УДК 004.942

ВИЗНАЧЕННЯ ВИМОГ ДО СИСТЕМИ АНАЛІЗУ ЗМІСТУ ЛИСТІВ ЕЛЕКТРОННОЇ ПОШТИ ЗА ОБРАНИМ НАПРЯМКОМ

Нестеренко О. В., Фаловський О. О.
(oleksandr_nesterenko@ieu.edu.ua, oleksandr_falovskyi@ieu.edu.ua)
Міжнародний європейський університет (Україна)

У статті розглянуто актуальне питання автоматизації процесу обробки листів електронної пошти. Показана доцільність розробка методологічних засад для побудови системи автоматизованого аналізу змісту листів електронної пошти та їх подальшого оцінювання для побудови тематичних добірок із найбільш прийнятних листів. Розглянуто варіант реалізації функціонально-структурної схеми такої системи.

Надзвичайно суттєвою частиною електронного документообігу у сучасному світі є листи електронної пошти. На відміну від месенджерів, соціальних мереж та сайтів електронна пошта є засобом цілеспрямованого надання інформації фізичним особам та організаціям. Обсяг листів, що надходять (сотні, або навіть тисячі на добу) може перевищити можливості людини обробити їх протягом прийнятного часу.

Отже зменшення навантаження з перегляду електронних листів шляхом обрання із важливих лише найсуттєвіших є надзвичайно бажаним з огляду економії часу власника поштової скриньки. Одночасно слід забезпечити «відсікання» спаму та матеріалів навколо-рекламного характеру. Одним із найпоширеніших шляхів розв'язання такої задачі є використання спам-фільтрів [1, 2] у найрізноманітніших варіаціях, які так чи інакше ґрунтуються на використанні методів Байєсової фільтрації. В основі такої фільтрації лежить достатньо просте припущення: «лист із спамом містить «заборонені» фільтром слова» – тобто наявність певних слів збільшує вірогідність того, що лист належить до спаму. Однак використання спам-фільтрів може призвести до «відсікання» значної частини листів від нових дописувачів, що може спричинити втрату цінної інформації.

Виходячи із ідеї визначання «неістотних» листів шляхом аналізу наявності у листі «заборонених» слів вбачається корисним використати аналогічний підхід із використанням «важливих» слів для побудови множини «корисних» листів. Тоді аналіз тексту листа може виконуватися одночасно у двох напрямках – на вміст слів із множин «заборонених» та «важливих» слів.

Множину «заборонених» слів можна обрати за загальноприйнятими методами. А от формування множини «важливих» слів слід віддати на розсуд власника поштової скриньки. До того ж користувач може визначити за певною шкалою *міру важливості* для нього тих або інших «важливих» слів, що дозволить побудувати *оцінку* кожного листа. Прикладом може слугувати надходження на поштову скиньку організації листів із резюме для участі у конкурсі на відповідну посаду. Тут множина «важливих» слів може містити такі слова, як «стаж», «досвід», «кваліфікація-…» та пов'язані з ними кількісні характеристики.

Процес аналізу тексту листа, що надійшов до поштової скриньки, за множинами обраних «заборонених» та «важливих» слів є нескладним і має оцінку $O = (N + n*M)$ [3], де N – довжина тексту листа у символах, M – довжина найдовшого слова-шаблону пошуку, n – кількість таких слів у відповідній множині. За умови, що для множини «важливих» слів $N < 1000$ (2-3 сторінки тексту), $n < 30$, а найбільше $M < 20$, стає зрозумілим, що такий пошук триватите лише частки секунди.

Підсумком такої обробки буде множина «корисних» листів, яка до того ж міститиме для кожного з них розраховану числову оцінку як добуток вектору мір важливості слів на вектор наявності у листі «важливих» слів та «анотацію» із усіма наявними «важливими» словами. У подальшому користувач повинен мати змогу відкоригувати міри важливості у відповідності до практичних результатів обробки відібраних «корисних» листів.

Таким чином на думку авторів основними вимогами до системи аналізу повинні бути :

- наявність зовнішніх критеріїв релевантності, організованої у вигляді матриці груп критеріїв: {«критерій-ключове слово»/ «значення/діапазон значень, пов'язаних із ключовим словом» / «рівень важливості критерія для загальної оцінки»} для співставлення із змістом листа;
- здатність аналізувати структуру листа – наприклад, наявність вкладень, гіперпосилань, виконуваних програм і т. ін.;
- здатність аналізувати зміст листа (та/або зміст безпечних вкладень) на предмет наявності ключових слів та (за наявності) значень, що їм відповідають;
- наявність бази даних для збереження «анотацій» отриманих листів із побудованими оцінками відповідності по групах критеріїв;
- можливість виконувати ранжування та динамічну зміну кількості, складу та характеристик груп критеріїв перегляду (листи, зміст яких не відповідає одному набору критеріїв – стовпцю матриці, можуть виявитися важливими для іншого набору);
- здатність використовувати різні мови та системи кодування літер для опису ключових слів у критеріях;
- можливість розрахунку «оцінки ваги» листа за обраним переліком критеріїв для обрання «найвагоміших»;
- можливість співставлення у подальшому збережених у базі даних «анотацій» листів із оцінками користувачів для «донавчання» механізму перегляду і регулювання відповідних коефіцієнтів для ключових слів;
- можливість одночасного (багатопотокового) перегляду масиву листів та побудови оцінок листів за матрицею критеріїв.

Реалізація набору визначених вимог може бути відображеною у вигляді узагальненої BPMN-діаграми (рис. 1). Необхідно зауважити, що наявність зворотного зв'язку з користувачем забезпечуватиме додаткову гнучкість та адаптивність системи.

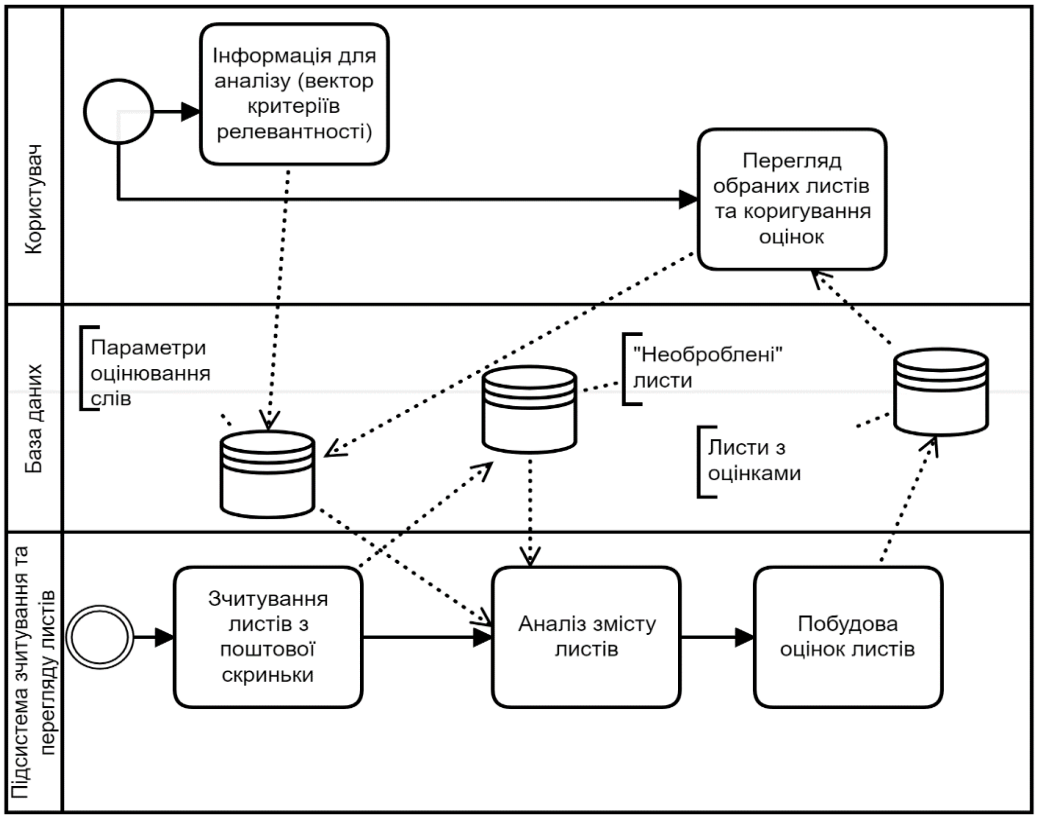

Рис. 1. Узагальнена модель бізнес-процесу для системи аналізу листів електронної пошти

Побудована у відповідності до наведених умов система буде здатна у автоматичному режимі зчитувати інформацію із поштової скриньки, уникати спаму, обраховувати оцінку за кожною групою критеріїв відповідності та формувати підсумковий звіт – перелік листів, що найкраще відповідають висунутим вимогам. Попереднє ранжування та нормування груп критеріїв (за визначеними користувачем уподобаннями) забезпечить ефективність оцінок. У підсумку листи з найвищими оцінками матимуть найбільшу питому вагу і, відповідно, складатимуть групу, що потраплятиме до підсумкового звіту.

Список використаної літератури

[1] S. Krishnamurthy, “SPAM: A Consumer Perspective”, In: *Spotts, H. (eds) Revolution in Marketing: Market Driving Changes. Developments in Marketing Science: Proceedings of the Academy of Marketing Science*. Springer, Cham. 2015. doi: 10.1007/978-3-319-11761-4_47

[2] J. Demsar, “Statistical Comparisons of Classifiers over Multiple Data Sets”, *Journal of Machine Learning Research*. 2006, 7, pp. 1–30.

[3] R.S. Boyer, J.S. Moore, “A fast string searching algorithm”, *Communication of the ACM*. 1977, 20, pp. 762-772.

ОСОБЛИВОСТІ ВИКОРИСТАННЯ ПАКЕТУ STATISTICA ТА MS EXCEL ДЛЯ ОБРОБКИ СТАТИСТИЧНИХ ДАНИХ

Янковий А., Радзіховська Л.
Вінницький торговельно-економічний інститут
Державного торговельно-економічного університету

Аналіз даних – це обов'язкова частина процесів дослідження економічних систем. Нині на ринку існує велика кількість різноманітного ПЗ для статистичної обробки даних. Розглянемо найбільш вживані: EXCEL та пакет STATISTICA.

Виходячи з поставленого круга завдань науково-дослідницької діяльності, користувачеві кожного разу необхідно обирати оптимальне і відповідне для нього ПЗ − статистичний пакет. Як правило, оптимальним є варіант, що комбінує в собі високий рівень продуктивності ПЗ, потрібні функціональні можливості і помірну ціну. При виборі важливо звернути увагу на наступні характеристики: відповідність комп’ютерного устаткування користувача системним вимогам ПЗ; відповідність можливостей ПЗ до параметрів поставлених завдань; об’єм даних для статистичного аналізу; кваліфікація (рівень знань) користувача в області статистики. Статистичний пакет повинен відповідати певним вимогам: модульність; можливість асистування при виборі способу обробки даних; використання простої проблемно-орієнтованої мови для формулювання завдання користувача; автоматична організація процесу обробки даних та зв'язків з модулями пакета; ведення банку даних користувача і складання звіту про результати зробленого аналізу; діалоговий режим роботи користувача з пакетом; сумісність з іншим програмами [1] .

Охарактеризуємо найпопулярніші та функціонально повні програмні продукти з наявними засобами статистичного аналізу даних.

MS Excel − найбільш поширений додаток з пакету офісних програм MS Office. MS Excel − це електронна таблиця з досить потужними математичними можливостями, в якій деякі статистичні функції є просто додатковими вбудованими формулами. MS Excel добре підходить для накопичення даних, проміжного перетворення, попередніх статистичних обчислень, для побудови деяких видів діаграм. Проте остаточний статистичний аналіз необхідно робити в програмах, які спеціально створені для цих цілей. Існують макроси-