Project 3: PageRank

# Members

1. **Krittin Chatrinan** ID 6088**022**
2. **Anon Kangpanich** ID 6088**053**
3. **Tanawin Wichit** ID 6088**221**

BONUS CREDIT REPORT

# OVERVIEW

This report contains a summary of a research paper which related to the application of PageRank algorithm. The research paper we will mention in this report is [**Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search.**](http://infolab.stanford.edu/~taherh/papers/topic-sensitive-pagerank-tkde.pdf)

# RESEARCH PAPER SUMMARY [1]

Traditional simple search engines ranks search results using simple similarity such as TF-IDF, Jaccard, and BM25 similarity; however**, these methods does not measure importance metrics such as Authenticity and Authority of each pages.** With the creation of PageRank algorithm which incorporates both metrics into its calculation, it allows people to create tools to rank importance of webpages; on the contrary, PageRank itself was not designed to rank webpages based on given queries or user’s information needs. Therefore, the researcher of this research paper tried to incorporate both relevance and importance ranking. There are many other research papers that tried to achieve the same goal; however, some of those improved PageRank by simply building a system in top of existing PageRank algorithm.

In this research, PageRank has been improved in such a way that it works well in web searching with relevant ranking. Therefore, this appoarch needs to sense topic of the user query. Therefore, the appoarch has 2 phases: **PageRank processing and query-time processing.**

**In this PageRank appoarch**, the original PageRank algorithm has been improved. Starting from webpage fetching process, each page is being classified into predefined categories from Open Dictionary Project. For example, an online shopping page is being classified as SHOPPING at a very high value, while it may be classified as NEWS or most other classes with a very low value. In the end, suppose we define 16 topics, a page would have scores for all 16 topics.

**During query-time**, after receiving a query from a user, the query will be classified using multinomial naïve Bayes. For each page, all scores from various topics then taken into account using conditional probabilities and later forming a query-sensitive importance score. The final search results are eventually sorted by the score in decending order.

The evaluation process of the research is using Web crawl from the Stanford WebBase which contains approximately 120 million pages. 280,000 URLs from Open Dictionary Project (ODP) and 35 sample queries were also used for the evaluation. Due to the fact that this Topic-Sensitive, PageRank involves with many parts that can be evaluated such as rankings between ordinary PageRank and Topic-Sensitive PageRank. **As for Query-Sensitive Scoring, precision measure is used**. **The researcher evaluated performance of their results using Precision at top-20 results.** The searcher evaluation was conducted by using five human volunteers. All volunteers were asked to mark all URLs that are relevant. Furthermore, they also were asked to mark which ranking is better. **In the end, out of 10 queries, the majority of users preferred Topic-Sensitive PageRank for the majority of queries. Overall, the Topic-Sensitive PageRank substantially yields better mean precision than traditional PageRank: 0.512 versus 0.276.** We have observed that the researcher decided not to use Recall in the evaluation because Recall involves the knowledge of total pages in the same categories which is not practical to achieve.

# TECHNIQUE APPLICATION TO A NOVEL PROBLEM

Our novel problem that suitable for Topic-Sensitive PageRank is **Topic-Sensitive Research Paper Citation Ranking.** Due to the fact that research papers are diverse in terms of categories and importances, similar to webpages, sometimes **searching for papers, which are reliable, relevant and fundamentally solid in terms of backgrounds can be painstaking and time consuming by requiring users to skim and scan several documents from search results.** Additionally, since research papers usually contain several citations some of which are more important than others, similar to hyperlinks in webpages; therefore, we can apply the same Topic-Sensitive PageRank methodology for this problem. To achieve the goal, we have to modify Topic-Sensitive PageRank in such a way that it can gather and read research papers to initialize PageRank process. Each research paper will be categorized into predefined topics and given scores accordingly. The modified PageRank could be used extensively in a academic research search engine**. When users search for result**, they will have an option to select the main topic of interest. When the searcher is given a query and the main topic, the result will be tailored with relevant and topic scoring in mind.

# REFERENCE

|  |  |
| --- | --- |
| [1] | T. H. Haveliwala, "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search," Stanford University, Stanford, 2003. |