Откриване на спам имейл съобщения чрез използване на методи за извличане на знания

Снежана Сълова

Finding a Spam Email Messages Using Data Mining Methods

Snezhana Sulova

Abstract

There are many software solutions that have been developed based on the use of various software technologies for identification of e-mail spam messages. This article presents how we may successfully use data mining methods for identifying spam messages. The proposed approach is based on Supervised Machine Learning methods - Support Vector Machines (SVM) and Naive Bayes (NB). Exemplary model for email messages extraction and classification is implemented in RapidMiner.

Keywords: Data Mining, Web Mining, classification, Support Vector Machines, Naive Bayes, Internet, e-mail, spam, RapidMiner

Въведение

Една от най-често използваните Интернет услуги е електронната поща. Този вид комуникации са предпочитани както за лични така и за бизнес цели, тъй като имат редица предимства пред конвенционалните средства. При тях писмата се разпращат бързо, обмяната на съобщенията се извършва без да е необходимо подателят и получателят да бъдат едновременно онлайн, няма времеви и пространствени ограничения, съобщенията се отличават с точност, яснота и могат да се препращат и адресират до много получатели.

Тези преимущества на комуникациите посредством имейл, обаче са предпоставка и за разпространеното на нежелани съобщения, на т.н. спам поща. Това най-често са реклами, търговски оферти, писма с агресивно съдържание и др., които не са желани от получателите. За откриването на спам съобщения, изпратени по електронна поща има множество софтуерни решения, които са разработени на базата на приложението на различни софтуерни технологии.

Цел на настоящата статия е да се представи как могат успешно да се използват методите за извличане на знания (Data Mining – DM) за идентифициране на нежелани съобщения и да се покаже апробация на предложения подход със софтуерния продукт RapidMiner.

Извличане на знания от неструктурирани данни

Един от начините за филтриране на електронната поща за спам съдържание е чрез извършване на анализ на нейното съдържание. По своето същество съобщенията, разпращани по имейл са под формата на обикновен текст или в html формат. За да се преобразуват тези съобщения във вид удобен за анализ е необходимо приложение на технологиите за обработка на естествен език (Natural Language Processing, NLP), които могат да се разглеждат като синтез между изкуствения интелект и компютърната лингвистика. Концепцията NLP е насочена към "пълното разбиране на текста, проверка на синтактичната и семантична валидност на лингвистичния вход, като се използват знанията за реалния свят за разбиране на целите и убежденията, а също и речевите действия, разговорите и дискурс за структурата"[1].

С различни аспекти на обработката на неструктурирани данни и по-конкретно на текст се занимават множество изследователи. Най-общо откриването на знания в

[1] Kumar, E. Natural Language Processing. New Delhi: I. K. International Pvt, 2011. p. 4.

неструктурирани данни в литературата е известно като извличане на текст (Text Mining, TM)[2]. Този процес се осъществява, чрез приложение на технологиите за извличането на знания върху неструктурирани текстови данни.

Извличането на знания (на английски терминът е Data Mining - DM) може да се определи като процес на изследване и откриване на скрити данни, знания, които преди това не са били известни, а са полезни за бизнеса чрез използване на техники и алгоритми от областта на изкуствения интелект[3].

Задачите, които се решават чрез DM методи е прието да се разделят на две основни категории: предсказващи (predictive), при които се получава модел, който може да се използва за извършване например на класификация и описателни (descriptive), при които най-важното е да се даде нагледно описание на скритите зависимости, връзки в анализираните данни[4] (вж. фиг.1).

Фигура 1. Методи за извличане на знания

Когато данните, от които се извличат знания са Интернет ресурси процесът се нарича Web Mining (WM). Един от първите изследователи на този процес е Ециони, който го определя като използване на DM техниките за изследване и извличане на информация от Интернет документи и услуги[5]. По-късно това определение се допълва като към него анализаторите вече включват и техниките за изследване и анализ на данните за използваемостта на уеб ресурсите и дефинират WM като "разкриване на закономерности, зависимости, шаблони в съдържанието, структурата и начина на използването на уеб документите, които показват как да се приложи извличането на знания от данни (Data Mining – DM) върху уеб базирани данни"[6]. В зависимост от данните, които ще се обработват има три основни вида извличане на знания от уеб ресурси[7]:

[2] Fayyad, M., Piatetsky-Shapiro and Smyth, P. From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3), 1996 p. 37-54, https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131, 23.09.2016.
Anne, A., Poteet, S. Natural Language Processing and Text Mining, Springer, 2007, p. 1.
[3] Барсегян, А. и др. Анализ даных и процессов. 3-е издание. Санкт-Петербург, 2009, с.68.
[4] Kantardzic, M. Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, New Jersey, 2011, p. 2.
[5] Etzioni, O. The World Wide Web: quagmire or gold mine?, Communications of the ACM, 11/1996, p. 65-68.
[6] Markov, Z. and D. Larosed. Data Mining the Web Uncovering Patterns in Web Content, Structure, and Usage. John Wiley & Sons, New Jersey, 2007.
[7] Cooley, R., Mobasher, B. and Srivastava, J. Web Mining: Information and Pattern Discovery on the World Wide Web, Proceedings, Ninth IEEE International Conference, https://www.researchgate.net/profile/Jaideep_Srivastava2/publication/3720481_Web_mining_Information_and_pattern_discovery_on_the_World_Wide_Web/links/0deec521b680228c6a000000.pdf, 10.9.2016.

- извличане на полезни знания от съдържанието на уеб документите (Web Content Mining, WCM);
- извличане на полезни знания на базата на структурата от входящи и изходящи връзки (топологията) в уеб (Web Structured Mining WSM);
- извличане на полезни знания на базата данните за използване на Интернет ресурсите (Web Usage Mining, WUM).

По проблемите на обработката на текстово съдържание от Интернет ресурси работят много изследователи[8]. Повечето проучвания са диференцирани в зависимост от конкретните изследователски задачи и типа на източниците, които се използват. Има и разработки, свързани с филтриране на електронната поща с цел откриване на спам съобщения, като те се основат най-вече на методите за класификация на текстови документи[9].

Подход за извличане и анализ имейл съобщения

Необходимо е да се отбележи, че за извличането на полезни знания от неструктурирани данни в литературата не съществува единен, универсален алгоритъм. Базирайки се на направените проучвания и експерименти установихме, че за текстови анализ успешно се прилагат методите за машинно самообучение с учител, където има примерен вход и параметри за желания изход, дадени от „учител“, такива са напр. методите за класификация.

За идентифициране на спам имейл съобщения успешно се прилагат линейният класификатор Support Vector Machines (SVM), алгоритъмът базиран на изчисляване на условната вероятност за принадлежността на обектите към даден клас Naive Bayes (NB), класификациите чрез дървета на решения и невронни мрежи[10].

Подходът, който предлагаме за идентифицирането на спам имейл съобщения се основава на класифициране на съобщенията на базата на анализ на тяхното съдържание. Използват се методите на машинното самообучение с учител и SVM и NB класификации, защото според направените проучвания те дават добри резултати при обработката на текст.

Известно е, че всяко електронно писмо се състои от две основни части:

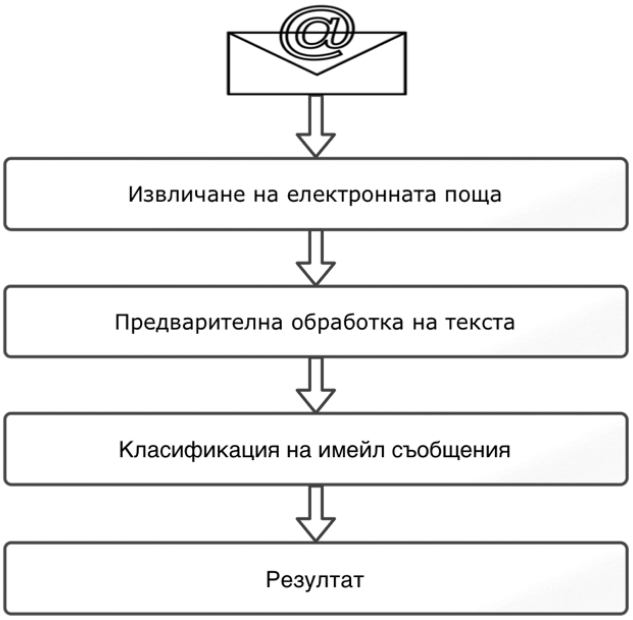
- хедър или заглавна част, която съдържа информация за изпращача, получателя, предмет на съобщението, дата и др.;
- съдържание на съобщението, което на практика представлява самото съобщение и е в текстов или html формат и може да съдържа изображения, хиперлинкове.

Тъй като както вече беше казано обработката на тест се базира на DM технологиите предлагаме процесът на извличане на знания от съдържанието на електроните писма да се осъществи в етапите, дадени на фиг. 2.

[8] Kosala, R., Blockeel, H. Web Mining Research: a survey. ACM SIGKDD Explorations Newsletter. 2(1). p. 1-15, http://www.kdd.org/exploration_files/kosala.pdf, 11.6.2016.
Navadiya, D., Patel, R. Web Content Mining Techniques-A Comprehensive Survey. International Journal of Engineering Research & Technology, 1(10). p. 1-6, http://www.ijert.org/view-pdf/1950/web-content-mining-techniques-a-comprehensive-survey, 11.9.2016.
Markov, Z. and D. Larosed. Data Mining the Web Uncovering Patterns in Web Content, Structure, and Usage. John Wiley & Sons, New Jersey, 2007.

[9] Rathi, M. Spam Mail Detection through Data Mining – A Comparative Performance Analysis, I.J. Modern Education and Computer Science, 12/2013, p. 31-39.
Shapoorjee1, J. et. al. Comparative Study of Data Mining Techniques for Classifying Spam Email, International Journal of Emerging Technology and Advanced Engineering, Vol. 5, Volume 5, Issue 10, 2015, p. 132-135.

[10] Tretyakov, K. Machine Learning Techniques in Spam Filtering, University of Tartu Data Mining Problem-oriented Seminar, 2004, page. 60-79.



Фигура 2. Етапи на идентифицирането на спам имейл съобщения

Процесът включва първоначално извличане на електронната поща от съответния пощенски сървър. Имейл съобщенията могат да се изтеглят директно от сървъра или от клиента чрез програмата, която той използва за достъп до електронната си поща. След това се осъществява предварителната обработка на текста от имейл съобщенията. Това е процес на подготовка на текста за същинската обработка. Известно е, че един от основните проблеми при анализа на текст е наличието на голямо количество думи, които не са носители на полезна информация и не е целесъобразно всички те да се обработват затова на този етап се прилагат техники за премахване на неинформативните текстовете. Основните такива са:

- токънизация - разделяне на пълния текст на списък от отделни думи;
- премахване на излишните думи (stop words) - това са спомагателни думи, които носят малко информация за съдържанието на текста, като например: “за”, “по-горе”, “през”, “след”, “така”, “по-късно”, “отново”, “срещу” и др.;
- стеминг (stemming) – извличане на корена на думата и преобразуването й в нормална форма;
- преобразуване на символите - да бъдат изписани само с малки букви;
- определяне на частите на речта, дали думите са прилагателни, съществителни, наречия, глаголи.

Следва построяването на модел за осъществяване на класификацията и прилагане на създадения модел върху реалните данни, за да се определи към кой клас се отнася изследвания обект. В нашия случай чрез модела се разпределят писмата към категорията към която принадлежат - спам или не спам съобщения.

Процесът завършва с визуализация на получения резултат.

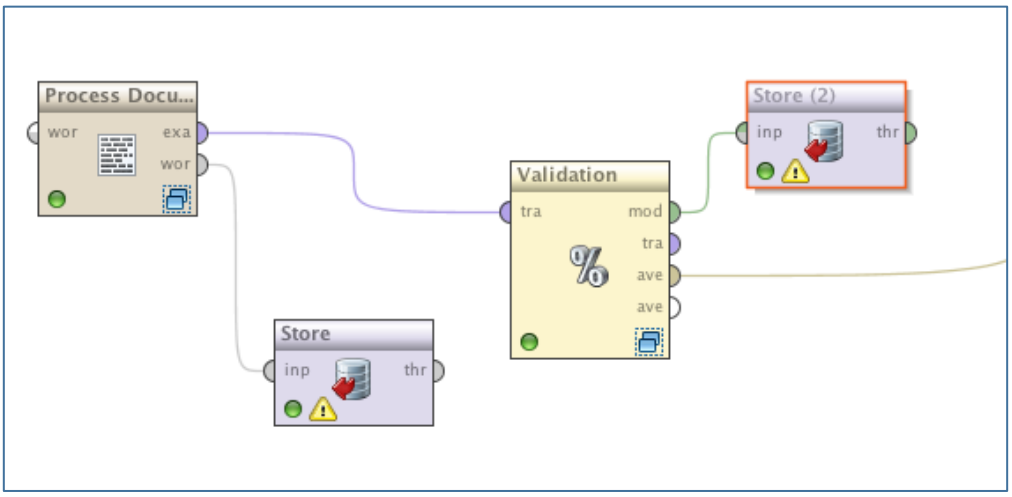
Реализация на предложения подход чрез RapidMiner

Предложеният модел ще реализираме чрез софтуера за извличане знания RapidMiner, които според представителното проучване на Garthner е сред лидерите в софтуерните решения от този клас[11]. Софтуерът е Java базиран и е разработен в университета в Дортмунд.

[11] Kart, L., Herschel, G., Linden, A., Hare, J., Magic Quadrant for Advanced Analytics Platforms, https://www.gartner.com/doc/reprints?id=1-2YEIILW&ct=160210&st=sb, 27.09.2016.

Разполага с интерактивен графичен потребителски интерфейс и средства за предварителна обработка, класификация, регресия, клъстеризация, асоциация и за графично представяне на моделите и резултатите.

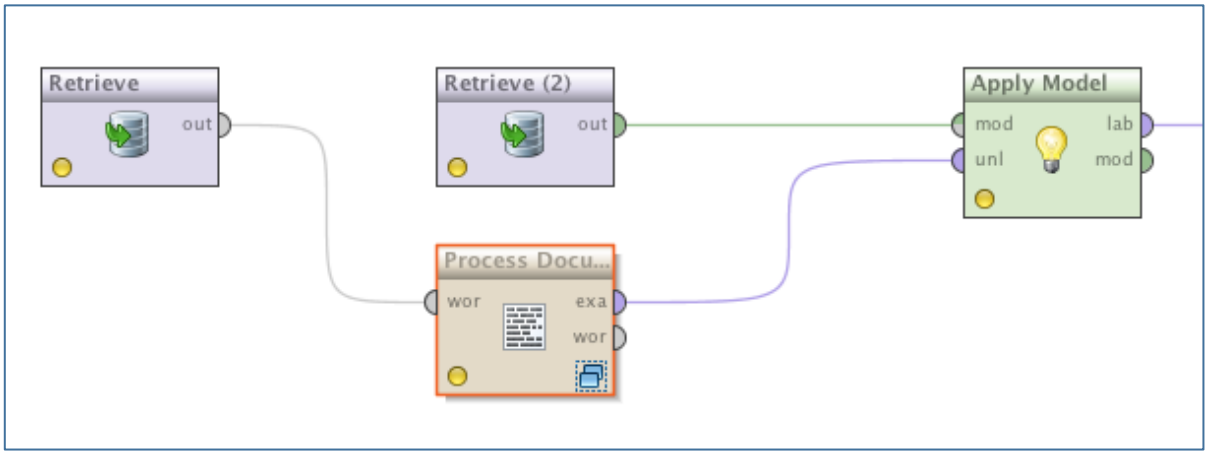
За да тестваме предложения подход сме подготвили две групи проверени предварително тренировъчни данни. Едната група се състои от текстови файлове със спам имейл съобщения, а другата с не спам съобщения. Моделът, който сме построили за да се обучи класификаторът е даден на фиг. 3.



Фигура 3. Модел за класификация

При него чрез оператора Process Documents from Files се извършва предварителната обработка на текста и се генерират вектори от думи от двете групи имейл съобщения спам и не спам. Резултатът се обработва с алгоритмите SVM или NB, за да може да се обучи моделът и след това да се използва за предвиждане на това новите примери към коя категория да се причислят. Това е етап на обучение, при който класификационният алгоритъм се прилага върху примерите с цел извличане на класификационни правила.

За осъществяване на самата обработка се създава нов модел (фиг. 4), при който се осъществява автоматично извличане на електронните пощи (Process Documents from Mail Store) директно от от мейл сървър чрез използване на един от протоколите IMAP или POP3 и след това се прилага модела за класификация за разделяне на съобщенията на две групи.



Фигура 4. Приложение на модела за класификация върху имейл съобщения

Чрез създадения модел успешно са тествани няколко нови имейл съобщения и резултатът показва, че те правилно се идентифицират дали са спам или не (фиг. 5).

Row No.	label	prediction(label)	confidence(nospam)	confidence(spam)
1	notclass	spam	0.465	0.535
2	notclass	spam	0.493	0.507
3	notclass	spam	0.464	0.536
4	notclass	spam	0.420	0.580
5	notclass	nospam	0.506	0.494
6	notclass	nospam	0.661	0.339
7	notclass	nospam	0.517	0.483
8	notclass	spam	0.450	0.550
9	notclass	nospam	0.574	0.426
10	notclass	spam	0.388	0.612

Фигура 5. Резултат от идентифицирането на спам имейл съобщения

Реализацията на предложения подход показва, че методите за класификация SVM и NB може успешно да се използват за изследване на реални данни. В нашия случай това са имейл съобщения, класифицирани в две основни групи, и чрез модел, който описва закономерностите в тези изследваните данни да се определи към кой клас може да се причислят новите некласифицирани съобщения.

Заключение

Идентифицирането на спам в електронната поща е обект на изследвания през последните години. Освен по-разпространените методи за борба със спама, които се базират на изследване на хедъра на имейл съобщението и блокиране на адресите, от които е известно, че се разпраща спам или анализ на предмета на съобщението, може да се твърди, че вече все по-успешно се използват и методите за анализ на съдържанието на писмото. В статията се предлага и апробира подход за идентифициране на спам имейл съобщения, в който се използват методите на извличане на знания от неструктурирани данни и се анализира текста в писмата разпращани по електронна поща. Класифицирането на пощата на спам и не спам е извършено с два алгоритъма - SVM и NB, като получените резултати са сходни, затова считаме, че и двата са приложими за анализ и откриване на спам имейл съобщения.

Използвана литература

1. Барсегян, А. и др. Анализ даных и процессов. 3-е издание. Санкт-Петербург, 2009, с.68.
2. Anne, A., Poteet, S. Natural Language Processing and Text Mining, Springer, 2007.
3. Cooley, R., Mobasher, B. and Srivastava, J. Web Mining: Information and Pattern Discovery on the World Wide Web, Proceedings, Ninth IEEE International Conference, https://www.researchgate.net/profile/Jaideep_Srivastava2/publication/3720481_Web_mining_Information_and_pattern_discovery_on_the_World_Wide_Web/links/0deec521b680228c6a000000.pdf, 10.9.2016.
4. Etzioni, O. The World Wide Web: quagmire or gold mine?, Communications of the ACM, 11/1996, p. 65-68.
5. Fayyad, M., Piatetsky-Shapiro and Smyth, P. From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3), 1996 p. 37-54, https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131, 23.09.2016.

6. Kantardzic, M. Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, New Jersey, 2011.

7. Kart, L., Herschel, G., Linden, A., Hare, J., Magic Quadrant for Advanced Analytics Platforms, https://www.gartner.com/doc/reprints?id=1-2YEIILW&ct=160210&st=sb, 27.09.2016.

8. Kosala, R., Blockeel, H. Web Mining Research: a survey. ACM SIGKDD Explorations Newsletter. 2(1). p. 1-15, http://www.kdd.org/exploration_files/kosala.pdf, 11.6.2016.

9. Kumar, E. Natural Language Processing. New Delhi: I. K. International Pvt, 2011.

10. Markov, Z. and D. Larosed. Data Mining the Web Uncovering Patterns in Web Content, Structure, and Usage. John Wiley & Sons, New Jersey, 2007.

11. Navadiya, D., Patel, R. Web Content Mining Techniques-A Comprehensive Survey. International Journal of Engineering Research & Technology, 1(10). p. 1-6, http://www.ijert.org/view-pdf/1950/web-content-mining-techniques-a-comprehensive-survey, 11.9.2016.

12. Rathi, M. Spam Mail Detection through Data Mining – A Comparative Performance Analysis, I.J. Modern Education and Computer Science, 12/2013, p. 31-39.

13. Shapoorjee1, J. et. al. Comparative Study of Data Mining Techniques for Classifying Spam Email, International Journal of Emerging Technology and Advanced Engineering, Vol. 5, Volume 5, Issue 10, 2015, p. 132-135.

14. Tretyakov, K. Machine Learning Techniques in Spam Filtering, University of Tartu Data Mining Problem-oriented Seminar, 2004, p. 60-79.

За контакти
доц. д-р Снежана Сълова
Икономически университет - Варна
ssulova@ue-varna.bg