PA164

Mining the Web to Determine Similarity Between Short Text Snippets

autor
Bc. Jiří Kremser

Původ

- Autoři - Mehran Sahami a Timothy D. Heilman
- Článek dostupný z
  http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.78.7807
- Prezentovaný na

  19th International FLAIRS Conference (FLAIRS), 2006.

  a

  15th International World Wide Web Conference (WWW), 2006.

- Google Inc.
  1600 Amphitheatre Parkway
  Mountain View, CA 94043 USA

Motivace

- Sémantická a syntaktická podobnost mezi krátkými texty
- ”Václav Klaus”, ”prezident ČR” a ”Santa Claus”
- Kosinová podobnost
- Podobnost založena na doménové znalosti, asociacích a jiných faktorech, které nejsou explicitně zmíněny.
- Ontologie a Word net (verze 3.0 ma 155 287 slov)
- Statické a centralizované řešení X concept drift
- Jasně definované vztahy ”být hypernymem/holonymem”

Metoda

1. Issue x as a query to a search engine S.
2. Let $R(x)$ be the set of (at most) n retrieved documents $d_1, d_2, \ldots, d_n$
3. Compute the TFIDF term vector v_i for each document $d_i \in R(x)$
4. Truncate each vector v_i to include its m highest weighted terms
5. Let $C(x)$ be the centroid of the L_2 normalized vectors v_i:

$$C(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{v_i}{\|v_i\|_2}$$

6. Let $QE(x)$ be the L_2 normalization of the centroid $C(x)$:

$$QE(x) = \frac{C(x)}{\|C(x)\|_2}$$

TF-IDF jednoho termu

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i})$$

L2 norma

$$L_1(B,R) = \frac{|B \cap R|}{|B| \cdot |R|}$$

Metoda

- Vzdálenost dvou útržků textu je

$$K(x,y) = QE(x).QE(y) \quad [0,1]$$

- Výhodou je míra konfidence výsledku
- K je kernel funkce => SVM
- Využití funkce QE ve vyhledávačích pro expanzi dotazu
- Reaguje perfektně na ”concept drift”

Výsledky

Parametry:

n … neznámé (konkrétné ”large amount”)

m = 50

K(”Microsoft CEO”,”Steve Ballmer”) = 0.838

K(”Microsoft CEO”,”Bill Gates”) = 0.317

K(”artificial intelligence”,”AI”) = 0.831

K(”UN Secretary-General”,”George W. Bush”) = 0.110

...

více v článku

Shrnutí

- Sémantická podobnost není ”exact match” ale spíše fuzzy míra
- Algoritmus je závislý na relevantních dotazech
- Výsledky jsou stále aktuální a nezastarávají jako v ontologiích
- Možné využití i pro ”našeptávač”, ačkoli to podporuje long tail fenomén

Konec

Děkuji za pozornost

Dotazy