УДК 519.765

# МЕТОДИКА РАСЧЕТА КОЭФФИЦИЕНТОВ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ МЕЖДУ НАУЧНЫМИ СТАТЬЯМИ

**Волков А.В.**
Российский экономический университет им. Г.В. Плеханова, г. Москва

*Поступила в редакцию 22.01.2015, после переработки 12.11.2015.*

В статье описывается методика расчета коэффициентов семантической близости между научными статьями, основанная на использовании семантического анализа текстовых элементов, а также анализе структурных элементов научной статьи. Научная статья представляется как нечеткое множество, элементами которого являются термины, где каждому термину присвоена степень принадлежности, характеризующая степень его важности для данного документа. В рамках описываемой методики предлагается новый способ расчета весовых коэффициентов терминов в тексте. Также вводится метрика измерения расстояния между научными статьями, которая учитывает неявные семантические связи между сравниваемыми документами. Для получения семантических характеристик терминов в исследовании в качестве тезауруса используется Викисловарь.

**Ключевые слова:** семантический анализ, мера близости документов, Викисловарь, семантическая сеть документа, нечеткие множества.

*Нечеткие системы и мягкие вычисления. 2015. Т. 10, № 2. С. 195–207.*

## Введение

В настоящее время развитие информационных технологий и среды Интернет приводит к ускоренному росту объема информации, которую необходимо анализировать и учитывать в практической и научной деятельности. Количество научных статей неуклонно растет, становится все труднее ориентироваться в больших массивах научной информации. Одной из актуальных задач в области компьютерной обработки научных статей является их автоматическая классификация по признаку близости тематики и научной направленности. Качественное решение этой задачи позволяет получить практически полезный эффект, например, ограничить поиск необходимой научной информации относительно небольшим подмножеством документов, выявить принципиально новые направления в науке, сократить время поиска аналогов проводимых исследований и повысить эффективность исследований за счет выявления работ, выполняемых независимо разными учеными по сходным тематикам [1].

Для того чтобы классифицировать научные статьи по признаку близости тематики необходимо научить вычислительную машину определять семантическую