Лабораторная работа №1

ФИО

Чижов Илья Сергеевич

Topic

Spam; Electronic Mail; Enron

Описание предметной области

Данный топик охватывает исследования в области фильтрации спама, основываясь на особенностях работы электронной почты и изучении набора данных Enron. Спам – анонимная массовая непрошенная рассылка. Электронная почта (email) – механизм обмена сообщениями между пользователями компьютерной сети. Корпус Enron – открытая база данных и электронных сообщений сотрудников компании Enron, сгенерированная в ходе расследования дела компании, связанного с мошенничеством.

Недостаток (Gap)

Существуют различные способы борьбы с различными прогрессирующими видами спама, но отсутствуют пути решения проблемы легитимного спама.

Идея

Производить фильтрацию, основываясь на качественном анализе сообщений, используя содержание в нём требуемых пользователем параметров и ключевых слов.

Краткий текст обзора

Электронная почта – неотъемлемая часть жизни любого современного человека, приносящая много пользы [1], прошедшая долгий путь развития. В течение этого пути всегда стоял вопрос касаемо проблемы нежелательных рассылок. Они могут носить разный характер как рекламный, так и мошеннический(фишинг [2]). Задача фильтрации спама далеко не нова, но остаётся актуальной по сей день. Зарождение истории развития спама датируется 1980-ми годами. На своём пути он прошел множество этапов эволюции [3]. Первоначально внимание уделялось частной блокировке пользователей, на которых поступали жалобы. По итогу “черные списки” не принесли должного эффекта. Во время их действия количество массовых рассылок возрастало также стремительно – на 866% в период с 2000 по 2004 год [4]. Для решения проблемы, стали активно разрабатываться более активные методы борьбы со спамом[5]. Их можно разделить на локальные (Байесовская фильтрация, методы на основе формальных протокольных правил, процедурные методы, проверка подлинности отправителя) и распределенные (контрольные суммы). Наибольшего внимания заслуживает Байесовский метод фильтрации, основанный на использовании байесовского наивного классификатора [6]. Данный вид фильтрации был большим шагом в борьбе со спамом. На основе его были и метода анализа ссылок [7] реализованы системы SpamAssassin и DSPAM, обладающие самообучающимися фильтрами. Так машинное обучение крепко закрепилось в фильтрации спама [8]. В его основе лежит метод опорных векторов [9]. Пригодность данного метода была проверена благодаря анализу библиотеки Ernon [10].

Таким образом, в современное время реализовано множество технических методов борьбы со спамом, но проблема легетимного спама остаётся не решенной.Все реализованные технические способы бессильны, если пользователю необходимо получить полезные сведения среди спам-сообщений в пределах одной рассылки, которую он согласен получать. Мы предлогаем метод фильтрации, основанный на анализе содержания сообщений, с помощью использования задаваемых пользователем поисковых параметров и ключевых слов.

References

[1] M. Sumner, «The impact of electronic mail on managerial and organizational communications», *ACM SIGOIS Bull.*, т. 9, вып. 2–3, Art. вып. 2–3, 1988, doi: 10.1145/966861.45421.

[2] M.-Y. Tseng, «The pragmatic act of fishing for personal details: From choice to performance», *J. Pragmat. - J Pragmat.*, т. 42, сс. 1982–1996, 2010, doi: 10.1016/j.pragma.2009.12.004.

[3] E. Ferrara, «The history of digital spam», *Commun. ACM*, т. 62, вып. 8, Art. вып. 8, 2019, doi: 10.1145/3299768.

[4] J. Jung и E. Sit, «An empirical study of spam traffic and the use of DNS black lists», в *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, Taormina Sicily, Italy: ACM, 2004, сс. 370–375. doi: 10.1145/1028788.1028838.

[5] S. Ghiam, «A Survey on Web Spam Detection Methods: Taxonomy», *Int. J. Netw. Secur. Its Appl.*, т. 4, вып. 5, Art. вып. 5, 2012, doi: 10.5121/ijnsa.2012.4510.

[6] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, и C. D. Spyropoulos, «An evaluation of Naive Bayesian anti-spam filtering», вып. arXiv:cs/0006013. arXiv, 2000 г. Просмотрено: 14 март 2024 г. [Онлайн]. Доступно на: http://arxiv.org/abs/cs/0006013

[7] L. Becchetti, C. Castillo, D. Donato, R. Baeza-YATES, и S. Leonardi, «Link analysis for Web spam detection», *ACM Trans. Web*, т. 2, вып. 1, Art. вып. 1, 2008, doi: 10.1145/1326561.1326563.

[8] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, и O. E. Ajibuwa, «Machine learning for email spam filtering: review, approaches and open research problems», *Heliyon*, т. 5, вып. 6, Art. вып. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[9] P. Pavlidis, I. Wapinski, и W. S. Noble, «Support vector machine classification on the web», *Bioinformatics*, т. 20, вып. 4, Art. вып. 4, 2004, doi: 10.1093/bioinformatics/btg461.

[10] B. Klimt и Y. Yang, «The Enron Corpus: A New Dataset for Email Classification Research», в *Machine Learning: ECML 2004*, J.-F. Boulicaut, F. Esposito, F. Giannotti, и D. Pedreschi, Ред., Berlin, Heidelberg: Springer, 2004, сс. 217–226. doi: 10.1007/978-3-540-30115-8\_22.