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**

<table>
<thead>
<tr>
<th>K-gram</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>San</td>
<td>14585</td>
</tr>
<tr>
<td>Antonio</td>
<td>285</td>
</tr>
<tr>
<td>San Antonio</td>
<td>2385</td>
</tr>
</tbody>
</table>

**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*: \( \frac{2385}{14585} \)

**Post-confidence** of *San Antonio*: \( \frac{2385}{2855} \)
If $k$-gram \{a_1 \ a_2 \ldots \ a_k\} for $k > 2$ is a concept, then at least one of the two sub-concepts: \{a_1 \ a_2 \ldots \ a_{k-1}\}, \{a_2 \ a_3 \ldots \ a_k\} is not a concept.

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

<table>
<thead>
<tr>
<th>$k$</th>
<th>Both Sub-Concepts “violating Claim 1”</th>
<th>1 or more sub-concepts “non-violating”</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>55.69 %</td>
<td>95.63 %</td>
</tr>
<tr>
<td>3</td>
<td>7.77</td>
<td>50.69</td>
</tr>
<tr>
<td>4</td>
<td>1.78</td>
<td>29.57</td>
</tr>
<tr>
<td>5</td>
<td>0.51</td>
<td>18.44</td>
</tr>
<tr>
<td>6</td>
<td>0.31</td>
<td>13.23</td>
</tr>
</tbody>
</table>
“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 \( \text{support}(s) < \text{support-threshold}(k) \)
    - Continue
  - If \( \min(\text{pre-conf}(s), \text{post-conf}(s)) > \text{threshold} \)
    - \( S = S \cup \{s\} - \{r, t\} \)
  - Elseif \( \text{pre-conf}(s) > \text{threshold} \) and \( t \in S \)
    - \( S = S \cup \{s\} - \{r\} \)
  - Elseif \( \text{post-conf}(s) > \text{threshold} \) and \( r \in S \)
    - \( S = S \cup \{s\} - \{t\} \)

Indicator 1

Indicator 2:
- \( r \) & \( t \) are not concepts
- \( r \) is not a concept
- \( t \) 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)
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

2-grams vs. Support Threshold

Support[3] = 20

Support[3] = 30

Support[3] = 20

No. of 2-grams

Precision Lowerbound

Support Threshold
On Varying Size of Log

$k$-grams vs. Percentage of log

No. of $k$-grams

Percentage of log

Precision

Number of $k$-grams
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