Исследование источников информационного влияния веб-ресурсов сети Интернет

© Додонов А.Г. © Ландэ Д.В.
Институт проблем регистрации информации Национальной академии наук Украины, Киев, Украина
dodonov@ipri.kiev.ua dwlande@gmail.com

Аннотация

В статье описывается технология построения сети влияния источников информации на основе анализа контекстных ссылок. Технология включает методы и средства, базирующиеся на контент-мониторинге глобальных сетей, концепциях Complex Networks и Text Mining. В отличие от методов анализа гиперссылок в сетевых документах, применяемых для анализа популярности веб-страниц в Интернете, предлагаемая технология учитывает взаимное влияние источников информации, выраженное в виде ссылок в тексте или перепечаток существенных фрагментов текста. Представлены методы и средства анализа сетей взаимного влияния источников информации, отражающих различные тематические срезы, а также информационные операции. Предложен подход к оперативному выявлению информационных операций на основе анализа сетей взаимных ссылок источников информации.

В работе также представлена технология выявления значимости информационного взаимного влияния различных источников информации – веб-ресурсов, а соответственно, и на конечных потребителей информации – пользователей сети Интернет. Данная технология базируется как на современных методах и инструментальных средствах контент-мониторинга глобальных сетей, так и на современных подходах Text Mining, распознавания образов, ранжирования узлов в информационных сетях, средствах анализа и визуализации информационных потоков.

Процедура исследования взаимного влияния источников информации, охватывает такие шаги, как получение репрезентативного массива публикаций по выбранной тематике; выявление контекстных ссылок и перепечаток в тематическом информационном потоке; построение сети влияния источников информации на основе анализа контекстных ссылок и перепечаток; исследование сети влияния источников информации; ранжирование узлов по степени влиятельности; выявление возможных информационных операций и построение сценария противодействия информационным операциям в сетевой среде.

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

Ключевые слова: веб-ресурсы, информационное влияние, контентные ссылки, информационные операции, ранжирование узлов сети.

1 Введение, постановка задачи

В настоящее время Интернет представляет собой значимый фрагмент информационного пространства, его влияние на людей постоянно возрастает. При этом следует учитывать различные механизмы распространения информации в сети, взаимного влияния источников информации, через которых и осуществляется воздействие на пользователей. Информационное пространство Интернета (веб-ресурсы, социальные сети) сегодня является мощнейшей площадкой для проведения информационных операций [1], признаки которых позволяют определять различные методики [2, 3].

В данной работе предлагается технология определения влиятельности сетевых источников информации на основе анализа контекстных ссылок. Представлена технология и методика ранжирования источников информации на основе оценки контекстных ссылок. В отличие от методов анализа гиперссылок в сетевых документах, применяемых для анализа популярности веб-страниц в Интернете, в предложенном подходе учитывается взаимное влияние источников информации, выраженное ссылками в тексте и перепечатками существенных фрагментов текстов. При этом предполагается, что влияние источников информации друг на друга определяется наличием контекстных ссылок или перепечаток (см. рис. 1) Также предложен подход к оперативному выявлению информационных операций на основе анализа сетей взаимных ссылок источников информации.

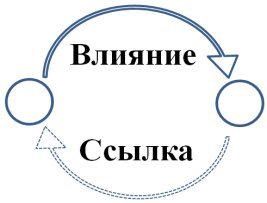

Рис. 1. Гипотеза о соотношении наличия контекстных ссылок и влияния источников информации

2 Технологические этапы исследования взаимного влияния источников информации

Для эффективного исследования взаимного влияния источников информации из сети Интернет (веб-ресурсов, социальных медиа) предлагается последовательность шагов, этапов обработки информации, каждый из которых сам по себе обеспечивает получение аналитического продукта. Совокупность таких этапов, базирующихся на использовании необходимых и доступных инструментальных средств, специальных приемов, можно рассматривать как процедуру проведения действий, нацеленных на получение аналитических материалов, включающих построение и анализ сети их взаимного влияния.

При проведении данных информационно-аналитических исследований на базе контент-мониторинга к таким задачам можно отнести:

- Нахождение релевантных публикаций по заданной тематике.
- Выявление взаимных контекстных ссылок и перепечаток в документах, представленных разными информационными источниками.
- Построение сети влияния, анализ и визуализация взаимосвязей информационных источников, в том числе ранжирование узлов построенной сети по степени влиятельности.
- Выявление возможных информационных операций и построение сценария противодействия информационным операциям в сетевой среде.

Соответственно процедура исследования взаимного влияния источников информации, охватывает такие шаги:

1. Получение репрезентативного массива публикаций по выбранной тематике.
2. Выявление контекстных ссылок и перепечаток в тематическом информационном потоке.
3. Построение сети влияния источников информации на основе анализа контекстных ссылок и перепечаток.
4. Исследование сети влияния источников информации, ранжирование узлов по степени влиятельности.
5. Выявление возможных информационных операций и построение сценария противодействия информационным операциям в сетевой среде.

Рассмотрим эти шаги подробнее на конкретных примерах.

3 Получение репрезентативного массива публикаций

Для получения репрезентативного массива публикаций по выбранной тематике необходимо выбрать систему контент-мониторинга, предоставляющую поток информационных сообщений по определенной тематике. Тематика может выражаться запросом на языке информационно-поисковой системы.

В качестве системы контент-мониторинга авторами была выбрана система InfoStream, которая в настоящее время охватывает 10 тыс. источников информации на русском и украинском 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. 2 примере показан фрагмент интерфейса системы, через который отрабатывался запрос, относящейся к обсуждению в январе 2016 года вопроса от ставки премьер-министра Украины А. Яценюка. В результате был сформирован тематический информационный массив, охватывающий 3196 документов.

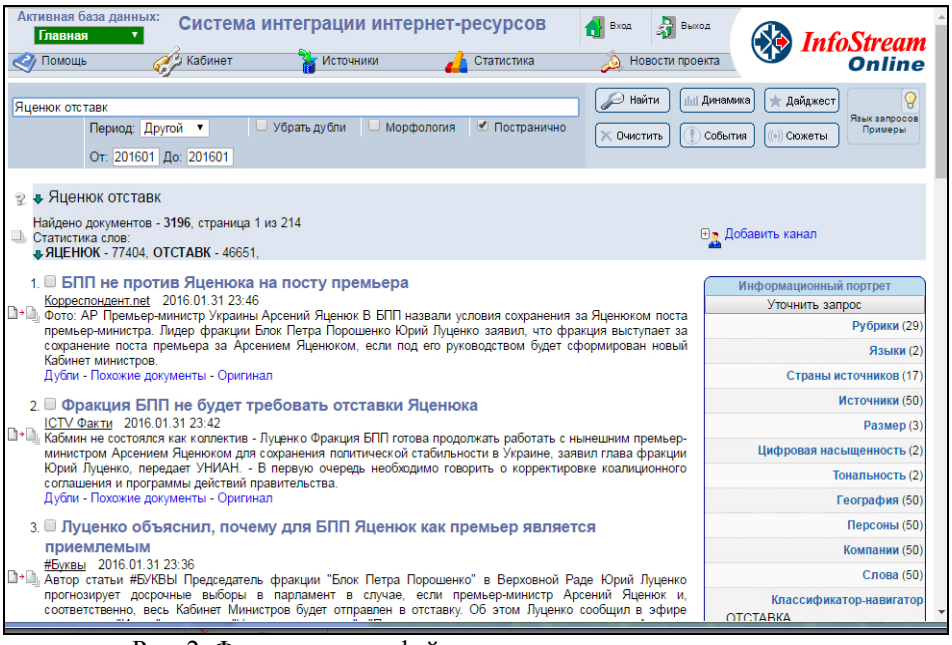

Рис. 2. Фрагмент интерфейса системы контент-мониторинга

4 Выявление контекстных ссылок

Основой построения сети влияния источников информации являются контекстные ссылки и перепечатки в тематическом информационном потоке. Контекстные ссылки выявляются путем идентификации шаблонов (Табл. 1) в документах выбранного информационного массива и признаков точных перепечаток, определяемых методами выявления плагиата [4, 5]. В свою очередь, сами шаблоны периодически определяются/дополняются экспертами в автоматизированном режиме путем анализа контекста потока документов системы контент-мониторинга методами Text Mining.

5 Построение сети влияния источников информации

Найденные в текстах контекстные ссылки и перепечатки позволяют сформировать матрицу цитирования, транспонируя которую в соответствии с приведенный выше гипотезой формируется матрица влияния. Данной матрице соответствует сеть влияния источников, пример визуализации которой для рассмотренного выше тематического информационного массива с помощью системы Gephi приведен рис. 3.

Таблица 1. Шаблоны названий информационных ресурсов (фрагмент)

№	*Код источника*	*Шаблон 1*	*Шаблон 2*
1	srd06193	Depo	"Депо"
2	srd03176	Українські національні новини	УНН
3	srd00045	Сегодня.ua	"Сегодня"
4	srd03076	ТСН.ua	"ТСН"
5	srd07509	112.ua	"112"
6	srd02348	Gazeta.ua	
7	srd00069	Корреспондент.net	"Корреспондент"
8	srd07487	Еспресо TV	"Еспресо ТВ"
9	srd02535	Телеканал новин "24"	"24"
10	srd06453	Телеграф.com.ua	"Телеграф"
11	srd01351	ЗІК	"ЗІК"
12	srd02514	РБК-Україна	РБК-Украина
13	srd04508	Українські Новини	
14	srd07686	"Антикор"	
15	srd00253	"Обозреватель"	
16	srd00057	Интерфакс	Інтерфакс
17	srd00404	ICTV Факти	ICTV
18	srd04125	РИА Новости Украина	
19	srd02732	УКРІНФОРМ	УКРИНФОРМ
20	srd00095	УНІАН	УНИАН
21	srd00094	Українська правда	
22	srd01408	Цензор.Нет	
23	srd00064	ЛІГАБізнесІнформ	
24	srd00039	Газета День	
25	srd07038	Вести.ua	

Найденные в текстах контекстные ссылки и перепечатки позволяют сформировать матрицу цитирования, транспонируя которую в соответствии с приведенной выше гипотезой формируется матрица влияния. Данной матрице соответствует сеть влияния источников, пример визуализации которой для рассмотренного выше тематического информационного массива с помощью системы Gephi приведен рис. 3.

6 Исследование сети влияния источников информации

Построенную сеть влияния источников информации можно исследовать как с помощью общепринятых инструментальных средств (например, с помощью системы Gephi были получены такие параметры построенной сети, как количество узлов: 141, ребер: 196, плотность графа: 0,01, средний коэффициент кластеризации: 0,026, средняя длина пути: 1,26 и т.д.).

Для содержательного анализа большое значение имеет вес узлов сети, список самых весомых узлов по критерию исходящей мощности приведен в Таблице 2.

Перспективным подходом к ранжированию источников по уровню влияния является алгоритм HITS, предложенный Дж. Кляйнбергом [6].

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

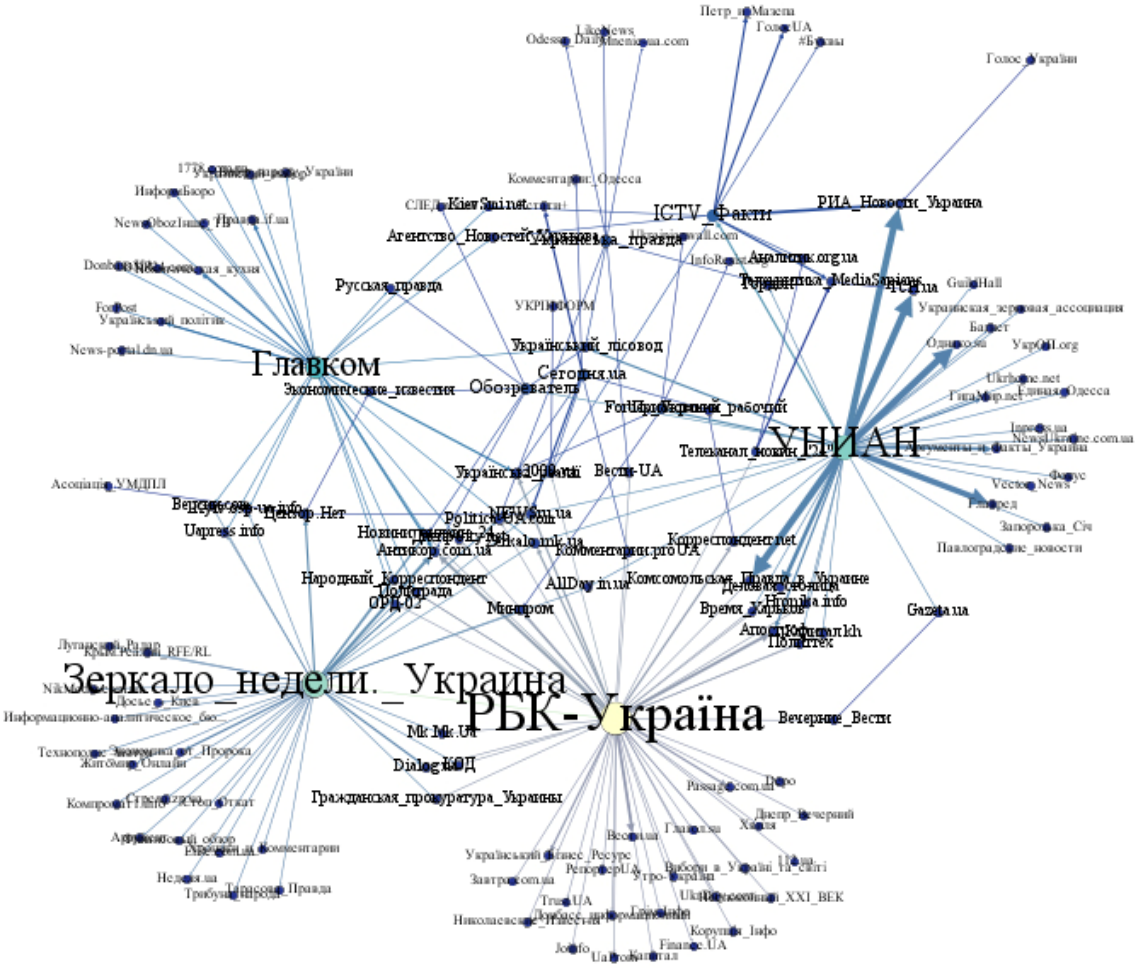

Рис. 3. Фрагмент сети связей источников по выбранной тематике

Таблица. 2. Наиболее влиятельные узлы по количеству цитирования

№	Веб-ресурс	Исходящая мощность
1	РБК-Україна	50
2	Зеркало недели	38
3	УНИАН	35
4	Главком	28
5	ICTV-Факты	10
6	Сегодня.ua	7
7	Українська правда	7
8	Обозреватель	6
9	Forbes-Украина	4
10	Цензор.Нет	3

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:

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

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

Перефразируя обозначения, приведенные в [6], а именно заменяя «авторство» на «подверженность влиянию», а «посредничество» на «влиятельность» можно с небольшими вычислительными затратами вычислять соответствующие характеристики узлов сети влияния.

Также для выявления информационных влияний большое значение имеет определение «скрытых» связей, т.е. когда прямых связей между узлами нет, но прослеживаются связи через другие (вторые, третьи и т.п. узлы). Методика определения скрытых связей, скрытых влияний приведена в работе [7].

7 Выявление возможных информационных операций

Сеть информационного влияния источников информации позволяет оперативно идентифицировать возможные информационные операции в соответствии с подходами, предложенными в работе [2]. Предполагается, что вероятность наличия 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. 4). Обратные явления, когда более влиятельные издания перепечатывают информацию у менее влиятельных, пусть и многочисленных, может являться признаком информационной операции, атаки (рис. 5). Именно такие картины наблюдались при сетевом анализе реальных тематических информационных потоков (рис.6).

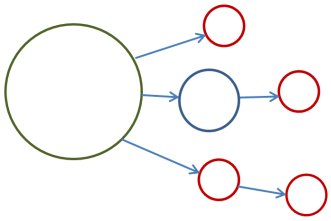

Рис. 4. Типовой сценарий распространения информации

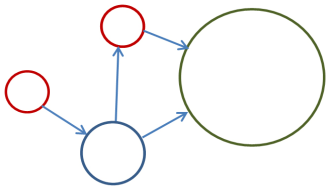

Рис. 5. Сценарий распространения информации, характерный для информационной операции

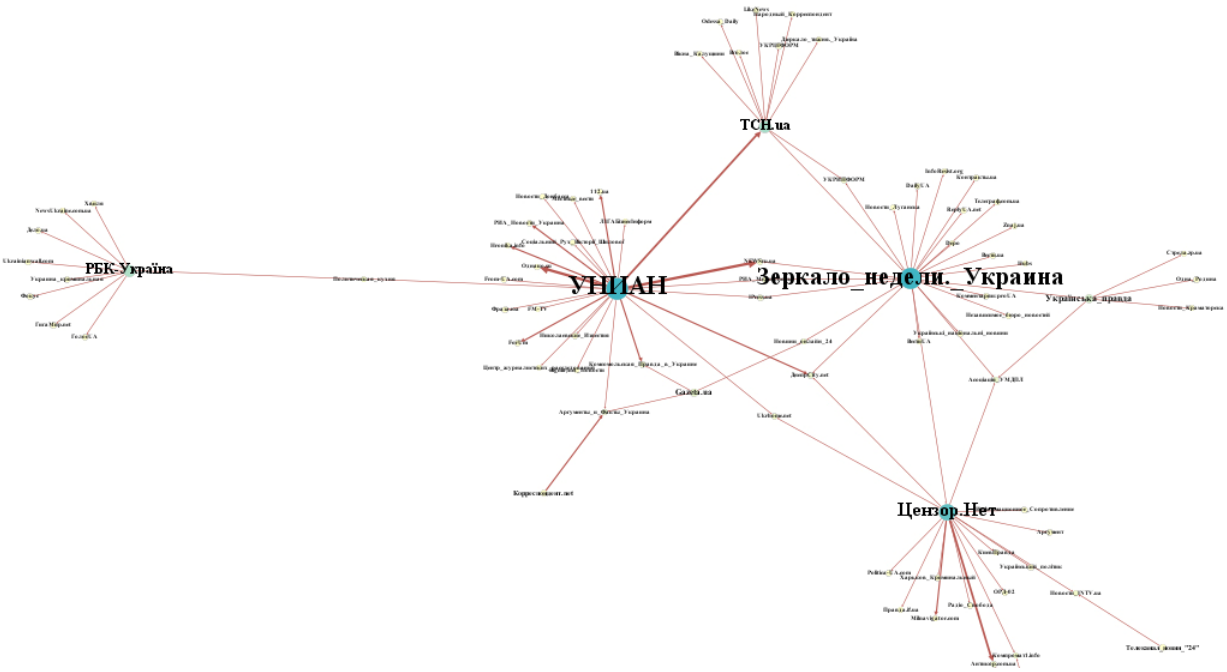

Рис. 6. Фрагмент сети связей источников по специальной тематике

В работе [1] приведены шаги противодействию выявленной (или возможной) информационной операции, некоторые из которых эффективно решаются в рамках предложенных технологических подходов, а именно:

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» – тех источников, которые первыми опубликовали негативную информацию.
5. С учетом реалий и публикаций оцениваются вероятные последствия.
6. Организуется информационное противодействие, диалог с наиболее влиятельными изданиями и т.д. Примеры публикаций в контексте информационного противодействия находятся в ретроспективной базе данных системы контент-мониторинга.

8 Выводы

Таким образом, представлена технология и методика ранжирования источников информации по влиятельности на основе оценки контекстных ссылок. Предложен подход к оперативному выявлению информационных операций на основе анализа сетей взаимных ссылок источников информации.

В работе также представлена технология выявления значимости информационного взаимного влияния различных источников информации – веб-ресурсов, а соответственно, и на конечных потребителей информации – пользователей сети Интернет. Данная технология базируется как на современных методах и инструментальных средствах контент-мониторинга глобальных сетей, так и на современных подходах Text Mining, распознавания образов, ранжирования узлов в информационных сетях, средствах анализа и визуализации информационных потоков.

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

Публикация содержит результаты исследований, проводимых при грантовой поддержке Государственного фонда фундаментальных исследований по конкурсному проекту Ф73 № 23558 “Разработка методов и средств поддержки принятия решений при обнаружении информационных операций”.

Литература

1. Горбулін В.П., Додонов О.Г., Ланде Д.В. Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія. – К.: Інтертехнологія, 2009. – 164 с.
2. Потемкин А.В. Распознавание информационных операций средств массовой информации сети Интернет // Науковедение, 2015. –Том 7, №3. – URL: http://naukovedenie.ru
3. Додонов А.Г., Ландэ Д.В., Прищепа В.В., Путятин В.Г. Конкурентная разведка в компьютерных сетях.– К.: ИПРИ НАН Украины, 2013. – 248 с.
4. Ланде Д.В. Елементи комп'ютерної лінгвістики в правовій інформатиці. – К.: НДІІП НАПрН України, 2014. – 168 с.
5. Зеленков Ю.Г, Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL, 2007. – Т. 1. – С. 166-174.
6. Kleinberg J.M. Authoritative Sources in a Hyperlinked Environment // Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998, and as IBM Research Report RJ 10076, May 1997.
7. Snarskii A.A., Zorinets D.I., Lande D.V. "Conjectural" links in complex networks // Physica A: Statistical Mechanics and its Applications, 2016. – Vol. 462. – pp. 266-273.

A Study of Informational Impact Sources of the World Wide Web

© Aleksandr G. Dodonov © Dmitry V. Lande
Institute for information recording of National academy of sciences of Ukraine,
Kyiv, Ukraine
dodonov@ipri.kiev.ua dwlande@gmail.com

Abstract

The paper describes a technology for building a network of information sources impact, based on analysis of contextual links. The technology incorporates methods and means, based on context-monitoring of global networks, Complex Networks and Text Mining concepts. In contrast to methods of hyperlink analysis in network documents, used to analyze popularity of Internet pages, the suggested technology utilizes mutual impacts of information sources, represented by references in the text or by reprints of existing text fragments. The paper presents methods and means of analysis of information sources’ mutual impact networks, reflecting various topical scopes as well as informational operations. An approach to efficient detection of informational operations based on analysis of information sources’ mutual impact networks is suggested.

The research also introduces a technology for defining the significance of mutual impacts of different information sources (web-resources), and, consequently, of their impacts upon the end users of information, i.e. Internet users. The technology is based on both modern methods and tools of global network content monitoring and present-day concepts of Text Mining, image recognition, ranking of information network nodes, means of analysis and visualization of information flows.

The procedure of analysis of information sources' mutual impacts involves the following phases: obtaining a representative set of publications within the selected topical scope; detection of contextual references and reprints in the topical information flow; construction of information source impact network based on contextual references and reprints; study of information source impact network; ranking of nodes according to impact degrees; detection of

potential informational operations, and development of a scenario for counteracting informational operations in the network environment.

The suggested technology can be used as basis for detection of different kinds of informational impact based on study of modern computer networks' content.

Keywords: web-resources, informational impact, content links, informational operations, ranking of network nodes.