MASARYKOVA UNIVERZITA
FAKULTA INFORMATIKY

PV070 – DIGITÁLNÍ KNIHOVNY

Technorati – searching the blog space

Bc. Petr Škoda
5. ročník

3. prosince 2008

Informace o projektu

- Název: **Technorati**
- URL: http://www.technorati.com
- Vytvořil: **Dave Sifry**
- Vlastník: Technorati, Inc.

Technorati je vyhledávač zaměřený na prohledávání blogů. Jedná se o první vyhledávač tohoto druhu, který kdy vznikl. Technorati se považuje, a také velkým množstvím subjektů je považován, za jakousi bránu do světa blogů. Umožňuje blogerům dát o sobě vědět a stejně tak ostatním zprostředkovává vždy to nejnovější o čem lidé píší na svých blozích. Dříve nevídaným přístupem posouvá vyhledávání do nové aktuálnější úrovně. Příspěvek, který byl před pár minutami napsán může být právě teď někým nalezen a komentován. Díky této interaktivitě je možné hledat zcela aktuální témata.

Vývoj Technorati začal v roce 2002 David Sifry jako „jednomužný" projekt s cílem vytvořit vyhledávač ušitý na míru dynamickému světu blogů, který byl v té době v USA na obrovském vzestupu. První verze Technorati byla spuštěna již v listopadu 2002 a od té doby se poskytované služby neustále rozšiřují o nové. Technorati do dnešní doby vyrostlo v plnohodnotnou mediální společnost poskytující služby a nástroje v oblasti blogů a sociálních médií.

Cíle projektu

Vytvořit vyhledávač, který bude schopen vyhledávat v oblasti blogů, protože klasické vyhledávače typu Google si s tímto fenoménem nedokáží poradit. Umožnit komunitě blogerů ještě větší semknutí, snazší vzájemné hodnocení a pohodlnější přístup k informacím ostatních. Rozvíjet „blogging" a sociální média.

Proč Yahoo nestačí?

Povaha klasického vyhledávání a hodnocení relevance odkazů ve všech dnešních Internetových vyhledávačích je založena na tom, že stránka na určité téma je právě tak kvalitní, kolik na ni odkazuje jiných stránek. Přičemž hodnocení je vyšší, pokud na stránku odkazují jiné stránky s vysokým hodnocením pro dané téma. Toto kritérium je výborné a osvědčilo se téměř bez výjimek. Proto všechny dnešní vyhledávače pracují na tomto principu, který byl objeven Googlem. Zajímavé je, že kritérium odkazujících stránek je výborné i pro blogy. V čem je tedy problém? Z vlastní zkušenosti vím, že při vyhledávání klasickými vyhledávači narážím na výsledky z blogů jen výjimečně. Obvykle jsou to odkazy na články steré i několik měsíců. A to je ten problém.

Povaha blogu je jiná, než běžného webu. Blog je dynamický a rychlý. Běžný web oproti tomu spíše usedlý a všichni počítají s tím, že jejich stránka se nestane slavnou, uznávanou a hojně navštěvovanou přes noc natož během pár hodin. Oproti tomu nejlépe hodnocení blo-

geři přispívají do svých blogů dvakrát až desetkrát denně! A se stejnou dynamikou musí také probíhat objevování těchto nových příspěvků. Běžnými postupy může jeden člověk objevit ty nejnovější příspěvky jen v okruhu blogů, které navštěvuje pravidelně (i vícekrát denně). Maximálně ještě může hledat témata sledováním určitého okruhu blogů pomocí RSS. Problém je, že tímto způsobem jen zřídka objevíme nové, dosud neznámé blogy a nejnovější příspěvky na nich. Přesto mohou tyto nenalezené příspěvky pro nás být v danou chvíli zcela relevantní a zajímavé. Nové blogy objevujeme pouze pomocí odkazů, které někdo uvede na námi navštěvovaném blogu. A tak spektrum blogů sledovaných jedním uživatelem je, a vždy bude, omezené. Bez kvalitního vyhledávače nám prostě chybí globálnější pohled na aktuální témata.

Kamenem úrazu pro Google-like vyhledávače je stále omílaná dynamika. Web totiž prohledávají sami od sebe tím, že jejich roboti (*crowlery*) tak často jak to jen jde navštěvují a indexují všechny stránky na Internetu. Tito roboti jsou vybaveni jistou inteligencí a sami umí odhadnout, které weby je třeba prohledávat týdně (protože se na nich nic nemění) a které je nezbytné prohledávat co nejčastěji (protože se jedná o uznávaný zpravodajský server). Problém však zůstává. Touto cestou není možné indexovat každý z několika milionů blogů čtyřikrát a dokonce ani dvakrát denně. Avšak ani větší četnost prohledávání by nevedla k tak dobrým výsledkům, jako má řešení od Technorati.

Kouzlo prostého rozšíření – ping

Takže proč vlastně používat automatizované roboty, aby se snažili objevit kde se stalo co nového, když místo toho o sobě může každý dát vědět? Původně byly roboti nasazeni právě proto, že bylo zřejmé, že existuje velká část Internetu, kterou nikdy nikdo nepřihlásí do vyhledávačů. Proti tomu snaha každého vyhledávače byla a je najít vše, co najít lze. Ve světě blogů však tolik nelpíme na tom, aby bylo indexováno vše, ale naopak chceme mít co nejvíce co nejaktuálnějších informací, proto vznikl takzvaný *ping*. Ping se postará o to, aby informoval vyhledávač o změnách na konkrétní stránce (blogu).

Vynález pingu (nejedná se teď o síťovou službu ICMP echo známou také jako ping) překvapivě nepatří Davidu Sifry ani Technorati jako společnosti. Před Davidem Sifry chtěl umožnit blogerům sledovat co nejjednodušeji novinky na blozích přátel Dave Winer, který vytvořil jeden z prvních programů pro blogování jménem *Radio Userland*. Tento program umožňoval pomocí nástroje *ping* informovat portál weblogs.com o každém novém příspěvku na blogu jednotlivých uživatelů. Weblogs.com pak tuto informci poskytuje čtenářům, kteří mohou novinky na blozích sledovat třeba pomocí RSS.

Sifry se svým nápadem tedy nejprve naprogramoval indexovacího robota, který stahoval informace o nových úpravách blogů z weblogs.com, ty „prolezl“ a hned indexoval. Právě to byly základy pozdějšího Technorati. Nedlouho po těchto prvních pokusech Sifry oslovil tvůrce různých blogovacích programů s nabídkou, že poud umožní ve svých produktech automaticky pingovat mimo weblogs.com i Technorati, budou blogy v jejich programech velice rychle zaindexovány a připraveny k nalezení jinými uživateli. Toto byla

samozřejmě pro tvůrce blogovacích nástrojů výbrná reklama a proto pingy hojně podporovali a podporují.

Technologie

Technorati funguje převážně na otevřeném software, základem je operační systém Linux, databáze MySQL a skriptovací jazyky PHP a Perl. Toto řešení se často nazývá LAMP. Aktuálně Technorati Media poskytuje i další služby mimo vyhledávání ve světě blogů. Jedná se například o systém umožňující blogerům vydělávat zobrazováním reklam.

Do technologie patří i čím dál sofistikovanější metody na potlačování nekalých aktivit, které se samozřejmě neustále snaží protlačit nerelevantní informace na první příčky vyhledávání, vytvářením takzvaných spam blogů (sblog), které na sebe různě odkazují a v některých tématech pak mohou mít vysoká hodnocení. Těmto pokusům se naštěstí daří zabraňovat.

A teď ještě jedno shrnutí celého principu vyhledávání v blozích. Kvalita stránky, takzvané *Authority*, je odvozeno z počtu odkazů v jiných blozích na jeden konkrétní příspěvek. Počet odkazů se obvykle pohybuje mezi 0 a 20, nejlepší příspěvky mají více než 100 na sebe odkazujících blogů. Technorati prohledává jen blogy (detekuje a odstraňuje falešně přidané stránky, které nejsou blogy), díky tomu je možné, aby se na horních příčkách výsledků vyhledávání objevovaly i stránky s malým počtem odkazů. Nejdůležitějším faktorem však je, že Technorati pomocí pingu indexuje nové příspěvky ihned po jejich vytvoření a tak už v první hodině po zvolení nového prezidenta USA můžeme zjistit, co si o tom lidé myslí.

Dosažené výsledky

Technorati lze považovat za velice úspěšný projekt. Po 6ti letech působení a postupného vylepšovaní je Technorati uznáváno za nejlepší vyhledávač ve světě blogů (*blogspace*). Bohužel v některých ohledech je z Technorati cítit jistá arogance [1] – „Svět blogů je Technorati, Technorati určuje kam bude blogging směřovat…“

Technorati hlásí v letošní zprávě o stavu světa blogů (*State of the Blogosphere*) 133 miliónů zaindexovaných příspěvků do blogů od svého vzniku v roce 2002. Celá letošní zpráva je k vidění v [2], ovšem doporučuji prostudovat i krátký kritický článek o této zprávě v [3] a případně i zprávy z minulých let. Čísla naznačují, že Technorati pravděpodobně úspěšně vyhledává ve většině blogů na Internetu.

Vlastní zhodnocení

V předchozím textu jsem Technorati vylíčil jako vesměs úžasný nástroj. Bohužel pravdou je, že celá tato sláva má i své stinné stránky. Už jen fakt, že na Wikipedii je minimum informací o technologiích a úspěších Technorati, naznačuje, že není vše jak by mělo být. Místo

těchto informací se na Wikipedii dočteme o mnoha problémech, které doprovázely (a některé ještě dnes doprovázejí) tento projekt.

Hlavní výtkou většiny oponentů je problematický postoj Technorati ke službě MySpace.com. MySpace je totiž nejen svými uživateli, ale i mnoha odborníky, považováno mimo jiné i za zprostředkovatele blogů. Technorati však MySpace neindexuje, jelikož podle přesvědčení mnoha non-myspace blogerů nejsou na MySpace blogy, ale jen jakési „příspěvky“ [4]. Další část zastánců neindexování MySpace zase tvrdí, že na MySpace je příliš mnoho nezletilých a navíc ještě velké množství nezletilých vydávajících se za zletilé. Z mého pohledu (nejsem vůbec bloger) jsou tyto argumenty nesmyslné. Pokud má MySpace alespoň z části povahu blogu, měl by tak být brán a indexován jako ostatní. Argument o mladosti lidí na MySpace je už vůbec scestný, protože říci, že něco není blog, protože to bylo napsáno člověkem, kterému nebylo 18 let, je dle mého názoru hodné hlupáka [5].

Musím také ještě jednou připomenout, že Technorati se chová značně arogantně a mnohdy své dobré výsledky zastíní silnými komerčními tlaky [1]. Informace poskytované Technorati jsou sice hodnotné, ale musíme na ně nahlížet s kritickým pohledem. V občasných zprávách o stavu světa blogů (*State of the Blogosphere*) se dočteme zajímavé informace, ale některé pravděpodobně jsou zkreslené [1]. Stejně tak vyhledávání v blozích je výborné, ale je třeba vědět, že nevyhledáváme v celém světě blogů, ale jen ve vybraných částech (seznam prohledávaných poskytovatelů blogů [6]). Přesto je Technorati jedním z hlavních podporovatelů blogů a výborným nástrojem pro vyhledávání informací a aktuálních témat.

Použité zdroje

[1] Information Architects Japan. *Technorati: Big Business with Bogus Data* [online]. 2006– . [cit. 2008-12-04]. < http://informationarchitects.jp/bogus-technoratiedelman-statistics/ >.

[2] Technorati. *State of the Blogosphere / 2008* [online]. 2008– . [cit. 2008-12-04]. < http://www.technorati.com/blogging/state-of-the-blogosphere >.

[3] KIRKPATRICK, Marshall. *State of the Blogosphere 2008: Technorati Numbers Indicate Blogging Is Niche and Slowing* [online]. 2008-09-22– . [cit. 2008-12-04]. < http://www.readwriteweb.com/archives/state_of_the_blogosphere_2008.php >.

[4] BRAZELL, Aaron. *Technorati Indexing MySpace Blogs* [online]. 2006-03-31– . [cit. 2008-12-04]. < http://technosailor.com/2006/03/31/technorati-indexing-myspace-blogs/ >.

[5] mobilejones. *The Site that ate the Blogosphere* [online]. 2006-02-17– . [cit. 2008-12-04]. < http://www.blogher.com/node/2509 >.

[6] Technorati Media. *Technorati Network* [online]. 2008– . [cit. 2008-12-04]. < http://technoratimedia.com/network/ >.

Dublin Core metadata

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Title" content="Technorati – searching the blog
   space" />
<meta name="DC.Creator" content="Bc. Petr Škoda" />
<meta name="DC.Subject" content="Technorati" />
<meta name="DC.Subject" content="blog" />
<meta name="DC.Subject" content="blogosphere" />
<meta name="DC.Subject" content="search" />
<meta name="DC.Description" content="Práce popisuje projekt
   Technorati, jehož cílem je indexovat blogy a umožnit vyhledávání
   ve světě blogů. Práce shrnuje cíle projektu, stav projektu,
   technologie, úspěchy a neúspěchy. V poslední části práce je také
   krátké shrnutí různých názorů hovořících proti kvalitě projektu
   Technorati a proti samotné společnosti Technorati, Inc." />
<meta name="DC.Date" content="4.12.2008" />
<meta name="DC.Type" content="text" />
<meta name="DC.Type" content="esej" />
<meta name="DC.Format" content="application/pdf" />
<meta name="DC.Format" content="computerFile" />
<meta name="DC.Identifier"
   content="http://www.pecashome.cz/uri/PV070-technorati-esej.pdf"
   />
<meta name="DC.Source" scheme="URL"
   content="http://www.guardian.co.uk/technology/2006/feb/16/newmedi
   a.weblogs" />
<meta name="DC.Source" scheme="URL"
   content="http://www.blogher.com/node/2509" />
<meta name="DC.Source" scheme="URL"
   content="http://informationarchitects.jp/bogus-technoratiedelman-
   statistics/" />
<meta name="DC.Source" scheme="URL"
   content="http://www.readwriteweb.com/archives/state_of_the_blogos
   phere_2008.php" />
<meta name="DC.Coverage" content="FI:PV070 Digitální knihovny" />
<meta name="DC.Language" content="cze" />
```