AUTOMATIC NAVIGATION ON PRIVATE WEBSITES

Radek Kliment
Master Degree Programme (2), FIT BUT
E-mail: xklime03@stud.fit.vutbr.cz

Supervised by: Zbyněk Křivka
E-mail: krivka@fit.vutbr.cz

Abstract: This paper deals with the navigation across web pages including the authentication to access their private sections and the user context management. It introduces the design of the mechanism for the automated navigation including new scripting language and tools for the visual description.

Keywords: navigation activity, web page, information extraction, automatic form filling, Jython

1 ÚVOD

Provádění různých aktivit na webových stránkách je pro mnoho lidí každodenní činností. Některé aktivity se stejným scénářem jsou v průběhu času prováděny opakovaně a mohou se skládat z posloupnosti kroků, z nichž se každý liší pouze několika parametry, například opakované odeslání požadovaných dat pomocí formuláře. V těchto případech může být užitečné tyto aktivity určitým způsobem automatizovat. Tato práce popisuje návrh mechanismu, který to umožňuje.

2 NAVIGACE NA WEBU

Aktivity na webových stránkách mohou být tvořeny zobrazením stránky na základě její URL adresy, využíváním odkazů pro přechod mezi jednotlivými stránkami nebo vyplňováním položek formulářů a jejich odesíláním. Velké množství stránek vyžaduje pro přístup do jejich určitých částí autentizaci (zadání uživatelského jména a hesla). Autentizaci je možné realizovat několika způsoby.

Jednou možností je HTTP autentizace [1], kdy se využije prostředků HTTP protokolu. Uživatelské jméno a heslo se zasílá jako součást hlavičky požadavku. Další možností je vytvoření přihlašovacího formuláře. Kromě polí pro jméno a heslo může obsahovat i další prvky pro dodatečné informace. Pro přihlášení je také možné využít Java applet nebo Flash objekt vložený do stránky. Z důvodu naprosté volnosti a nestandardnosti řešení však nebudou tyto případy brány dále v úvahu.

Vzhledem k použití bezstavového HTTP protokolu je po přihlášení nutné udržovat kontext uživatele. V případě HTTP autentizace si prohlížeč zadané údaje uloží a zasílá je spolu s každým požadavkem. Při využití přihlašovacího formuláře existuje více možností [2]. Ve všech případech je na serveru vygenerován unikátní identifikátor a dojde k jeho asociaci s uživatelem. Jednotlivé možnosti se pak liší způsobem předávání vygenerovaného identifikátoru. Lze využít tzv. *cookies*. Jedná se o malé množství dat obsahujících identifikátor, která jsou zaslána klientovi po prvním požadavku nebo např. po úspěšném přihlášení. Klient si je uloží a zasílá spolu s každým dalším požadavkem. Dále je možné využít skrytého formulářového pole, jehož hodnota je tvořena identifikátorem uživatele. Identifikátor může také být připojen přímo jako parametr k URL adrese.

3 EXISTUJÍCÍ NÁSTROJE

V současné době existuje několik nástrojů, které sice slouží hlavně pro extrakci dat ze stránek, ale umožňují i provedením několika navigačních kroků vedoucích na požadovanou stránku. Jmenovat

lze komerční nástroj Lixto [3]. Vytvoření navigačních kroků se provádí pomocí prvku zobrazujícího stránku. Uživatel může procházet web, jako by používal běžný prohlížeč, a jednotlivé kroky jsou zaznamenávány. Na cílové stránce může pomocí myši označit elementy s požadovanými informacemi. Nástroj ale není určen pro opakované odesílání dat, např. s využitím formulářů.

4 NÁVRH MECHANISMU PRO NAVIGACI

Mechanismus je navrhován tak, aby podporoval načtení stránky dle URL adresy, přechod mezi stránkami pomocí odkazů, práci s formuláři, přihlašování do privátních částí stránek a zpracování odpovědí pro kontrolu správnosti provedení kroku. Ze stránky je také možné extrahovat data a uložit je jako XML soubor. Pro označení dat je možné použít regulární výraz nebo XPath výraz. Mechanismus také umožňuje dávkové zpracování dat získaných ze stránky nebo XML či CSV souboru a jejich využití pro další navigaci, například opakované odeslání formuláře lišící se pouze zadávanými informacemi.

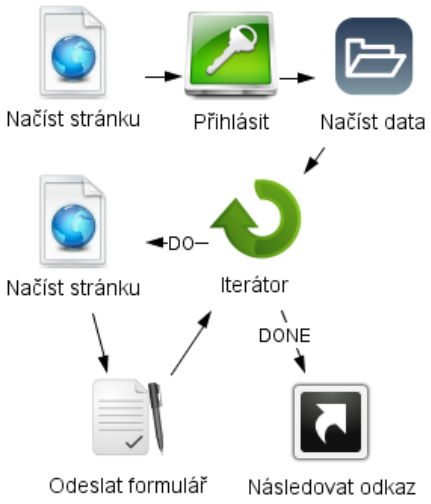
4.1 SKRIPTOVACÍ JAZYK PRO POPIS NAVIGACE

Aktivity prováděné na webových stránkách lze popsat pomocí navrženého skriptovacího jazyka. Poskytuje běžné řídicí konstrukce, práci s proměnnými a definování uživatelských funkcí. Pro provádění navigačních aktivit jsou dostupné vestavěné funkce, například pro načtení stránky, simulaci kliknutí na odkaz, vyplnění a odeslání formuláře a další. Jazyk je založen na projektu Jython, což je interpret jazyka Python implementovaný pro jazyk Java. Příklad popisu jednoduché aktivity je na obr. 1, v jehož pravé části je navíc doplněn popis provedený pomocí interaktivní vizuální reprezentace navigační aktivity, která je přiblížena v části 4.2.

```
loadPage("www.osoby.cz")
login("jmeno","heslo")
xml = loadXml("osoby.xml")
osoby = queryAllByXPath(xml,"//osoba")

for osoba in osoby:
  loadPage("www.osoby.cz/pridat.php")
  selectFormByPosition(1)
  jmeno = queryOneByXPath(osoba,"@jmeno")
  setInputByName("jmeno", jmeno)
  submitFormByBtnName("odeslat")

clickLinkByText("Odhlásit")
```



Obrázek 1: Příklad popisu navigace pomocí skriptovacího jazyka i vizuální reprezentace

Navigační aktivita na obr. 1 začíná načtením stránky na základě URL adresy, pokračuje přihlášením, kdy je automaticky nalezen formulář obsahující pole pro jméno a heslo. Následuje načtení externího XML souboru s daty a vybrání uzlů, jejichž data se použijí pro navigaci. Je iterováno přes všechny vybrané uzly. V každé iteraci se provede načtení stránky s formulářem, nastavení použití formuláře dle jeho pozice na stránce, vybrání dat z uzlu v aktuální iteraci, jejich vyplnění do pole ve formuláři a odeslání formuláře. Na konci je provedeno odhlášení kliknutím na odkaz obsahující text „Odhlásit“.

4.2 INTERAKTIVNÍ POPIS NAVIGACE

Navigační aktivity lze také popsat interaktivně s využitím vizuální reprezentace. Popis je tvořen grafem s uzly a hranami. Uzly reprezentují provedení jednoho kroku aktivity. Lze je přidávat z palety

dostupných uzlů, konfigurovat je v závislosti na jejich typu a propojovat je hranami, které reprezentují následnost prováděných kroků. Příklad je rovněž na obr. 1. Pro spuštění takto popsané navigační aktivity je nejprve vygenerován odpovídající zápis pomocí skriptovacího jazyka a ten je následně interpretován. Vygenerovaný zápis je možné ručně upravit. Provedené změny však nelze reflektovat zpět do vizuální reprezentace.

4.3 USNADNĚNÍ VYTVÁŘENÍ POPISU NAVIGACE

Vytváření popisu navigace, ať už s využitím skriptovacího jazyka, či pomocí vizuální reprezentace, lze usnadnit několika způsoby. Ve všech případech je ale nutné mít k dispozici stránku, která bude načtena před provedením aktuálně popisovaného kroku. Lze ji získat tak, že se navigační aktivita provede v ladicím režimu po předcházející krok.

Je možné zobrazit zdrojový kód načtené stránky. Je to užitečné v případě výběru nebo kontroly informací na stránce pomocí regulárního výrazu. Po jeho vytvoření ho lze pro kontrolu správnosti aplikovat na načtenou stránku a zvýraznit nalezené výskyty.

Dále lze zobrazovat aktuálně načtenou stránku pro kontrolu, zda krok proběhl dle očekávání, nebo například pro zvýraznění elementů vybraných pomocí zadaného XPath výrazu. Zobrazení stránky lze využít také pro vygenerování XPath výrazu vybraného elementu. Pro lepší orientaci ve struktuře stránky je rovněž možné zobrazit strom jejích elementů.

Při vytváření skriptu nebo při konfiguraci uzlu v případě vizuální reprezentace mohou být napovídány různé možnosti na základě aktuálně načtené stránky. Při výběru formuláře na základě jeho jména lze například napovědět jména všech formulářů vyskytujících se na stránce. V případě vytváření XPath výrazu se jedná o možnosti, jak na základě obsahu aktuální stránky v jeho tvorbě pokračovat.

5 ZÁVĚR

V tomto příspěvku byl představen návrh mechanismu pro automatizovanou navigaci na webových stránkách. V další fázi následuje implementace navrženého mechanismu jako součást desktopové aplikace založené na platformě Netbeans. S implementací souvisí také testování mechanismu na různých webových stránkách a zjišťování, které z uvedených způsobů usnadnění popisu navigace mají pro uživatele největší přínos. Při implementaci jim bude věnována zásadní pozornost.

PODĚKOVÁNÍ

Tato práce byla podpořena Evropským fondem regionálního rozvoje (ERDF) v rámci projektu Centra excelence IT4Innovations (CZ.1.05/1.1.00/02.0070), výzkumným záměrem MSM0021630528 a projektem specifického výzkumu FIT-S-11-2.

REFERENCE

[1] FRANKS, J., HALLAM-BAKER, P., HOSTETLER, J. et al. *HTTP Authentication: Basic and Digest Access Authentication, RFC 2617*. June 1999.

[2] *Session Tracking Methods* [online]. 31/05/2008 [cit. 2012-02-28]. Dostupné na: <http://javapapers.com/servlet/explain-the-methods-used-for-session-tracking>.

[3] BAUMGARTNER, R., FLESCA, S. a GOTTLOB, G. Visual Web Information Extraction with Lixto. In *Proceedings of the 27th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. S. 119–128. VLDB '01. ISBN 1-55860-804-4.