Using Documents to Search for Documents

Kyle Williams[1], C. Lee Giles[1,2]
Information Sciences and Technology[1], Computer Science and Engineering[2]
The Pennsylvania State University, University Park, PA, 16802, USA

Introduction

The need to find similar documents arises in many situations:

•Plagiarism detection
•Near duplicate detection
•Research paper recommendation

Traditionally, queries are constructed by users and submitted to search engines; however, it may not be obvious to users how they should construct such queries or overly complex to do so. Thus querying with a whole document and allowing a system to automatically construct queries can be useful.

This research investigates methods for automatically searching for similar documents using a sample document while dealing with issues such as scalability and multiple notions of similarity.

A Model and Algorithm for Similarity Search

Searching for similar documents with a sample document may not return all similar documents since it's possible for similar documents to not have any features in common. Swanson's ABC model [1] provides inspiration for a recursive search algorithm to overcome this shortcoming.

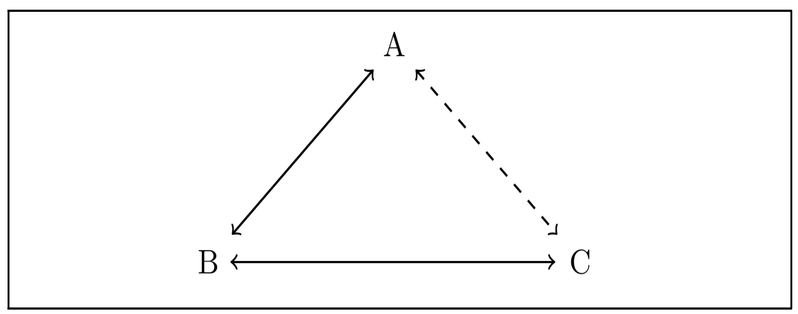

Figure: The ABC model. If A is related to B and B is related to C, then A may be related to C.

The search algorithm based on this model is:

1. Search with initial query document
2. Take the top *k* results and use them to search for more documents
3. Repeat step 2 a set number of times
4. Combine all results
5. Rank each result either by similarity to the initial query (target ranking) document submitted by the user or by the similarity to the query document that found it (local ranking)

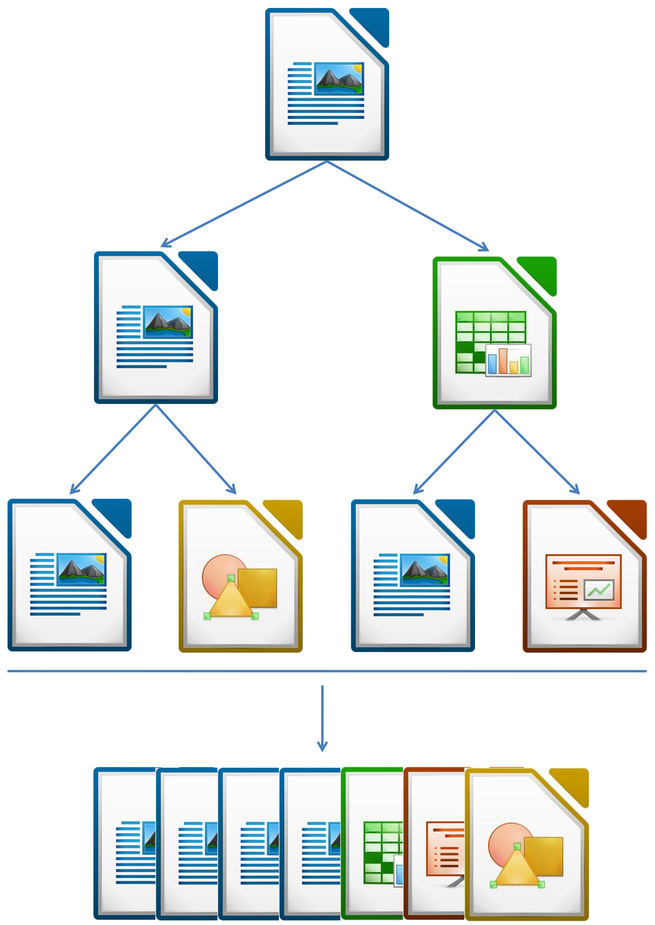

Figure: The recursive search strategy uses search result to find more results and then combines them.

SimSeer: A Similarity Search Engine

SimSeer [2] is a similarity search engine that incorporates the recursive similarity search algorithm. It currently supports three similarity functions, though it is possible to add more. They are:

•Key phrase similarity
•Shingle similarity [3] based on sequences of words in a document.
•Simhash similarity [4] based on all of the words in a document.

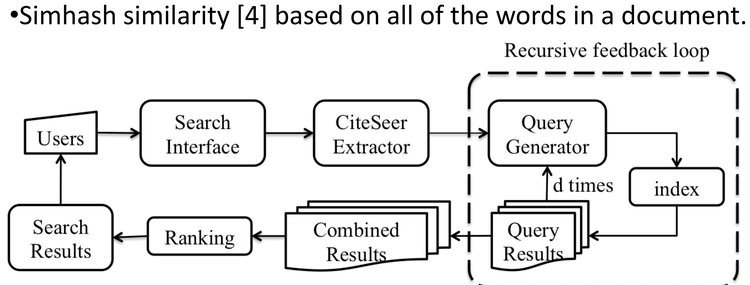

Figure: The SimSeer workflow.

Experiments

Experiments were performed on 3 datasets: documents belonging to the same category; fake computer science papers generated using the SCIGEN tool; and documents plagiarized from the Web. N = number of results, H = recursive depth, D = divider.

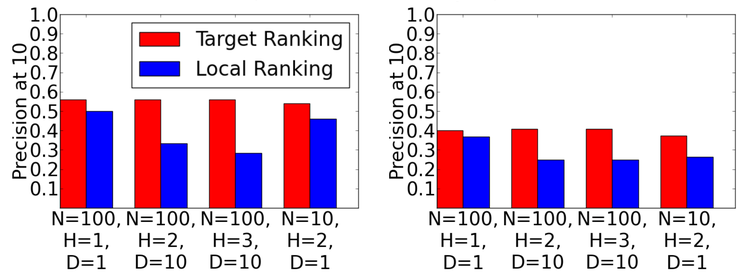

Figure: Same category search using key phrases and shingles

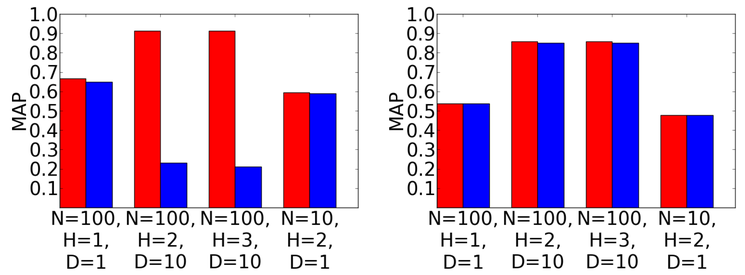

Figure: Fake paper search using key phrases and shingles

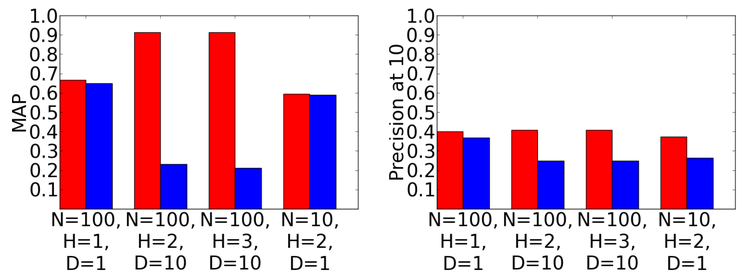

Figure: Plagiarism search using key phrases and shingles

Acknowledgments

We gratefully acknowledge support from the National Science Foundation under grant No. 1143921 and useful suggestions on evaluation by Dr. Hung-Hsuan Chen. Icons by The Document Foundation used under CC-BY-SA 3.0 Unported license.

References

[1] D. R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect. Biol. Med., 30(1):7–18, 1986.
[2] K. Williams, J. Wu, C. Lee Giles. 2014. SimSeer: A Similarity Search Engine. Under Review
[3] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. 1997. Syntactic clustering of the Web. Comput. Netw. ISDN Syst. 29, 8-13, 1157-1166.
[4] M. S. Charikar. 2002. Similarity estimation techniques from rounding algorithms. In Proceedings of the Symposium on Theory of computing, 380-388.