

Kľúčové slová v kontexte webového vyhľadávania

Pri vyhľadávaní informácií sa spočiatku webové prehliadače zameriavali hlavne na meta informácie v hlavičke HTML dokumentov. Slúžili k tomu hlavne meta elementy pre názov stránky, kľúčové slová a popis. Táto metóda sa však ukázala nespoľahlivá. Stačilo si do spomenutých objektov slová, pre ktoré chcela webová stránka dosahovať dobre výsledky a takto sa dali výsledok hľadania zmanipulovať. Aj z toho dôvodu od tejto metódy takmer všetky majoritné vyhľadávače odstúpili a prešli na metódy strojového učenia, ktoré sú schopné poskytnúť dostatočne presné výsledky.

Význam kľúčových slov sa tak presunul skôr na symbolickú úroveň, avšak nejaké (aj keď slabé) opodstatnenie si ešte zanechali. V súčasnosti sa využívajú v takmer všetkých vedeckých prácach a článkoch a niektorých biznis plátkoch (na vyhľadávanie podobných záverečných prác využíva kľúčové slová napr. aj informačný systém MU, ale výsledky sú nepostačujúce).

Moja myšlienka je, že pokiaľ sa KW zvolia vhodne, môžu priniesť spresnenie vyhľadávacieho procesu. Preto som začal pátrať po metodike extrahovania kľúčových slov z textu pomocou algoritmov a narazil som na oblasť extrahovania informácií (Information Extraction) nazvanú extrahovanie kľúčových slov (Key words minig). Informácie v týchto odsekoch čerpám zo zdrojov [1], [2] a [3]. Každý z nich sa venuje inej metodike a všetky vykázali slušné výsledky. Najviac ma zaujala metóda extrahovania pomocou quad-gramov, čo je v podstate N-gram o dĺžke 4. Zaujímavý je aj fakt, že kľúčové slová boli úspešne vytiahnuté v rôznych doménach (napr. biologické dáta, hovorová reč).

Zdroje:

[1] Ingrid Renz, Andrea Ficzy, Holger Hitzler: Keyword Extraction for Text Characterization.

[online] <http://subs.emis.de/LNI/Proceedings/Proceedings29/GI-Proceedings.29-19.pdf>

[2] Ying Liua, Martin Brandona, Shamkant Navathea, Ray Dingedinec, Brian J. Ciliac: Text Mining Functional Keywords Associated with Genes. 2004. [online]

http://www.cc.gatech.edu/classes/AY2005/cs4440_fall/4266Liu.pdf

[3] Josh Froelich, Sergei Ananyan, David L. Olson: The Use of Text Mining to Analyze Public Input. [online] <http://www.textminingnews.com/whitepapers/Megaputer%20-%20Text%20Mining%20for%20analyzing%20public%20input.pdf>