Data-Driven Behavioral Analytics: Observations, Representations and Models

Meng Jiang (UIUC)

Peng Cui (Tsinghua)

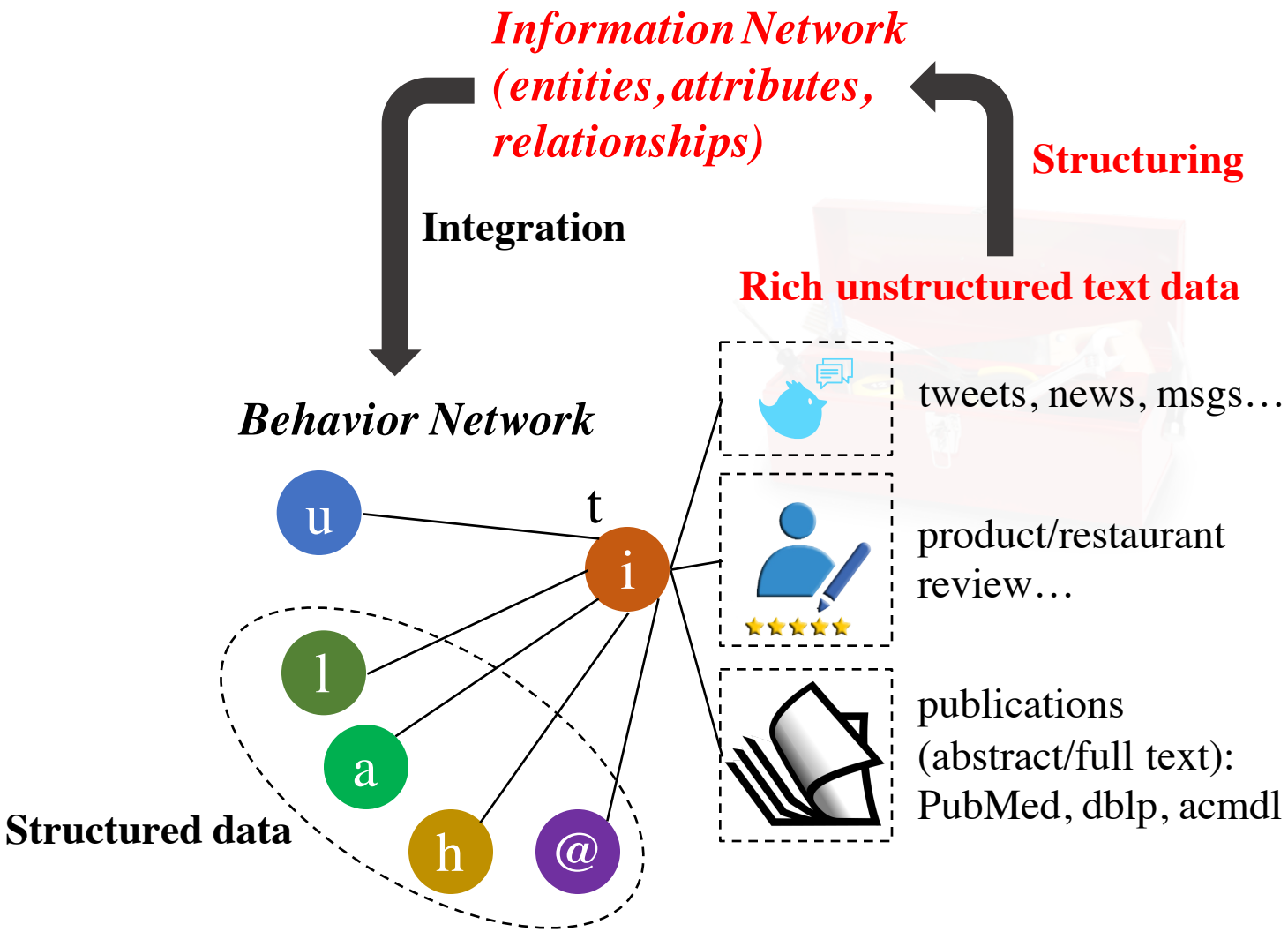
Jiawei Han (UIUC)

http://www.meng-jiang.com/tutorial-cikm16.html

Tutorial in CIKM 2016, October 24, Indianapolis, IN

II. Structuring behavioral content and integrating behavioral analysis with information networks

Data to Network to Knowledge

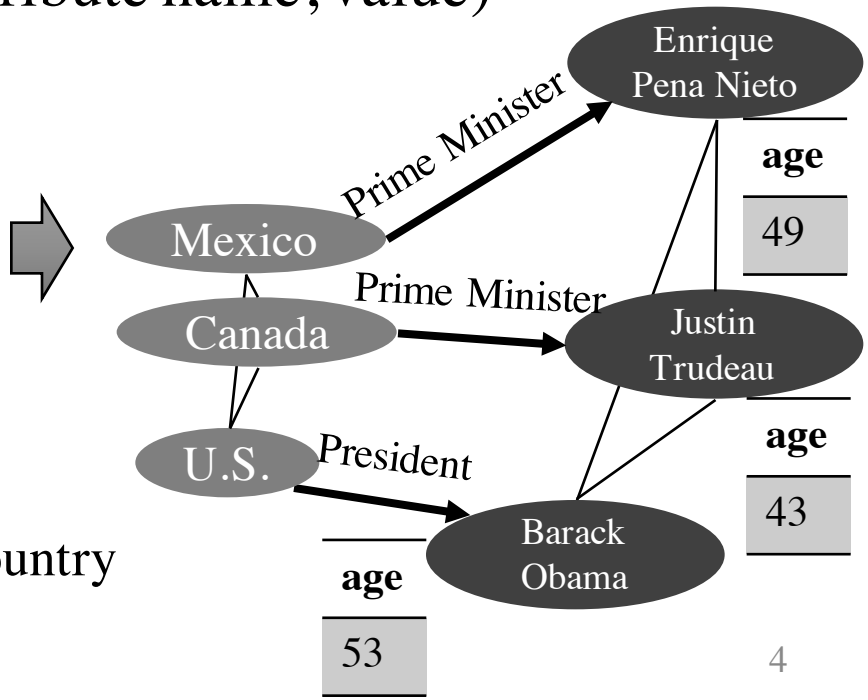


Construction of Heterogeneous Information Networks from Text

Philosophy: Not extensive "labeling" but exploring the power of massive text corpora!

- Mining phrases (the minimal semantic units)
- Entity recognition and typing
- Attribute discovery (entity, attribute name, value)

…here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie…of **Mexico's Enrique Pena Nieto, 49**, … **United States President Barack Obama, 53**, who…

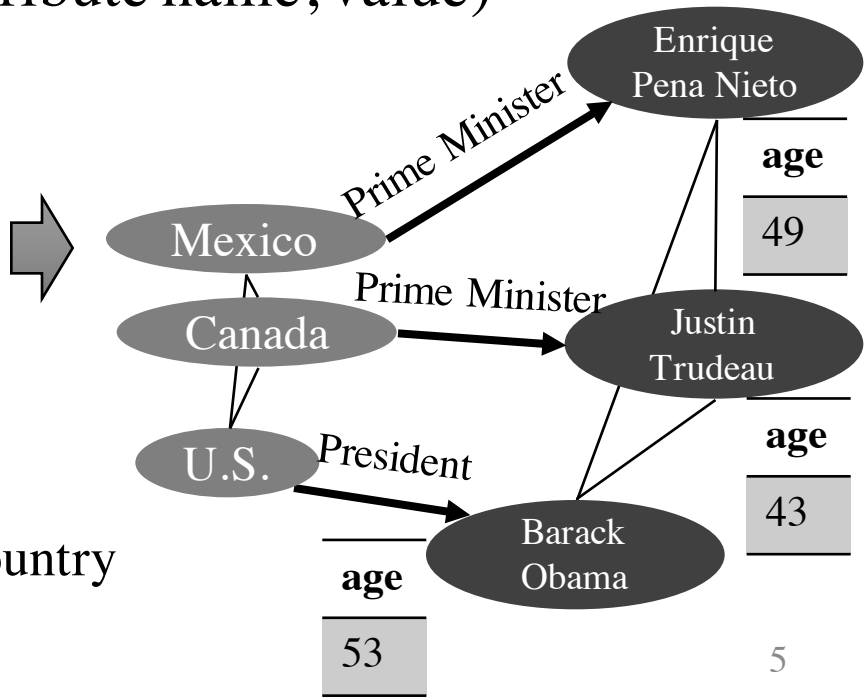


Construction of Heterogeneous Information Networks from Text

Philosophy: Not extensive "labeling" but exploring the power of massive text corpora!

- **Mining phrases (the minimal semantic units)** ⇦
- Entity recognition and typing
- Attribute discovery (entity, attribute name, value)

…here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie…of **Mexico's Enrique Pena Nieto, 49**, … **United States President Barack Obama, 53**, who…



Why Mining Phrases?

- **Unigrams** are *ambiguous but* **phrases** are natural, *unambiguous* semantic units
  - Ex.: "United" vs. United States, United Airline, United Parcel Service
- Mining semantically meaningful phrases
  - Transform text data from *word granularity* to *phrase granularity*
  - Enhance the power at manipulating unstructured data using information networks
- Phrase mining: Most NLP methods may need annotation and training
  - Annotate hundreds of documents as training data
  - Train a supervised model based on part-of-speech features
  - Limitations: High annotation cost
  - May not be scalable to domain-specific, dynamic, emerging applications
    - Scientific domains, query logs, or social media, e.g., Yelp, Twitter
- Minimal/no training but making good use of massing corpora

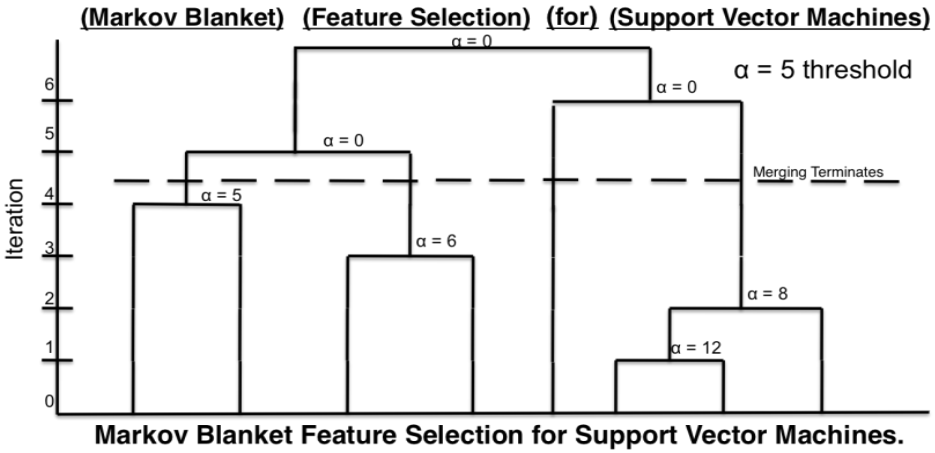
Strategies for Phrase Mining

- Strategy 1: Simultaneously inferring phrases and topics
  - Bigram topical model [Wallach'06], topical n-gram model [Wang, et al.'07], phrase discovering topic model [Lindsey, et al.'12]
  - High model complexity: Tends to overfitting; High inference cost: Slow
- Strategy 2: Post topic modeling phrase construction
  - Label topic [Mei et al.'07], TurboTopic [Blei & Lafferty'09], KERT [Danilevsky, et al.'14]
  - Words in the same phrase may be assigned to different topics
    - Ex. … knowledge discovery using least squares support vector machine …
- **Our solution 1: ToPMine** [El-kishky, et al., VLDB'15]
  - First Phrase Mining then Topic Modeling (No training data at all)
- **Our solution 2: SegPhrase+** [Liu, et al., SIGMOD'15]
  - Integrating phrase mining and document segmentation (with minimal training data)

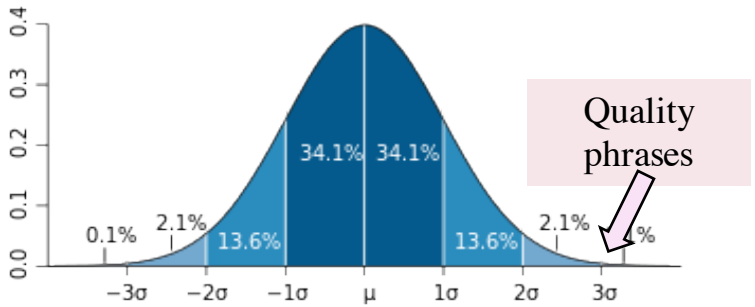
ToPMine: The Overall Phrase Mining Framework

- **ToPMine** [El-Kishky et al. VLDB'15]
  - First phrase construction, then topic mining
  - Contrast with KERT: First topic modeling, then phrase mining
- **The ToPMine Framework:**
  - Perform frequent ***contiguous pattern*** mining to extract candidate phrases and their counts
  - Perform agglomerative merging of adjacent unigrams as guided by a significance score—This segments each document into a ***"bag-of-phrases"***
  - The newly formed bag-of-phrases are passed as input to PhraseLDA, an extension of LDA, that constrains all words in a phrase to each sharing the same latent topic

Phrase Mining: Frequent Pattern Mining + Statistical Analysis



Markov Blanket Feature Selection for Support Vector Machines.



Based on significance score [Church et al.'91]:

$$\alpha(P_1, P_2) \approx (f(P_1 \bullet P_2) - \mu_0(P_1, P_2)) / f(P_1 \bullet P_2)^{1/2}$$

[Markov blanket] [feature selection] for [support vector machines]
[knowledge discovery] using [least squares] [support vector machine] [classifiers]
…[support vector] for [machine learning]…

Phrase	Raw freq.	True freq.
[support vector machine]	90	80
[vector machine]	95	0
[support vector]	100	20

What Kind of Phrases are of "High Quality"?

- Judging the quality of phrases
  - **Popularity**
    - "information retrieval" vs. "cross-language information retrieval"
  - **Concordance**
    - "powerful tea" vs. "strong tea"
    - "active learning" vs. "learning classification"
  - **Informativeness**
    - "this paper" (frequent but not discriminative, not informative)
  - **Completeness**
    - "vector machine" vs. "support vector machine"

ToPMine: Experiments on Yelp Reviews

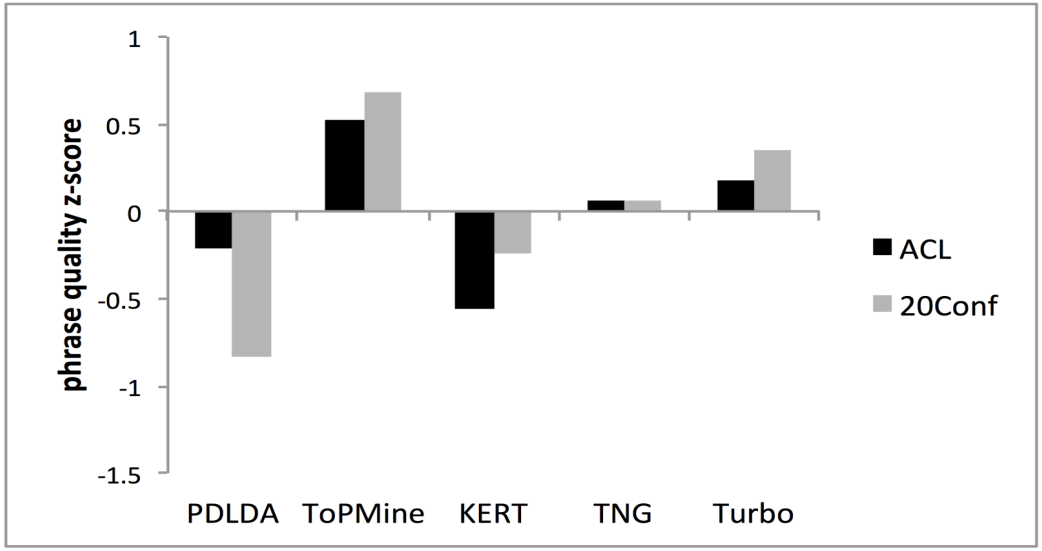
	Topic 1	*Topic 2*	*Topic 3*	*Topic 4*	*Topic 5*
unigrams	coffee	food	room	store	good
	ice	good	parking	shop	food
	cream	place	hotel	prices	place
	flavor	ordered	stay	find	burger
	egg	chicken	time	place	ordered
	chocolate	roll	nice	buy	fries
	breakfast	sushi	place	selection	chicken
	tea	restaurant	great	items	tacos
	cake	dish	area	love	cheese
	sweet	rice	pool	great	time
n-grams	ice cream	spring rolls	parking lot	grocery store	mexican food
	iced tea	food was good	front desk	great selection	chips and salsa
	french toast	fried rice	spring training	farmer's market	food was good
	hash browns	egg rolls	staying at the hotel	great prices	hot dog
	frozen yogurt	chinese food	dog park	parking lot	rice and beans
	eggs benedict	pad thai	room was clean	wal mart	sweet potato fries
	peanut butter	dim sum	pool area	shopping center	pretty good
	cup of coffee	thai food	great place	great place	carne asada
	iced coffee	pretty good	staff is friendly	prices are reasonable	mac and cheese
	scrambled eggs	lunch specials	free wifi	love this place	fish tacos

ToPMine: Faster and Generating Better Quality Phrases

Running time of different algorithms

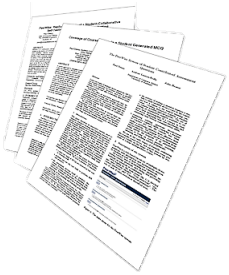
Method	*sampled dblp titles (k=5)*	*dblp titles (k=30)*	*sampled dblp abstracts*	*dblp abstracts*
PDLDA	3.72(hrs)	~20.44(days)	1.12(days)	~95.9(days)
Turbo Topics	6.68(hrs)	>30(days)*	>10(days)*	>50(days)*
TNG	146(s)	5.57 (hrs)	853(s)	NA†
LDA	**65(s)**	3.04 (hrs)	353(s)	13.84(hours)
KERT	68(s)	3.08(hrs)	1215(s)	NA†
ToP-Mine	67(s)	**2.45(hrs)**	**340(s)**	**10.88(hrs)**

Phrase quality measured by z-score

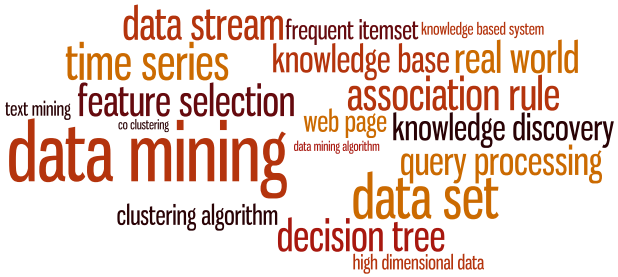


SegPhrase: From Raw Corpus to Quality Phrases and Segmented Corpus

Raw Corpus



Quality Phrases



Segmented Corpus

Document 1
Citation recommendation is an interesting but challenging research problem in data mining

Document 2
In this study, we investigate the problem in the context of heterogeneous information networks using data mining technique.

Document 3
Principal Component Analysis is a linear dimensionality reduction technique commonly used in machine learning applications.

+ A small set of labels or a general KB

Input Raw Corpus → **Quality Phrases** ↔ **Segmented Corpus**

Phrase Mining **Phrasal Segmentation**

Integrating phrase mining with phrasal segmentation

J. Liu et al. **Mining Quality Phrases from Massive Text Corpora.** SIGMOD, 2015 (**won Grand Prize in Yelp Dataset Challenge**).

Experiments: Interesting Phrases Generated
(From the Titles and Abstracts of SIGMOD)

Query	SIGMOD	
Method	**SegPhrase+**	**Chunking (TF-IDF & C-Value)**
1	data base	data base
2	database system	database system
3	relational database	query processing
4	query optimization	query optimization
5	query processing	relational database
…	…	…
51	sql server	database technology
52	**relational data**	database server
53	data structure	**large volume**
54	**join query**	performance study
55	web service	web service
…	…	…
201	**high dimensional data**	**efficient implementation**
202	location based service	sensor network
203	xml schema	**large collection**
204	**two phase locking**	**important issue**
205	deep web	frequent itemset
…	…	…

Only in SegPhrase+

Only in Chunking

Mining Quality Phrases in Multiple Languages

- Both ToPMine and SegPhrase+ are extensible to mining quality phrases in multiple languages
  - SegPhrase+ on Chinese (From Chinese Wikipedia)
  - ToPMine on Arabic (From Quran Fus7a Arabic)(no preprocessing)
    - Experimental results of Arabic phrases:

      اورفڪ → Those who disbelieve

      مسبر الله نمحرلا ميحرلا → In the name of God the Gracious and Merciful

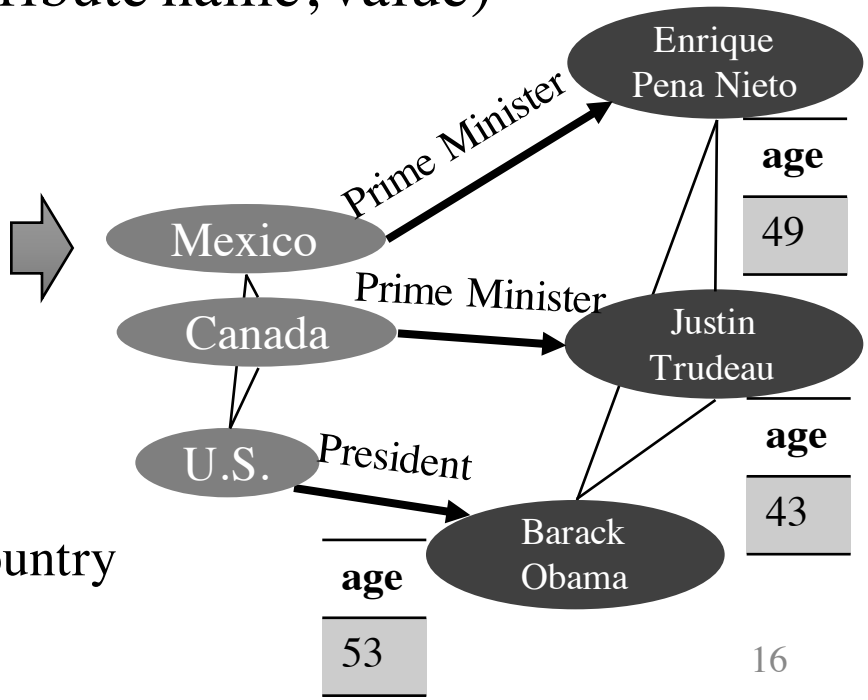
Rank	Phrase	In English
…	…	…
62	首席_执行官	CEO
63	中间_偏右	Middle-right
…	…	…
84	百度_百科	Baidu Pedia
85	热带_气旋	Tropical cyclone
86	中国科学院_院士	Fellow of Chinese Academy of Sciences
…	…	…
1001	十大_中文_金曲	Top-10 Chinese Songs
1002	全球_资讯网	Global Info Website
1003	天一阁_藏_明代_科举_录_选刊	A Chinese book name
…	…	…
9934	国家_戏剧_院	National Theater
9935	谢谢_你	Thank you
…	…	…

Construction of Heterogeneous Information Networks from Text

Philosophy: Not extensive "labeling" but exploring the power of massive text corpora!

- Mining phrases (the minimal semantic units)
- **Entity recognition and typing**
- Attribute discovery (entity, attribute name, value)

…here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie…of **Mexico's Enrique Pena Nieto, 49**, … **United States President Barack Obama, 53**, who…



Why Entity Recognition and Typing from Massive Corpora?

- Traditional named entity recognition systems are designed for major types (e.g., PER, LOC, ORG) and general domains (e.g., news)
  - Require additional steps to adapt to new domains/types
  - Expensive human labor on annotation
    - 500 documents for entity extraction; 20,000 queries for entity linking
  - Unsatisfying agreement due to various granularity levels and scopes of types
- Entities obtained by entity linking techniques have *limited* coverage and freshness
  - > 50% unlinkable entity mentions in Web corpus [Lin et al., EMNLP'12]
  - > 90% in our experiment corpora: tweets, Yelp reviews, …
- A new approach: ClusType: Entity Recognition and Typing by Relation Phrase-Based Clustering [Ren, et al., KDD 2015]
  - Recognizing entity mentions of target types with minimal/no human supervision and with no requirement that entities can be found in a KB (distant supervision)

Recognizing Typed Entities

Identifying token span as entity mentions in documents and labeling their types

Target Types

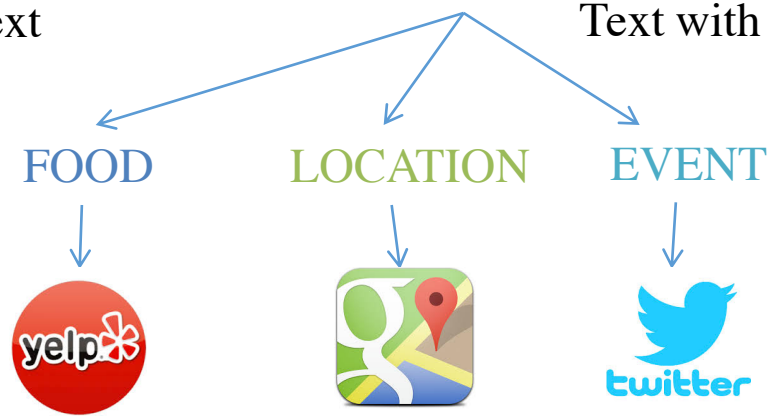
- FOOD
- LOCATION
- JOB_TITLE
- EVENT
- ORGANIZATION
- …

The best BBQ I've tasted in Phoenix! I had the pulled pork sandwich with coleslaw and baked beans for lunch. ... The owner is very nice. …

Plain text

The best BBQ:Food I've tasted in Phoenix:LOC ! I had the [pulled pork sandwich]:Food with coleslaw:Food and [baked beans]:Food for lunch. … The owner:JOB_TITLE is very nice. …

Text with typed entities

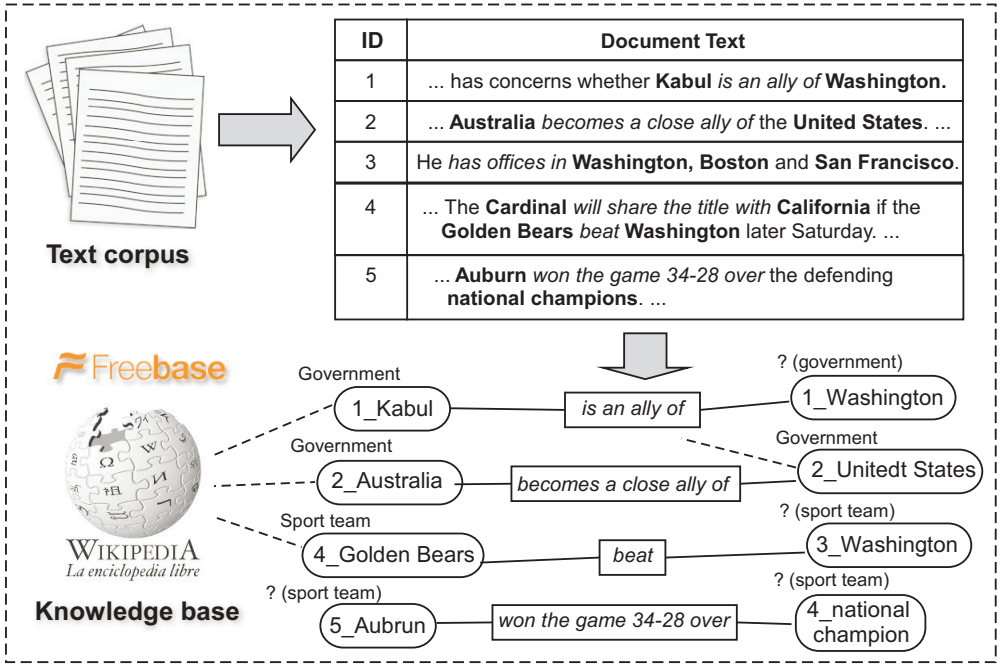


Enabling structured analysis of unstructured text corpus

ClusType: A Distant Supervision Framework

Problem: *Distantly-supervised entity recognition in a domain-specific corpus*

- Given: (1) a domain-specific corpus *D*, (2) a knowledge base (e.g., Freebase), (3) a set of target types (*T*) from a KB
- Detect candidate entity mentions in D, and categorize each candidate mention by target types or Not-Of-Interest (NOI)



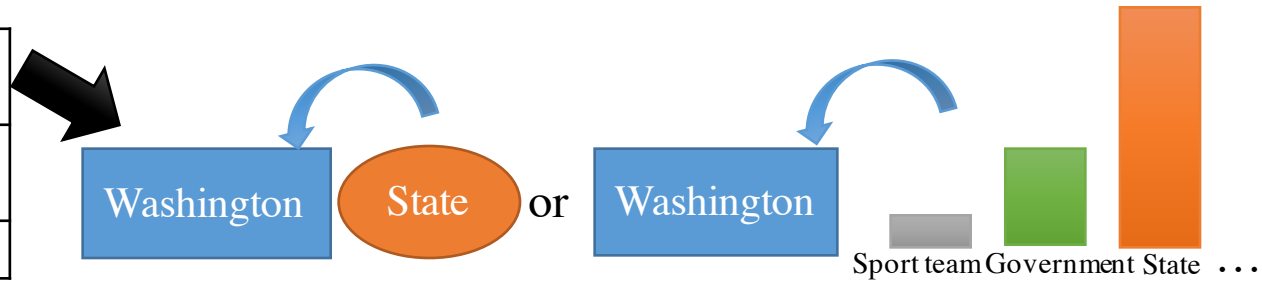
Solution:

- Detect entity mentions from text
- Map candidate mentions to KB entities of target types
- Use confidently mapped {mention, type} to infer types of remaining candidate mentions

Entity Recognition and Typing: Challenges and Solutions

- Challenge 1: Domain Restriction: Extensive training, use general-domain corpora, not work well on specific, dynamic or emerging domains (*e.g.*, tweets, Yelp reviews)
  - Solution: Domain-agnostic phrase mining: Extracts candidate entity mentions with minimal linguistic assumption (e.g., only use POS tagging)
- Challenge 2: Name ambiguity: Multiple entities may share the same surface name
  - Solution: Model each mention based on its surface name and context

While Griffin is not the part of Washington's plan on Sunday's game, …
… news from Washington indicates that the congress is going to…
It is one of the best state parks in Washington.



- Challenge 3: Context Sparsity: There are many ways to describe the same relation
  - Solution: cluster *relation phrase*, infer synonymous relation phrases

Sentence	Freq.
The magnitude 9.0 quake **caused widespread devastation in [Kesennuma city]**	12
… tsunami that **ravaged [northeastern Japan]** last Friday	31
The resulting tsunami **devastate [Japan]**'s northeast	244

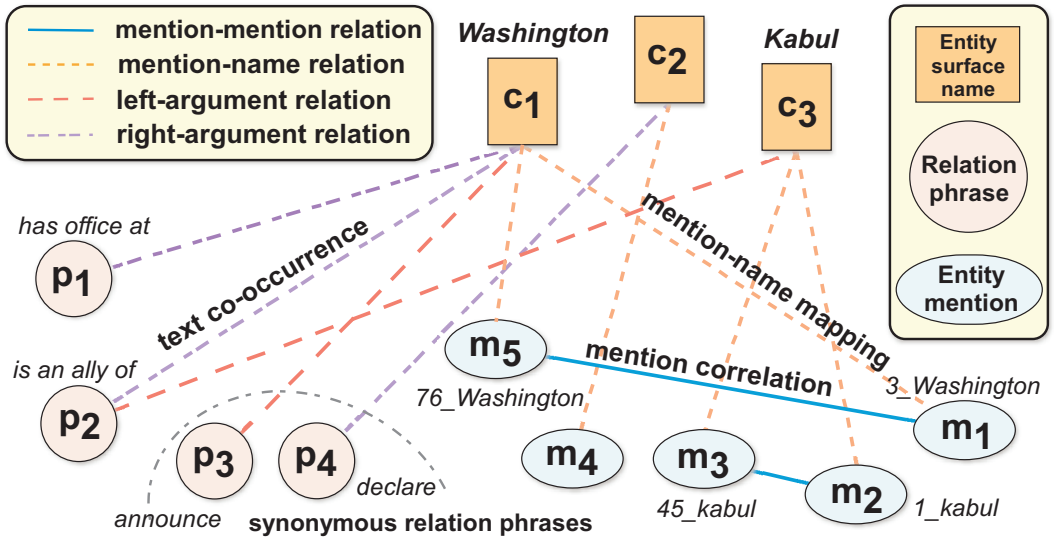
The ClusType Framework: Phrase Segmentation and Heterogeneous Graph Construction

- POS-constrained phrase segmentation for mining candidate entity mentions and relation phrases, simultaneously
- Construct a heterogeneous graph to represent available information in a unified form

Entity mentions are kept as individual objects **to be disambiguated**

Linked to entity surface names & relation phrases

Weight assignment: The more two objects are likely to share the same label, the larger the weight will be associated with their connecting edge

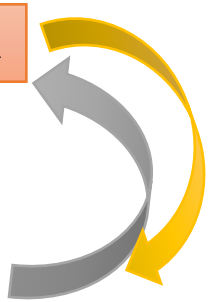


The ClusType Framework: Mutual Enhancement of Type Propagation and Relation Phrase Clustering

- With the constructed graph, formulate a graph-based semi-supervised learning of two tasks jointly:

Type propagation on heterogeneous graph

Derived entity argument types serve as **good feature** for clustering relation phrases



Multi-view relation phrase clustering

Propagate type information among entities bridges via synonymous relation phrases

Mutually enhancing each other; leads to quality recognition of unlinkable entity mentions

ClusType: A General Framework Overview

- **Candidate Generation**
  - Perform phrase mining on a POS-tagged corpus to extract candidate entity mentions and relation phrases
- **Construction of Heterogeneous Graphs**
  - Construct a heterogeneous graph to encode our insights on modeling the type for each entity mention
  - Collect seed entity mentions as labels by linking extracted mentions to the KB
- **Relation Phrase Clustering**
  - Estimate type indicator for unlinkable candidate mentions with the proposed type propagation integrated with relation phrase clustering on the constructed graph

Candidate Generation

- Phrase mining incorporating both *corpus-level statistics* and *syntactic constraints*
  - **Global significance score**: Filter low-quality candidates; **generic POS tag patterns**: remove phrases with improper syntactic structure
  - Extend ToPMine to partition corpus into segments which meet both significance threshold and POS patterns → candidate entity mentions & relation phrases

Relation phrase: Phrase that denotes a unary or binary relation in a sentence

Pattern	Example
V	disperse; hit; struck; knock;
P	in; at; of; from; to;
V P	locate in; come from; talk to;
VW*(P)	caused major damage on; come lately

V-verb; P-prep; W-{adv | adj | noun | det | pron}
W* denotes multiple W; (P) denotes optional.

Experiment: Entity detection: Performance comparison between our method and an NP chunker

Method	NYT		Yelp		Tweet	
	Prec	Recall	Prec	Recall	Prec	Recall
Our method	**0.469**	**0.956**	**0.306**	**0.849**	0.226	**0.751**
NP chunker	0.220	0.609	0.296	0.247	**0.287**	0.181

Recall is most critical for this step,since later we cannot detect the misses (i.e., false negatives)

Type Inference: A Joint Optimization Problem

$$\mathcal{O}_{\alpha,\gamma,\mu} = \mathcal{F}(\mathbf{C}, \mathbf{P}_L, \mathbf{P}_R) + \mathcal{L}_\alpha\big(\mathbf{P}_L, \mathbf{P}_R, \{\mathbf{U}^{(v)}, \mathbf{V}^{(v)}\}, \mathbf{U}^*\big) + \Omega_{\gamma,\mu}(\mathbf{Y}, \mathbf{C}, \mathbf{P}_L, \mathbf{P}_R). \quad (2)$$

$$\mathcal{F}(\mathbf{C}, \mathbf{P}_L, \mathbf{P}_R) = \sum_{i=1}^{n}\sum_{j=1}^{l} W_{L,ij} \left\| \frac{\mathbf{C}_i}{\sqrt{D^{(\mathcal{C})}_{L,ii}}} - \frac{\mathbf{P}_{L,j}}{\sqrt{D^{(\mathcal{P})}_{L,jj}}} \right\|_2^2 + \sum_{i=1}^{n}\sum_{j=1}^{l} W_{R,ij} \left\| \frac{\mathbf{C}_i}{\sqrt{D^{(\mathcal{C})}_{R,ii}}} - \frac{\mathbf{P}_{R,j}}{\sqrt{D^{(\mathcal{P})}_{R,jj}}} \right\|_2^2$$

Mention modeling & mention correlation

$$\Omega_{\gamma,\mu}(\mathbf{Y}, \mathbf{C}, \mathbf{P}_L, \mathbf{P}_R) = \big\|\mathbf{Y} - f(\Pi_{\mathcal{C}}\mathbf{C}, \Pi_L\mathbf{P}_L, \Pi_R\mathbf{P}_R)\big\|_F^2 + \frac{\gamma}{2}\sum_{c\in\mathcal{C}}\sum_{i,j=1}^{M_c} W^{(c)}_{ij} \left\| \frac{\mathbf{Y}_i}{\sqrt{D^{(c)}_{ii}}} - \frac{\mathbf{Y}_j}{\sqrt{D^{(c)}_{jj}}} \right\|_2^2 + \mu\|\mathbf{Y} - \mathbf{Y}_0\|_F^2$$

Type propagation between entity surface names and relation phrases

$$\mathcal{L}_\alpha\big(\mathbf{P}_L, \mathbf{P}_R, \{\mathbf{U}^{(v)}, \mathbf{V}^{(v)}\}, \mathbf{U}^*\big) = \sum_{v=0}^{d} \beta^{(v)} \big(\|\mathbf{F}^{(v)} - \mathbf{U}^{(v)}\mathbf{V}^{(v)T}\|_F^2 + \alpha\|\mathbf{U}^{(v)}\mathbf{Q}^{(v)} - \mathbf{U}^*\|_F^2\big). \quad (3)$$

Multi-view relation phrases clustering)

ClusType: Experiment Setting

- Datasets: 2013 New York Times news (~110k docs) [event, PER, LOC, ORG]; Yelp Reviews (~230k) [Food, Job, …]; 2011 Tweets (~300k) [event, product, PER, LOC, …]
- Seed mention sets: < 7% extracted mentions are mapped to Freebase entities
- Evaluation sets: manually annotate mentions of target types for subsets of the corpora
- Evaluation metrics: Follows named entity recognition evaluation (Precision, Recall, F1)
- Compared methods
  - **Pattern**: Stanford pattern-based learning; **SemTagger**:bootstrapping method which trains contextual classifier based on seed mentions; **FIGER**: distantly-supervised sequence labeling method trained on Wiki corpus; **NNPLB**: label propagation using ReVerb assertion and seed mention; **APOLLO**: mention-level label propagation using Wiki concepts and KB entities;
  - **ClusType-NoWm**: ignore mention correlation; **ClusType-NoClus**: conducts only type propagation; **ClusType-TwpStep**: first performs hard clustering then type propagation

Comparing ClusType with Other Methods and Its Variants

Performance comparison on three datasets in terms of Precision, Recall and F1 score

Table 5: Performance comparisons on three datasets in terms of Precision, Recall and F1 score.

Data sets	NYT			Yelp			Tweet		
Method	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Pattern [9]	0.4576	0.2247	0.3014	0.3790	0.1354	0.1996	0.2107	0.2368	0.2230
FIGER [16]	0.8668	0.8964	0.8814	0.5010	0.1237	0.1983	**0.7354**	0.1951	0.3084
SemTagger [12]	0.8667	0.2658	0.4069	0.3769	0.2440	0.2963	0.4225	0.1632	0.2355
APOLLO [29]	0.9257	0.6972	0.7954	0.3534	0.2366	0.2834	0.1471	0.2635	0.1883
NNPLB [15]	0.7487	0.5538	0.6367	0.4248	0.6397	0.5106	0.3327	0.1951	0.2459
ClusType-NoClus	0.9130	0.8685	0.8902	0.7629	0.7581	0.7605	0.3466	0.4920	0.4067
ClusType-NoWm	0.9244	0.9015	0.9128	0.7812	0.7634	0.7722	0.3539	**0.5434**	0.4286
ClusType-TwoStep	0.9257	0.9033	0.9143	0.8025	0.7629	0.7821	0.3748	0.5230	0.4367
ClusType	**0.9550**	**0.9243**	**0.9394**	**0.8333**	**0.7849**	**0.8084**	0.3956	0.5230	**0.4505**

- **vs. FIGER**: Effectiveness of our candidate generation and type propagation
- **vs. NNPLB** and **APOLLO**: ClusType utilizes not only semantic-rich relation phrase as type cues, but also cluster synonymous relation phrases to tackle context sparsity
- **vs.** our **variants**: (i) models mention correlation for name disambiguation; and (ii) integrates clustering in a mutually enhancing way

Comparing on Trained NER System

- Compare with Stanford NER, which is trained on general-domain corpora including ACE corpus and MUC corpus, on three types: PER, LOC, ORG

F1 score comparison with trained NER

Table 6: F1 score comparison with trained NER.

Method	NYT	Yelp	Tweet
Stanford NER [6]	0.6819	0.2403	0.4383
ClusType-NoClus	0.9031	0.4522	0.4167
ClusType	**0.9419**	**0.5943**	**0.4717**

[6] J. R. Finkel, T. Grenager and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In ACL'05.

- ClusType and its variants outperform Stanford NER on both dynamic corpus (NYT) and domain-specific corpus (Yelp)
- ClusType has lower precision but higher Recall and F1 score on Tweet → Superior recall of ClusType mainly come from domain-independent candidate generation

Example Output and Relation Phrase Clusters

Example output of ClusType and the compared methods on the Yelp dataset

ClusType	SemTagger	NNPLB
The best **BBQ:Food** I've tasted in **Phoenix:LOC** ! I had the **[pulled pork sandwich]:Food** with **coleslaw:Food** and **[baked beans]:Food** for lunch. ...	The best BBQ I've tasted in **Phoenix:LOC** ! I had the pulled **[pork sandwich]:LOC** with **coleslaw:Food** and **[baked beans]:LOC** for lunch. ...	The best **BBQ:Loc** I've tasted in **Phoenix:LOC** ! I had the pulled pork **sandwich:Food** with coleslaw and baked **beans:Food** for **lunch:Food**. ...
I only go to **ihop:LOC** for **pancakes:Food** because I don't really like anything else on the menu. Ordered **[chocolate chip pancakes]:Food** and a **[hot chocolate]:Food**.	I only go to ihop for pancakes because I don't really like anything else on the menu. Ordered **[chocolate chip pancakes]:LOC** and a **[hot chocolate]:LOC**.	I only go to ihop for pancakes because I don't really like anything else on the menu. Ordered chocolate chip pancakes and a hot chocolate.

- **Extracts more mentions and predicts types with higher**

Example relation phrase clusters and corpus-wide frequency from the NYT dataset

ID	Relation phrase
1	recruited by (5.1k); employed by (3.4k); want hire by (264)
2	go against (2.4k); struggling so much against (54); run for re-election against (112); campaigned against (1.3k)
3	looking at ways around (105); pitched around (1.9k); echo around (844); present at (5.5k);

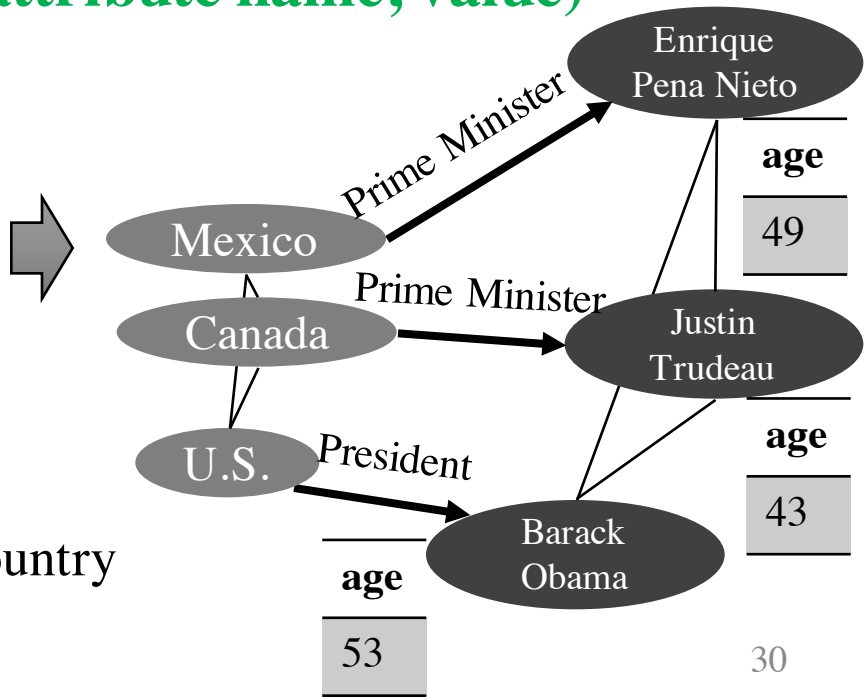
- Not only synonymous relation phrases, but also both sparse and frequent relation phrase can be clustered together
- → boosts sparse relation phrases with type information of frequent relation phrases

Construction of Heterogeneous Information Networks from Text

Philosophy: Not extensive "labeling" but exploring the power of massive text corpora!

- Mining phrases (the minimal semantic units)
- Entity recognition and typing
- **Attribute discovery (entity, attribute name, value)**

…here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie…of **Mexico's Enrique Pena Nieto, 49**, … **United States President Barack Obama, 53**, who…



Attributed Network Construction

- *Automatic* Attribute discovery: Given a class (*e.g.*, $Country)
  - **Feature** as a characteristic (*e.g.*, "population")
    - Value: the feature value (*e.g.*, $Digit or NULL)
  - **Relationship** with another class (*e.g.*, "prime minister")
    - Value: the other class (*e.g.*, $Person.Politician.PrimeMinister)
- Google's [VLDB'14, WWW'16] based on **fact-seeking** queries
  - Challenge 1: (Class, Attribute name, **Attribute value**)
  - Challenge 2: Just text documents (news, tweets, etc.). **NO query**.

"canada prime minister", "trudeau age",
"united states president", "obama age",
"mexico prime minister" …

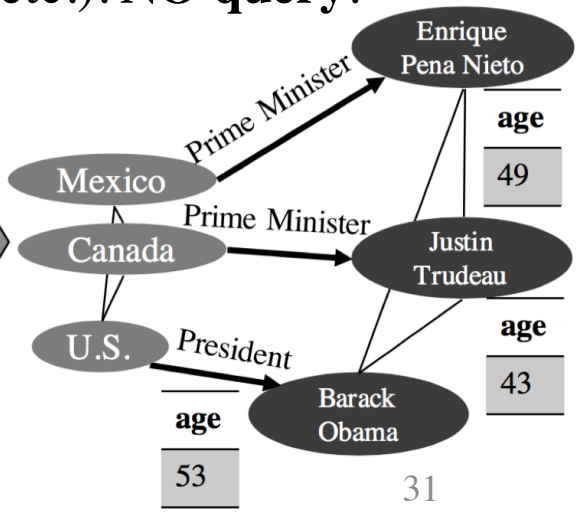
Unfortunately, we don't have the query data.

…here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie…of **Mexico's Enrique Pena Nieto, 49**, … **United States President Barack Obama, 53**, who…

Fortunately, we have large text corpus.

	Canada	Mexico
Prime Minister	Justin Trudeau	Enrique Pena Nieto

	Justin Trudeau	Enrique Pena Nieto	Barack Obama
age	43	49	53



Related Work

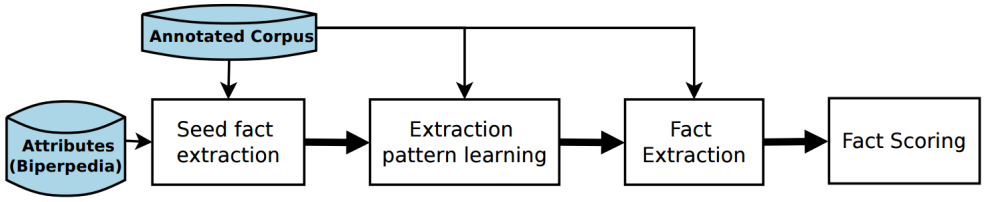
	Data source (√: available and unlimited with web crawler)	**Supervision level** (√: distant or unsupervised for massive data)	**Attribute discovery**		
			Fine-grained classes	Attribute names	Attribute values
CO-EM [6]	×, product description	?, semi-supervised	×	√	√
SE+R [2]	×, query log	?, weakly	×	√	×
DVSQ† [23]	×, query log	√, unsupervised	√	√	×
CAE† [22]	×, query log	√, unsupervised	√	√	×
WSIEQ† [20]	×, query log	?, weakly	√	√	×
HCAE† [21]	×, query log	?, weakly	√	√	×
WSEST† [24]	×, query log+HTML table	?, weakly	√	√	×
KNEXT [29]	√, text corpus	×, supervised	√	√	×
FAR [12]	×, HTML table	×, supervised	√	√	√
TYPICALITY [13]	×, query log+HTML table	×, supervised	√	√	×
BIPERPEDIA‡ [7]	×, query log+HTML table	√, distant	√	√	×
RENOUN‡ [30]	√, text corpus	×, supervised	√	×	√
ARI‡ [8]	×, query log	√, distant	√	√	×
UPSF [33]	√, text corpus	√, unsupervised	×	×	√
MetaPAD	**√, text corpus**	**√, distant**	√	√	√

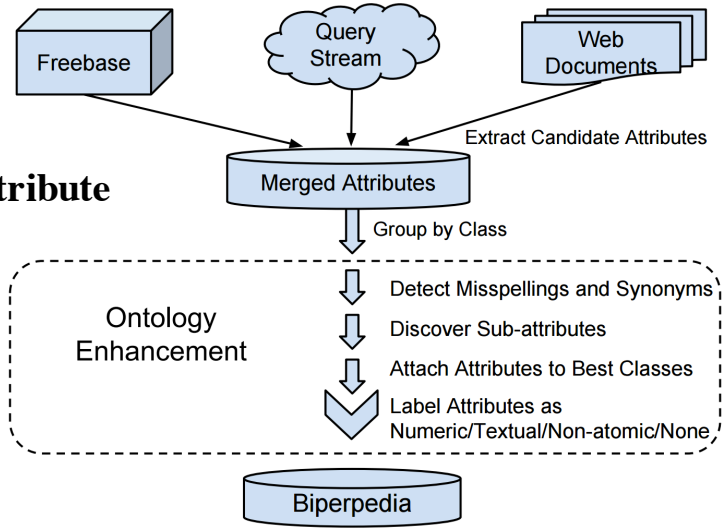
†These related papers were published by Dr. Marius Paşca et al., Google Inc. from 2007 to 2008.

‡These related papers were published by Dr. Alon Halevy et al., Google Inc. from 2014 to 2016.

Google's Approaches on Attribute Extraction

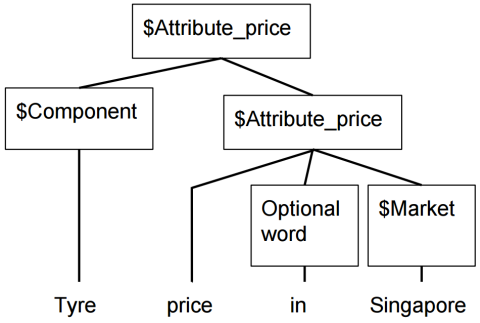
- Given Google's **query log**, web text and knowledge bases
  - "Obama wife name"… "Japan asian population", "Brazil female latino population", "Princeton economist"…
  - "Obama's wife, Michelle Obama, is a lawyer...", "Princeton economist Paul Krugman was awarded…"…
  - Obama: *$Person, $President*; Japan, Brazil: *$Location, $Country*; Princeton: *$Organization, $University*…
- Biperpedia (VLDB'14): **Attribute Name Extraction** from query log
  - $Person: wife name, daughter name
  - $Country: asian population, female latino population
  - $University: economist
- ReNoun (EMNLP'14): **Fact Extraction for Noun Phrase Attribute**
  - (Obama, wife, Michelle Obama)
  - (Princeton, economist, Paul Krugman)



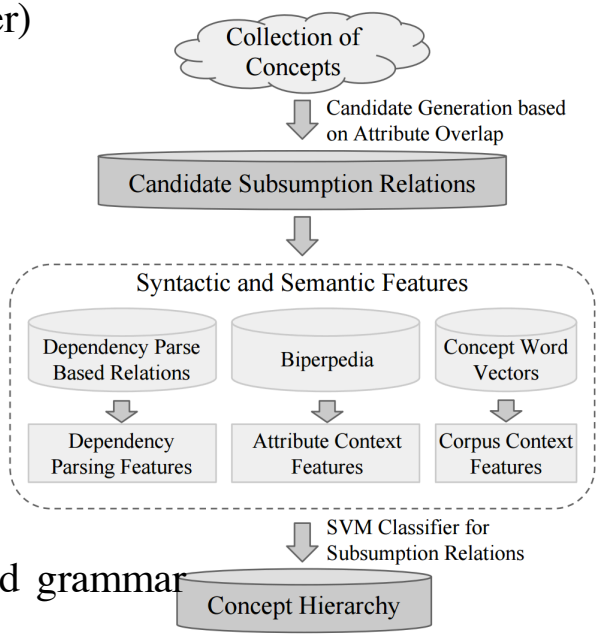


Google's Approaches on Attribute Extraction

- Latte (WebDB'15 Best Paper): **Concept (Type) Hierarchy Extraction** with attribute features
  - {country, address, zip code}: $University (sub) - $Location (super)
  - {online payment, non profit, tax return}: $University (sub) - $Organization (super)
  - {daughter name, wife name, age}: $President (sub) - $Person (super)




```
$Attribute_price ::= price
$Attribute_price ::= $Market $Attribute_price
$Attribute_price ::= $Attribute_price in $Market
$Attribute_price ::= $Component $Attribute_price
$Market          ::= Singapore | USA | Dubai | London | ...
$Component       ::= battery | bumper | tyre | door | ...
```



- ARI (WWW'16): **Attribute Name Structure Extraction** with rule-based grammar
  - Long-tail distribution of attribute names
  - $Person: $FamilyMember (name) - daughter, wife, mother, daughter name, wife name
  - $Country: ($Gender) ($Ethnicity) population - asian population, female latino population

Data-Driven: 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.

News:

…he's gotten older and grayer, and he's been eclipsed at an Asian economic forum here by **Canada Prime Minister Justin Trudeau, 43**, the so-called #APEChottie… He's also the youngest leader at the Asia Pacific Economic Cooperation forum, six years the junior of **Mexico's Enrique Pena Nieto, 49**, … **Obama, 53**, who becomes the elder statesman…

→ **1. $Person, $Digit,**
2. $Location.Country Prime_Minister
$Person.Politician.PrimeMinister

Tweets:

…Protestors march to **Gordon Square** for **12** -year-old **Tamir Rice**...

→ **1. protestors march to $Location.Square**
2. $Digit -year-old $Person.Victim

PubMed abstract:

… **Endocarditis** caused by **Streptococcus pneumoniae**… **Pericarditis** due to **Neisseria meningitidis** …

→ **$Cardiovasular_Diseases caused by $Bacteria**
$Cardiovasular_Diseases due to $Bacteria

MetaPAD Framework

Integrated Data-Driven Text Mining

Meta Pattern Mining

Attribute Extraction from Meta Patterns

… Canada Prime Minister Justin Trudeau …
… Barack Obama , 53, …

Quality phrase mining (SegPhrase, SIGMOD'15)

… Canada **Prime_Minister Justin_Trudeau** …
… **Barack_Obama** , 53, …

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

… **$Location** Prime_Minister **$Person** …
… **$Person** , **$Digit** , …

Fine-grained typing (PLE, KDD'16)

… **$Country** Prime_Minister **$PrimeMinister**…
… **$President**, **$Digit** , …

MetaPAD Framework

Integrated Data-Driven Text Mining

↓

Meta Pattern Mining

↓

Attribute Extraction from Meta Patterns

Quality Meta-Pattern Classifier

- **Frequency**
  - *"prime_minister $PrimeMinister" vs "young $PrimeMinister"*
- **Completeness**
  - *"$Country prime_minister $PrimeMinister" vs "$Country prime_minister"*
- **Informativeness**
  - *"$Person 's brother , $Person ," vs "$Person and $Person"*
- **Coverage**
  - *"$Person 's signature healthcare law": only "Barack Obama"*
- **Classifier: Random forest**

MetaPAD Framework

Integrated Data-Driven Text Mining

↓

Meta Pattern Mining

↓

Attribute Extraction from Meta Patterns

…xxx **$Country** Prime_Minister **$PrimeMinister** xxx…
…xxx **$President** , **$Digit** , xxx…

Quality Meta-Pattern Classifier

$Location Prime_Minister **$Person**
$Person, **$Digit** ,

$Country Prime_Minister **$PrimeMinister**
$President , **$Digit** ,

Synonym Meta-Pattern Detection
(1) Shared instances (2) J.W. similar words

$Location Prime_Minister **$Person**
$Location PM **$Person**
Prime_Minister **$Person** of **$Location**

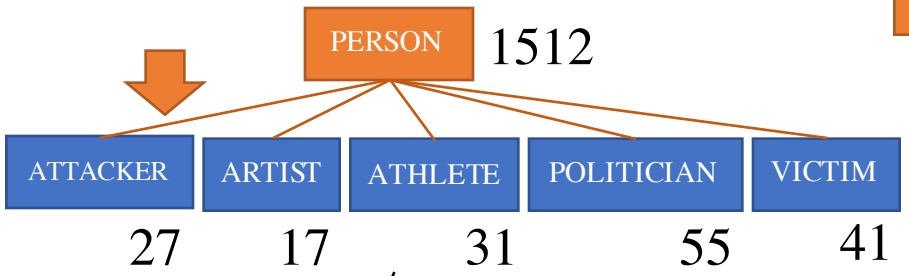
$Person , **$Digit** ,
$Person , a **$Digit** -year-old
$Person , age **$Digit**

Re-typing for Appropriate Granularity

$Country Prime_Minister **$PrimeMinister**
$Person , **$Digit** ,

Top-Down Re-Typing for Granularity

_____, a $Digit -year-old

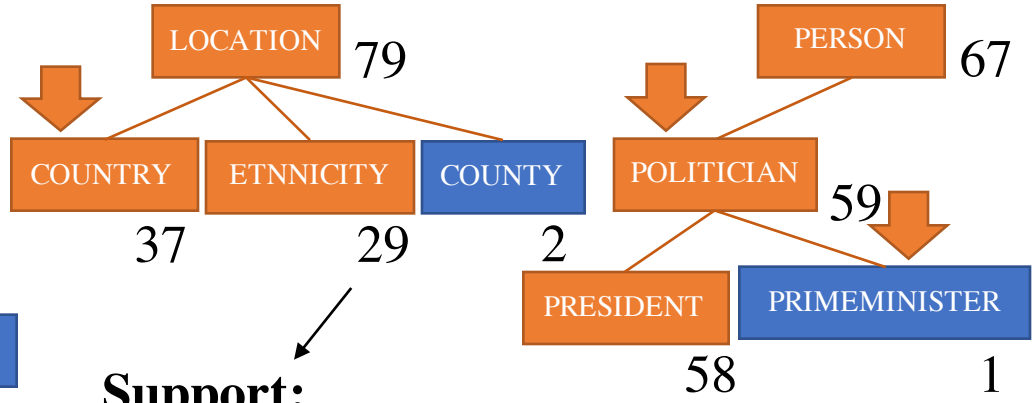


Graininess:

α=(27+17+…+41)/1512

small (< 0.8), stop going down

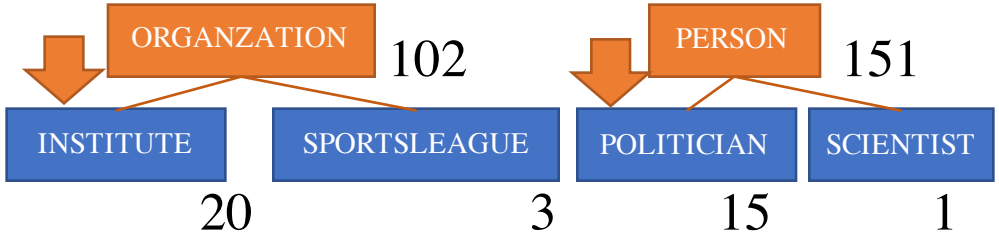
_____ President _____



Support:

β=29/max(37,29,2)

big (> 0.1), keep $Ethnicity



Similar for Bottom-Up…

Experimental Results

Class=$PERSON (METAPAD: 10,361 names, 4,839 pairs)		
Name:BIPERPEDIA	**(Name, -)**	**(Name, Value Type)**
Mr.	Dr.	(-year-old,$DIGIT)
Dr.	Mr.	(president,$ORGANIZATION)
president	president	(spokesman,$ORGANIZATION)
wife	director	(director,$ORGANIZATION)
-year-old	spokesman	(wife,$PERSON)
death	chief	(chairman,$ORGANIZATION)
coach	professor	(governor,$USSTATE)
love	head	(spokeswoman,$ORGANIZATION)
son	coach	(leader,$ORGANIZATION)
...	...	...
code case homicide	staff sergeant	(told reporters,$WEEKDAY)
snow pants	army chief	(board member,$ORGANIZATION)
fellow director	basketball coach	(hack,$COMPANY)

Class=$COUNTRY (METAPAD: 1,132 names, 3,930 pairs)		
Name:BIPERPEDIA	**(Name, -)**	**(Name, Value Type)**
president	president	(ambassador,$COUNTRY)
people	government	(president,$PRESIDENT)
government	war	(visit,$PERSON)
capital	border	(dead,$DIGIT)
visit	volcano	(prime minister,$PRIMEMINISTER)
economy	sanctions	(senator,$SENATOR)
prime minister	ambassador	(embassy,$COUNTRY)
part	earthquake	(condemn,$ORGANIZATION)
leaders	capital	(district judge,$PERSON)
...	...	...
nuclear dossier	volcano eruption	(protests,$NEWSAGENCY)
similar box	security	(-magnitude earthquake,$DIGIT)
episcopal oversight	parliament	(second biggest,$ORGANIZATION)

Class=$INSTITUTE (METAPAD: 402 names, 198 pairs)		
Name:BIPERPEDIA	**(Name, -)**	**(Name, Value Type)**
professor	professor	(professor,$PERSON)
students	students	(law professor,$PERSON)
president	graduate	(political science professor,$PERSON)
campus	law professor	(student,$PERSON)
law professor	campus	(grad,$PERSON)
graduate	degree	(signee,$PERSON)
director	dean	(economics professor,$PERSON)
study	faculty	(basketball coach,$PERSON)
researchers	expert	(finance professor,$PERSON)
...	...	...
foul	commitment	(class,$YEAR)
socialism speech	dorm	(superintendent,$PERSON)
good summary	program	(-year-old student,$DIGIT $PERSON)

Class=$BASKETBALLPLAYER (METAPAD: 58 names, 40 pairs)		
Name:BIPERPEDIA	**(Name, -)**	**(Name, Value Type)**
guard	forward	(points,$DIGIT)
star	points guard	(center,$TEAMNAME)
game	game	(freshman,$SPORTSLEAGUE)
forward	freshman	(forward,$TEAMNAME)
career	center	(point guard,$TEAMNAME)
teammate	get better	(all-star,$SPORTSLEAGUE)
point guard	basketball player	(games,$DIGIT)
points	full highlights	(rebounds,$DIGIT)
season	jumper	(ast,$DIGIT)
...	...	...
understudy	retirement	(PG,$TEAMNAME)
birthday boy	shoes	(career earnings,$DIGIT $DIGITUNIT)
injury meme	suspended without pay	(sue,$PERSON)

Experimental Results

Class=$Location; Value Type=$Month,$Day,$Year		
#	Meta Patterns	
1	$Location $Month $Day, $Year	
2	$Country, $Weekday, $Month $Day, $Year	
3	$Location on $Month $Day, $Year	
#	Entity	Attribute Value
1	Pearl Harbor	December 7, 1941
2	Green Bay	Sunday, Jan 11, 2015
3	Malta[1]	Friday, Nov 27, 2015
...	...	...
5862	Beijing[2]	October 11, 2013
5863	Finland[3]	April 8, 2015

Class=$Organization; Name="ceo"		
#	Meta Patterns	
1	$Organization CEO $Person	
2	$Company CEO $Businessperson	
3	$Organization's $Person	
#	Entity	Attribute Value
1	Apple	Tim Cook
2	Facebook	Mark Zuckerberg
3	Hewlett-Packard	Carly Fiorina
...	...	...
765	Boston Medical Center	Kate Walsh
766	Association of Private Sector Colleges and Universities	Steve Gunderson

Class=$Person; Name="-year-old"[7]		
#	Meta Patterns	
1	$Digit-year-old $Person	
2	$Person, $Digit,	
3	$Person, a $Digit-year-old	
#	Entity	Attribute Value
1	Tamir Rice	12
2	Bobbi Kristina Brown	21
3	Michael Brown	18
...	...	...
4993	Jay Nixon	58
4994	Xanana Gusmao	68

Class=$Person; Name="president"; Value Type=$Ethnicity		
#	Meta Patterns	
1	$Ethnicity President $President	
2	$Ethnicity leader $President	
3	$Ethnicity government of President $President	
#	Entity	Attribute Value
1	Vladimir Putin	Russian
2	Francois Hollande	French
3	Raul Castro	Cuban
...	...	...
254	Mohammed Morsi	Egyptian
255	Klaus Iohannis	Romanian

[1]Commonwealth Heads of Government Meeting. [2]UCI World Tour of Beijing. [3]Finnish parliamentary election begins.

Experimental Results

F1 score	WPB ('10, 100M)	CNA ('97-'10, 200M)	APR ('15, 200M)	TWT ('15, 1GB)
Total (vs Biperpedia -q)	↑67.7%	↑48.3%	↑189.5%	↑208.0%
w/ Meta pattern classifier	↑30.1%	↑27.0%	↑127.1%	↑195.6%
w/ Granularity	↑20.8%	↑15.6%	↑17.3%	↑3.1%
w/ Integrated text mining techs	↑13.8%	↑9.3%	↑13.0%	↑0.8%

$Cardiovasular_Diseases due to $Bacteria
$Cardiovasular_Diseases caused by $Bacteria

$Bacteria	$Cardiovascular_Diseases
Streptococcus pneumoniae	Endocarditis
Neisseria meningitidis	Pericarditis
Haemophilus paraphrophilus	Endocarditis
Proteus	Endocarditis
Listeria monocytogenes	Pericarditis
Corynebacterium	Endocarditis
Actinomyces	Endocarditis
Coxiella	Endocarditis
Pasteurella pneumotropica	Endocarditis
Cardiobacterium	Endocarditis

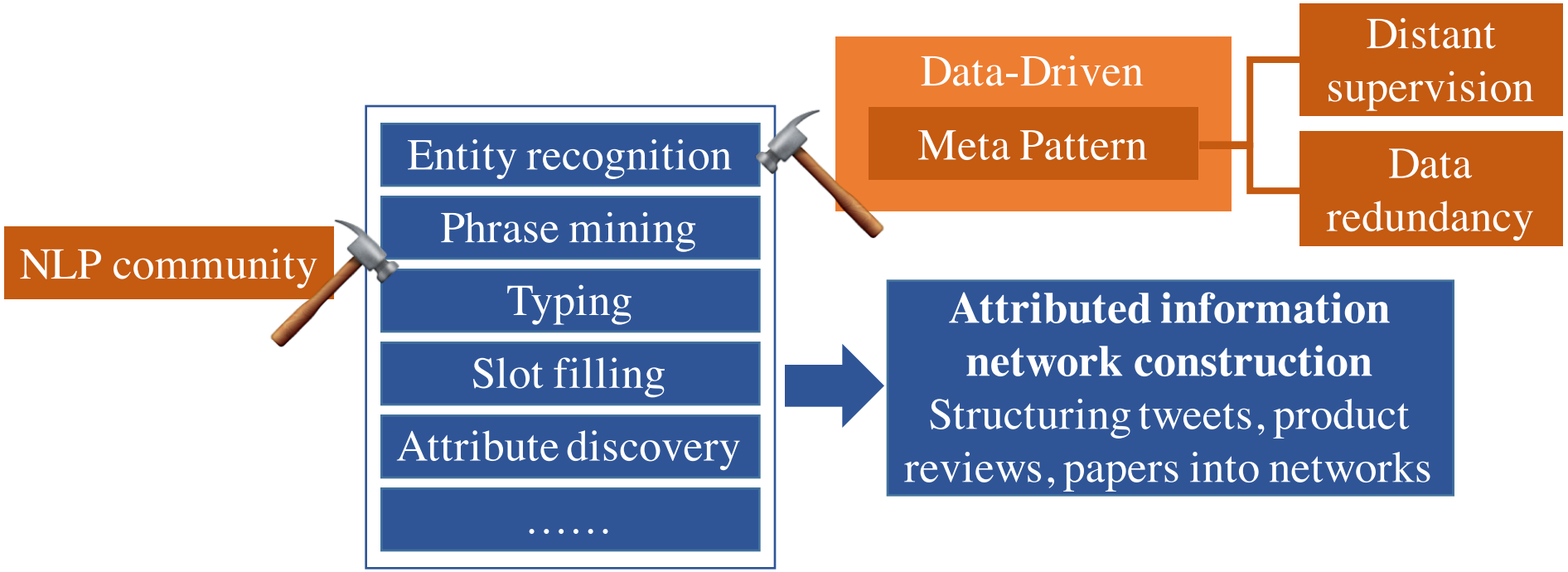
$Enzymes_and_Coenzymes inhibitor $Chemical

$Chemical	$Enzymes_and_Coenzymes
chelerythrine	protein kinase C
fondaparinux	Factor Xa
calphostin C	protein kinase C
bisindolylmaleimide	protein kinase C

$Diagnosis : $Digit +/- $Digit kg/m ($Digit)

$Diagnosis	$Digit $Digit $Digit
BMI	(31.0 , 6.4 , 2)
BMI	(26 , 4 , 2)
body mass index	(27 , 6 , 2)

Meta Pattern: Data-Driven Approaches for NLP Tasks

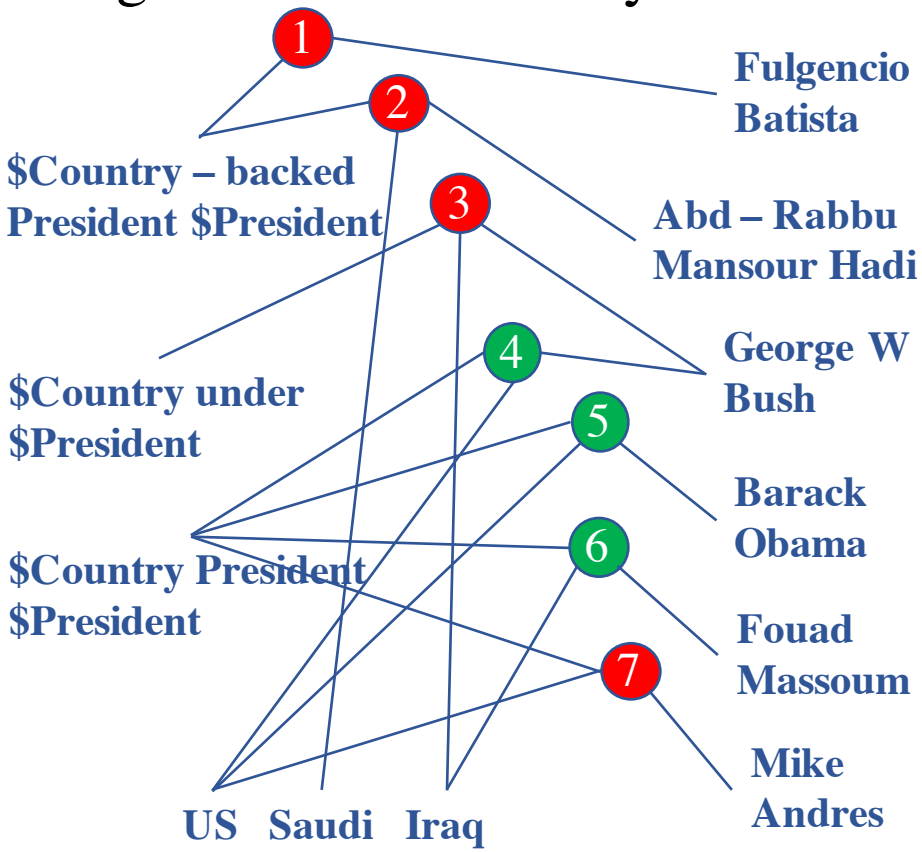


Finding Truth when Structuring

- Find trustworthy facts by modeling "source" reliability

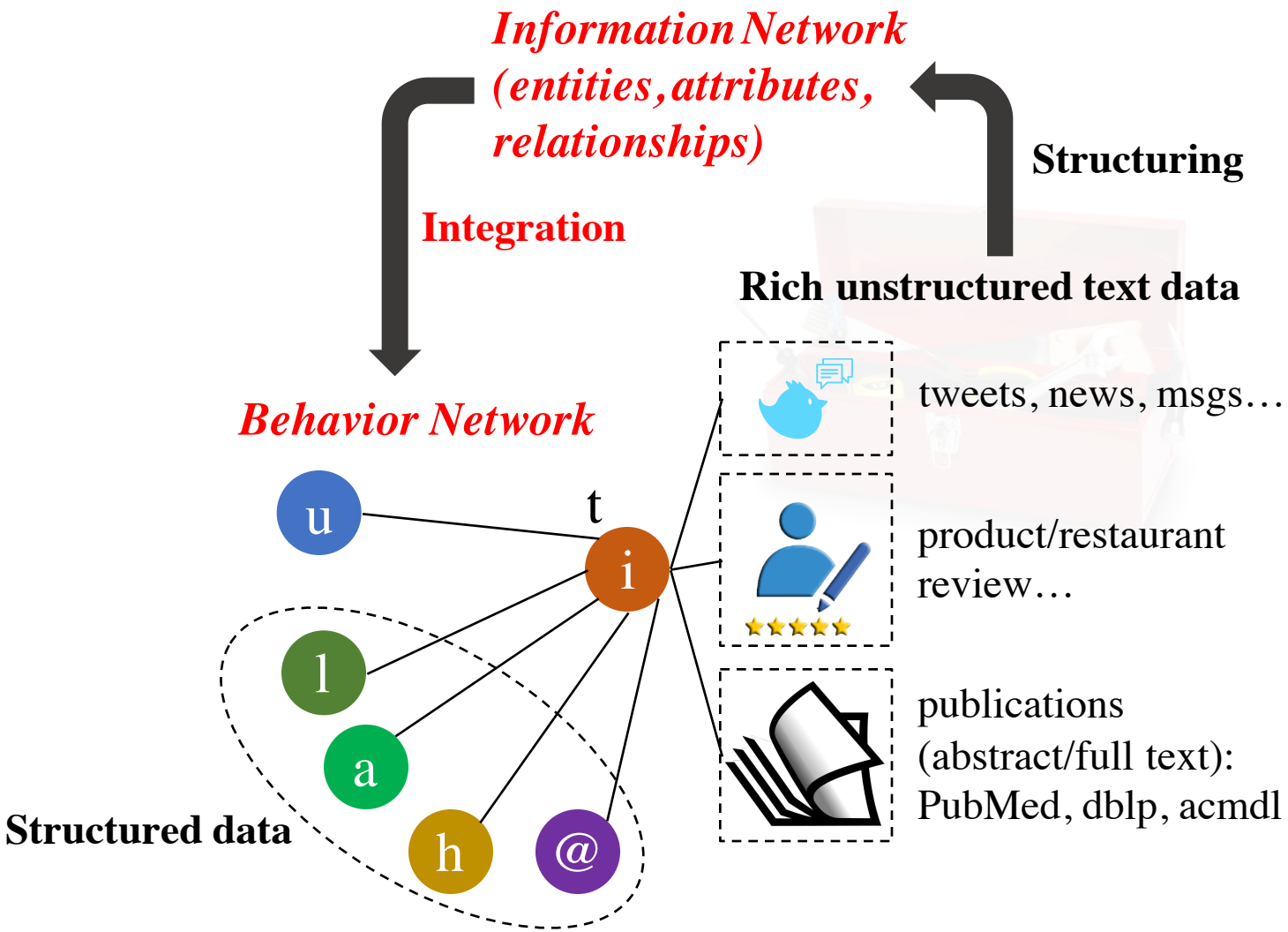
($Country, president, $President) : 0.829

Meta Pattern	Acc. (FP/P)
$Country 's President $President	0.984 (1/61)
President $President of $Country	1.000 (0/24)
$Country 's President $President ,	1.000 (0/16)
" $Country President $President	1.000 (0/7)
…	…
President $President said $Country	0.833 (1/6)
$Country President $President	0.807 (16/83)
$Country , President $President	0.650 (7/20)
$Country - backed President $President	0.500 (3/6)
$Country under $President	0.500 (1/2)



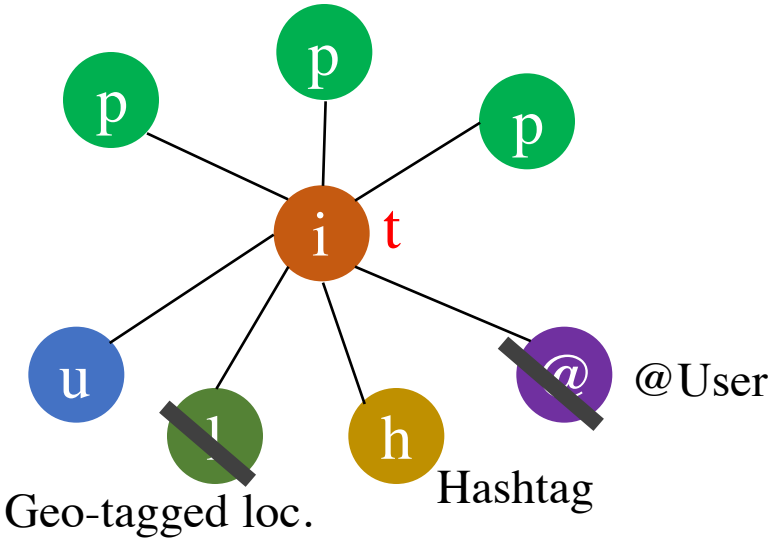
1. …*Fidel Castro and his brother Raul led winning a revolution toppling* ***US - backed President Fulgencio Batista*** .
2. …*control of the country and at reinstating* ***Saudi - backed President Abd - Rabbu Mansour Hadi*** .
3. …*was profoundly forward - leaning and outspoken about the importance of invading* ***Iraq under George W Bush*** .
7. …*better delivering on those expectations , " McDonald 's* ***US President Mike Andres*** *said in the announcement* .

Data to Network to Knowledge



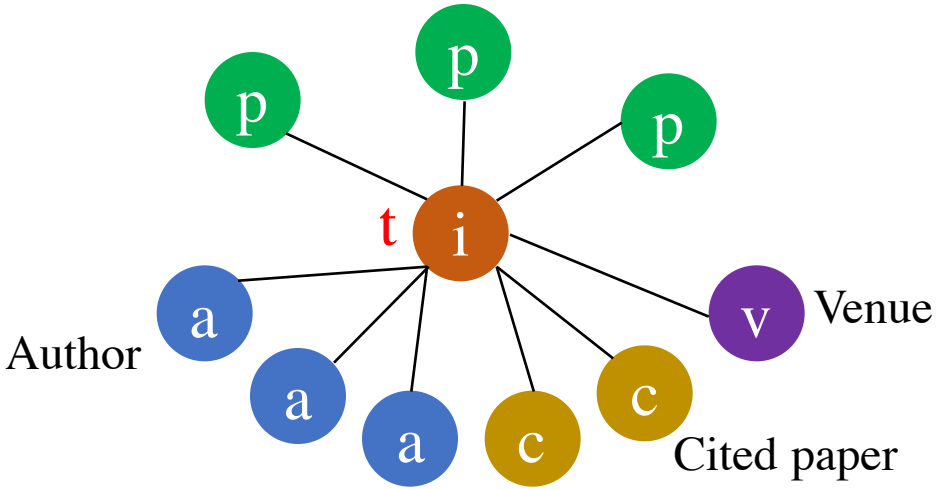
Bring Phrases to Behavior Modeling

- Tweeting behavior
  - Event **summary**



20:03:09 @ebekahwsm
this better be the best halftime show ever in the history of halftimes shows. ever. #SuperBowl

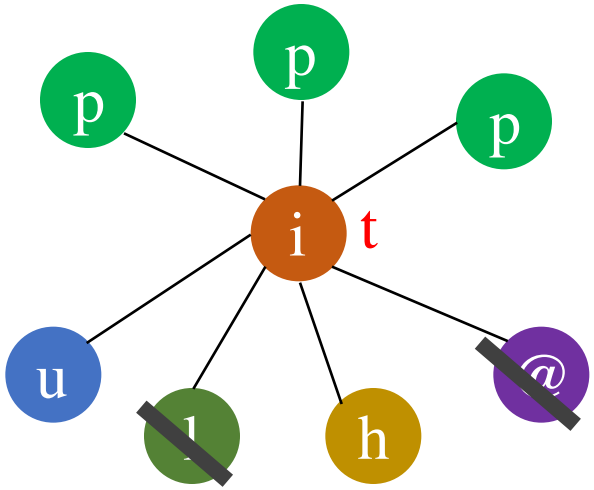
- Paper-publishing behavior
  - Research trend **summary**



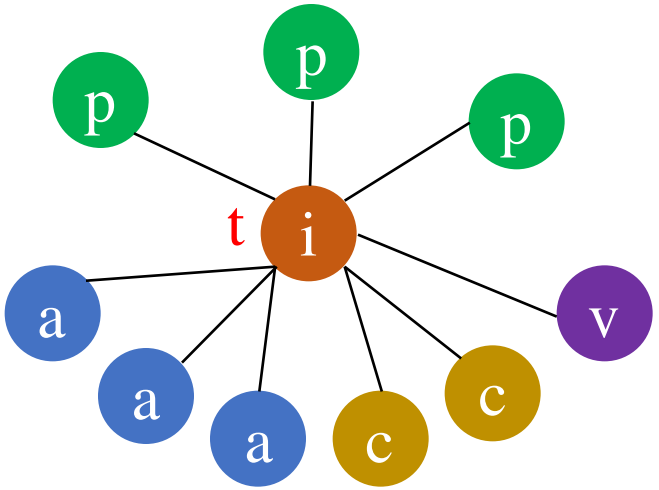
2009 P. Melville, W. Gryc, R. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification", KDD'09. Refs: p81623, p84395…

Tensor Fails

- Tweeting behavior
  - Event **summary**



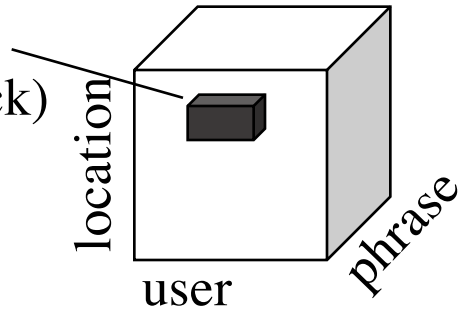
- Paper-publishing behavior
  - Research trend **summary**



Q: How to represent and summarize **dynamic multi-contextual** behaviors?

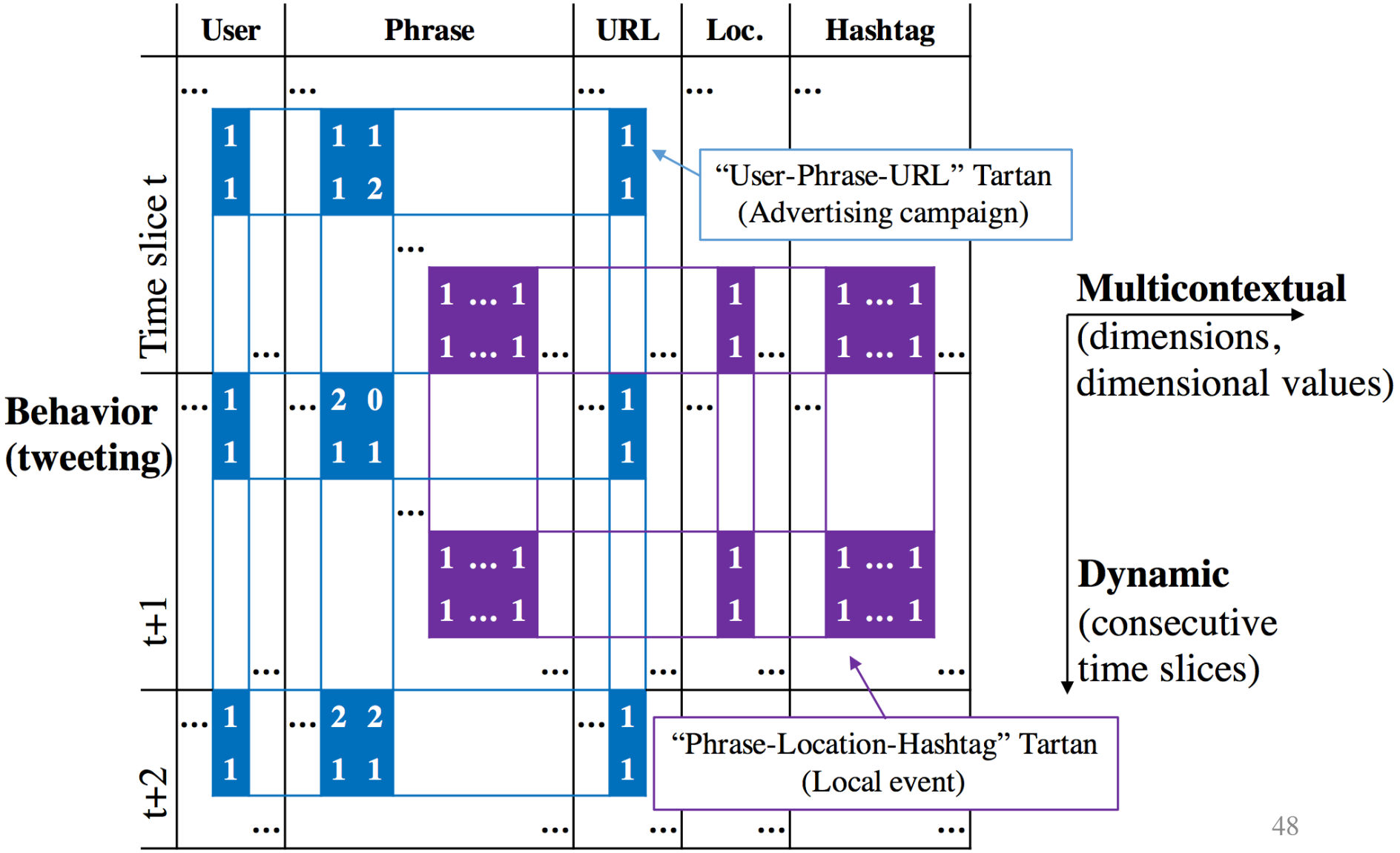
A set of values in dimensions (*one-guaranteed value*, *empty value*, *multi-values*)

summary
(dense block)

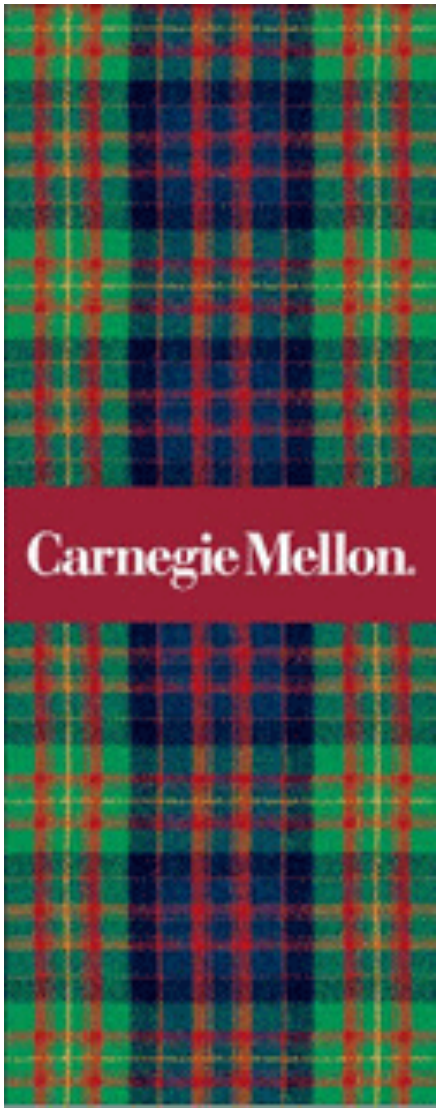


✔ ✘ ✘

Two-Level Matrix and "Tartan"



CMU Tartans


Carnegie Mellon.

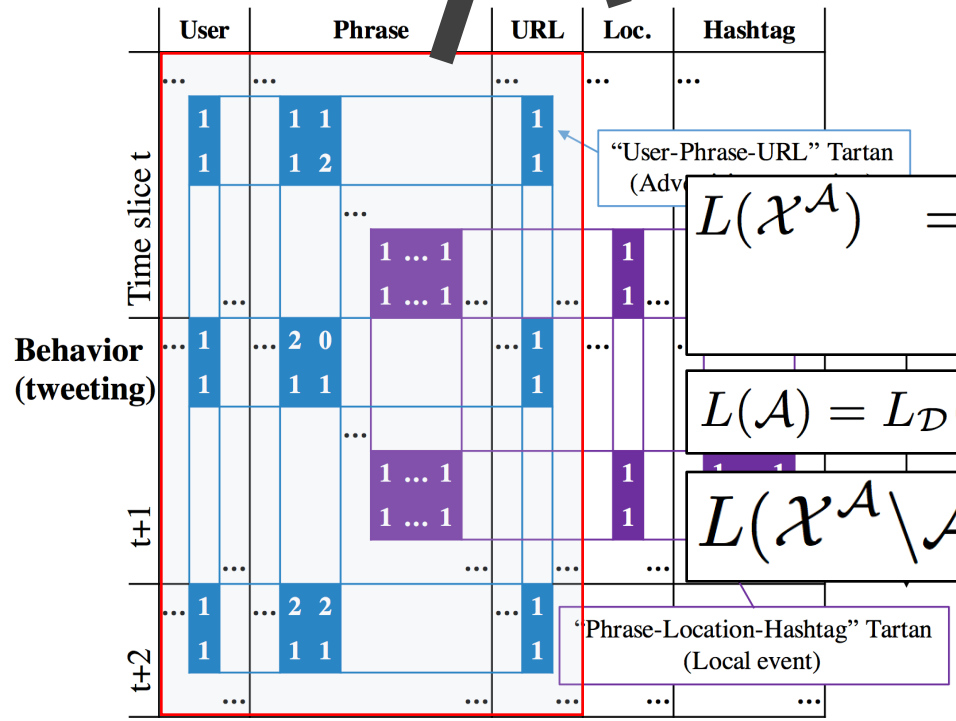

GO TARTANS!

CARNEGIE MELLON UNIVERSITY
PITTSBURGH PENNSYLVANIA 1900

Optimize with MDL Principle

- *Maximize* the number of bits by encoding the Tartan

$$f(\mathcal{A}, \mathcal{X}) = L(\mathcal{X}^{\mathcal{A}}) - L(\mathcal{A}) - L(\mathcal{X}^{\mathcal{A}} \backslash \mathcal{A}).$$

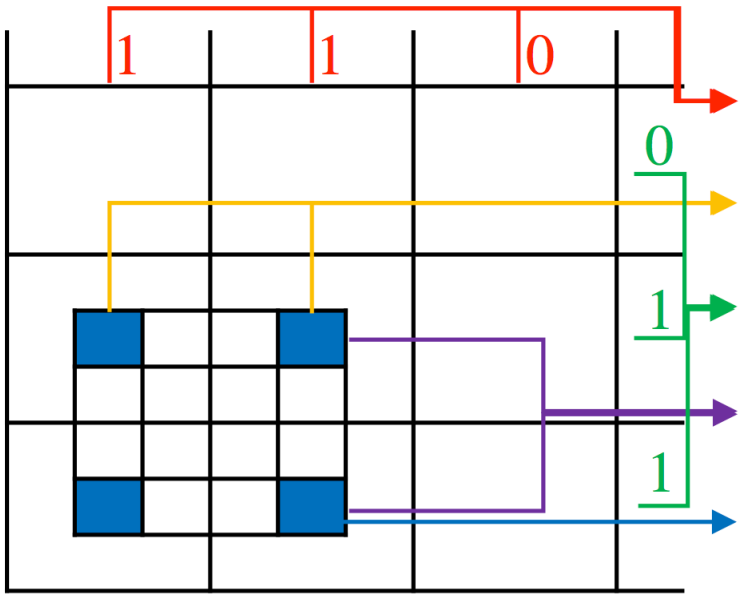


$$L(\mathcal{X}^{\mathcal{A}}) = g(V + C, C) + L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) + \sum_{d \in \mathcal{D}} \log^* N_d + \sum_{t \in \mathcal{T}} \log^* E^{(t)}.$$

$$L(\mathcal{A}) = L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{V}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) + L_{\mathcal{B}}(\mathcal{A}) + L_{\mathcal{A}}(\mathcal{A}).$$

$$L(\mathcal{X}^{\mathcal{A}} \backslash \mathcal{A}) = g(V + C - v - c, C - c);$$

Encoding Tartan: Dimensions



Encoding the dimensions

Encoding the dimensional values

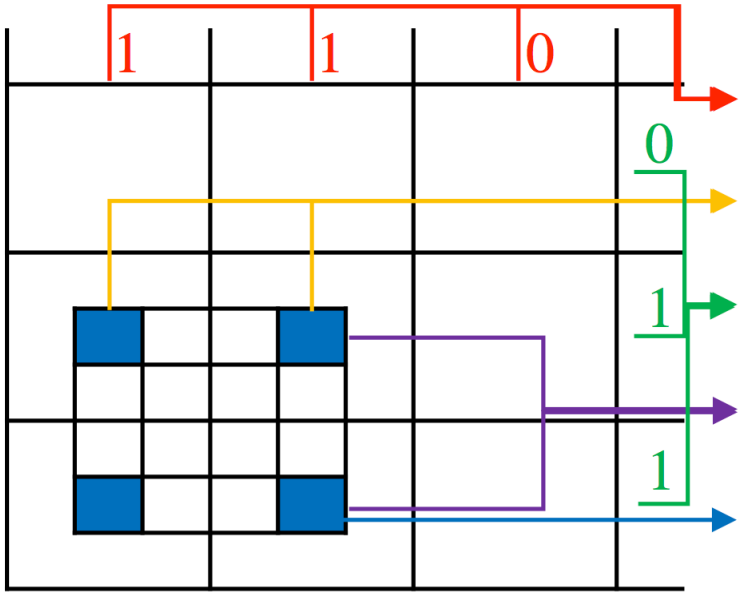
Encoding the time slices

Encoding the behaviors

Encoding the entries in the Tartan

$$H_{\mathcal{D}}(X) = -\sum_{x\in\{0,1\}} P(X=x)\log P(X=x)$$
$$= -\left(\frac{D^{\mathcal{A}}}{D}\log\frac{D^{\mathcal{A}}}{D} + \frac{D-D^{\mathcal{A}}}{D}\log\frac{D-D^{\mathcal{A}}}{D}\right).$$
$$L_{\mathcal{D}}(\mathcal{A}) = \log^* D + \log^* D^{\mathcal{A}} + D\cdot H_D(X)$$
$$= \log^* D + \log^* D^{\mathcal{A}} + g(D, D^{\mathcal{A}}),$$

Encoding Tartan: Dimensional Values



Encoding the dimensions

Encoding the dimensional values

Encoding the time slices

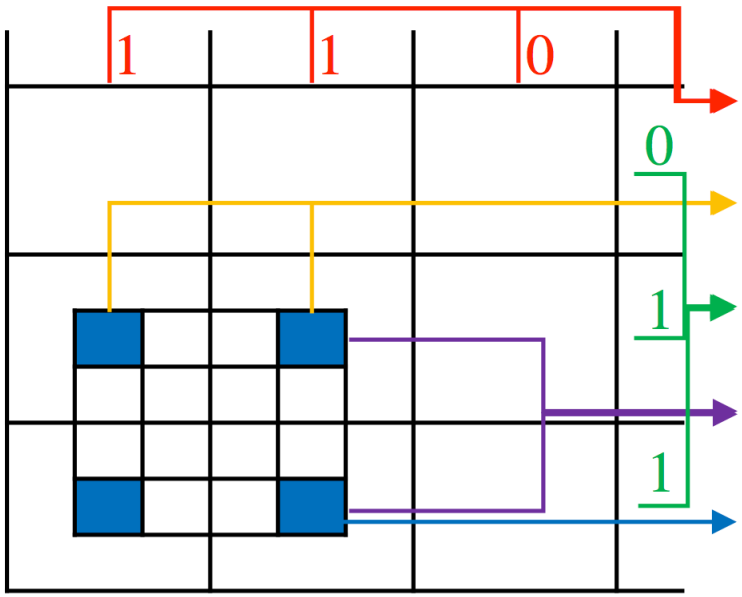
Encoding the behaviors

Encoding the entries in the Tartan

$$H_{\mathcal{V}_d}(X) = -\left(\frac{n_d}{N_d}\log\frac{n_d}{N_d} + \frac{N_d - n_d}{N_d}\log\frac{N_d - n_d}{N_d}\right).$$

$$L_{\mathcal{V}}(\mathcal{A}) = \sum_{d\in\mathcal{D}}\left(\log^* N_d + \log^* n_d + g(N_d, n_d)\right).$$

Encoding Tartan: Time Slices



Encoding the dimensions

Encoding the dimensional values

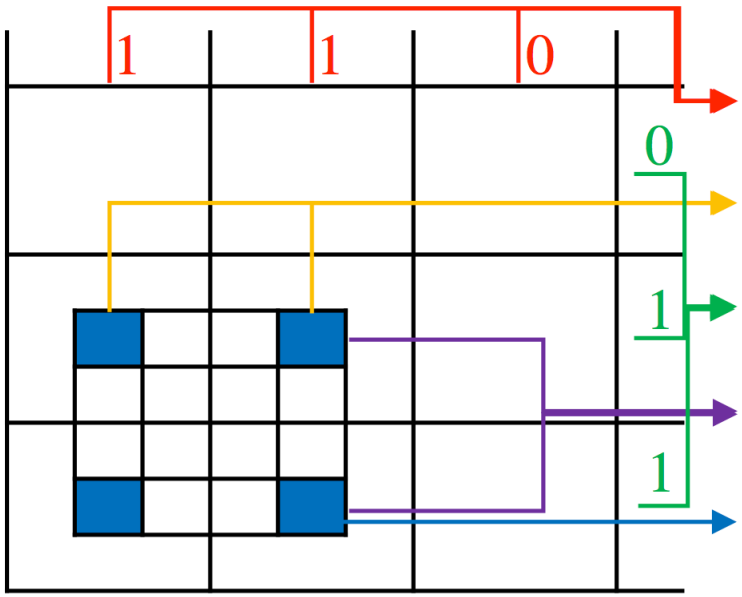
Encoding the time slices

Encoding the behaviors

Encoding the entries in the Tartan

$$L_{\mathcal{T}}(\mathcal{A}) = \log^* T + \log^* T^{\mathcal{A}} + \log^* t_{start}$$

Encoding Tartan: Behaviors



Encoding the dimensions

Encoding the dimensional values

Encoding the time slices

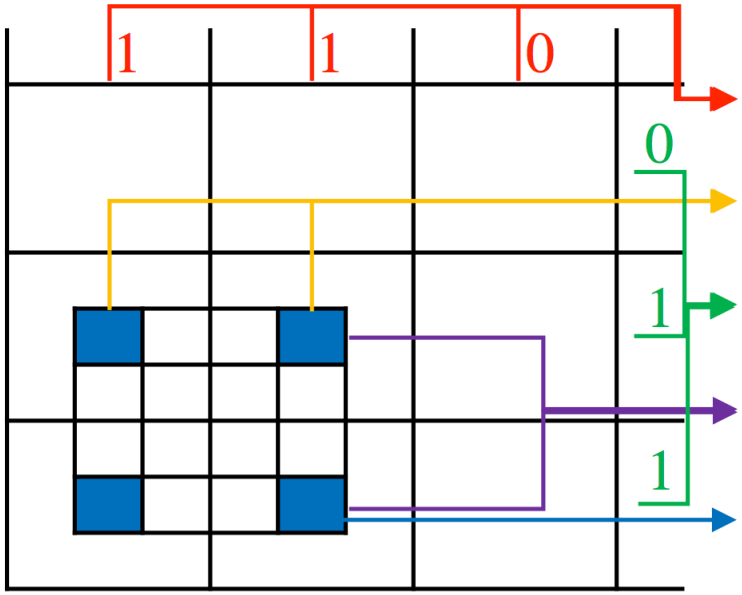
Encoding the behaviors

Encoding the entries in the Tartan

$$H_{\mathcal{B}^{(t)}}(X) = -\left(\frac{e^{(t)}}{E^{(t)}} \log \frac{e^{(t)}}{E^{(t)}} + \frac{E^{(t)}-e^{(t)}}{E^{(t)}} \log \frac{E^{(t)}-e^{(t)}}{E^{(t)}}\right).$$

$$L_{\mathcal{B}}(\mathcal{A}) = \sum_{t \in \mathcal{T}} \left(\log^* E^{(t)} + \log^* e^{(t)} + g(E^{(t)}, e^{(t)})\right).$$

Encoding Tartan: Entries



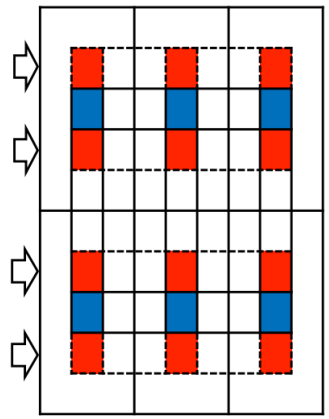
$$v = \left(\sum_{d\in\mathcal{D}} n_d\right)\left(\sum_{t\in\mathcal{T}} e^{(t)}\right).$$

$$c = \sum_{d\in\mathcal{D},t\in\mathcal{T}} \sum_{b\in\mathcal{B}^{(t)},i\in\mathcal{V}_d} \mathcal{X}_d^{(t)}(b,i).$$

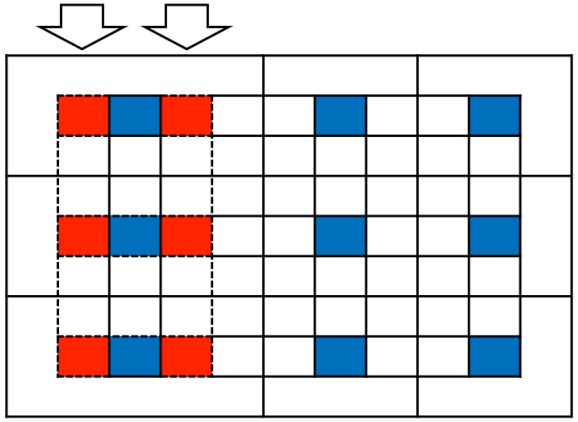
$$H_{\mathcal{A}}(X) = -\left(\frac{c}{v+c}\log\frac{c}{v+c} + \frac{v}{v+c}\log\frac{v}{v+c}\right).$$

$$L_{\mathcal{A}}(\mathcal{A}) = (v+c)H_{\mathcal{A}}(X) = g(v+c,c).$$

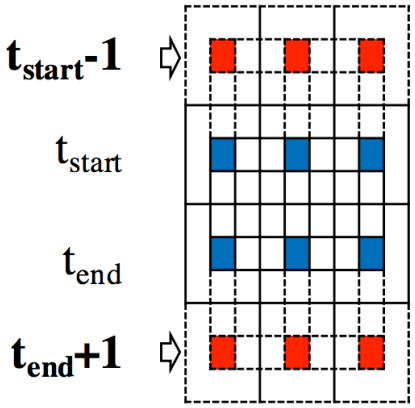
Greedy Search for the Local Optimum



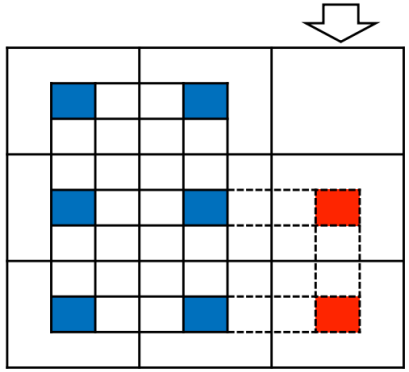
(a) Update the set of behaviors.



(b) Update the set of values.



(c) Update the consecutive time slices.



(d) Update the set of dimensions.

Time complexity:

$$\mathcal{O}(\sum_d N_d \log N_d + \sum_t E^{(t)} \log E^{(t)})$$

Experimental Results

- DM/ML research trend summaries with DBLP data

1997 2000 2003 2006 2009 2012

Author	Venue	Keyword	Cited	#Paper
76 Cheng-xiang Zhai Hui Fang S. Kambhampati	**7** SIGIR VLDB TKDE	**7** "information retrieval" "data integration" "text classification"	**68** p56743[1] p62995 p76869	**32** 2003- 2007

[1] "A language modeling approach to information retrieval"

Venue	Keyword	#Paper
5 ICML NIPS …	**6** "reinforcement learning" "machine learning"	**40** 1997- 2002

Author	Venue	Cited	#Paper
6 Jiawei Han Xifeng Yan	**1** SIG- MOD	**1** p76095[2]	**22** 2004- 2010

[2] "Frequent subgraph discovery"

Venue	Keyword	#Paper
3 ICDM AAAI TKDE	**1** "anomaly detection"	**25** 2005- 2013

Author	Venue	Keyword	#Paper
27 C. Faloutsos J. Pei P. S. Yu X. Lin C. Aggarwal…	**6** KDD ICDM ICDE TKDE …	**12** "large graphs" "data streams" "evolving data" "evolving graphs" …	**70** 2006- 2013

Author	Venue	Keyword	Cited	#Paper
12 Ryen White Hang Li Tie-Yan Liu Zhaohui Zheng…	**5** SIGIR WWW WSDM CIKM…	**3** "web search" "click-through data" "sponsored search"	**12** p82630[3] p116290 p103899 p106191…	**32** 2006- 2013

[3] "Optimizing search engines using clickthrough data"

Author	Venue	Keyword	#Paper
8 Qiang Yang Dou Shen Sinno Pan…	**3** KDD PAKDD AAAI	**6** "transfer learning" "data mining" "localization models"	**17** 2007- 2010

Experimental Results

- Event summaries with Super Bowl 2013 tweets

16:30 – 17:00

16:30:31 My prediction Ravens 34 Niners 31
16:30:57 Ready for the big game :D, my prediction 24-20 SF #SuperBowl
16:31:14 My prediction for superbowl.. 48.. Jets over Bears 17-13 Mark Sanchez MVP
16:32:24 I predict Baltimore Ravens will win 27 to 24 or 25 or 26. Basically it will be a close game.

"my prediction"

user	phrase	hashtag	URL	3,397 tweets
(3,325)	226	(0)	(0)	

Tartan #1: (1 dim) 16:30-17:30

17:30 – 18:00

17:30:51 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist
17:31:01 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist
17:31:16 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist
17:31:19 RT @LMAOTWITPICS: Make Your Prediction. Retweet For 49ers http://t.co/KKksEist

"make your prediction"

user	phrase	RT @user	URL	196 tweets
(196)	4	1	1	

Tartan #2: (3 dims) 17:00-18:00

18:00 – 18:30

18:55:03 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47
18:55:04 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47
18:55:44 RT @Ravens: David Akers is good from 36 yards to make the score 7-3 Ravens. Nice job by the defense to tighten up in the red zone.

"7-3", "1st Qtr"

user	phrase	RT @user	URL	215 tweets
(213)	21	3	(0)	

Tartan #3: (2 dims) 18:30-19:30

19:00 – 19:30

20:20:01 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6
20:20:02 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs
20:20:04 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6
20:20:05 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs

"halftime show"

user	phrase	RT @user	URL	617 tweets
(617)	11	4	4	

Tartan #4: (3 dims) 20:00-21:00

20:00 – 20:30

20:20:47 (Manhattan, NY)..and every one of those girls took #ballet #Beyonce #superbowl
20:22:01 (New York, NY) I have the biggest lady boner for Beyonce #BeyonceBowl #DestinyBowl #DestinysChild #SuperBowl
20:24:32 (Manhattan, NY) No one can ever top that performance by Beyonce. EVER. #Beyonce #superbowl #halftimeshow

"beyonce", #beyonce, #superbowl, #DestinysChild

location	phrase	hashtag	URL	166 tweets
2	55	17	(0)	

Tartan #5: (3 dims) 20:00-21:00

21:00 – 21:30

21:44:42 Ahora si pff #49ers 23-28 #Ravens
21:44:44 Baltimore #Ravens 28-23 San Francisco #49ers
21:44:50 FG Akers #49ers 23-28 #Ravens 3Q 3:10 #SuperBolwXLVII #SuperBowl #NFL

"28-23", #49ers, #Ravens

user	phrase	hashtag	URL	653 tweets
(650)	69	11	(0)	

Tartan #6: (2 dims) 21:00-22:00

22:00

22:42:27 Congratulations Ravens!!!!
22:42:43 Congratulations Ray Lewis and the Ravens.
22:42:43 Game over.! Ravens won ray got his retirement ring now all y'all boys and girls go to sleep .!
22:42:52 "@LetThatBoyTweet: Game over. Ravens win the Super Bowl."

"congratulations", "game over"

user	phrase	hashtag	URL	1,950 tweets
(1942)	248	(0)	(0)	

Tartan #7: (1 dim) 22:00-23:30

Summary

- Structuring text into heterogeneous information networks
- **Observations, Representations, Models**
  - **ToPMine/SegPhrase:** Quality phrase mining
  - **ClusType:** Entity recognition and typing
  - **MetaPAD:** Data-driven automatic attribute discovery for attributed network construction
    - Integrating text mining techniques
    - **Meta Pattern** Mining
- Integrating phrases into behavioral analysis
- **Observations, Representations, Models**
  - **CatchTartan:** Dynamic multicontextual. Tensor fails.

Acknowledgement

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

References

D. Blei, A. Ng, and M. Jordan. "Latent dirichlet allocation." JMLR, 2003.

J. Herlocker, J. Konstan, L. Terveen, J. Riedl. "Evaluating collaborative filtering recommender systems." ACM TOIS, 2004.

Y. Koren, R. Bell, C. Volinsky. "Matrix factorization techniques for recommender systems." Computer, 2009.

Y. Koren. "Factorization meets the neighborhood: A multifaceted collaborative filtering model." KDD, 2008.

Y. Koren. "Collaborative filtering with temporal dynamics." CACM, 2010.

M. Balabanovic and Y. Shoham. "FAB: Content-based, collaborative recommendation." CACM, 1997.

N. Liu and Q. Yang. "Eigenrank: A ranking-oriented approach to collaborative filtering." SIGIR, 2008.

N. Liu, M. Zhao, and Q. Yang. "Probabilistic latent preference analysis for collaborative filtering." CIKM, 2009.

References

H. Ma, H. Yang, M. Lyu, and I. King. "Sorec: Social recommendation using probabilistic matrix factorization." CIKM, 2008.

H. Ma, T. Zhou, M. Lyu, and I. King. "Improving recommender systems by incorporating social contextual information." ACM TOIS, 2011.

H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King. "Recommender systems with social regularization." WSDM, 2011.

J. Leskovec, A. Singh, and J. Kleinberg. "Patterns of influence in a recommendation network." PAKDD, 2006.

P. Massa and A. Paolo. "Trust-aware recommender systems." RecSys, 2007.

M. Jamali and E. Martin. "TrustWalker: A random walk model for combining trust-based and item-based recommendation." KDD, 2009.

H. Ma, I. King, and M. Lyu. "Learning to recommend with social trust ensemble." SIGIR, 2009.

H. Ma. I. King, and M. Lyu. "Learning to recommend with explicit and implicit social relations." ACM TIST, 2011.

References

M. Faloutsos, P. Faloutsos, and C. Faloutsos. "On power-law relationships of the internet topology." SIGCOMM, 1999.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Weiner. "Graph structure in the web." Computer Networks, 2000.

F. Chung and L. Lu. "The average distances in random graphs with given expected degrees." PNAS, 2002.

J. Kleinberg. "Authoritative sources in a hyperlinked environment." JACM, 1999.

H. Kwak, C. Lee, H. Park, and S. Moon. "What is Twitter, a social network or a news media?" WWW, 2010.

B. Hooi, H.A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos. "Fraudar: Bounding graph fraud in the face of camouflage." KDD, 2016.

C. Aggarwal and J. Han. "Frequent pattern mining." Springer, 2014.

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

References

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.

Y. Sun, R. Barber, M. Gupta, C. Aggarwar, and J. Han. "Co-author relationship prediction in heterogeneous bibliographic networks." ASONAM, 2011.

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

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

References

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.

E.E. Papalexakis, C. Faloutsos, N.D. Sidiropoulos. "ParCube: Sparse parallelizable tensor decompositions." PKDD, 2012.

D. Koutra, U. Kang, J. Vreeken, and C. Faloutsos. "VOG: Summarizing and understanding large graphs." SDM, 2014.

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.

References

Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation." SIGMOD, 2014.

Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. "A confidence-aware approach for truth discovery on long-tail data." VLDB, 2014.

F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation." KDD, 2015.

Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. "A survey on truth discovery." KDD Explorations Newsletter, 2016.

S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. "Modeling truth existence in truth discovery." KDD, 2015.

S. Kumar, R. West, and J. Leskovec. "Disinformation on the Web: Impact, characteristics, and detection of Wikipedia hoaxes." WWW, 2016.

S. Kumar, F. Spezzano, and V.S. Subrahmanian. "Identifying malicious actors on social media." ASONAM, 2016. (tutorial)

Thank you!

Data-Driven Behavioral Analytics: Observations, Representations and Models