



IIT-H

Web Mining

**Lecture 20: Entity Semantics Mining
(Part 2)**

Manish Gupta

9th Oct 2013

Slides borrowed (and modified) from

<http://research.microsoft.com/pubs/183817/www2013-acronym-mining.pptx>

Recap of Lecture 19: Entity Semantics Mining (Part 1)

- Entity Synonyms
- Entity Attribute Discovery and Augmentation
- Entity Linking

Announcements

- Happy midterms! We don't have a midterm for this course 😊
- We will meet next on Oct 23
 - Use this time to work on projects
 - 7% of coursework (Mid-project Report) is due on 27th Oct
 - motivation for the problem, problem definition, short summary of related papers you have read, and your methodology
 - 2-3 pages, single spacing, double column
 - Two main topics we will focus on for the next 6 lectures: Query log mining and crowdsourcing

Today's Agenda

- Entity Set Expansion
- Entity Acronym Expansion
- Entity Actions
- Entity Tagging

Today's Agenda

- **Entity Set Expansion**
- Entity Acronym Expansion
- Entity Actions
- Entity Tagging

Entity Set Expansion

- Yeye He, Dong Xin. SEISA: Set Expansion by Iterative Similarity Aggregation. WWW 2011.
- Aim is to extend set like (Canon, Nikon) with other related entities like Olympus.
- Data sources like web lists and user queries can be used, but they tend to be very noisy.
- Random walk approaches (Set Expander for Any Language - SEAL system) do not work well with noisy data sources.
- Hence a general framework based on iterative similarity aggregation is proposed.

What's New?

- Previous approaches (SEAL)
 - Build customized wrappers for each web page using left/right context that would enclose all given seeds
 - Model web pages, wrappers and candidate terms as graph nodes and perform random walk
 - Candidates close to the given seeds in the graph structure are more likely to belong to the same concept as the seeds
- Proposed approach
 - Set of expanded results is good if
 - Relevance: Set of produced entities are similar to the given seeds
 - Coherence: Set of produced entities are coherent in the sense that they represent a consistent concept

Bipartite Graph Data Model

- For web list data, weight on edge is 1.
- For query log data, weight on edge is mutual information between query term and query context.
 - Mutual information between term t and context c is $H(t, c) = \frac{P(t, c)}{P(t)P(c)}$
- Similarity measures
 - Jaccard Similarity
 - L_x and L_y be 2 sets of right hand side nodes that connect to nodes x and y
 - $Sim_{Jac}(x, y) = \frac{|L_x \cap L_y|}{|L_x \cup L_y|}$
 - Cosine Similarity
 - V_x and V_y be weight vectors that indicate weights of edges that connect web lists to node x and y
 - $Sim_{Cos}(x, y) = \frac{V_x \cdot V_y}{\|V_x\| \cdot \|V_y\|}$

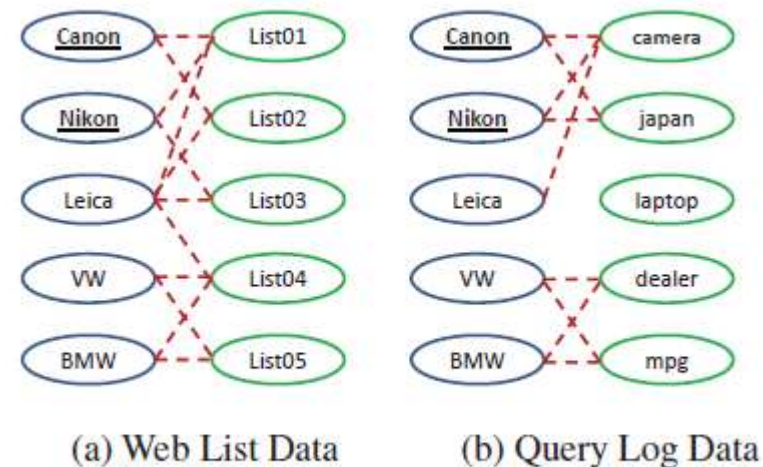


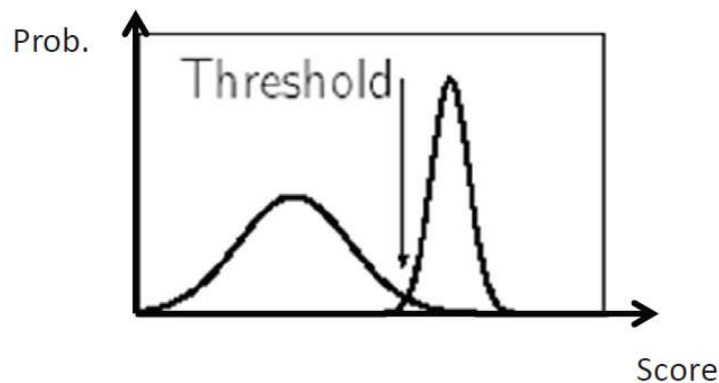
Figure 1: Bipartite graph data model

Relevance and Coherence

- Let U be the universe of all entities. S be seed set and R be the expanded set
- Relevance: $S_{rel}(R, S) = \frac{1}{|R||S|} \sum_{r \in R} \sum_{s \in S} Sim(r, s)$
- Coherence: $S_{coh}(R) = \frac{1}{|R||R|} \sum_{i=1}^{|R|} \sum_{j>i}^{|R|} Sim(r_i, r_j)$
- Quality of expanded seed set (ESS) R of size K is $Q(R, S) = \alpha S_{rel}(R, S) + (1 - \alpha) S_{coh}(R)$
- Individual terms t can then be ranked as follows
 - $g(t, R, S) = \frac{\alpha}{|S|} \sum_{i=1}^{|S|} Sim(t, s_i) + \frac{1-\alpha}{|R|} \sum_{i=1}^{|R|} Sim(t, r_i)$
- Given S and K , finding R that maximizes $Q(R, S)$ is NP hard

Iterative Similarity Aggregation Algorithm

- Static Thresholding Algorithm
 - Fix size of ESS R at the beginning, and then iteratively search for terms in R to maximize $Q(R,S)$
- Dynamic Thresholding Algorithm
 - Refines both size of R and contents of R at the same time in each iteration
- Threshold Computation given a vector
 - Otsu's algorithm
 - $T = \operatorname{argmin}_x f(x)$ where $f(t) = w_1 \sigma_1^2(t) + w_2 \sigma_2^2(t)$
 - w_1 and w_2 are the probabilities of the 2 classes and σ_1^2 and σ_2^2 are the variances of the 2 classes



Thresholding Algorithms

Static
Dynamic

Algorithm 1 Static Thresholding Algorithm

```

Static_Thresholding (seeds, graph)
  for each  $term_i$  in graph.terms do
     $Rel\_Score[i] \leftarrow S_{rel}(term_i, seeds)$ 
  end for
  sort  $term_i$  by  $Rel\_Score[i]$  desc
   $K \leftarrow \text{Pick\_Threshold}(Rel\_Score[i])$ 
   $R_0 \leftarrow$  the top  $K$  ranked terms by  $Rel\_Score[i]$ 
   $iter \leftarrow 1$ 
  while true do
    for each  $term_i$  in graph.terms do
       $Sim\_Score[i] \leftarrow S_{rel}(term_i, R_{iter-1})$ 
       $g(term_i) \leftarrow \alpha * Rel\_Score[i] + (1 - \alpha) * Sim\_Score[i]$ 
    end for
    sort  $term_i$  by  $g(term_i)$  desc
     $R'_{iter} \leftarrow$  the top  $K$  terms by  $g(term_i)$ 
    if  $R'_{iter} \neq R_{iter-1}$  then
      let  $r \in R'_{iter}$  be the top ranked term not in  $R_{iter-1}$ 
      let  $q \in R_{iter}$  be the last ranked term in  $R_{iter-1}$ 
       $R_{iter} \leftarrow (R_{iter-1} \cup \{r\}) - \{q\}$ 
    else
       $R_{iter} \leftarrow R_{iter-1}$ 
      break
    end if
     $iter++$ 
  end while
  return  $R_{iter}$ 

```

Algorithm 2 Dynamic Thresholding Algorithm

```

Dynamic_Thresholding (seeds, graph)
  for each  $term_i$  in graph.terms do
     $Rel\_Score[i] \leftarrow S_{rel}(term_i, seeds)$ 
  end for
   $K_0 \leftarrow \text{Pick\_Threshold}(Rel\_Score[i])$ 
  sort  $term_i$  by  $Rel\_Score[i]$  desc
   $R_0 \leftarrow$  the top ranked  $K_0$  terms by  $Rel\_Score[i]$ 
   $iter \leftarrow 1$ 
  while  $iter \leq MAX\_ITER$  do
    for each  $term_i$  in graph.terms do
       $Sim\_Score[i] \leftarrow S_{rel}(term_i, R_{iter-1})$ 
       $g(term_i) \leftarrow \alpha * Sim\_Score[i] + (1 - \alpha) * Rel\_Score[i]$ 
    end for
     $K_{iter} \leftarrow \text{Pick\_Threshold}(g(term_i))$ 
    sort  $term_i$  by  $g(term_i)$  desc
     $R'_i \leftarrow$  the top ranked  $K_{iter}$  terms by  $g(term_i)$ 
     $R_{iter} \leftarrow R'_{iter}$ 
     $iter++$ 
  end while
  return  $R_{iter}$ 

```

Today's Agenda

- Entity Set Expansion
- **Entity Acronym Expansion**
- Entity Actions
- Entity Tagging

The Popularity of Acronyms

- Acronym: abbreviations formed from the initial components of words or phrases
 - E.g., CMU, MIT, RISC, MBA, ...
- Acronyms are very commonly used in
 - Web search
 - Tweets
 - Text messages
 - ...
- Even more common on mobile devices

Acronym Characteristics

- Ambiguous: one acronym can have many different meanings
 - E.g., CMU can refer to “Central Michigan University”, “Carnegie Mellon University”, “Central Methodist University”, and many other meanings
- Disambiguated by context: the meaning is often clear when context is available
 - “cmu football” -> “Central Michigan University”
 - “cmu computer science” -> “Carnegie Mellon University”

Application Scenario

- Web Search
 - Acronym Queries
 - Suggest the different meanings of the input acronym, or expand to the most likely intended meaning
 - Acronym + Context Queries
 - Infer the most likely intended meaning given the context and then perform query alteration, e.g., “cmu football” -> “central michigan university football”

Problem Statement

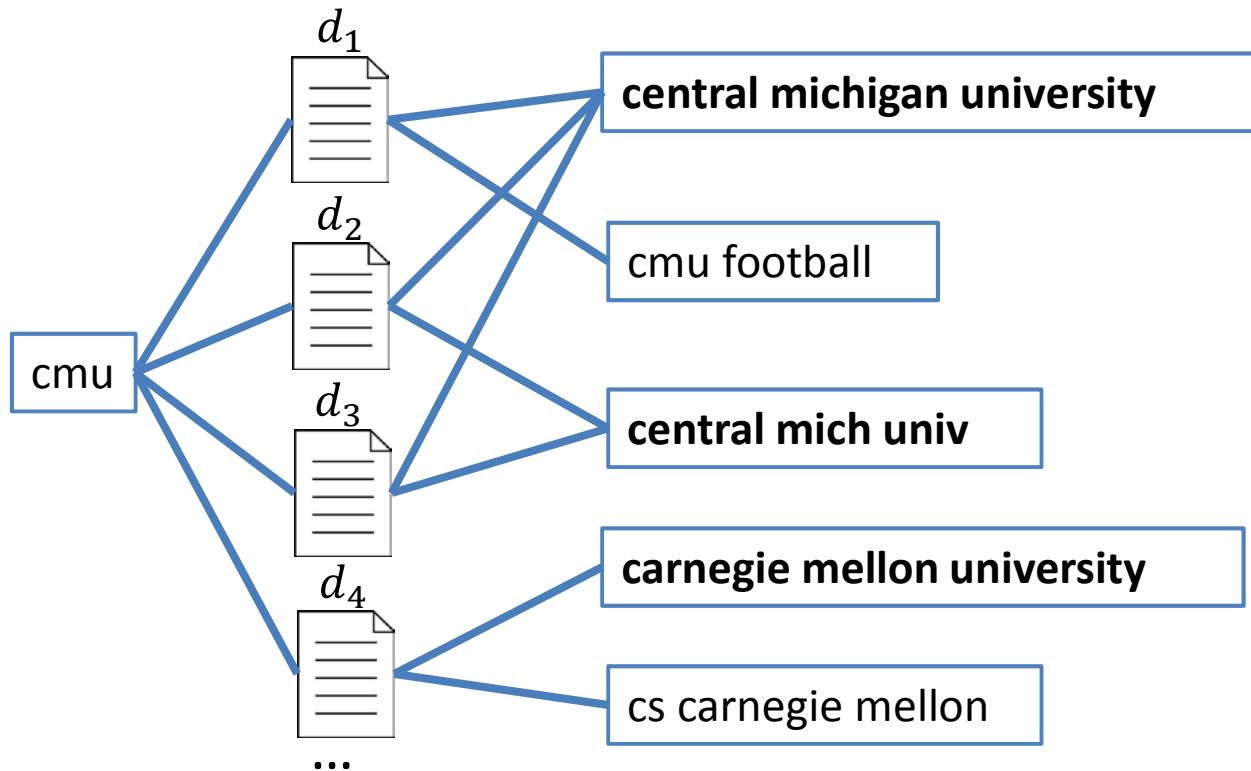
- Input: an acronym
- Output: the various different meanings of the acronym; each meaning is represented by its canonical expansion, a popularity score and a set of associated context words

Input
CMU



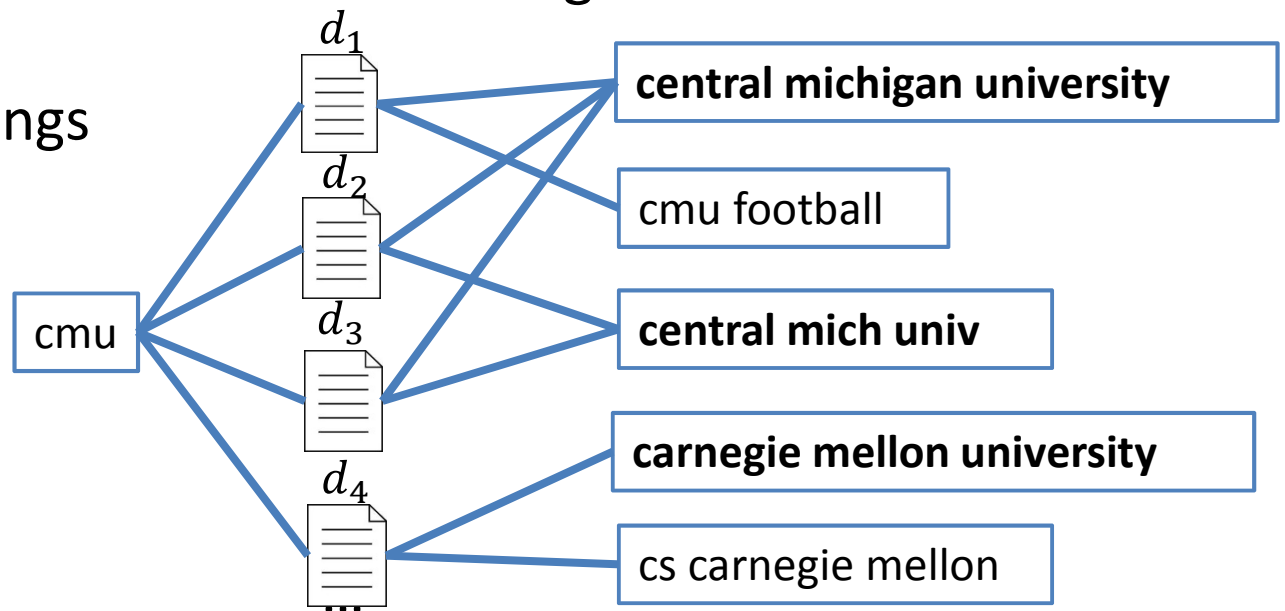
Meaning	Popularity	Context Words
central michigan university	0.615	michigan, athletics, football, ...
carnegie mellon university	0.312	pittsburgh, library, computer, ...
concrete masonry unit	0.045	block, concrete, cement, ...
central methodist university	0.017	fayette, central, missouri, ...
canton municipal utilities	0.004	court, docket, case, ...

Insight: Exploiting Query Co-click



Technical Challenges

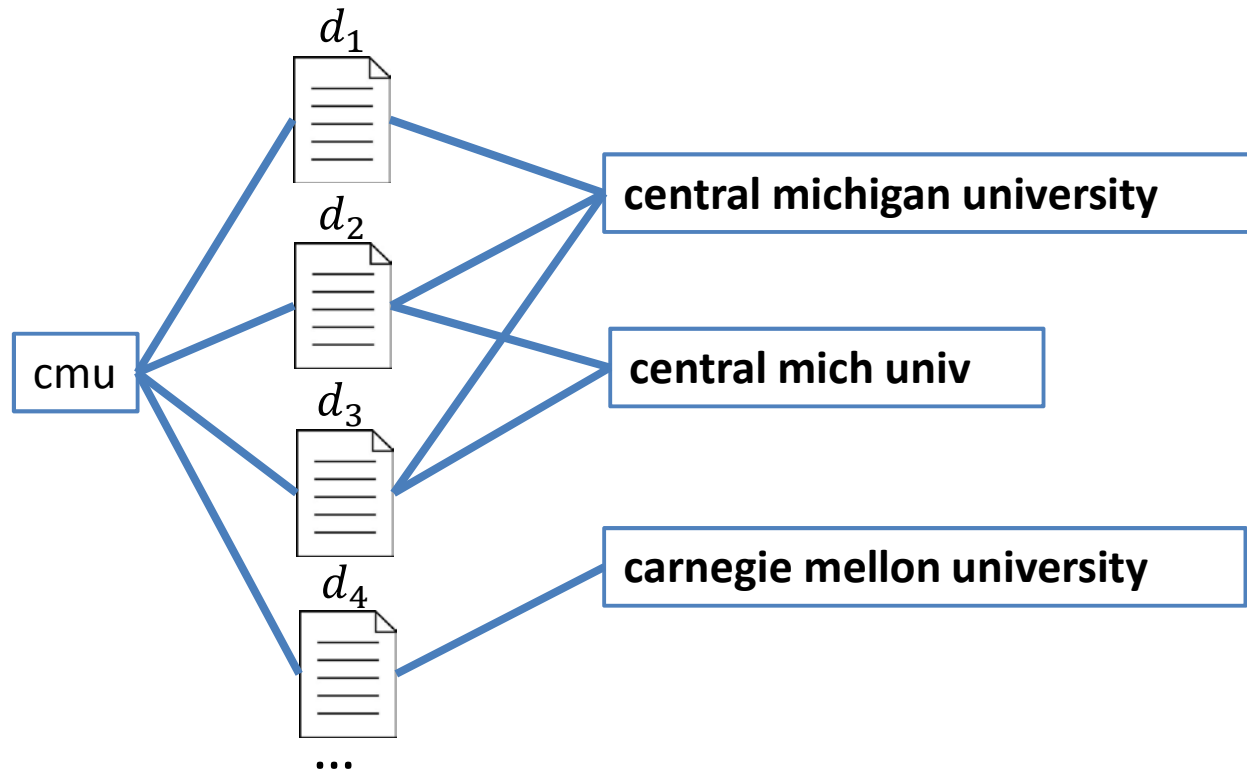
- Identify co-clicked queries that are expansions
 - Mined expansions are often noisy, containing variants for the same meaning
 - Identify context words for each meaning
 - Handle tail meanings



Mining Steps

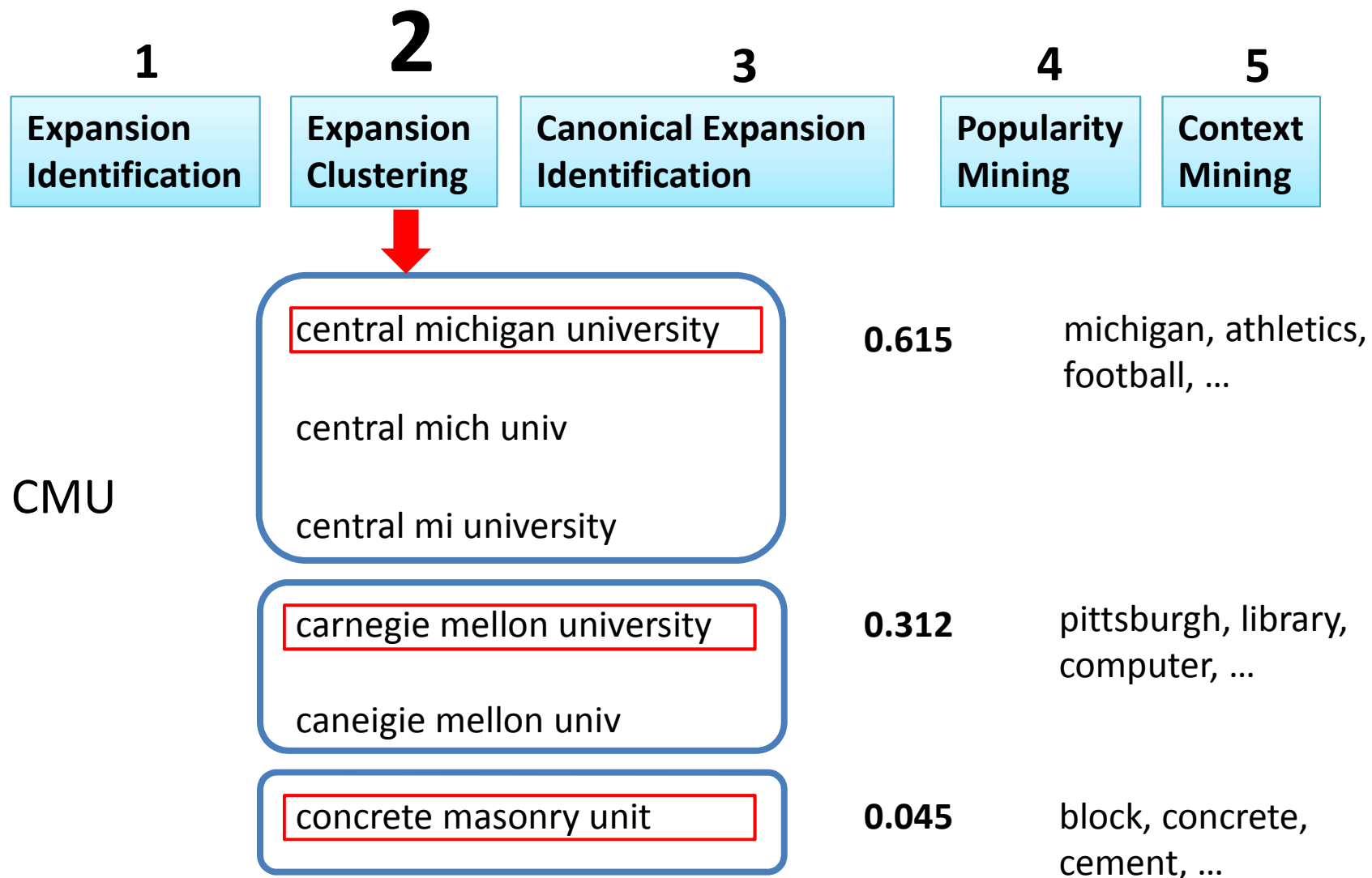
	1	2	3	4	5
	Expansion Identification	Expansion Clustering	Canonical Expansion Identification	Popularity Mining	Context Mining
CMU		<div>central michigan university</div> <div>central mich univ</div> <div>central mi university</div>		0.615	michigan, athletics, football, ...
		<div>carnegie mellon university</div> <div>caneigie mellon univ</div>		0.312	pittsburgh, library, computer, ...
		<div>concrete masonry unit</div>		0.045	block, concrete, cement, ...

Acronym Candidate Expansion Identification



- Rely on Acronym-Expansion Checking Function
 - Not a trivial task, e.g., “Hypertext Transfer Protocol” for “HTTP”, “Master of Business Administration” is for “MBA”

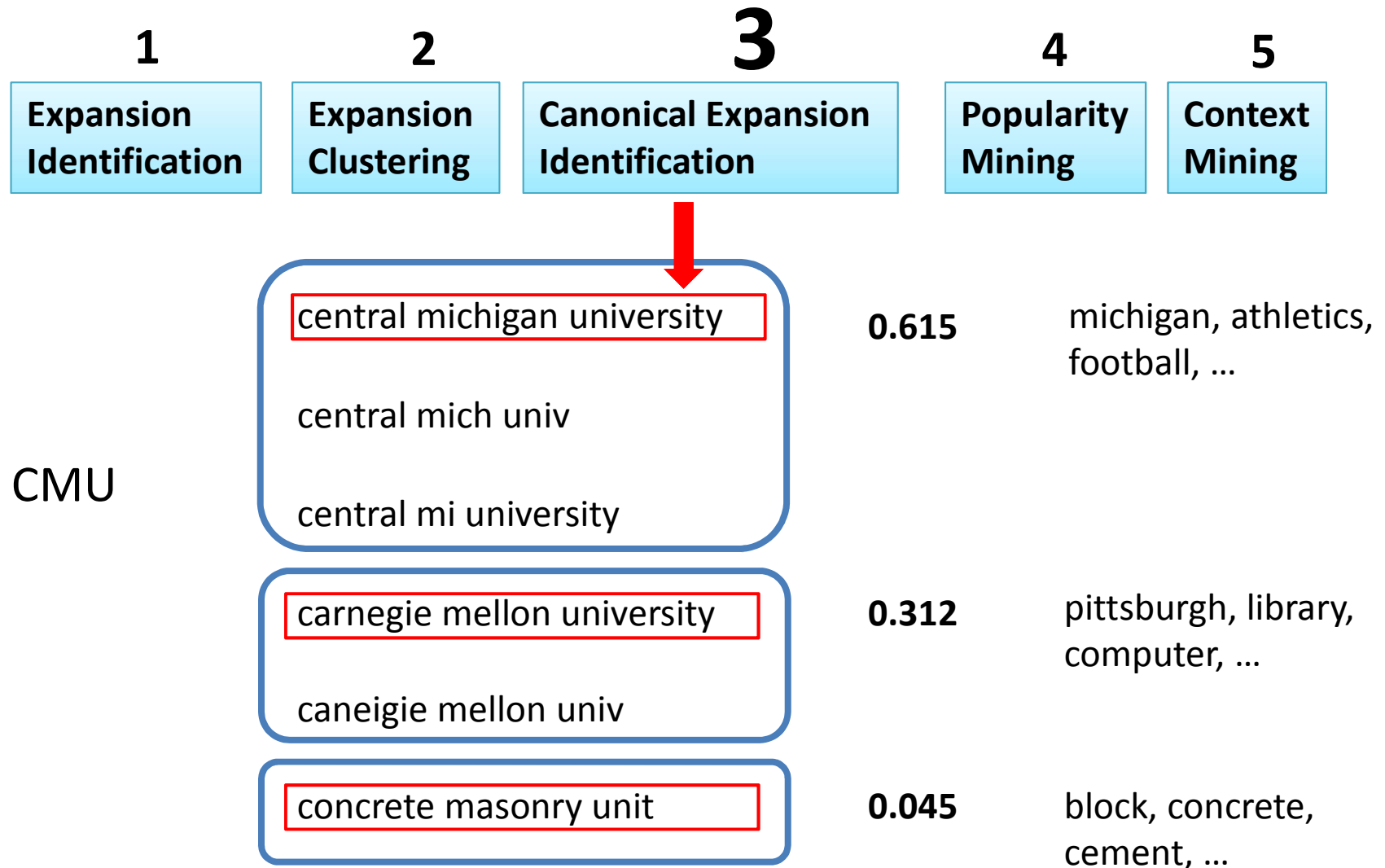
Mining Steps



Acronym Expansion Clustering

- Edit distance is inadequate
 - E.g, “central michigan university” and “central mich univ”
- Insight: leveraging clicked documents
 - Each document typically corresponds to a single meaning
 - Expansion of same meaning click on same set of documents, and expansion of different meanings click on different documents
- Clicked document based distance
 - Set distance (Jaccard distance)
 - Distributional distance (Sqrt of Jensen-Shannon Divergence)
 - $JSD(P || Q) = 0.5 KL(P || M) + 0.5 KL(Q || M)$
 - where $M = 0.5 (P + Q)$ and $KL(P || Q) = \sum_i \ln \left(\frac{P(i)}{Q(i)} \right) P(i)$

Mining Steps



Identifying Canonical Expansion

- The probability that a click of acronym query a on document d is intended for expansion e_k

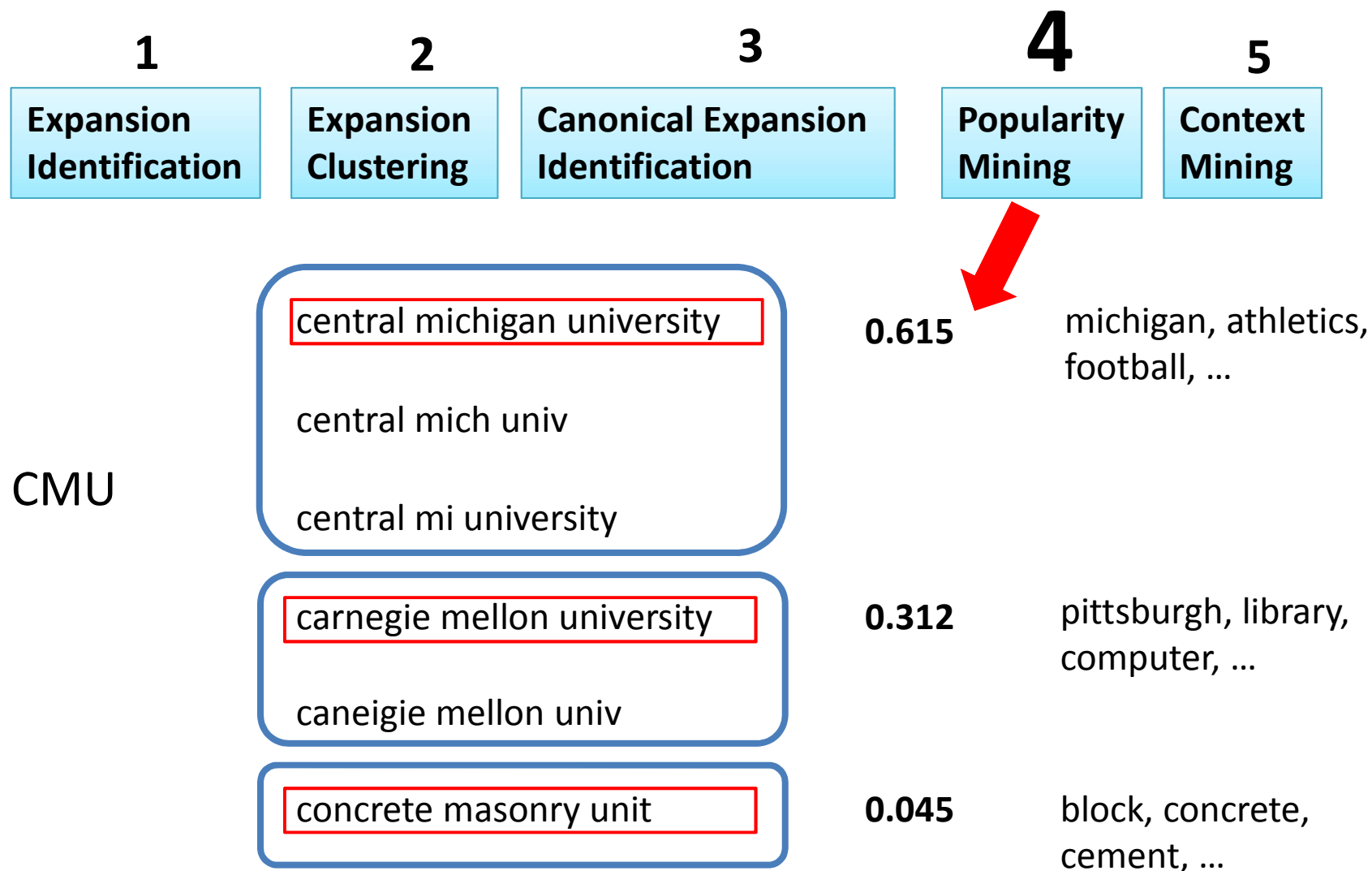
$$Pr(e_k, d) = \frac{F(e_k, d)}{\sum_{G_l \in \mathcal{G}} \sum_{e_j \in G_l} F(e_j, d)}$$

- The probability that acronym query a is intended for expansion e_k

$$e_k.p = \frac{\sum_{d \in D(a)} F(a, d) Pr(e_k, d)}{\sum_{d \in D(a)} F(a, d)}$$

- For each meaning group, canonical expansion is the one with the highest probability
- Note: $F(a, d)$ is click frequency of a for document d .

Mining Steps

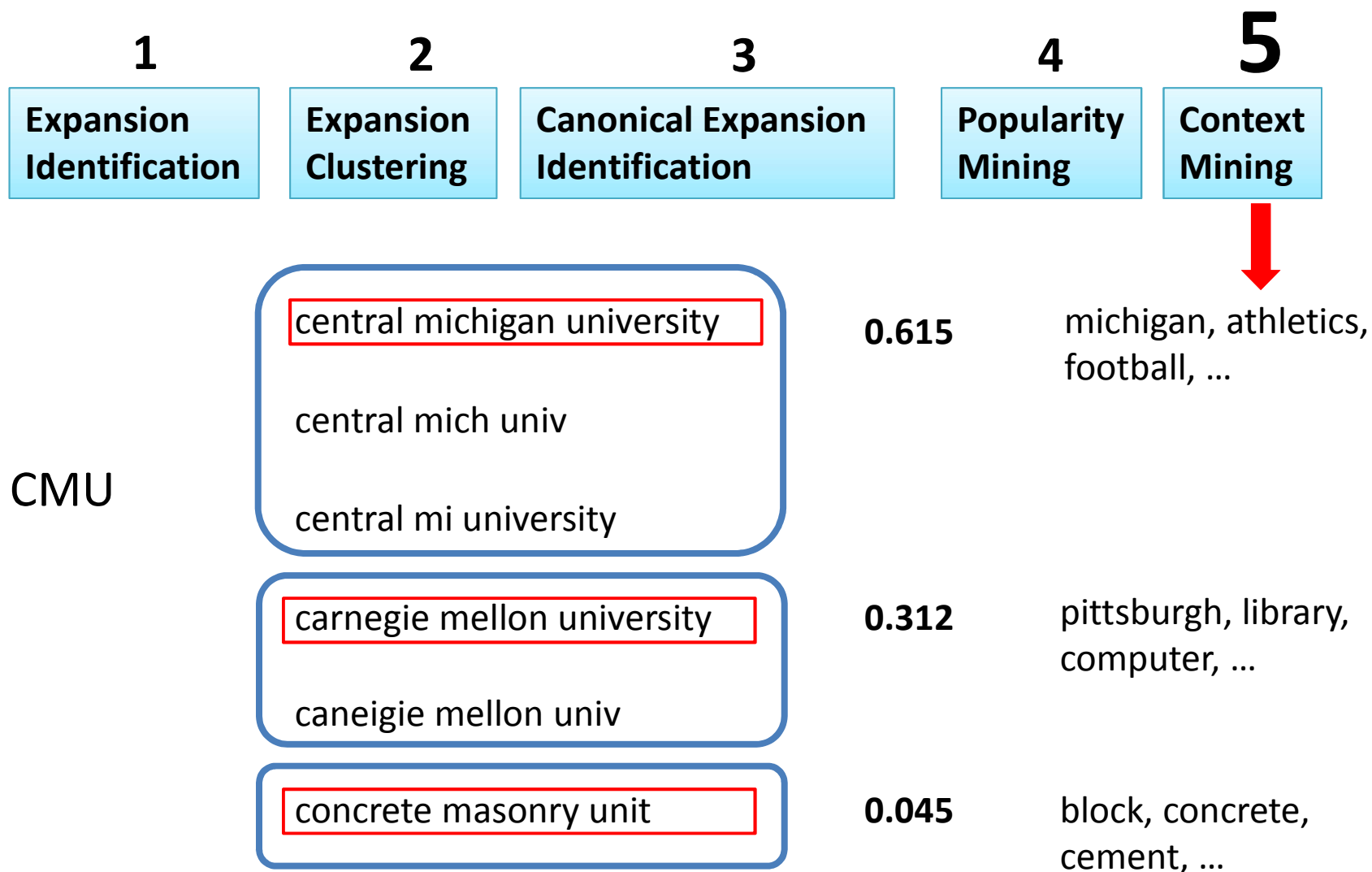


Measure Meaning Popularity

- Remember we mined the probability for an expansion when identifying the canonical expansion
- The popularity for a meaning M_i for acronym a is the aggregated popularity of all the expansions in its group

$$M_i.p = \sum_{e_k \in G_i} e_k.p$$

Mining Steps

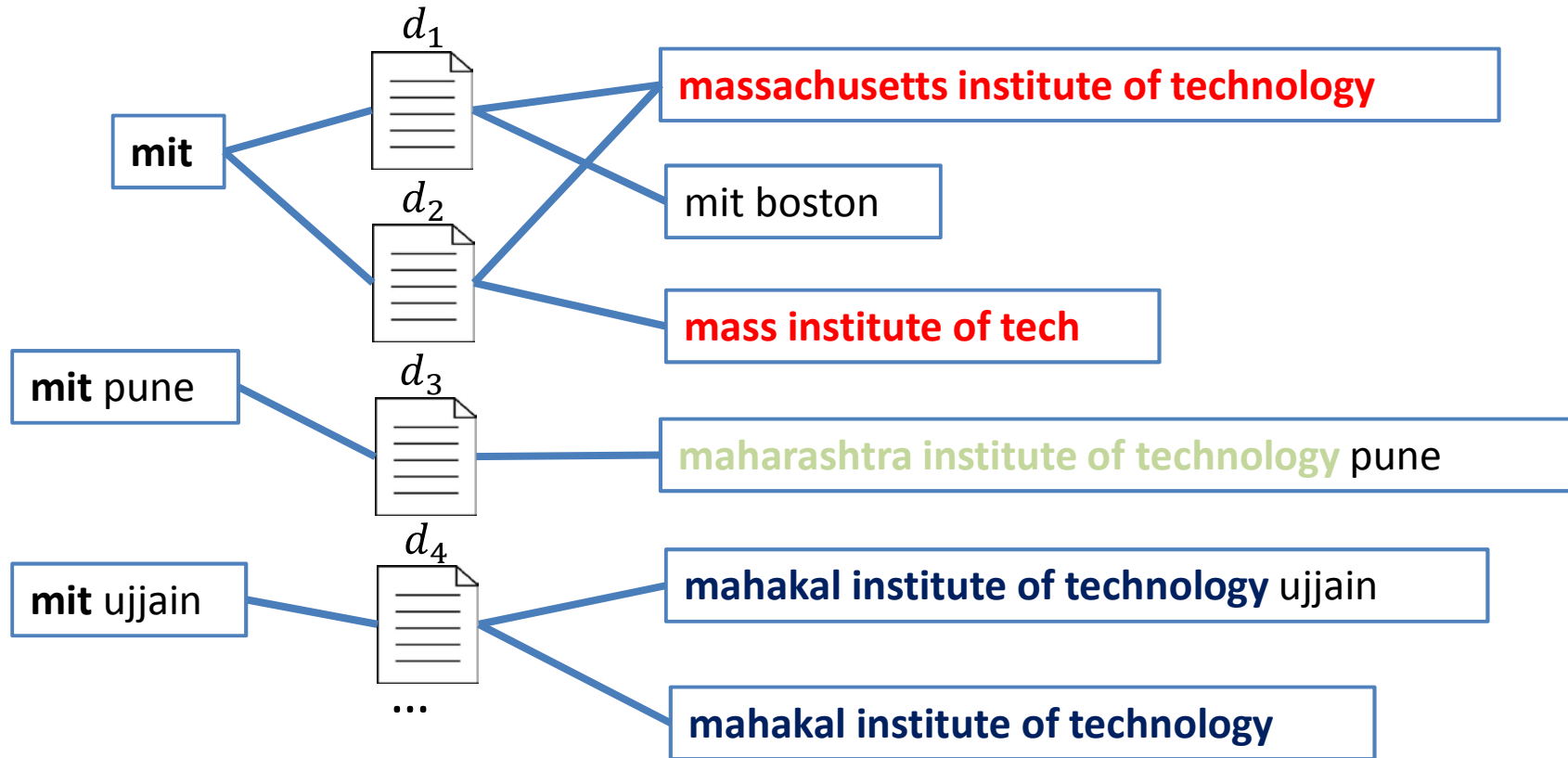


Compute Context Words for Each Meaning

- Consider the set of documents clicked by expansions in group G_i , we treat all the words from queries clicked on these documents as the context words for the meaning group
- Let $F(w, G_i)$ be the aggregated frequency of a word w in group G_i , the probability of a word given a meaning is:

$$Pr(w|M_i) = \frac{F(w, G_i)}{\sum_{w' \in M_i.C} F(w', G_i)}$$

Enhancement for Tail Meanings



While the set of co-clicked queries for the acronym *a* covers the popular meanings of *a*, it does not cover many of the less popular meanings (referred to as “tail meanings”).

Expansion Identification (Enhanced)

- Since users searching for tail meanings do not find the desired documents when they use only the acronym as a query, they use additional words to disambiguate the query.
- Consider acronym supersequence queries
 - E.g, “mit pune”, “mit ujjain”, etc.
- Identify expansions from the co-clicked queries of the acronym supersequence queries
 - E.g, “**maharashtra institute of technology** pune”, “**mahakal institute of technology** ujjain”, etc.

Expansion Clustering (Enhanced)

- The expansion supersequence queries that correspond to the same meaning will click on the same set of documents, whereas expansion supersequence queries corresponding to different meanings will click on different sets of documents.
- The same expansions can have multiple corresponding expansion supersequence queries; we need to compute the distance between two expansions by aggregating the distances between the corresponding expansion supersequence queries
- Need to aggregate across supersequence queries
 - E.g., “mahakal institute of technology ujjain”, “mahakal institute of technology india”, ...
- Distance aggregation
 - For each supersequence pair, compute the distance and then aggregate the distances over all supersequence pairs
- Click frequency aggregation
 - For each expansion, consider all the documents, including the ones clicked by supersequence queries, and then compute the distributional distance on the aggregated click distribution

Application: Online Meaning Prediction

- Given an acronym and context, predict the meaning of the acronym under that context
- Given a context word w , the probability that the intended meaning is M_i is calculated as follows:

$$Pr(M_i|w) = \frac{Pr(w|M_i)Pr(M_i)}{Pr(w)}$$

- This can be extended to handle context with multiple words

$$Pr(M_i|w_1, \dots, w_k) = \frac{Pr(w_1, \dots, w_k|M_i)Pr(M_i)}{Pr(w_1, \dots, w_k)}$$

Experiments

- Data: 100 input acronyms sampled from Wikipedia disambiguation pages
- Compared methods
 - Edit Distance based Clustering (**EDC**)
 - Jaccard Distance based Clustering (**JDC**)
 - Acronym Expansion Clustering (**AEC**)
 - Enhanced Acronym Expansion Clustering (**EAEC**)
- Ground Truth
 - Wikipedia meanings: Wikipedia disambiguation page
 - Golden standard meanings: manually captured from co-clicked queries

Evaluation Measures

- The algorithms output a set of groups $G(a) = \{G_1, \dots, G_n\}$ which maps to golden standard meanings $M = \{M_1, \dots, M_k\}$ for a given acronym a .
- Standard measures used for evaluating clustering, specifically
 - **Purity**: how pure are the meaning clusters
 - Let $N = \sum_{i=1}^n \sum_{j=1}^k |G_i \cap M_j|$ then $purity(G, M) = \frac{1}{N} \sum_{i=1}^n \max_j |G_i \cap M_j|$
 - **Normalized Mutual Information (NMI)**: considering both the quality of clusters and the number of clusters
 - **Recall**: number of meanings found with respect to the Golden Standard

Meanings, Popularity and Context Words

	Meaning	Probability	Context Words
CMU	central michigan university	0.615	michigan, university, athletics, campus, edu, football, chippewas
	carnegie mellon university	0.312	mellon, carnegie, pittsburgh, university, library, computer, engineering
	concrete masonry unit	0.045	block, concrete, cmu, masonry, cinder, cement, construction
	central methodist university	0.017	methodist, university, fayette, central, missouri, baseball
	canton municipal utilities	0.004	canton, court, municipal, docket, case, clerk, records
RISC	reduced instruction set computer	0.737	risc, instruction, set, computer, processor, architecture
	rice insurance services company	0.143	insurance, rice, risceo, services, real, estate
	rna induced silencing complex	0.046	complex, rna, silencing, gene, protein
	reinventing schools coalition	0.037	schools, coalition, inventing, alaska
	recovery industry services company	0.022	recovery, certified, specialist, matrix, educational
MBA	master of business administration	0.868	mba, business, gmat, administration, harvard, programs, degree
	mortgage bankers association	0.069	mortgage, bank, implode, amerisave, bankers, rates
	montgomery bell academy	0.022	bell, montgomery, academy, nashville, mba, school, edu
	metropolitan builders association	0.015	builders, homes, association, wisconsin, milwaukee
	military benefit association	0.006	military, armed, association, benefits, insurance, veterans

Mining Results

- AEC > JDE > EDC: weighting by click frequency helps

	Purity	NMI
EDC	0.956	0.862
JDC	0.999	0.918
AEC	0.998	0.999

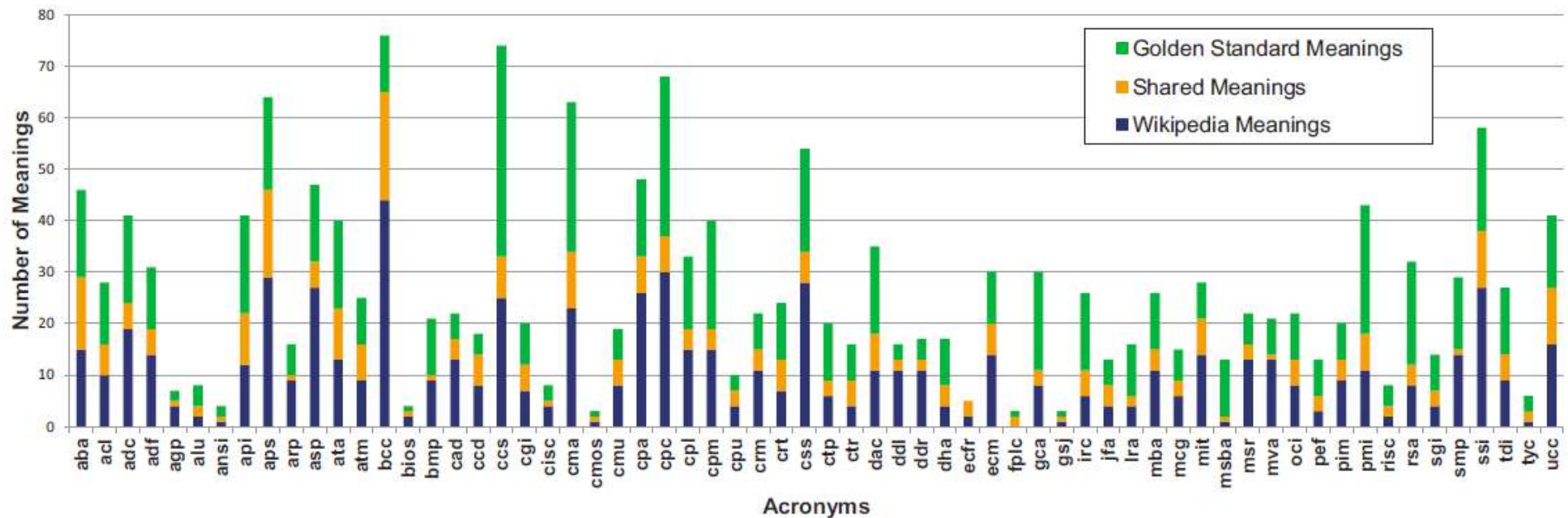
- EAEC > ACE: exploiting supersequence queries boost recall

	Purity	NMI	Recall
AEC	0.993	0.994	0.801
EAEC	0.996	0.995	0.996

Wikipedia and Golden Standard Meanings

	Only Wikipedia Meanings	Only Golden Standard Meanings	Shared Meanings
CMU	caribbean medical university chiang mai university california miramar university colorado mesa university coffman memorial union college music update complete music update communication management unit	central methodist university canton municipal utilities centrul medical unirea case management unit central mindanao university central missouri university	carnegie mellon university central michigan university canadian mennonite university concrete masonry unit couverture maladie universelle
RISC	rural infrastructure service commons research institute for symbolic computation	rice insurance services company reinventing schools coalition recovery industry services company rhode island statewide coalition	reduced instruction set computing rna induced silencing complex
MBA	maldives basketball association marine biological association metropolitan basketball association media bloggers association milwaukee bar association monterey bay aquarium macbook air main belt asteroid market basket analysis miss black america misty's big adventure	metropolitan builders association military benefit association master builders association mississippi basketball association mountain bike action massachusetts bar association mariana bracetti academy missionary baptist association morten beyer agnew mind body awareness memphis business academy	master of business administration mortgage bankers association montgomery bell academy mountain bothy association

Wikipedia vs. Golden Standard Meanings



Online Meaning Prediction Results

- Data: 7,612 acronym+context queries
- Each query is manually labeled to the most probable meaning by judges.
 - Examples:

Query	Label
cmu michigan	central michigan university
cmu robotics institute	carnegie mellon university
cmu pittsburgh	carnegie mellon university
cmu fayette missouri	central methodist university

- Average Precision: **94.1%**

Summary

- We introduce the problem of finding distinct meanings of each acronym, along with the canonical expansion, popularity score and context words
- We present a novel, end-to-end solution leveraging query click log
- We demonstrate the mined information can be used effectively for online queries in web search

Today's Agenda

- Entity Set Expansion
- Entity Acronym Expansion
- **Entity Actions**
- Entity Tagging

Active Objects: Objects with Actions

- Thomas Lin, Patrick Pantel, Michael Gamon, Anitha Kannan, Ariel Fuxman. Active Objects: Actions for Entity-Centric Search. WWW 2012.
- Beyond the Ten Blue Links

The image is a screenshot of a Yahoo! search result page for the movie 'Seven Samurai'. The search bar at the top contains the text 'Seven Samurai' and shows '82,800,000 result'. Below the search bar are tabs for 'WEB', 'IMAGES', 'VIDEO', 'SHOPPING', 'BLOGS', and 'MORE'. The 'WEB' tab is selected. The main content area displays 'Also try: [seven samurai movie](#), [seven samurai remake](#), [more...](#)'. Below this is a movie card for 'Seven Samurai (1954)' from 'movies.yahoo.com'. The card includes a movie poster, the running time '2 hr 40 min', the director 'Akira Kurosawa', and the starring cast 'Takashi Shimura, Toshiro Mifune, Yoshio Inaba, Seiji Miyaguchi, ... more'. To the left of the main content is a 'Structured Recommendations' section titled 'RELATED MOVIES' which lists several movies with their posters: 'Dreams', 'Ran', 'Dersu Uzala', 'Throne of Blood', 'Yojimbo', and 'Stray Dog'. Red callout boxes with lines pointing to specific elements are overlaid on the image: 'Entity Detection' points to the search bar; 'Structured Recommendations' points to the 'RELATED MOVIES' section; 'Play trailer' points to the movie poster; and 'Structured Data' points to the movie details box.

Entity Detection

YAHOO!

Seven Samurai

82,800,000 result

WEB IMAGES VIDEO SHOPPING BLOGS MORE

Also try: [seven samurai movie](#), [seven samurai remake](#), [more...](#)

Seven Samurai (1954)
[movies.yahoo.com](#)
Yahoos B+ A Japanese farming village, constantly beseiged and pillaged by an army of bandits, recruits seven independent... [more](#)

Running Time: 2 hr 40 min
Directed by: [Akira Kurosawa](#)
Starring: [Takashi Shimura](#), [Toshiro Mifune](#), [Yoshio Inaba](#), [Seiji Miyaguchi](#), ... [more](#)

Structured Recommendations

RELATED MOVIES

Dreams

Ran

Dersu Uzala

Throne of Blood

Yojimbo

Stray Dog

Play trailer

Structured Data

Stress on Structured Data Presentation (Entities!)

Google

brad pitt movies

manishg.iitb@gmail.com

+ Share


SafeSearch

Search tools


Web Images Maps Videos More Search tools

Brad Pitt movies


Most popular first




World War Z
2013




Fight Club
1999




Killing Them Softly
2012




Troy
2004




Inglourious Basterds
2009




Snatch
2000




Seven
1995




The Curious Case of Benjamin Button
2008



The Tree of Life
2011



Moneyball
2011



Legion
1999

[Brad Pitt - IMDb](http://www.imdb.com/name/nm0000093/)
www.imdb.com/name/nm0000093/

Movies. In Theaters; Top 250; US Box Office; Coming Soon; Trailer Gallery; Watch ...
actor **Brad Pitt's** most widely recognized role may be Tyler Durden in **Fight Club**. ...
Brad Pitt and Michael Fassbender at event of 12 Years a Slave Still of **Brad Pitt** ... Pitt
and Geena Davis in Thelma & Louise Still of **Brad Pitt** in **World War Z** ...
Killing Them Softly - 12 Years a Slave - Fury - The Tree of Life


[Brad Pitt - Wikipedia, the free encyclopedia](http://en.wikipedia.org/wiki/Brad_Pitt)
en.wikipedia.org/wiki/Brad_Pitt

Pitt first gained recognition as a cowboy hitchhiker in the road movie Thelma ... Four
years later, Pitt starred in the cult hit **Fight Club**. ... **William Bradley Pitt** was born in
Shawnee, Oklahoma, and is the son of Jane Etta Pitt at the German premiere of
Inglourious Basterds in July 2009 "Movie review: **Killing Them Softly**".
Brad Pitt filmography - List of awards and ... - List of most expensive - Jill Schoelen

Brad Pitt

Actor

William Bradley "Brad" Pitt is an American actor and film producer. Pitt has received four Academy Award nominations and five Golden Globe Award nominations, winning one Golden Globe. Wikipedia



Born: December 18, 1963 (age 49), Shawnee, Oklahoma, United States

Height: 1.80 m

Spouse: Jennifer Aniston (m. 2000–2005)

Upcoming movies: The Counselor, 12 Years a Slave, The Normal Heart, Voyage of Time, The Tiger, Fury, True Story

<https://www.google.ca/search?safe=off&q=the+tree+of+life+2011&stick=H4sIAA...>

Beyond Entities: Tasks and Task Completion

bing
Web

flights to maui

Web Flights More ▾

[Flights from Seattle, WA to Maui, HI](#)

FROM: Seattle, WA (SEA) - Seatl TO: Kahului, HI (OGG) - Kahul

LEAVE: 03/23/2012 RETURN: 03/25/2012 [Find flights](#)

[bing.com/travel](#)

WEEKEND ESTIMATES

✈	\$639	Depart March 09
✈	\$748	Depart March 16
✈	\$649	Depart March 23

Task completion


Price prediction

bing
Web

royal lahaina resort

Web Hotels Images More ▾

[Royal Lahaina Resort - Bing Travel](#)

 2780 Kekaa Dr · Lahaina
(808) 661-9469
[Details](#) · [Photos](#) · [Amenities](#) · [Directions](#) · [Website](#)
★★★★☆ (12)
Sample rate: \$230

Compare rates from major travel sites

CHECK IN: 03/15/2012 CHECK OUT: 03/17/2012 [Find rate](#)

[bing.com/travel/hotels/search?q=royal+lahaina+resort](#)

Task completion

Aggregate ratings

Tasks: Sequence of Actions on Entities

Web Images Videos Shopping News Maps More | MSN Hotmail

Thomas f Sign out Rewards

bing MS Beta

Web Pandorum

Web Images Videos Shopping Movies More

ENTITY

ALL RESULTS 1-10 of 116,000 results - Advanced

PANDORUM - Now Available on DVD & Blu-ray
PANDORUM Movie Official Site - Now Available on DVD & Blu-ray. Get Trailers and Clips, Synopsis, Games, Downloads, Gallery and more.
www.pandorummovie.com · Mark as spam

Pandorum - Wikipedia, the free encyclopedia
Plot · Cast · Production · Release
Pandorum is a 2009 German-British science fiction thriller film written by Travis Milloy, directed by Christian Alvart and produced by Paul W.S. Anderson.
en.wikipedia.org/wiki/Pandorum · Mark as spam

Pandorum (2009 Film)
Action/Horror/Sci-Fi · R · 108 min
With Dennis Quaid, Ben Foster, Cam Gigandet, Antje Traue. A pair of crew members aboard a spaceship wake up with no knowledge of ...
8729 · Mark as spam

IMDb 6.8/10 40,028 ratings

ACTIONS

Watch Trailer
Watch Film Online
Read User Reviews
Read Critics Reviews
Buy DVD/Blu-ray
Jinni Taste Match

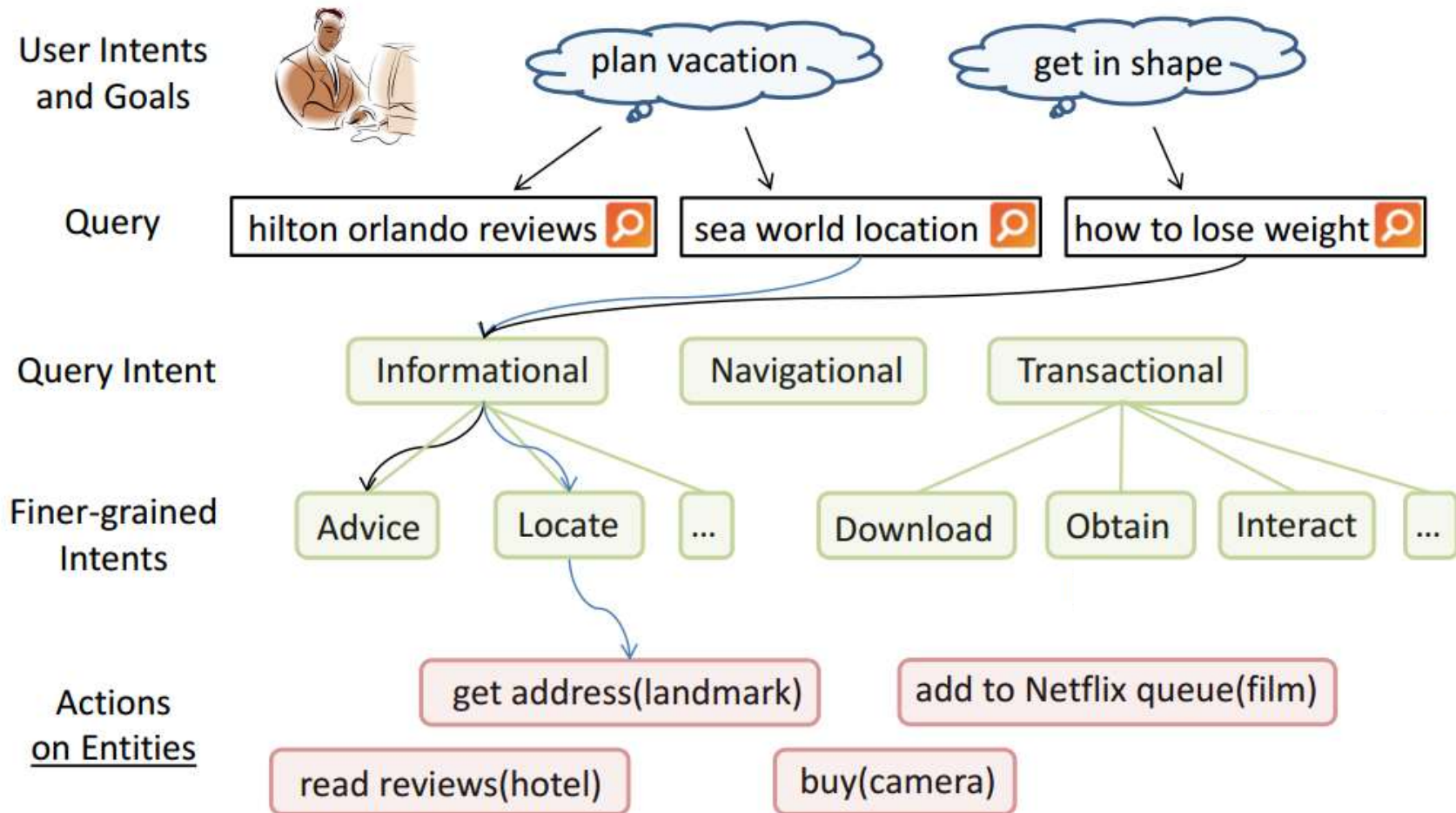
Netflix
Amazon Instant Video
Zune Store
iTunes
more ...

More

Recognize entity in query

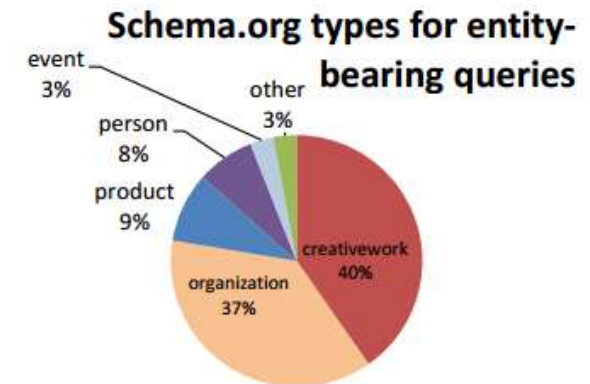
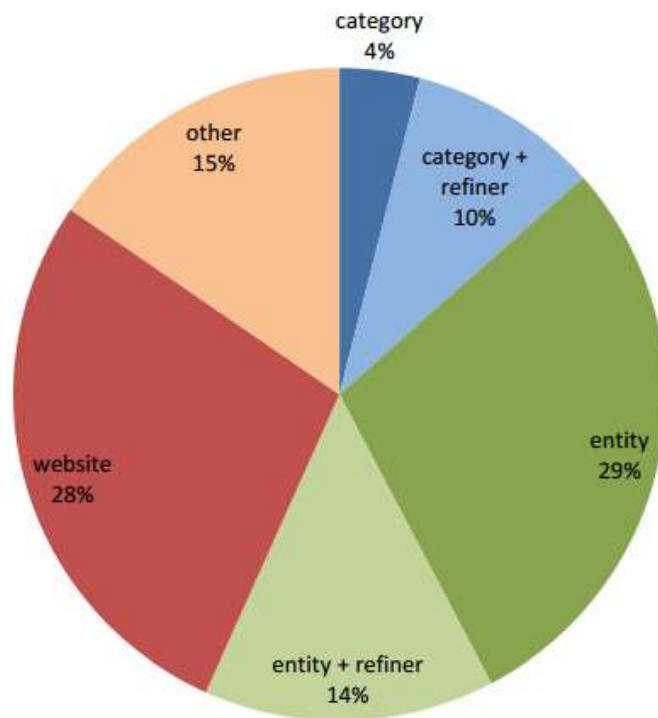
Actions easily accessible

Actions vs Intents



Do Web Queries contain Entities?

Entity Distribution in Web Search Queries



43% entity

(e.g., "GoldenEye", "Horne Auto")

14% entity category

(e.g., "golf cart battery", "global sim card")

15% no entity

(e.g., "xxx", "good reading quotes")

28% website

(e.g., "yahoo mail", "girlybox.com")

* From a query traffic-weighted sample

Actions on Small Dataset

Navigational

10x	Login Action (on a Website entity)
4x	Search Action (on a Website entity)
<u>Informational</u> (need satisfied by reading content, or could be satisfied by written transcript of content)	
1x	Find Location(s) (on an Organization entity)
1x	Find Lyrics (on a CreativeWork / MusicalTrack entity)
2x	Find Recipe For (on a food)
1x	Find Where to Buy (on a Product entity)
2x	Get Contact Information (on an Organization entity)
1x	Get Directions To (on an Organization / Location entity)
2x	Get Domain Information (on a Website entity)
1x	Get Event Details (on an Event entity)
2x	Get Event Results (on an Event entity)
4x	Product Detail (on a Product entity)
29x	Learn (on any entity)
6x	Learn / Educational (on a Person / Product / Organization entity)
1x	Learn / Trivia (on any entity)
1x	Operating Hours (on an Organization entity)
3x	Read Articles (on a News / Magazine entity)
1x	Read Guide (on a Product entity)
1x	Read Help (on a Product entity)
8x	Read News About (on any entity)
3x	Read Reviews (Shopping on a CreativeWork / Product / Service entity)
1x	Read Spoilers (on a CreativeWork)
8x	Research (focused information gathering, on any entity)
8x	Search Database of (e.g., obituaries, on an Organization / Website)
	See Menu (on a Restaurant)
3x	See Pictures (on a Person / Product / Organization entity)
	Side Effects / Safety (on a Product entity)
	Stock Price (on an Organization entity)

Transactional (navigating to a web-mediated action)

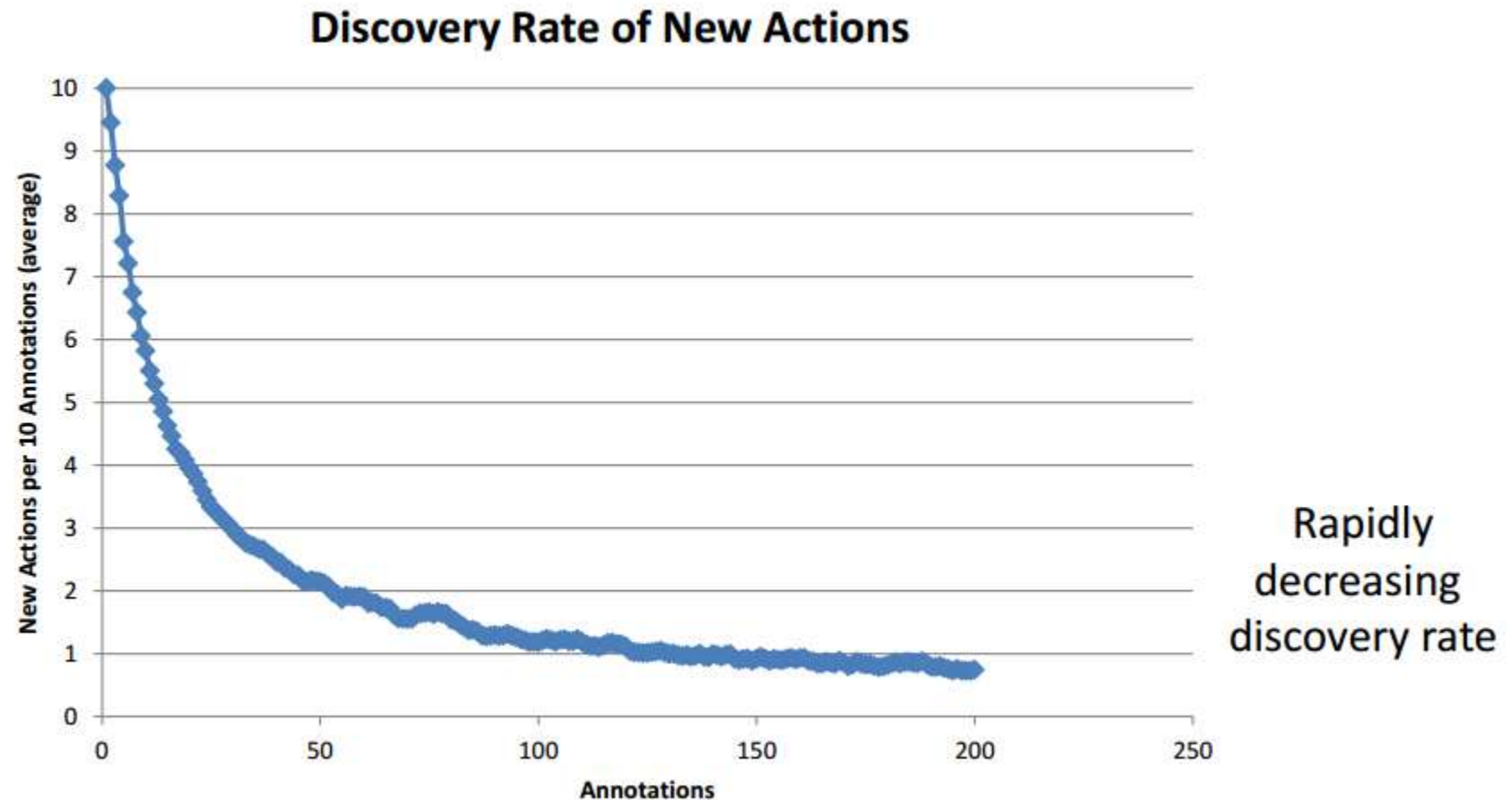
1x	Apply for Job (on a LocalBusiness / Organization entity)
	Buy (Shopping on a Product entity)
	Buy Tickets (on an Event / Product / Person entity)
3x	Content Creation (on a Website entity)
	Discuss Online (on any entity)
5x	Download (on a CreativeWork or Software entity)
1x	Listen to Music (on a CreativeWork or Website entity)
	Manage Account (on a Local Business / Website / Org entity)
	Pay Bill (on a Website / Organization entity)
14x	Play Game (on a Game entity)
	Rent (on a CreativeWork / Product entity)
2x	Reservation (on a Hotel entity)
	Schedule Appointment (on a LocalBusiness entity)
	Sell (Shopping on a Product entity)
1x	Use Service On (e.g., translate, on a Website)
6x	Watch Video About (on any entity)
1x	Web Chat

Other

13x	Shopping (category of actions including reviews and buying)
19x	Various/Unknown

Actions are tied to entity types
47 actions in current list

Is #Actions Limited?

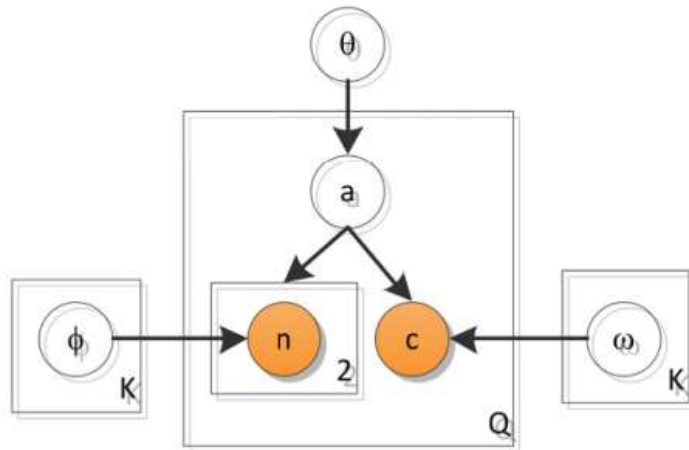


Dataset

- Bing Query Log
- Remove queries without click and without Freebase entities
- 21 Freebase entity types
- Keep only those queries which have context size as 0 or 1
- Remove navigational queries (ones with >98% clicks to same webhost)
- 235K entities, 129K context words, 58K hosts

website	product line	digital camera
consumer product	software	film
comp/video game	person	athlete
politician	actor	artist
employer	business operation	restaurant
location	travel destination	tourist attraction
sports facility	university	road

Generative Models: Model 1 (Context+Click)



Model 1: Generative model of actionable queries.

For each query q

action $a \sim \text{Multinomial}(\theta)$

l-context $n_1 \sim \text{Multinomial}(\phi_a)$

r-context $n_2 \sim \text{Multinomial}(\phi_a)$

click $c \sim \text{Multinomial}(\omega_a)$

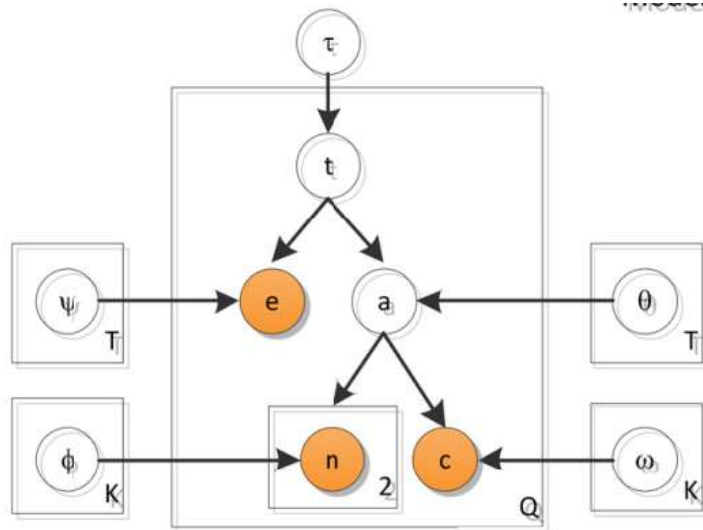
- Use EM to estimate model parameters
 - E step: Compute the posterior (i.e. $P(\text{latent variables} | \text{data}, \text{parameters})$) and hence compute the expected log likelihood function

$$\log P(Q) = \sum_{j=1}^N \sum_a P^j(a | q, c) \log P^j(q, c, a) \quad P(a, q=\{n_1, n_2\}, c | \theta, \Phi, \Omega) = P(a | \theta) P(n_1 | a, \Phi) P(n_2 | a, \Phi) P(c | a, \Omega)$$

- M step: Estimate parameters such that it maximizes the expected log likelihood

$$\theta_{\hat{a}} = \frac{\sum_{j=1}^N P^j(a=\hat{a} | q, c)}{\sum_{j=1}^N \sum_a P^j(a | q, c)} \quad \Phi_{\hat{a}, \hat{n}} = \frac{\sum_{j=1}^N P^j(a=\hat{a} | q, c) [I[n_1^j=\hat{n}] + I[n_2^j=\hat{n}]]}{2 \sum_{j=1}^N P^j(a=\hat{a} | q, c)} \quad \Omega_{\hat{a}, \hat{c}} = \frac{\sum_{j=1}^N P^j(a=\hat{a} | q, c) I[c^j=\hat{c}]}{\sum_{j=1}^N P^j(a=\hat{a} | q, c)}$$

Generative Models: Model 2 (Context+Click+Type+Entity)



Model 2: Generative model of actionable queries.

For each query q

type $t \sim \text{Multinomial}(\tau)$

action $a \sim \text{Multinomial}(\theta_t)$

entity $e \sim \text{Multinomial}(\psi_t)$

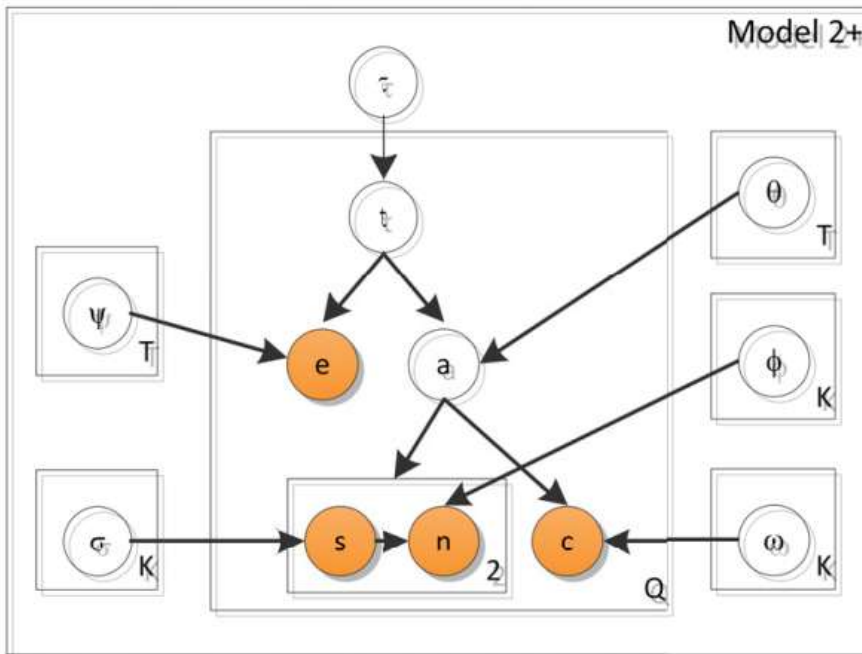
l-context $n_1 \sim \text{Multinomial}(\phi_a)$

r-context $n_2 \sim \text{Multinomial}(\phi_a)$

click $c \sim \text{Multinomial}(\omega_a)$

- Consider a user who is interested in reading a review about the movie “Inception” and who issues the query “inception review” and then clicks on “rottentomatoes.com”
- Model 1: Generate action “read reviews”; given this action choose refiner words ϕ and “review” and then generate a click on “rottentomatoes.com”
- Model 2: Generate type “film”; given the type, generate entity “inception” and then generate action “read reviews”. Given this action choose refiner words ϕ and “review” and then generate a click on “rottentomatoes.com”

Generative Models: Model 2+



Model X + Switch:

For each query q

...

l-context $n_l \sim \text{Multinomial}(\phi_a)$

r-context $n_r \sim \text{Multinomial}(\phi_a)$

switch $s_l \sim \text{Multinomial}(\sigma_a)$

switch $s_r \sim \text{Multinomial}(\sigma_a)$

if (s_l) l-context $n_l \sim \text{Multinomial}(\phi_a)$

if (s_r) r-context $n_r \sim \text{Multinomial}(\phi_a)$

...

Decoding (Handling a new Query)

- New search query: “new york city hotels”
- Run NER to find entity e : “new york city”
- Contexts $n_1 = \phi$ and $n_2 = \text{“hotels”}$ ($s_1 = \text{true}$, $s_2 = \text{false}$)
- Use historical data to compute $P(c|q)$ over all hosts $c \in H$ that received a click for query q
- Actions can be recommended as follows

$$P(a \mid q = \{n_1, e, n_2\}, c, s) = \sum_t \sum_{c \in H} P(a, t \mid q, c, s) P(c \mid q)$$

Web Actions to Web Action Phrases

- Action clusters are clusters of words defined by Φ
- How to translate action to action recommendation phrases?
 - Most probable context words for the action
 - Intersect context words with dictionary of verbs
 - Human involvement



$P(\text{context} | \text{action})$



Web Action words

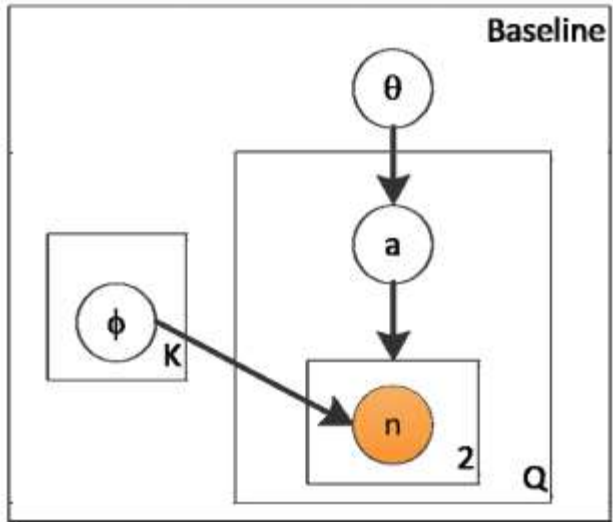


Other words

Download
Find Reviews of
Update
Get help for

Action Phrases

Experiments

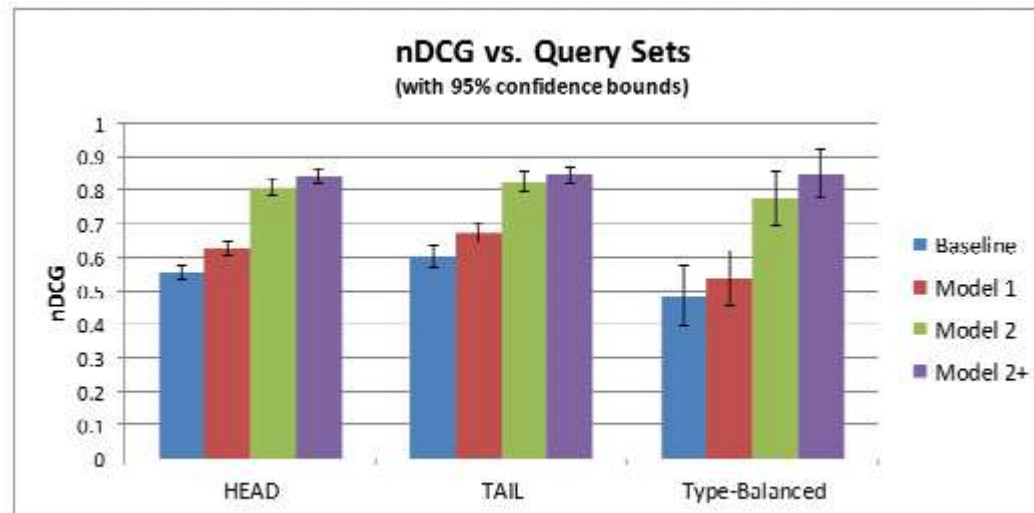


Experimental Configurations

- HEAD: 100 queries from a frequency-weighted random query sample
- TAIL: 100 queries from a uniform random sample
- Type-Balanced: HEAD with diverse entity types

Action Annotations

- Perfect
- Excellent
- Good
- Fair
- Bad



Normalized Discounted Cumulative Gain (nDCG)

Experiments

Table 2. Actions recommended by the various models for the query “Webster University”
Entity: “Webster University” Context: (\emptyset , \emptyset) Types: employer, university, and location

Baseline (context)	Model 1 (+click)	Model 2 (+type, +entity)	Model 2 ⁺ (+switch)
1.Torrent	1.Torrent	1.Read reviews of	1.Find address
2.Read biography	2.Read biography	2.See map of	2.See pictures of
3.Find adult pictures of	3.Read news about	3.Follow sports teams of	3.Find map of
4.Watch videos	4.See pictures of	4.Get weather in	4.Read news about
5.See picture of	5.Apply for jobs at	5.Apply for jobs at	5.Apply for jobs at
6.Get quotes from	6.Get quotes from	6.Find address of	6.See cost of
7.Apply for jobs at	7.See videos with	7.See tuition of	7.See ranking of

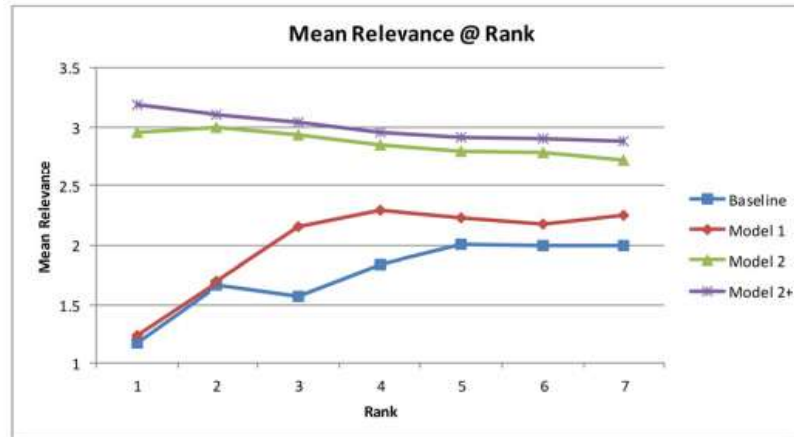


Figure 8. Mean relevance at action rank for our models.

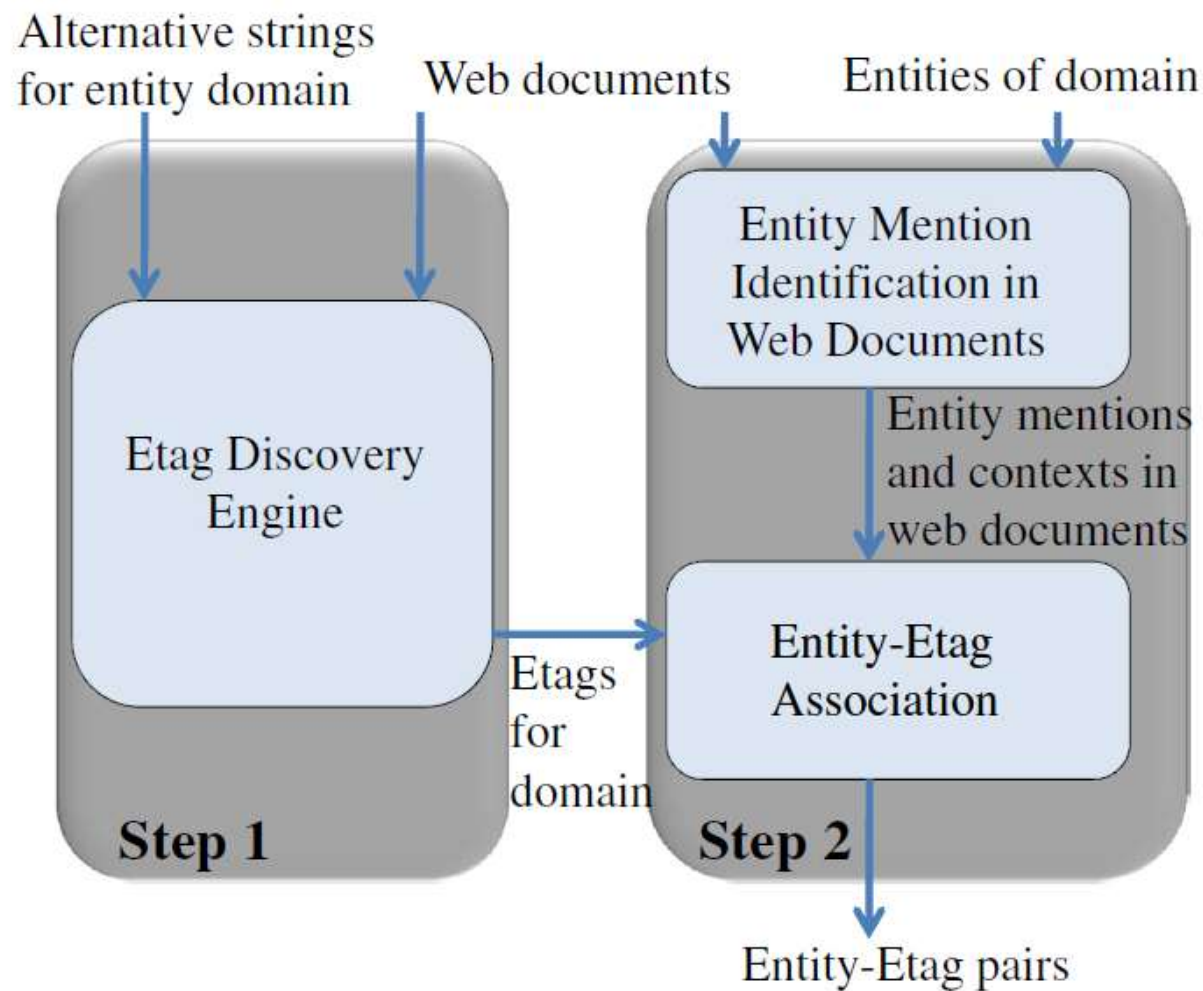
Today's Agenda

- Entity Set Expansion
- Entity Acronym Expansion
- Entity Actions
- **Entity Tagging**

Entity Tagging (Entity-ETag Association)

- Kaushik Chakrabarti, Surajit Chaudhuri, Tao Cheng, and Dong Xin. EntityTagger: automatically tagging entities with descriptive phrases. WWW 2011.
- Aim: Find descriptive tags (etags) for an entity
- Example: Associate etags such as “water resistant”, “rugged” and “outdoor” to entity *Ricoh G600 Digital Camera*
- Two step architecture
 - Leverage precise lexical patterns to find tags for domain
 - Associate tags with entities (only those that show statistically significant co-occurrence, and in proximity)
- Expected Co-occurrence Freq: $E(e, t) = \frac{Freq(e)Freq(t)}{N_W}$
 - $Freq(e)$ and $Freq(t)$ is #documents containing e and t resp.
 - N_W is number of documents in collection W
- Establish co-occurrence of entity with tag using Gtest of goodness of fit [Williams, 1976]
 - Has connections with TF IDF kind of computation
 - Compares expected co-occurrence frequency with observed co-occurrence frequency

E-Tagging System Architecture



Take-away Messages

- Last lecture we discussed about a few entity semantics mining tasks
 - Entity Synonyms
 - Entity Attribute Discovery and Augmentation
 - Entity Linking
- This lecture we covered more of such tasks
 - Entity Set Expansion
 - Entity Acronym Expansion
 - Entity Actions
 - Entity Tagging

Further Reading

- Yeye He, Dong Xin. SEISA: Set Expansion by Iterative Similarity Aggregation. WWW 2011.
- Kaushik Chakrabarti, Surajit Chaudhuri, Tao Cheng, and Dong Xin. 2011. EntityTagger: automatically tagging entities with descriptive phrases. In Proceedings of the 20th international conference companion on World wide web (WWW '11).
- Thomas Lin, Patrick Pantel, Michael Gamon, Anitha Kannan, and Ariel Fuxman. 2012. Active objects: actions for entity-centric search. In Proceedings of the 21st international conference on World Wide Web (WWW '12).
- Bilyana Taneva, Tao Cheng, Kaushik Chakrabarti, and Yeye He. 2013. Mining acronym expansions and their meanings using query click log. In Proceedings of the 22nd international conference on World Wide Web (WWW '13).

Preview of Lecture 21: Introduction to Web Search

Query Log Mining

- Search and browse logs
- Log mining applications
- Four data structures
- Query Statistics
- Query Classification

Disclaimers

- This course represents opinions of the instructor only. It does not reflect views of Microsoft or any other entity (except of authors from whom the slides have been borrowed).
- Algorithms, techniques, features, etc mentioned here might or might not be in use by Microsoft or any other company.
- Lot of material covered in this course is borrowed from slides across many universities and conference tutorials. These are gratefully acknowledged.

Thanks!