Инструментарий оценки инновационной активности на основе семантического анализа информационных текстов

С. В. Кулешов[1], А. А. Зайцева[2], А. Ю. Аксенов[3]
Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН)
[1]kuleshov@iias.spb.su , [2] cher@iias.spb.su,
[3] a_aksenov@mail.iias.spb.su

О. Н. Кораблева
Санкт-Петербургский государственный университет, НИУ ИТМО
on.korableva@gmail.com

О. В. Калимуллина
НИУ ИТМО
chemireva@mail.ru

***Аннотация*. В статье рассматривается оригинальный и эффективный инструментарий для оценки инновационной активности страны на основе технологий сематического анализа информационных текстов. Актуальность и научная новизна исследования обусловлена предметной областью, а также сочетанием применяемых инструментов для решения поставленных задач. Оригинальность исследования обусловлена практически полным отсутствием аналогичных работ в рамках предметной области инновационной активности, как в России, так и за рубежом. Полученный исследовательский прототип информационной системы доказал свою эффективность, что обеспечивает перспективы его применения исследователями инновационного развития и направлений экономического роста страны. Результаты исследования представляют огромный интерес для изучения динамического изменения направлений инновационной активности за период 2006–2017 гг., а также для создания прогнозов инновационного развития страны.**

Ключевые слова: семантический анализ; онтология; информационные тексты; инновационная активность

I. Введение

В контексте проведенного исследования инновационная активность страны рассматривается как комплексная деятельность по внедрению изобретений, разработок, последних научно-технических идей, а также включающая в себя совокупность мероприятий по обеспечению эффективности применения инноваций. Оценка инновационной активности на макроуровне представляет собой сложную задачу по анализу значительного объема как структурированной, так и неструктурированной информации [1].

Большинство исследований в области инновационной активности основано на анализе структурированных статистических данных, методология формирования которых определяется такими документами, как Руководство Фраскати, Руководство Осло, Руководство ЮНЕСКО и рядом других. Также на основе статистической информации формируются индексы и рейтинги инновационной активности. Тем не менее, неструктурированные текстовые данные, такие как новостные ленты, аналитические заметки, научные статьи, заявления официальных лиц и другие источники представляют собой особую специфическую базу неизученной информации. Важность работы с такого рода информацией определяется современными тенденциями перехода к качественной оценке экономических процессов. Таким образом, проводимое исследование является уникальным в своем роде применительно к предметной области анализа инновационной активности, а полученные результаты позволяют предположить, что анализ неструктурированной информации, в том числе текстов на естественном языке, обладает широкими прогностическими способностями и дает возможность выявить будущие тенденции инновационного развития страны. Применяемый инструментарий на основе онтологического моделирования и технологий семантического веб позволяет обрабатывать неструктурированную текстовую информацию, извлекая искомые данные об инновационной активности страны.

II. Исследование

Семантический анализ текстов широко применятся для самых различных задач. Так, исследователи Shanghai Key Laboratory of Financial Information Technology [2] провели семантический анализ деловых новостей для выявления тенденций рынка, осознания стратегии конкурентов и принятия инвестиционных решений. В исследовании предлагается новый подход к извлечению бизнес информации, объединяющий шаблоны, модели машинного обучения и технологию deep learning, который применяется для извлечения данных из онлайн-новостей Китая. В работе [3] была представлена разработка инструмента поиска направлений развития технологий с заданными характеристиками на

основе семантического анализа и визуализации патентной информации. Также представляет интерес применение инструментов семантического анализа текстов для решения экономических задач на примере изучения феномена социальной коммерции на основе анализа научных публикаций [4]. В статье [5] рассматривается комбинирование применения инструментария семантического поиска с нечеткой логикой. При этом методы нечеткой логики используются для обработки данных о поведении пользователей, а текстовые данные комментариев потребителей классифицируются на основе семантического анализа.

В рамках настоящего исследования был разработан Исследовательский прототип информационной системы, размещенный по адресу https://www.innoexp.ru/, и предоставляющий сервис по защищенному соединению SSL, что гарантирует доверие к сервису с точки зрения информационной безопасности.

Функциональная схема разработанной системы приведена на рис. 1.

Рис. 1. Функциональная схема работы системы

Система состоит из независимых модулей, каждый из которых обеспечивает реализацию определенных функций.

Модуль краулеров и парсеров обеспечивает автоматический сбор данных в Интернете, как всех подряд, так и на конкретную тему, а также парсинг HTML-страниц и поддержку разметки pdf, html, xls-документов, наполняя базу данных системы текстами, причем производится разделение полезного текста от служебной информации.

Ассоциативно-онтологический модуль обеспечивает динамическое формирование онтологий на основе текстов из сформированной базы данных и их визуализацию, как в автоматическом режиме, так и с учетом внешних данных (добавление новых ключевых понятий, изменение степени детализации онтологии, исключение некоторых терминов из формируемой онтологии «на ходу»). Экспертная онтология была ранее построена авторами и рассматривалась в статьях [6], [7].

Результаты работы модуля доводятся до сведения пользователя через интерактивный интерфейс, обеспечиваемый модулем WEB-сервиса. Данный модуль также обеспечивает работу созданного сервиса в сети Интернет. Функциональная схема работы модуля формирования базы данных из существующих источников для дальнейшей работы и построения онтологий приведена на рис. 2.

Рис. 2. Функциональная схема работы модуля формирования базы данных понятий и терминов

Разрабатываемая методика предполагает ассоциативное формирование онтологий в процессе работы системы, а также использование в качестве понятийного графа — графа ассоциативных связей, формируемого непосредственно на самих обрабатываемых текстах. Результаты структурной декомпозиции текста сохраняются с использованием JSON или XML-формата. Сохранению подлежит ранжирование по значимости используемых смысловых единиц (терминов или понятий) и их связей друг с другом. Пользователь может выбрать различные периоды выборки текстов из базы данных, за которые нужно получить информацию: год, месяц, день. После выбора интересующего нас периода, в котором присутствуют данные, можно просмотреть их. Для обработки выбранного объема текстов предназначены следующие команды:

- семантическое окружение – строится визуализация онтологии данной выборки текстов, с возможностью настройки степени детализации графа семантического окружения путем изменения порога фильтрации;
- тональность – система определяет положительную окраску текстов и отрицательную, в условных единицах, а затем вычисляет общий балл, показывающий, какую в целом эмоциональную окраску несет данная выборка текстов.

Необходимо учитывать, что при изменении порога фильтрации в семантическом окружении отсекаются термины, которые имеют в тексте недостаточное для данного уровня детализации количество связей. При анализе визуальной карты семантического окружения текстов за конкретный период времени можно управлять ее видом и формой, размещая термины на экране в том порядке, в котором удобно пользователю, а также убирая из онтологии понятия, которые, на взгляд эксперта, мешают пониманию общей картины. Убирать «лишние» термины из онтологии можно, зайдя в раздел «настройки» и вписав эти термины в лист дополнительных стоп-слов. В этом же разделе при необходимости можно сменить базу данных, по

которой ведется анализ. С помощью системы поддержки разметки pdf, html, xls-документов и модуля автоматизированного сбора данных были предварительно отобраны тексты по признаку их связи с областью «инновационная активность», «инновационный потенциал».

Ниже приведены результаты работы системы при обработке корпуса текстов отобранных по терминологическому ядру «инновационная активность», «инновационный потенциал» за разные годы с разной степенью детализации при визуализации семантического окружения. Так, на рис. 3 изображена визуализация семантического поля за 2006 г.

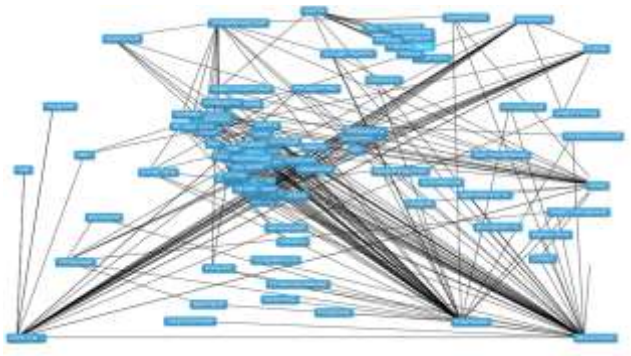

Рис. 3. Визуализация семантического поля за 2006 год, порог фильтрации 2

2006 г. – первый период, в котором появилось упоминание термина «инновационный» в контексте анализа экономических явлений. Выборка текстов мала, поэтому детализация связей между терминами очень высокая (порог фильтрации – 2). Судя по связям между терминами, инновации ассоциируются в основном с венчурными компаниями. На рис. 4 изображена детализация терминов за 2007 г.

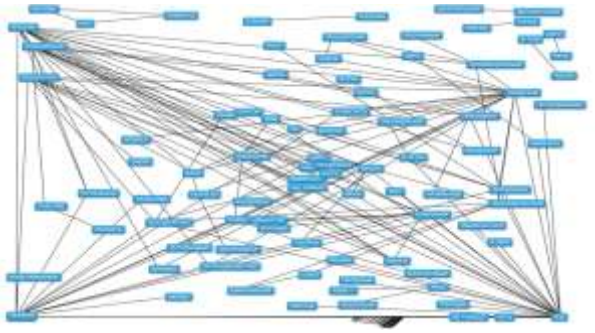

Рис. 4. Визуализация семантического поля за 2007 год

2007 год, порог фильтрации увеличен до 10 в связи со значительным увеличением объема выборки. По-видимому, в этом году понятие инновационного развития было тесно связано с экономическим развитием и с развитием международных связей России, в частности, с Казахстаном.

На рис. 5 изображена визуализация терминов за 2012 г.

При визуализации сохранены только термины, встречающиеся значительно чаще других. Количество разнородных связей у термина «инновационный» указывает на пик его популярности у экспертов. На рис. 6 представлена визуализация семантического поля за 2016 г. – текстов за этот год достаточно мало, поэтому порог фильтрации (5) с одной стороны обеспечивает высокий уровень детализации онтологии, а с другой допускает присутствие терминов, несущих общую лексическую нагрузку, например «год».

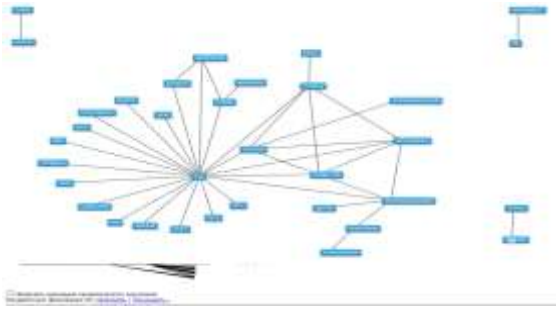

Рис. 5. Визуализация семантического поля за 2012 год., порог фильтрации 35

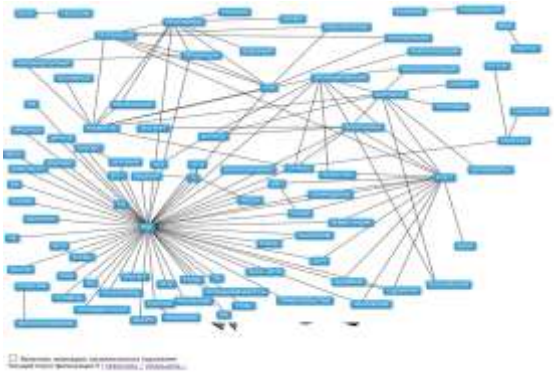

Рис. 6. Визуализация семантического поля за 2016 год., порог фильтрации 5

Одна из центральных связок онтологии показывает, что понятие инновационной продукции часто ассоциируется с различного рода программами развития, в том числе, в ВПК. Для эксперта это может означать, что, возможно, имеет смысл более подробно ознакомиться с государственными программами развития за этот год в контексте разработки инновационной продукции. На рис. 7 представлена визуализация семантического поля за 2017 год.

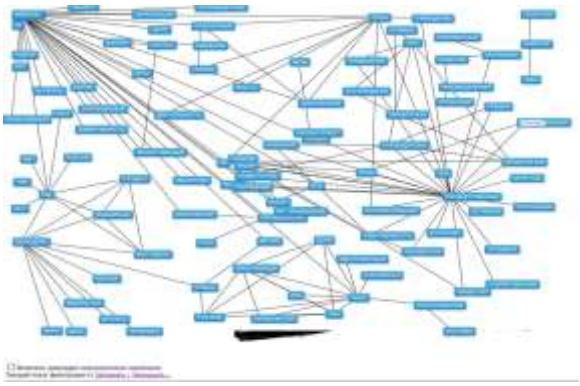

Рис. 7. Визуализация семантического поля за 2017 год., порог фильтрации 4

Текстов за 2017 этот год также достаточно мало, поэтому можно продолжать работать с семантическим окружением с достаточно большой степенью детализации (низким порогом фильтрации). В этом году, помимо ориентирования на фармакологическую промышленность в контексте инноваций, с этим понятием стали активно связывать молодежь и работу с молодежью на официальном уровне.

По результатам анализа долгосрочного мониторинга тематической области «инновационный потенциал» была сформирована база данных, в которой обобщены данные о связях и терминах по «соседним» годам (2007 год относительно 2006 года, 2008 относительно 2007 и т.д.), а также по наиболее детальным семантическим окружениям (с наименьшим порогом фильтрации). Это позволяет проследить изменение инновационного ландшафта по годам.

III. Заключение

Реализованный функционал исследовательского прототипа информационной системы представляет большой интерес для исследования инновационной активности, а именно возможность визуализации семантических карт, выявления самых употребляемых понятий, семантического окружения, а также определение тональности текста. Описанный функционал позволяет выявить основные направления инновационной активности, определить взаимосвязь понятий и явлений, корреляцию терминов в контексте с исследуемыми понятиями. Важным аспектом является анализ эмоциональной окрашенности текстов, поскольку он позволяет определить изменение отношение общества к исследуемому явлению и создает основу как для текущей аналитики, так и для прогнозирования. Таким образом, предлагаемый методологический подход и его реализация, адаптированные для исследования предметной области инновационного развития, представляют собой мощный инструмент анализа инновационной активности, обладающий, в том числе, прогностическими возможностями. Для повышения качества получаемых результатов в продолжении исследования планируется использование нескольких расширенных баз данных текстов, что приведет к их значительному увеличению, а также конкретизация отдельных направлений в рамках взаимосвязанной терминологии.

Список литературы

[1] Korableva O. N., Razumova I. A., Kalimullina O. V. (2017). Research of Innovation Cycles and the Peculiarities Associated with the Innovations Life Cycle Stages. Proceedings of 29th IBIMA Conference, Vienna, Austria, 3 - 4 May 2017, pp 1853-1862.

[2] Han, S., Hao, X., Huang, H. (2018) An event-extraction approach for business analysis from online Chinese news. Electronic Commerce Research and Applications, Volume 28, March 2018, Pages 244-260

[3] Yoon, B., Magee, C.L. (2018) Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction. Technological Forecasting and Social Change, in press

[4] Lin, X., Li, Y., Wang, X.(2017) Social commerce research: Definition, research themes and the trends. International Journal of Information Management. Volume 37, Issue 3, 1 June 2017, Pages 190-201

[5] Jiao, M.-H., Chen, X.-F., Su, Z.-H., Chen, X. (2018) Research on personalized recommendation optimization of E-commerce system based on customer trade behaviour data. Proceedings of the 28th Chinese Control and Decision Conference, August 2016, Pages 6506-6511, CCDC 2016; Rainbow Bridge Hotel Yinchuan; China.

[6] Кораблева О. Н., Митякова В. Н., Калимуллина О. В. Методология сбора данных об инновационной активности и ее влиянии на потенциал экономического роста на основе построения онтологий// Мир экономики и управления. 2018. Т. 18, No 1. С. 83–95.

[7] Кораблева О.Н., Митякова В.Н., Калимуллина О.В. Онтологическое моделирование инновационной активности и потенциала экономического роста // Вестник ВГУ. Серия: Экономика и управление. 2017. № 3. С. 160-167.