Перевод статьи Barbosa L., Freire J. An adaptive crawler for locating hidden-web entry points //Proceedings of the 16th international conference on World Wide Web. – ACM, 2007. – С. 441-450.

Переведено предисловие и часть первой главы. Русских знаков без пробелов = 4321.

**Адаптивный краулинг с целью нахождения скрытых входных точек в веб.**

**Предисловие.**

В данной бумаге мы описываем новую адаптируемую стратегию краулинга, для эффективного нахождения входных точек в скрытые веб-ресурсы. Тот факт, что скрытые веб-ресурсы распределены достаточно редко, делает задачу их нахождения довольно серьёзной. Мы решаем данную задачу за счёт использования содержимого веб-страниц, чтобы сфокусировать процесс краулинга на топиках; за счёт приоретизации обещающих ссылок в рамках топика; а также за счёт следования по ссылкам, которые могут не вести к незамедлительной пользе. Мы представляем новый фреймворк в котором краулеры автоматически изучают паттерны обещающих ссылок и адаптируют своё направление по мере работы краулера, что сильно уменьшает количество необходимых настроек для запуска и оттачивания качества. Наши эксперименты на настоящих веб-страницах из репрезентативного множества доменов показывает, что онлайн-обучение приводит к существенному приросту показателя качества – адаптируемые краулеры добывают до 3-х раз больше скрытых форм, нежели карулеры с чётко зафиксированной стратегией.

**Вступление.**

Скрытый веб растёт с огромной скоростью. Количество сейчас оценивается примерно в несколько миллионов скрытых сайтов. Это сайты, чей контент обычно находится в базах данных и выявляется только при запросе, по мере заполнения и отправки форм пользователями. По мере роста количества информации, увеличился интерес к техникам, которые позволяют пользователям и приложениям извлекать эту информацию. Примерами приложений, которые пытаются упростить доступ к информации скрытого веба являются: мета-поисковики, краулеры скрытого веба, директории онлайн баз данных и система интеграции веб-информации. Так как для каждого интересующего домена существует множество скрытых ресурсов, данные которых должны быть интегрированы и поддаваться поиску, требование ключа для данных веб приложений является возможностью для нахождения этих ресурсов. Однако сделать это в большом масштабе – является сложной задачей.

Учитывая динамичную структуру веба – с новыми, постоянно появляющимися ресурсами, и постоянным удалением и изменением старых ресурсов, очень важно автоматизировать нахождение форм с возможностью поиска, которые будут служить как точки входа в базы данных скрытого веба. Однако поисковые формы имеют редкое распределение по вебу, даже для небольших доменов. Например, краулер с алгоритмом первый-лучший и ориентированный на топики, находит лишь 94 формы поиска фильмов, после просмотра 100000 страниц, имеющих отношение к фильмам. Поэтому для эффективного сбора актуальной информации о ресурсах невидимого веба, стратегия краулинга должна реализовывать широкий поиск и одновременно избегать вхождения огромных ненужных регионов веба.

Краулер должен также предоставлять высококачественные результаты. Иметь гомогенный набор форм, которые ведут к базам данных в рамках того же домена, полезно и иногда необходима для некоторых приложений. Например, эффективность техник по интеграции с формами, может сильно уменьшиться, если множество входящих форм зашумлено и содержит формы, которые не ведут в тот же домен. Однако, автоматизированный процесс краулинга инвариантно извлекает различные множества форм. Сфокусированный топик может включать в себя страницы, которые содержат поисковые формы из баз данных множества различных доменов. Например, пока краулинг работает, чтобы найти Airfare поисковые интерфейсы, краулер скорее всего добудет огромное число форм из различных доменов, таких как отели или аренда машин, так как они часто сопоставлены с Airfare поисковыми интерфейсами на сайтах путешествий. Множество добытых форм также включает в себя много форм, не поддающихся поиску, которые не представляют записи баз данных, например, логин или подписка на e-mail рассылку, запросы на цитаты и веб-email-формы.

FFC – краулер сфокусированный на формы, был нашей первой попыткой для решения проблемы автоматического поиска онлайн баз данных. FFC – комбинирует техники направляющие краулинг на топик, и с классификатором ссылок, которые идентифицируют и приоритезируют ссылки, которые более вероятно приведут к формам поиска на одном или двух шагах. Наши предварительные результаты показали, что FFC гораздо лучше справляется с задачей, если учитывать число поисковых форм, которые он находит, по сравнению с краулером, которые лишь сосредоточен на поиск по топику. Этот подход, однако имеет существенные ограничения. Во-первых, это требует ручной настройки и подгонки, включая выбор подходящих фич и создание классификатора ссылок. Дополнительно, полученные результаты совершенно не зависят на качестве множества форм, используемых на этапе обучения классификатора ссылок.