Attribute Discovery in Massive Text Corpora:

Open IE Systems, Google's Projects, and Our Methodology

Meng Jiang
2130SC
www.meng-jiang.com

What is Attribute Discovery?

- Given the sentence "*President Blaise Compaore's government of Burkina Faso was founded…*",
- Find ⟨entity, attribute name, attribute value⟩
  - ⟨Burkina Faso, president, Blaise Compaore⟩
  - called "*attribute tuple extraction*"
- Find ⟨entity type, attribute name⟩
  - ⟨$Country, president⟩
  - called "*attribute name extraction*"

Attribute Tuple Extraction

- The state-of-the-art open information extraction systems, StanfordCoreNLP-OpenIE [ACL'15] and AI2-Ollie [EMNLP'12], learn syntactic and lexical patterns of expressing relationships
- But they *ignore the typing information*
  - Generate incorrect or imprecise extractions

President Blaise Compaore's government of Burkina Faso was founded…

⇩

⟨President Blaise Compaore, *have*, government of Burkina Faso⟩

Attribute Name Extraction

- Google's Biperpedia [VLDB'14, WWW'16] mines users' fact-seeking queries with "E-A" patterns.

"president of united states" ⇨ "A of E", "E 's A", "E A", "A, E"

- Google's ReNoun [EMNLP'15] learns annotated corpus for "S-A-O" patterns.

"Barack Obama, President of U.S., " ⇨ "O, A of S,", "S A O"

- But (1) query logs and annotations are often *unavailable or expensive*; (2) "E-A" patterns generate noisy attribute names *ignoring attribute value*: query log word distributions are highly constrained compared with ordinary written language.

"…Sunday night, Burkina Faso…" and "A, E" pattern ⇨ ⟨Burkina Faso, Sunday night⟩

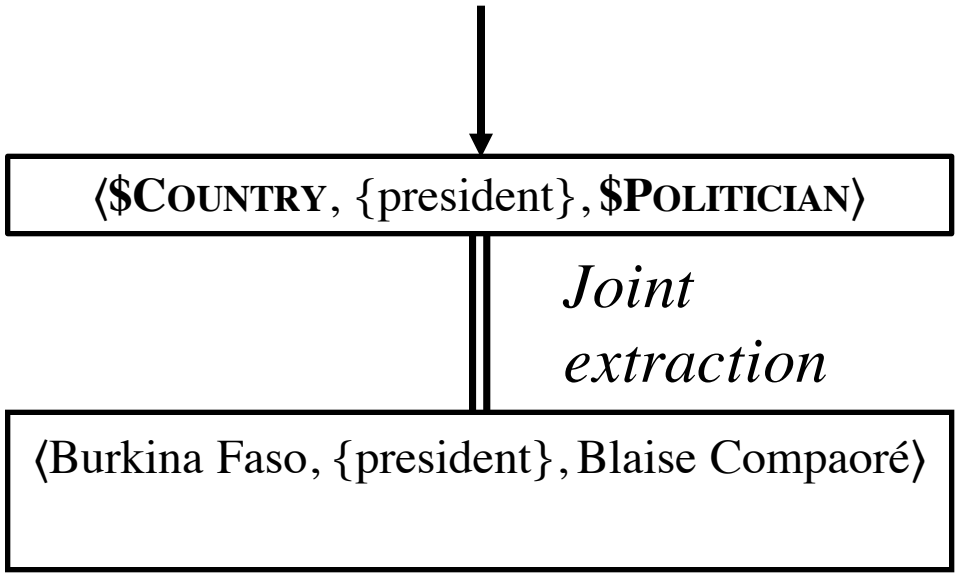
Automatic Attribute Discovery

- Problem definition: Given massive text corpora, can we discover attributes relying on NO linguistic assumptions NOR query logs NOR annotations?
- Our idea: **Joint** tuple extraction at type-level and entity-level
  - Results of ⟨entity type, attribute name⟩ *improve the precision* of ⟨entity, attribute name, attribute value⟩ extractions: ⟨$Country, president⟩ is more frequent than ⟨Burkina Faso, president⟩
  - Results of ⟨entity, attribute name, attribute value⟩ *improve the precision* of ⟨entity type, attribute name⟩ extractions: If we know the entities are {U.S., Burkina Faso…} in ⟨$Country, president⟩ and the entities are {Chicago, U.S., Illinois…} in ⟨$Location, population⟩

"Meta Pattern"

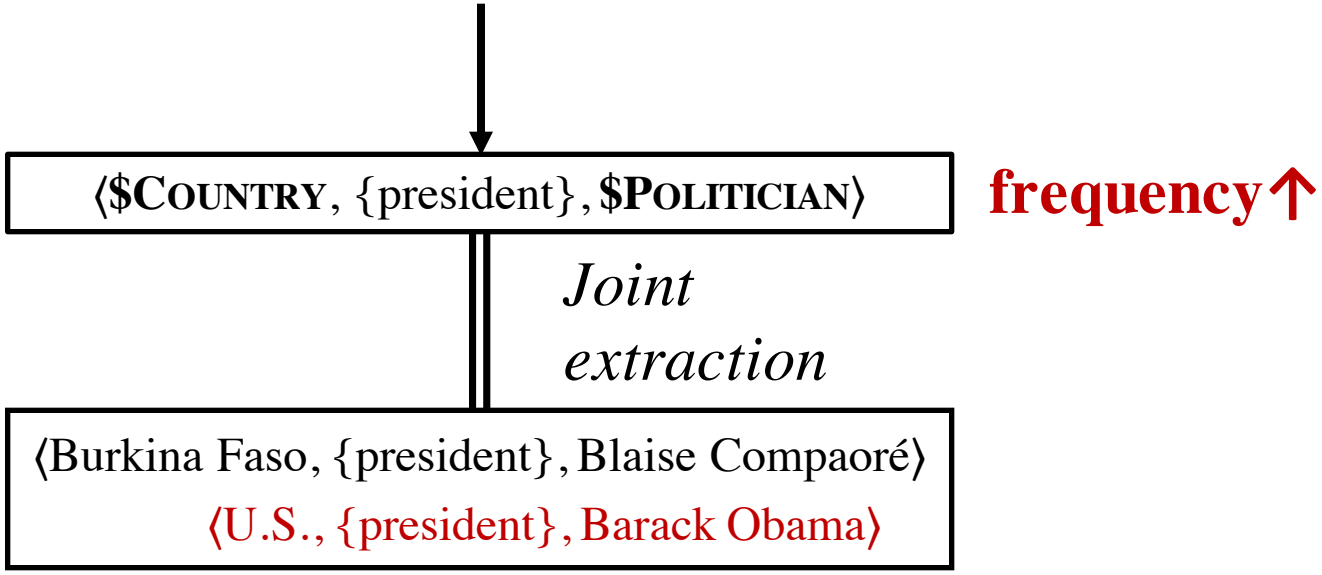
(#1) "President Blaise Compaoré's government of Burkina Faso was founded …"

⟨entity type, attribute name, **attribute value type**⟩



"Meta Pattern"

(#1) "President Blaise Compaoré's government of Burkina Faso was founded …"
(#2) "President Barack Obama's government of U.S. claimed that…"



"Meta Pattern"

(#1) "President Blaise Compaoré's government of Burkina Faso was founded …"
(#2) "President Barack Obama's government of U.S. claimed that…"

Meta pattern segmentation

Meta pattern:

frequency↑

⌈**president $PERSON.POLITICIAN 's government of $LOCATION.COUNTRY**⌋ was founded…

⟨**$COUNTRY**, {president}, **$POLITICIAN**⟩

Joint extraction

⟨Burkina Faso, {president}, Blaise Compaoré⟩
⟨U.S., {president}, Barack Obama⟩

"Meta Pattern"

(#1) "President Blaise Compaoré's government of Burkina Faso was founded …"
(#2) "President Barack Obama's government of U.S. claimed that…"
(#3) "U.S. President Barack Obama visited …"

Meta pattern segmentation

Meta patterns:

⌈**president $PERSON.POLITICIAN 's government of $LOCATION.COUNTRY**⌋ was founded…
⌈**$LOCATION.COUNTRY president $PERSON.POLITICIAN**⌋ …

frequency↑↑

⟨**$COUNTRY**, {president}, **$POLITICIAN**⟩

Joint extraction

⟨Burkina Faso, {president}, Blaise Compaoré⟩
⟨U.S., {president}, Barack Obama⟩

Group synonymous meta patterns

"Meta Pattern"

(#1) "President Blaise Compaoré's government of Burkina Faso was founded …"
(#2) "President Barack Obama's government of U.S. claimed that…"
(#3) "U.S. President Barack Obama visited …"

Meta pattern segmentation

Meta patterns:

⌈**president $PERSON.POLITICIAN 's government of $LOCATION.COUNTRY**⌋ was founded…
⌈**$LOCATION.COUNTRY president $PERSON.POLITICIAN**⌋ …

Adjust types for appropriate granularity
$LOCATION or $COUNTRY?

⟨**$COUNTRY**, {president}, **$POLITICIAN**⟩

Joint extraction

⟨**Burkina Faso**, {president}, Blaise Compaoré⟩
⟨**U.S.**, {president}, Barack Obama⟩

Group synonymous meta patterns

Meta Pattern Mining

- **Meta Pattern:** a sequence of *class symbols*, *words*, *phrases* and *punctuation marks* that appear contiguously in the text, and serves as a whole semantic unit.

Text:

U.S. President Barack Obama visited …
President Barack Obama of U.S. told …

Barack Obama's age is 55.
Yesterday, Barack Obama, 55, …
Michael Brown, an 18-year-old man…

Meta patterns:

⌈**$COUNTRY president $POLITICIAN**⌋ visited …
⌈**president $POLITICIAN of $COUNTRY**⌋ told …

⌈**$PERSON's age is $DIGIT**⌋ .
Yesterday, ⌈**$PERSON, $DIGIT,**⌋ …
⌈**$PERSON, an $DIGIT -year-old**⌋ man…

Type-level ⟨entity type C, attribute name A, value type T⟩ tuples:

⟨$COUNTRY, {president}, [$POLITICIAN]⟩

⟨$PERSON, {age, -year-old}, [$DIGIT]⟩

Entity-level ⟨entity e, attribute name a, attribute value v⟩ tuples (called "attribute tuples"):

⟨U.S., president, [Barack Obama]⟩

⟨Barack Obama, age, [55]⟩
⟨Michael Brown, age, [18]⟩

MetaPAD: Preprocessing

U.S. President Barack Obama and Prime Minister Justin Trudeau of Canada met in North America Leaders' Summit on June 29, 2016.

Quality phrase mining (SegPhrase, SIGMOD'15)

[u.s.] president [barack_obama] and [prime_minister] [justin_trudeau] of [canada] met in [north america leaders' summit] on [june] [29], [2016].

Entity recognition and typing with distant supervision (ClusType, KDD'15)

$LOCATION president $PERSON and prime_minister $PERSON of $LOCATION met in $EVENT on $MONTH $DAY , $YEAR.

Fine-grained typing (PLE, KDD'16)

$COUNTRY president $POLITICIAN and prime_minister $POLITICIAN of $COUNTRY met in $EVENT.SUMMIT on $MONTH $DAY , $YEAR.

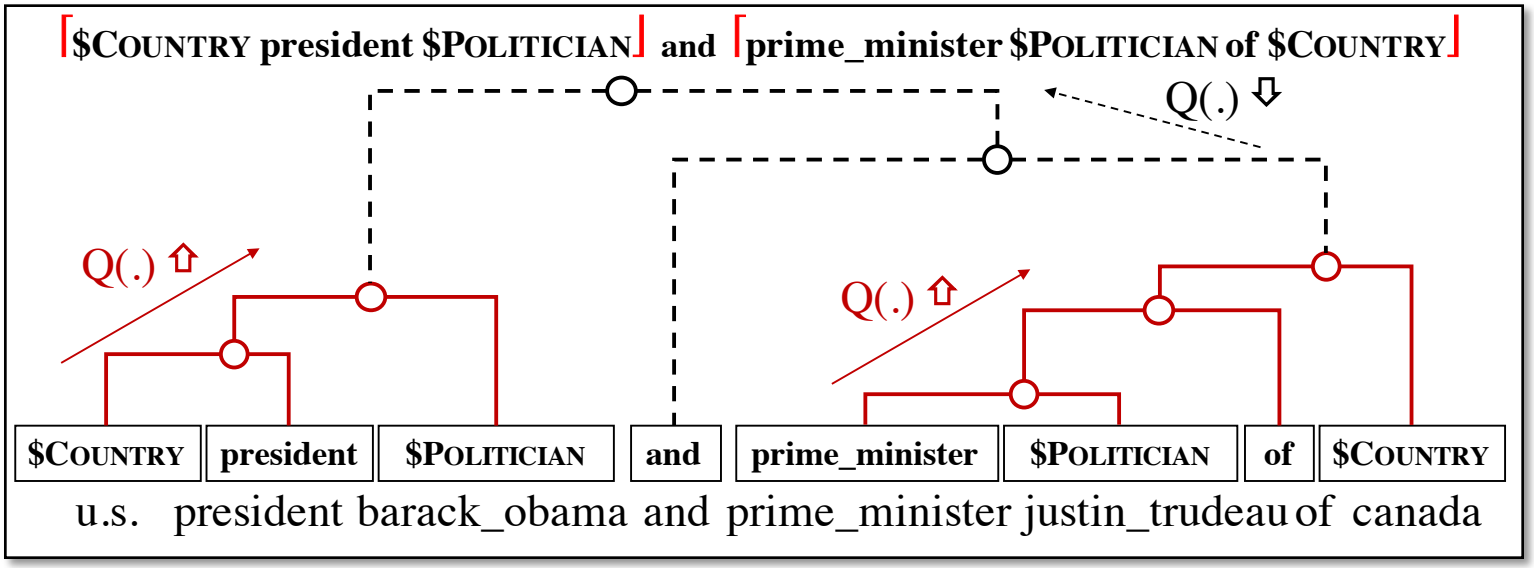
[$COUNTRY president $POLITICIAN] and [prime_minister $POLITICIAN of $COUNTRY] met in [$EVENT.SUMMIT on $MONTH $DAY , $YEAR].

MetaPAD: I. Meta-Pattern Segmentation with Quality Assessment

- **(Raw) Frequency**
  - *"prime_minister $Politician" vs "young $ Politician"*
- **Concordance**
  - *"prime_minister $Politician" vs "prime_minister"/"$Politician"*
- **Completeness**
  - *"$Country prime_minister $Politician" vs "$Country prime_minister"*
- **Informativeness**
  - *"$Person 's brother , $Person ," vs "$Person and $Person"*
- **Coverage**
  - *"$Politician 's signature healthcare law": only "Barack Obama"*
- **Classifier: Random Forest**

Rectified Frequency with Segmentation

- Segmentation

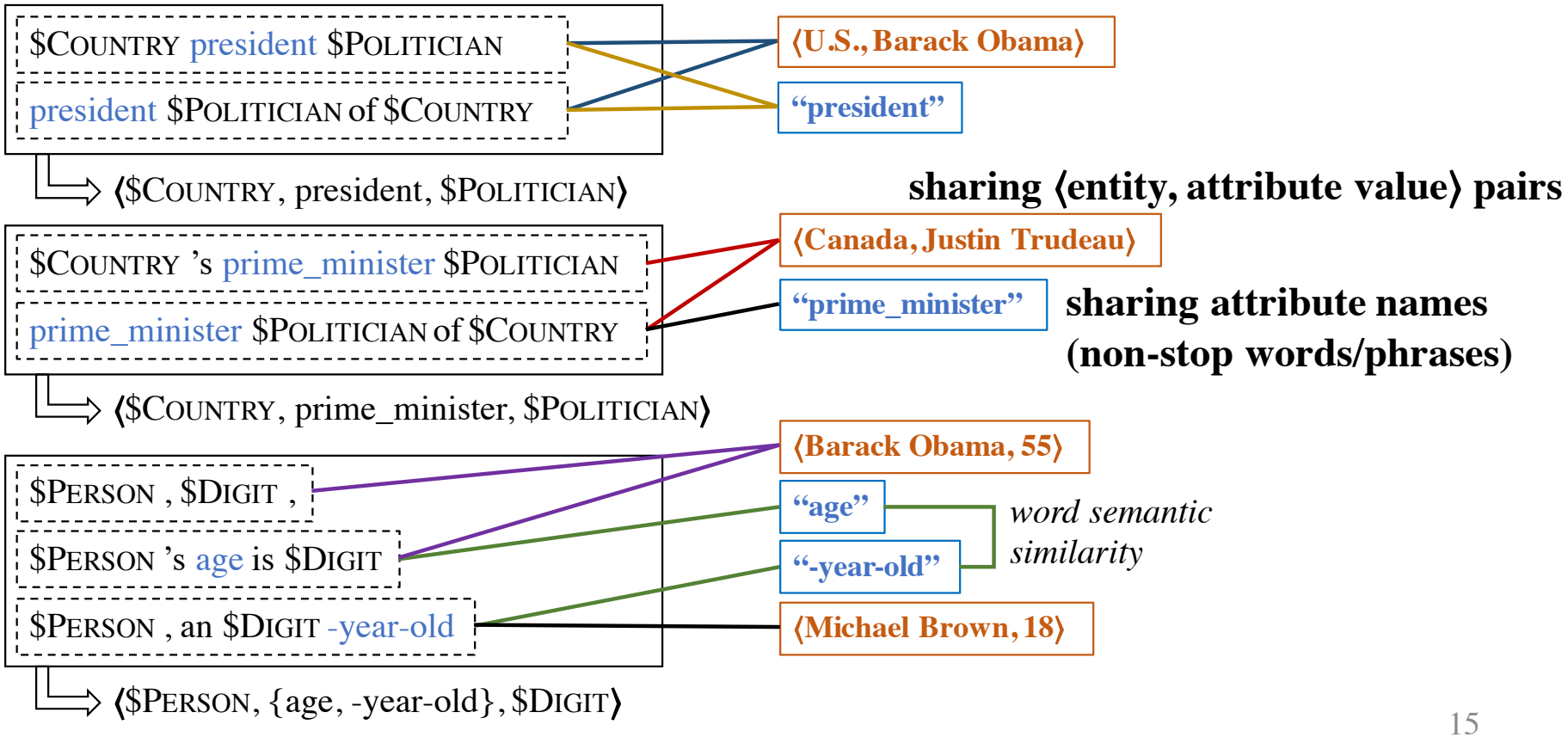


- Rectified frequency

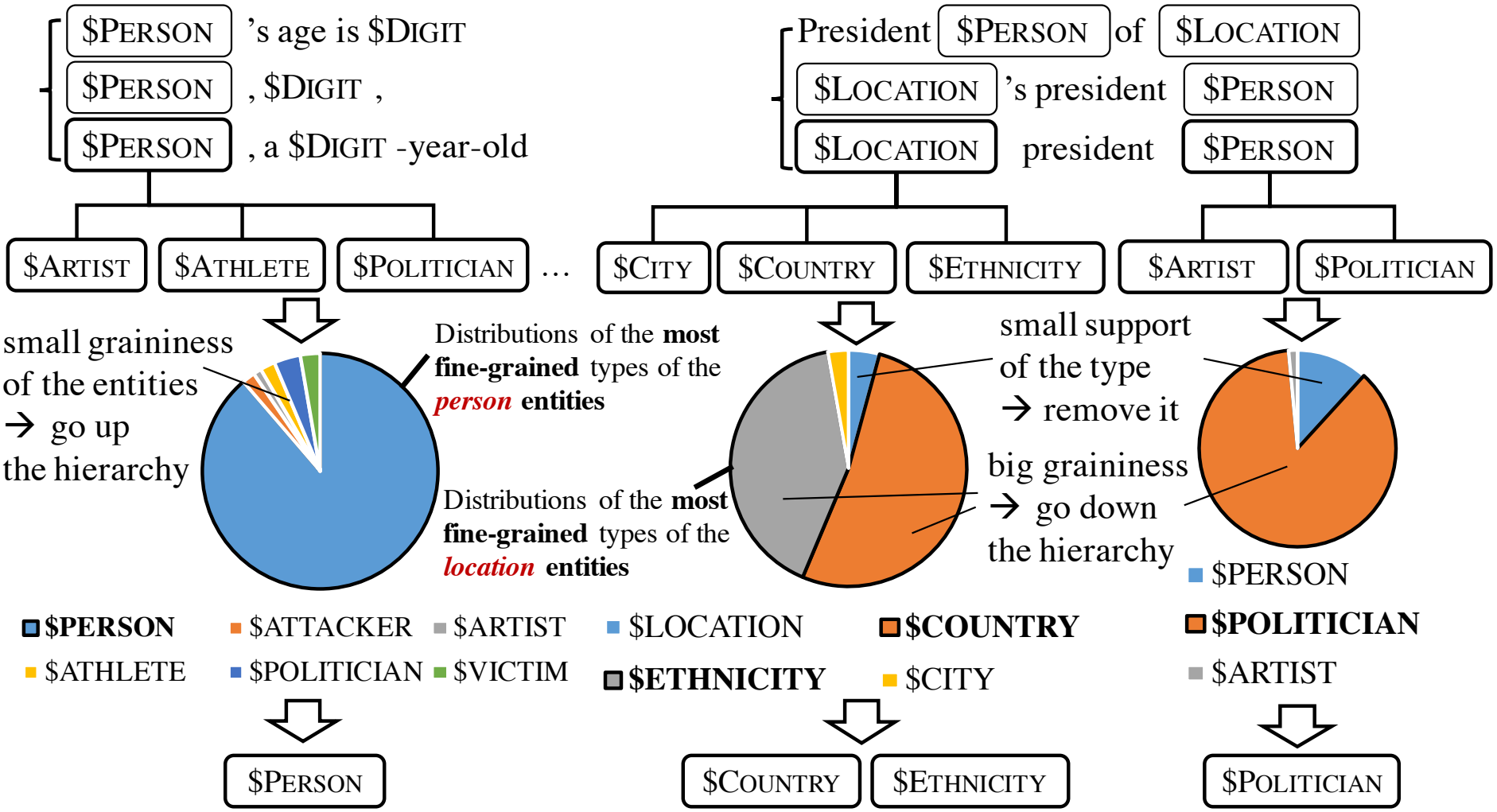
Contiguous pattern	Before segmentation Count	Quality	Rectified with segmentation features Count	Quality	Problem fixed by feedback
$COUNTRY president $POLITICIAN	2,912	0.93	2,785	0.97	N/A
prime_minister $POLITICIAN of $COUNTRY	1,285	0.84	1,223	0.92	slight underestimate
$POLITICIAN and prime_minister $POLITICIAN	532	0.70	94	0.23	overestimate

MetaPAD: II. Synonymous Meta-Pattern Detection

- *Definition*. A pair of meta patterns are synonymous if they can be represented as the same type-level ⟨entity type, attribute name, attribute value type⟩ tuple.

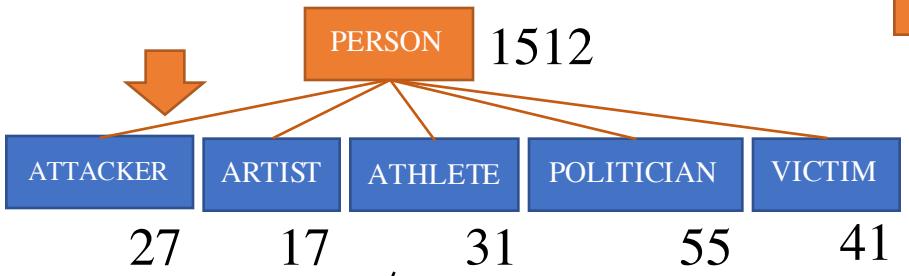


MetaPAD: III. Type Adjustment in Meta Patterns for Appropriate Granularity



Top-Down Scheme of Type Adjustment

_____, a $DIGIT -year-old

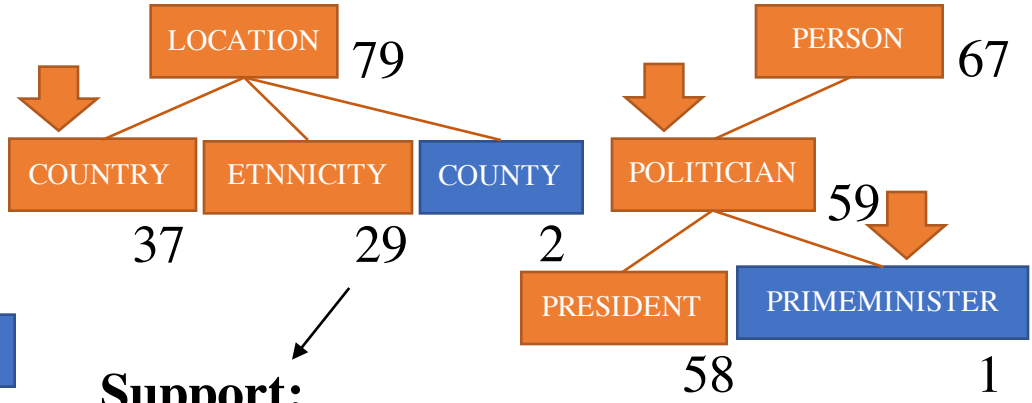


Graininess:

$\alpha=(27+17+\ldots+41)/1512$

Small, stop going down

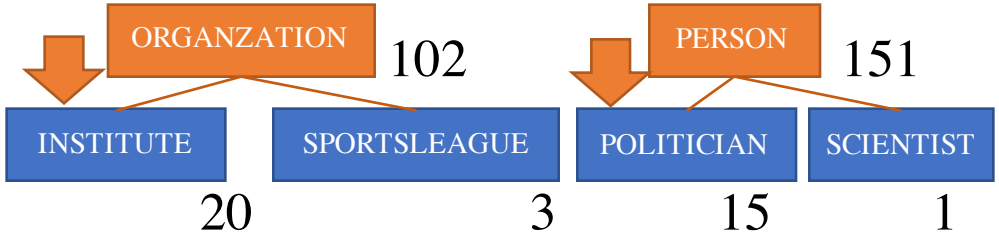
_____ President _____



Support:

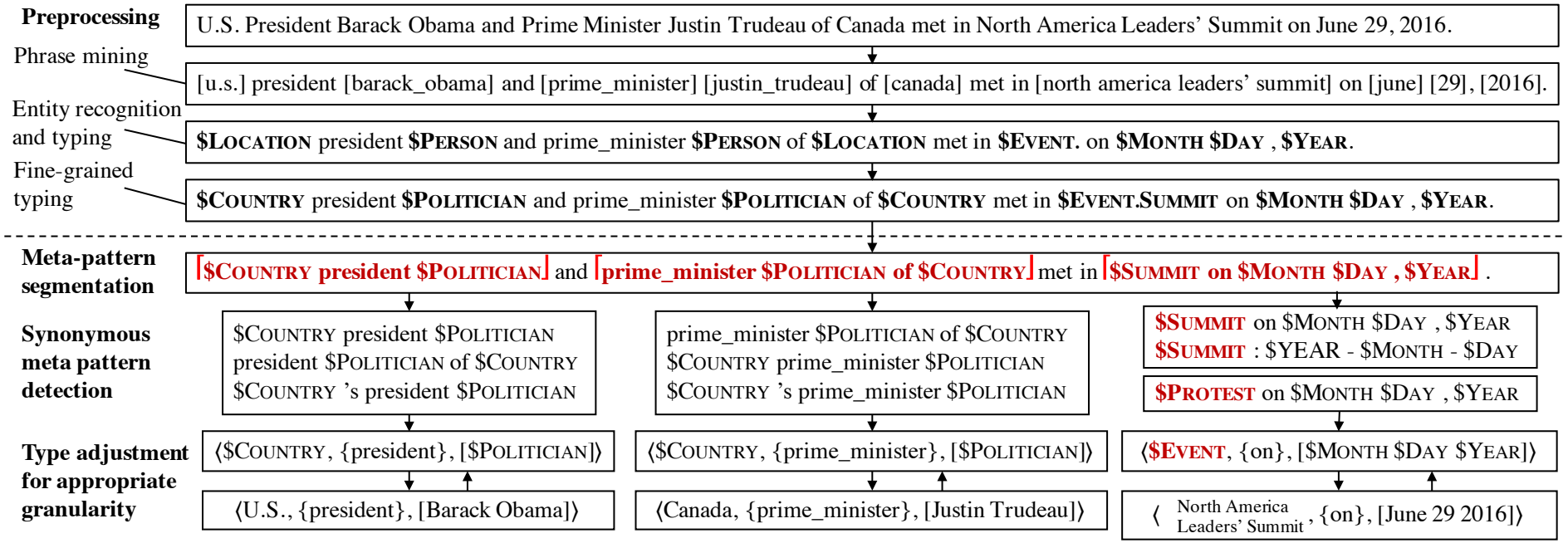
$\beta=29/\max(37,29,2)$

Big, keep $ETHNICITY



Similar for Bottom-Up…

Review: MetaPAD Framework



Experimental Results

- Datasets

Dataset	File Size	#Document	#Entity	#Entity Mention
WPB	105MB	26,143	96,402	1,200,536
CNA	214MB	144,803	165,979	4,299,112
APR	199MB	62,146	284,061	6,732,399
TWT	1.05GB	13,200,821	618,459	21,412,381
CVD	424MB	463,040	751,158	27,269,242

- Attribute name extraction

	WPB				CNA				APR				TWT			
	F1	Precision	Recall	AUC	F1	Precision	Recall	AUC	F1	Precision	Recall	AUC	F1	Precision	Recall	AUC
BIPERPEDIA	0.3000	0.2469	0.3821	0.1926	0.2519	0.2228	0.2899	0.1319	0.3236	0.2968	0.3557	0.2195	0.2421	0.2066	0.2924	0.1373
METAPAD-T	0.2523	0.3471	0.1982	0.1976	0.2316	0.1705	0.3611	0.1224	0.3426	0.2892	0.4202	0.2104	0.2376	0.1984	0.2961	0.1165
METAPAD-TA	0.3659	0.4022	0.3356	0.2763	0.3088	0.2417	0.4272	0.2099	0.4736	0.4130	0.5549	0.4020	0.3148	0.2562	0.4081	0.2263
METAPAD-TAS	0.4272	0.3714	**0.5026**	0.3200	0.3684	0.2951	0.4903	0.2411	0.4886	0.4333	0.5602	0.4051	**0.3244**	0.2475	0.4709	**0.2324**
METAPAD-B	0.3443	0.4937	0.2643	0.3191	0.2994	0.2092	**0.5263**	0.1987	0.4530	0.4913	0.4202	0.3746	0.2959	0.2279	0.4217	0.2169
METAPAD-BA	0.3526	**0.5456**	0.2605	0.3410	0.3060	0.2167	0.5205	0.2033	0.4824	**0.5641**	0.4214	0.4144	0.3105	0.2066	**0.6248**	0.2272
METAPAD-BAS	**0.4356**	0.4610	0.4129	**0.3499**	**0.3873**	**0.3580**	0.4218	**0.2530**	**0.4949**	0.4098	**0.6244**	**0.4201**	0.3209	**0.2655**	0.4054	0.2282

With component	Comparison	WPB F1	WPB AUC	CNA F1	CNA AUC	APR F1	APR AUC	TWT F1	TWT AUC
Meta-pattern segmentation	METAPAD-T vs Baseline	-15.9%	2.6%	-8.1%	-7.2%	5.9%	-4.1%	-1.9%	-15.1%
	METAPAD-B vs Baseline	14.8%	65.6%	18.8%	50.6%	40.0%	70.7%	22.2%	58.0%
Type adjustment for granularity	METAPAD-TA vs METAPAD-T	45.0%	39.8%	33.3%	71.5%	38.2%	91.1%	32.5%	94.3%
	METAPAD-BA vs METAPAD-B	2.4%	6.9%	2.2%	2.3%	6.5%	10.6%	4.9%	4.7%
Synonymous meta pattern detection	METAPAD-TAS vs METAPAD-TA	16.8%	15.8%	19.3%	14.8%	3.2%	0.8%	3.1%	2.7%
	METAPAD-BAS vs METAPAD-BA	23.5%	2.6%	26.6%	24.5%	2.6%	1.4%	3.3%	0.4%
Overall	METAPAD-TAS vs Baseline	42.4%	66.1%	46.2%	82.8%	51.0%	84.6%	34.0%	69.3%
	METAPAD-BAS vs Baseline	45.2%	81.6%	53.7%	91.9%	52.9%	91.4%	32.5%	66.2%

Case Study: Attribute Names

Class type C= $COUNTRY	
1,132 names $\mathcal{A}$ in (C, $\mathcal{A}$, T=∅)	**3,930 (name, value type) pairs ($\mathcal{A}$, T≠∅)**
{government}	({president}, [$PRESIDENT])
{citizens, residents}	({ambassador}, [$POLITICIAN])
{war, wars, conflict}	({capital, capital_city}, [$CITY])
{volcano, volcano_eruption}	({dead}, [$DIGIT])
{border}	({prime_minister}, [$MINISTER])
{earthquake}	({condemn}, [$ORGANIZATION])
{protests, protest}	({population, people}, [$DIGIT $DIGITUNIT])
...	...
{security, national_security}	({district_judge}, [$POLITICIAN])
{parliament, congress}	({-magnitude_earthquake}, [$DIGIT])

	Attribute names of $ORGANIZATION				Attribute names of $ORGANIZATION.COMPANY			
#	SOIE	OLLIE	BIPERPEDIA	**METAPAD**	SOIE	OLLIE	BIPERPEDIA	**METAPAD**
1	be	said	president	{leaders, leader}	have	said	account	{homepage, website}
2	have	is	chairman	{officials, official}	be	is	ceo	{ceo, chief executive,
3	be in	reported	members	spokesman				chief executive officer}
4	say in	national	secretary	director	report	has	pipeline	pipeline
5	call	affiliate	leader	{president, head, chairman}	be in	announced	post	shares
6	will	was	spokesman	{members, member}	make	was	president	stores
7	issue	told	member	spokeswoman	say in	says	chief executive	employees
8	say	told affiliate	director	statement	announce	reported	shares	engineer
9	release	say	head	agents	of	had	founder	{founder, co-founder}
10	of	says	rebels	decision	release	released	train	analyst
11	take	has	committee	committee	face	told	employees	shareholders
...	...	...	...	...	...	...	...	...

Experimental Results

- Attribute tuple extraction

	WPB			CNA			APR			TWT		
	F1	AUC	#Tuple	F1	AUC	#Tuple	F1	AUC	#Tuple	F1	AUC	#Tuple
SOIE	0.0360	0.0124	183	0.0399	0.0070	132	0.0353	0.0133	288	0.0094	0.0012	115
OLLIE	0.0941	0.0517	417	0.1490	0.0626	470	0.1309	0.0900	562	0.0821	0.0347	698
RENOUN	0.3062	0.2275	505	0.2974	0.2360	766	0.3085	0.2497	860	0.2029	0.1256	860
METAPAD-T	0.3329	0.2424	592	0.3711	0.2381	1,224	0.4120	0.3016	1,299	0.4059	0.2641	1,080
METAPAD-TAS	**0.3946**	0.2880	**758**	0.3742	**0.2445**	**1,317**	0.4156	0.3269	**1,355**	**0.4153**	0.2854	**1,111**
METAPAD-B	0.2456	0.2523	305	0.3663	0.2184	883	0.3684	0.3186	787	0.3809	0.2704	650
METAPAD-BAS	0.3654	**0.3090**	526	**0.3972**	0.2443	1,014	**0.4236**	**0.3525**	1,040	0.3827	**0.3408**	775

With component	Comparison	WPB F1	WPB AUC	WPB #Tuple	CNA F1	CNA AUC	CNA #Tuple
Meta-pattern segmentation	METAPAD-T vs Baseline	8.7%	6.5%	17.2%	24.8%	0.9%	59.8%
	METAPAD-B vs Baseline	-19.8%	10.9%	-39.6%	23.2%	-7.5%	15.3%
Type adjustment & synonymous detection	METAPAD-TAS vs METAPAD-T	18.5%	18.8%	28.0%	0.9%	2.7%	7.6%
	METAPAD-BAS vs METAPAD-B	48.8%	22.5%	72.5%	8.4%	11.9%	14.8%
Overall	METAPAD-TAS vs Baseline	28.9%	26.6%	50.1%	25.9%	3.6%	71.9%
	METAPAD-BAS vs Baseline	19.3%	35.8%	4.2%	33.6%	3.5%	32.4%

With component	Comparison	APR F1	APR AUC	APR #Tuple	TWT F1	TWT AUC	TWT #Tuple
Meta-pattern segmentation	METAPAD-T vs Baseline	33.5%	20.8%	51.0%	100.0%	110.3%	25.6%
	METAPAD-B vs Baseline	19.4%	27.6%	-8.5%	87.7%	115.3%	-24.4%
Type adjustment & synonymous detection	METAPAD-TAS vs METAPAD-T	0.9%	8.4%	4.3%	2.3%	8.1%	2.9%
	METAPAD-BAS vs METAPAD-B	15.0%	10.6%	32.1%	0.5%	26.1%	19.2%
Overall	METAPAD-TAS vs Baseline	34.7%	30.9%	57.6%	104.6%	127.3%	29.2%
	METAPAD-BAS vs Baseline	37.3%	41.1%	20.9%	88.6%	171.3%	-9.9%

Case Study: Attribute Tuples

Ranked meta patterns	Entity e	Attribute value v
$COUNTRY President $PRESIDENT	U.S.	Barack Obama
$COUNTRY's president $PRESIDENT	Russia	Vladimir Putin
President $PRESIDENT of $COUNTRY	France	Francois Hollande
...	...	...
$PRESIDENT, the president of $COUNTRY,	Comoros	Ikililou Dhoinine
$PRESIDENT's government of $COUNTRY	Burkina Faso	Blaise Compaoré

Ranked meta patterns	Entity e	Attribute value v
$COMPANY CEO $BUSINESSPERSON	Apple	Tim Cook
$COMPANY chief executive $BUSINESSPERSON	Facebook	Mark Zuckerberg
$BUSINESSPERSON, the $COMPANY CEO,	Hewlett-Packard	Carly Fiorina
...	...	...
$COMPANY former CEO $BUSINESSPERSON	Infor	Charles Phillips
$BUSINESSPERSON, the $COMPANY former CEO,	Afghan Citadel	Roya Mahboob

Ranked meta patterns	Entity *e* (C= $TREATMENT)	Attribute value v (T= [$DISEASE])
"$TREATMENT was used to treat $DISEASE",	zoledronic acid therapy	Paget 's disease of bone
"$DISEASE using the $TREATMENT",	bisphosphonates	osteoporosis
"$TREATMENT has been used to treat $DISEASE",	calcitonin	Paget 's disease of bone
"$TREATMENT of patients with $DISEASE"...	calcitonin ...	osteoporosis ...

Ranked meta patterns	Entity *e* (C= $BACTERIA)	Attribute value v (T= [$ANTIBIOTICS])
"$BACTERIA was resistant to $ANTIBIOTICS",	corynebacterium striatum BM4687	gentamicin
"$BACTERIA are resistant to $ANTIBIOTICS",	corynebacterium striatum BM4687	tobramycin
"$BACTERIA is the most resistant to $ANTIBIOTICS",	methicillin-susceptible S aureus	vancomycin
"$BACTERIA, particularly those resistant to $ANTIBIOTICS" ...	multidrug-resistant enterobacteriaceae ...	gentamicin ...

Case Study: Attribute Tuples

#	SOIE	RENOUN	METAPAD
1	⟨Kuiper Belt, small body orbit beyond Neptune⟩	⟨Dearborn, US Census⟩	⟨Syria, [23, million]⟩
2	⟨Richmond Hill, number of resident⟩	⟨Bosnian Muslim, Serbian⟩	⟨Nepal, [28, million]⟩
3	⟨Cubans, numbering about 2 million total 54 million US Latinos⟩	⟨Somali, Minnesota Public Radio⟩	⟨Nepal, [27.8, million]⟩
4	⟨South Africa, growth⟩	⟨Catholic, Francis⟩	⟨Nigeria, [170, million]⟩
5	⟨Cubans, numbering about 2 million of 54 million US Latinos⟩		⟨Sri Lanka, [21, million]⟩
6	⟨South Africa, natural growth⟩		⟨Egypt, [86, million]⟩
7	⟨Richmond Hill, number of resident from Caribbean⟩		⟨Kenya, [45, million]⟩
8	⟨Torrance, US Census Bureau estimate⟩		⟨India, [1.27, billion]⟩
9	⟨France, less than 5 million each⟩		⟨Myanmar, [51.5, million]⟩
10	⟨Vanuatu, 50,000 people⟩		⟨Philippines, [100, million]⟩
11	⟨Cubans, numbering⟩		⟨China, [1.3, billion]⟩
12	⟨Cairns, government statistics reveal⟩		⟨Macedonia, [2, million]⟩
13	⟨US Census Bureau, noncitizen⟩		⟨Australia, [24, million]⟩
14	⟨Minneapolis, US⟩		⟨India, [300, million]⟩
...	...		...

- Efficiency test

	WPB	APR	CNA	CVD	TWT
#Meta Pattern	7,855	19,034	25,632	41,539	156,338
Max Memory	3.4G	5.3G	6.4G	10.5G	31G
Time Cost	17min	29min	37min	72min	117min

Acknowledgement

NSF
NSFC
National Natural Science Foundation of China
Carnegie Mellon University
UNIVERSITY OF MARYLAND 1856
Microsoft Research
微软亚洲研究院
Tencent 腾讯

References

J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu. "FreeSpan: Frequent pattern-projected sequential pattern mining." KDD, 2000.

X. Yan and J. Han. "gspan: Graph-based substructure pattern mining." ICDM, 2003.

X. Yan and J. Han. "CloseGraph: Mining closed frequent graph patterns." KDD, 2003.

Y. Sun, J. Han, X. Yan, P.S. Yu, and T. Wu. "PathSim: Meta path-based top-k similarity search in heterogeneous information networks." VLDB, 2011.

Y. Sun, Y. Yu, and J. Han. "Ranking-based clustering of heterogeneous information networks with star network schema." KDD, 2009.

Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. "RankClus: Integrating clustering with ranking for heterogeneous information network analysis." EDBT, 2009.

A. El-Kishky, Y. Song, C. Wang, C.R. Voss, and J. Han. "Scalable topical phrase mining from text corpora." VLDB, 2014.

References

J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. "Mining quality phrases from massive text corpora." SIGMOD, 2015.

X. Ren, A. El-Kishky, C. Wang, F. Tao, C.R. Voss, and J. Han. "Effective entity recognition and typing by relation phrase-based clustering." KDD, 2015.

X. Ren, W. He, M. Qu, C.R. Voss, H. Ji, and J. Han. "Label noise reduction in entity typing by heterogeneous partial-label embedding." KDD, 2016.

C. Wang, M. Danilevsky, N. Desai, Y. Zhang, P. Nguyen, T. Taula, and J. Han. "A phrase mining framework for recursive construction of a topical hierarchy." KDD, 2013.

R. Gupta, A. Halevy, X. Wang, S.E. Whang, and F. Wu. "Biperpedia: An ontology for search applications." VLDB, 2014.

M. Yahya, S. Whang, R. Gupta, and A. Halevy. "ReNoun: Fact extraction for nominal attributes." EMNLP, 2014.

A. Halevy, N. Noy, S. Sarawagi, S.E. Whang, and X. Yu. "Discovering structure in the universe of attribute names." WWW, 2016.

Thank you!

www.meng-jiang.com