Aggregated Search: Motivations, Methods, and Milestones

Jaime Arguello
INLS 613: Text Data Mining
jarguell@email.unc.edu

November 19, 2014
Traditional Information Retrieval

- retrieving full-text documents from a single collection in response to a user’s query

\[ P(D|Q) \propto P(Q|D) \times P(D) \]

query-document similarity

document prior
Information Needs in Today’s World

news

News for **pittsburgh snow**
Western Pennsylvania blanketed by snow
5 hours ago
By Pittsburgh Tribune-Review Scattered snow showers and gusting winds could make roads hazardous as the lake effect snow continues today. ...
Tribune Review - 14 related articles
Cold, snow hit schools, travel in many states
Reuters - 2675 related articles
Snow bands crossing region could bring icy commute
Pittsburgh Post Gazette - 2 related articles - Shared by 5+

local business listings

Places for **apple store** near Pittsburgh, PA

- **Apple Store South Hills Village**
  - 5 reviews - Place page
  - www.apple.com - 301 South Hills Village, Pittsburgh - (412) 833-1840

- **Apple Store Shadyside**
  - 8 reviews - Place page
  - www.apple.com - 5508 Walnut Street, Pittsburgh - (412) 683-1186

- **Apple Store Ross Park**
  - Place page
  - www.apple.com - 1000 Ross Park Mall Drive, Pittsburgh - (412) 318-0656
Information Needs in Today’s World

Images

Videos
Information Needs in Today’s World

books

pittsburgh

Books for pittsburgh

Pittsburgh: a sketch of its early social life - Charles William Dahlinger - 1916 - 216 pages
Pittsburgh: 17582008 surveys the city's evolution from strategic fort in the wilderness ...
books.google.com

pittsburgh steelers

Related blogs about pittsburgh steelers

Behind the Steel Curtain - For Pittsburgh Steelers Fans
Your best source for quality Pittsburgh Steelers news, rumors, analysis, stats and scores from the ...
www.behindthesteelcurtain.com/

Nice Pick, Cowher | A Pittsburgh Steelers Blog
Pittsburgh Steelers running back Rashard Mendenhall spent his off day in Monroeville helping fans ...
nicepickcowher.com/

Pittsburgh Steelers NFL News – Steelers Football Blog
If you want to follow the Pittsburgh Steelers through the NFL schedule, you simply cannot afford to be without a ...
www.steelerslocker.com/

blogs
Information Needs in Today’s World

- **Weather**: Pittsburgh weather
  - 17°F | °C
  - Mon: Snow Showers
  - Tue: Rain
  - Wed: Mostly sunny
  - Thu: Mostly cloudy
  - Detailed forecast: The Weather Channel - Weather Underground - AccuWeather

- **Movies**: Pittsburgh movies
  - **The Tourist**: 1hr 43min - Rated PG-13 - Drama/Suspense/Thriller
  - **Harry Potter and the Deathly Hallows - Part 1**: 2hr 30min - Rated PG-13 - SciFi/Fantasy/Action/Adventure
  - **The Chronicles of Narnia: The Voyage of the Dawn Treader**: 1hr 52min - Rated PG - Action/Adventure/SciFi
  - More movies

- **Stock Information**: Yahoo! Inc. (NASDAQ)
  - YHOO
  - Google Finance
  - Yahoo Finance
  - MSN Money
  - Daily Finance
  - CNN Money
  - Reuters
  - 16.66 +0.03 (0.18%) Dec 15 10:25am ET
  - Open: 16.61
  - Volume: 1,741,609
  - High: 16.69
  - Low: 16.53
  - Avg Vol: 18,361,000
  - Mkt Cap: 21.72B
  - Disclaimer
Information Needs in Today’s World

- Flight information
- Social media updates
- Text-entry calculations
- Translation

Example:

- **Flight Information**:
  - United Flight 543
  - Status from Boston (BOS) to Chicago (ORD)
  - Estimated departure: 5:25 PM, Estimated arrival: 7:17 PM

- **Social Media Updates**:
  - @BetheChangeInc: President Obama creates a “WH Council for Community Solutions,” with lots of great thinkers and service friends http ...
  - ArtServe Michigan: Congratulations to Peter White Public Library and West Bloomfield Township Public Library on being honored by First Lady Michelle Obama for their ...
  - @meloobusiness: Obama brainstorming with 20 CEOs - Obama brainstorming with 20 CEOs - meloo.biz

- **Text-Entry Calculations**:
  - 1 mile in centimeters
    - 1 mile = 160,934.4 centimeters
  - Translate "nice to meet you" to Spanish
    - Nice to meet you - un placer conocerte

- **Translation**:
  - Translate.google.com
Information Needs in Today’s World

• different information needs are associated with:
  ‣ different retrievable items (representations)
  ‣ different definitions of relevance
  ‣ different information-seeking behavior

• they cannot be supported by a single search engine (in the traditional sense)

• the trend is towards specialization and integration
  ‣ highly specialized services brought together within a single search interface
Aggregated Search

- **maps**
  - Pittsburgh, PA maps.google.com
  - City of Pittsburgh, Pennsylvania - Pghgov.com
  - Images for pittsburgh

- **web**
  - www.city.pittsburgh.pa.us/
  - Pittsburgh - Wikipedia, the free encyclopedia
  - Books for pittsburgh

- **images**

- **books**
Aggregated Search

- aggregated search: providing users with integrated access to multiple specialized search services within a single interface
Background
Aggregated Search on the Web

- **vertical**: a search service that focuses on a particular domain or type of media

**Portal Interface**

- news
- books
- local
- maps
- images

“pittsburgh”
Why Aggregate Information?

- A user may not know that a vertical has relevant content (e.g., news)
- A user may want results from multiple verticals at once (e.g., planning a trip)
Task Decomposition

- news
- books
- local
- maps
- images

"pittsburgh"

portal interface
Task Decomposition

- **vertical selection**: predicting **which** verticals, if any, are relevant to the query

"pittsburgh"
Task Decomposition

- vertical results presentation: predicting where in the web results to present the vertical results

"pittsburgh"
Outline

- **vertical selection**
  (Arguello *et. al.* SIGIR 2009; Arguello *et. al.*, SIGIR 2010)

- **aggregated search evaluation**
  (Arguello *et al.* ECIR 2011)
Vertical Selection

- predicting **which** vertical(s), **if any**, are relevant to the query

"pittsburgh"

portal interface
Prior Research

federated search

- automatically searching across multiple distributed collections (of full-text documents)
Prior Research

federated search

- resource selection: predicting which collections to search

"pittsburgh"
Prior Research

**federated search**

- results merging: combining their results into a single merged ranking
Prior Research
federated search: unsupervised resource selection

- resource relevance as a function of sample relevance
Prior Research
federated search: unsupervised resource selection

1. combine cross-collection samples within a single index
Prior Research
federated search: unsupervised resource selection

2. conduct a retrieval to predict a set of relevant samples

query (q) → sample index → $R^s_q$
3. select resources as a function of sample relevance
Prior Research
federated search: limitations

- assume document type and retrieval algorithm homogeneity across resources
- derive evidence exclusively from collection content
Sources of Evidence for Vertical Selection

“pittsburgh hotels”

query

content

query-logs
Sources of Evidence for Vertical Selection

“pittsburgh pics”

query

web

content

query-logs

local

news

images

travel
Supervised Vertical Selection

- **machine learning approach**: predict vertical relevance as a function of a set of features
- learn a different model for each vertical
- **training data**: a set of queries with (positive and negative) relevance labels for each candidate vertical
Features

• vertical corpus features
  ‣ similarity between the query and sampled documents

• vertical query-log features
  ‣ similarity between the query and vertical query-traffic

• query features (vertical independent)
  ‣ the query’s topical category (e.g., travel-related)
  ‣ presence of a particular term (e.g., “pittsburgh pics”)
  ‣ geographic named entity types (e.g., city name)
Evaluation Methodology
Task Formulation
single vertical prediction

- given a query, predict a single relevant vertical or predict that no vertical is relevant
Task Formulation
single vertical prediction

- predict a single relevant vertical
Task Formulation
single vertical prediction

- predict that no vertical is relevant
Supervised Classification Framework

- $n + 1$ logistic regression models
- Use highest confidence prediction
- Default to no vertical if highest confidence is below threshold
Evaluation Metric
single-vertical precision

• percentage of queries for which we make a correct prediction
  ‣ correctly predict a vertical that is relevant
  ‣ correctly predict that no vertical is relevant
Queries

• about 25,000 queries drawn randomly from commercial Web traffic
• human annotators assigned between 0-6 relevant verticals per query
• 70% of queries assigned either one relevant vertical or none
Single-Evidence Baselines

- **redde**: content-based resource selection method
- **clarity**: query difficulty measure
- **qlog**: similarity to vertical query-traffic
- **soft.redde**: redde variant
- **no.rel**: always predict no vertical relevant
Experimental Results
Results
single-vertical precision

<table>
<thead>
<tr>
<th>Method</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>clarity</td>
<td>0.254</td>
</tr>
<tr>
<td>no.rel</td>
<td>0.263</td>
</tr>
<tr>
<td>soft.redde</td>
<td>0.324</td>
</tr>
<tr>
<td>redde</td>
<td>0.336</td>
</tr>
<tr>
<td>qlog</td>
<td>0.368</td>
</tr>
<tr>
<td>classification</td>
<td>0.583</td>
</tr>
</tbody>
</table>

▲ statistically significant improvement in performance (p < 0.05) compared to all worse-performing methods
58% improvement over best single-evidence baseline

Results

single-vertical precision

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>clarity</td>
<td>0.254</td>
</tr>
<tr>
<td>no.rel</td>
<td>0.263</td>
</tr>
<tr>
<td>soft.redde</td>
<td>0.324</td>
</tr>
<tr>
<td>redde</td>
<td>0.336</td>
</tr>
<tr>
<td>qlog</td>
<td>0.368</td>
</tr>
<tr>
<td>classification</td>
<td>0.583</td>
</tr>
</tbody>
</table>
Discussion

- is evidence integration helpful?

<table>
<thead>
<tr>
<th>Feature</th>
<th>Score</th>
<th>Decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.583</td>
<td></td>
</tr>
<tr>
<td>no.querylog</td>
<td>0.583</td>
<td>0.03%</td>
</tr>
<tr>
<td>no.boolean</td>
<td>0.583</td>
<td>-0.03%</td>
</tr>
<tr>
<td>no.clarity</td>
<td>0.582</td>
<td>-0.10%</td>
</tr>
<tr>
<td>no.geographical</td>
<td>0.577▼</td>
<td>-1.01%</td>
</tr>
<tr>
<td>no.redde</td>
<td>0.568▼</td>
<td>-2.60%</td>
</tr>
<tr>
<td>no.soft.redde</td>
<td>0.567▼</td>
<td>-2.67%</td>
</tr>
<tr>
<td>no.category</td>
<td>0.552▼</td>
<td>-5.33%</td>
</tr>
</tbody>
</table>

▼ statistically significant decrease in performance (p < 0.05) compared to the model that uses all features.
Discussion

• is evidence integration helpful?

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.583</td>
<td></td>
</tr>
<tr>
<td>no.querylog</td>
<td>0.583</td>
<td>0.03%</td>
</tr>
<tr>
<td>no.boolean</td>
<td>0.583</td>
<td>-0.03%</td>
</tr>
<tr>
<td>no.clarity</td>
<td>0.582</td>
<td>-0.10%</td>
</tr>
<tr>
<td>no.geographical</td>
<td>0.577▼</td>
<td>-1.01%</td>
</tr>
<tr>
<td>no.redde</td>
<td>0.568▼</td>
<td>-2.60%</td>
</tr>
<tr>
<td>no.soft.redde</td>
<td>0.567▼</td>
<td>-2.67%</td>
</tr>
<tr>
<td>no.category</td>
<td>0.552▼</td>
<td>-5.33%</td>
</tr>
</tbody>
</table>

- corpus
- query-log
- query

multiple types of evidence contribute to performance
Discussion

• is evidence integration helpful?

<table>
<thead>
<tr>
<th>Source</th>
<th>Score</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.583</td>
<td></td>
</tr>
<tr>
<td>no.querylog</td>
<td>0.583</td>
<td>0.03%</td>
</tr>
<tr>
<td>no.boolean</td>
<td>0.583</td>
<td>-0.03%</td>
</tr>
<tr>
<td>no.clarity</td>
<td>0.582</td>
<td>-0.10%</td>
</tr>
<tr>
<td>no.geographical</td>
<td>0.577▼</td>
<td>-1.01%</td>
</tr>
<tr>
<td>no.redde</td>
<td>0.568▼</td>
<td>-2.60%</td>
</tr>
<tr>
<td>no.soft.redde</td>
<td>0.567▼</td>
<td>-2.67%</td>
</tr>
<tr>
<td>no.category</td>
<td>0.552▼</td>
<td>-5.33%</td>
</tr>
</tbody>
</table>

most predictive source of evidence requires human supervision
Discussion

• per-vertical performance

- travel 0.842
- health 0.788
- music 0.772
- games 0.771
- autos 0.730
- sports 0.726
- tv 0.716
- movies 0.688
- finance 0.655
- local 0.619
- jobs 0.570
- shopping 0.563
- images 0.483
- video 0.459
- news 0.456
- reference 0.348
- maps 0.000
- directory 0.000
Discussion

• topically focused

- travel 0.842
- health 0.788
- music 0.772
- games 0.771
- autos 0.730
- sports 0.726
- tv 0.716
- movies 0.688
- finance 0.655
- local 0.619
- jobs 0.570
- shopping 0.563
- images 0.483
- video 0.459
- news 0.456
- reference 0.348
- maps 0.000
- directory 0.000
Discussion

- topically diverse

  travel 0.842
  health 0.788
  music 0.772
  games 0.771
  autos 0.730
  sports 0.726
  tv 0.716
  movies 0.688
  finance 0.655
  local 0.619
  jobs 0.570
  shopping 0.563
  images 0.483
  video 0.459
  news 0.456
  reference 0.348
  maps 0.000
  directory 0.000
Discussion

- text-impoverished

- travel 0.842
- health 0.788
- music 0.772
- games 0.771
- autos 0.730
- sports 0.726
- tv 0.716
- movies 0.688
- finance 0.655
- local 0.619
- jobs 0.570
- shopping 0.563
- images 0.483
- video 0.459
- news 0.456
- reference 0.348
- maps 0.000
- directory 0.000
Discussion

• highly dynamic content and user interests

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>travel</td>
<td>0.842</td>
</tr>
<tr>
<td>health</td>
<td>0.788</td>
</tr>
<tr>
<td>music</td>
<td>0.772</td>
</tr>
<tr>
<td>games</td>
<td>0.771</td>
</tr>
<tr>
<td>autos</td>
<td>0.730</td>
</tr>
<tr>
<td>sports</td>
<td>0.726</td>
</tr>
<tr>
<td>tv</td>
<td>0.716</td>
</tr>
<tr>
<td>movies</td>
<td>0.688</td>
</tr>
<tr>
<td>finance</td>
<td>0.655</td>
</tr>
<tr>
<td>local</td>
<td>0.619</td>
</tr>
<tr>
<td>jobs</td>
<td>0.570</td>
</tr>
<tr>
<td>shopping</td>
<td>0.563</td>
</tr>
<tr>
<td>images</td>
<td>0.483</td>
</tr>
<tr>
<td>video</td>
<td>0.459</td>
</tr>
<tr>
<td>news</td>
<td>0.456</td>
</tr>
<tr>
<td>reference</td>
<td>0.348</td>
</tr>
<tr>
<td>maps</td>
<td>0.000</td>
</tr>
<tr>
<td>directory</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Milestones

- traditional resource selection methods not well-suited for this environment
  - derive evidence exclusively from collection content
- a machine learning approach performs better
- multiple types of evidence contribute to performance
- the most predictive type of evidence (the query category) requires human supervision
• learn a model for a new vertical using only existing vertical training data
Outline

• vertical selection
  (Arguello et. al. SIGIR 2009; Arguello et. al., SIGIR 2010)

• aggregated search evaluation
  (Arguello et al. ECIR 2011)
Vertical Results Presentation

- predicting where in the web results to present the vertical results

“pittsburgh”
End-To-End Output

Search for "washington dc"

**Maps**
- Washington D.C., DC
- DC.gov

**Web**
- News for washington dc
- About Washington DC

**Images**
- Images for washington dc
How good is this presentation?

**Maps**

Washington D.C., DC  
maps.google.com  

**Web**

DC.gov  
The official web site of the Government of the District of Columbia. Includes news, web links, and information about the city and about local government ...  
www.dc.gov/ - Similar

**News**

News for washington dc  

Metro Detroiters to rally in Washington DC to support Egyptian ...  
16 hours ago  
By Niraj Warikoo  
About 250 Arab-Americans are leaving tonight on five buses from Michigan to rally Saturday in Washington DC in front of Egypt's embassy in ...  
Detroit Free Press - 15 related articles

Super Bowl XLV: Ice injures workers; weather messes with Texas  
Washington Post (blog) - 5535 related articles

CoStar Sells Washington, DC Building in Sign of Commercial Rebound  
Wall Street Journal (blog) - 16 related articles - Shared by 50+

About Washington DC  

Originally founded on July 16, 1790, Washington, DC is unique among American ...  
washington.org/about-washington-dc - Cached - Similar

**Images**

Images for washington dc - Report images
Is this one better?

News for washington dc

Metro Detroiters to rally in Washington DC to support Egyptian ...
By Niraj Warikoo About 250 Arab-Americans are leaving tonight on five buses from Michigan to rally Saturday in Washington DC in front of Egypt's embassy in ...
Detroit Free Press - 15 related articles
Super Bowl XLV: Ice injures workers; weather messes with Texas
Washington Post (blog) - 5535 related articles
CoStar Sells Washington, DC Building in Sign of Commercial Rebound
Wall Street Journal (blog) - 16 related articles - Shared by 50+

DC.gov
The official web site of the Government of the District of Columbia. Includes news, web links, and information about the city and about local government ...
www.dc.gov/ - Similar

Washington D.C., DC maps.google.com

About Washington DC
Officially founded on July 16, 1790, Washington, DC is unique among American ...
washington.org/about-washington-dc - Cached - Similar

Images for washington dc - Report images

images

news

web

maps
What about this one?

- Weather for Washington D.C., DC
  - Current: Drizzle
  - Wind: NE at 6 mph
  - Humidity: 93%
  - Sat: 40°F | 34°F
  - Sun: 45°F | 36°F
  - Mon: 50°F | 33°F
  - Tue: 38°F | 26°F

- DC.gov
  - The official web site of the Government of the District of Columbia. Includes news, web links, and information about the city and about local government...
  - www.dc.gov / Similar

- Washington D.C., DC
  - Officially founded on July 16, 1790, Washington, DC is unique among American cities...
  - washington.org/about-washington-dc - Cached - Similar

- Images for washington dc
1. define a set of layout constraints
   ‣ will define the set of all possible presentations

2. define an evaluation metric that can measure the quality of any possible presentation

3. validate the metric by ensuring that it correlates with human preferences
Layout Constraints and Task Formulation
Layout Constraints and Task Formulation

washington dc

Washington.org - Official Tourism Site of Washington, DC
Get the latest information on events, attractions, and create your own experience during your visit to Washington, DC.
100 Free (or Nearly Free) Things to ... - Attractions & Tours - Events Calendar - Maps
washington.org/ - Cached - Similar

About Washington DC
Officially founded on July 16, 1790, Washington, DC is unique among American ...
washington.org/about-washington-dc - Cached - Similar

Experience DC - Washington DC
Experience DC By Selecting a Category. ITINERARIES IDEAS. Washington, DC ...
washington.org/visiting/experience-dc - Cached - Similar

100 Free (or Nearly Free) Things to Do in DC - Washington DC
Find out what you can do in Washington, DC for free (or pretty close to that ...
washington.org/visiting/browse-dc/.../100-free-things-to-do -Cached - Similar

Washington, D.C. - Wikipedia, the free encyclopedia
Washington, D.C. formally the District of Columbia and commonly referred to as Washington, the District, or simply D.C., is the capital of the United States ...
History of Washington, D.C. - Geography - Demographics of Washington, DC
en.wikipedia.org/wiki/Washington,_D.C. - Cached - Similar

Tourist Information for Washington, DC - the District – Tourist ...
Tourist information and travel tips for Washington, DC visitors.
www.thedistrict.com/ - Cached - Similar

Washington DC Hotels, Attractions, Real Estate, Restaurants | City ...
Washington DC Travel & Tourism Guide specializing in Hotels, Attractions, Restaurants, Real Estate, Nightlife & Local Business Yellow Page Listings.
www.washingtondc.com/ - Cached - Similar
Layout Constraints and Task Formulation

web block 1

web block 2

web block 3

vertical slot 1

vertical slot 2

vertical slot 3
Layout Constraints and Task Formulation

image block

books block

weather block

news block

map block

web block 1

web block 2

imaginary end-of-results block
• formulate vertical results presentation as block ranking

• web blocks are always presented and maintain their natural order

• suppressed vertical blocks are effectively tied
Aggregated Search Evaluation Objectives

1. define a set of layout constraints
   ▸ will define the set of all possible presentations

2. define an evaluation metric that can measure the quality of any possible presentation

3. validate the metric by ensuring that it correlates with human preferences
How good is a presentation?

- images
- web 1
- weather
- web 2

- problem: a query with 10 blocks has 3,628,800 possible rankings/presentations
• given a query, collect human judgements on **individual blocks** and use these to derive a “ground truth” presentation

• evaluate alternative presentations based on their distance to this “ground truth”
Metric-Based Evaluation Approach

1. compose web and vertical blocks for the query

```
> washington dc

image block

books block

weather block

news block
```

```
web block 1

web block 2
```

```
map block
```
Metric-Based Evaluation Approach

2. collect preference judgements on all block pairs
   ▪ left is better, right is better, both are bad?
Metric-Based Evaluation Approach

2. collect preference judgements on all block pairs
   - left is better, right is better, both are bad?
3. use these pairwise preference judgements to construct a ground truth or reference presentation

schulze voting method (Schulze, 2010)
Metric-Based Evaluation Approach

4. evaluate a presentation based on its distance to the reference

\[ K(\sigma^*, \sigma_1) \]

generalized kendall’s tau
(Kumar et. al., 2010)
Metric-Based Evaluation Approach

1. compose web and vertical blocks for the query

2. collect redundant preference judgements on every block-pair

3. derive reference presentation
   schulze voting method

4. evaluate a presentation based on its distance to the reference
   generalized kendall’s tau
Aggregated Search Evaluation Objectives

1. define a set of layout constraints
   ‣ will define the set of all possible presentations

2. define an evaluation metric that can measure the quality of any possible presentation

3. validate the metric by ensuring that it correlates with human preferences
Empirical Metric Validation

- does the metric agree with human preferences on pairs of full presentations?
Empirical Metric Validation

- agreement: $\mathcal{K}(\sigma_1, \sigma^*) < \mathcal{K}(\sigma_2, \sigma^*)$
Methodology

- show assessors pairs of presentations and ask which they prefer
- sample presentation-pairs from particular regions of the metric-space
Materials

• 72 queries selected manually

• 13 verticals constructed from freely available Web APIs (Google, eBay, Yahoo, YouTube)

• human judgements also collected using Amazon’s Mechanical Turk

• 72 queries × 6 bin-combs. × 4 pairs per bin-comb. × 4 judges per pair = 6,912 judgements
Empirical Metric Validation Results
Inter-assessor Agreement
block-pairs

\[ \kappa_{\text{fleiss}} = 0.660 \]
(substantial)
Inter-assessor Agreement
presentation-pairs

\( \sigma_1 \)

images
web 1
weather
web 2
map

vs.

\( \sigma_2 \)

books
news
web 1
map
weather
web 2
Inter-assessor Agreement
presentation-pairs

<table>
<thead>
<tr>
<th>bin</th>
<th>$\kappa_{\text{fleiss}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-H</td>
<td>0.066</td>
</tr>
<tr>
<td>H-M</td>
<td>0.290</td>
</tr>
<tr>
<td>H-L</td>
<td>0.303</td>
</tr>
<tr>
<td>M-M</td>
<td>0.216</td>
</tr>
<tr>
<td>M-L</td>
<td>0.179</td>
</tr>
<tr>
<td>L-L</td>
<td>0.237</td>
</tr>
<tr>
<td>ALL</td>
<td>0.237</td>
</tr>
</tbody>
</table>

- fair agreement
- lower than agreement on block-pair judgements
Inter-assessor Agreement
presentation-pairs

<table>
<thead>
<tr>
<th>bin</th>
<th>$\kappa_{\text{fleiss}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-H</td>
<td>0.066</td>
</tr>
<tr>
<td>H-M</td>
<td>0.290</td>
</tr>
<tr>
<td>H-L</td>
<td>0.303</td>
</tr>
<tr>
<td>M-M</td>
<td>0.216</td>
</tr>
<tr>
<td>M-L</td>
<td>0.179</td>
</tr>
<tr>
<td>L-L</td>
<td>0.237</td>
</tr>
<tr>
<td>ALL</td>
<td>0.237</td>
</tr>
</tbody>
</table>

- very low agreement on H-H pairs
- these pairs were almost identical in terms of the top-ranked blocks
Inter-assessor Agreement
presentation-pairs

<table>
<thead>
<tr>
<th>bin</th>
<th>$\kappa_{\text{fleiss}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-H</td>
<td>0.066</td>
</tr>
<tr>
<td>H-M</td>
<td>0.290</td>
</tr>
<tr>
<td>H-L</td>
<td>0.303</td>
</tr>
<tr>
<td>M-M</td>
<td>0.216</td>
</tr>
<tr>
<td>M-L</td>
<td>0.179</td>
</tr>
<tr>
<td>L-L</td>
<td>0.237</td>
</tr>
<tr>
<td>ALL</td>
<td>0.237</td>
</tr>
</tbody>
</table>

• higher agreement on H-M and H-L pairs
• assessors agreed on pairs where one presentation was close to the reference and the other was far
## Metric Agreement  
metric vs. majority preference

<table>
<thead>
<tr>
<th>bin</th>
<th>% agree</th>
<th>bin</th>
<th>% agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-H</td>
<td>60.37 &quot;</td>
<td>H-H</td>
<td>65.96 &quot;</td>
</tr>
<tr>
<td>H-M</td>
<td>81.90 &quot;</td>
<td>H-M</td>
<td>87.37 &quot;</td>
</tr>
<tr>
<td>H-L</td>
<td>84.31 &quot;</td>
<td>H-L</td>
<td>91.75 &quot;</td>
</tr>
<tr>
<td>M-M</td>
<td>57.61 &quot;</td>
<td>M-M</td>
<td>58.67</td>
</tr>
<tr>
<td>M-L</td>
<td>50.80</td>
<td>M-L</td>
<td>54.93</td>
</tr>
<tr>
<td>L-L</td>
<td>63.37 &quot;</td>
<td>L-L</td>
<td>63.64 &quot;</td>
</tr>
<tr>
<td>ALL</td>
<td>67.07 &quot;</td>
<td>ALL</td>
<td>72.51 &quot;</td>
</tr>
</tbody>
</table>

▲ = statistically significant agreement based on a sign-test. The null hypothesis is that the metric selects the preferred presentation randomly with equal probability.
### Metric Agreement

#### metric vs. majority preference

<table>
<thead>
<tr>
<th>bin</th>
<th>pairs</th>
<th>% agree</th>
<th>3/4 majority or better</th>
<th>4/4 majority</th>
<th>% agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-H</td>
<td>164</td>
<td>60.37 ▲</td>
<td></td>
<td>H-H</td>
<td>65.96 ▲</td>
</tr>
<tr>
<td>H-M</td>
<td>210</td>
<td>81.90 ▲</td>
<td></td>
<td>H-M</td>
<td>87.37 ▲</td>
</tr>
<tr>
<td>H-L</td>
<td>204</td>
<td>84.31 ▲</td>
<td></td>
<td>H-L</td>
<td>91.75 ▲</td>
</tr>
<tr>
<td>M-M</td>
<td>184</td>
<td>57.61 ▲</td>
<td></td>
<td>M-M</td>
<td>58.67</td>
</tr>
<tr>
<td>M-L</td>
<td>187</td>
<td>50.80</td>
<td></td>
<td>M-L</td>
<td>54.93</td>
</tr>
<tr>
<td>L-L</td>
<td>202</td>
<td>63.37 ▲</td>
<td></td>
<td>L-L</td>
<td>63.64 ▲</td>
</tr>
<tr>
<td>ALL</td>
<td>1151</td>
<td>67.07 ▲</td>
<td></td>
<td>ALL</td>
<td>72.51 ▲</td>
</tr>
</tbody>
</table>

- when assessors agreed with each other, the metric agreed with the assessors
Metric Agreement
metric vs. majority preference

3/4 majority or better

<table>
<thead>
<tr>
<th>bin</th>
<th>pairs</th>
<th>% agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-H</td>
<td>164</td>
<td>60.37</td>
</tr>
<tr>
<td>H-M</td>
<td>210</td>
<td>81.90</td>
</tr>
<tr>
<td>H-L</td>
<td>204</td>
<td>84.31</td>
</tr>
<tr>
<td>M-M</td>
<td>184</td>
<td>57.61</td>
</tr>
<tr>
<td>M-L</td>
<td>187</td>
<td>50.80</td>
</tr>
<tr>
<td>L-L</td>
<td>202</td>
<td>63.37</td>
</tr>
<tr>
<td>ALL</td>
<td>1151</td>
<td>67.07</td>
</tr>
</tbody>
</table>

4/4 majority

<table>
<thead>
<tr>
<th>bin</th>
<th>pairs</th>
<th>% agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-H</td>
<td>47</td>
<td>65.96</td>
</tr>
<tr>
<td>H-M</td>
<td>95</td>
<td>87.37</td>
</tr>
<tr>
<td>H-L</td>
<td>97</td>
<td>91.75</td>
</tr>
<tr>
<td>M-M</td>
<td>75</td>
<td>58.67</td>
</tr>
<tr>
<td>M-L</td>
<td>71</td>
<td>54.93</td>
</tr>
<tr>
<td>L-L</td>
<td>77</td>
<td>63.64</td>
</tr>
<tr>
<td>ALL</td>
<td>462</td>
<td>72.51</td>
</tr>
</tbody>
</table>

- on average, the reference presentation was good
# Metric Agreement

metric vs. majority preference

## 3/4 majority or better

<table>
<thead>
<tr>
<th>bin</th>
<th>pairs</th>
<th>% agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-H</td>
<td>164</td>
<td>60.37</td>
</tr>
<tr>
<td>H-M</td>
<td>210</td>
<td>81.90</td>
</tr>
<tr>
<td>H-L</td>
<td>204</td>
<td>84.31</td>
</tr>
<tr>
<td>M-M</td>
<td>184</td>
<td>57.61</td>
</tr>
<tr>
<td>M-L</td>
<td>187</td>
<td>50.80</td>
</tr>
<tr>
<td>L-L</td>
<td>202</td>
<td>63.37</td>
</tr>
<tr>
<td>ALL</td>
<td>1151</td>
<td>67.07</td>
</tr>
</tbody>
</table>

## 4/4 majority

<table>
<thead>
<tr>
<th>bin</th>
<th>pairs</th>
<th>% agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-H</td>
<td>47</td>
<td>65.96</td>
</tr>
<tr>
<td>H-M</td>
<td>95</td>
<td>87.37</td>
</tr>
<tr>
<td>H-L</td>
<td>97</td>
<td>91.75</td>
</tr>
<tr>
<td>M-M</td>
<td>75</td>
<td>58.67</td>
</tr>
<tr>
<td>M-L</td>
<td>71</td>
<td>54.93</td>
</tr>
<tr>
<td>L-L</td>
<td>77</td>
<td>63.64</td>
</tr>
<tr>
<td>ALL</td>
<td>462</td>
<td>72.51</td>
</tr>
</tbody>
</table>

- less correlated for presentations far from the reference (low quality presentations)
Discussion
block-pair assessment interface

- in some cases, assessors favored a particular vertical only in the context of a full presentation
- suggests interactions between cross-vertical results
Discussion
block-pair assessment interface

- add context to the block-pair assessment interface?

\[ \sigma_1 \text{ weather} \quad \text{vs.} \quad \sigma_2 \text{ images} \]

\[ \text{images for washington dc} \quad \text{- Report images} \]

\[ \text{Weather for Washington D.C., DC - Add to iGoogle} \]

\[ 36^\circ F \quad | \quad 1^\circ C \]

\[ \text{Current: Drizzle} \]

\[ \text{Wind: NE at 6 mph} \]

\[ \text{Humidity: 93\%} \]

\[ 40^\circ F \quad | \quad 45^\circ F \quad | \quad 45^\circ F \quad | \quad 50^\circ F \quad | \quad 33^\circ F \quad | \quad 38^\circ F \quad | \quad 26^\circ F \]

\[ \text{Washington.org - Official Tourism Site of Washington, DC} \]

\[ \text{Get the latest information on events, attractions, and create your own experience during your visit to Washington, DC.} \]

\[ \text{100 Free (or Nearly Free) Things to Do in DC - Washington DC} \]

\[ \text{Experience DC - Washington DC} \]

\[ \text{Experience DC By Selecting a Category: ITINERARIES IDEAS. Washington, DC} \]

\[ \text{100 Free (or Nearly Free) Things to Do in DC - Washington DC} \]

\[ \text{Washington.org - Official Tourism Site of Washington, DC} \]

\[ \text{Get the latest information on events, attractions, and create your own experience during your visit to Washington, DC.} \]

\[ \text{100 Free (or Nearly Free) Things to Do in DC - Washington DC} \]
Discussion

crowd-sourcing relevance judgements

- 20% of the assessors did 80% of the work
  - “traps” can be used to detect careless assessors
  - determines reliability for 80% of the work
  - determining reliability for the other 20% is difficult
- increase fix costs: qualification tests
- lots of interesting research on deriving judgements (and learning models) from multiple noisy assessors
- TREC 2011 will offer a crowd-sourcing track
Milestones
metric-based evaluation approach

- with fewer than 100 judgements per query, we can evaluate any possible presentation
  - portable test collection
  - does not require live system with (many) users
  - facilitates model learning (my current research)
- correlates with human preferences (when humans agree with each other)
- general: bottom-up construction of "ground truth" + rank-based distance metric
Concluding Remarks
Information Needs in Today’s World

- IR systems will continuously support a wider-range of information needs
- different information needs require customized solutions
- the trend is towards specialization and integration
- aggregated search provides integrated access to specialized systems within a single search interface
  - predicting which back-ends to display
  - predicting how to combine their results
Aggregated Search
current challenges and opportunities

• dynamic content and dynamic user interests (e.g. news)
  ‣ implicit user feedback
  ‣ generate training data retrospectively

• improve cohesion between cross-vertical results
  ‣ capitalize on highly-confident verticals

• incorporate user-context and session-level evidence into aggregated search decisions
Aggregated Search
extensions

• aggregation in other environments
  ‣ mobile search, library search, enterprise search, personal information management, social network search

• search assistance and customized interactions
  ‣ search tools can be viewed as verticals
  ‣ predict when and how to make search tools available to a user
Aggregated Search
extensions: search assistance
Aggregated Search
extensions: customized interactions