Dokáže obsah HTML tagov zlepšiť výsledky ATR algoritmov?

MILAN LUČANSKÝ

Vedúci projektu
Ing. Marián Šimko

Cieľ projektu

Získať relevantné kľúčové slová z webovej domény

V skratke, ako by to mohlo fungovať...

… a teraz trochu podrobnejšie

Máme dokumenty, ale ako získať kľúčové slová?

- dolovaním v texte, pomocou ATR algoritmov

Relatívne úspešný spôsob, záleží však od kolekcie dokumentov a zvolených algoritmov.

Aj web stránky sú iba dokumenty

A je možné ich spracovávať ATR algoritmami, avšak musíme sa zbaviť nepotrebného obsahu.

Nie však všetky HTML tagy, pretože ...

… tagy v sebe skrývajú sémanticky zaujímavý obsah

Pán Hodgson skúmal stránky o logickom programovaní.

Zameral sa na text vo vnútri týchto tagov:

<a>, **<h1..6>**, **<meta>** a **<!-- -->**

Zistil, že 61% stránok obsahovalo v texte hypertextových odkazov sémanticky užitočný obsah.

Hodgson, J.: Do HTML Tags Flag Semantic Content? Published by the IEEE Computer Society, Vol. 5, January-February 2001, No. 1, pp. 20-25.

Searching Engine Optimization

Vyhľadávače vedia, ktoré tagy sú ako významné z hľadiska obsahu, napr.

<a> *užitočný obsah* ****

<h1> *Užitočný obsah* **</h1>**

<title> Užitočný obsah </title>

The Voting Algorithm[1]

Spája niekoľko ATR algoritmov, aby dostal čo najlepšie výsledky.

$$váha = \sum_{i}^{n} \frac{1}{R_i(k)} \cdot w_i \qquad w_i = \frac{P_i}{\sum_{i}^{n} P_i}$$

n – počet ATR algoritmov

$R_i(k)$ – skóre priradené algoritmom

w_i - „hodnovernosť“ algoritmu i

k – kandidát na kľúčové slovo

P_i – presnosť (angl. Precision)

i-teho algoritmu

[1] Z článku: A Comparative Evaluation of Term Recognition Algorithms. Department of Computer Science, University of Sheffield, 2008

Porovnanie úspešnosti ATR algoritmov

Výsledky nad Wikipédia korpusom pre 300 kandidátov pri porovnaním s hlasovacím algoritmom.

Rozhodca	TF-IDF	Weirdness	C-value	Glossex	Termex	Hlasovací
1.	0,57	0,81	0,55	0,86	0,9	0,95
2.	0,59	0,8	0,56	0,88	0,86	0,96
3.	0,64	0,77	0,61	0,88	0,9	0,97

Presnosť nad GENIA korpusom pre N kandidátov pri porovnaní s hlasovacím algoritmom.

N	TF-IDF	Weirdness	C-value	Glossex	Termex	Hlasovací
100	0,9	0,48	0,91	1	0,92	0,97
1 000	0,82	0,55	0,91	0,82	0,75	0,86
5 000	0,8	0,58	0,83	0,69	0,62	0,81
10 000	0,75	0,58	0,68	0,66	0,61	0,73
20 000	0,6	0,56	0,58	0,59	0,55	0,62

Z článku: A Comparative Evaluation of Term Recognition Algorithms. Department of Computer Science, University of Sheffield, 2008

Upravený „hlasovací“ algoritmus

Slová, ktoré nájde „hlasovací“ algoritmus, a ktoré sa budú nachádzať aj v niektorom z vyššie spomenutých tagov dostanú vyššiu váhu.

$$váha = \sum_{i}^{n} \frac{1}{R(k_i)} \cdot w_i \cdot TagIndex$$

TagIndex – číslo z intervalu <1 ; 2)

Ak sa slovo nájdené niektorým z ATR algoritmov nachádza aj v tagoch, zvýši sa jeho váha.

Hodnoty TagI*ndex*-u

TagyIndex pre jednotlivé tagy by mohl mať tieto hodnoty:

- Slová v **<title>** tagu – 1,66
- Slová v **<a>** tagu – 1,47
- Slová v **<h>** tagu – 1,35

Toto sú iba predbežné hodnoty. Bude potrebné ich experimentálne upresniť.

---

http://www.seomoz.org/article/search-ranking-factors#ranking-factors

Implementácia

Java platforma

Chcem využiť existujúce riešenia

- Nutch – crawling, HTML parsing, indexing
- knižnica ATR algoritmov dostupných na webe (Sheffield University)

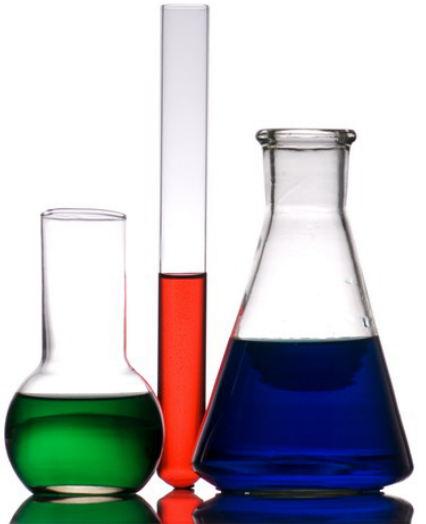
Experiment

Dátová vzorka 1 052 rôznych zvierat z en.wikipedia.org

- neobsahuje zoznam kľúčových slov, voči ktorým by sa dal výsledok porovnať
- problém s určením úspešnosti algoritmov

Náhodne vyberiem domény s anglickým obsahom

- overím získané kľúčové slová



Výsledok projektu

Overenie metódy, čo zahrňuje:

- stiahnutie stránok v rámci zadanej domény
- získanie textu zo stránok (odstráni reklamu, navigačné menu, atď.)
- získanie ováhovaných kandidátov na kľúčové slová spojením viacerých ATR algoritmov pomocou „hlasovacieho algoritmu“
- upravenie váh kandidátov na základe Tag*Indexu*
- vyhodnotenie relevantnosti kľúčových slov