Content Discovery from Deep Web using Large Scale Data Analytics Paradigm

Presented By:

Rao Muhammad Umer

Web: raoumer.github.io

Graduate Student, MSCS

Supervised by:

Dr. Muhammad Abid Mughal

Co-supervised by:

Dr. Fayyaz ul Amir Afsar Minhas

Outline

- What is the Deep Web?
- Use Cases of Deep Web?
- How is the Deep Web invisible to Search Engines?
- Objectives of the Thesis?
- DWX System Architecture?
- Web Form Classification?
- DWX Application Architecture & Implementation?
- Large Scale Data Extraction?
- Related Work?
- Conclusion?
- Future Work?
- References?

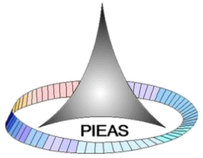
What is the Deep Web?

The Deep Web

- The **Deep Web** [1], **Invisible Web** [2], or **Hidden Web** [3] is the part of World Wide Web (WWW), whose web contents cannot be indexed by standard search engines for some reason.

- Search engines can access only **0.03%** of the World Wide Web. The remaining is known as the deep web [4].

- Recently estimated [5] that at least **4.72 billion** webpages are being indexed by typical search engines, while, **300 billion** webpages resides on the internet.



Use Cases of Deep Web

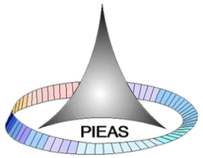
Use Cases of Deep Web

- www.grants.gov
- hdmoviespoint.com
- Search queries contain range values (e.g. laptop price from 10,000 to 20,000) on Standard Search Engines (i.e. Google, Yahoo or Bing, etc.)

How is the Deep Web invisible to Search Engines?

How is Deep Web invisible to search engines? [6]

- **Dynamic contents**
  - **returned in response to a query/ form**
- **Private web**
  - sites that require registration and login
- **Scripted content**
  - links produced by JavaScript
- **Non-HTML/text content**
  - specific file formats not handled by search engines; textual content encoded in multimedia
- **Limited access content**
  - sites that limit access to their pages in a technical way (Robots Exclusion Standard or CAPTCHAs)
- **Searchable databases**
  - **most valuable Deep Web resources**



Searchable Web Forms

[7]

Searchable Databases

- Searchable databases are those databases which reside on World Wide Web and are only accessible by directing the query through web form.

- Searchable databases are the most valuable resource of the deep web.

- Every searchable web form contains a searchable database behind it.

Objectives of the Thesis

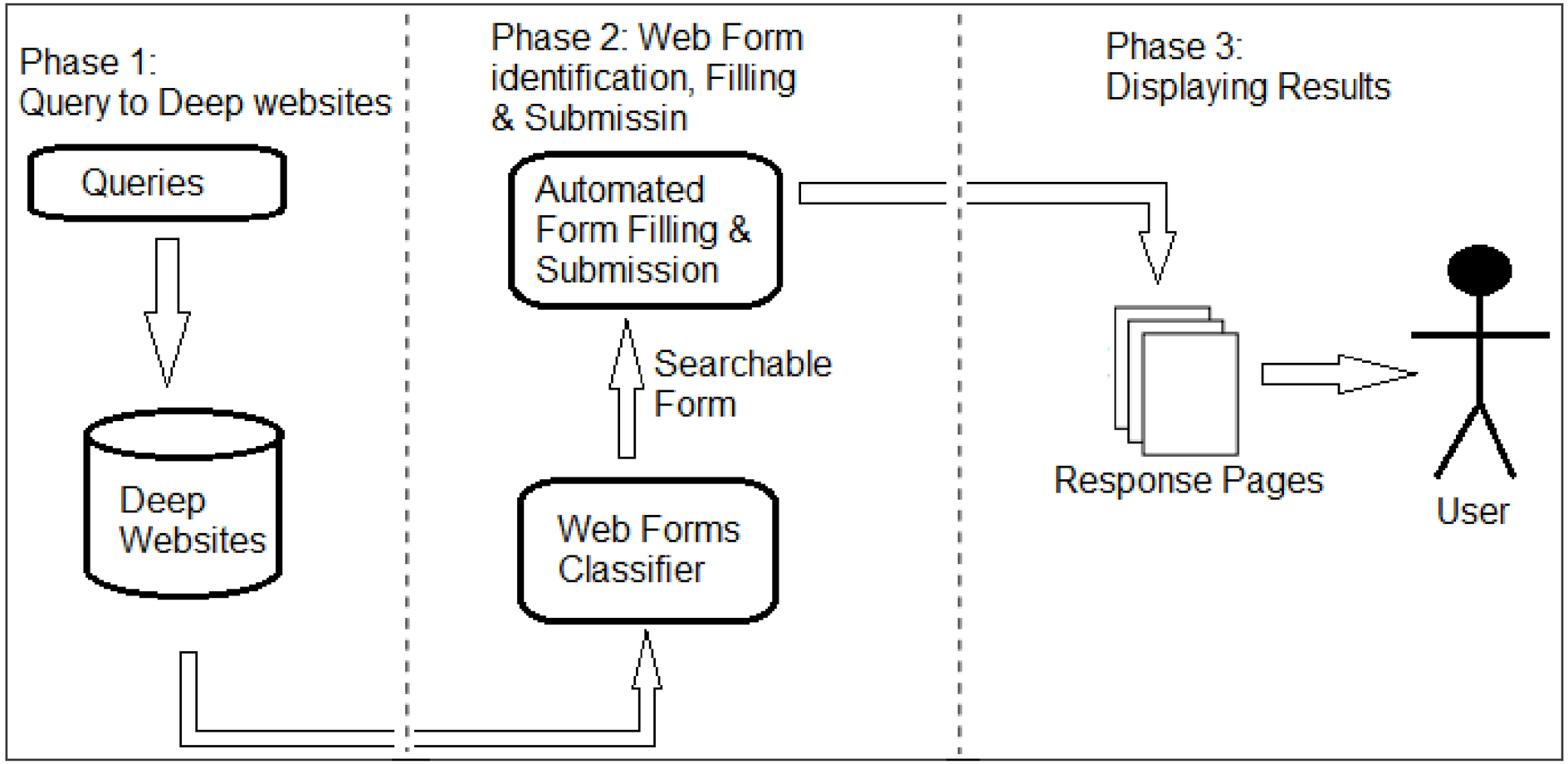
Objectives of the Thesis

The following are main objectives of this thesis:

- To discover and extract the deep web content of quality for web searchers.
- To discover automated means for identifying searchable web form interfaces and directing queries to them to dig out information.
- To build domain specific data repositories (e.g. real estate, health, newspapers, etc.) for purposeful analysis and uncovering hidden knowledge.
- To handle complex queries, like queries containing different range values, not entertained by traditional search engines.

DWX System Architecture

DWX System Architecture

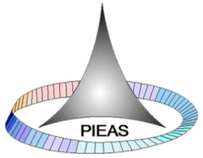


DWX System Architecture

- There are two main steps in developing a system like DWX:
  - **Resource Discovery:**

    In resource discovery, the task is to identify websites and databases that are expected to be relevant to the task.
  - **Content Extraction:**

    In content extraction, the task is to visit the identified sites to submit queries and extract data form webpages.
- Not handle the resource discovery task
- The related work [8-10] describes the resource discovery problem (i.e., identifying websites and webpages relevant to a specific task or topic).
- Here, we focus on content extraction task



Web Form Classification

Data Acquisition

- The empirical evaluation has been done on **1418** web forms dataset [11].
- It was originally collected form Alexa Top one million websites.

No. of form elements	Form Type	Form Label
415	Search	S
246	Login	L
165	Registration	R
143	Other	O
138	Contact	C
132	Join mailing list	M
105	Password / login recovery	P
74	Order / add to cart	B
Total Form: 1418		

Features Selection

Form Type Detection

The prominent features for form type detection are given as follow:

- A single search query field
- A search query field named "q" or "s" or something else
- Presence of "search" word in URL
- Presence of "search" word in submit button text
- Presence of "search" word in form CSS Class or ID
- Form method whether is a GET or POST
- etc.

Field Type Detection

The useful features have been included for field type detection as:

- Form type predicted by a form type classifier
- Field tag name
- Field value
- Text before and after field
- Field CSS class and ID
- Text of field label element
- Field title and placeholder attributes
- etc.

Machine Learning Models and Evaluation

- Logistic Regression Classifier
- SGD Classifier
- Linear SVM Classifier
- CRF Classifier

Logistic Regression Classifier

- Using Logistic Regression classifier, **88%** of web forms have been classified correctly.

Form type	Precision	Recall	f_1-score	Support
Search	0.90	0.96	0.93	415
Login	0.96	0.95	0.95	246
Registration	0.93	0.87	0.90	165
Password / login recovery	0.84	0.81	0.83	105
Join mailing list	0.87	0.87	0.87	132
Contact / comment	0.85	0.93	0.89	138
Other	0.66	0.68	0.67	143
Order / add to cart	0.96	0.61	0.74	74
Avg. / Total	0.88	0.88	0.88	1418

SGD Classifier

- Using SGD classifier, **88%** of web forms have been classified correctly.

Form type	Precision	Recall	f_1-score	Support
Search	0.90	0.96	0.93	415
Login	0.96	0.95	0.96	246
Registration	0.94	0.85	0.90	165
Password / login recovery	0.80	0.82	0.81	105
Join mailing list	0.89	0.86	0.88	132
Contact / comment	0.87	0.93	0.90	138
Other	0.65	0.71	0.68	143
Order / add to cart	0.98	0.59	0.74	74
Avg. / Total	0.88	0.88	0.88	1418

Linear SVM Classifier

- Using Linear SVM classifier, **89%** of web forms have been classified correctly.

Form type	Precision	Recall	f_1-score	Support
Search	0.92	0.96	0.94	415
Login	0.97	0.96	0.96	246
Registration	0.94	0.90	0.92	165
Password / login recovery	0.85	0.82	0.83	105
Join mailing list	0.88	0.86	0.87	132
Contact / comment	0.85	0.93	0.89	138
Other	0.67	0.71	0.69	143
Order / add to cart	0.94	0.62	0.75	74
Avg. / Total	0.89	0.89	0.89	1418

CRF Classifier

- Using CRF classifier, **86%** of web forms fields have been classified correctly.

Field type	Precision	Recall	f_1-score	Support
Search query	0.843	0.980	0.907	99
Email	0.945	0.987	0.966	156
Password	1.000	0.966	0.983	88
Product quantity	1.000	0.875	0.933	8
Submit button	0.895	1.000	0.944	68
Username	0.767	0.767	0.767	43
Password confirmation	1.000	1.000	1.000	24
Receive emails confirmation	0.909	0.370	0.526	27
First name	0.913	0.840	0.875	25
Last name	0.870	0.800	0.833	25
Organization name	1.000	0.417	0.588	12
Address	0.706	0.667	0.686	18
Avg. / Total	0.868	0.860	0.850	1182

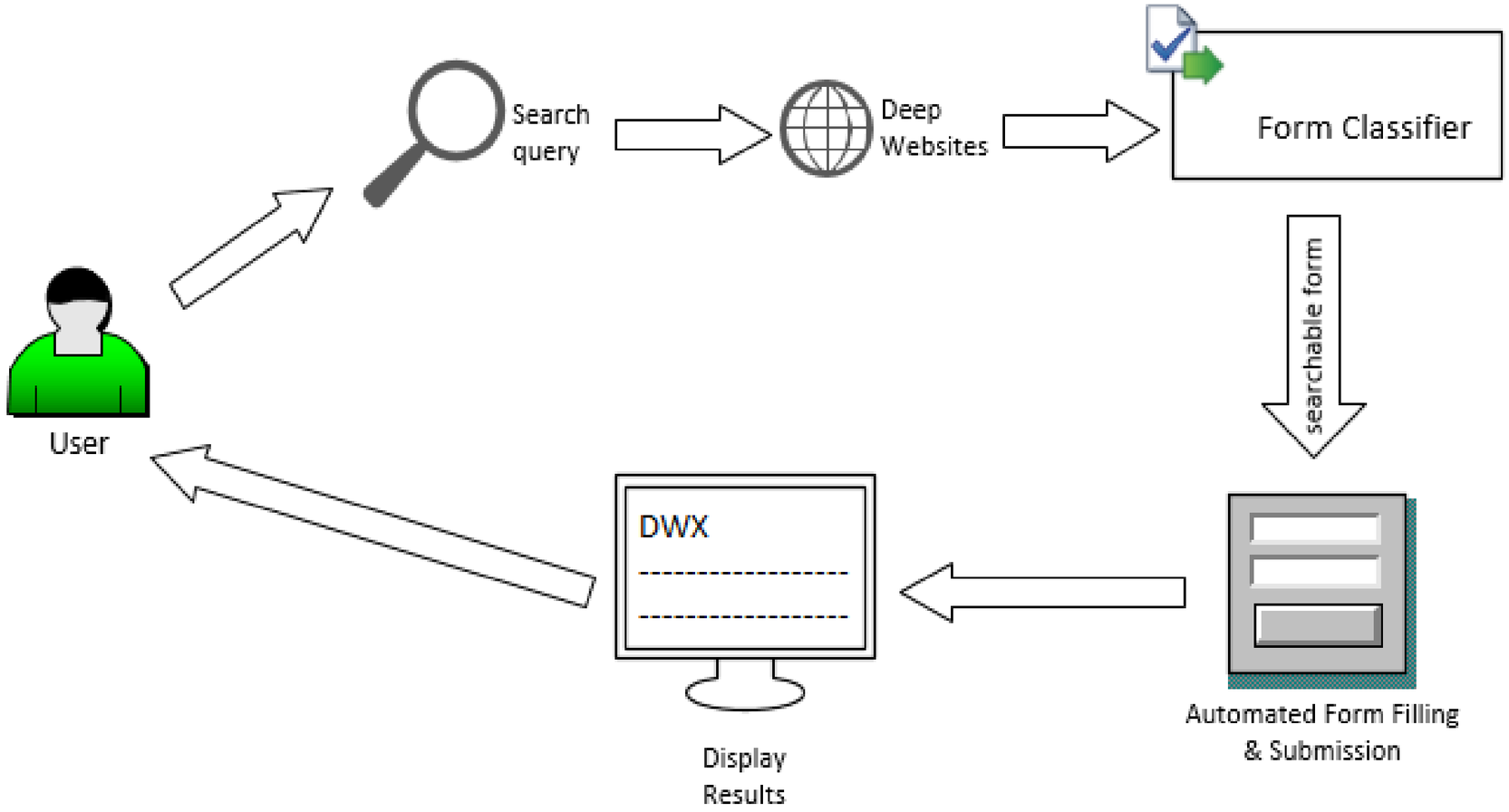
Results Comparison

- The empirical results in table revealed that Linear SVM classifier has better precision, recall and f_1-score than that of other two classifiers (i.e. Logistic Regression & SGD).
- Form type detection has been carried out by Linear SVM.
- Form field type detection has been done with CRF.

Evaluation Measure \ Classification Models	Logistic Regression	Stochastic Gradient Descent (SGD)	Linear SVM
Precision	0.88	0.88	0.89
Recall	0.88	0.88	0.89
f_1-score	0.88	0.88	0.89

DWX Application Architecture and Implementation

Application Architecture

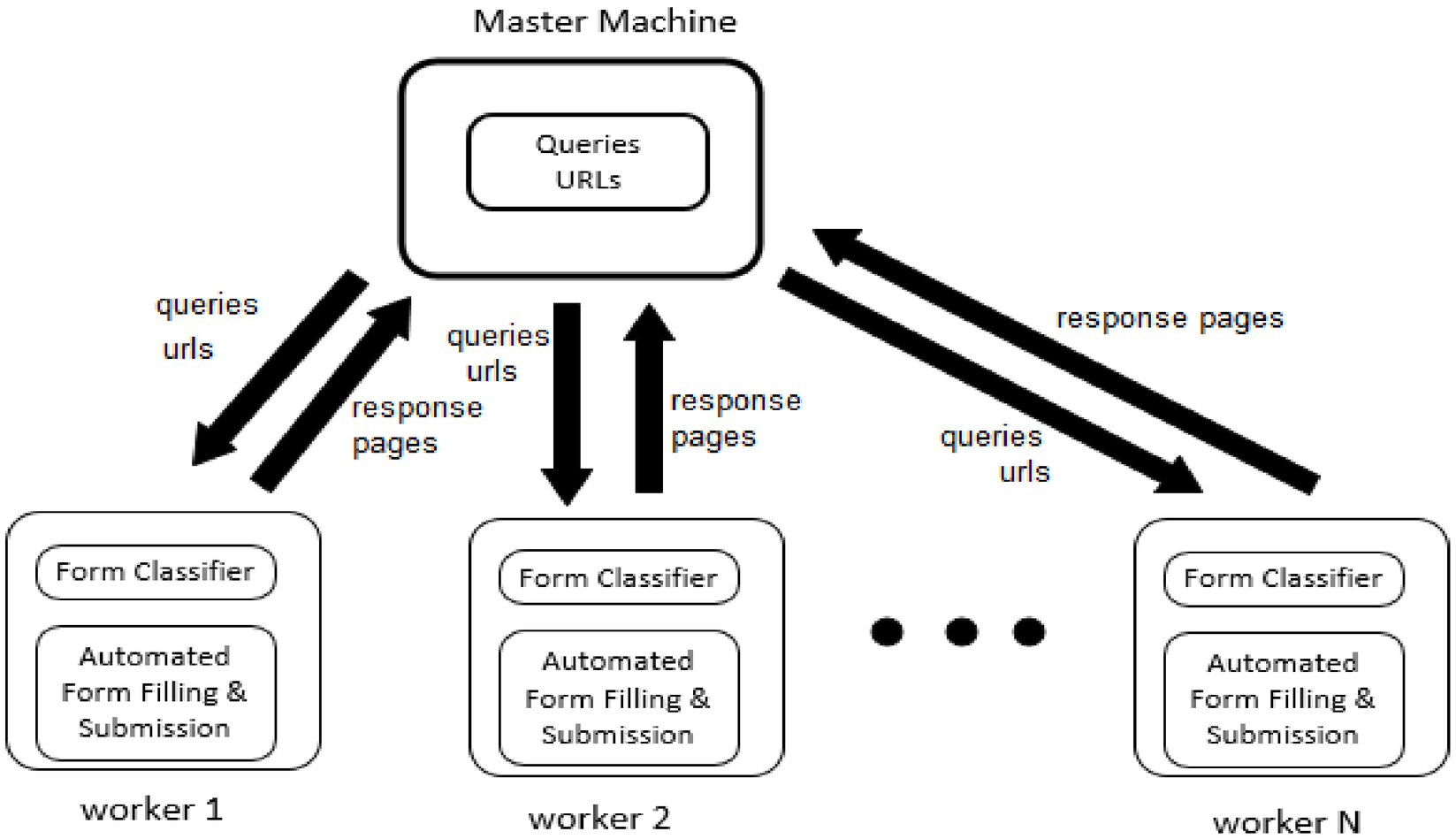


DWX Webapp

http://raoumer.github.io/dwx/

Large Scale Data Extraction

Large Scale Data Extraction



Related Work

Related Work

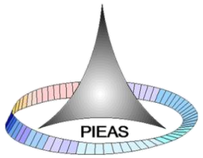
- **James Caverlee et al.** [12] developed the **“THOR”** system for discovering and extracting QA-Pagelets from the Deep Web resources.
- **RoadRunner** [13] proposed system automatically creates wrappers for extracting data from web pages and compares HTML pages. Similarly, **Arasu et al.** [14] developed a data extraction method that models template-generated pages and infers the unknown template used to generate the pages.
- **Panagiotis G. Ipeirotis et al.** [15] proposed an algorithm to originate content summaries from “uncooperative” databases by using focused query searching that extracts documents which are demonstrative of the topic exposure of the databases.

Related Work

- **Luciano Barbosa & Juliana Freire** [16] proposed a crawler to automatically locate hidden Web databases on a given topic by selecting links within a topic which are more probable to lead to webpages that contain searchable web forms. They also proposed adaptive crawler [17] to automatically learn patterns of promising links and to adapt topic focus of the crawl as it crawl more links.
- **Sriram Raghavan & Hector Garcia-Molina** [3] proposed a model of a hidden Web crawler for extracting contents from hidden web. Similarly, **Jayant Madhavan et al.** [18] proposed a system for surfacing Deep Web content, i.e. pre-evaluating submissions for each HTML form and adding the resulting HTML webpages into a search engine index.

DWX Approach

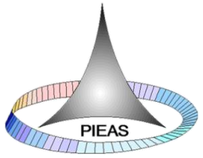
- Our work relating to data extraction from the Deep Web is different to above mentioned techniques, in such a way that it is directly passed user query to deep websites and extract data related to user query.
- Other developed systems overlooked about automating query generation and sometimes extract irrelevant data by submitting query into web form interfaces.



Conclusion

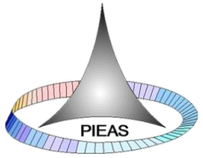
Conclusion

- Massive and high quality data from the Deep Web resource is useful to build a knowledge based databases (i.e. structured relational databases).
- Exploring and extracting the vast deep web content tends to a major research challenge for knowledge discovery and information retrieval community.
- An efficient **DWX** system has been introduced for extracting the contents from deep web resources.
- The DWX system heavily relies on **“form classifier”**, and **“automated form filling & submission”** modules.



Conclusion

- Form classifier module identifies the **searchable web interfaces** in the deep websites.
- Automated form filling and submission module **automatically fills the search form** field with query.
- The empirical results show that the DWX system is robust against changes in **“form interfaces”** and **“form fields”**.



Future Work

Future Work

- The form classifier in DWX is specifically capable of detecting web form type in which ***<form>* tag** element is present.
- In modern era, web form is created in **JavaScript** in which all action can be done in JavaScript function (e.g. onClick()) instead of *<form> tag* element.
- Form classifier is lacking to detect such type of form based on JavaScript.
- There is still a room to enhance the capability of “form classifier” by **detecting the JavaScript based forms** in future.
- Discover **hidden knowledge** from domain specific repositories.

Acknowledgement

- I feel my privilege and honor to express my sincere gratitude to my supervisor, **Dr. Muhammad Abid Mughal** and my co-supervisor, **Dr. Fayyaz-ul-Amir Afsar Minhas**.

- I would also like to thank **Department of Computer and Information Sciences (DCIS), Pakistan Institute of Engineering and Applied Sciences (PIEAS),** and **Pattern Recognition Lab** for providing very conducive educational environment.

- Finally, I would like to thank all those, whose names have not been mentioned here, but they have helped me in one or the other way in completion of my thesis work.

References

[1] N. Hamilton, "The Mechanics of a Deep Net Metasearch Engine," in *WWW*, 2003.

[2] J. Devine and F. Egger-Sider, "Beyond Google: the invisible web in the academic library," *The Journal of Academic Librarianship,* vol. 30, pp. 265-269, 2004.

[3] S. Raghavan and H. Garcia-Molina, "Crawling the hidden web," 2000.

[4] M. K. Bergman, "White paper: the deep web: surfacing hidden value," *Journal of electronic publishing,* vol. 7, 2001.

[5] A. Bosch, T. Bogers, and M. Kunder, "Estimating search engine index size variability: a 9-year longitudinal study," *Scientometrics,* vol. 107, pp. 839-856, 2016.

[6] (2016). *Deep web*. Available at: https://en.wikipedia.org/wiki/Deep_web

[7] (2016). *Github website page*. Available at: https://github.com

References

[8] S. Chakrabarti, M. Van den Berg, and B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery," *Computer Networks,* vol. 31, pp. 1623-1640, 1999.

[9] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori, "Focused Crawling Using Context Graphs."

[10] A. McCallumzy, K. Nigamy, J. Renniey, and K. Seymorey, "Building domain-specific search engines with machine learning techniques.“

[11] (2016), Web Forms Dataset. Available: https://github.com/TeamHG-Memex/Formasaurus/tree/master/formasaurus/data

[12] J. Caverlee, L. Liu, and D. Buttler, "Probe, cluster, and discover: Focused extraction of qa-pagelets from the deep web," in *Data Engineering, 2004. Proceedings. 20th International Conference on*, pp. 103-114.

References

[13] V. Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards automatic data extraction from large web sites," in *VLDB*, 2001, pp. 109-118.

[14] A. Arasu and H. Garcia-Molina, "Extracting structured data from web pages," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 2003, pp. 337-348.

[15] P. G. Ipeirotis and L. Gravano, "Distributed search over the hidden web: Hierarchical database sampling and selection," in *Proceedings of the 28th international conference on Very Large Data Bases*, 2002, pp. 394-405.

[16] L. Barbosa and J. Freire, "Searching for Hidden-Web Databases."

[17] L. Barbosa and J. Freire, "An adaptive crawler for locating hidden-web entry points," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 441-450.

[18] J. Madhavan, D. Ko, Ł. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy, "Google's deep web crawl," *Proceedings of the VLDB Endowment,* vol. 1, pp. 1241-1252, 2008.

Any Query?

Thank You for Your Patience!!!