INTELLIGENT MAILBOX

Antonín Pohlídal

Master Degree Programme (3), FIT BUT

E-mail: xpohli01@stud.fit.vutbr.cz

Supervised by: Petr Chmelař

E-mail: chmelarp@fit.vutbr.cz

Abstract: This paper describes the main principle of email classification based on text recognition and text classification. Naive Bayes and Support Vector Machines, two machine learning algorithms, are used. It is also shown how user can interacts with whole process of email classification and how it is important. At the conclusion, there are also mentioned problems that may occur during email classification.

Keywords: email classification, text mining, naive bayes, support vector machines, IMAP, apache james server, rapid miner, postgresql

1. ÚVOD

Emailová komunikace je v dnešní době nedílnou součástí našeho života. Denně se lidé potýkají s problémem třídění příchozí pošty. Zpočátku se zdálo jako vhodné řešení použití filtrovacích pravidel. Tento způsob je ale náročný na správu, nedokáže třídit neznámý typ příchozí pošty a s rostoucím počtem filtrovacích pravidel roste i chybovost při třídění zpráv. Z těchto důvodů mě proto zaujala možnost vytvoření emailové schránky, která by tyto problémy řešila.

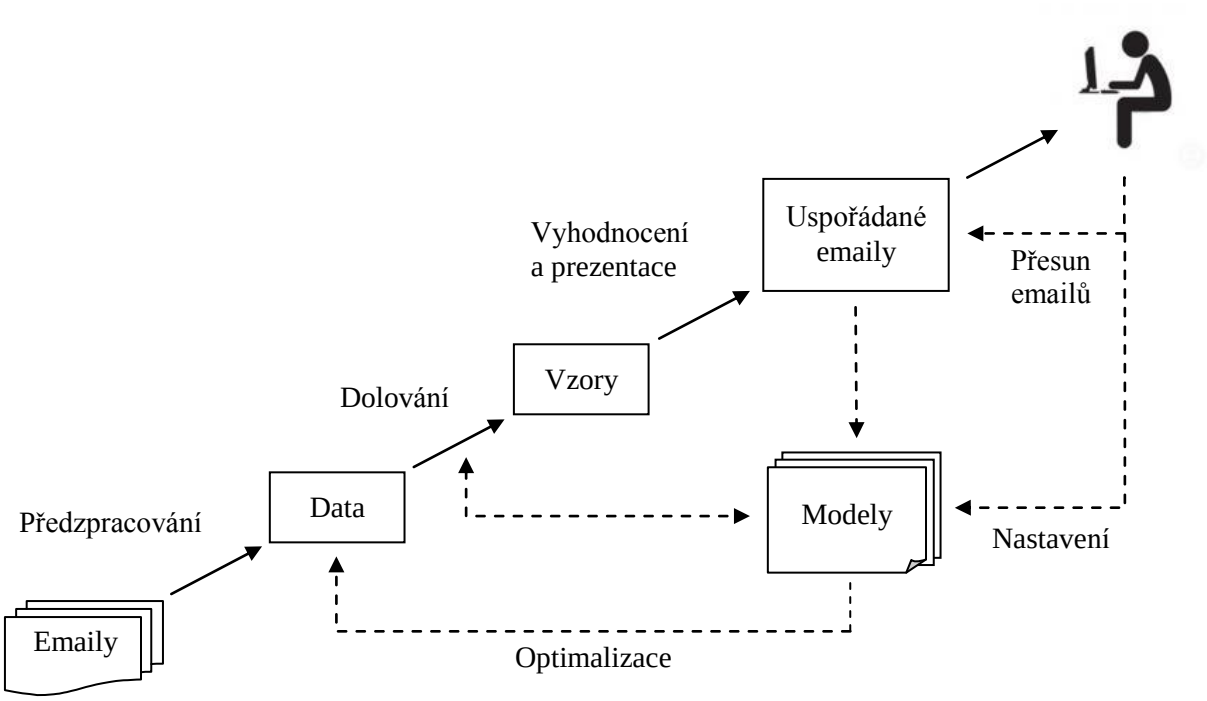
Cílem mé práce je proto vytvořit otevřený systém umožňující třídit příchozí poštu na základě rozpoznávání a klasifikace textu. Důležitou součástí je učení třídících návyků jednotlivých uživatelů a vhodná kombinace těchto modelů za účelem dosažení co nejlepších výsledků s malým procentem chybovosti. Použity jsou dva algoritmy strojového učení, jedná se o Naive Bayes a Support Vector Machines. Hlavním prvkem tohoto systému je emailový server Apache James Server. Jeho výhodou je, že je vyvíjen komunitou jako open-source. Pro účely získávání znalostí z textu je použit program Rapid Miner a pro perzistenci dat databáze PostgreSQL.

2. KLASIFIKACE EMAILŮ

Při klasifikaci emailů vycházím z principů klasifikace textu, jejímž účelem je podle [1] rozhodnout, do jaké třídy se má daný text zařadit. Může se jednat o jednoduché dělení na dvě třídy nebo o složitější s více různorodými třídami. Aby bylo možné klasifikovat určitý text, je potřeba provést získávání znalostí z textu. Jedná se o proces, který je ve své podstatě odvozen od získávání znalostí z databází. Podrobný popis získávání znalostí z databází je uveden v [1].

Podle [2] můžeme získávání znalostí z textu obecně definovat jako proces založený na znalostech, ve kterém uživatel pracuje s analytickými nástroji nad kolekcemi dokumentů. Podobně jako u získávání znalostí z databází se získávají užitečné informace z datových zdrojů na základě identifikace a zkoumání zajímavých vzorů. V tomto případě jsou za datový zdroj považovány kolekce dokumentů a zajímavé vzory se nehledají ve formalizovaných záznamech v databázi, ale v nestrukturovaných textových datech uložených v dokumentech, které tvoří kolekce.

Na níže uvedeném obrázku 2.1 je zobrazen celý proces získávání znalostí, jehož úkolem je roztřídit vstupní kolekci emailů do příslušných kategorií. Do tohoto procesu vstupují na začátku kolekce emailů uživatele.



Obrázek 2.1 Proces získávání znalostí z emailů

V počáteční fázi se provede předzpracování hlavičky a těla zprávy. Data se z původního zdroje převedou do vhodného formátu pro dolování. Ve zprávě jsou informace uloženy v přesně specifikovaném formátu podle standardu RFC 5322 definovaném v [4]. Hlavička zprávy se skládá z několika polí, která nejsou všechna povinná a některá mohou být nestrukturovaná a mohou se opakovat. Každé pole se sestává z jednoho řádku ASCII textu. Na začátku je uveden standardizovaný název pole ukončený dvojtečkou a za ní následuje hodnota. Tělo zprávy obsahuje jednotlivé řádky ASCII textu s maximální délkou každého řádku 1000 znaků.

Po předchozí fázi jsou data připravena pro dolování. Následuje hlavní fáze, ve které se provádí samotné dolování. To zahrnuje inkrementální objevování zajímavých vzorů, jejich spojování a případné provádění asociace. Jsou zde použity dva algoritmy Naive Bayes (NB) a Support Vector Machines (SVM).

NB je algoritmus, který je podle [3] založen na teorii pravděpodobnosti a je odvozen z Bayesovy rozhodovací teorie. Pro každý email hledá pravděpodobnost, s jakou patří do daných kategorií. Na závěr jsou vybrány kategorie přesahující určitou mez. Podrobněji v [3] na str. 39.

SVM je lineární klasifikační algoritmus. Podle [3] hledá hyperrovinu, která dělí emaily do dvou kategorií. To se provede výběrem dvou rovnoběžných hyperrovin, kde každá tato rovina je tečnou k alespoň jednomu vzorku její kategorie. Vzdálenost mezi dvěma tečnami rovin je okraj klasifikátoru a ten chceme následně maximalizovat. Podrobnější popis lze nalézt v [1] na str. 337.

Dále se provede fáze vyhodnocení a prezentace. Po této fázi jsou emaily uspořádány do kategorií a je vytvořen model pro daného uživatele. Uživatel má nyní možnost prohlédnout si příslušné kategorie a provést korekce špatně zařazených emailů případně upravit nastavení modelu. Tím dochází k úpravě vytvořeného modelu, čímž se ovlivní fáze dolování a při příští klasifikaci dojde ke zpřesnění výsledků.

V procesu získávání znalostí z emailů dochází také k optimalizaci modelů, tedy k potlačení nepotřebných informací (suppression), řazení (ordering), prořezávání (pruning), zobecňování (generalization) a shlukování (clustering).

Při klasifikaci emailů má každá kategorie vytvořený svůj klasifikátor, podle kterého se určí, jestli daný email spadá do této kategorie či nikoliv. Jedná se o tzv. multi-class klasifikaci, kde dokument patří do 0 až *n* kategorií. Email je proto při této klasifikaci přiřazen do všech tříd, které mají ohodnocení nad zvolenou mez.

Každý klasifikační algoritmus má dvě fáze - trénovací a testovací. Ve fázi trénování se vytváří klasifikátor na trénovacích datech, což jsou emaily, u kterých známe jejich zařazení. Testovací fáze poté ověřuje vytvořený klasifikátor na neznámých (testovacích) datech. Důležité je, aby testovací a trénovací data byla odlišná. Jestliže uživatel přesune své emaily do jiných kategorií, pak dojde k přetrénování, aby se při klasifikaci zohlednily tyto změny.

Při procesu klasifikace může nastat řada problémů ovlivňující výkon. Jedná se například o dlouhou dobu trénování, která může být neúnosná. Dalším problémem je přítomnost velkého počtu slov, která nemají sama o sobě význam. Tato slova způsobují šum ve vstupních datech. Při předzpracování emailů se proto použijí tzv. váhovací metody. Nejznámější metoda je TF-IDF (Term Frequency-Inverse Document Frequency). Podle [1] nejprve vypočítá frekvenci termu (slova) v emailu a poté počet výskytů slova v rámci kolekce emailů. Na závěr se z obou hodnot vypočítá váha daného termu. Podle této hodnoty se určí často vyskytující se slova a ta se mohou odstranit. Další možností je využití tzv. stop-listu, který obsahuje nevýznamová slova. Při předzpracování dat se poté každé slovo v dokumentu porovnává s tímto seznamem a je-li nalezena shoda, je slovo odstraněno.

3. ZÁVĚR

Emailová komunikace je v dnešní době nedílnou součástí našeho života. Problémy nám působí správné zařazení příchozí pošty do příslušné kategorie. Každý den proto ztrácíme drahocenný čas. U schránek pro osobní použití existuje řešení v podobě filtrovacích pravidel, ale tato metoda se stává neefektivní při větším množství různorodé příchozí pošty.

Práce představuje princip klasifikace emailů na základě rozpoznávání a klasifikace textu. Využity jsou dva algoritmy strojového učení, jedná se o Naive Bayes a Support Vector Machines. Je zde také ukázána interakce uživatele při třídění pošty. Právě uživatel je klíčovým prvkem, podle kterého se určuje, jak se mají emaily klasifikovat. Z tohoto důvodu má každý uživatel svůj model a ten se s časem a požadavky mění a upravuje. Na závěr jsou zmíněny problémy, které mohou nastat při procesu klasifikace. Jedná se o dlouho dobu při trénování klasifikačního algoritmu nebo přítomnost slov bez významu ve vstupních datech.

REFERENCE

[1] HAN, Jiawei a Micheline KAMBER. *Data Mining: Concepts and Techniques*. Second Edition. San Francisco: Morgan Kaufmann Publishers, 2006, 770 s. ISBN 1-55860-901-6.

[2] FELDMAN, Ronen a James SANGER. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press, 2007, 410 s. ISBN 0-521-83657-3.

[3] BERRY, Michael W. (editor) a Jacob KOGAN (editor). *Text Mining: Applications and Theory*. Chichester: Wiley, 2010, 207 s. ISBN 978-0-470-74982-1.

[4] RFC 5322. *Internet Message Format*. October 2008. Dostupné z: tools.ietf.org/html/rfc5322