Automatické prohledávání webových stránek (I)

Internetová síť připomíná obrovskou knihovnu bez katalogů a pátrání po informacích se dnes neobejde bez automatického prohledávání webových stránek, které v mnoha směrech připomíná identifikaci neznámého systému. Webové vyhledávače, včetně populárního Google využívají speciální hledací roboty, kteří zkoumají všechny dostupné webové servery. Neustále bloudí po jednotlivých webových stránkách navigováni pouze nalezenými hypertextovými odkazy a sestavují gigantické informační indexy, podle nichž vyhledávače odhadují, kde se nachází žádaná informace. Z takto nashromážděných dat občas vracejí i předlouhé seznamy možných stránek, z nichž málokdo prozkoumá víc než několik čelních referencí, a proto je možné si položit otázky: „Jakým způsobem vyhledávače získávají informace? Proč uvádějí některé adresy dříve, i když jde o méně zajímavé údaje?“

Indexace stránek

Umístění odkazu závisí na řadě faktorů a často ho ovlivní, kromě hodnotných elementů, na něž se zaměříme především, i rozličné triky, jimiž si jedinci obeznámení s činnosti prohledávačů vylepšují svoje pozice. Na konci článku bude zmíněn jeden takový způsob kvůli ilustraci obtížnosti kvalitního indexování, neboť serioznější prohledávače už mají zakomponovanou detekci mnohých podvůdků a penalizují podobné stránky jejich zařazením na vzdálenější pozice.

Pojem *stránka* bude v dalších odstavcích označovat jakýkoli textem reprezentovatelný dokument získaný z webového serveru, tedy nejen dokument v nejrozšířenějším formátu (X)HTML, ale i dokumenty v dalších formátech určených k publikování, u nichž má indexování smysl, třeba PDF nebo soubory DOC napsané ve editoru Word apod.

Vyhledávače odpovídají na dotazy na základě informací uložených ve svých databázích. Jejich obsah se aktualizuje během *indexace stránek*, jak se nazývá proces analýzy stažených stránek, extrahování údajů pro vyhledávání a jejich následný zápis do databází. Do těch se kromě základních charakteristik často ukládají i hypertextové odkazy nalezené na dané stránce a textové bloky příhodné pro fulltextové vyhledávání. Stránky uložené v databázi jsou považovány za *zaindexované*.

Proces stahování stránek

Indexování předchází stahování stránek. Každý vyhledávací systém potřebuje prohledávajícího robota, který získává informace o stránkách. Ten má mnoho názvů, poměrně vžité je anglické pojmenování *crawler*. Mnohé vyhledávací systémy používají vlastní názvy, např. robot vyhledávače Google se jmenuje *Googlebot*, což je složenina za slov Google a robot. V tomto článku zůstaneme u pouhého slova robot. Robot vždy začíná od určité, explicitně zadané adresy URL, která označuje první webovou stránku (*obr. 1*). Robot vydá požadavek na její stažení. Informace o dané webové stránce se mu vrátí z webového serveru ve dvou hlavních částech: v hlavičce HTTP (*header*), která tvoří součást přenosového protokolu HTTP, a ve vlastní stránce.

Hlavička HTTP není normálně v prohlížečích vidět a dovedou ji zobrazit jedině speciální programy (lze je nalézt na síti Internet zadáním hesla *view HTTP header*). Přenáší se v ní základní údaje o webové stránce, jako je typ obsahu, čas poslední modifikace obsahu, čas odeslání stránky nebo doba její platnosti, a rovněž proměnné relace mezi serverem a prohlížečem, tzv. *cookies*.

Protože vlastní stránka může mít rozličné formáty, před zpracováním se převádí do unifikovaného interního tvaru pomocí vhodného filtru. Výsledná data stránky se analyzují a vybírají se z nich údaje relevantní pro další prohledávání a budoucí vyhledávání informací, tedy odkazy a texty vhodné pro indexování.

Jakmile robot zpracuje první webovou stránku určenou zadanou adresou, zvolí se na základě údajů, které z ní extrahoval, a informací uložených v databázi, další stránky vhodné pro zpracování. Tím se proces zacyklí. Požadavky na stažení vybraných stránek odešle do fronty vhodného rozhraní, které jich zpracovává zpravidla i několik najednou. Stahování stránek může trvat různě dlouhou dobu, která je ovlivněna nejen velikostí dat a momentální přenosovou rychlostí, ale i chybami. Rozhraní musí detekovat třeba přerušení přenosu a opakovat nesplněný požadavek či rozeznat, že webový server je dočasně mimo provoz.

„Mazanější“ rozhraní dokonce rozkládají požadavky tak, aby na webový server přicházely z různých stran přes odlišné proxyservery, které jsou součástí sítě a plní řadu úkolů, mimo jiné pracují podobně jako vyrovnávací paměti pro stránky. Z důvodu zjednodušení *obr. 1* na něm tento jev chybí. Roboti tím obcházejí různá omezení, jimiž se webové servery brání přetížení od nadměrných požadavků posílaných jedním uživatelem.

Webový server může usměrnit činnost „slušných“ robotů, pokud ve svém kořenovém adresáři uloží pokyny pro ně v souboru *robots.txt*. Jeho formát lze nalézt například v [5].

Jaké překážky číhají na robota

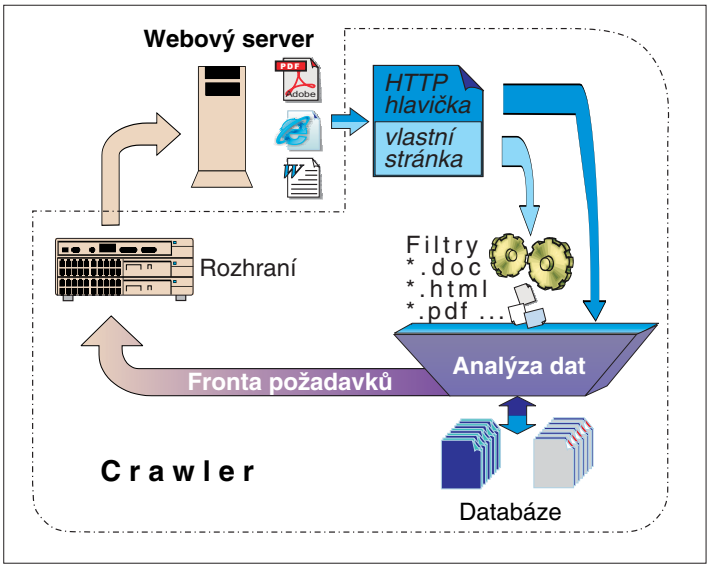
Práce robota vypadá zdánlivě jednoduše, ale komplikují ji četná úskalí. Na prvém místě stojí neurčitá vazba mezi adresou URL a korespondujícím obsahem. Odkaz na pomyslnou adresu *http://example.com/* specifikuje totiž pouze jistý interní podadresář na webovém serveru definovaný jeho konfigurací. Ta může například specifikovat, aby server v tomto případě hledal soubor *index.php* v daném adresáři, pak *index.html* nebo *index.htm*, a pokud nenalezne ani jeden z nich, aby provedl třeba přesměrování na úplně jinou adresu, resp. nahlásil chybu. Názvy uvedených souborů se samozřejmě mohou lišit, stejně jako popsaný postup.

Pokud dojde k přesměrování adresy, robot to ve většině případů vůbec nezjistí. Z principu neví, vedou-li adresy URL *http://example.com/ix.htm* a *http://example.com/* na stejné nebo zcela odlišné obsahy. Když se k tomu připočte i někdy používaná možnost mapování podadresářů či souborů na webovém serveru na zcela jiné, pak jenom pouhé rozhodnutí o identitě obsahů dvou webových stránek není vůbec triviální záležitostí. Prosté porovnání jejich obsahu nepomůže, protože webová stránka může obsahovat dílčí informace, které se mění při každém jejím čtení, jako třeba počítadla přístupů.

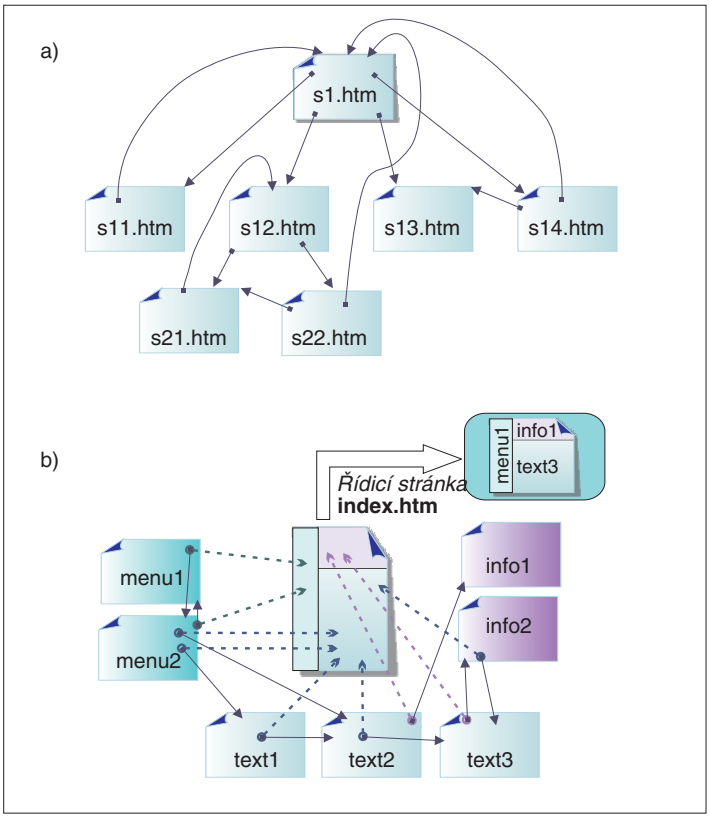
Indexování stránek dále komplikují různá kódování znaků na webových stránkách a hrubé chyby syntaxe použitých formátů (nevalidní dokumenty). Řada tvůrců webových stránek vychází z mylného přesvědčení, že když prohlížeč zobrazí webovou stránku, pak má správný formát. Tato implikace sice platí, ale pouze opačným směrem. Odpovědi na otázky tohoto typu bude obsahovat další díl tohoto článku v příštím čísle časopisu Automatizace.

Typy stránek z hlediska indexování

Indexování komplikuje i různorodost použitých technologií stránek na webových serverech. Pokud by roboti mohli vyprávět o své nekonečné sisyfovské práci, určitě by nejvíc chválili *statické stránky* znázorněné na *obr.* 2 nahoře. Ty jsou na serveru uloženy jako samostatné soubory a každé adrese URL odpovídá konkrétní soubor, takže existuje přímá korespondence mezi adresou URL a obsahem. Celá struktura statických



Obr. 1 Proces procházení adres



Obr. 2 Statické stránky (a) a rámce (b) – příklad jejich okamžitého zobrazení v prohlížeči

stránek připomíná konečný graf, jehož uzly tvoří jednotlivé stránky a hrany hypertextové odkazy, takže jeho procházení je prostě „radost sama“.

Horším případem bývají stránky s rámy (*frame design*), kdy je okno prohlížeče řídicí stránkou rozděleno na libovolný počet „podoken“, zpravidla dvě až tři, na kterých se zobrazují webové stránky, takže v prohlížeči vidíme více stránek najednou (*obr. 2* dole). Každý hypertextový odkaz specifikuje nejen stránku, ale může zvolit i rám, v němž se má stránka zobrazit, nebo nechat zobrazení na předdefinovaném nastavení. V obrázku je to naznačeno zdvojením šipek, které znázorňují hypertextové odkazy nalezené na stránce, jedna šipka odkazu specifikuje stránku, druhá určuje rám, do něhož ji má prohlížeč nahrát.

Koncepce rámů nedovoluje bohužel zadat odkaz z vnějšku na jinou kombinaci stránek než na výchozí, definovanou parametry v řídicí stránce rámu (tou je na obrázku *index.htm*), aniž by došlo k porušení celkového vzhledu. Někteří roboti rámce zcela ignorují. Jiní je sice procházejí, získávají z nich však leda dílčí stránky, takže objeví-li se v seznamech odkazů vyhledávače, uživatel spatří jen fragmenty původní složené stránky, jako třeba text bez nabídky (*menu*), či obráceně. Kvůli této nectnosti se rámy dnes prakticky nevyužívají.

Největší záludnosti pro indexování ukrývají dnes naopak hojně používané *dynamicky generované stránky*, které lze většinou poznat podle přípon php, asp nebo jsp. Program (skript) spuštěný na straně serveru vytváří webovou stránku teprve v okamžiku, kdy o ni uživatel požádá. Parametry potřebné pro činnost skriptu se přenášejí buď v HTTP hlavičce (post metoda) nebo jsou uvedeny v URL adrese příslušné stránky (get metoda), a to za otazníkem a vzájemně oddělené znaky & – například *citac.php?cnt=4&art=7&id=Karel* znamená předání tří parametrů *cnt*, *art* a *id* do skriptu (programu) v souboru *citac.php*.

Bohužel obsah vrácené stránky může, ale taky nemusí záviset na všech parametrech. Výše uvedený odkaz klidně vrátí článek (určený *art=7*), který bude ale v textu obsahovat opět odkaz na *art=7*, jenže s pomocným čítačem *cnt* zvýšeným o 1, čili na informačně stejný článek. Změna *id* třeba vůbec neovlivňuje textový obsah, protože slouží k nějaké interní identifikaci uživatele. Výhradně parametr *art* udává odlišný text. Chování takového záludného skriptu si můžete vyzkoušet na webové adrese [1].

Roboti samozřejmě neznají význam parametrů, a proto řeší vnořování do stránek různými postupy. Nejjednodušší z nich představuje pouhé omezení hloubky odkazů, ale to nezabrání duplicitě informací – i tak lze stejnou stránku obdržet mnohokrát, pokud se na ni odkazuje vícekrát ze stejné úrovně.

Sofistikovanější metody se opět zaměřují na již zmíněné porovnávání významných částí stránky, čímž zjistí, jestli byly nalezeny stránky se skutečně jiným obsahem, či uplatňují různé heuristické postupy, jako třeba analýzu adresy URL, nebo využívají apriorní informace o parametrech jim známých skriptů. Používané algoritmy se neustále mění. Například, roboti pracující pro Google dlouho ignorovali parametry obsahující id a začali zpracovávat až nedávno [2].

Nejhorší případy představují odkazy uvedené ve skriptech spouštěných na straně prohlížeče – tedy v programech javascript či vbscript, například můj odkaz. Podobné odkazy většinou roboti zcela ignorují, a tak neprohledávají stránky jimi určené, čehož se někdy využívá i záměrně pro nezvyšování hodnocení jiných stránek, které bude zmíněno v závěru článku. Chceme-li naopak, aby roboti zpracovávali odkazy v jazyku Javascript, musíme je napsat tak, aby fungovaly i při vypnutém jazyku Javascript, tedy třeba takto: můj odkaz. Další variantou jsou mapy webových stránek, jemuž bude věnováno pokračování článku v příštím čísle časopisu Automatizace.

Málokterá webová prezentace používá pouze jeden typ stránek. Mnohé obsahují rozličné kombinace všech zmíněných způsobů, takže část stránek prezentují jako statické stránky, kus mají v rámech, občas použijí dynamicky generované stránky a místy využívají i program v jazyku javascript.

Pro toho, kdo se právě rozhoduje, jaké webové stránky vytvořit, aby se výborně indexovaly prohledávači, bývají nejlepší volbou ryze statické stránky. Ty malé lze vytvořit ručně, velké je možné automaticky generovat z redakčního systému pomocí vhodného programu. Padne-li volba přece jen na dynamické stránky, pak je třeba do prvních dvou parametrů vložit podstatné informace, které určují odlišný obsah. Případně je možné doplnit informace v hlavičce HTTP tak, aby obsahovaly shodné atributy jako stránky statické (např. čas poslední změny obsahu).

Ohodnocení stránky – *page rank*

Vyhledávače pro svoje potřeby všechny indexované stránky ohodnocují. Kritéria hodnocení mohou být různého charakteru, od podmínky existence všech zadaných klíčových slov ve zbytku dokumentu až po ohodnocení (*page rank*) stránek na základě odkazů z jiných stránek.

Posledně jmenované kritérium má v dnešních vyhledávačích velký význam a citelně ovlivňuje umístění odkazu. Jeho algoritmus [3] byl poprvé použit ve vyhledávači Google, v němž se mu dodnes přikládá velká váha. Jedná se o iterační algoritmus, v němž se hodnoty automaticky aktualizují. Jednou za dva až tři měsíce se také upravuje aktualizační algoritmus, což ovlivňuje pořadí odkazů tak výrazně, že jednotlivé úpravy, obecně označované jako Google dance, dostávají i svá osobitá jména podobně jako hurikány – listopadová 2003 se nazývala Florida, lednová 2004 Austin. Končí spokojeností majitelů lépe umístěných stránek a protesty „poškozených“ majitelů a říká se, že se objevily i žaloby za ušlý zisk.

Kritérium vychází z předpokladu, že stránky s kvalitním obsahem (tj. stránky dobře ohodnocené) budou rovněž odkazovat jen na jiné podobně kvalitní stránky. Pokud nějaká stránka B_i odkazuje n_i jiných stránek, všem jim přispěje částí hodnoty svého ohodnocení *page rank*. Hodnota je rovnoměrně rozdělena mezi všechny odkazované stránky. Každý vyhledavač používá vlastní algoritmus. Činnost si můžeme demonstrovat na nejjednodušším vzorci pro výpočet hodnoty *page rank*:

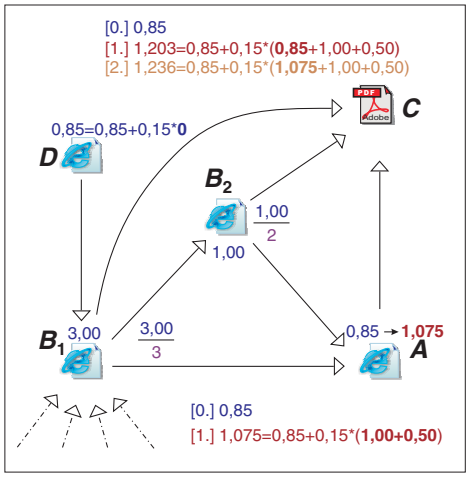
$$PR(A) = k + (1-k)\sum_{B_i}\left(PR(B_i)/n_i\right) \qquad (1)$$

kde k je konstanta algoritmu z intervalu 0 až 1, jejíž doporučovaná hodnota 0,85 zaručuje dobrou konvergenci. $PR(B_i)$ udává *page rank* stránky B_i, n_i udává počet stránek, na které stránka B_i sama odkazuje, sčítá se přes všechny stránky B_i, které odkazují na stránku A. Postup výpočtu hodnoty *page rank* (PR) objasňuje *obr. 3*, které šipky představují hypertextové odkazy vedoucí z jedné stránky na další. Je v něm použitá doporučená hodnota konstanty $k = 0{,}85$. Jestliže např. na stránku označenou jako B_1 vede hodně vnějších odkazů, může její $PR(B_1)$ dosáhnout hodnotu 3,00. Pro jednoduchost předpokládejme, že jde i o stabilní (dalšími iteracemi neměnnou) hodnotu. Stránka B_2 má rovněž stabilní hodnotu PR, protože na ni odkazuje jen stabilní B_1. $PR(B_2) = 0{,}85 + (1 - 0{,}85)\,(3{,}00/3) = 1$. Hodnota $PR(B_1) = 3{,}00$ byla vydělena třemi kvůli tomu, že B_1 sama ukazuje na tři další stránky. Naproti tomu stránka D, na kterou nic neodkazuje, má hodnotu $PR(D)$ určenou jen použitou konstantou k, čili 0,85. Tou se mohou inicializovat i všechny dosud nevypočítané stránky v nultém kroku.

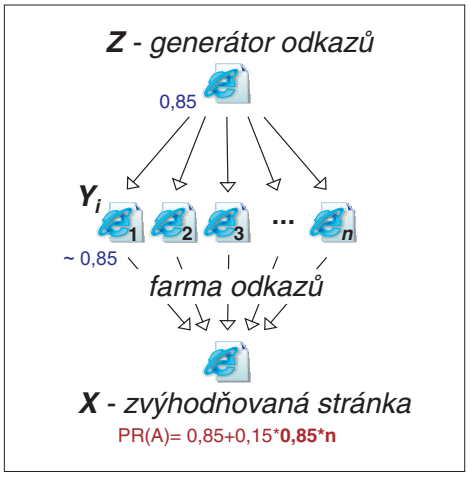
Dosud neurčené hodnoty $PR(A)$ a $PR(C)$ na *obr. 3* se vypočítají takto: Na stránku C směřují tři odkazy ze stránek A, $B1$ a $B2$. Jejich hodnoty PR dělené počtem různých odkazů, které jednotlivé stránky obsahují, se dosadí do vzorce (1):

$PR(C) = 0{,}85 + 0{,}15\,(PR(A) + PR(B1)/3 + PR(B1)/2)$; $PR(C) = 0{,}85 + 0{,}15\,(0{,}85 + 1{,}00 + 0{,}50) = 1{,}203$.

Podobným postupem vypočteme i $PR(A)$, jak je uvedeno na *obr. 3*.



Obr. 3 Iterace hodnot page rank



Obr. 4 Farma odkazů

Další iterace nemění hodnotu $PR(A)$, protože se nemění ani PR stránek B_1 a B_2, které na A odkazují, avšak ustálení hodnoty $PR(C)$ si žádá ještě další druhý krok, jelikož na stránku C odkazuje A a $PR(A)$ se zvýšila v prvním kroku. V druhém kroku vzroste proto i hodnota $PR(C)$.

Zájemcům lze na síti Internet doporučit kalkulátor hodnoty *page rank* [4], včetně detailního popisu.

Triky pro lepší *page rank* se nevyplácí

Tvůrci webových stránek brzy objevili, že výše popsanou metodu lze poměrně snadno zneužít – stačí vytvořit mnoho pomocných stránek, čím jich bude více, tím lépe. Postup ukazuje *obr. 4*.

Jednu pomocnou stránku označme jako Z a nazveme ji generátor odkazů, protože jejím účelem je pouze ukazovat na n dalších pomocných stránek Y_1 až Y_n. Stránka Z se zaregistruje u prohlížeče, aby o ní roboti věděli. Vzhledem k tomu, že vzorec pro výpočet hodnoty *page rank* dává konstantu k jako svou minimální hodnotu, pak, s ohledem na případné zaokrouhlení platí: $PR(Y_i) \cong k = 0{,}85$. Nyní přidáme na každou stránku Y_i odkaz na stránku X, kterou chceme zvýhodnit. Přírůstek ohodnocení stránky X lineárně roste s počtem „záškodnických“ stránek. Tato metoda bývá označována jako farma odkazů (*link farm*).

Takový způsob zvyšování hodnocení stránek se však nedoporučuje. Metoda je již známá a roboti, alespoň ti, kteří pracují pro dobré vyhledávače, se těmto technikám brání a znatelně penalizují stránky, u nichž podobné podvody zjistí. To však bohužel otevírá pole pro nasazení jiného triku, kdy se úmyslně vytvoří „záškodnická“ farma odkazů pro konkurenci, aby se poškodila v očích vyhledávačů. Roboti její stránky v dobré víře penalizují, čímž je původní záměr splněn. Samozřejmě i tento podvod je znám a také je již v současné době detekován.

Výše popsané způsoby jsou v článku uvedeny nikoli jako návody, ale pro ilustraci obtížnosti prohledávání webových stránek a četných úskalí, která se musí řešit, aby výsledná informace dávala relevantní výsledky. V pokračování článku, které se objeví v dalším čísle, budou popsány vlastní vyhledávače a uvedena i některá doporučení pro tvůrce stránek, jimiž lze korektně zvýšit umístění stránek.

Ing. Martin Řimnáč
Ing. Richard Šusta, Ph.D.
Ing. Jiří Živnůstka

LITERATURA

[1] Internet: *http://susta.cz/citac.php* – demonstrace cyklického odkazu zmíněná v textu článku

[2] *http://www.jakpsatweb.cz/weblog/archiv/2004-03.html#260023* – diskuse o id parametru u Google.

[3] HENZINGER, M. R.: Hyperlink Analysis for the Web. IEEE Internet Computing, 2001, January/February, s. 45–50.

[4] CRAVEN, P.: Google's PageRank Explained. Poslední revize 12. 3. 2004. <*http://www.webworkshop.net/pagerank.html*>

[5] Internet: *www.robotstxt.org/wc/norobots.html* – popis formátu souboru robots.txt