10.06 Information Retrieval Toolkits

NLP

Introduction to NLP

Information Retrieval Toolkits

Open Source IR Toolkits

- Smart (Cornell)
- MG (RMIT & Melbourne, Australia; Waikato, New Zealand),
- Lemur (CMU/Univ. of Massachusetts)
- Terrier (Glasgow)
- Clairlib (University of Michigan)
- Lucene/SOLR (Apache)

Smart

- The most influential IR system/toolkit
- Developed at Cornell since 1960’s
- Vector space model with lots of weighting options
- Written in C
- The Cornell/AT&T groups have used the Smart system to achieve top TREC performance

MG

- A highly efficient toolkit for retrieval of text and images
- Developed by people at Univ. of Waikato, Univ. of Melbourne, and RMIT in 1990’s
- Written in C, running on Unix
- Vector space model with lots of compression and speed up tricks
- People have used it to achieve good TREC performance

Lemur/Indri

- An IR toolkit emphasizing language models
- Developed at CMU and Univ. of Massachusetts in 2000's
- Written in C++, highly extensible
- Vector space and probabilistic models including language models
- Achieving good TREC performance with a simple language model

Lucene

- Open Source IR toolkit
- Initially developed by Doug Cutting in Java
- Now has been ported to some other languages
- Good for building IR/Web applications
- Many applications have been built using Lucene (e.g., Nutch and SOLR)

NLP