Web-based Inference Detection


Richard Chow
Philippe Golle
Jessica Staddon
PARC
UNCLASSIFIED//FOR OFFICIAL USE ONLY

(U//FOUO) The family

(U//FOUO) is a member of a large and wealthy Saudi family. The family patriarch came to the kingdom from Hadramout (South Yemen) sometime around 1930.¹

- In Saudi Arabia, the father became a construction magnate, completing prestigious projects such as the renovation of the holy mosques in Mecca and Medina. As a result, the family are a highly respected family both within the Saudi royal household and with the public.

(U//FOUO) There is some confusion as to the total number of siblings.

- Some cite that he is the youngest of some 20 sons,² while others claim he is the seventh son.³

- The total number of his siblings might be 50,⁴ 52,⁴ or 54.⁴ In an interview seemed unsure as well, citing that he had 25 brothers—although he could remember the names of only 20.⁷

- Nearly all of these siblings are half-brothers or half-sisters, as the father had multiple wives.⁵ He is cited as having only one son.⁸

(U//FOUO) The family has denounced repeatedly.

¹
²
³
⁴
⁵
⁶
⁷
⁸
Web search on: “sibling saudi magnate”

1. Bin Laden Sibling Gives DNA to CIA
   President Bush on Thursday ordered the CIA to make public the DNA profile of Osama bin Laden's brother, whom U.S. officials believe lived for years in Riyadh, Saudi Arabia.
   www.freerepublic.com/forum?news=691245/posts - 6k - Cached - More from this site

2. Sibling to defend Bin Laden if needed (I will pay for Osama's defence)
   ... and daughters of the late Saudi construction magnate Mohammed bin Laden, who had ... spent some time with Osama in Saudi Arabia between 1978 to 1981 before ...
   www.freerepublic.com/forum?news=1436396/posts - 19k - Cached - More from this site

3. NewsMax.com Inside Cover Story
   A sibling of Osama bin Laden has reportedly given the CIA a sample of ... Alia Ghanem and the ... Saudi construction magnate Mohammed bin Laden, he has ...
   www.newsmax.com/scripts/showinside.pl?a=2001/12/16/61344 - 10k - Cached - More from this site

4. Terrorist Attack - Information Only : Bin Laden Group
   ... of several firms approached by the Saudi-based conglomerate which is seeking to ... one of 57 ... Saudi construction magnate Mohammed bin Laden. ...
   www.suite101.com/discussion.cfm/investing/66790/565256 - 27k - Cached - More from this site

5. Terrorist 9/11 Attack Discussion 2,000 +
   ... 54 children of a Saudi Arabian construction magnate and reportedly has look ... A sibling of ... Osama bin Laden has reportedly given the CIA a sample of his DNA to ...
   www.suite101.com/discussion.cfm/investing/66773/843-852 - 118k - Cached - More from this site

6. Bin Laden Family Business Seeks To Improve Image
   ... firms approached by the Saudi-based conglomerate which is ... Osama bin Laden as one of 57 children of the late construction magnate Muhammad bin Laden. ...
   www.tensa.com/general17/binladenfamily.htm - 4k - Cached - More from this site
Observations

- Most web pages with terms “sibling saudi magnate” also contain terms “osama bin laden”
- Hence, deduce the inference:
  \{sibling saudi magnate\} $\rightarrow$ \{osama bin laden\}
- Get most valid inferences, since the Web is a proxy for all human knowledge
  - Not complete though!
- Idea: Deduce inferences from co-occurrence of terms on the Web
Conceptual Framework

• Consider any Boolean formula of terms, e.g.
  (saudi AND magnate AND sibling),
  (osama AND bin AND laden)

• Evaluates to TRUE or FALSE for each Web page
  – Or, for each paragraph in each Web page...

• Strength of inference: Conditional Probability
  – Given (PRECEDENT) is TRUE, what is probability that
    (CONSEQUENT) is TRUE?
  – Write: (PRECEDENT) IMPLIES (CONSEQUENT)

• From now on, restrict to special case: Conjunction of terms
  implying another conjunction of terms
  – Other cases may be of interest as well:
    (xxx) IMPLIES (Person1 OR Person2 OR …)
Traditional Association Rules

- **Problem:** Find market items that are commonly purchased together
  - Rules are of the form: \((A) \implies (B)\), \(A\) and \(B\) are sets of items
  - Legendary example: \((\text{diapers}) \implies (\text{beer})\)
- **Confidence of a rule:** \(\Pr (B | A)\)
  - Given that \(A\) is purchased, how likely is \(B\) to be purchased?
- **Support of a rule:** \(\Pr (A \text{ and } B)\)
  - What portion of all purchases contain both \(A\) and \(B\)?
- **Apriori (Agrawal et al):** well-known algorithm for this problem
  - Works for given confidence and support cutoffs
Web Association Rules

• Our problem: Find terms that are commonly found together on web pages

• Key differences from traditional association rules
  – Web is very large and unstructured
  – Natural Language Processing (NLP) may provide additional information since we are mining terms from text
  – More complex rules are of interest
    • Boolean formulae such as (A) IMPLIES (B OR C)
    • Linguistic patterns such as (a followed b) IMPLIES (C)

• Note that for privacy applications, need to find rules with very low support
  – Apriori algorithm not directly useful
Using search engines to estimate probabilities
Another Way

Probability is about 81/234
HIV Precision: Top 60 Inferences

• Precision: fraction of “correct” inferences produced
• Analyzed top precedents appearing in at least 100K documents
• Medical expert reviewed these inferences
  – 28 were “correct”
  – 3 not necessarily connected to HIV, but were related conditions
  – 29 unknown or did not indicate HIV
• Medical expert appropriate for medical records - note that appropriate reviewer depends on the application
  – “Montagnier” not considered “correct”, but was discoverer of the HIV virus
  – “Kwazulu” not considered “correct”, but this province of SA has one of the highest HIV infection rates in the world

| 1.049 | antiretroviral | 0.844 | lopinavir | 0.662 | chancroid | 0.604 | viremia |
| 0.969 | pmctt | 0.84 | microbicides | 0.658 | proviral | 0.594 | lymphotropic |
| 0.967 | cd4-gp120 | 0.834 | lipodystrophy | 0.651 | retroviruses | 0.583 | lymphohgranuloma |
| 0.964 | gp41 | 0.829 | unaids | 0.651 | pneumocystis | 0.578 | hemophiliacs |
| 0.953 | nrtis | 0.788 | saquinavir | 0.649 | cd4-positive | 0.578 | mtc |
| 0.953 | ccr5 | 0.784 | abacavir | 0.642 | ccr4 | 0.577 | lentivirus |
| 0.939 | nnrtis | 0.784 | ddf | 0.640 | htlv | 0.573 | trichomoniasis |
| 0.927 | efavirenz | 0.782 | didanosine | 0.638 | seroconversion | 0.557 | gonorrhoea |
| 0.922 | 8-antiretroviral | 0.779 | indinavir | 0.635 | acquire | 0.557 | seronegative |
| 0.902 | gp120 | 0.775 | nevirapine | 0.623 | retroviral | 0.538 | contagiosum |
| 0.892 | zidovudine | 0.758 | cointection | 0.619 | integrase | 0.535 | nucleoside |
| 0.888 | azt-3tc | 0.749 | virologic | 0.616 | anti-p24 | 0.519 | montagnier |
| 0.878 | aidsinfo | 0.741 | ritonavir | 0.610 | carinii | 0.517 | mononucleosis |
| 0.866 | syndr | 0.693 | immunodeficiency | 0.608 | kwazulu | 0.514 | postexposure |
| 0.854 | cd4-ccr5 | 0.662 | coreceptor | 0.607 | seropositive | 0.513 | gonococcal |
Inference Problem

• More and more publicly available data
  – Web 2.0 technologies becoming common
  – “long tail of the Internet”
• How to control the release of data?
  – What does the data reveal?
  – Need automated techniques
• Scenarios:
  – Individuals
    • Anonymous blogs or postings
    • Redaction of medical records
  – Corporations
    • News releases
    • Identification of content representing risk
  – Government
    • Declassification of government documents