Секция «Вычислительная математика и кибернетика»

Обучаемые алгоритмы выделения ссылок в научных документах

Полежаев Валентин Александрович
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия
E-mail: valentin.polezhaev@gmail.com

Автоматическое построение графа цитирований по коллекции научных документов требует решения следующей последовательности задач: (1) выделение одного или нескольких блоков библиографии в документе, (2) разбиение каждого блока на отдельные ссылки, (3) выделение в каждой ссылке полей авторов, названия и т.д., (4) идентификация ссылок и выявление ссылок-дубликатов. Известно, что применение методов машинного обучения — классификации, кластеризации, условных случайных полей CRF [2] для решения задач (3, 4) повышает качество распознавания цитирования по сравнению с простыми эвристическими методами. В то же время, для решения задач (1, 2) принято использовать «инженерный подход», что вполне оправдано для относительно небольших однородных коллекций. В данной работе для решения задач (1, 2) предлагаются методы классификации, ориентированные на обработку больших мультидисциплинарных мультиязычных коллекций, содержащих документы различных форматов и жанров.

Объектами в задаче классификации (1) являются строки текста. Признаками являются информативные характеристики строки, например, число чисел, число символов-разделителей, наличие определённых ключевых слов, наличие четырёхзначного числа, похожего на год, и т.д. Классов два — строка принадлежит или нет блоку библиографии. Для решения проблемы разрезанных строк используется суммирование признаков в скользящем окне, охватывающем несколько соседних строк. Предлагается метод оптимизации ширины окна. Для выделения блоков библиографии используется вторичная классификация на основе результатов базовой классификации. Объектами в задаче классификации (2) также являются строки текста; классов два — начало новой ссылки или продолжение предыдущей.

Для решения задач классификации предлагается использовать решающие деревья, их композиции и модель CRF. Все методы явным образом учитывают зависимости между последовательными строками. Рассматриваются также инкрементные решающие деревья и их композиции [1]. Их применение позволяет минимизировать затраты времени на разметку обучающих выборок благодаря тому, что эксперты размечают только те объекты, на которых алгоритм классификации допустил ошибку.

Предварительные эксперименты на коллекции разнородных документов, составленной из авторефератов ВАК, статей из англоязычных журналов, докладов с различных конференций, показывают, что предложенные методы позволяют добиться высокого качества распознавания строк библиографии (порядка 97%), не достижимого при использовании чисто эвристического подхода без обучения по прецедентам.

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

Литература

1. Кудинов П. Ю., Полежаев В. А. Композиция случайных инкрементных деревьев и восстановление структуры таблиц // Бизнес-информатика. 2011. Т.18. No. 4. С.39–46.

2. Lafferty, J., McCallum, A., Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data // Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann. 2001. С.282–289.

Слова благодарности

Автор выражает благодарность своему научному руководителю д.ф.-м.н. К.В. Воронцову за советы и конструктивные замечания.