Understanding Document Collections

Michael DiBernardo
November 7th, 2005
Purpose

• General description of the field
• Demonstrate challenges and techniques for:
  • A problem you’re familiar with
  • A problem you’re unfamiliar with
Purpose

• General description of the field

• Demonstrate challenges and techniques for:
  • A problem you’re familiar with
  • A problem you’re unfamiliar with
Some Tasks:

- Given a set of documents:
  - Find the subset that interests you
  - Categorize the set by composition
  - Compare the ‘utility’ of some subset
  - Quickly understand a single document
Purpose

• General description of the field
• **Demonstrate challenges and techniques for:**
  • A problem you’re familiar with
  • A problem you’re unfamiliar with
Purpose

• General description of the field
• Demonstrate challenges and techniques for:
  • A problem you’re familiar with
  • A problem you’re unfamiliar with
Scenario

- It’s almost Christmas
- We want some turkey
- We want some duck
- Maybe a duck... inside a turkey?
stuffing turkey with duck - Google Search

Web

Turducken - Thanksgiving or Christmas Eating
Detailed recipe for preparing the duck, chicken and turkey along with the different stuffings.
www.thesalmons.org/lynn/turducken.html - 11k - Cached - Similar pages

Roast Goose Recipe with Fruited Stuffing
Roast goose recipe is stuffed with a fruited bread stuffing with raisins and ...
Turkey> Duck and Goose Recipes> Roast Goose Recipe with Fruited Stuffing ... southernfood.about.com/od/duckandgooserecipes/r/bln71.htm - 24k - Cached - Similar pages

Turkey Basics - Stuffing
... and handling of chicken, turkey, duck, goose, and other poultry products. ...
For safety, prepare stuffing or dressing for the turkey according to these ...
www.fsis.usda.gov/Fact_Sheets/Poultry_Preparation_Fact_Sheets/index.asp - 38k - Cached - Similar pages

It's a turkey! It's a duck! It's a chicken! It's ... turducken!
... semi-boneless turkey -- the wings and drumsticks remain--that is stuffed with a ...
Would a duck stuffed with a hen stuffed with a quail be a danquail? ...
www.post-gazette.com/food/20000319turducken3.asp - 25k - Cached - Similar pages
Observation:

- It’s hard to tell which documents:
  - Are talking about turkey and duck
  - Are talking about a turducken
  - Contain recipes or histories
The Big Problem:

- Have to scan each document to see if it’s appropriate
- Ranking procedure is opaque
Hearst’s Idea

- When searching fulltext, user should know:
  - Relative lengths of documents
  - Frequency of query terms
  - *Distribution of terms* in the document
Proposed Solution

• CHI ‘96 video
Method

- Split document into chunks (topical, paragraph)
- Each chunk represented by a TextTile
- Darkness = frequency
Encodings

- Relative length of document = bar length
- Term frequency = TextTile darkness
- Term distribution = TextTile position
Strengths

• Simple and intuitive to use
• Based on accepted vis principles
• Can be integrated with ranking
• Documents can be ‘pre-chunked’
Weaknesses

- Long documents cause problems
- No discussion of how to chunk document
- No user study
- TextTiles give no indication of passage length
- Colour?
Applications

• Was incorporated in Berkeley’s e-library
• Clunky applet solution
• Addressed length problem
Applications

- Good candidate to supplement web search
- Would be interesting to see reactions
Before it can be used:

- Need to deal with length problem
- Decide how to position in search interface
Summary

- TileBars good summary of document structure and term frequency distribution
- Supplements ranked search
- Lightweight solution to a common problem
Purpose

• General description of the field
• Demonstrate challenges and techniques for:
  • A problem you’re familiar with
  • A problem you’re unfamiliar with
The Problem

- Understanding how the thematic composition of document collection changes over time
Observation

- People tend to abstract time as *relative motion*.
- Consider pithy sayings:
  - “Don’t let your life pass you by...”
  - “Your time will come...”
Idea

- Use visual metaphor of a river
Encodings

- Time = horizontal axis
- Thematic strength = height of stream
- Salient events plotted with time scale
Example

• We have 100,000 Associated Press articles from 1990

• How did media react to Iraqi invasion of Kuwait?
German economic and monetary union

OPEC agrees to raise oil price

NATO to redefine military strategy

Iraq invades Kuwait

Jul 1990     Aug 1990
Sample analysis

• What are those periodic ‘chokepoints’?
German economic and monetary union

NATO to redefine military strategy

OPEC agrees to raise oil price

Iraq invades Kuwait

nato(54)
germany(36)
communist(5)
bush(1)
oil(12)

baghdad(11)
saddam(10)
iraq(9)
kuwait(8)

Jul 1990

Aug 1990
Sample analysis

• On Sundays:
  • Is overall volume reduced?
  • Or are our themes less popular?
Plotting streams

- For a theme $T$:
  - Articles of $T$ are sorted into bins by time
  - For each bin, plot the *number* of articles
Plotting streams

- To keep curve smooth:
  - Interpolate between points
  - At $t'$ between bin $t$ and $t+1$:
    - $y(t) \leq y(t') \leq y(t+1)$
    - Or vice-versa if decreasing
Plotting streams

- Streams are then ‘stacked’ to create river
So what about Sundays?

• We don’t know how many articles were published *in total* on a given day

• So we can’t tell if
  
  • our themes were unpopular, or
  
  • Sundays have reduced distribution
Need more information

• Forced to plot histograms
Other problems?

- User study done to identify other problems
- Used collection of Castro’s speeches
- Compared to a plain histogram (!)
- Two users: No mention of competency
User comments

+ Users comfortable with river metaphor
+ Smooth curves aid navigation
+ Easily identified macro trends
+ Became inquisitive about theme changes
User comments

- Hard to compare micro changes
- Interpolation breeds mistrust
- Theme ordering
- Want access to more detail
User study

- User study highlighted some key strengths and weaknesses of approach
Strengths

- Comparing between small set of collections is fairly easy (without histograms)
Earthquake in Iran
 kills 40,000

NATO to redefine
 military strategy

earthquake
 apartheid
 mandela
 iran

supreme
 senate
 mandela
 bush

Jul 1990
Strengths

• Can easily relate events to macro trends (when close to related event!)
• Smooth curves easy to follow
• Performed (minimal) user study, and very forthcoming about weaknesses
Weaknesses

- Colour selection
- Differences in colour groups imply an ordering
germany (36)
unification (37)
gdr (38)
kohl (39)
ceausescu (40)
hungary (41)
iliescu (42)
bucharest (43)
Weaknesses

- Theme labels are hard to resolve
German economic and monetary union

NATO to redefine military strategy

OPEC agrees to raise oil price

Iraq invades Kuwait
Weaknesses

• Theme ordering matters
  • In how river looks (stacking)
  • In ease of theme comparison
• No discussion of theme selection
  • How to select good theme set?
Weaknesses

- Micro trends difficult to perceive
- Hard to select good “thematic strength”
- The “Sunday problem”
Weaknesses

- How to enable details-on-demand?
- Picking specific documents
- Seeing what types of documents compose streams
Applications

- Patterns of media coverage
- Discussion flows in meetings (minutes)
- Changes in patient records
- Juvenile criminal records
- Politics
- Assessing tech trends through patents
Summary

- Wanted to view thematic change over time
- Use river as visual metaphor
- Effective at identifying macro trends
- Need to address colour selection, ambiguities in thematic strength, enabling details on demand
In Closing

- Understanding Document Collections:
  - Search
  - Thematic composition
  - Document structure comparison
  - Document browsing
In Closing

- We saw two domain problems:
  - Understanding search results on full-length documents
  - Understanding thematic change over time
In Closing

• We saw two (attempted) solutions:
  • TileBars - succinct representation of document length and structure
  • ThemeRiver - visual metaphor to describe how themes change over time