# ЛАТЕНТНО-СЕМАНТИЧЕСКИЙ АНАЛИЗ ТЕКСТА

**Калинбет А.В., Безменов Н.И.**
***Национальный технический университет «Харьковский политехнический институт», г. Харьков***

Одной из самых распространенных форм представления знаний являются естественно-языковые тексты. На сегодняшний день объемы таких знаний превосходят возможности человека самостоятельно анализировать и обрабатывать их. Поэтому актуализируется необходимость в исследовании и разработке новых подходов обработки данных. Сложность обработки указанных данных заключается в том, что вычислительные машины легко обрабатывают формальные языки, подчиненные строгим и однозначным правилам построения, в отличие от естественных языков, смысл которых часто зависит от контекста.

Для обработки естественного языка вычислительная машина должна справляться со следующими задачами: распознавать структуру текста и извлекать его смысл. Латентно-семантический анализ направлен на решение второй задачи и заключается в выявление скрытых (латентных) ассоциативно-семантических связей между словами при помощи статистической обработки больших наборов текстовых данных.

Процесс выявления латентных связей происходит путем сокращения факторного пространства термы-на-документ, где термы – это слова или комбинации слов, а документы – наборы текстов. Краткое описание алгоритма латентно-семантического анализа можно представить в виде сингулярного разложения матрицы исходных данных, которая представляет собой матрицу термы-на-документы, описывающую набор данных, используемый для обучения системы. Элементы этой матрицы содержат, веса, учитывающие частоты использования каждого терма в каждом документе или вероятностные меры, основанные на независимом распределении. Согласно теореме о сингулярном разложении в самом простом случае матрица может быть разложена на произведение трех матриц: $A = U \cdot S \cdot V^T$, где матрицы $U$ и $V$ – ортогональные, а $S$ – диагональная матрица, значения на диагонали которой называются сингулярными значениями матрицы $A$.

Особенность такого разложения в том, что если в матрице S оставить только $k$ наибольших сингулярных значений, то линейная комбинация получившихся матриц $U_{lsa}, S_{lsa}$ будет наилучшим приближением исходной матрицы $A$ к матрице $\tilde{A}$ ранга $k : A \approx \tilde{A} = U_{lsa} \cdot S_{lsa} \cdot V_{lsa}^T$. В результате полученная матрица $\tilde{A}$ содержит только $k$ первых линейно независимых компонент исходной матрицы $A$, отражает структуру ассоциативных зависимостей, латентно присутствующих в исходной матрице. Структура зависимостей определяется весовыми функциями термов для каждого документа.

Таким образом, целью данного исследования является повышение процессов качества поиска, обработки и анализа информации, содержащейся в больших объемах естественно-языковых текстов.