Posudek diplomové práce

Michal Hrala – Automatická klasifikace podobných dokumentů 2012

(autor posudku – Ing. Jan Pluskal)

Diplomová práce „Automatická klasifikace dokumentů s podobným obsahem" Michala Hraly se, jak název napovídá, týká klasifikace dokumentů, a to především s ohledem na přiřazování tematických kategorií. Tato problematika je pro ČTK zajímavá v několika směrech, vlastní produkci tematické kategorie přiřazujeme, tento prvek metadat je velmi důležitý pro naše odběratele, a to i z obchodních důvodů dělení produkce na prodejní bloky. Uvítáme tedy možnost kontroly systému přiřazování. Zároveň do ČTK proudí spousta dokumentů (zpravodajství partnerských agentur, novinové články, atd.), u kterých kategorie nejsou přiřazeny vůbec nebo velice nedostatečně. Na vlastní přiřazení následně může navázat porovnání tematické struktury produkce celé ČTK a produkce obchodních partnerů – výstupem by mělo být srovnání, jak moc či málo odpovídá zpravodajství ČTK potřebám klientů a jak se tato shoda vyvíjí v čase.

Diplomová práce mi připadá napsaná dobře a vyváženě, teoretické části je věnován potřebný prostor a stejně tak části praktické i vlastní realizaci. Autor DP po úvodu popisuje v druhé kapitole typy klasifikačních úloh, existující algoritmy a metody klasifikace dokumentů. Popis všech metod je doplněn vzorci i textovým popisem, v DP jsou uvedeny odkazy na prameny, ze kterých autor čerpal. Na vhodných místech jsou vložený obrázky (např. popis metody SVM).

V kapitole 3 jsou uvedeny tři klasifikační metody, které byly použity pro praktickou část, a jsou zde popsány různé klasifikační nástroje dostupné jako volně šiřitelné i některé komerční. Z analýzy těchto nástrojů je patrné, že většina z nich obsahuje metody dvě, autorem vybraný Minorthird nabízí použití všech tří autorem vybraných klasifikátorů (a řadu dalších). Jednotlivé nástroje jsou přehledně srovnány v tabulce. Kapitola tři dále pojednává o databázích ČTK, je zde rozebrána struktura polí v databázích a v tabulce jsou uvedeny statistické přehledy dat dodaných pro zpracování. Konec kapitoly 3 je věnován způsobu měření úspěšnosti, jsou zde uvedeny jednotlivé metriky, opět včetně popisu a vzorců.

Následující kapitoly se již věnují praktické části, v kapitole 4 je popsán návrh aplikace, popis GUI, vstupů, výstupů a jednotlivých programových tříd. Kapitola 5 přestavuje dosažené výsledky, hodně prostoru je věnováno velikosti příznakového vektoru, autor správně bere ohled i na časovou náročnost zpracování. Text kapitoly je doplněn o grafy i tabulky, obojí vhodně a naprosto dostatečně popisuje naměřené výsledky.

V poslední kapitole 6 autor celou diplomovou práci uzavírá, shrnuje výsledky, ke kterým dospěl a uvádí některé možnosti rozšíření. Následují přílohy, seznam zkratek a seznam použité literatury.

Práce naprosto splňuje požadavky ČTK na automatickou kvalifikaci stejně tak jako celé vlastní zadání. Autorem navržené řešení s vysokou pravděpodobností použijeme. DP je dobře strukturovaná, v textu nejsou prakticky žádné pravopisné chyby, pokud jsem narazil na nějaké nedostatky, byly naprosto okrajové (drobná nesrovnalost v názvech jednotlivých databází ČTK), v teoretické části možná mohlo být věnováno více prostoru shlukovým analýzám, což by ale na druhou stranu bylo nad rámec této DP. Z uvedených důvodů hodnotím práci stupněm výborně a doporučuji k obhajobě.

Otázky:

Myslíte, že by bylo možné zvolené způsoby kvalifikace použít i pro jiné kvalifikátory než tematické kategorie? (např. sentiment dokumentu, prioritu a urgenci apod.)

Jakým způsobem byste řešil nastavení mezní hodnoty (treshold) pro získané míry pravděpodobnosti přiřazení kategorií u naprosto neznámých dokumentů, kde není znám počet kategorií?

V Praze dne 1. června 2012

Ing. Jan Pluskal