

Štatistické modelovanie hypertextu

Študijný program: SOFTVÉROVÉ INŽINIERSTVO

Autor: Bc. Ján Suchal

Vedúci diplomovej práce: Mgr. Gabriela Kosková PhD.

December 2006

Táto práca sa venuje skúmaniu, implementácii a testovaniu štatistického modelu na automatické zhľukovanie hypertextových dokumentov. Sústreďuje sa na pravdepodobnostnú latentnú sémantickú analýzu, ktorá bola realizovaná iteračným EM algoritmom. Formou experimentov so syntetickými a reálnymi dátami popisuje funkčnosť algoritmu a vysvetľuje jeho závislosť od zvolených parametrov. Výsledky experimentov poukazujú na nevýhodu tohto prístupu, ktorou je nutnosť empiricky určovať počet hľadaných zhľukov, ktorý zásadne ovplyvňuje výslednú kvalitu zhľukovania. Výsledky zhľukovania získané pravdepodobnostným algoritmom sú porovnané s výsledkami získanými algoritmi založenými na vzdialenostných metrikách. Na príklade je demonštrovaná neschopnosť populárneho k-means algoritmu riešiť problém s vysokou dimenzionalitou vstupných dát.

Výsledkom tejto práce je balík skriptov na zber, transformáciu, samotné zhľukovanie a analýzu výsledkov zhľukovania pre textové a hypertextové dokumenty. Priložená je aj časť dokumentov encyklopédie Wikipedia, na ktorej boli vykonané zdokumentované experimenty. V práci sa prezentuje metrika nazvaná kumulatívna entropia, pomocou ktorej je možné porovnať kvalitu výsledkov zhľukovania pre rôzne počty zhľukov.

Statistical hypertext modeling

Degree Course: SOFTWARE ENGINEERING

Author: Bc. Ján Suchal

Supervisor: Mgr. Gabriela Kosková PhD.

December 2006

This thesis examines, implements and tests a statistical model for automatic clustering of hypertext documents. Work focuses on probabilistic latent semantic analysis, which was implemented by an iterative EM algorithm. Functionality and relationships between various parameters are demonstrated by performing experiments on synthetic and real data sets. Results of experiments are showed that one of main drawbacks of this approach to clustering is the need to empirically set the number of clusters prior to clustering, which has radical impact on overall clustering quality. Probabilistic clustering algorithm is compared to standard distance-based clustering algorithm, explaining the inability of popular k-means algorithm to deal with high dimensional input data.

Result of this thesis is a package of scripts for collecting, transforming, clustering and analyzing results of clustering for text and hypertext documents with a partial set of documents gathered from worldwide encyclopedia Wikipedia. This thesis also presents a metric named cumulative entropy which can be used for evaluating clustering quality for different cluster counts.