In Search of Disruptive Ideas:
A Survey for Outlier Detection Techniques in Crowdsourcing Innovation Platforms
Overview

[research context / application area]

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

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**

"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]

**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, (i.e. low-end foothold; or non existent customers i.e. 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

IMS + winning ideas

IMS + disruptive ideas

IMS + outlier detection

---

**Outlier Detection + text** (subarea)

First (news) Story Detection/ TDT (98 - 04)
(novelty detection in stream)

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
- voice
- video

other media

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

---

**Innovation Management** (subarea)

TF-IDF + <approach x>

Theoretical models

- "Innovation management” A. Afuah, 1998

Case studies in enterprises

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

---

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

---

studies on selecting ideas based on legacy 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]
[ "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]
**Survey + alg picks**

### [different taxonomies / evaluations]

multitude of algorithms across years

→

many surveys / taxonomies to classify SoA

<table>
<thead>
<tr>
<th>CLASSIFICATION [Ji Zhang, 2013]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Statistical (probabilistic)</td>
</tr>
<tr>
<td>2. Distance (proximity) based</td>
</tr>
<tr>
<td>3. Density based</td>
</tr>
<tr>
<td>4. Clustering</td>
</tr>
<tr>
<td>5. High dimensional</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CLASSIFICATION [Aggarwal, 2013]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Extreme Value Analysis</td>
</tr>
<tr>
<td>2. Probabilistic and statistical</td>
</tr>
<tr>
<td>3. Linear</td>
</tr>
<tr>
<td>4. Proximity based</td>
</tr>
<tr>
<td>4.1. Clustering</td>
</tr>
<tr>
<td>4.2. Density</td>
</tr>
<tr>
<td>4.3. Nearest neighbour</td>
</tr>
<tr>
<td>5. Information Theory based</td>
</tr>
<tr>
<td>6. High dimensional</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CLASSIFICATION [Chandola, 2008]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Classification Based</td>
</tr>
<tr>
<td>2. Clustering Based</td>
</tr>
<tr>
<td>3. Nearest Neighbour Based</td>
</tr>
<tr>
<td>3.1. kNN</td>
</tr>
<tr>
<td>3.2. Density</td>
</tr>
<tr>
<td>4. Statistical</td>
</tr>
<tr>
<td>5. Information Theoretic</td>
</tr>
<tr>
<td>6. Spectral</td>
</tr>
</tbody>
</table>

**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]

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

3 innovation metrics:
- Implementation cost
- Potential profit
- Market size

1 overall rating
- Breakthrough

1 - 10 Likert scale based on innovation management theory

~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 (i.e., 1, 2, 3, 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 standpoint)

What makes a good ranking?

- [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

[ Evaluating Recommendation Systems, Guy Shani and Asela Gunawardana, Microsoft Research, 2009 ]
Results (IdeaStorm)

[correlation of algorithm rankings vs. manual picks]

**DISTANCE BASED OUTLIER DETECTION**

| TF-IDF + COSINE = 0.28 | MEDIUM* correlation with manual scoring |

**BEST RESULT**

| baseline | best performing legacy metric ranking |

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

<table>
<thead>
<tr>
<th>TF-IDF / COSINE DISTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF / MANHATTAN DISTANCE</td>
</tr>
<tr>
<td>TF-IDF / EUCLIDEAN DIST</td>
</tr>
<tr>
<td>WORD2VEC / COSINE DIST</td>
</tr>
<tr>
<td>WORD2VEC / MANHATTAN DIST</td>
</tr>
<tr>
<td>WORD2VEC / EUCLIDEAN DIST</td>
</tr>
<tr>
<td>dLDA / VEM / COSINE</td>
</tr>
<tr>
<td>dLDA / VEM / MANHATTAN</td>
</tr>
<tr>
<td>dLDA / VEM / EUCLIDEAN</td>
</tr>
<tr>
<td>dLDA / GIBBS / COSINE</td>
</tr>
<tr>
<td>dLDA / GIBBS / MANHATTAN</td>
</tr>
<tr>
<td>dLDA / GIBBS / EUCLIDEAN</td>
</tr>
<tr>
<td>pLDA / GIBBS</td>
</tr>
<tr>
<td>pLDA / VEM</td>
</tr>
<tr>
<td>TF-IDF / LOF / COSINE DIST</td>
</tr>
<tr>
<td>TF-IDF / LOF / MANHATTAN DIST</td>
</tr>
<tr>
<td>TF-IDF / LOF / EUCLIDEAN DIST</td>
</tr>
<tr>
<td>TF-IDF / KMEANS / EUCLIDEAN DIST</td>
</tr>
<tr>
<td>TF-IDF / KMEDOIDS / COSINE DIST</td>
</tr>
<tr>
<td>TF-IDF / KMEDOIDS / MANHATTAN DIST</td>
</tr>
<tr>
<td>TF-IDF / KMEDOIDS / EUCLIDEAN DIST</td>
</tr>
</tbody>
</table>

*Cohen correlation scale for social sciences (Cohen,xxxxx)
Results (Starbucks)

[correlation of algorithm rankings vs. manual picks]

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

*Cohen correlation scale for social sciences (Cohen,xxxxx)
Results (IdeaStorm)

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

<table>
<thead>
<tr>
<th>Algorithm Rankings</th>
<th>Precision@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF / COSINE DISTANCE</td>
<td>0.6</td>
</tr>
<tr>
<td>TF-IDF / MANHATTAN DISTANCE</td>
<td>0.5</td>
</tr>
<tr>
<td>TF-IDF / EUCLIDEAN DISTANCE</td>
<td>0.4</td>
</tr>
<tr>
<td>WORD2VEC / COSINE DISTANCE</td>
<td>0.3</td>
</tr>
<tr>
<td>WORD2VEC / MANHATTAN DISTANCE</td>
<td>0.2</td>
</tr>
<tr>
<td>WORD2VEC / EUCLIDEAN DISTANCE</td>
<td>0.1</td>
</tr>
<tr>
<td>dLDA / VEM / COSINE</td>
<td>0.0</td>
</tr>
<tr>
<td>dLDA / VEM / MANHATTAN</td>
<td>0.1</td>
</tr>
<tr>
<td>dLDA / VEM / EUCLIDEAN</td>
<td>0.2</td>
</tr>
<tr>
<td>dLDA / GIBBS / COSINE</td>
<td>0.3</td>
</tr>
<tr>
<td>dLDA / GIBBS / MANHATTAN</td>
<td>0.4</td>
</tr>
<tr>
<td>dLDA / GIBBS / EUCLIDEAN</td>
<td>0.5</td>
</tr>
<tr>
<td>pLDA / GIBBS</td>
<td>0.6</td>
</tr>
<tr>
<td>pLDA / VEM</td>
<td>0.7</td>
</tr>
<tr>
<td>TF-IDF / LOF / COSINE DISTANCE</td>
<td>0.8</td>
</tr>
<tr>
<td>TF-IDF / LOF / MANHATTAN DISTANCE</td>
<td>0.9</td>
</tr>
<tr>
<td>TF-IDF / LOF / EUCLIDEAN DISTANCE</td>
<td>1.0</td>
</tr>
<tr>
<td>TF-IDF / KMEANS / EUCLIDEAN DISTANCE</td>
<td>1.1</td>
</tr>
<tr>
<td>TF-IDF / KMEANS / COSINE DISTANCE</td>
<td>1.2</td>
</tr>
<tr>
<td>TF-IDF / KMEANS / MANHATTAN DISTANCE</td>
<td>1.3</td>
</tr>
<tr>
<td>TF-IDF / KMEANS / EUCLIDEAN DISTANCE</td>
<td>1.4</td>
</tr>
</tbody>
</table>

**BEST RESULT**

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

* (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**

\(^*\)(similar dLDA + COSINE; dLDA + MANHATTAN; K-MEDOIDS + COSINE)
Results / distance algorithms
[correlation of algorithm rankings vs. manual picks]

Observations:
- Best performance \( k = 100 \) dataset independent
- Overall behaviour dataset independent

closer look at best case
distance based algorithms
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