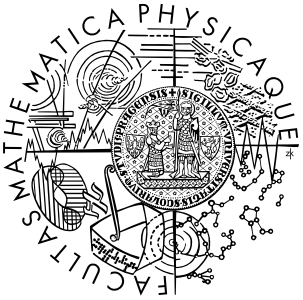
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Martin Pilát

Získávání a správa údajů o konferencích a workshopech

Katedra softwarového inženýrství

Vedoucí bakalářské práce: RNDr. Michal Žemlička, Ph.D.
Studijní program: Informatika, Obecná informatika

2007

Chtěl bych poděkovat RNDr. Michalu Žemličkovi, Ph.D. za vedení práce a cenné rady při jejím vypracování a Petře Šachové za korektury textu.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 23.5.2007 Martin Pilát

Obsah

Název práce: Získávání a správa údajů o konferencích a workshopech
Autor: Martin Pilát
Katedra: Katedra softwarového inženýrství
Vedoucí bakalářské práce: RNDr. Michal Žemlička, Ph.D.
E-mail vedoucího: zemlicka@ksi.mff.cuni.cz

Abstrakt: Cílem této práce je vytvořit aplikaci, která pomůže svým uživatelům vyznat se v záplavě informací o konaných konferencích, workshopech, kongresech a symposiích. Počet zobrazovaných událostí lze omezit pomocí pokročilého systému uživatelsky definovaných filtrů. Aplikace obsahuje propracovaný systém uživatelských práv a umožňuje vytváření osobních a skupinových kalendářů konferncí. Důležitou součástí je také modul pro automatické dolování informací o konaných akcích z příchozích zpráv. Tento modul používá jednoduché vyhledávání klíčových slov z ručně vytvořeného slovníku, k nalezení termínů, témat, názvu a dalších duležitých údají týkající se dané konference. Úspěšnost tohoto modulu je okolo 80% pro dolování termínů a okolo 70% při získávání dalších informací.

Klíčová slova: získávání informací z textu, konference, e-mailové zprávy, kalendář

Title: Mining and management of data on conferences and workshops
Author: Martin Pilát
Department: Department of Software Engineering
Supervisor: RNDr. Michal Žemlička, Ph.D.
Supervisor's e-mail address: zemlicka@ksi.mff.cuni.cz

Abstract: The goal of this work is to create an application, which would help its users to be acquianted with huge amounts of information on conferences, workshops, congresses, and symposiums. The number of shown events can be reduced by the use of an advanced system of user-defined filters. The application contains a sophisticated system of user privileges and provides support for creating personal and group calendars. An important part of the application is a module for automated retrieving of information from incoming e-mail messages. This module uses a simple pattern matching for extracting deadlines, topics, name, and other important data on a particular conference. These patterns are stored in a hand-crafted dictionary. The accuracy of this module is about 80% for mining the dates and about 70% for mining other types of data.

Keywords: text information extraction, conference, e-mail, calendar

Kapitola 1

Úvod

1.1 Motivace

Vědečtí pracovníci prezentují své výsledky na konferencích, workshopech, kongresech a symposiích po celém světě (dále jen konference příp. události). Aby si byli schopni udržet přehled o konajících se událostech, potřebují mít jejich kalendář. Existuje mnoho aplikací, které nějaký kalendář obsahují. Takové kalendáře ale nebývají určeny k hlídání konferencí a obvykle proto postrádají některé z funkcí, které jsou k danému účelu vhodné (např. uchovávání seznamů témat nebo vztahů mezi konferencemi).

Dále je potřeba sledovat různé zdroje, které informují o těchto událostech. Těmito zdroji jsou často e-mailové konference jako například [13, 9]. Informace z těchto zdrojů se ale týkají velkého počtu událostí, z nichž obvykle většina daného pracovníka nezajímá, a pročítání zpráv o nich mu jen zabírá čas. Proto je vhodné tyto informace nějakým způsobem sumarizovat, utřídit a umožnit v nich vyhledávání tak, aby každý mohl sledovat oznámení jen o událostech, které jej zajímají.

K tomuto účelu může sloužit například projekt [12], který umožňuje vytvoření personalizovaného kalendáře konferencí pro každého z uživatelů a který sloužil jako inspirace při vytváření této aplikace. Projekt obsahuje také skripty pro automatické dolování informací z příchozích zpráv, ale jejich účinnost není příliš vysoká. Rozhodli jsme se proto vytvořit nový projekt, který by obsahoval spolehlivější dolování informací z e-mailů, opravoval některé nedostatky zmíněného projektu a implementoval nové funkce, jako například podporu skupin uživatelů.

Kapitola 2

Řešení

2.1 Základní vlastnosti aplikace

Vlastní aplikace se skládá ze dvou částí. Jedna doluje důležité informace z příchozích zpráv elektronické pošty (e-mailů) a druhá zobrazuje vydolované informace na webových stránkách a tvoří vlastní uživatelské rozhraní.

2.1.1 Webová část

Tato část aplikace zobrazuje informace o konferencích uložených v databázi svým uživatelům. Podporuje tvorbu osobních a skupinových kalendářů, přidávání konferencí, kontrolu automaticky vydolovaných dat a mnoho dalších funkcí.

Tím, že se informace o konferencích sdílí mezi jednotlivými uživateli, aplikace šetří všem svým uživatelům mnoho času při jejich vyhledávání. To, že aplikace je webová a tedy přístupná z celého Internetu, zase umožňuje, že se do ní uživatelé mohou přihlásit odkudkoli z celého světa, i pokud zrovna nejsou doma nebo v kanceláři, a mají tak snadný přístup ke svému kalendáři kdykoliv a odkudkoliv.

Kompletní popis funkcí této části najdete v kapitole “Uživatelská dokumentace”.

2.1.2 Zpracování příchozích zpráv

S touto částí běžný uživatel nikdy nepřijde do styku. Dokonce ani nemusí vědět, že taková část existuje. Tato část zpracovává každý příchozí e-mail a doluje z něj data v něm obsažená, která nakonec uloží do databáze. Téměř všechno její nastavení (kromě parametrů připojení k databázi) se provádí přes webovou část aplikace.

Podrobný popis toho, jak tato část funguje, je v kapitole “Dolování informací”. Podrobnosti ohledně nastavování této části jsou uvedeny v kapitole “Uživatelská dokumentace”.

2.1.3 Použité prostředky

Část pro dolování informací je napsána v jazyce C++. Tento jazyk byl zvolen pro svou rychlost a snadnou přenositelnost mezi platformami na úrovni zdrojového kódu.

Důležitým důvodem k tomuto rozhodnutí také bylo, že jej lze provozovat i na serverech, které typicky nemají interprety vyšších jazyků, jakým je například Java.

Webová část je napsána pomocí skriptovacího jazyka PHP [2], který generuje stránky v jazyce HTML 4.01 Transitional. Data jsou uložena v MySQL databázi. Tato kombinace byla zvolena opět pro svou relativně snadnou dostupnost a rozšířenost.

2.1.4 Zabezpečení

Vzhledem k tomu, že aplikace neobsahuje žádné tajné informace, největším bezpečnostním rizikem by byl únik e-mailových adres jednotlivých uživatelů. Tomu také odpovídá míra zabezpečení, která je implementována. Toto zabezpečení je velmi jednoduché, používá autorizaci pomocí uživatelského jména a hesla. Navíc e-mailové adresy uživatelů jsou dostupné jen úzké skupině administrátorů.

Zabezpečení je řešeno pomocí technologie sessions (někdy též sezení). Tato technologie vyžaduje podporu cookies na straně serveru i uživatele, případně nastavení serveru, které předává identifikátor sezení v adrese URL.

Kapitola 3

Dolování informací

3.1 Zdroj informací

Informace o událostech (ať již se nazývají konferencemi, workshopy, či jinak) jsou dolovány z příchozích zpráv elektronické pošty. Existuje mnoho mail-listů (skupinových diskuzí probíhajících prostřednictvím hromadného rozesílání zpráv elektronické pošty), které poskytují tyto informace. Některé jsou obecné, jiné závislé na dané konferenci, nebo organizátorovi. Příchozí zprávy jsou většinou neformátované textové dokumenty, vzácněji mohou být v podobě html dokumentů, či v ještě jiném formátu. Zprávy jsou graficky formátovány tak, aby se v nich snadno vyznal člověk, ale nepočítá se s jejich automatickým zpracováním.

Některé mail-listy přidávají některé informace o událostech také do hlaviček zpráv (např. DBWorld [9]), zatímco jiné zdroje zpráv (např. SEWorld [13], či mail-listy pořadatelů konkrétních událostí) takovouto službu neposkytují. Údaje v hlavičce zpráv však bohužel obsahují více chyb než samotný nestrukturovaný text a neobsahují všechny důležité informace. Také se nedá očekávat, že by všechny zdroje využívaly stejného mapování údajů do hlaviček. Proto není možné použít údaje v hlavičkách zpráv jako jediný zdroj informací.

Písemná mezilidská komunikace využívá různé základní vzory (např. nejdůležitější informace bývají na začátku zprávy). Tyto vzory se snažíme využít v heuristikách pro dolování informací. Problém je, že stejná věc se dá vyjádřit mnoha způsoby, proto nelze použít jednoduché vyhledávání v textu. E-maily neobsahují pouze informace výzvy k účasti na konferencích, workshopech, či jiných událostech, ale také výzvy k pořádání wokshopů v rámci konferencí, informace o pracovních příležitostech, nových časopisech a knihách, nominacích i udílení prestižních cen a dalších novinek z oboru. Navíc každé události se obvykle týká více zpráv. Některé z těchto zpráv opravují pouze chyby ze zpráv odeslaných dříve, jiné přinášejí novou informaci. Často také jedna zpráva obsahuje informaci o více událostech, které spolu nějak souvisí.

3.2 Požadované a dolované informace

Vědci používají různé přístupy při vybírání správné konference pro prezentaci svých výsledků. Někdy již ze zkušenosti znají jména některých zajímavých konferencí, mohou ale také procházet všechna získaná data a hledat v nich tu správnou konferenci. Tento přístup je použitelný jen v případě, kdy stačí zkontrolovat jen několik málo konferencí. Konferencí jsou však stovky. Proto je nutné konference, které je potřeba zkontrolovat, nějak omezit. Proto dolujeme také informace o tématech, místě konání a způsobu publikace (zda bude sborník vydán před nebo až po konferenci, kým bude vydán...).

Při plánování je vhodné mít přehled o termínech dané konference – termíny pro odeslání abstraktu, článku, finální verze a data vyrozumění o přijetí, registrace a konání konference – a občas také další termíny týkající se sborníku vydávaného až po konferenci (tzv. post-proceedings).

Aby uživatelé mohli zkontrolovat data, případně zjistit další informace, dolujeme i odkaz na webové stránky konference.

Někdy je užitečné vědět, zda jsou nějaké jiné události kolokované s danou konferencí nebo zda je tato konference součástí nějaké větší události atd.

3.3 Postup

Zprávy zpracováváme v následujících třech fázích: předzpracování, dolování informací a přiřazení informací ke správné události.

3.3.1 Předzpracování

Zpráva se prochází a hledají se v ní jména měsíců. Najde-li se nějaké, je nahrazeno speciálním znakem a pořadí měsíce v roce se uloží, aby jej bylo možné použít v následující fázi. Při předzpracování se nerozlišují velká a malá písmena a hledají se pouze celá slova.

Tato fáze zjednodušuje zpracovávání různých formátů dat a datových intervalů, které se ve zprávě objevují.

3.3.2 Dolování informací

Předzpracované tělo e-mailu se znova prochází a hledají se definovaná klíčová slova. Dle významu nalezených slov se používají různé heuristiky pro dolování informací (tyto heuristiky jsou podrobněji probrány níže).

Jestliže algoritmus najde datum, nebo interval dat, zkonvertuje jej do ISO formátu, uloží jej a pokračuje v hledání. Najde-li známé klíčové slovo, uloží jeho význam. Konečně najde-li konec řádky, zjistí, jaké našel významy klíčových slov a jaká data, a zachová se podle příslušných heuristik.

3.3.3 Přiřazování informací k událostem

V poslední fázi aplikace kontroluje, zda existuje jiná zpráva s informacemi o téže konferenci, jako právě zpracovávaná zpráva; je-li taková zpráva nalezena, přiřadí novou zprávu ke stejné konferenci. Při tomto přiřazení se používá zkratka a ročník konference. Informace z těchto dvou zpráv neslučujeme, protože se domníváme, že kontrolování informací sloučených z více zpráv by bylo složitější, než je kontrolovat samostatně.

3.4 Heuristiky

Používáme různé heuristiky pro získávání různých typů informací. Tyto heuristiky jsou založeny na formátování zpráv, tak, aby byly dobře čitelné pro člověka. Například na začátku zprávy obvykle bývá uvedeno jméno události, datum a místo konání.

3.4.1 Termíny/data

Je-li nalezeno datum někde na začátku zprávy a zároveň na téže řádce není nalezeno jiné klíčové slovo, případně jen jméno státu, tak se toto datum (resp. časový interval) považuje za datum konání konference.

Je-li nalezeno datum a na stejné řádce klíčové slovo s významem nějakého termínu, je toto datum přiřazeno danému termínu.

3.4.2 Místo konání

Najde-li se jméno státu, celá řádka před ním se považuje za místo konání konference, pokud na této řádce byl nalezen ještě jiný údaj před jménem státu (např. datum), použije se část řádky začínající za tímto údajem.

3.4.3 Témata konference

Najde-li se řetězec s významem "Začátek sekce témat", následující řádky se přidávají k tématům, dokud není nalezen konec sekce (tj. buď řetězec s významem "konec sekce", nebo dvě prázdné řádky za sebou). Je-li nalezena jen jedna prázdná řádka, následující řádka se zkontroluje, zda začíná stejným znakem jako první řádka témat, není-li tomu tak, sekce témat končí, jinak pokračuje dál. Tato heuristika je motivována tím, že témata často obsahují odrážky, nebo jsou odsazena od začátku řádky.

3.4.4 Jméno konference

Najde-li se klíčové slovo s významem "Část jména konference" někde na začátku zprávy, a žádné jiné klíčové slovo se na stejné řádce nenajde, jsou tato a následující řádky považovány za název události. Název končí, pokud je nalezeno jiné klíčové slovo, nebo je delší než tři řádky. Při hledání jména konference se také používá pole

“Předmět” z hlavičky e-mailu. V tomto případě se z tohoto pole odstraní všechna klíčová slova s významem “Standardní část předmětu” (např. jméno mail-listu) a je-li zbytek delší než předem definovaný počet znaků, považuje se za jméno konference.

3.4.5 Zkratka

Je-li nalezeno krátké neznámé slovo na začátku zprávy, označí se jako jeden z kandidátů na zkratku konference. Zkratka se dále hledá v předmětu zprávy a ve jméně konference. V těch se hledají krátké části oddělené pomlčkou nebo dvojtečkou, případně uzavřené v závorkách.

Některé mail-listy uvádějí zkratku ve speciálním poli v hlavičce e-mailu.

3.5 Měření

Výsledky byly získány ručním ověřením informací vydolovaných naší aplikací ze 100 zpráv z DBWorldu [9] a 45 zpráv z SEWorldu [13]. Uvádíme celkové výsledky i výsledky rozdělené dle zdrojů.

Napřed jsem vydolovali 50 e-mailů z DBWorldu, potvrdili informace, které se v nich vyskytovaly, přidali nově získaná klíčová slova a opakovali stejný postup s dalšími 50 e-maily ze stejného zdroje.

Nakonec jsme otestovali úspěšnost aplikace na zprávách ze SEWorldu, abychom zjistili, jak je ovlivněna užitím různých zdrojů.

Statistiky pro jméno byly počítány, jen při první zmínce o dané konferenci, protože aplikace vyplňuje jméno podle informací, jaké již o dané události má. To pomáhá zachovat stejné jméno konference po celou dobu, kdy je v databázi. Jejich jména v e-mailech se často mění, přestože se jedná o stejnou událost. Z tohoto důvodu nejsou dostupné některé údaje, které se týkají jména konference.

3.6 Výsledky

Dosáhli jsme úspěšnosti okolo 80% při dolování termínů a okolo 65% při dolování ostatních druhů informací. Detaily jsou v Tabulkách 3.1, 3.2 a 3.3 a v diskuzi níže.

Sloupce ve všech tabulkách mají tyto významy.

správně – počet e-mailů, ve kterých informace byla a byla správně vydolována

nenalezeno – počet e-mailů, ve kterých informace byla, ale nebyla vůbec vydolována

chybně – počet e-mailů, ve kterých informace byla a byla vydolována špatně

dostupné – počet e-mailů, které obsahovaly danou informaci

vydolováno správně – *správně* + počet e-mailů, které neobsahovaly danou informaci a bylo správně označeno, že ji neobsahují

Popis	Správně	Nenalezeno	Chybně	Dostupné	Vydolováno		Poměr	Přesnost
					Správně	Chybně		
Abstrakt	20	1	0	21	99	1	0.95	0.99
Článek	75	15	4	94	81	19	0.80	0.81
Vyrozumění o přijetí	81	4	4	89	92	8	0.91	0.92
Finální verze	46	35	2	83	63	37	0.55	0.63
Registrce	14	2	1	17	97	3	0.82	0.97
Začátek	77	8	9	94	83	17	0.82	0.83
Konec	76	8	10	94	82	18	0.81	0.82
Celkově	389	73	30	492	597	103	0.79	0.85
Jméno	39	6	22	67	N/A	28	0.58	N/A
Témata	29	28	25	82	47	53	0.35	0.47
Stát	68	20	2	90	78	22	0.76	0.78
Město	58	18	11	87	71	29	0.67	0.71
Celkově	194	72	60	326	N/A	132	0.60	N/A

Tabulka 3.1: Přehled výsledků, DBWorld

Popis	Správně	Nenalezeno	Chybně	Dostupné	Vydolováno		Poměr	Přesnost
					Správně	Chybně		
Abstrakt	4	1	1	6	43	2	0.66	0.96
Článek	29	6	2	37	37	8	0.78	0.82
Vyrozumění o přijetí	32	1	2	35	42	3	0.91	0.93
Finální verze	24	6	1	31	38	7	0.77	0.84
Registrace	6	0	1	7	44	1	0.85	0.98
Začátek	35	7	3	45	35	10	0.78	0.78
Konec	34	9	2	45	34	11	0.76	0.76
Celkově	164	30	12	206	273	42	0.80	0.87
Jméno	24	3	17	44	N/A	20	0.55	N/A
Témata	23	10	4	37	31	14	0.62	0.69
Stát	39	2	2	43	41	4	0.91	0.91
Město	37	2	4	43	39	6	0.86	0.89
Celkově	123	17	27	167	N/A	44	0.74	N/A

Tabulka 3.2: Přehled výsledků, SEWorld

vydolováno chybně – celkový počet e-mailů − *vydolováno správně*, tj. počet e-mailů, ve kterých daná informace nebyla obsažena a aplikace přesto něco našla, nebo ve kterých daná informace byla, ale byla vydolována chybně případně vůbec

poměr – *správně / dostupné*,

přesnost – *přiřazeno správně*/celkový počet e-mailů

3.7 Diskuze

3.7.1 Data a termíny

Ve většině případů jsme při dolování tohoto typu dat dosáhli přesnosti okolo 80%. Jedinou vyjímkou je termín pro odeslání finální verze příspěvků získaný ze zpráv z DBWorldu. Toto je způsobeno velkým množstvím různých vyjádření, která mají tento význam. Přestože, jsme jich spoustu shromáždili ještě předtím, než jsme začali získávat statistiky, objevili jsme ještě mnoho dalších. Předpokládáme (a výsledky z SEWorldu nás v tom utvrzují), že po delší době užívání a shromáždění většího počtu těchto výrazů dosáhneme ještě vyšší účinnosti.

Popis	Správně	Nenalezeno	Chybně	Dostupné	Vydolováno		Poměr	Přesnost
					Správně	Chybně		
Abstrakt	24	2	1	27	142	3	0.89	0.99
Článek	104	21	6	131	118	27	0.79	0.81
Vyrozumění o přijetí	113	5	6	124	134	11	0.91	0.92
Finální verze	70	41	3	114	101	44	0.61	0.70
Registrace	20	2	2	24	141	4	0.83	0.97
Začátek	112	15	12	139	118	27	0.81	0.81
Konec	110	17	12	139	116	29	0.79	0.80
Celkově	553	103	42	698	870	145	0.79	0.79
Jméno	63	9	39	111	N/A	48	0.58	N/A
Témata	52	38	29	119	78	67	0.44	0.54
Stát	107	22	4	133	119	26	0.80	0.82
Město	95	20	15	130	110	35	0.73	0.76
Celkově	317	89	87	493	N/A	176	0.64	N/A

Tabulka 3.3: Přehled výsledků, celkově

Aplikace také získává informace o termínech pro odeslání finální verze příspěvků a vyrozumění o přijetí pro post-proceedings (pokud jsou jiné, než ty pro konferenci) a termín pro odeslání rozšířeného abstraktu. Ty ale nejsou uvedeny ve statistikách, neboť se vyskytují příliš zřídka na to, aby poskytly spolehlivé statistiky (každý z nich jsme objevili jen jednou).

3.7.2 Jméno, web, témata, stát, město

Dolování tohoto typu dat je mnohem obtížnější, a proto jsme dosáhli menší přesnosti. Dolování těchto informací je také výrazně ovlivněno formátováním zpráv. To je možné doložit analýzou úspěšnosti u zpráv ze SEWorldu. SEWorld je moderovaný mail-list, kde jsou povoleny zprávy jen v prostém textu, a kde zkratka nebo jméno konference jsou vždy uvedeny v předmětu zprávy a celkově jsou zprávy lépe čitelné i pro člověka.

Přesnost dolování témat konference je jen 35% pro zprávy z DBWorldu ale 62% pro zprávy ze SEWorldu. V tomto případě lepší formátování zpráv ze SEWorldu velmi pomohlo. Přesto tento druh informace není jednoduché najít. Občas je tato část odsazena, ale často tomu tak není. Ani najít konec této části není jednoduché. Nepodařilo se nám najít způsob, jak rozhodnout, kde část s tématy končí, protože je často rovnou následována dalším textem a oddělena pouze prázdnou řádkou, nebo dokonce bez ní. Nemůžeme ale použít prázdnou řádku jako konec této sekce, protože témata jsou často rozdělena do více částí a prázdná řádka se používá k tomuto účelu.

Najít tuto část je občas složité nejen pro počítač, ale i pro člověka, který jen zběžně pročítá e-mail.

Dosáhli jsme přesnosti okolo 60% při dolování jména konference. Zde se výsledky mezi oběma diskutovanými zdroji příliš neliší.

Naše jednoduchá metoda vyžaduje, aby jméno obsahovalo jedno z definovaných klíčových slov. Neobsahuje-li takové slovo, není název konference nalezen, pokud se nevyskytuje v předmětu e-mailu.

Tento způsob je poměrně účinný, protože většina událostí nějaké takové slovo obsahuje. Největším problémem je, že mnoho e-mailů obsahuje také jména kolokovaných nebo jinak příbuzných událostí, a potom je obtížné vydolovat správné jméno.

Občas také získáme pouze část jména, pokud je rozepsáno na více řádek. Naopak pokud se jméno vyskytuje například pouze v prvním odstavci, získáme i okolní text.

Většina chyb, které se vyskytly při dolování jména státu, kde se konference koná, vznikla, protože jméno nebylo v e-mailu vůbec uvedeno. Občas je uvedeno pouze město nebo oblast, kde se událost koná. Toto se nestává u zpráv ze SEWorldu, protože tam lidé, kteří zprávy píší, vždy zmiňují stát, kde se událost koná, a proto jsou statistiky získané z těchto zpráv mnohem lepší.

3.8 Podobná práce

Našli jsme několik aplikací, sloužících k podobnému účelu jako naše aplikace. Podrobněji rozebereme dvě. Message Parse [8] a Advanced Email Parser [14] umožňují dolování informací z e-mailů, ale zdá se, že Message Parse neumí dolovat bloky textu, což je potřeba například pro dolování témat a jména konference. V obou aplikacích je problém vyjádřit, že množina klíčových slov má stejný význam. Toto by se pravděpodobně dalo vyřešit díky jejich podpoře skriptování, ale to by se neprovádělo snadno a domníváme se, že by to také bylo neefektivní. Navíc zpracování různých formátů data by také nebylo triviální a přidání podpory pro nový formát by vyžadovalo přepsání skriptů. Pravděpodobně by bylo možné vytvořit aplikaci, která by skripty přepisovala automaticky, ale domníváme se, že její napsání by bylo téměř stejně obtížné, jako vytvoření nové aplikace.

Na druhou stranu naše aplikace dovoluje přidávání klíčových slov a formátů data jednoduše přes webové rozhraní pomocí dvou až třech kliknutí. To samé platí pro přidávání formátů data. To dělá administraci naší aplikace mnohem jednodušší.

Existuje mnoho článků (např. [5, 7, 11]), které se zabývají dobýváním informací z e-mailů. Ty se obvykle soustředí na mapování komunikace mezi odesilateli a příjemci, nebo na strukturu zpráv.

Například [4] a [6] uvádějí metodu podobnou té námi použité, ale používají pouze omezené množství formátů data a dolují méně údajů. Konfigurace přes webové rozhraní také není podporována.

3.9 Problémy

Mezi nejdůležitější problémy, které má náš přístup při hledání informací ve zprávách, patří, že občas zpráva obsahuje informace o více událostech, které mají obecně jiné termíny, témata atd. Potom je obtížné rozhodnout, které informace patří k jaké konferenci.

Dalším problémem je, že jméno konference je občas napsané jen v prvním odstavci zprávy, potom vydolujeme mnohem více textu, než jen jméno konference, především pokud toto jméno není v předmětu zprávy.

Také nalezení části s tématy konference (a ničeho navíc) je celkem obtížné, především v e-mailech z DBWorldu a jiných listů, které nemají přesná formátovací pravidla.

3.10 Zhodnocení

Námi použitá metoda vyhovuje účelu, za jakým byla vyvíjena. Šetří svým uživatelům čas při zadávání informací o konferencích, neboť je mnohem jednodušší informaci pouze upravit a potvrdit, než ji ručně zadávat.

Získáních dalších klíčových slov a ladění heuristik může přinést další zlepšení výsledků.

Kapitola 4

Uživatelská dokumentace

Mnoho možností zde popsaných závisí na konrétním nastavení vašich uživatelských práv. Nezobrazuje-li se vám nějaká zde popsaná možnost, znamená to, že nemáte práva příslušnou činnost provádět.

Sekce 4.1 až 4.9 jsou určeny všem uživatelům, sekce 4.10 je určena především pro správce aplikace.

4.1 O aplikaci

Aplikace slouží primárně k zobrazení a správě informací o konferencích, workshopech a dalších podobných událostech (dále jen konference, nebo události). Mezi její základní funkce patří přidávání a upravování údajů o konferencích a potvrzení informací, které byly automaticky vydolovány z e-mailů. Dále je možné nastavovat uživatelské filtry zobrazovaných konferencí. Každý uživatel má svůj kalendář hlídaných konferencí (tj. konferencí, které jsou pro něj jakkoli zajímavé), u každé z nich si může napsat krátký komentář a nastavit její stav.

Skupiny spolupracovníků si mohou vytvářet skupinové kalendáře. Stav konference je v každé skupině sdílen, stejně jako její komentář. Navíc u každé konference je jednoduchá skupinová diskuze.

Jako volitelnou součást lze nainstalovat i modul pro automatické zpracování údajů z příchozích e-mailových zpráv. Tyto údaje se pak zobrazují ve zvláštní sekci a lze je použít při přidávání konference do systému.

4.2 Požadavky na software

K provozování aplikace potřebujete libovolný internetový prohlížeč. Pro snadnější užívání je doporučena podpora JavaScriptu, ale většina funkcí funguje i bez ní.

4.3 Vznik uživatelského účtu

Je-li to povoleno administrátorem, v pravém horním rohu je odkaz na stránku s registrací. Není-li tam tento odkaz, musíte napsat administrátorovi, aby vám účet vytvořil.

Po kliknutí na příslušný odkaz se zobrazí stránka, kde se můžete zaregistrovat. Vyplňte uživatelské jméno a heslo. Je možné vyplnit vyplnit i e-mailovou adresu, na kterou se vám budou zasílat informace o změnách ve vašich hlídaných konferencích (pokud si toto zasílání zapnete v nastavení). Po kliknutí na tlačítko "Zaregistrovat" proběhne registrace a je-li vše v pořádku, můžete se přihlásit.

4.4 Jak přidat konferenci?

Detailní popis všech možností přidání konference naleznete níže v sekci "Přidání a editace konference". Formulář je na obrázku 4.8.

4.4.1 Ruční přidání

1. Výběrem "Přidat konferenci" v Menu přejděte na stránku pro přidání konference.
2. V zobrazeném formuláři vyplňte údaje o konferenci, data do polí pro termíny vyplňujte ve formátu, který máte aktuálně nastaven.
3. Kliknutím na příslušná tlačítka přidejte kolokované, nadřízené a podřízené konference.
4. Údaje uložte kliknutím na jedno z tlačítek pod formulářem.

4.4.2 Přidání konference z archívu

1. Výběrem "Přidat konferenci" v Menu přejděte na stránku pro přidání konference.
2. Klikněte na tlačítko "Vybrat konferenci z archívu" pod polem pro jméno konference.
3. V zobrazeném formuláři vyhledejte konferenci, kterou chcete přidat, vyberte ji kliknutím na "Vybrat".
4. Doplňte údaje stejně jako při ručním přidání konference.
5. Údaje uložte kliknutím na jedno z tlačítek pod formulářem.

4.4.3 Přidání konference z nepotvrzených konferencí

1. Výběrem "Nepotvrzené konference" v Menu přejděte na seznam nepotvrzených konferencí.

2. Vyhledejte konferenci, kterou chcete přidat, a klikněte na ikonku editace na příslušném řádku.

3. Nejde-li změnit jméno konference, znamená to, že údaje byly automaticky přiřazeny k již existující události. Zkontrolujte přiřazení a případně jej opravte kliknutím na "Vybrat jinou konferenci". Není-li konference ještě v systému (a přesto je k něčemu přiřazena), klikněte na "Nová konference".

4. Kliknutím na "Ukázat e-mail" můžete zobrazit zprávu, ze které byly vydolovány informace zobrazené ve formuláři.

5. Doplňte a opravte automaticky vydolované údaje stejně jako při ručním přidání.

6. Údaje uložte kliknutím na jedno z tlačítek pod formulářem.

4.5 Jak sloučit konference?

Formulář prou sloučení konferencí s popisky jednotlivých částí je na obrázku 4.9.

1. Vyberte "Editovat konferenci" v Menu.

2. Vyhledejte jednu z konferencí, které chcete sloučit a ikonkou editace přejděte na její editaci.

3. Klikněte na tlačítko "Sloučit s jinou konferencí".

4. V zobrazeném formuláři vyhledejte a vyberte konferenci, se kterou chcete slučovat.

5. Zaškrtnutím příslušných políček vyberte údaje, které mají být ve výsledku sloučení.

6. Uložte sloučenou konferenci.

4.6 Jak přidat konferenci mezi hlídané? Jak změnit její stav?

Popisovaný formulář je na obrázku 4.2.

1. Kliknutím na zkratku konference v Indexu, kalendáři, nebo seznamu konferencí pro editaci přejděte na detaily konference.

2. Klikněte na tlačítko “Hlídat konferenci”. Tím se konference přidá mezi hlídané.
3. Stav konference můžete změnit jeho výběrem v příslušném seznamu.
4. Ke konferenci si můžete připsat vlastní komentář jeho napsáním do pole “Komentář” a kliknutím na “Nastavit”.
5. Konferenci odeberete z hlídaných konferencí kliknutím na “Nehlídat konferenci”.

4.7 Jak vytvořit filtr konferencí?

Formulář najdete na obrázku 4.4.

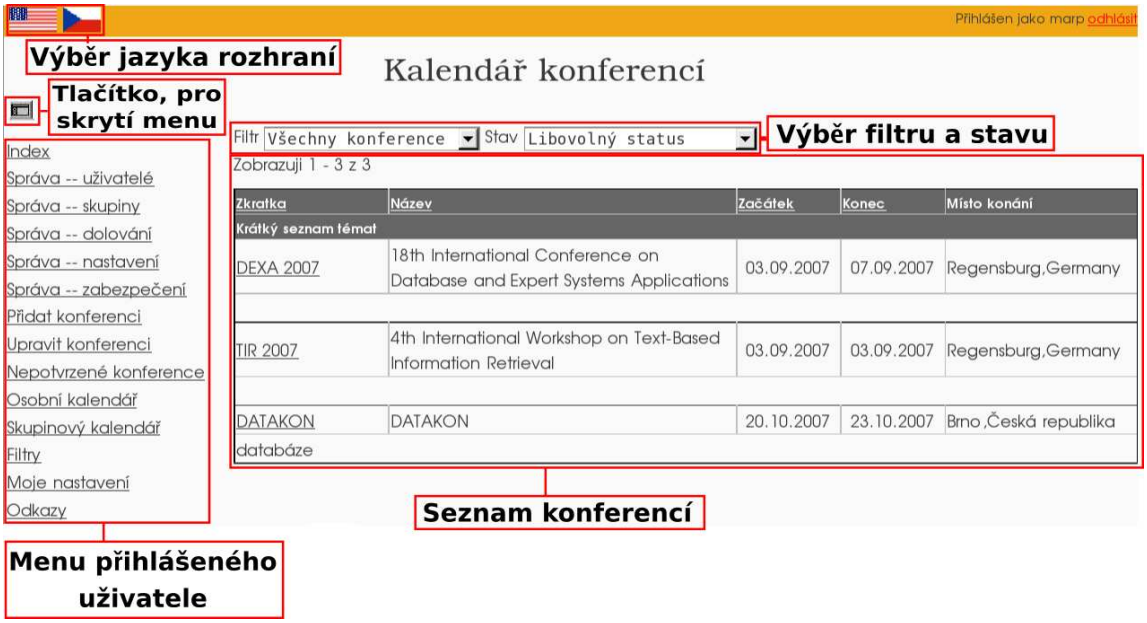
1. Kliknutím na “Filtry” přejděte do sekce pro editaci filtrů.
2. Do pole jméno napište jméno nového filtru. Nepovinně si můžete také napsat komentář.
3. V prvním sloupci vyberte atribut, podle kterého chcete filtrovat, nebo “filtr”.
4. Ve druhém sloupci nastavte operátor. (vybrali-li jste v předchozím kroku možnost “filtr”, nemá volba operátoru žádný vliv).
5. Do třetího sloupce napište hodnotu atributu, nebo vyberte svůj existující filtr ze seznamu (pokud jste vybrali v 3. kroku “filtr”).
6. Vyberte logickou spojku pro spojení s další podmínkou.
7. Jestliže potřebuje přidat další podmínku a již nemáte volnou řádku, klikněte na přidat podmínku.
8. Opakujte kroky 3-7, dokud nepřidáte všechny požadované podmínky.
9. Označte podmínky, které mají být aktivní (neaktivní podmínky jsou ignorovány).
10. Uložte filtr kliknutím na “Uložit”.

4.8 Jednotlivé části webové aplikace

4.8.1 Index

V této sekci uživatel má přístup k seznamu konferencí. Zobrazení je možné ovlivnit výběrem některého z filtrů, dále je možné vybrat jen konference, které jsou hlídané (volbou “Libovolný stav”), případně konference, které jsou ve zvoleném stavu.

Počet konferencí zobrazených na jedné stránce se nastavuje v sekci “Moje nastavení”. U nepřihlášeného uživatele toto nastavení provádí administrátor v sekci “Správa – nastavení”.



Obrázek 4.1: Index přihlášeného uživatele

4.8.2 Detaily konference

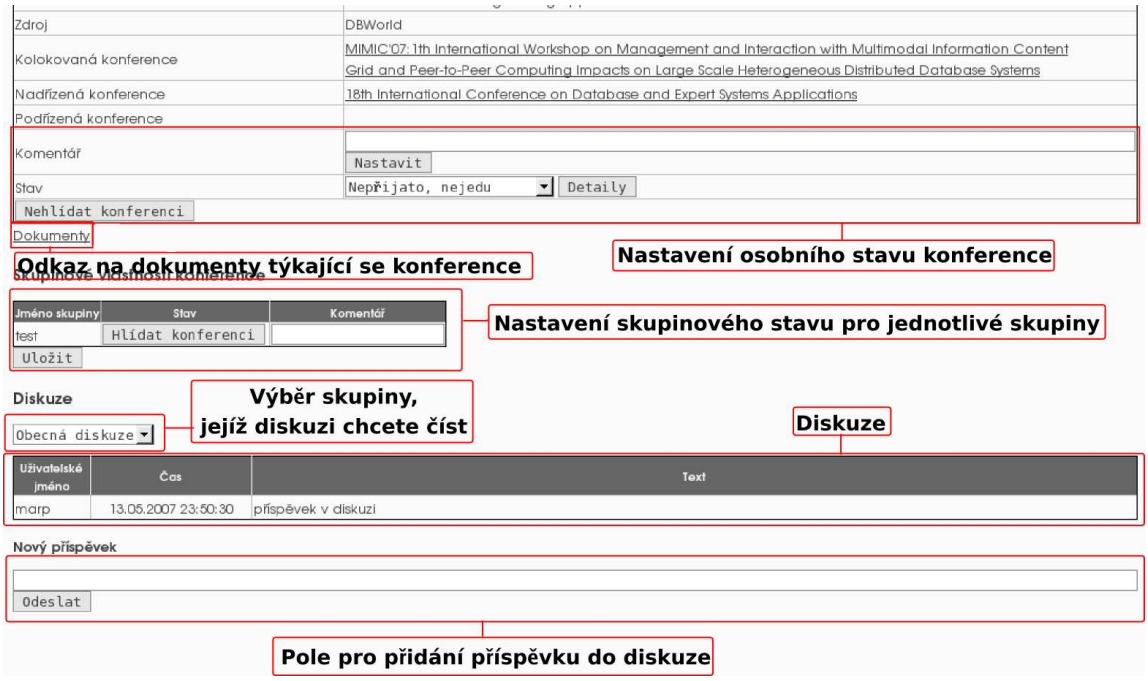
Po kliknutí na zkratku konference v seznamu konferencí v Indexu nebo Editaci konference se zobrazí stránka s detaily konference, uživatelským a skupinovým nastavením události a diskuzí týkající se dané konference. Můžete zde přidat konferenci mezi hlídané konference, můžete změnit její stav a případně si k ní napsat vlastní poznámku. U nastavení stavu je ještě možné nastavit jeho detaily, tzn. naléhavost jednotlivých termínů. Naléhavost se nastavuje čísly od 1 pro nejnižší důležitost až do 4 pro nejvyšší. Zvolená naléhavost určuje barvu, jakou se termín zobrazí v kalendáři. Hodnota 0 znamená, že se daný termín vůbec nezobrazuje. Nejsou-li detaily nastaveny, použije se výchozí nastavení, které může ovlivnit administrátor v sekci "Správa – nastavení".

4.8.3 Moje nastavení

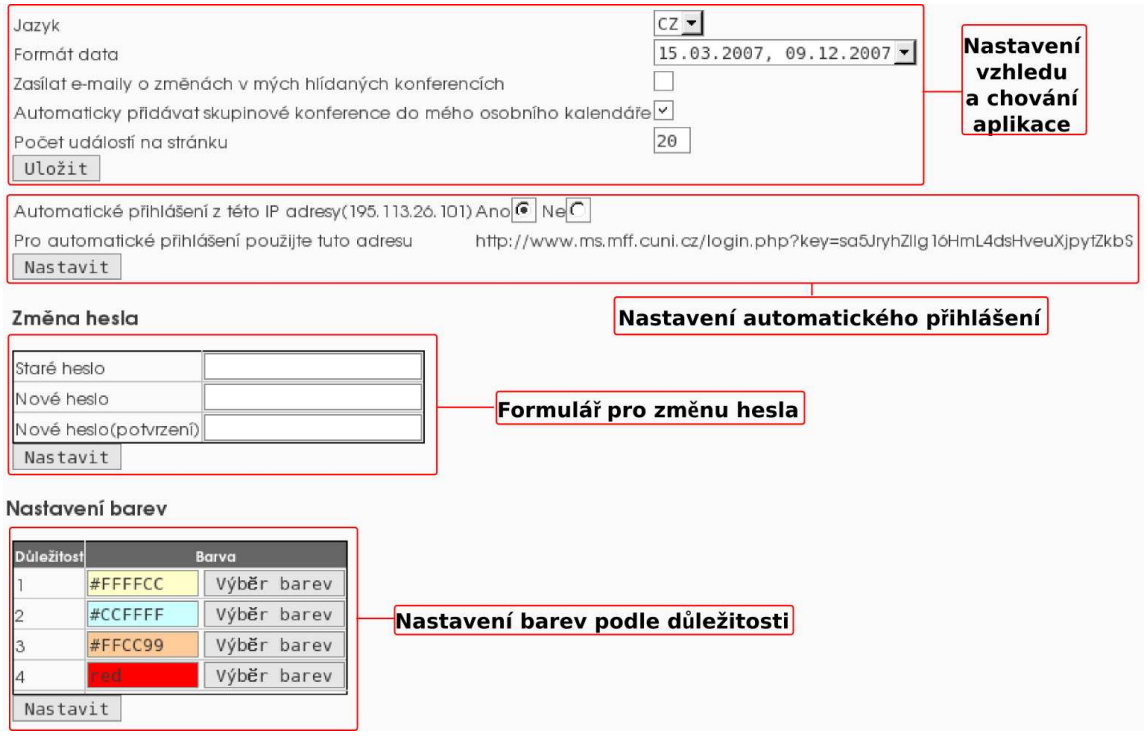
V této sekci můžete měnit své nastavení. Je zde možné změnit jazyk rozhraní, formát data a počet událostí na stránku. Můžete si zde nastavit posílání zpráv o změnách hlídaných konferencí a také to, jestli se skupinové konference vašich skupin mají automaticky přidat i do vašeho osobního kalendáře.

Dále zde můžete nastavit automatické přihlašování z určité IP adresy. Učiníte-li tak, aplikace vám zobrazí odkaz, jehož použitím se můžete kdykoliv později přihlásit do aplikace, aniž byste museli používat heslo. Ověření se pak provede podle jedinečného klíče v odkazu a IP adresy, ze které do aplikace přistupujete. Administrátor může při tomto způsobu přihlašení omezit vaše práva.

V této sekci také můžete změnit obvyklým způsobem své heslo.



Obrázek 4.2: Spodní část stránky s detaily konference



Obrázek 4.3: Stránka s osobním nastavením

Obrázek 4.4: Nastavení filtrů

Poslední volbou je nastavení barev používaných v kalendáři jako pozadí. Tyto barvy se nastavují v závislosti na důležitosti daného termínu. Barvu lze buď vybrat po kliknutí na tlačítko "Výběr barev", nebo ručně zadat jednu z předdefinovaných barev pro web.

4.8.4 Filtry

Tato sekce slouží k vytváření, upravování a mazání filtrů.

Každý filtr má své jméno, které se zobrazuje u Kalendáře a v Indexu. Dále si ke každému filtru můžete poznamenat svůj komentář.

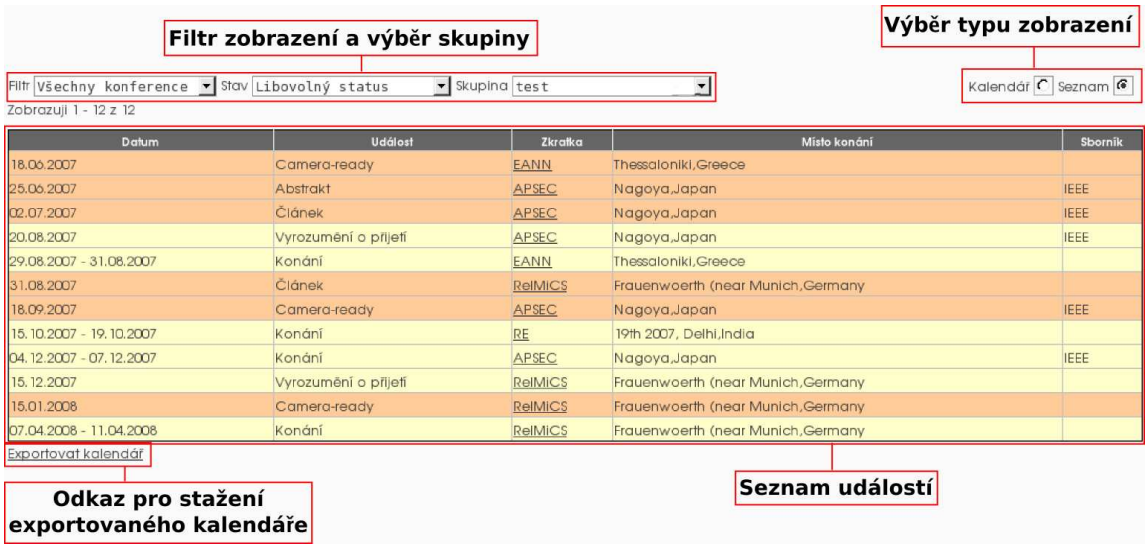
Každý filtr se skládá z libovolného počtu podmínek. Podmínka má tvar "Atribut", "Operátor", "Hodnota". "Atribut" může být libovolný údaj, který aplikace ukládá o konferenci. "Operátor" je jeden z $<$, $<=$, $==$, $>=$, $>$, *obsahuje*, *neobsahuje*, *obsahuje jako slovo*. "Hodnota" je buď datum (u termínů a dat), nebo libovolný textový řetězec. Poslední na řádce je logická spojka, která se použije mezi aktuální a následující podmínkou.

Jako "Atribut" je také možné zvolit filtr. Touto volbou se políčko na zadávání hodnoty změní na seznam vašich filtrů a můžete vybrat libovolný z nich. Tímto způsobem se dají vytvářet komplikovanější filtry z filtrů jednodušších. Také je tak možné vytvářet složité konstrukce, které by jinak vyžadovaly použití závorek.

U každé podmínky je také možnost zaškrtnout, zda má být aktivní. Není-li podmínka aktivní, filtr se chová jako by v něm vůbec nebyla. Toto se hodí chcete-li, například, dočasně nějakou část filtru odstranit.

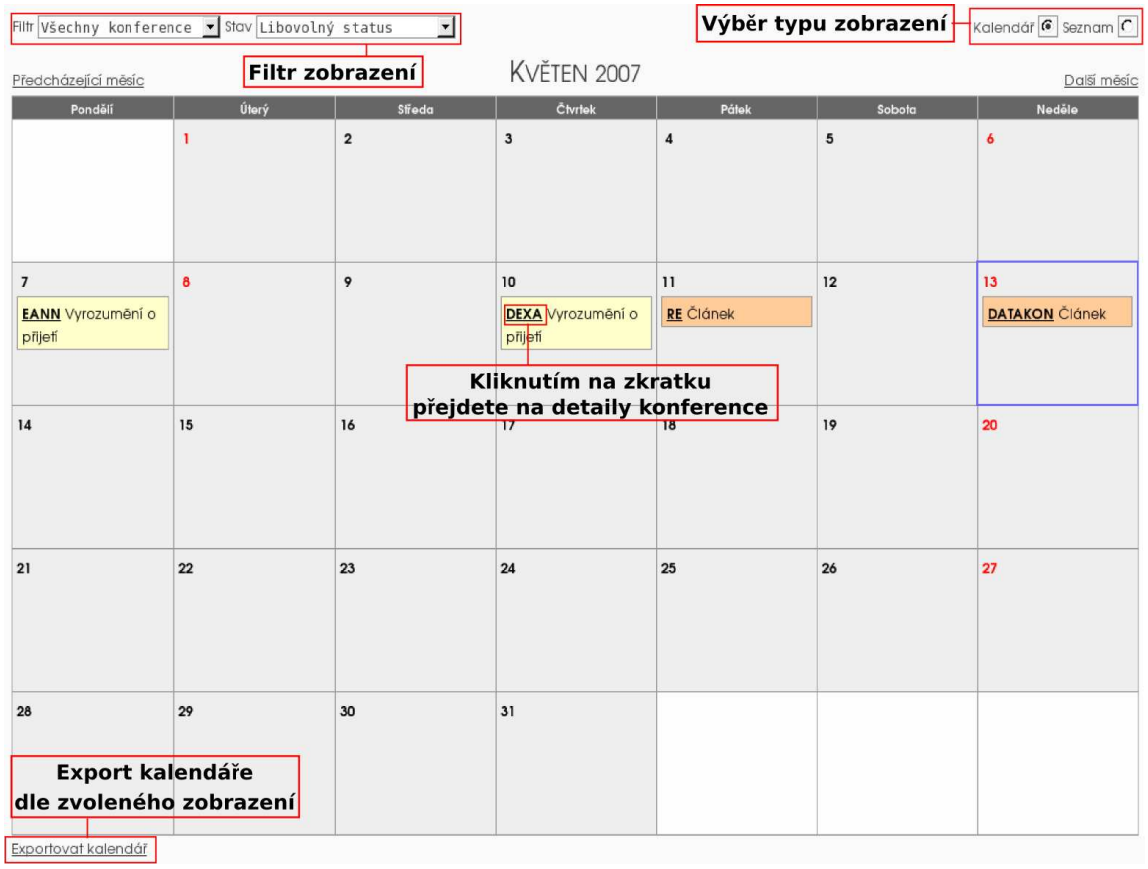
4.8.5 Kalendář

V kalendáři se zobrazují údaje o termínech konferencí. Je možné vybrat mezi zobrazením jako seznam, kde se zobrazuje více údajů, a klasickým kalendářovým zobrazením, kde se zobrazuje jen zkratka události a typ termínu.



Datum	Událost	Zkratka	Místo konání	Sborník
18.06.2007	Camera-ready	EANN	Thessaloniki,Greece	
25.06.2007	Abstrakt	APSEC	Nagoya,Japan	IEEE
02.07.2007	Článek	APSEC	Nagoya,Japan	IEEE
20.08.2007	Vyrozumění o přijetí	APSEC	Nagoya,Japan	IEEE
29.08.2007 - 31.08.2007	Konání	EANN	Thessaloniki,Greece	
31.08.2007	Článek	RelMiCS	Frauenwoerth (near Munich,Germany	
18.09.2007	Camera-ready	APSEC	Nagoya,Japan	IEEE
15.10.2007 - 19.10.2007	Konání	RE	19th 2007, Delhi,India	
04.12.2007 - 07.12.2007	Konání	APSEC	Nagoya,Japan	IEEE
15.12.2007	Vyrozumění o přijetí	RelMiCS	Frauenwoerth (near Munich,Germany	
15.01.2008	Camera-ready	RelMiCS	Frauenwoerth (near Munich,Germany	
07.04.2008 - 11.04.2008	Konání	RelMiCS	Frauenwoerth (near Munich,Germany	

Obrázek 4.5: Seznamové zobrazení skupinového kalendáře



Obrázek 4.6: Osobní kalendář

Obrázek 4.7: Seznam konferencí k editaci, seznam nepotvrzených konferencí vypadá stejně

Zobrazované události je možné filtrovat stejně jako v sekci Index. Zrušíte-li filtrování podle stavu konference, zobrazí se i konference, které nejsou hlídané.

Ve skupinovém kalendáři je navíc volba skupiny, jejíž kalendář chcete zobrazit.

Barvy na pozadí jednotlivých termínů v osobním kalendáři závisí na vámi nastaveném stavu konference a na nastavení barev. Ve skupinovém kalendáři barvy určuje skupinový stav konference.

Kalendář je také možné exportovat do formátu CSV, který je možné importovat například do aplikace Microsoft Outlook. Exportují se vždy všechny termíny podle zvoleného filtru zobrazení, které ještě neproběhly.

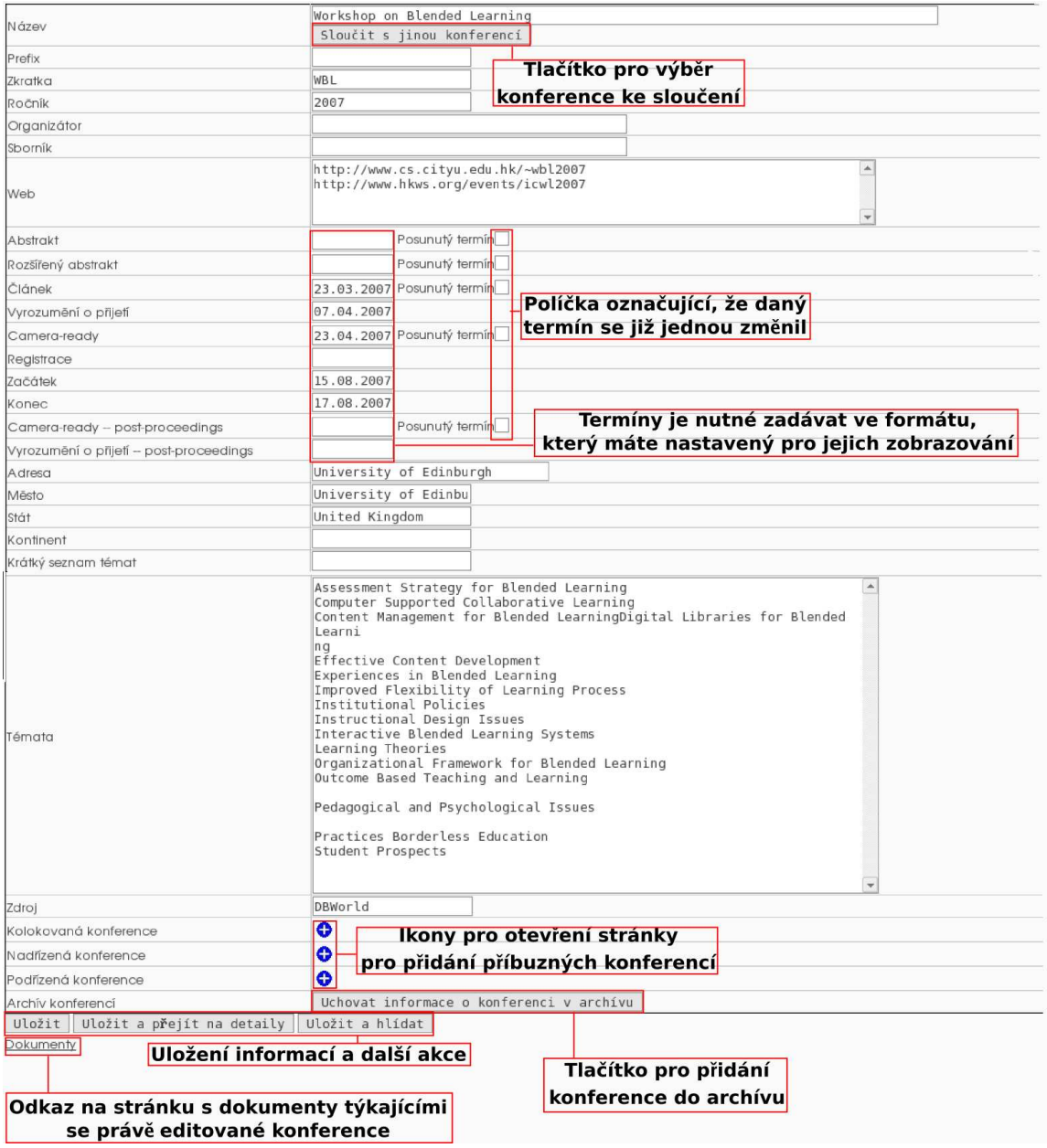
4.8.6 Přidání a editace konference

Přidat konferenci je možné dvěma způsoby: buď ručně v sekci Přidat konferenci, nebo potvrdit informace vydolované z e-mailu v sekci Nepotvrzené konference. Při volbě možnosti ručního přidání můžete ještě vybrat konferenci, která se konala v minulých letech, z archívu (pokud tam byla přidána příslušným tlačítkem). Pokud vyberete nějakou konferenci ze seznamu nepotvrzených konferencí zobrazí se formulář s předvyplněnými automaticky vydolovanými údaji. Zprávu, ze které byly tyto údaje získány, můžete zobrazit kliknutím na "Ukázat e-mail" ve spodní části stránky.

Při všech způsobech přidání vyplňujte datum podle vašeho aktuálního nastavení.

Po ukončení editace uložte konferenci. Kromě pouhého uložení můžete ještě stisknutím příslušného tlačítka přidat konferenci mezi svoje hlídané konference, nebo přejít na stránku s jejími detaily (např. pro přidání do skupinových konferencí).

E-maily v sekci Nepotvrzené konference jsou automaticky přiřazovány ke konferencím. Vyberete-li e-mail, který je přiřazen k již existující události, pole Jméno a Zkratka se vyplní podle již potvrzených dat. Je-li přiřazení chybné můžete jej



Obrázek 4.8: Formulář pro editaci/přidání konference

Obrázek 4.9: Formulář pro sloučení konferencí

změnit kliknutím na "Vybrat jinou konferenci" případně vytvořit novou konferenci kliknutím na "Nová konference". Potvrdíte-li informace o konferenci, která již existuje, informace se sloučí se známými informacemi (tj. chybějící informace se doplní a existující informace se aktualizují – pokud jsou starší než právě potvrzené informace). Jestliže by vám z jakéhokoliv důvodu tento způsob sloučení nevyhovoval, můžete konferenci uložit jako novou a informace sloučit postupem popsaným níže. K tomu můžete s výhodou využít tlačítko "Uložit a přejít na editaci" ve spodní části stránky.

Konference je možné editovat v sekci "Upravit konferenci". Kliknutím na ikonu editace zobrazíte stejný formulář jako při přidávání konference. V tomto formuláři lze změnit libovolné údaje o konferenci.

Konferenci je možné vložit do archívu dříve, než je uložena v aplikaci. To se může hodit například v případě, že přidáváte informace o více příbuzných konferencích. Potom je možné údaje, které jsou společné pro všechny tyto konference, vyplnit a vypněné údaje uložit do archívu. Při přidávání dalších událostí můžete vybrat z archívu předvyplněný formulář a doplnit pouze chybějící informace.

4.8.7 Sloučení konferencí

Při editaci je také možné konferenci sloučit s jinou konferencí. Tlačítko "Sloučit s jinou konferencí" zobrazí stránku, kde můžete vybrat konferenci, se kterou se má slučovat. Po tomto výběru se zobrazí fomulář pro sloučení konferencí. V tomto formuláři můžete zaškrtnutím příslušných políček vybrat, jaké informace mají být ve sloučené konferenci.

4.9 Přihlášení jen pro čtení

Kalendář libovolného uživatele je možné bez přihlášení zobrazit přístupem na adresu `login.php?uziv=<uzivatel>&view=<zobrazeni>&position=<pozice>`. Parametry "zobrazeni" a "pozice" jsou nepovinné. "zobrazeni" je buď "calendar" pro

kalendářové zobrazení nebo “list” pro zobrazení jako seznam. “pozice” je u kalendářového zobrazení měsíc a rok, který se má zobrazit ve tvaru rrrr-m (např. 2007-5). U zobrazení jako seznam je to řádek, od kterého má zobrazení začínat. Nejsou-li tyto dva parametry zadány, zobrazí se aktuální měsíc v kalendářovém zobrazení.

4.10 Správa aplikace

4.10.1 Instalace

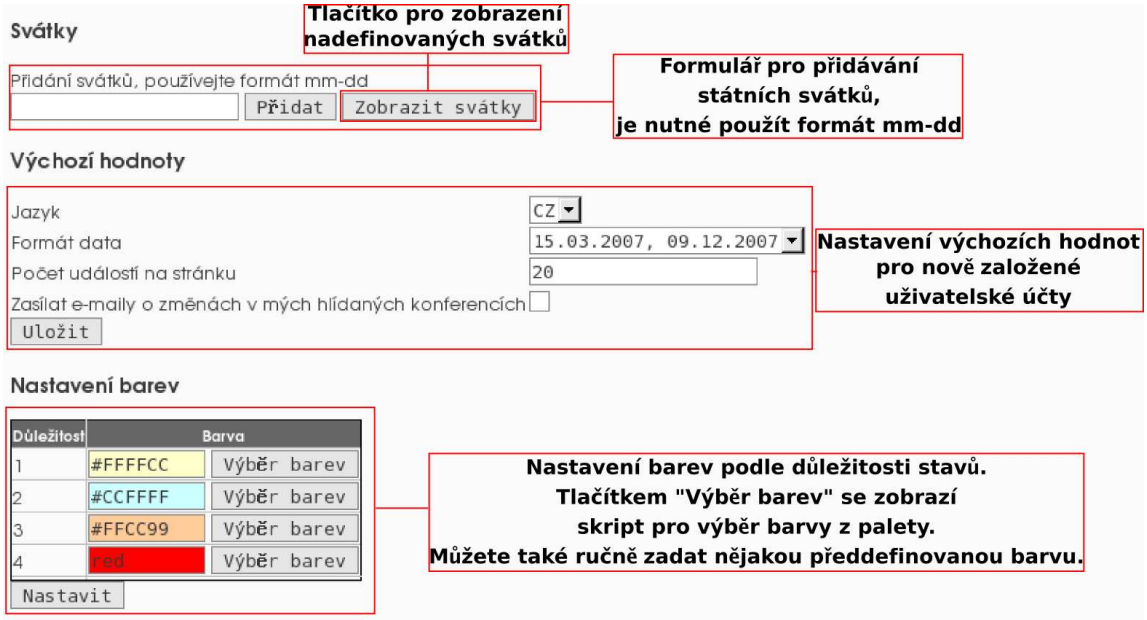
1. Rozbalte balíček `www.tar.gz` do libovolného adresáře na webovém serveru.
2. V prolížeči otevřete stránku `http://server/cesta/install.php`.
3. V zobrazeném formuláři vyplňte údaje o databázovém serveru.
4. Chcete-li nainstalovat i klíčová slova, zaškrtněte příslušné políčko.
5. Kliknutím na “Nainstalovat” se vytvoří tabulky v databázi a účet “admin” s heslem “admin”.
6. Pomocí zobrazeného odkazu stáhněte a uložte soubor s nastavením části pro dolování informací.
7. Rozbalte balíček `mailmine.tar.gz` do libovolného adresáře.
8. Pomocí příkazu `make` zkompilujte aplikaci.
9. Soubor `mailmine` a soubor s nastavením zkopírujte do libovolného adresáře.
10. Nastavte aplikaci jako filtr příchozích e-mailů v souboru `.forward`.

Aplikace vyžaduje, aby zpracovávané e-maily byly v prostém textovém formátu. Pro převod zpráv do tohoto formátu můžete použít skripty z adresáře `scripts`. Nejjednodušším způsobem je použití skriptu `mail_to_text.sh`, který dokáže zpracovat jak textové e-maily, tak e-maily ve formátu HTML a převést je do prostého textu. V některých případech ale může být efektivnější napsání vlastních skripů.

4.10.2 Uživatelská práva

Systém obsahuje propracovaný systém uživatelských práv. Práva může měnit administrátor, nebo uživatel, který k tomu má oprávnění. Práva se určují pro každého uživatele zvlášť. Je možné nastavit výchozí hodnoty. Práva uživatelů navíc může ovlivňovat i doména, ze které do aplikace přistupují.

Je možné nastavit následující práva: Editovat konference, Kontrolovat vydolované údaje, Přidávat konference, Mazat konference, Vytvářet skupiny, Vytvářet uživatele, Mazat uživatele, Přidávat uživatele do skupiny, Měnit nastavení dolování, Psát do diskuze, Měnit uživatelská práva, Mazat skupiny, Měnit nastavení a Měnit bezpečnostní nastavení.



Obrázek 4.10: Nastavení

Navíc každá skupina uživatelů má svého správce, který může měnit nastavení této skupiny, i když obecně právo měnit skupiny nemá.

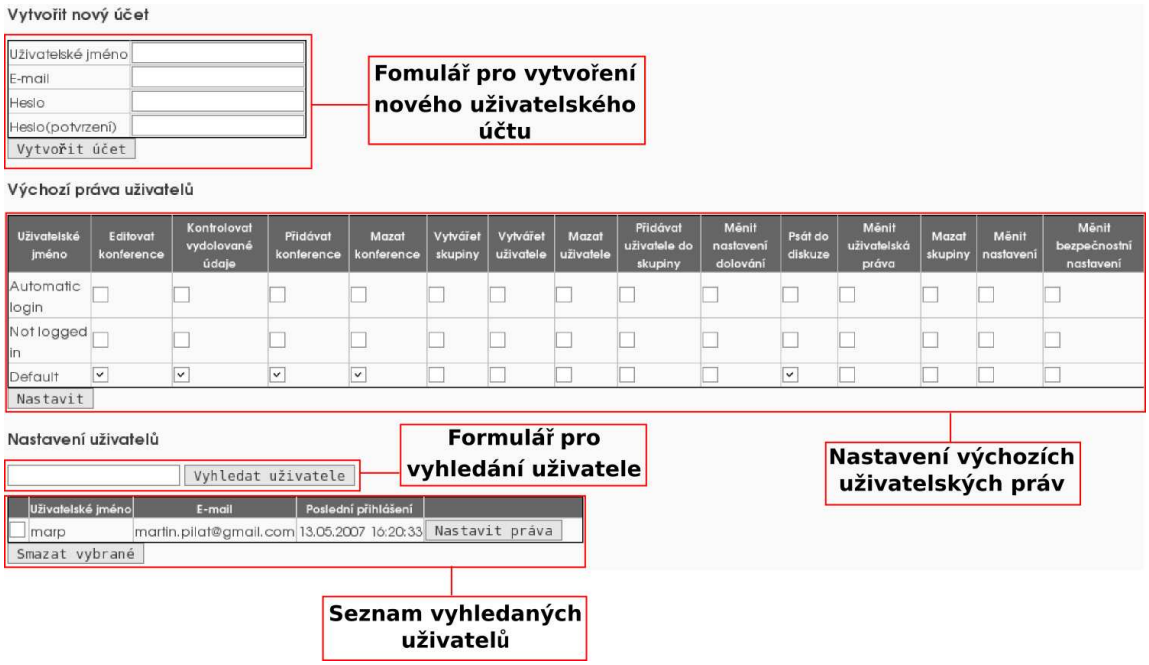
4.10.3 Jak změnit práva uživatele?

1. Vyberte “Správa – uživatelé” v Menu.
2. S části “Nastavení uživatelů” vyhledejte uživatele, jehož práva chcete změnit.
3. Klikněte na tlačítko “Nastavit práva” na řádce se jménem uživatele.
4. V zobrazeném formuláři nastavte práva uživatele a uložte je.

4.10.4 Správa – nastavení

V této sekci je možné měnit výchozí nastavení pro nepřihlášeného uživatele a pro nově zaregistrovaného uživatele. Je možné nastavit stejné údaje, jaké si může nastavit každý uživatel v sekci “Moje nastavení”.

Dále je v této sekci možné přidávat stavy konference včetně nastavení důležitosti jednotlivých termínů pro konference v daném stavu. Jméno stavu je nutné zadat ve všech jazycích, v jakých je aplikace provozována. Není-li v nějakém jazyce jméno stavu zadáno, stav se při použití tohoto jazyka nezobrazuje. Stavy lze také mazat, nicméně smazání stavu znamená jeho změnu pro všechny uživatele, kteří jej používali. Mazání stavu by se proto mělo používat jen ve výjimečných případech. Lepší řešení je přejmenování případně úprava existujícího podobného stavu.



Obrázek 4.11: Nastavení uživatelů

Také je zde možné přidávat státní svátky. Zadává se pouze měsíc a rok a to ve formátu mm-dd (tedy např. 09-28). Po kliknutí na tlačítko zobrazit svátky se zobrazí seznam aktuálně nadefinovaných svátků. Svátek potom lze také smazat.

4.10.5 Správa – uživatelé

Zde můžete vytvořit nový uživatelský účet za stejných podmínek jako při jeho samostatném vytváření uživateli při registraci.

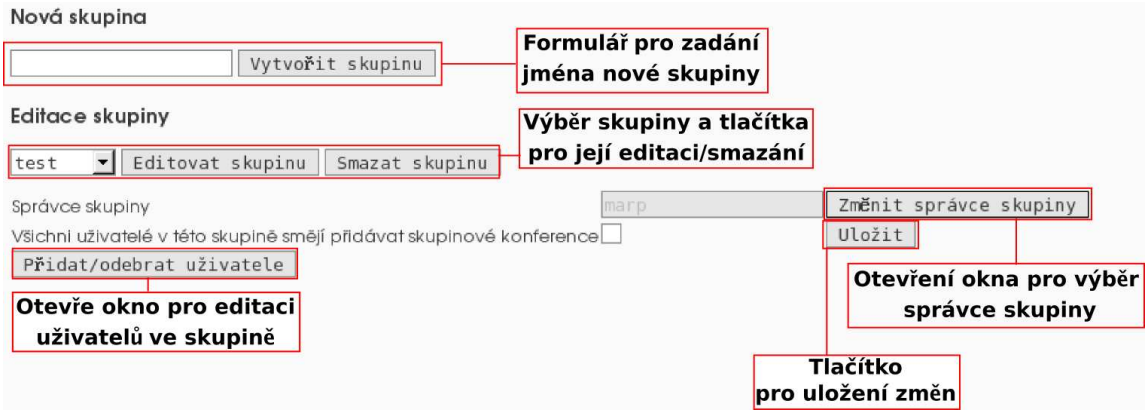
Také se zde dají nastavovat práva pro automatické přihlášení (tj. přihlášení bez nutnosti zadat uživatelské jméno a heslo), práva nepřihlášeného uživatele a práva, která získá uživatel po registraci.

Při automatickém přihlášení se práva uživatele určí jako průnik jeho práv a práv, která jsou nastavena pro tento způsob přihlášení.

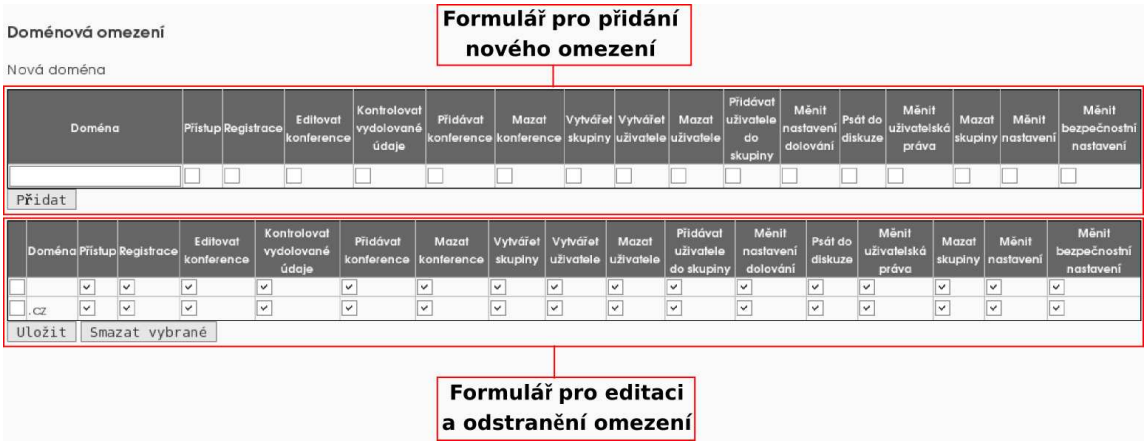
Ve spodní části stránky je možné vyhledat registrované uživatele. Po vyhledání se zobrazí seznam uživatelů s jejich e-mailovou adresou a časem posledního přihlášení. U každého uživatele je tlačítko, které zobrazí formulář pro editaci jeho práv. Uživatelské účty je zde možné i odstranit.

4.10.6 Správa – skupiny

Na tomto místě se dají vytvářet a upravovat skupiny. Každá skupina má svého správce, který může měnit všechna nastavení skupiny (tato nastavení může měnit také každý uživatel, který k tomu má práva). Správce může navíc přidávat skupinové konference u skupin, které nemají povoleno přidávání těchto konferencí všemi svými



Obrázek 4.12: Nastavení skupin



Obrázek 4.13: Nastavení zabezpečení

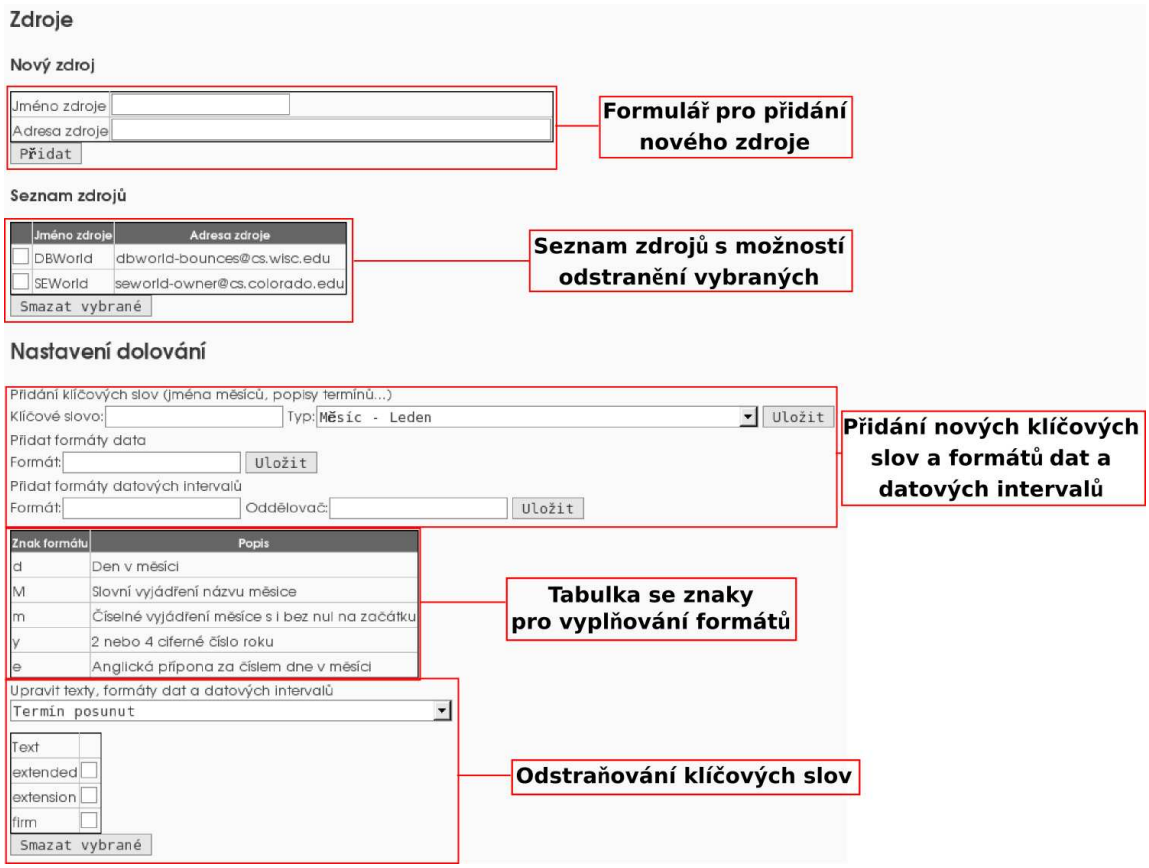
členy. Dole na stránce je tlačítko pro přidání/odebrání uživatelů ze skupiny. Po jeho stisknutí se zobrazí stránka, na které lze vyhledat uživatele dle jména a skupiny, ve které jsou, a přidat nebo odebrat je z právě editované skupiny.

4.10.7 Správa – zabezpečení

Tato sekce slouží k nastavování práv podle domény. Kromě všech uživatelských práv je zde ještě možnost zakázat přístup do aplikace případně registraci v závislosti na doméně, ze které uživatel do aplikace přistupuje.

Při přístupu do aplikace se použije nejspecifičtější doména, pro kterou je v systému nastavené nějaké omezení. To umožňuje například mít globálně povolenou registraci (z prázdné domény) a zakázat registraci pro uživatele z domény `.cz`, ale naopak ji zase povolit pro uživatele z domény `mff.cuni.cz`.

Uživatel získá při přihlášení práva, která jsou průnikem jeho práv a doménových práv, případně ještě práv v závislosti na typu přihlášení.



Obrázek 4.14: Nastavení dolování

Vzhledem k tomu, že je možné zakázat přístup do aplikace z libovolné domény, doporučuje se, aby alespoň dva správci měli své domény v seznamu a měli u nich povolena všechna práva, aby v případě chyby bylo možné ji snadno napravit. Také se doporučuje, aby v seznamu byla doména `localhost.localdomain` pro lokální přístup ze serveru.

4.10.8 Správa – dolování

Nastavení v této sekci ovlivňuje automatické dolování informací z e-mailů.

Nastavují se tu povolené zdroje. Zprávy, které nemají v jednom z polí From: nebo Sender: jednu z povolených adres, budou ignorovány.

Dále se tu nastavují klíčová slova a jejich významy, např. že "Paper submission deadline" znamená termín odeslání článku. Při nastavování těchto textových vyjádření je vhodné používat co nejdelší řetězce, aby dolování bylo co možná nejpřesnější. Pokuste se vyhýbat přidávání vyjádření, která jsou nejednoznačná.

Dále lze přidávat formáty dat a datových intervalů, které umí aplikace zpracovat. Při zadávání formátů používejte znaky, které najdete v tabulce přímo na stránce. Při zadávání datových intervalů je nutné zadat i oddělovač jednotlivých částí. Oddělovač může být libovolný znak, nebo řetězec.

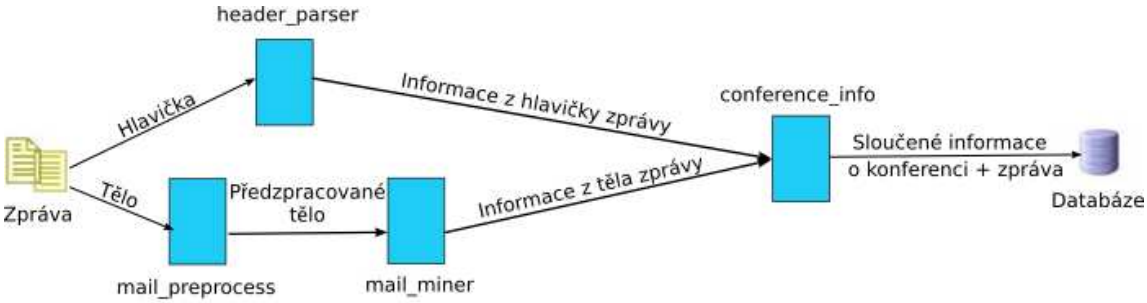
Všechny údaje zadané na této stránce lze smazat výběrem příslušného typu ve spodní části stránky. Po tomto výběru se zobrazí seznam všech vyjádření tohoto typu a máte možnost libovolný z nich smazat zaškrtnutím příslušného políčka a kliknutím na tlačítko “Smazat vybrané”.

Kapitola 5

Programátorská dokumentace

5.1 Část pro dolování informací z e-mailů

Základní schéma zpracování příchozích zpráv můžete vidět na obrázku 5.1.



Obrázek 5.1: Schéma zpracování příchozích zpráv

Účelem této části je dolování informací z příchozích zpráv. Vstupem je e-mailová zpráva včetně hlavičky. Tělo zprávy musí být v prostém textovém formátu. Na výstup je vypsána zpráva s vydolovanými údaji ve speciálních polích hlavičky. Vydolované údaje jsou uloženy do databáze.

Tato část programu je napsána v programovacím jazyce C++. Jejím základem jsou třídy `mail_preprocess` a `mail_miner`.

Obě tyto třídy jsou poděděny od třídy `aho_corasick`, která implementuje algoritmus Aho a Corasickové pro vyhledávání v textu [3]. Tento algoritmus je použit při vyhledávání klíčových slov. Mezi její základní metody patří `add_words` pro přidání klíčových slov a `search` pro vyhledávání. Po každém prohledaném znaku se volá metoda `report`, která oznamuje nalezení nějakého z klíčových slov a v potomcích této třídy provádí vlastní úpravy e-mailu, nebo dolování informací z něj.

Třída `mail_preprocess` provádí předzpracování těla e-mailu. Jako parametry dostane názvy měsíců a jejich pořadí v roce. Tyto měsíce potom vyhledává v textu zprávy a pokud nějaký najde, nahradí jej jedním speciálním znakem. Zároveň do své proměnné ukládá významy znaků na jednotlivých pozicích. Na konci své práce má

vytvořené předzpracované tělo e-mailu a právě zmíněný seznam. Základní metody jsou `add_words`, pro přidání klíčových slov a jejich významů, `search` pro spuštění vyhledávání, `get_months`, která vrací pozice a čísla nalezených měsíců a `get_ppmail` vracející předzpracované tělo e-mailu.

Třída `mail_miner` využívá výsledků třídy `mail_preprocess` a klíčových slov načtených z databáze (včetně jejich významů) a provádí vlastní dolování informací z textu. Při tom používá heuristiky popsané v kapitole "Dolování informací". Nakonec má vytvořenou strukturu `conference_info` se všemi vydolovanými informacemi o konferenci. Základní metody jsou `add_words`, která opět slouží pro přidání klíčových slov a jejich významů, `add_date_intervals` pro přidání datových intervalů, `set_months` pro přidání vyhledaných výskytů měsíců vrácených třídou `mail_preprocess`, `search` pro spuštění vyhledávání a `get_conf_info`, která vrací vyhledané informace o konferenci ve struktuře `conference_info`.

Tyto informace se ještě doplní o informace získané z hlavičky e-mailu a uloží se do databáze.

Struktura `conference_info` obsahuje všechny informace o konferenci jako textové řetězce ve formátu vhodném pro uložení do databáze. Tím odpadá potřeba je před uložením konvertovat.

Samotná hlavička e-mailu je zpracovávána třídou `header_parser`, která umí vracet jednotlivá pole z této hlavičky podle jejich jména. V konstruktoru si načte ze zadaného vstupního proudu hlavičku e-mailu, a potom prostřednictvím metod `get_header_field` a `get_arg_from_field` vrací hodnoty jednotlivých polí, respektive argumentů z těchto polí.

Podrobný přehled všech těchto tříd vygenerovaný systémem Doxygen najdete v adresáři `doc/developer` na přiloženém CD.

5.2 Webová aplikace

Tato část aplikace slouží k zobrazení, zadávání a úpravě informací o konferencích, nastavení obou částí aplikace a správě uživatelů a skupin.

Webová část projektu je napsána v jazyce PHP (testováno na verzi 5.1 a 5.2), informace jsou uloženy v databázi MySQL [1] (testováno na verzi 4.0 a 4.1). Zabezpečení je provedeno pomocí technologie sessions.

Všechny skripty napřed zkontrolují, zda má uživatel potřebná oprávnění pro činnost, kterou chce provádět, a pokud tato oprávnění jsou, tuto činnost provedou.

5.2.1 Důležité skripty

Všechna ID zde uvedená jsou celá kladná čísla.

conf_admin.php správa konferencí. Proměnná `$_GET['uncheck']` určuje, zda se mají zobrazovat potvrzené (je-li její hodnota 0) nebo nepotvrzené (1) konference. Proměnná `$_GET['c_id']` udává ID konference, která se má editovat. Jedná-li se o nepotvrzenou konferenci, určuje toto proměnná `$_GET['m_id']`, neboť informace

o těchto konferencích jsou ukládány podle e-mailů, ze kterých byly vydolovány. Není-li nastavena žádná z těchto proměnných, zobrazí se seznam potvrzených konferencí. `$_GET['cinfo']` je pole obsahující informace o konferenci, které byly vyplněné ve formuláři.

admin.php správa stránek a nastavení dolování. Proměnnou `$_GET['what']` se určuje, jaká část administrace se má zobrazit. Možné hodnoty jsou: *users* pro správu uživatelů, *groups* pro správu skupin, *mining* pro správu dolování, *security* pro nastavení zabezpečení a *other* pro nastavení aplikace. Je-li hodnota tohoto parametru jiná, skript vypíše chybovou hlášku a žádná akce se neprovede.

calendar.php zobrazení kalendáře. `$_GET['view']` udává typ zobrazení, pokud je její hodnota *calendar*, použije se kalendářové zobrazení, je-li tato hodnota *list*, zobrazí se seznam. `$_GET['month']` je měsíc, který se má zobrazit, ve tvaru rrrr-d u kalendářového zobrazení a `$_GET['offset']` je číslo prvního řádku, který se má zobrazit, v seznamovém zobrazení.

details.php zobrazení detailů konference, diskuze, nastavení osobního a skupinového stavu konference. ID konference jejíž detaily se mají zobrazit je v proměnné `$_GET['id']`.

filters_conf.php nastavení filtrů. Slouží k nastavení a editaci filtru. Před uložením filtru ověřuje, jestli všechny jeho části jsou z platných rozsahů, tzn. omezené možným výběrem na stránce.

index.php výchozí stránka. Zobrazuje seznam konferencí podle zvoleného filtru. Jméno filtru se předává v proměnné `$_GET['filter']`. Neexistuje-li filtr s takovým jménem, zobrazí se všechny konference. Směr řazení a parametr, podle kterého se mají konference řadit, se předávají v proměnných `$_GET['dir']` (možné hodnoty jsou *short*, *c_date_start* a *c_date_end* pro řazení podle zkratky, jména, začátku a konce konference) a `$_GET['order']`(*ASC* pro vzestupné řazení, *DESC* pro sestupné řazení). Nejsou-li tyto hodnoty uvedeny, nebo jsou jiné, než zde napsané, probíhá řazení vzestupně podle jména konference. Číslo stránky je v proměnné `$_GET['page']`.

login.php přihlašování. Skript dostane přihlašovací údaje uživatele v proměnných `$_GET['username']` a `$_GET['pass']`. Ověří je v databázi a provede přihlášení. Dostane-li proměnnou `$_GET['key']`, provádí se automatické přihlášení podle klíče v ní uloženého a IP adresy. Dostane-li proměnnou `$_GET['uziv']`, provede přihlášení jen pro čtení pro uživatele, jehož jméno je v této proměnné uloženo.

register.php registrace uživatele. Skript dostane údaje o uživateli v proměnných `$_GET['username']`, `$_GET['pass']`, `$_GET['pass2']` a `$_GET['email']`. Ověří údaje, zjistí, zda požadované uživatelské jméno již neexistuje, a vytvoří účet.

user_pref.php nastavení uživatele. Používá skript `js_color_picker_v2` ze stránek `www.dhtmlgoodies.com` pro výběr barev.

function.php funkce společné pro všechny skripty.

5.2.2 Významné proměnné

Popis významných proměnných, které se ve skriptech často vyskytují, najdete v Tabulce 5.1.

Proměnná	Popis
$TEXT	pole obsahující textové řetězce, které se vypisují na stránkách. Nastavuje se ve skriptech lang_XX.php , kde XX je zkratka jazyka, ve kterém se mají stránky zobrazovat.
$SESSION['u_id']	ID přihlášeného uživatele, 0 pokud uživatel není přihlášen
$SESSION['ro_log']	1, je-li uživatel přihlášen jen pro čtení
$SESSION['u_set']	pole s nastavením právě přihlášeného uživatele
$SESSION['u_right']	pole s oprávněními právě nastaveného uživatele
$SESSION['info']	pole s informacemi o provedených akcích
$SESSION['errors']	pole s chybami, které vznikly při vykonávání skriptu

Tabulka 5.1: Významné proměnné

5.2.3 Přidání jazykové verze

Přidání jazykové verze obnáší přeložení jednoho z existujících souborů `lang_XX.php` do nového jazyka. Dále je nutné upravit funkci pro vypisování hlavičky stránky tak, aby nabízela možnost přepnutí na novou jazykovou verzi a upravit soubor `lang_common.php` – doplnit nový jazyk do pole existujících jazyků. Nakonec je vhodné v nastavení přidat popisky pro stavy v nově vytvořeném jazyce, aby tyto stavy bylo možné používat.

5.3 Databáze

Pro uložení dat je použita MySQL databáze, pro přesný SQL popis tabulek viz soubor `tables.mysql`, který se nachází na přiloženém CD v adresáři `www`. Některé z tabulek používají pro udržení integrity a zjednodušení práce s databází tzv. cizí klíče, a proto je potřeba použít databázový engine, který je podporuje. Tím je v současné době v MySQL pouze InnoDB. Některé tabulky proto tento engine používají.

5.4 Rozhraní aplikace pro dolování informací

Této části aplikace je možné předat informace o konferenci prostřednictvím speciálních polí v hlavičce zprávy. Tato pole se použijí, pokud daná informace není ve

Jméno pole	Popis
x-mailmine-abstract	termín odeslání abstraktu
x-mailmine-cam-read	termín odeslání finální verze článku
x-mailmine-end	konec konání konference
x-mailmine-start	začátek konání konference
x-mailmine-ex-abstract	termín odeslání rozšířeného abstraktu
x-mailmine-address	adresa konání konference
x-mailmine-country	stát konání konference
x-mailmine-continent	kontinent, kde se konference koná
x-mailmine-city	město, kde se konference koná
x-mailmine-abbrev	zkratka konference
x-mailmine-name	jméno konference
x-mailmine-notification	datum vyrozumění o přijetí
x-mailmine-notification-pp	datum vyrozumění o přijetí pro post-proceedings
x-mailmine-organizer	organizátor konference
x-mailmine-paper	termín odeslání článku
x-mailmine-publisher	vydavatel článku, sborník
x-mailmine-registration	termín registrace
x-mailmine-url	odkazy na webové stránky konference, oddělené středníkem
x-mailmine-mid	m_id, se kterým byla konference uložena

Tabulka 5.2: Významy jednotlivých speciálních polí

zprávě nalezena. Naopak aplikace vypisuje na standardní výstup zprávu s těmito poli v hlavičce a umožňuje tak zpracování informací dalšími skripty nebo aplikacemi.

Formát polí odpovídá příslušnému RFC pro e-mailové zprávy [10].

Významy jednotlivých polí najdete v tabulce 5.2. Data termínů je potřeba předávat (a aplikace je vypisuje) ve formátu yyyymmdd.

Kapitola 6

Závěr

6.1 Problémy při řešení

Během tvorby systému bylo třeba učinit několik rozhodnutí.

Slučování automaticky vydolovaných informací – Informace se neslučují, neboť se domníváme, že kontrolování sloučených informací by bylo pracnější. Bylo by potřeba kontrolovat několik zpráv současně a porovnávat je s vydolovanými údaji.

Ovlivnění kalendáře jiných uživatelů – Přílišná svoboda by v tomto případě mohla znamenat, že se uživatelům budou měnit stavy jinak, než to sami uživatelé chtějí. Pokud by například byl uživatel ve více skupinách, které pracují na článku na stejnou konferenci, mohlo by nastavení jedné skupiny ovlivnit nastavení všech jejích členů, ačkoliv některý z nich mohl chtít ponechat stav, který je spojený s jinou skupinou. Proto je možnost ovlivnění kalendáře jiným uživatelům omezena na automatické přidání konference mezi hlídané v případě, že je přidána mezi skupinové konference nějaké skupiny, jejímž je členem.

Vylepšení filtrů konferencí – Bylo potřeba přidat podporu složitějších podmínek a logických operátorů. To s sebou přineslo problém uzávorkováni těchto podmínek. Toto bylo vyřešeno tím, že je možné vytvářet vnořené filtry. Vnitřní filtr je potom uzávorkován v rámci filtru vnějšího.

6.2 Nápady na zlepšení

Domníváme se, že aplikaci by šlo dále vylepšit. Některá vylepšení zde uvádíme.

Přidání oborů konference – systém by mohl obsahovat konference z více oborů a nerušilo by to uživatele, kteří o toto nemají zájem.

Kontrola počtu termínů během týdne, měsíce... – aby uživatel věděl, zda stihne udělat vše, co si naplánoval.

RSS kanál – uživateli by se v něm mohly zobrazovat informace o změnách v jím hlídaných konferencích.

Dolování informací – přidání dalších klíčových slov a vyladění stávajících heuristik by mohlo přinést zlepšení výsledků. Také by mohlo by být zajímavé zkusit začlenit některé z pokročilejších metod pro zpracování přirozeného jazyka.

Uživatelské rozhraní – vhodné rozvržení rozhraní a nastavení funkcí vyžaduje více času; mnoho problémů se projeví až po delším užívání; jiné problémy mohou vyvstat při používání větším počtem uživatelů.

Přidávání dalších událostí – např. termíny pro odeslání článků do časopisů, nebo osobní povinnosti uživatele. Aplikace by se stala obecnějším diářem.

6.3 Současný stav

Podařilo se vytvořit aplikaci, která velmi zjednodušuje správu informací o konferencích. Poloautomatické zpracování šetří mnoho času při vyhledávání a zpracování těchto informací, stejně jako spolupráce uživatelů na jejich upravování, potvrzování a doplňování.

Aplikace je v současnosti provozována na adrese [15] a obsahuje potvrzené informace o více než 230 konferencích z oborů softwarového inženýrství, databází, umělé inteligence a sémantického webu.

Seznam kalendářů konferencí

- http://dbms.uni-muenster.de/menu.php3?item=confs
- http://www.cs.wisc.edu/dbworld/browse.html
- http://www.netlib.org/confdb/Conferences.html
- http://www.cs.vu.nl/~gpierre/conf-cal/
- http://wume.cse.lehigh.edu/conferences.html
- http://campus.acm.org/calendar/
- http://www.ieee.org/conferencesearch/
- http://serl.cs.colorado.edu/~serl/seworld/
- http://www.computer.org/portal/site/ieeecs/
- http://csrc.nist.gov/events/index.html
- http://www.ece.ucsb.edu/Faculty/Banerjee/conferences.htm
- http://www.ams.org/mathcal/
- http://cnscenter.future.co.kr/menu/conference04.html
- http://www.kmining.com/info_conferences.html
- http://dsrg.mff.cuni.cz/
- http://confcal.vrvis.at/index.php

Literatura

[1] MySQL AB :: MySQL 3.23, 4.0, 4.1 reference manual. http://dev.mysql.com/doc/refman/4.1/en/.

[2] PHP: Hypertext preprocessor. http://cz2.php.net/.

[3] Alfred V. Aho and Margaret J. Corasick. Efficient string matching: an aid to bibliographic search. *Commun. ACM*, 18(6):333–340, 1975.

[4] Magnus Almgren and Jenny Berglund. Information extraction of seminar information, 2002. http://nlp.stanford.edu/courses/cs224n/2000/berglund/report.pdf.

[5] Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, and Anand Swaminathan. Mining email social networks. In *MSR '06: Proceedings of the 2006 international workshop on Mining software repositories*, pages 137–143, New York, NY, USA, 2006. ACM Press.

[6] Julie A. Black and Nisheeth Ranjan. Automated event extraction from email, 2004. http://nlp.stanford.edu/courses/cs224n/2004/jblack-final-report.pdf.

[7] Malcolm Corney, Olivier Y. de Vel, Alison Anderson, and George M. Mohay. Gender-preferential text mining of e-mail discourse. In *ACSAC*, pages 282–292, 2002.

[8] Cypress Technologies. Message parse. http://www.cypressnet.com/Products/msgparse/msgparse.htm.

[9] Dbworld. https://lists.cs.wisc.edu/mailman/listinfo/dbworld.

[10] N. Freed and N. Borenstein. Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies. RFC 2045 (Draft Standard), November 1996. Updated by RFCs 2184, 2231.

[11] Svetlana Kiritchenko and Stan Matwin. Email classification with co-training. In *CASCON*, page 8, 2001.

[12] Michal Podzimek. Správa termínů konferencí a workshopů. Bakalářská práce, Univerzita Karlova, Matematicko-fyzikální fakulta, Praha, 2006.

[13] Seworld. http://serl.cs.colorado.edu/~serl/seworld/.

[14] Tweak Marketing. Advanced email parser, 2007. http://www.tweakmarketing.com/products/aep/index.php.

[15] Domovská stránka aplikace. http://www.ms.mff.cuni.cz/~pilam4bm.

Seznam obrázků

Seznam tabulek