Overview of Text Data Access

컴퓨터과학과 천상진

csjin75@uos.ac.kr

INDEX

Intro

Text data access is the foundation for text analysis.

1) retrieval of the most relevant text data
   to a particular analysis problem

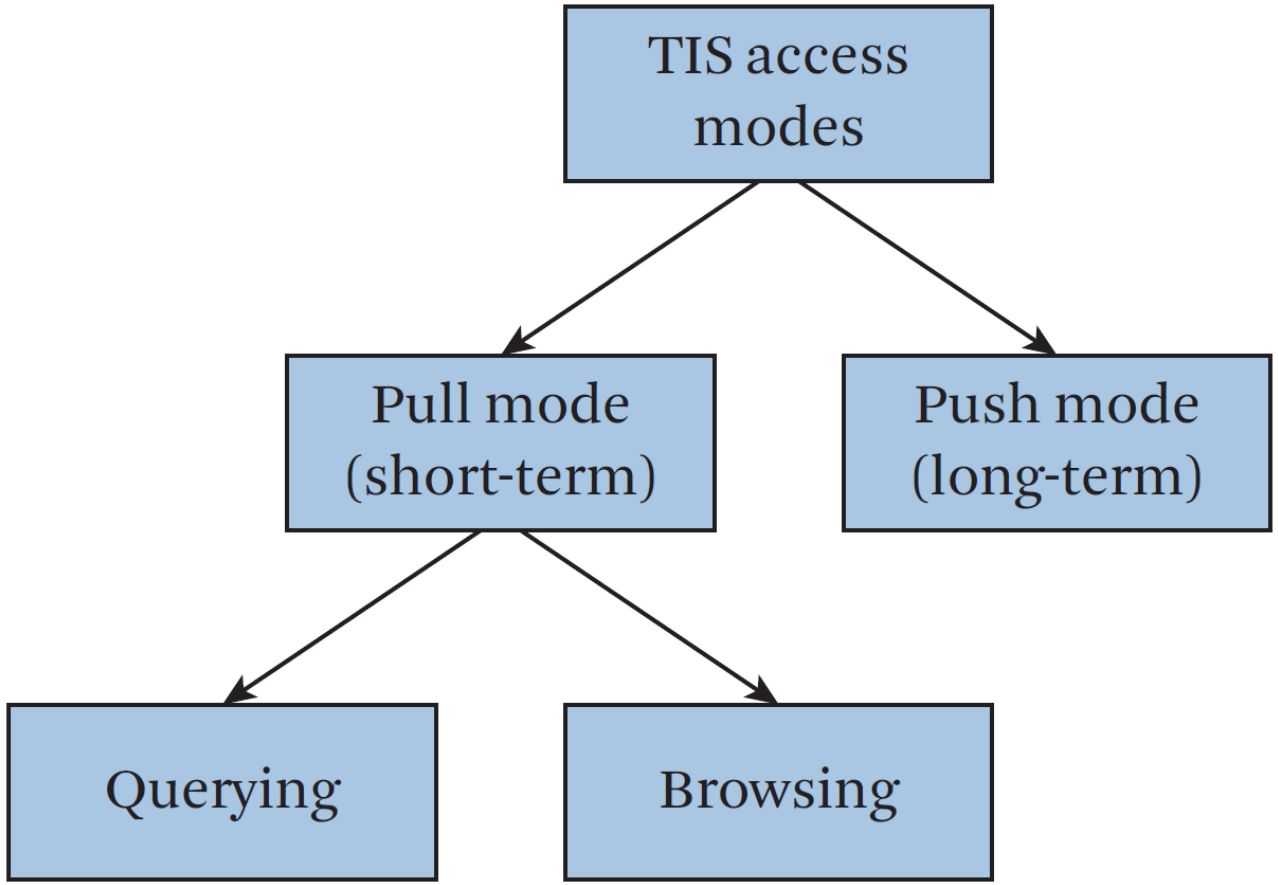
2) interpretation of any analysis results or discovered knowledge
   in appropriate context and provides data provenance

Intro

The general goal of text data access is to connect user with the **right information** at **the right time**

pull	*push*
the users take the initiative to fetch relevant information out from the system	the system takes the initiative to offer relevant information to users

5.1 Access Mode: Pull vs. Push



5.1 Pull mode *(short-term need)*

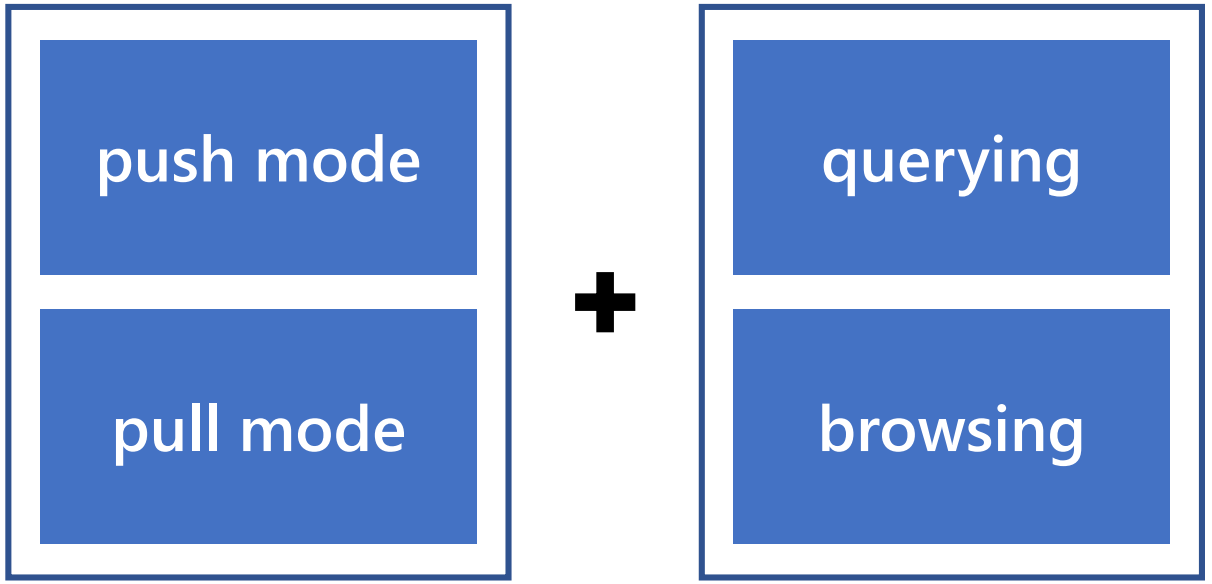
- ✓ the user initiates the access process to find **the relevant text data**, typically by using a search engine
- ✓ **querying** and **browsing** is two complementary ways of finding relevant information in the information space
- ✓ temporary and usually satisfied through **search** or **navigation** in the information space

5.1 Push mode *(long-term need)*

- ✓ the system initiates the process to **recommend** a set of relevant information items to the user
- ✓ be desirable to monitor any relevant text data about a topic related to the application
- ✓ satisfied through **filtering** or **recommendation** where the system would take the initiative to push the relevant information to a user

5.2 Multimode Interactive Access

multimode interactive access to relevant text data



Social Surfing

Search

TOPIC MAP for NAVIGATION

- table (0.000313848)
- dining (0.000139667)
  - dining table (center)
    - round dining table (0.0104634)
    - dining table bases (0.00554701)
    - asian dining table (0.00371214)
    - door dining table (0.00301223)
    - walnut dining table (0.00291244)
    - dining table scottsdale (0.00111778
  - dining tables (0.0554608)
  - dining chairs (0.00490978)
  - dining set (0.00191779)
  - dinning table (0.0015674)
  - patio table (0.00144361)
  - kitchen table (0.00119744)
  - buffet table (0.00056739)
  - pub table (0.000498793)
  - bar table (0.000487214)
  - coffee table (0.000159963)

Click-Through

http://www.diningtableguide.com/ (1)
http://www.diningtables.com/ (1)
http://www.flickr.com/photos/17425845@n00/121705244/in/pool-69453349@n00 (1)
http://www.glass-dining-table.net/ (1)

Search Result for "dining table"

Dining Tables - Overstock™ Shopping - The Best Prices Online

http://www.overstock.com/Home-Garden/Dining-Tables
/2021/subcat.htmlsa=Uei=hUUtVaabHMPAsAWd3IGIBgved=0CCwQFjAAusg=AFQjCNGkI

Dining Table and Chairs Dining Tables | Pottery Barn
www.potterybarn.com/shop/furniture-upholstery/tables/CachedSimilarPottery Barns dining tal
Choose from a wide range of dining tables and chairs in classic styles and...
http://www.potterybarn.com/shop/furniture-upholstery/tables
/sa=Uei=hUUtVaabHMPAsAWd3IGIBgved=0CDcQFjABusg=AFQjCNGC1FkPZSCWB5NI

Dining tables - Up to 4 seats Up to 6 seats - IKEA
www.ikea.com/us/en/catalog/categories/departments/dining/21825/CachedSimilarDining table
games, helping with homework or just lingering after a meal, theyre where you share...

5.2 Multimode Interactive Access

Querying (long-range jump)

→ when a user submits a new query,
the search results will be shown in the right pane and
the relevant part of a topic map will be shown in the left pane

Navigating on the map (short-range walk)

→ the left pane is to let a user navigate on the map

→ a user can zoom in or zoom out to nodes

→ the number attached to a node is a score for node ranking

5.2 Multimode Interactive Access

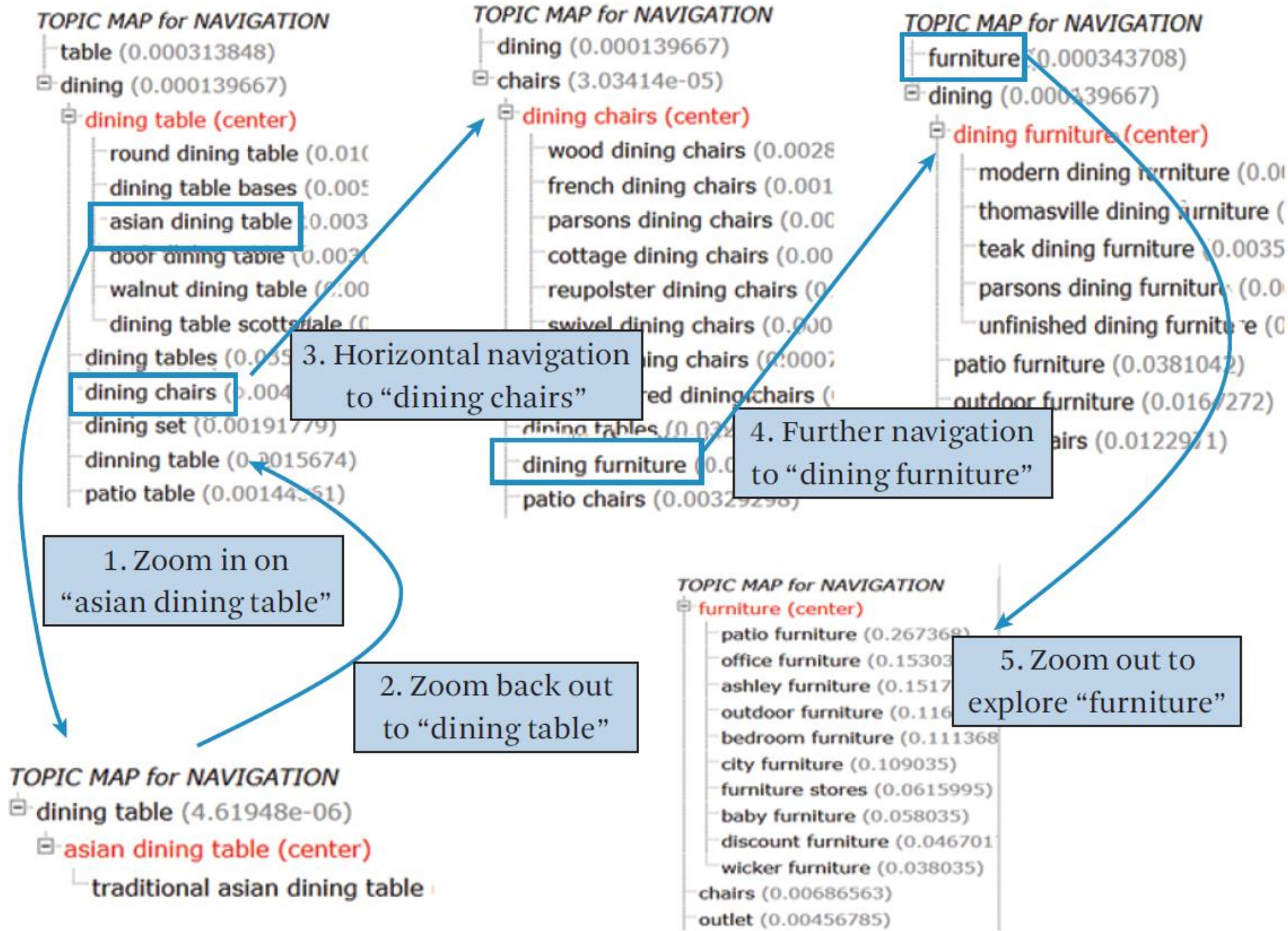
Viewing a topic region

→ a user may double-click on a topic node to view the documents

→ the result pane always shows the documents that the user is focused on

Viewing a document

→ a user can select any document to view


TOPIC MAP for NAVIGATION
table (0.000313848)
dining (0.000139667)
dining table (center)
round dining table (0.01(
dining table bases (0.005
asian dining table 0.003
door dining table (0.003(
walnut dining table (0.00
dining table scottsdale (0
dining tables (0.035
dining chairs (0.004
dining set (0.00191779)
dinning table (0.0015674)
patio table (0.00144561)
TOPIC MAP for NAVIGATION
dining (0.000139667)
chairs (3.03414e-05)
dining chairs (center)
wood dining chairs (0.0028
french dining chairs (0.001
parsons dining chairs (0.0(
cottage dining chairs (0.00
reupolster dining chairs (0.
swivel dining chairs (0.000
ing chairs (0.0007
red dining chairs (
dining tables (0.03
dining furniture (0.0
patio chairs (0.00329298)
TOPIC MAP for NAVIGATION
furniture (0.000343708)
dining (0.000139667)
dining furniture (center)
modern dining furniture (0.0(
thomasville dining furniture (
teak dining furniture (0.0035
parsons dining furniture (0.0
unfinished dining furniture (0
patio furniture (0.0381042)
outdoor furniture (0.0167272)
airs (0.0122971)
3. Horizontal navigation to "dining chairs"
4. Further navigation to "dining furniture"
1. Zoom in on "asian dining table"
2. Zoom back out to "dining table"
TOPIC MAP for NAVIGATION
dining table (4.61948e-06)
asian dining table (center)
traditional asian dining table
TOPIC MAP for NAVIGATION
furniture (center)
patio furniture (0.267368)
office furniture (0.15303
ashley furniture (0.1517
outdoor furniture (0.116
bedroom furniture (0.111368)
city furniture (0.109035)
furniture stores (0.0615995)
baby furniture (0.058035)
discount furniture (0.046701
wicker furniture (0.038035)
chairs (0.00686563)
outlet (0.00456785)
5. Zoom out to explore "furniture"

5.3 Text Retrieval

unstructured text retrieval vs. structured database retrieval

document *ranking* vs. document *selection*

the problem of TR:

to use a query to find relevant documents in a collection of text documents

→ users often have temporary ad hoc information needs for various task,
and would like to find the relevant information immediately

5.3 Why TR is difficult

1. a query is usually quite short and incomplete.

2. the information need is difficult to describe precisely, especially when the user isn't familiar with the topic.

3. precise understanding of the document content is difficult. (since what counts as the correct answer is subjective)

5.4 Text Retrieval vs. Database Retrieval

unstructured data → difficult for computers to understand	**data**	structured data → clearly defined meaning according to schema
keyword queries (a vague specification)	**query**	clearly specifies on the constraints on the fields of the data table
a set of relevant documents	**result**	very specific data elements

5.4 Text Retrieval vs. Database Retrieval

the challenges in building a search engine and a database

- ✓ to first figure out which documents should be returned for a query
- ✓ modeling a user's information need and search tasks

Text Retrieval

- ✓ how to find the answers as quickly as possible especially
- ✓ to maintain the integrity of data

Database Retrieval

5.4 Text Retrieval vs. Database Retrieval

performance measure in TR and DR

- ✓ no mathematical way to prove
- ✓ rely on empirical evaluation using some test collections and users
- ✓ the simulation may not accurately reflect the real applications

Text Retrieval

- ✓ efficiency can prove by analyzing the computational complexity
- ✓ simulation study(to determine which is faster), may not accurately reflect the real applications

Database Retrieval

5.4 Text Retrieval vs. Database Retrieval

problem of text retrieval is an *empirically defined problem.*

→ which method works better cannot be answered
by pure analytical reasoning or mathematical proofs.

5.5 Document Selection vs. Document Ranking

$V = \{w_1, \ldots, w_N\}$ vocabulary set of all the words

$q = q_1, q_2, \ldots, q_m$ user's query ($q_i \in V$)

$d_i = d_{i1}, \ldots, d_{im}$ document ($d_{ij} \in V$)

$C = \{d_1, \ldots, d_M\}$ text collection (set of text documents)

$R(q) \subset C$ which are relevant to the user's query q

5.5 Document Selection vs. Document Ranking

how can compute R'(q) which is approximation of R(q) ?

$$R'(q) = \{d | f(q, d) = 1, d \in C\}$$

binary classification function f

absolute relevance of documents

Document Selection

$$R'(q) = \{d | f(q, d) \geq \theta\}$$

ranking function f

relative relevance of documents

Document Ranking

5.5 Document Selection vs. Document Ranking

ranking is preferred to document selection for multiple reasons

1. the binary classifier is unlikely accurate
(due to the difficulty for a user to prescribe the exact criteria)

2. in case of over-constrained query:
forcing a binary decision may result in no delivery of any search result

3. in case of under-constrained query:
too many documents matching the query, resulting in over-delivery

4. relevance is a matter of degree

5.5 Document Selection vs. Document Ranking

the strategy of ranking is optimal theoretically under two assumptions:

1. The utility of a document to a user is independent of the utility of any other document.
2. A user will browse the results sequentially

"designing a good ranking function that can rank all the relevant documents on top of all the non-relevant ones."