Using Social Media To Understand Causal Relationships

Alankrit Johari
Causal Knowledge

Very important and hard to find.

Why important?

- To analyze claims like “Playing Video games result in violent behavior in children”

Why hard?

- Need domain expertise
Why do we need Causal Knowledge?

PREDICTION  EXPLANATION  INTERVENTION

Smoking  
Lung Cancer Rate  
Medication  
Side Effects  
New Product Release  
Higher Stocks
What I want to do?

Main focus: medical claims

Examples:

- Meditation is good for concentration.
- Watching too much television harms your eyes.
- Playing violent video games causes violent behavior in children.

Basically, claims that are myths and sounds true.
Right Variables

- What are we finding causes between?
- It can be in format: (Cause - Problem), (Problem - Cause)
- Not necessarily related to problems: (Solution - Effect)
- Example: “Mediation is good for Concentration”
  - Meditation - Solution
  - Concentration - Effect
Text Mining

There is so much information floating around on the web.

For these medical claims, people write:

- Blogs,
- Answers (e.g. Yahoo Answers, Quora),
- etc.

Use text mining techniques to analyze this text and understand correlations and verify causes.
Co-occurrence and Network Analysis Method

- I will plot the strength of co-occurrence on a graph.
- The closer the two variables the stronger is the correlation.
- The further the two variables the weaker is the correlation.
- The correlation will be based on count of co-occurrence over the web.
- Some weightage to the higher reliable sources?
Data

- Available public API’s
  - Yahoo Answers!
  - Blogger

- Information extraction easy from these public API’s

- Data extraction from Quora requires use of scraping tools
  - Selenium (for scraping)
  - BeautifulSoup (for scraping)
  - Mechanize (for navigating)
Verification

- Will compare the resulting graphs with various experimental studies.
- Correlation to Causal mapping
- Check results for highly unlikely pairs, eg. (good sleep - bad result on exams)
Challenges

- Representation of data?
- What about negative correlations.
- How closes to causes? Just correlations?
- Relevant data?
References

