OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822: 004.912

СЦЕНАРНЫЙ ПОДХОД ПРИ ИССЛЕДОВАНИИ ДИНАМИКИ ИНФОРМАЦИОННЫХ ПОТОКОВ В СЕТИ ИНТЕРНЕТ

Додонов А.Г., Ландэ Д.В., Бойченко А.В.

Институт проблем регистрации информации НАН Украины, г. Киев, Украина

dodonov@ipri.kiev.ua, dwlande@gmail.com, boychenko.a@gmail.com

Рассмотрен сценарный подход к исследованию динамики информационных потоков при анализе и прогнозировании социальных явлений. Определены и подробно описаны этапы информационно-аналитического исследования на основе анализа контента сети Интернет.
Ключевые слова: сценарный подход; динамика информационных потоков; контент-мониторинг; аналитические исследования.

Введение

Сценарии или сценарные модели является одним из видов логико-лингвистических моделей, предназначенных для отображения развернутых во времени последовательностей взаимосвязанных событий, операций или процессов. Сценарии могут иметь структуры, в которых установлены условия перехода к той или иной частной стратегии, или просто отражены возможные альтернативы без указания условий. Требование взаимосвязанности в сценарных моделях не является строгим и носит достаточно условный характер, так как устанавливается на основе субъективных суждений экспертов, а также определяется спецификой формулировки целей деятельности.

Понятие сценарной модели более широкое, чем понятие алгоритма. При этом сценарная модель накладывает менее суровые ограничения на характер причинно-следственных отношений.

Типовой сценарий аналитической деятельности состоит из следующих элементов:

– краткое описание сути отрабатываемой в сценарии функциональной задачи;
– ресурсы (участники сценария, базы данных, знаний, программы);
– операции, выполняемые в ходе сценария;
– этапы выполнения функциональной задачи;
– шаги сценария (включая временные требования);
– содержание действий участников на кожном шагу сценария;
– описание условий выполнения сценария;
– описание ограничений на выполнение;
– ссылки на экранные формы, иллюстрирующие содержание действий;
– ссылки на фрагменты документации, поясняющие содержание шага сценария;
– комментарии.

Сценарий аналитической деятельности можно представить в виде кортежа (S, P,R, E, Z), где:

– S – множество состояний системы,
– P – множество параметров,
– R – множество ресурсов,
– E – множество переходов между состояниями,
– $Z \subseteq S$ – множество результатов выполнения сценария.

Для построения сценария аналитик, используя средства реализации сценариев (обычно, АРМ некоторого моделирующего комплекса), выполняет обход графа, описывая действия, предписанные его дугам и, таким образом, решает задачу формирования последовательностей взаимодействия между участниками.

В настоящее время контент сети Интернет образует значимый динамический сегмент информационного пространства, информационные потоки, содержание и объемы которых необходимо учитывать при проведении аналитических исследований практически в любой области. Основным объектом анализа при этом являются событийные или тематические срезы этих потоков – массивы информационных сообщений, документов, соответствующих определенным событиям или тематикам. Динамика информационных потоков определяется комплексом как внутренних, так и

внешних нелинейных механизмов, которые отражаются, возможно, в неявном виде. Зачастую удовлетворительным оказывается упрощенное понимание информационного сюжета как некоторой зависимой от времени величины $n(t)$, поведение которой описывается нелинейными уравнениями [Додонов и др., 2009]. Таким информационным потокам можно ставить в соответствие временные ряды, для анализа которых все чаще обоснованно применяются формальные методы: статистического, фрактального, Фурье или вейвлет-анализа.

Постановка задачи

Для эффективного проведения информационно-аналитических исследований на основе анализа контента сети Интернет (а точнее ее веб-ресурсов) авторами предлагается последовательность шагов, этапов обработки информации, каждый из которых сам по себе обеспечивает получение продукта. Совокупность таких этапов, базирующихся на использовании необходимых и доступных инструментальных средств, специальных приемов, можно рассматривать как методику, процедуру проведения действий, нацеленных на получение аналитических материалов, которые могут использоваться для поддержки принятия решений.

Любая методика рассчитана на решение конкретных задач. При проведении информационно-аналитических исследований на базе интернет-контента к таким задачам можно отнести:

1. Нахождение релевантных публикаций по тематике.
2. Определение динамики тематических публикаций.
3. Определение критических точек в динамике тематических публикаций.
4. Выявление объектов мониторинга.
5. Выявление и визуализация взаимосвязей событий и объектов мониторинга, а также объектов мониторинга между собой.
6. Прогнозирование развития событий.

Этапы информационно-аналитического исследования

В соответствии с приведенными выше задачами предлагается выделить следующие этапы информационно-аналитического исследования:

1. Выбор системы интеграции интернет-документов.
2. Формирование запроса в среде выбранной системы. Нахождение тематических публикаций по запросу с помощью систем контент-мониторинга.
3. Определение динамики тематических публикаций по запросу.
4. Определение критических точек в динамике тематических публикаций.
5. Определение основных событий в критических точках.
6. Выявление объектов мониторинга.
7. Выявление и визуализация взаимосвязей.
8. Прогноз развития событий.

Рассмотрим эти этапы подробно.

Выбор системы интеграции интернет-документов

Для получения репрезентативной информации об объекте исследования необходимо воспользоваться поисковой системой с аналитическими функциями, охватывающей достаточный объем информации, относящейся к исследуемому объекту/событию. Для анализа динамики информационных потоков необходимо каким-то образом получить соответствующую статистику, представленную в виде временных рядов. Многие современные информационно-аналитические системы содержат в своем составе средства отображения статистики вхождения в базы данных понятий, соответствующих пользовательским запросам. В настоящее время существует несколько открытых информационных сервисов, в рамках которых можно наблюдать временную динамику объемов публикаций по тематикам, определяемым запросами. Так Google books Ngram Viewer (http://ngrams.googlelabs.com/), предоставляет визуализацию динамики количества книг, в которых упоминаются слова. Сервис «Яндекс пульс блогосферы» (http://blogs.yandex.ru/pulse/) также позволял отображать динамику публикаций в блогах, содержащих заданные пользователем ключевые слова, однако был закрыт ввиду малой посещаемости. Сегодня этот сервис доступен лишь по специальному разрешению компании «Яндекс». На сайте Национального корпуса русского языка (НКРЯ) в бета-режиме запущен сервис N-грамм (http://www.ruscorpora.ru/ngram.html), близкий по функциональности сервису Google books Ngram Viewer.

Безусловно, самыми эффективными для решения задач анализа динамического контента являются специализированные системы интеграции сетевого контента. В частности, в рамках исследований авторами использовалась система контент-мониторинга веб-ресурсов InfoStream (www.infostream.ua), реализующая необходимую функциональность и охватывающая около 100 тыс. документов в сутки с 7000 веб-сайтов.

Формирование запроса в среде выбранной системы

Массив тематических документов (тема – события, связанные с Евромайданом в Киеве 2013-2014 гг.) выбирается с помощью системы InfoStream путем ввода запроса на языке данной системы:

(майдан|евромайдан)&(избиен|разгон|штурм|беркут|молотов|титушк|погиб)&lang.RUS,

по которому в период с ноября 2013 года по март 2014 года было опубликовано свыше 200 тысяч ***тематических публикаций*** на веб-сайтах (рис. 1).

Рисунок 1 – Фрагмент ппоискового интерфейса системы InfoStream

Система InfoStream обеспечивает поиск, а также просмотр списка и полных текстов релевантных документов.

Определение динамики тематических публикаций по запросу

Режим динамики событий системы интеграции интернет-ресурсов позволяет получить данные о количестве публикаций по заданному запросу за указанный промежуток времени. Эти данные могут быть загружены в настольную систему обработки данных и отображаются в виде графика (рис. 2). В интерфейсе пользователя обеспечивается переход к просмотру релевантных документов по выбранной дате. После этого данные временной динамики за каждые сутки нормируются, т.е. в среде системы Excel формируется временной ряд, содержащий относительные значения, равные отношению количества тематических сообщений к общему потоку сообщений за сутки.

Это, в частности, позволяет избавиться от недельной периодичности в количестве тематических публикаций. Затем происходит переход к процедуре определения критических точек в данном временном ряде.

Определение критических точек в динамике сообщений

Критические точки как локальные максимумы временного ряда динамики публикаций можно определить, например, визуально по графику, представленному на рис. 2. Вместе с тем, существуют несколько научно-обоснованных методик, базирующихся на методах цифровой обработки сигналов.

Кривая цикла информационных операций

В результате анализа многочисленных диаграмм поведения ТИП, были выявлены наиболее типичные, базовые профили их поведения [Ландэ, 2013], [Додонов и др., 2013]. Предложенные модели полностью соответствуют реальным данным, которые экстрагируются системами контент-мониторинга. Поэтому приведенные зависимости могут быть использованы как шаблоны, например, для выявления информационных операций – как путем анализа ретроспективного фонда сетевых публикаций, так и для оперативного мониторинга появления некоторых их признаков в реальном времени.

Рисунок 2 – Режим «Динамика событий» системы интеграции

В частности, для выявления информационных операций [Горбулин и др., 2009] следует внимательно следить за динамикой публикаций по целевой теме и, если есть возможность, пользоваться доступными аналитическими средствами, средствами цифровой обработки данных и распознавания образов, например, вейвлет-анализом.

На рисунке 3 приведен обобщенная диаграмма, соответствующая всем этапам жизненного цикла информационных операций, обоснованная авторами в [Ландэ, 2008].

Для выявления степени «близости» фрагментов исследуемого временного ряда приведенной диаграмме в различных масштабах предлагается использовать так называемый «вейвлет-анализ» [Астафьева, 1996], который в настоящее время нашел широкое применение как в естественных науках, так и в социологии [Давыдов, 2008].

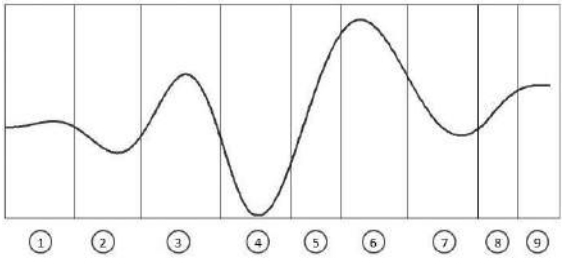

Рисунок 3 – Жизненный цикл информационных операций: 1 – фон; 2 – затишье; 3 – «артподготовка»; 4 – затишье; 5 – атака/триггер роста; 6 – пик завышенных ожиданий; 7 – утрата иллюзий организаторов; 8 – общественное осознание; 9 – продуктивность/фон

Главная идея вейвлет-преобразования заключается в том, что нестационарный временной ряд разделяется на отдельные промежутки (так называемые «окна наблюдения»), и на каждом из них вычисляется величина, показывающая степень

близости закономерностей исследуемых данных с разными сдвигами некоторого вейвлета (специальной функции) в разных масштабах. Вейвлет-преобразование генерирует набор коэффициентов, которые являются функциями двух переменных: времени и частоты, и потому образовывают поверхность в трехмерном пространстве.

Непрерывное вейвлет-преобразование для функции $f(t)$ строится с помощью непрерывных масштабных преобразований и переносов выбранного вейвлета $\psi(t)$ с произвольными значениями масштабного коэффициента a и параметра сдвига b:

$$W(a,b) = (f(t), \psi(t)) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \psi^* \left(\frac{t-b}{a} \right) dt .$$

Полученные вейвлет-коэффициенты можно представить в графическом виде, если по одной оси отложить сдвиг вейвлета (ось времени), а по другой – масштабы (ось масштабов), и раскрасить точки полученной схемы в зависимости от величины соответствующих коэффициентов (чем больше коэффициент, тем ярче цвета).

Эти коэффициенты, которые показывают, насколько поведение процесса в данной точке аналогично вейвлету в данном масштабе. Чем ближе от анализируемой зависимости в окрестности данной точки к виду вейвлета, тем большую абсолютную величину имеет соответствующий коэффициент. Применение этих операций, с учетом свойства локальности вейвлета в частотно-временной области, позволяет анализировать данные на разных масштабах и точно определять положение их характерных особенностей во времени.

На скейлограмме видные все характерные особенности исходного ряда: масштаб и интенсивность периодических изменений, направление и значение трендов, наличие, расположение и продолжительность локальных особенностей.

В работе [Ландэ, 2013] показано, что вейвлеты «мексиканская шляпа» и Морле (рис. 4) наиболее точно отражает динамику информационных операций, результаты применения этого вейвлета приведены на рис. 5, благодаря чему были выбраны три даты (2013.11.30, 2014.01.22, 2014.02.19), соответствующие критическим точкам исследуемого процесса.

Следует отметить, что инструменты построения вейвлет-спектограмм доступны как в ряде пакетов математических программ, например, в Matlab, так и через Интернет в режиме онлайн (http://ion.researchsystems.com/cgi-bin/ion-p?page=wavelet.ion).

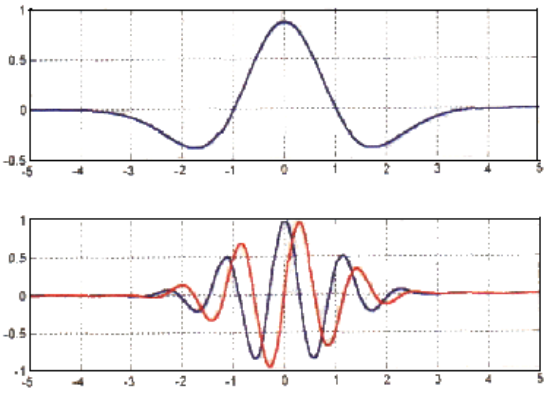

Рисунок 4 - Вейвлеты mexh, Морле

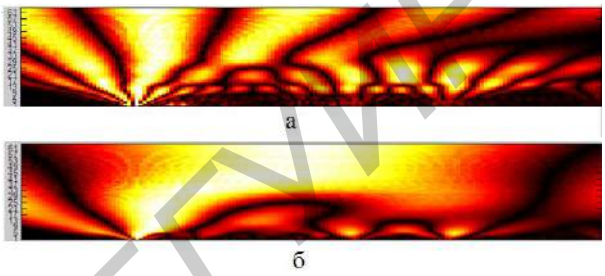

Рисунок 5 – Вейвлет-спектограммы исследуемого временного ряда (а – «мексиканская шляпа», б – вейвлет Морле)

Определение основных событий в критических точках

После определения критических точек с помощью системы контент-мониторинга выполняется построение основных сюжетных цепочек из сообщений, соответствующих запросу за выбранные даты. Таким образом определяются основные события за указанные даты (рисунок 6).

Для последующего анализа отбирается три массива сообщений, соответствующие трем выбранным датам, объекты из которых (в простейшем случае – персоны и веб-источники) могут рассматриваться как объекты мониторинга.

Выявление объектов мониторинга

С помощью методов экстрагирования фактических данных, применяющихся в системах интеграции интернет-ресурсов, в интерфейсе пользователя формируются так называемые «информационные портреты», охватывающие списки персон, топонимов, языков, компаний и т.п., содержащиеся в документах, релевантных некоторому заданному запросу.

В нашем случае из «информационного портрета», соответствующего тематическому запросу выбираются наиболее упоминаемые персоны и/или веб-ресурсы за выбранные даты (рис. 7 и 8). Эти списки могут агрегироваться, в результате чего возможно определение взаимосвязей событий и объектов (рис. 9).

В качестве системы визуализации авторами выбрана система анализа и отображения сетей Gephi (www.gephi.org).

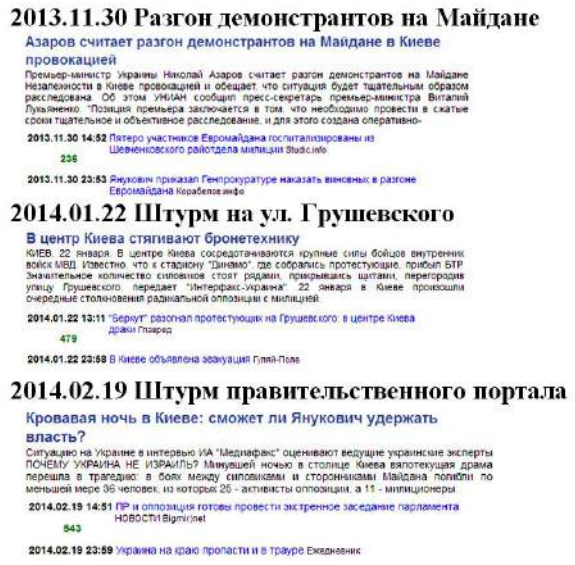

Рисунок 6 – Основные сюжетные цепочки за выбранные даты

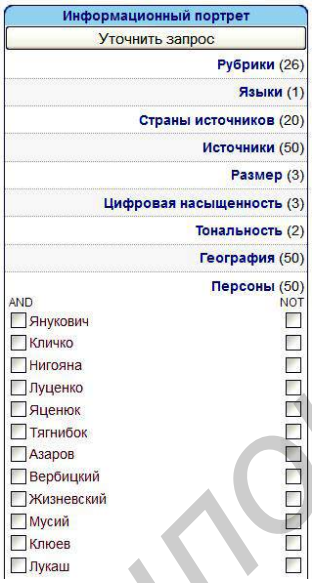

Рисунок 7 - Списки «доминирующих» персон

Эти же данные позволяют выявлять взаимосвязи между объектами, например, между указанными аналитиками веб-ресурсами и персонами (рис. 10).

На рис. 9 можно видеть, что каждому массиву (узлы, идентифицированные датами) соответствуют объекты. При этом в центральной части сети располагаются объекты, общие для нескольких событий (О-зона), а «гребешки» на периферии соответствуют специальным объектам, отражающим специфику конкретных событий (С-зоны).

Также можно предложить критерий релевантности события, связанного с конкретной датой, общей тематике: чем большая часть объектов из него попадает в О-зону, тем он более релевантен тематике. Формально значение этого критерия $k_{i,N}$ для сюжета i тематики s может быть записано следующим образом:

$$k_{i,N} = \frac{T_{i,N} \bigcap T_{S,N}}{N},$$

где N – количество объектов, $T_{i,N}$ – множество значимых объектов события i, – $T_{s,N}$ – множество значимых объектов для всей тематики.

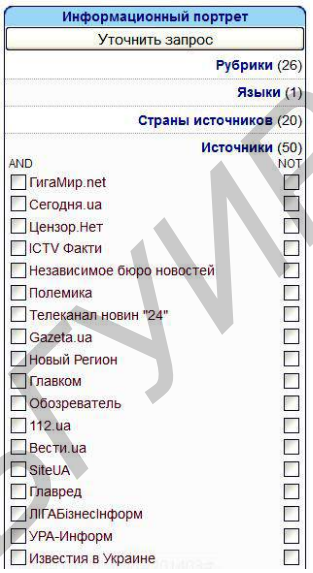

Рисунок 8 - Списки «доминирующих» веб-ресурсов

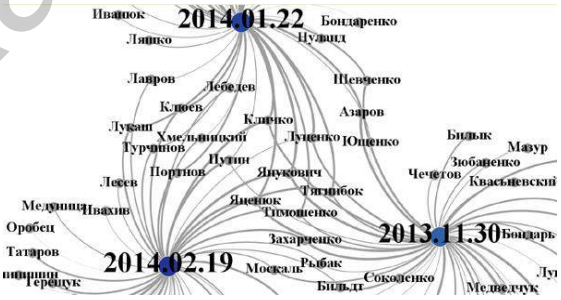

Рисунок 9 - Пример визуализации связей событий и объектов

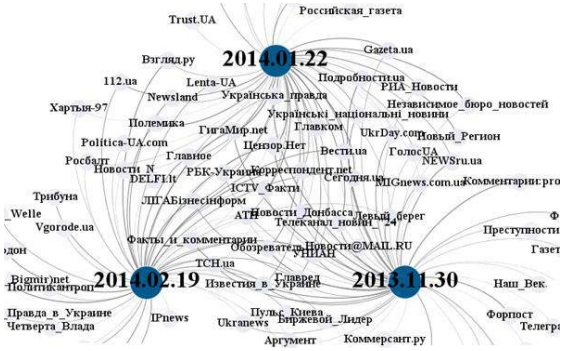

Рисунок 10 - Визуализация взаимосвязей веб-ресурсов и персон

Подход к прогнозу: R/S-анализ

Для решения задач прогнозирования перспективным представляется применение теории фракталов при анализе информационного пространства. Фрактальный анализ самоподобия информационных массивов может рассматриваться

как технология, предназначенная для осуществления аналитических исследований с элементами прогнозирования, пригодная к экстраполяции полученных зависимостей.

Важнейшей характеристикой рядов, которые имеют хаотичное поведение, является, как известно, показатель Херста (H), определяемый в результате так называемого R/S-анализа [Федер, 1991]. Этот показатель базируется на анализе нормированного разброса – отношения разброса значений исследуемого ряда R к стандартному отклонению S.

Достаточно часто, когда соотношение R/S

$$R/S = (N/2)^H,$$

имеет постоянный тренд, можно говорить о соотношении: где H – показатель Херста, который для достаточно широкого класса рядов связан с хаусдорфовой (фрактальнй) размерностью D постой формулой: $D+H=2$.

На рис. 11 представлено соотношения R/S для рассматриваемого в этой работе временного ряда. Как можно видеть, кривая нормированного размаха удовлетворительно аппроксимируется прямой в двойном логарифмическом масштабе. Численные значения H характеризуют разные типы коррелированной динамики (персистентности). При $H=0,5$ наблюдается некоррелированное поведение значений ряда, а значение $0,5 < H < 1$ соответствует степени автокорреляции ряда.

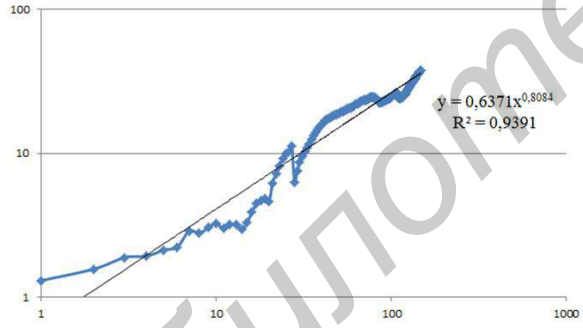

Рисунок 11 - Кривая *R/S* в двойной логарифмической шкале

Как можно видеть, значение показателя Херста для исследуемого информационного потока соответствует величине 0,81, что подтверждает предположение о самоподобии и итеративности рассматриваемых процессов в информационном пространстве. Это означает, что общая информационная напряженность остается на большом уровне, как только исчезает «шлейф» одного сюжета по выбранной тематике информационных, ему на смену возникает новый сюжет, т. е. его поведение в дальнейшем будет близко к предшествующему поведению.

Заключение

В докладе представлена методика аналитического исследования, которая базируется на сценарном подходе и использовании современных инструментальных средствах анализа и визуализации информационных потоков и временных рядов. Как показывает опыт, применение сценарного подхода при исследовании контента и структуры информационных ресурсов сети Интернет позволяет существенно повысить эффективность аналитической деятельности.

Предложенная методика, по мнению авторов, позволяет решить сформулированную задачу, ее можно использовать в качестве основы для проведения аналитической и прогнозной деятельности на основе исследования контента современных компьютерных сетей. В дальнейшем на основе полученных результатов планируется разработка моделирующего комплекса.

Библиографический список

[Астафьева, 1996] Астафьева Н.М. Вейвлет-анализ: основы теории и примеры применения // Успехи физических наук, 1996. – 166. – № 11. – Р. 1145-1170.

[Горбулин и др., 2009] Горбулін В.П., Додонов О.Г., Ланде Д.В. Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія. – К.: Інтертехнологія, 2009. – 164 с.

[Давыдов, 2008] Давыдов А.А. Системная социология. – М.: Издательство ЛКИ, 2008. – 192 с.

[Додонов и др., 2009] Додонов О.Г., Ланде Д.В., Путятін В.Г. Інформаційні потоки в глобальних комп'ютерних мережах – К: Наукова думка, 2009, – 295 с.

[Додонов и др., 2013] Додонов А.Г., Ландэ Д.В., Прищепа В.В., Путятин В.Г. Конкурентная разведка в компьютерных сетях.– К.: ИПРИ НАН Украины, 2013. – 248 с.

[Ландэ, 2008] Ланде Д.В. Новітні підходи й технології інформаційно-аналітичної підтримки прийняття рішень // Національна безпека: український вимір: щокв. наук. зб. / Рада нац. безпеки і оборони України, Ін-т пробл. нац. Безпеки, 2008. – Вип. 1-2 (20-21). – С. 87-105.

[Ландэ, 2013] Ланде Д.В. Тренди відображення інформаційних операцій в інформаційному просторі // Інформація і право, 2013. – N 1 (7). – С. 82-88.

[Федер, 1991] Федер Е. Фракталы. – М., Мир, 1991. – 261 с.

A SCENARIO APPROACH IN THE RESEARCH OF DYNAMICS OF INFORMATION STREAMS IN THE INTERNET

Dodonov A.G., Lande D.V.,Boychenko A.V.

Institute for Information Recording NAS of Ukraine, Kiev, Ukraine

dodonov@ipri.kiev.ua, dwlande@gmail.com

boychenko.a@gmail.com

A scenario approach to the study of the dynamics of information flows in the analysis and prediction of social phenomena are considered. Identified and described in detail the stages of information-analytical study based on the analysis of content on the Internet.

Keywords: scenario approach; dynamics of information streams; content-monitoring; analytical research.