



The Power of Clustering

***International Chemical Information Conference
Nimes, France***

**Dr. David A. Evans
Clairvoyance Corporation
October 20, 2003**



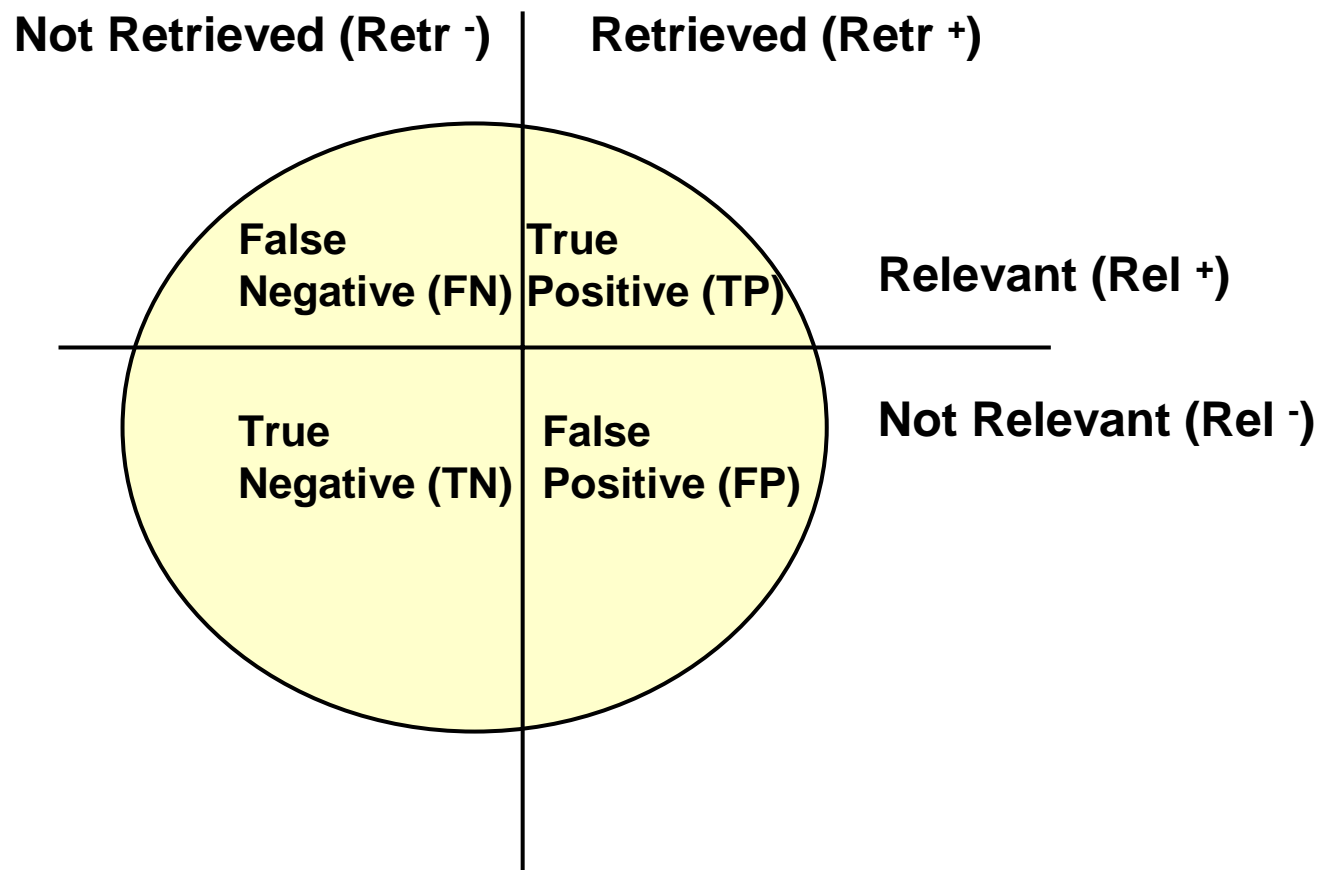
- **The Problem: Improving Retrieval**
 - What is the state of the art?
 - What contributes to “optimal” performance?
- **Pseudo-Relevance Feedback (PRF)**
 - Remarkably effective automatic technique
- **Clustering**
 - First significant improvement in PRF
 - As a means of revealing document relations
- **Conclusion**



State-of-the-Art IR



Evaluation Framework





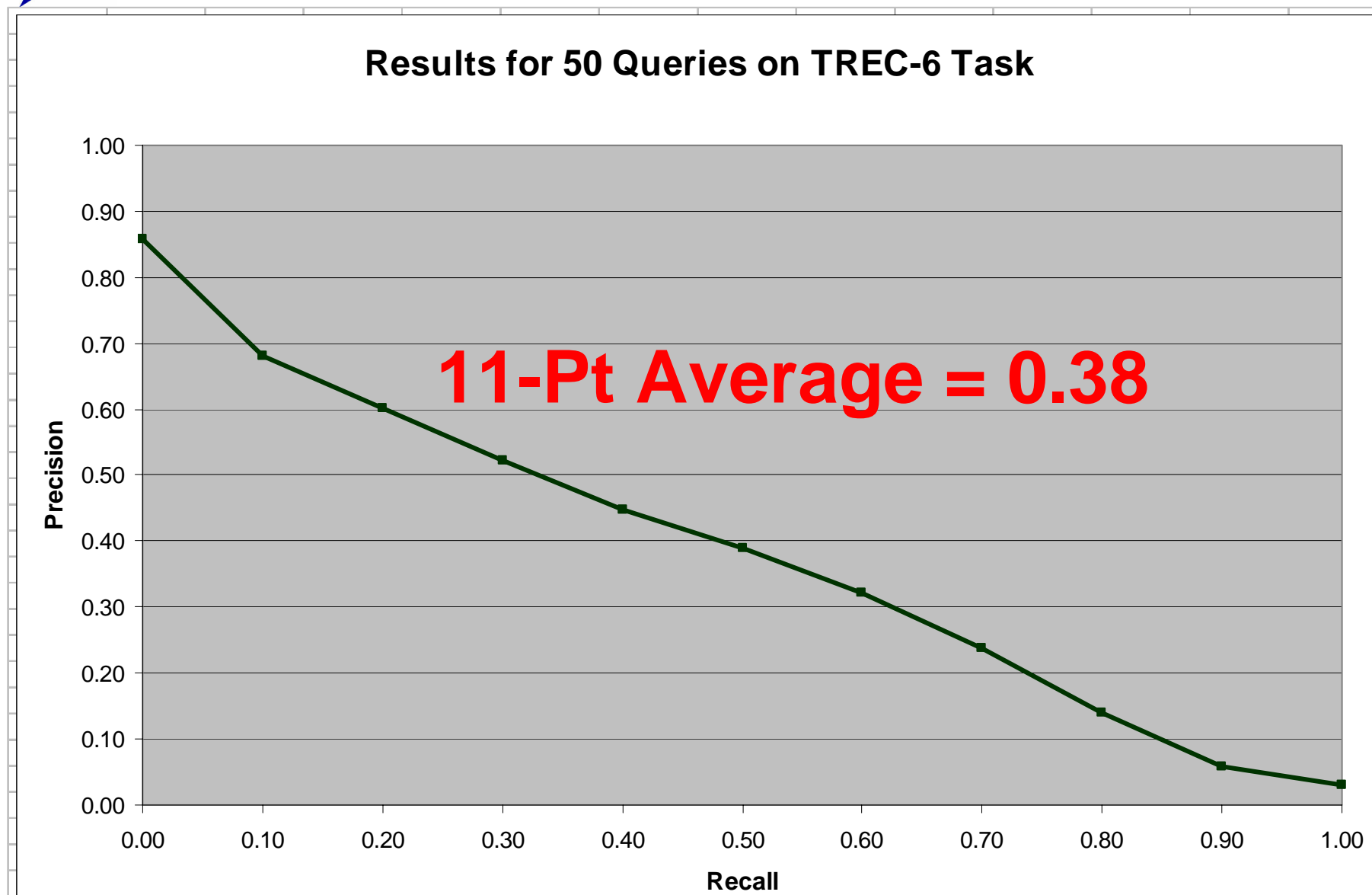
Evaluation Measures

| | Retrieved | Not Retrieved |
|--------------|-----------|---------------|
| Relevant | R^+ | R^- |
| Not Relevant | N^+ | N^- |

| Evaluation Measure | Definition |
|--------------------|--|
| Precision | $precision = \frac{R^+}{R^+ + N^+}$ |
| Recall | $recall = \frac{R^+}{R^+ + R^-}$ |
| T11U | $T11U = 2R^+ - N^+$ |
| T11SU | $T11SU = \frac{\max(T11NU, MinNU) - MinNU}{1 - MinNU}$ <p>where</p> $T11NU = \frac{T11U}{MaxU}$ $MinNU = -0.5$ |



P-R Curve





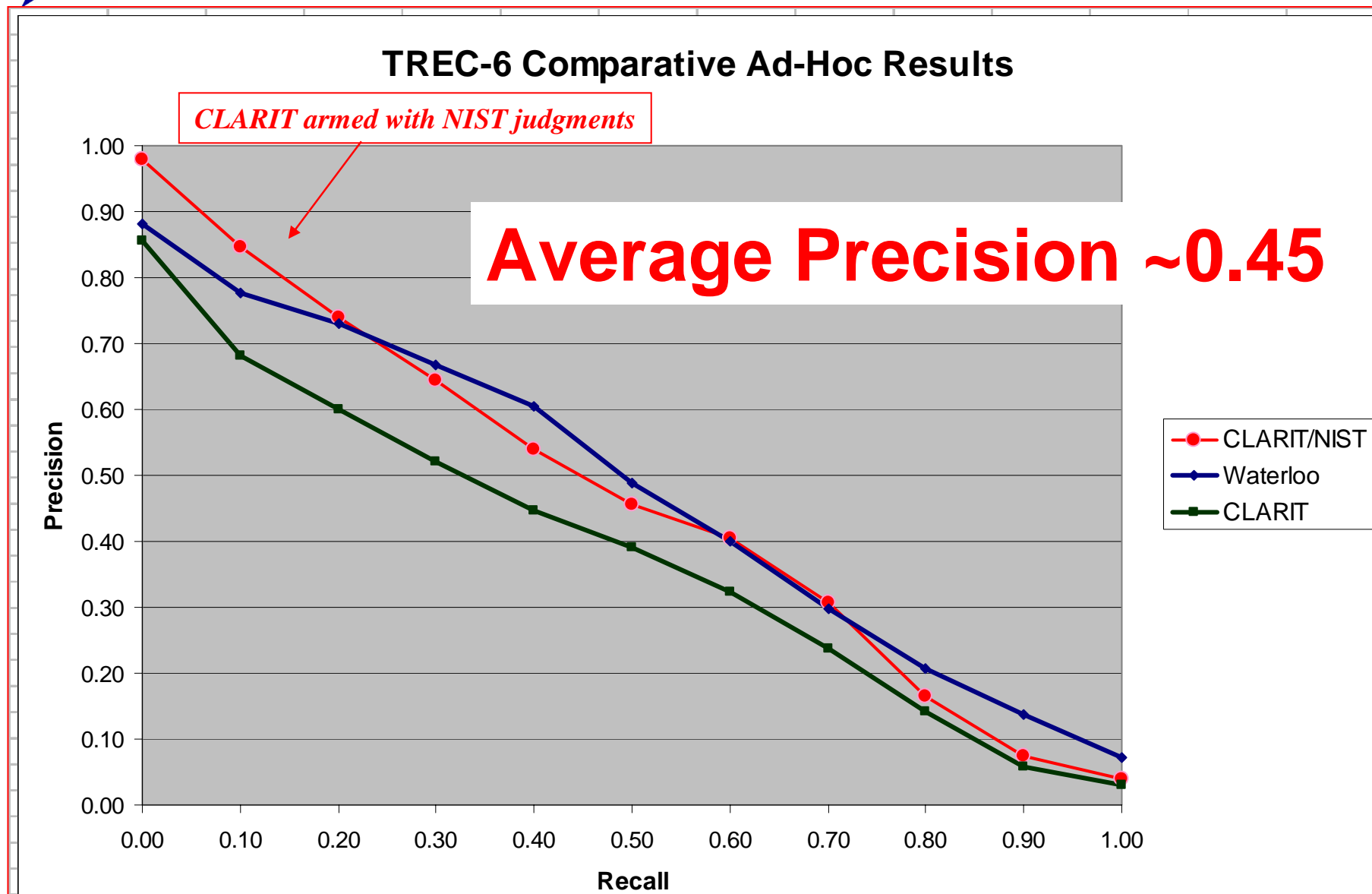
Comparative System Performance

TREC 2 (1993)

| | Average Precision | | Precision at 100 | | R-Precision | | Relative Precision | |
|--------------------|-------------------|--------|------------------|--------|-------------|--------|--------------------|--------|
| Statistical | INQ002 | 0.3565 | INQ002 | 0.5058 | INQ002 | 0.3954 | INQ002 | 0.5459 |
| | siems2 | 0.3408 | CLARTM | 0.4846 | siems2 | 0.3938 | VTcms2 | 0.5207 |
| | CLARTM | 0.3383 | VTcms2 | 0.4790 | CLARTM | 0.3741 | CLARTM | 0.5184 |
| | dortQ2 | 0.3340 | siems2 | 0.4690 | dortQ2 | 0.3722 | siems2 | 0.5109 |
| | Brkly3 | 0.3270 | dortQ2 | 0.4626 | VTcms2 | 0.3708 | TOPIC2 | 0.5020 |
| | cm1L2 | 0.3258 | TOPIC2 | 0.4624 | Brkly3 | 0.3697 | dortQ2 | 0.4972 |
| | VTcms2 | 0.3200 | Brkly3 | 0.4568 | cm1L2 | 0.3641 | Brkly3 | 0.4953 |
| KR-Based | Isiasm | 0.3018 | HNCad1 | 0.4520 | Isiasm | 0.3580 | HNCad1 | 0.4911 |
| | pircs4 | 0.2981 | pircs4 | 0.4494 | HNCad1 | 0.3474 | pircs4 | 0.4818 |
| | citri2 | 0.2874 | cm1L2 | 0.4406 | pircs4 | 0.3467 | cm1L2 | 0.4736 |
| | HNCad1 | 0.2787 | citri2 | 0.4354 | citri2 | 0.3388 | citri2 | 0.4682 |
| | CnQst2 | 0.2633 | Isiasm | 0.4306 | TOPIC2 | 0.3211 | Isiasm | 0.4672 |
| Boolean | schau1 | 0.2517 | CnQst2 | 0.4178 | schau1 | 0.3148 | CnQst2 | 0.4445 |
| | TOPIC2 | 0.2464 | schau1 | 0.4000 | CnQst2 | 0.3140 | schau1 | 0.4323 |
| | cityau | 0.2272 | gecrd2 | 0.3662 | cityau | 0.2849 | gecrd2 | 0.3961 |
| | gecrd2 | 0.2183 | cityau | 0.3512 | gecrd2 | 0.2816 | cityau | 0.3755 |
| | TMC8 | 0.1939 | erima2 | 0.3426 | TMC8 | 0.2612 | erima2 | 0.3639 |
| Boolean | erima2 | 0.1885 | TMC8 | 0.3342 | erima2 | 0.2494 | TMC8 | 0.3549 |
| | uicah | 0.1200 | uicah | 0.2784 | uicah | 0.1950 | prceo1 | 0.2955 |
| | prceo1 | 0.1120 | prceo1 | 0.2722 | prceo1 | 0.1809 | uicah | 0.2932 |
| | UREKA3 | 0.0819 | UREKA3 | 0.1688 | UREKA3 | 0.1231 | UREKA3 | 0.1779 |



What is the Limit?



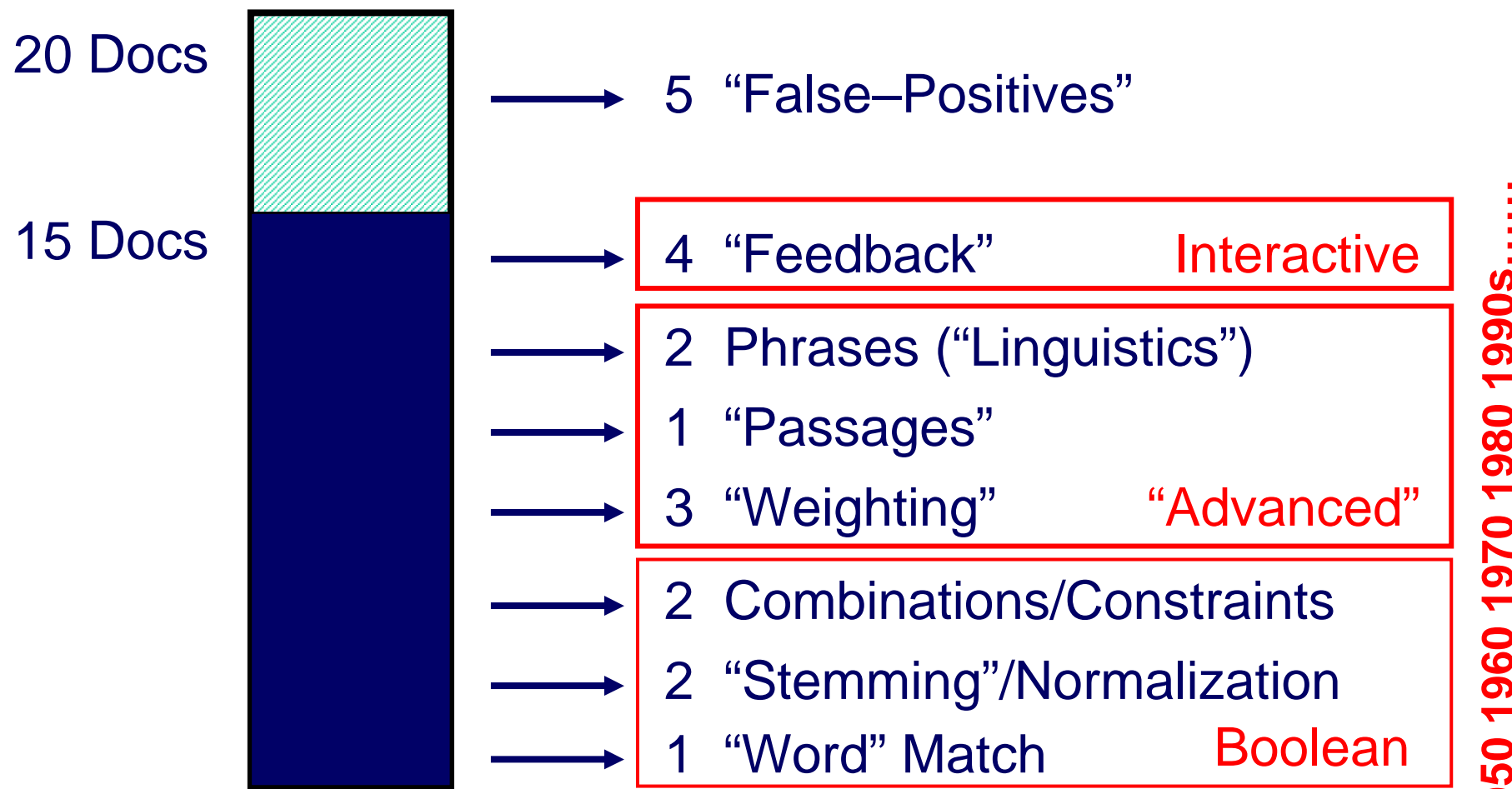


Optimizing IR



Components of “Quality”?

Good Queries (10+ Terms) / 1M Documents





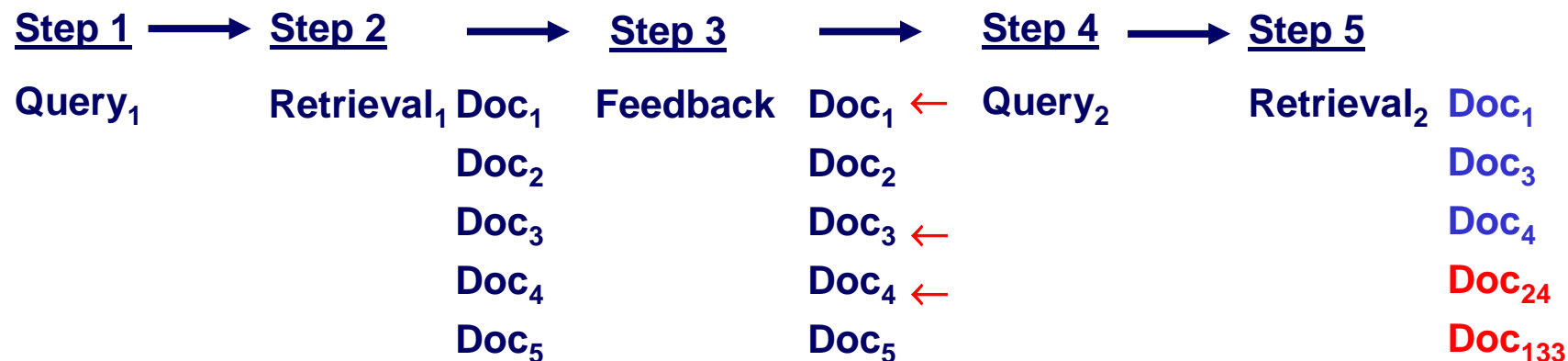
Pseudo-Relevance Feedback



1966: User Feedback

J.J. Rocchio

- Retrieve Documents
- User Identifies a Few Good Ones
- System Selects Terms from These Documents to add to the Query
- New Set of Documents is Retrieved

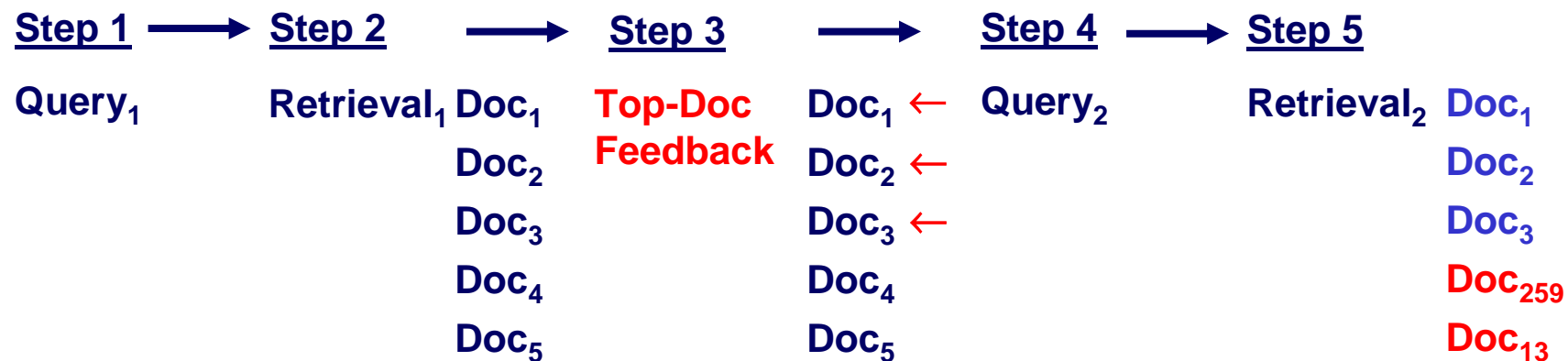




1993: Automatic Feedback

CLARIT

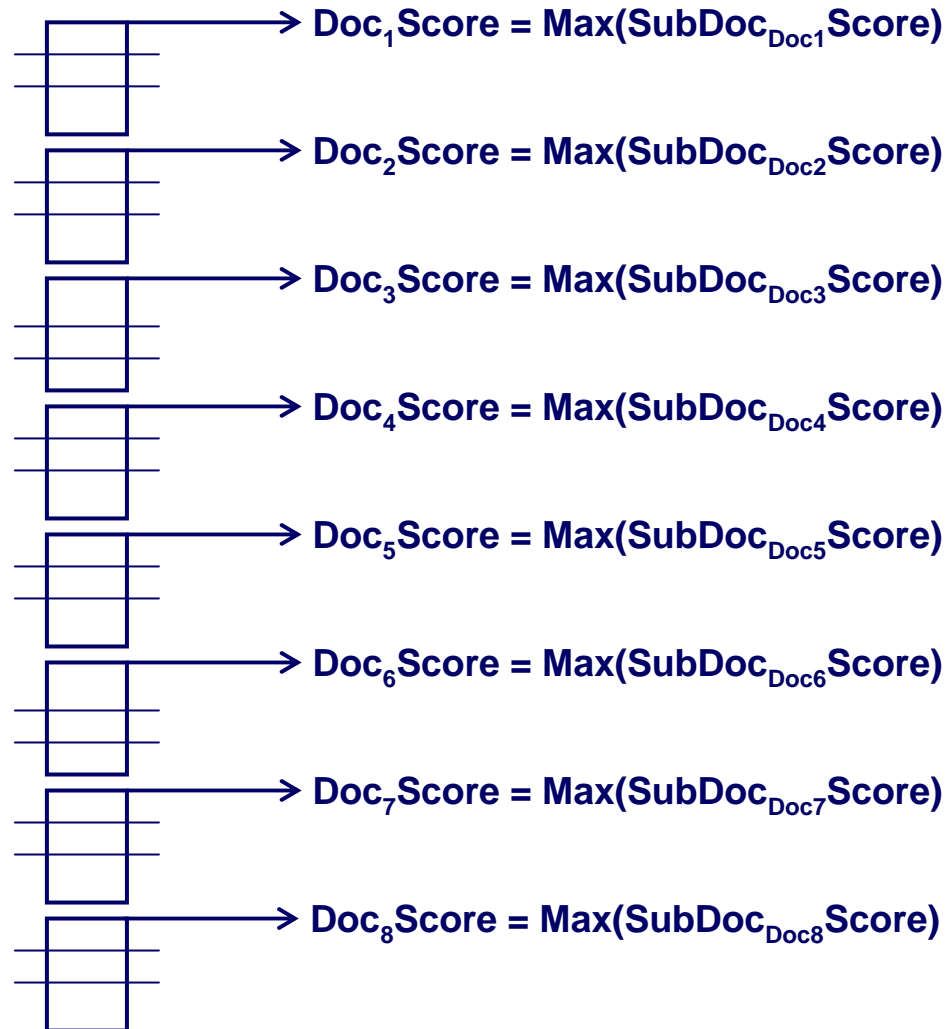
- Pseudo-Relevance Feedback





CLARIT Document-Scoring Strategy

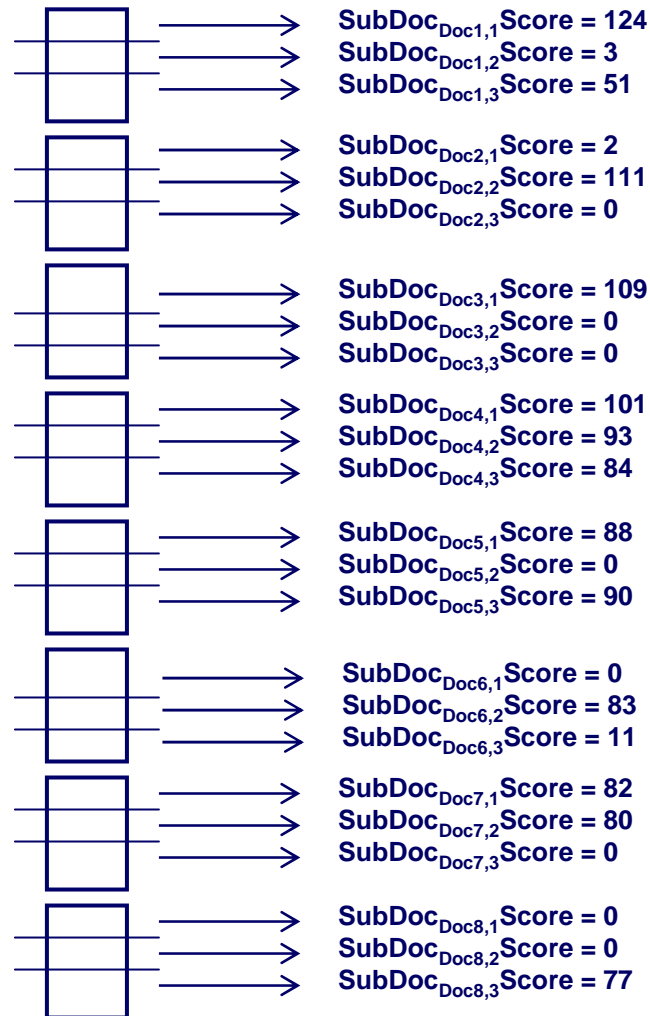
Query \Rightarrow





CLARIT SubDoc Scoring

Query \Rightarrow

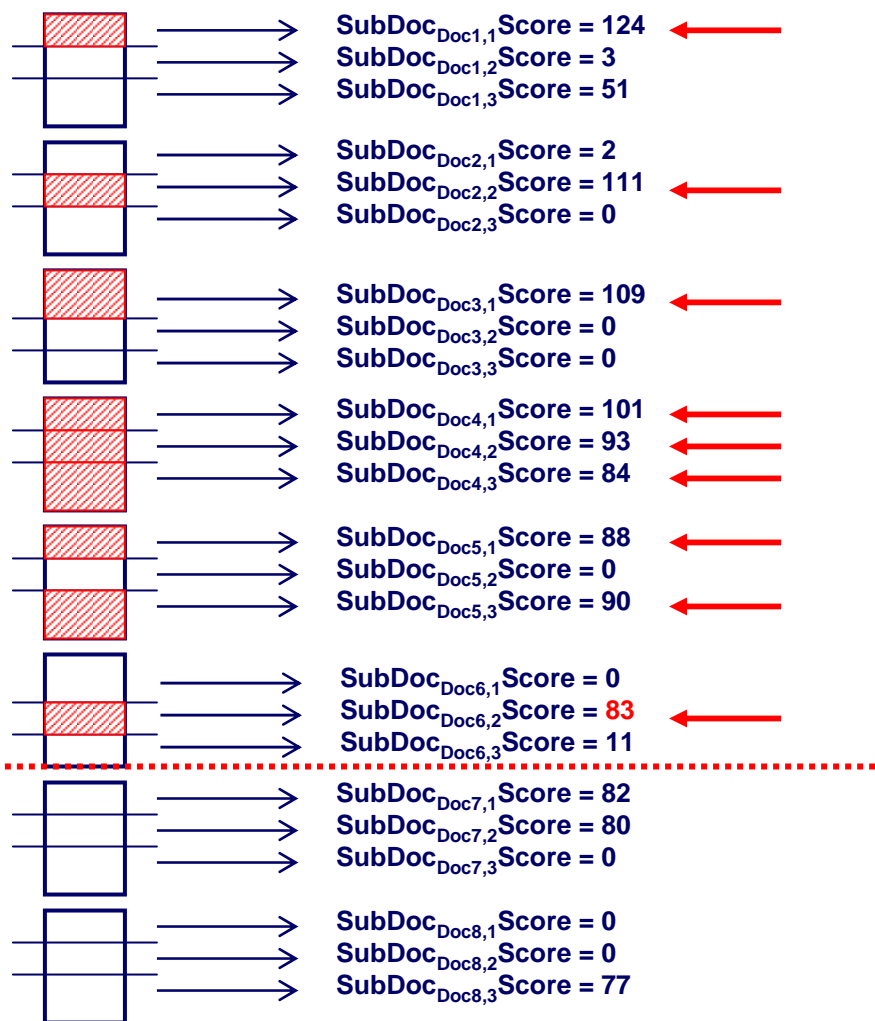




Feedback Selection (Detail)

Query ⇒

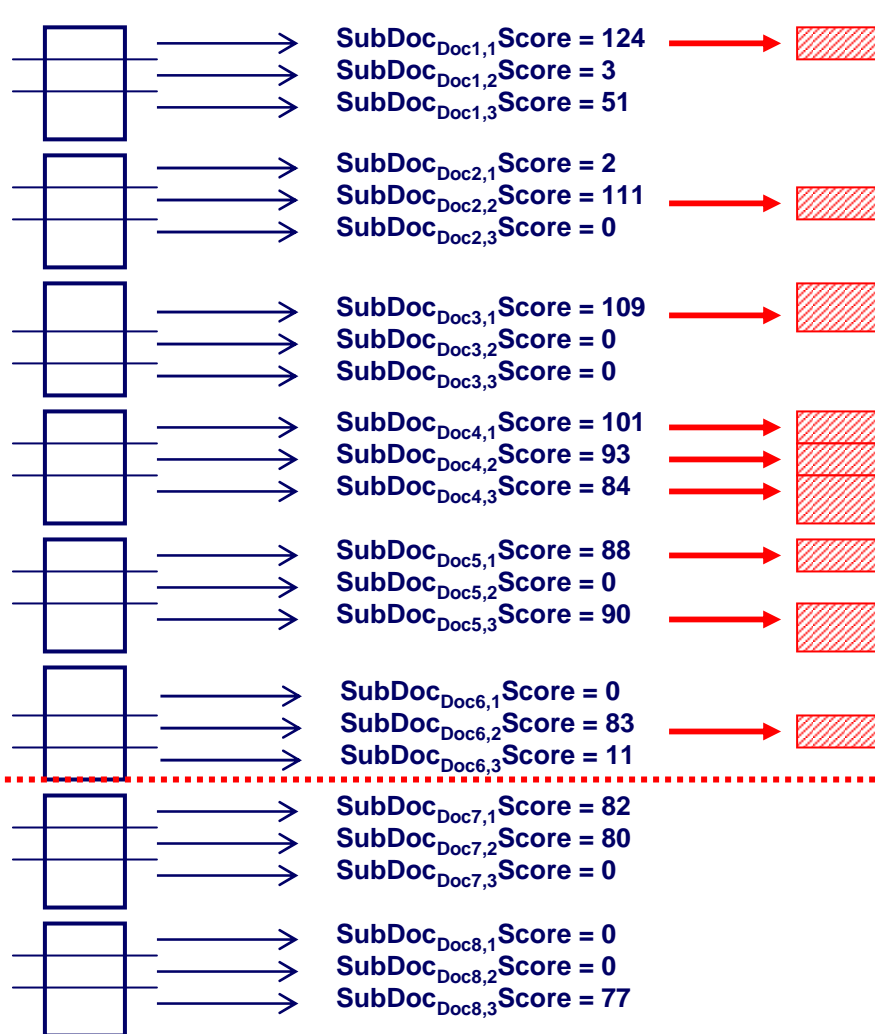
“Top 6”





Feedback Selection (Detail)

Query ⇒



“Top 6”

Thesaurus
Extraction
⇒
Feedback



Example: Response to Q310

**Q310:
“Evidence that radio
waves from radio
towers or car
phones affect brain
cancer occurrence”**

**1: (79319) FT931-11958
FT 30 JAN 93 / International Company News: US
mobile phone companies hit by
brain cancer scare (248.01)**

**2: (9824) FT921-7096
FT 25 FEB 92 / Making waves for the world: New
frequencies (193.45)**

**3: (109599) FT933-7815
FT 19 AUG 93 / Mobiles break into the big time:
Can cellular communications
replace traditional networks (139.03)**

**4: (55650) FT924-5286
FT 27 NOV 92 / Letter: 0 out of 10 (133.87)**

**5: (70797) FT931-4108
FT 12 MAR 93 / Technology: Guarding against
radio waves - Worth Watching (132.55)**

**6: (35765) FT922-15435
FT 01 APR 92 / Cancer link to power lines
'inconclusive' (127.00)**



Alternative Features from “Top 6”

19.4222 (Radio waves) **Prob2**

17.2147 (brain cancer)

15.3293 (Torremolinos)

14.2489 (Cancer link)

13.5272 (mobile)

12.7227 (cellular phones)

12.4477 (New frequencies)

12.4176 (frequencies)

12.2084 (phones)

12.0718 (McCaw)

11.9941 (far, delegates)

11.9941 (fixed telecommunications links)

11.9941 (brain cancer scare)

11.9941 (bouncing signals)

11.9941 (radio waves - Worth Watching)

11.9941 (Mobiles break)

11.9941 (Two main considerations)

11.9941 (powerful voting bloc)

11.9941 (rival agendas)

11.9941 (different technological approaches)

11.9941 (radio regulations)

11.9941 (various national interests)

11.9941 (future colony)

11.9941 (mobile phone companies hit)

11.9941 (global technical standards)

11.8535 (radio programmes)

11.4868 (car phones)

11.2483 (mobile phones)

11.2326 (use frequencies)

11.2326 (radio-based services)

...

3.10544 (Radio waves) **Rocchio**

2.50403 (mobile)

2.40001 (phones)

2.2953 (waves)

2.20213 (communications)

2.16951 (frequencies)

1.86238 (brain cancer)

1.67065 (cancer)

1.60654 (Torremolinos)

1.56585 (cellular phones)

1.50992 (technology)

1.48446 (cellular communications)

1.47164 (McCaw)

1.40887 (telephones)

1.39494 (mobile phones)

1.33967 (Cancer link)

1.29725 (scare)

1.27557 (services)

1.25709 (New frequencies)

1.23913 (new services)

1.19067 (car phones)

1.1675 (radio programmes)

1.12297 (radio waves - Worth Watching)

1.1217 (crowded and demands)

1.1217 (imaginative new services)

1.06692 (HDTV)

1.0588 (Mobiles break)

1.05039 (future)

1.0371 (European nations)

1.00999 (wire)

...

8.10115 (Radio waves) **RocchioFQ**

7.93649 (mobile)

7.84788 (cancer)

7.65708 (frequencies)

7.43219 (communications)

6.80806 (phones)

6.43268 (waves)

6.29029 (satellite)

4.82523 (telephones)

4.67565 (services)

4.58431 (brain cancer)

4.16557 (wire)

4.08937 (electro-magnetic fields)

3.84256 (technology)

3.70839 (cellular communications)

3.65359 (mobile communications)

3.57615 (fixed wire)

3.55222 (McCaw)

3.33014 (T)

3.28499 (earth's surface)

3.21504 (new services)

3.09427 (NRPB)

2.95355 (cellular phones)

2.93231 (radio programmes)

2.83377 (space communications)

2.67972 (HDTV)

2.64488 (scare)

2.61404 (smaller handsets)

2.61068 (network)

2.58593 (Handsets)

...



Pseudo-Relevance Feedback (Blind Feedback (BF)) Is Remarkably Robust!

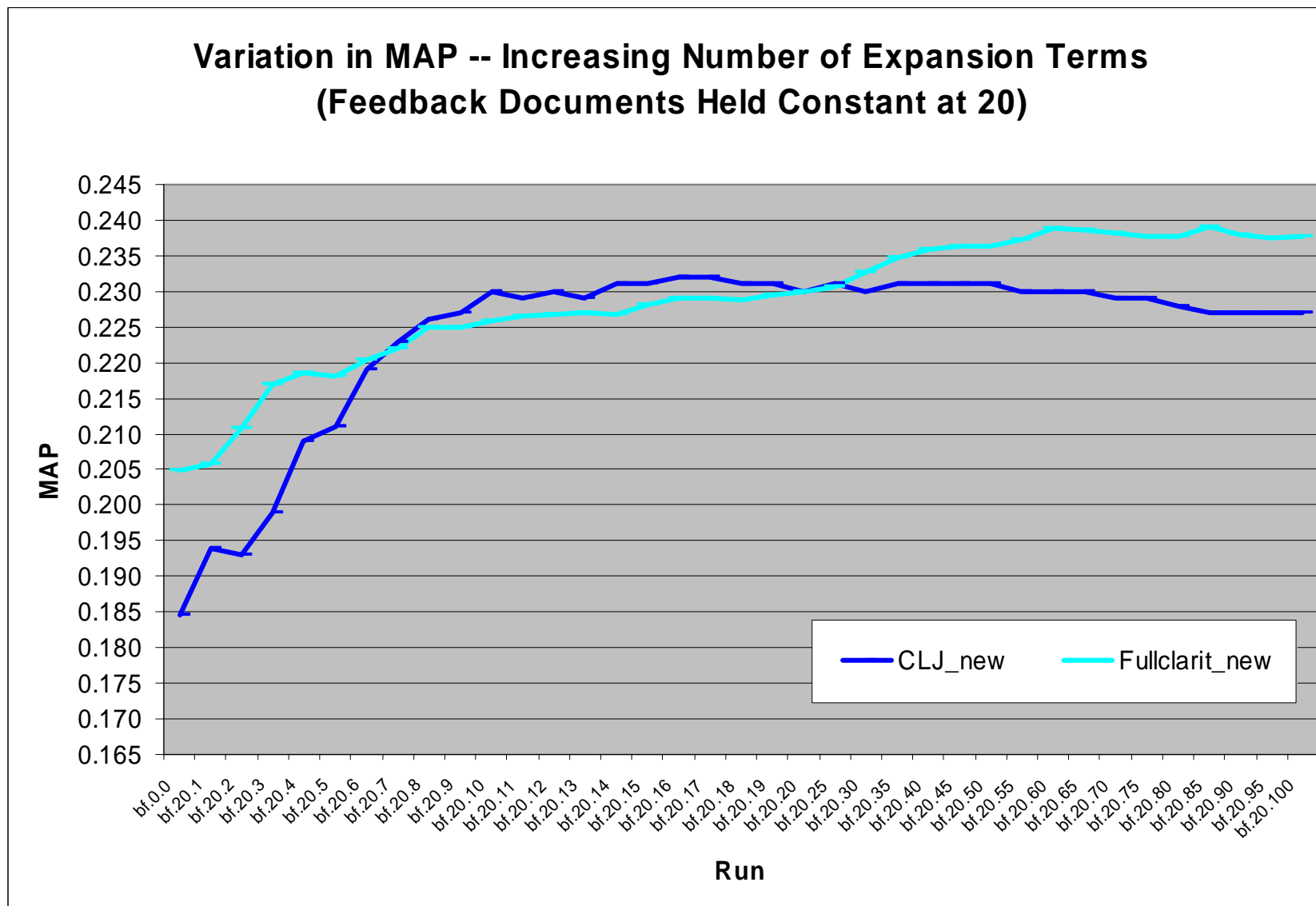


In-Depth Evaluation of BF Effects

- **Summer Workshop involving 8 Systems**
 - Many different retrieval paradigms represented
 - Teams explored variations in BF
 - All systems operated in fully automatic mode
- **150 Topics (Queries) over 1M+ DB**
 - A particularly difficult set of topics
 - Failure analysis of particularly hard topics

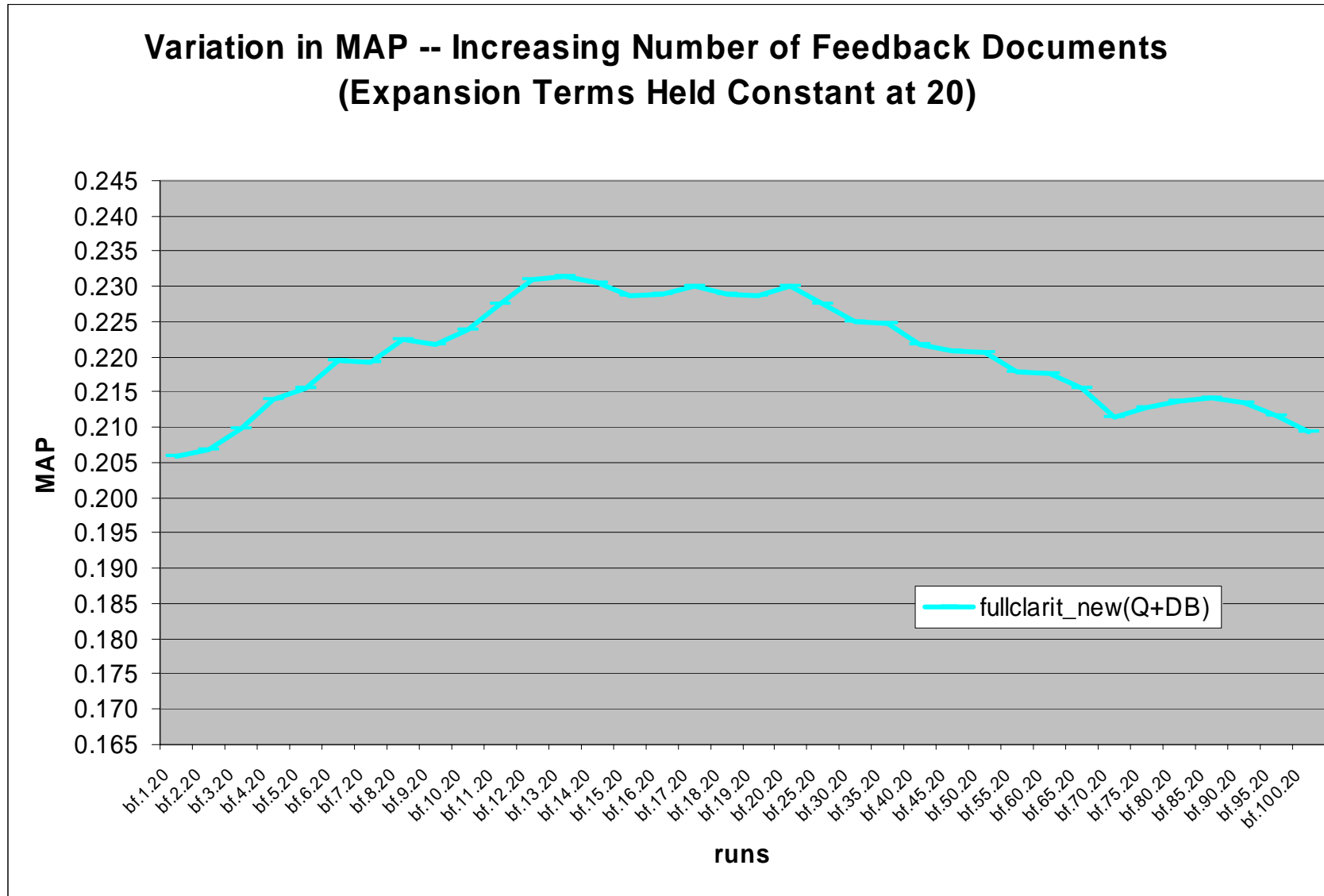


Results with Increasing Terms



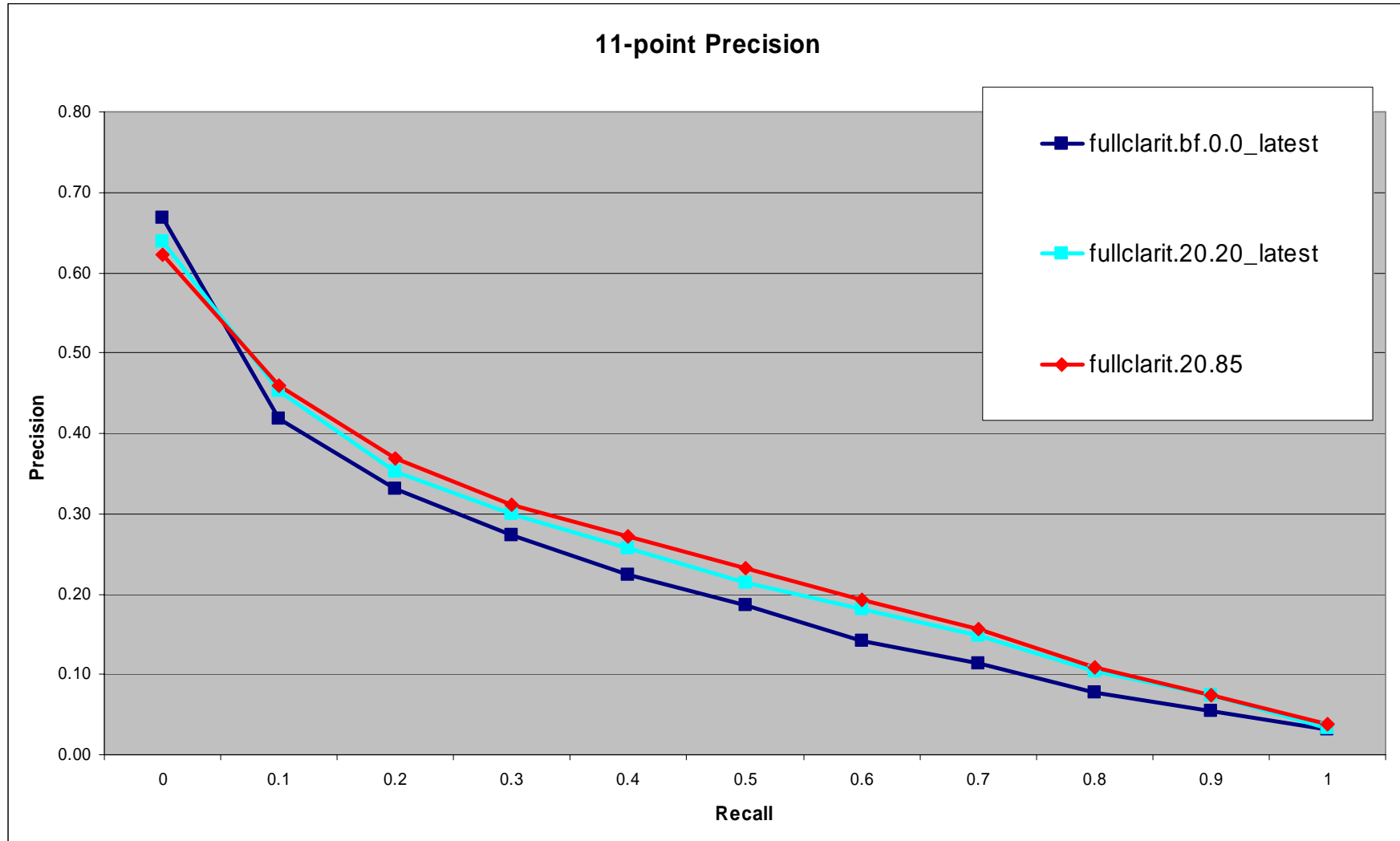


Results with Increasing Documents



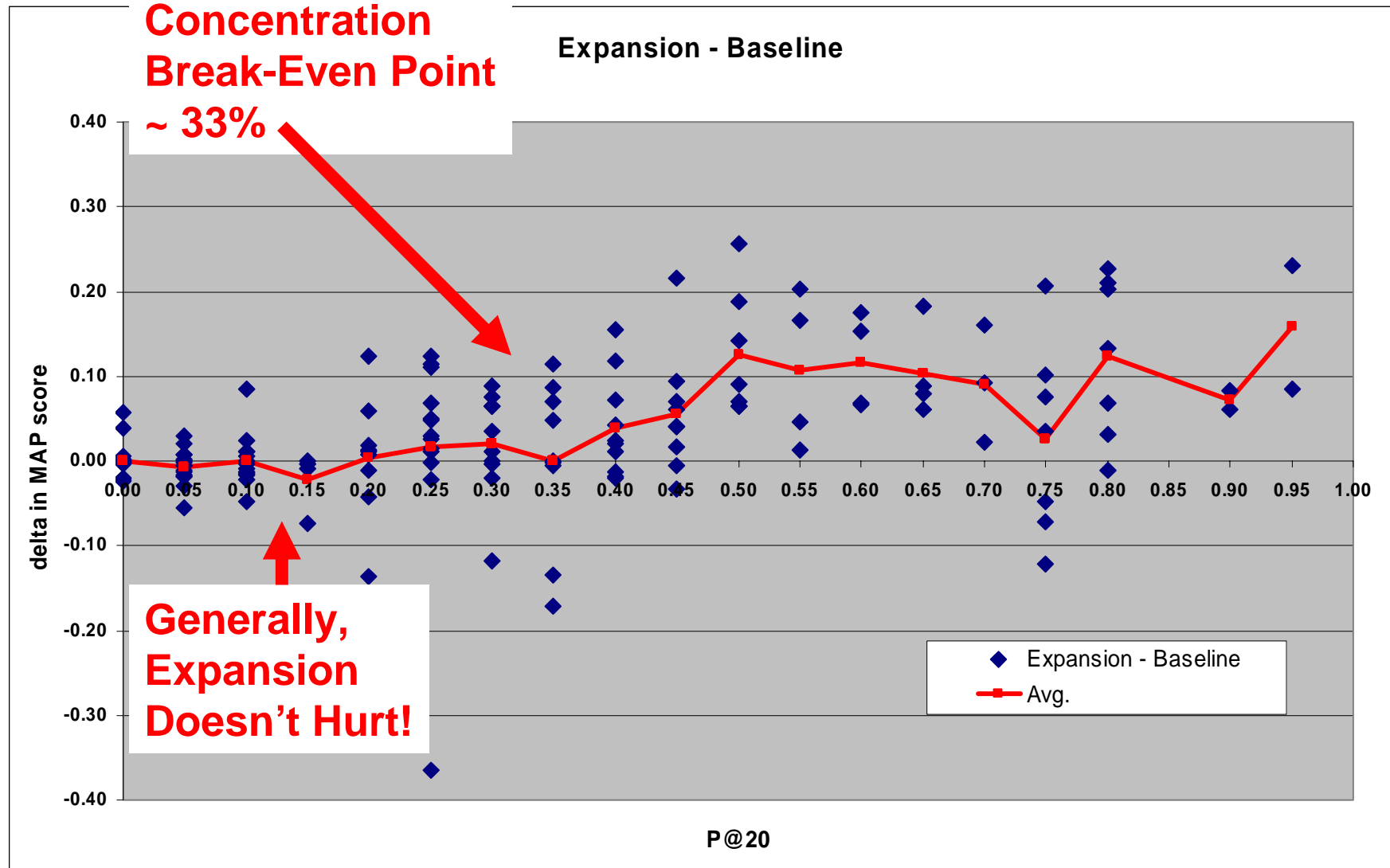


Comparative Average Precision





Effect of Concentration of Relevants





The Problem

- **If (pseudo-relevance) feedback helps, and if the effect is greater when more “true relevants” are in the feedback set, then how can we select sub-sets of responding documents that include more relevants?**

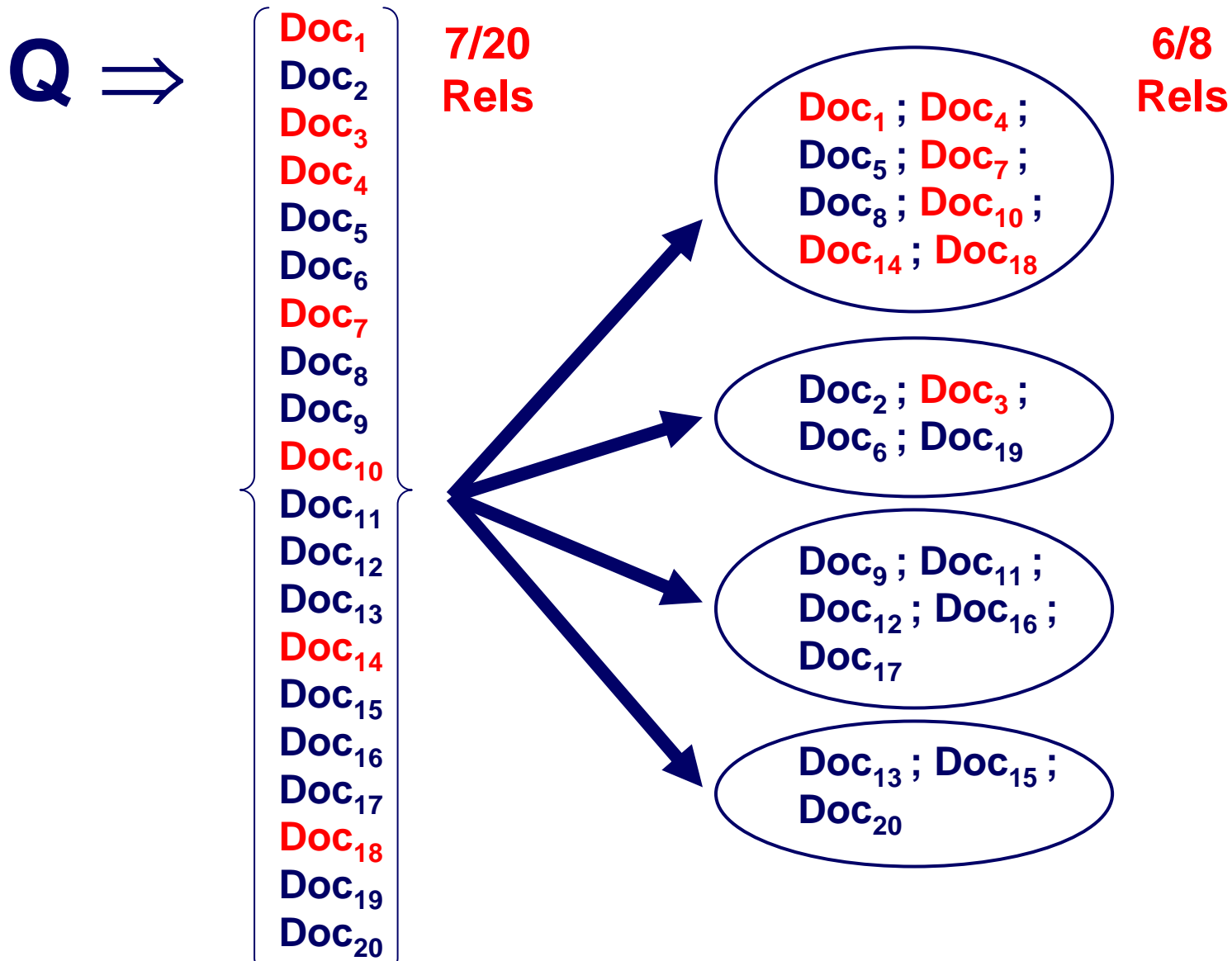


An Observation...

- **Relevant Documents tend to Cluster Together**
- **Clusters can “Concentrate” Relevant Documents**



Illustration of Clustering Effect





A Note on Clustering



What is Clustering ?

- **Clustering automatically groups objects based on their contents without any pre-defined classes**
 - objects within a cluster are similar to each other
 - objects in different clusters are different
- **Objects can be**
 - documents / subdocuments / sentences
 - concepts
 - terms



Document Clustering Procedures

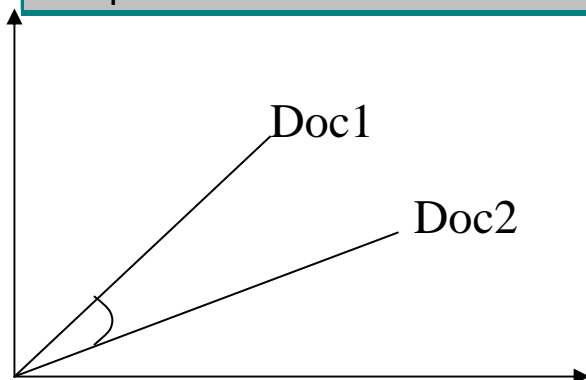
- **Document analysis**
 - segment documents into text components
 - extract features using NLP
 - select features based on statistics and selection methods
- **Clustering**
 - find the relationship among the documents based on the similarity function
 - group the documents into tree structure (Hierarchical algorithms) or flat categories (Partitioning Algorithms)
- **Tree slicing (Hierarchical algorithms only)**
 - flatten the tree to reflect the desired granularity



Document Similarity

The Vector Space Model

| | <i>Doc1</i> | <i>Doc2</i> | <i>Doc3</i> |
|------------|-------------|-------------|-------------|
| Money | 0.5 | 0.3 | |
| Credit | 0.7 | | |
| Stock | 0.5 | 0.6 | |
| Fund | | | |
| Technology | | 0.8 | 1 |
| Internet | | 0.5 | 0.6 |
| Computer | | | 0.4 |



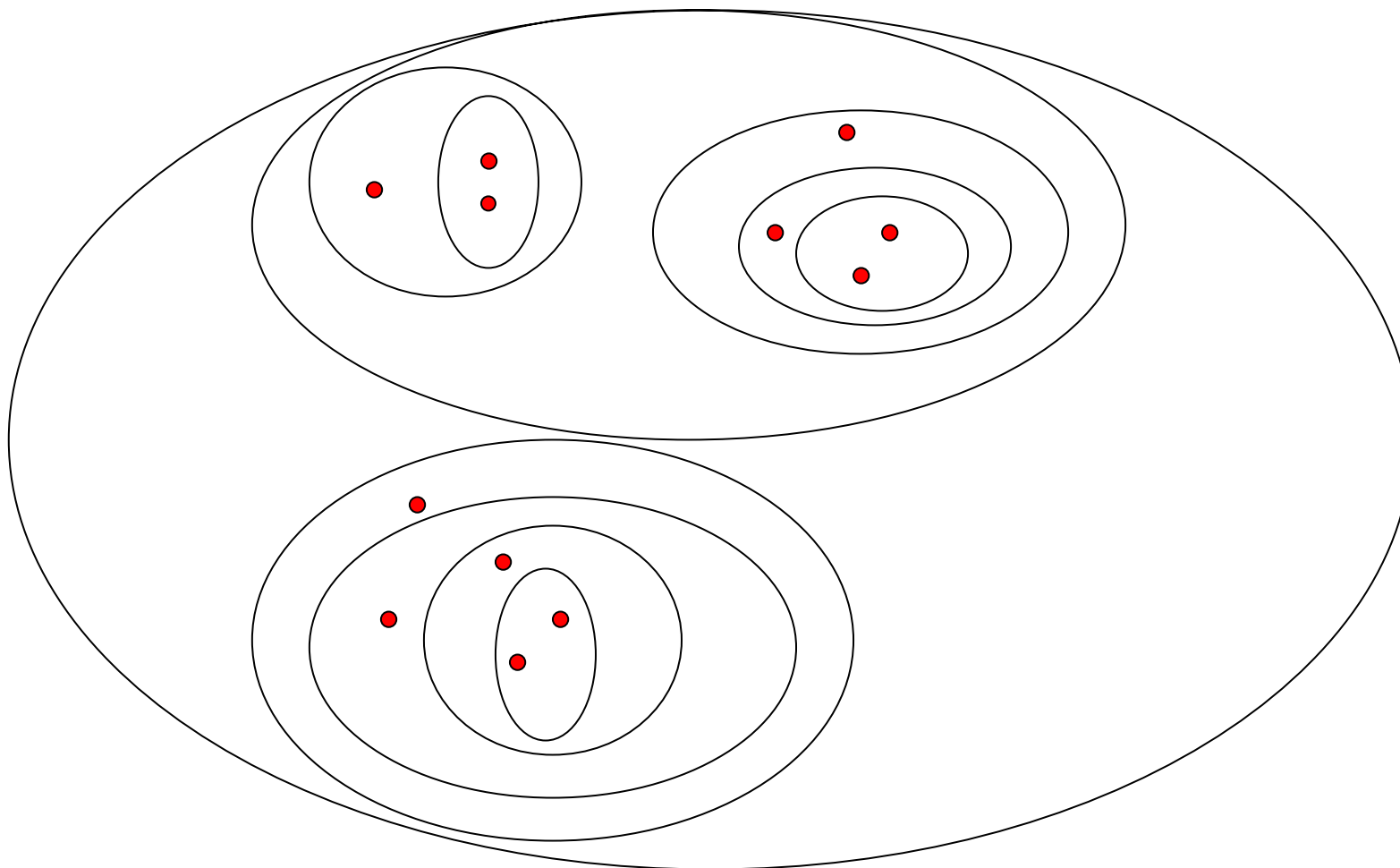
Documents are represented by vectors of weighted features in Vector Space

The proximity of two documents is computed by similarity functions, e.g., Cosine, Dice, Jaccard

Cosine measure is the angle between the two vectors



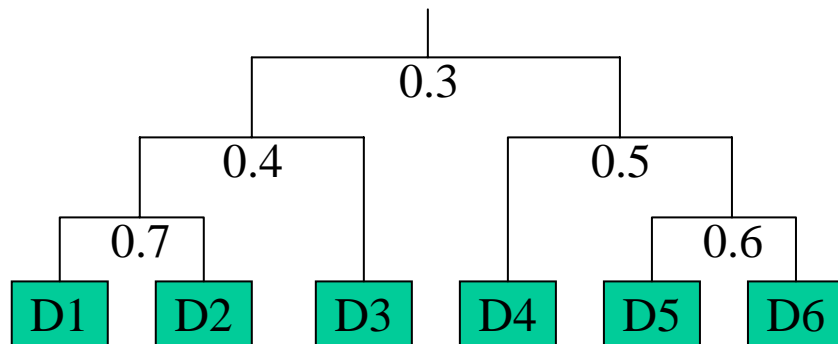
Hierarchical Clustering





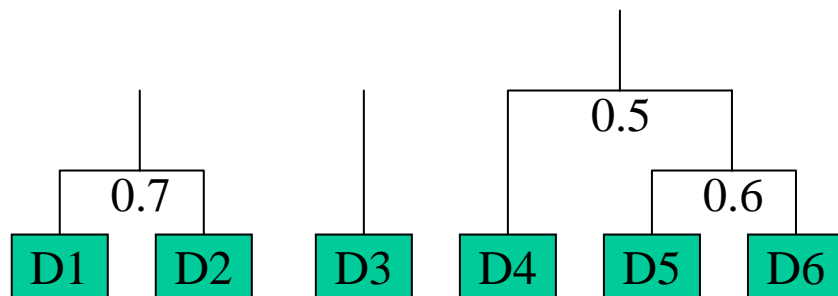
Tree Slicing

(Hierarchical Algorithms)



Slice the tree
with threshold
0.45

Cluster1 Cluster2 Cluster3



- The clustering results from hierarchical algorithms can be represented by a binary tree

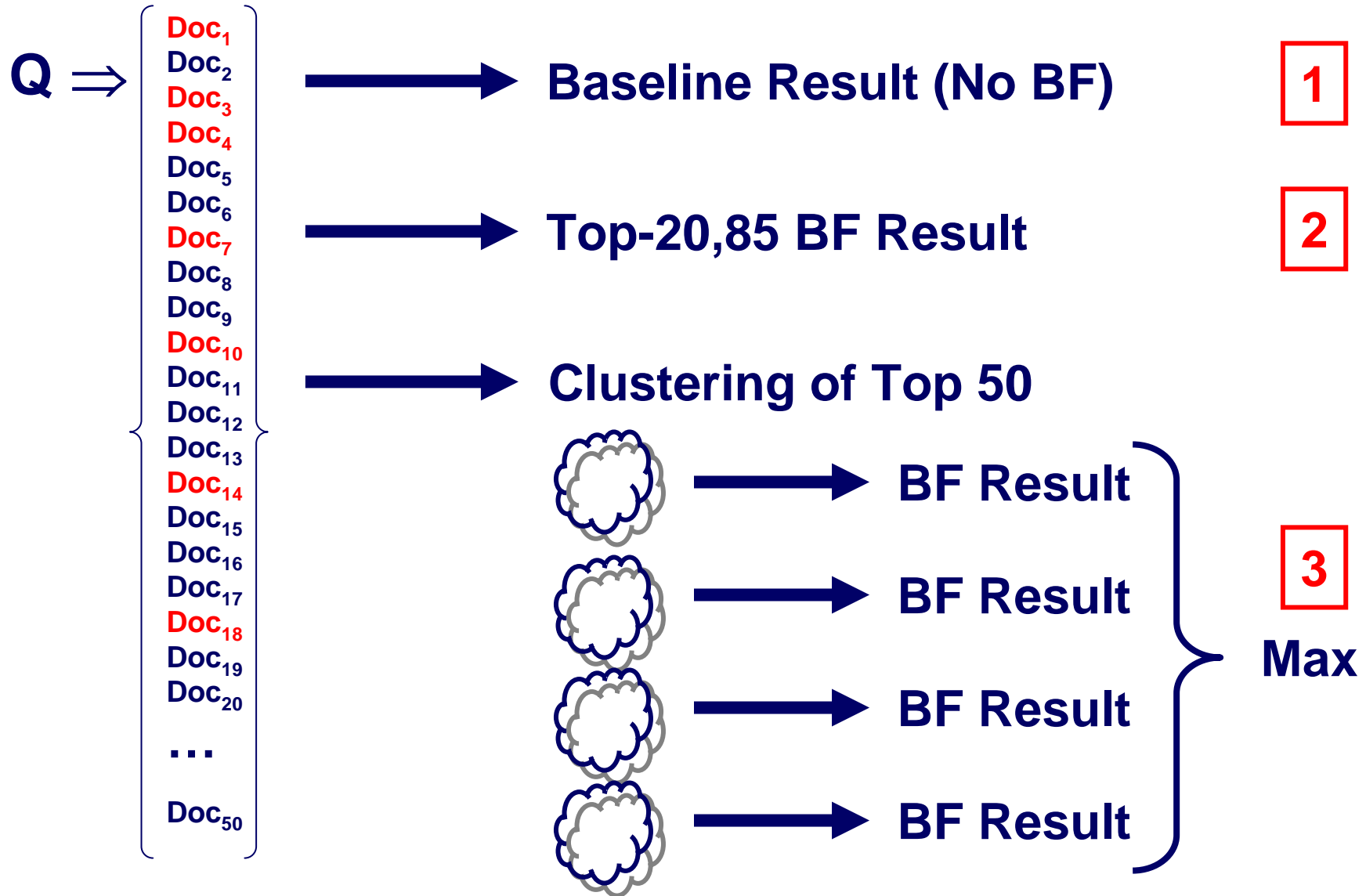
- Flattening the clustering tree using a threshold to reflect the user's desired granularity of the structure of the document set



IR Results Using Clustering



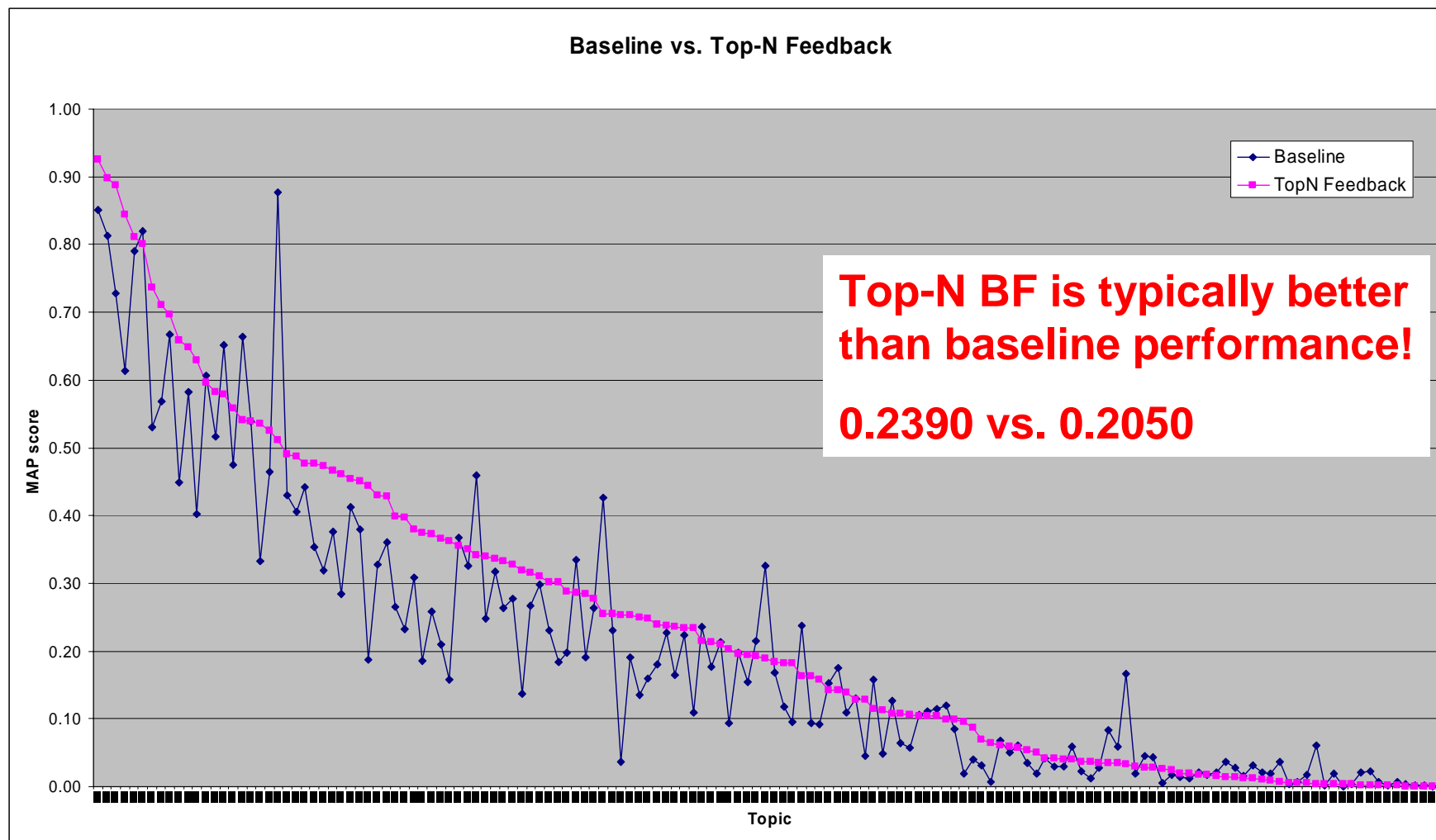
Can a Cluster Out-Perform Top-N?





Per-Topic Performance

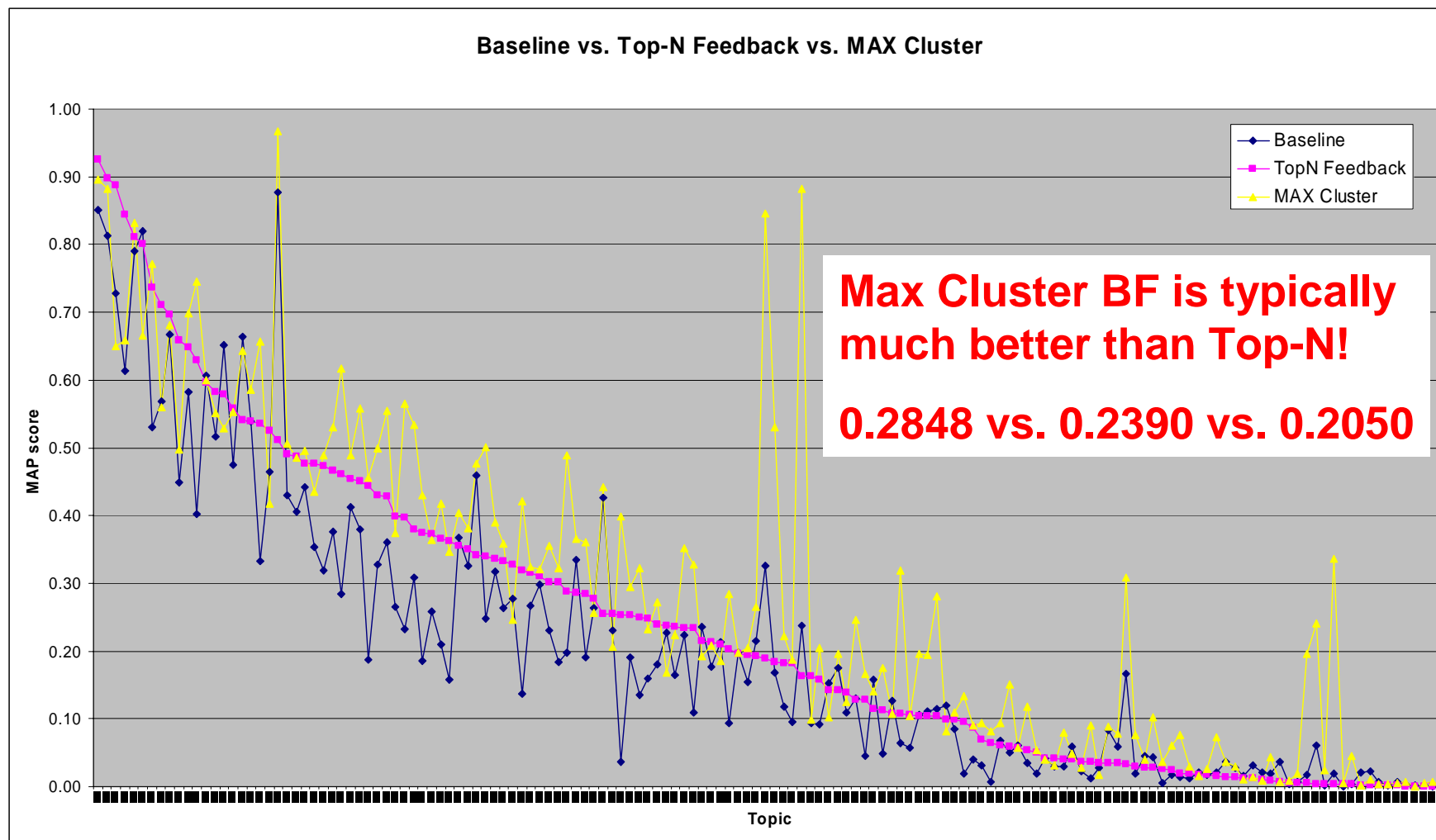
Sorted by Decreasing Performance of Top-N BF





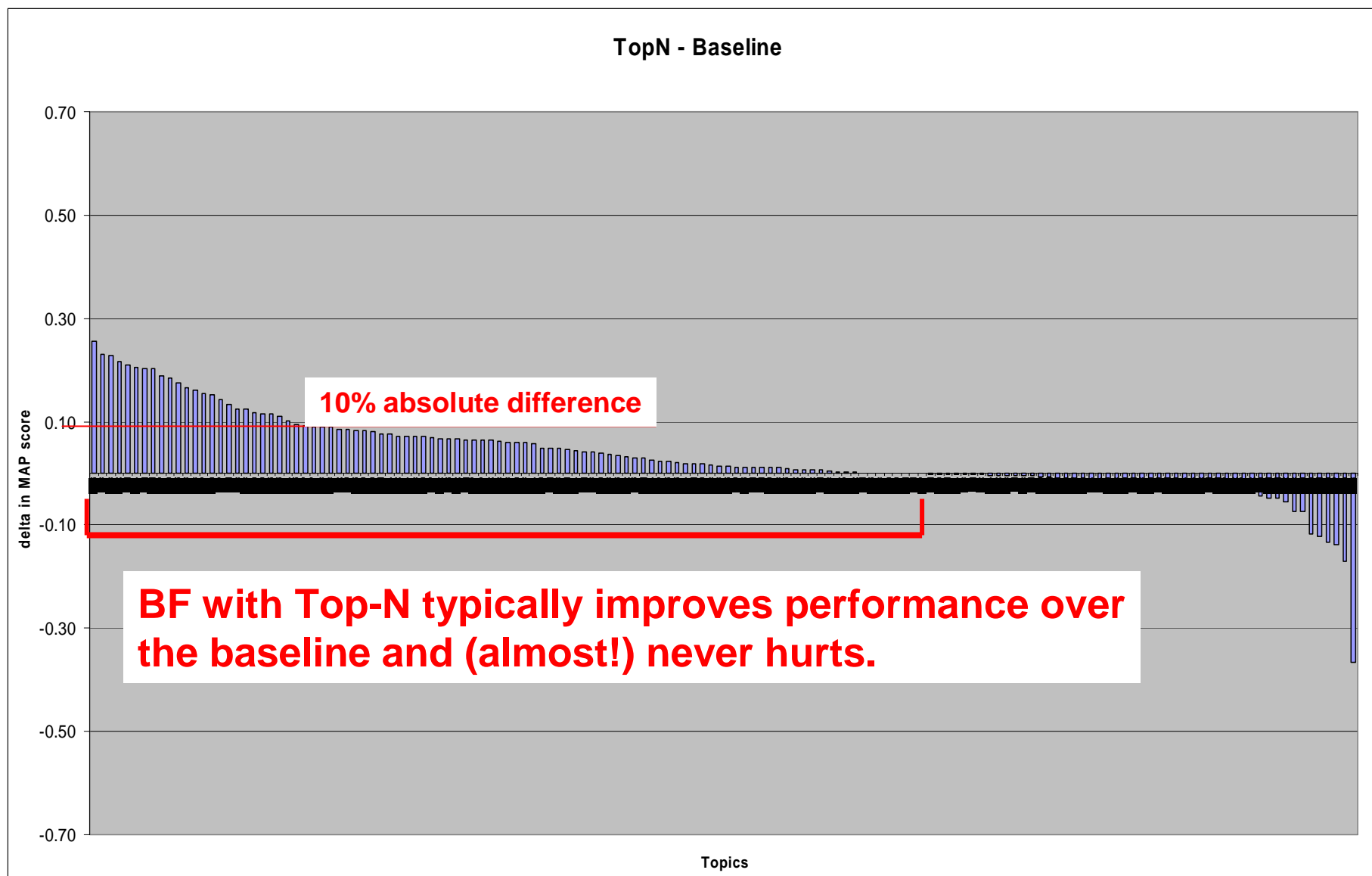
Per-Topic Performance

Sorted by Decreasing Performance of Top-N BF



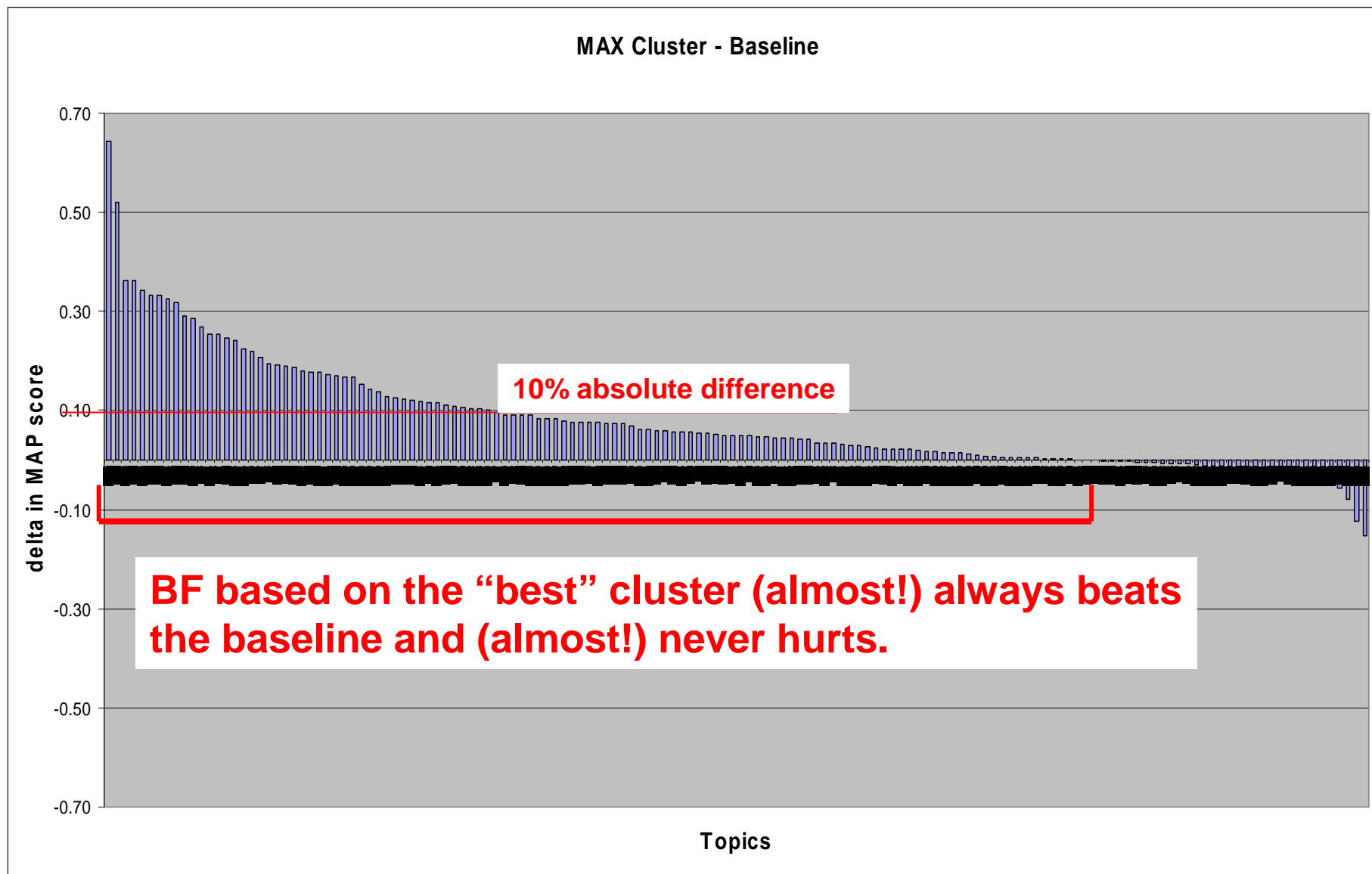


Per-Topic Performance



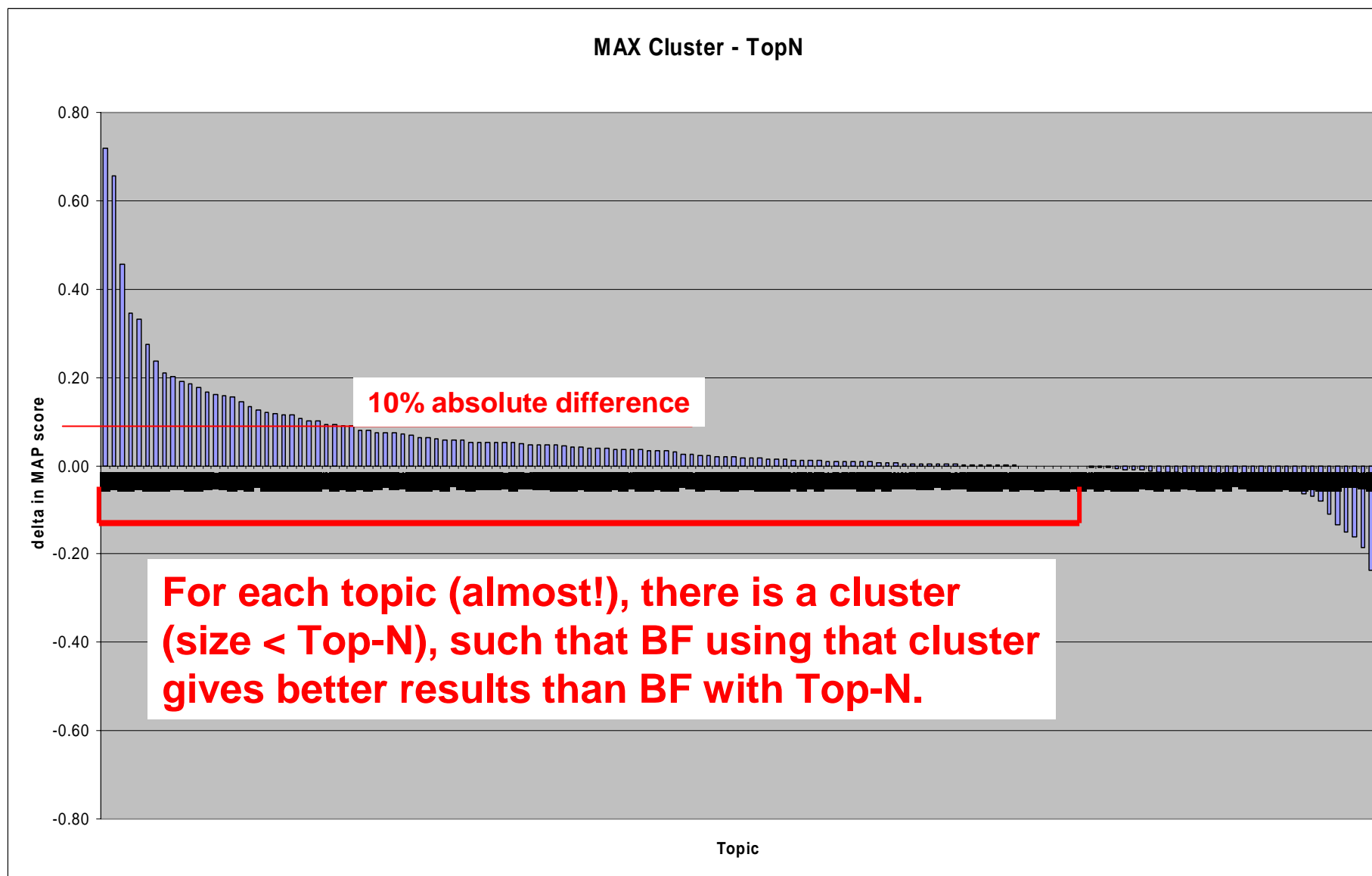


Per-Topic Performance





Per-Topic Performance





Challenge: Pick the Right Cluster!

- **Given that BF based on a “good” cluster can lead to better performance than BF based on Top-N, the problem is now, simply, *automatically*, to pick the “right” cluster...**
- **While working on this problem, we might observe the following:**
- **People are very good at recognizing clusters that contain relevant documents. This suggests that presenting retrieval results in clusters specifically for feedback choices is a good strategy.**



Clustering for Corpus Analysis

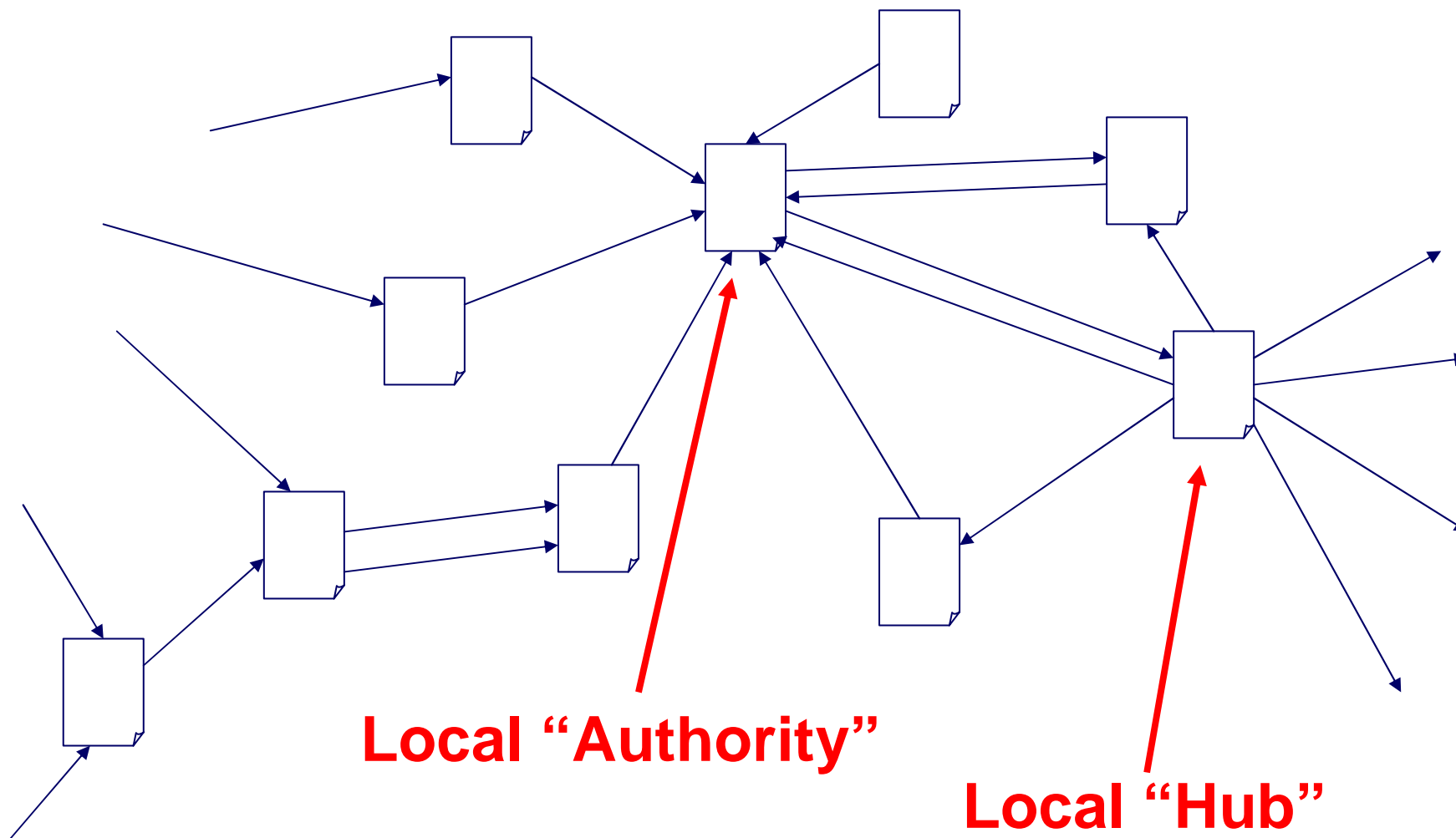


Documents are not Created Equal

- **In any DB of documents (and in any set of results to a query!), there are likely to be documents of different status for each topic.**
 - Surveys
 - Detailed reports
 - Abstracts
 - Critical Reviews
 - Technical Notes (vs. Published Articles)
 - Patents (vs. Papers)
 - News Articles
 - Etc.
- **Structure (links) among documents (as on the Web) can be exploited to help differentiate status, but what can we do with free-text documents in DBs?**



Hubs & Authorities Analysis



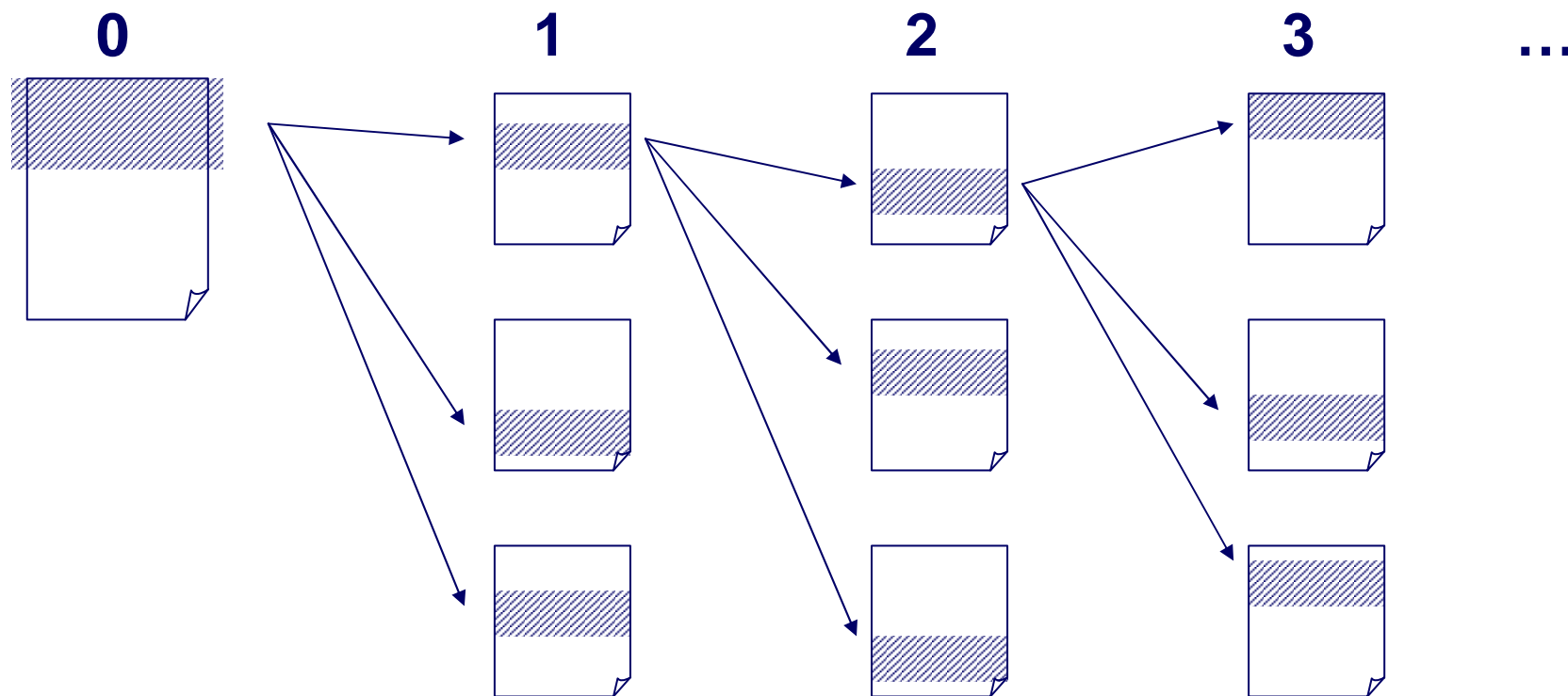


Structure Analysis in Free-Text DBs

- **Step 1: Connect each document to others via “similarity” links, based on sub-document units; keep track of the “in” and “out” links**
- **Step 2: Cluster the sub-documents**

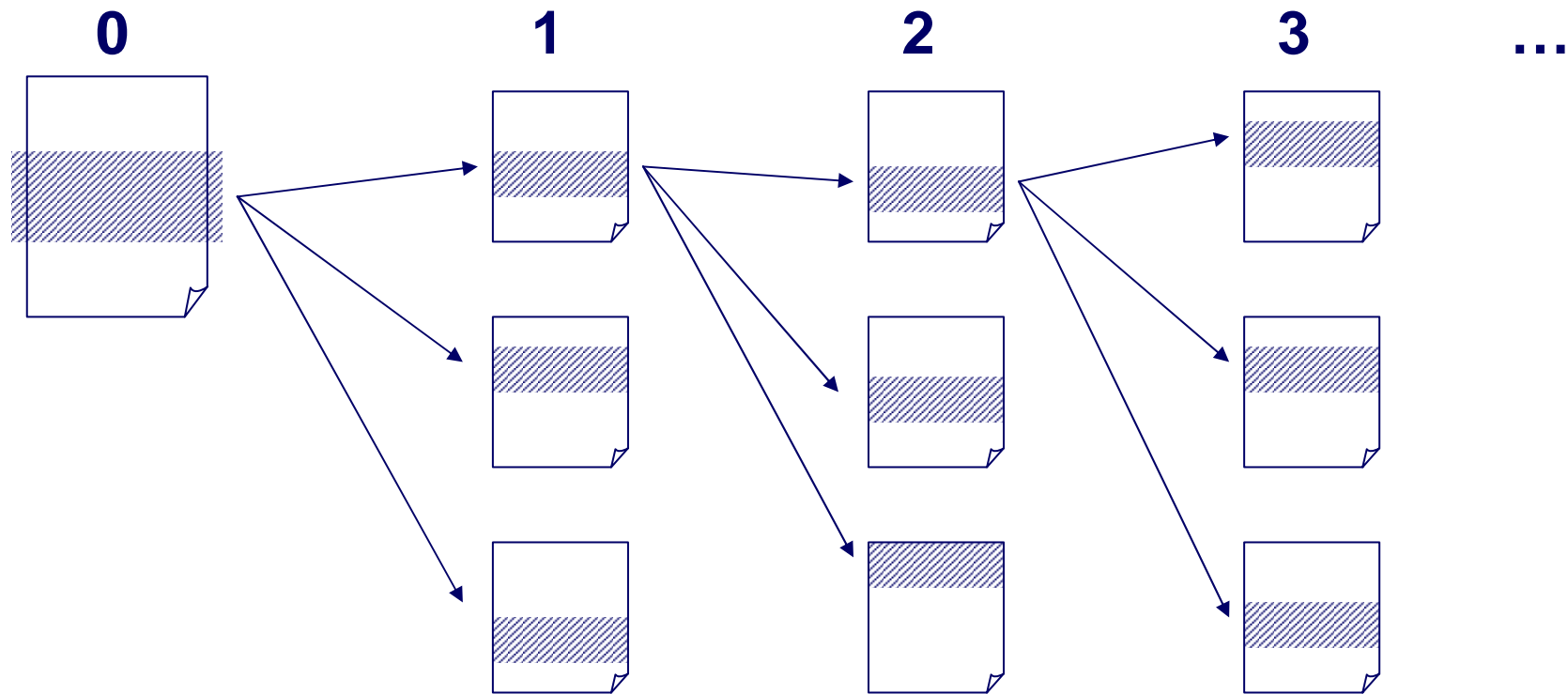


Similarity-Link Network



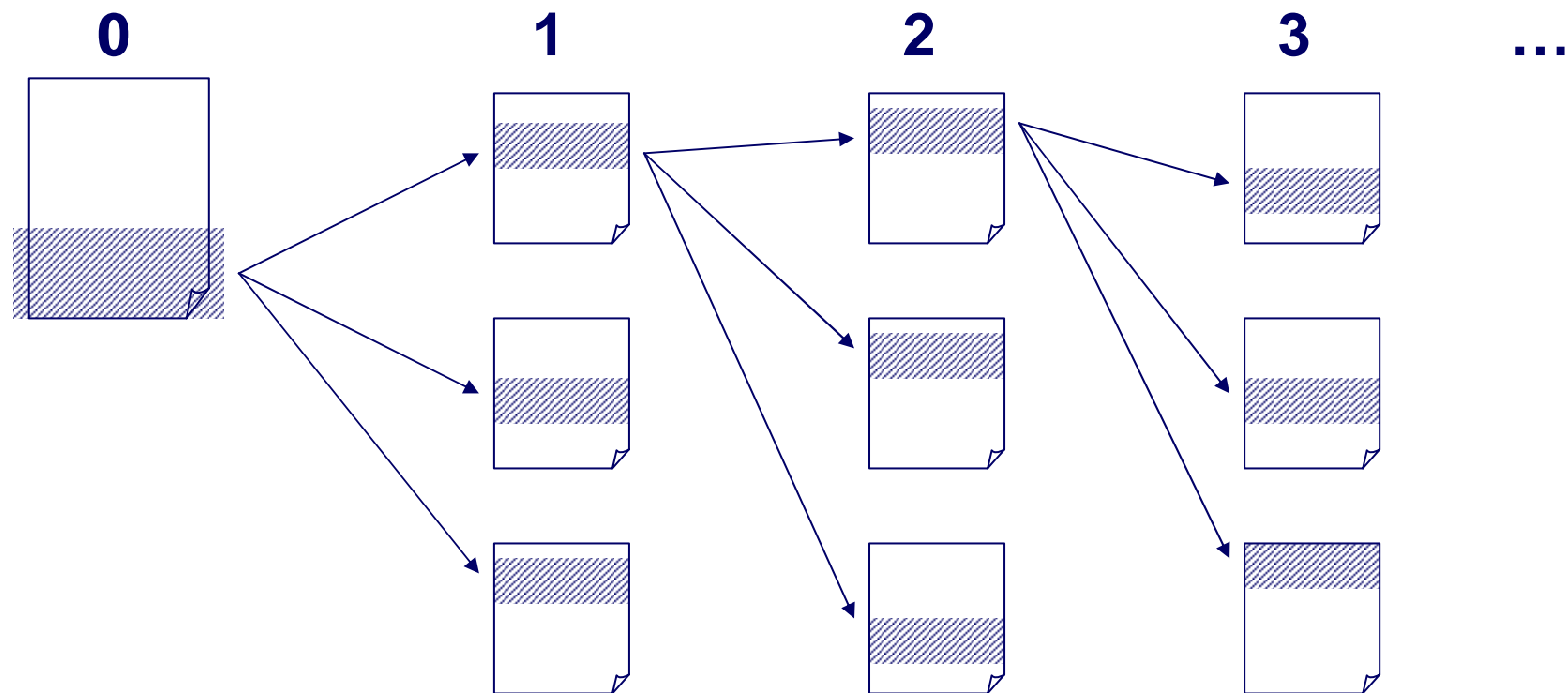


Similarity-Link Network



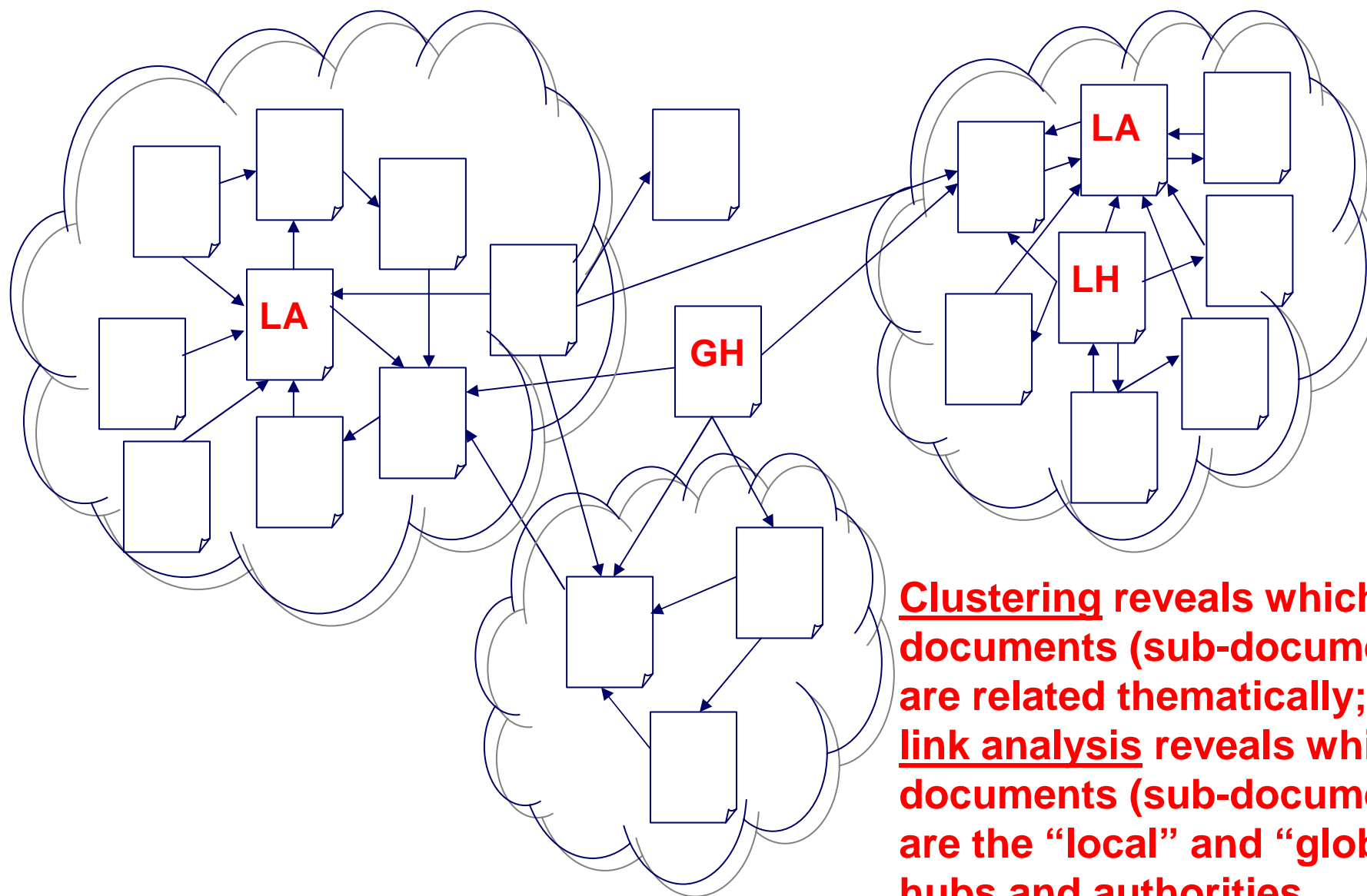


Similarity-Link Network





Similarity-Link H/A Analysis



Clustering reveals which documents (sub-documents) are related thematically; link analysis reveals which documents (sub-documents) are the “local” and “global” hubs and authorities.



An Illustration of Structure Discovery



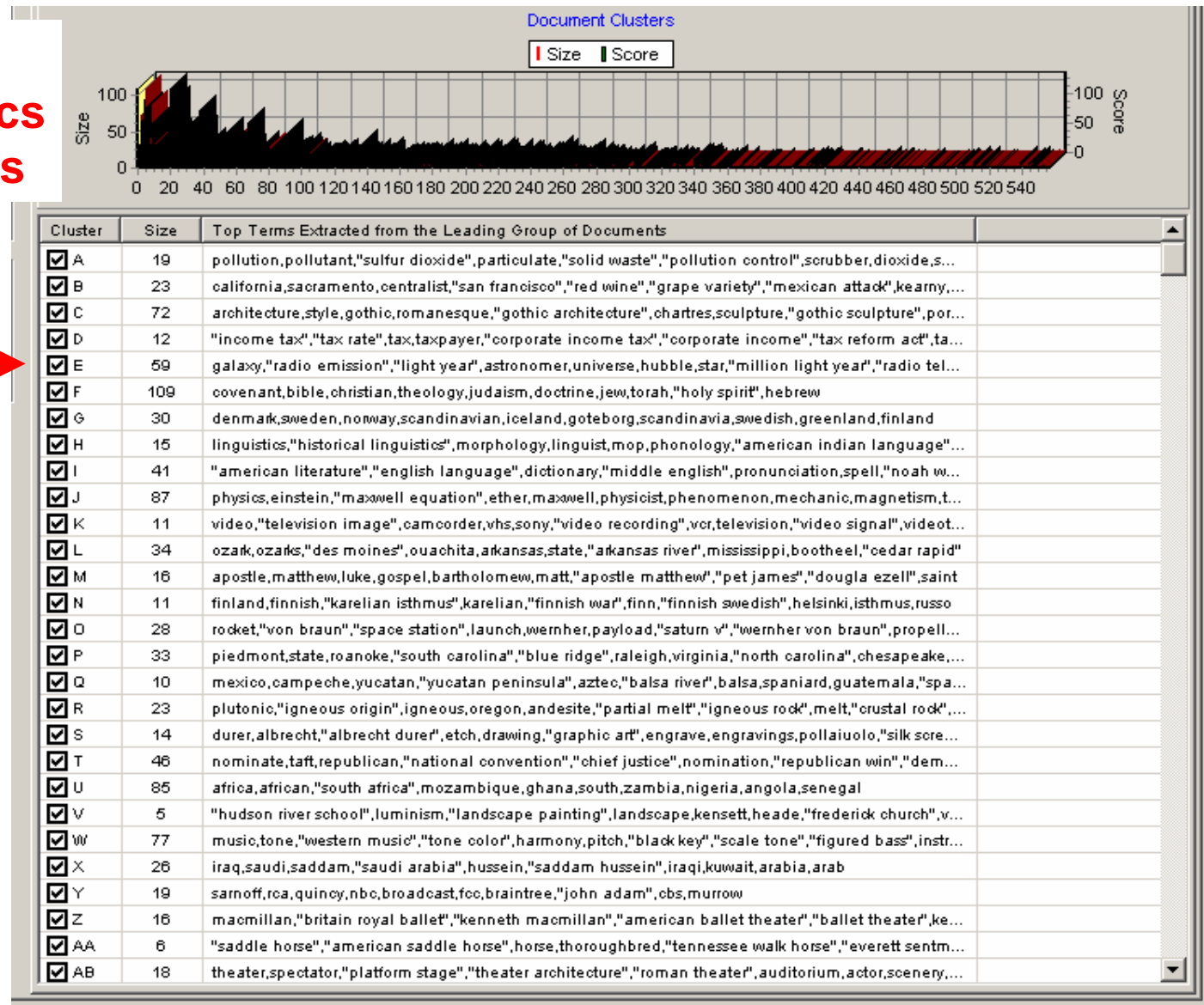
“Hard” / “Soft” Clustering

- **Form “Hard Clusters”—for Precision
(Use a method that clusters documents based on sub-documents)**
- **Make a “Classifier” (“Query”) for each Hard Cluster**
- **Apply Classifiers to the whole DB—
Each Result is a “Soft Cluster”
(Use a method that assigns documents to soft clusters based on sub-documents)**
- **Count Document Overlaps
(and Membership Scores)
across “Soft Clusters”**



Examples of Hard Clusters

**Gold-Standard
DB of 5,174 Docs
in 95 Categories**





Hard Cluster "E"

| Rank | Id | Title | galaxy,"radio emission","light year",astronomer,universe,hubble,star,"million light year","rad... |
|--------------------------|----|--|---|
| <input type="checkbox"/> | 1 | 1840... astronomy and astrophysics | solarsystem&astronomy:A052; 6.762 |
| <input type="checkbox"/> | 2 | 1017... extragalactic systems | solarsystem&astronomy:A052; 5.492 |
| <input type="checkbox"/> | 3 | 2354... radio astronomy | solarsystem&astronomy:A052; 5.001 |
| <input type="checkbox"/> | 4 | 1152... Galaxy, The | solarsystem&astronomy:A052; 4.714 |
| <input type="checkbox"/> | 5 | 7250... cosmology (astronomy) | solarsystem&astronomy:A052; 4.607 |
| <input type="checkbox"/> | 6 | 1493... interstellar matter | solarsystem&astronomy:A052; 3.776 |
| <input type="checkbox"/> | 7 | 2342... quasar | solarsystem&astronomy:A052; 3.588 |
| <input type="checkbox"/> | 8 | 1787... Magellanic Clouds | solarsystem&astronomy:A052; 3.186 |
| <input type="checkbox"/> | 9 | 3078... X-ray astronomy | solarsystem&astronomy:A052; 3.179 |
| <input type="checkbox"/> | 10 | 2354... radio galaxies | solarsystem&astronomy:A052; 3.023 |
| <input type="checkbox"/> | 11 | 2895... ultraviolet astronomy | solarsystem&astronomy:A052; 2.949 |
| <input type="checkbox"/> | 12 | 8609... distance, astronomical | solarsystem&astronomy:A052; 2.838 |
| <input type="checkbox"/> | 13 | 2676... star | solarsystem&astronomy:A052; 2.508 |
| <input type="checkbox"/> | 14 | 7715... Cygnus A | solarsystem&astronomy:A052; 2.434 |
| <input type="checkbox"/> | 15 | 1839... astronomy, history of | solarsystem&astronomy:A052; 2.339 |
| <input type="checkbox"/> | 16 | 1160... gamma-ray astronomy | solarsystem&astronomy:A052; 2.338 |
| <input type="checkbox"/> | 17 | 2155... Palomar Observatory | solarsystem&astronomy:A052; 2.197 |
| <input type="checkbox"/> | 18 | 2632... solar system | solarsystem&astronomy:A052; 2.083 |
| <input type="checkbox"/> | 19 | 2089... observatory, astronomical | solarsystem&astronomy:A052; 2.064 |
| <input type="checkbox"/> | 20 | 2385... relativity | physics&physicists:A051; 1.935 |
| <input type="checkbox"/> | 21 | 1828... astrochemistry | solarsystem&astronomy:A052; 1.893 |
| <input type="checkbox"/> | 22 | 2725... supernova | solarsystem&astronomy:A052; 1.853 |
| <input type="checkbox"/> | 23 | 7373... Crab nebula | solarsystem&astronomy:A052; 1.851 |
| <input type="checkbox"/> | 24 | 2566... Shapley, Harlow | astronomers:A050; 1.846 |
| <input type="checkbox"/> | 25 | 1790... magnitude | solarsystem&astronomy:A052; 1.794 |
| <input type="checkbox"/> | 26 | 1433... Humason, Milton La Salle | astronomers:A050; 1.738 |
| <input type="checkbox"/> | 27 | 2080... Nuffield Radio Astronomy Laboratories | solarsystem&astronomy:A052; 1.712 |
| <input type="checkbox"/> | 28 | 1478... inflationary theory (astronomy) | solarsystem&astronomy:A052; 1.694 |
| <input type="checkbox"/> | 29 | 1479... infrared astronomy | solarsystem&astronomy:A052; 1.601 |
| <input type="checkbox"/> | 30 | 2086... OAO | solarsystem&astronomy:A052; 1.578 |
| <input type="checkbox"/> | 31 | 2907... universe | solarsystem&astronomy:A052; 1.512 |
| <input type="checkbox"/> | 32 | 2013... National Geographic Society-Palomar Obs... | solarsystem&astronomy:A052; 1.352 |
| <input type="checkbox"/> | 33 | 1373... Herschel, Sir William | astronomers:A050; 1.352 |
| <input type="checkbox"/> | 34 | 2777... telescope, optical | solarsystem&astronomy:A052; 1.325 |
| <input type="checkbox"/> | 35 | 1381... aperture synthesis | solarsystem&astronomy:A052; 1.294 |
| <input type="checkbox"/> | 36 | 1980... Mullard Radio Astronomy Observatory | solarsystem&astronomy:A052; 1.267 |
| <input type="checkbox"/> | 37 | 2611... Slipher, Vesto Melvin | astronomers:A050; 1.266 |
| <input type="checkbox"/> | 38 | 2722... Sun | solarsystem&astronomy:A052; 1.202 |
| <input type="checkbox"/> | 39 | 1381... High Energy Astronomical Observatory | solarsystem&astronomy:A052; 1.181 |
| <input type="checkbox"/> | 40 | 2653... Special Astrophysical Observatory | solarsystem&astronomy:A052; 1.169 |
| <input type="checkbox"/> | 41 | 1425... Hoyle, Sir Fred | astronomers:A050; 1.157 |
| <input type="checkbox"/> | 42 | 2352... radar astronomy | solarsystem&astronomy:A052; 1.128 |
| <input type="checkbox"/> | 43 | 1256... gravitation | physics&physicists:A051; 1.125 |
| <input type="checkbox"/> | 44 | 2504... Saturn (planet) | solarsystem&astronomy:A052; 1.099 |
| <input type="checkbox"/> | 45 | 2016... National Radio Astronomy Observatory | solarsystem&astronomy:A052; 1.087 |
| <input type="checkbox"/> | 46 | 2577... Shklovsky, Iosif Samuilovich | astronomers:A050; 1.078 |
| <input type="checkbox"/> | 47 | 2910... Uranus (planet) | solarsystem&astronomy:A052; 1.055 |
| <input type="checkbox"/> | 48 | 1432... Hulst, Hendrik Christoffell van de | astronomers:A050; 0.988 |
| <input type="checkbox"/> | 49 | 2114... Oort, Jan Hendrik | astronomers:A050; 0.965 |



Soft Cluster "E"

Interactive Clustering | Hard Clustering | **Soft Clustering** | Source Generation | Cluster Evaluation | Cluster Log

| Cluster | Size | Top Terms Extracted from the Leading Group of Documents |
|---------|------|--|
| A | 129 | pollution,pollutant,"solid waste","pollution control","sulfur dioxide",particulate,scrubber,"air pollution",dioxide,smog |
| B | 157 | california,centralist,sacramento,"rio grande","san francisco","san diego",texas,"red wine","grape variety","mexican attack" |
| C | 394 | architecture,"gothic architecture",nave,gothic,romanesque,cathedral,sculpture,art,"gothic style",transepts |
| D | 65 | "tax rate",taxpayer,"income tax",tax,taxation,"tax reform act","corporate income tax","personal income tax","tax system","tax burden" |
| E | 222 | galaxy,"light year",luminosity,astronomer,star,quasar,"extragalactic system","red shift","milky way",nebula |
| F | 382 | judaism,jew,hebrew,talmud,testament,"hebrew bible",jesus,christian,torah,"holy spirit" |
| G | 150 | norway,denmark,sweden,scandinavian,"r","swedish,iceland,danish,"scandinavian country",norwegian |
| H | 97 | linguistics,"historical linguistics",linguist,morphology,"american indian language",bloomfield,morphology,language,"european langua... |

Show 100 docs

| | | | | | | | | |
|------------|--------------|-------------|----------|-------------|--------------|-------------|--------------|----|
| galaxy | comet | magnetism | rocket | geochemistr | archaeoastro | astrometry | almagest | ri |
| light year | short period | maxwell equ | launch | mantle | accurate tab | double star | ptolemy | h |
| luminosity | meteor | einstein | spacelab | geologic | astronomica | fundamenta | lunar theory | g |
| astronomer | meteor show | maxwell | shuttle | earth | glyph | fundamenta | epicycles | p |

GH 20 TD 25

astronomy and astrophysics

| | TD/GH | E-34.13 | CB-8.68 | J-5.39 | O-4.61 | BD-2.86 | PA-2.28 | SP-1.97 | BR-1.42 | LJ-1.35 | SH-1.20 |
|--|-------|---------|---------|--------|--------|---------|---------|---------|---------|---------|---------|
| extragalactic systems | | 1.000 | 0.134 | | | | | | | | |
| astronomy and astrophysics | 158/ | 0.980 | 0.261 | 0.221 | 0.280 | 0.061 | | | | 0.077 | |
| Galaxy, The | | 0.964 | 0.106 | | | | | | | | |
| quasar | | 0.915 | | 0.062 | | | | | | | |
| interstellar matter | | 0.901 | 0.083 | 0.066 | 0.083 | | | | | | |
| radio astronomy | 148/ | 0.871 | 0.167 | 0.089 | 0.063 | 0.068 | | | | | |
| cosmology (astronomy) | | 0.866 | 0.133 | 0.294 | | 0.085 | | | | | |
| ultraviolet astronomy | | 0.841 | 0.140 | | 0.095 | | | | | | |
| radio galaxies | | 0.835 | | | | 0.053 | | | | | |
| X-ray astronomy | | 0.832 | 0.117 | 0.083 | 0.131 | | | | | | |
| gamma-ray astronomy | | 0.780 | | 0.113 | 0.158 | | | | | | |
| distance, astronomical | | 0.683 | 0.196 | 0.075 | 0.113 | | | 0.103 | | | |
| Magellanic Clouds | | 0.659 | 0.073 | | | | | | | | |
| Palomar Observatory | | 0.576 | | | | | | | | | |
| Cygnus A | | 0.570 | | | | | | | | | |
| supernova | | 0.536 | 0.075 | | | 0.056 | | | | | |
| star | | 0.530 | 0.140 | 0.126 | 0.055 | | | 0.087 | | | |
| astronomy, history of | | 0.525 | 0.120 | 0.178 | 0.060 | 0.070 | 0.132 | 0.081 | 0.149 | 0.056 | |
| Crab nebula | | 0.474 | | | | | | | | | |
| Shapley, Harlow | | 0.466 | | | | | | | | | |
| astrochemistry | | 0.464 | 0.092 | | | | | | | | |
| High Energy Astronomical Observatory | | 0.434 | | 0.054 | 0.102 | | | | | | |
| OAO | | 0.410 | | | 0.243 | | | | | | |
| National Geographic Society-Palomar O... | | 0.407 | | | | | | | | | |
| relativity | | 0.403 | 0.172 | 0.463 | 0.120 | | | | | | |
| Special Astrophysical Observatory | | 0.401 | | | | | | | | | |
| magnitude | | 0.389 | 0.078 | | | | | | 0.065 | | |
| Mullard Radio Astronomy Observatory | | 0.378 | | | | | | | | | |
| Nuffield Radio Astronomy Laboratories | | 0.376 | 0.161 | | | | | | | | |
| observatory, astronomical | | 0.370 | 0.051 | | 0.174 | | 0.096 | 0.051 | | | |
| Herschel, Sir William | | 0.366 | 0.092 | | | | | | | | |
| Humason, Milton La Salle | | 0.355 | | | | | | | | | |
| Shklovsky, Iosif Samuilovich | | 0.351 | | | | | | | | | |
| telescope, optical | | 0.346 | 0.086 | 0.077 | | 0.052 | | 0.087 | | | |
| infrared astronomy | | 0.346 | 0.148 | | 0.053 | | | | | | |
| Sun | | 0.341 | 0.219 | 0.144 | 0.081 | 0.078 | | | | 0.080 | 0.052 |
| solar system | | 0.341 | 0.381 | 0.120 | 0.229 | 0.190 | 0.066 | | 0.051 | 0.147 | |
| Hulst, Hendrik Christoffell van de | | 0.321 | | | | | | | | | |
| Oort, Jan Hendrik | | 0.315 | 0.160 | | | | | | | | |
| Slipher, Vesto Melvin | | 0.311 | 0.080 | | | | | | | | |
| interferometer | | 0.292 | | 0.077 | | | | | | | |
| Wilson, Robert W. and Bessie Ann A | | 0.288 | | | | | | | | | |



Soft Cluster "E"—Sorted for Hubs

Interactive Clustering | Hard Clustering | **Soft Clustering** | Source Generation | Cluster Evaluation | Cluster Log

| Cluster | Size | Top Terms Extracted from the Leading Group of Documents |
|----------|------|--|
| A | 129 | pollution,pollutant,"solid waste","pollution control","sulfur dioxide",particulate,scrubber,"air pollution",dioxide,smog |
| B | 157 | california,centralist,sacramento,"rio grande","san francisco","san diego",texas,"red wine","grape variety","mexican attack" |
| C | 394 | architecture,"gothic architecture",nave,gothic,romanesque,cathedral,sulpture,art,"gothic style",transepts |
| D | 65 | "tax rate",taxpayer,"income tax",tax,taxation,"tax reform act","corporate income tax","personal income tax","tax system","tax burden" |
| E | 222 | galaxy,"light year",luminosity,astronomer,star,quasar,"extragalactic system","red shift","milky way",nebula |
| F | 382 | judaism,jew,hebrew,talmud,testament,"hebrew bible",jesus,christian,torah,"holy spirit" |
| G | 150 | norway,denmark,sweden,scandinavian,"r .","swedish,iceland,danish,"scandinavian country",norwegian |
| H | 97 | linguistics,"historical linguistics",linguist,morphology,"american indian language","bloomfield,mop.phonology,language,"european langua... |

Show 100 docs

| | | | | | | | | |
|------------|--------------|-------------|----------|-------------|--------------|-------------|--------------|----|
| galaxy | comet | magnetism | rocket | geochemistr | archaeoastro | astrometry | almagest | ri |
| light year | short period | maxwell equ | launch | mantle | accurate tab | double star | ptolemy | h |
| luminosity | meteor | einstein | spacelab | geologic | astronomica | fundamenta | lunar theory | g |
| astronomer | meteor show | maxwell | shuttle | earth | glyph | fundamenta | epicicles | p |

GH 20 TD 25

astronomy and astrophysics

| | TD/GH | E-34.13 | CB-8.68 | J-5.39 | O-4.61 | BD-2.86 | PA-2.28 | SP-1.97 | BR-1.42 | LJ-1.35 | SH-1.20 |
|--|-------|---------|---------|--------|--------|---------|---------|---------|---------|---------|---------|
| astronomy and astrophysics | 158/ | 0.980 | 0.261 | 0.221 | 0.280 | 0.061 | | | | 0.077 | |
| radio astronomy | 148/ | 0.871 | 0.167 | 0.089 | 0.063 | 0.068 | | | | | |
| space programs, national | -/322 | 0.267 | 0.140 | | 0.951 | 0.051 | | | | | |
| space exploration | -/258 | 0.207 | 0.212 | 0.091 | 0.572 | 0.091 | | | | 0.057 | |
| extragalactic systems | | 1.000 | 0.134 | | | | | | | | |
| Galaxy, The | | 0.964 | 0.106 | | | | | | | | |
| quasar | | 0.915 | | 0.062 | | | | | | | |
| interstellar matter | | 0.901 | 0.083 | 0.066 | 0.083 | | | | | | |
| cosmology (astronomy) | | 0.856 | 0.133 | 0.294 | | 0.085 | | | | | |
| ultraviolet astronomy | | 0.841 | 0.140 | | 0.095 | | | | | | |
| radio galaxies | | 0.835 | | | | 0.053 | | | | | |
| X-ray astronomy | | 0.832 | 0.117 | 0.083 | 0.131 | | | | | | |
| gamma-ray astronomy | | 0.760 | | 0.113 | 0.158 | | | | | | |
| distance, astronomical | | 0.683 | 0.196 | 0.075 | 0.113 | | | 0.103 | | | |
| Magellanic Clouds | | 0.659 | 0.073 | | | | | | | | |
| Palomar Observatory | | 0.576 | | | | | | | | | |
| Cygnus A | | 0.570 | | | | | | | | | |
| supernova | | 0.536 | 0.075 | | | 0.056 | | | | | |
| star | | 0.530 | 0.140 | 0.126 | 0.055 | | | 0.087 | | | |
| astronomy, history of | | 0.525 | 0.120 | 0.178 | 0.060 | 0.070 | 0.132 | 0.081 | 0.149 | 0.056 | |
| Crab nebula | | 0.474 | | | | | | | | | |
| Shapley, Harlow | | 0.466 | | | | | | | | | |
| astrochemistry | | 0.464 | 0.092 | | | | | | | | |
| High Energy Astronomical Observatory | | 0.434 | | 0.054 | 0.102 | | | | | | |
| DAO | | 0.410 | | | 0.243 | | | | | | |
| National Geographic Society-Palomar O... | | 0.407 | | | | | | | | | |
| relativity | | 0.403 | 0.172 | 0.453 | 0.120 | | | | | | |
| Special Astrophysical Observatory | | 0.401 | | | | | | | | | |
| magnitude | | 0.389 | 0.078 | | | | | | 0.065 | | |
| Mullard Radio Astronomy Observatory | | 0.378 | | | | | | | | | |
| Nuffield Radio Astronomy Laboratories | | 0.376 | 0.161 | | | | | | | | |
| observatory, astronomical | | 0.370 | 0.051 | | 0.174 | | 0.096 | 0.051 | | | |
| Herschel, Sir William | | 0.366 | 0.092 | | | | | | | | |
| Humason, Milton La Salle | | 0.365 | | | | | | | | | |
| Shklovsky, Iosif Samuilovich | | 0.351 | | | | | | | | | |
| telescope, optical | | 0.346 | 0.086 | 0.077 | | 0.052 | | 0.087 | | | |
| infrared astronomy | | 0.345 | 0.148 | | 0.053 | | | | | | |
| Sun | | 0.341 | 0.219 | 0.144 | 0.081 | 0.078 | | | | 0.080 | 0.052 |
| solar system | | 0.341 | 0.381 | 0.120 | 0.229 | 0.190 | 0.066 | | 0.051 | 0.147 | |
| Hulst, Hendrik Christoffell van de | | 0.321 | | | | | | | | | |
| Dort, Jan Hendrik | | 0.315 | 0.160 | | | | | | | | |
| Singer, Yorta Makin | | 0.311 | 0.080 | | | | | | | | |



Hard Cluster “AB”

| Rank | Id | Title | Type | Score |
|--------------------------|----|--|------------------|-------|
| <input type="checkbox"/> | 1 | 2794... theater, history of the | theater;A074; | 5.950 |
| <input type="checkbox"/> | 2 | 23:00... Abbey Theatre | theater;A074; | 4.957 |
| <input type="checkbox"/> | 3 | 2794... theater architecture and staging | theater;A074; | 3.706 |
| <input type="checkbox"/> | 4 | 1269... Gregory, Isabella Augusta, Lady | theater;A074; | 3.683 |
| <input type="checkbox"/> | 5 | 2742... Synge, John Millington | theater;A074; | 3.589 |
| <input type="checkbox"/> | 6 | 3086... Yeats, William Butler | NobelPrize;A095; | 3.456 |
| <input type="checkbox"/> | 7 | 193:0... acting | theater;A074; | 2.679 |
| <input type="checkbox"/> | 8 | 9590... Elizabethan playhouse | theater;A074; | 2.611 |
| <input type="checkbox"/> | 9 | 1071... Fitzmaurice, George | theater;A074; | 2.289 |
| <input type="checkbox"/> | 10 | 7267... costume and makeup, theatrical | theater;A074; | 2.099 |
| <input type="checkbox"/> | 11 | 1568... Kabuki | theater;A074; | 2.093 |
| <input type="checkbox"/> | 12 | 340:0... Aeschylus | theater;A074; | 2.061 |
| <input type="checkbox"/> | 13 | 1351... Antoine, Andre | theater;A074; | 1.756 |
| <input type="checkbox"/> | 14 | 2509... Saxe-Meiningen, George II, Duke of | theater;A074; | 1.550 |
| <input type="checkbox"/> | 15 | 1840... masque | theater;A074; | 1.383 |
| <input type="checkbox"/> | 16 | 1503... Irving, Sir Henry | theater;A074; | 1.129 |
| <input type="checkbox"/> | 17 | 2123... Oresteia | theater;A074; | 1.000 |
| <input type="checkbox"/> | 18 | 2061... No drama | theater;A074; | 0.987 |



Soft Cluster "AB"

Interactive Clustering | Hard Clustering | **Soft Clustering** | Source Generation | Cluster Evaluation | Cluster Log

| Cluster | Size | Top Terms Extracted from the Leading Group of Documents |
|-----------|------------|--|
| Z | 159 | ballet,"work leg",macmillan,dancer,"support leg","kenneth macmillan",makarova,"pas de","sleep beauty","one foot" |
| AA | 13 | "saddle horse","american saddle horse",horse,"show ring",thoroughbred,gait,"short strong back","showy trot","light strong fast","slow rock ch... |
| AB | 338 | theater,spectator,yeats,"platform stage",actor,"roman theater","theater architecture",auditorium,scenery,"scenic practice" |
| AC | 58 | foal,equus,ass,horse,donkey,equidae,stallion,hoof,"order perissodactyla",mule |
| AD | 153 | "ice age",paleozoic,"recent epoch","glacial stage",glaciation,carboniferous,glacial,permian,pleistocene,quaternary |
| AE | 66 | offender,crime,criminal,criminologist,enforcement,robbery,"law enforcement","law enforcement agency",police,offense |
| AF | 186 | "cape breton",quebec,canadian,"nova scotia",mackenzie,laurier,canada,"st. lawrence river","mackenzie king",province |
| AG | 13 | iuneau.anchorage,sitka,"cook inlet","seward peninsula".kenai."kenai peninsula".alaska.seward.fairbanks |

Show 100 docs

| | | | | | | | | |
|---------------|-------------|--------------|---------------|--------------|-----------|-------------|----------------|---|
| theater | bess | dance form | university wi | meyerhold | sophocles | olivier | freie buhne | c |
| spectator | musical com | dance | admiral mar | stanislavsky | euripides | richard iii | otto brahm | g |
| yeats | vaudeville | theater danc | thomas kyd | vsevolod em | electra | garrick | deutsches th n | |
| platform stad | porqy | folk dance | christopher r | seagull | passion | lear | german thea | c |

GH 20 TD 25

theater, history of the

| | TD/GH | AB-24... | BK-3.90 | BO-2.... | EO-2.... | FF-2.69 | CA-2.07 | FZ-2.00 | IA-1.80 | HA-1.76 | EO-1.... |
|---|-------|----------|---------|----------|----------|---------|---------|---------|---------|---------|----------|
| theater, history of the | 370A | 1.000 | 0.246 | 0.192 | 0.223 | 0.160 | 0.129 | | 0.200 | 0.054 | 0.105 |
| theater architecture and staging | | 0.871 | | 0.064 | | | | | | | |
| Abbey Theatre | | 0.684 | | | | 0.081 | | | | | |
| acting | | 0.636 | 0.053 | 0.054 | | 0.095 | 0.116 | | | 0.107 | 0.108 |
| Elizabethan playhouse | | 0.635 | | | | | | | | | |
| Synge, John Millington | | 0.579 | 0.055 | | | | | | | | |
| Gregory, Isabella Augusta, Lady | | 0.558 | | | | | | | | | |
| Yeats, William Butler | | 0.531 | | | | | | | | | |
| drama | -/512 | 0.440 | 0.151 | 0.103 | 0.408 | 0.059 | 0.250 | | 0.065 | 0.066 | 0.177 |
| Saxe-Meiningen, George II, Duke of | | 0.398 | | | | 0.067 | | | | | |
| Antoine, Andre | | 0.383 | | | | | | | | | |
| Stanislavsky, Konstantin | | 0.345 | | 0.055 | | 1.000 | | | | | |
| Meyerhold, Vsevolod Emilievich | | 0.341 | | | | 0.939 | | | | | |
| costume and makeup, theatrical | | 0.323 | | 0.052 | | | | | | | |
| Fitzmaurice, George | | 0.304 | | | | | | | | | |
| Irving, Sir Henry | | 0.284 | | | | | | | | | |
| Kabuki | | 0.277 | 0.105 | 0.095 | | | | | | | |
| Aeschylus | | 0.272 | | | | | 0.179 | | | | |
| opera houses | | 0.270 | 0.054 | | | | | | | | |
| Group Theatre | | 0.267 | | | | | | | | 1.000 | |
| stage lighting | | 0.260 | | | | | | | | | 0.123 |
| toy theaters | | 0.257 | | | | | | | | | |
| masks | | 0.256 | | | | | | | 0.073 | | |
| masque | | 0.252 | 0.064 | 0.126 | | | | | | | |
| directing | | 0.252 | 0.083 | | | | | | | 0.100 | |
| theatricalism | | 0.250 | | | | | 0.051 | | 0.073 | | |
| commedia dell'arte | | 0.246 | | | | | | | | | |
| Nemirovich-Danchenko, Vladimir Ivanovi... | | 0.241 | | | | 0.119 | | | | | |
| Moscow Art Theater | | 0.236 | | | | 0.170 | | | | | |
| Fortune Theatre | | 0.232 | | | 0.077 | | | | | | |
| Off-Broadway theater | | 0.226 | 0.083 | | | | | | | | |
| epic theater | | 0.225 | | | | | | | | | 0.233 |
| Kean (family) | | 0.219 | | | | | | | | | |
| tragedy | | 0.218 | | | | | 0.170 | | | 0.051 | |
| opera | | 0.216 | 0.392 | 0.092 | | | | | | | |
| Copeau, Jacques | | 0.214 | 0.053 | | | | | | | | |
| Jonson, Ben | | 0.212 | 0.053 | | 0.055 | | | | | | |
| Adler, Jacob, Stella, and Luther | | 0.211 | 0.057 | | | | | | | 0.088 | |
| Actors Studio | | 0.210 | | | | | | | | 0.125 | |
| Moliere | | 0.208 | 0.088 | | | | | | | | |
| Artaud, Antonin | | 0.207 | | 0.102 | | | | | | | |
| Radio City Music Hall | | 0.206 | 0.096 | | | | | | | | |



Soft Cluster "AB"—Sorted for Hubs

Interactive Clustering | Hard Clustering | **Soft Clustering** | Source Generation | Cluster Evaluation | Cluster Log

| Cluster | Size | Top Terms Extracted from the Leading Group of Documents |
|---------|------|--|
| Z | 159 | ballet,"work leg",macmillan,dancer,"support leg","kenneth macmillan",makarova,"pas de","sleep beauty","one foot" |
| AA | 13 | "saddle horse","american saddle horse",horse,"show ring",thoroughbred,gait,"short strong back","showy trot","light strong fast","slow rock ch... |
| AB | 338 | theater,spectator,yeats,"platform stage",actor,"roman theater","theater architecture",auditorium,scenery,"scenic practice" |
| AC | 58 | foal,equus,ass,horse,donkey,equidae,stallion,hooft,"order perissodactyla",mule |
| AD | 153 | "ice age",paleozoic,"recent epoch","glacial stage",glaciation,carboniferous,glacial,permian,pleistocene,quaternary |
| AE | 66 | offender,crime,criminal,criminologist,enforcement,robbery,"law enforcement","law enforcement agency",police,offense |
| AF | 186 | "cape breton",quebec,canadian,"nova scotia",mackenzie,laurier,canada,"st . lawrence river","mackenzie king",province |
| AG | 13 | iuneau.anchorage,sitka,"cook inlet","seward peninsula",kenai,"kenai peninsula",alaska.seward.fairbanks |

Show 100 docs

| | | | | | | | | |
|---------------|-------------|--------------|---------------|--------------|-----------|-------------|----------------|---|
| theater | bess | dance form | university wi | meyerhold | sophocles | olivier | freie buhne | c |
| spectator | musical com | dance | admiral mar | stanislavsky | euripides | richard iii | otto brahm | g |
| yeats | vaudeville | theater danc | thomas kyd | vsevolod em | electra | garrick | deutsches th n | |
| platform stad | porqy | folk dance | christopher r | seagull | passion | lear | german thea | c |

GH 20 TD 25

theater, history of the

| | TD/GH | AB-24... | BK-3.90 | BO-2.... | EO-2.... | FF-2.69 | CA-2.07 | FZ-2.00 | IA-1.80 | HA-1.76 | EO-1.... |
|---|--------|----------|---------|----------|----------|---------|---------|---------|---------|---------|----------|
| theater, history of the | 370/- | 1.000 | 0.246 | 0.192 | 0.223 | 0.129 | | | 0.200 | 0.054 | 0.105 |
| drama | -.512 | 0.440 | 0.151 | 0.103 | 0.408 | 0.059 | 0.250 | | 0.065 | 0.066 | 0.177 |
| dance | -.7405 | 0.200 | 0.180 | 1.000 | | | 0.076 | | | | |
| theater architecture and staging | | 0.871 | | 0.064 | | | | | | | |
| Abbey Theatre | | 0.684 | | | | 0.081 | | | | | |
| acting | | 0.636 | 0.053 | 0.054 | | 0.095 | 0.116 | | | 0.107 | 0.108 |
| Elizabethan playhouse | | 0.635 | | | | | | | | | |
| Synge, John Millington | | 0.579 | 0.055 | | | | | | | | |
| Gregory, Isabella Augusta, Lady | | 0.568 | | | | | | | | | |
| Yeats, William Butler | | 0.531 | | | | | | | | | |
| Saxe-Meiningen, George II, Duke of | | 0.398 | | | | 0.067 | | | | | |
| Antoine, Andre | | 0.383 | | | | | | | | | |
| Stanislavsky, Konstantin | | 0.345 | | 0.055 | | 1.000 | | | | | |
| Meyerhold, Vsevolod Emilievich | | 0.341 | | | | 0.939 | | | | | |
| costume and makeup, theatrical | | 0.323 | | 0.052 | | | | | | | |
| Fitzmaurice, George | | 0.304 | | | | | | | | | |
| Irving, Sir Henry | | 0.284 | | | | | | | | | |
| Kabuki | | 0.277 | 0.105 | 0.095 | | | | | | | |
| Aeschylus | | 0.272 | | | | | 0.179 | | | | |
| opera houses | | 0.270 | 0.054 | | | | | | | | |
| Group Theatre | | 0.267 | | | | | | | | 1.000 | |
| stage lighting | | 0.260 | | | | | | | | | 0.123 |
| toy theaters | | 0.257 | | | | | | | | | |
| masks | | 0.256 | | | | | | | 0.073 | | |
| masque | | 0.252 | 0.064 | 0.126 | | | | | | | |
| directing | | 0.252 | 0.083 | | | | | | | 0.100 | |
| theatricalism | | 0.250 | | | | | 0.051 | | 0.073 | | |
| commedia dell'arte | | 0.246 | | | | | | | | | |
| Nemirovich-Danchenko, Vladimir Ivanovi... | | 0.241 | | | | 0.119 | | | | | |
| Moscow Art Theater | | 0.236 | | | | 0.170 | | | | | |
| Fortune Theatre | | 0.232 | | | 0.077 | | | | | | |
| Off-Broadway theater | | 0.226 | 0.083 | | | | | | | | |
| epic theater | | 0.225 | | | | | | | | | 0.233 |
| Kean (family) | | 0.219 | | | | | | | | | |
| tragedy | | 0.218 | | | | | 0.170 | | | 0.051 | |
| opera | | 0.216 | 0.392 | 0.092 | | | | | | | |
| Copeau, Jacques | | 0.214 | 0.053 | | | | | | | | |
| Jonson, Ben | | 0.212 | 0.053 | | 0.055 | | | | | | |
| Adler, Jacob, Stella, and Luther | | 0.211 | 0.057 | | | | | | | 0.088 | |
| Actors Studio | | 0.210 | | | | | | | | 0.125 | |
| Moliere | | 0.208 | 0.088 | | | | | | | | |
| Adrian, Antonio | | 0.207 | | 0.102 | | | | | | | |



Hard Cluster “DW”

| Rank | Id | Title | | |
|----------------------------|------|-----------------------|--------------|-------|
| <input type="checkbox"/> 1 | 198 | butterflies and moths | insect:A004; | 6.000 |
| <input type="checkbox"/> 2 | 681 | Lepidoptera | insect:A004; | 3.143 |
| <input type="checkbox"/> 3 | 225 | caterpillar | insect:A004; | 2.499 |
| <input type="checkbox"/> 4 | 780 | mosquito (fly) | insect:A004; | 1.534 |
| <input type="checkbox"/> 5 | 960 | pupa | insect:A004; | 1.459 |
| <input type="checkbox"/> 6 | 1190 | tent caterpillar | insect:A004; | 1.098 |



Soft Cluster "DW"

Interactive Clustering | Hard Clustering | **Soft Clustering** | Source Generation | Cluster Evaluation | Cluster Log

| Cluster | Size | Top Terms Extracted from the Leading Group of Documents |
|---------|------|---|
| DU | 4 | "childhood disease", childhood, measles, disease, "developmental age process", "central nervous system damage", "eliminate worldwide", "d... |
| DV | 56 | spouse, marriage, bride, polygamy, polygyny, household, custom, "marriage contract", "domestic work", "wedding ceremony" |
| DW | 69 | moth, butterfly, lepidoptera, caterpillar, metamorphosis, larva, "sensory palpi", skipper, "bear spine", "second large insect order" |
| DX | 4 | triggerfish, "unusual trigger", "third spine", "lock mechanism", "second spine", "various colored bony scale", "spinous top fin", "strong tooth", "d... |
| DY | 18 | muskellunge, esox, "northern pike", esocidae, pickerel, pike, stizostedion, walleye, pickerels, "european perch" |
| DZ | 19 | "ramses ii", ramses, "r . bc", "ramses pursue", "prestige spaik revolt", "direct hostility", "final conclude", "trans jordan", "western coast road", "liby... |
| EA | 46 | homoptera, whitefly, "scale insect", strepsiptera, "minute insect", gnatlike, "suck sap", "serious pest", mealybugs, "greenhouse plant" |
| EB | 12 | "extend family", "nuclear family", "james lowell", "gibbs junior", "traditional prohibit", "clear institutionalize", "primary relative", "mother wife".... |

Show 100 docs

| | | | | | | | | |
|-------------|----------|------------|--------------|----------------|--------------|-------------|------------|---|
| moth | bee | salamander | tree species | shrimp | homoptera | ant lion | parakeet | c |
| butterfly | honeybee | caecilians | beetle | terrestrial wo | whitefly | dobsonfly | parrot | n |
| lepidoptera | ant | frog | meristematid | uropods | scale insect | odonata | maggot | b |
| caterpillar | nest | amphibian | bark | crustacean | strepsiptera | damselflies | body louse | c |

GH 20 TD 25

butterflies and moths

| | TD/GH | DW-7.... | DK-7.27 | CY-5.75 | BF-3.95 | FY-3.48 | EA-3.42 | JS-2.66 | BY-1.99 | AM-1.... | BW-1.... |
|-----------------------|----------|----------|---------|---------|---------|---------|---------|---------|---------|----------|----------|
| butterflies and moths | 178/1... | 1.000 | 0.133 | 0.101 | 0.148 | 0.064 | 0.055 | 0.072 | 0.068 | 0.067 | 0.123 |
| Lepidoptera | | 0.564 | 0.076 | 0.080 | | | | | | | |
| caterpillar | | 0.270 | | | | | | 0.070 | | | |
| fly | -/195 | 0.262 | 0.108 | 0.118 | 0.063 | | 0.099 | 0.104 | 0.816 | 0.093 | |
| insect | -/174 | 0.246 | 0.442 | 0.105 | 0.123 | 0.061 | 0.118 | 0.127 | 0.118 | 0.231 | 0.086 |
| pupa | | 0.232 | 0.099 | 0.101 | 0.067 | | | 0.055 | 0.076 | | |
| wasp | | 0.188 | 0.613 | 0.090 | 0.128 | | | 0.075 | | | |
| larva | | 0.184 | 0.094 | 0.471 | | | | | 0.091 | | |
| mosquito (fly) | | 0.170 | 0.157 | 0.142 | 0.054 | 0.050 | 0.070 | 0.095 | | | |
| tent caterpillar | | 0.147 | 0.061 | 0.076 | | | | | | | |
| beetle | | 0.145 | 0.077 | 0.068 | 0.638 | 0.054 | 0.106 | | 0.054 | | |
| cutworm | | 0.133 | 0.054 | 0.066 | 0.063 | | | | | | |
| arthropod | | 0.112 | 0.079 | 0.129 | 0.086 | 0.125 | 0.060 | 0.110 | | 0.212 | 0.387 |
| Orthoptera | | 0.111 | | | | 0.057 | 0.052 | | | | 0.052 |
| grasshopper | | 0.109 | 0.090 | 0.114 | | | 0.110 | 0.068 | 0.074 | | |
| mite | | 0.102 | 0.064 | | | | | | 0.073 | | 0.142 |
| worm | | 0.098 | | | | | | | | 0.053 | |
| whitefly | | 0.094 | 0.085 | | | | 0.933 | | 0.055 | | |
| bagworm | | 0.094 | | | 0.083 | | | | | | |
| gypsy moth | | 0.093 | 0.112 | | 0.086 | | | | | | |
| Hemiptera | | 0.093 | 0.063 | | 0.063 | | 0.119 | | | | |
| inchworm | | 0.093 | 0.063 | | 0.054 | | | | | | |
| flight | | 0.091 | | | | | | | | | |
| barnacle | | 0.090 | 0.054 | 0.117 | | 0.112 | | | | 0.055 | 0.130 |
| silkworm | | 0.087 | 0.069 | | | | | | | | |
| bee | | 0.085 | 1.000 | 0.107 | 0.091 | | | | | | 0.055 |
| termite | | 0.084 | 0.775 | | 0.058 | | 0.069 | 0.096 | 0.054 | 0.068 | |
| crustacean | | 0.084 | 0.096 | | | 0.936 | | 0.060 | | 0.188 | 0.095 |
| sawfly | | 0.083 | 0.323 | | 0.175 | | | | | | |
| shrimp | | 0.082 | | | | 1.000 | | | | | 0.136 |
| sea moth | | 0.082 | 0.054 | | | | | | | | |
| flea | | 0.082 | 0.084 | 0.128 | | | 0.066 | | 0.071 | | |
| pollination | | 0.076 | 0.136 | | 0.142 | | | | | | |
| ant | | 0.074 | 0.755 | 0.084 | 0.083 | | 0.153 | 0.101 | 0.062 | | |
| frog | | 0.073 | 0.098 | 0.730 | | | | | | | |
| mantisfly | | 0.072 | | 0.091 | | | 0.066 | 0.138 | | | 0.072 |
| krill | | 0.070 | | | | 0.128 | | | | | |
| butterflyfish | | 0.067 | | | | | | | | | |
| butterfly fish | | 0.067 | | | | | | | | | |
| tsetse fly | | 0.067 | 0.069 | | | | | | | | |
| thrips | | 0.067 | 0.060 | | | | 0.099 | | | | |
| terevil | | 0.066 | 0.058 | 0.050 | 0.082 | | | | 0.058 | | |



Soft Cluster “DW”—Sorted for Hubs

Interactive Clustering | Hard Clustering | **Soft Clustering** | Source Generation | Cluster Evaluation | Cluster Log

| Cluster | Size | Top Terms Extracted from the Leading Group of Documents |
|---------|------|---|
| DU | 4 | "childhood disease", childhood, measles, disease, "developmental age process", "central nervous system damage", "eliminate worldwide", "d... |
| DV | 56 | spouse, marriage, bride, polygamy, polygyny, household, custom, "marriage contract", "domestic work", "wedding ceremony" |
| DW | 69 | moth, butterfly, lepidoptera, caterpillar, metamorphosis, larva, "sensory palpi", skipper, "bear spine", "second large insect order" |
| DX | 4 | triggerfish, "unusual trigger", "third spine", "lock mechanism", "second spine", "various colored bony scale", "spinous top fin", "strong tooth", "d... |
| DY | 18 | muskellunge, esox, "northern pike", esocidae, pickerel, pike, stizostedion, walleye, pickerels, "european perch" |
| DZ | 19 | "ramses ii", ramses, "r . bc", "ramses pursue", "prestige spaik revolt", "direct hostility", "final conclude", "trans jordan", "western coast road", "liby... |
| EA | 46 | homoptera, whitefly, "scale insect", strepsiptera, "minute insect", gnatlike, "suck sap", "serious pest", mealybugs, "greenhouse plant" |
| EB | 12 | "extend family", "nuclear family", "james lowell", "gibbs iunior", "traditional prohibit", "clear institutionalize", "primary relative", "mother wife".... |

Show 100 docs

| | | | | | | | | |
|-------------|----------|------------|--------------|----------------|--------------|-------------|------------|---|
| moth | bee | salamander | tree species | shrimp | homoptera | ant lion | parakeet | c |
| butterfly | honeybee | caecilians | beetle | terrestrial wo | whitefly | dobsonfly | parrot | n |
| lepidoptera | ant | frog | meristematid | uropods | scale insect | odonata | maggot | b |
| caterpillar | nest | amphibian | bark | crustacean | strepsiptera | damselflies | body louse | c |

GH 20 TD 25

butterflies and moths

| | TD/GH | DW-7.... | DK-7.27 | CY-5.75 | BF-3.95 | FY-3.48 | EA-3.42 | JS-2.66 | BY-1.99 | AM-1.... | BW-1.... |
|-----------------------|----------|----------|---------|---------|---------|---------|---------|---------|---------|----------|----------|
| butterflies and moths | 178/1... | 1.000 | 0.133 | 0.101 | 0.148 | 0.064 | 0.055 | 0.072 | 0.068 | 0.067 | 0.123 |
| invertebrate | -/222 | 0.051 | 0.212 | 0.118 | 0.056 | 0.053 | | | | 0.246 | 0.098 |
| fly | -/195 | 0.262 | 0.108 | 0.118 | 0.063 | | 0.099 | 0.104 | 0.816 | 0.093 | |
| insect | -/174 | 0.246 | 0.442 | 0.105 | 0.123 | 0.061 | 0.118 | 0.127 | 0.118 | 0.231 | 0.086 |
| amphibians | -/169 | 0.057 | 0.144 | 1.000 | 0.105 | | | 0.058 | | 0.138 | |
| Lepidoptera | | 0.564 | 0.076 | 0.080 | | | | | | | |
| caterpillar | | 0.270 | | | | | | 0.070 | | | |
| pupa | | 0.232 | 0.099 | 0.101 | 0.067 | | | 0.055 | 0.076 | | |
| wasp | | 0.188 | 0.613 | 0.090 | 0.128 | | | 0.075 | | | |
| larva | | 0.184 | 0.094 | 0.471 | | | | | 0.091 | | |
| mosquito (fly) | | 0.170 | 0.157 | 0.142 | 0.054 | 0.050 | 0.070 | 0.095 | | | |
| tent caterpillar | | 0.147 | 0.061 | 0.076 | | | | | | | |
| beetle | | 0.146 | 0.077 | 0.068 | 0.638 | 0.054 | 0.106 | | 0.054 | | |
| cutworm | | 0.133 | 0.054 | 0.066 | 0.063 | | | | | | |
| arthropod | | 0.112 | 0.079 | 0.129 | 0.086 | 0.125 | 0.080 | 0.110 | | 0.212 | 0.387 |
| Orthoptera | | 0.111 | | | | 0.057 | 0.052 | | | | 0.052 |
| grasshopper | | 0.109 | 0.090 | 0.114 | | | 0.110 | 0.068 | 0.074 | | |
| mite | | 0.102 | 0.064 | | | | | | 0.073 | | 0.142 |
| worm | | 0.098 | | | | | | | | 0.053 | |
| whitefly | | 0.094 | 0.085 | | | | 0.933 | | 0.055 | | |
| bagworm | | 0.094 | | | 0.083 | | | | | | |
| gypsy moth | | 0.093 | 0.112 | | 0.086 | | | | | | |
| Hemiptera | | 0.093 | 0.063 | | 0.063 | | 0.119 | | | | |
| inchworm | | 0.093 | 0.063 | | 0.054 | | | | | | |
| flight | | 0.091 | | | | | | | | | |
| barnacle | | 0.090 | 0.054 | 0.117 | | 0.112 | | | 0.055 | 0.130 | |
| silkworm | | 0.087 | 0.069 | | | | | | | | |
| bee | | 0.085 | 1.000 | 0.107 | 0.091 | | | | | | 0.055 |
| termite | | 0.084 | 0.775 | | 0.058 | | 0.089 | 0.096 | 0.054 | 0.068 | |
| crustacean | | 0.084 | 0.096 | | | 0.936 | | 0.060 | | 0.188 | 0.095 |
| sawfly | | 0.083 | 0.323 | | 0.175 | | | | | | |
| shrimp | | 0.082 | | | | 1.000 | | | | | 0.136 |
| sea moth | | 0.082 | 0.054 | | | | | | | | |
| flea | | 0.082 | 0.084 | 0.128 | | | 0.066 | | 0.071 | | |
| pollination | | 0.076 | 0.136 | | 0.142 | | | | | | |
| ant | | 0.074 | 0.755 | 0.084 | 0.083 | | 0.153 | 0.101 | 0.062 | | |
| frog | | 0.073 | 0.098 | 0.730 | | | | | | | |
| mantisfly | | 0.072 | | 0.091 | | | 0.066 | 0.138 | | | 0.072 |
| krill | | 0.070 | | | | 0.128 | | | | | |
| butterflyfish | | 0.067 | | | | | | | | | |
| butterfly fish | | 0.067 | | | | | | | | | |
| tree fly | | 0.067 | 0.080 | | | | | | | | |

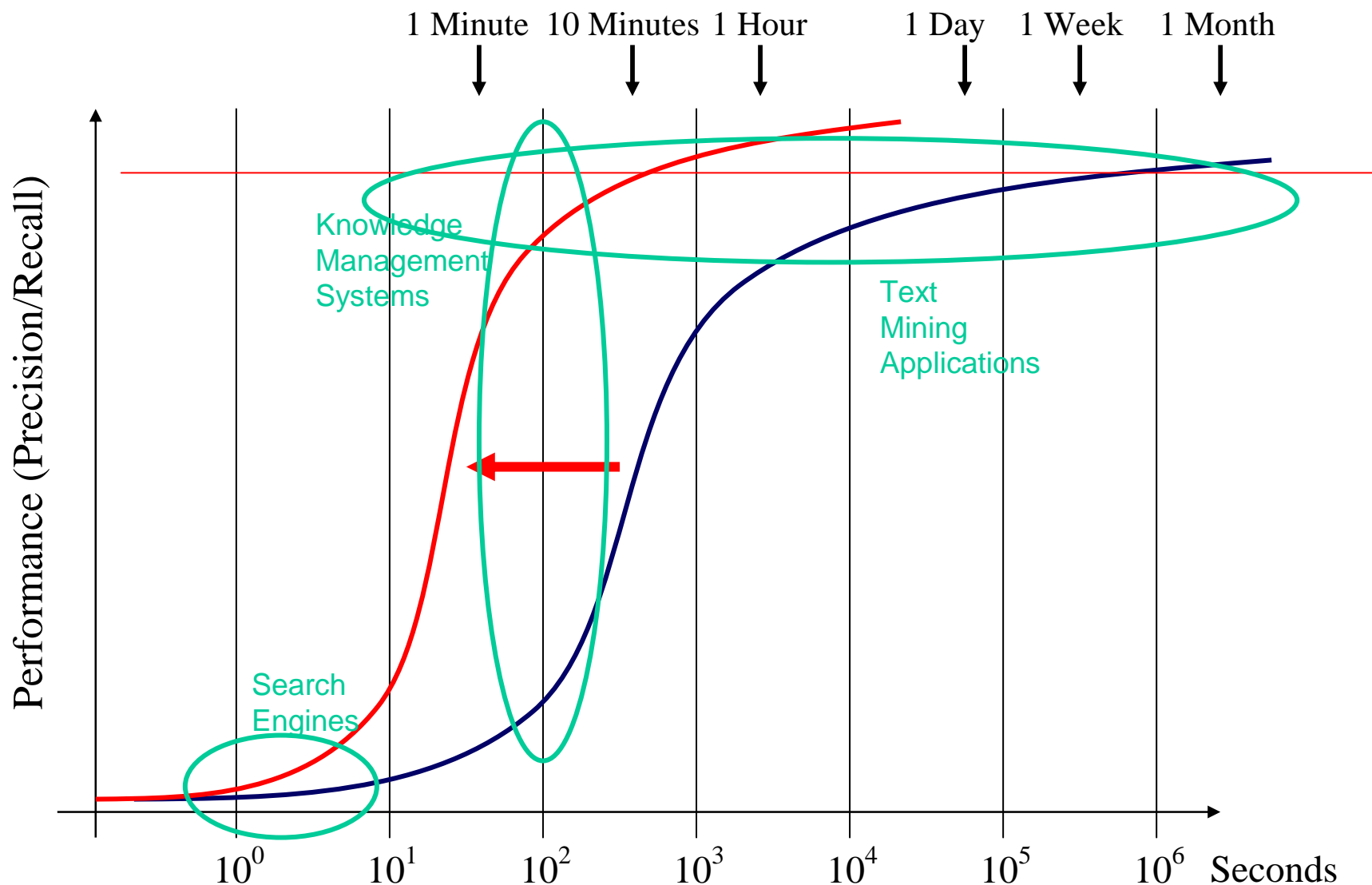


Conclusions



How Much 'Quality' is Possible?

Human Effort vs. Retrieval Performance





Conclusions

- **State-of-the-Art IR is quite good and significant improvements are hard to find.**
- **A remarkably effective technique—PRF—may be made stronger by clustering documents to concentrate relevants.**
- **The new challenge is to pick the “right” clusters automatically...**
- **New applications of clustering can also reveal document relations—making it possible discover document status in unstructured collections.**
- **These techniques have implications for IR, for user interface design (and user interaction), for text mining (e.g., patent collections), and for automating the discovery of latent semantics.**



The End