Kapitola 27

Státnice I3: Vyhledávání a extrakce informací

27.1 Informační systémy

- Faktografické vs. dokumentografické
- Zpřístupnění vs. dodání dokumentu
- Indexace nutná – termy
  - řízená, neřízená
  - tezaury
- Kritérium predikce + maxima
- Precision, recall

27.2 Vyhledávání v textu

- Triviální algoritmus
- Knuth-Morris-Pratt
- Aho-Corrasicková

27.3 Boolské informační systémy

- Dokument reprezentován množinou termů, které ho vystihují
- Dotazy: AND, OR, NOT, wildcards, víceslovné, proximitní omezení, tezaurus, lemmatizace
- Invertovaný indexový soubor (org. po termech)
- Uspořádání výsledků (DNF, počet splněných konjunkcí)
- Zpětná vazba

27.4 Vektorové informační systémy

- Každý z n dokument reprezentování m-složkovým vektorem vah důležitostí termů ($\in [0,1]$)
- Indexový soubor je matice vah $m \times n$
- Dotaz je taky vektor, vyhodnocení a řazení pomocí:
  - základní $Sim(\vec{w_i}, \vec{q}) = \sum_{k=1}^{n} w_{i,k} q_k$
  - vylepšení na délku vektorů (počet nenulových w_k) – dělení $\sum w_i + \sum q$, $\sum w_i + \sum q - 2\sum wq$ nebo $\sqrt{\sum w_i^2 \cdot \sum q^2}$
  - jiné – normalizace na jednotkovou délku vektorů

- Nerozlišuje se disjunkce a konjunkce
- Negace = přidání záporných vah do dotazů
- Indexace podle term frequency – $TF_{i,j} = \frac{t_j}{\sum_{i=1}^{m} t_i}$ (podíl počtu výskytů daného termu v dokumentu z celk. počtu termů v něm)
  - Normalizovaná $NTF = \frac{1}{2} + \frac{TF}{2\max(TF)}$ (do $\{0\} \cup [1/2, 1]$).
  - Inverzní $ITF_j = \log(n/k)$, pokud se term j vyskytuje v k dokumentech z n.
- Výpočet vah $w = \frac{NTF \cdot ITF}{Z}$ (Z je normalizace)
- Matice podobnosti termů – závislost a zastupitelnost termů

27.5 Induktivní systémy

- Dvouvrstvá neuronová síť se zpětnou aplikací vah (1. vrstva — termy, 2. — dokumenty)
- Laterální inhibice – zabránění nárůstu vah

27.6 Signaturové systémy

- Uložení na pomalých médiích – předstupeň k lepší metodě
- Každý dokument i search term má signaturu, která funguje jako maska (pokud je bitový and signatury dokumentu a termu nenulový, je dokument možná relevantní a použije se k detailnímu hledání)
- Přiřazení signatury – každý term: jedna jednička na nějakém místě / hashovací funkce
  - Zabránění příliš mnoha jedničkám v signaturách dokumentů – rozdělení na bloky (pevné délky nebo pevného počtu jedniček v signatuře)
- Wildcardy obecně nejsou možné, jen s monotónními signaturami

27.7 Rozšířená boolská logika

- Reprezentace stejná jako vektorový model
- Dotazy stejné jako s boolskou logikou, ale s váhami (pokud nejsou uvedeny, bere se 1)
- OR – vzdálenost od nulového dokumentu $DF = (0, \ldots, 0)$ jako $\sqrt[p]{\frac{q_a^p w_{i,a}^p + q_b^p w_{i,b}^p}{q_a + q_b}}$ (kde q_a, q_b jsou váhy dotazu)
- AND – vzdál. od jednotkového dokumentu jako $1 - \sqrt[p]{\frac{q_a^p (1-w_{i,a})^p + q_b^p (1-w_{i,b})^p}{q_a + q_b}}$
- Pro $p = 1$ je to vlastně vektorový model, pro $p \to \infty$ se blíží k boolskému

27.8 Rozlišovací hodnoty termů v indexu

- Informace o tom, jak dobře termy rozlišují dokumenty – co se stane, když nějaký z nich vyhodíme
- Rozlišovací hodnota $DV_k = Q^{(k)} - Q$, kde $Q = \frac{\sum_{i=1}^{n} Sim(d_i, C)}{n}$ je průměrná podobnost dokumentů s centroidem ("průměrným dokumentem" $C = \frac{\sum_{i=1}^{n} d_i}{n}$) a $Q^{(k)}$ je totéž, odstraníme-li k-tý dokument.
- Je možné použít jako IFT, má lepší vlastnosti než ten logaritmus (viz výše)

27.9 Přibližné hledání

- Detekce chyb, nalezení blízkých termů ve slovníku:
  - Počet společných digramů
  - Hammingova míra (počet operací replace při doplnění slova znakem λ na stejnou délku)
  - Levenshteinova míra (počet operací replace, insert nebo delete)
- Lze použít konečné automaty