Principy vyhledávání informací v prostředí internetu

LIBOR MĚSÍČEK – PAVEL PETRUS

Přírodovědecká fakulta UJEP – Pedagogická fakulta UJEP, Ústí nad Labem

V hodinách informatiky na středních školách se žáci seznamují s internetem. Cílem tohoto článku je seznámit žáky s fungováním internetových vyhledávačů a ukázat, jakým způsobem lze s nimi efektivně pracovat. Článek také okrajově popisuje nástroje umožňující analyzovat hledané fráze.

Principy fungování vyhledávačů

Internet tvoří v současné době přibližně 4,57 miliard viditelných stránek (údaj k 1. 6. 2016) [1]. Na první pohled je jasné, že není možné procházet všechny weby k nalezení požadované informace, ale je nutné k tomu použít nějaké nástroje. V podstatě máme k dispozici jen dva nástroje: katalogy a vyhledávače.

Katalog si můžeme přiblížit na příkladu telefonního seznamu typu „Zlaté stránky“, kde např. v kategorii instalatérství bylo možné najít instalatéry. Katalog je tedy internetová stránka (web), který obsahuje rozdělení dalších webů do jednotlivých kategorií (případně podkategorií) dle příslušnosti k dané kategorii (podkategorii). Weby jsou do katalogů umísťovány ručně a přední pozice v daných kategoriích lze zakoupit. Nejznámější české katalogy jsou Seznam.cz, Centrum.cz a Atlas.cz. Stránky Seznam.cz umožňují vyhledávat s pomocí katalogu, ale je na nich k dispozici i internetový vyhledávač.

Internetový vyhledávač (*search engine*) je naopak nástroj, který aktivně prohledává web. V současné době existuje několik vyhledávačů, např. Google.com, Seznam.cz, Yahoo.com, Bing.com. Vyhledávače pracují ve třech krocích.

V **prvním kroku** robot (pavouk) sbírá data. Robot (pavouk) je počítačový program, který slouží k prohledávání internetových stránek. Robot začne na nějaké stránce, např. to může být stránka katalogu. Na dané stránce najde další odkazy, které dále sleduje, tj. vstoupí na odkazované weby, stáhne si jejich obsah a hledá odkazy na další stránky. Robot si sa-

mozřejmě pamatuje stránky, které již navštívil. Jeho pohyb lze znázornit graficky tak, že odkazy jsou čáry mezi body, kde body představují jednotlivé weby (od toho název pavouk). Roboti se po určitém časovém úseku (dnů až měsíců) vrací na již navštívené stránky, aby zaznamenali změny na stránkách, odhalili nefunkční stránky, ale hlavně znovu přeindexovali dané stránky. Indexace stránek je ohodnocení daných stránek na základě celé řady parametrů (více v druhém kroku vyhledávače).

Výsledné ohodnocení se poté uloží do databáze daného vyhledávače. Indexace stránek probíhá při každém navštívení daného webu a tím dojde k aktualizaci ohodnocení příslušného webu. Každý vyhledávač má svého robota, kde ty nejznámější jsou SeznamBot (Seznam.cz), Googlebot (Google.com), Slurp (Yahoo.com) a Bingbot (Bing.com). Existují ovšem stránky, na které se robot vrací pravidelně i vícekrát denně, protože obsahují aktuální a často vyhledávané informace. Pro zajištění rychlejšího zachycení změn je možné používat jiné metody než klasický crawler (např. sledovat informační kanál RSS stránky a nové články pomocí něj indexovat). Příkladem jsou zpravodajské weby novinky.cz, idnes.cz, cnn.com a bbc.co.uk. Zveřejněný článek nebo informace jsou tak během hodin dohledatelné i ve vyhledávači. Pokud zadáme přímo titulek nebo frázi z článku, Google umístí jako první výsledek odkaz na plný text článku i s hlavní fotografií z článku.

Autoři webových stránek se mohou pokusit ovlivnit chování robota na svých stránkách pomocí dvou souborů: robots.txt a sitemap.xml. Některé vyhledávače ovšem ignorují nastavení a indexují bez ohledu na obsah souborů robots.txt a sitemap.xml.

Soubor robots.txt každý robot navštíví jako první na dané stránce (pokud existuje). V tomto souboru jsou informace pro robota, zda může dané stránky prohledávat, či zda jsou na těchto stránkách pro něj nějaká omezení (např. zakázané adresáře). Důvodem omezení může být např. citlivost informací, které jsou umístěny na dané stránce. Pokud soubor neexistuje, tak to roboti chápou tak, že nemají jakékoliv omezení v následném zpracování. Soubor má podobu prostého textového souboru, který je umístěn v kořenovém adresáři. Na jednotlivých řádcích lze přesně specifikovat, co může který robot přesně prohledávat [2].

Soubor sitemap.xml má na rozdíl od souboru robots.txt pomoci vyhledávačům s orientací na konkrétním webu. Obsahuje odkazy na stránky a metadata. Metadata jsou data o datech, které vyhledávači poskytují další informace o jednotlivých stránkách (informace pro roboty, klíčová slova

atd.). Má příponu XML a je rovněž umístěn v kořenovém adresáři webu. Soubor sitemap.xml lze napsat jako prostý text, či je možné ho vytvořit s pomocí generátorů (např. *Sitemap Generator*) [3]. Tento soubor najde využití hlavně u velkých webů, webů s rozsáhlým archivem stránek, které jsou izolovány, či jsou špatně propojené, web je relativně nový a směřuje na něj málo odkazů [4].

V **druhém kroku** dojde k indexování stránek. Stažené stránky se na základě celé řady faktorů (titulek, klíčová slova, popis, počet výskytu slov atd.) ohodnotí – je jim přiřazena váha *Google page rank (GPR)*. GPR (u Seznam.cz se nazývá S-rank) je číslo, které je přiřazeno ke každé stránce (URL). Má význam věrohodnosti (oblíbenosti) stránky. Hodnota GPR je v rozsahu 0 až 10, přičemž čím víc se GPR daného webu blíží k 10, tím je web oblíbenější [5]. Faktory, které ovlivňují hodnotu GPR, můžeme rozdělit do dvou skupin: on-page faktory a off-page faktory.

On-page faktor je vše, co se dá změnit na webu, aby se co nejlépe umístil mezi výsledky vyhledávání. Mezi ně patří titulek stránky, správně vyplněné meta značky description a keywords, obsah (ve smyslu vlastní text) stránky a řada dalších. Za nejdůležitější je považován titulek stránky, který by měl využívat klíčová slova a měl by být unikátní pro každou stránku. Uvádí se, že optimální délka titulku je kolem 50 znaků [6].

Off-page faktory jsou všechny faktory, které mají vliv na umístění mezi výsledky vyhledávání a provádí se mimo webové stránky. Jelikož jsou tyto faktory špatně ovlivnitelné, tak mají větší váhu než on-page faktory. Mezi hlavní off-page faktory patří zpětné odkazy, tj. z jakých stránek přicházejí, kolik jich je, jak jsou stránky relevantní, kde jsou umístěny a jaké je textové okolí odkazu [7].

Pro zlepšení pořadí stránky ve výsledcích hledání slouží SEO (*search engine optimization*), což je označení metod pro vytváření a úpravu webových stránek tak, aby jejich obsah a podoba byla vhodná pro internetové vyhledávače. Cílem SEO je získat co nejlepší pozici pro danou stránku ve výsledcích vyhledávání relevantních slov nebo frází. K tomu lze užít povolené metody (*white hat SEO*) anebo nepovolené techniky (*black hat SEO*). Mezi nepovolené metody patří např. *link farms* (síť propojených webů, které na sebe vzájemně odkazují), *doorway* (stránka je sice naplněná slovy a frázemi spojenými s hledaným tématem, ale neobsahuje pro uživatele žádné hodnotné informace), neviditelný text. Jejich cílem je zvýšit podvodně Google page rank webu a tím se dostat na vyšší pozice při vyhledávání. Všechny vyhledávače se snaží proti těmto metodám bojovat

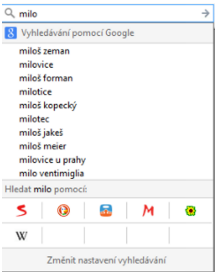
a pokud jsou odhaleny, tak dojde k penalizaci stránek, které těchto metod využívají [8]. Pro zajímavost dodejme, že SeznamBot indexuje pouze stránky napsané v českém jazyce [9], výsledky ostatních jazyků přejímá z jiných vyhledávačů. Googlebot používá více než 200 parametrů, které mu pomáhají rozhodnout, které stránky obsahují vámi hledanou informaci [10]. Hodnota GPR není zárukou, že stránka bude na prvních pozicích při vyhledávání, ale je to jen jeden z řady parametrů, které ovlivňují pořadí ve výsledcích. Přesný seznam kritérií a jejich vliv na výslednou váhu webu je přísně střeženým tajemstvím. Jednotlivé vyhledávače se stále vyvíjejí, a tudíž se počet a váha jednotlivých faktorů mění [11, 12].

V **třetím kroku** dojde k zpřístupnění výsledků uživatelům, kteří mohou zadávat dotazy do vyhledávače. Po zadání dotazu vyhledávač zobrazí SERP (search engine result page), což není nic jiného než stránka výsledků. Na SERP lze najít přirozené výsledky, které zobrazí vyhledávač, ale též reklamní odkazy (sponzorované odkazy) [11]. Při zadání dotazu vyhledávač prochází pouze své databáze, tudíž je schopen poskytnout výsledky v řádu desetin sekundy. Výsledky pro často vyhledávané fráze a slova pak vyhledávače mohou sestavit pouze jednou, a pak je na stejné dotazy pouze zobrazí. Tento seznam výsledků se aktualizuje několikrát denně.

V následujících třech kapitolách se seznámíme v krátkosti s Google Trends, Zeitgeist a poté s prací se dvěma nejznámějšími vyhledávači v České republice (Google a Seznam).

Google Trends a Zeitgeist

Témata a hesla vyhledávaná prostřednictvím vyhledávačů jsou samozřejmě sledována a zaznamenávána mj. za účelem analýzy zdrojů, sledování trendů a statistiky. Vyhledávače sledují, která slova jsou uživateli vyhledávána a na které stránky ze SERPu následně uživatel vstoupí. V minulosti zadané dotazy jsou cenným zdrojem pro našeptávání dotazů uživatelům. Obr. 1 ukazuje příklad našeptání dotazu, kdy stačí do příslušného pole prohlížeče Firefox zadat např. Miloš a prohlížeč prostřednictvím vyhledávače nabídne možná hledaná témata. Pokud uživatel najde v nabízeném seznamu hledaný dotaz, tak může přímo kliknutím dokončit dotaz a nemusí ho ručně celý vypisovat.



Obr. 1 Našeptávač v prohlížeči Firefox

Některé vyhledávače (např. Seznam, Google) sestavují každoročně žebříčky nejvyhledávanějších dotazů v kategoriích a komentují změny oproti

minulým letům. Přehled hledaných termínů od společnosti Google.com se jmenuje Zeitgeist (lze přeložit jako Duch doby) a lze ho nalézt na adrese http://www.google.cz/trends/2014/. Společnost Seznam.cz má svůj žebříček za kompletní rok na adrese http://skokani.seznam.cz/?report=2014. Pokud nás zajímají aktuálně hledané termíny, tak např. Google používá http://www.google.cz/trends. Po zadání dotazu lze sledovat z jakých lokalit byl vyhledáván, kdy a jak se vyvíjí míra hledání. Termín Vánoce je hledaný pravidelně v období prosince, naopak Velikonoce jsou pohyblivým svátkem, tudíž na grafu vyhledávání tohoto slova v jednotlivých letech je jeho vrchol v různých měsících.

Práce s vyhledávačem Google

Otevřete svůj oblíbený internetový prohlížeč (Firefox, Internet Explorer, Chrome, Opera atd.) a do adresního řádku napište www.google.cz a potvrďte stisknutím klávesy ENTER. Dostanete se na úvodní stránku, která vypadá jako na obr. 2.

Obr. 2 Úvodní stránka Googlu

Pro zadávání dotazů budeme využívat textové pole uprostřed obrazovky. Abychom mohli efektivně hledat informace na internetu, je nutné znát pravidla vyhledávače a všechny možnosti, jež jsou pro daný vyhledávač k dispozici. V následujícím souhrnu najdete nejdůležitější pravidla a omezení:

1. Google nerozlišuje mezi velkými a malými písmeny v dotazu, tudíž dotazy ve tvaru: král, KRÁL, Král nám dají stejné výsledky.

2. Interpunkce v dotazu se ignoruje až na dvě výjimky, a to apostrof a pomlčku. Pomlčka se zpracovává odlišně od ostatní interpunkce a je to zřejmé na dotazu ping-pong. Výsledky budou odpovídat těmto třem dotazů dohromady: ping pong, ping-pong a pingpong.

3. Při zadávání dotazu je nutné mít na paměti slova, která se nejpravděpodobněji budou nalézat na hledané stránce. Když se budeme zajímat o aktuální cenu rohlíku tak dotaz: rohlík cena, zajistí, že mezi prvními výsledky budou letáky hypermarketů.

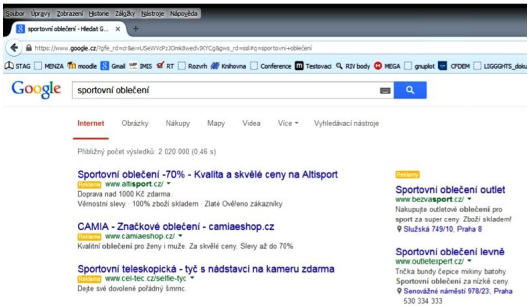
4. Google využívá celou paletu operátorů. Operátor je speciální výraz jímž ovliňujeme vyhledávání. Může mít podobu jednoho znaku či slova. Seznam a použití nejužívanějších operátorů lze nalézt v tabulce 1.

5. Operátory lze řetězit a tím blíže specifikovat dotaz. Pokud se zajímáme o jaguára jako o zvíře, tak nás nezajímají stránky o voze jaguár a můžeme tak v dotazu tyto stránky předem vyloučit: jaguár rychlost -auto [13].

Tabulka 1 Přehled nejpoužívanějších operátorů pro vyhledávání v Google [14, 15]

Operátor	Použití	Příklad
+	Hledá stránky Google+ nebo krevní skupiny.	AB+
@	Hledá označení na sociálních sítích.	@agoogler
$	Hledá ceny (vyhledává ceny jen v dolarech).	nikon $400
#	Hledá populární témata podle tzv. hashtagu.	#throwbackthursday
-	Když je použita před slovem nebo dalším operátorem, tak z výsledků vyloučí weby, které dané údaje obsahují. Hlavní využití je u slov, které mají více významů.	jaguár rychlost -auto
“”	Hledá se na stránkách výraz v přesném tvaru, který je uveden v uvozovkách.	„královno temných spádů“
*	Slouží jako zástupný znak pro neznámé výrazy.	„lepší * v hrsti než * na střeše“
..	Užívá se mezi dvěma čísly bez mezer. Výsledky jsou v rozsahu čísel, jež tento operátor obklopují.	fotoaparát 2 000..5 000 Kč
site:	Hledá na konkrétních webech či doménách.	studium site:ujep.cz
link:	Hledá weby, které odkazují na určitý web.	link:youtube.com
related:	Hledá weby podobné zadané webové adrese, kterou již znáte.	related: aktualne.centrum.cz
OR	Hledá weby, kde se vyskytuje jedno ze slov před a za operátorem.	maraton OR závod
AND	Hledá weby, kde jsou obě slova spojená tímto operátorem.	kočka AND pes
info:	Hledá informace o webové adrese, včetně archivované verze stránky, podobných stránek a stránek, které na daný web odkazují.	info:google.com
intitle:	Hledá klíčové slovo pouze v titulku.	intitle:source
allintitle:	Hledá celý výraz v titulku.	allintitle:„open source“
inurl:	Hledá klíčové slovo v URL stránky.	inurl:open
allinurl:	Hledá celý výraz v URL.	allinurl: „open community“
filetype:	Hledá v souborech daného typu.	filetype:pdf
define:	Hledá definici daného slova.	define:lion

Po zadání dotazu a stisknutí klávesy ENTER dostanete stránku s jednotlivými výsledky (SERP) a řadu dalších informací, jak ukazuje obr. 3.



Obr. 3 Výsledky vyhledávání na Googlu

Podívejme se detailněji na tuto stránku. Nejprve se tam objevila lišta, kde je červeně vyznačeno, že jsme hledali v celém internetu. Máme možnosti přepnout výsledky na Obrázky, Nákupy, Mapy, Videa. Dále je tam rolovací nabídka Více, která ukrývá položky: Zprávy, Knihy a Aplikace. Zajímavější je ale položka Vyhledávací nástroje, na kterou když klikneme, zobrazí se další lišta na upřesnění vyhledávání pro Jazyk, Časový parametr, Lokalizace a Přesná shoda. Alternativou k zadání dotazu s použitím operátorů je ozubené tlačítko vpravo nahoře a poté volba Rozšíření vyhledávání. V následném formuláři je pak možné zadat dotaz bez použití operátorů.

Před vlastními výsledky můžeme naleznout informaci o přibližném počtu výsledků (2 020 000) a době hledání (0,46 s). U některých výsledků můžeme najít v žlutém oválu slovo Reklama, čímž nám Google označuje sponzorované odkazy (jsou na začátku a konci stránek). Výsledky obsahují titulek výsledku, který je zároveň odkazem na příslušnou stránku. Pokud je výsledkem vyhledávání soubor, tak je tato informace znázorněna hranatými závorkami a typem souboru. Následuje adresa URL, poté můžeme najít popis stránky, který je převzat z daných stránek [13].

Práce s vyhledávačem Seznam

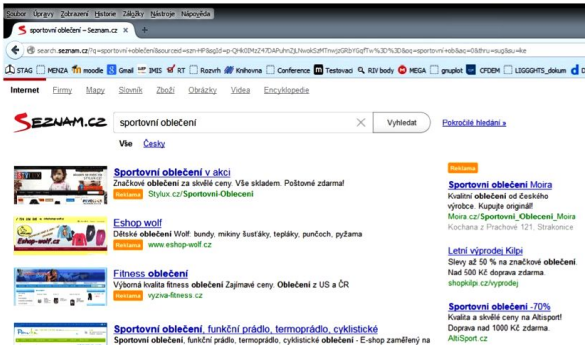
Práce s vyhledávačem Seznam je téměř totožná jako práce s Googlem. Největší rozdíl u Seznamu je ten, že indexuje primárně české stránky. Zahraniční výsledky jsou obvykle přejaty z jiného vyhledávače. Zásadní přínos vyhledávače Seznam spočívá v drobných optimalizacích určených

speciálně pro „český internet“. Pracuje s téměř stejnými operátory jako Google, ale najde se tam pár odlišností, které najdete v tabulce 2.

Tabulka 2 Přehled operátorů Seznamu, které se liší oproti operátorům Googlu [16]

Operátor	Použití	Příklad
,	Při hledání nezáleží na tom, jak daleko jsou slova od sebe v textu.	penzion, Krušné hory
intext:	Vyhledání zadaného slova přednostně v obsahu stránky.	intext:nápověda
+	Tento operátor není povinný. Používá se v případě, kdy chcete vynutit hledání určitého slova. Nalezená stránka tedy musí slovo, kterému plus předchází, obsahovat. Pokud operátor plus není zadán, chová se vyhledávač stejně, jako by zadaný byl.	čočka + recept
host:	Podobný jako site, ale na rozdíl od site nerozšiřuje hledání na subdomény.	host:seznam.cz
lang:	Omezuje vyhledávání na určitý jazyk.	cars lang:cs

Po zadání dotazu dostanete výsledky na Seznamu v jiné grafické podobě, jak se můžete přesvědčit na obr. 4.



Obr. 4 Výsledky vyhledávání na Seznamu

Prvním rozdílem oproti Googlu je zobrazení náhledu stránky u výsledku. Druhý rozdíl spočívá v pravém panelu reklamních odkazů. Třetí rozdíl je přítomnost odkazu na stránky až po popisku.

Náměty na samostatné úlohy

Zadání

1. Vytvořte dotaz, který vyhledá ve vyhledávači Google stránky k tématu Lednice, dále z výsledků vyloučí chladničky a penziony a bude se zaměřovat jen na stránky z let 2013 až 2015.

2. Zjistěte, ze které lokality je nejčastěji vyhledáván pojem Krakonoš, jak se vyhledávání vyvíjelo v čase a jaké další související vyhledávání byla prováděna?

3. Jaký výraz byl nejhledanějším na Google.com v ČR v roce 2014 v kategorii „Co je“?

Řešení

1. Do vyhledávacího pole napíšeme Lednice -chladničky -penziony a ve Vyhledávacích nástrojích omezíme interval hledání na 2013 až 2015 (vlastní časový úsek).

2. V Google Trends zadáme do vyhledávácího pole Krakonoš. Vidíme vývoj vyhledávání v čase a v dolní části i související hledání. Nejvíce hledaným je v Královehradeckém kraji a z měst v Praze.

3. Najdeme si na stránce https://www.google.cz/trends/ odkaz na žebříčky pro rok 2014. Zde je uveden jako první Instagram.

Závěr

Článek shrnul způsoby vyhledávání informací na internetu, principy fungování vyhledávačů, způsoby formulování složitějších dotazů, upozornil na žebříčky nejvyhledávanějších slov a frází.

Internetové vyhledávače mohou být velmi užitečnými a mocnými nástroji, ale je nutné zdůraznit, že všechny informace na internetu nemusí být přesné či pravdivé. K výsledkům vyhledávání je tedy nutné přistupovat s kritickým myšlením a ne s větou: „Bylo to na internetu, tak to musí být pravda.“ Pro sledování „žhavých“ témat pak vyhledávače sestavují žebříčky hledaných frází podle oblastí.

Poděkování

Děkujeme projektu „Mezioborové vazby a podpora praxe v přírodovědných a technických studijních programech UJEP“ (OPVK CZ.1.07/2.2.00/28.0296).

Literatura

[1] http://www.worldwidewebsize.com/
[2] http://www.jakpsatweb.cz/robots-txt.html
[3] https://www.interval.cz/clanky/google-sitemaps/
[4] https://support.google.com/webmasters/answer/156184?hl=cs
[5] http://pagerank.jklir.net/?p=pagerank
[6] http://www.propagacenainternetu.cz/optimalizace-webovych-stranek-on-page-faktory.
[7] http://www.corporateict.cz/koutek-redaktora/jaky-je-vyznam-off-page-faktor-na-uspnost-optimalizace-pro-vyhledavae.html.
[8] http://www.seoradce.cz/techniky-seo.html
[9] http://napoveda.seznam.cz/cz/fulltext-hledani-v-internetu/nez-zacnete-tvorit-web
[10] https://www.google.com/intl/cs_cz/insidesearch/howsearchworks/algorithms.html
[11] http://www.jakpsatweb.cz/vyhledavace.html
[12] http://n-host.cz/2014/12/jak-funguje-vyhledavac/
[13] *Iskra, J.*: Google: tipy a návody pro vyhledávač, Gmail, YouTube, Earth a další aplikace, Brno, Computer Press, 2008.
[14] https://support.google.com/websearch/answer/2466433?hl=cs
[15] http://opencommunity.cz/operatory-google
[16] http://napoveda.seznam.cz/cz/fulltext-hledani-v-internetu/pokrocile-hledani/

ZPRÁVY

57. ročník Mezinárodní matematické olympiády

57. ročník Mezinárodní matematické olympiády (IMO) se uskutečnil 6.–16. července 2017 v Hongkongu, kam se soutěž vrátila po 22 letech, poprvé do Hongkongu pod čínskou správou. Soutěže se zúčastnilo rekordních 602 studentů z rekordních 109 zemí pěti kontinentů. Nechyběli žádní tradiční účastníci, poprvé se IMO zúčastnili Egypt, Irák, Jamajka, Keňa, Laos, Madagaskar a Myanmar.

Příprava soutěže tohoto rozsahu je během na dlouhou trať. Hlavní organizátor, Hongkongský výbor pro mezinárodní matematickou olympiádu (The International Mathematical Olympiad Hong Kong Committee Limited), ve spolupráci s Hongkongskou univerzitou věd a technologií (The Hong Kong University of Science and Technology, HKUST) a Úřadem pro vzdělávání Vlády zvláštní administrativní oblasti Hongkong pracovali s rozpočtem 20 milionů hongkongských dolarů (60 milionů korun), zajišťovali ubytování a stravování pro 1200 soutěžících, organizátorů a hostů a soutěž připravovali několik let dopředu. Samotná příprava soutěžních úloh