Unsupervised Information Extraction by Text Segmentation

Eli Cortez

Advisor: Altigran Soares da Silva

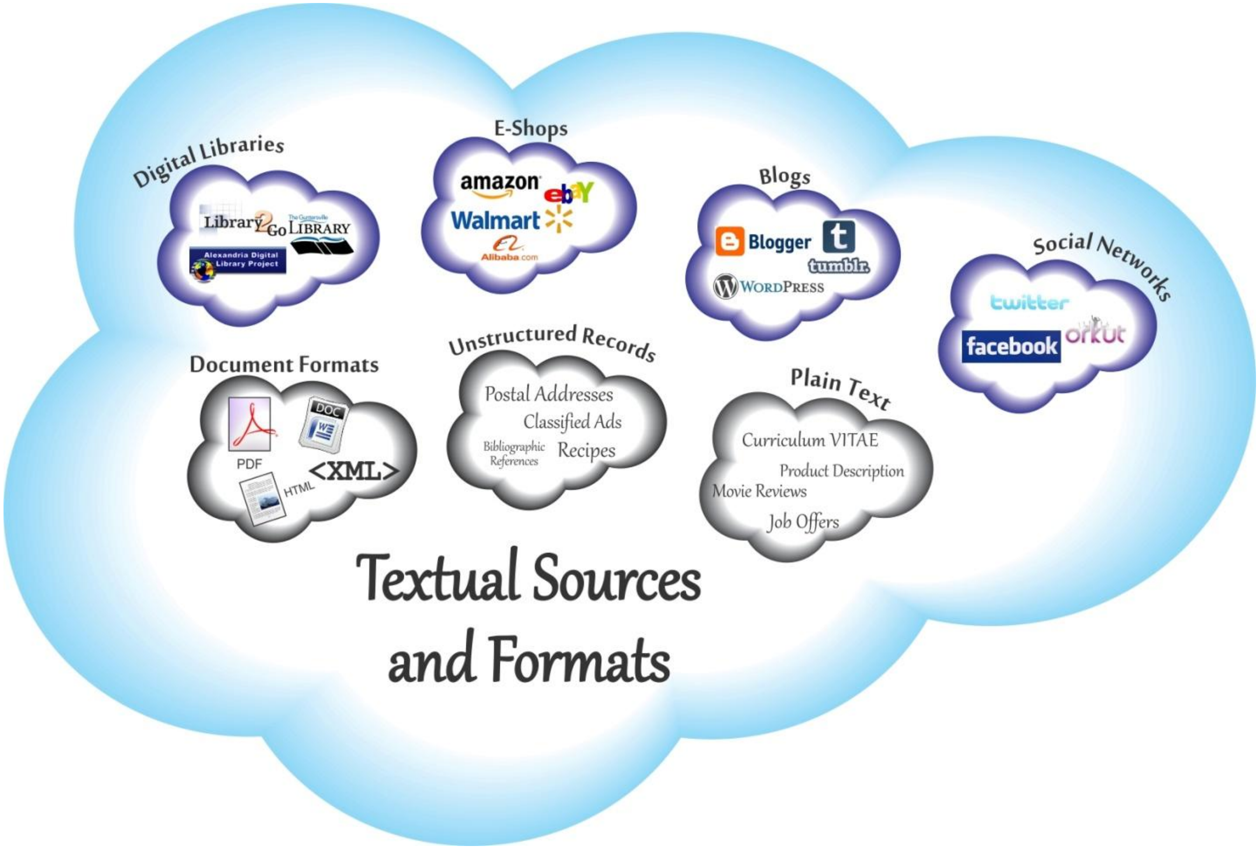
Universidade Federal do Amazonas

Agenda

- Introduction
  - Information Extraction by Text Segmentation (IETS)
  - Contributions
- Related Work
  - Web Extraction Methods and Tools
  - Probabilistic Graph-Based Methods
- Our Proposed Approach for IETS
  - Ondux
  - Judie
  - iForm
- Conclusions and Future Work

Introduction

- Steady increasing in the number and the types of sources of textual information available in the World-Wide Web

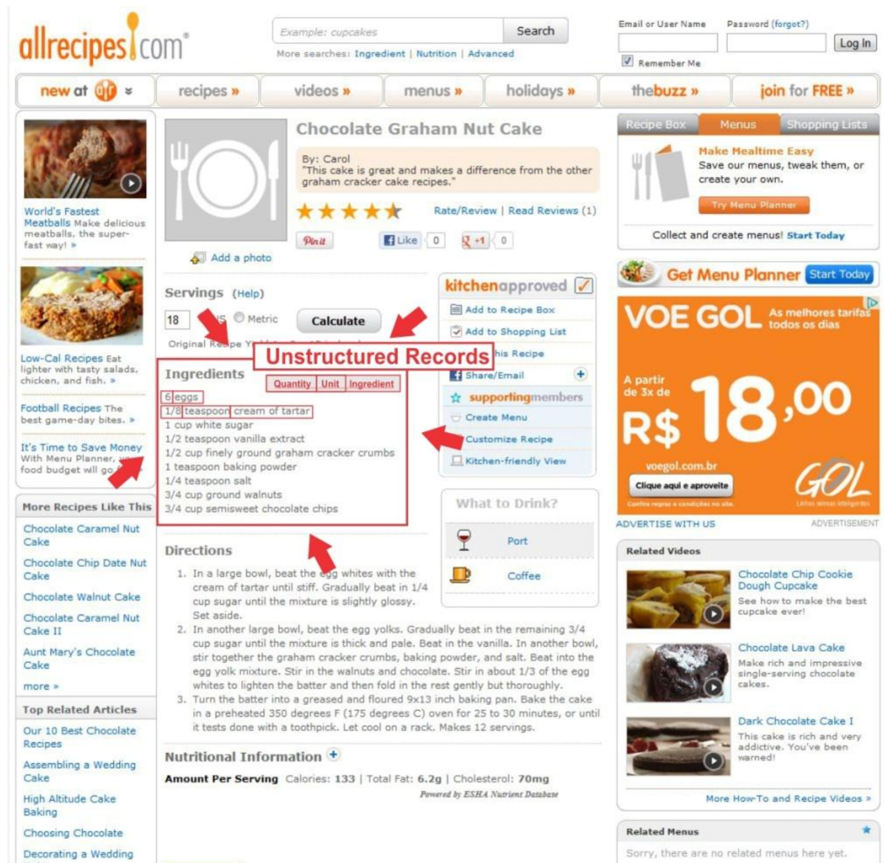


Introduction

- These sources constitute large repositories of valuable data on a variety of domains.
- Data referring to different "things" such as:
  - Personal Information;
  - Products;
  - Publication;
  - Companies;
  - Cities;

Introduction

- Important restrictions on the way data they contain can be manipulated.



- Text snippets (product descriptions, movie reviews) can hardly be subject to **automated processing**.
  - Difficult to automatically identify data of interest.

Introduction

- The Information Extraction (IE) Problem
  - Automatic extract **structured information** such as **entities**, **relationships** between entities, and **attributes** describing entities from **noisy unstructured sources**.

  - Named Entity Recognition;
  - Open Information Extraction;
  - Relationship Extraction;
  - **Information Extraction by Text Segmentation (IETS)**

Introduction

- Information Extraction by Text Segmentation (IETS)
  - The problem of **extracting attribute values** occurring in implicit **semi-structured data records** in the form of continuous text.

Eli Cortez - Rua 15 n 324 Japiim 1 - 69075 - Manaus

Name	Street	Number	Neigh.	Zip	City
Eli Cortez	Rua 15	n 324	Japiim 1	69075	Manaus

- Why is it important to extract information?
  - Query structured data; Data Mining; Record Linkage.

Introduction

- Contributions
  - In this work we tackle the Information Extraction by Text Segmentation Problem (IETS)
    - Important and practical problem frequently addressed in the recent literature.
      - Borkar@SIGMOD'01, McCallum@ICML'01, Agichtein@SIGKDD'04, Mansuri@ICDE'06, Zhao@SICDM'08, Cortez@JASIST'09
  - We propose and implement an **unsupervised approach** to this problem.
    - Relies on information available on pre-existing data.
    - Learn content-based features (i.e., *domain knowledge*).
    - Exploit content-based features to directly learn structure-based features (i.e., *source knowledge*) from test data.
    - Eliminate the need of a user involved in any source specific training process.

Introduction

- Contributions
  - Based on our approach we produced a number of results.
    - ONDUX – On-Demand Unsupervised Learning for IE
      - **SIGMOD'10, IDAR'10, SBBD'11**
    - JUDIE – Joint Unsupervised Structure Discovery and IE
      - **SIGMOD'11**
    - iForm – A Probabilistic Approach for Automatically Filling Form-Based Web Interfaces
      - **WWW'09, PVLDB'10**

Related Work

- **Language for Wrapper Development.**
  - Alternative to general purpose languages such as Perl and Java.
  - Minerva, WEB-OQL
- **Wrapper Induction Methods**
  - Machine Learning usage to semi-automatically induce wrappers.
  - WEIN, StalKer
- **NLP-based Methods**
  - Usage of Natural Language Processing techniques (semantic class, POS)
  - WHISK, TEXTRUNNER
- **Ontology-based Methods**
  - Usage of an ontology and conceptual description of the data of interest
- **HTML-aware Methods**
  - Explore the HTML Structure (Tags) and their representation (DOM)
  - RoadRunner, Webtables

Related Work

Methods	Disadvantages
Language for Wrapper Development.	Rely on the **Regularity of the HTML** format
Wrapper Induction Methods.	Rely on the **Regularity of the HTML** format
NLP-based Methods	Require **Linguistic and Grammatical** Elements
Ontology-based Methods	Require a **huge human effort** to manually create ontologies
HTML-aware Methods	Rely on the **Regularity of the HTML** format

These disadvantages precludes their usage in a large number of textual sources that are available on the Web.

Related Work

- Probabilistic Graph-Based Methods
  - Deal with the limitations of the extraction methods that are based on the HTML structure.
  - Based on probabilistic frameworks such as: Conditional Random Fields (CRF) and Hidden Markov Models (HMM)

  - **Supervised Methods**
    - Rely on human-created training sets to generate graphical models able to extract information
    - Require training data from each source

  - **Unsupervised Methods**
    - Rely on pre-existing datasets for easing the training process of probabilistic methods.
      - Dictionaries, Knowledge Bases

Related Work

- Probabilistic Graph-Based Methods
  - Supervised Methods

Regent Square $228,900 1028 Mifflin Ave.; 6 Bedrooms; 2 Bathrooms. 412-638-7273

Related Work

- Probabilistic Graph-Based Methods
  - Supervised Methods

Regent Square $228,900 1028 Mifflin Ave.; 6 Bedrooms; 2 Bathrooms. 412-638-7273

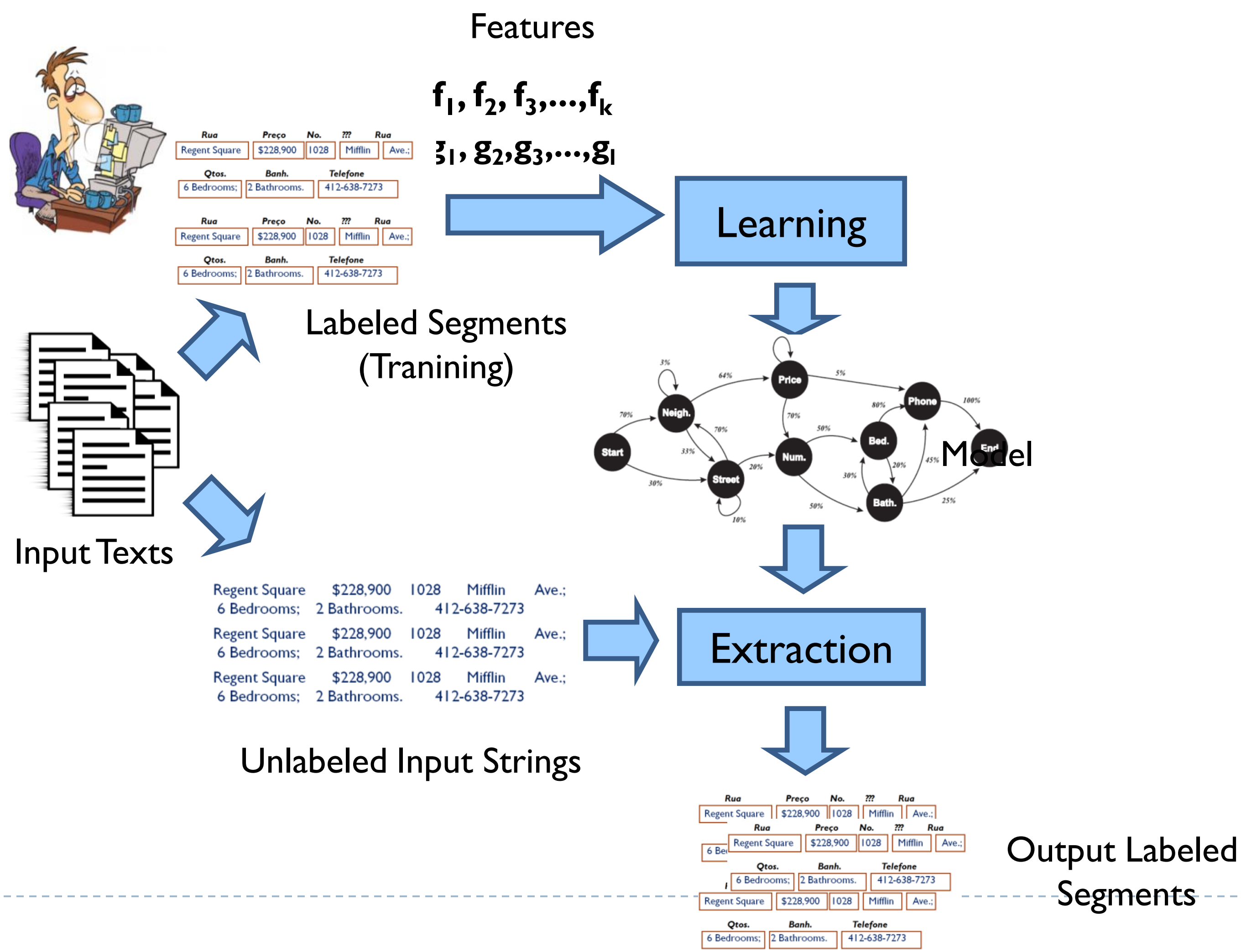
1. <Neighboorhood>**Regent Square**</Neighboorhood>
2. <Price>**$228,900**</Price>
3. <Number>**1028**</Number>
4. <Street>**Mifflin Ave.;**</Street>
5. <Bedroom>**6 Bedrooms**</ Bedroom>
6. <Bathroom>**2 Bathrooms**</Bathroom>
7. <Phone>**412-638-7273**</Phone>

CRF and HMM methods learn from given examples, lexical, style (content) positioning and sequencing (structure) features

Examples are source-dependent

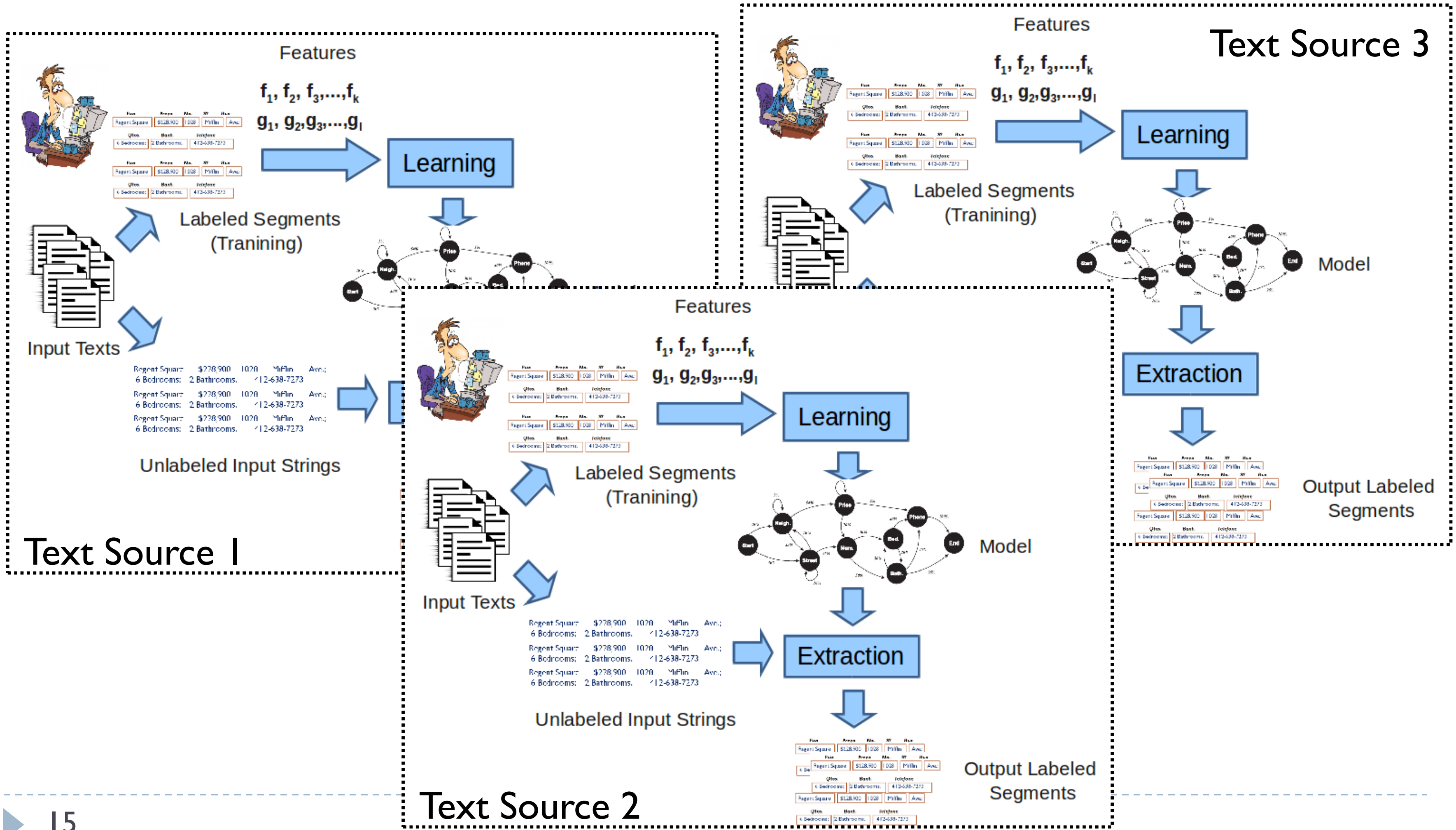
Related Work

- Supervised Methods



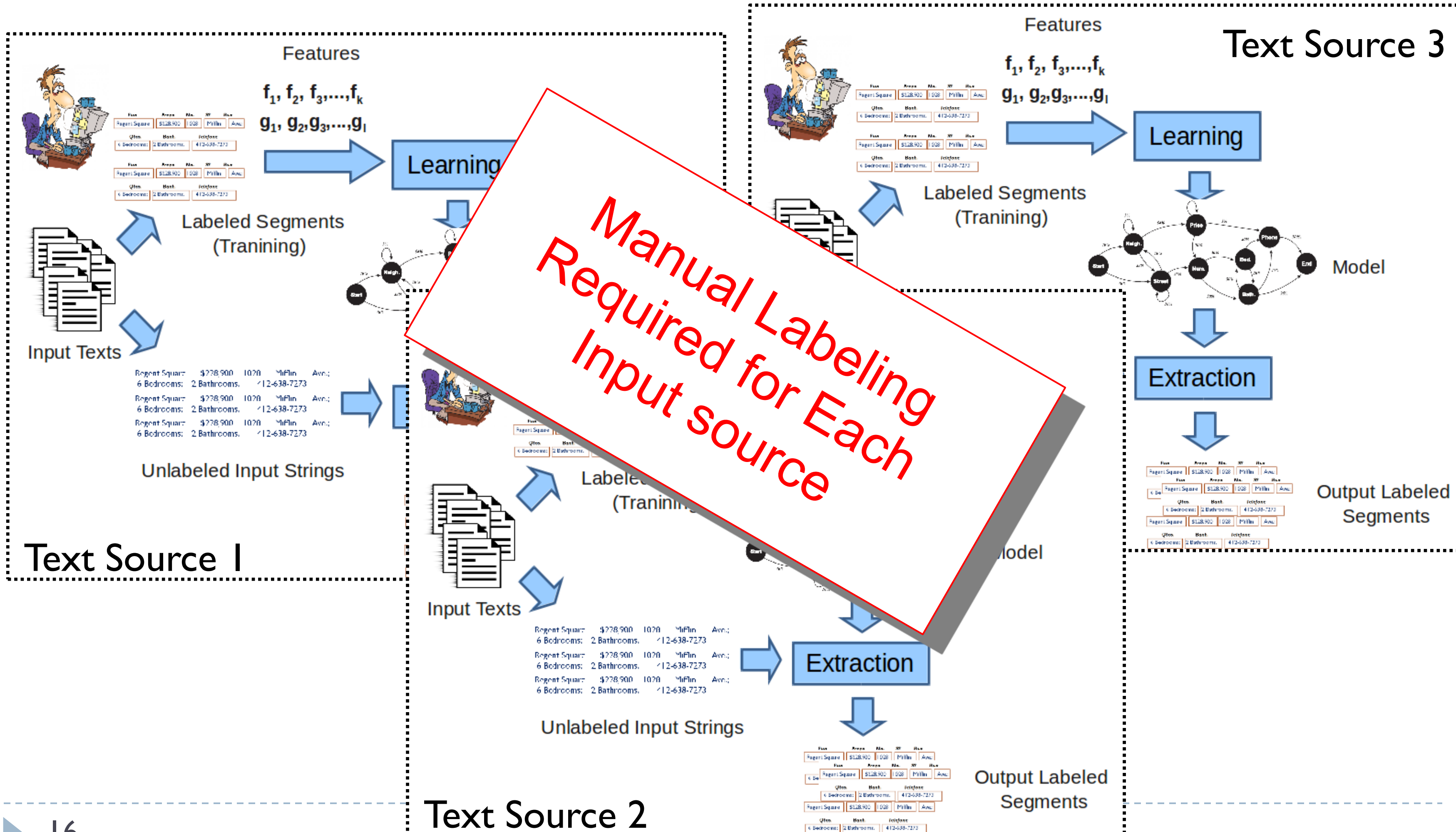
Related Work

- Supervised Methods



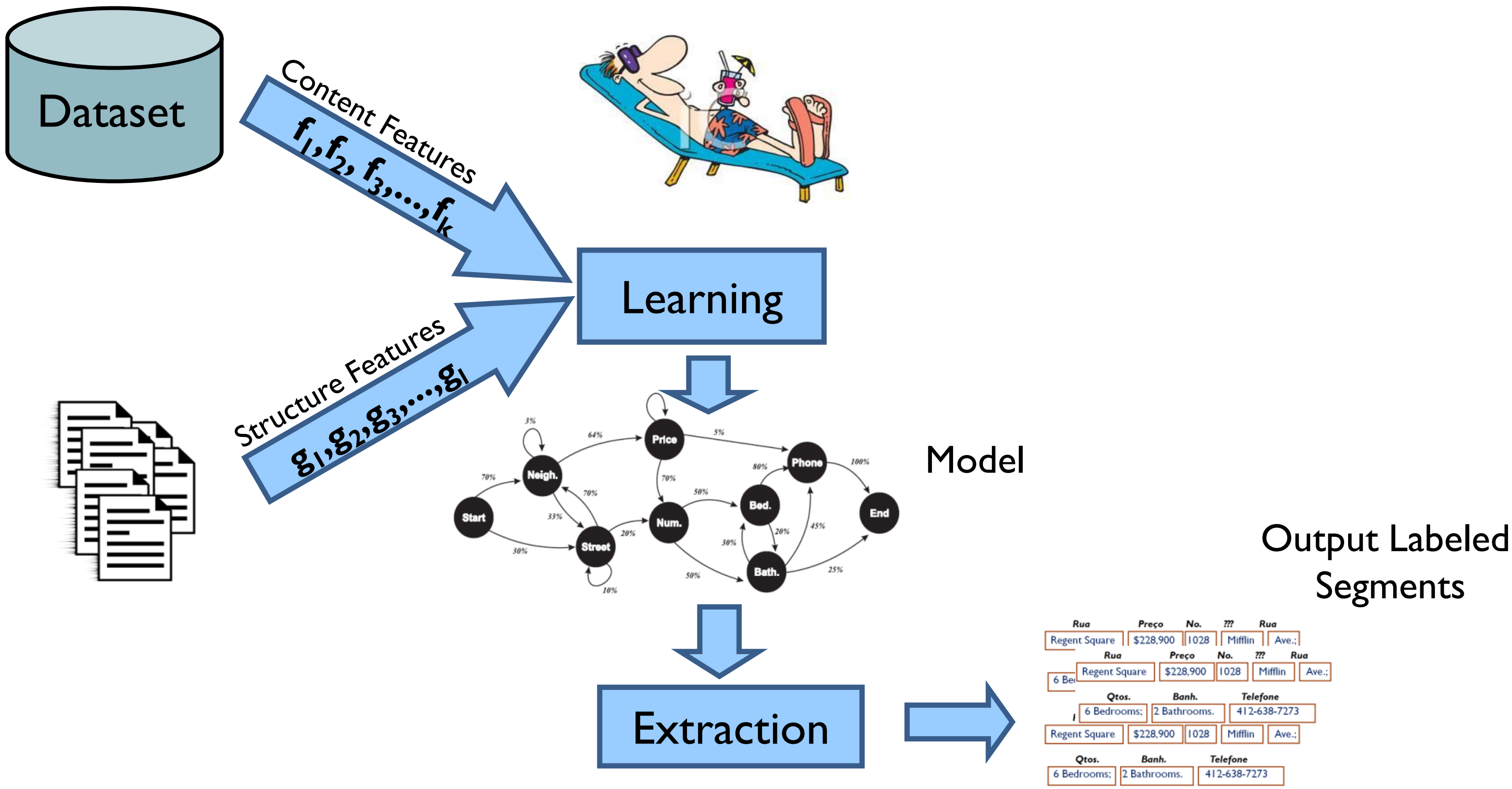
Related Work

- Supervised Methods



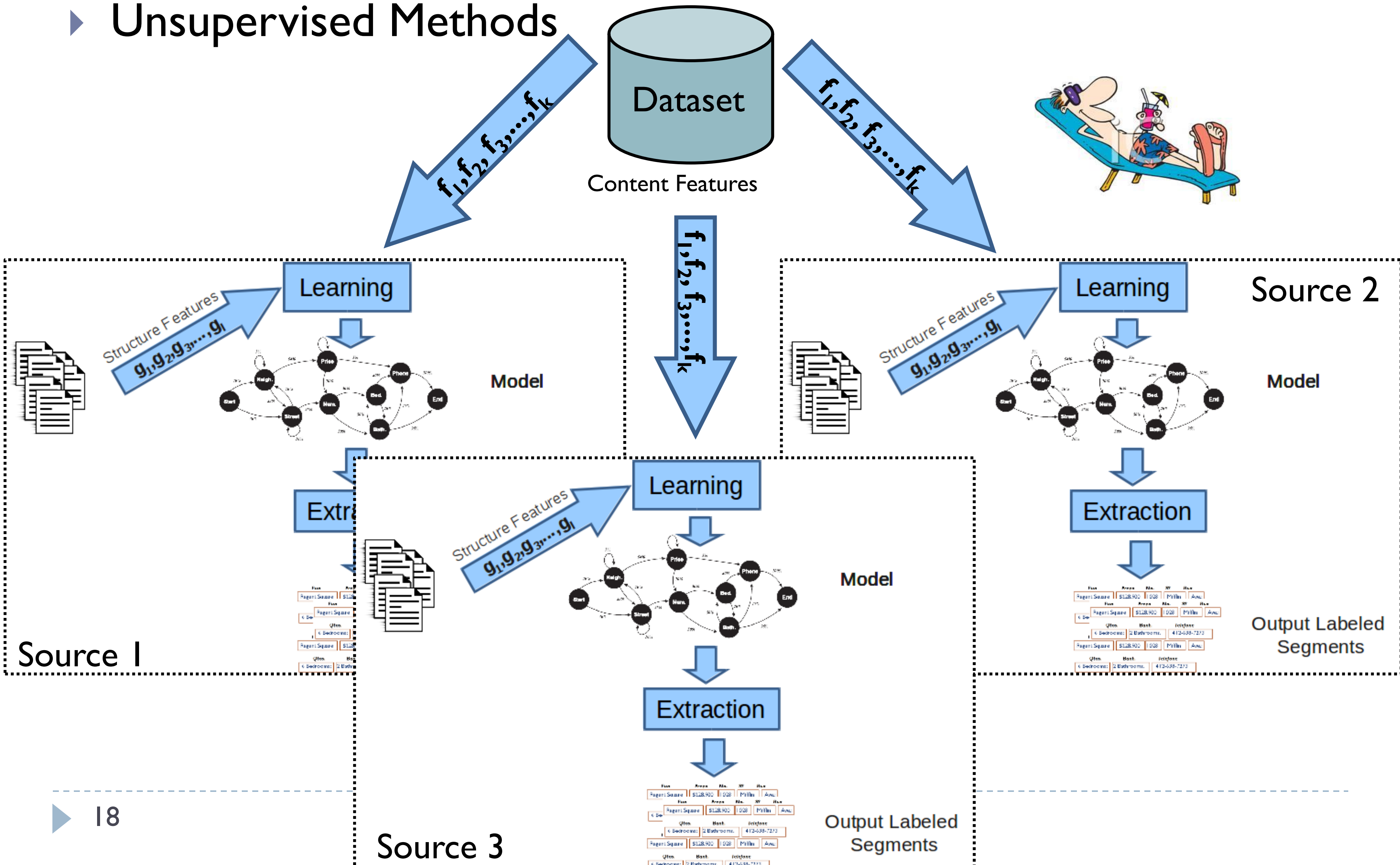
Related Work

- Unsupervised Methods



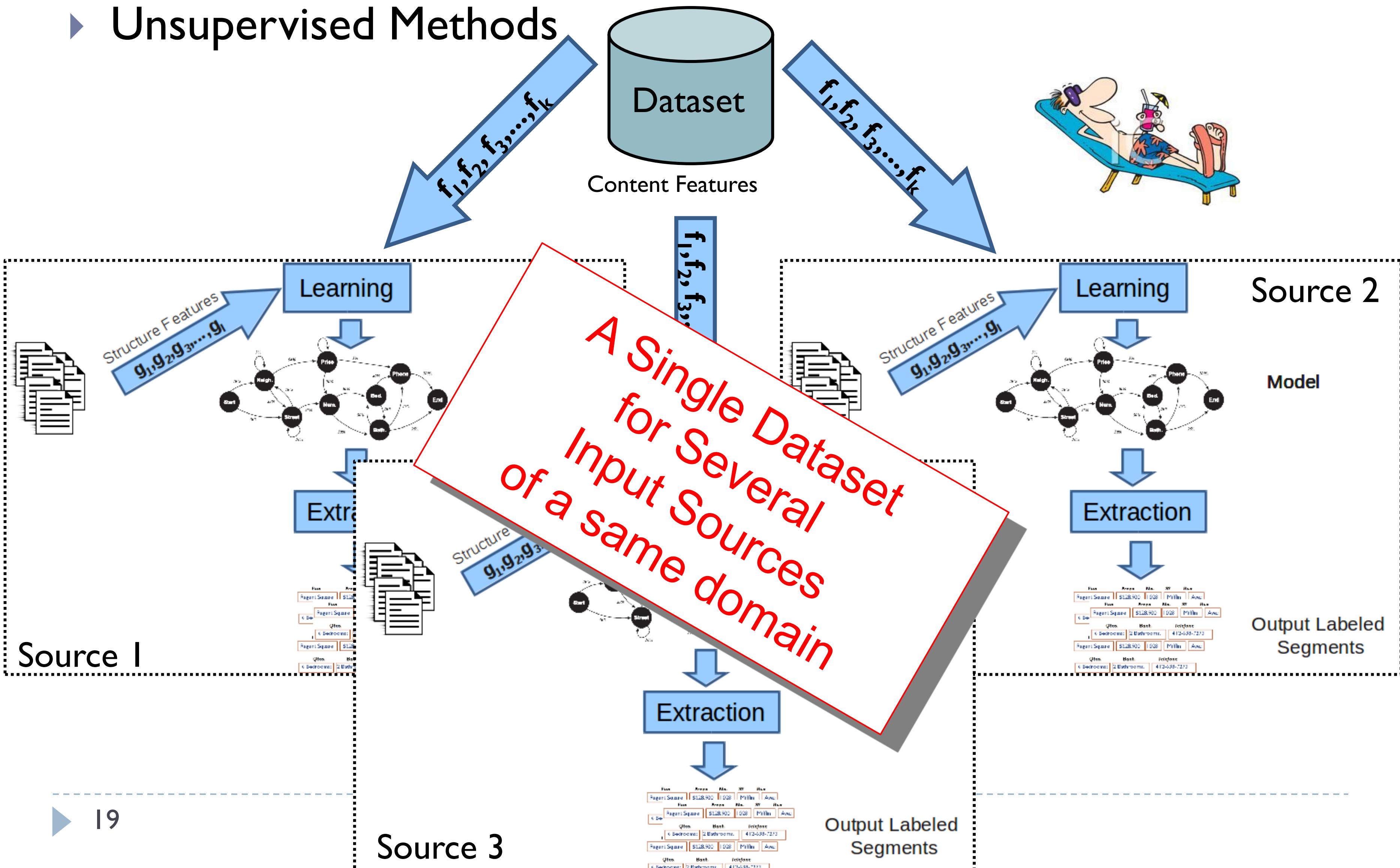
Related Work

- Unsupervised Methods



Related Work

- Unsupervised Methods



Related Work

- Probabilistic Graph-Based Methods

Supervised **X** **UNsupervised**

Supervised	UNsupervised
Hand-labeled examples	**Pre-existing information**
Source Dependent	**Source Indepedent**
Scalability Problem	**Easily adaptable**
Reusability	

Related Work

- Probabilistic Graph-Based Methods
  - Unsupervised Methods
  - [Agichtein et al @ SIGKDD 2004]
    - Usage of Reference Tables to create an unsupervised model using Hidden Markov Models (HMM)
  - [Zhao et al. @ SIAM ICDM 2008]
    - Usage of reference tables to create unsupervised CRF models - (U-CRF)
  - [Sarawagi et al. @ ICDE 2006]
    - Usage of pre-existing data and hand labeled training sets to create an semi-supervised model using CRF

Related Work

- Probabilistic Graph-Based Methods
  - Unsupervised Methods
  - [Agichtein et al @ SIGKDD 2004]
    - Usage of Reference Tables to create an unsupervised model using Hidden Markov Models (HMM)
  - [Zhao et al. @ SIAM ICDM 2008]
    - Usage of reference tables to create unsupervised CRF models - (U-CRF)

Both models assume single positioning and ordering of attributes in all test instances.

  - [Sarawagi et al. @ ICDE 2006]
    - Usage of pre-existing data and hand labeled training sets to create an semi-supervised model using CRF

Our Proposed Approach for IETS

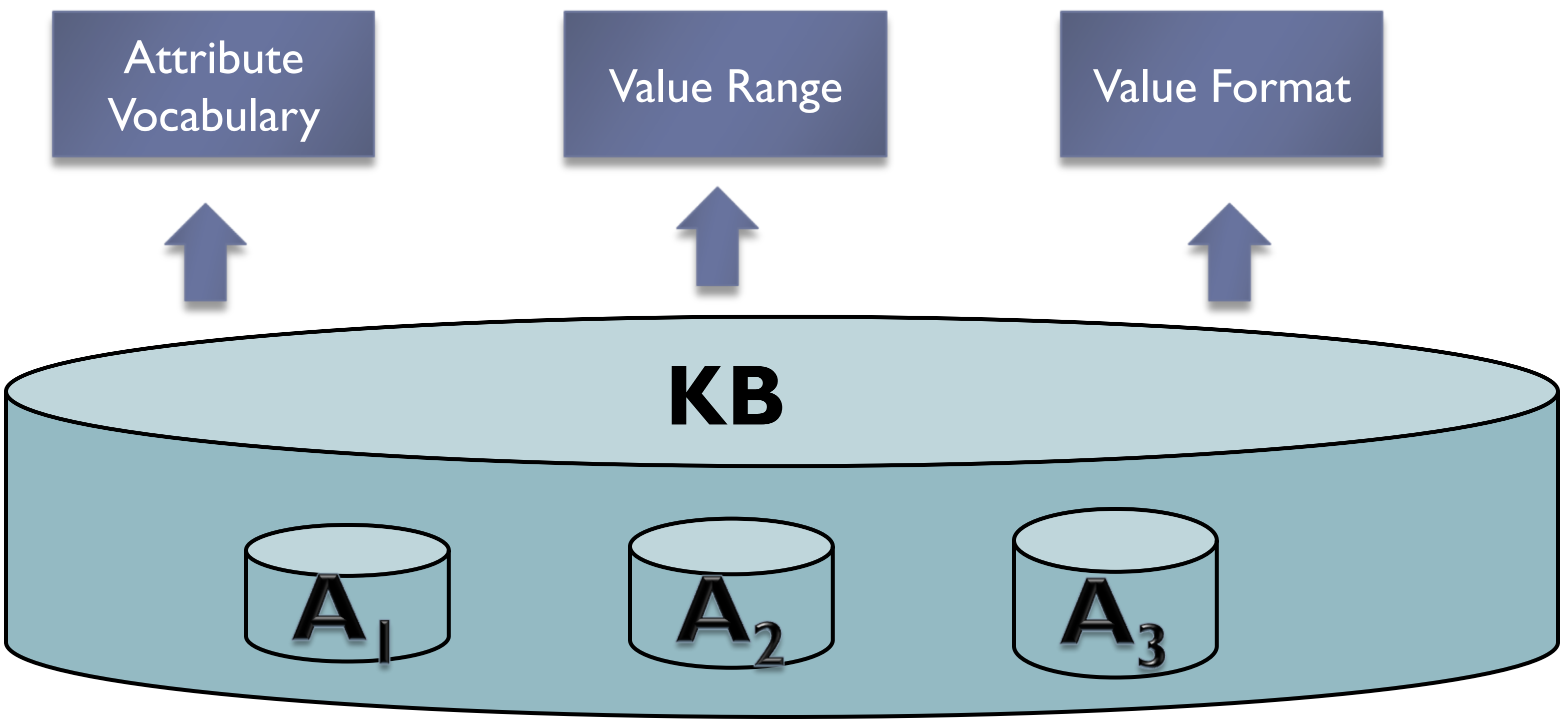
Our Proposed Approach for IETS

- Knowledge Bases
  - Set of pairs KB = $\{(m_1, O_1), \ldots, (m_n, O_n)\}$
  - Easily built from pre-existing sources
    - Bibliographic DBs, Freebase, **Wikipedia**, etc.

$$
\begin{aligned}
K &= \{ \quad \langle Author, O_{Author} \rangle, \langle Title, O_{Title} \rangle \} \\
O_{Author} &= \{ \quad \text{"J. K. Rowling","Galadriel Waters", "Beatrix Potter"} \} \\
O_{Title} &= \{ \quad \text{"Harry Potter and the Half-Blood Prince",} \\
& \qquad \text{"A Guide to Harry Potter", "The Rabbit's Halloween"} \}
\end{aligned}
$$

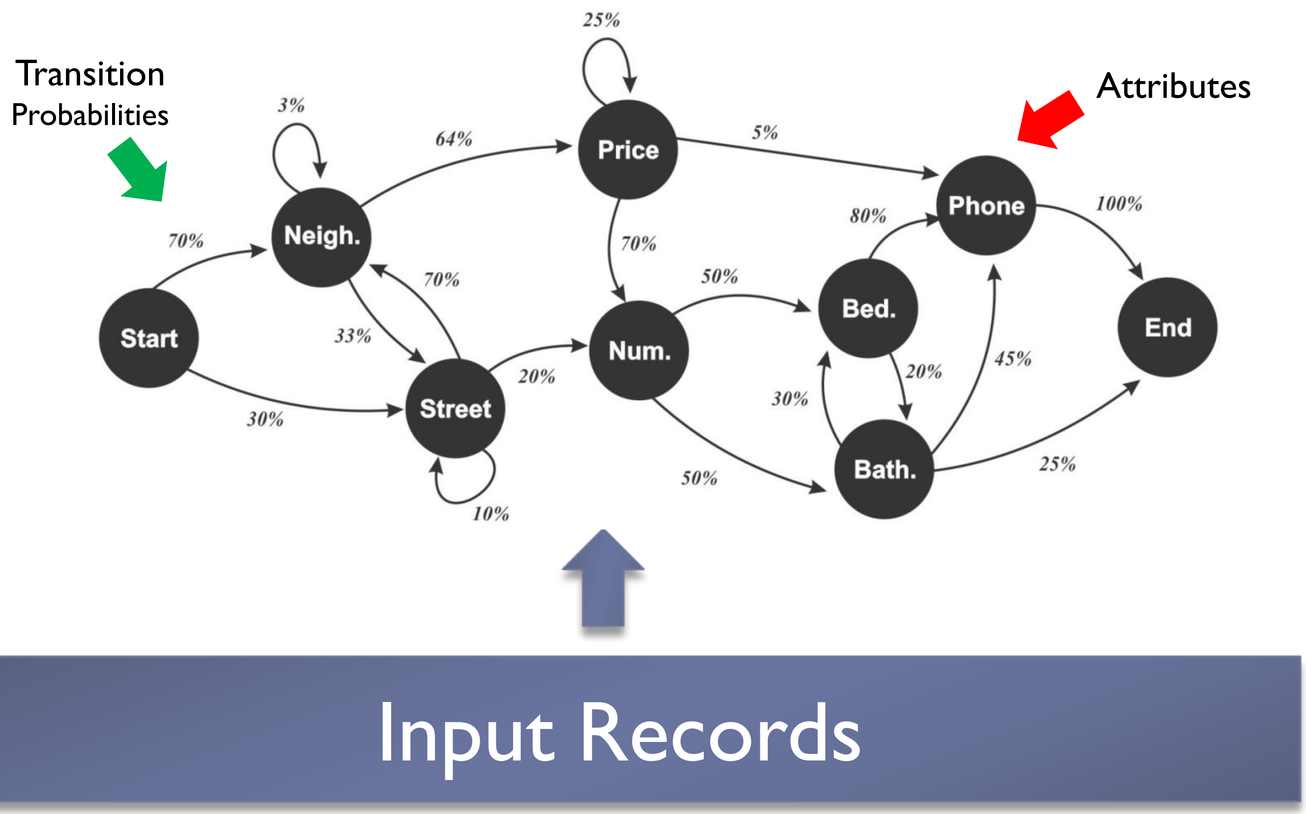
Our Proposed Approach for IETS

- Our approach relies on two types of features:
  - State or content-based features;



Our Proposed Approach for IETS

- Our approach relies on two types of features:
  - Transition or Structure-based Features;

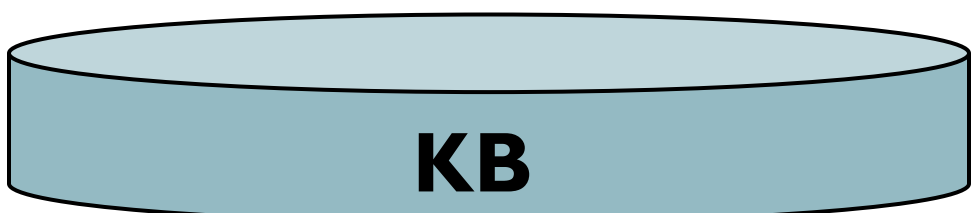


Our Proposed Approach for IETS

- Knowledge bases implicitly encode domain knowledge.
  - Very suitable source for learning content-based features

- Attribute Vocabulary
  - Exploit the common vocabulary often shared by values of textual attributes

$$AF(s,A) = \frac{\sum_{t \in T(A) \cap T(s)} fitness(t,A)}{|T(s)|} \qquad fitness(t,A) = \frac{f(t,A)}{N(t)} \times \frac{f(t,A)}{f_{max}(A)}$$

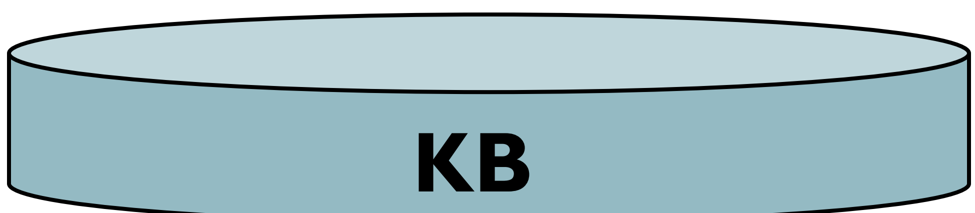


Our Proposed Approach for IETS

- Knowledge bases implicitly encode domain knowledge.
  - Very suitable source for learning content-based features

- Attribute Value Range
  - For the case of numeric candidate values, it measures the similarity between a numeric value and the set of values of a numeric attribute

$$NM(s, A) = e^{-\frac{v_s - \mu}{2\sigma^2}}$$

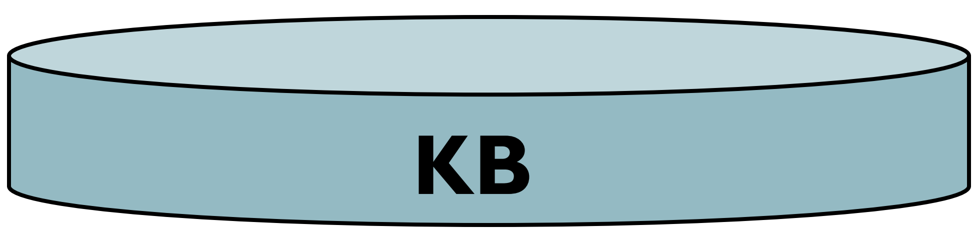


Our Proposed Approach for IETS

- Knowledge bases implicitly encode domain knowledge.
  - Very suitable source for learning content-based features

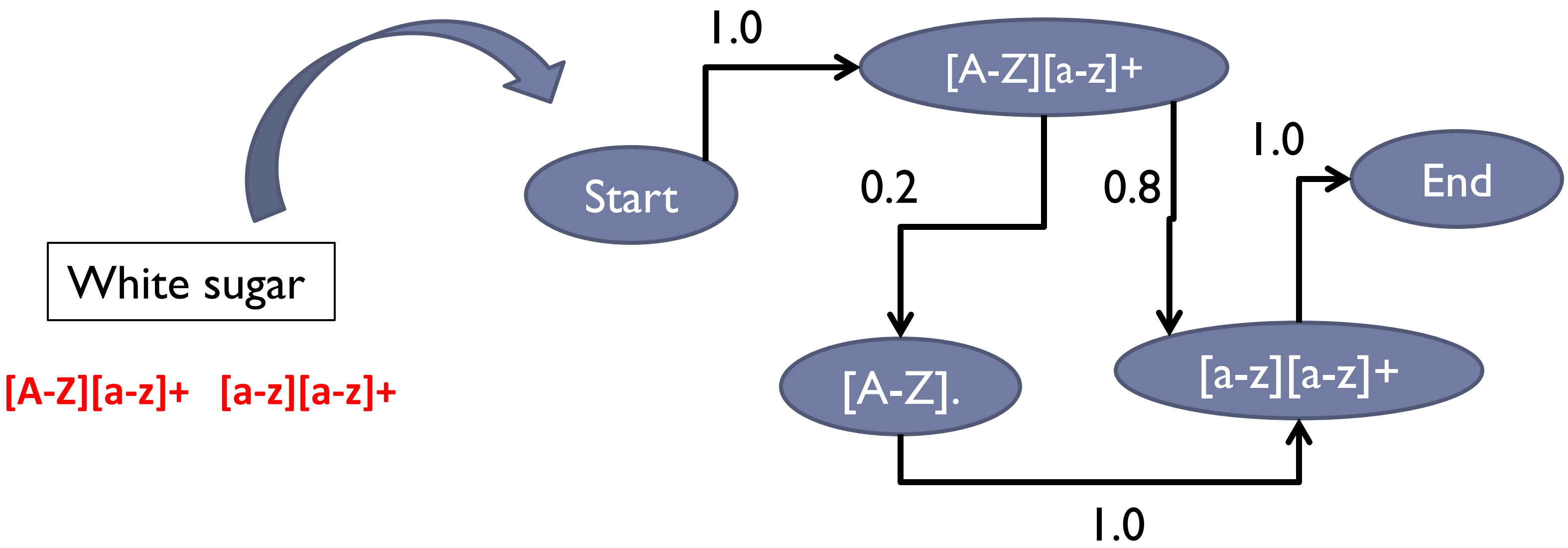
- Attribute Value Format
  - Exploits the common format often used to represent values of some attributes.

$$format(s, A) = \frac{\sum_{\langle n_x, n_y \rangle \in path(s)} w(n_x, n_y)}{|path(s)|}$$



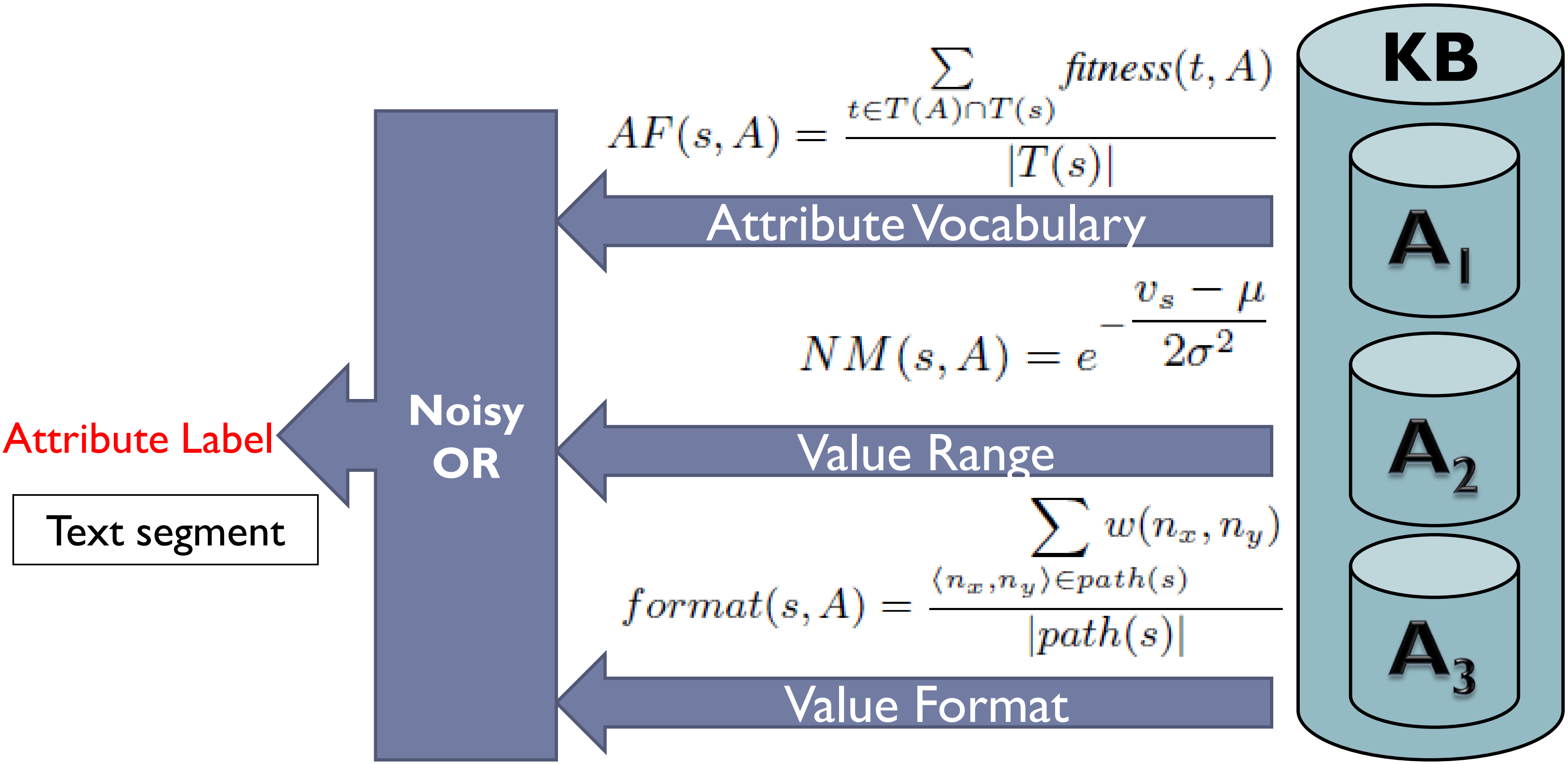
Our Proposed Approach for IETS

- Attribute Value Format (Style)
  - First a Markov model is generated for each attribute.
  - Computes the probability of the input mask sequence represents a path in each Markov model of each attribute.



Our Proposed Approach for IETS

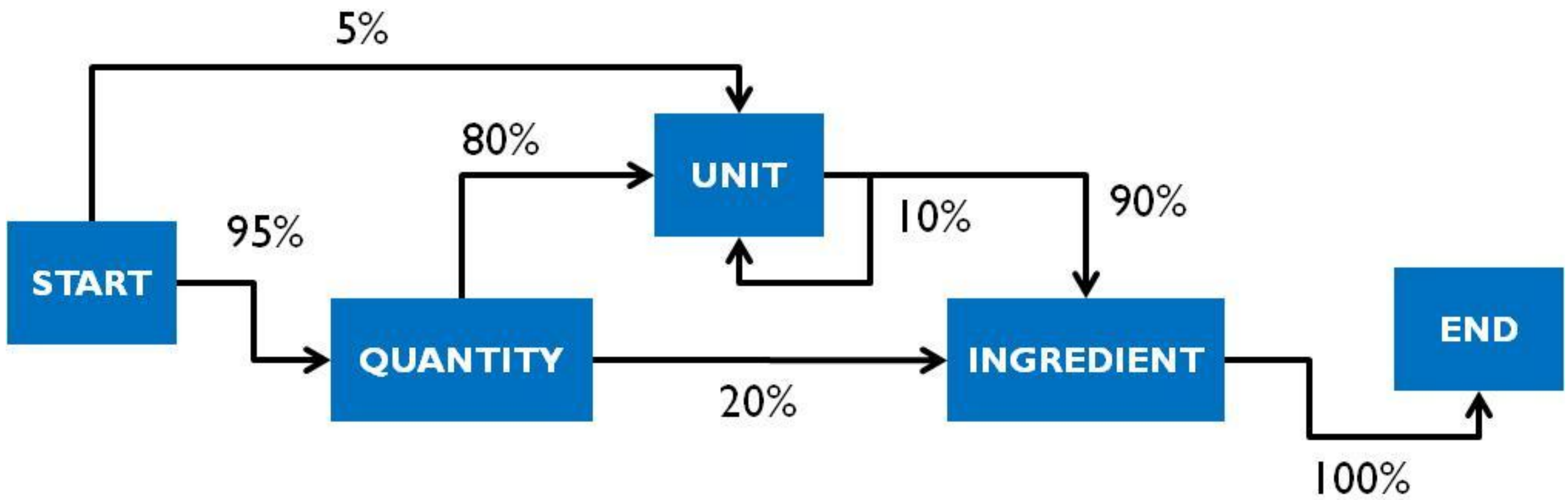
- Content-based Features



Our Proposed Approach for IETS

- Structure-based features are automatically induced from content-based Features
  - HMM-like graph called Positioning and Sequencing Model (PSM)
- Positioning and Sequencing Model
  - Automatically learned *On-Demand* from test instances
  - No *a priori* training required
- Structure-based features
  - Dependent of the placement of attributes values on the input
  - Thus, they are input-dependent

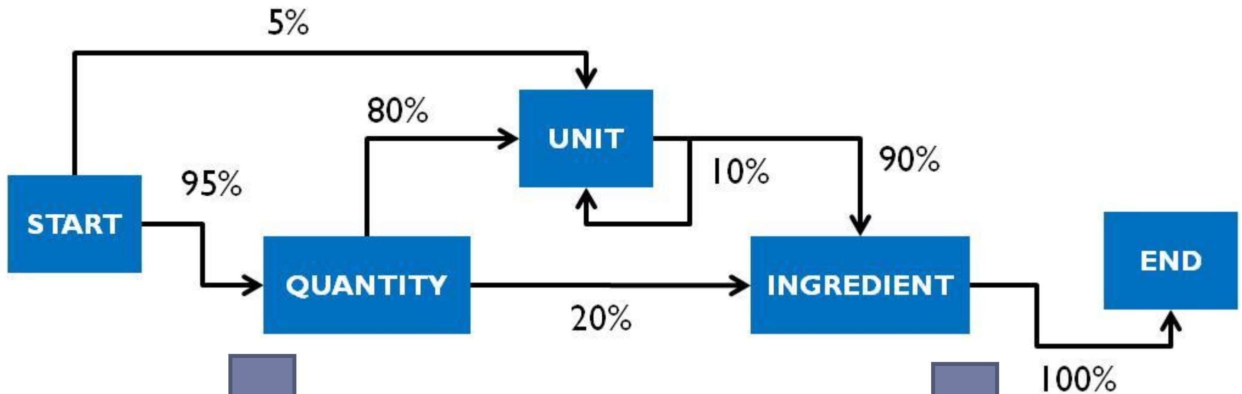
Our Proposed Approach for IETS



$$p_{i,k} = \frac{\text{\# of observations of } \ell_i \text{ in } k}{\text{Total \# of candidate values in } k} \quad t_{i,j} = \frac{\text{\# of transitions from } \ell_i \text{ to } \ell_j}{\text{Total \# of transitions out of } \ell_i}$$

$$pos(s_k, A_i, R) = p_{i,k} \qquad seq(s_k, A_i, R) = t_{i,j}$$

Our Proposed Approach for IETS



Content-related features

$$p_{i,k} = \frac{\#\text{ of observations of } \ell_i \text{ in } k}{\text{Total }\#\text{ of candidate values in } k}$$

$$pos(s_k, A_i, R) = p_{i,k}$$

$$t_{i,j} = \frac{\#\text{ of transitions from } \ell_i \text{ to } \ell_j}{\text{Total }\#\text{ of transitions out of } \ell_i}$$

$$seq(s_k, A_i, R) = t_{i,j}$$

Noisy OR

Attribute Label | Text Segment

Our Proposed Approach for IETS

- Combination Strategy
  - *Bayesian Noise-OR-Gate*

$$or(p_1,\ldots,p_n) = 1-((1-p_1)\times\ldots\times(1-p_n))$$

  - We assume that the features we use exploit different properties of the attributes of the KB, i.e., they are **independent.**

  - Probabilistic methods such as CRF and HMM deploy optimization process to combine their features.
    - Not using optimization can, in theory, lead to sub-optimal results, our experiments demonstrates that our combination works very well in practice.

Our Proposed Approach for IETS

- Based on our approach
  - We developed unsupervised information extraction by text segmentation methods
  - ONDUX
    - **On Demand Unsupervised Information Extraction**
  - JUDIE
    - **Joint Unsupervised Structure Discovery and Information Extraction**
  - iForm
    - **A Probabilistic Approach for Automatically Filling Form-Based Web Interfaces**

ONDUX

On-Demand Unsupervised Learning for Information Extraction

Cortez et al. - SIGMOD 2010, Cortez and Silva – IDAR 2010

ONDUX

- Deals with text documents containing implicit semi-structured data records
  - Addresses
  - Bibliographic References
  - Classified Ads
  - Product Descriptions

Postal Address

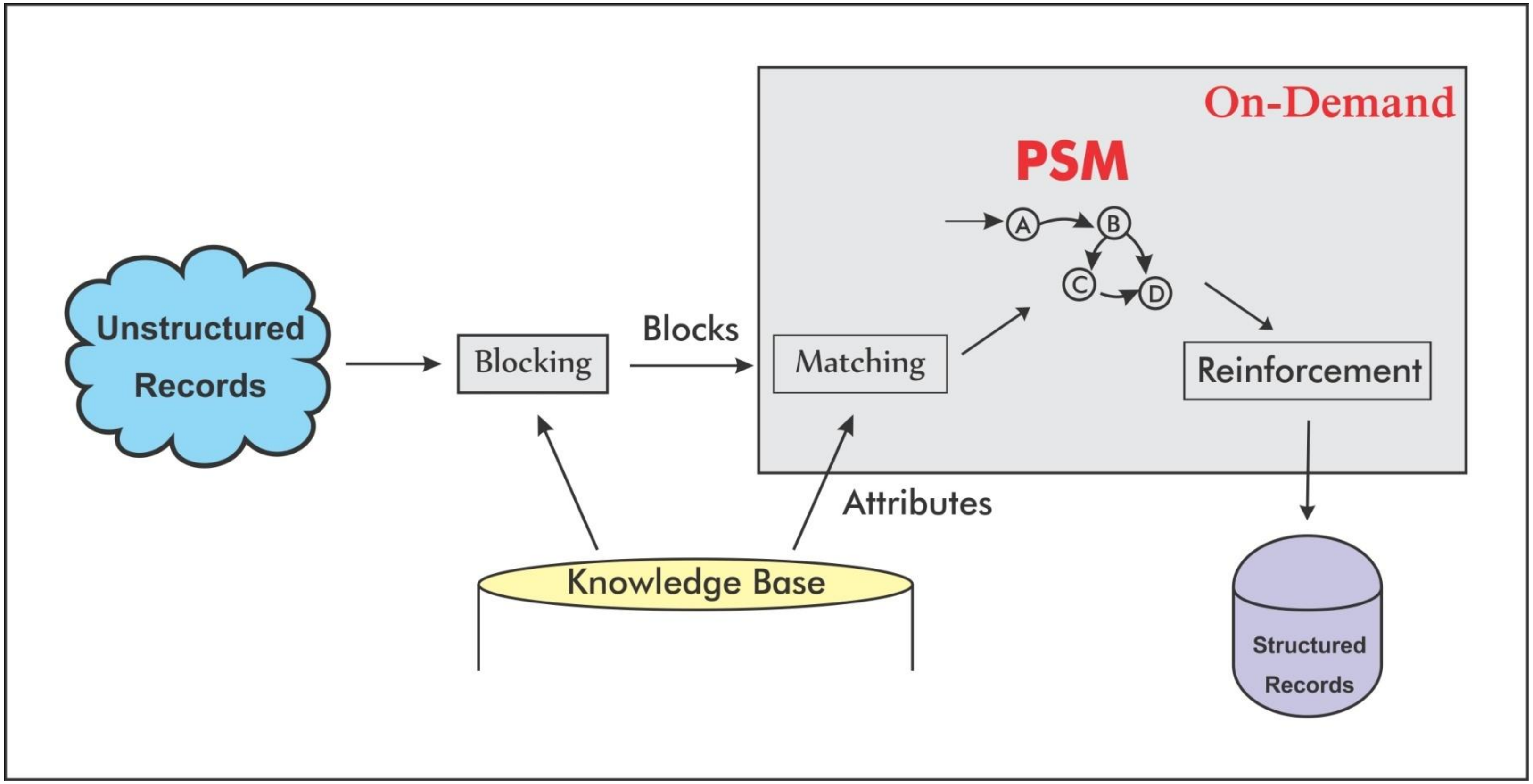
Dr. Robert A. Jacobson, 8109 Harford Road, Baltimore, MD 21214

Bibliographic Reference

Pável Calado, Marco Cristo, Marcos André Gonçalves, Edleno S. de Moura, Berthier Ribeiro-Neto, Nivio Ziviani. Link-based similarity measures for the classication of Web documents. JASIST, v. 57 n.2, p. 208-221, January 2006

ONDUX

- General View



ONDUX

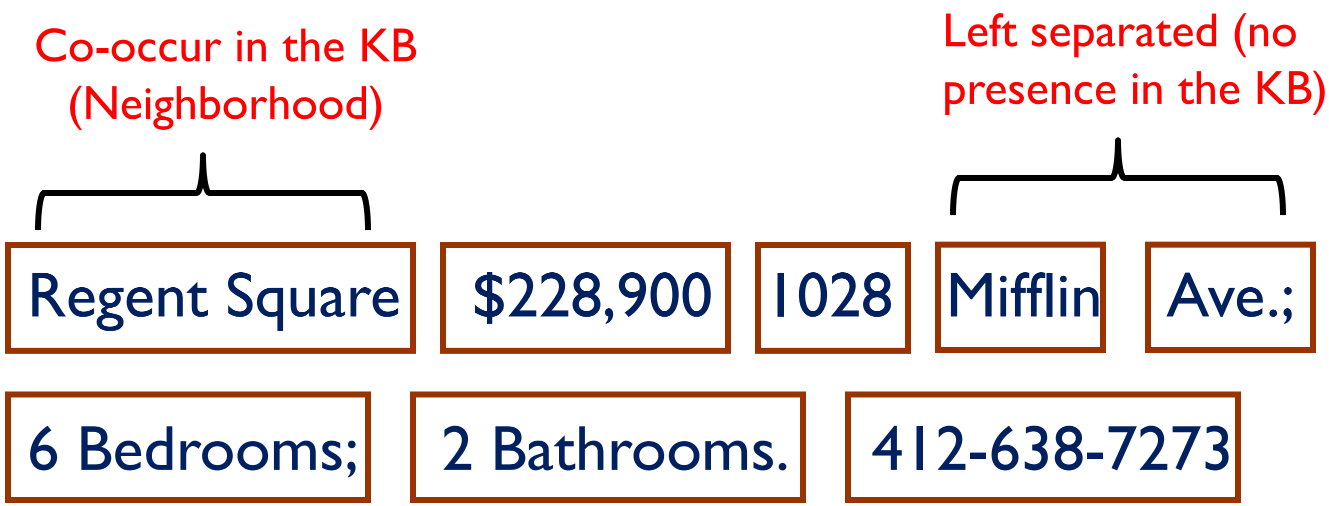
- Blocking
  - Split the input text in *substrings* called *blocks;*
  - Consider the co-occurrence of consecutive terms based on the KB

Regent Square $228,900 1028 Mifflin Ave.;

6 Bedrooms; 2 Bathrooms. 412-638-7273

ONDUX

- Blocking
  - Split the input text in *substrings* called *blocks;*
  - Consider the co-occurrence of consecutive terms based on the KB



ONDUX

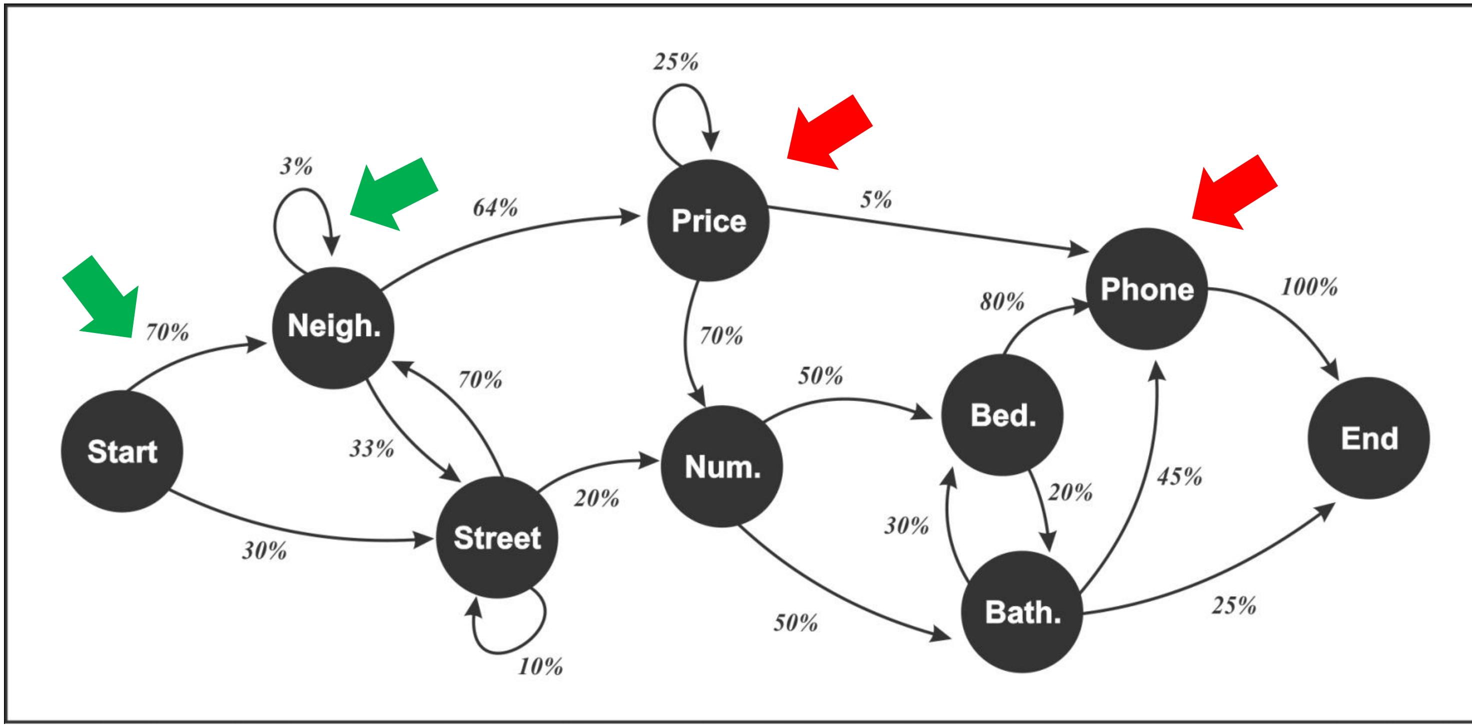
- Matching
  - Associate each blocks with attributes according to content-based features.
    - Attribute Vocabulary $AF(s, A) = \frac{\sum_{t \in T(A) \cap T(s)} fitness(t, A)}{|T(s)|}$
    - Value Range $NM(s, A) = e^{-\frac{v_s - \mu}{2\sigma^2}}$

Street	***Price***	***No.***	***???***	***Street***
Regent Square	$228,900	1028	Mifflin	Ave.;

Bed.	***Bath.***	***Phone***
6 Bedrooms;	2 Bathrooms.	412-638-7273

ONDUX

- Reinforcement – PSM



Ordering and Positioning Features are learned *On-Demand* based on the test instances trough the Matching Phase

ONDUX

- Reinforcement
  - Once the PSM is built, we combine the content-based and the structure-based features using the Bayesian operator *OR*.

Street	***Price***	***No.***	***???***	***Street***
Regent Square	$228,900	1028	Mifflin	Ave.;

Bed.	***Bath.***	***Phone***
6 Bedrooms;	2 Bathrooms.	412-638-7273

ONDUX

- Reinforcement
  - Once the PSM is built, we combine the content-based and the structure-based features using the Bayesian operator *OR*.

Neighborhood	***Price***	***No.***	***Street***
Regent Square	$228,900	1028	Mifflin Ave.;

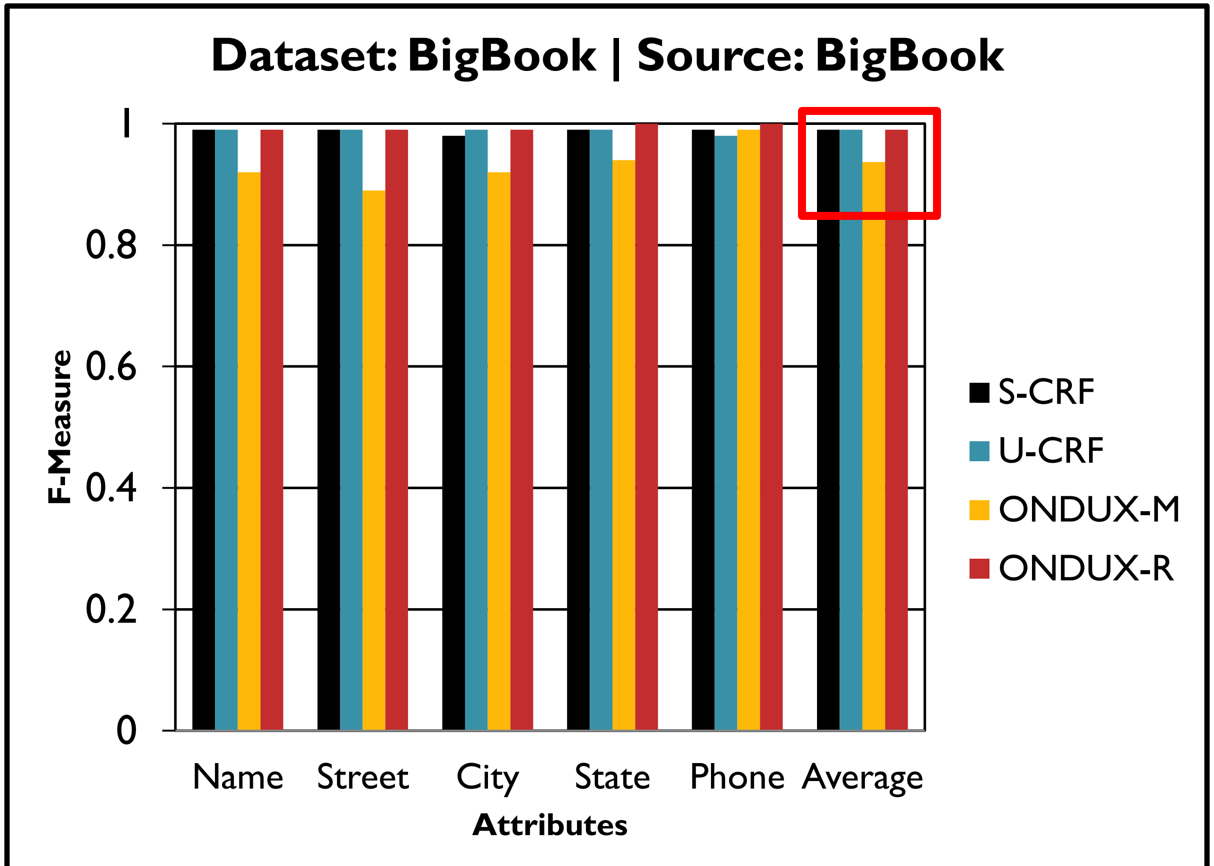
Bed.	***Bath.***	***Phone***
6 Bedrooms;	2 Bathrooms.	412-638-7273

ONDUX - Experiments

- Setup
  - We tested our proposed method with several sources from 3 distinct domains:
    - Addresses
    - Bibilographic Data
    - Classified Ads
  - Metrics
    - Precision, Recall and F-Measure
      - T-Test for the statistical validation of the results
  - Baselines
    - U-CRF and S-CRF

ONDUX - Experiments

- Extraction Quality



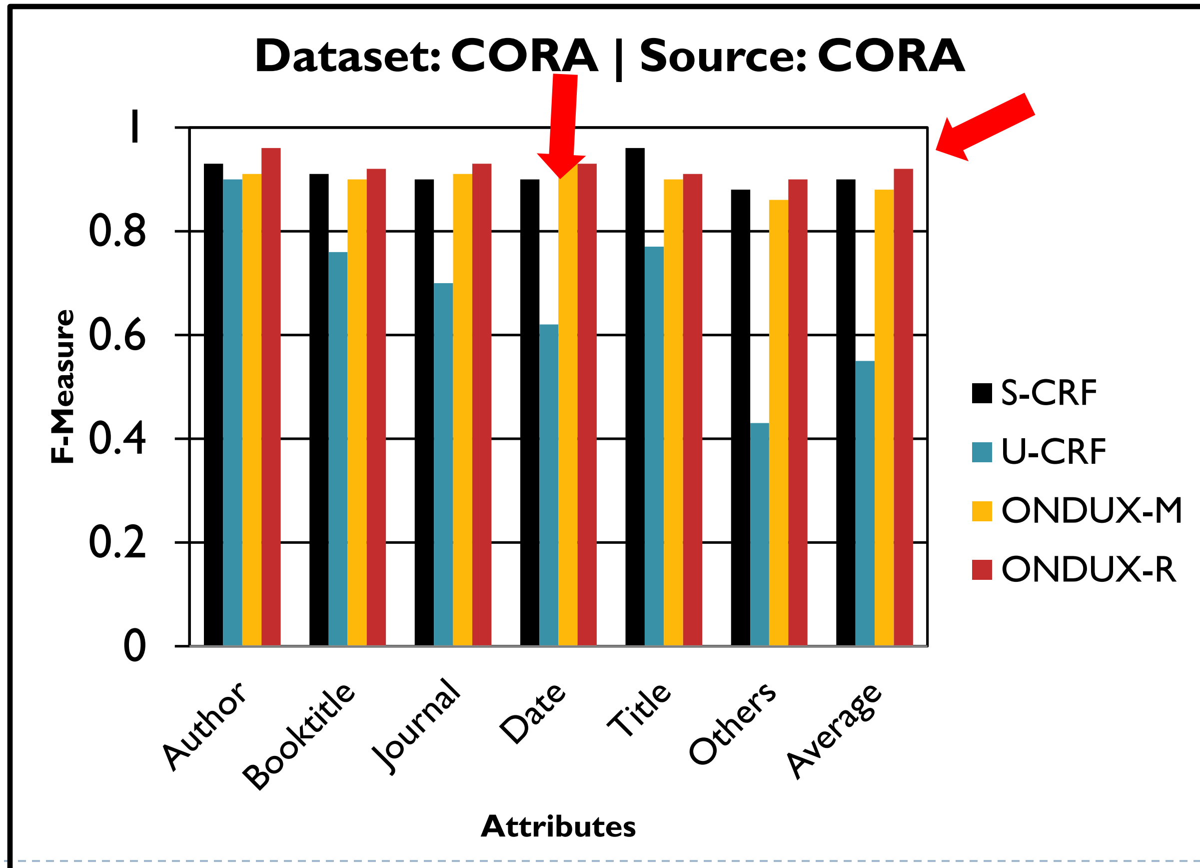
U-CRF results similar to Zhao@SICDM (validation)

Dataset follows the single order assumption

After Reinforcement ONDUX achieved similar quality

ONDUX - Experiments

- Extraction Quality



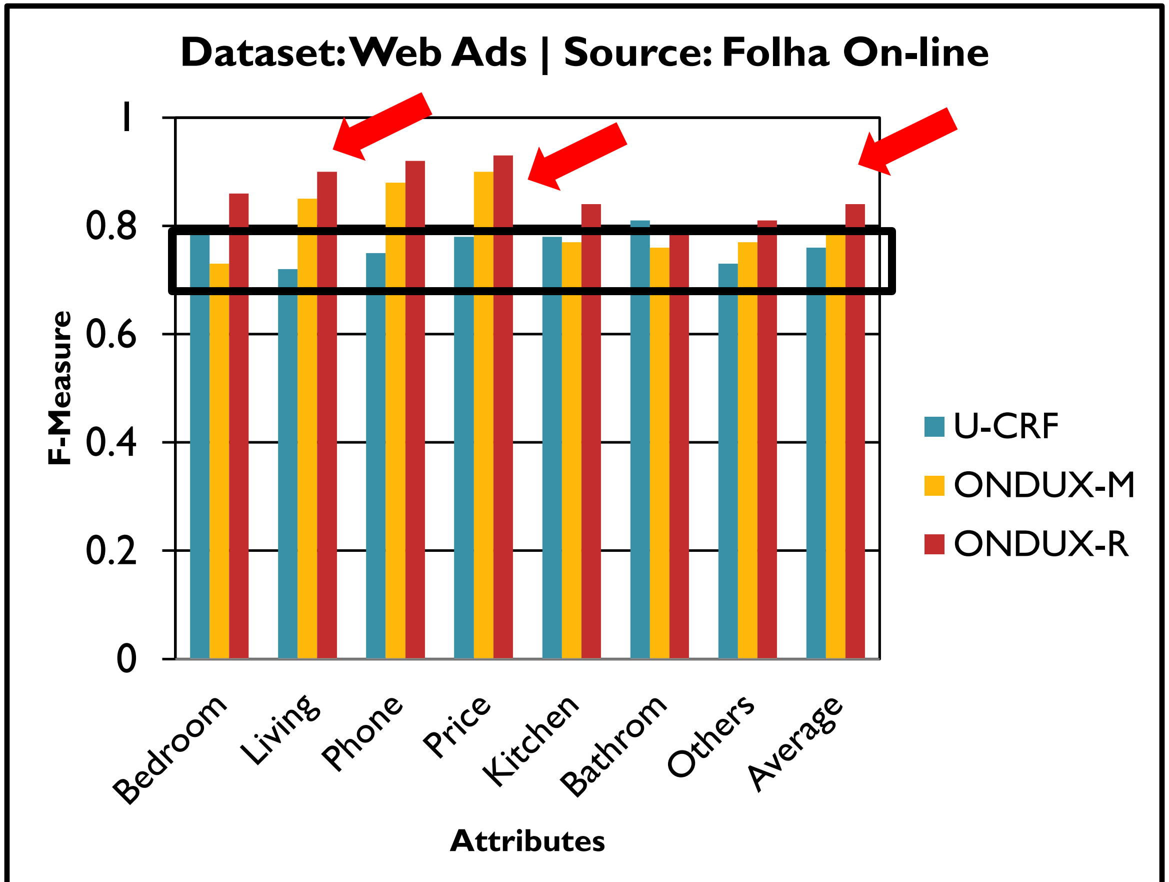
CORA includes a variety of citation styles (conference, journal, books, etc,)

S-CRF achieved results higher than U-CRF due to the hand-labeled training

In general, ONDUX outperformed CRF models

ONDUX - Experiments

- Extraction Quality



U-CRF presented a poor performance (very heterogeneous dataset)

Due to the Matching Phase and the PSM that is learned *On-Demand*, ONDUX achieve very high quality results

JUDIE

Joint Unsupervised Structure Discovery and Information Extraction

Cortez et al. - SIGMOD 2011

JUDIE

Chocolate Cake Recipe

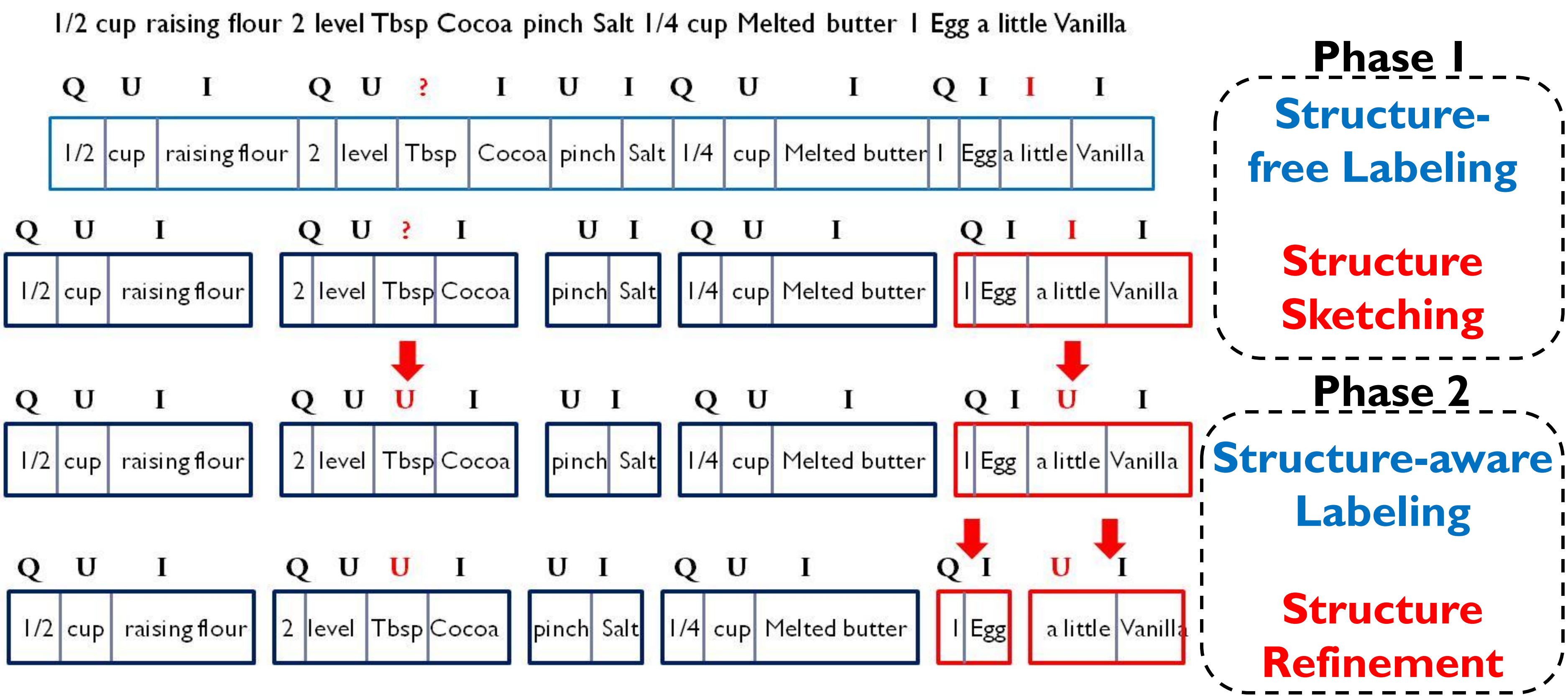
1/2 cup butter 2 eggs 4 cups white sugar ground cinnamon 2 tablespoons dark rum 6 chopped pecans 1/2 cup milk 1 1/2 cups applesauce 2 cups all-purpose flour 1/4 cup cocoa powder 2 teaspoons baking soda 1/8 teaspoon salt 1 cup raisins 1/4 cup dark rum

Quantity	Unit	Ingredient
1/2	cup	butter
2		eggs
4	cups	white sugar
		ground cinnamon
2	tablespoons	dark rum
6		chopped pecans

JUDIE

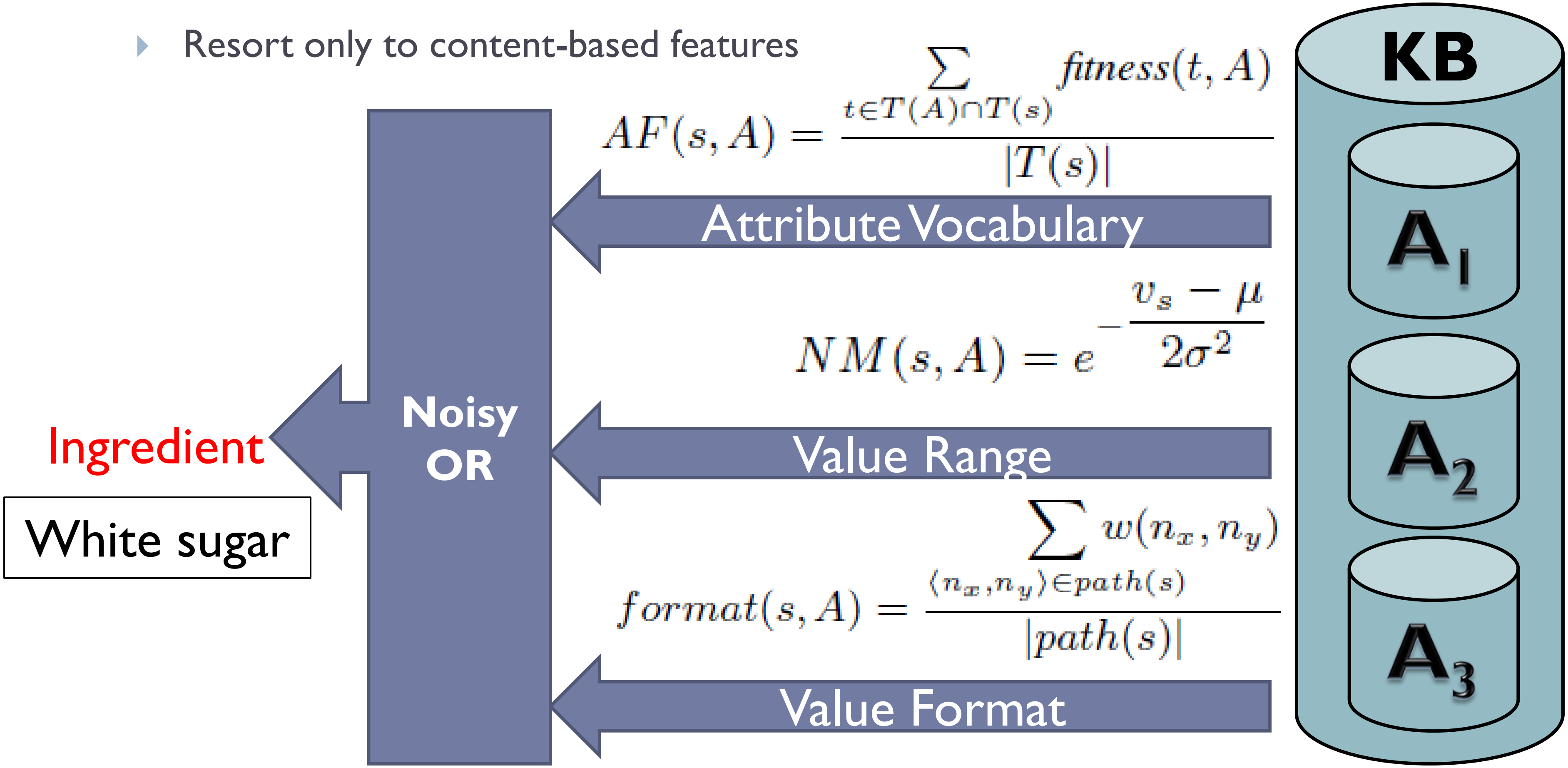
- Joint Unsupervised Structure Discovery and Information Extraction
  - Detects the structure of each individual record being extracted without any user intervention
  - Looks for frequent patterns of label repetitions or **cycles**
- Integrates this algorithm in the IE process
  - Accomplished by successive refinement steps that alternate information extraction and structure discovery.

JUDIE


1/2 cup raising flour 2 level Tbsp Cocoa pinch Salt 1/4 cup Melted butter 1 Egg a little Vanilla
Q U I Q U ? I U I Q U I Q I I I
1/2 cup raising flour 2 level Tbsp Cocoa pinch Salt 1/4 cup Melted butter 1 Egg a little Vanilla
Phase 1
Structure-free Labeling
Structure Sketching
Phase 2
Structure-aware Labeling
Structure Refinement

JUDIE – Structure-free Labeling

- *What is the best label for each segment?*
  - No information on the structure of the data records
  - Resort only to content-based features



JUDIE – Structure-free Labeling

- Initially labels potential values with attribute names.
  - No information on the structure of the data records
  - Resort only to content-based features
    - Learned from the pre-existing KB

1/2 cup raising flour 2 level Tbsp Cocoa pinch Salt 1/4 cup Melted butter 1 Egg a little Vanilla

Q	U	I	Q	U	?	I	U	I	Q	U	I	Q	I	I	I
1/2	cup	raising flour	2	level	Tbsp	Cocoa	pinch	Salt	1/4	cup	Melted butter	1	Egg	a little	Vanilla

Limitations: Unmatching: "Tbsp"
Mismatching: "a little"

The SD Algorithm (I)

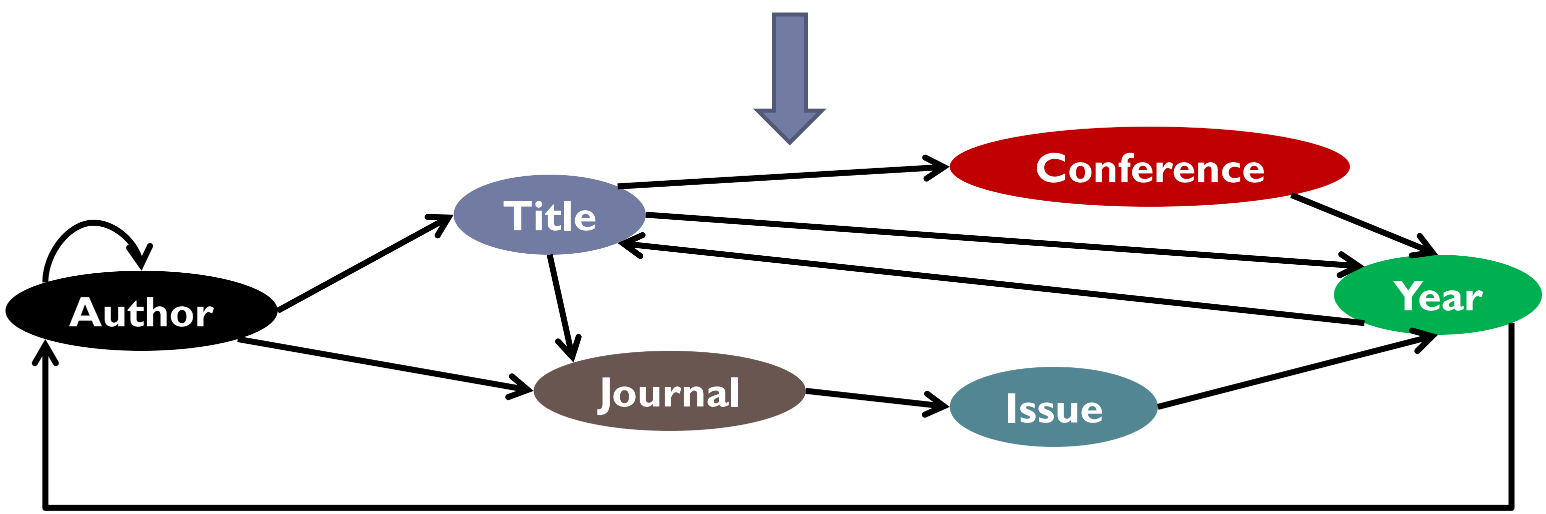
- Uncover the structure of implicit records from the input text.
  - Used in the Structure Sketching and Structure Refinement.

- Takes as input a sequence of labels and generates the structure of each record.
  - Assumption: It is possible to identify patterns of sequences by looking for cycles into a graph (Adjacency Graph) that models the ordering of labels.

The SD Algorithm (II)

- Consider a sequence of labels from a bibliographic reference input text.

Title Conference Year Author Author Title Conference Year Author Title Conference Year … Author Title Journal Issue Year Author Title Journal Issue Year Author Author Journal Issue Year Title Year … Author Title Conference Year Author Author Author Title Journal Issue Year



The SD Algorithm (V)

Dominant Cycles

- Given the set of Coincident cycles that are also viable, the Dominant Cycle are most frequent in the input

- Finally, the algorithm works by first identifying all dominant cycles in the adjacency graph and then processing each of these cycles, the largest cycles being processed first.

- In our given examples, the dominant cycles are:
    1. [Author, Title, Journal, Issue, Year]
    2. [Author, Title, Conference, Year]
    3. [Author, Journal, Issue, Year]
    4. [Title, Conference, Year]
    5. [Title, Year]

JUDIE – Structure Sketching

- Organizes the labeled candidate values into records
  - Induces a structure on the unstructured text input.
  - Outputs labeled values grouped into records
  - Uses a novel algorithm called Structure Discovery (SD)

Q	U	I	Q	U	?	I	U	I	Q	U	I	Q	I	I	I
1/2	cup	raising flour	2	level	Tbsp	Cocoa	pinch	Salt	1/4	cup	Melted butter	1	Egg	a little	Vanilla

Q	U	I
1/2	cup	raising flour

Q	U	?	I
2	level	Tbsp	Cocoa

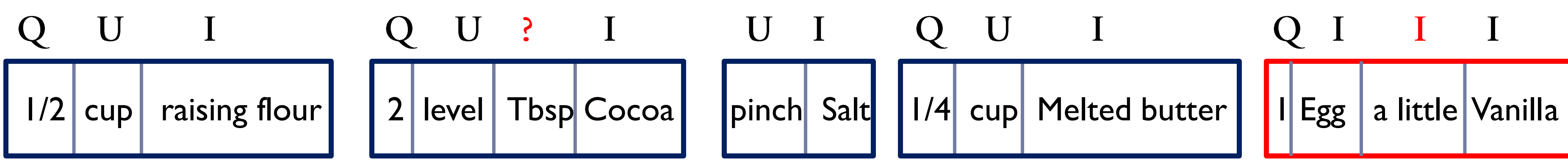
U	I
pinch	Salt

Q	U	I
1/4	cup	Melted butter

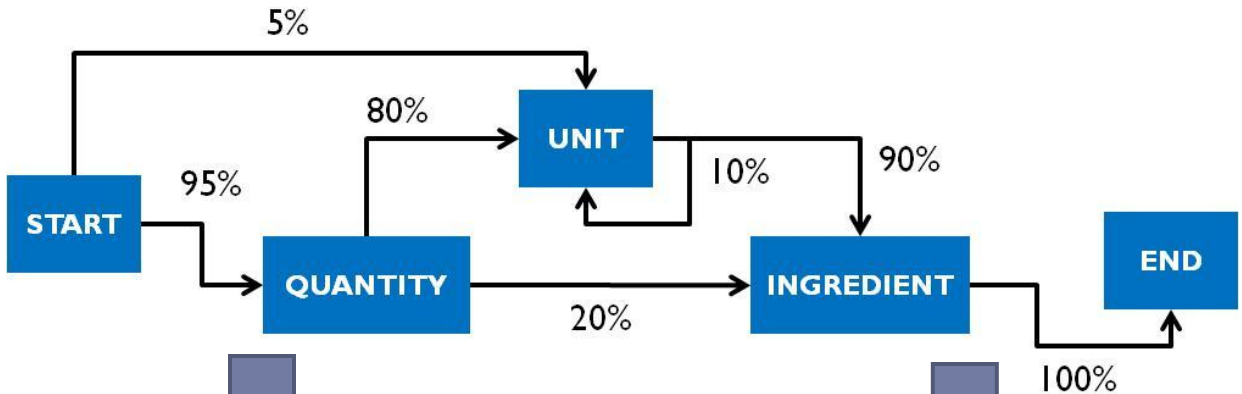
Q	I	I	I
1	Egg	a little	Vanilla

JUDIE – Structure-aware Labeling

- *Now, what is the best label for each segment?*
- We already know some structural information
- Re-labels segments considering **content-based features** and structure-based features
- Structure-based features learned using a graphical model (PSM)



JUDIE – Structure-aware Labeling



Content-related features

$$p_{i,k} = \frac{\text{\# of observations of } \ell_i \text{ in } k}{\text{Total \# of candidate values in } k}$$

$$pos(s_k, A_i, R) = p_{i,k}$$

$$t_{i,j} = \frac{\text{\# of transitions from } \ell_i \text{ to } \ell_j}{\text{Total \# of transitions out of } \ell_i}$$

$$seq(s_k, A_i, R) = t_{i,j}$$

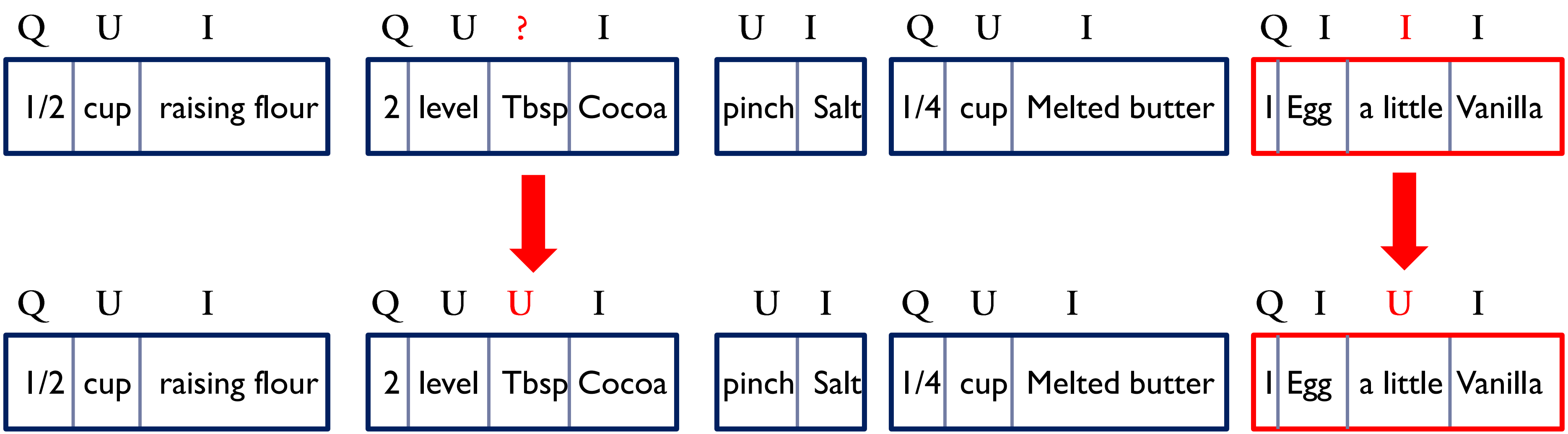
Noisy OR

Unit

A little

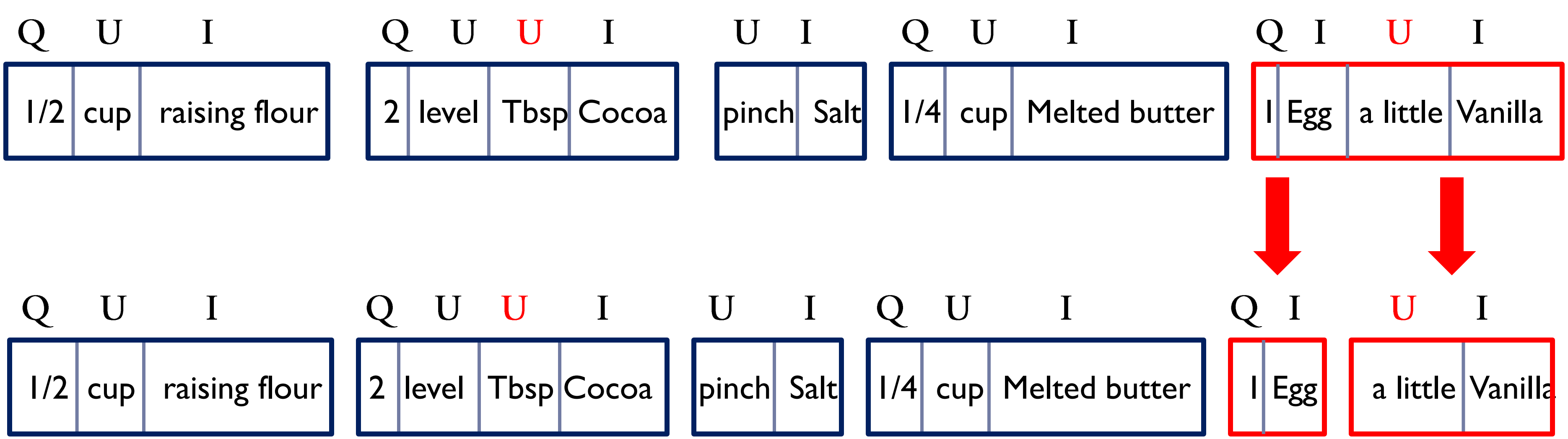
JUDIE – Structure-aware Labeling

- Labels textual values considering:
  - Uses a graphic model representing the likelihood of attribute transitions within the input text
  - Content-related features and structure-based features



JUDIE – Structure Refinement

- Applies again the SD algorithm
  - Considers the output of the structure-aware labeling
  - Fixes structural problems
    - Structure-aware labeling produces more precise results



Experiments

Domain	Dataset	Text Inputs	Attributes	Source	Attributes	Records
Cooking Recipes	*Recipes*	500	3	*FreeBase.com*	3	100
Product Offers	*Products*	10000	3	*Nhemu.com*	3	5000
Postal Addresses	*BigBook*	2000	5	*BigBook*	5	2000
Bibliography	*CORA*	500	3 to 7	*PersonalBib*	7	395
Classified Ads	*WebAds*	500	5 to 18	*Folha On-line*	18	125

- Metrics
  - Precision, Recall and F-Measure
    - T-Test for the statistical validation of the results
- Baselines
  - ONDUX and U-CRF

Evaluation – Record Level

Dataset	Phase 1	Phase 2	Gain (%)
Recipes	0.79	0.90	13.2
CORA	**0.69**	**0.83**	**19.3**
Web Ads	0.70	0.77	9.7

- Phase 1: acceptable. F ≈ 0.7
- Phase 2: positive impact. Gains > 9%
- In CORA, gains higher than 19%
  - Structural information led to significant improvements.

Comparison with baselines – Attribute Level

Attribute	JUDIE	ONDUX	U-CRF
Author	0.88	0.922	0.87
Title	0.70	0.79	0.69
Booktitle	0.86	0.89	0.56
Journal	0.84	0.90	0.55
Volume	0.90	0.96	0.43
Pages	0.86	0.84	0.50
Date	0.87	0.89	0.49
Average	**0.86**	**0.88**	**0.58**

CORA

Attribute	JUDIE	ONDUX	U-CRF
Bedroom	0.82	0.86	0.79
Living	0.89	0.90	0.72
Phone	0.87	0.92	0.75
Price	0.92	0.93	0.78
Kitchen	0.83	0.84	0.78
Bathroom	0.77	0.79	0.81
Others	0.73	0.79	0.71
Average	**0.84**	**0.85**	**0.76**

Web Ads

- Results very close to ONDUX and even better than U-CRF
- Recall: JUDIE faces a harder task.

iForm

A Probabilistic Approach for Automatically Filling Form-Based Web Interfaces

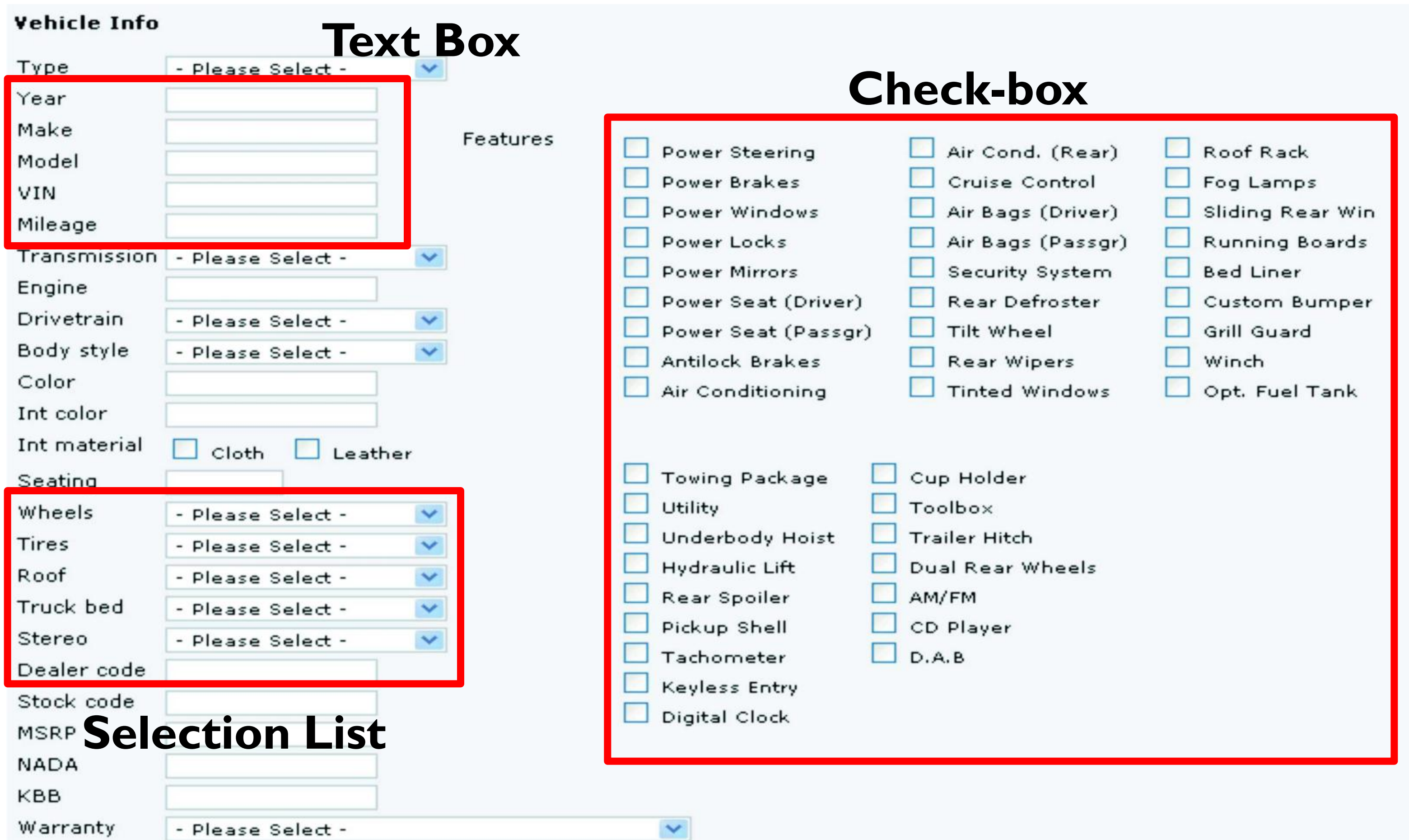
Toda et al. – WWW 2009, Toda et Al. – PVLDB 2010

The Form Filling Problem

- Goal:
  - To automatically fill out the fields of a given **form-based** interface with **values extracted** from a **data-rich free text document**.
    1. Extracting values from the input text;
    2. Filling out the fields of the target form using them.

Example

- Form-based interface



Example

- Data-rich free text document

2005 Honda new **Accord** Ex,Extra Clean, very **low Mileage**, Maintained By Dealer! Vechicle Located in Stockton, Ca. Ad Id# 28147

This is a brand new car with **automatic transmission**!

Car with Air Conditioning, clock, **Cruise Control**, Digital Info Center, Dual Zone Climate Control, Heated Seats, Leather Steering Wheel, Memory Seat Position, Power Driver's Seat, **Power Steering**, **Power Breaks**, Power Passenger Seat, **Power Windows**, **Cup Holder**, **Rear Air Conditioning**, **Sunroof**, Tilt Steering Wheel, Original Owner, **Alloy Wheels.**

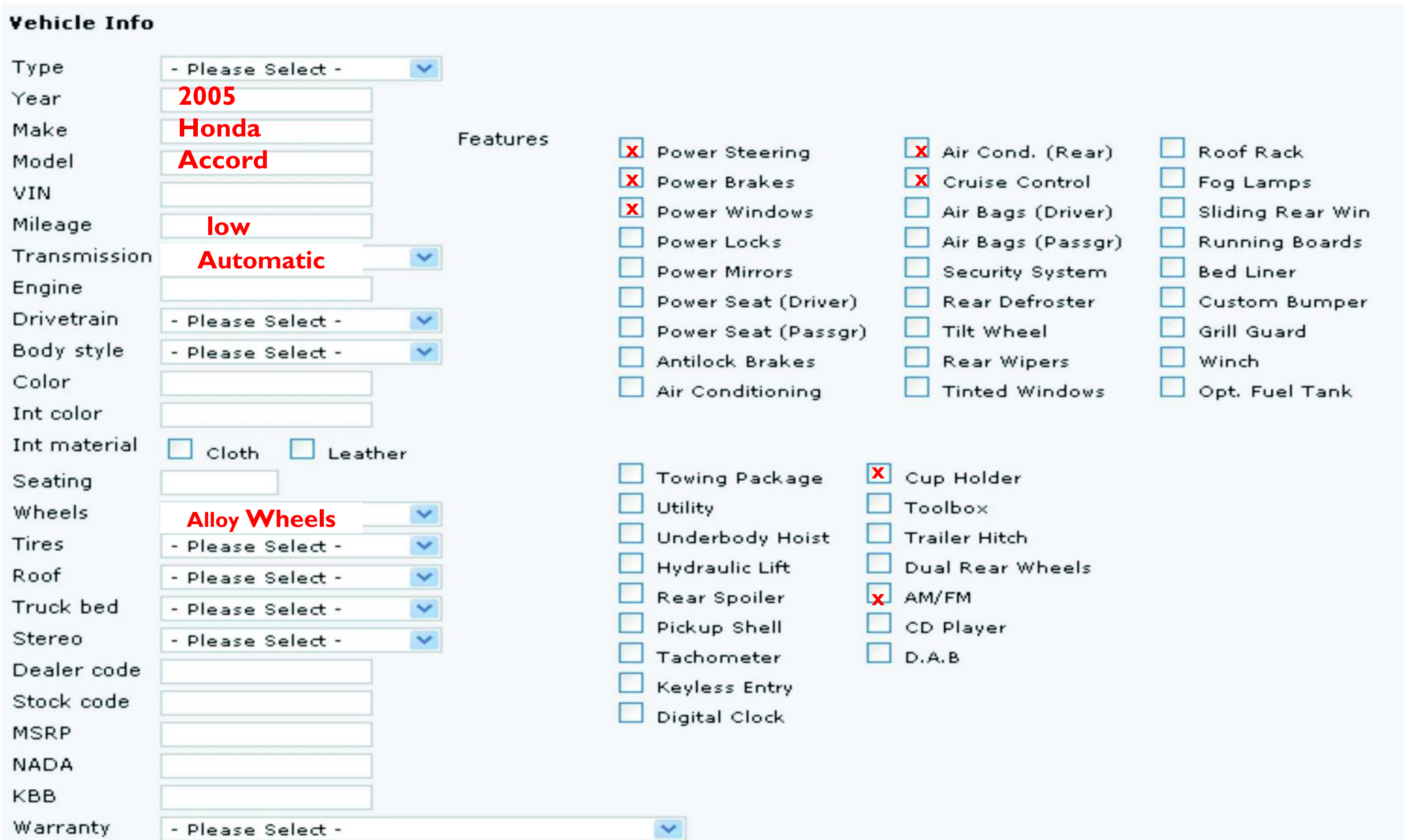
Am/fm, **Cd Changer**, Mp3, Satellite

Contact Us At XXX-XXXX-XXXX For More Information

Visit xxx xxx Motors

Example

- Form Filling



Vehicle Info

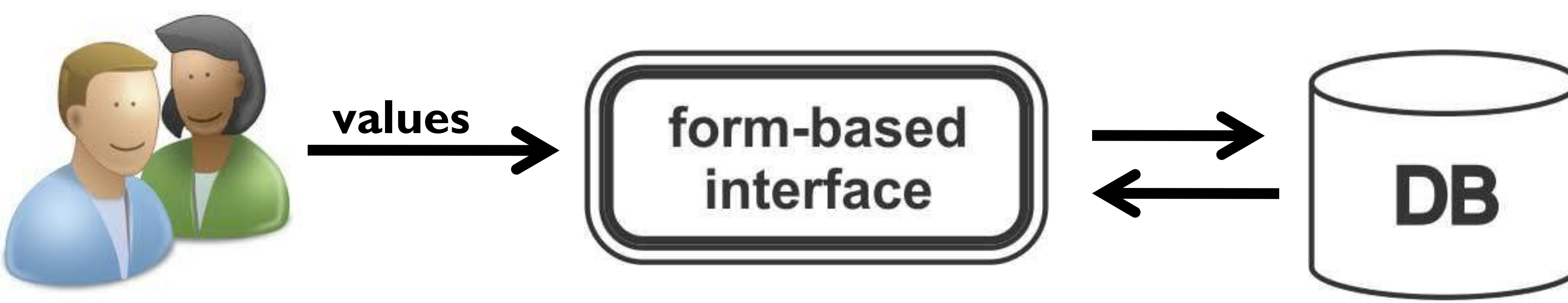
Type	- Please Select -
Year	2005
Make	Honda
Model	Accord
VIN	
Mileage	low
Transmission	Automatic
Engine	
Drivetrain	- Please Select -
Body style	- Please Select -
Color	
Int color	
Int material	[ ] Cloth [ ] Leather
Seating	
Wheels	Alloy Wheels
Tires	- Please Select -
Roof	- Please Select -
Truck bed	- Please Select -
Stereo	- Please Select -
Dealer code	
Stock code	
MSRP	
NADA	
KBB	
Warranty	- Please Select -

Features

- [x] Power Steering
- [x] Power Brakes
- [x] Power Windows
- [ ] Power Locks
- [ ] Power Mirrors
- [ ] Power Seat (Driver)
- [ ] Power Seat (Passgr)
- [ ] Antilock Brakes
- [ ] Air Conditioning
- [x] Air Cond. (Rear)
- [x] Cruise Control
- [ ] Air Bags (Driver)
- [ ] Air Bags (Passgr)
- [ ] Security System
- [ ] Rear Defroster
- [ ] Tilt Wheel
- [ ] Rear Wipers
- [ ] Tinted Windows
- [ ] Roof Rack
- [ ] Fog Lamps
- [ ] Sliding Rear Win
- [ ] Running Boards
- [ ] Bed Liner
- [ ] Custom Bumper
- [ ] Grill Guard
- [ ] Winch
- [ ] Opt. Fuel Tank
- [ ] Towing Package
- [ ] Utility
- [ ] Underbody Hoist
- [ ] Hydraulic Lift
- [ ] Rear Spoiler
- [ ] Pickup Shell
- [ ] Tachometer
- [ ] Keyless Entry
- [ ] Digital Clock
- [x] Cup Holder
- [ ] Toolbox
- [ ] Trailer Hitch
- [ ] Dual Rear Wheels
- [x] AM/FM
- [ ] CD Player
- [ ] D.A.B

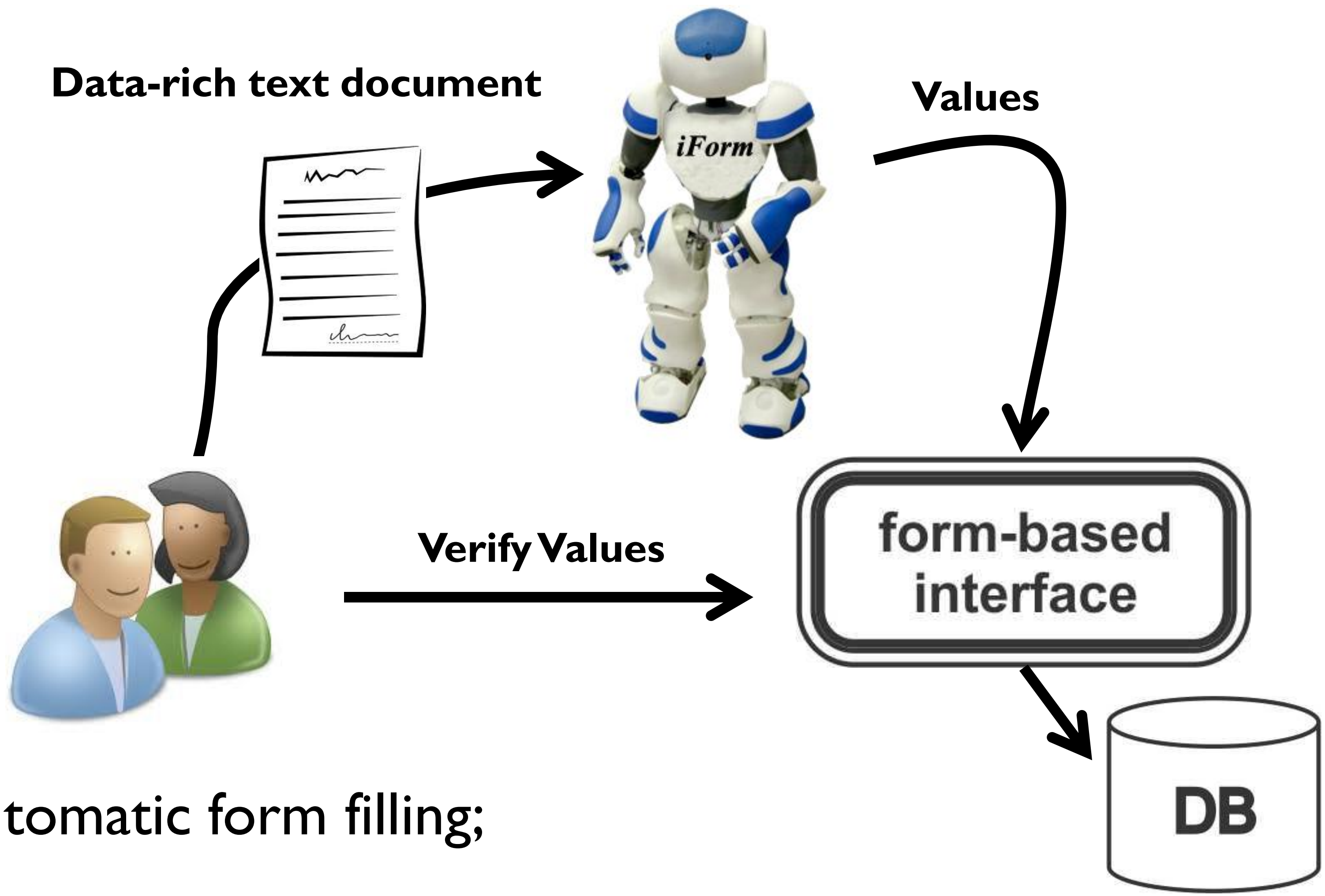
Common usage of Web Forms

- A user manually fills each form field
  - Text-box, selection list, check-box and radio button
- Tedious, error prone and repetitive process



iForm

- Information Extraction + Form Filling



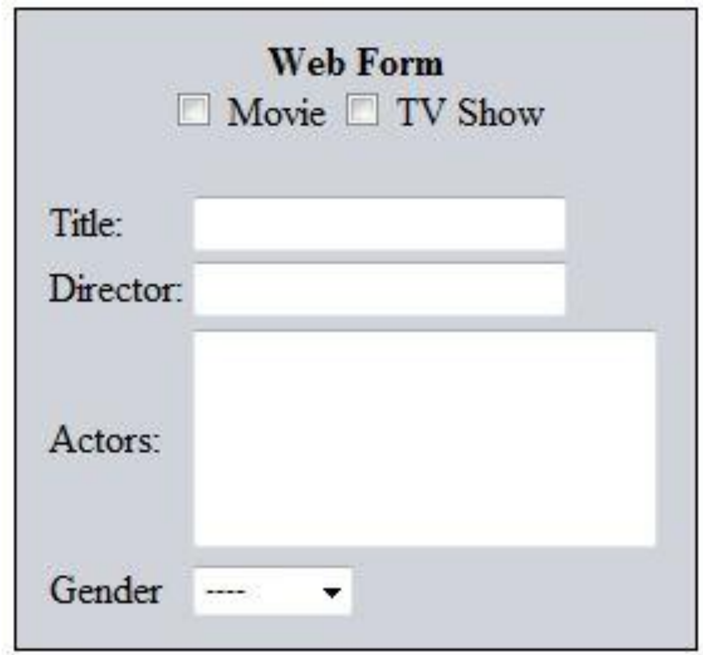
- Automatic form filling;

iForm - Scenario

Shutter Island is a 2010 American psychological thriller film directed by Martin Scorsese. The film is based on Dennis Lehane's 2003 novel of the same name . Starring Leonardo DiCaprio, Mark Ruffalo and Ben Kingsley.

Movie Review - Data-rich text

Web Form



iForm – Selecting plausible segments

- *Is this text segment a suitable value of a given field of the form?*

Shutter Island is a 2010 American psychological thriller film directed by Martin Scorsese. The film is based on Dennis Lehane's 2003 novel of the same name . Starring Leonardo DiCaprio, Mark Ruffalo and Ben Kingsley.

Shutter | **Shutter Island** | **Shutter Island is** | **Shutter Island is a**

...

Leonardo | **Leonardo DiCaprio** | **Kingsley.**

Redundant computation of several features can be avoided by using dynamic programming.

iForm - Features

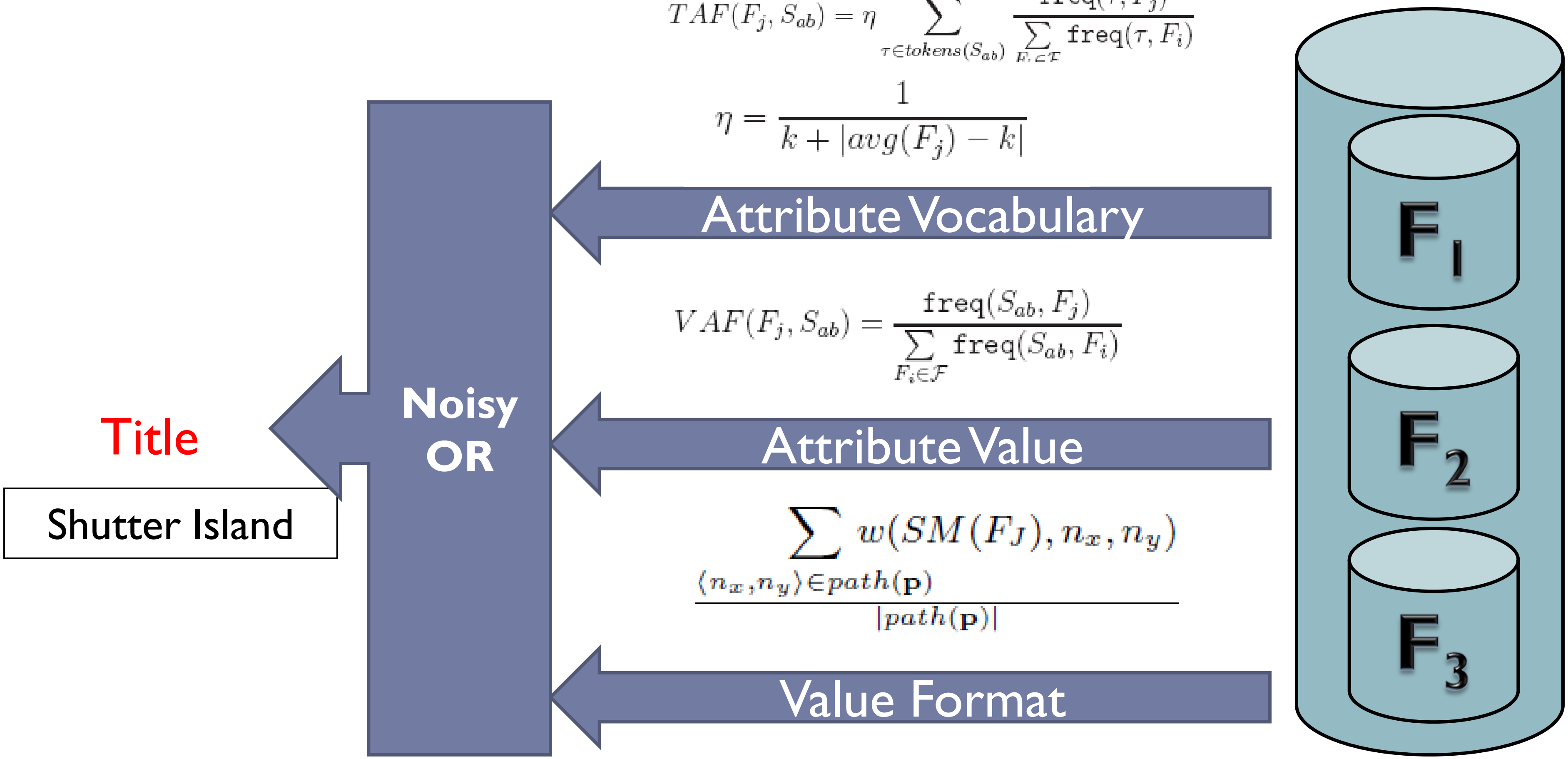
- Features Considered:

$$TAF(F_j, S_{ab}) = \eta \sum_{\tau \in tokens(S_{ab})} \frac{\mathtt{freq}(\tau, F_j)}{\sum_{F_i \in \mathcal{F}} \mathtt{freq}(\tau, F_i)}$$

$$\eta = \frac{1}{k + |avg(F_j) - k|}$$

$$VAF(F_j, S_{ab}) = \frac{\mathtt{freq}(S_{ab}, F_j)}{\sum_{F_i \in \mathcal{F}} \mathtt{freq}(S_{ab}, F_i)}$$

$$\frac{\sum_{\langle n_x, n_y \rangle \in path(\mathbf{p})} w(SM(F_J), n_x, n_y)}{|path(\mathbf{p})|}$$



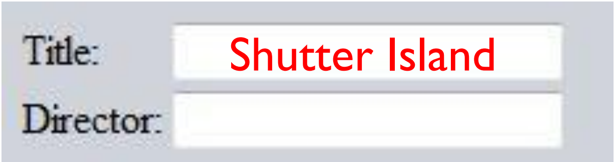
iForm – Mapping Segments to Fields

- Given the set of text segments such that theirs scores are above a threshold ε
  - iForm aims at finding a ***mapping*** between candidate values and form fields with a **maximum aggregate score**
    - Select non-overlaping segments.
- Accomplished by means of a two-phase procedure

iForm – Filling Form-based interfaces

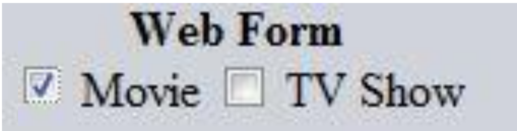
- Uses the final mapping to fill out the form fields
  - Text Boxes: Mapped text segments as a field values.

"Shutter Island" $\xrightarrow{\textit{title}}$



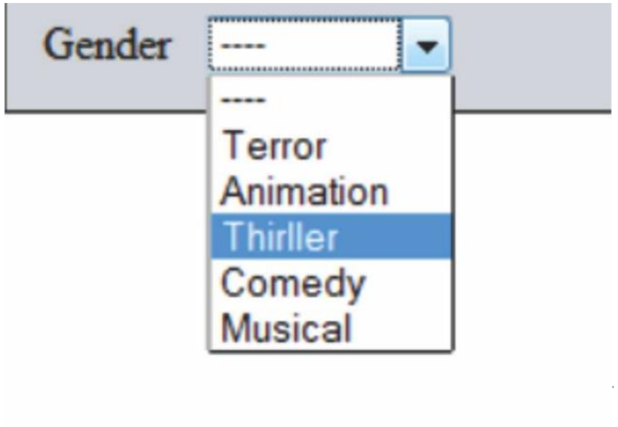
  - Check boxes: *Set true for mapped fields.*

"Movie" →



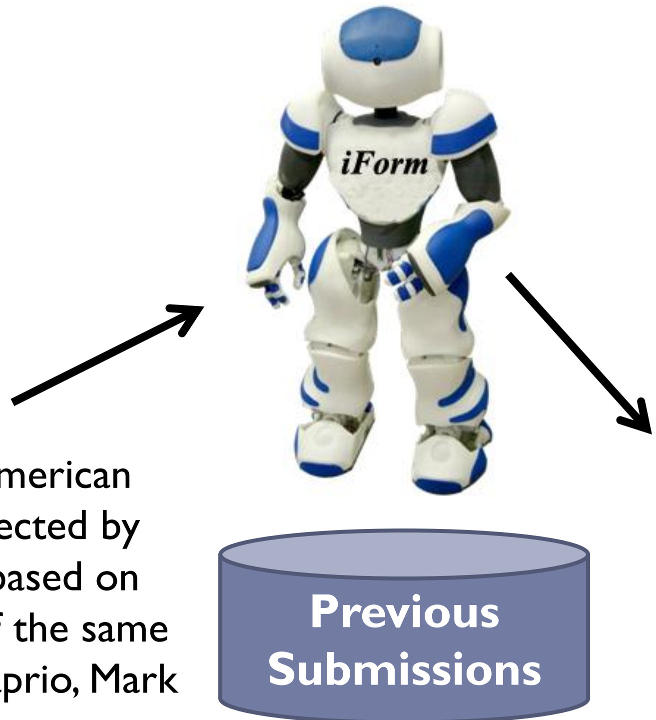
  - Selection List:

"psychological thriller" →



iForm - Overview

Shutter Island is a 2010 American psychological thriller film directed by Martin Scorsese. The film is based on Dennis Lehane's 2003 novel of the same name . Starring Leonardo DiCaprio, Mark Ruffalo and Ben Kingsley.



Web Form

Web Form

X Movie ☐ TV Show

Title: Shutter Island

Director: Martin Scorsese

Actors: Leonardo DiCaprio
Mark Ruffalo
Ben Kingslev

Gender Thriller

Experiments

Dataset	Test Data	Previous Data	# Fields	S - Test Data	S – Previous Data
Jobs	50	100	13	RISE	RISE
Movies	50	10000	4	IMDb	FreeBase / Wikipedia
Cars	50	10000	35	TodaOferta.com	TodaOferta.com
Cellphones	50	10000	37	TodaOferta.com	TodaOferta.com
Books 1	50	10000	5	Submarino.com	TodaOferta.com
Books 2	50	10000	4	Submarino.com	Ingenta
Books 3	50	10000	2	Submarino.com	Ourpress.com
Books 4	50	10000	3	Submarino.com	NetLibrary

- Baseline
  - iCRF - a method for interactive form filling based on CRF
  - The Jobs dataset was used for an experimental comparison between iForm and iCRF.

Evaluation – Multi-typed web forms

Movies

Type of Field	# Fields	P	R	F
Text Box	4	0.74	0.69	0.71
Submission-Level		**0.73**	**0.67**	**0.69**

iForm achieved high quality results in all datasets

Cellphones

Type of Field	# Fields	P	R	F
Text Box	2	0.89	0.69	0.78
Check Box	35	0.94	0.94	0.94
Average		0.94	0.93	0.93
Submission-Level		**0.96**	**0.94**	**0.95**

Filling quality above 0.90. In fact, more than 90% of each submission was correctly entered in the web form interface.

Evaluation – Comparison with iCRF

Jobs

Field	iForm	iCRF
Application	**0.82**	0.37
Area	0.18	0.23
City	**0.70**	0.65
Company	**0.41**	0.17
Country	0.77	0.87
Desired Degree	**0.57**	0.37
Language	**0.84**	0.69
Platform	**0.47**	0.38
Recruiter	**0.44**	0.22
Req. Degree	0.31	0.59
Salary	0.22	0.25
State	**0.85**	0.81
Title	**0.72**	0.49

iForm had superior F-measure levels in nine fields.

The lower quality obtained by iCRF is explained by the fact that segments to be extracted from typical free text inputs, such as jobs postings, may not appear in a regular context.

iForm was designed to conveniently exploit these field-related features from previous submissions

Conclusions

- This work proposes an unsupervised approach to the IETS problem.
  - Relies on information available on pre-existing data.
  - Exploit content-based features to directly learn from test data structure-based features.
  - Show that pre-existing datasets allow for the unsupervised learning of both content-based and structure-based features.
  - Eliminate the need of a user involved in any source specific training process.
- Information Extraction Methods:
  - ONDUX, JUDIE and iForm

Publications

- Thesis Core

1. Joint Unsupervised Structure Discovery and Information Extraction. **SIGMOD Conference** – 2011
2. Unsupervised Information Extraction with the ONDUX Tool. **Brazilian Symposium on Databases (SBBD)** – 2011
3. On Using Wikipedia to Build Knowledge Bases for Information Extraction by Text Segmentation. **Journal of Information and Data Management (JDIM)** – 2011
4. ONDUX: on-demand unsupervised learning for information extraction. **SIGMOD Conference**. - 2010
5. Unsupervised strategies for information extraction by text segmentation. **SIGMOD PhD Workshop on innovative Database Research (IDAR)** – 2010
6. A Probabilistic Approach for Automatically Filling Form-Based Web Interfaces. **Proceedings of the VLDB Endowment (PVLDB)** – 2010
7. Automatically filling form-based web interfaces with free text inputs. **International Conference on World Wide Web (WWW)** – 2009

Publications

- Related to the Information Extraction Problem

8. Building a research social network from individual perspective. **Joint Conference on Digital Libraries (JCDL)** – 2011
9. CiênciaBrasil – The Brazilian Portal of Science and Technology. **Integrated Seminar of Software and Hardware (Semish)**– 2011
10. A flexible approach for extracting metadata from bibliographic citations. **Journal of the American Society for Information Science and Technology (JASIST)** – 2009

Publications

- Other Publications

11. Lightweight methods for large-scale product categorization. **Journal of the American Society for Information Science and Technology (JASIST)** – 2011
12. Adaptive and Fexible blocking for record linkage tasks. **Journal of Information and Data Management (JDIM)** – 2010
13. Blocagem adptativa e flexível para o pareamento aproximado de registros. **Brazilian Symposium on Databases (SBBD)** – 2009

- Tutorials

14. Methods and techniques for information extraction by text segmentation. **Alberto Mendelzon International Workshop on Foundations of Data Management (AMW)** - 2012
15. Methods and techniques for information extraction by text segmentation. **Brazilian Symposium on Databases (SBBD)** - 2011

Future Work

- Generating transductive methods using domain knowledge
- Use our approach to extract information from HTML
- Query Extraction using our unsupervised approach
- Extraction Improvement Through User Feedback

Acknowledgments

UOL

CAPES

UNIVERSIDADE FEDERAL DO AMAZONAS
IN UNIVERSA SCIENTIA VERITAS
UFAM

FAPEAM
Fundação de Amparo à Pesquisa
do Estado do Amazonas

CNPq
Conselho Nacional de Desenvolvimento
Científico e Tecnológico

inweb

Unsupervised Information Extraction by Text Segmentation

Eli Cortez

Advisor: Altigran Soares da Silva

Universidade Federal do Amazonas