

Search Engine for Shoah Foundation

Presented
by
Ali Khodaei
(khodaei@usc.edu)


- ## Team Members
- Ali Khodaei
 - Kaveh Shahabi
 - Sangeetha U Santharam

- ## Project Motivation
- Existence of huge set of useful data
 - Over 50,000 video testimonies
 - Each divided to one-minute segments
 - Each segment tagged with set of keywords
 - Good amount of spatial and textual data
 - Lack of location-based search engine
 - Lack of an interface to ask for spatial data
 - Lack of ranking/scoring function to rank/score document based on space and text simultaneously

- ## Project Definition
- Robust, efficient and interactive search engine ranking testimonies based on combination of
 - Textual (regular) keywords
 - Spatial keywords
 - This search engine finds and ranks the most textually and spatially relevant testimonies (segments) according to
 - query keywords
 - query location

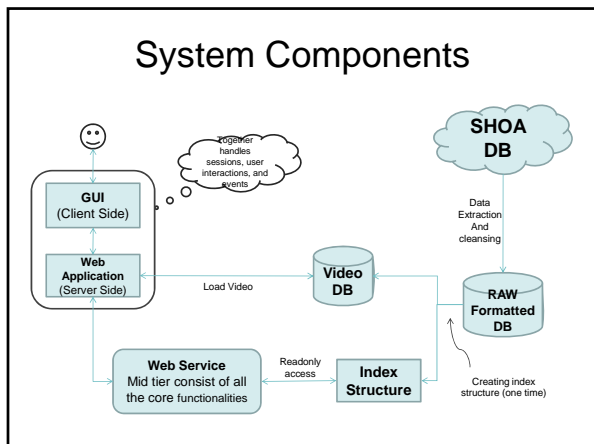
Input

- Query Keywords
 - Set of keywords inputted as text
- Query Location
 - A region drawn on the map OR
 - A spatial keyword inputted as text



Output





- ### Tasks
- 1- Data tier
- Data Cleansing
 - Understand / format / standardize the data
 - Geocoding / GeoTagging
 - Find missing lat/long information for some of spatial keywords
 - Assign appropriate geographical information to each testimony/segment
 - Index Construction
 - Create inverted files for regular keywords
 - Create inverted files for spatial keywords

- ### Tasks
- 2- Middle tier
- Intelligent web-services
 - Talk to interface
 - Receive input (query parameters)
 - Send output (query result)
 - Talk to data tier
 - Get data
 - Access index
 - Access video database
 - Perform necessary operations
 - Process data
 - Calculates scores
 - Format the results

- ### Tasks
- 3- Interface (GUI)
- User friendly interface to receive input from the user
 - Textbox for textual keywords
 - Map interface to draw/show query location
 - A textbox can be used to input a location's name
 - Displays the result dynamically and interactively
 - Results should be changed on-the-fly based on map location
 - Provides mechanism to show the testimonies from the interface
 - Show testimonies on the same page
 - Link to a new page for showing the testimonies

- ### Tasks
- 4- Research/Algorithm
- Hybrid index structure
 - captures spatial and textual keywords (probably using inverted files) simultaneously and efficiently
 - Relevance ranking function
 - Formulas for spatial and temporal scores
 - A combined scoring function with different weights for different features
 - Spatial representation of each segment and/or testimony's spatial data

- ### Break-down + Schedule
- Data tier
 - Understand / format / cleanse (/geocode) / transfer the data
 - 4 weeks **sangy + Ali**
 - Come up with index structure schema for the middle layer
 - 2 weeks **Ali**
 - Create/implement the actual index structure
 - 4weeks **Ali + sangy**
 - Integration/extra,...
 - 1 week **Ali**

Break-down + Schedule

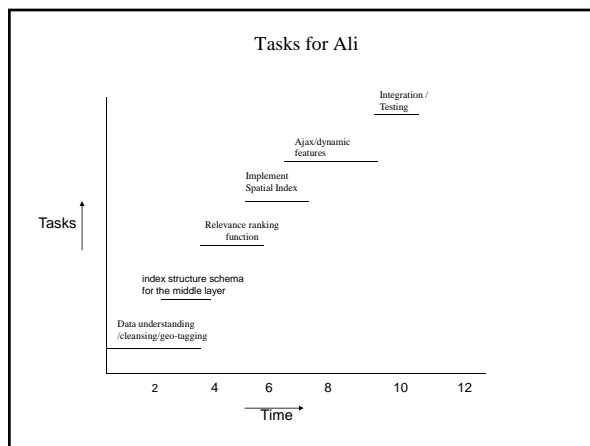
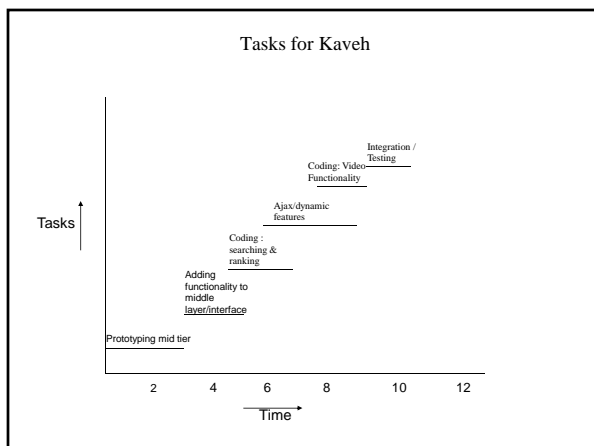
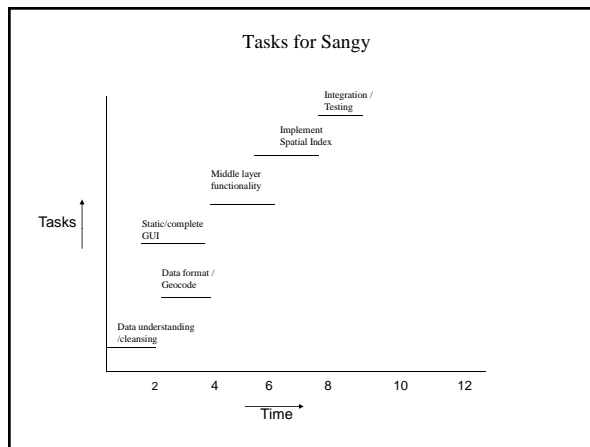
- Research / Algorithm
 - Spatial representation of each segment and/or testimony's spatial data
 - 1.5 weeks Ali + Sangy
 - Relevance ranking function, Formulas for spatial and textual scores
 - 2.5 weeks Ali

Break-down + Schedule

- Middle layer development
 - Creating prototypes /connectivity to the interface
 - 3 weeks Kaveh
 - [1.5 weeks wait for data tier]
 - Create code for ranking function
 - 2.5 weeks Kaveh
 - Create code for video
 - 2 weeks Kaveh
 - Integration/testing
 - 1 week Kaveh

Break-down + Schedule

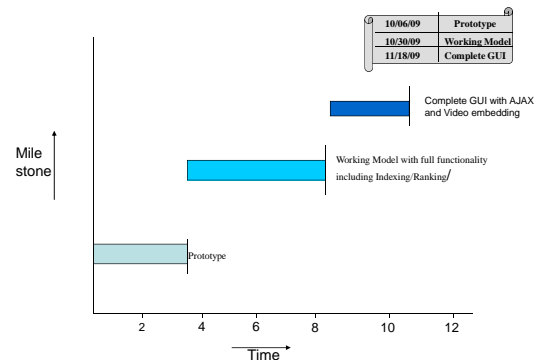
- Web-development
 - Static/complete GUI (no functionality) Sangy
 - 3 weeks
 - Adding functionality Sangy + Kaveh
 - 2 weeks
 - Adding Ajax and dynamic features Kaveh + Ali
 - 4 weeks
 - Integration/test Kaveh + Sangy + Ali
 - 1 week



Deliverables

- 1) Prototype of system having a static (non functional) interface
 - 4th week
- 2) System with actual ranking/index structure and end-to-end functionality
 - 9th week
- 3) (2) + Ajax + video embedding
 - 11th week

Milestones and Deliverables



Resources

- Data
 - Provided by Shoah Foundation
 - data stored in sysbase tables
 - Needs to be cleansed, formatted and transferred
- Software
 - MS Visual Studio .Net
 - Oracle 10g +
- Hardware
 - Windows Server (+IIS)