Enhance micro-blogging recommendations of posts with an homophily-based graph

Quentin Grossetti

Supervised by Cédric du Mouza, Camelia Constantin and Nicolas Travers

1LIP6 - Université Pierre Marie Curie - Paris, France
2CEDRIC Laboratory - CNAM - Paris, France

BDA - Novembre 2017
Introduction

Context

- Growth of microblogging platforms since 2000

- 700 millions of messages/day in 2017
- 300 millions of messages/day in 2017
- 70 millions of publications/day in 2017
- 70 millions of pictures/day in 2017
Enhance micro-blogging recommendations of posts with an homophily-based graph
Introduction

Real life examples

Finding Users of Interest in Micro-blogging Systems (EDBT 2016)

Enhance micro-blogging recommendations of posts with an homophily-based graph
Problem

How to connect users to relevant messages?

- Recommendation of messages
- 700M new messages every day
- 300M of users
- Real time
Table of contents

1 State of the art
2 Data Analysis
   1 Topology
   2 Retweets
   3 Homophily
3 Approach
   1 Similarity graph
   2 Propagation Model
4 Experiments
   1 Protocol
   2 Results
   3 Updating strategies
5 Conclusion
6 Annexes
Content-based [Lops (2011)]

<table>
<thead>
<tr>
<th>Method</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-based</td>
<td>No need of interactions</td>
<td>tweets are hard to describe</td>
</tr>
</tbody>
</table>
State of the art

Collaborative filtering [Schafer (2007)]

<table>
<thead>
<tr>
<th>Method</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-based</td>
<td>No need of interactions</td>
<td>tweets are hard to describe</td>
</tr>
<tr>
<td>Collaborative filtering</td>
<td>simple model and good results</td>
<td>too large matrix</td>
</tr>
</tbody>
</table>
Matrix Factorization [Koren (2009)]

\[
\begin{pmatrix}
  u_{11} & \cdots & u_{1r} \\
  \vdots & \ddots & \vdots \\
  u_{m1} & \cdots & u_{mr}
\end{pmatrix}
\begin{pmatrix}
  S_{11} & 0 & \cdots \\
  0 & \ddots & \vdots \\
  \vdots & \cdots & S_{rr}
\end{pmatrix}
\begin{pmatrix}
  v_{11} & \cdots & v_{1n} \\
  \vdots & \ddots & \vdots \\
  v_{r1} & \cdots & v_{rn}
\end{pmatrix}
\]

<table>
<thead>
<tr>
<th>Method</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-based</td>
<td>No need of interactions</td>
<td>tweets are hard to describe</td>
</tr>
<tr>
<td>Collaborative filtering</td>
<td>simple model and good results</td>
<td>too large matrix</td>
</tr>
<tr>
<td>Matrix Factorization</td>
<td>efficient to fight sparsity</td>
<td>matrix growing too fast</td>
</tr>
</tbody>
</table>
## State of the art

### Hybrid systems [Bostandjiev (2010)]

<table>
<thead>
<tr>
<th>Method</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-based</td>
<td>No need of interactions</td>
<td>tweets are hard to describe</td>
</tr>
<tr>
<td>Collaborative filtering</td>
<td>simple model and good results</td>
<td>too large matrix</td>
</tr>
<tr>
<td>Matrix Factorization</td>
<td>efficient to fight sparsity</td>
<td>matrix growing too fast</td>
</tr>
<tr>
<td>Hybrid systems</td>
<td>increase user engagement</td>
<td>hard to describe relationship</td>
</tr>
</tbody>
</table>
State of the art

Random walks models [Sharma (2016)]

<table>
<thead>
<tr>
<th>Method</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-based</td>
<td>No need of interactions</td>
<td>tweets are hard to describe</td>
</tr>
<tr>
<td>Collaborative filtering</td>
<td>simple model and good results</td>
<td>too large matrix</td>
</tr>
<tr>
<td>Matrix Factorization</td>
<td>efficient to fight sparsity</td>
<td>matrix growing too fast</td>
</tr>
<tr>
<td>Hybrid systems</td>
<td>increase user engagement</td>
<td>hard to describe relationship</td>
</tr>
<tr>
<td>Random walks models</td>
<td>very cheap</td>
<td>low memory</td>
</tr>
</tbody>
</table>
State of the art
Not only recommendations

- User recommendation (topology, content-based, demographic etc...)
- Hashtag (Bayesian model, euclidien...)
- Timeline Filtering (Deep Learning)
- Few papers on tweets recommendation except Twitter in 2016
Data Analysis

Dataset

Updated connected component from the graph found in [Kwak (2009)].

<table>
<thead>
<tr>
<th>No of nodes</th>
<th>2,182,867</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of edges</td>
<td>325,451,980</td>
</tr>
<tr>
<td>No of tweets</td>
<td>2,571,173,369</td>
</tr>
<tr>
<td>Avg. out-degree</td>
<td>57.8</td>
</tr>
<tr>
<td>Avg. in-degree</td>
<td>69.4</td>
</tr>
<tr>
<td>max out-degree</td>
<td>348,595</td>
</tr>
<tr>
<td>max in-degree</td>
<td>185,401</td>
</tr>
<tr>
<td>Diameter</td>
<td>15</td>
</tr>
<tr>
<td>Average shortest path</td>
<td>3.7</td>
</tr>
</tbody>
</table>

**Table** – Twitter dataset characteristics
Smallest path
Number of paths

Small world with average distance of 3.7

**Figure** – Twitter smallest paths distribution
Data Analysis

Retweets

Figure – Distribution of the number of retweets per tweet

- 1 retweet - 7%
- 2-5 retweets - 1%
- 6+ - 0.2%

Enhance micro-blogging recommendations of posts with an homophily-based graph
Data Analysis

Lifespan

Figure – Lifespan of a message

Enhance micro-blogging recommendations of posts with an homophily-based graph
Data Analysis

Homophily

<table>
<thead>
<tr>
<th>Distance</th>
<th>No of users</th>
<th>%</th>
<th>Mean similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3229</td>
<td>02,65</td>
<td>0,0085</td>
</tr>
<tr>
<td>2</td>
<td>32 668</td>
<td>26,86</td>
<td>0,0014</td>
</tr>
<tr>
<td>3</td>
<td>81 645</td>
<td>67,13</td>
<td>0,0009</td>
</tr>
<tr>
<td>4</td>
<td>3 820</td>
<td>03,14</td>
<td>0,0010</td>
</tr>
<tr>
<td>5</td>
<td>43</td>
<td>00,03</td>
<td>0,0014</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0,0008</td>
</tr>
<tr>
<td>Impossible</td>
<td>216</td>
<td>0,18</td>
<td>0,0017</td>
</tr>
</tbody>
</table>

**Table** – Evolution of the similarity score through distance in the network

\[
sim(u, v) = \frac{\sum_{i \in L_u \cap L_v} \frac{1}{\log(1 + pop(i))}}{|L_u \cup L_v|}
\]  

(1)

Enhance micro-blogging recommendations of posts with an homophily-based graph
Data Analysis

Homophily

**Table** – Link between distance in the network and position in the Top-N

<table>
<thead>
<tr>
<th>Rank</th>
<th>Average Distance</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,55</td>
<td>57,03</td>
<td>31,53</td>
<td>10,64</td>
<td>0,8</td>
</tr>
<tr>
<td>2</td>
<td>1,68</td>
<td>49,60</td>
<td>33,13</td>
<td>16,87</td>
<td>0,4</td>
</tr>
<tr>
<td>3</td>
<td>1,8</td>
<td>42,45</td>
<td>36,02</td>
<td>20,72</td>
<td>0,8</td>
</tr>
<tr>
<td>4</td>
<td>1,86</td>
<td>38,71</td>
<td>38,71</td>
<td>20,56</td>
<td>2,02</td>
</tr>
<tr>
<td>5</td>
<td>1,98</td>
<td>31,44</td>
<td>40,16</td>
<td>27,59</td>
<td>0,81</td>
</tr>
</tbody>
</table>

Enhance micro-blogging recommendations of posts with an homophily-based graph
Data Analysis

Conclusions

Many conclusions from this analysis:

- Freshness is crucial (Messages dies very fast)  
  $\Rightarrow$ real-time recommendation

- Few users have high similarity  
  $\Rightarrow$ use transitivity

- Distance 2 successfully gather important users  
  $\Rightarrow$ rely on this homophily
Similarity Graph

Building process

Figure – Twitter Graph
Graphe de similarité
Exemple de construction

Figure – Twitter Graph
Similarity Graph

Building process

\[ \text{Approach} \]

\[ \text{Similarity graph} \]

\[ \text{Building process} \]

\[ \text{Enhance micro-blogging recommendations of posts with an homophily-based graph} \]

\[ \text{Figure – Twitter Graph} \]
Graphe de similarité

Exemple de construction

Figure – Twitter Graph
Similarity Graph

Building process

\( \text{Figure} \) – Similarity Graph
Similarity Graph

Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Twitter Network</th>
<th>Similarity Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of nodes</td>
<td>2,182,867</td>
<td>1,149,374</td>
</tr>
<tr>
<td>No of edges</td>
<td>325,451,980</td>
<td>4,950,417</td>
</tr>
<tr>
<td>Avg. similarity score</td>
<td>57.8</td>
<td>0.008</td>
</tr>
<tr>
<td>Mean out-degree</td>
<td>5.9</td>
<td></td>
</tr>
</tbody>
</table>

Table – Similarity Graph Characteristics
Propagation Model

In a nutshell

\[ p(u, t) = \frac{\sum_{v \in Fu} p(u \leftarrow v, t)}{|Fu|} \]  \hspace{1cm} (2)

With \( Fu \) the set of users influential to \( u \) and \( p(u \leftarrow v, t) \) a probability estimation that \( u \) likes \( t \) determined by the behavior of the user \( v \).

\[ p(u \leftarrow v, t) = p(v, t) \times \text{sim}(u, v) \]  \hspace{1cm} (3)
Propogation Model

Example

\[
\begin{align*}
\text{U} & \quad 0.3 \quad \text{W} \\
\text{V} & \quad 0.1 \quad \text{Y} \\
\text{W} & \quad 0.5 \quad \text{X}
\end{align*}
\]

**Figure** – Propagation example
Propagation Model

Example

**Figure** – Propagation example - a tweet $t1$ is published
Propagation Model

Example

Figure – Propagation example - X shares/likes t1

\[ p(x, t1) = 1 \]
Propagation Model

Example

\[
p(w, t1) = \frac{\sum_{v \in F_w} p(w \leftarrow v, t)}{|F_w|} = \frac{0 + 1 \times 0.5}{2} = 0.25
\]
Propagation Model

Example

\[ p(u, t1) = \frac{0.25 \times 0.5}{2} = 0.0625 \]
Propagation Model

Convergence

Let $n$ be users $(u_1, u_2, ..., u_n)$:

\[
\begin{align*}
    a_{11} p_{u_1} + a_{12} p_{u_2} + \ldots + a_{1n} p_{u_n} &= b_1 \\
    a_{21} p_{u_1} + a_{22} p_{u_2} + \ldots + a_{2n} p_{u_n} &= b_2 \\
    \vdots &= \vdots \\
    a_{n1} p_{u_1} + a_{n2} p_{u_2} + \ldots + a_{nn} p_{u_n} &= b_n
\end{align*}
\]

Could also be written as $Ap = b$ with

\[
A = \begin{pmatrix}
    u_1 & u_2 & \ldots & u_n \\
    u_1 & a_{11} & a_{12} & \ldots & a_{1n} \\
    u_2 & a_{21} & a_{22} & \ldots & a_{2n} \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    u_n & a_{n1} & a_{n2} & \ldots & a_{nn}
\end{pmatrix},
\]

\[
p = \begin{pmatrix}
    p(u_1) \\
    p(u_2) \\
    \vdots \\
    p(u_n)
\end{pmatrix},
\]

\[
b = \begin{pmatrix}
    b_1 \\
    b_2 \\
    \vdots \\
    b_n
\end{pmatrix}
\]

Because $\forall u, v \ \text{sim}(u, v) \leq 1$, $|a_{jj}| \geq \sum_{j \neq i} |a_{ij}|$ for every $i$, the matrix $A$ is diagonally dominant.
Optimizations

- Speed up the convergence
  Let $\Delta(u, t1) = p(u, t)^{k+1} - p(u, t)^k$
  If $\Delta(u, t1) < \beta$ we stop the propagation
Propagation Model

Optimizations

- Speed up the convergence
  Let $\Delta(u, t1) = p(u, t)^{k+1} - p(u, t)^k$
  If $\Delta(u, t1) < \beta$ we stop the propagation

- Limitation of popular messages
  If $p(u, t) < f(t)$ no need to propagate.
  \[ f(t) = 1 - \frac{k^p}{k^p + \text{pop}(t)^p} \]
Experiments

Protocol

- 130 Millions of messages shared at least twice
- Split the ranked set 90% - 10%
- Compute recommendation during this 10% for 1500 random users (500 small, 500 medium, 500 big)
- Comparison with
  - CF : naive collaborative filtering
  - Bayes : probabilistic model
  - GraphJet : Twitter used solution
Experiments

**Experiments**

**Hits**

**Figure** – Hits pour 1500 utilisateurs

- Linear growth of *CF*
- Fast growth for *SimGraph*
- *GraphJet* stuck around 5000 hits
Experiments

Hits according to user profiles

**Figure – 500 small**

small $< 50$; medium $< 1000$; big $> 1000$

*Tendencies are very stables no matter the profile of users*
Experiments

Hits accuracy

Bayes targets close messages
GraphJet targets popular messages
CF and SimGraph are mixing both popular and close messages

Figure – Hits popularity
Experiments

F1 scores

- Figure – F1 Scores

Enhance micro-blogging recommendations of posts with an homophily-based graph
Experiments

Running time

<table>
<thead>
<tr>
<th>Method</th>
<th>init. (per user)</th>
<th>init total time (1,149,374 users)</th>
<th>time (per message)</th>
<th>total time (70 cores //) 13,238,941 Tweets (Trial period)</th>
<th>total time init + recos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes</td>
<td>10ms</td>
<td>0.04h</td>
<td>975ms</td>
<td>51.22h</td>
<td>51.26h</td>
</tr>
<tr>
<td>CF</td>
<td>8,583ms</td>
<td>39.40h</td>
<td>0.5ms</td>
<td>0.02h</td>
<td>41.01h</td>
</tr>
<tr>
<td>SimGraph</td>
<td>311ms</td>
<td>1.41h</td>
<td>38ms</td>
<td>2.00h</td>
<td>3.41h</td>
</tr>
<tr>
<td>GraphJet</td>
<td>0ms</td>
<td>0h</td>
<td>14ms</td>
<td>4.2h</td>
<td>4.2h</td>
</tr>
</tbody>
</table>

**Table** – Initialization and recommendation time (in ms)
How to update *SimGraph*?

- Split the last 10% in 2
- Evaluate hits prediction impact for the remaining 5%:
  - *do nothing*
  - *recompute everything*
  - *update only weights*
  - *crossfold*
Experiments

Updating strategies

Figure – Hits / updating strategies

- doing nothing is the same as updating weights
- crossfold (very cheap) works very well

Enhance micro-blogging recommendations of posts with an homophily-based graph
Experiments

Convergence property of the SimGraph

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Number of edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4,950,417</td>
</tr>
<tr>
<td>2</td>
<td>7,519,031</td>
</tr>
<tr>
<td>3</td>
<td>10,836,129</td>
</tr>
<tr>
<td>4</td>
<td>11,496,445</td>
</tr>
<tr>
<td>5</td>
<td>11,678,747</td>
</tr>
</tbody>
</table>

Table – Number of edges evolution through iterations
Conclusion

Contribution

- Construction and analysis of a large Twitter dataset
- Method relying on homophily to find nearest neighbors at low cost
- Construction and optimization of a convergent propagation model
- Comparison of the recommendations made by our model with state of the art solutions
- Possibility for the model to be updated at low cost
Conclusion

Future works

- Densify points of comparison between users
- Burst recommendation bubbles
- Work on the *crossfold* convergence of the model
- Add a popularity prediction optimization
Thanks for your attention!
ANNEXES
Annexes

Lifespan and popularity

Figure – Correlation between lifespan and popularity

- Strong correlation up to $10^3$ hours
- After a month, the correlation fades
Topology

**Figure** – Smallest path distribution for the similarity graph

- Diameter of 21 for an average path of 7.5

Enhance micro-blogging recommendations of posts with an homophily-based graph
Similarities

Figure – Score similarity evolution

- Really weak scores
- Breaks after the fifth most similar user
Figure – Parts of hits included in SimGraph

Enhance micro-blogging recommendations of posts with an homophily-based graph
Annexes

Number of recommendations

**Figure** – Recall capacity

- **CF** is less limited
- Other methods are bunched together
- Threshold effect for SimGraph and Bayes