Tweets Conceptualization and Classification

WANG Yue ywangby@ust.hk

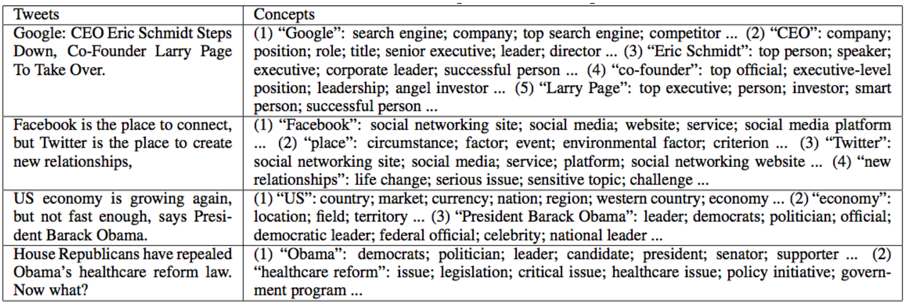
Department of Computer Science
Hong Kong University of Science and Technology

January 12, 2014

Problem Defination

Given a short piece of message such tweets, how to capture the semantic information of it, such as topic and subject information?

Twitter Conceptualization Examples



Tweets	Concepts
Google: CEO Eric Schmidt Steps Down, Co-Founder Larry Page To Take Over.	(1) “Google”: search engine; company; top search engine; competitor ... (2) “CEO”: company; position; role; title; senior executive; leader; director ... (3) “Eric Schmidt”: top person; speaker; executive; corporate leader; successful person ... (4) “co-founder”: top official; executive-level position; leadership; angel investor ... (5) “Larry Page”: top executive; person; investor; smart person; successful person ...
Facebook is the place to connect, but Twitter is the place to create new relationships,	(1) “Facebook”: social networking site; social media; website; service; social media platform ... (2) “place”: circumstance; factor; event; environmental factor; criterion ... (3) “Twitter”: social networking site; social media; service; platform; social networking website ... (4) “new relationships”: life change; serious issue; sensitive topic; challenge ...
US economy is growing again, but not fast enough, says President Barack Obama.	(1) “US”: country; market; currency; nation; region; western country; economy ... (2) “economy”: location; field; territory ... (3) “President Barack Obama”: leader; democrats; politician; official; democratic leader; federal official; celebrity; national leader ...
House Republicans have repealed Obama’s healthcare reform law. Now what?	(1) “Obama”: democrats; politician; leader; candidate; president; senator; supporter ... (2) “healthcare reform”: issue; legislation; critical issue; healthcare issue; policy initiative; government program ...

Figure : Tweets Conceptualization Example

Intuitive Steps in Tweets Clustering

For each tweet:

1. detect entities included in **knowledgebase**(longest entity)
2. get the concepts of the entities in knowledgebase
3. user the concept features to cluster the tweets

Description of Knowledgebase

The taxonomy is a **directed acyclic graph**.

- contains **isA** relationship
- *concept* → *instance*, *attribute*(extract syntactic hearst patterns from webpages)
- claims in knowledgebase is associated with scores such as *plausibility*, *typicality*.
- each concept may have multiple senses(meanings)
- hierarchical structure

Derive Concepts Based on Knowledgebase

Given a set of terms, how to derive their concepts.
Notation:

- candidate concepts: $C = \{c_k, k \in 1, \cdots, K\}$
- terms are instances: $E = \{e_i, i \in 1, \cdots, M\}$
- terms are attributes: $A = \{a_j, j \in 1, \cdots, N\}$
- terms are unknown types: $T = \{t_l, 1, \cdots, L\}$

Derive Concepts: Naive Bayes Model

When terms is a set of instances:

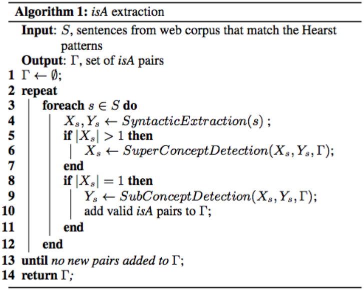
$$P(c_k|E) = \frac{P(E|c_k)P(c_k)}{P(E)} \propto P(c_k)\prod_{i=1}^{M} P(e_i|c_k)$$

Where:

$$P(e_i|c_k) = \frac{P(e_i, c_k)}{Pc_k}$$

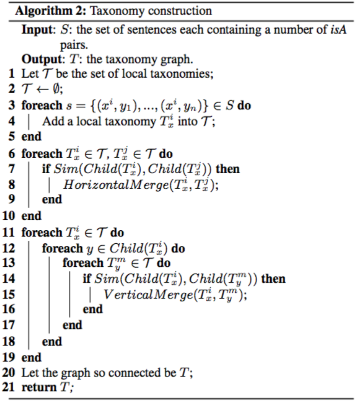
The concept with the largest posterior probability is ranked as the most possible concept to describe the observed instances. The process it similar to attrs and unknown terms.

How to Construct Probabilistic Taxonomy and Knowledgebase



```
Algorithm 1: isA extraction
  Input: $S$, sentences from web corpus that match the Hearst
         patterns
  Output: $\Gamma$, set of isA pairs
  $\Gamma \leftarrow \emptyset$;
  repeat
      foreach $s \in S$ do
          $X_s, Y_s \leftarrow SyntacticExtraction(s)$ ;
          if $|X_s| > 1$ then
              $X_s \leftarrow SuperConceptDetection(X_s, Y_s, \Gamma)$;
          end
          if $|X_s| = 1$ then
              $Y_s \leftarrow SubConceptDetection(X_s, Y_s, \Gamma)$;
              add valid isA pairs to $\Gamma$;
          end
      end
  until no new pairs added to $\Gamma$;
  return $\Gamma$;
```

Figure : step 1: extract isA pairs from sentences



```
Algorithm 2: Taxonomy construction
  Input: $S$: the set of sentences each containing a number of isA
         pairs.
  Output: $T$: the taxonomy graph.
  Let $\mathcal{T}$ be the set of local taxonomies;
  $\mathcal{T} \leftarrow \emptyset$;
  foreach $s = \{(x^i, y_1), ..., (x^i, y_n)\} \in S$ do
      Add a local taxonomy $T_x^i$ into $\mathcal{T}$;
  end
  foreach $T_x^i \in \mathcal{T}, T_x^j \in \mathcal{T}$ do
      if $Sim(Child(T_x^i), Child(T_x^j))$ then
          $HorizontalMerge(T_x^i, T_x^j)$;
      end
  end
  foreach $T_x^i \in \mathcal{T}$ do
      foreach $y \in Child(T_x^i)$ do
          foreach $T_y^m \in \mathcal{T}$ do
              if $Sim(Child(T_x^i), Child(T_y^m))$ then
                  $VerticalMerge(T_x^i, T_y^m)$;
              end
          end
      end
  end
  Let the graph so connected be $T$;
  return $T$;
```

Figure : step 2: construct the taxonomy as DAG