УДК 681.3

Д.В. Ландэ, В.Н. Фурашев

Объектно-статистический анализ информационных потоков

Информационный центр «ЭЛВИСТИ»,
Научно-исследовательский центр по вопросам правовой информатики
Академии правовых наук Украины

Динамика и постоянно увеличивающиеся объемы разноплановых публикаций в Интернет обусловливают проблему получения данных для информационно-аналитических исследований, как оперативных, так и ретроспективных по различным тематическим направлениям [1]. Обычными, традиционными, методами поиска и экстрагирования информации, необходимой для последующей обработки, уже не обойтись. На помощь могут прийти лишь системы контент-мониторинга, охватывающие тысячи информационных ресурсов, и позволяющие выявлять тенденции, сюжеты, объекты и их связи [2]. Вместе с тем, анализ процессов, которые имеют довольно значительные временные рамки, все еще ждет своего инструментария. Если вопросы визуализации результатов поиска информационного отображения подобных процессов освещаются в большом количестве работ [3-6], то анализу и визуализации объектного распределения отобранных информационных массивов больших объемов до сих пор не уделялось существенного внимания.

Предметной областью исследования авторов в данной работе является анализ и визуализация объектного распределения отобранных информационных массивов. на примере анализа динамики публикаций в Интернет-пространстве о деятельности системы избирательных комиссий в Украине по выборам Президента Украины и народных депутатов Украины за 2004-2006 годы. Эта динамика отражает реальный интерес общественности, через электронные средства информации, к избирательным процедурам, а также процессы, происходящие в ходе избирательных кампаний.

Система контент-мониторинга InfoStream на основании анализа около 3000 источников информации в сети Интернет позволила построить зависимость суточных объемов тематических публикаций за 3 года (1096 суток, общее количество – свыше 320 тысяч). Пики на графике (рис. 1), позволяют оценить интенсивность освещения в прессе как президентской избирательной кампании 2004 г., так выборов в Верховный Совет Украины в 2006 г.

Вместе с тем, для более детального анализа процессов, общепринятыми методиками является анализ Фурье и вэйвлет-анализ [7, 8]. Технология использования вэйвлетов (маленьких волн) позволяет выявлять одиночные и нерегулярные "всплески", резкие изменения значений количественных показателей в различные периоды времени, в частности, объемов тематических публикаций в Интернет. При этом могут выявляться моменты возникновения циклов, а также когда за периодами регулярной динамики следуют хаотические колебания. Метод вэйвлет–анализа используется также для декомпозиции, выделения сигнала из "шума", изучения динамики различных процессов, в том числе экономических и социальных. На рис. 2 приведена спектограмма - результат вэйвлет-анализа временного ряда, соответствующего изучаемому процессу.

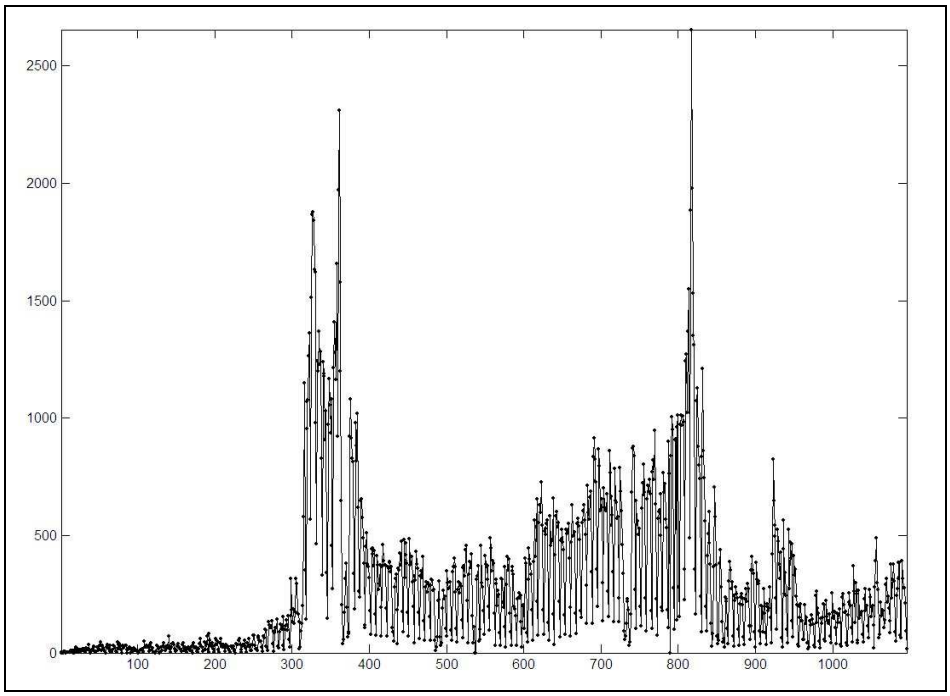

Рис.1. Количество тематических публикаций (ось Y) по дням (ось X)

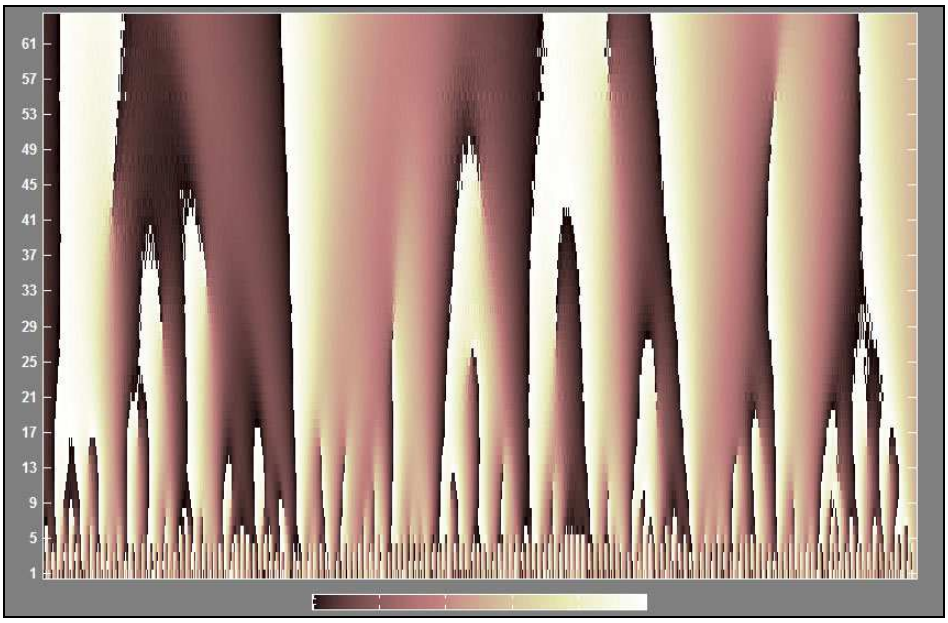

Рис. 2. Вэйвлет-спектограмма динамики тематического информационного потока (одномерное непрерывное вэйвлет преобразование, вэйвлет Гаусса), ось X – дни, ось Y – частоты

Прекрасно отражая спектральные характеристики сигналов, вэйвлет-анализ, однако, по своей природе, не может быть использован, когда информационный поток следует рассматривать с объектной точки зрения. В случае, рассматриваемом авторами, такими объектами выступали отдельные лица, определяемые в публикациях своими фамилиями, инициалами, должностями и т.п. В частности, с помощью средств экстрагирования информации системы InfoStream

из рассматриваемого потока было выявлено упоминание о более чем 40 тыс. лицах, в той или иной мере имеющих отношение к избирательному процессу. В экстрагированном виде каждая персона представлялась одним дескриптором. Для обеспечения учета и анализа распределения информационных потоков в разрезе интересующих персон был предложен оригинальный метод, так называемых вордлет-диаграмм. Эти диаграммы представляют собой форму визуального отображения информационного потока в разрезе объектов и дат, представляющая собой прямоугольную таблицу, ячейки которой заполнены значениями количества сообщений информационного потока за определенную дату, соответствующих определенному объекту. Столбцам этой таблицы соответствуют даты, а строкам – объекты, являющиеся своеобразными содержательными фильтрами исследуемого информационного потока. Объектам в рассматриваемом случае соответствуют определенные лица. Естественно, для визуального отображения из множества персон выбирается лишь несколько десятков интересующих исследователя.

Визуально вордлет-диаграмма представляет собой таблицу, ячейки которой закрашены оттенками серого цвета, в зависимости от значений объемов публикации по выбранному объекту в соответствующий день (большее значение соответствует более темному оттенку). Следует заметить, что многие строки вордлет-диаграммы обладают фрактальными свойствами, которые присущи им как количественным индикаторам тематических информационных потоков. В частности, для аналогичных временных рядов было экспериментально подтверждено наличие статистической корреляции на достаточно длительных интервалах [9-10].

Вордлет-диаграммы для относительно небольшого количества строк (несколько десятков) позволяют визуально выявлять группы наиболее связанных по датам и интенсивностям публикаций объектов. Для большего количества объектов в процессе построения Вордлет-диаграммы предлагается ее кластеризация путем перестановки строк (перегруппировки объектов) в соответствии с алгоритмом k-means [11]. При этом подразумевается, что например в случае разбиения на 2 кластера, основы кластеров *i* и *j* (центроиды), которые затем рекурсивно уточняются, выбираются для наибольших значений функции:

$$F_{ij} = \sum x_{ik} * \sum x_{jk} * R_{ij},$$

где x_{ik} – значения элементов таблицы, суммы берутся по всем датам *k*; R_{ij} – «расстояние» между строками *i* и *j*, определяемое формулой:

$$R_{ij} = \sum |x_{ik} - x_{jk}|.$$

Следует отметить, что кроме названной выше тематической задачи, были получены вордлет-диаграммы, соответствующие большим информационным потокам различной тематической направленности. В качестве параметров запросов для отбора объектов выбирались такие параметры, как ключевые слова, фамилии, географические названия, названия организаций.

На рис. 3 приведена вордлет-диаграмма первого уровня (превью), позволяющая визуально выявлять аномальные корреляции. На этой диаграмме, охватывающей информацию по 49 персонам, отчетливо видны циклы праздничных дней, а также корреляции отдельных объектов. С помощью

приведенной на рис. 4 уточняющей вордлет-диаграммы можно точно указать выявленные корреляции, например, персон с номерами 10, 11 и 26, 27 за последние 20 дней.

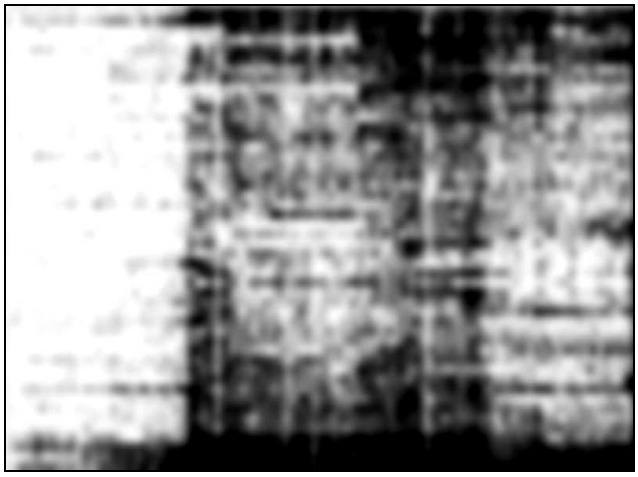

Рис. 3. Вордлет-диаграмма-превью (ось X – дни, ось Y – персоны)

В результате проведенных экспериментов, есть основания предположить, что использование таких средств визуализации, как вордлет-диаграммы, позволяет «разлагать» исходные временные ряды в соответствии с объектами, обнаруживать медиа-активность по выбранным объектам, выявлять взаимосвязи объектов в разрезе дат, определять детали медиа-активности каждого объекта или группы объектов. Вордлет-диаграмы позволяют более адекватно анализировать динамику публикаций в Интернет в разрезе интересующих объектов, предоставляя в наглядном виде важную информацию о динамике реальных процессов. Использование вордлет-диаграмм представляется важным дополнением к уже признанным методам исследований, таким как анализ Фурье, корреляционный и фрактальный анализ, а также вэйвлет-анализ.

Необходимо отметить, что представленный в подход к решению вопроса анализа и визуализации объектного распределения отобранных информационных массивов, несмотря на то, что он продемонстрирован на примере анализа динамики публикаций в Интернет-пространстве о деятельности системы избирательных комиссий в Украине по выборам Президента Украины и народных депутатов Украины, носит общий характер.

Данный подход применим для решения вопросов анализа и визуализации объектного распределения любых отобранных информационных массивов для процессов, которые интересуют исследователя и имеют довольно значительные временные рамки.

Кроме того, необходимо отметить, что представленный подход анализа информационных потоков носит объектно-статистический характер, который, в свою очередь, представляется как существенная составляющая методологической базы прогнозно-эмпирического анализа.

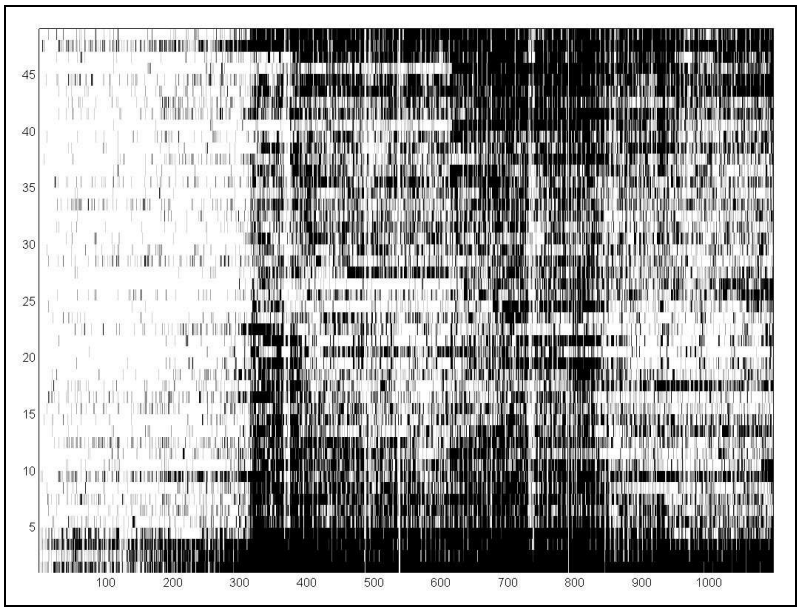

Рис. 4. Уточняющая вордлет-диаграмма (ось X – дни, ось Y – персоны)

Список литературы

1. Ландэ Д.В. Основы интеграции информационных потоков - К.: Инжиниринг, 2006. - 240 с.
2. Григорьев А.Н., Ландэ Д.В., Бороденков С.А., Мазуркевич Р.В., Пацьора В.Н. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие. – К.: ООО «Старт-98», 2007. – 40 с.
3. M.M. Knepper, R. Killam, K.L. Fox O. Frieder. Information Retrieval and Visualization using SENTINEL / TREC 1998: 336-340
4. Григорьев А.Н., Ландэ Д.В. Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream // Труды Международного семинара «Диалог'2005». – М.: Наука, 2005. – С. 109-111.
5. Григорьев А.Н. Многоуровневый классификатор-навигатор по откликам информационно-поисковой системы // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2006 – Москва: Наука, 2006. - С. 329-331.
6. Z. Junliang, Javed M., Himansu T. Information Retrieval by Semantic Analysis and Visualization of the Concept Space of D-Lib® Magazine // D-Lib Magazine October 2002 Volume 8 Number 10.
7. Давыдов. А. А. Системная социология. –М: КомКнига, 2006. - 192 с.
8. Чуи К. Введение в вэйвлеты.М.: Мир, 2001.
9. Иванов С.А. Стохастические фракталы в Информатике // Научно-техническая информация. Сер. 2, 2002. - № 8. - С. 7-18.
10. Ландэ Д.В. Фрактальные свойства тематических информационных потоков из Интернет // Регистрация, хранение и обраб. данных. - К., 2006. - Т. 8, № 2. - С. 93 - 99.
11. J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297.