

Scalable Uncertainty Management

01 – Introduction

Rainer Gemulla

April 20, 2012

Information & Knowledge Management Circa 1988

Databases

SQL

Datalog

Free text

Information retrieval

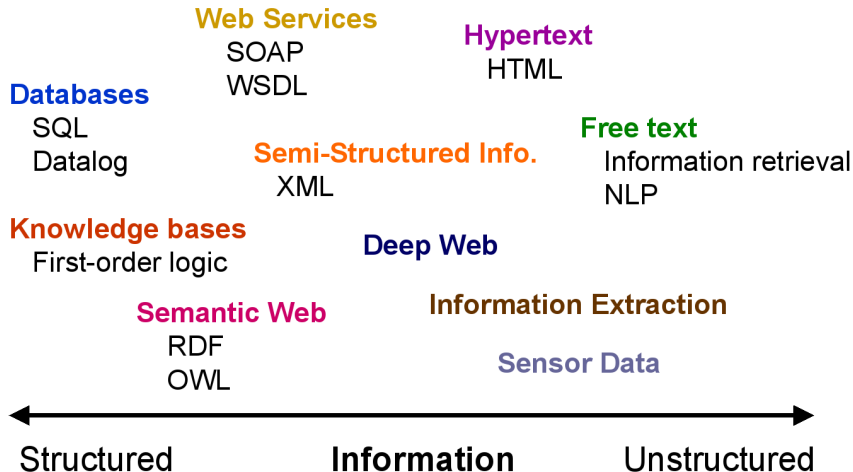
NLP

Knowledge bases

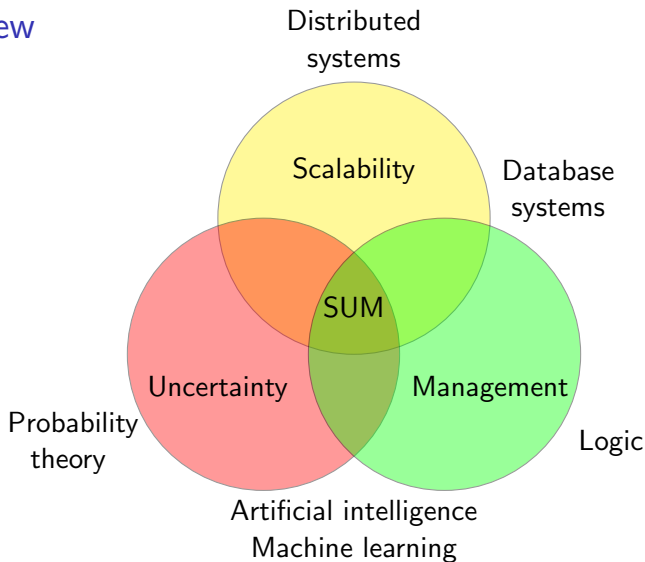
First-order logic



Information & Knowledge Management Today



Overview



SUM is about managing large amounts of uncertain data.

Outline

1 Uncertainty in the Real World

2 Managing Uncertainty

Sources of uncertainty

Certain data	Uncertain data	
The temperature is 25.634589 °C.	Sensor reported 25 ± 1 °C.	Precision of devices
Bob works for Yahoo.	Bob works for Yahoo or Microsoft.	Lack of information
MPH is located in Saarbrücken.	MPH is located in Saarland.	Coarse-grained information
Mary sighted a finch.	Mary sighted either a finch (80%) or a sparrow (20%).	Ambiguity
It will rain in Saarbrücken tomorrow.	There is a 60% chance of rain in Saarbrücken tomorrow.	Uncertainty about future
John's age is 23.	John's age is in [20,30].	Anonymization
Paul is married to Amy.	Paul is married to Amy. Amy is married to Frank.	Inconsistent data

Where does uncertainty arise?

Everywhere!

- Information extraction (D5 research)
- Sensor networks
- Business intelligence & predictive analytics
- Forecasting
- Scientific data management
- Privacy preserving data mining
- Data integration
- Data deduplication
- Social network analysis

Entity disambiguation (AIDA)

Disambiguate each mention of an entity in a piece of text.

Disambiguation Method:

☐ prior ☐ prior+sim ☒ prior+sim+coherence (graph)

Parameters: (default should be OK)

Similarity Impact:

Ambiguity degree:

Coherence threshold:

Mention Extraction:

☒ Stanford NER ☐ Manual

You can manually tag the mentions by putting them between `[]` and `||`. HTML Tables are automatically disambiguated in the manual mode.

Alexandria is an ancient city on the Mediterranean. It was famous for its lighthouse, one of the seven world wonders.

Input Type: TEXT

[Alexandria]Alexandria is an ancient city on the [Mediterranean Sea]Mediterranean. It was famous for its lighthouse, one of the seven world wonders.

Run Information Graph Removal Steps

0: Alexandria (solved by local sim. only)

37: Mediterranean

Candidate Entity	ME
Mediterranean_Sea	0.446966
Battle_of_the_Mediterranean	0.174937
Mediterranean_Basin	0.016009
Mediterranean_Fleet	0.11436
Yom_Kippur_War	0.087889
Napoleonic_Wars	0.418877
University_of_the_Mediterranean	1.87532
Mediterranean_sea_u0028oceanographyu0029	0.003499
Mediterranean_diet	0.008189
Mediterranean_naval_engagements_during_World_War_1	0.01274
Southern_Europe	0.033329
Mediterranean_race	0.01892
1991_Mediterranean_Games	0.00145
Mediterranean_Regionu002c_Turkey	0.581384
Israeli_coastal_plan	0.286977
Ecology_of_California	0.00611
Mediterranean_pass	0.00395
Mediterranean_Squadron_u0028United_Statesu0029	0.002111

Example

- Find web pages concerning “The King of Rock’n’Roll” (*entity search*)
- How much fuzz about “Santorum” in each month of 2012? (*entity tracking*)

Text segmentation

Segment a piece of text into fields. E.g., “52-A Goregaon West Mumbai 400 062”.

Id	House no
1	52
1	52-A
1	52-A
1	52

The screenshot shows the CiteSeerX beta interface. The document title is "The YAGO-NAGA approach to knowledge discovery [3 citations — 0 self]" by Gjergji Kasneci, Maya Ramanath, Fabian Suchanek, and Gerhard Weikum. The document is categorized as SIGMOD Rec and is available for download at <http://suchanek.name/work/publications/sigmodrec20>. The abstract states: "This paper gives an overview on the YAGO-NAGA approach to information extraction for building a conveniently searchable, large-scale, highly accurate knowledge base of common facts. YAGO harvests infoboxes and category names of Wikipedia for facts about individual entities, and it reconciles these with the taxonomic backbone of WordNet in order to ensure that all entities have proper classes and the class system is consistent. Currently, the YAGO knowledge base contains about 19 million instances of binary relations for about 1.95 million entities. Based on intensive sampling, its accuracy is estimated to be above 95 percent. The paper presents the architecture of the YAGO extractor toolkit, its distinctive approach to consistency checking, its provisions for maintenance and further growth, and the query engine for YAGO, coined NAGA. It also discusses ongoing work on extensions towards integrating fact candidates extracted from natural-language text sources. 1." The citations list includes: 155 Two-stage language models for information retrieval - Zhai, Lafferty - 2002; 149 Unsupervised named-entity extraction from the web: an experimental study - Etzioni, Cafarella, et al. - 2005; 106 DBpedia: A Nucleus for a Web of Open Data - Auer, Bizer, et al. - 2008. The popular tags section is empty, and the BibTeX entry is provided.

CiteSeer^x beta

Documents Authors Tables

Search

☐ Include Citations | Advanced Search | Help

Summary Related Documents Version History

The YAGO-NAGA approach to knowledge discovery [3 citations — 0 self]

by Gjergji Kasneci, Maya Ramanath, Fabian Suchanek, Gerhard Weikum
SIGMOD Rec
[Add To MetaCart](#)

DOWNLOAD:
<http://suchanek.name/work/publications/sigmodrec20>
CACHED:

[Add to Collection](#) [Correct Errors](#) [Monitor Changes](#)

Abstract:

This paper gives an overview on the YAGO-NAGA approach to information extraction for building a conveniently searchable, large-scale, highly accurate knowledge base of common facts. YAGO harvests infoboxes and category names of Wikipedia for facts about individual entities, and it reconciles these with the taxonomic backbone of WordNet in order to ensure that all entities have proper classes and the class system is consistent. Currently, the YAGO knowledge base contains about 19 million instances of binary relations for about 1.95 million entities. Based on intensive sampling, its accuracy is estimated to be above 95 percent. The paper presents the architecture of the YAGO extractor toolkit, its distinctive approach to consistency checking, its provisions for maintenance and further growth, and the query engine for YAGO, coined NAGA. It also discusses ongoing work on extensions towards integrating fact candidates extracted from natural-language text sources. 1.

Citations

155 Two-stage language models for information retrieval - Zhai, Lafferty - 2002
149 Unsupervised named-entity extraction from the web: an experimental study - Etzioni, Cafarella, et al. - 2005
106 DBpedia: A Nucleus for a Web of Open Data - Auer, Bizer, et al. - 2008

POPULAR TAGS

Add a tag:
[Submit](#)

No tags have been applied to this document.

BIBTEX | ADD TO METACART

```
@ARTICLE{Kasneci_theyago-naga,  
author = {Gjergji Kasneci and Maya Ramanath  
and Fabian Suchanek and Gerhard Weikum},  
title = {The YAGO-NAGA approach to knowledge  
discovery},  
journal = {SIGMOD Rec},  
year = {},  
pages = {03}}  
)
```

Example

- Send a promotion to customers in West Mumbai.
- Find all papers containing YAGO in the title (*faceted search*)

Relation extraction (NELL / Yago2)

Extract structured relations from the web.

Recently-Learned Facts twitter				Refresh	
Instance	iteration	date learned	confidence		
dried_squash_seeds is a nut	225	28-mar-2011	99.5		
sinnett_thorn_mountain_cave is a cave	225	28-mar-2011	99.7		
vail_road is a street	224	26-mar-2011	98.4		
harold_macmillan is a scientist	225	28-mar-2011	96.6		
n32207 is a ZIP code	224	26-mar-2011	99.4		
wday_tv collaborates with bbc_news	224	26-mar-2011	96.9		
times_controls friedman	227	03-apr-2011	96.9		
support_personnel is a profession that is a kind of professionals	224	26-mar-2011	96.9		
nbc_news is a newspaper in the city washington_dc	224	26-mar-2011	99.2		
twitter operates the website twitter.com	225	28-mar-2011	100.0		

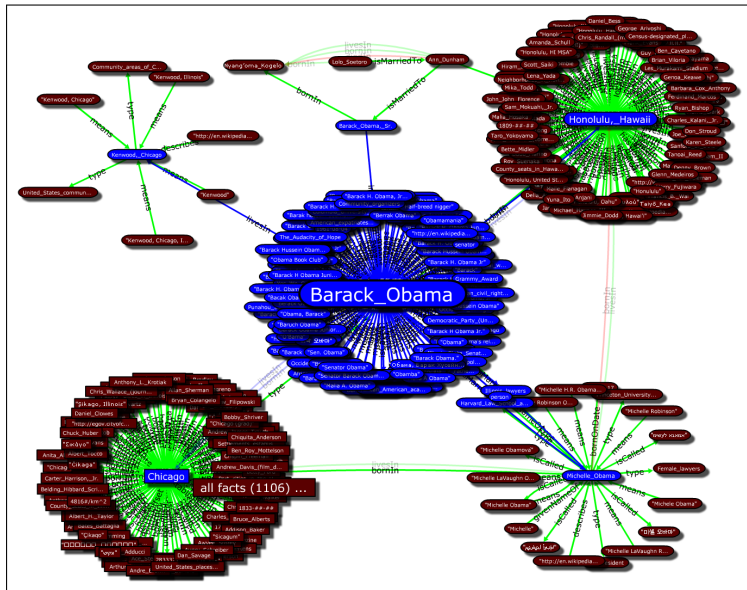
- [street](#) (98.4%)

- CPL @219 (98.4%) on 13-mar-2011 ["ramp onto _" "second right onto _" "first traffic light onto _" "bear left onto _" "off ramp onto _" "traffic light onto _"] using [vail_road](#)
- CPL @66 (87.5%) on 17-mar-2010 ["first traffic light onto _" "off ramp onto _" "bear left onto _"] using [vail_road](#)

Example

- What is known about Albert Einstein? (*fact search*)
- Who has won a Nobel Prize and is born in Ulm? (*question answering*)

Reasoning with uncertainty (URDF)



Google Squared (discontinued)

Find and describe items of a given category.

Google Squared Labs

comedy movies [Square it] [Add to this Square]

Item Name	Release Date	Genre	Director	Country	Language
☒ The Mask	29 July 1994	Comedy	Chuck Russell	USA	English
☒ Shrek 2	19 May 2004	Adventure			
☒ Scary Movie	7 July 2000	Comedy			
☒ Role Models	7 November 2008	Comedy			
☒ Road Trip	19 May 2000	Comedy			
☒ Old School	21 February 2003	Comedy			

Chuck Russell
Directed by for The Mask
www.freebase.com - [all 10 sources »](#)

Other possible values

- ☐ **Dean Koontz** Low confidence
Author for The Mask
www.freebase.com - [all 4 sources »](#)
- ☐ **Bob Engelman** Low confidence
Producer for The Mask
www.infibeam.com - [all 3 sources »](#)
- ☐ **Timothy Bond** Low confidence
Mask Comedy Movies and read product reviews. ... Director: Timothy Bond; Release Date: September 07, 2004. Add to list. Price At www.shopping.com - [all 2 sources »](#)

[Search for more values »](#)

Example

- Directors that directed at least one comedy movie?
- Birthplaces of directors of comedy movies with a budget of over \$20M?

Information integration

Same? {

Op.System	CustId	Name	City	State
1	C_1	John	San Francisco	CA
2	C_2	Johnny	San Jose	CA
1	C_3	Jack	San Francisco	CA
1	C_4	William	San Francisco	CA
2	C_5	Bill	San Jose	CA

} Which one?

(a) Customer Data

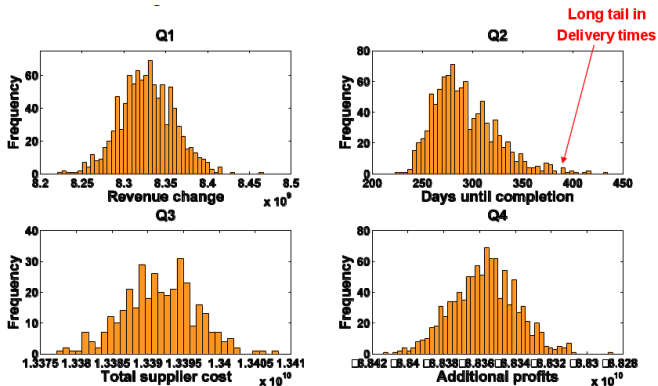
Op.System	TransID	CustID	Sales
1	Tr_1	C_1	\$15
1	Tr_2	C_1	\$5
2	Tr_3	C_2	\$30
2	Tr_4	C_2	\$20
1	Tr_5	C_3	\$30
1	Tr_6	C_4	\$90
2	Tr_7	C_5	\$25
2	Tr_8	C_5	\$15

(b) Transaction Data

Example

- Turnover in San Francisco? And in California? (*OLAP*)

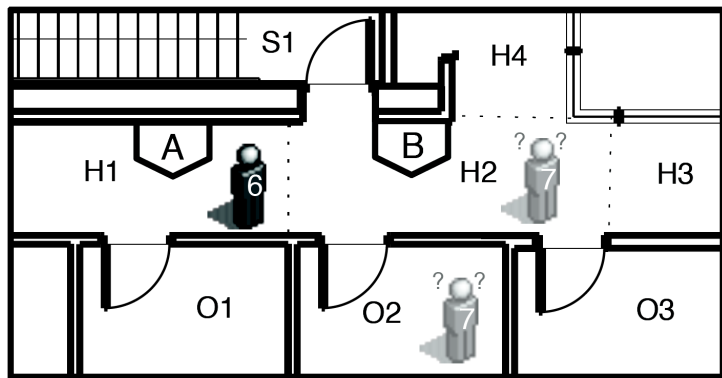
Predictive analytics



Example

- What is the effect of changing the price on future sales?
- What is the risk associated with my portfolio?

RFID & moving objects



Example

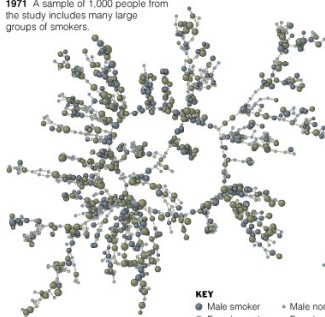
- How many people are attending John's lecture?
- Where are choke points when moving items through my storage facility?

Statistical & uncertain rules

Smoking and Quitting in Groups

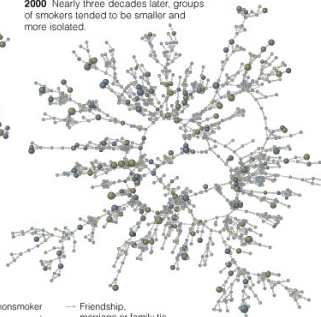
Researchers studying a network of 12,067 people found that smokers and nonsmokers tended to cluster in groups of close friends and family members. As more people quit over the decades, remaining groups of smokers were increasingly pushed to the periphery of the social network.

1971 A sample of 1,000 people from the study includes many large groups of smokers.



Sources: *New England Journal of Medicine*,
Dr. Nicholas A. Christakis; James H. Fowler

2000 Nearly three decades later, groups of smokers tended to be smaller and more isolated.



THE NEW YORK TIMES

KEY

● Male smoker • Male nonsmoker — Friendship, marriage or family tie
● Female smoker • Female nonsmoker

Circle size is proportional to the number of cigarettes smoked per day.

Example

- Does John smoke? (*social network analysis*)
- “Mississippi” most often refers to the state of Mississippi. (*entity disambiguation*)

Anonymized data

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Example

- Medical research, trend analysis, allocation of public funds, ...

Outline

1 Uncertainty in the Real World

2 Managing Uncertainty

How to deal with uncertainty? (1)

Clean it (then deny it)!

- E.g., data warehouse systems
- Advantages
 - ▶ Lots of expertise and tools for cleaning data
 - ▶ Can be stored and queried in traditional DBMS
- Disadvantages
 - ▶ Loss of information
 - ▶ No risk assessment
 - ▶ High expense of cleaning
 - ▶ New data may “break” the clean database
- Important, but not covered in this lecture!

Customers

Sys	Cust	Name	City	State
1	C ₁	John	SFO	CA
2	C ₂	Johnny	SJ	CA
1	C ₃	Jak	SFO	CA

Same! {

CleanedCustomers

Cust	Name	City	State
C ₁₂	Johnny	SFO	CA
C ₃	Jak	SFO	CA

How to deal with uncertainty? (2)

Manage it!

Approach I: Incomplete databases

- A data integration scenario

Customers

Same! {

Sys	Cust	Name	City	State
1	C ₁	John	SFO	CA
2	C ₂	Johnny	SJ	CA
1	C ₃	Jak	SFO	CA

Transactions

Sys	TransID	Cust	Sales
1	T ₁	C ₁	\$15
1	T ₂	C ₁	\$5
2	T ₃	C ₂	\$30
1	T ₄	C ₃	\$30

- Resolving entities via an incomplete database

ResolvedCustomers

Ent	Name	City	State
E ₁	John Johnny	SFO SJ	CA
E ₂	Jak	SFO	CA

ResolvedTransactions

TransID	Ent	Sales
T ₁	E ₁	\$15
T ₂	E ₁	\$5
T ₃	E ₁	\$30
T ₄	E ₂	\$30

- Some query results

Sales by city

City	Sum(Sales)	Status
SFO	\$30-\$80	guaranteed
SJ	\$50	non-guaranteed

Sales by state

State	Sum(Sales)	Status
CA	\$80	guaranteed

Approach II: Probabilistic databases

- Bird watcher's observations

Sightings

	Name	Bird	Species
t_1 :	Mary	Bird-1	Finch: 0.8 Toucan: 0.2
t_2 :	Susan	Bird-2	Nightingale: 0.65 Toucan: 0.35
t_3 :	Paul	Bird-3	Humming bird: 0.55 Toucan: 0.45

- Which species exist in the park?

ObservedSpecies

Species	
Finch: 0.8	? $(t_1, 1)$
Toucan: 0.714	? $(t_1, 2) \vee (t_2, 2) \vee (t_3, 2)$
Nightingale: 0.65	? $(t_2, 1)$
Humming bird: 0.55	? $(t_3, 1)$

DistinctSpecies

#	
1: 0.0315	? ...
2: 0.2230	? ...
3: 0.7455	? ...

- Observe: Cleaning up data by most likely choice would miss Toucan!

Approach III: Probabilistic graphical models

- Anna and Bob are friends. Anna smokes, but does not have cancer. What do we know about Bob?

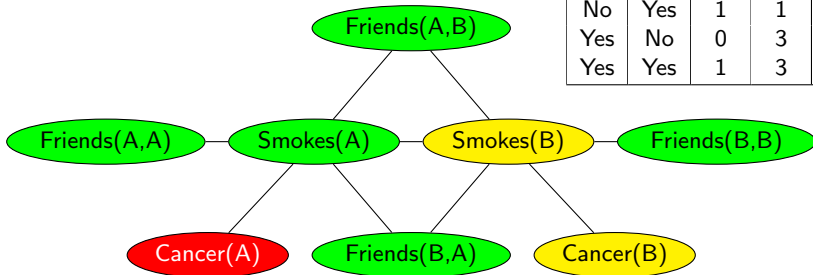
- Uncertain knowledge

1.5 { Smoking causes cancer
 $\forall x. \text{Smokes}(x) \implies \text{Cancer}(x)$

1.1 { Friends have similar smoking habits
 $\forall x. \forall y. \text{Friends}(x, y) \implies (\text{Smokes}(x) \iff \text{Smokes}(y))$

- Build a graphical model
& perform inference

S(B)	C(B)	#R1	#R2	w	Prob.
No	No	1	1	2.6	7.7%
No	Yes	1	1	2.6	7.7%
Yes	No	0	3	3.3	15.4%
Yes	Yes	1	3	4.8	69.2%



How to deal with uncertainty? (2)

Manage it!

- Advantages

- ▶ No or little loss of information
- ▶ Uncertainty might be resolved more accurately at query time
- ▶ Risk assessment is possible
- ▶ Less upfront effort
- ▶ Arrival of new data handled gracefully

- Disadvantages

- ▶ Increased cost of data processing
- ▶ Active research area with lots of open issues (and interesting results)
- ▶ No commercial DBMS systems available!

- This lecture!

Course overview

- Modelling uncertainty
 - ▶ Incomplete databases
 - ▶ Probabilistic databases
 - ▶ Probabilistic graphical models for relational data
- Managing uncertain data
 - ▶ Languages (relational algebra, datalog, relational calculus)
 - ▶ Provenance
 - ▶ Algorithms
 - ▶ Complexity
 - ▶ Approximation techniques
 - ▶ Systems
- Applications
 - ▶ Information extraction, sensor networks, business intelligence & predictive analytics, forecasting, scientific data management, privacy preserving data mining, data integration, data deduplication, social network analysis, ...

Suggested reading

- Charu C. Aggarwal (Ed.)
Managing and Mining Uncertain Data (Chapter 1)
Springer, 2009.
- Daphne Koller, Nir Friedman
Probabilistic Graphical Models: Principles and Techniques (Chapter 1)
The MIT Press, 2009
- Dan Suciu, Dan Olteanu, Christopher Ré, Christoph Koch
Probabilistic Databases (Chapter 1)
Morgan & Claypool, 2011
- Charu C. Aggarwal, Philip S. Yu
A Survey of Uncertain Data Algorithms and Applications
IEEE Transactions of Knowledge and Data Engineering, 21(5),
pp. 609–623, May 2009