RESEARCH STATEMENT

Rao Muhammad Umer (muhammad.umer@cs.uol.edu.pk)

My primary research interest lies at the intersection of machine learning and data science. Accordingly, my work has primarily focused on utilizing statistical machine learning models to extract semantic and syntactic information from plain text and to learn structure of web forms from unstructured data. I have a particular interest in problems of unstructured data to extract useful information. This interest is motivated by the fact that current data extraction and knowledge based systems support limited number of domain specific information and are hard to process with large amount of increasing web data. It is of great value to devise new learning paradigms, which minimize the costs of building applications for all human knowledgeable domains databases from Big Data.

Current Work

In my MS thesis, I proposed the **Deep Web Extractor (DWX)** system using statistical machine learning models for crawling and data discovery from the Deep Web (i.e., massive and quality portion of World Wide Web) to build knowledge based databases. The main objectives performed by this system as given below:

1. *To discover and extract the deep web's content of quality for web searchers.*
2. *To discover automated means for identifying search-able web form interfaces and directing queries to them to dig out information.*
3. *To build domain specific data repositories (e.g. real estate, newspapers, health, etc.) for purposeful analysis and building knowledge base databases.*
4. *To handle the complex queries, like queries containing different range values, not entertained by traditional search engines.*
5. *To facilitate Law and Enforcement Agencies to detect Fraudulent web user.*

The proposed architecture of Deep Web Extractor (DWX) system is shown in Figure 1:

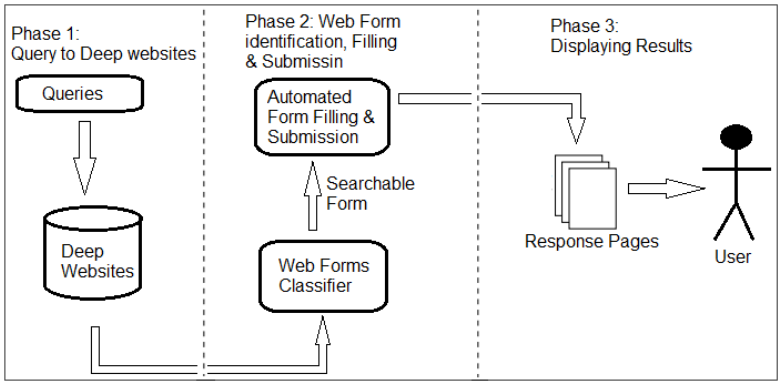

Figure 1: Deep Web Extractor System

Future Directions

Despite the achievements that have been already made by machine learning, particularly in information retrieval and data mining, I believe we have only scratched the surface of what intelligent systems could develop into. Next, I wish to address the issues that should be top priority for us to develop solutions for in the near future.

Recommendation Systems

Knowledge is encoded in a diverse set of formats; images, videos, text, speech, and databases. We take into consideration this prior knowledge when we take actions and make decisions.
Integrating these modalities into a single unified framework though trivial for a human is immensely difficult for a computer. It is possible to embed social networks and knowledgeable databases to build better predictors and annotators.

For example, integrating a **knowledge base (IMDB) database with a social graph** (your friends) will enable us to build smarter recommendation systems. We may discover that a set of movies are popular amongst your friends because they are produced by the same director. The system will recommend a new movie never seen before by your friends, given it is directed by their favored director.
A **visitor recommendation** system is another example of such system. If we integrate the knowledgeable database with visitor feedback, we will be able to recommend the best places, food, etc. to some stranger into new visiting place.

Never Ending Learning Systems

Learning systems require significant effort to develop and maintain. The problem appears infeasible to tackle once we are addressing diverse set of domains (languages) or when the problem has innumerable variations (set of categories to detect).
Our never ending learning system has to perform two main tasks:

- *Extract new instances of categories and relations among themselves*
- *Learn to read better than yesterday and to build better knowledgeable database*

The tool that I have developed, in the course of my graduate studies, are available as open source project[1]. Moreover, it is fast and web-scalable. I am interested in working with researchers in other fields to take advantage of the recent progress that has been made in machine learning, information retrieval and my research to push our sociological understanding of web content.

Research Plan Block Diagram

Figure 2 shows that my prospective research plan during my Phd studies.

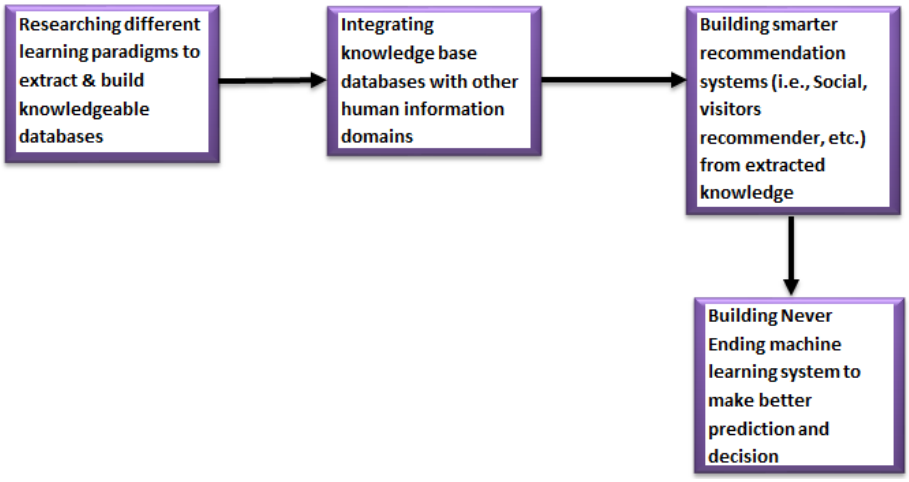

Figure 2: Deep Web Extractor System

Conclusion

In Conclusion, we will build a machine learning system that acquires the ability to extract structured information from unstructured web pages. If successful, this will result in a knowledge base (i.e., a relational database) of structured information that reflects the content of the Web and its learning is never ending because of improving itself day by day.

[1]DWX System: http://raoumer.github.io/dwx/