

A Brief History of Web Crawlers

Seyed M. Mirtaheri, Mustafa Emre Dinçtürk, Salman Hooshmand,
Gregor V. Bochmann, Guy-Vincent Jourdan

School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, Ontario, Canada

{*staheri,mdinc075, shooshmand*}@uottawa.ca, {*bochmann,gvj*}@eecs.uottawa.ca

Iosif Viorel Onut

Security AppScan® Enterprise, IBM
770 Palladium Dr, Ottawa, Ontario, Canada
vioonut@ca.ibm.com

Abstract

Web crawlers have a long and interesting history. Early web crawlers collected statistics about the web. In addition to collecting statistics about the web and indexing the applications for search engines, modern crawlers can be used to perform accessibility and vulnerability checks on the application.

Quick expansion of the web, and the complexity added to web applications have made the process of crawling a very challenging one. Throughout the history of web crawling many researchers and industrial groups addressed different issues and challenges that web crawlers face. Different solutions have been proposed to reduce the time and cost of crawling. Performing an exhaustive crawl is a challenging question. Additionally, capturing the model of a modern web application and extracting data from it automatically is another open question.

What follows is a brief history of different techniques and algorithms used from the early days of crawling up to the recent days. We introduce criteria to evaluate the relative performance and objective of web crawlers. Based on these criteria we plot the evolution of web crawlers.

1 Introduction

Crawling is the process of exploring web applications automatically. The web crawler aims at discovering the web pages of a web application by navigating through the application. This is usually done by simulating the possible user interactions considering just the client-side of the application.

As the amount of information on the web has been increasing drastically, web users increasingly rely on search engines to find desired data. In order for search engines to learn about the new data as it becomes available, the web crawler has to constantly crawl and update the search engine database.

1.1 Motivations for Crawling

There are several important motivations for crawling. The main three motivations are:

- Content indexing for search engines. Every search engine requires a web crawler to fetch the data from the web.
- Automated testing and model checking of the web application
- Automated security testing and vulnerability assessment. Many web applications use sensitive data and provide critical services. To address the security concerns for web applications, many commercial and open-source automated web application security scanners

have been developed. These tools aim at detecting possible issues, such as security vulnerabilities and usability issues, in an automated and efficient manner [1, 2].

They require a web crawler to discover the states of the application scanned.

1.2 The Evolution of Web Crawlers

In the literature on web-crawling, a web crawler is basically a software that starts from a set of seed URLs, and downloads all the web pages associated with these URLs. After fetching a web page associated with a URL, the URL is removed from the working queue. The web crawler then parses the downloaded page, extracts the linked URLs from it, and adds new URLs to the list of seed URLs. This process continues iteratively until all of the contents reachable from seed URLs are reached.

The traditional definition of a web crawler assumes that all the content of a web application is reachable through URLs. Soon in the history of web crawling it became clear that such web crawlers can not deal with the complexities added by interactive web applications that rely on the user input to generate web pages. This scenario often arises when the web application is an interface to a database and it relies on user input to retrieve contents from the database. The new field of *Deep Web-Crawling* was born to address this issue.

Availability of powerful client-side web-browsers, as well as the wide adaptation to technologies such as HTML5 and AJAX, gave birth to a new pattern in designing web applications called *Rich Internet Application* (RIA). RIAs move part of the computation from the server to the client. This new pattern led to complex client side applications that increased the speed and interactivity of the application, while reducing the network traffic per request.

Despite the added values, RIAs introduced some unique challenges to web crawlers. In a RIA, user interaction often results in execution of client side *events*. Execution of an event in a RIA often changes the state of the web application on the client side, which is represented in the form of a *Document Object Model* (DOM) [3]. This change in the state of DOM does not necessarily mean changing the URL. Traditional web crawlers rely

heavily on the URL and changes to the DOM that do not alter the URL are invisible to them. Although deep web crawling increased the ability of the web crawlers to retrieve data from web applications, it fails to address changes to DOM that do not affect the URL. The new and recent field of *RIA web-crawling* attempts to address the problem of RIA crawling.

1.3 Problem Definition

A web application can be modeled as a directed graph, and the *World Wide Web* can be modeled as a forest of such graphs. The problem of Web crawling is the problem of discovering all the nodes in this forest. In the application graph, each node represents a state of the application and each edge a transition from one state to another.

As web applications evolved, the definition of the state of the application evolved as well. In the context of traditional web applications, states in the application graph are pages with distinct URLs and edges are hyperlinks between pages i.e. there exist an edge between two nodes in the graph if there exist a link between the two pages. In the context of deep web crawling, transitions are constructed based on users input. This is in contrast with hyperlink transitions which always redirect the application to the same target page. In a deep web application, any action that causes submission of a form is a possible edge in the graph.

In RIAs, the assumption that pages are nodes in the graph is not valid, since the client side code can change the application state without changing the page URL. Therefore nodes here are application states denoted by their DOM, and edges are not restricted to forms that submit elements, since each element can communicate with the server and partially update the current state. Edges, in this context, are client side actions (e.g. in JavaScript) assigned to DOM elements and can be detected by web crawler. Unlike the traditional web applications where jumps to arbitrary states are possible, in a RIA, the execution of sequence of events from the current state or from a seed URL is required to reach a particular state.

The three models can be unified by defining the state of the application based on the state of the DOM as well as other parameters such as the page URL, rather than the URL or the DOM alone. A hyperlink in a traditional web application does not

Table 1: Different categories of web crawlers

Category	Input	Application graph components
Traditional	Set of seed URLs	Nodes are pages with distinct URL and a directed edge exist from page p_1 to page p_2 if there is a hyperlink in page p_1 that points to page p_2
Deep	Set of Seed URLs, user context specific data, domain taxonomy	Nodes are pages and a directed edge exists between page p_1 to page p_2 if submitting a form in page p_1 gets the user to page p_2 .
RIA	A starting page	Nodes are DOM states of the application and a directed edge exist from DOM d_1 to DOM d_2 if there is a client-side JavaScript event, detectable by the web crawler, that if triggered on d_1 changes the DOM state to d_2
Unified Model	A seed URL	Nodes are calculated based on DOM and the URL. An edge is a transmission between two states triggered through client side events. Redirecting the browser is a special client side event.

only change the page URL, but it also changes the state of the DOM. In this model changing the page URL can be viewed as a special client side event that updates the entire DOM. Similarly, submission of a HTML form in a deep web application leads to a particular state of DOM once the response comes back from the server. In both cases the final DOM states can be used to enumerate the states of the application. Table 1 summarizes different categories of web crawlers.

1.4 Requirements

Several design goals have been considered for web crawlers. *Coverage* and *freshness* are among the first [4]. Coverage measures the relative number of pages discovered by the web crawler. Ideally given enough time the web crawler has to find all pages and build the complete model of the application. This property is referred to as *Completeness*. Coverage captures the static behaviour of traditional web applications well. It may fail, however, to capture the performance of the web crawler in crawling dynamically created web pages. The search engine index has to be updated constantly to reflect changes in web pages created dynamically. The ability of the web crawler to retrieve latest updates is measured through *freshness*.

An important and old issue in designing web crawlers is called *politeness* [5]. Early web crawlers had no mechanism to stop them from bombing a server with many requests. As the result while crawling a website they could have lunched an inadvertent *Denial of Service*(DoS) attack and

exhaust the target server resources to the point that it would interrupt normal operation of the server. Politeness was the concept introduced to put a cap on the number of requests sent to a web-server per unit of time. A polite web crawler avoids launching an inadvertent DoS attack on the target server. Another old problem that web crawlers faced are *traps*. Traps are seemingly large set of websites with arbitrary data that are meant to waste the web crawler resources. Integration of *black-lists* allowed web crawlers to avoid traps. Among the challenges web crawlers faced in the mid 90s was *scalability* [6]. Throughout the history of web-crawling, the exponential growth of the web and its constantly evolving nature has been hard to match by web crawlers. In addition to these requirements, the web crawler’s model of application should be *correct* and reflect true content and structure of the application.

In the context of deep-web crawling Raghavan and Garcia-Molina [7] suggest two more requirements. In this context, *Submission efficiency* is defined as the ratio of submitted forms leading to result pages with new data; and *Lenient submission efficiency* measures if a form submission is semantically correct (e.g., submitting a company name as input to a form element that was intended to be an author name)

In the context of RIA crawling a non-functional requirement considered by Kamara et al. [8] called *efficiency*. Efficiency means discovering valuable information as soon as possible. For example states are more important than transitions and should be found first instead of finding transitions leading to

already known states. This is particularly important if the web crawler will perform a partial crawl rather than a full crawl.

This paper defines web crawling and its requirements, and based on the defined model classifies web crawlers.

A brief history of traditional web crawlers¹, deep web crawlers², and RIA crawlers³ is presented in sections II-IV. Based on this brief history and the model defined, taxonomy of web crawling is then presented in section V. Section VI concludes the paper with some open questions and future works in web crawling.

2 Crawling Traditional Web Applications

Web crawlers were written as early as 1993. This year gave birth to four web crawlers: *World Wide Web Wanderer*, *Jump Station*, *World Wide Web Worm* [11], and *RBSE spider*. These four spiders mainly collected information and statistic about the web using a set of seed URLs. Early web crawlers iteratively downloaded URLs and updated their repository of URLs through the downloaded web pages.

The next year, 1994, two new web crawlers appeared: *WebCrawler* and *MOMspider*. In addition to collecting stats and data about the state of the web, these two web crawlers introduced concepts of *politeness* and *black-lists* to traditional web crawlers. *WebCrawler* is considered to be the first parallel web crawler by downloading 15 links simultaneously. From *World Wide Web Worm* to *WebCrawler*, the number of indexed pages increased from 110,000 to 2 million. Shortly after, in the coming years a few commercial web crawlers became available: *Lycos*, *Infoseek*, *Excite*, *AltaVista* and *HotBot*.

In 1998, Brin and Page [12] tried to address the issue of scalability by introducing a large scale web crawler called *Google*. Google addressed the problem of scalability in several ways: Firstly it leveraged many low level optimizations to reduce disk access time through techniques such as compression and indexing. Secondly, and on a higher

level, Google calculated the probability of a user visiting a page through an algorithm called *PageRank*. *PageRank* calculates the probability of a user visiting a page by taking into account the number of links that point to the page as well as the style of those links. Having this probability, Google simulated an arbitrary user and visited a page as often as the user did. Such approach optimizes the resources available to the web crawler by reducing the rate at which the web crawler visits unattractive pages. Through this technique, Google achieved high *freshness*. Architecturally, Google used a master-slave architecture with a master server (called *URLServer*) dispatching URLs to a set of slave nodes. The slave nodes retrieve the assigned pages by downloading them from the web. At its peak, the first implementation of Google reached 100 page downloads per second.

The issue of scalability was further addressed by Allan Heydon and Marc Najork in a tool called *Mercator* [5] in 1999. Additionally *Mercator* attempted to address the problem of extendability of web crawlers. To address extensibility it took advantage of a modular Java-based framework. This architecture allowed third-party components to be integrated into *Mercator*. To address the problem of scalability, *Mercator* tried to solve the problem of *URL-Seen*. The *URL-Seen* problem answers the question of whether or not a URL was seen before. This seemingly trivial problem gets very time-consuming as the size of the URL list grows. *Mercator* increased the scalability of *URL-Seen* by batch disk checks. In this mode hashes of discovered URLs got stored in RAM. When the size of these hashes grows beyond a certain limit, the list was compared against the URLs stored on the disk, and the list itself on the disk was updated. Using this technique, the second version of *Mercator* crawled 891 million pages. *Mercator* got integrated into *AltaVista* in 2001.

IBM introduced *WebFountain* [13] in 2001. *WebFountain* was a fully distributed web crawler and its objective was not only to index the web, but also to create a local copy of it. This local copy was *incremental* meaning that a copy of the page was kept indefinitely on the local space, and this copy got updated as often as *WebFountain* visited the page. In *WebFountain*, major components such as the scheduler were distributed and the crawling was an ongoing process where the local copy of the web only grew. These features, as well as de-

¹ See Olston and Najork [4] for a survey of traditional web crawlers.

² See He et al. [9] for a survey of deep web crawlers.

³ See Choudhary et al. [10] for a survey of RIA crawlers.

ployment of efficient technologies such as the *Message Passing Interface* (MPI), made WebFountain a scalable web crawler with high freshness rate. In a simulation, WebFountain managed to scale with a growing web. This simulated web originally had 500 million pages and it grew to twice its size every 400 days.

In 2002, *Polybot* [14] addressed the problem of URL-Seen scalability by enhancing the batch disk check technique. Polybot used Red-Black tree to keep the URLs and when the tree grows beyond a certain limit, it was merged with a sorted list in main memory. Using this data structure to handle the URL-Seen test, Polybot managed to scan 120 million pages. In the same year, *UbiCrawler* [15] dealt with the problem of URL-Seen with a different, more peer-to-peer (P2P), approach. UbiCrawler used consistent hashing to distribute URLs among web crawler nodes. In this model no centralized unit calculates whether or not a URL was seen before, but when a URL is discovered it is passed to the node responsible to answer the test. The node responsible to do this calculation is detected by taking the hash of the URL and map it to the list of nodes. With five 1GHz PCs and fifty threads, UbiCrawler reached a download rate of 10 million pages per day.

In addition to Polybot and UbiCrawler, in 2002 Tang et al. [16] introduced *pSearch*. *pSearch* uses two algorithms called *P2P Vector Space Model (pVSM)* and *P2P Latent Semantic Indexing (pLSI)* to crawl the web on a P2P network. VSM and LSI in turn use vector representation to calculate the relation between queries and the documents. Additionally *pSearch* took advantage of *Distributed Hash Tables (DHT)* routing algorithms to address scalability.

Two other studies used DHTs over P2P networks. In 2003, Li et al. [17] used this technique to scale up certain tasks such as clustering of contents and bloom filters. In 2004, Loo et al. [18] addressed the question of scalability of web crawlers and used the technique to partition URLs among the crawlers. One of the underlying assumptions in this work is the availability of high speed communication medium. The implemented prototype requested 800,000 pages from more than 70,000 web crawlers in 15 minutes.

In 2005, Exposto et al. [19] augmented partitioning of URLs among a set of crawling nodes in a P2P architecture by taking into account servers ge-

ographical information. Such an augmentation reduced the overall time of the crawl by allocating target servers to a node physically closest to them.

In 2008, an extremely scalable web crawler called *IRLbot* ran for 41.27 days on a quad-CPU AMD Opteron 2.6 GHz server and it crawled over 6.38 billion web pages [20]. IRLbot primarily addressed the *URL-Seen* problem by breaking it down into three sub-problems: CHECK, UPDATE and CHECK+UPDATE. To address these sub-problems, IRLbot introduced a framework called *Disk Repository with Update Management (DRUM)*. DRUM optimizes disk access by segmenting the disk into several *disk buckets*. For each disk bucket, DRUM also allocates a corresponding bucket on the RAM. Each URL is mapped to a bucket. At first a URL was stored in its RAM bucket. Once a bucket on the RAM is filled, the corresponding disk bucket is accessed in batch mode. This batch mode access, as well as the two-stage bucketing system used, allowed DRUM to store a large number of URLs on the disk such that its performance would not degrade as the number of URLs increases.

3 Crawling Deep Web

As server-side programming and scripting languages, such as PHP and ASP, got momentum, more and more databases became accessible online through interacting with a web application. The applications often delegated creation and generation of contents to the executable files using *Common Gateway Interface (CGI)*. In this model, programmers often hosted their data on databases and used HTML forms to query them. Thus a web crawler can not access all of the contents of a web application merely by following hyperlinks and downloading their corresponding web page. These contents are *hidden* from the web crawler point of view and thus are referred to as *deep web* [9].

In 1998, Lawrence and Giles [21] estimated that 80 percent of web contents were hidden in 1998. Later in 2000, BrightPlanet suggested that the deep web contents is 500 times larger than what surfaces through following hyperlinks (referred to as *shallow web*) [22]. The size of the deep web is rapidly growing as more companies are moving their data to databases and set up interfaces for the users to access them [22].

Only a small fraction of the deep web is indexed by search engines. In 2007, He et al. [9] ran-

domly sampled one million IPs and crawled these IPs looking for deep webs through HTML form elements. The study also defined a depth factor from the original seed IP address and constrained itself to depth of three. Among the sampled IPs, 126 deep web sites were found. These deep websites had 406 query gateways to 190 databases. Based on these results with 99 percent confidence interval, the study estimates that at the time of that writing, there existed 1,097,000 to 1,419,000 database query gateways on the web. The study further estimated that Google and Yahoo search engines each has visited only 32 percent of the deep web. To make the matters worse the study also estimated that 84 percent of the covered objects overlap between the two search engines, so combining the discovered objects by the two search engines does not increase the percentage of the visited deep web by much.

The second generation of web crawlers took the deep web into account. Information retrieval from the deep web meant interacting with HTML forms. To retrieve information hidden in the deep web, the web crawler would submit the HTML form many times, each time filled with a different dataset. Thus the problem of crawling the deep web got reduced to the problem of assigning proper values to the HTML form fields.

The open and difficult question to answer in designing a deep web crawler is how to meaningfully assign values to the fields in a query form [23]. As Barbosa and Freire [23] explain, it is easy to assign values to fields of certain types such as radio buttons. The difficult field to deal with, however, is text box inputs. Many different proposals tried to answer this question:

- In 2001, Raghavan and Garcia-Molina [7] proposed a method to fill up text box inputs that mostly depend on human output.
- In 2002, Liddle et al. [24] described a method to detect form elements and fabricate a HTTP GET and POST request using default values specified for each field. The proposed algorithm is not fully automated and asks for user input when required.
- In 2004, Barbosa and Freire [23] proposed a two phase algorithm to generate textual queries. The first stage collected a set of data from the website and used that to associate

weights to keywords. The second phase used a greedy algorithm to retrieve as much contents as possible with minimum number of queries.

- In 2005, Ntoulas et al. [25] further advanced the process by defining three policies for sending queries to the interface: a random policy, a policy based on the frequency of keywords in a reference document, and an adaptive policy that learns from the downloaded pages. Given four entry points, this study retrieved 90 percent of the deep web with only 100 requests.
- In 2008, Lu et al. [26] map the problem of maximizing the coverage per number of requests to the problem of *set-covering* [27] and uses a classical approach to solve this problem.

4 Crawling Rich Internet Applications

Powerful client side browsers and availability of client-side technologies lead to a shift in computation from server-side to the client-side. This shift of computation, also creates contents that are often hidden from traditional web-crawlers and are referred to as "Client-side hidden-web" [28]. In 2013, Behfarshad and Mesbah studies 500 websites and found that 95 percent of the subject websites contain client-side hidden-web, and among the 95 percent web-sites, 62 percent of the application states are considered client-side hidden-web. Extrapolating these numbers puts almost 59 percent of the web contents at the time of this writing as client-side hidden-web.

RIA crawling differs from traditional web application crawling in several frontiers. Although limited, there has been some research focusing on crawling of RIAs. One of the earliest attempts to crawl RIAs is by Duda et al. [29–31] in 2007. This work presents a working prototype of a RIA crawler that indexed RIAs using a Breath-First-Search algorithm. In 2008, Mesbah et al. [32, 33] introduced *Crawljax* a RIA crawler that took the user-interface into account and used the changes made to the user interface to direct the crawling strategy. *Crawljax* aimed at crawling and taking a static snapshot of each AJAX state for indexing and testing. In the same year, Amalfitano et al. [34–37]

addressed automatic testing of RIAs using execution traces obtained from AJAX applications.

This section surveys different aspects of RIA crawling. Different strategies can be used to choose an unexecuted event to execute. Different strategies effect how early the web crawler finds new states and the overall time of crawling. Section 4.1 surveys some of the strategies studied in recent years. Section 4.2 explains different approaches to determine if two DOMs are equivalent. Section 4.3 surveys parallelism and concurrency for RIA crawling. Automated testing and ranking algorithms are explored in Sections 4.4 and 4.5, respectively.

4.1 Crawling Strategy

Until recent years, there has not been much attention on the efficiency requirement, and existing approaches often use either Breadth-First or a Depth-First crawling strategy. In 2007, 2008 and 2009, Duda et al. [29–31] used Breadth-First crawling strategy. As an optimization, the communication cost was reduced by caching the JavaScript function calls (together with actual parameters) that resulted in AJAX requests and the response received from the server. Crawljax [32, 33] extracted a model of the application using a variation of the Depth-First strategy. Its default strategy only explored a subset of the events in each state. This strategy explored an event only from the state where the event was first encountered. The event was not explored on the subsequently discovered states. This default strategy may not find all the states, since executing the same event from different states may lead to different states. However, Crawljax can also be configured to explore all enabled events in each state, in that case its strategy becomes the standard Depth-First crawling strategy.

In 2008, 2009 and 2010, Amalfitano et al. [34–36] focused on modelling and testing RIAs using execution traces. The initial work [34] was based on obtaining execution traces from user-sessions (a manual method). Once the traces are obtained, they are analyzed and an FSM model is formed by grouping together the equivalent user interfaces according to an equivalence relation. In later study [35] introduced CrawlRIA which automatically generated execution traces using a Depth-First strategy. Starting from the initial state, CrawlRIA executed events in a depth-first manner until

a DOM instance that is equivalent to a previously visited DOM instance was reached. Then the sequence of states and events was stored as a trace in a database, and after a reset, crawling continued from the initial state to record another trace. These automatically generated traces were later used to form an FSM model using the same technique that is used in [34] for user-generated traces.

In 2011, Benjamin et al. [8] present the initial version of the first model-based crawling strategy: the *Hypercube strategy*. The strategy makes predictions by initially assuming the model of the application to be a hypercube structure. The initial implementation had performance drawbacks which prevented the strategy from being practical even when the number of events in the initial state are as few as twenty. These limitation were later removed [38].

In 2012, Choudhary et al. [39,40] introduced another model-based strategy called the *Menu strategy*. This strategy is optimized for the applications that have the same event always leading to the same state, irrelevant of the source state [39]. Dincurk et al. [41] introduced a statistical model-based strategy. This strategy uses statistics to determine which events have a high probability to lead to a new state [38].

In the same year, Peng et al. [42] suggested to use a *greedy strategy*. In the greedy strategy if there is an un-executed event in the current state (i.e. the state which the web crawler’s DOM structure represents) the event is executed. If the current state has no unexplored event, the web crawler transfers to the closest state with an unexecuted event. Two other variants of the greedy strategy are introduced by the authors as well. In these variations, instead of the closest state, the most recently discovered state and the state closest to the initial state are chosen when there is no event to explore in the current state. They experimented with this strategy on simple test applications using different combinations of navigation styles to navigate a sequence of ordered pages. The navigation styles used are previous and next events, events leading to a few of the preceding and succeeding pages from the current page, as well as the events that lead to the first and last page. They concluded that all three variations of the strategy have similar performance in terms of the total number of event executions to finish crawling.

In 2013, Milani Fard and Mesbah [43] introduce

FeedEx: a greedy algorithm to partially crawl a RIAs. FeedEx differs from Peng et al. [42] in that: Peng et al. [42] use a greedy algorithm in finding the closest unexecuted event, whereas, FeedEx defines a matrix to measure the impact of an event and its corresponding state on the crawl. The choices are then sorted and the most impactful choice will be executed first. Given enough time, FeedEx will discover entire graph of the application.

FeedEx defines the impact matrix as a weighted sum of the following four factors:

- Code coverage impact: how much of the application code is being executed.
- Navigational diversity: how diversely the crawler explores the application graph.
- Page structural diversity: how newly discovered DOMs differ from those already discovered.
- Test model size: the size of the created test model.

In the test cases studied, Milani Fard and Mesbah [43] show that FeedEx beats three other strategies of Breadth-First search, Depth-First search, and random strategy, in the above-mentioned four factors.

4.2 DOM Equivalence and Comparison

In the context of traditional web applications it is trivial to determine whether two states are equal: compare their URLs. This problem is not as trivial in the context of RIAs. Different chains of events may lead to the same states with minor differences that do not effect the functionality of the state. Different researchers address this issue differently. Duda et al. [29–31] used equality as the DOM equivalence method. Two states are compared based on “the hash value of the full serialized DOM” [31]. As admitted in [31] this equality is too strict and may lead to too many states being produced.

Crawljax [32] used an edit distance (the number of operations that is needed to change one DOM instance to the other, the so-called Levenstein distance) to decide if the current DOM instance corresponds to a different state than the previous one. If

the distance is below a certain threshold the current DOM instance is considered equivalent to the previous one. Otherwise, the current DOM instance is hashed and its hash value is compared to the hash values of the already discovered states. Since the notion of distance is not transitive, it is not an equivalence relation in the mathematical sense. For this reason, using a distance has the problem of incorrectly grouping together client-states whose distance is actually above the given threshold.

In a later paper [33], Crawljax improves its DOM equivalency: To decide if a new state is reached, the current DOM instance is compared with all the previously discovered DOMs using the mentioned distance heuristic. If the distance of the current DOM instance from each seen DOM instance is above the threshold, then the current DOM is considered as a new state. Although this approach solves the mentioned problem with the previous approach, this method may not be as efficient since it requires to store the DOM-trees and compute the distance of the current DOM to all the discovered DOMs.

Amalfitano et al. [36] proposed DOM equivalence relations based on comparing the set of elements in the two DOM instances. According to this method, two DOM instances are equivalent if both contain the same set of elements. This inclusion is checked based on the indexed paths of the elements, event types and event handlers of the elements. They have also introduced two variations of this relation. In the first variation only visible elements are considered, in the other variation, the index requirement for the paths is removed.

In 2013, Lo et al. [44] in a tool called *Imagen*, consider the problem of transferring a JavaScript session between two clients. Imagen improves the definition of client-side state by adding the following items:

- JavaScript functions closure: JavaScript functions can be created dynamically, and their scope is determined at the time of creation.
- JavaScript event listeners: JavaScript allows the programmer to register event-handlers.
- HTML5 elements: Certain elements such as *Opaque Objects* and *Stream Resources*.

These items are not ordinarily stored in DOM. Imagen uses code instrumenting and other techniques to add the effect of these features to the state of the

application. To the best of our knowledge, Imagen offers the most powerful definition of a RIA state at the time of this writing. This definition has not been used by any web crawler yet, and its effect on the web crawler performance is an open research topic.

4.3 Parallel Crawling

To the best of our knowledge, at the time of this writing only one distributed RIA crawling algorithm exists. Mirtaheeri et al. [45] used the JavaScript events to partition the search space and crawl a RIA in parallel. Each web crawler, running on a separate computer, visits all application states, but only executes a subset of the JavaScript events in each state. If execution of an event leads to the discovery of a new state, the information about the new state is propagated to all the web crawlers. Together, the web crawlers cover every JavaScript events in every application states. The proposed algorithm is implemented and evaluated with 15 computers and a satisfactory speedup is demonstrated. Apart from this work, two algorithms are proposed to achieve a degree of concurrency:

- Matter [30], proposed to use multiple web crawlers on RIAs that use hyperlinks together with events for navigation. The suggested method first applies traditional crawling to find the URLs in the application. After traditional crawling terminates, the set of discovered URLs are partitioned and assigned to event-based crawling processes that run independent of each other using their Breadth-First strategy. Since each URL is crawled independently, there is no communication between the web crawlers.
- Crawljax [33] used multiple threads for speeding up event-based crawling of a single URL application. The crawling process starts with a single thread (that uses depth-first strategy). When a thread discovers a state with more than one event, new threads are initiated that will start the exploration from the discovered state and follow one of the unexplored events from there.

4.4 Automated Testing⁴

Automated testing of RIAs is an important aspect of RIA crawling. In 2008, Marchetto et al. [47] used a state-based testing approach based on a FSM model of the application. The introduced model construction method used static analysis of the JavaScript code and dynamic analysis of user session traces. Abstraction of the DOM states was used rather than the DOM states directly in order to reduce the size of the model. This optimization may require a certain level of manual interaction to ensure correctness of the algorithm. The introduced model produced test sequences that contained *semantically interacting events*⁵. In 2009, Marchetto and Tonella [48] proposed search-based test sequence generation using hill-climbing rather than exhaustively generating all the sequences up to some maximum length.

In 2009 and 2010, Crawljax introduced three mechanisms to automate testing of RIAs: Using *invariant-based* testing [49], security testing of interactions among web widgets [50], and regression testing of AJAX applications [50].

In 2010, Amalfitono et al. [35] compared the effectiveness of methods based on execution traces (user generated, web crawler generated and combination of the two) and existing test case reduction techniques based on measures such as state coverage, transition coverage and detecting JavaScript faults. In another study [37], authors used invariant-based testing approach to detect faults visible on the user-interface (invalid HTML, broken links, unsatisfied accessibility requirements) in addition to JavaScript faults (crashes) which may not be visible on the user-interface, but cause faulty behaviour.

4.5 Ranking (Importance Metric)

Unlike traditional web application crawling, there has been a limited amount of research in ranking states and pages in the context of RIA crawling. In 2007, Frey [31] proposed a ranking mechanism for the states in RIAs. The proposed mechanism, called *AjaxRank*, ordered search results by assign-

⁴For a detailed study of web application testing trends from 2000 to 2011 see Garousic et al. [46]

⁵Two events are semantically interacting if their execution order changes the outcome.

ing an importance value to states. AjaxRank can be viewed as an adaptation of the PageRank [51]. Similar to PageRank, AjaxRank is connectivity-based but instead of hyperlinks the transitions are considered. In the AjaxRank, the initial state of the URL is given more importance (since it is the only state reachable from anywhere directly), hence the states that are closer to the initial state also get higher ranks.

5 Taxonomy and Evolution of Web Crawlers

The wide variety of web crawlers available are designed with different goals in mind. This section classifies and cross-measures the functionalities of different web crawlers based on the design criteria introduced in Section 1.4. It also sketches out a rough architecture of web crawlers as they evolve. Sections 5.1, 5.2 and 5.3 explain the taxonomy of traditional, deep, and RIA web crawlers, respectively.

5.1 Traditional Web Crawlers

Figure 1 shows the architecture of a typical traditional web crawler. In this model *Frontier* gets a set of seed URLs. The seed URLs are passed to a module called *Fetcher* that retrieves the contents of the pages associated with the URLs from the web. These contents are passed to the *Link Extractor*. The latter parses the HTML pages and extracts new links from them. Newly discovered links are passed to *Page Filter* and *Store Processor*. Store Processor interacts with the database and stores the discovered links. Page Filter filters URLs that are not interesting to the web crawler. The URLs are then passed to *URL-Seen* module. This module finds the new URLs that are not retrieved yet and passes them to *Fetcher* for retrieval. This loop continues until all the reachable links are visited.

Table 2 summarizes the design components, design goals and different techniques used by traditional web crawlers.

5.2 Deep Web Crawlers

Figure 2 shows the architecture of a typical deep web crawler. In this model *Select Fillable* gets as input set of seed URLs, domain data, and user

specifics. *Select Fillable* then chooses the HTML elements to interact with. *Domain Finder* uses these data to fill up the HTML forms and passes the results to *Submitter*. Submitter submits the form to the server and retrieves the newly formed page. *Response Analyser* parses the page and, based on the result, updates the repository; and the process continues.

Table 3 summarizes the design components, design goals and different techniques used by deep web crawlers.

5.3 RIA Web Crawlers

Figure 3 shows the architecture of a typical RIA web crawler. *JS-engine* starts a virtual browser and runs a JavaScript engine. It then retrieves the page associated with a seed URL and loads it in the virtual browser. The constructed DOM is passed to the *DOM-Seen* module to determine if this is the first time the DOM is seen. If so, the DOM is passed to *Event Extractor* to extract the JavaScript events from it. The events are then passed to the *Strategy* module. This module decides which event to execute. The chosen event is passed to JS-Engine for further execution. This process continues until all reachable states are seen.

Table 4 summarizes the design components, design goals and different techniques used by RIA web crawlers.

6 Some Open Questions in Web-Crawling

In this paper, we have surveyed the evolution of crawlers, namely traditional, Deep and RIA crawlers. We identified several design goals and components of each category and developed a taxonomy that classifies different cases of crawlers accordingly. Traditional web crawling and its scalability has been the topic of extensive research. Similarly, deep-web crawling was addressed in great detail. RIA crawling, however, is a new and open area for research. Some of the open questions in the field of RIA crawling are the following:

- *Model based crawling*: The problem of designing an efficient strategy for crawling a RIA can be mapped to a graph exploration problem. The objective of the algorithm is to visit every node at least once in an unknown

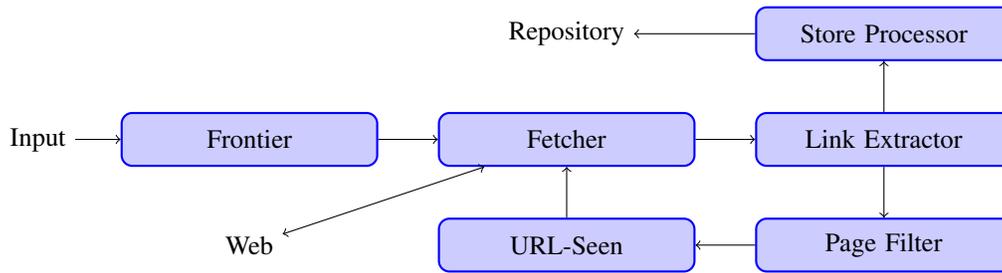


Figure 1: Architecture of a Traditional Web Crawler.

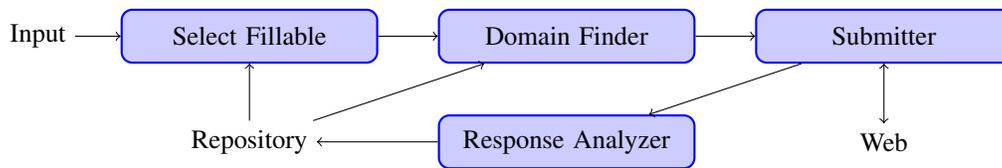


Figure 2: Architecture of a Deep Web Crawler.

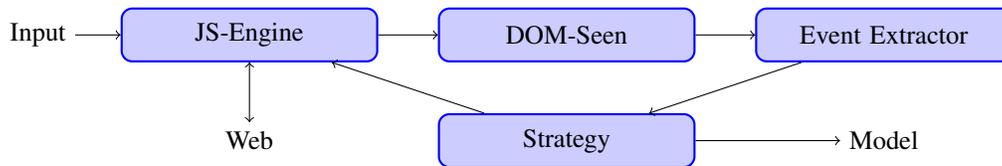


Figure 3: Architecture of a RIA Web Crawler.

Table 2: Taxonomy of Traditional Web Crawlers

Study	Component	Method	Goal
WebCrawler, MOM-spider [4]	Fetcher, Frontier, Page filter	Parallel downloading of 15 links, robots.txt, Black-list	Scalability, Politeness
Google [12]	Store processor, Frontier	Reduce disk access time by compression, PageRank	Scalability, Coverage, Freshness
Mercator [5]	URL-Seen	Batch disk checks, cache	Scalability
WebFountain [13]	Storage processor, Frontier, Fetch	Local copy of the fetched pages, Adaptive download rate, Homogenous cluster as hardware	Completeness, Freshness, Scalability
Polybot [14]	URL-Seen	Red-Black tree to keep the URLs	Scalability
UbiCrawler [15]	URL-Seen	P2P architecture	Scalability
pSearch [16]	Store processor	Distributed Hashing Tables (DHT)	Scalability
Exposto et al. [19]	Frontier	Distributed Hashing Tables (DHT)	Scalability
IRLbotpages [20]	URL-Seen	Access time reduction by disk segmentation	Scalability

Table 3: Taxonomy of Deep Web Crawlers

Study	Component	Method	Goal
HiWe [7]	Select fillable, Domain Finder, Submitter, Response Analyst	Partial page layout and visual adjacency, Normalization by stemming etc, Approximation matching, Manual domain, Ignore submitting small or incomplete forms, Hash of visually important parts of the page to detect errors	Lenient submission efficiency, Submission efficiency
Liddle et al. [24]	Select fillable, Domain Finder	Fields with finite set of values, ignores automatic filling of text field, Stratified Sampling Method (avoid queries biased toward certain fields), Detection of new forms inside result page, Removal of repeated form Concatenation of pages connected through navigational elements, Stop queries by observing pages with repetitive partial results, Detect record boundaries and computes hash values for each sentence	Lenient submission efficiency, Submission efficiency
Barbosa and Freire [23]	Select fillable, Domain Finder, Response Analysis	Single keyword-based queries, Based on collection data associate weights to keywords and uses greedy algorithms to retrieve as much contents with minimum number of queries, Considers adding stop-words to maximize coverage, Issue queries using dummy words to detect error pages	Lenient submission efficiency, Submission efficiency
Ntoulas et al. [25]	Select fillable, Domain Finder	Single-term keyword-based queries, Three policies: random, based on the frequency of keyword in a corpus, and an Adaptive policy that learn from the downloaded pages. maximizing the unique returns of each query	Lenient submission efficiency, Submission efficiency
Lu et al. [26]	Select fillable, Domain Finder	querying textual data sources, Works on sample that represents the original data source, Maximizing the coverage	Lenient submission efficiency, Scalability, Submission efficiency

Table 4: Taxonomy of RIA Web Crawlers

Study	Component	Method	Goal
Duda et al. [29–31]	Strategy, JS-Engine, DOM-Seen	Breadth-First-Search, Caching the JavaScript function calls and results, Comparing Hash value of the full serialized DOM	Completeness, Efficiency
Mesbah et al. [32,33]	Strategy, DOM-Seen	Depth-First-Search, Explores an event only once, New threads are initiated for unexplored events, Comparing Edit distance with all previous states	Completeness, State Coverage Efficiency, Scalability
CrawlRIA [34–37]	Strategy, DOM-Seen	Depth-First strategy (Automatically generated using execution traces), Comparing the set of elements, event types, event handlers in two DOMs	Completeness
Kamara et al. [8,52]	Strategy	Assuming hypercube model for the application, Using Minimum Chain Decomposition and Minimum Transition Coverage	State Coverage Efficiency
M-Crawler [53]	Strategy	Menu strategy which categorizes events after first two runs, Events which always lead to the same/current state has less priority, Using Rural-Postman solver to explore unexecuted events efficiently	State Coverage Efficiency, Completeness
Peng et al. [42]	Strategy	Choose an event from current state then from the closest state	State Coverage Efficiency
AjaxRank [31]	Strategy, DOM-Seen	The initial state of the URL is given more importance, Similar to PageRank, connectivity-based but instead of hyperlinks the transitions are considered hash value of the content and structure of the DOM	State Coverage Efficiency
Dincturk et al. [41]	Strategy	Considers probability of discovering new state by an event and cost of following the path to event's state	State Coverage Efficiency
Dist-RIA Crawler [45]	Strategy	Uses JavaScript events to partition the search space and run the crawl in parallel on multiple nodes	Scalability
Feedex [43]	Strategy	Prioritize events based on their possible impact of the DOM, Considers factors like code coverage, navigational and page structural diversity	State Coverage Efficiency

directed graph by minimizing the total sum of the weights of the edges traversed. The offline version of this problem, where the graph is known beforehand, is called the Asymmetric Traveling Salesman Problem (ATSP) which is NP-Hard. Although there are some approximation algorithms for different variations of the unknown graph exploration problem [54–57], not knowing the graph ahead of the time is a major obstacle to deploy these algorithms to crawl RIAs.

- *Scalability*: Problems such as URL-Seen may not exist in the context of RIA crawling. However, a related problem is the *State-Seen* problem: If a DOM state was seen before.
- *Widget detection*: In order to avoid state explosion, it is crucial to detect independent parts of the interface in a RIA. This can effect ranking of different states, too.
- *Definition of state in RIA*: some HTML5 elements such as *web sockets* introduce new challenges to the web crawlers. Some of the challenges are addressed by Imagen [44], however many of these challenges remain open.

In addition, combining different types of crawlers to build a unified crawler seems another promising research area.

Acknowledgements

This work is largely supported by the IBM® Center for Advanced Studies, the IBM Ottawa Lab and the Natural Sciences and Engineering Research Council of Canada (NSERC). A special thank to Anseok Joo and Di Zou.

Trademarks

IBM, the IBM logo, ibm.com and AppScan are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml. Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

References

- [1] J. Bau, E. Bursztein, D. Gupta, and J. Mitchell, “State of the art: Automated black-box web application vulnerability testing,” in *Security and Privacy (SP), 2010 IEEE Symposium on*. IEEE, 2010, pp. 332–345.
- [2] A. Doupé, M. Cova, and G. Vigna, “Why johnny can’t pentest: An analysis of black-box web vulnerability scanners,” in *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2010, pp. 111–131.
- [3] J. Marini, *Document Object Model*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 2002.
- [4] C. Olston and M. Najork, “Web crawling,” *Foundations and Trends in Information Retrieval*, vol. 4, no. 3, pp. 175–246, 2010.
- [5] A. Heydon and M. Najork, “Mercator: A scalable, extensible web crawler,” *World Wide Web*, vol. 2, pp. 219–229, 1999.
- [6] M. Burner, “Crawling towards eternity: Building an archive of the world wide web,” *Web Techniques Magazine*, vol. 2, no. 5, May 1997.
- [7] S. Raghavan and H. Garcia-Molina, “Crawling the hidden web,” in *Proceedings of the 27th International Conference on Very Large Data Bases*, ser. VLDB ’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 129–138.
- [8] K. Benjamin, G. Von Bochmann, M. E. Dincturk, G.-V. Jourdan, and I. V. Onut, “A strategy for efficient crawling of rich internet applications,” in *Proceedings of the 11th international conference on Web engineering*, ser. ICWE’11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 74–89.
- [9] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang, “Accessing the deep web,” *Commun. ACM*, vol. 50, no. 5, pp. 94–101, May 2007.
- [10] S. Choudhary, M. E. Dincturk, S. M. Mirtaheri, A. Moosavi, G. von Bochmann, G.-V. Jourdan, and I.-V. Onut, “Crawling rich internet applications: the state of the art.” in *CASCON*, 2012, pp. 146–160.
- [11] O. A. McBryan, “Genvl and www: Tools for taming the web,” in *In Proceedings of the First International World Wide Web Conference*, 1994, pp. 79–90.
- [12] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” in *Proceedings of the seventh international conference on World Wide Web 7*, ser. WWW7. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 1998, pp. 107–117.

- [13] J. Edwards, K. McCurley, and J. Tomlin, "An adaptive model for optimizing performance of an incremental web crawler," 2001.
- [14] V. Shkapenyuk and T. Suel, "Design and implementation of a high-performance distributed web crawler," in *In Proc. of the Int. Conf. on Data Engineering*, 2002, pp. 357–368.
- [15] P. Boldi, B. Codenotti, M. Santini, and S. Vigna, "Ubicrawler: A scalable fully distributed web crawler," *Proc Australian World Wide Web Conference*, vol. 34, no. 8, pp. 711–726, 2002.
- [16] C. Tang, Z. Xu, and M. Mahalingam, "psearch: Information retrieval in structured overlays," 2002.
- [17] J. Li, B. Loo, J. Hellerstein, M. Kaashoek, D. Karger, and R. Morris, "On the feasibility of peer-to-peer web indexing and search," *Peer-to-Peer Systems II*, pp. 207–215, 2003.
- [18] B. T. Loo, S. Krishnamurthy, and O. Cooper, "Distributed web crawling over dhts," EECS Department, University of California, Berkeley, Tech. Rep. UCB/CSD-04-1305, 2004.
- [19] J. Exposto, J. Macedo, A. Pina, A. Alves, and J. Rufino, "Geographical partition for distributed web crawling," in *Proceedings of the 2005 workshop on Geographic information retrieval*, ser. GIR '05. New York, NY, USA: ACM, 2005, pp. 55–60.
- [20] H. tsang Lee, D. Leonard, X. Wang, and D. Loguinov, "Irlbot: Scaling to 6 billion pages and beyond," 2008.
- [21] S. Lawrence and C. L. Giles, "Searching the world wide web," *SCIENCE*, vol. 280, no. 5360, pp. 98–100, 1998.
- [22] M. K. Bergman, "The deep web: Surfacing hidden value," September 2001.
- [23] L. Barbosa and J. Freire, "Siphoning hidden-web data through keyword-based interfaces," in *SBBD*, 2004, pp. 309–321.
- [24] S. W. Liddle, D. W. Embley, D. T. Scott, and S. H. Yau1, "Extracting Data behind Web Forms," *Lecture Notes in Computer Science*, vol. 2784, pp. 402–413, Jan. 2003.
- [25] A. Ntoulas, "Downloading textual hidden web content through keyword queries," in *In JCDL*, 2005, pp. 100–109.
- [26] J. Lu, Y. Wang, J. Liang, J. Chen, and J. Liu, "An Approach to Deep Web Crawling by Sampling," *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, vol. 1, pp. 718–724, 2008.
- [27] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. The MIT Press, 2009.
- [28] Z. Behfarshad and A. Mesbah, "Hidden-web induced by client-side scripting: An empirical study," in *Proceedings of the International Conference on Web Engineering (ICWE)*, ser. Lecture Notes in Computer Science, vol. 7977. Springer, 2013, pp. 52–67.
- [29] C. Duda, G. Frey, D. Kossmann, R. Matter, and C. Zhou, "Ajax crawl: Making ajax applications searchable," in *Proceedings of the 2009 IEEE International Conference on Data Engineering*, ser. ICDE '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 78–89.
- [30] R. Matter, "Ajax crawl: Making ajax applications searchable," Master's thesis, ETH Zurich, 2008, <http://e-collection.library.ethz.ch/eserv/eth:30709/eth-30709-01.pdf>.
- [31] G. Frey, "Indexing ajax web applications," Master's thesis, ETH Zurich, 2007, <http://e-collection.library.ethz.ch/eserv/eth:30111/eth-30111-01.pdf>.
- [32] A. Mesbah, E. Bozdag, and A. v. Deursen, "Crawling ajax by inferring user interface state changes," in *Proceedings of the 2008 Eighth International Conference on Web Engineering*, ser. ICWE '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 122–134.
- [33] A. Mesbah, A. van Deursen, and S. Lenselink, "Crawling ajax-based web applications through dynamic analysis of user interface state changes," *TWEB*, vol. 6, no. 1, p. 3, 2012.
- [34] D. Amalfitano, A. R. Fasolino, and P. Tramontana, "Reverse engineering finite state machines from rich internet applications," in *Proceedings of the 2008 15th Working Conference on Reverse Engineering*, ser. WCRE '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 69–73.
- [35] D. Amalfitano, A. Fasolino, and P. Tramontana, "Rich internet application testing using execution trace data," *Software Testing Verification and Validation Workshop, IEEE International Conference on*, vol. 0, pp. 274–283, 2010.
- [36] D. e. a. Amalfitano, "Experimenting a reverse engineering technique for modelling the behaviour of rich internet applications," in *Software Maintenance, 2009. ICSM 2009. IEEE International Conference on*, sept. 2009, pp. 571–574.
- [37] D. Amalfitano, A. Fasolino, and P. Tramontana, "Techniques and tools for rich internet applications testing," in *Web Systems Evolution (WSE), 2010 12th IEEE International Symposium on*, sept. 2010, pp. 63–72.

- [38] M. E. Dincturk, "Model-based crawling - an approach to design efficient crawling strategies for rich internet applications," Master's thesis, EECS - University of Ottawa, 2013, http://ssrg.eecs.uottawa.ca/docs/Dincturk-MustafaEmre_2013_thesis.pdf.
- [39] S. Choudhary, E. Dincturk, S. Mirtaheri, G.-V. Jourdan, G. Bochmann, and I. Onut, "Building rich internet applications models: Example of a better strategy," in *Web Engineering*, ser. Lecture Notes in Computer Science, F. Daniel, P. Dolog, and Q. Li, Eds. Springer Berlin Heidelberg, 2013, vol. 7977, pp. 291–305.
- [40] S. Choudhary, "M-crawler: Crawling rich internet applications using menu meta-model," Master's thesis, EECS - University of Ottawa, 2012, <http://ssrg.site.uottawa.ca/docs/Surya-Thesis.pdf>.
- [41] M. E. Dincturk, S. Choudhary, G. von Bochmann, G.-V. Jourdan, and I.-V. Onut, "A statistical approach for efficient crawling of rich internet applications," in *ICWE*, 2012, pp. 362–369.
- [42] Z. Peng, N. He, C. Jiang, Z. Li, L. Xu, Y. Li, and Y. Ren, "Graph-based ajax crawl: Mining data from rich internet applications," in *Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on*, vol. 3, march 2012, pp. 590–594.
- [43] A. Milani Fard and A. Mesbah, "Feedback-directed exploration of web applications to derive test models," in *Proceedings of the 24th IEEE International Symposium on Software Reliability Engineering (ISSRE)*. IEEE Computer Society, 2013, p. 10 pages.
- [44] J. Lo, E. Wohlstadter, and A. Mesbah, "Imagen: Runtime migration of browser sessions for javascript web applications," in *Proceedings of the International World Wide Web Conference (WWW)*. ACM, 2013, pp. 815–825.
- [45] S. M. Mirtaheri, D. Zou, G. V. Bochmann, G.-V. Jourdan, and I. V. Onut, "Dist-ria crawler: A distributed crawler for rich internet applications," in *In Proc. 8TH INTERNATIONAL CONFERENCE ON P2P, PARALLEL, GRID, CLOUD AND INTERNET COMPUTING*, 2013.
- [46] V. Garousi, A. Mesbah, A. Betin Can, and S. Mirshokraie, "A systematic mapping study of web application testing," *Information and Software Technology*, vol. 55, no. 8, pp. 1374–1396, 2013.
- [47] A. Marchetto, P. Tonella, and F. Ricca, "State-based testing of ajax web applications," in *Proceedings of the 2008 International Conference on Software Testing, Verification, and Validation*, ser. ICST '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 121–130.
- [48] A. Marchetto and P. Tonella, "Search-based testing of ajax web applications," in *Proceedings of the 2009 1st International Symposium on Search Based Software Engineering*, ser. SSBSE '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 3–12.
- [49] A. Mesbah and A. van Deursen, "Invariant-based automatic testing of ajax user interfaces," in *Software Engineering, 2009. ICSE 2009. IEEE 31st International Conference on*, may 2009, pp. 210 – 220.
- [50] D. Roest, A. Mesbah, and A. van Deursen, "Regression testing ajax applications: Coping with dynamism." in *ICST*. IEEE Computer Society, 2010, pp. 127–136.
- [51] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," 1998, standford University, Technical Report.
- [52] K. Benjamin, "A strategy for efficient crawling of rich internet applications," Master's thesis, EECS - University of Ottawa, 2010, <http://ssrg.eecs.uottawa.ca/docs/Benjamin-Thesis.pdf>.
- [53] S. Choudhary, M. E. Dincturk, G. von Bochmann, G.-V. Jourdan, I.-V. Onut, and P. Ionescu, "Solving some modeling challenges when testing rich internet applications for security," in *ICST*, 2012, pp. 850–857.
- [54] N. Megow, K. Mehlhorn, and P. Schweitzer, "Online graph exploration: new results on old and new algorithms," in *Proceedings of the 38th international conference on Automata, languages and programming - Volume Part II*, ser. ICALP'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 478–489.
- [55] S. Dobrev, R. Královič, and E. Markou, "Online graph exploration with advice," in *Structural Information and Communication Complexity*, ser. Lecture Notes in Computer Science, G. Even and M. Halldórsson, Eds. Springer Berlin Heidelberg, 2012, vol. 7355, pp. 267–278.
- [56] K.-T. Förster and R. Wattenhofer, "Directed graph exploration," in *Principles of Distributed Systems*, ser. Lecture Notes in Computer Science, R. Baldoni, P. Flocchini, and R. Binoy, Eds. Springer Berlin Heidelberg, 2012, vol. 7702, pp. 151–165.
- [57] R. Fleischer, T. Kamphans, R. Klein, E. Langetepe, and G. Trippen, "Competitive online approximation of the optimal search ratio," in *In Proc. 12th Annu. European Sympos. Algorithms, volume 3221 of Lecture Notes Comput. Sci.* Springer-Verlag, 2004, pp. 335–346.