DeepDive: Declarative Knowledge Base Construction

EPL 646: Advanced Topics in Databases

Christodoulou Eirini, Loizou Marios

Department of Computer Science
University of Cyprus

10 October 2017

Overview

Overview

Introduction

Knowledge Base Contribution (KBC) problem:

Need to populate a SQL database (DB) with information from unstructured data sources

- Goal
  - To obtain high-quality structured data from unstructured information
- Quality measures
  1. Precision: How often a claimed tuple is correct
  2. Recall: How many of the possible tuples to extract are actually extracted

Introduction

Knowledge Base Contribution (KBC) problem:

Need to populate a SQL database (DB) with information from unstructured data sources

- Goal
  - To obtain high-quality structured data from unstructured information
- Quality measures
  1. Precision: How often a claimed tuple is correct
  2. Recall: How many of the possible tuples to extract are actually extracted

Introduction

Knowledge Base Contribution (KBC) problem:

Need to populate a SQL database (DB) with information from unstructured data sources

- Goal
  - To obtain high-quality structured data from unstructured information
- Quality measures
  1. Precision: How often a claimed tuple is correct
  2. Recall: How many of the possible tuples to extract are actually extracted

Introduction

DeepDive

- System that combines DB and machine learning ideas to help develop KBC systems
- End-to-end framework for building KBC systems
- Uses a declarative language for rules that are manually created by the user (SQL, datalog)

Overview

Knowledge Base (klog)

- A knowledge base is a database used for knowledge sharing and management
- It promotes the collection, organization and retrieval of knowledge
- KBs are structured around artificial intelligence and not only store data

https://www.techopedia.com/definition/2511/knowledge-base-klog

Knowledge Base (klog)

- A knowledge base is a database used for knowledge sharing and management
- It promotes the collection, organization and retrieval of knowledge
- KBs are structured around artificial intelligence and not only store data

https://www.techopedia.com/definition/2511/knowledge-base-klog

Knowledge Base (klog)

- A knowledge base is a database used for knowledge sharing and management
- It promotes the collection, organization and retrieval of knowledge
- KBs are structured around artificial intelligence and not only store data

https://www.techopedia.com/definition/2511/knowledge-base-klog

Knowledge Base Construction (KBC)

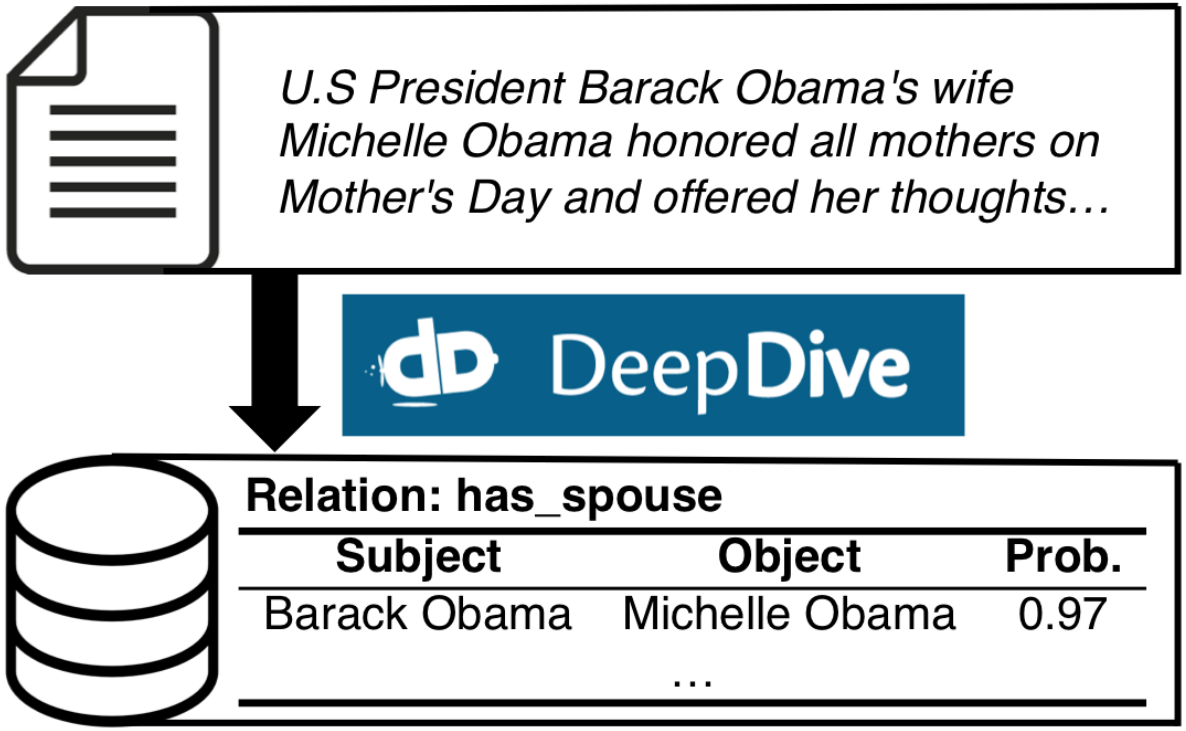


Figure: Knowledge Base Construction is the process of populating a structured relational knowledge base from unstructured sources

Knowledge Base Construction (KBC)

- Input $\longleftarrow$ a heterogeneous collection of data
  - unstructured
  - semi-structured
  - structured

  raging from texts documents to existing but incomplete KBs

- Output $\longrightarrow$ relational database
  - containing facts extracted from the input

Knowledge Base Construction (KBC)

- Input $\longleftarrow$ a heterogeneous collection of data
  - unstructured
  - semi-structured
  - structured

  raging from texts documents to existing but incomplete KBs

- Output $\longrightarrow$ relational database
  - containing facts extracted from the input

Knowledge Base Construction (KBC)

(a) Natural Language Text

... The Namurian Tsingyuan Formation from Ningxia, China, is divided into three members ...

Formation-Time (Location)

Formation	Time
Tsingyuan Fm.	Namurian

Formation	Location
Tsingyuan Fm.	Ningxia

Figure: A knowledge base with pairs of rock formations and locations that the rock formation is found

KBC Model

- A KBC system needs to interact with different types of data sources
- A KBC model is the abstraction of these data sources
- There are four types of objects that a KBC system seeks to extract from input documents
  1. *entities*
  2. *relations*
  3. *mentions*
  4. *relation mentions*

KBC Model

- A KBC system needs to interact with different types of data sources
- A KBC model is the abstraction of these data sources
- There are four types of objects that a KBC system seeks to extract from input documents
  1. *entities*
  2. *relations*
  3. *mentions*
  4. *relation mentions*

KBC Model

- A KBC system needs to interact with different types of data sources
- A KBC model is the abstraction of these data sources
- There are four types of objects that a KBC system seeks to extract from input documents
  1. *entities*
  2. *relations*
  3. *mentions*
  4. *relation mentions*

KBC Model

- A KBC system needs to interact with different types of data sources
- A KBC model is the abstraction of these data sources
- There are four types of objects that a KBC system seeks to extract from input documents
  1. *entities*
  2. *relations*
  3. *mentions*
  4. *relation mentions*

KBC Model

- A KBC system needs to interact with different types of data sources
- A KBC model is the abstraction of these data sources
- There are four types of objects that a KBC system seeks to extract from input documents
  1. *entities*
  2. *relations*
  3. *mentions*
  4. *relation mentions*

KBC Model

- A KBC system needs to interact with different types of data sources
- A KBC model is the abstraction of these data sources
- There are four types of objects that a KBC system seeks to extract from input documents
  1. *entities*
  2. *relations*
  3. *mentions*
  4. *relation mentions*

KBC Model

- A KBC system needs to interact with different types of data sources
- A KBC model is the abstraction of these data sources
- There are four types of objects that a KBC system seeks to extract from input documents
  1. *entities*
  2. *relations*
  3. *mentions*
  4. *relation mentions*

KBC Model

Entity

An entity is a real-world person, place, or thing. For example, the entity "Bill Gates 1" represents the actual entity for a person whose name is "Bill Gates".

Relation

A relation associates two (or more) entities, and represents the fact that there exists a relationship between these entities. For example, the entity "Microsoft Corp. 1" and "Bill Gates 1" participate in the FoundedBy relation, which indicates that Bill Gates is the founder of the Microsoft Corp.

KBC Model

Mention

A mention is a span of text in an input document that refers to an entity or relationship: "Gates" may be a mention of the entity "Bill Gates 1." An entity might have mentions of different from; for example, apart from "Gates", mentions of the form "B. Gates" or "Bill" can also refer to the same entity "Bill Gates 1." The process of mapping mentions to entities is called entity linking.

Relation Mention

A relation mention is a phrase that connects two mentions that participate in a relation, e.g., the phrase "is founded by" that connects the mentions "Microsoft Corp." and "Bill Gates".

KBC Model

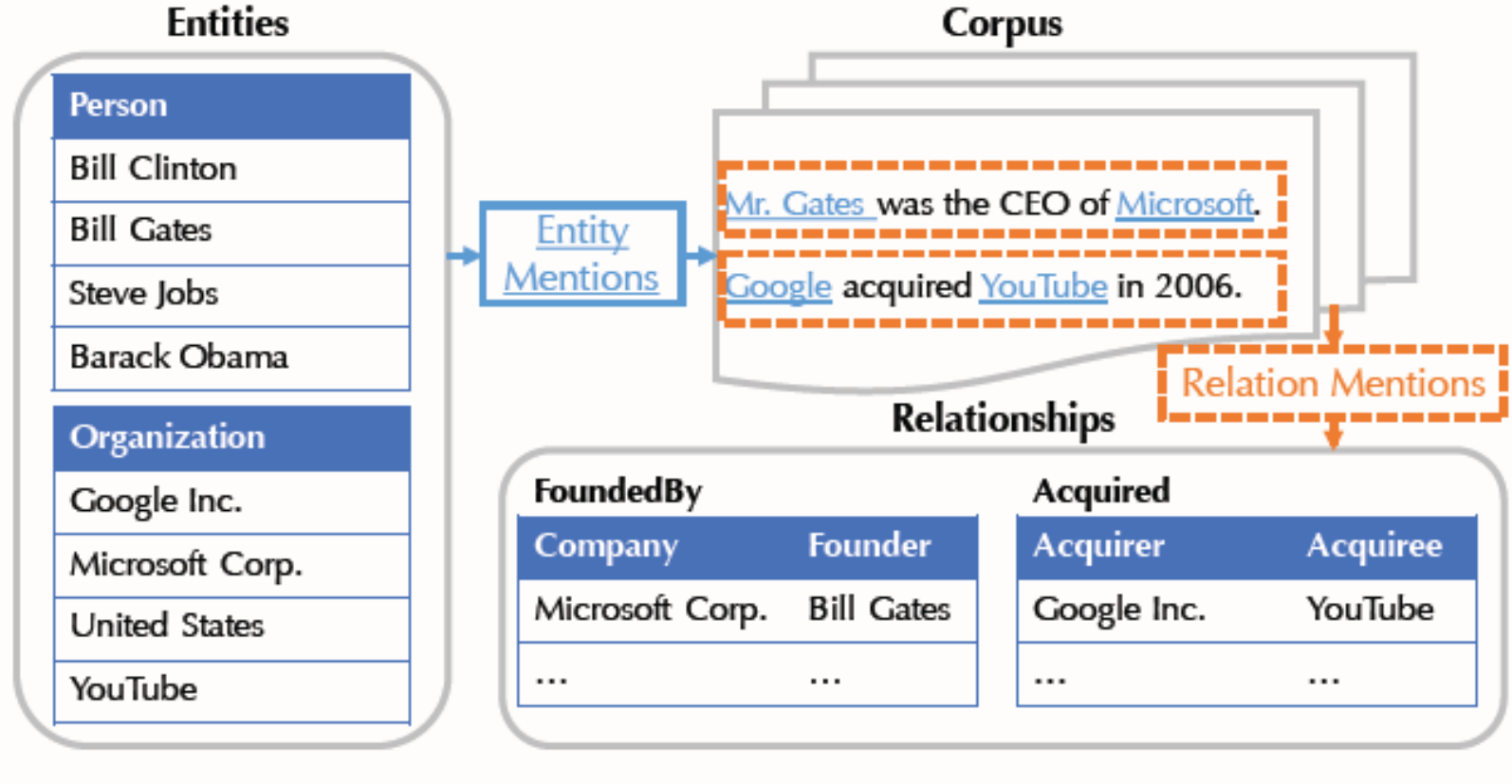
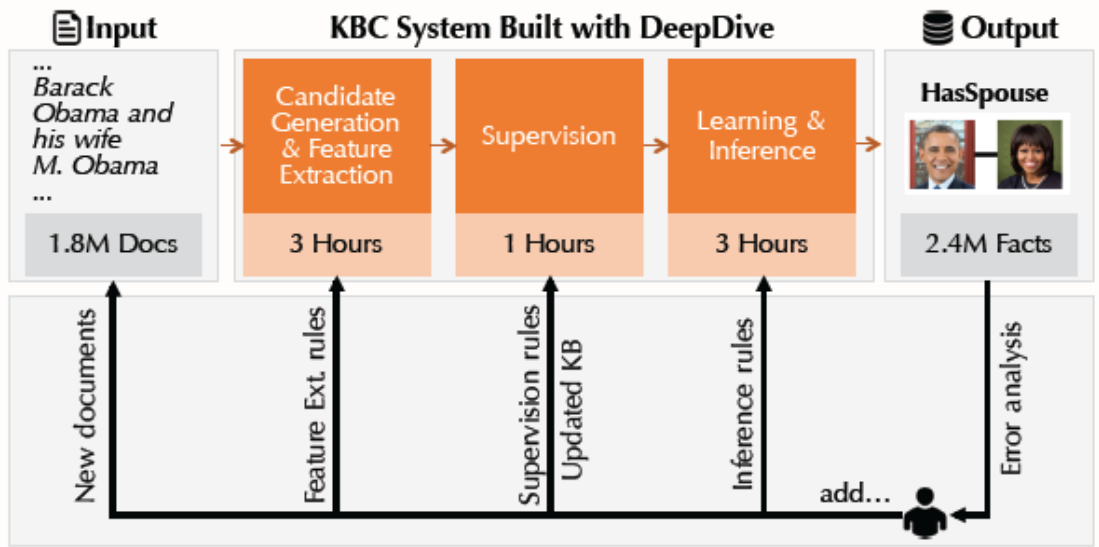


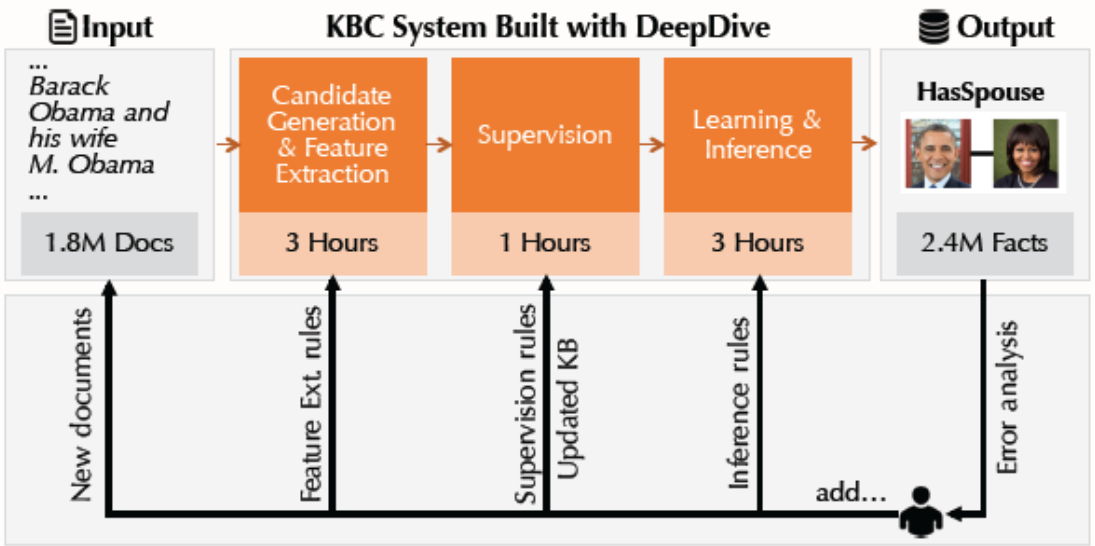
Figure: An illustration of the KBC model

The DeepDive Framework



- Description of DeepDive program, that the user writes to build a KBC system
- The user can specify features, domain knowledge and distant supervision, all in a unified framework

The DeepDive Framework



- Description of DeepDive program, that the user writes to build a KBC system
- The user can specify features, domain knowledge and distant supervision, all in a unified framework

Candidate Generation and Feature Extraction

- Loading phase to populate the database
- DeepDive stores all documents in the database in one sentence per row
- Two types of queries:
  1. **Candidate mappings:** SQL queries that produce possible mentions, entities and relations
  2. **Feature extractors:** Associate features to candidates,

Candidate Generation and Feature Extraction

- Loading phase to populate the database
- DeepDive stores all documents in the database in one sentence per row
- Two types of queries:
  1. **Candidate mappings:** SQL queries that produce possible mentions, entities and relations
  2. **Feature extractors:** Associate features to candidates,

Candidate Generation and Feature Extraction

- Loading phase to populate the database
- DeepDive stores all documents in the database in one sentence per row
- Two types of queries:
  1. **Candidate mappings:** SQL queries that produce possible mentions, entities and relations
  2. **Feature extractors:** Associate features to candidates,

Candidate Generation and Feature Extraction

- Loading phase to populate the database
- DeepDive stores all documents in the database in one sentence per row
- Two types of queries:
  1. **Candidate mappings:** SQL queries that produce possible mentions, entities and relations
  2. **Feature extractors:** Associate features to candidates,

Candidate Generation and Feature Extraction

- Loading phase to populate the database
- DeepDive stores all documents in the database in one sentence per row
- Two types of queries:
  1. **Candidate mappings:** SQL queries that produce possible mentions, entities and relations
  2. **Feature extractors:** Associate features to candidates,

Candidate Generation and Feature Extraction

- Candidate mappings are simply SQL queries with UDF (user defined functions), with high recall

Relation mention for every pair of candidate persons (mi) in the same sentence (s)

```
(R1)  MarriedCandidate(m1, m2):-
      PersonCandidate(s, m1), PersonCandidate(s, m2)
```

Candidate Generation and Feature Extraction

- Extract features with the use of (1) user defined functions and (2) weight tying

The phrase between two mentions may indicate whether two people are married

```
(FE1) MarriedMentions(m1, m2):-
      MarriedCandidate(m1, m2), Mention(sid, m1)
      Mention(sid, m2), Sentence(sid, sent)
      weight = phrase(m1, m2, sent)
```

- The relationship between m1 and m2 mentions is influenced by the phrase between those mention pairs

Supervision

- Each user relation is associated with an evidence relation with the same schema and an additional field that indicates whether the entry is true or false

Distant supervision

```
(S1) MarriedMentions_Ev(m1, m2, true):-
     MarriedCandidate(m1, m2), EL(m1, e1)
     EL(m2, e2), Married(e1, e2)
```

- The idea is to use an incomplete KB of married entity pairs to heuristically label (as True evidence) all relation mentions that link to a pair of married entities

KBC procedure

(1a) Unstructured Information

B. Obama and Michelle were married in 1992.

Malia and Sasha Obama attended the state dinner.

Tunicia and Raleigh Hall were married in June.

(1b) Structured Information

HasSpouse (Freebase)

Person1	Person2
Barack Obama	Michelle Obama
...	...

(2) User Schema

Sentence (documents)

SID	Content
S1	B. Obama and Michelle were married in 1992.

PersonCandidate

SID	MID
S1	M2

Mentions

SID	MID	WORD
S1	M2	Michelle

Married

EID1	EID2
Barack Obama	Michelle Obama

EL

MID	EID
S2	Michelle Obama

MarriedMentions_Ev

MID1	MID2	VALUE
M1	M2	true

MarriedCandidate

MID1	MID2
M1	M2

(3a) Candidate Generation and Feature Extraction

```
(R1) MarriedCandidate(m1,m2) :- PersonCandidate(s,m1), PersonCandidate(s,m2)

(FE1) MarriedMentions(m1,m2) :- MarriedCandidate(m1,m2), Sentence(sid,sent),
                                Mentions(s,m1,w1), Mentions(s,m2,w2),
                                weight=phrase(m1,m2,sent)
```

(3b) Supervision Rule

```
(S1) MarriedMentions_Ev(m1,m2,true) :- MarriedCandidate(m1,m2), Married(e1,e2),
                                       EL(m1,e1), EL(m2,e2),
```

Figure: An example DeepDive program for a KBC system

Factor Graph

- Probabilistic graphical model
- Nodes:
  1. Variables: Evidence (known), Query (value should be predicted)
  2. Factors: Relationships between variables
- Factor function: defines the relationship between factors
- Weights: Describes how much influence the factor has on variables

Factor Graph

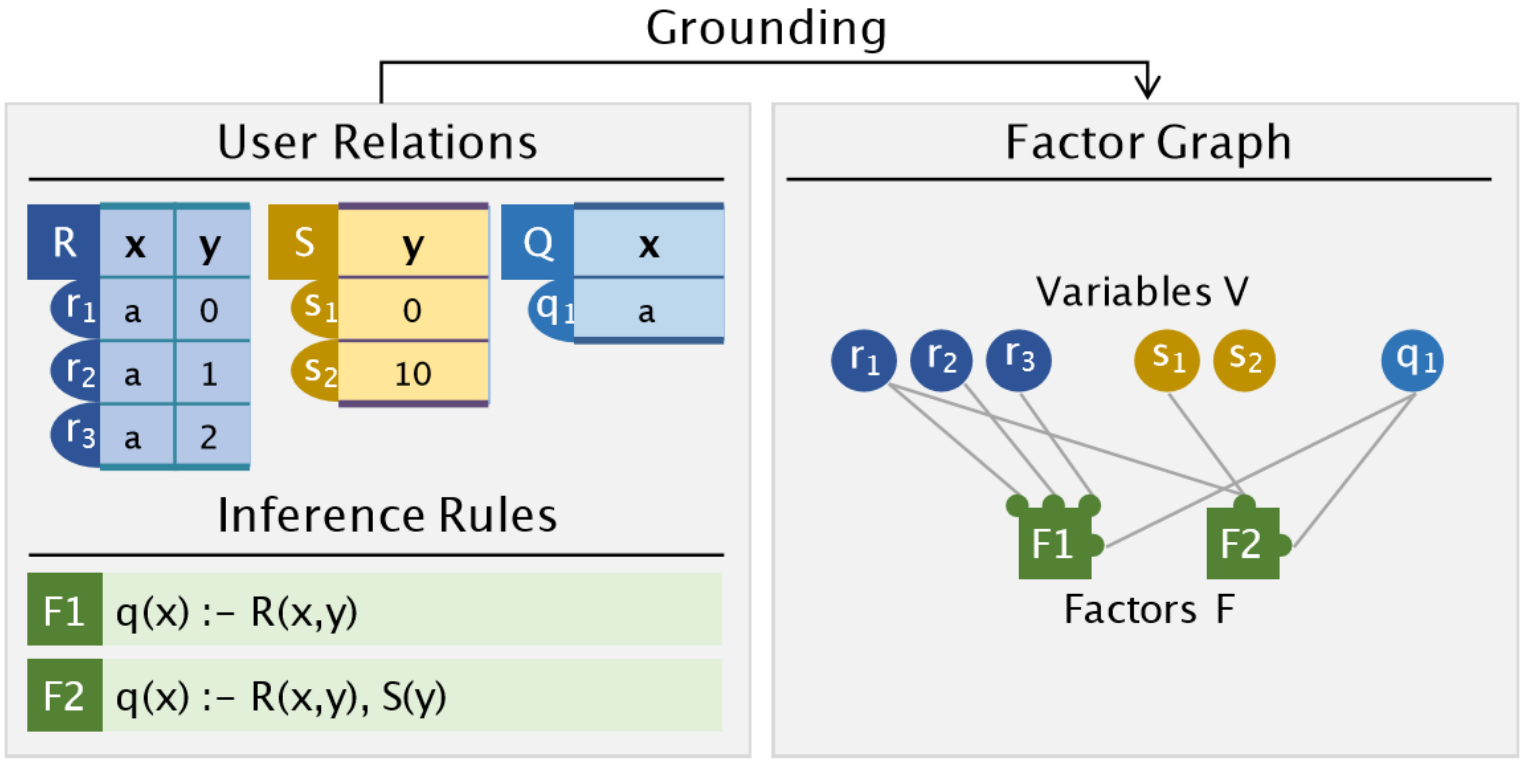


Figure: Each tuple corresponds to a Boolean random variable and node in the factor graph. One factor is created for every inference rule.

Learning and Inference & Error Analysis

- Possible world $\longleftarrow$ assignment to every variable in a factor graph
- **Learning process:** Chooses the weights by maximizing the probabilities of possible worlds
  1. Starts from a random possible world and iterates over each variable v
  2. **Error Analysis:** Update value be taking into account:
     - Factor functions connected to v
     - Values of connected variables
  3. Estimated probability of the variable taking a value $=$ number of iterations which the variable has the specific value / total iterations

Discussion of Design Choices

- No reference to the underlying machine learning algorithms
  $\Rightarrow$ DeepDive programs are declarative in a strong sense
- Allows users to write feature extraction code (UDFs) in familiar languages (Python, SQL)
- Fits into the familiar SQL stack, which allows standard tools to inspect and visualize data
- The user constructs an end-to-end system and then refines the quality of the system in a pay-as-you-go way

Discussion of Design Choices

- No reference to the underlying machine learning algorithms $\Rightarrow$ DeepDive programs are declarative in a strong sense
- Allows users to write feature extraction code (UDFs) in familiar languages (Python, SQL)
- Fits into the familiar SQL stack, which allows standard tools to inspect and visualize data
- The user constructs an end-to-end system and then refines the quality of the system in a pay-as-you-go way

Discussion of Design Choices

- No reference to the underlying machine learning algorithms $\Rightarrow$ DeepDive programs are declarative in a strong sense
- Allows users to write feature extraction code (UDFs) in familiar languages (Python, SQL)
- Fits into the familiar SQL stack, which allows standard tools to inspect and visualize data
- The user constructs an end-to-end system and then refines the quality of the system in a pay-as-you-go way

Discussion of Design Choices

- No reference to the underlying machine learning algorithms $\Rightarrow$ DeepDive programs are declarative in a strong sense
- Allows users to write feature extraction code (UDFs) in familiar languages (Python, SQL)
- Fits into the familiar SQL stack, which allows standard tools to inspect and visualize data
- The user constructs an end-to-end system and then refines the quality of the system in a pay-as-you-go way

The DeepDive Framework

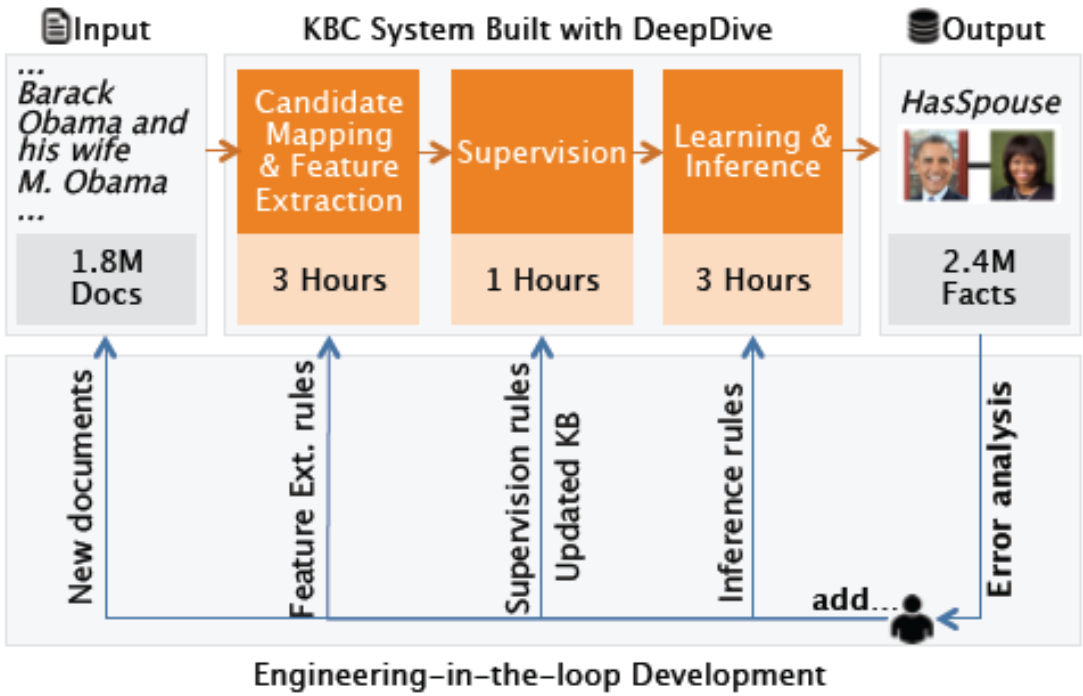


Figure: A KBC system takes as input unstructured documents and outputs a structured knowledge base

KBC - Challenges

1. Joint statistical inference
   - Diverse sources of input (e.g. text, tables, figures, images, articles)
   - Noisy information
   - Need of models that is able to ingest whatever sources present themselves
     1. Opportunistically
     2. Simultaneously
2. Scalability and Efficiency
   - Performance is a major challenge
   - Input 300K journal articles, size exceeds 2TB (PaleoDeep-Dive)
   - Hundreds of thousands of machine hours for process the raw data
   - The output is used by DeepDive to construct factor graphs
   - More than 300 million candidates variables

Joint Statistical Inference

DeepDive approach

- Bayesian probabilistic approach
- Treats all these information sources as one joint probabilistic inference problem, with all predictions modelled as random variables within a factor graph model
- This probabilistic framework ensures all facts that are produced by DeepDive are associated with a marginal probability

KBC - Challenges

1. Joint statistical inference
   - Diverse sources of input (e.g. text, tables, figures, images, articles)
   - Noisy information
   - Need of models that is able to ingest whatever sources present themselves
     1. Opportunistically
     2. Simultaneously
2. Scalability and Efficiency
   - Performance is a major challenge
   - Input 300K journal articles, size exceeds 2TB (PaleoDeepDive)
   - Hundreds of thousands of machine hours for process the raw data
   - The output is used by DeepDive to construct factor graphs
   - More than 300 million candidates variables

Scalability and Efficiency

DeepDive approach

- Focus on speed up probabilistic inference

1. **Hardware Efficiency:** How efficient is each step
   - Takes into account architecture of Non-uniform memory access (NUMA)
   - Decrease the communication across different nodes
2. **Statistical Efficiency:** How many steps needed for converge
   - Parallelized gradient descent
   - Lock-free execution

Overview

DeepDive Applications

1. Paleontology
2. Human Trafficking / MEMEX
3. Medical Genetics
4. Pharmacogenomics
   - DeepDive is used to extract relations between genes, diseases, and drugs in order to predict novel pharmacological relationships
5. Text Analysis Conference - Knowledge Base Population (TAC-KBP)
   - Research competition where the task is to extract common properties of people and organizations

Paleontology

- Paleontology studies fossils and ancient organisms
- Main purpose is to construct a knowledge base about fossils from scientific publications
  1. PaleoDB
     - Compiled by **human** volunteers
     - **Expensive** and time consuming
  2. PaleoDeepDive
     - Input $\longleftarrow$ PDF documents
     - Challenges: Optical character recognition, Natural language processing, Information extraction
- PaleoDeepDive has achieved comparable and sometimes better quality than PaleoDB

DeepDive Applications

1. Paleontology
2. Human Trafficking / MEMEX
3. Medical Genetics
4. Pharmacogenomics
   - DeepDive is used to extract relations between genes, diseases, and drugs in order to predict novel pharmacological relationships
5. Text Analysis Conference - Knowledge Base Population (TAC-KBP)
   - Research competition where the task is to extract common properties of people and organizations

Human Trafficking / MEMEX

- MEMEX is a Defence Advanced Research Projects Agency (DARPA) program
- Supports the fight against human trafficking
  - Traffickers use physical or financial force to coerce victims into providing labour
  - Such labour is often advertised online
  - These online advertisements contain crucial information about workers

DeepDive

Tool for structured data extraction that allow users to transform the raw set of advertisements into a single clean structured DB table

Input $\longleftarrow$ Portion of the publicly-index and "dark" web

Output $\longrightarrow$ Extracted evidential data

DeepDive Applications

1. Paleontology
2. Human Trafficking / MEMEX
3. Medical Genetics
4. Pharmacogenomics
   - DeepDive is used to extract relations between genes, diseases, and drugs in order to predict novel pharmacological relationships
5. Text Analysis Conference - Knowledge Base Population (TAC-KBP)
   - Research competition where the task is to extract common properties of people and organizations

Medical Genetics

- Literature has been growing at accelerating speeds
- Online Mendelian Inheritance in Man (OMIM)
  - Authoritative database of human genes and genetic disorders
  - Started in 1960s and curated by humans
  - So far contains about 6,000 hereditary diseases or phenotypes
  - Growing rate: 50 records / month for many years

DeepDive

- Extract mentions of genes, diseases, and phenotypes
- Statistically infer their relationships
- Applied to clinical genetic diagnostics and reproductive counselling

DeepDive Applications

1. Paleontology
2. Human Trafficking / MEMEX
3. Medical Genetics
4. Pharmacogenomics
   - DeepDive is used to extract relations between genes, diseases, and drugs in order to predict novel pharmacological relationships
5. Text Analysis Conference - Knowledge Base Population (TAC-KBP)
   - Research competition where the task is to extract common properties of people and organizations

DeepDive Applications

1. Paleontology
2. Human Trafficking / MEMEX
3. Medical Genetics
4. Pharmacogenomics
   - DeepDive is used to extract relations between genes, diseases, and drugs in order to predict novel pharmacological relationships
5. Text Analysis Conference - Knowledge Base Population (TAC-KBP)
   - Research competition where the task is to extract common properties of people and organizations

The End

Thank you!!!

Bibliography

Stanford, **DeepDive**, http://deepdive.stanford.edu/ (last visited on 07/10/2017).

Ce Zhang, **DeepDive: A Data Management System for Automatic Knowledge Base Construction.** A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Computer Sciences), University of Wisconsin - Madison, August 2015.

Christopher De Sa et al, **DeepDive: Declarative Knowledge Base Construction.** SIGMOD Record, Vol. 45, No. 1, March 2016. https://sigmodrecord.org/2016/04/19/deepdive-declarative-knowledge-base-construction/