TextScope:
Enhance Human Perception via Text Mining

ChengXiang ("Cheng") Zhai
Department of Computer Science
(Carl R. Woese Institute for Genomic Biology
School of Information Sciences
Department of Statistics)
University of Illinois at Urbana-Champaign

czhai@illinois.edu  http://czhai.cs.illinois.edu/

IEEE BigData 2017, Boston, Dec. 12, 2017
Text data cover all kinds of topics

Topics:
People
Events
Products
Services, ...

Sources:
Blogs
Microblogs
Forums
Reviews ,...

45M reviews
53M blogs
1307M posts
65M msgs/day
115M users
10M groups
Humans as **Subjective & Intelligent “Sensors”**

Real World ➔ **Sense** ➔ Sensor ➔ **Report** ➔ Data

- **Weather** ➔ Thermometer ➔ 3°C, 15°F, …
- **Locations** ➔ Geo Sensor ➔ 41°N and 120°W, …
- **Networks** ➔ Network Sensor ➔ 01000100011100

Perceive ➔ **Express** ➔ “Human Sensor”
Unique Value of Text Data

• Useful to all big data applications
• Especially useful for mining knowledge about people’s behavior, attitude, and opinions
• Directly express knowledge about our world: Small text data are also useful!

Data ➔ Information ➔ Knowledge

Text Data
Opportunities of Text Mining Applications

1. Mining knowledge about language
2. Mining content of text data
3. Mining knowledge about the observer
4. Infer other real-world variables (predictive analytics)

Real World → Perceive (Perspective) → Observed World → Express (English) → Text Data

+ Non-Text Data + Context
However, NLP is difficult!

“A man saw a boy with a telescope.” (who had the telescope?)

“He has quit smoking” ➜ he smoked before.

How can we leverage imperfect NLP to build a perfect general application?

Answer: Having humans in the loop!
TextScope to enhance human perception

Microscope

Telescope

TextScope

Intelligent Interactive Retrieval & Text Analysis for Task Support and Decision Making
TextScope in Action: intelligent interactive decision support

Predicted Values of Real World Variables

Optimal Decision Making

Real World

Sensor 1

Sensor k

Non-Text Data

Learning to interact

Domain Knowledge

Prediction

Text + Non-Text

Interactive text analysis

Interactive information retrieval

Natural language processing

TextScope

Text + Non-Text Data

Interactive information retrieval

Natural language processing

Optimal Decision Making

Real World

Sensor 1

Sensor k

Non-Text Data

Learning to interact

Domain Knowledge

Prediction

Text + Non-Text

Interactive text analysis

Interactive information retrieval

Natural language processing

TextScope
Microsoft (MSFT), Google, IBM (IBM) and other cloud-computing rivals of Amazon Web Services are bracing for an AWS "partnership" announcement with VMware expected to be announced Thursday.

<table>
<thead>
<tr>
<th>RANK</th>
<th>DATE</th>
<th>EXCERPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9/29/2000</td>
<td>Expect earning will be far below</td>
</tr>
<tr>
<td>2</td>
<td>12/8/2000</td>
<td>$4 billion cash in company</td>
</tr>
<tr>
<td>3</td>
<td>10/19/2000</td>
<td>Disappointing earning report</td>
</tr>
<tr>
<td>4</td>
<td>4/19/2001</td>
<td>Dow and Nasdaq soar after rate cut by Federal Reserve</td>
</tr>
<tr>
<td>5</td>
<td>7/20/2001</td>
<td>Apple's new retail store</td>
</tr>
</tbody>
</table>

...
Application Example 1: Medical & Health

Joint Mining of Non-Text and Text

Multiple Predictors (Features)

Doctors, Nurses, Patients...

Predictive Model

Non-Text Data

Text Data

Optimal Decision Making

Diagnosis, optimal treatment
Side effects of drugs, ...

Medical & Health

Predicted Values

Sensor 1

Sensor k
1. Extraction of Adverse Drug Reactions from Forums

Green: Disease symptoms
Blue: Side effect symptoms
Red: Drug

Drug: Cefalexin
ADR: panic attack
faint
....
A Probabilistic Model for ADR Extraction [Wang et al. 14]

• **Challenge**: how do we separate treated symptoms from side-effect symptoms?

• **Solution**: leverage knowledge about known symptoms of diseases treated by a drug

• **Probabilistic model**
  – Assume forum posts are generated from a mixture language model
  – Most words are generated using a background language model
  – Treated (disease) symptoms and side-effect symptoms are generated from separate models, enabled by the use of external knowledge about known disease symptoms
  – Fitting the model to the forum data allows us to learn the side-effect symptom distributions

Sample ADRs Discovered [Wang et al. 14]

<table>
<thead>
<tr>
<th>Drug(Freq)</th>
<th>Drug Use</th>
<th>Symptoms in Descending Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zoloft</td>
<td>antidepressant</td>
<td><strong>weigh gain, weight</strong>, depression, side effects, mgs, gain weight, anxiety, nausea, head, brain, pregnancy, pregnant, headaches, depressed, tired</td>
</tr>
<tr>
<td>Ativan</td>
<td>anxiety disorders</td>
<td>Ativan, sleep, Seroquel, doc prescribed seroqual, raising blood sugar levels, anti-psychotic drug, diabetic, constipation, diabetes, 10mg, benzo, addicted</td>
</tr>
<tr>
<td>Topamax</td>
<td>anticonvulsant</td>
<td>Topmax, liver, side effects, migraines, headaches, weight, Topamax, pdoc, neurologist, supplement, sleep, fatigue, seizures, liver problems, kidney stones</td>
</tr>
<tr>
<td>Ephedrine</td>
<td>stimulant</td>
<td>dizziness, stomach, Benadryl, dizzy, tired, lethargic, tapering, tremors, panic attach, head</td>
</tr>
</tbody>
</table>

2. Analysis of Electronic Medical Records (EMRs)

EMRs (Patient Records)

Disease Profile

Typical symptoms: \( P(\text{Symptom} | \text{Disease}) \)

Typical treatments: \( P(\text{Treatment} | \text{Disease}) \)

Effectiveness of treatment

Subcategories of disease

… …
The Conditional Symptom-Treatment Model [Wang et al. 16]

\[
Pr(s, h|t) = \sum_{d \in D_t} Pr(d|t) Pr(s|d) Pr(h|d)
\]

Likelihood of patient t having disease d

Typical symptoms of disease d

Typical treatments of disease d

Symptom

Treatment

Disease

All diseases of patient t

Observed Patient Record

Evaluation: Traditional Chinese Medicine (TCM) EMRs

• 10,907 patients TCM records in digestive system treatment
• 3,000 symptoms, 97 diseases and 652 herbs
• Most frequently occurring disease: chronic gastritis
• Most frequently occurring symptoms: abdominal pain and chills

• Ground truth: 27,285 manually curated herb-symptom relationship.
Output of the model

Disease distribution $P(d \mid t)$ of patient $t$

- $P(\text{Anemia} = 0.3)$
- $P(\text{Gastritis} = 0.6)$
- $P(\text{Hypertension} = 0.1)$

- Dizziness 0.3
- Vomiting 0.3
- Hypodynamia 0.3
- Abdominal 0.1
- Distension

- Medlar 0.4
- Ginseng 0.3
- Dates 0.3

- Stomachache 0.5
- Vomiting 0.3
- Abdominal 0.2
- Distension

- Bitter Orange 0.4
- Medlar 0.3
- Dates 0.3

Symptom cluster $P(s \mid d)$ of Anemia
Herb cluster $P(h \mid d)$ of Anemia
Symptom cluster $P(s \mid d)$ of Gastritis
Herb cluster $P(h \mid d)$ of Gastritis
“Typical Symptoms” of 3 Diseases: $p(s|d)$

<table>
<thead>
<tr>
<th>chronic gastritis</th>
<th>constipation</th>
<th>reflux esophagitis</th>
</tr>
</thead>
<tbody>
<tr>
<td>胃胀</td>
<td>餐便困难</td>
<td>痞满</td>
</tr>
<tr>
<td>epigastric distention</td>
<td>difficult bowel movements</td>
<td>上腹部胀满和满胀</td>
</tr>
<tr>
<td>胃脘腹寒</td>
<td>腹胀</td>
<td>反酸 Acid reflux</td>
</tr>
<tr>
<td>epigastric chills</td>
<td>abdominal distension</td>
<td>腹胀</td>
</tr>
<tr>
<td>痞满</td>
<td>大便干燥</td>
<td>烧心 Heartburn</td>
</tr>
<tr>
<td>upper abdominal distention and fullness</td>
<td>dry, hard stools</td>
<td>烧心 Heartburn</td>
</tr>
<tr>
<td>烧心</td>
<td>乏力</td>
<td>呃气 Belching</td>
</tr>
<tr>
<td>heartburn</td>
<td>hypodynamia</td>
<td>呃气 Belching</td>
</tr>
<tr>
<td>反酸</td>
<td>腹部畏寒</td>
<td>咽部异物感 Paresthesia pharynges</td>
</tr>
<tr>
<td>acid reflux</td>
<td>abdominal chills</td>
<td>咽部异物感 Paresthesia pharynges</td>
</tr>
</tbody>
</table>
“Typical Herbs” Prescribed for 3 Diseases: $p(h|d)$

<table>
<thead>
<tr>
<th>Disease</th>
<th>Herb</th>
<th>Herb</th>
<th>Herb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperlipidemia</td>
<td>荷叶 - lotus leaf</td>
<td>荷叶 - lotus leaf</td>
<td>桔梗 - Chinese bellflower root</td>
</tr>
<tr>
<td></td>
<td>生山楂 - raw hawthorn</td>
<td>茵陈 - Oriental wormwood</td>
<td>紫苏叶 - perilla root</td>
</tr>
<tr>
<td></td>
<td>生地黄 - R. glutinosa root</td>
<td>牡丹皮 - peony root bark</td>
<td>连翘 - weeping forsythia fruit</td>
</tr>
<tr>
<td></td>
<td>醋莪术 - turmeric vinegar</td>
<td>虎杖 - Japanese knotweed</td>
<td>炒牛蒡子 - stir-fried burdock</td>
</tr>
<tr>
<td></td>
<td>虎杖</td>
<td>栀子 - gardenia fruit</td>
<td>玄参</td>
</tr>
</tbody>
</table>
## Algorithm-Recommended Herbs vs. Physician-Prescribed Herbs

<table>
<thead>
<tr>
<th>Model</th>
<th>Prescription from our model</th>
<th>Different herbs</th>
<th>Shared herbs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>炒牛蒡子 (stir-fried greater burdock),</td>
<td>陈皮 (Pixie tangerine peel),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>生槟榔 (betelnut),</td>
<td>柴胡 (Chinese thorowax root),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>麦炒白术 (white atractyloides</td>
<td>蜜甘草 (licorice root),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rhizome),</td>
<td>麥炒朽实 (stir-fried immature</td>
</tr>
<tr>
<td></td>
<td></td>
<td>黃芩 (baikal skullcap root),</td>
<td>bitter orange in bran),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>黃芪 (astragalus root),</td>
<td>炒白芍 (fried white peony root),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>枳壳 (bitter orange),</td>
<td>当归 (Chinese angelica root)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>茯苓 (tuckahoe)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prescription from physician</td>
<td>生地黃 (R. glutinosa),</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>大腹皮 (areca nut shell),</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>桃仁 (peach seed),</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>酒苁蓉 (desert broomrape),</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>炒杏仁 (Siberian apricot seed),</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>火麻仁 (hemp fruit)</td>
<td></td>
</tr>
</tbody>
</table>
Application Example 2: Business intelligence

Optimal Decision Making

Predictive Model

Joint Mining of Non-Text and Text

Non-Text Data

Text Data

Sensor 1

Sensor k

Business analysts, Market researcher...

Multiple Predictors (Features)

Products

Business intelligence
Consumer trends...

Predicted Values

Consumer Variables

Business intelligence
Market researcher...

Predictive Model

Multiple Predictors (Features)
Latent Aspect Rating Analysis (LARA) [Wang et al. 10]

Hotel Palomar Chicago: Traveler Reviews

Great location+spacious room = happy traveler

Stayed for a weekend in July. Walked everywhere, enjoyed the comfy bed and quiet hallways. more
terrific service and gorgeous facility

I stayed at the Palomar with my young daughter for three nights June 17-20, 2010 and absolutely loved the hotel. The room was one of the nicest I’ve ever stayed in (My daughter loved the Fuji jetted tub so much that she wanted to take 2 baths a day!) in terms of decor, design, and size. It compared favorably to... more

How to infer aspect weights?

How to infer aspect ratings?

Solving LARA in two stages:
Aspect Segmentation + Rating Regression

<table>
<thead>
<tr>
<th>Reviews + overall ratings</th>
<th>Aspect segments</th>
<th>Term Weights</th>
<th>Aspect Rating</th>
<th>Aspect Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>location:1</td>
<td>β_i 0.0</td>
<td>Si 3.9</td>
<td>α_d 0.2</td>
<td></td>
</tr>
<tr>
<td>amazing:1</td>
<td>2.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>walk:1</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>anywhere:1</td>
<td>0.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>room:1</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nicely:1</td>
<td>1.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>appointed:1</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>comfortable:1</td>
<td>3.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nice:1</td>
<td>2.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>accommodating:1</td>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smile:1</td>
<td>1.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>friendliness:1</td>
<td>2.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>attentiveness:1</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Aspect Segmentation + Latent Rating Regression
Latent Rating Regression

Aspect segments
- location: 1
- amazing: 1
- walk: 1
- anywhere: 1
- room: 1
- nicely: 1
- appointed: 1
- comfortable: 1
- nice: 1
- accommodating: 1
- smile: 1
- friendliness: 1
- attentiveness: 1

Term Weights
- amazing: 0.9
- walk: 0.1
- anywhere: 0.3
- nicely: 0.7
- accommodating: 0.1
- comfortable: 0.9
- nice: 0.6
- friendly: 0.8
- smile: 0.7
- friendliness: 0.8
- attentiveness: 0.9

Aspect Rating
- location: 1.3
- amazing: 0.9
- walk: 0.1
- anywhere: 0.3
- room: 1.8
- nicely: 0.7
- appointed: 0.6
- comfortable: 0.8
- nice: 3.8
- accommodating: 0.8
- smile: 0.7
- friendliness: 0.8
- attentiveness: 0.9

Aspect Weight
- location: 0.2
- amazing: 0.2
- walk: 0.6
- anywhere: 0.2
- room: 0.6
- nicely: 0.2
- appointed: 0.6
- comfortable: 0.8
- nice: 3.8
- accommodating: 0.8
- smile: 0.7
- friendliness: 0.8
- attentiveness: 0.9

Conditional likelihood

\[ P(r|d) = P(r_d|\mu, \Sigma, \delta^2, \beta, W_d) \]

\[ = \int p(\alpha_d|\mu, \Sigma)p(r_d|\sum_{i=1}^{k} \alpha_{d_i} \sum_{j=1}^{n} \beta_{dij} W_{dij}, \delta^2) d\alpha_{d} \]
Excellent location in walking distance to Tiananmen Square and shopping streets. That’s the best part of this hotel! The rooms are getting really old. Bathroom was nasty. The fixtures were falling off, lots of cracks and everything looked dirty. I don’t think it worth the price. Service was the most disappointing part, especially the door men. this is not how you treat guests, this is not hospitality.

Sample Result 1: Rating Decomposition

- Hotels with the same overall rating but different aspect ratings

(All 5 Stars hotels, ground-truth in parenthesis.)

<table>
<thead>
<tr>
<th>Hotel</th>
<th>Value</th>
<th>Room</th>
<th>Location</th>
<th>Cleanliness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grand Mirage Resort</td>
<td>4.2(4.7)</td>
<td>3.8(3.1)</td>
<td>4.0(4.2)</td>
<td>4.1(4.2)</td>
</tr>
<tr>
<td>Gold Coast Hotel</td>
<td>4.3(4.0)</td>
<td>3.9(3.3)</td>
<td>3.7(3.1)</td>
<td>4.2(4.7)</td>
</tr>
<tr>
<td>Eurostars Grand Marina Hotel</td>
<td>3.7(3.8)</td>
<td>4.4(3.8)</td>
<td>4.1(4.9)</td>
<td>4.5(4.8)</td>
</tr>
</tbody>
</table>

- Reveal detailed opinions at the aspect level
Sample Result 2: Comparison of reviewers

- Reviewer-level Hotel Analysis
  - Different reviewers’ ratings on the same hotel
  - Reveal differences in opinions of different reviewers

<table>
<thead>
<tr>
<th>Reviewer</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr. Saturday</td>
<td>3.7(4.0)</td>
</tr>
<tr>
<td>Salsrug</td>
<td>5.0(5.0)</td>
</tr>
</tbody>
</table>

(Review: "Good Price for what we got"

Riu Palace Punta Cana

Salsrug 13 contributions
Marylander
Oct 27, 2008 | Trip type: Family

We stayed for six days, five nights. Overall, we had a very good time. The restaurant was very friendly. They definitely do not skip on the free alcohol. The room was a little smelly, which we had read about, but it's not a problem. They only thing I had an issue with was the little bugs. They were like gnats or fleas but they didn't bite me. I had some candy and popcorn which we bought from the States to munch on. I left it out on the table and within 40 minutes, the bag was infested. DO NOT keep any open food in your room. Also we ended up having to wash all of our clothes (clean and dirty) and airing out our luggage when we got home because we could still smell the room on them. For the price we paid, we really did have an excellent time besides these small things. The pool was awesome and the beach was spectacular. Out of the nearby resorts that we saw, Riu Palace Punta Cana was the best (it was also the nicest out of the other Riu's on Punta Cana). We went on the 1/2 day Outback Safari and had a great time. We got coffee and souvenirs cheaper than other places and the hotel. General - not good or bad just things that we noticed - There were a lot of topless sunbathers. The crowd is middle aged (35 - 55) so we were on the younger side and the majority of the people were European or Brazilian. It helps to know some Spanish but it's not a necessity.

Liked — The beach was excellent.
Disliked — Room smell and little bugs.

My ratings for this hotel

- 5
- 5
- 3
- 4
- 3

Value
Rooms
Location
Cleanliness
Check in / front desk

Hotel Riu Palace Punta Cana)
Sample Result 3: Aspect-Specific Sentiment Lexicon

<table>
<thead>
<tr>
<th>Value</th>
<th>Rooms</th>
<th>Location</th>
<th>Cleanliness</th>
</tr>
</thead>
<tbody>
<tr>
<td>resort 22.80</td>
<td>view 28.05</td>
<td>restaurant 24.47</td>
<td>clean 55.35</td>
</tr>
<tr>
<td>value 19.64</td>
<td>comfortable 23.15</td>
<td>walk 18.89</td>
<td>smell 14.38</td>
</tr>
<tr>
<td>excellent 19.54</td>
<td>modern 15.82</td>
<td>bus 14.32</td>
<td>linen 14.25</td>
</tr>
<tr>
<td>worth 19.20</td>
<td>quiet 15.37</td>
<td>beach 14.11</td>
<td>maintain 13.51</td>
</tr>
<tr>
<td>bad -24.09</td>
<td>carpet -9.88</td>
<td>wall -11.70</td>
<td>smelly -0.53</td>
</tr>
<tr>
<td>money -11.02</td>
<td>smell -8.83</td>
<td>bad -5.40</td>
<td>urine -0.43</td>
</tr>
<tr>
<td>terrible -10.01</td>
<td>dirty -7.85</td>
<td>road -2.90</td>
<td>filthy -0.42</td>
</tr>
<tr>
<td>overprice -9.06</td>
<td>stain -5.85</td>
<td>website -1.67</td>
<td>dingy -0.38</td>
</tr>
</tbody>
</table>

Uncover sentimental information directly from the data
# Sample Result 4: User Rating Behavior Analysis

<table>
<thead>
<tr>
<th></th>
<th>Expensive Hotel</th>
<th></th>
<th>Cheap Hotel</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 Stars</td>
<td>3 Stars</td>
<td>5 Stars</td>
<td>1 Star</td>
</tr>
<tr>
<td>Value</td>
<td>0.134</td>
<td>0.148</td>
<td>0.171</td>
<td>0.093</td>
</tr>
<tr>
<td>Room</td>
<td>0.098</td>
<td>0.162</td>
<td>0.126</td>
<td>0.121</td>
</tr>
<tr>
<td>Location</td>
<td>0.171</td>
<td>0.074</td>
<td>0.161</td>
<td>0.082</td>
</tr>
<tr>
<td>Cleanliness</td>
<td>0.081</td>
<td>0.163</td>
<td>0.116</td>
<td>0.294</td>
</tr>
<tr>
<td>Service</td>
<td><strong>0.251</strong></td>
<td>0.101</td>
<td>0.101</td>
<td>0.049</td>
</tr>
</tbody>
</table>

People like expensive hotels because of good service.

People like cheap hotels because of good value.
Sample Result 5: Personalized Recommendation of Entities

Query: 0.9 value
0.1 others

<table>
<thead>
<tr>
<th>Hotel</th>
<th>Overall Rating</th>
<th>Price</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majestic Colonial</td>
<td>5.0</td>
<td>339</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Agua Resort</td>
<td>5.0</td>
<td>753</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Majestic Elegance</td>
<td>5.0</td>
<td>537</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Grand Palladium</td>
<td>5.0</td>
<td>277</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Iberostar</td>
<td>5.0</td>
<td>157</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Elan Hotel Modern</td>
<td>5.0</td>
<td>216</td>
<td>Los Angeles</td>
</tr>
<tr>
<td>Marriott San Juan Resort</td>
<td>4.0</td>
<td>354</td>
<td>San Juan</td>
</tr>
<tr>
<td>Punta Cana Club</td>
<td>5.0</td>
<td>409</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Comfort Inn</td>
<td>5.0</td>
<td>155</td>
<td>Boston</td>
</tr>
<tr>
<td>Hotel Commonwealth</td>
<td>4.5</td>
<td>313</td>
<td>Boston</td>
</tr>
</tbody>
</table>
Application Example 3: Prediction of Stock Market

- Market volatility
- Stock trends, ...

Optimal Decision Making

- Events in Real World

- Sensor 1
- Sensor k

Predictive Model

- Non-Text Data
- Text Data

Multiple Predictors (Features)

Joint Mining of Non-Text and Text

Stock traders
Text Mining for Understanding Time Series [Kim et al. CIKM’13]

What might have caused the stock market crash?

Sept 11 attack!

TextScope

Any clues in the companion news stream?


Heuristic Optimization of Causality + Coherence

<table>
<thead>
<tr>
<th>AAMRQ (American Airlines)</th>
<th>AAPL (Apple)</th>
</tr>
</thead>
<tbody>
<tr>
<td>russia russian putin</td>
<td>paid notice st</td>
</tr>
<tr>
<td>europe european</td>
<td>russia russian europe</td>
</tr>
<tr>
<td>germany</td>
<td>olympic games olympics</td>
</tr>
<tr>
<td>bush gore presidential</td>
<td>she her ms</td>
</tr>
<tr>
<td>police court judge</td>
<td>oil ford prices</td>
</tr>
<tr>
<td>airlines airport air</td>
<td>black fashion blacks</td>
</tr>
<tr>
<td>united trade terrorism</td>
<td>computer technology software</td>
</tr>
<tr>
<td>food foods cheese</td>
<td>internet com web</td>
</tr>
<tr>
<td>nets scott basketball</td>
<td>football giants jets</td>
</tr>
<tr>
<td>tennis williams open</td>
<td>japan japanese plane</td>
</tr>
<tr>
<td>awards gay boy</td>
<td></td>
</tr>
<tr>
<td>moss minnesota chechnya</td>
<td></td>
</tr>
</tbody>
</table>

Topics are biased toward each time series

---

“Causal Topics” in 2000 Presidential Election

Top Three Words in Significant Topics from NY Times

- **tax cut**
- screen
- pataki
guiliani
enthusiasm
door
symbolic
- **oil energy**
prices
news w top
pres al vice
love tucker presented
partial **abortion**
privatization
court supreme **abortion**
gun control nra

Issues known to be important in the 2000 presidential election

Time Series: Iowa Electronic Market
http://tippie.uiowa.edu/iem/
A general TextScope to support many different applications?

Predicted Values of Real World Variables

Optimal Decision Making

Medical & Health
E-Commerce
Stocks & Financial
Education
Security

Sensor 1
...
Sensor k

Non-Text Data

Learning to interact

Domain Knowledge

Text + Non-Text

Interactive text analysis

Interactive information retrieval

Natural language processing

TextScope

Joint Mining of Non-Text and Text Data

Predictive Model

Multiple Predictors (Features)

Predicted Values of Real World Variables

Real World

Sensor 1
...
Sensor k

Interactive information retrieval

Natural language processing
Major Challenges in Building a General TextScope

• Different applications have different requirements  ➔ Need abstraction
  – What are the common analysis operators shared by multiple text analysis tasks?
  – How can we design a general text analysis language covering many applications?

• Retrieval and analysis need to be integrated  ➔ A unified operator-based framework
  – How can we formalize retrieval and analysis functions as multiple compatible general operators?
  – How can we manage workflow?

• How can we optimize human-computer collaboration?
  – How can TextScope adapt to a user’s need dynamically and support personalization?
  – How can humans train/teach TextScope with minimum effort?

• How can we perform joint analysis of text and non-text data?

• Implementation Challenges: Architecture of a general TextScope? Real-time response?
Some Possible Analysis Operators

- Select
- Split
- Intersect
- Union
- Ranking
- Topic
- Compare
- Interpret

<table>
<thead>
<tr>
<th>Common</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Formalization of Operators

- $C=\{D_1, \ldots, D_n\}$; $S, S_1, S_2, \ldots, S_k$ subset of $C$
- Select Operator
  - Querying (Q): $C \rightarrow S$
  - Browsing: $C \rightarrow S$
- Split
  - Categorization (supervised): $C \rightarrow S_1, S_2, \ldots, S_k$
  - Clustering (unsupervised): $C \rightarrow S_1, S_2, \ldots, S_k$
- Interpret
  - $C \times \theta \rightarrow S$
- Ranking
  - $\theta \times S_i \rightarrow$ ordered $S_i$
Compound Analysis Operator: Comparison of K Topics

\[
\text{Interpret}(\text{Compare}(\text{Select}(T_1,C), \text{Select}(T_2,C), \ldots, \text{Select}(T_k,C)), C)
\]
Compound Analysis Operator: Split and Compare

\[
\text{Interpret(} \text{Compare(} \text{Split}(S,k),C) \text{)}
\]
BeeSpace: Analysis Engine for Biologists [Sarma et al. 11]

Summary

- **Human as Subject Intelligent Sensor ➔ Special value of text for mining**
  - Applicable to all “big data” applications
  - Especially useful for mining human behavior, preferences, and opinions
  - Directly express knowledge (small text data are useful as well)

- **Difficulty in NLP ➔ Must optimize the collaboration of humans and machines, maximization of combined intelligence of humans and computers**
  - Let computers do what they are good at (statistical analysis and learning)
  - Turn imperfect techniques into perfect applications

- **Vision of TextScope**: many applications & many new challenges
  - Integration of intelligent retrieval and text analysis
  - Joint analysis of text and non-textual (context) data
  - How to optimize the collaboration (combined intelligence) of computer and humans?
Beyond TextScope: Intelligent Task Agent, DataScope

Intelligent Task Agents

- Learning to explore
- Learning to collaborate

Real World

Sensor 1

Sensor k

Learning to interact

Domain Knowledge

Prediction

Text + Non-Text

Analysis of non-text data

Interactive text analysis

Interactive information retrieval

Natural language processing

predicted values of real world variables

Joint mining of non-text and text data

Multiple predictors (features)

Optimal decision making

Interactive information retrieval

Natural language processing

Learning to interact

Learning to collaborate

Domain Knowledge

Text + Non-Text

Interactive text analysis

Interactive information retrieval

Natural language processing

Joint mining of non-text and text data

Multiple predictors (features)

Optimal decision making

Interactive information retrieval

Natural language processing

Joint mining of non-text and text data

Multiple predictors (features)

Optimal decision making
General Open Research Challenges

• **Grand Challenge:** How to maximize the **combined intelligence** of humans and machines instead of intelligence of machines alone

• How to optimize the “cooperative game” of **human-computer collaboration**?
  – Machine learning is just one way of human-computer collaboration
  – What are other forms of collaboration? How to optimally divide the task between humans and machines?

• How to **minimize** the **total effort of a user** in finishing a task?
  – How to go beyond component evaluation to measure task-level performance?
  – How to optimize sequential decision making (reinforcement learning)?
  – How to model/predict user behavior?
  – How to minimize user effort in labeling data (active learning)?
  – How to explain system operations to users?

• How to **minimize** the **total system operation cost**?
  – How to model and predict system operation cost (computing resources, energy consumption, etc)?
  – How to optimize the tradeoff between operation cost and system intelligence?

• **Robustness Challenge:** How to manage/mitigate risk of system errors? Security problems?
Acknowledgments

• Collaborators: Many former and current TIMAN group members, UIUC colleagues, and many external collaborators

• Funding
References


Thank You!

Questions/Comments?