ACL 2011
Portland, Oregon

June 2011

Web Search Queries as a Corpus

Tutorial at the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)

Marius Paşca

Google Inc.
mars@google.com

Overview

- Part One: Introduction
- Part Two: Queries as a Corpus
- Part Three: Extraction from Queries

Part One: Introduction

- Open-domain information extraction
- Instances, concepts, relations

Unweaving the World Wide Web of Facts

- The Web is a repository of implicitly-encoded human knowledge
  - some text fragments contain easier-to-extract knowledge
- More knowledge leads to better answers
  - acquire facts from a fraction of the knowledge on the Web
  - exploit available facts during search
- Open-domain information extraction
  - extract knowledge (facts, relations) applicable to a wide range, rather than closed, pre-defined set of domains (e.g., medical, financial etc.)
  - no need to specify set of concepts and relations of interest in advance
  - rely on as little manually-created input data as possible

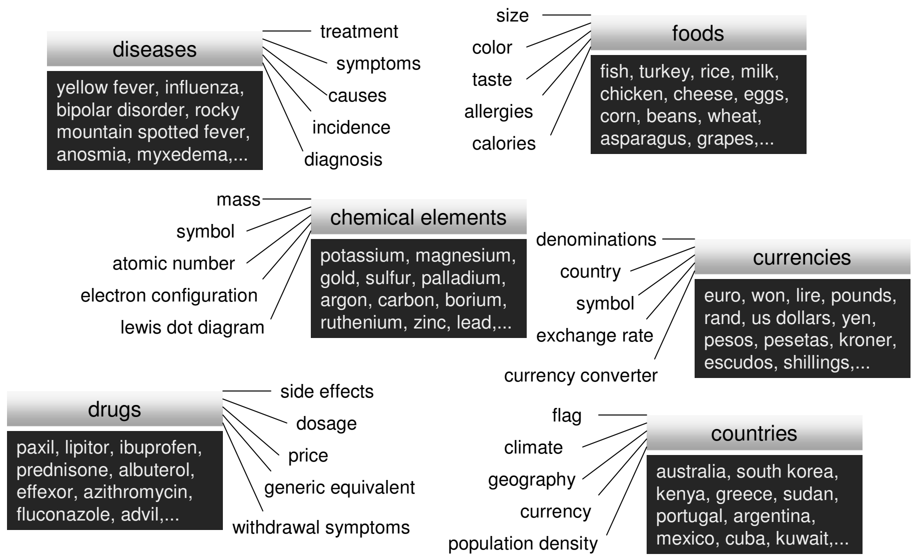
Instances, Concepts and Relations

- A concept (class) is a placeholder for a set of instances (objects) that share similar properties
  - set of instances
    - {matrix, kill bill, ice age, pulp fiction, inception, cidade de deus,...}
  - class label
    - movies, films
  - definition
    - a series of pictures projected on a screen in rapid succession with objects shown in successive positions slightly changed so as to produce the optical effect of a continuous picture in which the objects move (Merriam Webster)
    - a form of entertainment that enacts a story by sound and a sequence of images giving the illusion of continuous movement (WordNet)

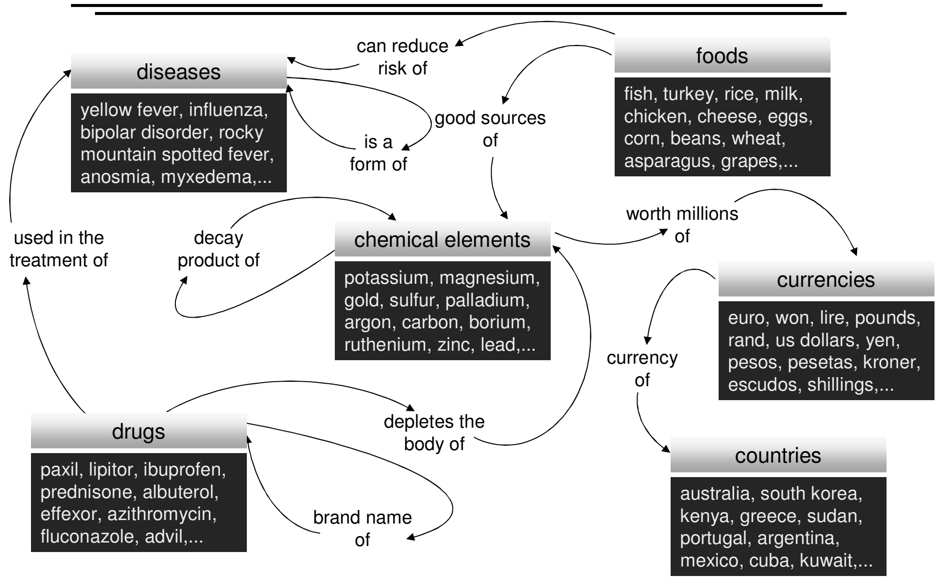
Instances, Concepts and Relations

- Relations are assertions linking two (binary relation) or more (n-ary relation) concepts
  - actors-act in-movies; cities-capital of-countries
- Facts are instantiations of relations, linking two or more instances
  - leonardo dicaprio-act in-inception; cairo-capital of-egypt
- Attributes correspond to facts capturing quantifiable properties of a class or an instance
  - actors --> awards, birth date, height
  - movies --> producer, release date, budget

Open-Domain Information



Open-Domain Information



Terminology and Scope

- Terminology
  - concept vs. class: used interchangeably
  - instance vs. entity: used interchangeably
- Scope
  - discussing methods using queries to extract open-domain information
  - not discussing methods using queries in other tasks such as Web search in general (e.g., query suggestion, spelling correction, improving search results)

Sources of Open-Domain Information

- Human-compiled knowledge resources
  - resources created by experts
  - resources created collaboratively by non-experts
- Sources of textual data
  - text documents (unstructured or semi-structured text)
  - (Web) search queries

Expert Resources

- WordNet
  - [Fel98]: C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press 1998.
  - lexical database of English created by experts
  - wide-coverage of upper-level conceptual hierarchies
  - replicated or extended to other languages
- Cyc
  - [Len95]: D. Lenat. CYC: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 1995.
  - knowledge base of common-sense knowledge created by experts over 100+ person-years
  - terms and assertions capturing ground assertions and (inference) rules

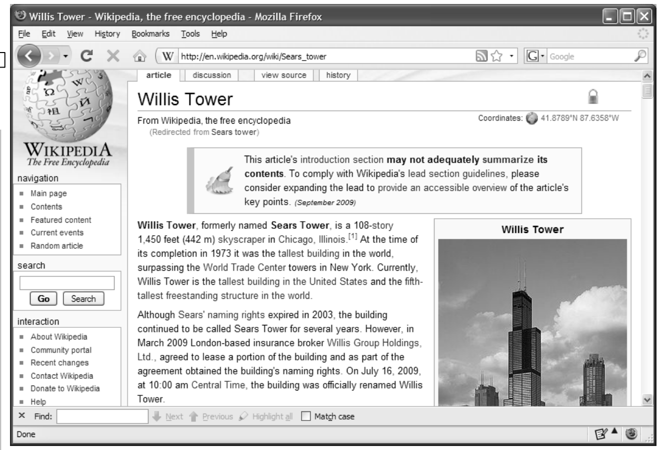
Collaborative, Non-Expert Resources

- Wikipedia
  - [Rem02]: M. Remy. Wikipedia: The Free Encyclopedia. Journal of Online Information Review 2002.
  - free online encyclopedia developed collaboratively by Web volunteers
  - among top 20 most popular Web sites (according to comScore: Top 50 US Web Properties, Aug 2009)
- DBpedia
  - [BLK+09] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer et al. DBpedia – A Crystallization Point for the Web of Data. Journal of Web Semantics 2009.
  - community effort to convert Wikipedia articles into structured data
  - manually-created ontology, mappings from subset of Wikipedia infoboxes to ontology, mappings from Wikipedia articles to WordNet concepts
- Freebase
  - [BEP+08]: K. Bollacker, C. Evans, P. Paritosh et al. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. SIGMOD-08.
  - repository for storing structured data from Wikipedia and other sources, as well as from user contributions
  - collaboratively created, structured and maintained
- Open Mind
  - [SLM+02]: P. Singh, T. Lin, E. Mueller, G. Lim, T. Perkins and W. Zhu. Open Mind Common Sense: Knowledge Acquisition from the General Public. Lecture Notes In Computer Science 2002.
  - collect common-sense knowledge from non-expert Web users
  - unlike Cyc, collect and represent knowledge in natural language rather than through formal assertions

Wikipedia

Wikipedia infobox

Preceded by	World Trade Center
Surpassed by	Petronas Twin Towers
Information	
Location	233 S. Wacker Drive Chicago, Illinois 60606 United States
Status	Complete
Constructed	1970-1973
Use	Office, observation, communication
Height	
Antenna/Spire	1,730 feet (527 m)
Roof	1,451 ft (442 m)[1]
Technical details	
Floor count	108[1]
Floor area	4.56 million sq ft. (3.81 million sq ft. rentable) 418,064 m² (353,961 m² rentable)[2]
Elevator count	104, with 16 double-decker elevators, made by Westinghouse, modernized by Schindler Group
Companies	
Architect	Skidmore, Owings and Merrill



Wikipedia article

Categories: 1973 architecture | Buildings and structures on U.S. Route 66 | Visitor attractions along U.S. Route 66 | Former world's tallest buildings | Landmarks in Chicago, Illinois | Skyscrapers in Chicago, Illinois | Skyscrapers over 350 meters | Retail company headquarters in the United States | Office buildings in Chicago, Illinois | Skidmore, Owings and Merrill buildings

DBpedia, Freebase

Wikipedia infobox

Preceded by	World Trade Center
Surpassed by	Petronas Twin Towers
Information	
Location	233 S. Wacker Drive Chicago, Illinois 60606 United States
Status	Complete
Constructed	1970-1973
Use	Office, observation, communication
Height	
Antenna/Spire	1,730 feet (527 m)
Roof	1,451 ft (442 m)[1]
Technical details	
Floor count	108[1]
Floor area	4.56 million sq ft. (3.81 million sq ft. rentable) 418,064 m² (353,961 m² rentable)[2]
Elevator count	104, with 16 double-decker elevators, made by Westinghouse, modernized by Schindler Group
Companies	
Architect	Skidmore, Owings and Merrill

Wikipedia infobox source code

```
{{Infobox Skyscraper
|building_name=Willis Tower
|image=[[Image:Sears Tower ss.jpg|center|256px]]
|year_built=1974
|previous_building = [[World Trade Center]]
|surpassed_by_building = [[Petronas Twin Towers]]
|year_highest=1974
|year_end=1998
|location=233 S. Wacker Drive<br />[[Chicago]], [[Illinois]] 60606<br />{{USA}}
|use=Office, observation, communication
|height_stories=108<ref name="emporis">The tower has 108 stories as counted by
standard methods, though the building's owners count the main roof as 109 and
the mechanical penthouse roof as 110. [http://www.emporis.com/en/wm
/bu/?id=117064 Emporis.com] Retrieved on June 7, 2008</ref>
|construction_period=1970-1973
|emporis_id=117064
|roof=1,451 ft (442 m)<ref name="emporis" />
|top_floor=
|antenna_spire={{convert|1730|ft|m|0}}
|floor_area=4.56 million sq ft. (3.81 million sq ft. rentable) <br/> 418,064 m²
(353,961 m² rentable)<ref name="autogenerated1">{{cite
web|url=http://www.searstower.org/home.html |title=Welcome to Sears Tower
(Chicago, Illinois) |publisher=Searstower.org |date= |accessdate=2009-09-14}}
```

<Sears_Tower, previous_building, World_Trade_Center>
<Sears_Tower, construction_period, 1970-1973>
...

DBpedia entries

Quantitative Comparison of Human-Compiled Resources

- Wikipedia
  - 3.5+ million articles in English
  - articles also available in 200+ other languages
- DBpedia
  - 2.5+ million instances, 250+ million relations
- Freebase
  - 20+ million instances, 300+ million relations
- Cyc
  - ResearchCyc: 300,000+ concepts and 3+ million assertions
  - OpenCyc 2.0: add mappings from Cyc concepts to Wikipedia articles
- Open Mind
  - 800,000+ facts in English
  - facts also available in other languages

Sources of Open-Domain Information

- Human-compiled knowledge resources
  - resources created by experts
  - resources created collaboratively by non-experts
- <u>Sources of textual data</u>
  - text documents (unstructured or semi-structured text)
  - (Web) search queries

Documents

Unstructured text

Semi-structured text

Preceded by	World Trade Center
Surpassed by	Petronas Twin Towers
Information	
Location	233 S. Wacker Drive Chicago, Illinois 60606 United States
Status	Complete
Constructed	1970–1973
Use	Office, observation, communication
Height	
Antenna/Spire	1,730 feet (527 m)
Roof	1,451 ft (442 m)[1]
Technical details	
Floor count	108[1]
Floor area	4.56 million sq ft. (3.81 million sq ft. rentable) 418,064 m² (353,961 m² rentable)[2]
Elevator count	104, with 16 double-decker elevators, made by Westinghouse, modernized by Schindler Group
Companies	
Architect	Skidmore, Owings and Merrill

Documents

Semi-structured text

F

- Faroe Islands - Tórshavn
- Finland - Helsinki
- France - Paris

G

- Georgia - Tbilisi
- Germany - Berlin
- Greece - Athens

Semi-structured text

English Short Name	English Long Name	Domestic Short Name	Domestic Long Name	Capital
France	French Republic	French: *France*	French: *République française*	Paris
Georgia[1]	Republic of Georgia		Georgian: საქართველო Georgian Transliteration: *Sakartvelo*	Tbilisi Georgian: თბილისი
Germany	Federal Republic of Germany	German: *Deutschland*	German: *Bundesrepublik Deutschland*	Berlin

Alternative to Documents

- Conventionally: data for textual information extraction is available as (some sort of) a document collection
  - documents capture knowledge, or assertions about the world
  - assertions are often "hidden" in expository text
  - the goal is to derive some of that knowledge from text
- Alternatively: textual information extraction may be pursued even without a document collection
  - to find new knowledge within a document collection, users formulate their search queries based on the knowledge that they already possess at the time of the search
    - --> query logs collectively capture knowledge, through requests that may be answered by knowledge asserted in document collections

Next Topic

- Part One: Introduction
- <u>Part Two: Queries as a Corpus</u>
- Part Three: Extraction from Queries

Queries as a Corpus

- Structure of queries
- Comparison with other textual sources
- Usage, demographics and privacy

Structure of Queries

- [SW07]: S. Bergsma and Q. Wang. Learning Noun Phrase Query Segmentation. EMNLP-07.
  - identify segments of contiguous query tokens corresponding to semantic concepts, using manually annotated queries as training data
- [TP08]: B. Tan and F. Peng. Unsupervised Query Segmentation Using Generative Language Models and Wikipedia. WWW-08.
  - identify segments of contiguous query tokens corresponding to semantic concepts, using evidence from queries and from Wikipedia documents
- [BJR08]: C. Barr, R. Jones and M. Regelson. The Linguistic Structure of English Web-Search Queries. EMNLP-08.
  - identify structural characteristics of queries in the task of part of speech tagging
- [ML09]: M. Manshadi and X. Li. Semantic Tagging of Web Search Queries. ACL-IJCNLP-09.
  - classify queries into domains, and identify query fragments corresponding to pre-specified, per-domain schema of tags
- [GXC+09]: J. Guo and G. Xu and X. Cheng and H. Li. Named Entity Recognition in Query. SIGIR-09.
  - detect instances within queries, and classify instances into coarse-grained classes
- [Li10]: X. Li. Understanding the Semantic Structure of Noun Phrase Queries. ACL-10.
  - represent noun phrase queries as a combination of intent heads and intent modifiers, and identify those components automatically

Finding Structure in Queries

- [BJR08]: C. Barr, R. Jones and M. Regelson. The Linguistic Structure of English Web-Search Queries. EMNLP-08.

Part-of-Speech Tags of Query Tokens

- Task
  - investigate the task of part-of-speech (POS) tagging when applied to queries
- Input data
  - set of 3.2K (2.5K unique) Web search queries, after automatic spell checking and tokenization
- Manual annotation of POS tags of query tokens is unreliable
  - inter-annotator agreement: 0.79 (token-level), 0.65 (query-level)
  - main cause of annotation errors (70% of cases): actual query ambiguity (e.g., download may be a noun or a verb) rather than human annotation mistakes
- POS tags have a different distribution in queries than in documents
  - in documents (Brown corpus): ~90 distinct tags, of which 15 for determiners, and 35 for verbs
  - in queries: ~20 distinct tags are sufficient, of which 1 for determiners and 1 for verbs

Suggested Part-of-Speech Tags

Part-of-Speech Tag	Example Token	Percentage of Query Tokens
proper noun	texas	40.2%
common noun	pictures	30.9%
adjective	big	7.1%
URI	ebay.com	5.9%
preposition	in	3.7%
unknown	y	2.5%
verb	get	2.4%
...	...	...

(Courtesy R. Jones)

- Nouns are predominant in queries
  - most frequent tags in documents: 13% of tokens are common nouns
  - most frequent tags in queries: 40% of tokens are proper nouns, 71% of tokens are common nouns or proper nouns
- Verbs are infrequent in queries
  - in documents: at least one verb in most sentences
  - in queries: less than 3% of tokens

Part-of-Speech Tagging Experiments

- Use of capitalization in queries is inconsistent
  - 17% queries contain capitalization, of which 4% are all-caps
  - when a query contains mixed capitalization, first-letter token capitalization is indicative of an actual proper noun for 73% of cases
  - other uses of capitalization in queries: acronyms, capitalization for first token of query, first-letter capitalization for all tokens

  --> cannot rely on capitalization to identify proper nouns in queries

Experimental Setting	Per-Token Tagging Accuracy
tagger that assigns most frequent tag (over separate training lexicon) of each token	65.4%
tagger trained on annotated documents	48.2%
tagger trained on annotated queries	69.7%
tagger trained and evaluated on queries with perfect capitalization	89.4%
tagger trained and evaluated on queries with automatically-induced capitalization	70.9%

Comparison with Other Textual Sources

- [CGC+09]: M. Carman, R. Gwadera, F. Crestani and M. Baillie. A Statistical Comparison of Tag and Query Logs. SIGIR-09.
  - investigate similarity between vocabularies of tokens from search queries vs. tags assigned by users to Web documents
- [GNL+10]: J. Gao, P. Nguyen, X. Li, C. Thrasher, M. Li and K. Wang. A Comparative Study of Bing Web N-gram Language Models for Web Search and Natural Language Processing. SIGIR 2010, Web N-gram Workshop.
  - generate a repository of n-grams from Web data, including from queries, and evaluate it in various text processing tasks

Characteristics of Documents vs. Queries

Characteristic	Data Source	
	Document Sentences	Queries
Type of medium	text	text
Purpose	convey info.	request info.
Available context	surrounding text	self-contained
Average quality	high (varies)	low
Grammatical style	natural language	bag of keywords
Average length	25 words or more	2-3 words

Queries vs. Other Textual Sources

- [CGC+09]: M. Carman, R. Gwadera, F. Crestani and M. Baillie. A Statistical Comparison of Tag and Query Logs. SIGIR-09.

Queries vs. Tags

- Task
  - investigate the similarity between query logs and user-generated tags (entered by users to annotate documents)
- Input data
  - from query logs containing click-through data, and from Delicious (social bookmark) tags, select queries and tags associated with a set of 4K Web documents
  - each document clicked at least 50 times, and associated with a tag at least 20 times
  - generate respective vocabularies (i.e., sets) of tokens for tags and queries, after removing stop words and stemming all tokens with the Porter stemmer

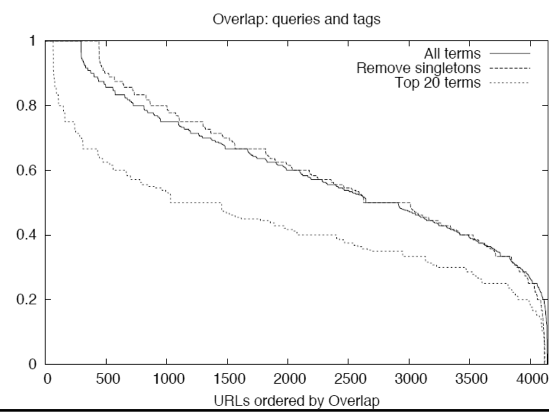
Metric	Token Occurrences		Vocabulary Size	
	Queries	Tags	Queries	Tags
Mean	955.3	1105.8	17.6	139.6
Std deviation	6464.7	1533.4	12.8	137.7
Median	278.0	393.0	15.0	83.0

Query vs. Tag Vocabulary

- Compute overlap between query tokens and tag tokens

$$Overlap(url) = \frac{|V_q \cap V_\tau|}{\min(|V_q|, |V_\tau|)}$$

- Optionally, remove low frequency tokens or keep high frequency tokens



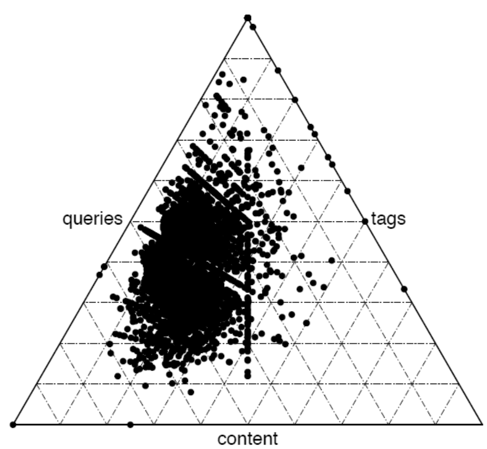
- Over more than half of documents, overlap ≥ 0.5

--> query vocabulary is very similar to tag vocabulary

(Courtesy M. Carman)

Query vs. Tag vs. Document Vocabulary

- Include vocabulary of Web documents in comparison of relative overlap



- Similarity between query and document vocabulary is higher than between query and tag vocabulary
  - since documents are clicked search results, they are likely to contain query tokens
- Similarity is lowest between tag and document vocabulary
  - users do not necessarily enter tags that appear in document content

(Courtesy M. Carman)

Repositories of Distilled Query Data

- [GNL+10]: J. Gao, P. Nguyen, X. Li, C. Thrasher, M. Li and K. Wang. A Comparative Study of Bing Web N-gram Language Models for Web Search and Natural Language Processing. SIGIR 2010, Web N-gram Workshop.

Web N-Gram Collection

- Language models of n-grams, from Web documents and search queries

N-gram Length	Documents			Queries
	Body	Anchor Text	Title	
1-grams	1.2B	60.3M	150M	251.5M
2-grams	11.7B	464.1M	1.1B	1.3B
3-grams	60.0B	1.4B	3.1B	3.1B
4-grams	148.5B	2.3B	5.1B	4.6B
5-grams	230.0B	N/A	N/A	N/A

- Language models found to be more similar between queries and document title (and queries and document anchor text) than between queries and document body

Queries as a Corpus

- Structure of queries
- Comparison with other textual sources
- Usage, demographics and privacy

Usage, Demographics and Privacy

- [MC08]: Q. Mei and K. Church. Entropy of Search Logs: How Hard is Search? With Personalization? With Backoff? WSDM-08.
  - investigate Web search from the perspective of entropy in search logs, and assess the impact of aggregated data about users (e.g., from IP addresses) on the outcome of Web search
- [JBS08]: B. Jansen and D. Booth and A. Spink. Determining the Informational, Navigational, and Transactional Intent of Web Queries. Journal of Information Processing and Management 2008.
  - investigate the distribution of queries from the point of view of intent type (and subtypes), and automatically classify queries accordingly
- [JBS09]: B. Jansen, D. Booth and A. Spink. Patterns of Query Reformulation During Web Searching. Journal of the American Society for Information Science and Technology 2009.
  - develop models to classify various types of query reformulations and identify the most frequent ones among Web users
- [WC10]: Ingmar Weber and Carlos Castillo. The Demographics of Web Search. Sigir-10.
  - study the impact of various user demographics factors on the users' choice of queries
- [JKP+07]: R. Jones, R. Kumar, B. Pang and A. Tomkins. "I Know What You did Last Summer": Query Logs and User Privacy. CIKM-07.
  - study the possibility of uncovering user identity from query logs, despite attempts to remove basic personally identifiable information from queries
- [GBG+10]: S. Goel, A. Broder, E. Gabrilovich and B. Pang. Anatomy of the Long Tail: Ordinary People with Extraordinary Tastes. WSDM-10.
- [KKM+09]: A. Korolova, K. Kenthapadi, N. Mishra and A. Ntoulas. Releasing Search Queries and Clicks Privately. WWW-09.
  - investigate methods to generate modified query log data that preserves user privacy

Query Usage

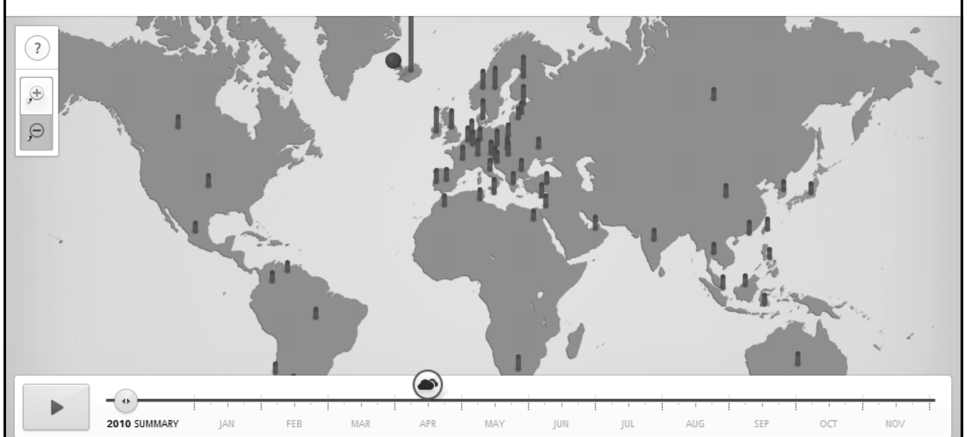
- Search zeitgeist
  - capture "the general intellectual, moral, and cultural climate of an era" (Merriam Webster), as reflected in the aggregation of search queries submitted by Web users

Top Global Events (2010)	Top Rising Queries (2010)	
	Entertainment	Consumer Electronics
world cup	justin bieber	ipad
olympics	shakira	iphone 4
haiti earthquake	eminem	nokia 5530
oil spill	netflix	htc evo 4g
ash cloud	youtube videos	nokia n900

(Google Zeitgeist)

Geographical Distribution

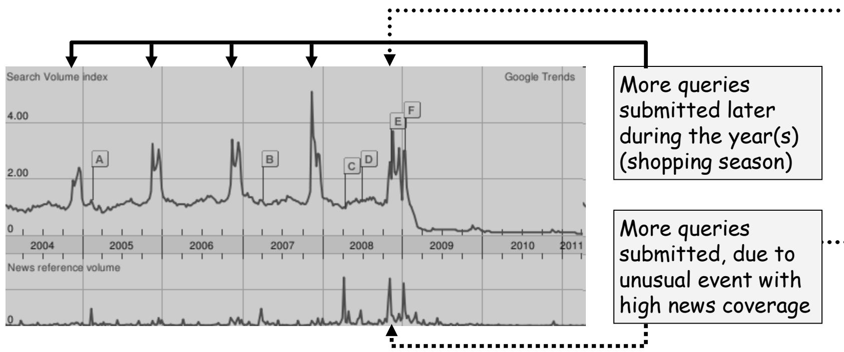
- For: ash cloud



(Google Zeitgeist)

Temporal Distribution

- For: circuit city



(Google Trends)

Query Demographics

- [WC10]: Ingmar Weber and Carlos Castillo. The Demographics of Web Search. Sigir-10.

Query Demographics

- Task
  - investigate impact of user demographics on Web search
- Input data
  - user profile data (birth year, gender, zip code)
  - set of pairs of (query, clicked URL) from query logs
  - census demographic data for various zip codes

(Courtesy I. Weber)

Feature	Query Log Data					US Avg.
	20%	40%	60%	80%	Avg.	
Per-capita income ($k)	16.0	18.9	22.4	27.7	22.7	21.6
Below poverty (%)	4.5	7.2	10.9	16.5	11.1	12.4
BA degree (%)	12.8	18.1	25.6	37.6	25.5	24.4
White (%)	61.9	78.8	88.1	94.4	76.9	75.1
Afric. Amer. (%)	0.9	2.4	5.7	15.5	4.0	12.3
Asian (%)	0.4	1.1	2.3	5.1	4.0	3.6
Non-English (%)	4.5	7.9	14.0	27.3	17.3	17.9
Year of birth	1956	1966	1974	1982	1968	1974

Role of Demographics in Web Search

- Highly-discriminant queries for various user demographics

Feature	Query
Per-capita income ($k)	chris jordan electric candle warmer www.popsugar.com ns4w.org
Below poverty (%)	www.unitnet.com slaker kipasa www.tokbox.com
BA degree (%)	spencer stuart executive search insight venture partners federal circuit four seasons jackson hole

(Courtesy I. Weber)

Role of Demographics in Web Search

- Highly-discriminant queries for various user demographics

Feature	Query
White (%)	pulloff.com central boiler wood furnace firewood processors midwest super cub
Afric. Amer. (%)	trey songz bio def jam records address s2s magazine madinaonline
Asian (%)	sina big bang lyrics tvb series jay chou lyrics

Role of Demographics in Web Search

- Highly-discriminant queries for various user demographics

Feature	Query
Year of birth, old	www.johnshopkinshealthalerts.com www.envisionreports.com/vz yahoo free bridge games bnymellon.mobular.net/bnymellon/frp
Year of birth, young	free teen chatrooms wet seal tottaly layouts photofiltre brushes

Queries and User Privacy

- [JKP+07]: R. Jones, R. Kumar, B. Pang and A. Tomkins. "I Know What You did Last Summer": Query Logs and User Privacy. CIKM-07.

Queries and User Privacy

- Task
  - investigate the vulnerability of narrowing down the identify (demographics) of users submitting search queries, even after removal of personally identifiable information (names, numbers) from query logs
- Input data
  - from user profile data (anonymized id, birth year, gender, zip code), select 100M profiles
  - from query logs, select query sessions issued by users with available profile data, for 744K users
- Assessment of vulnerability
  - arrange data into buckets by age, gender, zip code
  - arrange buckets into bins, by conjunctions of age, gender, zip code
  - smaller bin size makes it easier to identify a particular user from the bin (especially when additional information, e.g., hobbies, is available about the user)
  - e.g., if input data is arranged into bins that share gender bucket, age bucket, and first 3 of 5 zip code digits (e.g., males, age 25-29, living in zip code 950xx) --> almost 100K of the 744K users fit into a bin of 100 users or less

Deriving Demographics from Queries

- Identifying user gender and age
  - classifiers using bag-of-words features
  - gender identification: accuracy of 83.8%
    - examples of discriminative features: {bridal, makeup, hair, women's,..} for women; {nfl, poker, male, compusa,..} for men
  - age identification: absolute error of 7 years (predicted vs. actual), better than always guessing the middle age point
    - examples of discriminative features: {myspace, pregnancy, wikipedia, mall,..} for lower age; {aarp, lottery, amazon.com, senior, repair,..} for higher age
  - if personally identifiable information (names and numbers) are removed from queries, both gender and age classification remain about as accurate
- Identifying location (zip code)
  - existing classifier for locations: given query as input, output list of locations
  - convert list of locations into zip code buckets of known first 3, 4 or 5 digits

Known Digits of Zip Code	First 5	First 4	First 3
Correct at top one	6.2%	13.7%	34.9%
Correct among top three	13.1%	251.%	54.1%

  - if personally identifiable information (names and numbers) are removed from queries, location classification becomes much less accurate

Deriving Queries from Known Information

- Identifying query sessions submitted by a known user
  - use demographics, conversations with, lifestyle changes of user, in order to guess queries that may have been submitted by user
  - as an approximation, manually create a set of guessed queries

Category	Common	Rare
Cars	volkswagen beetle (478) honda odyssey (1504) toyota prius (1070)	triumph tr23 (23) e-type jaguar (5)
Sports	skiing (9618) football (123802)	bassmaster (388) skulling (17)
Food	pizza (104888) italian restaurant (4998) brie (39325)	assam (747)
Books	harry potter (27838) danielle steele (238) freakonomics (574)	holly lisle (20) elizabeth moon (27)

Knowing that a user submitted the query e-type jaguar narrows down the identity of the user to a bin of 5 possible users

  - use combinations of guessed queries

(Courtesy R. Jones)

Deriving Queries from Known Information

Query Combination	Bin Size
harry potter, pizza	4855
football, skiing	2430
italian restaurant, pizza	1441
harry potter, volkswagen beetle	27
...	...
pizza, triumph tr3	2
brie, holly lisle, pizza	1
danielle steele, volkswagen beetle	1

--> even if individual bits of information are far from unique among users, putting them together can uniquely identify a user

Next Topic

- Part One: Introduction
- Part Two: Queries as a Corpus
- <u>Part Three: Extraction from Queries</u>

Extraction Methods

- Methods for extraction of:
  - instances and concepts
  - attributes and relations

Instances and Concepts

diseases
yellow fever, influenza, bipolar disorder, rocky mountain spotted fever, anosmia, myxedema,...

foods
fish, turkey, rice, milk, chicken, cheese, eggs, corn, beans, wheat, asparagus, grapes,...

chemical elements
potassium, magnesium, gold, sulfur, palladium, argon, carbon, borium, ruthenium, zinc, lead,...

currencies
euro, won, lire, pounds, rand, us dollars, yen, pesos, pesetas, kroner, escudos, shillings,...

drugs
paxil, lipitor, ibuprofen, prednisone, albuterol, effexor, azithromycin, fluconazole, advil,...

countries
australia, south korea, kenya, greece, sudan, portugal, argentina, mexico, cuba, kuwait,...

Instances and Concepts

- [Pas07]: M. Paşca. Weakly-Supervised Discovery of Named Entities using Web Search Queries. CIKM-07.
  - expand sets of instances using Web search queries
- [VP08]: B. Van Durme and M. Paşca. Finding Cars, Goddesses and Enzymes: Parametrizable Acquisition of Labeled Instances for Open-Domain Information Extraction. AAAI-08.
  - extract labeled sets of instances from Web documents, by merging clusters of distributionally similar phrases with IsA pairs extracted with lexico-syntactic patterns
- [PP09]: M. Pennacchiotti and P. Pantel. Entity Extraction via Ensemble Semantics. EMNLP-09.
  - expand sets of instances using multiple sources of text including queries
- [AHH09]: E. Alfonseca and K. Hall and S. Hartmann. Large-Scale Computation of Distributional Similarities for Queries. NAACL-HLT-2009.
  - apply vector-space model of distributional similarities to queries rather than documents
- [JP10]: A. Jain and P. Pantel. Open Entity Extraction from Web Search Query Logs. COLING-10.
  - extract clusters of distributionally similar phrases from Web search queries and click-through data

Instances and Concepts

- [VP08]: B. Van Durme and M. Paşca. Finding Cars, Goddesses and Enzymes: Parametrizable Acquisition of Labeled Instances for Open-Domain Information Extraction. AAAI-08.

Extraction from Documents and Queries

- Input
  - target relation, available as a small set of extraction patterns
    - e.g., <C [such as|including] I>
- Data sources
  - collection of Web documents
  - collection of anonymized Web search queries
- Output
  - sets of instances, each set associated with a class label
    - e.g., marine animals = {whales, seals, dolphins, turtles, sea lions, fishes, penguins, squids, pacific walrus, aquatic birds, comb jellies, starfish, florida manatees, walruses,...}
  - each set also associated with lists of attributes

Acquisition of Open-Domain Classes

- Define a closed vocabulary of potential class instances, as the set of most frequently-submitted Web search queries
  - textual data source: Web query logs
  - output: noisy set of potential class instances
- Acquire class labels for potential class instances, via hand-written extraction patterns
  - textual data source: Web documents
  - <C [such as|including] I>, where C is a potential class label (e.g., zoonotic diseases) and I is a potential instance (e.g., brucellosis)
  - output: noisy pairs of an instance and a class label
- Organize potential class instances into sets of distributionally similar phrases
  - output: noisy sets of distributionally similar instances

Merge into labeled sets of instances

Extraction of Labeled Instances

Input: - pairs of an instance and a class label
- unlabeled sets of distributionally similar instances

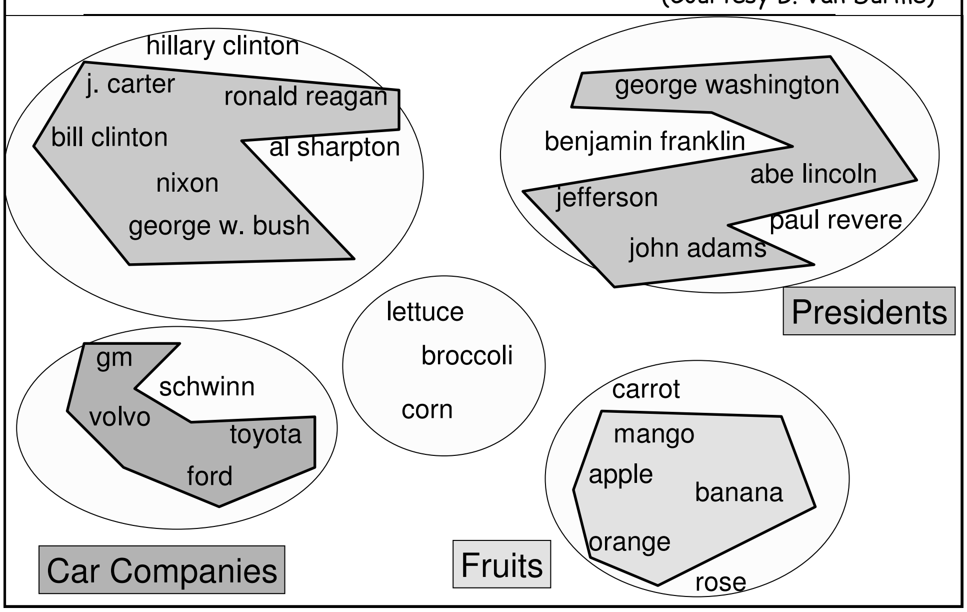
Output: - sets of instances, each set associated with a class label

```
For each unlabeled set of distributionally-similar instances S
    For each class label L assigned to some instance(s) of set S
tf  →  A=set of instances of S whose class label is L
idf →  B=set of sets that contain some instance(s) whose label is L
       If |A| > J×|S|:
          If |B| < K:
             Collect instances of A, associated with the class label L
```

- Note: J, K are weighting parameters controlling precision/recall
  - J in [0,1); higher J --> higher precision
  - K is non-negative integer; lower K --> higher precision

Patterns and Distributional Similarities

(Courtesy B. Van Durme)



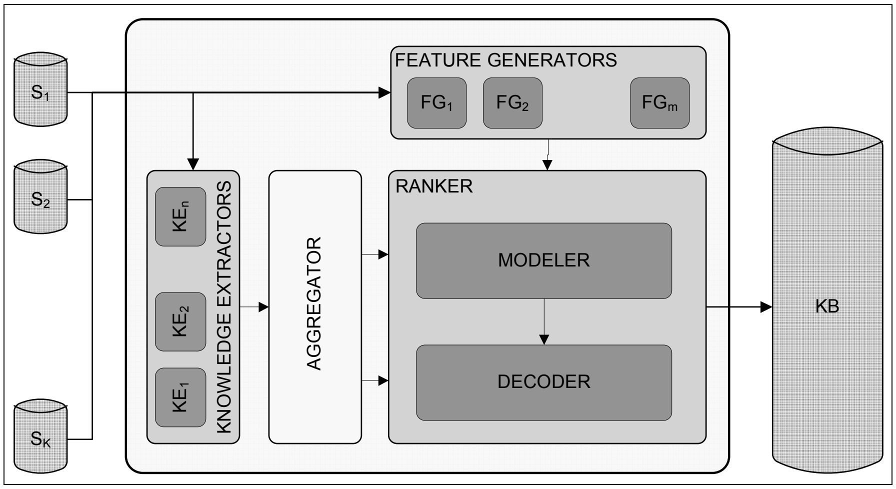
Instances and Concepts

- [PP09]: M. Pennacchiotti and P. Pantel. Entity Extraction via Ensemble Semantics. EMNLP-09.

Extraction from Multiple Sources

- Input
  - target classes, available as small sets of seed instances
    - e.g., {jodie foster, humphrey bogart, anthony hopkins} for Actor
  - target classes, also available as small sets of seed relations with other classes
    - e.g., < leonardo dicaprio, inception>, <nicole kidman, eyes wide shut> for Actor (corresponding to relation Actor-act in-Movie)
- Data sources
  - collection of Web documents
  - collection of Web search queries
  - HTML tables identified within the collection of Web documents
  - collection of articles from Wikipedia
- Output
  - ranked lists of instances, one per class
    - e.g., [gordon tootoosis, rosalind chao, john hawkes, jeffrey dean morgan,...] for Actor

Ensemble Semantics



(Courtesy P. Pantel, M. Pennacchiotti)

Extraction Components

- Sources (S_1, S_2,..., S_k)
  - data sources from which instances and their relevant features are extracted
- Knowledge extractors (KE_1, KE_2,..., KE_n)
  - extract candidate instances from sources, using various algorithms
- Feature generators (FG_1, FG_2,..., FG_m)
  - collect evidence/features relevant to deciding whether candidate instances are correct or not
- Aggregator
  - combine evidence available from multiple sources for candidate instances
- Ranker
  - rank candidate instances extracted by knowledge extractors, based on features available from feature generators

Ranking Features

- Collected by feature generators
  - 4 feature families: from Web documents, queries, tables, Wikipedia
  - 5 feature types: frequency, co-occurrence, distributional, pattern, termness (i.e., checking whether extracted terms are well-formed)

Family	*Type*		*Features*
Web (w)	Frequency	(wF)	term frequency; document frequency; term frequency as noun phrase
	Pattern	(wP)	confidence score returned by KE_{pat}; pmi with the 100 most reliable patterns used by KE_{pat}
	Distributional	(wD)	distributional similarity with the centroid in KE_{dis}; distributional similarities with each seed in $\mathcal{S}$
	Termness	(wT)	ratio between term frequency as noun phrase and term frequency; pmi between internal tokens of the instance; capitalization ratio
Query log (q)	Frequency	(qF)	number of queries matching the instance; number of queries containing the instance
	Co-occurrence	(qC)	query log pmi with any seed in $\mathcal{S}$
	Pattern	(qP)	pmi with a set of trigger words $\mathcal{T}$ (i.e., the 10 words in the query logs with highest pmi with $\mathcal{S}$)
	Distributional	(qD)	distributional similarity with $\mathcal{S}$ (vector coordinates consist of the instance's pmi with the words in $\mathcal{T}$)
	Termness	(qT)	ratio between the two frequency features F
Web table (t)	Frequency	(tF)	table frequency
	Co-occurrence	(tC)	table pmi with $\mathcal{S}$; table pmi with any seed in $\mathcal{S}$
Wikipedia (k)	Frequency	(kF)	term frequency
	Co-occurrence	(kC)	pmi with any seed in $\mathcal{S}$
	Distributional	(kD)	distributional similarity with $\mathcal{S}$

(Courtesy P. Pantel, M. Pennacchiotti)

Extraction Results

- Input data = collection of 600 million Web documents; tables identified within the documents; one year of queries; 2 million Wikipedia articles
- Evaluate lists of instances extracted for 3 classes: Actor, Athlete and Musician
  - create gold standard from samples of 500 instances selected randomly for each class
  - compute precision of extracted lists of instances, relative to and over the gold standards
- Average precision: 0.860 (Actor), 0.915 (Athlete), 0.788 (Musician)
- Precision@100: 0.99 (Athlete)
- Estimated precision@22000: 0.97 (Athlete)

Instances and Concepts

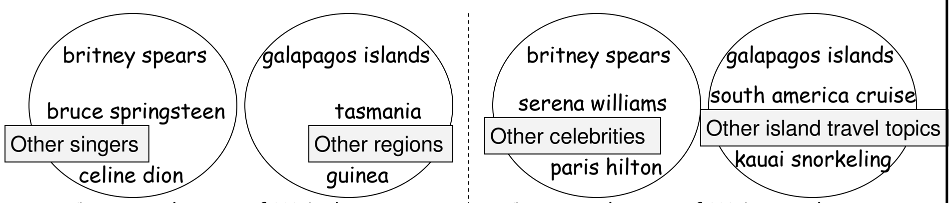
- [JP10]: A. Jain and P. Pantel. Open Entity Extraction from Web Search Query Logs. COLING-10.

Extraction from Queries

- Data sources
  - anonymized search queries along with frequencies and click-through data (clicked search results)
  - Web documents
- Output
  - clusters of similar instances
    - e.g., {basic algebra, numerical analysis, discrete math, lattice theory, nonlinear physics, ...}, {aaa insurance, roadside assistance, personal liability insurance, international driving permits, ...}
- Steps
  - collect set of candidate instances from queries
  - cluster instances using context in queries or click-through data or both

Similarity in Documents vs. Queries

- Contextual space of Web documents
  - an instance is represented by the contexts in which it appears in text documents
  - instances are modeled "objectively", according to descriptions of the world
- Contextual space of Web search queries
  - an instance is represented by the contexts in which it appears in a search queries
  - instances are modeled "subjectively", according to users' perception of the world



Contextual space of Web documents | Contextual space of Web search queries

Extraction of Instances

- Identify candidate instances
  - intuition: in queries composed by copying fragments from Web documents and pasting them into queries, capitalization of instances is preserved
  - from queries containing capitalization, extract contiguous sequences of capitalized tokens as instances

Queries	Candidate Instances
Britney Spears new song -->	Britney Spears
travel to Italy Roma -->	Italy Roma
restaurant Cascal in Mountain View -->	Cascal, Mountain View

- Retain set of best candidate instances
  - first criterion: promote candidate instances whose capitalization is frequent in Web documents
  - second criterion: promote candidate instances that occur as full-length queries

$$r_w(E) = \frac{|\gamma(E)|}{\sum_{i \in O(E)} |\gamma(i)|} \qquad s_q(E) = \frac{|Q == E|}{|queries\ that\ contain\ E|}$$

  - retain set of candidate instances that score highly (above some thresholds) according to both criteria

$$r_w(E) \geq \tau_r \ and\ s_q(E) \geq \tau_s$$

(Courtesy A. Jain)

Clustering of Instances

- Induce unlabeled classes of instances, by clustering instances using features collected from queries
  - as an alternative to collecting features from unstructured text in documents
  - for efficiency, no attempt to parse the queries
- Context features
  - vector of elements corresponding to contexts, where a context is the prefix and postfix around the instance, from queries containing the instance
- Click-through features
  - vector of elements corresponding to documents, where a document is one that is clicked by a user submitting the instance as a full-length query
- Hybrid features
  - normalized combination of context and click-through vectors

Impact of Clustering Features

- Given an instance, manually judge each co-clustered instance:
  - "If you were interested in instance I, would you also be interested in instance Ic in any intent?"
  - also, annotate with type of relation between instance and co-clustered instance
- Compute precision, over a set of evaluation instances
  - CL-CTX: context
  - CL-CLK: click-through
  - CL-HYB: hybrid
  - CL-Web: context collected from Web documents rather than queries

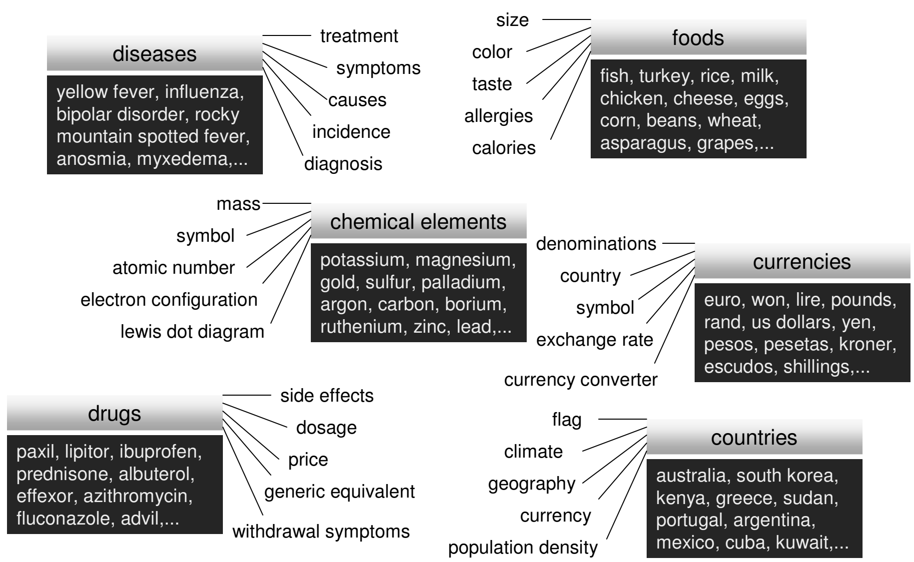
Method	Precision
CL-Web	0.73
CL-CTX	0.46
CL-CLK	0.81
CL-HYB	0.85

Relation Type	Method			
	CL-Web	CL-CTX	CL-CLK	CL-HYB
topic	0.27	0.46	0.46	0.40
sibling	0.72	0.43	0.29	0.32
parent	-	0.09	0.13	0.09
child	0.01	-	0.01	0.02
synonym	0.01	0.03	0.12	0.16

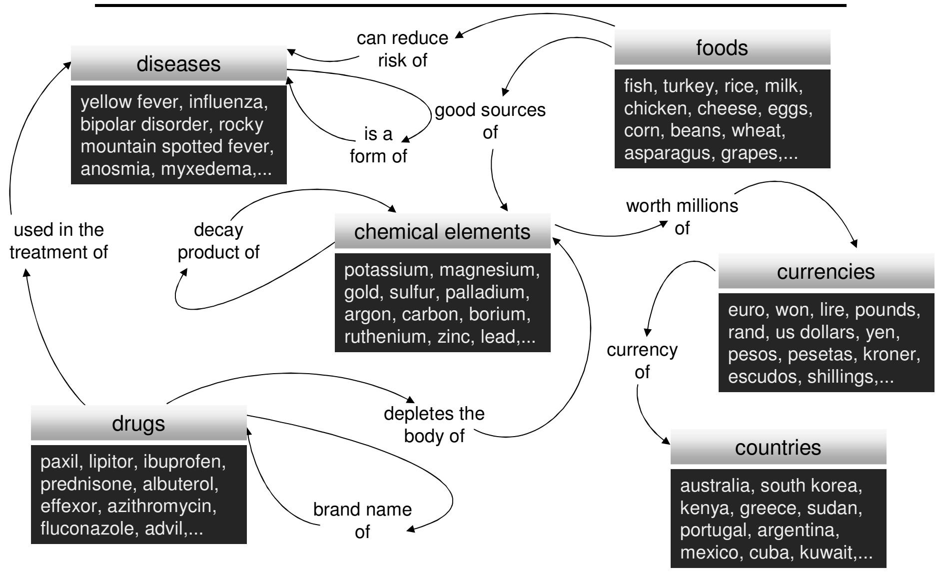
Extraction Methods

- Methods for extraction of:
  - instances and concepts
  - <u>attributes and relations</u>

Attributes and Relations



Attributes and Relations



Attributes and Relations

- [PV07]: M. Paşca and B. Van Durme. What You Seek is What You Get: Extraction of Class Attributes from Query Logs. IJCAI-07.
  - apply small set of patterns to extract attributes from queries
- [PVG07]: M. Paşca, B. Van Durme and N. Garera. The Role of Documents vs. Queries in Extracting Class Attributes from Text. CIKM-07.
  - apply patterns to extract attributes from unstructured text in documents vs. queries
- [Pas07]: M. Paşca. Organizing and Searching the World Wide Web of Facts - Step Two: Harnessing the Wisdom of the Crowds. WWW-07.
  - expand sets of seed attributes using queries
- [LWA09]: X. Li, Y. Wang and A. Acero. Extracting Structured Information from User Queries with Semi-Supervised Conditional Random Fields. SIGIR-09.
  - detect relevant fields in product-search queries, using click data and document content
- [PER+10]: M. Paşca, E. Alfonseca, E. Robledo-Arnuncio, R. Martin-Brualla and K. Hall. The Role of Query Sessions in Extracting Instance Attributes from Web Search Queries. ECIR-10.
  - extract attributes of instances, from sequences of queries within query sessions
- [YTT10]: X. Yin, W. Tan and Y. Tu. Automatic Extraction of Clickable Structured Web Contents for Name Entity Queries. WWW-10.
  - given a query containing an instance, extract structured data from click data and contents of subsequently visited documents
- [SJY11]: A. Das Sarma, A. Jain and C. Yu. Dynamic Relationship and Event Discovery. WSDM-11.
  - acquire temporally-anchored relations that apply within a given set of instances, using queries and (news) documents

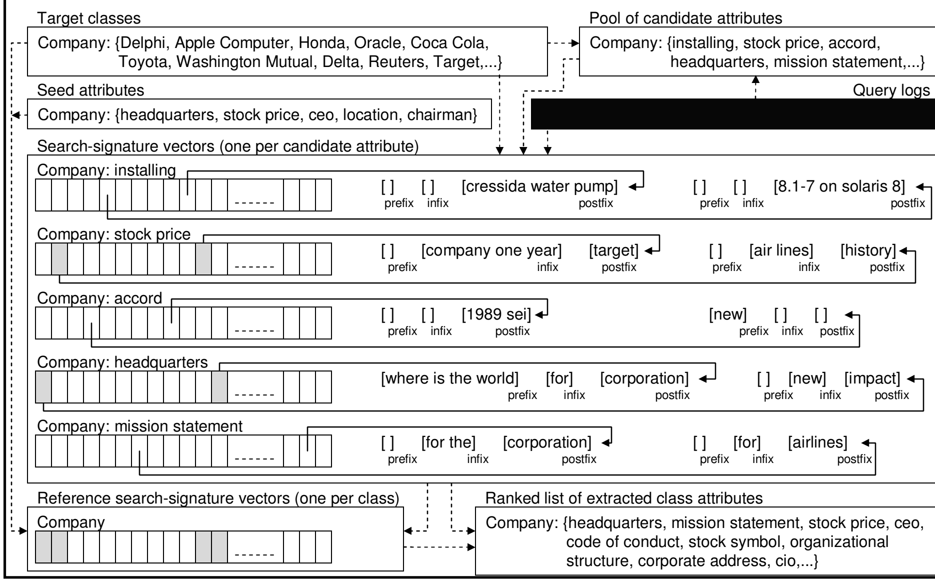
Attributes and Relations

- [Pas07]: M. Paşca. Organizing and Searching the World Wide Web of Facts - Step Two: Harnessing the Wisdom of the Crowds. WWW-07.

Extraction from Queries

- Input
  - target classes, available as sets of representative instances
    - e.g., {Delphi, Apple Computer, Honda, Oracle, Coca Cola, Toyota, Washington Mutual, Delta, Reuters, Target, ...} for Company
  - small sets of seed attributes, one per class
    - e.g., {headquarters, stock price, ceo, location, chairman} for Company
- Data source
  - anonymized search queries along with frequencies
- Output
  - ranked (longer) lists of attributes, one per class
    - e.g., {headquarters, mission statement, stock price, ceo, code of conduct, stock symbol, organizational structure, corporate address, cio, ...} for Company
- Steps
  - select candidate attributes, from queries containing an instance
  - create internal representation of candidate attributes, from queries containing an instance and a candidate attribute
  - rank candidate attributes, from similarity between internal representation of a candidate attribute and combined internal representation of all seed attributes

Class Attribute Extraction



Top Extracted Attributes

	Class	Top Extracted Attributes
1	Actor	awards, height, age, date of birth, weight, b** ****, birthdate, birthplace, cause of death, real name
2	AircraftModel	weight, length, history, fuel consumption, interior photos, specifications, photographs, interior pictures, seating arrangement, flight deck
3	Award	recipients, date, winners list, result, gossip, printable ballot, nominees, winners, location, announcements
4	BasicFood	calories, color, size, allergies, taste, carbs, nutritional information, nutrition facts, nutritional value, nutrition
5	CarModel	transmission, top speed, acceleration, transmission problems, owners manual, gas mileage, towing capacity, stalling, maintenance schedule, performance parts
6	CartoonChar	costume, voice, creator, first appearance, funny pictures, origins, cartoon images, cartoon pics, color pages
7	CellPhoneModel	features, battery life, retail price, mobile review, specification, price list, functions, ratings, tips, tricks
...	...	...

Top Extracted Attributes

	Class	Top Extracted Attributes
...	...	...
34	Stadium	location, seating capacity, architect, address, seating map, dimensions, tours, pics, poster, box office
35	TerroristGroup	attacks, leader, goals, meaning, website, leadership, photos, images, definition, flag
36	Treaty	countries, ratification, date, definition, summary, purpose, pros, cons, members, picture
37	University	alumni, mascot, dean, economics department, career center, graduation 2005, department of psychology, school colors, tuition costs, campus map
38	VideoGame	price, system requirements, creator, official site, official website, free game download, concept art, download demo, pc cheat codes, reviews
39	Wine	vintage, color, cost, style, taste, vintage chart, pronunciation, shelf life, wine ratings, wine reviews
40	WorldWarBattle	date, location, significance, images, importance, timeline, summary, pics, maps, photographs

Extraction Results

- Input data = 50 million anonymized queries
- Evaluate attributes extracted with hand-written patterns vs. based on seeds

	Class	Precision					
		@10		@20		@50	
		Patt	Seed	Patt	Seed	Patt	Seed
1	Actor	0.85	1.00	0.82	1.00	0.74	0.96
2	AircraftModel	0.80	0.80	0.77	0.85	0.68	0.71
3	Award	0.30	0.95	0.15	0.77	0.24	0.69
4	BasicFood	1.00	1.00	0.90	0.95	0.65	0.86
...	...	...	...	...	...	...	...
37	University	0.90	0.85	0.82	0.85	0.65	0.74
38	VideoGame	0.70	0.90	0.57	0.90	0.44	0.90
39	Wine	0.40	1.00	0.42	0.87	0.29	0.57
40	WorldWarBattle	0.00	0.85	0.00	0.82	0.00	0.66
	Average (40 Classes)	0.72	0.90	0.64	0.85	0.53	0.76

Summary

- Do ask, do tell
  - if knowledge is prominent, someone will eventually write about it
  - if knowledge is prominent, someone will eventually ask about it
  - Web search queries are cursory reflections of knowledge encoded deeply within unstructured and structured content available in documents
- Queries are useful in open-domain information extraction
  - each user searches for something; collectively, all users search for many (most?) things
  - queries often reflect the relative popularity of people, topics, events etc.

  --> useful in the extraction and ranking of instances, classes and relations