FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2014

Klasifikace textů využitím Linked Data a PageRanku

Michal Nykl[1], Martin Dostal[1]

1 Úvod

Klasifikace textů je nedílnou součástí systémů pro správu dokumentů a úloh zpracování textů. Současné metody klasifikace jsou obvykle statistické a obsahují dvě fáze: trénování a klasifikaci. Ve fázi trénování se snažíme nalézt vztahy mezi termy obsaženými v dokumentu a danými klasifikačními třídami. Časté je použití poměrně komplexních korpusů, které obsahují mnoho dokumentů s určenými klasifikačními třídami. Protože je zapotřebí velké množství dat, jsou obvykle příprava trénovacích množin dokumentů a výběr metody pro nalezení reprezentativních termů dokumentu úlohami pouze pro doménové specialisty.

Náš postup, prezentovaný v článcích Dostal et al. (2014) a Nykl et al. (2013), využívá sémantických informací získaných z Linked Data pro rozšíření základních rozpoznaných klíčových slov dokumentu. Například term *MySQL* může využitím Linked Data expandovat na term *databáze* bez potřeby explicitního výskytu tohoto termu v obsahu dokumentu. Při klasifikaci pak můžeme použít tyto nadřazené pojmy pro správné spárování dokumentu s klasifikační třídou.

Dále budou popsány základní principy Linked Data a algoritmu PageRank a představena metoda nalezení nadřazených reprezentativních termů pro daný dokument.

2 Linked Data a PageRank

Linked Data představil Berners-Lee (2006) jakožto koncept pro doplnění sémantických informací do webových stránek. Definována byla čtyři pravidla pro vytvoření strojově čitelného obsahu webu:

- Identifikujte věci využitím URI
- Používejte HTTP URI, aby se věci dali vyhledat
- Při HTTP požadavku na URI poskytněte datovou reprezentaci věci (RDF, SPARQL)
- Umožněte objevení dalších věcí uvedením jejich URI

Iterační algoritmus PageRank byl původně vyvinut autory Brin a Page (1998) pro určování významnosti webových stránek v rámci vyhledávacího stroje (v současnosti je nedílnou součástí vyhledávače Google). Obecně PageRank slouží k určení významnosti vrcholů grafu využitím vstupních hran a významnosti vrcholů, ze kterých tyto hrany vedou. Ve vzorci PageRanku (1) je $P_x(a)$ hodnota PageRanku vrcholu a v iteraci x, d je damping faktor (obvykle $d=0,85$), V je množina všech vrcholů grafu, U je množina vrcholů, které mají vstupní hranu do vrcholu a, D je množina vrcholů, které nemají výstupní hrany, a w_{ua} je váha hrany vedoucí z vrcholu u do vrcholu a.

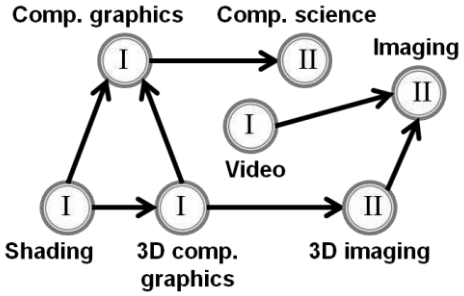
$$P_{x+1}(a) = \frac{(1-d)}{|V|} + d * \left(\left(\sum_{u \in U} \frac{P_x(u) * w_{ua}}{\sum_{v \in V} w_{uv}}\right) + \frac{\sum_{s \in D} P_x(s)}{|V|}\right) \qquad (1)$$

[1] studenti doktorského studijního programu Inženýrská informatika, obor Informatika a výpočetní technika, specializace Data mining, e-mail: [nyklm, madostal]@kiv.zcu.cz

3 Metoda nalezení reprezentativních termů dokumentu

Nalezení reprezentativních termů je nejdůležitější součástí naší metody klasifikace. Vybrané termy jsou použity ve fázi trénování i následně ve fázi klasifikace. Jednotlivé kroky našeho algoritmu pro nalezení reprezentativních termů jsou:

1) Nalezení základních termů pro každý dokument využitím TFIDF (nebo χ^2).
2) Namapování základních termů na vrcholy v Linked Data a vytvoření grafu.
3) Expanze vrcholů grafu o jeden krok využitím vazeb z Linked Data. (Tím získáme první verzi grafu – příklad ilustruje obr. 1, kde základní termy/vrcholy jsou označeny *I* a nově přidané vrcholy jsou označeny *II*.)
4) Výpočet algoritmu PageRank pro celý graf.
5) Když alespoň jeden vrchol z nově přidaných vrcholů má skóre PageRanku vyšší, než mají všechny staré vrcholy, pokračuj krokem 3), jinak algoritmus ukonči a vrcholy s nejvyššími skóre PageRanku prohlas za nejvíce reprezentativní termy dokumentu.



Obrázek 1: Expanze grafu využitím Linked Data

4 Závěr

Ve vlastní práci bylo testováno několik typů vážených grafů a v kroku 5) zmíněného algoritmu byl vždy vybrán pouze jeden term s nejvyšším skóre. Vyhodnocení bylo provedeno využitím 20 News Groups - kolekce novinových článků rozdělených do 20 tříd dle oblasti zájmu. Nejlepších výsledků klasifikace bylo dosaženo, když byly pro trénování použity malé množiny dokumentů (10 dokumentů na 1 kategorii). Při použití větších množin dokumentů docházelo k mírnému přetrénování, což je problém, který chceme v budoucnu vyřešit.

Naše metoda klasifikace textů poskytuje slibné výsledky v případech, kdy máme neadekvátní trénovací množiny (málo dokumentů atd.) nebo chceme rychle filtrovat existující dokumenty. Velkou výhodou metody je, že může být využita neprofesionálními uživateli, kteří pouze zvolí základní termy dokumentu (např. klíčová slova).

Zde zmíněný postup lze použít i pro shlukování nebo štítkování, viz Dostal et al. (2013).

Literatura

Berners-Lee, T., 2006. *Linked Data – Design Issues.* [on-line, cit. 7. 5. 2014] Dostupné z: http://www.w3.org/DesignIssues/LinkedData.html

Brin, S., and Page, L., 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, vol. 30 pp. 107–117.

Dostal, M., Nykl, M., and Ježek, K., 2013. Cluster labeling with Linked Data. *Journal of Theoretical and Applied Information Technology*, vol. 53. pp 340–345.

Dostal, M., Nykl, M., and Ježek, K., 2014. Exploration of Document Classification with Linked Data and PageRank. In: *Intelligent Distributed Computing VII – IDC 2013, Prague (CZ), September 2013*. Springer: Studies in Computational Intelligence, vol. 511.

Nykl, M., Ježek, K., Dostal, M., and Fiala, D., 2013. Linked Data and PageRank based classification. In: *IADIS International Conference TPMC 2013*. Praha: IADIS Press.