

# **Vyhledávání a grafová struktura Webu**

Václav Snášel, Dušan Húsek,  
Hana Řezanková

# Osnova

- Využití strukturních vlastností webu
- Topologická analýza webového prostoru
  - Motivace
  - Grafová analýza
  - Algoritmy pro hledání relevantních dokumentů
- Experimentální výsledky a vizualizace výsledků
  - Zhodnocení výsledků
  - Vizualizace výsledků
- Závěr

# Využití strukturních vlastností webu

Rozsah a dynamičnost webu vede k tomu, že je třeba hledat nové možnosti, jak vyhledávat a jak prezentovat výsledky vyhledávání.

<b>Služba</b>	<b>Počet dotazů (v milionech)</b>
Google	250
Overture	167
Inktomi	80
LookSmart	45
FindWhat	33
Ask Jeeves	20
AltaVista	18
FAST	12

# Využití strukturních vlastností webu

Struktura, jenž se nabízí na první pohled, je grafová struktura webu.

Web můžeme chápat jako rozsáhlý graf.

Znalost grafové struktury webu je důležitá zejména z následujících důvodů:

- návrh nových strategií procházení (crawl) webu,
- analýza chování webu,
- vyhledávání a vizualizace webu,
- predikce nových vlastností webu.

# Topologická analýza webového prostoru

- Přítomnost hypertextových odkazů v rámci kolekce webových stránek nám dovoluje interpretovat „plochou“ kolekci dokumentů jako „plastickou“ strukturu orientovaného grafu, kde webové stránky představují uzly a hypertextové odkazy představují hrany. Rozsáhlé grafy jsou v současné době zkoumány v mnoha oblastech matematiky a informatiky.

# Motivace

- Topologické vlastnosti grafů webových stránek mohou poskytnout důležité informace (metadata) nejen o samotných webových stránkách (uzlech grafu), ale také o podgrafech splňujících různé vlastnosti, kde tyto podgrafy můžeme chápat jako zcela nové, vyšší entity (komunity) uvnitř internetové sítě.
- Tyto nové informace přirozeně poskytují lepší možnosti pro vyhledávání na internetu.

# Grafová analýza

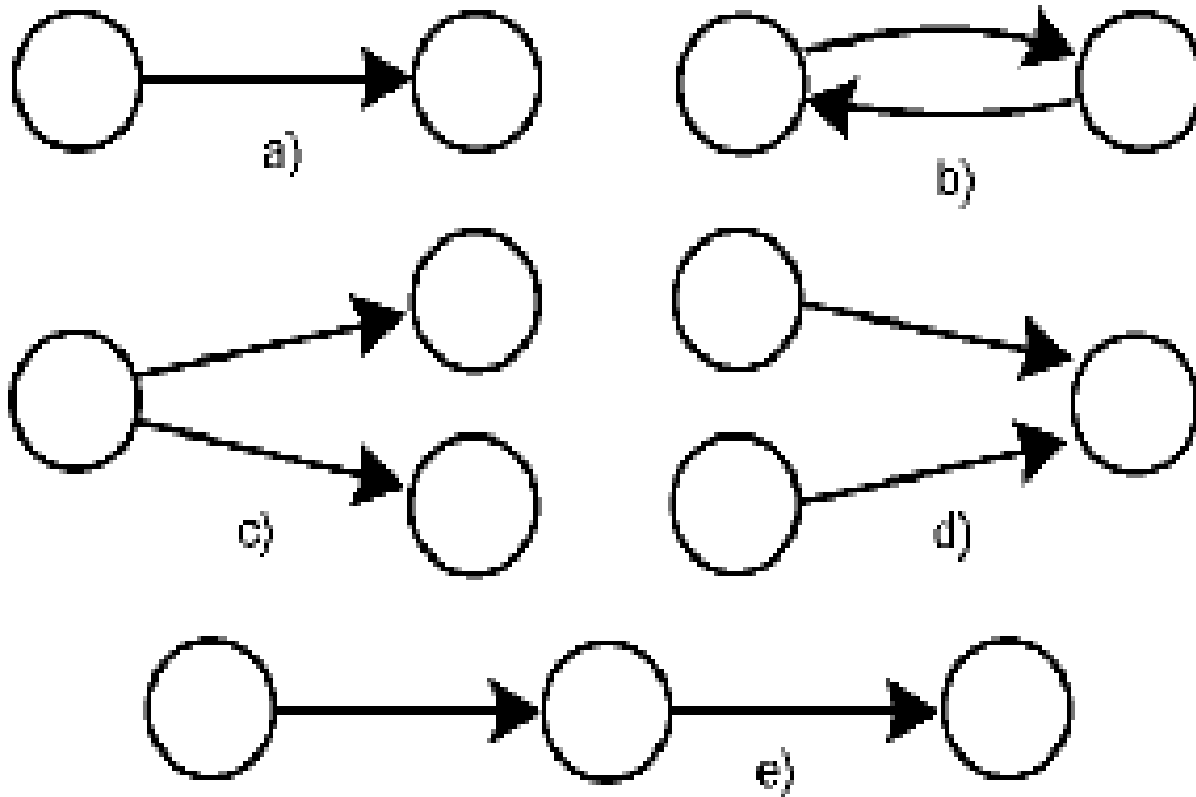
- V hypertextovém prostředí WWW se ustálila vlastní terminologie popisující různá vzájemná propojení stránek, v důsledku lze ale tyto termíny převést na analogické pojmy v kontextu grafové analýzy.
- Každá stránka obsahuje určitý počet odkazů na jiné stránky – *forward links*, přičemž pravděpodobnost, že vrchol má stupeň  $i$ , je proporcionální  $1/i^x$  pro nějaké  $x > 1$ . Odkazy, které směřují na danou stránku z jiných stránek, se nazývají *back links*. Při stažení webové stránky můžeme jednoduše zjistit její *forward links*, ale nezjistíme z ní její *back links*.

# Grafová analýza

Základní tvary v grafu webu jsou následující:

- *Potvrzení stránky* (endorsement) – stránka obsahuje odkaz na jinou (potvrzenou) stránku (viz obr. a).
- *Relevantní stránky* – navzájem se potvrzující stránky (viz obr. b).
- *Společná citace* (co-citation) – stránka se odkazuje na více různých stránek (viz obr. c).
- *Společná volba* (social choice) – na stránku se odkazuje více stránek (obr. d).
- *Nepřímé potvrzení* (transitive endorsement) – stránka  $s_1$  se odkazuje na  $s_2$  a  $s_2$  na  $s_3$ . Potom  $s_1$  (slabě) potvrzuje  $s_3$ . (viz obr. e).

# Základní tvary v grafu webu

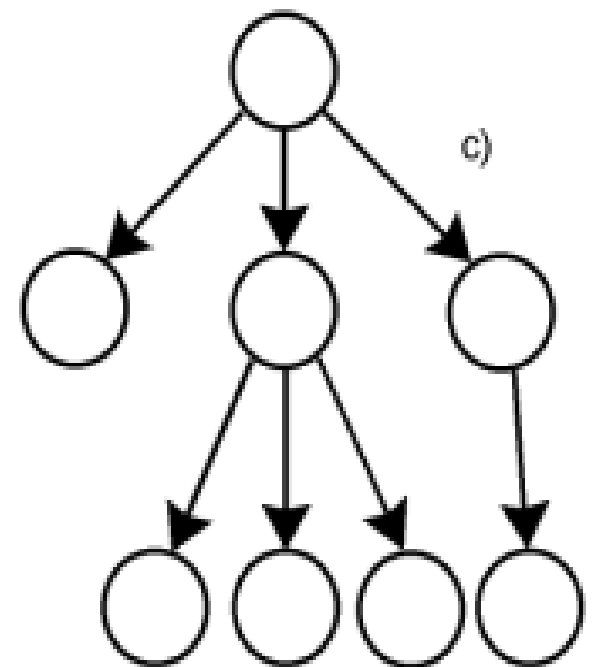
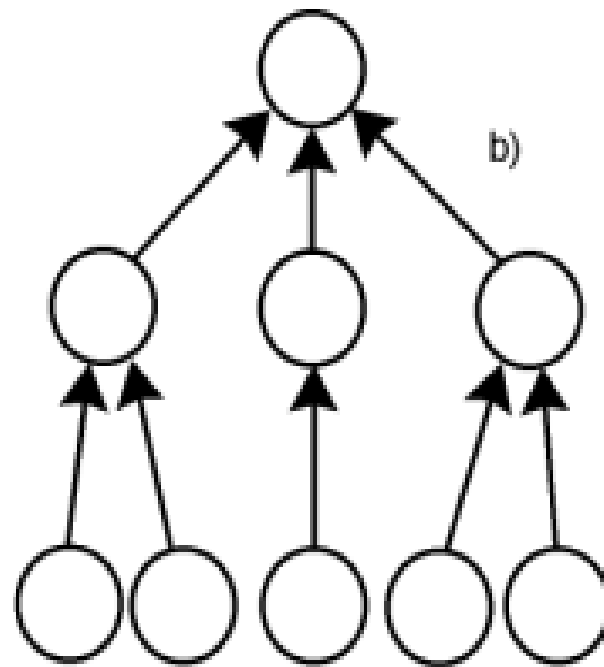
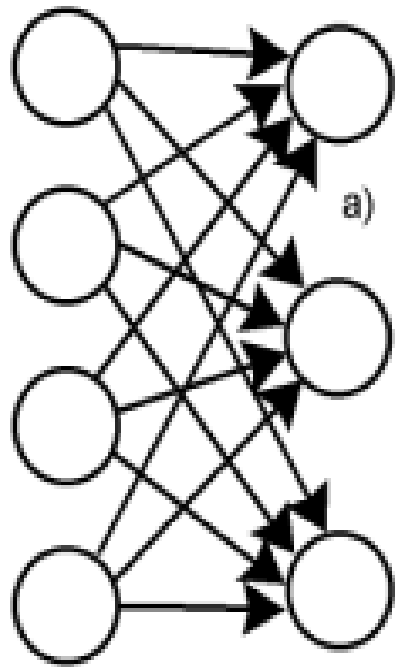


# Grafová analýza

Pro studium topologie webu jsou zajímavé složitější grafové struktury, které již můžeme interpretovat jako přímé informace o webu.

- *Bipartitní graf* – skládá se ze dvou množin stránek, stránky jedné množiny odkazují na stránky druhé množiny. (viz obr. a).
- *Sbíhavá hvězdice* (in-tree) – zobecnění “společné volby” (viz obr. b).
- *Rozbíhavá hvězdice* (out-tree) – zobecnění “společné citace” (viz obr. c).

# Složitější grafové struktury



# Grafová analýza

V kontextu WWW pavučiny můžeme všem těmto útvarům přiřadit základní pravděpodobný význam.

Struktura *bipartitního grafu* napovídá, že stránky v něm obsažené patří do jedné webové komunity.

Dokumenty v horních vrstvách *sbíhavé hvězdice* můžeme chápat jako důležitý zdroj informací k nějakému tématu, neboť se na něj odkazuje mnoho stránek.

Dokumenty v horních vrstvách *rozbíhavé hvězdice* odkazující na relevantní stránky mohou představovat dobré rozcestníky pro hledání monotematických informací.

# Grafová analýza

- Z jiného pohledu jsou pro vyhledávání v prostředí internetu důležité dva typy stránek *rozcestníky* a *autority* (hubs and authorities).
  - *Rozcestník* je stránka, která odkazuje na mnoho autorit
  - *Autorita* je stránka, na kterou odkazuje mnoho rozcestníků
- Charakteristické tvary grafů, stejně tak jako podgrafy tvořené rozcestníky a autoritami, mohou být použity k identifikaci komunit webových stránek stejného tématu.
- Alternativní metodou, jak identifikovat webové komunity, je hledání množiny uzlů, pro které je hustota odkazů mezi sebou vyšší, než hustota odkazů mezi uzly okolní sítě.

# Algoritmy pro hledání relevantních dokumentů

Jako příklad uveďme algoritmus HITS

- skládá ze dvou hlavních kroků.
- První krok se nazývá vzorkování (sampling) a je zaměřen na vytvoření množiny webových stránek, která by měla být bohatá na důležité autority.
- Druhým krokem je propagace vah (weight-propagation), která má za úkol iterativním postupem přiřadit každé stránce ohodnocení, které určuje, jak důležitou autoritou a rozcestníkem je.

Stránky s nejvyšším daným ohodnocením jsou pak vybrány jako rozcestníky nebo autority pro dané téma.

# Algoritmy pro hledání relevantních dokumentů

- Prvním krokem algoritmu HITS je vytvoření takového grafu, který by byl bohatý na relevantní stránky, mezi kterými by se pak mohly hledat rozcestníky a autority. Tento graf vytvoříme pomocí takzvané „kořenové množiny“ (root set), která by měla obsahovat okolo 200 stránek. Tuto kořenovou množinu získáme jako výsledek dotazu pro vyhledávací stroj obsahující klíčová slova. Tato množina nemusí obsahovat všechny důležité stránky. Pomocí kořenové množiny získáme rozšířenou množinu, a to tak, že přidáme všechny stránky, které jsou odkazovány z této množiny. Velikost rozšířené množiny se obvykle pohybuje okolo 1000 až 3000 stránek.

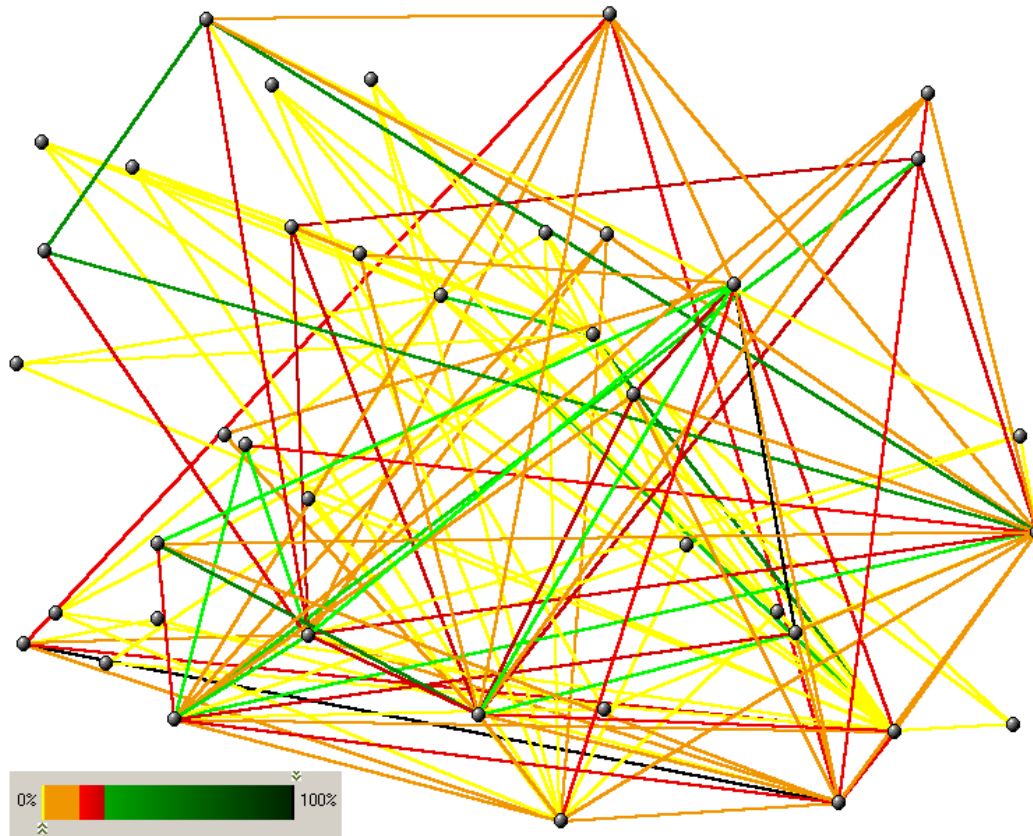
# Experimentální výsledky a vizualizace výsledků

- Výsledky vycházejí z prací [4,7,8] a jsou součástí projektu AmphorA viz [www.cs.vsb.cz/arg](http://www.cs.vsb.cz/arg). Testování webových útvarů probíhalo náhodným zvolením několika výchozích URL adres, ze kterých se postupným vnořováním do webové struktury (prohledáváním do hloubky) získávaly informace o struktuře. K nalezení struktur se používaly výše zmíněné grafové algoritmy (2-souvislé komponenty, hledání cyklů, společná citace). Druhou částí experimentů byla vizualizace relevance vztahů uvnitř grafů a vizualizace samotných grafů.

## 2- souvislé komponenty

- Výchozí adresou, ze které se provádělo hledání těchto útvarů, byla URL <http://www.kourim-mu.cz>. Adresa patří městskému úřadu města Kouřim, Česká republika. Na obrázku 3 je znázorněna 2-souvislá komponenta, která se skládá ze 134 hran.
- Barvy jednotlivých hran určují míru podobnosti dokumentů, které na sebe odkazují. Grafická škála v levém dolním rohu obrázku znázorňuje míru podobnosti od 0% do 100%, přičemž podobnost stránek se měřila pomocí kosinové míry viz [8]. Žluté hrany spojují málo podobné dokumenty, zatímco tmavě zelené hrany spojují hodně podobné dokumenty

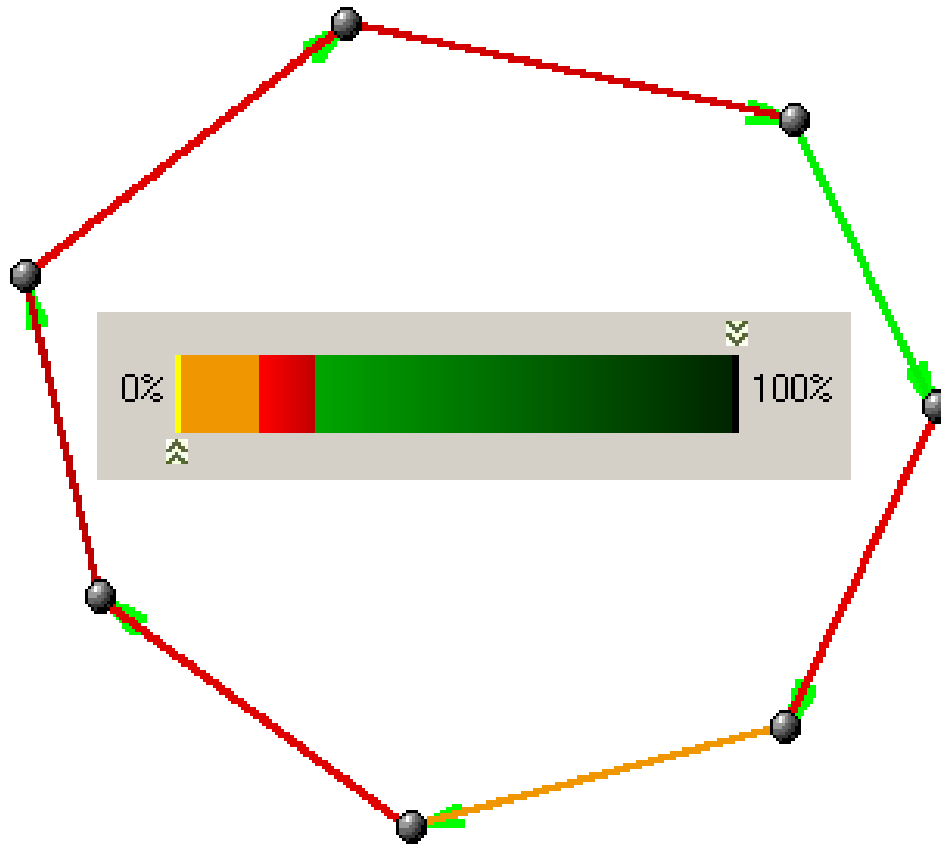
# 2- souvislé komponenty



# Cykly

- U hledání cyklů stojí za zmínku cykly s největším počtem hran, kde na obrázku 4 je vidět, že i dokumenty obsažené v obyčejném cyklu mají dostatečnou míru podobnosti.
- Za zmínku stojí říci, že cyklů s větším počtem hran (5 a více) se ve zmíněném grafu našlo poměrně málo.

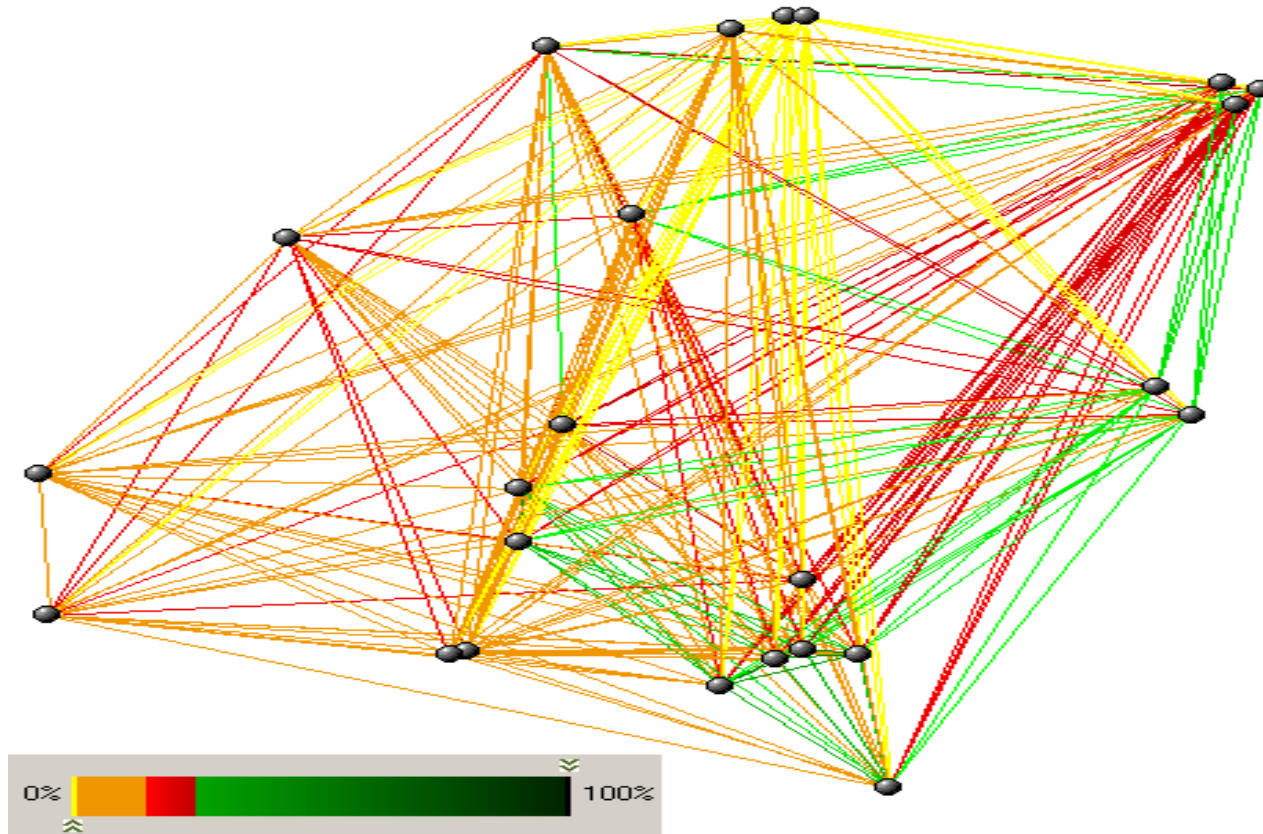
# Cyklus v grafu



# Společná citace

- K vizualizaci byl použit mírně obměněný algoritmus “společná citace”. Změna se spočívala v obohacení algoritmu o hledání rodičů dětí daného výchozího uzlu. Algoritmus tedy hledá jak děti rodičů uzlu, tak rodiče dětí výchozího uzlu. Grafová struktura na obrázku 5 reprezentuje výsledek algoritmu. Je vidět, že podobnost vybraných dokumentů je vysoká, ale na druhé straně se těchto útvarů v dané struktuře vyskytuje poměrně hodně. To je dáno hlavně tím, že většina dokumentů na internetu má nějakého rodiče nebo potomka.

# Společné citace



# Zhodnocení výsledků

Po stažení dat následném vytvoření grafu webu jsme měli k dispozici graf o 171 uzlech a 383 hranách. Hloubka stahování byla v tomto případě 16. Po analýze grafu jsme obdrželi přehled, kolik grafových útvarů bylo nalezeno.

U grafových útvarů „Společná citace – rodiče a děti“ musela být k nalezení splněna ještě jedna podmínka, jež se týkala počtu vrcholů, které mohly být zařazeny do útvaru. Tento počet byl stanoven konstantou 8.

Počet nalezených útvaru		
komponenty	cykly	společné citace
5	23	135

# Zhodnocení výsledků

Následující tabulka nám dává informaci o jednotlivých komponentách. Informace jsou zaměřeny především na míru podobnosti, v jaké se vyskytovaly jednotlivé dvojice dokumentů v daných komponentách.

Připomeňme, že podobnost byla měřena pomocí vektorového modelu a je vyjádřena příslušností různých barev k jednotlivým hranám.

# Zhodnocení výsledků

KOMPONENTY						
Název útvaru	počet uzlů	počet hran	0%-15%	16%-25%	26%-50%	51%-99%
komponenta č1	7	10	7	1	2	0
komponenta č2	11	20	10	5	2	3
komponenta č3	9	35	13	21	1	0
komponenta č4	39	134	93	23	9	9
komponenta č5	10	28	12	14	2	0

Charakteristika komponent

# Zhodnocení výsledků

Jak je možno vidět, podobnost dokumentů ve 2-souvislé komponentě není příliš velká ve srovnání například s podobností dokumentů nalezených obměněným algoritmem „společná citace“. Zde by stála za zmínku myšlenka nalezení ještě silnější 2-souvislé komponenty. Nejprve by se odstranily hrany v grafu 2-souvislé komponenty, které by nepředstavovaly velkou podobnost. Poté by se na tento graf aplikoval opět algoritmus nalezení 2-souvislé komponenty. Jako výsledek bychom dostali dokumenty, které se v této komponentě nacházejí a mají vyšší míru relevance oproti původnímu grafu. Tato metoda postupného ubírání nepodstatných hran a opětovné aplikace algoritmu je dobrým příkladem jak z velkého grafu vybrat zajímavější podgraf, jehož dokumenty si budou velmi podobné. Rovněž tato metoda byla vyzkoušena.

# Zhodnocení výsledků

Následující tabulka popisuje údaje o nalezených cyklech. První část tabulky ukazuje, jak velký počet cyklů a s kolika hranami byl v daném cyklu nalezen. V druhé části tabulky jsou shromážděna data o jednom vybraném cyklu. Tento cyklus má délku 7 a byl vybrán, jelikož byl v tomto případě (hloubka stahování 16) nejdelší. Každá hrana tohoto cyklu je procentuálně ohodnocena a vyjadřuje míru podobnosti mezi dvěma dokumenty v tomto cyklu. Ve zbylé části tabulky je uveden seznam URL adres dokumentů, které tvoří právě tento cyklus. I když se cyklus může jevit jako nejslabší grafový útvar, v tomto případě, co se týká podobnosti dokumentů, činilo průměrné ohodnocení podobnosti sousedních dokumentů okolo 22%.

# Zhodnocení výsledků

CYKLY							
<i>délka cyklu</i>	3	4	5	6	7	8	9
počet nalezených	11	6	3	2	1	0	0
cyklus délky 7							
hrany mezi uzly	1-2	2-3	3-4	4-5	5-6	6-7	7-1
podobnost	20%	25%	20%	15%	19%	34%	22%
uzel	URL						xml soubor
1	<a href="http://www.wwwx.cz/http.html">http://www.wwwx.cz/http.html</a>						145
2	<a href="http://www.wwwx.cz/poslani.html">http://www.wwwx.cz/poslani.html</a>						144
3	<a href="http://www.wwwx.cz/reseni.html">http://www.wwwx.cz/reseni.html</a>						143
4	<a href="http://www.wwwx.cz/realizace.html">http://www.wwwx.cz/realizace.html</a>						142
5	<a href="http://www.wwwx.cz/technologie.html">http://www.wwwx.cz/technologie.html</a>						141
6	<a href="http://www.wwwx.cz/webworx.html">http://www.wwwx.cz/webworx.html</a>						140
7	<a href="http://www.wwwx.cz/kontakty.html">http://www.wwwx.cz/kontakty.html</a>						137

## Charakteristika cyklů

# Zhodnocení výsledků

Třetí nejzajímavější strukturou jsou dokumenty nalezené pomocí obměněného algoritmu „společná citace“. Zajímavé je to, že množiny, které vzniknou po aplikaci tohoto algoritmu na graf, obsahují dokumenty, jejichž míra podobnosti je relativně vysoká. Pro příklad bylo vybráno několik množin, které se vyznačovaly svými extrémy (velký počet citací, malý počet uzlů a naopak), viz tabulka 5. Údajem citace se zde myslí součet jednak odkazů vstupujících do vybrané množiny a jednak odkazů z množiny vystupujících.

# Zhodnocení výsledků

SPOLEČNÁ CITACE											
<i>z uzlu</i>	<i>citaci</i>	uzlu	hran	0%	1%-15%	16%-25%	26%-50%	51%-75%	76%-99%	100%	citace/uzel
156	111	31	465	75	230	88	37	15	16	4	3,58
103	148	44	946	350	391	103	39	15	17	31	3,36
95	53	17	136	42	2	0	0	1	0	91	3,12
111	124	48	1128	386	574	102	19	25	13	9	2,58
159	53	30	435	72	211	81	37	15	16	3	1,77

Výsledky získané aplikací algoritmu „společná citace“

# Vizualizace výsledků

- Velmi zajímavý je projekt KartOO viz <http://www.kartoo.com>. Tento vyhledávač umožňuje prezentovat výsledky vyhledávání velmi přehlednou formou. Tato forma respektuje odkazy, které propojují nalezené stránky.
- Výsledky dosažené v projektu KartOO ukazují, že výraznou roli budou hrát shlukovací algoritmy

# Vizualizace výsledků

The screenshot displays the KartOO visual meta search engine interface within a Microsoft Internet Explorer browser window. The browser title is "KartOO visual meta search engine - Microsoft Internet Explorer" and the address bar shows "http://www.kartoo.com/flash.php3". The search bar contains the query "snasel".

The main area features a network visualization of search results. Nodes are represented by orange and red spheres of varying sizes, connected by lines. The central node is "vaclav", which is connected to "mathematica" and "czech republic". Other nodes include "university computer", "palacky olomouc", "www.math.cas.cz", "www.ceur-ws.org", "www.computer.org", "www.cpt.univ-mrs.fr", "www.iarerelative.com", "www.siam.org", "www.ejbiochem.org", "www.fee.vutbr.cz", "www.emis.de", "www.vutbr.cz", "www.math.cas.cz", "www.ceur-ws.org", "www.computer.org", "www.cpt.univ-mrs.fr", "www.iarerelative.com", "www.siam.org", "www.ejbiochem.org", "www.fee.vutbr.cz", "www.emis.de".

Labels associated with the nodes include: "document", "charles", "methods", "text", "jaroslav", "fajkus", "karel", "publications", "university computer", "palacky olomouc", "czech republic", "research", "department", "science", "workshop", "ostrava", "vaclav", "www.cs.vsb.cz", "dvorsky", "www.ejbiochem.org", "list", "bohemia", "analysis", "skopal", "kratky", "jewel.morgan.edu", "papers", "www.fee.vutbr.cz", "www.emis.de".

On the left side, there are two sections: "Top Sites" and "Topics".

**Top Sites**

- 1) www.ejbiochem.org
- 2) www.cs.vsb.cz
- 3) www.ceur-ws.org
- 4) jewel.morgan.edu
- 5) www.computer.org

**Topics**

- "snasel vaclav"
- "vaclav snasel"
- "czech republic"
- "charles university"
- "text documents"
- "based compression method..."
- "mathematica bohemia"
- "palacky university"
- vaclav
- czech
- republic
- pokorny
- mathematica
- university
- computer

At the bottom left, there is a logo for "KartOO v4" with a timer showing "00:00:00" and an "OK!" button. The status bar at the bottom indicates "Done" and "Internet".

# Závěr

Relativně novým problémem je využití informace skryté ve vazbách mezi dokumenty, případně částmi těchto dokument. V tradičním pohledu jsou jednotlivé dokumenty chápány jako izolované elementy. Tento pohled v mnohých případech již nevyhovuje současným potřebám. Vizualizace těchto vazeb a následná navigace ve výsledné struktuře vede ke značnému zlepšení vyhledávacího procesu. Rozsah vazeb motivuje i ke studiu dalších problémů svázaných s reprezentací takto rozsáhlých grafů.

# Závěr

Tato problematika není v současné době plně zvládnutá. Existuje mnoho otevřených problémů jako například:

- kombinace různých typů informací obsažených na webu, text, otázky, ...,
- výzkum dalších statistických vlastností grafové struktury webu,
- vizualizace výsledků vyhledávání,
- využití dalších grafových algoritmů.