User Behavior Analysis in Big Data

Morteza(Mori) Zihayat
About Myself

- Mitacs Elevate Postdoctoral Research Fellow
  - Faculty of Information, University of Toronto
- Big Data Scientist
  - Spectrum Computing, IBM
  - Globe and Mail
- Main research interests
  - User modeling
  - Big Data Mining and Engineering
    - Finding Meaningful Patterns from Structured and Unstructured Big Data
    - Parallel and Distributed Data Mining
  - Social Network Analysis
Outline

- Introduction to Big Data
- User Modeling in Digital Media
- Depression Acuity Detection
- Cogniciti: An Online Brain Health Assessment
- Conclusion
Outline

- Introduction to Big Data
- User Modeling in Digital Media
- Depression Acuity Detection
- Cogniciti: An Online Brain Health Assessment
- Conclusion
Big Data

50 MILLION TWEETS PER DAY

150 Exabytes
HEALTHCARE DATA

200 Gigabytes
THE SIZE OF A SINGLE SEQUENCED HUMAN GENOME

50 MILLION DEVICES
EACH DEVICE GENERATES 156 MB OF DATA PER DAY

0.5% ever analyzed and used
(MIT Technology review)
Big Data to Data Science

Data Scientist:
The Sexiest Job of the 21st Century

Harvard Business Review

(c) 2012 Biocomicals
by Dr. Alper Uzon
Challenges

- Read/Write to disk is slow
  - Use multiple disks for parallel read
- Hardware failure
  - Single machine/disk failure
  - Keep multiple copies of data
- How do we merge data from different reads
  - Distributed processing or Hadoop MapReduce
Apache Hadoop

- Hadoop is an open-source software framework written in Java
  - Distributed storage
  - Computer clusters

- Two main components
  - HDFS (Hadoop Distributed File System)
    - Provides Distributed Storage
  - MapReduce (Distributed Data Processing Model)
    - Provides Distributed Processing
MapReduce

- MapReduce is a method for distributing a task across multiple nodes
- Consists of two phases
  - Map
  - Reduce
- The data is processed in the form of <key, value>
- Each map task processes a discrete portion of the overall data
- After all Maps are complete, the system distributes the intermediate data to nodes which perform Reduce phase (aggregation)
Example: Word counting

• Counting the number of occurrences of each word in a large collection of documents
  • Input: documents
  • Output: \(<\text{word}, \text{frequency}\>\)

1. Divide collection of documents among the class.
2. Each student gives count of individual word in a document. Repeats for assigned quota of documents.
3. Sum up the counts from all the documents to give final answer.
Word Count Execution

Input

the quick brown fox
the fox ate the cow
how now brown cow

Output

brown, 2
fox, 2
how, 1
now, 1
the, 3

Student 1

the, 1
brown, 1
fox, 1
quick, 1

Student 1

brown, 1

Student 2

the, 1
fox, 1
the, 1
ate, 1
cow, 1

Student 2

brown, 1

Student 3

how, 1
now, 1
brown, 1
cow, 1

Student 2

ate, 1
cow, 2
quick, 1
Word Count Execution

Input

the quick brown fox

the fox ate the cow

how now brown cow

Map tasks

Map

the, 1
brown, 1
fox, 1
quick, 1

Map

the, 1
fox, 1
the, 1
ate, 1
cow, 1

Map

how, 1
now, 1
brown, 1
cow, 1

Reduce tasks

brown, 1

brown, 1

brown, 1

Reduce

Output

brown, 2
fox, 2
how, 1
now, 1
the, 3

ate, 1
cow, 2
quick, 1
Problem #1

- MapReduce I/O sandbags runtime for advanced analytics.
  - Must persist results after each pass through data
  - Advanced analytics often requires multiple passes through data
Example

- Goal: find best line separating two sets of points

![Diagram showing a random initial line and a target line with data points.

Bar chart showing running time and number of iterations with Hadoop.

Running Time (s) vs Number of Iterations:
- 1: 0
- 5: 500
- 10: 1000
- 20: 2000
- 30: 4500

18/11/2016
Apache Spark

- A response to limitations in the MapReduce
- UC Berkeley’s AMP Lab (2009)
- Fully open sourced in 2010
Spark Performance

- Machine Learning
  - 100x faster than MapReduce

- Queries (Shark)
  - 100x faster than Hive

- Streaming
  - 2X throughput of Storm

- Graph (GraphX)
  - 10X faster than MapReduce
Outline

- Introduction to Big Data
- User Modeling in Digital Media
- Depression Acuity Detection
- Cogniciti: An Online Brain Health Assessment
- Conclusion
Ottawa promises to overhaul mental-health services for military
• A Must-Read: Trump reaches out to distance from his turbulent years

Trump taps Flynn, Sessions and Pompeo for top positions
• A Must-Read: Trump reaches out to distance from his turbulent years

Teenager Tony wins Ontario by-election, Liberals hold Ont. in Warner

Raising doses payments needed to battle ironing sickness: CIR: CEO

Global bonds poised for steepest two-week loss in quarter century

IT MITE: Is pay, says who be a Canadian CEO

Top military physician skeptical about toxicity of malaria medication

U.S. ELECTION: With so many unknowns, it’s a challenge to avoid jarring over Donald Trump’s plans for Muslims

JOHAN WEBB: All people who use opioids are dying every day. Why won’t Ottawa talk to us?

GLOBE EDITORIAL: Call Canada’s Iqra mission by its real name: It’s war

GARY MAGIC: Not so progressive: Trump’s politics aap into Alberta

MUST WATCH:

Try Globe Unlimited 99¢ per week for the first 4 weeks

Already a subscriber? Log In
Digital Media

- Dataset Characteristics:
  - Attribute: 246
  - Size: 2,842 GB

- Year 2014 (Jan-Jul):
  - Number of Records: 264,735,412
  - Number of Visits: 51,748,518
  - Number of Users: 19,760,853
Preprocessing

- Data Preprocessing and Cleaning
  - Filtering out irrelevant hits
  - Extracting the user types
  - Extracting the event of interest
  - Computing the time spent
  - Roll-up form hit to visit and the user
  - Converting to expressive format (i.e., json)
  - ...
Frequency-based News Recommendation

- News articles are not independent
  - Pattern (behavior): a list of visited articles
  - Finding sets of co-occurrence news articles

Web Clickstreams

Offline
- Discover **Frequent Patterns** from user logs
  - \(<\text{News1},\text{New2}>\)
- Extract rules from the patterns
  - \(\text{news1} \rightarrow \text{news2}\)

Phase 1

Phase 2

Online

Phase 3

- Recommend articles
Frequent Pattern Mining

- Frequent Pattern Mining (FPM)
  - FPM is a fundamental research topic in data mining
  - Example application
    - Discover sets of items (i.e., itemsets) that are frequently purchased together by customers

Minimum support threshold: 60%

\[
\text{Sup}(\{\text{Bread, Milk}\}) = \frac{4}{6} = 66.6\%
\]

\{\text{Bread, Milk}\} is a frequent itemset
Domain Driven Actionable Knowledge Discovery

- The identified patterns are handed over to business people

- They cannot interpret the patterns for business use
  
  - There are many patterns/not informative
  
  - Not interested to business needs.
  
  - How to interpret the patterns to business actions
Insufficiency of Frequent Pattern Mining

- In Market Analysis
  - Business objective: Increase Revenue
  - May lose infrequent but valuable patterns
  - May present too many frequent but unprofitable patterns
  - Cannot find patterns having high profits

Which patterns can bring high profits and make high revenue?
## A Motivation Example

<table>
<thead>
<tr>
<th>TID</th>
<th>Transaction</th>
<th>Item</th>
<th>Unit Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>{Bread(1), Milk(1)}</td>
<td>Bread</td>
<td>20</td>
</tr>
<tr>
<td>$T_2$</td>
<td>{Bread(1), Milk(1)}</td>
<td>Milk</td>
<td>30</td>
</tr>
<tr>
<td>$T_3$</td>
<td>{Bread(1), Milk(1), Diaper(3), Beer(6)}</td>
<td>Diamond</td>
<td>1,000</td>
</tr>
<tr>
<td>$T_4$</td>
<td>{Bread(1), Milk(1), Diaper(3), Beer(6)}</td>
<td>Necklace</td>
<td>300</td>
</tr>
<tr>
<td>$T_5$</td>
<td>{Diamond(1), Necklace(1)}</td>
<td>Diaper</td>
<td>300</td>
</tr>
<tr>
<td>$T_6$</td>
<td>{Diamond(1), Necklace(1)}</td>
<td>Beer</td>
<td>70</td>
</tr>
</tbody>
</table>

{Bread, Milk}: $200  
{Diamond, Necklace}: $2600  
{Diaper, Beer}: $2640
High Utility Sequential Pattern Mining

- Given a set of sequences: find all **sequences** whose **utility** is > a user-specified minimum threshold

  - Each item has quantity in a transaction
  - Each item has a value (e.g., price)

<table>
<thead>
<tr>
<th>Items</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milk</td>
<td>$3</td>
</tr>
<tr>
<td>Egg</td>
<td>$2</td>
</tr>
<tr>
<td>Birthday Cake</td>
<td>$20</td>
</tr>
<tr>
<td>Birthday Card</td>
<td>$10</td>
</tr>
<tr>
<td>Bread</td>
<td>$1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CID</th>
<th>TID</th>
<th>Transaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>T1</td>
<td>{(Bread,2), (Milk,6)}</td>
</tr>
<tr>
<td>C1</td>
<td>T2</td>
<td>{(Birthday Card,2)}</td>
</tr>
<tr>
<td>C1</td>
<td>T3</td>
<td>{(Birthday Cake,2), (egg,3)}</td>
</tr>
<tr>
<td>C2</td>
<td>T3</td>
<td>{(Birthday Cake,2), (Birthday Card,2), (Yoghurt,3), (Tuna,5)}</td>
</tr>
<tr>
<td>C2</td>
<td>T4</td>
<td>{(egg,5), (Pizza,4), (Juice,2)}</td>
</tr>
<tr>
<td>C3</td>
<td>T5</td>
<td>{(Birthday Cake,2), (Birthday Card,2), (Yoghurt,4), (Milk,3)}</td>
</tr>
<tr>
<td>C3</td>
<td>T6</td>
<td>{(Milk,1), (cheese,2)}</td>
</tr>
</tbody>
</table>
What is utility?

- Utility of item in a transaction = internal utility (quantity of items in the transaction) x external utility (profit of the item).
  - $U(\text{Milk}, T1) = 3 \times 6 = 18$

- Utility of itemset in a sequence = sum of utilities of its items:
  - $U(\{\text{Bread, Milk}\}, C1) = 2 \times 1 + 6 \times 3 = 20$

- Utility of sub-sequence in a sequence = sum of its itemsets’ utilities
  - $U(\langle \{\text{Milk}\}, \text{egg} \rangle, C1) = 3 \times 6 + 3 \times 2 = 24$
  - If more than one occurrence, then maximum value among occurrences

<table>
<thead>
<tr>
<th>Items</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milk</td>
<td>$3</td>
</tr>
<tr>
<td>Egg</td>
<td>$2</td>
</tr>
<tr>
<td>Birthday Cake</td>
<td>$20</td>
</tr>
<tr>
<td>Birthday Card</td>
<td>$10</td>
</tr>
<tr>
<td>Bread</td>
<td>$1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CID</th>
<th>TID</th>
<th>Transaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>T1</td>
<td>{(Bread,2), (Milk,6)}</td>
</tr>
<tr>
<td>C1</td>
<td>T2</td>
<td>{(Birthday Card,2)}</td>
</tr>
<tr>
<td>C1</td>
<td>T3</td>
<td>{ (Birthday Cake,2), (egg,3)}</td>
</tr>
<tr>
<td>C2</td>
<td>T3</td>
<td>{(Bread,2), (Milk,4), (Yoghurt,3), (Tuna,5)}</td>
</tr>
<tr>
<td>C2</td>
<td>T4</td>
<td>{(eggs,5), (Pizza,4), (Juice,2)}</td>
</tr>
<tr>
<td>C3</td>
<td>T5</td>
<td>{ (Bread,2),( Yoghurt,4), (Milk,3)}</td>
</tr>
<tr>
<td>C3</td>
<td>T6</td>
<td>{(Milk,1), (cheese,2)}</td>
</tr>
</tbody>
</table>
Problem 1: Actionability

- **Frequent patterns**
  - 75% frequency
  - 2 minutes

- **Actionable patterns**
  - 30% frequency
  - 12 minutes

News1

News2

News3

GAP

New s1

New s2

New s3

New s6

New s7

Morteza Zihayat
Problem 2: News cold-start

Recent news

Visited articles

Reading behavior

Newly-published articles
PENSYS: The Proposed Framework

**First Stage:** Utility-based recommendation rules

**Second Stage:** Topic-based recommendation rules
FIRST STAGE: NEWS LEVEL

Morteza Zihayat
Stage 1: Utility-based Pattern Mining

Engagement Measures
- Time Spent
- Cursor movement

Internal Utility Function (F)

External knowledge
- Editorial Board
- Freshness

External Utility Function (G)

Utility = \( U(F, G) \)

Patterns with Utility ≥ Threshold

Attractive news reading behavior

\[ Utility(news) = \text{Time spent} \times \text{Freshness} \]
Stage 1 Overview

- Discover High Utility Patterns from user logs
  - \(<\text{News1, New2}>\)

- Extract Utility-Based rules from the patterns
  - \(\text{news1} \rightarrow \text{news2}\)

- Recommendation

Web Clickstreams

Phase 1

Phase 2

Phase 3

Offline

Online
# Top-2 High Utility Patterns

<table>
<thead>
<tr>
<th>Num.</th>
<th>Set of news</th>
<th>Time (mins)</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Vigil held for daughter of Conservative Party president</td>
<td>2537</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>MH17: Disaster ratchets up Russia-Ukraine tensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Retiree, 60, wonders how long her money will last , Which is better, a RRIF or an annuity You may be surprised</td>
<td>1473</td>
<td>102</td>
</tr>
</tbody>
</table>
# Top-2 Frequent Patterns

<table>
<thead>
<tr>
<th>Num.</th>
<th>Set of news</th>
<th>Time (mins)</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Target faces calls to withdraw from Canada , Mike Duffy facing 31 charges</td>
<td>144</td>
<td>254</td>
</tr>
<tr>
<td></td>
<td>from Senate expenses scandal, RCMP says</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Florida police say , La Prairie, Quebec mayor dies from wasp stings</td>
<td>105</td>
<td>286</td>
</tr>
<tr>
<td></td>
<td>Canadian professor was killed in targeted attack</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Recommendation Rules

<table>
<thead>
<tr>
<th>Top 3 of our Utility-Based Association Rules, sorted by uconf, time spent in descending order</th>
<th>Time Spent (minute)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Researchers find dozens of genetic links to schizophrenia] (\Rightarrow) [Anti-Semitic graffiti found in Thornhill, days after Islamic centre defaced] [Video: Fearless 93-year-old tackles CN Tower EdgeWalk] [Teen use of human growth hormone on the rise] [Canadian poised to become terror tourist given 10-year sentence] [Ontario Liberals waste no time playing hardball with opposition] [The Gaza war has done terrible things to Israeli society]</td>
<td>239.98</td>
</tr>
<tr>
<td>[Researchers find dozens of genetic links to schizophrenia] (\Rightarrow) [Anti-Semitic graffiti found in Thornhill, days after Islamic centre defaced] [Teen use of human growth hormone on the rise] [Canadian poised to become terror tourist given 10-year sentence] [Ontario Liberals waste no time playing hardball with opposition] [The Gaza war has done terrible things to Israeli society]</td>
<td>239.93</td>
</tr>
<tr>
<td>[Researchers find dozens of genetic links to schizophrenia] (\Rightarrow) [Anti-Semitic graffiti found in Thornhill, days after Islamic centre defaced] [Video: Fearless 93-year-old tackles CN Tower EdgeWalk] [Canadian poised to become terror tourist given 10-year sentence] [Ontario Liberals waste no time playing hardball with opposition] [The Gaza war has done terrible things to Israeli society]</td>
<td>239.8</td>
</tr>
</tbody>
</table>
Performance

- Only 116000 records
  - Over 4 GB memory usage in average
  - Around 45 mins run time
Big Data Framework

Sequence Dataset

Initialization
Local HUSPs Mining
Global Candidate Generation
Global HUSPs Mining

Mapper 1
Reducer 1

Mapper 2
Reducer 2

Mapper 3
Reducer 3

Mapper n
Reducer n

Mapper 1
Reducer 1

Mapper 2
Reducer 2

Mapper 3
Reducer 3

Mapper n
Reducer n

Apply GCP to prune low Global Candidates

Global Candidates

Global HUSPs and their Utilities

Apache Spark

IBM Platform Symphony

Morteza Zihayat
# Results

\[ m = \text{Minutes}, h = \text{Hours} \]

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( \delta ) (%)</th>
<th>BigHUSP</th>
<th>BigHUSP\text{Basic}</th>
<th>BigHUSP\text{SA}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Globe</strong></td>
<td>0.09</td>
<td>1.6 m</td>
<td>3.6 m</td>
<td>0.99 h</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>2.3 m</td>
<td>4.4 m</td>
<td>1.4 h</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>3.1 m</td>
<td>6.6 m</td>
<td>2.2 h</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>5.0 m</td>
<td>11.0 m</td>
<td>3.3 h</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>9.2 m</td>
<td>20.7 m</td>
<td>4.5 h</td>
</tr>
<tr>
<td><strong>synthDS1</strong></td>
<td>0.05</td>
<td>3.0 m</td>
<td>10.0 m</td>
<td>1.1 h</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>4.26 m</td>
<td>14.4 m</td>
<td>1.2 h</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>6.23 m</td>
<td>17.9 m</td>
<td>1.6 h</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>9.9 m</td>
<td>27.2 m</td>
<td>1.9 h</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>14.3 m</td>
<td>29.4 m</td>
<td>3.2 h</td>
</tr>
<tr>
<td></td>
<td>0.009</td>
<td>37.6 m</td>
<td>76.5 m</td>
<td>7.8 h</td>
</tr>
<tr>
<td><strong>ChainStore</strong></td>
<td>0.09</td>
<td>15.0 m</td>
<td>33.4 m</td>
<td>6.4 h</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>19.9 m</td>
<td>56.2 m</td>
<td>12.0 h</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>25.0 m</td>
<td>77.0 m</td>
<td>13.4 h</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>34.8 m</td>
<td>107.7 m</td>
<td>14.6 h</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>38.7 m</td>
<td>159.8 m</td>
<td>17.4 h</td>
</tr>
<tr>
<td><strong>synthDS2</strong></td>
<td>0.09</td>
<td>13.1 m</td>
<td>26.3 m</td>
<td>7.7 h</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>16.3 m</td>
<td>34.5 m</td>
<td>9.3 h</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>20.6 m</td>
<td>47.2 m</td>
<td>15.7 h</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>23.8 m</td>
<td>51.8 m</td>
<td>17.8 h</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>32.2 m</td>
<td>85.3 m</td>
<td>21.4 h</td>
</tr>
</tbody>
</table>
SECOND STAGE: TOPIC LEVEL

Morteza Zihayat
PENSYS: The Proposed Framework

**First Stage: Utility-based recommendation rules**
- **Utility-based Reading Behavior Discovery**
- News reading behaviors
- **Rule Engine**
- Attractive article-based Rules

**Second Stage: Topic-based recommendation rules**
- **Probabilistic Language Model**
- Article distribution over topics
- **User Based Relational Topic Model**
- Attractive Topic based Rules

**News Corpus**

**Recommended Articles**
Stage 2: A Semantic Relational Topic Model

- News cold start problem
  - Content based recommendation systems
  - Similar topics
- User behavior?
  - Hybrid approach

<table>
<thead>
<tr>
<th>Topic 1: Health</th>
<th>Topic 2: Debt</th>
<th>Topic 3: Travel</th>
<th>Topic 4: Sport</th>
<th>Topic 5: Politics</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w, p(w</td>
<td>\text{Topic1}) )</td>
<td>( w, p(w</td>
<td>\text{Topic2}) )</td>
<td>( w, p(w</td>
</tr>
<tr>
<td>Drug, 0.5</td>
<td>Debt, 0.5</td>
<td>Hotel, 0.5</td>
<td>Soccer, 0.5</td>
<td>Obama, 0.5</td>
</tr>
<tr>
<td>Health, 0.4</td>
<td>Bond, 0.4</td>
<td>Travel, 0.4</td>
<td>Sport, 0.4</td>
<td>Politics, 0.4</td>
</tr>
<tr>
<td>People, 0.25</td>
<td>Credit, 0.25</td>
<td>Park, 0.25</td>
<td>NBA, 0.25</td>
<td>Election, 0.25</td>
</tr>
<tr>
<td>Disease, 0.15</td>
<td>Investors, 0.15</td>
<td>Mountain, 0.15</td>
<td>Ronaldo, 0.15</td>
<td>War, 0.15</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( nw_1 )</th>
<th>( nw_2 )</th>
<th>( nw_3 )</th>
<th>( nw_4 )</th>
<th>( nw_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t, p(t</td>
<td>nw_1) )</td>
<td>( t, p(t</td>
<td>nw_2) )</td>
<td>( t, p(t</td>
</tr>
<tr>
<td>Topic 1, 0.6</td>
<td>Topic 2, 0.7</td>
<td>Topic 3, 0.6</td>
<td>Topic 4, 0.8</td>
<td>Topic 5, 0.6</td>
</tr>
<tr>
<td>Topic 3, 0.2</td>
<td>Topic 3, 0.4</td>
<td>Topic 1, 0.4</td>
<td>Topic 5, 0.4</td>
<td>Topic 3, 0.4</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
</tbody>
</table>
Stage 2: A Semantic Relational Topic Model

Attractive News Reading Behavior
\[ p(nw_m|nw_n) \]

Topic model
\[ p(w_i|t) \]
\[ p(t_i|nw) \]

Aggregate

Semantic relational topic model
\[ p(t_i|t_j) \]
Topic-based Rules Example

- University/College (0.24)
  - Conf: 70%
  - Conf: 90%
  - Conf: 81%
  - Confidence: 82%

- Bollywood (0.24)
  - Conf: 74%
  - Conf: 67%

- Motor Racing (0.35)

- Homelessness (0.40)

- Urban Planning/Development (0.34)

- Armed Forces (0.68)

- Economic Growth (0.23)

- Election

- Sport
Outline

- Introduction to Big Data
- User Modeling in Digital Media
- Depression Acuity Detection
- Cogniciti: An Online Brain Health Assessment
- Conclusion
Depression

- Recalling events
- Affected by current mental state
- Low quality of data

Problems:

Diagnosed by Screening questionnaire
Depression Acuity Detection

- Mental state
- Physical wellness

Learning models
- Sooner
- More Accurate
Proposed Framework

Apache Spark

Learning Phase
- Type
- Time stamp
- Duration

Learning Models
- Physical behavioral patterns
- Language model
- Emotion model

Students

Clinicians

- Linguistic styles
- Sentiment analysis

Social Media - Type - Time stamp - Duration - Linguistic styles - Sentiment analysis

18/11/2016
Morteza Zihayat
Results

1. Depression terms
   - 65% higher

2. Mood classification
   - 69% accuracy

3. Physical wellness indicators
   - Time
   - Duration
   - Sequential order of events matters
Outline

- Introduction to Big Data
- User Modeling in Digital Media
- Depression Acuity Detection
- Cogniciti: An Online Brain Health Assessment
- Conclusion
Brain Health

- **Motivation**
  - 47 million people suffer from dementia (Alzheimer’s Society of Canada)
  - Early detection of dementia: $219 billion saving

- **Solution**
  - An early warning test for brain health
Assessments

Demographic Test

Shape match test

Stroop interference

Unified Score

Letter-number alternation

Face-name association
Strong consumer response from one Canadian media release

- Test re-test reliability: 72%
- Site Visits: 153,000
- Completed Assessments: 41,000
Conclusions

- User behavior analysis in big data is an important area of research

- We have done some (hopefully) interesting work in this area
  - Utility-based pattern discovery in big data streams
  - Depression acuity detection
  - Online Brain Health Assessment

- A lot more research needs to be done!
User Behavior Analysis in Big Data

mori.zihayatkermani@utoronto.ca

Morteza Zihayat