Rohail Syed and Kevyn Collins-Thompson
School of Information, University of Michigan

Abstract
We know people use Web search to learn but we don’t know what types of documents are best suited for such goals. We train regression models on vocabulary learning outcomes using document/user features. Our models show strong ability to predict learning outcomes even when using only document features. We further conduct the first longitudinal crowdsourced study of learning from Web documents. Our retrieval model outperforms Google search results by over 92% in long-term retention of learning gains.

RQ1: Document Features are Strong Predictors of Learning
- Common positive features for multiple learning measures:
  - Use more images (excluding ads and navigational)
  - Document sets should have longer paragraphs.
  - Document sets should have more coverage of relevant keywords compared to other words.

RQ2: Personalizes Improves Long-term Retention
- Our personalized retrieval model (P) showed 92% better long-term retention of gains (for high-difficulty vocabulary) compared to Google baseline (Web):
  - Retained Gains: # words learned 9 months prior and still known.
  - Robust Gains: # words unknown 9 months prior and now known.

Methodology
Learning from Documents Data
RQ1: Short-term learning
RQ2: Long-term learning

Feature Extraction
Images
Keywords
Paragraphs
Easeiness

Study Long-term Retention Of Learning
- Re-test participants
- 9 months delay

Compare Retention By Retrieval Models
Model 1
Model 2

Discussion
In evaluating our two Research Questions, we found the following general results:

RQ 1 Supported: We were able to fit models with/without user-specific features that showed strong predictive power. Trained on multiple types of learning outcomes.

RQ 2 Supported: Strong evidence that personalized retrieval model yields far better long-term learning retention.

Applications
- Possible integration with existing search systems to better serve queries with learning intents.
- Intelligent tutoring systems can use trained models to predict what learning material to provide students.
- May guide website designers to structure and write content that may improve learning outcomes.

Conclusions
1. Personalized retrieval model could outperform commercial baseline in long-term retention of knowledge.
2. Regression models showed strong ability to predict short-term learning, even without user-specific features.

References