A Mass Assignment Approach to Granular Association Rules for Multiple Taxonomies

Trevor Martin, Yun Shen and Ben Azvine

AI Group
Engineering Mathematics
University of Bristol

Intelligent Systems Research Lab,
BT

Currently : *Senior Research Fellow*,
Intelligent Systems Research Lab, BT

URSW, November 12, 2007

Overview of presentation

- what is the problem ?
- research background and the need for fuzziness
- mass assignment-based approach
- demonstrator application
- summary

The aim is to give a high level overview, not to describe the specific methods step-by-step.
For detail, see the paper or talk offline.

Digital obesity needs machine assistance

- **Digital obesity (personal and corporate)**
  - Britons carry an average of 20GB of data in their pockets (toshiba)
  - estimated 5 exabytes of new data produced globally in 2002, much as text (sims)
  - 15% of enterprise info is in a core database, the rest is on lap/desk tops (HP labs)
  - 80% of a typical "new" document is recycled from existing documents (HP labs)
  - *there is a need for automatic assistance in managing this information*

- **Machine-based solutions must be understandable …**

  "never trust anything that thinks if you can't see where it keeps its brain" *
  - google uses syntactic features (presence of words, links, …)
  - computational linguistics / natural language processing aims to use grammar and deep structure / meaning
  - humans use natural language - machines cannot understand NL but can process it
  - engineering approach - be consistent with (fuzzy) humans

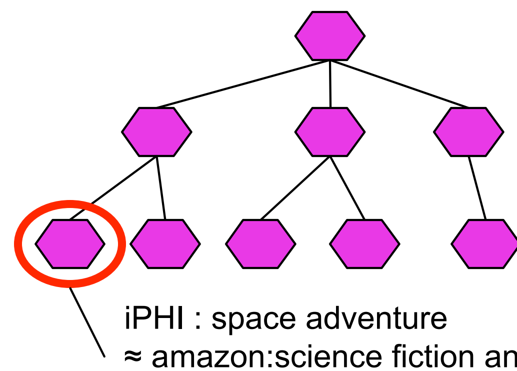
* *Arthur Weasley , UK Ministry of Magic - Harry Potter and the Chamber of Secrets*

Four Steps

Summary : approximate relations (class - class)

Most films in the "space adventure" category are directed by Lucas or Spielberg

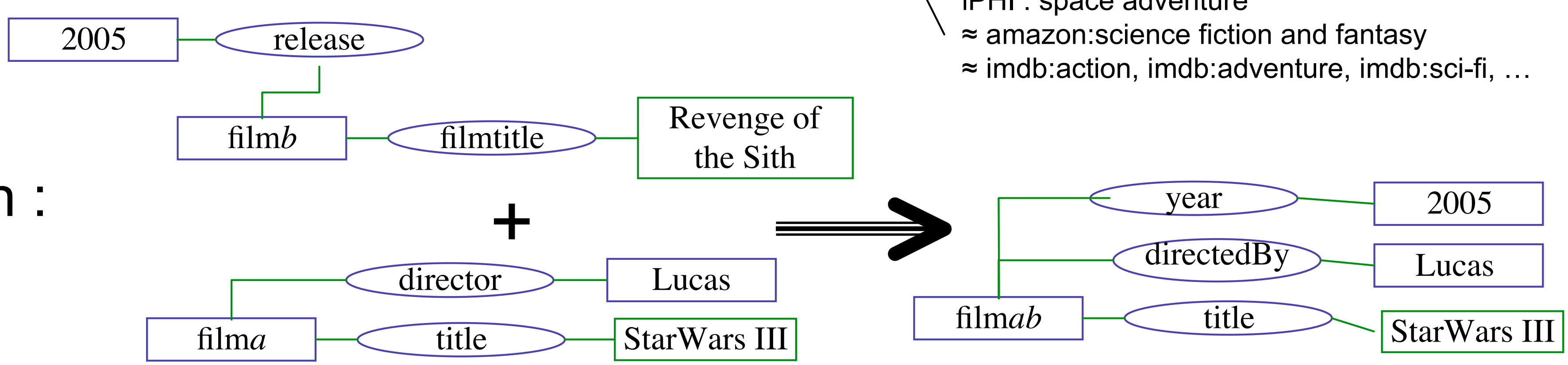
Organisation / Granulation: iPHI



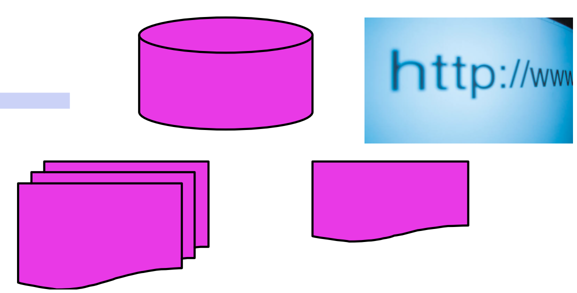
iPHI : space adventure
≈ amazon:science fiction and fantasy
≈ imdb:action, imdb:adventure, imdb:sci-fi, …

Fusion :

SOFT

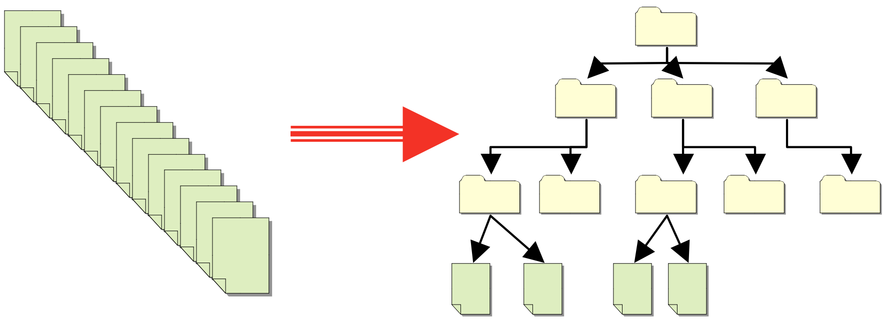


pre-processing : fuzzy grammars, entity tagging



Granulation and Hierarchies

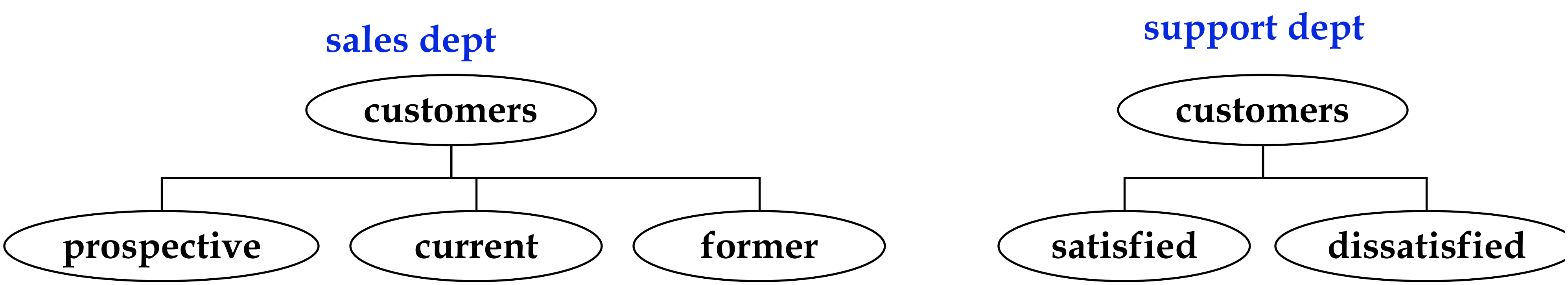
- hieararchical organisation is widespread



- each category in the hierarchy is a granule
    - e.g. shapes, geography, wine, species, books, movies, documents, …
- There is rarely a unique hierarchy, and categories are rarely crisp
- leaf nodes are often categorical values in a database

Equivalence of hierarchies

- it is helpful to know how different hierarchies (views) are related
  - enables reuse of categorised information
  - enables combination of information from different sources



- "DL approach" based on binary logic, requires equivalence / strict subset ... $\forall$, $\exists$

- better approach - look for strong associations
  - "most current customers are satisfied customers"

Association Rules

- most prominent use - transaction analysis
  - how often do customers buy cheese or crackers when they buy wine or beer?

internet retailers : "*customers who purchased A and B also purchased Y and Z*"

	beer	wine	cheese	crackers	milk	…
1	✓			✓	✓	✓
2		✓	✓	✓		
3		✓	✓		✓	✓
…	…	…	…	…	…	…

X = set of items

Tr = set of transactions

$\underbrace{\text{beer, wine}}_{s = \{beer,wine\}}$ $\underbrace{\text{cheese, crackers}}_{t = \{cheese,crackers\}}$

s = {beer,wine} t = {cheese,crackers}

$S, T \subseteq Tr$ are (multi-)sets of transactions containing items in s, t respectively

$$Support(S,T) = |S \cap T| \qquad Conf(S,T) = \frac{|S \cap T|}{|S|}$$

Fuzzy Categories

- words mean what we agree that they mean
  - *wicked, cool, bling, chav, rubbish*
  - light snowfall, bright colour, rock music
  - what makes a White Christmas in the UK? "*That one flake of snow will fall on Met office monitoring stations over the 24 hr period of the 25th of December*"

    www.mybetting.co.uk/white-christmas-betting.htm

- communication is made more efficient by use of loose definitions
- over-precision is not user-friendly- it needs soft definitions
  - most relations / categories / attributes / ... are not defined by unbreakable rules,
  - data can be missing, inconsistent, unreliable, ...
  - should we adapt our thinking to the computer *or*

adapt the computer to our thinking

An "illogical" hierarchy - dairy-free spread

Bakery
Beers, Wines, Spirits
Beverages, Hot Drinks
Breakfast Cereals
Clothing
Confectionery, Biscuits, Cakes
Cooking/Baking Ingredients
Crisps, Nuts, Snacks
Dairy
Delicatessen
Easter Confectionery
…
Pickles, Preserves, Oils, Spreads
…

Animal and Vegetable Fats
Artificial Sweeteners
Colouring and Decoration
Custard and Cornflour
Dried Fruit
Flour - Other
…

Cheese - American
Cheese - Canadian
…
Cheese - Snacking
…
Spreads (Butter, Margarine, etc.)
…
Yogurt - Twin Pots

No luck …

Dairy Free Spread 500g

dairy-free classified as *dairy*.
Not logic!
But it works.

www.sainsburystoyou.com - 2005

Fuzzy Sets in Information Systems

- Logic approaches are too “black and white”
  - data must be known precisely or not at all (confusion over NULL values)
  - the real world is rarely as neat and tidy as this
  - uncertainty in data values -
    e.g. John is *quite heavy*, John’s car is a *small hatchback*
  - uncertainty in relations -
    e.g. John hates Microsoft Word *a lot*
  - both
    e.g. driving in a *small car* at *high speed* is *very uncomfortable*
  - uncertainty in deduction
    a car *usually* has *high* running costs if its list price is *expensive*

- how can the logical model be extended ?
  - incorporation of uncertainty in attribute values, relations, rules, inference, and queries
  - fuzzy - lack of adequate definition e.g. what is meant by *large number*
  - probabilistic - lack of information e.g. will dice score be 6 ?

Possibility Distributions
vs
Monadic Fuzzy Relations

- speed of car is *fast* == μ_{fast} *(70) : speed of car is 70 mph OR*
  μ_{fast} *(71) : speed of car is 71 mph OR ...*

- John is speaking *L* == μ_L *(spanish) : John is speaking spanish OR*
  μ_L *(portugese) : John is speaking portugese OR ...*
  - single value, not known precisely

- legal speed is *about55* == χ_{a55} *(55) : legal speed is 55 mph AND*
  χ_{a55} *(56) : legal speed is 56 mph AND ...*

- John speaks *L* fluently == χ_L *(spanish) : John speaks spanish AND*
  χ_L *(portugese) : John speaks portugese AND ...*
  - multiple values, all satisfy predicate to some degree

Fuzzy Association Rules

$$Support(S,T) = |S \cap T| \qquad Conf(S,T) = \frac{|S \cap T|}{|S|}$$

- **straightfoward fuzzification**
  - **allow *S, T* to be fuzzy (multi-)sets (monadic relations)**
  - **use t-norm (min) for intersection and sigma-count for cardinality**

$$Conf(S,T) = \frac{\sum_{x \in X} \mu_{S \cap T}(x)}{\sum_{x \in X} \mu_S(x)}$$

but

$$S = [x_1/1, x_2/0.01, x_3/0.01, \ldots, x_{1000}/0.01]$$
$$T = [x_1/0.01, x_2/1, x_3/0.01, \ldots, x_{1000}/0.01]$$

$$Conf(S,T) = \frac{1000 \times 0.01}{1 + 999 \times 0.01} \approx 0.91$$

name	sales	salary
a	100	1000
b	80	400
c	50	800
d	20	700

S = goodSales = [a/1, b/0.8, c/0.5, d/0.2]
T = highSalary = [a/1, b/0.4, c/0.8, d/0.7]

$$Conf(S,T) = \frac{1 + 0.4 + 0.5 + 0.2}{1 + 0.8 + 0.5 + 0.2} = 0.84$$

[Martin-Bautista, M. J., M. A. Vila, H. L. Larsen, and D. Sanchez, "Measuring Effectiveness in Fuzzy Information Retrieval," presented at Flexible Query Answering Systems (FQAS), 2000]

Mass assignment-based approach

Fuzzy set $A = \{a/1, b/0.8, c/0.3, d/0.2\}$

$\Rightarrow$ *alpha cuts* $\{a\}/1$, $\{a, b\}/0.8$, $\{a, b, c\}/0.3$, $\{a, b, c, d\}/0.2\}$

$\Rightarrow$ *mass assignment*

$$M(A) = \{\{a\}:0.2, \{a,b\}:0.5, \{a,b,c\}:0.1, \{a,b,c,d\}:0.2\}$$

- **interpretation (as a possibility distribution)**
  - **value is in** *{a}* **with probability mass 0.2**
  - **value is in** *{a, b}* **with probability mass 0.5**
  - **etc**

  Mass can be distributed between elements of a set (restriction) - may not correspond to fuzzy set
  - **e.g.** $\{a, b, c, d\} : 0.2 \Rightarrow \{a, b, c\} : 0.1$ *and* $\{a, c\} : 0.1$

$$M_R(A) = \{\{a\}:0.2, \{a,b\}:0.5, \{a,b,c\}:0.2, \{a,c\}:0.1\}$$

  - **Combination of 2 mass assignments - re-distribute mass in a way that is consistent with both original mass assignments**

- **cardinality** $$p(|A| = n) = \sum_{\substack{A_i \subseteq A \\ |A_i| = n}} m_A(A_i)$$

$p(|A| = 1) = 0.2$, $p(|A| = 2) = 0.5$, etc

Least prejudiced distribution (LPD)

-split mass equally between elements

$\{a, b, c, d\} : 0.2 \Rightarrow \{a\} : 0.05, \{b\} : 0.05\}$ *etc*

Mass Assignment Association Rules

S = goodSales = [a/1, b/0.8, c/0.5, d/0.2]   *T = highSalary = [a/1, b/0.4, c/0.8, d/0.7]*

		0.2	0.1		0.3			0.4			
		a	*a*	*ac*	*a*	*ac*	*acd*	*a*	*ac*	*acd*	*abcd*
0.2	*a*	0.2									
0.3	*a*		0.1								
	ab		0.1					0.2			0.2
0.3	*a*				0.3						
	ab										
	abc				0.3						
0.2	*a*							0.2			
	ab										
	abc										
	abcd							0.2			

Assign mass in a way consistent with both components

Many possibilities !

Calculate support, confidence in standard (crisp) way

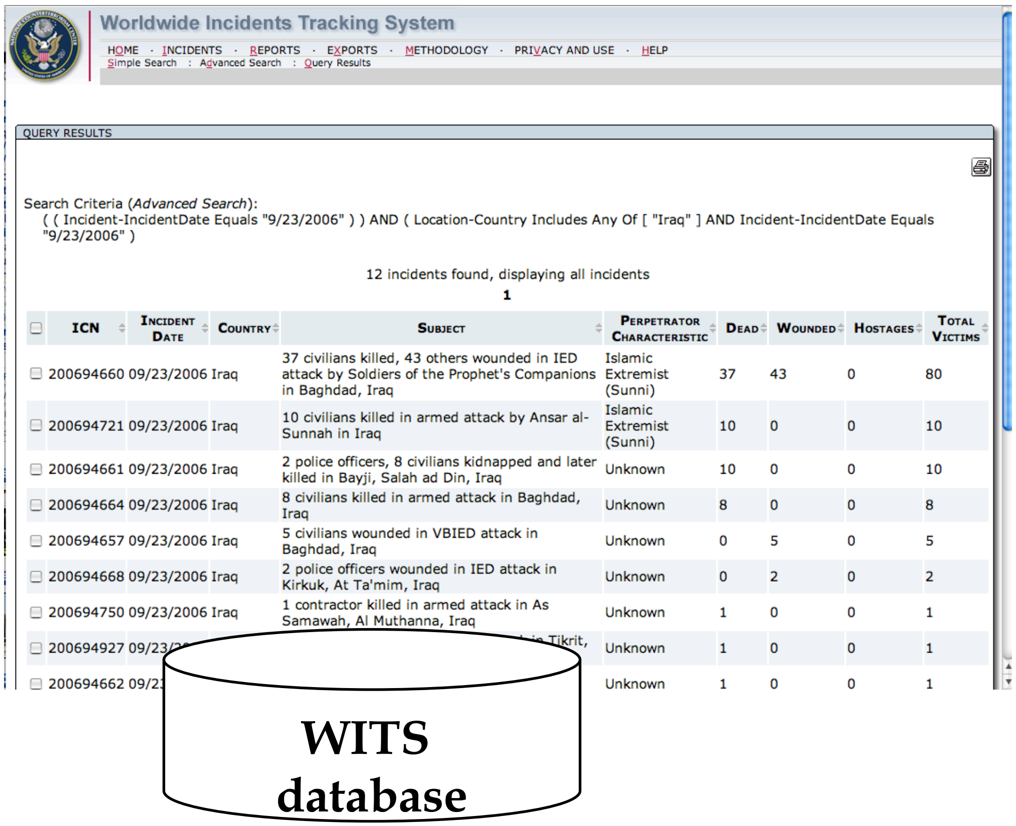
min and max give interval [0.4, 1] in this case   (potentially expensive computation)

Alternative - use least prejudiced distribution - very fast (see paper)

Demonstrator Application

Sample questions:

"who are the most active terrorist groups in countries near to iran?"
"who are the main targets, and has that answer changed recently?"



Search Criteria (*Advanced Search*):
((Incident-IncidentDate Equals "9/23/2006")) AND (Location-Country Includes Any Of ["Iraq"] AND Incident-IncidentDate Equals "9/23/2006")

12 incidents found, displaying all incidents

1

ICN	Incident Date	Country	Subject	Perpetrator Characteristic	Dead	Wounded	Hostages	Total Victims
200694660	09/23/2006	Iraq	37 civilians killed, 43 others wounded in IED attack by Soldiers of the Prophet's Companions in Baghdad, Iraq	Islamic Extremist (Sunni)	37	43	0	80
200694721	09/23/2006	Iraq	10 civilians killed in armed attack by Ansar al-Sunnah in Iraq	Islamic Extremist (Sunni)	10	0	0	10
200694661	09/23/2006	Iraq	2 police officers, 8 civilians kidnapped and later killed in Bayji, Salah ad Din, Iraq	Unknown	10	0	0	10
200694664	09/23/2006	Iraq	8 civilians killed in armed attack in Baghdad, Iraq	Unknown	8	0	0	8
200694657	09/23/2006	Iraq	5 civilians wounded in VBIED attack in Baghdad, Iraq	Unknown	0	5	0	5
200694668	09/23/2006	Iraq	2 police officers wounded in IED attack in Kirkuk, At Ta'mim, Iraq	Unknown	0	2	0	2
200694750	09/23/2006	Iraq	1 contractor killed in armed attack in As Samawah, Al Muthanna, Iraq	Unknown	1	0	0	1
200694927	09/23/[illegible]	[illegible]	[illegible] Tikrit,	Unknown	1	0	0	1
200694662	09/2[illegible]	[illegible]	[illegible]	Unknown	1	0	0	1

US National Counterterrorism Center
Worldwide Incident Tracking System

BAGHDAD, Iraq, Sept. 23, 2006 (Page 1 of 2)

Iraqi civilians inspect the site of a bomb explosion in Baghdad's poor neighborhood of Sadr City Sept. 23, 2006. **(AFP/Getty Images)**

(CBS/AP) A bomb claimed by a little-known Sunni Arab extremist group killed at least 37 Shiites and wounded another 40 in Baghdad on Saturday as they stocked up fuel for Ramadan, just days after the U.S. military warned sectarian violence would surge during the Islamic holy month.

Al Qaeda in Iraq's leader also reappeared in an old video posted on the Internet just as Sunni Arabs declared the start of Ramadan.

QUOTE

"I went into the flames just to get anyone left out of the fire. I saw a mother holding her child, both of them burned and dead."

Dhiyaa Ali, 24
Sadr City resident

Abu Ayyoub al-Masri, also known as Abu Hamza al-Muhajer, was shown killing a Turkish hostage — a possible signal to his Sunni Arab followers. In early September, he had called on Sunnis to step up attacks against American forces.

WHAT DO YOU THINK?

Go To Comments

Iraq's armed forces said they made some headway against groups affiliated with al Qaeda in Iraq, announcing the arrest of a senior leader of Ansar al-Sunnah — a radical Sunni group responsible for deadly attacks against U.S. forces, kidnappings and beheadings.

A Sunni extremist group, Jamaat Jund al-Sahaba — or Soldiers of the Prophet's Companions — claimed responsibility for the attack against Shiites in Sadr City, a sprawling slum that is home to more than two million people and a stronghold

...g to retaliate for a Friday attack by a ... and mosques that killed four people in a

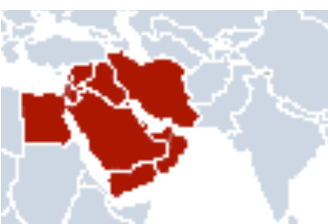
...ng sectarian violence that has forced

Simple iPHI hierarchy

incidents in all regions

middle east | south asia | etc

Iran, Iraq, …



Afghanistan, Pakistan, India, …

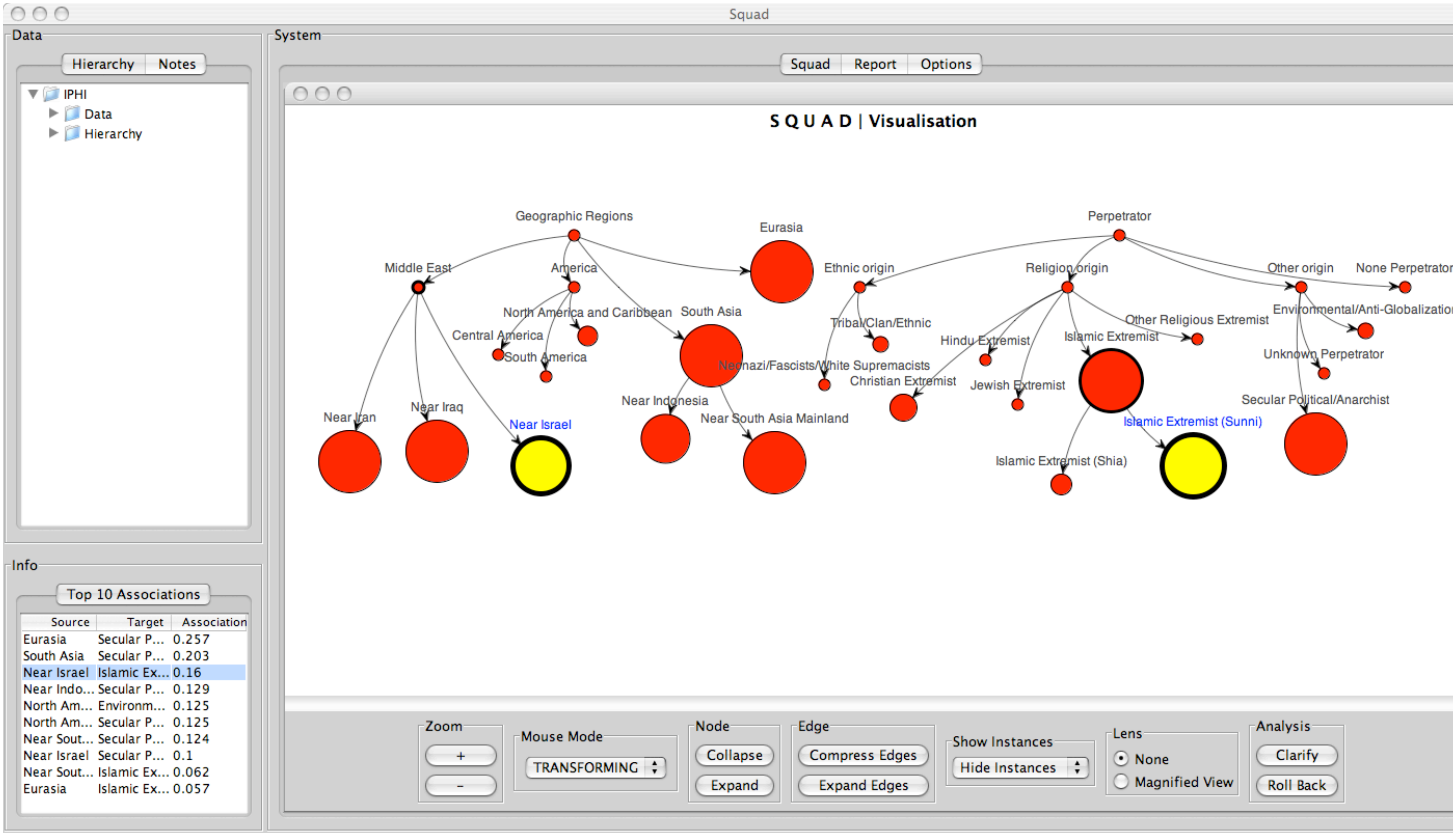
source data uses a given hierarchical categorisation - iPHI transforms to a more convenient (and fuzzy) sub-division

incidents in all regions

near Iran | near Pakistan | etc.

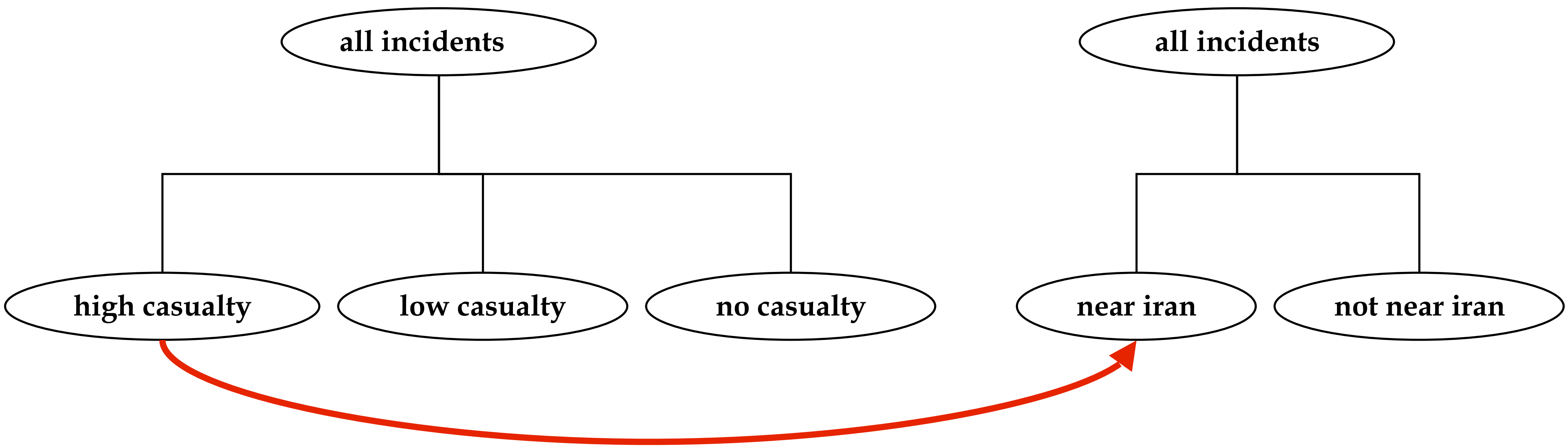
JAXB converts XML representation of category nodes and instances into executable code (java)

Demo


Squad
Data
Hierarchy
Notes
IPHI
Data
Hierarchy
System
Squad
Report
Options
S Q U A D | Visualisation
Geographic Regions
Eurasia
Perpetrator
Middle East
America
Ethnic origin
Religion origin
Other origin
None Perpetrator
North America and Caribbean
South Asia
Central America
South America
Tribal/Clan/Ethnic
Other Religious Extremist
Hindu Extremist
Islamic Extremist
Unknown Perpetrator
Neonazi/Fascists/White Supremacists
Christian Extremist
Jewish Extremist
Secular Political/Anarchist
Near Iran
Near Iraq
Near Israel
Near Indonesia
Near South Asia Mainland
Islamic Extremist (Sunni)
Islamic Extremist (Shia)
Info
Top 10 Associations
Source Target Association
Eurasia Secular P... 0.257
South Asia Secular P... 0.203
Near Israel Islamic Ex... 0.16
Near Indo... Secular P... 0.129
North Am... Environm... 0.125
North Am... Secular P... 0.125
Near Sout... Secular P... 0.124
Near Israel Secular P... 0.1
Near Sout... Islamic Ex... 0.062
Eurasia Islamic Ex... 0.057
Zoom
+
-
Mouse Mode
TRANSFORMING
Node
Collapse
Expand
Edge
Compress Edges
Expand Edges
Show Instances
Hide Instances
Lens
None
Magnified View
Analysis
Clarify
Roll Back

Summary

- soft relation = extended form of association rule
  - transaction analysis - e.g. do customers generally buy crisps or nuts when beer, lager or wine are bought?
  - rule S ⇒T (e.g. S = beer/lager/wine T = crisps/nuts), rule confidence is the conditional probability of *T* given *S*
  - extended to cope with fuzzy categories S, T e.g. do customers generally buy *salty snacks* when *alcoholic drinks* and *soft drinks* are bought?
  - fast calculation if we use *lpd*
  - *now looking at changes in association strength over time*



Thank you for your attention