Large-Scale Modeling of Diverse Paths using Structured k-DPPs

Jennifer Gillenwater **Alex Kulesza** **Ben Taskar**
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
{jengi,kulesza,taskar}@cis.upenn.edu

Abstract

Large, interrelated document collections are becoming increasingly available. Tools like search have made these collections useful for the average person; however, search tools require prior knowledge of likely document contents to construct a query, and typically reveal no relationship structure among the returned documents. Often a more valuable result might be a small, structured set of documents that attempts to cover the content space of the collection. In this work we consider structure expressed in "threads", i.e., singly-connected chains of documents. For example, in response to a search for news articles from a particular time period, we might want to show the user the most significant stories from that period, and for each such story provide a timeline ("thread") of its major events.

Following that intuition, we formalize collection threading as the problem of finding diverse paths in a directed graph, where the nodes correspond to items in the collection (e.g., news articles), and the edges indicate relationships (e.g., word distribution similarity). A path in this graph describes a thread of related items. A diverse set of high-quality paths then forms a cover for the most important threads in the collection.

The task of finding a diverse set of high-quality paths in a graph is relatively novel, but many of its subtasks have been addressed in previous work. The largest body of related material is centered around the Topic Detection and Tracking (TDT) program (Wayne, 2000). Relevant examples of TDD-related work include that of Blei and Lafferty (2006) and Leskovec et al. (2009). In addition, there has been some previous work on selecting a *single* thread in a graph, given some prior knowledge about the thread content (Shahaf and Guestrin, 2010). Other relevant work in the single-thread vein includes that of Chieu and Lee (2004), Nallapati et al. (2004), and Mei and Zhai (2005).

In this work, to model sets of paths in a way that allows for repulsion (and hence diversity), we employ the structured determinantal point process (SDPP) framework (Kulesza and Taskar, 2010), incorporating k-DPP extensions to control the size of the produced threads (Kulesza and Taskar, 2011). The SDPP framework provides a natural model over sets of structures where diversity is preferred, and offers polynomial-time algorithms for normalizing the model and sampling sets of structures.

However, even these polynomial-time algorithms can be too slow when dealing with many real-world datasets, since they scale quadratically in the number of features. We address this problem using random feature projections, which reduce the dimensionality to a manageable level. Furthermore, we show that this reduction yields a close approximation to the original SDPP distribution, using a result of Magen and Zouzias (2008).

We demonstrate our model using two real-world datasets. The first is the Cora research paper dataset, where we extract research threads from a set of about 30,000 computer science papers. The second is a multi-year corpus of news text from the New York Times, where we produce timelines of the major events over six month periods. We demonstrate the superiority of our method over multiple baselines using human-produced news summaries as references.

References

D. Blei and J. Lafferty. 2006. Dynamic topic models. In *Proc. ICML*.

H. Chieu and Y. Lee. 2004. Query based event extraction along a timeline. In *Proc. SIGIR*.

A. Kulesza and B. Taskar. 2010. Structured determinantal point processes. In *Proc. NIPS*.

A. Kulesza and B. Taskar. 2011. k-DPPs: fixed-size determinantal point processes. In *Proc. ICML*.

J. Leskovec, L. Backstrom, and J. Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proc. KDD*.

A. Magen and A. Zouzias. 2008. Near optimal dimensionality reductions that preserve volumes. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 523–534.

W. Mei and C. Zhai. 2005. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proc. KDD*.

R. Nallapati, A. Feng, F. Peng, and J. Allan. 2004. Event threading within news topics. In *Proc. CIKM*.

D. Shahaf and C. Guestrin. 2010. Connecting the dots between news articles. In *Proc. KDD*.

C. Wayne. 2000. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Proc. LREC*.