FAKULTA PRÍRODNÝCH VIED
UNIVERZITA KONŠTANTÍNA FILOZOVA V NITRE

ZÍSKAVANIE ZNALOSTÍ Z ELEKTRONICKÝCH INFORMAČNÝCH ZDROJOV

Innovation and support of doctoral study program

Michal Munk

Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327

Katedra informatiky
FPV UKF v Nitre

esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
University of Pardubice
INVESTMENTS IN EDUCATION DEVELOPMENT

CIEĽ A OBSAH PREDNÁŠKY

Cieľ:

- Prezentovať fázy procesu získavania znalostí a ich špecifiká s dôrazom na web log mining

Obsah:

- Úvod do problematiky získavania znalostí
- Fázy procesu objavovania znalostí na základe používania webu
  - Definícia cieľovej úlohy - určenie typu problému
  - Získanie relevantných dát o používaní webu a porozumenie dátam
  - Predspracovanie dát - čistenie dát, identifikácia používateľov/sedení, rekonštrukcia aktivít
  - Dolovanie z dát - aplikácia analytických metód
  - Interpretácia a evalvácia nájdených znalostí
  - Aplikácia získaných znalostí

PROCES OBJAVOVANIA ZNALOSTÍ

- Dnešná doba - charakteristická množstvom elektronicky dostupných dát na jednej strane - často nedostatkom znalostí na strane druhej (Paralič, 2003)
- Obrovské množstvo údajov - slabá výpovedná hodnota
- Koncept objavovania znalostí (Knowledge Discovery)

Objavovanie znalostí chápeme ako **proces**, ktorý zahrňuje **výber dát, predspracovanie dát, transformáciu dát, analýzu dát a interpretáciu výsledkov** (Fayyad et al., 1996)

- Knowledge Discovery in Databases
- Knowledge Discovery in Texts/Text Mining
- Web Mining

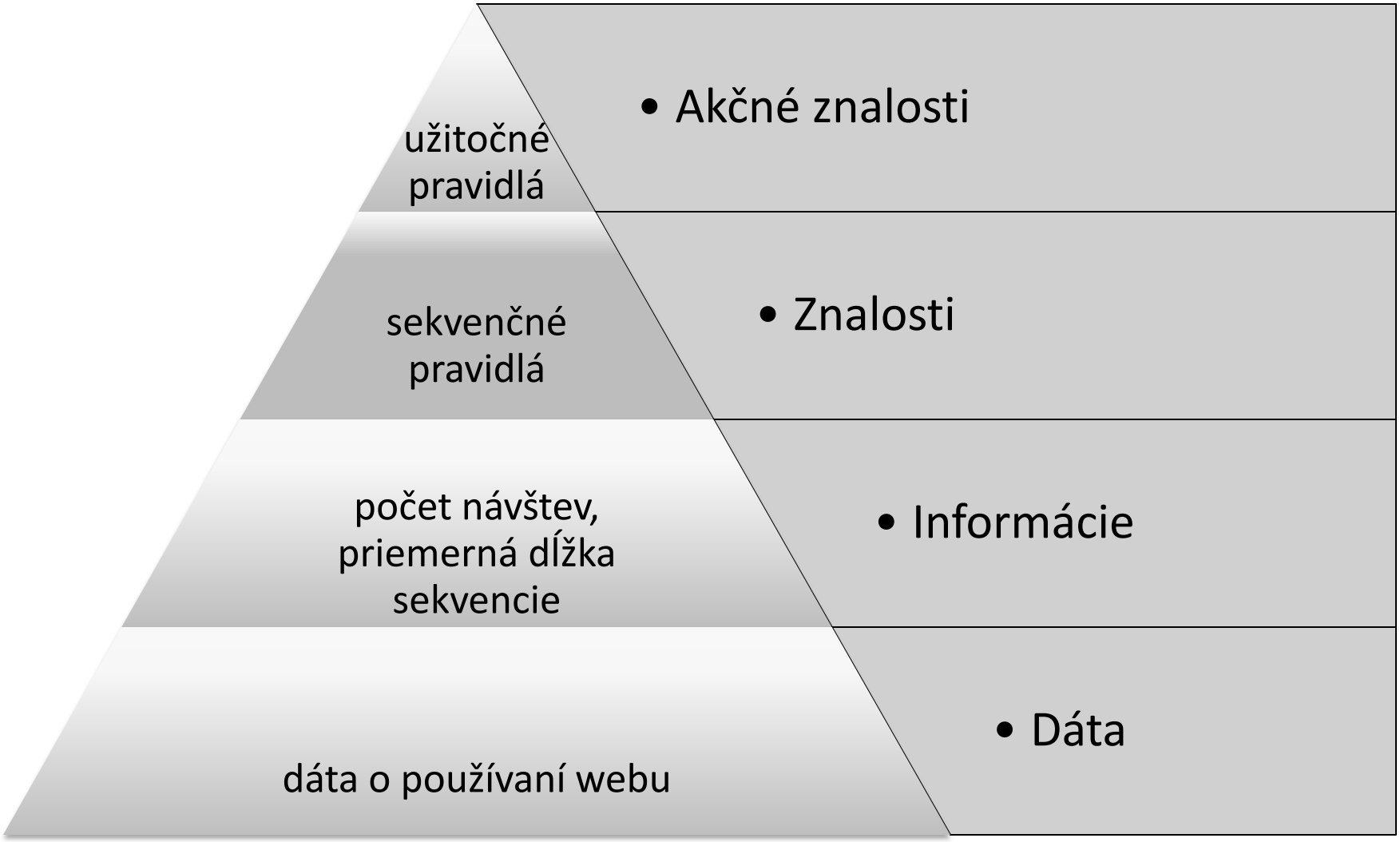
PROCES OBJAVOVANIA ZNALOSTÍ

Objavovanie znalostí z webu (web mining) môžeme definovať ako **extrakciu zaujímavých a potenciálne užitočných znalostí a informácií z aktivít súvisiacich s webom** (Liu, 2007)

- Web Content Mining
- Web Structure Mining
- Web Log Mining/Web Usage Mining

Cieľom objavovania znalostí na základe používania webu (web log mining) je **analýza správania sa používateľov pri prechádzaní webu** (Srivastava et al., 2000; Berka, 2003; Romero et al., 2009)

PRINCÍP OBJAVOVANIA ZNALOSTÍ

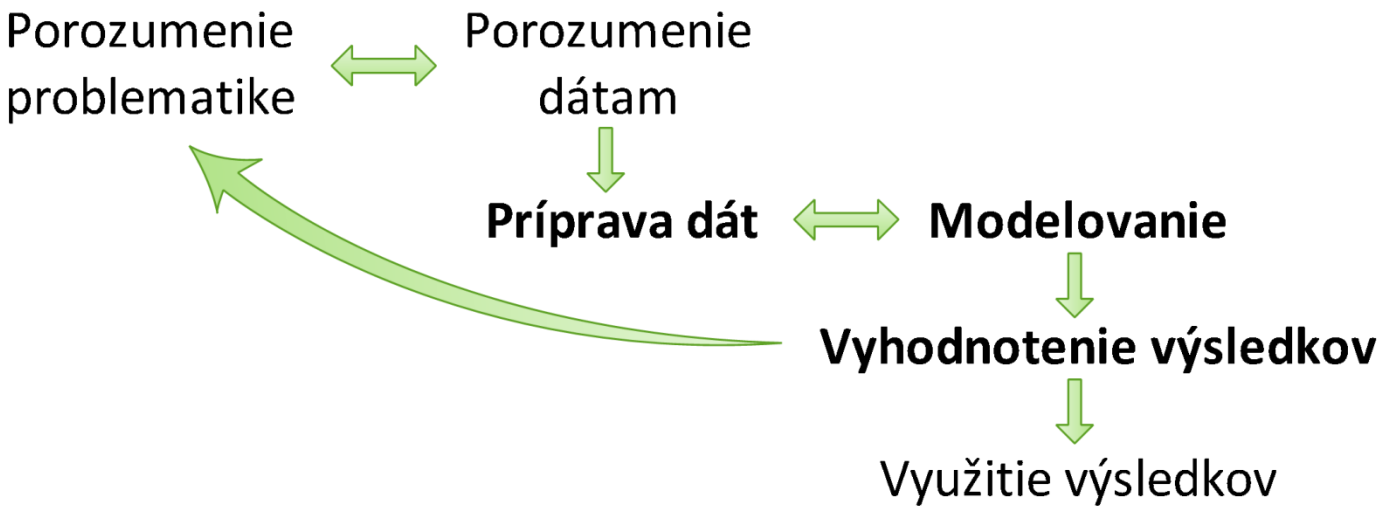


FÁZY PROCESU WEB LOG MININGU

Proces objavovania znalostí na základe používania webu sa rozdeľuje do troch základných častí:

Predspracovanie ⇒ **Objavovanie vzorov** ⇒ **Analýza vzorov**

Metodika CRISP-DM, fázy procesu:



Porozumenie problematike

Typy problémov - úloh:

- Deskripcia dát a sumarizácia
- Segmentácia
- Deskripcia konceptov
- Klasifikácia
- Predikcia
- Analýza závislostí

Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327

esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
University of Pardubice

INVESTMENTS IN EDUCATION DEVELOPMENT

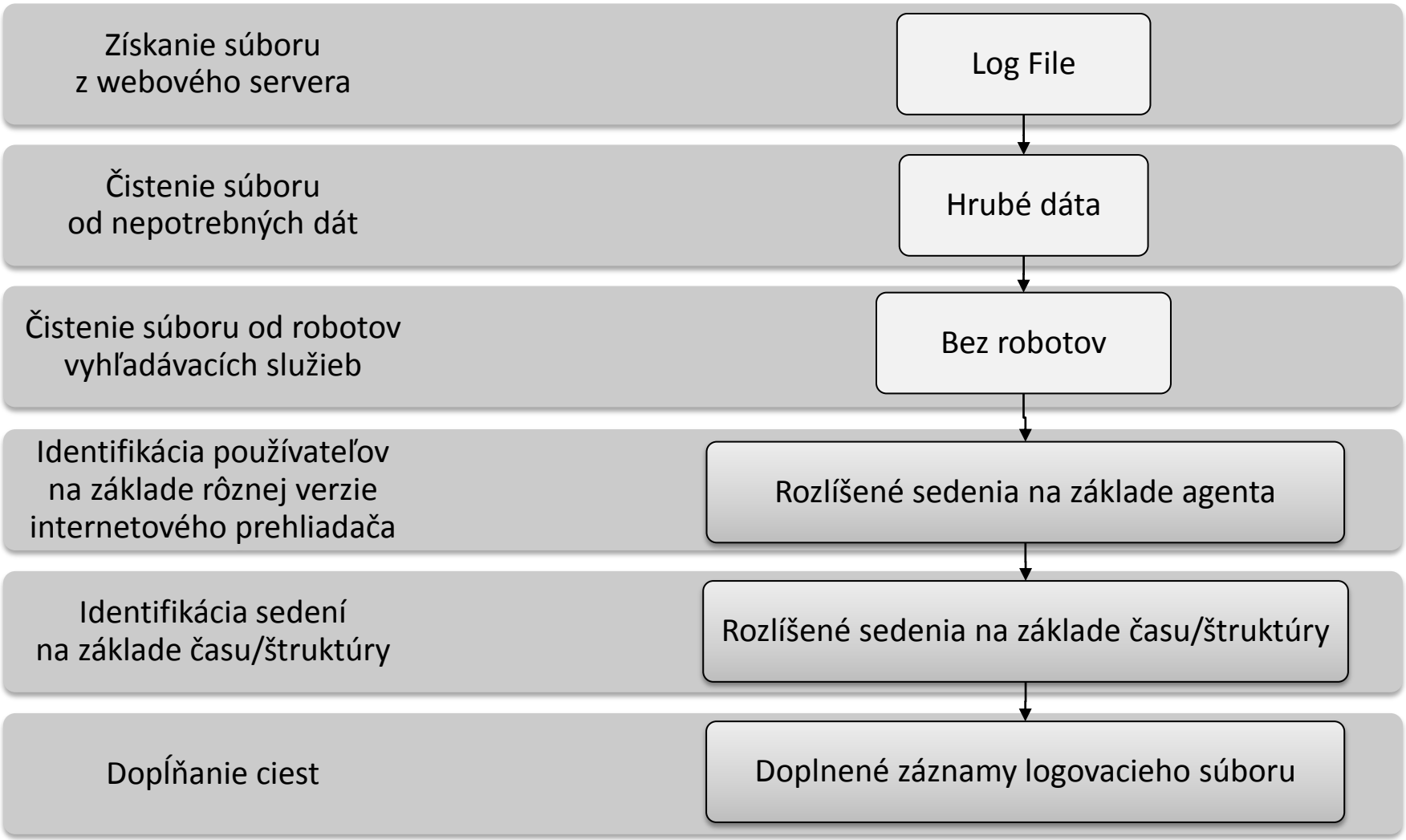
Porozumenie dátam

- Informačné systémy - dáta vo vlastnej štruktúre, organizované v databáze
- Web/proxy servery - dáta v spoločnej štandardnej štruktúre, v textovom formáte

178.41.1.187 - - [16/Dec/2011:18:16:26 +0100] "GET /studium/akreditovane-programy HTTP/1.1" 200 15364 "http://www.ukf.sk/studium" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sk; rv:1.9.2.13) Gecko/20101203 Firefox/3.6.13 BLNGBAR"

...

Príprava dát



Príprava dát

```
GET / HTTP/1.1

GET /favicon.ico HTTP/1.1

GET /templates/system/css/system.css HTTP/1.1
GET /templates/system/css/general.css HTTP/1.1
GET /templates/newukf/js/hs.css HTTP/1.1
GET /templates/newukf/css/template.css HTTP/1.1
GET /templates/newukf/css/editor_content.css HTTP/1.1

GET /templates/newukf/js/ja.script.js HTTP/1.1
GET /media/system/js/mootools.js HTTP/1.1
GET /media/system/js/caption.js HTTP/1.1
GET /templates/newukf/js/highslide.js HTTP/1.1

GET /templates/newukf/images/user-increase.png HTTP/1.1
GET /templates/newukf/images/ukf_header_logo.png HTTP/1.1
GET /templates/newukf/images/en.gif HTTP/1.1
GET /dokumenty/images/reklama/e-prihlaska.gif HTTP/1.1
GET /templates/newukf/images/ukf_budova_jesen.jpg HTTP/1.1
```

Príprava dát

Identifikácia robotov:

- na základe poľa User-Agent, resp. IP adresy
- na základe prístupu k súboru robots.txt v adresári servera
- na základe prístupu k skrytému odkazu

Agent
msnbot/1.1 (+http://search.msn.com/msn**bot**.htm)
ia_archiver (+http://www.alexa.com/site/help/webmasters; **crawler**@alexa.com)
Mozilla/5.0 (compatible; Google**bot**/2.1; +http://www.google.com/bot.html)

IP adresa	Host
65.55.110.205	msn**bot**-65-55-110-205.search.msn.com
74.6.18.226	llf520148.**crawl**.yahoo.net
77.88.27.25	**spider**27.yandex.ru

Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327
esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
University of Pardubice
INVESTMENTS IN EDUCATION DEVELOPMENT

Príprava dát

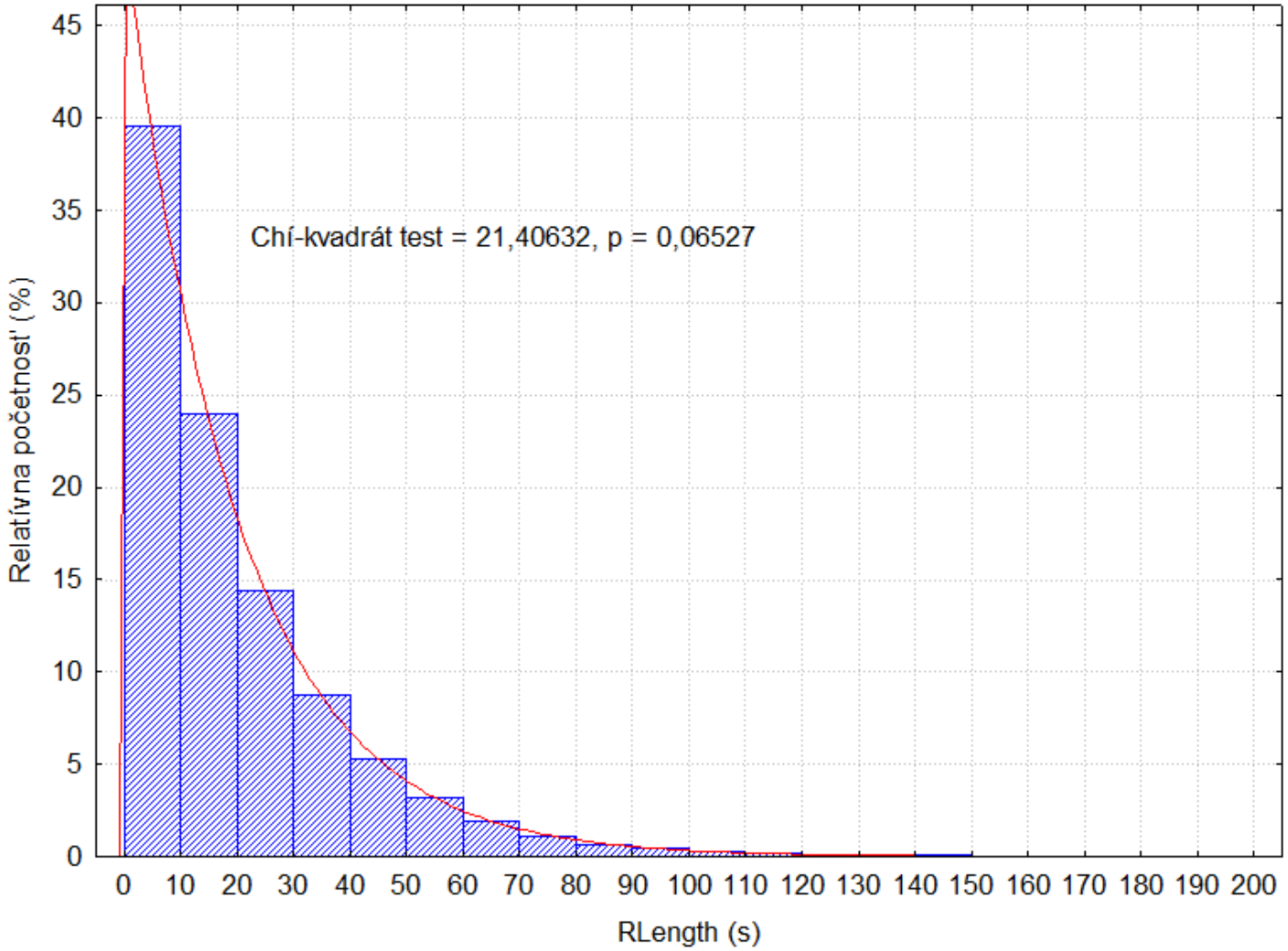
Identifikácia používateľov/sedení:

- na základe rôznej verzie internetového prehliadača
- na základe štruktúry prehľadávaných stránok
- na základe času

```
178.41.1.187 - - [16/Dec/2011:18:16:26 +0100] "GET " ..
178.41.1.187 - - [16/Dec/2011:18:19:47 +0100] "GET /studium " ..
178.41.1.187 - - [16/Dec/2011:18:52:37 +0100] "GET /studium/akreditovane-programy "..
178.41.1.187 - - [16/Dec/2011:18:54:25 +0100] "GET /prijimaciekonanie " ..
178.41.1.187 - - [16/Dec/2011:18:59:58 +0100] "GET /prijimaciekonanie/podmienky "..
```


Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327
esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
University of Pardubice
INVESTMENTS IN EDUCATION DEVELOPMENT

Príprava dát



Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327

Príprava dát

Premenná $RLength$ má exponenciálne rozdelenie

$$f(RLength) = \lambda e^{-\lambda RLength},$$
$$F(RLength) = 1 - e^{-\lambda RLength},$$

kde $RLength \geq 0$

Ak je p relatívna početnosť navigačných stránok, na odhad hraničného času C môžeme použiť kvantilovú funkciu

$$F^{-1}(p, \lambda) = C = \frac{-\ln(1-p)}{\lambda}$$

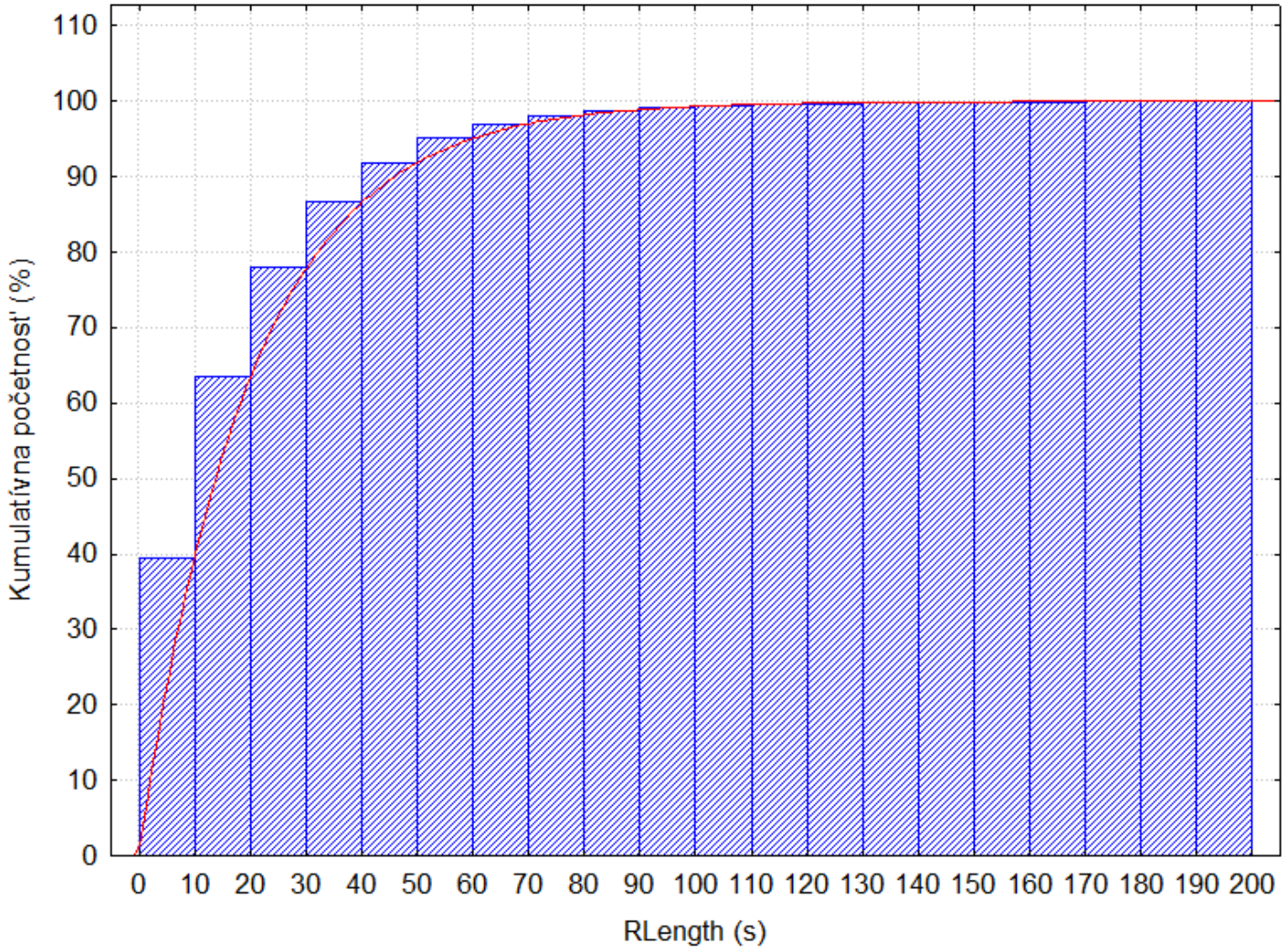
pre $0 \leq p < 1$

Maximálne vierohodný odhad parametra λ (priemerná intenzita udalostí) je

$$\hat{\lambda} = \frac{1}{\overline{RLength}},$$

kde $\overline{RLength}$ je pozorovaný priemer dĺžky návštev

Príprava dát



Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327

Príprava dát

Ak máme odhad hraničného času C, tak sedenie je sekvencia navštívených stránok s časovou známkou, pre ktorú platí:

$$\langle USID, \langle URL_1, DTime_1, RLength_1 \rangle, \ldots, \langle URL_k, DTime_k, RLength_k \rangle \rangle,$$

$$RLength_i \leq C, \qquad (1)$$

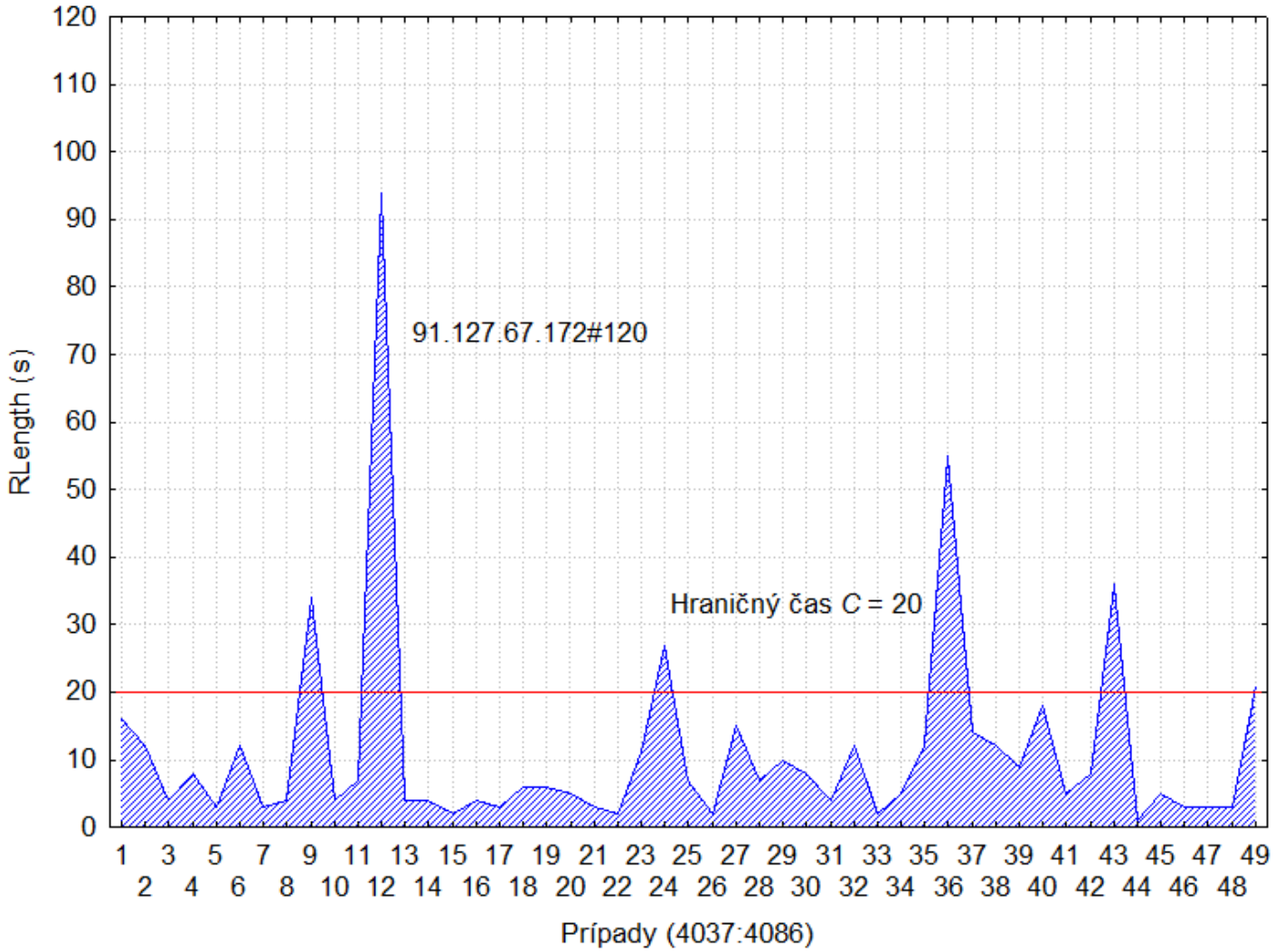
kde $1 \leq i < k$ a pre poslednú stránku sedenia platí:

$$RLength_k > C \qquad (2)$$

Od stránky s vlastnosťou (2) je definované ďalšie sedenie

Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327
INVESTMENTS IN EDUCATION DEVELOPMENT

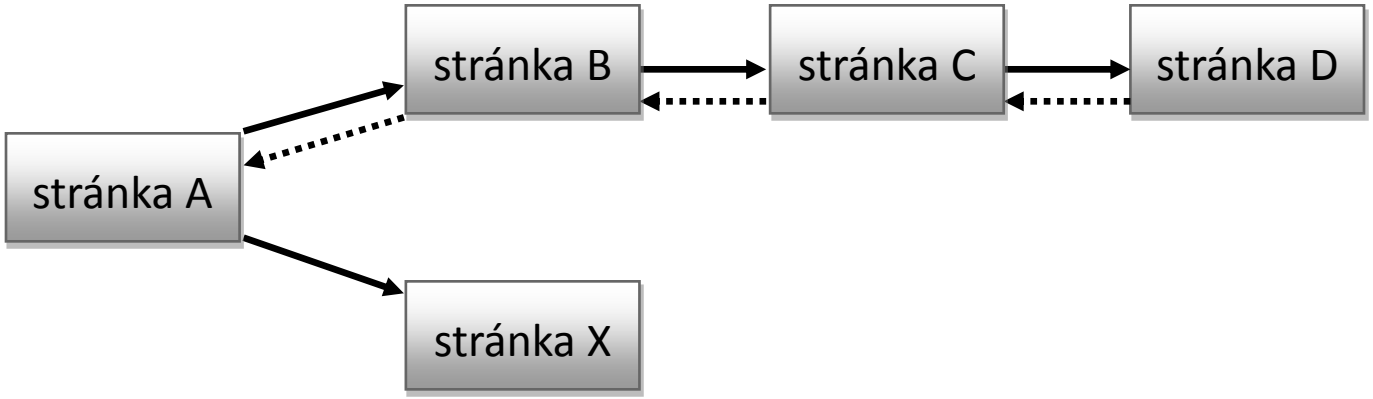
Príprava dát



Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327

Príprava dát

Aktuálna navigácia



Mapa webu

URL adresa	Odkazujúca stránka
B	A
X	A
C	B
D	C

Záznamy logovacieho súboru

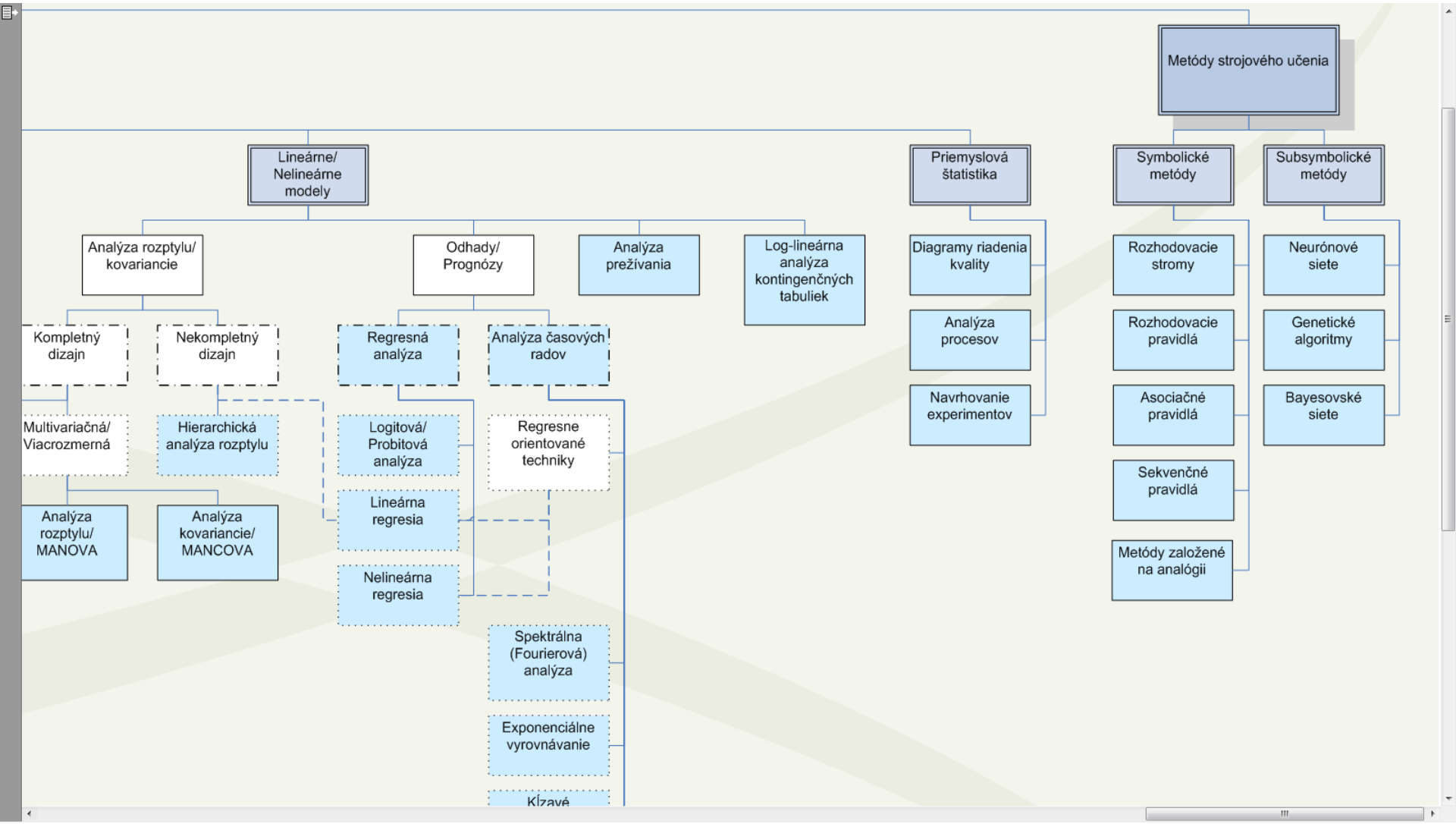
194.160.10.10#5#2	12:00	stránka A
194.160.10.10#5#2	12:01	stránka B
194.160.10.10#5#2	12:04	stránka C
194.160.10.10#5#2	12:08	stránka D
194.160.10.10#5#2	12:15	stránka X

Dopĺňanie ciest

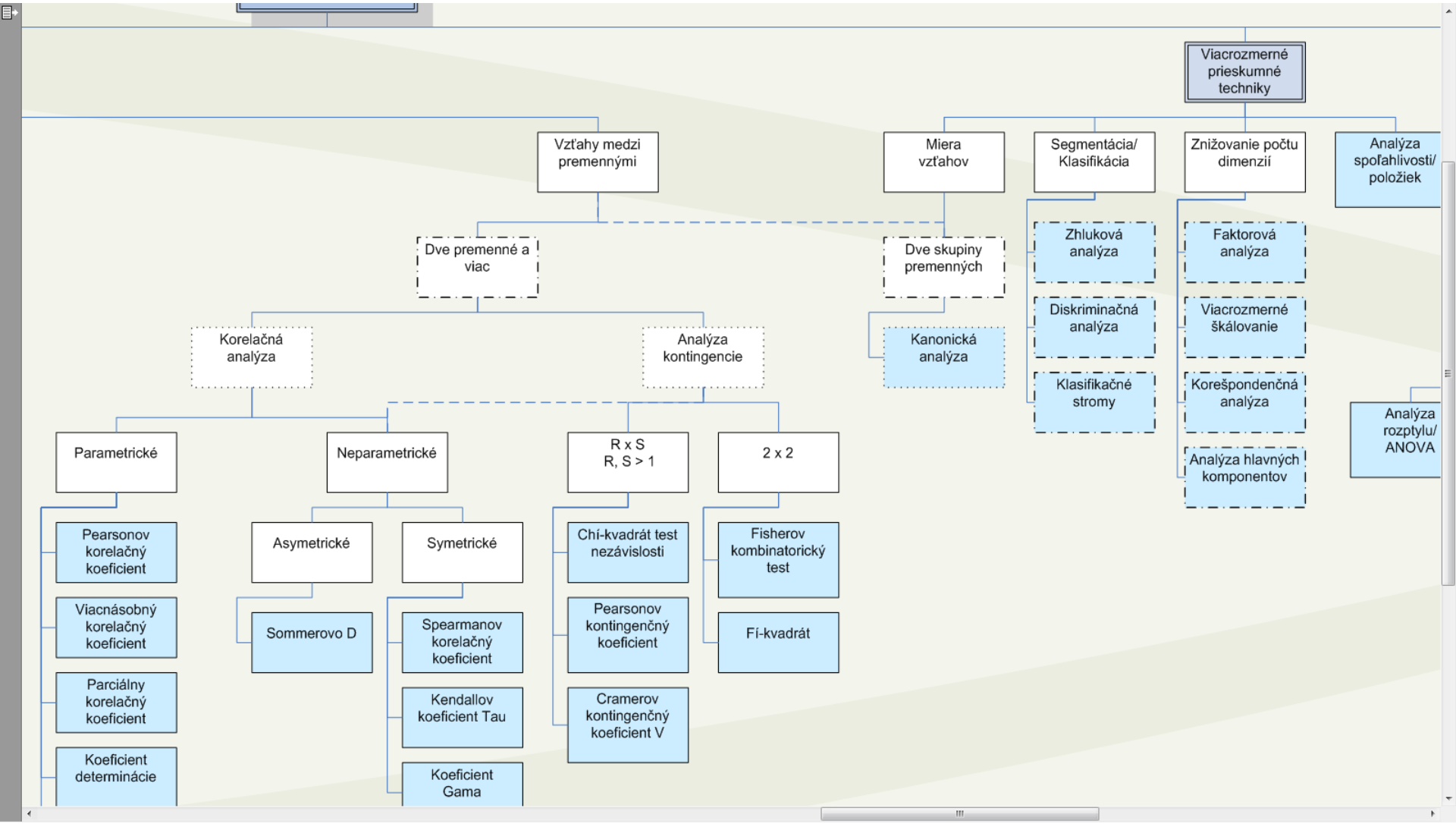
A→B→C→D→X
A→B→C→D→**C**→X
A→B→C→D→**C**→**B**→X
A→B→C→D→**C**→**B**→**A**→X

Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327
esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
University of Pardubice
INVESTMENTS IN EDUCATION DEVELOPMENT

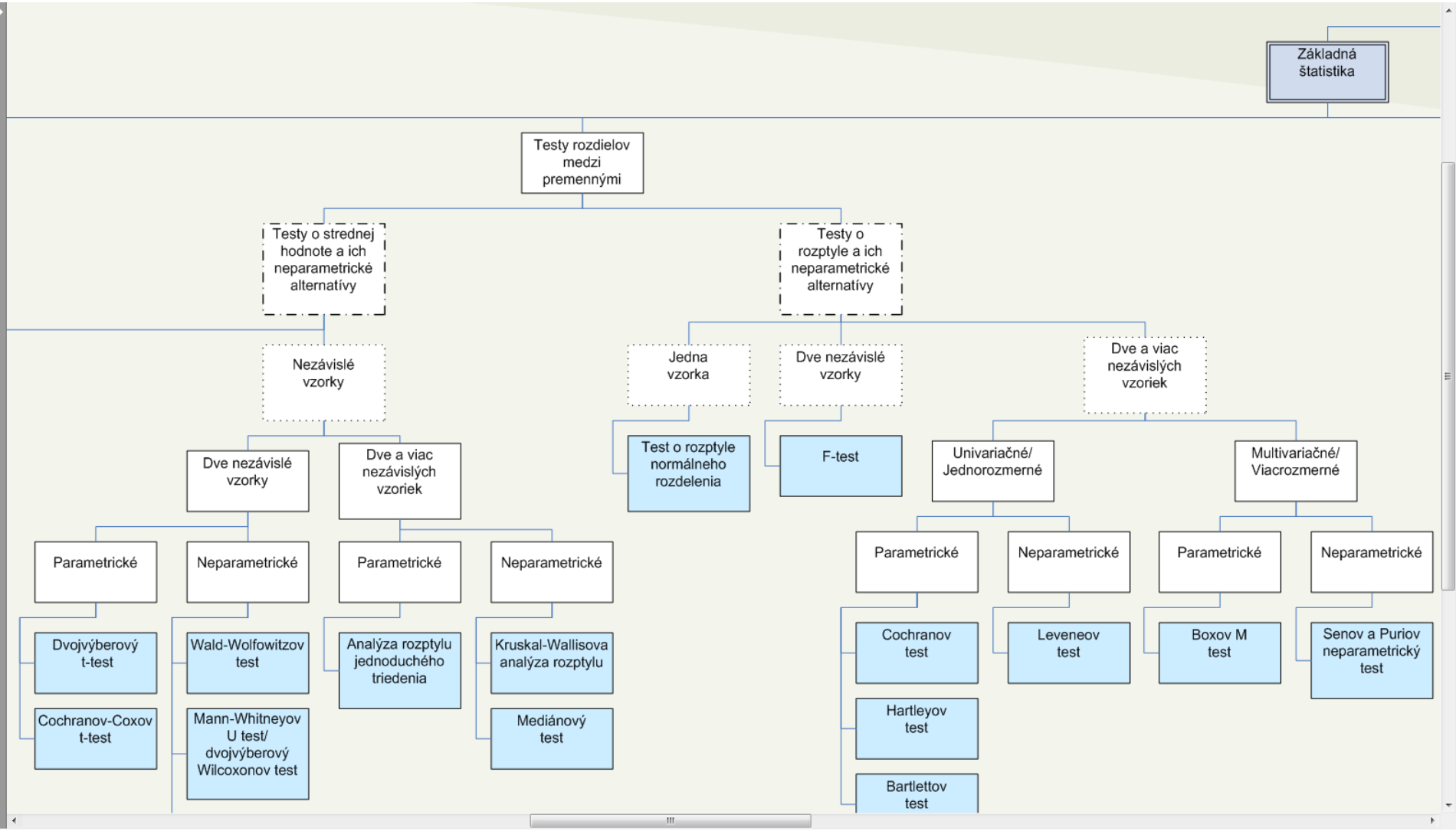
Modelovanie a vyhodnotenie výsledkov



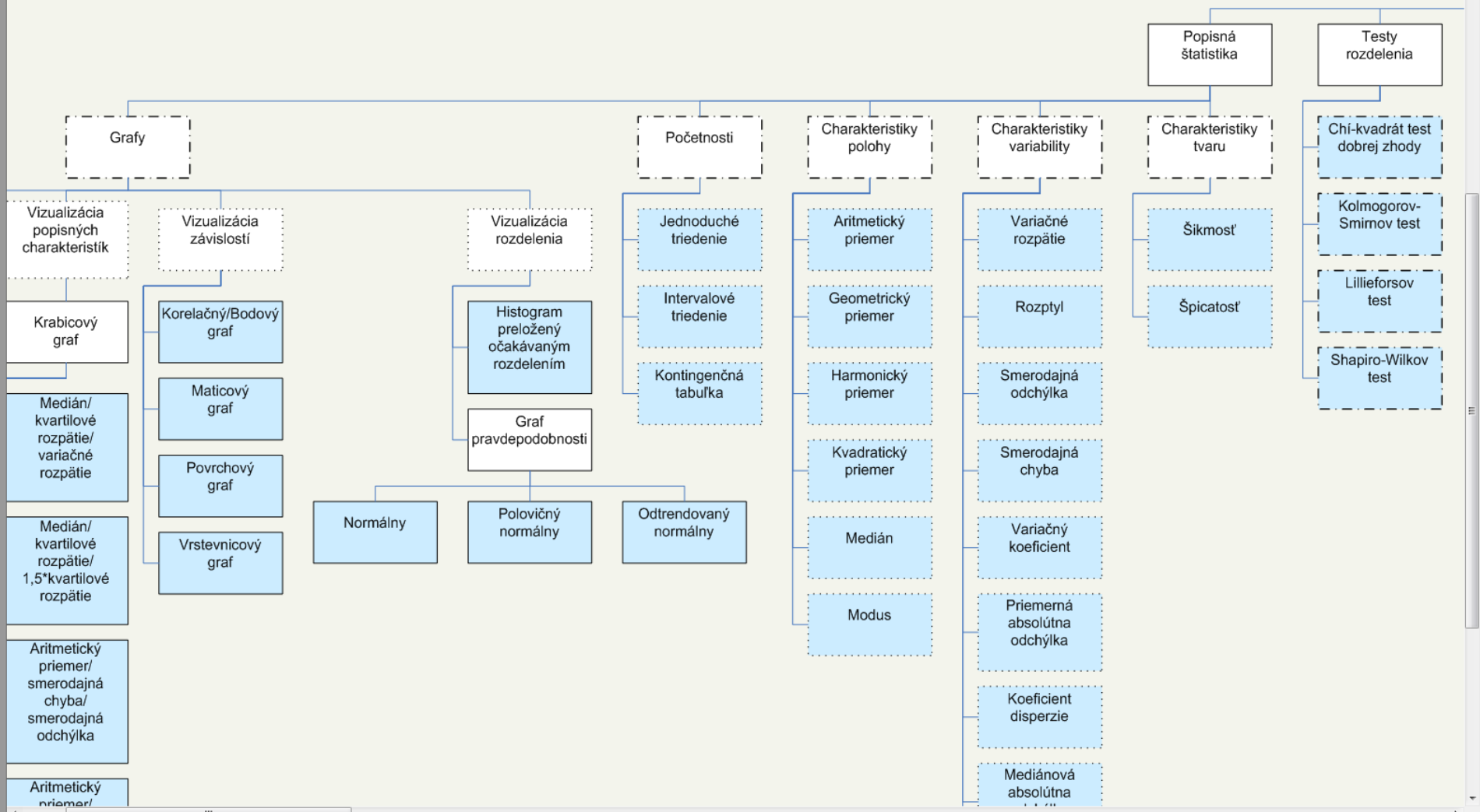
Modelovanie a vyhodnotenie výsledkov



Modelovanie a vyhodnotenie výsledkov



Modelovanie a vyhodnotenie výsledkov



Využitie výsledkov

Všeobecne môžeme aplikácie web log miningu rozdeliť do piatich kategórií: personalizácia, zlepšovanie systémov, modifikácia webových stránok, bussines intelligence a charakteristika používania (Srivastava et al., 2000)

Analýza návštevnosti - Analýza závislostí

Plánovanie údržby - Predikcia

Reštrukturalizácia - Analýza závislostí

Personalizácia - Segmentácia, Klasifikácia, Analýza závislostí

SÚHRN

- Cieľom prednášky bolo prezentovať fázy procesu získavania znalostí
- Zamerali sme sa na zdroje dát o používaní webu
- Podobne by sme postupovali aj pri riadení procesu objavovania znalostí z databáz, či textov
- Najväčšie rozdiely medzi oblasťami objavovania znalostí pri riadení procesu metodikou CRISP-DM sú vo fáze prípravy dát
- Príprava dát predstavuje časovo najnáročnejšiu fázu v rámci celého procesu objavovania znalostí
- Zložitosť prípravy dát závisí od použitého zdroja dát

Fáza modelovania

1. Určenie modelu

Pravdepodobnostné rozdelenie počtov prístupov Y_{ij} v čase i na kategóriu j s pozorovaniami y_{ij}, ak je daný počet prístupov $n_i = \sum_j y_{ij}$ v čase i, je

$$P\left[Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots, Y_{iJ} = y_{iJ}\right] = \frac{n_i!}{y_{i1}!\, y_{i2}! \dots y_{iJ}!} \pi_{i1}^{y_{i1}} \pi_{i2}^{y_{i2}} \dots \pi_{iJ}^{y_{iJ}}$$

Pretože $\sum_{j=1}^{J} \pi_{ij} = 1$, potrebujeme odhadnúť $J - 1$ neznámych pravdepodobností

Odhady sa vypočítajú metódou maximálnej vierohodnosti

ln-funkcia vierohodnosti: $\sum_i \sum_{j=1}^{J} y_{ij} \ln \pi_{ij}$

Logitová transformácia: $\eta_{ij} = \ln \frac{\pi_{ij}}{\pi_{iJ}}, \eta_{iJ} = 0, \eta_{ij} = \alpha_j + \mathbf{x}_i^T \boldsymbol{\beta}_j$

Inverzná transformácia: $\pi_{iJ} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\eta_{ij}}}, \pi_{ij} = e^{\eta_{ij}} \pi_{iJ}, j = 1, 2, \dots, J-1,$

$$\pi_{iJ} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\alpha_j + \mathbf{x}_i^T \boldsymbol{\beta}_j}}, \pi_{ij} = \frac{e^{\alpha_j + \mathbf{x}_i^T \boldsymbol{\beta}_j}}{1+\sum_{j=1}^{J-1} e^{\alpha_j + \mathbf{x}_i^T \boldsymbol{\beta}_j}}$$

Fáza modelovania

2. Odhad parametrov modelu $\alpha_j, \boldsymbol{\beta}_j$ maximalizáciou logaritmu multinominálnej funkcie vierohodnosti

$$\text{H0: } \alpha_j = 0, \text{H0: } \beta_{kj} = 0, \text{ kde } k \text{ je počet prediktorov}$$

3. Odhad logitov η_{ij} pre všetky hodnoty nezávislých premenných

$$\hat{\eta}_{ij} = a_j + \mathbf{x}_i^T \mathbf{b}_j, j = 1, 2, \ldots, J-1$$

4. Odhad pravdepodobností prístupov π_{iJ} v čase i pre referenčnú webovú časť J

$$\hat{\pi}_{iJ} = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\hat{\eta}_{ij}}}$$

5. Odhad pravdepodobností prístupov π_{ij} v čase i pre webovú časť j

$$\hat{\pi}_{ij} = e^{\hat{\eta}_{ij}} \hat{\pi}_{iJ}, j = 1, 2, \ldots, J-1$$

6. Vizualizácia pravdepodobností výberu webovej časti j v čase i,

$$j = 1, 2, \ldots, J, i \in \{0, 1, \ldots, 23\}$$

Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327
esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
University of Pardubice
INVESTMENTS IN EDUCATION DEVELOPMENT

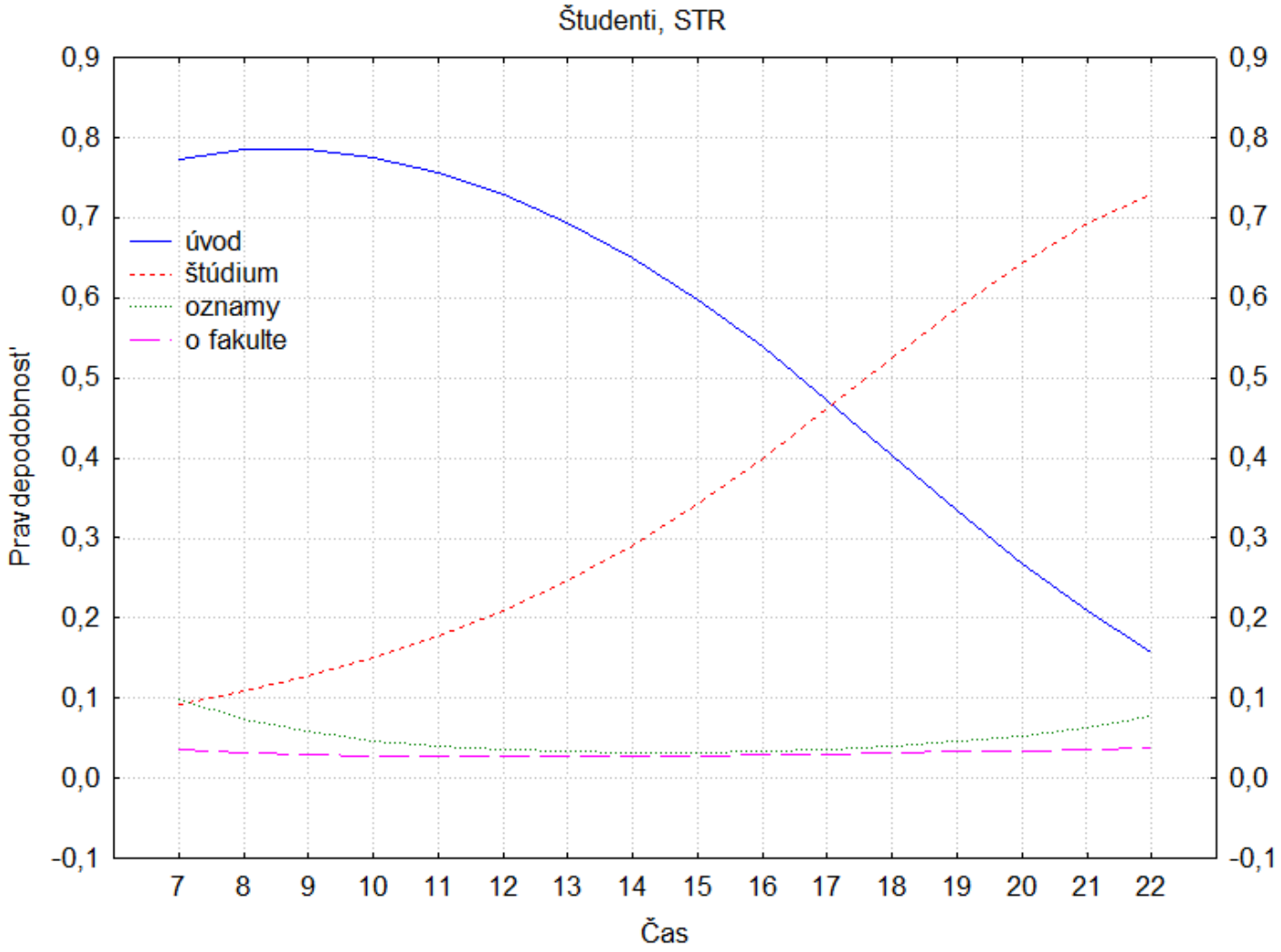
Fáza modelovania

$$\eta_{ij} = \alpha_j + \beta_{1j} t_i + \beta_{2j} t_i^2 + \gamma_{1j} PO_i + \gamma_{2j} UT_i + \gamma_{3j} STR_i + \gamma_{4j} STVR_i + \gamma_{5j} PIA_i$$

	sv	Wald. št.	p
Abs. člen	3	46,0604	0,0000
t	3	30,6124	0,0000
t²	3	35,8017	0,0000
PO	3	78,2983	0,0000
UT	3	103,9821	0,0000
STR	3	48,4787	0,0000
STVR	3	54,7024	0,0000
PIA	3	11,2170	0,0106

	Kategória	Odhad par.	Sm. chyba	Wald. Št.	p
t	*úvod*	0,3418	0,1718	3,9571	0,0467
t²	*úvod*	-0,0156	0,0060	6,8487	0,0089
PO	*úvod*	-0,6335	0,2975	4,5358	0,0332
UT	*úvod*	1,1031	0,4581	5,7978	0,0160
STR	*úvod*	0,1138	0,3684	0,0954	0,7575
STVR	*úvod*	-0,4991	0,3611	1,9108	0,1669
PIA	*úvod*	0,6868	0,4075	2,8402	0,0919
t	*štúdium*	0,3845	0,1802	4,5525	0,0329
t²	*štúdium*	-0,0087	0,0062	1,9773	0,1597
PO	*štúdium*	0,9552	0,3453	7,6541	0,0057
UT	*štúdium*	3,0282	0,4891	38,3302	0,0000
STR	*štúdium*	1,5483	0,4127	14,0730	0,0002
STVR	*štúdium*	1,2215	0,4055	9,0757	0,0026
PIA	*štúdium*	1,2077	0,4594	6,9102	0,0086
t	*oznamy*	-0,3834	0,2150	3,1811	0,0745
t²	*oznamy*	0,0125	0,0074	2,8724	0,0901
PO	*oznamy*	1,6510	0,5167	10,2086	0,0014
UT	*oznamy*	3,0491	0,6270	23,6456	0,0000
STR	*oznamy*	1,5906	0,5846	7,4017	0,0065
STVR	*oznamy*	0,1425	0,6768	0,0443	0,8332
PIA	*oznamy*	-1,2419	1,1522	1,1617	0,2811

Fáza modelovania



Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327

Fáza vyhodnotenia výsledkov

Za predpokladu, že očakávané početnosti sú dostatočne veľké pre porovnanie aktuálneho modelu so saturovaným modelom, ktorý odhaduje pravdepodobnosti nezávisle pre $i = 1, 2, \ldots, 23$ môžeme použiť

$$LR(\hat{\pi}) = 2 \sum_{i=0}^{23} \sum_{j=1}^{J} y_{ij} \, ln \frac{y_{ij}}{\hat{y}_{ij}},$$

$$sv = 10\,476; LR\ stat. = 5\,460{,}54; LR\ stat./sv = 0{,}5212$$

1. Určenie empirických početností prístupov y_{ij}

2. Odhad teoretických početností prístupov

$$\hat{y}_{ij} = \hat{\pi}_{ij} \sum_{j} y_{ij}$$

3. Vizualizácia rozdielov empirických a teoretických početností

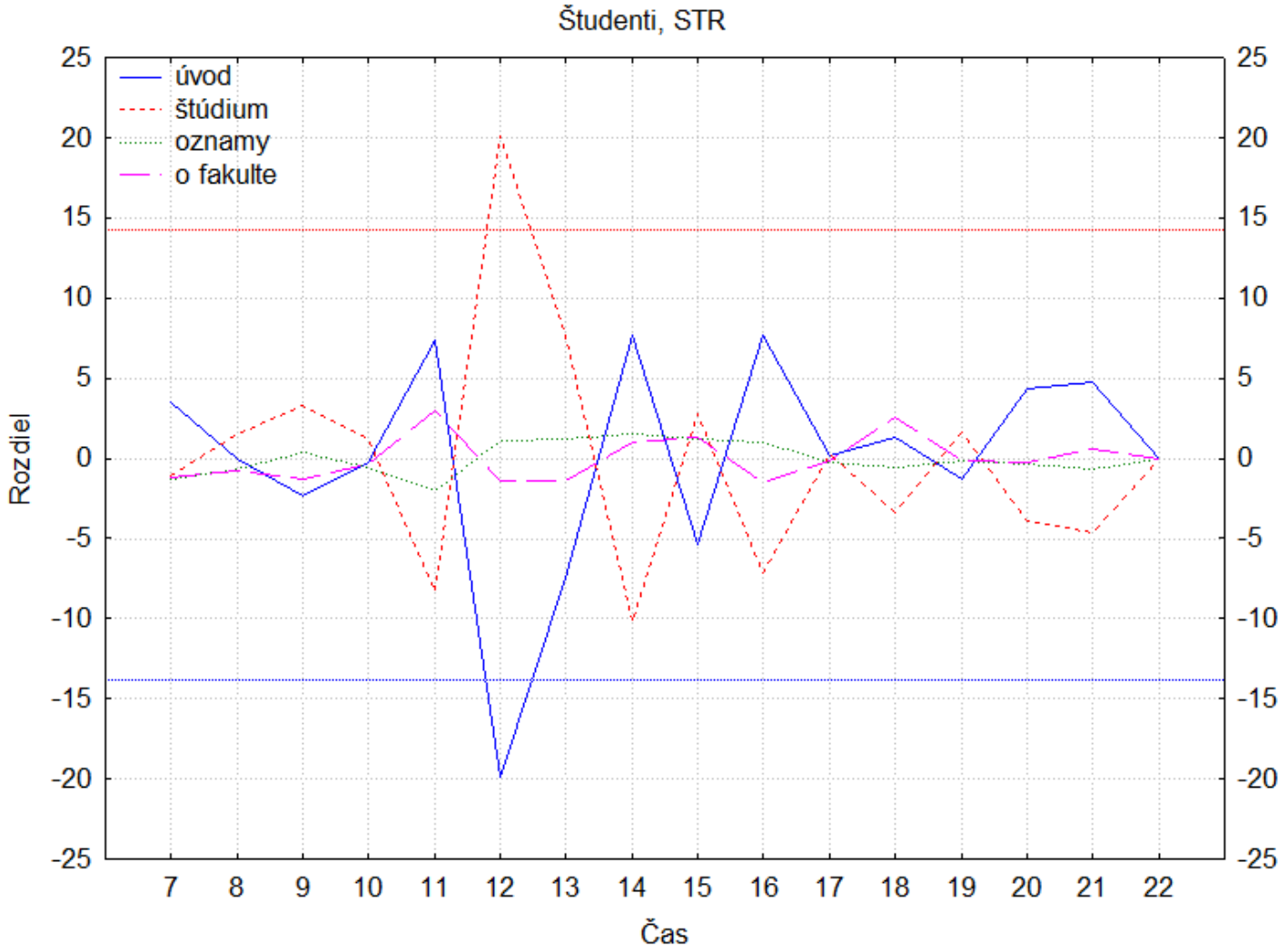
$$d_{ij} = y_{ij} - \hat{y}_{ij}$$

4. Identifikácia extrémnych hodnôt

$$d_{ij} > \bar{d}_j \pm 2s$$

PRÍKLAD:

Fáza vyhodnotenia výsledkov



Fáza vyhodnotenia výsledkov

5. Výpočet empirických relatívnych početností prístupov

$$p_{ij} = \frac{y_{ij}}{\sum_j y_{ij}}$$

6. Porovnanie rozdelenia pravdepodobnosti empirických relatívnych početností prístupov a odhadnutých pravdepodobností výberu webovej časti j v čase i

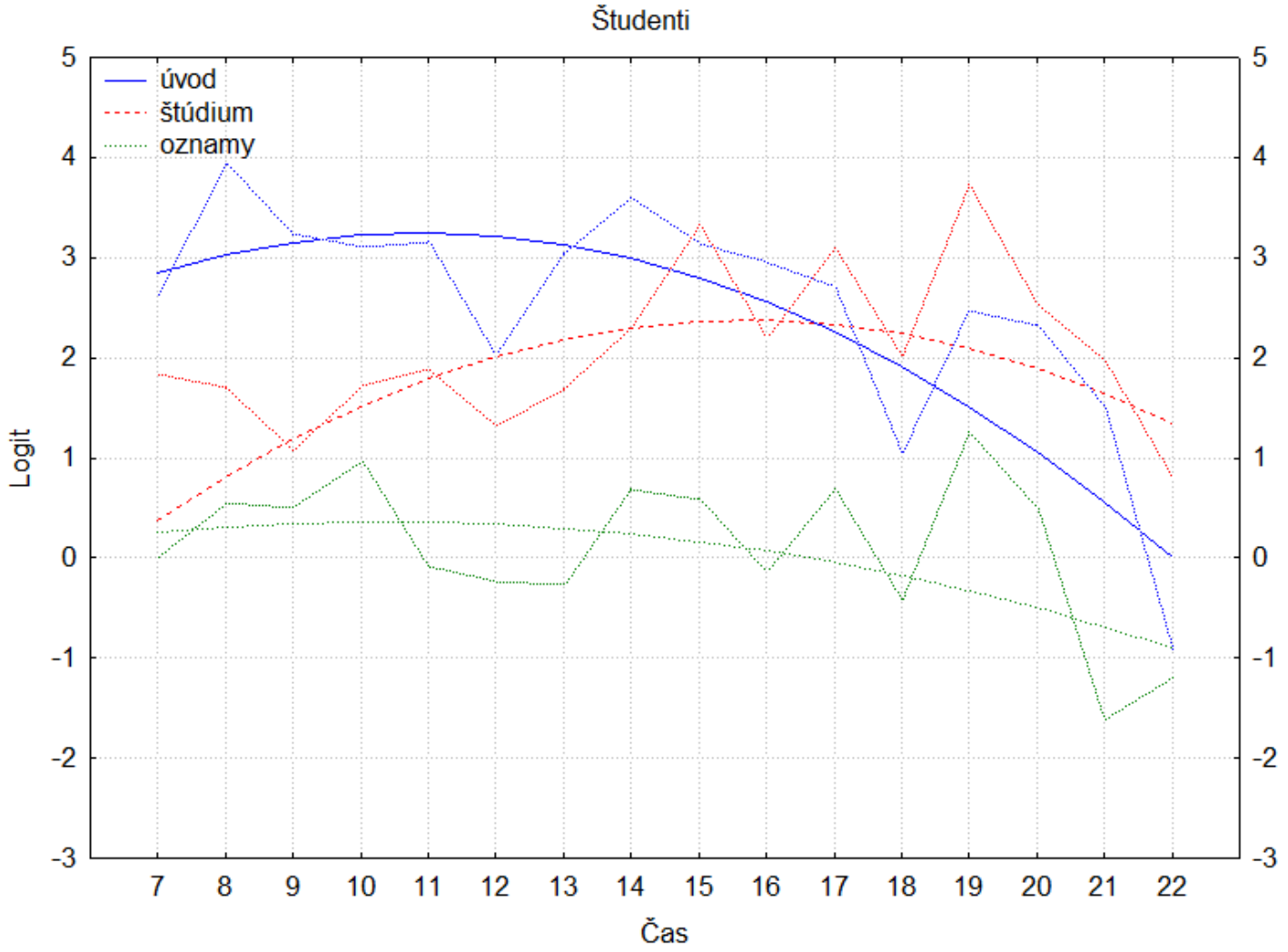
$$r_{ij} = p_{ij} - \hat{\pi}_{ij}, \mathrm{H0}: F(-r) = 1 - F(r)$$

7. Výpočet empirických logitov

$$h_{ij} = \ln \frac{p_{ij}}{p_{iJ}}, j = 1, 2, \ldots, J-1, i \in \{0, 1, \ldots, 23\}$$

8. Vizualizácia empirických a teoretických logitov pre jednotlivé webové časti, okrem referenčnej

Fáza vyhodnotenia výsledkov



Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327

ĎAKUJEM ZA POZORNOSŤ

Innovation and support of doctoral study program no. CZ.1.07/2.2.00/28.0327

Katedra informatiky
FPV UKF v Nitre

esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
University of Pardubice
INVESTMENTS IN EDUCATION DEVELOPMENT