IITB-Monash Research Academy

Indian Institute of Technology Bombay | MONASH University

An Indian-Australian research partnership

Project Title:	**Speeding up complex analytics and natural language processing tasks using the inverted index**
Project Number	**IMURA0262**

Monash Supervisor(s)	Prof. Mark James Carman, Prof. Yuan-Fang Li	*Full names and titles*
Monash Primary Contact:	mark.carman@monash.edu, yuanfang.li@monash.edu	*Email, phone*
IITB Supervisor(s)	Prof. Ganesh Ramakrishnan	*Full names and titles*
IITB Primary Contact:	ganesh@cse.iitb.ac.in	*Email, phone*

Research Academy Themes:

Highlight which of the Academy's Theme(s) this project will address?

(Feel free to nominate more than one. For more information, see www.iitbmonash.org)

1. Advanced computational engineering, simulation and manufacture
2. Infrastructure Engineering
3. Clean Energy
4. Water
5. Nanotechnology
6. Biotechnology and Stem Cell Research

The research problem

Information Extraction and analytics are emerging as key enabling requirements for search based on deeper semantics: for example, a search on 'John's address', that returns matches to all entities annotated as an address that co-occur with 'John'. A dominant paradigm adopted by NLP processes rule-based named entity annotators is to annotate a document at a time. The complexity of this approach varies linearly with the number of documents and the cost for annotating each document, which could be prohibiting for large document corpora. The research problem is to scale up the statistical NLP processes used for Information Extraction and parsing problems to web-scale, with no or minimal loss of semantics.

In 2006 we proposed an alternative paradigm for rule-based entity annotation, which operates on the inverted index of a document collection and achieves an order of magnitude speed-up over the document-based counterpart. In addition the index based approach permits collection level optimisation of the order of index operations required for the annotation process. In 2007-2008, we developed a polynomial time algorithm that, based on estimated cost, can optimally select between different logically equivalent evaluation plans for a given rule.

Project aims

We believe that this work has only explored the tip of the iceberg. As a part of some informal class projects at IITB, we explored the translation of parsing using CYK algo for CFG and the probabilistic MAP inference problems for HMM to operations either purely on the inverted index or a combination of the inverted index along with some auxilliary data structure (such as the Next-word index, etc.), to speed up the extraction of syntactic parse trees and part of speech tags respectively by at least an order of magnitude. Very preliminary initial experiments have been extremely encouraging.

In this project, we aim to formally address the speed-up of annotation processes using Hidden Markov Models, Conditional Random Fields, Context Free Grammars, Probabilistic Context Free Grammas, Dependency Grammars, etc, using operations on the inverted index.

Expected outcomes

1] New data mining techniques for scalable NLP
2] New data structures for enabling web-scale analytics
3] 2 PhD student completions
4] Conference and Journal Papers
5] Software for scaling up NLP and analystics

How will the project address the Goals of the above Themes?

This project directly addresses the theme of Advanced Computational Engineering

---

Capabilities and Degrees Required

Btech or Mtech in a good university with excellent capabilities in:
1] Programming
2] Handling systems
3] Data structures and algorithms
3] Good inclination toward statistics and probability theory