УДК 004.77

А. В. Тихонов

Компания «Яндекс»
Московский физико-технический институт (национальный исследовательский университет)

Использование навигационных спектров для оценки медийности сайтов сети Интернет

Рекомендательные системы играют важную роль в современной сети Интернет, осуществляя поставку пользователям интересующей их информации. Развитие рекомендательных систем требует как понимания поведения пользователей, так и постоянного совершенствования качества предоставляемой информации. В настоящей работе показано, как использование метода навигационных спектров позволяет существенно улучшить качество оценки медийности сайта. Описывается, как применение данного метода позволило построить для более чем ста различных стран списки сайтов-кандидатов, использованных в дальнейшем в реальной рекомендательной системе.

Ключевые слова: анализ сети Интернет, пользовательское поведение, навигация в сети Интернет, классификация сайтов сети Интернет, рекомендательные системы.

A. V. Tikhonov

Company «Yandex»
Moscow Institute of Physics and Technology

Using navigation spectra for Internet media sites classification

Recommendation systems play an important role in the modern Internet, providing users with the information they need. The development of recommendation systems requires both an understanding of user behaviour and continuous improvement in the quality of the information provided. In this paper we demonstrate how the use of the method of navigation spectra can significantly improve the quality of media sites classification. We describe how the application of this method allows us to build whitelists of media sites for more than a hundred countries. These lists are later used in the real recommendation system.

Key words: internet analysis, user behaviour, internet navigation, web sites classification, recommendation systems.

1. Введение

Быстрый рост размера и сложности сети Интернет приводит к постепенному изменению пользовательского поведения и используемых механизмов навигации и поиска новой информации. Ручная навигация давно вытеснена специальными инструментами, в первую очередь – сервисами поддержки информационного поиска: поисковыми системами и рекомендательными системами.

При большом внешнем многообразии рекомендательных систем, большинство из них объединяет общий паттерн использования – пользователю предлагается непрерывный поток материалов, специально подобранных на основании его персональных интересов. Действуя как агрегаторы, подобные системы анализируют большое количество различных

материалов из разных источников, выбирая из них качественные и формируя индивидуальные предложения для пользователей системы. Одной из важнейших задач в работе рекомендательного агрегатора является обеспечение должного уровня качества материалов, в первую очередь, путём поиска и оценки как отдельных материалов, так и их возможных источников.

Размеры сети Интернет и темпы её роста делают неэффективным, если не невозможным, ручной поиск и анализ качества новых источников материалов. Так, если в 2009 году в сети функционировало более 100 миллионов различных сайтов [1], то по оценке 2019 года [2] их число составило уже более полутора миллиардов. Поэтому важной практической задачей является создание систем, поддерживающих автоматический анализ новых сайтов-кандидатов.

В настоящей работе описывается проект, в рамках которого разработанный автором метод навигационных спектров для описания страниц сети Интернет [3] был использован для построения факторов для классификатора медийных сайтов в рамках специализированного рекомендательного сервиса компании Яндекс. Полученный классификатор применялся для автоматического отбора сайтов-кандидатов для включения в рекомендательную систему в качестве источника.

2. Краткое описание метода навигационных спектров

Метод навигационных спектров развивает стандартный подход описания страниц, основанный на анализе источников переходов на страницу. Как показано автором в [3], учёт не только непосредственного источника перехода, но и полного предшествующего маршрута пользователя от начала сеанса работы до достижения целевой страницы позволяет более полно и точно описать роль страницы в сети Интернет. Предложенный способ выделения и классификации обобщённых маршрутов, компактно хранящих ключевую информацию обо всём пути пользователя с акцентами на точке начала маршрута, на точке достижения целевого сайта и на странице, непосредственно предшествующей целевой, позволяет сжато хранить эту информацию в виде навигационных спектров.

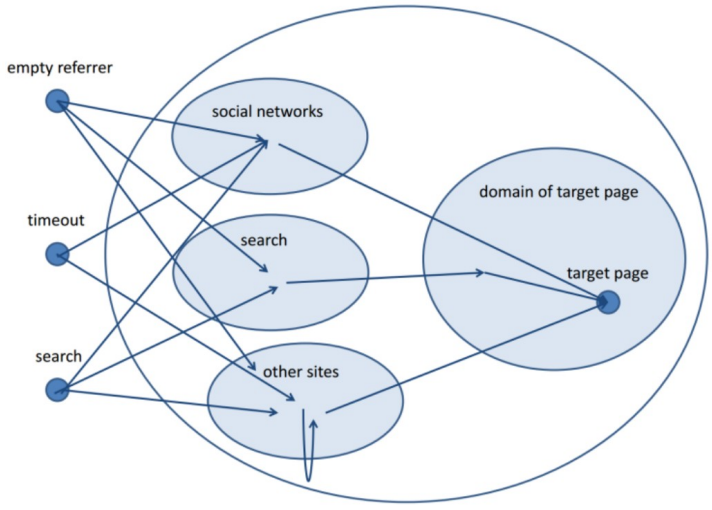

Рис. 1. Схема типов обобщённых маршрутов

Каждый визит на страницу имеет не более одного источника, содержащего адрес предыдущей посещенной страницы, при этом значение источника может отсутствовать (например, в случаях прямого ввода адреса страницы в браузер, открытия страницы из закладок и т.п.). Для унификации используется несколько выделенных типов источников: поисковые системы, социальные сети, страница на том же сайте, что и целевая, страница на внешнем

сайте, пустой источник. Выделение поисковых систем и социальных сетей в отдельные источники обусловленно тем, что они играют большую роль в навигации по современной сети Интернет, а сами пользователи, находясь на страницах социальных сетей, демонстрируют специфическое поведение [8]. Поисковыми считались страницы, принадлежащие домену одной из трех популярнейших поисковых систем в России: yandex.ru, google.ru и mail.ru (вместе данные системы покрывают около 97% всего российского поискового трафика). Страница была отнесена к социальным в случае, если она принадлежит домену одной из 22 наиболее популярных в России социальных сетей.

Согласно методу, описанному в [3], для каждой из страниц сначала строится множество содержащих её навигационных сессий, затем на их основе выстраивается реферальный лес, в рамках которого выделяются обобщённые маршруты, ведущие к данной странице (рис. 1). Вектор частот реализаций обобщённых маршрутов к странице составляет навигационный спектр данной страницы.

3. Прикладная задача и использованные данные

В рамках запуска рекомендательного сервиса компании Яндекс в новых странах стояла задача построения первичных списков сайтов-источников для более чем сотни различных стран. Список каждой страны должен был содержать несколько сотен отобранных сайтов. Сайты, включенные в список, должны были соответствовать ряду требований:

- искомые сайты должны содержать ленты актуальный информационный или развлекательный материал;
- публикуемый на них материал не должен быть вторичным (то есть требуется исключить сайты-агрегаторы);
- публикуемый материал должен быть доступен без оплаты;
- сайты должны иметь заметную аудиторию пользователей;
- публикация контента должна быть достаточно регулярной.

Такое количество стран делало ручное формирование и проверку подобных списков неэффективным. Традиционно подобные задачи принято решать с помощью классификации сайтов на основании их содержимого. Однако сильные различия в культурных и языковых особенностях разных стран не позволяли использовать этот подход для создания единого решения, применимого к сотне стран сразу. Для устранения этого препятствия было решено использовать дополнительные поведенческие факторы классификатора, рассчитанные на основании навигационных спектров страниц сайтов-кандидатов.

Формирование первичных списков сайтов-кандидатов и расчёт значений всех необходимых факторов происходили на основании массива данных онлайн-панели SimilarGroup за период с ноября 2015 по май 2016 года. Панель предоставляет навигационную статистику десятка миллионов пользователей по всему миру и охватывает около 200 стран. Статистика панели является коммерческим продуктом компании интернет-измерителя SimilarWeb и, согласно информации на её официальном сайте, обладает следующими свойствами:

- данные панели получены путём объединения четырёх типов источников:
  - базовая интернет-панель, содержащая около 10 миллионов человек по всему миру;
  - данные некоторых интернет-провайдеров;
  - сайтоцентричные счётчики посещаемости, размещённые на крупных порталах;
  - данные интернет-роботов, собирающих дополнительную статистику;

- типичная запись массива данных включает в себя следующие поля:
  - уникальный обезличенный идентификатор пользователя;
  - код страны пользователя в формате ISO Numeric Country Code;
  - значение HTTP Referer, то есть адрес предыдущей страницы, с которой был осуществлён переход на данную;
  - временную метку посещения страницы.

Таким образом, этот массив данных содержал навигационную статистику нескольких десятков миллионов пользователей по всему миру в формате, допускающем извлечение маршрутов и построения навигационного спектра, и охватывает около 200 стран.

В первой фазе проекта на основании исходного массива данных были построены страновые списки сайтов-кандидатов, удовлетворяющих базовым требованиям – доступности сайта из заданной страны и достаточной популярности сайта среди пользователей из заданной страны. В общей сложности в списки сайтов-кандидатов вошло 1 076 117 сайтов.

Затем с помощью нескольких привлечённых экспертов были вручную построены эталонные списки для девяти стран: Франции, Германии, Индии, Индонезии, Мексики, Бразилии, Турции, Италии и Таиланда. Эксперты размечали по 1000 наиболее популярных сайтов в каждой стране, проверяя по единой инструкции все требования к медийным сайтам (в том числе оригинальность и бесплатность контента и частоту его публикации) и разделяя их на две категории – медийные и прочие. В дальнейшем полученная разметка была использована для обучения и оценки классификатора «медийности» сайтов.

4. Построение классификатора и оценка вклада факторов, извлеченных из навигационных спектров

В качестве опоры для классификации использовались статистические и поведенческие факторы, построенные на основе описанных выше данных панели SimilarGroup, для чего для каждого сайта из списка сайтов-кандидатов были построены следующие наборы показателей:

- n-граммные факторы, построенные на основании имени домена и названий 50 страниц данного сайта, случайно выбранных с учётом частоты их посещения пользователями (выборка производилась с помощью алгоритма reservoir sampling [5]);
- структурные факторы, такие как средняя глубина страницы (число символов «/» в адресе страницы) и частота появления дат в адресах страниц этого сайта – с и без взвешивания по частоте посещаемости;
- частотные поведенческие характеристики посетителей данного сайта, такие как:
  - число разных пользователей панели, посетивших сайт за период;
  - среднее число дневных пользователей по всем дням в периода;
  - число посещений страницы сайта;
  - число визитов, то есть таких посещений страницы сайта, что предыдущая страница была не на этом сайте;
  - число сессий с участием данного сайта;
  - число страниц на сессию;
  - частота случаев, когда сайт оказался последним в сессии;
  - средняя длина сессии во времени;
  - медианная длина сессии во времени;

  - частота посещения главной страницы сайта;
  - среднее число просмотренных страниц за визит;

- факторы, построенные на основе навигационного спектра страниц данного сайта – в том числе статистика обобщённых маршрутов и отдельные флаги типов источника страницы, домена и сессии.

Тестировались различные алгоритмы классификации (в том числе, классификатор на основе логистической регрессии, SVM, KNN и другие), но, в конечном итоге, по результатам оценки по схеме, описанной ниже, была выбрана к использованию проприетарная реализация алгоритма Фридмана по построению деревьев решений на основе стохастическому градиентного бустинга [27].

Обучение и валидация классификатора происходили с помощью классического подхода кросс-валидации [41]. Для обеспечения независимости данных в обучающей и в тестовой выборках проводилась 7-шаговая кросс-валидация на данных 7 стран (Франции, Германии, Индии, Индонезии, Мексии, Бразилии и Турции), при которой на каждом шаге одна из этих стран выносилась в тестовую выборку, а остальные 6 использовались для обучения. Оценка качества классификатора проводилась традиционно с помощью метрики ROC AUC [32]. Две оставшиеся страны (Италия и Таиланд) использовались для финальной валидации качества работы подхода после подбора всех параметров, определяющих состав факторов и способ их нормализации. При этом для обучения использовались все 7 первоначальных стран, а для тестирования – одна выбранная.

С целью унификации и обеспечения сравнимости сайтов из списков-кандидатов разных стран все факторы, имеющие абсолютные значения, нормировались на число различных пользователей в панели, замеченных в данной стране. Затем для каждого из факторов строилось преобразование, приводящее множество его значений на тестовой выборке к распределению со средним в нуле и стандартным отклонением, равным единице. Это преобразование применялось ко всем данным, то есть и к обучающей, и к тестовой, и к валидационной выборкам.

Таблица 1

Оценка вклада факторов на основе навигационного спектра

Выборка	ROC AUC без факторов спектра	ROC AUC с факторами спектра
Франция	0,87	0,94
Германия	0,83	0,90
Индия	0,82	0,95
Индонезия	0,84	0,89
Мексика	0,85	0,93
Бразилия	0,85	0,93
Турция	0,81	0,93
Среднее	0,84	0,92
Италия	0,84	0,93
Таиланд	0,80	0,89

Специально для оценки важности вклада факторов, построенных на основе навигационного спектра было обучено два классификатора – с использованием и без использования данной группы факторов. В табл. 1 представлены результирующие значения метрики ROC AUC для отдельных стран и среднее на кросс-валидации, для случаев без и с использованием факторов на основе навигационного спектра; а сами кривые приведены на рис. 2 и рис. 3.

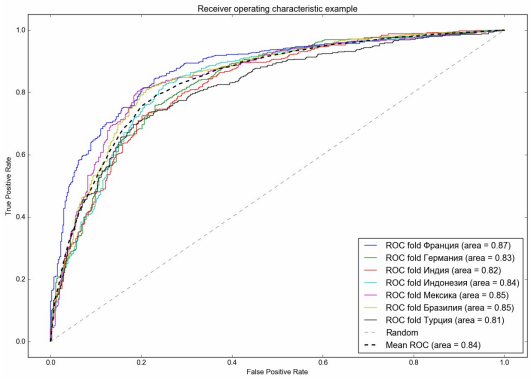

Рис. 2. ROC классификатора без использования факторов на основе навигационного спектра

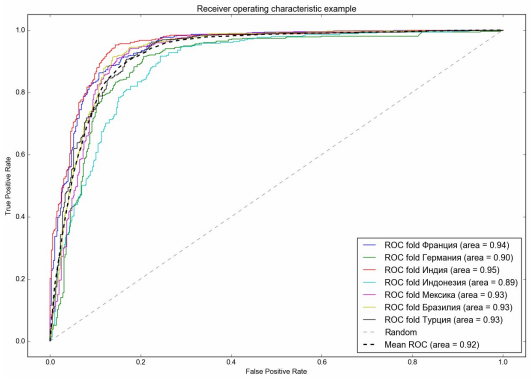

Рис. 3. ROC классификатора с использованием факторов на основе навигационного спектра

5. Заключение и выводы

В данной работе описано применение метода навигационных спектров для получения дополнительных факторов, описывающих роль сайта в сети Интернет. Использование этих факторов позволило увеличить точность страно-независимого классификатора медийности сайтов в среднем с 84% до 92% . Затем данный классификатор был использован для построения списков медийных сайтов для более чем 100 различных стран. В конечном итоге из 1076117 сайтов, вошедших в списки сайтов-кандидатов, были отобраны 37498 медийных сайтов, использованных в дальнейшем в работе рекомендательного сервиса компании Яндекс.

Литература

1. «Domain Counts & Internet Statistics». Name Intelligence. Retrieved 17 May 2009.
2. «January 2019 Web Server Survey». Netcraft. Retrieved 27 May 2019.
3. *Тихонов А.В.* Анализ структуры сети Интернет с помощью обобщенных маршрутов // УБС. 2016. 63. С. 38–70.
4. *Leskovec J., Backstrom L., Kumar R., Tomkins A.* Microscopic evolution of social networks // Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining. 2008. P. 462–470.
5. *Babcock B., Datar M., Motwani R.* Sampling from a moving window over streaming data // Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms. 2002. P. 633–634.
6. *Friedman J.H.*. Stochastic gradient boosting // In Comput. Stat. Data Anal. 2002. V. 38(4). P. 367–378.
7. *Kohavi R.* A study of cross-validation and bootstrap for accuracy estimation and model selection // Ijcai. 1995. V. 14. N 2.
8. *Hanley J.A., McNeil B.J.* The meaning and use of the area under a receiver operating characteristic (ROC) curve // Radiology. 1982. 143(1). P. 29–36.

References

1. «Domain Counts & Internet Statistics». Name Intelligence. Retrieved 17 May 2009.
2. «January 2019 Web Server Survey». Netcraft. Retrieved 27 May 2019.
3. *Tikhonov A.V.* Analysis of Web structure using generalized navigational routes. UBS. 2016. 63. P. 38–70. (in Russian).

4. *Leskovec J., Backstrom L., Kumar R., Tomkins A.* Microscopic evolution of social networks. Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining. 2008. P. 462–470.

5. *Babcock B., Datar M., Motwani R.* Sampling from a moving window over streaming data. Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms. 2002. P. 633–634.

6. *Friedman J.H.*. Stochastic gradient boosting. In Comput. Stat. Data Anal. 2002. V. 38(4). P. 367–378.

7. *Kohavi R.* A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai. 1995. V. 14. N 2.

8. *Hanley J.A., McNeil B.J.* The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982. 143(1). P. 29–36.

Поступила в редакцию 27.05.2019