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

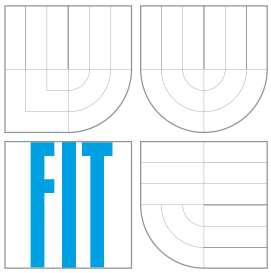
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

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
4. ledna 2011

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). Konkrétně jsou zde popsány stávající vlastnosti implementace [odkud se cerpalo], architektura protokolu XMPP obecně [9, 10] a další detaily ptotokolu [1, 11, 7], které se vztahují k požadavkům na data mining popisovaný v této práci. Další informace byly čerpány z [13, 5, 2, 4, 3].

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 5 platí i pro tento protokol.

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 a e-mail[1]. S každou zde jmenovanou službou má Jabber něco společného. Co se týče charakteristiky se vydal

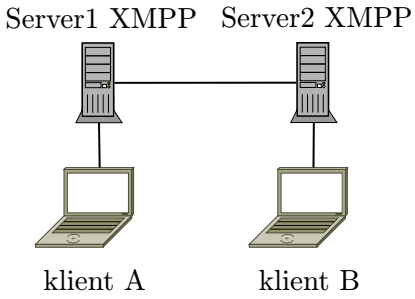
[1] internetový systém elektronické pošty

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[2], který je popsán v podkapitole Jabber ID 2.3, a spamování. Obrázek 2.1 zobrazuje přenos zprávy mezi klientem *A* jehož účet vlastní *server1* a klientem *B* s účtem na *serveru2*.



Obrázek 2.1: Přenos zprávy

Klient

Klient je především plně ovládatelný grafický program podporující jednoduché odesíláni zpráv, ale v této práci jej zastupuje bot s konzolovým rozhraním. XMPP svou architekturou vnucuje, aby byl co nejjednodušší. Vlastnosti, které by měl mít jsou shrnuty do tří bodů:

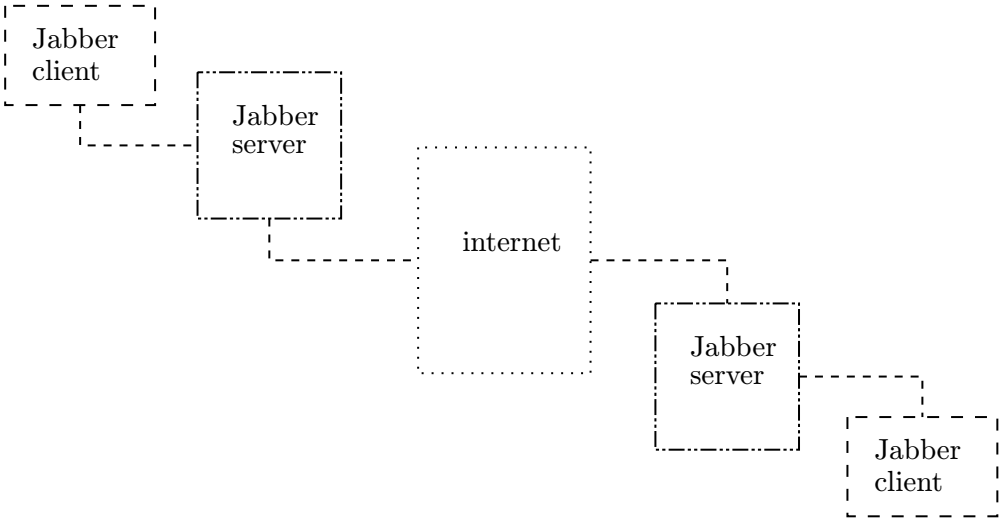
1. komunikace se serverem pomocí TCP soketu
2. rozparsrování a následná interpretace příchozí XML zprávy „stanza“(kapitola 2.3)
3. porozumění sadě zpráv z Jabber jádra

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 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[3] tudíž používá jména na rozdíl od IP protokolu[4].

Obrázek 2.2 znázorňující distribuovanou architekturu Jabberu byl převzat z [7].

[2] uživatelské jméno
[3] Domain Name System
[4] Internet Protocol



Obrázek 2.2: Distribuovaná architektura Jabberu

ejabberd

Openfire

jabberd2

jabberd14

Prosody

Tigase

V tabulce 2.2 jsou shrnuty informace o serverech Jabbru. První sloupec tvoří jméno, následuje programovací jazyk v němž je napsán. Většina je vydávána pod licencí GPL[5], kromě ejabbred a Prosody. Ejabbred používá GPLv2, což je GPL licence druhé verze a Prosody licenci MIT/X11. U všech software byla zkoumána nejaktuálnější verze. Její číslo naleznete ve třetím sloupci. Pouze u software jabberd14 byl ukončen vývoj. Všechny servery leze provozovat na operačním systému Linux a Windows. Na platformě Mac OS mohou být použity všechny zde jmenované vyjma jabberd2. Pět z šesti zde představených software pro server Jabber jsou ve stavu vývoje, tedy kromě jabberd14.

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.

[5]General Public License-všeobecná veřejná licence GNU

Server	Jazyk	Verze	XEP-0060	XEP-0163
ejabberd	Erlang/ Top	2.1.6	ANO	ANO
Openfire	java	3.6.4	ANO	ANO
jabbred2	c	2.2.11	NE	NE
jabbred14	c, c++	1.6.1.1	ANO[6]	NE
Prosody	lua	0.7.0	NE [7]	ANO
Tigase	java	5.0.0	ANO	ANO

Tabulka 2.2: Přehled Jabber serverů

2.3 Stanza

Jabber ID

Jabber ID (JID) je jednoznačný virtuální identifikátor uživatele na síti. Není case-sensitive a je složen ze dvou částí. Takzvané *Jabber bare* neboli čisté ID a *resource*. První je část na první pohled připomíná e-mailovou adresu *user@server*. Druhá část slouží k přesné identifikaci jednotlivých spojení. Je použit ke směrování trafiku s uživateli v případě otevření vícero spojení pod jedním uživatelem. Společně Jabber bare a resource tvoří tzv. *full JID* — *user@server/resource*.

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. V tabulka 2.3 jsou pro nejznámější programovací jazyky zobrazeny dostupné knihovny. Níže budou některé rozebrány a vyzdviženy jejich hlavní přednosti.

Programovací jazyk	knihovna
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
Perl	Net-Jabber
Ruby	XMPP4R, Jabber4R, Jabber::Simple, Jabber::Bot

Tabulka 2.3: Přehled Jabber knihoven

iksemel

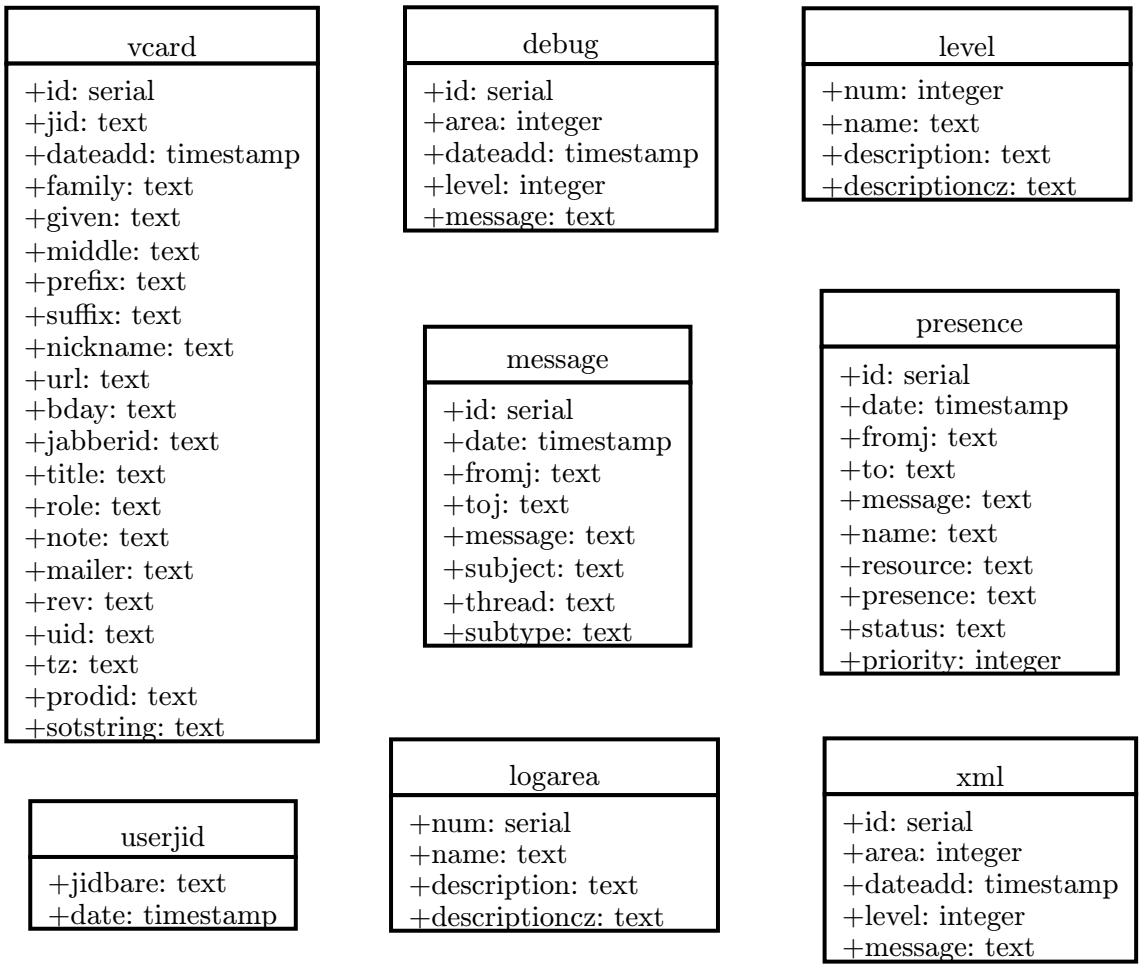
JabberBeans

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

Kapitola 5

Závěr

[12, 14, 16, 19, 15, 17, 18, 6, 8]

Literatura

[1] Adams, D.: *Programming jabber*. Sebastopol: O'Reilly, první vydání, 2002, 455 s., iSBN 05-960-0202-5.

[2] Fred, H.: *Computer networking and the internet*. Edinburg: Addison-Wesley Publishing Company, první vydání, 2005, 803 s., iSBN 03-212-6358-8.

[3] Kolektiv autorů: Extensible Markup Language (XML) 1.0. [online], 26-11-2008, [cit. 4. ledna 2011].
URL http://www.w3.org/TR/2008/REC-xml-20081126/

[4] Kosek, J.: *XML pro každého : podrobný průvodce*. Praha: Grada, první vydání, 2000, 163 s., iSBN 80-716-9860-1.

[5] Kurose, J. F.; Ross, K. W.: *Computer networking : top-down approach featuring the internet*. Boston: Addison-Wesley Publishing Company, druhé vydání, 2003, 752 s., iSBN 03-211-7644-8.

[6] Millard, P.; Saint-Andre, P.; Meijer, R.: XEP-0060: Publish-Subscribe. [online], 12-07-2010, [cit. 4. ledna 2011].
URL http://xmpp.org/extensions/xep-0060.html

[7] Moore, D.; Wright, W.: *Jabber developer's handbook*. Indianapolis: Sams Publishing, první vydání, 2004, 487 s., iSBN 06-723-2536-5.

[8] Saint-Andre, P.; Smith, K.: XEP-0163: Personal Eventing Protocol. [online], 12-07-2010, [cit. 4. ledna 2011].
URL http://xmpp.org/extensions/xep-0163.html

[9] Saint-André, P.: Extensible Messaging and Presence Protocol (XMPP): Core. [online], 10-2004, [cit. 4. ledna 2011].
URL http://tools.ietf.org/html/rfc3920

[10] Saint-André, P.: Extensible Messaging and Presence Protocol (XMPP): Instant Messaging and Presence. [online], 10-2004, [cit. 4. ledna 2011].
URL http://tools.ietf.org/html/rfc3921

[11] Saint-André, P.; Smith, K.; Troncon, R.: *XMPP : the definitive guide : building real-time applications with jabber technologies*. Sebastopol: O'Reilly, první vydání, 2009, 287 s., iSBN 978-059-6521-264.

[12] Schröter, J.: Gloox API Dokumentace. [online], 31-11-2009 , [cit. 4. ledna 2011].
URL http://camaya.net/api/gloox-1.0/index.html

[13] Stevens, W.; Fenner, B.; M.Rudoff, A.: *UNIX Network Programming*. Boston: Addison-Wesley Publishing Company, třetí vydání, 2004, 991 s., iSBN 01-314-1155-1.

[14] WWW Stránky: Ejabberd the Erlang Jabber/XMPP deamon community site. [online], 2010 , [cit. 4. ledna 2011].
URL http://www.ejabberd.im/

[15] WWW Stránky: Jabberd14 the original Jabber server implementation. [online], 2010 , [cit. 4. ledna 2011].
URL http://jabberd.org/

[16] WWW Stránky: Openfire is a real time collaboration server. [online], 2010 , [cit. 4. ledna 2011].
URL http://www.igniterealtime.org/projects/openfire/

[17] WWW Stránky: Prosody a study in simplicity. [online], 2010 , [cit. 4. ledna 2011].
URL http://prosody.im/

[18] WWW Stránky: Tigase is the website of the Tigase XMPP/Jabber Server. [online], 2010 , [cit. 4. ledna 2011].
URL http://www.tigase.org/

[19] WWW Stránky - Matthias Wimmer: Jabberd2 - XMPP Server. [online], 2010 , [cit. 4. ledna 2011].
URL http://codex.xiaoka.com/wiki/jabberd2:start

Příloha A

Obsah CD

Příloha B

Manual

Příloha C

Konfigrační soubor

Příloha D

Slovník výrazů

XMPP — dkshckdsjvlsdjvodsvjdfokj

IM služby — dkshckdsjvlsdjvodsvjdfokj

XEP — dkshckdsjvlsdjvodsvjdfokj

JEP — dkshckdsjvlsdjvodsvjdfokj

SASL — dkshckdsjvlsdjvodsvjdfokj

TLS — dkshckdsjvlsdjvodsvjdfokj

GPG — dkshckdsjvlsdjvodsvjdfokj

XML — dkshckdsjvlsdjvodsvjdfokj

WWW — dkshckdsjvlsdjvodsvjdfokj

jabber — dkshckdsjvlsdjvodsvjdfokj

JID — dkshckdsjvlsdjvodsvjdfokj

presence — dkshckdsjvlsdjvodsvjdfokj

stanza — dkshckdsjvlsdjvodsvjdfokj

klient — dkshckdsjvlsdjvodsvjdfokj

server — dkshckdsjvlsdjvodsvjdfokj

e-mail — dkshckdsjvlsdjvodsvjdfokj

DNS — dkshckdsjvlsdjvodsvjdfokj

TCP — dkshckdsjvlsdjvodsvjdfokj

IP — dkshckdsjvlsdjvodsvjdfokj

vCard — dkshckdsjvlsdjvodsvjdfokj