

# FINA- A Fast Index News Archive

Mariana Barca, Guadalupe Canahuat, Laura Stoia and Rodi Tountcheva

Department of Computer and Engineering Science

The Ohio State University

2015 Neil Ave, 395 Dreese Labs

Columbus, Ohio USA

barca, canahuat, stoia, tountcheva@cse.ohio-state.edu

## Abstract

This paper presents the Fast Index News Archive (FINA) system, a content based web-search engine over a database of news articles. The system uses Latent Semantic Indexing (LSI), which allows us to retrieve articles similar in meaning but not necessarily containing the keywords that have been input in the query. We discuss the challenges that arise from using a large database of articles and our solutions to these challenges.

## 1 Motivation

The recent growth of the World Wide Web has provided researchers with a rich resource of data. Organizing and utilizing this data is still a big challenge for the research community. An example of such data are news articles. Daily, thousands of news articles are published by various web sites. The purpose of the FINA project is to gather these articles in a huge database and provide a fast search engine based on content similarity. The database will be kept up to date by a web crawler that gathers the news that have been published recently. The friendly web interface allows the users to do various searches, including title, dates, keywords and meaning similarity. Searches within results are also enabled.

The development of the FINA project addresses a lot of research questions from various fields (computational linguistics, database design, efficiency issues) and its implementation provides a useful real-life application.

## 2 Our work

We have implemented an initial version of FINA that is fully functional but uses a relatively small set of news articles collected from the web using a web crawler made for this project. Currently, the database has over 10 thousand records. Latent Semantic Indexing algorithms

have been applied for this data and specific tables and procedures that employ LSI searches have been implemented. The web interface provides searches for both keywords and meaning-related articles. The algorithms have been design for efficiency with large datasets, and our future work involves populating our database with more articles while dealing with the problems that arise with large datasets.

## 3 Conclusions

The FINA project applies algorithms from various fields of computer science to a real-world problem. Its friendly user interface and exciting research questions involving retrieving documents with similar meaning, evaluating document similarities and database design for efficiency would inspire further research in these areas.