Концепция построения систем анализа и фильтрации Интернет трафика на основе методов интеллектуального анализа данных

Машечкин И. В., Петровский М. И., Глазкова В. В., Масляков В. А.
Москва, МГУ им. М. В. Ломоносова

Проблема контроля доступа к Интернет ресурсам возникает при решении следующих важных задач: блокирование доступа к нелегальной (экстремистской, антисоциальной и другой) информации; пресечение утечек конфиденциальной информации через Интернет; ограничение использования Интернет ресурсов не по назначению, в частности, блокирование доступа к развлекательным ресурсам в рабочее время.

Традиционно в существующих системах анализа и фильтрации Интернет информации применяется подход, основанный на применении *экспертных баз знаний* адресов Интернет ресурсов, где для каждого ресурса эксперт задает набор релевантных тем (категорий). Однако такие системы обладают рядом недостатков:

— не поддерживается анализ содержимого трафика в реальном времени, что необходимо, поскольку контент одного и того же ресурса может динамически изменяется;
— не поддерживается контентный анализ исходящего Интернет трафика (для предотвращения утечки конфиденциальной информации);
— необходимо использовать обновляемые извне базы знаний, что может быть недопустимо по соображениям безопасности, и в целом, качество фильтрации зависит от оперативности работы организаций, поддерживающих подготовку обновлений;
— невозможно классифицировать Интернет ресурс, данных о котором нет в текущей базе знаний.

Для преодоления данных недостатков авторами предлагается подход, основанный на *применении методов машинного обучения для анализа содержимого Интернет трафика в режиме реального времени*, что позволяет построить систему фильтрации, обладающую такими свойствами, как: *адаптируемость* — способность дообучаться и анализировать содержимое Интернет ресурсов в динамике; *автономность* — независимость от внешних баз знаний и экспертов; независимость системы от языка анализируемых ресурсов.

В процессе обучения на основе обучающей совокупности, состоящей из заранее рубрицированных HTML-документов, строится классификатор, который позволяет определять релевантные категории для произвольных ресурсов аналогичного содержания (Рис. 1). Впоследствии

класификатор может *дообучаться* на новых ресурсах. В предлагаемом подходе учитывается, что Интернет ресурсы являются *многотемными (multi-label)*, то есть каждый Интернет ресурс может быть отнесен к нескольким категориям. Для решения задачи многотемной классификации реализован подход на основе декомпозиции multi-label проблемы в набор задач бинарной классификации на основе подходов «каждый-против-остальных» и «каждый-против-каждого» [1]. Для начального обучения бинарных классификаторов используется метод SVM, а для дообучения — Kernel Perceptron.

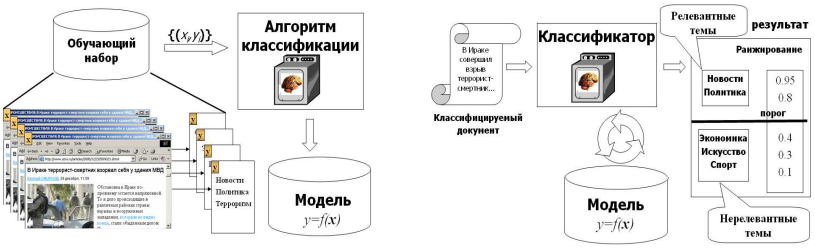

Рис. 1. Процесс обучения (слева) и процесс классификации (справа).

В процессе классификации построенный классификатор для новомого документа (или фрагмента) выдает релевантности всех тем (из предопределённого на этапе обучения набора). Далее находится *порог отсечения по релевантности* и отбрасываются наименее релевантные темы (Рис. 1). Пороговое значение, также определяется на основе методов машинного обучения, т. е. на обучающей совокупности строится модель, предсказывающая для классифицируемого документа значение порога отсечения по релевантности [2].

Для представления HTML-документов реализованы подходы на основе ключевых слов (с поддержкой стемминга) и n-грамм. В качестве меры сходства используется частотная мера сходства типа TF-IDF, а также модифицированная нами мера сходства на основе k-spectrum kernel. Кроме того, учитывается ссылочная структура HTML-документов. Для этого гиперссылки в данном документе заменяются на идентификаторы тем, релевантные документу, на который указывает ссылка (если эти темы известны, т. е. если документ, на который указывает ссылка, уже был классифицирован).

Авторами была предложена и реализованна в виде прототипа *архитектура системы интеллектуального анализа Интернет трафика* (Рис. 2).

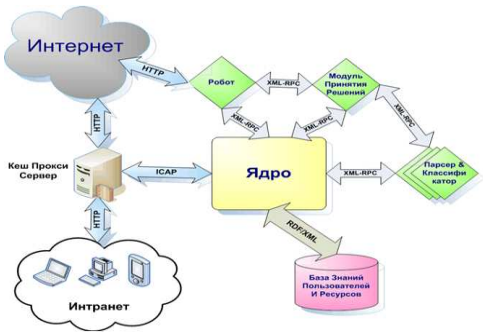

Рис. 2. Архитектура системы.

Модуль лексического разбора (парсинга) и классификации осуществляет преобразование HTML-документов во внутреннее представление и осуществляет многотемную (multi-label) классификацию. *Кеш-прокси сервер* предназначен для кеширования и фильтрации трафика локальной сети и перенаправляет запрашиваемые ресурсы ядру системы. *Ядро* координирует организацию работы модулей системы, а также сохраняет в *автоматически формируемой базе знаний* параметры ресурсов результаты анализа и классификации. *Модуль принятия решений* осуществляет разрешения или блокирования доступа к Интернет ресурсу с учетом результатов классификации и заданных политик доступа для конткретного пользователя. *Робот* используется в системе для скачивания содержимого ресурсов для последующего обучения на них и для отложенной классификации с целью пополнения базы знаний и классификации документов, ссылки на которые были выявлены ранее в процессе анализа других документов.

Работа выполнена при поддержке РФФИ, проект №06-01-00691, гранта Президента РФ МК-4264.2007.9, а также в рамках госконтракта с Федеральным агентством по науке и инновациям №02.514.11.4026.

Литература

[1] *Petrovskiy M. I.* Paired Comparisons Method for Solving Multi-label Learning Problem. — Hybrid Intelligent Systems, IEEE Press, 2006. — Pp. 42–48.

[2] *Petrovskiy M. I., Glazkova V. V.* Linear Methods for Reduction from Ranking to Multilabel Classification. — Springer-Verlag, LNAI, 2006. — V. 4304 — Pp. 1152–1156.