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

Есикова Т.Н., Гордин М.С.

Новосибирский национальный исследовательский государственный университет

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 построению гипотез на базе анализа информационных потоков

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

С одной стороны, существуют мощные потоки статистических данных, которые собираются различными официальными органами и специализированными организациями. Это многомерные динамические потоки, регулярно обновляющиеся. Они отражают различные аспекты и факторы развития территориальных экономических систем разного уровня иерархии. На работу с этими статистическими потоками данных ориентирован подход, опирающийся на возможности эмпирических эконометрических моделей. В данном случае для нас 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.

Такого типа исследования формируют базу для построения гипотез о возможных сценариях развития страны и открывающихся возможностях, благодаря росту ее экономического потенциала и совокупной мощи государства в целом.

С другой стороны, существует мощный информационный пласт с накопленными знаниями о возможных изменениях мирохозяйственной системы в целом, о тенденциях развития различных сторон территориальных экономических систем (разных уровней иерархии).

В этом пласте информации 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.

Наиболее подходящими для обработки этого мощного информационного пласта (информационных потоков) являются методы и инструментарий обработки Big Data. При этом очевидно, что интересующие исследователей информационные потоки определенной природы 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, например, поиск в Google по запросам о темпах экономического роста это генерирует только поток ссылок в несколько десятков миллионов. Детальное ознакомление с этим информационным потоком требует обработки порядка 30 ГБ информации: по запросу «темпы роста регионов России» - 8 млн. результатов, «темпы роста ВРП регионов России» - 10 млн. результатов, «темпы развития России до 2030» еще 8 млн. результатов и так далее. Не менее внушительный поток информации по запросам о неравномерности и асимметричности экономического развития страны – 44 ГБ.

Эти объемы информации нужно не только быстро обработать, но и оперативно и максимально качественно извлечь полезную информацию для поставленных целей исследований.

Информация и знания, содержащиеся в ней, могут быть неявными и даже неожиданными, так называемые «черные лебеди», поэтому игнорировать эти информационные потоки не следует. Априори же, по названиям источников нельзя предугадать полезность той или иной статьи, например для понимания уровня рисков для стабильного и устойчивого развития страны сохранении или усилении тех или иных тенденций экономического развития, в частности сохранения неоднородности и неравномерности развития регионов.

Обработка этих слабоструктурированных и неструктурированных информационных потоков сопряжена с колоссальными затратами времени, причем не только на ее анализ, но и на отсеивание информационного мусора. Традиционный подход (чтение, анализ текстов, извлечение необходимых данных) к обработке и анализу информационных массивов таких объемов малопродуктивен и малопригоден.

Во-первых, эта информация по сути разноплановая и разноформатная (аналитические доклады, обзоры и статьи, дискуссии в журналах, статьи экспертов и аналитиков в специализированных журналах, монографии и т.п).

Во многих источниках, посвященных именно сценариям и стратегиям развития, нередко отсутствует количественный материал (явные знания) по заложенным гипотезам развития и не приводятся какие-либо конечные ориентиры развития. Это довольно типичная ситуация. Сложно получить данные о преодолении или усилении неравномерности развитии регионов России по весьма емким и объемным источникам, посвященным прогнозу социально-экономического развития на 2030 г. (более 300 страниц текста) [2] или по проекту концепции**,** рассматривающей стратегии пространственного развития страны в тот же прогнозный период времени (более 100 страниц) [3].

Во-вторых, со временем объем и размерность этих потоков не только не уменьшается, а, наоборот увеличивается. При этом появляется масса новых материалов. С еще более высокими темпами нарастает информационный мусор (многочисленные копипасты на разных интернет-источниках).

В-третьих, во многих источниках, посвященных именно сценариям и стратегиям, нередко отсутствуют явные знания (количественный материал по гипотезам). необходимая информация (например, для оценки уровня риска того или иного экономического явления) не всегда содержится в источниках в цифровом выражении, практически готовом для понимания сути проблем и возможностей их решения. Чаще всего содержится довольно много оценочной информации.

Все выше сказанное свидетельствует о том, что средств и инструментария, который традиционно используется в экономических исследованиях, для проектирования гипотез явно недостаточно. В этом случае может оказаться полезным опыт использования инструментария для обработки больших данных.

Целесообразна разработка и использование алгоритмов извлечения «описательно», «эмоциональной», «неколичественной» информации на базе таких слабоструктурированных и неструктурированных информационных потоков для конструирования гипотез в традиционном

виде для экономических исследований виде. Неявные и явные знания являются существенными компонентами при разработке таких алгоритмов.

2 Использование латентно-семантического анализа для конструирования гипотез

Метод латентно-семантического анализа (ЛСА) предназначен для выявлять значения слов с учетом контекста [4]. Документы и отдельные слова отражаются в «семантическое пространство», которое является основным для всех дальнейших сравнений. Модель представления текста довольно близка к восприятию информации, статей любым человеком при определенных допущениях. Например, каждое слово имеет только одно значение или порядок слов не принципиален.

Притягательность метода латентно-семантического анализа сопряжена с тем, что он, во-первых, предназначен не просто для выявления значений слов, но с учетом того контекста, в котором они используются.

Второе: латентно-семантический анализ 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 [4-7].

Применение латентно-семантического анализа предполагает в нашем случае следующие шаги:

- извлечение информации из электронных инфопотоков с целью формирования семантического пространства;
- обработка семантического пространства с целью удаления шумов, не представляющих значимости для последующего анализа (стэминг);
- формирование множества ключевых слов, представляющих значимость и предопределяющих целевую область;
- исчисления эмоциональной оценки элементов множества ключевых слов, формирующих восприятие потенциальных геоэкономического сдвигов;
- исчисления итоговой оценки с учетом эмоциональной окраски и сформированных ранее знаний, находящихся в системе базы знаний;
- визуализация полученных результатов с учетом алгоритма кластеризации k-means;
- пополнение базы знаний на основе полученных результатов.
- исчисления итоговой оценки.

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 (формирование инфопотока).

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 того или иного составного фактора на темпы экономического развития путем анализа тестового содержимого источников информационного потока.

Этап 4. Настройка базы знаний:

- формирование набора ключевых слов (нулевого уровня),
- сопоставление каждому ключевому слову нулевого уровня (признак фактора для построения гипотезы) 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. В качестве основанного подхода взят подход, основанный на тональных словарях, который может быть подкорректирован с учетом веса и рейтинга источника информации;
- формирования набора правил для учета зависимости тональности от контекста публикаций.

Основные предположения при настройке база знаний предлагаемого подхода следующие: слова всегда можно понять из контекста; для того, чтобы найти семантически схожие слова следует найти слова, которые употребляются в таком же контексте; и др.

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.

Рассмотрим более подробно этап формирования инфопотока. Признано целесообразным предварительная подготовке информационного потока, с учетом конкретных целяй исследования.

Инфопоток, которые представляются наиболее продуктивным и полезным для анализа при построении гипотез оценки пространственной неоднородности регион, задается через список адресов html-страниц, Формирование единого инфопотока обеспечивается путем последовательной загрузки html-страницы с того или иного сайта с помощью средств библиотеки urllib.

Тело сайта преобразуется в дерево элементов. Поиск по заданным элементам осуществляется через XML Path Language (xpath). В нашем случае XML Path Language это наиболее удобный метод запроса к дереву XML документов. Древовидная структура XML позволяет легко извлекать нужный элемент. При этом на выходе имеет пару вида address - xpath (адрес электронного ресурса – элемент), которая используется при извлечении текста для анализа.

Каждый сайт имеет собственную структуру. Для того чтобы формализовать и унифицировать правила парсинга для ручного режима предложен следующий алгоритм:

1) На вход подается адрес электронного ресурса.
2) В поле инструментария выводится html-составляющая указанного ресурса, которая позволяет выбрать и установить xpath-элемента, необходимого для извлечения текстовой информации на данном сайте.
3) Осуществляется извлечение текстовой информации по xpath-элементу в html-составляющей с последующей записью.
4) По извлеченному тексту проводится поиск возможных факторов для построения гипотезы (ключевое слово 0-го уровня в базе знаний) через регулярное выражение regular'[А-Я]+\b'. Для этого из базы знаний извлекается очередное поисковое слово-фактор и ведется его поиск в связных фрагментах текста.

Если выбранное ключевое слово найдено, то далее следует извлечение из базы знаний, сопряженных с ним слов (ключевые слова уровня 1 и 2), значимы для построения гипотезы.

Эти слова используются для поиска в выявленных фрагментах текста. При наличии вхождения их в соответствующий фрагмент используются алгоритм для генерации эмоциональной или смысловой оценки ключевого слова 0-го уровня

Для сокращения избыточности инфопотока, производится исключение из рассмотрения все слова, длиной меньше трех символов.

5) Из выбранного xpath-поля берется текст и записывается в поле результата.
6) Для формирования матрицы латентно-семантического анализа используются обновленный список факторов, по которым осуществляется анализ.

В связи с тем, что в нашем случае ценность представляет как построение гипотезы по данным какого отдельного источника (наиболее значимого для исследователя источника), так и сформированной совокупности источников.

Это решается путем введения двух режимов обработки информационного потока. В автоматическом режиме поиск осуществляется по всему списку выбранных электронных ресурсов. В ручном режиме поиск осуществляется по указанному адресу и записывается в историю поиска.

Этап разработка алгоритма оценки влияния того или иного составного фактора на темпы экономического развития путем анализа тестового содержимого источников информационного потока.

Множество всех слов, встречающихся в данной совокупности предложений, образует множество вершин графа. Ребра графа отображают факт непосредственного следования одного слова за другим. Традиционно направление ориентации каждого ребра задается от последующего слова к предыдущему. Вес ребра — это характеристика количества вхождений данной пары слов, следующих друг за другом в данном порядке, в весь текст S.

Вес каждого слова можно определить по формуле:

$$(2) \quad S(V_i) = \frac{1-\lambda}{N} + \lambda \sum_{V_j \in IN(V_i)} \frac{S(V_j)}{|OUT(V_j)|},$$

$S(V_i)$ – ранг вершины (вес слова) V_I;

$S(V_J)$ – ранг вершины (вес слова), V_J из которой направлена связь в вершину V_I;

$|OUT(V_0)|$ – количество потомков вершины V_J;

N– количество вершин графа;

λ – коэффициент затухания (dampingfactor), в формуле используется фиксированная величина, равная 0,85.

Таким образом, можем вычленить из контекста того или иного информационного источника наиболее важные ключевые слова, эмоциональная оценка которых отражает влияние того или иного фактора на изменение темпов развития территориальных экономических систем.

Сходство или различие между объектами классификации устанавливается в зависимости от выбранного метрического расстояния между ними. Если каждый объект описывается N свойствами (признаками), то он может быть представлен как точка в N-мерном пространстве. Такой подход, при котором каждое предложение или слова рассматривают как точка N-мерного пространства, называют дистрибутивные (распределенные) методы измерения семантического расстояния между словами или предложениями. В качестве расстояний или меры близости используется Евклидово расстояние.

Алгоритм исчисления итоговой оценки. Исследование семантической близости термов — это неотъемлемая и важная часть теории обработки текстов на естественном языке. Семантическая близость между двумя сущностями с течением времени может изменяться, что обусловливается изменениями корпусов и словарей. Как было отмечено выше, поддержание словарей в актуальном состоянии является достаточно сложной задачей, потому что предметные области меняются и очень часто создаются новые слова и новые смыслы приписываются новым словам.

Для вычисление семантическая близость термов может быть также использована специализированных баз данных и статистических корпусов, так же огромное число современных подходов к вычислению семантической близости основано на вычислении расстояний между словами в известной семантической сети WordNet.

3 Использование алгоритмов анализов текстов для конструирования гипотез

Существует множество разработок, связанных с анализом текстов и довольно развитый инструментарий [4-9]: 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 по следующим моментам.

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 использование этих алгоритмов не позволяет решить задачи, стоящие на данном этапе исследований

Для построения гипотезы о возможных темпах развития территориальных экономических систем разного уровня иерархии исчисляется эмоциональная оценка влияния фактора, генерируя на базе анализа соответствующих фрагментов текста, которая впоследствии нормируется:

$$(3) \quad S_j = Q_i \frac{\sum_1^{\tau} s_{ij}}{n}$$

где

S_j – нормированная оценка влияния фактора j на анализируемый параметр (темп роста, темп прироста и др.),

S_{ij}- эмоциональная оценка влияния фактора j на базе анализа всех сопряженных фрагментов статье i;

Q_i - вес данного источника i-го информации источника из информационного потока.

В связи с тем, что оценки могут быть отрицательными, то для построения гипотез об изменениях темпов развития (или других параметров, влияющих на экономический потенциал или

на мощь государства) на данном этапе исследований используется следующая последовательность действий:

Если в исходном информационном потоке найдены количественные оценки факторов, влияющих на темпы развития, то выдается картеж этих значений по каждой территориальной экономической системе.

Если же в процессе анализа исходного информационного потока сгенерированы только эмоциональные оценки, то при построении гипотез о возможностях изменениях учитывается возможность увеличение темпов роста (от 0,1% до 0,5%);при положительной итоговой эмоциональной оценке фактора, о снижении темпов рост от - 0,1 до -2 при значения ниже нуля (по определенной шкале).

Этот алгоритм один из возможных. Ибо предложить универсальный алгоритм для измерения сложно, так как информация, сведения и знания истолковываются индивидом (определенным слоем общества) исходя из личной точки зрения. Они не беспристрастны, ибо все подчиняется определенными предположениями, предубеждения относительно того или иного аспекта жизни, т.е. пропускается через призму личных представлений.

Следует учитывать, что автор той или иной статьи «выхватывает» из всего фактов, нередко только те свидетельства, источники которые укладываются в его мировозрение и представление о возможном ходе развития событий и работают не его гипотезы о будущем развитии событий. То, что противоречит, как правило, иногда на подсознательном уровне исключается из рассмотрения.

Можно использовать и иные подходы при построении гипотез о влиянии тех или иных факторов на значимые параметры развития территориальных экономических систем.

Например, использовать инструментарий, работающий на базе сформированных наборов правил, для формирования представления о тональности текста того или иного источника. В самом лучшем случае этот подход потребует только больших временных затрат. Необходимо прописать в той или иной форме весьма представительный набор правил, которые смогли бы полностью покрыть все пространство эмоциональных оценок.

Это означает, что сначала требуется вручную проанализировать большой тестовый информационный поток. Сформировать на основе его анализа первоначальный набор правил. Потом сформировать несколько отладочных информационных потоков, на примере которых провести тестировании предлагаемого набора правил. После этого при необходимости провести корректировку правил и повторить эти итерации до получения приемлемого результата.

При этом для каждого фактора, который может знаково сказать на судьбе территориальных экономических систем необходимо эту настроечную работу проводить индивидуально.

Третий подход - машинное обучение с учителем. Этот тип чаще всего используется при проведении исследований. В этом случае обучение машинного классификатора осуществляется на наборе заранее размеченных текстов. После чего полученную модель используют для анализа новых текстов или документов.

Заключение

Начата разработка алгоритмов генерации гипотез по информационным потокам интернет-пространства для оценки усиления/ослабления пространственной асимметрии и неоднородности. 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. Прогноз долгосрочного социально–экономического развития Российской Федерации на период до 2030 года. [Электронный ресурс]. //URL: http://static.government.ru/media/files/41d457592e04b76338b7.pdf
3. Проект концепции Стратегии пространственного развития Российской Федерации на период до 2030 г. Москва, 2016. [Электронный ресурс] Kontseptsiya_SPR.pdf

4. Методы реализации продукционных правил [Электронный ресурс]. //URL: http://homehelper.in.ua/blog/expertsystems/110.html.
5. Using latent semantic indexing for information filtering / P. Foltz // In ACM Conferenceon Office Information Systems (COIS), 1990. –PP. 40–47.
6. *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.
7. *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.
8. *Маннинг К., Рагхаван П., Шютце Х.* Введение в информационный поиск. М.: ООО «И.Д.Вильямс», 2014. 528 с.
9. *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.