Identifikace tématických sociálních sítí

Jiří Jelínek

Katedra managementu informací
Fakulta managementu J. Hradec
Vysoká škola ekonomická Praha

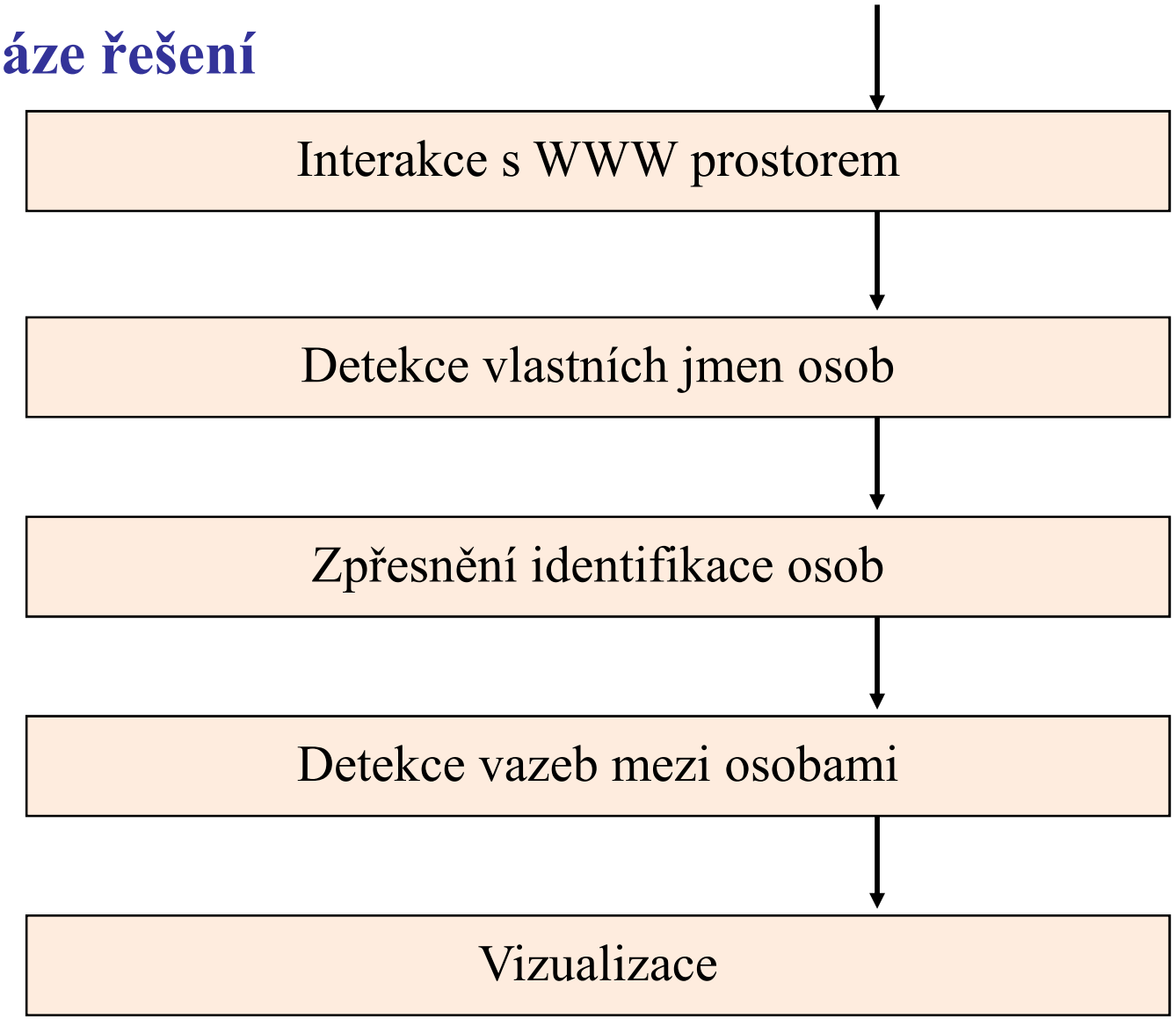
Obsah prezentace

- Cíl
- Fáze řešení a navržené postupy
- Prototyp a výsledky
- Další postup

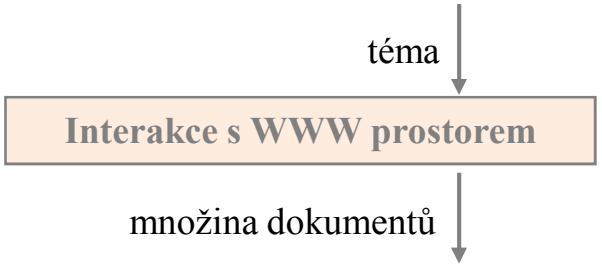
Cíl

- Kdo se tím zabývá?
  - identifikace osob spojených s daným tématem, oblastí či problematikou
  - vzájemné vazby osob
- Tématická sociální síť
- Souhrn metod a postupů pro praktické užití
- Využití WWW prostoru jako zdroje dat

Fáze řešení

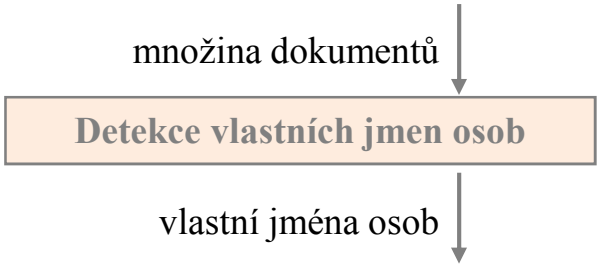

Interakce s WWW prostorem
Detekce vlastních jmen osob
Zpřesnění identifikace osob
Detekce vazeb mezi osobami
Vizualizace

Interakce s WWW prostorem



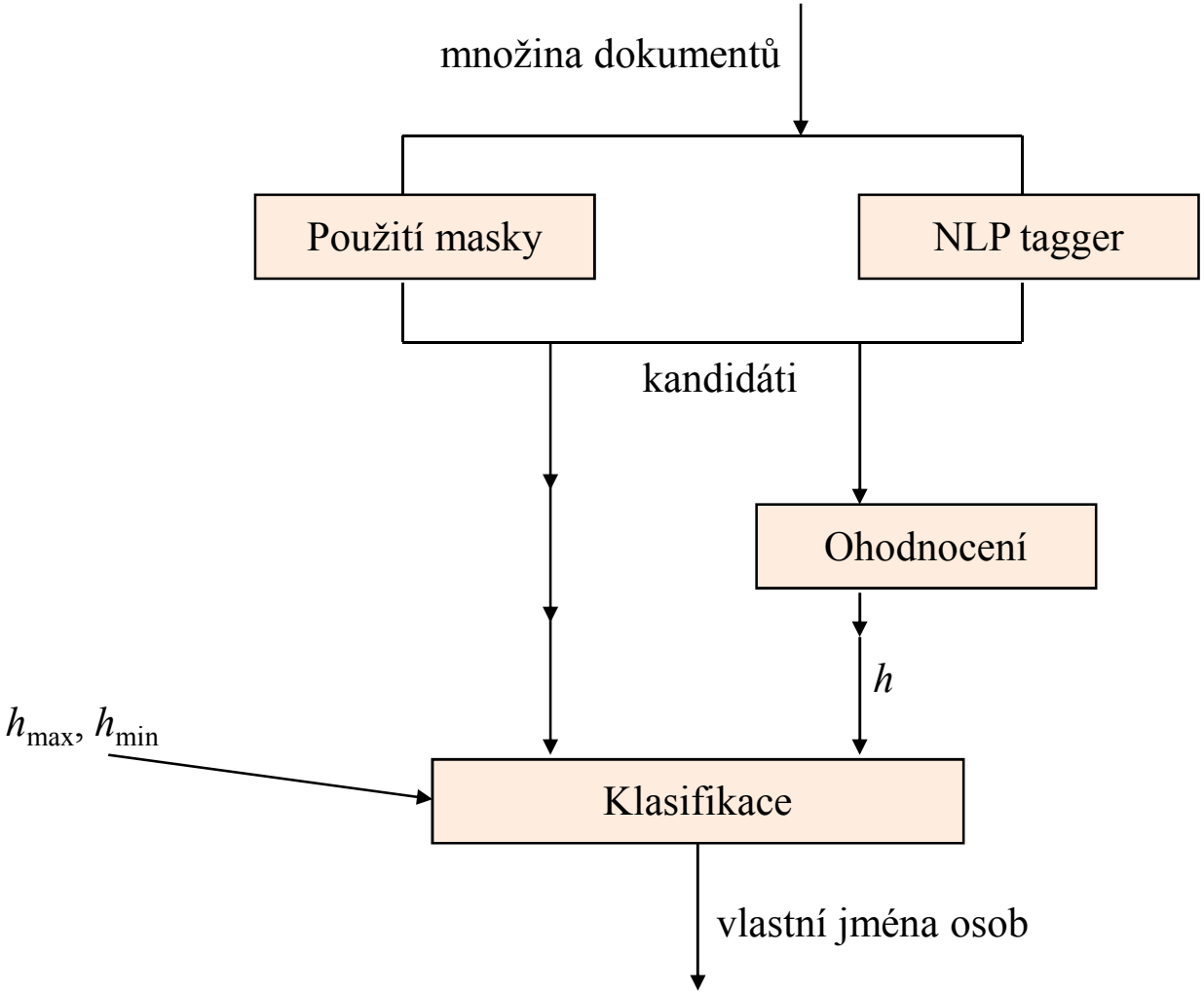
- Téma reprezentováno klíčovými slovy
- Výstupem množina textových dokumentů
- Použití WWW vyhledávačů

Detekce vlastních jmen osob

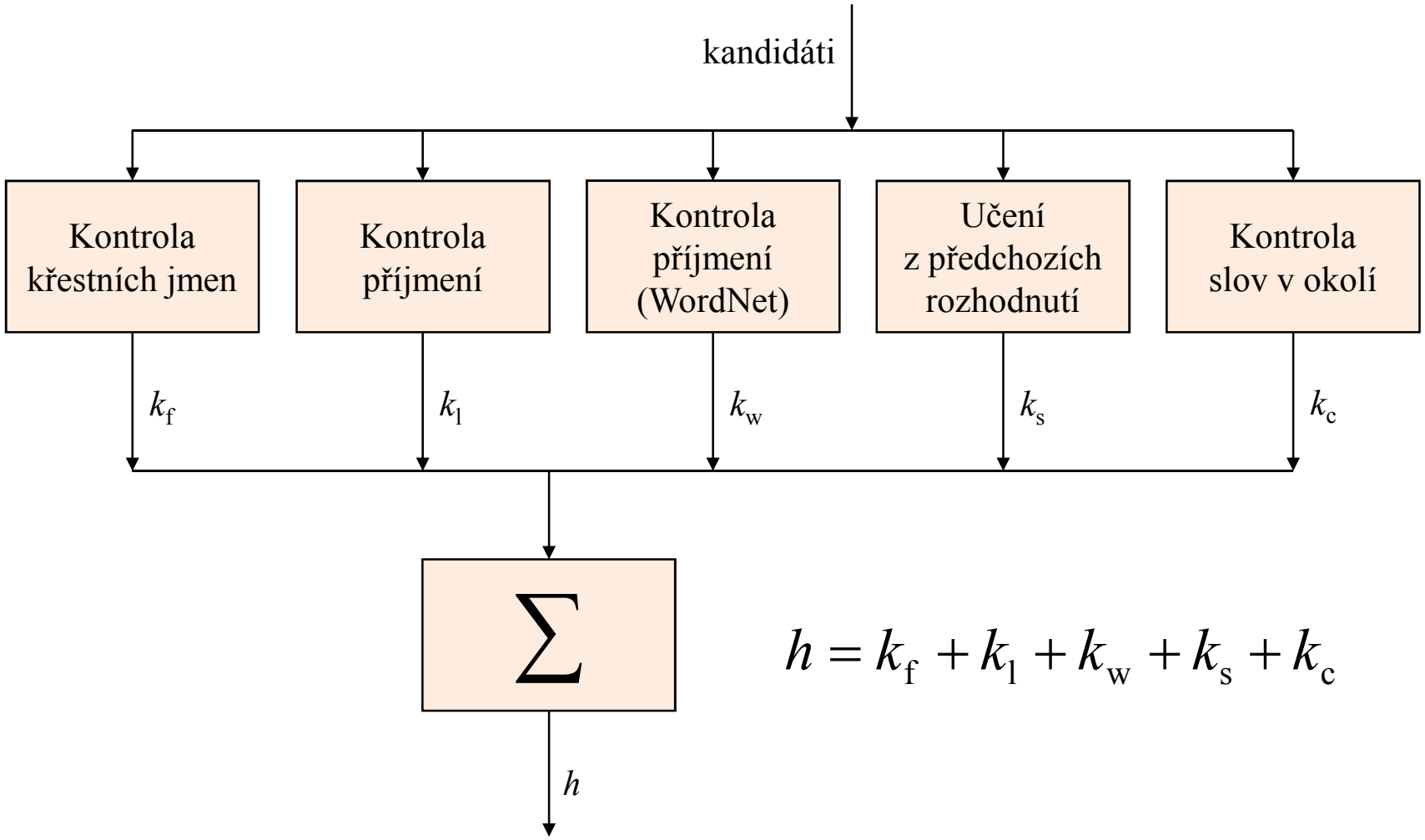


- Metody NLP
- Statistický přístup
- Slovníky vlastních jmen osob
- Využití kontextu

Detekce vlastních jmen osob



Detekce vlastních jmen osob



$$h = k_f + k_l + k_w + k_s + k_c$$

Detekce vlastních jmen osob

- Kontrola křestních jmen – 60 tis. jmen
  - Behind the Name - the Etymology and History of First Names
  - DBLP Bibliography
- Kontrola příjmení – 217 tis. příjmení
  - Frequently Occurring Names from the 1990 Census
  - ICU Project at the Data Privacy Laboratory
  - DBLP Bibliography

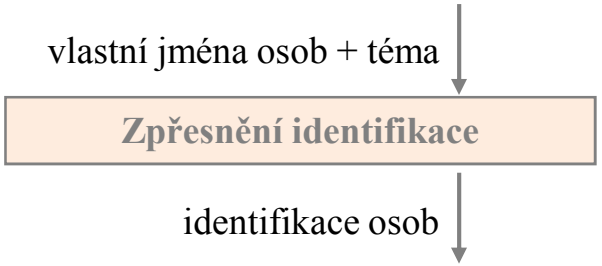
Detekce vlastních jmen osob

- Kontrola příjmení (Wordnet) – 143 tis. slov
  - kladné ohodnocení, pokud kandidát není ve Wordnetu
- Učení z předchozích rozhodnutí
  - lze na křestní jména i příjmení

$$k_s = k_{sm} \frac{c_p - c_n}{c_p + c_n}$$

- Kontrola slov v okolí
  - sledování „okolí“ kandidáta (3 slova před a po)
  - jako u učení

Zpřesnění identifikace

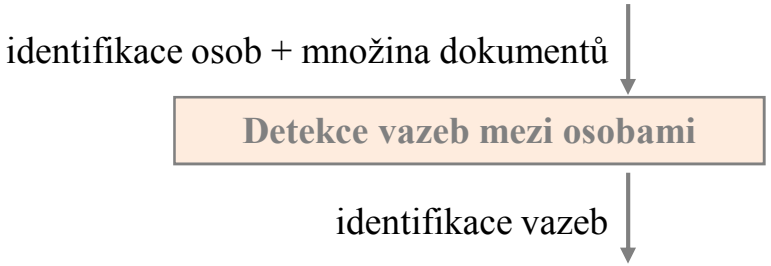


- Čištění vstupních dat
  - zpracování gramatických chyb a různých zápisů křestních jmen i příjmení
  - koef. shody (soundex, levenshtein)

Zpřesnění identifikace

- Odlišit osoby se stejným vlastním jménem
  - užití tématu jako doplňkové informace
  - v tématu jedna osoba s daným vlastním jménem
- Identifikace osob s různou formou zápisu vlastního jména
  - John Smith – J. W. Smith – Smith, J. William – John William Smith
  - z možných zápisů vybrán ten nejúplnější

Detekce vazeb mezi osobami



- Množina dokumentů S - jedno téma nebo sjednocení dokumentů z vybraných témat
- Váha w_{iS} termu i vzhledem k S
  - množina významných termů V - $w_{iS} > práh$

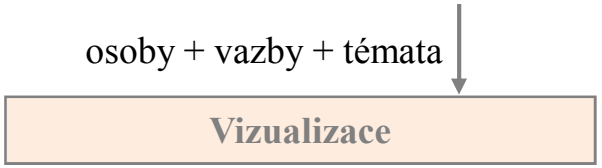
Detekce vazeb mezi osobami

- Detekce vazeb mezi termy ve V
  - podle společného výskytu ve vstupních dokumentech
  - síla vazby p_{ijS} mezi termy i a j nad množinou S

- Významnost vazby h_{ijS} mezi termy
  - vazby významných osob nebo pevné týmy – koef. k
  - dále jen vazby, kde $h_{ijS} > práh$

$$h_{ijS} = k(w_{iS} + w_{jS}) + (1-k)\,p_{ijS}$$

Vizualizace



- Knihovna Graphwiz
  - algoritmus NEATO
- Vzdálenost uzlů sítě úměrná $1/h_{ijS}$
  - na vstupu vizualizace, nelze vždy
- Barva uzlů podle w_{iS}

Prototyp a dosažené výsledky

- Aplikace v PHP + MySQL
- Interakce s WWW prostorem
  - Google pro získání množiny dokumentů
  - přímé načtení zadaného URL a jeho přidání k tématu
- Detekce vlastních jmen osob
  - ohodnocení nalezených kandidátů
    - ruční v úvodních fázích – návrh $h_{\min}$ a $h_{\max}$
    - později automatické

Prototyp a dosažené výsledky

- Zpřesnění identifikace
  - bez čištění dat
- Detekce vazeb mezi termy
  - analýza množiny témat podle zadaných prahových hodnot w_{iS} a h_{ijS} nebo podle požadovaného počtu zobrazených vazeb
- Grafické zobrazení výstupu
  - zobrazení osob v tématu podle příjmení
  - vazby vybraného jedince podle příjmení
  - seznam dosud načtených témat a dokumentů

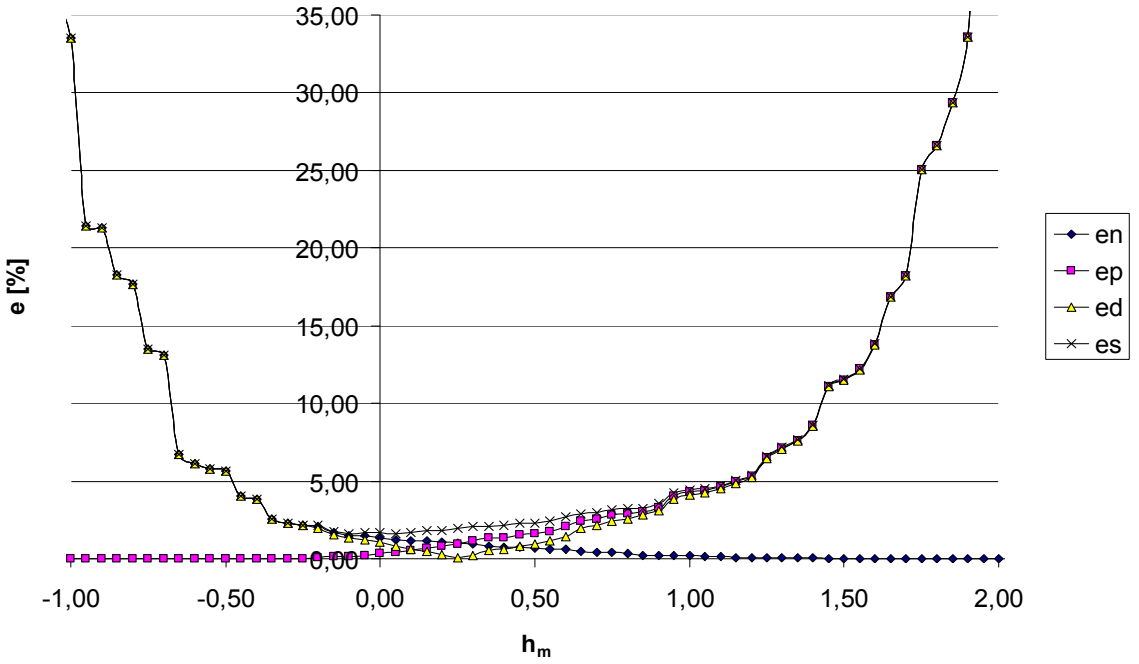
Nastavení vstupních hodnot

- Volba k_f , k_l , k_w , k_{sm} , k_{cm}
  - tak, aby byla co nejlépe odlišitelná vlastní jména osob od ostatních slovních spojení
  - zvoleno $h_{min} = h_{max} = h_m$
  - pro dané hm
    - e_n – chybně ohodnocené negativní příklady
    - e_p – chybně ohodnocené pozitivní příklady

$$e_n = \frac{c_{n+}}{c_n} \qquad e_p = \frac{c_{p-}}{c_p}$$

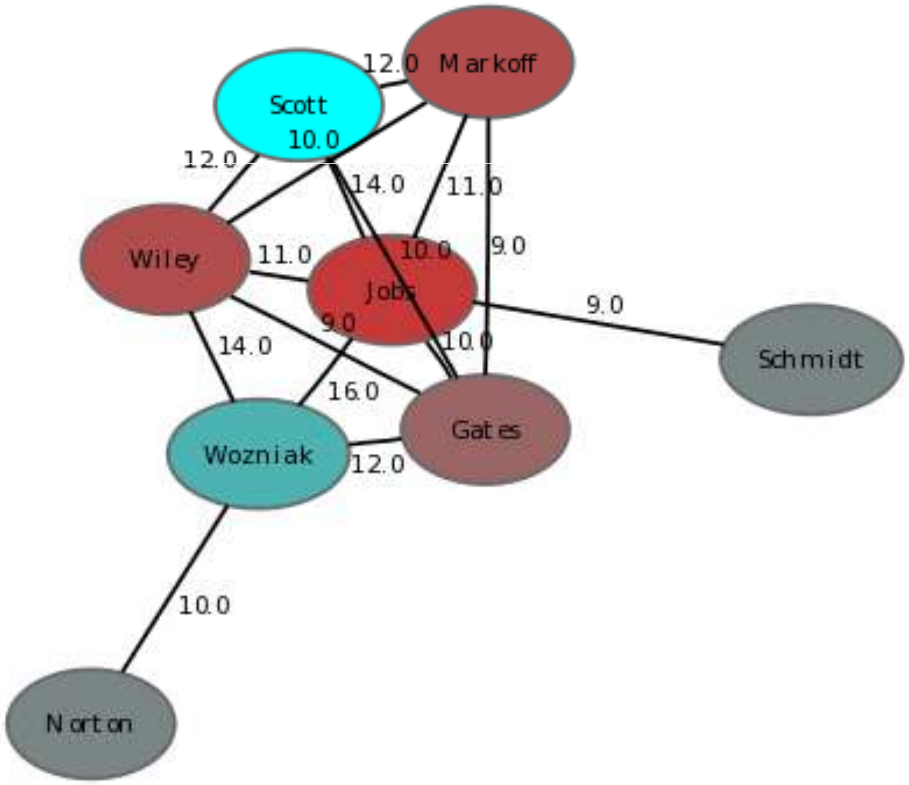
Nastavení vstupních hodnot

- Volba h_m
  - podle $e_d = | e_p - e_n |$ (stejná velikost chyb e_p a e_n)
  - podle $e_s = e_p + e_n$ (minimální souhrnná chyba)



Výsledky

- 16671 klasifikovaných termů, z toho 3079 pozitivně hodnocených jako vlastní jména osob
- 17 témat
- 397 WWW stránek



Další postup

- Základ pro další výzkum
- Rozšíření záběru a zpřesnění dosahovaných výsledků
  - rozšíření vstupních importních filtrů - citační servery
- Stanovení vhodných hodnot k_f , k_l, k_w, k_{sm}, k_{cm}
  - genetické algoritmy
  - $F(k_f , k_l, k_w, k_{sm}, k_{cm}) = e_d (k_f , k_l, k_w, k_{sm}, k_{cm})$

Další postup

- Metodika automatizovaného stanovení mezních hodnot $h_{\min}$ a $h_{\max}$ (event. h_{m})
  - výběr vhodného kritéria
- Zpřesnění identifikace osob
  - využití WWW zdrojů
- Vizualizace výstupů
  - 3D zobrazení pomocí jazyka VRML, X3D

Závěr

- Příspěvek do oblasti detekce potenciálních sociálních sítí
- Navrženy metody práce s vlastními jmény osob
- Funkční prototyp aplikace, prakticky otestován
- Výstupy použitelné všude, kde je potřeba identifikovat tématicky definované sociální sítě
  - výzkum, finančnictví, kriminalistika, ekonomika, atd.

Děkuji za pozornost

Otázky?