Improved Algorithms for Topic Distillation in a Hyperlinked Environment
Krishna Bharat, Monika R. Henzinger

The core idea of this paper is to find authoritative web pages on broad topics. It introduces algorithms for topic distillation, focusing on how to identify hubs and authorities, focused on improving on Kleinberg's HITS algorithm from Lecture 4, specifically to overcome mutual reinforcement issues with that paper. The goal of topic distillation in this context is to identify a small set of hubs and authorities for a given topic from a large web graph without selecting irrelevant pages that manipulate rankings through things like mutual reinforcement. The new approach to the HITS algorithm includes weight normalization and fractional edge weights. This prevents a single website from dominating the rankings as well reducing the effects of mutual reinforcement. They also include an experiment in the paper, showing that the modified algorithms improve precision by reducing irrelevant authoritative pages. The authors assume that hyperlink structures can be leveraged to identify authoritative sources. In other words, they assume links represent endorsements, meaning if many pages link to a specific document, it indicates that document's importance. They also assume that documents from the same host are authored by a single entity, influencing the computation of hub and authority scores. They also assume that mutually reinforcing relationships between pages distort true authority scores, leading them to propose adjustments like fractional edge weights. One candidate extension to this paper could be to include supervised learning models that predict authority based on a combination of link structure, relevance, and user interaction metrics. Some of the strengths of the paper are the identification of key issues in Kleinberg's hub and authorities algorithm, building off of previous research, and proposed modifications such as addressing mutually reinforcing links and cycles. However, the paper doesn't scale well as the response time could get up to 30 minutes in some cases depending on the nature of the data fetching.

Analysis of a Very Large AltaVista Query Log
Craig Silverstien, Monika Henzinger, Hannes Marais, Michael Moricz

This paper presents an analysis of web search behavior empirically, using a relatively large query log from AltaVista with the goal of understanding how users search. The authors focus on patterns like query lengths and session behaviors. The key findings challenge the assumptions from information retrieval research (hubs and authorities). The authors found that most queries are short and are repeated many times. The authors found that about three quarters of sessions consist of only one query, suggesting users are typically seeking information quickly, not pouring deep into the depths of a particular topic. The authors then used a chi-squared test to find relationships between terms, concluding that users often type phrases without quotation marks and that correlated search terms improve search relevance. They also found that users rarely refine queries and that those who do usually just add or delete terms. They also found that users focus on the first page of results, rarely clicking beyond the first 10 results. What this means is that HITS which assumes long, complex queries oftentimes, doesn't reflect real web search behavior, nor is it conducive to users' behavior with fast searching as they do not want to wait for an algorithm to run, they'd rather receive results quickly at the compromise of some relevance. The authors assume that a gap of more than 5 minutes between queries from the same user indicates a new search session. They also assume that cookies provide an accurate way to track unique users. Additionally, the dataset is assumed to be representative of typical web search behavior despite being specific to AltaVista users during a certain period. A possible extension to this paper would be the inclusion of some discussion of real-time data to observe how query patterns evolve with breaking news for example. The strengths of this paper are the large scale of the dataset using real-world search engine logs as well as exploring possible limitations and edge cases like duplication, modification patterns, etc. However, the paper only focuses on AltaVista, failing to capture the entire web search population. Additionally, the 5 minute cutoff is practical but likely fails to accurately reflect the user's intentions in every case.