VizCommender

Computing Text-Based Similarity in Visualization Repositories for Content-Based Recommendations

Michael Oppermann, Robert Kincaid, and Tamara Munzner

Conference Talk, IEEE VIS 2020

michaeloppermann.com/work/viz-commender
Large-scale visualization repositories
Users have difficulty discovering relevant content.

Users often start from scratch instead of reusing content.
Recommendation systems are increasingly used to assist users by surfaced relevant content.
Visual encoding recommendation

Tableau ShowMe, Voyager, Draco, Data2Vis, …
Visual encoding recommendation
Tableau ShowMe, Voyager, Draco, Data2Vis, …
Visualization workbook recommendations based on content features
Recommendation Systems

Content-based filtering

Collaborative filtering
Recommendation Systems

Content-based filtering

- Focus of our work
- Finding relevant items based on their actual content
- Less diverse but more accurate recommendations
- Allows identification of near-duplicate items

Collaborative filtering
Recommendation Systems

Content-based filtering

Collaborative filtering

- Recommendations based on user interactions
- Requires no domain knowledge, allows fast computation, serendipitous recommendations
- *Cold start* problem for new items or new users
Recommendation Systems

Content-based filtering

Collaborative filtering

Hybrid system
Recommendation Systems

- Content-based filtering
- Collaborative filtering

→ Hybrid system
Which content features are most informative for comparisons?

What techniques can we use for comparing and ranking viz specifications?
Text-based similarity measure

- Content-based recommendations
- Facilitate information seeking
Overview

Close collaboration with the Recommender Systems Group at Tableau
Overview

Close collaboration with the Recommender Systems Group at Tableau

**VizCommender**

- Extract content from viz specifications
- Analysis & feature engineering
- Proof-of-concept interface
Overview

Close collaboration with the *Recommender Systems Group* at Tableau

**VizCommender**
- Extract content from viz specifications
- Analysis & feature engineering
- Proof-of-concept interface

**User study**: Crowdsourced human text similarity judgements

**Comparative model analysis**
VizCommender prototype
Tableau Visualization Workbook

Sales Commission

New quota: $500K
Base salary: 50,000

Estimated Quota Attainment

- Barbara Davis: 35%
- Betty Clark: 26%
- Carol Allen: 16%
- Charles Lee: 59%
- Christopher: 78%
- Daniel Gonzalez: 53%
- David Thompson: 62%

Commissions

Overview

Sales and Profit by Customer

- Count of Customers
  - West
  - East
  - Central
  - South

Number of Shipments

- Profits
  - $0
  - $5,000
  - $10,000
  - $15,000
  - $20,000
  - $25,000

Customer Analysis
Workbook

Sales and Profit by Customer

Visualization Specification

```xml
<worksheet name="Customer Analysis">
  <layout-options>
    <title>
      <formatted-text>
        <run>Sales and Profit by Customer</run>
      </formatted-text>
    </title>
  </layout-options>
  <table>
    <rows>[sum:Profit:qk]</rows>
    <cols>[sum:Sales:qk]</cols>
    ...
  </table>
</worksheet>
```
**Workbook**

**Visualization Specification**

```xml
<worksheet name="Customer Analysis">
  <layout-options>
    <title>
      <formatted-text>
        <run>Sales and Profit by Customer</run>
      </formatted-text>
    </title>
  </layout-options>
  <table>
    <rows>...
      <sum column="Profit"/>
    </rows>
    <cols>...
      <sum column="Sales"/>
    </cols>
  </table>
  ...
</worksheet>
```
Workbook

Sales and Profit by Customer

Visualization Specification

```xml
<worksheet name="Customer Analysis">
    <layout-options>
        <title>
            <formatted-text>
                <run>Sales and Profit by Customer</run>
            </formatted-text>
        </title>
    </layout-options>
    <table>
        <rows>[sum:Profit:qk]</rows>
        <cols>[sum:Sales:qk]</cols>
    </table>
</worksheet>
```
Workbook

Sales and Profit by Customer

Visualization Specification

```xml
<worksheet name="Customer Analysis">
  <layout-options>
    <title>
      <formatted-text>
        <run>Sales and Profit by Customer</run>
      </formatted-text>
    </title>
  </layout-options>
  <table>
    <rows>[sum:Profit:qk]</rows>
    <cols>[sum:Sales:qk]</cols>
  </table>

  <datasource>
    <metadata-records>
      <metadata-record class="column">
        <remote-name>City</remote-name>
      </metadata-record>
      <metadata-record class="column">
        <remote-name>Customer Name</remote-name>
      </metadata-record>
      <metadata-record class="column">
        <remote-name>Sales</remote-name>
      </metadata-record>
    </metadata-records>
  </datasource>
</worksheet>
```
Initial Experiments

Extracted text

Visual encodings
Different data, same encoding

Governor WA

Bryant

Inslee

Sales

Office Supplies

Furniture
Features: Leaving Out Visual Encodings
Features: Leaving Out Visual Encodings

For design inspiration and learning
  ‣ Need visual style
Features: Leaving Out Visual Encodings

For design inspiration and learning
  ‣ Need visual style

Our primary task: Information seeking
  ‣ Core enterprise task
  ‣ Subject matter of a workbook
  ‣ **Do not need** visual style (marks, colors, layout properties, ...)
Data Challenges
Data Challenges

Very limited text
Data Challenges

**Very limited text**

Additional challenges:

- Multi-sheet workbooks and nested visualizations
Data Challenges

Very limited text

Additional challenges:

- Multi-sheet workbooks and nested visualizations
- Incomplete workbooks
Data Challenges

**Very limited text**

Additional challenges:

- Multi-sheet workbooks and nested visualizations
- Incomplete workbooks
- Multiple versions
Data Challenges

**Very limited text**

Additional challenges:

- Multi-sheet workbooks and nested visualizations
- Incomplete workbooks
- Multiple versions
- Out-of-vocabulary words
Extracted text

customer analysis sales profit discount
commission segment ratio ranking count ship
performance target furniture office home
supplies city drilldown late early product
category forecast order quantity target ...
Extracted text

customer analysis sales profit discount
commission segment ratio ranking count ship
performance target furniture office home
supplies city drilldown late early product
category forecast order quantity target ...

Numeric document representation

| 0.37546 | 0.13540 | 0.01713 | 0.04225 | 0.01993 | ...

Transform?
Extracted text

customer analysis sales profit discount commission segment ratio ranking count ship performance target furniture office home supplies city drilldown late early product category forecast order quantity target ...

Transform?

Numeric document representation

| 0.37546 | 0.13540 | 0.01713 | 0.04225 | 0.01993 | ... |

Comparisons?
Extracted text

customer analysis sales profit discount commission segment ratio ranking count ship performance target furniture office home supplies city drilldown late early product category forecast order quantity target ...

Transform?

Numeric document representation

| 0.37546 | 0.13540 | 0.01713 | 0.04225 | 0.01993 | ... |

Recommendations

Comparisons?

39
NLP Models

- TF-IDF & cosine similarity
- Latent semantic indexing (LSI) & cosine similarity
- Latent dirichlet allocation (LDA) & Jensen-Shannon divergence
- Word embeddings (Doc2Vec, GloVe) & cosine similarity
Overview

VizCommender
- Extract content from viz specifications
- Analysis & feature engineering
- Proof-of-concept interface

User study: Crowdsourced human text similarity judgements

Comparative model analysis
Overview

VizCommender

- Extract content from viz specifications
- Analysis & feature engineering
- Proof-of-concept interface

User study: Crowdsourced human text similarity judgements

Comparative model analysis
Crowdsourced human similarity judgements
2-Alternative Forced Choice Experiment

135 Triplets

Reference

Alternative 1

Alternative 2
2-Alternative Forced Choice Experiment

135 Triplets

Reference

Alternative 1

Alternative 2

75 Participants
2-Alternative Forced Choice Experiment

- 135 Triplets
  - Reference
  - Alternative 1
  - Alternative 2

- 75 Participants

- NLP models
Experimental Stimulus

<table>
<thead>
<tr>
<th>Flight incidents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q2</strong></td>
</tr>
<tr>
<td>SUM(Survived)</td>
</tr>
<tr>
<td>pclass survived</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>National Parks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sheet 7</strong></td>
</tr>
<tr>
<td>Park Name</td>
</tr>
<tr>
<td>Park</td>
</tr>
<tr>
<td>Park Name (copy)</td>
</tr>
</tbody>
</table>

**Is A or B more similar to the Reference?**

Reference

<table>
<thead>
<tr>
<th>Sheet 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>airplane safety</td>
</tr>
<tr>
<td>AVG(Age)</td>
</tr>
<tr>
<td>pclass</td>
</tr>
</tbody>
</table>
Experimental Stimulus

Is A or B more similar to the Reference?

Reference

Baseball Story Final
Sheet 9
Height & Batting Average & Handedness

<table>
<thead>
<tr>
<th>height</th>
<th>AVG(avg)</th>
<th>handedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>name, handedness, height, weight, avg, HR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Final Visualization

Olympics 2016 - Which Athlete and Sport are you

<table>
<thead>
<tr>
<th>Event</th>
<th>SUM(Number of Athletes)</th>
<th>Sport</th>
</tr>
</thead>
<tbody>
<tr>
<td>id, name, nationality, sex, date_of_birth, height, weight, sport, gold, silver, bronze, year_of_birth, Bronze (copy), Gold (copy), Height (copy), id, nationality (copy), Silver (copy), Weight (copy), year_of_birth (copy), year_of_birth</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A

B
# Agreement Scores

<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>TF-IDF</th>
<th>GloVe</th>
<th>Doc2Vec</th>
<th>LSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>.978</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GloVe</td>
<td>.978</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc2Vec</td>
<td>.912</td>
<td>.889</td>
<td>.889</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSI</td>
<td>.935</td>
<td>.956</td>
<td>.956</td>
<td>.848</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>.914</td>
<td>.892</td>
<td>.892</td>
<td>.871</td>
<td>.852</td>
</tr>
</tbody>
</table>
## Agreement Scores

<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>TF-IDF</th>
<th>GloVe</th>
<th>Doc2Vec</th>
<th>LSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>.978</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GloVe</td>
<td>.978</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc2Vec</td>
<td>.912</td>
<td>.889</td>
<td>.889</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSI</td>
<td>.935</td>
<td>.956</td>
<td>.956</td>
<td>.848</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>.914</td>
<td>.892</td>
<td>.892</td>
<td>.871</td>
<td>.852</td>
</tr>
</tbody>
</table>
Agreement Scores

<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>TF-IDF</th>
<th>GloVe</th>
<th>Doc2Vec</th>
<th>LSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>.978</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GloVe</td>
<td>.978</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc2Vec</td>
<td>.912</td>
<td>.889</td>
<td>.889</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSI</td>
<td>.935</td>
<td>.956</td>
<td>.956</td>
<td>.848</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>.914</td>
<td>.892</td>
<td>.892</td>
<td>.871</td>
<td>.852</td>
</tr>
</tbody>
</table>

- Very good alignment between human similarity judgements and off-the-shelf model predictions
- LDA performed slightly better
Gov. Gavin Newsom declared a state of emergency Tuesday in response to wildfires in California, as the state gave evacuation orders and battled the effects of a sweltering heat wave, rolling blackouts and the coronavirus pandemic.

By early Wednesday morning, the state fire authorities had ordered residents to evacuate in parts of Santa Cruz, San Mateo, Napa and Sonoma Counties, in Northern California, where thunderstorms brought lightning strikes this week.

The largest fire in the region, called the SCU Lightning Complex, had spread to 35,000 acres in several counties east of San Jose and was 4 percent contained. Another fire, called the LNU Lightning Complex fire, was quickly growing north of the Bay Area, with 32,000 acres burned by about 9:30 Tuesday night.

That fire forced evacuations in parts of Napa and Sonoma, with the authorities warning of an “immediate threat to life” in some places. Local news outlets showed structures consumed by flames in Vacaville, about 35 miles southwest of Sacramento, and fire overtaking a camera meant to help spot wildfires on Mount Vaca. Photos and videos on social media showed flames lapping at the road and, in the hours before dawn, some images showed a glowing red sky, as the fire lit up dense smoke.

To the south, residents in Oakland and San Francisco could smell smoke as they woke up on Wednesday morning. The authorities around Northern California warned of poor air quality in addition to the rising heat …
Gov. Gavin Newsom declared a state of emergency Tuesday in response to wildfires in California, as the state gave evacuation orders and battled the effects of a sweltering heat wave, rolling blackouts and the coronavirus pandemic. By early Wednesday morning, the state fire authorities had ordered residents to evacuate in parts of Santa Cruz, San Mateo, Napa and Sonoma Counties, in Northern California, where thunderstorms brought lightning strikes this week. The largest fire in the region, called the SCU Lightning Complex, had spread to 35,000 acres in several counties east of San Jose and was 4 percent contained. Another fire, called the LNU Lightning Complex fire, was quickly growing north of the Bay Area, with 32,000 acres burned by about 9:30 Tuesday night. That fire forced evacuations in parts of Napa and Sonoma, with the authorities warning of an “immediate threat to life” in some places. Local news outlets showed structures consumed by flames in Vacaville, about 35 miles southwest of Sacramento, and fire overtaking a camera meant to help spot wildfires on Mount Vaca. Photos and videos on social media showed flames lapping at the road and, in the hours before dawn, some images showed a glowing red sky, as the fire lit up dense smoke. To the south, residents in Oakland and San Francisco could smell smoke as they woke up on Wednesday morning. The authorities around Northern California warned of poor air quality in addition to the rising heat …
Proof-of-concept implementation
Are people visiting the national parks of America?

Click on a decade to filter


Which states had the most recreational visitors in the 2010's?

The 20 most visited parks overall
Visitors since the 1920's

Golden Gate  35,604
Blue Ridge Parkway  33,298
Great Smoky Mountains  31,139
George Washington MEM  30,241
Gateway  29,786
Lincoln Memorial  27,942
Lake Mead  27,265
Natchez Trace Parkway  25,981
Grand Canyon  25,004
Gulf Island  24,981
Cape Co  24,765
Delaware Water Gap  23,769
Yosemite  23,041
Independence  21,581
Yellowstone  20,934
Statue of Liberty  18,765
Rocky Mountain  18,364
Colonial  13,461
Acadia  12,461

Top 20 park development over decades

Interactive workbook
Are people visiting the **national parks** of America?

Click on a decade to filter:

- 1920's
- 1930's
- 1940's
- 1950's
- 1960's
- 1970's
- 1980's
- 1990's
- 2000's
- 2010's

Which states had the most recreational visitors in the 2010's?

The 20 most visited parks overall:

- Golden Gate
- Blue Ridge Parkway
- Great Smoky Mountains
- George Washington
- National Park
- Yosemite
- Denali National Park
- Yosemite National Park
- Gulf Coast
- Great Lakes
- Yellowstone
- Zion National Park
- Glacier
- Yosemite National Park
- Grand Canyon
- Grand Teton
- Acadia
- Shenandoah
- Death Valley
- Rocky Mountain

Top 20 park development over decades

**Recommendation panel**
Similarity facets for different user tasks

- Related workbooks
- Similar data
- Similar versions
Similarity facets for different user tasks

- Related workbooks
- Similar data
- Similar versions
Similarity facets for different user tasks

- Related workbooks
- Similar data
- Similar versions
Similarity facets for different user tasks

- Related workbooks
- Similar data
- Similar versions
Generalizable to other viz specifications
Generalizable to other viz specifications
VizCommender

Computing Text-Based Similarity in Visualization Repositories for Content-Based Recommendations

Michael Oppermann, Robert Kincaid, and Tamara Munzner

michaeloppermann.com/work/viz-commender

Contributions

- Challenges for content-based visualization recommendations
- Design and implementation of a proof-of-concept pipeline
- Analysis of applicable NLP techniques and a user study assessing the alignment with human judgements of similarity
Interactive Visual Analysis Tool
<table>
<thead>
<tr>
<th>ID</th>
<th>TIMING</th>
</tr>
</thead>
<tbody>
<tr>
<td>187</td>
<td>17min (19)</td>
</tr>
<tr>
<td>184</td>
<td>12min (19)</td>
</tr>
<tr>
<td>183</td>
<td>10min (10)</td>
</tr>
<tr>
<td>182</td>
<td>17min (21)</td>
</tr>
<tr>
<td>181</td>
<td>6min (6)</td>
</tr>
<tr>
<td>179</td>
<td>7min (11)</td>
</tr>
<tr>
<td>176</td>
<td>7min (9)</td>
</tr>
<tr>
<td>175</td>
<td>10min (13)</td>
</tr>
<tr>
<td>174</td>
<td>11min (13)</td>
</tr>
<tr>
<td>173</td>
<td>8min (10)</td>
</tr>
<tr>
<td>172</td>
<td>6min (7)</td>
</tr>
<tr>
<td>170</td>
<td>9min (11)</td>
</tr>
<tr>
<td>169</td>
<td>7min (9)</td>
</tr>
<tr>
<td>168</td>
<td>14min (16)</td>
</tr>
<tr>
<td>167</td>
<td>13min (17)</td>
</tr>
<tr>
<td>166</td>
<td>20min (28)</td>
</tr>
<tr>
<td>165</td>
<td>14min (17)</td>
</tr>
<tr>
<td>164</td>
<td>4min (7)</td>
</tr>
<tr>
<td>163</td>
<td>14min (15)</td>
</tr>
<tr>
<td>162</td>
<td>7min (9)</td>
</tr>
<tr>
<td>161</td>
<td>30min (32)</td>
</tr>
<tr>
<td>159</td>
<td>6min (9)</td>
</tr>
<tr>
<td>158</td>
<td>7min (9)</td>
</tr>
<tr>
<td>157</td>
<td>18min (20)</td>
</tr>
</tbody>
</table>

| TOTAL | AVG. | MEDIAN | MIN. | MAX. | POS A | POS B | INTRARATER AGREEMENT | INTER-RATER AGREEMENT | EXPERT AGREEMENT | DIFFICULTY | BOTH RELEVANT | NONE RELEVANT | DEMOGRAPHICS |
|-------|------|--------|------|------|-------|-------|-----------------------|-----------------------|----------------|------------|-------------|--------------|--------------|---------------|
| 112   | 22s  | 18s    | 4s   | 2min | 21    | 26    | 100%                  | 96%                   | 97.87%        | 3/5        | 2/5         | 2/5          | 54, female, Training |
| 108   | 16s  | 13s    | 7s   | 37s  | 21    | 26    | 100%                  | 91%                   | 93.62%        | 2/5        | 2/5         | 2/5          | 57, female, Training |
| 100   | 12s  | 7s     | 2s   | 2min | 22    | 25    | 50%                   | 84%                   | 82.98%        | 3/5        | 4/5         | 2/5          | 30, male, College, no degree |
| 98    | 12s  | 7s     | 2s   | 2min | 22    | 25    | 50%                   | 81%                   | 86.36%        | 4/5        | 4/5         | 2/5          | 34, female, Bachelor's degree |
| 92    | 11s  | 7s     | 3s   | 3min | 21    | 26    | 100%                  | 96%                   | 96.74%        | 1/5        | 3/5         | 1/5          | 40, male, Bachelor's degree |
| 87    | 10s  | 6s     | 3s   | 32s  | 26    | 21    | 100%                  | 98%                   | 95.74%        | 5/5        | 3/5         | 3/5          | 26, male, Bachelor's degree |
| 84    | 13s  | 10s    | 5s   | 37s  | 25    | 22    | 100%                  | 96%                   | 93.62%        | 2/5        | 2/5         | 1/5          | 63, male, Bachelor's degree |
| 82    | 12s  | 10s    | 5s   | 37s  | 25    | 22    | 100%                  | 98%                   | 95.74%        | 5/5        | 3/5         | 1/5          | 31, male, College, no degree |
| 80    | 11s  | 8s     | 3s   | 3min | 17    | 30    | 100%                  | 98%                   | 95.11%        | 2/5        | 5/5         | 1/5          | 32, female, Bachelor's degree |
| 77    | 8s   | 7s     | 3s   | 29s  | 27    | 20    | 100%                  | 98%                   | 95.74%        | 2/5        | 1/5         | 2/5          | 35, female, Bachelor's degree |
| 76    | 11s  | 9s     | 3s   | 31s  | 24    | 23    | 100%                  | 91%                   | 89.36%        | 3/5        | 4/5         | 1/5          | 28, male, College, no degree |
| 74    | 10s  | 9s     | 4s   | 20s  | 22    | 25    | 100%                  | 94%                   | 91.49%        | 2/5        | 2/5         | 2/5          | 30, male, College, no degree |
| 73    | 18s  | 7s     | 3s   | 3min | 27    | 20    | 100%                  | 74%                   | 76.6%         | 4/5        | 2/5         | 5/5          | 60, male, Associate's degree |
| 71    | 16s  | 14s    | 4s   | 1min | 22    | 25    | 100%                  | 83%                   | 89.36%        | 2/5        | 3/5         | 2/5          | 40, female, Bachelor's degree |
| 69    | 18s  | 7s     | 3s   | 4min | 22    | 25    | 100%                  | 91%                   | 89.36%        | 2/5        | 2/5         | 2/5          | 41, female, Bachelor's degree |
| 66    | 5s   | 5s     | 2s   | 10s  | 25    | 22    | 100%                  | 89%                   | 91.49%        | 1/5        | 2/5         | 2/5          | 36, male, Bachelor's degree |
| 65    | 18s  | 15s    | 4s   | 44s  | 22    | 25    | 50%                   | 89%                   | 89.36%        | 4/5        | 2/5         | 3/5          | 36, male, College, no degree |
| 61    | 9s   | 8s     | 4s   | 20s  | 28    | 19    | 50%                   | 98%                   | 98.77%        | 3/5        | 2/5         | 2/5          | 37, male, Bachelor's degree |
| 59    | 35s  | 13s    | 1min | 29    | 18    | 50%                   | 91%                   | 89.36%        | 6/5        | 5/5         | 2/5          | 69, female, College, no degree |
| 58    | 10s  | 7s     | 3s   | 37s  | 24    | 23    | 23%                   | 87%                   | 85.11%        | 2/5        | 2/5         | 1/5          | 33, female, Bachelor's degree |
| 57    | 9s   | 8s     | 4s   | 27s  | 24    | 23    | 100%                  | 85%                   | 87.23%        | 1/5        | 2/5         | 1/5          | 40, female, High school |
| 56    | 23s  | 17s    | 6s   | 2min | 23    | 24    | 100%                  | 94%                   | 95.74%        | 3/5        | 2/5         | 3/5          | 39, female, Bachelor's degree |