АЛГОРИТМ АВТОМАТИЧЕСКОГО СОЗДАНИЯ ФИЛЬТРА ДЛЯ СОРТИРОВКИ ЭЛЕКТРОННЫХ АНГЛОЯЗЫЧНЫХ ДЕЛОВЫХ ПИСЕМ НА ОСНОВЕ КЛЮЧЕВЫХ СЛОВ И ФРАЗ

Волошина Е. Ю.
Харьковский национальный университет
«Харьковский политехнический институт»
г. Харьков, ул. Пушкинская 79/1, тел. +380994377700,
e-mail: kislaya.katia@yandex.ua

В информационный век новейшие технологии вытеснили некоторые привычные нам вещи. Так, электронная почта стала новым этапом эпистолярного жанра, сменив устаревшую переписку. Главным образом – по причине своей практичности, которая заключается в дешевизне и в большей степени – в быстроте [1].

Задача автоматического сортирования входящей корреспонденции сегодня решается в практически каждой программе-клиенте электронной почты, путем применения фильтров по ключевым фразам, которые задает пользователь.

Как сказано выше, фильтры могут создаваться вручную, однако в связи с этим возникает как минимум следующая проблема: относительно невысокий процент распознавания тематики письма и относительно высокая погрешность, так как для создания адекватного фильтра от пользователя требуются определенные навыки и знания особенностей английской деловой корреспонденции.

Наиболее популярные почтовые программы для фильтрации спама и сортировки писем: The Bat! 1.60, The Bat! 3.x, Microsoft Outlook Express 6.0, Microsoft Outlook 2002, Microsoft Outlook 2003,Mozilla Thunderbird и т.д.

Однако, в связи с этим возникает проблема создания единого фильтра, способного распознавать тематику делового письма и осуществлять сортировку почты, без участия пользователя почты.

Одной из причин актуальности создания подобного фильтра следующая: высокий процент деловых писем в электронном виде. Особенно в компаниях, на почтовый ящик которых ежедневно приходит более сотни писем. Подобная сортировка сэкономит время на обработку писем.

Для решения подобной проблемы можно использовать следующий алгоритм автоматизированной сортировки писем по тематическому признаку, основный на работе спам-фильтра.

Принцип работы алгоритма заключается в сравнении частоты появления той или иной ключевой фразы в каждой тематике письма. Если в одной тематике фраза встречается значительно чаще, чем в другом письме, при этом не менее определенного количества раз, которое можно задать, то можно считать, что оно является ключевой фразой.

Порядок выделения ключевых фраз таков:

1) создание списка всех слов во всех письмах каждого типа;

2) определение частоты появления слова в типе письма;

3) выявление и пометка кандидатов в ключевые выражения – выражений, для которых частота встречаемости выше определенного количества раз;

4) для всех таких кандидатов - добавление в список выражений;

5) повторение шагов 2-4 для всех новых появившихся выражений, пока не перестанут появляться новые кандидаты в ключевые выражения на шаге 3.

6) сравнение частоты встречаемости кандидатов в ключевые выражения вне и внутри типа письма: если их количество не больше определённого значения, то выражение считается ключевым [3].

Итак, результатом работы алгоритма является набор ключевых выражений, каждое из которых способно распознать принадлежность или непринадлежность нового входящего делового письма к какому-либо типу.

Сфера наиболее эффективного применения алгоритма – почтовые ящики, на которые ежедневно приходит множество писем.

Задача данной работы – лишь показать принцип действия, позволяющий генерировать ключевые фразы для автоматической сортировки деловых писем по тематическому признаку.

Дальнейшее развитие алгоритма может предполагать взаимодействие ручной и машинной обработки входящей корреспонденции и одновременное автоматическое накопление/обобщение опыта ручной сортировки. Если система не в состоянии автоматически определить по критериям категорию сообщения с заданной степенью достоверности, она передает его на обработку оператору. Действия оператора являются основой для дальнейшей автоматической работы системы.

Такой фильтр можно интегрировать в уже существующие программы-клиенты, а так же включить как один из фильтров в существующие электронные почты.

Список литературы:

1. *Гадасин В. А., Конявский В. А* "От документа - к электронному документу. Системные основы" - М.: "РФК-Имидж Лаб", 2001 -190 с.

2. *Кучернюк П.В., канд. техн. наук, Сушко К.В.* Электроника и связь 1’2009 Методы обработки и фильтрации нежелательных сообщений электронной почты.

3. *Свердлов М.И., Андриенко П.В.* Алгоритм автоматического создания фильтров для сортировки входящей почты. Электронный многопредметный научный журнал «ИССЛЕДОВАНО В РОССИИ» 161/021114 .