Visualizing Dispersed Risk Signals for A Specific Emerging Technology: A Novel Approach of Keywords Aggregation across Topics (KAAT)

School of Business,
South China University of Technology, China
Prof. Munan Li
2018-09
Email: limn@scut.edu.cn
Research Background

• A typical interdisciplinary topic query:
  • “Risk for Graphene, Additive Manufacturing” could involve social science, management, business, environment science, engineering and so forth.

• Research Question:
  • For a specific emerging technology, Can we timely and efficiently discovery the risk signal / relevant works, especially in early period?
• **Research Design**

• **Goal:** Attempt to find a simple algorithm of machine learning to improve the capability and efficiency of discovery the risk signals of a specific emerging technology from publications, patents, and even Internet space.
• **Keywords Aggregation across Topics (KAaT):**
  • **KAaT could be or could be similar with a simple algorithm of machine learning**

• **Basic Philosophy (Components) of Machine Learning:**

  • (1) Computation logic (e.g. Non-linear programming model)
  • (2) Training & Optimize parameters
  • (3) Run Algorithm
  • (4) Verification & Feedback
• KAaT could be taken into account a naïve machine learning.

• Why is it naïve?

• Because this algorithm only utilize the basic philosophy of machine learning, and not involve such complicated topic modeling: LDA, LSA and etc.
• **Phase I: To find the synonyms of risk**

Qualitative tool
(Synonyms query from dictionaries in *general semantic context*)
- Oxford English Dictionary
- Cambridge Online
- Wiki Pedia
- Collins Dictionary

Quantitative tool
(Co-word analysis in *academic context*)
- Collect original corpus
- Co-word analysis (NLP)
- Compute Frequency & Centrality etc.
- Weighting & Ranking (TF*IDF, TOPSIS)
- Output synonyms in Academic context

Integrate into the synonyms of risk
• **KAaT:** Produce Training Sample in Phase II
  
  • TS="emerging technolog**" AND
  
  • TS=(risk* OR unsafe OR uncertainty OR danger* OR peril OR threat* OR menace OR fear OR unpredictab* OR precarious* OR instability* OR insecurity* OR perilousness OR venture OR jeopardy OR loss OR chancy OR toxic OR poison* OR vulnerability OR injury OR hazard* OR misfortune OR endanger OR jeopardize OR peril) AND TS =(environment* OR health* OR security OR safety OR ecosystem OR “air pollution” OR “soil contaminat**” OR “water pollution” OR “water contaminat**”)

  • AND DOCUMENT TYPES: (Article)

<table>
<thead>
<tr>
<th>Training Sample</th>
<th>Signal</th>
<th>Noise</th>
<th>Accuracy(%)</th>
<th>Presumed Recall(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>15</td>
<td>68</td>
<td>18.07</td>
<td>100</td>
</tr>
</tbody>
</table>

Algorithm training is to find the optimized keywords combinations that can efficiently identify signal and noise.
Training Results:

<table>
<thead>
<tr>
<th>Training Times (Attempt the different combinations of Terms)</th>
<th>Recall &gt;0 &amp; Accuracy &gt;0</th>
<th>Recall &gt; 20%</th>
<th>Accuracy &gt; 50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>11372</td>
<td>170</td>
<td>13</td>
<td>126</td>
</tr>
</tbody>
</table>

If the Recall (%) is prior to Accuracy (%), such keywords as: risk, health, environment, toxicity; and those 126 combination of keywords whose Accuracy is larger than 50% are selected to the next computation.
Case Study: The Risk Discovery of 3D printing/Additive Manufacturing

<table>
<thead>
<tr>
<th>Query Formula</th>
<th>Result</th>
<th>Type</th>
<th>Refine Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1    TS=((3D OR 3-D OR &quot;3 dimension*&quot; OR &quot;three dimension*&quot; OR additive) NEAR/2 (print* OR fabricat* OR manufactur* OR product*))</td>
<td>8477</td>
<td>ARTICLE (5,321)</td>
<td>DOCUMENT TYPES: ( ARTICLE OR PROCEEDINGS PAPER OR REVIEW ) Timespan: 2015-2016. Indexes: SCI-EXPANDED, SSCI, A&amp;HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC.</td>
</tr>
<tr>
<td>#2    TS=((3D OR 3-D OR &quot;3 dimension*&quot; OR &quot;three dimension*&quot; OR additive) NEAR/2 (print* OR fabricat* OR manufactur* OR product*)) AND TS=(risk* OR pathogen OR &quot;Scenario planning&quot; OR environment* OR health OR toxic*)</td>
<td>995</td>
<td>ARTICLE (609)</td>
<td>PROCEEDINGS PAPER (340) REVIEW (72)</td>
</tr>
</tbody>
</table>

(1) query formula refers the work of Yin Huang et al., which is published in 2017. and (2#) query formula combines 1# with training results.
Descriptive results on 3D printing studies:
Descriptive results on 3D printing studies:

Top 15 Countries/Regions on 3D printing papers in WOS

- USA: 4789 papers
- SOUTH KOREA: 515 papers
- AUSTRALIA: 511 papers
- CANADA: 476 papers
- PEOPLE'S REPUBLIC OF CHINA: 1938 papers
- ENGLAND: 1129 papers
- FRANCE: 468 papers
- JAPAN: 396 papers
- SPAIN: 322 papers
- ITALY: 645 papers
- SINGAPORE: 419 papers
- NETHERLANDS: 298 papers
- SWITZERLAND: 291 papers
- INDIA: 268 papers
Descriptive results on 3D printing studies:
Finally, based on the 126 rules, 99 publications are selected from 995 samples; and 23 are matched signal, another are noise; therefore, the recall is 100%, accuracy is just 23%. 

Naïve Machine Learning Result:
KAaT: a naïve algorithm of machine learning

Conclusion or Implication

In summary, based on the introduced NML (Naïve Machine Learning), a method or algorithm for discovery the risk signal of a specific emerging technology are explored. And NML can compress the noise space, and bring a moderate accuracy of identification.

Also, in the future research, more complicated semantic modeling can be integrated into NML to further improve the accuracy.
Research Extension: Can the above question be transformed into a linear classifier question?

\[ f(x, w, b) = \text{sign}(w \cdot x - b) \]

How would you classify this data?

A classic linear classifier: Support Vector Machine (SVM)
How would you classify this data?

Support Vector Machines

\[ f(x, w, b) = \text{sign}(w \cdot x - b) \]
Support Vector Machines

$f(x, w, b) = \text{sign}(w \cdot x - b)$

Any of these would be fine..

..but which is best?
Support Vector Machines

\[ f(x, w, b) = \text{sign}(w \cdot x - b) \]

Linear Classifier

*margin*: the distance between hyper surface and the nearest samples (dots)
Find the biggest margin

\[ f(x, w, b) = \text{sign}(w \cdot x - b) \]

Support Vectors: those nearest points to hyper surface

This is a simple SVM, and also named as linear SVM

Linear SVM
Hyperplane and the margin

- **Plus-Plane**
- **Minus-Plane**
- **Classifier Boundary**

```
zone
“Predict Class = +1”

“Predict Class = -1”
zone
```

Diagram with axes $x_1$ and $x_2$, showing points classified as positive (circles) and negative (squares) with respect to the hyperplane and the margin.
Compute margin

How to use $\mathbf{w}$ and $b$ to compute margin?

- **Plus-plane** = \( \{ \mathbf{x} : \mathbf{w} \cdot \mathbf{x} + b = +1 \} \)
- **Minus-plane** = \( \{ \mathbf{x} : \mathbf{w} \cdot \mathbf{x} + b = -1 \} \)
So, the above question could be transformed into Optimization question (Quadratic Programming)

Minimize \( \frac{1}{2} \mathbf{w} \cdot \mathbf{w} \)

subject to \( y_k (\mathbf{w} \cdot \mathbf{x}_k + b) \geq 1 \) for \( k = 1, 2, \ldots, n \)
Research Extension II: Can be transformed into Bipartite Graph? (And then, graph theory and those related algorithms could be helpful)

**U:** the **relevant** papers on risk analysis (Environment, Health and Safety etc.) for a specific emerging technology

**V:** the **irrelevant** papers on risk analysis (Environment, Health and Safety etc.) for a specific emerging technology

**U + V:** All papers on the specific emerging technology (e.g. 3D printing, synthetic biology, Graphene, etc.)
Thank you!