CS 410: Technology Review

7 November 2021

The Lemur Project

The Lemur Project was started in a collaborative effort between the Center for Intelligent Information Retrieval at the University of Massachusetts and the Language Technologies Institute at Carnegie Mellon University. Started in 2000, the Lemur Project develops software tools, search engines, and resources that support the research and development of information retrieval and text mining. Throughout the years, the Lemur Project has developed a number of search engines, datasets, and libraries that are widely used in both scientific studies as well as in research.

Indri is one of the most notable search engines developed and released by the Lemur Project. This state-of-the-art search engine provides text search and structured query language with capacity for large text collections. The query language allows to index data or structure documents using simple command line instructions with the ability to query millions of documents. This search engine was built to support various programming and scripting languages including C++, Java, C#, and PHP as well as the ability to parse various different document types. Other notable features of this search engine include its efficiency in implementation, effectiveness, and ability to make use of multiple document representations.

Galago in another search engine that has been developed and released through the Lemur Project. This search engine is a great tool for experimentation with text search as it is based on small components that are easily interchangeable during both indexing as well as retrieval. Galago is built in such a way that allows users to control nearly all aspects of the system through specifying configurations and also enables users to modify and even implement their own classes to incorporate into the system. Due to its ability to support experimentation, the Lemur Project has extended the Galago Search Engine and has since modified many components in order to support experimentation.

RankLib is a library released by the Lemur Project of learning-to-rank algorithms. Currently, RankLib houses eight algorithms that have been implemented. These algorithms are MART (Multiple Additive Regression Trees), RankNet, RankBoost, AdaRank, Coordinate Ascent, LambdaMART, ListNet, and Random Forests. As well as implementing these algorithms, RankLib also implements retrieval metrics as well as provide various ways to carry out evaluations.

These are just the tip of the iceberg of all the different tools, libraries, and softwares that the Lemur Project has developed and released to the public. The Lemur Project is an ongoing initiative and still lists a number of active contributors, mainly from the two Universities that initially started the project.

Sites used

Galago Documentation (sourceforge.io)
Lemur Project Home
Lemur Project - Wikipedia