CKM: A Shared Visual Analytical Tool for Large-Scale Analysis of Audio-Video Interviews

Lu Xiao, Yan Luo
The University of Western Ontario, Canada

Steven High
Concordia University, Canada

Oct 8th 2013
Outline

• Introduction: Research Context and Approach
• Data: Stories Matter - An Oral History Database
• Prototype: Clock-based Keyphrases Map
• Future Work
• Q/A
Introduction
How to Discover Patterns?

A Hybrid Solution

Data

Patterns & Knowledge: Categories Cluster Model Rules...

Visualization

Interactive Graphs and Images

Exploration and Analysis

End Users

New Knowledge

Data Mining/Natural Language Processing

Discover

Interactive Visual Analytics

Adjust and Refine the visualization and data mining strategy

Perception

Interactive Visual Analytics
Our Research Context

• Research Problem
  – To develop a tool that facilitates analysis of large scale collections of audio-video recordings in human rights research – Stories Matter database

• Tool Development
  – To develop a *collaborative interactive visual analytical tool* that combines *artificial intelligent algorithms* (e.g., text data mining, natural language processing) and *interactive visualization techniques*
Stories Matter - An Oral History Database
Stories Matter Database

- Centre for Oral History and Digital Storytelling, Concordia University (http://www.stories-matter.com/)

- Allow researchers to interact directly with the original human rights audio-video interviews

- Researchers/Users generated content (200 researchers group/public in the future)
Characteristics of Baseline System

- Meta-data Rich
- Well-Structured
- Mixture of Audio-Text
- Human Tagging
- Series Processing
- Multi-Languages
Design Goals

• Information Retrieval
  • Better access and retrieve the desired information

• Pattern Identification
  • Shared point of views or content among different interviewees
  • Topics of interviews evolution

• Information Sharing and Collaboration
  • Sharing of information among different researchers
CKM: Clock-based Keyphrases Map
Data source management

Automatic Keyphrases Extraction

Keyphrases Map Visualization

Stories Matter Database

Data Entry

Metadata extraction

Text data preprocessing

Generating Training Data

KEA Training

Generating Model

Model

KEA predicting

Top rank keyphrases

Key phrases

GMLs

Visualization

graphs

Data Entry

Extracting time, people, location, etc

Interactions and new knowledge from users

Data Entry

Generating Training Data

Generate Testing Data

Testing Data

Visual Entities

Visualization

Generating Model Testing Data

Generating Model

Model

KEA predicting

Top rank keyphrases

Key phrases

GMLs

Visualization

graphs
Automatic Keyphrases Extraction

• Keyphrases Extraction Algorithm (KEA)
  – TF & IDF and First Occurrence as two main features (more)
  – Naïve Bayes as the discriminative model

• Rationales for KEA:
  – Start point
  – Computational effective and efficient
  – A flexible framework for future extension and refinement
Keyphrases Map Visualization

Determine Node Position

Entity selection

Key phrases list
Node Localization Strategy

• Visual entity on the clock are anchor points

• The node positions are calculated based on the relative weights (distance) on each visual entity

• Weight is calculated based on the commonness/uniqueness to each visual entity (Levenshtein distance)

• Built a global collision avoidance

• Most shared nodes are placed at the center

• Private nodes are placed near each entity
Main User Interactions

- Highlighting
- Zoom and Pan
- Filtering and Reconfiguration
- Context Navigation
- Collaborative Annotation and Feedback
Future Work
Future Work

• User Study and Systemic Evaluation

• Topic modelling and its visual representation

• Intelligent Scaling
Acknowledgement

• Funding support: Social Sciences and Humanities Research Council in Canada (SSHRC)

• Personnel: Kim Moore and Eve-Lyne Cayouette Ashby
Questions and Answer
Hybrid Solution Example: TextFlow, Cui et al. 2011

Fig. 1. Selected topic flows of VisWeek publication data with thread weaving patterns related to primary keywords “graph” and “document” (All keywords overlaid on the threads are manually labeled).
Hybrid Solution Example: Introne and Drescher, 2013

**Topic 3 (Billy)**

- **P1**: billy was VERY withholding of information
- **P2**: Has lied to police and was caught twice.

**Topic 13 (Eddie)**

- **P1**: eddie said he heard the loud muffler, and said it was billie's. was he trying to frame him?
Participatory Design Process

• Requirements analysis
  – A meeting with users of Stories Matter to understand the history of Stories Matter and common use of the database
  – Paper prototyping of three design ideas to solicit users’ feedback on activity and information design

• Iterative design
  – Current prototype – Clock-based Keyphrases Map (CKM)
Contributions

• Visual analytical tool that facilitates the qualitative analysis of collections of audio/video interview recordings about human rights

• Concept of proof of combining text mining and information visualization techniques

• It is applicable to different kinds of time-series sequential data stream