Секция «Вычислительная математика и кибернетика»

Согласование иерархической тематической модели с внешним рубрикатором

Исупова Ольга Олеговна
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия
E-mail: ihoho89@gmail.com

[12pt,a4paper]article
[cp1251]inputenc
[russian]babel

Согласование иерархической тематической модели с внешним рубрикатором

Вероятностные тематические модели, такие как PLSA (вероятностный латентный семантический анализ) или LDA (латентное размещение Дирихле) широко применяются для выявления тематики в больших коллекциях научных документов. Они позволяют оценивать условные вероятности $p(w|t)$ того, что термин w относится к теме t, и условные вероятности $p(t|d)$ того, что термины документа d относятся к теме t. Иерархические тематические модели наряду с этим восстанавливают отношение «общее–частное» между темами. Число уровней иерархии и число тем в каждом узле либо задаётся априори с помощью внешнего рубрикатора, либо определяется исходя из статистических свойств коллекции документов. В первом случае рубрикатор может оказаться неполным, субъективным или устаревшим. Во втором случае статистические закономерности коллекции могут противоречить общепринятым представлениям о структуре предметной области; кроме того, возникает проблема интерпретации и наименования каждой темы.

Предлагается итерационный метод согласования иерархических моделей PLSA и LDA с внешним рубрикатором. Для коллекции научных документов рассматривается подмножество рубрикатора УДК. Первоначально структура иерархии определяется рубрикатором и используется информация о принадлежности некоторых документов определённым рубрикам. Модель выявляет случаи, когда структура рубрикатора не согласуется со статистическими свойствами коллекции. Рассматриваются следующие типы несогласия: раздел неоднороден и должен быть разбит на подразделы; подразделы однородны и могут быть объединены в новый раздел; документ или раздел может быть отнесён к другим надразделам, возможно, одновременно к нескольким. Для проверки этих условий используются статистические критерии согласия и однородности для мультиномиальных распределений. Для выявления отношения «общее–частное» между темами используется расстояние Кульбака–Лейблера между распределениями вида $p(w|t)$ [1]. Выявленные несогласия предъявляются экспертам в виде списка, ранжированного по «степени несогласия». Изменения в структуру иерархии вносятся только в том случае, когда эксперты их подтверждают, что позволяет отслеживать актуальность структуры, избегая произвольных изменений, которые могла бы допускать модель в случае полностью автоматического режима работы.

Работа выполнена при поддержке Министерства образования и науки РФ в рамках Государственного контракта 07.524.11.4002.