CLUSTERING THE TAGGED WEB

Daniel Ramage - dramage@cs.stanford.edu
Paul Heymann - heymann@cs.stanford.edu
Christopher D. Manning - manning@cs.stanford.edu
Hector Garcia-Molina - hector@cs.stanford.edu

Second ACM International Conference on Web Search and Data Mining
February 2009

Presented By,
Vipin Batchu
Rohit Surve
Chetan Nagarkar
Introduction

• This paper deals with the Automatic clustering of web pages/web documents into semantic groups
• User queries are ambiguous – A keyboard can mean a musical keyboard or a computer keyboard.
• The searchable content on the Internet is increasing at phenomenal rate.
• Clustering similar web documents (based on user tags, tagged data in clustering) can make the search efficient.

“associations between web documents convey information about the relevance of documents to the request”
Example

“Google Engineers curse out NSA over data center intrusions” – GIGAOM

User tags – #ITrevolution, #googleBold, #NSAthrashed

“Apple Comes Clean On How Much Data It's Giving The Government” – HuffingtonPost

User tags – #appleNews, #ITrevolution

• Use the related tags to cluster these two web documents together
• The main idea – using tagged data in web clustering
Process Overview

1. Create groups or clusters from a set of documents (words & tags) using candidate clustering algorithms.
2. Derive the gold standard classes; Open Directory Project
3. Compare the results from step 1) with the gold standard using an evaluation metric.

The clustering algorithms take the document set as input and produce the clusters as output.

Vector Space Model : K-Means clustering
Probabilistic Model : Multi-Multinomial LDA
Dataset

• Original Dataset referred: Stanford Tag Crawl Dataset; contained 2.5 million unique URL’s from 1 month feed of delicious (social bookmarking website)
• Limited consideration to 15K documents that are also present in ODP (for comparison in the end) and ignoring the ‘Regional’ category
• 2000 records were set aside for testing and the rest 13K for implementing the algorithm.

Cluster – F1 Evaluation metric: Harmonic mean of precision and recall.

\[ P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \]

\[ F1 : \frac{2PR}{P+R} \]
K-Means Clustering

- Main Idea – Iteratively reassigning each document to its nearest cluster
- How to find distance of a web document? You can’t actually.
- VSM: all documents are vectors in real-valued space with dimensionality being the entire vocabulary.

- Words Only: (1) \( V(w_1, w_2, \ldots, w_n) \) (2) L2 normalized
- Tags Only: (1) \( V(t_1, t_2, \ldots, t_n) \) (2) L2 normalized
- Words + Tags: \( V_{w+t} = \langle \sqrt{\frac{1}{2}} V_w, \sqrt{\frac{1}{2}} V_t \rangle \).
- Tags as Words times n: \( B_w \cup (B_t \times n) \)
- Tags as new words: tags are different words
K-Means contd.

- Assign weights using term weighting functions.
- Term frequency (tf) : each dimension/# of occurrences.
- Tf – Inverse Document Frequency (tf-idf) : Term frequency downweighted by the log of ratio of total docs to the docs containing the term.

<table>
<thead>
<tr>
<th></th>
<th>tf</th>
<th>tf-idf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>.131</td>
<td>.152</td>
</tr>
<tr>
<td>Tags</td>
<td>.201</td>
<td>.154</td>
</tr>
<tr>
<td>Words+Tags</td>
<td>.209</td>
<td>.168</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>.139</td>
</tr>
<tr>
<td>Tags as Words ×1</td>
<td>.158</td>
</tr>
<tr>
<td>Tags as Words ×2</td>
<td>.176</td>
</tr>
<tr>
<td>Tags as New Words</td>
<td>.154</td>
</tr>
<tr>
<td>Words+Tags</td>
<td>.225</td>
</tr>
</tbody>
</table>
Multi-Multinomial LDA

- This is a variation of Latent Dirichlet Allocation.
- Clustering model with probabilistic semantics to model words and tags as set of observations.
- LDA allows this set of observations to be explained by a set of unobserved variables that explain similarity between data.
- Topic mining: LDA models documents as a set of hidden topic (distribution over words) variables
- Topics are distinct even though the documents are not.

- MM LDA includes both words and tags as distinct sets of observations.

<table>
<thead>
<tr>
<th></th>
<th>(MM-)LDA</th>
<th>K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>0.260</td>
<td>.139</td>
</tr>
<tr>
<td>Tags</td>
<td>0.270</td>
<td>.219</td>
</tr>
<tr>
<td>Words + Tags</td>
<td>0.307</td>
<td>.225</td>
</tr>
</tbody>
</table>
Our Thoughts

• Social tagged data is a useful data source for IR applications
• Accuracy of two algorithms improved by using tagged data.

• There is a performance tradeoff. MM LDA executes 4 times slower than K-Means algorithm modified to accept tag data.

• Further research: Does including anchor text with plain text make tags redundant?
Crime Data Analysis Tool

• This is a Datamart application
• The dataset contains historical data of several incidences/crimes that have occurred in the several areas in San Francisco on specific dates.
• The data consist of several attributes such as the Street-address, PD. District, CrimeType, Description, Severity, DayOfWeek, Resolution, Time, Location, X and Y co-ordinates and so on.
How to use?

• We identified CrimeType, Severity, PD District, DayOfWeek, Resolution as useful attributes to summarize important statistical information pertaining to the crimes prevalent in specific areas.
• Additionally, we can help classify different areas based on the severity of Crimes as Safe or Unsafe.
• We can also provide a statistical report displaying the Specific Crime time and its increase/decrease in count with respect to a specific District/area or with respect to a specific time period.
Who will use?

- This can be used by the SF police to keep a track of the several crimes and their increase/decrease with respect to several SF areas and for specific time period.
- This Mined data can be used by the SF Police Department to identify the Security requirements in different areas of SF.
- This can be used by Real Estate agents and even Home Buyers in selection of safe neighborhood to buy New houses.
- This data can be used by the Government to study the historical Crime data and to alter / introduce new and strict laws based on their severity.
- This can be used by the Hospitals to identify the severity of crimes based on the areas so that they can have the appropriate medical facilities available in case of an emergency.
Tool Design

- Backend: MySQL database
- The schema of the data warehouse is a fact constellation (multiple fact tables for multiple business requirements)
- Frontend: Web application developed in HTML/PHP served over Apache.
Schedule

• Week 10 : Data Preprocessing
• Week 11 : Constructing the Database
• Week 12, 13 : Building the Datamart
References

• Research Paper
  • http://dl.acm.org.proxy.lib.csus.edu/citation.cfm?id=1498759.1498809&coll=DL&dl=ACM&CFID=376898649&CFTOKEN=56207807
  • http://www.huffingtonpost.com/2013/11/06/apple-nsa_n_4224670.html
  • http://gigaom.com/2013/11/06/google-engineers-curse-out-nsa-over-data-center-intrusion/

• Project
  • https://data.sfgov.org/Public-Safety/SFPD-Reported-Incidents-2003-to-Present/dyj4-n68b
  • http://www.google.com/publicdata/directory
  • http://www.homeandlearn.co.uk/php/php1p3.html
THANK YOU 😊