Interactive Hierarchical Tag Clouds
--for Summarizing Spatiotemporal Social Contents

Wei Kang, Anthony K.H. Tung, Feng Zhao and Xinyu Li
Motivation

• Various social networks

• Various info
  – Status
  – Check-in
  – Tagging
  – Photo/video…

• Many analytics applications
  – User relationship
  – Community discovery
  – Popular topics
  – Trend prediction
  – Food/POI suggestion
  – …
Motivation

- **Social (network) contents**
  - A huge number of contents published each day
  - E.g., Twitter was seeing 400 million tweets per day in June 2012

- **Social contents analysis**
  - Extract useful patterns from huge amounts of social contents effectively and efficiently

- **Summarization of contents to disclose underlying knowledge in social network contexts**
Motivation

• **Purpose of summarization of contents**
  – Organize contents into groups
  – Reveal topics of interest
  – Explore oceans of social media data easily

• **Enrich the summarization with multi-faceted features**
  – Spatial
  – Temporal
  – Visualization
  – etc.
Motivation

• User may ask questions like
  – What did people in Chicago mainly talk about during the President election?
  – What are people in Malaysia discussing on Twitter during the missing plane incident?

• The above questions involve 3 parts
  – Where? (Chicago, Malaysia)
  – When? (during the President election/missing plane incident)
  – What topics? (summarizations of social contents)

• By analyzing spatiotemporal social contents, we aim to provide answers for the above questions
  – Objective 1: summarize social contents for a certain geographic region during a certain time period
Motivation

• How to present the extracted summaries?

• Various visualization tools

• Tag cloud for visualizing summaries of social contents

• Objective 2: visualize the summaries in a hierarchical manner by using “hierarchical tag cloud”
What is hierarchical tag cloud?

(a) Parameters

(b) The first level

(c) The second level

(d) The fourth level
Why hierarchical tag cloud?

• **Purpose of summarization of contents**
  – Organize contents into groups
  – Reveal topics of interest
  – Explore oceans of social media data easily

• **Purpose of hierarchical tag cloud**
  – Organize tags of a summary hierarchically
  – Visualize revealed topics of interest
  – Explore summaries and contents interactively
Challenges

• Develop efficient methods to summarize social contents
  – Existing works often with low efficiency
    • e.g., LDA

• Generate a tag hierarchy to visualize a summary in hierarchical tag cloud

• Further enhance system scalability
Outline

• Motivation
• **Summarization by Biclustering Social Contents**
• Partition and Merge (PM) Scheme
• System Architecture
• Experimental Study
• Conclusion
Summarization by Biclustering Social Contents

- Biclustering
- Bicluster(genes, conditions)

- Bicluster(tags, contents)

  Summary

- Density of a Bicluster
  - $0 < \text{density} \leq 1$
Summarization by Biclustering Social Contents

- **Pros and cons of ordinary biclustering methods**

- **Pros**
  - Tags and contents in a bicluster generated simultaneously
  - Tags serve as a summary of the contents
  - Contents provide more context to supplement tags

- **Cons**
  - Unsupervised, iteration over rows & columns
  - NP-complete to find the largest
  - Not suitable for social network contents (sparse, big data)
Biclustering by Formal Concept Analysis

- **Formal Concept Analysis (FCA)**
- **Formal Concept**
- **Two properties**
  - Fullness
  - Maximum
- **Formal concepts viewed as “full-density” biclusters**
  - Small size
  - Relaxation
  - Merge (PM scheme)
Biclustering by Formal Concept Analysis

• **Formal context**: a triplet \((\mathcal{O}, \mathcal{A}, I)\)
  - \(\mathcal{O}\): object set
  - \(\mathcal{A}\): attributes set
  - \(I\): relation, \(I \subseteq \mathcal{O} \times \mathcal{A}\)

• **A formal concept of a context**: \((A, O)\)
  - \(O \subseteq \mathcal{O}, A \subseteq \mathcal{A}\)
  - \(\forall o \in O, a \in A: (o, a) \in I\)
  - \(\forall o \notin O, \exists a \in A: (o, a) \notin I\), vice versa

• \((A, O)\) is full & maximal

• **A formal concept can be perceived as a bicluster with no empty values** (i.e., density=1)

• Our aim is to find all the formal concepts given social network contents (corresponding to \(\mathcal{O}\)) and the tags (corresponding to \(\mathcal{A}\)) in them
Biclustering by Formal Concept Analysis

• **Galois operators**
  - For any $O \subseteq \hat{O}$, $A \subseteq \hat{A}$
  - $A' = \{o \in \hat{O} | (o, a) \in I, \text{ for any } a \in A\}$
  - $O' = \{a \in \hat{A} | (o, a) \in I, \text{ for any } o \in O\}$
  - $A'' = (A')'$, $O'' = (O')'$

  - For any $O = \{o\}$, $A = \{a\}$
  - $a'$, $a''$, $o'$, $o''$ are shown in the right figure
Biclustering by Formal Concept Analysis

- For any \((o, a) \in I\), we can find 4 different biclusters
  - \((a'', o'')\): too tight
  - \((a'', a')\): low fat, has more objects
  - \((o', o'')\): tall thin, has more attributes
  - \((o', a')\): may be sparse, density \(\neq 1\)

- The first three are formal concepts while the last one is not

- Empirically, 2\(^{nd}\) & 3\(^{rd}\) are good choice
- 1\(^{st}\) generates biclusters with very small size
- 4\(^{th}\) tends to generate sparse biclusters

- To remove duplication (avoid generating the same biclusters multiple times), we don’t allow overlapping between biclusters
Biclustering by Formal Concept Analysis

• Example

• Formal Concept
  – Tag-based (vertical scan)
    • \(\{a_2,a_3\}, \{o_2,o_4,o_5\}\)
  – Content-based (horizontal scan)
    • \(\{a_2,a_3,a_4\}, \{o_2,o_4\}\)

• Two properties
  – Fullness
  – Maximum

• Relaxation
Biclustering by Formal Concept Analysis

• Example

• Formal Concept
  – Tag-based (vertical scan)
    • \( \{a_2, a_3\}, \{o_2, o_4, o_5\} \)
  – Content-based (horizontal scan)
    • \( \{a_2, a_3, a_4\}, \{o_2, o_4\} \)

• Two properties
  – Fullness
  – Maximum

• Relaxation
Biclustering by Formal Concept Analysis

• **Example**

• **Formal Concept**
  – Tag-based (vertical scan)
    • \(\{a_2, a_3\}, \{o_2, o_4, o_5\}\)
  – Content-based (horizontal scan)
    • \(\{a_2, a_3, a_4\}, \{o_2, o_4\}\)

• **Two properties**
  – Fullness
  – Maximum

• **Relaxation**
  – \(\{a_2, a_3, a_4\}, \{o_2, o_4, o_5\}\)
Outline

• Motivation
• Summarization by Biclustering Social Contents

• Partition and Merge (PM) Scheme
  – Offline Partitioning
  – Offline Pre-computation
  – Online Merging
  – Mismatch of PM Scheme

• System Architecture
• Experimental Study
• Conclusion
Partition and Merge (PM) Scheme

• **What is PM scheme?**

• **Why?**
  – **Big data**
    • Billions of contents, tens of thousands of tags
    • Gigabytes/terabytes of data storage
  – **Response time**
    • Online computation of biclustering is costly
  – **Scalability**
    • Disk-based
    • Partition data space
    • Process in parallel
Offline Partitioning

- **3 dimensions for spatiotemporal space**
  - Slice data on a daily basis
  - Split (daily) geographic space adaptively accord to data density
    - **Kd-tree/quadtree**
  - Partitioning layout varies daily

![Diagram of offline partitioning](image)
Offline Pre-computation

• Pre-computation is done for each partition of social contents

• Pre-computation of summaries
  – Proposed biclustering approach
  – Assumption: globally interesting summaries are also interesting in certain partitions

• Pre-computation of topic hierarchies
  – Hierarchical Latent Dirichlet allocation (hLDA)
Pre-computation of Topic Hierarchies

- **hLDA: extension of LDA**
  - The topic hierarchy of a partition is generated from the collection of social contents in the generated biclusters of that partition
  - Tree, with a few closely related tags in each node
  - Higher-level tags more general, lower-level tags more specific

- **Example (by applying hLDA to a group of tweets)**
Online Merging

• Merging Biclusters
  – Given time range $R_{\text{tim}}$ and geographic region $R_{\text{geo}}$, biclusters in partitions within the two ranges should be merged to produce “unified” summaries
  – Biclusters with common tags are merged
  – Density(new bicluster) > $\delta_{\text{den}}$

• Example

(a) Biclusters

(b) Inverted list
Online Merging

• **Merging Topic hierarchies**
  
  – topic hierarchies for partitions having merged biclusters also need to be merged to form a *tag hierarchy* so as to visualize the new bicluster

  – **Tag-Level matrix**

    |       | level₁ | level₂ | ...  | levelᵣ |
    |-------|---------|---------|------|---------|
    | *tag₁*| *c₁₁*  | *c₁₂*  | ...  | *c₁ᵣ*  |
    | *tag₂*| *c₂₁*  | *c₂₂*  | ...  | *c₂ᵣ*  |
    | ...   | ...     | ...     | ...  | ...     |
    | *tagᵣ*| *cᵣ₁*  | *cᵣ₂*  | ...  | *cᵣᵣ*  |

  – weight(*tagᵢ*,levelᵢ) = \((c_{ij}/\sum_j c_{ij})*(c_{ij}/\sum_i c_{ij})\)

  – Draw tags for each level without replacement to determine a final tag hierarchy probabilistically
Mismatch of PM Scheme

- Biclusters (topic hierarchies) are merged when corresponding partitions fall into \( R_{\text{tim}} \) and \( R_{\text{geo}} \)

- Two mismatch cases for \( R_{\text{geo}} \)

- Purpose
  - approximate the mismatch quantitatively
  - verify & tune partitioning parameter: \( \delta_{\text{cnt}} \)
Outline

• Motivation
• Summarization by Biclustering Social Contents
• Partition and Merge (PM) Scheme
• System Architecture
• Experimental Study
• Conclusion
System Architecture

- Vesta (Visual exploration of social network data via tag clouds):
  http://db128gb-b.ddns.comp.nus.edu.sg/kangwei/bicluster/
Outline

• Motivation
• Summarization by Biclustering Social Contents
• Partition and Merge (PM) Scheme
• System Architecture
• Experimental Study
• Conclusion
Experimental Study

• Data sets

  – Real-world geo-coded tweets crawled by using Twitter Streaming API

  – Four data sets: 0.1M, 0.7M, 2.5M, 4.3M

  – 4.3M is a proper approximation of # of geo-coded tweets published daily
Experimental Study

• Performance comparison

• Precision and Recall
Experimental Study

- **Assumption Validation**

  ![Graph](image)

  (a) Number of top biclusters
  (b) Half-matched percentage

- **Mismatch Evaluation**

  ![Graph](image)
Experimental Study

- **Offline System Scalability**
  
  ![Graphs showing time and memory scaling for offline system](image)

- **Online System Scalability**
  
  ![Graphs showing time and bicluster number decrease for online system](image)
Conclusion

• We build a system which enables interactive exploration of summaries of spatiotemporal social contents
• We propose a summarization method by biclustering social contents, and extend it to a PM scheme for better scalability purpose
• We generate and merge topic hierarchies so as to visualize summaries in hierarchical tag clouds
Question & Answer

Thank you!