# МАТРИЧНОЕ СИНОНИМИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ КОРПУСОВ ЭЛЕКТРОННЫХ ТЕКСТОВ В ИНФОРМАЦИОННЫХ ПОИСКОВЫХ СИСТЕМАХ

*Чалая Л.Э., Шевякова Ю.Ю.*
*Харьковский национальный университет радиоэлектроники*

Интеллектуализация систем поиска текстовой информации требует учета ее смыслового содержания. Классические проблемы поиска документов – это синонимия (одно и то же понятие может быть выражено с использованием различных синонимичных терминов) и полисемия (один и тот же термин, используемый в анализируемых документах, может иметь различные значения в зависимости от конкретных контекстов) [1].

В докладе рассматривается одна возможная модель представления текстовой информации, основанная на использовании синонимических словарей.

Синонимические словари описывают слова, разные по звучанию и написанию, но тождественные или близкие по значению. Такое определение синонимов следует считать рабочим, поскольку оно не претендует на всесторонность охвата сущности синонимии. Существуют различные определения синонимии, множественность которых объясняется, прежде всего, особенностями самого предмета рассмотрения, его многообразием, а также существованием различных типов семантических сближений [2].

Входные данные в рассматриваемой задаче задаются лингвистическими корпусами. Лингвистический корпус – это совокупность текстов, собранных в соответствии с определенными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой. Такие корпусы могут бытьсформированы по определенному признаку – например, в соответствии с универсальной десятичной классификацией (УДК). УДК – система классификации информации, широко используемая для систематизации произведений науки, литературы и искусства, периодической печати, различных видов документов и организации картотек. Лингвистический корпус, использующий универсальную десятичную классификациею, может быть задан следующим набором данных:

$$K^{U.D.C.} = \{d_1, d_2, d_3, ..., d_n\},$$

где $K^{U.D.C.}$ – лингвистический корпус по УДК (например, $K^{004.8}$ – корпус текстов, относящихся к УДК 004.8 – искусственный интеллект); $d_1, d_2, d_3, ..., d_n$ – тексты, входящие в корпус $K^{U.D.C.}$.

На основе входных данных лингвистического корпуса на первом этапе предлагаемого алгоритма вычисляется характеристика *TF* (term frequency), представляющая собой отношение числа вхождения некоторого слова к общему количеству слов документа. Значимость слова $w_i$ в пределах отдельного документа может быть определена следующей *TF*-характеристикой:

$$TF(w,d) = \frac{n_i}{\sum_k n_k},$$

где $n_i$ – число вхождений слова $w_i$ в документ $d$; $\sum_k n_k$ – общее число слов в данном документе.

На втором этапе лингвистический корпус для определенного УДК представляется матрицей, имеющей следующую структуру:

$$M(K^{U.D.C.}) = \begin{pmatrix} n_{11} & n_{12} & \dots & n_{1l} \\ n_{21} & n_{22} & \dots & n_{2l} \\ \dots & \dots & \dots & \dots \\ n_{k1} & n_{k2} & \dots & n_{kl} \end{pmatrix}.$$

Элементами $n_{ij}$ этой матрицы являются частота слова в пределах отдельного документа, т.е. $TF(w,d)$. Количество столбцов и строк в этой матрице соответствует количеству слов и документов в корпусе.

Использование словаря синонимов позволяет существенно уменьшить количество элементов матрицы путем наложения частот синонимов, встречающихся в текстах. После суммирования частот синонимов снижение размерности матрицы пропорционально уменьшению количества столбцов.

Обозначив слова-синонимы символом $p$ , получаем:

$$n'_{ij} = \begin{cases} p = p_1 + p_2 + p_3 + \dots + p_n, \text{если } w_{ij} \text{имеет синонимы} \\ n_{ij}, \text{если } p = 0, \text{ синонимы отсутствуют} \end{cases}.$$

Таким образом, объединение синонимических рядов дает возможность повысить оперативность анализа электронных документов, классифицируемых по УДК.

Список литературы:

1. Мисуно И.С. Поиск текстовой информации с помощью векторных представлений [Текст] / И.С. Мисуно, Д.А. Рачковский, С.В. Слипченко, А.М. Соколов. – Проблеми програмування, 2005, №4 – С. 50–59

2. http://www.ruslang.ru/agens.php?id=text_noss2_title

## ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ МЕТОДОМ SSA С УЧЕТОМ РИСКА

*Чистякова А. А., Шамша Б. В.*

*Харьковский национальный университет радиоэлектроники*

В настоящее время эффективность функционирования предприятий существенным образом зависит от процессов получения, накопления, передачи данных и выявления в них закономерностей, которые существенно помогут построить модель прогнозирования основных технико-экономических показателей. К таким предприятиям следует отнести: банки, медицинские учреждения, предприятия телекоммуникации и связи, страховые компании, метеорологические станции и др.

В докладе основное внимание уделяется вопросам построения модели прогнозирования при помощи разрабатываемой информационной технологии, которая позволит не только оценить закономерности данных, но и формализовано подойти к выбору метода прогнозирования.

В работе отмечается, что на сегодняшний день множество методов построения моделей временных рядов имеют свои предпосылки и предположения их использования. Так, в частности, метод экспоненциального сглаживания требует стационарности ряда, метод ARIMA – стационарности временного ряда при некоторой трансформации, которая выражается взятием разности определенного порядка, методы регрессионного анализа требуют выполнение ряда своих специальных требований. Поэтому возникает проблема