In Search of Disruptive Ideas:

A Survey for Outlier Detection Techniques in Crowdsourcing Innovation Platforms

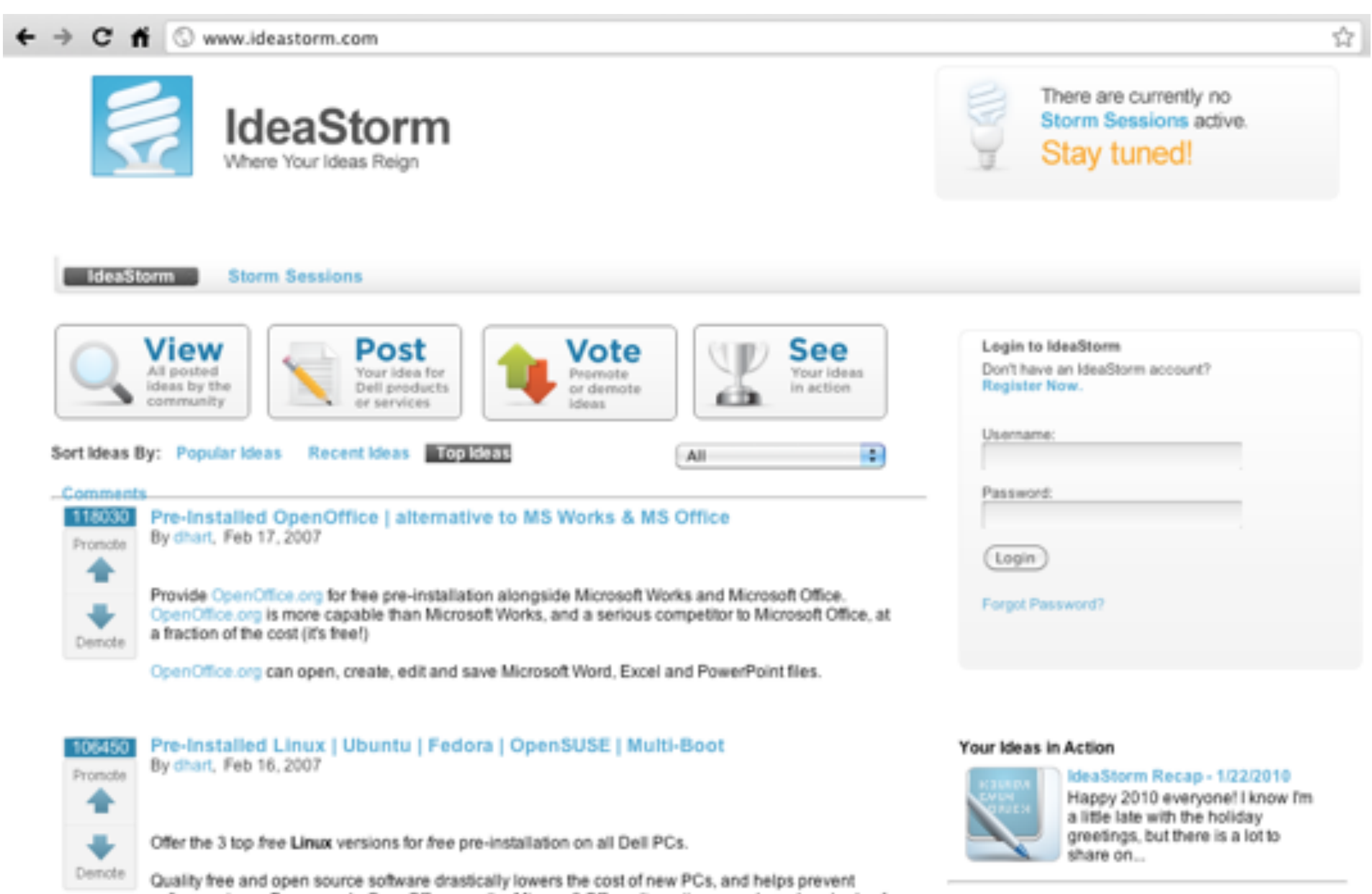
GI2MO
Semantically Empowered Idea Management

A*STAR
Institute for Infocomm Research

Overview

[research context / application area]

Idea Management Systems = online collaborative tool to **collect ideas from many people** (e.g. clients of a company)



Overview

[research context / application area]

Idea Management Systems problems

- Lots of contributions
- Lots of duplicates, similar ideas etc.
- Lots of simple or obvious input

difficult to choose the best innovations

Overview

[problem - hypothesis - approach]

- **problem:** pick best candidates for interesting/ disruptive innovations
- **hypothesis:** good ideas are rare outliers that stand out from the majority of other proposals
- **approach:** use outlier detection algorithms on idea text to detect the most anomalous ideas

what is disruptive innovation / how to find it ?

[hypothesis theoretical grounding]

innovation literature

disruptive
origin

“ **disruptors** deliver innovations for overlooked market segments , while **market leaders** address their most demanding customers via incremental innovation “

["What is disruptive innovation?" Christensen, Raynor, Harvard Business Review, 2015]

disruptive
vs. success

“ **not all disruptive ideas have to lead to success** “

["What is disruptive innovation?" Christensen, Raynor, Harvard Business Review, 2015]

IMS hypothesis

standard criteria/ metrics of IMS:

- **favour success as perceived by** the entrenched **market leader** point of view
- **could overlook disruptive ideas** that get **no popular support**
- metrics bring up ideas of **most vocal customers**
- less-demanding customers are less vocal and not equally participatory in IMS, (*ie. low-end foothold; or non existent customers ie. new market-footholds [Harvard Review]*)

Approach: creating metric for disruptiveness of ideas

[how to find and evaluate the best outlier detection]

1. **survey** available outlier detection algorithms
2. **pick** the most representative candidates based on previous applications
3. **apply + eval** algorithms for idea management problem using two different public datasets (alg picks vs. manual annotation)
4. **compare** results of different algorithms
5. **recommend** the best approach

Work so far

[related / past work]

Idea Management Systems (key area)

IMS + metrics
["Idem: a prediction market for idea management" Bothos et al.,2008]
["Assessing the management of innovation with software tools" S.J. Conn,2009]
["Semantic innovation management across the extended enterprise," K. Ning, 2006]

different approaches to rating ideas:
- *stock market imitation*
- *based on supporting enterprise data*
- *innovation theory metrics*

IMS + winning ideas
["A review of technologies for open innovatIon" Hrastinski et al.,2010]
["Steal my idea! Organizational adoption of user innovations" Gangi et al., 2009]
["An Ontology-based Co-creation Enhancing System for Idea Recommendation"K. Choi et al., 2015]

studies on selecting ideas based on legacy metrics

IMS + disruptive ideas
NONE

IMS + outlier detection
NONE

Outlier Detection + text (subarea)

First (news) Story Detection/ TDT (98 - 04)
(novelty detection in stream) *["news filtering, topic detection and tracking" J. Allen et al.,2004]*

TF-IDF + <approach x>

new techniques
["Unified Analysis of Streaming News" A. Ahmed et al.,2011]
- *stochastic approach*
- *.....*

new domains
["Streaming First Story [...] application to Twitter" S. Petrovic et al.,2010]

narrow down problem
- *short text only (e.g. twitter novelty)*
- *specific domain (e.g. terrorist threats)*

other media
- *voice*
- *video*

["Video scene detection using closed caption text" G. Smith et al.,2009]

Innovation Management (subarea)

Theoretical models
["Innovation management" A. Afuah, 1998]

Case studies in enterprises

- ***rich and mature area***
- ***mostly theoretical***
- ***applied approaches are based on business studies***

Survey + alg picks

[different taxonomies / evaluations]

multitude of algorithms across years

many surveys / taxonomies to classify SoA

CLASSIFICATION [Ji Zhang, 2013]
1. Statistical (probabilistic)
2. Distance (proximity) based
3. Density based
4. Clustering
5. High dimensional

CLASSICATION [Aggarwal, 2013]
1. Extreme Value Analysis
2. Probabilistic and statistical
3. Linear
4. Proximity based
4.1. Clustering
4.2. Density
4.3. Nearest neighbour
5. Information Theory based
6. High dimensional

CLASSICATION [Chandola, 2008]
1. Classification Based
2. Clustering Based
3. Nearest Neighbour Based
3.1. kNN
3.2. Density
4. Statistical
5. Information Theoretic
6. Spectral

choose the categories that repeat across surveys
pick one algorithm per each category to evaluate

Evaluated Algorithms

[evaluation outline]

1. **Distance Based:** kNN (k Nearest Neighbours)
    1. Feature vector generation:
        1. TF-IDF
        2. WORD2VEC
        3. LDA / VEM
        4. LDA / Gibbs
    2. Distance measures:
        1. Cosine
        2. Manhattan
        3. Euclidean
2. **Probabilistic / statistical:** LDA (Latent Dirichlet Allocation)
3. **Density Based:** LOF (Local Outlier Factor)
    1. Distance measures: Cosine, Manhattan, Euclidean
4. **Clustering:** kMeans / kMedoids
    1. Distance Measures: Cosine, Manhattan, Euclidean

Evaluation datasets

[two different scenarios]

1. **Dell IdeaStorm:**
   1. **Ideas:** new equipment, software for PC manufacturer business
   2. **Innovators:** customers
   3. **Stats:**
      - 15,000 ideas (207 implemented)
      - 2,000 users

2. **Starbucks Ideas:**
   1. **Ideas:** new drinks, food, changes in offering for coffee chain
   2. **Innovators:** customers / store owners
   3. **Stats:**
      - 10,000 ideas (1069 implemented)
      - 3,000 users

Evaluation - dataset labels

[manual annotation]

Idea title

Idea textual desc

manual annotation →

3 innovation metrics:

- Implementation cost
- Potential profit
- Market size

1 overall rating

- Breakthrough

based on **innovation management** theory

1 - 10 Likert scale

Which ideas to annotate ?

legacy metrics ranking

- Vote count
  - 10 top
  - 10 middle
  - 10 bottom
- Comment count
  - 10 top
  - 10 middle
  - 10 bottom
- 10 implemented (random pick)
- 10 unimplemented (random pick)

outlier metrics ranking

- 10 top for every algorithm / configuration tested

~1000 ideas annotated per dataset

Evaluation metrics

[assessment of results quality]

What makes a good ranking ?

correlation with manual eval results

shows if the overall ordering reflects the expected one (ie 1,2,3,4... 5000 etc. if idea count = 5000 , as ranked by breakthrough rating)

precision@10 vs. manual ranking

shows how well the outlier ranking works for the top outliers (most important ones for organization stand point)

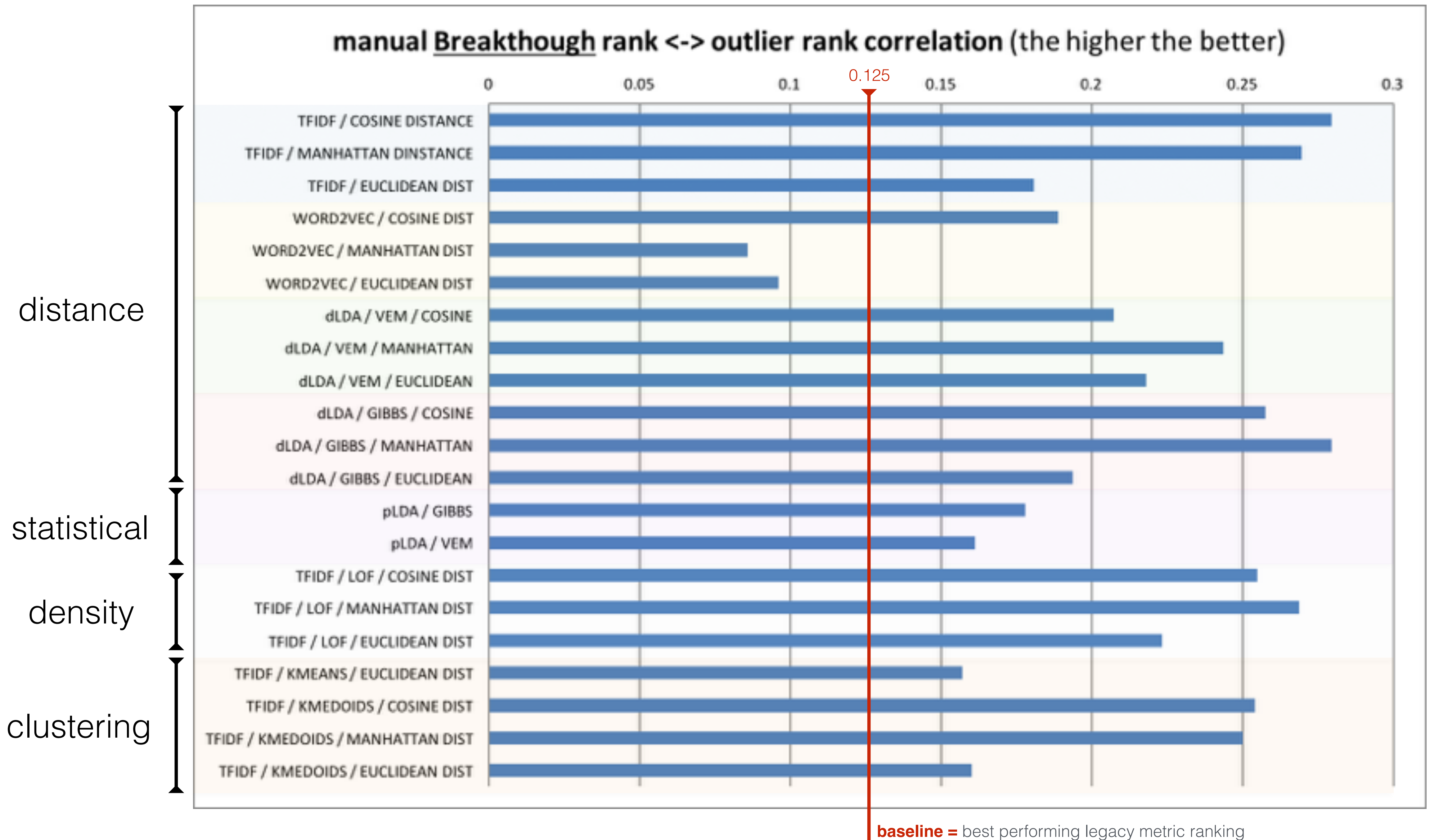
*[**Evaluating Recommendation Systems,** Guy Shani and Asela Gunawardana**, Microsoft Research,** 2009]*

additional extended analysis

- [distance/density] comparison of effectiveness for different neighbourhood settings
- [probabilistic] comparison for different topic optimisation settings
- [clustering] comparison for different cluster sizes / iterations / feature vectors

Results (IdeaStorm)

[correlation of algorithm rankings vs. manual picks]



BEST RESULT

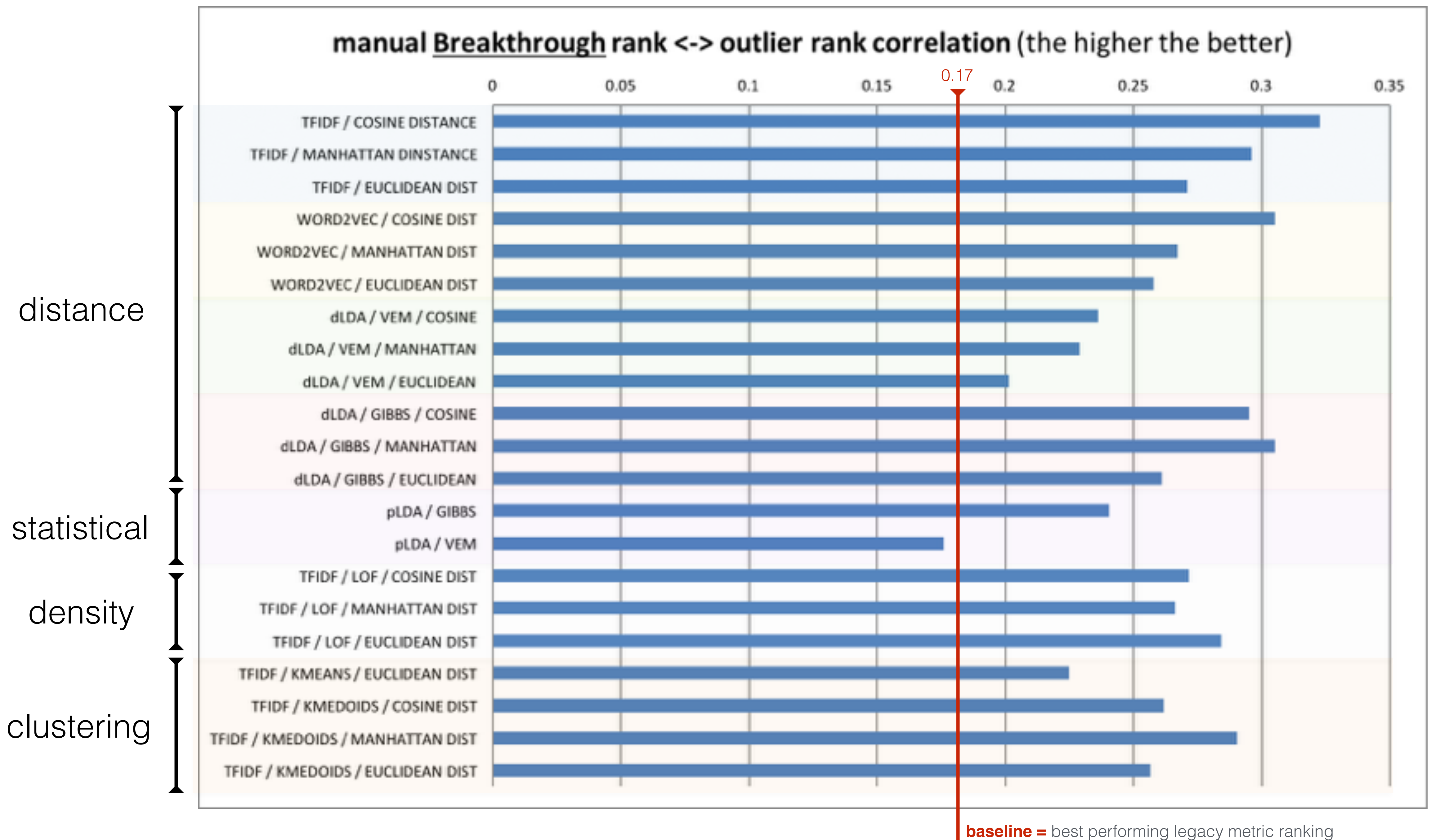
DISTANCE BASED OUTLIER DETECTION | **TF-IDF + COSINE = 0.28** -> **MEDIUM*** correlation with manual scoring

VS. **0.12** legacy score -> **WEAK***correlation with manual scoring

***Cohen** correlation scale for social sciences (Cohen,xxxx)

Results (Starbucks)

[correlation of algorithm rankings vs. manual picks]



BEST RESULT

DISTANCE BASED OUTLIER DETECTION | **TF-IDF + COSINE = 0.32** -> **MEDIUM*** correlation with manual scoring

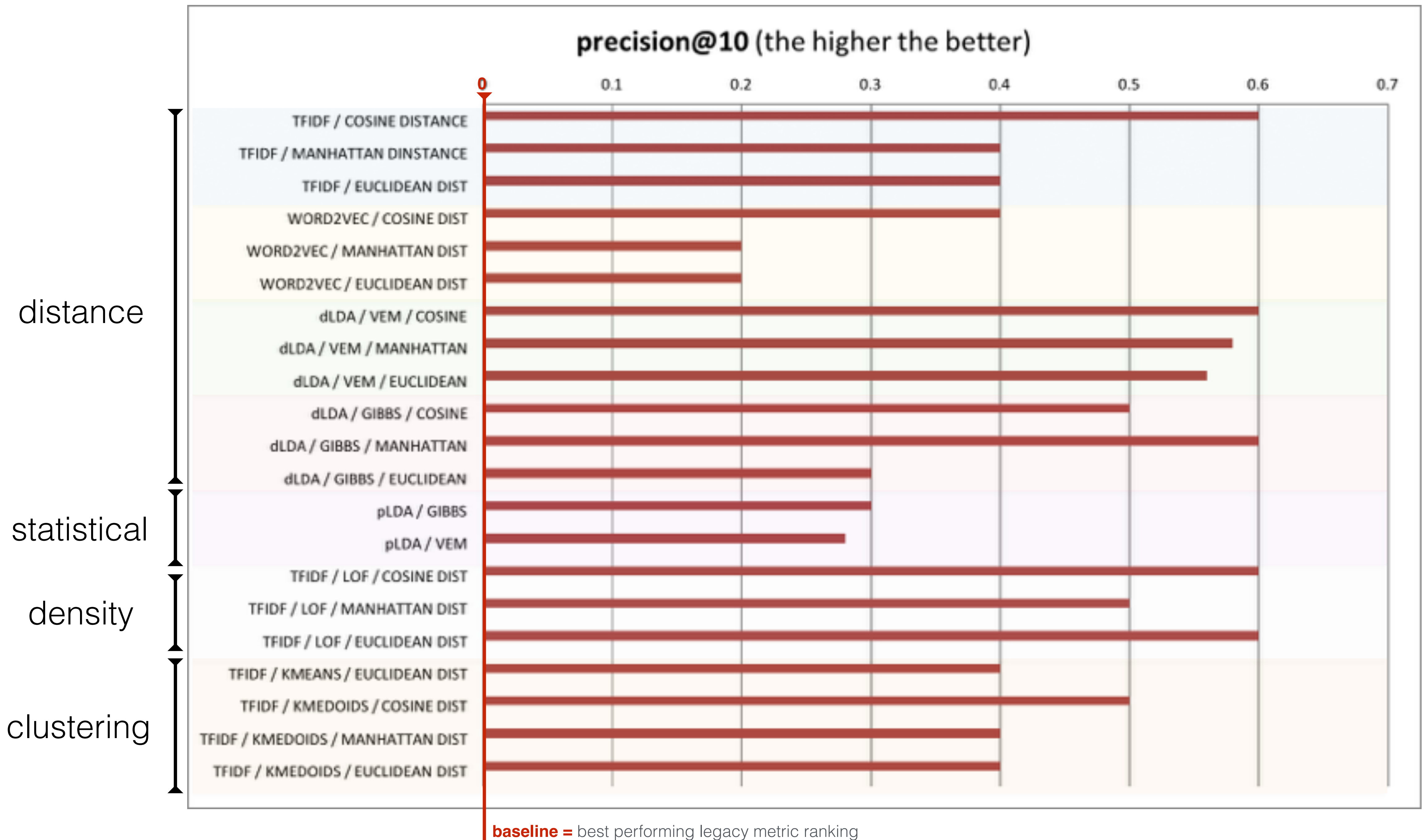
vs.

0.17 legacy score -> **WEAK***correlation with manual scoring

***Cohen** correlation scale for social sciences (Cohen,xxxx)

Results (IdeaStorm)

[precision@10 for algorithm rankings vs. manual picks]



BEST RESULT

DISTANCE BASED OUTLIER DETECTION | TF-IDF + COSINE* = 0.6

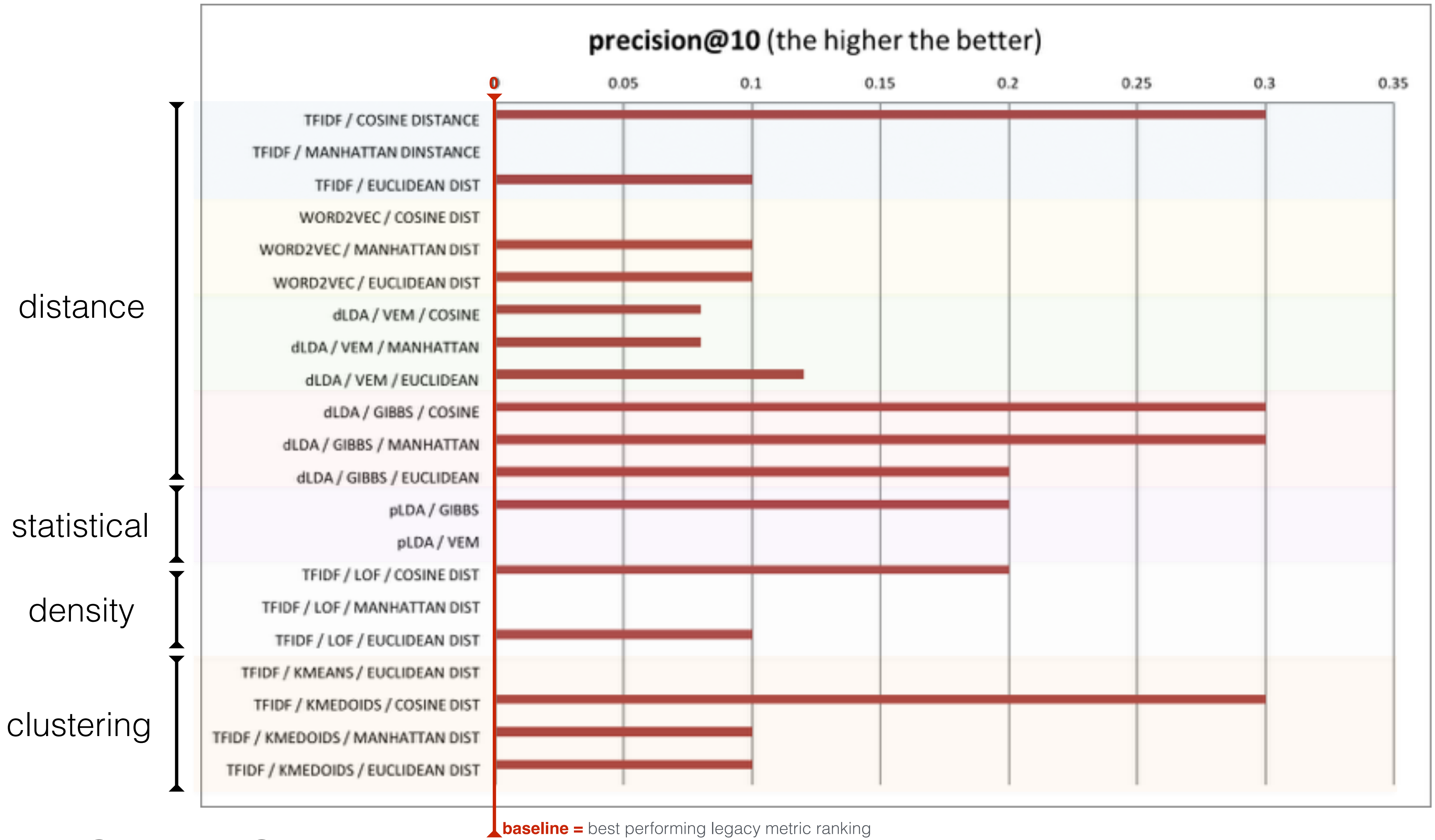
vs.

0 legacy score

*(similar dLDA + COSINE; dLDA + MANHATTAN; LOF + COSINE; LOF EUCLIDEAN)

Results (Starbucks)

[precision@10 of algorithm rankings vs. manual picks]



BEST RESULT

DISTANCE BASED OUTLIER DETECTION | **TF-IDF + COSINE*** = 0.3

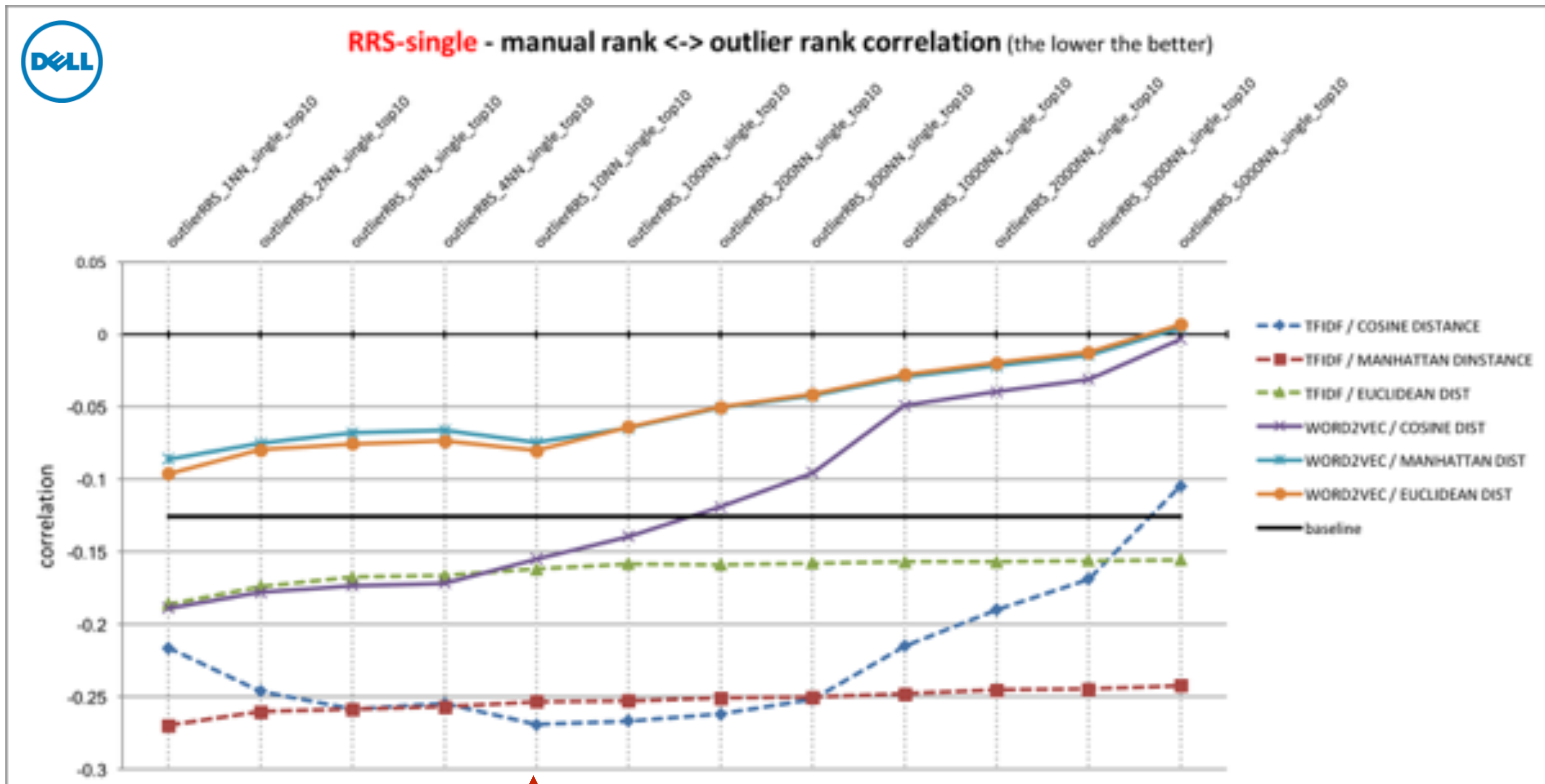
vs.

0 legacy score

*(similar **dLDA + COSINE; dLDA + MANHATTAN; K-MEDOIDS + COSINE**)

Results / distance algorithms

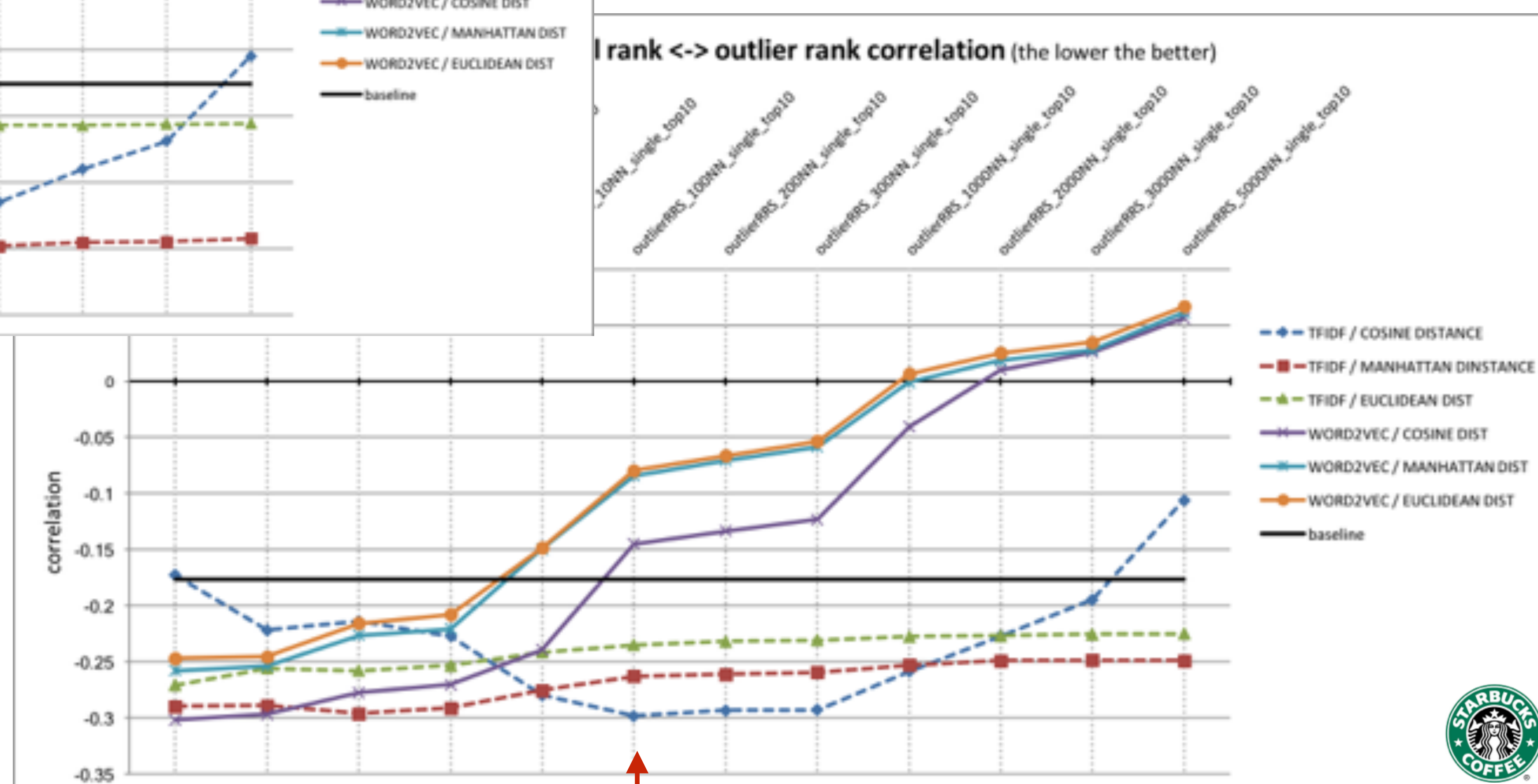
[correlation of algorithm rankings vs. manual picks]



best performance

closer look at best case

distance based algorithms



best performance

Observations:

- **Best performance I k = 100 I** dataset independent
- **Overall behaviour I** dataset independent

Conclusions

- **Distance based algorithms**
  perform best for particular problem discussed

- **Statistical outlier detection**
  performs worst and is also **hardest to tune**

- **Clustering algorithms**, contrary to distance algorithms
  significance of "k" outweighs any other parameter by big margin (in terms of accuracy impact)

- **All cases** (almost) regardless of approach outlier detection brings **new metric quality** to Idea Management System