Beyond Search: Statistical
Topic Models for Text Analysis

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign
http://www.cs.uiuc.edu/homes/czhai
Search is a means to the end of finishing a task.

Search → Information Synthesis & Analysis → Task Completion

Decision Making
Learning...

Information Interpretation

Information Synthesis

Potentially iterate...

Multiple Searches
Example Task 1: Comparing News Articles

Vietnam War
- CNN
- Before 9/11

Afghan War
- Fox
- During Iraq war

Iraq War
- BBC
- Post-Iraq war

United nations
- ... (Vietnam)
- ... (Afghan)
- ... (Iraq)

Death of people
- ... (Vietnam)
- ... (Afghan)
- ... (Iraq)

What’s in common? What’s unique?

<table>
<thead>
<tr>
<th>Common Themes</th>
<th>“Vietnam” specific</th>
<th>“Afghan” specific</th>
<th>“Iraq” specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>United nations</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Death of people</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Example Task 2: Compare Customer Reviews

<table>
<thead>
<tr>
<th>Common Themes</th>
<th>“IBM” specific</th>
<th>“APPLE” specific</th>
<th>“DELL” specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>Battery Life</td>
<td>….</td>
<td>…</td>
<td>….</td>
</tr>
<tr>
<td>Hard disk</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Speed</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

Which laptop to buy?
Example Task 3: Identify Emerging Research Topics

What’s hot in database research?
Example Task 4: Analysis of Topic Diffusion

One Week Later

How did a discussion of a topic in blogs spread?
Sample Task 5: Opinion Analysis on Blog Articles

What did people like/dislike about “Da Vinci Code”? 

- Tom Hanks, who is my favorite movie star act the leading role.
- a good book to past time.
- protesting... will lose your faith by watching the movie.
- ... so sick of people making such a big deal about a fiction book

Query=“Da Vinci Code”
Questions

• Can we model all these analysis problems in a general way?  
• Can we solve these problems with a unified approach?  
• How can we bring users into the loop?

Yes!  
Yes!  
Yes!

Solutions: Statistical Topic Models
Rest of the talk

• Overview of Statistical Topic Models
• Contextual Probabilistic Latent Semantic Analysis (CPLSA)
• Text Analysis Enabled by CPLSA
• From Search Engines to Analysis Engines
What is a Statistical LM?

• A probability distribution over word sequences
  – $p(\text{"Today is Wednesday"}) \approx 0.001$
  – $p(\text{"Today Wednesday is"}) \approx 0.0000000000001$
  – $p(\text{"The eigenvalue is positive"}) \approx 0.00001$

• Context/topic dependent!

• Can also be regarded as a probabilistic mechanism for “generating” text, thus also called a “generative” model
The Simplest Language Model
(Unigram Model)

• Generate a piece of text by generating each word **independently**
• Thus, \( p(w_1 \ w_2 \ \ldots \ w_n) = p(w_1)p(w_2)\ldots p(w_n) \)
• Parameters: \( \{p(w_i)\} \quad p(w_1)+\ldots+p(w_N)=1 \) (\( N \) is voc. size)
• Essentially a multinomial distribution over words
• A piece of text can be regarded as a sample drawn according to this word distribution
Text Generation with Unigram LM

(Unigram) Language Model \( \theta \)

\( p(w|\theta) \)

Sampling

Document \( d \)

Topic 1: Text mining

- text 0.2
- mining 0.1
- association 0.01
- clustering 0.02
- food 0.00001

Topic 2: Health

- food 0.25
- nutrition 0.1
- healthy 0.05
- diet 0.02

Given \( \theta \), \( p(d|\theta) \) varies according to \( d \)

Text mining paper

Food nutrition paper
Estimation of Unigram LM

(Unigram) Language Model $\theta$  
$p(w|\theta)=?$

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>text</td>
<td>10/100</td>
</tr>
<tr>
<td>mining</td>
<td>5/100</td>
</tr>
<tr>
<td>association</td>
<td>3/100</td>
</tr>
<tr>
<td>database</td>
<td>3/100</td>
</tr>
<tr>
<td>query</td>
<td>1/100</td>
</tr>
</tbody>
</table>

Total #words = 100

language model as topic representation?
Language Model as Text Representation: Early Work

• 1961: H. P. Luhn’s early idea of using relative frequency to represent text [Luhn 61]

• 1976: Robertson & Sparck Jones’ BIR model [Robertson & Sparck Jones 76]

• 1989: Wong & Yao’s work on multinomial distribution representation [Wong & Yao 89]

Language Model as Text Representation:  
Two Important Milestones in 1998~1999

- **1998:** Language model for retrieval (i.e., query likelihood scoring [Ponte & Croft 98] (and also independently [Hiemstra & Kraaij 99])

- **1999:** Probabilistic Latent Semantic Analysis (PLSA) [Hofmann 99]

Probabilistic Latent Semantic Analysis (PLSA)

\[ P(w) = \sum_{i=1}^{K} P(z = i)P(w|Topic_i) \]

- **Topic 1**
  - Apple iPod
  - iPod 0.15
  - Nano 0.08
  - Music 0.05
  - Download 0.02
  - Apple 0.01

- **Topic 2**
  - Harry Potter
  - Movie 0.10
  - Harry 0.09
  - Potter 0.05
  - Actress 0.04
  - Music 0.02

Example sentence: I downloaded the music of the movie Harry Potter to my iPod nano.
Parameter Estimation

- Maximizing data likelihood:
  \[ \Lambda^* = \arg \max_{\Lambda} \log(P(Data | Model)) \]
- Parameter Estimation using EM algorithm

Guess the affiliation

Estimate the params

Prior set by users

Pseudo-Counts

I downloaded the music of the movie harry potter to my ipod nano
Context Features of a Document

Weblog Article

- Author
- Author’s Occupation
- Time
- Location
- Source
- Communities
A General View of Context

- Partition of documents
- Any combination of context features (metadata) can define a context

papers written in 1998

papers written by Bruce Croft
Empower PLSA with Context [Mei & Zhai 06]

- Make topics depend on context variables
- Text is generated from a contextualized PLSA model (CPLSA)
- Fitting such a model to text enables a wide range of analysis tasks involving topics and context

Qiaozhu Mei, ChengXiang Zhai, A Mixture Model for Contextual Text Mining, Proceedings of the 2006 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD'06), pages 649-655
Contextual Probabilistic Latent Semantics Analysis

Choose a theme: Criticism of government response to the hurricane primarily consisted of criticism of its response to … The total shut-in oil production from the Gulf of Mexico … approximately 24% of the annual production and the shut-in gas production … Over seventy countries pledged monetary donations or assistance. …

government 0.3 response 0.2…

donate 0.1 relief 0.05 help 0.02 …
city 0.2 new 0.1 orleans 0.05 …

Choose a view

Theme coverages: Texas July 2005 document

Choose a Coverage
### Comparing News Articles

**Iraq War (30 articles) vs. Afghan War (26 articles)**

The common theme indicates that “United Nations” is involved in both wars.

<table>
<thead>
<tr>
<th>Common Theme</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>united nations 0.04</td>
<td>killed month 0.035</td>
<td>( \ldots )</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>month 0.032</td>
<td>( \ldots )</td>
</tr>
<tr>
<td></td>
<td>( \ldots )</td>
<td>deaths 0.023</td>
<td>( \ldots )</td>
</tr>
<tr>
<td></td>
<td>killed</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td></td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
</tbody>
</table>

**Iraq Theme**

<table>
<thead>
<tr>
<th></th>
<th>n 0.03</th>
<th>troops 0.016</th>
<th>( \ldots )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weapons</td>
<td>0.024</td>
<td>hoon 0.015</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>Inspections</td>
<td>0.023</td>
<td>sanches 0.012</td>
<td>( \ldots )</td>
</tr>
<tr>
<td></td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
</tbody>
</table>

**Afghan Theme**

<table>
<thead>
<tr>
<th></th>
<th>Northern alliance 0.04</th>
<th>taleban rumsfeld 0.026</th>
<th>( \ldots )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>kabul 0.03</td>
<td>hotel 0.02</td>
<td>( \ldots )</td>
</tr>
<tr>
<td></td>
<td>taleban 0.025</td>
<td>front 0.012</td>
<td>( \ldots )</td>
</tr>
<tr>
<td></td>
<td>aid 0.02</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
</tbody>
</table>

Collection-specific themes indicate different roles of “United Nations” in the two wars.
Spatiotemporal Patterns in Blog Articles

- **Query** = “Hurricane Katrina”

- **Topics in the results:**

<table>
<thead>
<tr>
<th>Government Response</th>
<th>New Orleans</th>
<th>Oil Price</th>
<th>Praying and Blessing</th>
<th>Aid and Donation</th>
<th>Personal</th>
</tr>
</thead>
<tbody>
<tr>
<td>bush 0.071</td>
<td>city 0.063</td>
<td>price 0.077</td>
<td>god 0.141</td>
<td>donate 0.120</td>
<td>i 0.405</td>
</tr>
<tr>
<td>president 0.061</td>
<td>orleans 0.054</td>
<td>oil 0.064</td>
<td>pray 0.047</td>
<td>relief 0.076</td>
<td>my 0.116</td>
</tr>
<tr>
<td>federal 0.051</td>
<td>new 0.034</td>
<td>gas 0.045</td>
<td>prayer 0.041</td>
<td>red 0.070</td>
<td>me 0.060</td>
</tr>
<tr>
<td>government 0.047</td>
<td>louisiana 0.023</td>
<td>increase 0.020</td>
<td>love 0.030</td>
<td>cross 0.065</td>
<td>am 0.029</td>
</tr>
<tr>
<td>fema 0.047</td>
<td>flood 0.022</td>
<td>product 0.020</td>
<td>life 0.025</td>
<td>help 0.050</td>
<td>think 0.015</td>
</tr>
<tr>
<td>administrate 0.023</td>
<td>evacuate 0.021</td>
<td>fuel 0.018</td>
<td>bless 0.025</td>
<td>victim 0.036</td>
<td>feel 0.012</td>
</tr>
<tr>
<td>response 0.020</td>
<td>storm 0.017</td>
<td>energy 0.018</td>
<td>lord 0.017</td>
<td>organize 0.022</td>
<td>know 0.011</td>
</tr>
<tr>
<td>brown 0.019</td>
<td>resident 0.016</td>
<td>company 0.018</td>
<td>jesus 0.016</td>
<td>effort 0.020</td>
<td>something 0.007</td>
</tr>
<tr>
<td>blame 0.017</td>
<td>center 0.016</td>
<td>market 0.016</td>
<td>will 0.013</td>
<td>fund 0.019</td>
<td>guess 0.007</td>
</tr>
<tr>
<td>governor 0.014</td>
<td>rescue 0.012</td>
<td>gasoline 0.012</td>
<td>faith 0.012</td>
<td>volunteer 0.019</td>
<td>myself 0.006</td>
</tr>
</tbody>
</table>

- **Spatiotemporal patterns**
Theme Life Cycles ("Hurricane Katrina")

(a) Theme life cycles in Texas (Hurricane Katrina)

(b) Theme "New Orleans" over states (Hurricane Katrina)

Oil Price

- price 0.0772
- oil 0.0643
- gas 0.0454
- increase 0.0210
- product 0.0203
- fuel 0.0188
- company 0.0182

New Orleans

- city 0.0634
- orleans 0.0541
- new 0.0342
- louisiana 0.0235
- flood 0.0227
- evacuate 0.0211
- storm 0.0177

...
Theme Snapshots ("Hurricane Katrina")

Week 1: The theme is the strongest along the Gulf of Mexico

Week 2: The discussion moves towards the north and west

Week 3: The theme distributes more uniformly over the states

Week 4: The theme is again strong along the east coast and the Gulf of Mexico

Week 5: The theme fades out in most states
Theme Life Cycles (KDD Papers)

[Diagram showing normalized strength of themes over time, highlighting themes such as biology data, web information, time series, classification, association rules, clustering, business, marketing, customer, model, and rules.]

- Gene: 0.0173
- Expressions: 0.0096
- Probability: 0.0081
- Microarray: 0.0038
- Marketing: 0.0087
- Customer: 0.0086
- Model: 0.0079
- Business: 0.0048
- Rules: 0.0142
- Association: 0.0064
- Support: 0.0053
Theme Evolution Graph: KDD

1999 2000 2001 2002 2003 2004

SVM 0.007
criteria 0.007
classification 0.007
linear 0.005
...

decision 0.006
tree 0.006
classifier 0.005
class 0.005
Bayes 0.005
...

web 0.009
classification 0.007
features 0.006
topic 0.005
...

mixture 0.005
random 0.006
cluster 0.006
clustering 0.005
variables 0.005
...

topic 0.010
mixture 0.008
LDA 0.006
semantic 0.005
...

Classifica-
tion 0.015
text 0.013
unlabeled 0.012
document 0.008
labeled 0.008
learning 0.007
...

Informa-
tion 0.012
web 0.010
social 0.008
retrieval 0.007
distance 0.005
networks 0.004
...

Keynote at SIGIR 2011, July 26, 2011, Beijing, China
### Multi-Faceted Sentiment Summary (query="Da Vinci Code")

<table>
<thead>
<tr>
<th>Facet 1: Movie</th>
<th>Neutral</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>... Ron Howards selection of Tom Hanks to play Robert Langdon.</td>
<td>Tom Hanks stars in the movie, who can be mad at that?</td>
<td>But the movie might get delayed, and even killed off if he loses.</td>
<td></td>
</tr>
<tr>
<td>Directed by: Ron Howard Writing credits: Akiva Goldsman ...</td>
<td>Tom Hanks, who is my favorite movie star act the leading role.</td>
<td>protesting ... will lose your faith by ... watching the movie.</td>
<td></td>
</tr>
<tr>
<td>After watching the movie I went online and some research on ...</td>
<td>Anybody is interested in it?</td>
<td>... so sick of people making such a big deal about a FICTION book and movie.</td>
<td></td>
</tr>
<tr>
<td>Facet 2: Book</td>
<td>I remembered when i first read the book, I finished the book in two days.</td>
<td>Awesome book.</td>
<td>... so sick of people making such a big deal about a FICTION book and movie.</td>
</tr>
<tr>
<td>I'm reading “Da Vinci Code” now. ...</td>
<td>So still a good book to past time.</td>
<td>This controversy book cause lots conflict in west society.</td>
<td></td>
</tr>
</tbody>
</table>
Separate Theme Sentiment Dynamics

“book”

“religious beliefs”

The diagrams show the sentiment dynamics over time for the themes “book” and “religious beliefs.” The x-axis represents time, and the y-axis represents the strength of sentiment. The lines indicate neutral, positive, and negative sentiments.
Event Impact Analysis: IR Research

Theme: retrieval models

term 0.1599
relevance 0.0752
weight 0.0660
feedback 0.0372
independence 0.0311
model 0.0310
frequent 0.0233
probabilistic 0.0188
document 0.0173
...

vector 0.0514
concept 0.0298
extend 0.0297
model 0.0291
space 0.0236
boolean 0.0151
function 0.0123
feedback 0.0077
...

xml 0.0678
email 0.0197
model 0.0191
collect 0.0187
judgment 0.0102
rank 0.0097
subtopic 0.0079
...

probabilist 0.0778
model 0.0432
logic 0.0404
ir 0.0338
boolean 0.0281
algebra 0.0200
estimate 0.0119
weight 0.0111
...

SIGIR papers

Publication of the paper “A language modeling approach to information retrieval”

Starting of the TREC conferences

1992

1998

year

model 0.1687
language 0.0753
estimate 0.0520
parameter 0.0281
distribution 0.0268
probable 0.0205
smooth 0.0198
markov 0.0137
likelihood 0.0059
...

The Database and Information Systems Laboratory
at The University of Illinois at Urbana-Champaign
Core: Scalable Information Management

Many Other Variations

- **Latent Dirichlet Allocation (LDA) [Blei et al. 03]**
  - Impose priors on topic choices and word distributions
  - Make PLSA a generative model

- **Many variants of LDA!**

- **In practice, LDA and PLSA variants tend to work equally well for text analysis** [Lu et al. 11]

---


Other Uses of Topic Models for Text Analysis

- **Topic analysis on social networks** [Mei et al. 08]
- **Opinion Integration** [Lu & Zhai 08]
- **Latent Aspect Rating Analysis** [Wang et al. 10]


Topic Modeling + Social Networks: who work together on what?

Authors writing about the same topic form a community
Separation of 3 research communities: IR, ML, Web
Topic Model for Opinion Integration

How to digest all?

4,773,658 results

190,451 posts

Google Blog Search

Hillary Rodham Clinton

From Wikipedia, the free encyclopedia

(Redirected from Hillary Clinton)

Hillary Diane Rodham Clinton, from New York, and a candidate. She is married to Bill Clinton, the President of the United States from 1993 to 2001. A native of Illinois, Hillary Rodham began her career as a lawyer and as an attorney, she moved to Arkansas in 1975. She was later named the first female partner at Rose L.
Two Kinds of Opinions

**Expert opinions**
- CNET editor’s review
- Wikipedia article
- Well-structured
- Easy to access
- Maybe biased
- Outdated soon

**Ordinary opinions**
- Forum discussions
- Blog articles
- Represent the majority
- Up to date
- Hard to access
- Fragmental

How to benefit from both?
Generate an Integrative Summary

Input

Topic: iPod
Expert review with aspects
Text collection of ordinary opinions, e.g., Weblogs

Output

Design
Battery
Price

Similar opinions

cute... tiny...
last many hrs
could afford it
die out soon
still expensive

Supplementary opinions

..thicker..

Extra Aspects Review Aspects

iTunes 
... easy to use...

warranty 
...better to extend...

Integrated Summary

Keynote at SIGIR 2011, July 26, 2011, Beijing, China
Methods

• Semi-Supervised Probabilistic Latent Semantic Analysis (PLSA)
  – The aspects extracted from expert reviews serve as clues to define a conjugate prior on topics
  – Maximum a Posteriori (MAP) estimation
  – Repeated applications of PLSA to integrate and align opinions in blog articles to expert review
### Results: Product (iPhone)

- **Opinion Integration with review aspects**

<table>
<thead>
<tr>
<th>Review article</th>
<th>Similar opinions</th>
<th>Supplementary opinions</th>
</tr>
</thead>
<tbody>
<tr>
<td>You can make emergency calls, but you can't use any other functions...</td>
<td>N/A</td>
<td>... methods for <strong>unlocking</strong> the iPhone have emerged on the Internet in the past few weeks, although they involve tinkering with the iPhone hardware...</td>
</tr>
<tr>
<td>rated battery life...</td>
<td>Up to <strong>8 Hours of Talk Time</strong>, 6 Hours of Internet Use, <strong>7 Hours of Video</strong> Playback or <strong>24 Hours of Audio Playback</strong></td>
<td>Playing relatively high bitrate VGA H.264 videos, our iPhone lasted almost exactly <strong>9 freaking hours</strong> of continuous playback with cell and WiFi on (but Bluetooth off)...</td>
</tr>
</tbody>
</table>

**Activation**

**Battery**

**Unlock/hack iPhone**

**Confirm the opinions from the review**

**Additional info under real usage**

- **Rated battery life**: 8 hours talk time, 24 hours of music playback, 7 hours of video playback, and 6 hours on Internet use.
Results: Product (iPhone)

- Opinions on extra aspects

<table>
<thead>
<tr>
<th>support</th>
<th>Supplementary opinions on extra aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>You may have heard of iASign—an iPhone Dev Wiki tool that allows you to <strong>activate</strong> your phone without going through the iTunes rigamarole.</td>
</tr>
<tr>
<td>13</td>
<td><strong>Cisco</strong> has owned the <strong>trademark</strong> on the name &quot;iPhone&quot; since 2000, when it acquired InfoGear Technology Corp., which originally registered the name.</td>
</tr>
<tr>
<td>13</td>
<td>With the imminent availability of Apple's iPhone, a look at 10 things current smartphones like the <strong>Nokia N95</strong> have been able to do for a while and that the <strong>iPhone</strong> can't currently match...</td>
</tr>
</tbody>
</table>

Another way to activate iPhone

iPhone trademark originally owned by Cisco

A better choice for smart phones?
Results: Product (iPhone)

- Support statistics for review aspects

People care about price

Controversy: activation requires contract with AT&T

People comment a lot about the unique wi-fi feature
Latent Aspect Rating Analysis

Hotel Palomar Chicago: Traveler Reviews

Great location + spacious room = happy traveler

Stayed for a weekend in July. Walked everywhere, enjoyed the comfy bed and quiet hallways. more

terrific service and gorgeous facility

I stayed at the Palomar with my young daughter for three nights June 17-20, 2010 and absolutely loved the hotel. The room was one of the nicest I’ve ever stayed in (my daughter loved the Fuji jetted tub so much that she wanted to take 2 baths a day!) in terms of decor, design, and size. (It compared favorably to... more

---

How to infer aspect ratings?

How to infer aspect weights?
Solution: Latent Rating Regression Model

Aspect Segmentation + Latent Rating Regression

Reviews + overall ratings

Aspect segments

Term weights

Aspect Rating

Aspect Weight

Location: 1
Amazing: 1
Walk: 1
Anywhere: 1
Room: 1
Nicely: 1
Appointed: 1
Comfortable: 1
Nice: 1
Accommodating: 1
Smile: 1
Friendliness: 1
Attentiveness: 1

0.0
0.1
0.3
0.0
0.1
0.7
0.1
0.9
0.6
0.8
0.7
0.8
0.9
0.9
0.7
0.8
0.9
0.1
0.1
0.3
0.9
0.6
0.8
0.7
0.8
0.9
0.1
0.1
0.3
0.9
0.6
0.8
0.7
0.8
0.9
0.1
0.1
0.3
0.9
0.6
0.8
0.7
0.8
0.9

Topic model for aspect discovery
Table 6: Aspect-based Comparative Summarization (Hotel Max in Seattle)

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Summary</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Value</strong></td>
<td>Truly unique character and a great location at a reasonable price Hotel Max was an excellent choice for our recent three night stay in Seattle.</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>Overall not a negative experience, however considering that the hotel industry is very much in the impressing business there was a lot of room for improvement.</td>
<td>1.7</td>
</tr>
<tr>
<td><strong>Room</strong></td>
<td>We chose this hotel because there was a Travelzoo deal where the Queen of Art room was $139.00/night.</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>Heating system is a window AC unit that has to be shut off at night or guests will roast.</td>
<td>1.2</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td>The location, a short walk to downtown and Pike Place market, made the hotel a good choice.</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>when you visit a big metropolitan city, be prepared to hear a little traffic outside!</td>
<td>2.1</td>
</tr>
<tr>
<td><strong>Business Service</strong></td>
<td>You can pay for wireless by the day or use the complimentary Internet in the business center behind the lobby though.</td>
<td>2.7</td>
</tr>
<tr>
<td></td>
<td>My only complaint is the daily charge for internet access when you can pretty much connect to wireless on the streets anymore.</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Reviewer Behavior Analysis & Personalized Ranking of Entities

Table 4: User behavior analysis

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Expensive Hotel</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 Star</td>
<td>3 Star</td>
<td>5 Star</td>
<td>1 Star</td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>0.134</td>
<td>0.148</td>
<td>0.171</td>
<td>0.093</td>
<td></td>
</tr>
<tr>
<td>Room</td>
<td>0.098</td>
<td>0.162</td>
<td>0.126</td>
<td>0.121</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>0.171</td>
<td>0.074</td>
<td>0.161</td>
<td>0.082</td>
<td></td>
</tr>
<tr>
<td>Cleanliness</td>
<td>0.081</td>
<td>0.163</td>
<td>0.116</td>
<td>0.294</td>
<td></td>
</tr>
<tr>
<td>Service</td>
<td>0.251</td>
<td>0.101</td>
<td>0.101</td>
<td>0.049</td>
<td></td>
</tr>
</tbody>
</table>

People like expensive hotels because of good service

People like cheap hotels because of good value

Query: 0.9 value 0.1 others

Table 10: Personalized Hotel Ranking

<table>
<thead>
<tr>
<th>Hotel</th>
<th>Overall Rating</th>
<th>Price</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majestic Colonial</td>
<td>5.0</td>
<td>339</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Agua Resort</td>
<td>5.0</td>
<td>753</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Majestic Elegance</td>
<td>5.0</td>
<td>537</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Grand Palladium</td>
<td>5.0</td>
<td>277</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Iberostar</td>
<td>5.0</td>
<td>157</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Elan Hotel Modern</td>
<td>5.0</td>
<td>216</td>
<td>Los Angeles</td>
</tr>
<tr>
<td>Marriott San Juan Resort</td>
<td>4.0</td>
<td>354</td>
<td>San Juan</td>
</tr>
<tr>
<td>Punta Cana Club</td>
<td>5.0</td>
<td>409</td>
<td>Punta Cana</td>
</tr>
<tr>
<td>Comfort Inn</td>
<td>5.0</td>
<td>155</td>
<td>Boston</td>
</tr>
<tr>
<td>Hotel Commonwealth</td>
<td>4.5</td>
<td>313</td>
<td>Boston</td>
</tr>
</tbody>
</table>

Non-Personalized

Personalized
How can we extend a search engine to leverage topic models for text analysis?

How should we extend a search engine to support text analysis in general?
Analysis Engine based on Topic Models

Search + Analysis Interface

Query

Search Engine

Results

Information Synthesis
Comparison
Summarization
Categorization ...

Workspace

Topic Models
Beyond Search: Toward a General Analysis Engine

Analysis Engine

Task Completion

Search 1

Search 2

Search

Information Synthesis & Analysis

Decision Making Learning

...
Challenges in Building a General Analysis Engine

- What is a “task” and how can we formally model a task? (task vs. intent vs. information needs)
- How to design a task specification language?
- How do we design a set of general analysis operators to accommodate many different tasks?
- What does ranking mean in an analysis engine (ranking terms, documents, topics, operators)?
- What should the user interface look like?
- How can we seamlessly integrate search and analysis?
- How should we evaluate an analysis engine?
- …
Analysis Operators

- Select
- Split
- Intersect
- Union
- Ranking
- Topic
- Interpret
- Compare

<table>
<thead>
<tr>
<th>Common</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
</tr>
<tr>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
</tbody>
</table>
Examples of Specific Operators

- $C=\{D_1, \ldots, D_n\}; \quad S, S_1, S_2, \ldots, S_k$ subset of $C$

- **Select Operator**
  - Querying (Q): $C \rightarrow S$
  - Browsing: $C \rightarrow S$

- **Split**
  - Categorization (supervised): $C \rightarrow S_1, S_2, \ldots, S_k$
  - Clustering (unsupervised): $C \rightarrow S_1, S_2, \ldots, S_k$

- **Interpret**
  - $C \times \theta \rightarrow S$

- **Ranking**
  - $\theta \times S_i \rightarrow$ ordered $S_i$
Compound Analysis Operator: Comparison of K Topics

Interpret(Compare(Select(T1,C), Select(T2,C),...,Select(Tk,C)),C)
Compound Analysis Operator: Split and Compare

\[ \text{Interpret}(\text{Compare}(\text{Split}(S,k)), C) \]
BeeSpace System

Automation-Confidence (AC) Tradeoff

Automation of task

Return Raw Search Results

Confidence in service

Multi-Resolution Information Delivery

Deliver Actionable Knowledge

Goal
Automation-Generality (AG) Tradeoff

Automation of task

Complete support for special tasks

Operator-Based Analysis Engine

Search Engine

Scalability/Generality

Goal
**Automation-Confidence Tradeoff: Dining Analogy**

**Serve Raw-Food**
Need further processing, but flexible for making different dishes

**Serve Cooked Dishes**
Directly useful for a task, But would be worse if it’s not the right dish
Automation-Generality Tradeoff: Dining Analogy

What’s the right paradigm? Need both paradigms?

**Buffet Paradigm**
Basic Components + Infinite Combination

**Food Court Paradigm**
Finite Choices of Complete Packages
Summary

• Statistical topic models are promising general tools for supporting text analysis
• Next-generation search engines should go beyond search to seamlessly support text analysis and better help users complete their tasks
• Many challenges to be solved:
  – Task modeling
  – Task specification language
  – New analysis operators
  – New ranking models
  – New interface issues
  – New evaluation challenges
  – Automation-Generality (AG) tradeoff & Automation-Confidence (AC) tradeoff
  – …
Looking Ahead...

Text Analysis/Mining

Databases & Data Mining

Visualization

Natural Language Processing

Information Retrieval
Acknowledgments

• Collaborators: Qiaozhu Mei, Yue Lu, Hongning Wang, Jiawei Han, Bruce Schatz, and many others

• Funding
Thank You!

Questions/Comments?