VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

DATAMINING Z JABBERU

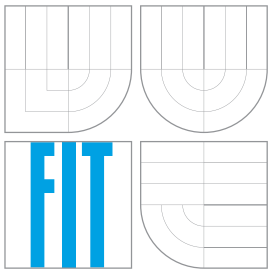
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

JAROSLAV SENDLER

BRNO 2011

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

DATAMINING Z JABBERU

DATAMINING FROM JABBERU

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — JAROSLAV SENDLER
AUTHOR

VEDOUCÍ PRÁCE — Ing. JOZEF MLÍCH
SUPERVISOR

BRNO 2011

Abstrakt

Výtah (abstrakt) práce v českém jazyce.

Abstract

Výtah (abstrakt) práce v anglickém jazyce.

Klíčová slova

Klíčová slova v českém jazyce.

Keywords

Klíčová slova v anglickém jazyce.

Citace

Jaroslav Sendler: Datamining z jabberu, bakalářská práce, Brno, FIT VUT v Brně, 2011

Datamining z jabberu

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana ...

.....
Jaroslav Sendler
15. prosince 2010

Poděkování

Zde je možné uvést poděkování vedoucímu práce a těm, kteří poskytli odbornou pomoc.

Obsah

Kapitola 1

Úvod

Kapitola 2

XMPP

Pro usnadnění a lepší pochopení budou v následující kapitole rozebrány základní stavební kameny protokolu Extensible Messaging and Presence Protocol (XMPP). Samotný protokol je datován do roku 2004 (březen), kdy na něj byl přejmenován jabber. Původní projekt jabber byl vytvořen roku 1998 autorem Jeremie Miller, jež ho založil na popud nesvobodných uzavřených IM služeb. Měl mít tři základní vlastnosti -jednoduchost a srozumitelnost pro implementaci, jednoduše rozšiřitelný a otevřený. Základní vlastnosti a výhody klientů a serverů budou popsány níže. Roku 1999, 4.ledna byl vytvořen první server se jménem jabber. Komunita vývojářů se chopila iniciativy a napsala klienty pro různé platformy (Linux, Macintosh, Windows), kteří dokázali se serverem komunikovat. Roku 2004 byl přidán mezi RFC (request of comments - žádost o komentáře) dokumenty. Základní normy jsou RFC 3920 (obecná specifikace protokolu) a RFC 3921 (samotný instant messaging a zobrazení stavu). Další zdokumentovaná rozšíření jsou vydávána v podobě tzv. XEP (XMPP Extension Protocol) dokumentů, starším jménem JEP (Jabber Enhancement Proposal). Dnešní počet těchto norem se blíží k číslu 300. Každý XEP obsahuje status, stav vývoje (schválení), ve kterém se zrovna nachází. Jako bezpečnostní prvky jsou zde podporovány SASL, TLS a GPG. XMPP protokol je postaven na obecném značkovacím jazyce XML, proto vlastnosti popsané v kapitole 2.2 na straně 6 platí i pro tento protokol.

Jabber

Dnes známý jako komunikační platforma založená na protokolu XMPP. Vzdáleně jej lze přirovnat k dnes již na slávě upadajícímu software ICQ (I Seek You) využívající protokol OSCAR (Open System CommunicAtion in Realtime - otevřený systém pro komunikaci v reálném čase).

2.1 Architektura

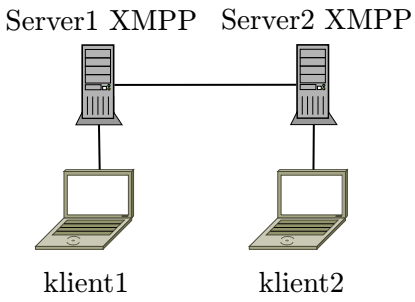
Dobře navržená architektura tvoří základ pro správně fungující internetovou technologii. XMPP protokol využívá decentralizované klient-server složení. Tato struktura se nejvíce podobá struktuře posílání e-mailů. V tomto případu je decentralizace sítě chápána jako inteligentní nezávislost mezi vývojáři klientů a serverů. Každý z nich se může zaměřit na důležité části svého vývoje. Server na spolehlivost a rozšiřitelnost, klient na uživatele. Každý server pracuje samostatně, chod ani výpadek jiné datové stanice nijak neovlivní jeho běh, pouze bude nedostupný seznam kontaktů a služby, kterými server disponuje.

V tabulce č.2.1 jsou shrnuty rozdíly v architektuře Jabber, WWW[1] a e-mail[2]. S každou zde jmenovanou službou má Jabber něco společného. Co se týče charakteristiky se vydal střední cestou. Na rozdíl od e-mailu, nepoužívá vícenásobný hops a v porovnáním s WWW využívá mezi-doménové připojení.

Charakteristika	**WWW**	**Email**	**Jabber**
mezi-doménové připojení	Ne	Ano	Ano
vícenásobný hops	N/A	Ano	Ne

Tabulka 2.1: Srovnání služeb WWW, Email a Jabber

Tyto vlastnosti jsou zárukou pro bezpečný přenos zpráv, znemožňují "krádeže"JID[3] viz. 2.1 a spamování. Obrázek 2.1 zobrazuje přenos zprávy mezi klientem jehož účet vlastní server1 a klientem2 s účtem na serveru2.



Obrázek 2.1: Přenos zprávy

Klient

Klient je především plně ovládatelný grafický program podporující jednoduché odesíláni zpráv. XMPP svou architekturou vnucuje, aby byl co nejjednodušší. Vlastnosti, které by měl mít jsou shrnuty do tří bodů:

1. komunikace se serverem pomocí TCP soketu
2. rozpársrování a následná interpretace příchozí XML zprávy „stanza"(viz 2.3)
3. porozumění sadě zpráv z Jabber jádra

V následující tabulce jsou zobrazeni nejpoužívanější Jabber klienti a jejich funkčnost na operačním systému.

Server

Hlavní vlastnost již není jako u klienta jednoduchost, ale stabilita a bezpečnost. Standardně běží na TCP portu 5222. Komunikace mezi servery je realizována pře port 5269. Každý

[1] World wide web
[2] internetový systém elektronické pošty
[3] uživatelské jméno

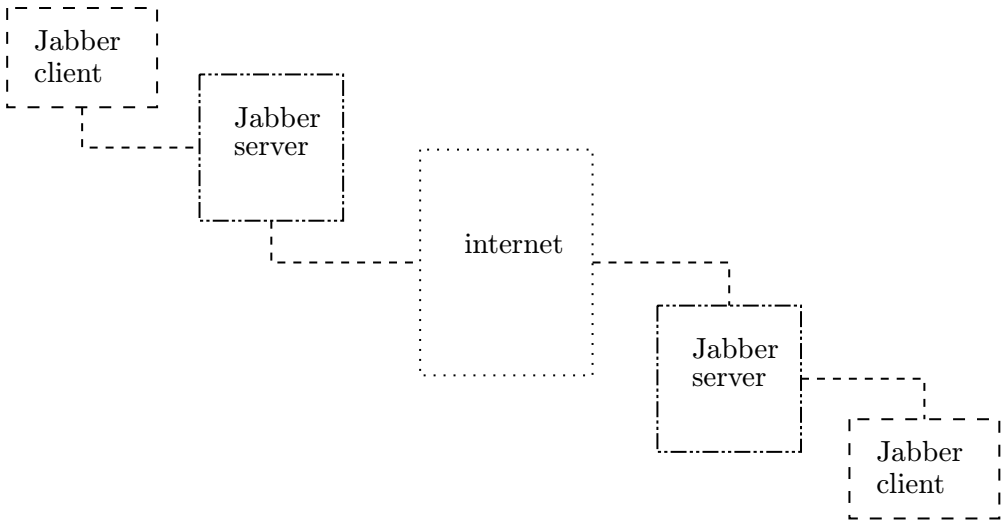
Klient	Windows	Linux	Mac OS	mobilní zařízení
Pidgin	Ano	Ano	Ano	Ne
Empathy	Telephaty[4]	Ano	Ne	Ne
Gajim	Ano	Ano	Ne	Ne
Jabbim	Ano	Ano	Ne	Ne
Meebo[5]	Ano	Ano	Ano	Ano
Miranda	Ano	Ano	Ne	Ne
Psi	Ano	Ano	Ano	Ne
QIP	Ano	Ne	Ne	Ano
Trilian	Ano	Ne	Ne	Ne
kopete	Ne	Ano	Ne	Ne

Tabulka 2.2: Jabber klienti v závislosti na operačním sytému

server uchovává seznam zaregistrovaných uživatelů, kteří se do sítě mohou připojovat pouze přes něj. Tento seznam nemá žádný jiný server. To zajišťuje nemožnost „krádeže“ účtu. Protože XMPP komunikace probíhá přes síť, musí mít každá entita adresu, tady nazvána JabberID. XMPP spoléhá na DNS[6] tudíž používá jména na rozdíl od IP protokolu[7].

Jabber ID

Jabber ID (JID), je jednoznačný identifikátor uživatele. Na první pohled připomíná e-mailovou adresu *user@server*. Je složen z několika částí, takzvané *Jabber bare* neboli čisté ID, je část složena z místa před zavináčem a místa za.



Obrázek 2.2: Distribuovaná architektura Jabberu

[6]Domain Name System
[7]Internet Protocol

ejabberd

Openfire

jabberd2

jabberd14

Prosody

Tigase

V tabulce 2.3 jsou shrnuty informace o serverech Jabbru. První sloupec tvoří jméno, následuje programovací jazyk v němž je napsán. Většina je vydávána pod licencí GPL[8], kromě ejabbred a Prosody. Ejabbred používá GPLv2, což je GPL licence druhé verze a Prosody licenci MIT/X11. Pátý sloupeček *Platforma* obsahuje zkratky platforem pro které je software vyvíjen (L - linux, W - Microsoft windows, M - Mac Os, S - zdrojový kód, So - Solaris, B - BSD). Nakonec je z tabulky viditelné v jakém stavu je vývoj (v - vývoj probíhá, s - vývoj ukončen).

Server	**Jazyk**	**Licence**	**Verze**	**Platforma**	**Domovská stránka**	**stav**
ejabberd	Erlang/ Top	GPLv2	2.1.6	L,W,M,S	ejabberd.im	v
Openfire	java	GPL	3.6.4	L,W,M,U	igniterealtime.org/ projects/openfire	v
jabbred2	c	GPL	2.2.11	L,W,B,So	codex.xiaoka.com/ wiki/jabberd2:start	v
jabbred14	c, c++	GPL	1.6.1.1	AIX,B,HP-UX,IRIX, L,M,So	jabberd.org	s
Prosody	lua	MIT/X11	0.7.0	L,W,M,So	prosody.im	v
Tigase	java	GPL	5.0.0	L,W,M	tigase.org	v

Tabulka 2.3: Přehled Jabber serverů

2.2 XML

Jazyk XML (eXtensible Markup Language), metajazyk pro deklaraci strukturovaných dat, je jádrem protokolu XMPP. Samotný jazyk vznikl rozšířením metajazyka SGML, jež slouží pro deklaraci různých typů dokumentů. Základní vlastností je jednoduchá definice vlastních značek (tagů). Dokument XML se skládá z elementů, jež můžeme navzájem zanořovat. Vyznačujeme je pomocí značek - počáteční a ukončovací.

Základní jednotkou komunikace je stanza. Obsahuje 3 elementy *message*, *presence* a *iq*, jež každý má svůj jednoznačný význam.

[8] General Public License-všeobecná veřejná licence GNU

2.3 Stanza

Message

IQ

Presence

2.4 Knihovny

Jabber je realizován jako otevřený XML standart pro instant messaging formát, proto existuje mnoho programovacích jazyků. Většina z nich disponuje několika knihovnami, usnadňující práci s protokolem XMPP. Mezi nejznámější patří C (iksemel, libstrophe, Loudmoutn), C++ (gloox, Iris), JAVA (JabberBeans, Smack, JSO, Feridian, Emite, minijingle), .NET (Jabber-Net, agsXMPP SDK), Python (JabberPy, PyXMPP, SleekXMPP, Twisted Words), Ruby (XMPP4R, Jabber4R, Jabber::Simple, Jabber::Bot), Perl (Net-Jabber). Níže budou některé rozebrány a vyzdviženy jejich hlavní přednosti.

iksemel

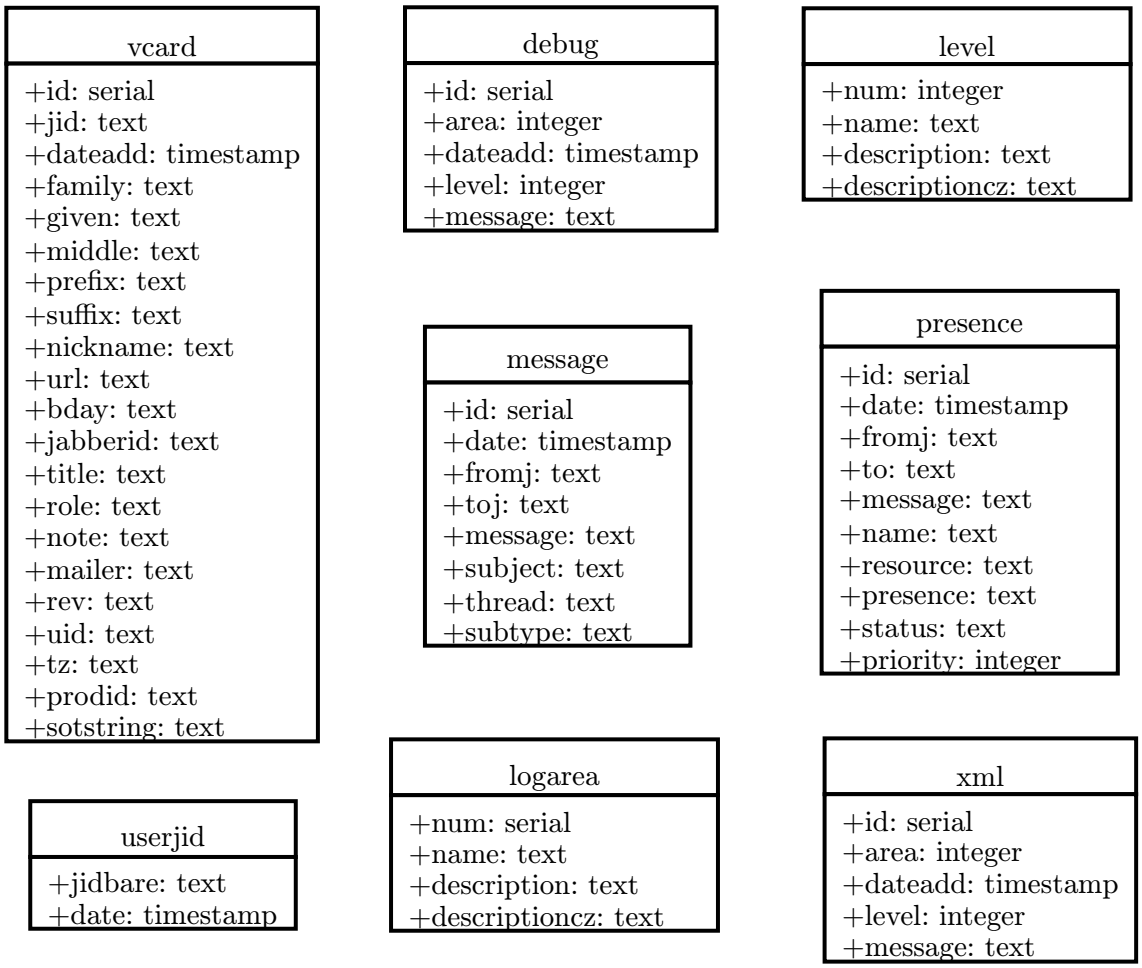
JabberBeans

Jabber-Net

JabberPy

Kapitola 3

Dataming



Obrázek 3.1: Struktura databáze

3.1 Metody dolovaní dat

3.2 Dolovaní znalosti z databází

3.3 Programy

Kapitola 4

Developer

4.1 Slepá ulička

4.2 Knihovny, jazyk

4.3 Jiné produkty

Pravidla pro psaní zkratek jsou uvedena v Pravidlech českého pravopisu [1]. I z jiných důvodů je vhodné, abyste tuto knihu měli po ruce.

Kapitola 5

Závěr

Literatura

[1] Kolektiv autorů: *Pravidla českého pravopisu*. Academia, 2005, iSBN 80-200-1327-X.