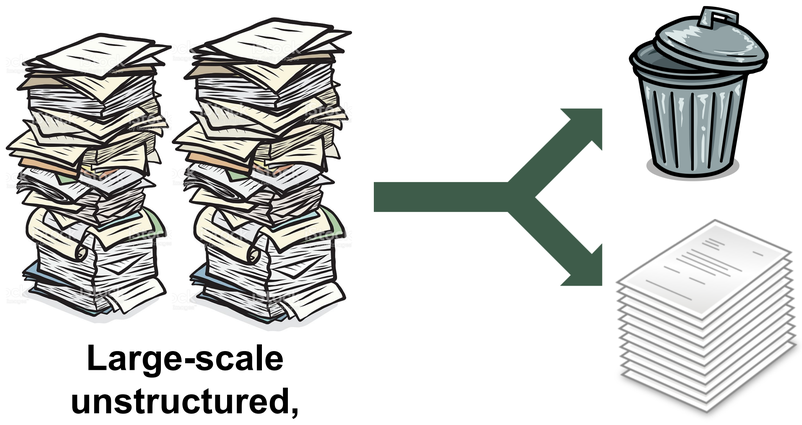
High-Recall Document Retrieval from Large-Scale Noisy Documents via Visual Analytics based on Targeted Topic Modeling

Hannah Kim[1], Jaegul Choo[2], Alex Endert[1], and Haesun Park[1]

[1]Georgia Tech, [2]Korea University

hannahkim@gatech.edu VA.lab

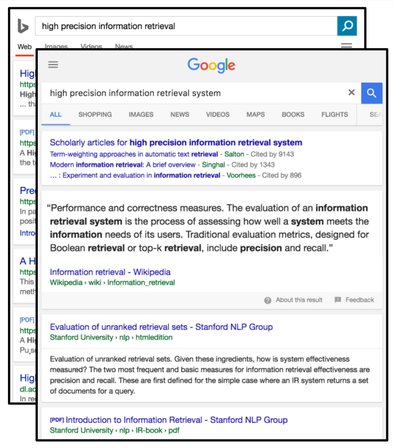
Information Retrieval



Large-scale unstructured, noisy text data

Relevant documents (about a particular event, subject, or product)

High Precision vs. High Recall



Traditional retrieval systems:

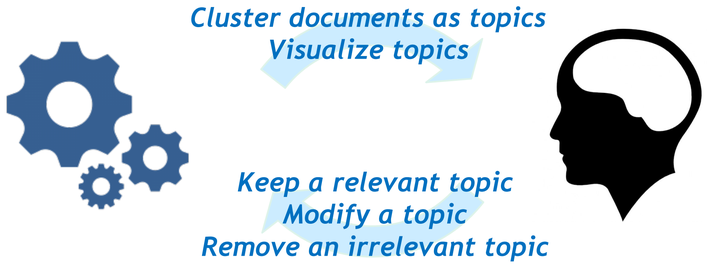
- Focus only on high precision
- Retrieve a number of most relevant documents
- Ex) Google, Bing, Twitter, PubMed

Our system:

- Focuses on high precision and high recall
- Retrieves *ALL* relevant documents
- Is suited when missing any relevant item is critical
- Ex) marketing, social media, legal cases, medical cases, literature review, etc

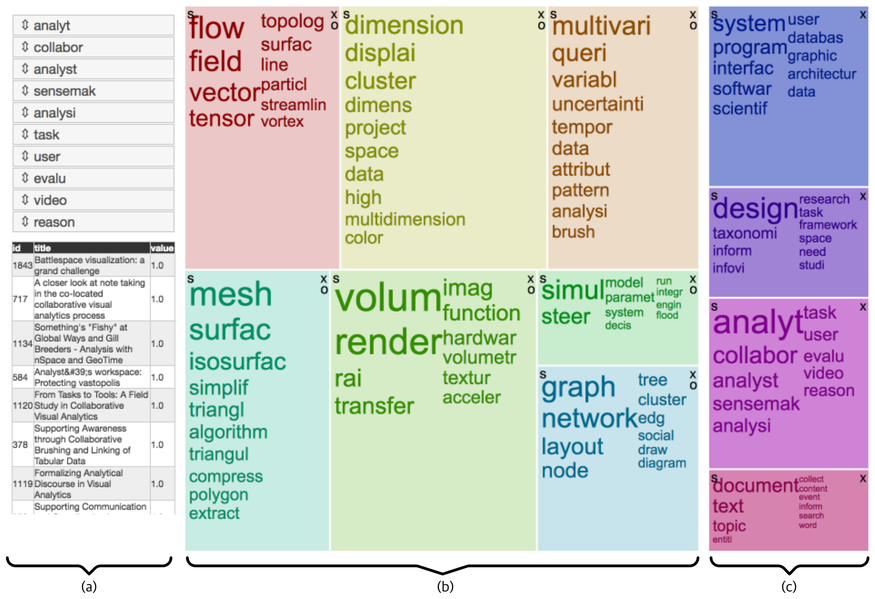
Interactive Visual Retrieval System

Workflow



1. Documents are clustered w.r.t. their topics.
2. Topics are visualized and used as exploration units.
3. Users can inspect a topic and keep/modify/remove it iteratively.

System Design



(a) The topic detail panel with an interactive list of keywords and a document table

(b) Main topic treemap visualization for interactive topic exploration where each cell represents a topic

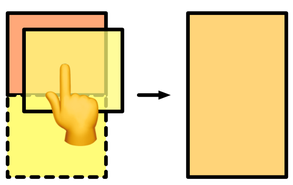
(c) Confirmed topic treemap visualization which shows relevant topics that are confirmed by users

In (b) and (c), semantically similar topics are placed closer.

Supported Interaction

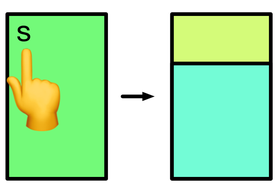
Getting topic details

Click a topic cell to see its details in (a).



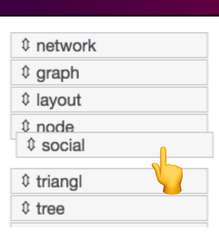
Merging topics.

When two topics are similar, combine them by dragging one topic cell onto another.



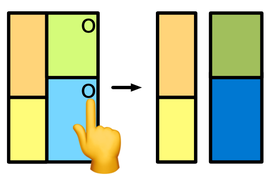
Splitting a topic.

When a topic is too big or too general, split it into subtopics by clicking ‘s’



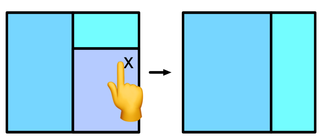
Refining a topic.

To modify a topic, re-order its keywords in (a) to (de/)emphasize a particular aspect.



Keeping/marking a topic.

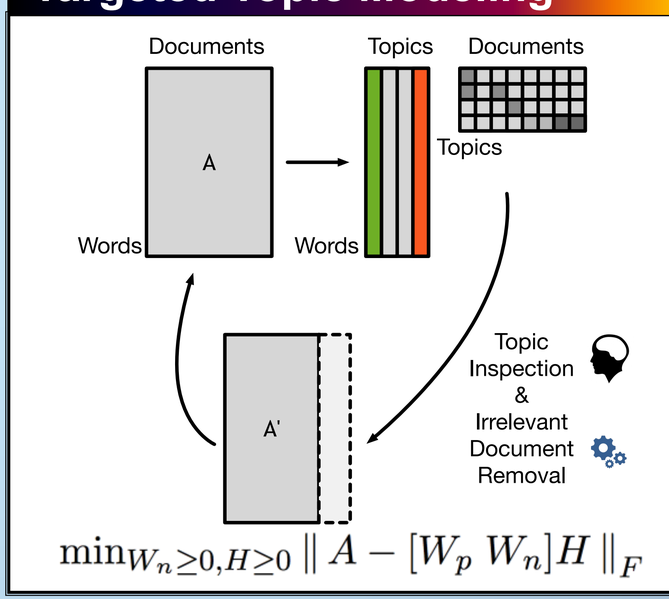
When a topic is interesting and relevant, keep them separate in (c) from unexplored topics in (b) by clicking ‘o’



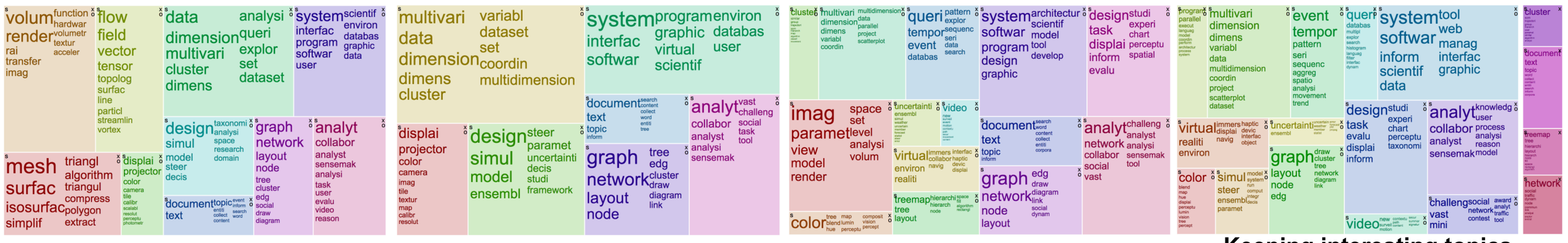
Removing a topic.

When a topic is irrelevant or uninteresting, remove it by clicking ‘x’

Targeted Topic Modeling



Usage Scenario with IEEE VIS publication dataset



Initial topic overview

After removing SciVis topics

Several split/merge/refine interactions

Keeping interesting topics