УДК 00.005

Проведение предварительного семантического анализа больших массивов текстов на этапе информационного поиска

Е.А. Машина (Университет ИТМО)
Научный руководитель – к.т.н., П.В. Балакшин
(Университет ИТМО)

В работе приводится описание способа проведения предварительного семантического анализа больших массивов текстов с целью предварительной оценки их значимости при проведении информационного поиска.

Цель работы – обоснование метода оценки вида содержащейся в тексте информации на основе семантического анализа структурных частей научного материала с применением семантических шаблонов и ключевых терминов-маркеров.

Как показал опыт обработки научных источников в «прорывных направлениях», ориентироваться в чрезмерно большом количестве научных публикаций, выходивших в первые месяцы практически любых инновационных исследований, а также проводить сравнительную оценку представленных в них результатов достаточно сложно [1]. Это может быть объяснено несколькими причинами:

- значительное количество материалов, публикуемых в начальный (прорывной) период исследований содержит, как правило, большое количество не прошедших необходимых проверок гипотез,
- вновь возникающая тема научных дискуссий провоцирует большое количество исследователей всего лишь высказать собственное мнение по актуальной проблеме, не приводя сколько-нибудь существенных фактов по описываемой теме,
- существенная часть работ на начальном этапе исследований носит описательный характер и сводится к перечислению и начальной систематизации уже известной фактографии.

Однако, начальный этап исследовательских работ в новой для авторов тематике исследований практически всегда требует предварительной обработки достаточно большого количества исходной информации по теме. При этом следует отметить, что существующие методы смысловой оценки содержания текстов, описывающих инновационную область, к сожалению, не позволяют с высокой степенью вероятности отобрать наиболее значимые работы [2], поскольку автоматизированная обработка больших объемов информационных массивов инновационной информации оказывается часто затруднена в связи с отсутствием детально проработанной онтологической модели рассматриваемой проблемы, что не позволяет в полном объеме применить разработанные автоматические методы семантического анализа текстов, что в свою очередь не позволяет в автоматическом режиме ранжировать весь информационный массив работ по его научной значимости. При этом и хорошо применимые методы сравнительного ранжирования научной значимости того или иного материала, основанные на индексах цитируемости авторов [3] на начальном этапе исследований в достаточно новой предметной области исследований также не особенно применимы, поскольку маркетинговые стратегии исследовательских коллективов напрямую заставляют авторов постоянно и как можно скорее демонстрировать свое присутствие в области инновационных исследований.

Таким образом наиболее обоснованным подходом для предварительного семантического анализа научных статей на предмет принадлежности к тому или иному виду по их информационной наполненности автору настоящей работы видится разработанный им формальный структурный анализ текста научной статьи на основе существенных словоформ, семантических шаблонов и ключевых терминов-маркеров.

Суть предлагаемого метода формальной оценки состоит в проведении последовательного анализа всех существенных частей научного текста на наличие в них терминов-маркеров и словоформ, семантически предполагающих принадлежность

исследуемого текста к тому или иному информационному виду. При этом предполагается, что все виды информации, содержащейся в научных текстах, можно подразделить на следующие группы:

- описательно-регистрационные работы, содержащие описание отдельных явлений, их отношений и свойств,
- аналитические обзоры, характеризующиеся высоким уровнем обработки информации, полученной ранее,
- разработка новых концептов или описание усовершенствований существующих сущностей,
- описание исследований, посвященных созданию новых методов и устройств,
- информация, содержащая теоретическое решение проблем, на основании которых строятся новые закономерности.

В связи с тем, что научный материал представляет собой, как правило, хорошо структурированный набор текстовых данных [4]: заглавие, ключевые слова, аннотация, введение, основное содержание, заключение, для каждой из перечисленных шести структурных частей научной статьи путем предварительного анализа могут быть построены наборы ключевых терминов и словоформ, характеризующих принадлежность работы к тому или иному информационному виду.

В качестве тестового информационного массива данных для отработки предложенного метода была использована тестовая база статей, состоящая из научных работ по вирусологии и компьютерным наукам с уже известной степенью обобщения описанной информации.

Для построения семантических шаблонов использовались методы лингвистического анализа текстов [5] и алгоритмы технологии машинного перевода, основанного на правилах. Автоматизированное выделение характерных терминов и словоформ происходило с использованием технологии выявления ключевых слов, с использование технологий слеминга и частотного анализа. При этом для каждого из созданных шаблонов строился онтологический граф семантических отношений, характеризующий отношения подчиненности между ключевыми терминами-маркерами.

Проведенное тестирование описанного метода показало высокое качество проведения семантического анализа больших массивов текстов, что позволяет рекомендовать его к использованию для семантического анализа на начальном этапе информационного поиска.

Литература:

1. Schmidt GA, Girard TD, Kress JP, Morris PE, Ouellette DR, Alhazzani W, et al. Official executive summary of an American Thoracic Society/American College of Chest Physicians clinical practice guideline: liberation from mechanical ventilation in critically ill adults. Am J Respir Crit Care Med. 2017; vol. 195(1), pp. 115–119.
2. Costa R.V., Ramos A.P. Designing an AHP Methodology to Prioritize Critical Elements for Product Innovation: An intellectual capital perspective. International Journal of Business Science and Applied Management, 2015, vol. 10, iss. 1, pp. 15–34.
3. Glänzel W. Towards a model for diachronous and synchronous citation analyses // Scientometrics. 2004. Vol. 60. No. 3. P. 511-522.
4. Davis, M. Scientific Papers and Presentations, 2nd Edition// Academic Press, 2005, 384 P.
5. Priss U., Old L.J. Modelling Lexial Databases with Formal Conept Analysis // J. Univer. Comput. Sci. - 2004. V. 10, No 8. - P. 967- 984.