Towards the Web of Concepts: Extracting Concepts from Large Datasets

Lisa Miller
ICS 624
3/14/2011

Paper and Slides by:

Aditya G. Parameswaran

Stanford University

Joint work with:

Hector Garcia-Molina (Stanford) and Anand Rajaraman (Kosmix Corp.)

Motivating Examples

Lord of the rings

Lord of the

Of the rings

Microsoft Research Redmond

Microsoft Research

Research Redmond

Computer Networks

Computer

Networks

The Web of Concepts (WoC)

Concepts are:

Entities, events and topics people are searching for

Search:	Japanese restaurants in Palo Alto
Return:	Homma's Sushi

Web of concepts contains:

Concepts

Relationships between concepts

Metadata on concepts

Hours: M-F 9-5
Expensive

How does the WoC help us?

- Improve search
- Find concepts the query relates to
- Return metadata
  - E.g., *Homma's Sushi Hours, Phone No., …*
- Return related concepts
  - E.g., *Fuki Sushi, …*
- Rank content better
- Discover intent

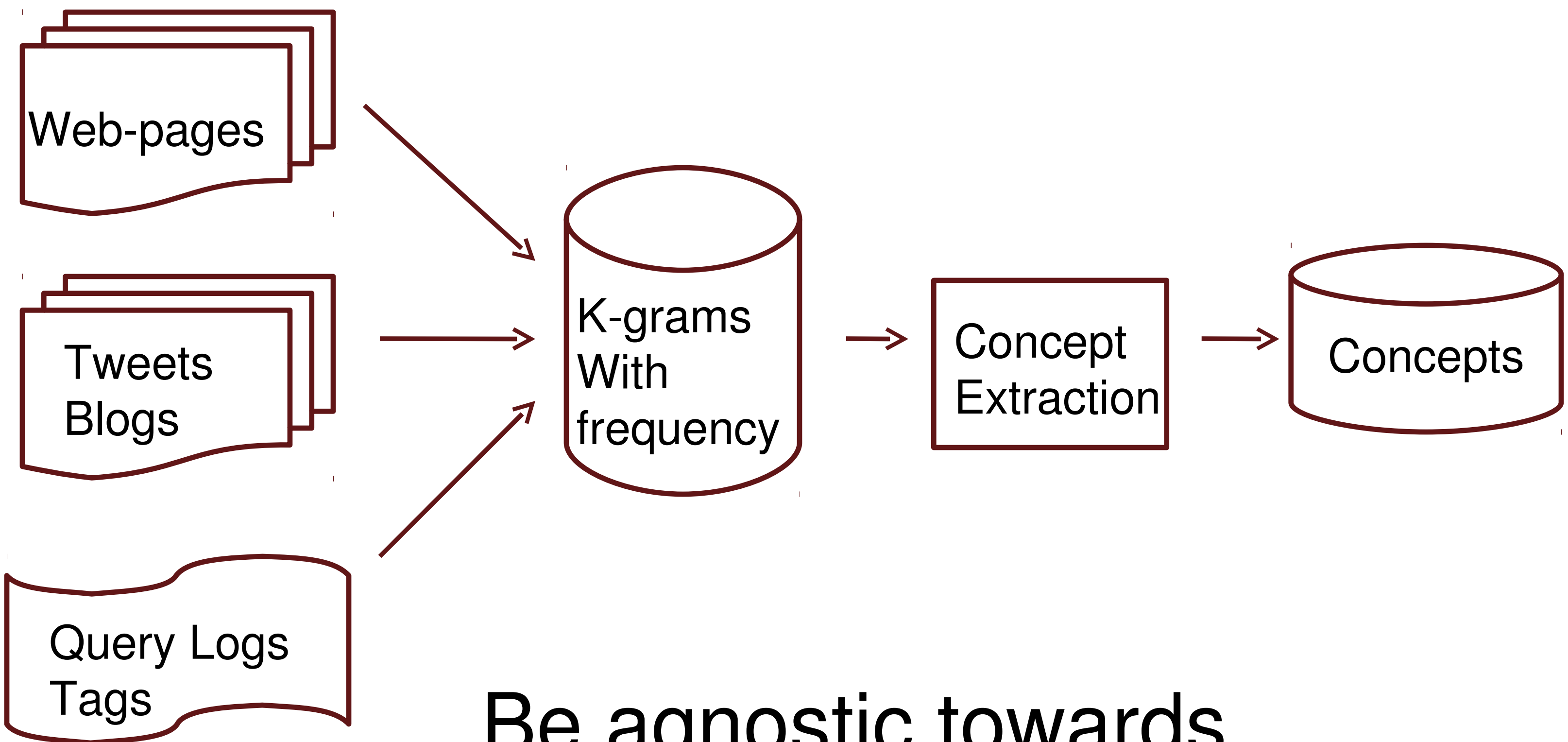
How to construct the WoC?

- Standard sources
  - *Wikipedia, Freebase,* …
- Small fraction of actual concepts
  - Missing: restaurants, hotels, scientific concepts, places, …
- Updating the WoC is *critical*

  Timely results

  New events, establishments, …,
- Old concepts not already known

Desiderata



Be agnostic towards

- Context
- Natural Language

Our Definition of Concepts

Concepts are:

k-grams representing

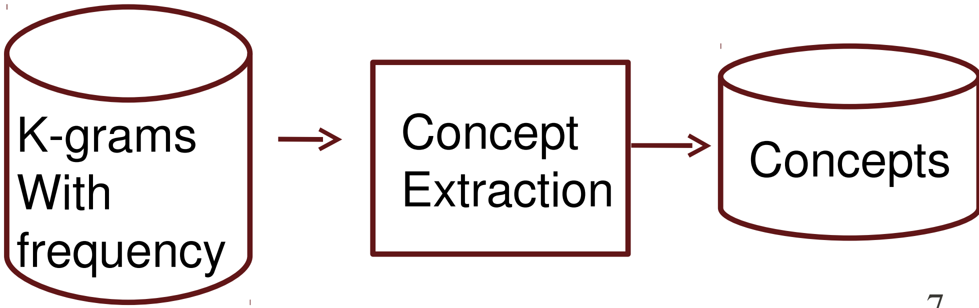
- Real / imaginary entities, events, … that
- People are searching for / interested in

Concise

- E.g., *Harry Potter* over *The Wizard Harry Potter*
- Keeps the WoC small and manageable

Popular

- Precision higher



Previous Work

Frequent Item-set Mining

- Not quite frequent item-sets
  - *k*-gram can be a concept even if *k*-1-gram is not
- Different support thresholds required for each k
- But, can be used as a first step

Term extraction

- IR method of extracting terms to populate indexes
- Typically uses NLP techniques, and not popularity
- One technique that takes popularity into account

Notation

K-gram	Frequency
San	14585
Antonio	285
San Antonio	2385

Sub-concepts of *San Antonio*: “San”, “Antonio”

Sub-concepts of *San Antonio Texas* : “San Antonio” ,“Antonio Texas”

Super-concepts of *San* : “San Antonio”, “San Diego”, etc.

Support (*San Antonio****)*** = 2385

Pre-confidence of *San Antonio*: 2385 / 14585

Post-confidence of *San Antonio*:   2385 / 2855

Empirical Property

Observed on Wikipedia

Lord of the Rings
Manhattan Acting School
Microsoft Research Redmond
Computer Networks

If *k*-gram {*a1 a2 … ak* } for *k* > 2 is a concept, then at least one of the two sub-concepts: *{a1 a2 … ak-1}, {a2 a3 … ak}* is not a concept.

Table 1: Percentage of Wikipedia Title concepts violating/not violating "Claim 1"

k	Both Sub-Concepts "violating Claim 1"	1 or more sub-concepts "non-violating"
2	55.69 %	95.63 %
3	7.77	50.69
4	1.78	29.57
5	0.51	18.44
6	0.31	13.23

“Indicators” that we look for

- Popular
- Scores highly compared to sub- and super-concepts
  - *“Lord of the rings”* better than *“Lord of the”* and *“Of the rings”*.
  - *“Lord of the rings”* better than *“Lord of the rings soundtrack”*
- Does not represent part of a sentence
  - *i.e. “Barack Obama Said Yesterday”*
  - “Not required for tags, query logs” ?

Outline of Approach

S = {}

For k = 1 to n

- Evaluate all k-grams w.r.t. k-1-grams
  - Add some k-grams to S
  - Discard some k-1-grams from S
- Precisely k-grams until $k = n-1$ that satisfy indicators are extracted
  - Under perfect evaluation of concepts w.r.t. sub-concepts
  - Proof in Paper

Detailed Algorithm

$S = \{\}$

For $k = 1$ to n

- For all k-grams s (two sub-concepts r and t)
  - If $support(s) < support\text{-}threshold(k)$
    - Continue
  - If $min\ (pre\text{-}conf(s), post\text{-}conf(s)) >$ threshold
    - $S = S$ □ $\{s\} - \{r, t\}$
  - Elseif $pre\text{-}conf(s) >$ threshold & >> $post\text{-}conf(s)$ & $t \in$ S
    - $S = S$ □ $\{s\} - \{r\}$
  - Elseif $post\text{-}conf(s) >$ threshold & >> $pre\text{-}conf(s)$ & $r \in$ S
    - $S = S$ □ $\{s\} - \{t\}$

Indicator 1

Indicator 2:
r & t are not concepts

Indicator 2:
r is not a concept

Indicator 2:
t is not a concept

Experiments: Methodology

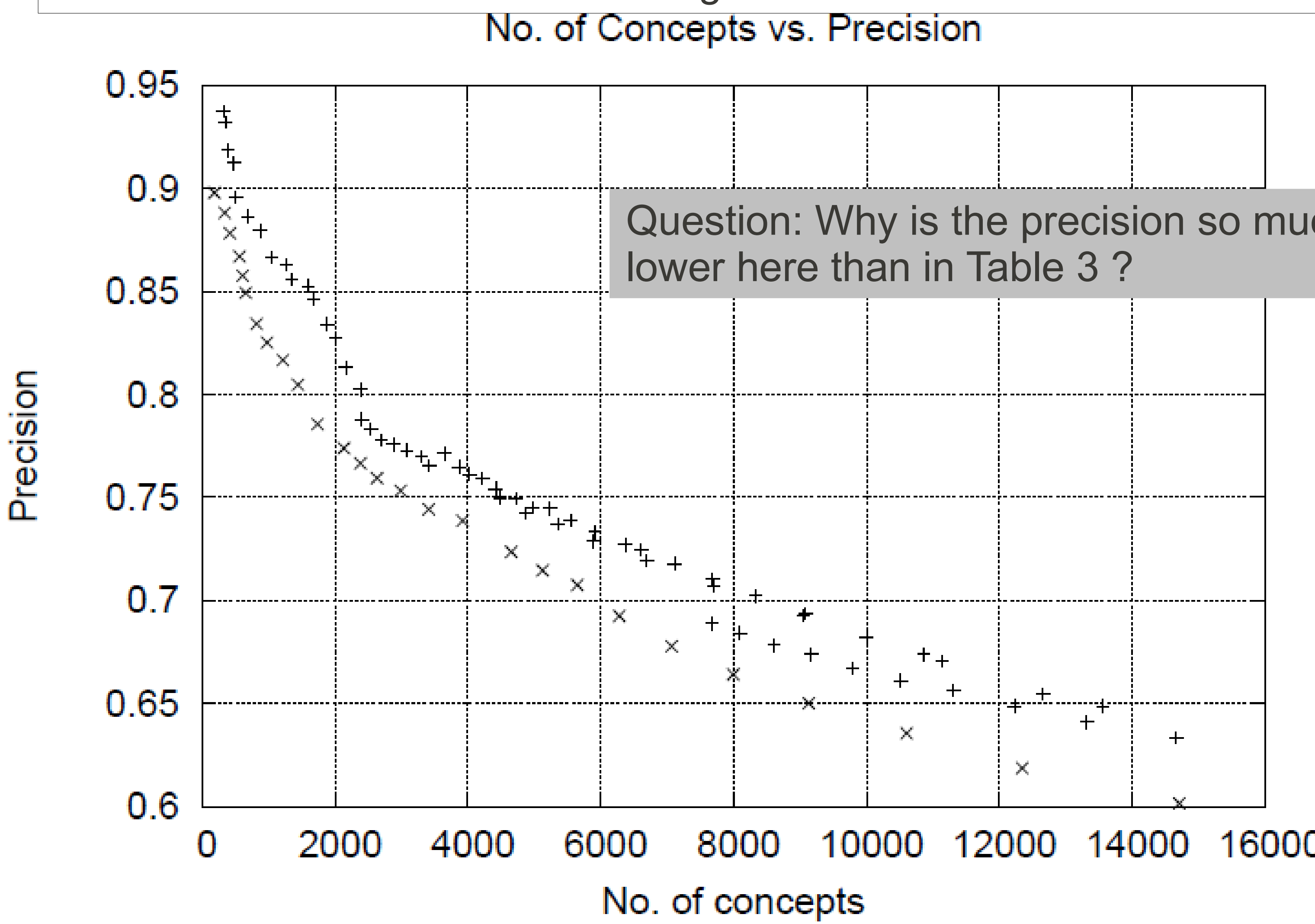
- AOL Query Log Dataset
  - 36M queries and 1.5M unique terms.
  - Evaluation using Humans (Via M.Turk)
  - Plus Wikipedia
    - (For experiments on varying parameters)
  - Experimentally set thresholds
- Compared against
  - C-Value Algorithm:
    - a term-extraction algorithm with popularity built in
  - Naïve Algorithm:
    - simply based on frequency

Raw Numbers

- 25882 concepts extracted
- Absolute precision of 0.95 rated against Wikipedia and Mechanical Turk.
- For same volume of 2, 3, and 4-gram concepts, our algorithm gave
  - Fewer absolute errors (369) vs. C-Value (557) and Naïve (997)
  - Greater Non-Wiki Precision (0.84) vs. C-Value (0.75) and Naïve (0.66)

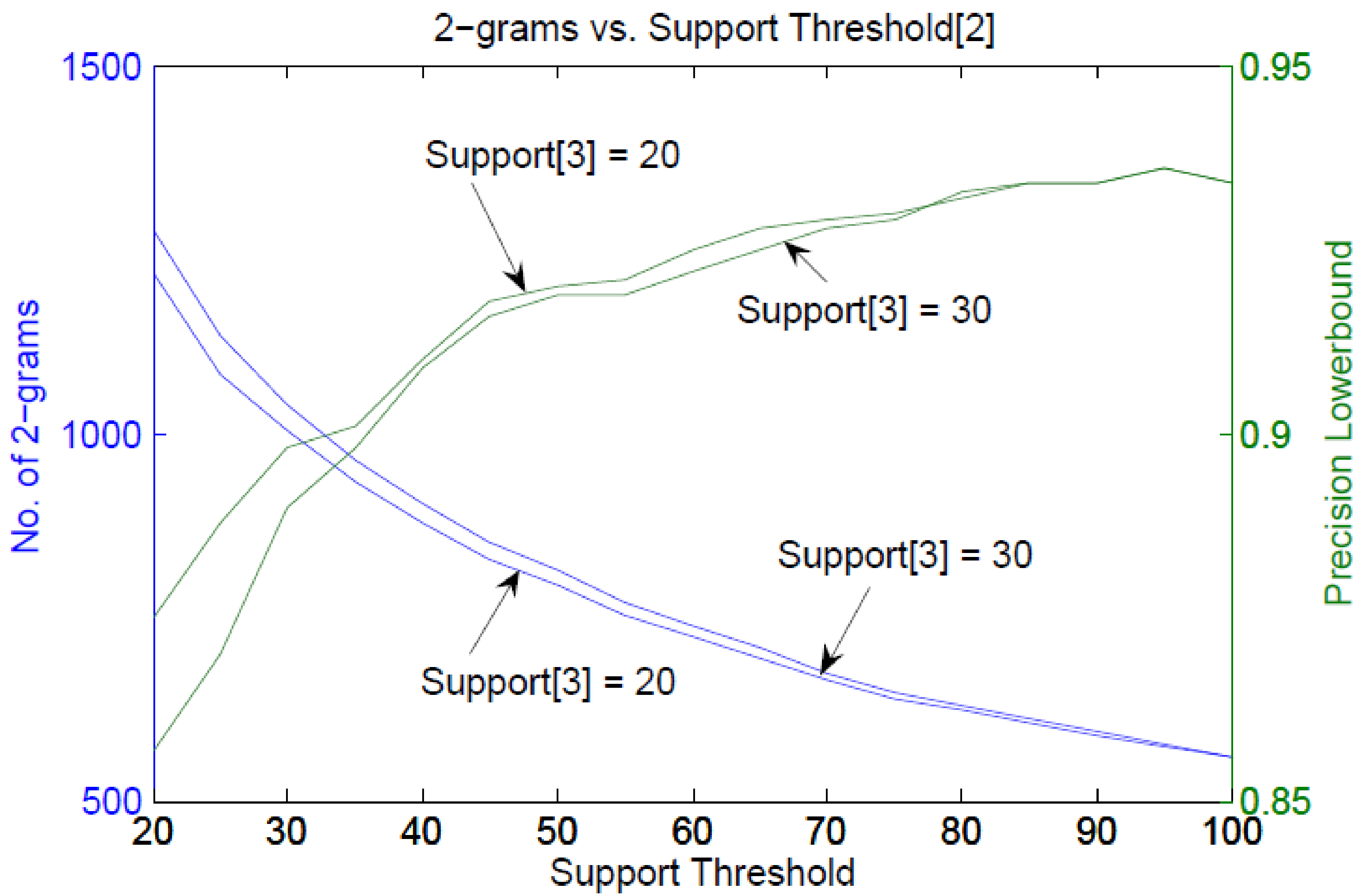
Head-to-head Comparison

Figure 2: Variation of precision vs. volume of concepts extracted
X - C-Value
+ - our algorithm

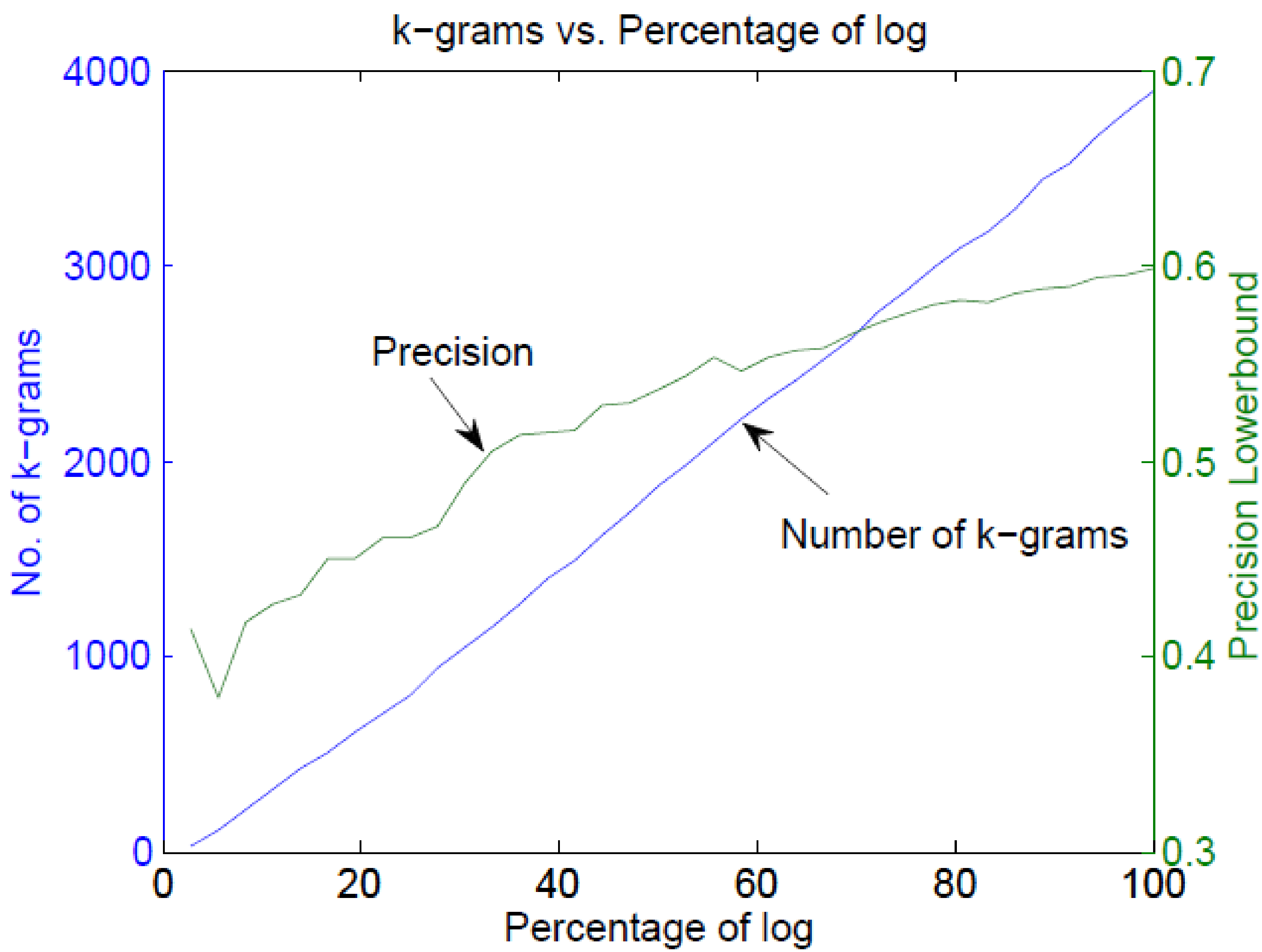


Question: Why is the precision so much lower here than in Table 3 ?

Experiments on varying thresholds



On Varying Size of Log



Ongoing Work
(with A. Das Sarma, H. G.-Molina, N. Polyzotis and J. Widom)

How do we attach a new concept *c* to the web of concepts?

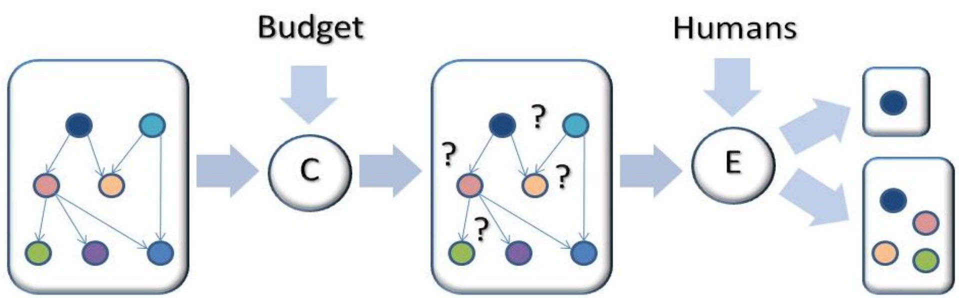
Via human input

But: costly, so need to minimize # questions

Questions of the form: Is *c* a kind of X?

Equivalent to Human-Assisted Graph Search

Algorithms/Complexity results in T.R.



Questions

- What did they really accomplish?
  - Only worked for log of queries, already concepts in general
- What about ordering of words?
  - San Antonio Japanese restaurant vs. Japanese restaurant San Antonio