SOFTWARE SECURITY RESEARCH GROUP

In Collaboration With IBM

Pdist-RIA Crawler: A P2P architecture to crawl RIAs

Seyed M. Mirtaheri, Gregor v. Bochmann, Guy-Vincent Jourdan, Iosif Viorel Onut

School of Information Technology and Engineering - University of Ottawa

Introduction

Rich Internet Applications (RIAs) allow better user interaction and responsiveness than traditional web applications.
Thanks to new technologies like AJAX (Asynchronous JavaScript and XML), Rich Internet Applications can communicate with the server asynchronously. This allows continuous user interactions.

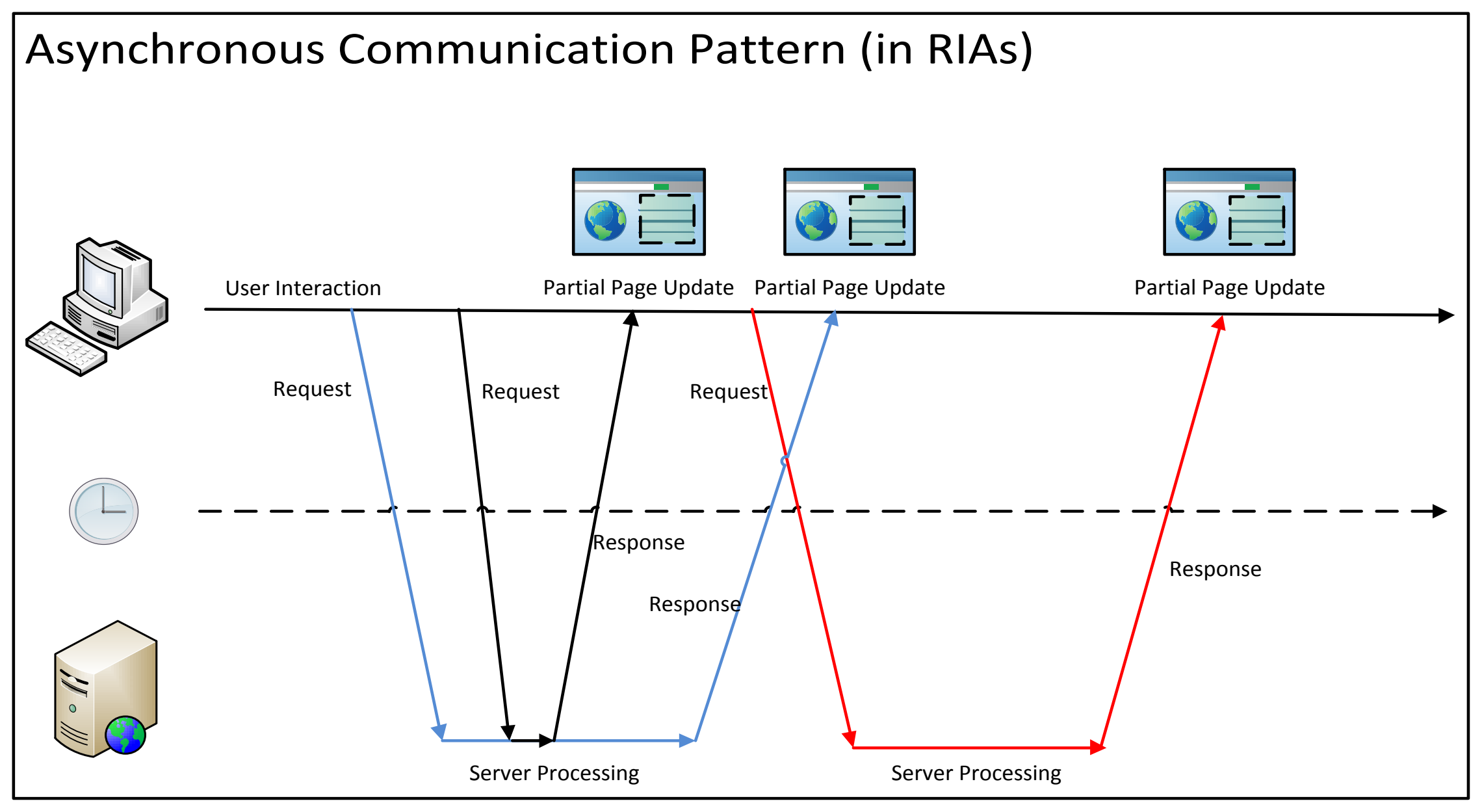


Figure 1: AJAX enabled RIA communication pattern.

Security of RIA and automating security testing are important, ongoing, and growing concerns. One important aspect of this automation is the crawling of RIAs i.e. reaching all possible states of the application from the initial state. Being able to do so automatically is also valuable for search engines and accessibility assessment.

Crawling of RIA applications is an expensive and time consuming process due to their large number of states. To accelerate this operation we distribute the operation over many nodes in an elastic cloud environment.

Background

- **Crawling Strategy**:
  - ➢ Breath-First Search
  - ➢ Bounded Depth-First Search
  - ➢ Based on page weight
  - ➢ Model Based Crawling
    - ➢Hypercube Model
    - ➢Menu Model
  - ➢ Greedy algorithm
  - ➢ Probabilistic model
- **Partitioning strategies**: Mostly use server related matrix as primary tool to partition search space:
  - ➢ Page URL
  - ➢ Server IP address
  - ➢ Server geographical location
  - ➢ [Loo 2004] describes distributed web crawling by hashing the URL

Architecture of AJAX Crawl (Appeared in [Duda 2009])

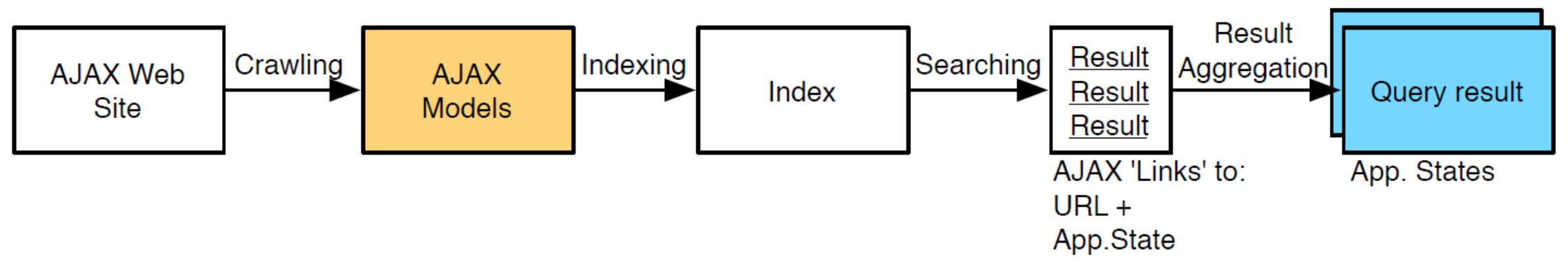


Figure 2: [Loo 2004] System Architecture

References:

- [Amalfitano 2010] D. Amalfitano, A.R. Fasoline, P. Tramontana, Techniques and tools for Rich Internet Application testing, in Proc. WSE, 2010, pp.63-72.
- [Mesbah 2008] A. Mesbah and A. van Deursen, A Component- and Push-based Architectural Style for Ajax Applications. 2008, Journal of Systems and Software (JSS) 81(12):2194-2209
- [Duda 2009] C. Duda, G. Frey, D. Kossmann, R. Matter, C. Zhou, AJAX Crawl: Making AJAX Applications Searchable, Proc. IEEE Intern. Conf. on Data Engineering, Shanghai, 2009, pp. 78-
- [Loo 2004] B.T. Loo, O. Cooper, S.Krishnamurthy, Distributed Web Crawling over DHTs, Technical report, University of California, Berkeley, 2004, http://www.eecs.berkeley.edu/Pubs/TechRpts/2004/5370.html
- [Exposto 2008] J. Exposto, J. Macedo, A. Pina, A. Alves and J. Rufino, Efficient partitioning strategies for distributed web crawling, in Information Networking: Towards Ubiquitous Networking and Services, Springer LNCS 5200 (2008), pp. 544-553.
- [Boldi 2003] P. Boldi, B. Codenotti, M. Santini, S.Vigna, UbiCrawler: A Scalable Fully Distributed Web Crawler, Software: Practice & Experience, Vol. 34, 2003, p. 2004.
- [Dincturk 2014] Dincturk, Mustafa Emre and Jourdan, Guy-Vincent and Bochmann, Gregor von and Onut, Iosif Viorel: Model Based Crawling, ACM Transactions on the WEB
- [Dincturk 2012] Dincturk, Mustafa Emre and Choudhary, Suryakant and Bochmann, Gregor von and Jourdan, Guy-Vincent and Onut, Iosif Viorel: A Statistical Approach for Efficient Crawling of Rich Internet Applications, Proceedings of the 12th international conference on Web engineering

Motivation and Aim

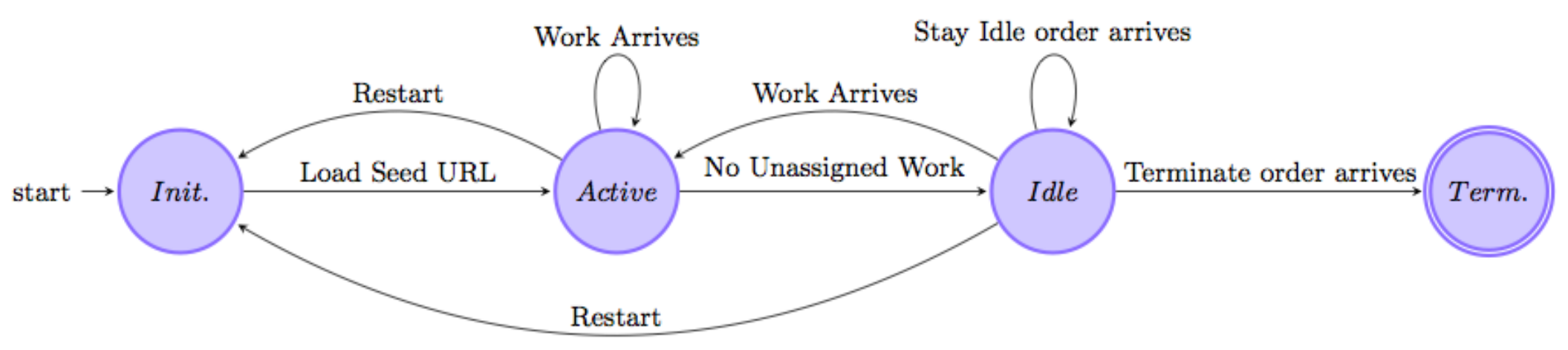
- **Non-URL-Based Crawling strategy**:
  - ➢ In a RIA one URL corresponds to many states of DOM. Unlike traditional websites in which every call to server would change the whole DOM and the page URL, RIA relies on small AJAX updates that does not necessarily modify the page URL:
  - ➢ Traditional distributed crawlers rely heavily on URL in order to partition the search space. Underlying assumption for this strategy is a one to one correspondence between the URL and the state of DOM which does not hold in RIA.
  - ➢ Therefore we propose to partition the search space based on events.
- **Crawling Strategy:** Reduce the workload by choosing the events to execute using Greedy algorithm.
- **Crawling Efficiency**: Discover states as soon as possible, using Probabilistic model.

Reference:

- [Benjamin 2010] K. Benjamin, G. v. Bochmann, G.-V. Jourdan and V. Onut, Some modeling challenges when testing Rich Internet Applications for security, First Intern. Workshop on Modeling and Detection of Vulnerabilities (MDV 2010), Paris, France, April 2010. 8 pages.

Proposed Architecture

- Nodes act autonomously and independently. Each node starts at Init state, when get tasks go to Active state, when has nothing to do goes to idle state, and finally terminates when termination order arrives.



- A virtual ring is created based on breath-first-search traversal of the nodes. A termination token goes around this wring that keeps the list of states IDs and the number of states visited by each node. When all states are visited on all nodes, a termination order is broadcasted.

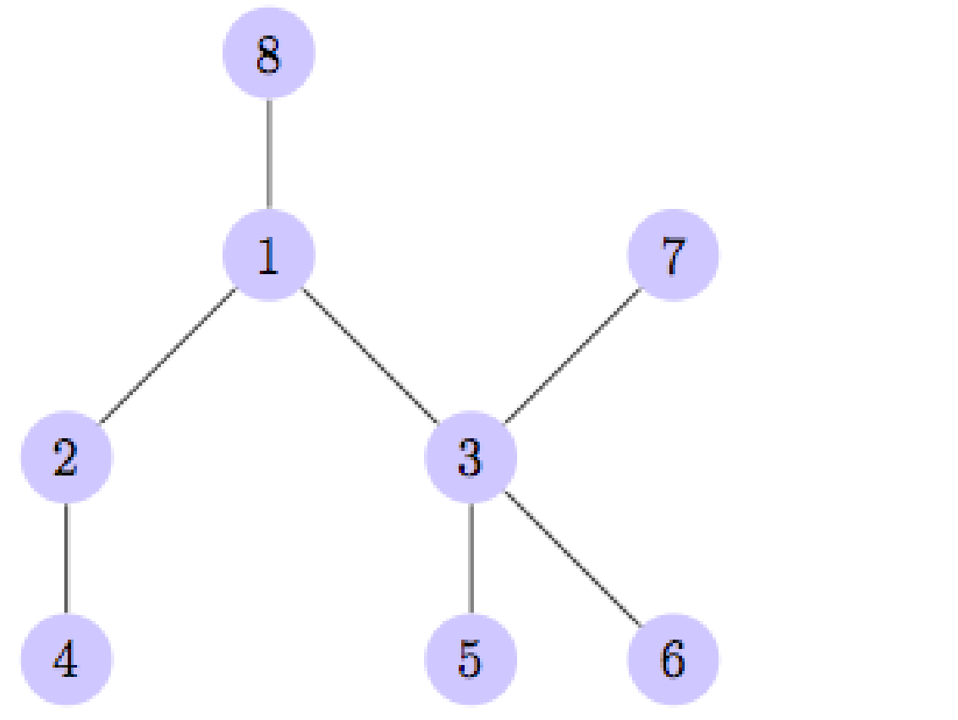


Figure 5.1: Sample virtual spanning tree

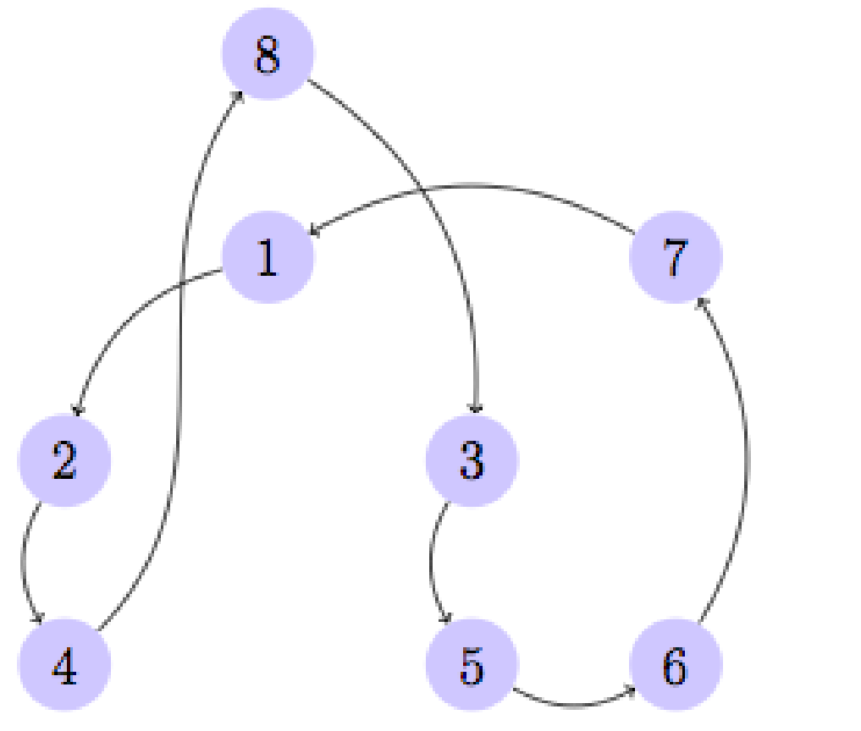
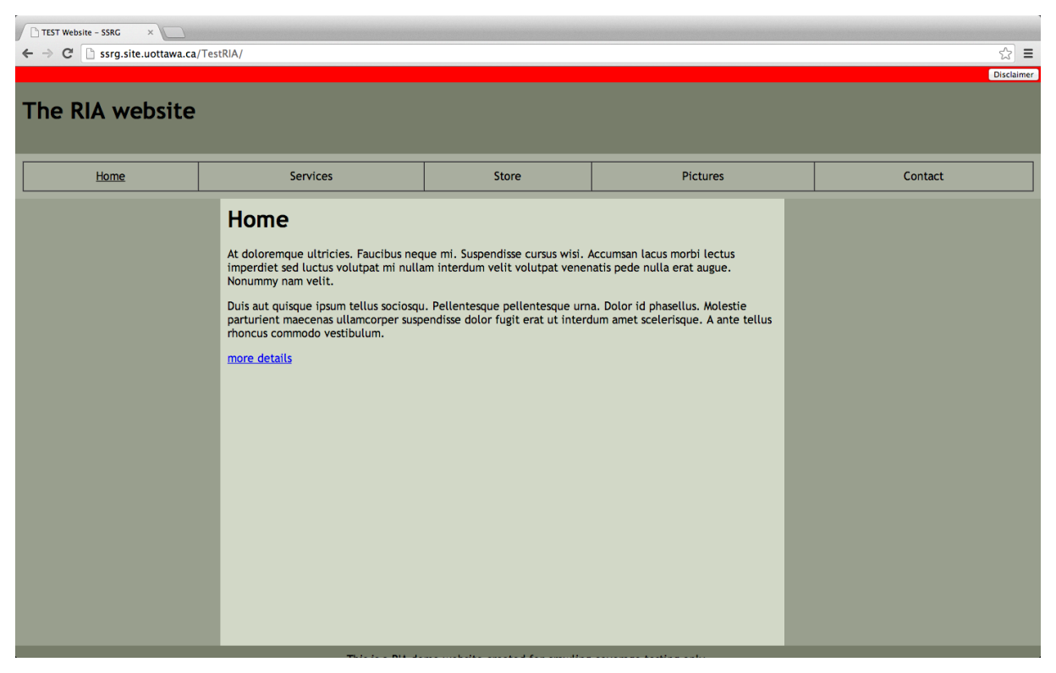


Figure 5.2: The termination virtual ring

Test Beds

- Four target web applications used to measure the performance of the distributed web crawler:


TEST Website - SSRG
ssrg.site.uottawa.ca/TestRIA/
Disclaimer

The RIA website

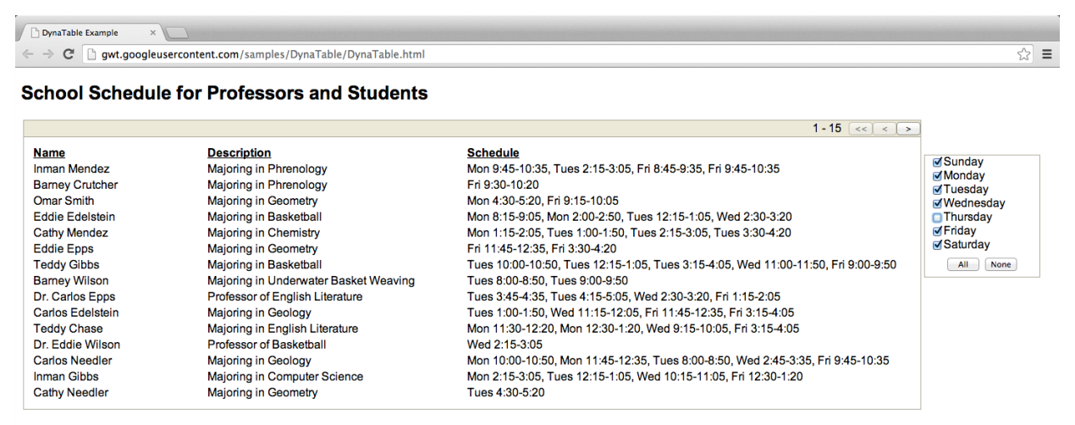
Home | Services | Store | Pictures | Contact

Home

At doloremque ultricies. Faucibus neque mi. Suspendisse cursus wisi. Accumsan lacus morbi lectus imperdiet sed luctus volutpat mi nullam interdum velit volutpat venenatis pede nulla erat augue. Nonummy nam velit.

Duis aut quisque ipsum tellus sociosqu. Pellentesque pellentesque urna. Dolor id phasellus. Molestie parturient maecenas ullamcorper suspendisse dolor fugit erat ut interdum amet scelerisque. A ante tellus rhoncus commodo vestibulum.

more details

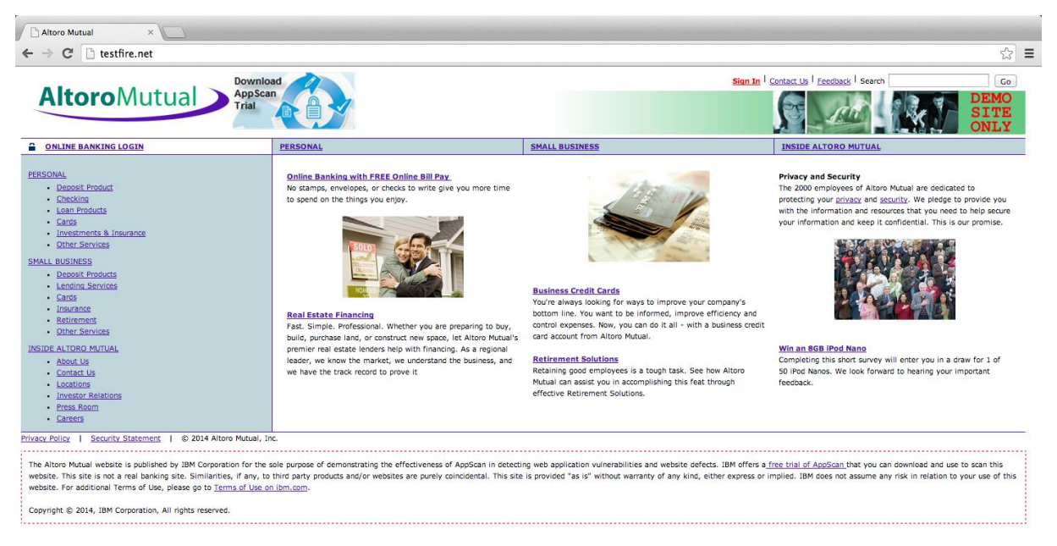

DynaTable Example
gwt.googleusercontent.com/samples/DynaTable/DynaTable.html

School Schedule for Professors and Students

1 - 15

Name	Description	Schedule
Inman Mendez	Majoring in Phrenology	Mon 9:45-10:35, Tues 2:15-3:05, Fri 8:45-9:35, Fri 9:45-10:35
Barney Crutcher	Majoring in Phrenology	Fri 9:30-10:20
Omar Smith	Majoring in Geometry	Mon 4:30-5:20, Fri 9:15-10:05
Eddie Edelstein	Majoring in Basketball	Mon 8:15-9:05, Mon 2:00-2:50, Tues 12:15-1:05, Wed 2:30-3:20
Cathy Mendez	Majoring in Chemistry	Mon 1:15-2:05, Tues 1:00-1:50, Tues 2:15-3:05, Tues 3:30-4:20
Eddie Epps	Majoring in Geometry	Fri 11:45-12:35, Fri 3:30-4:20
Teddy Gibbs	Majoring in Basketball	Tues 10:00-10:50, Tues 12:15-1:05, Tues 3:15-4:05, Wed 11:00-11:50, Fri 9:00-9:50
Barney Wilson	Majoring in Underwater Basket Weaving	Tues 8:00-8:50, Tues 9:00-9:50
Dr. Carlos Epps	Professor of English Literature	Tues 3:45-4:35, Tues 4:15-5:05, Wed 2:30-3:20, Fri 1:15-2:05
Carlos Edelstein	Majoring in Geology	Tues 1:00-1:50, Wed 11:15-12:05, Fri 11:45-12:35, Fri 3:15-4:05
Teddy Chase	Majoring in English Literature	Mon 11:30-12:20, Mon 12:30-1:20, Wed 9:15-10:05, Fri 3:15-4:05
Dr. Eddie Wilson	Professor of Basketball	Wed 2:15-3:05
Carlos Needler	Majoring in Geology	Mon 10:00-10:50, Mon 11:45-12:35, Tues 8:00-8:50, Wed 2:45-3:35, Fri 9:45-10:35
Inman Gibbs	Majoring in Computer Science	Mon 2:15-3:05, Tues 12:15-1:05, Wed 10:15-11:05, Fri 12:30-1:20
Cathy Needler	Majoring in Geometry	Tues 4:30-5:20

☑ Sunday
☑ Monday
☑ Tuesday
☑ Wednesday
□ Thursday
☑ Friday
☑ Saturday
All | None


Altoro Mutual
testfire.net

AltoroMutual
Download AppScan Trial
Sign In | Contact Us | Feedback | Search Go
DEMO SITE ONLY

ONLINE BANKING LOGIN | PERSONAL | SMALL BUSINESS | INSIDE ALTORO MUTUAL

PERSONAL: Deposit Product, Checking, Loan Products, Cards, Investments & Insurance, Other Services
SMALL BUSINESS: Deposit Products, Lending Services, Cards, Insurance, Retirement, Other Services
INSIDE ALTORO MUTUAL: About Us, Contact Us, Locations, Investor Relations, Press Room, Careers

Online Banking with FREE Online Bill Pay
No stamps, envelopes, or checks to write give you more time to spend on the things you enjoy.

Real Estate Financing
Fast. Simple. Professional. Whether you are preparing to buy, build, purchase land, or construct new space, let Altoro Mutual's premier real estate lenders help with financing. As a regional leader, we know the market, we understand the business, and we have the track record to prove it

Business Credit Cards
You're always looking for ways to improve your company's bottom line. You want to be informed, improve efficiency and control expenses. Now, you can do it all - with a business credit card account from Altoro Mutual.

Retirement Solutions
Retaining good employees is a tough task. See how Altoro Mutual can assist you in accomplishing this feat through effective Retirement Solutions.

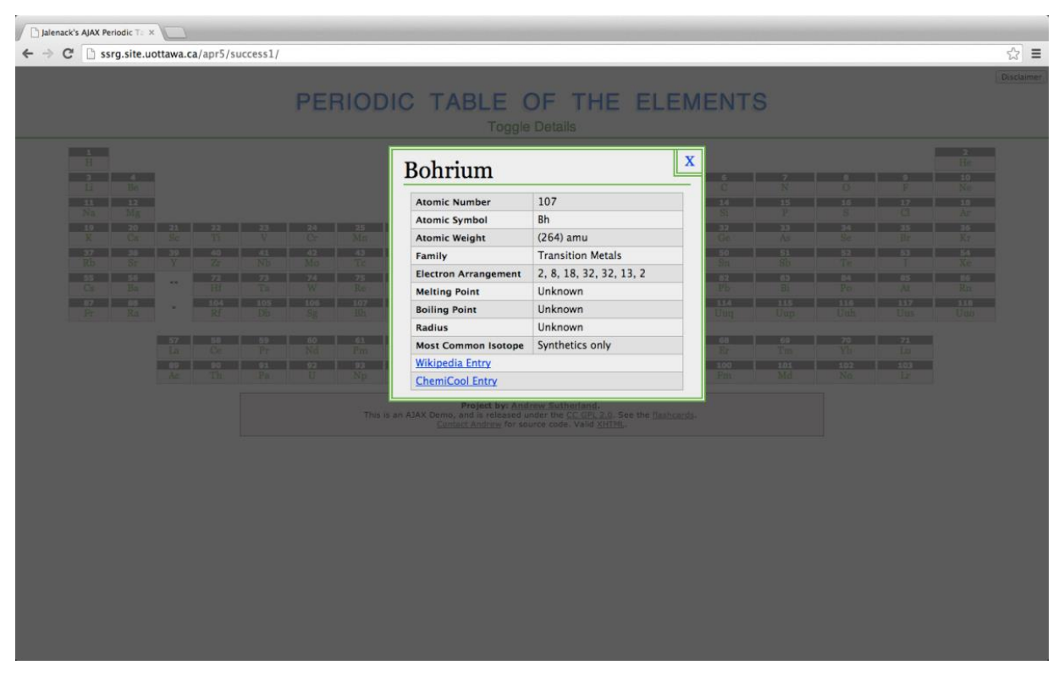
Privacy and Security
The 2000 employees of Altoro Mutual are dedicated to protecting your privacy and security. We pledge to provide you with the information and resources that you need to help secure your information and keep it confidential. This is our promise.

Win an 8GB iPod Nano
Completing this short survey will enter you in a draw for 1 of 50 iPod Nanos. We look forward to hearing your important feedback.

Privacy Policy | Security Statement | © 2014 Altoro Mutual, Inc.

The Altoro Mutual website is published by IBM Corporation for the sole purpose of demonstrating the effectiveness of AppScan in detecting web application vulnerabilities and website defects. IBM offers a free trial of AppScan that you can download and use to scan this website. This site is not a real banking site. Similarities, if any, to third party products and/or websites are purely coincidental. This site is provided "as is" without warranty of any kind, either express or implied. IBM does not assume any risk in relation to your use of this website. For additional Terms of Use, please go to Terms of Use on ibm.com.

Copyright © 2014, IBM Corporation, All rights reserved.


Jalenack's AJAX Periodic T
ssrg.site.uottawa.ca/apr5/success1/

PERIODIC TABLE OF THE ELEMENTS
Toggle Details

Bohrium

Atomic Number	107
Atomic Symbol	Bh
Atomic Weight	(264) amu
Family	Transition Metals
Electron Arrangement	2, 8, 18, 32, 32, 13, 2
Melting Point	Unknown
Boiling Point	Unknown
Radius	Unknown
Most Common Isotope	Synthetics only

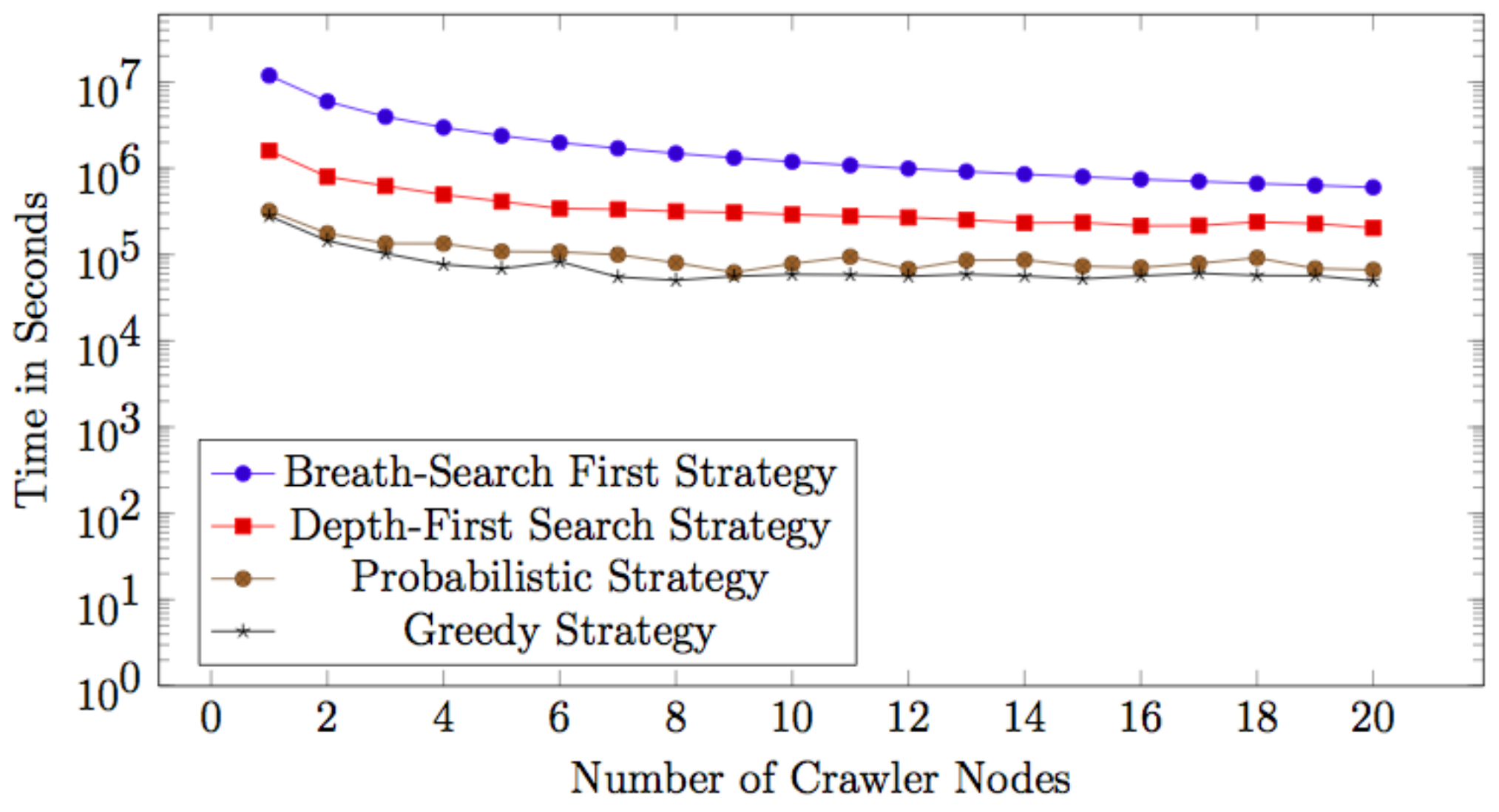
Wikipedia Entry
ChemiCool Entry

Algorithm

Algorithm 1 Crawling Algorithm (As Executed at Each Node)

```
GetCredentialsFromCoordinator()
nodeStatus ← Active
while (nodeStatus is not Terminated) do
    if workingStates is Empty then
        GetNewStatesFromCoordinator()
        if workingStates is Empty then
            nodeStatus ← Done
            SendNodeStatusToCoordinator()
        else
            nodeStatus ← Active
        end if
    else
        $stateToVisit$ ← PickState(workingStates)
        $eventToExecute$ ← PickUnExecutedEvent($stateToVisit$)
        ExecuteEvent($stateToVisit, eventToExecute$)
        if currentState is not in discoveredStates then
            push currentState to discoveredStates
            push Assign(currentState) to currentState.unExecutedEvents
            push currentState to workingStates
            SendNewStateToCoordinator( currentState )
        end if
        $stateToVisit$.RemoveUnExecutedEvent($eventToExecute$)
        if $stateToVisit$.unExecutedEvents is empty then
            workingStates.RemoveState($stateToVisit$)
        end if
    end if
end while
```

Experimental Results



Crawling Efficiency

- Crawling efficiency measures how early in the crawl new states are discovered:

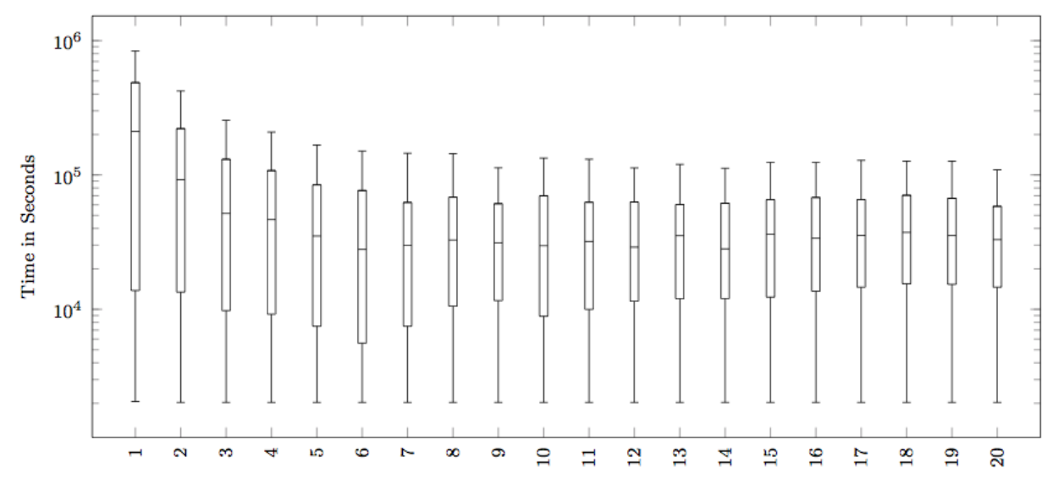


Figure 8.12: Dyna-Table: Cost of Discovering States using Depth-First Search Strategy

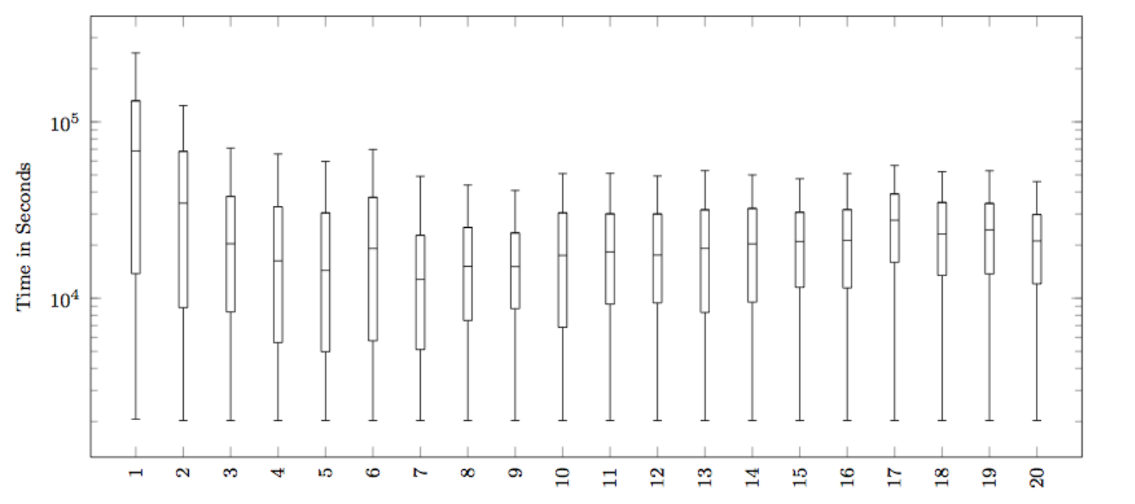


Figure 8.13: Dyna-Table: Cost of Discovering States using Greedy Strategy

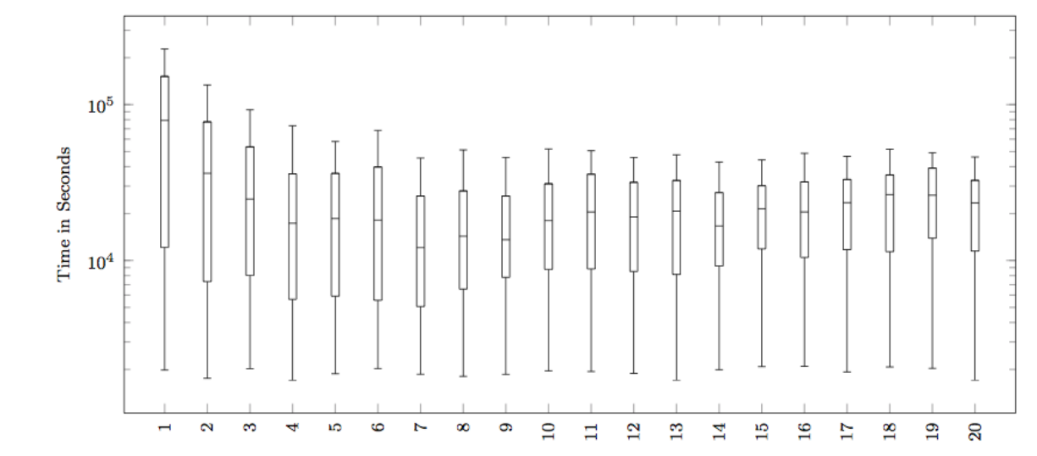
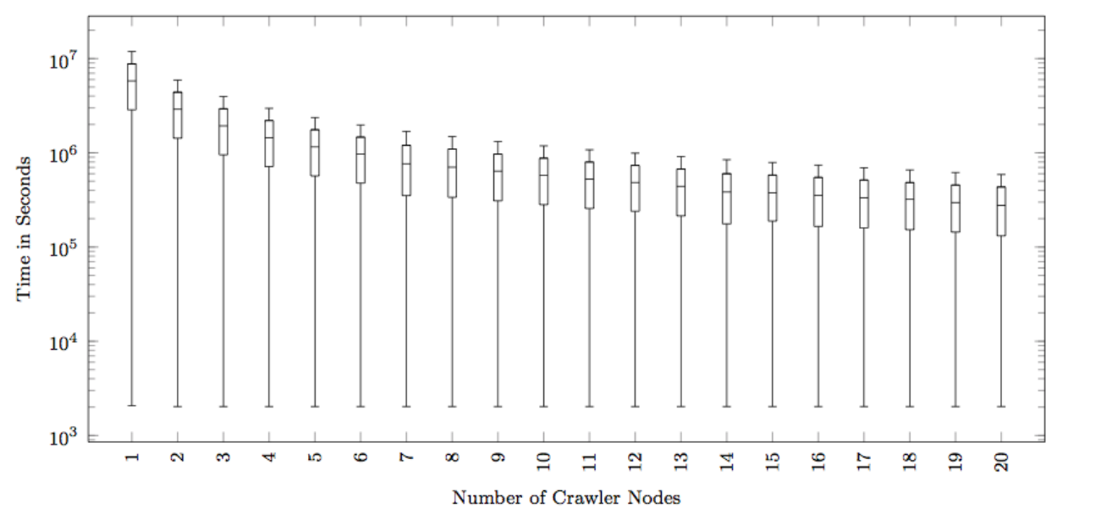


Figure 8.14: Dyna-Table: Cost of Discovering States using Probabilistic Strategy



Conclusion

- Distributed Greedy algorithm has the best performance in terms to total time it takes to crawl a website.
- Distributed Probabilistic model is the most efficient algorithm and discovers states early in the crawl.

Future Work

- We are currently working on distributed crawling of RIA in a cloud environment.
- We plan to add fault tolerance to our strategy so that if some of the nodes crash rest of the node continue without interruption.
- Once we have a working implementation of the system we plan to optimize it based on different infrastructure parameters such as cost of communication or the processing power available to different nodes.

Acknowledgments

This work is supported in part by Center for Advanced Studies, IBM Canada.

DISCLAIMER
The views expressed in this poster are the sole responsibility of the authors and do not necessarily reflect those of the Center for Advanced Studies of IBM.