Document Snippet Generation

Troya Çağıl Köylü 21002018

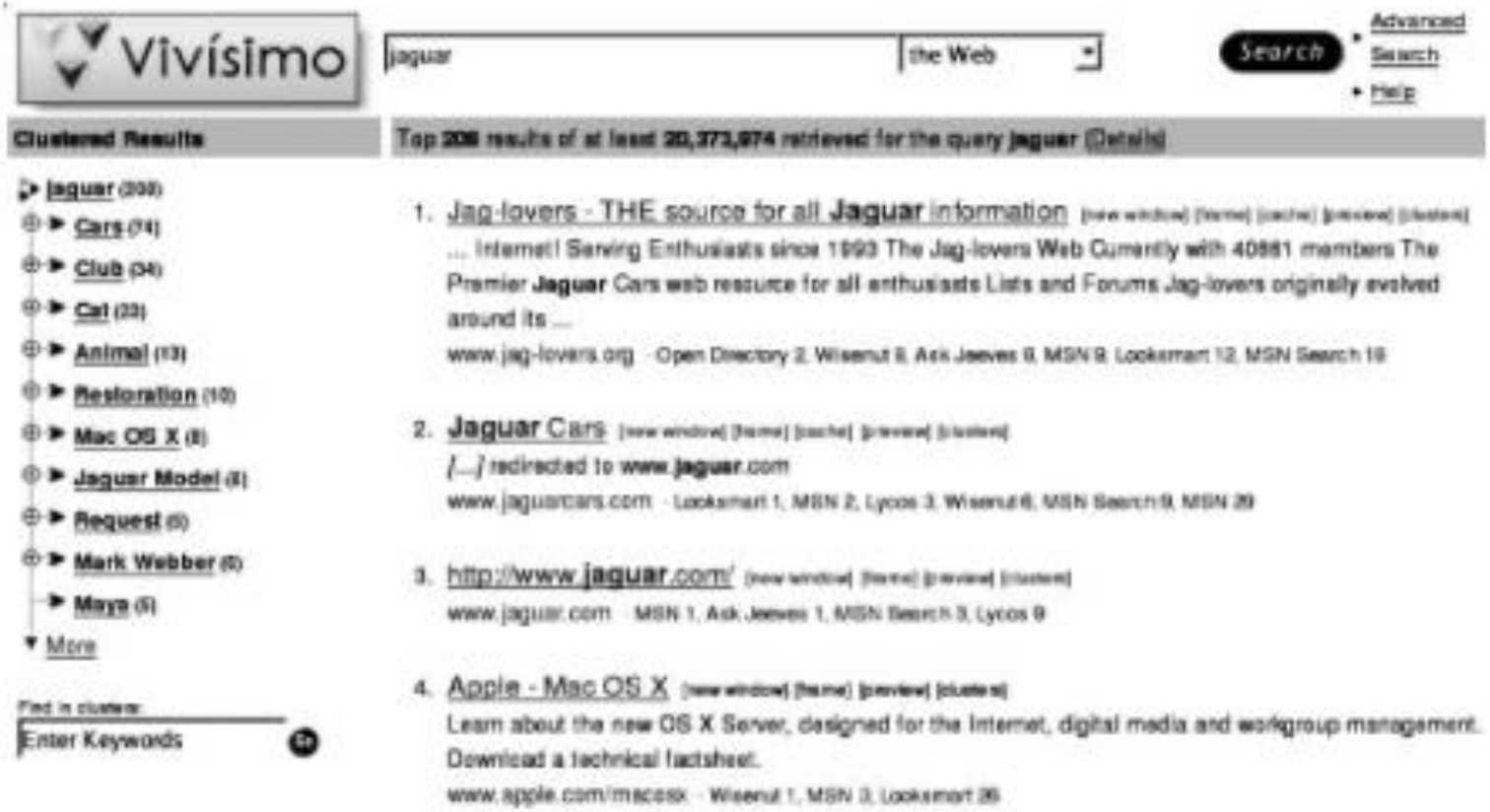
Outline

- Introduction
- Phases
  - Preprocessing
  - Clustering
  - Labeling
  - Performance Measures
- Future Work
- Sample Run
- References

Introduction

Aim is to present retrieved documents in clustered & labeled manner to enable:

- Better topic understanding
- Ease of access


Vivísimo
jaguar
the Web
Search
Advanced Search
Help
Clustered Results
Top 208 results of at least 20,373,974 retrieved for the query jaguar (Details)
jaguar (208)
Cars (74)
Club (34)
Cat (23)
Animal (13)
Restoration (10)
Mac OS X (8)
Jaguar Model (8)
Request (6)
Mark Webber (6)
Maya (5)
More
Find in cluster:
Enter Keywords
1. Jag-lovers - THE source for all Jaguar information
... Internet Serving Enthusiasts since 1993 The Jag-lovers Web Currently with 40881 members The Premier Jaguar Cars web resource for all enthusiasts Lists and Forums Jag-lovers originally evolved around its ...
www.jag-lovers.org - Open Directory 2, Wisenut 8, Ask Jeeves 8, MSN 9, Looksmart 12, MSN Search 19
2. Jaguar Cars
[...] redirected to www.jaguar.com
www.jaguarcars.com - Looksmart 1, MSN 2, Lycos 3, Wisenut 6, MSN Search 9, MSN 29
3. http://www.jaguar.com/
www.jaguar.com - MSN 1, Ask Jeeves 1, MSN Search 3, Lycos 9
4. Apple - Mac OS X
Learn about the new OS X Server, designed for the Internet, digital media and workgroup management. Download a technical factsheet.
www.apple.com/macosx - Wisenut 1, MSN 3, Looksmart 26

Phases

- Preprocessing
- Clustering
- Labeling
- Performance Measures
  - Clustering evaluation
  - Labeling evaluation

Preprocessing

- Database selection & preparation
- Document/Term matrix creation
  - Using Porter Stemmer
  - Binary or weighted?

year > year
make > make
making > make
coward > coward
cowardly > cowardli
note > note
noted > note
life > life
live > live
like > like
likely > like
unlikely > unlik
propose > propose
proposes > proposes
cat > cat
cats > cat

Clustering

- Cover Coefficient-based Clustering
  - Number of clusters
  - Initial cluster centers
- k-means Clustering
  - Using term vectors to calculate cluster means
  - Stopping if no centroid changes

Labeling

- Using document titles to label clusters [2]
- Labeling via term weighting [1]

Performance Measures

- Clustering evaluation
  - F-measure (that uses recall & precision)
- Labeling evaluation
  - Human evaluation (ground truth)
  - sim-F measure [1]

Future Work

- Remaining phases
  - Labeling
  - Performance evaluation
- Modifications, changes & additions

Sample Run

Number of documents: 51

Number of terms: 2304

Number of clusters: 21

Number of iterations in k-means: 168

Execution time: ~2 seconds

Cluster 1: [world news in brief: brussels rioting]

Cluster 2: [world news in brief: cocaine ring broken, world news in brief: newspaper pays up]

Cluster 3: [world news in brief: malawi frees poet]

Cluster 4: [observer: nobbled]

Cluster 5: [world news in brief: population warning]

Cluster 6: [world news in brief: booby-trap murder]

Cluster 7: [london stock exchange: equity futures and options trading, london stock exchange: ft-se 2,500 lost in nervous trading, world stock markets (america): bond recovery fails to bring joy to equities, government bonds: gilts follow treasuries' fall in auction]

Cluster 8: [observer: free wheeling, the lex column: british airways, leading article: a compromise path to emu]

Cluster 9: [us economic hopes rise as output level stabilises, foreign exchanges: dollar and yen lose ground, commodities and agriculture: lme monitors copper market as squeeze tightens, commodities and agriculture: fuel cell use is predicted for platinum, international company news: fletcher challenge shares placed, uk consumer credit growth at six-year low, the lex column: uk credit]

Cluster 10: [international company news: sa brewing expands in us water heating, bush pledge on chemical weapons]

Cluster 11: [international company news: usf&g to bolster capital by dollars 300m stock issues, international capital markets: citicorp in asset-backed launch worth dollars 1.4bn, severn trent to pay pounds 212m for bet unit, drexel details settlement plan, letter: government cover for private export insurance]

Cluster 12: [international company news: labatt to reduce its food processing operations, commodities and agriculture: farmers' viewpoint - machinery, the lex column: severn trent places a bet on waste]

Cluster 13: [uk company news: geevor merger hits rocks over pre-conditions, jubilee of a jet that did what it was designed to do, commodities and agriculture: total in algerian gas field, international company news: temporary rescue for french, hurd opposes wider role for ec on labour issues, leading article: a charter for consumers]

Cluster 14: [international company news: contigas plans dm900m east german project, international company news: apple computer upgrades]

Cluster 15: [the lex column: stock lending]

Cluster 16: [brent walker shares drop ahead of loss, uk company news: business spread behind power corp's advance, tvs unveils plan to raise equity as it falls into red]

Cluster 17: [world news in brief: khmer rouge snub]

Cluster 18: [the lex column: uk markets]

Cluster 19: [world news in brief: soviet deadlock]

Cluster 20: [international company news: toyoda machinery sales surge, international company news: nordstrom boosts earnings, green profits]

Cluster 21: [international company news: two venezuelan airlines register losses, international company news: aerolneas argentinas sale enters final phase]

References

- [1] A. Türel and F. Can. A new approach to search result clustering and labeling. AIRS'11 Proceedings of the 7th Asia conference on Information Retrieval Technology, 1(1):283–292, December 2011.
- [2] M. Alam and K. Sadaf. Labeling of Web Search Result Clusters using Heuristic Search and Frequent Itemset. ICICT 2014, 1(1):216-222, April 2014