Е.В. Шарапова
Муромский институт (филиал) Владимирского государственного университета
Россия, Владимирская обл., г. Муром, ул. Орловская, 23
E-mail: sharapovamivlgu@gmail.com

Использование структурной сигнатуры документов для обнаружения нечетких дубликатов *

В настоящее время в Интернет активно используются нежелательные почтовые сообщения (спам). Эти сообщения содержат рекламу различных товаров и услуг, политическую рекламу, используются для фишинга и распространения вирусов. В начале 2018 года доля спама в почтовом трафике в России составила 56,72%. Другими словами – более половины почтовых писем является спамом.

Спам – анонимная незапрошенная массовая рассылка электронной почты. Миллионы копий электронных писем одновременно отправляются различным пользователям. Часто копии отличаются друг от друга приветствием (например, автоматическим указанием имени отправителя из словаря – Леонтий Людвигович, Ядвига Святославовна) или цепочкой символов (например, 1c3790b4b8ad11e8aa21e41d2d101530). Уникальность сообщений обеспечивается автоматическим путем, то есть случайные последовательности символов, приветствия и так далее [1]. Таким образом, подобные сообщения можно считать нечеткими дубликатами.

Существует множество путей обнаружения (фильтрации) спама. Проводится проверка подлинности отправителя, анализ заголовков почтовых сообщений, используются адреса-ловушки, анализируется содержания текстов писем [2, 3, 4, 5, 6]. Проверка производится как на почтовых серверах, так и на стороне клиента-получателя.

Одним из способов быстрой проверки сообщений, незначительно отличающихся по содержания, может являться использование структурной сигнатуры документов. Структурная сигнатура – это основа гипертекстовой разметки почтового сообщения с удаленной из нее содержательной частью.

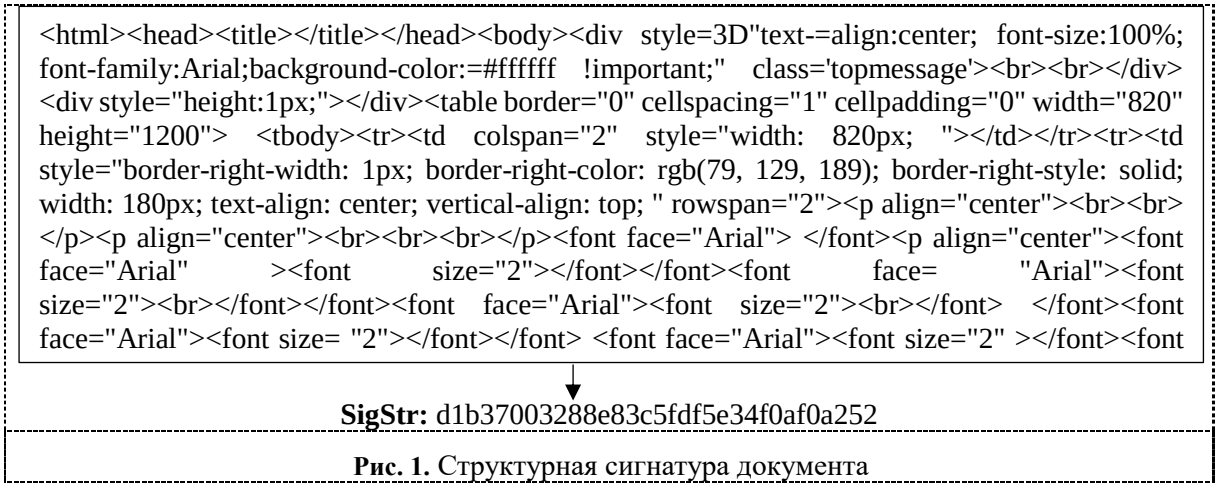

Рис. 1. Структурная сигнатура документа

Для полученной структуры документа вычисляется хэш-код. Сравнение хэш-кодов структурных сигнатур писем позволяет быстро обнаружить подобные сообщения. Структурные сигнатуры могут использоваться совместно с другими сигнатурами, например, построенными по содержанию.

Литература

1. Lee S.M., Kim D.S., Kim J.H., Park J.S. Spam detection using feature selection and parameters optimization // Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on, 2010, pp. 883-888.

2. Ляпичева Н. Г. Проблемы защиты от почтового спама: влияние облачных технологий // Вестник ЦЭМИ РАН. 2018. Выпуск 1

3. Bogers T., Van den Bosch A. Using language models for spam detection in social bookmarking // Proceedings of 2008 ECML/PKDD Discovery Challenge Workshop, 2008, pp. 1-12.

4. Subramaniam T., Jalab H.A., Taqa A.Y. Overview of textual anti-spam filtering techniques // Int. J. Phys. Sci, vol. 5, pp. 1869-1882, 2010.

5. Ahmed A. Abdo, Salim N. Ligand-based Virtual screening using Fuzzy Correlation Coefficient // International Journal of Computer Applications, vol. 19, pp. 38-43, 2011.

6. Beiranvand A. Osareh, Shadgar B. Spam Filtering By Using a Compound Method of Feature Selection // Journal of Academic and Applied Studies, vol. 2, pp. 25-31, 2012.

*** Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00692**