EXTRAKCE DAT ZE ZPRÁV ELEKTRONICKÉ POŠTY

Jan Kosnar[1], Miloslav Konopík[2]

1 ÚVOD

Elektronická pošta je běžným způsobem výměny informací mezi lidmi za pomoci počítače či mobilního zařízení. Velmi často obsahuje údaje, které je možno (při určité míře strojového porozumění) dále zpracovávat a umožnit tak uživateli jejich komfortnější správu a využití.

Elektronická komunikace probíhá v přirozeném jazyce, tj. v jazyce, kterým se lidé běžně dorozumívají. Pro zpracovaní e-mailových zpráv je tedy potřeba implementovat mechanismy, které umožní přirozenému jazyku porozumět. Porozuměním a strojovou interpretací přirozeného jazyka se zabývá vědní oblast zvaná Natural Language Understanding (NLU, porozumění přirozenému jazyku).

Automatická extrakce dat z textů e-mailů v přirozeném jazyce otevírá značné možnosti pro rozšíření funkcionality již používaných programů a celkové zkvalitnění procesu správy elektronické pošty.

2 ZÍSKÁVÁNÍ INFORMACÍ O SCHŮZKÁCH

Oblastí dat, tzv. domén, které lze z e-mailů extrahovat a způsobů jejich využití je celá řada. Aplikace je proto navrhována s důrazem na modularitu. Jako základní oblast však byla zvolena problematika „vyhledávání informací o schůzkách“. Cílem celého projektu je tedy vytvořit jednoduchý mechanismus, který umožní uživatelům el. pošty získávat ze zpráv informace o místě, datu a čase konání schůzky a tyto údaje dále využívat, například k automatickému uložení do uživatelského kalendáře.

3 INTEGRACE DO EXISTUJÍCÍCH APLIKACÍ

Pro zajištění dostatečného uživatelského komfortu je vhodné aplikaci pro extrakci dat integrovat do již existujících řešení, které uživatelé běžně používají ke správě elektronické pošty a kalendáře. Příkladem takové integrace může být začlenění programu do open source aplikace Mozilla Thunderbird nebo přímo do on-line řešení (tzv. webmailu) libovolného poskytovatele. Aplikací pro správu kalendáře existuje rovněž velké množství, pro příklad uveďme on-line plánovač společnosti Google – Google Calendar.

4 SÉMANTICKÁ ANALÝZA

K sémantické analýze lze přistupovat za pomoci tzv. empirických nebo stochastických metod. Empirickým přístupem je označováno expertní vytváření pravidel, na jejichž základě

[1] Bc. Jan Kosnar, student navazujícího studijního programu Inženýrská informatika, obor Softwarové inženýrství, e-mail: jkosnar@students.zcu.cz

[2] Ing. Miloslav Konopík, Ph.D., ZČU v Plzni, FAV, Katedra informatiky a výpočetní techniky, Univerzitní 22, 306 14 Plzeň, tel.: +420 377632418, e-mail: konopik@kiv.zcu.cz (vedoucí práce)

jsou v textu vyhledávány určité struktury. Stochastické algoritmy naopak nevyžadují tak úzkou specifikaci a pracují na principu statistických metod a automatického učení.

Na příkladu vyhledávaní informací o schůzkách je možné tyto dva přístupy dobře odlišit. Empirický přístup je vhodné využít pro vyhledávání data a času konání schůzky – Data a časy jsou většinou specifikovány číselně v ustáleném formátu nebo za pomoci definovatelných slovních spojení (zítra, večer, …). Pro určení místa konání schůzky je naopak vhodné použít spíše metod založených na automatickém učení, neboť specifikovat pro ně úplná a konkrétní pravidla je velmi obtížné. Dále je také třeba zohlednit fakt, že v textu zprávy může být obsaženo značné množství slov a slovních spojení, která určují datum a čas – z nich je nutné automaticky vybrat ty záznamy, jež odpovídají požadované realitě.

Pro ilustraci je uveden následující jednoduchý příklad:

Schůzka ke dni otevřených dveří, který se koná 5. 6. 2010 se uskuteční již zítra. Sraz realizačního týmu je v 14:00 v aule. Předpokládaný konec – 16:30.

V této zprávě je nutno identifikovat správný nález, tj. „*zítra, 14:00“* a převést jej do podoby, která může být uložena v kalendáři.

5 REALIZOVANÉ ALGORITMY

Během projektu bylo vytvořeno a specifikováno několik empirických i stochastických metod zaměřených zejména na identifikaci správného data a času. Takzvaný „bodovací“ algoritmus vychází z tzv. principu lokality, který byl identifikován na základě analýzy e-mailových zpráv a poté experimentálně ověřen. Tento princip říká, že datum, čas a místo schůzky bývá určeno „na jednom místě“, v jedné větě. Tento algoritmus ohodnocuje zdrojový text na základě celé řady pravidel a poskytuje velmi obstojné výsledky. Dále implementovaný statistický algoritmus pracuje na principu naivního Bayesovského klasifikátoru. Jeho výsledky jsou závislé na rozsahu a charakteru natrénovaných dat. Pro množinu zpráv menší než 50 metoda neposkytuje kvalitní výsledky. Postupným zvětšováním datové základny dochází ke zlepšování výsledků algoritmu, které umožní správné zpracování i těch typů zpráv, jež nereflektuje bodovací algoritmus. Určení dostatečného rozsahu dat je stále předmětem dalšího zkoumání.

6 ZÁVĚR

Aplikace je vyvíjena za podpory Laboratoře inteligentních komunikačních systémů Katedry informatiky a výpočetní techniky. Výsledky poskytované výše zmíněnými algoritmy, se jeví jako dostačující pro základní určení data a času – reálná úspěšnost se pohybuje okolo 80%. Prostor pro další vylepšování projektu zůstává zejména v oblasti extrakce informace o místě konání schůzky. Aplikace jako celek je integrována do prostředí poštovního klienta Mozilla Thunderbird v podobě jediného tlačítka a umožňuje rovněž uložení nalezených dat do aplikace Google Calendar.

LITERATURA

Jurafsky, D., and Martin, James H., 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. 1 edition. Prentice Hall. ISBN 0-13-095069-6.

Allen, James. 1994. *Natural Language Understanding*. 1 edition. Redwood City, Benjamin/Cummings Publishing Company, Inc. ISBN 0-8053-0334-0.