Fulltext je mrtev, ať žije Googlebot

Almad

Vyhledávání fulltextem je nezbytnost každého webu. Proč tedy není a ani nebude?

Obsah

Fulltext: Základ webu

Fulltextové vyhledávání je bezpochyby jednou ze základních vlastností informačního webu. Po pravdě řečeno si myslím, že to je jedna z největších nevýhod současné verze - totiž, že není možné opravdové vyhledávání v příspěvcích. Nebudu vysvětlovat, proč není ve verzi současné, místo toho se pokusím objasnit, proč přes výše uvedené prohlášení ani nebude.

Druhy vyhledávání

Existují dva základní druhy vyhledávání: klasické a full-textové (bohužel neznám žádný český ekvivalent tohoto slova, ví-li někdo, ať se mi prosím ozve). Klasické neobsahuje mnoho možností - prostě vezme zadaný řetězec a prohledá databázi, zda na její *přesný* výskyt. Naproti tomu vyhledávání fulltextové má mnoho dalších možností: dokáže používat logické operátory (AND, OR, NOT), tím pádem dokáže vyhledat i výskyty jednotlivách slov, nejen celého řetězce, dokonce jsou schopny rozeznat i přechylování (čili kromě výskytu slova "auto" vyhledá i "auta", "autu", atd.). Na drtivě většině stránek funguje pouze vyhledávání klasické, neboť je asi vidět, že fulltextové vyhledávání je velmi náročné na kód (algoritmy opravdu nejsou jednoduchou záležitostí), ale i na výkon počítače.

Proto jsem se rozhodl, že výjimečně nebudu prorážet hlavou zeď, ale projdu brankou.

DocBook a Googlebot

Ještě bych nastínil jeden problém. Pokud má server sloužit jako centrála informací, pak je nutné, aby se na informace v něm uložené bylo možné jednoduše odkazovat, což je v současné verzi dosti problematické (URL typu dracidoupe.cz?rub=clanek&id=123 nepatří mezi nejhezčí) - a v nové verzi by to bylo ještě horší. Krom toho je možnost omylem zkopírovat vaše session id, což je malinká bezpečnostní chybka) - a v nové verzi by to bylo ještě horší (adresy budou mít ještě ošklivější formát).

Jak jsem již psal dříve, články v nové verzi budou uchovávány ve formátu DocBook. Ten umožňuje jednoduchý export např. do xhtml, ovšem formátovaného dle předepsaných direktiv (tohoto kroku budete ušetřeni, provádět ho bude server, takže netřeba se děsit). Proto jsem se rozhodl následujícím způsobem.

Pro každý článek bude vygenerována externí HTML a PDF verze, která bude umístěna na pěkné, trvalé a veřejné adrese (něco jako http://www.dracidoupe.cz/export/clanky/20/index.html), na kterou bude u článků umístěn odkaz. Na tuto adresu se budete moci libovolně odkazovat, když se budete chtít pochlubit svými dílky. Krom toho bude mnohem lépe dostupná indexovatelná fulltextovými internetovými prohlížečí, zejména tím asi nejlepším - Googlebotem. Krom toho, že stránka bude mít zřejmě poměrně vysoké hodnocení, podporuje google direktivu site:dracidoupe.cz, což povede k vyhledávání pouze na našich stránkách - a vlastně tak bude dostupný i ten fulltext.

Ukázku toho, jak takový (zatím nezformátovaný, tuhle část jsem ještě neudělal ;o)) export bude vypadat můžete vidět třeba na mé oblíbené nestvůrce "Hvězdička noční" od Errica [HTML [http://v2.dracidoupe.cz/fulltext-nebude/hvezdicka/index.html]] [PDF [http://v2.dracidoupe.cz/fulltext-nebude/hvezdicka/hvezdicka.pdf]] (mimochodem, převod z html do docbooku byl záležiotstí cca 10 -15 min. včetně exportu).