УДК 004.4′6
В. В. ВИШНЕВСКИЙ, доктор техн. наук, профессор,
Государственный университет телекоммуникаций, Киев;
С. В. ПРИЩЕПА; Д. В. ЛАНДЭ,
Институт проблем регистрации информации НАН Украины, Киев

РАНЖИРОВКА ВЕБ-САЙТОВ СОБЫТИЙНОГО ХАРАКТЕРА

Описана технология построения сети авторства информации на веб-сайтах событийного характера, основанная на анализе контекстных ссылок, с учетом взаимного влияния источников информации, выраженного в виде ссылок в тексте или перепечаток существенных фрагментов текста. Указанные признаки позволяют определять первоисточники информации. Представлены методы и средства анализа сетей источников информации событийного характера.

Ключевые слова: информационные операции; информационные потоки; выявление событий; кластерный анализ; модель предметной области.

Введение

Современное веб-пространство представляет собой значимый фрагмент мирового информационного пространства, содержащий репрезентативную информацию о событиях и процессах, происходящих в обществе. При формировании веб-сайтов необходимо учитывать различные механизмы распространения информации в сети Интернет, а также способы взаимного влияния и авторство источников информации, посредством которых и осуществляется воздействие на пользователей.

Рост объемов информации в интернете, информационный шум, огромное количество фейка — все это приводит к увеличению временных и других затрат на поиск и выявление достоверной событийной информации, ценной как для различных государственных органов, так и для представителей конкурентной бизнес-разведки. Разработка и внедрение системы автоматического выявления событий, их субъектов и объектов — актуальная проблема, требующая углубленного исследования.

Информационное пространство интернета (веб-ресурсы, социальные сети), являясь мощнейшей площадкой для отражения действительности, в частности событий и общественных процессов, становится все более важным фактором воздействия на общество путем проведения информационных операций [1].

Цель статьи — представить оригинальную технологию ранжировки веб-сайтов событийного характера, предусматривающую определение «авторства» сетевых источников информации на основе анализа контекстных ссылок. В отличие от существующих подходов к анализу гиперсвязей в сетевых документах, применяемых для анализа популярности веб-страниц, в данной статье предлагается учитывать авторство веб-сайтов, выраженное ссылками в тексте и цитированиями некоторых фрагментов текстов. Предлагается также методология оперативного выявления информации о событиях благодаря анализу взаимных ссылок веб-ресурсов.

Актуальность разработки методов экстрагирования событий из информационных потоков растет во всем мире. Подтверждением этому является, например, количество научных публикаций и патентов по указанной теме за прошедший год. В рамках данной статьи под ***событием*** будем понимать значительное происшествие, явление или результат иной деятельности, представляющий собой факт общественной или личной жизни [2]. В данной статье описан ***метод сетевой ранжировки источников сообщений*** на основе анализа слабо структурированных текстовых данных из веб-пространства.

Этапы исследования авторства источников событийной информации

Для определения авторства источников информации из сети Интернет (веб-ресурсов) рассматривается последовательность этапов, базирующихся на использовании доступных инструментальных средств, а также ряда специальных приемов. Указанную последовательность можно трактовать как процедуру, включающую в себя действия по построению и анализу ***сети авторства источников информации***.

При проведении указанных исследований на базе системы контент-мониторинга, в частности системы InfoStream (http://infostream.ua), к задачам определения источников событийной информации можно отнести:

- нахождение релевантных событийных публикаций по заданной тематике;
- выявление контекстных ссылок и цитирований в документах, представленных на различных веб-сайтах;
- построение сети авторства источников информации на основе анализа и визуализации взаимосвязей информационных источников, в том числе ранжировки узлов построенной сети по степени авторства;

• выявление возможных информационных операций путем анализа рангов первоисточников информации.

Процедура исследования авторства источников информации включает в себя пять этапов.

1. Получение репрезентативного массива публикаций по выбранным событиям.

2. Выявление контекстных ссылок и перепечаток в сформированном информационном потоке.

3. Построение сети авторства источников информации на основе анализа контекстных ссылок и перепечаток.

4. Исследование сети авторства источников информации с ранжировкой узлов данной сети.

5. Выявление возможных информационных операций по результатам анализа рангов первоисточников информации.

Рассмотрим указанные шаги на конкретных примерах.

Получение репрезентативного массива публикаций

Для получения репрезентативного массива публикаций по выбранной тематике необходимо выбрать систему контент-мониторинга, отражающую поток информационных сообщений, связанных с определенным событием. Событие может формально выражаться запросом на языке информационно-поисковой системы.

В качестве системы контент-мониторинга авторами была выбрана система InfoStream, охватывающая ныне 10 тыс. источников информации на русском, украинском и английском языках. В базы данных системы ежедневно поступает свыше 100 тыс. документов. Система InfoStream обеспечивает поиск, а также просмотр списка и полных текстов релевантных документов.

В приведенном на рис. 1 примере представлен фрагмент интерфейса системы, через который отрабатывался запрос, относящийся к проведенному в 2016 году референдуму по выходу Великобритании из Европейского Союза, известному как Brexit (от *Britain* — Британия и *Exit* — выход). Период исследования: июнь–июль 2016-го. В результате был сформирован тематический информационный массив, охватывающий 43 697 документов.

Выявление контекстных ссылок

Основой построения сети авторства событийных источников информации являются контекстные ссылки и перепечатки в тематическом информационном потоке. Контекстные ссылки обнаруживаются идентификацией шаблонов в документах выбранного информационного массива и признаков точных перепечаток, определяемых методами выявления плагиата [3; 4]. При этом сами шаблоны периодически определяются/дополняются экспертами в автоматизированном режиме путем анализа контекста потока документов системы контент-мониторинга методами Text Mining.

Построение сети авторства источников информации

Найденные в текстах контекстные ссылки и перепечатки позволяют сформировать матрицу цитирования. Данной матрице соответствует сеть авторства событийных источников, связанных с событием Brexit. Пример визуализации указанной сети для рассмотренного ранее тематического информационного массива приведен на рис. 2.

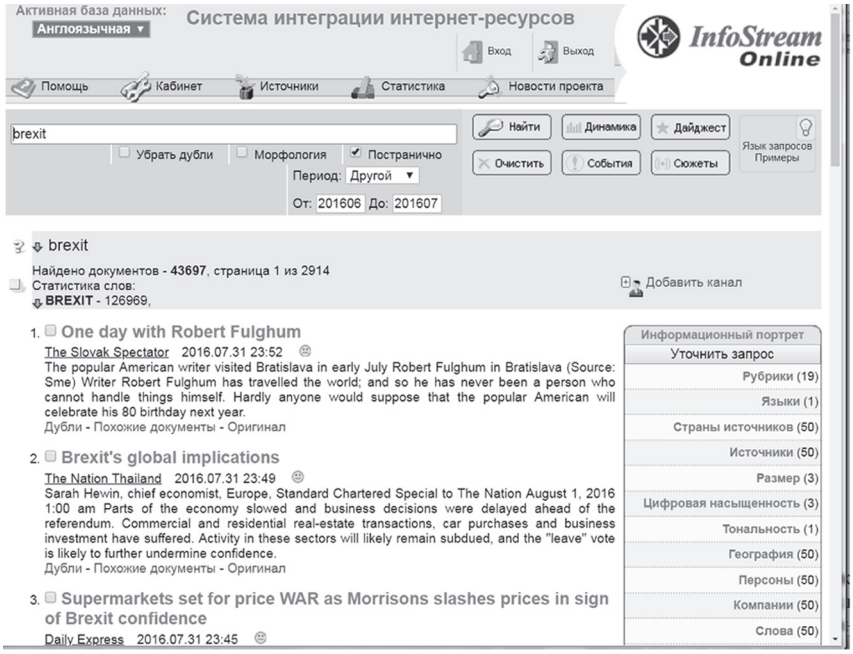

Рис. 1. Фрагмент интерфейса системы контент-мониторинга

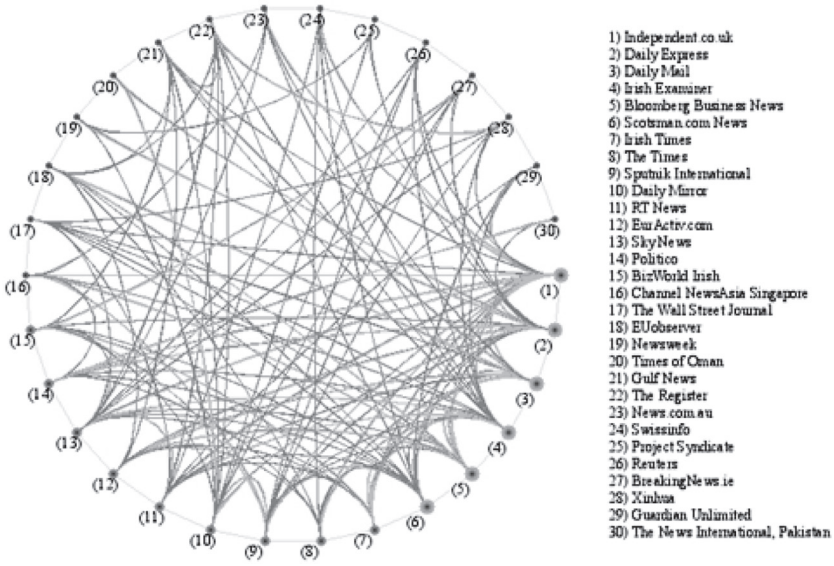

Рис. 2. Фрагмент сети связей источников событийной информации

Исследование сети авторства источников событийной информации

Построенную сеть авторства источников информации можно исследовать при помощи общепринятых инструментальных средств. Например, с использованием системы Gephi были получены такие параметры построенной сети, как средний коэффициент кластеризации, равный 0,01; средняя длина пути, равная 1,27, и др.

Для содержательного анализа большое значение имеет вес узлов сети. Список самых весомых узлов приведен в таблице.

Перспективный подход к ранжировке источников по уровню влияния реализует алгоритм HITS, предложенный Дж. Кляйнбергом [5].

Алгоритм HITS обеспечивает выбор из сети лучших «авторов» (узлов, на которые ведут ссылки) и «посредников» (узлов, от которых идут ссылки включения). Узел является хорошим посредником, если от него идут связи на другие важные узлы, и, в свою очередь, хорошим автором, если к нему ведут ссылки от важных узлов.

Таблица 1

Наиболее весомые узлы по количеству цитирования

№ п/п	Веб-ресурс	Исходящая мощность
1	Independet.co.uk	53
2	Daily Express	38
3	Daily Mail	35
4	Irish Examiner	31
5	Bloomberg Business News	28
6	Scotman.com	21
7	Irish Times	20
8	The Times	17
9	Sputnik International	11
10	Daily Mirror	9

В соответствии с алгоритмом HITS для каждого узла сети v_j рекурсивно вычисляется его значимость как автора $a(v_j)$ и посредника $h(v_j)$ по формулам:

$$a(v_j) = \sum_{j \to i} h(v_i); \;\; h(v_j) = \sum_{j \to i} a(v_i).$$

В данных формулах суммирование осуществляется по всем узлам, которые ссылаются (или на которые ссылаются — во второй формуле) на данный узел.

Перефразируя обозначения, приведенные в [5], а именно заменяя «авторство» на «подверженность влиянию», а «посредничество» на «влиятельность», можно с небольшими вычислительными затратами определять соответствующие характеристики узлов сети влияния.

В частности, для определения информационных влияний важное значение имеет определение «скрытых» связей в ситуации, когда прямых связей между узлами нет, но прослеживаются связи через другие (вторые, третьи и т. д.) узлы. Методика определения скрытых связей и скрытых влияний приведена в [6].

Выявление возможных информационных операций

Сеть авторства источников информации позволяет оперативно идентифицировать возможные информационные операции в соответствии с подходами, предложенными в [2]. Предполагается, что вероятность информационной операции мала, если информация о событии вначале зарождается во влиятельном информационном источнике, а затем перепечатывается (со ссылками или без них) менее влиятельными источниками (рис. 3). Обратные явления, когда более влиятельные издания перепечатывают информацию из менее влиятельных источников, пусть и многочисленных, могут

служить признаком информационной операции, атаки (рис. 4). Именно такие картины наблюдались при сетевом анализе реальных тематических информационных потоков.

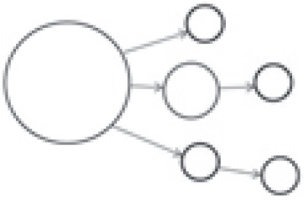

Рис. 3. Типовой сценарий распространения информации

Рис. 4. Сценарий распространения информации, характерный для информационной операции

В [1] приведены шаги по противодействию выявленной (или возможной) информационной операции. Некоторые из указанных шагов эффективно решаются в рамках предложенных далее технологических подходов.

1. Сбор информации с публикациями об объекте атаки.

2. Анализ контента публикаций в критических точках.

3. Определение источников, публикующих негативную информацию об объекте атаки.

4. Определение «первоисточников», т. е. тех источников, которые первыми опубликовали негативную информацию.

5. С учетом реалий и публикаций оценка вероятных последствий.

6. Организация информационного противодействия, диалога с наиболее влиятельными изданиями и пр.

Примеры публикаций в контексте информационного противодействия находятся в ретроспективной базе данных системы контент-мониторинга, а ранжировка, «вес» источников информации может определяться на основе предложенной методики.

Выводы

♦ Представлены технология и методика ранжировки веб-сайтов по авторству на основе оценки контекстных ссылок и перепечаток информации. Предложен подход к оперативному выявлению информационных операций на основе анализа сетей авторства источников информации.

♦ Описана также технология определения значимости различных источников информации — веб-ресурсов, а соответственно, и влияния на конечных потребителей информации — пользователей сети Интернет. Эта технология базируется как на современных методах и инструментальных средствах контент-мониторинга глобальных сетей, так и на современных подходах Text Mining, в часности ранжировки узлов в информационных сетях, средствах анализа и визуализации информационных потоков.

Данную технологию можно использовать в качестве основы для обнаружения различных видов информационного влияния на основе исследования контента современных компьютерных сетей.

Список использованной литературы

1. ***Горбулін В. П., Додонов О. Г., Ланде Д. В.*** *Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія. Київ, 2009. 164 с.*

2. ***Прищепа С. В.*** *Выявление события, его субъекта и объекта в текстовых документах // Информационные технологии и безопасность: материалы XVI Междунар. науч.-практ. конф. ИТБ-2016. Киев: ИПРИ НАН Украины, 2017. С. 121–129.*

3. ***Ланде Д. В.*** *Елементи комп'ютерної лінгвістики в правовій інформатиці. Київ: НДІІП НАПрН України, 2014. 168 с.*

4. ***Зеленков Ю. Г, Сегалович И. В.*** *Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Тр. 9-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL, 2007. Т. 1. С. 166–174.*

5. ***Kleinberg J. M.*** *Authoritative Sources in a Hyperlinked Environment // Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.*

6. ***Snarskii A. A., Zorinets D. I., Lande D. V.*** *Conjectural links in complex networks // Physica A: Statistical Mechanics and its Applications, 2016. Vol. 462. P. 266–273.*

Рецензент: канд. техн. наук, доцент **А. П. Бондарчук,** Государственный университет телекоммуникаций, Киев.

В. В. Вишнівський, С. В. Прищепа, Д. В. Ланде

РАНЖУВАННЯ ВЕБ-САЙТІВ ПОДІЄВОГО ХАРАКТЕРУ

У статті описано технологію побудови мережі авторства інформації на веб-сайтах подієвого характеру за результатами аналізу контекстних посилань. Запропонована технологія базується на врахуванні взаємного впливу джерел інформації, вираженого у вигляді посилань у тексті або передруків істотних фрагментів тексту — ознак, що дозволяють визначати першоджерела інформації. Запропоновано методи та засоби аналізу мереж джерел подієвої інформації.

Ключові слова: інформаційні операції; інформаційні потоки; виявлення подій; кластерний аналіз; модель предметної області.

V. V. Vyshnivskyi, S. V. Prishchepa, D. V. Lande

RANKING OF EVENT WEB-SITES

The article describes the methodology of automatic detection of the event basis of information operations, reflected in thematic information flows. The implementation of this methodology allows you to determine the time frame of the information operation, identify the main events that accompany the information operation, and see the techniques of information impact. Information operations in practice are supported by numerous events, processes, actions. It is assumed that a systematic violation of the typical dynamics of some thematic information flows in the open information space may indicate information operations. In the study of information operations, much attention is also paid to the analysis of the dynamics of information flows, a typical template of information operations, which allows using available analytical tools, for example, correlation analysis, to identify them. The proposed methodology for identifying the event basis of information operations covers the following stages of the research: the formation of a thematic information flow on the topic under study; investigation of the dynamics of the received flow, identification of signs of an information operation; narrowing the timeframe of the information flow, obtaining a representative sample of documents; the definition of the terminological basis for the description of events within the scope of the subject area under study. Identify sources of information about events (if necessary); cluster analysis, identify the main events that accompany the information operation. It is suggested to use as reference primary centroids of clusters certain support words that express the essence of events. The proposed approach allows simple implementation of clustering within the capabilities of the content monitoring system InfoStream. Clusters formed in this way reflect the main events occurring during information operations, provide an opportunity to identify participants in information confrontations, and methods of informational influence that disclose the technique for their implementation.

Keywords: information operations; information flows; detection of events; cluster analysis; model domain.

УДК 004.8+65.05+681.5

Ю. І. КАТКОВ, канд. техн. наук, доцент,
Державний університет телекомунікацій, Київ

АНАЛІЗ ПРИЧИН КРИТИЧНИХ СИТУАЦІЙ В ІНФОРМАЦІЙНО-ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ

Розглянуто логічний ряд понять, пов'язаних із категорією «небезпека», що дозволить ранжувати спектр можливих загроз безпеці функціонування складної організаційно-технічної системи (СОТС) у критичних ситуаціях і сприятиме розробці затребуваних адекватних технологій протидії впливам зовнішнього середовища на інформаційну систему із мінімізацією негативних наслідків і, зрештою, дасть змогу вдосконалювати існуючу систему забезпечення безпеки СОТС. З огляду на згаданий логічний ряд понять з'ясовано причини виникнення критичних ситуацій в інформаційно-інтелектуальних системах і запропоновано класифікацію факторів-стресорів, зумовлюваних виникненням критичних знань, котрі, у свою чергу, призводять до критичного стану СОТС.

Ключові слова: інтелектуальні системи; критичний стан системи; фактор-стресор; невизначеність; ризик; виклик; загроза.

ВСТУП

Сьогодні неможливо уявити ***складну організаційно-технічну систему*** **(СОТС)**, що функціонує без упровадження ***інформаційних систем*** **(ІС).** Наприклад, енергетичні та транспортні системи, телекомунікаційні мережі, PDM-система (система управління даними про виріб), виробничі організації тощо, які належать до складу СОТС, неодмінно включають у себе ІС.

Відомо, що ІС у складі СОТС призначено для отримання, обробки, зберігання, відображення чи реєстрації даних про технічний стан конструкцій, систем, елементів, їхні властивості, особливості функціонування заради забезпечення інформаційних потреб користувачів у разі прийняття рішень стосовно управління технологічними процесами, а також для адаптації елементів СОТС до впливів зовнішнього середовища.

Сучасну ІС можна розглядати як окремий вид організаційно-технічної системи щодо обробки інформації за допомогою технічних і програмних засобів, котра для підвищення якості функціонування СОТС дедалі більше насичується інтелектуальними технологіями на основі штучного інтелекту, з упровадженням до її складу різноманітних ***інтелектуальних систем*** **(ІнтС).**

ІнтС використовуються для генерування варіантів рішень, з яких особа, що приймає рішення, вибирає потрібне. Отже, метою ІнтС є моделювання розумових процесів організації знань, притаманних людині при прийнятті рішень у різних галузях соціально-економічної сфери суспільства на основі засобів і методів штучного інтелекту. Тому сучасні ІС неухильно стають інформаційно-інтелектуальними системами.