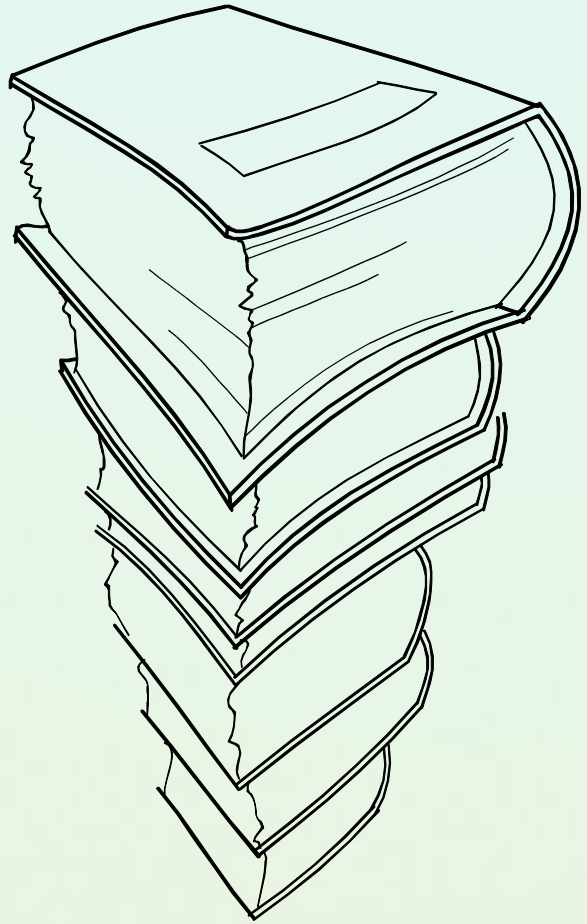
Search Results Clustering in Polish: Evaluation of Carrot

DAWID WEISS
JERZY STEFANOWSKI

Institute of Computing Science
Poznań University of Technology

Introduction

- search engines – tools of everyday use
- poor knowledge about search techniques
- presentation of search results
  - „Baudelaire?”



Limitations of ranked list presentation

What is Search Results Clustering?

Search Results Clustering is about efficient identification of meaningful thematic groups of documents in a search result and their concise presentation

- benefits gained from SRC
  - faster identification of relevant groups of documents
  - identification of topics range covered by the search result
- SRC does not cure
  - SRC is not a query answering system

Our research

- general influence of data pre-processing on the quality of clustering
  - ignoring stop-words
  - stemming
- clustering inflectionally rich languages (Polish)
- Suffix Tree Clustering algorithm’s thresholds and quality of results
- new search results clustering algorithms

Suffix Tree Clustering algorithm

- Snippet similarity based on recurring phrases
- utilizes suffix trees for clustering (theoretically linear complexity)
- one of the first approaches dedicated to search results clustering

All the real knowledge which we possess, depends on methods by which we distinguish the similar from the dissimilar.

- Genera plantarum, Linnaeus

Example

(1) “cat ate cheese”

(2) “mouse ate cheese too”

(3) “cat ate mouse too”

Base clusters:

[a] (1,3) cat ate

[b] (1,2,3) ate

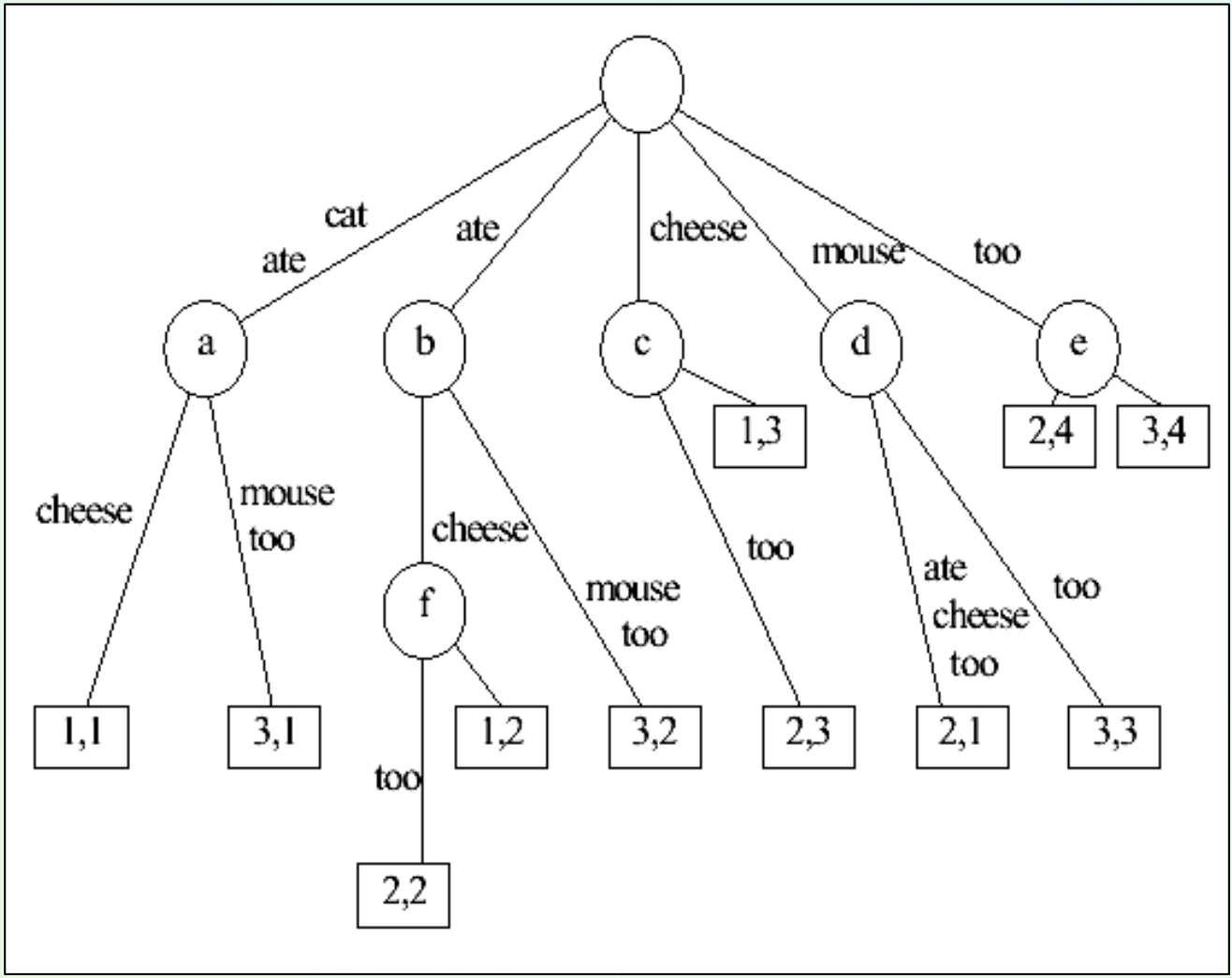
[f] (1,2) ate cheese

[c] (2,3) too

…

- some base clusters will be removed because they contain stop words, *np.* [c]

- for each cluster we calculate a *base cluster score*



$$s(m) = |m| \cdot f(|m_p|) \cdot \Sigma\ \mathrm{tfidf}(w_i)$$

Example (contd)

- base clusters merging

$$\text{sim}(m_i, m_j) = 1 \quad \text{if} \quad |m_i \cap m_j| / |m_i| > \alpha \text{ and } |m_i \cap m_j| / |m_j| > \alpha$$

$$\text{sim}(m_i, m_j) = 0 \quad \text{otherwise}$$

- binary similarity measure
- all connected sub graphs become clusters
- many limitations of the merging method

Data pre-processing (in STC and not only)

- ignoring frequently occurring terms (stop words)
- stemming

- how we addressed the above for Polish?
  - stop words – public sources and private word frequency list (Rzeczpospolita)
  - SAM
  - custom stemming and lemmatization methods: quasi-stemmer i lametyzator

Quasi-stemmer

- very simple
- head-word (lexeme) is not explicit
  - the terms share identical prefix (*k* characters)
  - after removing the prefix, the remainders for both terms exists in the lookup table of allowed suffixes
- suffixes table from Rzeczpospolita corpus
- weaknesses of the method
  - does not handle alternations
  - relation of ‘stem’ equality not transitive

[Lame]tyzator

- inflected and base forms generated using *ispell-pl* dictionary
- compressed to a finite state automaton
- advantages
  - very fast
  - large word coverage (1.4 million? src: *ispell-pl*)
  - open source (dictionary: GPL, Java code: free)
- weaknesses
  - only words in the dictionary can be analyzed
  - contains erroneous entries (betoniarka [beton])
  - no tags (stemming only)

The experiment: measuring clustering quality

- existing approaches
  - precision/ recall – lack of test data
  - user surveys – subjective, hard to involve large number of participants
  - user interface efficiency measures (Zamir)

The experiment: measuring clustering quality

- Byrona E. Dom measure of clustering quality
  - entropy-based
  - measures differences between the ‘ideal’ and given clustering
  - $Q2=1$ → *C* i *K* are identical
  - $Q2=0$ → groups in *K* do not carry any information about groups in *C*

The experiment: assumptions

- clustering of 1:1 type (partitioning)
- binary document-to-cluster membership
- flat structure of clusters (no hierarchy)

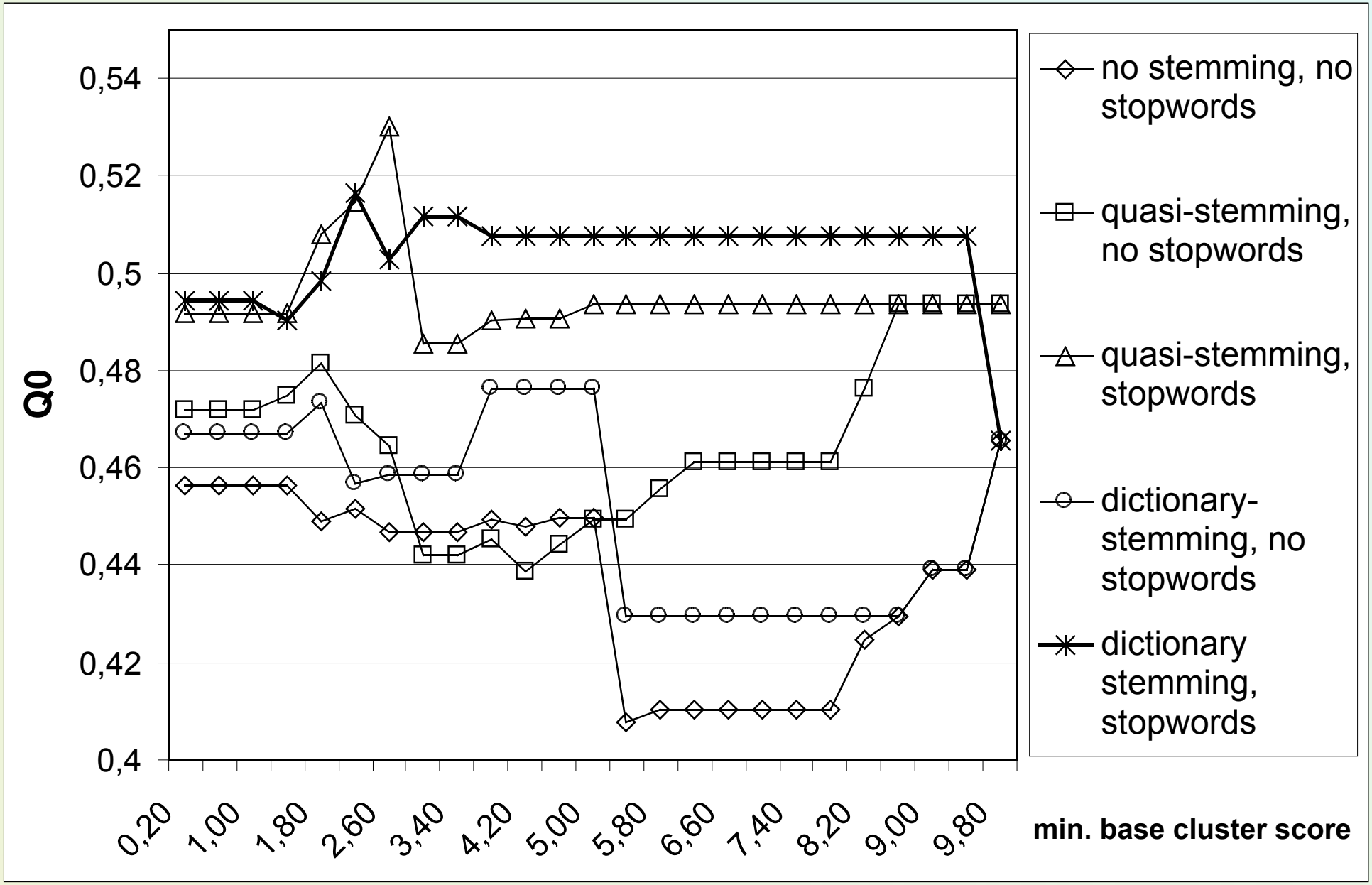
The experiment: input data and ground truth

- A set of 100 results for two queries (*inteligencja* and *odkrywanie wiedzy*) were downloaded
- Manual clustering of this set was performed by 5 individuals (experts)
- Ground truth set was obtained by unifying the results from each expert
- A large number of inconsistencies in manual clustering only proves the problem is indeed difficult (only about 50% of assignments fully consistent among all experts)
- Experiment has been later extended to cover more queries (2 in Polish and 4 in English)

The experiment: configurations

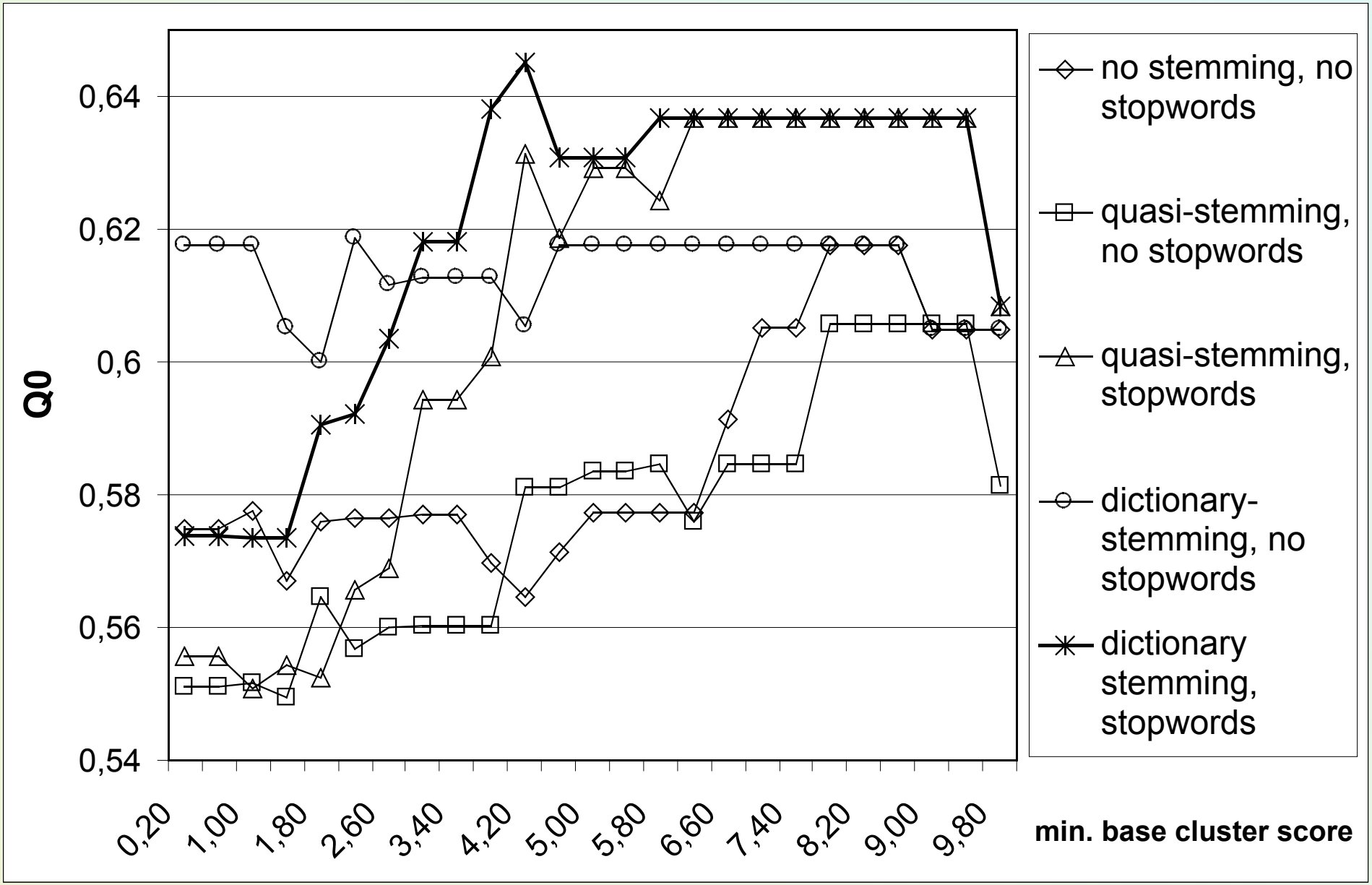
- pre-processing configurations
  - for Polish:
    - no stemming, all words
    - quasi-stemmer, all words
    - quasi-stemmer, stop words ignored
    - lametyzator, all words
    - lametyzator, stop words ignored
  - for English:
    - as above, Porter algorithm used for stemming
- wide spectrum of values for control thresholds (*minimum base cluster score* and *merge threshold)*

Results



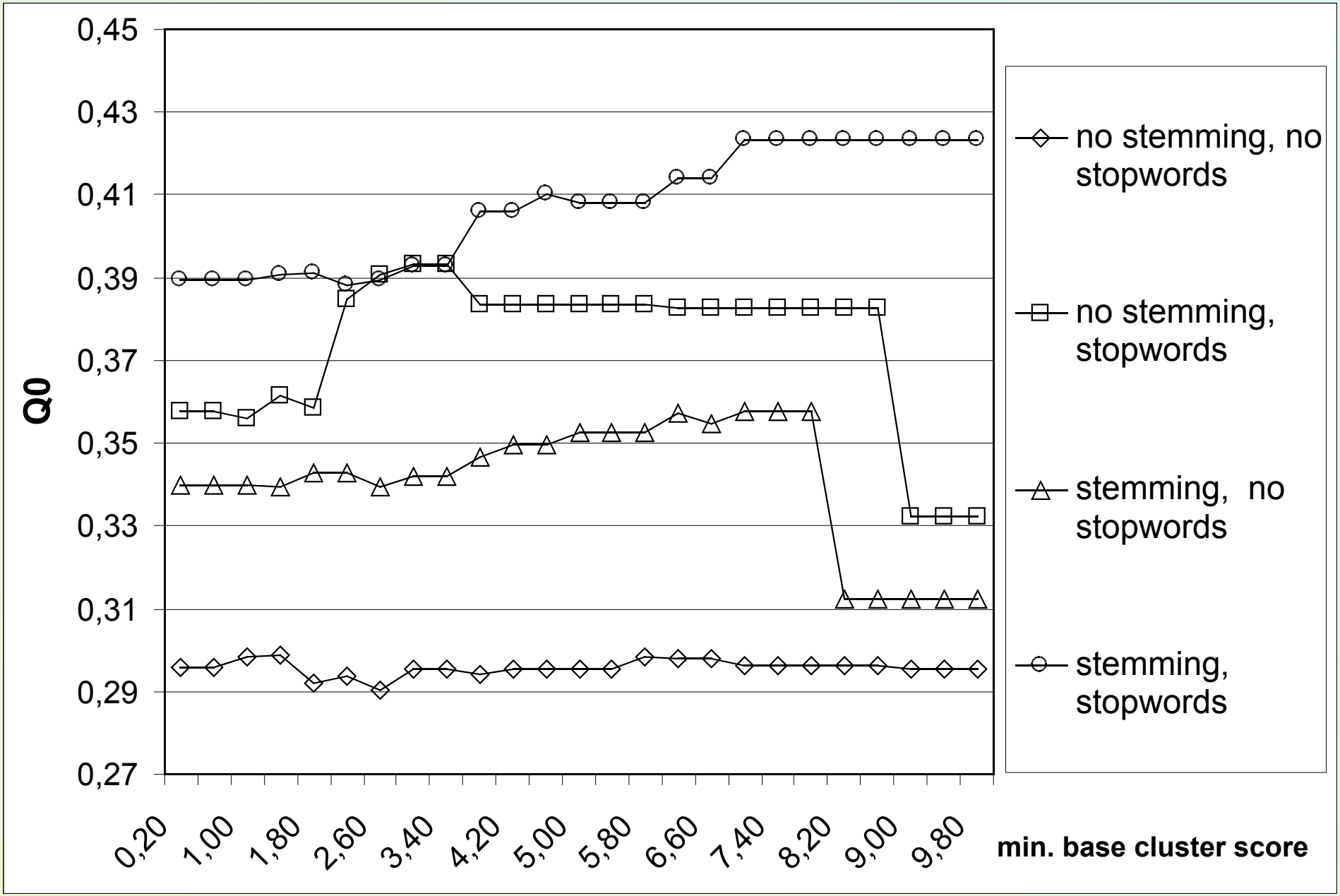
Distribution of Q0, constant merge threshold (0.6), query: inteligencja

Results (contd)



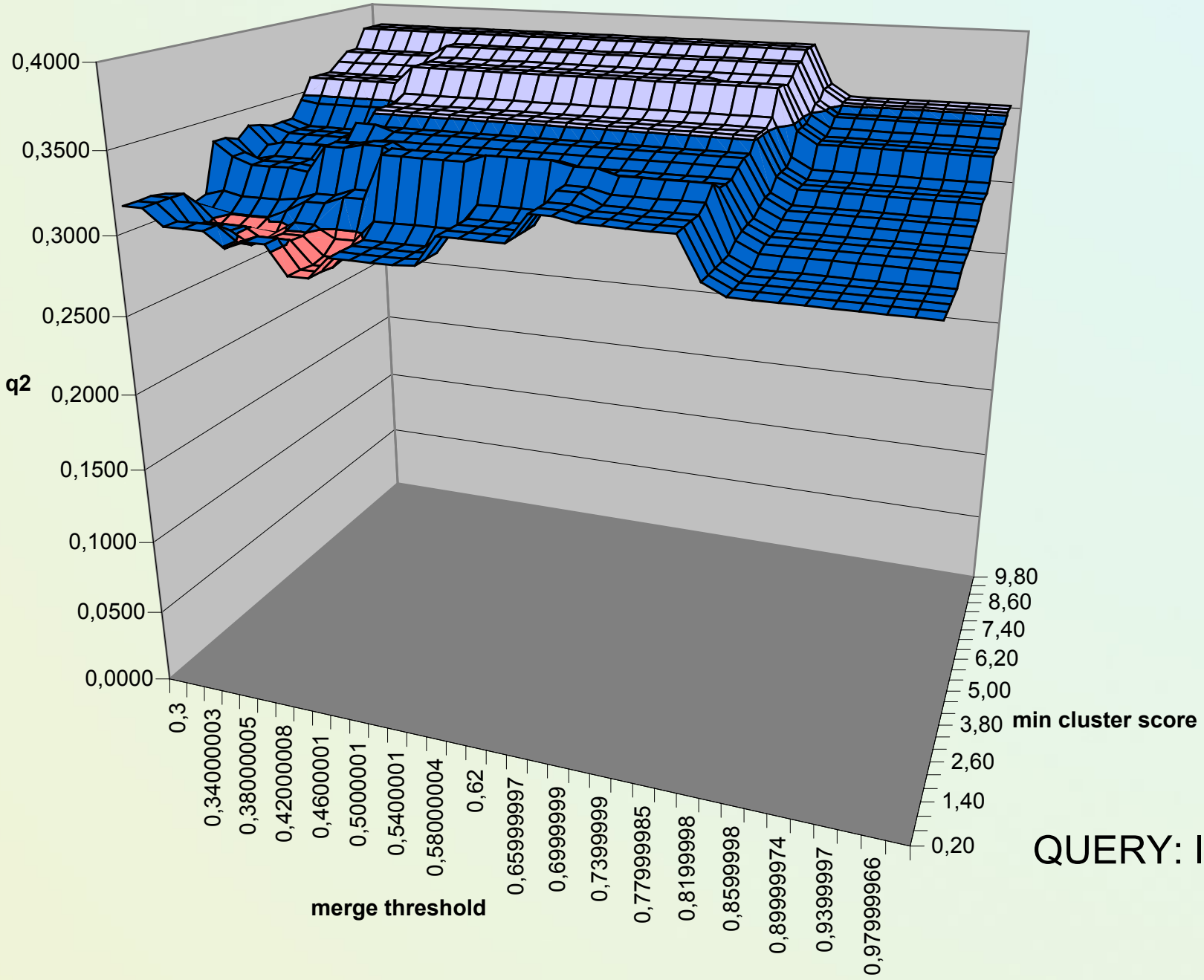
Distribution of Q0, constant merge threshold (0.6), query: odkrywanie wiedzy

Results (contd)



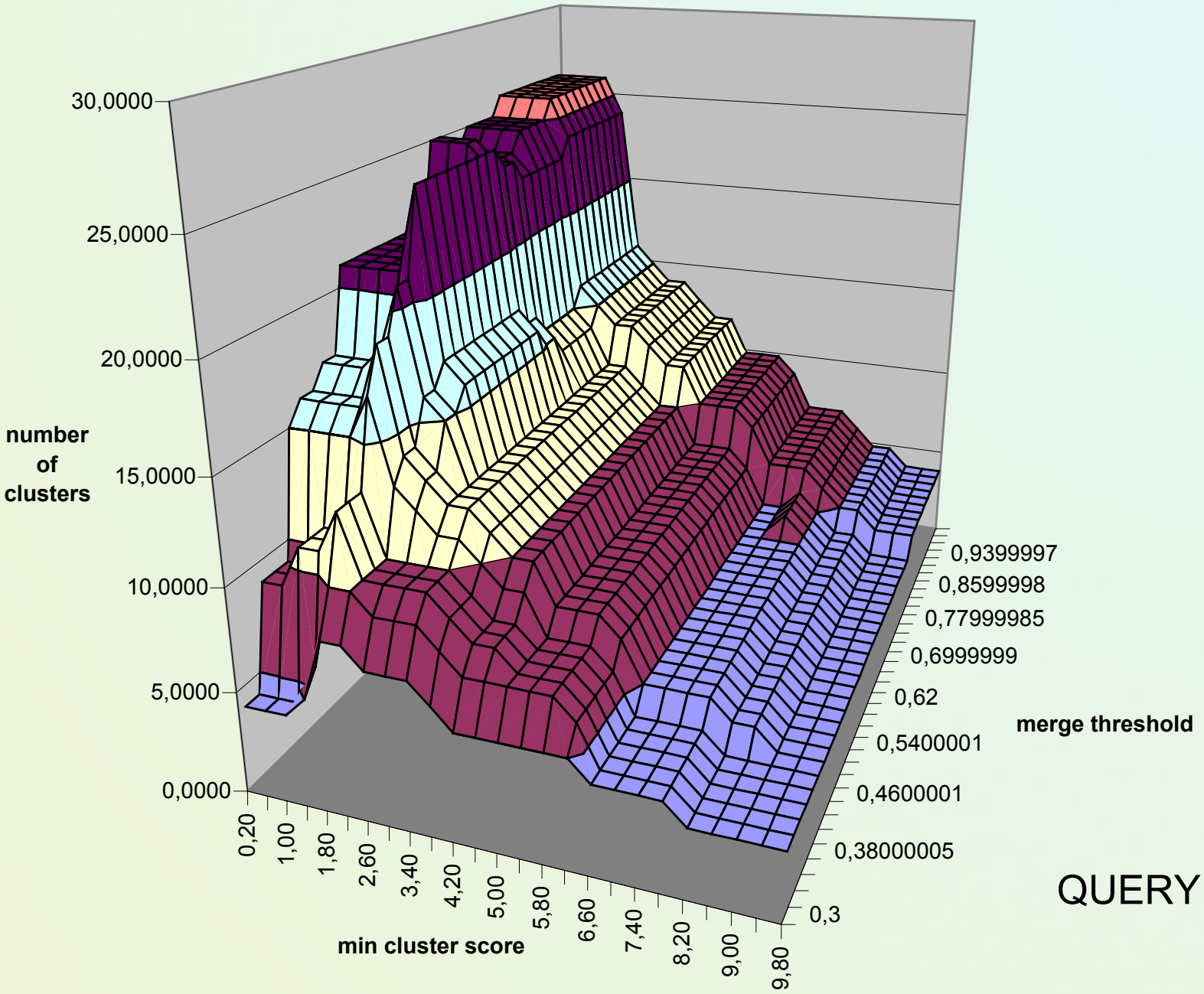
Distribution of Q0, constant merge threshold (0.6), query: salsa

Results – thresholds and quality



QUERY: logika rozmyta

Results – thresholds and clusters number



QUERY: logika rozmyta

Conclusions (general)

- STC seems to be sensitive to languages with rich inflection
- stemming and ignoring stop words improved the quality of results (within our assumptions and quality measure)
- even simple pre-processing methods yielded significant improvement (quasi-stemmer)

Conclusions (STC-specific)

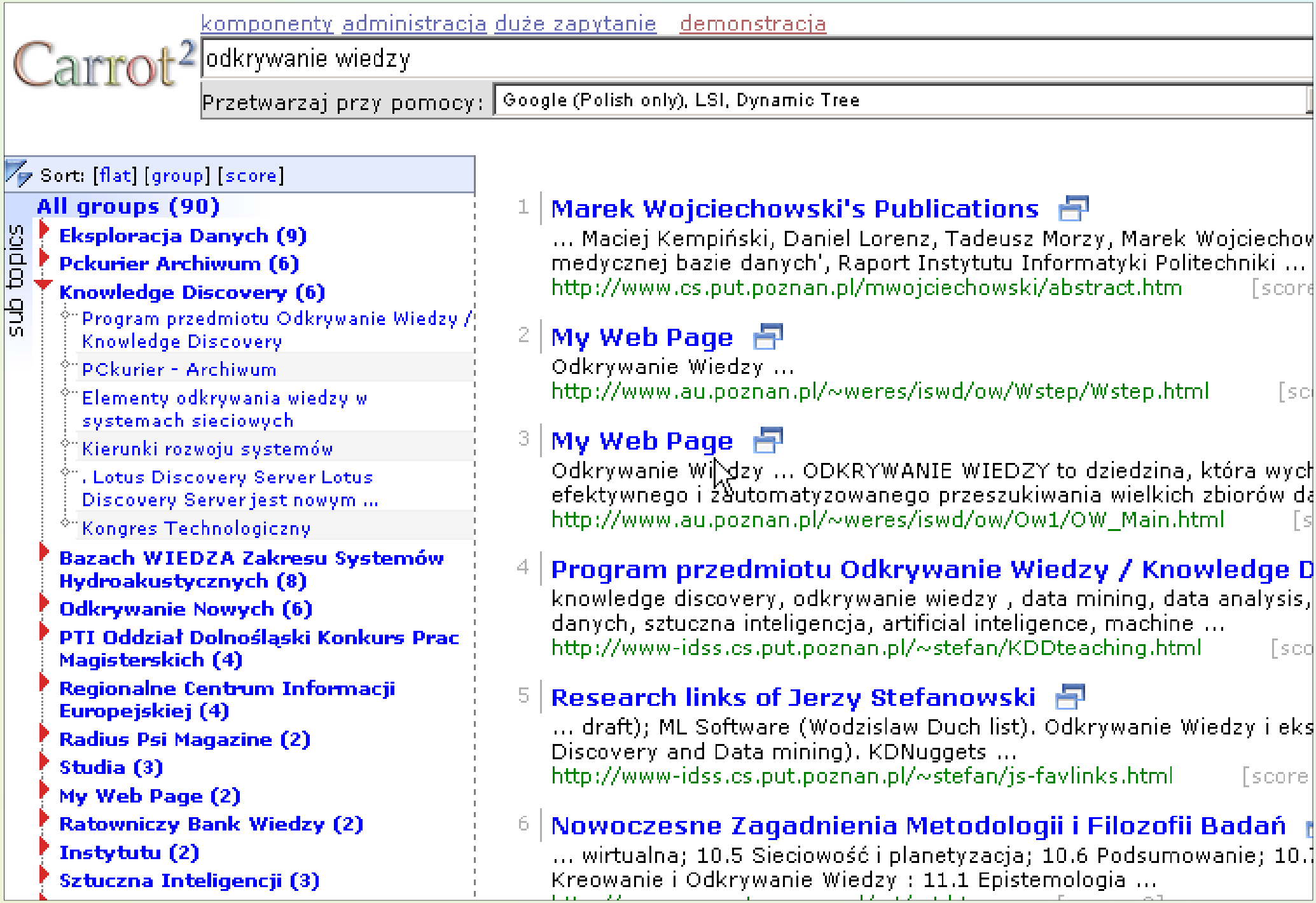
- low base cluster score and merge threshold decrase the stability of quality measure
- base cluster score strongly affects the number of final clusters
- high base cluster score leads to highly distinctive, but potentially obvious, clusters

Current work

- other algorithms (not phrase-based)
  - derived from Latent Semantic Indexing
  - hierarchical methods
- search results clustering framework – Carrot^2

Carrot^2

- in the beginning…
  - reference STC implementation
- now
  - many algorithms
  - distributed architecture
  - data-driven components (XML)
  - ease of debugging and component integration
  - active open source project


komponenty administracja duże zapytanie demonstracja
Carrot2
odkrywanie wiedzy
Przetwarzaj przy pomocy: Google (Polish only), LSI, Dynamic Tree
Sort: [flat] [group] [score]
sub topics
All groups (90)
Eksploracja Danych (9)
Pckurier Archiwum (6)
Knowledge Discovery (6)
Program przedmiotu Odkrywanie Wiedzy / Knowledge Discovery
PCkurier - Archiwum
Elementy odkrywania wiedzy w systemach sieciowych
Kierunki rozwoju systemów
. Lotus Discovery Server Lotus Discovery Server jest nowym ...
Kongres Technologiczny
Bazach WIEDZA Zakresu Systemów Hydroakustycznych (8)
Odkrywanie Nowych (6)
PTI Oddział Dolnośląski Konkurs Prac Magisterskich (4)
Regionalne Centrum Informacji Europejskiej (4)
Radius Psi Magazine (2)
Studia (3)
My Web Page (2)
Ratowniczy Bank Wiedzy (2)
Instytutu (2)
Sztuczna Inteligencji (3)
1 Marek Wojciechowski's Publications
... Maciej Kempiński, Daniel Lorenz, Tadeusz Morzy, Marek Wojciechow
medycznej bazie danych', Raport Instytutu Informatyki Politechniki ...
http://www.cs.put.poznan.pl/mwojciechowski/abstract.htm
2 My Web Page
Odkrywanie Wiedzy ...
http://www.au.poznan.pl/~weres/iswd/ow/Wstep/Wstep.html
3 My Web Page
Odkrywanie Wiedzy ... ODKRYWANIE WIEDZY to dziedzina, która wych
efektywnego i zautomatyzowanego przeszukiwania wielkich zbiorów da
http://www.au.poznan.pl/~weres/iswd/ow/Ow1/OW_Main.html
4 Program przedmiotu Odkrywanie Wiedzy / Knowledge D
knowledge discovery, odkrywanie wiedzy , data mining, data analysis,
danych, sztuczna inteligencja, artificial inteligence, machine ...
http://www-idss.cs.put.poznan.pl/~stefan/KDDteaching.html
5 Research links of Jerzy Stefanowski
... draft); ML Software (Wodzislaw Duch list). Odkrywanie Wiedzy i eks
Discovery and Data mining). KDNuggets ...
http://www-idss.cs.put.poznan.pl/~stefan/js-favlinks.html
6 Nowoczesne Zagadnienia Metodologii i Filozofii Badań
... wirtualna; 10.5 Sieciowość i planetyzacja; 10.6 Podsumowanie; 10.
Kreowanie i Odkrywanie Wiedzy : 11.1 Epistemologia ...

Become part of the project

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