

ПЕРСПЕКТИВЫ МЕТОДА "ГРАФОВ ВИДИМОСТИ" В КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

В рамках теорий цифровой обработки сигналов и сложных сетей существует несколько методов построения графов на основе временных рядов, среди которых можно назвать несколько методов построения графов видимости, в частности, так называемый граф горизонтальной видимости. Эти подходы также позволяют строить сетевые структуры на основании текстов, в которых отдельным словам или словосочетаниям специальным образом поставлены в соответствие числовые весовые значения. В качестве функции, ставящей в соответствие слову число, можно рассматривать, например, порядковый номер уникального слова в тексте, длину слова, «вес» слов в текстах, общепринятую оценку TFIDF или ее варианты, а также другие весовые оценки.

В докладе авторами предлагается алгоритм компактифицированного графа горизонтальной видимости (КГГВ). На его основе строятся сети слов различных текстов, показано, что они являются безмасштабными, а также, что среди узлов с наибольшими степенями имеются слова, определяющие не только структуру связности текста, но и его информационную структуру.

Использование выявленных одновременно информационно и структурно важных лексических единиц как опорных слов для различных текстовых жанров, документов, представленных на разных языках, позволяет:

- формировать информационные портреты текстов;
- выполнять автоматическое реферирование текстов;
- формировать цепочки подобных документов, используя выявленные слова в качестве дескрипторов;
- выявлять возможное содержательное дублирование документов, представленных на различных языках;
- составлять словари опорных слов, формировать тезаурусы и онтологии предметных областей.