A Smarter Process for Sensing the Information Space

Scott Spangler
IBM Almaden Services Research
Outline

• What is Exploratory Analytics
• The Practice of Exploratory Analytics
• Towards a Smarter Process
• Corporate Brand Reputation Analysis
• Looking Forward
• Further Reading
Exploratory Analytics

• **Exploratory data analysis (EDA)** is an approach to analysing data for the purpose of formulating hypotheses worth testing, complementing the tools of conventional statistics for testing hypotheses[1]. It was so named by John Tukey to contrast with Confirmatory Data Analysis, the term used for the set of ideas about hypothesis testing, p-values, confidence intervals etc. which formed the key tools in the arsenal of practicing statisticians at the time. (wikipedia)

• Less Formally: Exploratory Analytics is the process of analyzing data to learn about what you don’t know to ask.
Words to Live By…

If we need a short suggestion of what exploratory data analysis is, I would suggest that

1. It is an attitude AND
2. A flexibility AND
3. Some graph paper and transparencies

No catalogue of techniques can convey a willingness to look for what can be seen, whether or not anticipated. Yet this is at the heart of exploratory data analysis. The graph paper - and transparencies - are there, not as a technique, but rather as recognition that the picture-examining eye is the best finder we have of the wholly unanticipated.

-- John Tukey
## Comparison

<table>
<thead>
<tr>
<th>Confirmatory</th>
<th>Exploratory</th>
</tr>
</thead>
<tbody>
<tr>
<td>• More data is “better”</td>
<td>• Topical Collections</td>
</tr>
<tr>
<td>• Predefined Annotations</td>
<td>• Discovered Features</td>
</tr>
<tr>
<td>• Batch processing</td>
<td>• Dynamic processing</td>
</tr>
<tr>
<td>• Supervised Learning</td>
<td>• Unsupervised Learning</td>
</tr>
<tr>
<td>• Single Hierarchical Approach</td>
<td>• Multiple user defined taxonomies</td>
</tr>
<tr>
<td>• Precision/Recall</td>
<td>• Interestingness/Utility</td>
</tr>
</tbody>
</table>

### Reductionism vs. Holism

"The whole is more than the sum of its parts" -- Aristotle
Limits of Reductionism

- Ground Truth – what does it mean in practice
- Precision/Recall – misses the interesting “wrong” answer
- Repeatability – not practical
- Optimization – one person’s “optimum” is another person’s “useless”
- The perils of “overfit”
Confirmatory Analytics Pitfalls

• Atlantic magazine article: “Lies, Damned Lies, and Medical Science” “I realized even our gold-standard research had a lot of problems,” [Dr. John Ioannidis] says. Baffled, he started looking for the specific ways in which studies were going wrong. And before long he discovered that the range of errors being committed was astonishing…

• This array suggested a bigger, underlying dysfunction, and Ioannidis thought he knew what it was. “The studies were biased,” ….” Researchers headed into their studies wanting certain results—and, lo and behold, they were getting them. We think of the scientific process as being objective, rigorous, and even ruthless in separating out what is true from what we merely wish to be true, but in fact it’s easy to manipulate results, even unintentionally or unconsciously.

• The odds are that in any large database of many nutritional and health factors, there will be a few apparent connections that are in fact merely flukes,

• Medical research is not especially plagued with wrongness. Other meta-research experts have confirmed that similar issues distort research in all fields of science
Organization, Analysts, and Data

- Organizational Attributes
  - Goals
  - Limitations
  - Vocabulary

- Analysts Characteristics
  - Knowledge
  - Biases
  - Algorithm and Report preferences
  - Scope of Responsibility

- Data Features
  - Reliability
  - Coverage
  - Timeliness
The Practice of Exploratory Analytics

Jam Analytics: $25 Million impact

BIW: $34 Million impact

COBRA: $50 Million impact

SIMPLE: $15 Million impact

Total Impact: $124 Million and counting….
A process for Exploratory Analytics

1) Parenthesis: (Localize Data)
2) Expunge Noise
3) Multiple Feature Spaces
4) Divide and Conquer
5) Add Connections
6) Subtract the Uninteresting Material

Algebraic order of operations: (, **, *, /, +, - or Please Excuse My Dear Aunt Sally)
Tools/Techniques

• Parenthesis: Localize
  – Search
  – Query
  – Proximal Search
  – Snippets
  – Set Operations
  – Sampling
Proximal Search
Tools/Techniques

• Expunge Noise
  – Tools for customizing how we spot and eliminate irrelevant text
    • Junk filters
    • Boiler plate detectors
    • Eliminate garbage text of all kinds
  – Unsupervised and Supervised methods for selection of relevant text content.
    • Clustering
    • Classifiers
Intuitive Clustering

- Rank order terms based on a “cohesion” score
- Terms that score relatively high with this measure tend to be those with a significant number of examples having many words in common. Adjusting the $n$ parameter tends to surface more general terms with larger matching sets, while adjusting it upward gives more specific terms.
- The greedy algorithm selects enough of the most cohesive terms to get 80-90% of the data categorized. Terms are selected in cohesive order, skipping those terms in the list that do not add a significant number (e.g. more than 3) of additional examples to those already categorized with previous terms.
- The algorithm halts when at least X% of the data has been categorized and the uncategorized examples are placed in a “Miscellaneous” category.
- The resulting categories are then refined using a single iteration of k-means (i.e. each document is placed in the category of the nearest centroid as calculated by the term membership just described).

\[
\text{cohesion}(T, n) = \frac{\sum \text{cos}(\text{centroid}(T), x)}{|T|}
\]
Tools and Techniques

• Multiple Feature Spaces
  – Generated Words and Phrases
  – Feature spaces generated from advanced annotators and user supplied dictionaries
  – User input allowed in creating automatically generated features
  – Regular expression patterns for advanced user created features.
    • Social Security Numbers
    • Customer complaints
User Defined Features in COBRA

Name: CRUELTY

Include Patterns:
- cruel
- abuse
- abusive
- mistreat
- cause
- (0,10)suffer
- euthanize
- (0,10)truth
- an abuse of
- had
- abused

Exclude Patterns:
- child
- abuse
- shouting
- abuse
- shout
- abuse
- cruel
- truth
- an abuse of
- had
- abused

All Sources:

Blogs:

MessageBoard:

News:

ImpactViewer

Save Changes
Tools/Techniques

• Divide and Conquer
  – Intuitive Clustering
  – Keyword Based Taxonomies
  – Time based Taxonomies
  – Sentiment Analysis
  – Iterative Taxonomy Browsing/Editing
  – Visualization
Visualization
Tools/Techniques

- Add connections
  - Co-occurrence table
  - Trend Chart
  - Dictionary Analysis
  - Word Trends
  - Industry Landscape
  - Networks
<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>PFIZER</th>
<th>AstraZeneca</th>
<th>AMGEN</th>
<th>Genentech</th>
<th>Novartis</th>
<th>Merck</th>
<th>Bristol-Myers</th>
<th>Johnson &amp; Johnson</th>
</tr>
</thead>
<tbody>
<tr>
<td>alzheimer</td>
<td>321</td>
<td>204 (0.0)</td>
<td>7 (1.0)</td>
<td>18 (6.7538)</td>
<td>1 (1.0)</td>
<td>12 (1.0)</td>
<td>36 (1.0)</td>
<td>43 (0.27108)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>anti-infam...</td>
<td>367</td>
<td>115 (1.3328)</td>
<td>11 (0.55007)</td>
<td>10 (0.88916)</td>
<td>5 (1.0)</td>
<td>50 (0.87330)</td>
<td>130 (0.4197)</td>
<td>41 (1.0)</td>
<td>5 (0.521284)</td>
</tr>
<tr>
<td>arthritis</td>
<td>577</td>
<td>232 (1.3305)</td>
<td>29 (8.89878)</td>
<td>36 (2.71071)</td>
<td>57 (1.0)</td>
<td>36 (1.0)</td>
<td>128 (1.0)</td>
<td>59 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>asthma</td>
<td>552</td>
<td>213 (1.4128)</td>
<td>36 (8.12254)</td>
<td>17 (0.48193)</td>
<td>52 (1.0)</td>
<td>28 (1.0)</td>
<td>164 (1.0)</td>
<td>52 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>breast</td>
<td>1384</td>
<td>62 (1.0)</td>
<td>4 (1.0)</td>
<td>3 (1.0)</td>
<td>1238 (0.0)</td>
<td>13 (1.0)</td>
<td>41 (1.0)</td>
<td>23 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>cancer</td>
<td>785</td>
<td>253 (6.0049)</td>
<td>19 (1.0)</td>
<td>41 (2.63881)</td>
<td>115 (1.0)</td>
<td>30 (1.0)</td>
<td>233 (1.0)</td>
<td>94 (0.64142)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>cardiovas...</td>
<td>416</td>
<td>156 (2.8534)</td>
<td>15 (0.15002)</td>
<td>1 (1.0)</td>
<td>23 (1.0)</td>
<td>6 (1.0)</td>
<td>140 (0.9309)</td>
<td>76 (1.01206)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>cartilage</td>
<td>474</td>
<td>8 (1.0)</td>
<td>0 (1.0)</td>
<td>1 (1.0)</td>
<td>449 (0.0)</td>
<td>4 (1.0)</td>
<td>7 (1.0)</td>
<td>5 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>cervical</td>
<td>985</td>
<td>5 (1.0)</td>
<td>0 (1.0)</td>
<td>0 (1.0)</td>
<td>979 (0.0)</td>
<td>0 (1.0)</td>
<td>0 (1.0)</td>
<td>1 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>coding_seq...</td>
<td>1782</td>
<td>11 (1.0)</td>
<td>1 (1.0)</td>
<td>12 (1.0)</td>
<td>1740 (0.0)</td>
<td>9 (1.0)</td>
<td>7 (1.0)</td>
<td>2 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>colon</td>
<td>1307</td>
<td>48 (1.0)</td>
<td>2 (1.0)</td>
<td>7 (1.0)</td>
<td>215 (0.0)</td>
<td>8 (1.0)</td>
<td>7 (1.0)</td>
<td>20 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>delivery</td>
<td>268</td>
<td>37 (1.0)</td>
<td>13 (0.0138)</td>
<td>25 (3.62070)</td>
<td>14 (1.0)</td>
<td>26 (1.0)</td>
<td>108 (0.0167)</td>
<td>34 (0.52423)</td>
<td>11 (4.89102)</td>
</tr>
<tr>
<td>dna</td>
<td>2473</td>
<td>47 (1.0)</td>
<td>2 (1.0)</td>
<td>187 (0.0)</td>
<td>1907 (0.0)</td>
<td>77 (1.0)</td>
<td>196 (1.0)</td>
<td>57 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>gastrointes...</td>
<td>397</td>
<td>177 (2.7908)</td>
<td>33 (8.39105)</td>
<td>9 (1.0)</td>
<td>7 (1.0)</td>
<td>15 (1.0)</td>
<td>116 (1.0)</td>
<td>40 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>gene</td>
<td>1169</td>
<td>72 (1.0)</td>
<td>16 (1.0)</td>
<td>75 (3.03985)</td>
<td>745 (0.0)</td>
<td>83 (1.0)</td>
<td>129 (1.0)</td>
<td>48 (1.0)</td>
<td>1 (1.0)</td>
</tr>
<tr>
<td>growth_hor...</td>
<td>312</td>
<td>67 (0.26885)</td>
<td>4 (1.0)</td>
<td>8 (1.0)</td>
<td>113 (4.11372)</td>
<td>2 (1.0)</td>
<td>112 (0.3564)</td>
<td>6 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>heart</td>
<td>461</td>
<td>207 (0.0)</td>
<td>2 (1.0)</td>
<td>3 (1.0)</td>
<td>37 (1.0)</td>
<td>32 (1.0)</td>
<td>132 (1.0)</td>
<td>44 (1.0)</td>
<td>4 (1.0)</td>
</tr>
<tr>
<td>immune</td>
<td>352</td>
<td>109 (8.5430)</td>
<td>9 (1.0)</td>
<td>18 (0.00292)</td>
<td>82 (5.85665)</td>
<td>7 (1.0)</td>
<td>85 (1.0)</td>
<td>50 (0.10208)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>kinase</td>
<td>245</td>
<td>48 (0.82418)</td>
<td>10 (0.11448)</td>
<td>26 (2.54970)</td>
<td>44 (0.55085)</td>
<td>30 (1.0)</td>
<td>44 (1.0)</td>
<td>41 (0.00900)</td>
<td>2 (1.0)</td>
</tr>
<tr>
<td>liver</td>
<td>1329</td>
<td>47 (1.0)</td>
<td>2 (1.0)</td>
<td>12 (1.0)</td>
<td>1204 (0.0)</td>
<td>8 (1.0)</td>
<td>41 (1.0)</td>
<td>15 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>lung</td>
<td>1466</td>
<td>76 (1.0)</td>
<td>5 (1.0)</td>
<td>13 (1.0)</td>
<td>1268 (0.0)</td>
<td>17 (1.0)</td>
<td>68 (1.0)</td>
<td>19 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>pain</td>
<td>529</td>
<td>268 (0.0)</td>
<td>51 (2.14510)</td>
<td>21 (0.04642)</td>
<td>2 (1.0)</td>
<td>27 (1.0)</td>
<td>138 (1.0)</td>
<td>19 (1.0)</td>
<td>3 (1.0)</td>
</tr>
<tr>
<td>rheumatoid</td>
<td>425</td>
<td>154 (6.2229)</td>
<td>25 (7.12472)</td>
<td>30 (5.78526)</td>
<td>53 (1.0)</td>
<td>22 (1.0)</td>
<td>99 (1.0)</td>
<td>42 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>stroke</td>
<td>405</td>
<td>116 (0.0)</td>
<td>19 (0.00463)</td>
<td>16 (0.08691)</td>
<td>12 (1.0)</td>
<td>12 (1.0)</td>
<td>87 (1.0)</td>
<td>48 (0.80072)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>tumor</td>
<td>1908</td>
<td>98 (1.0)</td>
<td>10 (1.0)</td>
<td>42 (1.0)</td>
<td>333 (0.0)</td>
<td>42 (1.0)</td>
<td>183 (1.0)</td>
<td>200 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>vaccine</td>
<td>178</td>
<td>41 (0.17239)</td>
<td>3 (1.0)</td>
<td>7 (0.265987)</td>
<td>17 (1.0)</td>
<td>3 (1.0)</td>
<td>107 (3.5736)</td>
<td>0 (1.0)</td>
<td>0 (1.0)</td>
</tr>
<tr>
<td>vascular</td>
<td>350</td>
<td>118 (1.6240)</td>
<td>9 (0.944719)</td>
<td>2 (1.0)</td>
<td>91 (1.56106)</td>
<td>17 (1.0)</td>
<td>82 (1.0)</td>
<td>28 (1.0)</td>
<td>3 (1.0)</td>
</tr>
<tr>
<td>virus</td>
<td>317</td>
<td>68 (0.26957)</td>
<td>0 (1.0)</td>
<td>10 (0.53945)</td>
<td>27 (1.0)</td>
<td>33 (1.0)</td>
<td>137 (2.0156)</td>
<td>42 (0.31203)</td>
<td>0 (1.0)</td>
</tr>
</tbody>
</table>

Total: 17701

count: 3370

445

462

2930

3262

5922

2028

182
Tools/Techniques

• Subtraction – Focus on Value
  – Most Typical Sorting
  – Value Estimation
  – Modeling
  – Prediction
Patent Value Estimation

<table>
<thead>
<tr>
<th>Patent</th>
<th>Published</th>
<th>Cited</th>
<th>Main US Class</th>
<th>Distinctive word/phrase</th>
<th>Timeline</th>
<th>1st use</th>
<th>Days</th>
<th>Patents</th>
</tr>
</thead>
<tbody>
<tr>
<td>US6822286</td>
<td>2001-08-28</td>
<td>26</td>
<td>379</td>
<td>internet telephone</td>
<td>2000-02-29</td>
<td>545</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>US6633914</td>
<td>2003-10-14</td>
<td>11</td>
<td>709</td>
<td>greetings prior</td>
<td>2001-08-28</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>US6650661</td>
<td>2003-11-18</td>
<td>6</td>
<td>370</td>
<td>server threads</td>
<td>2003-02-04</td>
<td>251</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>US8968340</td>
<td>2005-11-22</td>
<td>1</td>
<td>707</td>
<td>region relative</td>
<td>2003-11-18</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Copyright © 2009 IBM
COrporate Brand and Reputation Analysis -- COBRA

An end-to-end social media mining solution with comprehensive analytics

- Dynamic dashboard overview and reporting
- Early warning detection and monitoring & alerting
- Emerging topics and theme detection
- Sentiment and opinion mining
- Competitive analysis
- Relationships and trends
- Social network analysis and influencer identification
Orthogonal Filtering – the power of independent rules

Data of interest lies at the intersection of multiple independent business rules.

Rules that independently have high recall but low precision can be combined to produce a rule that has high precision and only slightly lower recall.
COBRA brings it all together by extracting content and using Text Analytics to detect Signals from social media conversation.
COBRA Backend Process

Query Management

Data Loading

Models/Hotwords

Generate System
**Type of Analysis**
Combines the capability of keyword, faceted search with a visual results list, which can be manipulated by the end user.

**What can the user perform**
- Visual results list based on users search request
- Visualize “points of view” by focusing in on sources, models or hotword categories
- Visualize psychographics trends like Geo, Age, Gender, Sentiment, Topic Text Classifications
Type of Analytics

Taxonomy is the classification of document postings based on their content. Some of the taxonomies are static like Age, Source, Gender, and some of the taxonomies are dynamic like Text Clustering which is based on the words or concepts in the documents.

What can the user perform

Allow the user to select one or more of the taxonomies and see the entire document posting associated with these taxonomies along with their trends.

Allow the user to build a matrix of text cluster topics and look at the document which associated the intersection of the topics. Then they can see the entire document posting associated with this intersection along with their trends.
**Type of Analysis**

Relationships are similar to Taxonomies in that they provide the user the capability to generate a matrix which can be used to discovery document posting for intersection. The major difference is it allow you create a matrix for any of the taxonomies, both static (age, source, gender) or word based (text clusters).

**What can the user perform**

Allow the user to select one or more of the taxonomies and see the entire document posting associated with these taxonomies along with their trends.

Allow the user to build a matrix for any taxonomy and look at the document which associated the intersection of the topics. Then they can see the entire document posting associated with this intersection along with their trends.
**Type of Analysis**

Sentiment measures the “tonality” of the posting, either negative or positive. This is done based on the words in the postings and is presented as trends.

**What can the user perform**

- Capability to select any of the taxonomy dimensions (age, brand, text cluster) and visualize the sentiment trends
- Select a specific dimension and visualize a side by side comparisons of negative / positive sentiment
- View all of the posting by the sentiment categories (negative or positive)
Type of Analysis

Key Influencers are measured by the depth of the conservation about a topic and popularity of the site which is measured by the number of sites who link to these site.

What can the user perform

- Identify who are the key influencer by topic
- Identify what site “BasCitation” are listening to the key influencers sites
- Key influencers sites can be used during “relationship” and “sentiment” analysis
**Type of Analysis**

Using orthogonal models identify social media postings (Blogs, Message Boards, News) which match the user defined topics description with a 95% precision

**What can the user perform**

- A list of document which are grouped by the user defined topics base on models / hotwords
- Trend chart for the selected topics compared to the total of all document for this period
- Drill down to view full text of document and with like to original source
Looking Forward

• Analytic Recipes
• Goal Directed Analytics
• Exploratory Analytics on the Index
• Exploration and Hadoop
• Modeling
• Exploratory Analytics of Structured Data
Further Reading


• **Mining the talk: unlocking the business value in unstructured information.** S Spangler, J Kreulen, IBM Press, 2007.
