



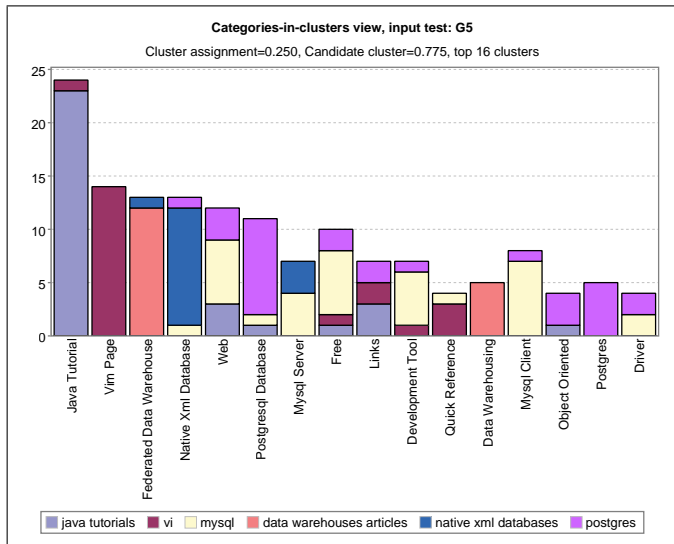
Ortho category (outlier), XMLDB consumed by MySQL!

→ Is Lingo able to highlight outliers and “minorities”?



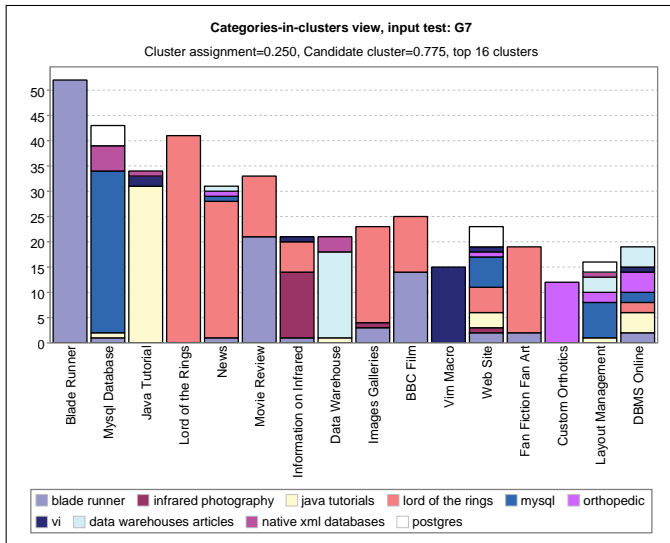
Infra category (outlier)

→ Is Lingo able to highlight outliers and “minorities”?



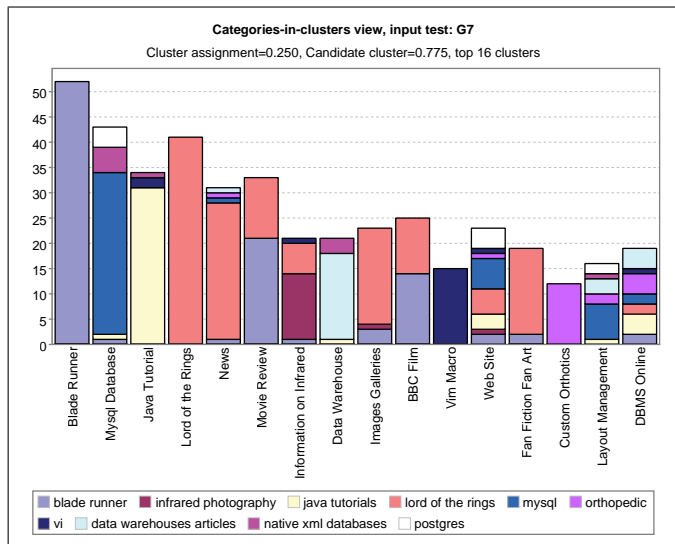
XMLDB category (outlier)

→ Is Lingo able to capture generalizations?



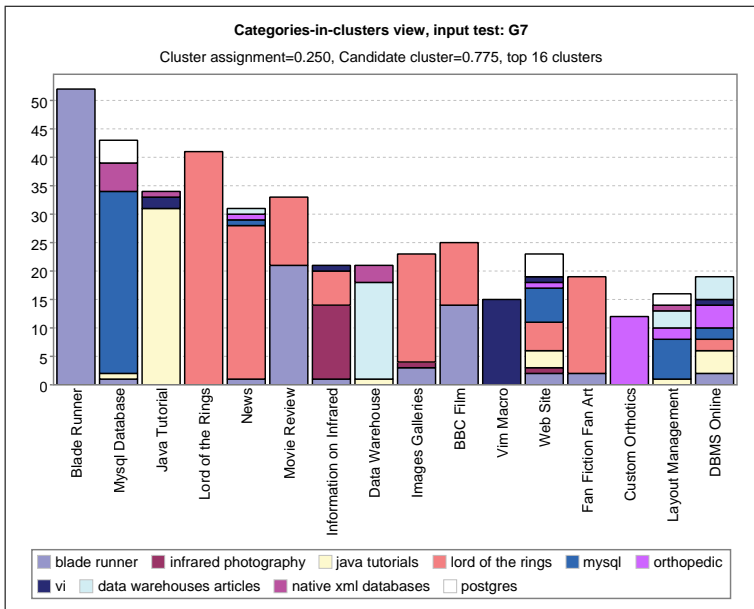
“movie review” cluster is a generalization, but. . .

→ Is Lingo able to capture generalizations?

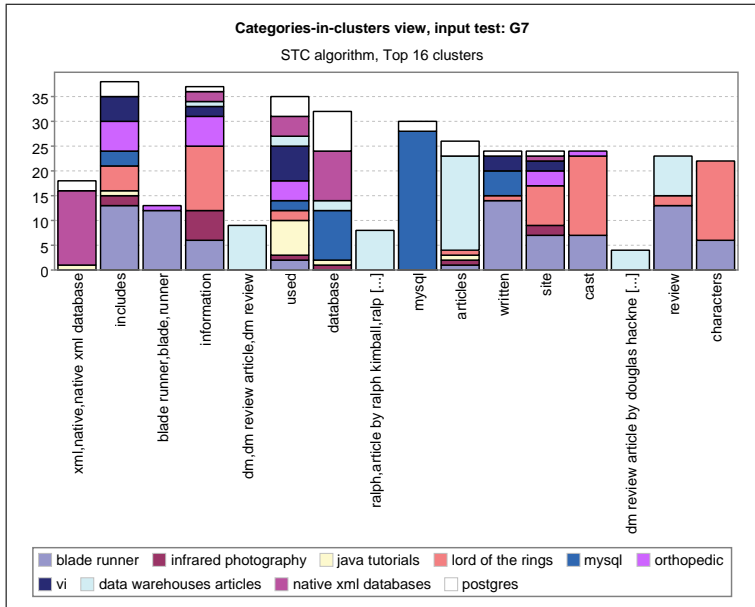


Clusters are usually orthogonal with SVD, so no good results should be expected in this area.

→ How does Lingo compare to Suffix Tree Clustering?



→ How does Lingo compare to Suffix Tree Clustering?



Key differences between Lingo and STC

- Size-dominated clusters in STC
- Cluster labels much less informative
- Common-term clusters in STC

Cluster labels quality

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

A word about analytical comparison 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.



komponenty administracja duże zapytanie demonstracja czym jest Carrot?

Proces: Google, English stemming, STC, Dynamic Tree

Pobrać

Sort: [flat] [group] [score]

All groups (74)

sub topics

- 陳正然 (5)
- 爾維經理 (9)
- 蕃薯藤知識長 表示 (5)
- 蕃薯藤 (4)
- 新出 (4)
- 小蕃薯 關於小蕃薯 小蕃薯大事記 (2)
- 新浪雜誌 商業周刊 838 Mon Dec (2)
- 網站內容分鐘之靈犀觀察 (2)
- 網路100強 中國領先 台灣勁旅的100強 (2)
- PC Home Online 網路家庭-重要新聞 (2)
- 蕃薯藤數位科技知識長 (3)
- 開拓文教基金會 (3)
- Yahoo (2)
- (Other) (55)

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, 陳正然、蕭景燧, 四月, 慶祝兒童
<http://home.yam.org.tw/jan/>

5 | P080

... 照原價絕, 當然有正面的意義,」蕃薯藤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>

6 | 新新聞510期:網路國發會的推動背景與前景

... 也針對柯林頓連任之後的使命窮追猛打。據蕃薯藤網站成員蕭景燧指出, 最近的美、國網
<http://www.new7.com.tw/weekly/old/510/article089.html>

7 | 資訊圖書館

... 名人專訪- 黎小萍 (Accenture台灣區總裁), 名人專訪-蕃薯藤知識長 蕭景燧 (下), 名人
<http://infolib.ncl.edu.tw/info/in6.asp>

9 | 開拓文教基金會/關於我們

... 董事袁姍姍, 淡水河廣播電台台長, 董事蘇煥智, 立法委員, 董事蕭景燧, 蕃薯藤數位科技知
<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」

... 兩地應該很佳, 你是 誰? 那本有無線上網站設置, 蕃薯藤知識長 蕭景燧, 世界水網

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

Thank you. Questions?