

Toward an Intelligent Crawling Scheduler for Archiving News Websites Using Reinforcement Learning

Course: CS 6604

Instructor: DR. EDWARD A FOX

Team: Web Archive

Members: Xinyue Wang, Naman Ahuja, Ritesh

Bansal, Siddharth Dhar, Nathaniel Llorens

Date: 12/10/2019

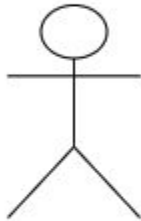


Motivation

- The Web is growing rapidly, but so too is information frequently disappearing from the WWW. Preserving the WWW is crucial to record the history of human society.
- How to preserve the web?
 - Crawling and saving
- The existing crawling model in the webarchive communities mainly adopts predefined crawling plans
 - Does not have the flexibility to capture the web dynamics
 - Needs a lot human efforts for designing the rules
- Can we have a smart web crawling scheduler reduce the human efforts efficiently?



Motivation



Web Archive Collector

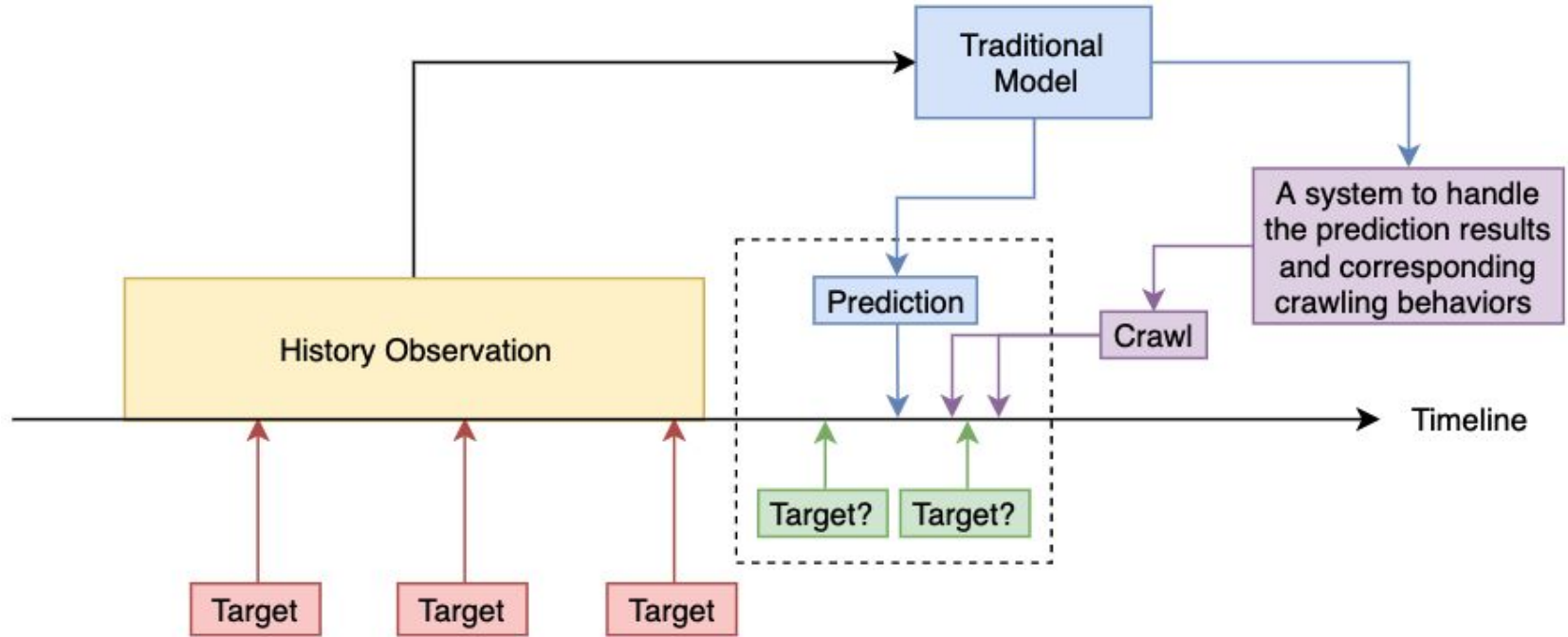




A smart web crawling scheduler

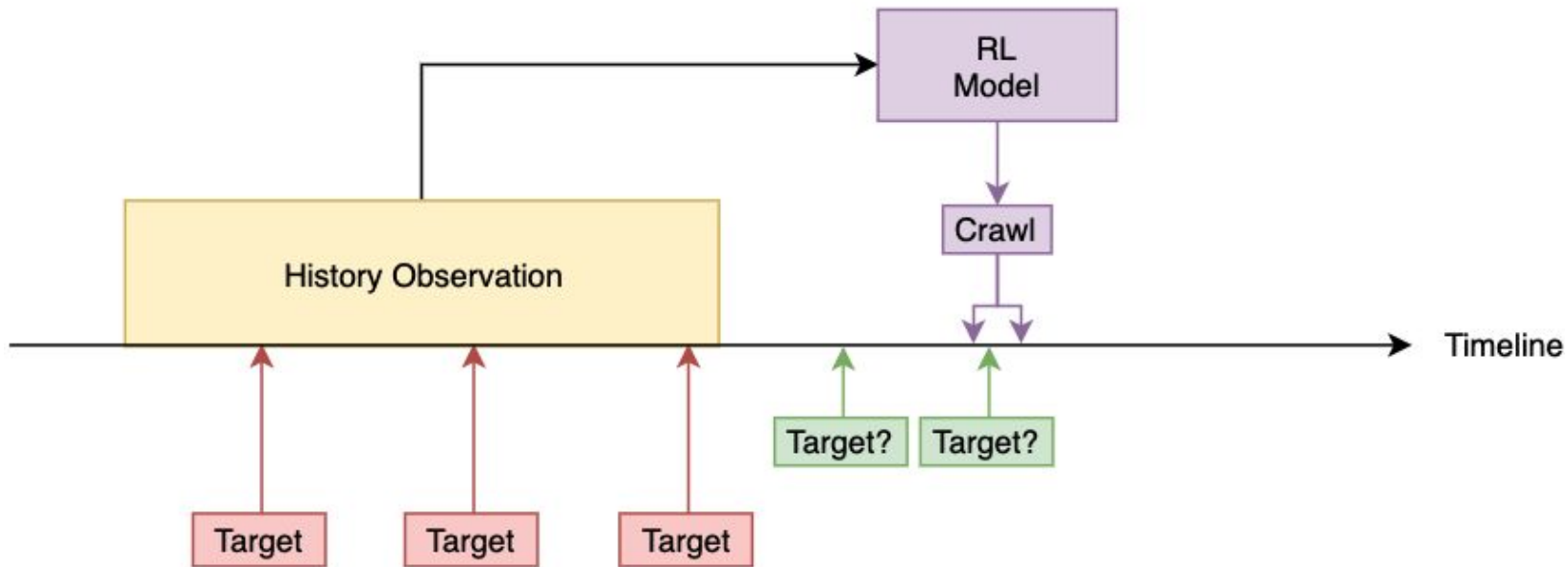
- A scheduler that can automatically **predict** the web content changes and generate corresponding plans to capture the content
 - Web page content change
 - Web site structure change
- The traditional way: predict when the web content would change in the future
 - Use machine learning model to track the changes and make predictions
 - SVM, random forest, logistic regression etc.
 - Then build the scheduler based upon the prediction results
- What if a smart model generate the crawling plans directly?
- We propose to explore **reinforcement learning (RL)**

Traditional Method



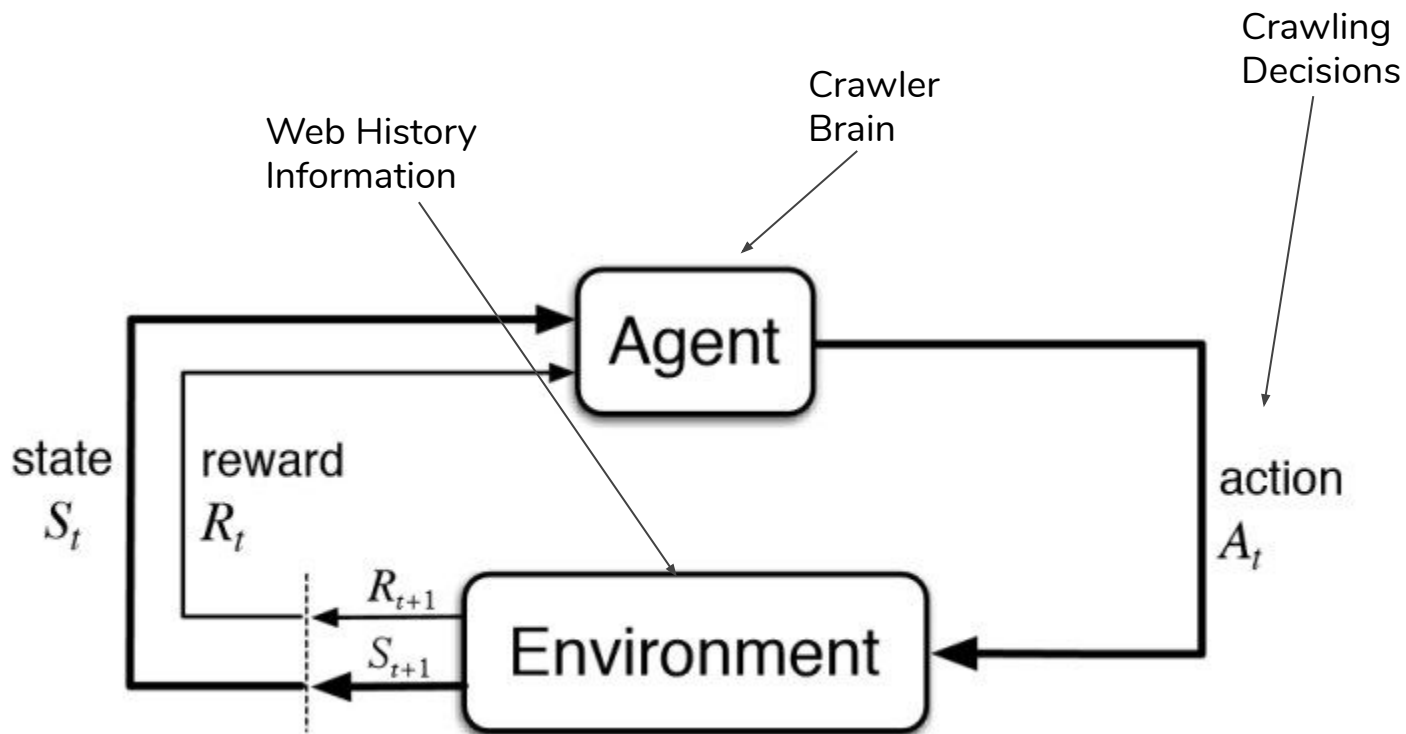


A RL idea





RL model basic



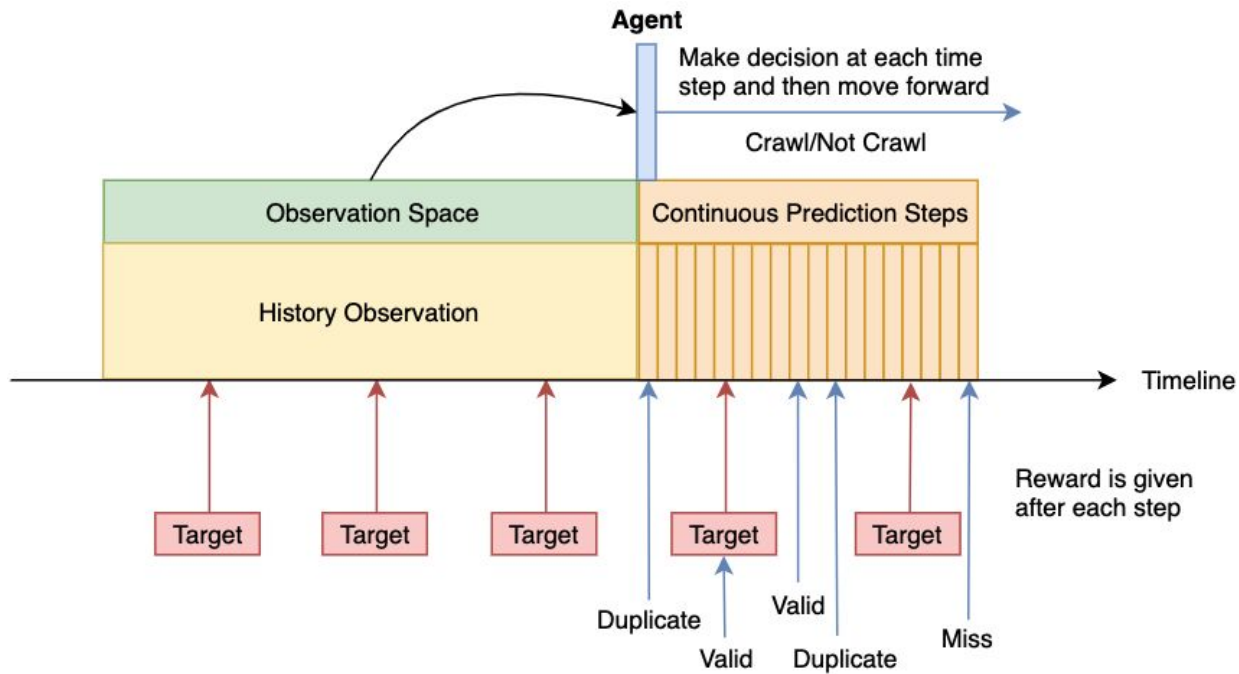


RL model design - web page change

- Environment
 - A continuous sequence from the data set, which represents the historical records of web pages
- Action
 - Continuous prediction model
 - Sparse prediction model
- Reward
 - Reward the correct actions
 - Punish the wrong actions



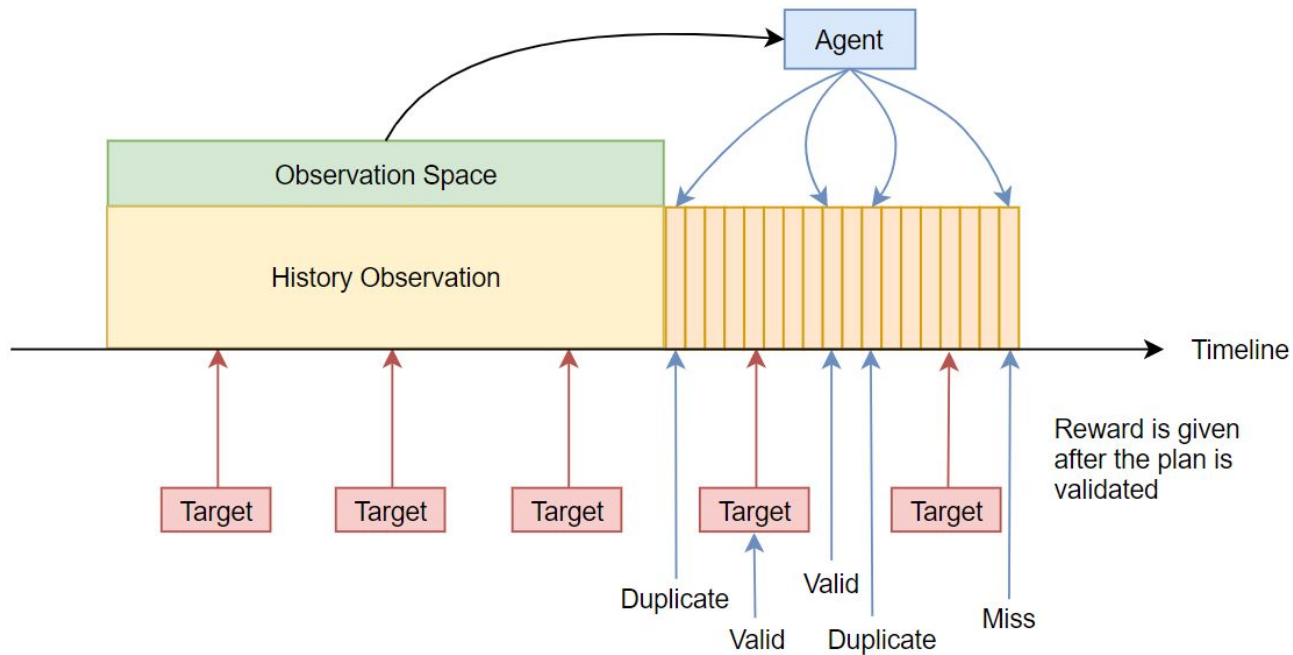
Continuous Prediction





Sparse Prediction

1. Pick number of crawls (N) to be chosen
2. Generate N positions within the range (crawling plan)





Rewarding

Result Type	Continous Prediction	Sparse Prediction
Duplicate (Crawl)	Negative	Negative
Valid	Positive	Positive
Miss	Negative	Negative
Duplicate (Not Crawl)	Positive	N/A

- A RL model tend to avoid negative rewarded actions
- A RL model tend to stack more positive rewards as much as possible
- Reward is give after each episode (crawling plan) is generated



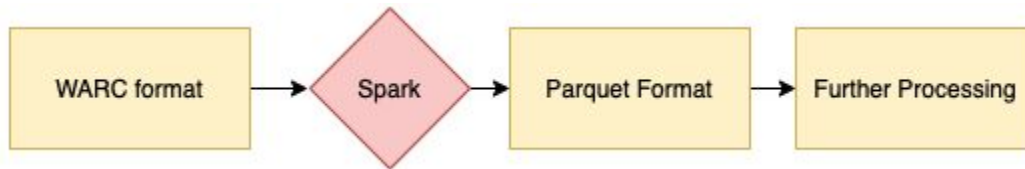
Data Collection

- CNN news: <http://cnn.com/>
- Web Archive collection from Archive.org: CNNfocuscrawls
 - From May 2019 to Oct 2019
 - Contains both CNN domain crawls and external links in CNN web sites
 - Crawling Tool: Heritrix
- **Our own focus crawl**
 - From Nov 6th to now
 - Contains only CNN domain crawls with 2 levels depth, crawled every hour
 - Crawling Tool: Web2Warc
- Issue: gap from Archive.org and our own crawls
 - We plan to compare the differences of our own crawl and the Archive.org collection to validate the capture accuracy of Archive.org
 - Will need until Archive.org publish the Nov data



Data Preprocessing

- The raw web archive files are in WARC format
- WARC -> Parquet
 - WARC is not flexible to use
 - WARC is inefficient to process
- We use Spark framework to batch converting WARC to Parquet efficiently





WARC and Parquet

```
1 |WARC/1.0
2 |WARC-Type: response
3 |WARC-Record-ID: <urn:uuid:ffc6e90-13d6-11e7-873d-525400672438>
4 |WARC-Target-URI: http://www.biography.com/people/seung-hui-cho-235991
5 |WARC-Date: 2017-03-28T16:52:59Z
6 |WARC-Payload-Digest: sha1:NSEZCE2K3EMRRJF6VXPBNY6GXUQLIE6M
7 |WARC-Block-Digest: sha1:BUBXYQ65JZ6KDTSZYLJUDF3LVF3F005K
8 |Content-Type: application/http; msgtype=response
9 |Content-Length: 319638
10
11 |HTTP/1.0 200 OK
12 |Cache-Control: max-age=30
13 |Content-Encoding: gzip
14 |Content-Type: text/html; charset=utf-8
15 |ETag: W/"4dd68-12Mt5MN6ZGRwf6q42PdzrA"
16 |X-Powered-By: Passion
17 |X-Recruiting: We are hiring! Come write HTTP headers with us! http://bit.ly/1vk8BEP1
18 |Via: 1.1 varnish
19 |Via: 1.1 varnish
20 |Fastly-Debug-Digest: 46b7ac7544ef20ed03c2e28cc89ec513eb71d96b09a65f274329c006f9156fd2
21 |X-SayCDN-TTL: 600.000
22 |X-Say-Cacheable: YES
23 |X-Say-TTL: 600.000
24 |X-SayCDN-UA: normal
25 |X-SayCDN-DevType: desktop
26 |Content-Length: 44993
27 |Accept-Ranges: bytes
28 |Date: Sun, 19 Mar 2017 00:59:01 GMT
29 |Age: 0
30 |Connection: keep-alive
31 |X-Served-By: cache-sea1020-SEA, cache-iad2644-IAD
32 |X-Cache: MISS, MISS
33 |X-Cache-Hits: 0, 0
34 |X-Timer: S1489885140.005043,V50,V296
35 |Vary: X-Say-ConfigVersion, X-SayCDN-UA, X-SayCDN-DevType, Accept-Encoding
36
37 |<!DOCTYPE html><html class="no-js" id="phx-wrapper" lang="en-us"><head phx-view-meta><
38 |</style><script>(function(d) {
39 |   var config = {
40 |     kitId: "nnd5wkr",
41 |     scriptTimeout: 3000,
42 |     async: true
43 |   },
```



WARC and Parquet

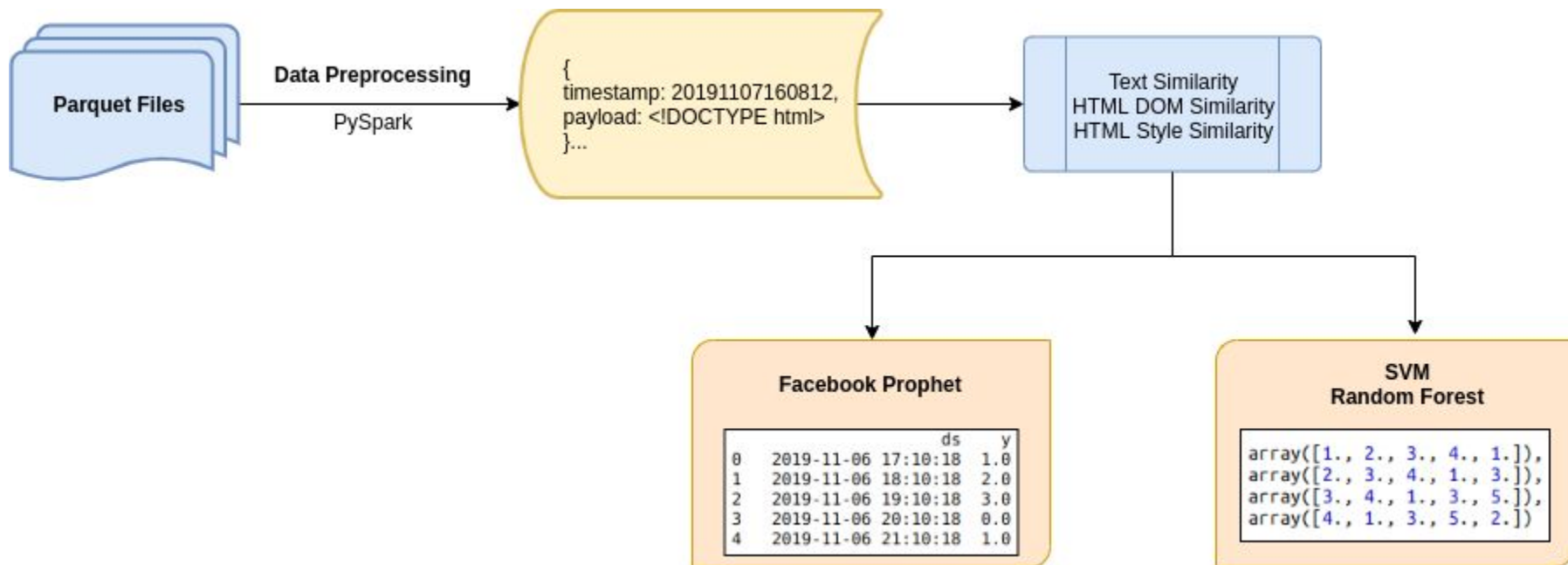
key	surtUrl	timestamp	originalUrl	mime	status	digest	redirectUrl	meta	contentLength	offset	filename	allheader	payload
http://(info,7ba,...	http://(info,7ba,...	20180520	http://0.7ba.info...	text/html	200	-	-	-	4947	855	CC-MAIN-201805201...	application/xhtml...	HTTP/1.1 200 OK
...
http://(cjp,ne,aki...	http://(cjp,ne,aki...	20180520	http://007412-002...	text/html	200	-	-	-	18962	4069	CC-MAIN-201805201...	text/html	sha1:AS... HTTP/1.1 200 OK
...
http://(com,021cc...	http://(com,021cc...	20180520	http://021cctv.co...	text/html	200	-	-	-	3613	10262	CC-MAIN-201805201...	text/html	sha1:B5... HTTP/1.1 200 OK
...
http://(cjp,nikki...	http://(cjp,nikki...	20180520	http://05a37y.nik...	text/html	200	-	-	-	58476	12832	CC-MAIN-201805201...	text/html	sha1:LI... HTTP/1.0 200 OK
...
http://(info,with...	http://(info,with...	20180520	http://06.withtub...	text/html	200	-	-	-	23773	23958	CC-MAIN-201805201...	text/html	sha1:RC... HTTP/1.1 200 OK
...
http://(com,080ut...	http://(com,080ut...	20180520	http://080ut.com/...	text/html	200	-	-	-	77425	31420	CC-MAIN-201805201...	text/html	sha1:RC... HTTP/1.1 200 OK



Find Web Content Change

- Two types of web content change:
 - **Web page change**
 - **Website structure change**
- Get the dataset labeled
 - Find unique web copies for a web page from the web archive
 - Find unique website treemap from the web archive

Baselines for predicting web content change - WebPage





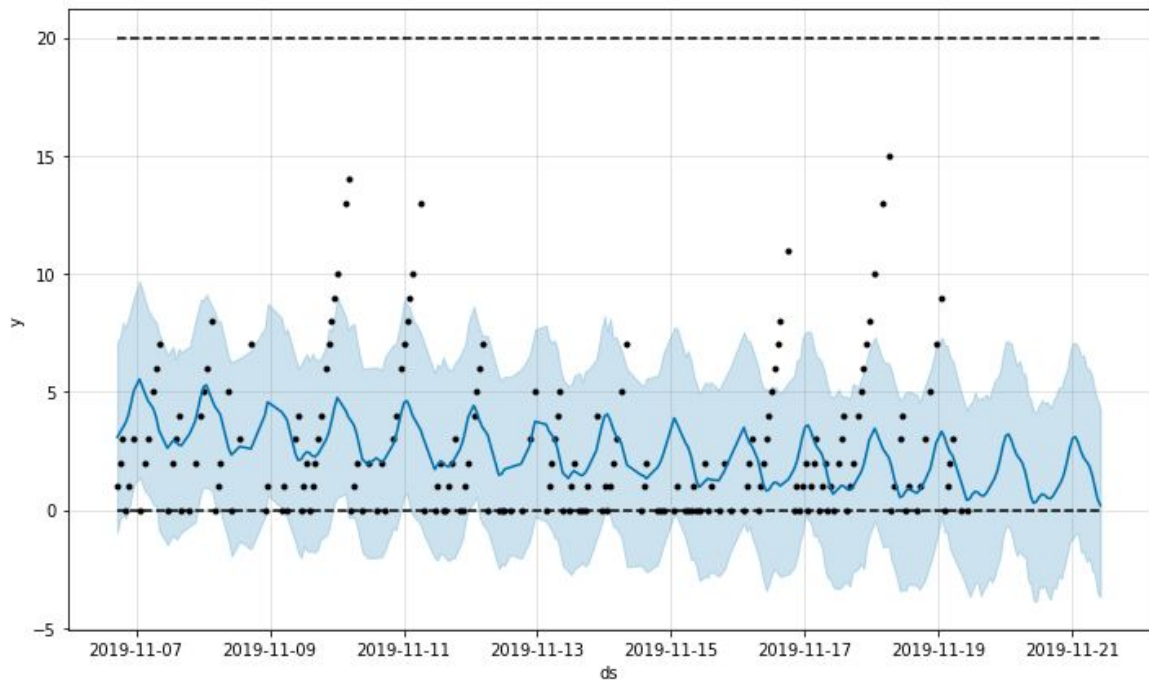
Baselines for predicting web content change - WebPage

Model	Accuracy	Precision	Recall	F1 Score
Support Vector Machine	0.769	0.74	0.77	0.75
Random Forest	0.615	0.60	0.62	0.61

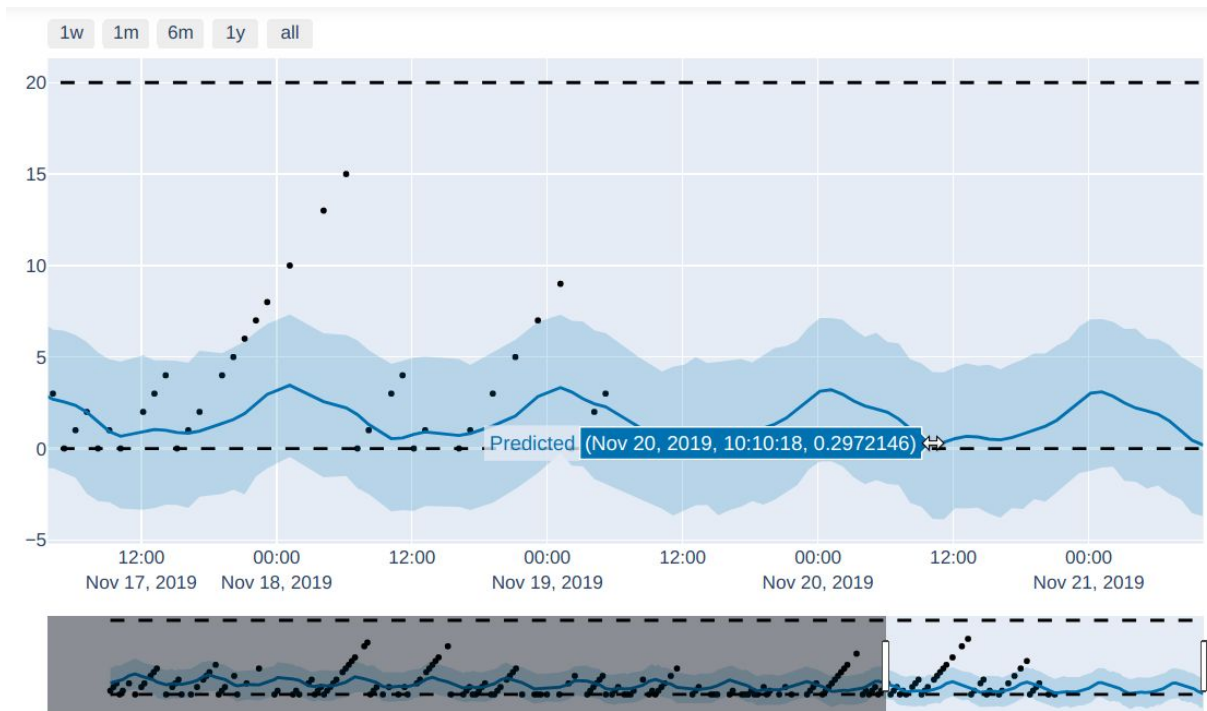
- The SVM model is able to solve nonlinear estimation problems; successful in time series forecasting.
- Random Forests generally don't fit very well for time series trends and seasonalities.



Predicting web page change - Facebook Prophet

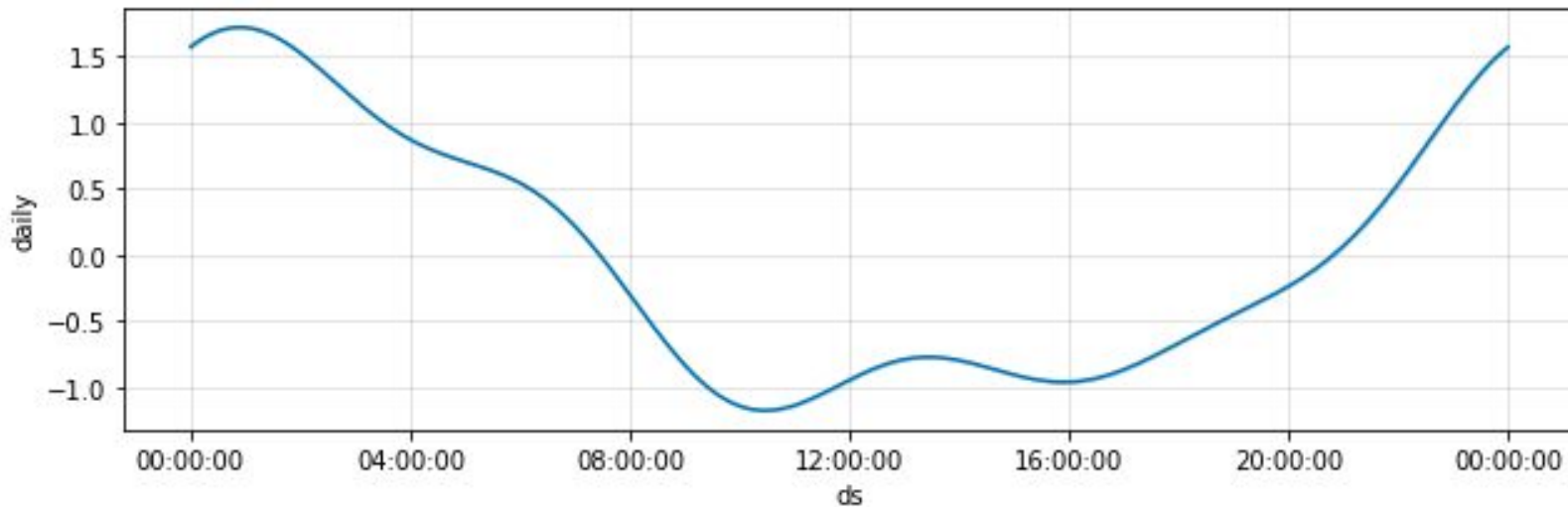


Predicting web page change - Facebook Prophet





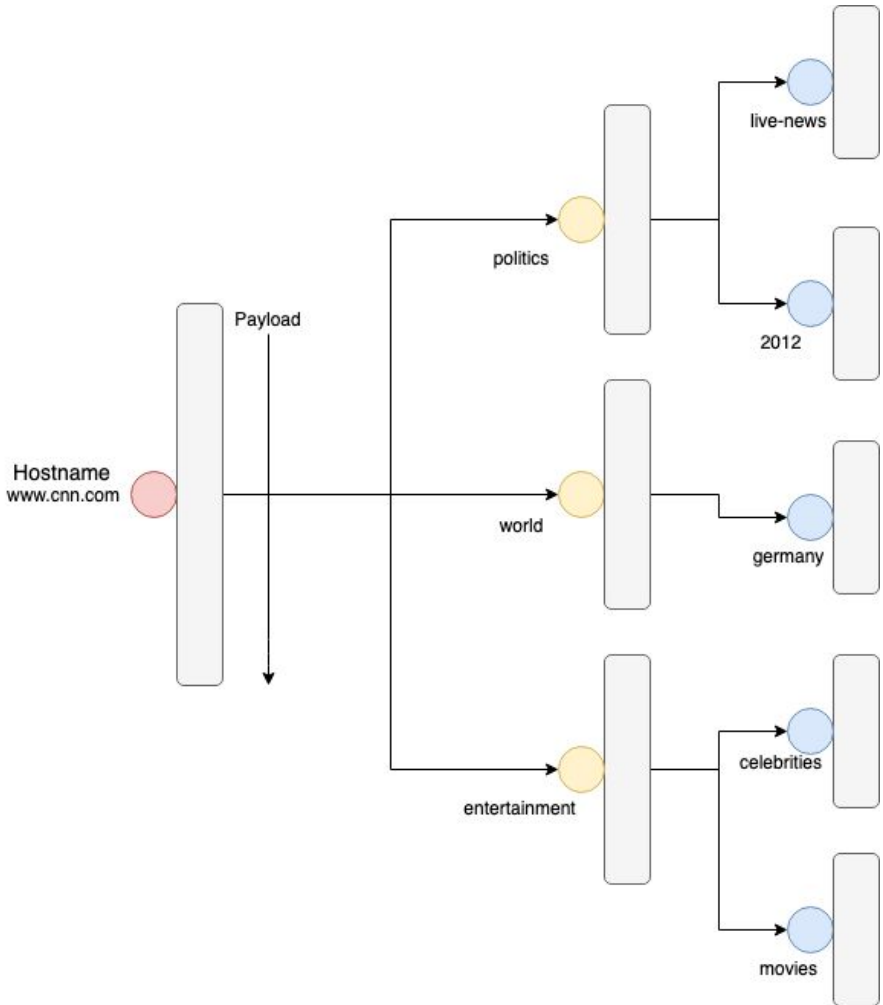
Predicting web page change - Facebook Prophet





Constructing unique web site treemap from the web archive

1. **Page level** prediction will help decide to **crawl particular** web page.
2. Consider the case when **multiple new web pages** are added, maybe during a particular time and day. We may want to crawl all these pages.
3. **Sitemap prediction** will help **decide** to crawl **whole website** or not.
4. **Data archived** contains data of **multiple media-type** and **status code**.
5. **Initial preprocessing** is required to build sitemap, 'text/html' and '200'.
6. **Tree-Based Structure**, 'hostname' is 'root node', 'path act' as 'nodes'.





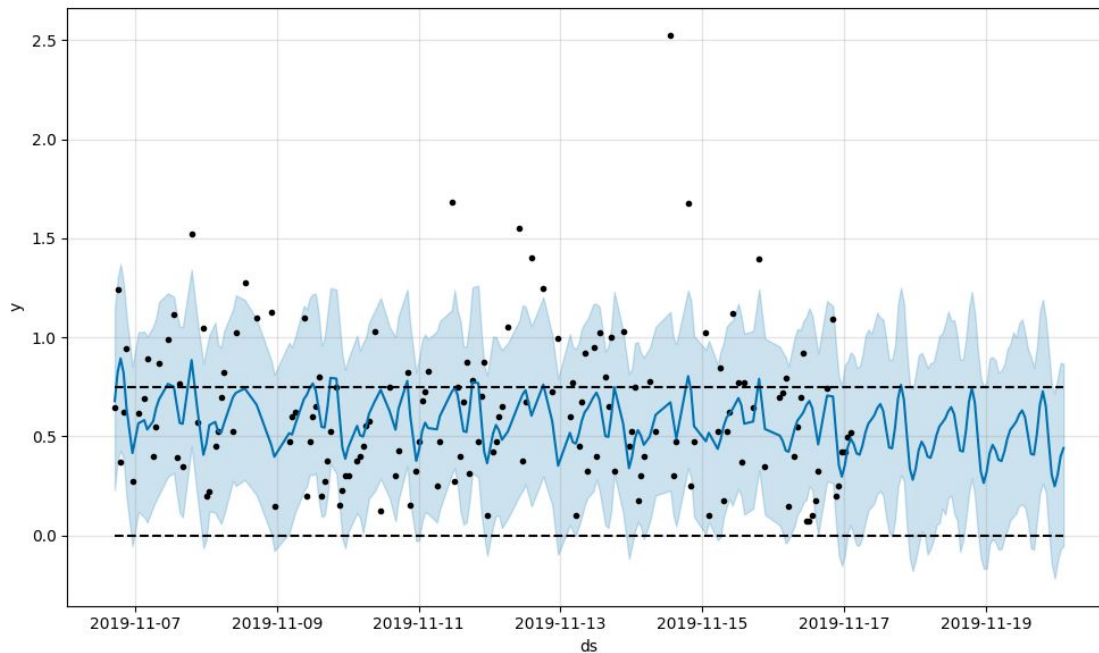
Baselines for predicting web structure change - TreeMap

1. Built **training data** based on the window size of **5 hours**.
2. Used **percentage of nodes added** from the last crawled datapoint to evaluate the **labels**.
3. Number of nodes added will act as the input feature.
4. Used SVM and Random Forest models to predict if the website should be crawled.
5. Prediction is if we want “to crawl whole website in the next hour”

	Accuracy	Precision	Recall	F1-Score
SVM	0.71	0.53	0.71	0.61
RF	0.61	0.59	0.59	0.59

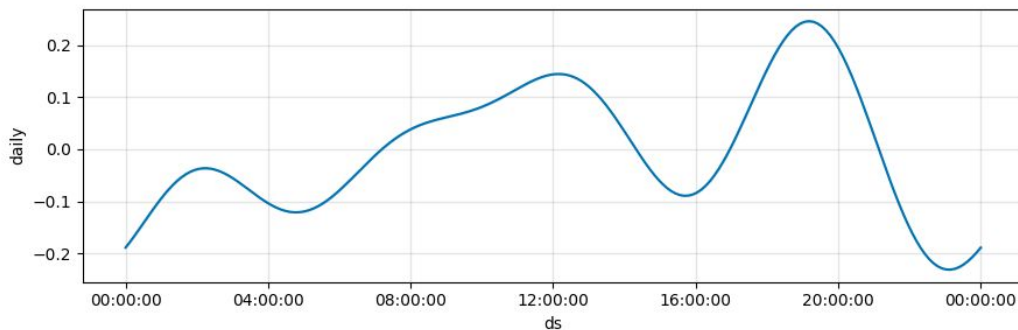
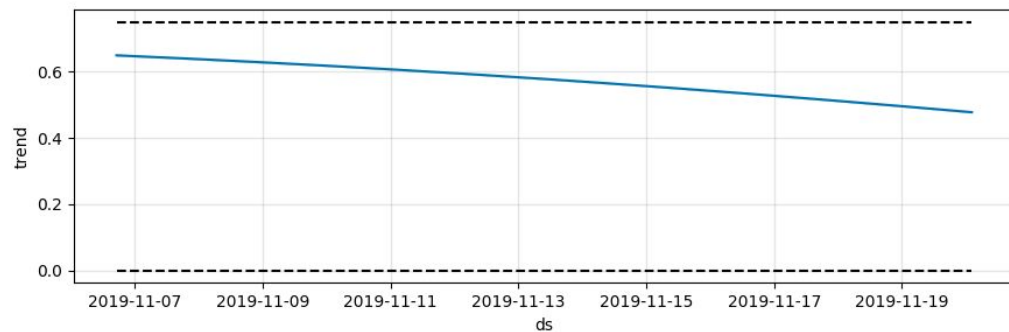


Predicting SiteMap change - Prophet





Predicting SiteMap change - Prophet

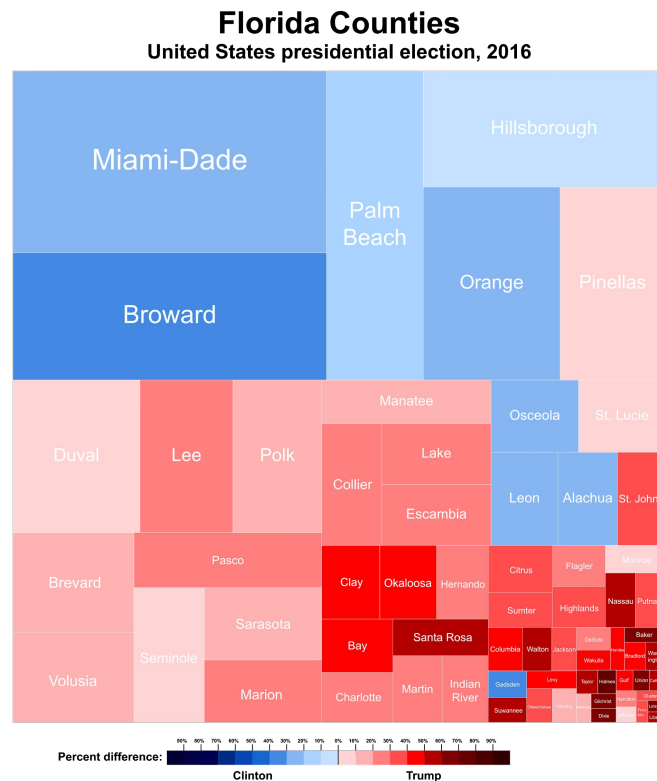




Constructing unique web site treemap from the web archive (continued)

- The sitemaps generated by our algorithms are important for predicting the change of a site's tree structure, but they are also useful for visualizing how and where a site is changing
- Algorithm, when given input of two sitemaps, can show differences between them in a visualization known as a treemap (unrelated example to the right)
- Allows a user to see how a site is changing between snapshots of a site

Credit: <https://en.wikipedia.org/wiki/Treemapping>





RL model evaluation - web page change

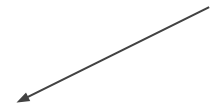
- Use the same dataset as web page prediction baselines
- 100000 training steps for each model
- 8:2 training, testing ratio
- Two RL learning policies: DQN, PPO
- Two types of rewarding functions: fixed score, scaled score



RL model evaluation - web page change

Result Type	Fixed Score	Scaled Score
Duplicate (Crawl)	-2	-2
Valid	15	$20 \cdot (1 - D_t / D_{max})$
Valid (Exact Match)	20	N/A
Miss	-10	-10
Duplicate (Not Crawl)	2	N/A

Reward
Setting



Learning
Policy Used



	Max Fixed Score	DQN	PPO2	Max Scaled Score	DQN	PPO2
continuous prediction	50	-28.5	27.6	50	-26.8	30.8
sparse prediction	20	N/A	9.7	20	N/A	10.5



Conclusion and future work

- At current stage, we can not compare the RL model with baselines directly since they are different tasks
 - Further experiments are needed to build a system to work with baselines for crawling plan generation
- We also need more training data with a longer period of various of types
- RL model has the potential to solve our web crawling
- Quantitative evaluation is needed to reflect the actual performance of the model compared with baselines or predefined policies
- Investigate the web site structure change model in the future
- The web page change and web site structure change may also be combined as a single model