Building A Search Engine

By Sahil Mutneja
under the mentorship of Dr. Dileep A.D.

What is a search engine?

- Basically a type of program that uses keywords to search for documents that relate to these keywords
- Results found are then put in the order of relevance to the topic that was searched for
- Examples
  - Google
  - Alta Vista

Importance of Search Engines

- Web graph constitutes about 8 billion web pages.
- Searching for a specific information on this scale is almost impossible
- Search Engines filter the wide range of information present on the Internet into something meaningful
- The final results are then easily accessed and used by the users within the matter of seconds.

Basic Problem Statement

- Search Engine that provides the user with a list of experts when queried against a specialization
  - The experts will be from the premium Institutes(IIT’s, NIT’s) spanning across India
- The list will be such that the most relevant of the result will be towards the top.

Final Outcome

- The result will comprise of the web pages of the specialist, relevant papers, top-notch work done on the queried specialization and much more.
- Structure of the formed web graph/network will be such that the nodes will be representing the specialization and edges will be the specialist.

Description of the project

Decomposing the Problem Statement gives us three subparts:

1. Web Crawling
2. Indexing
3. Searching

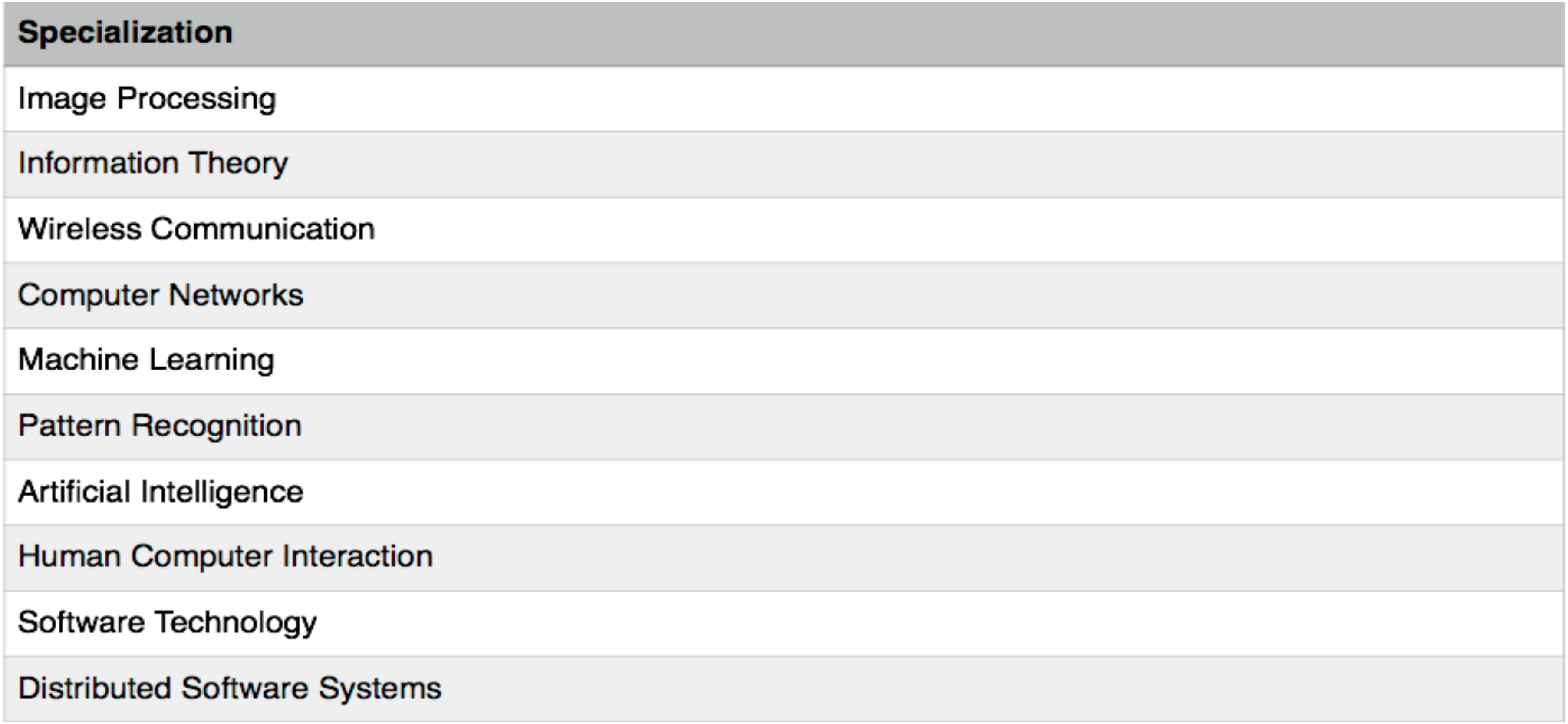
Web Crawling and Parser

- Crawler task is to browse the web, download the documents (pages) and save them.
- It will start from a seed page and extend in such a manner that all the web pages related to the experts are crawled.
- All the pages crawled will then be downloaded and saved for further processing.
- All the fetched content will be in an unorganized manner which for further processing needs to be fixed via Parser
- Parser will convert the unorganised content into a format which will be suitable for Indexing.

Indexer

- For the initial testing purpose, a list of specialization is formed which will form the basis of Indexing.
- Structures will be created by the Indexer after processing the documents passed to it by the crawler.
- Various text processing algorithms will be used in order to implement it with high precision.
- After the final processing we will be having a dictionary kind of data structure with
    - key as the specialization
    - value as the list of documents with key given as the prime importance.

Specialization for the purpose of Indexing



Specialization
Image Processing
Information Theory
Wireless Communication
Computer Networks
Machine Learning
Pattern Recognition
Artificial Intelligence
Human Computer Interaction
Software Technology
Distributed Software Systems

Ranker and Searcher

- Ranker is the part that arranges the documents in some particular order
- The main aim here is to put the most relevant documents on top of the results page
- Some of the famous ranking algorithms namely PageRank or Hubs and authorities can be used to rank the pages according to their merit
- Searcher takes as input the query entered by the user, transform and parses it somehow and sends it for further processing
- After the initial processing it finds the documents containing words of the transformed query via the help of indexer and finally display the results

Course of Action with Time Frame (7th and 8th Sem included)

Work to be Completed	**Tentative Date**
Build a Crawler	11th October 2014
Formatting via Parser	7th November 2014
Research on ML Algorithms, Indexer and Algorithms for Ranker	In Winter Vacations
Document Classification and Indexer	10th March 2015
Pre-Requisites for Ranker to be in place	21st March 2015
Implementation of Ranker and Searcher	17th April 2015
Further Enhancement and Improvements	28th April 2015

Work Carried Out Till Now

- Almost done with the Crawler
  - Build a program that crawls around a 1000 pages, covering almost all the experts from all the premium institutes
  - Testing of all the downloaded/crawled pages is in process on the local machine
- Started working on the Parser
  - Working on formatting the content into some meaningful data
- Analysis and testing of various ranking algorithms, namely PageRank, Topic Sensitive PageRank and Hubs & Authorities already done with more than 90% accuracy.