Univerzita Hradec Králové
Fakulta informatiky a managementu
Katedra informatiky a kvantitativních metod

Matematické metody pro získávání a zpracování informací

Diplomová práce

Autor: Šárka Křížková
Studijní obor: Informační management

Vedoucí práce: doc. RNDr. Tatiana Gavalcová, CSc.

Hradec Králové duben 2015

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně a s použitím uvedené literatury.

V Hradci Králové dne 10. 4. 2015 _____

Šárka Křížková

Poděkování

Touto cestou bych chtěla poděkovat vedoucí diplomové práce doc. RNDr. Tatianě Gavalcové, CSc. za ochotu a čas, který věnovala mé práci, a rovněž za cenné rady a odborné připomínky při její tvorbě.

Anotace

Diplomová práce podává vhled do problematiky získávání a zpracování informací - Information Retrieval (IR). Tímto termínem se označují způsoby a postupy vyhledávání a následné zpracování informací uložených ve strukturovaných databázích; jejich zpracování má být relevantní pro uživatele a odpovídat jeho potřebám, které uživatel vyjádří formou žádosti nebo dotazu. Diplomová práce vysvětluje základní principy a metody používané ve vyhledávání a zpracování informací, uvádí strategie v těchto činnostech, popisuje modely používané v IR a jejich podstatu, rozdělení jednotlivých úloh v modelu nebo metod, které IR systémy využívají.

Diplomová práce je členěna na tři části. První část uvádí vývoj disciplíny IR a poskytuje její teoretický přehled. Uvádí klasifikaci modelů používaných v IR a charakterizuje základní principy získávání informací a následné zpracování pomocí metod specifických pro tyto modely. Ve druhé části jsou uvedeny jejich aplikace a praktická část diplomové práce je zaměřena na matematické modely na bázi některých algebraických struktur. Třetí část obsahuje strategie vyplývající z teorie modelů v IR, aplikované na softwary, jejichž účelem je detekce plagiátů; jsou v ní uvedeny programy Big Brother, EVE2, TurnItIN, Jplag, Moss, MyDrobBox, SHERLOCK, Theses a Odevzdej.

Annotation

Title: Mathematical methods in Information Retrieval

This Diploma Thesis provides the insight into Information Retrieval (IR). The term Information Retrieval denotes ways and methods of a search, finding and returning of information stored in structured databases; these databases should be evaluated and the relevant ones returned to the user, depending on the user's needs as expressed in his/her request or query. The Diploma Thesis provides an overview of the basic principles and methods applied in IR, working strategies, and descriptions of models and models' roles used in IR.

The Diploma Thesis consists of three chapters. The basic methods and the elementary principles of IR are defined in the first chapter as a theoretical review, together with notes on the development of IR. A classification of models used in IR is given, and methods and principles for the search and return of information specific for those models are characterized. The second chapter studies the practical applications of IR, focused on mathematical methods based on some special algebraic structures. The model theory used in IR is applied in the third part of the Diploma Thesis. As examples of using mathematical methods in IR for the purpose of the detection plagiarism, softwares as Big Brother, EVE2, TurnItIN, Jplag, Moss, MyDrobBox, SHERLOCK, Theses and Odevzdej are listed.

Obsah

Úvod

Diplomová práce se zabývá problematikou získávání a zpracování informací - Information Retrieval (IR). Termínem Information Retrieval se označují způsoby a postupy vyhledávání a následného zpracování informací uložených ve strukturovaných databázích; zpracování má být relevantní pro uživatele a odpovídat jeho potřebám, které jsou vyjádřeny formou žádosti nebo dotazu uživatele. V rámci diplomové práce je detailně prostudován Information Retrieval jako koncept a soubor nástrojů pro zpracování a získávání informací včetně různých přístupů a jsou navrženy základní metody, které jsou využívány napříč technologiemi. Teoretický rámec diplomové práce poskytuje základní konceptuální model včetně charakteristik vyhledávané a zpracovávané informace, zpracovávaných úkolů a úloh, hypotéz a znalostí. Na obecnou teorii navazují základní modely IR systémů. V další části se diplomová práce věnuje systémům používajícím specifické matematické metody. Byly vybrány tři hlavní modely, které využívají matematické metody, a to pravděpodobnostní model, Booleovský model a algebraický model. Aplikace výše zmíněných metod navazuje na teoretická východiska, a to v použití IR při odhalování plagiátorství. Diplomová práce uvádí konkrétní příklady modelů a způsoby jejich využití a získání výsledku v detekci plagiátorství. V práci jsou uvedeny nejznámější antiplagiátorské softwary a jejich porovnání.

Diplomová práce je členěna na tři části.

V první části je popsána teorie Information Retrieval včetně jejího časového, koncepčního a částečně i technologického vývoje. Jsou popsány základní koncepty včetně jejich hlavního zástupce, a to konceptuálního modelu. Tato část obsahuje popis všech důležitých složek postupu zpracování a získávání informací (komplexnost úlohy, charakteristiky úkolů, typy informací, typy informačních zdrojů, systematizace znalostí, mapování reality). Je v ní detailně rozebráno, co přesně IR znamená, a jsou zde uvedeny dvě hlavní taxonomie IR systémů (vertikální, horizontální).

Druhá část je zaměřena na vybrané matematické modely používané v IR a na jejich aplikaci. Jedná se o pravděpodobnostní model, Booleovský model a algebraický model založený na použití algebraických struktur. Aplikace modelů je uvedena a popisována

na těchto systémech a službách: Westlaw (Booleovský model), Mooerův model, FaIR model, BR-Explorer systém, Rajapakse-Denham systém a FooCA systém (vše algebraické modely). U algebraických modelů v diplomové práci jedná především o algebraické struktury – svazy, a jsou v ní vyčleněny speciální třídy svazů, které se hodně využívají ve zmíněných systémech.

Třetí část se věnuje aplikaci výše zmíněných metod na detekci plagiátů včetně uvedení různé klasifikace přístupů. Diplomová práce se zabývá programy Big Brother, EVE2, TurnItIN, Jplag, Moss, MyDrobBox, SHERLOCK, Theses a Odevzdej. Diplomová práce obsahuje také příklady detekovaného textu, který by mohl být plagiát. V této části jsou srovnány klady a zápory jednotlivých softwarů.

V závěru práce jsou vyhodnoceny výsledky a shrnuté základní poznatky. Studovaná problematika poskytla vhled do charakteru i komplexity úkolů, kterými se zabývá Information Retrieval a také do strategií, metod, používaných modelů jako nástrojů na jejich naplnění. Aplikace potvrzují, že jde o disciplínu vyžadující souhru konceptů a postupů jak z oblasti práce s informacemi, tedy v aktuálních a specifických oborech informatiky, dále také z oblasti matematických struktur, ale i z dalších rozmanitých oborů.

I. Teoretická část

1 Information retrieval (IR) jako koncept a soubor nástrojů

Žijeme v době, která nás doslova zahlcuje informacemi. Proto je dnes nutností umět se v tomto přesycení informacemi orientovat a umět za velmi krátký časový úsek vyhledat relevantní data. Tyto informace jsou uložené na různých servrech a knihovnách po celém světě. Proto je tady pro většinu uživatelů GoogleSearch, který vyhodnotí více než bilion webových stránek a vrátí výsledek na dotaz za méně než půl vteřiny. Google zpracuje více než stovky milionů dotazů každý den, využívá stovky tisíc počítačů, sofistikovanou počítačovou vědu a nejmodernější vyhledávací a vyhodnocovací algoritmus, který je přesně definovaný a záměrně vytvořený pro řešení vyhledávacích problémů. Základ těchto algoritmů je v dnešní době stále založen na information retrieval modelu (model pro získávání a zpracování informací). Důvod pro neustálé využívání IR modelu, i když je znám již dlouho, je ten, že jsou webové stránky a dotaz uživatele indexovány. Požadavek (dotaz) je tak snadno spárován ke konkrétní webové stránce (Teufel, 2012). Zájem o IR modely se mnohonásobně zvýšil za poslední desetiletí a to nejen z řad vědců, ale také z řad firem.

1.1 Vývoj získávání a zpracování informací (information retrieval)

Model pro vyhledání informací byl vymyšlen kolem roku 1950, což je dříve než objev internetu. Oproti internetu se nedočkal tak masivního rozvoje hned ve svých počátcích. Teprve během posledních 20 let se rozvinul zájem o internetový vyhledávací nástroj a vyhledávače. Práce s nimi se brzy staly nezbytnou součástí práce s internetem. Se zvýšením úložného prostoru počítačů a zvýšením rychlosti procesorů se zvýšila kapacita a rychlost vyhledávání (zvýšil se počet prohledaných stránek a snížila se doba, za kterou výsledek uživatel obdrží). Vývoj v možnostech počítačového systému odráží rychlý přechod od získávání, indexování a vyhledávání informací založeném na manuálním vyhledávání k modelům založených na plně automatických metodách. Mezi hlavní důvody skokového přechodu patří neschopnost katalogizační techniky zvládat neustále se zvětšující objem informací. Podle Moorova zákona, který hovoří o počtu tranzistorů zapojených v procesoru a nárůstu jejich počtu (zhruba každých 18 měsíců se zdvojnásobí počet zapojených tranzistorů),

dochází k neustálému zdvojování kapacity digitálního úložiště, a to každé dva roky (Sanderson a Croft, 2012). Pro ilustraci počet bitů nesoucích informace na 6,5cm^2 (1čtverečný palec) povrchu disku vzrostl z 2000 bitů v roce 1956 na 100 bilionů bitů v roce 2005 (Walter, 2005). S růstem počtu digitálních nestrukturovaných informací a s rozvojem vysokorychlostních sítí (internetu) najednou mnoho lidí získalo přístup k nepředstavitelně velkému množství informací. A jediné možné řešení pro vyhledání relevantní položky z ohromné světové databáze je dodnes používaný IR systém.

1.2 Vývoj IR systému v rámci mechanických a elektromechanických zařízení

Obvyklý přístup k velkým souborům informací vycházel ze zkušeností v knihovnách, kde se položky (knihy nebo články) popisovaly pomocí platných standardů (autor, název dokumentu, rok a místo vydání, název nakladatelství, edice, pořadí vydání, fyzický popis a číslo ISBN). V knihovnách se využíval katalogizační systém k vyhledání určité knihy pomocí zadaných hesel, např. autor. Na urychlení vyhledávání takto uskladněných položek se výzkum zaměřil ke konci 19. století. V roce 1918 patentoval Soper zařízení s katalogizačními kartičkami (Soper, 1918). Mezi roky 1920-1930 byla patentována Goldbergem série zařízení, která vyhledávala dle vzoru děr nebo písmen skrz celý katalog položek (Goldberg, 1931). Další mechanická technologie byla založena na principu děrovaných štítků, světla a fotobuňky. Tento systém byl schopen prohledat 600 štítků za minutu. (Luhn, 1956). V tomto období byl poprvé použit výraz information retrieval (IR) a to na konferenci v roce 1950. Jeden z příspěvků byl: „Problém, který diskutujeme, je vyhledávání z úložiště dle zadaných kritérií pomocí přístroje (information retrieval). “ (Mooers, 1960).

1.3 Vývoj IR systému s využitím počítačů

Již v roce 1948 Holmstrom popsal zařízení nazvané Univac, které je schopno vyhledávat reference (odkazy) na text související s vyhledávaným tématem. Kód a text byly uloženy na magnetickém pásku (Holmstrom, 1984). Mitchell přišel s modelem využívajícím zařízení Univac, které bylo schopno vyhledat 1 000 000 indexovaných záznamů dle šesti předmětových kódů. Předpokládaný čas na vyhledání těchto záznamů byl 15 hodin (Gupta, Saini a Saxena, 2013). Výzkum IR se rychle rozvíjel. Do popředí se dostaly otázky týkající se indexování dokumentů a jejich následného získávání (vyhledání). Klasický přístup indexování využíval

hierarchické klasifikační schéma dle předmětu, např. Deweyův číselný systém. Tento logický systém využívá číselný kód k určení obsahu dokumentů při katalogizaci. Dewey rozdělil celý seznam literatury do deseti hlavních oborů (religionistika, sociální vědy, jazyky, přírodní vědy atd.), obory do deseti podtříd a každou podtřídu do dalších deseti částí. Pro ilustraci číslo 421.3.7 měla sedmá kniha filologie anglického jazyka ve třetí posloupnosti. Jeho systém třídění využívá přes 135 zemí světa, převážně anglosaských (Dopitová, 2004). Jako alternativní přístup byl navržen Uniterm systém. Základní indexování bylo v Unitermu navrženo na systému seznamu s klíčovými slovy (Taube, Gull a Wachtel, 1952). Tento přístup se později projevil jako důležitý pro další rozvoj vyhledávacích metod. Později bylo provedeno důkladné porovnání metody využívající počítač Uniterms a klasické metody založené na číselném kódu. Závěr srovnání byl, že Uniterms je srovnatelný, ne-li lepší než klasické metody (Cleverdon, 1970). Styl používáný při vyhledávání na elektromechanických a počítačově založených zařízeních se začal nazývat „Boolean retrieval“. Dotaz položený uživatelem byl převeden na logickou kombinaci termínů, které zabezpečovaly přesnou shodu vyhledané položky s dotazem. Kolem roku 1950 bylo pevně stanoveno, že počítače se budou výhradně používat jako nástroj pro vyhledávání. A tak následoval obrovský růst výzkumu v komerční sféře a IR proces se stal důležitou oblastí zkoumání.

V šedesátých letech 20. století se na Harvardské a Cornellově univerzitě formovaly výzkumné skupiny, které vytvořily mnoho konceptů využívaných dodnes. Jedna z mnoha oblastí zájmu těchto skupin byla formalizace algoritmů s cílem vyhodnotit dokumenty, které se vztahují k dotazům položených uživateli během vyhledávání.

Jeden z mnoha přístupů, se kterými tyto skupiny pracovaly, využívá nástroje specifické algebry - vektorových prostorů k formalizaci dokumentu. Dokumenty a dotazy jsou viděny jako N-dimenzionální vektory. Vektory mají dimenzi N, kde N je počet jednotlivých výrazů v dokumentu (Sanderson a Croft, 2012).

Další z mnoha věcí objevených v šedesátých letech bylo uvědomění si závažnosti zpětné vazby. Tento postup znamenal opakování ve vyhledávání, kde dokumenty získané během dřívějšího vyhledávání byly označeny jako relevantní k dotazu v IR systému. Modifikované verze založené na zpětné vazbě se využívají v moderních vyhledávačích, jako jsou „související články“ odkazy na Google Scholar. Konkrétně v roce 1959 existoval pouze obecný, ale velmi intenzivní zájem o podporu, vyčištění a

zdokonalování dřívějšího postupu pro skladování a získávání informací. Úsilí bylo založeno na předpokladu, že lze využít metodu pro manuální organizaci složek a dokumentů a převést tento proces na automatické metody indexování a získávání informací (Harmon, 2008). Zároveň nebylo vzato v úvahu, že nové technologie nejsou pouhé urychlení či vylepšení starých, ale jedná se o úplně nové, dříve neznámé technologie. Proto také Bill Goffman se svým týmem přišel s otázkou, zda je dobré spoléhat se na Booleovskou logiku. Dospěli totiž k závěru, že Booleovské operátory byly pro manuální a automatické získávání dat v podstatě neoptimální a neefektivní. Například sjednocením subjektů „a“ a „b“ riskujeme nalezení mnoha nepodstatných informací. Mezitím při průniku „a“ a „b“ riskujeme vynechání mnoha žádaných informací (Verhoeff, Goffman a Belzer, 1961). Goffmanův tým přišel s myšlenkou, že relevantnost dokumentů je měřítko informací v jednotlivém dokumentu ve vztahu k určitému dotazu. Goffman uvedl, že při využití relevantnosti jako měřítka musí být jasně definována relevantnost ve vztahu k úplné sadě dokumentů a ne k jednotlivému dokumentu. Goffman byl dlouho přehlížen navzdory perspektivním navrhovaným změnám.

Klíčový rozvoj nastal kolem roku 1970, kdy Luhnova práce o váze frekvence termínů (založena na počtu výskytu daného slova v rámci dokumentu) byla dokončena Jonesovou prací o výskytu slov v rámci dokumentu a kolekce. V článku je uvedena myšlenka, že frekvence výskytu slova v kolekci dokumentů je nepřímo úměrná k významnosti ve vyhledávání. Méně běžná slova předávají více specifický obsah, který je důležitější při vyhledávání (Jones, 2004).

Jedna z firem, která si jako první vyhradila právo nabízet nástroj pro vyhledávání, byla firma Dialog v roce 1966. Ta vznikla při výzkumu IR systému pro NASA (Bjorner a Ardito, 2003). Pozoruhodným faktem v tehdejší době byla nízká úroveň komunikace mezi komerční oblastí a sférou výzkumu. Navzdory opakované demonstraci vědců, že využití algoritmu pro vyhledávání založeném na pořadí je nejlepší, v komerční oblasti se stále využívala metoda založená na Booleovském vyhledávání. Teprve kolem roku 1990 se situace změnila a to díky novému systému WESTLAW´s WIN a růstu počtu webových vyhledávačů. V této době se vědci začali zabývat analýzou záznamu požadavků. Z nichž odvodili, že různí uživatelé s různě dostupnými informacemi použijí někdy shodného požadavku při získání informace (Verhoeff, Goffman a Belzer, 1961). Systém založený na IR by měl být schopen obsloužit rozdílné potřeby díky

nalezením „jinak příslušných“ dokumentů. Mnoho dalších vyhledávacích modelů, metod a procesů bylo vyvinuto po roce 1990. Konkrétně se jednalo o úvod k metodě založené na pravděpodobnosti s pomocí využití jazykových modelů. Díky novému pohledu na zpracování a vyhledání shody mezi dokumentem a požadavkem prokázal přístup založený na jazykovém modelu nové porozumění pro velký rozsah IT zpracování jako například relevantnost odpovědi, tvoření shluku dokumentů a pojem nezávislost. Mnoho výzkumů IR algoritmu se zaměřilo na krátké dotazy, které jsou tvořeny pouze lingvistickou strukturou (typicky se jedná o jednoslovné výrazy). Několik výzkumů se zaměřilo na delší uživatelské dotazy, které jsou více přirozené. Mnoho nápadů a inovací se zrodilo na Text REtrieval Conference (TREC, konference o vyhledání textu). Na jednom z mnoha workshopů se pracovníci zabývali tím, jak snadno najít jednoduchou odpověď v textu při limitovaném počtu otázek (jako například „wh“ otázky v anglickém jazyce „who“ a „when“). Tento nápad přivedl vědce k hlubšímu prozkoumání této otázky. Vědci vynalezli techniku, která poskytuje více konkrétních odpovědí na detailní otázky. Úspěch aplikací jako Siri (Apple), Watson (IBM) a Answer (Yahoo) je založen právě na tomto postupu.

Od roku 1993 se webové vyhledávání plně rozvinulo se vzrůstem počtu webových stránek. Dříve známé nápady byly využity a implementovány v komerčním sektoru. Vývojáři algoritmů si brzy uvědomili, že mohou využít spojení mezi webovými stránkami a uživateli. Mohou tak vytvořit roboty, které sesbírají většinu požadovaných webových stránek na Internetu. Tento přístup ale nezajistil, aby vybrané kolekce obsahovaly pouze spolehlivý materiál. Bezohlední vývojáři a prodejci brzy zjistili, že pomocí manipulace obsahu stránky mohou změnit pořadí stránky ve vyhledávačích. Metody, které by vylučovaly takovou manipulaci a zároveň identifikovaly nejlepší webové stránky, se staly velmi žádoucí. Dva důležité body k dosažení těchto metod bylo propojení analýzy a vyhledávání ukotveného textu, což znamenalo vyhledat obsah webové stránky a text spojení. Ukotvený text (obvykle krátký souhrn stránky) byl brzy rozpoznán jako důležitý zdroj informací (McBryan, 1994). Využití ukotveného textu je klíčové pro Google vyhledávací algoritmus již od začátku jeho vývoje spolu s využitím významných analytických metod (PageRank vyvinutá Googlem a metoda HITS vyvinutá Kleinbergem). Propojení analýzy a reprezentace vícenásobného textu, který byl promítnut do existujícího dokumentu, znamenalo, že vnitřní algoritmus IR systému byl komplexním. Kolem roku 1990 se

také začalo používat automatické získávání informací ze záznamu vyhledávače. Přínos takto získaných informací ze záznamů vzrůstá s počtem lidí, kteří používají webové vyhledávače. Ohodnocení uživatelských dotazů, klikacích vzorů a reformulace dotazů umožnily vědcům vyvinout efektivnější techniku pro zpracování dotazů, která je založena na porozumění uživatelského záměru. Příkladem je automatická korekce chyb, automatické rozšíření dotazů a větší přesnost v očistění dat (Peng, Ahmed, Li a Lu, 2007). Jako klíčový pro rozvoj IR systému se jevil systém od autorů Carbonell a Goldstein, kteří vytvořili „maximální mezní relevantní odlišný systém“ (Sanders a Croft, 2012).

2 Koncepty pro vyhledávání a získávání informací

Dle Wilsona existuje několik konceptů pro modely vyhledávání a získávání informací. Některé z nich jsou sumarizačního typu a některé analytické. Typy modelů se rozdělují na modely informačního chování (Wilson, 1981), informačně vyhledávacího chování (Dervin a Nilan, 1986) a informačního vyhledávání a získávání (Ingswersen, 1996). Modely informačního chování jsou oproti ostatním modelům sice již zastaralé, ale stále mohou nabídnout pomoc v teoretické fázi výzkumu, naznačit vztah nebo stav, který je vhodný k otestování. Nevýhodou těchto modelů je, že nepopisují faktory chování a navrhují testovanou hypotézu velmi nepřesně. Přesto jsou tyto modely prospěšné, protože poukazují na mezery ve výzkumu v oblasti získávání informací (Wilson, 1999).

2.1 Konceptuální modely a jejich využití

Všechny práce přikládají význam modelu, který je založen na entitách a vztazích. Tyto modely se nazývají konceptuální rámce (Engelbart, 1962) nebo konceptuální modely. Ve vývoji byly neustále porovnávány modely a rámce, jejich výhody a využití pro různé oblasti výzkumu. Nastavení pro ohodnocení těchto modelů jsou důležitá. Pro rozvoj konceptuálních modelů je rozhodující popsat základní objekty a komponenty, rozpoznat a zorganizovat vztahy mezi objekty, rozpoznat jak změny objektů a jejich vztahy ovlivní funkcionalitu systému, stanovit si zavazující cíle a metody výzkumu (Engelbart, 1962). Konceptuální model je primární a širší než jakýkoliv jiný, protože nabízí konceptuální a metodologické nástroje pro formulování hypotéz a teorií. Reprezentuje také myšlenky, principy a hodnoty vědecké komunity kolem roku 1990. Vědecké teorie jsou nutné pro systematizaci znalostí, vedení výzkumu a mapování

určité části reality (Bunge 1967). Dle některých názorů jsou tyto funkce vhodné i jako funkce konceptuálních modelů. Konceptuální modely jasně mapují realitu, vedou výzkum, systematizují znalosti pomocí integrace a navrhují hypotézy. Modely poskytují pracovní strategii, schéma s hlavními koncepty a jejich vzájemné vztahy. Modely vedou výzkum pomocí sad vědeckých otázek. Konceptuální modely nemohou být přímo ohodnoceny empiricky, protože jejich formulace je postavena na bázi formulování empirického testování vědeckých otázek a hypotéz. V modelech jsou pouze instrumentální (nástroj) a heuristické hodnoty. Obvykle jsou hodnoty získány při ohodnocení vědecké strategie a jejích výsledků (Vakkari, 1998). Pro měření jsou důležité metriky (způsoby a nástroje měření) konceptuálních modelů, které stojí na základě obecných vědeckých principů (hypotézy by měly být studovány ve všech situacích a extrémních podmínkách a rámec by měl být smysluplně omezen). Pokud jsou použity tyto principy pro metriky konceptuálních modelů, pro hodnocení modelu mohou být použita níže uvedená kritéria (Byström a Järvelin, 1995):

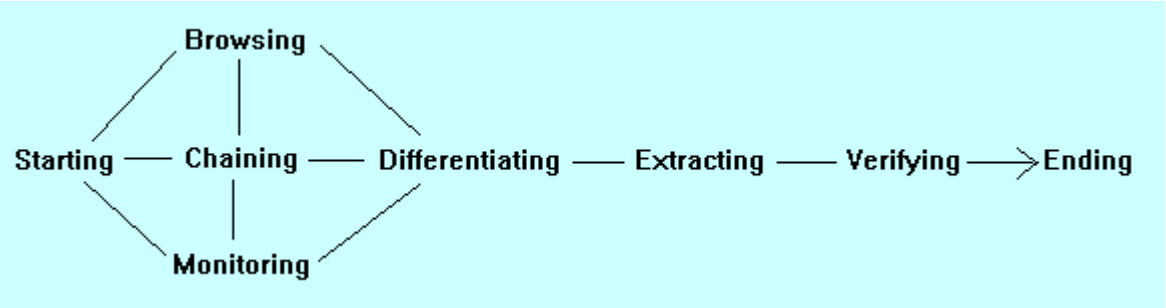
- jednoduchost (simplicity) – jednodušší systém je lepší, pokud jsou systémy jinak stejné
- přesnost (accuracy) - přesnost a jednoznačnost jsou bezpodmínečně nutné pro omezení konceptuálního modelu
- rozsah (scope) – širší rozsah je lepší, protože nezahrnuje omezení
- systematická síla (systematic power) – schopnost organizovat koncept, vztahy a data systematickou cestou
- vysvětlující síla (explanatory power) – schopnost objasnit a předpovídat jevy
- spolehlivost (reliability) – schopnost poskytovat správnou reprezentaci v celém rozsahu možných situací
- správnost (validity) – schopnost poskytovat pravdivou reprezentaci a zjištění
- plodnost (fruitfulness) – schopnost navrhnout problém k vyřešení a hypotézy k testování

Teoretický rozvoj nebo konstrukce nového konceptuálního modelu obvykle zahrnuje rozvoj v konceptu a terminologii. Důležité je, aby vyjádřené návrhy dodržely základní požadavky na přesnost, jednoznačnost, jednoduchost, obecnost a vhodnost. Dobrý koncept reprezentuje základní vlastnosti objektů, vztahů a událostí v prohledávané oblasti. Konceptuální modely by na sebe měly systematicky navazovat. Existují dva základní rámce navržené Ellisem (Ellis, Cox a Hall, 1993) a Ingwersenem (Ingwersen,

1996). Ellisovo zpracování různých typů chování zahrnuje šest základních vlastností (stádií):

- začátek (starting): uživatel začíná vyhledávat informace
- řetězení (chaining): následující poznámky pod čarou a citace materiálu svázané ze známých položek skrz citační index
- prohlížení (browsing): polovedené nebo polostrukturované vyhledávání
- rozlišování (differentiating): rozlišení známých zdrojů
- monitorování (monitoring): udržování vyhledávání k aktuálnímu datu nebo k nejnovějšímu ohodnocení
- extrahování (extracting): selektivní identifikování odpovídajícího materiálu
- ověřování (verifying): ověřování přesnosti informací
- konec (ending): konečný výsledek

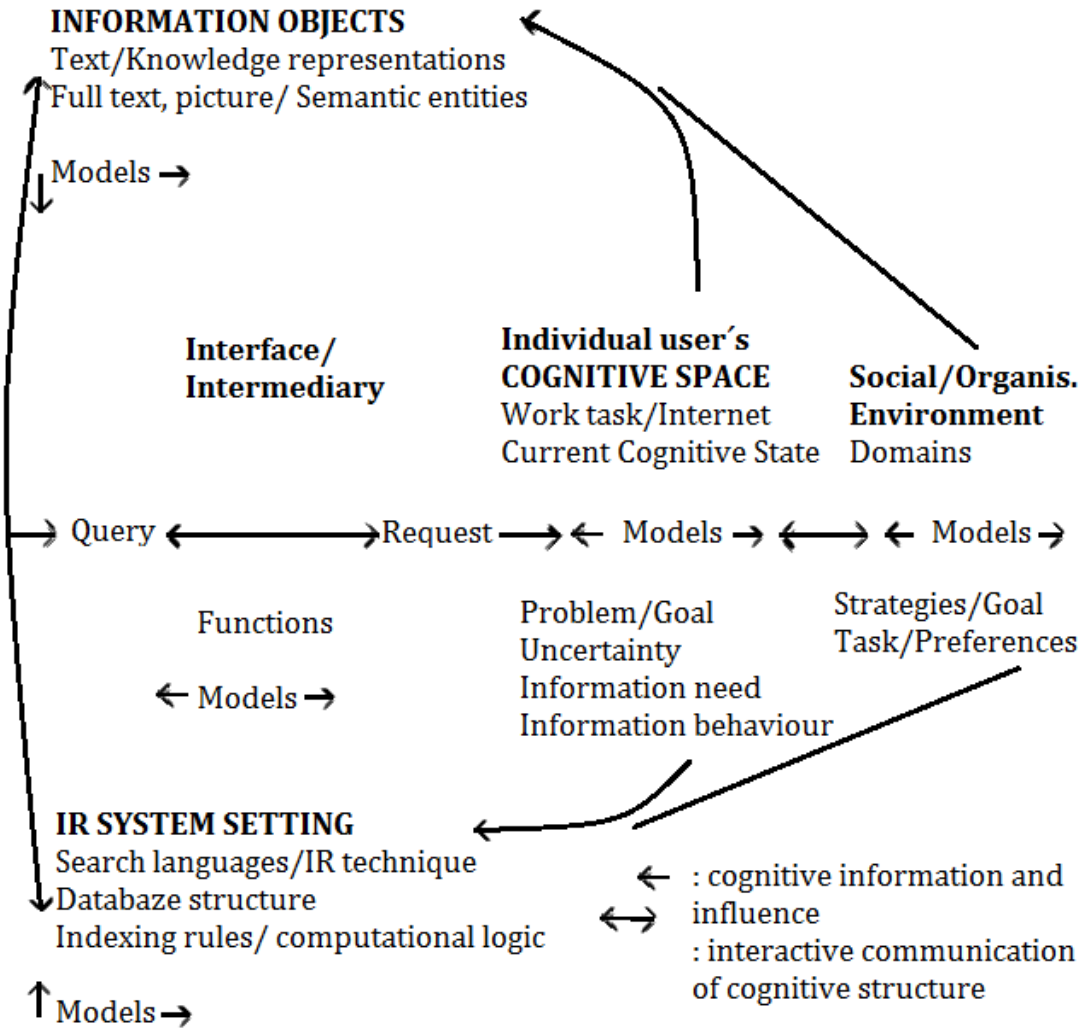
Síla Ellisova modelu spočívá v tom, že je založen na empirickém vyhledávání a byl testován několika studiemi. Co se týká funkčnosti systému, tak jakýkoliv vztah a vzájemná interakce funkcí v jakémkoliv individuálním vzorci závisí na jedinečných okolnostech vyhledávací aktivity uživatele zajímajícího se v určitý časový okamžik (Ellis, 1989). Návrh na propojení jednotlivých stádií je naznačen v obr. 1, který zobrazuje provázanost jednotlivých jevů v modelovém rámci.



Obr. 1: Fáze zpracovávání Ellisova rámce chování (A process version of Ellis´s behavioural framework) (Wilson, 1999)

Pomocí naznačeného modelu lze popsat jakoukoliv vyhledávací aktivitu pomocí Ellisových stádií. Jsou natolik obecné, že jsou vhodné pro většinu empirických situací. Selžou v ojedinělém případu, kdy se vysvětlení pracovního úkolu subjektu zabývá znalostmi úkolu. Stádia nejsou dostatečně explicitně spojená s externími příčinnými faktory. Ellisův model rozeznává stádia v různých situacích zahrnující různé typy uživatelů během úspěšného vyhledávání. Například uživatel v určité roli vyžaduje

více nebo méně monitorování než jiný uživatel. Tento přístup vede k nalezení podmínky, která způsobuje tyto odlišnosti. Na obr. 2 je zobrazen zjednodušený Ingwersenův model. Na tomto obrázku je znázorněn vztah mezi Ingwersenovým modelem a jinými modely s informačně vyhledávacím chováním. Prvky uživatelského kognitivního prostoru a sociálního/organizačního prostředí se podobají osobě v obsahu a environmentálním faktorům ve Wilsonově modelu (Wilson, 1996).



Obr. 2: Ingwersenův model IR procesu (Ingwersen´s model of the IR process) (Ingweresen, 1996) (Wilson, 1999)

Obecně orientovaný dotaz daný na IR systém je zaměřený na aktivní vyhledávání. Toto je metoda, kterou používá většina vyhledávačů. Ačkoliv je v tomto modelu explicitní počet ostatních prvků v rámci prohledávané oblasti modelu, tak funkčnost uživatele, autor dokumentu, prostředník, rozhraní a IR systém jsou výsledky

explicitního nebo implicitního kognitivního modelu dané doménou zájmu. Uživatelé tedy mají model práce-úkol s potřebnými informacemi nebo uživatelskými cíli, které jsou obvykle implicitní, ale velmi často jsou schopny dalšího vysvětlení. IR systém tedy vysvětluje navržený kognitivní model, jaké jsou možnosti systému a jakou by měl funkčnost. Ingwersen navrhl IR systém jako obraz komprese modelu informačního chování obsahující nutný detail, kdy se objekt podílí na dotazu. Také ukázal, že transformace probíhá z reálného světa, kde uživatelé definují problémy nebo cíle z určité situace, k ukazatelům objektů, které mohou uspokojivě vyhledat a identifikovat dané objekty. Ingwersen zmínil potřebu kognitivní struktury a její transformaci, aby mohla efektivně komunikovat se systémem, který zahrnuje uživatele (autor a IR systém designer). Všechny Ingwersenovy zákony jsou všeobecně platné.

Nevýhoda tohoto přístupu je ta, že neumožňuje testování a ani nenabízí aplikaci k ohodnocení IR systému (Saracevic, 1996). Nedávno však byla vytvořena na základě Ingwersenova modelu a testována hodnotící strategie, která ukázala sílu a hodnotu právě tohoto modelu pro testování interaktivních IR systémů (Borlund, 2000). Potenciální nevýhoda může spočívat v tom, že jiné chování než informační vyhledávání není výslovně analyzováno. A tak, když uživatel dospěje do bodu výzkumu a jeho kognitivní struktury jsou ovlivněny procesem získávání rozhodnutí jak a kdy se posunout ve vyhledávání vpřed, tak může být vyhledání ztraceno. V tomto modelu je několik objektů prezentujících souhru a jejich relevantní rysy jsou explicitně vyjádřeny. Souhrnný model nabízí vhled do zkoumané domény a seznam faktorů ovlivňujících funkce. Model nemůže nabídnout analýzu faktorů příčiny jevů v IR systému bez detailních komponent, a proto nelze navrhnout hypotézu k testování (Järvelin, 1987).

2.2 Komplexnost úlohy

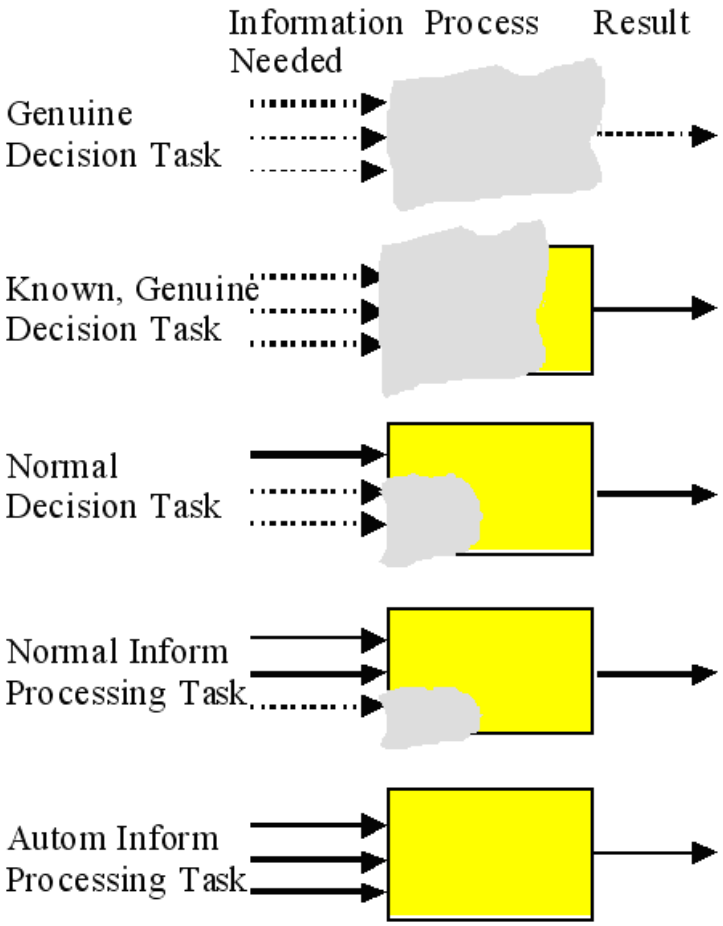
Práce se sestává z úkolů (tasks), které jsou složeny z několika úrovní postupně menších podúkolů. Úkoly jsou dané nebo jsou identifikovány pracovníkem. Každý úkol má rozpoznatelný konec a začátek, formu obsahující stimuly, podněty, cíle nebo měření, které musí být dosaženo (Hackman, 1969). Každý sebevětší úkol se rozkládá na podúkoly, které mohou být brány jako samostatné úkoly. Tato relativita v definici je důležitá pro pořadí při analyzování úkolu v odlišných úrovních komplexnosti. IR systémy se soustředí na úkoly související s informacemi, proto jsou úlohy viděny jako

subjektivní nebo objektivní úkol. Vztahy mezi úkoly vycházející z psychologie organizace, kdy je popis úkolu založen na vědomém úkolu, jsou obvykle neplatné z mnoha důvodů (Hackman, 1969). Ve vyhledávání musíme považovat úkol za vědomý, protože každý uživatel reprezentuje úkol odlišně. Vědomý úkol zahrnuje hrubý základ pro formu úkolu, potřebné informace o úkolu a výběr akcí (činností) k naplnění úkolů. Autoři navrhují mnoho charakteristik spojených s komplexností úkolu: opakovatelnost, analyzovatelnost, předem danou definovatelnost, počet alternativních cest úkolu, aktuálnost závěru, počet cílů, závislost konfliktů na cílech, nejistotu během zpracování a stanovení cílů, počet vstupů, kognitivní a dovednostní požadavky a časovou variabilitu podmínek úkolu (Cambell, 1988), (Daft, Sormunen a Parks, 1988), (Fisher, 1979) a (Zeffane a Gul, 1993). Je mnoho dalších v literatuře uváděných charakteristik, které patří do dvou hlavních skupin a to charakteristiky spojené s předem danou určitelností úkolu a charakteristiky spojené s rozměrem úkolu. Järvelin navrhuje jednoduchou jednodimenzovou kategorizaci založenou na úhlu pohledu uživatele (pracovníka), předem dané definovatelnosti nebo nejistotě o úkolu, výsledky úkolu, zpracování a požadavcích na informace (Byström a Järvelin, 1995). Dimenze se vztahuje k následujícím charakteristikám úkolu: opakování, analyzovatelnost, předem daná určitelnost, počet alternativních cest provedení úkolu a aktuálnost výstupu. Podobné modely založené na jedné dimenzi použili i dva další autoři. V jednoduchém úkolu jsou vstup, zpracování a výsledek předem ovlivněny, zatímco složitost nebo úplnost úkolů jsou neznámé (Tiamiyu, 1992). Opravdový rozhodovací úkol je ten, který nemůže být stanovený předem (Van de Ven a Ferry, 1980). Tato generalizace je obecná a proto široce aplikovatelná na mnoho typů úkolů a domén (domain).

2.2.1 Kategorizace úkolů určitelnost = determinovanost

Järveling klasifikoval pět základních kategorií zahrnujících od automatického informačního zpracování úkolu po obecně rozhodovací úkol. Rozdělení je založeno na předpokladu determinovanosti (předem daná strukturovatelnost) úkolu a je svázáno se složitostí nebo komplexností úkolu. Komplexnost úkolu závisí na stupni neurčitosti vstupu, zpracování a výstupu (Van de Ven a Ferry, 1980). Automatické zpracování úkolu, typ výsledku, zpracování a typ použitých informací byly již detailněji popsány v předchozí kapitole. V obecně rozhodovacích úkolech nemohou být úkoly, výsledky nebo informace popsány před procesem zpracování. V kategorizaci uvedené v obr. 3

dle Järvelina a Wilsona jsou informace reprezentovány šipkami a zpracování úkolu naznačeno žlutými obdélníky (Järvelin a Wilson, 2003). Determinovatelná část je naznačena plnými šipkami. Nedeterminovatelná část je naznačená přerušovanými šipkami. Stínované části obdélníků a přerušované šipky znázorňují atributy založené na konkrétním úkolu. Levá strana znázorňuje vstup třemi šipkami, protože je obvykle potřeba mnoho vstupů, které mají různou míru určitelnosti. Výsledek je potom samozřejmě ovlivněn těmito vstupy a má různou míru určitelnosti.



Obr. 3: Rozdělení úkolů do kategorií (Task categories) (Anon, 1974)

2.2.2 Charakteristika jednotlivých úkolů

Automatické zpracování úkolu je plně determinované a může být automatizováno. Příkladem je výpočet čistého platu pracovníka, kde výsledkem jsou reálná čísla získána na základě rozmezí známého hrubého platu a požadavků na daně.

Běžné zpracování úkolu je skoro determinovatelné, ale jen s požadavky, které mají atributy definované na daném případě např. dostatečný počet informací k danému tématu. Tedy součástí zpracování je jakási neurčitelnost. Příkladem je daňový zákon,

který závisí převážně na pravidlech, ale v některých případech je nutná případová verifikace.

Normální rozhodovací úkol je stále částečně strukturovaný, ale atributy definovatelné v daném případu převládají. Příkladem je přijímání zaměstnanců nebo ohodnocení seminární práce studenta.

Ve známých, skutečných rozhodovacích úkolech je typ a struktura výsledkem předem známých, ale stále se dynamicky vyvíjejících postupů pro zpracování úkolu, který se objevuje. Zpracování je ryze neurčité stejně tak jako požadavky na informace. Příkladem je rozhodování o umístění nové továrny.

Skutečný rozhodovací úkol je neočekávaný, nový a nestrukturovaný stejně tak jako výsledek. Zpracování ani požadavky na informace nejsou známy dopředu. První krok je tedy strukturování úkolu. Příkladem může být pád SSSR a zpracování této události jinými vládami.

2.2.3 Typy informací potřebných v úkolech

V designu expertního systému jsou informace klasifikovány do třech částí (Barr a Feigenbaum, 1981):

- informací o problému
- informací o doméně (oblasti)
- informací řešící problém

Informace o problému popisují strukturu, vlastnosti a požadavky na problém. Příkladem jsou nutné informace při stavbě mostu např. o typu, důvodu stavby, postup konstrukce, které obsahují informace o problému. Tyto informace jsou typické pro prostředí problému, ale v některých případech jsou dostupné z dokumentu.

Informace o doméně (stavba mostů) obsahují známá fakta, koncepty, zákonitosti, informace o síle a tepelné roztažnosti železa patřící do domény. Takové typy informací jsou vědecké a technologické podstaty. Jsou uvedené v normách, článcích, knihách a dalších odborných textech.

Informace řešící problém pokrývají metody, které problém vyjasňují. Popisují, jak jsou vytvářeny a formulovány problémy, jaké informace o problému a doméně mají být použity. Při konstrukci mostu se jedná o informace popisující pozitiva a negativa vybraných architektur mostů. Jsou to užitečné informace a jsou obvykle dostupné pouze od člověka pracujícího v oboru (Järvelin a Repo, 1984).

Tyto tři typy informací na sebe navazují, protože reprezentují tři odlišné dimenze a odlišné role v řešení problému. Všechny výše klasifikované informace jsou potřebné k řešení problému. Závisí pouze na úkolu a odlišném stupni, který je dostupný uživateli provádějící úkol. Proto i kanály, kterými jsou prezentovány tyto informace, jsou různé. Na obr. 3 představují plné šipky vstup informací, které jsou viděny jako předem jasně dané, a přerušované šipky reprezentují všechny neurčité informace, obvykle jsou to problém řešící informace.

2.2.4 Typy informačních zdrojů (klasifikace dle Byström a Järvelin, 1995)

Zdroj orientovaný na fakta

a. registry – manuálně a počítačově zpracované katalogy a soubory
b. komerční databáze

Zdroj orientovaný na problém

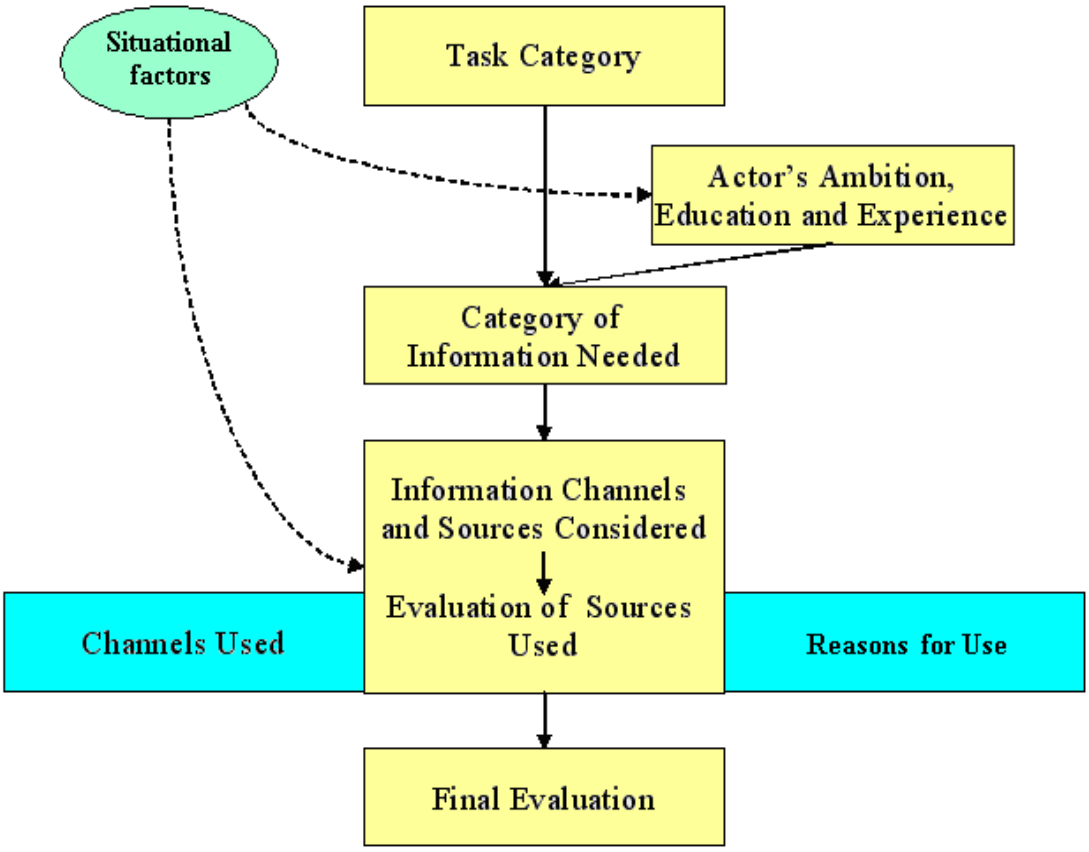
a. lidé související s problémem (podané návrhy, lidé ovlivnění problémem, administrativní činnost)
b. oficiální dokumenty (zákony, standarty, normy, předpisy, dopisy, aplikace, mapy, nepublikované dokumenty)

Zdroj se všeobecným účelem

a. experti (kolegové se znalostmi)
b. literatura (knihy, články, reporty, noviny)
c. osobní kolekce (osobní poznámky, provedené výpočty atd.)

Zdroje mohou být rozlišeny také na externí a interní dle organizace firmy a pracovníků.

Pomocí užití klasifikace úkolů, informací a informačních zdrojů je analýza dat strukturována viz. obr. 4.

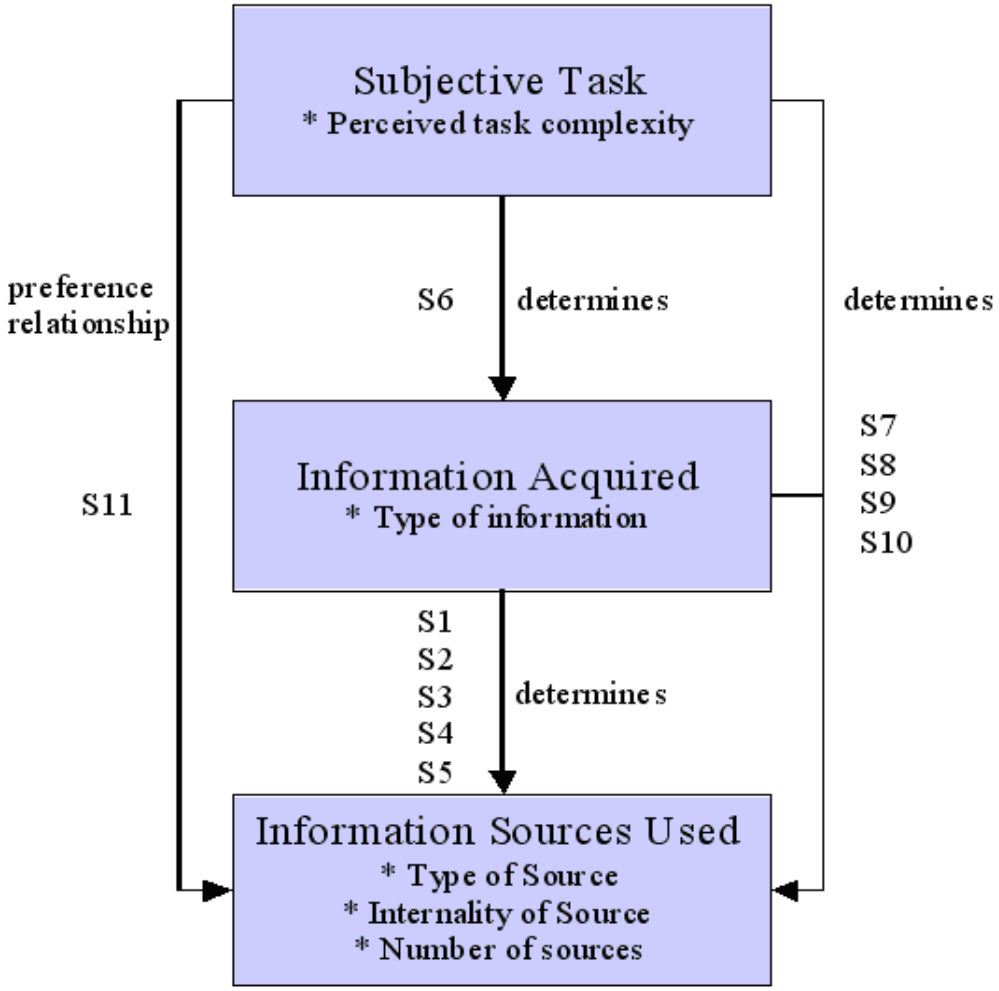


Obr. 4: Struktura dat (The work chart structure) (Byström a Järvelin, 1995)

Kombinací výše uvedených klasifikací jsou navrženy následující hypotézy typu: „Úkol týkající se komplexnosti typu X požaduje informace typu Y dostupné ze zdroje typu Z." Tedy klasifikace navrhuje analytický vztah mezi proměnnými. V modelu byla rozvinuta kvalitativní metoda pro analýzu efektů na komplexnost úkolu v informačním vyhledávání nad všemi úrovněmi úkolů. Tyto efekty jsou systematické a logické. Specifický výzkum se zaměřil na následující otázku: „Jaké typy informací jsou hledány pomocí jakých kanálů, z jakých zdrojů a jakých úkolů se to týká" (Murtonen, 1992). Výsledkem výzkumu bylo několik následujících zjištění. Nárok na úplnost informací potřebných pro provedení úkolu vzrůstá. Nárok na informační doménu a problém řešící informace také vzrůstá. Stejně tak nárok na sdílení zdrojů s obecným účelem (experti, literatura, osobní sbírky) vzrůstá. Sdílení zdrojů orientovaných na problém a fakta klesá. Oproti tomu vzrůstá tok v kanálech a počet používaných zdrojů.

Kontrast mezi jednoduchým a komplexním úkolem podtrhuje důležitost a následky komplexnosti úkolu. Pro porozumění a formulace problému jsou nezbytně nutné

rozdílné typy a více komplexních informací, které jsou získávány různými typy kanálů z různých zdrojů. Na základě této studie byl navržen konečný model založený na systému úkolů naznačený v obr. 5. Model obsahuje 11 prvků (S1-S11) (Byström, 1999)



Obr. 5: Konečný model založený na úkolech (A model of task-based information seeking) (Byström, 1999)

Tento model byl upraven tak, aby byl plně využit potenciál konceptuálního rámce v IR. Příkladem je vztah mezi typy a zdroji informací, který nebyl plně rozvinut.

2.2.5 Systematizace znalostí pomocí:

a. Začlenění dřívějších oddělených částí (studie komplexnosti úkolu z výzkumu je integrována do vyhledávací studie)
b. Zevšeobecnění a vysvětlení nižších abstraktních znalostí (pozorování, dat) pomocí konstrukce vyšší úrovně (specifické informace potřebují důkladné analýzy)
c. Rozšíření znalostí odvozením nového návrhu, který je založen na vybraných počátečních východiscích a získaných informací (pozdější empirický a teoretický vývoj jasně rozšiřuje původní návrh)
d. Vylepšení testovatelnosti hypotéz pomocí kontroly obsahu a to poskytovaná systémem hypotéz (klasifikace navrhuje mnoho podobných hypotéz S1-S11)

2.2.6 Provádění hypotéz pomocí:

a. Vytyčení důležitých problémů
b. Navržení kolekcí dat, které by nikdy nebyly sebrány bez předchozí teorie (rámec navrhuje data, která mají být získána kvůli úplnosti dat, vyhledávání založeném na úkolech a potřebných informacích)
c. Navržení úplně nového úhlu pohledu

2.2.7 Mapování reality pomocí:

a. Reprezentace nebo modelování objektů a vztahů v dané oblasti místo pouhé sumarizace dat, rámec navrhuje úkol a typy vysvětlující jevy
b. Poskytování nástrojů pro získávání nových dat (rámec je užitečný nástroj pro poskytování hypotéz a sdružuje vyhledávací metody pro produkci potřebných dat)

3 Definice IR

Ve světle komplexnosti předchozích úvah, strategií a úkolů je na místě odpovědět na otázku, co tedy znamená pojem information retrieval (IR)? Dle definice se jedná o nalezení materiálů (dokumentů) s nestrukturovanou podstatou (většinou textem), který uspokojí potřebu najít dané informace z rozsáhlého souboru, umístěného většinou na počítači (Manning, Raghavan a Schütze, 2008). IR řeší nalezení, organizaci, uložení, získávání a ohodnocení informací relevantních k uživatelskému dotazu.

Při tom se stalo skutečností, že IR nevrací hledanou informaci, ale pouze dokument obsahující požadovanou informaci. Systém pro získávání a zpracování informací nehledá uživateli předmět jeho dotazu, ale pouze uživatele informuje o existenci (neexistenci) daného předmětu a o umístění dokumentu obsahujícího vyhledávaný předmět. IR systém většinou vyhledává v souboru nestrukturovaných a polostrukturovaných dat (webové stránky, dokumenty, obrázky a videa).

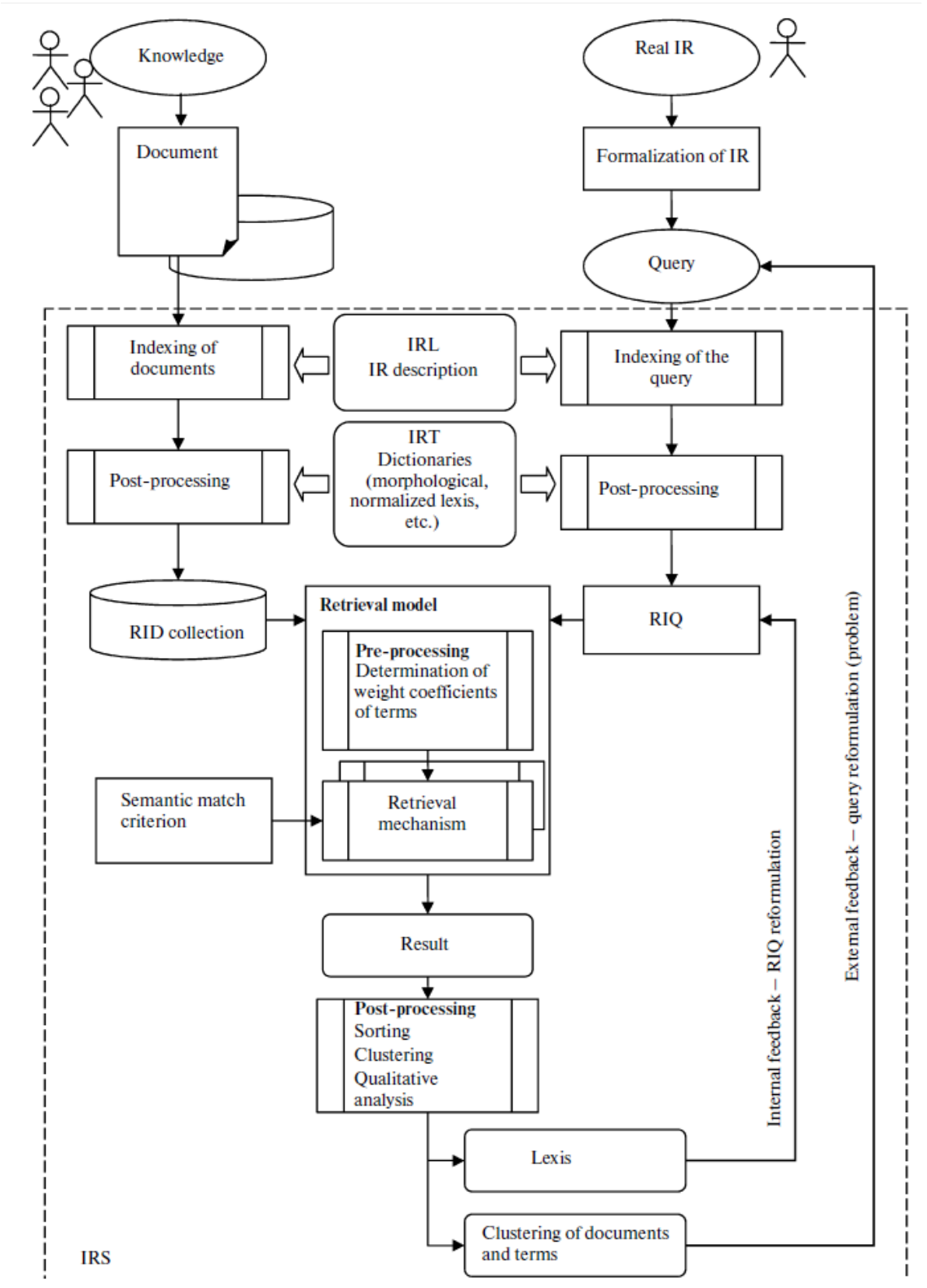
Další definice považuje IR systém za způsob podpory informací během vyhledávacího procesu, který se výrazně liší od původního systému obsahujícího data a knihovnu. Původní systém pracoval na principu faktografického získávání objektů (Golitsyna a Maksimov, 2011).

IR systém hraje podstatnou roli při generalizování vztahu člověk – stroj. Vyhledávací adresy domácích zařízení získávají a zpracovávají informace pro vytvoření nového znalostního systému s povolením ukládat zdroje. Zdroje jsou důležité také při změně stavu objektu při vykonání aktivity. Informace není sama o sobě dostatečný objekt pro charakteristiku aktivity.

Efektivita IR systému je definována přínosem řídící aktivity (principal activity PA). Konkrétně se jedná o komponenty, které se podílejí na zvýšení efektivity IR systému. Jsou to komponenty zajišťující ukládání popisu řešení, které bylo použito na vyřešení řídící aktivity (nalezené v IR systému) namísto konstrukce aktivit a ukládání popisu aktivit získaných během vyhledávání v IR systému (Maksimov, 2001).

3.1 Na jakém principu funguje IR systém?

Za účelem kontroly vyhledávacího procesu uživatelem je model neustále rozkládán na jednotlivé kroky zpracování (dotaz-odpověď) dle uživatele koncové úlohy viz obr. 6.



Obr. 6: Princip fungování IR systému (Information Retrieval models) (Järvelin a Wilson, 2003)

Funkční dekompozice, která je organizačně vyjádřena odděleně v bloku, nabízí příležitost pro úspěšné odstranění nejistot různých typů. Získávání dat je proces, který může být redukován na pouhou selekci pomocí porovnání jednotlivých položek nalezených s každým předmětem uloženým v datovém poli (Golitsyna a Maksimov, 2011). V tomto případě máme dva faktory, které se využívají při porozumění metod, jakými získáváme data. Tyto dva faktory jsou následující:

- Preferuje se srovnání dobře strukturovaných popisů namísto získaných obrazů předmětů
- Zpracování je komplexní a je obvykle vykonáváno různými typy operací

První faktor vychází z komunikativní povahy problému, která předvídá řešení na bázi lingvistického významu. Druhý faktor vychází z technické povahy problému, který předpokládá, že úkol je během procesu zpracován a redukován na úkoly týkající se struktury dat a vyhledávacího algoritmu.

Z uživatelského pohledu jsou dva typy v řídícím operačním zařízení (principal operational facilities) – dotazy a dokumenty. Dokumenty a dotazy jazykově reprezentují nedílné sémantické fragmenty předmětové domény. Další funkční vybavení, a to technologické (technological operational facilities) obsahuje nezávislé předměty nesoucí informace s meta-hodnotami předmětů, které jsou odvozeny od klíčových předmětů. Cílem technologických objektů je lokalizovat, odstranit, přemístit nebo opravit nejistoty různých druhů, stejně tak jako poskytovat efektivní zpracování dat. Například během konstrukce dotazu použít slova přirozeného jazyka, který je charakteristický synonymy. Jazyk ovlivňuje meta informace předmětů pomocí užití různých slov (thesaurus, slovníky a ontologie) (Maksimov, 2001).

První krok generalizování, který je naznačen na obrázku č. 1 (jednotlivé fáze procesu zpracování), je lokace reálného požadavku (strukturně-logická dekompozice předmětové domény), během níž jsou detekovány charakteristické koncepty a vztahy jako možné aspekty k posouzení při vyhledávání předmětu. Dalším krokem (lingvistická úroveň) je nalezení množiny se jmény konceptů a s výrazy jazyka IR, které formují třídu pomocí pravidla podmíněné rovnocennosti pro každý jednotlivý objekt s jeho specifickým aspektem. Zjednodušeně se jedná o formování možných variant terminologických struktur, které vyjadřují podstatu IR v možném posuzovaném aspektu. Komponenta s meta informacemi poskytuje sémantiku použitých slov charakteristických svým jazykem, který odráží jejich užití během

procesu indexování dokumentu. Tento krok byl automatizován velmi efektivně s využitím strojových slovníků, metod a algoritmů na analýzu textu a strojových překladačů. Je však nutno podotknout, že taková metoda nereflektuje sémantiku reálného požadavku. Obecně platí, že uživatel nemůže získat kompletní odpověď v jednokrokovém vyhledávacím procesu s využitím „inteligentního“ systému. Stejně tak pravděpodobnostní charakter při zpracování a konstrukci podmíněné rovnocennosti tříd předpokládá ověření hypotézy konzistentnosti (důsledku). Díky takovému charakteru je nutné, aby uživatel ohodnotil adekvátnost výrazů dle stupně jejich důležitosti v dokumentu, který byl nalezen. Vyhledávací stádium zpracování je konstrukce množiny dokumentů, která koresponduje s vyhledávaným obrazem požadavku (retrieval image of query = RIQ) s použitím jakéhokoliv vyhledávacího modelu. Zde je nutné vzít v úvahu sérii transformací znalostí (knowledge). Vyhledávaný obraz dokumentu (retrieval image of document = RID) a problémová situace (problem situation) se mění na vyhledávaný obraz požadavku (RIQ). Tyto transformace představují pouze malý počet variant. Rozmanitost reprezentace dokumentů a požadavků znamená, že nepřesnost či nesprávnost popisu jakéhokoliv požadavku může způsobit nadměrný výstup, který postupně vede ke zvyšování konečného výstupu a obvykle snižuje jeho přesnost. Nechtěné výsledky lze dosáhnout reformulací nebo obohacením požadavku přidáním jiného terminologického systému. Podstatou takové metody je jednoduše zvýšit možnosti vstupu, který doplňuje již používaný lingvistický systém stejně tak, jako využít správnou metodu rozhraní pro formulaci požadavku (slovníky, list klasifikovaných nadpisů a synonymický slovník).

Druhá metoda je založena na užití více vyhledávacích mechanismů, včetně těch s různě silnými kritérii pro důležitost odpovědi hodnocené pomocí zpětné vazby. Tato metoda je definována různými variantami operačních objektů a kontrolních parametrů. Existují dva typy zpětné vazby kontrolující formy vyhledávacích požadavků a zpracovávacího procesu:

A. Interní zpětná vazba (úroveň lingvistická) umožňuje jednu reformulaci vyhledávacího obrazu požadavku (RIQ), aby byl sladěn uživatelský pohled s terminologií IR systému. Toto pravidlo odráží zvláštnost využití jazyka při zpracování tam, kde je jazyk reprezentovaného statickými charakteristikami založený na indexu frekventovanosti výrazů oproti sémantice jazyka. Hodnoty výrazů při konstrukci větších struktur (věty v

dokumentech) jsou definovány dosti přesně uživatelem a obvykle bez jasně vymezených meta informací. Význam výrazu v množině dokumentů je definován méně kompletněji než uživatel uvede. Statistické parametry pro využití slova mohou být použity při ohodnocení kvality efektivity požadavku v dané předmětové doméně.

B. Externí zpětná vazba (úroveň sémantiky předmětové domény) je založena na analýze struktury obrazu předmětové domény, který je postupně získán při vyhledávacím procesu. Vazba koresponduje s uživatelským pohledem během fáze lokalizace problému.

Model IR systému obsahuje dvě komponenty:

A. model s výrazy, které reprezentují metody ošetřující spolupráci dat, které obsahují vyhledávané obrazy dokumentů (RID)
B. model, který vyhodnocuje sémantické rozdílnosti

Spolupráce dvou modelů na základě indexů (metoda založená na klíčovém obsahu dokumentu nebo požadavku). Je samozřejmostí, že klíčová slova nejsou v tomto případě propojena s ostatními. V tomto bodu koresponduje n-dimenzionální sémantický prostor s jednotlivými výrazy. Spolupráci výrazů lze reprezentovat různými výrazy.

A. Termíny specifikují charakteristické rysy popisovaných objektů pomocí kombinací metod (povinné/volitelné, zaměnitelné, asociativní atd.), které korespondují s Booleovskou logikou. Smysl termínů je definován predikátem, který má formu logické formule (množiny výrazů).
B. Výrazy získaných obrazů dokumentů (RID) jsou považovány za vektory, které obvykle shrnují formy s novými výrazy v n-dimenzionálním sémantickém prostoru.
C. Podíl výrazů na klíčovém významu je definován subjektivně formou váženého koeficientu.
D. Význam a role výrazu je objektivně definován jako výsledek použití slova.

Tradiční vyhledávací modely pracují nad vyhledávacím obrazem dokumentu nebo požadavku, který reprezentuje podmnožinu nezávislých výrazů v adresáři IR systému (množinu lexikálních položek v poli dokumentu).

3.2 Jazyk v IR systémech

Popisný IR jazyk je nejrozšířenější a nejvíce adekvátní jazyk, jehož slova jsou používána při popisu množin. Gramatika odráží metody vytváření vyhledávaného obrazu dokumentu (RID) pomocí interakce mezi popisy některých předmětů (Mikhailov, Chernii a Giliarevskii, 1968).

3.3 Typologie vyhledávacích úloh

A. Vyhledávání za známých podmínek jako například vyhledávání faktů nebo činností konkrétního autora. Toto vyhledávání je úloha prvního typu – **předmětné nebo přívlastkové vyhledávání**.
B. Druhý typ úloh vyhledávání je **tematické vyhledávání**, které zahrnuje sbírku informací o jednom konkrétním tématu. Jedná se například o přezkoumání vědeckého problému, odůvodnění či vyhledání metody k vyřešení problému. Úloha spočívá v hledání popisu aktuálního nebo teoreticky existujícího objektu, jehož vlastnosti jsou kompletně definovány již stávající množinou atributů, ale jejíž hodnota je neznámá.
C. Třetím typem úloh obsahující problém vyhledávání, kde základem je klíčová komponenta k vytvoření postupu, se nazývá **vyhledání problému**. Úloha je o hledání popisu objektu nebo jeho komponentu, které potencionálně existují v předmětové doméně (subject domain SD). Celkově komponenty vytvoří novou oblast, jejíž obsah (vlastnosti) nebude tvořit pouze souhrn vlastností komponent. Takže vlastní atributy oblasti nebudou korespondovat s vlastnostmi dříve nalezených atribut. Nově vzniklý obsah oblasti může být množina, který obsahuje kombinací již známých atributů. V tomto případě nejistota při procesu získávání předmětové domény je charakteristikou tohoto vyhledávání. Nejistota se hlavně týká předmětů, protože nejistotu způsobuje celkové pochopení struktury předmětové domény, které nemusí korespondovat s realitou.

Koncept systému je vždy nějakým způsobem svázán s určitou nejistotou při zpracování a při očekávání výsledku. Jestliže se jedná o řízené zpracování s výběrem, je činnost systému založena na porovnání dat obdržených z venku (částečně využito IR systémem s dostupnými informacemi). Nejistota výběru informací je podmínkou pro úspěšnou transformaci informací skrz řetězec porozumění, vyjádření a formalizaci do dvou řetězců:

A. **znalost – informace – dokument – vyhledávací obrázek dokumentu**

B. **řešení problému – úloha – dotaz – vyhledávací obrázek požadavku**

3.4 Typologie vyhledávacích požadavků (information requirement)

Dříve, než je požadavek adresován systému, prochází čtyřmi základními stádii (Taylor, 1968). Nejdříve uživatel zadá uvědomělý požadavek na potřebné informace nutné k vyřešení problému – **reálný požadavek** (real information requirement). Z důvodu vnímání a prezentace dostupných znalostí (výběr známých a neznámých) je řešení problému redukováno na stádium problému nebo úlohy (problem or task), které jsou sémanticky reprezentovány v jazyce předmětové domény. Tehdy je reálný požadavek transformován do stádia **vnímaného požadavku** (perceived IR). Pro získávání informací z vnějšího zdroje by měl být požadavek vědomě vyjádřen v jazyce komunikace mezi osobou a počítačem ve formě vyhledávacího dotazu. Tomuto stádiu se říká **vyjádřený požadavek** (expressed IR). Poslední stádium je transformace vyjádřeného požadavku do stádia **formalizovaného požadavku** (formalized IR), kde je dotaz složen za podmínek IR systému do formy vyhledávacího obrázku daného dotazu například binární reprezentací výrazu a vztahu. Každá transformace je jiná a se svojí vlastní neurčitostí. Během transformace z reálného požadavku je definováno kritérium informační účelné produktivity pomocí měření důležitosti. Transformace se objevují i v lidském vědomí, ale jsou špatně strukturované a sestavené. Formování výrazu vyhledávacího požadavku je doprovázeno jazykovou neurčitostí, protože podstatou transformace je převést smysl (význam) požadavku na text. Tato transformace je nepřesná a existuje mnoho variant její přesnosti a celistvosti. Vyhledávací proces představuje transformaci požadavků nad množinou dokumentů, které se setkávají s reálnými požadavky. To znamená, že předmětová doména úlohy obsahuje získané informace. Ale i absence požadovaných informací může znamenat pozitivní výsledek.

4 Základní modely IR systémů

Charakteristika IR modelů je složitá, protože jeden objekt může být založen na více modelech a na druhou stranu model může být založen na více než jednom objektu, a tak existuje dvojí pohled na IR modely (Canfora a Cerulo, 2004). Jedná se o:

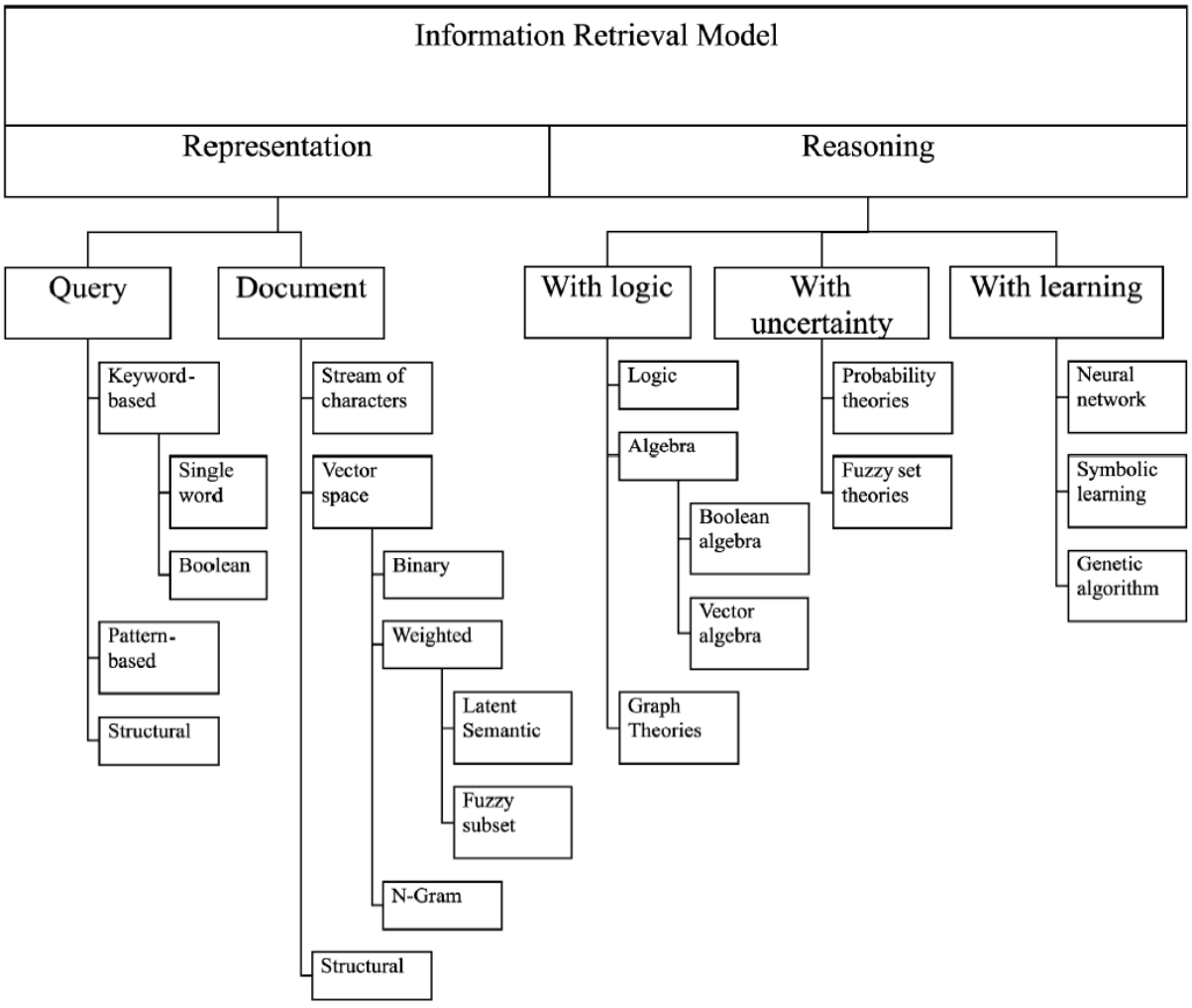
- horizontální taxonomii, která je založena na IR objektech

- vertikální taxonomii, která se zabývá IR modely

Modelování vyhledávacího procesu je složité, protože mnoho jeho částí je obtížné formulovat vzhledem k jejich povaze. Lidská složka hraje důležitou roli a mnoho konceptů jako relevantnost výsledků a potřeba vyhledaných informací je subjektivních. Proto jsou modely i jejich klasifikace velmi komplexní a složité.

4.1 Vertikální taxonomie

Navzdory složitosti klasifikace lze najít některé společné rysy. Jedním z nich je reprezentace dokumentů a požadovaných informací. Z této reprezentace je odvozena uvažovaná strategie (reasoning strategy), která řeší problém s podobností reprezentace dokumentů pomocí propočtu kvantifikace relevantnosti dokumentu vůči požadavku. Různé strategie byly představeny s cílem zlepšit vyhledávací proces. Pro popsání modelu lze využít reprezentaci dokumentů (representation) a uvažované strategie (reasoning) jak jsou uvedeny na obr. 7.



Obr. 7: Vertikální taxonomie (Vertical taxonomy) (Canfora a Cerulo, 2004)

Dle vertikální taxonomie lze model popsat jako čtveřici {D, Q, F, R(q, d)}, kde D je množina logických hledisek na dokumenty v kolekci (reprezentační komponenta), Q je množina logických hledisek na informace požadované uživatelem (reprezentační komponenta), F je rámec pro modelování prezentace dokumentů, požadavků a jejich vztahů (účelová komponenta) a R (q, d) je funkce ohodnocení. Funkce R (q, d) stanovuje určité reálné číslo v závislosti na požadavek q, kde q∈Q a dokument d, kde d∈D (Baetza-Yates a Ribeiro, 1999).

Model je popsán jako uspořádaná dvojice <Rp, Rs>, kde Rp reprezentuje model dokumentů a požadavků a Rs je rámec, který modeluje vztahy mezi dokumenty a požadavky. Toto reprezentuje proces, při kterém se formuje používaná strategie. Každá komponenta může být při zpracování rozdělena na menší komponenty a pro každou takovou komponentu lze vytvořit strom možných řešení a účelů viz. obrázek č. 2. Definováním účelu pro každou komponentu se identifikuje IR model. Nyní detailněji o jednotlivých komponentách.

4.1.1 Reprezentace požadavků

Požadavek reprezentuje informaci požadovanou uživatelem. Tyto informace pramení z problému, které jsou nutné pro úspěšné řešení problému. Jsou bezpodmínečně vytvořené v mysli uživatele a jejich existence je nutná k překlenutí znalostní mezery. Potřebné informace jsou rozděleny na tři typy (Mizzaro, 1996):

A. známé položky potřebných informací (uživatel vyhledává nebo ověřuje existenci dokumentu, který zná)
B. vědomé položky potřebných informací (uživatel vyhledává dokument, který nezná, ale s ohledem na již známý subjekt)
C. neznámé položky potřebných informací (uživatel nezná dokument ani subjekt)

V modelu jsou také prezentovány tři základní třídy reprezentace požadavků:

A. požadavky založené na klíčovém slově
B. založené na vzoru
C. založené na strukturálních požadavcích

Nejjednodušší je forma založená na **klíčovém slovu**. Je složena z klíčového slova a dokumentu, jenž obsahuje vyhledávaná slova. Tyto požadavky jsou nejvíce populární, protože jsou intuitivně a snadno vyjádřitelné. Obvykle jsou vyjádřeny jedním slovem nebo komplexněji kombinací Booleovských operací aplikovaných nad několika slovy.

Jednoslovný požadavek může být v systému vyjádřen textem, záleží na účelu komponenty. Výsledkem takto provedeného vyhledávání je obvykle soubor dokumentů obsahující minimálně jedno z vyhledávaných slov. Oproti tomu nejstarší a zatím nejrozšířenější metoda formulace požadavku pomocí Booleovským operátorů je tvořena kombinací klíčových slov. Takto tvořený požadavek obsahuje:

A. prvky, které jsou klíčovými slovy
B. Booleovské operátory
C. zápis prvků pomocí dříve provedených (precedent notation) operací

Kromě základních Booleovských operátorů (AND a OR) byly přidány i nové jako NEAR operátor (umožňuje vyhledávat pomocí kontextu) a fuzzy Booleovské operátory (které přidávají nový význam pro základní operátory AND a OR).

Dalším typem, který umožňuje více specifickou formulaci požadavku, je formulace, která dovolí další upřesnění textu s určitými vlastnostmi. Tato formulace je založena na **vzoru**. Vzor je množina syntaktických vlastností, které se musí objevit ve vybraném textu. Takto vybraný text je porovnán se vzorem, dle kterého je vyhledáváno.

Poslední typ formulace je **strukturální požadavek**, což je mechanismus, který zlepšuje kvalitu vyhledaných strukturálních informací. Tento mechanismus je postaven na základních dotazech s přidanými strukturálními omezeními například blízkost slova a další omezení zaměřená na prvky v dokumentu. Strukturální dotazy jsou rozděleny do tří kategorií:

A. fixní struktury – dotaz je nejjednodušší, a proto také nejvíce omezující. Dokument je rozdělen na pole, kde každé obsahuje určitý text. Takový dotaz vyhledává striktně text obsažený v textových polích.
B. hypertexty - dotaz pomocí hypertextem je nejflexibilnější forma. Je to orientovaný graf, kde uzly jsou text a vazby jsou spojení mezi uzly. V tomto požadavku probíhá transformace z vyhledávací na navigační aktivitu.
C. hierarchická struktura - struktura je středně pokročilý mechanismus a reprezentuje přirozenou dekompozici mnoha textových kolekcí (knihy, články atd.).

Jako příklad lze uvést XML, který je předním strukturálním modelem (Calvenese, Giacomo a Lenzerini, 1999) a Xpath, který je dotazovacím jazykem pro adresování částí obsahu v hierarchické struktuře (Hidders, 2005).

4.1.2 Reprezentace dokumentu

Dokument je vyhledávaný element v prostoru dokumentů IR systému. Je považován za minimálně nutný zdroj, kde může IR systém vyhledávat. Nejdříve byly dokumenty reprezentovány množinou výrazů nazvané jako klíčové slovo. Ta byla obvykle extrahována z textu, který byl vložen autorem. Existují tři základní typy reprezentace dokumentu a to pomocí:

A. posloupnosti znaků
B. vektorového tvaru (binárního, váženého, N-rozměrného a N-složkového vektoru)
C. struktury dokumentu

Při využití znaků je text reprezentován jako **posloupnost znaků** bez naznačení struktury nebo sémantického obsahu.

Vektorový prostor je množina, kde je každý dokument popsán jako vektor včetně sémantického obsahu dokumentu. Tradiční přístup vektorového prostoru používá klíčová slova (indexované výrazy), ale existují i jiné metody například N-gramy. Indexy jsou slova, která pomáhají sémanticky popsat hlavní témata dokumentu. Samozřejmě ne všechny výrazy jsou využity pro popsání obsahu dokumentu a indexy jsou nejvíce vhodné pro popis. Je složité vybrat adekvátní výrazy při hledání indexů. V rozsáhlé kolekci dokumentů, ve které se v každém dokumentu objevuje stále jedno slovo, je toto slovo nepoužitelné jako index, protože nerozlišuje jednotlivé dokumenty. Na druhou stranu takové slovo, které se vyskytuje pouze v jednom dokumentu, je vhodné využít jako výraz popisující obsah dokumentu (Luhn, 1957). Vektorová reprezentace může být rozdělena na tři subkategorie:

A. Binární vektorová reprezentace, kde je text v dokumentu reprezentován jako binární vektor výrazů. Každá složka vektoru zastupuje výraz s hodnotou 1, pokud se objeví v dokumentu. Pokud prvek není v dokumentu, pak vektor samozřejmě nabývá ve všech složkách hodnoty 0.
B. Vážená vektorová prezentace, kde hodnoty složek vektorů nabývají reálných čísel mezi 1 a 0, se nazývají vážená hodnota výrazu. Hodnota vyjadřuje příbuznost mezi výrazem a dokumentem. Nejrozšířenější metoda počítání hodnot váženého výrazu obsahuje dva faktory a to frekvenci výrazu (term frequency = TQ) a inverzní frekvenci výrazu (inverse frequency term = IFQ) (Van Rijsbergen, 1979). Frekvence výrazů měří, jak správně výraz popisuje

obsah dokumentu. Obrácená frekvence výrazu měří, jak termín může diskriminovat dokument v souboru dokumentů. Tyto pojmy platí pro obecný soubor, ve kterém se výrazy nacházejí mezi frekvencí a obrácenou frekvencí výrazu. (Salton a Bucklet, 1988).

C. Metoda založená na N-gramu je mezikrok v metodě vektorového prostoru. V tradičním pojetí jsou dimenze prostoru dokumentů pro daný soubor slova nebo fráze, které se objevují v souboru. V N-gram metodě jsou dimenze prostoru dokumentu N-gramy a řetězce n-členů s následně jdoucími znaky získaných z textu bez ohledu na délku slova a bez hranic slov. N-gram může být označen za špatný statistický přístup, který měří statistické vlastnosti řetězce textu v daném souboru a nebere ohled na slovní zásobu, lexikální a sémantické vlastnosti přirozeného jazyka, ve kterém je napsán dokument. Délka N-gramu (n) a metody pro extrahování N-gramů z dokumentu závisí pouze na autorovi. Například Damashek použil N-gramy o délce řetězce 5 a 6 pro shromažďování textu podle řeči a témat, při čemž využil metodu klouzavého okna. N-gramy jsou získávány posouváním okna o n znacích skrz celým dokumentem nebo požadavkem (jeden znak za jedno časové období) (Damashek, 1995). Jiní autoři používají N-gramy, které překračují hranice slova, začínají jedním slovem a končí slovem následujícím. Mezera odděluje jinak po sobě jdoucí slova (Yannakoudakis, Goyal a Huggill, 1982).

Poslední metoda je **strukturální přístup** k dokumentu, který je podobný přístupu aplikovanému nad dotazy. Metody vylepšuje vyhledávací mechanismus. Hlavním cílem je obohatit dokument přídavnými informacemi, které umožní počítači vytvořit určitou část sémantického obsahu. XML je standart pro modelování přídavných informací (Calvenese, Giacomo a Lenzerini, 1999).

4.1.3 Uvažovaná strategie (Reasoning)

Vertikální taxonomie nabízí množství metod, modelů a technologií, které pomáhají při hledání požadovaného dokumentu dotazem a jsou reprezentovány ve vyhledávacím úkolu. Primární úkol vyhledávacího systému je získání dokumentu odpovídajícího dotazu (požadavku). Uvažované komponenty definované v rámci strategie měří relevantnost mezi dokumentem a dotazem. Klíčem ke správnému adresování a pořadí komponent je správná definice relevantnosti (náležitosti).

Definice je stále velký problém mezi širokou komunitou zabývající se IR systémy. Zatím za nejrozšířenější definici je považováno (Saracevic, 2007):

Relevantnost je A (měřítko) B (spokojenosti) existující mezi C (dokumentem) a D (potřebnými informacemi) určená pomocí E (požadavku), kde

A = měřítko, odhad, propočet, posudek…

B = užitek, přínos, shoda, spokojenost…

C = dokument, reprezentace dokumentu, poskytnuté informace…

D = otázka, reprezentace otázky, potřebné informace…

E = požadavek, dotaz, zprostředkovatel, prostředník, export…

Jedna z mnoha podrobnějších definicí, dle navržených pravidel, začíná precizní analýzou vazby mezi uživatelem a systémem. Podobně byla definována důležitost, dle které je možné definovat pořadí spojení. (Mizzaro, 1998).

Klasické modely se zaměřují na uvažování s logikou a uvažování za nejistoty. Do první kategorie spadají metody založené na logice prvního řádu (Bernd, 2000), metody založené na Booleovské a vektorové algebře (Wong, Zhiarko, Raghavan a Wong, 1985). Do druhé kategorie spadají metody s nepřesnými a neurčitými aspekty vyhledávání, které jsou založeny na metodě pravděpodobnosti (Callan, 1996) a fuzzy logice (Bookstein, 1980). Počítače využívající tyto techniky se staly hitem teprve v posledních dvaceti letech (Chen, 1995). V nedávné době se objevily nové přístupy založené na teorii grafů (Brin, Page, Motwani a Winograd, 1998) a formálních ontologiích (Guarino, Masolo a Vetere, 1999).

4.1.4 Rozhodování s logikou

Metoda založená na použití logiky. Vytvořená strategie je formulována logickou formulí P (d→n), kde šipka vyjadřuje implikaci, která je samozřejmě logickou operací.

Definice: Nechť P je predikát (výrok). Pak je reprezentace dokumentu d relevantní k požadovaným informacím n.

Hlavním problémem této metody je správný výběr implikace, což znamená výběr správné logické operace, která bude nejlépe zrcadlit jejich relevantnost (Sebastiani, 1998). Nejrozšířenější logické operace jsou realizovány pomocí algebry. Pod tímto si lze představit účelovou strategii, která je založena na množině operací definovaných v algebraickém prostoru, kde algebraický prostor je množinou R uspořádaných n-tic reálných čísel s operacemi sčítání a násobení reálným číslem.

4.1.5 Booleovská algebra

Konvenční metoda využívá Booleovskou algebru s Booleovskými dotazy ke zjištění, zda nalezený dokument uspokojuje dotaz (dokument je nebo není relevantní). Podstatná limitace pro tuto metodu je, že nelze vytvořit pořadí vyhledávaných dokumentů. Mnoho Booleovských modelů bylo modifikováno za účelem možnosti řazení. Takto rozšířené modely se nazývají Booleovské operátory, protože v model byly pozměněny právě operátory (Lee, 1994).

4.1.6 Vektorová algebra

Použití váženého schématu pro reprezentaci dokumentů a požadavků je typické pro metody vektorového prostoru.

Definice: Vektorový prostor je množina vektorů nad polem R, kde R je množina reálných čísel (vektorů), množina je neprázdná a jsou definovány dvě základní operace: sčítání vektorů a násobení vektorů skalárem. Pod pojmem vektor A, rozumíme uspořádanou entici A=(a_1, a_2,…, a_n), kde a_i jsou reálná čísla. Prvky pole se nazývají skaláry. Vzdálenost dvou vektorů je měřena pomocí skalárního součinu dvou vektorů.

Příklad: Vektorový prostor nad polem P je množina V, ve které jsou definovány binární operace VxV → V, (a,b) ↦ a+b sčítání a zobrazení PxV→ V, (p,b) ↦ p*a násobení p pro libovolné prvky a, b, c ∈ V a p, q∈ P platí (Marvan, 2000):

a+b=b+a	1·a=a
a+(b+c)= (a+b)+c	p·(q·a)= (p·q) ·a
a+0=a	(p+q) ·a= (p·a) +(q·a)
a+(-a)=0	p· (a+b)= (p·a) +(p·b)

Každý dokument v modelu je indexovaný pomocí n výrazů a model bere v úvahu počet výskytu výrazů v dokumentu. Měření podobnosti dokumentu lze vyjádřit funkcí: Sim (Q, Di) = $\sum_{k=1\ldots n} (q_k * w_{ik})^2$, kde w_{ik} jsou váhy výrazů v dokumentu a q_k je požadavek. Toto vyjádření nerespektuje závislost výrazů na délce dokumentu, a proto lze zapsat podobnost přesněji dle kosinové míry cos (ϱ) (Rijsenberg, 1979).

$$\text{Sim (Q, Di)} = \frac{\sum_{k=1\ldots n} (q_k * w_{jk})}{\sqrt{\sum_{k=1\ldots n} (w_{ik})^2}} * \sum_{k=1\ldots n} (q_k)^2 = \cos\ (\varrho)$$

A podobnost dokumentu s dotazem pro binární hodnoty lze vypočítat pomocí Diceova a Jaccardova koeficientu.

$$\text{Sim (Q, Di)} = \frac{\sum_{k=1\ldots n} (q_k * w_{jk})}{\sum_{k=1\ldots n} (q_k * w_{ik}) + (\sum_{k=1\ldots n} |q_k - w_{ik}|)}$$

Funkce Sim řadí dokumenty ve výsledku a to sestupně podle výsledků Sim funkce. (Salton, 1988).

4.1.7 Teorie grafů

Poslední přístup uplatňuje teorii grafů, které pracují se strukturami složených z vrcholů a hran. Hrany tvoří spojnice některých vrcholů. Aplikace grafů na vyhledávací proces začala být zajímavá až s příchodem webu. Webové zdroje jsou snadno modelovatelné pomocí grafů, ve kterých dokumenty reprezentují vrcholy a dotazy hrany. První metody založené na grafech byly aplikovány na bibliografické dokumenty. Byly založeny na jednoduchých principech citace a bibliografického párování. Některé z těchto metod jsou využívány v prohlížečích. Jsou to algoritmy PageRank, na kterém je založen vyhledávací systém Googlu, HITS (Kleinberg, 1998), a SAE (Pirolli, Pitkow a Rao, 1996).

4.1.8 Rozhodování za nejistoty

Tato metoda zahrnuje dva základní přístupy a to pravděpodobnostní teorii a fuzzy-množinovou teorii. Pravděpodobnostní teorie byla představena v roce 1979 (Robertson a Sparck-Jones, 1976). Základní metoda je založena na následujícím tvrzení: Sjednocením uživatelského dotazu a dokumentu v souboru se pravděpodobnostní model pokusí odhadnout pravděpodobnost, s jakou uživatel najde potřebný dokument. Existuje další přístup založený na Bayesovských sítích. Model odvozený ze sítí se používá v systému INQUERY (Broglio, Callan, Croft a Nachbar, 1995). Fuzzy-množinová teorie modelů překonala limitace Booleovského IR modelu, částečně totiž překlenula nejasné a nekompletní formulace uživatelských dotazů. Modely jsou rozšířené struktury Booleovských modelů. To znamená, že Booleovské modely mohou být rozšířeny bez narušení jejich komplexnosti. Standartní modely vyžadují přesnou shodu mezi dotazem a dokumentem, kde se množina dokumentů skládá z vyhledaných a zamítnutých dokumentů. Výsledkem příliš restriktivního chování může být odmítnutí i užitečných položek a místo nich je vyhledáno mnoho neužitečných dokumentů jako odpověď na příliš obecný dotaz. Uvolnění pravidel pro vyhledávací aktivity, které ohodnocují získané položky ve vzestupném pořadí dle shodnosti k uživatelskému dotazu, napomáhá zvýšení

efektivity ve fuzzy systému. Tento cíl lze dosáhnout modifikací Booleovského systému (Kraft a Buell, 1983). Ve fuzzy rozšíření reprezentace dokumentu je cílem poskytnout více specifickou a úplnou reprezentaci obsahu dokumentu ve snaze zamezit neúplnosti a nepřesnosti při Booleovskému systému indexování. Dokument je tedy reprezentován fuzzy množinou výrazů. Ve fuzzy generalizování dotazů je dokument vyjádřen přesným jazykem. Poskytuje také uživateli prostor pro nejasné vyjádření při tvorbě dotazu jako zjednodušení komunikace. Bylo navrženo mnoho variací, které navrhují zjemnění spojení vybraných kritérií (Bordogna a Pasi, 1993). Varianta je charakteristická svým parametrickým chováním, které je nastaveno mezi dvěma extrémy AND (A) a OR (NEBO). Jiný přístup generalizuje dotaz pomocí agregačních operátorů např. lingvistické kvantifikátory AT LEAST (nejméně) nebo ABOUT (asi, okolo).

4.1.9 Rozhodování učením

Několik autorů navrhuje použití v IR systému teorii učení. Nejužitečnější přístup zahrnuje učení pomocí více vrstev (Chen, 1993), neuronové sítě (Rumelhart, Hinton a Williams, 1985), symbolický a induktivní učící algoritmus jako ID3 (Quinlan, 1983) a ID5R (Utgoff, 1989) a algoritmus založený na evoluci – genetický algoritmus (Koza, 1992). Metoda neuronových sítí je velmi dobře aplikovatelná na konvenční modely IR systému jako je pravděpodobnostní a vektorový model. Jedna z prvních aplikací byla založena na třívrstvé neuronové síti (autoři, indexy a dokumenty). Systém používá zpětnou vazbu uživatelů ke změně reprezentace uživatelů, indexů a dokumentů během celého procesu (Belew, 1989). Rozvoj této aplikace probíhal pomocí využití Hebbova učení (založeno na myšlence, že váhové hodnoty při spojení mezi dvěma neurony ve stavu ON narůstají a naopak) (Kwok, 1989). Neuronové sítě byly použity pro specifičtější úkoly v mnoha jiných aplikacích. V Kohonenově samo-organizační mapě vlastností byly použity neuronové sítě na vytvoření samo organizující reprezentaci sémantických vztahů mezi dokumenty (Lin, Soergei a Marchionini, 1991). Také byl vytvořen algoritmus pro shlukování dokumentů v neuronových sítích (MacLEod a Robertson, 1991). Pro vyhledávání založeném na obsahu dokumentu byla použita Hopfieldova neuronová síť založená na uzlech (jednotkách) (Chen, Lynch, Basu a Ng, 1993).

IR systémy založené na symbolickém učení jsou více limitovány v porovnání k ostatním systémům, protože využívají jiný typ učení. V této metodě je využita

technika pro automatickou klasifikaci textu. (Blosseville, Herbail, Monteil a Penot, 1992). Zpracování prezentuje numerickou klasifikaci výsledků dle pravidel „IF-THEN“ výrazu (Fuhr, Lustig, Schwanter, Knorz, Hartmann a Knorz, 1991). V regresní metodě byla použita implementovací technika založená na indexování dle vlastností. ID3 a ID5R jsou adaptace tohoto algoritmu (Chen a She, 1994). Oba algoritmy jsou schopny využít ukázku požadovaného dokumentu navrženou uživatelem k vytvoření rozhodovacího stromu pomocí importování klíčových slov, které ale nemohou reprezentovat uživatelský dotaz. Bylo vytvořeno mnoho genetických algoritmů ve spojení s IR systémy. Jeden z algoritmů je založen na přístupu indexování dokumentu a využívá mutace a křížení operandů. V tomto algoritmu je klíčové slovo gen, dokument je vektor klíčového slova reprezentující jednotlivé dokumenty i kolekci. Výchozí uživatel reprezentuje počáteční populaci (Gordon, 1988). Na základě Jaccardovy spojité funkce (koeficient) se počáteční populace vyvíjí v průběhu generací a vývoj konverguje k optimální zlepšené populaci. Podobná metoda je použita při shlukování dokumentů (Gordon, 1991). Souhrn modelů a jejich reprezentace je naznačen v tab. 1, která představuje metody, na nichž jsou modely založeny. Z tabulky je vidět, že nejvíce používaný přístup je založen na reprezentaci dotazu pomocí klíčového slova a dokumentu pomocí vektorového prostoru. Při určování vyhledávací strategie už takto jasně převládající přístup není.

Information Retrieval Model	Vetical Taxonomy													
	Representation						Reasoning							
	Query (Požadavek)			Document (Dokument)			With logic			With Uncertainty		With Learning		
	Keyword-	Pattern-	Structural	Stream of	Vector	Structural	Logic	Algebra	Graph	Probability theories	Fuzzy set theories	Neutral	Symbolic	Genetic
Pattern match (Baeza-Yetas a Gonnet, 1992)		X		X				X						
Vector (Salton a Lesk, 1965)	X				X			X						
Probabilistic	X				X					X				
Fuzzy	X				X						X			
Fuzzy with learning	X				X						X	X		
Fuzzy thesaurus					X		X				X			
Genetic	X				X			X						X
Neural Network	X				X			X				X		
Probabilistic Neural Network	X				X					X		X		
Neural Network Clustering					X	X		X				X		
Symbolic learning	X				X				X				X	
Browsing						X			X					
Rule-based and logic	X			X			X							

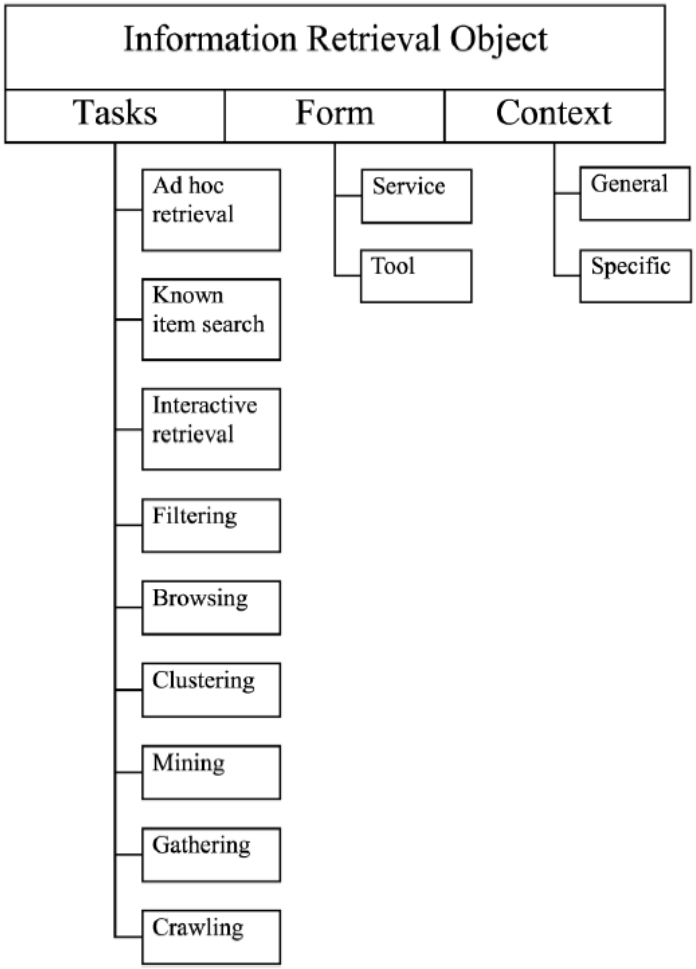
Tab. 1: Vertikální taxonomie s modely a jejich využitými přístupy (Vertical taxonomy of a set of Information Retrieval Models) (Canfora a Cerulo, 2004)

Vertikální taxonomie sama o sobě není dostatečná, aby mohla popsat a obsáhnout všechny objekty, které jsou zahrnuty pod IR systémy. Obvykle uživatelé nekomunikují přímo s modelem, ale využívají software, který řeší informačně vyhledávací problém.

Díky tomuto přístupu nazvanému horizontální taxonomie byla vytvořena nová dimenze s novým úhlem pohledu nazvaný horizontální taxonomie.

4.2 Horizontální taxonomie

Dle nového přístupu vertikální taxonomie jsou klasifikovány vyhledávané objekty jako artefakty, které řeší více nebo méně obecné problémy IR systému. Objekt je definován třemi komponentami a to úkoly či úlohami (tasks), formou (form) a obsahem (context), jak je naznačeno v obr. 8.



Obr. 8: Horizontální taxonomie objektů (Horizontal taxonomy) (Canfora a Cerulo, 2004)

4.2.1.1 Úkol

Úkol se zabývá částečným aspektem informačního vyhledávání odvozeného z uživatelského pohledu. Neměl by být zaměňován s úkolem během zpracování jako například formulace požadavku, rozšíření požadavku, porovnání, ohodnocení a reprezentace dokumentu. Získaný objekt může obsahovat jeden nebo více úkolů, úkol

může být osamocený nebo může být zapojen do procesu zpracování většího úkolu. Dle horizontální taxonomie jsou definovány následující úlohy:

A. ad hoc vyhledávání (ad hoc retrieval)
B. vyhledávání známých položek (known item search)
C. interaktivní vyhledávání (interactive retrieval)
D. filtrování (filtering)
E. procházení (browsing)
F. shlukování (clustering)
G. dolování (mining)
H. setkávání (gathering)
I. dolaďování nových nebo stávajících informací (crawling)

V jiné literatuře se lze setkat s jinou terminilogií, protože každá používá svoje vlastní názvy. Principy úloh jsou ale stejné.

A. Ad hoc vyhledávací úloha je charakterizována použitím libovolného objektu pro vyhledávání a krátkým trváním. Je to typické vyhledávání v knihovně. V tomto prostředí systém zná prohledávanou sadu dokumentů, ale nemůže předvídat téma, které bude zkoumáno (Voorhess a Harman, 2001). Odpověď systému na ad hoc vyhledávání je soubor dokumentů seřazený dle vzrůstající shodnosti k dotazu. Internetový vyhledávač je dobrým příkladem informačního objektu, který může provádět ad hoc vyhledávání.
B. Vyhledávání známých položek je podobné jako ad hoc vyhledávání, ale cílem vyhledávání je jediný dokument (nebo malá sada dokumentů), protože vyhledávání probíhá nad existujícími kolekcemi a je nutné pouze daný dokument najít (Voorhess a Harman, 2001). Objekt, který provádí takový úkol, obvykle velmi precizně implementuje jazyk dotazu, s nímž je schopen najít část dokumentu se známou strukturou nebo sémantikou. Hlavním příkladem takového úkolu je knihovní systém, kde uživatel vyhledává všechny články daného autora.
C. Interaktivní vyhledávání je úkol, kde se uživatelské rozhodnutí o užitečnosti dokumentu může lišit během vyhledávací aktivity (Kuhltau, 1991). Rozhodnutí je zachyceno pomocí interakce vyhledávacího úkolu. Během interakce se systém pokouší vnímat, jak uživatel reaguje, a mění důležitost dokumentu dle reakce. Systém proto během procesu modifikuje vyhledávací strategii (Robins,

2000). Klasické vyhledávání založené na zpětné vazbě je nejranější forma interaktivního vyhledávání (Rocchio, 1971). Uživatelské ohodnocení je zachyceno jako ANO/NE rozhodnutí o důležitosti dokumentu. Systém používá toto ohodnocení k rozšíření nebo změně dotazu (Haines a Croft, 1993)

D. Filtrování neboli selektivní šíření kombinuje vyhledávání textu a kategorizaci textu. Filtrování textu probíhá v reálném čase a je odhodnoceno 0 nebo více třídami (kategoriemi). Stejně tak jako získaný text, tak i kategorie jsou striktně asociovány s informacemi potřebnými určitou uživatelskou skupinou (malou či velkou). Každý uživatel (skupina) může přidávat, odebírat nebo modifikovat dotazy a profily dle jejich potřeby. Příklady filtrování jsou NewSieve, client/server USENET (nový filtrovací systém, který může být použitý jako základní prostředí), NewWeeder (experimentální systém) USENET a SIFT (Standart Information Filtering Tools), který zahrnuje vybrané modely šíření (jeden pro vědecko-technické reporty, druhý pro nové USENET články).

E. Pokud uživatel nepředloží konkrétní dotaz na systém, ale investuje čas na procházení prostoru s dokumenty a hledání zajímavých referencí, pak místo hledání pouze prochází. Existují tři typy procházení: **plochá, vedená strukturou a hypertextem**. Hlavní myšlenkou v **plochém procházení** je, že uživatel objevuje prostor s dokumenty s plochou organizací. Například se jedná o složky v adresáři. Ve **strukturně vedeném** procházení je uživatel veden hierarchickou strukturou, ve které jsou dokumenty organizovány do kategorií a podkategorií. **Hypertextový model** představuje navigační strukturu, která povoluje projít text krokovým způsobem. Web je nejrozsáhlejší příklad modelu úlohy procházení.

F. Shlukování je automatické poznávání a generování entit, které mohou být textovými dokumenty. Shlukování je založeno na měření podobnosti mezi dokumenty stejně tak jako na explicitní/implicitní definici odlišnosti mezi skupinou dokumentů. Je využíván ke zlepšení vyhledávacího procesu, protože vyhledávání může být omezeno na sady dokumentů, o které je zájem. Ve spojení s touto metodou se používá kategorizování. To představuje poznání, ohodnocení a přiřazení dokumentů do jedné nebo více již existujících kategorií. Příkladem kategorizování je CORA (Computer Science Research Paper Search Engine), která představuje nástroj pro automatizaci kategorizování vědeckých

článků. Za kategorizační servis lze považovat Yahoo Adresář, ve kterém je kategorizace vytvářena lidskými experty.

G. Dolování je proces automatické extrakce klíčových informací z textu dokumentu. Takové informace mohou být jazykové identifikátory, extrakce vlastností, terminologie převládajících témat, zkratek a vztahů. LEXA je příklad základního zpracovávajícího systému. IBM využívá nástroj pro dolování, který pracuje na základě spolupráce mezi homonymy v textu.
H. Matching je úloha, která zahrnuje aktivní přírůstek informací pomocí různorodých zdrojů. Meta výzkum je částečnou ukázkou setkávacího úkolu, například MetaCrawler a InFind fungují na tomto principu. Oba kombinují výstupy mnoha vyhledávacích nástrojů. Prezentují výsledky, tak, jako by byly vytvořeny pouze jedním vyhledávacím nástrojem.
I. Crawling je aktivita vybírající nové nebo aktualizující stávající zdroje informací, které byly úspěšně zpracovány jinými aktivitami (dolování nebo shlukování). Jedná se o zpracování indexů v kontextu k webu takzvaný spidering. Známými příklady jsou Scooter, ArchitextSpider, Sidewinder, Slurp, Guliver, Altavista, Excite, Infoseek, Inktomi a Northernlight.

4.2.1.2 Forma

Forma odkazuje na cestu, jakou jsou objekty nabízeny uživateli. Lze je nabídnout jako nástroj nebo servis. Když je předmět implementován jako softwarový produkt, forma je nástroj. Tato forma existuje, protože ji některé firmy kupují, prodávají a instalují. Pokud objekt existuje v jedné nebo málo instancích použitých pro doručení objektu uživateli, poté se jedná o servis. Příkladem servisu je webový prohlížeč.

4.2.1.3 Obsah

Obsah odkazuje na formu získaného objektu s odvoláním na doménu aplikace. Obsah může být obecný nebo specifický. Obecný důvod získaného objektu operuje nad různorodou doménou a obsahem. Na druhou stranu obsahově specifický systém operuje nad kolekcemi dokumentů. Ve specifické doméně se jedná o právní, obchodní dokumenty a technické výkresy. Neopomenutelným příkladem je mechanismus webového vyhledávače, kde vysoká různorodost informací požaduje velmi obecný přístup. Příkladem takového přístupu je Google, Altavista a Infoseek. Specifický vyhledávací systém je takový, který částečně vytváří domény v mysli. LEXIS-NEXIS systém je příklad, kde systém poskytuje přístup do velmi rozsáhlé kolekce právních a

obchodních dokumentů. Stejně tak jako ResearchIndex servis poskytuje volný přístup do velké kolekce vědeckých článků.

V tab. 2 je uvedena vertikální taxonomie vybraných služeb, které jsou popsány výše.

Information retrieval Objects	**Vertical Taxonomy**													
	Representation						Reasoning							
	Query			Document			With Logic			With Uncertainty		With learning		
	Keyword-based	Pattern-based	Structural	Stream of characters	Vector space	Structural	Logic	Algebra	Graph theories	Probability theories	Fuzzy set theories	Neutral network	Symbolic learning	Genetic algoritmus
CORA	X					X			X					
TACHIR						X			X					
SIFT	X				X			X						
NewsWeeder	X				X			X						
Grep		X		X				X						
LEXA		X		X				X						
OCP		X		X				X						
INQUERY	X				X					X				
SMART	X				X			X						
ILA	X		X			X	X			X				
WebLearner	X				X									
Isearch	X		X		X			X						
Google	X				X				X	X	X			
ResearchIndex	X				X	X		X	X					
Glimpse, Agrep	X			X	X			X						
Scatter/Gather	X				X			X	X					
Amalthaea	X				X				X					X
WEBSOM	X				X			X				X		

Tab. 2: Ukázka vertikální projekce vybraných služeb (Vertical projections) (Canfora a Cerulo, 2004)

II. Praktická část

Dle Golitsyna a Maksimova existují tři hlavní typy modelů: Booleovské modely, algebraické modely a pravděpodobnostní modely (Golitsyna a Maksimov, 2011).

5 Pravděpodobnostní model

Model byl poprvé navržen až v roce 1995 autory S. E. Robertsonem a K Spark-Johnsem. Nesporná výhoda oproti vektorovému modelu je, že vychází z teorie pravděpodobnosti. Pravděpodobnostní model lze proto charakterizovat jistou rozhodovací veličinou (Manning, Raghavan a Schütze, 2008). Model je založen na předpokladu, že význam výrazu je definován jeho užitím. Kalkulace pravděpodobnosti je provedena na základě poměru mezi zastoupením výrazu dotazu ve zkoumané množině, která je plně relevantní, a předchozím irelevantním dokumentem.

V teorii pravděpodobnosti existují dva přístupy. Jeden z nich je založen na využití vzorů k předpovídání důležitosti (Maron a Kuhns, 1960). Druhý z nich je založen na užití výrazu v dotazu jako klíči k určení, zda je dokument důležitý (Robertson a Sparck Jones, 1976). Využití vážených termínů je založeno na principu ohodnocení pravděpodobnosti (probability ranking principle = RPR), který předpokládá, že optimální ohodnocení dokumentu závisí na odhadu pravděpodobnosti důležitosti dokumentu ve vztahu k požadavku (Robertson, 1977). Klíčovým úkolem je ohodnocení jednotlivých komponent požadavku, které jsou využity k výpočtu konečné pravděpodobnosti dokumentu příslušného k dotazu. Výraz v dotazu je ohodnocen váhou, která koresponduje s pravděpodobností částečného termínu, který souhlasí s daným dotazem. Takový výraz pak získá požadovaný dokument. Váha každého výrazu v dotazu je kombinací konečného měření důležitosti. Na následujícím příkladu je znázorněn konkrétní výpočet pomocí Bayesovy věty.

Definice (Rychlý, 2003): Pokud A_1, A_2, až A_k jsou jevy s nenulovou pravděpodobností, jejich sjednocení je jev jistý a C je jev s nenulovou pravděpodobností, pak pro libovolné přirozené j≤k platí:

$$P\ (Aj|\mathrm{C}) = \frac{P(C|\mathrm{Aj})P(\mathrm{Aj})}{\sum P(C|\mathrm{Aj})P(\mathrm{Aj})}$$

Příklad: Pravděpodobnost, že tým vyhraje, když je pěkně, je 0.75% anebo když je spojka (hráč mezi první a druhou metou v softballu) dobrá, tak je 0.6%. Pravděpodobnost, že tým vyhraje za obou podmínek, je P (výhra|pěkně, spojka). P lze tedy vypočítat pomocí užití Bayesovy věty, kde označme pravděpodobnosti:

P (výhra|pěkně, spojka) = α P (výhra|pěkně) = ß = 0.75 P (výhra|spojka) = γ = 0.6

$$\alpha = \frac{P(\text{výhra, pěkně, spojka})}{P(\text{pěkně, spojka})} = \frac{P\ (\text{pěkně, spojka}|výhra)P(výhra)}{P(\text{pěkně, spojka})}$$

$$\frac{\alpha}{1-\alpha} = \frac{P\ (\text{pěkně, spojka}|výhra)P(výhra)}{P\ (\text{pěkně, spojka}|prohra)P(prohra)}$$

$$\frac{ß}{1-ß} = \frac{P\ (\text{pěkně}|výhra)P(výhra)}{P\ (\text{pěkně}|prohra)P(prohra)}$$

$$\frac{\gamma}{1-\gamma} = \frac{P\ (\text{spojka}|výhra)P(výhra)}{P\ (\text{spojka}|prohra)P(prohra)}$$

$$\frac{\alpha}{1-\alpha} = \left(\frac{ß}{1-ß}\right)\left(\frac{P(prohra)}{P(výhra)}\right)\left(\frac{\gamma}{1-\gamma}\right)\left(\frac{P(prohra)}{P(výhra)}\right)\left(\frac{P(výhra)}{P(prohra)}\right)$$

$$\frac{\alpha}{1-\alpha} = \left(\frac{ß}{1-ß}\right)\left(\frac{\gamma}{1-\gamma}\right)\left(\frac{P(prohra)}{P(výhra)}\right)$$

$$\frac{\alpha}{1-\alpha} = \left(\frac{0.6}{0.4}\right)\left(\frac{0.75}{0.25}\right)\left(\frac{0.5}{0.5}\right) = 4.5$$

$$\alpha = \frac{9}{11} = 0.818$$

Při splnění zároveň obou podmínek je výhra týmu jistější než při splnění jedné z nich. Nezávislé podmínky jsou klíčem k úspěchu, což v našem případě podmínky jsou. Počasí je neměnné v závislosti na tom, jak dobře hraje hráč na pozici spojky. Pokud by tyto podmínky byly závislé, mohli bychom tvrdit, že hráč hraje lépe za slunečného počasí. Podmínka nezávislosti požaduje, aby byly nezávislé i na výsledku týmu (zda vyhraje nebo ne). Pro vyhledávácí požadavek může být výraz viděn jako indikátor důležitosti dokumentu. Přítomnost nebo nepřítomnost výrazu A může být brána jako předpověď, zda je nebo není dokument důležitý. Tedy po nějaké době pozorování lze odhadnout x-procentuální důležitost dokumentu pro určitý požadavek, když je výraz A přítomný v dokumentu i v požadavku. Tato pravděpodobnost je pak přisouzena výrazu A. Takto může být ohodnocen každý výraz v požadavku. A tak výpočet celého dotazu lze použít pro výpočet pravděpodobnosti relevantnosti dokumentu.

Nevýhodou je, že lze těžko splnit podmínku absolutní nezávislosti, protože nezávislost by měla být také mezi výpočty jednotlivých pravděpodobností (Copper, 1991). Například: Požadavek (dotaz) q obsahuje dva výrazy q_1 a q_2. Po zpracování je získáno pět dokumentů D_1 až D_5, ale pouze dva z nich jsou relevantní D_2 a D_4. Z tohoto ohodnocení je vypočtena relevantnost dokumentu v závislosti obsažení výrazů v q_1 a taktéž pro q_2. Tedy důležitost dokumentu je závislá na pravděpodobnosti významu každého výrazu v dokumentu. Daný výraz v dokumentu přispívá k odhadnutí důležitosti (I) pomocí výpočtu: $I = \frac{P(Wi|rel)}{P(Wi|nonrel)}$. Vyhledaný dokument D_i obsahuje t výrazů ($w_1 \ldots w_t$), P(Wi | rel) je odhad pravděpodobnosti pro výraz i, že právě pomocí něho se najde relevantní dokument a P(Wi | nonrel) je odhad pravděpodobnosti pro výraz i, že nebude vyhledán relevantní dokument při použití výrazu i. Pro vyhledávání je důležité splnění dvou základních podmínek:

- distribuce výrazů v relevantním a nerelevantním dokumentu je nezávislá
- distribuce dokumentů je nezávislá

Stejně tak pro metody je důležité: pravděpodobnost relevantnosti je založena na přítomnosti vyhledávaného výrazu v dokumentu a pravděpodobnost relevantnosti je založena na přítomnosti a také nepřítomnosti vyhledávaného výrazu v dokumentu (Robertson a Sparck-Kones, 1976).

6 Booleovské modely

Booleovský model je jednoduchý vyhledávací model založený na teorii množin a Booleovské algebře. Koncept je intuitivní a je založen na rámci, který je jednoduše pochopitelný i pro nové uživatele i díky tomu, že dotazy jsou vytvářeny pomocí přesné sémantiky (Baeza-Yates a Ribeiro-Neto, 1999). V modelech je požadavek reprezentován logickým výrazem, které se chovají jako operandy. Během procesu vyhledávání jsou nejdříve porovnány vyhledávané obrazy dokumentů (RID) s vyhledávanými obrazy požadavků (RIQ). Tímto krokem je přítomnost nebo nepřítomnost požadavku v dokumentu. Druhý krok vyhodnotí pravdivost logické formule, která spojuje požadavek a dokument. V jednoduchém Booleovském modelu se předpokládá, že hodnotu 1 nabývá logická proměnná v přítomnosti výrazu a hodnotu 0 v nepřítomnosti výrazu v dokumentu. Dokument je formálně považován za relevantní k požadavku, když hodnota logické formule nabývá 1. V rozšířeném

Booleovském modelu je každý výraz v dokumentu ohodnocen vahou a vzdáleností mezi dokumentem a požadavkem s využitím pravidel disjunkce a konjunkce. Model zjistí, zda je výraz přítomen nebo nepřítomen a proměnná váha (index terms weight) může nabývat pouze hodnoty 0 nebo 1: $w_{i,j} \in \{0,1\}$. Výrazy v požadavku q jsou propojené třemi operátory a to NOT (NE), AND (A) a OR (NEBO). Dokument pak může nabývat pouze hodnoty ANO (relevantní) nebo NE (nerelevantní). Nikdy nemůže nastat situace, kde by dokument nabýval hodnoty *„částečná shoda"*. Nejvýraznějšími charakteristickými rysy modelu jsou snadná formalizovanost a jednoduchost, protože booleovské vyhledávání je založeno na teorii Booleovské algebry s jasnými pravidly. Tento model nabízí rámec, který je určen i pro běžného uživatele IR systému. Dotazy jsou specifikovány Booleovskými výrazy, které mají přesně danou sémantiku.

Model má nedostatek, tj. vyhledávací strategie je založena na binárním rozhodovacím kritériu (dokument je nebo není relevantní) bez další stupnice. Booleovské výrazy jsou čistě sémantické, ale někdy může být obtížné převést informace do těchto výrazů. Pro uživatele je náročné vyjádřit svůj složitější požadavek dle přesně daných pravidel v Booleovské logice, protože tyto výrazy jsou mnohem jednodušší a nesplňují takový rozsah, který by uživatel chtěl. I přes tyto nedostatky jsou Booleovské modely nejvyužívanějšími modely v komerčních databázích a poskytují velmi dobrý základ pro pozdější rozvoj IR modelů (Baeza-Yates a Ribeiro-Neto, 1999). Příkladem komerčního využití je vyhledávací služba Westlaw, která je založena na Booleovském modelu s přidanou možností ohodnocení vyhledaného dokumentu (Manning, Raghavan a Schütze, 2008).

7 Algebraické modely

První model tohoto typu byl vektorový model dle G. Saltona v roce 1975. Model umožňuje i částečné párování, což je jeho největší výhoda oproti Booleovskému modelu. Částečného párování lze dosáhnout pomocí ohodnocení výrazů v dotazu s využitím nebinárních hodnot (weights). Tyto hodnoty jsou použity pro výpočet podobnosti mezi každým dokumentem uloženým v paměti a uživatelským dotazem. Vyhledávání je uspořádáno vzestupně od největší shody, a tak vektorový model umožní i vyhledání dokumentu s částečnou shodou. Proto je vektorové vyhledávání přesnější (Baeza-Yates a Ribeiro-Neto, 1999). Vyhledávaný obraz dokumentů (RID) a vyhledávaný obraz požadavků (RIQ) jsou v modelu vektory, které jsou zastoupeny

výrazy v prostoru adresáře IR systému. Vzdálenost mezi dokumentem a požadavkem je měřena jako skalární součin mezi vektory. Uvažujeme vektor V (d) odvozený z dokumentu d obsahující jednu komponentu pro každý termín z požadavku.
Metody používající algebraické struktury jsou Booleovské algebry, svazy a lineární prostory. Metoda založená na používání Booleovské algebry byla zmíněna v předchozí kapitole viz. Booleovské modely, další metody budou popsány v následující kapitole.

7.1 Svazy

Definice: Svaz je množina M, na které je definována relace uspořádání mezi prvky tak, že ke každým dvěma prvkům existuje supremum i infimum (Kučera, 2010).
Libovolná množina L se všemi svými podmnožinami a binárními operacemi (∩ a ∪) je příkladem svazu. Pro operace (∩ a ∪) jsou používány stejné symboly jako pro průnik a sjednocení u teorie množin. Tato podobnost není náhodná, protože množina všech podmnožin je univerzální množina U, která je reprezentována svazem podmnožin uspořádaných pomocí vztahu „být podmnožinou“. Tento vztah je relací částečného uspořádání. Proto platí: a ∩ b = a ∨ b a ∪ b= a ∧ b.
Příklad: Množina L se skládá z prvků a a b. Průnik a ∩ b se nazývá nejnižší možná mez neboli infimum (the greatest lower bound). Pro infimum platí, že existuje právě jeden prvek, který je menší než a, b, ale zároveň je největší s touto vlastností. Takové prvky budeme označovat a ∨ b. Sjednocení a ∪ b se nazývá nejvyšší možná mez neboli supremum (the least upper bound). Pro supremum platí, že je to prvek, který je větší než a, b a navíc je nejmenší s touto vlastností. Pro a, b i libovolné c v množině L musí operace průnik a sjednocení splnit následující axiomy:

- a∩b ≤ a, a∩b≤ b
- jestliže c je prvek v L, pro který platí c≤a, pak c≤ a∩b
- a ≤ a∪b, b ≤ a∪b
- jestliže c je prvek v L, pro který platí a≤c a b≤c, pak a∪b ≤c

Svaz se nazývá ohraničený, jestliže v něm existuje horní ⊤ a dolní ⊥ohraničení všech jeho prvků. Pro jakýkoliv prvek d náležící svazu platí: ⊤ ≤ d ≤ ⊥. Všechny konečné svazy jsou ohraničené, nekonečné jsou ohraničené jednou mezí. Ve svazu všech podmnožin dané množiny U je sama o sobě horní hranicí U a prázdná množina ∅ je dolní hranicí. Protože jsou ∪ a ∩ tolik podobné množinovém průniku a sjednocení, mají tyto operátory mnoho stejných vlastností a platí pro ně pravidla:

- Idempotence (Idempotency): A∩A = A a A∪A = A
- Komutativnost (Commutativity): A∩B = B∩A

  A∪B = B∪A
- Asociativnost (Associativity): A∩ (B∩C) = (A∩B)∩C

  A∪ (B∪C) = (A∪B) ∪C
- Absorpce (Absorption): A∩ (A∪B) = A∩ (A∪B)

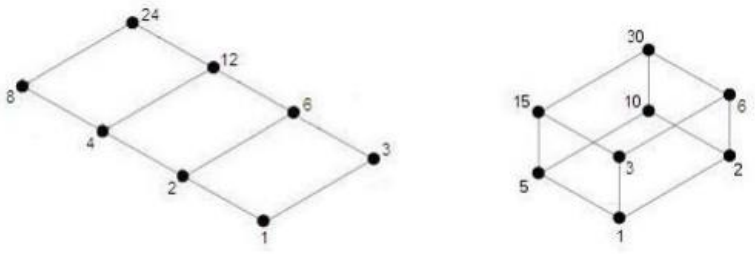
Tyto vlastnosti platí pro všechny svazy. Jsou zde určité třídy svazů, pro které platí více zákonů např. u distributivního svazu distributivní zákony. Nechť (G, ∧, ∨) je svaz a pak pro libovolné prvky a, b, c ∈ G platí:

- a ∨ (b∧c) = (a∨b) ∧ (a∨c)
- a ∧ (b∨c) = (a∧b) ∨ (a∧c)

Dalšími příklady, pro něž platí i jiné zákony, jsou komplementární svazy, pro které platí v operacích pro sjednocení a průnik De Morganovy zákony. Komplementární svaz je takový, že ke každému prvku ve svazu existuje alespoň jeden doplněk neboli komplement tohoto prvku. Doplněk množiny B k množině A je množina, která obsahuje všechny prvky z A, které zároveň nejsou v B. Nechť (G, ∧, ∨) je svaz s nejmenším prvkem 0 a největším 1 a nechť x´∈G se nazývá komplement prvku (Kučera, 2010). De Morganovy zákony určují vztah mezi sjednocením, průnikem a doplňkem množiny se systémem podmnožin. Pokud máme množinu A, B a doplněk označený „ ´ “, pak De morganovy zákony uvádějí:

- (A∪B)´ = A´∩ B´
- (A∩B)´= A´∪ B´

Příklad: Nechť n je přirozené číslo P (n), kde P je množina všech celočíselných dělitelů čísla n. Pak P(n) (nejmenší společný násobek, největší společný dělitel dvou prvků) je svaz. Pro n=24 a n=30 jsou tyto svazy znázorněny diagramem (obr. 9).



Obr. 9: Znázornění svazu pro n=24 a n=30 (Kučera, 2010)

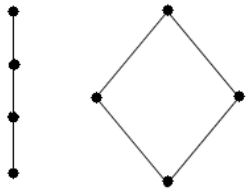
7.1.1 Reprezentace svazů

Obrázek č. 9 graficky reprezentuje zmiňované svazy. Jak již bylo zmíněno výše, svazy se využívají jako matematické modely pro objekty a vztahy mezi nimi, struktury dokumentů, vztahy mezi dokumentem a výrazem atd.. Každý z těchto svazů má jiný účel např. požadavek, dokument. Vyhledání je pak definováno jako shoda mezi požadavkem a dokumentem neboli shoda mezi svazem dokumentu a požadavku. Svazy, které při vyhledávání používají, jsou složité a jejich konstrukce je obtížná. Každý svaz lze přetvořit do částečně uspořádané množiny P (partially ordered set = poset). Pokud svaz (L, ∧, ∨) a vztah (relation)≤ je definován jako (A≤B) ⇔(A∧B=A), potom svaz lze považovat za částečně uspořádanou množinu. Svaz lze poté vizualizovat pomocí Hasseova diagramu (Dominich, 2008) následujícím způsobem:

Jakýkoliv prvek svazu je obvykle reprezentován bodem, kruhem, čtvercem nebo trojúhelníkem (záleží na tom, jaká reprezentace je vhodnější pro konkrétní problém), do něj jsou vpisována data. Dva prvky jsou spojeny čarou tak, že A≤B a neexistuje žádný prvek C ≠A, B a A≤C≤B viz. obr. 10.

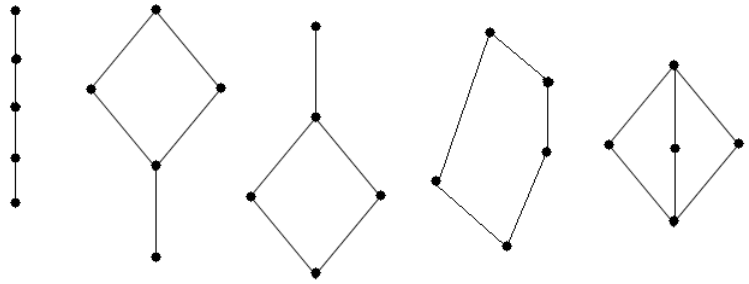
Obr. 10: Hasseho diagram reprezentující svaz, kde A≤C≤B (Dominich, 2008)

Hasseho diagram lze vytvořit využitím n =1,…, 5 prvků. Pro n = 1 je diagram svazem s jedním bodem. Pro n =2 je diagram dvou-prvkový svaz znázornitelný jako úsečka se dvěma body. Pro n = 3 je diagram znázorněn na obr. 11. Pro n = 4 je možné diagram svazu prezentovat dvěma způsoby viz. obr. 11.



Obr. 11: Diagramy se 4 prvky (Dominich, 2008)

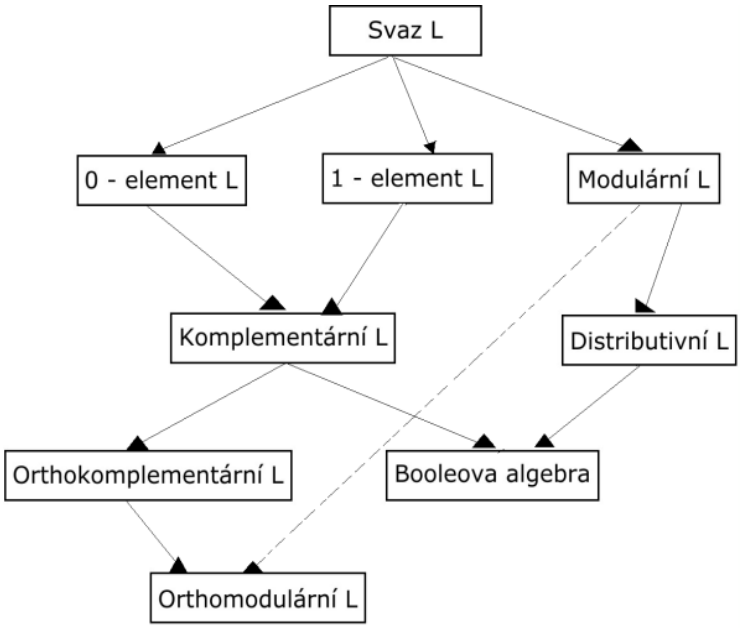
Hasseho diagram s n=5 je vizualizován na obr. 12.



Obr. 12: Diagramy s 5 prvky (Dominich, 2008)

Pro doplnění každý svaz je částečně uspořádaná množina, ale ne každá částečně uspořádaná množina je svaz. Příkladem je množina {A, B, C}, kde A ≤B a C≤B. Pokud máme vyčíslit počet uzlů ve svazu H, který má každý svaz, potom $H = O(n * 2^k)$. O vyjadřuje supremum (nejvyšší možná mez), n je počet dokumentů a k je počet dokumentů/výrazů (Godin, Missoaoui a April, 1998).

Při odlišné reprezentaci svazu pomocí definice, která říká, že svaz je množina T = {t_1, t_2, ..., t_n}. Kde T je množina prvků (výrazů, dokumentů atd.). Potom lze pomocí T definovat několik typů svazů (diagram všech typů svazů je naznačen v obr. 13): atomární a úplný, Booleovská algebra, komplementární, nemodulární, nekomplementární a modulární.



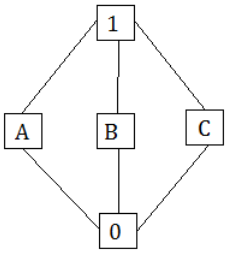
Obr. 13: Diagram typů svazů (Dominich, 2008)

7.1.1.1 Komplementární a atomární svaz

Svaz L lze považovat za úplný, pokud každá neprázdná množina v L má infinum a supremum. Jakýkoliv úplný svaz má nejmenší prvek vyjádřen 0 a největší prvek vyjádřen 1. Je tedy zřejmé, že každý konečný svaz je úplný. Za atom svazu L lze považovat takový prvek A, který splňuje: $(0 \leq B \leq A) \Leftrightarrow (0=B \text{ nebo } B=A)$. Každý svaz s 0 lze považovat za atomární pro každý prvek složený z atomů, pokud splňuje: $\forall x \in L, x \neq 0 \Rightarrow \exists \text{atom } a \neq 0$, právě tak, že $a \leq x$ (Dominich, 2008).

7.1.1.2 Modulární svaz

Jakýkoliv svaz s následujícími vlastnostmi lze považovat za slabě distributivní svaz: $A \vee (B \wedge C) \leq (A \vee B) \wedge (A \vee C)$, $\forall A, B, C \in L$. Svaz s nejmenším prvkem 0 a největším prvkem 1 je komplementární, jestliže ke každému $a \in L$ existuje komplement a^c ne nutně jediný tak, že $a \wedge a^c = 0$ a $a \vee a^c = 1$.



Obr. 14: Modulární svaz, který není distributivní, je znázorněn pomocí Hasseho diagramu (Dominich, 2008)

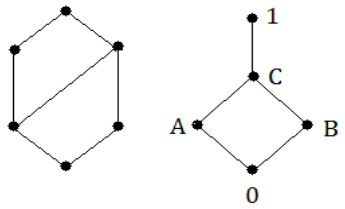
Platí, že každý podsvaz modulárního svazu je modulární a každý distributivní svaz je modulární (Dominich, 2008).

7.1.1.3 Podsvaz

Nechť L je svaz, kde $\emptyset \neq S \subseteq L$ poté S se nazývá podsvaz svazu L, pokud platí: $\forall A, B \in S$, $A \vee B \in S$ a $A \wedge B \in S$. Platí, že každá jednoprvková množina je jeho podsvazem, prázdná množina je podsvazem libovolného prvku, každý svaz je svým podsvazem (Dominich, 2008).

7.1.1.4 Distributivní svaz

Svaz je distributivní (obr. 15), když platí pro $\forall A, B, C \in L$: $A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C)$. Je tedy zřejmé, že každý distributivní svaz je modulární, když $A \leq C$ a $A \vee C = C$ poté $A \vee (B \wedge C) = (A \vee B) \wedge C$. Ne každý modulární svaz je distributivní.

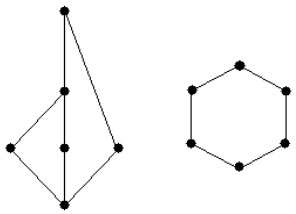


Obr. 15: Distributivní svaz znázorněn pomocí Hasseho diagramu (Dominich, 2008)

7.1.1.5 Komplementární a orthomodulární svaz

Předpokládejme, že svaz L je svaz s 0 a 1, kde je možné definovat zobrazení z L do L takové, že každému prvku A přiřadíme prvek A^c s vlastností:

$A \mid \rightarrow A^c$, $\forall A \in L$ tak abych $A \wedge A^c = 0$ a $A \vee A^c = 1$ pro každé A. $A^c \in L$ se nazývá doplněk A. Může se stát, že prvek má více než jeden doplněk. Komplementární (obr. 16) svaz se nazývá orthokomplementární (obr. 16) pokud platí: $A^{cc} = A$ a $A \leq B \Leftrightarrow B^c \leq A^c$.

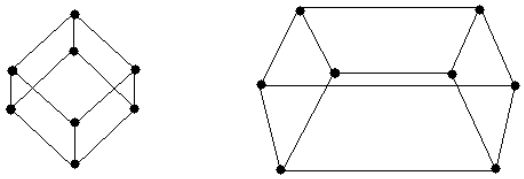


Obr. 16: Komplementární svaz a orthokomplementární svaz znázorněn pomocí Hasseho diagramu (Dominich, 2008)

Orthokomplementární svaz se nazývá orthomodulárním pokud je splněna podmínka modularity pro $B = A^c$: $A \leq C \Rightarrow A \vee (A^c \wedge C) = C$.

7.1.1.6 Booleovská algebra

Komplementární a distributivní svazy (L, ∧, ∨) se nazývají Booleovskou algebrou (obr. 17), jestliže jakýkoliv prvek a v L má jediný doplněk a^c.



Obr. 17: Booleova algebra (dva rozdílné diagramy stejné Booleovy algebry) (Dominich, 2008)

Příkladem vyhledávacích systémů, které využívají svazy, jsou Moors, FaIR, BR-Explorer, FooCA a Rajapakse-Denham.

7.1.1.7 Mooerův model

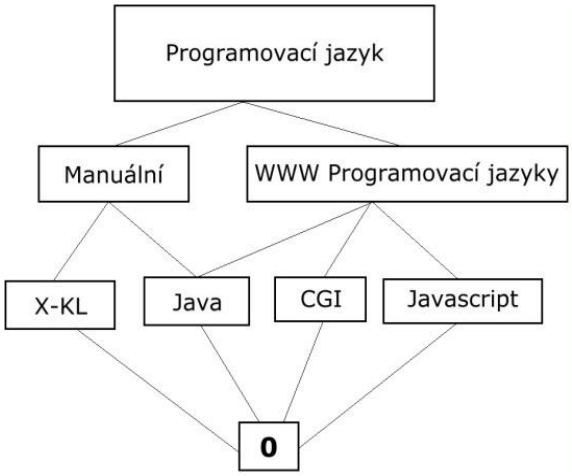
Model je založen na svazech a jako jeden z prvních zachycuje vztah mezi výrazy v dokumentu. Model obsahuje symboly (v dnešní době známé jako výrazy), které tvoří dohromady požadavek uživatele. Podmnožina dokumentů se vytváří ze souboru (knihovny). Jestliže podmnožinu označíme A a celý soubor dokumentů $L = \{D_1, .., D_n\}$, potom podmnožiny dokumentů $A \subseteq D$ vytvářejí struktury ($\mathbf{\varrho}(L)$, ∩, ∪, |) s ohledem na pravidla pro množinový průnik ∩, množinové sjednocení ∪ a komplement množiny | a kde $\mathbf{\varrho}(L)$ vytváří množinu ze všech podmnožin množiny L neboli struktura ($\mathbf{\varrho}(L)$, ∩, ∪, |) je komplementární nebo distributivní svaz. Ve svazu je požadavek Q, který obsahuje jeden nebo více výrazů. Model pracuje i s určitou hierarchií výrazů (Moors, 1959).

7.1.1.8 FaIR model

Model je založen na svazech a využívá znalostní doménu (tezaurus) k vytvoření výrazů, které mají podobu svazu (Priss, 2000). Tezaurus obsahuje množinu T s výrazy, které jsou segmentovány do tříd (faset). Fasety jsou tedy svazy a každý uzel ve svazu reprezentuje výraz (slovo nebo frázi). Každý takový svaz nebo faseta je konceptuálně kompletní, tudíž výrazy v něm obsažené představují pouze jeden pojem. Dílčí části fasety (prvky) si lze představit jako prvky, které jsou pod nimi (jdoucí až k atomům). Každý dokument je zmapovaný jako jednotný koncept skrz všechny fasety. Dokument je reprezentován tolika výrazy, kolik je jich potřeba, ale nanejvýš

jedním výrazem z každé fasety. Dokument, který obsahuje několik výrazů z fasetyje mapován pomocí spojení pojmů (obr. 18).

Příklad:



Obr. 18: Fasety svazu programovacího jazyka (Priss, 2000)

Požadavek Q je Booleovské vyjádření výrazu, kde Q='Java' vyhledá dokumenty, které přesně a pouze obsahují 'Java' při výlučném vyhledávání. Při rozšířeném vyhledávání požadavek vyhledá také dokumenty, které obsahují více obecný pojem 'WWW programovací jazyk'.

7.1.1.9 BR-Explorer systém

Vyhledávací systém je založen na svazu. Booleovský svaz výrazů je nejprve transformován na svaz pojmů L. Požadavek Q je tvořen množinou výrazů (atributů). Dle pořadí odpovědí na požadavek Q jsou výrazy vloženy do množiny pojmů L. Důležitost dokumentu d k požadavku Q je definována tehdy, když spolu sdílejí alespoň jeden atribut. Model obsahuje seřazené dokumenty dle jejich důležitosti k požadavku (Messai, Devignes, Napoli a Smail-Tabbone, 2006).

7.1.1.10 Rajapakse-Denham systém (RD systém)

V RD systému jsou dokumenty a požadavky reprezentovány individuálními svazy. Pojmy jsou extrahovány z dokumentu a jsou využity pro tvorbu svazu. Dokument nebo požadavek je tvořen strukturou objektů, atributů a vztahů mezi nimi a z toho je také vytvořen svaz. Atomy jsou prvky, které představují objekty. Objekty obsahují identické atributy. Prvky jsou objekty, které sdílejí více atributů. Nejmenší prvek svazu je prázdný prvek. Na druhou stranu největší prvek je sjednocení všech objektů. Důležitost dokumentu k požadavku je posuzována na základě jejich společných

pojmů. Hodnotu důležitosti získáme pomocí porovnávání uzlů ve svazu „požadavek“ a uzlů ve svazu „dokument“. Částečná shoda mezi svazem požadavku a svazem dokumentu je definována jako shoda mezi odpovídajícími objekty a atributy. Pokud je množina shody prázdná, poté je uplatněno klíčové slovo. V systému je využita také zpětná vazba. Proces zpětné vazby funguje na principu přidání všech výrazů obsažených v požadavku do relevantního dokumentu, pokud již v něm nejsou obsaženy. K utřídění výsledků je v modelu využit princip důležitosti dokumentu (Rajapakse a Denham, 2006).

7.1.1.11 FooCA systém

FooCA systém uplatňuje obvyklé pojetí k rozšíření webového vyhledávání pomocí svazů. Systém pracuje na platformě LINUX a je napsán v jazyce PERL. Systém umožní uživateli vložit dotaz, který zasílá na Google (lze použít i jiné prohlížeče). Výsledek je vrácen uživateli jako tabulka, kde řádky korespondují s objektem a sloupce s atributy. Pokud uživatel klikne na řádek, je automaticky přesměrován na příslušnou stránku, kde najde vyhledaný dotaz. Tabulka může být také použita pro opětovné vyhledávání, pokud uživatel klikne na příslušný atribut nebo příslušný atribut může přidat/odebrat z původního dotazu (Koester, 2006).

Svaz ve vyhledávání je dále využíván ke zdokonalení požadavku. Zde je požadavek vyjádřen pomocí Booleovské logiky a dokumenty, které přesně odpovídají požadavku, jsou vyhledány jako první. Poté je množina společných výrazů použita k vytvoření svazu pojmů. Dotaz může být zdokonalen, pokud jsou vybrány nejobecnější výrazy, které obsahují všechny výrazy z požadavku.

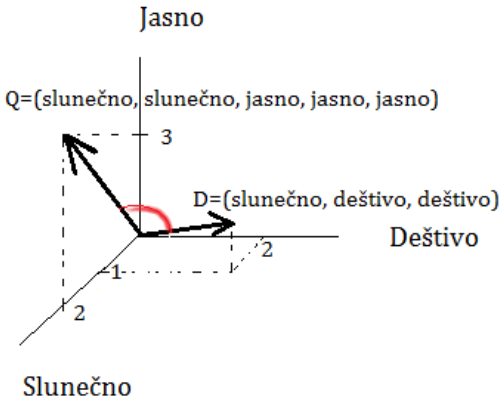
Svaz lze při vyhledávání dále využít při reprezentaci tezaurus jako svaz pojmů, který respektuje pořadí navržené tezaurusem. Svaz pojmů, který je reprezentován souborem dokumentů, může pro svoji strukturu využít cluster (svazek) jako předlohu ke svému vytvoření. Požadavek je pak zapojen do vzniklého svazu. Každý dokument je ohodnocen dle nejkratší cesty mezi požadavkem a dokumentem s pojmy. Svaz pojmů může být také použit při vyhledání v omezeném prostoru nebo při navigaci (Carpineto a Romano, 2005).

7.2 Lineární prostory

Poslední metoda využívající algebraické struktury se nazývá lineární prostor neboli vektorový prostor. V předchozích kapitolách již byl vektorový model několikrát zmíněn v souvislosti s reprezentací dokumentu a požadavku, v návaznosti na pravděpodobnostní model nebo při vyhledávání informací a rozhodnutí s logikou. Následně je uvedena pouze základní definice a přesný výpočet pro porovnání dokumentů při vyhledávání.

Definice: Jak dokument D = (t_0, w_{d0};t_1, w_{d1};…;t_t,w_{dt}) tak požadavek Q=(q_0, wq_{d0};q_1, wq_{d1};…;q_t,wq_{dt}) jsou prvky Euklidovského prostoru E_n, n= t+1 neboli formální struktura (formal framework) vektorového prostoru je E_n. Výraz w_{dt} nebo wq_{dt} vyjadřuje váhu každého výrazu t_i v dokumentu nebo požadavku. Každý výraz t_i odpovídá vektoru e_i v prostoru E_n. Stupeň důležitosti r dokumentu D závisí na vektoru w a vektoru q, který reprezentuje požadavek Q. Numerickou hodnotu tohoto stupně důležitosti získáme pomocí skalárního součinu <w,q>. Pokud je <w,q>=0, potom dokument není relevantní a není tedy vyhledán. Na druhou stranu pokud <w,q>≠0, potom je dokument relevantní a je vyhledán.

Příklad: Množina T = {slunečno, deštivo, jasno}, dokument D = {slunečno, deštivo, deštivo, jasno} a požadavek Q = {slunečno, slunečno, jasno, jasno, jasno} jsou reprezentovány vektory. Odpovídající vážené vektory jsou w=(1,2,1) a q=(2,0,3). Skalární součin (obr. 19) mezi dokumenty měří míru podobnosti mezi dokumentem a požadavkem.



Obr. 19: Skalární součin mezi vektory w a q (Dominich, 2008)

8 Aplikace metod se zaměřením na detekci plagiátorů

Již v předchozí kapitole je uvedeno, v jakých softwarech se jednotlivé modely využívají nebo pro jaké systémy byly modely předlohou. Mnoho článků je zaměřeno na téma metod vyhledávání vedoucích ke zdokonalení systému, kde cílem systému je vyhledání nejadekvátnější odpovědi na dotaz. Jen několik prací je zaměřeno na odhalování plagiátorství a fungování systému. A to i přesto, že se k odhalování plagiátů používají systémy, které jsou založeny na matematických metodách. Cíle detekce plagiátů jsou rozlišeny na dvě hlavní kategorie:

- externí, které mají za úkol identifikovat zdroj nebo původ dokumentu, který je podezřelý
- vnitřní, kde není zdrojový dokument k dispozici a opsaný text je identifikován pomocí stylistických nesrovnalostí od zbytku textu

8.1 Definice plagiarismu

Problém plagiarismu je v nepřesné interpretaci textu, který autor uvádí jako své dílo i přesto, že autorovi je jasné, že původní autor textu je někdo jiný. Zároveň si autor uvědomuje, že čtenář by se neměl danou skutečnost dozvědět a doufá, že bude mít prospěch z čtenářovy neznalosti (Samuelson, 1994). Jiná definice uvádí plagiarismus jako nepovšimnuté kopírování dokumentů, notových zápisů, programů či jejich částí. Objevuje se v mnoha souvislostech např. v průmyslu se jedná o získání konkurenční výhody, v akademické sféře se jedná o výhodu oproti kolegům a v komerční sféře se jedná o zneužití softwaru a opětovné využití kódu (Joy a Luck, 1999). Další definice uvádí plagiarismus jako neautorizované užití nebo blízká imitace nápadů, jazyka a výrazů, které autoři prezentují jako svoje vlastní.

Plagiarismus je tedy velmi úzce svázán s intelektuálním vlastnictvím (Hannabuss, 2001). Plagiáty lze dle formy rozlišit na:

- doslovné (přímé zkopírování frází nebo celých pasáží z publikovaného textu bez správného citování)
- parafrázující (slova a fráze jsou pozměněny, ale originální text je snadno rozpoznatelný)
- plagiát z druhotného zdroje (originální zdroj je citován z druhotného zdroje bez předchozího přečtení originálního textu)
- plagiát struktury (formy) původního zdroje (doslovně přepsaná forma)

- plagiát myšlenky (využití myšlenky původního textu bez využití slov nebo formy originálního textu)
- autorské (podpis pod text, který napsal jiný autor) (Nawab, Stevenson a Clough, 2011)

K nejjednodušším formám, které lze rozpoznat, jsou plagiáty doslovné nebo opětné využití slov či frází. K jejich detekci stačí nejjednodušší systém s automatickými metodami. Další formy se odhalují již hůře, hlavně pokud autor použije pouze myšlenku, ale ne obsah. Nejhorší případ plagiarismu, který se velmi těžko dokazuje, je tkz. ghost-writting, kdy za autora napíše práci někdo jiný. Tento případ se velmi těžko odhaluje.

8.2 Detekce plagiátů

Manuální detekce uvnitř samostatného, jednotného textu je založena na identifikaci nesrovnalostí autorova stylu psaní nebo v odhalení velmi dobře známých pasáží. Detekce ve více textech zahrnuje nález podobností, které jsou více než náhodné a mohou být výsledkem kopírování nebo spolupráce několika autorů. V první fázi je samostatný text přečten a jsou nalezeny jisté znaky, které odhalují možné plagiátorství. Ve druhé fázi jsou získány potřebné nástroje k odhalení on-line zdrojů nebo ruční odhalení pomocí nedigitálních zdrojů. Mezi základní signály, které vedou k důkladnějšímu prověření prací, zejména u studentů, patří:

- využití pokročilé slovní zásoby (neodpovídá autorovi)
- mnohonásobné zlepšení stylu psaní oproti minulým pracím
- nesrovnalosti v samotném textu (změna výrazů, stylu)
- nesouvislost textu (tok textu není souvislý)
- velká podobnost obsahu nebo stylu mezi dvěma a více odevzdanými pracemi
- sdílení chyb nebo omylů mezi více prácemi
- chybné reference (reference se objevují v textu, ale ne v bibliografii)
- používání odlišného stylu referencí

8.2.1 Detekce externího plagiátorství

Detekce externího plagiátorství navržená univerzitním týmem ze Sheffieldu (Velké Británie) se skládá ze tří stádií:

- předzpracování a indexace (pre-processing)
- výběr dokumentu (candidate document selection)
- detailní analýza pomocí RKR-GST (Running Karb-Rabin Greedy String Tiling.

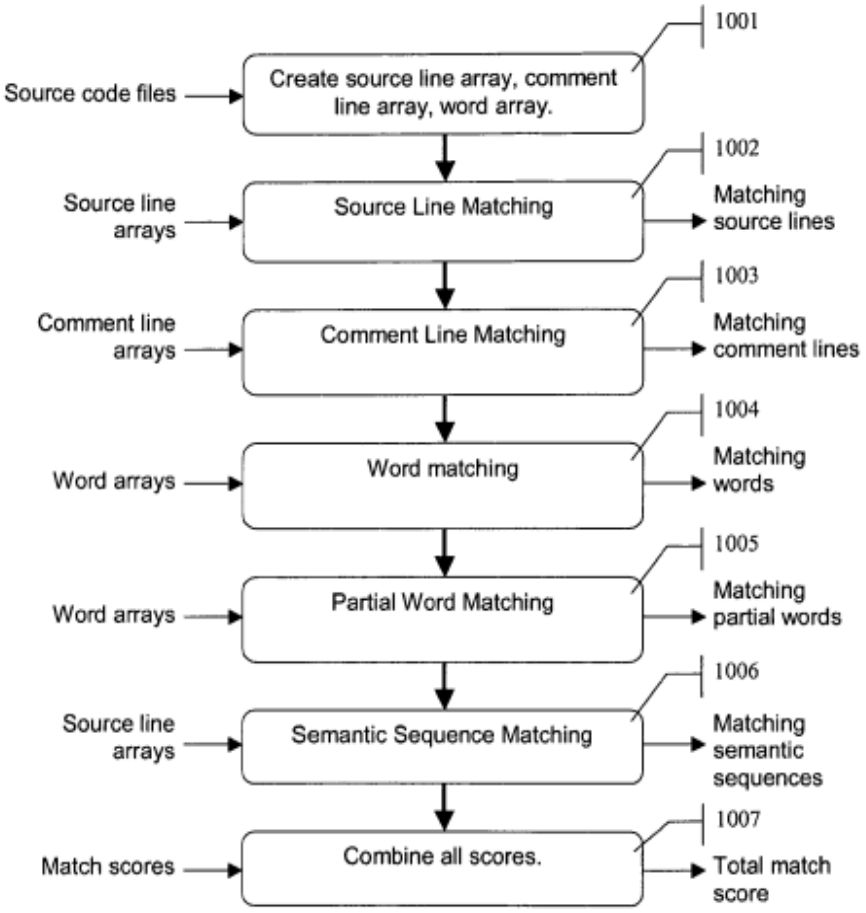
V prvním stádiu je každý dokument v souboru a v možném podezřelém souboru rozdělen do vět pomocí využití NLTK detektoru (Nawab, Stevenson a Clough, 2011). NLTK je platforma pro vývoj programů v Pythonu pro práci s daty, která jsou napsána člověkem. Poskytuje snadno použitelné rozhraní. Text je dále transformován a všechny ostatní nealfanumerické znaky jsou odstraněny. Dokument ve zdrojovém souboru je potom indexován pomocí Terrier IR systému. Terrier je flexibilní, efektivní a účinný zdrojový kód snadno rozvinutelný na rozsáhlých souborech dat.

Indexování pomocí Terrier probíhá následovně. Nejprve je v souboru dokumentů extrahován základní obsah každého dokumentu. Obsah je předán do předmětu dokumentu (Dokument objekt – název třídy v kódu). V této třídě jsou odebrány nechtěné obsahy dokumentů a výsledek je předán do třídy Token (Tokeniser object). Nakonec třída Token převede text do streamu Token, který reprezentuje obsah dokumentu (Ounis, Amati, He a MacDonald, 2014). V dalším stádiu je snaha přiřadit zdrojový dokument každému podezřelému dokumentu. Druhé stádium je velmi důležité při mnohokrokové detekci, protože pokud nenalezneme zdrojový dokument v tomto stádiu, tak v dalších krocích zpracování již nebudeme mít s čím pracovat. Ve druhém kroku se snažíme získat co nejvíce dokumentů, ačkoliv je zde určitý limitní počet pro další zpracování dokumentů.

Proces vyhledání vhodných dokumentů probíhá následovně. V prvním kroku je podezřelý dokument rozdělen do vět, které použil uživatel. Index je dotázán na každou větu, aby vyhledal potencionální zdrojové nebo původní dokumenty. Je vybrán určitý počet dokumentů ke každému požadavku. Vznikne tak soubor možných zdrojových dokumentů. V poslední fázi, kde jsou dokumenty důkladně zanalyzovány, se využívá algoritmus založený na sekvencích (RKR-GST). Tento systém byl využit na třetí mezinárodní soutěži v odhalování plagiátorů a dokáže detekovat plagiáty pouze v jednom jazyku (Nawab, Stevenson a Clough, 2011).

8.2.2 Programy na odhalování plagiátů

Nástroj pro odhalování plagiátů byl patentován až v roce 2012 a to Robertem Zeidmanem ve Spojených státech amerických. Jak software postupuje v jednotlivých fázích je naznačeno na obr. 20 (Zeidman, 2012).



Obr. 20: Základní princip softwaru (Zeidman, 2012)

Jeden z nejznámějších případů, kdy byl použit software na odhalování plagiátů, je příběh hry Edward III. publikované v roce 1596. Profesor literatury Sir Brian Vickers z Londýna použil software na detekci plagiátů k odhalení faktu, že hra Edward III. patří Shakespearovi. Pomocí softwaru určil 200 řetězců tří a více slov použitých v této hře, které odpovídají frázím v jiných Shakespearových hrách. Práce od dvou autorů obvykle mají pouze 20 shodných řetězců. Kromě přínosného využití softwaru pro určení autora neznámých her se software využívá hlavně pro detekci plagiátů. Zejména pro studenty vysokých škol jsou tyto softwary pomůckou, aby si u svých prací mohli ověřit správné citace včetně originality své práce.

8.2.3 Klasifikace vyhledávačů (engine) na detekci plagiátů

Existuje pět přístupů klasifikace vyhledávačů, což jsou programy, které porovnávají dokumenty z různých zdrojů s možností vyhledání podobnosti nebo úplné shody. Klasifikace je hlavně založena na metrikách, které programy využívají, a tak se klasifikace vyhýbá velké nejednoznačnosti, která vzniká při využití klasifikace na základě systému počítání znaků nebo systému metrik struktury (Lancaster a Culwin, 2005).

- **Tradiční klasifikace**

Tradiční přístup byl prvotně zamýšlen pro použití vyhledávačů pouze na zdrojový kód. Proto klasifikuje vyhledávače na dva typy:

- systémy počítání znaků (attribute counting system)
- systému metrik struktury (structure metric system)

Systém počítání znaků měří vlastnosti individuálního systému. Příklad je výsledekem, který zahrnuje počet objevujících se operandů v programu nebo počet možných cest programem. Systém metrik z velké části již nahradil systém počítání a lze tak nazvat jakýkoliv vyhledávač, který je založen na systému vyhledání podobnosti. Za metriku lze považovat pravidlo, které je použito pro převedení podaného dokumentu na numerickou hodnotu. Ta numericky reprezentuje podobnost. Numerické vyjádření podobnosti je známé jako míra shodnosti či podobnosti (similarity score). Tradiční klasifikaci nelze použít pro rozlišení vyhledávačů, zda používají nebo ne techniku předzpracování (tokenization), kde je v původním textu nahrazena část určitým symbolem (token) a všechna klíčová slova jsou nahrazena stejným symbolem bez ohledu na původní typ.

Příklad tokenizace textu (Manning, Raghavan a Schütze, 2008):

Vstup: Friends, Romans, Countrymen, lend me your ears

Výstup: Friends, Romans, Countrymen, lend, me, your, ears

- **Klasifikace dle typu souboru (corpora) zpracovaného vyhledávačem**

Některé vlastnosti vyhledávačů jsou velmi dobře zachyceny v předchozí klasifikaci. Jedna důležitá vlastnost není obsažena a to, zda vyhledávač operuje se zdrojovým kódem nebo volným textem. A proto systémy můžeme klasifikovat jako textové nebo netextové. S druhem obsahu je také svázaný jazyk, kde ve zdrojovém kódu vyhledávač operuje s programovacími jazyky jako C, Java nebo Prolog, a v textu se jedná o různé

jazyky. Je zřejmé, že vyhledávač pracuje s více jazyky a také tak získá nejlepší výsledek, pokud nebude zaměřen pouze na jeden jazyk nebo jeden typ jazyka.

- **Klasifikace dle dostupnosti vyhledávače**

Vyhledávače lze klasifikovat dle umístění a to následovně:

- první typ zpracovává texty na lokálním stroji
- druhý typ pracuje online na webu a všechny texty jsou podány online

Dle umístění vyhledávače lze také klasifikovat dostupnost a to na veřejné (je přístupný komukoliv) a privátní (je přístupný pouze autorizovaným osobám např. studentům určité univerzity).

- **Klasifikace dle počtu zpracovaných podání (submissions) pomocí využití metriky**

Technika, která napomáhá rozlišovat vyhledávače dle počtu zpracování, rozlišuje dvě základní metriky, a to jednoduché (singular) metriky a párové (paired) metriky. Singulární metriky pracují s jedním dokumentem a generují číselnou hodnotu. Párové metriky pracují se dvěma dokumenty a generují více informací, které jsou jednoduše sebrané, kombinované a spočítané z jednoduchých metrik. Pro doplnění se v této klasifikaci uvádějí také metriky korpolární (corpal) a vícerozměrné (multi-dimensional). V korpolární metrice musí být každý dokument uvnitř souboru zpracován a dohromady reprezentují určitou číselnou hodnotu daného souboru. Vícerozměrná klasifikace je taková, kde je zpracováno n dokumentů v souboru a dohromady jsou zastoupeny číselnou hodnotou. V tab. 3 jsou znázorněny příklady jednotlivých metrik aplikovaných na zdrojový kód a text (Lancaster a Culwin, 2005).

	Zdrojový kód (Source code)	Text (Free text)
Jednoduchá metrika (Singular metrics)	Znamená počet znaků na řádku, poměr cyklů „while“ k „for“	Znamená počet slov ve větě, poměr využití „there“ k „their“
Párová metriky (Paired metrics)	Počet klíčových společných slov ve dvou zdrojových kódech, délka nejdelšího tokenu podřetězce (substring) společného oběma kódům	Počet slov začínající velkým písmenem společné oběma textům, délka nejdelšího podřetězce společné oběma textům
Corporal metrics	Poměr klíčových společných slov v množině podání	Poměr slov vybraných ze společné skupiny množiny podání
Vícerozměrná metriky (Multi-dimensional metrics)	Poměr využití klíčového slova „while“ (zatímco) v podaném kódu	Poměr využití slova „hence“ (tedy) v podaném textu

Tab. 3: Příklad metrik ve zdrojovém kódu a textu (Lancaster a Culwin, 2005)

- **Klasifikace dle složitosti využité metriky**

Klasifikace je založena na výpočtu složitosti metody, která vyhledává podobnost. Existují dvě skupiny a to mělké (superficial) metriky a strukturální (structural) metriky. Mělké metriky vytvářejí numerickou reprezentaci dokumentu nebo množiny dokumentů, přičemž není nutná znalost vlastností jazyka. Měří podobnost, která může být změřena jednoduše při pohledu na jednu nebo více prací. Strukturální metriky popisují tradiční cestou počet vlastností dokumentu nebo dokumentů, kde je potřebná znalost jejich struktury.

	Zdrojový kód (Source code)	Text (Free text)
Mělké metriky (Superficial metrics)	Počet vyhrazených klíčových slov „while“ („zatímco“)	Počet sérií pěti slov společných pro oba texty
Strukturální metriky (Structural metrics)	Počet operačních cest skrz program	Velikost syntaktického stromu (parse tree)

Tab. 4: Příklad metrik ve zdrojovém kódu a textu (Lancaster a Culwin, 2005)

Tyto klasifikace jsou založeny na osvědčených principech a je mezi nimi velmi tenká hranice, která říká, jaký vyhledávač patří do dané klasifikace. V níže uvedené tab. 5 jsou vyznačeny současné vyhledávače a metriky.

	Typ souboru		Přístup		Dostupnost		Metriky			
Engine name (Jméno vyhledávače)	Zdrojový kód	Text	Lokální umístění	Online	Veřejný	Privátní	Jednoduchá	Párová	Strukturální	Mělká
Big Brother	X	X	X					X	X	
EVE2		X	X					X		X
TurnItIN		X		X				X	X	
JPlag	X			X				X	X	
Moss	X			X				X	X	
MyDrobBox		X		X	X			X	X	
SHERLOCK	X	X						X		X

Tab. 5: Příklad metrik ve vyhledávačích (Lancaster a Culwin, 2005)

8.2.4 Zahraniční vyhledávače

- **Turnitin**

Turnitin je jeden z nejvíce používaných softwarů na odhalení plagiátů na univerzitách ve Velké Británi. Turnitin neurčuje přesnou míru plagiarismu, ale pouze míru podobnosti s ostatními dokumenty v databázích, které obsahují přes 45miliard online odkazů (i ty, které již nejsou dostupné) a více než 337 milionů již odevzdaných prací v odborných a komerčních publikacích. Databáze jsou obnovovány a přidávány např. Elsevier databáze přidala přes milion stránek. Výsledekem turnitinu je text, který obsahuje odkazy na články, kde jsou shodné věty nebo odstavce. A tak záleží na dalším posouzení člověka, aby rozhodl, zda se jedná o plagiát.

- **Moss**

Moss je automatický systém pro odhalení podobnosti programů. Aplikace detekuje plagiáty hlavně mezi jednotlivými třídami v programu a v kódu algoritmů jednotlivých softwarů. Moss sice detekuje podobnost v kódu, ale již bez informací, proč je kód podobný. Je tedy nutné, aby někdo tuto podobnost porovnal a zjistil, zda se jedná o plagiarismus. Program je schopen analyzovat kód napsaný v C, C++, Java, C#, Python, Visual Basic, Javascript, FORTRAN, ML, Haskell, Lisp, Scheme, Pascal, Modula2, Ada, Perl, TCL, Matlab, VHDL, Verilog, Spice, MIPS assembly, a8086 assembly, MIPS assembly, HCL2. Moss je velmi jednoduchý systém, který je nabízen

jako servis na internetu. Výsledkem serveru je list obsahující odkaz HTML stránek, které obsahují podobný kód. Zároveň výsledek eliminuje shody kódu, u kterých se shoda předpokládá (knihovny).

Další nástroje volně dostupné na internetu pro kontrolu plagiarismu textu nebo softwarového kódu: PlagiarismChecker.com, DupliChecker.com, Viper, EVE2, JPlag (systém pro detekci plagiarismu v softwaru), PlagiarismDetection.org a Glatt Plagiarism Services (nástroj využívaný univerzitou v Chicagu).

8.2.5 Vyhledávače používané v ČR

- **Theses**

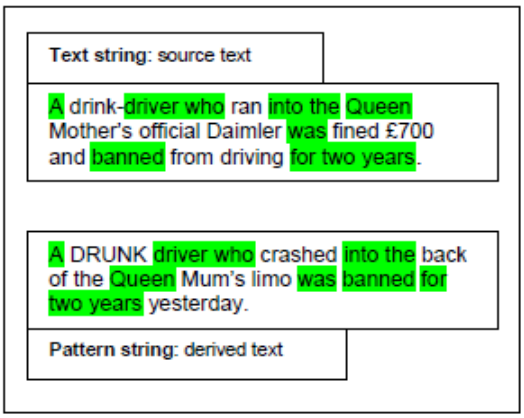
Theses je systém na odhalení plagiátů a je provozován Masarykovou univerzitou jako národní registr a jako databáze prací. Systém umožní vyhledávání mezi jednotlivými pracemi a mezi přidanými záznamy o textech včetně metadat.

- **Odevzdej**

Odevzdej je systém podobný Theseu. Rozdíl je pouze v tom, že k nástroji Odevzdej má přístup široká veřejnost a v databázi jsou kromě univerzitních prací také další seminární práce.

8.2.6 Příklad vyhledaného textu

Jedna z nejúspěšnějších metod pro odhalení plagiarismu, která byla prověřena velkým počtem dokumentů, je založena na překrytí vhodných subsekvencí a podřetězců (substring) délky ≥ n, kde n je odvozeno empiricky (měření shodnosti mezi více texty). Na obr. 21 je naznačen příklad vyhledání plagiátu založený na n-gramech. Metoda je založena na porovnání délky vybraných n-gramů, které rozlišují mezi původním a odvozený textem. Jediný problém odhalení plagiátů je v přístupu k citacím nebo ke specifické terminologii v daném oboru.



Obr. 21: Příklad vyhledání podobných textů pomocí n-gramů (The longest common substrings computed between two sentences using GST) (Clough, 2003)

Většina softwarů detekující plagiáty je schopna rozpoznat a navrhnout dokumenty, ze kterých autor určitou část převzal. Nikde však neuvádějí, proč byla práce, nápad nebo kód převzat ani v jakém smyslu. Stále je tedy nutno tyto softwary vyvíjet, aby se staly přesnějšími a mohlo by se na ně více spolehnout. Protože v dnešní době, pokud software najde 60%ní míru podobnosti s originálními dokumenty, stále to neznamená, že autorova práce je z 60% plagiát. Software (systém nebo pouze nástroj pro odhalení plagiarismu) uvádí, že 60% textu práce obsahuje shodná slova, věty nebo dokonce jen pořadí čísel s jinou prací v databázích. Zatím není vyvinutý žádný software, který by spolehlivě definoval plagiarismu bez další kontroly člověka, stejně tak jako není vyvinutý spolehlivý program na detekci využití shodných nebo podobných tabulek, diagramů, grafů, vědeckých experimentů, hudby nebo dalších netextových souborů (Lancaster a Culwin, 2005). Pro další výzkum je nezbytné se zaměřit na kontrolu skrz více jazyků (např. původní text bude napsán v jiném jazyce než je kontrolovaný text), rozšíření a větší kontrola databází s texty (články), rozšíření využití textů také v přirozeném jazyku a zlepšení metod pro analýzu textů.

9 Závěr

Jak z předchozích kapitol vyplývá, tato diplomová práce provádí komplexní a ucelený přehled matematických metod pro získávání a zpracování dat s možností konkrétní aplikace v plagiarismu.

V první části diplomové práce je uvedena teorie Information Retrieval včetně jejího časového, koncepčního a částečně i technologického vývoje. Jsou popsány základní koncepty IR včetně jejich hlavního zástupce, a to konceptuálního modelu. Druhá část je zaměřena na matematické modely (pravděpodobnostní model, Booleovský model a algebraický model) používané v IR a jejich aplikace. Aplikace modelů je uvedena a popsána na těchto systémech a službách: Westlaw (Booleovský model), Mooerův model, FaIR model, BR-Explorer systém, Rajapakse-Denham systém a FooCA systém (vše algebraické modely). Třetí část se věnuje aplikaci výše zmíněných metod na detekci plagiátů včetně uvedení různé klasifikace přístupů. Diplomová práce se zabývá programy Big Brother, EVE2, TurnItIN, Jplag, Moss, MyDrobBox, SHERLOCK, Theses a Odevzdej. Diplomová práce obsahuje také příklady detekovaného textu, který by mohl být plagiát. V této části jsou srovnány klady a zápory jednotlivých softwarů.

Diplomovou práci by bylo možné rozvinout v otázce plagiarismu v netextových souborech a přidáním konkrétní ukázky kódu. Také samotný IR nabízí mnoho příležitostí k dalšímu rozvoji nejen v otázce metod, které IR využívá, ale také v oblasti využití vyhledávání v sociálních sítích. Nové vyhledávání by mělo zahrnovat oblasti jako uživatelské označení (tagging), vyhledávání konverzací a součinné vyhledávání. Takovéto vyhledávání by znamenalo novou dimenzi ve vedení osobních a sociálních informací.

Seznam použité literatury

1. ANDERBERG, M. R. *Cluster analysis for applications*. New York: Academic Press, 1973. ISBN 0120576503 9780120576500.
2. ANON. Tietosysteemin rakentaminen: Information system design. *Journal of Documentation*. 1974, vol. 53, issue 4, s. 404-426.
3. BAEZA-YATES, Ricardo a Berthier de Araújo Neto RIBEIRO. *Modern information retrieval*. New York: ACM Press, 1999, xx, 513 s. ISBN 02-013-9829-X. Dostupné z: http://people.ischool.berkeley.edu/~hearst/irbook/print/chap10.pdf
4. BAEZA-YATES, Ricardo a Gaston H. GONNET. A new approach to text searching. *Communications of the ACM* [online]. 1992, vol. 35, issue 10, s. 74-82 [cit. 2014-10-01]. DOI: 10.1145/135239.135243. Dostupné z: http://portal.acm.org/citation.cfm?doid=135239.135243
5. BARR, Avron, Edward A FEIGENBAUM a Paul R COHEN. *The Handbook of artificial intelligence*. Stanford, Calif.: HeurisTech Press, 1981, v. <1-2, 4 >. ISBN 08-657-6004-7. Dostupné z: https://ia601202.us.archive.org/3/items/handbookofartific01barr/handbookofartific01barr.pdf
6. BELEW, R. K. Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents.*Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '89* [online]. New York, New York, USA: ACM Press, 1989, s. 11-20 [cit. 2014-10-01]. DOI: 10.1145/75334.75337. Dostupné z: http://portal.acm.org/citation.cfm?doid=75334.75337
7. BELZER, Jack, William GOFFMAN a John VERHOEFF. Inefficiency of the use of Boolean functions for information retrieval systems. *Communications of the ACM*. 1961, roč. 4, č. 12, s. 557–559.
8. BERND, Thomas. Token-Templates and Logic Programs for Intelligent Web Search. *Journal of Intelligent Information Systems* [online]. 2000, roč. 14, 2/3, s. 241-261 [cit. 2014-09-04]. DOI: 10.1023/A:1008792020665. Dostupné z: http://link.springer.com/10.1023/A:1008792020665

9. BJØRNER, Susanne a Stephanie C. ARDITO. Online Before the Internet, Part 1: Early Pioneers Tell Their Stories. *Information Today Inc.* [online]. 2003, roč. 11, č. 6 [cit. 2014-08-17]. Dostupné z: http://www.infotoday.com/searcher/jun03/ardito_bjorner.shtml
10. BLOSSEVILLE, Marie-Joëlle, HEBRAIL, Marie-Gaëlle MONTEIL a Nadine PENOT. Automatic document classification: natural language processing, statistical analysis, and expert system techniques used together. *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, ACM*. 1992, č. 1, s. 51-58.
11. BOOKSTEIN, Abraham. Fuzzy requests: An approach to weighted boolean searches. *Journal of the American Society for Information Science* [online]. 1980, roč. 31, č. 4, s. 240-247 [cit. 2014-09-05]. DOI: 10.1002/asi.4630310403. Dostupné z: http://doi.wiley.com/10.1002/asi.4630310403
12. BORDOGNA, Gloria a Gabriella PASI. A fuzzy linguistic approach generalizing Boolean Information Retrieval: A model and its evaluation. *Journal of the American Society for Information Science* [online]. 1993, roč. 44, č. 2, s. 70-82 [cit. 2014-09-06]. DOI: 10.1002/(SICI)1097-4571(199303)44:2<70::AID-ASI2>3.0.CO;2-I. Dostupné z: http://neuron.csie.ntust.edu.tw/homework/93/fuzzy/%E6%97%A5%E9%96%93%E9%83%A8/homework_1/D9315003/A%20fuzzy%20linguistic%20approach.pdf
13. BORLUND, Pia. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation* [online]. 2000, vol. 56, issue 1, s. 71-90 [cit. 2014-10-03]. DOI: 10.1108/EUM0000000007110. Dostupné z: http://www.emeraldinsight.com/10.1108/EUM0000000007110
14. BRIN, S., L. PAGE, R. MOTWANI a T. WINOGRAD. The PageRank Citation Ranking: Bringing Order to the Web. [online]. 1999, 1-17 [cit. 2014-09-05]. Dostupné z: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf
15. BROGLIO, J., J. P. CALLAN, W.B. CROFT a D. W. NACHBAR. Document retrievaland routing using INQUERY. *Proceeding of the 3rd Retrieval Conference TREC*. 1995, č. 3, s. 29-38.
16. BUNGE, M. A. Scientific research. *Heidelberg: Springer Verlag*. 1967, vol. 2.

17. BYSTRÖM, K. *Task complexity, information types and information sources.* Tampere, 1999. Doctoral Dissertation. Tampere: University of Tampere, Acta Universitatis Tamperensis.
18. BYSTRÖM, Katriina a Kalervo JÄRVELIN. Task complexity affects information seeking and use.*Information Processing* [online]. 1995, vol. 31, issue 2, s. 191-213 [cit. 2014-10-05]. DOI: 10.1016/0306-4573(94)00041-Z. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/030645739400041Z
19. CALLAN, J. Document filtering with interface network. *Proceedings of the 19th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*. 1996, s. 262-269 [cit. 2014-09-05].
20. CALVANESE, D, G DE GIACOMO a M LENZERINI. Representing and reasoning on XML documents: a description logic approach. *Journal of Logic and Computation* [online]. 1999-06-01, roč. 9, č. 3, s. 295-318 [cit. 2014-09-04]. DOI: 10.1093/logcom/9.3.295. Dostupné z: http://logcom.oupjournals.org/cgi/doi/10.1093/logcom/9.3.295
21. CARPINETO, Claudio a Giovanni ROMANO. Using Concept Lattices for Text Retrieval and Mining. *Formal Concept Analysis* [online]. 2005, č. 3626, pp.161-170 [cit. 2015-01-22]. Dostupné z: http://search.fub.it/claudio/pdf/ICFCA03.pdf
22. CANFORA, Gerardo a Luigi CERULO. A Taxonomy of Information Retrieval Models and Tools.*Journal of Computing and Information Technology - CIT 12* [online]. 2004, č. 3, 175–194 [cit. 2014-07-29]. Dostupné z: file:///C:/Users/krizksa1/Downloads/OJS_file.pdf
23. CLEVERDON, Cyril. Evaluation Tests of Information Retrieval Systems. *Journal of Documentation*[online]. 1970, roč. 26, č. 1, s. 55-67 [cit. 2014-08-17]. DOI: 10.1108/eb026487. Dostupné z: http://www.emeraldinsight.com/10.1108/eb026487
24. COOPER, William S. Some inconsistencies and misnomers in probabilistic information retrieval. *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '91* [online]. New York, New York, USA: ACM Press, 1991, s. 57-61 [cit. 2014-10-19]. DOI: 10.1145/122860.122866. Dostupné z: http://portal.acm.org/citation.cfm?doid=122860.122866

25. DAFT, Richard L, Juhani SORMUNEN a Don PARKS. Chief executive scanning, environmental characteristics, and company performance: An empirical study. *Strategic Management Journal*[online]. 1988, vol. 9, issue 2, s. 123-139 [cit. 2014-10-04]. DOI: 10.1002/smj.4250090204. Dostupné z: http://doi.wiley.com/10.1002/smj.4250090204
26. DAMASHEK, M. Gauging Similarity with n-Grams: Language-Independent Categorization of Text. *Science* [online]. 1995-02-10, roč. 267, č. 5199, s. 843-848 [cit. 2014-09-04]. DOI: 10.1126/science.267.5199.843. Dostupné z: http://www.sciencemag.org/cgi/doi/10.1126/science.267.5199.843
27. DERVIN, B. a M. NILAV. Information needs and uses. *Annual review of information science and technology* [online]. 1986, vol.37, issue 21, s. 3-33 [cit. 2014-10-03]. Dostupné z: https://comminfo.rutgers.edu/~tefko/Courses/612/Articles/zennezdervinnilan86arist.pdf
28. DOPITOVÁ, Simona. Selekční jazyky: Deweyho desetinné třídění. *Vědecké informace a knihovnictví (VIK) a Pomocných věd historických (PVH) na Masarykově univerzitě v Brně* [online]. 2004 [cit. 2014-12-27]. Dostupné z: http://www.phil.muni.cz/~dopitova/DDT.html
29. DOMINICH, Sándor. *The modern algebra of information retrieval*. Berlin: Springer, 2008, xiv, 327 p. ISBN 978-354-0776-598.
30. DOUGLAS, Engelbart C. *AUGMENTING HUMAN INTELLECT: A CONCEPTUAL FRAMEWORK*. SRI Summary Report AFOSR-3223, Prepared for: Director of Information Sciences, Air Force Office of Scientific Research., 1962. AFOSR-3233: Summary Report.
31. ELLIS, DAVID. A BEHAVIOURAL APPROACH TO INFORMATION RETRIEVAL SYSTEM DESIGN.*Journal of Documentation* [online]. 1989, vol. 45, issue 3, s. 318-338 [cit. 2014-10-03]. DOI: 10.1108/eb026843. Dostupné z: http://www.emeraldinsight.com/10.1108/eb026843
32. ELLIS, DAVID, DEBORAH COX a KATHERINE HALL. A COMPARISON OF THE INFORMATION SEEKING PATTERNS OF RESEARCHERS IN THE PHYSICAL AND SOCIAL SCIENCES. *Journal of Documentation* [online]. 1993, vol. 49, issue 4, s. 356-369 [cit. 2014-10-03]. DOI: 10.1108/eb026919. Dostupné z: http://www.emeraldinsight.com/10.1108/eb026919

33. FISCHER, William A. The acquisition of technical information by R. *IEEE Transactions on Engineering Management* [online]. 1979, EM-26, issue 1, s. 8-14 [cit. 2014-10-04]. DOI: 10.1109/TEM.1979.6447429. Dostupné z: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6447429
34. FUHR, Norbert, Gerhard LUSTIG, Michael SCHWANTNER, Knorz KNORZ, Gerhard KNORZ a Stephan HARTMANN. AIR/X - a Rule-Based Multistage Indexing System for Large Subject Fields. *PROCEEDINGS OF RIAO'91* [online]. 1991, č. 1, s. 1-18 [cit. 2014-10-01]. DOI: 10.1.1.18.9415. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.18.9415&rep=rep1&type=pdf
35. GODIN, Robert, Rokia MISSAOUI a Alain APRIL. Experimental comparison of navigation in a Galois lattice with conventional information retrieval methods. *International Journal of Man-Machine Studies*. 1998, č. 38, pp.747-767. Dostupné z: http://www.labunix.uqam.ca/~godin_r/ijmms93.pdf
36. GOLDBERG, Emanuel. *Statistical machine* [patent]. USA. MN 223,60 W, US1838389 A. Uděleno 29. prosinec 1931. Dostupné z: http://www.google.com/patents/US1838389
37. GOLITSYNA, O. L. a N. V. MAKSIMOV. Information retrieval models in the context of retrieval tasks. *Automatic Documentation and Mathematical Linguistics* [online]. 2011, vol. 45, issue 1, s. 20-32 [cit. 2014-09-02]. DOI: 10.3103/S0005105511010079. Dostupné z: http://www.springerlink.com/index/10.3103/S0005105511010079
38. GORDON, Michael D. Probabilistic and genetic algorithms in document retrieval. *Communications of the ACM* [online]. 1988, vol. 31, issue 10, s. 1208-1218 [cit. 2014-10-01]. DOI: 10.1145/63039.63044. Dostupné z: http://portal.acm.org/citation.cfm?doid=63039.63044
39. GORDON, Michael D. User-based document clustering by redescribing subject descriptions with a genetic algorithm.*Journal of the American Society for Information Science* [online]. 1991, roč. 42, č. 5, 311–322 [cit. 2014-10-01]. DOI: 10.1002/(SICI)1097-4571(199106)42:5<311::AID-ASI1>3.0.CO;2-J. Dostupné z: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(199106)42:5%3C311::AID-ASI1%3E3.0.CO;2-J

40. GUARINO, N., C. MASOLO a G. VETERE. OntoSeek: content-based access to the Web. *IEEE Intelligent Systems*[online]. 1999, roč. 14, č. 3, s. 70-80 [cit. 2014-09-30]. DOI: 10.1109/5254.769887. Dostupné z: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=769887
41. GUPTA, Yogesh, Ashish SAINI a A.K. SAXENA. A Review on Important Aspects of Information Retrieval. *World Academy of Science, Engineering and Technology International Journal of Computer: Information, Systems and Control Engineering* [online]. 2013, roč. 7, č. 12 [cit. 2014-08-17]. Dostupné z: http://waset.org/publications/9997190/a-review-on-important-aspects-of-information-retrieval
42. HACKMAN, J. Richard. Toward understanding the role of tasks in behavioral research. *Acta Psychologica* [online]. 1969, vol. 31, s. 97-128 [cit. 2014-10-04]. DOI: 10.1016/0001-6918(69)90073-0. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/0001691869900730
43. HAINES, David a W. Bruce CROFT. Relevance feedback and inference networks. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '93* [online]. New York, New York, USA: ACM Press, 1993, roč. 16, s. 2-11 [cit. 2014-10-01]. DOI: 10.1145/160688.160689. Dostupné z: http://portal.acm.org/citation.cfm?doid=160688.160689
44. HANNABUSS, Stuart. Contested texts: issues of plagiarism. *Library Management*. 2001, vol. 22, 6/7, s. 311-318. DOI: 10.1108/EUM0000000005595. Dostupné z: http://www.emeraldinsight.com/doi/abs/10.1108/EUM0000000005595
45. HARMON, Glynn. Remembering Wiliam Goffmann: Mathematical information science pioneer. *Information Processing and Managemen*. 2008, roč. 14, č. 4, s. 1634-1647. DOI: 10.106/j.ipm.2007.12.004. Dostupné z: http://garfield.library.upenn.edu/papers/goffman.pdf
46. HIDDERS, Jan. Satisfiability of XPath Expressions. GEORG LAUSEN, Dan Suciu. *Database Programming Languages* [online]. Berlin: Springer, 2005, s. 21-35 [cit. 2014-09-04]. ISBN 9783540246077.

47. HOLMSTROM, John Edwin. Section III. Opening plenary session. *The Royal Society Scientific Information Conference: 21 červen -2 červenec 1948: Reports and Paper submitted*. 1948.
48. CHEN, Hsinchun. Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms.*Journal of the American Society for Information Science* [online]. 1995, roč. 46, č. 3, s. 194-216 [cit. 2014-10-01]. DOI: 10.1002/(SICI)1097-4571(199504)46:3<194::AID-ASI4>3.0.CO;2-S. Dostupné z: http://arizona.openrepository.com/arizona/bitstream/10150/106427/1/chen27.pdf?origin=publication_detail
49. CHEN, H., K. J. LYNCH, K. BASU a T.D. NG. Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE Expert* [online]. 1993, vol. 8, issue 2, s. 25-34 [cit. 2014-10-01]. DOI: 10.1109/64.207426. Dostupné z: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=207426
50. CHEN, H. a L. SHE. Inductive Query by Examples (IQBE): A Machine Learning Approach. *In Proceedings of HICSS*[online]. 1994, č. 3, s. 428-437 [cit. 2014-10-01]. Dostupné z: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(199806)49:8%3C693::AID-ASI4%3E3.0.CO;2-O/abstract
51. INGWERSEN, P. Cognitive perspective of information retrieval interaction. *Journal of Documentation*. 1996, vol.31, issue 1, s. 3-50.
52. JÄRVELIN, Kalervo. Kaksi yksinkertaista jäsennystä tiedon hankinnan tutkimista varten. *Kirjastotiede ja informatiikka* [online]. 1987, vol.6, issue 1, 18—24 [cit. 2014-10-04]. Dostupné z: http://ojs.tsv.fi/index.php/inf/article/view/2344/2182
53. JÄRVELIN, Kalervo a T. D. WILSON. On conceptual models for information seeking and retrieval research. *Information Research* [online]. 2003, vol. 9, issue 1 [cit. 2014-10-05]. Dostupné z: http://www.informationr.net/ir/9-1/paper163.html
54. JÄRVELIN, K. a A. REPO. A taxonomy of knowledge work support tools. *Proceedings of the Annual Meeting of the American Society for Information Science*. 1984, vol.21, s. 59-62.

55. JONES, Karen Spärck. A statistical interpretation of term specificity and its application in retrieval.*Journal of Documentation* [online]. 2004, roč. 60, č. 5, s. 493-502 [cit. 2014-08-17]. Dostupné z: http://nlp.cs.swarthmore.edu/~richardw/papers/sparckjones1972-statistical.pdf
56. KLEINBERG, Jon M. Authoritative sources in a hyperlinked environment. *Journal of the ACM* [online]. 1998, roč. 46, č. 5, s. 604-632 [cit. 2014-09-05]. DOI: 10.1145/324133.324140. Dostupné z: http://portal.acm.org/citation.cfm?doid=324133.324140
57. KOESTER, B. FooCA Web information retrieval with formalconcept analysis. In: *Verlag, Allgemeine Wissenschaft*. Muhltal, 2006.
58. KOZA, J. R. Genetic Programming: On the Programming of Computers by Means of Natural Selection. *Cambridge, MA: The MIT Press.* 1992.
59. KRAFT, Donald H. a Duncan A. BUELL. Fuzzy sets and generalized Boolean retrieval systems.*International Journal of Man-Machine Studies* [online]. 1983, roč. 19, č. 1, s. 45-56 [cit. 2014-09-06]. DOI: 10.1016/S0020-7373(83)80041-8. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0020737383800418
60. KUČERA, Radan. Základy teorie svazů. *Masarykova univerzita, Brno* [online]. 2010 [cit. 2015-01-18]. Dostupné z: http://www.math.muni.cz/~kucera/texty/Svazy2010.pdf
61. KUHLTHAU, Carol C. Inside the search process: Information seeking from the user's perspective.*Journal of the American Society for Information Science*. 1991, vol. 42, issue 5, 361–371. DOI: 10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-#.
62. KWOK, K. L. A neural network for probabilistic information retrieval. *ACM SIGIR Forum* [online]. 1989, roč. 23, SI, s. 21-30 [cit. 2014-10-01]. DOI: 10.1145/75335.75338. Dostupné z: http://portal.acm.org/citation.cfm?doid=75335.75338
63. LANCASTER, Thomas a Fintan CULWIN. Classifications of plagiarism detection engines. *Innovation in Teaching and Learning in Information and Computer Sciences* [online]. 2005, vol. 4, issue 2, s. - [cit. 2015-02-28]. DOI:

10.11120/ital.2005.04020006. Dostupné z: http://journals.heacademy.ac.uk/doi/abs/10.11120/ital.2005.04020006

64. LEE, Joon Ho. Properties of extended Boolean models in information retrieval. *Proceeding SIGIR '94 Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* [online]. 1994, s. 182-190 [cit. 2014-09-05]. Dostupné z: http://www.cis.famu.edu/~riggs/IR/p182-lee.pdf
65. LIN, X, D SOERGEI a G MARCHININI. A self-organizing semantic map for information retrieval. *Proceeding of the 14th Annual Itn. ACM SIGIR Conference on Research and Development*. 1991, č. 1, s. 262-269.
66. LUHN, Hans P. *Table lookup mechanisms* [patent]. USA. G06K7/06, US 2741429 A. Uděleno 10. duben 1956. Dostupné z: http://www.google.com/patents/US2741429
67. LUHN, H. P. A Statistical Approach to Mechanized Encoding 1 and Searching of Literary Information. *IBM Journal of Research and Development* [online]. 1957, č. 1, s. 309-317 [cit. 2014-09-04]. Dostupné z: http://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Institute/Informationswissenschaft/downloadcenter/infocenter/Informationretrieval/Luhn_1957_statistical_approach.pdf
68. MACLEOD, Kevin J. a W. ROBERTSON. A neural algorithm for document clustering. *Information Processing* [online]. 1991, vol. 27, issue 4, s. 337-346 [cit. 2014-10-01]. DOI: 10.1016/0306-4573(91)90088-4. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/0306457391900884
69. MANNING, Christopher D, Prabhakar RAGHAVAN a Hinrich SCHÜTZE. *Introduction to information retrieval*. New York: Cambridge University Press, 2008, xxi, 482 p. ISBN 05-218-6571-9. Dostupné z: http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf
70. MAKSIMOV, N. V. Components and technologies of interactive search for document information. *MFD*. 2001, roč. 3, s. 16-23 [cit. 2014-07-01].
71. MARON, M. E a J. L. KUHNS. On Relevance, Probabilistic Indexing and Information Retrieval.*Journal of the ACM* [online]. 1960, vol. 7, issue 3, s. 216-244 [cit. 2014-10-17]. DOI: 10.1145/321033.321035. Dostupné z: http://portal.acm.org/citation.cfm?doid=321033.321035

72. MARVAN, Michal. Vektorové prostory. *Matematický ústav Slezské univerzity v Opavě* [online]. 2000 [cit. 2015-01-18]. Dostupné z: http://www.slu.cz/math/cz/knihovna/docs/algebra1/9.-vektorove-prostory
73. MCBRYAN, O. A. GENVL and WWWW: Tools for taming the Web. *Computer Networks and ISDN Systems* [online]. 1994, vol. 27, issue 2, s. 308- [cit. 2014-08-31]. DOI: 10.1016/S0169-7552(94)90149-X. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S016975529490149X
74. MESSAI, Nizar, Marie-Dominique DEVIGNES, Amedeo NAPOLI a Malika SMAIL-TABBONE. BR-Explorer: An FCA-based algorithm for Information Retrieval. *Https://hal.archives-ouvertes.fr/inria-00103913/document* [online]. 2006, pp. 1-7 [cit. 2015-01-20]. Dostupné z: https://hal.archives-ouvertes.fr/inria-00103913/document
75. MIKHAILOV, A. M., A. I. CHERNII a R. S GILIAREVSKII. Osnovy Informatiki: Information Science Fundamentals. *Nauka*. 1968, s. 756-766.
76. MIZZARO, Stefano. A Cognitive Analysis of Information Retrieval. *Proceeding of CoLIS2* [online]. 1996, s. 234-250 [cit. 2014-03-04]. Dostupné z: http://sole.dimi.uniud.it/~stefano.mizzaro/research/papers/colis.pdf
77. MIZZARO, Stefano. How many relevances in information retrieval?. *Interacting with Computers* [online]. 1998, roč. 10, č. 3, s. 303-320 [cit. 2014-09-04]. DOI: 10.1016/S0953-5438(98)00012-5. Dostupné z: http://iwc.oxfordjournals.org/cgi/doi/10.1016/S0953-5438(98)00012-5
78. MOOERS, Calvin N. The next twenty years in information retrieval; some goals and predictions. *American Documentation*[online]. 1960, roč. 11, č. 3, s. 229-236 [cit. 2014-08-16]. DOI: 10.1002/asi.5090110306. Dostupné z: http://doi.wiley.com/10.1002/asi.5090110306
79. MOOERS, C. N. A mathematical theory of language symbols in retrieval. *Proceedings of International Conference of Scientific Information*., pp. 1327-1352.
80. MURTONEN, K. Tuloksellisempaan tiedonhankintatutkimukseen : prosessianalyysi tiedontarpeiden ja tiedonhankinnan tutkimuksessa [Toward more effective information seeking studies : use of process-analysis in information needs and information seeking research. *Thesis for the Degree of*

Licentiate of Social Sciences: University of Tampere, Department of Information Studies [online]. 1992, č. 1 [cit. 2014-10-06]. DOI: 10.1.1.20.3317. Dostupné z: http://citeseerx.ist.psu.edu/showciting?cid=1896330

81. NAWAB, Rao Muhammed Adeel, Mark STEVENSON a Paul CLOUGH. External Plagiarism detection using Information Retrieval and Sequence Alignment: Notebook for PAN at CLEF 2011. In: *Proceedings of the 5th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse. CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation* [online]. Holandsko, 2011 [cit. 2015-02-07]. Dostupné z: http://eprints.whiterose.ac.uk/78502/7/WRRO_78502.pdf
82. OUNIS, Iadh, G. V. P. AMATI, B. HE a C. Johnson MACDONALD. Terrier information retrieval platform. *Proceedings of the 27th European Conference on IR research*. 2005, pp. 517-519. Dostupné z: http://www.academia.edu/2687763/Terrier_information_retrieval_platformP ENG, Fuchun, Nawaaz AHMED, Xin LI a Yumao LU. Context sensitive stemming for web search.*Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07* [online]. New York, New York, USA: ACM Press, 2007, s. 639-646 [cit. 2014-09-01]. DOI: 10.1145/1277741.1277851. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=DC54D23128F9D5749D2E483C9D1588D5?doi=10.1.1.86.3910&rep=rep1&type=pdf
83. PIROLLI, Peter, James PITKOW a Ramana RAO. Silk from a sow's ear. *Proceedings of the SIGCHI conference on Human factors in computing systems common ground - CHI '96* [online]. New York, New York, USA: ACM Press, 1996, č. 1, s. 118-125 [cit. 2014-09-30]. DOI: 10.1145/238386.238450. Dostupné z: http://portal.acm.org/citation.cfm?doid=238386.238450
84. PRISS, Uta. Lattice-based Information Retrieval. *School of Library and Information Science, Indiana University Bloomington,* [online]. 2000, pp.1-15 [cit. 2015-01-20]. Dostupné z: http://callisto.nsu.ru/documentation/CSIR/qmeta/ko00.pdf
85. QUINLAN, J. R. Learning Efficient Classification Procedures and Their Application to Chess End Games. *Machine Learning An Artificial Intelligence*

Approach. Palo Alto: Tioga Publishing Company, 1983, s. 463-482. ISBN 9783662124079.

86. RAJAPAKSE, R. K. a M. DENHAM. Text retrieval with more realistic concept matching and reinforcement learning.*Information Processing and management*. 2006, č. 42, pp. 1260-1275.
87. RIJSENBERG, C. J. *Information retrieval*. London: Butterworths, 1979.
88. ROBERTSON, S. E. THE PROBABILITY RANKING PRINCIPLE IN IR. *Journal of Documentation*[online]. 1977, vol. 33, issue 4, s. 294-304 [cit. 2014-10-17]. DOI: 10.1108/eb026647. Dostupné z: http://www.emeraldinsight.com/10.1108/eb026647
89. ROBERTSON, S. E. a K. Sparck JONES. Relevance weighting of search terms. *Journal of the American Society for Information Science* [online]. 1976, vol. 27, issue 3, s. 129-146 [cit. 2014-10-19]. DOI: 10.1002/asi.4630270302. Dostupné z: http://doi.wiley.com/10.1002/asi.4630270302
90. ROBINS, David. Interactive Information Retrieval: Context and Basic Notions. *Informing Science*[online]. 2000, vol. 3, issue 2 [cit. 2014-10-01]. Dostupné z: http://www.inform.nu/Articles/Vol3/v3n2p57-62.pdf
91. ROCCHIO, J.J. Relevance feedback and query expansion. *Prentice Hall*. 1971.
92. RUMELHART, D. E., G. E. HINTON a R. J. WILLIAMS. Learning Internal Representation by Error Propagation. *Parallel Distributed Processing*. 1986, s. 318-362, The MIT Press, Cambridge.
93. RYCHLÝ, Marek. Klasifikace a predikce. *Ústav informačních systémů, VUT, Brno* [online]. 2003 s. 1-11 [cit. 2015-01-16]. Dostupné z: http://www.fit.vutbr.cz/~rychly/public/docs/classification-and-prediction/classification-and-prediction.pdf
94. SANDERSON, M. a W. B. CROFT. The History of Information Retrieval Research. *Proceedings of the IEEE* [online]. 2012, vol. 100, Special Centennial Issue, s. 1444-1451 [cit. 2014-08-16]. DOI: 10.1109/JPROC.2012.2189916. Dostupné z: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6182576
95. SALTON, Gerard. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading, Mass.: Addison-Wesley, 1988, xiii, 530 p. ISBN 02-011-2227-8.

96. SALTON, Gerard a Christopher BUCKLEY. Term-weighting approaches in automatic text retrieval. *Information Processing*[online]. 1988, roč. 24, č. 5, s. 513-523 [cit. 2014-09-04]. DOI: 10.1016/0306-4573(88)90021-0. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/0306457388900210
97. SALTON, G. a M. E. LESK. The SMART automatic document retrieval systems---an illustration. *Communications of the ACM* [online]. 1965, vol. 8, issue 6, s. 391-398 [cit. 2014-10-01]. DOI: 10.1145/364955.364990. Dostupné z: http://portal.acm.org/citation.cfm?doid=364955.364990
98. SAMUELSON, P. Self-plagiarism or Fair Use?. *Communications of the ACM* [online]. 1994, roč. 8, č. 37, pp. 21-25 [cit. 2015-01-25]. Dostupné z: http://people.ischool.berkeley.edu/~pam/papers/SelfPlagiarism.pdf
99. SARACEVIC, T. Modeling Interaction in Information Retrieval (IR): A Review and Proposal.*Proceedings of the ASIS Annual Meeting* [online]. 1996, issue 33, s. 3-9 [cit. 2014-10-03]. Dostupné z: http://eric.ed.gov/?id=EJ557152
100. SARACEVIC, Tefko. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II. *Journal of the American Society for Information Science and Technology* [online]. 2007, roč. 58, č. 13, s. 1915-1933 [cit. 2014-09-04]. DOI: 10.1002/asi.20682. Dostupné z: http://doi.wiley.com/10.1002/asi.20682
101. SEBASTIANI, Fabrizio. On the role of logic in information retrieval. *Information Processing* [online]. 1998, roč. 34, č. 1, s. 1-18 [cit. 2014-09-05]. DOI: 10.1016/S0306-4573(97)00055-1. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0306457397000551
102. SOWA, John F. Mathematical Background. *C* [online]. 2007, c [cit. 2014-12-14]. Dostupné z: http://www.jfsowa.com/logic/math.htm#Lattice
103. SOPER, Herbert Edward. *Means for compiling tabular and statistical data* [patent]. USA. G06K21/04, US 1351692 A. Uděleno 31. srpen 1920. Dostupné z: http://www.google.com/patents/US1351692
104. TAYLOR, R. S. Question-Negotiation and Information Seeking in Libraries. *Question-Negotiation and Information Seeking in Libraries* [online]. 1968, č. 29, s. 178-194 [cit. 2014-03-27]. Dostupné z: https://www.ideals.illinois.edu/bitstream/handle/2142/38236/crl_29_03_178_opt.pdf?sequence=2

105. TAUBE, M., GULL, C. D. a WACHTEL, I. S. Unit terms in coordinate indexing. *American Documentation* [online]. 1952, vol. 3, issue 4, s. 213-218 [cit. 2014-08-17]. DOI: 10.1002/asi.5090030404. Dostupné z: http://doi.wiley.com/10.1002/asi.5090030404

106. TIAMIYU, M.A. The relationships between source use and work complexity, decision-maker discretion and activity duration in Nigerian government ministries. *International Journal of Information Management* [online]. 1992, vol. 12, issue 2, s. 130-141 [cit. 2014-10-05]. DOI: 10.1016/0268-4012(92)90019-M. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/026840129290019M

107. UTGOFF, Paul E. Incremental Induction of Decision Trees. *Machine Learning* [online]. 1989, roč. 4, č. 2, s. 161-186 [cit. 2014-10-01]. DOI: 10.1023/A:1022699900025. Dostupné z: http://link.springer.com/10.1023/A:1022699900025

108. VAN RIJSBERGEN, C. *Information retrieval*. 2. vyd. Boston: Butterworths, 1979, ix, 208 p. ISBN 04-087-0929-4.

109. VAN DE VEN, Andrew H a Diane L FERRY. *Measuring and assessing organizations*. New York: Wiley, 1980, xvii, 552 p. ISBN 04-710-4832-1.

110. VAKKARI, Pertti. Growth of theories on information seeking: An analysis of growth of a theoretical research program on the relation between task complexity and information seeking. *Information Processing* [online]. 1998, vol. 34, issue 2-3, s. 361-382 [cit. 2014-10-03]. DOI: 10.1016/S0306-4573(97)00074-5. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0306457397000745

111. VERHOEFF, J., W. GOFFMAN a Jack BELZER. Inefficiency of the use of Boolean functions for information retrieval systems. *Communications of the ACM* [online]. 1961, roč. 4, č. 12, s. 557-558 [cit. 2014-09-01]. DOI: 10.1145/366853.366861. Dostupné z: http://portal.acm.org/citation.cfm?doid=366853.366861

112. VOORHESS, E. M. a D. HARMAN. Overview of TREC 2001. *National Institute of Standart and Technology*. 2001.

113. WALTER, Chip. Kryder's Law. *Scientific American* [online]. 2005, vol. 293, issue 2, s. 32-33 [cit. 2014-08-16]. DOI: 10.1038/scientificamerican0805-

32. Dostupné z: http://www.nature.com/doifinder/10.1038/scientificamerican0805-32

114. WILSON, T. D. ON USER STUDIES AND INFORMATION NEEDS. *Journal of Documentation* [online]. 1981, vol. 37, issue 1, s. 3-15 [cit. 2014-10-03]. DOI: 10.1108/eb026702. Dostupné z: http://www.emeraldinsight.com/10.1108/eb026702

115. WILSON, T. D. Models in information behaviour research. *Journal of Documentation*. 1999, vol. 55, issue 3, s. 249-270.

116. WONG, S. K. M., W. ZHIARKO, V. V. RAGHAVAN a P.C.N. WONG. On Extending the Vector Space Model for Boolen Query Processing. *Proceedings of the 8th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*. 1985, s. 18-25 [cit. 2014-09-05].

117. YANNAKOUDAKIS, E. J., P. GOYAL a J. A. HUGGILL. The generation and use of text fragments for data compression.*Information Processing* [online]. 1982, roč. 18, č. 1, s. 15-21 [cit. 2014-09-04]. DOI: 10.1016/0306-4573(82)90047-4. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/0306457382900474

118. ZEFFANE, Rachid M. a Ferdinand A. GUL. The effects of task characteristics and sub-unit structure on dimensions of information processing. *Information Processing* [online]. 1993, vol. 29, issue 6, s. 703-719 [cit. 2014-10-04]. DOI: 10.1016/0306-4573(93)90100-R. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/030645739390100R

119. ZEIDMAN, Robert M. SOFTWARE ANALYSIS AND FORENSIC ENGINEERING CORP. *Software tool for detecting plagiarism in computer source code* [patent]. USA. US 8261237 B2, US 12/217,711. Uděleno 4. září 2012. Dostupné z: http://www.google.com/patents/US8261237

Seznam obrázků

Seznam tabulek