Západočeská univerzita v Plzni
Katedra informatiky a výpočetní techniky
Univerzitní 8
30614 Plzeň

Aplikace sémantických prostorů v úloze rozšiřování dotazu

Odborná práce ke státní doktorské zkoušce

Lubomír Krčmář

Technical Report No. DCSE/TR-2011-02
Květen, 2011

Technical Report No. DCSE/TR-2011-02
Květen 2011

Aplikace sémantických prostorů v úloze rozšiřování dotazu

Lubomír Krčmář

Abstrakt

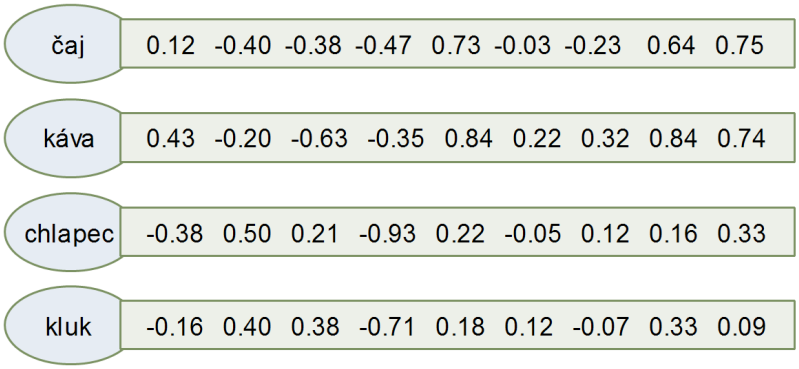
Práce je zaměřená na sémantické prostory a jejich využití ve vyhledávání informací v úloze rozšiřování dotazu. V práci jsou nejprve stručně popsány principy nejznámějších algoritmů pro vytváření sémantických prostorů, mezi které patří HAL a LSA. Popsány jsou rovněž způsoby testování a vyhodnocování sémantických prostorů. Na toto téma navazuje porovnání výsledků sémantických prostorů vytvářených různými algoritmy. Dále se práce věnuje problematice vyhledávání informací a používaným technikám v této oblasti, mezi které patří rozšiřování dotazu. V předkládaném textu lze nalézt některé zdokumentované dosavadní úspěchy aplikace sémantických prostorů v úloze rozšiřování dotazu.

Stěžejní součástí práce je pak popis výzkumných záměrů, které budou předmětem navazující disertační práce. Mezi tyto plány patří vytvoření systému pro vyhodnocování úspěšnosti nasazení sémantických prostorů v úloze rozšiřování dotazu. Dále je to zkoumání způsobů nasazení sémantických prostorů na tuto úlohu a aplikace algoritmů pro vytváření sémantických prostorů na český jazyk.

Copies of this report are available on
http://www.kiv.zcu.cz/publications/
or by surface mail on request sent to the following address:

University of West Bohemia in Pilsen
Department of Computer Science and Engineering
Univerzitni 8
30614 Pilsen
Czech Republic

Obsah



Obrázek 1: Ukázka sémantického prostoru znázorňující, že významově blízká slova (čaj a káva; chlapec a kluk) mají podobné vektory.

1 Úvod

Pojem sémantický prostor uvedl na svět Tannenbaum v [1]. Sémantický prostor je slovník, ve kterém je ke každému slovu[1] nebo konceptu přiřazen vektor čísel. Vektory čísel jsou pro všechna slova slovníku stejně dlouhé. Jejich délka je dimenzí (rozměrem) sémantického prostoru. Porovnáváním vektorů se určuje míra podobnosti slov, které vektorům odpovídají. Sémantické prostory se vytvářejí pomocí různých algoritmů využívajících statistické metody, které jsou zpravidla aplikované na velké textové korpusy. Hypotetický sémantický prostor je zobrazen na obrázku 1.

Sémantické prostory se začaly zkoumat v souvislosti s lingvistikou a v souvislosti se sémantickou pamětí, která je jednou z několika druhů pamětí člověka. Algoritmy vytvářející sémantické prostory jsou založené na Harrisově bučním teorému [2]. Harris již v roce 1968 publikoval tezi o významu slova z pohledu lingvistiky. Harris řekl, že slova jsou si podobná do té míry, do jaké sdílejí podobný lingvistický kontext.

Tato práce se nejprve zabývá sémantickými prostory v souvislosti se sémantickou pamětí, způsoby vytváření sémantických prostorů a možnostem vyhodnocování jejich kvality. V práci jsou také popsány základní pricipy vyhledávání informací v textových datech. Oblasti sémantické prostory a vyhledávání jsou propojeny úlohou automatického rozšiřování dotazu, která je hlavním tématem prezentovaného textu.

Následující kapitola se věnuje především souvislosti sémantických prostorů s pamětí. V kapitole lze také nalézt stručné popisy nejznámějších algoritmů pro tvoření sémantických prostorů. Kapitola 3 popisuje principy dvou vybraných algoritmů, kterými jsou LSA a HAL. Kapitola 4 se věnuje způsobům vyhodnocení kvality sémantických pro-

[1] V celém následujícím textu je pojem slovo zaměňován za nejmenší zpracovávanou jednotku textu často označovanou jako term. Term označuje kromě slov i interpunkci či čísla.

storů. V kapitole 5 lze nalézt základní principy vyhledávání informací a používané způsoby vyhodnocení vyhledávacích systémů. Kapitola 6 se zabývá úlohou, ve které se používají sémantické prostory. Touto úlohou je automatické rozšiřování dotazu. Předposlední kapitola 7 diskutuje směry zamýšleného výzkumu v navazující disertační práci. Poslední kapitola 8 práci shrnuje.

2 Sémantické prostory

Sémantické prostory úzce souvisí se sémantickou pamětí. Sémantickou paměť řadí psychologové mezi dlouhodobé deklarativní[2] paměti. V této paměti se ukládají fakta a obecné znalosti nezávisle na okolnostech, při kterých je člověk nabyl. Okolnostmi můžou být čas, místo nebo třeba nálada. Příkladem toho, co je v sémantické paměti uloženo může být: „Váza je nádoba, kam se dávají květiny.“ Doplňkovou pamětí k sémantické, je paměť episodická. Do episodické paměti jsou ukládány osobní prožitky spojené s výše uvedenými okolnostmi. Např.: „Včera se stala na dálnici D1 ve 12:10 nehoda.“ Více o sémantické a episodické paměti se lze dočíst v [3].

První sémantické prostory vznikaly právě jako modely sémantické paměti člověka. Psychologové přišli s několika modely popisujícími, jak by mohlo fungovat ukládání faktů a znalostí do sémantické paměti. Mezi modely, kterými se zabývá tato práce, patří sémantické sítě a modely statistické.

2.1 Sémantické sítě

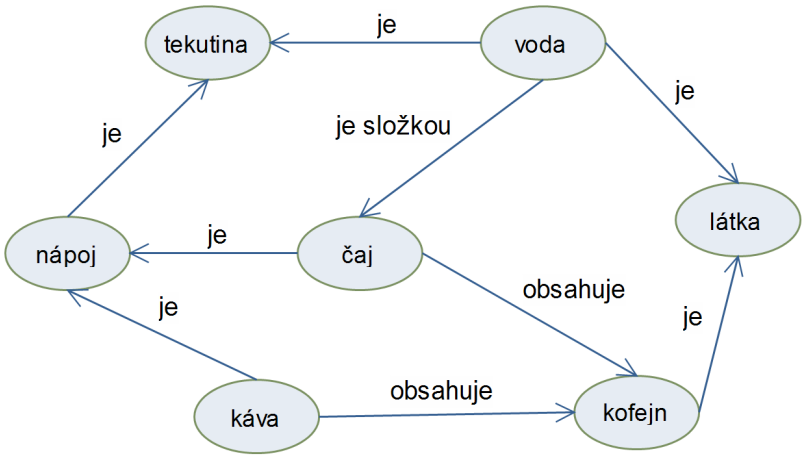
Sémantická síť je orientovaný nebo neorientovaný graf, ve kterém jsou znázorněny vztahy mezi koncepty (abstraktními pojmy). Koncepty jsou uzly grafu. Souvislosti mezi koncepty jsou vyjádřeny hranami grafu. Příklad sémantické sítě je znázorněn na obrázku 2. Sémantickými sítěmi se mimo jiné dlouho zabýval Sowa [4]. Nejznámější sémantickou sítí je dnes Wordnet [5] (pro češtinu: [6]).

Uzly sítě Wordnet představují tzv. synsety. Každý ze synsetů je tvořen skupinou synonym, které reprezentují stejný koncept. Dalšími vztahy mezi koncepty vyjádřenými ve Wordnetu jsou hyperonyma, hyponyma, holonyma a meronyma.

2.2 Přehled algoritmů pro vytváření sémantických prostorů

Zajímavými modely sémantické paměti jsou modely statistické. Tyto modely jsou založené na statistickém vyhodnocení četností a druhů kontextů jednotlivých slov konceptů (slov). Koncepty jsou si podobné, vyskytují-li se v podobných kontextech v po-

[2] deklarativní paměti se také říká explicitní



Obrázek 2: Ukázka sémantické sítě

dobném distribučním rozdělení[3]. Statistické modely sémantické paměti jsou sémantickými prostory. Sémantické prostory je možné vytvářet mnoha různými algoritmy.

Mezi první badatele zabývající se sémantickými prostory patří Lund a Burgess [7]. Tito zkoumali kontexty slov velkého textového korpusu[4]. K tomuto účelu použili algoritmus, který nazvali HAL[5]. Výsledky pokusů potvrdily množství informace o významu konceptů, které poskytují jejich kontexty. Lundovi a Burgessovi se pomocí HAL podařilo zachytit významy konceptů a vztahy mezi nimi automatickou analýzou textu [7]. Princip algoritmu HAL je popsán v odstavci 3.1.

Jedni z prvních výzkumníků, kteří se zabývali reprezentací konceptů v mnohorozměrném prostoru, jsou také Landauer a Dumais [8]. Ti navrhli algoritmus LSA[6]. Pomocí LSA vytvořili sémantický prostor, který úspěšně aplikovali na klasickou úlohu výběru jednoho synonyma ze čtyř možností k danému slovu. Tato úloha je popsána v oddílu 4.1. Princip algoritmu LSA je popsán v odstavci 3.2..

Mezi další algoritmy, pomocí nichž lze vytvářet sémantické prostory, patří také COALS[7], RI[8] nebo BEAGLE[9]. Kromě těchto vznikají i modifikace starších algoritmů jako jsou pLSA nebo pHAL. Zajímavým algoritmem je také LRA[10] popsaný v [9]. Algoritmus LRA však již neslouží k porovnávání souvislosti či podobnosti konceptů,

[3] jinak interpretovaný Harrisův teorém [2]
[4] korpus článků USENET - 160 miliónů anglických slovních tvarů
[5] Hyperspace Analogue to Language
[6] Latent Semantic Analysis
[7] Correlated Occurrence Analogue to Lexical Semantic
[8] Random Indexing
[9] Bound Encoding of the Aggregate Language Environment
[10] Latent Relational Analysis

ale k porovnávání toho, do jaké míry jsou si podobné vztahy mezi koncepty.

Algoritmus COALS je popsaný v [10]. Z článku vyplývá, že COALS má společné základy s algoritmem HAL. Navíc také využívá principu redukce dimenze pomocí SVD jako LSA. Zajímavou myšlenkou COALS je využití korelace pro transformaci matice získané ze spoluvýskytů slov ve vstupním korpusu dat. V [10] jsou prezentovány velmi dobré výsledky COALS v porovnání s ostatními algoritmy. Výsledky COALS jsou rovněž diskutovány v kapitole 4 v této práci.

Zajímavým algoritmem je také algoritmus RI, jehož princip je vysvětlený v [11]. V článku se píše, že algoritmus RI dosahuje podobných výsledků jako LSA s tím, že není zdaleka tak výpočetně náročný. Základní myšlenkou RI je přiřazení náhodných vektorů předem dané délky všem slovům korpusu před výpočtem. Následuje vypočítávání kontextů slov podobně jako v metodě HAL. Na rozdíl od HAL jsou sémantické vektory slov tvořeny součty předem vytvořených náhodných vektorů slov, které se vyskytují v jejich kontextech.

Vytváření náhodných vektorů pro slova využívá také poměrně nový algoritmus BEAGLE popsaný v [12]. Tento algoritmus kombinuje principy LSA, HAL i RI. Autoři v článku tvrdí, že pro vyšetřování významů slov jsou podstatné nejen jejich kontexty, ale i uspořádání, ve kterém se vyskytují. V článku je popsán algoritmus spojující tyto dva druhy informací do sémantických vektorů slov. Podstatná odlišnost BEAGLE od HAL je, mimo jiné, zpracovávaná jednotka textu. Tou není v BEAGLE pevně dané okénko okolo slova, ale věta.

Algoritmy pro vytváření sémantických prostorů stále vznikají. Často jsou přebírány principy z algoritmů starších. Příkladem takových algoritmů jsou pravděpodobnostní LSA a HAL, zkráceně pLSA a pHAL. Tyto algoritmy, jak název napovídá, využívají pravděpodobnost. Algoritmus pLSA na rozdíl od LSA se tak vyhýbá použití výpočetně náročné matematické operaci SVD. Algoritmu pLSA se věnuje [13], o algoritmu pHAL se lze dočíst v [14].

3 Vytváření sémantických prostorů

V této kapitole jsou popsány principy prvních algoritmů pro vytváření sémantických prostorů, kterými jsou LSA a HAL. Jedním ze zásadních rozdílů mezi LSA a HAL je to, že LSA na rozdíl od HAL nebere v úvahu pořadí slov ve větách. LSA zpracovává dokumenty[11] jako množiny slov. LSA také na rozdíl od HAL využívá náročnou matematickou operaci SVD[12]. Výstupem obou algoritmů jsou sémantické prostory konstruované velmi odlišným způsobem.

[11] nebo menší jednotku textových dat - například odstavec

[12] Singular Value Decomposition

Tabulka 1: Příklad matice vytvořené algoritmem HAL pro větu: „Někdo pije kávu a někdo raději čaj.“ při zvolené šířce okénka 5.

	raději	čaj	pije	a	kávu	někdo
.	4	5	0	2	1	3
raději	0	0	2	4	3	6
čaj	5	0	1	3	2	4
pije	0	0	0	0	0	5
a	0	0	4	0	5	3
kávu	0	0	5	0	0	4
někdo	0	0	3	5	4	2

3.1 Princip algoritmu HAL

Vstupním parametrem algoritmu HAL je velikost okénka n. Po spuštění algoritmu HAL dojde nejprve k vytvoření nulové matice, jejíž každý řádek i sloupec odpovídá různému slovu zpracovávaného textového korpusu.

Následně je HALem procházen korpus slovo po slově. Pro každé slovo x v korpusu je vyhledán jeho řádkový vektor v matici. Po vyhledání vektoru je zvětšeno n jeho hodnot, které odpovídají slovům před slovem x ve zpracovávaném místě korpusu do vzdálenosti n. Tyto hodnoty jsou zvětšovány o lineárně sestupnou hodnotu závislou na vzdálenosti od slova x. Nachází-li se slovo y bezprostředně před slovem x, dojde ke zvětšení hodnoty v řádkovém vektoru slova x odpovídající slovu y o n. Nachází-li se slovo y o 2 pozice před slovem x, dojde ke zvětšení hodnoty v řádkovém vektoru slova x odpovídající slovu y o (n - 1), atd. Takto jsou algoritmem HAL zaznamenávány informace o kontextu právě zpracovávaného výskytu slova.

Výstupem algoritmu HAL je „matice spoluvýskytů slov“. V této matici je v každém řádku uložen sémantický vektor pro jedno slovo, který je tvořen všemi kontexty získanými výše uvedeným postupem pro všechny výskyty tohoto slova. Ukázková matice pro korpus obsahující jen jeden dokument s jednou větou „Někdo pije kávu a někdo raději čaj.“ je k nahlédnutí v tabulce 1. Tento příklad je převzatý[13] z [7]. Princip algoritmu HAL je také vysvětlený v [10].

Přestože je výstup algoritmu HAL („matice spoluvýskytů slov“) sémantickým prostorem, často se používá sémantický prostor, který lze z této matice přímo získat. V „matici spoluvýskytů“ je totiž každý řádek slova x popsán jen hodnotami, které vytvářejí slova předcházející slovu x. Z hodnot ve sloupcích ovšem vyplývají i hodnoty pro slova, které za slovem x následují. Používaný typ sémantického prostoru získávaného z „matice spoluvýskytů“ zobrazuje tabulka 2.

[13] přesněji: v [7] se nachází anglická verze příkladu

Tabulka 2: Používaný typ sémantického prostoru získaného z „matice spoluvýskytů“ vytvořeného algoritmem HAL pro větu: „Někdo pije kávu a někdo raději čaj.“ při zvolené šířce okénka 5.

	raději	čaj	pije	a	kávu	někdo	.	raději	čaj	pije	a	kávu	někdo
.	4	5	0	2	1	3	0	0	0	0	0	0	0
raději	0	0	2	4	3	6	4	0	5	0	0	0	0
čaj	5	0	1	3	2	4	5	0	0	0	0	0	0
pije	0	0	0	0	0	5	0	2	1	0	4	5	3
a	0	0	4	0	5	3	2	4	3	0	0	0	5
kávu	0	0	5	0	0	4	1	3	2	0	5	0	4
někdo	0	0	3	5	4	2	3	6	4	5	3	4	2

Existují různé modifikace algoritmu HAL. Všechny sloupce v sémantickém prostoru vytvářeného pomocí HAL nenesou stejné množství informace. Sloupce lze seřadit podle množství informace, které nesou, pomocí entropie[14]. Proto je možné nejméně důležité sloupce odstranit a přinejmenším tak zredukovat velikost sémantického prostoru. Dalšími možnými modifikacemi jsou změna velikosti okénka nebo změna vah slov v okénku[15].

3.2 Princip algoritmu LSA

Základní myšlenkou algoritmu LSA je promítání slov a konceptů do obecnějších (možná lépe řečeno nadřazenějších) konceptů, latentních prostorů či témat. Matematická operace, která toto promítání umožňuje, se nazývá singulární rozklad (dále SVD). Pomocí SVD dochází k redukci dimenze dat.

Algoritmus LSA začíná vytvořením matice spoluvýskytů[16], jejíž řádky odpovídají různým slovům v korpusu a sloupce dokumentům korpusu. Pro LSA je uspořádání slov v dokumentech nepodstatné. Matice jednoduše obsahuje počty výskytů slovních tvarů v jednotlivých dokumentech. Po vytvoření matice spoluvýskytů dojde k její normalizaci. Pro normalizaci se používají techniky jako je TFIDF[17] popsané například v [15].

Dalším krokem algoritmu je rozklad normalizované matice spoluvýskytů pomocí SVD. Výsledkem SVD rozkladu jsou tři matice, jejichž součinem je matice původní. První matice ze tří reprezentuje, jak jsou slova (řádky) obsažena v jednotlivých latentních prostorech (sloupce). Druhá matice je diagonální a obsahuje seřazená singulární

[14] Více o entropii se lze dozvědět v [16]

[15] změna důležitosti sousedících slov vzhledem k právě zkoumanému slovu

[16] narozdíl od HAL není v LSA vytvářena matice slova na slova, ale slova na dokumenty

[17] term frequency, inverse document frequency

čísla reprezentující míru zastoupení jednotlivých témat v celém korpusu. Třetí matice říká, do jaké míry dokumenty pojednávají o latentních prostorech.

Získané matice jsou redukovány nastavením všech singulárních čísel, kromě prvních[18] k, na hodnotu 0. Číslo k bývá nastaveno na hodnotu 300, která se ukazuje jako optimální [8]. Číslo k často bývá vstupním parametrem implementovaného algoritmu LSA. Po redukci matic dojde k jejich pronásobení. Součinem je již matice výsledná, která je sémantickým prostorem pro slovní tvary korpusu[19].

Princip algoritmu LSA nastiňuje obrázek 3. Obrázek znázorňuje vytvoření sémantických vektorů pro slova „člověk“ a „uživatel“. Nejprve je vytvořena matice slova na dokumenty. Na obrázku 3 jsou z této matice zobrazeny vektory pro zkoumaná slova. Je vidět, že slova se nevyskytují v žádném dokumentu společně. Po aplikaci SVD rozkladu a po redukci dimenze jsou vytvořeny sémantické vektory pro zkoumaná slova. Tyto vektory jsou si velmi podobné. Důvodem je fakt, že zkoumaná slova se vyskytovala v podobných kontextech. Obrázek znázorňuje příklad, který je převzat z [8], kde je podrobněji vysvětlen.

4 Vyhodnocování sémantických prostorů

Vytváření sémantických prostorů je obtížné. Mezi překážky patří „řídkost dat“. Řídkostí dat je myšleno to, že slova se v žádném přirozeném korpusu nevyskytují se všemi slovy, se kterými mohou souviset. Dalšími možnými problémy jsou nekvalitní korpusy (více informací o kvalitě korpusů lze nalézt v [17]) nebo korpusy zaměřené na určitou doménu. Proto také vyhodnocování sémantických prostorů není jednoduchou úlohou. Tato kapitola ukazuje, jak se sémantické prostory testují a některé výsledky těchto testů.

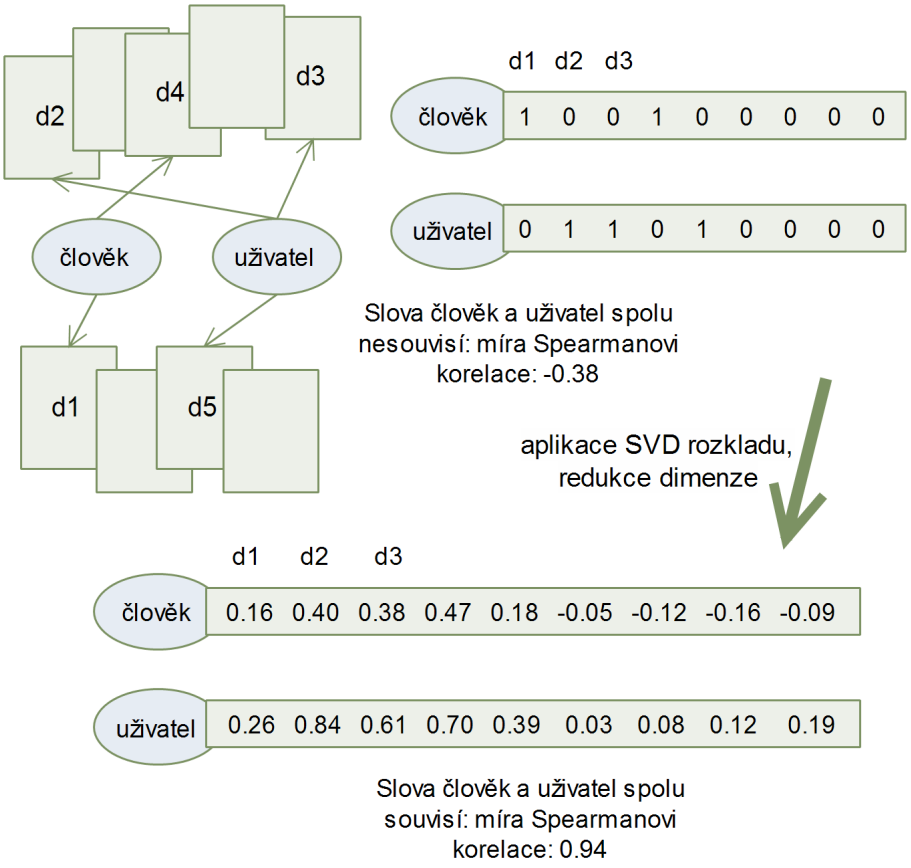
4.1 Způsoby vyhodnocování

První vytvořené sémantické prostory se testovaly na to, jestli vůbec nesou sémantickou informaci o slovech. Přítomnost sémantické informace ukázali již Lund a Burgess při testování algoritmu HAL [7]. Použili k tomu 3 různé experimenty.

Prvním experimentem je hledání nejpodobnějších slov k danému slovu v sémantickém prostoru. Lund a Burgess ve svém experimentu vybrali 20 náhodných slov s přibližně průměrnou frekvencí výskytů v korpusu. Měřením euklidovských vzdáleností vektorů těchto 20 slov se všemi ostatními vektory sémantického prostoru a následným porovnáváním těchto vzdáleností byli získáni „nejbližší sousedé“ ke vstupním 20 slovům.

[18] v diagonální matici jsou singulární čísla seřazena podle důležitosti

[19] násobení 3. maticí pro zjišťování vztahů mezi slovy není nutné, tento postup je v příkladu použitý, aby bylo možné určovat i vztahy mezi dokumenty



Obrázek 3: Princip algoritmu LSA

Nalezená nejbližší slova byla ke každému odpovídajícímu původnímu slovu významově i asociativně blízká.

Podstatou druhého experimentu byla projekce vektorů slov s velkou dimenzí do plochy[20]. Projekce slov byla vykonána na slovech patřících do jedné ze tří kategorií. První kategorií byly názvy zvířat, druhou části lidského těla a třetí zeměpisná místa. Ve výsledné projekci byly projekcí vytvořeny tři shluky, které odpovídaly jednotlivým kategoriím.

Třetí experiment souvisel s reakční dobou dobrovolníků, kteří měli určit, zda spolu páry slov, které jim byly ukazovány, souvisí. Výsledky pokusů potvrdily očekávanou korelaci mezi reakční dobou dobrovolníků a mírou podobnosti slov v sémantickém prostoru.

Další způsoby vyhodnocování se zaměřují na množství informace, které sémantické prostory nesou, nebo na to, do jaké míry informace v sémantických prostorech koreluje s informacemi v sémantické paměti člověka. Landauer a Daumais ve svém článku o LSA [8] testují vytvořený sémantický prostor pomocí TOEFL[21] testu. TOEFL test skládá většina studentů hlásících se na univerzity, kde se hovoří anglicky. Landauer a Daumais pomocí LSA vytvořili takový sémantický prostor, pomocí kterého byl TOEFL test správně vyplněn ze 64%. Tato hodnota přibližně odpovídá průměrnému skóre studentů, kteří TOEFL test skládali.

TOEFL test spočíval[22] ve výběru nejlepšího synonyma k danému slovu ze 4 možností. Výběr nejbližšího slova ze 4 nepředstavuje v sémantickém prostoru žádný problém. Stačí porovnat vektory těchto 4 slov se slovem zadaným. Stejný TOEFL test a jemu podobné (ESL[23] nebo RDWP[24]) se dnes stále používají k vyhodnocování sémantických prostorů.

Populární způsobem pro vyhodnocování sémantických prostorů se později stalo použití výsledků Rubenstein-Goodenough testu (dále RG test). RG test je popsaný v [18]. RG test se skládá z 65 párů slov. Slova v těchto párech se opakují a jsou běžná v anglickém jazyce. Těchto 65 párů bylo ohodnoceno 21 respondenty, kteří určovali, do jaké míry spolu slova v párech souvisí. Míra podobnosti byla vyjadřována hodnotou 0 až 4[25]. Výsledky byly statisticky zpracovány a finálním produktem se stal seřazený seznam párů slov s číselnými hodnotami, které vyjadřují míru souvislosti slov těchto párů.

Princip testování sémantických prostorů pomocí RG testu spočívá v určení míry korelace mezi lidským úsudkem a vytvořeným sémantickým prostorem. Určuje se korelace

[20] plocha - 2 dimenze
[21] Test of English as a Foreign Language
[22] dnes se v TOEFL testu testují jiné dovednosti
[23] English as a Second Language
[24] Reader's Digest Word Power
[25] 0 - nejméně souvisí, 4 - nejvíce souvisí

mezi hodnotami z RG testu a hodnotami, které jsou získávány měřením vzdáleností vektorů odpovídajících slov v sémantickém prostoru. Na stejném principu jsou založeny další používané testy, jako jsou Miller & Charles test [19] nebo WordSim353 test [20].

Další způsoby vyhodnocování sémantických prostorů lze nalézt v [17]. Mezi ně patří test syntaktické kategorizace. Sémantický prostor v tomto testu uspěje, jsou-li k předkládaným slovům v sémantickém prostoru nacházena jako nejpodobnější ta slova, která patří do stejné syntaktické kategorie. Dalším testem, který je v článku zmíněn, je test založený na porovnávání vzdáleností. Tento test je podobný TOEFL testu. Sémantický prostor se testuje na to, jestli vyhodnotí jako nejpodobnější slovo k předkládanému určité slovo před dalšími náhodně vybranými slovy ze specifikované množiny.

Sémantické prostory se také testují na úspěšnost jejich nasazení na úlohy, ve kterých jsou užitečné. Takovou úlohou může být například rozšiřování dotazu. O sémantických prostorech a jejich využití v oblasti rozšiřování dotazu pojednává oddíl 6.1.

4.2 Výsledky vyhodnocování

Tento oddíl stručně diskutuje výsledky vyhodnocování některých sémantických prostorů, které byly vytvářeny pomocí algoritmů popsaných v oddíle 2.2. Vyhodnocování bylo prováděno pomocí testů popsaných v oddíle 4.1. Výsledky testů jsou převzaty z [10] a [21].

Tabulka 3 znázorňuje nejlepší výsledky, kterých dosáhly algoritmy LSA, HAL a COALS s různým nastavením. Algoritmy byly trénovány na korpusu USENET[26]. Tyto výsledky publikoval Rohde v článku [10], ve kterém je představen a vyhodnocován nový algoritmus COALS. Výsledky ukazují, že algoritmus COALS překonává algoritmy HAL a LSA ve všech testech. U TOEFL testu je algoritmus COALS dokonce o více než 30% lepší než druhý nejlepší HAL.

Z tabulky 3 je také vidět, že algoritmus HAL v jednotlivých testech příliš úspěšný není. Zatímco jeho konkurenti podávají v testech poměrně stabilní výkony, algoritmus HAL za nimi velmi ztrácí v testu ESL (26% je velmi špatný výsledek) i v testu R-G. Naopak algoritmus LSA dosahuje dobrých výsledků, ve většině testů jen o málo horších než COALS. V TOEFL testu LSA dosahuje výkonu 53,4%. Tato hodnota je přibližně o 10% horší než ta, které dosáhl Landauer [8].

Tabulka 4 pro srovnání ukazuje výsledky, kterých dosáhl Jurgens [21]. Článek [21] byl publikován později než [10] od Rohde. V tabulce 4 lze nalézt i výsledky pro jiné algoritmy, kterými jsou BEAGLE nebo RI. Jurgens používal při vytváření sémantických prostorů korpusy TASA[27] a Wiki[28].

[26] přibližně 120 mil. příspěvků stažených ze systému elektronických diskusních skupin
[27] 44, 486 tématicky zaměřených esejí
[28] 387, 082 dokumentů obsahujících kolem 917 mil. slov

Tabulka 3: Výsledky testů sémantických prostorů publikované v [10]. Číslo za názvem algoritmu udává počet dimenzí výsledného sémantického prostoru. Obsahuje-li název zkratku SVD, je redukce dimenzí prováděna pomocí SVD.

Algoritmus	korpus	TOEFL	ESL	RDWP	R-G	WordSim353
COALS-14K	USENET	86.2	52.0	65.5	68.2	62.6
COALS-SVD-800	USENET	88.8	68.0	66.8	67.3	65.7
COALS-SVD-200	USENET	86.2	58.0	60.8	64.7	65.3
HAL-14K	USENET	56.2	26.0	37.9	14.6	28.2
HAL-400	USENET	53.8	26.0	35.7	15.3	31.1
LSA	USENET	53.4	43.0	40.6	65.6	59.9

Tabulka 4: Výsledky testů sémantických prostorů publikované v článku [21].

Algoritmus	Korpus	TOEFL	ESL	RDWP	R-G	WordSim353
BEAGLE	TASA	46.03	35.56	46.99	0.431	0.342
COALS	TASA	65.33	60.42	93.02	0.572	0.478
HAL	TASA	44.00	20.83	50.00	0.173	0.180
HAL	Wiki	50.00	31.11	43.44	0.261	0.195
LSA	TASA	56.00	50.00	45.83	0.652	0.519
LSA	Wiki	60.76	54.17	59.20	0.681	0.614
RI	TASA	42.67	27.08	34.88	0.224	0.201
RI	Wiki	68.35	31.25	40.80	0.226	0.315

Z tabulky 4 lze také vyčíst závislost výsledků sémantických prostorů na použitém korpusu dat. Všechny algoritmy (HAL, LSA, RI) dosahují lepších výsledků, jsou-li trénované na větším korpusu Wiki. Bohužel v [21] nejsou publikovány výsledky algoritmů BEAGLE a COALS trénovaných na tomto korpusu.

Výsledky v tabulkách 3 a 4 ukazují, že nejlepším algoritmem pro vytváření sémantických prostorů by mohl být COALS. Dobrých výsledků dosahuje také algoritmus LSA. Naopak všechny ostatní algoritmy: HAL, BEAGLE i RI, příliš úspěšné nejsou. Nutno však dodat, že výkony algoritmů závisí na jejich nastavení, na korpusu, na kterém jsou trénovány, ale i na způsobu předzpracování korpusů [21]. Proto jsou i výsledky ve výše diskutovaných tabulkách různé. Zatímco Rohde [10] hledal optimální nastavení algoritmu COALS, Jurgens [21] pravděpodobně použil standardní nastavení algoritmů, které převzal z dřívějších prací věnujících se sémantickým prostorům. Motivací této myšlenky je fakt, že Jurgens v [21] použité nastavení algoritmů nepublikuje.

5 Vyhledávání informací

Problém s vyhledáváním informací vyvstal s možností jejich ukládání. Dnes se informace ukládají a publikují velmi snadno prostřednictvím internetu. Jak však efektivně zařídit, aby ten, kdo informace potřebuje, je na internetu snadno dohledal, je stále řešený problém.

Následující oddíl 5.1 popisuje základní principy a modely používané ve vyhledávání v čistě textových netříděných datech. Oddíl 5.2 se věnuje způsobům vyhodnocení efektivity vyhledávacích systémů. Zmíněné techniky a principy jsou podrobněji popsány v [22].

5.1 Základní principy vyhledávání

Základním používaným principem, na kterém jsou založeny vyhledávací systémy, je vytvoření a udržování invertovaného indexu. Ukázkovým invertovaným indexem je rejstřík v knize, kde jsou uložena klíčová slova a stránky, které se o nich zmiňují. Na stejném principu jsou založeny vyhledávací systémy, pomocí nichž je možné hledat informace v korpusu dat nebo na internetu. Vyhledávací portály, jakými jsou například Google[29] nebo Altavista[30] vytvářejí obrovské invertované indexy.

Při vytváření invertovaných indexů se využívají techniky, jako jsou odstranění stopslov nebo lemmatizace. Odstranění stopslov spočívá ve vynechání slov při vytváření invertovaného indexu, která jsou velmi běžná v daném jazyce nebo mají ve větách stavební funkci. Tato slova se pak neberou v úvahu i během vyhledávání. Mezi stopslova jsou řazeny například spojky, předložky a další neohebná slova. V anglickém jazyce jsou mezi stopslova řazena především členy „the“ a „a, an“.

Technika lemmatizace spočívá v převodu slova na jeho základní tvar. Pro český jazyk, ve kterém existuje pro většinu slov mnoho tvarů, je lemmatizace klíčovou technikou. Využívá jí například známý český vyhledávač Seznam.cz[31]. Pomocí lemmatizace je možné vyhledávat všechny možné slovní tvary daného slova, a přinejmenším tak zvětšit objem nalezených dat.

Samotné vyhledávání většinou spočívá v zadání uživatelského dotazu a odpovědi vyhledávacího systému ve formě seřazeného seznamu relevantních odpovědí. Jakým způsobem vyhledávací systém vybírá relevantní dokumenty, je určeno především vyhledávacím modelem, který systém používá. Nejznámějším a nejjednodušším je model booleovský. Za relevantní dokumenty v booleovském modelu jsou považovány ty, které obsahují všechna slova dotazu. Samotný booleovský model neumožňuje řadit nalezené relevantní dokumenty. Tento nedostatek překonává model vektorový.

[29] www.google.com
[30] www.altavista.com
[31] www.seznam.cz

Podstatou vektorového modelu je rozšíření invertovaného indexu o váhy slov. Do indexu mohou být jednoduše ukládány počty výskytů slov v daných dokumentech. Většinou jsou však tyto počty ještě vynásobeny číslem, které udává „důležitost slova“ v daném korpusu. Za důležitá slova jsou považována ta, co se vyskytují v málo dokumentech. Takto vytvářený vektorový model využívá techniky TFIDF.

Existuje více alternativ boolovského i vektorového modelu i další modely, mezi které patří například model pravděpodobnostní popsaný v [23]. V poslední době se stále více používá model jazykový. Základním principem tohoto vyhledávacího modelu je vytvoření jazykového modelu pro každý dokument korpusu. Následně je zkoumána pravděpodobnost vygenerování uživatelského dotazu jednotlivými jazykovými modely. Dokumenty v odpovědi jsou nakonec řazeny podle toho, s jakou pravděpodobností jejich jazykový model generuje uživatelský dotaz. Studii jazykového modelu pro vyhledávání lze nalézt v [24].

Mezi další techniky používané ve vyhledávání patří ohodnocování důležitosti webových stránek pomocí algoritmu PageRank. Tento algoritmus tvoří základ vyhledávacího systému Google. Internetový vyhledávací systém tak bývá tvořen 3 komponentami. První komponentou je invertovaný index, druhou vyhledávací model a třetí algoritmus PageRank nebo jeho obdoba.

Řešenými problémy v oblasti vyhledávání jsou problém synonymních výrazů a problém mnohoznačnosti slov. Problém synonymních výrazů spočívá v tom, že stejnou skutečnost, událost či fakt je možné popsat různými slovy se stejným významem. Často tak dochází k tomu, že není nalezen relevantní dokument k dotazu, který obsahuje jiná slova než ta, která uživatel do dotazu zadá. Využití základních modelů pro vyhledávání tento problém označovaný jako „informační mišmaš“ neřeší. Problém mnohoznačnosti slov spočívá v tom, že některá slova mají více významů, přestože se píší stejně.

Možným způsobem řešení problému mnohoznačnosti slov v diskutované problematice je určení správného významu slova z dalších slov v dotazu. Problém synonymních výrazů ve vyhledávání lze řešit rozšiřováním dotazu. Rozšiřování dotazu je technika, která spočívá ve vytváření alternativních dotazů k původnímu dotazu. Alternativní dotazy jsou vytvářeny ze slov, která souvisí se slovy v původním dotazu. Problematice rozšiřování dotazu se věnuje kapitola 6.

5.2 Vyhodnocování úspěšnosti vyhledávacích systémů

Aby bylo možné porovnávat úspěšnost vyhledávacích systémů, je nutné mít data, na kterých budou testovány, a stanovit metriky pro vyhodnocování těchto systémů. Tento oddíl zmiňuje data spojená s konferencí TREC[32]. Dále také nastiňuje, co je to efektivita vyhledávacího systému a důležité metriky, které slouží k měření této efektivity.

[32]Text Retrieval Conference

Tabulka 5: Tabulka znázorňující rozdělení dokumentů kolekce do 4 skupin po jejich vyhodnocení vyhledávacím systémem z hlediska relevance k dotazu.

	relevantní	**nerelevantní**
vyhledané	správně vyhledané (sv)	špatně vyhledané (šv)
nevyhledané	špatně nevyhledané (šn)	správně nevyhledané (sn)

TREC je každoroční konference zaměřená na oblast vyhledávaní informací. Hlavním cílem konference TREC je porovnání a vyhodnocování vyhledávacích systémů. Vyhodnocování vyhledávacích systémů umožňují ručně označkovaná data. Kolekce, které jsou součástí TREC, obsahují 3 typy dat. Prvním typem jsou dokumenty. Druhým typem dat jsou dotazy na dokumenty. Třetím a zásadním typem dat jsou soubory, které obsahují ke každému dotazu ručně vytvořený seznam dokumentů, které jsou k dotazu relevantní. Kolekce TREC tak umožňují vyhodnocování efektivity vyhledávacích systému pomocí dále uvedených metrik.

Efektivní vyhledávací systém je takový, který na uživatelův dotaz odpovídá relevantními dokumenty k dotazu a to nejlépe všemi. Neefektivní vyhledávací systém, naopak, na dotaz odpovídá nerelevantními dokumenty nebo jen zlomkem relevantních dokumentů. Základními metrikami, které berou v potaz tyto uživatelské nároky, jsou přesnost a úplnost.

Původ pojmů přesnost a úplnost plyne z principu činnosti vyhledávacích systémů. Vyhledávací systémy rozdělují dokumenty kolekce na dokumenty, které jsou podle nich relevantní k dotazu a na dokumenty, které jsou podle nich k dotazu nerelevantní. Dokumenty kolekce je tak možné, po jejich vyhodnocení vyhledávacím systémem, z hlediska relevance k dotazu rozdělit do 4 skupin, které znázorňuje tabulka 5.

Efektivní vyhledávací systém odpovídá na položený dotaz s velkou přesností P, kterou je podíl správně vyhledaných dokumentů ku všem vyhledaným dokumentům.

$$P = sv/(sv + šv)$$

Efektivní vyhledávací systém také odpovídá s velkou úplností U, kterou je podíl správně vyhledaných dokumentů ku všem relevantním dokumentům.

$$U = sv/(sv + šn)$$

Protože 100% úplnosti lze dosáhnout výběrem všech dokumentů kolekce a protože velké přesnosti lze dosáhnout pečlivým výběrem jen malého množství dokumentů, byly zavedeny další metriky, které tento problém odstraňují. Používanou metrikou v oblasti vyhledávání informací se stala F-míra. F-míra je váženým harmonickým průměrem přesnosti a úplnosti. Základní vzorec pro výpočet F-míry neupřednostňuje ani

důležitost přesnosti ani úplnosti: F-míra s parametrem $\beta = 1$.

$$F_{\beta=1} = \frac{2PR}{P+R}$$

Metriky přesnost, úplnost a F-míra jsou základní. Používají se na neseřazené množiny dokumentů. Aby bylo možné ohodnotit i vyhledávací systémy, které řadí dokumenty podle relevance k dotazu, byly vymyšleny další metriky. Mezi ně patří měření přesnosti v 11 bodech, které představují stupeň úplnosti. Stupně úplnosti jsou hodnoty úplnosti $U = 0.0; 0, 1..$ až $U = 1.0$.

Zajímavou a uznávanou metrikou je také MAP[33]. MAP, stejně jako měření přesnosti v 11 bodech, předpokládá, že množina dokumentů, která je odpovědí vyhledávacího systému na dotaz, je seřazená. MAP se počítá jako průměr z průměrných přesností vyhledávacího systému k jednotlivým dotazům $q_j \in Q$. Průměrná přesnost vyhledávacího systému k dotazu q_j se počítá jako průměr z přesností, které jsou vypočítávány po každém nalezeném relevantním dokumentu z množiny všech relevantních dokumentů $\{d_1, ...d_{m_j}\}$ k dotazu q_j. Je-li nalezen k-tý relevantní dokument, je dílčí přesnost k dotazu q_j počítána z množiny o velikosti R_{jk}.

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Přesnost(R_{jk})$$

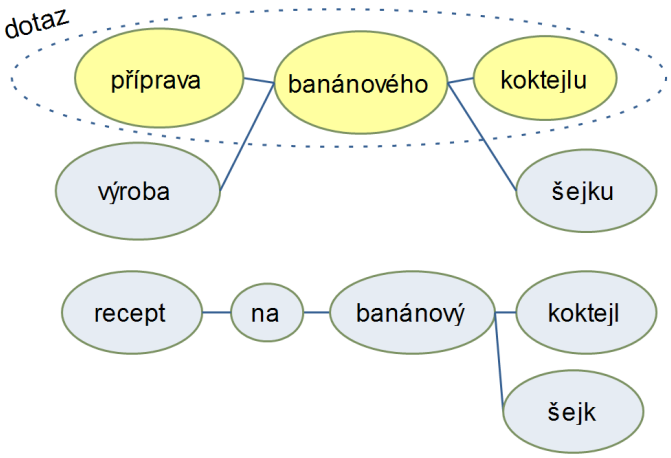
6 Automatické rozšiřování dotazu

Tato kapitola popisuje techniku rozšiřování dotazu. Rozšiřování dotazu se používá v oblasti vyhledávání informací. Cílem rozšiřování dotazu je obejít skutečnost, že různé věci se dají vyjádřit různými způsoby. Rozšiřování dotazu může probíhat automaticky na straně vyhledávacího systému a spočívá v modifikaci dotazu za účelem zvýšení přesnosti a úplnosti odpovědi.

Modifikace dotazu probíhá formou jeho rozšíření o alternativy. Alternativy původního dotazu jsou dotazy, které obsahují jiná slova se stejným nebo podobným významem (nebo slova, která se slovy původního dotazu souvisí). Ukázkový dotaz s jeho alternativami znázorňuje obrázek 4.

Způsobů, jak vytvářet alternativní dotazy, existuje více. Následující oddíl 6.1 je stručně popisuje. Navazující oddíl 6.2 pojednává o experimentech, které tyto způsoby aplikují.

[33] Mean avarage precision



Obrázek 4: Ukázka alternativních dotazů k dotazu „příprava banánového koktejlu“

6.1 Techniky pro rozšiřování dotazu

Možným způsobem rozšiřování dotazu je využití zpětné vazby. Princip využití zpětné vazby spočívá ve výběru slov, o která bude dotaz rozšířen, z dokumentů, kterými vyhledávací systém odpověděl na původní dotaz. Vybírají se ta slova, která jsou pro tyto dokumenty typická. Typická slova jsou taková, která se v těchto dokumentech vyskytují častěji než v ostatních a nejsou stopslovy. K výběru těchto slov je možné použít například TFIDF. Více informací o zpětné vazbě v oblasti rozšiřování dotazu se lze dočíst v [25].

Dalším způsobem pro výběr slov alternativních dotazů je využití manuálně vytvořených sítí slov nebo různých tezaurů, ze kterých lze určovat vztahy mezi slovy. Často využívanou sítí slov je anglický Wordnet [5]. Využitím Wordnetu v úloze rozšiřování dotazu se zabývá např. [26].

Kromě manuálně vytvářených sítí slov, lze využít i vztahy mezi slovy získávanými automaticky. Je možné využít sémantické prostory. Zajímavým článkem, který motivuje využití sémantických prostorů ve vyhledávání, je článek Google and the Mind [27].

Autoři [27] diskutují analogii lidské paměti a vyhledávacích systémů. Lidská paměť a vyhledávací systémy podle autorů čelí stejným úkolům. Na dotazy vyhledávají relevantní odpovědi, které jsou uložené ve velké síti vzájemně propojených uzlů, kde jsou informace uloženy. Autoři článku zkoumají hlavně techniku PageRank jako analogii k asociacím v paměti. V závěru však autoři obecně sdělují, že cesta k vytvoření lepších vyhledávacích systémů by mohla vést přes hlubší prozkoumání lidské paměti.

Po té, co se vyberou slova, o které bude dotaz rozšiřován, se vytváří samotný rozšířený

dotaz. Ten je obvykle tvořen disjunkcí původního dotazu a nově vytvářených alternativních dotazů. Vytváření alternativních dotazů není triviální úlohou. Často dochází k tomu, že alternativní dotaz nemá smysl. Tento jev může být způsoben tím, že nahrazovaná slova původního dotazu nejsou jednoznačnými synonymy (mohou mít v různých kontextech jiné významy) nebo tím, že jsou ke slovům původního dotazu vztažená jiným způsobem. Jedním možným řešením tohoto problému je filtrace alternativních dotazů založená na četnosti jejich výskytů v korpusu. Další možností je např. využití principu informačního toku.

Filtrace alternativních dotazů je využita např. v [28]. Dotazy jsou filtrovány využitím informace z korpusu, kde jsou hledána sousloví, které vznikají v alternativních dotazech. Alternativní dotazy jsou v [28] řazeny podle toho, jak často se sousloví, ze kterých se skládají, vyskytují v korpusu.

Jiné řešení, využití informačního toku, spočívá v rozšíření dotazu o slova, která souvisí s více slovy původního dotazu. Využitím informačního toku v oblasti rozšiřování dotazu se zabývá např. Bruza [29]. Názorným příkladem, kde by tato technika jistě pomohla, je dotaz „počítačová myš“. Při rozšiřování tohoto dotazu jsou jistě nežádoucí slova jako „pastička“, „sýr“ nebo „sklep“, se kterými by slovo „myš“ mohlo v automaticky vytvářeném sémantickém prostoru souviset.

Kromě výše zmíněných způsobů pro rozšiřování dotazu se rozvíjejí i další. Zajímavý model pro rozšiřování dotazu je popsán v [30]. Základní myšlenkou [30] je vytvoření sítě vztahů mezi slovy využitím více zdrojů informací. Zdroji informací jsou Wordnet [5], statistika spoluvýskytů slov nebo lexikální vztahy mezi slovy, které jsou určeny například stejnými základními tvary slov.

6.2 Související práce

V tomto oddíle jsou prezentovány výsledky některých prací, které se zabývají využitím sémantických prostorů v oblasti rozšiřování dotazu. Vybrané práce názorně ukazují variabilitu přístupů k využití sémantických prostorů v úloze rozšiřování dotazu. Diskutované články využívají k vyhodnocování výsledků korpusy pocházející z TREC. Jako model vyhledávání používají jazykový model.

Azzopardi [14] se zabývá tím, jakých výsledků dosahuje algoritmus pHAL oproti HAL v úloze rozšiřování dotazu. Azzopardi využitím pHAL nedosahuje statisticky významného zlepšení oproti HAL. Jeho práce však ukazuje, že různě nastavený a konstruovaný HAL v úloze rozšiřování dotazu pomáhá. Azzopardi dosahuje využitím HAL vylepšení MAP z původních 25,3% na 27,4%, tedy výsledku, který je o 8,3% lepší.

Práce Baie [31] porovnává využití algoritmu HAL, principu informačního toku a techniky zpětné vazby k rozšiřování dotazu. Nasazením algoritmu HAL Bai dosahuje vylepšení o 2 - 4% v průměrné přesnosti při současném zvýšení úplnosti o 2 - 3%. Výsledné hodnoty vylepšení závisí na použitém korpusu a na nastavení způsobu vyhlazo-

vání v jazykovém modelu. Pomocí informačního toku se Baiovi daří dosahovat lepších výsledků. Přesnost vyhledávacího systému je vylepšována až o 12 - 27% při současném nárůstu úplnosti o 5 - 44%. Pomocí informačního toku v kombinaci s technikou zpětné vazby Bai dosahuje nejlepších výsledků. Přesnost je vylepšena o 25 - 36% a úplnost o 11 - 52%.

Zajímavou prací je také článek [32] od Yan, který se zabývá modifikacemi HAL využitím syntaktických informací. V [32] je publikováno vylepšení výsledků vyhledávání aplikací algoritmu HAL. Testována je hodnota MAP, která je využitím originálního HAL vylepšena z 0,2015 na 0,2299 (14,09%), z 0,2290 na 0,2738 (19,56%) a z 0,2242 na 0,2346 (4,64%) v závislosti na použitém testovacím korpusu. Modifikované algoritmy nazvané eHAL-1 a eHAL-2 dosahují ještě o přibližně 5% procent lepších výsledků.

Jiný sémantický prostor pak zkoušel využít např. Wei-jiang [33]. Algoritmus pLSA v jeho testech překonává metodu zpětné vazby. Autor dosahuje vylepšení průměrné přesnosti v 11 bodech úplnosti o 44,69% u jednoho korpusu, o 65,20% u druhého. Filtrování slov, o která je dotaz rozšiřován, využívá pravděpodobnosti výskytů slov v dotazech kolekcí. Kromě jiného, autor [33] zdůrazňuje, že způsob rozšiřování dotazu by měl být založen spíše na automaticky vytvářených vztazích mezi slovy než na využívání tezaurů nebo slovníků typu Wornet [5]. Autor opodstatňuje své tvrzení tím, že Wordnet [5] a další slovníky nejsou vytvářené pro aplikaci ve vyhledávání.

7 Další práce

Sémantické algoritmy a jimi vytvářené sémantické prostory jsou používány i v jiných úlohách než v rozšiřování dotazu. LSA je možné použít např. i pro sumarizaci textů [34] nebo k indexování ve vyhledávacích systémech [35]. Úlohu rozšiřování dotazu však vnímáme jako velmi důležitou z hlediska významnosti pro dnešní informační společnost a doposud málo prozkoumanou.

Následující oddíl 7.1 stručně popisuje způsob, jakým chceme sémantické prostory zkoumat v souvislosti s rozšiřováním dotazu. Navazující oddíl 7.2 se stručně zabývá zamýšleným výzkumem spojeným s vytvářením sémantických prostorů. Nakonec oddíl 7.3 popisuje stav našeho výzkumu sémantických algoritmů aplikovaných na český korpus.

7.1 Sémantické prostory v úloze rozšiřování dotazu

V odborných pracích, které jsme zatím prozkoumali, objevujeme, jak je možné sémantické prostory na rozšiřování dotazů aplikovat. Prostudované práce se však vždy

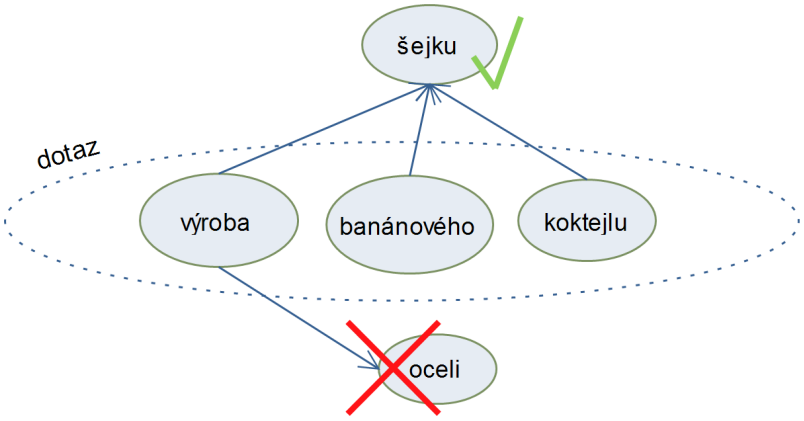
zabývají pouze jedním typem sémantického prostoru[34]. Často je také využití sémantických prostorů kombinováno s dalšími technikami, mezi které patří princip zpětné vazby nebo získávání syntaktických informací.

Nejsme si vědomi žádné studie, která by se využitím sémantických prostorů v oblasti rozšiřování dotazu zabývala systematicky. V odborných článcích a dalších pracích jsme našli zatím jen pokusy o jejich nasazení. Tato nasazení, jak dokazuje oddíl 6.2 jsou však úspěšná. Domníváme se proto, že sémantické prostory si v oblasti vyhledávání zaslouží větší pozornost.

Jedním z našich cílů je proto vytvořit nástroj, který by umožňoval porovnání aplikace různých sémantických prostorů[35] na úlohu rozšiřování dotazu. Chtěli bychom přitom úlohu rozšiřování dotazu s využitím sémantických prostorů dekomponovat na části:

- vytváření sémantických prostorů
  - vstupem je korpus, výstupem sémantický prostor
  - zahrnuje předzpracování textů korpusu
  - zahrnuje použití různých algoritmů pro vytváření sémantických prostorů s různým nastavením
- vytváření modelů pro rozšiřování dotazu
  - výstupem je model pro rozšiřování dotazu
  - zahrnuje vytvoření postupů pro výběr slov pro různé sémantické prostory
  - zahrnuje prožezávání[36] vybraných slov
- aplikace v různých modelech vyhledávání
  - výstupem jsou seznamy relevantních dokumentů k dotazům
  - zahrnuje výběr vyhledávacích modelů, případně jejich kombinování
  - zahrnuje možné využití principu zpětné vazby
- testování
  - výstupem jsou výsledky vyhledávání s rozšiřováním dotazu
  - zahrnuje testování pomocí standardních metrik - přesnost, úplnost, MAP a případné další
  - zahrnuje testování na ručně anotovaných kolekcích typu TREC

[34] vytvářeného stejným algoritmem s různým nastavením
[35] vytvářené různými algoritmy nebo stejnými, ale použitím různého nastavení
[36] jaká slova nebo sousloví v rozšířeném dotazu ponechat a jaká odstranit



Obrázek 5: Rozšiřování dotazu

Chceme vytvořit modulární systém, který bude možné testovat jako celek. Moduly systému budou zejména představovat algoritmy pro vytváření sémantických prostorů, techniky a modely ve vyhledávání. Chtěli bychom tak například vyzkoušet sémantický prostor vytvořený algoritmem BEAGLE, který, podle našich znalostí, na úlohu rozšiřování dotazu doposud aplikován nebyl. Dále chceme porovnávat výkony různých sémantických prostorů aplikovaných v různých vyhledávacích modelech v úloze rozšiřování dotazu.

Důležitou částí našeho výzkumu je také hlubší výzkum sémantických prostorů z hlediska delších slovních spojení. Jako mnoho dalších výzkumníků v oblasti rozšiřování dotazu ani my nechceme původní dotazy rozšiřovat o slova, která souvisí jen s jedním slovem dotazu. Raději bychom dotaz rozšiřovali o slova, která souvisí s více slovy původního dotazu. Důvod ilustruje obrázek 5. Chtěli bychom proto také prostudovat techniku informačního toku a zamyslet se nad souvislostí informačního toku se sémantickými prostory.

7.2 Výzkum spojený se sémantickými prostory

Chtěli bychom také vyzkoušet vytváření sémantických prostorů, jejichž slovníkem by nebyla jen slova, ale i delší slovní spojení. Tato slovní spojení bychom ovšem museli nějakým způsobem vybírat. Domníváme se, že by bylo zajímavé víceslovné výrazy identifikovat například četností výskytů nebo využitím syntaxe. Pomocí sémantických prostorů bychom pak mohli určovat souvislosti mezi těmito souslovími. Jsme přesvědčeni, že sousloví nesou užší význam, a proto i vztahy mezi nimi by mohly být v jistém smyslu přesnější. Na druhou stranu, jsme si vědomi, že určování vztahů mezi souslovími je problematičtější kvůli problému „řídkosti dat“.

Přemýšlíme také nad samotnými algoritmy pro vytváření sémantických prostorů. Zdá se nám, že mají několik nedostatků. Nelíbí se nám například to, že hledají vztahy mezi slovy vždy stejným způsobem. Klademe si otázky, jestli by algoritmy neměly brát v úvahu například to, s jakými slovními druhy zrovna pracují. Neměla by být třeba velikost HAL okénka pro předložky být jinak veliká než pro přídavná jména nebo slovesa?

Otázkou také je, jaké informace využívat při vytváření sémantických prostorů. Možná by bylo vhodné využívat znalostí o slovních druzích slov, možná znalostí o jejich četnostech. Zaráží nás také to, že neexistují algoritmy pro vytváření sémantických prostorů, které prochází data vícekrát a jinými způsoby. Podobným způsobem to přeci dělá člověk. Zamýšlíme se dále nad tím, že vektory slov v sémantickém prostoru mají složky, mezi nimiž neexistují žádné vztahy. Na složky vektorů můžeme při tom nahlížet jako na vlastnosti slov v sémantickém prostoru. Zřejmým nedostatkem sémantických vektorů je, podle nás, ignorování toho, že nějaké vlastnosti jsou důležitější než ostatní.

Myslíme si, že sémantické prostory jsou vhodnou strukturou pro ukládání vztahů mezi slovy. Avšak hlavně proto, že do sémantických prostorů je možné uložit libovolnou grafovou strukturu. Určitě ne proto, že současné sémantické prostory jsou věrnou kopií sémantické paměti člověka a dají se snadno využívat.

7.3 Čeština a sémantické prostory

Dalším naším výzkumným záměrem je aplikace sémantických prostorů na české korpusy. Nevíme o žádné práci, kromě naší, která by se zabývala využitím sémantických prostorů k zjišťování vztahů mezi českými slovy. Výsledky naší dosavadní práce jsou publikovány v [36], kde jsme se zabývali testováním algoritmů LSA, HAL a COALS aplikovaných na češtinu.

Využili jsme při tom korpusu ČTK[37] 1998, který obsahuje 131, 956 novinových článků publikovaných na internetu. Při práci na [36] jsme používali nástroje [37] pro morfologické značkování českého jazyka. Ukázalo se, že lepších výsledků dosahují algoritmy na lemmatizovaných datech.

Algoritmy jsme vyhodnocovali pomocí přeloženého RG testu, který jsme si museli sami vytvořit. Testy podobné anglickému TOEFL nebo ESL se nám zatím bohužel nepodařilo získat. Rovněž testy podobné RG nejsou pro český jazyk k dispozici. Dalším úskalím našeho výzkumu je velikost českého korpusu. Jsme přesvědčeni, že s větším korpusem [38] bychom dosahovali lepších výsledků.

Při testování algoritmů jsme zjistili, že algoritmus LSA není pro češtinu zdaleka tak

[37] Česká tisková kancelář

[38] řádově desítky miliónů slov

úspěšný jako pro anglický jazyk. Výsledky také ukazovaly, že použití lemmatizace pro češtinu zlepšuje výsledky. Nejlepších výsledků jsme někdy dosahovali, když jsme z korpusu před vytvářením sémantických prostorů vymazali nevýznamová slova jako zájmena, předložky, spojky, částice a interpunkci.

Také jsme objevili, že když odstraňujeme páry z RG testu obsahující slova, která jsou málo četná v korpusu, všechny sémantické prostory dosahují lepších výsledků. Jsme proto přesvědčeni, že výsledky „kazí“ páry, které se skládají ze slov, o kterých algoritmy naleznou v korpusu málo informací.

Kromě algoritmů LSA, HAL a COALS jsme žádné další na český jazyk zatím nevyzkoušeli. Naší další snahou bude proto vyzkoušet i algoritmy BEAGLE nebo RI. Plánujeme algoritmy aplikovat na větší korpusy a jejich výsledky porovnávat pomocí dalších testů.

8 Závěr

Využití vztahů mezi slovy pro vyhledávání je dnes populární myšlenka. Vytváření sémantických prostorů umožňuje vztahy vypočítat automaticky. Práce výzkumníků v oblasti vyhledávání informací, konkrétně v technice rozšiřování dotazu, dokazují, že využití sémantických prostorů má smysl. Přesto doposud nevíme o žádné rozsáhlejší studii v této oblasti.

Proto chceme sémantickým prostorům porozumět do hloubky a aplikovat je na úlohu rozšiřování dotazu. Plánujeme při tom využít moderních přístupů v oblasti vyhledávání, kterými jsou např. jazykový model vyhledávání nebo princip informačního toku.

Také se chceme zabývat problematikou delších slovních spojení v souvislosti se sémantickými prostory. Uvažujeme nad vyhledáváním vztahů mezi delšími slovními spojeními pomocí algoritmů pro vytváření sémantických prostorů.

Dalšími výzkumnými záměry jsou hledání a odstranění nedostatků algoritmů pro vytváření sémantických prostorů. Zamýšlíme se nad vylepšením těchto algoritmů. Chtěli bychom při tom využívat například další informace o slovech, jakými jsou morfologické nebo syntaktické značky.

V neposlední řadě chceme aplikovat sémantické prostory nejen na anglické, ale také na české texty. Doposud nevíme o žádné práci zaměřené na český jazyk, která by se úlohou rozšiřování dotazu s využitím sémantických prostorů zabývala. Nevíme ani o práci, která by využívala sémantických prostorů pro zjišťování vztahů mezi českými slovy.

Reference

[1] Osgood, C., Suci, G., and Tannenbaum, P. (1957). *The measurement of meaning*, University of Illinois Press.

[2] Harris, Z. (1954). Distributional structure. (J. Katz, Ed.) *Word Journal Of The International Linguistic Association, 10*(23), 146-162. Oxford University Press.

[3] Tulving, E. (1983). *Elements of episodic memory*. (D. Broadbent, L. McGaugh James, J. Mackintosh Nicoloas, I. Posner Michael, E. Tulving, & L. Weiskrantz, Eds.), *Behavioral and Brain Sciences* (Vol. 7, pp. 223-268). Oxford University Press.

[4] Sowa, J. (1991). *Principles of Semantic Networks*. (J. Sowa, Ed.) (pp. 157-190). Morgan Kaufmann.

[5] George A. Miller. 1995. Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41. ACM.

[6] Pala K., Smrž P. (2004), Building Czech Wordnet. *Romanian Journal of Information Science and Technology, Romanian Academy, 7*, 1-2, p. 79-88.

[7] Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav Res Methods Instrum Comput, 28*(2), 203-208 203–208.

[8] Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2), 259-284. Routledge.

[9] Turney, P. (2006). Similarity of Semantic Relations. *Computational Linguistics, 32*(3), 379-416. MIT Press.

[10] Rohde, D. T., Gonnerman, L., & Plaut, D. (2004). An improved method for deriving word meaning from lexical co-occurrence. *Cognitive Science*.

[11] Sahlgren, M. (2002). Vector-based Semantic Analysis: Representing Word Meaning Based on Random Labels. *ESSLI Workshop on Semantic Knowledge Acquistion and Categorization*. In ESSLI Workshop on Semantic Knowledge Acquistion and Categorization.

[12] Jones, M., Kintsch, W., & Mewhort, D. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55*(4), 534-552. Elsevier.

[13] Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 99*, 50-57. ACM Press

[14] Azzopardi, L., Girolami, M., & Crowe, M. (2005). *Probabilistic hyperspace analogue to language. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 05* (p. 575).

[15] Dumais, S. (1992). Enhancing performance in latent semantic indexing (LSI) retrieval. *Bellcore System Journal*, (TM-ARH-017527), 1-19.

[16] Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. (M. M. T. Cambridge, Ed.) *Reading* (Vol. 26, pp. 277-279). MIT Press.

[17] Bullinaria, J., & Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods, 39*(3), 510-526. Psychonomic Society Publications.

[18] Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8*(10), 627-633. ACM Press.

[19] Miller, G., & Charles, W. (1991). Contextual Correlates of Semantic Similarity. *Language & Cognitive Processes, 6*(1), 1-28. Psychology Press.

[20] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Transactions on Information Systems, 20*(1), 116-131.

[21] Jurgens and Stevens, (2010). The S-Space Package: An Open Source Package for Word Space Models. In *System Papers of the Association of Computational Linguistics*.

[22] Mogotsi, I. (2010). Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze: Introduction to information retrieval. *Information Retrieval, 13*(2), 192-195.

[23] Fuhr, N. (1992). Probabilistic Models in Information Retrieval. *The Computer Journal, 35*(3), 243-255. Oxford University Press.

[24] Zhai, C. (2008). Statistical Language Models for Information Retrieval. (C. Manning, Ed.) *Synthesis Lectures on Human Language Technologies, 1*(1), 1-141. Cambridge University Press.

[25] Xu, J., & Croft, W. (1996). Query expansion using local and global document analysis. (H. Frei, D. Harman, P. Schaüble, & R. Wilkinson, Eds.) *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 96*, (Zurich, Switzerland), 4-11. ACM Press.

[26] Parapar, D., Barreiro, A., Losada, D.E. (2005). Query expansion using wordnet with a logical model of information retrieval. In IADIS AC 487-494.

[27] Griffiths, T., Steyvers, M., & Firl, A. (2007). Google and the Mind. *Psychological Science, 18*(12), 1069-1076.

[28] Zukerman, I., & Road, B. (2002). Lexical Query Paraphrasing for Document Retrieval. (S. Tseng, T. Chen, & Y. Liu, Eds.) *Proceedings of the 19th international conference on Computational linguistics*, 1–7.

[29] Bruza, P., & Song, D. (2002). Inferring Query Models by Computing Information Flow.

[30] Collins-Thompson, K., & Callan, J. (2005). Query expansion using random walk models. *Framework*, 704-711. ACM.

[31] Bai, J., Song, D., Bruza, P., Nie, J., & Cao, G. (2005). Query expansion using term relationships in language models for information retrieval. *Knowledge Creation Diffusion Utilization, 05pages*, 688-695. ACM.

[32] Yan, T., Maxwell, T., Song, D., Hou, Y., & Zhang, P. (2010). Event-Based Hyperspace Analogue to Language for Query Expansion. *Proceedings of the ACL 2010 Conference Short Papers* (pp. 120-125).

[33] Li Wei-jiang, Zhao Tie-jun, Zang Wen-mao (2009). PLSA-Based Query Expansion. International Conference on Computer and Information Science (icis 2009), pp. 400-405.

[34] Steinberger, J., Ježek K. (2004). Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. *Proceedings of ISIM '04*, pages 93–100.

[35] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

[36] Krčmář L., Konopík M., Ježek K. (2011) Exploration of Semantic Spaces Obtained from Czech Corpora. *Proceedings of the DATESO 2011 Conference*, pp. 97-107.

[37] J. Hajič, A. Böhmová, E. Hajičová, B. Vidová Hladká, The Prague Dependency Treebank: A Three-Level Annotation Scenario. In A. Abeillé (ed.): *Treebanks Building and Using Parsed Corpora*. pp. 103-127. Amsterdam, The Netherlands: Kluwer, 2000.