Automated Synthetic Exploratory Monitoring of Dynamic Web Sites Using Selenium

Marcelo M. De Barros

marcelod@microsoft.com

Abstract

Web search engines are very dynamic in nature; not only are the backend and data powering the site evolving, but the frontend is always adapting to different browsers, devices and form-factors, and experiments are often running in production. In fact, when it comes to User Experience (UX), it is likely that users are always falling into some live experiment in production: variation of colors, fonts, typography, different Java Scripts and so on. Issues can occur on the live site for very particular contexts, where a context is defined as a particular configuration of browser, market and experiment. As an example, a JavaScript error can occur on a certain page, for certain types of queries, against a certain market on a particular browser. Creating a priori monitoring for all these different contexts is not feasible.

We at the Microsoft Bing Experiences Team developed a concept of synthetic exploratory monitoring that can focus on the important features on the sites and pages, and use invariants (conditions that should always hold true, or always hold false, for specific contexts) to detect potential anomalies in the current context. We make use of stochastic models to ensure maximum relevant coverage of contexts and devices. We use the power of the Selenium testing framework to drive end-to-end automation on browsers and devices, the notion of exploratory tests based on simple finite-state machines, and a set of heuristics and invariants (text-based and image-based) that can auto-detect problems on the live site in very particular contexts.

We implemented the idea explained in this paper to monitor large-scale web sites such as Bing Search Engine where alerts are generated automatically whenever the anomaly conditions are detected. The solution is easily expandable to other sites. We envision, as future work, moving this technology to the cloud that would allow easy customization of all parameters (browsers used, definition of the finite-state machine, heuristics and invariants).

This paper explains the fundamental principles to create a stochastic monitoring model and demonstrates how to apply the principles to large-scale web sites and services. We will utilize Bing Search Engine to illustrate the techniques explained here.

Biography

Marcelo De Barros is a Principal Software Test Manager at Bing, Microsoft. A 15-year veteran at Microsoft, Marcelo has worked on a variety of different services, from Xbox to Search Engines. He has given many talks across the industry, most recently speaking at the Swiss Testing Day 2014 about the evolution of tests in the context of agile systems and organizations [1]. He holds 6 U.S. Patents and 9 U.S. Patent Applications, a M.Sc. degree in Computer Science by the University of Washington and a B.S. degree in Computer Science by UFPE (Federal University of Pernambuco, Brazil). In his spare time Marcelo is an avid programmer, being ranked in the top 1% of all Project Euler (projecteuler.net) members. He also actively writes an algorithms blog (http://anothercasualcoder.blogspot.com/).

1 Scaling Systems to Devices, Browsers and Markets

In today's world whenever a new online system is launched it is usually available across several devices (devices that display web contents), browsers and markets instantaneously and simultaneously. This poses a significant development challenge since:

a) Different browsers, devices and markets have specific requirements and resources that may differ from each other, and there is no enforced global standard across them.
b) Support for CSS (Cascading Style Sheet) [2] and HTML5 compatibility and support [3] vary significantly from browser to browser.
c) The form-factor for the different devices varies significantly. Because of smaller screens, code might need to be optimized to show the user different information or presentation of the information.
d) Markets are also another important dimension given the differences in language grammars as well as geo-cultural differences. Large-scale systems such as Google, Bing and Facebook are always dealing with such challenges.

Figure 1: different devices on which Bing is enabled

Many large-scale web sites are now making use of "flights" or "experiments" [13]. An experiment is a way to expose a percentage of the site's users to a different treatment of the site (which can be differences in the UI, middle-tier, backend or even data differences) in order to collect early feedback and then make an informed decision about the upcoming features for the system. For example, a search engine might want to expose 2% of its users to a search results page (or SERP) that shows only eight "blue links" by default instead of ten blue links. The telemetry for that experiment is then collected and analyzed against the "control" (the ten blue links) and data analysts work on distilling the positive and negative aspects of the experiment. Experiments can overlap with each other. At any point in time there might be tens or even hundreds of experiments running in production environments.

2 Monitoring Complexities

Since the code is somewhat customized to different user experiences (experiments, browsers, devices and markets), there is a possibility of encountering specific issues on any of these and worst, on combination of these dimensions: a specific problem may only happen on an experiment, on a given browser, on a given device and for a particular market.

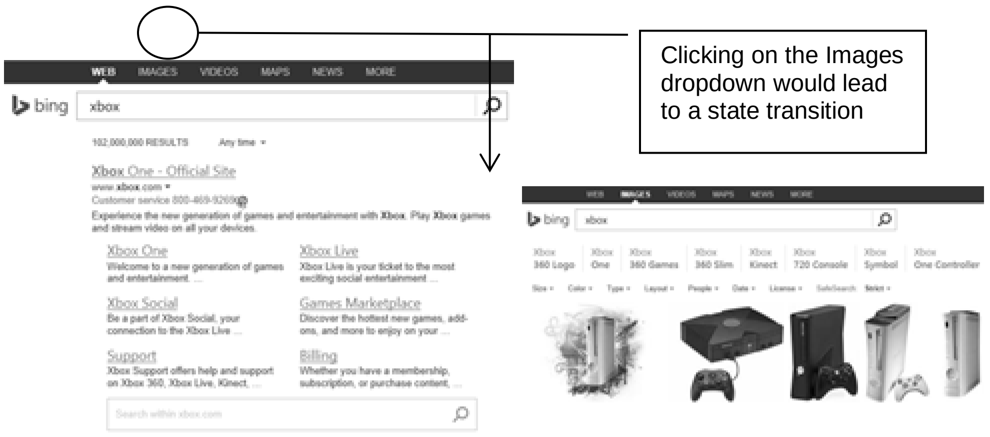
Some simple lower-bound calculations show the complexity and the scale of this problem. If we have around 30 experiments, 30 browsers, 30 devices and 200 markets, the number of possible combinations (assuming no overlaps on the experiments) becomes 30*30*30*200 = 5,400,000 different permutations. Even using well known monitoring techniques such as Gomez [4] or Keynote [5], it becomes impossible to monitor all these variations. In reality, though, most of these contexts are either not significantly crucial to the business or are not valid at all (for example, most of the time experiments are limited to either a group of markets or a group of browsers), hence understanding the valid and important permutations can prune the combinatorial space considerably.

3 Using Markov Chains

We use Markov Chains [6] to model the behavior of the system, limiting the monitoring space to the most probable paths. A Markov Chain is a type of stochastic model based on the concept of Finite-State Machines [14]) that undergoes transitions from one state to another on a state space. It is a random process usually characterized as memory-less: the next state depends only on the current state and not on the sequence of events that preceded it.

For search engines, the states in a Markov Chain are web site landing pages, such as: the search engine Home Page, the search engine Web Results Page, Videos Results Page, Images Results Page, Settings Page, and any the other page type included in the search engine substrate.

The actions that lead to a state transition are the different actions that can be performed by the end user, mainly Searches, Clicks, Tabs, Hovers, and so on. With enough anonymous log-based information about the different states and actions, one can build a comprehensive Markov Chain diagram modelling the proper behavior of the average user of the web system in questions. The assumption is that most web sites nowadays log information about their users' iterations with the page (in an anonymized manner). The picture below (Figure 2) gives an example of a state transition, and the table below (Table 1) gives an example of a simple Markov Chain. Notice that the key aspect here is that each action is associated with a certain probability (the "Probability Weightings" column in table 1), calculated based on the number (percentage) of users who triggered that respective action based on captured data.



Xbox One - Official Site
www.xbox.com
Customer service 800-469-9269
Experience the new generation of games and entertainment with Xbox. Play Xbox games and stream video on all your devices.

Xbox One
Welcome to a new generation of games and entertainment. ...

Xbox Live
Xbox Live is your ticket to the most exciting social entertainment ...

Xbox Social
Be a part of Xbox Social, your connection to the Xbox Live ...

Games Marketplace
Discover the hottest new games, add-ons, and more to enjoy on your ...

Support
Xbox Support offers help and support on Xbox 360, Xbox Live, Kinect, ...

Billing
Whether you have a membership, subscription, or purchase content, ...

Figure 2: Example of state transitions

Initial State	Final State	Action	Probability Weightings
Home Page	Search Results Page	Typed Query	20%
Home Page	Search Results Page	Clicked on “Top News”	15%
Search Results Page	Search Results Page	Typed Query	60%
Search Results Page	Non-Search Page	Clicked on Ads	15%
Search Results Page	Non-Search Page	Clicked on Algo Result	33%

Table 1: example of Markov Chains states and weighted transitions

The granularity of the states and the actions is something that varies depending on the applications. In the example above, the Typed Query action could certainly be further refined by specifying the category/class of query being typed, such as “Local Query”, “Adult Query” or “Electronics”. Likewise the “clicked on” event can be grouped into categories (such as “clicked on Algo Results”) or further refined down to the domain of the link being clicked (such as “clicked on an amazon.com link”). The important aspect is to create the chain in such a way that it truly encompasses the users’ behavior but keeping it concise enough to prune the overall search space. For our project, we also added some random aspects to our testing in order to provide extra coverage. For example, when the action is “Send a new query” we take a random query from a pool of pre-defined queries, usually a combination of head and tail queries (See “Web Search Queries” [9]).

Some states are outside the scope of the pages being tested. For example, if the scope being tested is all the pages under the bing.com domain, any site outside that domain would be considered an out-of-scope state. It is important to model the chain in such a way that once out of the scope, actions will lead to in-scope states (such as clicking the back button, or navigating back to the initial state).

With the Markov Chain created, the monitoring approach can be tweaked to randomly follow the paths and probabilities specified by the chain. Notice that the approach will necessarily focus on the most probable paths (assuming a random distribution), which is the desired approach.

In addition to using a Markov Chain for transitions, another important aspect that needs to be taken into consideration is the overall distribution of browsers, devices, markets and flights (experiments).

There are two different approaches to integrating Markov Chains for these additional dimensions into the monitoring system:

1) Create the Markov Chain to take into account browsers, devices, markets and flights (experiments). In such cases there can be multiple Markov Chains for each dimension, or combination of dimensions, or one Chain where states and transitions take into account these dimensions; or, alternatively
2) Create the Markov Chain without the particular data about browsers, devices, markets and flights, and use an orthogonal table with the distribution of the population across these dimensions, and randomly switch to a certain dimension as you navigate the chain.

The approach we have taken is the second one. The Markov Chain is created with the overall usage pattern across all the users in the system. At the same time we get the distribution of users across all browsers, devices, markets and experiments. In the following hypothetical example (Table 2), we see several user context distributions across browsers, devices, markets and experiments. We then combine these two sources of data (the Markov Chain and the User Context Distributions) in order to come up with the proper stochastic model for the exploratory tests. Section 5 explains the details of how these two data sources come together.

Browsers Distribution

Browser	**Percentage of users**
IE7	6%
IE8	8%
IE11	15%
Firefox	9%
Others	*62%*

Devices Distribution

Device	**Percentage of users**
Windows Phone	34%
iPhone	17%
Kindle Fire	17%
Android	9%
Others	*23%*

Markets Distribution

Market	**Percentage of users**
United States	52%
China	17%
Brazil	4.5%
Canada	7%
Others	*19.5%*

Experiments Distribution

Experiment	**Percentage of users**
Experiment #1: light-blue background color	2%
Experiment #2: larger font size for titles	3%
Experiment #3: larger images	20%
Experiment #4: new relevance ranker	1%

Table 2: example of user context distributions

4 Selenium

Selenium [15] is a portable software testing framework for web applications that provides a record/playback tool for authoring tests without learning a test scripting language (Selenium IDE). It also provides a test domain-specific language (Selenese) to write tests in a number of popular programming languages, including Java, C#, Groovy, Perl, PHP, Python and Ruby. The tests can then be run against most modern web browsers. Selenium deploys on Windows, Linux, and Macintosh platforms. The way we use Selenium for exploratory tests and monitoring is thru Selenium WebDrivers. Selenium WebDriver accepts commands and sends them to a browser. This is implemented through a browser-specific browser driver, which sends commands to a browser, and retrieves results. Most browser drivers actually launch and access a browser application (such as Firefox or Internet Explorer). Selenium WebDriver does not need a special server to execute tests. Instead, the WebDriver directly starts a browser instance and controls it. There is an ongoing effort by the inventors of Selenium to make it an internet standard [7].

Selenium provides an easy interface to interact with the browser, and the same test scripts can be used against many supported browsers. The ability to perform clicks, hovers, navigation manipulation, simulate different keyboard commands to the browser, scroll, change the browser settings and even detect and manipulate pop-up windows make it ideal for web automation.

In order to provide extra reliability, one can make use of a Selenium Grid. Selenium Grid is a server that allows tests to use web browser instances running on remote machines. With Selenium Grid, one server acts as the hub. Tests contact the hub to obtain access to browser instances. The hub has a list of servers that provides access to browser instances (WebDriver nodes), and lets tests use these instances. Selenium Grid allows running tests in parallel on multiple machines, and to manage different browser versions and browser configurations centrally (instead of in each individual test).

5 Exploratory Runs

The term Exploratory Runs here is loosely used to define the process of semi-randomly exploring different parts of a system while performing different verifications and validations that are pertinent to the current part of the system in question. The semi-random nature is accomplished via two methods: walking the generated Markov Chain, and modifying the context based on the users' distribution of markets, browsers, devices and experiments. The process usually starts at the initial page of the system, such as the user's home page. At that point a frequency-weighted random set of actions gets triggered based on the weight (probability) of the actions in the Markov Chain. It continues from that point on following the same approach indefinitely or until a certain time amount elapses. The transition of the states is implemented via commands in Selenium. Figure 3 below illustrates a simple Markov Chain being walked probabilistically:

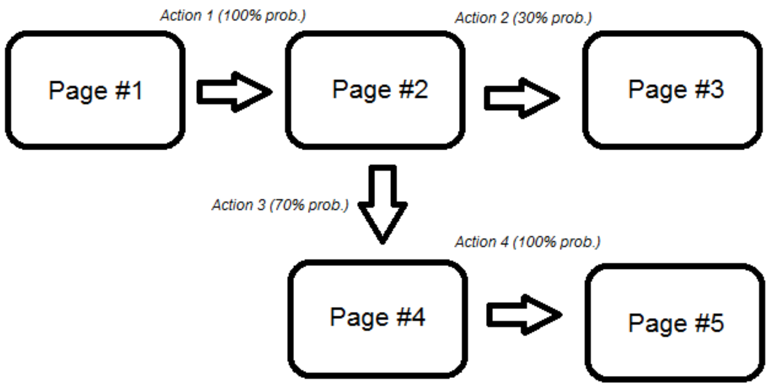


Figure 3: schema depicting a simple Markov Chain

Orthogonally to the walk of the Markov Chain, we make use of the contexts distribution in the following manner:

a) Markov Chain traversal keeps happening randomly for a period of time (say N minutes)
b) After that period of time elapses, a change of context happens based on the distribution table

We use N = 30 minutes, which is based on our observations with real Bing user data, 30 minutes is the average time for a user web session. After 30 minutes, the contexts in which the tool is running may change: browser, device, market or experiment. The change is random but weighted based on the distribution tables. We utilize a number of Selenium Grids, one for each type of IE browser (from version 7 to the latest version), and all the grids also contain other browsers, such as Chrome and Firefox. Markets also change based on a set of pre-defined markets (around 200 in our case). The device is simulated on the desktop browsers by manipulating the user-agent [8]. This simulation isn't ideal as some issues only appear or repro on the actual devices, but it is a good stopgap solution to catch some types of issues (like features being under/over triggered). We also force the exploratory run to fall into one (or combination of) experiments by using test hooks (in our case query-string parameters that are only enabled/visible inside the Microsoft corporate network). The automation keeps running indefinitely as a monitoring mechanism against production.

6 Subscription-Based Validation Modules

It is common to see the schema of a validation module (or test case) as a self-contained unit that performs all the steps necessary to set up the proper pre-validation before the validation takes place, followed by the validation itself, culminating with the post-validation (or teardown).

Schematically we have:

```
SampleTestCase()
{
  Pre-ValidationSetup();
  Validation();
  Post-ValidationSetup(); //Teardown
}
```

There are many advantages of such scheme: simplicity, standard pattern, readability, reproducibility, determinism, to name a few. However, such a model does not fit well into the exploratory runs mentioned previously. Instead, what we want is a subscription-based model where the test case subscribes to the current state (or action) if the current state (or action) meets certain criteria pertinent to that test.

Schematically, subscription-based test cases have the following format:

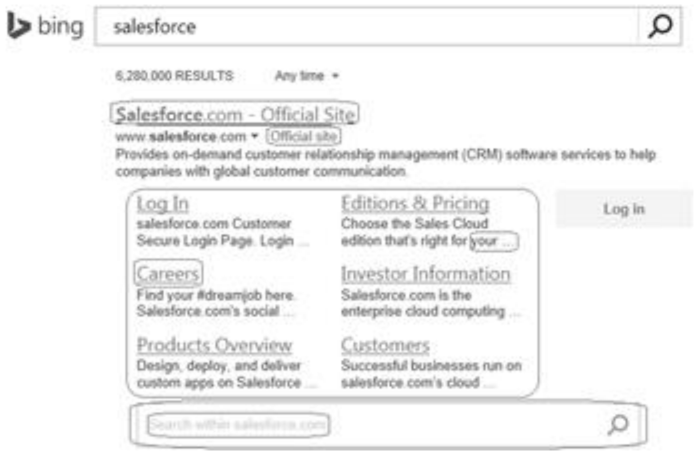
```
SubscriptionBasedSampleTestCase()
{
  If(IsRelevantState(this.CurrentState))
    Validation();
}
```

In essence we're proposing a separation of the validation method from the configuration. The test becomes opportunistic rather than deterministic: if we reach a situation during the traversal of the Markov Chain where the test is applicable, then it runs; otherwise it ignores the current state.

An example of a subscription-based test case would be the following: suppose that we want to write a test case to validate behaviors for a certain segment of queries called navigational queries, which are queries that seek a single website or web page of a single entity (See "Web Search Queries" [9]). A query such as "sales force" is a navigational query. There are several types of validation that can be performed for navigational queries. As per the example in Figure 4, when searching on "sales force", we can base validation on:

a) Correctness of the algorithmic first result returned
b) Proper attribution for "Official Site"
c) Proper number, format, truncation for deep-links [10]
d) Proper placement and usage of inner-search boxes

The picture below (Figure 4) depicts the items that can be subjected to validation:



bing salesforce

6,280,000 RESULTS Any time

Salesforce.com - Official Site
www.salesforce.com Official site
Provides on-demand customer relationship management (CRM) software services to help companies with global customer communication.

Log In
salesforce.com Customer Secure Login Page. Login ...

Editions & Pricing
Choose the Sales Cloud edition that's right for your ...

Careers
Find your #dreamjob here. Salesforce.com's social ...

Investor Information
Salesforce.com is the enterprise cloud computing ...

Products Overview
Design, deploy, and deliver custom apps on Salesforce ...

Customers
Successful businesses run on salesforce.com's cloud ...

Log in

Search within salesforce.com

Figure 4: Validation aspects for web-search deep-links

There are two types of tests that can be used in this model:

1) <u>Custom Tests</u> are specific for only certain states (or actions). For instance, the deep-links validation shown above is an example of custom test since it only applies to pages originated from navigational queries
2) <u>Invariant Tests</u> verify general invariants that should always be true (or always be false) no matter what state we are

Invariant Tests are very powerful since they apply to all states (or actions). It is important and recommended that the product being tested be properly instrumented with test hooks in order to enable invariant conditions that can then be tested thru invariant tests. An example of an invariant test would be a test that looks for java script errors. No state (or action) should lead to a java script error on the page. We instrumented some of the Bing pages so that whenever inside the Microsoft corporate network and when a certain query string parameter is passed in the URL, any java script error is caught via a global try/catch and written into a hidden HTML div tag [11]. With such instrumentation implemented, the invariant test for java script errors becomes trivial – basically checking for the presence of the java script error div tag. Other types of invariant tests are:

a) Links: no links should lead to 404 pages
b) Server Error: no state/action should lead to server errors
c) Security: no state/action should expose any security flaw
d) Overlapping: no state/action should contain overlapped elements

Selenium also provides a capability of taking the screenshot of the current page. This allows the engineers to implement image-based test methods, some of which can be custom methods (such as the rendering and placement of some objects on the page specific to certain contexts) or invariants (such as the space between blocks on the page). Also, it is important to notice that some of the methods only apply to certain contexts (browsers, devices, markets or experiments). In such cases the test needs to verify that the current context is relevant for the test in question to be executed.

Methodology

Combining all the approaches described in this paper, we come up with the following methodology for synthetic exploratory testing or monitoring of large-scale web systems:

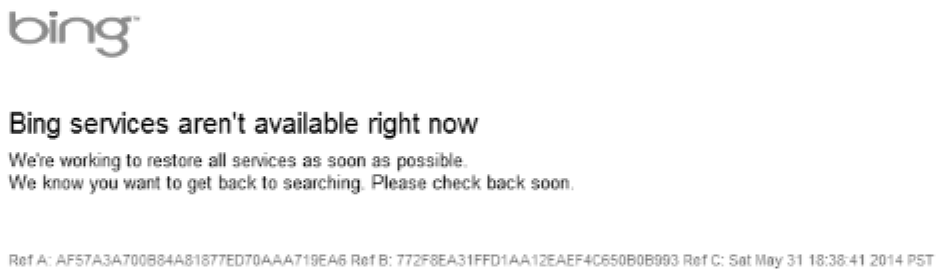
1) Mine the logs to create the user's profile Markov Chain
2) Retrieve the percentage distribution of different contexts (browsers, devices, markets and experiments)
3) Create custom and invariant tests that adhere to the subscription-based model

4) Stochastically run through the Markov Chain using Selenium or Selenium Grid. When testing search engines a key aspect is the generation of relevant queries to be used. It can be a combination of top queries based on frequency as well as segment-specific queries (such as queries that trigger local results or movie results)
5) Sporadically (time-based) switch contexts based on the distribution from #2
6) At each state (and action), apply the subscription-based tests from the library (#3). Alert in case of failures.

We differentiate monitoring from testing in terms of running the tests post-production and pre-production, respectively. The approach can be used for either one. However, we prefer to have deterministic tests as a pre-production mechanism, leaving the non-deterministic ones (such as the stochastic ones based on Markov Chains) as a monitoring mechanism (post-production). Also, the different tests have different priorities, so not all the tests will lead to an escalation (usually the invariant ones are deemed higher priority than the custom ones).

As the approach above executes, over time the critical monitoring paths will certainly be covered. Given that the approach follows a weighted-probability model, the critical paths will be covered more often than the non-critical ones. That is desirable since in today's fast-pace development environment of large-scale web systems, only the critical problems (the ones affecting the vast majority of users) get real attention, others are treated as low priority. The stochastic model is an elegant way to ensure highly-probable coverage of critical scenarios, and yet also cover some low-key scenarios.

Below are two examples of invariant failures when the model was applied to Bing.com. We used a set of 5 high-end servers executing around 1,000,000 state transitions per day, and running over 100 validation methods (of which 15% were invariant ones). The first example (figure 5) is an invariant that looks for HTTP 500 server errors, in this case generated by a combination of experiment and different interactions with the site:


bing

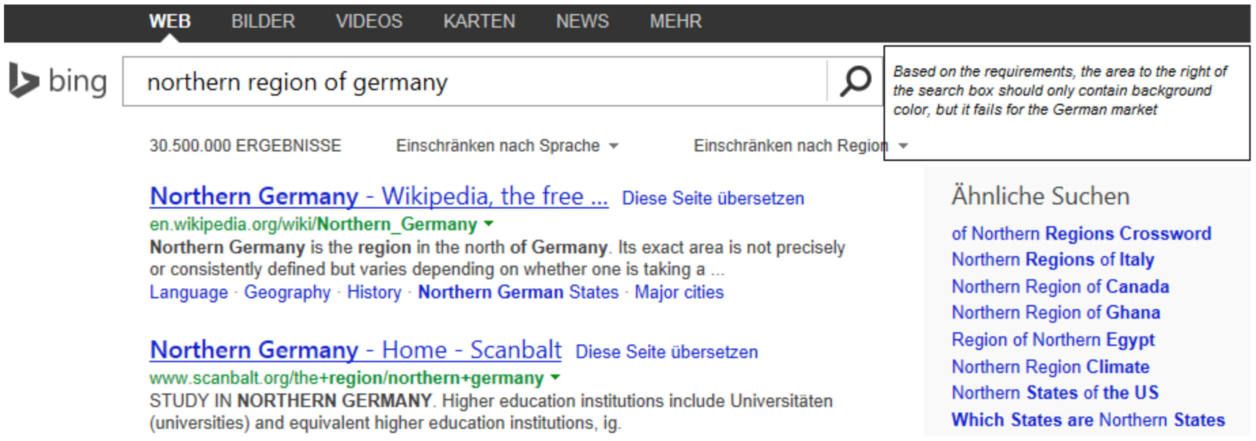
Bing services aren't available right now

We're working to restore all services as soon as possible.
We know you want to get back to searching. Please check back soon.

Ref A: AF57A3A700B84A81877ED70AAA719EA6 Ref B: 772F8EA31FFD1AA12EAEF4C650B0B993 Ref C: Sat May 31 18:38:41 2014 PST

Figure 5: Issue discovered thru an invariant test method

The second one is a low-priority invariant test based on image processing. In this case the area to the right of the end of the search box should always contain background color only. But in the case of the German market, whenever search filters are present due to the long words in German, the placement of the filters are going beyond the limits of the search box, breaking the pre-specified requirement. Figure 6 shows an example of such an issue:


WEB BILDER VIDEOS KARTEN NEWS MEHR

bing northern region of germany

Based on the requirements, the area to the right of the search box should only contain background color, but it fails for the German market

30.500.000 ERGEBNISSE Einschränken nach Sprache Einschränken nach Region

Northern Germany - Wikipedia, the free ... Diese Seite übersetzen
en.wikipedia.org/wiki/Northern_Germany
Northern Germany is the **region** in the north **of Germany**. Its exact area is not precisely or consistently defined but varies depending on whether one is taking a ...
Language · Geography · History · Northern German States · Major cities

Northern Germany - Home - Scanbalt Diese Seite übersetzen
www.scanbalt.org/the+region/northern+germany
STUDY IN **NORTHERN GERMANY**. Higher education institutions include Universitäten (universities) and equivalent higher education institutions, ig.

Ähnliche Suchen
of Northern Regions Crossword
Northern Regions of Italy
Northern Region of Canada
Northern Region of Ghana
Region of Northern Egypt
Northern Region Climate
Northern States of the US
Which States are Northern States

Figure 6: Example of an image-based error related to markets

Notice that the use of Markov Chains and context distributions allows the monitoring system to be highly adaptive: as the user patterns and context distributions change over time, the system will adapt itself based on the new data. The other important aspect is that the validation and monitoring mechanisms can certainly be extended to more than functional use, such as covering security concerns. At each step during the traversal of the Chain, we can also plug-in penetration tests [12] which would be characterized as invariant methods.

Summary

Monitoring large-scale dynamic web sites across multiple browsers, devices, markets and experiments is a very complex task. In this paper we have proposed a way to model the users' behavior via stochastic methods such as Markov Chains and use Selenium to recreate the same conditions experienced by real users in production. In addition, the validation approach is also changed from self-contained validation methods to a subscription-based model where the validation method subscribes to only the applicable states. Finally, validations can be invariant ones (applicable to all states) or custom ones (applicable to specific states).

References

[1] M. De Barros, Chap Alex, 2014. *Agile quality-centric development process of large-scale web systems*, Swiss Testing Day

[2] Comparison of layout engines - CSS, http://en.wikipedia.org/wiki/Comparison_of_layout_engines_(CSS) (accessed July 16, 2014)

[3] Comparison of layout engines – HTML5, http://en.wikipedia.org/wiki/Comparison_of_layout_engines_(HTML5) (accessed July 16, 2014)

[4] Gomez Network, https://www.gomeznetworks.com/?g=1 (accessed July 16, 2014)

[5] Keynote, http://www.keynote.com/ (accessed July 16, 2014)

[6] M. De Barros et al, 2007. *Web Services Wind Tunnel: On Performance Testing Large-Scale Stateful Web Services*, 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07)

[7] Selenium Software, http://en.wikipedia.org/wiki/Selenium_(software) (accessed July 16, 2014)

[8] Browser User Agents, http://en.wikipedia.org/wiki/User_agent (accessed July 16, 2014)

[9] Web Search Queries, http://en.wikipedia.org/wiki/Web_search_query (accessed July 16, 2014)

[10] Deep linking, http://en.wikipedia.org/wiki/Deep_linking (accessed July 16, 2014)

[11] The HTML <div> tag, http://www.w3schools.com/tags/tag_div.asp (accessed July 16, 2014)

[12] Penetration Tests, http://en.wikipedia.org/wiki/Penetration_test (accessed July 16, 2014)

[13] A/B Testing, http://en.wikipedia.org/wiki/A/B_testing (accessed July 16, 2014)

[14] Finite-State Machine, http://en.wikipedia.org/wiki/Finite-state_machine (accessed July 16, 2014)

[15] Selenium HQ Browser Automation, http://docs.seleniumhq.org/ (accessed July 16, 2014)