МОДЕЛИ АНАЛИЗА НАУЧНЫХ ТЕКСТОВ И СОПОСТАВЛЕНИЯ АКТОРОВ НАУЧНОЙ ДЯТЕЛЬНОСТИ

М.Г. Крейнес, Е.М. Крейнес
ООО «БАЗИСНЫЕ ТЕХНОЛОГИИ»

Методология и методы объективного анализа, оценивания и сопоставления конкретных научных результатов и акторов научной деятельности (отдельных ученых, коллективов, организаций, научных журналов, конференций, издателей научной литературы и т. п.) сегодня является предметом научной и экспертной дискуссии.

Распространенные в настоящее время методы анализа, оценивания и сопоставления текстов, представляющих научные результаты, и акторов научной деятельности имеют существенные недостатки, осознаваемые всеми заинтересованными сторонами. Широко тиражируемые библиометрические и наукометрические методы (на базе которых созданы и функционируют масштабные высокодоходные международные бизнес-проекты, например, Web of Science и Scopus) оказались подверженными манипулированию (со стороны бизнеса и со стороны акторов научной деятельности) и не всегда научно обоснованы. Экспертные методы дорогостоящи, субъективны и также легко поддаются манипуляциям.

Ранее были предложены оригинальные математические модели содержания (семантики) текстов на естественных языках (русском и английском) и тематики текстовых коллекций (обзор соответствующих работ см. в [1]). На их основе построены модели содержательной близости тестов, тематической близости текстовых коллекций и характеристик содержания научного текста (оценок информативности и содержательной независимости) относительно текстовой коллекции [2]. Эти модели позволили развить методологию и методы вычислительного объективного анализа, оценивания и сопоставления научных текстов и акторов научной деятельности, которые представлены коллекциями порожденных и/или опубликованных ими текстов, сочетающие содержательный анализ текстов с данными наукометрии и библиометрии [3, 4].

Предложенные методология и методы позволяют вычислительно решать следующие задачи:

- поиск содержательно схожих текстов и выявление среди них сопоставимых научных текстов (коллекции текстов, адекватных для сравнения с анализируемым текстом),
- объективное формирование оценки текста по характеристикам содержания и библиометрическим показателям анализируемого и сопоставимых с ним текстов, выявление содержательно необоснованного («дружественного») цитирования по формальным расчетным критериям,
- оценка актора научной деятельности по объективным оценкам текстов, включенных в коллекцию, ассоциированную с актором, и выявление акторов научной деятельности, злоупотребляющих дружественным цитированием,
- поиск акторов, порождающих тематически близкие текстовые коллекции, по формируемым вычислительно моделям тематики текстовых коллекций, ассоциированных с акторами научной деятельности,
- объективное оценивание актора научной деятельности по результатам сравнения с сопоставимыми акторами.

В докладе будут приведены результаты вычислительных экспериментов с миллионом русскоязычных научных статей (2009 – 2015 г.г. из коллекции elibrary.ru) и 180 тысячами англоязычных статей, опубликованных в журнале Science с начала издания журнала до 1996 г. (данные предоставлены НЭИКОН):

- примеры адекватных коллекций для научных текстов на русском и английском языках, формируемые оценки качества и случаи выявления дружественного цитирования,
- пример текста, для которого не удалось сформировать адекватную коллекцию [5],
- примеры формирования подборок тематически близких научных журналов на русском языке и сравнения их библиометрических и наукометрических показателей.

Практическое использование разработанных методологии и методов основано на реализации оригинальных технологических вычислительных сервисов содержательного поиска, смысловой категоризации и анализа текстов и текстовых коллекций в масштабных электронных

хранилищах научно-технических документов на естественных языках. Вычислительное формирование необходимых математических моделей текстов и текстовых коллекций в настоящее время реализовано на технологическом уровне. Накоплен опыт практического технологического использования моделей для поиска содержательно близких научно-технических документов [5] и тематически близких текстовых коллекций на естественных языках (русском и английском). Разработаны и экспериментально апробированы методология и методы объективной оценки качества научных текстов, сочетающих вычислительный анализ содержания текстов и данных наукометрии и библиометрии.

ЛИТЕРАТУРА

1. Крейнес М.Г. Модели текстов и текстовых коллекций для поиска и анализа информации// Труды Московского физико-технического института (государственного университета). – 2017. – т. 9. – № 3 (35). – С. 132-142.

2. Крейнес М.Г. Методы вычислительного анализа моделей семантики для оценки качества научных текстов // Известия РАН. Теория и системы управления. – 2013. – № 2. – С. 64-75.

3. Крейнес Е.М., Крейнес М.Г. Модель управления выбором референтных коллекций для объективной оценки качества научно-технических публикаций по библиометрическим и наукометрическим показателям// Известия РАН. Теория и системы управления. – № 5. – 2016. – с. 73-89.

4. Крейнес Е.М., Крейнес М.Г. Модель управления построением оценки качества научно-технических документов на основе анализа их содержательного контекста// Известия РАН. Теория и системы управления. – № 6. – 2016. – с. 97-106.

5. Бартини Р.О. Некоторые соотношения между физическими константами// Доклады Академии наук СССР. – 1965. – т. 163. - № 4. – С. 861-864.

6. Петров А.Н., Крейнес М.Г., Афонин А.А. Семантический поиск неструктурированной текстовой информации на естественных языках в задачах организации экспертизы при реализации научно-технических программ // Информатизация образования и науки. – 2013. – No. 2 (18). – С. 54-67.

7. Петров А.Н., Крейнес М.Г., Афонин А.А. ВЫЧИСЛИТЕЛЬНЫЕ МОДЕЛИ СЕМАНТИКИ ТЕКСТОВЫХ ИСТОЧНИКОВ ИНФОРМАЦИИ ДЛЯ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОГО ОБЕСПЕЧЕНИЯ НАУЧНО-ТЕХНИЧЕСКОЙ ЭКСПЕРТИЗЫ // МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ. – 2016. – т. 28 – № 6 – С. 33-52.