Towards Computational Assessment of Idea Novelty

Kai Wang\textsuperscript{1}  Boxiang Dong\textsuperscript{2}  Junjie Ma\textsuperscript{1}

\textsuperscript{1}School of Management and Marketing  
Kean University  
Union NJ

\textsuperscript{2}Department of Computer Science  
Montclair State University  
Montclair, NJ

Jan 11, 2019
Companies collect ideas from a large number of people to improve existing offerings [AT12, WN17].
Manually selecting the most innovative ideas from a large pool is not effective.
Manually selecting the most innovative ideas from a large pool is not effective.

It would be very helpful to automate the evaluation of creative ideas.
Idea Novelty Assessment

- Idea Similarity Comparison
  - Latent Semantic Analysis (LSA)
  - Latent Dirichlet Allocation (LDA)
- Proposal Novelty Evaluation
  - Term Frequency-Inverse Document Frequency (TF-IDF)

However, none of these approaches have been validated through the comparison with human judgment.
Our Contribution

- Three computational idea novelty evaluation approaches
  - LSA
  - LDA
  - TF-IDF
- Three sets of ideas
- Comparison with human expert evaluation
**Input** Idea by word matrix

**Output** Idea by topic matrix

**Key Idea** Apply *Singular Value Decomposition (SVD)* on the input matrix.

\[
\begin{align*}
T & \quad \text{Word by Idea Matrix} \quad (m \times n) \\
K & \quad \text{Word by Topic Matrix} \quad (m \times z) \\
S & \quad \text{Topic by Topic Matrix} \quad (z \times z) \\
D^T & \quad \text{Idea by Topic Matrix} \quad (n \times z)
\end{align*}
\]
**Input**  
Idea by word matrix

**Output**  
Idea by topic matrix

**Key Idea**

- Each idea is represented as a mixture of latent topics.
- Each topic is characterized as a distribution over words.

\[
P(w|d) \quad \text{Idea Distribution over Words} \quad (m \times n) \\
P(t|d) \quad \text{Idea Distribution over Topics} \quad (k \times m) \\
P(w|t) \quad \text{Topic Distribution over Words} \quad (n \times k) \\
\]

\[P(w|d) = x \times P(w|t)\]
Background - TF-IDF [WB13]

**Input**  Idea by word matrix

**Output**  Idea by word tf-idfs

**Key Idea**  Determine how important a word is to an idea.

\[
\text{tf-idf}(w_i, d_j) = tf(w_i, d_j) \times \log\left(\frac{n}{df(w_i)}\right)
\]

- \(tf(w_i, d_j)\): \# of times that \(w_i\) appears in \(d_j\)
- \(df(w_i)\): \# of ideas that include \(w_i\)
- \(n\): \# of ideas
We use Amazon Mechanical Turk (www.mturk.com) to employ crowd workers to collect three set of ideas.

**Alarm**  Ideas about a mobile app of an alarm clock.

**Fitness**  Ideas to improve physical fitness.

**Advertising**  Ideas to promote TV advertising.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Ideas</th>
<th>Avg. # of Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alarm</td>
<td>200</td>
<td>555</td>
</tr>
<tr>
<td>Fitness</td>
<td>240</td>
<td>586</td>
</tr>
<tr>
<td>Advertising</td>
<td>300</td>
<td>307</td>
</tr>
</tbody>
</table>
Methods - Human Expert Evaluation

We hire a group of human experts to evaluate the collected ideas.

- Each idea is evaluated by at least two human experts.
- Novelty is defined by using a Likert scale of 1 to 7 (1 being not novel at all, 7 being highly novel).
- Human experts demonstrate reasonable level of agreement in the ratings (Intraclass correlation coefficient is higher than 0.7).
- We take the average of human ratings as the ground truth of idea novelty.
Methods - Computational Novelty Evaluation

**LSA**  Cosine distance to average

**LDA**  
- Use Gibbs sampling with 2,000 iterations
- Cosine distance to average

**TF-IDF**  Sum of all tf-idfs in an idea
Experiments

We compare the following methods with the ground truth.

- LSA
- LDA
- TF-IDF
- Crowd

We hire 20 crowd workers to manually evaluate the idea novelty, and take their average.
Experiments

LSA correlates well with the ground truth on the Fitness and TV Advertising datasets.

LDA and TF-IDF performs well on all three datasets.

Crowd evaluation correlates with expert evaluation better than all the three computational methods.

<table>
<thead>
<tr>
<th>Ideation Tasks</th>
<th>Alarm Clock App (n=200)</th>
<th>Fitness App (n=240)</th>
<th>TV Advertising (n=300)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert Evaluated Novelty (Mean±SD)</td>
<td>4.55±1.01</td>
<td>4.64±1.38</td>
<td>3.97±1.20</td>
</tr>
<tr>
<td>Correlation of LSA measure with expert evaluation</td>
<td>0.114 (p=0.107)</td>
<td>0.230 (p&lt;0.001)</td>
<td>0.230 (p&lt;0.001)</td>
</tr>
<tr>
<td>Correlation of LDA measure (k=10) with expert evaluation</td>
<td>0.184 (p=0.009)</td>
<td>0.231 (p&lt;0.001)</td>
<td>0.198 (p&lt;0.001)</td>
</tr>
<tr>
<td>Correlation of LDA measure (k=20) with expert evaluation</td>
<td>0.253 (p&lt;0.001)</td>
<td>0.190 (p=0.003)</td>
<td>0.229 (p&lt;0.001)</td>
</tr>
<tr>
<td>Correlation of LDA measure (k=30) with expert evaluation</td>
<td>0.226 (p=0.001)</td>
<td>0.219 (p&lt;0.001)</td>
<td>0.235 (p&lt;0.001)</td>
</tr>
<tr>
<td>Correlation of TF-IDF measure with expert evaluation</td>
<td>0.340 (p&lt;0.001)</td>
<td>0.319 (p&lt;0.001)</td>
<td>0.307 (p&lt;0.001)</td>
</tr>
<tr>
<td>Correlation of crowd evaluation and expert evaluation</td>
<td>0.748 (p&lt;0.001)</td>
<td>0.501 (p&lt;0.001)</td>
<td>0.648 (p&lt;0.001)</td>
</tr>
</tbody>
</table>
Experiments

- Crowd evaluation identifies more top-10 novel ideas than all computational approaches.
- Crowd evaluation resulted in significant point-biserial correlation for all three ideation tasks.

### Table 2. Top ten novel ideas according to different measures.

<table>
<thead>
<tr>
<th>Ideation Tasks</th>
<th>Alarm Clock App</th>
<th>Fitness App</th>
<th>TV Advertising</th>
</tr>
</thead>
<tbody>
<tr>
<td>True novelty of the top ten ideas by LSA measure (Mean±SD)</td>
<td>4.38±0.82</td>
<td>5.45±1.01</td>
<td>5.1±1.07</td>
</tr>
<tr>
<td>Number of correctly identified top ten ideas by LSA measure</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Point-biserial correlation comparing top 10 novel ideas by LSA measure and the remaining ideas</td>
<td>-0.039 (p=0.587)</td>
<td>0.122 (p=0.059)</td>
<td>0.176 (p=0.002)</td>
</tr>
<tr>
<td>True novelty of the top ten ideas in LDA measure (k=20) (Mean±SD)</td>
<td>5.13±0.82</td>
<td>5.30±0.90</td>
<td>4.55±1.21</td>
</tr>
<tr>
<td>Number of correctly identified top ten ideas by LDA (k=20) measure</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Point-biserial correlation comparing top 10 novel ideas by LDA (k=20) measure and the remaining ideas</td>
<td>0.132 (p=0.062)</td>
<td>0.099 (p=0.125)</td>
<td>0.09 (p=0.119)</td>
</tr>
<tr>
<td>True novelty of the top ten ideas in TF-IDF measure (Mean±SD)</td>
<td>4.98±0.97</td>
<td>5.40±0.49</td>
<td>5.35±0.90</td>
</tr>
<tr>
<td>Number of correctly identified top ten ideas by TF-IDF measure</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Point-biserial correlation comparing top 10 novel ideas by TF-IDF measure and the remaining ideas</td>
<td>0.098 (p=0.168)</td>
<td>0.130 (p=0.045)</td>
<td>0.215 (p&lt;0.001)</td>
</tr>
<tr>
<td>True novelty of the top ten ideas in crowd evaluation (Mean±SD)</td>
<td>5.73±0.66</td>
<td>5.9±0.44</td>
<td>5.95±0.61</td>
</tr>
<tr>
<td>Number of correctly identified top ten ideas by crowd evaluation</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Point-biserial correlation comparing top 10 novel ideas by crowd evaluation and the remaining ideas</td>
<td>0.269 (p&lt;0.001)</td>
<td>0.190 (p=0.003)</td>
<td>0.300 (p&lt;0.001)</td>
</tr>
</tbody>
</table>
We experimentally compare three computational novelty evaluation approaches with ground truth.

- TF-IDF outperforms LSA and LDA in matching expert evaluation.
- All three computational approaches fall far behind crowd evaluation.
- Much more research is needed to automate the evaluation of creative ideas.
Crowdsourcing as a solution to distant search. 

The importance of iteration in creative conceptual combination. 

Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. 

[TN16] Olivier Toubia and Oded Netzer. 
Idea generation, creativity, and prototypicality. 

A text mining approach to evaluate submissions to crowdsourcing contests. 

A literature review on individual creativity support systems. 

Crowdsourced idea generation: the effect of exposure to an original idea. 
2013.
Thank you!

Questions?