Shallow Information Extraction for the Knowledge Web

Denilson Barbosa, Univ. of Alberta, denilson@ualberta.ca
Haixun Wang, MSR Asia, haixunw@microsoft.com
Cong Yu, Google Research NYC, congyu@google.com

Outline

- Motivation
  - Why doing all this in the first place?
  - Define what shallow means – no deep linguistic analysis
  - Emphasizing why the need for shallow extraction techniques
- PART I: shallow extraction techniques
  - Entity extraction
  - Relation extraction
  - Application Social text mining
- Part II: Bring Knowledge to Search
- Part III: Real life knowledge base, scalability and probability

Deep vs shallow NLP for Information Extraction

- DISCLAIMER – Natural Language Processing is a broad field, with many more applications than discussed here
- Broadly speaking, Information Extraction (IE) is concerned with finding entities and facts/relations about these entities in text

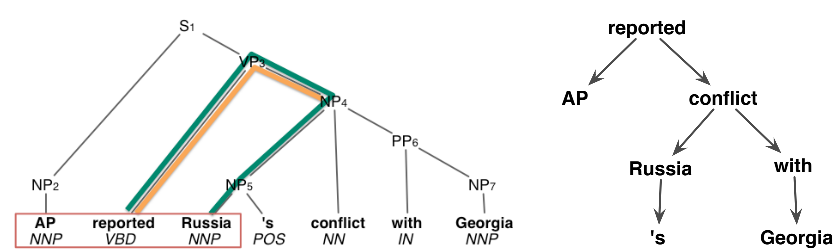
Brisbane travel guide - Wikitravel
wikitravel.org/en/Brisbane
Get in · Do · Buy · Eat · Sleep
Brisbane is the capital of the state of Queensland. It has a population of about 2 million people, making it the third-largest city in Australia.

Entities:
Brisbane, Queensland, Australia

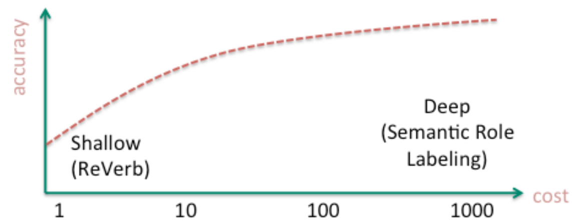
Relations:
city(Brisbane)
state(Queensland)
capitalOf(Brisbane, Queensland)
populationOf(Brisbane, 2M)
cityIn(Brisbane, Australia)
…

Deep vs shallow NLP (for Information Extraction)

- Deep == sophisticated == expensive (e.g., parsing, figuring out dependencies among words)
- Shallow == heuristic == cheap
  - Ex: assume every verb between two entities define a relation



Why shallow methods?



- The difference in computational cost is several orders of magnitude [Christensen et al., K-Cap 2011]
- Shallow methods can be deployed at Web scale
- Probabilistic methods filter out individual mistakes as noise

Where do we draw the line?

- Shallow vs deep depends on many factors, but generally we assume that
  - POS tagging and chunking are shallow
  - Parsing and beyond are considered deep
- Virtually all NLP toolkits out there will handle theses tasks
  - GATE (http://gate.ac.uk/)
  - LingPipe (http://alias-i.com/lingpipe)
  - Apache OpenNLP (http://opennlp.apache.org/)
  - Apache UIMA--Unstructured Information Management Applications (http://uima.apache.org/)
  - …

Roadmap—Part I

- Finding entities
  - Shallow "ontology" extraction
  - Entity identification and Co-reference resolution
- Finding (binary) relations
  - One sentence at a time
  - All sentences "at once" with clustering
- Applications
  - Social media aggregation/analytics

Finding entities and classes with Hearst Patterns

- Succinct list of syntactic patterns expressing hyponymy (i.e., sub-classes or instances of a class) [Hearst, ACL 1992]
- Ex:

NP "such as" NP*
cities such as London, Paris, and Rome

"such" NP "as" NP* "or" | "and" NP
works by such authors as Herrick and Shakespeare

NP, NP* "or" | "and" other NP
bruises, wounds, broken bones or other injuries

NP "especially" | "including" NP*
all common-law countries including Canada and England

KnowItAll

- [Etzioni et al., 2005]
- Multiple classes: extracting, validating, and generalizing rules. KnowItAll project at U. Washington
- Extracts both entities and relations
  - Pattern learning: relies on a small number of templates, which are instantiated as good extractions are found
  - Subclass extraction: aims at finding (part of) the concept hierarchy (e.g., physicist IS-A scientist)
- Generate-and-test loop: apply extraction patterns and test the plausibility of the extracted results—using Point-wise Mutual Information (PMI)

KnowItAll

template

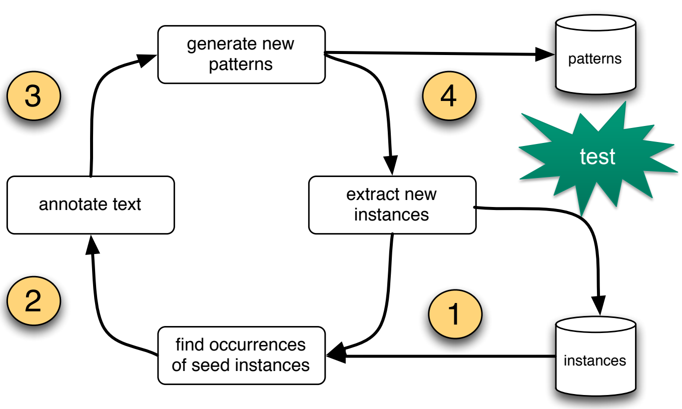
Predicate:	Class1
Pattern:	NP1 "such as" NPList2
Constraints:	head(NP1)= plural(label(Class1)) & properNoun(head(each(NPList2)))
Bindings:	Class1(head(each(NPList2)))

rule

Predicate:	City
Pattern:	NP1 "such as" NPList2
Constraints:	head(NP1)= "cities" & properNoun(head(each(NPList2)))
Bindings:	City(head(each(NPList2)))
Keywords:	"cities such as"

Generate-and-test loop

- General idea that has been used over and over again: finding entities, relations, etc.

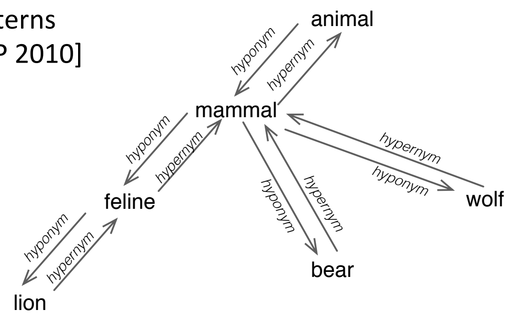


Entity extraction—other pointers

- Single class: expanding a list of known entities using a search engine
  - [Pasca, 2007] Pasca, M. (2007). Weakly-supervised discovery of named entities using web search queries. In Proc. of the sixteenth ACM conference on Conference on information and knowledge management, pages 683–690. ACM.
  - [Wang and Cohen, 2008] Wang, R. C. and Cohen, W. W. (2008). Iterative Set Expansion of Named Entities Using the Web. In 2008 Eighth IEEE International Conference on Data Mining, pages 1091–1096. IEEE.
- Single class from multiple sources: combining extractors
  - [Pennacchiotti and Pantel, 2009] Pennacchiotti, M. and Pantel, P. (2009). Entity Extraction via Ensemble Semantics. Language, 96(August):238–247.
- Multiple classes: extracting, validating, and generalizing rules
  - [Etzioni et al., 2005] Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised named-entity extraction from the Web: An experimental study. Artificial Intelligence, 165(1):91–134.

Finding IS-A relationships

- Iteratively apply Hearst patterns [Kozareva and Hovy, EMNLP 2010]
- Optimization problem: removing redundant edges



"animals such as lion and ? " → {lion, tiger, jaguar}

" ? such as lion and tiger" → feline

" ? such as feline" → "Big Predatory Mammals"

"mammals such as felines and ? " → {felines, bears, wolves}

Adding some depth: [Snow et al., NIPS 2005]

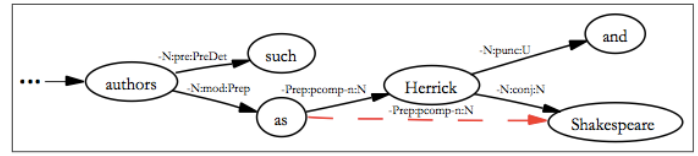
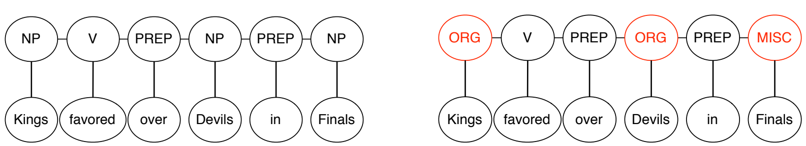


Figure 1: MINIPAR dependency tree example with transform

- Corpus: 6 million sentences from several News corpora
- Parsing provides more reliable features, including part of speech tags and structural dependencies, compared to the syntactic features of the Hearts patterns
- The resulting concept hierarchy was shown to be more precise and more detailed than Wordnet
  - Although Wordnet is designed to be general

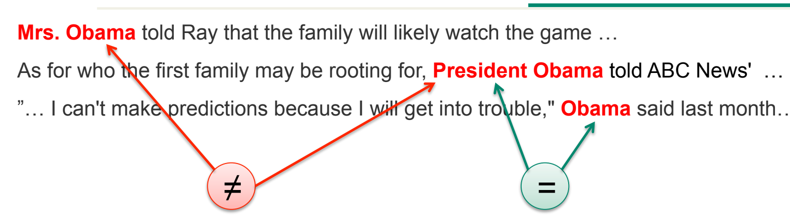
NER as sequence labeling problem

- Start with annotated example sentences, indicating where the entities are, and their types (e.g., ORG, PER, LOC,…)



- Labeled sequence prediction: sampling from a trained CRF model: Stanford NER tool [Finkel et al., ACL 2005]
  - Uses both local and global information; features include POS tags, previous and following words, n-grams, …
- Invaluable open source, stand-alone tool
  - Off the shelf, it comes trained for news articles, but can be easily trained for other domains

De-duplication and disambiguation of entities



- Once references to named entities are identified, detect whether any of them refer to the same real world entity and which don't
- One intra-document co-reference resolution tool provided by the GATE framework: Orthomatcher [Bontcheva et al., TALN 2002]

Orthomatcher [Bontcheva et al., 2002]

- Rule-based: inexpensive, ad-hoc, but shown to perform well in many tasks
  - Gazetteers: known entities, common abbreviations (Ltd., Inc., …), synonyms (New York = the big apple, …), and ad-hoc list for the specific domain
- Proper name co-reference resolution
  - Orthographical matches (James Jones = Mr. Jones); Token Re-ordering and abbreviations (University of Sheffield = Sheffield U.)
  - Non-transitivity and exclusion triggers in some rules (BBC News ≠ News);
- Pronominal co-reference resolution
  - Ad-hoc rules from empirical observation (e.g., 80% of all 'he, his, she, her' mentions refer to the closest person in the text)
- Fairly accurate on news articles
  - Well-written text

Text type	OM precision	recall
broadcast news	94%	92%
newswire	98%	92%
newspapers	98%	95%

Entity identification and linking [Cucerzan, 2007]

- Performs entity identification, in-document co-reference resolution, and cross-document co-reference resolution
  - Instead of relying on heuristics, the disambiguation rules come from statistical analysis of Wikipedia (>1.4 million entities) and a large corpus of Web searches
- Surface forms, entities, tags and context words
  - Mentions to entities are compared against the knowledge base using other terms nearby in the sentence that indicate context
  - Spread-activation like algorithm
- Building the knowledge base
  - Parse Wikipedia's entity pages, redirecting pages, disambiguation pages, and list pages

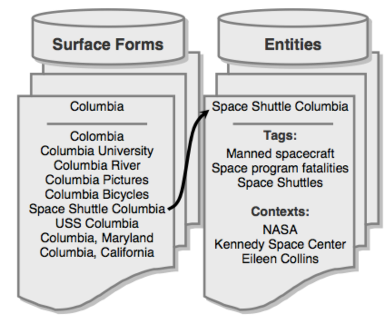


Figure 1. The model of storing the information extracted from Wikipedia into two databases.

Courtesy of S. Cucerzan

Entity identification and linking [Cucerzan, 2007]

- Entity recognition builds on capitalization rules, so the document processing starts with splitting sentences and finding the correct case for all words in each sentence
  - Builds on a large (1B words) corpus of Web documents
- Resolves structural ambiguity (e.g., [[Barnes and Noble]] or [[Barnes]] and [[Noble]]), possessives, and prepositional attachments using the surface forms extracted from Wikipedia (or the Web corpus for entities not in Wikipedia)
- Disambiguation based on vector space model similarity of mentions in the text and mentions in the knowledge base
- Evaluation of 756 surface forms, of which 127 were non-recallable, from news text shows accuracy of 91.4%

Entity linking

Mrs. Obama told Ray that the family will likely watch the game …

As for who the first family may be rooting for, **President Obama** told ABC News' …

"… I can't make predictions because I will get into trouble," **Obama** said last month…

en.wikipedia.org/wiki/Michelle_Obama

en.wikipedia.org/wiki/Barack_Obama

- Linking mentions in the corpus to a knowledge base (e.g., Wikipedia) would accomplish both co-reference resolution and disambiguation
- Cluster entities that cannot be linked (e.g., are not on Wikipedia)

Collective entity linking

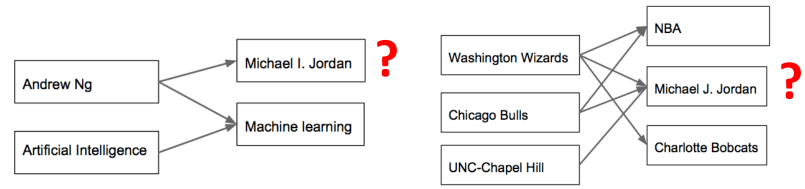
Surface Form	Entity
Michael Jordan	Michael Jeffery Jordan
Michael J. Jordan	
MJ	
MJ23	
Jordan	
Air Jordan	

Surface Form	Entity
Michael Jordan	Michael Jordan (mycologist)
	Michael Jordan (footballer)
	Michael Jordan (English businessman)
	Mike Jordan(English racing driver)
	Michael Jordan (Irish politician)
	Michael I. Jordan (Researcher)

- Multiple surface forms for the same entity, and multiple entities with the same "canonical" surface form
- Often it is easier to link all named entities at once

Michael Jordan to headline celebrity hoops fundraiser event for Obama re-election campaign
...
The first player that Obama, a longtime Chicago Bulls fan's mouth, referenced in doing so was, of course, Michael Jordan.
...
In addition to Jordan, the "2012 Obama Classic" will also feature Hall of Famer Patrick Ewing, New York Knicks All-Star Carmelo Anthony, and multiple other NBA and WNBA stars past and present,

Collective entity linking



Michael Jordan to headline celebrity hoops fundraiser event for Obama re-election campaign
...
The first player that Obama, a longtime Chicago Bulls fan's mouth, referenced in doing so was, of course, Michael Jordan.
...
In addition to Jordan, the "2012 Obama Classic" will also feature Hall of Famer Patrick Ewing, New York Knicks All-Star Carmelo Anthony, and multiple other NBA and WNBA stars past and present,

- NP-hard optimization
- Match entities in the text to nodes in the KB and search for a compact sub-graph that best "covers" the text
- Take into account:
  - Similarity of mentions and KB entries
  - Prior frequency of entities
  - Some notion of coherence for the edges

Roadmap—Part I

- Finding entities
  - Shallow "ontology" extraction
  - Entity identification and Co-reference resolution
- Finding (binary) relations
  - One sentence at a time
  - All sentences "at once" with clustering
- Applications
  - Social media aggregation/analytics

Open/Closed Relation extraction

	Closed	Open
Number of target relations	1	All
Relation-specific training	Yes	No
Cost	Linear on the number of relations	Constant

- Closed relation extraction == binary classification problem: does the sentence express a relation (YES/NO)?
  - Supervised systems: [Culotta and Sorensen, ACL 2004]; [Bunescu and Mooney, EMNLP 2005]; [Zelenko et al., JMLR 2003]
  - Bootstrapping approaches: DIPRE [Brin, WWW 1998]; Snowball [Agichtein and Gravano, DL 2000];
    KnowItAll [Etzioni et al., WWW 2004]
  - Distant supervision: [Mintz et al., ACL 2009]

Open relation extraction

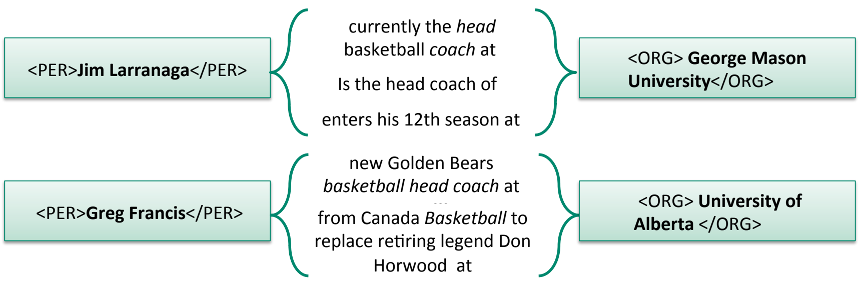
- Term coined by Banko and Etzioni (2008) to mean learning both the relations and the instances from the data
- Some landmark systems/papers
  - TextRunner uses a classifier based on Conditional Random Fields (CRFs) over sentences [Banko and Etzioni, ACL 2008]
  - ReVerb is a manually refined version of TextRunner focusing a subset of relation patterns [Fader et al., ACL 2011]
  - StatSnowBall builds on the SnowBall system [Zhu et al., WWW 2009]
  - [Hasegawa et al., ACL 2004] introduced an unsupervised method based on text clustering

ORE "one sentence at a time"

Frequency	Pattern	Example
38%	E_1 Verb E_2	X established Y
23%	E_1 NP Prep E_2	X settlement with Y
16%	E_1 Verb Prep E_2	X moved to Y
9%	E_1 Verb to Verb E_2	X plans to acquire Y

- Relation extraction as a sequence prediction task
  - TextRunner: [Banko and Etzioni, 2008] a small list of part-of-speech tag sequences that account for a large number of relations in a large corpus
  - ReVerb: [Fader et al., 2011] use an even shorter list of patterns, extracting verb-based relations
- The ReVerb/TextRunner tools have extracted over one billion facts from the Web
- Efficient: no need to store the whole corpus
- Brittle: multiple synonyms of the the same relation are extracted

ORE on "all sentences" at once



- [Hasegawa et al., 2004] use hierarchical agglomerative clustering of all triples (E_1, C, E_2) in the corpus
  - E_1, C, E_2 where the context C derives from all sentences connecting the entities
  - The clustering is done on the context vectors (not the entities)
- All triples (and thus, entity pairs) in the same cluster belong to the same relation

SONEX

- Offline (HAC clustering)
  - [Merhav et al., 2012] Merhav, Y., de Sá Mesquita, F., Barbosa, D., Yee, W. G., and Frieder, O. (2012). Extracting information networks from the blogosphere. ACM Transactions on the Web.
- Online (buckshot): cluster a sample and classify the remaining sentences, one at a time
  - No discernible loss in accuracy, but much higher scalability
  - Also allows the same entity pair to belong to multiple relations

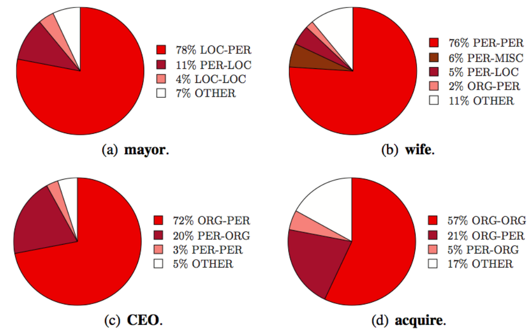
SONEX: Clustering features

- Clustering features derived from the words between entities
  - **Unigrams**: stemmed words, excluding stop words.
    - [Allan, 1998] Allan, J. (1998). Book Review: Readings in Information Retrieval edited by K. Sparck Jones and P. Willett. Inf. Process. Manage., 34(4):489–490.
  - **Bigrams**: sequence of two (unigram) words (e.g., Vice President).
  - **Part of Speech Patterns**: small number of relation-independent linguistics patterns from TextRunner [Banko and Etzioni, 2008]
- Verbal and non-verbal relations
- Weights: Term frequency (*tf*), inverse document frequency (*idf*) and **Domain frequency** (*df*) [Merhav et al., SIGIR 2010]

$$df_i(t) = \frac{f_i(t)}{\sum_{1 \le j \le n} f_j(t)}$$

SONEX: importance of domain

- DF works really well except when MISC types are involved
- Example: **coach**
  - LOC–PER domain: (England, Fabio Capello); (Croatia, Slaven Bilic)
  - MISC–PER domain: (Titans, Jeff Fisher); (Jets, Eric Mangini)



- DF alone improved the f-measure by 12%

SONEX: From Clusters to Relations

- Clusters are sets of entity pairs with similar contexts
- We find **relation names** by looking for prominent terms in the context vectors
  - Most frequent term
  - Centroid of cluster

Relation	Cluster
Campaign Chairman	McCain : Rick Davis Obama : David Plouffe
Strongman President	Zimbabwean : Robert Mugabe Venezuelan : Hugo Chavez
Chief Architect	Kia Behnia : BMC Software Brendan Eich : Mozilla
Military Dictator	Pakistan : Pervez Musharraf Zimbabwe : Robert Mugabe
Coach	Tennessee : Rick Neuheisel Syracuse : Jim Boeheim

SONEX: from clusters to relations

- Evaluate relations by computing the agreement between the Freebase term and the chosen label
  - Scale: 1 (no agreement) to 5 (full agreement)

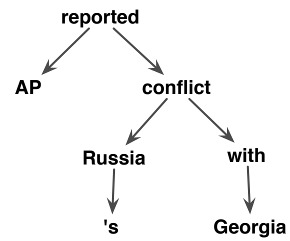
Relation	Omiotis			Manual		
	Centroid	SDEV	Difference	Centroid	SDEV	Difference
Capital	5.00	5.00	*0.00*	5.00	3.00	*2.00*
Governor	3.97	2.03	*1.94*	3.98	3.02	*0.96*
Athlete Repr.	1.00	4.76	*3.76*	3.82	3.82	*0.00*
Marriage	3.97	3.97	*0.00*	3.92	3.92	*0.00*
Author	3.92	2.96	*0.96*	3.88	4.84	*0.96*
Headquarters	4.55	1.15	*3.40*	4.55	3.75	*0.80*
President	4.60	4.60	*0.00*	4.55	4.55	*0.00*
Prime Minister	4.89	4.89	*0.00*	3.00	3.00	*0.00*
Mayor	5.00	5.00	*0.00*	4.87	4.87	*0.00*
Founder	5.00	5.00	*0.00*	5.00	4.00	*1.00*
Average	4.19	3.94	*1.00*	4.26	3.88	*0.57*

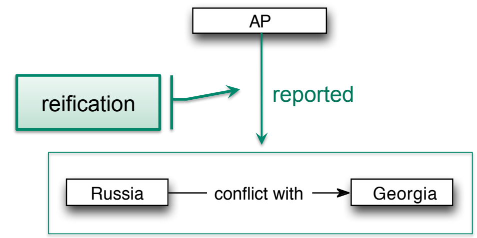
SONEX vs ReVerb—clustering analysis

Systems	Purity	Inv. Purity
ReVerb	**0.97**	0.22
SONEX	0.96	**0.77**

- Purity: homogeneity of clusters
  - Fraction of instances that belong together
- Inv. purity: specificity of clusters
  - Maximal intersection with the relations
- Also known as overall f-score
  - [Larsen and Aone, 1999] Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In Proc. of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 16–22. ACM.

Meta CRF: deep vs shallow cost/benefit





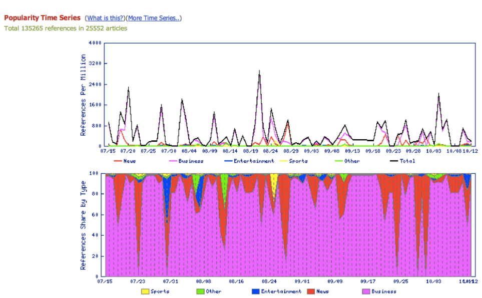
- CRF taking into account structural features of the tree to label direct and indirect (meta) relations [Mesquita and Barbosa, ICWSM 2011]
- Outperformed the baseline by
  - 190% on meta relations and
  - 86% on statements with direct relations

Roadmap—Part I

- Finding entities
  - Shallow "ontology" extraction
  - Entity identification and Co-reference resolution
- Finding (binary) relations
  - One sentence at a time
  - All sentences "at once" with clustering
- Applications
  - Social media aggregation/analytics

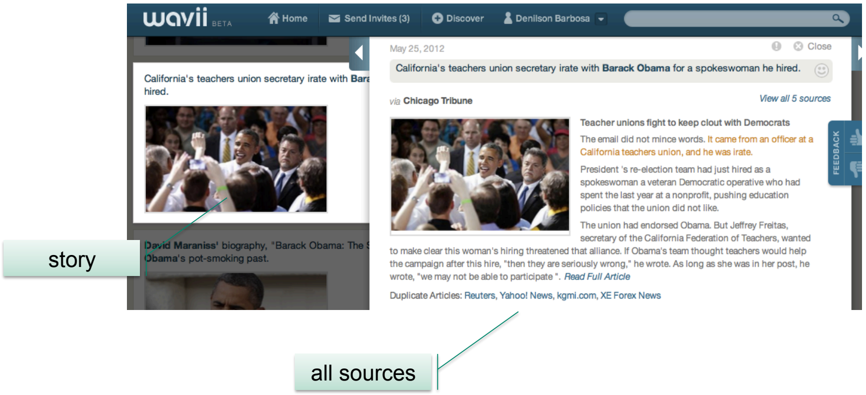
Popularity/hit counts over time

- Source: http://www.textmap.com entry about Barack Obama (around July 2008)
- Task: entity recognition
  - Identifying that the articles mention Barack Obama
- Task: entity disambiguation
  - Figuring out all "surface" forms for the same entity
- Task: entity disambiguation
  - Figuring out which Barack Obam the articles mention
- Task: clustering
  - Grouping the article sources by kind (sports, business, entertainment, …)



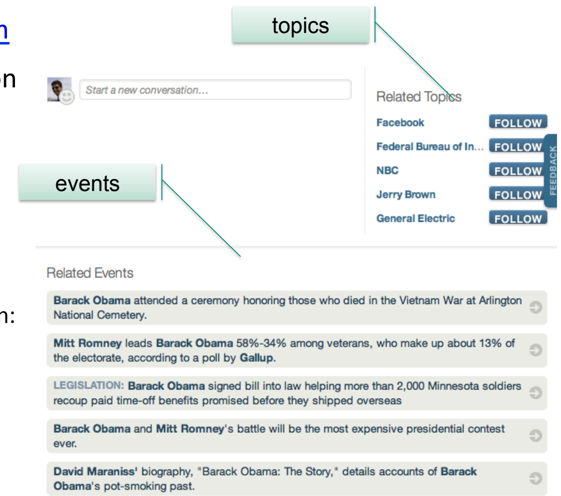
Social delivery—story centered

- Wavii http://wavii.com story (May 2012)
- News aggregator building on preferences and your social network



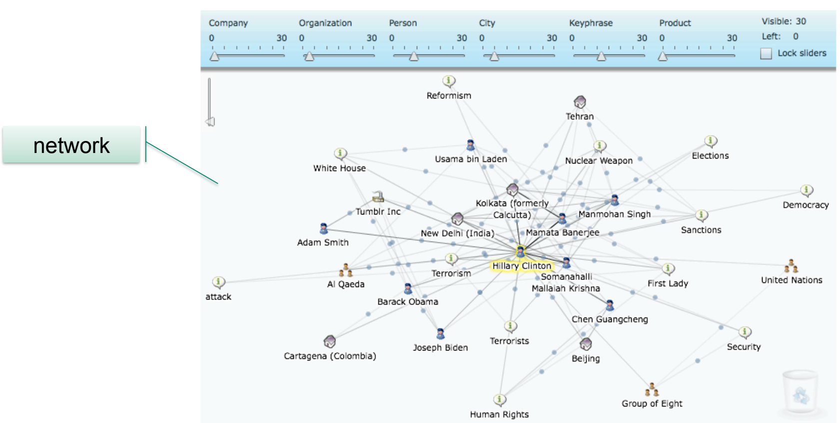
Social delivery—story centered

- Wavii: http://wavii.com
- Tasks: news aggregation
  - Identifying topics and related news
- Task: content filtering
  - Identifying preferences
- Task: event extraction
  - Similar to summarization: finding a sentence that captures the news item (e.g., its title?)



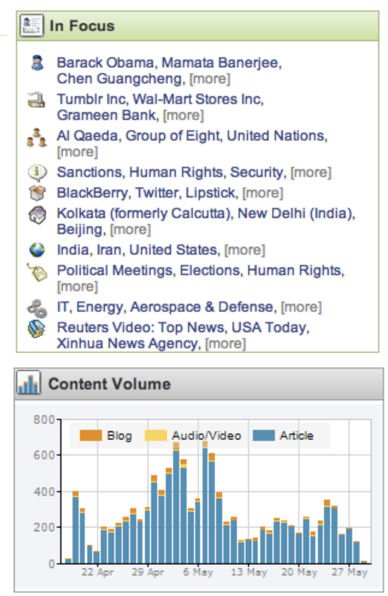
Information networks in Social Media Analysis

- Source: http://www.silobreaker.com network around Hilary Clinton
- Integrates news, blogs, audio/video feeds, press releases…



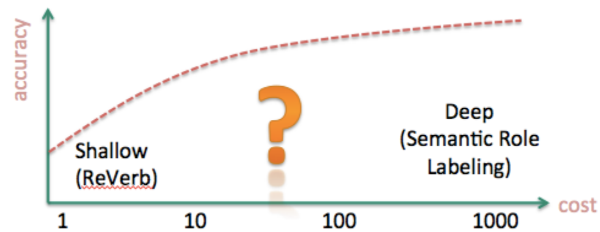
SILO Breaker

- Extracting information networks from text
- Task: entity recognition
  - Figuring out which entities are mentioned in the corpus (and their kinds), and the relations among entities
- Task: entity disambiguation
  - Figuring out all "surface" forms for the same entity
- Tasks: relation extraction
  - Determining which entities and/or relations are most relevant to the entity on the spotlight



Summary

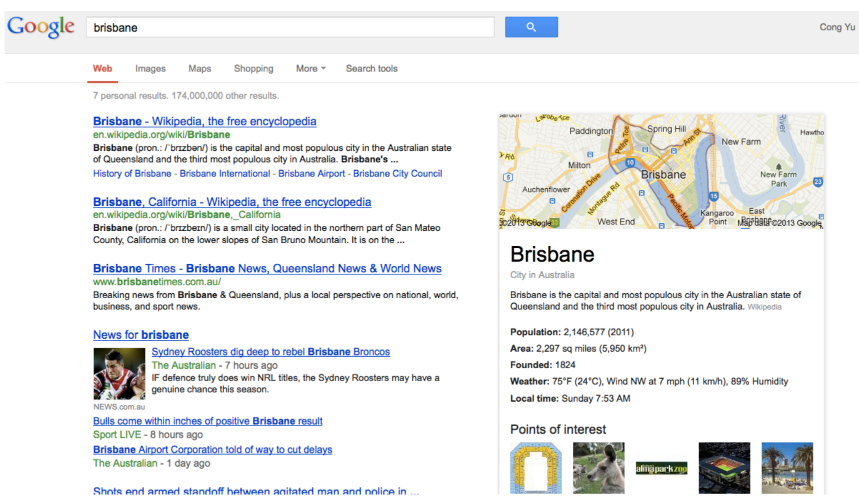
- Large-scale open information extraction is an active and exciting area, with many impressive results and ongoing projects
  - YAGO (Max Planck Institute), KnowItAll (U. Washington), NELL (Carnegie Mellon U.), Google's Knowledge Graph, Microsoft's Satori, Probase
  - …
- Challenges/future work:
  - Plug and play NLP????
  - Evaluation



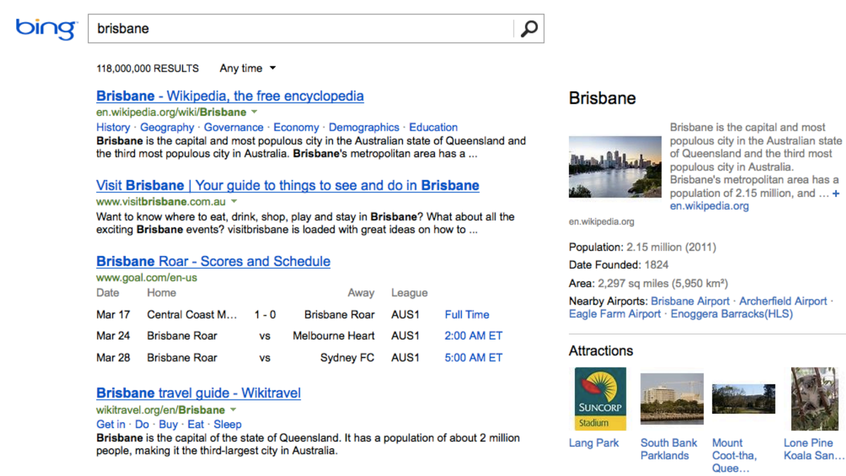
Outline

- Motivation
  - Why doing all this in the first place?
  - Define what shallow means – no deep linguistic analysis
  - Emphasizing why the need for shallow extraction techniques
- PART I: shallow extraction techniques
  - Entity extraction
  - Relation extraction
  - Application Social text mining
- Part II: Bring Knowledge to Search
- Part III: Real life knowledge base, scalability and probability

Knowledge is Becoming Part of Search



Knowledge is Becoming Part of Search



Knowledge is Becoming Part of Search

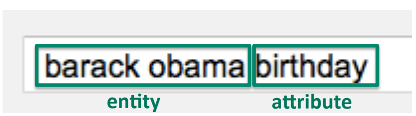
- "… a new breed of search experiences … the user is saved the burden of culling documents from a results list and laboriously extracting information buried within them."
- - Baeza-Yates & Raghavan on Next Generation Web Search
- All major search engines have started incorporating "knowledge" into search results
- Search users do respond, albeit slowly, to the capabilities of search engines → leading to more innovations on integrating knowledge and search.

Leveraging Knowledge for Search

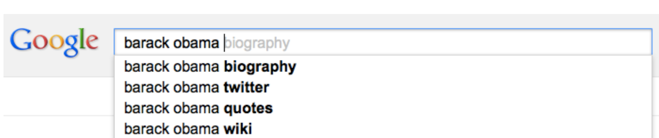
- Improving web search
  - **Query enrichment**
  - Entity navigation
- Shallow knowledge search
  - Search over knowledge bases
  - **Knowledge search over Web**
- AI-ish knowledge search
  - Question answering
  - Natural language search
  - Survey by *Lopez, Uren, Sabou, Motta*, Semantic Web 2(2):125

Improving Web Search via Query Enrichment

- Goal: better query understanding by associating semantics with user queries
  - Complimentary to using query logs to learn classes and attributes
- Case studies:
  - Query tagging

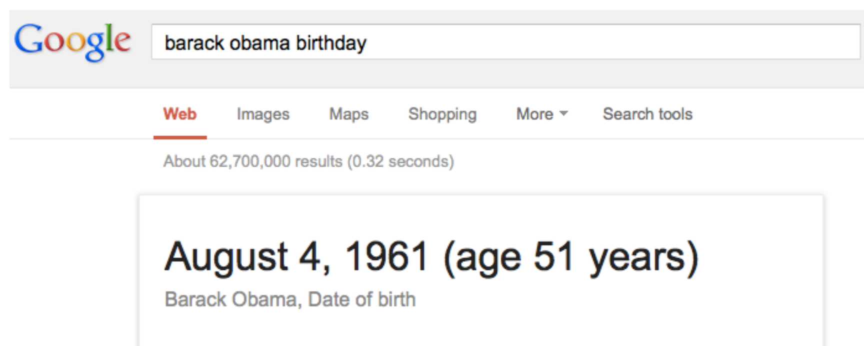


  - Query suggestion



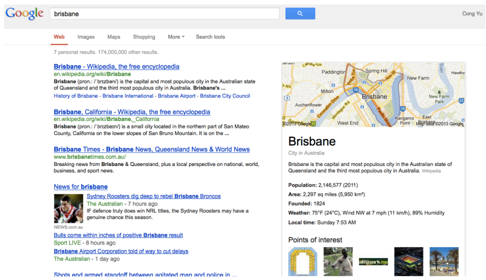
Query Enrichment is Critical for Knowledge Search

- Query tagging is the first step toward knowledge search
- Query suggestion can guide users toward queries they otherwise assume can not be handled



From Entity to the Information Box, via Knowledge Base

- Industrial focus has been on KB construction
  - Entity navigation is considered to be relatively simple
  - Not true, but KB construction poses more challenges now
- However, some challenges are already hard to ignore:
  - Information selection
  - Information visualization
  - Information freshness



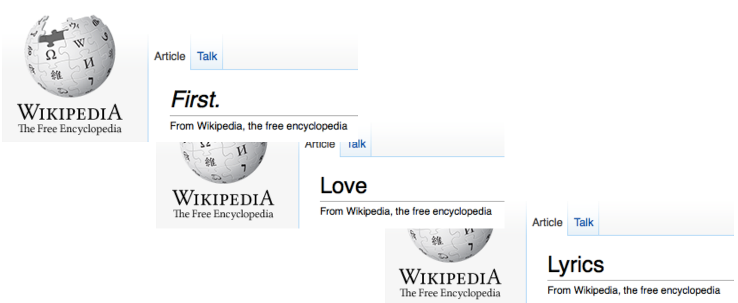
Recent Studies on Query Tagging

- Named entity recognition
  - **[Guo et al, SIGIR 2009]**
- Rich interpretation
  - **[Li, Wang, Acero, SIGIR 2009]**
- Template mining:
  - [Agarwal, Kabra, Chang, WWW 2010]

The Query Tagging Problem

- Even simple named entity tagging is not easy

"first love lyrics"



➔ The real entity is "first love" of class "song"

Tagging Named Entities in Queries [Guo et al, SIGIR 2009]

- Challenges:
  - Short text
  - Fewer language features: e.g., no punctuation, no capitalization
- Intuition:
  - Use context to disambiguate
  - Use query logs to learn probabilities
  - Gather class labels on seed entities

Probabilistic Model

- Model the tagging problem as computing the probabilities of all possible triples, (e, t, c), that can represent the query
  - e: entity
  - t: context
  - c: class label for the entity
- "first love lyrics"
  - ("first love", "# lyrics", song), or
  - ("first", "# love lyrics", album), or
  - ("love" "first # lyrics", emotions), or
  - ("lyrics", "first love #", music)
- The one with the highest probability can be considered as the correct tagging.

Probabilistic Model

- The learning problem is:

$$\max \prod_{i=1}^{N} \Pr(e_i, t_i, c_i)$$

- Pr(e, t, c) can be estimated as:

$$\begin{aligned} \Pr(e,t,c) &= \Pr(e)\Pr(c|e)\Pr(t|e,c) \\ &= \Pr(e)\Pr(c|e)\Pr(t|c) \end{aligned}$$

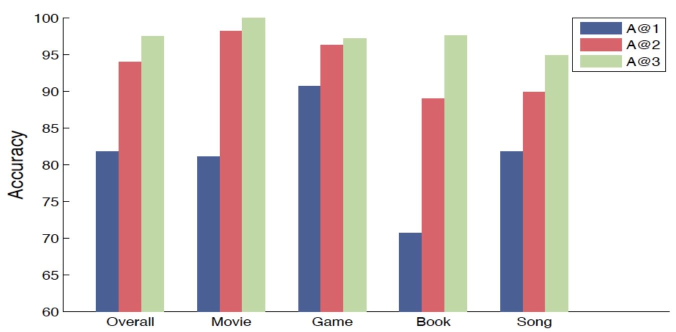
  - Simplification: context only depends on the class
    - I.e., "lyrics" is more likely associated with songs, regardless which song

Training and Prediction

- Step 1: Gather seed set of (entity, class) pairs
- Step 2: Match the seed set with query log and gather their contexts: (e_{seed}, t)
- Step 3: Use the contexts gathered from step 2, match with the query log again and gather expanded entities: ($e_{expanded}$, t)
- Conditional probability estimates:
  - Pr(e): occurrence frequency in logs
  - Pr(t|c): learned from Step 2.
  - Pr(c|e): learned from Step 3 with fixed Pr(t|c) from Step 2.
- Prediction: apply to all possible query segmentations

Results

- 12 Million unique queries ➔ tagged 0.15 Million
  - Recall is quite low
- Sampled 400 queries for evaluation
  - 111 movie, 108 game, 82 book, 99 song



Challenging Queries

- "american beauty company"
  - Highly popular entity that is wrong
- "lyrics for forever by brown"
  - Multiple contexts
- "canon sd350 camera" or "canon vs nikon"
  - Multiple entities

The Query Tagging Problem

- Tagging just the entities is not enough

"canon powershot sd850 camera silver"

Rich interpretation:

"canon" ➔ brand

"powershot sd850" ➔ model

"camera" ➔ type

"silver" ➔ attribute

Rich Interpretation [Li et al, SIGIR 2009]

- Fully interpret the query instead of just tagging a single entity
- Handles multi-entity and multi-context queries
- Limited within a specific domain
- Challenges:
  - Which learning model to use:
    - Query is no longer treated as bag of words but a sequence instead
  - Training labels are harder to generate
    - Each query can have multiple labels co-exist
- Sequential learning model is easy to find: Conditional Random Fields (CRF)

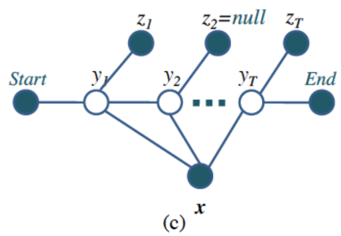
Automatically Obtaining Labels (Shopping)

- Target schema
- Leverage click log to find (query, product) pairs
  - Focus on queries that led to clicks on product listing pages
- Extract metadata from those products to produce (query, metadata) events
  - Relatively easy since product pages are well-structured (within MSN shopping)
- Map metadata to target to produce (query, target) pairs
- Conservative automatic labeling
  - Only query tokens mapped to exactly one target field are labeled
- Complementing automatic labels with manual labels
  - E.g., "cheap" ➔ SortOrder

Fields	**Example use in queries**
Brand	**canon** powershot sd850
Model	canon **powershot sd800**
Type	canon **digital cameras** silver
Attribute	canon digital cameras **silver**
Merchant	digital cameras at **best buy**
SortOrder	**best** digital cameras
BuyingIntent	**buy** canon digital cameras
ResearchIntent	digital cameras **review**
Other	digital cameras **at** best buy

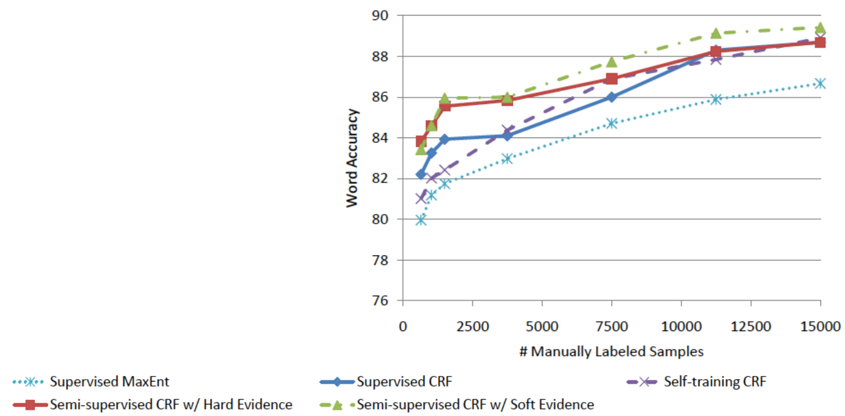
Using Automatic Labels in CRF

- Automatic labels can often be wrong ➔ Adopt them as soft evidences
- The true labels are created as hidden
- The automatic labels on the query terms are created as observed variables to bias the true label selections



Results

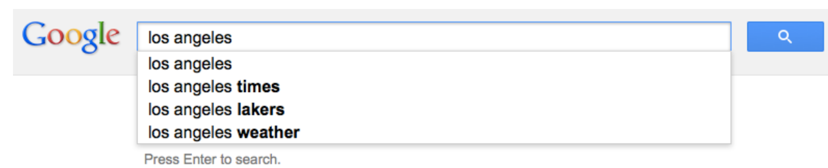
- Training labels
  - Automatic: 50K labels for clothing; 20K labels for electronics
  - Enhanced by 4K manual labels for clothing and 15K manual labels for electronics



Recent Studies on Query Suggestion

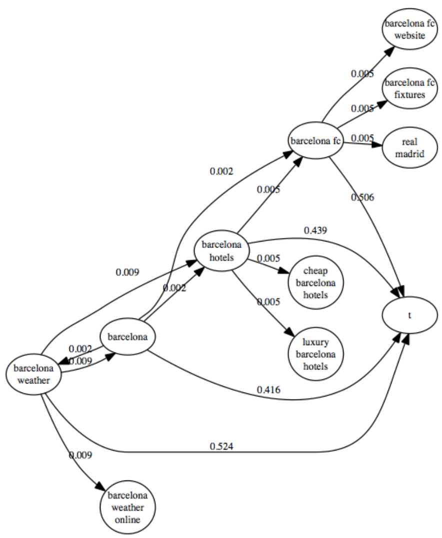
- Query to Query
  - **[Szpektor, Gionis, Maarek, WWW 2011]**
- Entity to Query
  - **[Bordino et al, WSDM 2013]**

The Query Suggestion Problem



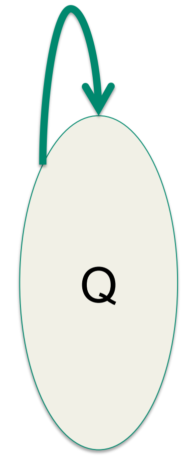
- Enormously popular with the users
  - Works very well for head queries
- Approaches
  - Query similarity
    - E.g., Cosine similarity, edit distance
  - Query flow graph
    - Leveraging co-occurrences of queries in the same query session

Query Flow Graph [Boldi et al, CIKM 2008]



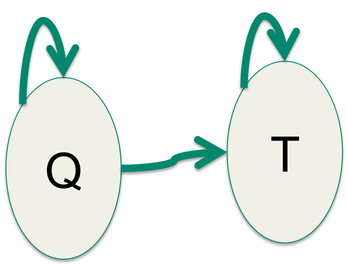
Query Flow Graph

- Constructed from query session logs
- Nodes are queries
- Create an edge (q1, q2) if:
  - q2 appeared as a reformulation of q1 in a session
- Edge weights can be assigned in many ways
  - Pr(q2|q1) = f(q1, q2) / f(q1)
  - PMI(q1, q2) = log(f(q1,q2) / f(q1)f(q2))



Query Template Flow Graph [Szpektor et al, WWW 2011]

- Intuition:
  - If, users often search “new york restaurants” after searching for “new york hotels” and similarly for other popular cities such as “shanghai”, “paris”, etc.
  - Then, “pkoytong restaurants” can be a good recommendation candidate for “pkoytong hotels” since “pkoytong” is a also a city



Query to Template Edges

- Query entity tagging techniques made query template generation possible!
  - “new york restaurants” ➔ “<city> restaurants”
  - Essentially, knowledge can be used to enrich the query to address many issues associated with long tail queries
- Computing edge weights between query and template
  - Assuming a hierarchical ontology
  - S(q, t) is computed based on where in the hierarchy the query entity in q is matched

$$S_{\mathrm{qt}}(q,t) = \alpha^{d(z,e)}$$

Template-to-Template Edges

- Creating edges between templates
  - A template-to-template edge occurs if and only if a query-to-query edge occurs and the two queries match the two templates, respectively, with the same entity
- Computing edge weights between templates
  - The more supporting query-to-query pairs there are, the higher the weights

$$S_t(t_1, t_2) = \sum_{(q_1,q_2)\in \mathrm{Sup}(t_1,t_2)} s_{qq}(q_1, q_2),$$

$$s_{tt}(t_1, t_2) = \frac{S_t(t_1, t_2)}{\sum_t S_t(t_1, t)}.$$

Generating Recommendations

- Let S(x, y) be the probability of reaching y from x in the graph
  - E.g., product of all the edge weights on the path from x to y.
- The score of r(q1, q2) can be computed as
  - S(q1, q2) based on original query flow graph, plus
  - SUM_{ij}((q1, t_i), S(t_i, t_j), S(q2, t_j)) based on the query template flow graph
- Tough cases remain
  - E.g., entities that do not appear in the ontology hierarchy, which is much more common in long tail queries

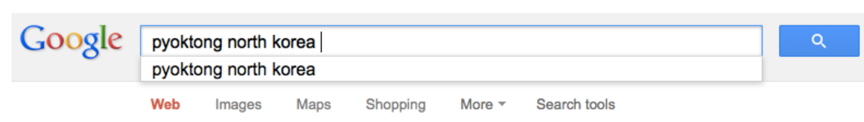
Results

- The query template graph: 95M queries, 60 candidate templates per query
  - Number of edges is linear to the number of nodes

	pair occurrences				
	QFG		QTFG		relative increase
total in test-set		3134388		3134388	
upper-bound coverage	(22.65%)	709832	(28.17%)	882851	24.37%
# in top-100	(16.97%)	531854	(25.49%)	799001	50.23%
# in top-10	(9.49%)	297462	(20.74%)	649939	118.49%
# ranked highest	(2.86%)	89740	(10.01%)	313638	249.5%
MAP		0.050		0.137	
avg. position		18.35		8.3	

The Query Suggestion Problem

- More challenging for long tail queries



- Query similarity often lead to wrong suggestions
- By definition, they are very rare in query log
- Proposed approaches:
  - Dropping non-critical terms [Jain, Ozertem, Velipasaoglu, SIGIR 2011]
  - More interestingly, query templates!

Entity Query Graph [Bordino et al WSDM 2013]

- Recommending queries when a user shows interests in an entity, e.g.:
  - When a user is visiting a Wikipedia page
  - When a user searches for an entity
  - When a user's profile has an entity
- Similar idea to extend query flow graph, but using entities instead of templates
- Again, enabled by query tagging techniques

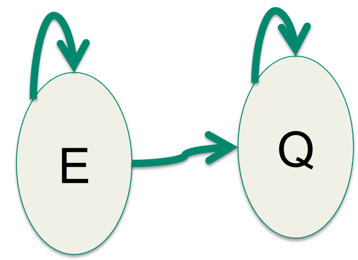
Entity Query Graph

- Entity-to-query edges

$$w_{\mathcal{EQ}}(e \to q) = \frac{f(q)}{\sum_{q_i | e \in \mathcal{X}_{\mathcal{E}}(q_i)} f(q_i)},$$

- Entity-to-entity edges

$$w_E(e_u \to e_v) = 1 - \prod_{i=1,\ldots,r} (1 - p_{q_{i_s} \to q_{i_t}}(e_u \to e_v)).$$



Results

- Generate recommendations based on personalized PageRank over the Entity Query Graph
- Data: 200M queries; 100M entities; linear number of edges again

Testset	Label	EQGraph	Reverse IR
Wikipedia pages	Related and interesting	62.7%	33%
	Related but obvious	3.3%	41.5%
	Unrelated	34%	25.5%
Yahoo! News + Yahoo! Finance	Related and interesting	52%	40%
	Related but obvious	2.3%	34.3%
	Unrelated	45.7%	25.7%
Full testset	Related and interesting	58%	36.1%
	Related but obvious	2.9%	38.4%
	Unrelated	39.1%	25.5%

Leveraging Knowledge for Search

- Improving web search
  - **Query enrichment**
  - Entity navigation
- Shallow knowledge search
  - Search over knowledge bases
  - **Knowledge search over Web**
- AI-ish knowledge search
  - Question answering
  - Natural language search

Shallow Knowledge Search

- Shallow == Queries are represented as
  - Simple keywords
  - Shallowly tagged with structural annotations
    - Nothing resembles the full structure-ness of SQL/SPARQL
- Approaches can be classified on knowledge representation
- Search over knowledge bases
  - Assume the presence of structured knowledge bases
    - Relational databases
    - Semi-structured databases
    - Ontologies such as YAGO [Suchanek, Kasneci, Weikum, WWW 2007]
- Knowledge search over Web
  - Assume only structured annotations on Web documents

Search over Knowledge Bases

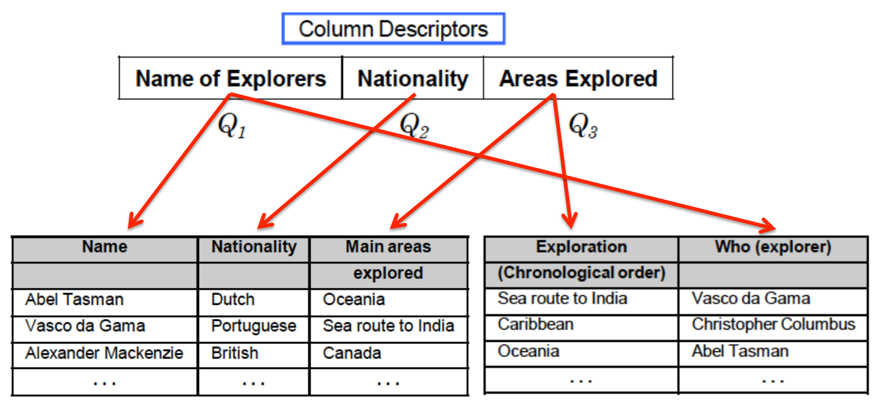
- An "ancient" topic in database community (incomplete list)
  - Relational:
    - DBXplorer [Agrawal, Chaudhuri, Das, ICDE 2002]
    - BANKS [Bhalotia et al, ICDE 2002]
    - DISCOVER [Hristidis, Papakonstantinou, VLDB 2002]
    - ObjectRank [Balmin, Hristidis, Papakonstantinou, VLDB 2004]
  - Semi-structured:
    - XRANK [Guo et al, SIGMOD 2003]
    - TIX [Al-Khalifa, Yu, Jagadish, SIGMOD 2003]
    - XSEarch [Cohen et al, VLDB 2003]
    - PIX [Amer-Yahia et al, VLDB 2003]
    - Schema-Free XQuery [Li, Yu, Jagadish, VLDB 2004]
  - Ontology:
    - NAGA [Kasneci et al, ICDE 2008]
- Semantic query to semantic search in IR / Semantic Web community
  - Combining full-text with ontology [Bast et al, SIGIR 2007]
  - Falcon [Cheng, Qu, Int. J. Semantic Web Inf. Syst., 2009]
  - Semplore [Wang et al, J. Web Semantics 2009]
  - Sig.ma [Tummarello et al, J. Web Semantics, 2010]
  - Search over RDF data [Blanco, Mika, Vigna, ISWC 2011]

Knowledge Search over Web Documents

- Searching for entities and objects (incomplete list)
  - Object level ranking [Nie et al, WWW 2005]
  - Object finder queries [Chakrabarti et al, SIGMOD 2006]
  - EntityRank [Cheng, Yan, Chang, VLDB 2007]
  - Entity package finder [Angel et al, EDBT 2009]
  - Concept search [Giunchiglia, Kharkevich, Zaihrayeu, ESWC 2009]
  - TEXplorer [Zhao et al, CIKM 2011]
- Searching for tabular data (very few studies so far)
  - Studies on table annotation with search as a motivating application
    - [Cafarella et al, VLDB 2008] [Venetis et al, VLDB 2011]
    - [Limaye, Sarawagi, Chakrabarti, VLDB 2010]
    - [Wang et al, ER 2012]
  - *Answering table queries [Pimplikar, Sarawagi, VLDB 2012]*
  - Entity enrichment using tabular data: [Yakout et al, SIGMOD 2012]

Tabular Data Search

- Query: consists of a set of component queries, each correspond to a search for a column
- Answer: combining multiple tables



Name	Nationality	Main areas explored
Abel Tasman	Dutch	Oceania
Vasco da Gama	Portuguese	Sea route to India
Alexander Mackenzie	British	Canada
...	...	...

Exploration (Chronological order)	Who (explorer)
Sea route to India	Vasco da Gama
Caribbean	Christopher Columbus
Oceania	Abel Tasman
...	...

Focusing on Column Mapping

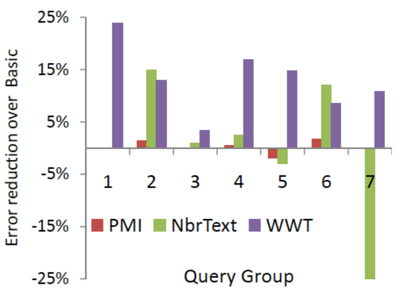
- Naïve Approach
  - Finding relevant tables based on the whole query
  - Match component queries to columns individually
- Global Approach
  - The more relevant the table, the more likely a column can be matched, and vice versa
  - The more relevant the column, the more likely other columns in the same table can be matched
- Solution: Jointly determining query-table, query-column and column-column associations using a graphical model.
  - Nodes in the graphical model are *column* variables: assignable to one of the component queries, plus *relevant* or *irrelevant*

Some Details

- The graphical model takes care of the global modeling, node and edge potentials are modeled using a feature based framework
- Matching columns to component queries
  - Fuzzy matching between tokens in the component query and column header or table context
- Associating columns
  - Based on column content
- Table-level constraints, e.g.:
  - One component query can only match one column per table
  - Each relevant table must match at least n component queries
- Approximate inference

Results

- 59 queries collected using AMT with column splitting based on Google search
  - Single column: “dog breed”
  - Two columns: “country | currency”
  - Three columns: “fast cars | company | top speed”
- 25 million tables from 500 million pages



Grp	**Basic**
1	98.8%
2	72.7%
3	54.1%
4	45.1%
5	31.8%
6	19.7%
7	4.2%

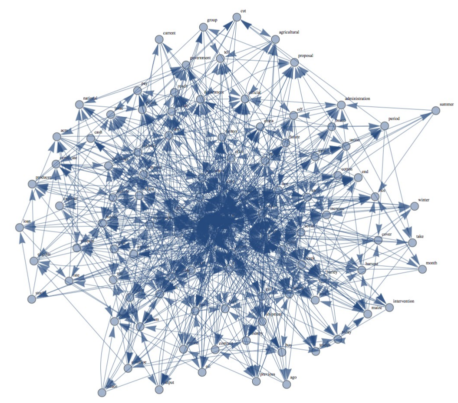
Knowledge Search Summary

- Lots of work are happening in knowledge search
- Lots of challenges remain:
  - Knowledge base maintenance
  - Information selection
  - Search beyond simple entities
    - Some of which are being addressed by Q/A and NLP search

Outline

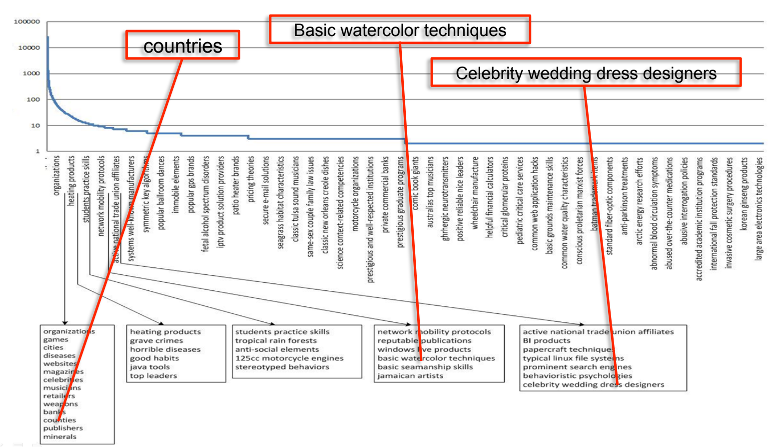
- Motivation
  - Why doing all this in the first place?
  - Define what shallow means – no deep linguistic analysis
  - Emphasizing why the need for shallow extraction techniques
- PART I: shallow extraction techniques
  - Entity extraction
  - Relation extraction
  - Application Social text mining
- Part II: Bring Knowledge to Search
- Part III: Real life knowledge base, scalability and probability

Probase: a probabilistic semantic network



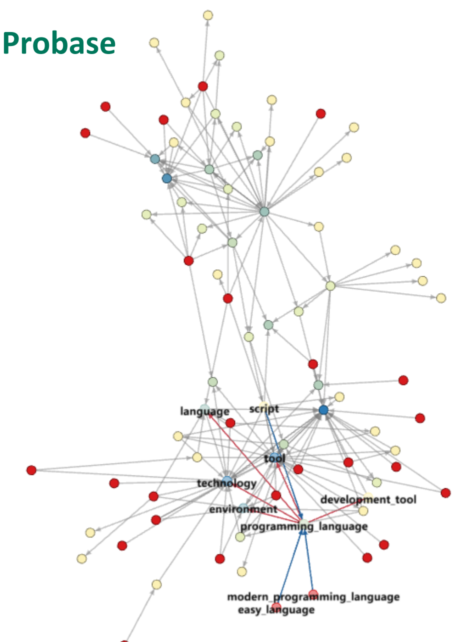
Concepts Entities isA isPropertyOf Co-occurrence

Probase Concepts (2+ millions)

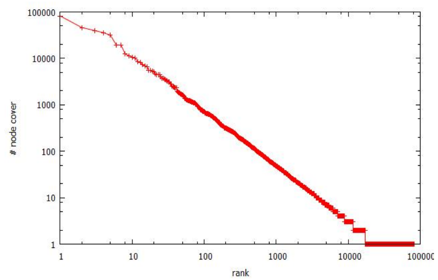


Probase isA error rate: <%1 @1 and <10% for random pair

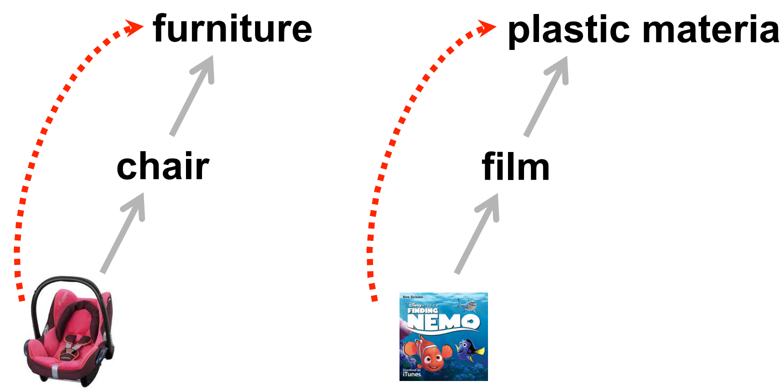
"python" in Probase



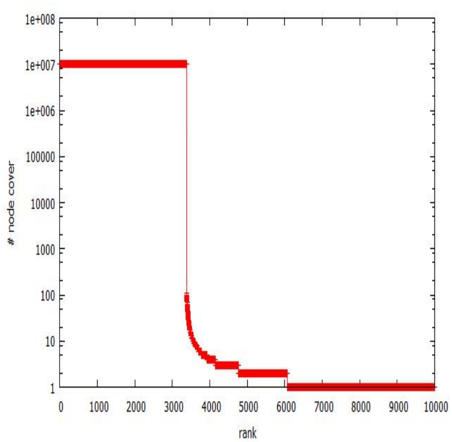
# of descendants (WordNet)



Transitivity does not always hold



# of descendants (early version of Probase)



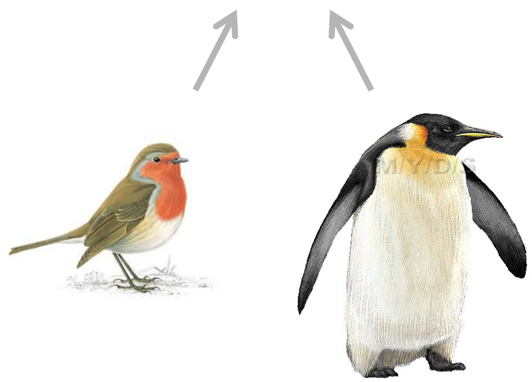
Probase Scores

- Typicality
- Vagueness
- Representativeness
- Ambiguity
- Similarity

foundation for inferencing

Typicality

bird



$$P(e|c) = \frac{n(c,e) + \alpha}{\sum_{e_i \in c} n(c,e_i) + \alpha N}$$

$$P(c|e) = \frac{n(c,e) + \alpha}{\sum_{e \in c_i} n(c_i,e) + \alpha N}$$

"robin" is a more *typical* bird than a "penguin" ➡ $p(robin|bird) > p(penguin|bird)$

Representativeness (basic level of categorization)

software company

$\max_c$ p(c|e) x p(e|c)

company

high typicality p(c|e)

. . . **?** . . .

largest OS vendor

high typicality p(e|c)

Microsoft

Vagueness

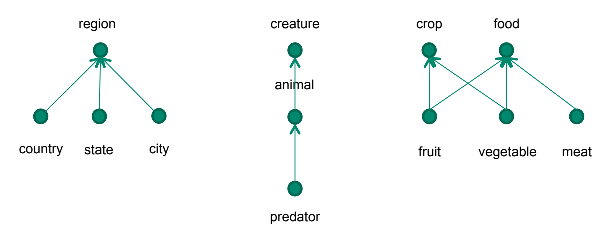
key players
factors
items
things
reasons
…

$$V(C) = \frac{|\{e_i | P(C|e_i) \geq c, \forall e_i \in C\}|}{N(C)}$$

(Do people whom you regard highly regard you highly?)

Ambiguity

- Probase defines 3 levels of ambiguity
  - Level 0 (1 sense): apple juice
  - Level 1 (2 or more related senses): Google
  - Level 2 (2 or more senses): python
- Concepts form clusters, clusters form senses (through isa relation)



Similarity

- microsoft, ibm ➡ 0.933
- google, apple ➡ 0.378 ??

$$sim(t_1, t_2) = \max_{x,y} cosine\ (c_x(t_1), c_y(t_2))$$

Applications

- Query Understanding
  - Head/Modifier/Constraint detection
- ...
- SRL (semantic role labeling) with FrameNet
  - e.g. Tom broke the window.

agent patient

Example: FrameNet

Frame: Apply_heat

FE1 FE2 FE3 FE4

She was FRYING eggs and bacon and mushrooms on a camp stove in Woolley 's billet .

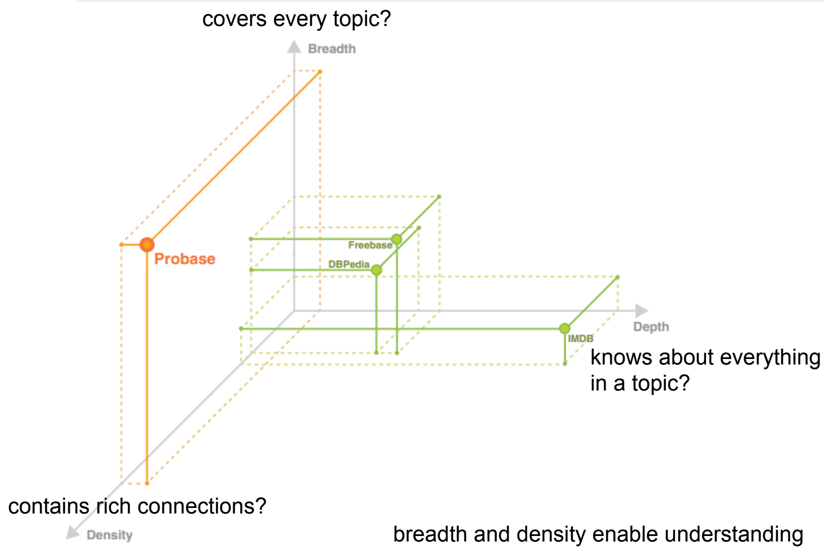
Concept	P(c|FE)
heat source	0.19
Large metal	0.04
Kitchen appliance	0.02

Instance	P(w|FE)
Stove	0.00019
Radiator*	0.00015
Oven	0.00015
Grill*	0.00014
Heater*	0.00013
Fireplace*	0.00013
Lamp*	0.00013
Hair dryer*	0.00012
Candle*	0.00012

Knowledge Bases

	WordNet	Wikipedia	Freebase	Probase
Cat	Feline; Felid; Adult male; Man; Gossip; Gossiper; Gossipmonger; Rumormonger; Rumourmonger; Newsmonger; Woman; Adult female; Stimulant; Stimulant drug; Excitant; Tracked vehicle; ...	Domesticated animals; Cats; Felines; Invasive animal species; Cosmopolitan species; Sequenced genomes; Animals described in 1758;	TV episode; Creative work; Musical recording; Organism classification; Dated location; Musical release; Book; Musical album; Film character; Publication; Character species; Top level domain; Animal; Domesticated animal; ...	Animal; Pet; Species; Mammal; Small animal; Thing; Mammalian species; Small pet; Animal species; Carnivore; Domesticated animal; Companion animal; Exotic pet; Vertebrate; ...
IBM	N/A	Companies listed on the New York Stock Exchange; IBM; Cloud computing providers; Companies based in Westchester County, New York; Multinational companies; Software companies of the United States; Top 100 US Federal Contractors; ...	Business operation; Issuer; Literature subject; Venture investor; Competitor; Software developer; Architectural structure owner; Website owner; Programming language designer; Computer manufacturer/brand; Customer; Operating system developer; Processor manufacturer; ...	Company; Vendor; Client; Corporation; Organization; Manufacturer; Industry leader; Firm; Brand; Partner; Large company; Fortune 500 company; Technology company; Supplier; Software vendor; Global company; Technology company; ...
Language	Communication; Auditory communication; Word; Higher cognitive process; Faculty; Mental faculty; Module; Text; Textual matter;	Languages; Linguistics; Human communication; Human skills; Wikipedia articles with ASCII art	Employer; Written work; Musical recording; Musical artist; Musical album; Literature subject; Query; Periodical; Type profile; Journal; Quotation subject; Type/domain equivalent topic; Broadcast genre; Periodical subject; Video game content descriptor; ...	**Instance of**: Cognitive function; Knowledge; Cultural factor; Cultural barrier; Cognitive process; Cognitive ability; Cultural difference; Ability; Characteristic; **Attribute of:** Film; Area; Book; Publication; Magazine; Country; Work; Program; Media; City; ...

Knowledge Bases



breadth and density enable understanding

Concept Learning

China Brazil India

$\underbrace{\hspace{12em}}$

emerging market
country

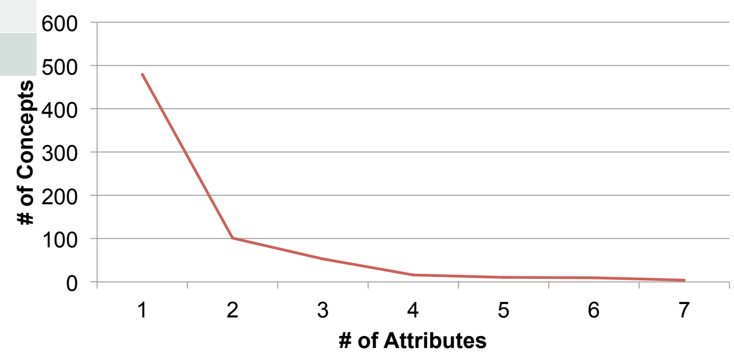
body smell taste

$\underbrace{\hspace{12em}}$

wine

Understanding Web Tables

	website	president	city	motto	state	type	director



china population

country

collector of fine china

earthenware

Bayesian

$$P(c_k|E) = \frac{P(E|c_k)P(c_k)}{P(E)} \quad \propto P(c_k)\prod_{i=1}^{M} P(e_i|c_k).$$

- For a mixture of instances and properties: Noisy-Or model

$$P(c|t_l) = 1 - \big(1 - P(c|t_l, z_l = 1)\big)\big(1 - P(c|t_l, z_l = 0)\big)$$

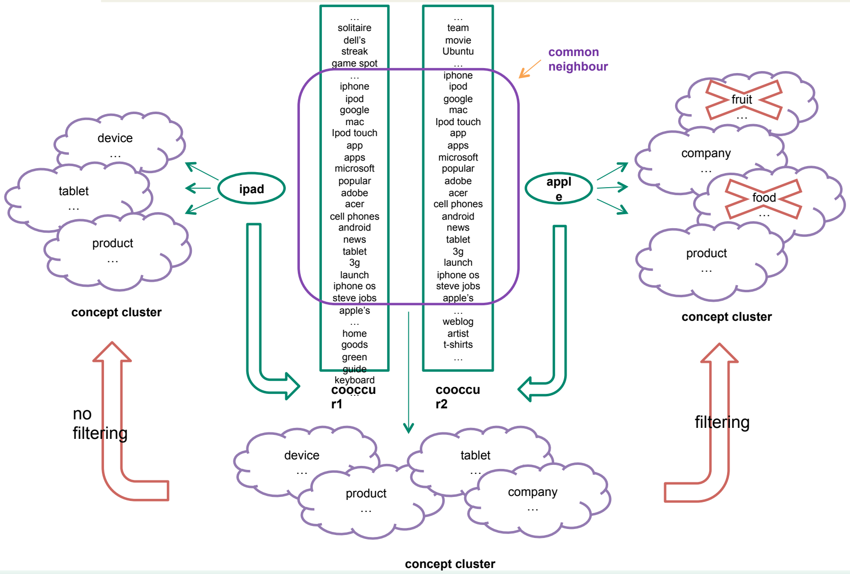
Where $z_l = 1$ indicates t_l is an entity, $z_l = 0$ indicates t_l is a property

- Bayesian rule gives:

$$P(c|T) \;\propto P(c)\prod_l^L P(t_l|c) \propto \frac{\prod_l P(c|t_l)}{P(c)^{L-1}}$$

apple iPad

company device

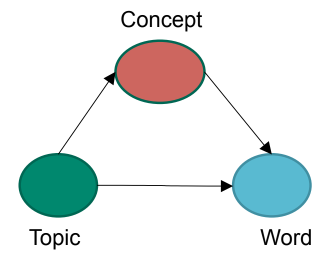


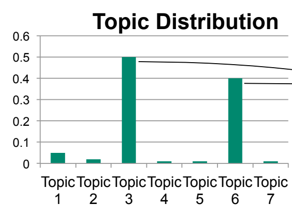
Modeling Co-occurrence

Probase

+ LDA model

Wikipedia

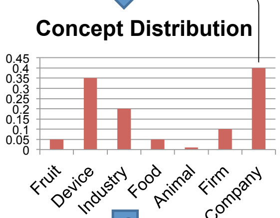




$p(c|t)$

Company	Prob
Apple	0.214123
Google	0.122754
Microsoft	0.089717
affiliate company	0.073715
Sony	0.048214
ISP	0.038612
Internet Service Provider	0.036651
web host	0.036341
Nintendo	0.034999
HP	0.033760
Blizzard	0.031798
Toyota	0.028598

$p(w|c)$



$p(w|c)$

$p(w|topic)$

Apple, iPhone

Topic 3	Prob	Topic 6	Prob
software	0.0260	company	0.1068
windows	0.0224	business	0.0454
system	0.0184	companies	0.0186
version	0.0175	inc	0.0167
file	0.0172	corporation	0.0139
user	0.0141	market	0.0138
support	0.0115	founded	0.0136
microsoft	0.0114	based	0.0136
os	0.0098	sold	0.0132
computer	0.0097	industry	0.0127
based	0.0089	products	0.0126
available	0.0088	firm	0.0125
mac	0.0085	group	0.0124
source	0.0081	owned	0.0112
linux	0.0079	first	0.0111
operating	0.0077	largest	0.0101
open	0.0073	management	0.0091
released	0.0072	new	0.009
server	0.0069	million	0.009
release	0.0066	acquired	0.0085

$p(w|topic)$

- Infer topics z from text s using collapsed Gibbs sampling:

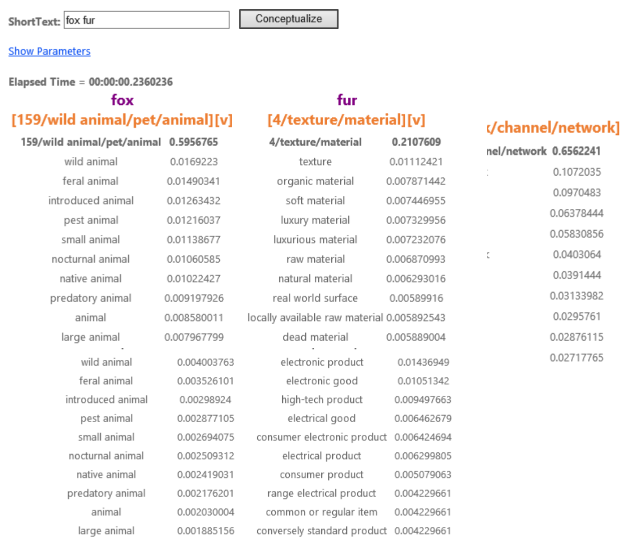
$$p(z_i = k|\vec{s}, z_{-i}, C) \propto (n_{\cdot k} + \alpha) \times \frac{C_{s_i k} + n_{s_i k} + \beta}{\sum_w C_{wk} + n_{wk} + |W|\beta},$$

- Estimate the concept distribution for each term w in s:

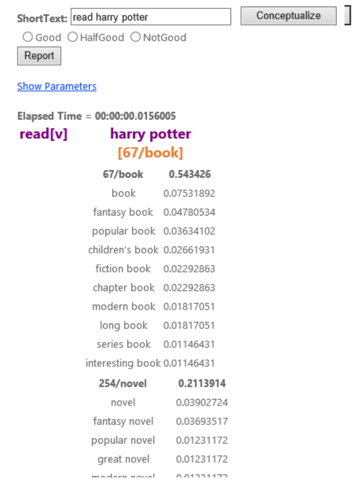
$$p(c|w, z) \propto p(c|w) \sum_k \pi_{wk} \phi_{ck},$$

$$\phi_{ck} = \frac{C_{ck} + \beta}{\sum_w C_{wk} + |W|\beta},$$

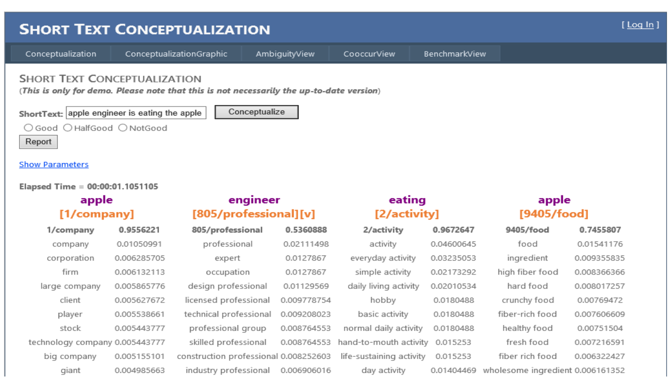
Examples



Examples



Examples



Similarity between Two Short Texts

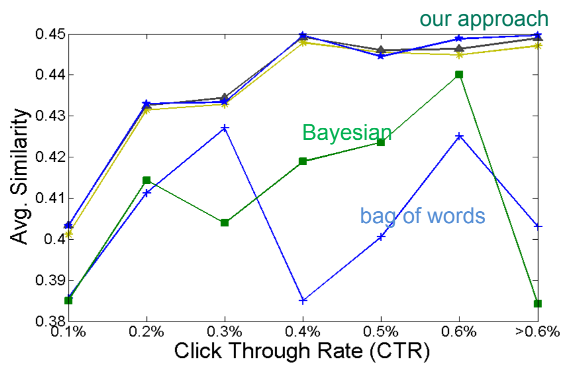
Search and URL title:

	Bayesian	LDA	LDA +Probase
T100	0.31 (0.29)	0.55 (0.31)	0.42 (0.39)
T200		0.52 (0.31)	0.42 (0.39)
T300		0.50 (0.31)	0.43 (0.40)

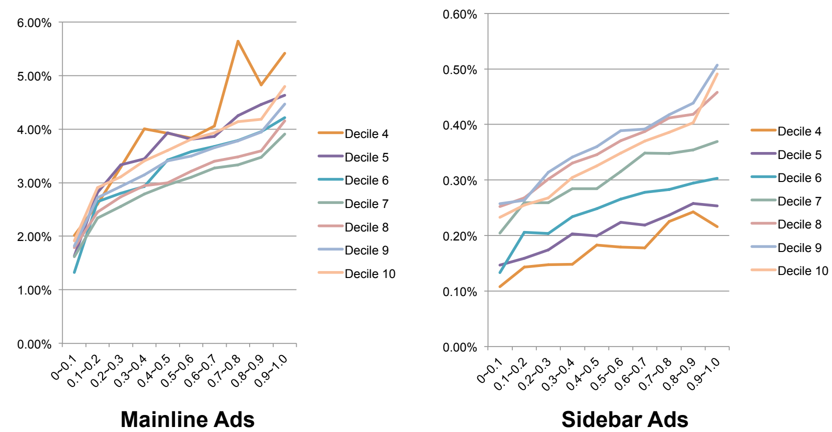
Two random searches:

	Bayesian	LDA	LDA +Probase
T100	0.02	0.24	0.03
T200		0.21	0.03
T300		0.19	0.03

CTR and search/ads similarity



CTR and search/ads similarity (torso and tail queries)



FrameNet Sentences

	Basic	Context Sensitive T100	T200	T300
Fold 1	-4.716	-3.401	-3.385	-3.378
Fold 2	-4.728	-3.409	-3.393	-3.389
Fold 3	-4.741	-3.432	-3.417	-3.410
Fold 4	-4.727	-3.413	-3.399	-3.392
Fold 5	-4.740	-3.433	-3.417	-3.413

Log-likelihood of frame elements with five-fold validation.

Conclusion

- Knowledge is needed in learning
- Knowledge is probabilistic
- (Short) Text understanding
  - Syntax (from NLP parser)
  - Dictionary (from an entity store)
  - Probabilistic knowledge