Automatic Task-Centric Organization of Analyst’s Documents using Instrumentation and Journaling data

Paul Jones, Shivani Sharma, Changsung Moon, Nagiza F. Samatova
Recommender & Big Data Analytics Groups, Computer Science Department, NC State University

Motivation

- Conducted survey on how analysts work with their computers to find information... responses indicated that most of them initially drop files in a default location (e.g. Desktop) and intend to sort them later but often don’t.

Problem Formulation

- Assume analysts are working in a collaborative, multi-tasking environment.
- Each user \( u \) accesses a sequence of documents \( d \). Each has task label \( y \) (which may be supplied by Journaling) with confidence \( c \).
- Problem is to predict \( y_{n+1} \) given a new document \( d_{n+1} \) and the sequence.

Network Fusion Approach

- Desired algorithm characteristics:
  - Handles heterogeneous associations,
  - Supports multiple users,
  - Probabilistic (provides confidence values),
  - Adaptable (handles dynamic add/delete/update),
  - Scalable from very small to very large graphs,
  - Ability to operate in a streaming environment.

Approach:

- Construct multi-layer graph, generate random walks on graph, and embed nodes using Skipgram with negative sampling.
- Add extra edges and nodes to try to improve the embeddings (i.e. by reducing random walk length between related nodes)
  - User-User edges, Task-Task edges, Doc-Doc edges
  - User||Task nodes, Task||Time nodes, User||Task||Time nodes

Workflow Dashboard Interface

- Recent-work dashboard interface developed in 2015 to organize documents according to the task(s) that they relate to, and save analyst’s time when resuming previous tasks.
- Previously, populated from Instrumentation and Journaling data only.
- Now, we can learn associations between documents and tasks in a semi-supervised manner.
- Currently active task at top of dashboard and highlighted in yellow.
- Collects feedback from analysts on automatically predicted task-document associations.
  - will enable dynamic updates to edges in the multi-layer graph and to confidence values on task labels.

Journaling Datasets

Tested using heterogeneous real-world workflows:

- Study A: CSC791 class
  - Feb-Apr 2015
- Study B: P5598 class
  - Oct-Nov 2015
- Study C: LAS analysts
  - Aug-Oct 2015

Evaluation Results

- With approx 30% labelled data, micro-F1>0.6 in all cases (e.g. Precision 0.9, Recall 0.5):
  - Overall accuracy >90% if 3-5 ‘guesses’ are allowed:
  - Small improvements explained by improved embeddings of concatenated Task nodes within the Document space, as shown by t-SNE plots:

Future Work and Potential Mission Impact

- Evaluation of a fully streaming (incremental) implementation.
- Improved embeddings by biasing the random walks (node2vec).
- Build recommender algorithms to ‘suggest’ documents from other users.
- Quantify time savings for mission analysts from use of the recent-work dashboard.