УДК 004.912

Оценка информационных потоков интернет пространства для прогнозирования геоэкономических сдвигов МХС на основе латентно-семантического анализа

Девочкин Юрий Вадимович
Новосибирский государственный университет
Бакалавр

Есикова Татьяна Николаевна
Институт экономики и организации промышленного производства
Канд. экон. наук, ведущий научный сотрудник

Аннотация
Рассматриваются возможности метода латентно-семантического анализа для получения знаний из информационных потоков интернет пространства о будущих вариантах мироустройства. Предложен алгоритм формирования базы знаний на основе эмоциональной окраски текста с учетом сформированной матрицы индексируемости геоэкономических субъектов.
Ключевые слова: латентно-семантический анализ, мирохозяйственная система, алгоритм формирования базы знаний

Evaluation of information flows of the Internet space to predict geo-economic changes of MHS based on latent-semantic analysis

Devochkin Yurii Vadimovich
Novosibirsk State University
Bachelor

Yesikova Tatyana Nikolaevna
Institute of Economics and Industrial Engineering
Candidate of Economics

Abstract
We are consider the possibilities of the method of latent-semantic analysis for obtaining knowledge from information streams of the cyberspace about variants of the future world order . We are propose an algorithm for forming a knowledge base based on the emotional coloring of a text, taking into account the formed matrix of index ability of geo-economic subjects.
Keywords: latent-semantic analysis, geo-economic system, knowledge base formation algorithm

Введение

Продуктивность прогнозирования экономических процессов предопределяется не только развитостью экономико-математического аппарата, адекватностью его применения к проблемам того или иного уровня, но и качеством, надежностью исходной информации. В настоящее время особую роль начинает играть информация о будущих сценариях мироустройства. Ибо, наблюдаемое переформатирование мирового пространства приводит не только к формированию новых центров влияния, разнообразных коалиций стран (Чимерика, БРИКС, ИРИ, Меркосур и т.п.), усилению или ослаблению внешнеэкономических связей между ними, и, соответственно, предопределяет их возможности к развитию, усилению или ослаблению экономической мощи государств и т.п.[5].

Данный тип задач сопряжен с двумя моментами. С одной стороны, это гигантские информационные потоки, находящиеся в динамике. С другой стороны, это отсутствие инструментария, позволяющего оценить эти информационные потоки, их продуктивность и весомость для прогнозирования трансформации экономического активного пространства Азиатской России.

Существующие методы анализа и сбора информации должным образом не могут отразить существующие проблемы и ситуацию в мирохозяйственной системе. И создает предпосылки для принятия неверных решений, которые могут иметь негативное влияние на экономическую систему, вплоть до ее разрушения. Частично, эти процессы наблюдаются и проявляются на мировой арене, например во взаимоотношения России, Китая, США и некоторых стран первой двадцатки. Все эти экономические субъекты создают окончательные контуры будущего расклада сил в мировой системе, предопределяемые множеством факторов (стратегические цели и устремление их транснациональных компаний, экономический потенциал, военная мощь государств и др.).

При этом судить о возможном развитии событий можно, ибо в информационном пространстве необходимой прогнозной информации более чем достаточно. Другое дело, что эти информационные потоки характеризуются большим и дублирующим объемом информации. Кроме того все события, происходящие в МХС находятся в динамическом состоянии и являются неупорядоченными и представляют собой набор не структурированных фактов, знаний, аналитических обзоров, прогнозов, экспертных оценок и т.п.

Это типичная неформализованная или плохо формализованная задача [3].Например, поиск в Yandex только по трем поисковым запросам (мирохозяйственная систем, варианты мироустройства, БРИКС) формирует информационный поток объемом минимум в 60ГБ (соответственно, 70 млн. результатов (23 ГБ),18 млн. результатов (6 ГБ), 108 млн. результатов (36 ГБ)).Обработать «в ручную» такой объем информации и сделать по ней полновесные и корректные выводы практически невозможно. Это предопределяет необходимость разработки методического подхода и

инструментария, использующего наработки методологий формирования знаний. А именно, разработать инструментарий позволяющий взаимодействовать с интересующими нас ресурсами и извлечением полезной информации из инфопотоков. Под инфопотоком понимается перечень электронных ресурсов интернет пространства, которые целесообразно проанализировать для решения выше поставленной задачи.

Анализ существующих решений

Представляется, что для разработки такого инструментария необходимо использование следующих существующих инструментов.

SQL– язык структурированных запросов, позволяющий работать с базами данных. С помощью SQL можно создавать и модифицировать данные, а управлением массива данных занимается соответствующая система управления базами данных.

NoSQL– совокупность подходов, направленных на реализацию базы данных, имеющих отличия от моделей, используемых в традиционных, реляционных СУБД. Их удобно использовать для динамической, постоянно меняющей не только по составу стран, но и ключевыми атрибутами мирохозяйственной системы.

MapReduce– модель распределения вычислений, которая предназначена для распараллеливания вычислений над структурами данных большого объема (петабайты* и более).

SAP HANA– обеспечивает высокую скорость обработки запросов, что привлекательно при работе с большеразмернымиинфопотоками данных.

Storm — это система обработки инфопотоков с открытым исходным кодом, главным отличием которой состоит в распределенной обработке данных в режиме реального времени и ее не зависимость от языка программирования.

Реализация

В соответствии с поставленной задачей предположено следующее решение, отображенное на рис. 1.

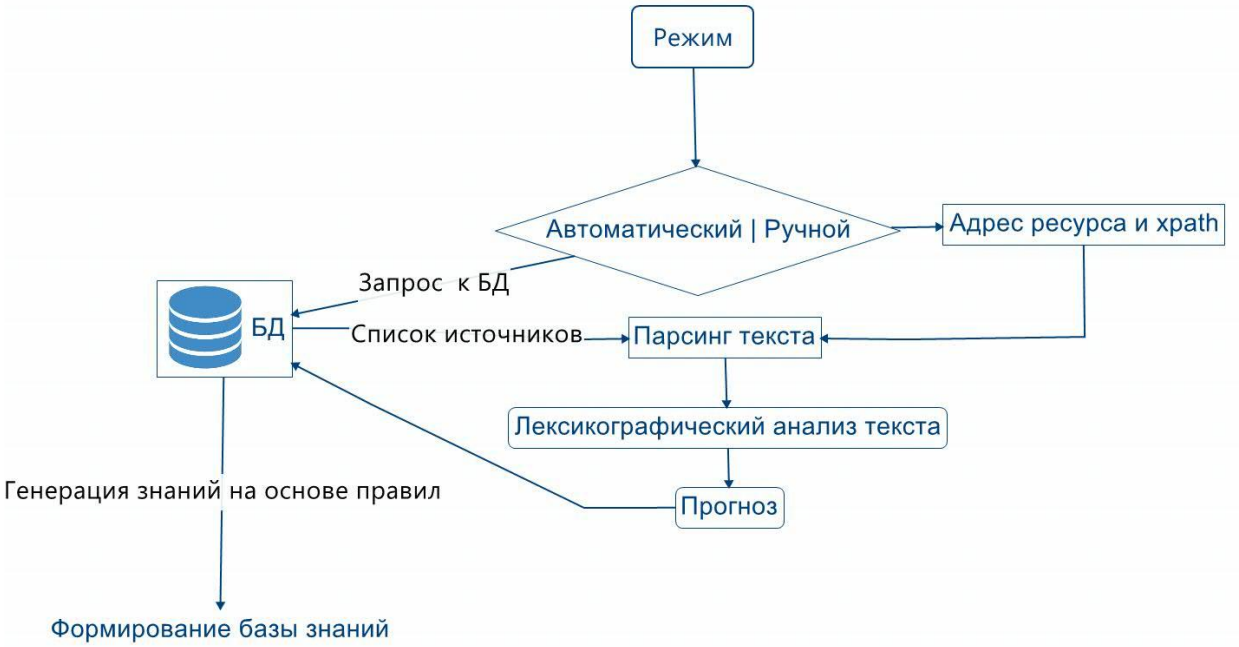

Рис. 1 Блок схема алгоритма работы инструментария (block.jpg)

Для прогнозирования геоэкономических сдвигов МХС (на базе латентно-семантического анализа инфопотоков) реализована следующая логическая схема:

1. Спецификация (выбор) режима обработки инфопотока интернет пространства.
2. Формирование инфопотокаизинтернетпространства по заданным ресурсам (сайтам) с учетом ключевых слов.При формировании запроса необходимо учесть а) требования к извлекаемой информации, характеризующие описания различных вариантов мироустройства и отдельных его компонент, субъектов мирохозяйственной системы, б) перечень необходимых атрибутов.
3. Проведение латентно-семантического анализа информационного потока.
4. Генерация полученных знаний на основе результатов латентно-семантического анализа.
5. Пополнение базы знаний с учетом структуры базы данных.
6. Визуализация полученных результатов и формирование отчетов по запросу.

Рассмотрим более подробно шаги логической схемы.

Шаг 1. Выбор режима обработки инфопотока интернет пространства

Необходимость данного шага обусловливается тем, что сложно заранее точно и однозначно определить те сегменты интернет пространства, в которых может содержаться необходимая для данного исследования информация. Основная причина, по которой те или иные виды «больших данных» не работают, как раз и заключается в элементарном отсутствии какой-либо минимально выстроенной системы сбора этих данных[1].

Для того, чтобы иметь возможность выбирать необходимые сегменты интернет пространства, предусмотрено два режима парсинга: ручной и автоматический. Ручной режим позволяет в интерактивном режиме указывать перечень электронных ресурсов, которые необходимо обработать (произвести парсинг текста), с последующим извлечением интересующей информации. Адрес ресурса задается в специальном поле инструментария, по которому проводится парсинга.В автоматическом режиме выбор перечня необходимых электронных ресурсов осуществляется через запрос к БД.

Шаг 2.Формирование инфопотока

Для формирования инфопотока необходимо загружать html-страницы соответствующих сайтов. Загрузка html-страницы реализована с помощью библиотеки urllib. Тело сайта преобразуется в дерево элементов и поиск по заданным элементам осуществляется через xpath.По нашему мнению xpath это наиболее удобный метод запроса к дереву XML документов. XML имеет древовидную структуру, что позволяет уникально извлекать нужный элемент. На выходе получаем пару вида address (адрес электронного ресурса) – xpath (элемент), которая используется при извлечении текста для анализа.

Поскольку ценность представляет анализнак отдельного источника, так и всей совокупности источников (по которым поиск осуществлялся ранее), то предусмотрено два режима выбора ресурсов для анализа – автоматический и ручной. В автоматическом режиме поиск осуществляется по списку электронных ресурсов указанных ранее. В ручном режиме поиск осуществляется по указанному адресу и записывается в историю поиска.

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 мирохозяйственных объединений через регулярное выражение regular'[А-Я]+\b'. Для сокращения избыточности инфопотока, считаем необходимым отсеивать) все слова длиной меньше двух символов и б) не строчные слова (не являющиеся аббревиатурой). Извлеченный список формирований используется для пополнения перечня стран первой двадцатки.
5) Из выбранного xpath-поля берется текст и записывается в поле результата.
6) Для формирования матрицы латентно-семантического анализа используются обновленный список геосубъектов, по которым осуществляется анализ.

Шаг 3. Латентно-семантический анализ инфопотока

Реализованный в инструментарии алгоритм латентно-семантического анализа[6, 7] включает в себя следующие этапы: генерация матрицы индексируемых слов по ключевым словам и текстам, выбранными для анализа на шаге 1; выделение связных групп; «эмоциональная» оценка взаимоотношений внутри групп [2,4].

Для демонстрации работоспособности алгоритма используется пример отладочного информационного потока, который включает в себя три текста из интернет-пространства: «БРИКС призывает США ратифицировать документы по реформам валютного фонда», «Китай отметил 60-летие дипотношений с Россией», «НАТО поблагодарило Украину и Россию».

Составление частотной матрицы индексируемых слов для отладочного примера. В этой матрице строки соответствуют экономическим субъектам (страны, коалиции, формирования и др.), а столбцы – атрибуты текстовой информации (заголовки публикаций, смысловые фрагменты текста, авторство парадигмы и т.п.). В каждой ячейке матрицы указывается какое количество раз субъект удовлетворяет следующему условию f(sub) = 1, если wf> 0.

Итоговая матрица индексируемости содержит значения 1 и 0, показывающие наличие или отсутствие соответствующих субъектов в заданном инфопотоке, который в нашем случае состоит из трех инфопотоков: инфопоток 1 – «БРИКС призывает США ратифицировать документы по реформам валютного фонда», инфопоток 2 – «Китай отметил 60-летие дипотношений с Россией», инфопоток 3 – «НАТО поблагодарило Украину и Россию».На рисунке 2 представлена полученная матрица индексируемости на основе выше перечисленных инфопотоков.

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Рис.2 Матрица индексируемости

Шаг 4 – 6. Генерация полученных знаний и формирование базы знаний

Следующим шагом является формирование базы знаний о представлении взаимоотношений между геосубъектами мирохозяйственной системе. В нашем случае, база знаний содержитперечень правил вывода, определяющих «эмоциональную окраску» о взаимоотношениях внутри выделенных групп. Следует воспользоваться результатами латентно-семантического анализа, дополнив его «эмоциональной окраской».

Для этого необходимо преобразить текст и отдельные слова в так называемое «семантическое пространство», используемого для дальнейшего сравнения. При этом нами делаются следующие предположения:

Предположение 1. Взаимоотношения внутри групп между геосубъектами можно определить через «эмоциональную окраску», задаваемую определенными наборами слов.

Предположение 2. Семантическое значение текста, посвященного геосубъектам, определяется набором ключевых слов, которые задаются при поиске.

Предположение 3. Знаковая информация содержится не только по странам первой двадцатки, но и в аббревиатурах, за которыми стоят геоэкономические объединения.

После этого осуществляется сингулярное разложение полученной матрицы на три составляющих. Исходную матрицу M представляем в следующем виде:

$$M = U*\sum*V^t$$

где U и V^t– ортогональные матрицы, а $\sum$ – диагональная матрица. Диагональные элементы матрицы $\sum$ упорядочены в порядке убывания.

$$M = U.\Sigma.V^*$$

where

$$M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

$$U = \begin{pmatrix} 0.127737 & 0.632456 & -0.763992 & 0 \\ 0.484288 & 0.632456 & 0.604537 & 0 \\ 0.612025 & -0.316228 & -0.159454 & -0.707107 \\ 0.612025 & -0.316228 & -0.159454 & 0.707107 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 2.1889 & 0 & 0 \\ 0 & 1.41421 & 0 \\ 0 & 0 & 0.45685 \\ 0 & 0 & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} 0.559207 & -0.447214 & -0.69806 \\ 0.279604 & 0.894427 & -0.34903 \\ 0.780454 & 0 & 0.625213 \end{pmatrix}$$

Рис. 3 Результаты сингулярного разложения

Анализ результатов расчетов на базе инструментария позволяет выделить сформированные группы, отображающие степень связи между геосубъектами (рис. 4). Первая группа представлена США и БРИКС (красные точки). Вторая группа – Китай, Россия, Украина, НАТО. В свою очередь, эти геосубъекты были сгенерированы в процессе анализа на основе трех инфопотоков.

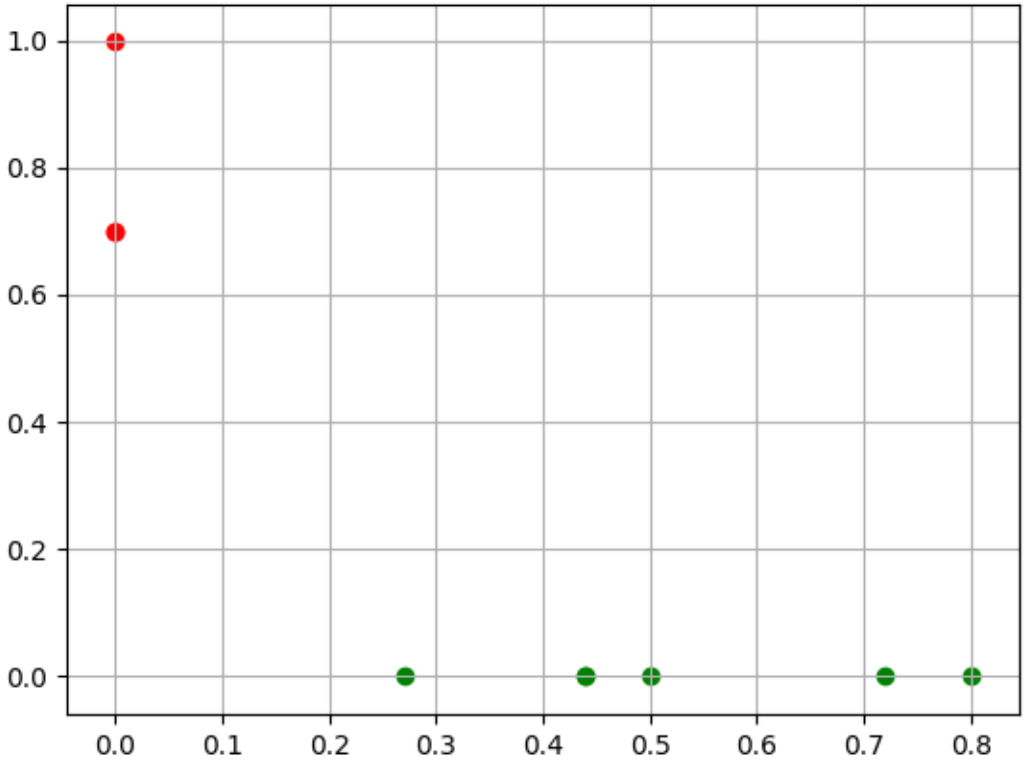

Рис. 4 Графическое представление результатов латентно-семантического анализа

Это позволяет приступить к формированию базы знаний. Для формирования которой необходимо производить автоматическое доказательство (вывод) на основе приближенных результатов, которые были получены в ходе латентно-семантического анализа. Для определения возможной ситуации внутри группы текст для каждого геоэкономического субъекта прогоняется через PageRank Algorithm [8], через который высчитывается вес для каждого слова. На вход подается список предложений $S =< S_1, S_2, ..., S_n >$ которым сопоставляется граф. Множество всех слов, встречающихся в данной совокупности предложений, образует множество вершин графа. Ребра графа отображают факт непосредственного следования одного слова за другим. При этом направление ориентации ребра задается от последующего слова к предыдущему. Вес ребра – это количество вхождений данной пары слов, следующих друг за другом в данном порядке, в весь текст *S*. Вес каждого слова можно определить по формуле:

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

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

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

$|OUT(V_j)|$ – количество потомков вершины V_j;

N – количество вершин графа;

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

После этого текст прогоняется через словарь, содержащий список слов имеющий позитивный и негативный оттенок. Далее на основе веса каждого предложения в тексте, высчитывается суммарная оценка.

Полученные данные хранятся для последующей работы с ними.В инструментарий была интегрирована база данных Sqlite, которая базируется на файле и предоставляет довольно широкий набор инструментов для работы с ней, по сравнению с сетевыми СУБД.

Данные в базе данных хранятся в следующем виде:

- ID – уникальный идентификатор,
- Structure – структура в который упоминается та или иная организация, государство, структура,
- Text – текст публикации или статьи,
- Author – автор публикации,
- Positive_Evaluation– позитивная эмоциональная оценка,
- Negative_Evaluation– негативная эмоциональная оценка,
- Neutral_Evaluation– нейтральная эмоциональная оценка.

Апробация алгоритма на условном примере продемонстрировала работоспособность предлагаемого методического подхода и разработанного инструментария.

Заключение

Как показали проведенные исследования, метод латентно-семантического анализа продуктивен для получения знаний о будущих вариантах мироустройстваиз информационных потоков интернет пространства.Разработанный инструментарий позволяет а) формировать запросы в интерактивном режиме для анализа инфопространства в соответствии с целями исследования; б) осуществлять анализ данных с выделением связных групп геосубъектов на базе латентно-семантического анализа, основанного на матрице индексируемости; в) визуализировать полученные кластеры; г) формировать базу знаний по «эмоциональной окрасу».

Библиографический список

1. Батура Т. В. Формальные методы определения авторства текста // Вестн. НГУ. Серия: Информационные технологии. 2012. Т. 10, вып. 4. С. 81–94.
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. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2014. 528 с.
4. 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.
5. Поляков И.В., Соколова Т.В., Чеповский А.А., Чеповский А.М. Проблема классификации текстов и дифференцирующие признаки // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2015. Т. 13. № 2. С. 55–63.
6. Discourse Processes / T. Landauer, P. Foltzand D. Laham // An introductionto Latent Semantic Analysys, 1998. Т.25. 1998. С.259–284.
7. Using latent semantic indexing for information filtering / P. Foltz // ACM Conferenceon Office Information Systems (COIS), 1990. С. 40–47.
8. Батура Т.В. Методы повышения эффективности поиска информации на основе синтаксического анализа: моногр. / Т.В. Батура Ф.А. Мурзин, А.А. Перфильев, Т.В. Шманина; Рос. акад. наук, Сиб. отд-ние, Ин-т систем информатики им. А.П. Ершова. Новосибирск: Изд-во СО РАН, 2014. 76 с.