Automatic Construction of Topic Maps for Navigation in Information Space

ChengXiang ("Cheng") Zhai

Department of Computer Science
University of Illinois at Urbana-Champaign
http://www.cs.uiuc.edu/homes/czhai

My Group: TIMAN@UIUC

Text Data

WWW Desktop
Blog
Literature Intranet
Email

We develop general models, algorithms, systems for

Text Data Access

Pull:
- Retrieval models
- Personalized search
- Topic map for browsing

Push:
- Recommender Systems

Text Data Mining

Contextual topic mining
Opinion integration and summarization
Information trustworthiness

Applications
in multiple domains

http://timan.cs.uiuc.edu

Today's talk

12 Ph.D. students
5 MS students
5 Undergraduates
Information Seeking as Sightseeing

- **Know the address of an attraction site?**
  - Yes: take a taxi and go directly to the site
  - No: walk around or take a taxi to a nearby place then walk around

- **Know what exactly you want to find?**
  - Yes: use the right keywords as a query and find the information directly
  - No: browse the information space or start with a rough query and then browse

**When query fails, browsing comes to rescue...**
Current Support for Browsing is Limited

- **Hyperlinks**
  - Only page-to-page
  - Mostly manually constructed
  - Browsing step is very small
- **Web directories**
  - Manually constructed
  - Fixed categories
  - Only support *vertical* navigation

**Beyond hyperlinks?**

**Beyond fixed categories?**

How to promote browsing as a “first-class citizen”?

---

Sightseeing Analogy Continues…

- **Region**
- **Horizontal navigation**
- **Zoom in**
- **Zoom out**
Topic Map for Touring Information Space

Multiple resolutions

Zoom in

Topic regions

Zoom out

Horizontal navigation

Topic-Map based Browsing

Exploratory Search

Multi-Resolution Topic Map

Parents

Current Position

Horizontal Neighbors

Querying

Topic Region

Demo

Click-Through:
- http://www.nadaguides.com/ (17)
- http://www.kbb.com/ (17)
- http://www.c-domande.com/ (11)
- http://www.c-domande.com/kbb/ (10)
- http://yedok.yahoo.com/ (8)

Search Result for “used car”
How can we construct such a multi-resolution topic map automatically?

Multiple possibilities…

Rest of the talk

• Constructing a topic map based on user interests
• Constructing a topic map based on document content
• Summary & Future Directions
Search Logs as Information Footprints

Footprints in information space

User 2722 searched for "national car rental" [!] at 2006-03-09 11:24:29
User 2722 searched for "military car rental benefits" [!] at 2006-03-10 09:33:37 (found http://www.valoans.com)
User 2722 searched for "military car rental benefits" [!] at 2006-03-10 09:33:37 (found http://benefits.military.com)
User 2722 searched for "military car rental benefits" [!] at 2006-03-10 09:33:37 (found http://www.avis.com)
User 2722 searched for "enterprise rent a car" [!] at 2006-04-05 23:37:42 (found http://www.enterprise.com)
User 2722 searched for "meineke car care center" [!] at 2006-05-02 09:12:49 (found http://www.meineke.com)
User 2722 searched for "car rental" [!] at 2006-05-25 15:54:36
User 2722 searched for "autosave car rental" [!] at 2006-05-25 23:26:54 (found http://eautosave.com)
......

Information Footprints → Topic Map

• Challenges
  – How to define/construct a topic region
  – How to control granularities/resolutions of topic regions
  – How to connect topic regions to support effective browsing

• Two approaches
  – Multi-granularity clustering [Wang et al. CIKM 2009]
  – Query editing [Wang et al. CIKM 2008]


Multi-Granularity Clustering

Star clustering

Multi-Granularity Clustering

Star clustering
Multi-Granularity Clustering

Control granularity

Star clustering

Adding horizontal links

Control granularity

Star clustering
Star Clustering [Aslam et al. 04]

1. Form a similarity graph
   - TF-IDF weight vectors
   - Cosine similarity
   - Thresholding

2. Iteratively identify a “star center” and its “satellites”

“Star center” query serves as a label for a cluster

Simulation Experiments

Could the user have browsed into \( C_1, C_2, \) and \( C_3 \) with a map without using \( Q_2, ..., Q_k \)?
Browsing can be more effective than query reformulation

![Graph showing P@5 and P@10 for different query editing methods]

Topic Map as Systematic Query Editing

Query Term Substitution

**Addition**

- insurance
- rental
- car
- auto
- loan
- car:parts
- car:used
- car:blue+book
- rental:boat
- car:rental
- car:pictures
- national+car+rental
- enterprise+car+rental
- alamo+car+rental
- exotic+car+rental

**Substitution**

- 0.03
- 0.03
- 0.03
- 0.03
- 0.03
- 0.03
- 0.03
- 0.03
- 0.03
Map Construction
= Mining Query-Editing Patterns

- Context-sensitive term substitution
  
  \[
  \text{auto} \rightarrow \text{car} \mid _\text{wash}
  \]
  
  \[
  \text{yellowstone} \rightarrow \text{glacier} \mid _\text{park}
  \]

- Context-sensitive term addition

  \[
  +\text{sale} \mid \text{auto}_\text{quotes}
  \]

  \[
  +\text{progressive} \mid _\text{auto insurance}
  \]

Dynamic Topic Map Construction

Search logs

Offline

Task 1: Contextual Models

Task 2: Translation Models

Task 3: Pattern Retrieval

q = auto wash

\[
\text{auto} \rightarrow \text{car} \mid _\text{wash}
\]

\[
\text{auto} \rightarrow \text{truck} \mid _\text{wash}
\]

\[
+\text{southland} \mid _\text{auto wash}
\]

\[
\text{car wash}
\]

\[
\text{truck wash}
\]

\[
\text{southland auto wash}
\]
Examples of Contextual Models

- Left and Right contexts are different
- General context mixed them together

<table>
<thead>
<tr>
<th>$w=\text{car}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
</tr>
<tr>
<td>rental</td>
</tr>
<tr>
<td>rent</td>
</tr>
<tr>
<td>rentals</td>
</tr>
<tr>
<td>enterprise</td>
</tr>
<tr>
<td>national</td>
</tr>
<tr>
<td>prices</td>
</tr>
<tr>
<td>audio</td>
</tr>
<tr>
<td>budget</td>
</tr>
<tr>
<td>insurance</td>
</tr>
<tr>
<td>dealers</td>
</tr>
</tbody>
</table>

Examples of Translation Models

- Conceptually similar keywords have high translation probabilities
- Provide possibility for exploratory search in an interactive manner

<table>
<thead>
<tr>
<th>$w=\text{computer}$</th>
<th>$w=\text{leg}$</th>
<th>$w=\text{chinese}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s$</td>
<td>$t(s</td>
<td>w)$</td>
</tr>
<tr>
<td>computer</td>
<td>0.00155</td>
<td>leg</td>
</tr>
<tr>
<td>computers</td>
<td>0.00014</td>
<td>abdominal</td>
</tr>
<tr>
<td>laptop</td>
<td>0.00011</td>
<td>stomach</td>
</tr>
<tr>
<td>pc</td>
<td>0.00009</td>
<td>legs</td>
</tr>
<tr>
<td>notebook</td>
<td>0.00009</td>
<td>muscle</td>
</tr>
</tbody>
</table>
## Sample Term Substitutions

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Reworded query</th>
</tr>
</thead>
<tbody>
<tr>
<td>auto → car</td>
<td>car wash</td>
</tr>
<tr>
<td>car → auto</td>
<td>auto trade</td>
</tr>
<tr>
<td>children → kids</td>
<td>kids games</td>
</tr>
<tr>
<td>kids → children</td>
<td>children clothing</td>
</tr>
<tr>
<td>driving → maps</td>
<td>google maps</td>
</tr>
<tr>
<td>military → army</td>
<td>army acu</td>
</tr>
<tr>
<td>birthday → greeting</td>
<td>greeting cards</td>
</tr>
<tr>
<td>lotto → lottery</td>
<td>florida lottery results</td>
</tr>
<tr>
<td>interpretation → meanings</td>
<td>meanings of dreams</td>
</tr>
<tr>
<td>music → song</td>
<td>song lyrics</td>
</tr>
</tbody>
</table>

## Sample Term Addition Patterns

Given: \( q = \text{“wedding”} \)

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Refined Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>+dresses</td>
<td>wedding dresses</td>
</tr>
<tr>
<td>+cakes</td>
<td>wedding cakes</td>
</tr>
<tr>
<td>+invitations</td>
<td>wedding invitations</td>
</tr>
<tr>
<td>+songs</td>
<td>wedding songs</td>
</tr>
<tr>
<td>+favors</td>
<td>wedding favors</td>
</tr>
<tr>
<td>+flowers</td>
<td>wedding flowers</td>
</tr>
<tr>
<td>+gowns</td>
<td>wedding gowns</td>
</tr>
<tr>
<td>+rings</td>
<td>wedding rings</td>
</tr>
<tr>
<td>+toasts</td>
<td>wedding toasts</td>
</tr>
<tr>
<td>+vows</td>
<td>wedding vows</td>
</tr>
</tbody>
</table>
Effectiveness of Query Suggestion

Our method

[Jones et al. 06]

Rest of the talk

• Constructing a topic map based on user interests
• Constructing a topic map based on document content
• Summary & Future Directions
Document-Based Topic Map

- **Advantages over user-based map**
  - More complete coverage of topics in the information space
  - Can help satisfy long-tail information needs

- **Construction methods**
  - Traditional clustering approaches: hard to capture subtopics in text
  - Generative topic models: more promising and able to incorporate non-textual context variables

- **Two cases:**
  - Construct topic map with probabilistic latent topic analysis
  - Construct topic evolution map with probabilistic citation graph analysis

Contextual Probabilistic Latent Semantics Analysis

[Mei & Zhai KDD 2006]

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.
Theme Evolution Graph: KDD [Mei & Zhai KDD 2005]

Joint Analysis of Text Collections and Associated Network Structures [Mei et al., WWW 2008]
### Topics from Pure Text Analysis

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>term</td>
<td>peer</td>
<td>visual</td>
<td>interface</td>
</tr>
<tr>
<td>question</td>
<td>patterns</td>
<td>analog</td>
<td>towards</td>
</tr>
<tr>
<td>protein</td>
<td>mining</td>
<td>neurons</td>
<td>browsing</td>
</tr>
<tr>
<td>training</td>
<td>clusters</td>
<td>vlsi</td>
<td>xml</td>
</tr>
<tr>
<td>weighting</td>
<td>stream</td>
<td>motion</td>
<td>generation</td>
</tr>
<tr>
<td>multiple</td>
<td>frequent</td>
<td>chip</td>
<td>design</td>
</tr>
<tr>
<td>recognition</td>
<td>e</td>
<td>natural</td>
<td>engine</td>
</tr>
<tr>
<td>relations</td>
<td>page</td>
<td>cortex</td>
<td>service</td>
</tr>
<tr>
<td>library</td>
<td>gene</td>
<td>spike</td>
<td>social</td>
</tr>
</tbody>
</table>

### Topical Communities Discovered from Joint Analysis

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieval</td>
<td>mining</td>
<td>neural</td>
<td>web</td>
</tr>
<tr>
<td>information</td>
<td>data</td>
<td>learning</td>
<td>services</td>
</tr>
<tr>
<td>document</td>
<td>discovery</td>
<td>networks</td>
<td>semantic</td>
</tr>
<tr>
<td>query</td>
<td>databases</td>
<td>recognition</td>
<td>services</td>
</tr>
<tr>
<td>text</td>
<td>rules</td>
<td>analog</td>
<td>peer</td>
</tr>
<tr>
<td>search</td>
<td>association</td>
<td>vlsi</td>
<td>ontologies</td>
</tr>
<tr>
<td>evaluation</td>
<td>patterns</td>
<td>neurons</td>
<td>rdf</td>
</tr>
<tr>
<td>user</td>
<td>frequent</td>
<td>gaussian</td>
<td>management</td>
</tr>
<tr>
<td>relevance</td>
<td>streams</td>
<td>network</td>
<td>ontology</td>
</tr>
</tbody>
</table>
Constructing Topic Evolution Map with Probabilistic Citation Analysis [Wang et al. under review]

- Given research articles and citations in a research community
- Identify major research topics (themes) and their spans
- Construct a topic evolution map

For each topic, identify milestone papers

---

Probabilistic Modeling of Literature Citations

- Modeling the generation of literature citations
  - Document: bag of “citations”
  - Topic: distribution over documents
  - To generate a document:
    - draw topic sample $z \sim D_{doc.topic}(z; d)$
    - draw document sample $c \sim D_{topic.doc}(c; z)$
  - Any topic model can be used
Citation-LDA

- Document-topic distribution: $\theta_d \sim \text{Dir}(\alpha)$
- Topic-Document distribution: $\varphi_z \sim \text{Dir}(\beta)$

- To generate citations in document $d_i$
  - sample a topic $k \sim \text{Multi}(\theta_i)$
  - sample a document to cite $c \sim \text{Multi}(\varphi_k)$

Summarization of a Topic

- Milestone papers: The topic-document $\{\hat{\varphi}_{k,j}\}$ distribution provides a natural ranking of papers
- Topic Key Words: weighted word counts in document titles
- Topic Life Span: Expected Topic Time:

$$t(z = k) = \mathbb{E}_{z = k}[t(d)] = \sum_i \Pr(d = i | z = k) t(d = i)$$ (5)
Citation Structure and Topic Evolution

• **Topic-level citation distribution:**

\[
\Pr(k^{(1)} \rightarrow k^{(2)} | k^{(1)})
= \sum_i \Pr(c = i | z = k^{(1)}) \Pr(z = k^{(2)} | d = i)
= \sum_i \phi_{k^{(1)},i} \theta_{i,k^{(2)}}
\]  

• **Theme Evolution Patterns**

- Branching
- Merging
- Shifting
- Fading-out

Sample Results: Major Topics in NLP Community

<table>
<thead>
<tr>
<th>Topic</th>
<th>Weight</th>
<th>$E(t)$</th>
<th>Top Word Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>94</td>
<td>0.02806</td>
<td>2005.16</td>
<td>dependency parsing, non-projective, shared tasks, multilingual</td>
</tr>
<tr>
<td>89</td>
<td>0.02761</td>
<td>2004.64</td>
<td>sentiment classification, opinion analysis, orientation, learning</td>
</tr>
<tr>
<td>8</td>
<td>0.02509</td>
<td>1991.37</td>
<td>word sense disambiguation, lexical semantics</td>
</tr>
<tr>
<td>92</td>
<td>0.02428</td>
<td>2004.98</td>
<td>machine translation, phrase-based models, alignment</td>
</tr>
<tr>
<td>96</td>
<td>0.02277</td>
<td>2005.45</td>
<td>machine translation, online, margin, discriminative learning</td>
</tr>
<tr>
<td>84</td>
<td>0.02093</td>
<td>2003.94</td>
<td>semantic role labeling, shared tasks</td>
</tr>
<tr>
<td>80</td>
<td>0.02069</td>
<td>2003.44</td>
<td>machine translation, reordering, alignment</td>
</tr>
<tr>
<td>73</td>
<td>0.01965</td>
<td>2002.76</td>
<td>discriminative parsing, sequential labeling, part-of-speech</td>
</tr>
<tr>
<td>50</td>
<td>0.01908</td>
<td>2000.87</td>
<td>machine translation, minimum error rate training, BLEU evaluation</td>
</tr>
<tr>
<td>72</td>
<td>0.01804</td>
<td>2002.74</td>
<td>coreference resolution, machine learning, anaphora, pronoun</td>
</tr>
</tbody>
</table>

ACL Anthology Network (AAN)
Papers from NLP major conferences from 1965 - 2011
18,041 papers
82,944 citations
Citation Structure

NLP-Community Topic Evolution

• Topic Evolution: (green: newer, red: older)

25: Spelling correction (1997)
95: Dependency parsing (2005)
34: Statistical parsing (1998)
73: Discriminative-learning parsing (2002)
18: Prepositional phrase attachment (1994)
27: Prepositional phrase attachment (1994)
50: min-error-rate approaches (2000)
29: decoding, alignment, reordering (2000)
89: Sentiment Analysis (2004)
8: Word sense disambiguation (1991)
Component 1
Component 2
Component 3
Branching

Fading-out

Shifting
### Rest of the talk

- **Constructing a topic map based on user interests**
- **Constructing a topic map based on document content**
- **Summary & Future Directions**

#### Detailed View of Topic “Statistical Machine Translation”

<table>
<thead>
<tr>
<th>Topic</th>
<th>Year</th>
<th>ID</th>
<th>Paper Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>1990</td>
<td>P1</td>
<td>A Statistical Approach To Machine Translation</td>
</tr>
<tr>
<td></td>
<td>1991</td>
<td>P2</td>
<td>A Program For Aligning Sentences In Bilingual Corpora</td>
</tr>
<tr>
<td>$T_2$</td>
<td>1996</td>
<td>P4</td>
<td>HMM-Based Word Alignment In Statistical Translation</td>
</tr>
<tr>
<td></td>
<td>1997</td>
<td>P5</td>
<td>Decoding Algorithm In Statistical Machine Translation</td>
</tr>
<tr>
<td></td>
<td>1999</td>
<td>P6</td>
<td>Improved alignment models for statistical machine translation</td>
</tr>
<tr>
<td></td>
<td>2002</td>
<td>P8</td>
<td>Discriminative Training And Maximum Entropy Models For Statistical Machine Translation</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>P9</td>
<td>Minimum Error Rate Training In Statistical Machine Translation</td>
</tr>
<tr>
<td>$T_4$</td>
<td>2003</td>
<td>P10</td>
<td>Statistical Phrase-Based Translation</td>
</tr>
<tr>
<td></td>
<td>2005</td>
<td>P11</td>
<td>A Hierarchical Phrase-Based Model For Statistical Machine Translation</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>P12</td>
<td>Hierarchical Phrase-Based Translation</td>
</tr>
</tbody>
</table>
Summary

• Querying & Browsing are complementary ways of navigating in information space
• General support for browsing requires a topic map
• It’s feasible to automatically construct topic maps
  – Search logs \(\rightarrow\) multi-resolution topic map
  – Document content + context \(\rightarrow\) contextualized topic map
  – Citation graph \(\rightarrow\) topic evolution map
• Topic maps naturally enable collaborative surfing

Collaborative Surfing

New queries become new footprints

Navigation trace enriches map structures

Clickthroughs become new footprints

Browse logs offer more opportunities to understand user interests and intents
Future Research Questions

- How do we evaluate a topic map?
- How do we visualize a topic map?
- How can we leverage ontology to construct a topic map?
- A navigation framework for unifying querying and browsing
  - Formalization of a topic map
  - Algorithms for constructing a topic map
  - Topic maps with multiple views
- A sequential decision model for optimal interactive information seeking
  - Optimal topic/region/document ranking
  - Learn user interests and intents from browse logs + query logs
  - Intent clarification
- Beyond information access to support knowledge service (information space $\rightarrow$ knowledge space)

Future: Towards Multi-Mode Information Seeking & Analysis

**Multi-Mode Text Access**
- Pull: Querying + Browsing
- Push: Recommendation

**Multi-Mode Text Analysis**
- Topic extraction & analysis
- Sentiment analysis
- ...

**Interactive Decision Support**

Big Raw Data $\rightarrow$ Small Relevant Data $\rightarrow$ Knowledge

Need to develop a general framework to support all these
IKNOWX: Intelligent Knowledge Service (collaboration with Prof. Ying Ding)

Future knowledge service systems

Inferences
- Text summarization
- Document Linking
- Passage Linking

Question Answering
- Entity-relation summarization
- Entity Resolution
- Relation Resolution

Document Retrieval
- Document Retrieval
- Passage Retrieval
- Entity Retrieval
- Relation Retrieval

Current Search engines

Acknowledgments

- Contributors: Xuanhui Wang, Xiaolong Wang, Qiaozhu Mei, Yanen Li, and many others
- Funding
Thank You!

Questions/Comments?