



## **Filtrování internetových stránek podle obsahu 602LAN SUITE 2004 Content Filter**



**Přístup odepřen** \_\_\_\_\_

Braňte Vaším zaměstnancům či studentům v přístupu  
na internetové stránky s nevhodným obsahem!

## Obsah

<b>Kontakty .....</b>	<b>2</b>
<b>Úvod do problematiky filtrování stránek podle obsahu .....</b>	<b>3</b>
Internet jako zdroj informací .....	3
Proč filtrovat přístup na některé stránky.....	3
Co říkají statistiky .....	3
<b>Význam filtrování stránek podle obsahu .....</b>	<b>4</b>
Zvýšení produktivity práce .....	4
Zamezení přístupu k nevhodnému obsahu .....	4
Zefektivnění využívání internetového připojení .....	4
<b>Srovnání metod používaných pro filtrování stránek .....</b>	<b>5</b>
Statické metody .....	5
Dynamické metody.....	5
<b>Jak funguje Content Filter v 602LAN SUITE.....</b>	<b>5</b>
Upozornění .....	5
Rozpoznávání obsahu – Artificial Content Recognition™ .....	5
Princip zpracování požadavků uživatelů .....	6
Základní konfigurace – definice pravidel.....	7
Filtrování stránek podle 22 kategorií obsahu .....	8
Manuálně definované blokace .....	8
Výjimky z filtrování .....	8
Podpora standardů ICRA™ a SafeSurf™ .....	9
Přehledné statistiky .....	9
Licencování.....	9

## Kontakty

Další informace naleznete na <http://www.602.cz/>

Software602 a.s.

P.O. Box 1, 140 00 Praha 4

infolinka: 222 011 602, [info@602.cz](mailto:info@602.cz)

# Úvod do problematiky filtrování stránek podle obsahu

## Internet jako zdroj informací

V dnešní době se stalo využívání internetu pro většinu uživatelů počítačů naprostou samozřejmostí. Řada užitečných informací začíná být k dispozici kromě klasické papírové podoby i v elektronické formě, a některé informace jsou dokonce dostupné již jen elektronicky. Bez schopnosti vyhledávat a získávat informace na internetu se proto dnes v podstatě nikdo neobejde.

Internet lze bez nadsázky charakterizovat jako zdroj nepředstavitelného množství informací nejrůznějšího charakteru, často volně přístupných bez jakéhokoliv omezení, což je jak jeho hlavní výhodou, tak také nevýhodou.

## Proč filtrovat přístup na některé stránky

Přístup na internet je dnes nutností jak ve firmách, tak ve školách a celé řadě dalších organizací. Problém ale tkví v tom, jak umožnit zaměstnancům či studentům volný přístup k rozsáhlým informačním zdrojům internetu, ale jak je i vlastní organizaci na druhé straně chránit před potenciálními riziky, která číhají na internetových stránkách s pochybným nebo nevhodným obsahem.

### Firmy a organizace

Jedním z důvodů, proč filtrovat přístup na určité stránky může být produktivita práce. Ta zákonitě klesá už tím, že se zaměstnanci v pracovní době věnují místo práce surfování po internetu. Další rizika ale spočívají v možnosti napadení inkriminovaného počítače a následně i celé počítačové sítě škodlivým software (např. spyware nebo počítačovými viry), který se často na stránkách s pochybným nebo dokonce nelegálním obsahem vyskytuje. Následné řešení této situace a odstraňování podobných napadení, která mohou někdy vyústit až v poškození nebo ztrátu důležitých dat, může negativně ovlivnit činnost celé firmy.

### Školy

Stejně jako firmám a organizacím, hrozí také školám napadení jejich počítačových sítí škodlivým software ze stránek s pochybným obsahem. Také jistě není žádoucí, aby se studenti místo studia věnovali prohlížení internetových stránek nebo třeba „chatovali“ se svými kamarády. Na rozdíl od firem, jejichž zaměstnanci jsou dospělí lidé, kteří mají plnou zodpovědnost za své konání a jsou většinou schopni rozlišit pravdivost informací, které na internetu naleznou, musí mít školy možnost účinně blokovat přístup studentů na stránky se závadným obsahem – ať už se jedná o pornografii nebo třeba stránky propagující rasismus.

### Shrnutí

Z předchozích odstavců vyplývá, že bez ohledu na to, jak Vy sami pojem „stránka s nevhodným obsahem“ definujete, musíte mít k dispozici takové nástroje, abyste mohli zajistit efektivní využívání Vašeho internetového připojení. Jednoduše řečeno – abyste měli možnost ovlivňovat kdo, kdy a k jakému druhu informací má, či spíše nemá mít přístup. Nástroj, který se pro tento účel používá, se nazývá **filtr obsahu** nebo také **content filter**.

## Co říkají statistiky

Žebříčky nejčastěji vyhledávaných slov a frází za loňský rok zveřejněné známými vyhledávači Google a Seznam jednoznačně ukazují, že většina lidí na internetu nejčastěji vyhledává informace, které se jen stěží vztahují k náplni jejich práce nebo studia.

Tab. 1 - Žebříčky nejvyhledávanějších slov na internetu

<u>Google Popular Queries 2004</u>	<u>Seznam TOP 10 za rok 2003</u>
1. britney spears	1. sms
2. paris hilton	2. hry
3. christina aguilera	3. Nokia
4. pamela anderson	4. práce
5. chat	5. seznamka
6. games	6. chat
7. carmen electra	7. Praha
8. orlando bloom	8. CD
9. harry potter	9. MP3
10. mp3	10. mobil
<u>Google Popular Tech Stuff 2004</u>	<u>Seznam – Slavné ženy za rok 2004</u>
1. wallpaper	1. Ester Ládová
2. kazaa	2. Aneta Langerová
3. mp3	3. Britney Spears
4. spybot	4. Avril Lavigne
5. linux	5. Šárka Vaňková

## Význam filtrování stránek podle obsahu

### Zvýšení produktivity práce

Z výsledků nedávno provedeného průzkumu mezi uživateli internetu vyplynulo, že přes 60 % z nich využívá internet k soukromým účelům i během pracovní doby. Zavedením filtrování stránek podle obsahu lze výrazně zredukovat možnosti neefektivního trávení pracovní doby při čtení zpráv, prohlížení sportovních výsledků nebo třeba sledování stránek s erotickým obsahem. Rovněž lze takto omezit přístup k soukromým e-mailům, se kterými uživatelé pracují pomocí webových rozhraní.

### Zamezení přístupu k nevhodnému obsahu

Ačkoliv nelze nevhodný obsah jednoznačně specifikovat, neboť míra nevhodnosti se liší v závislosti na kontextu a konkrétním uživateli, lze pomocí vhodně nakonfigurovaného filtrování stránek podle obsahu účinně omezit přístup uživatelů (zaměstnanců, studentů apod.), ke stránkám obsahujícím pro ně nevhodný obsah. Například zaměstnancům lze omezit přístup na stránky umožňující download nelegálního software nebo hudby, zatímco studenty je možné navíc chránit před přístupem na stránky s pornografií, propagací drog nebo rasismu.

### Zefektivnění využívání internetového připojení

Filtrování přístupu na stránky podle jejich obsahu může mít také pozitivní vliv na efektivitu využívání internetového připojení. Uvážíme-li, že velikost dnešních internetových stránek včetně všech stylů a grafických prvků dosahuje řádově 100 KB, a vynásobíme-li toto číslo počtem uživatelů a dobou, kterou stráví na internetu za jiným než pracovním účelem, pak je snadné dovodit, že důsledné zavedení restrikcí může ve výsledku vést ke značnému snížení objemu „zbytečně“ přenášených dat a ke zrychlení přístupu na internet pro ostatní uživatele.

## Srovnání metod používaných pro filtrování stránek

Pro filtrování stránek podle obsahu se v zásadě používají dva druhy metod – statické a dynamické. Statické metody jsou založené na databázi kategorizovaných URL, zatímco dynamické metody provádějí v reálném čase rozbor obsahu požadovaných stránek.

### Statické metody

Filtrování stránek na základě **databáze klasifikovaných URL** poskytuje teoreticky vždy 100% správné výsledky. Slabinou tohoto řešení ale je, že jeho kvalita je přímo úměrná kvalitě obsahu používané databáze URL. I ta největší databáze URL totiž pokrývá pouze zlomek volně přístupných stránek, a to zpravidla těch, které lze nalézt pomocí vyhledávačů. Stránky, které nejsou vyhledávači zaindexovány anebo jsou přístupné například pouze přes heslo, tyto databáze většinou neobsahují.

Databáze URL obvykle vznikají za pomoci automatizovaného software, jehož výsledky jsou kombinovány s prací lidských editorů, kteří klasifikují obsah jednotlivých serverů a stránek. Z toho vyplývá, že je prakticky nemožné získat takovou databázi URL, která by dostatečně pokrývala všechny dostupné internetové stránky, a problém rovněž nastává se zajištěním aktualizace obsahu databáze.

### Dynamické metody

Dynamické metody, na rozdíl od statických, **analyzují a kategorizují obsah internetových stránek v reálném čase** tak, jak na ně jednotliví uživatelé přistupují. Nezáleží tedy na tom, zda daná stránka na internetu existuje již 5 měsíců nebo jen 5 minut. Právě schopnost analýzy v reálném čase umožňuje dynamickým filtrům obsahu držet krok se současným exponenciálním rozvojem internetu. Analýza každé stránky je přitom prováděna pomocí složitých algoritmů a matematických postupů, které zohledňují nejen prostý textový obsah stránky, ale i uspořádání objektů na stránce a její grafickou podobu, a které úspěšně minimalizují možnost chybné klasifikace analyzované stránky.

Na rozdíl od statických metod nemůže také při použití dynamických filtrů obsahu nastat situace, ve které by přístup na některou požadovanou stránku byl zablokovaný jen proto, že byla označena jako neklasifikovaná, neboť se nenacházela v databázi klasifikovaných URL.

## Jak funguje Content Filter v 602LAN SUITE

### Upozornění

Aby bylo filtrování stránek modulem Content Filter v 602LAN SUITE funkční, je třeba pro přístup na internet využívat služeb proxy serveru obsaženého v 602LAN SUITE. Internetové prohlížeče na stanicích musí být tedy nakonfigurovány tak, aby pro přístup na internet tuto proxy využívaly.

### Rozpoznávání obsahu – Artificial Content Recognition™

Jádro modulu Content Filter v 602LAN SUITE tvoří technologie Artificial Content Recognition™ (ACR) od firmy PureSight Inc. Technologie ACR se skládá z celé řady algoritmů na bázi umělé inteligence, pomocí kterých provádí analýzu a klasifikaci internetových stránek v reálném čase.



Obr. 1 - Schéma zpracování internetové stránky technologií ACR

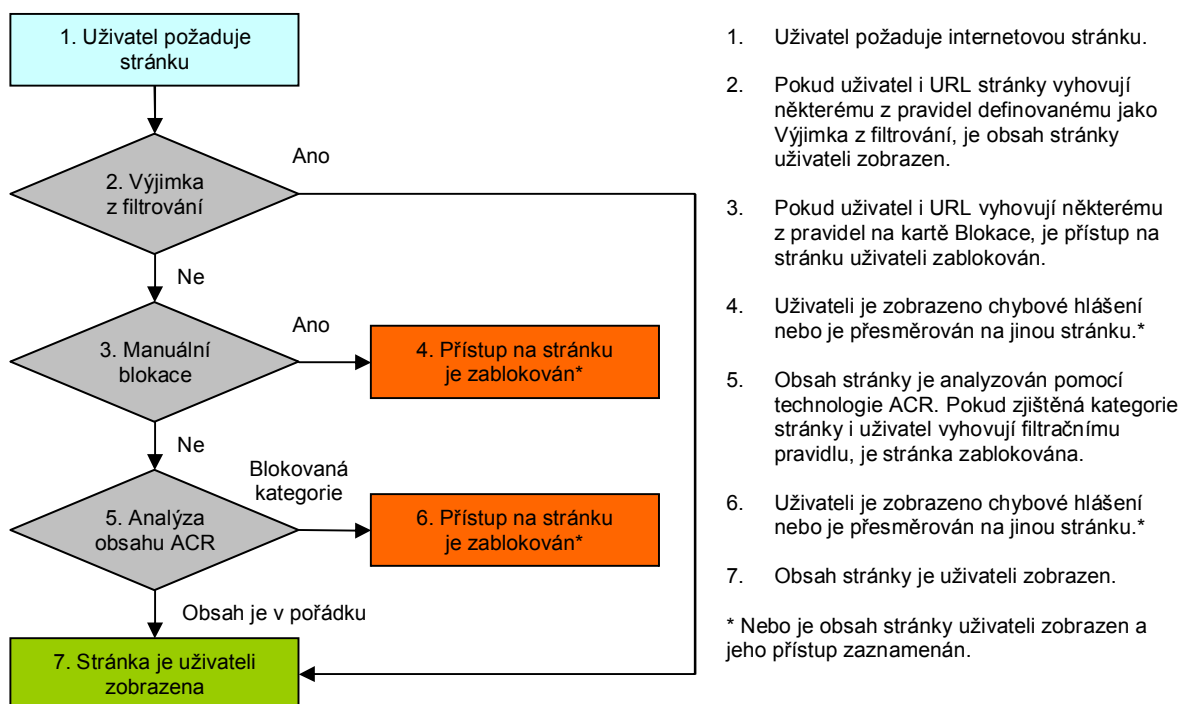
## Princip zpracování internetové stránky pomocí ACR

1. Každá požadovaná internetová stránka je nejprve pomocí HTML parseru převedena do podoby vektoru o stovkách numerických parametrů. V této fázi jsou analyzovány mj. následující parametry HTML stránky:  
**Ne-textové informace:** barva pozadí, druhy písma, barva písma, velikost písma, počet odkazů, počet obrázků, velikost obrázků, počet rámců, průměrný počet slov, počet slov, speciální znaky, meta tagy.  
**Textové informace:** název URL, slova obsažená na stránce, obsah meta tagů.
2. Tyto parametry jsou dále zpracovány pomocí algoritmů umělé inteligence, které hledají závislosti mezi jednotlivými parametry, a jejichž výsledkem je další vektor obsahující již jen 20 až 25 specifických vlastností.
3. Tento vektor je na závěr zpracován pomocí neuronové sítě, přičemž výsledkem je ohodnocení dané HTML stránky jejími koordináty ve vícerozměrném prostoru, ve kterém se již nachází několik „oblaků“ koordinátů reprezentujících jednotlivé kategorie obsahu stránek (např. sport, hazard, pornografie a další).
4. Porovnáním získaných koordinátů s koordinátami předdefinovaných „oblaků“ zjistí ACR kategorii požadované stránky. V závislosti na konfiguraci poté 602LAN SUITE buď danému uživateli přístup na požadovanou stránku umožní nebo naopak zablokuje.

## Přesnost zjišťování kategorie stránky pomocí ACR

Firma PureSight Inc. věnuje „trénování“ technologie ACR velkou pozornost. Aby technologie ACR správně rozlišovala například mezi stránkami s pornografickým obsahem a stránkami obsahujícími informace týkající se sexuální výchovy, je nejprve „trénována“ na vzorcích HTML stránek z obou těchto kategorií. U každé takové stránky je v průběhu „učení“ označeno, do jaké kategorie právě analyzovaná stránka patří. Po ukončení tohoto „tréninku“ pak ACR umí správně rozlišovat mezi stránkami, na kterých se sice vyskytuje podobný obsah, ale patří do naprosto rozdílných kategorií. Technologie ACR navíc obsahuje sady slovníkových výrazů v řadě jazyků, které umožňují přesnější převod obsahu stránek na numerické parametry a vyhodnocení závislostí mezi nimi.

## Princip zpracování požadavků uživatelů



## Základní konfigurace – definice pravidel

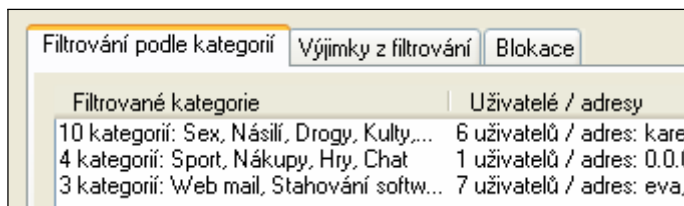
Konfigurace modulu Content Filter se provádí pomocí tzv. pravidel, která určují kdo, kdy a kam má nebo nemá mít povolen přístup. Těchto pravidel lze vytvořit prakticky libovolné množství pro různé kategorie stránek i různé uživatele tak, aby výsledná konfigurace zcela odpovídala požadavkům a struktuře organizace.

### Pro koho bude pravidlo platit

V rámci každého pravidla lze snadno specifikovat, pro koho bude dané pravidlo platit. To lze definovat buď pomocí konkrétních uživatelských jmen nebo pomocí IP adres jednotlivých stanic v síti, přičemž lze tyto možnosti vzájemně kombinovat. Kromě konkrétních IP adres je možné platnost pravidla omezit na rozsah IP adres nebo jednotlivé podsítě.

Je třeba si uvědomit, že pravidla definovaná pro konkrétní uživatele fungují, pouze pokud je 602LAN SUITE nakonfigurována tak, aby používala tzv.

autentifikaci do proxy (přístup na internet přes jméno a heslo). Pravidla definovaná pomocí IP adres počítačů je možné použít vždy, ovšem zvědavější uživatel by mohl změnou IP adresy své stanice tento parametr „obejít“.



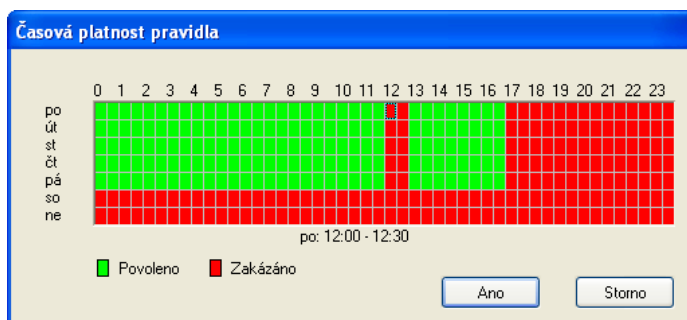
Filtrované kategorie	Uživatelé / adresy
10 kategorií: Sex, Násilí, Drogy, Kulty,...	6 uživatelů / adres: karel
4 kategorií: Sport, Nákupy, Hry, Chat	1 uživatelů / adres: 0.0.0
3 kategorií: Web mail, Stahování softw...	7 uživatelů / adres: eva,

Obr. 2 - Seznam zadaných pravidel

### Časová platnost pravidla

Platnost každého pravidla lze kromě uživatelů/IP adres navíc omezit časem. Standardně jsou všechna pravidla platná neomezeně, ale v případě potřeby je možné omezit jejich platnost během týdne v půlhodinovém intervalu.

Na příkladu vpravo vidíte pravidlo, které platí pouze od pondělí do pátku, a to vždy od 0:00 do 12:00 hodin, a pak od 13:00 do 17:00 hodin. Přes poledne, večer a ani o víkendu se tedy dané pravidlo nepoužívá.



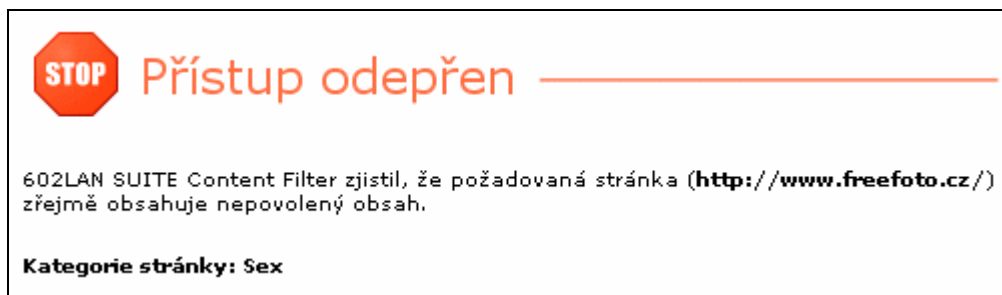
Obr. 3 - Nastavení časového omezení platnosti

### Akce

U pravidel, která mají za úkol blokovat přístup na některé stránky, lze navíc specifikovat akci, která nastane, pokud budou splněny ostatní podmínky zadané v definici pravidla.

Při splnění zadaných podmínek je možné přístup na požadovanou stránku zablokovat a uživateli zobrazit chybové hlášení nebo lze uživatele přesměrovat na jinou stránku (např. na firemní server nebo třeba na vysvětlující stránku vytvořenou za tímto účelem správcem sítě). Poslední možností je pouhé sledování nežádoucích aktivit uživatele prostřednictvím zápisu do logu s tím, že uživatel obsah požadované stránky normálně obdržel.

Aktivitu uživatele zaznamenává Content Filter do logu vždy, tj. i v případě, že je nastavena libovolná jiná akce. Z logu je tedy možné zpětně sledovat aktivitu jednotlivých uživatelů prostřednictvím přehledných statistik generovaných doplňkovou aplikací ActiveReports.



Obr. 4 - Hlášení zobrazované při zablokování přístupu na stránku

## Filtrování stránek podle 22 kategorií obsahu

Content Filter v 602LAN SUITE podporuje filtrování až 22 kategorií obsahu stránek, které je založeno na rozpoznávání obsahu technologií ACR. Tyto kategorie lze navíc rozšířit o manuálně udržované seznamy blokových stránek a serverů nebo je možné udělit pro přístup na některé stránky tzv. výjimky.

### Přehled podporovaných kategorií pro filtrování stránek

- Sex
- Hazardní hry
- Sport
- Hledání zaměstnání
- Nákupy
- Akcie
- Násilí
- Web mail
- Zbraně
- Nenávist
- Drogy
- Hry
- Zprávy
- Stahování software
- Cestování
- Kulty
- Warez
- Aukce
- Seznamování
- Chat
- Zdraví
- Diskuze

### Definice filtrovacího pravidla

Konfigurace filtrování stránek podle obsahu technologií ACR se provádí pomocí pravidel, jejichž společné parametry již byly popsány výše. V rámci každého filtrovacího pravidla stačí pouze zaškrtnout kategorie stránek, na které chcete omezit přístup zvolených uživatelů.

Filtrovat přístup k webovým stránkám zařazeným do těchto kategorií:		
<input checked="" type="checkbox"/> Sex	<input checked="" type="checkbox"/> Zbraně	<input checked="" type="checkbox"/> Warez
<input checked="" type="checkbox"/> Hazardní hry	<input checked="" type="checkbox"/> Nenávist	<input type="checkbox"/> Aukce
<input type="checkbox"/> Sport	<input checked="" type="checkbox"/> Drogy	<input checked="" type="checkbox"/> Seznamování
<input type="checkbox"/> Hledání zaměstnání	<input type="checkbox"/> Hry	<input checked="" type="checkbox"/> Chat
<input type="checkbox"/> Nákupy	<input type="checkbox"/> Zprávy	<input type="checkbox"/> Zdraví
<input type="checkbox"/> Akcie	<input type="checkbox"/> Stahování software	<input type="checkbox"/> Diskuze
<input checked="" type="checkbox"/> Násilí	<input type="checkbox"/> Cestování	
<input type="checkbox"/> Web mail	<input checked="" type="checkbox"/> Kulty	

Obr. 5 - Definice kategorií v rámci filtrovacího pravidla

## Manuálně definované blokace

Kromě filtrování stránek podle obsahu technologií ACR lze v modulu Content Filter také manuálně zadat adresy stránek nebo serverů, na které má být přístup pro určité uživatele blokován. Na rozdíl od filtrovacích pravidel se zde zadávají URL konkrétních stránek nebo serverů, které mají být blokovány. Pro větší variabilitu lze při zadávání URL používat i zástupné znaky hvězdička (\*) a otazník (?). Při pokusu o přístup na takto blokovanou stránku se opět provede zadaná akce (zablokování, přesměrování nebo jen zaznamenání přístupu).

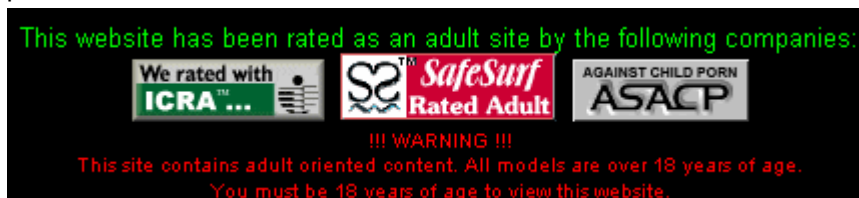
## Výjimky z filtrování

I přes maximální snahu o přesnost technologie ACR se může stát, že Content Filter někdy mylně kategorizuje obsah zcela legitimní stránky a zablokuje na ni přístup, ač má být tato stránka uživatelům normálně přístupná. Nebo může být zapotřebí udělit některému z uživatelů výjimku z nedefinovaných pravidel. Právě pro řešení podobných situací slouží tzv. výjimky z filtrování, které se zadávají také ve

formě pravidel. Jejich výjimečnost spočívá v tom, že pravidla definovaná jako výjimky mají při zpracování přednost před ostatními (filtrovacími a blokovacími) pravidly.

## Podpora standardů ICRA™ a SafeSurf™

Kromě výše popsaných metod podporuje Content Filter filtrování stránek také podle klasifikačních značek organizací [ICRA™](#) a [SafeSurf™](#), kterými někteří autoři internetových stránek sami klasifikují charakter obsahu informací na svých stránkách a které je zejména na zahraničních stránkách poměrně rozšířené.



Obr. 6 - Příklad stránky s klasifikací

## Přehledné statistiky

Doplňková aplikace 602LAN SUITE ActiveReports poskytuje přehledné statistiky provozu a umožňuje snadno monitorovat nežádoucí aktivity jednotlivých uživatelů. Kromě toho slouží k podrobné analýze a zobrazení objemu a struktury dat, která přenesou jednotlivé stanice v síti přes komunikační server. Výsledky analýzy lze pro větší přehlednost zobrazit ve formě tabulek i grafů. Pokud některý z uživatelů překročí stanovené denní limity, může být o tom administrátor automaticky informován prostřednictvím elektronické pošty.

Uživatel	Čas	Velikost [B]	Soubor (URL)
novak	30.3.2005 17:20:09	blokován-Kategorie: Sex (ACR)	http://www5.sex-mission.com/zpornstars/008a/03-30-05/set03.html
novak	30.3.2005 17:20:22	blokován-Kategorie: Sex (UCC)	http://www.hugeboobsclub.com/series/patrol236as/pichm.html
novak	30.3.2005 17:20:30	blokován-Kategorie: Sex (ACR)	http://www.porno-realm.com/pages/oldgal41/galli.html
novak	30.3.2005 17:20:32	blokován-Kategorie: Sex, Sport (ACR)	http://www.freshfat.com/yf78/pichunter.html
novak	30.3.2005 17:20:44	blokován-Kategorie: Sex (STL)	http://vis.sexlist.com:81/
novak	30.3.2005 17:20:45	blokován-Kategorie: Sex (STL)	http://cgi.sexlist.com/counter.cgi
novak	30.3.2005 17:20:50	blokován-Kategorie: Sex, Hazardní hry, ...	http://www.madthumbs.com/
novak	30.3.2005 17:21:00	blokován-Kategorie: Sex (ACR)	http://www.bdsmite.com/bdsm/bl28mar/index148.html
novak	30.3.2005 17:21:19	blokován-Kategorie: Sex (DEM STL)	http://www.ceskyholky.cz/
novak	30.3.2005 17:21:25	blokován-Kategorie: Sex (DEM STL)	http://www.extremy.cz/
novak	30.3.2005 17:29:41	blokován-Kategorie: Sport (ACR)	http://1.im.cz/szn/img/logo_sport.gif
novak	30.3.2005 17:29:45	blokován-Kategorie: Web mail (STL)	http://email.seznam.cz/index.py/embeddedLogin
novak	30.3.2005 17:29:47	blokován-Kategorie: Web mail (STL)	http://email.seznam.cz/
novak	30.3.2005 17:29:56	blokován-Kategorie: Aukce (DEM STL)	http://www.aukro.cz/
novak	30.3.2005 17:30:01	blokován-Kategorie: Akcie (DEM STL)	http://www.akcie.cz/
novak	30.3.2005 17:30:12	blokován-Kategorie: Web mail (DEM STL)	http://mail.atlas.cz/
novak	30.3.2005 17:30:23	blokován-Kategorie: Hry (DEM STL)	http://bonusweb.idnes.cz/obrazek/sims2_idnes
novak	30.3.2005 17:30:42	blokován-Kategorie: Stahování software ...	http://www.download.com/

Obr. 7 - Statistika přístupů na blokované stránky

## Licencování

Modul Content Filter, sloužící pro filtrování přístupu na internetové stránky podle jejich obsahu, je licencován ve shodných krocích jako komunikační server 602LAN SUITE, jehož je součástí. K dispozici jsou tedy varianty pro 3, 10, 25, 100 a neomezený počet uživatelů