УДК 20.19.27

ИЗВЛЕЧЕНИЕ КОНЦЕПТОВ ИЗ ВЬЕТНАМСКИХ ТЕКСТОВ

Данг Хань (Университет ИТМО).
Научный руководитель – профессор, д.т.н Бессмертный Игорь Александрович
(Университет ИТМО)

Аннотация
Работа посвящена проблеме автоматического извлечения знаний из естественно-языковых текстов (*text mining*). разработка и исследование метода извлечения концептов в предметной области на основе подсчета частоты встречаемости слов в тексте. Предложение метода использования контрастной коллекции для повышения точности задачи.

Введение

В зависимости от характеристик каждого естественного языка обработка естественного языка, имеющего определенные трудности и преимущества. Поэтому необходимо исследовать и выбрать методы оптимизации процесса обработки естественного языка с наилучшей оптимизацией и точностью. Для вьетнамского языка из-за неоднозначности семантики, сложности структуры слов, извлечение концептов во вьетнамских текстов до сих пор сталкивается со многими трудностями. В работе рассматривается использование метода извлечения терминов в тексте, основанного на подсчете частоты встречаемости слов в текстов. Исследование и предложение использовать наборы контрастов для повышения точности задачи.

Преимущество извлечения концептов в тексте путем подсчета частоты слов состоит в том, что его легко выполнить, но недостатком является появление стоп-слов. Производительность удаления стоп-слов влияет на точность задачи.

Использованием контрастной коллекции создает корпус слов, в котором содержит очень большое количество стоп-слов. Операцией над множеством будет удалять окончания слова и слова, не имеющие значения в рассматриваемой предметной области.

Основной результат

Метод подсчета встречаемости слов использовался для извлечения концептов во вьетнамских текстах. После использования контрастного набора данных стоп-слова были удалены, точность метода повышена.

Выводы

Экспериментальные результаты были показан исследование и разработки метода извлечения терминов в вьетнамских текстах в предметной области, основанного на сочетании метода подсчета частоты слов и использования контрастных наборов данных. Результаты исследования, применяющиеся для задачах извлечения знаний при обработке естественного языка. Однако в работе есть ошибки из-за процесса разделения слов на вьетнамском языке.

Данг Хань (Автор)
Бессмертный Игорь Александрович (Научный руководитель)