Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Disertační práce

2015 **Ing. Michal Nykl**

Západočeská univerzita v Plzni
Fakulta aplikovaných věd

HODNOCENÍ VÝZNAMNOSTI VARIANTAMI PAGERANKU

Ing. Michal Nykl

disertační práce
k získání akademického titulu doktor
v oboru Informatika a výpočetní technika

Školitel: Prof. Ing. Karel Ježek, CSc.

Katedra informatiky a výpočetní techniky

Plzeň 2015

University of West Bohemia
Faculty of Applied Sciences

EVALUATION OF SIGNIFICANCE BASED ON PAGERANK VARIANTS

Ing. Michal Nykl

doctoral thesis
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in specialization Computer Science and Engineering

Supervisor: Prof. Ing. Karel Ježek, CSc.

Department of Computer Science and Engineering

Pilsen 2015

Prohlášení

Předkládám tímto k posouzení a obhajobě svou disertační práci, která vznikla v závěru mého doktorského studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni, a prohlašuji, že jsem tuto práci vypracoval samostatně s použitím výhradně citované odborné literatury.

V Plzni dne 25. 11. 2015.

……

Ing. Michal Nykl

Věnováno lidem,
kteří mění svět k lepšímu.

Touto formou bych zvláště rád poděkoval profesoru Karlu Ježkovi za dlouholeté ochotné a vstřícné vedení a za čas, který se mnou v průběhu uplynulých let strávil. Jeho dobré rady pro mě byly přínosem jak na poli vědy a výuky, tak i v osobním životě. Dále bych chtěl poděkovat všem členům *Text-Mining Research Group* na Katedře informatiky a výpočetní techniky ZČU v Plzni za občasnou pomoc a kolegiální náladu na pracovišti. Poděkování patří zejména Martinu Dostalovi, Michalu Camprovi, Lubomíru Krčmářovi a Daliboru Fialovi, kteří byli mými blízkými kolegy. Závěrem bych chtěl také poděkovat i všem zbylým členům katedry za jejich otevřenost a dobrou náladu, se kterou jsem se na katedře často setkával.

Abstrakt

Tato práce se zabývá výzkumem metod pro hodnocení významnosti vrcholů v rozsáhlých grafových strukturách. Navržené metody jsou aplikovány při vyhodnocení citačních sítí a sítí vytvořených z Linked Data. V úvodu práce jsou popsány cíle, které nás k návrhu nových metod vedly. Následně lze text práce pomyslně rozdělit na dvě části, z nichž první a obsáhlejší část je věnována návrhu metod pro hodnocení autorů vědeckých publikací a druhá část je věnována návrhu metody pro určení klíčových slov textového dokumentu. Společnou vlastností všech navržených metod je použitý algoritmus PageRank.

V první části práce je nejprve shrnut aktuální stav poznání v oblasti citační analýzy a zmíněny nejznámější bibliografické databáze a algoritmy, které bývají při citační analýze používány. Zvláštní prostor je věnován popisu algoritmu PageRank, který jsme při výzkumu používali a dále upravovali. Následně první část obsahuje popis návrhu nových metod pro hodnocení významnosti autorů a popis experimentálního ověření jejich kvality. Pro experimenty byly použity datové kolekce CiteSeer, DBLP a WoS, přičemž výsledky získané z kolekce WoS byly, vzhledem k jejím vlastnostem, prohlášeny za nejdůvěryhodnější. Poté, co se prokázala vhodnost nově navržených metod pro hodnocení autorů, jsme provedli další experimenty, jejichž cílem bylo metody ještě více vylepšit. Zde se pro hodnocení autorů ukázalo nejvhodnější parametrizovat PageRank aplikovaný na citační síť publikací významností časopisů, ve kterých byly publikace zveřejněny. Vhodnost navržených metod a platnost vyvozených závěrů byly ověřeny také vyhodnocením specializovaných kategorií WoS.

V druhé části práce jsou nejprve zmíněny významné práce z oblasti klasifikace textových dokumentů a z oblasti využití PageRanku pro extraktivní sumarizaci obsahu dokumentu. Následně je popsán návrh naší metody pro volbu klíčových slov textového dokumentu. Tato metoda využívá PageRank a Linked Data, čímž dokáže určit k textu dokumentu vysoce relevantní klíčová slova, která v textu nemusejí být explicitně uvedena. Kvalita navržené metody byla experimentálně ověřena jejím použitím v klasifikátoru dokumentů, který byl aplikován na dokumenty z kolekce diskusních článků *20 Newsgroups* a na dokumenty z vlastní kolekce konferenčních *Call-for-Papers*. Určená klíčová slova byla použita jako vlastnosti dokumentů. Závěrem bylo, že navržená metoda je vhodná zejména v situacích, kdy máme malé množství dat pro natrénování klasifikátoru.

Autorovy vědecké přínosy, které jsou popsány v této práci, byly publikovány formou pěti vědeckých článků, z nichž dva byly zveřejněny v časopisech a tři v konferenčních sbornících.

Klíčová slova: *dolování dat, citační analýza, PageRank, hodnocení autorů, volba vlastností textových dokumentů*.

Abstract

This thesis deals with the research of methods of evaluating the significance of nodes in large graph structures. The proposed methods are applied to evaluating citation networks and networks created from Linked Data. The introduction describes the goals that led us to propose the new methods. The text is divided into two parts, while the first one deals with the suggestion of methods of evaluating the authors of scientific publications, the second part is dedicated to the suggestion of a method of determining text document keywords. The common feature of all the proposed methods is the use of the PageRank algorithm.

The first part provides the summary of the current state of knowledge in citation analysis and there are mentioned the best known bibliographic databases and algorithms that are used in the citation analysis. A special section is devoted to the description of the PageRank algorithm, which we used and further modified in our research. Subsequently, the first part contains the description of the new evaluation methods of author's significance and the description of the experimental verification of their quality. For the experiments, we used the CiteSeer, DBLP and WoS data collections, while the results obtained from the WoS collection have been declared as the most accurate, due to its characteristics. After proving the suitability of the newly developed evaluation methods of authors, we performed additional experiments aimed at their further improvement. The most appropriate author's evaluation method proved to be PageRank applied to the citation network of publications and parameterized with the significance of journals in which the publications were published. The suitability of the proposed methods and the validity of the drawn conclusions were also verified by the evaluation of WoS specialized categories.

In the second part we first mention the most significant works in the field of text documents classification and in the field of PageRank using for extractive summarization of the document content. Then we describe our suggested method for the text document keywords selection. This method uses PageRank and Linked Data, so that it can identify the most relevant keywords from the text, which may not even be explicitly present. The quality of the proposed method was experimentally verified by using it in a document classifier, which has been applied to the documents from the collection of *20 Newsgroup*s discussion articles and also on documents from our own collection of conference *Call-for-Papers*. The identified keywords have been used as document features. The conclusion is that the method is particularly suitable in situations where we have a small amount of data for training the classifier.

The author's scientific contributions that are described in this thesis have been published in the form of five scientific articles, two of which were in journals and three in conference proceedings.

Keywords: *data-mining, citation analysis, PageRank, author evaluation, feature selection for textual documents*.

Obsah

1 Úvod

Tato práce shrnuje naše[1] stěžejní výsledky publikované ve 2 časopiseckých (Nykl et al. 2014, 2015) a 3 konferenčních (Nykl a Ježek 2012; Nykl et al. 2013; Dostal et al. 2014a) článcích. Obsah je soustředěn na problematiku určování významných vrcholů grafu. Ta v počítačových vědách patří do oblasti dolování dat (*data mining*), přičemž příslušné metody bývají používány pro dolování struktury grafu (*graph structure mining*). Graf obvykle představuje určitou oblast znalostí, přičemž jeho vrcholy zastupují zúčastněné entity (webové stránky, publikace, instituce, autory či obecně osoby atd.) a hrany vyjadřují určitý vztah (tok informací, společný výskyt, známost apod.). Na základě vypočtených hodnot významnosti vrcholů lze příslušné entity porovnávat a vybírat entity pro další zpracování.

Jedním z používaných algoritmů je algoritmus PageRank, který pro určení hodnoty vrcholu používá hodnoty vrcholů, které na daný vrchol odkazují. Vypočtená hodnota vrcholu bývá označována jako významnost, vliv, autoritativnost nebo podobně a používána např. při řazení výsledků ve vyhledávači webových stránek, při porovnávání či vyhledávání významných osob, institucí, časopisů atd. Protože PageRankem bývají často vyhodnocovány citační grafy, tak bývá také označován jako nástroj citační analýzy.

Od svého vzniku v roce 1998 byl PageRank vylepšován pro potřeby jeho adaptace na různé druhy grafů nebo pro urychlení jeho výpočtu. Našimi cíli, které jsou shrnuty v této práci, bylo navrhnout nové, na PageRanku založené, metody a to jednak pro potřeby bibliometrie, a dále pak pro potřeby zpracování textů. V bibliometrii jsme navrhli metody, které umožňují hodnotit, porovnávat a vyhledávat významné autory vědeckých publikací, a porovnali jsme je s některými stávajícími metodami. V úloze zpracování textů slouží námi navržená metoda pro určení klíčových slov textového dokumentu. Klíčová slova mohou být dále použita při klasifikaci, shlukování či štítkování dokumentů. Detailnější popis našich cílů obsahuje část 1.1 a stručný popis jednotlivých kapitol práce část 1.2. Jednotlivé části práce mohou sloužit jako podpůrný zdroj při výuce, proto je práce napsána v českém jazyce. Pro účely vytvoření české terminologie v dané vědní oblasti jsou také u některých algoritmů zavedeny odpovídající české názvy.

1.1 Cíle práce

Základním cílem popisovaného výzkumu bylo prověření schopností algoritmu PageRank při hodnocení významnosti vrcholů grafu. Z uvedeného základního cíle vzniklo několik odvozených cílů, které byly námětem výzkumů, jejichž výsledky byly uveřejněny v publikovaných článcích. Oblastmi, které jsme pro ověření použitelnosti PageRanku zvolili, byly:

a) Bibliometrie – úloha hodnocení autorů vědeckých publikací.
b) Zpracování textů – úloha volby klíčových slov textového dokumentu.

[1] Přestože cílem této disertační práce je shrnout vědecké přínosy Michala Nykla, tak v práci bude při popisu dosažených výsledků použito množné číslo „my“, protože všechny práce vznikly pod odborným dozorem profesora Karla Ježka a s pomocí kolektivu „Text mining research group“, viz http://textmining.zcu.cz

V oblasti bibliometrie jsme, vzhledem k nám nejbližšímu oboru, hodnotili autory, kteří publikují v počítačových vědách, přičemž našimi cíli bylo:

(a1) Navržení metody pro automatické hodnocení autorů, která bude hodnotit autory z počítačových věd s výsledky obdobnými hodnocením organizací *Association for Computing Machinery* (ACM) a *Institute for Scientific Information* (ISI), a analýza vhodnosti použití datových kolekcí CiteSeer (2005), DBLP (2004) a WoS (1996-2005) pro hodnocení autorů.

(a2) Porovnání navržených metod s neiteračními metodami.

(a3) Zjištění, jaký vliv na kvalitu hodnocení autorů mají použité citační sítě publikací či autorů, samocitace autorů a váhy hran v citační síti autorů.

(a4) Zjištění, jaký vliv na kvalitu hodnocení autorů mají způsoby rozdělení hodnot publikací jejich autorům, a posouzení vhodnosti zvýhodňování prvních či korespondujících autorů publikací.

(a5) Ověření vlivu parametrizace PageRanku charakteristikami autora či publikace na kvalitu hodnocení autorů.

(a6) Ověření použitelnosti navržených metod v případě změny rozsahu vyhodnocovaného oboru.

V oblasti zpracování textových dokumentů byly naše cíle:

(b1) Navržení metody, která využitím Linked Data a PageRanku dokáže automaticky určit klíčová slova pro daný textový dokument. Tato slova se nemusejí explicitně vyskytovat v textu dokumentu, ale měla by daný dokument reprezentovat lépe, než slova určená pouze statisticky.

(b2) Ověření kvality navržené metody při klasifikaci textových dokumentů.

1.2 Struktura práce

Ve 2. kapitole je popsán aktuální stav poznání v oblasti citační analýzy. Představena je její historie, nejpoužívanější bibliografické databáze, vytvářené grafy a nejznámější neiterační metody pro hodnocení časopisů a autorů. Dále je zde popsán iterační algoritmus PageRank a jemu podobné algoritmy, které byly navrženy pro použití v citační analýze.

Ve 3. kapitole je popsán návrh našich metod pro hodnocení autorů a experiment s hodnocením autorů v kolekcích CiteSeer a DBLP, který byl publikován v (Nykl a Ježek 2012). Jsou zde uvedeny postupy vytvoření námi používaných citačních sítí, včetně charakteristických vlastností sítí vytvořených z kolekcí CiteSeer a DBLP. Dále jsou popsány manuálně vytvořené referenční seznamy oceněných autorů a v závěru kapitoly je diskutována kvalita námi navržených metod.

Ve 4. kapitole je uveden experiment s aplikací ve 3. kapitole navržených metod na kolekci WoS, který byl publikován v (Nykl et al. 2014). Popsány jsou cíle zmíněného experimentu, charakteristické vlastnosti sítí vytvořených z kolekce WoS a odpovídající referenční seznamy autorů. K navrženým metodám je pro porovnání navíc přidána neiterační metoda počítající citace. V závěru kapitoly je opět posouzena kvalita navržených metod.

Kapitola 5. popisuje náš aktuálně poslední výzkum v oblasti bibliometrie, publikovaný v (Nykl et al. 2015). Cílem výzkumu bylo navrhnout další možné zdokonalení metody pro hodnocení autorů a prověřit vliv míry specifičnosti zpracovávané oblasti na hodnocení autorů. V textu jsou popsány charakteristické vlastnosti kategorií *Umělá inteligence* a *Hardware*, které jsme z kolekce WoS vyextrahovali, abychom ověřili kvalitu našich metod ve specializovaných oblastech výzkumu. Následuje detailní popis návrhu našich nových metod a experimentálního ověření jejich kvality při hodnocení autorů z kolekce WoS a zvolených kategorií. V závěru kapitoly je diskutována kvalita metod a to jak pro případ hodnocení autorů z celé kolekce WoS či zvolených kategorií, tak i pro případ předpovědi laureátů vědeckých ocenění.

V 6. kapitole jsou shrnuty naše experimenty s určováním klíčových slov pro textové dokumenty, které byly publikovány v (Nykl et al. 2013) a v (Dostal et al. 2014a). Protože jsme kvalitu navržené metody experimentálně ověřili jejím použitím v klasifikátoru dokumentů, tak jsou v této kapitole také zmíněny významné práce z oblasti klasifikace textových dokumentů. Dále jsou zde zmíněny relevantní práce z oblasti extrakce klíčových slov či frází z textů PageRanku podobnými algoritmy. Následně jsou popsány použité datové kolekce a koncept Linked Data. Více prostoru je věnováno návrhu naší metody pro získání klíčových slov, která mohou zastupovat daný dokument. Závěr kapitoly obsahuje posouzení kvality navržené metody.

V 7. kapitole je popsáno splnění cílů této práce. Jsou zde shrnuty vědecké přínosy autora, které byly v této práci publikovány, a uvedena doporučení pro budoucí práce.

Příloha A obsahuje aktuální výčet publikačních výsledků autora. Uvedeny jsou reference na publikované články autora a na články, které je citují.

2 Citační analýza

Tato kapitola seznamuje s aktuálním stavem poznání v oblasti citační analýzy. Historie citační analýzy je stručně shrnuta v části 2.1. V části 2.2 jsou zmíněny druhy grafů, které lze vytvořit z bibliografických záznamů, nejznámější bibliografické databáze a možnosti porovnání strojově vytvořených pořadí bibliografických entit. V části 2.3 jsou popsány dvě nejznámější neiterační metody pro hodnocení časopisů a autorů, jejichž hodnoty bývají aktuálně zobrazovány bibliografickými databázemi. Těmito metodami jsou Impact Factor (pro hodnocení časopisů) a h-index (pro hodnocení autorů). Dále jsou v této části shrnuty míry centrality, které bývají v sociálních sítích používány pro určení centrálnosti vrcholů. Iterační algoritmus PageRank, který je naším stěžejním algoritmem, je důkladně popsán v části 2.4. Některé bibliografické databáze už PageRank či jeho modifikaci také používají pro hodnocení časopisů. Modifikace PageRanku a jemu podobné algoritmy jsou detailně sepsány v části 2.5.

2.1 Historie citační analýzy

Jedním ze zakladatelů citační analýzy je Eugen Garfield. Ten jako první navrhl systematické indexování vědecké literatury a citací v ní obsažených za účelem tvorby citačního indexu, který slouží k hodnocení vědeckých časopisů. Navrženou metodu pro hodnocení časopisů nazval *Impact Factor* (Garfield 1955a). Cílem návrhu bylo použití Impact Factoru pro zhodnocení vlivu vybraných vědeckých časopisů na literaturu a výzkum ve zvoleném období. K vlivnosti časopisů může být přihlíženo např. při nákupu časopisů do vědeckých knihoven. Garfield poznamenává, že Impact Factor indikuje vliv časopisů více, než celkový počet publikací, který dříve použili Lehman (1954) a Dennis (1954) pro hodnocení autorů. Dále uvádí, že je podobný metodě počítání citací, kterou pro hodnocení významu vědeckých časopisů z oblasti chemie použili Gross a Gross (1927). Ti ale při výpočtu použili reference uvedené pouze v jednom časopise. V oblasti hodnocení autorů vědeckých publikací lze za nejznámější hodnotící metodu považovat *h-index* (Hirsch 2005). Detaily výpočtu Impact Factoru a h-indexu budou zmíněny v části 2.3.

Obecně citační analýza slouží k nalezení významných bibliografických entit (článků, autorů, časopisů, institucí, témat atd.) využitím algoritmů nebo metod, které pracují s bibliografickými záznamy a citačním grafem. Tento problém lze zapsat takto: *na vstupu máme bibliografické záznamy o publikacích z určené vědní oblasti (např. počítačové vědy) a na výstupu chceme získat hodnoty významnosti prvků zvolené entity (např. autorů), dle kterých můžeme prvky seřadit*.

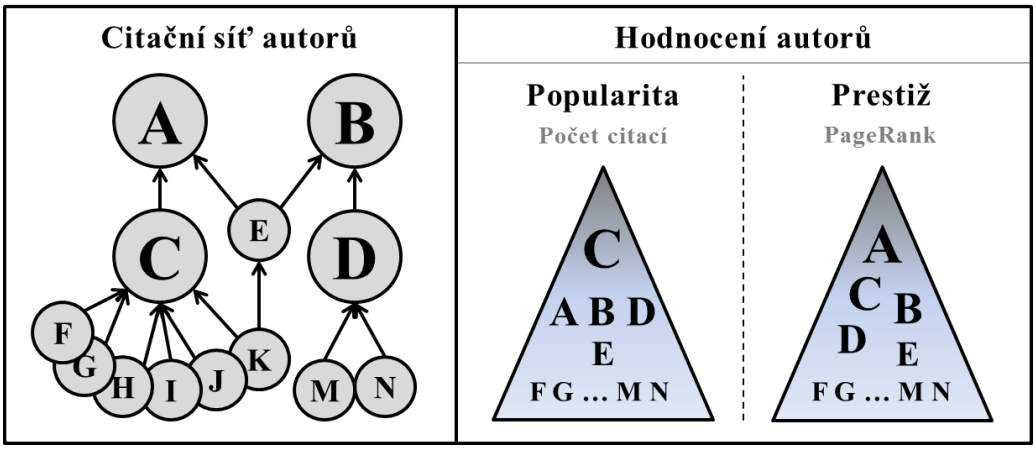
Jedním z aktuálních cílů citační analýzy je odlišení populárních a prestižních autorů. Ding (2011a) zmiňuje skutečnost, že pojem *populární* pochází z latinského výrazu *popularis*[2], kterému lze rozumět jako „*milovaný lidmi*“, kdežto pojem *prestižní*, z latinského *praestigious*[3], vyjadřuje „*mající oslnivý vliv*“. Autorka uvádí pěkný příklad, když říká, že autor, který ve své práci shrnuje aktuální stav poznání v určité oblasti, může být hodně citován začínajícími autory v dané oblasti, ale již méně těmi, kteří jsou v dané oblasti experty – autor je populární. Naopak autor referátu, který představuje inovativní metodu, může být citován experty, ale již méně laiky – autor je prestižní (pozn.: autor může být populární, ale nemusí být prestižní a naopak). Z toho autorka vyvozuje, že populární autor je hodně citovaný a popularitu tedy lze měřit *počtem citací*. Naopak prestižní autor je citovaný významnými

[2] Výklad slova „populární“ - http://www.etymonline.com/index.php?term=popular
[3] Výklad slova „prestižní“ - http://www.etymonline.com/index.php?term=prestigious

autory a prestiž tedy lze měřit *počtem citací od významných autorů* (to ale vyžaduje vědět, kdo je významný). Stejný koncept zmínili také Bollen et al. (2006) při hodnocení časopisů.

S ohledem na výše uvedené odlišení pojmů *populární* a *prestižní* se v citační analýze pozvolna přechází od metod, které pro hodnocení používají pouze kvantitativní vlastnosti (např. počet citací), k metodám používajícím i vlastnosti odvozené. Tyto metody obvykle používají významnosti citujících entit a tak dokáží určit, zda citace pochází z významného zdroje (Bollen et al. 2006; Ding 2011a). Často používán je algoritmus PageRank (Brin a Page 1998), který určuje významnost bibliografických entit (např. publikací, autorů atd.) na základě významnosti entit, které je citují, přičemž výpočet je iterační. Jednou z dobrých vlastností PageRanku je např. jeho schopnost odhalit články, které obsahují převratné výsledky, ale jsou méně citované (Chen et al. 2007; Maslov a Redner 2008). Za zmínku stojí, že v bibliografických databázích ISI Web of Science a Scopus jsou upravené varianty PageRanku dnes již používány pro hodnocení časopisů. V ISI Web of Science jsou to Eigenfactor™ Metrics (Bergstrom 2007; Bergstrom et al. 2008; West et al. 2008, 2010) a ve Scopus je to SCImago Journal Rank (González-Pereira et al. 2010). Tyto metody budou popsány v části 2.5.

Ideu odlišení popularity a prestiže znázorňuje obrázek 2.1, kde autora A můžeme označit za prestižního a autora C za populárního. Autor C je hodně citován necitovanými pracemi, ale jeho práce je založena na práci autora A. Autor A je také prestižnější než autor B.



Obrázek 2.1: Rozdíl mezi popularitou (počet citací) a prestiží (PageRank).

Pořadí autorů, vytvořené dle hodnot jejich významnosti, může být použito při vyhledávání nebo porovnávání expertů ve zvolené oblasti, např. pro účely výběrových řízení, udílení odměn nebo ocenění atd. Vedle hodnocení autorů (Sidiropoulos a Manolopoulos 2005a; Fiala et al. 2008; Ding et al. 2009; Radicchi et al. 2009; Ding 2011a; Fiala 2012b; West et al. 2013) lze využitím citační analýzy určovat významnost časopisů (Garfield 1972; Bollen et al. 2006; González-Pereira et al. 2010; West et al. 2010) a následně dle ní vybírat časopisy do vědeckých knihoven či bibliografických databází nebo vybírat časopisy, ve kterých bychom chtěli publikovat své vědecké výsledky. Se stejným záměrem můžeme hodnotit konference (Sidiropoulos a Manolopoulos 2005b). Publikace mohou být také vyhodnocovány s cílem určení jejich významnosti, či pro zjištění jejich vědeckého přínosu (Sidiropoulos a Manolopoulos 2005a; Chen et al. 2007; Ma et al. 2008; Maslov a Redner 2008; Li a Willett 2009; Sayyadi a Getoor 2009). Publikační významnost výzkumných institucí nebo univerzit či jejich oddělení (Fiala 2013; Ho 2013; Mryglod et al. 2013; West et al. 2013) lze využít při rozdělování finančních prostředků, přičemž zahrnuta může být např. do státního systému pro hodnocení

výzkumných institucí, což používá Česká republika (Úřad vlády ČR 2012, 2013), Austrálie (ERA 2009) a Velká Británie (HEFCE 2009). Porovnání dalších systémů pro hodnocení vědy je uvedeno např. v (Abramo et al. 2010). Pořadí významnosti univerzit či jejich oddělení mohou využívat také např. studenti při výběru univerzity, nebo osoby z vedení a správy jednotlivých institucí. Využitím citační analýzy lze dále vytvářet pořadí států a porovnávat tak jejich přínos k celosvětovému vědeckému rozvoji (Ma et al. 2008; Fiala 2012a; Leydesdorff 2013). Také vědní oblasti mohou být vyhodnoceny citační analýzou (Banks 2013). Zde se obvykle ptáme, která oblast byla nejvíce rozvíjena či přínosná ve sledovaném období.

Více základních informací o citační analýze lze nalézt např. v (Moed 2005; Bellis 2009).

2.2 Bibliografické grafy a uznávané databáze

Cílem této části je ukázat, které informace z bibliografických databází můžeme použít pro tvorbu grafu. Vyhodnocovaný druh grafu udává vlastnost či vlastnosti, které jsou hodnotící metodou měřeny. Následně jsou zmíněny nejznámějších bibliografické databáze a popsány možnosti porovnání vypočtených pořadí.

2.2.1 Druhy bibliografických grafů

Bibliografickým grafem rozumíme graf vytvořený z bibliografických záznamů, ve kterém vrcholy představují prvky zvolené entity (publikace, autory, instituce atd.) a hrany jejich vzájemnou interakci. Hodnocení vrcholů grafu můžeme rozdělit na:

- vyhodnocení „homogenního“ grafu – všechny vrcholy a hrany jsou pouze jednoho typu;
- mnohorozměrné (*multidimensional*) vyhodnocení (Yu et al. 2012) – vyhodnocení, které pracuje s více druhy homogenních grafů současně;
- vyhodnocení heterogenního (Yan et al. 2011) grafu – graf obsahuje vrcholy a/nebo hrany různého typu.

V některých případech může vyhodnocovaný homogenní graf vzniknout kombinací více homogenních grafů a určení, o který typ vyhodnocení se jedná, není jednoznačné, což ale obvykle není příliš důležité. Jedním z faktorů ovlivňujících tvorbu některých grafů autorů a dalších z publikací odvozených entit je, zda použijeme vždy pouze prvního autora publikace nebo použijeme všechny autory publikace (Zhao 2005; Ding 2011a).

Základními bibliografickými entitami jsou publikace (článek, kniha, referát atd.) a základní interakcí jejich vzájemné citace. Z těchto dat lze vytvořit *citační graf publikací*, kde vrcholy jsou publikace a každá hrana/citace je orientována od citující publikace k citované. Ze záznamů o publikacích lze obvykle získat informace o dalších bibliografických entitách (autoři, časopisy, místa publikování, instituce či státy nebo témata) a vytvořit z nich citační grafy, ve kterých lze hodnotit popularitu či prestiž, jak již bylo zmíněno v části 2.1. Pokud graf splňuje definici (Ryjáček 2001): *„Síť je orientovaný graf s kladným reálným ohodnocením hran a s reálným (připouštíme i záporné hodnoty) ohodnocením uzlů.“*, tak lze hovořit o síti.

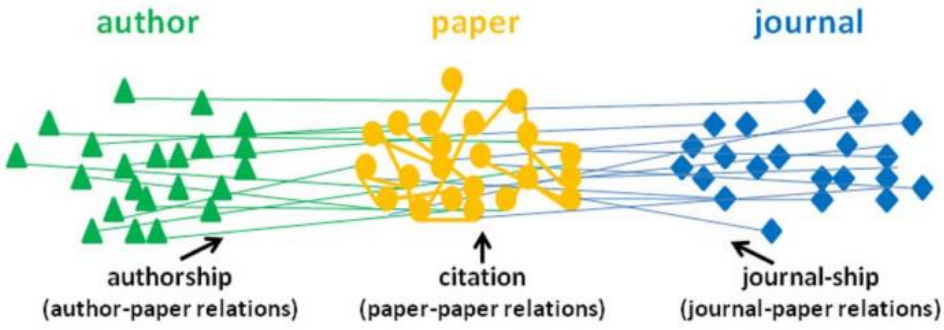
Vedle citačních grafů lze z bibliografických záznamů vytvářet *grafy spoluautorství, spolupráce* či *společného výskytu*, kde mezi entitami vede neorientovaná hrana, pokud se nacházejí ve stejném záznamu o publikaci. Tímto způsobem lze vytvářet grafy spolupráce autorů, institucí nebo států a

grafy společného výskytu témat, klíčových slov či slov obsažených v názvu publikace. Vyhodnocením grafu spoluautorství autorů můžeme např. měřit míru ochoty jednotlivých autorů spolupracovat (Liu et al. 2005; Yan a Ding 2009). Vyhodnocení grafu společného výskytu slov v dokumentu může sloužit např. pro extrakci klíčových slov z dokumentu (Erkan a Radev 2004; Mihalcea a Tarau 2004).

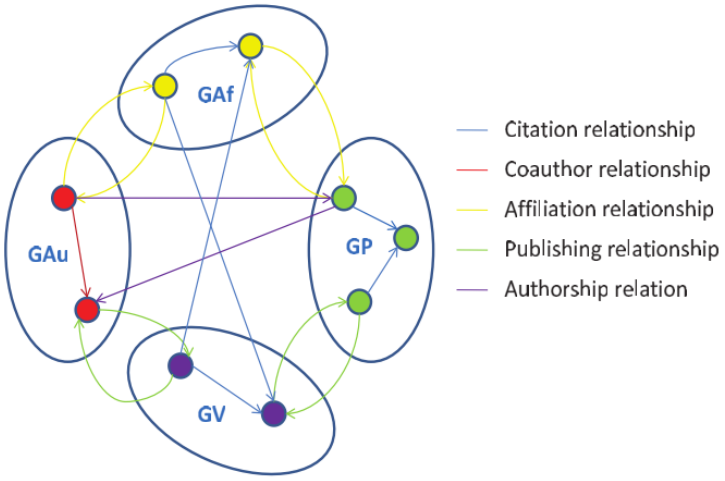
Dalšími vytvářenými grafy jsou *grafy společně citovaných* (*co-citation* nebo *co-cited*) a *grafy společně citujících* (*co-citing* nebo *co-reference*) entit. V grafu společně citovaných vede mezi dvěma entitami neorientovaná hrana, pokud byly obě citovány ve stejné publikaci. V grafu společně citujících vede mezi dvěma entitami neorientovaná hrana, pokud obě citují stejnou publikaci. Vyhodnocením grafu společně citovaných autorů můžeme měřit např. míru toho, jak často byl autor citován společně s vysoce citovanými autory (Ding et al., 2009). Vzájemné porovnání těchto a některých dalších druhů grafu zmiňují Yan a Ding (2012).

Za mnohorozměrné metody pro hodnocení bibliografických entit můžeme označit takové metody, které pracují s více druhy grafů současně. Sayyadi a Getoor (2009) s využitím PageRanku vyhodnocují citační graf publikací a následně aplikují algoritmus HITS (viz část 2.5.3) na bipartitní graf autorství (tj. autoři a jejich publikace), aby získali současně hodnocení autorů i publikací. Yu et al. (2012) s využitím soustavy rovnic hodnotí současně publikace, autory, komentáře a zdroje, tj. časopisy a konference.

Vyhodnocování heterogenního grafu je spíše idea, protože většinou se jedná o mnohorozměrné vyhodnocení. Graf je heterogenní, pokud obsahuje více typů vrcholů a/nebo více typů hran. Tuto vlastnost autoři obvykle ve svých pracích nastíní a ukáží vytvořený graf, ale poté tento graf vyhodnocují po částech, tj. vyhodnocují několik grafů, stejně jako u mnohorozměrného vyhodnocení. Částečnou výjimku tvoří bipartitní grafy, ve kterých ale hrany nikdy nevedou mezi vrcholy stejné množiny. Vyhodnocení heterogenního grafu ukazují např. Yan et al. (2011), kteří používají graf (viz obrázek 2.2) složený z citačního grafu publikací, bipartitního grafu autorství a bipartitního grafu vydávání publikací (tj. časopisy a publikace v nich obsažené). Takto vytvořený graf ale následně vyhodnocují po částech s využitím právě zmíněných tří grafů. Stejným postupem pracují Yang et al. (2010), kteří ukazují heterogenní graf (viz obrázek 2.3) vytvořený spojením citačních grafů publikací (G_P), autorů (G_{Au}), institucí (G_{Af}) a míst publikování (G_V) a grafů spolupráce autorů a institucí prostřednictvím bipartitních grafů publikace-autoři, publikace-instituce, publikace-místa publikování, autoři-instituce, autoři-místa publikování a instituce-místa publikování. Protože obě zmíněné práce následně vyhodnocují dílčí podgrafy, můžeme tyto přístupy označit také za mnohorozměrné vyhodnocení.



Obrázek 2.2: Heterogenní graf, který vnikl spojením citačního grafu publikací, bipartitního grafu autorství (autoři-publikace) a bipartitního grafu vydávání publikací (časopisy-publikace). Přejato z (Yan et al. 2011).



Obrázek 2.3: Nástin heterogenního grafu přejatý z (Yang et al. 2010). Heterogenní graf v sobě kombinuje graf publikací G_P, autorů G_{Au}, institucí G_{Af} a míst publikování G_V vztahy citování (modrá), spoluautorství (červená), příslušnost k instituci (žlutá), publikování (zelená) a autorství (fialová). Pro přehlednost obrázku je v grafu mnoho hran vynecháno.

2.2.2 Bibliografické databáze

Bibliografické záznamy obsahují minimálně jména publikací a jejich autorů a seznam referencí, které jsou v publikaci uvedené. Dále bývají obsaženy rok a místo publikování, časopis či sborník, afiliace autorů apod. Záznamy bývají sdruženy do kolekce, která je udržována tzv. bibliografickou databází. Ta se stará o sběr nových záznamů a případně o aktualizaci těch stávajících. Nejznámějšími bibliografickými databázemi (dále jen databáze) jsou Web of Science, Scopus, Google Scholar, CiteSeer, DBLP, Microsoft Academic Search a arXiv. Následující informace o databázích jsou čerpány z oficiálních webů databází a z (Bar-Ilan 2007; Bellis 2009; Fiala 2011).

Web of Science[4] (WoS), multioborová databáze Ústavu pro vědecké informace (*Institute for Scientific Information* – ISI) udržovaná firmou *Thomson Reuters*, je jednou z nejstarších a nejuznávanějších databází nejen vědeckých článků. Databáze vznikla v roce 1955, aby naplnila ideu, kterou zmínil Garfield (1955a, 1955b)[5]. Aktuálně shromažďuje vědecké články z více než 12000 vlivných časopisů a více než 150000 konferenčních sborníků a pokrývá tak zhruba 250 vědních disciplín. Indexovány jsou publikace od roku 1945, přičemž všechny indexované časopisy a sborníky podléhají přijímacímu řízení. WoS byl mnohokrát použit pro citační analýzu, jak ukazují např. (Yan a Ding 2009, 2011, 2012; Ding 2011a; Fiala 2012b, 2013, 2014; Zhu a Guan 2013; Nykl et al. 2014, 2015; Fiala et al. 2015).

Scopus[6], který vnikl v roce 2004 a je udržován firmou *Elsevier*, je multioborovou databází, která obsahuje více než 50 miliónů záznamů o vědeckých publikacích z více než 21000 zdrojů (časopisy a konference) od zhruba 5000 vydavatelů. Indexovány jsou manuálně vložené publikace ze všech vědních oborů od roku 1960. Pro citační analýzu Scopus použili např. (Elkins et al. 2010; Haddow a Genoni 2010; Franceschini et al. 2013).

[4] Databáze *Web of Science* - http://www.webofknowledge.com (*Web of Science* je dnes také znám jako *Web of Knowledge*)

[5] Historie Web of Knowledge - http://wokinfo.com/about/whoweare/

[6] Databáze *Scopus* - http://www.scopus.com

Google Scholar[7] (GS) společnosti *Google Inc.* je automatický systém shromažďující informace o vědeckých článcích, který vznikl v roce 2004. Indexovány jsou články ze všech vědních oborů od vydavatelů, kteří poskytují alespoň abstrakt článků zdarma. Počet indexovaných článků ani rozsah jejich let není znám, ale přístup do vyhledávání je zdarma. Použit v citační analýze byl GS např. v (Bar-Ilan 2007; Mingers a Lipitakis 2010; Amara a Landry 2012; Harzing 2013).

CiteSeer[8] byl prvním autonomním systémem, který indexuje vědecké publikace v elektronické podobě (Giles et al. 1998). *CiteSeer* byl vyvinut v *NEC Research Institute* (USA), je zaměřen na oblast počítačových věd, přístup do vyhledávání poskytuje zdarma a dle (Fiala 2011) v roce 2010 obsahoval téměř 33 miliónů záznamů. Rozsah indexovaných let není z vyhledávání jednoznačně patrný, protože se zde projevují nedůslednosti v datech – některé články obsahují místo čtyřciferného údaje o roku publikování údaj pouze dvouciferný. Nový CiteSeerX je stále ve verzi beta. Uplatnění CiteSeeru v citační analýze nalezneme např. v (Sidiropoulos a Manolopoulos 2005b; Zhou et al. 2007; Fiala 2011, 2012a; Nykl a Ježek 2012).

DBLP[9] (DataBases and Logic Programming) je databáze University v Trieru (Německo), která vznikla v roce 1993 (Ley 1993) a původně obsahovala pouze články z oblasti databázových systémů a logického programování. Dnes se DBLP soustředí na celou oblast počítačových věd. Vyhledávání v databázi, která aktuálně obsahuje téměř 2,4 miliónů manuálně vložených záznamů od roku 1936 [10], je přístupné zdarma. Některé části databáze lze také stáhnout v podobě XML souborů. DBLP byla mnohokrát použita v citační analýze, viz např. (Liu et al. 2005; Sidiropoulos a Manolopoulos 2005a, 2006; Fiala et al. 2008; Di Caro et al. 2012; Nykl a Ježek 2012).

Microsoft Academic Search[11] (MAS) společnosti *Microsoft* vnikl v roce 2009 a obsahuje více než 48 miliónů publikací od více než 20 miliónů autorů ze 14 oblastí výzkumu. Lze v něm nalézt např. i články Isaaca Newtona z roku 1672. Indexace publikací je automatická a přístup do vyhledávání je zdarma. MAS v citační analýze použil např. Jacsó (2011).

arXiv[12], který vznikl v roce 1991 pod záštitou knihovny Cornellovy univerzity (Ithaca, NY, USA) jako automatizovaný elektronický archív a distribuující server vědeckých článků, zahrnuje 6 oblastí výzkumu (fyzika, matematika, statistika, počítačové vědy, kvantitativní biologie a nelineární vědy[13]) a obsahuje články od roku 1992. Přístup do vyhledávání je zdarma, ale počet indexovaných článků není uveden[14]. Použití arXiv v citační analýze lze nalézt např. v (Sayyadi a Getoor 2009).

Vedle výše zmíněného základního porovnání těchto databází lze v literatuře nalézt i jejich porovnání při použití v citační analýze. Mingers a Lipitakis (2010) porovnávají WoS a GS v oblasti byznysu a managementu a docházejí k závěru, že GS pokrývá tuto oblast více než WoS. Harzing (2013) využitím

---

[7] Databáze *Google Scholar* - http://scholar.google.com
[8] Databáze *CiteSeer* (dnes označována jako *CiteSeerX*) - http://www.citeseer.com
[9] Databáze *DBLP* - http://dblp.uni-trier.de
[10] Statistiky vztahující se k databázi DBLP - http://dblps.uni-trier.de/~mwagner/statistics/
[11] Databáze *Microsoft Academic Search* - http://academic.research.microsoft.com
[12] Databáze *arXiv* - http://www.arxiv.org
[13] Nelineární vědy (*Nonlinear Sciences*) v arXiv obsahují kategorie: *Adaptation and Self-Organizing Systems*, *Cellular Automata and Lattice Gases*, *Chaotic Dynamics*, *Exactly Solvable and Integrable System*, *Pattern Formation and Solitons*.
[14] Statistiky vztahující se k databázi *arXiv* - http://arxiv.org/help/stats/2012_by_area/index

držitelů Nobelovy ceny porovnává WoS a GS z pohledu indexování vědních oborů a dochází k závěru, že GS je méně zaujatý než WoS a může např. napravit znevýhodněné postavení sociálních věd v bibliografických databázích. Bar-Ilan (2007) porovnává výpočet h-indexu Izraelských vědců na základě dat získaných z WoS, Scopus a GS, ale její závěr není jednoznačný.

Za zmínku stojí, že dle nařízení Úřadu vlády České republiky pro roky 2013 až 2015 (Úřad vlády ČR 2013) se pro hodnocení výzkumných organizací v České republice v části publikačních výsledků používají vědecké publikace zaznamenané v RIV (Rejstřík informací o výsledcích), které se nacházejí v databázích WoS, Scopus nebo ERIH (humanitní obory), či jsou v časopisech uvedených na seznamu Českých recenzovaných neimpaktovaných periodik. Při rozdělování bodů za vědecké publikace se u časopiseckých publikací přihlíží k Impact Factoru, pokud je časopis indexován ve WoS, nebo k SCImago Journal Ranku, pokud časopis není ve WoS, ale je ve Scopus.

2.2.3 Možnosti porovnání vytvořených pořadí

Pokud jsme vytvořili požadovaný graf, vyhodnotili ho zvolenými metodami a získali několik pořadí prvků zvolené entity, tak nás obvykle zajímá, jak lze získaná pořadí porovnat. Častým cílem je buďto pouhé zjištění podobnosti jednotlivých pořadí, nebo určení, která z použitých metod poskytuje v porovnání s referenčním seznamem lepší výsledné pořadí prvků. Dále pro názornost uvažujme porovnání dvou získaných pořadí autorů vědeckých publikací.

První možností porovnání pořadí je určení jejich statistické podobnosti. K tomuto účelu lze použít koeficienty korelace, přesněji Spearmanův (Spearman 1904) nebo Kendallův (Kendall 1938) koeficient pořadové korelace, které měří statistickou závislost dvou veličin. Veličinou zde rozumíme posloupnost prvků s určeným pořadím, přičemž obě zkoumané veličiny musí obsahovat totožné prvky. Porovnání je následně závislé pouze na vytvořeném pořadí a ne na hodnotách, dle kterých pořadí vzniklo. Koeficient pořadové korelace, který může nabývat hodnot z intervalu <+1;-1>, udává, do jaké míry jsou na sobě obě sledované veličiny funkčně závislé:

(+1) – obě veličiny jsou na sobě zcela funkčně závislé;
(0) – mezi zkoumanými veličinami není žádná funkční závislost;
(-1) – veličiny mají opačnou funkční závislost, tj. prvek, který je v první veličině na první pozici, je ve druhé veličině na pozici poslední atd.

Nejčastěji používaným koeficientem korelace pro porovnání výsledků citační analýzy je Spearmanův koeficient. Jeho použití nalezneme např. ve (Fiala et al. 2008; Ma et al. 2008; Ding et al. 2009) i jinde.

Chceme-li určit, která metoda hodnocení poskytuje „lepší“ pořadí, musíme zvolit referenční pořadí či hodnocení, které prohlásíme za nejlepší, a porovnávat, jak blízké je námi vytvořené pořadí k tomuto referenčnímu pořadí. V oblasti hodnocení časopisů či institucí narazíme na problém, že žádné referenční hodnocení neexistuje, vyjma žebříčku univerzit[15] (který je ale výsledkem kombinace mnoha faktorů, které se pro hodnocení univerzit používají). V oblasti hodnocení autorů lze jako referenční hodnocení použít různá ocenění udílená za vědeckou a publikační činnost, jako např. Nobelova cena udílená ve zkoumané oblasti. Pokud námi zkoumaná oblast výzkumu jsou počítačové

[15] Web s hodnocením univerzit z celého světa - http://www.webometrics.info

vědy, tak můžeme použít Turingovu cenu (*ACM A.M. Turing Award*[16]), Coddovu cenu (*ACM SIGMOD E.F. Codd Innovations Award*[17]), cenu *VLDB 10 Year Award*[18], cenu *ACM Test of Time*[19] nebo jiná podobná ocenění. Ceny *VLDB 10 Year Award* a *ACM Test of Time* mohou být použity i pro porovnání vytvořených pořadí publikací. Jako příklad můžeme uvést, že Nobelovu cenu pro určení kvality vytvořených pořadí autorů použil Harzing (2013), Turingovu cenu použili Fiala (2012b), Nykl et al. (2014) a Fiala et al. (2015), Coddovu cenu použili Sidiropoulos a Manolopoulos (2005a), Fiala et al. (2008) a Nykl et al. (2014) a ceny *VLDB 10 Year Award* a *ACM Test of Time* použili Sidiropoulos a Manolopoulos (2005a, 2006). Jinou možností by bylo využití osob z redakčních rad časopisů (Fiala et al. 2015) nebo z programových výborů konferencí (Liu et al. 2005). Vytvořená pořadí autorů mohou být následně porovnána na základě součtu, průměru, mediánu, minima či maxima z pozic, které ve vytvořeném pořadí obsadili držitelé zvoleného ocenění. Dále se můžeme zaměřit na porovnání pouze několika nejlepších pozic autorů, např. prvních dvacet. Zde se ptáme, kolik oceněných autorů je na nejlepších pozicích ve vytvořeném pořadí, viz např. (Yan a Ding 2009). Neposlední možností je využít úpravu metody zvané *Ranked Normalized Impact Factor* (viz část 2.3.1), která umožňuje na základě několika zvolených prvků porovnat i pořadí, která neobsahují shodný počet prvků.

2.3 Nejznámější metody citační analýzy

První oblastí zájmu při automatizované analýze bibliografických záznamů bylo hodnocení vědeckých časopisů na základě obdržených citací, přesněji na základě Impact Factoru (Garfield 1955a, 1955b). Další oblastí zájmu je hodnocení autorů, ve kterém je jednou z nejznámějších metod h-index (Hirsch 2005). Výhodou obou metod je jejich snadný neiterační výpočet, ale naopak nevýhodou může být, že při výpočtu nejsou využívány významnosti citující entit. Z toho důvodu můžeme říci, že obě metody měří popularitu (viz část 2.1). Popisu Impact Factoru je věnována následující část 2.3.1 a popisu h-indexu část 2.3.2. V obou částech jsou popsány dané metody a některé jejich modifikace. Iterační metody, které obvykle počítají prestiž bibliografických entit (např. SCImago Journal Rank), budou popsány v části 2.5.

Protože 2. kapitola shrnuje nejznámější metody pro měření významnosti vrcholů v bibliografických grafech, tak další skupinou neiteračních metod, kterou popíšeme v části 2.3.3, jsou míry centrality. Míry centrality zavedl Bavelas (1948), když se zabýval komunikací v malých skupinách osob a poukázal na vztah mezi strukturální centralitou a vlivem ve skupinových procesech. Centralita tedy je, vedle popularity a prestiže, další mírou, kterou lze hodnotit vrcholy grafu. Přestože míry centrality pocházejí z oblasti sociologie, tak, jak shrnují např. Yan a Ding (2009), byly již také mnohokrát použity v bibliometrii.

2.3.1 Impact Factor a jeho modifikace

Impact Factor[20] byl jednou z prvních metod pro měření významnosti časopisů, kterou *Institute for Scientific Information* (ISI) aplikoval v databázi *Web of Science* (WoS) a výsledky zobrazil v *Journal Citation Reports* (JCR). První zmínku o Impact Factoru nalezneme v (Garfield 1955a, 1955b), kde autor

[16] Web *ACM A. M. Turing Award* - http://amturing.acm.org
[17] Web *ACM SIGMOD Edgar F. Codd Innovations Award* - http://www.sigmod.org/sigmod-awards
[18] Web *VLDB 10 Year Award* - http://www-nishio.ist.osaka-u.ac.jp/vldb/archives/public/10year/10year.html
[19] Web *ACM Test of Time* - http://www.sigmod.org/sigmod-awards/sigmod-awards#time
[20] Journal Impact Factor a 5-Year Journal Impact Factor na webu ISI
- http://admin-apps.webofknowledge.com/JCR/help/h_impfact.htm

přichází s myšlenkou indexování článků obsažených ve vědeckých časopisech pro účely hodnocení významnosti časopisů. Dále se autor o Impact Factoru zmiňuje v (Garfield 1972, 1999). Impact Factor časopisu vyjadřuje, jak bylo vědecké smýšlení v daném roce ovlivněno články publikovanými v daném časopise dva roky před tím.

Impact Factor (IF, faktor vlivu) časopisu *j* v roce *y* (např. 2011) je počet citací z roku *y* na všechny články publikované v časopise *j* dva roky před tím (tj. 2010 a 2009) dělený počtem všech podstatných článků (tj. bez redakčních poznámek, úvodních článků, recenzí atd.) publikovaných v těchto dvou letech v časopise *j*. IF časopisu je tedy průměrným počtem citací, které v daném roce obdržely články publikované v předchozích dvou letech v daném časopise, a proto dle něj lze porovnávat různě objemné časopisy. Také jím lze odhalit časopisy obsahující pouze recenze (tyto časopisy s neúměrně vysokým IF nejsou zařazovány do WoS). JCR vedle hodnot IF časopisů, které byly vypočítány včetně samocitací časopisů, ukazuje i hodnoty IF vypočítané bez těchto samocitací (pozn.: pokud je rozdíl hodnot „příliš velký“, tak časopis obvykle bývá vyřazen z dalšího indexování).

IF a některé jeho další varianty lze zapsat vzorcem (2.1), kde $IF(j)_y$ je hodnota časopisu *j* v roce *y*, $C(j)^{y}_{<IntPub>}$ udává počet citací z roku *y* na články publikované v časopise *j* v rozmezí let daném intervalem *IntPub* a $P(j)_{<IntPub>}$ je počet článků publikovaných v časopise *j* v rozmezí let daném intervalem *IntPub*. Pokud chceme vzorcem (2.1) vyjádřit Impact Factor, tak *IntPub = <y-1; y-2>*.

$$IF(j)_y = \frac{C(j)^{y}_{<IntPub>}}{P(j)_{<IntPub>}} \tag{2.1}$$

V JCR můžeme nalézt také *5-Year Journal Impact Factor*, který používá publikace z pěti let (*IntPub = <y-1; y-5>*), a *Immediacy Index*[21] (index bezprostřednosti), který je jednoletou obdobou IF (*IntPub = <y; y>*) a indikuje, jak rychle jsou články v časopise citovány.

Extended Impact Factor (rozšířený faktor vlivu) můžeme nalézt v (Haddow a Genoni 2010), kde ho autoři na konkrétním příkladu pro rok 2007 definují jako: *„počet citací z let 2001 až 2007 na články publikované v časopise j v letech 2001 až 2006 dělený počtem článků časopisu j z let 2001 až 2006*“. Tato verze IF vyjadřuje průměrnou citovanost článků daného časopisu v rozmezí sedmi let, což ale celkem dlouho znevýhodňuje nové časopisy.

Modified Journal Diffusion Factor (modifikovaný faktor rozptylu časopisu), publikovaný v (Haddow a Genoni 2010), využívá stejného výpočtu jako Extended Impact Factor, ale neuvažuje počty souhlasných citací, tj.: pokud byl v časopise A citován časopis B, tak B získá od A jednu citaci, bez ohledu na to, kolikrát byl časopis B v časopise A citován. Původní Journal Diffusion Factor, jehož výpočet je složitější, než výpočet zde popsaný, lze nalézt v (Rowlands 2002) a další jeho modifikace v (Frandsen 2004; Sanni a Zainab 2011).

Aggregate Impact Factor of a Field (AIFF, sloučený faktor vlivu oblasti) je klasický Impact Factor, který je ale počítán pro celou zvolenou oblast či kategorii WoS (např. *Computer Science: Artificial Intelligence*), tj. využívá všechny citace z daného roku na články publikované ve zvolené oblasti dva roky před tím a počet těchto článků (Dorta-González a Dorta-González 2012). JCR vedle Aggregate

[21] Immediacy Index na webu ISI - http://admin-apps.webofknowledge.com/JCR/help/h_immedindex.htm

Impact Factoru uvádí i *Aggregate Immediacy Index* (sloučený index bezprostřednosti). Na základě těchto metod lze zjistit, které vědní oblasti jsou ve sledovaném roce nejvíce rozvíjené.

Egghe a Rousseau (2003) představili *Global Impact Factor* (globální/souhrnný faktor vlivu) oblasti a s jeho pomocí zavádli *Relative Impact Factor* (relativní faktor vlivu) časopisu, který lze použít pro porovnání časopisů z různých vědních oblastí. Global Impact Factor oblasti je počítán obdobně jako AIFF, ale není implicitně definováno, z jakého rozsahu let jsou použity citace a citované publikace. Zvolené rozsahy let se použijí i pro výpočet obdoby Extended Impact Factoru časopisu. Relative Impact Factor časopisu je následně podílem takto vzniklého Extended Impact Factoru časopisu a Global Impact Factoru oblasti, ve které se časopis nachází.

Ranked Normalized Impact Factor (RNIF, pořadím normalizovaný faktor vlivu) navrhli Abrizah et al. (2013) pro porovnání postavení časopisu v různých bibliografických databázích. RNIF je počítán dle vzorce (2.2), kde $RNIF_j^d$ je hodnota Ranked Normalized Impact Factoru časopisu *j* v databázi *d*, K_j^d je počet časopisů v kategorii časopisu *j* databáze *d* a R_j^d je pozice časopisu *j* v pořadí jeho kategorie v databázi *d.* Pořadí časopisů v kategorii je vytvořeno dle Impact Factoru (ve WoS) nebo dle SCImago journal ranku (ve Scopus, viz část 2.5.8). Abrizah et al. (2013) využitím RNIF porovnávají WoS a Scopus a uvádějí příklad časopisu, který je na 60 pozici v JCR kategorii obsahující 77 časopisů a na 48 pozici ve Scopus kategorii obsahující 128 časopisů. Zvolený časopis má ve WoS RNIF=0,234, což znamená, že v dané WoS kategorii je 76,6% časopisů na lepší pozici, než zvolený časopis. Ve Scopus je lepších pouze 36,7% časopisů. (Pozn.: obdobou RNIF lze také porovnat pozice zvoleného prvku či prvků libovolné bibliografické entity v pořadích vytvořených libovolnými metodami. Porovnávaná pořadí navíc nemusejí obsahovat shodný počet prvků.)

$$RNIF_j^d = \frac{K_j^d - R_j^d + 1}{K_j^d} \tag{2.2}$$

Cited Half-Life[22] (poločas obdržených citací) a *Citing Half-Life*[23] (poločas obsažených citací) jsou dalšími metodami, které lze nalézt v JCR. Cited Half-Life udává počet roků (počítáno od aktuálního roku), ve kterých časopis obdržel 50% všech citací z aktuálního roku. Výpočet lze lépe pochopit z konkrétního příkladu: *„pokud je v roce 2012 Cited Half-Life hodnota časopisu 5,25, tak 50% všech citací, které časopis obdržel v roce 2012, směřuje na jeho články z let 2012 až 2008 (5 let) a z roku 2007 je použita čtvrtina citací"*. Obdobně je tomu u Citing Half-Life, pouze se nepočítají citace, které časopis obdržel (vstupní hrany), ale citace, které časopis obsahoval (výstupní hrany). Obě metody měří aktuálnost obsahu časopisu – Citing Half-Life z pohledu informací obsažených v článcích daného časopisu (použité zdroje) a Cited Half-Life z pohledu využití článků, které daný časopis obsahoval.

2.3.2 H-index a jeho modifikace

Vedle metod pro hodnocení časopisů vznikly i metody primárně určené pro hodnocení autorů vědeckých článků. Nejznámější metodou je Hirsch-index či jen h-index, který navrhl Hirsch (2005) pro účely kvantifikování individuálního vědeckého přínosu. H-index je definován takto: *„Autor má h-index*

[22] Cited Half-Life na webu ISI - http://admin-apps.webofknowledge.com/JCR/help/h_ctdhl.htm
[23] Citing Half-Life na webu ISI - http://admin-apps.webofknowledge.com/JCR/help/h_ctghl.htm

o velikosti h, pokud h z jeho publikací obdrželo alespoň h citací a žádná další jeho publikace nemá více než h citací". Samocitace autorů by při výpočtu neměly být použity. Nejlepší publikace autora, které určily velikost jeho h-indexu, tvoří množinu nazývanou *h-jádro*. Protože h-index autora v průběhu let pouze stagnuje nebo roste, tak lze o h-indexu hovořit jako o míře vyspělosti autora. *Normalizovaný h-index* (Sidiropoulos a Katsaros 2008) je h-index autora dělený celkovým počtem článků autora.

Ch-index (či *citer index*) uvedený v (Ajiferuke a Wolfram 2009) uvažuje pouze počty autorů (každý počítán jen jednou), kteří publikaci citovali. Zbytek výpočtu je stejný jako u h-indexu. H-index a ch-index porovnali Franceschini et al. (2010) a zjistili, že ch-index není citlivý na samocitace a opakující se citace a je také méně citlivý na chyby v bibliografické databázi (např. duplicitní záznamy) než h-index.

Egghe (2006, 2013) představil *g-index*, jehož výpočet je blízký výpočtu h-indexu, s tím rozdílem, že se používá druhá mocnina souhrnného počtu citací: „*autor má g-index o velikosti g, jestliže g z jeho top článků obdrželo v součtu alespoň g^2 citací".* Jak Egghe (2006) poznamenává, platí $g \geq h$, přičemž g-index přebírá všechny dobré praktiky h-indexu a navíc zohledňuje množství citací nejlepších článků autora (pozn.: h-index nezohledňuje skutečnost, že nejcitovanější článek autora může mít daleko více citací, než jeho další články). Tol (2008) představuje *successive g_1-index* (g_1-index „úspěšnosti"), který je počítán pro výzkumné oddělení nebo skupinu tak, že: „*skupina má g_1-index úspěšnosti o velikosti g_1, pokud g_1 z jejích výzkumníků má g-index o velikosti alespoň g_1*".

A-index (*average index*; průměrný h-index), publikovaný v (Jin et al. 2007), je počítán jako součet citací článků, které náleží do h-jádra autora, dělený velikostí h-indexu. Další metodou představenou Jin et al. (2007) je *R-index*, který je počítán jako odmocnina ze součtu citací článků náležících do h-jádra autora. Autoři dále ukazují *AR-index* či *age-dependent R-index* („na věku závislý" R-index), který je počítán jako odmocnina ze součtu podílů citací článků náležících do h-jádra autora a věku článku (pozn.: věk je celočíselný počet let existence článku, tj. nejmenší věk je 1). AR-index zamezuje neustálému zvyšování hodnoty autora v průběhu let, protože pozvolna znevýhodňuje starší články.

Autor h-indexu představil také jeho variantu zvanou *ħ-index* („h s pruhem"), viz (Hirsch 2010), kterou definuje takto: „*vědec má index o velikosti ħ, pokud jeho ħ publikací náleží do jeho ħ-jádra. Publikace náleží do autorova ħ-jádra, pokud má alespoň ħ citací a navíc náleží do ħ-jádra všech svých autorů*". Výpočet začíná s vypočtenými h-indexy autorů, přesněji s jejich h-jádry, ze kterých se postupně odstraňují publikace, které nejsou v h-jádrech všech svých autorů. Tím se může snížit h-index autora a do utvářeného ħ-jádra se tak mohou dostat publikace, které mají méně citací než publikace v původním h-jádře, ale jsou v ħ-jádrech všech svých autorů (platí $\hbar \leq h$). Výhodou ħ-indexu je, že produktivnějším autorům penalizuje spolupráci se začínajícími autory.

Z uvedených variant h-indexu lze usoudit, že h-index lze snadno modifikovat pro různé účely hodnocení, přičemž modifikace se často týkají míry zohlednění citovanosti či produktivity autora ve výsledném hodnocení. Odkazy na další modifikace h-indexu, jejich studie a použití pro hodnocení autorů, výzkumných skupin, univerzit, časopisů, témat, států apod. obsahuje např. (Alonso et al. 2009).

2.3.3 Míry centrality

Koncept centrality více rozpracoval Freeman (1977), který pro účely určení centrálních vrcholů v sociální síti definoval sadu metod či měr centrality (*Centrality Measures*) založených na

betweenness (volně přeloženo: „mezilehlost“). V sociální síti vrcholy obvykle zastupují osoby nebo jejich skupiny a hrany určitý akt jejich vzájemné interakce (např. spřízněnost). Centralita vrcholu udává, do jaké míry je daný vrchol schopen ovlivnit probíhající dění (např. tok informací mezi osobami). V následující práci Freeman (1979) shrnul základní míry centrality, kterými jsou: degree („stupeň“), closeness („blízkost“) a betweenness centralita. Dobrý popis měr centrality a odkazy na jejich uplatnění ve vyhodnocení bibliografických grafů lze nalézt např. v (Yan a Ding 2009), kde autoři aplikují míry centrality na graf spoluautorství autorů a určují nejvíce centrální autory. Různé další úpravy měr centrality lze nalézt např. v (Hanneman a Riddle 2005). Poznamenat můžeme, že míry centrality se obvykle aplikují na neorientované grafy, ale neorientovanost grafu není podmínkou.

Degree centralita (Freeman 1979; Yan a Ding 2009) je počtem hran nebo součtem vah hran, které se váží na daný vrchol. Rozlišovat můžeme „prosté“ degree nebo vážené degree (*weighted degree*). Pokud je graf orientovaný, lze dále rozlišovat *in-degree* a *out-degree* centralitu, kde „in“ zastupuje vstupní hrany a „out“ hrany výstupní. Obecně uvažujeme, že vrchol s vysokým počtem hran je ve struktuře grafu více centrální a má tak větší schopnost ovlivňovat ostatní. V orientovaném grafu lze vrchol, na který vede mnoho hran (vysoké in-degree), označit za prominentní či přední. V analogii s počítáním obdržených citací lze vrchol s vysokým in-degree také označit za populární. Naopak vrchol, ze kterého vede mnoho hran (vysoké out-degree), lze označit za vlivný vrchol – má vyšší šanci ovlivnit ostatní. Vždy ale záleží na konkrétním významu hrany a její orientace.

Vrcholy různě velkých grafů můžeme s využitím hodnot degree porovnávat až po jejich normalizaci. Tu provedeme tak, že degree všech vrcholů vydělíme maximálním možným počtem hran, které vrchol může v příslušném grafu mít, tj. *(n-1)*, kde *n* je počet všech vrcholů grafu (Freeman 1979; Ferrara 2012). Vedle základní Freemanovy varianty degree centrality existuje i její varianta, zmíněná např. v (Hanneman a Riddle 2005), která při výpočtu používá i vazby sousedních vrcholů. V česky psané literatuře bývá degree centralita také označována jako *centralita měřená stupněm vrcholu*.

Closeness centralitu (Freeman 1979; Yan a Ding 2009) lze chápat jako míru toho, jak blízko je vrchol ke všem ostatním vrcholům grafu, což může být interpretováno např. jako míra schopnosti vrcholu rychle rozšířit informaci po celém grafu. Closeness centralitu lze zapsat vzorcem (2.3), kde $C_c(u)$ je hodnota closeness centrality vrcholu *u, V* je množina všech vrcholů grafu či jeho zvolené souvislé komponenty, viz dále, a $d(u,v)$ je délka nejkratší cesty z vrcholu *u* do vrcholu *v*. Čím blíže je vrchol všem ostatním vrcholům grafu, tím má vyšší hodnotu closeness centrality. Pokud celý graf není jednou souvislou komponentou, tak je potřeba vypočítat closeness centrality vrcholů v každé jeho souvislé komponentě zvlášť a následně vypočítané hodnoty vrcholů normalizovat velikostí komponent, tj. v případě closeness centrality je vynásobit *(n-1)*, kde *n* je počet vrcholů komponenty, ve které se vrchol nachází (Freeman 1979). Normalizovanou closeness centralitou lze porovnávat i vrcholy z různě velkých grafů.

$$C_C(u) = \frac{1}{\sum_{v \in V} d(u,v)} \tag{2.3}$$

Pokud je graf vážený, tak při výpočtu closeness centrality musíme znát význam vah hran. Jestliže váhy hran vyjadřují vzdálenost (tj. čím větší váhu hrana má, tím jsou její koncové vrcholy vzdálenější, např. vzdálenost dvou měst), tak výpočet neměníme. Pokud ale váhy hran vyjadřují spříznění či blízkost (tj. čím větší váhu hrana má, tím bližší si jsou její koncové vrcholy, např. počet společných publikací

autorů), tak při výpočtu délek nejkratších cest *d(u,v)* musíme sčítat obrácené hodnoty vah hran. Vzorec closeness centrality byl také dále zkoumán a vznikly i jeho další varianty, které jsou zmíněny např. v (Hanneman a Riddle 2005). V česky psané literatuře bývá closeness centralita také nazývána *centralita měřená blízkostí polohy ke středu*.

Betweenness centralita (Freeman 1979; Yan a Ding 2009) vyjadřuje schopnost vrcholu propojovat (rozdílné) skupiny vrcholů. V sociální síti přátel může být příkladem osoby s vysokou betweenness centralitou osoba *O*, která navštěvuje dva zájmové kroužky (např. volejbal a fotbal) a tím propojuje dvě skupiny osob – pokud sociální síť přátel obsahuje pouze osoby z těchto dvou kroužků a osoba *O* je jedinou osobou, která navštěvuje oba kroužky, tak má osoba *O* největší betweenness centralitu. Tato osoba může do značné míry ovlivňovat probíhající dění např. blokováním (nežádoucích) zpráv, vybíráním poplatků „za spojení" nebo izolováním osob, které nemají jinou možnost, jak se dostat ke sdílené informaci.

Při výpočtu betweenness centrality nás zajímá, na kolika nejkratších cestách mezi všemi dvojicemi různých vrcholů daný vrchol leží. Výpočet znázorňuje vzorec (2.4), kde $C_B(u)$ je betweenness centralita vrcholu *u*, $g_{j,u,k}$ je počet nejkratších cest mezi vrcholy *j* a *k*, které vedou přes vrchol *u*, a $g_{j,k}$ je počet všech nejkratších cest mezi vrcholy *j* a *k* (tj. počet všech cest mezi *j* a *k*, které mají totožnou vzdálenost a jsou mezi danými dvěma vrcholy nejkratší).

$$C_B(u) = \sum_{j,k \in V; j \neq k \neq u} \frac{g_{j,u,k}}{g_{j,k}} \qquad (2.4)$$

Porovnávat hodnoty betweenness centrality vrcholů z různě velkých grafů lze opět až po jejich normalizaci. Tu provedeme vydělením hodnot betweenness centrality vrcholů maximální možnou hodnotou betweenness centrality, kterou by v daném grafu mohl vrchol získat. Maximální hodnotu betweenness centrality získá vrchol, přes který v grafu, který obsahuje cesty mezi všemi dvojicemi vrcholů, vedou všechny nejkratší cesty. Maximální betweenness centralita vrcholu je tedy *(n-1)*(n-2)* pro orientovaný graf a *(n-1)*(n-2)/2* pro graf neorientovaný, kde *n* je počet vrcholů grafu (Freeman 1979). Je-li graf vážený, tak při zjišťování nejkratších cest musíme opět brát v úvahu význam vah hran a správně určit nejkratší cesty. Další varianta betweenness centrality (Freeman et al. 1991) při výpočtu používá všechny cesty mezi dvojicemi vrcholů, na kterých daný vrchol leží, a tím uvažuje, že při komunikaci mezi dvěma vrcholy nemusí být vždy použity pouze nejkratší cesty. Tento postup je ale výpočetně náročný. V česky psané literatuře bývá betweenness centralita také nazývána *centralita měřená středovou mezipolohou*.

Mezi nejznámější programy pro analýzu sociálních sítí či obecně grafů patří *UCINET*[24] (Borgatti et al. 2002), *Pajek*[25] (Batagelj a Mrvar 1998) a *Gephi*[26] (Bastian et al. 2009), které vedle výpočtu měr centrality umožňují i výpočet základních statistik pro porovnání celých grafů, jako jsou např. poloměr grafu[27], hustota grafu[28], rozložení stupňů vrcholů[29] apod. Pajek XXL je navíc použitelný pro analýzu

[24] Program *UCINET* - https://sites.google.com/site/ucinetsoftware/home
[25] Program *Pajek* - http://mrvar.fdv.uni-lj.si/pajek/
[26] Program *Gephi* - https://gephi.org
[27] V grafu jsou určeny délky nejkratších cest mezi všemi dvojicemi vrcholů a následně je mezi těmito délkami vyhledána maximální délka, která je poloměrem daného grafu.

objemných grafů a Gephi pro vizualizaci grafů. Různé další metody, které lze použít k analýze sociálních sítí, shrnují např. (Hanneman a Riddle 2005; Aggarwal 2011; Ferrara 2012).

2.4 Algoritmus PageRank

Algoritmus PageRank (Brin a Page 1998; Page et al. 1999), dále také jen PageRank, byl původně vyvinut pro webové vyhledávače, ve kterých slouží pro určení významností webových stránek. Tyto významnosti jsou následně používány při řazení webových stránek ve výsledcích vyhledávání. Dnes PageRank používá např. webový vyhledávač Google.com a některé další vyhledávače. Jak Page et al. (1999) uvádějí, koncept PageRanku vychází z citační analýzy. Měli bychom také zmínit, že PageRank je aplikací Markovova řetězce, viz detailněji např. v (Langville a Meyer 2006), a PageRanku podobný koncept navrhli už dříve Pinski a Narin (1976), ale potenciál tohoto konceptu využili až Brin a Page (1998), když navrhli PageRank.

PageRank při určování významnosti webové stránky používá hypertextové odkazy, které na stránku odkazují, a významnost stránek, ze kterých tyto odkazy vedou. Z matematického hlediska je vyhodnocován graf, jehož vrcholy reprezentují webové stránky a hrany vyjadřují, že z jedné webové stránky vede hypertextový odkaz na stránku jinou. Interní hypertextové odkazy (lze použít označení z bibliografie „samocitace"), tj. odkazy, které odkazují na stránku, na které se nacházejí, se při vyhodnocování Webu nepoužívají.

2.4.1 Matematický popis algoritmu PageRank

Algoritmus PageRank lze popsat buďto maticovým zápisem, který je užitečný pro matematické zkoumání algoritmu (např. jeho konvergence, urychlení výpočtu atd.), nebo zápisem výpočtu pro jeden prvek, který je užitečný pro „snazší" pochopení a implementaci PageRanku. Dále jsou ukázány oba dva typy zápisů, přičemž použity byly maticové zápisy a důkazy uvedené v (Langville a Meyer 2006). Výpočet PageRanku je iterační a končí po zvoleném počtu iterací nebo po dosažení zvolené přesnosti PageRankových hodnot jednotlivých vrcholů.

Základní vzorec PageRanku, který Page et al. (1999) navrhli, je vzorec (2.5), kde $PR_x(A)$ je hodnota PageRanku vrcholu A v iteraci x, U_A je množina všech vrcholů odkazujících na vrchol A a N_u je počet výstupních hran vrcholu u. Suma ve vzorci (2.5) zastupuje citační prestiž vrcholu. Tomuto zápisu odpovídá maticový zápis uvedený ve vzorci (2.6), kde $\boldsymbol{\pi}^{(x)}$ je vektor PageRankových hodnot všech vrcholů grafu v iteraci x a $\boldsymbol{H}$ je řádkově normalizovaná matice sousednosti zastupující vyhodnocovaný graf.

$$PR_{x+1}(A) = \sum_{u \in U_A} \frac{PR_x(u)}{N_u} \tag{2.5}$$

$$\boldsymbol{\pi}^{(x+1)T} = \boldsymbol{\pi}^{(x)T}\boldsymbol{H} \tag{2.6}$$

---

[28] *Hustota grafu* je počet hran grafu dělený počtem všech hran, které by graf mohl obsahovat, tj. *n*(n-1)* v orientovaném grafu a *n*(n-1)/2* v neorientovaném grafu, přičemž *n* je počet vrcholů grafu.

[29] *Rozložení stupňů vrcholů* je znázorněno diagramem, který má na vodorovné ose stupně vrcholů a na svislé ose počty vrcholů, které mají daný stupeň.

Nyní použijeme analogii s hodnocením webových stránek, zavedeme pojem webový surfař (tj. uživatel Webu, který klikáním na hypertextové odkazy prochází Web – dále jen surfař) a budeme uvažovat, že se Webem pohybuje nekonečně mnoho surfařů. Pokud PageRanku nastavíme $\forall PR_0(A) = 1/|V|$, kde $|V|$ je velikost množiny všech webových stránek vyhodnocovaného grafu, tak vypočtená hodnota PageRanku webové stránky udává pravděpodobnost, s jakou se surfaři po nekonečně mnoho krocích nacházejí právě na dané stránce. Součet hodnot PageRanku všech vrcholů grafu je roven 1, tj. 100%, což znamená, že každý surfař je v některém z vrcholů grafu.

Prvním problémem, se kterým se vzorce (2.5) a (2.6) potýkají, je úbytek celkové hodnoty PageRanku vlivem vrcholů, které nemají žádné výstupní hrany, tzv. **slepé vrcholy** (*dangling nodes)*. V analogii se surfaři lze říci, že surfaři, kteří jsou aktuálně ve slepých vrcholech, nemohou pro svůj pohyb využít žádný hypertextový odkaz, a proto budou v následující iteraci PageRanku na stálo „vyloučeni" z grafu. Z toho důvodu se součet hodnot PageRanku všech vrcholů grafu přestane rovnat jedné, přičemž celkový úbytek hodnoty PageRanku značí, jaká část surfařů už je z grafu vyloučena. Problém slepých vrcholů lze řešit např. těmito způsoby:

1) vytvoření stoku - vytvoříme v grafu nový vrchol (tzv. stok) se smyčkou (samocitační hranou) a všem slepým vrcholům přidáme výstupní hranu směřující na tento vrchol.
2) normalizace – po každé iteraci PageRanku normalizujeme hodnoty všech vrcholů grafu tak, aby jejich součet byl roven jedné.
3) rovnoměrné rozdělení – každému slepému vrcholu přidáme výstupní hrany na všechny vrcholy grafu (i na sebe sama).

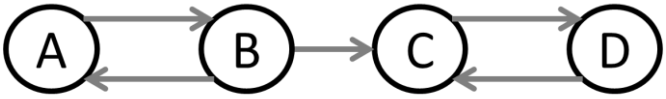
První způsob (vytvoření stoku) není příliš vhodný, protože může vést k situaci, kdy nově vytvořený vrchol získá celou hodnotu PageRanku, tj. bude mít hodnotu 1. Druhý způsob (normalizace) lze, po odstranění problému Rank sink (viz dále), použít, ale jeho nevýhodou je, že každému vrcholu je normalizací přidána jiná hodnota. Z těchto důvodů se obvykle používá způsob třetí, tj. rovnoměrné rozdělení hodnot slepých vrcholů. V maticovém zápisu doplníme výstupní hrany slepým vrcholům přímo do vyhodnocovaného grafu, který je zastoupen řádkově normalizovanou maticí sousednosti $\boldsymbol{H}$, čímž vznikne matice $\boldsymbol{S}$, jak ukazuje vzorec (2.7). Nově vzniklou maticí $\boldsymbol{S}$ nahradíme matici $\boldsymbol{H}$ ve vzorci (2.6). Matice $\boldsymbol{S}$ zastupuje vyhodnocovaný graf, ve kterém jsou všem slepým vrcholům doplněny hrany na všechny vrcholy grafu. Ve vzorci (2.7) pro výpočet matice $\boldsymbol{S}$ je $\boldsymbol{a}$ vektor „sleposti" vrcholů, kde a_i je rovno jedné, pokud je vrchol *i* slepým vrcholem, jinak je a_i rovno nule, a $\boldsymbol{e}$ je jednotkový vektor.

$$\boldsymbol{S} = \boldsymbol{H} + \boldsymbol{a}\left(\frac{1}{|V|}\boldsymbol{e}^T\right) \tag{2.7}$$

Při výpočtu PageRanku využitím matematického zápisu výpočtu pro jeden prvek nemusíme výstupní hrany ze slepých vrcholů doplňovat přímo do grafu, ale stačí s nimi pouze počítat, jak ukazuje námi navržený vzorec (2.8). Protože hodnota předávaná od slepých vrcholů se v průběhu iterace nemění, tak výpočet lze urychlit. Ve vzorci (2.8) je *D* množina všech slepých vrcholů grafu, *V* je množina všech vrcholů grafu a $|V|$ velikost množiny *V*. První suma ve vzorci (2.8) zastupuje část prestiže, kterou vrchol získá díky svým vstupním hranám, a druhá suma část prestiže, kterou vrchol získá ze slepých vrcholů.

$$PR_{x+1}(A) = \sum_{u \in U_A} \frac{PR_x(u)}{N_u} + \frac{1}{|V|} \sum_{s \in D} PR_x(s) \tag{2.8}$$

Vzorce (2.5) až (2.8) ovšem stále neřeší problém zvaný *Rank sink*, který se projevuje, pokud vrcholy ve skupině odkazují samy na sebe, ale neodkazují vně skupiny, přičemž skupina je odkazována z vnější. Rank sink ilustruje obrázek 2.4 s příkladem, ve kterém vrcholy A a B po nekonečně mnoho iteracích výpočtu PageRanku předají své hodnoty PageRanku vrcholům C a D a PageRank vrcholů A a B bude (díky zaokrouhlovacím chybám) roven nule. Dalším problémem je, že pokud vrcholy C a D nebudou mít každý přesně polovinu celkové hodnoty PageRanku, tak si v každé iteraci vymění své hodnoty a nikdy nenastane ustálený stav, tj. algoritmus nebude konvergovat.



Obrázek 2.4: Příklad grafu, ve kterém při použití některého ze vzorců (2.5) až (2.8) vznikne Rank sink.

Problém Rank sink Brin a Page (1998) vyřešili navržením modelu náhodných webových surfařů, kteří se Webem pohybují klikáním na hypertextové odkazy nebo použitím tzv. teleportu, tj. přechodem na náhodnou webovou stránku zadáním její URL adresy přímo do webového prohlížeče. Analýzou chování reálných uživatelů Webu autoři zjistili, že teleport uživatelé využívají průměrně jednou za 7 kroků. Proto ve svém algoritmu stanovili užití teleportu s pravděpodobností 15% ≈ 1/7. Do algoritmu PageRank byl model náhodných surfařů vložen konstantou *d* zvanou **faktor tlumení**. Surfaři tak s pravděpodobností *d* následují hypertextové odkazy nebo s pravděpodobností *(1-d)* použijí teleport. Faktor tlumení je tedy obvykle nastaven na hodnotu 0,85, ale tato hodnota může být změněna např. při využití personalizace (neuniformní úprava náhodného teleportu, viz část 2.4.2).

Vzorec (2.9) ukazuje, jak byl vzorec (2.8) doplněn o faktor tlumení a hodnota, kterou každý vrchol získá díky teleportu, normalizována počtem všech vrcholů grafu. První část vzorce (2.9) zastupuje hodnotu, kterou vrchol získá díky náhodnému teleportu (jedná se o statickou část PageRanku), a druhá část vzorce zastupuje hodnotu prestiže, kterou vrcholu získá díky vyhodnocení grafu (dynamická část PageRanku).

$$PR_{x+1}(A) = \frac{(1-d)}{|V|} + d \cdot \left(\sum_{u \in U_A} \frac{PR_x(u)}{N_u} + \frac{1}{|V|} \sum_{s \in D} PR_x(s) \right) \tag{2.9}$$

Do maticového zápisu PageRanku zakomponujeme náhodný teleport upravením matice $\boldsymbol{S}$, což ukazuje vzorec (2.10), a nově vzniklou maticí $\boldsymbol{G}$ nahradíme matici $\boldsymbol{H}$ ve vzorci (2.6). Matice $\boldsymbol{G}$ zastupuje vyhodnocovaný graf, ve kterém je zakomponován náhodný teleport a všem slepým vrcholům jsou doplněny hrany na všechny vrcholy grafu.

$$\boldsymbol{G} = \frac{(1-d)}{|V|} \boldsymbol{e}\boldsymbol{e}^T + d\boldsymbol{S} \tag{2.10}$$

Poslední nepřesností vzorce (2.9) je, že každá hrana vyhodnocovaného grafu, má ve výpočtu PageRanku stejnou váhu, tj. jsou-li na stránce např. 4 rozdílné odkazy, tak uvažujeme, že každý z nich bude použit s pravděpodobností ¼. V prostředí Webu obvykle počet stejných odkazů zanedbáváme, ale pokud bychom chtěli některý z odkazů zvýhodnit (např. počtem výskytů), tak musíme vzorec PageRanku doplnit o váhy hran. To ukazuje vzorec (2.11), kde w_{utoA} je váha hrany vedoucí z vrcholu *u* do vrcholu *A* a w_{uout} je součet vah všech výstupních hran vrcholu *u*. Naší úpravou v prezentovaných vzorcích, které popisují výpočet PageRanku pro jeden prvek, bylo přidání části s ošetřením slepých vrcholů. Tato část umožňující urychlení výpočtu nebyla v původních pracích (Brin a Page 1998; Page et al. 1999; Langville a Meyer 2006) použita.

$$PR_{x+1}(A) = \frac{(1-d)}{|V|} + d \cdot \left(\sum_{u \in U_A} \frac{PR_x(u) \cdot w_{u\text{toA}}}{w_{u\text{out}}} + \frac{1}{|V|} \sum_{s \in D} PR_x(s) \right) \qquad (2.11)$$

V maticovém zápisu žádnou úpravu, přidávající do výpočtu váhy hran, provádět nemusíme, protože váhy hran můžeme zapsat přímo do matice sousednosti ***H*** předtím, než ji řádkově znormalizujeme.

S odkazem na matematické důkazy v (Langville a Meyer 2006) můžeme říci, že výpočet PageRanku vzorci (2.9) až (2.11) konverguje k jedinečnému výsledku bez ohledu na výchozí nastavení PageRankových hodnot. Přesto se ale obvykle před první iterací nastavují hodnoty vrcholů na hodnotu $1/|V|$, či nejlépe na hodnoty blízké konečnému výsledku.

2.4.2 Personalizace PageRanku

V některých případech můžeme požadovat, aby určité vrcholy byly algoritmem PageRank v průběhu výpočtu zvýhodněny. Toho můžeme docílit neuniformním rozdělením faktoru tlumení (část *1-d*), což bývá označováno jako **personalizace**. Pojem personalizace zavedli Page et al. (1999), když do výpočtu PageRanku chtěli zakomponovat různé potřeby či vlastnosti uživatelů. PageRank doplněný o personalizaci znázorňuje vzorec (2.12), kde *P* je množina personalizací všech vrcholů grafu a p_A je hodnota personalizace vrcholu *A*. První část vzorce (2.12) zastupuje hodnotu vrcholu získanou díky personalizaci (statická část PageRanku) a druhá část zastupuje hodnotu prestiže vrcholu získanou vyhodnocením grafu (dynamická část PageRanku). Do jaké míry bude konečná hodnota PageRanku tvořena statickou či dynamickou částí udává faktor tlumení.

$$PR_{x+1}(A) = \frac{(1-d) \cdot p_A}{\sum_{p \in P} p} + d \cdot \left(\sum_{u \in U_A} \frac{PR_x(u) \cdot w_{u\text{toA}}}{w_{u\text{out}}} + \frac{1}{|V|} \sum_{s \in D} PR_x(s) \right) \qquad (2.12)$$

Ve vzorci (2.12) předpokládáme, že nenastane situace, kdy $\sum p = 0$. Pokud by tato situace mohla nastat, tak můžeme využít např. náš vzorec (2.13), ale možností řešení je více a vždy záleží na konkrétním použití algoritmu.

$$PR_{x+1}(A) = \frac{(1-d) \cdot (1+p_A)}{|V| + \sum_{p \in P} p} + d \cdot \left(\sum_{u \in U_A} \frac{PR_x(u) \cdot w_{u\text{toA}}}{w_{u\text{out}}} + \frac{1}{|V|} \sum_{s \in D} PR_x(s) \right) \qquad (2.13)$$

V maticovém zápisu PageRanku zakomponujeme personalizaci odpovídající vzorci (2.12) do matice sousednosti $\boldsymbol{G}$ namísto náhodného teleportu, jak ukazuje vzorec (2.14), který je úpravou vzorce (2.10), a nově vzniklou maticí $\boldsymbol{G_p}$ opět nahradíme matici sousednosti $\boldsymbol{H}$ ve vzorci (2.6). Ve vzorci (2.14) je $\boldsymbol{v}$ normalizovaný vektor personalizací a $\boldsymbol{S}$ matice sousednosti získaná vzorcem (2.7). Matice $\boldsymbol{G_p}$ zastupuje vyhodnocovaný graf, ve kterém je zakomponována personalizace a všem slepým vrcholům doplněny hrany na všechny vrcholy grafu.

$$\boldsymbol{G_p} = (1-d)\boldsymbol{e}\boldsymbol{v}^T + d\boldsymbol{S} \tag{2.14}$$

Pěkný příklad použití personalizace v citační analýze ukazují např. Yan a Ding (2011), kteří PageRankem vyhodnocují graf spoluautorství autorů a personalizaci používají tak, že p_A představuje počet citací, které obdržel autor *A*, tj. popularitu autora A (viz část 2.1). Využitím faktoru tlumení *d*=0,55 autoři vkládají popularitu do výpočtu PageRanku hodnotícího spoluautorství.

2.4.3 Citlivost PageRanku na změnu parametrů

Informace o citlivosti algoritmu PageRank na změnu parametrů, přejaté z (Langville a Meyer 2006), jsou důležité zejména, pokud vyhodnocujeme graf s dynamicky se měnící strukturou, např. Web.

Výpočet PageRanku značně ovlivňuje velikost faktoru tlumení *d*, který určuje, do jaké míry se ve výpočtu PageRanku zohlední struktura grafu a do jaké míry se zohlední personalizace. Pokud je faktor tlumení malý, tak výpočet PageRanku není příliš citlivý na jeho malou změnu a je více citlivý na změnu personalizace (tj. změnu vektoru $\boldsymbol{v}$). Naopak, pokud je faktor tlumení velký, tak je výpočet PageRanku hodně citlivý na jeho malou změnu a je také více citlivý na změnu struktury grafu (tj. změnu matice $\boldsymbol{H}$). Navíc, čím blíže jedné faktor tlumení je, tím více iterací je potřeba k dosažení zvolené přesnosti výsledku (Langville a Meyer 2006; Nykl 2011).

2.5 Další metody pro měření významnosti vrcholů grafu

V této části 2. kapitoly je shrnut aktuální stav poznání v úloze určování významných vrcholů grafu iteračními metodami, přičemž důraz je kladen na metody, které byly nebo mohou být uplatněny v bibliometrii. Tyto metody obvykle vyhodnocují citační grafy a pro určení hodnot významnosti vrcholů používají hodnoty vrcholů, které na ně odkazují. Jak bylo zmíněno v části 2.1, takto vypočtenou významnost vrcholu můžeme nazývat prestiž. Metody, které jsou v následujících částech 2.5.1 až 2.5.12 zmíněny, jsou variací nebo obdobou algoritmu PageRank, který upravují tak, aby při vyhodnocení bibliografického grafu vyzdvihl určitou vlastnost a poskytl tak specifické hodnocení prestiže vrcholů. Použitím metod pro výpočet prestiže by mělo být získáno lepší ohodnocení vrcholů, než použitím metod měřících popularitu, které jsou zmíněny v části 2.3.

2.5.1 Vážený PageRank a AuthorRank

Od vzniku PageRanku bylo navrženo několik jeho variant, které autoři označili jako vážený PageRank (*Weighted PageRank* - WPR). Mezi jednodušší varianty patří WPR prezentovaný Ding (2011a), který používá pouze personalizaci a nepoužívá váhy hran. Autorka zmiňuje, že personalizace vrcholu, který zastupuje autora, může obsahovat počet citací, počet publikací, počet publikací, kde byl autor uveden jako první autor nebo h-index. Autorka následně při vyhodnocení citačního grafu autorů používá v personalizaci počet citací a počet publikací, přičemž personalizace počtem citací jí poskytuje na základě porovnání s oceněnými autory nejlepší pořadí autorů. Variantu PageRanku, která využívá

váhy hran a nevyužívá personalizaci, použitou na graf spoluautorství s vahami hran vyjadřujícími frekvence spolupráce jednotlivých dvojic autorů, nazvali Liu et al. (2005) *AuthorRank*. Dle Liu et al. (2005) tato metoda hodnotí míru ochoty jednotlivých autorů ke spolupráci efektivněji než betweenness centralita, která je pro objemné grafy výpočetně náročná. Měli bychom zmínit, že klasickou variantou WPR je vzorec (2.12), který uvažuje váhy hran i personalizaci.

Xing a Ghorbani (2004) v souvislosti s vyhodnocením Webu říkají, že významný vrchol zastupující webovou stránku je hodně provázaný s ostatními vrcholy, protože ostatní vrcholy chtějí na tento vrchol odkazovat a chtějí jím být odkazovány. Proto ve svém vzorci (2.15) používají vstupní i výstupní hrany vrcholu. Ve vzorci (2.15) je $WPR_x(u)$ vážený PageRank vrcholu u v iteraci x, d je faktor tlumení, B_u je množina vrcholů, které odkazují na vrchol u, R_v je množina vrcholů, na které odkazuje vrchol v, I_u je počet vstupních hran vrcholu u a O_u je počet výstupních hran vrcholu u. Xing a Ghorbani (2004) říkají, že jejich WPR umožňuje identifikovat větší množství k dotazu relevantních webových stránek než klasický PageRank.

$$WPR_{x+1}(u) = (1-d) + d \cdot \sum_{v \in B_u} \left(WPR_x(v) \cdot \frac{I_u}{\sum_{p \in R_v} I_p} \cdot \frac{O_u}{\sum_{p \in R_v} O_p} \right) \tag{2.15}$$

2.5.2 Bibliografický PageRank a Time-aware PageRank

Pojmenování bibliografický PageRank (*bibliographic PageRank*) použili Fiala et al. (2008), když do svých metod pro hodnocení autorů PageRankem zakomponovali předpoklad, že citace od spoluautora je méně významná, než citace od jiného autora. K tomu účelu upravili váhy hran v citačním grafu autorů tak, aby váhy zohlednily frekvenci spolupráce citovaného autora s citujícím autorem, viz vzorec (2.16), kde $w_{v,k}$ je počet citování autora k autorem v, $c_{v,k}$ je počet společných publikací autorů v a k a $b_{v,k}$ je jednou z následujících sedmi variant: **(a)** nula; **(b)** počet všech publikací obou autorů; **(c)** počet všech spoluautorů autora v a autora k, přičemž každý spoluautor se počítá tolikrát, kolikrát byl danému autorovi spoluautorem; **(d)** počet různých spoluautorů autora v a autora k, tj. každý spoluautor je pro každého autora počítán pouze jednou; **(e)** počet publikací autora v a autora k, přičemž se počítají pouze ty publikace, které obsahují alespoň jednoho spoluautora; **(f)** počet všech spoluautorů ve společných publikacích autorů v a k, přičemž každý spoluautor je počítán tolikrát, kolikrát byl autorům v a k spoluautorem; **(g)** počet různých spoluautorů ve společných publikacích autorů v a k.

Váhy hran Fiala et al. (2008) vložili do vzorce váženého PageRanku bez personalizace, viz vzorec (2.17), a kvalitu navržených metod experimentálně ověřili na kolekci DBLP. Porovnání získaných pořadí autorů na základě seznamu držitelů Coddovy ceny ukázalo, že pro hodnocení autorů je nejlepší použít varianty vah hran (d) a (e). Ve vzorci (2.17) je $PR_x(u)$ hodnota PageRanku vrcholu u v iteraci x, d je faktor tlumení, V je množina všech vrcholů v citačním grafu autorů a E je množina všech hran, (v,u) je hrana vedoucí z vrcholu v do vrcholu u a $\sigma_{v,u}$ je konstantní váha přiřazená hraně (v,u) dle některé varianty vzorce (2.16).

$$\sigma_{v,k} = \frac{w_{v,k}}{\frac{c_{v,k}+1}{b_{v,k}+1} \cdot \sum_{(v,j) \in E} w_{v,j}} \tag{2.16}$$

$$R_{x+1}(u) = \frac{(1-d)}{|V|} + d \cdot \sum_{(v,u)\in E} \frac{PR_x(v) \cdot \sigma_{v,u}}{\sum_{(v,k)\in E} \sigma_{v,k}} \tag{2.17}$$

V následující práci Fiala (2012b) doplnil do výše zmíněných variant vah hran (a - f) čas publikování a vytvořil bibliografický PageRank podporující čas (*Time-aware PageRank*), viz vzorce (2.18) a (2.19), kde $\sigma_{v,u}$, $c_{v,k}$ i $b_{v,k}$ získaly horní index t značící rok, do kterého se daná veličina počítá, např. $c^t_{v,k}$ představuje počet společných publikací autorů *v* a *k* vydaných před rokem *t*. V experimentu s kolekcí WoS autor porovnal pořadí autorů vytvořená Time-aware PageRankem a bibliografickým PageRankem na základě držitelů Turingovy ceny a držitelů Coddovy ceny, přičemž nejlepší pořadí autorů poskytl Time-aware PageRank s vahami hran (d) nebo (c). Varianta (d) tedy byla nejlepší v obou těchto typech PageRanku.

$$PR_{x+1}(u) = \frac{(1-d)}{|V|} + d \cdot \sum_{(v,u)\in E} \frac{PR_x(v) \cdot \sum_{i=1}^{w_{v,u}} \sigma^i_{v,u}}{\sum_{(v,k)\in E} \sum_{i=1}^{w_{v,k}} \sigma^i_{v,k}} \tag{2.18}$$

$$\sigma^i_{v,k} = \frac{1}{\frac{c^i_{v,k}+1}{b^i_{v.k}+1} \cdot \sum_{(v,j)\in E} 1} \tag{2.19}$$

2.5.3 HITS

Algoritmus HITS (*Hypertext Induced Topic Search*), který navrhl Kleinberg (1999), je podobný PageRanku, ovšem se dvěma zásadními rozdíly. Jak uvádějí Langville a Meyer (2006), prvním rozdílem je, že PageRank použitý ve vyhledávači webových stránek obvykle není závislý na dotazu (je tzv. *query-independent*), kdežto HITS bývá na dotazu závislý (tzv. *query-dependent*), tj. PageRank se počítá pro celý webový graf (tzv. off-line) a vypočítané hodnoty webových stránek jsou přiřazeny k výsledkům webového vyhledávání, kdežto HITS se počítá pro graf vytvořený z výsledků vyhledávání. Využitím PageRankových nebo HITS hodnot webových stránek se výsledky vyhledávání seřadí. Implementaci algoritmu HITS používá např. vyhledávač Teoma.com[30]. Druhým rozdílem algoritmů HITS a PageRank je, že HITS pomyslně rozděluje vrcholy na autority (*authorities*) a rozcestníky (*hubs*). Kleinberg (1999) říká, že: „*dobré rozcestníky jsou ty vrcholy, které odkazují na dobré autority, a dobré autority jsou ty vrcholy, které jsou odkazovány dobrými rozcestníky*“. V duchu této kruhové definice je HITS i počítán a hodnotí u každého vrcholu dvě vlastnosti – autoritativnost (*authoritativeness*) a „rozcestníkovost“ (*hubness*). Každému vrcholu grafu jsou tedy přiřazeny dvě hodnoty, jak ukazují vzorce (2.20), ve kterých $\boldsymbol{x}^{(k)}$ je vektor hodnot autoritativnosti jednotlivých vrcholů v iteraci *k*, $\boldsymbol{y}^{(k)}$ je vektor hodnot rozcestníkovosti vrcholů v iteraci *k* a $\boldsymbol{L}$ je matice sousednosti.

[30] Vyhledávač Teoma používá HITS pro řazení výsledků vyhledávání - http://www.teoma.com

$$\boldsymbol{x}^{(k)} = \boldsymbol{L}^T \boldsymbol{y}^{(k-1)}$$
$$\boldsymbol{y}^{(k)} = \boldsymbol{L}\boldsymbol{x}^{(k)} \tag{2.20}$$

Vzorce (2.20) lze zbavit kruhové závislosti, jak je ukázáno ve vzorcích (2.21).

$$\boldsymbol{x}^{(k)} = \boldsymbol{L}^T \boldsymbol{L}\boldsymbol{x}^{(k-1)}$$
$$\boldsymbol{y}^{(k)} = \boldsymbol{L}\boldsymbol{L}^T \boldsymbol{y}^{(k-1)} \tag{2.21}$$

Pro hodnocení bibliografických entit se obvykle používají HITS hodnoty autoritativnosti vrcholů. Přestože byl HITS při citační analýze mnohokrát použit, viz např. (Borodin et al. 2005; Sidiropoulos a Manolopoulos 2005a, 2006; Fiala et al. 2008, 2015; Fiala 2011, 2012b), tak ale obvykle sloužil pouze pro porovnání s nově vytvořenými metodami a neposkytl nejlepší hodnocení. Úpravy algoritmu HITS a zvláště pak jeho kombinace s algoritmem PageRank, které napravují některé nedostatky obou algoritmů, zmiňují např. části 2.5.4, 2.5.5 a 2.5.10. Více o konvergenci, citlivosti, urychlení výpočtu a silných a slabých stránkách algoritmu HITS lze nalézt v (Langville a Meyer 2006).

2.5.4 FutureRank

Algoritmus *FutureRank* (Sayyadi a Getoor 2009) je kombinací algoritmů PageRank a HITS pro účely souběžného hodnocení autorů a publikací na základě citačního grafu publikací a bipartitního grafu autorství (tj. autoři a jejich publikace), který by měl umožnit výpočet budoucích hodnot PageRanku publikací. Při iteračním výpočtu se střídají dva kroky:

1) vypočtení hodnot publikací na základě hodnot jejich autorů, hodnot citujících publikací a stáří publikace (kombinace PageRanku a HITSu), viz vzorec (2.22).
2) rozdělení hodnot publikací jejich autorům (obdoba HITS), viz vzorec (2.24).

Hodnoty publikací jsou počítány dle vzorce (2.22), který obsahuje tři faktory tlumení α, β a γ (musí platit $\alpha+\beta+\gamma \leq 1$) pro určení míry, s jakou se v hodnocení publikací projeví jednotlivé hodnocené vlastnosti:

- faktor α tlumí vliv předávání hodnot mezi publikacemi (dynamická část PageRanku) – použit je citační graf publikací reprezentovaný maticí $\boldsymbol{M}^C$ a vektor hodnot publikací $\boldsymbol{R}^P$,
- faktor β tlumí vliv hodnot předaných publikacím od autorů (obdoba HITS) – použit je bipartitní graf autoři-publikace reprezentovaný maticí $\boldsymbol{M}^A$ a vektor hodnot autorů $\boldsymbol{R}^A$,
- faktor γ tlumí vliv stáří publikace (personalizace PageRanku) – $\boldsymbol{R}^{Time}$ je vektor, jehož hodnoty se snižují dle stáří publikace.

Stáří R_i^{Time} publikace *i* je počítáno vzorcem (2.23), kde $T_{current}$ je současný rok nebo rok dotazu (pokud je vzorec použit ve vyhledávači) a T_i je rok uveřejnění publikace *i*. Hodnota ρ byla experimentálně stanovena na 0,62. Počet hodnocených publikací je značen *n*. Hodnoty autorů $\boldsymbol{R}^A$ jsou počítány dle vzorce (2.24), který je obdobou vzorce HITS pro výpočet rozcestníkových hodnot vrcholů.

$$\boldsymbol{R}^P = \alpha \cdot \boldsymbol{M}^C \cdot \boldsymbol{R}^P + \beta \cdot \boldsymbol{M}^{A^T} \cdot \boldsymbol{R}^A + \gamma \cdot \boldsymbol{R}^{Time} + (1 - \alpha - \beta - \gamma)/n \tag{2.22}$$

$$R_i^{Time} = e^{-\rho \cdot (T_{current} - T_i)} \tag{2.23}$$

$$\boldsymbol{R}^A = \boldsymbol{M}^A \cdot \boldsymbol{R}^P \tag{2.24}$$

Pro výpočet hodnot publikací autoři použili několik variant nastavení faktorů tlumení ve vzorci (2.22), přičemž, jak autoři uvádějí, největší přesnosti výsledků dosáhli s α=0,19, β=0,02 a γ=0,79. Autoři dále uvádějí, že FutureRank s β=0 je podobný algoritmu *CiteRank* (Walker et al. 2007) a FutureRank s γ=0 je podobný algoritmu *CoRank* (Zhou et al. 2007), který také současně hodnotí publikace a autory. Algoritmy CiteRank a CoRank jsou zmíněny v části 2.5.12.

2.5.5 SALSA

Algoritmus SALSA (*the Stochastic Approach for Link-Structure Analysis*), publikovaný v (Lempel a Moran 2000, 2001), je kombinací algoritmů HITS a PageRank pro účely webového vyhledávání. SALSA, obdobně jako HITS, počítá hodnoty rozcestníkovosti a autoritativnosti vrcholů, ale navíc z PageRanku přejímá koncept Markovova řetězce. Tato kombinace odstraňuje z algoritmu HITS těsnou provázanost autorit a rozcestníků a poskytuje tak lepší hodnocení webových stránek než HITS. Jak uvádějí Langville a Meyer (2006), SALSA vytváří dva Markovovy řetězce a výpočet lze zjednodušit využitím matice sousednosti $\boldsymbol{L}$, ze které se normalizací nenulových řádků vytvoří matice $\boldsymbol{L}_r$ a normalizací nenulových sloupců matice $\boldsymbol{L}_c$. Následně použitá matice provázanosti rozcestníků $\boldsymbol{H}$ obsahuje všechny nenulové řádky z matice $\boldsymbol{H'}$ a matice provázanosti autorit $\boldsymbol{A}$ všechny nenulové sloupce z matice $\boldsymbol{A'}$. Matice $\boldsymbol{H'}$ a $\boldsymbol{A'}$ byly vytvořeny využitím vzorců (2.25).

$$\begin{aligned} \boldsymbol{H}' &= \boldsymbol{L}_r \boldsymbol{L}_c^T \\ \mathbf{A}' &= \boldsymbol{L}_c^T \boldsymbol{L}_r \end{aligned} \tag{2.25}$$

Hodnoty rozcestníkovosti i autoritativnosti vrcholů jsou počítány zvlášť pro každou souvislou komponentu $\boldsymbol{C}$ matic $\boldsymbol{H}$ a $\boldsymbol{A}$ dle vzorce (2.26), kde $\boldsymbol{\pi^{(k)}(C)}$ je vektor hodnot vrcholů z komponenty $\boldsymbol{C}$ v iteraci k. Tento vzorec je obdobou vzorce (2.6) pro výpočet PageRanku, který je zde ale počítán pro jednotlivé souvislé komponenty. Globální hodnoty rozcestníkovosti a autoritativnosti vrcholů jsou dány poměrným sloučením příslušných komponent, tj. pokud má matice např. 5 prvků a je tvořena dvěma souvislými komponentami s 2 a 3 prvky, tak výsledný vektor s hodnotami vrcholů obsahuje prvky první komponenty s hodnotami vynásobenými 3/5 a prvky druhé komponenty s hodnotami vynásobenými 2/5.

$$\boldsymbol{\pi}^{(k+1)T}(\boldsymbol{C}) = \boldsymbol{\pi}^{(k)T} \boldsymbol{C} \tag{2.26}$$

2.5.6 *Eigenfactor Metrics používané databází ISI Web of Science*

Eigenfactor™ Metrics (Bergstrom 2007; West et al. 2008) obsahují dvě metody implementované společností *Eigenfactor.org*[31] pro analýzu libovolné úrovně citačního grafu (tj. časopisy, instituce, autoři, články atd.). Tyto metody jsou významné zejména proto, že je databáze ISI Web of Science používá pro hodnocení časopisů a vypočtené hodnoty zobrazuje v JCR. Metoda Eigenfactor Score je kombinací 5-Year Impact Factoru a PageRanku. Její hodnota časopisu udává procento celkového počtu vážených citací, které v aktuálním roce obdržela vydání daného časopisu z předchozích pěti let. Aby byly hodnoty Eigenfactor Score časopisů porovnatelné s hodnotami 5-Year Impact Factoru časopisů, tak je potřeba vypočíst průměrnou hodnotu významnosti článků v daném časopise, k čemuž slouží metoda Article Influence Score (hodnota vlivnosti časopisu).

Při výpočtu Eigenfactor Score se používá matice sousednosti (či citovanosti) časopisů $\boldsymbol{Z}$, přičemž prvek Z_{ij} je počtem citací z časopisu *j* ve zvoleném roce (např. 2012) na články časopisu *i* publikované 5 let před tím (tj. v letech 2007 až 2011). Samocitace časopisů jsou při výpočtu ignorovány. Z matice $\boldsymbol{Z}$ vytvoříme sloupcovou normalizací matici $\boldsymbol{H}$. Dále vytvoříme vektor počtu článků $\boldsymbol{a}$, kde a_i je počet článků publikovaných v časopise *i* v daném pětiletém okénku (tj. v letech 2007 až 2011), a celý vektor znormalizujeme. Následně ošetříme problém slepých vrcholů (tj. časopisů, které nikoho necitují), což jsou nulové sloupce matice $\boldsymbol{H}$, tak, že tyto sloupce nahradíme vektorem $\boldsymbol{a}$ a získáme tak matici $\boldsymbol{H'}$. Nyní obdobně, jako při vytváření matice pro PageRank s personalizací, viz vzorec (2.14), vytvoříme matici $\boldsymbol{G}$ dle vzorce (2.27), kde *d* je faktor tlumení (obvykle *d*=0,85) a $\boldsymbol{e}$ je jednotkový vektor. Poté PageRankem vypočítáme vektor hodnot významnosti časopisů značený $\boldsymbol{\pi}$, viz vzorec (2.28), kde $\boldsymbol{\pi}^k$ je vektor významnosti časopisů v iteraci *k* a $\boldsymbol{G}$ je upravená matice sousednosti časopisů.

$$\boldsymbol{G} = d\boldsymbol{H}' + (1-d)\boldsymbol{a}\boldsymbol{e}^T \tag{2.27}$$

$$\boldsymbol{\pi}^{(k+1)T} = \boldsymbol{\pi}^{(k)T}\boldsymbol{G} \tag{2.28}$$

Eigenfactor™ Score je definováno dle vzorce (2.29), kde $\boldsymbol{E}$ je vektor Eigenfactor Score hodnot časopisů, *J* je množina všech zkoumaných časopisů a H_j řádek matice $\boldsymbol{H}$ odpovídající časopisu *j* (pozn.: matice $\boldsymbol{H'}$ zde použita není).

$$\boldsymbol{E} = 100\frac{\boldsymbol{H\pi}}{\sum_{j\in J}[H\pi]_j} \tag{2.29}$$

Metoda *Article Influence™ Score* rozděluje hodnotu Eigenfactor Score časopisu všem článkům daného časopisu, a proto lze její výsledky porovnat s výsledky 5-Year Impact Factoru. Zapsat ji lze vzorcem (2.30), kde A_i je hodnota Article Influence Score časopisu *i*.

$$A_i = 0{,}01\frac{E_i}{a_i} \tag{2.30}$$

[31] Web Eigenfactor.org - http://www.eigenfactor.org

E-factor či *Energyfactor* (Prathap 2010) je součinem Eigenfactor Score a Article Influence Score navrženým za účelem získání lepší míry prestiže časopisů. Yin et al. (2009) využitím Eigenfactor Score a h-indexu časopisů graficky klasifikovali časopisy do čtyř kvadrantů a ukázali, že pouze časopisy *Nature* a *Science* mají vysoké hodnoty Eigenfactor Score i h-indexu. Pozn.: h-index časopisu je počítán obdobně jako h-index autorů, viz (Braun et al. 2006).

2.5.7 Y-factor

Y-factor (Bollen et al. 2006) byl také navržen pro hodnocení časopisů. Autoři definují populární časopisy jako: *„časopisy často citované časopisy s malou prestiží"* a říkají, že tyto časopisy mají vysoký Impact Factor a nízký vážený PageRank. Prestižní časopisy definují jako: *„časopisy, které nejsou často citované, ale jejichž citace pocházejí z prestižních časopisů"* a říkají, že tyto časopisy mají nízký Impact Factor a vysoký PageRank. Představený Y-factor, viz vzorec (2.31), je součinem Impact Factoru zatupujícího popularitu časopisu a PageRanku, který zastupuje prestiž časopisu. Impact Factor, PageRank a Y-factor Bollen et al. (2006) testují na datech z kolekce WoS (JCR 2003) a ukazují, že Y-factor jako jediný poskytuje pořadí časopisů s časopisy *Nature* a *Science* na prvních dvou pozicích. Ve vzorci (2.31) je *Y(j)* hodnota Y-factoru časopisu *j*, *IF(j)* je hodnota Impact Factoru časopisu *j* a *WPR(j)* je hodnota váženého PageRanku časopisu *j* vypočítaná z citačního grafu časopisů vzorcem PageRanku (2.11).

$$Y(j) = IF(j) \cdot WPR(j) \tag{2.31}$$

2.5.8 Metody pro hodnocení zdrojů používané databází Scopus

SCImago Journal Rank, Impact per Publication a Source Normalized Impact per Paper jsou významné metody proto, že je používá databáze Scopus[32] pro hodnocení zdrojů (časopisy, sborníky apod., dále jen časopisy). Všechny metody se vyznačují tím, že používají pouze časopisecké a konferenční články a recenze (dále jen publikace). *SCImago Journal Rank*[33] (SJR), viz (González-Pereira et al. 2010), je obdobou PageRanku pro hodnocení prestiže časopisů. Jeho výpočet je rozdělen do dvou kroků:

1) výpočet tzv. prestižního SJR (*Prestige SJR* – PSJR) – míra vyjadřující celkovou prestiž časopisu v závislosti na počtu publikací, které časopis obsahuje.
2) výpočet SJR rozdělením PSJR časopisu jeho publikacím – míra prestiže časopisu, která není závislá na počtu publikací časopisu a může být použita pro porovnání různě objemných časopisů.

Krok s výpočtem SJR znázorňuje vzorec (2.32), kde $PSJR_i^{(k)}$ je hodnota PSJR časopisu *i* v iteraci *k*, *d*=0,9 a *e*=0,0999 jsou faktory tlumení, *V* je množina všech časopisů, A_i je počet publikací obsažených v časopise *i*, C_{ji} je počet referencí na časopis *i* uvedených v časopise *j*, C_j je celkový počet referencí uvedených v časopise *j*, *CF* je „faktor korekce" počítaný vzorcem (2.33) a *D* je množina všech slepých vrcholů (tj. časopisů, které neobsahují reference). Při výpočtu jsou používány publikační záznamy ze tří let a počáteční nastavení $\forall PSJR_i^0 = 1/|V|$, kde $|V|$ je velikost množiny všech časopisů. Vzorec (2.32), který je podobný vzorci (2.13) PageRanku s personalizací, lze pomyslně rozdělit na tři části,

[32] Popis metod používaných bibliografickou databází Scopus - http://www.journalmetrics.com
[33] Popis výpočtu SJR uvedený na webu Scopus - http://www.journalmetrics.com/sjr.php

přičemž hodnoty prvních dvou částí se pro daný časopis v průběhu výpočtu nemění, a proto je lze před-vypočítat.

Význam jednotlivých částí vzorce (2.32) lze popsat následovně:

- první zlomek vyjadřuje minimální hodnotu, kterou časopis získá za to, že se nachází ve zvolené datové kolekci (jedná se o ošetření problému Rank sink, viz část 2.4.1).
- druhý zlomek je podíl publikační produktivity určený na základě počtu publikací, které časopis obsahuje, a počtu všech publikací obsažených v kolekci (jedná se o personalizaci časopisu dle počtu publikací).
- zbytek vzorce obsahuje dvě složky, z nichž první zastupuje citační prestiž časopisu určenou na základě významnosti citujících časopisů, a druhá přiřazuje časopisu část z prestiže slepých vrcholů/časopisů, určenou na základě publikační produktivity.

$$PSJR_i^{(k+1)} = \frac{(1-d-e)}{|V|} + e \cdot \frac{A_i}{\sum_{j \in V} A_j} + d \cdot \left(\sum_{j \in V} \frac{PSJR_j^{(k)} \cdot C_{ji}}{C_j} \cdot CF^{(k)} + \frac{A_i}{\sum_{j \in V} A_j} \cdot \sum_{d \in D} PSJR_d^{(k)} \right) \quad (2.32)$$

$$CF^{(k)} = \frac{1 - \sum_{d \in D} PSJR_d^{(k)}}{\sum_{u \in V} \sum_{v \in V} C_{vu} \cdot \frac{PSJR_v^{(k)}}{C_v}} \quad (2.33)$$

Ve druhém kroku z PSJR vypočteme SJR dle vzorce (2.34), kde *c* je konstanta použitá k navýšení hodnot SJR tak, aby nebyly „příliš malé“.

$$SJR_i = c \cdot \frac{PSJR_i}{A_i} \quad (2.34)$$

Další metodou, kterou Scopus používá pro hodnocení časopisů, je *Impact per Publication*[34] (IPP, vliv publikací) či též *Raw Impact per Paper* (RIP, surový vliv článků), viz (Moed 2010), který je tříletou obdobou Impact Factoru, tj. je mírou popularity. IPP daného časopisu je počet citací z daného roku na publikace vydané během předchozích tří let v daném časopise, dělený počtem těchto publikací. Nevýhodou IPP (stejně jako Impact Factoru) je, že neodráží odlišné praktiky citování v různých oblastech výzkumu, a proto s ním nelze porovnávat vědecké časopisy z různých oblastí (např. počítačové a společenské vědy). Pokud je IPP časopisu normalizován relativním citačním potenciálem vědní oblasti daného časopisu v použité kolekci (*Relative Database Citation Potential* – RDCP), tak vzniká *Source Normalized Impact per Paper*[35] (SNIP, zdrojem normalizovaný vliv článků), kterým již lze porovnávat časopisy z různých oblastí výzkumu, viz (Moed 2010; Waltman et al. 2013).

[34] Popis výpočtu IPP uvedený na webu Scopus - http://www.journalmetrics.com/ipp.php
[35] Popis výpočtu SNIP uvedený na webu Scopus - http://www.journalmetrics.com/snip.php

Výpočet SNIP, RDCP a citačního potenciálu databáze (*Database Citation Potential* - DCP) znázorňují vzorce (2.35), kde Ω je vědecká oblast, do které patří časopis *i*, DCP_Ω a $RDCP_\Omega$ jsou DCP a RDCP hodnoty vědecké oblasti Ω, *P* je množina všech publikací ze zvoleného roku (např. 2012) publikovaných ve vědní oblasti Ω a obsažených v kolekci dat, p_r je počet referencí v publikaci *p* na všechny publikace vydané v předchozích třech letech (tj. v letech 2009 až 2011), n_Ω je počet publikací vydaných v oblasti Ω v přechozích třech letech (tj. v letech 2009 až 2011), *DCP* je množina hodnot DCP_Ω všech oblastí a *median(DCP)* je hodnota mediánu v množině *DCP*, IPP_i je hodnota IPP časopisu *i* a $SNIP_i$ je hodnota SNIP časopisu *i*.

$$DCP_\Omega = \frac{\sum_{p \in P} p_r}{n_\Omega}$$

$$RDCP_\Omega = \frac{DCP_\Omega}{median(DCP)} \tag{2.35}$$

$$SNIP_i = \frac{IPP_i}{RDCP_\Omega} \ , \qquad i \in \Omega$$

Hodnocení časopisů a států, vypočítaná na základě dat databáze Scopus, lze nalézt na webu *SCImago Journal & Country Rank*[36]. Na webu *CWTS Journal Indicators*[37] lze nalézt následující porovnání Journal Impact Factoru (JIF), IPP a SNIP:

- SNIP a IPP jsou počítané z dat Scopus, kdežto JIF je počítán z dat Web of Science,
- SNIP a IPP používají tříleté okénko citovaných publikací, kdežto JIF používá okénko dvouleté,
- SNIP a IPP používají pouze citace z vybraných typů dokumentů, kdežto JIF používá citace ze všech indexovaných dokumentů,
- SNIP koriguje vliv odlišných trendů citování v různých vědních oblastech, kdežto IPP a JIF tento vliv zanedbávají.

2.5.9 SCEAS

Algoritmus SCEAS, součást stejnojmenného systému pro hodnocení vědeckých kolekcí (*Scientific Collection Evaluator by using Advanced Scoring*)[38], svou koncepcí vychází z algoritmu PageRank. První zmínky o SCEAS lze nalézt v (Sidiropoulos a Manolopoulos 2005a), kde jsou publikace a autoři hodnoceni na základě vyhodnocení citačních sítí publikací, které byly vytvořeny z dat databáze DBLP. Hodnota autora je stanovena jako průměrem z hodnot jeho 25 nejlepších publikací. Kvalitu metod Sidiropoulos a Manolopoulos (2005a) určili porovnáním získaných pořadí publikací na základě ocenění *VLDB 10 Year Award* a *SIGMOD Test of Time Award* a porovnáním získaných pořadí autorů na základě ocenění *SIGMOD E. F. Codd Innovations Award*, přičemž ve všech případech jim SCEAS poskytl lepší výsledky než PageRank a HITS.

[36] Web SCImago Journal & Country Rank - http://www.scimagojr.com
[37] Web metodologie CWTS Journal Indicators - http://www.journalindicators.com/methodology
[38] Web systému SCEAS, který obsahuje statistiky z hodnocení bibliografických entit - http://sceas.csd.auth.gr

Ve SCEAS, viz vzorec (2.36), je posílen vliv přímých citací (tj. vstupních hran vrcholu) konstantou *b* a tlumen vliv nepřímých citací (tj. hran, které leží na cestách směřujících do daného vrcholu, ale nejsou vstupními hranami daného vrcholu) mocninami konstanty *a*. Vlivem tlumení nepřímých citací změna hodnoty vrcholu *i* ovlivní hodnotu vrcholu *j*, který je *x*-tým vrcholem v řadě (tj. mezi vrcholy *i* a *j* je *x*-1 vrcholů), s faktorem a^{-x}. Autoři jako konstantu *a* používali Eulerovo číslo *e*. Výhodou SCEAS oproti PageRanku a HITSu je, že výpočet hodnot vrcholů je více ovlivněn přímým citováním a je méně citlivý na přidání nového vrcholu do grafu. Také konvergence algoritmu SCEAS je velmi rychlá. Ve vzorci (2.36) je $S_j^{(k)}$ SCEAS hodnota vrcholu *j* v iteraci *k*, *d* je faktor tlumení, U_j je množina všech vrcholů, ze kterých vede hrana na vrchol *j*, N_u je počet výstupních hran vrcholu *u* (váhy hran se zde nepoužívají), *b* je faktor prosazení přímého citování a *a* faktor rychlosti, se kterou prosazení nepřímého citování konverguje k nule. Protože vzorec (2.36) algoritmu SCEAS neošetřuje slepé vrcholy, tak po každé iteraci algoritmu je potřeba provést normalizaci vypočtených hodnot. Autoři při experimentech používali dvě varianty SCEAS a to SCEAS1 s *d*=1 a *b*=1 a SCEAS2 s *d*=0,85 a *b*=0, přičemž zanedbatelně lepší pořadí poskytoval SCEAS1.

$$S_j^{(k+1)} = (1-d) + d \cdot \sum_{u \in U_j} \frac{S_u^{(k)} + b}{N_u} a^{-1} \tag{2.36}$$

2.5.10 B-HITS, B-SALSA a varianty SCEAS

Sidiropoulos a Manolopoulos (2006) ve své další práci analyzují kvalitu algoritmů PageRank, HITS, SALSA a SCEAS při hodnocení publikací a autorů a navrhují jejich úpravy, které by hodnocení měly zlepšit. Hodnoty publikací jsou opět vypočteny z citační sítě a hodnoty autorů jsou stanoveny jako průměr z hodnot jejich nejlepších 25 a 30 publikací. Kvalita získaných pořadí byla stanovena na základě pozic oceněných publikací (*VLDB 10 Year Award*, *SIGMOD Test of Time Award*) nebo autorů (*SIGMOD E.F. Codd Innovations Award*). V práci jsou nejprve definovány počet citací (*Citation Count* – CC), tj. počet vstupních hran vrcholu, a vyvážený počet citací (*Balanced Citation Count* – BCC), což je součet částí, které vrchol získá od vrcholů, které ho citují. Výpočet BCC ukazuje vzorec (2.37), kde BCC_x je vyvážený počet citací vrcholu *x*, U_x je množina všech vrcholů, ze kterých vede hrana na vrchol *x*, a N_u je počet výstupních hran vrcholu *u*. Míry CC i BCC autoři kritizují, protože se při jejich výpočtu nevyužívá významnost citujících vrcholů, tj. jedná se o míry popularity.

$$BCC_x = \sum_{u \in U_x} \frac{1}{N_u} \tag{2.37}$$

Autoři následně zavádí míru *Prestiž*, kterou definují jako součet *Prestiží* citujících vrcholů, jak ukazuje vzorec (2.38), kde $P_x^{(k)}$ je *Prestiž* vrcholu *x* v iteraci *k*. Tuto míru kritizují, protože hodnoty vrcholů neúčastnících se žádného cyklu konvergují k nule a pokud existuje v grafu cesta, na které vrchol *x* cituje vrchol *y*, tak hodnota vrcholu *x* nebude nikdy větší než hodnota vrcholu *y*. (Pozn.: mimo tyto problémy je se vzorcem (2.38) spojeno několik dalších problémů, které byly řešeny při návrhu algoritmu PageRank, např. problém Rank sink, viz část 2.4.1).

$$P_x^{(k+1)} = \sum_{u \in U_x} P_u^{(k)} \tag{2.38}$$

Autoři také kritizují PageRank, protože vrcholy účastnící se cyklů získají největší hodnoty PageRanku, a algoritmy SALSA a HITS, protože uvažují rozcestníky a autority, což příliš neodpovídá hodnocení publikací ani autorů. Z těchto důvodů navrhují nové algoritmy B-HITS, B-SALSA a různé alternativy algoritmu SCEAS.

Balanced (vyvážený) *HITS* či *B-HITS* při výpočtu hodnot autoritativnosti vrcholů vedle hran „vedoucích z rozcestníků“ uvažuje i hrany „vedoucí z autorit“, čímž kombinuje PageRank a HITS, jak ukazují vzorce (2.39), kde $BHA_x^{(k)}$ je B-HITS hodnota autoritativnosti vrcholu *x* v iteraci *k*, $BHH_x^{(k)}$ je B-HITS hodnota rozcestníkovosti vrcholu *x* v iteraci *k*, U_x je množina vrcholů, ze kterých vede hrana na vrchol *x*, W_x je množina vrcholů, na které vedou hrany z vrcholu *x*, a *p* je faktor tlumení či míra, se kterou se hodnoty autoritativnosti vrcholů navzájem ovlivňují ($0 < p < 1$).

$$BHA_x^{(k)} = (1-p) \cdot \sum_{u \in U_x} BHH_u^{(k-1)} + p \cdot \sum_{u \in U_x} BHA_u^{(k-1)}$$
$$BHH_x^{(k)} = \sum_{w \in W_x} BHA_w^{(k)} \tag{2.39}$$

Balanced (vyvážená) *SALSA* či *B-SALSA* obsahuje podobnou úpravu jako B-HITS. Znázorněna je vzorci (2.40), kde $BSA_x^{(k)}$ je B-SALSA hodnota autoritativnosti vrcholu *x* v iteraci *k*, $BSH_x^{(k)}$ je B-SALSA hodnota rozcestníkovosti vrcholu *x* v iteraci *k*, M_w je počet vstupních hran vrcholu *w* a ostatní parametry jsou stejné jako v B-HITS.

$$BSA_x^{(k)} = (1-p) \cdot \sum_{u \in U_x} \frac{BSH_u^{(k-1)}}{N_u} + p \cdot \sum_{u \in U_x} \frac{BSH_u^{(k-1)}}{N_u}$$
$$BSH_x^{(k)} = \sum_{w \in W_x} \frac{BSA_w^{(k)}}{M_w} \tag{2.40}$$

Metody pro hodnocení publikací a autorů Sidiropoulos a Manolopoulos (2006) testovali na kolekci DBLP a kvalitu metod určovali na základě pozic oceněných publikací nebo autorů ve vytvořených pořadích. Průměrně nejlepší pořadí publikací i autorů poskytly PageRank a SCEAS-BPS (což je *SCEAS Balanced Publication Score*, tj. SCEAS vyvážené publikační hodnocení) a nejhorší pořadí poskytly *Prestiž* (více jak 3x horší) a BHA část B-HITS (téměř 3x horší). Autoři dále zmiňují, že SCEAS1 a SCEAS2, popsané v části 2.5.9, konvergují nejrychleji ze všech použitých metod. Výsledky hodnocení publikací zde popsanými metodami lze nalézt na webu systému SCEAS[39].

[39] Statistiky z hodnocení publikací metodami z částí 2.5.9 a 2.5.10 na webu systému SCEAS - http://sceas.csd.auth.gr/php/stats.php4

Pro úplnost uvádíme vzorec (2.41) algoritmu SCEAS-BPS, ve kterém $BPS_x^{(k)}$ je SCEAS-BPS hodnota vrcholu *x* v iteraci *k*, *b* je faktor prosazení přímého citovaní (viz část 2.5.9) a zbylé parametry byly již popsány výše.

$$BPS_x^{(k+1)} = \sum_{u \in U_x} \frac{BPS_x^{(k)} + b}{N_u} \tag{2.41}$$

2.5.11 Hodnocení konferencí

Sidiropoulos a Manolopoulos (2005a) navrhli také několik nových metod primárně určených pro hodnocení konferencí. Některé z těchto metod jsou zajímavé zvláště tím, že hodnotí jednotlivé ročníky konferencí. Metody autoři aplikovali na kolekci DBLP a výsledky zobrazili na webu systému SCEAS [40]. Referenční pořadí konferencí autoři nevytvářeli, a proto ani nediskutovali kvalitu jednotlivých metod.

Plain Score (prosté hodnocení) udává průměrný počet citací článků obsažených ve sbornících dané konference. Počítáno je dle vzorce (2.42), kde S_c je Plain Score hodnota konference *c*, P_c je počet publikací obsažených ve sbornících konference *c*, *K* je množina konferencí a $N_{i \to c}$ je počet referencí na články ze sborníků konference *c*, které obsahují sborníky konference *i*.

$$S_c = \frac{1}{P_c} \sum_{i \in K} N_{i \to c} \tag{2.42}$$

Plain Score per Year (PSY, prosté roční hodnocení) omezuje výpočet Plain Score na jednotlivé ročníky konference. Ve vzorci (2.43) je $SY_{c,y}$ Plain Score per Year hodnota konference *c*, která se konala v roce *y*, $P_{c,y}$ je počet publikací ve sborníku konference *c* z roku *y* a $N_{i \to c,y}$ je počet referencí na články ze sborníku konference *c* z roku *y*, které obsahují sborníky konference *i*.

$$SY_{c,y} = \frac{1}{P_{c,y}} \sum_{i \in K} N_{i \to c,y} \tag{2.43}$$

Inverted Impact (či jen *I-Impact*) *Score per Year* (obrácené roční hodnocení vlivu) je jakýmsi „obráceným" Impact Factorem, který namísto počítání citací, které obdržely sborníky dané konference z předchozích *k* let, počítá citace, které sborník dané konference obdržel v následujících *k* letech, což dle autorů lépe vystihuje vliv konference z daného roku na vědecké smýšlení v následujících *k* letech (obvykle v následujících 2 nebo 5 letech, dle Impact Factoru a 5-Year Impact Factoru). $IISY_{c,y}$ ve vzorci (2.44) je I-Impact Score per Year hodnota konference *c* pořádané v roce *y*, *k* je počet let po konání konference *c* v roce *y* a $N_{i,z \to c,y}$ je počet referencí na články ze sborníku konference *c* z roku y, které obsahuje sborník konference *i* z roku *z*. Protože Plain Score, PSY a

[40] Statistiky z hodnocení konferencí metodami z části 2.5.11 na webu systému SCEAS - http://sceas.csd.auth.gr/php/ranking.php

I-Impact Score per Year nepoužívají významnosti citujících vrcholů, tak je lze považovat za míry popularity.

$$IISY_{c,y} = \frac{1}{P_{c,y}} \sum_{i \in K} \sum_{z=y}^{y+k} N_{i,z \to c,y} \tag{2.44}$$

Weighted Score (vážené hodnocení) je obdobou Plain Score, která používá významnosti citujících vrcholů, a proto ji lze do určité míry považovat za míru prestiže, viz vzorec (2.45), ve kterém $WS_{c,l}$ je hodnota Weighted Score konference *c* v *l*-té iteraci výpočtu a $W_{i,l}$ je hodnota konference *i* v *l*-té iteraci. Výpočet začíná s $\forall W_{i,0} = 1$ a před každou další iterací se hodnoty konferencí stanovují na základě rozdělení konferencí do shluků dle $WS_{c,l}$ (autoři testovali více variant počtu shluků a jejich ohodnocení).

$$WS_{c,l} = \frac{1}{P_c} \cdot \frac{\sum_{i \in K} W_{i,l-1} \cdot N_{i \to c}}{\sum_{i \in K} W_{i,l-1}} \tag{2.45}$$

Weighted Score per Year (WSY, vážené roční hodnocení) je kombinací Weighted Score a PSY, která pro výpočet vlivu konference v následujících letech používá významnosti citujících vrcholů, viz vzorec (2.46). Lze říci, že WSY udává, jak byla konference z daného roku považována za prestižní v následujících letech. Výpočet začíná od aktuálního/nejvyššího roku a pokračuje k nižším rokům tak, aby při výpočtu hodnot pro rok *y* byly známy hodnoty z let {*y+1*; …; *aktuální rok*}. Pro každý rok se iteruje, dokud nedojde k ustálení hodnot. Ve vzorci (2.46) je $WSY_{c,y,l}$ Weighted Score per Year hodnota konference *c* pořádané v roce *y* v *l*-té iteraci výpočtu a $W_{i,z,\infty}$ je ustálená hodnota konference *i* z roku *z*.

$$WSY_{c,y,l} = \frac{1}{P_{c,y}} \cdot \frac{\sum_{i \in K} (W_{i,y,l-1} \cdot N_{i,y \to c,y} + \sum_{z=y+1}^{akt.rok} W_{i,z,\infty} \cdot N_{i,z \to c,y})}{\sum_{i \in K} (W_{i,y,l-1} + \sum_{z=y+1}^{akt.rok} W_{i,z,\infty})} \tag{2.46}$$

2.5.12 Další PageRanku podobné algoritmy pro měření významnosti

Další algoritmy pro hodnocení bibliografických entit, které se svou podstatou podobají některému z výše zmíněných algoritmů, ukazují např. Borodin et al. (2005), kteří upravují algoritmus HITS a vytvářejí jeho nové varianty, které zmírňují některé jeho nedostatky. Algoritmus *CiteRank* (Walker et al. 2007) je variantou PageRanku pro hodnocení publikací, ve které autoři uvažují skutečnost, že nová citace (znamenající, že článek je významný v aktuální linii výzkumu) má větší váhu než citace stará, a proto exponenciálně snižují váhy citací v závislosti na jejich stáří. Yan a Ding (2010) říkají, že prestiž publikace může být tvořena významem citujících časopisů, přičemž novější citace mají větší význam než citace starší, a proto autoři určují prestiž publikací na základě významnosti citujících časopisů a časového intervalu citování. Tento postup následně rozšiřují v článku (Yan et al. 2011), kde vytvářejí algoritmus *P-Rank*, s jehož využitím hodnotí současně publikace, autory a časopisy. Obdobný postup hodnocení více bibliografických entit současně s využitím provázaných vzorců lze nalézt v (Yu et al. 2012), ovšem tento postup je aktuálně pouze teoretický, protože pro hodnocení chybí některá data, jako např. komentáře uživatelů. Algoritmus *CoRank* (Zhou et al. 2007) hodnotí současně autory a

publikace. Pro vyhledávání expertů v online znalostní komunitě Wang et al. (2013) navrhli algoritmus *ExpertRank*, který určuje významnost či relevantnost autorů na základě relevance textů jejich publikací k uživatelovu dotazu.

Vedle uplatnění PageRanku ve webovém vyhledávání a v bibliometrii byl PageRank od svého vzniku adaptován na grafy získané z různých oblastí výzkumu. Pro úlohu určování významných genů vznikl algoritmus *GeneRank* (Morrison et al. 2005; Benzi a Kuhlemann 2012) a graf interakce proteinů byl vyhodnocován algoritmem *PageRank Affinity* (Voevodski et al. 2009). Hodnocení reputace uživatelů v P2P[41] sítích s využitím PageRanku ukazují Chirita et al. (2004). Použití PageRanku pro predikci vzniku nové hrany mezi dvěma vrcholy v sociální síti můžeme nalézt v (Liben-Nowell a Kleinberg 2007), určování reputace uživatelů sociální sítě PageRankem v (Han et al. 2012; Hao et al. 2012), hlasovací systém využívající sociální síť a PageRank v (Boldi et al. 2009) a vyhledání vůdčích osob na základě vyhodnocení firemní e-mailové komunikace PageRankem v (Berchenko et al. 2011). Adaptace PageRanku pro úlohy zpracování textů, ve kterých byl použit pro extrakci klíčových slov nebo významných vět, budou více popsány v části 6.1.2.

[41] Pojmem *Peer-to-Peer* (P2P) síť bývá označována počítačová síť, jejíž klienti využívají komunikační protokol, který umožňuje přímou komunikaci dvou klientů. Protokoly tohoto typu lze označit za protokoly s komunikací klient-klient.

3 Návrh metod pro hodnocení autorů

Jak již bylo uvedeno ve 2. kapitole, hodnocení vědeckých pracovníků na základě jejich publikační činnosti může být použito při vyhledávání nebo porovnávání expertů ve zvolené oblasti vědy a výzkumu, při výběrových řízeních, udíleních grantů nebo ocenění apod. Lze podotknout, že např. o grant se obvykle žádá v určité vědní oblasti (např. počítačové vědy), pro kterou buďto existují odpovídající záznamy v bibliografické databázi, nebo lze bibliografickou datovou kolekci vytvořit z článků obsažených v časopisech a konferenčních sbornících, které se zabývají danou oblastí vědy. Aktuálním trendem v hodnocení autorů vědeckých publikací je používání iteračních algoritmů, které při hodnocení významnosti dokáží využít vlastnosti citační sítě. Často používaným algoritmem, který bývá dále upravován, je algoritmus PageRank (Brin a Page 1998; Page et al. 1999), protože dle Bollen et al. (2006) a Ding (2011a) lze PageRankem měřit významnost lépe, než dříve používaným počtem citací. Významnost určená PageRankem bývá označována jako prestiž, kdežto významnost určená počtem citací bývá označována jako popularita (odlišení viz část 2.1).

V této kapitole je popsán náš návrh metod pro hodnocení prestiže autorů vědeckých publikací na základě citační analýzy, který byl publikován v (Nykl a Ježek 2012). Navržené metody pro hodnocení autorů byly testovány na citačních sítích vytvořených z datových kolekcí databází CiteSeer a DBLP, které shromažďují bibliografické záznamy z oblasti počítačových věd, viz část 2.2.2. V rámci experimentů byly vyhodnoceny citační sítě autorů a publikací, které různým způsobem penalizovaly samocitace a spolupráci autorů. Experimenty měly jednak ověřit, zda má smysl při hodnocení autorů odstraňovat samocitace, a dále posoudit, zda je dobré penalizovat spolupráci autorů. Dále jsme testovali, zda lze výsledky hodnocení autorů zlepšit využitím personalizace PageRanku, tj. zakomponováním další míry kvality do PageRanku. Autory jsme s využitím personalizace zvýhodnili počtem jejich publikací (tj. produktivitou) a publikace počtem jejich autorů (modelujícím vynaložené úsilí). Kvalita získaných pořadí autorů byla stanovena jejich porovnáním na základě manuálně vytvořených seznamů držitelů významných ocenění, která jsou udílena v oblasti počítačových věd. Zvolena byla ocenění *ACM Turing Award* (Turingova cena) a *ACM Codd Award* (Coddova cena) a dále seznamy *ACM Fellows* (významné osoby ACM) a *ISI Highly Cited Researchers* (vysoce citovaní výzkumníci ISI).

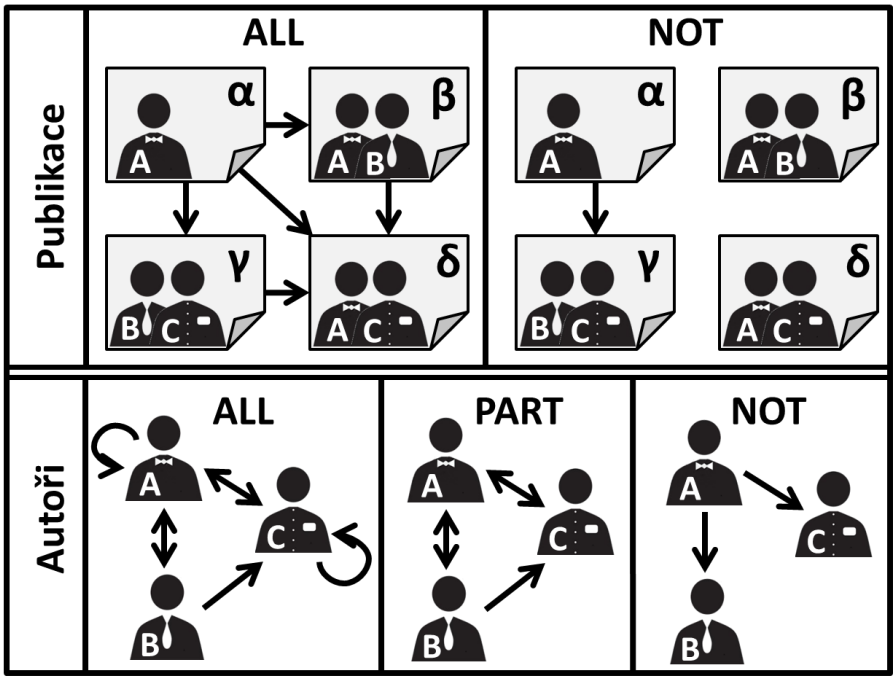
V části 3.1 jsou popsány naše postupy vytváření citačních sítí z bibliografických záznamů, v části 3.2 navržené metody pro hodnocení autorů a v části 3.3 datové kolekce CiteSeer a DBLP a referenční seznamy držitelů významných ocenění v oblasti počítačových věd. Diskuse výsledků je provedena v části 3.4 a vyvozené závěry shrnuty v části 3.5.

3.1 Vytváření citačních sítí s ohledem na samocitace a spoluautorství

Jedním z aspektů při hodnocení autorů na základě publikační činnosti je, kolika autorům publikace je publikace započítána. Například Ding (2011a) a Zhao (2005) použili pouze první či korespondující autory publikací a ostatní autory ignorovali, zatímco např. Fiala et al. (2008) použili všechny autory publikací. My při experimentech použili také všechny autory publikací. Dalším aspektem při hodnocení je míra toho, jak moc jsme ochotni připustit, aby významnost autora, či jiných bibliografických entit, byla ovlivněna autorovým citováním sama sebe, tzv. samocitacemi. My jsme testovali tři varianty použití samocitací:

- Varianta ***ALL*** používá samocitace jako plnohodnotné citace, což je nejvíce benevolentní.
- Varianta ***PART***, kterou jsme použili pouze pro citační sítě autorů, odstraňuje z citační sítě autorů vytvořené variantou *ALL* všechny smyčky. Lze říci, že samocitace jsou odstraněny na úrovni autorů. Stejný postup odstranění samocitací následně použili např. West et al. (2013).
- Varianta ***NOT*** odstraňuje samocitace autorů na úrovni publikací a to tak, že odstraní citace mezi publikacemi, které mají společného alespoň jednoho autora. Tuto variantu lze považovat za nejvíce striktní. Stejný postup použili např. Fiala et al. (2008).

Příklad citačních sítí vytvořených s použitím zmíněných variant samocitací demonstruje obrázek 3.1, na kterém jsou v horní části zobrazeny varianty citační sítě publikací a v dolní části odpovídající varianty citační sítě autorů.



Obrázek 3.1: Námi použité varianty samocitací autorů v citační síti publikací a v citační síti autorů (publikace jsou značeny α, β, γ, δ a autoři A, B, C).

Váhy hran v síti vyhodnocované algoritmem PageRank udávají proporcionalitu rozdělení hodnot vrcholů mezi jejich výstupní hrany. Váhy hran v citační síti autorů mohou být s ohledem na spoluautorství autorů stanoveny různými postupy, viz např. práce (Fiala et al. 2008; Fiala 2012b) popsané v části 2.5.2. My použili dvě základní varianty značené *N* a *1* a variantu *1/N*, která redukuje narůstání významnosti autora vlivem psaní článků s mnoha spoluautory. Naším předpokladem bylo, že článek, který autor napsal sám, má na hodnotu autorovy významnosti větší vliv, než článek, který autor napsal s mnoha spoluautory. V citační síti publikací jsme použili pouze variantu *1*. Popis použitých variant vah hran v citační síti autorů obsahuje následující seznam:

- Varianta ***N*** přiřazuje hranám váhy, které odpovídají počtu citací autora v citační síti publikací. Například pokud je autor A jedním z autorů dvou publikací citujících publikace, ve kterých je jedním z autorů autor B, tak v citační síti autorů existuje hrana z A do B s váhou 2.

- Ve variantě ***1/N*** je v citační síti autorů hodnota každé publikace rovnoměrně rozdělena mezi citované publikace. Například pokud autor A ve své publikaci cituje publikaci autorů B a C a publikaci autora E, tak v citační síti autorů existují hrany z A do B a z A do C, obě s váhou ½, a hrana z A do E s váhou 1. Souběžné hrany mezi dvojicemi autorů se následně sjednotí a jejich váhy se sečtou.
- Varianta ***1*** přiřazuje všem hranám váhu 1 a součet vah vstupních hran autora tak udává počet různých autorů, kteří autora citovali. Tuto variantu lze považovat za nejvíce přísnou, protože nezohledňuje počet citací od jednotlivých autorů.

Jakým způsobem byly přiřazeny váhy hranám v citačních sítích autorů zobrazených na obrázku 3.1, ukazuje tabulka 3.1. Pozn.: pokud z vrcholu vede pouze jedna výstupní hrana, tak při vyhodnocení PageRankem nezáleží na váze této hrany, protože využitím této hrany vrchol předá celou svou hodnotu odkazovanému vrcholu, viz např. hrana (C,A) z vrcholu C do vrcholu A ve variantě citační sítě autorů PART-1/N na obrázku 3.1. Z použitých variant samocitací a vah hran lze vydedukovat, že nejvíce benevolentní varianta citační sítě autorů je *ALL-N* a nejvíce striktní varianta je *NOT-1*.

Tabulka 3.1: Přiřazení vah hranám v citačních sítích autorů zobrazených na obrázku 3.1.

Hrana	ALL			PART			NOT		
	N	1/N	1	N	1/N	1	N	1/N	1
(A, B)	2	1	1	2	1	1	1	0,5	1
(A, C)	3	1,5	1	3	1,5	1	1	0,5	1
(B, A)	2	1	1	2	1	1	- - -	- - -	- - -
(B, C)	2	1	1	2	1	1	- - -	- - -	- - -
(C, A)	1	0,5	1	1	0,5	1	- - -	- - -	- - -
(A, A)	3	1,5	1	- - -	- - -	- - -	- - -	- - -	- - -
(C, C)	1	0,5	1	- - -	- - -	- - -	- - -	- - -	- - -

3.2 Metody pro hodnocení autorů založené na PageRanku

Citační sítě autorů a publikací jsme vyhodnotili jak PageRankem bez personalizace, tak i PageRankem, který je doplněný o personalizaci. Jako personalizaci autorů jsme použili hodnoty jejich produktivity zastoupené počtem jejich publikací, přičemž jsme předpokládali, že produktivita je jednou z významných vlastností autorů. Pro zajímavost můžeme uvést alternativní výklad k teorii náhodných surfařů, která byla zmíněna v části 2.4.1 u problému Rank sink: *„při nastavení personalizace autorů dle produktivity autorů lze z pohledu algoritmu PageRank říci, že čtenář/surfař si jednou za sedm ‘kroků‘ náhodně vybere autora (‘teleportuje se‘) s pravděpodobností úměrnou podílu počtu publikací autora vůči počtu všech publikací, které jsou obsažené v kolekci“*. Nejsnáze získatelnou informací o publikacích, kterou jsme použili jako jejich personalizaci, pro nás byl počet autorů publikace. V tomto případě jsme předpokládali, že počet autorů publikace odráží kvalitu publikace z pohledu využitých zdrojů (vědců a investovaných prostředků). Jak zmiňuje např. Glänzel (2001), publikace napsané autory z více států jsou citovány častěji. Nás zajímalo, zda větší počet autorů publikace přispívá ke zvýšení kvality publikace a tím k lepšímu hodnocení autorů.

Na tomto místě zopakujeme použité vzorce PageRanku, které byly v části 2.4 značeny jako vzorce (2.11) a (2.12). PageRank bez personalizace popisuje vzorec (3.1), kde $PR_x(A)$ je hodnota PageRanku vrcholu A v iteraci x, d=0,85 je faktor tlumení, $|V|$ je velikost množiny všech vrcholů grafu, U_A je

množina vrcholů odkazujících na vrchol *A*, w_{utoA} je váha hrany vedoucí z vrcholu *u* do vrcholu *A*, w_{uout} je součet vah všech výstupních hran vrcholu *u* a *D* je množina slepých vrcholů (tj. vrcholů bez výstupních hran). PageRank s personalizací popisuje vzorec (3.2), kde je navíc *P* množina personalizací všech vrcholů a p_A je hodnota personalizace vrcholu *A*. Oba vzorce obsahují námi navržené ošetření slepých vrcholů, které urychluje výpočet hodnot PageRanku, viz část 2.4. Pro výpočet hodnot vrcholů jsme vždy použili 50 iterací, což je pro konvergenci PageRanku dostatečné (Nykl 2011).

$$PR_{x+1}(A) = \frac{(1-d)}{|V|} + d \cdot \left(\sum_{u \in U_A} \frac{PR_x(u) \cdot w_{u\text{toA}}}{w_{u\text{out}}} + \frac{1}{|V|} \sum_{s \in D} PR_x(s) \right) \tag{3.1}$$

$$PR_{x+1}(A) = \frac{(1-d) \cdot p_A}{\sum_{p \in P} p} + d \cdot \left(\sum_{u \in U_A} \frac{PR_x(u) \cdot w_{u\text{toA}}}{w_{u\text{out}}} + \frac{1}{|V|} \sum_{s \in D} PR_x(s) \right) \tag{3.2}$$

Hodnoty publikací vypočtené z citační sítě publikací zmíněnými variantami PageRanku jsme rozdělili jejich autorům. Rozdělení hodnot bylo provedeno dvěma způsoby. Buď jsme autorovi sečetli celé hodnoty jeho publikací (varianta *SUM*), viz vzorec (3.3), nebo jsme mu sečtli pouze rovnoměrné díly hodnot jeho publikací, určené na základě počtu autorů každé publikace (varianta *DIV*), viz vzorec (3.4). Ve vzorcích (3.3) a (3.4) je *SUM(A)* a *DIV(A)* hodnota autora *A* získaná daným rozdělením, Ψ_A je množina všech publikací autora *A*, *VAL(Q)* je hodnota publikace *Q* a Q_N je počet autorů publikace *Q*.

$$SUM(A) = \sum_{Q \in \Psi_A} VAL(Q) \tag{3.3}$$

$$DIV(A) = \sum_{Q \in \Psi_A} \left[\frac{1}{Q_N} \cdot VAL(Q) \right] \tag{3.4}$$

3.3 Zvolené datové kolekce a seznamy významných autorů

Všechny navržené metody mohou být použity pro hodnocení autorů v libovolných bibliografických kolekcích. My pro experimentální ověření kvality navržených metod použili kolekce bibliografických databází CiteSeer (Giles et al. 1998) a DBLP (Ley 1993), které jsme zvolili, aby naše výsledky mohly být porovnány s pracemi (Sidiropoulos a Manolopoulos 2005a, 2005b; Fiala et al. 2008; Fiala 2011). Obě databáze shromažďují bibliografické záznamy z oblasti počítačových věd, viz část 2.2.2. CiteSeer je udržován strojově a my použili jeho kolekci z roku 2005 (pozn.: nová verze CiteSeerX[42] nebyla v době psaní článku k dispozici). DBLP[43] je udržována manuálně a my použili její kolekci z roku 2004, protože kolekce z let 2006 a 2009 obsahují téměř totožný počet citací jako verze z roku 2004, ale daleko více

[42] Web CiteSeerX - http://citeseerx.ist.psu.edu
[43] Web DBLP - http://dblp.uni-trier.de

publikací a autorů, viz (Heller et al. 2011), což pro nás není zajímavé (pozn.: aktuálně je dostupná kolekce DBLP z roku 2015[44]).

V dále popisovaných citačních sítích *vrcholy* reprezentují publikace nebo autory a *hrany* reprezentují citace. Jednotlivé typy vrcholů jsou:

- ***Slepé*** vrcholy, které nikoho necitují,
- ***Citující*** vrcholy, které někoho citují,
- ***Necitované*** vrcholy, které nejsou nikým citovány,
- ***Citované*** vrcholy, které jsou někým citovány,
- ***Izolované*** vrcholy, které jsou slepé a necitované.

Kvantitativní údaje shrnující charakteristické vlastnosti citačních sítí vytvořených ze zvolených kolekcí CiteSeer a DBLP obsahuje tabulka 3.2, ve které jsou uvedeny počty vrcholů, hran a vrcholů jednotlivých typů, které jsou obsaženy v jednotlivých sítích. Tyto údaje mohou být použity pro porovnání námi zvolených kolekcí s jinými kolekcemi a také pro porovnání našich výsledků s výsledky podobných experimentů. Za povšimnutí stojí velmi malý počet citací obsažených v kolekci DBLP (průměrně 0,22 citací na publikaci, kdežto v CiteSeer je průměrně 2,38 citací na publikaci) a s tím související malý počet citujících a citovaných vrcholů, tj. vrcholů, které ovlivňují výpočet PageRanku. Pozn.: pokud při výpočtu PageRanku není použita personalizace, tak všechny necitované vrcholy získají stejnou minimální hodnotu PageRanku a budou na poslední pozici ve vytvořeném pořadí.

Tabulka 3.2: Kvantitativní údaje citačních sítí vytvořených z bibliografických kolekcí CiteSeer 2005 a DBLP 2004.

DB	Typ sítě	Vrcholů	Samoc.	Hran	Slepé	Citující	Necitované	Citované	Izolované
CiteSeer (2005)	publikace	648897	ALL	1542165	343081	305816	469086	179811	299460
			NOT	1221024	375952	272945	505245	143652	331937
	autoři	406465	ALL	5398239	183541	222924	244289	162176	154826
			PART	5337600	184140	222325	245278	161187	155073
			NOT	4702724	199340	207125	266675	139790	169056
DBLP (2004)	publikace	469804	ALL	100621	461856	7948	451524	18280	448965
			NOT	84804	462060	7744	454103	15701	451071
	autoři	315485	ALL	331245	308314	7171	301222	14263	299546
			PART	328150	308318	7167	301267	14218	299548
			NOT	295531	308493	6992	302551	12934	300307

Pro úplnost můžeme uvést, že v CiteSeer byly publikace průměrně napsány 2,5 autory a autoři průměrně napsali 4 publikace. V DBLP byly publikace průměrně napsány 2,3 autory a autoři průměrně napsali 3,4 publikace.

3.3.1 Seznamy držitelů významných ocenění

Pořadí autorů, která jsme získali vyhodnocením kolekcí CiteSeer a DBLP navrženými metodami pro hodnocení autorů, jsme porovnali na základě manuálně vytvořených seznamů držitelů významných

[44] Datová kolekce DBLP 2015 - http://dblp1.uni-trier.de/xml/

ocenění v oblasti počítačových věd. K tomuto účelu byli použiti držitelé prestižní Turingovy nebo Coddovy ceny a osoby uvedené na seznamu významných osob ACM nebo na seznamu vysoce citovaných výzkumníků ISI. Náš postup porovnání vytvořených pořadí autorů na základě držitelů významných ocenění je podobný postupu, který použili Sidiropoulos a Manolopoulos (2006), kteří k porovnání vytvořených pořadí autorů použili držitele Coddovy ceny. Námi použitá ocenění jsou popsána v následujícím seznamu:

- ***ACM A. M. Turing Award***[45], *Turingova cena* – je nejprestižnější technickou cenou udílenou ACM za významné přispění s dlouhotrvajícím vlivem v oblasti počítačových věd.
- ***ACM SIGMOD Edgar F. Codd Innovations Award***[46]*, Coddova cena* – je udílena za inovativní a významné příspěvky trvalé hodnoty k vývoji, pochopení nebo používání databázových systémů a databází.
- ***ACM Fellows***[47]*, významné osoby ACM* – program ACM Fellows vznikl v roce 1993 s cílem rozpoznat a ocenit vynikající členy ACM za jejich úspěchy v oboru počítačových věd a informačních technologií a za jejich významné přispění k poslání ACM.
- ***ISI Highly Cited Researchers***[48]*, vysoce citovaní výzkumníci ISI* – seznam těchto výzkumníků vytvořil mezi lety 2000 až 2008 tým Thomson Reuters na základě analýzy článků obsažených v bibliografické kolekci databáze ISI Web of Science z let 1981 až 2003. (Pozn.: aktuálně existuje nová verze z roku 2014.)

Držitele Coddovy ceny jsme pro jejich malý počet nalezli v kolekcích manuálně. Protože kolekce CiteSeer 2005 dle našeho zjištění obsahuje indexovací chyby ve jménech autorů (např. špatné pořadí jmen a jejich neúplnost), tak pro nalezení držitelů ostatních ocenění jsme použili následující postup:

1) Nalezli jsme jména autorů, která obsahují příjmení oceněných autorů na libovolné pozici, tj. nerozlišovali jsme, zda je příjmení uvedeno na první nebo jiné pozici ve jméně uvedeném v kolekci.
2) Odstranili jsme příjmení oceněných autorů, která se v kolekci nacházela ve více jak dvaceti variantách jmen.
3) Pro každé ze zbylých příjmení jsme nastálo určili jedno odpovídající zástupné jméno z kolekce a to to, které bylo ve vytvořených pořadích autorů na nejvyšší pozici (pozn.: ve všech případech se vždy jednalo o stejná jména).

Tabulka 3.3: Počty nalezených oceněných autorů v kolekcích CiteSeer 2005 a DBLP 2004.

	ACM Turing	ACM Codd	ACM Fellows	ISI High.Cit.
Oceněných osob celkem	57	19	809	364
Nalezeno v CiteSeer (2005)	16	18	247	146
Nalezeno v DBLP (2004)	12	19	192	91

[45] Web *ACM A.M. Turing Award* - http://amturing.acm.org
[46] Web *ACM SIGMOD E.F. Codd Innovations Award* - http://www.sigmod.org/sigmod-awards/
[47] Web *ACM Fellows* - http://fellows.acm.org
[48] Web *ISI Highly Cited Researchers*, dříve jen *ISI Highly Cited* - http://www.highlycited.com/?dtyear=2011

Tabulka 3.3 zobrazuje počty autorů oceněných jednotlivými oceněními do roku 2012 (pozn.: rok, ve kterém jsme experiment prováděli) a počty jim odpovídajících jmen nalezených v kolekcích CiteSeer 2005 a DBLP 2004. Odstranění víceznačnosti jmen autorů (*name disambiguation*) provedeno nebylo.

3.4 Diskuse výsledků vyhodnocení kolekcí CiteSeer a DBLP

Ze záznamů obsažených v kolekcích CiteSeer 2005 a DBLP 2004 jsme vytvořili všechny typy citačních sítí, které jsme popsali v části 3.1, aplikovali jsme na ně metody pro hodnocení autorů, které byly uvedeny v části 3.2, a dle hodnot autorů jsme pro každou metodu vytvořili pořadí autorů. Autor, který byl metodou hodnocen nejlépe, obsadil 1. pozici ve vytvořeném pořadí autorů. Pokud mělo více autorů shodnou hodnotu, tak jejich pozice byla stanovena jako průměr z obsazených pozic (např. pokud 4 autoři se stejnou hodnotou sdílejí 7. až 10. pozici, tak každému z nich je přidělena pozice 8,5). Ve vytvořených pořadích autorů jsme nalezli oceněné autory a z jejich pozic v pořadí jsme pro každou metodu vypočetli průměrnou pozici oceněných autorů ve vytvořeném pořadí. Tyto průměrné pozice oceněných autorů udávají kvalitu jednotlivých metod použitých pro hodnocení autorů, která je určena na základě schopnosti dané metody vyzdvihnout v hodnocení autory uvedené na konkrétním seznamu oceněných autorů.

Naše metody jsou porovnány v tabulkách 3.4 a 3.5, viz dále, ve kterých jsou ve sloupcích ***průměr*** zobrazeny hodnoty průměrných pozic oceněných autorů ve vytvořených pořadích autorů. Pokud metody seřadíme dle sloupce *průměr* od nejmenších (nejlepších) hodnot k nejvyšším hodnotám a v tomto pořadí očíslujeme, tak získáme pořadí úspěšnosti jednotlivých metod pro hodnocení autorů ve vyzdvižení oceněných autorů. Toto pořadí je zobrazeno ve sloupci ***p.***. Nejmenší průměrná pozice autorů oceněných daným oceněním, kterou naše metody vypočetly, je pro každý seznam oceněných autorů zobrazena v řádku ***Minimum*** *(m_{best})*. Procentuální odlišení průměrných pozic oceněných autorů v pořadích získaných všemi metodami od nejmenší průměrné pozice, která byla získána metodou s *p.*=1, ukazují sloupce ***$m_{\%}$***. V každém sloupci jsou navíc tři nejlepší metody s $p. \in [1,2,3]$ zvýrazněny černým pozadím s bílým písmem a tři nejhorší metody s $p. \in [24,25,26]$ jsou zvýrazněny šedým pozadím s tučným černým písmem.

V tabulkách 3.4 a 3.5 je PageRank bez personalizace značen ***BezP.*** (*bez personalizace*), PageRank s personalizací dle počtu publikací je značen ***P.P.P.*** (*personalizace počtem publikací*) a PageRank s personalizací dle počtu autorů publikace je značen ***P.P.A*** (*personalizace počtem autorů*).

3.4.1 Hodnocení autorů z kolekce CiteSeer

Analýza výsledků aplikování navržených metod pro hodnocení autorů na kolekci CiteSeer 2005, viz tabulka 3.4, ukázala, že pro téměř všechny seznamy oceněných osob (*ACM Turing*, *ACM Fellows* a *ISI Highly Cited*) byly nejlepší metody, které hodnoty významnosti autorů stanovily na základě vyhodnocení citační sítě autorů s odstraněnými samocitacemi (*NOT*) PageRankem bez personalizace (*BezP.*). V tomto případě tedy personalizace autorů počtem publikací (tj. produktivitou) zhoršila výpočet prestiže. Zvolený typ vah hran měl na hodnocení autorů nejmenší vliv, ale mírně lepších výsledků bylo dosaženo, když byly zanedbány počty citací od jednotlivých autorů a všechny hrany v síti měly váhu 1. Naší nejlepší metodou, která autory uvedené na seznamech oceněných autorů při hodnocení autorů nejvíce vyzdvihla, byla nejpřísnější metoda značená *Autoři-NOT-1-BezP.*. Tato metoda v citační síti autorů nepoužívá samocitace, rozdílné váhy hran, ani personalizaci. Metody pro hodnocení autorů, které pracují s citační sítí publikací, měly v kolekci CiteSeer nejhorší výsledky.

Porovnání získaných pořadí autorů na základě seznamu *ACM Codd* neposkytlo zcela stejné závěry, jako porovnání pořadí na základě ostatních seznamů oceněných osob. To může být způsobeno tím, že Coddova cena je udílena v oblasti databázových systémů a databází, ale CiteSeer indexuje články z celé oblasti počítačových věd. Pro seznam *ACM Codd* byla nejlepší pořadí autorů vytvořena metodami, které také vyhodnocují citační síť autorů s vahami hran *1*, ale navíc autory zvýhodňují počtem jejich publikací (*P.P.P.*). V tomto případě tedy využití personalizace zlepšilo výsledky. Nejmenší vliv na hodnocení autorů zde měl použitý typ samocitací, přičemž nejlepší metoda pro seznam *ACM Codd* byla metoda *Autoři-PART-1-P.P.P.*, která odstraňuje samocitace na úrovni autorů. Za povšimnutí stojí, že 4. nejlepší metoda pro seznam *ACM Codd* hodnotila autory na základě hodnot jejich publikací, viz metoda *Publikace-NOT-SUM-P.P.A.*.

Tabulka 3.4: Výsledky metod pro hodnocení autorů v kolekci CiteSeer 2005 (tři nejlepší metody s p. ∈ [1,2,3] jsou zvýrazněny černým pozadím s bílým písmem a tři nejhorší metody s p. ∈ [24,25,26] jsou zvýrazněny šedým pozadím s tučným černým písmem).

Kolekce CiteSeer 2005 obsahuje 406465 odlišných autorů vědeckých publikací																
Síť	**Samocit.**	**Váhy**	**Person.**	**ACM Turing (16 osob)**			**ACM Codd (18 osob)**			**ACM Fellows (247 osob)**			**ISI Highly Cited (146 osob)**		
				průměr	p.	$m_{\%}$	průměr	p.	$m_{\%}$	průměr	p.	$m_{\%}$	průměr	p.	$m_{\%}$
Autoři	**NOT**	**1**	BezP.	**52168**	**1**	**0%**	4888	16	32%	**43151**	**1**	**0%**	**42086**	**1**	**0%**
			P.P.P.	59408	6	14%	**3930**	**3**	**6%**	47503	6	10%	45726	7	9%
		1/N	BezP.	**54640**	**2**	**5%**	**6115**	**26**	**65%**	**45022**	**3**	**4%**	44194	4	5%
			P.P.P.	64220	14	23%	4617	14	24%	52015	15	21%	49978	15	19%
		N	BezP.	**54654**	**3**	**5%**	**5751**	**25**	**55%**	**44216**	**2**	**2%**	**42873**	**2**	**2%**
			P.P.P.	63624	12	22%	4411	9	19%	50040	13	16%	47647	12	13%
	PART	**1**	BezP.	58673	5	12%	4541	12	22%	46437	4	8%	**44018**	**3**	**5%**
			P.P.P.	63814	13	22%	**3714**	**1**	**0%**	48645	8	13%	46439	8	10%
		1/N	BezP.	61086	7	17%	5398	22	45%	48782	9	13%	46669	10	11%
			P.P.P.	67397	17	29%	4370	8	18%	52551	17	22%	51267	17	22%
		N	BezP.	61384	8	18%	4942	18	33%	48238	7	12%	45597	6	8%
			P.P.P.	67297	15	29%	4051	5	9%	51451	14	19%	49154	14	17%
	ALL	**1**	BezP.	58461	4	12%	4676	15	26%	47046	5	9%	44708	5	6%
			P.P.P.	63485	11	22%	**3795**	**2**	**2%**	49069	10	14%	46944	11	12%
		1/N	BezP.	61970	10	19%	5500	23	48%	49757	12	15%	47728	13	13%
			P.P.P.	68050	19	30%	4437	11	19%	53563	18	24%	52383	18	24%
		N	BezP.	61609	9	18%	5045	19	36%	49078	11	14%	46494	9	10%
			P.P.P.	67376	16	29%	4110	6	11%	52215	16	21%	50161	16	19%
Publikace	**NOT**	**DIV**	BezP.	76873	22	47%	5301	21	43%	**72575**	**26**	**68%**	**71431**	**26**	**70%**
			P.P.A.	76757	21	47%	4901	17	32%	**69370**	**24**	**61%**	**68432**	**24**	**63%**
		SUM	BezP.	67508	18	29%	4113	7	11%	56872	20	32%	56449	19	34%
			P.P.A.	77397	23	48%	4001	4	8%	61944	22	44%	61513	21	46%
	ALL	**DIV**	BezP.	**79105**	**25**	**52%**	**5723**	**24**	**54%**	**70155**	**25**	**63%**	**68877**	**25**	**64%**
			P.P.A.	**77520**	**24**	**49%**	5219	20	41%	66207	23	53%	65505	23	56%
		SUM	BezP.	70616	20	35%	4585	13	23%	56455	19	31%	56658	20	35%
			P.P.A.	**79792**	**26**	**53%**	4417	10	19%	60759	21	41%	61700	22	47%
Minimum (m_{best})				52168			3714			43151			42086		

3.4.2 Hodnocení autorů z kolekce DBLP

Diskusi výsledků získaných z vyhodnocení kolekce DBLP 2004, viz tabulka 3.5, začneme sloupcem *ACM Codd*, protože autoři ocenění Coddovou cenou (oblast databázových systémů) byli naší metodou *Autoři–ALL–N–P.P.P.* umístěni na nejlepší průměrnou 30. pozici. Z toho lze usoudit, že DBLP se stále soustředí na oblast databázových systémů a tato oblast je v ní dobře zastoupena. Přihlédneme-li ke skutečnosti, že DBLP obsahuje 315 486 rozdílných autorů, tak lze říci, že naše metoda dokázala 19 držitelů Coddovy ceny ve vytvořeném pořadí autorů vyzdvihnout velmi dobře. Z analýzy výsledků metod *Autoři–ALL–N–?*, které mají v porovnání se seznamem *ACM Codd* nejlepší výsledky, můžeme poukázat na vliv článků držitelů Coddovy ceny na vědeckou literaturu indexovanou v DBLP – citace článků držitelů tohoto ocenění a jejich následovníků jsou v DBLP obsaženy ve větším počtu, což je v našich metodách zahrnuto ve vahách hran (*N)* a v použití všech samocitací (*ALL*).

Tabulka 3.5: Výsledky metod pro hodnocení autorů v kolekci DBLP 2004 (tři nejlepší metody s p. ∈ [1,2,3] jsou zvýrazněny černým pozadím s bílým písmem a tři nejhorší metody s p. ∈ [24,25,26] jsou zvýrazněny šedým pozadím s tučným černým písmem).

Kolekce DBLP 2004 obsahuje 315485 odlišných autorů vědeckých publikací																
Síť	**Samocit.**	**Váhy**	**Person.**	**ACM Turing (12 osob)**			**ACM Codd (19 osob)**			**ACM Fellows (192 osob)**			**ISI Highly Cited (91 osob)**		
				průměr	p.	$m_{\%}$	průměr	p.	$m_{\%}$	průměr	p.	$m_{\%}$	průměr	p.	$m_{\%}$
Autoři	NOT	1	BezP.	3389	21	108%	47	22	57%	3802	10	14%	4018	12	17%
			P.P.P.	2382	13	46%	41	13	36%	**3348**	**1**	**0%**	**3421**	**1**	**0%**
		1/N	BezP.	3059	18	88%	41	14	37%	3626	5	8%	3795	7	11%
			P.P.P.	2197	9	35%	39	11	30%	3564	4	6%	3669	4	7%
		N	BezP.	3425	22	110%	35	6	15%	3764	8	12%	4017	11	17%
			P.P.P.	2310	12	42%	32	4	8%	3645	6	9%	3769	6	10%
	PART	1	BezP.	3479	23	113%	47	20	56%	4195	15	25%	4431	15	30%
			P.P.P.	2384	14	46%	41	12	36%	**3535**	**2**	**6%**	**3556**	**2**	**4%**
		1/N	BezP.	3074	19	89%	38	10	25%	4000	13	19%	4167	13	22%
			P.P.P.	2260	10	39%	36	8	20%	3702	7	11%	3711	5	8%
		N	BezP.	**3535**	**24**	**117%**	33	5	11%	4208	16	26%	4483	16	31%
			P.P.P.	2409	15	48%	**31**	**2**	**2%**	3840	11	15%	3886	9	14%
	ALL	1	BezP.	**3543**	**25**	**117%**	47	21	57%	4246	17	27%	4487	17	31%
			P.P.P.	2417	16	48%	41	15	37%	**3553**	**3**	**6%**	**3576**	**3**	**5%**
		1/N	BezP.	3175	20	95%	37	9	24%	4061	14	21%	4234	14	24%
			P.P.P.	2307	11	42%	36	7	19%	3793	9	13%	3885	8	14%
		N	BezP.	**3630**	**26**	**123%**	**32**	**3**	**7%**	4270	18	28%	4551	18	33%
			P.P.P.	2453	17	50%	**30**	**1**	**0%**	3872	12	16%	3913	10	14%
Publikace	NOT	DIV	BezP.	**1630**	**1**	**0%**	**81**	**26**	**172%**	**6534**	**25**	**95%**	**7022**	**26**	**105%**
			P.P.A.	**1646**	**3**	**1%**	79	23	162%	6105	22	82%	6444	21	88%
		SUM	BezP.	2013	5	24%	45	19	50%	5603	19	67%	5775	20	69%
			P.P.A.	2160	7	33%	45	18	49%	6443	23	92%	6493	22	90%
	ALL	DIV	BezP.	**1643**	**2**	**1%**	**80**	**25**	**168%**	**6537**	**26**	**95%**	**6975**	**25**	**104%**
			P.P.A.	1662	4	2%	**79**	**24**	**163%**	6079	21	82%	**6594**	**24**	**93%**
		SUM	BezP.	2033	6	25%	44	17	47%	5612	20	68%	5711	19	67%
			P.P.A.	2182	8	34%	43	16	44%	**6500**	**24**	**94%**	6514	23	90%
Minimum (m_{best})				1630			30			3348			3421		

Z porovnání výsledků našich metod v kolekci DBLP lze dále vidět, že autoři ocenění *ACM Fellows* nebo *ISI Highly Cited* jsou ve vytvořených pořadích autorů nejvíce vyzdviženi metodami *Autoři-?–1–P.P.P.*, které v síti autorů nevyužívají rozdílné váhy hran (tj. váhy hran jsou *1*) a autory zvýhodňují počtem publikací (*P.P.P.*). Nejmenší vyliv na vytvoření pořadí autorů měl použitý typ samocitací, ale nejlepší bylo odstranit samocitace autorů na úrovni publikací (*NOT*). Nejlepší metoda pro tyto dva seznamy oceněných autorů byla tedy téměř stejná jako v kolekci CiteSeer, ale v kolekci DBLP se projevila vhodnost naší úpravy personalizace PageRanku počtem publikací (*P.P.P.*) autorů.

Seznam *ACM Turing* byl jediným seznamem oceněných autorů, pro který jsme nejlepší pořadí autorů získali metodami pracujícími s citační sítí publikací, přičemž téměř nezáleželo na použitém typu samocitací (*ALL* nebo *NOT*) ani na tom, zda byla či nebyla použita personalizace publikací dle počtu jejich autorů (*BezP.* nebo *P.P.A.*). Lepší pořadí autorů ale byla získána rovnoměrným rozdělením hodnot publikací jejich autorům (*DIV*), nežli sčítáním celých hodnot publikací (*SUM*). Z dobrých výsledků, které pro seznam *ACM Turing* poskytly metody pracující s citační sítí publikací, jsme usoudili, že tyto metody by neměly být vynechány z dalších experimentů.

3.5 Závěry z hodnocení autorů z kolekcí CiteSeer a DBLP

V této kapitole jsme navrhli metody pro hodnocení autorů a analyzovali výsledky z hodnocení autorů obsažených v kolekcích CiteSeer 2005 a DBLP 2004. Pro hodnocení autorů jsme navrhli a použili dvě úpravy personalizace PageRanku, které měly zvýhodnit buď produktivnější autory (tj. autory s větším počtem publikací), nebo kvalitnější publikace (tj. v našem případě publikace napsané více autory, což mělo vyjadřovat, že na jejich vytvoření bylo vynaloženo více úsilí a zdrojů). Vyhodnocované citační sítě autorů různým způsobem penalizovaly samocitace a spolupráci autorů. Pokud byla použita citační síť publikací, tak jsme hodnoty publikací rozdělili jejich autorům tak, že jsme buď každému autorovi sečetli celé hodnoty jeho publikací, nebo jsme mu sečetli rovnoměrné díly těchto hodnot, určené na základě počtu autorů publikace. Protože kolekce CiteSeer a DBLP obsahují bibliografické záznamy z oblasti počítačových věd, tak pro účely určení kvality a porovnání získaných pořadí autorů jsme vytvořili seznamy autorů, kteří jsou držiteli významného ocenění v oblasti počítačových věd. Zvoleni byli držitelé Turingovy nebo Coddovy ceny a autoři, kteří jsou uvedeni na seznamu významných osob ACM nebo na seznamu vysoce citovaných výzkumníků ISI.

Při hodnocení autorů z kolekce CiteSeer 2005, kterou jsme kritizovali, protože obsahuje indexovací chyby, měly naše metody téměř stabilní výsledky, tj. pořadí úspěšnosti použitých metod ve vyzdvižení oceněných autorů se příliš nelišilo a pro téměř všechny seznamy oceněných autorů byla nejlepší shodná metoda pro hodnocení autorů. Tato metoda, značená *Autoři-NOT-1-BezP.*, aplikovala PageRank bez personalizace na citační síť autorů, ve které byly samocitace autorů odstraněny na úrovni publikací a váhy všech hran byly 1. V tomto případě se tedy vhodnost navržených úprav personalizace PageRanku nepotvrdila. Výjimku tvořilo pouze porovnání použitých metod na základě seznamu autorů oceněných Coddovou cenou. Zde byla navržená personalizace autorů počtem publikací vhodná. Důvodem, proč není pro seznam držitelů Coddovy ceny nejlepší stejná metoda jako pro zbylé seznamy oceněných autorů, může být skutečnost, že Coddova cena je udílena v oblasti databázových systémů a databází, zatímco CiteSeer indexuje vědecké články z celé oblasti počítačových věd.

Při analýze kvality metod pro hodnocení autorů aplikovaných na kolekci DBLP 2004, která dle našeho zjištění stále indexuje převážně vědecké články z oblastí databází a logického programování, jsme se

nejprve zaměřili na porovnání získaných pořadí autorů na základě držitelů Coddovy ceny, která je udílena v oblasti databází. Pro tento seznam oceněných autorů jsme metodou *Autoři-ALL-N-P.P.P.* získali velmi pěkný výsledek – autoři ocenění Coddovou cenou (19 osob) ve vytvořeném pořadí autorů obsadili mezi všemi hodnocenými autory průměrnou 30. pozici, přičemž DBLP 2004 obsahuje 315 486 rozdílných autorů. Tato metoda aplikuje na citační síť autorů se všemi samocitacemi a váhami hran, které vyjadřují počet citování, PageRank s personalizací autorů zastoupenou počtem jejich publikací, což potvrzuje vhodnost naší úpravy personalizace. Z analýzy této metody jsme usoudili, že držitelé Coddovy ceny měli značný vliv na výběr článků indexovaných v DBLP a na kvalitu indexace jejich citací.

Autoři ze seznamů významných osob ACM a vysoce citovaných výzkumníků ISI byli v DBLP nejlépe hodnoceni podobnou metodou jako v CiteSeer, pouze se zde nově potvrdila vhodnost zakomponování produktivity autorů do jejich personalizace. Seznam autorů oceněných Turingovou cenou byl jediným seznamem, pro který bylo lepší vyhodnotit síť publikací než síť autorů. Z toho důvodu jsme doporučili nevynechávat sítě publikací z dalších experimentů. S ohledem na experimenty se seznamy držitelů Coddovy a Turingovy ceny můžeme databázi DBLP doporučit spíše pro strojové hodnocení autorů specializujících se na oblast databázových systémů, nežli pro hodnocení autorů, kteří publikují v jiných oblastech počítačových věd.

Na základě provedených experimentů nelze obecně určit, zda navržené personalizace PageRanku umožňují lepší hodnocení autorů než PageRank bez personalizace. V CiteSeer byla personalizace vhodná pouze pro 1 ze 4 seznamů oceněných autorů a v DBLP pro 3 ze 4 seznamů oceněných autorů. Odchylky mohou být způsobeny jednak malým počtem indexovaných citací a specializovaností DBLP, ale i vlivem indexovacích chyb v CiteSeer. Z důvodu nejednoznačného závěru jsme se rozhodli otestovat naše metody na kvalitnější bibliografické kolekci, za kterou považujeme kolekci databáze ISI Web of Science. Experiment s touto kolekcí je popsán v následující 4. kapitole. V rámci experimentu jsme navíc k metodám pro hodnocení prestiže autorů přidali metody pro hodnocení popularity, abychom ukázali, že využitím prestiže lze hodnotit autory lépe, než využitím popularity.

4 Ověření kvality navržených metod v kolekci ISI Web of Science

V předchozí kapitole jsme navrhli několik metod pro hodnocení vědeckých pracovníků na základě jejich publikační činnosti a na datech z bibliografických kolekcí CiteSeer 2005 a DBLP 2004 jsme testovali jejich schopnost vyzdvihnout při hodnocení osoby, které byly oceněny významným oceněním. K tomuto účelu jsme vytvořili seznamy osob, které obdržely Turingovu nebo Coddovu cenu anebo se nacházejí na seznamu významných osob ACM nebo na seznamu vysoce citovaných výzkumníků ISI. Nicméně, v závěru kapitoly jsme zmínili, že obě datové kolekce obsahují chybné nebo nekompletní informace, protože CiteSeer obsahuje chyby v automaticky indexovaných záznamech o vědeckých publikacích (chybná jména a duplicitní záznamy) a specializovaná DBLP obsahuje málo citací. Z těchto důvodů jsme v následujícím experimentu otestovali námi navržené metody pro hodnocení autorů na datové kolekci, kterou jsme koupili od provozovatelů bibliografické databáze ISI Web of Science (WoS). Tento experiment byl prezentován v článku (Nykl et al. 2014) a je popsán v této kapitole. Do databáze WoS jsou týmem Thomson Reuters manuálně vkládány bibliografické záznamy o publikacích z vybraných vědeckých zdrojů (časopisy, sborníky apod.), a proto by její kolekce měla být z pohledu indexovaných údajů kvalitnější než kolekce CiteSeer a DBLP. Z toho důvodu by i hodnocení autorů založené na kolekci WoS mělo být věrohodnější. Zakoupena byla kolekce WoS, která obsahuje bibliografické záznamy z kategorií počítačových věd.

V části 4.1 jsou zmíněny cíle a předpoklady, které jsme měli při experimentování s datovou kolekcí WoS. V části 4.2.1 jsou v krátkosti zopakovány naše způsoby vytváření citačních sítí, určování vah hran v síti autorů a způsoby rozdělení hodnot publikací jejich autorům. Také jsou zde ukázány charakteristické kvantitativní údaje citačních sítí vytvořených z kolekce WoS a zmíněny rozdíly mezi kolekcemi WoS, CiteSeer a DBLP. Popis manuálně vytvořených seznamů držitelů významných ocenění obsahuje část 4.2.2. V části 4.3 jsou popsány metody použité pro hodnocení autorů vědeckých publikací, které pracují s citační sítí autorů nebo publikací. Výsledky aplikování našich metod pro hodnocení autorů na citační sítě vytvořené z kolekce WoS jsou ukázány a diskutovány v části 4.4 a hlavní závěry shrnuty v části 4.5.

4.1 Cíle experimentu s datovou kolekcí ISI Web of Science

V prezentovaném experimentu jsme hodnotili autory, kteří jsou obsaženi v bibliografické datové kolekci WoS, metodami hodnotícími jejich prestiž a popularitu. Jak jsme zmínili v části 2.1, prestiž je zastoupena PageRankem, který iteračně počítá významnosti vrcholů v citační síti na základě významností vrcholů, které na ně odkazují. Popularita je zastoupena počítáním citací. My použitím popularity i prestiže chtěli potvrdit, že prestiž je lepší mírou pro hodnocení autorů než popularita.

Naším hlavním výzkumným záměrem bylo ověřit vhodnost metod, které jsme navrhli a použili pro vyhodnocení kolekcí DBLP a CiteSeer ve 3. kapitole. Dalším záměrem bylo zjistit, zda lepší pořadí autorů (měřeno oceněnými autory) získáme vyhodnocením citační sítě publikací nebo vyhodnocením citační sítě autorů, ve které jsou pominuty některé informace (např. časový sled publikování a citování). Kvalita metod pro hodnocení autorů byla opět určena porovnáním získaných pořadí autorů na základě osob uvedených na manuálně vytvořených seznamech držitelů významných ocenění. Použiti byli držitelé Turingovy nebo Coddovy ceny a osoby zapsané na seznamu významných osob ACM nebo na seznamu vysoce citovaných výzkumníků ISI. Dále jsme, stejně jako ve 3. kapitole, testovali, jaký vliv na vytvořené pořadí autorů měly různé způsoby penalizace samocitací či spolupráce autorů.

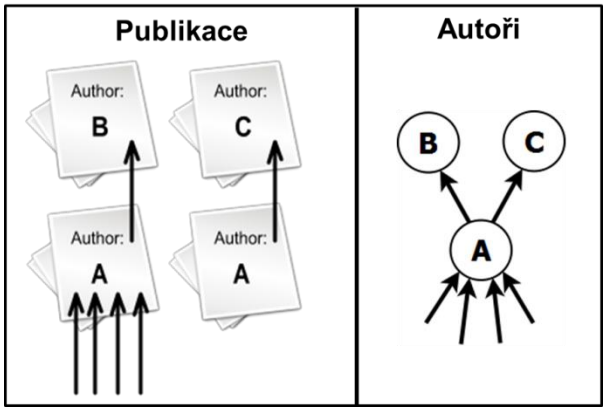
4.2 Datová kolekce, citační sítě a ocenění autoři

V této části je popsána datová kolekce WoS, způsoby vytváření citačních sítí a jejich charakteristiky a porovnání s citačními sítěmi vytvořenými z kolekcí CiteSeer a DBLP. Dále jsou zde popsány manuálně vytvořené seznamy držitelů významných ocenění v oblasti počítačových věd, které byly použity pro určení kvality našich metod pro hodnocení autorů.

4.2.1 ISI Web of Science a citační sítě

Všechny námi navržené metody pro hodnocení autorů mohou být aplikovány na citační sítě vytvořené z libovolné bibliografické kolekce. My jsme nyní použili kolekci WoS zakoupenou od Thomson Reuters, se kterou již pracoval Fiala (2012a, 2014). Tato kolekce obsahuje záznamy o všech článcích publikovaných v letech 1996 až 2005, které byly použity při výpočtu Journal Citation Reports 2009 a jsou v kategoriích počítačových věd databáze ISI Web of Science klasifikovány jako „časopisecký článek“. Obsaženy jsou články ze všech sedmi WoS kategorií počítačových věd[49], kterými jsou: Umělá inteligence, Kybernetika, Hardware a architektura, Informační systémy, Mezioborové aplikace, Softwarové inženýrství a Teorie a metody. Jednou z vlastností odborné komunikace v počítačových vědách je, že určité množství odborných výsledků je publikováno na předních konferencích a ne v časopisech. My ale konferenční články nezakoupili z důvodu jejich nepravidelného zařazování do databáze WoS.

Hodnoty významnosti autorů mohou být vypočítány přímo z jejich citační sítě nebo rozdělením hodnot publikací. Zde bychom měli zmínit, že citační síť publikací obsahuje informace o časové posloupnosti vydávání publikací a následnosti citací, zatímco síť autorů tyto informace pomíjí. Tento rozdíl může při vyhodnocení zvýhodnit některé autory, např. na obrázku 4.1 je v síti publikací autor B prestižnější než autor C, kdežto v síti autorů jsou oba autoři stejně prestižní. Z toho důvodu jsme předpokládali, že hodnocení autorů na základě vyhodnocení citační sítě publikací poskytne lepší pořadí autorů, než hodnocení autorů na základě vyhodnocení citační sítě autorů. Přestože v kolekcích CiteSeer a DBLP se tento předpoklad nepotvrdil, tak zde popsané experimenty s kolekcí WoS ukázaly, že byl správný.



Obrázek 4.1: Rozdíl v hodnocení prestiže autorů založeném na citační síti publikací nebo na citační síti autorů – autoři B a C jsou v síti autorů stejně prestižní, kdežto v síti publikací je prestižnější autor B.

[49] WoS Computer Science categories: Artificial Intelligence, Cybernetics, Hardware & Architecture, Information Systems, Interdisciplinary Applications, Software Engineering and Theory & Methods.

Využitím bibliografických záznamů z kolekce WoS jsme vytvořili stejné citační sítě publikací a autorů jako ve 3. kapitole, tj. použili jsme stejné způsoby penalizace samocitací autorů a určení vah hran. Zrekapitulovat můžeme, že varianta samocitací ***ALL*** používá rovnocenně všechny citace a je proto nejvíce benevolentní. Nejpřísnější varianta samocitací je ***NOT***, která odstraňuje citace mezi publikacemi, které mají alespoň jednoho společného autora. V citační síti autorů jsme navíc použili variantu ***PART***, která z citační sítě autorů vytvořené variantou *ALL* odstraňuje všechny smyčky. Varianty, které by samocitacím přidělovaly menší váhy, viz např. Yan et al. (2011), jsme nepoužili. Použité způsoby určení vah hran v síti autorů budou zrekapitulovány dále.

Pro popis charakteristických vlastností citačních sítí vytvořených z kolekce WoS jsme použili stejné značení jako v části 3.3. Počty vrcholů (publikací nebo autorů), hran (citací) a vrcholů jednotlivých typů v citačních sítích vytvořených z naší kolekce WoS shrnuje tabulka 4.1. Tyto kvantitativní údaje mohou být použity pro porovnání našeho experimentu s jinými experimenty. Pro úplnost můžeme uvést, že celá naše kolekce WoS obsahuje záznamy o 149 347 článcích z 386 časopisů obsažených ve WoS kategoriích počítačových věd. Publikace byly průměrně napsány 2,5 autory a mají průměrně 1,3 citací ve variantě *ALL* a 1 citaci ve variantě *NOT*. Autoři průměrně napsali 2,4 publikací a obdrželi průměrně 6,8 citací v *ALL*, 6,6 citací v *PART* a 5,4 citací v *NOT* variantě.

Tabulka 4.1: Kvantitativní údaje citačních sítí vytvořených z bibliografické kolekce WoS (1996-2005). (Obsaženy jsou záznamy o všech časopiseckých článcích ze všech kategorií počítačových věd databáze WoS, které byly použity při výpočtu JCR 2009.)

DB	Typ sítě	Vrcholů	Samoc.	Hran	Slepé	Citující	Necitované	Citované	Izolované
ISI WoS (1996-2005)	publikace	149347	ALL	191447	79571	69776	90901	58446	49774
			NOT	145372	92694	56653	103312	46035	64517
	autoři	157440	ALL	1062886	71354	86086	83146	74294	48333
			PART	1039339	71843	85597	83662	73778	48482
			NOT	852356	82170	75270	94094	63346	56907

Pokud porovnáme charakteristické vlastnosti citačních sítí vytvořených z kolekce WoS z let 1996 až 2005 (viz tabulka 4.1) a charakteristické vlastnosti sítí vytvořených z kolekcí CiteSeer 2005 a DBLP 2004 (viz tabulka 3.2 ve 3. kapitole), tak můžeme konstatovat, že ve variantě použití samocitací *ALL*:

- WoS obsahuje 4x méně publikací a 2,5x méně autorů než CiteSeer a 3x méně publikací a 2x méně autorů než DBLP.
- Ve WoS je téměř 40% publikací citováno, zatímco v CiteSeer je jich citováno jen cca 30% a v DBLP jsou citována pouhá 4% publikací.
- Ve Wos, stejně jako v CiteSeer, 47% publikací cituje jiné publikace. V DBLP pouhá 2% publikací citují jiné publikace.
- Izolovaných publikací obsahuje WoS 33%, zatímco CiteSeer 46% a DBLP 96%.
- Publikace ve WoS průměrně obdržely 1,3 citací, v CiteSeer obdržely 2,4 citací a v DBLP jen 0,2 citací.
- Autoři ve WoS napsali průměrně 2,4 publikace, v CiteSeer 4 publikace a v DBLP 3,4 publikace.

- Publikace mají podobný průměrný počet autorů, což je očekávatelné, protože nejvíce publikací bývá napsáno dvěma autory.

Z uvedeného porovnání kolekcí WoS, CiteSeer a DBLP lze usoudit, že kolekce WoS obsahuje z pohledu počtu indexovaných citací kvalitnější záznamy než kolekce DBLP. Tyto záznamy by také měly být kvalitnější než záznamy z kolekce CiteSeer, protože jsou do databáze vkládány manuálně a neobsahují chyby vzniklé automatickým indexováním. Citační sítě vytvořené z kolekce WoS mají také složitější strukturu, než sítě vytvořené z kolekcí CiteSeer a DBLP, protože jejich vrcholy jsou více provázány[50].

Nyní zrekapitulujeme způsoby určení vah hran v citační síti autorů a způsoby rozdělení hodnot publikací jejich autorům, které byly použity ve 3. kapitole. Ve variantě vah hran ***N*** váha hrany zastupuje počet citací, které autor získal v síti publikací (např. pokud autor A ve dvou publikacích citoval autora B, tak v síti autorů existuje hrana z A do B s váhou 2). Ve variantě vah hran ***1/N*** jsou hodnoty publikací rovnoměrně rozděleny mezi jejich reference (např. pokud autor A ve své publikaci citoval publikaci autorů B a C a současně citoval publikaci autora D, tak v citační síti autorů existují hrany z A do B a z A do C s váhami ½ a hrana z A do D s váhou 1 – souběžné hrany se následně sjednotí a jejich váhy sečtou). Ve variantě ***1*** je všem hranám sítě přiřazena váha 1.

Pro hodnocení autorů na základě hodnot jejich publikací, které jsme určili vyhodnocením citační sítě publikací, jsme použili rozdělení *SUM* a *DIV*. Rozdělení ***SUM***, viz vzorec (3.3) v části 3.2, sečte autorům celé hodnoty jejich publikací bez ohledu na počet autorů publikace. Rozdělení ***DIV***, viz vzorec (3.4) v části 3.2, sečte autorům rovnoměrné díly hodnot jejich publikací, které jsou pro každou publikaci určeny vydělením její hodnoty počtem jejích autorů – např. pokud má autor 2 spoluautory ve své první publikaci a 4 spoluautory v druhé publikaci, tak ve variantě *DIV* získá ⅓ z hodnoty první publikace a ⅕ z hodnoty druhé publikace.

Další varianty rozdělení hodnot publikací jejich autorům, které jsme v tomto experimentu nepoužili, mohou různým způsobem zohledňovat pozice autorů ve výčtu autorů, který je uveden v záznamu o publikaci, viz např. (Assimakis a Adam 2010). Tyto varianty rozdělení jsme použili až v experimentu, který je popsán v 5. kapitole. Jinou možnost využití výčtu autorů použil např. Zhao (2005), který při vytváření sítě autorů testoval použití pouze prvního autora publikace, prvních *K* autorů publikace či všech autorů publikace.

4.2.2 Seznamy oceněných autorů

Kvalitu pořadí autorů vytvořených našimi metodami pro hodnocení autorů jsme určili, podobně jako např. Sidiropoulos a Manolopoulos (2006) nebo Lin et al. (2013), jejich porovnáním na základě manuálně vytvořených seznamů významných autorů. Tyto seznamy významných autorů jsme stejně jako ve 3. kapitole vytvořili z držitelů Turingovy (*ACM A.M. Turing Award*) nebo Coddovy (*ACM SIGMOD E.F. Codd Innovations Award*) ceny a z osob uvedených na seznamu významných osob ACM (*ACM Fellows*) nebo na seznamu vysoce citovaných výzkumníků ISI (*ISI Highly Cited Researchers*).

[50] Citační síť publikací se všemi citacemi (*ALL*) vytvořená z kolekce CiteSeer má hustotu $3{,}7*10^{-6}$, síť vytvořená z kolekce DBLP má hustotu $4{,}6*10^{-7}$ a síť vytvořená z kolekce WoS má hustotu $8{,}6*10^{-6}$. Hustota je počítána jako $h/(n*(n-1))$, kde h je počet hran sítě a n je počet vrcholů sítě, viz poznámka pod čarou v části 2.3.3.

Ocenění autoři byli v kolekci WoS vyhledáni podle příjmení a prvních písmen křestních jmen (takto jsou jména autorů ve WoS uložena), ale odstranění víceznačnosti jmen autorů provedeno nebylo.

V naší kolekci WoS jsme nalezli 39 autorů oceněných *ACM Turing Award*, 15 autorů oceněných *ACM Codd Award*, 576 autorů ze seznamu *ACM Fellows* a 280 autorů ze seznamu *ISI Highly Cited Researchers*. Z těchto autorů jsme vytvořili odpovídající seznamy oceněných osob. Sjednocení těchto seznamů osob obsahuje 805 jmen autorů, kteří byli nalezeni v kolekci WoS. Počty shodných jmen autorů v jednotlivých seznamech oceněných autorů ukazuje tabulka 4.2.

Tabulka 4.2: Počty shodných jmen na seznamech oceněných autorů, které byly manuálně vytvořeny pro kolekci WoS.

Seznam oceněných autorů	ACM Turing	ACM Codd	ACM Fellows	ISI Highly Cited
ACM Turing	39	0	17	10
ACM Codd	0	15	8	5
ACM Fellows	17	8	576	74
ISI Highly Cited	10	5	74	280

Kvalitu metod pro hodnocení autorů, které budou popsány v následující části, jsme využitím manuálně vytvořených seznamů oceněných autorů určili tak, že jsme v získaných pořadích autorů (vytvořena dle vypočtených hodnot autorů) nalezli oceněné autory. Z pozic oceněných autorů byla vypočítána jejich průměrná pozici v daném pořadí. Průměrná pozice oceněných autorů v pořadí autorů udává kvalitu dané metody při vyzdvižení osob z daného seznamu oceněných autorů. Protože nejlepší autor je v pořadí autorů na 1. pozici, tak nejlepší metoda je ta s nejmenší průměrnou pozicí oceněných autorů ve vytvořeném pořadí autorů.

4.3 Výpočet popularity a prestiže

Pro hodnocení autorů na základě citační analýzy jsme použili stejné varianty PageRanku jako ve 3. kapitole, viz vzorce (3.1) a (3.2) v části 3.2, přičemž pro výpočet hodnot vrcholů bylo vždy použito 50 iterací. Varianty PageRanku hodnotí prestiž autorů. Abychom výsledky porovnali s méně důmyslnou metodou pro hodnocení významnosti vrcholů grafu, tak jsme navíc použili neiterační metodu in-degree (viz části 2.3.3), která určuje hodnoty vrcholů na základě součtu vah jejich vstupních hran, čímž hodnotí jejich popularitu. Výpočet in-degree je popsán vzorcem (4.1), kde *InD(A)* je hodnota in-degree vrcholu *A*, U_A je množina vrcholů, které odkazují na vrchol *A*, a $w_{u\text{to}A}$ je váha hrany vedoucí z vrcholu *u* do vrcholu *A*.

$$InD(A) = \sum_{u \in U_A} w_{u\text{to}A} \tag{4.1}$$

Vedle PageRanku bez personalizace, viz vzorec (3.1), jsme použili i PageRank s personalizací, viz vzorec (3.2), který trvale zvýhodňuje některé vrcholy grafu. PageRank bez personalizace je dále značen ***BezP.***. Autoři byli opět zvýhodněni na základě jejich produktivity – personalizace počtem publikací (značena ***P.P.P.***). Publikace byly zvýhodněny dle jejich kvality stanovené na základě počtu autorů, kteří se na vytvoření publikace podíleli – personalizace počtem autorů (značena ***P.P.A.***).

4.4 Diskuse výsledků vyhodnocení kolekce ISI Web of Science

Všechny námi použité metody pro hodnocení autorů jsou porovnány v tabulce 4.3, kde řádky zastupují jednotlivé metody s parametry:

- *síť* – autorů nebo publikací,
- *algoritmus* – in-degree nebo PageRank,
- *samocitace* – zahrnuty nebo vyloučeny (*ALL*, *PART* nebo *NOT* v síti autorů a *ALL* nebo *NOT* v síti publikací),
- *váhy* – váhy hran v síti autorů (*1*, *1/N* nebo *N*) nebo způsoby rozdělení hodnot publikací jejich autorům (*DIV* nebo *SUM*),
- *personalizace* – bez personalizace (*BezP.*), personalizace autorů počtem publikací (*P.P.P.*) nebo personalizace publikací počtem autorů (*P.P.A.*).

Každá hodnota ve sloupci ***průměr*** v tabulce 4.3 reprezentuje průměrnou pozici autorů oceněných daným oceněním v pořadí autorů, které vytvořila daná metoda. Protože nejlépe hodnocený autor je na první pozici ve vytvořeném pořadí autorů, tak nejnižší hodnota průměrné pozice oceněných autorů určuje naší nejlepší metodu pro hodnocení autorů a je navíc zobrazena v řádku ***Minimum*** (m_{best}). Pokud metody vzestupně seřadíme podle průměrných pozic autorů oceněných daným oceněním a v tomto pořadí očíslujeme (1-39), tak získáme pořadí úspěšnosti jednotlivých použitých metod, které je zobrazeno ve sloupci ***p.***. Opět platí, že čím menší hodnotu ve sloupci *p.* metoda má, tím je lepší. Procentuální odlišení průměrných pozic oceněných autorů od nejlepší dosažené průměrné pozice znázorňuje sloupec $\boldsymbol{m_{\%}}$. Nejlepší průměrné pozice oceněných autorů jsou pro jednotlivé sítě (autoři/publikace) a použité algoritmy hodnocení (in-degree/PageRank) zvýrazněny.

Naše hlavní otázka byla: lze vyhodnocením citační sítě publikací získat lepší pořadí autorů, než vyhodnocením citační sítě autorů? Průměrné pozice oceněných autorů (sloupce *průměr* v tabulce 4.3) mají metody využívající síť publikací obvykle lepší, než metody využívající síť autorů. Pro *Sjednocení* (sjednocení všech seznamů oceněných autorů) je průměrná pozice metod (ve sloupci *p.*) využívajících síť publikací 12,3, kdežto průměrná pozice metod využívajících síť autorů je 23,4.

Algoritmus PageRank také obvykle předčil jednodušší in-degree, což jsme očekávali, protože PageRank měří prestiž autorů, zatímco in-degree měří pouze jejich popularitu. Pro síť publikací je ve sloupci *Sjednocení* průměrná pozice metod využívajících PageRank 5,8, kdežto metody in-degree zde mají průměrnou pozici 25,3.

Z tabulky 4.3 je dále patrné, že lepších výsledků dosahují metody, které PageRankem vyhodnocují sítě publikací s odstraněnými samocitacemi autorů (*NOT*), nežli metody, které používají všechny citace (*ALL*). Pro *Sjednocení* je průměr pozic metod využívajících PageRank a citační síť publikací bez samocitací 4,3 a průměr pozic metod využívajících PageRank a všechny citace v síti publikací je 7,3. Dále můžeme vidět, že po vyhodnocení sítí publikací bez samocitací PageRankem je lepší autorům sčítat rovnoměrné díly z hodnot jejich publikací určené na základě počtu autorů publikace (*DIV*), nežli jim sčítat celé hodnoty jejich publikací (*SUM*). Pro rozdělení *DIV* je zde průměrná pozice metod ve sloupci *Sjednocení* 1,5, kdežto pro *SUM* je 7. Jedinou výjimku, pro kterou bylo lepší *SUM*, tvoří sloupec s Coddovou cenou. Důvody mohou být: 1) Coddova cena je udílena v oblasti databázových systémů, ale naše kolekce WoS obsahuje bibliografické záznamy z celé oblasti počítačových věd; 2) počet použitých oceněných autorů je velmi malý (pouze 15 osob).

Tabulka 4.3: Porovnání kvality našich metod při hodnocení autorů z kolekce WoS na základě seznamů významných autorů (nejlepší průměrné pozice oceněných autorů jsou pro jednotlivé citační sítě autorů či publikací a použité algoritmy in-degree či PageRank zvýrazněny).

Kolekce WoS (1996-2005) obsahuje 157440 odlišných autorů vědeckých publikací																							
Síť	**Alg.**	**Samocit.**	**Váhy**	**Person.**	**ACM Turing (39 osob)**			**ACM Codd (15 osob)**			**ACM Fellows (576 osob)**			**ISI Highly C. (280 osob)**			**Sjednocení (805 osob)**		
					průměr	p.	$m_{\%}$	průměr	p.	$m_{\%}$	průměr	p.	$m_{\%}$	průměr	p.	$m_{\%}$	průměr	p.	$m_{\%}$
Autoři	In-degree	NOT	1		41087	16	30%	39614	34	102%	30571	25	21%	26081	34	15%	31500	30	19%
			1/N		**38812**	**10**	**23%**	38573	31	97%	**30190**	**22**	**19%**	**25579**	**31**	**13%**	**30960**	**23**	**17%**
			N		41288	19	30%	40165	35	105%	30619	26	21%	26058	33	15%	31510	31	19%
		PART	1		44335	31	40%	**34488**	**16**	**76%**	31321	35	24%	26575	36	17%	32199	36	21%
			1/N		42069	24	33%	36222	21	85%	30862	29	22%	25993	32	15%	31627	32	19%
			N		45073	34	42%	35727	18	82%	31676	38	25%	26770	38	18%	32493	38	22%
		ALL	1		44555	33	41%	34987	17	79%	31346	36	24%	26632	37	18%	32240	37	21%
			1/N		42518	27	34%	36746	24	88%	31010	32	22%	26148	35	15%	31780	35	20%
			N*		45463	35	44%	36332	23	85%	31786	39	25%	26854	39	18%	32602	39	23%
	PageRank	NOT	1	BezP.	38442	7	21%	41533	38	112%	29837	19	18%	24255	19	7%	30409	19	15%
				P.P.P.	40180	11	27%	33231	12	70%	27014	7	7%	23570	10	4%	28243	8	6%
			1/N	BezP.	**37083**	**5**	**17%**	40397	37	106%	29723	18	17%	24252	18	7%	30268	18	14%
				P.P.P.	38502	8	22%	32060	11	64%	**26800**	**6**	**6%**	**23514**	**8**	**4%**	**28010**	**6**	**6%**
			N	BezP.	38418	6	21%	41833	39	114%	29860	20	18%	24227	17	7%	30416	20	15%
				P.P.P.	40213	12	27%	33437	13	71%	27015	8	7%	23517	9	4%	28225	7	6%
		PART	1	BezP.	41258	18	30%	36307	22	85%	30860	28	22%	24523	21	8%	31176	26	17%
				P.P.P.	42166	25	33%	**26035**	**4**	**33%**	28160	14	11%	23902	13	5%	29085	14	10%
			1/N	BezP.	40339	13	27%	38431	30	96%	30620	27	21%	24618	22	9%	31038	24	17%
				P.P.P.	40823	14	29%	29900	8	53%	27818	12	10%	23881	12	5%	28858	11	9%
			N	BezP.	41852	23	32%	37215	28	90%	31076	33	23%	24727	23	9%	31382	28	18%
				P.P.P.	42712	29	35%	26751	6	37%	28286	15	12%	24071	14	6%	29214	15	10%
		ALL	1	BezP.	41715	21	32%	37064	26	89%	31119	34	23%	24827	25	10%	31436	29	18%
				P.P.P.	42539	28	34%	26640	5	36%	28402	16	12%	24121	16	6%	29311	16	10%
			1/N	BezP.	40938	15	29%	39188	33	100%	30904	30	22%	24913	26	10%	31308	27	18%
				P.P.P.	41304	20	30%	30299	9	55%	28071	13	11%	24088	15	6%	29079	13	10%
			N	BezP.	42428	26	34%	38017	29	94%	31356	37	24%	25005	28	10%	31648	34	19%
				P.P.P.	43204	30	36%	27362	7	40%	28538	17	13%	24270	20	7%	29436	17	11%
Publikace	In-degree	NOT	DIV		**38643**	**9**	**22%**	38994	32	99%	**29968**	**21**	**18%**	**24771**	**24**	**9%**	**30568**	**21**	**15%**
			SUM		41100	17	30%	40318	36	106%	30223	23	19%	24919	27	10%	30887	22	16%
		ALL	DIV		41781	22	32%	37167	27	90%	30542	24	21%	25200	29	11%	31175	25	17%
			SUM		44336	32	40%	**36911**	**25**	**88%**	31007	31	22%	25477	30	12%	31643	33	19%
	PageRank	NOT	DIV	BezP.	**31680**	**1**	**0%**	34134	15	74%	25476	2	1%	22782	2	1%	26646	2	0%
				P.P.A.	32798	3	4%	33746	14	72%	**25341**	**1**	**0%**	**22665**	**1**	**0%**	**26544**	**1**	**0%**
			SUM	BezP.	47017	36	48%	30388	10	55%	26424	5	4%	22943	3	1%	27771	5	5%
				P.P.A.	52407	38	65%	**19589**	**1**	**0%**	27067	9	7%	23100	6	2%	28257	9	6%
		ALL	DIV	BezP.	32656	2	3%	36066	20	84%	25964	4	2%	23005	5	2%	27090	4	2%
				P.P.A.	34202	4	8%	35793	19	83%	25879	3	2%	22990	4	1%	27082	3	2%
			SUM	BezP.	48170	37	52%	25712	3	31%	27087	10	7%	23437	7	3%	28338	10	7%
				P.P.A.	53637	39	69%	21284	2	9%	27699	11	9%	23588	11	4%	28900	12	9%
Minimum (m_{best})					31680			19589			25341			22665			26544		

Nejlepší pořadí autorů bylo získáno použitím PageRanku s personalizací dle počtu autorů publikací (*P.P.A.*). Metoda, která nám poskytla nejlepší výsledky, je značena *Publikace-PageRank-NOT-DIV-P.P.A.*. Minimální odlišnost vykazují pouze porovnání s Turingovou a Coddovou cenou, přičemž tato odlišnost může být způsobena malým počtem oceněných autorů. Lze si všimnout, že použití

navržených personalizací poskytlo ve většině případů lepší hodnocení autorů než nepoužití personalizace (výjimku tvoří *ACM Turing* a některé metody využívající rozdělení *SUM*). Poznamenat také můžeme, že metoda *Autoři–In-degree–ALL–N**, která hodnotí autory na základě počtu citací, které obdržely jejich publikace, poskytovala obvykle nejhorší pořadí autorů.

V tabulkách 4.4 a 4.5 jsou detailněji porovnána pořadí autorů vytvořená metodami pro hodnocení autorů, které aplikující PageRank na citační síť publikací. K těmto metodám byla přidána metoda počítající autorům citace jejich publikací a nejlepší z metod pracujících s citační sítí autorů (*Autoři-PR-NOT-1/N-P.P.P.*). V obou tabulkách je zvýrazněno 24 nejvyšších hodnot. Tabulka 4.4 obsahuje Spearmanovy koeficienty pořadové korelace vynásobené stem, které na intervalu <+100; -100> udávají, nakolik jsou funkce určující pořadí autorů navzájem funkčně závislé. Vidět v ní lze např. menší korelace mezi pořadími autorů, která byla získána metodami vyhodnocujícími citační síť publikací, a pořadími, která poskytly metoda vyhodnocující síť autorů a metoda počítající autorům citace jejich publikací. Tabulka 4.5 ukazuje, kolik stejných jmen autorů obsahují na prvních/nejlepších 100 pozicích dvojice získaných pořadí. V rámci experimentu jsme také zjišťovali počty shodných jmen na prvních 1000 pozicích ve dvojicích získaných pořadí, přičemž výsledky byly srovnatelné s výsledky v tabulce 4.5. Hodnoty prezentované v tabulkách 4.4 a 4.5 ukazují, že nejvíce podobná pořadí autorů poskytují PageRank bez personalizace (*BezP.*) a PageRank s personalizací dle počtu autorů publikace (*P.P.A.*), tj. využití personalizace v tomto případě zlepšilo výsledky pouze minimálně. Dále lze vidět, že metoda počítající citace autorů (*Autoři–In-D.–ALL–N**) se od ostatních metod hodnocení odlišuje nejvíce.

Tabulka 4.4: Porovnání pořadí autorů vytvořených vybranými metodami pro hodnocení autorů na základě Spearmanových koeficientů pořadové korelace (koeficienty jsou vynásobeny stem a 24 nejvyšších hodnot je zvýrazněno).

Síť	Algoritmus	Samocit.	Váhy	Person.	Publikace PageRank								Autoři	
					NOT				ALL				In-D. ALL N*	PR NOT 1/N P.P.P.
					DIV		SUM		DIV		SUM			
					BezP.	P.P.A.	BezP.	P.P.A.	BezP.	P.P.A.	BezP.	P.P.A.		
Publikace	PageRank	NOT	DIV	BezP.	x	**100**	92	88	**99**	**98**	89	85	56	84
				P.P.A.	**100**	x	93	90	**99**	**98**	91	88	56	86
			SUM	BezP.	92	93	x	**99**	91	92	**98**	**96**	66	88
				P.P.A.	88	90	**99**	x	87	90	**97**	**98**	63	87
		ALL	DIV	BezP.	**99**	**99**	91	87	x	**100**	92	88	56	83
				P.P.A.	**98**	**98**	92	90	**100**	x	94	91	56	85
			SUM	BezP.	89	91	**98**	**97**	92	94	x	**99**	64	86
				P.P.A.	85	88	**96**	**98**	88	91	**99**	x	62	84
Aut.	In-D., ALL, N*				56	56	66	63	56	56	64	62	x	61
	PR, NOT, 1/N, P.P.P.				84	86	88	87	83	85	86	84	61	x

Tabulka 4.5: Porovnání vybraných metod pro hodnocení autorů dle počtu stejných jmen autorů na nejlepších 100 pozicích ve vytvořených pořadích (24 nejvyšších hodnot je zvýrazněno).

Síť	Algoritmus	Samocit.	Váhy	Person.	Publikace PageRank								Autoři	
					NOT				ALL				In-D. ALL N*	PR NOT 1/N P.P.P.
					DIV		SUM		DIV		SUM			
					BezP.	P.P.A.	BezP.	P.P.A.	BezP.	P.P.A.	BezP.	P.P.A.		
Publikace	PageRank	NOT	DIV	BezP.	x	**96**	69	67	**79**	**79**	66	62	47	51
				P.P.A.	**96**	x	72	70	**81**	**81**	70	66	50	52
			SUM	BezP.	69	72	x	**96**	65	67	**86**	**85**	56	40
				P.P.A.	67	70	**96**	x	63	65	**84**	**84**	55	39
		ALL	DIV	BezP.	**79**	**81**	65	63	x	**97**	69	65	47	40
				P.P.A.	**79**	**81**	67	65	**97**	x	72	68	50	41
			SUM	BezP.	66	70	**86**	**84**	69	72	x	**96**	57	37
				P.P.A.	62	66	**85**	**84**	65	68	**96**	x	58	36
Aut.	In-D., ALL, N*				47	50	56	55	47	50	57	58	x	41
	PR, NOT, 1/N, P.P.P.				51	52	40	39	40	41	37	36	41	x

Tabulka 4.6 zobrazuje 20 nejlepších pozic z pořadí autorů, která byla vytvořena pěti nejlepšími metodami pro hodnocení autorů. Autoři, kteří jsou v alespoň jednom sloupci na nejlepších třech pozicích, jsou zvýrazněni. Autoři, kteří nejsou mezi dvaceti nejlepšími autory v odpovídajícím sloupci, ale jsou mezi dvaceti nejlepšími autory v některém ze zbylých sloupců, jsou vypsáni ve spodní části tabulky. Jak je z tabulky 4.6 zřejmé, jména na nejlepších pozicích ve vybraných pořadích autorů se více liší jen ve sloupci *Publikace-PR-SUM-NOT-BezP.*. Nejlépe hodnoceno bylo jméno *Jain, AK*, které ve všech pěti zobrazených pořadích obsadilo vždy některou z prvních tří pozic. Zajímavá byla např. jména *Setiono, R* a *Dannenberg, RB*, která obsadila dobré pozice v metodách používajících všechny citace (*ALL*), ale mírně horší pozice v metodách odstraňujících samocitace (*NOT*). Vysvětlením může být časté používání samocitací samotnými autory nebo jejich spoluautory. Lepší pozice těchto autorů v metodě používající rovnoměrné rozdělení hodnot publikací (*DIV*), než v metodě používající celé hodnoty publikací (*SUM*), mohou navíc naznačovat, že autoři publikovali své články s pouze malým počtem spoluautorů.

Tabulka 4.6: Nejlepších 20 pozic v pořadích autorů vytvořených našimi pěti nejlepšími metodami pro hodnocení autorů (nejlepší tři autoři z libovolného sloupce jsou vždy zvýrazněni).

Pozice	Síť publikací vyhodnocena PageRankem				
	DIV				SUM
	NOT		ALL		NOT
	BezP.	P.P.A.	BezP.	P.P.A.	BezP.
1	Simon, DR	Simon, DR	Setiono, R	Setiono, R	Jain, AK
2	Breiman, L	Breiman, L	Jain, AK	Jain, AK	Vazirani, U
3	Jain, AK	Jain, AK	Yager, RR	Breiman, L	Bernstein, E
4	Moltenbrey, K	Yager, RR	Breiman, L	Yager, RR	Simon, DR
5	Yager, RR	Moltenbrey, K	Simon, DR	Simon, DR	Breiman, L
6	Robertson, B	Vazirani, U	Moltenbrey, K	Vazirani, U	Tanaka, K
7	Vazirani, U	Bernstein, E	Vazirani, U	Moltenbrey, K	Yager, RR
8	Bernstein, E	Zadeh, LA	Bernstein, E	Bernstein, E	Kim, J
9	Zadeh, LA	Robertson, B	Pedrycz, W	Pedrycz, W	Lee, J
10	Pedrycz, W	Pedrycz, W	Robertson, B	Zadeh, LA	Chang, CC
11	Amari, S	Hyvarinen, A	Zadeh, LA	Robertson, B	Lee, S
12	Hyvarinen, A	Amari, S	Amari, S	Amari, S	Pedrycz, W
13	Chang, CC	Oja, E	Wang, J	Wang, J	Wang, J
14	Oja, E	Chang, CC	Hyvarinen, A	Hyvarinen, A	Osher, S
15	Wang, J	Tanaka, K	Oja, E	Oja, E	Kim, JH
16	Tanaka, K	Wang, J	Chang, CC	Lee, J	Wang, Y
17	Lee, J	Burges, CJC	Lee, J	Chang, CC	Moltenbrey, K
18	Burges, CJC	Lee, J	Tanaka, K	Tanaka, K	Bennett, CH
19	Lee, S	Lee, S	Egghe, L	Lee, S	Oja, E
20	Zhang, J	Kim, J	Dannenberg, RB	Picard, RW	Wang, HO
	(21) Kim, J	(21) Zhang, J	(22) Lee, S	(22) Dannen..	(24) Zhang, J
	(26) Kim, JH	(25) Kim, JH	(23) Picard, RW	(26) Kim, JH	(28) Amari, S
	(32) Picard, ..	(31) Wang, Y	(27) Zhang, J	(27) Egghe, L	(34) Picard, ..
	(33) Wang, Y	(33) Picard, ..	(28) Kim, JH	(28) Zhang, J	(41) Hyvarin..
	(41) Egghe, L	(52) Egghe, L	(31) Kim, J	(31) Kim, J	(46) Robert..
	(56) Wang, ..	(53) Wang, ..	(35) Burges, CJC	(32) Burges, ..	(57) Zadeh, ..
	(59) Osher, S	(56) Osher, S	(36) Wang, Y	(36) Wang, Y	(128) Burge..
	(68) Setiono, ..	(75) Setiono, ..	(52) Osher, S	(41) Osher, S	(172) Setion..
	(85) Bennett..	(92) Bennett..	(65) Wang, HO	(64) Wang, ..	(204) Egghe, ..
	(1401) Dann..	(1585) Dann..	(100) Bennett, ..	(106) Bennet..	(3948) Dann..

4.5 Shrnutí závěrů z hodnocení autorů z kolekce WoS

V této kapitole jsme metodami pro hodnocení autorů vyhodnocovali citační sítě vytvořené z kolekce WoS (oblast počítačových věd), která je dle našeho názoru kvalitnější, co se indexovaných údajů týká, než kolekce CiteSeer a DBLP. Proto výsledky, které jsme v této kapitole prezentovali, považujeme za věrohodnější. V rámci experimentu jsme testovali, zda lepší pořadí autorů získáme vyhodnocením citační sítě publikací nebo vyhodnocením citační sítě autorů, která pomíjí některé informace. Dále jsme testovali, jak je výpočet hodnot významnosti autorů ovlivněn použitým typem vah hran v síti

autorů nebo způsobem rozdělení hodnot publikací jejich autorům. Navíc jsme naše metody hodnotící prestiž autorů porovnali s metodami pro hodnocení popularity, abychom potvrdili, že prestiž je pro hodnocení autorů lepší než popularita.

Naší nejlepší metodou pro hodnocení autorů v kolekci WoS byla metoda, která aplikuje PageRank na citační síť publikací s odstraněnými samocitacemi autorů a následně rovnoměrně rozděluje PageRankové hodnoty publikací jejich autorům. Pokud je současně použita personalizace publikací počtem jejich autorů, tak se výsledky hodnocení autorů ještě mírně zlepší. Na základě zde popsaného experimentu tedy můžeme pro hodnocení autorů, které využívá kolekci WoS, doporučit metodu značenou *Publikace-PageRank-NOT-DIV-P.P.A.*. Zatímco v kolekcích CiteSeer a DBLP nebyla vhodnost použití navržených úprav personalizace PageRanku jednoznačně prokázána, tak v kolekci WoS personalizace ve většině případů zlepšila hodnocení autorů. Také se zde nově ukázalo vhodnější hodnotit autory na základě hodnot jejich publikací, nežli na základě citační sítě autorů, která pomíjí informaci o časovém sledu vydávání publikací. Dále se potvrdila vhodnost odstranění samocitací autorů a rovnoměrného rozdělení hodnoty publikace mezi její autory. Pro úplnost můžeme uvést, že metody, které byly založeny na in-degree (tj. hodnotily popularitu), poskytovaly, v porovnání s naší nejlepší metodou využívající PageRank (tj. hodnotící prestiž), značně horší pořadí autorů (ve sloupci *Sjednocení* průměrně o 19% horší).

Protože použití kvality publikací při výpočtu prestiže autorů zlepšilo hodnocení autorů, tak v našem dalším experimentu s kolekcí WoS, který je popsán v 5. kapitole, jsme do metod pro hodnocení autorů na základě hodnot jejich publikací integrovali hodnoty Impact Factoru a PageRanku časopisů. Tím jsme chtěli ukázat, že využití hodnot významnosti časopisů může zlepšit kvalitu metod pro hodnocení autorů. Protože rovnoměrné rozdělení hodnot publikací jejich autorům nám v této kapitole poskytlo lepší pořadí autorů, než když byly autorům sčítány celé hodnoty jejich publikací, tak neméně zajímavé se nám v dalším experimentu jevilo testování nerovnoměrných rozdělení hodnot publikací jejich autorům. Tato rozdělení zvýhodňují autory, kteří jsou ve výčtu autorů publikace na předních pozicích.

5 Varianty personalizace PageRanku pro hodnocení autorů

Vyhodnocení kolekce WoS, které bylo popsáno ve 4. kapitole, potvrdilo, že při hodnocení významnosti autorů je vhodné použít navržené úpravy PageRanku, kterými byly personalizace autorů počtem publikací a personalizace publikací počtem autorů. Také se potvrdil náš předpoklad, že použitím citační sítě publikací lze hodnotit autory lépe, než použitím citační sítě autorů, která pomíjí některé informace. Dále jsme ověřili, že PageRank, který hodnotí prestiž autorů, poskytuje lepší hodnocení autorů, než pouhé počítání citací, které hodnotí popularitu autorů. Potvrzeno také bylo, že samocitace autorů je nejlepší odstranit na úrovni publikací. Za naší obecně nejlepší metodu jsme po předchozích experimentech s kolekcí WoS prohlásili metodu, která aplikuje PageRank s personalizací dle počtu autorů publikace (tj. využívá kvalitu publikací) na citační síť publikací s odstraněnými samocitacemi autorů a následně hodnoty publikací rovnoměrně rozděluje jejich autorům.

Protože se v kolekci WoS při hodnocení autorů potvrdila vhodnost použití námi navržených personalizací PageRanku, tak jsme při dalším experimentu, který je popsán v (Nykl et al. 2015) a v této kapitole, navrhli a testovali různé další personalizace, které by mohly hodnocení autorů ještě více zlepšit. Pro experimentální ověření kvality navržených metod jsme opět použili kolekci WoS a některé seznamy oceněných autorů. Navržené metody byly navíc testovány při hodnocení autorů ve specializovaných kategoriích počítačových věd, přičemž naším cílem bylo ověřit jejich kvalitu při změně rozsahu vyhodnocované vědní oblasti. Také jsme prověřili možnost predikovat laureáty významných ocenění. Protože se ve 4. kapitole ukázalo vhodnější sčítat autorům rovnoměrné díly z hodnot publikací, nežli celé hodnoty publikací, tak jsme nově zkusili rozdělovat hodnoty publikací autorům nerovnoměrně. Naším cílem bylo zjistit, zda má smysl při hodnocení zvýhodňovat autory, kteří jsou ve výčtu autorů publikace na předních pozicích.

V následující části 5.1 je zmíněna spojitost mezi naším a předchozími vlastními i cizími experimenty a popsány motivace, které nás vedly k vytvoření nových metod pro hodnocení autorů. V části 5.2 je popsán postup vyextrahování zvolených kategorií z kolekce WoS a ukázány charakteristické údaje vytvořených citačních sítí. Dále je zde zmíněn postup vytvoření referenčních seznamů významných autorů, které odpovídají zvoleným kategoriím. Část 5.3 obsahuje popis nově navržených metod a shrnutí všech metod, které jsme pro hodnocení autorů použili. Jsou zde zrekapitulovány informace o PageRanku a použitých způsobech vytváření citačních sítí, dále jsou zde popsány nově navržené personalizace PageRanku a způsoby rozdělení hodnot publikací jejich autorům. Popsán je i postup získání hodnot významnosti časopisů a jejich použití při hodnocení autorů. Výsledky provedeného experimentu jsou diskutovány v části 5.4 a vyvozené závěry a doporučení shrnuty v části 5.5.

5.1 Návaznost na předchozí experimenty

Inspirováni skutečnostmi, že použití personalizace PageRanku zlepšuje hodnocení autorů a modifikace personalizace umožňuje vytvoření mnoha dalších metod pro hodnocení autorů, které dosud nebyly zkoumány, jsme navrhli nové varianty personalizací. Naším cílem bylo na základě referenčních seznamů oceněných autorů porovnat různé druhy personalizací a zjistit, která z nich poskytuje nejlepší pořadí autorů. Za nejvíce zajímavé úpravy personalizace jsme považovali doplnění personalizace o významnost časopisů (1), h-index (2) nebo počet citací (3). Další prezentované metody byly použity pro porovnání nově vytvořených metod s metodami, které jsme použili dříve. Kromě různých variant personalizace jsme také chtěli otestovat nerovnoměrná rozdělení hodnot

publikací jejich autorům (4), pro která nás inspirovali Assimakis a Adam (2010). Naše nejzajímavější metody pro hodnocení autorů a předpoklady, které nás vedly k jejich navržení, byly:

1) Použití hodnot významnosti časopisů, ve kterých byly publikace vytištěny, v personalizaci publikací. Naším předpokladem bylo, že úspěšně absolvované recenzní řízení věhlasného časopisu odráží kvalitu publikovaného článku. Z toho důvodu jsme očekávali, že použití hodnot časopisů poskytne, v porovnání se všemi námi testovanými metodami pro hodnocení autorů, nejlepší pořadí autorů.
2) Použití h-indexu (Hirsch 2005) v personalizaci autorů, protože h-index je jednou z nejznámějších neiteračních metod pro hodnocení autorů. Cílem tohoto postupu bylo zjistit, zda použití vyspělosti autorů (h-index) při výpočtu jejich prestiže poskytne lepší pořadí autorů, než použití jejich produktivity (počet publikací).
3) Použití počtu citací v personalizaci publikací, protože počet citací svědčí o popularitě publikací. Podobný postup použila (Ding 2011a), která ale vyhodnocovala pouze citační sítě vytvořené z autorů, kteří jsou v publikaci uvedeni na první pozici.
4) Použití nerovnoměrných rozdělení hodnot publikací jejich autorům. Tato rozdělení jsme testovali proto, že náš předchozí výzkum (viz 4. kapitola) ukázal, že rozdělení hodnot publikací jejich autorům zlepšuje kvalitu vytvořených pořadí autorů.

V experimentech s celou datovou kolekcí jsme provedli několik hodnocení autorů z celé oblasti počítačových věd (tato hodnocení dále nazýváme „globální hodnocení autorů v počítačových vědách“ nebo jen „globální hodnocení autorů“) a vytvořená pořadí autorů jsme porovnali na základě seznamů držitelů významného ocenění v oblasti počítačových věd (zvolena byla ocenění *ACM Fellows* a *ISI Highly Cited Researchers*). Alternativou k použití oceněných autorů by bylo použití členů programových výborů vědeckých konferencí, viz např. (Liu et al. 2005), nebo členů editorských rad vědeckých časopisů, viz např. (Fiala et al. 2015). Pro určení kvality navržených metod jsme se snažili použít pokud možno větší referenční seznamy významných autorů, protože lze předpokládat, že použití seznamů s malým počtem významných autorů, viz např. (Sidiropoulos a Manolopoulos 2006; Ding 2011a; Fiala 2012b; Fiala et al. 2015), může vést ke zkreslení výsledků. Abychom otestovali platnost našich zjištění ve specifičtějších oblastech výzkumu, tak jsme nově vyzkoušeli hodnotit autory v kategoriích *Umělá inteligence* a *Hardware*. Tyto kategorie jsme zvolili z důvodu jejich jasného vyhranění a dobré korespondence kategorií počítačových věd v databázi *ISI Web of Science* s kategoriemi *Special Interest Group*s (SIGs) v *Association for Computing Machinery* (ACM). SIGs byly zvoleny proto, že poskytují dostatečně obsáhlé seznamy oceněných osob. Přiřazení kategorií jednotlivým časopisům z naší kolekce WoS jsme provedli na základě *Journal Citation Reports* (JCR), stejně jako Fiala et al. (2015). Jinou možností by bylo vyhledání konkrétních publikací na základě dotazu (Haveliwala 2003; Ding 2011b), ale tento postup je složitější. Hodnocení autorů v kategoriích může být prospěšné při vyhledávání expertů se zvolenou specializací.

Na závěr této části musíme poznamenat, že v recenzním řízení při vytváření článku (Nykl et al. 2015) byla jedním recenzentem vznesena opodstatněná námitka, že porovnávat pořadí autorů, která vznikla citační analýzou naší kolekce WoS, na základě seznamu *ISI Highly Cited*, který také vznikl citační analýzou kolekce WoS, je cyklické a tudíž nevhodné vyhodnocení. Proto bylo porovnání našich výsledků se seznam *ISI Highly Cited* z článku (Nykl et al. 2015) odstraněno. Protože ale závěry získané využitím *ISI Highly Cited* nevykazují, v porovnání s *ACM Fellows*, přílišné odchylky, tak je v této

kapitole pro názornost v některých částech zmíníme také, ale bude jim věnována minimální pozornost.

5.2 Zvolená data

V části 5.2.1 je uveden detailní popis naší kolekce WoS a zvláště pak jejích dvou kategorií, které byly použity pro testování navržených metod ve specializovaných oblastech výzkumu. Zmíněné informace mohou být použity jednak pro porovnání naší kolekce a jejích částí s jinými kolekcemi (např. z jiných oblastí výzkumu), ale i pro porovnání výsledků budoucích experimentů s našimi výsledky. Kvalita navržených metod pro hodnocení autorů byla určena vyhledáním oceněných autorů v získaných pořadích autorů. Popis manuálního vytvoření seznamů držitelů prestižního ocenění je popsán v části 5.2.2.

5.2.1 Datová kolekce ISI Web of Science a zvolené kategorie

Pro experimenty s našimi metodami pro hodnocení autorů jsme použili stejnou kolekci WoS z let 1996 až 2005 jako ve 4. kapitole. Tato zakoupená kolekce obsahuje záznamy o všech publikacích, které jsou v kategoriích počítačových věd databáze ISI Web of Science klasifikovány jako časopisecký článek. Zakoupenou kolekci WoS považujeme za kvalitnější, než datové kolekce získané vlastními silami, viz např. (Ding et al. 2009; Fiala 2012b; Fiala et al. 2015), které obvykle nejsou kompletní nebo mohou být odvozeny pouze od jednoho konkrétního článku, viz např. (Li a Willett 2009).

Celou kolekci WoS obsahující 386 časopisů se 149 347 publikacemi od 157 440 různých autorů jsme použili pro experimenty s globálním hodnocením autorů v počítačových vědách. Při volbě specializovaných kategorií, ve kterých jsme také testovali kvalitu našich metod pro hodnocení autorů, jsme se zaměřili na kategorie WoS, které korespondují kategoriím speciálně zaměřených skupin ACM (*Special Interest Groups of ACM*; dále jen SIGs). SIGs jsme zvolili proto, že jednotlivé SIGs udílejí autorům významná ocenění ve svých oblastech zájmu a my tato ocenění použili pro určení kvality vytvořených pořadí autorů. Výčet ACM kategorií SIGs a WoS kategorií počítačových věd obsahuje následující seznam (porovnatelné kategorie SIGs a WoS jsou označeny kurzívou):

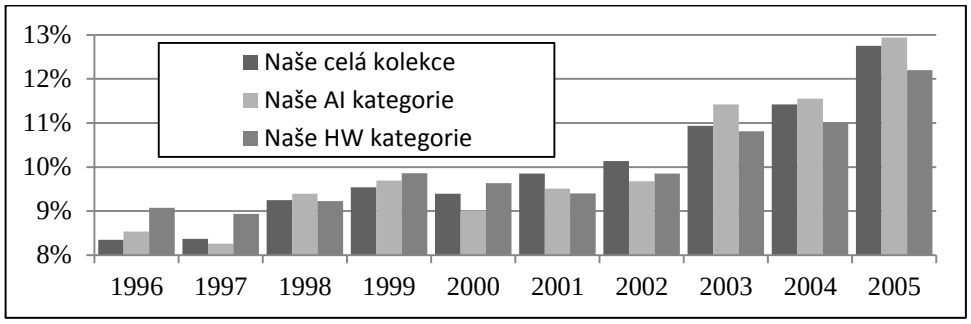
- ACM kategorie SIGs[51]: *Umělá inteligence*, *Aplikace*, Digitální obsah, Vzdělávání, *Návrh hardwaru*, Interakce, Počítačové sítě, *Software*, Management a řízení provozu, Výkonnost, *Teorie*.
- WoS kategorie počítačových věd[52]: *Umělá inteligence*, Kybernetika, *Hardware a architektura*, Informační systémy, *Mezioborové aplikace*, *Softwarové inženýrství*, *Teorie a metody*.

Pro experimenty jsme použili pouze kategorie *Umělá inteligence* (dále značena *AI*) a *Hardware a architektura* (dále jen *Hardware* či značení *HW*), protože kategorie *Mezioborové aplikace*, *Softwarové inženýrství* a *Teorie a metody* mají širokou oblast zájmu, která se často prolíná i do ostatních oblastí.

[51] ACM categories of SIGs: *Artificial Intelligence*, *Applications*, Digital Content, Education, *Hardware Design*, Interaction, Networking, *Software*, Ops & Management, Performance, and *Theory*.
Web ACM SIGs - http://www.acm.org/sigs

[52] WoS Computer Science categories: *Artificial Intelligence*, Cybernetics, *Hardware & Architecture*, Information Systems, *Interdisciplinary Applications*, *Software Engineering*, and *Theory & Methods*.

JCR 2013 na webu[53] WoS obsahovalo v kategorii *Umělá inteligence* 121 časopisů, kterým odpovídá 84 časopisů v naší kolekci WoS. To znamená, že JCR 2013 obsahuje o 37 časopisů z oblasti *Umělá inteligence* více než JCR 2009 (některé časopisy mohly být z indexování odebrány, ale jiné do něj byly přidány). Těchto 84 časopisů nalezených v naší kolekci WoS obsahuje 31 749 publikací od 39 891 různých autorů. V kategorii *Hardware* je v JCR 2013 uvedeno 50 časopisů, kterým v naší kolekci WoS odpovídá 40 časopisů s 18 917 publikacemi od 29 243 různých autorů. Jediný časopis, který je v naší kolekci v obou kategoriích je *IEEE Transaction on Neural Networks and Learning Systems*. Rozdělení publikací z naší kolekce WoS dle roku vydání znázorňuje graf na obrázku 5.1. Z grafu lze vypozorovat, že objem vědecké produktivity v počítačových vědách a/nebo její indexace ve WoS se za sledované desetiletí zvýšily více než o třetinu.



Obrázek 5.1: Zastoupení roků publikování článků v naší kolekci WoS.

Z bibliografických záznamů kolekce WoS jsme stejnými postupy jako v části 3.1 vytvořili citační sítě publikací (tj. článků) a autorů a navíc i citační sítě časopisů. Zopakovat můžeme, že varianta samocitací *ALL* v citační síti používá rovnocenně všechny citace; varianta *PART*, kterou jsme použili pouze pro citační sítě autorů, ze sítí vytvořených variantou *ALL* odstraňuje smyčky vrcholů; a varianta *NOT* odstraňuje citace mezi publikacemi, které mají alespoň jednoho společného autora. Jinou možností by bylo použít menší váhu pro samocitační hrany, s čímž experimentovali např. Yan et al. (2010), ale my tuto variantu nepoužili. Také jsme nevytvářeli sítě časopisů ze specializovaných kategorií.

Porovnat citační sítě vytvořené z celé kolekce WoS nebo ze zvolených kategorií lze na základě kvantitativních údajů, které pro vytvořené sítě považujeme za charakteristické. Těmito údaji jsou počty vrcholů (publikací, autorů nebo časopisů), hran (citací) a vrcholů jednotlivých typů (popis viz část 3.3), které shrnuje tabulka 5.1. Za povšimnutí stojí, že sítě z kategorie *Hardware* mají méně citací a polovina jejich vrcholů je izolována, kdežto v sítích vytvořených z kategorie *Umělá inteligence* nebo z celé kolekce WoS je izolována „pouze“ třetina vrcholů. To nás vede k domněnce, že výsledky, které ze sítí z kategorie *Hardware* získáme, mohou vykazovat mírné odchylky a nemusí se shodovat s výsledky získanými z kategorie *Umělá inteligence* nebo z celé kolekce WoS.

Průměrně každý časopis naší kolekce WoS vydal za zkoumaných 10 let 387 publikací od 727 různých autorů. *AI* časopisy průměrně vydaly 378 publikací (*HW* časopisy 473) od 722 různých autorů (*HW* časopisy od 948 autorů). Je zřejmé, že časopisy z kategorie *HW* obsahují vyšší než průměrný počet

[53] Web databáze ISI Web of Science - http://www.webofknowledge.com (pozn.: Web of Science je také znám jako Web of Knowledge).

publikací a autorů, v porovnání s ostatními časopisy z kolekce WoS, kdežto časopisy z kategorie *AI* obsahují publikací a autorů průměrný počet. Publikace byly napsány obecně podobným průměrným počtem autorů (2,5 v celé kolekci a kategorii *AI* a 2,7 v kategorii *HW*), ale publikace z kategorie *HW* mají podprůměrný počet citací (1,3 je průměrný počet citací článků v celé kolekci, 1,5 v *AI*, ale pouze 0,6 v *HW*). Menší počet citací v kategorii *HW* je zřejmě důsledkem obecně menšího trendu citování v této kategorii anebo malého počtu zastoupených publikací. Pro úplnost, průměrně každý autor z celé naší kolekce WoS napsal 2,4 publikace (autor v *AI* průměrně 2 publikace a autor v *HW* 1,8 publikací) a má 6,8 citací (v *AI* mají autoři průměrně 5,6 citací a v HW 3,2 citací), což by odpovídalo již zmíněným dedukcím.

Tabulka 5.1: Charakteristické kvantitativní vlastnosti citačních sítí vytvořených z kolekce WoS a jejích kategorií Umělá inteligence a Hardware.

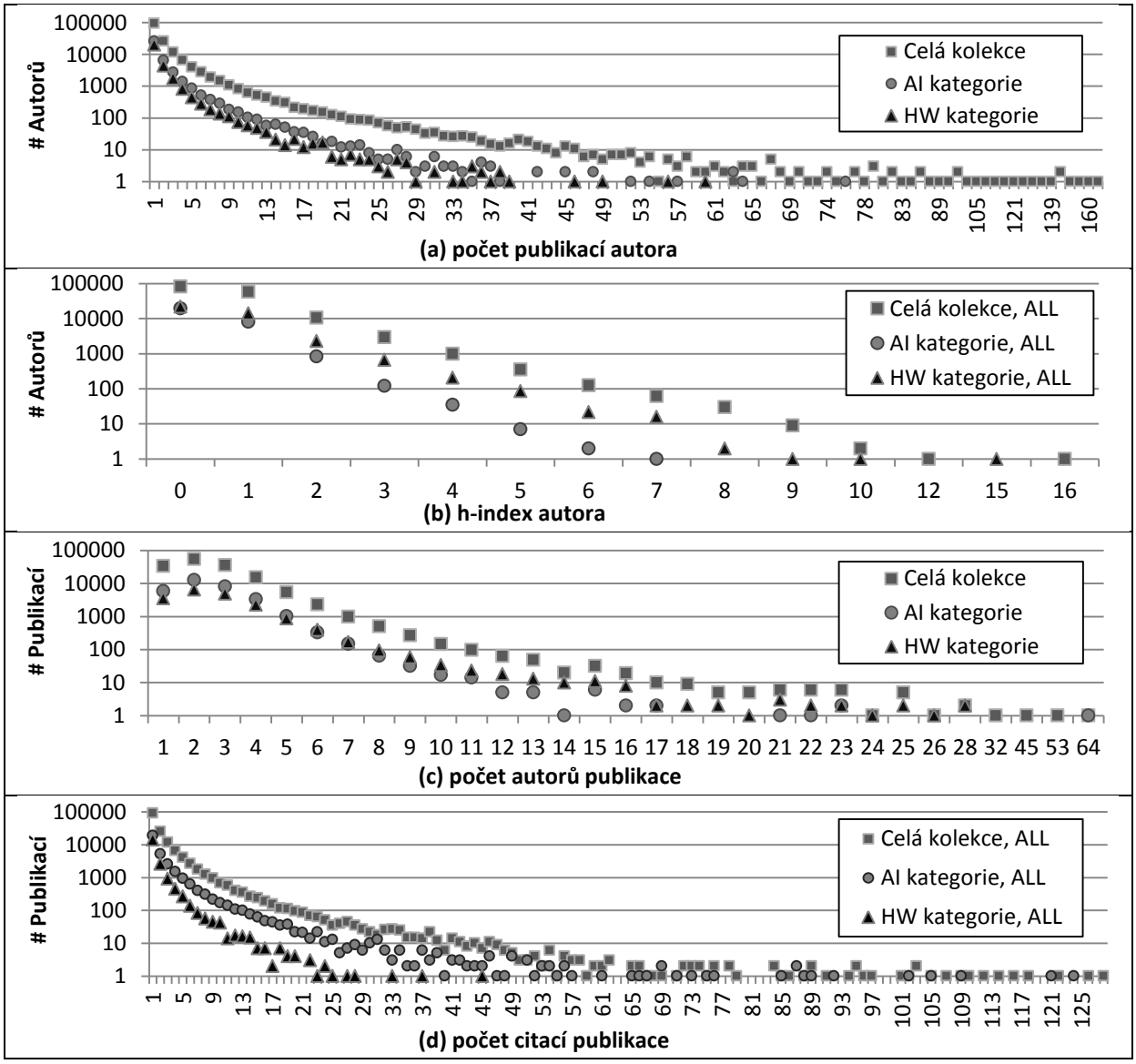
Kateg.	Typ sítě	Vrcholů	Samoc.	Hran	Slepé	Citující	Necitované	Citované	Izolované
Všechny WoS kategorie počítačových věd	Publikace	149347	ALL	191447	79571	69776	90901	58446	49774
			NOT	145372	92694	56653	103312	46035	64517
	Autoři	157440	ALL	1062886	71354	86086	83146	74294	48333
			PART	1039339	71843	85597	83662	73778	48482
			NOT	852356	82170	75270	94094	63346	56907
	Časopisy	386	ALL	20488	3	383	21	365	2
			NOT	20135	3	383	26	360	2
WoS kategorie Umělá inteligence	Publikace	31749	ALL	46203	15946	15803	18866	12883	9767
			NOT	36381	18428	13321	21252	10497	12394
	Autoři	39891	ALL	226898	18648	21243	22278	17613	12882
			PART	221713	18738	21153	22386	17505	12912
			NOT	188461	20961	18930	24669	15222	14748
WoS kategorie Hardware a architektura	Publikace	18917	ALL	10765	13200	5717	14159	4758	10064
			NOT	7967	14460	4457	15350	3567	11834
	Autoři	29243	ALL	94663	18384	10859	19955	9288	14330
			PART	92278	18439	10804	20011	9232	14350
			NOT	68958	20094	9149	21793	7450	16008

Protože naším cílem bylo testovat, jaký vliv na vytvoření pořadí autorů mají různé personalizace PageRanku, tak nás zajímalo, jak jsou autoři a publikace dle zvolených hodnot kvality členěni do skupin. Autory jsme personalizovali počtem jejich publikací a h-indexem a publikace počtem jejich autorů a počtem citací. Rozdělení autorů a publikací do skupin dle těchto hodnot demonstrují grafy na obrázku 5.2, které mají na vodorovné ose hodnoty dané metody a na svislé ose příslušný počet autorů nebo publikací s danou hodnotou (prázdné množiny jsou vynechány). Detailní popis vytvořených skupin autorů a publikací obsahuje následující výčet:

a) Počet publikací rozdělil autory do 101 skupin (v *AI* do 47 a v *HW* do 41) s nejvyšší hodnotou 181 (v *AI* 76 a v *HW* 60).
b) H-index se všemi citacemi autory rozdělil do 13 skupin (v *AI* do 12 a v *HW* do 8) s nejvyšší hodnotou 16 (v *AI* 15 a v *HW* 7). H-index bez samocitací autory rozdělil také do 13 skupin (v *AI* do 11 a v *HW* do 7) s nejvyšší hodnotou 16 (v *AI* 14 a v *HW* 6).

c) Počet autorů publikace rozdělil do 31 skupin (v *AI* do 21 a v *HW* do 27) s nejvyšší hodnotou 64 (v *AI* 64 a v *HW* 28).
d) Počet citací rozdělil publikace do 99 skupin (v *AI* do 75 a v *HW* do 29) s nejvyšší hodnotou 262 (v *AI* 189 a v *HW* 44). Pokud se odstranily samocitace autorů, tak byly publikace počtem citací rozděleny do 98 skupin (v *AI* do 69 a v *HW* do 27) s nejvyšší hodnotou 258 (v *AI* 185 a v *HW* 43).

Na obrázku 5.2 lze vidět, že pouze počet autorů publikace má mírně odlišný průběh rozdělení publikací do skupin, protože nejvíce publikací bylo napsáno dvěma autory. Počet publikací autora a počet citací publikace mají velmi podobný průběh rozdělení autorů resp. publikací do skupin a vytvořily také podobný počet skupin. O h-indexu lze říci, že má malou rozlišovací schopnost, protože autoři byli rozděleni maximálně do 13 skupin.



Obrázek 5.2: Velikosti skupin autorů a publikací vytvořených dle zvolených metod (všechny grafy mají logaritmické měřítko a nezobrazují prázdné množiny).

5.2.2 Referenční seznamy prestižních autorů

Strojově vytvořená pořadí autorů jsme opět porovnali na základě referenčních seznamů, které jsme manuálně vytvořili z držitelů prestižních ocenění v oblasti počítačových věd. V těchto seznamech jsou autoři reprezentováni svým příjmením a iniciály křestních jmen, stejně jako v kolekci WoS. Odstranění víceznačnosti jmen provedeno nebylo, pouze jsme odstranili nekompletní jména. Stejné vyhledání oceněných autorů v kolekci WoS bylo použito ve 4. kapitole a ve (Fiala 2012b; Fiala et al. 2015).

Experimenty s globálním hodnocením autorů v počítačových vědách byly vyhodnoceny na základě seznamu významných osob ACM (*ACM Fellows*[54]) a seznamu vysoce citovaných výzkumníků ISI (*ISI Highly Cited Researchers*[55]), které jsme použili už ve 4. kapitole. V naší kolekci WoS jsme nalezli 576 autorů uvedených na seznamu *ACM Fellows* a 280 autorů uvedených na seznamu *ISI Highly Cited*. Tyto seznamy jsou stejné jako ve 4. kapitole, což umožňuje porovnání výsledků.

Abychom mohli vyhodnotit experimenty s hodnocením autorů ve specializovaných kategoriích WoS, tak jsme pro zvolené ACM kategorie *Umělá inteligence* a *Návrh hardwaru* manuálně nalezli osoby oceněné příslušnými SIGs. Ve zvolených kategoriích ACM jsou následující SIGs:

- SIGs v ACM kategorii Umělá inteligence: Artificial Intelligence (SIGAI), Electronic Commerce (SIGecom), Genetic and Evolutionary Computation (SIGEVO), Information Retrieval (SIGIR), Knowledge Discovery in Data (SIGKDD), Hypertext and the Web (SIGWEB).
- SIGs v ACM kategorii Návrh hardwaru: Computer Architecture (SIGARCH), Embedded Systems (SIGBED), Design Automation (SIGDA), Mobility of Systems, Users, Data and Computing (SIGMOBILE), Microarchitecture (SIGMICRO).

V tabulce 5.2 jsou uvedeny názvy ocenění, která udílejí jednotlivé SIGs, a počty oceněných osob. Ceny jsou udíleny za dlouhotrvající nebo mimořádný přínos v oblasti zájmu dané SIG, za současnou práci, nejlepší článek nebo test-of-time článek (udíleno 10 nebo více let zpětně). Ceny za studentské práce použity nebyly. Sloupec *Oceněných* v tabulce 5.2 obsahuje počty různých jmen autorů, která jsme nalezli na webech daných ocenění. Kolik z těchto autorů jsme nalezli v celé naší kolekci WoS ukazuje sloupec *V celé WoS* a kolik z nich bylo v odpovídající kategorii naší kolekce WoS ukazuje sloupec *V kategorii*.

Celkově jsme v oceněních od SIGs z kategorie *Umělá inteligence* nalezli 354 různých jmen autorů, 224 z nich bylo rozpoznáno v naší kolekci WoS a 133 v odpovídající WoS kategorii. V oceněních od SIGs z kategorie *Návrh hardwaru* bylo nalezeno 158 různých jmen autorů, 85 z nich bylo rozpoznáno v naší kolekci WoS a 76 v odpovídající WoS kategorii. Pro vyhodnocení našich experimentů s hodnocením autorů ve zvolených kategoriích byla použita jména, která jsme nalezli v odpovídajících WoS kategoriích. Tabulka 5.3 ukazuje malé počty shodných jmen v použitých seznamech oceněných autorů. Vidět v ní lze, že pouze jedna osoba byla oceněna v oblasti *Umělé inteligence* i v oblasti *Hardwaru* a že osoby oceněné v oblasti *Umělé inteligence* jsou na seznamech *ACM Fellows* a *ISI Highly Cited* zastoupeny méně než osoby oceněné v oblasti *Hardwaru*.

[54] Web *ACM Fellows* - http://fellows.acm.org
[55] Web *ISI Highly Cited Researchers* - http://www.highlycited.com

Tabulka 5.2: Počty držitelů ocenění, která udílejí SIGs z vybraných ACM kategorií.

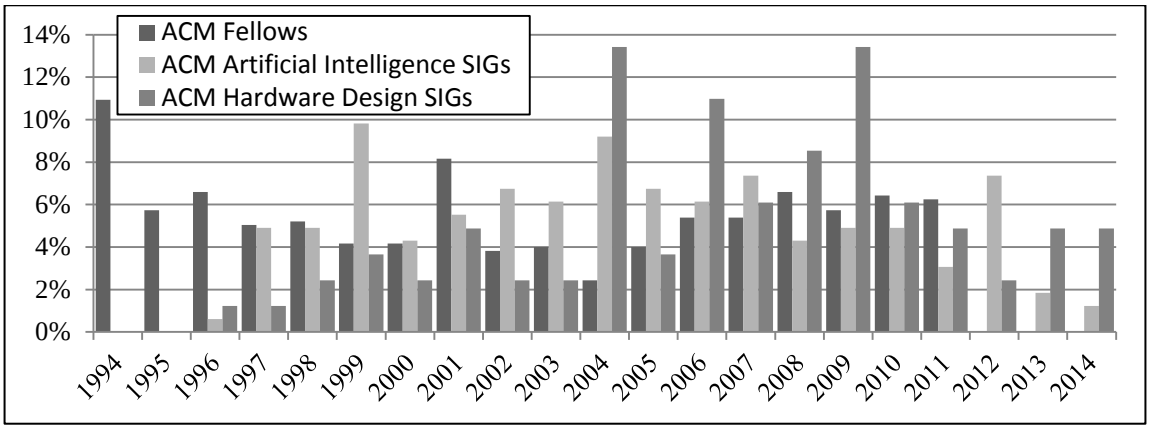
Kategorie	Název SIG	Název ocenění	Oceněných	V celé WoS	V kategorii
Umělá inteligence (354 jmen oceněných autorů)	**SIGAI**	Allen Newell Award (1994-2012)	21	15	11
		The ACM/SIGAI Autonomous Agents Research Award (2001-14)	14	12	12
		A.M. Turing Award (AI, 2010-11)	2	2	2
	SIGecom	The SIGecom Best Paper Awards (2006-14)	37	14	9
	SIGEVO	SIGEVO Impact Award (2011-13)	8	2	2
	SIGIR	Gerard Salton Award (1983-2012)	10	8	0
		Best paper award (1996-2014)	51	36	15
		Test of Time Award 2014 (2002-3)	13	9	6
	SIGKDD	SIGKDD Innovation Award (2000-14)	14	13	13
		SIGKDD Service Award (2000-14)	12	10	7
		SIGKDD Best Research Paper Awards (1997-2007)	48	37	22
		SIGKDD Best Application Paper Award (1997-2007)	52	37	29
	SIGWEB	Hypertext Douglas Engelbart Best Paper Award (1996-2013)	46	22	9
		SIGWEB/SIGIR Vannevar Bush Award (1998-2013)	47	27	11
Návrh hardwaru (158 jmen oceněných autorů)	**SIGARCH**	ACM/IEEE Eckert-Mauchly Award (1979-2014)	36	0	0
		ACM SIGARCH Maurice Wilkes Award (1998-2014)	17	16	14
		ACM SIGARCH Distinguished Service Award (2008-14)	6	4	3
		ACM SIGARCH/IEEE-CS TCCA Influential ISCA Paper Award (2003-14)	46	31	30
	SIGBED	SIGBED EMSOFT Best Paper Award (2008-13)	18	8	7
	SIGDA	SIGDA Outstanding New Faculty Award (2004-14)	13	11	10
	SIGMOBILE	The SIGMOBILE Distinguished Service Award (2001-3)	3	3	0
		The SIGMOBILE Outstanding Contribution Award (1996-2014)	15	14	12
		The SIGMOBILE RockStar Award (2013-14)	2	2	2
	SIGMICRO	---			

Tabulka 5.3: Počty shodných jmen v použitých seznamech oceněných autorů.

Seznamy oceněných autorů	ACM Fel.	ISI HC	ACM AI	ACM HW
ACM Fellows	576	74	13	21
ISI Highly Cited	74	280	3	6
ACM Artificial Intelligence SIGs	13	3	133	1
ACM Hardware SIGs	21	6	1	76

Na obrázku 5.3 jsou pro každý vytvořený seznam oceněných autorů znázorněny četnosti udílení ocenění v jednotlivých letech (graf z obrázku 5.3 lze porovnat s grafem na obrázku 5.1). Seznam *ISI Highly Cited* na obrázku 5.3 není, protože neobsahoval explicitně vyjádřené roky udělení cen. Z grafu je zřejmé, že pro vyhodnocení experimentu bylo použito i hodně osob oceněných po roce 2005 (poslední rok obsažený v naší kolekci WoS). Tyto osoby jsme použili mimo jiné proto, abychom

otestovali, jestli naše metody pro hodnocení autorů dokáží odhalit autory, kteří teprve budou oceněni v blízké budoucnosti (2006-2014). To odpovídá potřebě vyhledávat jak aktuálně nejlepší vědce, tak i ty, kteří budou jedni z nejlepších v blízké budoucnosti. Jak je diskutováno dále v části 5.4.4, nejlepší z našich metod byla nejlepší i v tomto testu.



Obrázek 5.3: Četnosti zastoupení roků v seznamech oceněných autorů.

V závěru této části bychom měli zmínit, že přesto, že ACM je největší organizace zabývající se počítačovými vědami, tak (na rozdíl od Nobelovy ceny, která má globální rozsah) zaměření ACM může být do značné míry omezeno na Spojené státy americké (USA) a ACM proto může favorizovat vědce z USA. Protože naše kolekce WoS obsahuje záznamy o vědeckých publikacích z celého světa (oblast počítačových věd), tak porovnání výsledků na základě osob oceněných od ACM může vést k možnému nesouladu.

5.3 Úpravy personalizace PageRanku pro účely hodnocení autorů

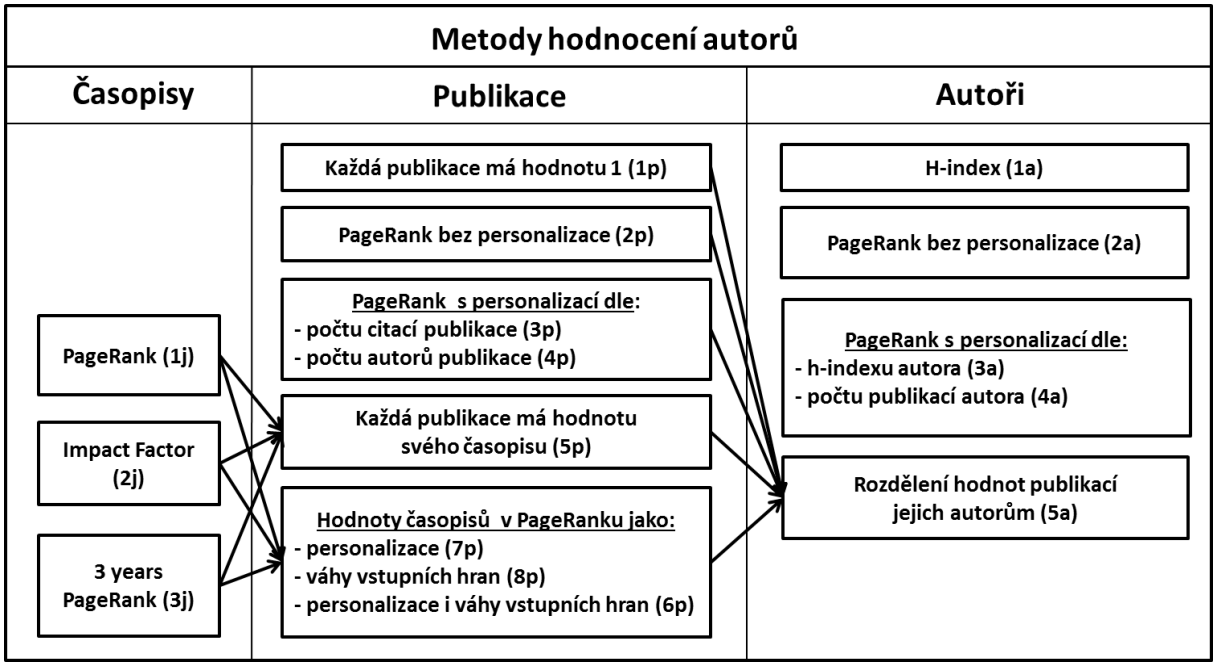
Hlavním algoritmem, který jsme použili pro hodnocení autorů, je PageRank s personalizací, který byl v části 2.4.2 popsán vzorcem (2.12). Protože jsme se při návrhu metod zaměřili na modifikaci personalizace PageRanku, tak můžeme zopakovat, že míru toho, jak moc jsou vypočtené PageRankové hodnoty vrcholů tvořeny personalizací, ovlivníme velikostí faktoru tlumení d, $0 < d < 1$ (pozn.: pokud je d=1, tak PageRank nemusí konvergovat, a pokud je d=0, tak jsou hodnoty vrcholů tvořeny pouze personalizací, viz část 2.4.3). Brin a Page (1998) pro vyhodnocení citačního grafu webových stránek použili d=0,85. Chen et al. (2007) zmiňují, že pro vyhodnocení „odborných" citačních sítí je vhodné použít d=0,5. K hodnotě d=0,55 dospěli také Yan a Ding (2011), když aplikovali PageRank s personalizací dle počtu citací na síť spoluautorů. My testovali $d \in [0{,}15; 0{,}25; \ldots ; 0{,}85]$. Pokud není při výpočtu PageRanku použita personalizace (respektive personalizace všech vrcholů je rovna 1 nebo je použit vzorec (2.11)), tak má velikost faktoru tlumení minimální vliv na vytvořené pořadí vrcholů (mohou se projevit zaokrouhlovací chyby) a my proto použili klasickou hodnotu d=0,85. Výpočty PageRanku byly vždy ukončeny po 50 iteracích.

Poznámka: obecně může vzorec PageRanku obsahovat více částí propojených damping faktory, jejichž součet je 1 (Sayyadi a Getoor 2009), nebo může být personalizace nahrazena dynamickou částí (Sidiropoulos a Manolopoulos 2006), ale to nebylo obsahem experimentů, které jsme prováděli.

V části 5.1 jsme zmínili, že jsme do experimentů zahrnuli h-index a počet citací publikace, které jsou definovány následovně:

- *Autor má h-index o velikosti h, pokud h z jeho publikací obdrželo alespoň h citací a žádná další jeho publikace nemá více než h citací*, viz (Hirsch 2005) a část 2.3.2. H-index autora by měl vzrůstat s každým rokem jeho vědecké činnosti, proto ho lze chápat jako hodnotu autorovy vyspělosti.
- Počet citací lze definovat jako součet vah všech vstupních hran vrcholu v citačním grafu, což dobře odráží popularitu daného vrcholu (Ding 2011a), viz část 2.1.

Na obrázku 5.4 jsou shrnuty a v následujících seznamech detailně popsány konkrétní metody, které jsme navrhli a použili pro hodnocení autorů. Šipky na obrázku 5.4 ukazují, jak se ve výpočtu hodnot významnosti autorů uplatňují hodnoty publikací či časopisů. Například metoda (5a7p1j) pro hodnocení autorů použije PageRankové hodnoty časopisů (1j) v personalizaci PageRanku, který aplikuje na citační síť publikací (7p), a vypočtené hodnoty publikací rozdělí jejich autorům (5a).



Obrázek 5.4: Metody použité pro hodnocení autorů.

Následující popisy metod jsou doplněny o důvody, které nás k návrhu metod vedly, a o symbolické zápisy určení hodnot významnosti, personalizace nebo vah vstupních hran autorů, publikací a časopisů. V symbolických zápisech A zastupuje autora, P publikaci a J časopis. *Významnost(A)* je významnost autora A (popř. publikace P nebo časopisu J) a *Personalizace(A)* je personalizace autora A (pozn.: ve vzorci PageRanku (2.12) je personalizace značena p_A). Ψ_A je množina všech publikací autora A a *Díl(Významnost(P))* je díl z hodnoty významnosti publikace P určený některým ze způsobů rozdělení hodnot publikací jejich autorům, které jsou popsány v samostatné části 5.3.2. *Váhy(P)* jsou váhy vstupních hran vrcholu, který zastupuje publikaci P. Pozn.: v sítích autorů jsou váhy hran určeny třemi způsoby, viz část 5.3.1, v síti časopisů jsou použity váhy hran vyjadřující počet citování, viz část 5.3.4, a v síti publikací, není-li řečeno jinak, mají všechny hrany váhu 1.

První seznam metod (1a - 5a) obsahuje metody pro získání hodnot autorů. Metoda (5a) vyžaduje určení hodnot publikací některou z možností (1p - 8p). Metody (5a5p - 5a8p) používají hodnoty časopisů, které jsou stanoveny některou z možností (1j - 3j).

Metody (1a – 5a) pro výpočet hodnot významnosti autorů:

(1a) *H-indexem*, který zastupuje vyspělost autora. *Významnost(A) = h-index(A)*;

(2a) *PageRankem bez personalizace*, který zastupuje prestiž autora a byl použit jako baseline pro metody pracující se sítí autorů. *Personalizace(A)* = 1;

(3a) *PageRankem s h-index personalizací*, který do výpočtu prestiže autora zahrnuje autorovu vyspělost. *Personalizace(A) = h-index(A)*;

(4a) *PageRankem s personalizací dle počtu publikací autora*, který do výpočtu prestiže zahrnuje autorovu produktivitu. *Personalizace(A) = počet_publikací(A)*;

(5a) *Součtem dílů hodnot publikací*, které autor napsal, viz část 5.3.2 a možnosti (1p – 8p) pro určení hodnot významnosti publikací. *Významnost(A)* = $\sum_{P \epsilon \Psi_A} Díl(Významnost(P))$;

Možnosti (1p – 8p) pro určení hodnot významnosti publikací, které jsou použity metodou (5a):

(1p) *Všem publikacím je nastavena hodnota 1*, což po převodu na autory zastupuje různé způsoby počítání produktivity autorů. *Významnost(P)* = 1;

(2p) *PageRankem bez personalizace*, který počítá prestiž publikací a ta po převodu na autory poskytuje prestiž autorů. Tato metoda byla použita jako baseline pro metody pracující se sítí publikací. *Personalizace(P)* = 1;

(3p) *PageRankem s personalizací dle počtu citací publikace*, který do výpočtu prestiže publikací zahrnuje jejich popularitu. *Personalizace(P) = počet_citací(P)*;

(4p) *PageRankem s personalizací dle počtu autorů publikace*, který do výpočtu prestiže publikací zahrnuje jejich kvalitu. Zde jsme předpokládali, že čím více autorů na publikaci spolupracovalo, tím více pracovního úsilí a peněz do ní bylo investováno a publikace je proto kvalitnější. *Personalizace(P) = počet_autorů(P)*;

(5p) *Všem publikacím je nastavena hodnota dle časopisu*, ve kterém byly vytištěny, viz možnosti (1j – 3j) pro určení hodnot významnosti časopisů. To po převodu na autory udává autorovu produktivitu založenou na významnosti časopisů, ve kterých autor publikoval. *Významnost(P) = Významnost(J)*;

(6p) *PageRankem s personalizací a vahami vstupních hran nastavenými dle hodnot časopisů*, který do výpočtu prestiže publikací zahrnuje významnosti jejich časopisů.
Personalizace(P) = Významnost(J); *Váhy(P) = Významnost(J)*;

(7p) *PageRankem s personalizací dle hodnot časopisů*. Tato a následující metoda byly použity, abychom zjistili, zda je lepší uplatnit významnosti časopisů ve výpočtu PageRanku publikací jako personalizace nebo jako váhy vstupních hran.
Personalizace(P) = Významnost(J); *Váhy(P) = 1*;

(8p) *PageRankem bez personalizace, ale s vahami vstupních hran nastavenými dle hodnot časopisů* (např. pokud byla publikace vytištěna v časopise s Impact Factorem 12, tak všechny vstupní hrany vrcholu, který zastupuje danou publikaci, mají váhu 12). Porovnání této metody s baseline bylo použito pro zjištění, zda má smysl měnit váhy hran v citační síti publikací. *Personalizace(P) = 1*; *Váhy(P) = Významnost(J)*;

Možnosti (1j – 3j) pro určení hodnot významnosti časopisů, které používají metody (5a5p – 5a8p):

(1j) *PageRankem bez personalizace*, který počítá prestiž časopisů.
Významnost(J) = *pagerank(J); Personalizace(J)* = 1;

(2j) *Impact Factorem*, který počítá popularitu časopisů v obdobích vydávání publikací, viz část 5.3.4. *Významnost(J)* = *impact_factor(J)*;

(3j) *3 years PageRankem*, který počítá prestiž časopisů v obdobích vydávání publikací, viz část 5.3.4. *Významnost(J)* = *3_years_pagerank(J)*; *Personalizace(J)* = 1;

Určení kvality navržených metod pro hodnocení autorů bylo provedeno na základě průměrné pozice oceněných autorů (z referenčních seznamů) ve vytvořených pořadích. Autor s nejlepším hodnocením obsadil ve vytvořeném pořadí autorů první pozici, a proto nejnižší průměrná pozice oceněných autorů ve vytvořeném pořadí určuje naší nejlepší metodu pro hodnocení autorů. Baseliny (2a) a (5a2p) stanovily minimální průměrné pozice oceněných autorů, které mohou při výpočtu prestiže PageRankem ocenění autoři obsadit ve vytvořeném pořadí autorů.

Všechny navržené metody pro hodnocení autorů jsme testovali při globálním hodnocení autorů v počítačových vědách. Abychom využili zvolené kategorie *Umělá inteligence* a *Hardware*, tak jsme také použili dva způsoby jejich vyhodnocení, které jsme pojmenovali *category-independent* (vyhodnocení nezávislé na kategorii) a *category-dependent* (vyhodnocení závislé na kategorii). Tato terminologie je analogií k terminologii používané v oblasti využití PageRanku pro řazení relevantních výsledků ve vyhledávači webových stránek, kde je PageRank obvykle počítán před položením dotazu (je tzv. *query-independent*), kdežto např. algoritmus HITS (viz část 2.5.3) bývá počítán až nad výsledky dotazu (je tzv. *query-dependent*). Více informací obsahuje např. (Langville a Meyer 2006).

Použité způsoby získání pořadí autorů ve zvolených kategoriích lze detailněji popsat takto:

- *Category-independent* způsob hodnocení autorů používá pořadí autorů, která byla vytvořena globálními hodnoceními autorů v počítačových vědách. V těchto pořadích jsou vyhledáni autoři, kteří alespoň jednou publikovali ve zvolené WoS kategorii. Tito autoři jsou vybráni včetně vypočtených hodnot, dle kterých se vytvoří jejich pořadí v dané kategorii. Výhodou tohoto postupu je, že využívá informace z celé naší kolekce WoS, tj. ze všech oblastí počítačových věd.
- *Category-dependent* způsob hodnocení autorů vyhodnocuje navrženými metodami citační sítě publikací nebo autorů, které vznikly na základě článků publikovaných v časopisech ze zvolené WoS kategorie. Výhodou tohoto postupu je, že hodnocení autorů je založeno pouze na jejich publikační činnosti ve zvolené WoS kategorii.

V následujících částech jsou detailně popsány jednotlivé metody pro hodnocení autorů. Nejprve jsou popsány metody používající citační síť autorů (část 5.3.1), následně možnosti rozdělení hodnot publikací jejich autorům (část 5.3.2) a metody používající citační síť publikací (část 5.3.3). Na závěr je ukázáno, jakými způsoby lze získat a využít hodnoty významnosti časopisů (část 5.3.4).

5.3.1 Experimenty se sítí autorů

Pro experimenty s citačními sítěmi autorů jsme zvolili stejné tři varianty stanovení vah hran jako ve 3. kapitole:

- Varianta *N* hranám přiřazuje váhy vyjadřující počet citací, které autor obdržel v citační síti publikací.
- Varianta *1/N* rozděluje hodnoty publikací rovnoměrně mezi uvedené reference. To znamená, že pokud autor A ve své publikaci citoval publikaci autorů B a C a publikaci autora E, tak citační síť autorů obsahuje hrany z A do B a z A do C s váhou ½ a hranu z A do E s váhou 1. Souběžné hrany jsou sloučeny a jejich váhy sečteny.
- Varianta *1* přiřadí všem hranám váhu 1.

PageRank bez personalizace (2a) nám při vyhodnocení sítí autorů sloužil jako baseline, přičemž jsme předpokládali, že bude mít lepší výsledky než samotný h-index (1a). Metoda (4a) využívající personalizaci autorů jejich produktivitou (počtem publikací autora) poskytovala ve 4. kapitole lepší hodnocení autorů na základě jejich citační sítě než PageRank bez personalizace. Ve zde prezentovaném experimentu jsme předpokládali, že PageRank využívající vyspělost autorů (h-index personalizace, 3a) poskytne ještě lepší pořadí autorů než metoda (4a).

Obdobu metody (4a) použili také West et al. (2013), kteří ji zakomponovali do výpočtu Eigenfactor Score na úrovni autorů (obdoba Eigenfactor Score z části 2.5.6). Touto metodou hodnotili autory obsažené v kolekci, kterou vytvořili na základě *Social Science Research Network*[56] (ve vyhodnocované citační síti autorů byly použity podobné váhy hran, jako používá naše varianta *1/N*). Následně ukázali výsledky z hodnocení autorů (a odvozených hodnocení institucí a států) na základě navržené metody a dále na základě počtu citací, referencí a stažení (*download*) jejich článků. Autoři sice v závěru říkají, že jimi navržená metoda poskytuje lepší hodnocení autorů než zbylé testované metody, ale, protože autoři výsledky neporovnali s žádným referenčním seznamem autorů a ani s jinou metodou počítající prestiž, tak nelze určit, jak kvalitní hodnocení autorů jejich metoda poskytuje. Citační síť autorů, s podobným nastavením vah hran jako má naše varianta *1/N*, vyhodnocovali také Radicchi et al. (2009), kteří hodnotili autory obsažené ve fyzikálních přehledových časopisech[57]. Svou metodu pro hodnocení autorů, která je podobná naší metodě (4a), ale její zápis je komplikovanější, porovnali s metodami počítajícími citace a ukázali, že jejich metoda počítající prestiž autorů poskytuje lepší hodnocení než metody popularity – na předních pozicích v pořadích autorů bylo více oceněných osob. Nicméně autoři netestovali, zda jejich metoda poskytuje lepší hodnocení autorů než jiné metody pro výpočet prestiže. Naším přínosem v této části tedy je porovnání metody (4a) s dalšími metodami pro výpočet prestiže autorů v kvalitní datové kolekci WoS z oblasti počítačových věd.

5.3.2 Rozdělování hodnot publikací jejich autorům

Pokud chceme určovat významnost autorů na základě hodnot jejich publikací (metoda (5a)), tak máme několik možností, jak hodnoty publikací mezi jejich autory rozdělit. Protože ve 4. kapitole se ukázalo vhodnější rozdělovat autorům rovnoměrné díly hodnot jejich publikací, než sčítat jim celé hodnoty publikací, tak jsme experimenty nyní doplnili o další způsoby rozdělení hodnot publikací

[56] Web *Social Science Research Network* - http://www.ssrn.com

[57] Datové kolekce fyzikálních přehledových časopisů (*the datasets of Physical Review journals*) - http://journals.aps.org/datasets

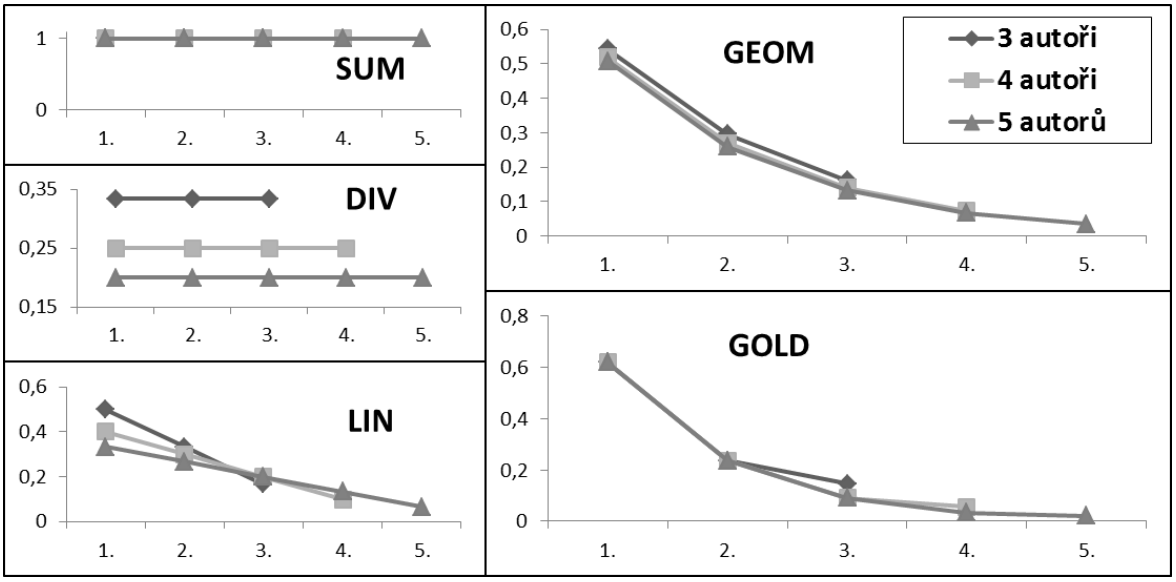
jejich autorům. Tyto způsoby rozdělení bývají v anglicky psané literatuře označovány jako *Counting methods* (Gauffriau a Larsen 2005), což lze přeložit jako „metody počítání“, ale my nadále použijeme pojem „rozdělení“. Pro experimenty jsme použili rozdělení popsaná vzorci (5.1) až (5.7), pro jejichž výběr nás inspirovali Assimakis a Adam (2010), kteří diskutovali možnosti rozdělení publikace mezi její autory a pro účely neměnného zvýhodnění prvních autorů publikací navrhli své vlastní rozdělení, které nazývají *Golden Productivity Index* (dále značen *Zlaté rozdělení*).

V úvodu této části můžeme ještě zmínit, že ve článcích (Lindsey 1980; Van Hooydonk 1997; Egghe et al. 2000; Gauffriau a Larsen 2005; Assimakis a Adam 2010; Hagen 2010), které jsou v této části citovány, autoři buď pouze navrhli zmíněná rozdělení, ale netestovali je na reálných datech, nebo je testovali tak, že každá publikace měla stejnou hodnotu, viz metoda (5a1p). O této metodě počítání shodně oceněných publikací, která počítá produktivitu autorů, bylo již dříve prokázáno, že hodnotí autory hůře, než metody popularity nebo prestiže. Přínosem našeho experimentu v tomto případě je, že jsme navržená rozdělení otestovali na reálných datech a navíc v situaci, při které jsou publikace ohodnoceny metodami prestiže (5a3p – 5a8p). Pozn.: výjimku tvoří metoda (5a5p2j), která určuje hodnoty publikací na základě popularity jejich časopisů, viz část 5.3.4.

Ve vzorcích (5.1) až (5.7) je:

- Ψ_A množina všech publikací autora *A*,
- *VAL(Q)* hodnota publikace *Q*,
- Q_N počet autorů publikace *Q,*
- Q_{Aj} pozice autora *A* ve výčtu autorů publikace *Q*,
- *SUM(A)*, *DIV(A)*, *LIN(A)*, *GEOM(A)* a *GOLD(A)* daným rozdělením získaný součet dílů z hodnot publikací autora *A*.

Na obrázku 5.5 je ukázáno, jak lze dle zvolených rozdělení rozdělit hodnotu publikace mezi 3, 4 nebo 5 autorů. Na vodorovné ose je vždy zobrazeno pořadí autora ve výčtu autorů publikace a na svislé ose díl z hodnoty publikace, který autor získá.



Obrázek 5.5: Způsoby rozdělení hodnoty publikace mezi 3, 4 nebo 5 autorů.

Zopakovat můžeme, že *součtové rozdělení* (*SUM)* sčítá autorům celé hodnoty jejich publikací, viz vzorec (5.1), a *rovnoměrné rozdělení (DIV)* jim sčítá rovnoměrné díly hodnot publikací, viz vzorec (5.2). Součtové rozdělení je v anglicky psané literatuře označováno jako *Normal*, *Total* nebo *Whole counting* a rovnoměrné rozdělení jako *Fractional* nebo *Adjusted counting* (Lindsey 1980; Egghe et al. 2000; Gauffriau a Larsen 2005).

$$SUM(A) = \sum_{Q \in \Psi_A} VAL(Q) \tag{5.1}$$

$$DIV(A) = \sum_{Q \in \Psi_A} \left[\frac{1}{Q_N} \cdot VAL(Q) \right] \tag{5.2}$$

Lineární rozdělení (*LIN)*, známé jako *Arithmetic* nebo *Proportional counting* (Van Hooydonk 1997), používá pro určení dílů publikace lineární funkci, která využívá pozice autorů ve výčtu autorů publikace, viz vzorec (5.3).

$$LIN(A) = \sum_{Q \in \Psi_A} \left[\left(\frac{-2 \cdot Q_{Aj}}{Q_N \cdot (Q_N + 1)} + \frac{2}{Q_N} \right) \cdot VAL(Q) \right] \tag{5.3}$$

Geometrické rozdělení (*GEOM*) dle Assimakis a Adam (2010) rozděluje hodnoty publikací autorům vzorcem (5.4), ve kterém je λ určena dle pozice autora ve výčtu autorů publikace rovnicí (5.5) nebo (5.6). Rovnice (5.5) a (5.6) mohou mít více reálných kořenů, ale nás zajímá pouze kořen splňující kritérium $0 < \lambda < 1$. Tento kořen mají obě rovnice stejný. Assimakis a Adam (2010) zmiňují, že takový kořen existuje vždy právě jeden. Egghe et al. (2000) podobný postup rozdělení hodnot publikací jejich autorům nazývají *Pure geometric count*. Porovnání *DIV*, *LIN* a *Pure geometric count* ukazují např. (Egghe et al. 2000; Hagen 2010).

$$GEOM(A) = \sum_{Q \in \Psi_A} [\lambda^{Q_{Aj}} \cdot VAL(Q)] \tag{5.4}$$

$$\lambda^{Q_N} + \lambda^{Q_N - 1} + \cdots + \lambda^1 = 1 \tag{5.5}$$

$$\lambda^{Q_N + 1} - 2 \cdot \lambda + 1 = 0 \tag{5.6}$$

Zlaté rozdělení (GOLD), které navrhli Assimakis a Adam (2010), používá pro určení dílů hodnot publikací vzorec (5.7), kde φ je konstanta získaná z rovnice $\varphi^2 + \varphi = 1$, která má právě jedno kladné reálné řešení a to 0,618. Zatímco geometrické rozdělení s přibývajícím počtem autorů publikace mění všem autorům velikost dílu, který z hodnoty publikace získají, tak zlaté rozdělení mění díl hodnoty publikace vždy pouze aktuálně poslednímu autorovi. Například, pokud má publikace tři nebo více autorů, tak první autor vždy získá 61,8% z její hodnoty a druhý autor 23,6%.

$$GOLD(A) = \sum_{Q \in \Psi_A} [VAL(Q) \cdot \Gamma(A)]$$

$$\Gamma(A) = \begin{cases} 1 & Q_N = 1 \\ \varphi^{2 \cdot Q_{Aj} - 1} & Q_{Aj} = 1, \dots, (Q_N - 1);\ Q_N > 1 \\ \varphi^{2 \cdot Q_{Aj} - 2} & Q_{Aj} = Q_N \end{cases} \tag{5.7}$$

Další možností při rozdělování hodnot publikací autorům by bylo použít např. pouze první autory publikací, jak popisují (Zhao 2005; Ding 2011a), ale my si myslíme, že je spravedlivé, aby každý autor publikace získal alespoň malou část z její hodnoty. Postupu, při kterém jsou použiti pouze první autoři publikací, by v našem případě mělo být nejbližší zlaté rozdělení.

5.3.3 Experimenty s hodnocením autorů na základě hodnot jejich publikací

Pro stanovení hodnot autorů na základě hodnot jejich publikací metodou (5a) jsme použili rozdělení hodnot publikací, která byla zmíněna v předchozí části. Hodnoty publikací byly určeny osmi různými postupy. Postup, při kterém má každá publikace hodnotu 1 (5a1p), měl jako první naznačit, zda má smysl na reálných datech používat nerovnoměrná rozdělení hodnot publikací jejich autorům. Navíc, rozdělení *SUM* v tomto případě představovalo hodnocení autorů dle počtu jejich publikací. PageRank bez personalizace měl ukázat, jakého zlepšení výsledků lze dosáhnout vyhodnocením citační sítě publikací (5a2p), oproti vyhodnocení sítě autorů (2a) nebo oproti počítání publikací (5a1p), a byl proto použit jako druhý baseline. Protože počet citací, zastupující popularitu, a PageRank, zastupující prestiž, nejvíce odrážejí vliv publikace na vědeckou komunitu, tak jsme dále testovali PageRank s personalizací dle počtu citací publikace (5a3p).

Naše nejlepší metoda pro globální hodnocení autorů v počítačových vědách ze 4. kapitoly, PageRank s personalizací dle počtu autorů publikace aplikovaný na citační síť publikací (5a4p), byla určena porovnáním vytvořeného pořadí autorů se seznamy držitelů prestižních ocenění. Proto můžeme hodnocení autorů používající hodnoty PageRanku jejich publikací popsat jako: „*Prestiž autorů založená na prestiži jejich publikací (která má v sobě navíc dodatečně zakomponovanou kvalitu publikací)*“. Personalizace publikací dle počtu jejich autorů v našem případě poprvé zohlednila kvalitu publikace, kterou jsme popsali jako míru prostředků vynaložených na její vytvoření. V tomto experimentu byla metoda (5a4p) použita jako indikátor, zda nově testované metody zlepšují hodnocení autorů a poskytují tak jejich lepší pořadí.

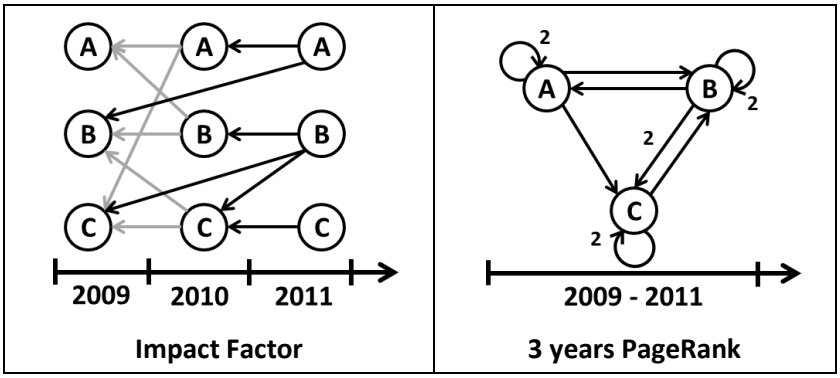
Další ukazatel kvality publikace, jehož využití pro hodnocení autorů jsme testovali, byla hodnota významnosti časopisu, ve kterém byla publikace vytištěna. Nejlepší publikace jsou obvykle přijímány do nejvýznamnějších časopisů daného vědního oboru, proto jsme předpokládali, že metoda pro hodnocení autorů, která použije publikace zvýhodněné na základě hodnot časopisů, poskytne naše nejlepší pořadí autorů. Hodnoty časopisů jsme stanovili a použili více způsoby, proto jsme těmto metodám hodnocení autorů (5a5p – 5a8p) věnovali samostatnou následující část.

5.3.4 Použití významnosti časopisů při hodnocení autorů

Hodnoty časopisů jsme vypočetli přímo z našich dat primárně dvěma metodami: Impact Factorem a PageRankem. Vyhodnoceny byly citační sítě časopisů se všemi citacemi (*ALL*) a sítě s odstraněnými samocitacemi autorů na úrovni publikací (*NOT*). Váhy hran vždy vyjadřovaly počet citací (varianta *N*).

Jinou možností by bylo použití Impact Factorů uvedených v JCR z jednotlivých let, ale tento postup by pro nás byl složitější, protože jsme neměli historické záznamy JCR.

Zopakovat můžeme, že Impact Factor[58] (více detailů viz část 2.3.1) časopisu *j* v daném roce (např. 2011) je počet citací z tohoto roku na všechny články publikované v časopise *j* dva roky před tím (tj. 2010 a 2009), který je dělený počtem všech „podstatných" článků (tj. bez redakčních poznámek, úvodních článků, recenzí atd.) publikovaných v těchto dvou letech v časopise *j*. Z toho lze vypozorovat, že Impact Factor citujících časopisů se při výpočtu nepoužívá, a proto lze Impact Factor považovat za metodu měřící popularitu. Jak je zřejmé, Impact Factor používá citace z daného roku, pro který je počítán, na publikace vydané dva roky před tím, ale PageRank používá celou citační síť. Pro alespoň částečné porovnání PageRanku a Impact Factoru jsme PageRankem vyhodnotili také citační pod-sítě časopisů vytvořené dle tříletých časových okének, např. 1996-1998, 1997-1999 atd. Tento typ hodnocení časopisů budeme nazývat *3 years PageRank*. Rozdíl mezi sítěmi pro Impact Factor a pro 3 years PageRank zobrazuje obrázek 5.6, na kterém je ukázáno, že šedé citace nejsou při výpočtu Impact Factoru použity, protože nepocházejí z roku, pro který je Impact Factor počítán.



Obrázek 5.6: Rozdíl mezi sítí časopisů pro Impact Factor a sítí pro 3 years PageRank (šedé citace se při výpočtu Impact Factoru nepoužijí).

Hodnoty časopisů byly použity ve čtyřech našich metodách pro hodnocení publikací. První metoda přidělila každé publikaci hodnotu časopisu, ve kterém byla uveřejněna (5a5p). Tato metoda, v porovnání s metodou (5a1p), která přiděluje všem publikacím hodnotu 1, měla jako první naznačit, zda se vyplatí při hodnocení autorů použít hodnoty časopisů. Ve zbylých třech metodách byl o hodnoty časopisů obohacen PageRank, který jsme aplikovali na citační síť publikací. Metoda (5a7p) použila hodnoty časopisů jako personalizace publikací, metoda (5a8p) je použila jako váhy vstupních hran vrcholů zastupujících publikace a metoda (5a6p) použila obojí.

5.4 Diskuse výsledků navržených metod

Všechny výše popsané metody pro hodnocení autorů jsme aplikovali na citační sítě vytvořené z kolekce WoS a získaná pořadí autorů jsme porovnali na základě referenčních seznamů významných autorů, kteří jsou držiteli prestižního ocenění. Tabulka 5.4 shrnuje úspěšnost námi testovaných metod. Její sloupce ***p.*** ukazují pořadí metod v úspěšnosti vyzdvižení významných autorů z jednotlivých seznamů oceněných autorů ve vytvořeném pořadí autorů. Čím menší hodnotu ve

[58] Výpočet Journal Impact Factoru v databázi WoS - http://admin-apps.webofknowledge.com/JCR/help/h_impfact.htm

sloupci *p.* metoda má (1 – 12), tím lépe ve vytvořeném pořadí hodnotí významné autory z daného seznamu. Nejmenší a současně nejlepší průměrná pozice významných autorů z daného seznamu, v pořadí autorů vytvořeném metodou s *p.*=1, je pro každý seznam oceněných autorů uvedena v posledním řádku tabulky s označením m_{best}. Protože každá naše metoda má několik variant (např. různá nastavení vah hran a samocitací, použitá rozdělení hodnot publikací autorům atd.), tak jsme v tabulce 5.4 pro každou metodu a seznam oceněných autorů zobrazili výsledky pouze té varianty dané metody, která poskytla nejlepší průměrnou pozici oceněných autorů (detailnější vyhodnocení obsahují části 5.4.1 a 5.4.2). Jak se v hodnocení autorů od nejlepší metody lišily ostatní metody, demonstrují sloupce $m_{\%}$ s procentuálními rozdíly průměrných pozic oceněných autorů ve vytvořených pořadích. Hodnotu průměrné pozice oceněných autorů *avg(m)* v pořadí vytvořeném nejlepší variantou dané metody *m*, lze rekonstruovat vzorcem (5.8), který používá hodnotu průměrné pozice oceněných autorů v pořadí vytvořeném naší nejlepší metodou (řádek m_{best}) a procento odlišení (sloupec $m_{\%}$) metody *m* od m_{best}.

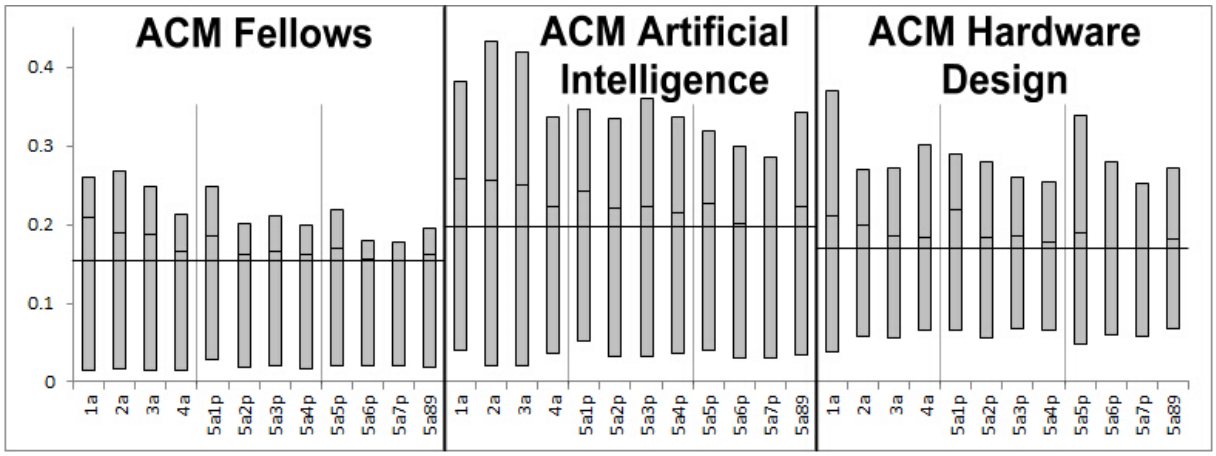
$$avg(m) = m_{best} \cdot \left(1 + \frac{m_{\%}}{100}\right) \tag{5.8}$$

Tabulka 5.4: Porovnání kvality našich metod při hodnocení autorů z kolekce WoS (sloupce p. zobrazují pořadí úspěšnosti jednotlivých metod a sloupce $m_{\%}$ procento jejich odlišení od minimální dosažené průměrné pozice oceněných autorů, která byla dosažena příslušnou nejlepší metodou s p.=1 a jejíž hodnota je uvedena v řádku m_{best}).

Autoři ocenění v letech 1994 až 2014												
Síť		**Metoda** (Oceněno - / Hodnoceno -)	**Global**				**Category-indep.**				**Category-dep.**	
			Fellows 576 / 157440		**ISI HC** 280 / 157440		**AI** 133 / 39891		**HW** 76 / 29243		**AI** 133 / 39891	**HW** 76 / 29243
			p.	**$m_{\%}$**	**p.**	**$m_{\%}$**	**p.**	**$m_{\%}$**	**p.**	**$m_{\%}$**	**p.**	**p.**
Autoři	1a	H-index	12	35%	12	36%	12	31%	11	24%	12	38%	11	29%
	2a	PageRank bez personalizace	11	23%	11	22%	11	29%	10	17%	11	29%	10	25%
	3a	Personalizace h-indexem	10	21%	9	21%	10	26%	8	10%	10	29%	8	21%
	4a	Personalizace počtem publikací	**6**	**7%**	**7**	**15%**	**7**	**13%**	**5**	**8%**	**5**	**11%**	**5**	**14%**
Publikace	5a1p	Všechny mají hodnotu 1	9	20%	10	30%	9	22%	12	29%	9	25%	12	54%
	5a2p	PageRank bez personalizace	5	5%	6	14%	4	12%	6	9%	4	11%	7	19%
	5a3p	Personalizace počtem citací	7	8%	8	19%	6	13%	7	10%	7	13%	**4**	**9%**
	5a4p	Personalizace počtem autorů	**3**	**4%**	**4**	**13%**	**3**	**8%**	**3**	**5%**	**3**	**7%**	9	21%
	5a5p	Hodnoty dle hodnot časopisů	8	10%	3	10%	8	15%	9	12%	8	19%	3	6%
	5a6p	Hodnoty časopisů v pers. i hranách	2	0,4%	2	0,4%	2	2%	2	0,1%	2	1,5%	**1**	**0%**
	5a7p	Hodnoty časopisů v personalizaci	**1**	**0%**	**1**	**0%**	**1**	**0%**	**1**	**0%**	**1**	**0%**	2	1,0%
	5a8p	Hodnoty časopisů ve vstup. hranách	4	4%	5	14%	5	12%	4	7%	6	12%	6	17%
Minimální průměrná pozice oceněných autorů (m_{best})			24315		19934		7892		4949		9515	4715		

Výsledky metod získané z globálního hodnocení autorů v počítačových vědách a z hodnocení autorů v kategoriích *Umělá inteligence* a *Hardware* category-independent způsobem, jsou na základě krabicového diagramu porovnány také na obrázku 5.7 (porovnání metod na základě seznamu *ISI Highly Cited* zobrazeno není). Jednotlivé sloupce diagramu zobrazují relativní pozice oceněných

autorů ve vytvořených pořadích autorů. Relativní pozici každého oceněného autora jsme vypočítali vydělením jeho pozice počtem hodnocených autorů. Nejlepší a nejhorší čtvrtiny oceněných autorů nejsou pro přehlednost diagramů zobrazeny, protože testované metody obvykle dokázaly alespoň jednoho oceněného autora umístit mezi nejlepší autory (např. do patnácté nejlepší pozice, viz část 5.4.4) a naopak alespoň jednoho autora ze seznamu *ACM Fellows* vždy umístily mezi téměř nejhorší autory. Každý sloupec diagramu je tedy shora ohraničen 3. kvartilem a zespodu 1. kvartilem, mezi kterými se nachází čára, která v našem případě zastupuje průměr (tj. ne obvyklý medián) z relativních pozic všech oceněných autorů ve vytvořeném pořadí autorů. Průměr jsme zobrazili proto, že jsme ho použili při vyhodnocení kvality našich metod. Dlouhá vodorovná čára v každé sekci značí nejnižší dosaženou průměrnou relativní pozici oceněných autorů v pořadí vytvořeném naší nejlepší metodou.



Obrázek 5.7: Porovnání metod navržených pro hodnocení autorů na základě relativních pozic oceněných autorů ve vytvořených pořadích (každý sloupec je ohraničen 1. a 3. kvartilem, čára uprostřed sloupce je průměr z relativních pozic všech oceněných autorů ve vytvořeném pořadí a dlouhá vodorovná čára v každé sekci značí nejnižší dosaženou průměrnou relativní pozici všech oceněných autorů).

Analýza našich výsledků ukázala, že:

- Vyhodnocením citační sítě publikací lze získat lepší pořadí autorů, než vyhodnocením citační sítě autorů (PageRank bez personalizace aplikovaný na citační síť publikací (5a2p) poskytl mírně lepší pořadí autorů, než naše nejlepší vyhodnocení citační sítě autorů metodou (4a) – metoda (5a2p) byla lepší o 0 až 2 procentní body). Výjimku tvoří porovnání výsledků metod se seznamem oceněných autorů *ACM HW*. Důvody odlišnosti našich metod při porovnání se seznamem *ACM HW* jsou diskutovány v závěru této části.
- Použití autorovy vyspělosti v personalizaci PageRanku (h-index, 3a) poskytlo průměrně o 10 procentních bodů horší pořadí autorů, než pokud byla v personalizaci použita jeho produktivita (počet publikací, 4a). Nicméně metoda (3a) byla průměrně o 3 procentní body lepší než baseline (2a).
- Samostatný h-index (1a) byl nejhorší metodou z námi použitých metod pro hodnocení autorů a byl tedy horší než pouhé počítání publikací (5a1p), což je zřejmě způsobeno malou rozlišovací schopností h-indexu. Jak zmiňuje část 5.2.1 a ukazuje obrázek 5.2, h-index byl schopen autory z celé naší kolekce WoS rozdělit pouze do 13 skupin, kdežto dle počtu publikací bylo vytvořeno 101 skupin autorů. Konkrétním příkladem může být, že nejlepších 105 autorů bylo h-indexem rozděleno pouze do 6 skupin, kdežto počtem publikací bylo 105

nejlepších autorů rozděleno do 51 skupin. To je důvod, proč říkáme, že h-index má malou rozlišovací schopnost.

Metody (5a1p – 5a8p), které používají významnosti publikací, ukázaly, že:

- Metoda (5a3p), která pro hodnocení autorů používá personalizaci publikací dle jejich popularity (počet citací), jako jediná poskytla horší pořadí autorů než příslušná baseline (5a2p). To nás vede k přesvědčení, že dodatečné přidání popularity do výpočtu prestiže má na hodnocení významnosti autorů negativní vliv, a proto je lepší do výpočtu prestiže zahrnovat jiné míry kvality. Až na výjimku ve sloupci *ACM HW* je metoda (5a3p) o 1 až 5 procentních bodů horší než baseline (5a2p).
- Použití libovolné významnosti časopisů (1j-3j) ve variantách metody (5a5p), která počítá produktivitu autorů, zlepšilo hodnocení autorů (5a1p byla porovnána s 5a5p, přičemž 5a5p je o 6 až 48 procentních bodů lepší než 5a1p).
- Metoda (5a7p) používající významnosti časopisů v personalizaci PageRanku, kterým vyhodnocuje citační síť publikací, poskytla nejlepší pořadí autorů ze všech námi testovaných metod. Tím se potvrdil náš hlavní předpoklad, že nejlepší hodnocení autorů získáme PageRankem využívajícím hodnoty časopisů.
- Nastavení vah vstupních hran publikací dle hodnot jejich časopisů (5a8p) vedlo, v porovnání s baseline (5a2p), k mírnému zlepšení výsledků (průměrně o 0,7 procentních bodů).
- Současné použití hodnot časopisů v personalizaci i vahách vstupních hran (5a6p), ale mírně zhoršilo naše nejlepší pořadí autorů vytvořené bez použití vah hran metodou (5a7p) – průměrné zhoršení o 0,6 procentních bodů.
- Pro úplnost: in-degree publikací poskytlo lepší pořadí autorů než in-degree autorů, ale i tak mělo průměrně o 24% horší výsledky než naše nejlepší metoda (5a7p). H-index měl výsledky horší průměrně o 32%.

Z výsledků hodnocení nelze jednoznačně určit, zda lepší pořadí autorů v kategoriích lze získat z category-independent nebo z category-dependent způsobu vyhodnocení. Protože ale category-dependent vyhodnocení mělo v kategorii *Hardware* větší úchylky, které mohou být způsobeny větším počtem izolovaných vrcholů a malým počtem citací v příslušných sítích (popř. malým počtem oceněných osob), tak pro další použití raději doporučujeme category-independent způsob vyhodnocení. Ten pro určení hodnoty významnosti autora využívá informace z celé citační sítě (obdobně jako Google.com při řazení výsledků vyhledávání) a autorova hodnota významnosti tak věrněji odpovídá jeho celkovému publikačnímu snažení. Na druhou stranu, může existovat autor, který je hodně věhlasný v celé použité datové kolekci, ale který např. v kategorii Hardware napsal pouze jednu publikaci. Tento autor bude v kategorii HW vyhledán a řazen velmi pozitivně (v category-dependent vyhodnocení by se to nestalo). Tuto vlastnost lze odstranit např. některou z následujících možností, které nebyly v našich experimentech testovány, a proto by mohly být námětem pro další práci:

- Hodnoty publikací jsou vypočteny z celé citační sítě publikací, ale hodnoty autorů jsou určeny pouze na základě publikací ze zvolené kategorie.

- Hodnoty autorů určené na základě všech publikací jsou vynásobeny podílem počtu publikací, které autor publikoval ve zvolené kategorii, vůči všem jeho publikacím. To by mělo zohlednit míru autorovy produktivity v dané kategorii.

Naše metody mají několik stupňů volnosti (různé váhy hran, způsoby rozdělení hodnot publikací autorům, typy samocitací, faktory tlumení apod.) a jejich výsledky, které jsme porovnali v tabulce 5.4, jsou nejlepšími výsledky, které lze těmito metodami získat, tj. byly získány konkrétním nastavením metod. To může vést k fenoménu, který je v oblastech statistiky a strojového učení znám jako přeučení (*overfitting*[59]), viz (Hawkins 2004). Nicméně díky skutečnosti, že jsme kvalitu získaných pořadí autorů určovali na základě tří rozdílných referenčních seznamů oceněných autorů a výsledky se příliš nelišily (viz tabulka 5.4 a dále tabulka 5.6), se přeučení neobáváme. Citlivost PageRanku na změnu jednotlivých parametrů detailněji zmiňují např. Langville a Meyer (2006) a část 2.4.3.

5.4.1 Diskuse výsledků metod, které pracují se sítí autorů

Metody pracující s citační sítí autorů nám neposkytly nejlepší pořadí autorů, ale pro úplnost v této části uvádíme, jak dle našich experimentů každou z těchto metod nastavit tak, aby poskytla co nejlepší možné pořadí autorů. Nejlepší je metody nastavit tak, jak je uvedeno v tomto seznamu:

(1a) H-index – samocitace odstranit na úrovni publikací (h-index bez samocitací).
(2a) PageRank bez personalizace – d=0,85.
(3a) PageRank s personalizací dle h-indexu autorů – d=0,45, h-index bez samocitací.
(4a) PageRank s personalizací dle počtu publikací autorů – d=0,55, samocitace odstraněny na úrovni publikací.

V seznamu nejlepších parametrů pro jednotlivé metody pracující s citační sítí autorů lze vidět, že pokud byla použita personalizace, tak bylo vhodné nastavit faktor tlumení na hodnotu blízkou 0,5, což potvrdilo závěry (Chen et al. 2007; Yan a Ding 2011), zmíněné v části 5.3. Seznam neobsahuje parametry, které nelze pro danou metodu jednoznačně určit (typ samocitací nebo vah hran), protože pro různé seznamy oceněných autorů se tyto parametry lišily. Proto zde zmíníme, jak jsou metody na tyto parametry citlivé. Rozdíly dané metody v pořadích autorů, která získáme změnou parametrů metody, lze demonstrovat Spearmanovými koeficienty pořadové korelace, počty shodných osob na nejlepších pozicích, popř. počtem oceněných osob na nejlepších pozicích. Na obrázku 5.8 je na základě Spearmanových koeficientů ukázáno, jak se změnilo danou metodou vytvořené pořadí autorů, když jsme metodě změnili parametry vyhodnocované sítě. Parametry, které nejsou na obrázku 5.8 zobrazeny, byly nastaveny dle předchozího seznamu. Na obrázku 5.8 jsou Spearmanovy koeficienty vynásobeny stem, a čím větší rozdíl v pořadích koeficienty představují, tím mají tmavší pozadí.

Ze Spearmanových koeficientů korelace lze vyvodit, že při vyhodnocení sítě autorů měl na vytvořené pořadí autorů větší vliv použitý typ samocitací než typ vah hran. Stejný závěr můžeme učinit, pokud pořadí porovnáme na základě počtu oceněných autorů na nejlepších pozicích. Zajímavé je, že pokud varianty metod porovnáme na základě počtu stejných autorů na nejlepších pozicích ve vytvořených

[59] Přeučený statistický model nebo přeučený model strojového učení obvykle poskytuje přesné predikce pro příklady z trénovací množiny (model si zapamatoval trénovací data), ale neposkytuje dobré predikce pro nové, dosud nepoužité příklady.

pořadích autorů, tak větší rozdíly mají varianty se stejným typem vah hran, nežli varianty se stejným typem samocitací, ale tomuto zjištění nepřikládáme příliš velký význam. Na obrázku 5.8 je dále vidět, že od všech pořadí autorů se nejvíce odlišovala pořadí získaná použitím sítě autorů s odstraněnými samocitacemi na úrovni publikací (*NOT*). S přihlédnutím ke skutečnosti, že pořadí autorů získaná použitím různých typů vah hran měla nejnižší korelaci 0,98, lze konstatovat, že na použitém typu vah hran v síti autorů příliš nezáleželo. Z rozdílů koeficientů korelace jednotlivých variant metody lze vyvodit, že nejvíce stabilní pořadí autorů poskytovala metoda s personalizací autorů dle počtu publikací (4a), jejíž varianty mají nejnižší korelaci 0,96. Největší rozdíly ve vytvořených pořadích autorů (korelace 0,88) měla metoda využívající PageRank bez personalizace (2a). Pořadí autorů získaná h-indexem se všemi citacemi a h-indexem bez samocitací měla korelaci 0,95, stejných všech deset nejlepších autorů a ze sta nejlepších autorů jich měla 95 stejných.

(2a) PageRank bez personalizace

		ALL			PART			NOT		
		N	1/N	1	N	1/N	1	N	1/N	1
ALL	N	x	99	100	99	98	98	89	88	89
	1/N	99	x	98	97	99	97	89	90	89
	1	100	98	x	99	98	99	90	88	90
PART	N	99	97	99	x	99	100	91	89	90
	1/N	98	99	98	99	x	98	90	91	90
	1	98	97	99	100	98	x	91	89	91
NOT	N	89	89	90	91	90	91	x	99	100
	1/N	88	90	88	89	91	89	99	x	99
	1	89	89	90	90	90	91	100	99	x

(3a) Personalizace h-indexem autorů

		ALL			PART			NOT		
		N	1/N	1	N	1/N	1	N	1/N	1
ALL	N	x	99	100	100	99	99	94	93	94
	1/N	99	x	99	99	100	98	94	94	93
	1	100	99	x	99	99	100	94	93	94
PART	N	100	99	99	x	99	100	94	93	94
	1/N	99	100	99	99	x	99	94	95	94
	1	99	98	100	100	99	x	94	94	95
NOT	N	94	94	94	94	94	94	x	99	100
	1/N	93	94	93	93	95	94	99	x	99
	1	94	93	94	94	94	95	100	99	x

(4a) Personalizace počtem publikací autorů

		ALL			PART			NOT		
		N	1/N	1	N	1/N	1	N	1/N	1
ALL	N	x	99	100	100	99	99	96	96	96
	1/N	99	x	99	99	100	98	96	97	96
	1	100	99	x	100	99	100	96	96	96
PART	N	100	99	100	x	99	100	97	96	96
	1/N	99	100	99	99	x	99	96	97	96
	1	99	98	100	100	99	x	97	96	97
NOT	N	96	96	96	97	96	97	x	99	100
	1/N	96	97	96	96	97	96	99	x	99
	1	96	96	96	96	96	97	100	99	x

Obrázek 5.8: Vliv změn parametrů metod pro hodnocení autorů na vytvořené pořadí autorů (Spearmanovy koeficienty pořadové korelace jsou vynásobeny stem a čím větší rozdíl v pořadích koeficienty představují, tím mají tmavší pozadí).

Při porovnání variant metod dle počtu oceněných autorů na 10 a 100 nejlepších pozicích ve vytvořených pořadích autorů byly nejlepší ty varianty vyhodnocení, které odstraňují samocitace na úrovni publikací. Při použití 1000 nejlepších pozic byly mírně lepší varianty s odstraněnými samocitacemi na úrovni autorů. To potvrzuje úsudek, že na samocitace by při vyhodnocení citačních sítí autorů neměl být brán zřetel, tj. měly by být odstraněny.

5.4.2 Diskuse výsledků metod, které pracují se sítí publikací

Výsledky experimentů s rozdělováním hodnot publikací autorům metodou (5a) ukázaly, že pro globální hodnocení autorů bylo v našem případě obvykle nejlepší rozdělení *DIV*, ale pro hodnocení v kategoriích bylo vždy nejlepší rozdělení *SUM*. Protože obvykle nejhorší rozdělení *GEOM* je dle Assimakis a Adam (2010) blízké způsobu hodnocení autorů, ve kterém jsou použiti pouze první autoři publikací, tak na základě našich výsledků můžeme říci, že je vždy lepší používat všechny autory publikace a každému z nich přiřadit stejnou hodnotu publikace – buď celou hodnotu (*SUM*) nebo rovnoměrný díl této hodnoty (*DIV*). Tato dvě rozdělení navíc neznevýhodňují autory, kteří jsou vlivem alfabetického výčtu autorů publikace odsunuti na horší pozice, přestože, jak zmiňuje Waltman (2012), mohli k vytvoření publikace přispět nejvíce.

Na obrázku 5.9 jsou porovnána jednotlivá rozdělení hodnot publikací jejich autorům v metodách (5a1p) a (5a5p), které počítají autorovu produktivitu. Pro porovnání metod byly použity Spearmanovy koeficienty korelace a v metodě (5a5p) byly jako hodnoty publikací použity PageRankové hodnoty jejich časopisů (1j). Poznamenat můžeme, že použití hodnot časopisů zlepšilo hodnocení autorů na základě produktivity. Z obrázku 5.9 lze vypozorovat, že rozdělení *DIV* má vyšší korelace

s neuniformními rozděleními než s rozdělením *SUM*, přestože rozdělení *SUM* a *DIV* měla při hodnocení autorů lepší výsledky než neuniformní rozdělení. To odpovídá závěru, že nemá smysl zvýhodňovat některé autory publikace.

		Publikace mají hodnotu 1 (5a1p)					Publikace mají hodnotu dle PageRanku časopisu (5a5p1j)				
		SUM	DIV	LIN	GEOM	GOLD	SUM	DIV	LIN	GEOM	GOLD
Hodnota 1 (5a1p)	SUM	x	81	70	73	76	58	54	51	50	51
	DIV	81	x	95	96	95	67	73	70	70	71
	LIN	70	95	x	98	96	64	73	76	77	77
	GEOM	73	96	98	x	99	65	71	73	74	75
	GOLD	76	95	96	99	x	64	69	71	72	74
PageRank časopisu (5a5p1j)	SUM	58	67	64	65	64	x	95	91	89	87
	DIV	54	73	73	71	69	95	x	97	96	94
	LIN	51	70	76	73	71	91	97	x	99	98
	GEOM	50	70	77	74	72	89	96	99	x	100
	GOLD	51	71	77	75	74	87	94	98	100	x

Obrázek 5.9: Porovnání vlivu způsobu rozdělení hodnot publikací autorům v metodách (5a1p) a (5a5p) na vytvořené pořadí autorů (Spearmanovy koeficienty korelace jsou vynásobeny stem a čím větší rozdíl v pořadích koeficienty představují, tím mají tmavší pozadí).

Jak již bylo zmíněno dříve, hodnoty časopisů bylo nejlepší použít jako personalizace publikací. Jejich současné použití ve váhách vstupních hran vrcholů výsledky hodnocení autorů mírně zhoršilo (viz tabulka 5.4). Pro seznam oceněných autorů z kategorie *AI* bylo zanedbatelně lepší použít Impact Factor časopisů (2j), ale pro ostatní seznamy oceněných autorů bylo nejlepší použít PageRank časopisů (1j). Přesto, že získaná pořadí autorů se příliš nelišila, jsme doporučili používat při hodnocení autorů PageRank časopisů. Použití hodnot PageRanku, které byly vypočteny z celé citační sítě časopisů (1j), také poskytlo lepší výsledky, než použití hodnot 3 years PageRanku (3j), což je zřejmě způsobeno větším počtem použitých dat. Porovnání pořadí časopisů, která byla vytvořena Impact Factorem a 3 years PageRankem, obsahuje obrázek 5.10. Nejnižší korelace rozdílných metod jsou mezi Impact Factorem *NOT* a 3 years PageRankem *ALL* (průměrná hodnota korelace je 0,55) a nejvyšší mezi Impact Factorem *NOT* a 3 years PageRankem *NOT* (průměrná hodnota korelace je 0,70).

rok	1998				1999				2000				2001			
M.	PR_A	PR_N	IF_A	IF_N	PR_A	PR_N	IF_A	IF_N	PR_A	PR_N	IF_A	IF_N	PR_A	PR_N	IF_A	IF_N
PR_A	x	87	71	59	x	88	66	53	x	85	68	55	x	86	60	48
PR_N	87	x	73	75	88	x	69	69	85	x	64	68	86	x	62	65
IF_A	71	73	x	88	66	69	x	87	68	64	x	86	60	62	x	87
IF_N	59	75	88	x	53	69	87	x	55	68	86	x	48	65	87	x
rok	2002				2003				2004				2005			
M.	PR_A	PR_N	IF_A	IF_N	PR_A	PR_N	IF_A	IF_N	PR_A	PR_N	IF_A	IF_N	PR_A	PR_N	IF_A	IF_N
PR_A	x	86	63	53	x	89	66	57	x	88	66	55	x	88	65	56
PR_N	86	x	68	70	89	x	67	71	88	x	66	69	88	x	66	71
IF_A	63	68	x	87	66	67	x	88	66	66	x	86	65	66	x	87
IF_N	53	70	87	x	57	71	88	x	55	69	86	x	56	71	87	x

Obrázek 5.10: Korelace pořadí časopisů, která byla vytvořena dle Impact Factoru a dle 3 years PageRanku (PR_A – 3 years PageRank ALL; PR_N – 3 years PageRank NOT; IF_A – Impact Factor ALL; IF_N – Impact Factor NOT; koeficienty jsou vynásobeny stem a nejnižší korelace mají černá pozadí).

Použití samocitací při vyhodnocení citační sítě časopisů mělo na hodnocení autorů minimální vliv. To bylo zřejmě způsobeno skutečností, že použitá síť časopisů měla málo vrcholů (pouze 386), mezi které byl vždy rozdělen velký počet citací (buď 191 447 všech citací, nebo 145 372 citací, pokud byly samocitace autorů odstraněny na úrovni publikací). Odstranění samocitací autorů v citační síti časopisů ovlivnilo převážně jen váhy hran (ze sítě časopisů bylo odstraněno pouze 353 z 20 488 hran) a to nejspíš tak, že významné časopisy zůstaly stále významné. Protože ocenění autoři jistě publikovali ve významných časopisech, tak se jejich pozice ve vytvořeném pořadí autorů příliš nezměnila. Proto zřejmě příliš nezáleží na použitém typu samocitací autorů v síti časopisů. Nicméně v našich experimentech s globálním hodnocením autorů v počítačových vědách bylo mírně lepších výsledků dosaženo použitím všech citací (tj. i samocitací) v síti časopisů.

Pro úplnost jsou na obrázku 5.11 porovnány metody pro hodnocení autorů, které PageRankem vyhodnocují citační síť publikací s odstraněnými samocitacemi autorů (*NOT*). Faktor tlumení je na základě našich nejlepších variant pro hodnocení autorů nejlepší nastavit následovně:

- 5a2p, 5a8p → *d*=0,85
- 5a3p, 5a4p → *d*=0,75
- 5a6p, 5a7p → *d*=0,55

Lze vidět, že v metodách (5a6p) a (5a7p), které v personalizaci používají hodnoty časopisů, byla potvrzena vhodnost použití faktoru tlumení s hodnotou přibližně 0,5. Při nastavení personalizace dle počtu citací publikace (5a3p) nebo dle počtu autorů publikace (5a4p) byly nejlepší výsledky získány s větším faktorem tlumení (*d*=0,75). To znamená, že personalizace dle hodnot časopisů přispívá k lepšímu hodnocení autorů více, než personalizace dle počtu citací nebo autorů.

Z koeficientů korelace (viz obrázek 5.11) je patrné, že metody (5a6p) a (5a7p) využívající personalizaci publikací dle hodnot časopisů se od ostatních metod odlišovaly nejvíce. Tyto metody měly vyšší korelaci s metodou, která nepoužívá personalizaci, ale používá hodnoty časopisů jako váhy vstupních hran (5a8p), než s metodami (5a3p) a (5a4p) používajícími jiné personalizace. Tento závěr potvrdila i porovnání vytvořených pořadí autorů na základě počtu shodných autorů na nejlepších pozicích. Metody (5a6p) a (5a7p) také poskytly téměř totožná pořadí autorů (korelace 0,9985).

		SUM						DIV					
		5a2p	5a8p	5a3p	5a4p	5a6p	5a7p	5a2p	5a8p	5a3p	5a4p	5a6p	5a7p
SUM	5a2p	x	97	97	97	87	88	91	89	90	94	80	81
	5a8p	97	x	94	97	91	91	91	91	89	94	83	84
	5a3p	97	94	x	94	86	87	87	85	92	90	78	79
	5a4p	97	97	94	x	87	88	85	83	84	89	75	76
	5a6p	87	91	86	87	x	100	82	83	83	85	92	91
	5a7p	88	91	87	88	100	x	82	83	83	86	91	91
DIV	5a2p	91	91	87	85	82	82	x	99	97	99	90	90
	5a8p	89	91	85	83	83	83	99	x	96	99	92	92
	5a3p	90	89	92	84	83	83	97	96	x	97	89	89
	5a4p	94	94	90	89	85	86	99	99	97	x	90	90
	5a6p	80	83	78	75	92	91	90	92	89	90	x	100
	5a7p	81	84	79	76	91	91	90	92	89	90	100	x

Obrázek 5.11: Porovnání pořadí autorů, která byla vytvořena metodami pro hodnocení autorů pracujícími s citační sítí publikací (Spearmanovy koeficienty korelace jsou vynásobeny stem a čím větší rozdíl v pořadích koeficienty představují, tím mají tmavší pozadí).

5.4.3 Nejlepší autoři ve vytvořených pořadích autorů

Oblíbeným způsobem prezentování výsledků metod pro hodnocení autorů je tabulka, která zobrazuje několik nejlepších pozic z pořadí autorů vytvořeného testovanou metodou hodnocení. My jsme na ukázku zvolili pět metod, jejichž parametry jsou nastaveny dle variant, které ve vytvořeném pořadí autorů nejvíce vyzdvihly autory oceněné *ACM Fellows* (nejvíce oceněných autorů). Poznamenat můžeme, že seznam *ISI Highly Cited* měl ve vytvořených pořadích autorů nejvíce oceněných autorů na nejlepších 10 pozicích, což může být způsobeno tím, že seznam *ISI Highly Cited* vznikl na základě citační analýzy. V tabulce 5.5 je pro každou vybranou metodu zobrazeno prvních patnáct pozic z vytvořeného pořadí autorů. Ve všech vybraných metodách jsou v sítích autorů a publikací samocitace autorů vždy odstraněny na úrovni publikací. Pokud je použita síť časopisů, tak obsahuje všechny citace. Zvoleny byly metody s následujícími parametry:

(5a7p1j) V síti publikací je aplikován PageRank s d=0,55 a s personalizací nastavenou dle PageRankových hodnot časopisů. Hodnoty publikací jsou převedeny na autory rozdělením *SUM*.

(5a6p1j) V síti publikací je aplikován PageRank s d=0,55 a s personalizací a vahami vstupních hran nastavenými dle PageRankových hodnot časopisů. Hodnoty publikací jsou převedeny na autory rozdělením *SUM*.

(5a8p1j) V síti publikací je aplikován PageRank bez personalizace s d=0,85 a s vahami vstupních hran nastavenými dle PageRankových hodnot časopisů. Hodnoty publikací jsou převedeny na autory rozdělením *DIV*.

(5a4p) V síti publikací je aplikován PageRank s d=0,75 a s personalizací nastavenou dle počtu autorů publikace. Hodnoty publikací jsou převedeny na autory rozdělením *DIV*.

(4a) V síti autorů s typem vah hran *1/N* je aplikován PageRank s d=0,55 a s personalizací nastavenou dle počtu publikací autora.

Z tabulky 5.5 je opět patrné, že metody (5a7p1j) a (5a6p1j), které používají personalizaci publikací dle hodnot časopisů, se od ostatních metod značně liší. Například *HOLZMANN*, který byl metodou (4a) zařazen na 6. pozici, byl metodou (5a7p1j) zařazen na 1145. pozici. Naopak o sloupci metody (5a8p1j), která používá hodnoty časopisů pouze jako váhy vstupních hran, lze říci, že nejlépe sjednocuje ostatní sloupce, protože autoři, kteří jsou v ostatních sloupcích mezi patnácti nejlepšími autory, jsou zde umístěni na nejlepších 63. pozicích. Pokud v tabulce 5.5 vyhledáme oceněné autory (označeni *), tak zjistíme, že nejvíce oceněných autorů na patnácti nejlepších pozicích ve vytvořeném pořadí autorů poskytlo vyhodnocení sítě autorů metodou (4a). Z toho plyne, že naše nejlepší metoda (5a7p1j) určená na základě průměrných pozic oceněných autorů neměla v porovnání s ostatními použitými metodami nejvíce oceněných autorů na patnácti nejlepších pozicích. Nicméně ale poskytla pořadí autorů, ve kterém byla celá skupina tvořená všemi námi použitými oceněnými autory vyzdvižena nejvíce. Toto zjištění aktuálně nepovažujeme za příliš podstatné, ale v budoucnu by jistě bylo zajímavé zjistit, co je příčinou tohoto jevu.

*Tabulka 5.5: Patnáct nejlepších pozic z pořadí autorů, která byla vytvořena zvolenými metodami. (Ocenění autoři jsou zvýrazněni * a autoři, kteří jsou alespoň jednou mezi třemi nejlepšími autory, jsou zvýrazněni tučným písmem. Autoři, kteří nejsou mezi patnácti nejlepšími autory v daném sloupci, ale jsou mezi patnácti nejlepšími autory v jiném sloupci, jsou vypsáni ve spodní části tabulky.)*

	5a7p1j	5a6p1j	5a8p1j	5a4p	4a
1.	**JAIN, AK***	**JAIN, AK***	**SIMON, DR**	**BREIMAN, L***	**BREIMAN, L***
2.	**OSHER, S**	**OSHER, S**	**BREIMAN, L***	**JAIN, AK***	**JAIN, AK***
3.	**PEDRYCZ, W**	**PEDRYCZ, W**	**JAIN, AK***	**MOLTENBREY, K**	**ZADEH, LA***
4.	WANG, J	AMARI, S	**MOLTENBREY, K**	YAGER, RR*	HYVARINEN, A
5.	KIM, J	WANG, J	YAGER, RR*	**SIMON, DR**	BURGES, CJC
6.	AMARI, S	KIM, J	VAZIRANI, U	ROBERTSON, B	HOLZMANN, GJ*
7.	YAGER, RR*	YAGER, RR*	BERNSTEIN, E	**ZADEH, LA***	YAGER, RR*
8.	TANAKA, K	TANAKA, K	**ZADEH, LA***	**PEDRYCZ, W**	AMARI, S
9.	KIM, JH	KITTLER, J	ROBERTSON, B	CHANG, CC	PAXSON, V*
10.	YAN, H	LI, J	**PEDRYCZ, W**	WANG, J	OJA, E
11.	LI, J	ZHU, SC	AMARI, S	HYVARINEN, A	CIMINO, JJ
12.	SAPIRO, G	KIM, JH	DIETTERICH, TG*	LEE, J	TANAKA, K
13.	WANG, Y	YAN, H	HYVARINEN, A	AMARI, S	PENTLAND, A
14.	LEE, J	KANADE, T*	TANAKA, K	LEE, S	HARTLEY, RI
15.	KITTLER, J	LEE, J	CHANG, CC	KIM, J	**PEDRYCZ, W**
	(19) CHANG, CC	(16) SAPIRO, G	(16) WANG, J	(16) OJA, E	(16) DIETTERICH,..*
	(22) KANADE, T*	(17) WANG, Y	(17) LEE, J	(17) TANAKA, K	**(21) SIMON, DR**
	(23) ZHU, SC	(24) CHANG, CC	(19) OJA, E	(18) BURGES, CJC	(23) KIM, J
	(25) LEE, S	(27) PENTLAND, A	(21) KIM, J	(20) KIM, JH	(27) LEE, J
	(26) BREIMAN, L*	(29) LEE, S	(22) LEE, S	(21) VAZIRANI, U	(29) KANADE, T*
	(29) OJA, E	**(34) BREIMAN, L***	(24) HARTLEY, RI	(22) YAN, H	(30) WANG, J
	(31) PENTLAND, A	(35) OJA, E	(25) HOLZMANN,..*	(23) CIMINO, JJ	(33) SAPIRO, G
	(55) HYVARINEN, A	(52) HYVARINEN, A	(28) YAN, H	(24) WANG, Y	(35) CHANG, CC
	(114) HARTLEY, RI	(88) HARTLEY, RI	(29) KIM, JH	(26) BERNSTEIN, E	(37) ZHU, SC
	(124) ZADEH, LA*	**(118) ZADEH, LA***	(31) PAXSON, V*	(27) DIETTERICH,..*	(38) LEE, S
	(140) MOLTENBREY, K	(138) VAZIRANI, U	(34) LI, J	(29) LI, J	(46) YAN, H
	(150) VAZIRANI, U	**(153) MOLTENBREY, K**	(39) PENTLAND, A	(30) HOLZMANN,..*	(49) KIM, JH
	(235) BERNSTEIN, E	(223) BERNSTEIN, E	(43) WANG, Y	(40) PENTLAND, A	(52) KITTLER, J
	(239) BURGES, CJC	(347) ROBERTSON, B	(48) CIMINO, JJ	(41) PAXSON, V*	(53) LI, J
	(311) ROBERTSON, B	(426) CIMINO, JJ	(49) BURGES, CJC	(46) HARTLEY, RI	(57) WANG, Y
	(362) CIMINO, JJ	(494) PAXSON, V*	(51) ZHU, SC	(51) SAPIRO, G	**(59) OSHER, S**
	(611) PAXSON, V*	(540) BURGES, CJC	**(55) OSHER, S**	(54) KITTLER, J	**(77) MOLTENBREY, K**
	(616) DIETTERICH, TG*	(556) DIETTERICH, TG*	(57) KANADE, T*	**(59) OSHER, S**	(87) VAZIRANI, U
	(642) SIMON, DR	**(590) SIMON, DR**	(59) KITTLER, J	(76) KANADE, T*	(111) BERNSTEIN, E
	(1145) HOLZMANN,..*	(997) HOLZMANN, GJ*	(63) SAPIRO, G	(91) ZHU, SC	(164) ROBERTSON, B

5.4.4 Predikce laureátů významných ocenění

Protože naše seznamy oceněných autorů obsahovaly autory oceněné po roce 2005 (poslední rok obsažený v naší kolekci WoS), tak jsme otestovali, jak jsou tito autoři hodnoceni našimi metodami. Zajímalo nás, zda by naše metody pro hodnocení autorů mohly být použity pro predikci v budoucnu oceněných autorů. V celé naší kolekci WoS z let 1996 až 2005 jsme nalezli 206 držitelů *ACM Fellows*, kteří byli oceněni po roce 2005. V kategorii *Umělá inteligence* jsme nalezli 62 a v kategorii *Hardware* 49 autorů oceněných po roce 2005. Výsledky našich metod pro hodnocení autorů jsou využitím seznamů „v budoucnu“ oceněných autorů porovnány v tabulce 5.6, která má stejnou strukturu jako tabulka 5.4.

Tabulka 5.6: Porovnání kvality našich metod při hodnocení autorů oceněných v letech 2006 až 2014 (sloupce p. zobrazují pořadí úspěšnosti jednotlivých metod a sloupce $m_{\%}$ jejich procentuální odlišnost od minimální dosažené průměrné pozice oceněných autorů v pořadí autorů vytvořeném naší nejlepší metodou s p.=1 a jejíž hodnota je uvedena v řádku m_{best}).

Pouze autoři ocenění v letech 2006 až 2014											
Síť		Metoda (Oceněno - Hodnoceno -)	Global		Category-indep.				Category-dep.		
			Fellows 206 / 157440		AI 62 / 39891		HW 49 / 29243		AI 62 / 39891		HW 49 / 29243	
			p.	$m_{\%}$	p.	$m_{\%}$	p.	$m_{\%}$	p.	$m_{\%}$	p.	$m_{\%}$
Autoři	1a	H-index	12	40%	11	23%	11	26%	12	57%	8	15%
	2a	PageRank bez personalizace	11	25%	12	25%	10	23%	11	45%	9	16%
	3a	Personalizace h-indexem	9	24%	10	19%	8	15%	10	45%	6	10%
	4a	Personalizace počtem publikací	**4**	**3%**	**7**	**9%**	**7**	**13%**	**7**	**21%**	**5**	**9%**
Publikace	5a1p	Všechny mají hodnotu 1	8	20%	8	15%	12	29%	9	35%	12	56%
	5a2p	PageRank bez personalizace	7	4%	5	7%	6	12%	5	17%	10	17%
	5a3p	Personalizace počtem citací	5	3%	4	6%	5	11%	4	14%	**1**	**0%**
	5a4p	Personalizace počtem autorů	**3**	**0,6%**	**3**	**5%**	**3**	**3%**	**3**	**13%**	11	19%
	5a5p	Hodnoty dle hodnot časopisů	10	25%	9	17%	9	18%	8	22%	4	7%
	5a6p	Hodnoty časopisů v pers. i hranách	**1**	**0%**	2	3%	**1**	**0%**	2	2%	**2**	**3%**
	5a7p	Hodnoty časopisů v personalizaci	2	0,04%	**1**	**0%**	2	0,6%	**1**	**0%**	3	5%
	5a8p	Hodnoty časopisů ve vstup. hranách	6	4%	6	7%	4	10%	6	18%	7	15%
Minimální průměrná pozice oceněných autorů (m_{best})			17973		7404		5466		7623		5579	

Protože z tabulky 5.6 lze vyvodit téměř stejné závěry jako z tabulky 5.4, tak lze konstatovat, že naše nejlepší metoda (5a7p1j) pro hodnocení autorů (popř. metoda (5a6p1j)) by mohla být použita i pro predikci laureátů významných ocenění. Výsledky metod se více liší pouze u metody s personalizací PageRanku dle popularity publikací (počet citací, 5a3p), která byla v tomto případě lepší než PageRank bez personalizace (5a2p), což při použití všech osob ze seznamů oceněných autorů neplatilo.

5.4.5 Je prestiž lepší než popularita?

Nejen naše výsledky dokládají, že prestiž je lepší mírou pro hodnocení autorů než popularita, viz např. (Ding 2011a; Ding a Cronin 2011). To ale zpochybňují Fiala et al. (2015), kteří říkají, že: *„není důkaz o tom, že metody pro hodnocení autorů, které jsou podobné PageRanku, překonávají jednodušší počítání citací“*. V reakci bychom toto chtěli krátce diskutovat. Hlavním problémem dle našeho názoru je, že pro vyhodnocování kvality metod pro hodnocení autorů neexistuje referenční pořadí autorů, a proto si ho každý, kdo chce metody vyhodnotit, obvykle vyrábí sám. Sidiropoulos a Manolopoulos (2006) k tomuto účelu použili držitele Coddovy ceny, tj. držitele významného či prestižního ocenění, a práce, které po nich následovaly, také často používaly držitele různých významných ocenění (což je i náš případ). Vedle toho vznikly i práce, např. (Liu et al. 2005), které pro vyhodnocení používaly členy programových výborů zvolených konferencí.

My si myslíme, že hlavním závěrem, ke kterému Fiala et al. (2015) dospěli, je „pouze“, že jimi zvolení členové redakčních rad vybraných časopisů jsou lépe hodnoceni mírami popularity (počítání citací) než mírami prestiže (metody založené na PageRanku). Ale, jak sami autoři ukazují, držitelé významné Turingovy ceny jsou lépe hodnoceni mírami prestiže. Z toho lze vyvodit, že držitelé významných ocenění jsou správně považováni za prestižní a jsou dobře využitelní pro vyhodnocení metod měřících prestiž. Kdežto členy redakčních rad časopisů lze považovat spíše za populární. Správnost použití oceněných autorů pro vyhodnocení metod měřících prestiž dokládají, vedle např. Ding (2011a), Fiala (2012), Sidiropoulos a Manolopoulos (2006) a dalších, také naše experimenty. Jedním z našich závěrů je, že metody počítající citace publikací poskytují lepší pořadí autorů (měřeno oceněnými autory), než metody počítající citace autorů, ale i tak jsou o 24% horší než naše nejlepší metoda používající PageRank (5a7p1j). Určitý vliv na kvalitu vyhodnocení mohou mít také malé počty autorů na referenčních seznamech autorů (32 v kategorii *Umělá inteligence*, 12 v kategorii *Softwarové inženýrství* a 17 v kategorii *Teorie a metody*), které Fiala et al. (2015) použili.

5.5 Závěry z testování nově navržených metod pro hodnocení autorů

V této kapitole jsme popsali naše experimenty s novými metodami pro hodnocení autorů vědeckých publikací na základě citační analýzy, které jsme původně publikovali v (Nykl et al. 2015). Ze zakoupené bibliografické kolekce ISI Web of Science (kategorie počítačových věd, 1996-2005) jsme vytvářeli citační sítě publikací, autorů a časopisů a hledali jsme metodu pro hodnocení autorů, která poskytuje pořadí autorů nejbližší lidskému úsudku. Pro tyto účely jsme testovali metody, které vyhodnocují sítě autorů, sítě publikací nebo sítě publikací, které byly obohaceny o hodnoty časopisů. V metodách jsme také zkoušeli zohlednit další ukazatele kvality autora nebo publikace, kterými byly h-index autora, počet publikací autora, počet citací publikace a počet autorů publikace. Kvalitu získaných pořadí autorů (vytvořena dle jejich hodnot) jsme stanovili na základě průměrné pozice oceněných autorů, což nám umožnilo pořadí porovnat. Pro určení kvality metod pracujících s citačními sítěmi vytvořenými z celé oblasti počítačových věd jsme použili seznamy *ACM Fellows* a *ISI Highly Cited*. Abychom otestovali, jak naše metody hodnotí autory ve specializovaných kategoriích počítačových věd, tak jsme použili WoS kategorie *Umělá inteligence* a *Hardware* a vytvořená pořadí autorů jsme porovnali na základě držitelů ocenění udílených od SIGs z odpovídajících kategorií ACM. Autory ve zvolených kategoriích jsme navíc zkoušeli hodnotit dvěma způsoby aplikování našich metod a to:

- category-independent způsobem – v pořadí autorů získaném globálním hodnocením autorů v počítačových vědách byli nalezeni autoři, kteří patří do zvolené kategorie, a z nich bylo vytvořeno nové pořadí.
- category-dependent způsobem – pracuje pouze s publikacemi, které patří do zvolené kategorie.

Výsledky ukázaly, že pro hodnocení autorů je v našem případě nejlepší metoda (5a7p1j), která:

- používá PageRank, namísto neiteračních postupů (počet publikací, h-index, in-degree);
- vyhodnocuje citační síť publikací s odstraněnými samocitacemi autorů, namísto sítě autorů;
- používá PageRankové hodnoty časopisů v personalizaci publikací, nežli jiné personalizace;
- hodnoty autorů stanovuje sčítáním rovnoměrných dílů z hodnot publikací (stanoveny dle počtu autorů publikace), nežli sčítáním nerovnoměrných dílů z hodnot publikací.

Zajímavým zjištěním je, že tyto závěry potvrdilo i vyhodnocení, při kterém jsme použili pouze autory oceněné po roce 2005 (poslední rok obsažený v naší kolekci WoS). Proto na základě výsledků můžeme říci, že naše nejlepší metoda pro hodnocení autorů by mohla být stejně dobře prospěšná pro predikci laureátů významných ocenění.

Pro hodnocení autorů ve specializovaných kategoriích jsme doporučili použití category-independet způsobu hodnocení (používá celou datovou kolekci), ale zmínili jsme potřebu dále experimentovat s možnostmi získání hodnot autorů v rámci tohoto hodnocení. Při získávání hodnot autorů by se měla více zohlednit jejich produktivita v dané kategorii a tím eliminovat nárůst hodnot autorů vlivem publikací spadajících do jiných kategorií počítačových věd.

Experimenty dále ukázaly, že nastavení personalizace dle popularity publikací (počet citací) nebo vyspělosti autorů (h-index) neposkytuje lepší hodnocení autorů, než personalizace použité ve 4. kapitole (počet autorů publikace a počet publikací autora). Použití h-index personalizace sice poskytlo lepší pořadí autorů, než nepoužití personalizace, ale vytvořené pořadí autorů bylo horší než pořadí, které bylo získáno personalizací PageRanku dle produktivity autorů (počet publikací). Personalizace dle počtu citací publikace poskytla horší výsledky než PageRank bez personalizace, proto lze říci, že dodatečné přidání popularity (počet citací) do výpočtu prestiže je nežádoucí.

Protože v této kapitole byl popsán náš aktuálně poslední výzkum v oblasti hodnocení významnosti autorů a v další kapitole je popsán náš experiment s využitím PageRanku v úloze zpracování textů, tak na závěr této kapitoly můžeme říci, že další práce v oblasti hodnocení autorů by mohly být zaměřeny na experimentování:

- s různými variantami získání hodnot autorů v category-independent způsobu hodnocení;
- s kritériem stanovení kvality vytvořených pořadí autorů;
- s využitím klasifikačního systému (např. *ACM Computing Classification System*[60]) nebo ontologie (např. *DBpedie*[61]), které by umožnily specifičtější vyhledání autorů;
- s hodnocením významnosti pracovních skupin, oddělení či institucí.

[60] Web *ACM Computing Classification System* - http://www.acm.org/about/class/
příklad CCS z roku 2012 - http://dl.acm.org/ccs_flat.cfm
[61] Web *DBpedie* - http://www.DBpedia.org
příklad *Data mining* - http://DBpedia.org/page/Category:Data_mining

6 PageRank jako podpůrný nástroj při klasifikaci dokumentů

V této kapitole je popsán náš výzkum prezentovaný v (Nykl et al. 2013) a mírně rozšířený v (Dostal et al. 2014a)[62]. V článcích jsme publikovali náš nový přístup ke klasifikaci dokumentů, který klasifikaci obohatil o sémantické informace získané PageRankem z Linked Data. Alternativním zdrojem sémantických informací mohla být ontologie. Protože naše metoda pro volbu vlastností (*feature selection*) či klíčových slov dokumentů využívá sémantické informace a vlastnosti jsou v lidmi čitelné a pochopitelné formě, tak tyto vlastnosti mohou být interpretovány i neprofesionálními uživateli. V této kapitole bude větší důraz kladen na popsání úlohy PageRanku, který je použit nejprve pro rozšíření základních vlastností dokumentu (zvolených použitím TF-IFD, viz část 6.4) o obecnější vlastnosti, k čemuž používá sémantickou síť získanou z Linked Data, a následně pro zvolení nejreprezentativnějších vlastností dokumentu. Navržená metoda pro volbu vlastností dokumentů může být zakomponována do fáze učení i do fáze klasifikace klasických klasifikačních algoritmů. My jsme v článcích publikovali jednoduchou použitelnost navržené metody v klasifikátoru dokumentů a její slibné výsledky, které jsme získali při experimentech s volně dostupnou kolekcí diskusních článků *20 Newsgroups* a s vlastní kolekcí konferenčních *Call-for-Papers*.

V následující části 6.1 je zmíněn stručný úvod do klasifikace dokumentů a relevantní práce z oblasti klasifikace dokumentů (část 6.1.1) a z oblasti použití PageRanku pro zpracování přirozeného jazyka (část 6.1.2). Základní principy Linked Data jsou popsány v části 6.2 a datové kolekce, které jsme použili pro vyhodnocení kvality klasifikace dokumentů, v části 6.3. Naše metoda pro volbu vlastností dokumentů s využitím Linked Data a PageRanku je navržena v části 6.4. Diskuse kvality naší metody je provedena v části 6.5 a závěry shrnuty v části 6.6.

6.1 Úvod do klasifikace dokumentů

Klasifikace dokumentů je důležitou součástí systémů pro správu dokumentů i jiných služeb zpracovávajících texty. Jejím úkolem je zařadit dokumenty na základě obsahu do odpovídajících klasifikačních tříd. K tomu účelu jsou používány metody strojového učení s učitelem, které si ve fázi učení na základě učitelem předpřipravené množiny „trénovacích“ dokumentů nastaví vnitřní funkce, s jejichž využitím následně ve fázi klasifikace dokáží dosud nespatřené dokumenty zařadit do odpovídající klasifikační třídy. Pro vyhodnocení kvality klasifikace se používá učitelem předpřipravená množina „testovacích“ dokumentů. Důležitým faktorem ovlivňujícím kvalitu klasifikace je zvolená metoda předzpracování dokumentů a volby vlastností, které každý dokument v průběhu klasifikace reprezentují. Více základních informací o klasifikaci a dalších úlohách získávání informací z textů lze nalézt např. v (Manning et al. 2008).

[62] Protože v tomto článku jsem nebyl prvním autorem, tak považuji za vhodné uvést, že klasifikací dokumentů se zabýval Martin Dostal a já se zabýval určováním významných vrcholů v grafu slov. Práce na článku byla rozdělena tímto způsobem: Martin Dostal předzpracoval kolekce dokumentů a provedl mapování základních vlastností dokumentů na vrcholy v Linked Data. Já navrhl metodu pro určení klíčových slov dokumentu využitím PageRanku a Linked Data a aplikoval jsem ji na dokumenty z kolekcí. Určená klíčová slova byla použita jako vlastnosti dokumentů. Martin Dostal následně provedl experiment s klasifikací dokumentů a vyhodnotil kvalitu navržené metody. Text článku byl převážně dílem Marina Dostala, ale na konferenci jsem článek prezentoval já. Profesor Karel Ježek byl v pozici školitele odborným dozorem. V rámci této kapitoly budou navíc uvedeny doplňující informace o metodě pro volbu klíčových slov, se kterou jsem já experimentoval.

Současné metody klasifikace jsou obvykle statistické, a proto ve fázi učení vyžadují velké množství dat. Ale příprava dostatečně reprezentativní množiny trénovacích dokumentů a výběr metody pro volbu vlastností, které dokumenty zastupují, jsou náročné úkoly i pro doménové specialisty. Problém s vytvořením množiny trénovacích dokumentů bývá často řešen použitím relativně ucelených a vyvážených korpusů, které obsahují velké množství dokumentů rozdělených do odpovídajících klasifikačních tříd či kategorií. Statistické metody ve fázi učení určují vztahy mezi vlastnostmi dokumentů, které reprezentují v dokumentu obsažené významné termíny, a klasifikačními třídami. Naše metoda volby vlastností tento postup vylepšuje tak, že v dokumentu nalezené významné termíny využitím sémantických informací získaných z Linked Data doplňuje o obecnější termíny, např.: na základě nalezení termínu *MySQL* můžeme vlastnosti dokumentu doplnit o termín *Databáze* bez explicitního výskytu tohoto slova v obsahu dokumentu. Odvozené obecnější termíny dle našeho předpokladu následně umožnily vytvoření lepších vztahů mezi vlastnostmi dokumentů a klasifikačními třídami a tím zlepšily klasifikaci dokumentů.

6.1.1 Relevantní práce z oblasti klasifikace dokumentů

Pro klasifikaci dokumentů existuje mnoho metod strojového učení s učitelem, které v sobě obsahují algoritmy jako Naive Bayes, Support Vector Machine (Burges 1998), Boosting (Schapire a Singer 2000) anebo používají pravidla (Cohen a Singer 1999), latentní sémantickou analýzu (Deerwester et al. 1989), vektorové prostory (Salton 1971) či postupy (např.: algoritmy Rocchio (Rocchio 1971) a k-nejbližších sousedů (Cover a Hart 1967)) nebo maximální entropii.

My za zdroj Linked Data prezentovaných v této kapitole zvolili DBpedii[63], což je sémanticky obohacená Wikipedie[64]. Kvůli efektivnosti byla použita její lokální kopie uložená v relační databázi. Alternativou pro přístup k Linked Data by bylo použití SPARQL endpointu[65]. DBpedie byla již dříve úspěšně použita pro výpočet sémantické podobnosti dokumentů, viz např. algoritmus WikiRelate! (Strube a Ponzetto 2006), který kombinuje metody založené na cestách, obsahu a textovém překrytí. Na Wikipedii založená explicitní sémantická analýza (Gabrilovich a Markovitch 2007) používá techniky strojového učení pro přímou reprezentaci významu textu ve formě váženého vektoru.

Wang et al. (2005) ve své metodě pro klasifikaci dokumentů navrhli term-graf model, což je vylepšená verze modelu vektorového prostoru, viz (Salton 1971). Cílem tohoto modelu je reprezentovat obsah dokumentu s využitím vztahů mezi klíčovými slovy, což následně umožňuje definovat míry podobnosti a použít PageRanku podobné algoritmy. Vektor PageRankových hodnot je vypočítán pro každý dokument a hodnoty korelace a vzdálenosti termínů jsou použity pro zařazení dokumentu do odpovídající klasifikační třídy. Alternativní metody klasifikace dokumentů používají nadřazená obecnější slova a další přímo související koncepty (Ramakrishnanan a Bhattacharyya 2003; Bloehdorn a Hotho 2004). Další metoda rozšiřuje vlastnosti dokumentu na základě přidání

[63] Web *DBpedie* – http://www.DBpedia.org

[64] Web *Wikipedie* - http://www.wikipedia.org

[65] SPARQL (*Simple Protocol and RDF Query Language*) je sémantický dotazovací jazyk pro data uchovaná ve formátu RDF, který byl standardizován konsorciem W3C. Jako Endpointy jsou označovány webové stránky či služby, které nad poskytnutými daty umožňují vykonávat SPARQL dotazy.

RDF (*Resource Description Framework*) se používá pro modelování informací zapsaných v různých syntaxích, což umožňuje interpretovat informace webových zdrojů identifikovaných využitím URI.

URI (*Uniform Resource Identifier*) slouží v prostředí internetu k přesné specifikaci zdrojů informací (dokumentů nebo služeb) pro účel jejich strojového zpracování.

sémantických informací z ontologie (De Melo a Siersdorfer 2007). Tato metoda ale pro mapování termínů do konceptů používá externí znalosti. Pro zkoumání souvisejících konceptů je použito procházení grafu.

6.1.2 Relevantní práce z oblasti použití PageRanku pro zpracování přirozeného jazyka

Protože 2. kapitola byla věnována algoritmům citační analýzy a zvláště pak algoritmům podobným PageRanku, tak za doplnění stojí, že v úlohách zpracování textů vznikly algoritmy *TextRank*, *LexRank* a *MFSRank*, které slouží pro extrakci klíčových slov nebo důležitých vět z textů dokumentů. Tyto algoritmy nepoužívají strojové učení (tj. nevyžadují obsáhlou množinu trénovacích dokumentů), protože jsou založeny pouze na PageRanku, který určuje významné vrcholy v grafu slov zastupujícím daný dokument. TextRank (Mihalcea a Tarau 2004) slouží pro extrakci volitelně dlouhých slovních spojení nebo významných vět, které mohou být použity pro extraktivní sumarizaci[66]. PageRank je aplikován na text dokumentu reprezentovaný grafem, ve kterém vrcholy zastupují slova (či slovní spojení), hrany vyjadřují, že se slova v textu nacházejí v těsné blízkosti (volitelné okolí, např. 3 slov) a jejich váhy udávají, kolikrát se slova v této blízkosti v textu vyskytují. V případě vět váhy hran udávají počet stejných slov, která věty obsahují. LexRank (Erkan a Radev 2004), který slouží pro nalezení významných vět, které lze použít pro extraktivní sumarizaci, je TextRanku velmi podobný, s tím rozdílem, že určuje váhy hran na základě jiné míry podobnosti vět. LexRank byl navíc použit v rámci většího sumarizačního systému, který LexRankem vypočítané hodnoty kombinoval s dalšími vlastnostmi vět (např. s pozicí věty v dokumentu), a byl také použit při multi-dokumentové sumarizaci. MFSRank (López et al. 2011) pro extrakci významných frází z textů používá graf, jehož vrcholy jsou nejvíce relevantní části textu (sekvence slov) a váhy hran jsou určeny na základě součinu počtu společných výskytů ve větách a hodnoty sémantické podobnosti, která je získána z ConceptNet[67]. Zakomponováním ConceptNet byla extrakce významných frází obohacena o sémantické informace, přičemž stále nevyžaduje strojové učení.

6.2 Koncept Linked Data

Linked Data navrhl Berners-Lee (2006) pro účely vytvoření sémantického Webu, který poskytuje svůj obsah ve strojově čitelné podobě. Autor stanovil čtyři pravidla:

- používejte URI pro pojmenování zdrojů[68],
- používejte HTTP URI, aby zdroje byly vyhledatelné,
- poskytněte užitečné informace použitím standardů (tj. RDF nebo SPARQL) komukoliv, kdo vyhledá dané URI,
- zahrňte odkazy na jiná URI, která mohou poskytnout detailnější poznatky.

---

[66] Extraktivní sumarizace je (automatizovanou) metodou zpracování textů, která z předloženého textu vyjme nejdůležitější informace (často celé věty) a poskládá je do souhrnu tak, že informace zůstanou v původní podobě.

[67] *ConceptNet* je volně dostupná sémantická síť slov a krátkých frází, které jsou propojeny na základě vzájemných vztahů, např. „vytvořeno z“, „použito pro“, „motivováno cílem“ apod.
Web ConceptNet - http://conceptnet5.media.mit.edu

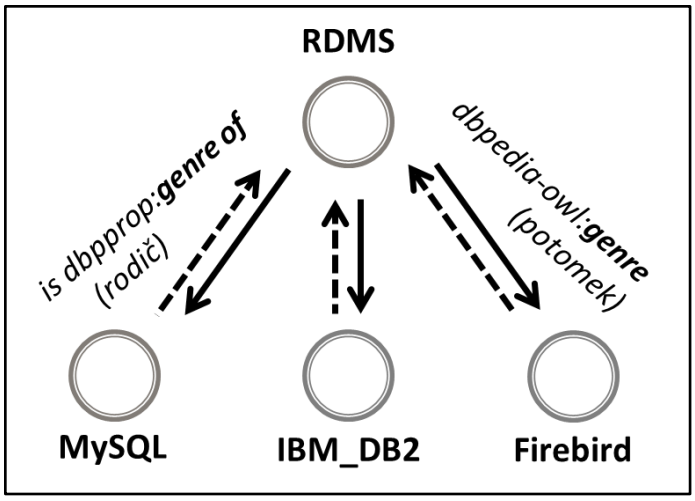
[68] Anglická formulace používá pojem *things*, což může být chápáno jako: věci, potřeby, náležitosti, nástroje atd. My použili pojem *zdroje*, kterým, dle našeho názoru, lze také popsat sady informací obsažených na Webu.

Podnětem k zavedení Linked Data byl rostoucí počet RDF dokumentů a dalších strojově čitelných zdrojů informací (např. webových stránek), z nichž je mnoho volně dostupných on-line. Více specifická je idea Linked Open Data, která je založena na předpokladu volně dostupných dat bez omezení přístupu a bez poplatků. S Linked Data jsou spojeny dva základní problémy, které se týkají duplicitních zdrojů, ze kterých mohou být Linked Data pořízena: nejednoznačnost (*disambiguation*) a rozlišení stejných referencí (*co-reference*). Tyto problémy jsou diskutovány v (Jaffri et al. 2008). V akademickém výzkumu bývají častým zdrojem Linked Data kolekce DBpedia a DBLP (popsána v části 2.2.2 a použita ve 3. kapitole).

Linked Data obecně obsahují informace o zdrojích a odkazy na další relevantní zdroje, přičemž zdroje obsahují konkrétní informace. Dvěma základními typy odkazů, které mohou být metodou pro volbu vlastností textových dokumentů přímo použity, jsou vztahy:

- hyperonymum-hyponymum (tj. vztah nadřazenosti, např. rodič-potomek)
- synonymum (tj. vztah podobnosti, např. sourozenec)

Tyto vztahy jsou obousměrné, tj. potomek může nalézt své rodiče a rodič může nalézt své potomky. Vztahy jsou v Linked Data popsány ontologickými predikáty, kterými jsou např.: *DBpedia-owl:genre*, *skos:broader*, *dcterms:subject*. Přesto, že se význam těchto predikátů mírně liší, tak je ale můžeme použít stejným způsobem, tj. ve významu rodič-potomek. Příklad hierarchických vztahů mezi zdroji v Linked Data ukazuje obrázek 6.1. Podobné zdroje či synonyma jsou spojena ontologickým vztahem *owl:sameAs* a související koncepty vztahem *skos:related*.



Obrázek 6.1: Příklad hierarchických vztahů mezi zdroji v Linked Data
(odkaz na potomka je zobrazen plnou čárou a odkaz na rodiče čárou přerušovanou;
RDMS je *Relation database management system*).

6.3 Zvolené kolekce dokumentů

Kvalita námi navržené metody pro volbu vlastností byla stanovena na základě jejího použití v klasifikátoru dokumentů. Pro testování klasifikace jsme použili dvě odlišné datové kolekce: vlastní kolekci *Call-for-Papers* a kolekci *20 Newsgroups*. Kolekce *Call-for-Papers*, kterou jsme vytvořili z 15 000 konferenčních oznámení, byla zvolena jako jednoduchý prostředek vyhodnocení, protože konference mohou být dle témat relativně jednoduše rozděleny do kategorií či klasifikačních tříd.

Z kolekce byla odstraněna oznámení o konferencích, které patří do více rozdílných kategorií, a zbylým oznámením byly přiřazeny odpovídající kategorie na základě seznamů konferencí, které jsme nalezli na internetu. Téměř každé použité oznámení obsahovalo množství klíčových slov souvisejících s danou kategorií, což nám umožnilo relativně rychle natrénovat klasifikační třídy.

Volně dostupná kolekce *20 Newsgroups*[69] (Lang 1995) obsahuje téměř 19 000 článků z diskusního systému *Uneset*, které jsou rozděleny do 20 kategorií zobrazených v tabulce 6.1.

Tabulka 6.1: Kategorie diskusních článků obsažených v 20 Newsgroups.

Oblast	**Název kategorie diskusí**[70]	**Označení**
Počítače	Počítačová grafika	comp.graphics
	Operační systém MS Windows	comp.os.ms-windows.misc
	Hardware od IBM	comp.sys.ibm.pc.hardware
	Hardware od Macintosh	comp.sys.mac.hardware
	X Windows či X11 systém	comp.windows.x
Záliby	Automobily	rec.autos
	Motocykly	rec.motorcycles
	Baseball	rec.sport.baseball
	Hokej	rec.sport.hockey
Věda	Kryptografie	sci.crypt
	Elektronika	sci.electronics
	Medicína	sci.med
	Vesmír	sci.space
Politika	Politické diskuse všeho druhu	talk.politics.misc
	Politika vlastnictví a užití zbraní	talk.politics.guns
	Události na blízkém východě	talk.politics.mideast
Náboženství	Náboženské diskuse všeho druhu	talk.religion.misc
	Ateistické diskuse	alt.atheism
	Křesťanské diskuse	soc.religion.christian
Obchodování	Inzeráty	misc.forsale

6.4 Naše metoda pro volbu klíčových slov textového dokumentu

Návrh metody pro automatickou volbu vlastností či klíčových slov textových dokumentů byl nejdůležitější částí naší práce. Navržená metoda používající Linked Data a PageRank obsahuje následující kroky:

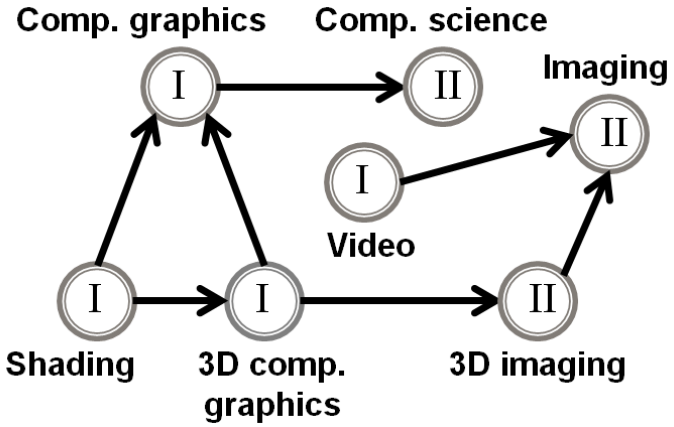
1) Použitím TF-IDF, viz dále, z dokumentu vybereme nejdůležitější termíny (použita může být i jiná metoda, např. χ^2), které prohlásíme za základní.
2) Základní termíny namapujeme na Linked Data tak, aby každý termín byl zastoupen jedním vrcholem, který je identifikován URI. Protože vrcholy v Linked Data jsou vzhledem k jejich důležitosti a oblíbenosti obvykle pojmenovány s využitím několika jazyků, tak je mapování založeno na úplné nebo alespoň částečné shodě mezi daným termínem a jménem vrcholu

[69] Web 20 Newsgroups – http://qwone.com/~jason/20Newsgroups/

[70] Pro získání názvů kategorií diskusí v *20 Newsgroups* byly použity názvy z Google Groups, viz https://groups.google.com

v Linked Data v odpovídajícím jazyce. Základní termíny, které se nepodařilo namapovat, označíme za nevýznamné a odstraníme. Vrcholy propojíme využitím vztahů z Linked Data, čímž dokument reprezentujeme grafem, ve kterém má každý důležitý termín jednoznačnou pozici.

3) Využitím vazeb z Linked Data provedeme jeden krok rozšíření grafu dokumentu a doplníme graf o další relevantní termíny z Linked Data. Cílem tohoto postupu je získat nové k textu dokumentu vysoce relevantní termíny a to zvláště takové, které nebyly explicitně rozpoznány v textu dokumentu. Jeden krok rozšíření grafu ilustruje obrázek 6.2, na kterém jsou základní důležité termíny označeny (I) a nově přidané termíny jsou označeny (II).
4) Vytvořený graf vyhodnotíme PageRankem, čímž určíme významnost vrcholů, která udává míru relevance jednotlivých termínů k obsahu dokumentu. Pro vyhodnocení grafu použijeme PageRank bez personalizace, který používá váhy hran a je v části 2.4.1 zapsán vzorcem (2.11).
5) Pokud libovolný vrchol z vrcholů, které byly do grafu přidány při jeho posledním rozšiřování, získal nejvyšší hodnotu PageRanku, tak pokračujeme krokem 3), tj. dalším rozšířením grafu, jinak ukončíme metodu pro volbu vlastností a termín zastoupený vrcholem s nejvyšší hodnotou PageRanku prohlásíme za nejreprezentativnější vlastnost dokumentu. Samozřejmě na základě nejvyšších hodnot PageRanku může být za vlastnosti pro reprezentaci dokumentu zvoleno i více termínů.

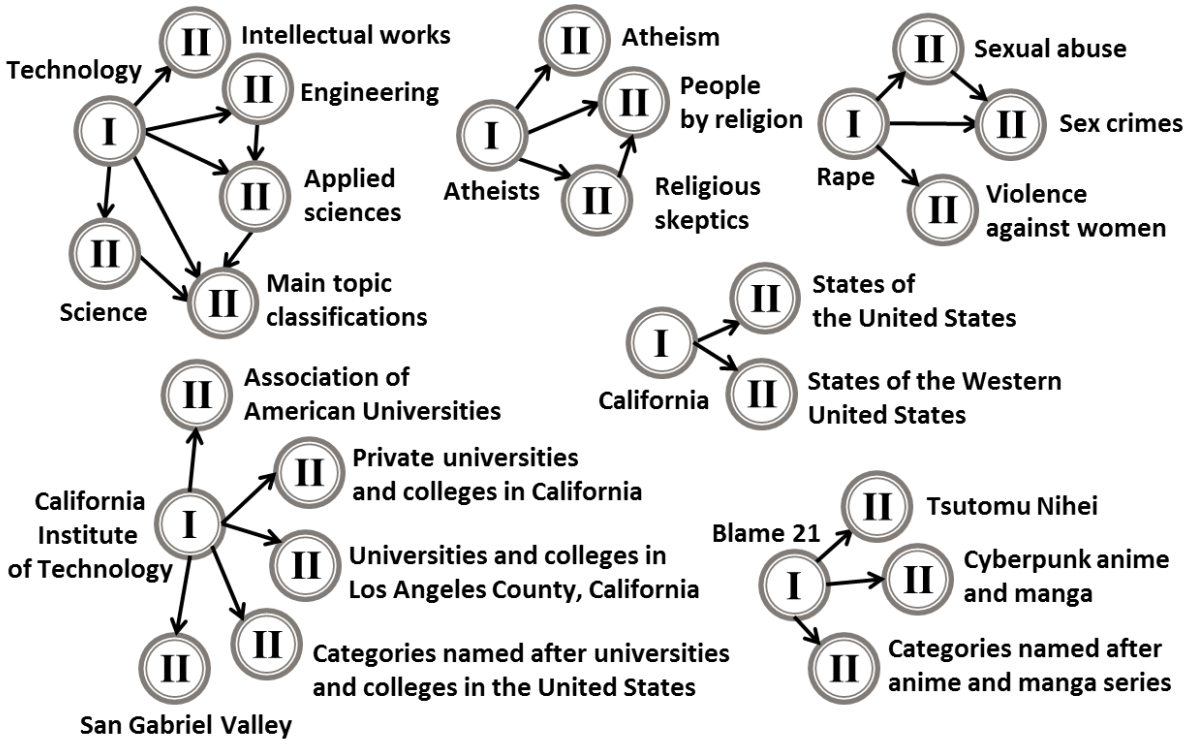


Obrázek 6.2: První krok rozšíření grafu termínů s použitím Linked Data – konference I3D (vrcholy značené (I) jsou namapované základní nejdůležitější termíny dokumentu a vrcholy značené (II) jsou termíny odvozené na základě vazeb z Linked Data).

Jedním z problémů, který jsme při návrhu naší metody pro volbu vlastností textového dokumentu řešili, bylo, jak zachovat dynamičnost metody a současně zajistit, aby se rozšiřování grafu „včas" zastavilo a nejreprezentativnější vybraná vlastnost nebyla příliš obecná. Zde jsme nejprve zkusili ukončit rozšiřování grafu ve chvíli, kdy celý graf vytvořil jednu souvislou komponentu. To se ukázalo dobře použitelné pro kolekci *Call-for-Papers*, protože rozšiřováním původních klíčových slov obsažených v dokumentu s oznámením konference rychle vznikl souvislý graf (v téměř nejhorším případě např. různé pojmy z počítačových věd brzo odkázaly na vrchol *Počítačové vědy* a tím vznikla jedna souvislá komponenta). Reálný příklad pro konferenci I3D je zobrazen na obrázku 6.2. Ideálním termínem či vlastností zastupující konferenci I3D je *Počítačová grafika*.

Pro kolekci *20 Newsgroups* nebyl tento způsob ukončení rozšiřování grafu příliš použitelný, protože např. kriminální článek obsahuje informace o místě zločinu a typu zločinu, které jsou z různých oblastí

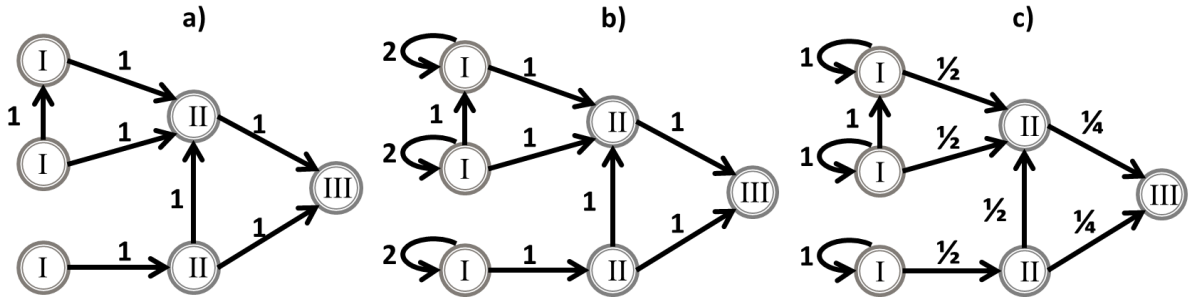
použitých Linked Data. Příkladem může být článek, ve kterém pojem *univerzita+jméno* vedl k doplnění pojmů týkajících se vzdělávacích institucí v daném městě a současně pojem *znásilnění* vedl k doplnění pojmů týkajících se sexuálních zločinů, viz reálný příklad na obrázku 6.3 – pokud bychom graf tohoto článku rozšiřovali, dokud nevytvoří jednu souvislou komponentu, tak bychom do něj zahrnuli spoustu nadbytečných termínů a následně určená nejrelevantnější vlastnost by byla buďto velmi obecná, nebo dokonce zcestná.



Obrázek 6.3: Příklad prvního kroku rozšiřování základních vlastností dokumentu využitím Linked Data. Dokumentem je reálný kriminální článek pojednávající o znásilnění na Kalifornské univerzitě (zobrazeno je rozšíření pouze vybraných vlastností).

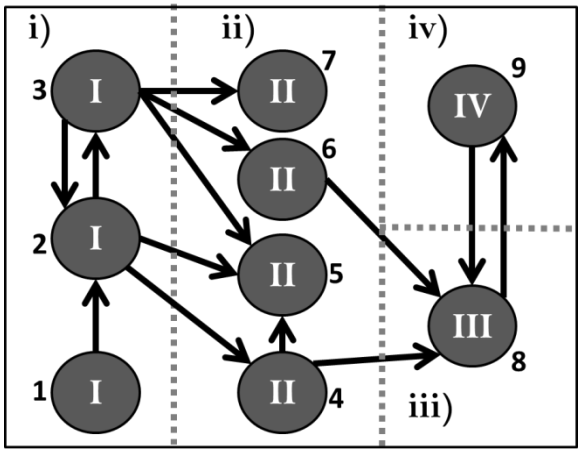
Abychom zamezili nežádoucímu rozšiřování grafu, tak jsme použili způsob, který v předchozím postupu shrnují kroky 3) až 5), a to rozšiřování grafu pouze v případě, že libovolný termín z naposledy doplněných termínů získal nejvyšší hodnotu PageRanku, viz krok 5). Protože tento způsob samostatně nevedl k efektivnímu ukončení rozšiřování grafu, tak jsme dále zkoušeli vylepšit konstrukci grafu. Zde jsme vyzkoušeli 3 způsoby konstrukce, viz následující seznam a obrázek 6.4:

a) původní varianta - všechny hrany mají váhu 1,
b) vrcholům zastupujícím základní termíny byla přidána smyčka s váhou 2,
c) vrcholům zastupujícím základní termíny byla přidána smyčka a všem hranám byly stanoveny váhy rovné $1/2^{Q-1}$, kde Q je číslo iterace, ve které byl do grafu přidán odkazovaný vrchol – vrcholy označené (I) byly přidány v první iteraci, vrcholy označené (II) v iteraci druhé atd. Pozn.: testovali jsme i jiné způsoby snižování vah hran, ale tento způsob byl nejefektivnější.



Obrázek 6.4: Varianty konstrukce grafu, který zastupuje textový dokument.

Dle našeho předpokladu, testování variant konstruování grafu ukázalo, že varianta c) poskytuje vlivem efektivního ukončení rozšiřování grafu nejlepší výsledky. Proto jsme ji při klasifikaci dokumentů použili v naší metodě pro volbu vlastností. Efektivitu ukončení metody pro volbu vlastností ilustruje tabulka 6.2, která zobrazuje hodnoty PageRanku jednotlivých vrcholů z grafu na obrázku 6.5 v jednotlivých iteracích rozšiřování grafu – nejvyšší hodnota PageRanku v každé iteraci je tučná a pokud nejvyšší hodnotu získal vrchol přidaný do grafu aktuální iterací rozšiřování grafu, tak byla provedena další iterace rozšíření grafu. Na obrázku 6.5 jsou znázorněny jednotlivé iterace rozšiřování grafu, přičemž římské číslice značí, ve které iteraci byly vrcholy z Linked Data do grafu doplněny. Nejprve je tedy vyhodnocen graf tvořený vrcholy 1-7, které byly přidány v 1. a 2. iteraci vytváření grafu. Ve variantách konstrukce grafu a) a b), viz tabulka 6.2, získá nejvyšší hodnotu PageRanku vrchol 5, který byl přidán v aktuálně poslední 2. iteraci. Proto je provedena další iterace rozšíření grafu a do grafu je doplněn vrchol 8. Pokud je použita varianta konstrukce grafu b) a vzniklý graf je vyhodnocen PageRankem, tak nejvyšší hodnotu PageRanku nyní získá vrchol 2. Protože tento vrchol nebyl do grafu přidán v poslední iteraci rozšíření grafu, tak je metoda pro volbu vlastností ukončena a vlastnost, kterou zastupuje vrchol 2, je prohlášena za nejreprezentativnější vlastnost daného dokumentu. V tabulce 6.2 lze vidět, že pokud je použita varianta konstrukce grafu c), tak vrchol 2 je zvolen už po prvním rozšíření grafu, které je zastoupeno vrcholy přidanými v 1. a 2. iteraci vytváření grafu.



Obrázek 6.5: Příklad grafu pro testování variant rozšiřování grafu.

Tabulka 6.2: Závislost hodnot PageRanku na variantách konstrukce grafu při vyhodnocení grafu z obrázku 6.5 (nejvyšší hodnoty PageRanku v každé iteraci rozšíření grafu jsou tučné).

krok	vrchol	varianta a)			varianta b)		varianta c)
		I + II	I + II+ III	I+II+III+IV	I + II	I + II + III	I + II
I	1	0,08	0,07	0,027	0,16	0,14	0,12
I	2	0,18	0,15	0,059	0,20	**0,18**	**0,22**
I	3	0,13	0,11	0,043	0,14	0,13	0,17
II	4	0,13	0,11	0,043	0,10	0,09	0,10
II	5	**0,27**	0,18	0,071	**0,21**	0,15	0,21
II	6	0,11	0,09	0,036	0,09	0,08	0,09
II	7	0,11	0,09	0,036	0,09	0,08	0,09
III	8		**0,19**	**0,355**		0,16	
IV	9			0,329			

Protože pro nalezení základních nejdůležitějších termínů obsažených v dokumentu jsme použili TF-IDF, tak ho nyní v krátkosti popíšeme. Frekvence termínu (*Term Frequency* – TF) je frekvence výskytu termínu v dokumentu, která udává důležitost daného termínu pro daný dokument, viz vzorec (6.1), kde *TF(t,d)* je frekvence výskytu termínu *t* v dokumentu *d*, n_t^d je počet výskytů termínu *t* v dokumentu *d* a n^d je počet všech podstatných termínů v dokumentu *d* (tj. všech termínů, pro které se TF počítá). Inverzní frekvence dokumentu (*Inverse Document Frequency* – IDF) je dána výskytem daného termínu v dokumentech z celé datové kolekce. IDF udává důležitost daného termínu pro celou kolekci tak, že hodnoty termínů snižuje v závislosti na frekvenci jejich výskytu v dokumentech, viz vzorec (6.2), kde *IDF(t,D)* je inverzní frekvence výskytu termínu *t* v kolekci dokumentů *D*, N^D je počet dokumentů v kolekci *D* a N_t^D je počet dokumentů z kolekce *D*, ve kterých se vyskytuje termín *t*. TF-IDF následně udává důležitost daného termínu pro daný dokument v závislosti na zbylých dokumentech kolekce, viz vzorec (6.3), kde *TFIDF(t,d,D)* je důležitost termínu *t* pro dokument *d* v kolekci *D*. Podrobnější informace a uplatnění TF-IDF v textových úlohách zmiňují např. Manning et al. (2008).

$$TF(t,d) = \frac{n_t^d}{n^d} \tag{6.1}$$

$$IDF(t,D) = log\frac{N^D}{N_t^D} \tag{6.2}$$

$$TFIDF(t,d,D) = TF(t,d) \cdot IDF(t,D), \quad d \in D \tag{6.3}$$

6.5 Diskuse kvality naší metody pro volbu klíčových slov dokumentu

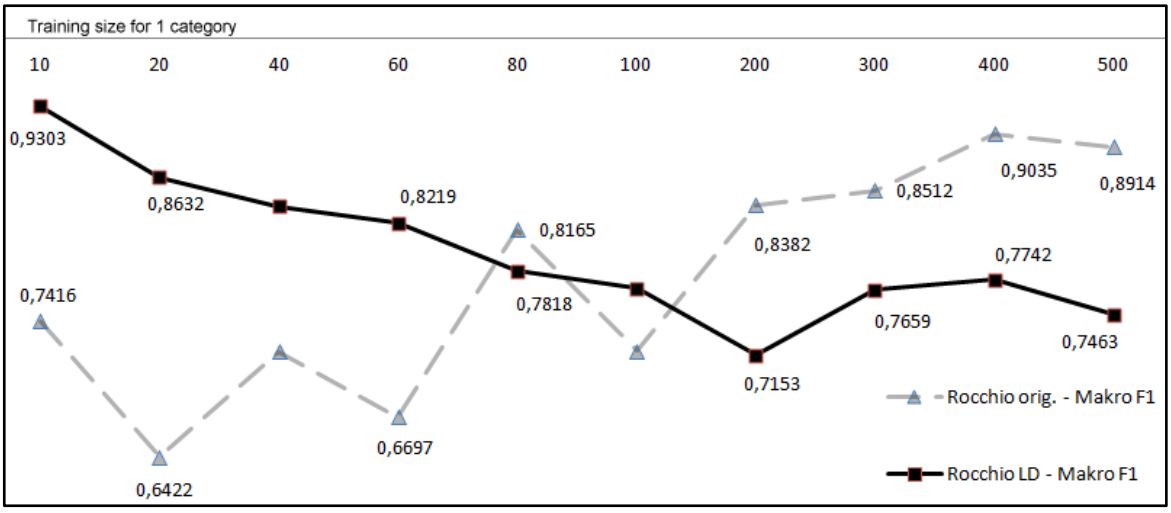
Abychom určili kvalitu námi navržené metody pro volbu vlastností, tak jsme ji otestovali v klasifikátoru, který jsme použili na dvě rozdílné kolekce dokumentů – kolekci diskusních článků *20 Newsgroups* a vlastní kolekci konferenčních *Call-for-Papers*. Pro porovnání naší metody pro volbu vlastností se standardním statistickým přístupem jsme použili stejný klasifikační algoritmus (v našem případě algoritmus Rocchio, který používá vektorový prostor). Na obrázku 6.6 jsou porovnány

úspěšnosti Rocchio klasifikátoru při natrénování s využitím vlastností získaných naší metodou pro volbu vlastností (souvislá čára) a téhož klasifikátoru při natrénování s využitím základních vlastností získaných statistickou metodou TF-IDF (přerušovaná čára). Porovnání je provedeno na základě makro-průměrování[71] (*macro-averaging*) F_1 míry, viz vzorec (6.4), kde F_1 je F_1 makro-průměr provedené klasifikace, K je množina klasifikačních tříd a $|K|$ její velikost, P_α je přesnost[72] klasifikace pro třídu α a R_α je úplnost[73] klasifikace pro třídu α. Počet testovacích dokumentů byl stanoven jako 20% z dokumentů obsažených v každé kategorii použité kolekce.

$$F_1 = \frac{1}{|K|}\sum_{\alpha \in K}\left(\frac{2 \cdot P_\alpha \cdot R_\alpha}{P_\alpha + R_\alpha}\right) \tag{6.4}$$

V kolekci *Call-for-Papers* bylo pro natrénování každé klasifikační třídy postačujících 10 „kvalitních" dokumentů, přičemž bylo dosaženo téměř konstantního makro-průměrování ($F_1 \doteq 0{,}9$). Při trénování klasifikátoru s použitím 10 až 100 trénovacích dokumentů pro každou klasifikační třídu byla F_1 míra stále téměř konstantní, ale pokud bylo použito více jak 100 trénovacích dokumentů nebo 2000 vlastností, tak jsme zaznamenali problém s přetrénováním.

Kolekce diskusních článků *20 Newsgroups* byla použita pro simulování častého případu užití klasifikace, ale my použili jen část z jejích dat, protože náš zdroj Linked Data nebyl schopen rozlišit mezi podobnými kategoriemi, jako jsou např. comp.os.ms-windows.misc (články o operačním systému MS Windows) a comp.windows.x (články o systému X Window pro vytváření GUI). Při použití přibližně 100 trénovacích dokumentů pro natrénování každé klasifikační třídy se při vyhodnocení kolekce *20 Newsgroups* opět projevilo přetrénování klasifikátoru, viz obrázek 6.6.



Obrázek 6.6: Makro-průměr F_1 míry pro Rocchio klasifikátor při použití 20 Newsgroups (přerušovaná čára značí použití základních vlastností a plná čára použití námi zvolených vlastností dokumentů).

[71] Při makro-průměrování F_1 míry je vypočtena F_1 míra pro každou klasifikační třídu zvlášť a F_1 míra celé klasifikace je určena jako průměr z vypočtených hodnot.

[72] *Přesnost klasifikace* je podíl počtu dokumentů, které byly správně zařazeny do dané klasifikační třídy, a počtu všech dokumentů, které byly do dané třídy zařazeny.

[73] *Úplnost klasifikace* je podíl počtu dokumentů, které byly správně zařazeny do dané klasifikační třídy, a počtu všech dokumentů, které měly být do dané třídy správně zařazeny.

6.6 Vyhodnocení experimentu s volbou klíčových slov dokumentu

Naše metoda pro volbu klíčových slov či vlastností dokumentů s využitím PageRanku a Linked Data je slibná zejména pro úlohy klasifikace, které pro natrénování klasifikátoru používají malou množinu trénovacích dokumentů, nebo pro rychlé filtrování dokumentů. V těchto případech může být fáze učení pro uživatele drahá a náročná na čas. Naše metoda volby vlastností či klíčových slov umožňuje definovat funkce pro přiřazení dokumentů do klasifikačních tříd s využitím pouze malého počtu trénovacích dokumentů nebo s využitím např. jen jednoho vrcholu z Linked Data a jeho automatického rozšíření. Navržená metoda byla v experimentu použita jak pro natrénování modelu ve fázi učení, tak i pro zařazení dokumentů do konkrétních tříd ve fázi klasifikace. Nicméně, pro běžné úlohy klasifikace dokumentů je nutné, aby v Linked Data existovala klíčová slova z odpovídajících oblastí či kategorií. V budoucím výzkumu bychom jednak chtěli eliminovat problém s přetrénováním klasifikátoru, a dále pak vytvořit metodu, která bude dokumenty klasifikovat přímo na základě grafové analýzy.

7 Shrnutí dosažených výsledků

V první části práce je uveden vyčerpávající popis současného stavu poznání v oblasti citační analýzy. Jsou zde popsány nejznámější neiterační metody citační analýzy, důkladně vysvětlen iterační algoritmus PageRank a shrnuty PageRanku podobné algoritmy, které byly navrženy pro potřeby citační analýzy. Následně jsou v textu práce shrnuty vlastní přínosy autora v oblasti bibliometrie a v oblasti zpracování textů, přičemž společným prvkem navržených metod je použitý algoritmus PageRank.

V části práce s vlastními přínosy autora v oblasti bibliometrie jsou představeny nově navržené metody pro strojové hodnocení autorů vědeckých publikací umožňující hodnotit autory v souladu s hodnoceními, která poskytují organizace ACM a ISI. Kvalita navržených metod je experimentálně ověřena jejich aplikováním na citační sítě vytvořené z bibliografických kolekcí CiteSeer, DBLP a WoS. Poté, co byly za nejvěrohodnější výsledky prohlášeny výsledky získané z kolekce WoS, jsou navrženy a porovnány další varianty metod pro hodnocení autorů. Celkově nejlepší pořadí autorů vytvořila metoda, která hodnotí autory na základě vyhodnocení citační sítě publikací PageRankem s personalizací dle hodnot časopisů, ve kterých byly publikace zveřejněny. Hodnoty publikací metoda rovnoměrně rozděluje jejich autorům. Tato metoda má lepší výsledky, než námi dříve použité metody, a je značně lepší než metody neiterační. Navržené metody a výsledky provedených experimentů byly publikovány ve 2 časopiseckých a 1 konferenčním článku.

V části práce, která je věnována zpracování textů, je popsán návrh naší metody pro určování klíčových slov textového dokumentu. Navržená metoda umožňuje využitím Linked Data doplnit množinu slov, která byla z textu dokumentu získána statistickým postupem, o další relevantní slova a následně využitím PageRanku z celé této množiny slov vybrat jedno či více klíčových slov, která dokument nejlépe reprezentují. Vybraná klíčová slova mohou být použita jako vlastnosti dokumentu při klasifikaci, shlukování či štítkování dokumentů. Kvalita navržené metody je experimentálně ověřena jejím použitím v klasifikátoru textových dokumentů. Pro klasifikaci je použita datová kolekce diskusních článků *20 Newsgroups* a vlastní kolekce konferenčních *Call-for-Papers*. Výsledky ukazují, že námi navržená metoda je vhodná pro situace, ve kterých máme málo dat pro klasický způsob natrénování klasifikátoru, protože umožňuje kvalitní natrénování klasifikátoru s použitím pouze např. deseti dokumentů pro jednu klasifikační třídu. Výsledky tohoto experimentu byly publikovány ve 2 konferenčních článcích.

V následující části 7.1 je popsáno splnění cílů práce, v části 7.2 jsou shrnuty hlavní přínosy autora a v části 7.3 zrekapitulovány možné budoucí práce v oblasti hodnocení autorů a v oblasti zpracování textových dokumentů.

7.1 Splnění cílů práce

Splnění cílů této práce v oblasti bibliometrie (cíle jsou kurzívou):

(a1) Navržení metody pro automatické hodnocení autorů, která bude hodnotit autory z počítačových věd s výsledky obdobnými hodnocením organizací ACM a ISI, a analýza vhodnosti použití datových kolekcí CiteSeer (2005), DBLP (2004) a WoS (1996-2005) pro hodnocení autorů.

- Metody pro automatické hodnocení autorů byly navrženy ve 3. a 5. kapitole. Kolekce CiteSeer a DBLP byly pro experimentální ověření kvality metod použity ve 3. kapitole a kolekce WoS byla použita ve 4. a 5. kapitole.
- Kvalita jednotlivých metod byla určena na základě jejich schopnosti vyzdvihnout ve vytvořeném pořadí autory oceněné od ACM (Turingova cena, Coddova cena, seznam významných osob ACM a ceny udílené ACM v kategoriích *Umělá inteligence* a *Hardware*) nebo od ISI (seznam vysoce citovaných výzkumníků ISI), viz 3. až 5. kapitola.
- Ve 3. kapitole je diskutována menší kvalita kolekcí CiteSeer (obsahuje indexovací chyby) a DBLP (specializuje se na oblast databázových systémů a logického programování a mimo tyto oblasti obsahuje velmi málo indexovaných citací). Ve 4. a 5. kapitole je prověřena vhodnost použití zakoupené kolekce WoS pro hodnocení autorů z celé oblasti počítačových věd nebo z kategorií *Umělá inteligence* a *Hardware*. Výsledky, které jsme z kolekce WoS získali, považujeme, vzhledem k vlastnostem této kolekce, za nejvěrohodnější. Porovnání použitých kolekcí z pohledu kvality obsažených dat zmiňuje část 4.2.

(a2) Porovnání navržených metod s neiteračními metodami.

- Ve 4. kapitole je k iteračním metodám přidána neiterační metoda počítající citace autorů a publikací (tj. měřící popularitu) a v 5. kapitole je přidána metoda počítající publikace autorů (tj. měřící produktivitu) a metoda počítající h-index autorů (tj. měřící vyspělost). Všechny tyto méně důmyslné metody poskytují horší výsledky než metody založené na PageRanku.

(a3) Zjištění, jaký vliv na kvalitu hodnocení autorů mají použité citační sítě publikací či autorů, samocitace autorů a váhy hran v citační síti autorů.

- Ve 4. a 5. kapitole je prokázáno, že metody pracující s citační sítí publikací dosahují v hodnocení autorů lepších výsledků, než metody pracující s citační sítí autorů, která pomíjí některé informace (např. časový sled publikování).
- Samocitace autorů je nejlepší odstranit na úrovni publikací, tj. odstranit citace mezi publikacemi se stejným autorem, viz 3. až 5. kapitola.
- Ve 4., 5. a částečně i ve 3. kapitole je diskutována skutečnost, že námi použité váhy hran v citační síti autorů mají na hodnocení autorů minimální vliv. Nicméně v 5. kapitole je ukázáno, že použití hodnot časopisů, ve kterých byly zveřejněny publikace, jako vah vstupních hran v citační síti publikací umožňuje, v porovnání s baseline, mírně lepší hodnocení autorů.

(a4) Zjištění, jaký vliv na kvalitu hodnocení autorů mají způsoby rozdělení hodnot publikací jejich autorům, a posouzení vhodnosti zvýhodňování prvních či korespondujících autorů publikací.

- V 5. kapitole je použito pět způsobů rozdělení hodnot publikací autorům a prokázáno, že není vhodné hodnoty publikací rozdělovat autorům nerovnoměrně a tím zvýhodňovat první či korespondující autory publikací.

(a5) Ověření vlivu parametrizace PageRanku charakteristikami autora či publikace na kvalitu hodnocení autorů.

- Ve 3. kapitole jsou do personalizace PageRanku zakomponovány produktivita autorů (počet publikací) a kvalita publikací, která je zastoupena počtem autorů publikace (měřícím vynaložené úsilí). V 5. kapitole je v případě autorů do personalizace PageRanku zakomponována jejich vyspělost (h-index) a v případě publikací jejich popularita (počet citací) a kvalita časopisu, ve kterém byla publikace vytištěna.
- Metoda využívající produktivitu autorů v jejich personalizaci poskytuje při vyhodnocení citační sítě autorů nejlepší výsledky. Vyspělost autorů v personalizaci PageRanku poskytuje výsledky horší než předchozí metoda, ale lepší než baseline, tj. PageRank bez personalizace. Nejlepší výsledky v hodnocení autorů ze všech námi testovaných metod má metoda využívající v personalizaci publikací kvalitu časopisů. Metoda, která využívá kvalitu publikací danou počtem autorů, má výsledky horší než předchozí metoda, ale lepší než baseline. Horší výsledky než baseline poskytuje pouze metoda využívající v personalizaci publikací jejich popularitu (počet citací). Tyto závěry jsou podrobně diskutovány v části 5.4.

(a6) Ověření použitelnosti navržených metod v případě změny rozsahu vyhodnocovaného oboru.

- V 5. kapitole jsou navržené metody mimo jiné testovány při hodnocení autorů v kategoriích *Umělá inteligence* a *Hardware.* V obou případech mají metody téměř totožné pořadí kvality, jako při hodnocení autorů z celé oblasti počítačových věd. Nejlepší metoda používá hodnoty časopisů v personalizaci publikací. Stejné závěry byly vyvozeny i při predikci laureátů vědeckých ocenění.

Splnění cílů této práce v oblasti zpracování textů (cíle jsou kurzívou):

(b1) Navržení metody, která využitím Linked Data a PageRanku dokáže automaticky určit klíčová slova pro daný textový dokument. Tato slova se nemusejí explicitně vyskytovat v textu dokumentu, ale měla by daný dokument reprezentovat lépe, než slova určená pouze statisticky.

- V 6. kapitole je navržena metoda, která využitím Linked Data získaných z DBpedie a PageRanku dokáže určit klíčová slova textového dokumentu. Tato klíčová slova se nemusejí vyskytovat v textu dokumentu, ale přesto, jak ukazuje provedený experiment, reprezentují dokument lépe, než slova získaná z dokumentu statisticky.

(b2) Ověření kvality navržené metody při klasifikaci textových dokumentů.

- V 6. kapitole je navržená metoda použita pro volbu vlastností při klasifikaci dokumentů z kolekce diskusních článků *20 Newsgroups* a z vlastní kolekce konferenčních *Call-for-Papers*. Z výsledků je zřejmé, že navržená metoda vylepšuje klasifikaci zvláště v případech, kdy máme pro natrénování klasifikátoru menší množinu trénovacích dokumentů.

7.2 Hlavní vědecké přínosy této práce

Hlavní vědecké přínosy v oblasti bibliometrie jsou:

- Zjištění, že metody, které pracují s citační sítí publikací, poskytují dokonalejší hodnocení autorů, než metody pracující s citační sítí autorů.
- Navržení parametrizace PageRanku personalizací autorů (počtem publikací, h-indexem) nebo publikací (počtem autorů, počtem citací, hodnotou časopisu) pro účely hodnocení autorů.
- Experimentální porovnání navržených metod a zjištění, že zakomponování hodnot časopisů do PageRanku zlepšuje hodnocení autorů, kdežto dodatečné zakomponování popularity (počet citací) do PageRanku hodnocení autorů zhoršuje.
- Zjištění, že není vhodné rozdělovat hodnoty publikací autorům nerovnoměrně a tím zvýhodňovat autory, kteří jsou v publikacích ve výčtech autorů uvedeni na předních pozicích.
- Potvrzení, že metody založené na PageRanku hodnotí autory lépe, než metody založené na počítání citací, a že samocitace autorů je vhodné odstranit na úrovni publikací (tj. odstranit citace mezi publikacemi, které mají stejného alespoň jednoho autora).
- Vyhodnocení kolekcí CiteSeer, DBLP a WoS z oblasti počítačových věd a vyvození závěru, že CiteSeer a DBLP nejsou pro citační analýzu příliš vhodné, protože obsahují nekompletní údaje.
- Ověření funkcionality navržených metod a platnosti vyvozených závěrů na úžeji specializovaných WoS kategoriích (ověřeno na kategoriích *Umělá inteligence* a *Hardware*) a při testování schopnosti metod predikovat laureáty vědeckých ocenění.

Hlavní vědecké přínosy této práce v oblasti zpracování textů jsou:

- Navržení metody, která využitím PageRanku a Linked Data dokáže generovat klíčová slova reprezentující daný dokument. Tato slova jsou na základě Linked Data odvozena od slov, která byla v dokumentu vyhledána statisticky. Navržená metoda, v porovnání s pouhým statistickým přístupem, umožňuje lepší klasifikaci dokumentů a to zvláště v případech, kdy máme pro natrénování klasifikátoru menší množství dat. Pro natrénování jedné klasifikační třídy postačuje např. jen deset dokumentů.

7.3 Budoucí práce

Budoucí práce, které by mohly být řešeny v oblasti hodnocení autorů PageRankem:

- Protože v části 5.4 jsme zmínili skutečnost, že v category-independent způsobu hodnocení autorů v kategoriích jsou zvýhodňováni autoři, kteří jsou významní v celé kolekci dat (oblast počítačových věd), ale nemusí být významní v dané kategorii, tak v další práci lze experimentovat se získáním hodnot autorů v category-independent způsobu hodnocení autorů. Zajímavé by bylo vyzkoušet tyto varianty:
    - Vypočítat hodnoty publikací na základě citační sítě publikací, která je vytvořena ze všech publikací v kolekci, ale hodnoty autorů určit pouze na základě hodnot publikací ze zvolené kategorie.
    - Hodnoty autorů určené na základě hodnot všech jejich publikací vynásobit podílem počtu publikací, které autor publikoval ve zvolené kategorii, vůči všem jeho publikacím.

- Zjistit, které kritérium stanovení kvality strojově vytvořených pořadí autorů na základě pozic oceněných autorů je nejlepší. Použít lze průměr (viz experimenty v 3. až 5. kapitole), medián, minimum či maximum, viz (Fiala 2012b), nebo se lze zaměřit pouze na několik nejlepších pozic ve vytvořeném pořadí autorů, např. 20 nebo 100 nejlepších pozic, viz (Ding 2011a).
- Experimentovat s využitím klasifikačního systému nebo ontologie, které by umožnily vyhledávání autorů se zvolenou specializací. Součástí vyhledání autorů je vytvoření jejich pořadí a to ideálně naší nejlepší metodou pro hodnocení autorů, která byla popsána v 5. kapitole.
- Experimentovat s hodnocením významnosti pracovních skupin, oddělení či institucí.
- Ověřit kvalitu naší metody pro hodnocení autorů v jiných oblastech výzkumu.
- Na základě navržených metod porovnat různé referenční seznamy významných autorů a rozdělit tak tyto seznamy na seznamy populárních a seznamy prestižních autorů.

Při budoucím určování klíčových slov textových dokumentů PageRankem a jejich využití při klasifikaci dokumentů by bylo zajímavé:

- Navrhnout nový způsob konstrukce grafu slov, který zastupuje daný dokument, za účelem zlepšení výběru klíčových slov (případně lze také experimentovat s počtem zvolených klíčových slov).
- Analyzovat a eliminovat problém s přetrénováním klasifikátoru, který se projevuje, pokud je pro trénování klasifikační třídy použito přibližně 100 a více dokumentů.
- Navrhnout metodu, která by dokumenty klasifikovala pouze na základě analýzy grafu vytvořeného z Linked Data. Pro každou třídu by byl určen jeden vrchol (či více vrcholů) v Linked Data, který ji reprezentuje. V prvním kroku klasifikace (zastupujícím volbu vlastností dokumentů) by byla nalezena klíčová slova dokumentu stejným způsobem jako v 6. kapitole. V druhém kroku klasifikace (zastupujícím samotnou klasifikaci dokumentů) by byly určeny vzdálenosti od vrcholů reprezentujících klíčová slova dokumentu k vrcholům zastupujícím klasifikační třídy. Na základě nejmenší vzdálenosti by byl dokument zařazen do odpovídající klasifikační třídy.

Literatura

ABRAMO, Giovanni, D'ANGELO, Ciriaco Andrea a VIEL, Fulvio, 2010. Peer review research assessment: a sensitivity analysis of performance rankings to the share of research product evaluated. *Scientometrics*. 85(3), 705–720. ISSN 01389130. Dostupné z: doi:10.1007/s11192-010-0238-0

ABRIZAH, A., ZAINAB, A. N., KIRAN, K. a RAJ, R. G., 2013. LIS journals scientific impact and subject categorization: a comparison between Web of Science and Scopus. *Scientometrics*. 94(2), 721–740. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-012-0813-7

AGGARWAL, Charu C., 2011. *Social Network Data Analytics*. New York: Springer US. ISBN 978-1-4419-8461-6. Dostupné z: doi:10.1007/978-1-4419-8462-3

AJIFERUKE, Isola a WOLFRAM, Dietmar, 2009. Citer analysis as a measure of research impact: library and information science as a case study. *Scientometrics*. 83(3), 623–638. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-009-0127-6

ALONSO, S., CABRERIZO, F.J., HERRERA-VIEDMA, E. a HERRERA, F., 2009. H-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*. 3(4), 273–289. ISSN 17511577. Dostupné z: doi:10.1016/j.joi.2009.04.001

AMARA, Nabil a LANDRY, Réjean, 2012. Counting citations in the field of business and management: why use Google Scholar rather than the Web of Science. *Scientometrics*. 93(3), 553–581. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-012-0729-2

ASSIMAKIS, N. a ADAM, M., 2010. A new author's productivity index: p-index. *Scientometrics*. 85(2), 415–427. ISSN 01389130. Dostupné z: doi:10.1007/s11192-010-0255-z

BANKS, Michael G., 2013. An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics*. 69(1), 161–168. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-006-0146-5

BAR-ILAN, Judit, 2007. Which h-index? — A comparison of WoS, Scopus and Google Scholar. *Scientometrics*. 74(2), 257–271. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-008-0216-y

BASTIAN, Mathieu, HEYMANN, Sebastien a JACOMY, Mathieu, 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. In: *Third International AAAI Conference on Weblogs and Social Media*. San Jose (USA): AAAI Publications, s. 361–362.

BATAGELJ, Vladimir a MRVAR, Andrej, 1998. Pajek – program for large network analysis. *Connections*. 21(2), 47–57. ISSN 0226-1776. Dostupné z: http://vlado.fmf.uni-lj.si/pub/networks/doc/pajek.pdf

BAVELAS, Alex, 1948. A Mathematical Model for Group Structures. *Human Organization*. 7(3), 16–30. ISSN 0018-7259. Dostupné z: doi:10.17730/humo.7.3.f4033344851gl053

BELLIS, Nicola De, 2009. *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Lanham, Toronto, Plymouth: The Scarecrow Press. ISBN 9780810867130.

BENZI, Michele a KUHLEMANN, Verena, 2012. Chebyshev acceleration of the GeneRank algorithm. *Electronic Transactions on Numerical Analysis*. 40, 311–320.
Dostupné z: http://mathcs.emory.edu/~benzi/Web_papers/geneRank.pdf

BERGSTROM, Carl T., 2007. Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*. 68(5), 314–316. Dostupné z: http://crln.acrl.org/content/68/5/314.full.pdf+html

BERGSTROM, Carl T., WEST, J. D. a WISEMAN, M. A., 2008. The EigenfactorTM Metrics. *Journal of Neuroscience*. 28(45), 11433–11434. ISSN 0270-6474. Dostupné z: doi:10.1523/JNEUROSCI.0003-08.2008

BERCHENKO, Yakir, DALIOT, Or a BRUELLER, Nir N., 2011. Intra-firm information flow: a content-structure perspective. In: *Advances in Intelligent Data Analysis X*. Porto: Springer Berlin Heidelberg, s. 34–42. Dostupné z: http://link.springer.com/chapter/10.1007/978-3-642-24800-9_6

BERNERS-LEE, Tim, 2006. *Linked Data - Design Issues*.
Dostupné z: http://www.w3.org/DesignIssues/LinkedData.html

BLOEHDORN, Stephan a HOTHO, Andreas, 2004. Boosting for Text Classification with Semantic Features. In: *WebKDD*. s. 149–166. ISBN 978-3-540-47127-1. Dostupné z: doi:10.1007/11899402_10

BOLDI, Paolo, BONCHI, Francesco, CASTILLO, Carlos a VIGNA, Sebastiano, 2009. Voting in social networks. In: *CIKM'09*. New York, USA: ACM Press. ISBN 9781605585123. Dostupné z: doi:10.1145/1645953.1646052

BOLLEN, Johan, RODRIQUEZ, Marko A. a VAN DE SOMPEL, Herbert, 2006. Journal status. *Scientometrics*. 69(3), 669–687. Dostupné z: doi:10.1007/s11192-006-0176-z

BORGATTI, S P, EVERETT, M G a FREEMAN, L C, 2002. Ucinet for Windows: Software for Social Network Analysis. *Harvard Analytic Technologies*. Dostupné z: http://www.analytictech.com/downloaduc6.htm

BORODIN, Allan, ROBERTS, Gareth O., ROSENTHAL, Jeffrey S. a TSAPARAS, Panayiotis, 2005. Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology*. 5(1), 231–297. ISSN 15335399. Dostupné z: doi:10.1145/1052934.1052942

BRAUN, Tibor, GLÄNZEL, Wolfgang a SCHUBERT, András, 2006. A Hirsch-type index for journals. *Scientometrics*. 69(1), 169–173. ISSN 01389130. Dostupné z: doi:10.1007/s11192-006-0147-4

BRIN, Sergey a PAGE, Lawrence, 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*. 30(1-7), 107–117. ISSN 01697552.
Dostupné z: doi:10.1016/S0169-7552(98)00110-X

BURGES, Christopher J.C., 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*. 2(2), 121–167. ISSN 1573-756X. Dostupné z: doi:10.1023/A:1009715923555

COHEN, William W. a SINGER, Yoram, 1999. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems*. 17(2), 141–173. Dostupné z: doi:10.1145/306686.306688

COVER, T. a HART, P., 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. 13(1), 21–27. ISSN 0018-9448. Dostupné z: doi:10.1109/TIT.1967.1053964

DE MELO, G a SIERSDORFER, S, 2007. Multilingual Text Classification Using Ontologies. *Advances in Information Retrieval*. 4425, 541–548. Dostupné z: http://www.springerlink.com/index/m1q3347j9t16541n.pdf

DEERWESTER, S C, DUMAIS, S T, FURNAS, G W, HARSHMAN, R A, LANDAUER, T K, LOCHBAUM, K E a STREETER, L A, 1989. Computer information retrieval using latent semantic structure. 1989. Dostupné z: http://www.google.com/patents/US4839853

DENNIS, Wayne, 1954. Bibliographies of eminent scientists. *The Scientific Monthly*. 79, 180–183.

DI CARO, Luigi, CATALDI, Mario a SCHIFANELLA, Claudio, 2012. The d-index: Discovering dependences among scientific collaborators from their bibliographic data records. *Scientometrics*. 93(3), 583–607. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-012-0762-1

DING, Ying, 2011a. Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology*. 62(2), 236–245. Dostupné z: doi:10.1002/asi.21452

DING, Ying, 2011b. Topic-based PageRank on author cocitation networks. *Journal of the American Society for Information Science and Technology*. 62(2), 449–2011. ISSN 15322882. Dostupné z: doi:10.1002/asi.21467

DING, Ying a CRONIN, Blaise, 2011. Popular and/or Prestigious? Measures of Scholarly Esteem. *Information Processing & Management*. 47(1), 80–96. ISSN 03064573. Dostupné z: doi:10.1016/j.ipm.2010.01.002

DING, Ying, YAN, Erjia, FRAZHO, Arthur a CAVERLEE, James, 2009. PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*. 60(11), 2229–2243. Dostupné z: doi:10.1002/asi

DORTA-GONZÁLEZ, P. a DORTA-GONZÁLEZ, M. I., 2012. Comparing journals from different fields of science and social science through a JCR subject categories normalized impact factor. *Scientometrics*. 95(2), 645–672. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-012-0929-9

DOSTAL, Martin, NYKL, Michal a JEŽEK, Karel, 2013. Cluster labeling with Linked Data. *Journal of Theoretical and Applied Information Technology*. 53(3), 340–345. ISSN 1992-8645.

DOSTAL, Martin, NYKL, Michal a JEŽEK, Karel, 2014a. Exploration of Document Classification with Linked Data and PageRank. In: *Intelligent Distributed Computing VII*. Praha (CZE): Springer International Publishing, s. 37–43. ISBN 978-3-319-01570-5. Dostupné z: doi:10.1007/978-3-319-01571-2_6

DOSTAL, Martin, NYKL, Michal a JEŽEK, Karel, 2014b. Semantic analysis of software specifications with Linked Data. *Journal of Theoretical and Applied Information Technology*. 67(2), 368–376. ISSN 1992-8645.

EGGHE, Leo, 2006. An improvement of the h-index: The g-index. *ISSI newsletter*. 2(1), 8–9. Dostupné z: http://pds4.egloos.com/pds/200703/08/11/g_index.pdf

EGGHE, Leo, 2013. Theory and practise of the g-index. *Scientometrics*. 69(1), 131–152. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-006-0144-7

EGGHE, Leo a ROUSSEAU, Ronald, 2003. A General Framework for Relative Impact Indicators. *Canadian Journal of Information and Library Science*. 27(1), 29–48. ISSN 1195-096X.
Dostupné z: https://uhdspace.uhasselt.be/dspace/handle/1942/767

EGGHE, Leo, ROUSSEAU, Ronald a VAN HOOYDONK, Guido, 2000. Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *Journal of the American Society for Information Science and Technology*. 51(2), 145–157. ISSN 15322882. Dostupné z: doi:10.1002/(SICI)1097-4571(2000)51:2<145::AID-ASI6>3.0.CO;2-9

ELKINS, Mark R., MAHER, Christopher G., HERBERT, Robert D., MOSELEY, Anne M. a SHERRINGTON, Catherine, 2010. Correlation between the Journal Impact Factor and three other journal citation indices. *Scientometrics*. 85(1), 81–93. ISSN 01389130. Dostupné z: doi:10.1007/s11192-010-0262-0

ERA, Australian Government, 2009. *Excellence in Research for Australia: Evaluation Guidelines for the 2009*. Canberra (AUS): Commonwealth of Australia. ISBN 9780980620412.

ERKAN, Günes a RADEV, Dragomir R., 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*. 22, 457–479.
Dostupné z: http://www.aaai.org/Papers/JAIR/Vol22/JAIR-2214.pdf

FERRARA, Emilio, 2012. *Mining and Analysis of Online Social Networks*. Messina. University of Messina. Dostupné z: http://www.emilio.ferrara.name/wp-content/uploads/2011/06/thesis.pdf

FIALA, Dalibor, 2011. Mining citation information from CiteSeer data. *Scientometrics*. 86(3), 553–562. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-010-0326-1

FIALA, Dalibor, 2012a. Bibliometric analysis of CiteSeer data for countries. *Information Processing & Management*. 48(2), 242–253. ISSN 03064573. Dostupné z: doi:10.1016/j.ipm.2011.10.001

FIALA, Dalibor, 2012b. Time-aware PageRank for bibliographic networks. *Journal of Informetrics*. 6(3), 370–388. ISSN 17511577. Dostupné z: doi:10.1016/j.joi.2012.02.002

FIALA, Dalibor, 2013. Suborganizations of Institutions in Library and Information Science Journals. *Information*. 4(4), 351–366. ISSN 2078-2489. Dostupné z: doi:10.3390/info4040351

FIALA, Dalibor, 2014. Sub-organizations of institutions in computer science journals at the turn of the century. *Malaysian Journal of Library & Information Science*. 19(2), 53–68. ISSN 1394-6234. Dostupné z: http://majlis.fsktm.um.edu.my/document.aspx?FileName=1491.pdf

FIALA, Dalibor, ROUSSELOT, François a JEŽEK, Karel, 2008. PageRank for bibliographic networks. *Scientometrics*. 76(1), 135–158. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-007-1908-4

FIALA, Dalibor, ŠUBELJ, Lovro, ŽITNIK, Slavko a BAJEC, Marko, 2015. Do PageRank-based author rankings outperform simple citation counts? *Journal of Informetrics*. 9(2), 334–348. ISSN 17511577. Dostupné z: doi:10.1016/j.joi.2015.02.008

FRANCESCHINI, Fiorenzo, MAISANO, Domenico a MASTROGIACOMO, Luca, 2013. The effect of database dirty data on h-index calculation. *Scientometrics*. 95(3), 1179–1188. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-012-0871-x

FRANCESCHINI, Fiorenzo, MAISANO, Domenico, PEROTTI, Anna a PROTO, Andrea, 2010. Analysis of the ch-index: an indicator to evaluate the diffusion of scientific research output by citers. *Scientometrics*. 85(1), 203–217. ISSN 01389130. Dostupné z: doi:10.1007/s11192-010-0165-0

FRANDSEN, Tove Faber, 2004. Journal diffusion factors – a measure of diffusion? *Aslib Proceedings*. 56(1), 5–11. ISSN 0001-253X. Dostupné z: doi:10.1108/00012530410516822

FREEMAN, Linton C., 1977. A set of measures of centrality based on betweenness. *Sociometry*. 40(1), 35–41. ISSN 00380431. Dostupné z: doi:10.2307/3033543

FREEMAN, Linton C., 1979. Centrality in social networks conceptual clarification. *Social networks*. 1(3), 215–239. ISSN 0378-8733. Dostupné z: doi:10.1016/0378-8733(78)90021-7

FREEMAN, Linton C., BORGATTI, Stephen P. a WHITE, Douglas R., 1991. Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks*. 13(2), 141–154. ISSN 03788733. Dostupné z: doi:10.1016/0378-8733(91)90017-N

GABRILOVICH, Evgeniy a MARKOVITCH, Shaul, 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: *IJCAI International Joint Conference on Artificial Intelligence*. s. 1606–1611. ISBN 9781450307178. Dostupné z: doi:10.1145/2063576.2063865

GARFIELD, Eugene, 1955a. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*. 122, 108–111. ISSN 0036-8075. Dostupné z: doi:10.1126/science.122.3159.108

GARFIELD, Eugene, 1955b. Science Citation Index. *Library*. 122(3159), 108–111. ISSN 0363-0277. Dostupné z: http://thomsonreuters.com/products_services/science/science_products/a-z/science_citation_index

GARFIELD, Eugene, 1972. Citation analysis as a tool in journal evaluation. *Science*. 178(60), 471–479. ISSN 0036-8075. Dostupné z: doi:10.1126/science.178.4060.471

GARFIELD, Eugene, 1999. Journal impact factor: a brief review. *Canadian Medical Association journal*. 161(8), 979–980. ISSN 0820-3946. Dostupné z: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1230709

GAUFFRIAU, Marianne a LARSEN, Peder Olesen, 2005. Counting methods are decisive for rankings based on publication and citation studies. *Scientometrics*. 64(1), 85–93. ISSN 01389130. Dostupné z: doi:10.1007/s11192-005-0239-6

GILES, C. Lee, BOLLACKER, Kurt D. a LAWRENCE, Steve, 1998. CiteSeer: an automatic citation indexing system. In: *Proceedings of the third ACM conference on Digital libraries - DL '98*. New York: ACM Press, s. 89–98. ISBN 0897919653. Dostupné z: doi:10.1145/276675.276685

GLÄNZEL, Wolfgang, 2001. National characteristics in international scientific co-authorship relations. *Scientometrics*. 51(1), 69–115. ISSN 1588-2861. Dostupné z: doi:10.1023/A:1010512628145

GONZÁLEZ-PEREIRA, Borja, GUERRERO-BOTE, Vicente P. a MOYA-ANEGÓN, Félix, 2010. A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*. 4(3), 379–391. ISSN 17511577. Dostupné z: doi:10.1016/j.joi.2010.03.002

GROSS, P. L. K. a GROSS, E. M., 1927. College Libraries and Chemical Education. *Science*. 66(1713), 385–389. ISSN 0036-8075. Dostupné z: doi:10.1126/science.66.1713.385

HADDOW, Gaby a GENONI, Paul, 2010. Citation analysis and peer ranking of Australian social science journals. *Scientometrics*. 85(2), 471–487. ISSN 01389130. Dostupné z: doi:10.1007/s11192-010-0198-4

HAGEN, Nils T, 2010. Harmonic publication and citation counting: sharing authorship credit equitably - not equally, geometrically or arithmetically. *Scientometrics*. 84(3), 785–793. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-009-0129-4

HAN, Yo-Sub, KIM, Laehyun a CHA, Jeong-Won, 2012. Computing user reputation in a social network of web 2.0. *Computing and Informatics*. 31(2), 1001–1016. ISSN 1335-9150. Dostupné z: http://air.changwon.ac.kr/wp-content/uploads/2012/01/2012CAI.pdf

HANNEMAN, Robert A. a RIDDLE, Mark, 2005. *Introduction to social network methods* [vid. 16. srpen 2015]. Dostupné z: http://faculty.ucr.edu/~hanneman/nettext/

HAO, Fei, PEI, Zheng, ZHU, Chunsheng, WANG, Guojun a YANG, Laurence T., 2012. User attractor: An operator for the evaluation of social influence. *Future Generation Computer Systems*. ISSN 0167739X. Dostupné z: doi:10.1016/j.future.2012.04.005

HARZING, Anne-Wil, 2013. A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. *Scientometrics*. 94(3), 1057–1075. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-012-0777-7

HAVELIWALA, T H, 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*. 15(4), 784–796. ISSN 10414347. Dostupné z: doi:10.1109/TKDE.2003.1208999

HAWKINS, Douglas M., 2004. The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*. 44(1), 1–12. ISSN 00952338. Dostupné z: doi:10.1021/ci0342472

HEFCE, Higher Education Funding Council for England, 2009. Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework. *The Research Excellence Framework*. 39, 1–24. Dostupné z: http://www.hefce.ac.uk/pubs/hefce/2009/09_39/09_39.pdf

HELLER, Petr, NYKL, Michal a JEŽEK, Karel, 2011. PageRank and analysis of citation cycles. In: *ITAT 2011*. Vrátna Dolina (SVK): CEUR-WS.org, s. 89–90. ISBN 978-80-89557-01-1.

HIRSCH, J. E., 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*. 102(46), 16569–16572. ISSN 1091-6490. Dostupné z: doi:10.1073/pnas.0507655102

HIRSCH, J. E., 2010. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*. 85(3), 741–754. ISSN 01389130. Dostupné z: doi:10.1007/s11192-010-0193-9

HO, Yuh-Shan, 2013. The top-cited research works in the Science Citation Index Expanded. *Scientometrics*. 94(3), 1297–1312. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-012-0837-z

CHEN, P, XIE, H, MASLOV, S a REDNER, S, 2007. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*. 1(1), 8–15. ISSN 17511577. Dostupné z: doi:10.1016/j.joi.2006.06.001

CHIRITA, Paul Alexandru, NEJDL, Wolfgang, SCHLOSSER, Mario a SCURTU, Oana, 2004. Personalized Reputation Management in P2P Networks. In: *ISWC Workshop on Trust, Security, and Reputation on the Semantic Web*. Hiroshima, JP: CEUR-WS.org. Dostupné z: http://www.l3s.de/~chirita/publications/chirita04personalized.pdf

JACSÓ, Péter, 2011. The pros and cons of Microsoft Academic Search from a bibliometric perspective. *Online Information Review*. 35(6), 983–997. ISSN 1468-4527. Dostupné z: doi:10.1108/14684521111210788

JAFFRI, Afraz, GLASER, Hugh a MILLARD, Ian, 2008. URI disambiguation in the context of linked data. In: *Linked Data on the Web*. Being (CHN): CEUR-WS.org. ISSN 1613-0073.

JIN, BiHui, LIANG, LiMing, ROUSSEAU, Ronald a EGGHE, Leo, 2007. The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*. 52(6), 855–863. ISSN 1001-6538. Dostupné z: doi:10.1007/s11434-007-0145-9

KENDALL, M. G., 1938. A new measure of rank correlation. *Biometrika*. 30(1), 81–93. ISSN 0006-3444. Dostupné z: doi:10.2307/2332226

KLEINBERG, Jon M., 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*. 46(5), 604–632. ISSN 00045411. Dostupné z: doi:10.1145/324133.324140

LANG, Ken, 1995. NewsWeeder: Learning to Filter Netnews. In: *Proceedings of the 12th Conference on Machine Learning*. s. 331–339. Dostupné z: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.22.6286

LANGVILLE, Amy N. a MEYER, Carl D., 2006. *Google's PageRank and Beyond The Science of Search Engine Rankings*. Princeton, NJ, USA: Princeton University Press. ISBN 9780691152660.

LEHMAN, Harvey C., 1954. Men's creative production rate at different ages and in different countries. *The Scientific Monthly*. 78, 321–326. ISSN 00963771.

LEMPEL, Ronny a MORAN, Shlomo P., 2000. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks*. 33(1-6), 387–401. ISSN 13891286. Dostupné z: doi:10.1016/S1389-1286(00)00034-7

LEMPEL, Ronny a MORAN, Shlomo P., 2001. SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*. 19(2), 131–160. ISSN 10468188. Dostupné z: doi:10.1145/382979.383041

LEY, Michael, 1993. *DBLP.uni-trier.de: Computer Science Bibliography*. Dostupné z: http://dblp.uni-trier.de/

LEYDESDORFF, Loet, 2013. An evaluation of impacts in „Nanoscience & nanotechnology": steps towards standards for citation analysis. *Scientometrics*. 94(1), 35–55. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-012-0750-5

LI, J a WILLETT, P, 2009. ArticleRank: a PageRank-based alternative to numbers of citations for analysing citation networks. *Aslib Proceedings*. 61(6), 605–618. ISSN 0001-253X. Dostupné z: doi:10.1108/00012530911005544

LIBEN-NOWELL, David a KLEINBERG, Jon, 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*. 58(7), 1019–1031. ISSN 15322882. Dostupné z: doi:10.1002/asi.20591

LIN, Lili, XU, Zhuoming, DING, Ying a LIU, Xiaozhong, 2013. Finding topic-level experts in scholarly networks. *Scientometrics*. 97(3), 797–819. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-013-0988-6

LINDSEY, D., 1980. Production and Citation Measures in the Sociology of Science: The Problem of Multiple Authorship. *Social Studies of Science*. 10(2), 145–162. ISSN 0306-3127. Dostupné z: doi:10.1177/030631278001000202

LIU, Xiaoming, BOLLEN, Johan, NELSON, Michael L. a VAN DE SOMPEL, Herbert, 2005. Co-Authorship Networks in the Digital Library Research Community. *Information Processing & Management*. 41(6), 1462–1480. Dostupné z: doi:10.1016/j.ipm.2005.03.012

LÓPEZ, Roque Enfique, BARREDA, Dennis, TEJADA, Javier a CUADROS, Ernesto, 2011. MFSRank: an unsupervised method to extract keyphrases using semantic information. In: *Advances in Artificial Intelligence*. Puebla: Springer Berlin Heidelberg, s. 338–344. Dostupné z: http://www.springerlink.com/index/41157QW828172131.pdf

MA, Nan, GUAN, Jiancheng a ZHAO, Yi, 2008. Bringing PageRank to the citation analysis. *Information Processing & Management*. 44(2), 800–810. ISSN 03064573. Dostupné z: doi:10.1016/j.ipm.2007.06.006

MANNING, Christopher D., RAGHAVAN, Prabhakar a SCHÜTZE, Hinrich, 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press. ISBN 0521865719. Dostupné z: http://dspace.cusat.ac.in/dspace/handle/123456789/2538

MASLOV, Sergei a REDNER, Sidney, 2008. Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *The Journal of neuroscience*. 28(44), 11103–11105. ISSN 0270-6474. Dostupné z: doi:10.1523/JNEUROSCI.0002-08.2008

MIHALCEA, Rada a TARAU, Paul, 2004. TextRank: Bringing order into texts. In: *Proceedings of EMNLP*. Barcelona (ESP): ACL, s. 404–411. Dostupné z: http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf

MINGERS, John a LIPITAKIS, Evangelia, 2010. Counting the citations: a comparison of Web of Science and Google Scholar in the field of business and management. *Scientometrics*. 85(2), 613–625. ISSN 01389130. Dostupné z: doi:10.1007/s11192-010-0270-0

MOED, Henk F., 2005. Citation Analysis in Research Evaluation. In: *Information Knowledge and Science Management*. Dordrecht, NL: Springer, Information science and knowledge management, v. 9, s. 333. ISBN 1402037139. Dostupné z: doi:10.1007/1-4020-3714-7

MOED, Henk F., 2010. Measuring contextual citation impact of scientific journals. *Journal of Informetrics*. 4(3), 265–277. ISSN 17511577. Dostupné z: doi:10.1016/j.joi.2010.01.002

MORRISON, Julie L, BREITLING, Rainer, HIGHAM, Desmond J a GILBERT, David R, 2005. GeneRank: using search engine technology for the analysis of microarray experiments. *BMC bioinformatics*. 6(1), 233–247. ISSN 1471-2105. Dostupné z: doi:10.1186/1471-2105-6-233

MRYGLOD, O., KENNA, R., HOLOVATCH, Yu. a BERCHE, B., 2013. Absolute and specific measures of research group excellence. *Scientometrics*. 95(1), 115–127. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-012-0874-7

NYKL, Michal, 2011. *Vyhodnocování informačních sítí*. Diplomová práce. Plzeň. Západočeská univerzita v Plzni.

NYKL, Michal, CAMPR, Michal a JEŽEK, Karel, 2015. Author ranking based on personalized PageRank. *Journal of Informetrics*. 9(4), 777–799. ISSN 1751-1577. Dostupné z: doi:10.1016/j.joi.2015.07.002

NYKL, Michal a JEŽEK, Karel, 2012. Varianty použití PageRanku pro citační analýzu. In: *DATAKON 2012*. Mikulov (CZE): Technická univerzita v Košiciach, s. 87–97. ISBN 978-80-553-1049-7.

NYKL, Michal, JEŽEK, Karel, DOSTAL, Martin a FIALA, Dalibor, 2013. Linked Data and PageRank based classification. In: *IADIS International Conference Theory and Practice in Modern Computing 2013 (part of MCCSIS 2013)*. Praha (CZE): IADIS Press, s. 61–64. ISBN 978-972893994-6.

NYKL, Michal, JEŽEK, Karel, FIALA, Dalibor a DOSTAL, Martin, 2014. PageRank variants in the evaluation of citation networks. *Journal of Informetrics*. 8(3), 683–692. ISSN 17511577. Dostupné z: doi:10.1016/j.joi.2014.06.005

PAGE, Lawrence, BRIN, Sergey, MOTWANI, Rajeev a WINOGRAD, Terry, 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. 1999. Stanford: Stanford InfoLab. [vid. 18. prosinec 2012]. Dostupné z: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

PINSKI, Gabriel a NARIN, Francis, 1976. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*. 12(5), 297–312. ISSN 03064573. Dostupné z: doi:10.1016/0306-4573(76)90048-0

PRATHAP, Gangan, 2010. The iCE approach for journal evaluation. *Scientometrics*. 85(2), 561–565. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-010-0239-z

RADICCHI, Filippo, FORTUNATO, Santo, MARKINES, Benjamin a VESPIGNANI, Alessandro, 2009. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*. 80(5), 056103. ISSN 1539-3755. Dostupné z: doi:10.1103/PhysRevE.80.056103

RAMAKRISHNANAN, Ganesh a BHATTACHARYYA, Pushpak, 2003. Text Representation with WordNet Synsets Using Soft Sense Disambiguation. *Ingénierie des systèmes d'information*. 8(3), 55–70. ISSN 16331311. Dostupné z: doi:10.3166/isi.8.3.55-70

ROCCHIO, J. J., 1971. Relevance feedback in information retrieval. In: *THE SMART RETRIEVAL SYSTEM: Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall Inc., s. 313–323. ISBN 9780138145255/0138145253. Dostupné z: http://ci.nii.ac.jp/naid/10000036124/en/

ROWLANDS, Ian, 2002. Journal diffusion factors: a new approach to measuring research influence. *Aslib Proceedings*. 54(2), 77–84. ISSN 0001-253X. Dostupné z: doi:10.1108/00012530210435211

RYJÁČEK, Zdeněk, 2001. *Teorie grafů a diskrétní optimalizace 1*. Plzeň [vid. 25. září 2013]. Západočeská univerzita v Plzni. Dostupné z: http://cam.zcu.cz/~ryjacek/students/ps/TGD2.pdf

SALTON, G., 1971. *The SMART Retrieval System*. Englewood Cliffs, NJ: Prentice-Hall, Inc. ISBN 9780138145255/0138145253. Dostupné z: http://dl.acm.org/citation.cfm?id=1102022

SANNI, SA a ZAINAB, AN, 2011. Measuring the influence of a journal using impact and diffusion factors. *Malaysian Journal of Library & Information Science*. 16(2), 127–140. ISSN 1394-6234. Dostupné z: http://arxiv.org/abs/1301.5383

SAYYADI, Hassan a GETOOR, Lise, 2009. FutureRank: Ranking Scientific Articles by Predicting their Future PageRank. In: *The Ninth SIAM International Conference on Data Mining*. Nevada: SIAM, s. 533–544. Dostupné z: doi:10.1137/1.9781611972795.46

SCHAPIRE, Robert E. a SINGER, Yoram, 2000. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*. 39(2-3), 135–168. ISSN 1573-0565. Dostupné z: doi:10.1023/A:1007649029923

SIDIROPOULOS, Antonis a KATSAROS, Dimitrios, 2008. Unfolding the full potentials of H-index for bibliographic ranking. *Unpublished work*. Dostupné z: http://delab.csd.auth.gr/papers/TRhindex06_sk.pdf

SIDIROPOULOS, Antonis a MANOLOPOULOS, Yannis, 2005a. A citation-based system to assist prize awarding. *ACM SIGMOD Record*. 34(4), 54–60. ISSN 01635808. Dostupné z: doi:10.1145/1107499.1107506

SIDIROPOULOS, Antonis a MANOLOPOULOS, Yannis, 2005b. A new perspective to automatically rank scientific conferences using digital libraries. *Information Processing & Management*. 41(2), 289–312. ISSN 03064573. Dostupné z: doi:10.1016/j.ipm.2003.09.002

SIDIROPOULOS, Antonis a MANOLOPOULOS, Yannis, 2006. Generalized comparison of graph-based ranking algorithms for publications and authors. *Journal of Systems and Software*. 79(12), 1679–1700. ISSN 01641212. Dostupné z: doi:10.1016/j.jss.2006.01.011

SPEARMAN, C., 1904. The proof and measurement of association between two things. *The American journal of psychology*. 15(1), 72–101. ISSN 0002-9556. Dostupné z: http://www.jstor.org/stable/10.2307/1412159

STRUBE, Michael a PONZETTO, Simone Paolo, 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In: *Proceedings of the National Conference on Artificial Intelligence*. s. 1419 – 1424. ISBN 9781577352815. Dostupné z: doi:10.1.1.231.9545

TOL, R.S.J, 2008. A rational, successive g-index applied to economics departments in Ireland. *Journal of Informetrics*. 2(2), 149–155. ISSN 17511577. Dostupné z: doi:doi:10.1016/j.joi.2008.01.001

ÚŘAD VLÁDY ČR, 2012. *Metodika hodnocení výsledků výzkumných organizací a hodnocení výsledků ukončených programů (platná pro léta 2010 a 2011 a rok 2012)*. 2012.

ÚŘAD VLÁDY ČR, 2013. *Metodika hodnocení výsledků výzkumných organizací a hodnocení výsledků ukončených programů (platná pro léta 2013 až 2015)*. 2013.

VAN HOOYDONK, G., 1997. Fractional counting of multiauthored publications: Consequences for the impact of authors. *Journal of the American Society for Information Science*. 48(10), 944–945. ISSN 0002-8231. Dostupné z: doi:10.1002/(SICI)1097-4571(199710)48:10<944::AID-ASI8>3.0.CO;2-1

VOEVODSKI, Konstantin, TENG, Shang-Hua a XIA, Yu, 2009. Spectral affinity in protein networks. *BMC systems biology*. 3(1), 112–125. ISSN 1752-0509. Dostupné z: doi:10.1186/1752-0509-3-112

WALKER, Dylan, XIE, Huafeng, YAN, Koon-Kiu a MASLOV, Sergei, 2007. Ranking Scientific Publications Using a Simple Model of Network Traffic. *Journal of Statistical Mechanics: Theory and Experiment*. 2007(6), 4. ISSN 1742-5468. Dostupné z: doi:10.1088/1742-5468/2007/06/P06010

WALTMAN, Ludo, 2012. An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics*. 6(4), 700–711. ISSN 17511577. Dostupné z: doi:10.1016/j.joi.2012.07.008

WALTMAN, Ludo, VAN ECK, Nees Jan, VAN LEEUWEN, Thed N. a VISSER, Martijn S., 2013. Some modifications to the SNIP journal impact indicator. *Journal of Informetrics*. 7(2), 272–285. ISSN 17511577. Dostupné z: doi:10.1016/j.joi.2012.11.011

WANG, G. Alan, JIAO, Jian, ABRAHAMS, Alan S., FAN, Weiguo a ZHANG, Zhongju, 2013. ExpertRank: A topic-

aware expert finding algorithm for online knowledge communities. *Decision Support Systems*. 54(3), 1442–1451. ISSN 01679236. Dostupné z: doi:10.1016/j.dss.2012.12.020

WANG, Wei, DO, Diep Bich a LIN, Xuemin, 2005. Term Graph Model for Text Classification. In: *Advanced Data Mining and Applications*. Berlin, Heidelberg: Springer, s. 19–30. ISBN 978-3-540-27894-8. Dostupné z: doi:10.1007/11527503_5

WEST, Jevin D, BERGSTROM, Theodore C a BERGSTROM, Carl T, 2010. The Eigenfactor Metrics TM: A network approach to assessing scholarly journals. *College & Research Libraries*. 71(3), 236–244. ISSN 0010-0870. Dostupné z: doi:10.1016/j.joi.2010.03.002

WEST, Jevin D., ALTHOUSE, Ben, ROSVALL, Martin, BERGSTROM, Carl T. a BERGSTROM, Theodore C, 2008. *Eigenfactor Score and Article Influence Score: Detailed methods*. Dostupné z: http://www.eigenfactor.org/methods.pdf

WEST, Jevin D., JENSEN, Michael C., DANDREA, Ralph J., GORDON, Gregory J. a BERGSTROM, Carl T., 2013. Author-level Eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *Journal of the American Society for Information Science and Technology*. 64(4), 787–801. ISSN 15322882. Dostupné z: doi:10.1002/asi.22790

XING, Wenpu a GHORBANI, Ali, 2004. Weighted PageRank algorithm. In: *Proceedings of the Second Annual Conference on Communication Networks and Services Research*. Fredericton, CA: IEEE, s. 305–314. ISBN 0-7695-2096-0. Dostupné z: doi:10.1109/DNSR.2004.1344743

YAN, Erjia a DING, Ying, 2009. Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*. 60(10), 2107–2118. ISSN 15322882. Dostupné z: doi:10.1002/asi.21128

YAN, Erjia a DING, Ying, 2010. Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*. 61(8), 1635–1643. ISSN 15322882. Dostupné z: http://onlinelibrary.wiley.com/doi/10.1002/asi.21349/full

YAN, Erjia a DING, Ying, 2011. Discovering author impact: A PageRank perspective. *Information Processing & Management*. 47(1), 125–134. ISSN 03064573. Dostupné z: doi:10.1016/j.ipm.2010.05.002

YAN, Erjia a DING, Ying, 2012. Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*. 63(7), 1313–1326. ISSN 15322882. Dostupné z: doi:10.1002/asi.22680

YAN, Erjia, DING, Ying a SUGIMOTO, Cassidy R., 2011. P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*. 62(3), 467–477. ISSN 15322882. Dostupné z: doi:10.1002/asi.21461

YANG, Zaihan, HONG, Liangjie a DAVISON, Brian D., 2010. Topic-driven multi-type citation network analysis. In: *RIAO '10 Adaptivity, Personalization and Fusion of Heterogeneous Information*. s. 24–31. Dostupné z: http://dl.acm.org/citation.cfm?id=1937055.1937062

YIN, Chun-Yang, ARIS, Mohd Jindra a CHEN, Xi, 2009. Combination of Eigenfactor TM and h-index to evaluate scientific journals. *Scientometrics*. 84(3), 639–648. ISSN 01389130. Dostupné z: doi:10.1007/s11192-009-0116-9

YU, Kun, CHEN, Xiaobing a CHEN, Jianhong, 2012. A multidimensional PageRank algorithm of Literatures. *Journal of Theoretical and Applied Information Technology*. 44(2), 308–315. ISSN 1992-8645.

ZHAO, Dangzhi, 2005. Going beyond counting first authors in author co-citation analysis. In: *Proceedings of the American Society for Information Science and Technology*. Charlotte: Association for Information Science and Technology. ISSN 2373-9231. Dostupné z: doi:10.1002/meet.14504201210

ZHOU, Ding, ORSHANSKIY, Sergey A., ZHA, Hongyuan a GILES, C. Lee, 2007. Co-ranking authors and documents in a heterogeneous network. In: *Seventh IEEE International Conference on Data Mining*. Omaha, USA: IEEE, s. 739–744. Dostupné z: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4470320

ZHU, Wenjia a GUAN, Jiancheng, 2013. A bibliometric study of service innovation research: based on complex network analysis. *Scientometrics*. 94(3), 1195–1216. ISSN 0138-9130. Dostupné z: doi:10.1007/s11192-012-0888-1

Příloha A – Soupis publikovaných článků autora k datu 26. 10. 2015

Nejvýznamnějšími články autora, ze kterých vznikla tato disertační práce, jsou (Nykl a Ježek 2012; Nykl et al. 2014, 2015), při jejichž vytváření byl autor hlavním řešitelem popisovaného problému, a články (Nykl et al. 2013; Dostal et al. 2014a), při jejichž vytváření byl autor jedním ze dvou hlavních řešitelů a byl zodpovědný za část týkající se PageRanku (druhým hlavním řešitelem byl Martin Dostal; vymezení podílu v přínosech práce je uvedeno v úvodu 6. kapitoly). Jedním ze dvou hlavních řešitelů byl autor také v článku (Heller et al. 2011), nicméně tento článek nepovažujeme za příliš důležitý, a proto nebyl v práci popisován. Na vytvoření dalších dvou zmíněných článků (Dostal et al. 2013, 2014b) se autor také podílel, ale nebyl hlavním řešitelem v nich popisovaných problémů.

A.1 Publikace v časopisech

NYKL, Michal, JEŽEK, Karel, FIALA, Dalibor a DOSTAL, Martin, 2014. PageRank variants in the evaluation of citation networks. *Journal of Informetrics*. 8(3), 683–692. ISSN 17511577. Dostupné z: doi:10.1016/j.joi.2014.06.005
Indexováno v: ISI Web of Science (IF=2,412), Scopus (SJR=1,44).

NYKL, Michal, CAMPR, Michal a JEŽEK, Karel, 2015. Author ranking based on personalized PageRank. *Journal of Informetrics*. 9(4), 777-799. ISSN 17511577 Dostupné z: doi:10.1016/j.joi.2015.07.002
Indexováno v: ISI Web of Science (IF=2,412), Scopus (SJR=1,44).

DOSTAL, Martin, **NYKL, Michal** a JEŽEK, Karel, 2013. Cluster labeling with Linked Data. *Journal of Theoretical and Applied Information Technology*. 53(3), 340–345. ISSN 1992-8645.
Indexováno v: Scopus (SJR=0,151).

DOSTAL, Martin, **NYKL, Michal** a JEŽEK, Karel, 2014b. Semantic analysis of software specifications with Linked Data. *Journal of Theoretical and Applied Information Technology*. 67(2), 368–376. ISSN 1992-8645.
Indexováno v: Scopus (SJR=0,151).

A.2 Publikace ve významných sbornících

DOSTAL, Martin, **NYKL, Michal** a JEŽEK, Karel, 2014a. Exploration of Document Classification with Linked Data and PageRank. In: *Intelligent Distributed Computing VII*. Praha (CZE): Springer International Publishing, s. 37–43. ISBN 978-3-319-01570-5. Dostupné z: doi:10.1007/978-3-319-01571-2_6
Indexováno v: ISI Web of Science, Scopus (SJR=0,235).

NYKL, Michal, JEŽEK, Karel, DOSTAL, Martin a FIALA, Dalibor, 2013. Linked Data and PageRank based classification. In: *IADIS International Conference Theory and Practice in Modern Computing 2013 (part of MCCSIS 2013)*. Praha (CZE): IADIS Press, s. 61–64. ISBN 978-972893994-6.
Indexováno v: Scopus

A.3 Ostatní publikace

NYKL, Michal a JEŽEK, Karel, 2012. Varianty použití PageRanku pro citační analýzu. In: *DATAKON 2012*. Mikulov (CZE): Technická univerzita v Košiciach, s. 87–97. ISBN 978-80-553-1049-7.

HELLER, Petr, **NYKL, Michal** a JEŽEK, Karel, 2011. PageRank and analysis of citation cycles. In: *ITAT 2011*. Vrátna Dolina (SVK): CEUR-WS.org, s. 89–90. ISBN 978-80-89557-01-1.

A.4 Citace

Publikace: **NYKL, Michal**, JEŽEK, Karel, FIALA, Dalibor a DOSTAL, Martin, 2014. PageRank variants in the evaluation of citation networks. *Journal of Informetrics*. 8(3), 683–692. ISSN 17511577.

1. citace: LI, Weimin, YE, Zhengbo, XIN, Minjun a JIN, Qun, 2015. Social recommendation based on trust and influence in SNS environments. *Multimedia Tools and Applications*. [on-line]. ISSN 1573-7721. Dostupné z: doi: 10.1007/s11042-015-2732-0

2. citace: FIALA, Dalibor, ŠUBELJ, Lovro, ŽITNIK, Slavko a BAJEC, Marko, 2015. Do PageRank-based author rankings outperform simple citation counts? *Journal of Informetrics*. 9(2), 334–348. ISSN 17511577. Dostupné z: doi:10.1016/j.joi.2015.02.008

3. citace: JACOB, Nibu, EKBIA, Kia a BENTUM, Samuel, 2015. *Mapping the Academic Landscape of Numerical Cognition through Interactive Citation Networks*. [on-line] Dostupné z: http://jacobnibu.info/articles/Visualizing the Academic Landscape of Numerical Cognition.pdf

4. citace: IQBAL, Muhammad a ABID, Malik Muneeb, 2015. Combating anainst Web Spam through Content Features. *International Journal of Computer Science Issues*. 12(4), [on-line]. ISSN 1694-0814 Dostupné z: http://ijcsi.org/papers/IJCSI-12-4-36-44.pdf

Publikace: DOSTAL, Martin, **NYKL, Michal** a JEŽEK, Karel, 2013. Cluster labeling with Linked Data. *Journal of Theoretical and Applied Information Technology*. 53(3), 340–345. ISSN 1992-8645.

1. citace: ALAM, Mansaf a SADAF, Kishwar, 2015. Labeling of Web Search Result Clusters Using Heuristic Search and Frequent Itemset. *Procedia Computer Science*. 46(2015), 216-22. ISSN 1877-0509. Dostupné z: doi:10.1016/j.procs.2015.02.014

Příloha B – Seznam vzorců

Příloha C – Seznam obrázků

Příloha D – Seznam tabulek