Querying Term Associations and their Temporal Evolution in Social Data

Vassilis Plachouras   Yannis Stavrakas

IMIS / "ATHENA" R.C.
Greece

August 31, 2012

ΙΠΣΥ IMIS
Ε.Κ. Αθηνά
Athena R.C.

Motivation

- Many applications use data from OSNs or microblogging services
  - Data collected by searching for terms related to the application domain
- Selection of terms can have significant impact on results
- Important to be able to explore the context and associations of terms

Objective

- Aim to develop a platform that enables definition of data analysis campaigns from OSNs
- Example: a journalist explores Twitter data can issue the following query concerning the financial crisis:

  For the period during which there is a strong association between hashtags #crisis and #protest, which other hashtags are associated to both #crisis and #protest? Which are the relevant tweets?

Preliminaries

- Model applies to any temporally evolving collection of documents
    - We focus on tweets
- Downloaded tweets are processed at regular time instances $t = 1, 2, \ldots, i$
- At time instance $t = i$, we process tweets downloaded between $i - 1$ and i
    - load tweets in relation TT with attributes tweet id, publication time and term
    - build model for tweets published between $i - 1$ and i

Model definition

Model $\mathcal{M}$ is a set of quintuples

$$\mathcal{M} = \{\langle n, c, w, T, g\rangle\}$$

where

- n and c are *target* and *context* nodes, respectively, corresponding to terms
- T is the set of time instances for which the tuple is valid
- g is the time granularity
- $w = P_T(n \to c) = \frac{\sum_{n,c} \frac{1}{|tw|-1}}{\sum_{n \in tw} 1}$ or
  $w = P_T(n \to n) = \frac{\sum_{n \in tw, |tw|=1} 1}{\sum_{n \in tw} 1}$

Example of Model

Build model $\mathcal{M}$ for the tweets tw_i in two time instances

$$t = 1 : tw_1 = \{a\}, tw_2 = \{a\}, tw_3 = \{a, b\}, tw_4 = \{c\}, tw_5 = \{a, c\}$$

$$t = 2 : tw_6 = \{a\}, tw_7 = \{a, c\}$$

Example of Model

Build model $\mathcal{M}$ for the tweets tw_i in two time instances

$$t = 1 : tw_1 = \{a\}, tw_2 = \{a\}, tw_3 = \{a, b\}, tw_4 = \{c\}, tw_5 = \{a, c\}$$
$$t = 2 : tw_6 = \{a\}, tw_7 = \{a, c\}$$

- For tuple $\langle a, b, w, \{1\}, 1\rangle \in \mathcal{M},\ w = 1/4 = 0.25$

Example of Model

Build model $\mathcal{M}$ for the tweets tw_i in two time instances

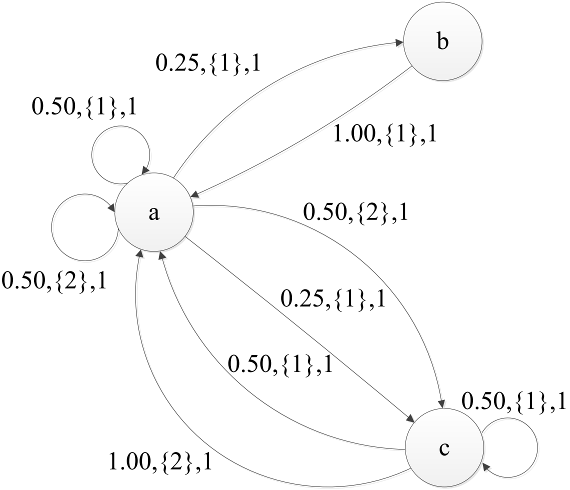
$$t = 1 : tw_1 = \{a\}, tw_2 = \{a\}, tw_3 = \{a, b\}, tw_4 = \{c\}, tw_5 = \{a, c\}$$
$$t = 2 : tw_6 = \{a\}, tw_7 = \{a, c\}$$

- For tuple $\langle a, b, w, \{1\}, 1\rangle \in \mathcal{M}$, $w = 1/4 = 0.25$

The model $\mathcal{M}$ is

$$\begin{aligned}\mathcal{M} = \{&\langle a, b, 0.25, \{1\}, 1\rangle, \langle a, c, 0.25, \{1\}, 1\rangle, \langle b, a, 1.00, \{1\}, 1\rangle,\\ &\langle c, a, 0.50, \{1\}, 1\rangle, \langle a, a, 0.50, \{1\}, 1\rangle, \langle c, c, 0.50, \{1\}, 1\rangle,\\ &\langle a, c, 0.50, \{2\}, 1\rangle, \langle c, a, 1.00, \{2\}, 1\rangle, \langle a, a, 0.50, \{2\}, 1\rangle\}\end{aligned}$$

Model as a graph



Ε.Κ. Αθηνά
Athena R.C.

Query operators

Manipulating the quintuples of models with operators

- *filter*
- *fold*
- *jump*
- *merge*
- *join*

Filter operator

Notation

$\textit{filter}(\mathcal{M}, \textit{cond})$

Input

- Model $\mathcal{M}$
- Condition *cond*

Returns

Set of quintuples in $\mathcal{M}$ that satisfy *cond*

Example

$$\mathcal{M}_2 = \textit{filter}(\mathcal{M}_1, T \text{ inside } \{5 \ldots 12\} \wedge w \in \textit{top}(10))$$

Fold operator

Notation

$fold(\mathcal{M}, g)$

Input

- Model $\mathcal{M}$
- integer $g = g_o/g_i$ where g_o and g_i are the time granularities of the output and input models respectively

Returns

Set of folded quintuples with time granularity $g \times g_i$

Fold operator

Example

For the input model $\mathcal{M}_1$

$$\mathcal{M}_1 = \{\langle n_1, c_1, w_1, \{1\}, 1\rangle, \langle n_1, c_1, w_2, \{2\}, 1\rangle, \langle n_1, c_1, w_3, \{3\}, 1\rangle, \langle n_2, c_1, w_4, \{1\}, 1\rangle, \langle n_2, c_1, w_5, \{4\}, 1\rangle\}$$

the operation $\mathcal{M}_2 = fold(\mathcal{M}_1, 3)$ returns

$$\mathcal{M}_2 = \{\langle n_1, c_1, w_6, \{1,2,3\}, 3\rangle, \langle n_2, c_1, w_4, \{1,2,3\}, 3\rangle, \langle n_2, c_1, w_5, \{4,5,6\}, 3\rangle\}$$

where $w_6 = P_{\{1,2,3\}}(n_1 \to c_1)$

Jump operator

Notation

$jump(\mathcal{M}, k)$

Input

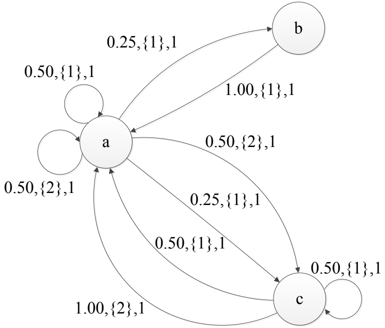
- Model $\mathcal{M}$
- integer k

Output

A model with expanded contexts and weights equal to the probability of a path of length k between two nodes

Jump operator

Example

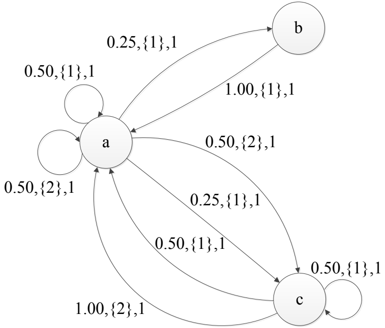


For $t = 1$ the transition matrix

$$P_{\{1\}} = \begin{pmatrix} 0.50 & 0.25 & 0.25 \\ 1.00 & 0.00 & 0.00 \\ 0.50 & 0.00 & 0.50 \end{pmatrix}$$

Jump operator

Example



For $t = 1$ the transition matrix

$$P_{\{1\}} = \begin{pmatrix} 0.50 & 0.25 & 0.25 \\ 1.00 & 0.00 & 0.00 \\ 0.50 & 0.00 & 0.50 \end{pmatrix}$$

For $\mathcal{M}' = jump(\mathcal{M}, 2)$ the weight w of tuple $\langle a, a, w, \{1\}, 1\rangle \in \mathcal{M}'$ is $w = p^2_{\{1\}}(1, 1)$

Merge operator

Notation

$merge(\mathcal{M})$

Input

- Model $\mathcal{M}$

Output

A model where all tuples with the same n and c are aggregated

Merge operator

Example

If the input model is

$$\mathcal{M}_1 = \{\langle n_1, c_1, w_1, T_1, g\rangle, \langle n_2, c_1, w_2, T_1, g\rangle, \langle n_1, c_1, w_3, T_2, g\rangle\}$$

then the output model $\mathcal{M}_2 = merge(\mathcal{M}_1)$ is

$$\mathcal{M}_2 = \{\langle n_1, c_1, w_4, T_1 \cup T_2, g\rangle, \langle n_2, c_1, w_2, T_1, g\rangle\}$$

Join operator

Notation

$join(\mathcal{M}_1, \mathcal{M}_2, cond)$

Input

- Models $\mathcal{M}_1$ and $\mathcal{M}_2$
- Condition *cond*

Output

A subset of $\mathcal{M}_1$ which satisfies condition *cond* on variables of $\mathcal{M}_1$ and $\mathcal{M}_2$

Join operator

Example

Given $\mathcal{M}_1$

$$\mathcal{M}_1 = \{\langle n_1, c_1, 0.5, \{1,2\}, 1\rangle, \langle n_1, c_2, 0.5, \{1,2\}, 1\rangle,$$
$$\langle n_1, c_1, 0.7, \{3,4\}, 1\rangle, \langle n_1, c_2, 0.3, \{3,4\}, 1\rangle\}$$

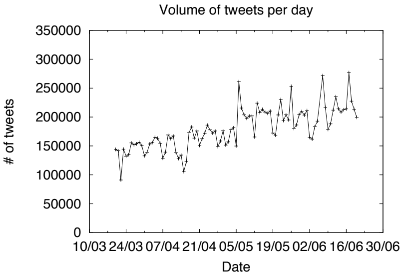
a query, which asks for the tuples with increasing weight over time

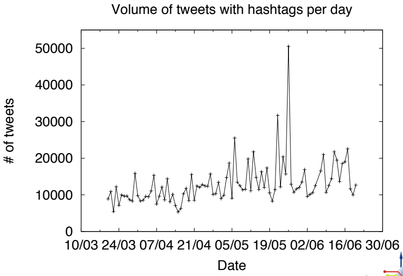
$$join(\mathcal{M}_1 \text{ as } m, \mathcal{M}_1 \text{ as } m', m.n = m'.n \wedge m.c = m'.c$$
$$\wedge\, min(m.T) > max(m'.T) \wedge m.w > m'.w)$$

returns $\mathcal{M}_2 = \{\langle n_1, c_1, 0.7, \{3,4\}, 1\rangle\}$

Dataset

- Set of 16.5 million tweets
  - tracking a set of 74 Greek stop-words
  - collected between March 20 and June 20, 2012
  - processed every 4 hours
- Two most frequent hashtags are #ff and #elections12





Example query

Query

Find the hashtags that are associated with #ekloges12 and for which the association weight increases for two consecutive weeks.

Example query

Query expressed with operators

$$
\begin{aligned}
\mathcal{M}_2 =& \mathit{filter}(\mathcal{M}_1, n = \#\text{ekloges12}) \\
\mathcal{M}_3 =& \mathit{fold}(\mathcal{M}_2, 42) \\
\mathcal{M}_4 =& \mathit{join}(\mathcal{M}_3 \text{ as } m, \mathcal{M}_3 \text{ as } m', \mathit{cond}) \\
\mathcal{M}_5 =& \mathit{join}(\mathcal{M}_4 \text{ as } m, \mathcal{M}_4 \text{ as } m', \mathit{cond})
\end{aligned}
$$

where

$$
\begin{aligned}
\mathit{cond} = m.n <> m.c \wedge m.n = m'.n \wedge m.c = m'.c \\
\wedge m.w > m'.w \wedge \mathit{min}(m.T) = \mathit{max}(m'.T) + 1
\end{aligned}
$$

Query processing and intermediate results

Intermediate results for n=#ekloges12 and c=#eklogesgr

Quintuple	Models
$\langle$#ekloges12, #eklogesgr, 0.0048, {169 ... 210}, 42$\rangle$	$\mathcal{M}_3$
$\langle$#ekloges12, #eklogesgr, 0.0015, {211 ... 252}, 42$\rangle$	$\mathcal{M}_3$
$\langle$#ekloges12, #eklogesgr, 0.0031, {253 ... 294}, 42$\rangle$	$\mathcal{M}_3, \mathcal{M}_4$
$\langle$#ekloges12, #eklogesgr, 0.0004, {295 ... 336}, 42$\rangle$	$\mathcal{M}_3$
$\langle$#ekloges12, #eklogesgr, 0.0036, {337 ... 378}, 42$\rangle$	$\mathcal{M}_3, \mathcal{M}_4$
$\langle$#ekloges12, #eklogesgr, 0.0136, {379 ... 420}, 42$\rangle$	$\mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5$
$\langle$#ekloges12, #eklogesgr, 0.0011, {421 ... 462}, 42$\rangle$	$\mathcal{M}_3$
$\langle$#ekloges12, #eklogesgr, 0.0032, {463 ... 504}, 42$\rangle$	$\mathcal{M}_3, \mathcal{M}_4$
$\langle$#ekloges12, #eklogesgr, 0.0030, {505 ... 546}, 42$\rangle$	$\mathcal{M}_3$
$\langle$#ekloges12, #eklogesgr, 0.0010, {547 ... 588}, 42$\rangle$	$\mathcal{M}_3$

Tuples with highest weight for example query

$\langle$#ekloges12, #pasok,	0.08794, $\{421 \ldots 462\}$, 42$\rangle$
$\langle$#ekloges12, #samaras,	0.06469, $\{505 \ldots 546\}$, 42$\rangle$
$\langle$#ekloges12, #syriza,	0.04663, $\{463 \ldots 504\}$, 42$\rangle$
$\langle$#ekloges12, #ekloges2012,	0.04537, $\{253 \ldots 294\}$, 42$\rangle$
$\langle$#ekloges12, #2012ek,	0.02956, $\{463 \ldots 504\}$, 42$\rangle$
$\langle$#ekloges12, #cpel2012,	0.02859, $\{379 \ldots 420\}$, 42$\rangle$
$\langle$#ekloges12, #ekloges2012,	0.02780, $\{421 \ldots 462\}$, 42$\rangle$
$\langle$#ekloges12, #cpel2012,	0.02140, $\{337 \ldots 378\}$, 42$\rangle$
$\langle$#ekloges12, #mega,	0.01724, $\{463 \ldots 504\}$, 42$\rangle$
$\langle$#ekloges12, #eklogesgr,	0.01361, $\{379 \ldots 420\}$, 42$\rangle$

Concluding remarks

Introduced model and query operators for exploring term associations in social data

- with varying time granularities, forming complex queries

Next steps include

- Handling temporal properties of nodes
- Experimenting with alternative definitions of associations
- Providing user-defined weighting functions
- Experimenting with larger datasets

Thank you!