Posudek oponenta diplomové práce

Michal VYKYPĚL: Testovací platforma pro Webové roboty

Cílem této práce bylo prozkoumat techniky, které dovolují předikovat parametry Webového prostoru a následně navrhnout a implementovat nástroje pro simulaci topologie celého webu na jediném počítači.

Druhá kapitola práce se věnuje definicím pojmů z teorie grafů. Definice v ní uvedené by však mohly být rozhodně přesnější a především by měly pokrývat to, co se v práci dále používá.

- V definici 2.1.2 na str. 7 se definuje výstupní stupeň grafu jako počet jistých hran e, ale v popisu požadovaných vlastností už se pracuje s dvojicemi (u,v), nikoli s hranami e.

- Totéž se opakuje v definici 2.1.3 pro vstupní stupeň grafu.

Třetí kapitola, popisující modely náhodných grafů a jejich souvislosti s vlastnostmi webového grafu již zavedené symboly pro vstupní a výstupní stupně uzlů grafu nepoužívá a operuje s neorientovanými grafy a obecnými stupni uzlů $deg(u)$, který pro změnu v práci definován není. Není tedy jasné, zda se například v počítání bipartitních jader orientace hran zanedbávají, nebo se vyžadují hrany vedoucí oběma směry.

Kapitola 4, popisující implementaci, začíná svůj popis někde zprostředka. Sekce 4.1 zmiňuje jakési soubory atributů, aniž by bylo v té chvíli zřejmé, jaké atributy je potřebné (a zda vůbec) si v grafu webu pamatovat. Na straně 13 v téže sekci se čtenář jen tak mimochodem dozví, že souborů uzlů je více než jeden, aniž by chápal, proč by jich pro modelování grafu webu mělo být více. To, že může být více i souborů pro seznamy vstupních a výstupních hran se čtenář nedozví vůbec.

Předpokládám rovněž, že adresa atributu v souboru atributů u N-tého uzlu není *N*počet_atributů*, jak je uvedeno na str. 13, ale *N*počet_atributů+pořadové_číslo_atributu*. Možná offset také závisí na velikosti reprezentací jednotlivých atributů. Rozhodně nejsou všechny jednobajtové, ale vícebajtové a zřejmě také různých velikostí.

Příklady, které by měly ozřejmit konkrétní řešení použitá v implementaci jsou leckdy spíše zavádějící.

- Na straně 19 se hovoří o průchodu grafu do šířky s využitím prioritní fronty, řazené dle výstupního stupně uzlů, ale bezprostředně navazující příklad ve skriptovacím jazyce prochází graf do šířky s využitím fronty standardní. Více matoucí je informace v příkladu na str. 24, kde se uvádí, že se prioritní fronta uzlů, řazená podle atributu *pagerank*, deklaruje pomocí jedno-parametrického typu *queue(pagerank)*. Podle popisu syntaxe skriptovacího jazyka na str. 23 se však jedná o frontu celých čísel, reálných čísel nebo logických hodnot a jako hodnota parametru jsou proto povoleny jen hodnoty *asc* a *desc*. Prioritní fronta uzlů se dle popisované syntaxe musí deklarovat pomocí dvou-parametrického typu *queue(pagerank,asc)*.

- V popisu pole front na straně 25 se v příkladu uvádí, že po stažení stránky *a1* a po projití jeho následníků *a2*, *a3*, *c4*, *c3* budou ve frontách uloženy jen prvky *a3*, *c4* a *c3* s tím, že se začne zpracovávat *a3*. Proč se první následník (zde *a2*) do polí neukládá?

- V příkladu převodu čísla uzlu na fiktivní URL se uvádí „*Pokud budeme stránky pojmenovávat například **aaaa.html** až **zzzz.html**, ...*“, ale stránka s číslem 987 má podle textu v příkladu přiřazeno jméno ***00abcd.html***, což do uvedeného čtyřpísmenného značení nezapadá. Také není vůbec jasné, proč by se mělo uváděné číslo dokumentu 987 přemapovat v daném intervalu právě na hodnotu abcd. Pokud bych číslo vyjádřil v 26-kovém základu, jak napovídá popis značení, vyšlo by 987 = **0***26^3 + **1***26^2 + **11***26^1 + **25***26^0 a stránka by se měla tedy jmenovat spíše ***ablz.html***. V sekci 5.4 na straně 37 je pro změnu uveden příklad s URL ve tvaru ***000127.html***. O tom, jak se mapování čísla uzlu v doméně na jméno html souboru provádí, se čtenář nic nedozví ani z textu práce, ani z příkladu.

V popisu sestavení obsahu webové stránky na straně 30 se uvádí, že se generátor náhodných čísel pro výběr stránky inicializuje číslem stránky v dané doméně (zde číslem 987). To zřejmě není pravda. Pokud ano, znamenalo by to, že všechny stránky se stejným názvem ve všech doménách budou mít stejný vygenerovaný obsah.

Nemá být žádost na DNS překlad obvykle posílána nameserveru podle údajů v `/etc/resolv.conf` a nikoli v `/etc/hosts` (Simulace DNS na str. 31)?

Bývá rovněž dobrým zvykem odkázat se na veškerou uvedenou literaturu přímo z textu, aby bylo zřejmé, kde byl daný zdroj použit. Zde jsem nenalezl explicitní odkazy na AAR01, DC06, DD05, DD06, JT01 a WA00.

Nikde jsem v práci nenašel popsáno, na co jsou dobré externí knihovny boost a libmicrohttpd, přiložené na DVD ani jak se v simulátoru webu používají. Stejně tak jsem nenašel zmínku o tom, jak se pojmenovávají soubory, tvořící emulovaný web. Na DVD jsou soubory `outlinks001`, `inlinks001`, `001_01` a `001_02`. Předpokládám, že `outlinks001` je soubor následníků, zmiňovaný v sekci 4.1 na str. 12, soubor `inlinks001` je soubor předchůdců. Zbylé dva soubory jsou zřejmě (snad) soubory uzlů.

S překladem tří na DVD přítomných projektů *WebInterpret*, *WebReader* a *WebSimulator* jsem pomocí přiložených `Makefile` souborů uspěl pouze na distribuci Ubuntu. Stejný pokus na distribuci *Gentoo*, i když s přibližně stejnými verzemi *g++*, skončil neúspěšně. Neúspěšný pokus na platformě *Red Hat* nepočítám, tam byly k dispozici evidentně zastaralé nástroje *bison*, *flex*, i *g++*.

Celkově na mě práce, především její text, působí značně nedokončeným dojmem. S problémem samotným se autor nějak vypořádal a nabízí na jediném PC simulaci i velmi rozlehlého webu, splňujícího základní statistické parametry rozložení hran. Text práce však ponechává čtenáři k domyšlení až příliš mnoho informací.

Nejsem si proto jist, zda práce jako celek může být uznána za práci diplomovou a ponechávám toto rozhodnutí na komisi.

V Praze dne 12. 5. 2009

RNDr. Michal Kopecký, Ph.D.
KSI MFF UK