September 2015

Knowledge Acquisition for Web Search

Marius Paşca

Google Inc.
mars@google.com

Overview

- Part One: Introduction
- Part Two: Acquisition of Open-Domain Knowledge
- Part Three: Role of Knowledge in Information Retrieval

Part One: Introduction

- Open-domain information extraction
- Instances, concepts, relations
- Impact on Web search

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

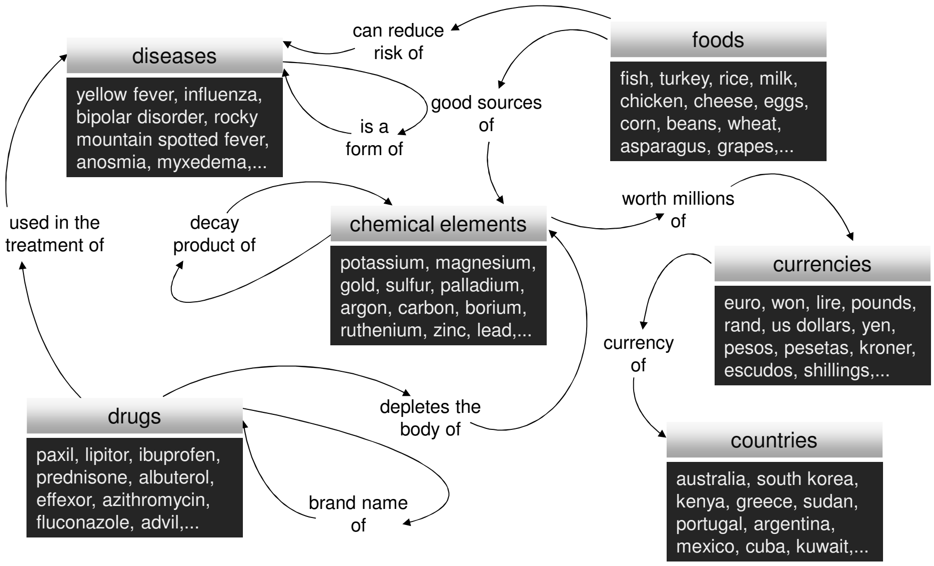
Instances, Concepts and Relations

- 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)

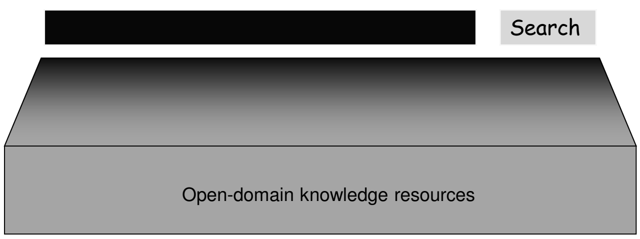
Instances, Concepts and Relations

- Relations are assertions linking two (binary relation) or more (n-ary relation) concepts or instances
  - 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 concept or an instance
  - actors --> awards, birth date, height
  - movies --> producer, release date, budget
- Terminology
  - concept vs. class: used interchangeably
  - instance vs. entity: used interchangeably

Open-Domain Knowledge



Usefulness in Information Retrieval



Next Topic

- Part One: Introduction
- Part Two: Acquisition of Open-Domain Knowledge
- Part Three: Role of Knowledge in Information Retrieval

Part Two: Acquisition of Knowledge

- Human-compiled knowledge resources
  - created by experts:
  - [Fel98]: C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press 1998.
  - [Len95]: D. Lenat. CYC: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 1995.
  - created collaboratively by non-experts:
  - [Rem02]: M. Remy. Wikipedia: The Free Encyclopedia. Journal of Online Information Review 2002.
  - [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.
  - [BEP+08]: K. Bollacker, C. Evans, P. Paritosh et al. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. SIGMOD-08.
  - [VK14]: Denny Vrandecic, Markus Krotzsch. Wikidata: A Free Collaborative Knowledgebase. Communications of the ACM 2014.
  - [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.
  - [LS04]: H. Liu and P. Singh. ConceptNet – a Practical Commonsense Reasoning Tool-Kit. BT Technology Journal 2004.

Quantitative Comparison of Human-Compiled Resources

- Wikipedia
  - 5+ million articles in English
  - articles also available in 250+ other languages
- DBpedia
  - 4+ million instances in English, 250+ million relations
- Freebase
  - 40+ million instances, 3+ billion 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
- ConceptNet
  - 1.5+ million assertions in multiple languages

Extraction of Open-Domain Knowledge

Methods for extraction of:

- concepts and instances as:
  - flat sets of unlabeled instances
  - flat sets of labeled instances, associating instances with class labels
  - conceptual hierarchies
- relations and attributes over:
  - flat concepts
  - conceptual hierarchies

Instances Within Unlabeled Concepts

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

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

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

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

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

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

Instances Within Unlabeled Concepts

- [PTL93]: F. Pereira, N. Tishby and L. Lee. Distributional Clustering of English Words. ACL-93.
  - extract clusters of distributionally similar words from text documents
- [LP02]: D. Lin and P. Pantel. Concept Discovery from Text. COLING-02.
  - extract clusters of distributionally similar phrases from text documents
- [WC08]: R. Wang and W. Cohen. Iterative Set Expansion of Named Entities using the Web. ICDM-08.
  - expand sets of instances using Web documents via search engines
- [VP09]: V. Vyas and P. Pantel: Semi-Automatic Entity Set Refinement. NAACL-09.
  - improve expansion of sets of instances using Web documents, by providing as input a small set of negative examples (i.e., extractions that would be incorrect)
- [PP09]: M. Pennacchiotti and P. Pantel. Entity Extraction via Ensemble Semantics. EMNLP-09.
  - expand sets of instances using multiple sources of text
- [LW09]: D. Lin and X. Wu. Phrase Clustering for Discriminative Learning. ACL-IJCNLP-09.
  - extract clusters of distributionally similar phrases from Web 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
- [SZY+10]: S. Shi, H. Zhang, X. Yuan and J. Wen. Corpus-Based Semantic Class Mining: Distributional vs. Pattern-Based Approaches. COLING-10.
  - compare and select between extraction patterns and distributional similarities, in the task of expanding sets of instances
- [HX11]: Y. He and D. Xin. Seisa: Set Expansion by Iterative Similarity Aggregation.WWW-11.
  - expand sets of instances using Web documents and queries

Instances Within Unlabeled Concepts

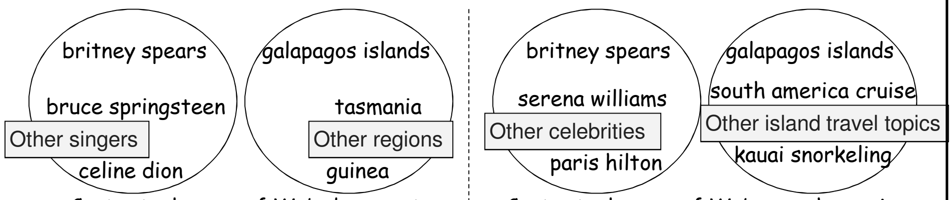
- [JP10]: A. Jain and P. Pantel. Open Entity Extraction from Web Search Query Logs. COLING-10.

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

Next Topic

Methods for extraction of:

- concepts and instances as:
  - flat sets of unlabeled instances
  - <u>flat sets of labeled instances, associating instances with class labels</u>
  - conceptual hierarchies
- relations and attributes over:
  - flat concepts
  - conceptual hierarchies

Instances Within Labeled 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 Within Labeled Concepts

- [Hea92]: M. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. COLING-92.
  - extract IsA pairs (i.e., pairs of an instance and a class label) from text documents using a set of lexico-syntactic patterns
- [RJ99]: E. Riloff and R. Jones. Learning Dictionaries for Information Extraction by Multi-level Bootstrapping. AAAI-99.
  - expand set of IsA pairs by iteratively identifying extraction patterns and conservatively growing the set of IsA pairs by a small number of new IsA pairs
- [RP04]: P. Pantel and D. Ravichandran. Automatically Labeling Semantic Classes. HLT-NAACL-04.
  - assign class labels to pre-extracted sets of instances
- [SJN05]: R. Snow, D. Jurafsky and A. Ng. Learning syntactic patterns for automatic hypernym discovery. NIPS-05.
  - learn extraction patterns for extracting IsA pairs from text documents
- [ECD+05]: O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D, Weld and A. Yates. Unsupervised Named-Entity Extraction from the Web: an Experimental Study. Journal of Artificial Intelligence 2005.
  - instantiate generic rule templates to extract instances within various concepts via search engines
- [TBL+06]: P. Talukdar, T. Brants, M. Liberman and F. Pereira. A Context Pattern Induction Method for Named Entity Extraction. CoNLL-06.
  - expand set of IsA pairs from text documents, by exploiting pairs extracted for other classes as negative examples to improve the quality of the induced patterns and extracted IsA pairs

Instances Within Labeled Concepts

- [KRH08]: Z. Kozareva, E. Riloff and E. Hovy. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. ACL-08.
  - expand set of instances and associated class label (also given as input) from Web documents via search engines
- [YTK+09]: I. Yamada, K. Torisawa, J. Kazama, K. Kuroda, M. Murata, S. De Saeger, F. Bond and A. Sumida. Hypernym Discovery Based on Distributional Similarity and Hierarchical Structures. ACL-IJCNLP-09.
  - extract IsA pairs from Web documents, by using lexico-syntactic patterns and distributional similarities, and attach extracted pairs to Wikipedia categories
- [WC09]: R. Wang and W. Cohen. Automatic Set Instance Extraction using the Web. ACL-IJCNLP-09.
  - extract instances of given class labels, from Web documents via search engines
- [TP10]: P. Talukdar and F. Pereira. Experiments in Graph-Based Semi-Supervised Learning Methods for Class-Instance Acquisition. ACL-10.
  - extract IsA pairs from manually-created or automatically-extracted repositories, via graph propagation, by incorporating structured data derived from Wikipedia
- [SHL10]: S. Singh, D. Hillard and C. Leggetter. Minimally-Supervised Extraction of Entities from Text Advertisements. ACL-10.
  - extract instances within around 30 class labels, from corpus of Web sponsored ads

Instances Within Labeled Concepts

- [ZSL+11]: F. Zhang. S. Shi, J. Liu, S. Sun and C. Lin: Nonlinear Evidence Fusion and Propagation for Hyponymy Relation Mining. ACL-11.
  - extract IsA pairs from Web documents, by using lexico-syntactic patterns then propagating class labels among similar instances
- [DCC12]: B. Dalvi, W. Cohen and J. Callan. WebSets: Extracting Sets of Entities from the Web Using Unsupervised Information Extraction. WSDM-12.
  - extract labeled sets of instances from Web documents, by using lexico-syntactic patterns and clusters of instances from Web tables
- [PL14]: P. Pasupat and P. Liang. Zero-shot Entity Extraction from Web Pages. ACL-14.
  - given a class label, extract set of instances of the class from Web documents
- [WCH+15]: C. Wang, K. Chakrabarti, Y. He, K. Ganjam, Z. Chen and P. Bernstein. Expansion of Tail Concepts Using Web Tables. WWW-15.
  - given a class label and a small set of seed instances, extract larger set of instances of the class from Web tables

Instances Within Labeled Concepts

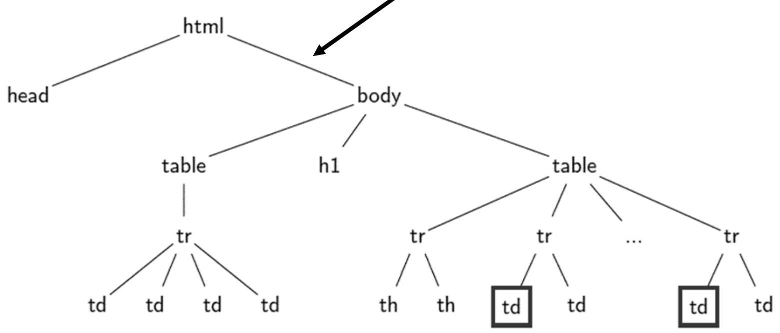
- [PL14]: P. Pasupat and P. Liang. Zero-shot Entity Extraction from Web Pages. ACL-14.

Extracting Instances from Documents

- Input
  - class label for which instances should be extracted (e.g., hiking trains near Baltimore)
- Data source
  - Web documents accessed via general-purpose Web search engine
- Output
  - instances of the class label (e.g., Avalon Super Loop)
- Steps
  - construct extraction predicates based on the Html representation of documents, targeting table columns, lists, headers at the same level
  - apply extraction patterns to Web documents relevant to the input class label
  - extract instances of the class label

Extraction Predicates

Html tree of Web document

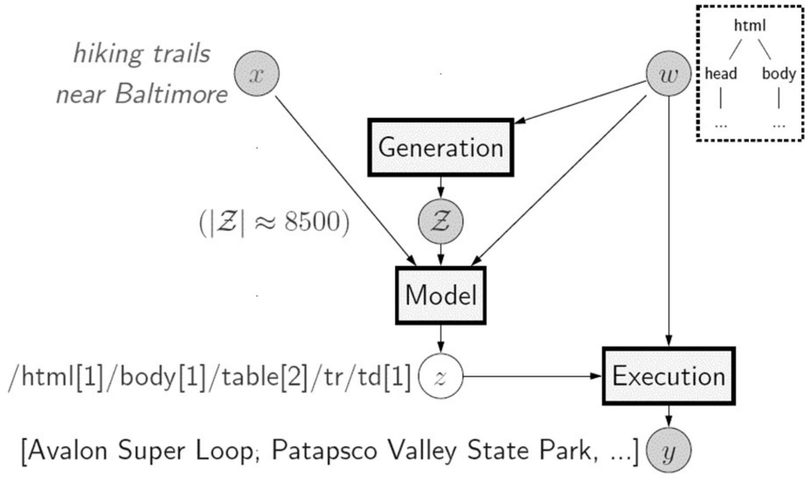


$z = $ /html[1]/body[1]/table[2]/tr/td[1]

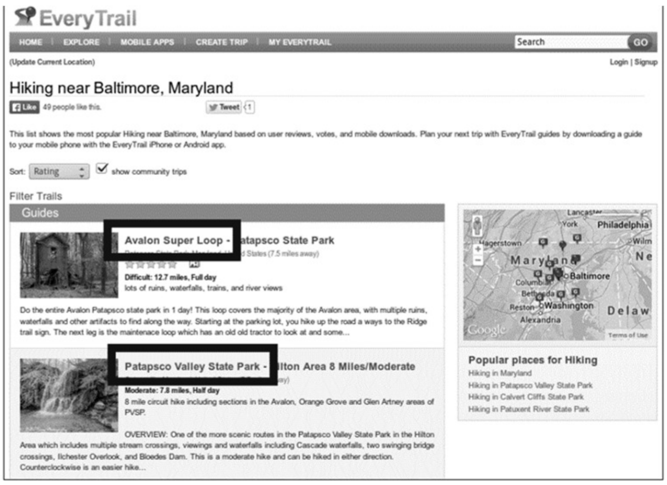
(Courtesy P. Pasupat)

Extraction predicate

Extraction Model



Extraction from Web Documents



Extraction Features

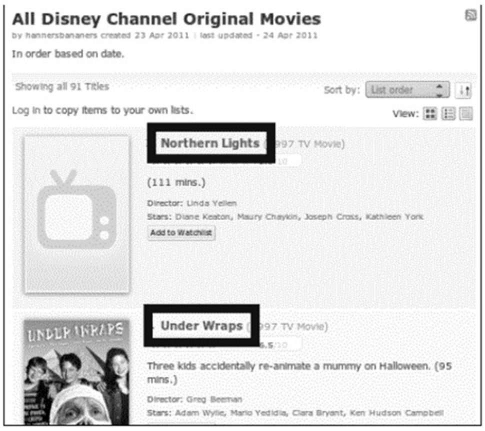
	George Washington John Adams Thomas Jefferson James Madison ... (39 more) ... Barack Obama	John Adams John Adams John Adams John Adams John Adams John Adams ... (100 more) ... John Adams	Blog Photos and Video Briefing Room In the White House Mobile Apps Contact Us
Desired Classification	(good)	(bad)	(bad)
Feature: Identity	(diverse)	(identical)	(diverse)

Extraction Features

	NNP NNP NNP NNP NNP NNP NNP NNP ... (39 more) ... NNP NNP	NNP NNP NNP NNP NNP NNP NNP NNP NNP NNP NNP NNP ... (100 more) ... NNP NNP	NN NNS CC NNP NN NN IN DT NNP NNP NNP NNPS NN PRP
Desired Classification	(good)	(bad)	(bad)
Feature: Identity	(diverse)	(identical)	(diverse)
Feature: Parts of speech	(identical)	(identical)	(diverse)

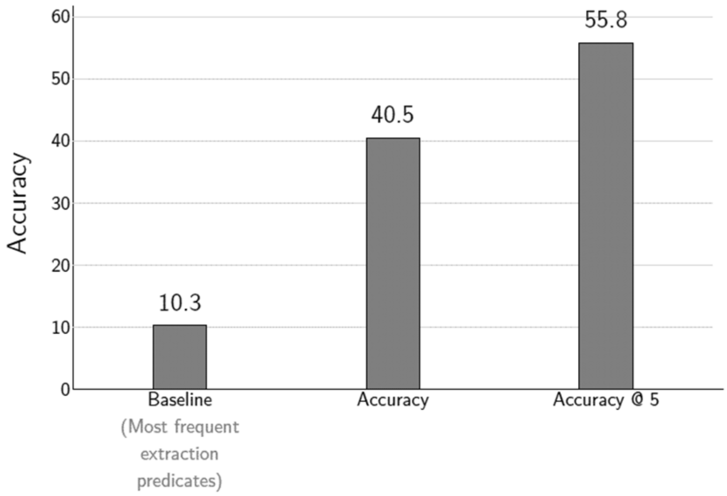
Extracted Instances

Query: *disney channel movies*



/html[1]/body/div[2]/div/div/div[3]/div[1]/div/div/div/div/b

Extracted Instances



Next Topic

Methods for extraction of:

- concepts and instances as:
  - flat sets of unlabeled instances
  - flat sets of labeled instances, associating instances with class labels
  - conceptual hierarchies
- relations and attributes over:
  - flat concepts
  - conceptual hierarchies

Conceptual Hierarchies

- [Wid03]: D. Widdows. Unsupervised Methods for Developing Taxonomies by Combining Syntactic and Statistical Information. HLT-NAACL-03.
  - insert new phrases into an existing hierarchy
- [SJN06]: R. Snow, D. Jurafsky and A. Ng. Semantic Taxonomy Induction from Heterogeneous Evidence. ACL-06.
  - extend WordNet with IsA pairs extracted from text
- [PS07]: S. Ponzetto and M. Strube. Deriving a Large Scale Taxonomy from Wikipedia. AAAI-07.
  - apply filters to network of Wikipedia categories to extract hierarchy of categories
- [YC09]: H. Yang and J. Callan. A Metric-Based Framework for Automatic Taxonomy Induction. ACL-IJCNLP-09.
  - incrementally cluster set of phrases into an hierarchy, using co-occurrence, syntactic dependencies and lexico-syntactic patterns
- [PN09]: S. Ponzetto and R. Navigli. Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. IJCAI-09.
  - map Wikipedia categories to WordNet synsets, and use mappings to restructure the hierarchy generated in [PS07]
- [KH10]: Z. Kozareva and E. Hovy. A Semi-Supervised Method to Learn and Construct Taxonomies Using the Web. EMNLP-10.
  - organize concepts extracted from Web documents via search engines, into hierarchies created from scratch

Conceptual Hierarchies

- [FVP+14]: T. Flati, D. Vannella, T. Pasini and R. Navigli. Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. ACL-14.
  - from network of Wikipedia articles and Wikipedia categories, extract hierarchy of articles and categories
- [SSF+15]: Y. Sun, A. Singla, D. Fox and A. Krause. Building Hierarchies of Concepts via Crowdsourcing. IJCAI-15.
  - automatically generate informative questions to ask users, such that their answers serve in incrementally constructing and refining conceptual hierarchies

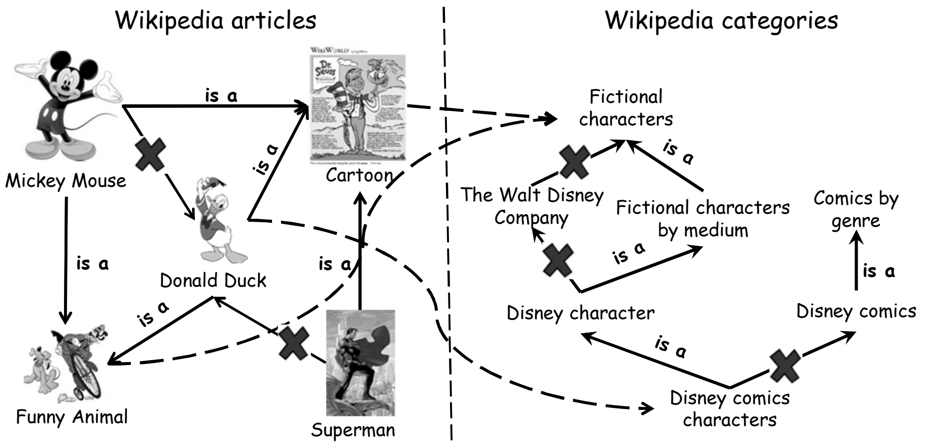
Conceptual Hierarchies

- [FVP+14]: T. Flati, D. Vannella, T. Pasini and R. Navigli. Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. ACL-14.

Constructing Hierarchies from Text

- Data source
  - category network in Wikipedia, containing edges from articles to other articles, from articles to their parent categories, and from categories to their own parent categories
- Output
  - an hierarchy containing IsA edges between Wikipedia articles, and another hierarchy containing IsA edges between Wikipedia categories
- Steps
  - analyze the text of each article to select a candidate hypernym phrase for the article (e.g., character for Mickey Mouse)
  - disambiguate the hypernym phrase into another article corresponding to the desired sense of the hypernym phrase (e.g., character into Character (arts))
  - organize pairs of an article and its disambiguated hypernym article into hierarchy of articles
  - starting from the hierarchy of articles (e.g., Real Madrid IsA Football club), exploit Wikipedia edges between articles and categories to iteratively infer IsA edges between categories of articles (e.g., Football clubs in Madrid IsA Football clubs), and then IsA edges between articles of categories (e.g., Atletico Madrid IsA Football club)
  - expand the hierarchy of categories to increase its coverage

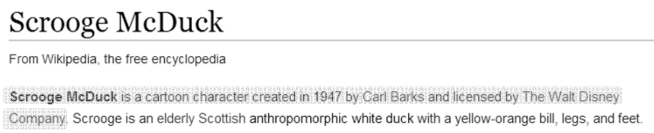
Edges in Category Network



(Courtesy T. Flati)

Constructing Hierarchy of Articles

- Select main (first) sentence in article (Courtesy T. Flati)


Scrooge McDuck

From Wikipedia, the free encyclopedia

Scrooge McDuck is a cartoon character created in 1947 by Carl Barks and licensed by The Walt Disney Company. Scrooge is an elderly Scottish anthropomorphic white duck with a yellow-orange bill, legs, and feet.

- From the dependency parse of main sentence, select candidate hypernym phrases

Scrooge McDuck is a character [...] ⟶ Scrooge McDuck is a character[...] ⟶ Hypernym phrase: character

nn nsubj cop

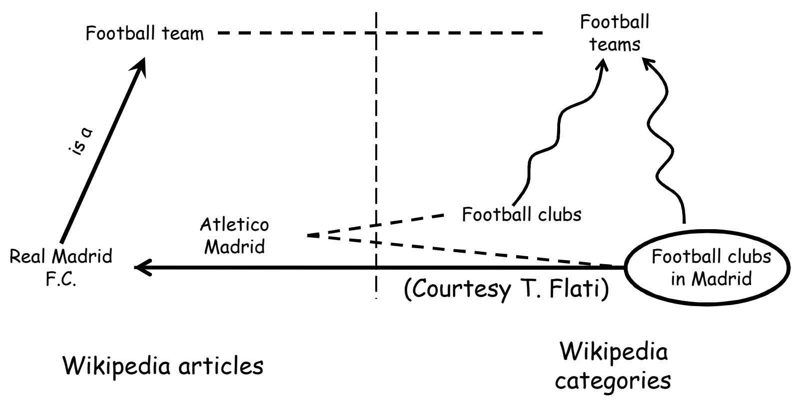
- Using heuristics, disambiguate the hypernym phrase into another article corresponding to the desired sense
  - heuristics rely on links among articles, common categories, context around links

Hypernym phrase: character ⟶ Character (arts)

- Retain pairs of an article and its hypernym article, as IsA edges

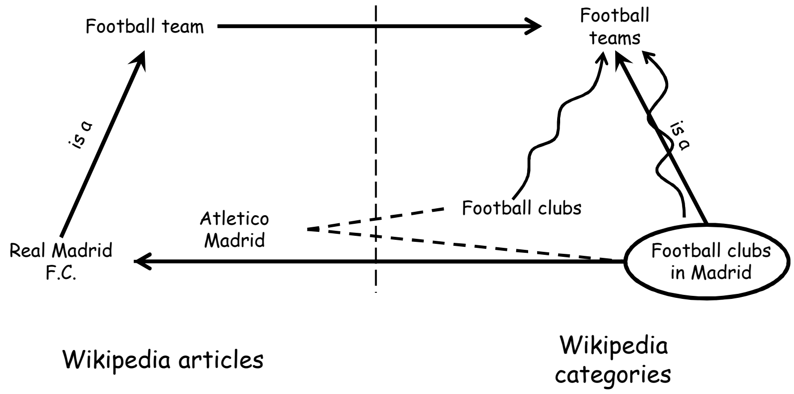
Scrooge McDuck ⟶ Character (arts)

Constructing Hierarchy of Categories



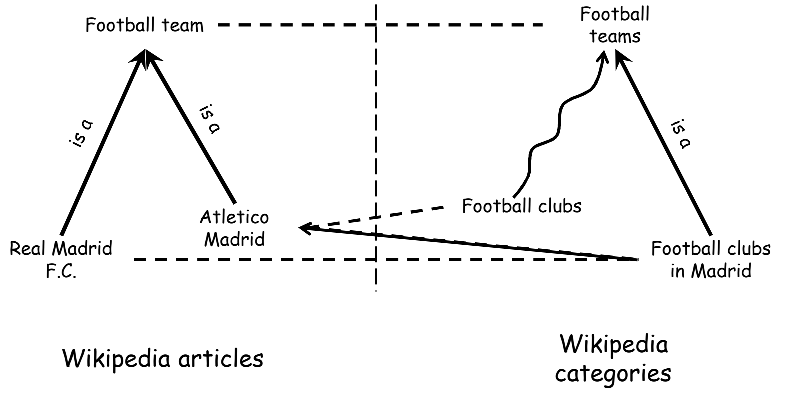
Start from the hierarchy of articles

Constructing Hierarchy of Categories



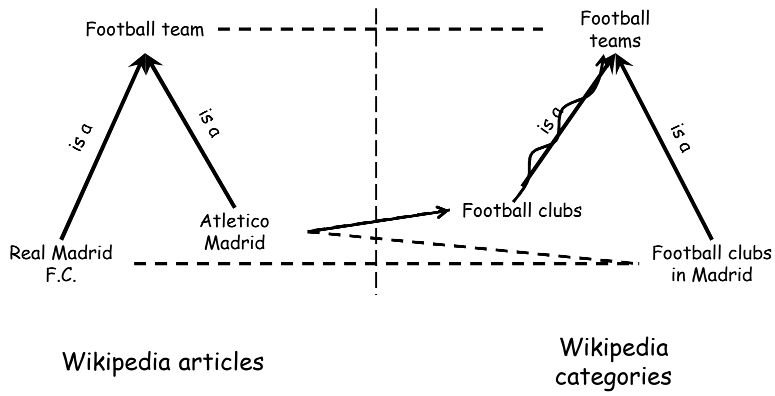
Exploit the article to category edges to infer hypernym relations in the category hierarchy

Iteratively Refining the Hierarchies



Traverse article to category edges to infer back is-a relations in the hierarchy of articles

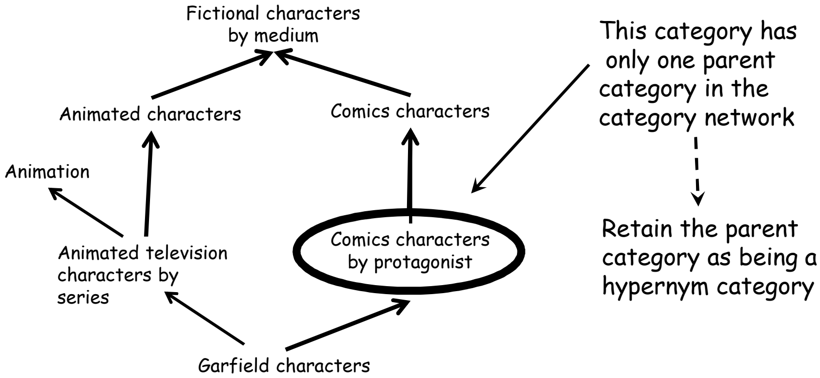
Iteratively Refining the Hierarchies



Iteratively use the relations found in previous step to infer new hypernym edges

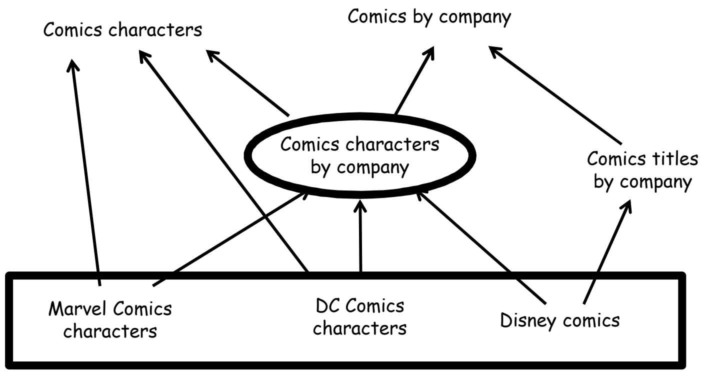
Expanding the Hierarchy of Categories

- Expand for categories with a single parent category in the category network



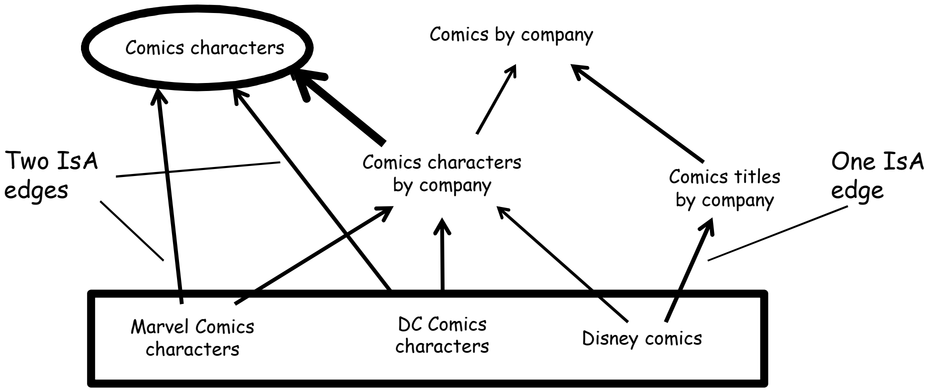
Expanding the Hierarchy of Categories

- Infer hypernym categories of a category, from hypernym categories already extracted for child categories in the category network



Expanding the Hierarchy of Categories

- Infer hypernym categories of a category, from hypernym categories already extracted for child categories in the category network

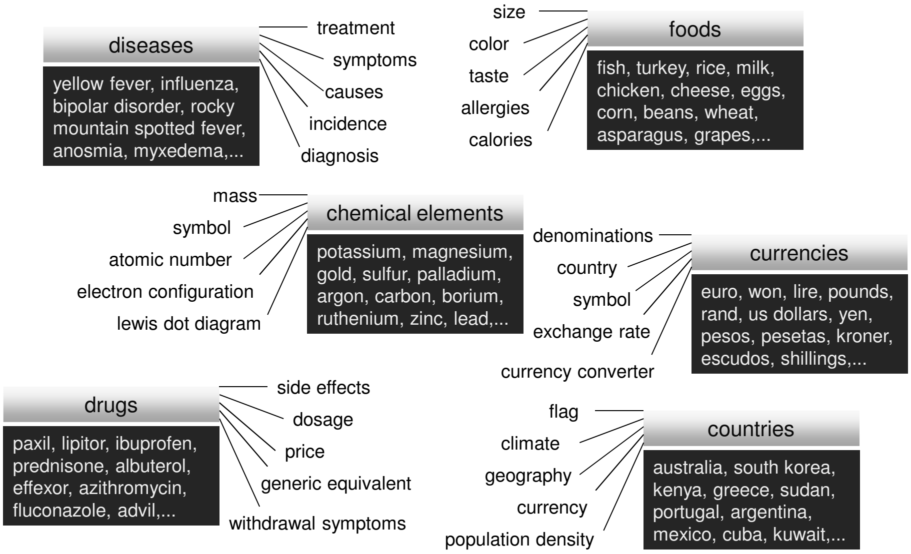


Next Topic

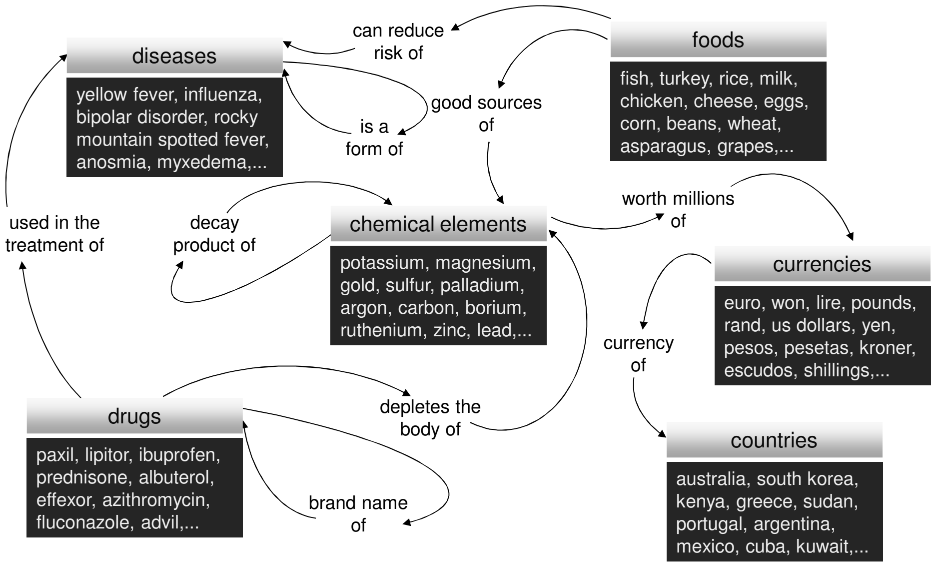
Methods for extraction of:

- concepts and instances as:
  - flat sets of unlabeled instances
  - flat sets of labeled instances, associating instances with class labels
  - conceptual hierarchies
- <u>relations and attributes over:</u>
  - flat concepts
  - conceptual hierarchies

Attributes and Relations



Attributes and Relations



Relations over Flat Concepts

- [AP04]: A. Almuhareb and M. Poesio. Attribute-Based and Value-Based Clustering: an Evaluation. EMNLP-04.
  - examine the role of attributes vs. values in acquiring concept descriptions via search engines
- [PE05]: A. Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews. EMNLP-05.
  - use product features (including attributes) found in text, to extract and rank opinions about products
- [TKT05]: K. Tokunaga, J. Kazama and K. Torisawa. Automatic Discovery of Attribute Words from Web Documents. IJCNLP-05.
  - apply small set of patterns to extract attributes from unstructured text in a small Web collection
- [CDS+05]: M. Cafarella, D. Downey, S. Soderland and O. Etzioni. KnowItNow: Fast, Scalable Information Extraction from the Web. HLT-EMNLP-05.
  - extract open-ended facts, without specifying the concepts or relations of interest in advance
- [SS06]: Y. Shinyama and S. Sekine: Preemptive Information Extraction using Unrestricted Relation Discovery. HLT-NAACL-06.
  - extract clusters of relations from parsed text, without specifying relations of interest in advance
- [PP06]: P. Pantel and M. Pennacchiotti. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. ACL-06.
  - expand seed set of relations from text documents via iteratively induced extraction patterns
- [WW07]: F. Wu and D. Weld. Autonomously Semantifying Wikipedia. CIKM-07.
  - extend Wikipedia infoboxes with attributes and values inferred from text
- [YT07]: N. Yoshinaga and K. Torisawa. Open-Domain Attribute-Value Acquisition from Semi-Structured Texts. Workshop on Ontolex-07.
  - extract attributes and associated values from semi-structured text via search engines

Relations over Flat Concepts

- [BM07]: R. Bunescu and R. Mooney. Learning to Extract Relations from the Web using Minimal Supervision. ACL-07.
  - exploit small sets of positive and negative seeds, to extract relations from text via search engines
- [PGK+07]: K. Probst, R. Ghani, M. Krema and A. Fano. Semi-Supervised Learning of Attribute-Value Pairs from Product Descriptions. IJCAI-07.
  - extract attributes and associated values of products
- [BCS+07]: M. Banko, M. Cafarella, S. Soderland, M. Broadhead and O. Etzioni. Open Information Extraction from the Web. IJCAI-07.
  - extract relations in a single pass over collection of Web documents, without any manual input
- [DRK07]: D. Davidov, A. Rappoport and M. Koppel. Fully Unsupervised Discovery of Concept-Specific Relationships by Web Mining. ACL-07.
  - extract relevant relations for given concepts, from unstructured text via search engines
- [VQS08]: B. Van Durme, T. Qian and L. Schubert. Class-Driven Attribute Extraction. COLING-08.
  - extract attributes via more complex representations of parsed text
- [CHW08]: M. Cafarella, A. Halevy and Z. Wang. WebTables: Exploring the Power of Tables on the Web. VLDB-08.
  - identify and exploit high-quality relational data available in tables

Relations over Flat Concepts

- [BE08] M. Banko and O. Etzioni. The Tradeoffs Between Open and Traditional Relation Extraction. ACL-08.
  - investigate the mapping of open-ended relations into relation-independent lexico-syntactic patterns
- [WLW08]: T. Wong, W. Lam and T. Wong. An Unsupervised Framework for Extracting and Normalizing Product Attributes from Multiple Web Sites. SIGIR-08.
  - extract attributes of products from semi-structured text within Web documents
- [NS08]: V. Nastase and M. Strube. Decoding Wikipedia Categories for Knowledge Acquisition. AAAI-08.
  - from categories and category network, derive relations among categories or instances, including attributes of categories
- [MBS+09]: M. Mintz, S. Bills, R. Snow and D. Jurafsky. Distant Supervision for Relation Extraction Without Labeled Data. ACL-IJCNLP-09.
  - using tuples already available for the relation in Freebase, extract additional relations from unstructured text
- [YOM+09]: Y. Yan, N. Okazako, Y. Matsuo et al. Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. ACL-IJCNLP-09.
  - identify relevant relations for Wikipedia categories, from parsed Wikipedia articles and from Web documents via search engines
- [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
- [CBW+10]: A. Carlson and J. Betteridge and R. Wang and E. Hruschka Jr. and T. Mitchell. Coupled Semi-Supervised Learning for Information Extraction. WSDM-10.
  - expand seed sets provided for each target concept and relation, enhancing extractions of individual concepts/relations using extractions for other concepts/relations

Relations over Flat Concepts

- [BZ10]: R. Blanco and H. Zaragoza. Finding Support Sentences for Entities. SIGIR-10.
  - loosely identify the relation between given a query and a given instance, in the form of explanatory sentences collected from Wikipedia articles
- [JP10b]: A. Jain and P. Pantel. FactRank: Random Walks on a Web of Facts. COLING-10.
  - improve quality of individually extracted facts, by global analysis of common arguments (instances) shared among the facts
- [DR10]: D. Davidov and A. Rappoport. Extraction and Approximation of Numerical Attributes from the Web. ACL-10.
  - given an instance and an attribute whose value is numerical, extract the value from Web documents via search engines
- [KH10b]: Z. Kozareva and E. Hovy. Learning Arguments and Supertypes of Semantic Relations using Recursive Patterns. ACL-10.
  - given an extraction pattern expressing a relation, and a seed instance for one argument of the relation, infer additional pairs of arguments for the same relation as well as the types of those arguments, from Web documents via search engines
- [LME10]: T. Lin, Mausam and O. Etzioni. Identifying Functional Relations in Web Text. EMNLP-10.
  - given relations extracted from Web documents, identify relations that connect the first argument to a unique value
- [SEW+10]: S. Schoenmackers, O. Etzioni, D. Weld and J. Davis. Learning First-Order Horn Clauses from Web Text. EMNLP-10.
  - acquire inference rules and apply them to expand a set of relations extracted from Web documents
- [BMI10]: D. Bollegala, Y. Matsuo and M. Ishizuka. Relation Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web. WWW-10.
  - model relations through combination of patterns expressing the type, and phrase pairs expressing the arguments

Relations over Flat Concepts

- [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
- [WW10]: F. Wu and D. Weld. Open Information Extraction Using Wikipedia. ACL-10.
  - from unstructured text, extract relations whose types are derived from Wikipedia
- [FSE11]: A. Fader, S. Soderland and O. Etzioni. Identifying Relations for Open Information Extraction. EMNLP-11.
  - enforce lexical and syntactic constraints on relations extracted from text, to improve their quality
- [DG13]: L. Del Corro and R. Gemulla. ClausIE: Clause-Based Open Information Extraction. WWW-13.
  - apply a small set of general-purpose patterns to parse trees over unstructured text, to extract higher-precision relations
- [WGM+14]: R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta and D. Lin. Knowledge Base Completion via Search-Based Question Answering. WWW-14.
  - extract missing values of attributes of instances within an existing knowledge repository, from Web search result snippets returned for automatically-generated questions
- [TMW14]: N. Tandon, G. de Melo and G. Weikum. Acquiring Comparative Commonsense Knowledge from the Web. AAAI-14.
  - from unstructured text, extract relations among disambiguated instances, where the relations compare the respective pairs of arguments along relevant dimensions
- [DGH+14]: X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao and K. Murphy. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. KDD-14.
  - create a knowledge repository, based on relations extracted from Web documents and knowledge from available repositories

Relations over Flat Concepts

- [DMS15]: A. Dutta, C. Meilicke and H. Stuckenschmidt. Enriching Structured Knowledge with Open Information. WWW-15
  - given relations extracted from Web documents, convert arguments and relations from ambiguous strings into disambiguated entries from a knowledge repository
- [NRC15]: A. Neelakantan, B. Roth and A. McCallum. Compositional Vector Space Models for Knowledge Base Completion. ACL-15.
  - infer missing relations based on relations already available in a knowledge repository
- [AJM15]: G. Angeli, M. Johnson Premkumar and C. Manning. Leveraging Linguistic Structure For Open Domain Information Extraction. ACL-15.
  - reduce document sentences deemed relevant to shorter clauses, then apply small set of patterns to clauses to extract relations

Relations over Flat Concepts

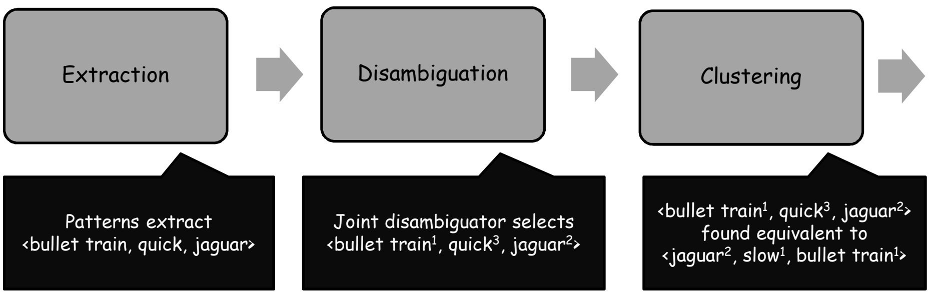
- [TMW14]: N. Tandon, G. de Melo and G. Weikum. Acquiring Comparative Commonsense Knowledge from the Web. AAAI-14.

From Relations to Explanatory Sentences

- Data source
  - collection of Web documents
- Output
  - facts capturing comparative knowledge, e.g., <snow, less dense, rain>
- Steps
  - apply a small set of extraction patterns to Web documents to extract comparative facts
  - disambiguate arguments and predicate within the extracted facts relative to WordNet
  - organize into groups of equivalent facts

Acquisition from Web Documents

(Courtesy N. Tandon)



Argument1	Relation/Adjective	Argument2
snow-n-2	less dense-a-3	rain-n-2
marijuana-n-2	more dangerous-a-1	alcohol-n-1
little child-n-1	happier (happy-a-1)	adult-n-1
private school-n-1	more expensive-a-1	public institute-n-1
peaceful resistance-n-1	more effective-a-1	violent resistance-n-1
wet wood-n-1	softer (soft-a-1)	dry wood-n-1

Next Topic

Methods for extraction of:

- concepts and instances as:
  - flat sets of unlabeled instances
  - flat sets of labeled instances, associating instances with class labels
  - conceptual hierarchies
- relations and attributes over:
  - flat concepts
  - <u>conceptual hierarchies</u>

Relations over Conceptual Hierarchies

- [PP06b]: M. Pennacchiotti and P. Pantel. Ontologizing Semantic Relations. ACL-06.
  - attach relations extracted from text to WordNet hierarchies, by identifying the WordNet concepts to which the arguments of the relation correspond
- [SKW07]: F. Suchanek, G. Kasneci and G. Weikum. Yago: a Core of Semantic Knowledge Unifying WordNet and Wikipedia. WWW-07.
  - map Wikipedia categories to WordNet to generate hybrid resource of concepts and relations
- [WW08]: F. Wu and D. Weld. Automatically Refining the Wikipedia Infobox Ontology. WWW-08.
  - extend Wikipedia infoboxes with additional attributes and values, by mapping templates of Wikipedia infoboxes to WordNet
- [SSW09]: F. Suchanek, M. Sozio and G. Weikum. Sofie: A Self-Organizing Framework for Information Extraction. WWW-09.
  - extend existing repositories of relations like Wikipedia, with facts acquired from unstructured text
- [HZW10]: R. Hoffmann, C. Zhang and D. Weld. Learning 5000 Relation Extractors. ACL-10.
  - extract relations from unstructured text within Wikipedia articles, via dynamic lexicons acquired from semi-structured text within Web documents
- [NP10]: R. Navigli and S. Ponzetto. BabelNet: Building a Very Large Multilingual Semantic Network. ACL-10.
  - link Wikipedia articles to WordNet concepts and apply machine translation, to create a multi-lingual repository of relations
- [MHM11]: T. Mohamed., E. Hruschka and T. Mitchell. Discovering Relations between Noun Categories. EMNLP-11.
  - given hierarchically-organized concepts associated with their sets of instances, extract relations among the concepts from unstructured text

Relations over Conceptual Hierarchies

- [HSB+13]: J. Hoffart, F. Suchanek, K. Berberich and G. Weikum. YAGO2: a Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Artificial Intelligence Journal.
  - combine WordNet, Wikipedia and other sources into hierarchically organized instances and their relations, where the data is anchored in time and space
- [VMT+15]: N. Voskarides, E. Meij, M. Tsagkias, M. de Rijke and W. Weerkamp. Learning to Explain Entity Relationships in Knowledge Graphs. ACL-15.
  - from Web documents, extract textual descriptions of relations between entries in pairs of entries from a knowledge repository
- [MC15]: D. Movshovitz-Attias and W. Cohen. KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts. ACL-15.
  - from Web documents, extract hierarchy of concepts and relation types, and also extract relations filling in the hierarchy

Relations over Conceptual Hierarchies

- [VMT+15]: N. Voskarides, E. Meij, M. Tsagkias, M. de Rijke and W. Weerkamp. Learning to Explain Entity Relationships in Knowledge Graphs. ACL-15.

From Relations to Explanatory Sentences

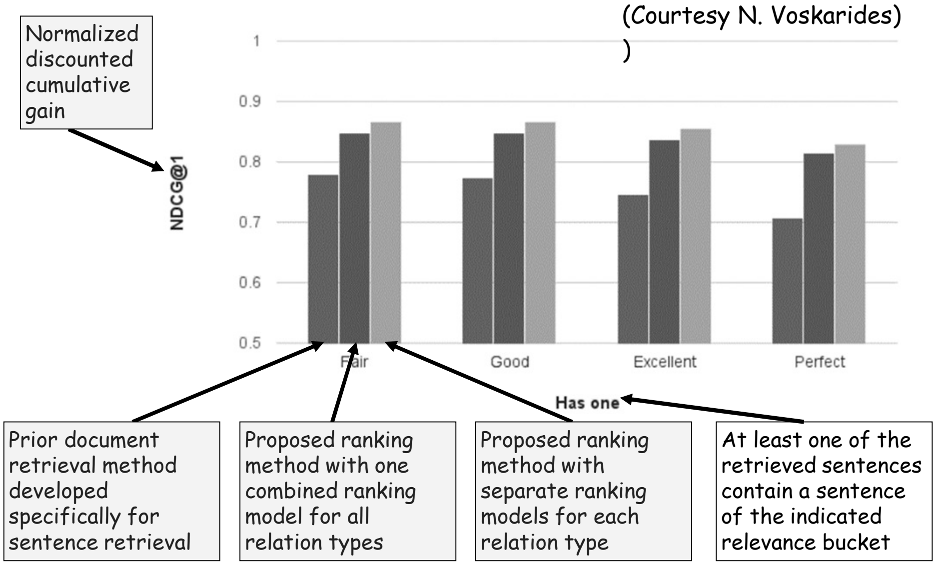
- Input
  - pairs of instances connected by a relation, as available in a knowledge repository (e.g., the pair Brad Pitt, Seven (1995 movie))
- Data source
  - Wikipedia articles corresponding to instances from the knowledge repository
- Output
  - ranked list of sentences extracted from documents, which explain the relations within the pairs of instances (e.g., "Brad Pitt gave critically acclaimed performances in the crime thriller Seven")
- Steps
  - extract surface forms for each instance in a pair of instances (e.g., Brad Pitt, Brad, Pitt)
  - extract candidate sentences from Wikipedia articles, using both surface forms and instances to which surface forms are disambiguated
  - rank candidate sentences using a variety of features

Ranking Candidate Sentences

- Textual features
  - length of candidate sentence
  - fractions of sentence tokens that are verbs vs. nouns vs. adjectives
  - ...
- Instance features
  - count of instances in candidate sentence
  - distance in tokens between last match of the two input instances in the candidate sentence
  - ...
- Relation features
  - whether candidate sentence contains tokens of the input relation
  - ...
- Document features
  - position of candidate sentence in source document
  - whether the source document of the candidate sentence is the article of one of the two input instances
  - ...

Accuracy of Explanatory Sentences

(Courtesy N. Voskarides)
)



Next Topic

- Part One: Introduction
- Part Two: Acquisition of Open-Domain Knowledge
- Part Three: Role of Knowledge in Information Retrieval

Role of Knowledge in Search

- [Voo94]. E. Voorhees. Query Expansion Using Lexical-Semantic Relations. SIGIR-94.
  - investigate the impact of manual and automatic expansion of queries on search results, using lexico-semantic relations available in WordNet
- [LZZ+06]: M. Li, M. Zhu, Y. Zhang and M. Zhou. Exploring Distributional Similarity Based Models for Query Spelling Correction. ACL-06.
  - take advantage of a repository of distributionally similar phrases acquired from search queries, in order to suggest correctly spelled queries in response to misspelled queries
- [Fan08]: H. Fang. A Re-Examination of Query Expansion Using Lexical Resources. ACL-08.
  - propose an alternative term weighting scheme for query expansion using lexico-semantic relations available in WordNet
- [HWL+09]: J. Hu, G. Wang, F. Lochovsky, J. Sun and Z. Chen. Understanding User's Query Intent with Wikipedia. WWW-09.
  - model query intent domains as areas in the Wikipedia category network situated around manually-provided seed articles in Wikipedia, and map queries into those domains
- [YTL11]: X. Yin, W. Tan and C. Liu. FACTO: a Fact Lookup Engine Based on Web Tables. WWW-11.
  - in response to fact-seeking queries, return facts identified in tuples of an instance, attribute and value extracted from tables within Web documents
- [JOV11]: A. Jain, U. Ozertem and E. Velipasaoglu. Synthesizing High Utility Suggestions for Rare Web Search Queries. SIGIR-11.
  - return synthetic query suggestions in response to long-tail queries for which few or no query suggestions would be otherwise available

Role of Knowledge in Search

- [RRD+11]: L. Ratinov, D. Roth, D. Downey and M. Anderson. Local and Global Algorithms for Disambiguation to Wikipedia. ACL-11.
  - compare the impact of algorithms for disambiguating instances mentioned in a document relative to articles in Wikipedia, using evidence available locally for each mention vs. globally for all mentions
- [PF11]: P. Pantel and A. Fuxman. Jigs and Lures: Associating Web Queries with Structured Entities. ACL-11.
  - compute mappings from queries into instances from a structured database, for the purpose of identifying relevant products from a product catalog and recommending them in response to queries
- [HMT+11]: K. Haas, P. Mika, P. Tarjan and R. Blanco. Enhanced Results for Web Search. SIGIR-11.
  - extend search results with multimedia objects, for the purpose of improving search experience and aiding users in determining the relevance of search results
- [SMF+12]: U. Scaiella, A. Marino, P. Ferragina and M. Ciaramita. Topical Clustering of Search Results. WSDM-12.
  - take advantage of mappings from instances mentioned in documents to Wikipedia articles, in order to cluster search results and their result snippets into sets associated with descriptive labels
- [WUG12]: I. Weber, A. Ukkonen and A. Gionis. Answers, not Links: Extracting Tips from Yahoo Answers to Address How-To Queries. WSDM-12.
  - in response to queries with how-to intent, return relevant tips extracted from a collaborative question-answering repository containing pairs of a question and an answer
- [KZ12]: A. Kotov and C. Zhai. Tapping into Knowledge Base for Concept Feedback: Leveraging ConceptNet to Improve Search Results for Difficult Queries. WSDM-12.
  - improve the search results returned for poorly performing queries, by expanding the queries with concepts derived from a large knowledge repository

Role of Knowledge in Search

- [HMB13]: L. Hollink, P. Mika and R. Blanco. Web Usage Mining with Semantic Analysis. WWW-13.
  - compute mappings from query fragments into instances from an existing knowledge repository, to better identify patterns of Web usage
- [GYS+13]: M. Gamon, T. Yano, X. Song, J. Apacible and P. Pantel. Identifying Salient Entities in Web Pages. CIKM-13.
  - extract the most salient instances mentioned in Web documents
- [YV14]: X. Yao and B. Van Durme. Information Extraction over Structured Data: Question Answering with Freebase. ACL-14.
  - in response to fact-seeking questions, extract answers from unstructured text from Web documents and from relations available in a knowledge repository
- [DAD14]: J. Dalton, J. Allan and L. Dietz. Entity Query Feature Expansion using Knowledge Base Links. SIGIR-14.
  - compute mappings from query fragments into instances from an existing knowledge repository, to expand queries for better search results
- [BMH+15]: B. Bi, H. Ma, B. Hsu, W. Chu, K. Wang and J. Cho. Learning to Recommend Related Entities to Search Users. WSDM-15.
  - given a query, compute and recommend related entries from a knowledge repository
- [BOM15]: R. Blanco, G. Ottaviano and E. Meij. Fast and Space-Efficient Entity Linking in Queries. WSDM-15.
  - compute mappings from query fragments into instances from an existing knowledge repository, under strong latency constraints
- [FBJ15]: J. Foley, M. Bendersky and V. Josifovski. Learning to Extract Local Events from the Web. SIGIR-15.
  - extract and convert mentions of local events within Web documents into structured, searchable calendar entries

Role of Knowledge in Search

- Document analysis and understanding
  - mapping of document terms into concepts [RRD+11]
  - clustering of search results [SMF+12]
  - extraction of salient instances [GYS+13]
- Query analysis and understanding
  - understanding intent, query categorization [HWL+09]
  - Web usage analysis [HMB13]
  - mapping of queries into concepts [BOM15], product recommendation [PF11]
  - query suggestion [JOV11], recommendation of related queries [BMH+15]
  - spell checking [LZZ+06]
- Matching of queries onto documents
  - query expansion [Voo94, Fan08, KZ12, DAD14]
- Onebox search results
  - retrieval of answers for queries with how-to intent [WUG12]
  - retrieval of answers for fact-seeking queries [YTL11, YV14]
  - retrieval of multimedia objects [HMT+11]
- ...
- ...

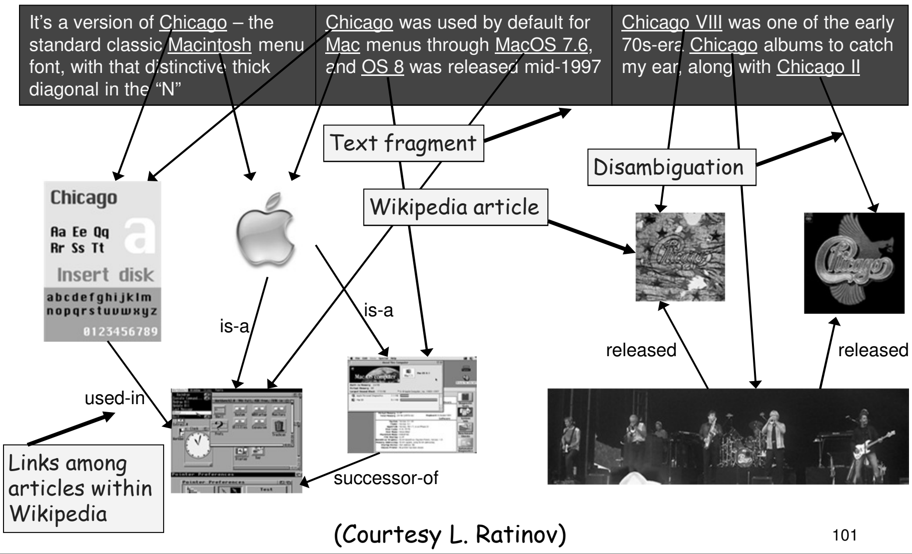
Document Understanding

- [RRD+11]: L. Ratinov, D. Roth, D. Downey and M. Anderson. Local and Global Algorithms for Disambiguation to Wikipedia. ACL-11.

Disambiguation to Wikipedia

- Task
  - given a text fragment containing mentions (substrings) to be disambiguated, "wikifi" the mentions by identifying the Wikipedia article, if any, corresponding to each mention
  - mapping from mentions to Wikipedia articles relies on evidence available in the text fragment
- Scope of available evidence
  - local: separately available for each mention in the text fragment
  - global: collectively available for all mentions in the text fragment
- Goal
  - investigate impact of local vs. global evidence on accuracy of disambiguation

Mapping Mentions to Wikipedia



(Courtesy L. Ratinov)

Disambiguation Strategy

Algorithm: Disambiguate to Wikipedia
Input: document d, Mentions $M = \{m_1, \ldots, m_N\}$
Output: a disambiguation $\Gamma = (t_1, \ldots, t_N)$.
1) Let $M' = M \cup \{$ *Other potential mentions in* $d\}$
2) For each mention $m'_i \in M'$, *construct a set of disambiguation candidates* $T_i = \{t^i_1, \ldots, t^i_{k_i}\}, t^i_j \neq$ null
3) **Ranker**: *Find a solution* $\Gamma = (t'_1, \ldots, t'_{|M'|})$, *where* $t'_i \in T_i$ *is the best non-null disambiguation of* m'_i.
4) **Linker**: *For each* m'_i, *map* t'_i *to* null *in* Γ iff *doing so improves the objective function*
5) Return Γ *entries for the original mentions* M.

- Two stages
  - ranker: compute best Wikipedia article that potentially disambiguates the mention
  - linker: determine whether the mention should be mapped to the Wikipedia article or should not be mapped to any article

Ranker: Local vs. Global Disambiguation

Accuracy: fraction of mentions for which ranker identifies correct disambiguation

Dataset	Baseline	Baseline+ Lexical	Baseline+ Global Unambiguous	Baseline+ Global NER	Baseline+ Global, All Mentions
ACE	94.05		94.56	96.21	96.75
MSNBC News	81.91		84.46	84.04	88.51
AQUAINT	93.19		95.40	94.04	95.91
Wikipedia Test	85.88		89.67	89.59	89.79

Previous methods

Ranker: Local vs. Global Disambiguation

Accuracy: fraction of mentions for which ranker identifies correct disambiguation

Dataset	Baseline	Baseline+ Lexical	Baseline+ Global Unambiguous	Baseline+ Global NER	Baseline+ Global, All Mentions
ACE	94.05	96.21			96.75
MSNBC News	81.91	85.10			88.51
AQUAINT	93.19	95.57			95.91
Wikipedia Test	85.88	93.59			89.79

Local disambiguation

Global disambiguation

Over test set of Wikipedia documents, local performs better than global

Overall: Local vs. Global Evidence

Combined precision and recall (F1 score)

Dataset	Baseline	Baseline+ Lexical	Baseline+ Lexical+ Global
ACE	94.05	96.21	97.83
MSNBC News	81.91	85.10	87.02
AQUAINT	93.19	95.57	94.38
Wikipedia Test	85.88	93.59	94.18

(Comparing set of Wikipedia articles output by algorithm for a document, with gold set of Wikipedia articles for the document)

(Result) Document Understanding

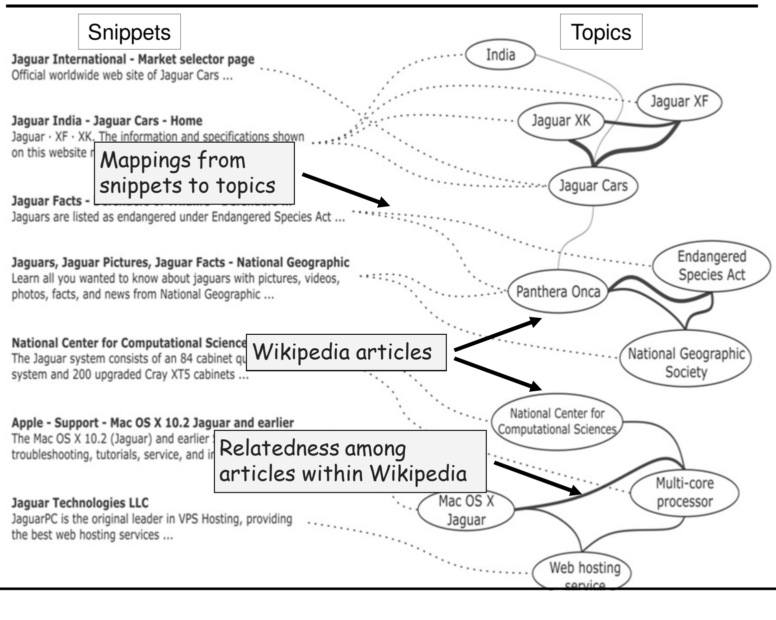
- [SMF+12]: U. Scaiella, A. Marino, P. Ferragina and M. Ciaramita. Topical Clustering of Search Results. WSDM-12.

Clustering of Search Results

- Input
  - search results and their snippets, returned in response to queries
- Data source
  - Wikipedia articles and categories, connected via the category network
- Output
  - decomposition of search results into topically coherent subsets associated with labels derived from Wikipedia
  - on the fly, without analysis of full content of search results
- Steps
  - annotate snippets with corresponding Wikipedia articles ("topics")
  - analyze graph of snippets and topics, to determine most significant topics
  - partition graph around most significant topics, and cut into ~10 clusters
  - for each cluster, select centroid topic as label for the entire cluster

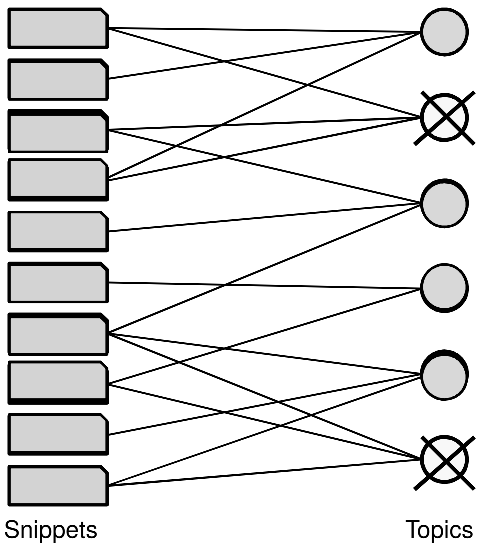
Annotation of Snippets with Topics

(Courtesy P. Ferragina)

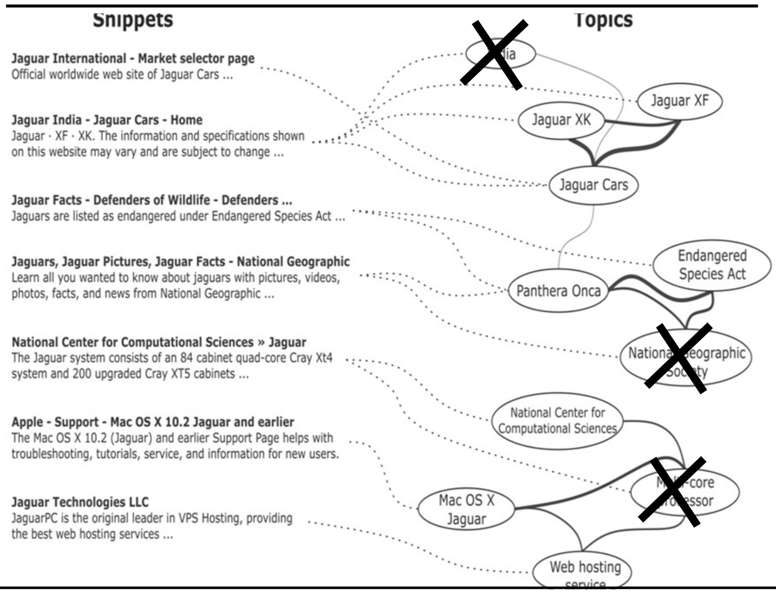


Selection of Significant Topics

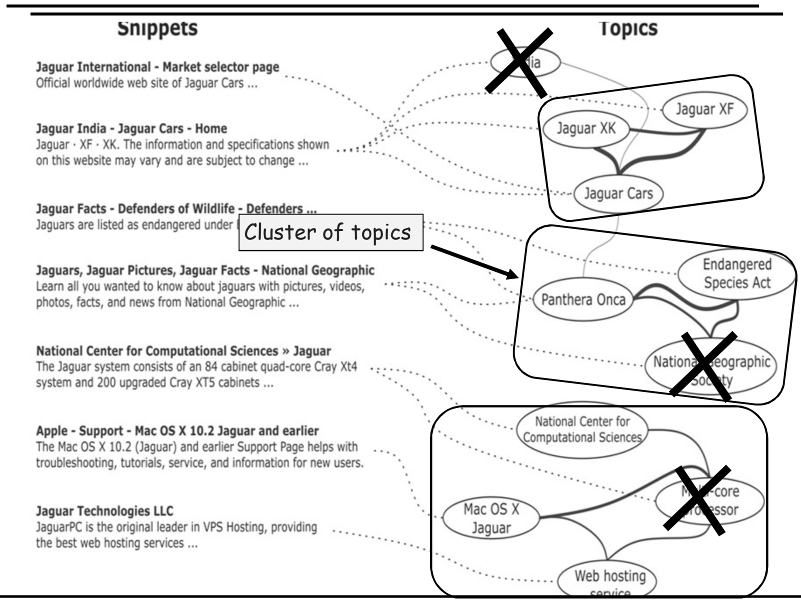
- Exploit weights assigned during annotation to mappings from snippets to topics
- Process topic nodes iteratively, in order of the sum of weights of connecting edges
- As topic nodes are marked as significant, ignore corresponding snippet nodes and their outgoing mappings in subsequent iterations



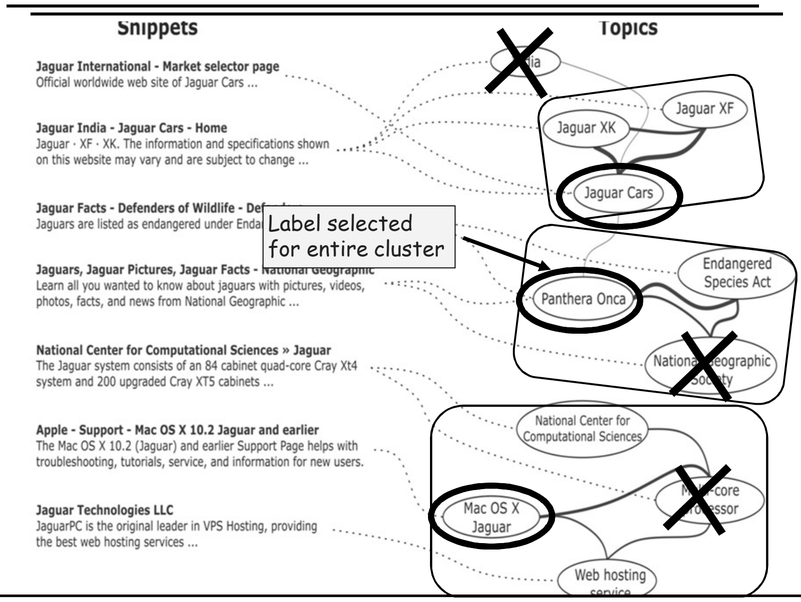
Selection of Significant Topics



Partition into Clusters



Selection of Cluster Labels



Next Topic

- Document analysis and understanding
- Query analysis and understanding
- Matching of queries onto documents
- Onebox search results

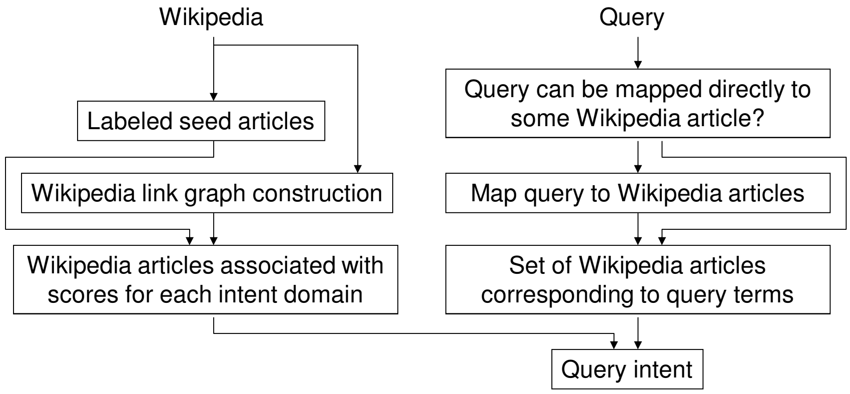
Query Understanding

- [HWL+09]: J. Hu, G. Wang, F. Lochovsky, J. Sun and Z. Chen. Understanding User's Query Intent with Wikipedia. WWW-09.

Modeling Query Intent with Wikipedia

- Input
  - queries
- Data source
  - Wikipedia articles and categories, connected via the category network
- Output
  - intent domains identified for queries, modeled as areas in the Wikipedia category network situated around manually-provided seed articles in Wikipedia
- Steps
  - independently from input queries, manually identify a small set of seed queries for each domain of interest
  - given set of seed queries, manually identify seed Wikipedia articles that correspond to the domain of interest
  - for each domain, expand seed Wikipedia articles into more Wikipedia articles, using connections between articles (article links, category network)
  - map queries into intent domains, taking into consideration manually-provided mappings from sets of seed queries

System Architecture



Modeling of Intent Domains

- Construct link graph for Wikipedia articles
  - nodes: Wikipedia articles, Wikipedia categories
  - edges: links between articles, links in Wikipedia category network between articles and categories; edges added between two nodes only when bi-directional links exist between the two nodes
  - edge weights: counts of links between the two nodes
- Associate Wikipedia articles with score for each intent domain
  - manually select seed Wikipedia articles deemed to belong to intent domain

Intent Type	Examples of Seed Queries	# Seed Queries
Travel	travel, hotel, tourism, airline tickets, expedia	2389
Person Name	britney spears, david beckham, george w. bush	10000
Employment	employment, monster, career	2543

  - iteratively propagate intent from seed articles to their neighbors articles in the link graph, assigning gradually lower intent scores

Determining Query Intent

- Case 1: query can be mapped directly to a Wikipedia article
  - retrieve intent domain whose intent score associated with the Wikipedia article is highest
- Case 2: query cannot be mapped directly to a Wikipedia article
  - map query into its more related Wikipedia articles, by disambiguating ("wikifying") mentions (substrings) from query to corresponding Wikipedia articles
  - retrieve intent domain for which the combination of intent scores, associated with the related Wikipedia articles, is highest

Query	Top Articles to Which Query is Mapped	Query Intent
employment guide	employment website, job search engine, careerlink, job hunting, eluta.ca, types of unemployment, airline tickets, expedia	Employment
job builder	job search engine, jobserve, falcon's eye, careerbuilder, eluta.ca, monster (website)	Employment

Query Understanding

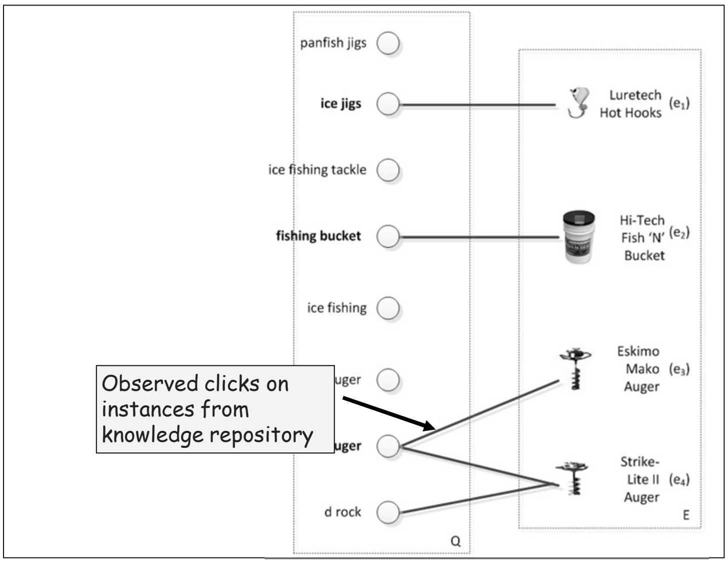
- [PF11]: P. Pantel and A. Fuxman. Jigs and Lures: Associating Web Queries with Structured Entities. ACL-11.

Mapping Queries to Structured Entities

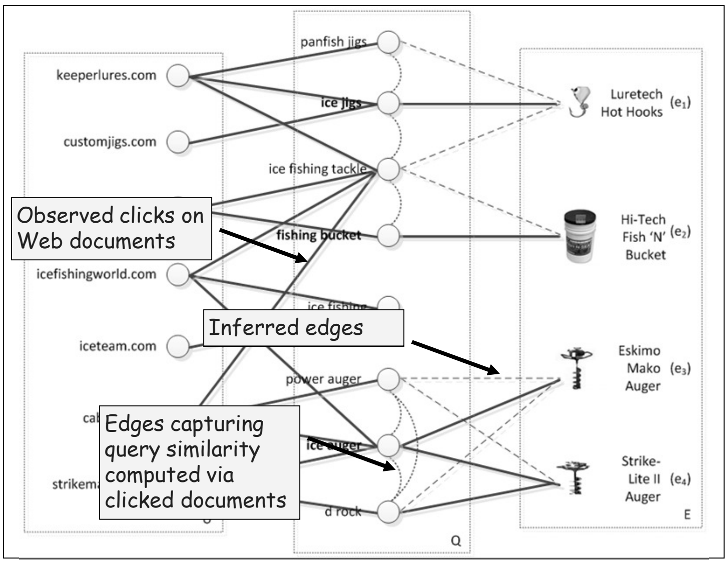
- Input
  - queries
  - click data for search results returned in response to queries
  - click data for structured instances returned in response to queries
- Data source
  - collection of instances available within a structured knowledge repository (e.g., Freebase, IMDB, product catalog)
- Output
  - list of instances from knowledge repository deemed relevant to the query
  - similar to query suggestion, but suggestions are instances not strings
- Steps
  - create click graph connecting queries, instances and clicked documents
  - exploit edges between queries and clicked documents, and similarity edges between queries capturing the overlap of their sets of clicked documents, to extend graph with new edges between queries and instances
  - transfer weights from existing edges to newly added edges
  - apply resulting graph to suggest relevant instances for each query

Construction of Click Graph

(Courtesy P. Pantel)



Construction of Click Graph



Associations from Instances to Queries

Query	$\hat{P}_{mle}$	$\hat{P}_{intp}$	Query	$\hat{P}_{mle}$	$\hat{P}_{intp}$
Garmin GTM 20 GPS			Canon PowerShot SX110 IS		
garmin gtm 20	0.44	0.45	canon sx110	0.57	0.57
garmin traffic receiver	0.30	0.27	powershot sx110	0.48	0.48
garmin nuvi 885t	0.02	0.02	powershot sx110 is	0.38	0.36
gtm 20	0	0.33	**powershot sx130 is**	0	0.33
garmin gtm20	0	0.33	**canon power shot sx110**	0	0.20
nuvi 885t	0	0.01	**canon dig camera review**	0	0.10
Samsung PN50A450 50" TV			Devil May Cry: 5th Anniversary Col.		
samsung 50 plasma hdtv	0.75	0.83	devil may cry	0.76	0.78
samsung 50	0.33	0.32	**devilmaycry**	0	1.00
50" hdtv	0.17	0.12	High Island Hammock/Stand Combo		
samsung plasma tv review	0	0.42	high island hammocks	1.00	1.00
50" samsung plasma hdtv	0	0.35	**hammocks and stands**	0	0.10

via observed and inferred clicks on instances

via observed clicks on instances

Associations from Queries to Instances

- Instances suggested via observed clicks and inferred "clicks" on instances

Query	Product Recommendation
wedding gowns	27 Dresses (Movie Soundtrack)
wedding gowns	Bridal Gowns: The Basics of Designing, [...] (Book)
wedding gowns	Wedding Dress Hankie
wedding gowns	The Perfect Wedding Dress (Magazine)
wedding gowns	Imagine Wedding Designer (Video Game)
low blood pressure	Omron Blood Pressure Monitor
low blood pressure	Healthcare Automatic Blood Pressure Monitor
low blood pressure	Ridgecrest Blood Pressure Formula - 60 Capsules
low blood pressure	Omron Portable Wrist Blood Pressure Monitor
'hello cupcake' cookbook	Giant Cupcake Cast Pan
'hello cupcake' cookbook	Ultimate 3-In-1 Storage Caddy
'hello cupcake' cookbook	13 Cup Cupcakes and More Dessert Stand
'hello cupcake' cookbook	Cupcake Stand Set (Toys)
1 800 flowers	Todd Oldham Party Perfect Bouquet
1 800 flowers	Hugs and Kisses Flower Bouquet with Vase

Query Understanding

- [BMH+15]: B. Bi, H. Ma, B. Hsu, W. Chu, K. Wang and J. Cho. Learning to Recommend Related Entities to Search Users. WSDM-15.

Recommending Related Instances

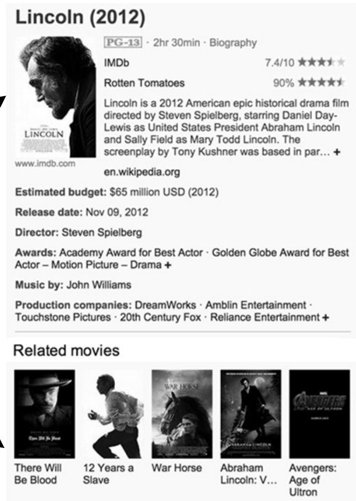
- Input
  - query submitted by a user
  - main structured instance, if any, returned by search engine as a side result for the query
- Data source
  - click data for search results returned in response to queries
  - click data for main instances, if any, returned as side results in response to queries
  - collection of instances available within a structured knowledge repository (e.g., Freebase)
- Output
  - list of instances from knowledge repository deemed relevant to the query and the main instance, based on click data available for the user
  - similar to query suggestion, but suggestions are instances not strings, and suggestions are for instances rather than queries
- Steps
  - exploit three types of evidence, namely edges between instances within knowledge repository; edges between main instances in side results and clicked documents; and edges between main instances in side results and clicked related instances in side results
  - given a main instance, recommend a list of related entities based on the user's interests

Main Instance and Related Instances

lincoln movie | Search

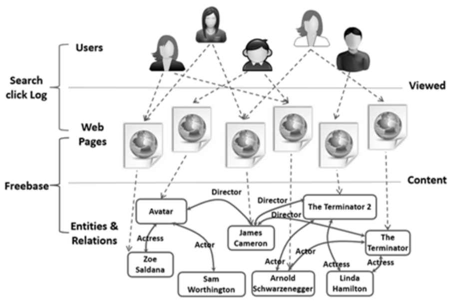
Main instance returned by search engine

Related instances



(Courtesy B. Bi)

Evidence Towards Related Instances



Edges between instances and clicked documents

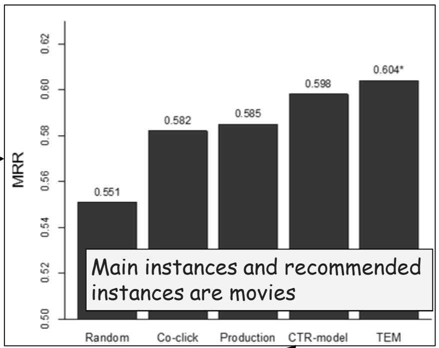
Edges between instances within knowledge repository

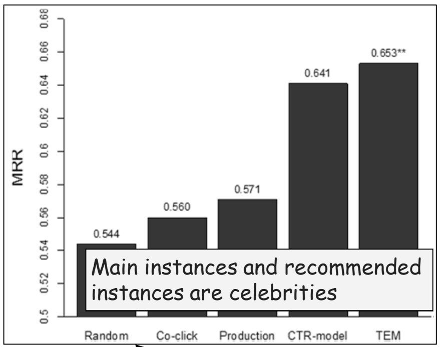
Edges between instances and clicked related instances in side results

User ID	Time	Main entity	Related entity	Rank	Click
32	7/9/2013 10:32:26	Pacific Rim	Man of Steel	1	0
32	7/9/2013 10:32:26	Pacific Rim	The Wolverine	2	0
32	7/9/2013 10:32:26	Pacific Rim	The Lone Ranger	3	1
498	6/16/2013 15:16:41	Leonardo DiCaprio	Kate Winslet	1	0
498	6/16/2013 15:16:41	Leonardo DiCaprio	Johnny Depp	2	1

Recommended Instances

Mean reciprocal rank of relevant instance in computed ranked list of related instances





Co-click: evidence from user clicks on both main instance and related instance

CTR-model: evidence from click-through rate for related instances being returned

TEM: all sources of evidence

Query Understanding

- [HMB13]: L. Hollink, P. Mika and R. Blanco. Web Usage Mining with Semantic Analysis. WWW-13.

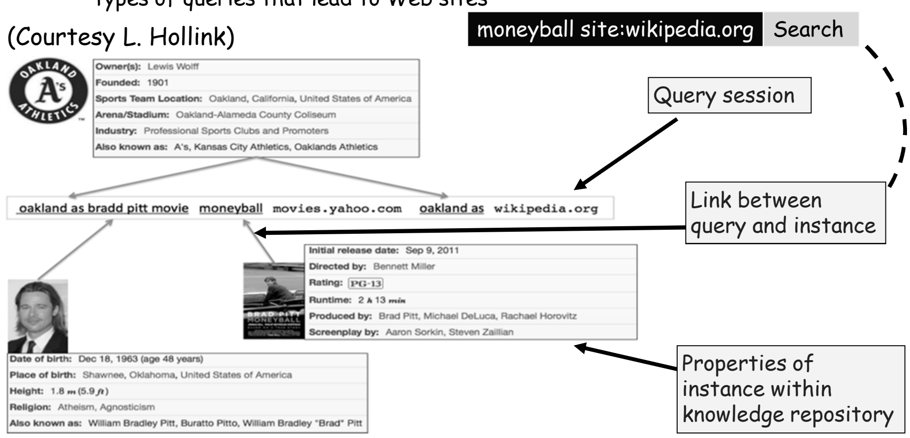
Web Usage Understanding

- Data source
  - query sessions including queries and clicked documents
  - collection of instances available within a structured knowledge repository (Freebase and DBpedia)
- Output
  - semantic rather than lexical patterns of Web usage mining
- Steps
  - annotate queries, by linking query fragments to corresponding instances from knowledge repository
  - use properties available for instances within knowledge repository to generalize and categorize queries

Linking Queries to Instances

- Link queries to instances
  - using types available for instances in the knowledge repository, compute types of queries that lead to Web sites

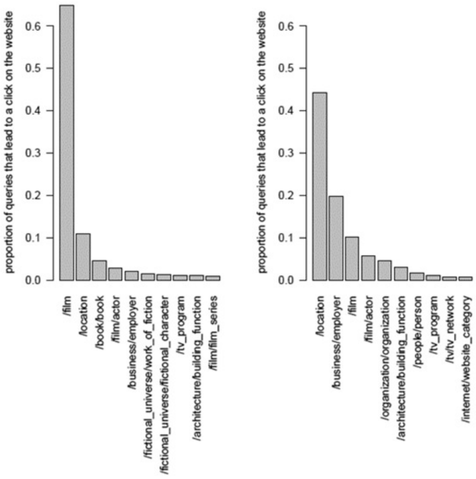
(Courtesy L. Hollink)



- Generalize from linked queries to query types
  - using instance types from knowledge repository, generalize queries into query types that lead to Web sites

Patterns of Web Usage

- Estimate the differences in the content of two Web sites, by comparing the top query types that lead to clicks on the sites



The first Web site specializes in movies, whereas the second is broader as it covers content related to movies, TV and people

Patterns of Web Usage

- Compare query patterns obtained by generalizing from queries linked to instances that are relatively recent movies; vs. instances that are relatively older movies

Level	Pattern	Support
L1	*movie*	0.396
	2011	0.15
	trailer	0.103
	movies	0.095
	moviedict : 2011	0.063
	cast	0.053
	new	0.049
	dvd	0.035
	release	0.032
	reviews	0.028
L2	*movie* ⟶ *movie*	0.165
	2011 ⟶ 2011	0.042
	movie ⟶ 2011	0.04
	2011 ⟶ *movie*	0.038
	movies ⟶ *movies*	0.038
	trailer ⟶ *trailer*	0.027
	movie ⟶ *movie* 2011	0.026
	movies ⟶ *movie*	0.025
	movie ⟶ *trailer*	0.024
	movie 2011 ⟶ *movie*	0.023

vs.

Level	Pattern	Support
L1	*movie*	0.391
	cast	0.096
	movies	0.091
	quotes	0.038
	trailer	0.034
	new	0.027
	2011	0.027
	free	0.023
	soundtrack	0.021
	watch	0.021
L2	*movie* ⟶ *movie*	0.169
	movies ⟶ *movies*	0.038
	cast ⟶ *cast*	0.028
	movies ⟶ *movie*	0.025
	movie ⟶ *movies*	0.023
	quotes ⟶ *quotes*	0.019
	movie ⟶ *cast*	0.018
	movie ⟶ *trailer*	0.012
	movie ⟶ *moviecast*	0.011
	new ⟶ *new*	0.01

Next Topic

- Document analysis and understanding
- Query analysis and understanding
- Matching of queries onto documents
- Onebox search results

Matching of Queries onto Documents

- [Voo94]. E. Voorhees. Query Expansion Using Lexical-Semantic Relations. SIGIR-94.

Query Expansion Using Lexical Resources

- Goal
  - investigate the role of concepts and relations available in WordNet in the expansion of queries, for the purpose of improving the quality of retrieved documents
- Procedure
  - manually or automatically identify WordNet concepts corresponding to query terms
  - collect expansion terms from among the synonym, more general and more specific concepts of the identified concepts
  - expand queries using the expansion terms
- Findings
  - with manual identification of WordNet concepts, the expansion of queries improves results for underspecified queries, and does not improve results for well-specified queries
  - with automatic identification of WordNet concepts, the expansion of queries degrades results

Matching of Queries onto Documents

- [Fan08]: H. Fang. A Re-Examination of Query Expansion Using Lexical Resources. ACL-08.

Query Expansion Using Lexical Resources

- Goal
  - revisit the task of query expansion using concepts and relations available in WordNet
- Procedure
  - focus on the assignment of appropriate weights to expansion terms, such that terms selected for expansion are strongly related to query terms
  - weights capture similarity among query terms and expansion terms
  - term similarity functions use synonym vs. more general vs. more specific concepts vs. overlap of concept definitions
- Findings
  - the expansion of queries with terms from WordNet improves results
  - improvement is largest when similarity between terms is computed as the overlap of their definitions in WordNet
  - combining multiple similarity functions gives no additional improvement
  - query expansion using WordNet is not better than query expansion using expansion terms that co-occur with query terms in the document collection (pseudo-relevance feedback using global analysis)

Matching of Queries onto Documents

- [KZ12]: A. Kotov and C. Zhai. Tapping into Knowledge Base for Concept Feedback: Leveraging ConceptNet to Improve Search Results for Difficult Queries. WSDM-12.

Query Expansion Using Semantic Sources

- Goal
  - investigate the role of concepts and relations available in ConceptNet in the expansion of queries, for the purpose of improving the quality of retrieved documents
  - focus on difficult (i.e., poorly performing) queries
- Procedure
  - manually or automatically identify ConceptNet concepts related to query terms
  - collect expansion terms from among concepts available in the ConceptNet graph of concepts and relations, within a certain distance away from the identified concepts
  - expand queries using the expansion terms
- Findings
  - with manual identification of ConceptNet concepts, there is some possible expansion of queries that improves results, for all difficult queries
  - expansion terms manually selected from ConceptNet give better results than expansion terms automatically selected from top results (pseudo-relevance feedback using local analysis)

Impact of Query Expansion

- Manual selection of ConceptNet concepts for expansion

(Courtesy A. Kotov)

	KL	KL-PF	CF-1	CF-2	CF-3
AQUAINT	0.0521	0.0429	0.1247	0.1622	0.1880
ROBUST04	0.0509	0.0788	0.1061	0.1539	0.1823
GOV	0.0748	0.0447	0.1830	0.3481	0.4826

Concepts from top retrieved documents (pseudo-relevance feedback)

Concepts in ConceptNet within radius 2 of the identified concepts

Impact of Query Expansion

- Automatic selection of ConceptNet concepts for expansion

	KL	KL-PF	QPATH	RWALK	LR-2	LR-PF-2	LR-3	LR-PF-3
AQUAINT	0.0521	0.0429	0.0538	0.0534	0.0535	0.0662	0.0571	0.0776
ROBUST04	0.0509	0.0788	0.0542	0.0559	0.0604	0.0844	0.0588	0.0837
GOV	0.0748	0.0447	0.1034	0.1179	0.1293	0.1119	0.1236	0.0914

Heuristic-based selection of concepts from ConceptNet

Learning-based selection of concepts from ConceptNet

Combination of learning-based selection of concepts from ConceptNet and pseudo-relevance feedback

Next Topic

- Document analysis and understanding
- Query analysis and understanding
- Matching of queries onto documents
- Onebox search results

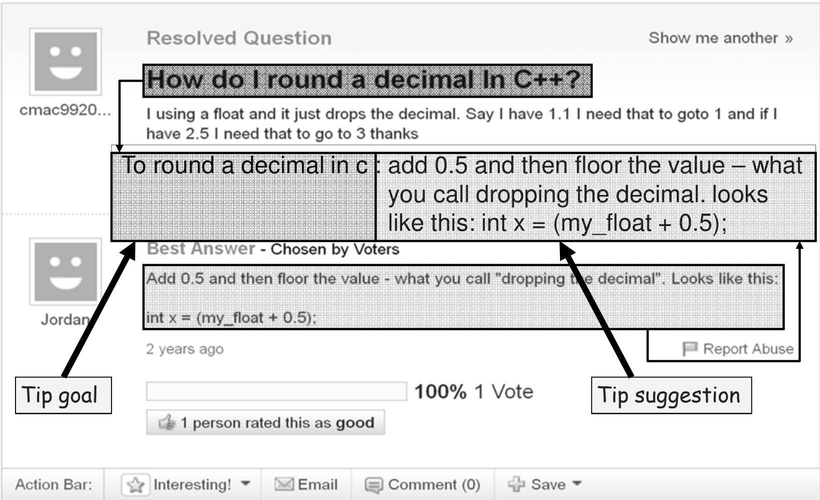
Retrieval of OneBox Results

- [WUG12]: I. Weber, A. Ukkonen and A. Gionis. Answers, not Links: Extracting Tips from Yahoo Answers to Address How-To Queries. WSDM-12.

Extracting and Retrieving How-To Tips

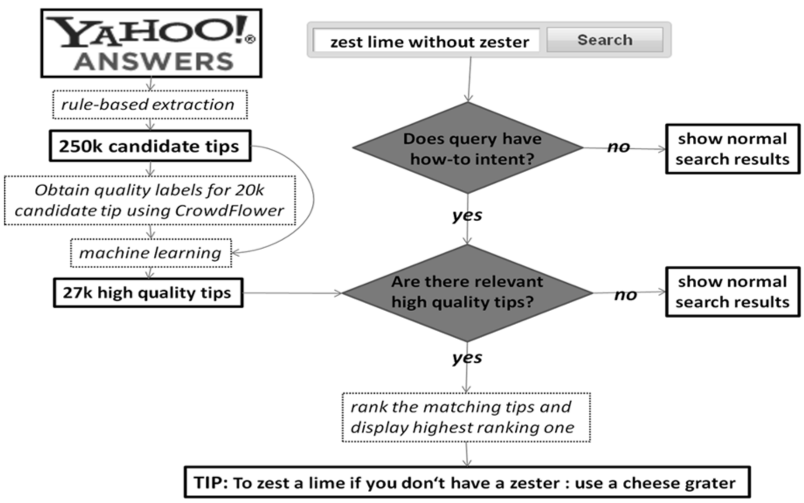
- Input
  - queries
- Data source
  - collaboratively-created collection of pairs of a question and an answer
- Output
  - a tip (to round a decimal in c: add 0.5 and then floor the value) in the format (tip goal: tip suggestion), selected from a set of tips extracted in advance from the question-answer pairs
  - returned only for queries deemed to have how-to intent (how to round a decimal in c, how do you fix keys on a laptop, clean iphone screen)
- Steps
  - construct tips, from pairs of a "how to" question and its answer
  - for queries with how-to intent, retrieve tip whose goal best matches the queries

Tips from Question-Answer Pairs



(Courtesy I. Weber)

Answering How-To Queries



Retrieval of OneBox Results

- [YTL11]: X. Yin, W. Tan and C. Liu. FACTO: a Fact Lookup Engine Based on Web Tables. WWW-11.

Extracting and Retrieving Facts

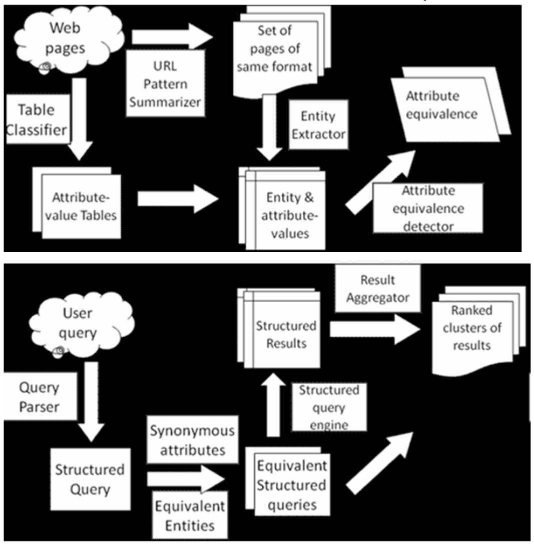
- Input
  - queries
- Data source
  - collection of tables identified within Web documents
- Output
  - one-box search result containing the fact (paris; bay city), if any, deemed to most confidently answer the user's query (france capital; where was madonna born)
  - selected from a set of facts (tuples of an instance, attribute and value) extracted in advance from Web tables
- Steps
  - identify subset of Web tables containing attribute-value pairs
  - from attribute-value pairs in a table, and the instance identified to be the main topic of the document containing the table, extract instance-attribute-value tuples
  - if query is deemed to be a fact lookup query, retrieve the value with the highest confidence among values, if any, present in the tuples for the instance and attribute specified in the query

System Architecture

(Courtesy X. Yin)

Data extraction

Query answering



Extraction of Factual Tuples

- Table classifier
  - distinguish attribute-value tables from other types of tables

	Attribute-Value Tables	Relational Tables
% among all tables	6.6%	1.6%
avg # of instances	1	10.3
avg # of attributes	14.3	3.6
avg. # of data elements	14.3	38.8
% of numerical data elements	8.8%	62.8%

- Pattern summarizer
  - analyzing sets of documents with the same format, identify and discard spurious attribute-value tables

Log in	Contact us
Help	Customer services
About us	Store locations

Britney Spears	Paris Hilton
Jennifer Lopez	Jessica Simpson
Madonna	Jessica Alba

Extraction of Factual Tuples

- Entity extractor
  - extract the main instances about which the source Web documents, and the attribute-value tables that they contain, are about

... → repository of instance-attribute-value tuples

- Attribute equivalence detector
  - attributes that have the same value for the same instance tend to be equivalent

address	phone	price	weight
location	telephone	list price	gewicht
addresse	phone number	regular price	poids
dirección	admissions	our price	peso
street address	tel	your price	waga

Query Answering

- Query parser
  - match queries against small set of manually-written rules ("E A", "E's A", "who was the A of E", "when was E born")
- Instance equivalence detector
  - instances whose vectors of search-result click counts are very similar to one another are deemed equivalent
  - considered very similar, when vectors have cosine ≥ 0.5:

Equivalent?	Cause of Error	Pct	Example of Instance Pair
Yes	N/A	87%	australian job vs. job in australia
No	One is more specific than the other	7%	flightless bird vs. large flightless bird
No	One is an aspect of the other	5%	will county vs. map of will county
No	Different	1%	1972 chevrolet suburban vs. 1968 chevrolet suburban

Query Answering

- Structured query engine
  - given a query, generate instance-attribute pairs by replacing entity with equivalent entity or attribute with equivalent attribute
  - lookup instance-attribute pairs in instance-attribute-value tuples
- Result aggregator
  - single or multiple lookups may result in retrieval of multiple values
  - select value if extracted from more Web domains, and if similar to more of the other values

Type of Answering Error	Example of Query – Erroneous Answer
answer is wrong	turkey language - english
answer is incomplete	george bush date of birth – 1946
answer is relevant for another query	how santa monica college was founded – 1929
query is an instance	microsoft publisher – (any)
query is navigational	lil wayne myspace – (any)
query should not trigger an answer	watch free movies – (any)

Retrieval of OneBox Results

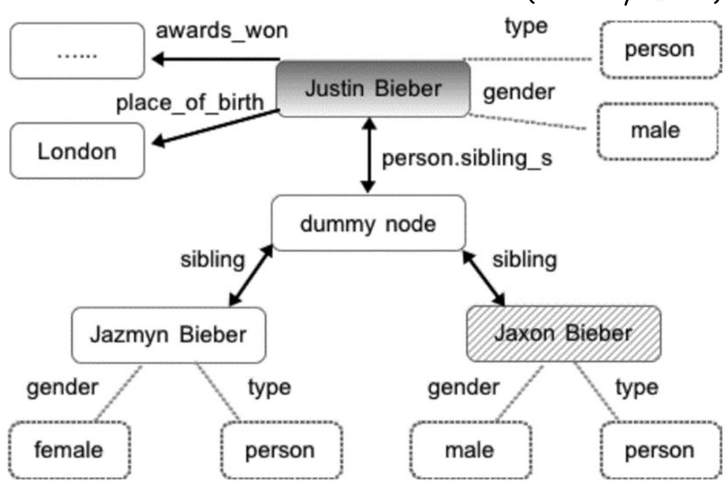
- [YV14]: X. Yao and B. Van Durme. Information Extraction over Structured Data: Question Answering with Freebase. ACL-14.

Question Answering as Binary Classification

- Input
  - natural-language questions (What is the name of the Justin Bieber's brother?)
- Data source
  - knowledge repository of inter-connected topics (Freebase)
  - collection of Web documents
- Output
  - topics that answer the questions (Jaxon Bieber)
- Steps
  - convert the question into a question graph
  - based on the node from the question graph corresponding to the question topic (Justin Bieber), assemble a topic graph of inter-connected topics up to a few hops away from the question topic
  - using individual and combination features from question graph and topic graph, determine whether each node from the topic graph is or is not an answer to the question

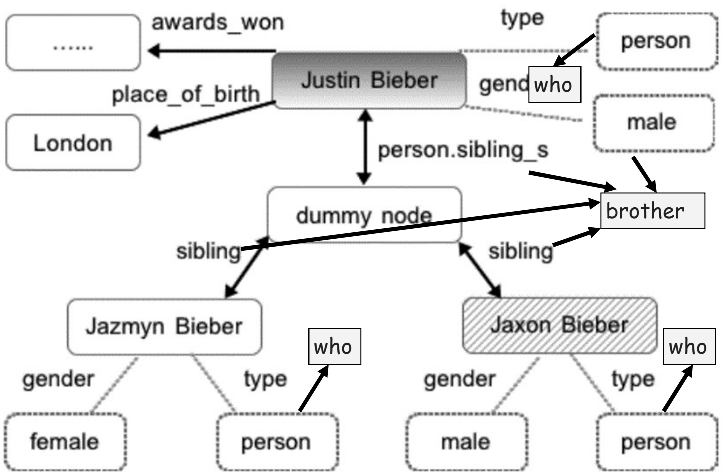
Answers from Knowledge Repositories

(Courtesy X. Yao)



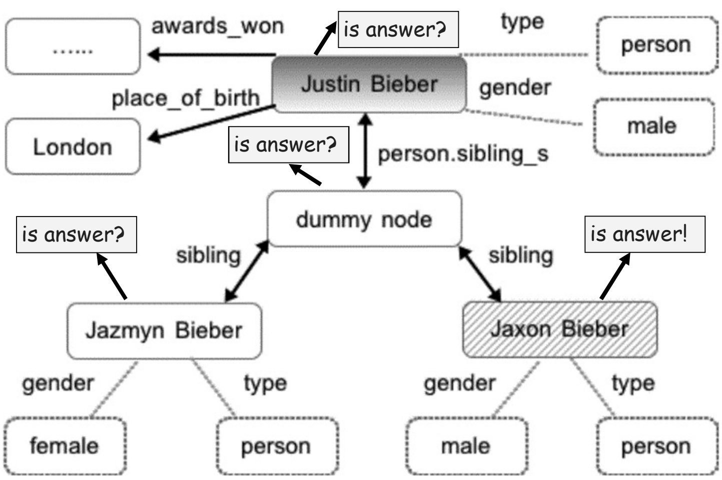
Question: What is the name of Justin Bieber's brother?

Answers from Knowledge Repositories



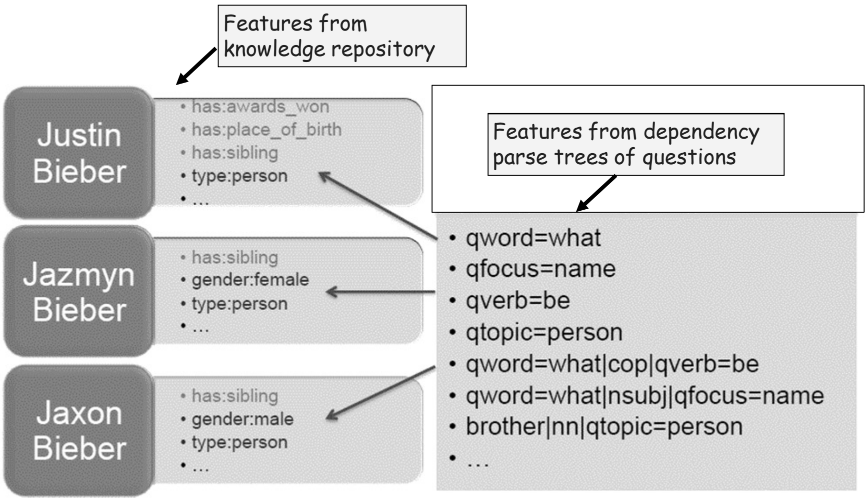
Question: What is the name of Justin Bieber's brother?

Question Answering as Classification

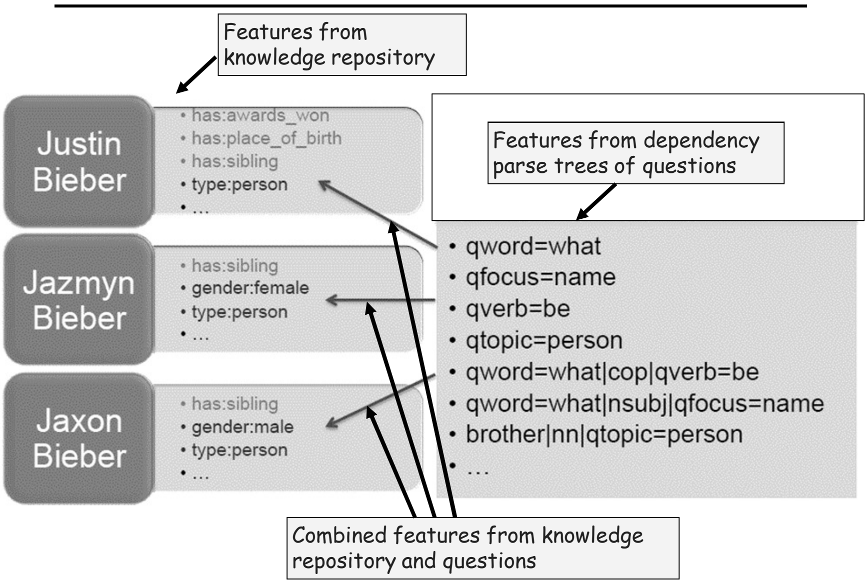


Question: What is the name of Justin Bieber's brother? Answer: Jaxon Bieber

Features for Classification



Features for Classification



Estimated Utility of Features

Justin Bieber	expected weights
• has:awards_won | qword=what	• medium
• has:place_of_birth | qword=what	• low
• has:sibling | qfocus=name	• low
• type:person | qfocus=name	• medium
• …	• …

Jazmyn Bieber	expected weights
• has:sibling | brother | nn | qtopic=person	• high
• gender:female | brother | nn | qtopic=person	• low
• type:person | brother | nn | qtopic=person	• high
• …	• …

Jaxon Bieber (is answer)	expected weights
• has:sibling | brother | nn | qtopic=person	• high
• gender:male | brother | nn | qtopic=person	• high
• type:person | qword=what | nsubj | qfocus=name	• high
• …	• …

Mapping Relations to Phrases

- Align relations from knowledge repository to phrases that may express the relations in document sentences
  - film/starring (Gravity, Sandra Bullock)
    vs.
  - Sandra then was cast in Gravity, a two actor spotlight film
  - Sandra Bullock plays an astronaut hurtling through space in new blockbuster "Gravity"
  - Sandra Bullock stars/acts in Gravity
  - Sandra Bullock conquered her fears to play the lead in Gravity
- Use alignments to predict relevant relations when answering questions

feature	weight	feature	weight
qfocus=religion|type=Religion	8.60	qword=when|type=datetime	5.11
qfocus=money|type=Currency	5.56	qverb=border|rel=location.adjoins	4.56
qverb=die|type=CauseOfDeath	5.35	qverb=go|qtopic=location|type=Tourist attraction	2.94

Summary

- Knowledge and its acquisition from textual data have the potential to enhance Web search
  - sources of textual data: documents, queries
  - impact on content understanding: query and document analysis, query-document matching
  - impact on alternative search interfaces: structured search, answer retrieval