

---

# Search Results Clustering in Polish: Evaluation of Carrot

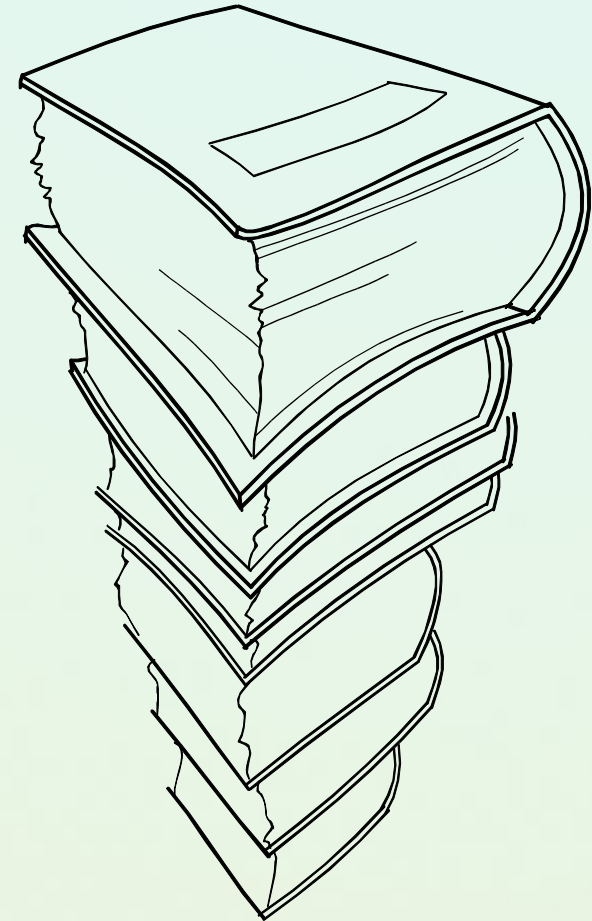
DAWID WEISS  
JERZY STEFANOWSKI

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



The screenshot shows the alltheweb search engine interface. At the top left is the logo 'alltheweb' with the tagline 'find it all'. To the right are links for 'advanced search', 'customize preferences', 'submit site', and 'help'. A search box contains the text 'salsa' and a 'SEARCH' button. Below the search box, it says 'Results in: Any Language' (selected) and 'Polish and English'. A navigation bar includes 'Web', 'News', 'Pictures', 'Video', 'Audio', and 'FTP files'. A blue banner indicates '1 - 10 of 3,705,187 Results for salsa'. The results list includes news items, a product index for 'SALSACYCLES2003', a magazine website 'Salsaweb.com', a UK website 'salsa.co.uk', and a dance website 'PLANET SALSA'. Each result includes a description and a file size.

alltheweb  
• • • find it all • • •

advanced search :: customize preferences :: submit site :: help

salsa SEARCH

Results in:  Any Language  Polish and English

Web News Pictures Video Audio FTP files

1 - 10 of 3,705,187 Results for **salsa**

News : [Cato Salsa Experience ut av garasjen CATO SALSA EX...](#) - [Aftenposten](#) - Found: 18 hours ago  
[Salsa hver fredag TOM I. ANDERSEN 28.5.2003 04:00](#) - [Harstad Tidende](#) - Found: 4 hours ago

[SALSACYCLES2003](#)  
**SALSA** CYCLES | FRAMES | COMPONENTS | ACCESSORIES | APPAREL | CONTACT  
**Description:** Product index, **salsa** recipes, catalog requests, mission statement and contact information.  
<http://www.salsacycles.com/> - 8 KB

[Salsaweb.com The World's Largest Salsa Magazine, the World's Largest Latin...](#)  
... Entertainment Online Magazine, The World's Largest **Salsa** Dancing Magazine ... com The World's Largest Onli  
Videos Mailing List ...  
**Description:** On-line source for Latin dance. Advice, humour and information. Includes links to **salsa** sites and danc  
<http://www.salsaweb.com/> - 59 KB

~ [salsa.co.uk](#)  
... Love **Salsa** ... JOIN **SALSA.CO.UK** ... an email to [join@salsa.co.uk](mailto:join@salsa.co.uk) ...  
**Description:** **salsa salsa.co.uk**  
<http://www.salsa.co.uk/> - 28 KB

[PLANET SALSA, from the ORIGINS of CLAVE to MILLENNIUM MAMBO; Salsa-Mambo Da](#)  
**Salsa-Mambo** Dance, Latin dancing, **salsa** Music, **salsa** Dance Clubs, Live **salsa** Music Events and Music Rhyth  
Clave and Mambo Dance via University of **Salsa P L A N E T** ...  
**Description:** A fun and informative site loaded with **Salsa/Mambo** Dance and music news, dancer interviews, events  
University of **Salsa**.  
[more hits from: http://www.planetsalsa.com/](http://www.planetsalsa.com/) - 51 KB

# 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

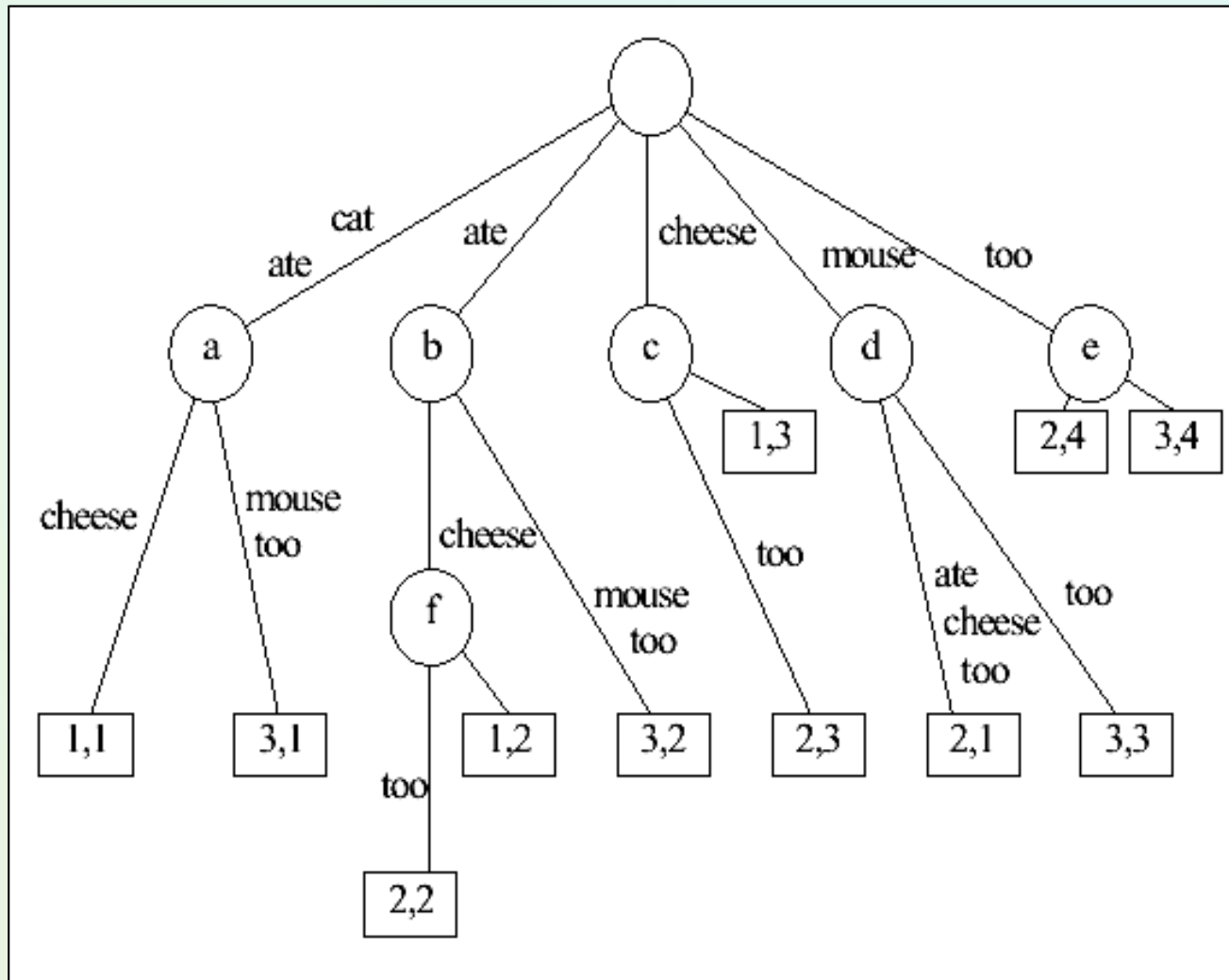
[f] (1,2) ate cheese

[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**



$$s(m) = |m| \cdot f(|m_p|) \cdot \sum \text{tfidf}(w_i)$$

# Example (contd)

---

- base clusters merging

$$\begin{aligned} \text{sim}(m_i, m_j) &= 1 && \text{if } |m_i \cap m_j| / |m_i| > \alpha \text{ and } |m_i \cap m_j| / |m_j| > \alpha \\ \text{sim}(m_i, m_j) &= 0 && \text{otherwise} \end{aligned}$$

- 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
  - $Q_2=1 \rightarrow C \text{ i } K \text{ are identical}$
  - $Q_2=0 \rightarrow \text{groups in } K \text{ do not carry any information about groups in } C$

# 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)

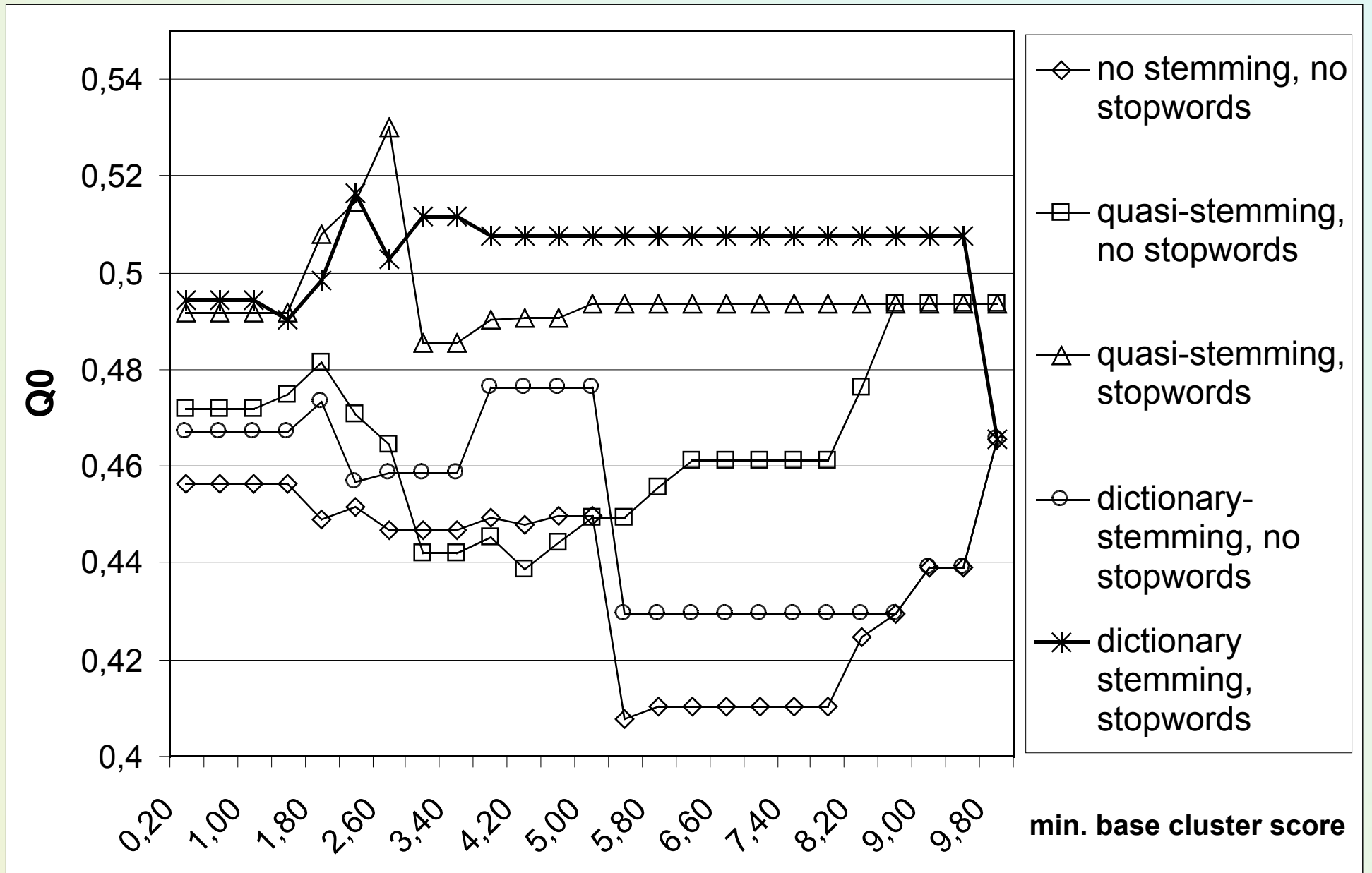
# The experiment: configurations

---

- 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*)

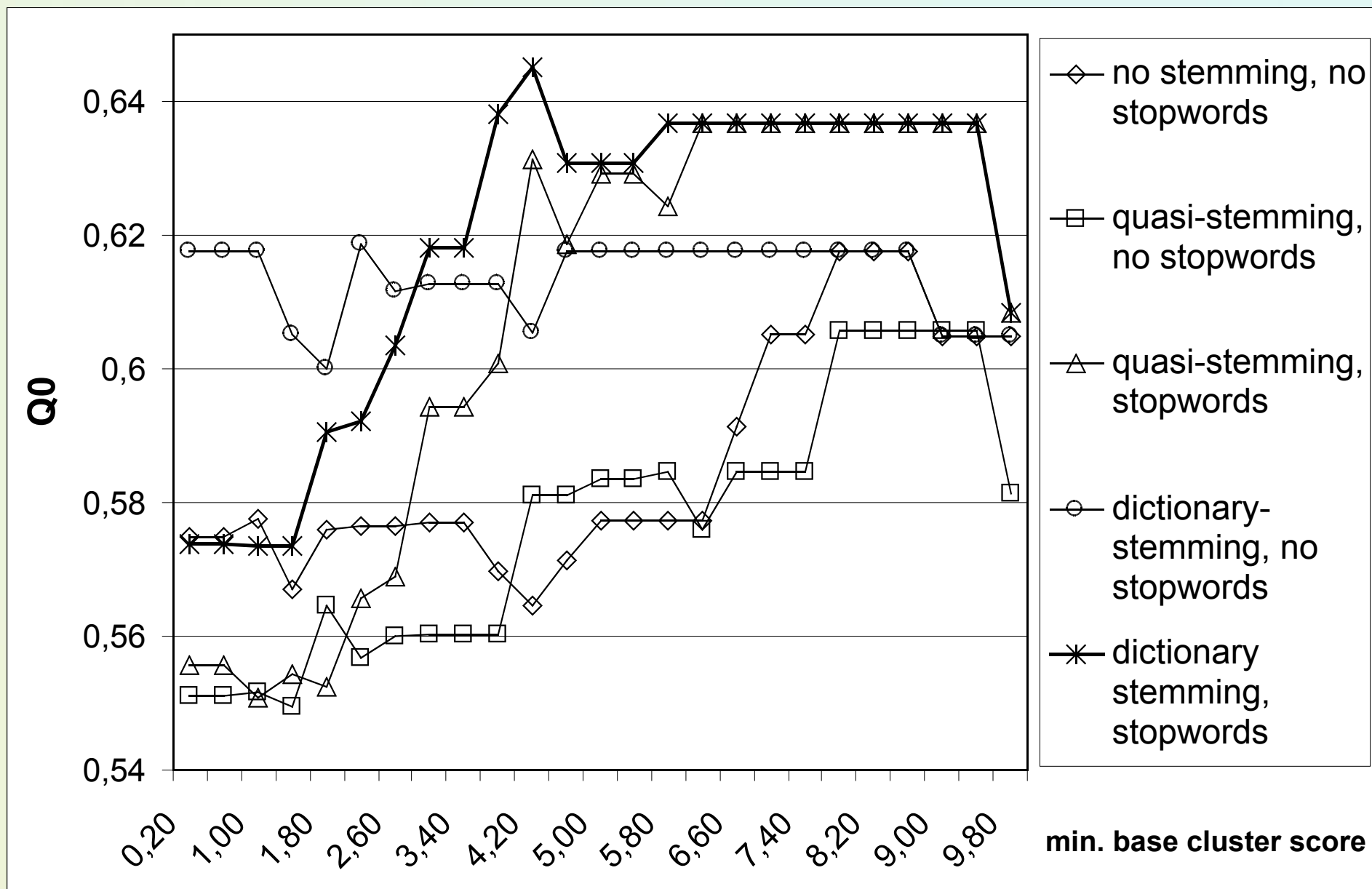


# Results



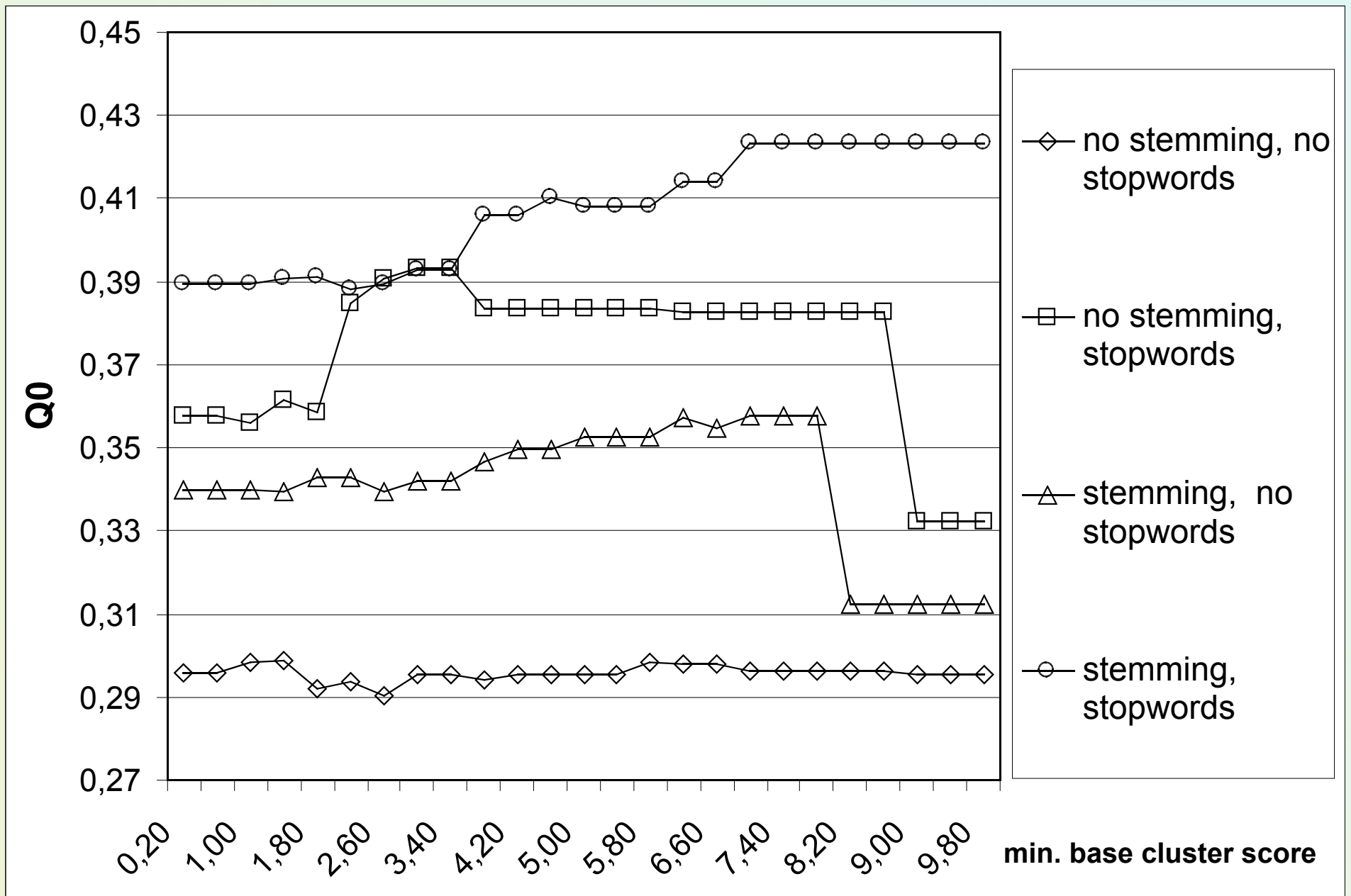
Distribution of Q0, constant merge threshold (0.6), query: inteligencja

# Results (contd)



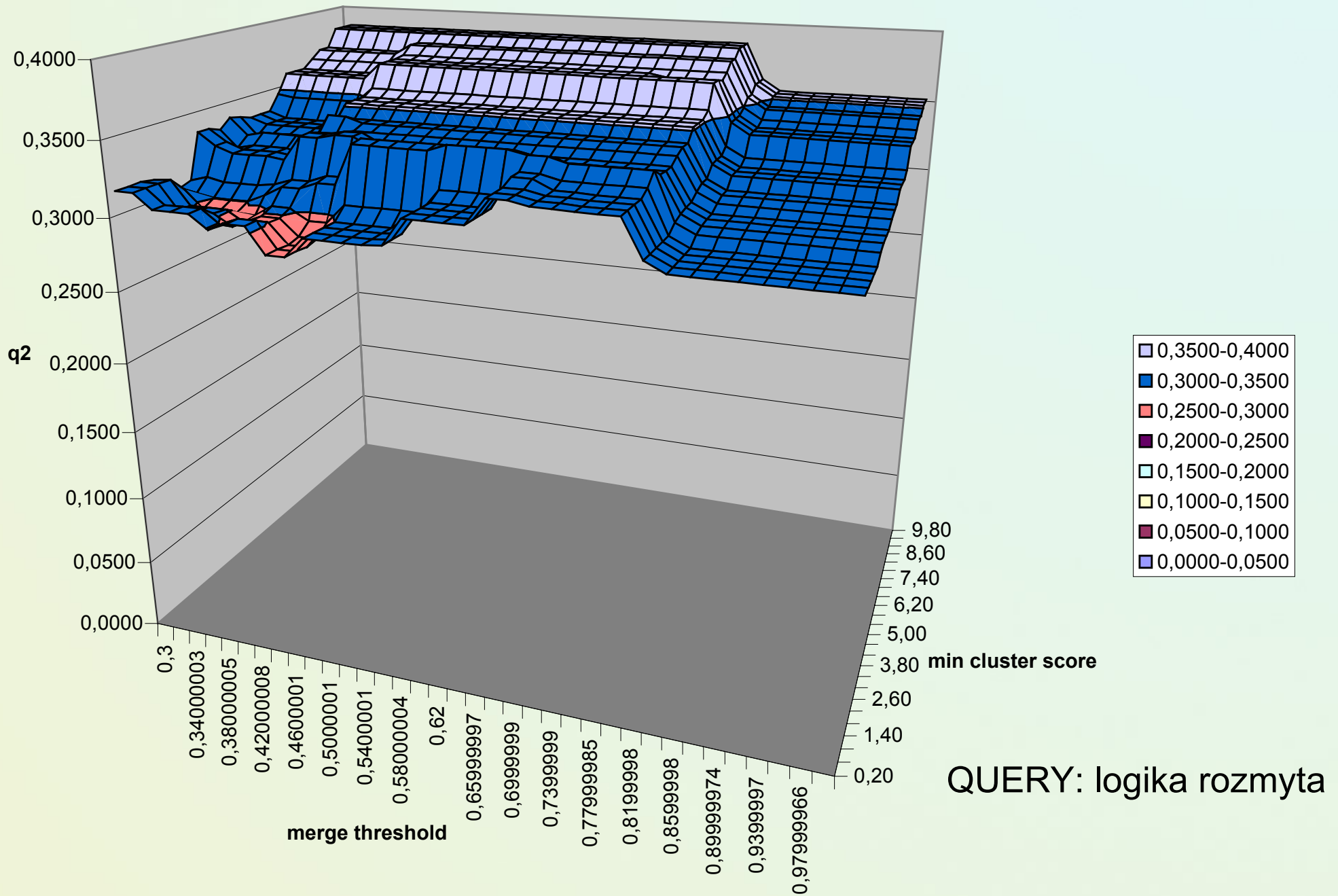
Distribution of Q0, constant merge threshold (0.6), query: odkrywanie wiedzy

# Results (contd)

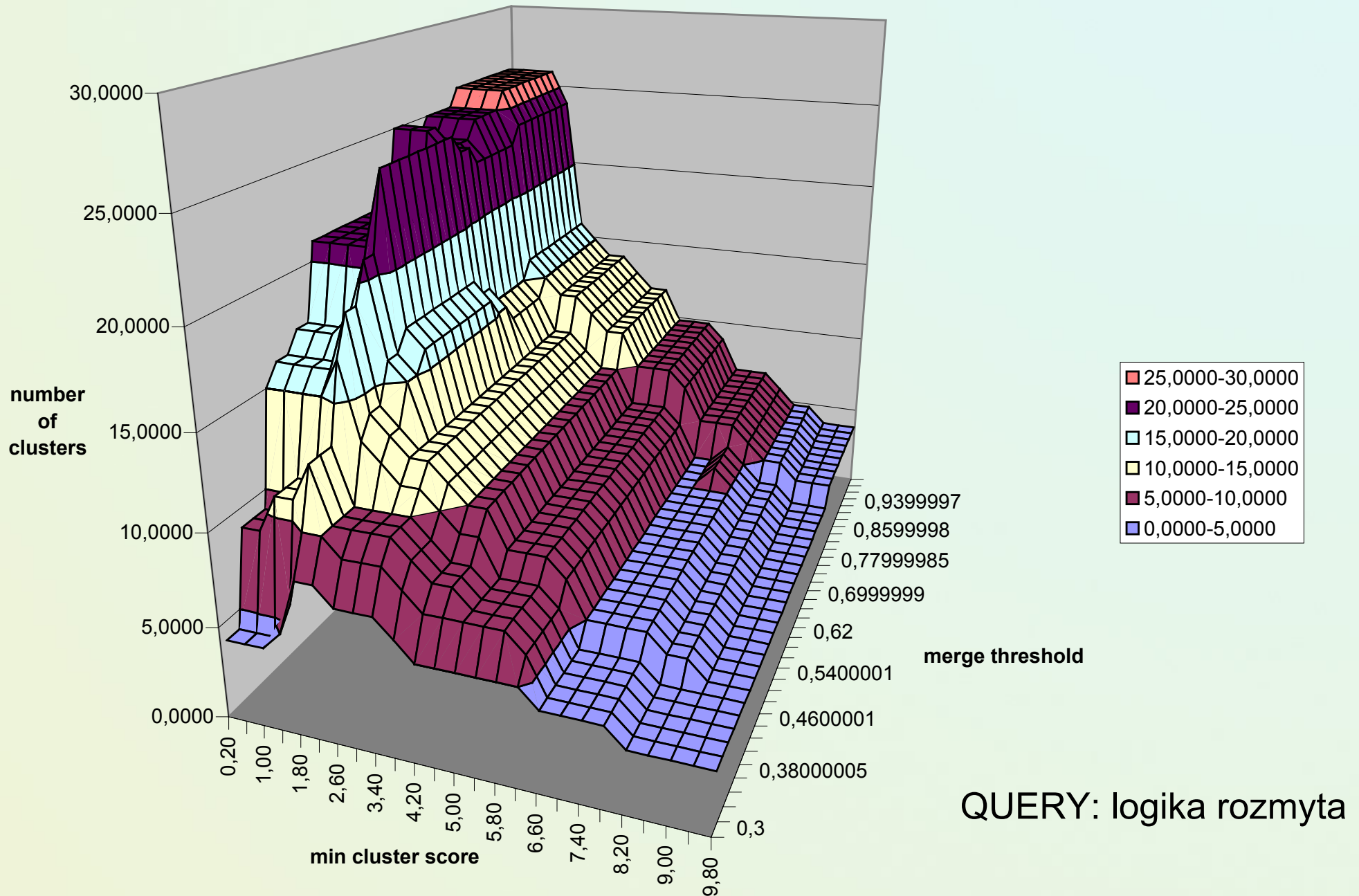


Distribution of Q0, constant merge threshold (0.6), query: salsa

# Results – thresholds and quality



# Results – thresholds and clusters number



# 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 decrease 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<sup>2</sup>



# Carrot<sup>2</sup>

---

- 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

Sort: [flat] [group] [score]

## All groups (90)

sub topics

## 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 Kempniński, Daniel Lorenz, Tadeusz Morzy, Marek Wojciechowski 'Odkrywanie wiedzy w medycznej bazie danych', Raport Instytutu Informatyki Politechniki ...  
<http://www.cs.put.poznan.pl/mwojciechowski/abstract.htm> [score]
- 2 | **My Web Page** 

Odkrywanie Wiedzy ...  
<http://www.au.poznan.pl/~weres/iswd/ow/Wstep/Wstep.html> [score]
- 3 | **My Web Page** 

Odkrywanie Wiedzy ... ODKRYWANIE WIEDZY to dziedzina, która wychodzi poza ramy tradycyjnego i zautomatyzowanego przeszukiwania wielkich zbiorów danych ...  
[http://www.au.poznan.pl/~weres/iswd/ow/Ow1/Ow\\_Main.html](http://www.au.poznan.pl/~weres/iswd/ow/Ow1/Ow_Main.html) [score]
- 4 | **Program przedmiotu Odkrywanie Wiedzy / Knowledge Discovery**

knowledge discovery, odkrywanie wiedzy , data mining, data analysis, data mining, sztuczna inteligencja, artificial intelligence, machine learning ...  
<http://www-idss.cs.put.poznan.pl/~stefan/KDDteaching.html> [score]
- 5 | **Research links of Jerzy Stefanowski** 

... draft); ML Software (Wodzislaw Duch list). Odkrywanie Wiedzy i eksploracja danych (Discovery and Data mining). KDNuggets ...  
<http://www-idss.cs.put.poznan.pl/~stefan/js-favlinks.html> [score]
- 6 | **Nowoczesne Zagadnienia Metodologii i Filozofii Badań** 

... wirtualna; 10.5 Sieciowość i planetyzacja; 10.6 Podsumowanie; 10.7 Kreowanie i Odkrywanie Wiedzy : 11.1 Epistemologia ...  
<http://www-idss.cs.put.poznan.pl/~stefan/nowoczesne-zagadnienia-metodologii-i-filozofii-badan.html> [score]

# Become part of the project

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

[komponenty](#) [administracja](#) [duże zapytanie](#) [demonstracja](#)

Carrot<sup>2</sup>

iipwm



Przetwarzaj przy pomocy:

Sort: [\[flat\]](#) [\[group\]](#) [\[score\]](#)

sub topics

## All groups (116)

- ▶ [Intelligent Information Systems IIS \(30\)](#)
- ▶ [Nd call for Papers \(14\)](#)
- ▶ [Linguist List show Conference \(14\)](#)
- ▶ [CFP \(10\)](#)
- ▶ [KAW List Contribution \(8\)](#)
- ▶ [Euromap Events \(8\)](#)
- ▶ [Forthcoming Meetings \(8\)](#)
- ▶ [Machine Learning \(4\)](#)
- ▶ [BRUJULA El Primer Buscador Argentino \(2\)](#)
- ▶ [Wojtek Jamroga Home Page \(4\)](#)
- ▶ [Brian EDAIM \(2\)](#)
- ▶ [Mieczysław Kłopotek \(2\)](#)
- ▶ [\(Other\) \(10\)](#)

- 1 | [New Trends in Intelligent Information Processing and](#)  
<http://iipwm.ipipan.waw.pl/> [score: 0.19]
- 2 | [Registration for IIS: IIPWM '2003](#)   
... IIS: IIPWM '2003, International Conference Intelligent Informat  
Mining Zakopane, 2-5 June 2003 Daglezja Hotel. REGISTRATION F  
<http://iipwm.ipipan.waw.pl/2003/iipwm2003latereg.html> [score]
- 3 | [IIPWM - Call for Papers](#)   
... IIPWM - Call for Papers. From: IIPWM Conf. Office; Subject: IIP  
Date: Sun, 18 Aug 2002 10:08:26 -0700. ... IIPWM - Call for Paper  
<http://www.mail-archive.com/inductive@listserv.unb.ca/msg00641>
- 4 | [Machine Learning List: Vol. 14, No. 6](#)   
... at Stanford University CFP: Genetic and Evolutionary Computati  
[2nd CFP] ES2002 workshop: AI for Intelligent Business IIPWM - C  
<http://www.mail-archive.com/ml@isle.org/msg00005.html> [score]
- 5 | [KAW List Contribution](#) 