The Draft Design Issues of iZENELib

Yingfeng Zhang, Kevin Hu

November 10, 2008

Abstract

This document presents the draft design issue of the project *iZENELib*, especially focuses on the information retrieval subpart.

Contents

1 Design Goal

iZENELib plans to provide a collection of utilities which could be used in the search engine developing process. *iZENELib* is expected to be composed of two parts—the part taking charge of storage(AM-Lib), and the part in charge of information retrieval and machine learning(IR-Lib). Generic design would be adopted largely in *iZENELib*.

AM-Lib is expected to be composed of two sub-parts, the one which provides common utilities for data storage, and the one providing corpus data management which serves for the IR-Lib.The former one has been involved by Kevin's initial report, therefore, more details about it would not be included in this document.

2 Corpus Management

Corpus management component of AM-Lib provides storage services for IR-Lib. It is necessary because each component of IR-Lib will read data from corpus and output the results to a generic structs, therefore refactoring this common part into a single library will improve the system's reusage.

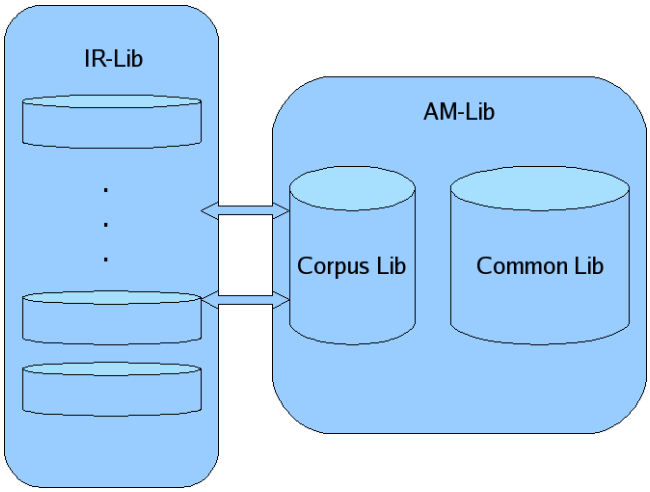

Figure 1: AM-Lib and IR-LIB

Existing project as *SML* provides similiar components as DocumentBag,etc. We can base corpus management on SML. What's more, IR-Lib needs a powerful matrix component to store the middle temporary computation outputs and the ultimate results. Through research, such a matrix component is expected to be included in *iZENELib*:

- A general matrix that has both memory and file version.

  Memory version of matrix has been provided by *Boost*, and file version has been implemented by *Clustering-framework*. Therefore it is necessary to provide the general matrix with an improvement on them.

- SVD Decomposition, QR Decomposition, LU Decomposition, Eigenvalue Decomposition.

  These utilities are extremly useful in machine learning and information retrieval.

- Hessian matrix

  Hessian matrix is useful in Bayesian inference and decision theory.

During the developing process of *iZENELib*, more components of matrix would perhaps be included if necessary.

3 Components of IR-Lib

IR-Lib is a collection of algorithms in machine learning and information retrieval, together with the Corpus Management component of AM-Lib, it could provide a generic framework for search applications. Both machine learning and information retrieval have covered lots of fields, therefore, the main purpose of IR-Lib is to provide a scalable framework together with general algorithms, then in future, more advanced algorithms could be added easily.

The relationship between machine learning and information retrieval is very close and machine learning could be seen as the lower layer to provide methods for information retrieval's usage. Therefore IR-Lib could be composed of two layers. In addition, there exists some fields in information retrieval that has not adopted methods provided by machine learning, such as recommendation systems, preprocessing, etc. We will talk about all the components of IR-Lib one by one.

4 Machine learning Components

- Supervised Learning.
  Wisenut-classifier has implemented most of the basic supervised learning methods. We plan to replace the interface to *SML* with the interface to new corpus management component of AM-Lib, and then refactor it to the generic design.

- Unsupervised Learning.
  Clustering-framework has already doen a good job of it. Therefore, the relavant job of this component is to make *Clustering-framework* suit for the whole framework of IR-Lib.

- Learning Complex Models.

  1. EM(Expectation—Maximization), which is also a basic learning approach in semi-supervised learning.
  2. Hidden Markov Models.
  3. Sampling method, including MCMC.
  4. Graphical Models, graphical models including following directions, each of which is under hot research, we are not sure whether it is possible to implement all of them, just try to do that.
     (a) Bayesian Network.
     (b) Markov Random Fields.
     (c) Conditional Random Fields.

- Dimensionality Reduction. We plan to implement PCA at first, more approaches could be done in future if possible.

In summary, machine learning are still under fast developing process, therefore only some basic directions would be included into this library, we hope a good design framework could be provided in order that more learning approaches could be included into this library easily in future.

5 Information Retrieval Components

- Text Pre-Processing
  Text pre-processing techniques are mature and have been implemented by existing projects, therefore we can refactor them from existing code.

  1. Stopword Removal

2. Stemming
3. TF-IDF
4. Tokenization
5. Feature Selection
6. Duplicate Detection

- Language Models
  It is necessary to refactor and integrate Jinglei's work into the library.
- Topic Modeling
  Topic modelling is a hot research direction and lots of new approaches appear continuously. We only plan to provide some topic modelling methods including LSI, LDA and 4-level PAM. We hope more topic modeling approaches could be easily added to this library.

6 Schedule

Kevin and I will take main charge of implemeting *iZENELib*, however, since both of us have other projects to maintain and therefore we could not guarantee the certian time to finish this project. We hope to finish it before Spring Festival at the end of Jan,2009.

MileStone	Finish Date	In Charge	Description
AM-Lib	2008-12-15	Kevin	AM-Lib has been finished
Basic IR-Lib	2008-12-15	Yingfeng	Basic IR-Lib could be provided, including corpus management, basic supervised and unsupervised learning.
Pre-Processing	2008-12-31	Kevin	Text pre-processing component is encapsulated.
Complex machine learning	2009-01-15	Yingfeng	Learning methods for complex models have been finished.
Language models and topic models	2009-01-31	Yingfeng	Finish IR-Lib.