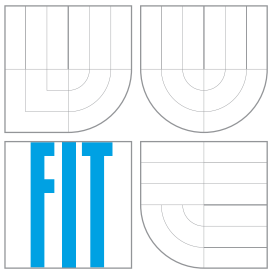
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÉ VYHLEDÁVÁNÍ RELEVANTNÍCH PUBLIKACÍ NA ZÁKLADĚ ANALÝZY CITACÍ

AUTOMATIC SEARCH FOR RELEVANT PUBLICATIONS BY MEANS OF CITATION ANALYSIS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — TOMÁŠ HOLÍK
AUTHOR

VEDOUCÍ PRÁCE — Doc. RNDr. PAVEL SMRŽ, Ph.D.
SUPERVISOR

BRNO 2016

Abstrakt

Cílem této práce je navrhnout a implementovat systém pro automatické vyhledávání relevantních publikací na základě citační analýzy lokálně uložených publikací. V prácí se rozebírá několik metrik pro určení podobnosti článků.

Abstract

The aim of this work is to design and implement system for automatic search of relevant publications by means of citation analysis based on localy saved publications. This work analyse several metrics for assessment of relevance of publication.

Klíčová slova

citační analýza, citace, biblografická citace, vyhledávání, relevantní dokumenty, citační síť

Keywords

citation analysis, citation, reference, search, relevant document, citation network

Citace

HOLÍK, Tomáš. *Automatické vyhledávání relevantních publikací na základě analýzy citací.* Brno, 2016. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Smrž Pavel.

Automatické vyhledávání relevantních publikací na základě analýzy citací

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana doc. RNDr. Pavla Smrže, Ph.D. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Tomáš Holík
18. května 2016

Poděkování

Chtěl bych poděkovat vedoucímu panu doc. Pavlu Smržovi za odborné vedení a pomoc s prací. Dále bych chtěl poděkovat Ondřeji Kurákovi za pomoc s přípravou datové sady pro testovaní.

Obsah

Kapitola 1

Úvod

V dnešní době dochazí k největsímu nárustů dat v celé historii a proto se na popředí dostává vyhledávání, bez kterého si již téměř nedokážeme život představit. Pomáhá nám utřídit znalosti a najít relevantní výsledky v co nejkratším čase. Většína lidi zná vyhledávaní na webu pomoci webových prohlížečů, ale jakým způsobem najít elektronický dokument podobný tomu, který náš zaujal a chceme se o daném tématu dozvědet více? Jednou z možností zpracovat citace v daném článku a najít ostatní publikace, které mají s těmito citacemi nějaký vztah. Abychom mohli určovat tyto vztahy, tak je potřeba mít vytvořenou citační sít pro ucelenou sadu dokumentů. S roustoucí velkostí této sady roste přesnost, ale také náročnost zpracovaní a vyhledávání. Tato práce se zabývá zpracováním velkého množství lokálně uložených publikací a pomocí několika metod určuje podobnost publikací.

Ze začátku se čtenář seznámí s problematikou citační analýzy, její důležitosti ve vědě a metrikami, které se využívají. V třetí kapitole je popsáná testovací datová sada, která byla využita. Čtvrtá kapitola obsahuje seznam využitých technologií a zabývá se návrhem systému a jeho jednotlivých komponent. V páté kapitole jsou popsány implementační detaily a šestá kapitola popisuje na závěr vyhodnocení vytvořeného systému.

Kapitola 2

Analýza problému

2.1 Citace a jejich smysl

Citace jsou již dlouhou dobu nezbytným a důležitým prvkem při tvorbě odborných textů. Slouží nejenom k uznání původního zdroje myšlenky, ale také ke zpětnému dohledání literatury, ze které autor vycházel. Pomáhají nám určit jakým způsobem autor došel k daným závěrům. Jedná se o cenný zdroj informací pro čtenáře i recenzenty textu.

Existuje několik dalších důvodů proč citovat. Ve vědě a tvorbě odborných prací se často navazuje na předchozí práce a základní prameny pro dané téma. Citace tedy určují z jakých předpokladů vycházíme a pokud jsou tyto předpoklady chybné, tak pomohou lokalizovat, kde nastal problém. Pro dodržení citační etiky autor musí zveřejnit všechny původní zdroje, ze kterých čerpal informace a pokud tak neučiní, tak mu může hrozit porušení autorského zákona.

Podle Kábrta [6] správná citační praxe nejen usnadňuje identifikaci citovaného díla, ale zároveň usnadňuje i provádění citační analýzy (vysledování v citačním řetězci počátek objevu, pracovní vazby vědeckých týmů, vědecké fronty apod.), která je vítaným pomocníkem při zkoumání dějin a rozvoje vědy i odhadů budoucího vývoje.

2.2 Citační analýza

Jedná se o matematicko-statistickou bibliometrickou metodu, která se zabývá citovaností dokumentů a četností citací v dalších vědeckých publikacích.

Podle toho kolikát byl článek citován v jiných odborných publikacích lze zjistit jeho dopad na danou disciplínu/téma. Jestliže je toto číslo velké, tak to může znamenat, že článek se stal objektem diskuze nebo i kritiky. Používá se několik metrik pro zpracování citační analýzy, které jsou zmíněny v kapitole 2.5.

Analýza citací se také používá pro detekci plagiátu publikací, ale zatím se nejedná o rozšířenou metodu pro určování plagiátu, i když existuje několik odborných prací zabývající se touto tématikou a jejími výhodami [3]. Tato práce se zabýva využitím analýzy citací pro hledání podobných dokumentů za pomocí několika metod.

Citační analýza se běžně používá v akademické sféře pro vyhodnocení fakultního výzkumu a publikací jednotlivých lidí. Na kvalitu a kvantitu těchto údajů se bere ohled při příjmacím procesu nových kandidátů, určování nástupního platu a i povýšení, ale také mohou pomoc při zíkávání grantů [8]. Při interpretaci výsledků citační analýzy je třeba

přihlížet k možnostem a omezením, za které bývá kritizována a je poukazováno na její nedostatky, které plynou i z toho, že autoři nedodržují citační etiku a tím se dopouští chyb při citování.

Mezi nejčastější chyby se řadí takzvané autocitace tedy citování vlastních děl bez souvislosti s daným tématem. Nezahrnutí všech zdrojů, ze který autor získal informace. Citování publikací, které autor nepoužil, ale jsou například považovány za důležité v daném oboru [9]. Jako další prohřešek se uvádí, když autor záměrně cituje či necituje dílo na základě sympatií k autorovi tohoto díla. V neposlední řadě zde patří i nepřesné citování, které znemožnuje identifikaci díla.

2.3 Metody citování a odkazování

2.3.1 Citace

Citací se označuje část textu, myšlenky nebo závěru, který je přebírán z jiného zdroje. Dělí se na přímé, kterými rozumíme převzetí části textu z cizího díla zcela bez úprav, a nepřímé, u kterých prezentujeme využité myšlenky či fakta v upravené podobě, jako je například parafráze.

2.3.2 Bibliografická citace

Bibliografická citace je záznam popisující zdroj, ze kterého je citace přebírána. Standartně jsou bibliografické citace u odborných prací umístěny na konci dokumentu. Výhodou tohoto umístění je přehlednost, protože všechny použité zdroje jsou na jednom místě v pořadí, v jakém je na ně odkazováno v textu. Seznam použitých bibliografických citací se označuje jako soupis bibliografický citací, seznam použité literatury a nebo seznam použitých zdrojů.

Struktura bibliografické citace se skládá z následujících několika údajů, které jsou seřazeny tak, jak se po sobě uvádí v citaci. Podtržené údaje jsou povinné, stejně jako psaní názvů kurzívou [11].
Struktura citace:

Primární odpovědnost. *Název díla: podnázev díla.* Alternativní odpovednost; Sekundární odpovědnost. Označení vydání. Místo vydání: Jméno nakladatele, Rok vydání. Rozsah díla. Edice. Poznámky. Standardní číslo.

2.3.3 Odkazy na citace v textu

Primární funkcí odkazu je propojení konkrétní citace s odpovídající bibliografické citací. Následující metody se používájí pro odkazování v textu [11].

Forma průběžných poznámek

Dokumenty jsou odkazovány pomocí čísla poznámky. Citace jsou uvedené v poznámce pod čarou, ale mohou být uvedený také na konci dokumentu v soupisu. Číslování může začínat na každe stránce od čísla 1, a nebo může být číslovaní průběžné v celém dokumentu.

Např.: **Text s poznámkami**

. . . jak ukazuje VANĚK[1] i někteří další autoří. . .

V poznámkách pod čarou vypadají záznamy následovně:

---

[1] VANĚK, Jiří. Obecná, ekonomická a informační etika. Praha: Wolters Kluwer Česká republika, 2010, 252 s. : il., portréty. ISBN 9788073575045.

[2] BEAZLEY, David M a Brian K JONES. Python cookbook. 3nd ed. Beijing: O'Reilly, 2013, xvi, 687 s. ISBN 9781449340377.

Forma číselného odkazu

Odkaz na bibliografické citace je tvořen číslem, které je pro danou práci v textu vždy stejné, uvozeným v kulatých závorkách, hranatých závorkách nebo horním indexu.

Např.: **Text s odkazy**

. . . jak uvádí ve své práci(1). . .

V soupisu bibliografický citací vypadají záznamy následovně:

1. VANĚK, Jiří. Obecná, ekonomická a informační etika. Praha: Wolters Kluwer Česká republika, 2010, 252 s. : il., portréty. ISBN 9788073575045.

2. BEAZLEY, David M a Brian K JONES. Python cookbook. 3nd ed. Beijing: O'Reilly, 2013, xvi, 687 s. ISBN 9781449340377.

Forma uvádění prvního prvku a data

První prvek a rok vydání odkazovaného dokumentu jsou uvedeny v textu. *Beayley(2010)*

V případě, že se jméno autora přirozeně vyskytuje v textu, tak následuje pouze rok v kulatých závorkách.

. . . například Beayley (2010) tvrdí . . .

Mají-li dva nebo více dokumentů stejný první prvek a rok, tak se pro odlišení přidají za rokem vydání malá písmena abecedy (a-z).

Beayley(2010a) a *Beayley(2010b)*

Pokud má bibliografická citace více autorů, tak se v odkazu uvádí všichni tito autoři.

Beayley a Jones(2010)

V soupisu citací vypadají záznamy následovně:
Beazley, D.M. & Jones, B.K., 2013. Python cookbook 3nd ed., Beijing: O'Reilly.

2.4 Existující řešení

V této oblasti existuje několik řešení, které se liší jak množstvím zpracovaných článků, způsobem jakým indexují data a také službami, které navíc uživateli poskytují. Klasické

digitální knihovny jako například ACM Digital Library, IEEE Xplore a PubMed využívají model, kdy výzkumní pracovníci nemají vliv na to, jestli jejich článek bude naindexovaný a přístupný v jejich digitální knihovně. Musí publikovat v publikaci, kterou vydavatel indexoval.

Oproti tomu vědecké vyhledávače jako Google Scholar, CiteSeerX a Microsoft Academic Research indexují elektronické soubory z jakýchkoli zdrojů dostupných na webu i takových, které nemusí být ověřené. V důsledku toho je množství naindexovaný článků v těchto digitálních knihovnách mnohem vyšší. Oproti klasickým digitálním knihovnám zde vědci mohou ovlivnit, jestli je jejich článek indexován. Tento model zjednodušuje přístup k vědeckým článkům a zároveň je nezávislý na vydavateli. S volností tohotu modelu ale klesá jeho bezpečnost. Bylo zjištěno, že vědecké vyhledávače nejsou odolné vůči spamu [2]. Je tedy možné naindexovat články, které byly vygenerovány a obsahují absolutní nesmysly. Tímto lze uměle navyšovat počet citací a některé metriky pro výpočet relevance článku mohou být tímto velice ovlivněny a měly by být brány s rezervou.

V zájmu vědeckých pracovníků je, aby jejich články byly indexovány v co nejvíce vědeckých vyhledávačích a digitálních knihovnách, protože jim to pomůže, aby jejich práce získala pozornost co největší vědecké komunity. Zároveň je důležité, jak vysoko jsou umístěny v žebříčku vyhledávání, protože publikace, které jsou umístěny mezi prvními jsou častěji citovány. Vzniká takzvaný Matthew Effect [10], který upozorňuje, že publikace, které získaly velkou pozornost a oblíbenost, bývají mnohem více citovány kvůli tomu, že jsou umístěny na horních pozicích ve vyhledávání. Někteří autoři díky tomuto automaticky předpokládají, že by je měli citovat také. V důsledku toho, určitá část publikací bývá mnohem více citována než ostatní.

Tento souhrn webových aplikací není popis všech existujících řešení, ale pouze vybraná část.

2.4.1 CiteSeerX

CiteSeerX vychází z původní služby CiteSeer, která byla jedna z prvních digitálních knihoven, jež poskytovala automatický systém pro indexovaní citací. Tento systém autonomně indexoval akademickou a vědeckou literaturu v elektronickém formátu.

Dříve se indexovaly hlavně články ze specifických žurnálů a muselo se manuálně zasahovat u indexovaní citací. Tento přístup byl však neflexibilní a také byl kritizován kvůli tomu, že některé důležité články byly například pouze ve sborníku konferencí a nemusely být vůbec naindexovány, protože se neobjevily v žulnálu [7].

CiteSeerX automaticky prochází a indexuje dokumenty, které jsou volně dostupné na webu stejně jako Google Scholar. Ostatním volně poskytuje naindexovaná data a metadata článku, které lze stáhnout pomocí programu. Toto je možné díky tomu, že CiteSeerX podporuje standart Open Archives Initiative Protocol for Metadata Harvesting.

Podobné dokumenty lze dohledat pomocí funkce hledání dokumentů s kocitační vazbou 2.5.2.

2.4.2 Google Scholar

Obsahuje vědeckovýzkumné články z placených i volně přístupných zdrojů jako například Pubmed, JSTOR a i Elsevier. Zároveň prochází web a indexuje odborné články. Většinou

Google Scholar najde více citovaných článků, protože indexuje data z více zdrojů. Odhadované množství obsahovaných článků, je přibližne jeden milión. Uživatel má možnost vytvořit profil, pomocí kterého získá přehled o svých pracích a kým byly citovány. Publikace jsou děleny do kategorií a podkategorií s možností řažení podle relevantnosti na základě jejich metrik. Jako metriky jsou zde použity h-index, h5-index a i10-index, které kladou hlavně důraz na to, jak často je publikace citována. Kvůli tomuto býva kritizována za posilování [12] Matthew Effect.

Obsahuje také funkci zobrazení relevantních publikací, která určuje primárně podle podobnosti článků, ale také bere v potaz relevanci každého článku. Přesnou metriku se mi nepodařilo zjistit.

2.4.3 Scopus

Scopus je víceoborová bibliografická a citační databáze firmy Elsevier, která je konkurentem databáze Web of Science v oblasti scientometrie, tedy hodnocení vědeckých výstupů na základě citačních ohlasů. Zahrnuje informační zdroje z největších patentových databází, webu a dalších zdrojů. Přístup na ni není volný, a proto je potřeba mít licenci. Scopus API umožnuje integrovat indexovaná data do vlastních aplikací.

Umožňuje vyhledávání podobných dokumentů sdílející bibliografické citace, autory a nebo klíčové slova s daným článkem.

2.4.4 Web of Science

Multioborová bibliografická a citační databáze se zaměřením na získávání zdrojových dat pro bibliometrii. Databáze Web of Science je součástí portálu Web of Knowledge, který provozuje firma Thomson Reuters. Stejně jako Scopus, není volně přístupný.

Umožňuje vyhledání podobných dokumentů na základě stejných bibliografických citaci.

2.5 Metriky podobnosti dokumentů

Metriky pro určování míry podobnosti dokumentů, spadají hlavně do dvou kategorií. Metriky založené na podobnosti textů, které měří syntaktickou nebo sémantickou shodu dokumentů a metriky využívající citační analýzu vycházející z bibliografického párování 2.5.1 a kocitace 2.5.2.

Z analýzy výše zmíněných existujících řešení vyplývá, že se nejčastěji používají metriky pro určení impaktu článků, autora či žurnálu. Rozmezí metrik pro doporučování podobných publikací na základě citací je ale značně omezené.

Byly vybrány následující metriky pro určování podobnosti dokumentů.

2.5.1 Bibliografické párování

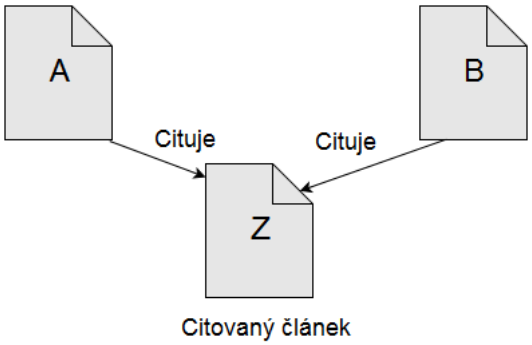
Jedná se o jednu z nejstarších metod citační analýzy jejíž koncept byl vytvořen už v roce 1962 M. M. Kesslerem [5]. V této metrice se podobnost dvou dokumentů určuje pomocí toho, jestli tyto dokumentdy citují společně alespoň jeden další dokument. Čím více společných bibliografických citací tyto dokumenty sdílejí, tím mají mezi sebou větší vazbu.

Pro určení váhy podobnosti dvou článků P_x a P_q, $Sim_{bib}(P_x, P_q)$ byla použila definice [1]:

$$Sim_{bib}(P_q, P_x) = (\text{počet společných bibliografický citací mezi } P_x \text{ a } P_q)/MaxB$$

kde $MaxB$ je číslo určující nejvyšší počet sdílených bibliografických citací mezi dvěma libovolnými články ve využité datové sadě.

Bibliografické párování je považováno za retrospektivní metriku pro určování podobnosti, protože bibliografické citace se u publikace postupem času měnit nemohou, jelikož informace, na základě které se vyhodnocuje vztah mezi dvěma dokumenty je již neměnná.



Obrázek 2.1: Bibliografická vazba mezi článkem A a B

Vyhodnocování podobnosti publikací na základě bibliografického párování je používána ve Scopus a Web Of Science.

2.5.2 Kocitační vazba

V kocitační analýze se dokumenty považují za podobné, pokud jsou citovány společně alespoň v jedné publikaci. Pro určování míry podobnosti dvou děl se používá kocitační index, který je odvozen z toho, jak často jsou tyto dvě publikace citovány společně. Kocitací je tedy míněn vztah mezi dvěma publikacemi, bez přímé vazby mezi sebou.

Výhoda oproti bibliografickému párování je v tom, že míra kocitační vazby mezi dokumenty se může měnit. Navíc výpočet kocitační míry je založen na názoru více autorů a je tedy považována za přesnější indikátor pro určení podonosti článků.

Pro určení váhy podobnosti dvou článků P_x a P_q, $Sim_{bib}(P_x, P_q)$ byla použila definice:

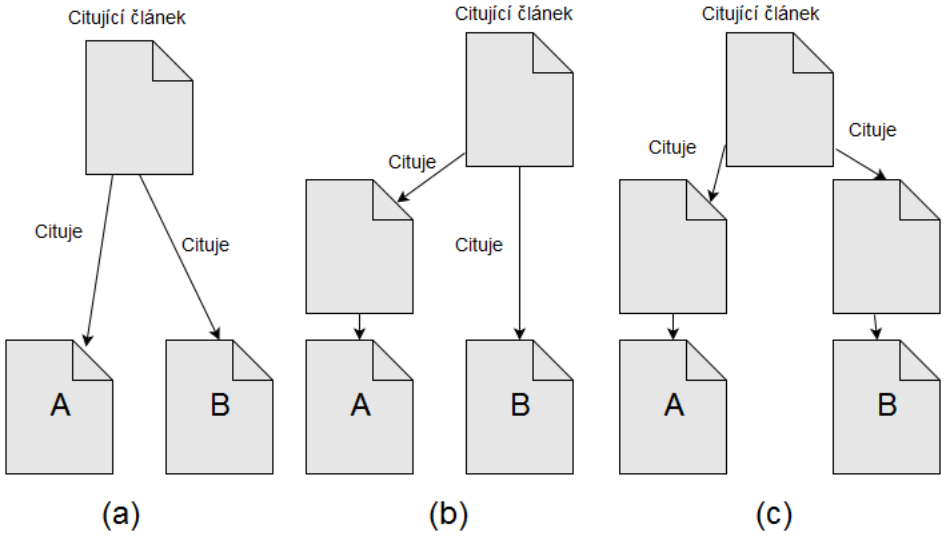
$$Sim_{cocit} = |C_Q \cup C_X|/MaxN$$

kde C_Q a C_X je množina publikací, které citují zaroveň P_q i P_x. $MaxN$ je číslo, které označuje kolikrát nejvíce jsou dvě publikace z datové sady citovány společně.

Z analyzovaných existujících řešení tuto metriku využívá CiteSeerX.

Víceúrovňová kocitační analýza

Zatím byla popsáná takzvaná jednoúrovňová kocitační vazba, ale počet úrovní není ničím omezený a lze jich nastavit libovolný počet. Víceúrovňovou kocitací rozumíme vztah mezi dvěma publikacemi, které nemají přímou vazbu mezi sebou a zároveň nemusí být ani společně citovány ve stejné publikaci.



Obrázek 2.2: Kocitační vazba mezi článkem A a B, (a) jednoúrovňová (b) dvouúrovňová (c) tříúrovňová

S každou další úrovní se exponenciálně zvyšuje náročnost na výpočet, a proto se s touto variantou běžně nesetkáváme.

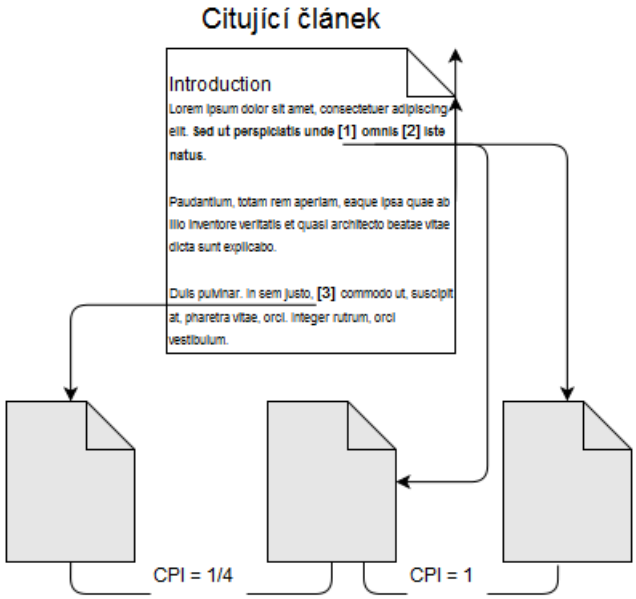
2.5.3 Citation Proximity Analysis

Problém u kocitační analýzy 2.5.2 stejně jako u bibliografické 2.5.1 je v tom, že dávají stejnou váhu všem článkům. Tento problém částečně zmírňuje metrika Citation Proximity Analysis (dále jen CPA), která vychází z kocitační analýzy, ale navíc pro vyhodnocení využívá citačního kontextu. Gipp a Bell [4] navrhnuli CPA za předpokladu, že jsou dva dokumenty kocitovány ve stejné větě, tak jsou si bližší, než kdyby byly kocitovány jen ve stejném odstavci. Tato metrika tedy lze využít pouze pro dokumenty, ke kterým máme jejich fulltextovou reprezentaci.

Tabulka 2.1: CPI váhy podle blízkosti citací

Výskyt	Váha
věta	1
odstavec	1/2
kapitola	1/4
časopis	1/8
konference	1/16

Míra podobnosti dvou dokumentů je vyjadřena pomocí Citation Proximity Index (dále jen CPI). Při návrhu pro určení vah 2.1 podle blízkosti dvou citací jsem vycházel z řešení [4], které je již otestováno v praxi a poskytuje uspokojivé výsledky.



Obrázek 2.3: Ukázka principu Citation Proximity Analysis

2.6 Vstupní formát dat

Digitální knihovny poskytují články hlavně ve formátu Portable Document Format (ve zkratce PDF a dále už jen PDF), což je souborový formát, který ukládá data nezávisle na hardwaru a softwaru, na kterém byly pořízeny.

Kapitola 3

Datová sada ACL anthology

Aby datová sada vyhovovala pro účel citační analýzy, muselo být splněno několik kritérií. U článků bylo potřeba přesně určit jejich metadata jako například název článků, autory, rok vydaní a také musela být k dispozici jejich fulltextová podoba. Zároveň bylo vhodné, aby články citovaly nebo byly citovány co nejvíce ostatními články z této datové sady tzv. vnitřní odkaz.

Těmto požadavkům vyhovovala datová sada, kterou poskytuje zdarma pro výzkumné použití Association for Computational Linguistics (ACL) Anthology. Obsahuje více než 20,000 tisíc článků z oblasti výpočetní lingvistiky od roku 1965 do 2013. U článků je navíc informace o tom, kde přesně byly zvěřejněny. Hlavní kategorie, pod které články spadají jsou konference, workshopy a žurnály. Články byly manuálně rozděleny do podkategorií, které jsou zobrazeny v tabulce.

Zkratka	Název	Vnitřní odkaz
ACL	Annual Meeting of the Association of Computational Linguistics	32
EACL	Annual Meeting of The European Chapter of The Association of Computational Linguistics	24
NAACL	Annual Conference of the North American Chapter of the Association for Computational Linguistics	31
COLING	International Conference on Computational Linguistics	27
CoNLL	International Conference on Computational Natural Language Learning	43
ANLP	Applied Natural Language Processing Conference	23
EMNLP	Conference on Empirical Methods in Natural Language Processing	43
IJCNLP	International Joint Conference on Natural Language Processing	31
HLT	Human Language Technologies	32
INLG	International Conference on Natural Language Generation	13
LREC	International Conference on Language Resources and Evaluation	24
MUC	Message Understanding Conference	21
CL	Computational Linguistics Journal	30

Tabulka 3.1: ACL Anthology rozdělení článků do kategorií

Jednotlivé články v datové sadě obsahují vlastní unikátní identifikátor, který je uložen i v názvu PDF souboru. Je tedy jednoduše možné získat metadata pro zpracovávaný článek.

Kapitola 4

Návrh řešení

Cílem teto práce je vytvořit řešení, které je schopné vytvořit citační síť z velkého množství lokálně uložených elektronických článků a dokáže určovat podobnost dokumentů na základě citační analýzy. K vyhodnocení relevantnosti článku bude využito několik metrik, které jsou probrány v sekci 2.5. Většína digitálních knihoven včetně naší vybrané testovací sady, obsahují a poskytují články v elektronickém formátu PDF a z tohoto důvodu bude povolený formát pro vstupní data pouze PDF.

Celý projekt lze rozdělit na dva hlavní nezávislé moduly. První modul 4.1, který obstarává konverzi elektronických souborů do textového formátu, extrakci citací a indexaci dat. Druhý modul 4.2, který se skládá z webové aplikace, kterou může uživatel využívat pro zobrazování naidexovaných článků, ale také pro vyhledávání podobných článků.

4.1 Extrakce a indexace dat

Při návrhu systému je třeba vycházet z předpokladu, že množství článků pro zpracování se bude pohybovat v tisících. Je tedy potřeba brát ohled na rychlost zpracování.

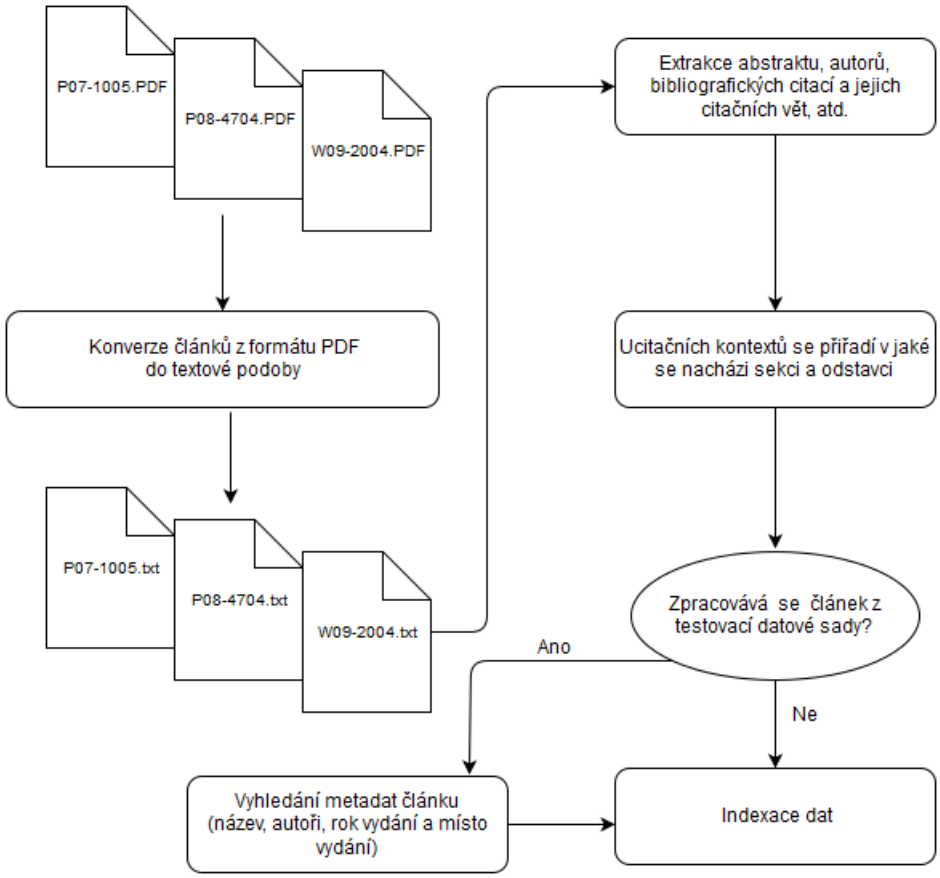
Proces zpracování článků se bude skládat z několika navazujících kroků, které jsou zobrazeny na diagramu 4.1.

Prvním krokem bude převod článků z formátu PDF do jednotné textové podoby. Pro tuto funkci se nabízí využít nějaký nástroj pro konverzi pomocí OCR[10], který funguje na principu optického rozpoznávání znaků a nebo vybrat jeden z volně dostupných nástrojů. Při analýze RRS knihovny jsem narazil na nástroj sloužící k převodu PDF do textu s využitím OCR. Cílem bude vyzkoušet obě varianty a vybrat tu, která bude lépe vyhovovat požadavkům.

Dále bude následovat extrakce jednotlivých informací z textu, která se bude provádět pomocí RRS knihovny. Jedná se o knihovnu vyvinutou v rámci školního projektu ReReSearch (dále jen RRS), která poskytuje širokou paletu funkcí pro extrakci dat. Cílem tedy bude vylepšit úspěšnost extrakce informací, potřebných pro citační analýzu a implementovat novou funkčnost pokud bude potřeba.

Jestliže se právě budou zpracovávat články z datové sady, tak dojde k vyhledání metadat pro daný článek pomocí identifikátoru získaného z názvu právě zpracovávaného souboru.

[1] Optical Character Recognition



Obrázek 4.1: Proces zpracování vstupních článků

Mezi získané metadata patří název článku, autoři, rok publikace a místo vydaní. Pomocí místa vydaní byly články rozděleny do jednotlivých kategorií 3.1.

Jakmile budou všechny informace o článku k dispozici, dojde k indexaci dokumentu.

Systém by měl být navržen s ohledem na rychlost a nejspíš využít paralelní zpracování dat, jelikož proces zpracování bude pro všechny články stejný.

4.2 Webová aplikace

Uživatelé na úvodní stránce budou moci vyhledávat mezi všemi naindexovanými dokumenty. Vyhledávání bude řešeno jako fulltextové, kdy uživatelé mohou speficikovat vyhledávání dle více kritérií. Například vyhledávání v názvu článku, autora či abstraktu.

Hlavní stránka se tedy bude skládat ze seznamu všech článků které vyhovují hledaným kritériím. Pro přehlednost se nebudou zobrazovat všechny, ale pouze část s možností listování jednotlivými stránkami výsledků. Články zde budou mít zobrazeny pouze hlavní informace jako například název článku nebo autoři a bude je možné řadit podle různých kritérií. U každého článku bude odkaz, který přesměruje uživatele na detail daného článku.

Na stránce detailu článku si uživatel bude moci navíc zobrazit abstrakt, bibliografické citace, rok vydání a seznam publikací, které tento článek citovali. Bude zde možnost hledání podobných publikací, kdy si uživatel vybere, která metrika(viz. odkaz) se má použít pro

hledání relevantních publikací.

4.3 Použité technologie

Tato kapitola obsahuje popis použitých technologií a jejich základní popis. Při volbě technologií, které budou využity pro vývoj, jsem upřednostnil ty, se kterými jsem již měl zkušenosti z předchozích projektů.

4.3.1 Symfony2

Symfony2[1] je sada samostatných a znovupoužitelných komponent v jazyce PHP, které řeší běžné problémy spojené s vývojem webových aplikací. Po projení těchto komponent je Symfony2 také plnohodnotný webový aplikační framework pro vývoj webových aplikací. Nejedná se přímo o framework založený na architektuře Model View Controller, protože nechává vývojáři velkou volnost při tom, jakým způsobem bude s modelem dat pracovat. Díky své flexibilitě se hodí pro malé i vetší projekty. Zárověn má výbornou dokumentaci a nástroje pro debugování.

4.3.2 AngularJS

Jedná se o ucelený JavaScriptový webový framework navržený pro programátory, jelikož jeho hlavní autor je původně Java vývojář. Má aktuální dokumentaci, velkou komunitu a nabízí prvky pro testovatelnost. Aplikace tvořené v AngularJS[2] rozšiřují HTML[3] atributy s takzvanýma direktivami a spojují data s HTML pomocí výrazů. Díky tomu lze definovat dynamické šablony přímo v HTML. Obsahuje navíc tzv. two way data binding, který zajisťuje obousměrnou synchronizaci dat mezi prezentační a datovou vrstvou.

4.3.3 Bootstrap

Bootstrap[4] je knihovna kaskádových stylů a JavaScriptových komponent, která značně usnadňuje a urychluje tvorbu responzivních webových stránek. Obsahuje styly pro všechny základní HTML elementy jako například tabulky, formuláře a tlačítka. Navíc nabízí i pokročilejší komponenty sloužící ke stránkování nebo vyskakovací dialogy. Díky tomu je možné vytvořit slušně vypadající webovou stránku v krátkém čase.

4.3.4 Elastic Search

Elastic search[5] je fulltextový vyhledáváč implementovaný v Javě. Je postavený na Apache Lucene[7], které je jeho jádrem a zprostředkovává funkce například pro vyhledávání a indexování dokumentů. Elastic search tedy zapouzdřuje funkčnost Lucene a poskytuje jednoduché

[1] https://symfony.com/
[2] https://angularjs.org/
[3] HyperText Markup Language
[4] http://getbootstrap.com/
[5] https://www.elastic.co/
[7] https://lucene.apache.org/

a více použitelné API. Hlavní protokol přes který elastic search komunikuje je HTTP, formát zpráv je JSON a disponuje RESTful rozhraním, takže díky tomu je projení elastic search a webových aplikací jednoduché.

4.3.5 D3.js

D3.js je open source JavaScriptová knihovna, která uživatelům umožňuje vytváření dynamické a interaktivní vizualizace dat přímo ve webovém prohlížeči.

Kapitola 5

Implementace

Po analýze existujících řešení a počátečním návrhu 4 modulů aplikace začala implementace. Jako první jsem začal vytvářet skripty pro automatizaci převodu PDF souborů do textu a testoval jsem vybrané konvertory.

Dalším krokem byla extrakce bibliografických citací a citačních vět. Analyzoval jsem chyby, ke kterým docházelo při extrakci a snažil jsem se je opravit. U tohoto kroku jsem ze začátku strávil hodně času, což se ukázalo jako chybný přístup. Nyní bych postupoval tak, že bych se snažil dopracovat k prototypu celého systému co nejrychleji a až poté ladit jednotlivé části.

Následovalo vytvoření struktury pro ukládání článků v elastic search a indexace získaných dat. Nad naindexovanými daty jsem začal provádět experimenty, vytvořil skripty v pythonu s algoritmy obsahující metriky pro hledaní podobných článků.

Nakonec jsem vytvořil webovou aplikaci a ladil jednotlivé komponenty.

5.1 Extrakce a indexace dat

Modul pro extrakci a indexaci dat je vytvořen jako konzolová aplikace pomocí programovacího jazyka Python verze 2.7.6. Při vývoji jsem postupně narážel na limity původního návrhu a rozhodl jsem se ho změnit5.1, aby lépe vyhovoval aktuálním potřebám.

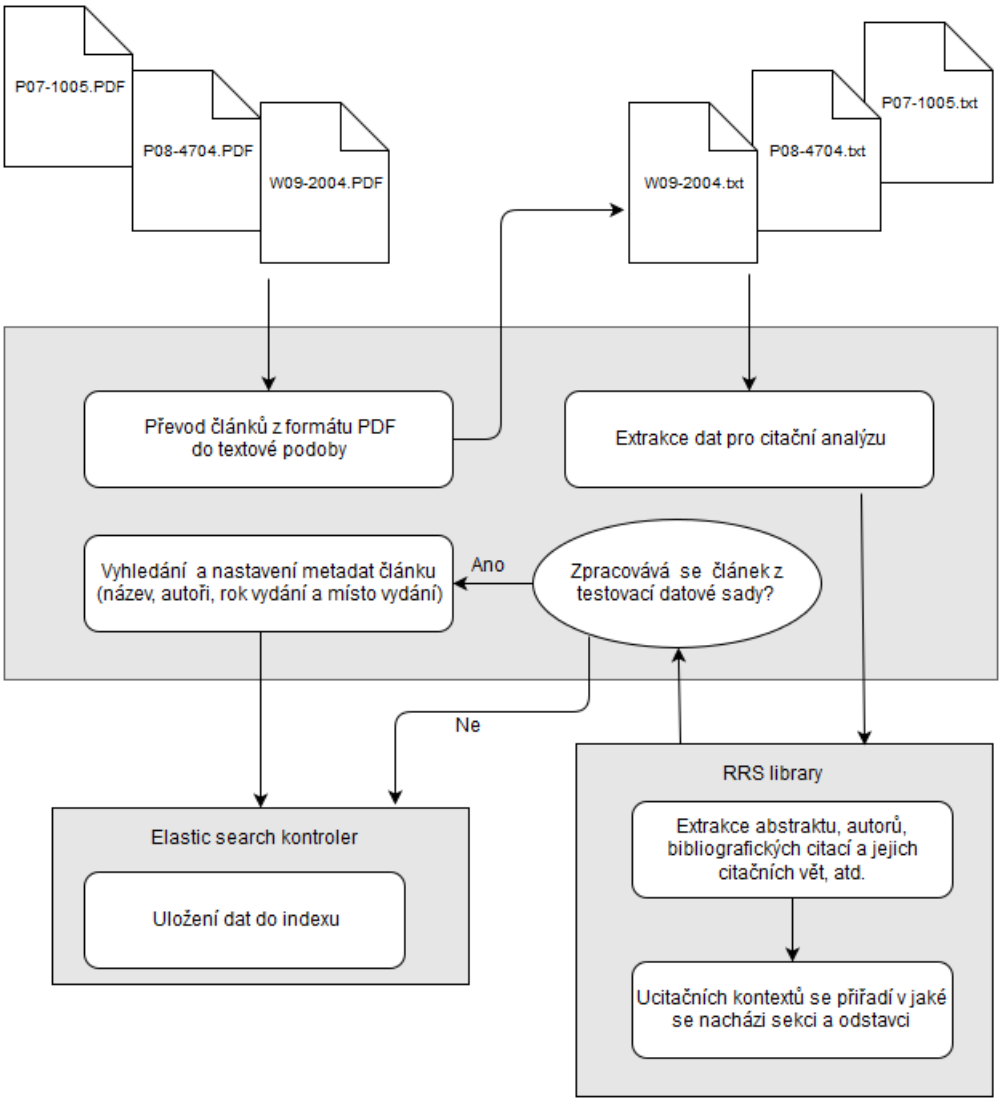
Důvody pro změnu byly například to, že pro testovací účely jsem využíval několik samostatných skriptů a bylo potřeba centralizovat nastavení a přístup k elastic search. Dále jsem chtěl docílit toho, aby jednotlivé komponenty byly znovupoužitelné i samostatně.

5.1.1 Popis činnosti hlavního programu

Po spuštění skriptu dojde k vytvoření několika procesů, jejichž množství může uživatel ovlivit nastavením argumentu, které začnou paralelně provádět hlavní program. Postupně dochází k převodu PDF souborů z vybrané složky do textové podoby, extrakci jednotlivých informací ze článku a nakonec se uloží získané data o článku do indexu.

Jelikož se jedná o paralelní zpracovaní dat, je nutné vyřesit problém synchronizace mezi procesy, aby nedocházelo k tomu, že jeden soubor bude zpracován zbytečně několikrát. Tento problém je vyřešen, jak je znázorněné v diagramu 5.1, pomocí sdílených front, které jsou instance třídy Queue[1]8 z balíčku multiprocessing.

[8] https://docs.python.org/2/library/multiprocessing.html



Obrázek 5.1: Upravená architektura systému pro extrakci a indexaci dat

Hlavní program může také sloužit pouze jako nástroj pro převod dokumentů z formátu PDF do textu a nebo je možne také přeskočit převod souborů PDF a extrahovat data přímo ze souboru v textové podobě.

Pro převod článků jsem původně využíval OCR konvertor z RRS knihovny. Převedené textové dokumenty neobsahovali mnoho chyb, ale kvůli dlouhé době zpracování jsem začal používat nástroj pdftotext**??**, jehož průměrná doba převodu dokumentu je výrazně rychlejší.

Skript implementuje kontrolní mechanismus, který testuje vstupní soubory, jestli již nebyly v minulosti zpracovány a pokud ano, tak je vyřadí. Tento mechanismus je poměrně potřebný, protože zpracování článků je časově náročné a pokud by došlo k přerušení skriptu, například v důsledku výpadku proudu či restartu počítače na kterém skript běží, tak nechceme, aby zpracování muselo běžet zcela od začátku. Navíc to slouží jako kontrola proti vytváření duplikátu.

Dále došlo k úpravám, aby RRS knihovna umožňovala získání všech dat nutných pro výpočet metriky CPA 2.5.3. Původně byla možná extrakce pouze jedné citační věty pro

každou bibliografickou citaci a bylo tedy nutné ji upravit, aby získávala všechny citační věty. Extrakce citačních vět a jejich přiřazení ke správným bibliografickým citacím byla nejméně spolehlivá, pokud byla použita možnost odkazování na citace z textu 2.3.3 pomocí formy uvádění prvního prvku a data 2.3.3. U tohoto způsobu odkazování docházelo hlavně k chybám, pokud odkaz na citaci obsahoval více autorů.

Aby mohl být určen vztah mezi jednotlivými citacemi pro potřeby metriky CPA 2.5.3, tak byla implementována nová funkčnost, která k jednotlivým citačním větám přiřazuje, ve které sekci a odstavci se nachází.

5.1.2 Ukládání článku do indexu

Ze začátku při práci s elastic search, jsem vytvořil návrh mapování článku, i když to není nutné, protože elastic search umožňuje dynamické mapování. Důležité z hlediska nastavení bylo u atributů typu řetězec nastavit, aby se indexovali dvěma způsoby. Klasickým analyzátorem a také v neanalyzované podobě. Bylo to z toho důvodu, aby mohlo být na řetězce aplikované fulltextové vyhledávání a zároveň mohl být daný atribut použit pro agregaci dat. Použité mapování je k nahlédnutí v příloze A.

5.2 Webová aplikace

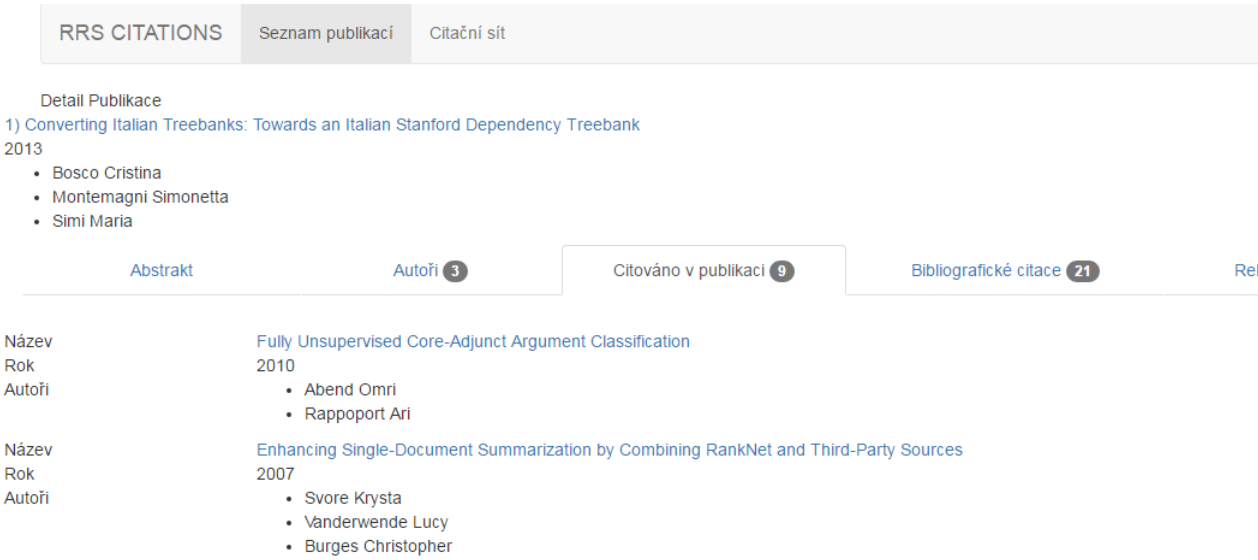
Webová aplikace se skládá ze dvou komponent. Servrové částí, která zpracovává dotazy, komunikuje s elastic search a obsahuje algoritmy navržené podle metrik pro vyhledávání podobných publikací.

Klient komunikuje se servrem pouze pomocí REST API a o existenci elastic search vůbec neví, což je vhodné z hlediska bezpečnosti.

Původně se nabízela možnost vynechat servrovou část a navrhnout systém tak, aby klient přímo komunikoval s elastic search, jelikož pro AngularJS existují knihovny, které by práci s elastic search usnadnili. Problém tohoto návrhu spočíval v tom, že některé použité algoritmy pro vyhledávání podobných článků jsou náročné a nebylo by vhodné je implementoval v JavaScriptu. Také zde vznikalo bezpečnostní riziko, protože díky tomuto přímému přístupu by uživatel mohl provádět neoprávněné operace jako například mazaní dat.

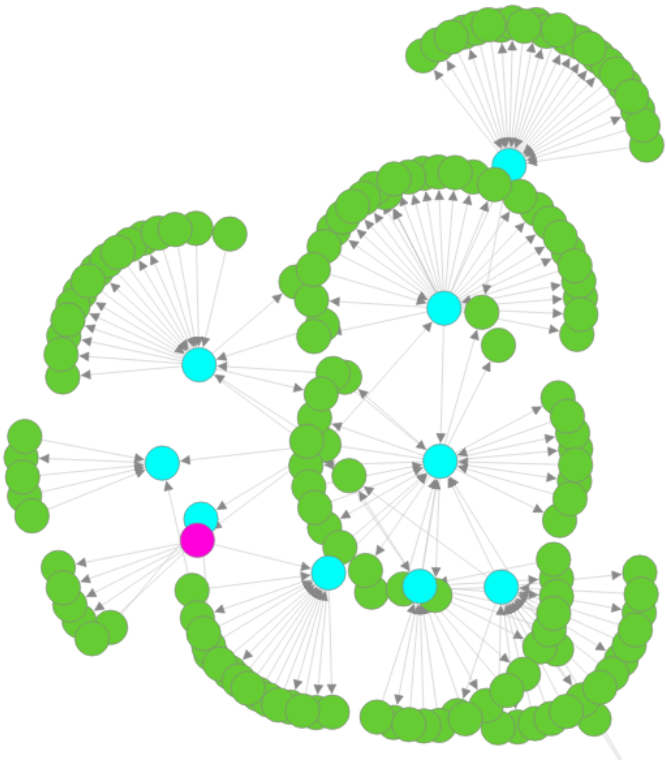
5.2.1 Veřejná část aplikace

Pro vytvoření webového uživatelského rozhraní jsem se inspiroval již existujícími řešeními2.4, které poskytují podobné služby. Aplikace se skládá ze dvou hlavních stránek. Stránka se seznamem všech naindexovaných dat s možností vyhledávání a detail publikace, který obsahuje všechny důležité data o článku.



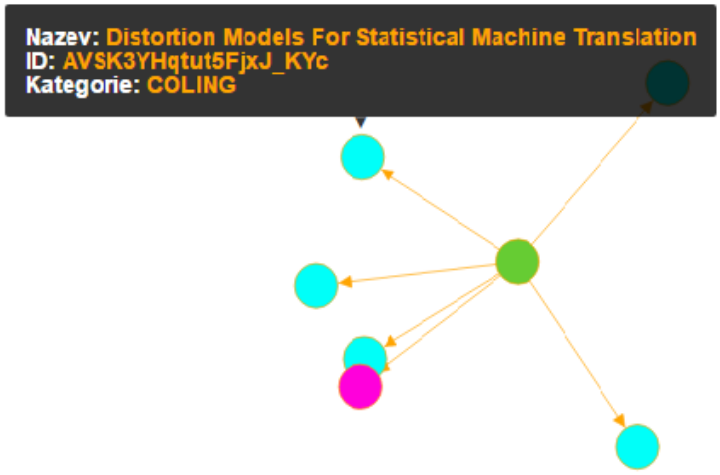
Obrázek 5.2: Stránka obsahující detail článku

Vyhledávání podobných publikací je k dispozici v záložce Relevantní publikace. Uživatel má možnost vybrat, která metrika 2.5 se využije pro vyhledání podobných článků. Výsledek je možné zobrazit v tabulce a nebo na samostané stránce v podobě interaktivní citační sítě, kde jsou zobrazeny vztahy mezi články.



Obrázek 5.3: Zobrazení citační sítě pro publikaci pří využití kocitační analýzy

V citační síti jsou nalezené relevantní publikace barevně odlišeny stejně tak jako článek pro který se to vyhledává. Po kliknutí na některý uzel grafu se zobrazí pouze ty články, které jsou s tímto uzlem propojené.



Obrázek 5.4: Detail citační sítě

Kapitola 6

Vyhodnocení systému

V této kapitole se zaměřím na vyhodnocení systému a jeho komponent. Provedl jsem několik druhu testů od úspěšnosti extrakce dat pro citační analýzu až po náročnost zpracování.

6.1 Paměťová a procesorová náročnost

Pro měření náročnosti zpracování byl využit multiprocessing skládající se z 24 procesorů typu Intel(R) Xeon(R) CPU E5-2630 v2 @ 2.60GHz. V tabulce 6.3 je zobrazna náročnost zpracování pro jeden proces, který extrahuje data pro citační analýzu ze článku pomocí modulu popsaného v kapitole 5.1.

Pruměrná pamětová náročnost[Mb]	115
Průměrné vytížení procesoru[%]	98.4

Tabulka 6.1: Náročnost zpracování extrakce dat pro citační analýzu

6.2 Rychlost extrakce dat

V této sekci se zabývám srovnáním originální verze RRS knihovny a upravené verze u které došlo k několika optimalizacím, ale zároveň se přidala nová funkčnost, která navíc zatěžuje zpracování. Rychlost byla měřena při extrakci dat pro citační analýzu ze všech článků testovací datové sady 3.

Počet zpracovaných článků	20989

	Celkový čas[h]	Celkový čas[h] s využitím paralelního zpracování s 10 procesy	Průměrný čas[s]
Originální verze RRS knihovny	458,26	45,83	78,6
Upravená verze RRS knihovny	280,44	28	48,1

Tabulka 6.2: Porovnání rychlosti zpracování extakce dat

‘

Pruměrná pamětová náročnost[Mb]	115
Průměrné vytížení procesoru[%]	98.4

Tabulka 6.3: Náročnost zpracování extrakce dat pro citační analýzu

6.3 Zpřesnění extrakce bibliografických citací

Při testování zlepšení extrakce bibliografických citací jsem ručně náhodně procházel indexované články a analyzoval jsem chyby, ke kterým nejčastěji dochází. Porovnání opět proběhlo jako srovnání originální a upravené RRS knihovny.

Počet kontrolovaných článků	350
Celkový počet manuálně nalezených bibliografických citací	3880

	Celkový počet nalezených bibliografických citací	Počet správně extrahovaných bibliografických citací
Originální verze RRS knihovny	1536	956
Upravená verze RRS knihovny	2388	1924

Tabulka 6.4: Porovnání zpřesnění extrakce bibliografických citací

Popis chyby	Výskyt chyb v originální verzi RRS knihovny [%]	Výskyt chyb v upravené verzi RRS knihovny [%]
Část bibliografické citace chybí	46,7	9
Sloučení dvou bib. citací dohromady	15	61
Bib. citace je rozdělena na několik částí	9,1	3,9
Poslední bib. citace chybí	5,4	4,5
Bib. citace sloučena s číslem stránky	1,1	1.5
Problém převodu PDF do textové podoby	17,17	16,2

Tabulka 6.5: Přehled chyb ke kterých dochází při extrakci bibliografických citací

6.4 Experimenty na datech ACl anthology

Ve webové aplikaci jsem vytvořil stránku, pomocí které lze zobrazit citační síť pro jednotlivé konference 3.1, které jsou barevně od sebe odlišeny. Samozřejmostí je filtr, protože některé konference obsahují tísíce článků, pro výběr počtu článků, ze kterých se citační síť vytvoří. Tento přehled může sloužit ke zkoumání citační sítě a její porovnávání. Například je možné zjistit,jestli články, které patří do stejné konference citují a jsou citovány převážně články z této skupiny.

Kapitola 7

Závěr

Zadání této bakalářské práce bylo analyzovat problematiku citační analýzy. Navrhnout a implementovat systém, který bude schopný vytvořit citační síť, na základě zpracování velkého množství lokálně uložených elektronických dokumentů a pomocí metrik založených na citační analýze vyhledávat podobné publikace.

Výsledný systém umožňuje paralelní zpracování vstupních dat, které dosahuje poměrně slušných výsledků. Uživatel může využít několik metrik pro vyhledání relevantních publikací a systém výsledek zobrazuje v přehledné a interaktívní podobě. Celková úspěšnost extrakce bibliografických citací nesplnila očekávání a je zde ještě značný prostor pro zlepšení.

Aplikaci je v budoucnu možné rozšířit o další metriky citační analýzy. Webová aplikace by mohla poskytovat či odkazovat na stažení publikace, Z uživatelského hlediska by bylo vhodné vytvořit pokročilejší filtr pro seznam publikací.

Literatura

[1] Bani-Ahmad, S.; Cakmak, A.; Özsoyoglu, G.; aj.: Evaluating Publication Similarity Measures. *IEEE Data Eng. Bull.*, ročník 28, č. 4, 2005: s. 21–28.

[2] Beel, J.; Gipp, B.: Academic Search Engine Spam and Google Scholar's Resilience Against it. *Journal of Electronic Publishing*, December 2010.

[3] Gipp, B.: *Citation-based plagiarism detection: detecting disguised and cross-language plagiarism using citation pattern analysis*. Springer, 2014.

[4] Gipp, B.; Beel, J.: Citation Proximity Analysis (CPA)-A new approach for identifying related work based on Co-Citation Analysis. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, ročník 2, Rio de Janeiro (Brazil): International Society for Scientometrics and Informetrics, 2009, s. 571–575.

[5] Kessler, M. M.: Bibliographic coupling between scientific papers. *American Documentation*, ročník 14, č. 1, 196301: s. 10–25, ISSN 0096946X, doi:10.1002/asi.5090140103.

[6] Kábrt, J.: *Slovník teorie a metodiky bibliografie*. Národní knihovna, druhé vydání, 1990, ISBN 80-7050-064-6.

[7] Lawrence, S.; Giles, C. L.; Bollacker, K.: Digital libraries and autonomous citation indexing. *IEEE COMPUTER*, ročník 32, č. 6, 1999: s. 67–71.

[8] Leslie, F.: *Win Friends and Influence Faculty: Methods for Citation Analysis*. 2011, [Online; navštíveno 25.4.2016].
URL http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1236&context=charleston

[9] Macroberts, M.; Macroberts, B.: Problems of Citation Analysis. *Journal of the American Society for Information Science (1986-1998)*, ročník 40, č. 5, 1989: str. 342, ISSN 00028231.

[10] Merton, R. K.: The Matthew effect in science. *Science*, ročník 159, č. 3810, 1968: s. 56–63.

[11] Petr, B.: *Bibliografické citace dokumentů podle ČSN ISO 690 a ČSN ISO 690-2: Část 1 – Citace: metodika a obecná pravidla [online]*. 2004-11-11, [Online; navštíveno 2.5.2016].
URL http://www.boldis.cz/citace/citace1.pdf

[12] Serenko, A.; Dumay, J.: Citation classics published in Knowledge Management journals. Part II: studying research trends and discovering the Google Scholar Effect. *Journal of Knowledge Management*, ročník 19, č. 6, 2015: s. 1335–1355.

Přílohy

Seznam příloh

Příloha A

Mapování článku v elastic search

```
1
2
3 stringTypeSetting = {
4         "type" : "string",
5         "analyzer": "english",
6         "fields": {
7                 "raw" : {
8                 "type": "string",
9                 "index": "not_analyzed"
10                },
11                "std": {
12                        "type": "string",
13                        "analyzer": "standard"
14                }
15        }
16 }
17
18 "title" : stringTypeSetting,
19 "id" : {"type" : "string"},
20 "time_processed" : {"type" : "string"},
21 "original_file_path" : {"type" : "string"},
22 "original_file_name" : {"type" : "string"},
23 "authors" : {
24        "type" : "nested",
25        "properties" : {
26                "full_name":stringTypeSetting,
27                "first_name":stringTypeSetting,
28                "last_name":stringTypeSetting,
29        }
30 },
31 "year" : {"type" : "short"},
32 "pages" : {"type" : "short"},
33 "text" : {"type" : "string"},
```

```
"abstract" : {"type" : "string"},
"references" : {
        "type" : "nested",
        "properties" : {
                "title" : stringTypeSetting,
                "content" : {"type" : "string"},
                "year" : {"type" : "short"},
                "referenced_pub_id" : {"type" : "string"},
                "referenced_pub_relevance" : {"type" : "float"
                   },
                "referenced_pub_script_version" : {"type" : "
                   string"},
                "authors" : {
                        "type" : "nested",
                        "properties" : {
                                "full_name":stringTypeSetting,
                                "first_name":stringTypeSetting
                                   ,
                                "last_name":stringTypeSetting,
                        }
                },
                "citation_contexts" : {
                        "type" : "nested",
                        "properties" : {
                                "context" : stringTypeSetting,
                                "from_position" : {"type" : "
                                   short"},
                                "to_position" : {"type" : "
                                   short"},
                                "section" : stringTypeSetting,
                        }
                }
        }
}
```

Příloha B

Obsah CD

Přiložené DVD obsahuje všechny dokumenty a soubory tykající se práce. Adresářová struktura DVD je následující:

- skripty
- web
- rrs_citations_xholik09.pdf
- plakat.png
- README.txt
- LICENSE.txt