РАЗРАБОТКА ИНСТРУМЕНТАРИЯ ПОСТРОЕНИЯ ПРЕДВАРИТЕЛЬНЫХ ГИПОТЕЗ ДЛЯ ОЦЕНКИ ПРОСТРАНСТВЕННОЙ НЕОДНОРОДНОСТИ РЕГИОНОВ (НА ОСНОВЕ ИНФОРМАЦИОННЫХ ПОТОКОВ ИНТЕРНЕТ-ПРОСТРАНСТВА)[35]

Есикова Т.Н., Гордин М.С.

Новосибирский национальный исследовательский государственный университет, Россия, г. Новосибирск, улица Пирогова, 1

T.N.Yesikova@gmail.com, gordin.ms@gmail.com

Аннотация: Рассмотрены целесообразность и перспективность использования современных технологий обработки больших неструктурированных потоков данных для генерации рабочих гипотез регионального развития. Предложены алгоритмы генерации гипотез по информационным потокам интернет-пространства для оценки усиления/ослабления пространственной асимметрии и неоднородности.

Ключевые слова: построение гипотез, алгоритмы, информационные потоки, анализ интернет-пространства, территориальные экономические системы, регионы, пространственная неоднородность, асимметрия.

Введение

Глобализация экономического пространства обусловливает усиление внимание к проблемам неравномерности, асимметричности экономического развития отдельных стран и их регионов. Это сопряжено с тем, что превышение порогового уровня дифференциации развития регионов (макрорегионов, стран) имеет критическое значение не только для сохранения суверенности отдельных территориальной системы, но и стабильности глобальной экономики. Именно неоднородность экономического пространства, усиление дифференциация внутри него рассматривается как один из основных вызовов и угроз экономической безопасности [1].

Для выявления таких угроз и рисков недостаточно опираться на официально декларируемые цели и задачи стратегий и программ развития, заявленных в них темпов развития. Во-первых, потому они носят индикативный характер и редко выполняются на практике. Во-вторых, стандартные процедуры их разработки, опираются на предыдущий опыт. Это приводит к недоучету новых реалий и появления возможностей развития. В-третьих, из внимания выпадают альтернативные варианты изменения совокупных мощностных потенциалов территориальных систем.

Все это приводит к усилению неоднородности экономического пространства, обострению противоречий и возникновению кризисных ситуаций. Это можно было бы избежать, расширив возможности инструментария построения предварительных, за счет использования технологий обработки информационных потоков.

1 Общая характеристика подходов

К выявлению рисков усиления пространственной неоднородности и неравномерности развития территориальных экономических систем (макрорегионов, стран, отдельны регионов) можно подходить с различных сторон.

Эмпирические эконометрические модели. В данном случае представляют интерес работы, связанные с анализом влияния экономики знаний, которые показывают не только возможности ускорения развития территориальной системы в целом, но и локальных региональные изменения. Именно экономика знаний рассматривается как движетель на современном этапе развития.

Аппарат эмпирических эконометрических моделей панельной регрессии с фиксированными эффектами и с набором гипотез о роли разных факторов [1] ориентирован на детальный анализ влияния различных сегментов экономики знаний. Исследования базируются на данные статистики по 80 российских регионов. В частности были сконструированы и проверены гипотезы влияние на динамику роста ВРП в целом (и на душу населения) таких факторов, как

а) увеличение затрат на финансирования науки, высшего образования и здравоохранения в том или ином регионе;

б) принятие решения о перетоке финансовых ресурсов между регионами и сегментами экономики знаний (наука, высшего образование, здравоохранение);

в) принятие решения о перетоке знаний между высшими учебными заведениями регионов; и др.

[35] *Исследование выполнено при финансовой поддержке РФФИ/РНФ в рамках научного проекта № 17-02-00060-ОГН*

Такого типа исследования создают базу для построения гипотез о возможных сценариях развития страны и открывающихся возможностях для изменения позиций страны, благодаря росту ее совокупной мощи.

С другой стороны, существует мощный информационный пласт с накопленными знаниями о возможных изменениях мирохозяйственной системы в целом, отдельных территориальных экономических систем разных уровней иерархии. В нем в той или иной форме содержатся а) прогнозы официальных структур экономического профиля, б) прогнозы мозговых центров, всевозможных аналитических и исследовательских организаций, и т.п.; в) прогнозы аналитиков, экспертов, отдельных исследовательских групп; г) журналистские расследования и т.п.

Этот пласт информации может содержать ценную информацию для понимания степень угроз, всевозможных рисков для страны при сохранение или усилении неоднородности (неравномерности) пространственного развития страны. Однако необходимая информация для оценки уровня риска не всегда содержится в готовом (цифровом) виде. Ее можно извлечь из смыслового контекста, по эмоциональной окраске того или иного фрагмента информации. Все это предполагает использование специальных технологий, как для извлечения информации, так и для ее перевода в количественные параметры(«цифровизации»).

Наиболее подходящими для обработки этого мощного информационного пласта (информационных потоков) является методы и инструментарий обработки Big Data.

При этом очевидно, что интересующие нас информационные потоки не растут с такой же скоростью, как и весь поток информации в интернет-пространстве. По размерам они существенно меньше, но это все равно большие объему массивы абсолютно разнотипной информации (несколько десятков гигабайтов**)**. Так, поиск в Google по запросам о темпах роста это информации порядка 30 ГБ (темпы роста регионов России - 8 млн. результатов, темпы роста ВРП регионов России - 10 млн. результатов, темпы развития России до 2030 - 8 млн. результатов), по запросам о неравномерности и асимметричности экономического развития страны – 44 ГБ и т.п.

Со временем объем и размерность этих потоков не только не уменьшается, а, наоборот увеличивается. Эти объемы информации нужно а) обработать качественно и быстро и б) извлечь из них для анализа необходимую и достаточную информацию для поставленных целей исследований. Это требует разработки специального инструментария, ориентированного на данный тип задач.

2 Использование латентно-семантического анализа и алгоритмов анализов текстов для конструирования гипотез

Существует множество разработок, связанных с анализом текстов и довольно развитый инструментарий [2 - 7]: в том числе представляющих интерес с позиции нашего исследования: извлечение ключевых слов (AskNet, TerMine, TextAnalyst и др.), анализ тональности (BrandSpotter, Eureka Engine, TextBlob и т.п.), морфологический анализ: AskNet, RussianPOSTagger, Solarix и др.). Использование их (даже в комбинации) для целей нашего исследования затруднительно. Во-первых, потому что некоторые пакеты не имеет поддержку русского языка(TerMine), а использование других требует получение коммерческой лицензии (Eureka Engine, TextAnalyst, AskNet). Во-вторых, даже комбинированное их использование не покрывает всех стоящий на данном этапе исследований задач.

Притягательность метода латентно-семантического анализа (ЛСА) сопряжена с тем, что он во-первых, предназначен не просто для выявления значений слов, а с учетом того контекста, в котором они используются. Второе: ЛСА ориентирован на обработку больших текстовых массивов информации. При этом модель представления текста, которая задействована в латентно-семантическом анализе, воспроизводит логику восприятия этого текста самим человеком. В качестве исходной информации выступает матрица индексируемости, которая описывает частоту терминов, встречающихся в данной коллекции анализируемых документов: столбцы это документы в коллекции, строки – термины.

Для анализа инфопотока с целью извлечения из них значимой информации и конструирования на ее основе гипотез о возможных темпах развития отдельных территориальных систем предлагается следующая логическая схема:

Этап 1. Формирование перечня ресурсов для извлечения исходных данных с последующим формированием лексикографического пространства мирохозяйственной системы.

Этап 2. Приведение данных к форме, пригодной для алгоритма формирования (на базе анализа информационных потоков) множества ключевых слов, используемых для характеристики геоэкономических сдвигов.

Этап 3. Разработка алгоритма оценки влияния сдвигов мирохозяйственной системы на базе информационных потоков источника.

Этап 4. Настройка базы знаний:

- формирование набора ключевых слов (нулевого уровня) и связных с ними слов (ключевые слова 1-го и 2-го уровня), по которым можно судить о наличии в инфопотоке (статье, фрагменте статьи) информации о тенденциях развития территориальных систем;
- оценки тональности ключевых слов в контексте,
- формирования набора правил для учета зависимости тональности от контекста публикаций.

Этап 5. Оценка работоспособности формализованных алгоритмов прогнозирования.

Этап 6. Корректировка алгоритма, словарей и правил на основе результатов предыдущих этапов.

Этап 7. Проектирование гипотезы и проведение сценарных расчетов по источнику.

Для исчисления итоговой оценки учитывается семантическая близость связанных слов. на базе семантической близости. Данное множество можно сформировать на базе словарей и анализа сопоставимых корпусов текстов.

Основные предположения предлагаемого подхода следующие: слова всегда можно понять из контекста; для того, чтобы найти семантически схожие слова, необходимо найти слова, которые употребляются в таком же контексте; семантическая близость между двумя сущностями может изменяться, что обусловливается изменением корпусов текстов, а также словарей. Способ вычисления семантической близости, основан на предположении, что семантически близкие термы употребляются в одинаковых или схожих контекстах.

При расчете семантической близости может использоваться не только связность между словами в контексте, но и подходы к вычислению семантической близости, базирующиеся на вычислении расстояний между словами в известной семантической сети WordNet. Для получения более широкой картины целесообразно использование специального алгоритм анализа текстов с учетом специфики предметной области, особенностей подачи материала авторами и источниками.

Заключение

Начата разработка алгоритмов генерации гипотез по информационным потокам интернет-пространства для оценки усиления/ослабления пространственной асимметрии и неоднородности были апробированы. Предварительные экспертные расчеты подтвердили целесообразно использования методов и инструментарий обработки Big Data для построения рабочих гипотез развития территориальных систем и подходов к оценкам потенциальных рисков.

Литература

1. *Унтура Г.А., Морошкина О.Н.* Оценка динамики экономического роста: влияние компонентов экономики знания регионов РФ // XX Апрельская международная научная конференция по проблемам развития экономики и общества. 9-12 апреля 2019 г. Москва [Электронный ресурс] : Программа секций / Нац. исслед. ун-т «Высшая школа экономики», Всемирный банк. - : НИУ ВШЭ, 2019. - Сессия S-08. Наука и инновации: количественные оценки. - Режим доступа (29.04.19) [9 с.]. Электронный ресурс (pdf)
2. *Клековкина М. В., Котельников Е. В.* Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Тр. 14-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL-2012. Переславль-Залесский, 2012, с. 81–86.
3. *Пазельская А. Г., Соловьев А. Н.* Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). М.: Изд-во РГГУ, 2011. Вып. 10 (17). с. 574–586.
4. *Маннинг К., Рагхаван П., Шютце Х.* Введение в информационный поиск. М.: ООО «И.Д.Вильямс», 2014. 528 с.
5. *Балашов О.В., Круглов В.В.* Подход к извлечению продукционных правил для систем поддержки принятия решений. [Электронный ресурс]: //URL: http://www.smolensk.ru/user/sgma/MMORPH/N-12-html/borisov/balashov-2/balashov-2.htm.