Výber distribúcie Linuxu

Peter Zelenay
Ivana Šupalová

25.5.2006

Predspracovanie dát (1)

- Zdroj dát : www.abclinuxu.cz - výsledky ankety o najobľúbenejšiu distribúciu linuxu roku 2006
- Počet záznamov: 6625
- Tvar záznamu: .xml subor

```
<anketa>
<screen id="START">
<distribuce>fedora</distribuce>
<fedora_verze>5</fedora_verze>
<platforma>x86</platforma>
<osbitu>32</osbitu>
<archbitu>32</archbitu>
<zamestnani>zamestnanec</zamestnani>
<vek>26-30</vek>
<pohlavi>muz</pohlavi>
<vzdelani>maturita</vzdelani>
<pouziti>desktop</pouziti>
<znameni>vahy</znameni>
<jineos>WindowsNT</jineos>
<casto>napul</casto>
<doba>dva roky</doba>
</screen>
</anketa>
```

Predspracovanie dát (2)

- Pomocou DOM parseru vytvorená sql databáza tvaru:

dist	znamenie	casto	jinyos	doba	pouziti	zamestnani	vzdelani	vek	pohlavi
debian	blizenec	vyhradne	WindowsNT	rok	notebook	zamestnanec	maturita	20-25	muz
	rak	vyhradne	NetBSD	sedm let		podnikatel	maturita	31-40	muz
fedora		prevazne	WindowsNT	dele	desktop	zamestnanec	doktorat	31-40	muz
gentoo	byk	napul	FreeBSD	dva roky	desktop	student	maturita	20-25	muz
debian		vyhradne	zadny	rok	desktop	zamestnanec	vysoka skola	26-30	muz
debian	lev	prevazne	WindowsNT	tri roky	desktop	zamestnanec	vysoka skola	20-25	muz
mandriva	rak	prevazne	WindowsNT	pet let	desktop	zamestnanec	vysoka skola	31-40	zena
mandriva	blizenec	prevazne	WindowsNT	ctyri roky		zamestnanec	vysoka skola	20-25	muz
suse	strelec	obcas	WindowsNT	pul roku	desktop	student	maturita	15-19	muz
gentoo	strelec	prevazne	WindowsNT	dva roky	desktop	student	maturita	20-25	muz

Predspracovanie dát (3)

- Odfiltrovanie záznamov s chýbajúcimi údajmi, zostalo 5000

dist	znamenie	casto	jinyos	doba	pouziti	zamestnani	vzdelani	vek	pohlavi
ubuntu	kozoroh	obcas	WindowsNT	pul roku	server	zamestnanec	vysoka skola	26-30	muz
mandriva	beran	prevazne	WindowsNT	dva roky	desktop	zamestnanec	maturita	20-25	zena
mandriva	beran	obcas	WindowsNT	pul roku	desktop	student	maturita	20-25	muz
debian	vodnar	napul	WindowsNT	rok	desktop	zamestnanec	vysoka skola	26-30	muz
slackware	beran	obcas	WindowsNT	rok	desktop	zamestnanec	maturita	26-30	muz
suse	panna	minimalne	WindowsNT	tri roky	desktop	zamestnanec	vysoka skola	20-25	muz
debian	lev	minimalne	Windows95	dva roky	desktop	zamestnanec	vysoka skola	31-40	muz
gentoo	ryby	prevazne	WindowsNT	dva roky	notebook	student	zakladni	15-19	muz
gentoo	byk	prevazne	Solaris	deset let	desktop	duchodce	zakladni	15-19	zena
ubuntu	blizenec	napul	FreeBSD	pet let	desktop	podnikatel	maturita	51-60	muz

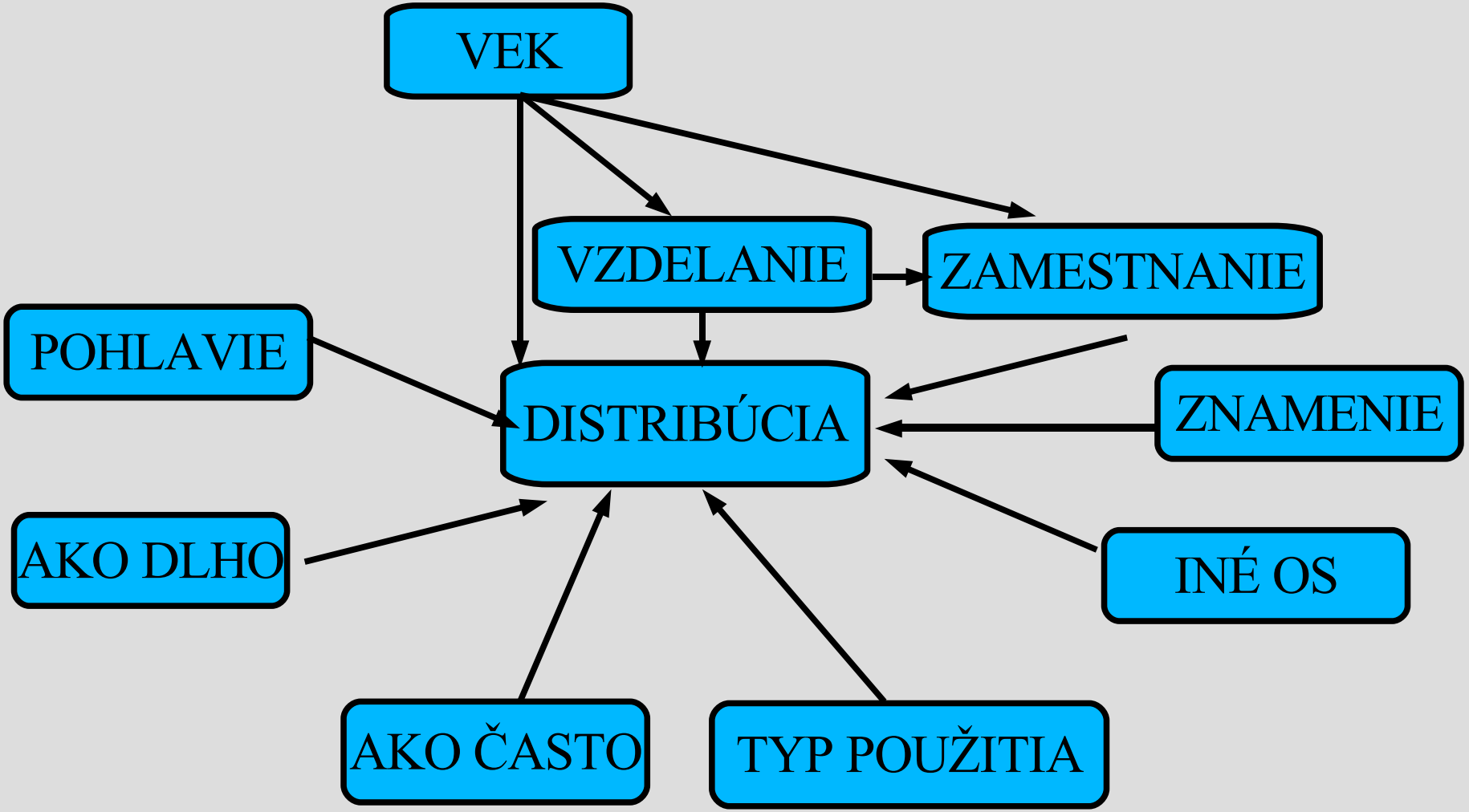
Naivní Bayesovský klasifikátor

- Předpoklad:
  - Jednotlivá pozorování jsou podmíněně nezávislá
- Hledáme maximum z:

$$\max_{t} \left(P(H_t) \prod_{k=1} P(E_k | H_t) \right)$$

- Výsledek:
  - Úspěšnost pouze 30% :-(

Bayesovská sieť (1)



Bayesovská sieť (2)

- Združená pravdepodobnostná distribúcia celej siete: $P(u_1, \ldots, u_n) = \prod P(u_i | rodičia(u_i))$

P(iné OS, ako dlho, ako často, pohlavie, znamenie, typ použitia, vek, vzdelanie, zamestnanie, distribúcia) = P(iné OS)*P(ako dlho)*P(ako často)*P(pohlavie)* *P(znamenie)*P(typ použitia)*P(vek)* *P(vzdelanie | vek)*P(zamestnanie | vek, vzdelanie)* *P(distribúcia | iné OS, ako dlho, ako často, pohlavie, znamenie, typ použitia, vek, vzdelanie, zamestnanie)

Kauzálna inferencia (1)

- Otázka: Aká je pravdepodobnosť výberu určitej distribúcie linuxu na základe vlastností užívateľa?
- Metóda: kauzálna inferencia, známa štruktúra siete, veličiny plne pozorovateľé
- Cieľ: Nájsť maximálne vierohodný odhad vzhľadom k trénovacím dátam

$$P(H_t|E_1,\dots,E_k)=\frac{(n(H_t\cap E_1\cap,\dots,\cap E_k))}{(n(E_1\cap,\dots,\cap E_k))}$$

Kauzálna inferencia (2)

- Hypotéza H : distribúcia

  Debian, SUSE, Mandriva, Ubuntu, Gentoo, Fedora, Slackware, Red hat, Arch, Cent OS, Aurox, ostatné

- Pozorovania E1,...,Ek :

  zamestnanie – *dôchodca, nezamestnaný, podnikateľ, študent, zamestnanec*

  ako často používa linux -*minimálne, občas, polovične, prevažne, výhradne*

  koľko rokov používa linux – *0.5,1,...,10, dlhšie*

  aké ďalšie OS používa – *WindowsNT, Windows95, FreeBSD,DOS, Solaris, MacOS X, HPUX, AIX, OpenBSD, NetBSD, IRIX, MacOS, žiadny*

Kauzálna inferencia (3)

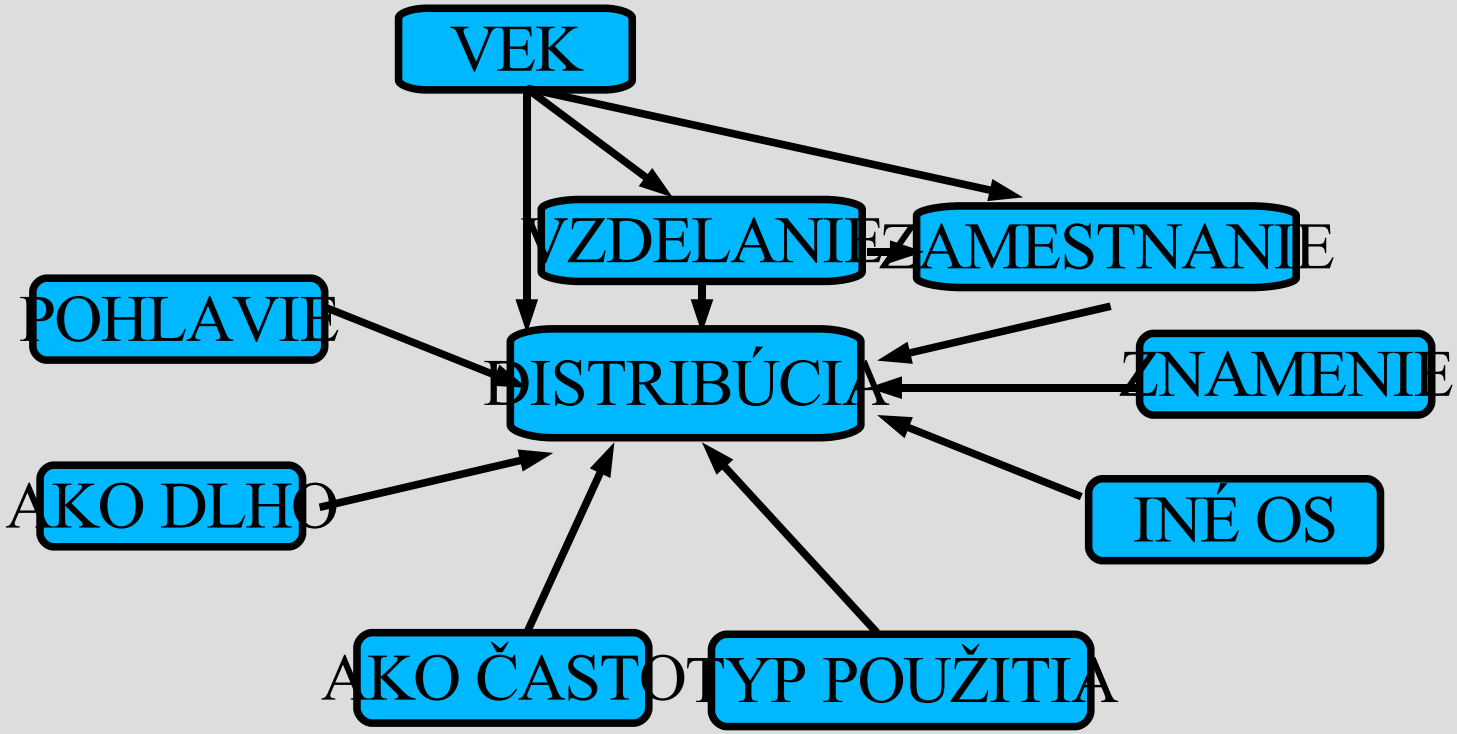
pohlavie – *muž, žena*
vzdelanie – *základné, učeň, maturita, vysoká škola, doktorát*
vek - *<15,15-19,20-25,26-30,31-40,41-50,51-60, >60*
typ použitia – *desktop, notebook, server*
znamenie zverokruhu

Implementace (1)

- Třída Net
  - Spočte sdruženou pravděpodobnostní distribuci sítě pro každou distribuci a vráti distribuci s maximální pravděpodobností.

Implementace (2)

- Třída Point
  - Spočte hodnotu jedného uzlu sítě



Výsledok

- Testovacie dáta – vopred vytriedené záznamy pri filtrovaní (recyklácia)

distribúcia	
skutočne zvolená	**nami odporúčaná**
gentoo	gentoo
ubuntu	ubuntu
suse	suse
gentoo	debian
debian	debian
ubuntu	ubuntu
mandriva	mandriva
gentoo	gentoo
mandriva	mandriva

Záver

- Podľa výsledkov vychádza úspešnosť vhodnosti nami odporúčanej distribúcie pre klienta až **88%** !!!
- Toto číslo by bolo pravdepodobne ešte vyššie, keby sme do siete zahrnuli viac závislostí
- Výsledky by bolo možné využiť pri závádzaní nových features do distribúcií (pre určitú charakteristickú skupinu užívateľov)

Zdroje

- dáta : http://www.abclinuxu.cz/
- inšpirácia: http://www.zegeniestudios.net/ldc/