

**Федеральное государственное автономное  
образовательное учреждение высшего образования  
«Национальный исследовательский университет  
«Высшая школа экономики»**

**Факультет компьютерных наук  
Основная образовательная программа  
«Прикладная математика и информатика»**

## **ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**Программный проект на тему  
Разработка и внедрение  
метрики скорости  
реакции Поиска Яндекса  
на события**

**Выполнил студент группы БПМИ191, 4 курса,  
Козлов Денис Михайлович**

**Руководитель от НИУ ВШЭ:  
Доцент, Соколов Евгений Андреевич**

**Соруководитель:  
Приглашенный преподаватель, Макоева Берта Ермаковна**

**Москва 2023**

# 1. Оглавление

1.	<i>Оглавление .....</i>	<i>2</i>
2.	<i>Аннотация .....</i>	<i>3</i>
3.	<i>Введение .....</i>	<i>3</i>
4.	<i>Обзор литературы .....</i>	<i>4</i>
4.1.	<i>Выводы.....</i>	<i>6</i>
5.	<i>Требования к метрике .....</i>	<i>6</i>
6.	<i>Тривиальная версия метрики .....</i>	<i>8</i>
7.	<i>Первый прототип оффлайн-метрики .....</i>	<i>8</i>
7.1.	<i>Определение свежих запросов.....</i>	<i>8</i>
7.2.	<i>Поиск информации о событии .....</i>	<i>9</i>
7.3.	<i>Вычисление метрики .....</i>	<i>10</i>
7.4.	<i>Результаты работы первого прототипа .....</i>	<i>10</i>
8.	<i>Второй прототип оффлайн-метрики.....</i>	<i>11</i>
8.1.	<i>Определение свежих запросов.....</i>	<i>11</i>
8.2.	<i>Поиск информации о событии .....</i>	<i>11</i>
8.3.	<i>Вычисление метрики .....</i>	<i>12</i>
8.4.	<i>Результаты работы второго прототипа .....</i>	<i>12</i>
9.	<i>Финальная версия метрики.....</i>	<i>12</i>
9.1.	<i>Определение свежих запросов.....</i>	<i>12</i>
9.2.	<i>Поиск информации о событии .....</i>	<i>13</i>
9.3.	<i>Вычисление метрики .....</i>	<i>13</i>
10.	<i>Заключение .....</i>	<i>13</i>
11.	<i>Список источников .....</i>	<i>14</i>

## 2. Аннотация

Сегодня поисковые системы являются одними из самых важных сервисов для пользования интернетом. В этой работе описывается разработка метрики, созданной для измерения качества Поиска Яндекса по отношению к свежим новостям. Эта метрика должна дать разработчикам поиска возможность проще следить за эффектом своих изменений.

Ключевые слова: поисковые системы, тематическое моделирование, анализ данных, машинное обучение

Nowadays, search engines are among the most important services on the Web. This paper describes a development of a metric, used for measuring search quality of Yandex Search regarding fresh news queries. The metric is aimed to help developers monitor their progress.

Keywords: search engines, topic models, data mining, machine learning

## 3. Введение

Задача поисковых систем — находить качественные документы, релевантные запросам пользователей. Яндекс это одна из самых популярных поисковых систем с миллионами пользователей ежедневно.

Заметной долей запросов являются запросы, связанные с событиями, которые произошли совсем недавно: так называемые «свежие» запросы. Чтобы правильно обработать такие запросы, нужно качественно сработали несколько частей Яндекса:

- Поисковый робот должен обойти эти новые документы и добавить их в свои базы
- Алгоритмы Поиска должны определить, что пользовательский запрос связан с новостными событиями
- Из новых документов нужно найти качественные, подходящие по теме, и показать их на выдаче.

Все эти задачи сложны сами по себе, и ими постоянно занимаются команды разработчиков, у которых есть свои метрики, измерения. Однако Поиску также нужны метрики, оценивающие всю систему полностью.

В этой работе описывается разработка метрики скорости реакции поиска. Эта метрика создается для того, чтобы оценивать, как быстро качественные новостные документы оказываются на выдаче. Улучшение такой метрики будет значить, что больше людей могут в первые минуты новостного события задать вопросы в Яндекс и получить релевантную выдачу. Метрика должна учитывать события разного масштаба на разные темы, учитывать разнообразие пользователей и их интересов, а результаты должны быть подробными и интерпретируемыми. Вычисление значений метрики должно быть быстрым и сравнительно дешевым.

На конкретные подходы в реализации метрики влияли работы из области информационного поиска (information retrieval) и методов обнаружения полезных данных (data mining). Для разработки и вычисления значений метрики используются различные внутренние и внешние сервисы, в частности YTsaurus, Toloka.

В работе описан процесс разработки метрики: разные подходы, их результаты, выводы. Сейчас метрика скорости реакции внедрена и уже работает, и активно продолжает развиваться и совершенствоваться.

## 4. Обзор литературы

**Поисковые системы** — вид программного обеспечения, направленный на удобный поиск по документам в интернете. Поисковые системы существуют десятки лет и являются одними из самых популярных сайтов в интернете. Несмотря на зрелость технологии, поисковые системы постоянно развиваются чтобы улучшить пользовательский опыт и качество ответов, например в последнее время многие поисковые системы начинают встраивать в свой поиск нейросети для лучшего взаимодействия с пользователями.

Для работы современных поисковых систем задействуется множество различных сервисов и технологий, но для этой работы стоит выделить следующие:

**Поисковый робот** — интернет-боты, задача которых обходить всемирную сеть и находить новые вебсайты для добавления в базу. Данные обойденных сайтов обрабатываются и добавляются в базу поиска. Страниц в интернете крайне много, поэтому чтобы с достаточной полнотой обходить свежие новости, алгоритмы роботов должны быть написаны качественно и аккуратно протестированы. Технология продолжает активно развиваться [1].

**Ранжирование веб страниц** — задача поиска максимально релевантных ссылок из поисковой базы на запрос пользователя. В случае больших поисковых систем задача крайне трудна, ведь подходящие документы нужно выбирать из миллиардов вариантов, а время работы и вычислительные затраты должны быть низкими, чтобы успеть обслужить огромный поток пользователей. Для этого поисковыми системами используется множество различных факторов запроса, документов из базы, и используются крайне оптимизированные алгоритмы [2].

Для вычисления значений метрики нужно обрабатывать большие объемы данных. Чтобы обработка больших распределенных данных проходила быстро, используется модель распределенных вычислений **MapReduce** [3]. В Яндексе используется платформа с открытым кодом **YTsauros** [4], с помощью которой обработка происходит быстро и надежно. В YTsauros предложен свой язык запросов YQL, который активно используется в работе для обращения к данным.

**Toloka** — краудсорсинговая платформа, используемая для разметки данных, в которой исполнители зарабатывают деньги решая простые задания. Краудсорсинг уже давно используется в работах по машинному обучению [5]. В этом проекте платформа Toloka используется для решения задач, которые не получается с должным качеством решать автоматически. Официальная документация платформы содержит множество полезной информации и примеров.

В этой работе проводится работа со свежими новостями, которые возникают, эволюционируют и затухают крайне непредсказуемо. Чтобы получать больше

информации можно обратиться к теме тематического моделирования (**topic modelling**) — использования статистических моделей для выявления тем из набора документов. Это полезно, чтобы определять хронологическую последовательность изменений в событиях [6, 7].

## 4.1. Выводы

Поисковые системы являются огромными сервисами с десятками внутренних сервисов, команд, метрик. Каждая компания реализовывает и оценивает внутренние сервисы по-разному. Данная работа описывает метрику, разработанную специально для использования внутри Яндекс Поиска, поэтому не имеет конкретных аналогов, с которыми было бы возможно провести сравнение, однако на тему исследования и моделирования поведения новостей существует немало научных работ, результатами которых можно воспользоваться в разработке.

## 5. Требования к метрике

Прежде чем приступать к описанию реализации метрики, стоит понять, что на нее влияет, на каких запросах мы ее хотим вычислять и другие детали.

Из новостных событий мы выделяем **внезапные** и **ожидаемые**. Ожидаемым событием может быть, например, футбольный матч или запланированное выступление политика — время этих событий известно заранее. Внезапные события отличаются тем, что о них заранее неизвестно — это может пост знаменитости в соцсетях или авария на дороге. В этом проекте рассматриваются только внезапные события, ведь именно для них критически важно быстро показывать на выдаче релевантные ссылки.

Чтобы правильно понимать точку отсчета начала события, требуется его определить. В нашем проекте было решено «началом события» называть время возникновения самого первого источника в интернете. Смысл в том, что основной поток интересующихся событием пользователей узнает о событии из интернета, а значит сможет задать вопросы только лишь после появления информации в сети. Также, если делать отсчет от настоящего времени события,

то на скорость реакции помимо технологий Яндекса будет влиять скорость, с которой журналисты узнают и описывают событие. Это особенно заметно в случаях, например смерти знаменитостей: близкие могут сколько-то времени хранить смерть в тайне, и только спустя какое-то время сообщить поклонникам об утрате. Действительно, в таком случае скорость реакции было бы некорректно вычислять времени смерти знаменитости, а лучше начинать замер от публичного сообщения.

Еще одним важным критерием является **масштабность** события.

- Во-первых, хочется учитывать как новости уровня страны и мира (землетрясение в Турции), так и новости на более нишевые тематики (новая песня не очень популярного исполнителя). В этих ситуациях много чего меняется: разный состав и количество новостных источников, которые пишут о событии; новостные источники сами могут с разной скоростью реагировать на новость. С другой стороны и количество интересующихся в первые минуты пользователей будет разное: случай, когда мы быстро отработали на глобальную тематику, но медленно на нишевую заметно отличается от обратного, потому что в обратном случае было бы больше неудовлетворенных пользователей, которым была показана еще не релевантная выдача.
- Во-вторых, можно пробовать учитывать скорость реакции не только на сами события, но и на изменения событий, такие как подтверждения или опровержения, появление дополнительной информации и так далее. Например, для события «пожар на складе в Москве», обновлениями будут сообщения о том, что пожар локализован, погашен, о том, что найден нарушитель и другие. Действительно, в момент, когда пожар уже потушен, выдача должна содержать в себе документы о тушении, а не только о том, что склад горит.

Имея этот набор условий, было решено приступить к работе над метрикой. В следующих главах будут описаны разные варианты реализации метрики.

## **6. Тривиальная версия метрики**

Пусть несколько сотрудников будут следить за свежими новостями, и когда они увидят новую новость — пусть начинают смотреть на выдачу Яндекса по разным подходящим поисковым запросам и периодически перепроверять, не появились ли релевантные документы на выдаче.

К сожалению, такое решение будет дорогим, а результаты шумными, и нерепрезентативными: вручную получится следить лишь за небольшим количеством источников, из-за чего могут потеряться многие темы, а какие-то — проявиться слишком поздно для полезности замера. Каждый замер будет редким и очень дорогим, что не позволит совершать долгие периодические замеры и делать по ним выводы.

После обсуждений с командой было принято решение разрабатывать метрику в оффлайн-режиме: будут анализироваться логи использования поиска, и значение метрики будет вычисляться по ним.

## **7. Первый прототип оффлайн-метрики**

В алгоритме вычисления скорости реакции поиска можно выделить три основные части:

### **7.1. Определение свежих запросов**

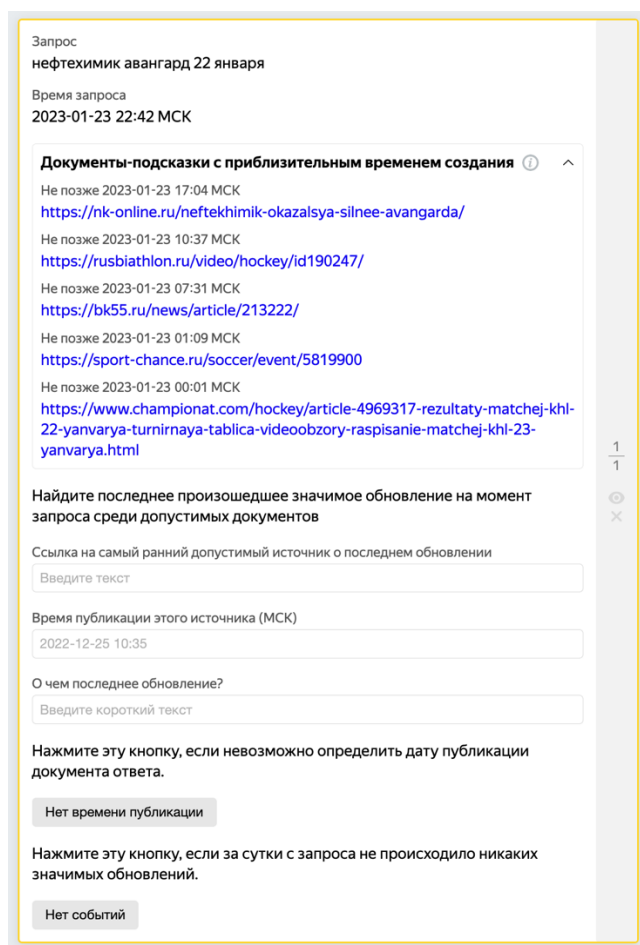
Для экономии времени было решено переиспользовать репрезентативную подборку свежих запросов, которая создается каждый день в рамках других внутренних процессов. Это решение имеет ограничения: в ежедневной подборке достаточно мало запросов, а значительная часть из них относится к событиям, которые не подходят по выбранным ранее критериям. Например, в подборке могут оказываться запросы про спортивные матчи, которые являются предсказуемыми событиями, и потому в метрике не используются, могут попадать запросы на общие темы. Из-за малого количества запросов в выборках, подходящие для вычисления метрики запросы выбирались силами разработчиков.



## 7.2. Поиск информации о событии

В первой реализации метрики было решено учитывать как начала событий, так и их обновления в том числе (не просто «произошла авария», но и «объявили количество пострадавших»). Для этого был разработан процесс в Toloka.

Исполнителям показывается запрос, время задания запроса, набор полезных ссылок, а задача исполнителя — определить последнее изменение в событии, найти первый источник информации и время его публикации и кратко описать суть события или обновления.



Запрос  
нефтехимик авангард 22 января

Время запроса  
2023-01-23 22:42 МСК

Документы-подсказки с приблизительным временем создания ⓘ ^

Не позже 2023-01-23 17:04 МСК  
<https://nk-online.ru/neftekhimik-okazalsya-silnee-avangarda/>

Не позже 2023-01-23 10:37 МСК  
<https://rusbiathlon.ru/video/hockey/id190247/>

Не позже 2023-01-23 07:31 МСК  
<https://bk55.ru/news/article/213222/>

Не позже 2023-01-23 01:09 МСК  
<https://sport-chance.ru/soccer/event/5819900>

Не позже 2023-01-23 00:01 МСК  
<https://www.championat.com/hockey/article-4969317-rezultaty-matchej-khl-22-yanvarya-turnirnaya-tablica-videoobzory-raspisanie-matchej-khl-23-yanvarya.html>

Найдите последнее произошедшее значимое обновление на момент запроса среди допустимых документов

Ссылка на самый ранний допустимый источник о последнем обновлении

Введите текст

Время публикации этого источника (МСК)

2022-12-25 10:35

О чем последнее обновление?

Введите короткий текст

Нажмите эту кнопку, если невозможно определить дату публикации документа ответа.

Нет времени публикации

Нажмите эту кнопку, если за сутки с запроса не происходило никаких значимых обновлений.

Нет событий

Рис. 7.2.1 Интерфейс задания поиска информации о событии

Полезные ссылки выбираются с помощью офлайн-данных, находя документы, которые были показаны Поиском Яндекса по аналогичным запросам, а также при условии, что документы были созданы близко по времени к запросу. Смысл полезных документов в том, что они помогут исполнителю быстрее понять контекст, суть происходящего.

В этой работе было решено допускать в качестве первоисточника почти все документы, а не только публикации в официальных СМИ: в наше время очень

много новостей сначала публикуются в соцсетях (ВК, Телеграм, Дзен и другие) [8], и только потом они оказываются на новостных сайтах.

### 7.3. Вычисление метрики

Зная время публикации источника, мы смотрим по логам на поисковые выдачи, которые по этому запросу показывались пользователям рядом с моментом начала события. Задача — определить, как быстро на выдаче начали появляться релевантные документы. Для этого из поисковых выдач выбираются показанные документы, из них фильтруются только документы, созданные недавно, и запускается разметка в Toloka. Исполнителям демонстрируется информация о событии (запрос, описание события, ссылка на первоисточник), а также документ из выдачи. Задача исполнителя — определить, сообщает этот документ о событии или нет. Когда исполнителями найден первый релевантный документ, задача оказывается решена — и мы можем вычислить, как много времени прошло с момента события до появления первой актуальной выдачи.

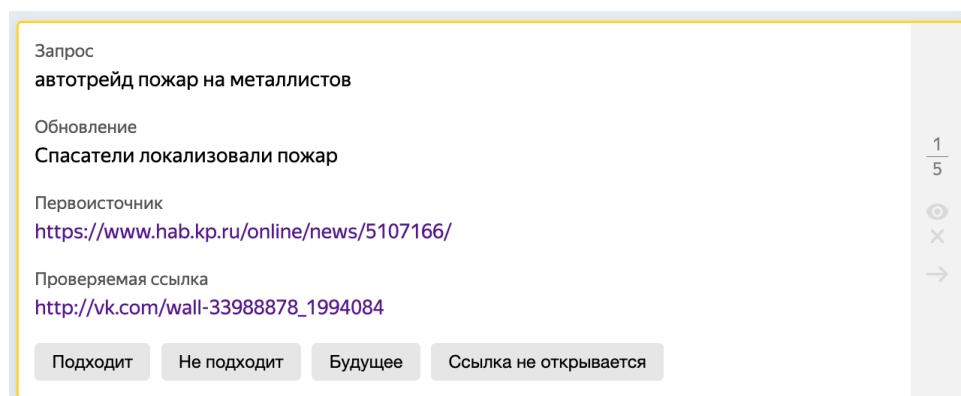


Рис. 7.3.1 Интерфейс задания проверки указанного обновления

### 7.4. Результаты работы первого прототипа

Прототип был запущен на 8 днях. Каждый раз все результаты разметки перепроверялись командой для поиска ошибок, уточнений.

В результате было обнаружено, что запросов в переиспользуемой нами чужой подборке не хватает, их нужно сильно больше для каких-либо точных и интерпретируемых результатов.

Было обнаружено, что задание по поиску информации о событии очень сложное: для выполнения требуется слишком много действий, и они все

непростые, из-за чего работать над заданием брались лишь единицы исполнителей, и все равно ошибались в своей разметке. Особенно сложно было разработать критерий для значимости обновления: какие обновления стоит игнорировать, а какие — нет. После множества споров внутри команды было решено в следующий раз работать без обновлений. К тому же качество выполнения задания было бы невозможно автоматически оценивать из-за множества разнообразных верных ответов.

## **8. Второй прототип оффлайн-метрики**

### **8.1. Определение свежих запросов**

Чтобы получать больше запросов для вычисления метрики было принято брать потоковую выборку запросов и из нее выбирать подходящие. Как и в первом прототипе, пока что это выполнялось силами разработчиков.

Смена подхода позволила лучше сформулировать критерии для следующих усовершенствований этапа, избавила команду от привязки к другому процессу с неподходящими запросами.

### **8.2. Поиск информации о событии**

Так как скорость реакции было решено вычислять от появления первой новости в интернете, задача сводится к поиску первой релевантной новости. Так как про запросы известно, что они свежие, то, по определению свежести, релевантные новостные документы могли быть созданы лишь незадолго до задания запроса.

С помощью внутренних нейросетей Яндекса выбираются большой набор кандидатов на релевантные пары запрос-документ. Запросы берутся из предыдущего этапа, а документы — из недавно созданных, согласно внутренним инструментам Поиска.

С помощью разметки в сервисе Toloka и внутренних алгоритмов из этих пар выбираются ссылки-первоисточники на все запросы. Дата создания первоисточника и является моментом отсчета для метрики.

### **8.3. Вычисление метрики**

Имея дату создания первого релевантного документа (то есть «начала события»), как и в первом прототипе рассматриваются поисковые выдачи на наличие актуальных релевантных документов. В этом прототипе было решено перейти на другое измерение: в периоде  $N$  часов с начала события проверяются все выдачи и замеряется доля поисковых выдач, на которых есть релевантные документы. Это показывает долю «удовлетворенных пользователей», которые могли увидеть нужные им новости на выдаче. Для большей интерпретируемости показаний метрики замеряются разные периоды времени от полчаса до 6 часов.

### **8.4. Результаты работы второго прототипа**

Второй прототип показал превосходство подхода над первым, исполнители в Toloka были больше удовлетворены заданиями, выполняли их быстрее, затраты на разметку были меньше. Сильное ускорение и упрощение разметки позволило получать достаточно подробные результаты менее чем за сутки, что продемонстрировало возможность запуска процесса каждый день.

Определение свежих запросов оставалось этапом с долгой ручной работой, поэтому было решено разработать модель машинного обучения, которая заменит ручную разметку. Запуски второго прототипа помогли собрать много данных для обучения этой будущей модели.

Благодаря качественной настройке процесса, улучшения, описанные в следующем разделе, могли применяться постепенно, не требуя переписывать процесс с нуля. Это сильно упростило разработку и позволило более явно следить за улучшениями.

## **9. Финальная версия метрики**

### **9.1. Определение свежих запросов**

С помощью набранных данных при тестировании прототипов с помощью внутренних сервисов Яндекса была разработана и обучена модель машинного

обучения, которая из потоковых запросов определяет лишь подходящие для процесса.

На вход модели помимо формулировки запроса подаются также множество синтетических показателей, в частности показатели, основанные на исторических данных. Например, крайне важной оказалась частотность запроса в прошлом: если сегодня запрос задают в разы чаще, чем на прошлой неделе, то скорее всего действительно какое-то событие произошло.

## **9.2. Поиск информации о событии**

На некоторые запросы внутренние нейросети выдавали крайне много кандидатов пар запрос-документ. Это связано как со спецификой обучения модели, так и с тем, что какие-то запросы являются более громким инфоповодом, и про них действительно пишут больше материалов. Чтобы экономить деньги и время на разметку в Toloka, был разработан ряд эвристик, фильтрующий точно неподходящие пары.

## **9.3. Вычисление метрики**

По результатам собранных данных за время тестирования прототипов было решено избавиться от этапа проверки релевантности свежих поисковых выдач. Гипотеза в том, что из-за высокого качества алгоритмов ранжирования в Поиске, все документы, созданные после начала события, а также попавшие на первую страницу выдачи, считаются релевантными. Действительно, лишь в единичных ситуациях на выдаче находились свежие, но нерелевантные документы.

Это позволило избавиться от дорогого и долгого этапа разметки в Toloka, заменив на быстрый алгоритм в парадигме MapReduce. Также это дало возможность получать более подробные результаты.

## **10. Заключение**

В этой работе описана разработка метрики скорости реакции Поиска Яндекса, которая поможет разработчикам оценивать эффект от своих улучшений. В результате получился готовый процесс, который периодически запускает

измерения и дает полезные интерпретируемые результаты. Несмотря на то, что метрика уже внедрена, она продолжает совершенствоваться и ускоряться.

## **11. Список источников**

[1] Контогианнис А. [Kontogiannis A.] “Tree-based Focused Web Crawling with Reinforcement Learning”, arXiv:2112.07620, Дек. 2021

[2] Ченг С., ЮнТао П. [S. Cheng, P. YunTao], и др. "PageRank, HITS and Impact Factor for Journal Ranking," 2009 WRI World Congress on Computer Science and Information Engineering, Лос Анджелес, США, 2009, С. 285-290, doi: 10.1109/CSIE.2009.351.

[3] Дин Дж., Гемават С. [J. Dean and S. Ghemawat], ‘MapReduce: Simplified Data Processing on Large Clusters’, OSDI’04: Sixth Symposium on Operating System Design and Implementation, 2004, С. 137–150.

[4] “YTsauros is a distributed storage and processing platform for Big Data,” ytsaurus.tech. URL:<https://ytsaurus.tech/> (дата обращения: 05.04.2023).

[5] Вортман Дж. [J. Wortman], ‘Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research’, JMLR, том. 18, N 193, С. 1–46, Сент. 2018.

[6] Чен [Chen] и др. Modeling Emerging, Evolving and Fading Topics Using Dynamic Soft Orthogonal NMF with Sparse Representation, 2015. С. 61-70. doi:10.1109/ICDM.2015.96.

[7] Лин [Lin] и др. Generating event storylines from microblogs. 2012, С. 175-184. doi:10.1145/2396761.2396787.

[8] Ньюман Н. [Newman N.] и др. Social Media in the Changing Ecology of News: The Fourth and Fifth Estates in Britain // Society and the Internet 2012 С. 135-148. doi:10.1093/acprof:oso/9780199661992.003.0009