Zkušenosti ze stáže – El-Paso

Lukáš Havrlant

DEPARTMENT OF COMPUTER SCIENCE
PALACKÝ UNIVERSITY, OLOMOUC

esf evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Základní informace o stáži

- Datum konání stáže: 2. 5. 2014–25.5.2014
- Navštívené pracoviště: UTEP, El Paso, USA
- Zahraniční garant: prof. Vladik Kreinovich

Informace o univerzitě

- University of Texas at El Paso byla založena v roce 1914
- V současnosti ji navštěvuje 23 003 studentů, z toho 3 532 doktorandů a postdoců.
- Součástí univerzity je fakulta College of Engineering.
- Ta dále obsahuje 8 kateder, mezi nimi i katedru Computer Science, na které jsem byl.
- Nejspíše nejvýznamnější výzkumná skupina je *Army High Performance Computing Research Center*

Průběh stáže

- Přednesl jsem přednášku na téma *Search Engine Based on Formal Concept Analysis*.
- S prof. Kreinovichech jsme napsali článek *A Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (tf-idf) Heuristic (and Variations Motivated by This Explanation)*, ve kterém jsme se pokusili nalézt teoretické vysvětlení principu tf-idf algoritmu.
- Článek byl odeslán do časopisu *International journal of general systems*.

Popis tf-idf algoritmu

- Algoritmus hodnotí význam slov v dokumentu v dané sadě dokumentů.
- Pomocí tfidf algoritmu lze seřadit dokumentu od nejvíce po nejméně relevantní vzhledem k nějakému dotazu od uživatele.
- Popis algoritmu: mějme sadu D dokumentů a funkce:
  - $\mathrm{tf}(t, d)$ – *term frequency*: vrací počet slov t v dokumentu d.
  - $\mathrm{df}(t)$ – *document frequency*: vrací počet dokumentů, které obsahují slovo t.
  - $\mathrm{idf}(t)$ – *inverse document frequency*: $\mathrm{idf}(t) = \log(|D| / \mathrm{df}(t))$.
- Funkci $\mathrm{tfidf}(t, d)$ definujeme jako

$$\mathrm{tfidf}(t, d) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t)$$

Popis tf-idf algoritmu #2

- Vyšší hodnota $\mathrm{tfidf}(t, d)$ funkce znamená, že slovo t je v dokumentu d významnější.

- Pokud položíme dotaz q, kterému odpovídají dokumenty

$$D_q = \{d_1, \ldots, d_n\} \subseteq D,$$

  můžeme je seřadit od nejvíce po nejméně relevantní dokumenty $\langle d_{a_1}, \ldots d_{a_n} \rangle$, kde $a_i \in \{1, \ldots, n\}$ a přitom

$$\mathrm{tfidf}(q, d_{a_1}) \geq \mathrm{tfidf}(q, d_{a_2}) \geq \cdots \geq \mathrm{tfidf}(q, d_{a_n}).$$

- Pomocí tfidf algoritmu můžeme také vygenerovat klíčová slova ke každému dokumentu – jsou to ta slova, která mají nejvyšší tfidf hodnotu mezi všemi slovy daného dokumentu.

Problém: proč tfidf funguje?

- Algoritmus tfidf *funguje* v tom smyslu, že výsledky jsou použitelné pro běžné nenáročné potřeby.
- Algoritmus se navíc dočkal mnoha modifikací, které jeho úspěšnost dále zvyšují.
- Otázkou zůstává: *proč* algoritmus funguje?

Teoretický vhled

- Přepíšeme tfidf algoritmus do řeči teorie pravděpodobnosti.
- Představme si náhodnou sadu dokumentů D. Jaká je pravděpodobnost p, že dokument d obsahuje právě $k = \mathrm{tf}(t, d)$ slov t?

$$p = \binom{n}{k} \cdot \left(\frac{1}{N}\right)^k \cdot \left(1 - \frac{1}{N}\right)^{n-k}.$$

- Přitom nás ale zajímají jen případy, kdy

$$1 \ll k \ll n \ll N.$$

- Protože $n \ll N$, dostaneme $n - k \ll N$, odtud $(n - k) \cdot \dfrac{1}{N} \ll 1$, čímž získáme

$$\left(1 - \frac{1}{N}\right)^{n-k} \approx 1.$$

Teoretický vhled #2

- Pomocí Stirlingovy věty a několika dalších úprav dostaneme

$$p \approx \frac{n^k \cdot e^k}{k^k} \cdot \left(\frac{1}{N}\right)^k.$$

- Dalším výsledkem je vložení logaritmů do předchozího vzorce. Opět po několika úpravách dostáváme: $-\ln(p) \approx k \cdot \ln\left(\frac{N}{n}\right) + k \cdot \ln(k)$.

- Původní vzorec tfidf algoritmu lze přepsat jako $k \cdot \ln\left(\frac{N}{\widetilde{n}}\right)$, přitom vzorec s pravděpodobností lze dále upravit, protože pro $\frac{N}{n} \gg k$ máme $\ln(k) \ll \ln\left(\frac{N}{\widetilde{n}}\right)$, takže

$$-\ln(p) \approx k \cdot \ln\left(\frac{N}{\widetilde{n}}\right).$$

- $=>$ dostáváme původní tfidf vzorec.