Automatické prohledávání webových stránek II

První díl článku (Automatizace č. 4/2004) byl zaměřen převážně na činnost robotů, jejichž úkolem je stahovat a analyzovat webové stránky. Tento díl objasní, jak vyhledávače využívají výsledky jejich práce. V závěru jsou pak uvedena i některá doporučení pro psaní webových stránek, aby se dobře prohledávaly.

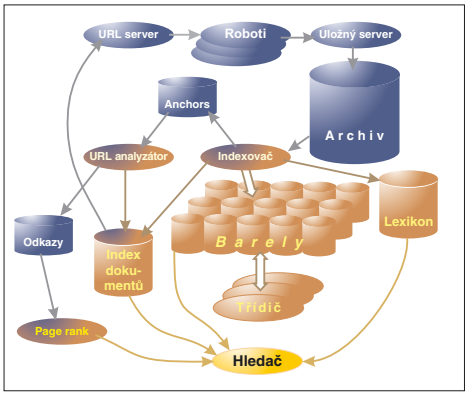
Google – přísně tajné

Úvodem je třeba poznamenat, že provozovatelé vyhledávačů nemají zájem zveřejňovat cokoliv konkrétnějšího, čím by konkurenci usnadnili cestičku ke stvoření jiného ještě úspěšnějšího vyhledávače, případně poskytli podklady pro možné domněnky o protěžování nějakých stránek. Proto nelze žádnému zasvěcenému pojednání o struktuře některého z populárních vyhledávačů, jako třeba o Google, beze zbytku věřit (tedy ani tomu, který právě čtete). Vyhledávače si dnes osvojily roli jakýchsi internetových reklamních agentur se všemi dopady, které z podobného postavení vyplývají, včetně kritik ze strany veřejnosti. Pouhá úprava algoritmu použitého pro ohodnocení stránek – *Google Dance* (popsaný v předchozím dílu článku) – vyvolává občas tak bouřlivé reakce, že se někdy píše i o firmách jím poškozených, jako kdyby přišla živelná pohroma. A možná ještě něco horšího, neboť proti povodním je možné se pojistit, ale proti *Google Dance* nikoliv, aspoň ne v dnešní době…

Přesnou strukturu a algoritmy vyhledávačů halí pečlivé mlčení, a tak je nutné spokojit se s pouhými náznaky. Pro Google lze najít nějaké informace v dokumentačním projektu [1], který napsali jeho tvůrci Sergey Brinem a Lawrence Page. V něm ale popisují pouze svůj prvotní návrh – současný stav se však může výrazně lišit. Další informace lze posbírat po rozličných článcích a debatních fórech věnovaných problematice vyhledávačů, kde různí nadšenci zveřejňují své experimenty nebo teoretické studie. Podle těchto zdrojů sestavená struktura vyhledávače Google jen velmi pravděpodobná možnost z mnoha dalších a jiní autoři se mohou klonit k jiným názorům.

Jak funguje Google

Google vznikl v roce 1998 na Standfordské univerzitě, kde v počátcích i hostoval. Své jméno obdržel podle čísla *googolplex*, jehož zápis se skládá z jedničky následované



Obr. 1 Náznak struktury vyhledávače Google

googol nulami, přičemž *googol* se rovná 10^{100} (názvy obou čísel zavedl matematik Edward Kasner, 1878–1955). Číslo *googol* převyšuje svou velikostí i počet všech částic v celém vesmíru, odhadovaný mezi 10^{72} až 10^{87}. Ještě větší číslo *googolplex*, rovnající se 10^{googol}, se pak už zcela vymyká jakýmkoliv lidským představám.

Vyhledávač Google se svou aktivitou úspěšně blíží těmto nesmírně vysokým číslům, neboť odpovídá už na více jak 200 milionů dotazů denně, čili v průměru vyřídí přes 2 300 žádostí každou sekundu. Pracuje pro něj více než 10 000 serverů, vybavených operačním systémem Linux a umístěných ve více než třinácti výpočetních centrech. Při generování odpovědí vychází z podkladů získaných analýzou více než 3,3 miliardy webových stránek. K jejich přečtení by jeden člověk potřeboval asi 20 000 let, avšak Google je prý prohledá za 0,5 sekundy.

Základní struktura Google, naznačená na *obr. 1*, není nikterak složitá. Server URL adres předává tyto adresy robotům (*crawlers*) a jimi získané webové stránky se poté posílají úložnému serveru (*StoreServer*), který je komprimuje pomocí metody Zlib a ukládá do archivu (*Repository*). Každá stránka při ukládání obdrží svůj identifikátor (*docID*) a obsah archivu pro danou stránku lze zobrazit, když se ve výpisu vyhledávače zvolí odkaz *archiv*.

Archivují se kompletní kódy stránek HTML, ale pouze do omezené velikosti. Google v současné době, a to podle údajů jím nedávno potvrzených, omezuje data z běžné stránky na 101 kB, z toho se zhruba 1 kB předpokládá pro hlavičku HTTP a 100 kB pro text vlastní stránky bez obrázků. Pokud však Google vyhodnotí dokument jako zajímavý, zpracuje se celá stránka, zejména v případě souborů PDF. Odlišně se přistupuje i ke katalogům, čili seznamům odkazů vytvářených uživateli katalogu ve spolupráci s jeho správci. Katalogy se sice neukládají do archivu celé, ale vždy se zpracovávají všechny odkazy na nich uvedené.

Indexovač (*indexer*) zajišťuje řadu funkcí. Vybírá jednotlivé stránky z archivu a zpracovává je. Z načtené stránky vytvoří množinu slov a u každého slova zjistí počet jeho výskytů, jeho umístění v textu a důležitost odvozenou podle prvku (např. element HTML), kde bylo slovo použité. Tyto údaje mají společný název HIT a schraňují se v množině barelů (*barrels*).

Indexovač zjišťuje rovněž informace o odkazech v dané stránce. Google váže odkazy nejen ke stránce, na níž se vyskytuje, ale i k té, na kterou směřuje, což se nazývá *anchor* (kotva, zarážka) a ukládá je do speciálních souborů. Spojení odkazu s cílovou stránkou přináší četné výhody. Popisky u odkazů, zejména u těch, které vytvořila nějaká nezávislá strana, dávají často mnohem výstižnější informaci o významu cílové stránky a jejím obsahu než stránka samotná. Kromě toho existují i pro dokumenty, které nelze z technických důvodů prozkoumat textově orientovanými prohledávači, například obrázky, programy a multimediální soubory. Nicméně *anchor* umožní, aby vyhledávač vracel referenci na tyto soubory, ačkoliv nebyly vůbec prohledané.

Analyzátor URL (*URLResolver*) přetváří relativní odkazy URL na absolutní a těm přiřazuje jedinečné identifikátory. Dále generuje seznam dokumentů a podklady pro výpočet *page rank* probíraného minule. Nakonec třídič (*Sorter*) vybírá data z barelů, setříděná podle identifikátoru dokumentů, a řadí je podle slov. Generuje tak invertovaný seznam, uložený v lexikonu, který dovoluje rychle najít dokumenty, v nichž se slovo vyskytuje.

Jak vypadá taková odpověď na dotaz složený z jednotlivých slov a logických spojek? Jeho vyhodnocení prochází následujícími kroky. Nejprve se rozloží na jednotlivá slova, pro ta se zjistí jedinečné identifikátory a určí se začátky seznamů HIT v barelech. Jejich postupným prohledáním se najdou dokumenty vyhovující celému dotazu. Když se některý takový nalezne, je vypočtena jeho *relevance*, čili důležitost vzhledem k danému dotazu. Podle té jsou nakonec výsledky utříděny, a to sestupně, od dokumentu s nejvyšší relevancí dolů.

Pokud se uživatel zajímá o pojem, který se vyskytuje příliš často, vyhledávač nezkoumá všechny reference, ale přeruší pátrání, jakmile získá dostatečné množství relevantních referencí. Z těch odhadne, kolik dokumentů by asi mohlo najít při úplném prohledání – vždyť většinou lidé stejně přečtou jen pár prvních stránek výsledků.

Zapeklitá sémantická analýza

Obecně známá struktura vyhledávače Google nedává odpověď na řadu klíčových otázek. Ty zůstávají tajemstvími. Pominemeli všechny realizační problémy, tzn. problémy spojené s efektivitou práce a distribucí problému vyhledávání na více počítačích, zůstávají také nejasnosti ohledně ohodnocení stránek a určení jejich *relevance* vůči dotazu, čili stanovení vhodného pořadí ve výpisu tak, aby stránky pro tazatele zajímavé byly na předních místech. Právě toto totiž rozhoduje o kvalitě vyhledávače.

Zhruba se ví, že relevanci dokumentu ovlivňuje to, kolikrát se na něm vyskytuje hledané slovo, v případě víceslovného dotazu pak vzdálenost mezi výskyty jednotlivých slov dotazu, a dále ohodnocení stránek (*page rank*), pro který je znám základní vzorec, avšak nikoliv přesný – ten tvůrci vyhledávače Google pečlivě tají. Odhaduje se jen, že jeho vzorec zahrnuje různé penalizace a detekce podvodných pokusů o zvýšení relevance a také nově zaváděný *tematický page rank*, který upravuje ohodnocení stránky podle odkazů z hodnověrných externích katalogů. Výsledná relevance je pak váženým součtem všech těchto dílčích kritérií.

Ale ani takto složitý vzorec nestačí k nalezení dokumentů zajímavých pro tazatele. Kdo chce skutečně něco vyhledat, nemůže zadávat příliš obecné dotazy. Například slovo „PID“ vede k záplavě referencí, z nichž jen zlomek souvisí s zamýšlenou PID regulací. Zkratka PID se totiž používá pro další pojmy: označuje i jednu nemoc, zařízení na sběrnici, program pro identifikaci a pražský dopravní integrovaný systém. Zúžíme-li však dotaz, třeba na vazbu PID regulace, pak prostým porovnáním slov by se nenašly dokumenty obsahující spojení „PID regulátor“ nebo „PID regulování“, eventuálně používající synonymum „PID řízení“.

Vyhledávání by tak snadno mohlo *jednu neurčitost nahradit jinou* – bylo by sice možné vypátrat, kde se požadovaný údaj nachází, ale jedině při znalosti přesné otázky. Modernější vyhledávače se proto snaží zadaný výraz analyzovat, odhadnout obor, o který se tazatel zajímá, a vypočítat relevanci dokumentů nejen z počtu výskytů hledané fráze, ale vzít v úvahu i její ekvivalenty, v našem případě „PID regulátor“ a „PID řízení“.

Webová stránka se kvůli tomu musí během indexování zpracovávat ještě sémanticky (*obr. 2*). Nejprve se provádí jeho *lexikální analýza*, při níž se text dělí na jednotlivá slova, čísla nebo speciální znaky. Problémy nastávají s interpretováním určitých slovních spojeních. Například pouhé rušení rozdělujících znamének v textu není vůbec jednoduché. Pokud se třeba ve slově „sci-fi“ odstraní pomlčka, vzniknou dvě slova se zcela odlišným významem. Obdobné problémy nastávají v reprezentaci telefonních čísel a letopočtů. Zvláštní přístupy vyžadují i speciální entity (X)HTML jako „ , &, ©“ a jiné.

Výsledek lexikální analýzy může postoupit do sekce zvané *eliminace stop slov*, kde se ze seznamu slov eliminují slova (především spojky nebo předložky), která se vyskytují ve většině dokumentů. Jejich vynecháním se zmenší velikost ukládaných dat, ale ztěžuje se hledání podle frází. Google například v anglických dokumentech vynechává příliš častá slova jako například „I“, „of“, „for“, „with“ a další. A při zúžení hledání jen na české stránky se zase vyloučí spojky „a“ a „i“.

Slova dále procházejí *lemmatizérem*, tzn. modulem zabývajícím se převodem slov na slovníkové tvary, jehož základní část tvoří *stemmer* (stem – kořen slova). Pro převod na základní tvar lze použít tabulkový model nebo metody založené na odstraňování přípon – Porterův algoritmus [2] a v neposlední řadě program *ispell*. Ten ve své podstatě představuje slovník obsahující všechna slova spolu s příznaky, podle kterých se zjistí všechny další možné tvary. Slovník *ispell* pro většinu jazyků lze nalézt na [3].

Při analýze se dále hodně využívá *thesaurus*, který obsahuje synonyma daných slov. Matematický popis lze nalézt na [4]. Výsledné ohodnocení dokumentu pak může záviset nejen na počtu nalezených výskytů vazby „PID regulace“ a příslušných vyskloňovaných a odvozených tvarech, jako „PID regulacemi“, „PID regulátor“ apod., ale současně i na výskytu synonym, třeba „řízení“, či oborově blízkých pojmů, jako třeba „termoregulace“ a „stabilizace“.

Velmi důležitým modulem je rovněž *detektor triků*. Již v prvním dílu článku byly popsány podvodné snahy o lepší ohodnocení stránek (*page rank*) pomocí farem odkazů. Podvádí se pochopitelně i v oblasti výskytu slov. Všechny níže zmíněné triky jsou již známé a vyhledávač podobné stránky penalizuje snížením jejich relevance.

Častým trikem bývají hlavně bezvýznamné texty, které nejsou vidět při zobrazení stránky v prohlížeči, ale obsahují pojmy zvyšující hodnocení. Sem patří třeba bílé písmo na bílém pozadí, eventuálně text šikovně překrytý jinými prvky, nebo zamaskovaný kresbou, kterou generuje JavaScript. Jinou možnost představuje umístění falešných textů až na konec stránky a jejich oddělení od informací významných pro návštěvníka velkou prázdnou mezerou nebo vhodnou značkou, která vzbuzuje dojem, že už nenásleduje nic zajímavého. Detektory triků musí proto brát v úvahu nejen výskyty slov, ale i jejich viditelnost pro uživatele.

Mnohem problematičtěji se zjišťuje opačná situace – tajení nevhodného obsahu. Vše, co mělo uniknout robotům, se vloží jako obrázky, k nimž se uvedou jiné textové popisky, než uživatel skutečně uvidí, aby stránka získala jiné ohodnocení, než by si zasloužila.

Jaké stránky se dobře vyhledávají?

Zásady pro tvorbu webových stránek, které se budou dobře vyhledávat, se často označují pojmem SEO (*Search Engine Optimization*). Všechny tyto zásady nelze vyjmenovat, už proto, že mnohé se rychle mění. Jak již bylo uvedeno v předešlé části článku, Google dlouhou dobu ignoroval parametr *id* u dynamicky generovaných stránek – bral jej jako unikátní označení návštěvníka, tzv. *sessions* – a teprve nedávno ho začal zpracovávat. Hojně uváděné pravidlo „nepoužívat *id*“ tak ztratilo svůj význam doslova ze dne na den.

I uznávaná zásada, že prohledávací roboti neprocházejí JavaScript, může brzy padnout, protože Google začíná údajně již experimentovat s jeho analýzou. V následujících statích se proto zmíníme jen o hlavních pravidlech, u nichž lze předpokládat dlouhodobější platnost.

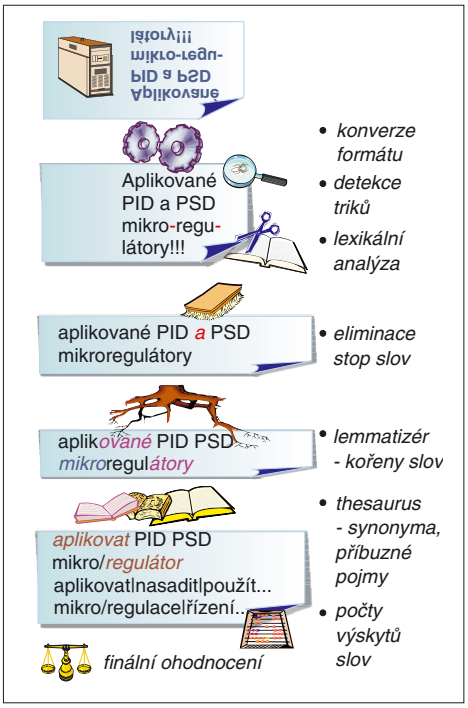
Zveřejňujte na stránkách zajímavé informace, a to i v textové formě. Právě informace přitahují návštěvníky. Vzhled je sice také důležitý, avšak je třeba mít na paměti, že grafické prvky, jako obrázky, Flash či kresby generované jazykem JavaScript, nejsou viditelné pro textové prohledávající roboty – pro ně neexistují. Ke každému grafickému prvku by proto měla existovat jeho textová analogie. U obrázků je vhodné vždy vyplňovat atribut *alt*. Pokud jsou na stránce prezentovány důležité informace pomocí programů Macromedia Flash či JavaScript, je třeba k nim vytvořit i textovou analogii, třeba pod nabídkou „textová verze“.

Nevytvářejte příliš velké stránky. Architektuře vyhledávače Google vyhovují stránky menší než 101 KB. Mezi výjimky mohou patřit pouze soubory PDF, resp. katalogy odkazů.

Vyhýbejte se rámům, jelikož nedovolují zadat libovolnou sestavu zobrazených stránek. Blíže se o problémech s jejich prohledáváním psalo v minulém dílu.

Uveďte výstižné titulky stránek definované v (X)HTML v elementu <*title*>. Mnohdy titulek chybí nebo mívá shodnou hodnotu na všech stránkách, takže potencionální zájemce vidí v seznamu odkazů vyhledávače jen nic neříkající reference, jako třeba pouhý název firmy. Má-li být název firmy uveden v titulcích, je vhodné jej připojit až na konec, na začátek je lepší uvádět popis obsahu stránky.

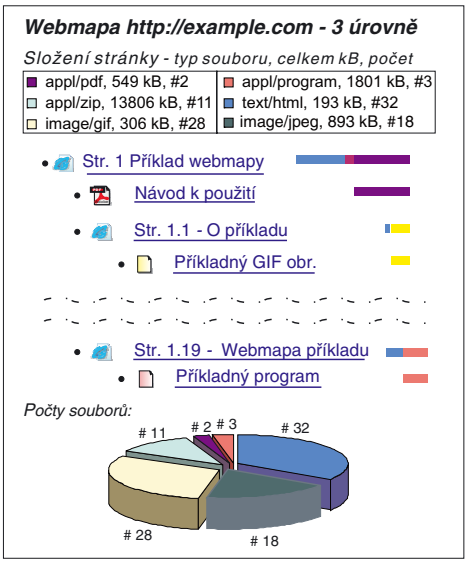
Dbejte na svůj page rank. Pomocí vyhledávače je možné zjistit, z kolika stránek vede odkaz na sledovanou stránku. Ve vyhledávači Google je doporučeno vyhledávat odkazy příkazem „link:http://*example.com*/“, kde se text napsaný kurzívou nahradí jménem domény sledované stránky. Tento příkaz však nevrací všechny odkazy směřující na sledovanou stránku, ale jen ty z významných stránek. Zde nejlépe pomůže prosté vyhledání jména dané domény. Je nutné vzít přitom



Obr. 2 Analýza dokumentu

v úvahu nejen počet nalezených odkazů, ale i jejich kvalitu. Pokud na sledovanou stránku ukazuje jiná stránka, která má sice extrémně vysoké ohodnocení (*page rank*), ale sama odkazuje na mnoho dalších stránek, pak to sledované stránce příliš neprospěje. Mnohem výhodnější je, když na ni ukazuje stránka, která má sice nízké ohodnocení, ale ukazuje pouze na ni.

Další důležitou věcí je i umisťování odkazů směřujících mimo stránky. Zde je výhodné domluvit se o vzájemné výpomoci, kdy si dvě firmy nebo jednotlivci umístí na své stránky odkazy jeden na druhého. Vzájemné odkazy mají prohledávací roboti rádi, protože dokazují, že se nejedná o pochybné techniky popsané v předchozím dílu. Pozor



Obr. 3 Webmapa usnadní orientaci návštěvníkům i prohledávacím robotům

ale, aby roboti uměli prohledávat odkaz umístěný protistranou a služba nebyla jednostranná.

Využijte robots.txt, soubor obsahující pokyny pro prohledávací roboty. Hodí se v případě, je-li na stránce použito přesměrování, nebo pokud nemají být určité stránky indexovány. Bližší popis formátu souboru lze nalézt například v [7].

Zaregistrujte stránku v seriózních katalozích. Odkazům, které se v nich vyskytují, mohou vyhledávače přikládat větší váhu, protože se předpokládá, že uvedené údaje aspoň částečně ověřují správci katalogů.

Pište validní stránky, tedy syntakticky správné a vyhovující normám, což si lze snadno překontrolovat třeba bezplatnou službou [5]. Bezchybná syntaxe (X)HTML dokumentů tvoří základní předpoklad pro jejich správné indexování. Navíc upozorní na základní prohřešky, jako například chybějící atributy *alt* u obrázků či vynechané specifikování normy, podle níž byl dokument napsaný, takže její kontrola se vřele doporučuje. Je třeba také poukázat na problematickou konverzi z některých textových editorů, nejznámější jsou potíže s dokumenty napsanými v editoru MS Word a uloženými jako soubory HTML – v jejich luštění se často vyzná jedině MS Explorer. Takové dokumenty je vhodné vždy překontrolovat *validátory*.

Na stránky přidejte údaje důležité pro prohledávací roboty. Prohledávání je proces plně automatický a roboti potřebují rozhodně znát jazykovou verzi stránky a kódování, ve kterém je uložena. Kromě toho je nutné definovat i hlavní uživatelské meta atributy, jejich popis najdete třeba v [6]. Doporučuje se hlavně popisek stránky, meta atribut *<meta name="description" …>*, kterému vyhledávače přikládají větší váhu než jinému textu. Někdy se přidává i meta atribut s klíčovými slovy *<meta name="keywords" …>*. Jejich seznam by neměl být delší než 1 000 znaků a současně je třeba dbát na to, aby se uvedená klíčová slova hojně vyskytovala v textu příslušné stránky. Je-li mezi klíčovými slovy pojem, který se na ní vůbec neobjeví, může naopak hrozit snížení hodnocení stránky. Seznamy klíčových slov se totiž v minulosti až příliš často zneužívaly.

Používáte-li dynamicky generované stránky, v prvních třech parametrech předávejte podstatné informace určující odlišný obsah stránky. Mají-li webové prezentace složitější strukturu, je pro jistotu vhodné vytvořit její *statickou webmapu*, čili jakousi analogii obsahu knihy. Stačí pouhý seznam odkazů na webové stránky. Webmapa může být samozřejmě i mnohem podrobnější s využitím grafických prvků (*obr. 3*).

Podobná webmapa nejen usnadní orientaci návštěvníkům, ale především zaručí, že roboti prohledají všechny stránky. Může rovněž posloužit i jako náhradní reference *pro ošetření chyby 404* (stránka nenalezena), na kterou se často neoprávněně zapomíná. Problematika automatického generování webmap je podrobněji pojednána v [10 a 11].

Co říci na závěr?

Všechna uvedená doporučení pro psaní webových stránek představují opravdu jen základní pravidla, pouhou kapku v oceánu problémů optimalizace pro automatické prohledávání stránek. Kdo chce mít opravdu dobré webové stránky, musí zajistit, aby jejich správce neustále sledoval novinky týkající se problémů se SEO, které jsou uvedeny např. na doméně v [6] či na některém diskusním fóru věnovaném tomuto tématu, třeba na [8].

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

LITERATURA

[1] BRIN, S. – PAGE, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. *www.internet-marketing-branding.com/Google/Google-secret.html*

[2] Internet: *http://sourceforge.org/* – projekt Snowball popisující Porterův algoritmus

[3] Dictionaries for International Ispell. *http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell-dictionaries.html*

[4] PÁNEK, K.: Jehla v kupce sena: Thesaurus. *www.lupa.cz/clanek.php3?show=2213.*

[5] Internet: *http://validator.w3.org/* – ověření syntaktické správnosti stránky

[6] Internet: *www.jakpsatweb.cz/meta_tagy.htm* – popis kódů HTML stránek

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

[8] Internet: *http://seo.nawebu.cz* – diskusní SEO fórum

[9] Internet: *www.webpronews.com/insiderreports/searchinsider/wpn-49-20040406GoogleIndexesDocumentsFirst101k.html*

[10] ŘIMNÁČ, M.: Mapa webové sítě. Dipl. práce. Praha, ČVUT FEL 2004. *http://dce.felk.cvut.cz/dolezilkova/diplomky.htm*

[11] ŽIVNŮSTKA, J.: PHP rozhraní pro vyhledávání a správu mapy sítě. Dipl. práce. Praha, ČVUT FEL 2004. *http://dce.felk.cvut.cz/dolezilkova/diplomky.htm*