A Unsupervised Information Extraction with the ONDUX Tool

André Porto, Eli Cortez, Altigran S. da Silva,
Edleno S. de Moura

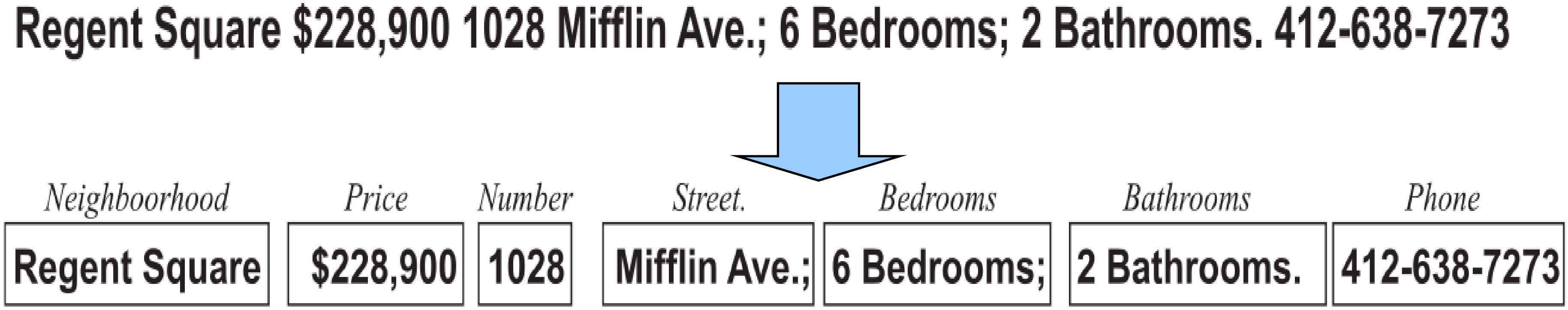
Univ. Fed. do Amazonas (UFAM) - Brazil

Presented by André Porto

SBBD 2011
Florianópolis, Brazil

The IETS Problem

- **Information Extraction by Text Segmentation**
- Goal:
  - To extract attribute values occurring in implicit semi-structured data records
- Current IETS methods predict labels for sequence of text segments corresponding to attribute values
  - HMM – Borkar et al. (SIGMOD01), CRF – Laferty et al. (ICML01), ONDUX – Cortez et. al (SIGMOD10)



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

Neighboorhood	*Price*	*Number*	*Street.*	*Bedrooms*	*Bathrooms*	*Phone*
Regent Square	$228,900	1028	Mifflin Ave.;	6 Bedrooms;	2 Bathrooms.	412-638-7273

Motivation

- Abundance of on-line sources of text documents
  - Postal Addresses, Classified Ads, Bibliography references...
- Necessity of storing these data in structured format
  - Relational DB, XML,
- Unsupervised Methods rely on attribute values from pre-existing data sources to perform extraction task
  - Knowledge Bases

Examples

Product Descriptions

Apple iPad 2 Wi-Fi + 3G 64 GB - Apple iOS 4 1 GHz - Black $589

LG - 32LE5300 - 32" LED-backlit LCD TV - 1080p (FullHD) - $400

Samsung - UN55D7000 - 55" Class (54.6" viewable) LED-backlit LCD ... $2,048

Mixter Max Accessory Plasma TV Rack Tilt Bracket 248-A05 $65

HP Deskjet 3050 All-in-One Color Ink-jet - Printer / copier / scanner $50

Bibliographic Citations

L. Barbosa and J. Freire. Using Latent-structure to Detect … In Proc. of the 13th WeDB, pages 1–6, 2010.

A. Doan et. al. Information Extraction Challenges in Managing .. SIGMOD Record, 37(4):14–20, 2008.

J. Pearl and G. Shafer. Probabilistic reasoning in intelligent systems: Morgan Kaufmann, 1988.

Classified Ads

$1106 / 2br - Luxury 2 BR, 1 BA apartment loaded with amenities - (Bothell)

$1945 / 2br - Beautiful HighPoint Community "Built Green" 2 BR 2.5 Bth Town Home! - (West Seattle)

$735 / 1br - Top floor 1 bedroom apt available just minutes from downtown!! - (Seattle,Burien,Highline)

$820 / 1br - Lovely 1 bedroom 1k sq ft! Nearly a 2 bdrm! - (Federal Way,Edgewood,Milton,Tacoma)

$895 / 2br - ****Lovely 2-Bedroom/2-Bathroom Condo with a View! FREE RENT!!!**** - (Monroe)

Related Work – IETS Approaches/Methods

- Probabilistic – Supervised
  - Hidden Markov Models (HMM)
    - Borkar et al.@SIGMOD'01;McCallum et al.@AAAI'00
  - Conditional Random Fields (CRF)
    - Lafferty et al.@ICML'01;McCallum et al.@IPM'06)

- Require labeled training instances

<Neighboorhood>Regent Square </Neighboorhood> <Price> $228,900 </Price>
<No>1028 </No><Street>Mifflin Ave, </Street> <Bed>6 Bedrooms </Bed>
<Bath> 2 Bathrooms </Bath> <Phone>412-638-7273 </Phone>

Related Work - IETS Approaches / Methods

- Probabilistic – Unsupervised
  - Rely on previously built datasets
  - Unsup. HMM (Agichtein et al.@SIGKDD '04)
    - Rely on records in references tables
    - Still requires a few training instances
  - Unsup. CRF (Zhao et al. @SIAM ICDM'08)
    - Also reference tables
    - Batches of fixed-order records as input
  - ONDUX (Cortez et al. @SIGMOD'10)
    - Knowledge-base: sets of typical values per attribute – no records

Tool Overview

- Implements an information extraction method called ONDUX (On-Demand Unsupervised Information Extraction) [E. Cortez et. al. 2010] for extracting information from unstructured data records.

- Should allow:
  - Easy to Use
  - Use by final users
  - Conducting experiments with ONDUX
  - Support the teaching of extraction techniques

Tool Overview

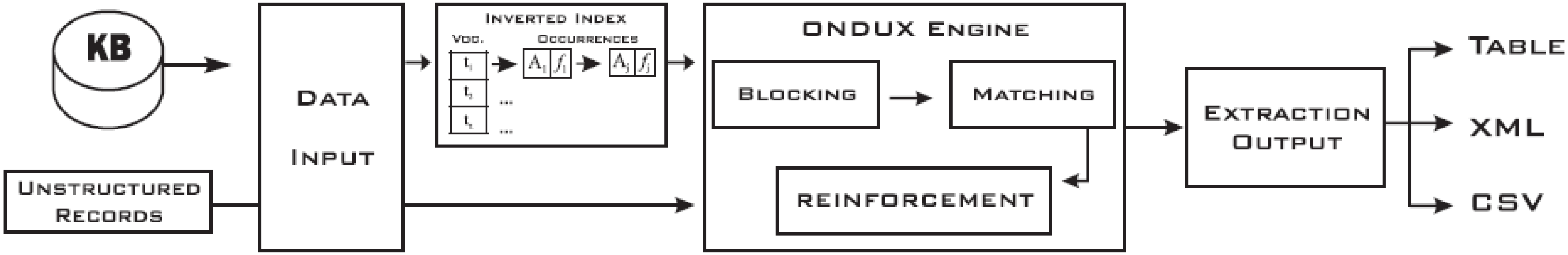
- An effective implementation of an unsupervised probabilistic approach for information extraction by text segmentation;

- A friendly graphical user interface that allows non expert users to easily carry out information extraction tasks;

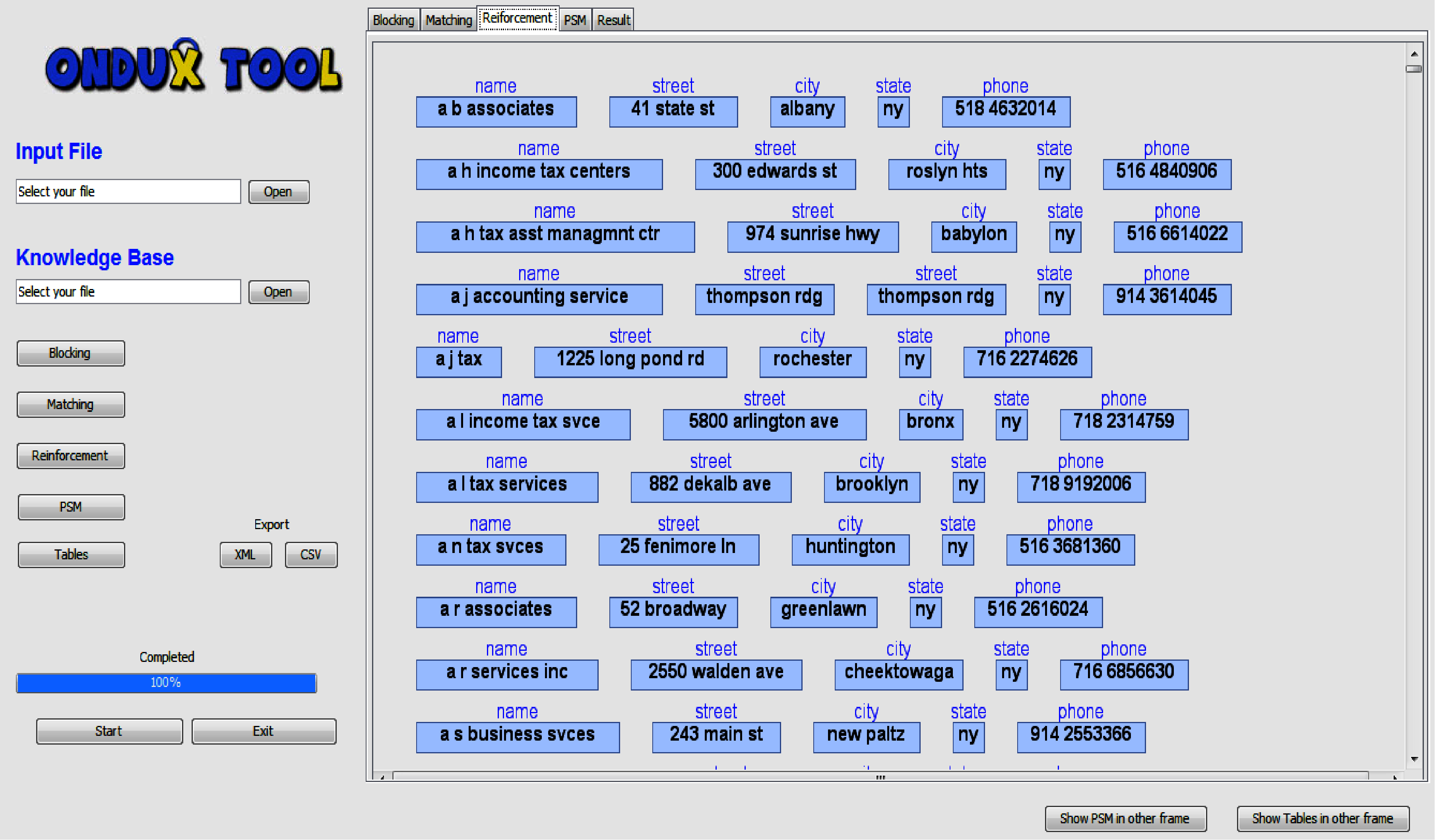
- Visualization facilities that allow users to follow a understand all the steps involved in the extraction process.

Tool Overview

- Architecture
  - Data Input: handles the KB and the input file
  - ONDUX Engine: implements the 3 main steps of the ONDUX method:| Blocking, Matching and Reinforcement
  - Extraction Output: presents the extraction results to the user and exports it to several formats



Ondux Tool



Input File

Select your file

Knowledge Base

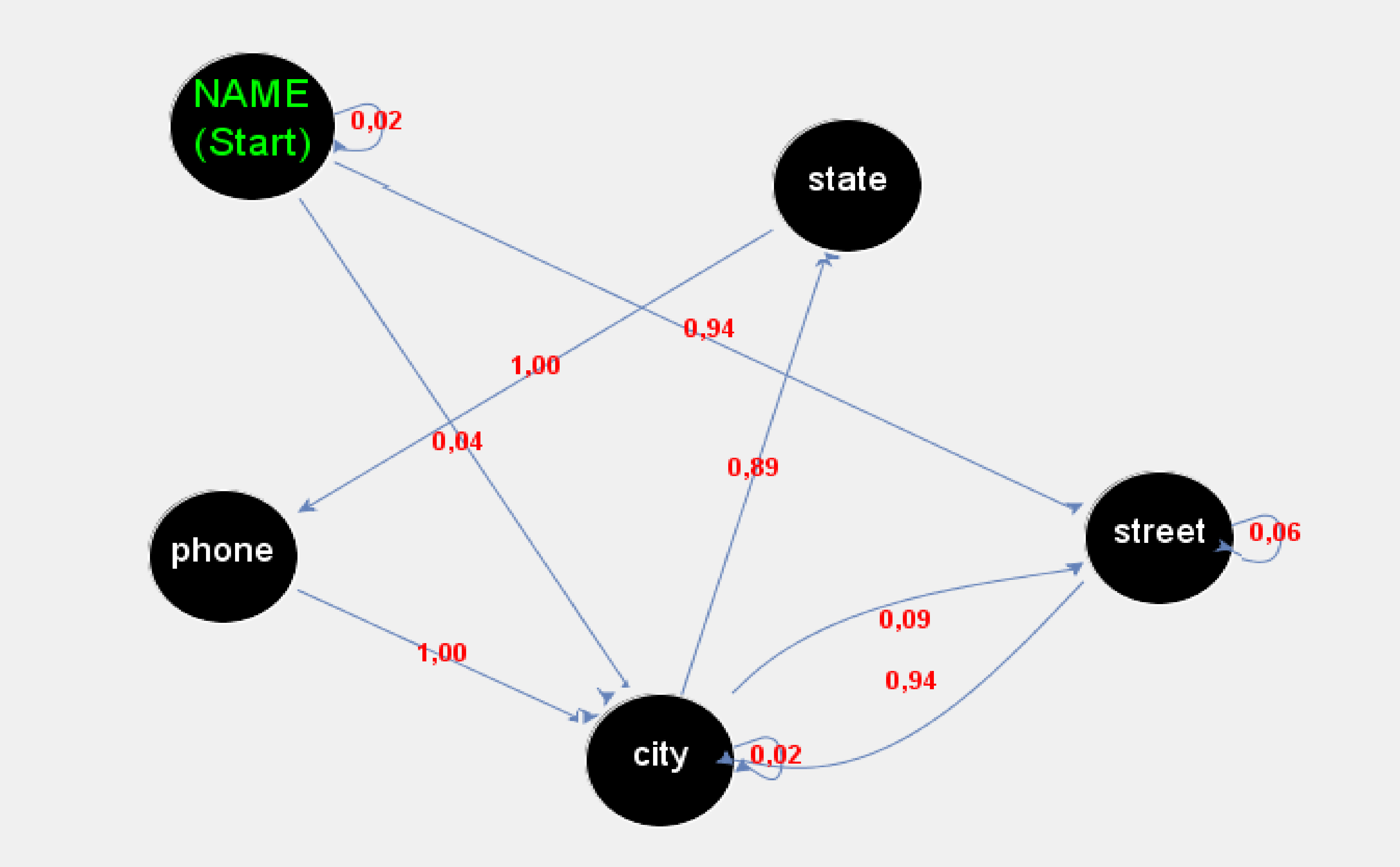
Select your file

Completed

100%

name	street	city	state	phone
a b associates	41 state st	albany	ny	518 4632014
name	street	city	state	phone
a h income tax centers	300 edwards st	roslyn hts	ny	516 4840906
name	street	city	state	phone
a h tax asst managmnt ctr	974 sunrise hwy	babylon	ny	516 6614022
name	street	street	state	phone
a j accounting service	thompson rdg	thompson rdg	ny	914 3614045
name	street	city	state	phone
a j tax	1225 long pond rd	rochester	ny	716 2274626
name	street	city	state	phone
a l income tax svce	5800 arlington ave	bronx	ny	718 2314759
name	street	city	state	phone
a l tax services	882 dekalb ave	brooklyn	ny	718 9192006
name	street	city	state	phone
a n tax svces	25 fenimore ln	huntington	ny	516 3681360
name	street	city	state	phone
a r associates	52 broadway	greenlawn	ny	516 2616024
name	street	city	state	phone
a r services inc	2550 walden ave	cheektowaga	ny	716 6856630
name	street	city	state	phone
a s business svces	243 main st	new paltz	ny	914 2553366

PSM Graph


NAME
(Start)
0,02
state
0,94
1,00
0,04
0,89
phone
street
0,06
0,09
1,00
0,94
city
0,02

Tool Functions

- Blocking
  - Segments the input string into **blocks**
- Matching
  - Matches blocks against known attribute values in the KB
  - From the best match, a label is derived
- Reinforcement
  - Relabel mismatched and umatched blocks
- PSM
  - Show PSM Graph
- Export Data
  - XML and CSV

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 Adresses	Big Book	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

- Datasets: Used for test
- Source: Used for building each KB

Conclusions

- In this demo we presented a tool that implements ONDUX

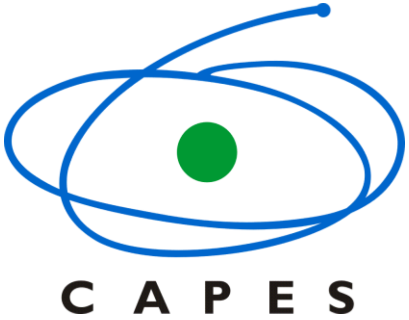
- The ONDUX Tool allows non expert users to easily perform information extraction tasks and export the extraction result in different formats.

- Can be used as a teaching tool

Acknowledgments

UNIVERSIDADE FEDERAL DO AMAZONAS
IN UNIVERSA SCIENTIA VERITAS
UFAM

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico
60 ANOS



Thank you!

A Unsupervised Information Extraction with the ONDUX Tool

André Porto, Eli Cortez,
Altigran S. da Silva, Edleno S. de Moura

Univ. Fed. do Amazonas (UFAM)
Brazil

Presented by André Porto

SBBD
Florianópolis, Brazil - 2011