Объектная визуализация тематических информационных массивов♣

© Ландэ Д.В. Григорьев А.Н. Брайчевский С.М. Дармохвал А.Т. Снарский А.А.

Информационный центр «ЭЛВИСТИ» НТТУ «КПИ»
dwl@visti.net gri@visti.net smb@visti.net hval@visti.net asnarskii@gmail.com

Аннотация

Описан подход к визуализации тематических информационных массивов электронных публикаций. Предлагается использование так называемых Wordlet-диаграмм, которые формируются путем учета распределения объемов публикаций, соответствующих выбранным информационным объектам.

Использование Wordlet-диаграмм представляется важным визуальным дополнением к системам интеграции информационных ресурсов.

В настоящее время объемы и динамика обновления информации в Интернет достигли такого уровня, что позволяют говорить об информационных потоках [1, 2]. В то же время, навигация в этих потоках является проблематичной. Предполагается, что часть проблем, связанных с ориентацией пользователей в информационных ресурсах Интернет решится за счет средств визуализации, создания своеобразных «пультов наблюдения» за информационными потоками.

Вопросы визуализации результатов поиска посвящено большое количество научных работ и практических разработок [3-6]. Вместе с тем, визуализации тенденций и объектного распределения информационных потоков больших объемов, имеющих значительные временные рамки и доступных пользователям Интернет уже сегодня, не уделяется существенного внимания.

Предметной областью исследования авторов в данной работе является анализ и визуализация объектного распределения отобранных массивов электронных документов. В качестве примера авторами был выбран анализ динамики публикаций в сети Интернет, посвященных теме «коррупция».

Система контент-мониторинга InfoStream на основании анализа около 3000 источников информации в сети Интернет позволила построить зависимость суточных объемов тематических публикаций за 456 дней (с 1 января 2007 г. по 31 марта 2007 г.), на котором видно распределение количества сообщений по дням (рис. 1). Общий объем рассматриваемого тематического информационного массива составил около 86 тыс. сообщений.

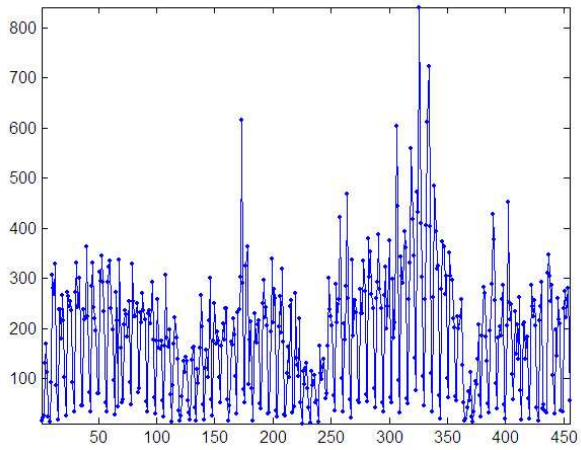

Рис.1. Количество тематических публикаций (ось Y) по дням (ось X)

Для более детального анализа процессов, общепринятыми методиками является анализ Фурье и вейвлет-анализ [7, 8]. Технология использования вейвлетов (маленьких волн) позволяет выявлять одиночные и нерегулярные "всплески", резкие изменения значений количественных показателей в различные периоды времени, в частности, объемов тематических публикаций в Интернет. Метод вейвлет–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 вейвлет-анализа временного ряда, соответствующего изучаемому процессу.

Вместе с тем, методы Фурье и вейвлет-анализа обеспечивают лишь формальную, частотную фильтрацию временных рядов. Авторы же своей целью ставили создание интерфейса содержательного анализа этих рядов на основании дополнительной информации о распределении объектов, каждый из которых может ассоциироваться с соответствующим информационным запросом.

В случае, рассматриваемом авторами, такими объектами выступали отдельные лица, определяемые в публикациях своими фамилиями, инициалами, должностями и т.п. В частности, с

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.

помощью средств экстрагирования информации системы InfoStream из рассматриваемого потока было выявлено упоминание о более чем 26 тыс. лицах, присутствующих в публикациях по выбранной теме. Как промежуточный результат исследования был выявлен тот факт, что ранговое распределение фамилий с большой точностью удовлетворяют закономерности Ципфа (рис. 3).

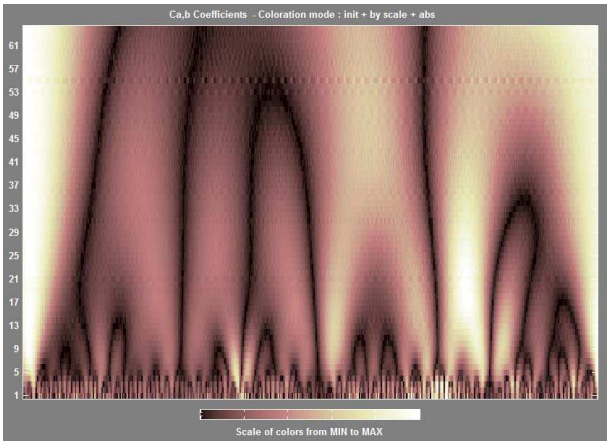

Рис. 2. Вейвлет-спектограмма динамики тематического информационного потока (одномерное непрерывное вейвлет-преобразование, вейвлет Гаусса), ось X – дни, ось Y – частоты

Авторами предложена форма визуального отображения информационного потока в разрезе объектов и дат, представляющая собой прямоугольную таблицу, будем называть ее Wordlet-диаграммой, ячейки которой заполнены количеством публикаций, соответствующих выбранному объекту в разрезе дат. Т.е. столбцам этой таблицы соответствуют даты, а строкам – объекты, выступающие своеобразными содержательными фильтрами информационного потока. Точнее фильтрами являются запросы на информационно-поисковом языке, соответствующие объектам.

Визуально Wordlet-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 Wordlet-диаграммы обладают фрактальными свойствами [9, 10], в частности, для аналогичных временных рядов было экспериментально подтверждено наличие статистической корреляции на длительных интервалах.

Wordlet-диаграммы для относительно небольшого количества строк (несколько десятков) позволяют без дополнительной обработки выявлять группы наиболее связанных по датам и интенсивностям публикаций объектов визуально. Для большего количества объектов в процессе построения Wordlet-диаграммы предлагается ее кластеризация путем перестановки строк (перегруппировки объектов) в соответствии с алгоритмом k-means [11]. При этом для отнесения объектов к тому или иному кластеру определяется евклидова мера близости между основой кластера i (центроидом – объектом μ_i) и объектом x:

$$sim(\mu_i,x) = \sum (\mu_{ik} - x_k)^2,$$

где сумма берется по всем датам $k = 1, \ldots, N$.

В ходе исследований были получены Wordlet-диаграммы, соответствующие большим информационным потокам (свыше 50 тыс. сообщений) различной тематической направленности. В качестве параметров запросов для отбора объектов выбирались такие параметры, как ключевые слова, фамилии, географические названия, бренды. В частности, рассматривался ранжированный по частоте встречаемости список фамилий лиц, упоминаемых в текстах сообщений.

На приведенной на рис. 4. Wordlet-диаграмме, охватывающей информацию по 40 персонам, например, отчетливо видны циклы праздничных дней, а также корреляции отдельных объектов.

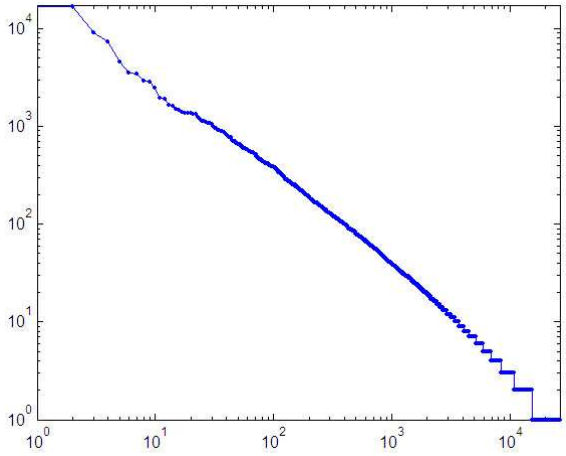

Рис. 3. Ранжированный список фамилий в текстах сообщений тематического информационного массива в двойном логарифмическом масштабе (ось X – номер фамилии, ось Y – частота встречаемости)

В результате проведенных экспериментов, есть основания предположить, что использование таких средств визуализации, как Wordlet-диаграммы, позволяет «разлагать» исходные временные ряды в соответствии с составом и особенностями объектов, обнаруживать активность публикаций о выбранных объектах, выявлять взаимосвязи объектов в разрезе дат, определять детали динамики встречаемости в информационном потоке каждого объекта или группы объектов. Использование Wordlet-диаграмм представляется важным дополнением к уже признанным методам исследований, таким как анализ Фурье, корреляционный, фрактальный и вейвлет-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 по теме «коррупция», носит общий характер.

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

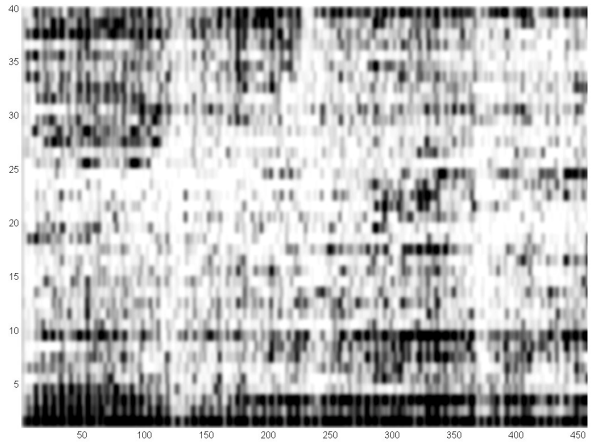

Рис.4. Пример превью Wordlet-диаграммы (ось X – дни, ось Y – персоны).

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

Литература

[1] Gianna M. Del Corso, Antonio Gullí, Francesco Romani. Ranking a stream of news. Proceedings of the 14th international conference on World Wide Web. Chiba, Japan. – 2005. - P. 97 - 106.

[2] С.М. Брайчевский, Д.В. Ландэ. Современные информационные потоки: актуальная проблематика // Научно-техническая информация. Сер. 1, 2005. - № 11. - С. 21-33.

[3] M.M. Knepper, R. Killam, K.L. Fox O. Frieder. Information Retrieval and Visualization using SENTINEL / TREC 1998: 336-340.

[4] 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.

[5] Ландэ Д.В. Присмотритесь внимательнее или "Изюминки" поисковой визуализации // hiTech Pro. - К., 2006. - декабрь, - С. 94-95. (http://dwl.kiev.ua/art/hitech/)

[6] 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.

[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. (http://dwl.kiev.ua/art/frak-ip/)

[11] J. B. MacQueen. 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. – 1997.

Object visualization of thematic informational arrays

Lande, D.V., Grigorjev A.N., Brajchevskiy, S.M., Darmokhval, A.T., Snarskii, A.A.

Approach to visualization of thematic informational arrays of electronic publications is described. Use of so called Wordlet-diagrams which are formed by consideration of distribution of the volumes of publications, corresponding to the chosen informational objects, is proposed.

The use of Wordlet-diagrams is presented by the important visual complement of the systems of integration of the informational resources.

♣ Исследование является компонентой НИР, поддержанной компанией «Яндекс» в рамках конкурса «Интернет-математика»