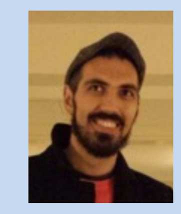
The Posts Recommendation Algorithm for Providing Exploration Guidance in dExplorer

Drashko Nakikj[1], MS, MA and Tom Effland[2], MS

1. *Department of Biomedical Informatics, Columbia University*
2. *Department of Computer Science, Columbia University*



Research topics:
- social computing
- online health communities
- collective sensemaking
- discussion visualization

Research topics:
- machine learning
- natural language processing
- document modeling, mining, and recommendation

Research Questions

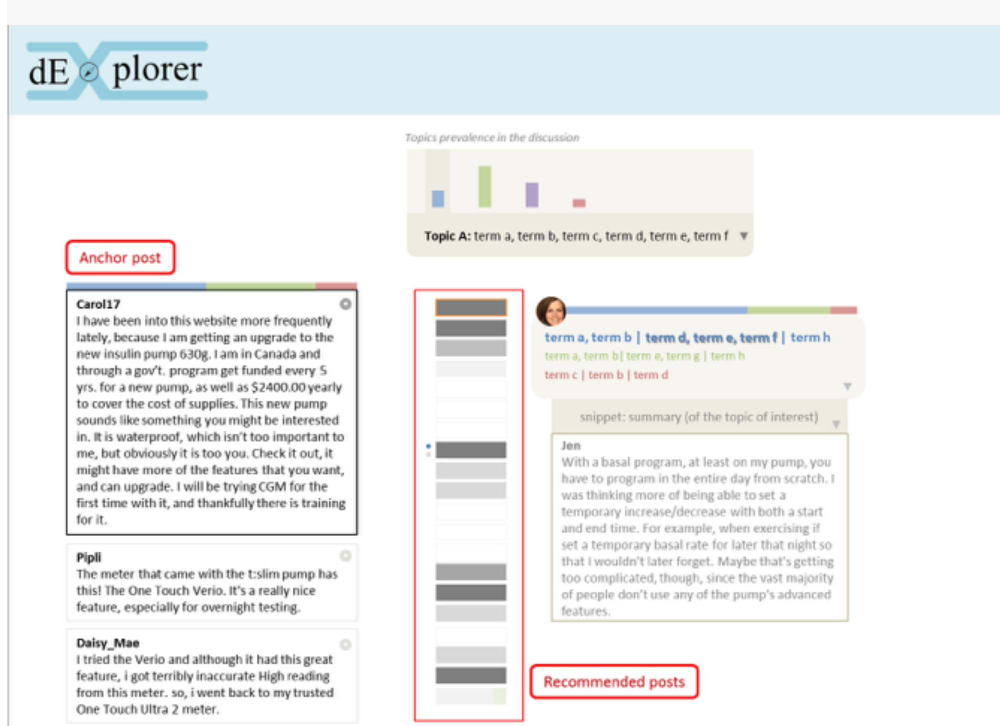
1. What could be potential ways to support forum discussions exploration in OHCs ?
2. Can we build a successful posts recommendation algorithm to support the guidance in OHCs discussion exploration ?

DESIGN PROTOTYPE

The Problem

- Online health communities (OHCs) forums are valuable repositories of vast amounts of information and opinions on health related issues and diseases.
- However, the units that store that information i.e. discussions, constantly increase in volume, go through topic shifts, arguments and heated debates around them, often interwoven with off-topic social interactions.
- Exploring this space of dispersed yet much relevant information to the user, intermixed with redundancy and noise is a challenging task.

Tool prototype design



dExplorer at Work

Supports discussion exploration through addressing three steps in the exploration process: *overview*, *guidance* and *focus*.

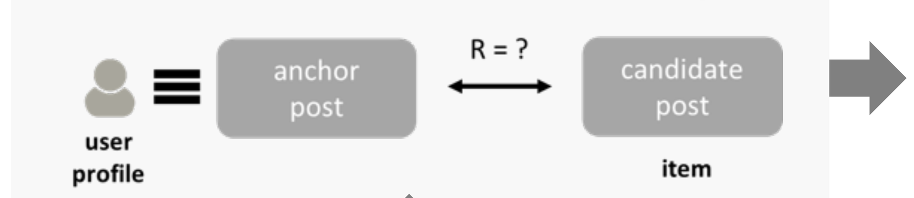
- Overview: high level visual summary of the discussion based on topics prevalence in the discussion and its distribution across different posts.
- <u>Guidance (A):</u> post recommendation algorithm that ranks the relevant posts with respect to an user selected anchor post at 4 levels of relevance, which are then visually encoded with different shades of gray, the darker being more relevant.
- Focus: verbatim posts' text as appeared in the discussion

SYSTEM MODELING

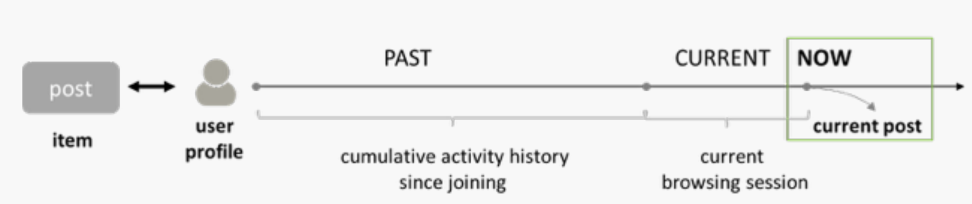
Posts Recommendation

- The user's profile is represented by the *current post of their interest*.
- The user's profile does not include:
  - previous posts interested in during the current exploration session.
  - any history of activity like: all the posts they have contributed, replied to or liked in the past, for example.
- Because INs change very dynamically due to: disease progression, comorbidities interactions in unclear ways, high variety of treatments and medications, leaving and returning into the OHC with different medical profile.

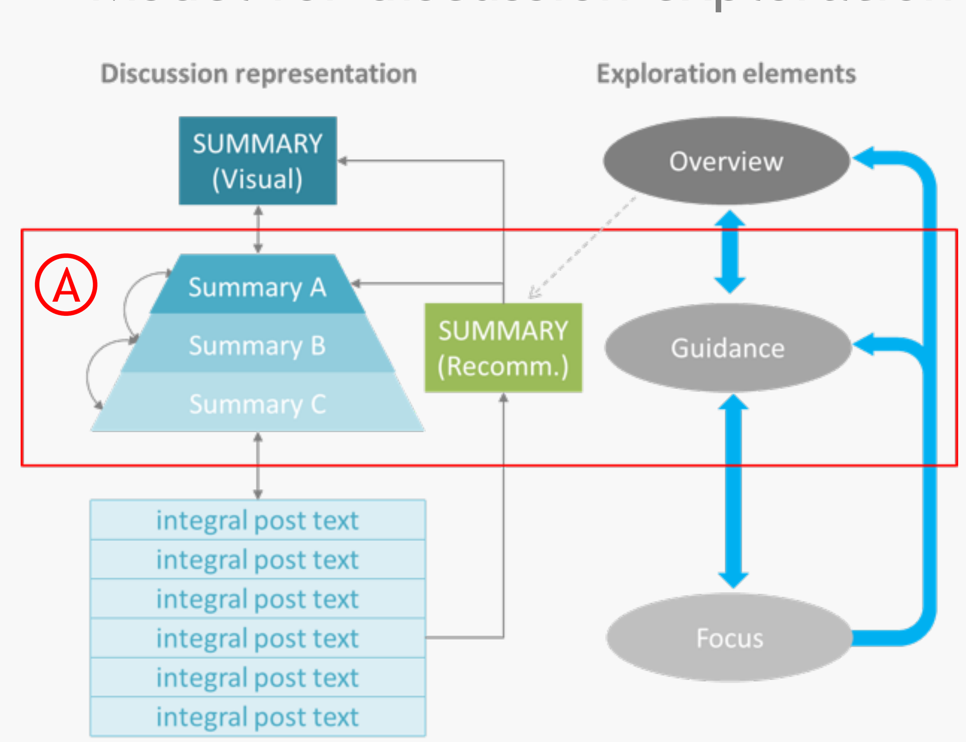
Relevance between two posts



Model for posts recommendation



Model for discussion exploration



Discussion Exploration

1. Overview – quick impression: deciding fast whether the discussion looks promising to satisfy some information need(s)

2. <u>Guidance (A)</u> – quickly increase the likelihood estimation of finding the information that can satisfy an information need(s)

3. Focus : once likelihood passes a satisfying threshold, get to the right information fast

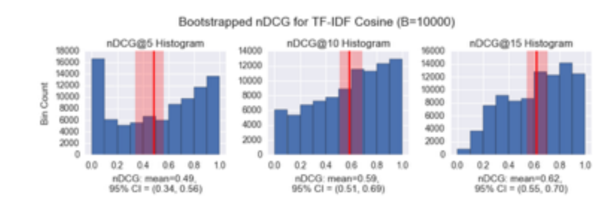
(A) explained:
There need to be multiple levels of summary abstractions between which the user can freely move. The posts recommendation feature functions primarily as a guidance mechanism for what could be potentially the next best steps in the exploration, but can also serve as a form of discussion summary.

MODELING POSTS AND RELEVANCE METRICS

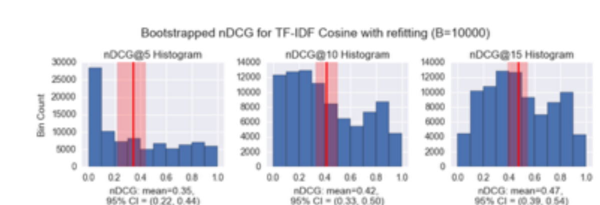
The Dataset

Source: Tudiabetes.org, an online health community for diabetes self-management
Training data: 20,404 discussions; 296118 posts; 14.51 posts/discussion; *sd* = 45.27
Test data: 3 discussions; 3 anchors per; 122 posts; 40.67 posts/discussion; *sd* = 6.55

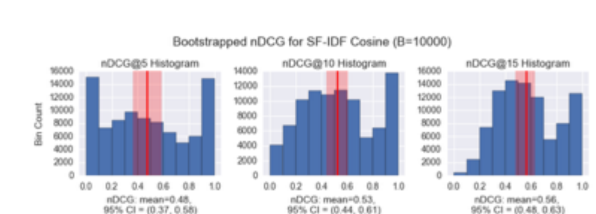
All models use bag-of-posts model of the discussion due to unintended use of imposed discussion structure and variation of forum platforms (forum affordances).



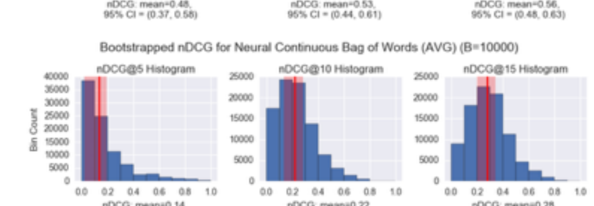
1.a TF-IDF Cosine (weighted Bag-of-Words)
Term-Frequency, Inverse Document Frequency



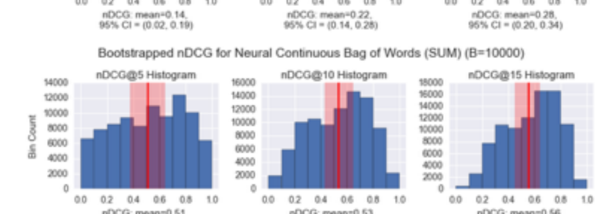
1.b TF-IDF Cosine with refitting
TF-IDF where IDF is recomputed for each discussion



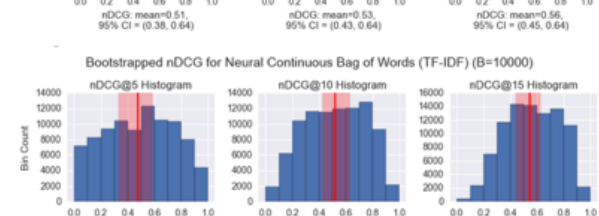
2. SF-IDF Cosine (Related Work)
Replace terms with their WordNet "synsets" for better approximation of word meaning



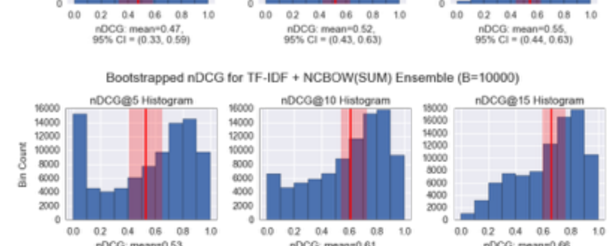
3.a NCBOW (AVG)
Represent document by average of GloVe word vectors, trained on Common Crawl



3.b NCBOW (SUM)
Represent document by sum of GloVe word vectors, trained on Common Crawl



3.c NCBOW (TF-IDF)
Represent document by TF-IDF weighted sum of Glove word vectors, trained on Common Crawl



4. TF-IDF and NCBOW (SUM) Ensemble
Committee of both classifiers. Most confident prediction is taken as decision.

MODEL	nDCG@5	nDCG@10	nDCG@15
Random Baseline	0.14	0.21	0.19
Neural Continuous Bag of Words (AVG)	0.14	0.22	0.28
LSA Cosine (latent dim=10)	0.31	0.33	0.40
LSA Cosine (latent dim=50)	0.32	0.36	0.43
LSA Cosine (latent dim=100)	0.32	0.38	0.43
TF-IDF Cosine with refitting	0.35	0.42	0.47
Neural Continuous Bag of Words (TF-IDF)	0.47	0.52	0.55
SF-IDF Cosine	0.48	0.53	0.56
Neural Continuous Bag of Words (SUM)	**0.51**	0.53	0.56
TF-IDF Cosine	0.49	**0.59**	**0.62**
TF-IDF and NCBOW (SUM) Ensemble	**0.53**	**0.61**	**0.66**

- Sum of pretrained dense word vectors (NCBOW (SUM)) has advantage that different words with similar meaning can be accurately compared for relevance, so topically similar posts match well.
- TF-IDF performs well when posts discuss the same topics exactly in part, but also discuss additional, different topics. These situations cause ambiguity for NCBOW because the differences of the topics are averaged.
- The ensemble performs better than both individual models because these are distinct situations: some posts are on topic but have variation in word use, while others discuss the same topic, but transition to other topics.
- Obtain crowd-sourced labeled data in large quantities to allow for semi-supervised models and more statistically robust evaluation
- Introduce a multi-term relevance measure (+overlap purity, +overlap importance, etc.) and train that model on the labeled data.

Conclusion

We believe there's room to utilize posts recommendation as a form of guidance for discussion exploration. Our method allows for relevant-post recommendation without relying on any form of supervision (labeled relevance or domain-specific ontologies, etc.) regardless of the discussion forum structure; it is a plug and play solution. Our best model shows promising results, but further evaluation is needed.