Západočeská univerzita v Plzni

Fakulta aplikovaných věd

VYHLEDÁVÁNÍ INFORMACÍ V ŘEČI A VYUŽITÍ SLEPÉ ZPĚTNÉ VAZBY

Ing. Lucie Skorkovská

Disertační práce
k získání akademického titulu doktor
v oboru Kybernetika

Školitel: Prof. Ing. Josef Psutka, CSc.
Katedra kybernetiky

Plzeň 2016

University of West Bohemia

Faculty of Applied Sciences

SPOKEN DOCUMENT RETRIEVAL AND THE USE OF BLIND RELEVANCE FEEDBACK

Ing. Lucie Skorkovská

Doctoral thesis

submitted for the degree Doctor of Philosophy
in the field of Cybernetics

Supervisor: Prof. Ing. Josef Psutka, CSc.
Department of Cybernetics

Pilsen 2016

Poděkování

Tato práce vznikla za odborného vedení mého školitele Prof. Ing. Josefa Psutky, CSc. Dále bych chtěla poděkovat za odborné rady a konzultace Doc. Ing. Pavlu Ircingovi, Ph.D.

Ráda bych poděkovala také své rodině za to, že věřili v důležitost studia a od dětství mě vedli k přípravě na budoucí studium na vysoké škole. Jsem jim vděčná za jejich podporu a vytvoření dobrých studijních podmínek, provázejících mne jak na střední, tak následně i na vysoké škole. Také bych chtěla poděkovat kolegům z oddělení umělé inteligence katedry kybernetiky za cenné rady při vypracovávání této práce.

Prohlášení

Prohlašuji, že jsem tuto disertační práci vypracovala samostatně, s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

V Plzni dne Lucie Skorkovská

Anotace

Díky rychlému rozvoji počítačové techniky je stále více informací ukládáno ve formě multimediálních databází, ve velké míře dostupných prostřednictvím internetu. Prohlížení takovýchto rozsáhlých databází manuálně není možné, proto v současné době dochází k rychlému rozvoji vyhledávání informací v řeči jako určité nadstavby již běžně používaného vyhledávání informací v textu. Pro úspěšné vyhledávání v řeči je nutné propojení systému pro automatické rozpoznávání řeči a systému pro vyhledávání informací. V procesu vyhledávání informací často dochází k efektu označovanému jako slovníkový problém, tedy že dokumenty a dotazy nejsou psány stejnou formou, nepoužívají stejná slova a dochází tak ke zhoršení výsledků vyhledávání. Tento problém může být ještě umocněn v případě vyhledávání informací v řeči, kdy automatické rozpoznávání řeči může vnášet další rozdíly v použitém slovníku, případně i chyby. Metody rozšíření dotazu, zvláště pak použití zpětné vazby, se ukázaly jako jedny z nejpřirozenějších a nejúspěšnějších postupů, jak tento problém řešit pomocí vytvoření nového, úspěšnějšího dotazu. Tato práce prezentuje způsoby jak lze vyhledávat informace v řeči, používané metody a postupy, a dále se věnuje zapojení zpětné vazby, zejména pak slepé zpětné vazby, do procesu vyhledávání informací.

Anotation

With the rapid development of the computer technology the ever increasing amount of information is stored in the form of multimedia databases, widely available through the Internet. Browsing such a large database manually is not possible, therefore a rapid development in the area of spoken document retrieval as a certain extension of already commonly used text information retrieval occurs recently. To search successfully in speech the connection between the automatic speech recognition system and the information retrieval system is needed. In the process of information retrieval an effect often appears referred to as the vocabulary problem, namely that the documents and queries are not written in the same form, they do not use the same words, and this leads to a deterioration in the search results. This problem can be magnified in the case of speech information retrieval, in which automatic speech recognition can bring other differences in vocabulary usage, or even errors. Query expansion methods, especially the use of relevance feedback has proven to be one of the most natural and successful techniques how to solve this problem by creating a new, more useful query. This thesis presents the possibilities of retrieving information from the speech data, commonly used methods and procedures, and addresses the incorporation of the relevance feedback, especially the blind relevance feedback, into the information retrieval process.

Obsah

Seznam zkratek

Acc	Accuracy
ASR	Automatic Speech Recognition
BIM	Binary Independence Model
BM25	Best Match 25
BRF	Blind Relevance Feedback
CL-SDR	Cross-Language Spoken Document Retrieval
CL-SR	Cross-Language Speech Retrieval
CLEF	Cross-Language Education and Function
CN	Confusion Networks
CoN	Cohort Normalization
D	Dirichlet (vyhlazování)
DMM	Dirichlet Mixture Model
EM	Expectation-Maximization
FPR	False Positive Rate
GAP	Generalized Average Precision
GTMN	General topic model normalization
HMM	Hidden Markov Model
IDF	Inverse Document Frequency
JM	Jelinek-Mercer (vyhlazování)
KLD	Kullback-Leibler Distance
LVCSR	Large-Vocabulary Continuous Speech Recognition
MALACH	Multilingual Access to Large Spoken Archives
MAP	Mean Average Precision
mGAP	mean Generalized Average Precision
MIT	Massachusetts Institute of Technology
MLE	Maximum Likelihood Estimate
MpSD	Mean plus Standard Deviation
NGSW	National Gallery of the Spoken Word
NTCIR	NII Testbeds and Community for Information access Research
OOV	Out Of Vocabulary
OSTI-SI	Open-Set Text-Independent Speaker Identification
P	Precision
PDT	the Prague Dependency Treebank
PLSA	Probabilistic Latent Semantic Analysis
PSPL	Position Specific Posterior Lattice
QL	Query Likelihood model
R	Recall
RM	Relevance Model
ROC	Receiver Operating Characteristic

RSV	Robertson Selection Value
SCR	Spoken Content Retrieval
SDR	Spoken Document Retrieval
SMM	Simple Mixture Model
STD	Spoken Term Detection
SUR	Spoken Utterance Retrieval
SVM	Support Vector Machine
T-norm	Test normalization
TALE	Time-Anchored Lattice Expansion
TF	Term Frequency
TF-IDF	Term Frequency - Inverse Document Frequency
TMI	Time-based Merging for Indexing
TPR	True Positive Rate
TREC	Text REtrieval Conference
TREC/SDR	Text REtrieval Conference / Spoken Document Retrieval
TS	Two-Stage (vyhlazování)
UBMN	Universal Background Model Normalization
UCN	Unconstrained Cohort Normalization
VMR	Video Mail Retrieval
WCN	Word Confusion Networks
WER	Word Error Rate
WMN	World Model Normalization
ZV	Zpětná Vazba

Seznam tabulek

Seznam obrázků

Kapitola 1

Úvod

Vyhledávání informací jako vědní obor se poprvé objevilo v odborných textech v padesátých letech dvacátého století, tento termín - *vyhledávání informací*[1] zavedl Calvin Mooers v roce 1948 ve své diplomové práci [102] na MIT[2]. Hlavním cílem vyhledávání informací je nalézt informace požadované uživatelem a ty mu poté předložit seřazené podle předpokládané relevance. Zpočátku se metody vyhledávání informací uplatňovaly pouze v knihovnických systémech, s rozvojem internetu na konci dvacátého století se ale tento obor stává objektem stále většího zájmu jak expertů z oblasti informatiky, tak i veřejnosti.

Vyhledávání informací v textu je již dlouhou dobu velice důležitým oborem mezi informačními vědami, rozvíjejícím se v posledních desetiletích zejména díky rychlému vzestupu internetu a vývoji v oblasti počítačového hardwaru. Rychlý vývoj internetu také umožnil pohlížet na něj jako na univerzální informační médium, spíše než textovým zdrojem informací se tak stává zdrojem multimediálním. Čím dál více informací je ukládáno nejen v textové podobě, ale i jako audio či video záznamy. Rozsáhlé kolekce audio-vizuálních dat z různých oblastí, jako je historie, umění či kultura, jsou v současné době dostupné online, je tedy přirozené, že se výzkum v oblasti vyhledávání informací rozšiřuje také na vyhledávání informací v multimédiích, v tomto případě v řeči. Lze říci, že vyhledávání v textu je již běžnou součástí každodenního života, oproti tomu vyhledávání v řečových záznamech je ve většině případů stále ještě omezeno na textové vyhledávání v ručně přepsaných nahrávkách, případně pouze v jejich popisech - metadatech. Ať se jedná o archívy televizních nahrávek, záznamy konferencí nebo přednášek, cílem vyhledávání informací v řeči je zpřístupnit informace v nich obsažené bez nutnosti jejich manuálního přepisu.

Systém pro vyhledávání informací v řeči vzniká spojením dvou částí, první částí je systém pro automatické rozpoznávání řeči - ASR[3], na který navazuje systém pro vyhledávání informací. Vzhledem k tomu, že vývoj kvalitního systému pro automatické rozpoznávání řeči je komplikovaná a časově náročná úloha, bývá propojení těchto dvou systémů spíše volné. Je tedy možné říci, že úkolem vyhledávání informací v řeči je co nejlepším způsobem využít informace získané z ASR systému a zaměřit se na vývoj metod umožňujících dosáhnout takové úspěšnosti vyhledávání, jaké by bylo možné dosáhnout při vyhledání v přesném přepisu řečového záznamu do textu.

Hlavním problémem, způsobujícím špatné výsledky vyhledávání, je nesoulad mezi slovy dotazu a slovy vyskytujícími se v kolekci dokumentů. Tento problém je často nazýván jako *slovníkový problém*. Jednou z jeho příčin v oblasti vyhledávání informací v řeči je použití

[1] Information Retrieval
[2] Massachusetts Institute of Technology
[3] Automatic Speech Recognition

ASR systému pro přepis řečových dokumentů. I přes to, že současné ASR systémy používají velké slovníky, může se stát, že některé slovo bude ve slovníku chybět a bude tedy následně chybět i v kolekci dokumentů. V případě vysoké chybovosti ASR systému mohou být do automatických přepisů dokumentů zaneseny další chyby v podobě špatně rozpoznaných slov. Druhou z příčin slovníkového problému je, že uživatel systému pro vyhledávání informací nepoužívá stejných výrazových prostředků, jaké se vyskytují v kolekci.

První problém může být částečně zmírněn použitím jiných reprezentací řeči ze systému ASR než jen nejlepších přepisů, například použitím slovních mřížek nebo jejich reprezentací, a chybovost ASR systému tak může být zmenšena. Pro řešení druhého problému se při vyhledávání v textu ukázaly jako nejlepší metody automatického rozšíření dotazu, zejména použitím slepé zpětné vazby lze dosáhnout největšího zlepšení výsledků vyhledávání [176, 18, 81]. V experimentech s rozšířením dotazu a použitím slepé zpětné vazby ve vyhledávání informací v řeči se ukázalo, že použití těchto metod může zároveň také výrazně zmenšit vliv chybovosti ASR systému a chybějících slov v jeho slovníku. Chybovost automatických přepisů řeči i při WER[4] kolem 35% nemá v tomto případě zásadní vliv na výsledky vyhledávání [118, 71, 23, 79, 110]. Zpětná vazba při použití ve vyhledávání informací v řeči využívá vzájemný výskyt slov v prvních vyhledaných dokumentech, které mívají malou chybovost [137], a pomocí těchto slov je pak možné vyhledat i dokumenty, kde bylo slovo z dotazu špatně rozpoznáno [81].

Výzkum v oblasti vyhledávání informací v řeči se v současné době upírá dvěma směry. První z nich je zaměřen na využití bohatších reprezentací řeči získaných z ASR systému a jeho cílem jsou zejména situace, kdy dochází k velké chybovosti automatických přepisů (nad 50% WER) například u špatných nahrávek konverzační řeči. Druhým směrem výzkumu je zaměření se na vylepšení metod vyhledávání informací a zejména na větší zapojení metod rozšíření dotazu a slepé zpětné vazby a na vylepšení těchto metod. Tato práce se bude zabývat zejména druhým směrem výzkumu, tedy snahou o vylepšení metod slepé zpětné vazby a jejího zapojení do vyhledávání informací v řeči.

Principem zpětné vazby ve vyhledávání informací je využít informaci získanou z relevantních dokumentů z prvního průchodu vyhledávacího sytému a použít ji na rozšíření dotazu novými slovy pro druhý průchod vyhledávání. Ve většině případů nemá systém informaci o relevanci vyhledaných dokumentů, zvolí tedy tyto dokumenty „slepě“ a považuje je za relevantní. Tato volba je většinou provedena pouze výběrem prvních k dokumentů a z těchto dokumentů je poté vybráno několik slov pro rozšíření dotazu. Většina metod slepé zpětné vazby, od těch základních až po ty nejsložitější, má jedno společné - výběr počtu k relevantních dokumentů je realizován na základě předchozích experimentů, či určitého zvyku a je nastaven pro celý systém, tedy všechny dotazy, stejně. Tento způsob není vhodný, nerespektuje rozdílnosti jednotlivých dotazů ani úspěšnost vyhledávání daného dotazu. Jedním z cílů této práce je tedy navrhnout metodu pro dynamické stanovení množství dokumentů použitých pro slepou zpětnou vazbu.

1.1 Cíle disertační práce

Cílem této práce je popsat a vysvětlit základy oboru vyhledávání informací, používané metody a postupy, a ukázat aplikaci těchto metod v oboru vyhledávání informací v řeči. Práce se teoreticky zabývá jak využitím informací získaných z ASR systému a popisem upravených metod pro vyhledávání v různých řečových reprezentacích, tak i popisem metod zpětné a slepé zpětné vazby a popisem jejího vlivu na úspěšnost vyhledávání. Práce

[4] Word Error Rate

je dále zaměřena na porovnání jednotlivých metod pro vyhledávání informací při použití v oblasti vyhledávání v řeči a zejména na aplikaci metod slepé zpětné vazby a nastavení jejích parametrů. Práce se zabývá vlivem těchto nastavení na výsledky vyhledávání a důkladným otestováním jejich vzájemného ovlivnění, zejména co se týká doporučeného nastavení počtu dokumentů a termů pro použití ve slepé zpětné vazbě. Základním problémem všech metod slepé zpětné vazby je špatné nastavení výběru pseudo relevantních dokumentů, v práci je ukázáno, jak tento výběr zásadně ovlivní výsledky vyhledávání. Z tohoto důvodu jsou navrženy nové metody pro výběr dokumentů pro slepou zpětnou vazbu a následně jsou odvozeny pro zapojení do hlavních metod pro vyhledávání informací.

Cíle práce se dají shrnout do několika bodů:

- Popsat metody a principy používané v oblasti vyhledávání informací, prozkoumat možnosti vyhledávání informací v řeči. Popsat a prostudovat metody slepé zpětné vazby a vlivu na vylepšení vyhledávání při jejím zapojení do vyhledávacího systému.
- Provést experimentální porovnání jednotlivých metod na české kolekci dokumentů pro vyhledávání informací v řeči, zjistit nejlepší možná nastavení a úpravy metod. Zaměřit se hlavně na použití slepé zpětné vazby, experimentálně ověřit vliv a vzájemné vazby nastavení jejích parametrů.
- Vylepšit zapojení zpětné vazby do vyhledávacího systému, navrhnout metodu pro automatický výběr počtu dokumentů pro zpětnou vazbu a experimentálně ověřit její výsledky.

1.2 Struktura práce

Tato práce se zabývá vyhledáváním informací v řeči a zapojením slepé zpětné vazby do tohoto systému. V kapitole 2 je uveden stručný úvod do historie tohoto oboru a představení úlohy řeči ve vyhledávání informací, popis systému pro vyhledávání v řeči a také zapojení zpětné vazby do tohoto systému a motivace k použití metod slepé zpětné vazby.

Kapitola 3 uvádí základní míry používané pro ohodnocení úspěšnosti jednotlivých metod pro vyhledávání informací. Tyto metody jsou představeny v kapitole 4. V kapitole 5 je nejprve stručně popsán systém pro automatické rozpoznávání řeči a poté jsou popsány možnosti vyhledávání v jednotlivých reprezentacích řeči získaných z ASR systému a upravené metody pro vyhledávání v nich.

Kapitola 6 obsahuje popis metod zpětné vazby, popsány jsou jak přístupy klasické zpětné vazby, tak metody slepé zpětné vazby z nich vycházející. Text se zabývá vývojem metod od prvních počátků až k popisu současného stavu oboru. Součástí kapitoly je také shrnutí jednotlivých přístupů z hlediska základních problémů aplikace a nastavení metod.

V kapitole 7 jsou uvedeny vlastní experimenty a navržené metody. Na začátku je popsána česká kolekce pro vyhledávání informací v řeči, na které jsou experimenty provedeny, následuje popsání experimentů s jednotlivými metodami pro vyhledávání a jejich úpravami. Navazují experimenty se zapojením slepé zpětné vazby a průzkum jejích možností. Hlavní část výzkumu práce se zabývá ověřením výsledků nově navržené metody pro nastavení slepé zpětné vazby, tato metoda je představena a jsou popsány experimenty s jejím zapojením do nejčastěji používaných metod pro vyhledávání informací. Na závěr jsou také popsány experimenty s navrženou metodou v příbuzné oblasti detekce témat textu.

Závěr práce, kapitola 8 obsahuje shrnutí práce a dosažených výsledků.

Kapitola 2

Vyhledávání informací v řeči a slepá zpětná vazba

Cílem této kapitoly je přiblížit obor vyhledávání informací v řeči, jeho historii a typy úloh, které do tohoto oboru spadají. V další části kapitoly bude navázáno popisem možností slepé zpětné vazby a motivací jejího zapojení do vyhledávání informací v řeči.

2.1 Úlohy řeči ve vyhledávání informací

Dalo by se říci, že existují tři různé úlohy vzniklé ze spojení řečových dat a vyhledávání informací.

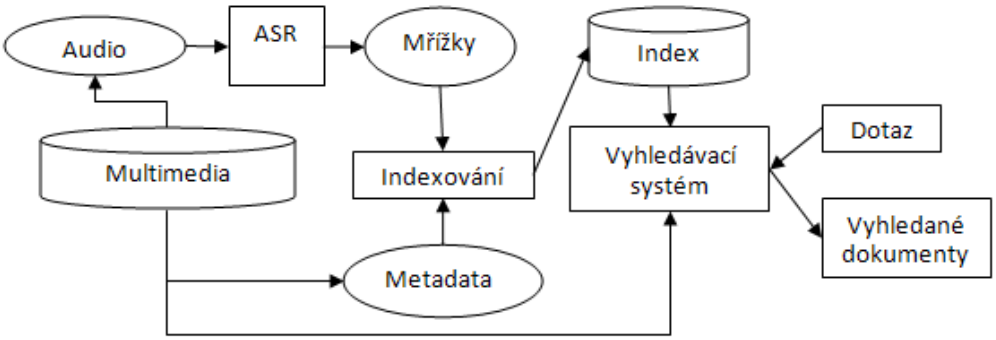
2.1.1 Vyhledávání textových dotazů v řečových datech

Tato úloha bývá označována jako vyhledávání informací v řeči (Spoken Document Retrieval[1] a Spoken Utterance Retrieval[2]). Cílem je umožnit vyhledávání v rozsáhlých archivech řečových nahrávek stejně tak, jako je možné vyhledávat v textových dokumentech. Na řečová data je použit systém na rozpoznávání řeči, jeho výstupem mohou být slovní nebo subslovní mřížky, seznam nejlepších přepisů či pouze jeden nejlepší přepis. V tomto rozpoznaném přepisu je pak vyhledáván textový dotaz pomocí systému na vyhledávání informací.

2.1.2 Vyhledávání v textových dokumentech pomocí řečových dotazů

Vyhledávání probíhá v existujících textových databázích pomocí mluvených dotazů. Dotaz je pomocí systému pro automatické rozpoznávání řeči rozpoznán a se stanovenou neurčitostí použit pro vyhledávání. Tato úloha bývá nazývána řečové vyhledávání (Voice Search) [168], většinou je součástí dialogového systému pro odpovídání dotazů. Příkladem může být automatický telefonní seznam a adresář, tedy systém poskytující adresu a telefonní číslo na požadovanou firmu, restauraci, či jednotlivce jako odpověď na řečový dotaz [181].

[1] SDR
[2] SUR



Obrázek 2.1: Struktura typického sytému pro vyhledávání informací v řeči

2.1.3 Vyhledávání v řečových datech řečovými dotazy

Tento případ je složitější než předchozí dva, protože jak dotaz tak dokumenty jsou v řečové formě, vzniká tedy určitá nejistota na obou stranách. Úloha se řeší různými přístupy: od přímého porovnání zvukového signálu dotazu a dokumentů [178] po přístup dotazu příkladem (query-by-example) [28], kde jsou dotaz a dokumenty rozpoznány systémem pro automatické rozpoznávání řeči a nejlepší přepis nebo celá mřížka jsou pak porovnány.

2.2 Popis vyhledávání informací v řeči

Tato práce se bude věnovat první úloze, tedy vyhledávání informací v řečových dokumentech pomocí textových dotazů. Schéma typického systému pro tuto úlohu je vidět na obrázku 2.1. Tuto úlohu můžeme ještě rozdělit na dvě podúlohy, podle toho, co je cílem vyhledávání:

2.2.1 Vyhledávání řečových dokumentů (SDR)

Jako řečové dokumenty lze chápat například jednotlivé nahrávky zpravodajských reportáží, kde bude cílem vyhledat celou nahrávku. Je ale také možné vyhledávat v rozsáhlé řečové nahrávce, kde bude cílem vyhledat úsek týkající se předloženého dotazu. V takovém případě se většinou rozdělí nahrávka na segmenty, buď podle předem stanovené délky (podle času nebo počtu slov) nebo podle shodného tematického obsahu, a tyto segmenty jsou poté vyhledávány. Někdy se tento přístup označuje také jako vyhledávání pasáží[3] [118, 36]. Cílem vyhledávání není nalézt přímo výskyt slov z dotazu v nahrávce, ale nalézt dokument (pasáž), který pojednává o uživatelem požadované informaci vyjádřené formou dotazu. Obě tyto varianty bývají někdy také souhrnně nazývány jako vyhledávání řečového obsahu[4].

[3] Passage retrieval
[4] SCR - Spoken Content Retrieval

2.2.2 Vyhledávání řečových promluv (SUR)

Cílem úlohy je v řečové nahrávce vyhledat přesné místo výskytu nějaké promluvy, většinou se jedná o krátké slovní spojení [138]. Tato úloha je rozšířením úlohy vyhledání klíčového slova[5], kde je cílem vyhledat přesný výskyt dotazovaného slova v řečové nahrávce.

2.3 Vliv chybovosti systému pro automatické rozpoznávání řeči

V oboru vyhledávání informací v řeči je nutné propojit systém pro automatické rozpoznávání řeči se systémem pro vyhledávání informací. Nedochází však k přímému napojení těchto dvou systému, ale k využití informace ze systému pro automatické rozpoznávání řeči systémem pro vyhledávání informací.

Základním přístupem je získat automatický přepis řeči do textu a na ten použít již existující systém pro vyhledávání v textu (viz podkapitola 5.2). Při zpracování TREC[6] úloh v letech 1997 - 2000 [38] bylo ukázáno, že i pro systémy s chybovostí kolem 30% (zpracovávající například zpravodajské nahrávky) lze dosáhnout téměř stejné přesnosti vyhledávání jako při použití manuálních přepisů řeči. Tyto závěry byly následně potvrzeny i v dalších pracích kde byl zkoumán vliv chybovosti ASR systému na výsledky vyhledávání [118, 71, 23, 79].

Pro systémy s chybovostí větší, okolo 50% a více, přesnost vyhledávání klesá. V práci [71] byl testován vliv různé úrovně chybovosti ASR systému na výsledky vyhledávání, při vyhledávání v přepisech s chybovostí kolem 25% bylo dosaženo 14% relativního zlepšení oproti ASR s chybovostí 61,5%. Na druhou stranu výsledky systému s chybovostí 25% byly jen o 3,7% relativně horší než pro manuální přepisy. V případě systému s velkou chybovostí je tedy nutné buď vylepšit systém ASR (pokud je to možné), nebo alespoň využít co nejvíce informací, které tento systém může poskytnout. Tedy například použít místo nejlepších přepisů slovní mřížky či jejich reprezentace (viz podkapitola 5.3). Příkladem dat, pro která dosahují systémy takové chybovosti, jsou například nahrávky spontánní řeči, telefonní rozhovory, audio konference nebo hlasová pošta.

Na druhou stranu, pokud systém dosahuje dostatečně malé chybovosti přepisu, není vhodné se věnovat výzkumu těchto metod, ale je lepší zaměřit se na vylepšení algoritmů vyhledávání informací a metod rozšíření dotazu. V citované práci [71] bylo ukázáno, že při použití metod slepé zpětné vazby dochází ke konstantnímu zlepšení výsledků vyhledávání, nehledě na prvotní výkon vyhledávacího systému, ovlivněný rozdílnou chybovostí ASR systému. K podobnému závěru došel Chen a kol. v práci [24], kde byl testován vliv akustických (například použití mřížkových reprezentací) a lingvistických (například slepá zpětná vazba) vylepšení vyhledávání informací v řeči. Ukázalo se, že při použití lingvistických vylepšení se dá dosáhnout mnohem většího zlepšení vyhledávání (zlepšení o 21,3%, respektive 12,3% na druhé kolekci) než při použití akustických vylepšení (zlepšení o 3,5%, respektive o 2,8%).

Můžeme rozlišit dva důvody, proč dochází k chybám v automatickém rozpoznávání řeči.

[5] Spoken Term Detection (STD)
[6] Text REtrieval Conference

2.3.1 Chybný výběr nejlepší hypotézy

Systém pro automatické rozpoznávání řeči při generování nejlepšího přepisu vybírá ze všech možných slovních hypotéz tu s největší aposteriorní pravděpodobností. Pokud ale největší pravděpodobnosti dosáhne slovo, které se v původní řeči nevyskytovalo, získáme chybný přepis a správné slovo bude v přepisu chybět. Pokud právě toto chybějící slovo bude jedním ze slov dotazu, pak dokument s takto chybným přepisem nebudeme moci nalézt, nebo dosáhne menšího ohodnocení podobnosti.

2.3.2 Chybějící slovo ve slovníku

Druhým problémem vyskytujícím se v automatických přepisech řeči jsou slova, která nejsou obsažena ve slovníku systému pro automatické rozpoznávání řeči. Takové slovo se nebude vyskytovat ve slovní mřížce, nepomůže tedy zpracování všech možných hypotéz. Místo slova nevyskytujícího se ve slovníku určíme z hypotéz jiné slovo s podobnou výslovností. Pokud v dotazu bude obsaženo slovo, které není ve slovníku, nebudeme moci nalézt na jeho základě relevantní dokumenty.

Problémy se špatným určením nejlepší hypotézy lze zmenšit použitím všech slovních hypotéz, tedy vyhledáváním ve slovní mřížce ze systému pro automatické rozpoznávání řeči, nebo nějaké její reprezentaci (viz podkapitoly 5.3 a 5.4). Problém se slovy chybějícími ve slovníku se snaží minimalizovat přístup vyhledávání v subslovních mřížkách nebo jejich reprezentacích (viz podkapitola 5.5), případně kombinovaný přístup, používající vyhledávání ve slovních i subslovních mřížkách (viz podkapitola 5.6). V současné době se většina systémů snaží výskyt těchto slov minimalizovat použitím ASR s co největším slovníkem, takzvaného LVSCR[7] systému.

2.4 Využití slepé zpětné vazby

V experimentech s rozšířením dotazu a použitím slepé zpětné vazby ve vyhledávání informací v řeči se ukázalo, že kromě efektu vylepšení vyhledávání stejně jako ve vyhledávání v textu, může použití těchto metod také zároveň výrazně zmenšit vliv chybovosti ASR systému a chybějících slov v jeho slovníku. Slepá zpětná vazba při použití ve vyhledávání informací v řeči může dosáhnout vylepšení výsledků vyhledávání díky tomu, že používá pro rozšíření dotazu slova vyskytující se v prvních vyhledaných dokumentech. Ukázalo se, že při vyhledávání informací v řeči jsou první vyhledané výsledky ty dokumenty, které byly rozpoznány s malou chybovostí [137]. Slepá zpětná vazba využívá informaci o vzájemném výskytu slov v těchto dokumentech a pomocí těchto slov je pak možné vyhledat i dokumenty, kde bylo slovo z dotazu špatně rozpoznáno, ale okolní slova byla rozpoznána správně [4, 81]. Díky využití tohoto kontextu tak mohou být vyhledány i dokumenty rozpoznané s větší chybovostí.

Experimenty ukázaly, že chybovost automatických přepisů řeči i při WER kolem 35% nemá v tomto případě zásadní vliv na výsledky vyhledávání [118, 71, 23, 79, 110], zejména s využitím metod rozšíření dotazu a využití slepé zpětné vazby lze dosáhnout největšího zlepšení výsledků vyhledávání [176, 18, 81].

[7] Large-Vocabulary Continuous Speech Recognition

2.5 Historie vyhledávání informací v řeči

První práce v oblasti vyhledávání informací v řeči se objevovaly od roku 1990 - 1991: Ulrike Glavitsch a Peter Schäuble pracovali na systému pro vyhledávání v rádiových nahrávkách [42, 139]. Později začala na Cambridge University práce na projektu VMR (Video Mail Retrieval) - vyhledávání v řeči z video nahrávek [67]. Na Carnegie Mellon University byl vyvíjen systém pro vyhledávání a získávání multimediálních dat v projektu Infomedia [54, 53]. Cílem tohoto projektu bylo především vyhledávání televizních zpráv dle požadavku uživatele na konkrétní obsah.

V dalších letech (1997 - 2000) se velké množství výzkumných týmů soustředilo v rámci konference TREC/SDR[8] na vývoj a testování algoritmů pro vyhledávání v řečových datech [38]. Cílem jejich práce bylo otestovat do jaké míry ovlivní chyby v automatických přepisech řeči přesnost vyhledávání. Závěrem experimentů bylo, že při dosažení dostatečně malé chybovosti v nejlepším přepisu (kolem 30%) lze dosáhnout přesnosti vyhledávání blížící se vyhledávání v textových datech. Problém vyhledávání informací v řeči byl označen za vyřešený. Výzkum by se měl nadále zaměřovat na oblasti, které se vymykají charakteristikám této úlohy, tedy ASR systémy s velkou chybovostí, vyhledávání pomocí krátkých dotazů a v krátkých dokumentech.

Od roku 2003 do roku 2007 byla součástí kampaně CLEF[9] úloha zabývající se vyhledáváním informací v řečových datech, nejprve to byla úloha CL-SDR[10], která navazovala na výsledky úlohy TREC/SDR. Byla použita kolekce z úlohy TREC-9 a doplněna o dotazy v jiných jazycích pro mezijazykové vyhledávání [35, 34].

V roce 2005 na tuto úlohu navázala CL-SR[11] úloha, která se lišila tím, že vyhledávání probíhalo ve spontánní řeči a data nebyla rozdělena na dokumenty. Byla použita kolekce výpovědí svědků holocaustu [106], získaná z projektu MALACH (viz podkapitola 2.6.5). V roce 2006 a 2007 bylo součástí této úlohy vyhledávání v české kolekci spontánní řeči [62], též získané z projektu MALACH.

V roce 2008 byla v rámci CLEF založena úloha vyhledávání videa, VideoCLEF, z které se poté v roce 2010 vyvinula samostatná úloha MediaEval [80], zaměřující se na uživatelská videa dostupná na internetu. Tato úloha pokračuje až do současné doby, se zaměřením na multimediální obsah na internetu a zejména v sociálních sítích.

Úloha vyhledávání v neformální řeči byla představena v roce 2011 na 9. workshopu NTCIR[12], kde bylo cílem vyhledávat v japonských přednáškách [2].

2.6 Aplikace vyhledávání informací v řeči

Vyhledávání informací v řeči je rozvíjející se obor, dosažené výsledky zatím nevykazují spolehlivost a univerzálnost potřebnou pro reálné komerční aplikace. Přesto lze nalézt velké množství příkladů aplikací, které umožňují uživateli pro konkrétní úlohu vyhledávat v rozsáhlých řečových archívech a tím zpřístupnit jiným způsobem těžko dosažitelné informace.

Následující aplikace jsou příklady využití vyhledávání informací v řeči v různých oblastech: vyhledávání v historických audiovizuálních archivech, v záznamech přednášek,

[8] Text REtrieval Conference / Spoken Document Retrieval - http://trec.nist.gov/
[9] Cross-Language Evaluation Forum - http://clef-campaign.org/
[10] Cross-Language Spoken Document Retrieval
[11] Cross-Language Speech Retrieval
[12] NII Testbeds and Community for Information access Research

v hlasové poště a v televizních a rozhlasových archivech.

2.6.1 SpeechFind

Archivy National Gallery of the Spoken Word[13] [48] obsahují velké množství zvukových záznamů z dvacátého století - projevy vědců i státníků, vysílání zpravodajských relací. Prezentovaný systém SpeechFind[14] bude vyhledávat v online databázi obsahující přes 60 000 hodin zvukových záznamů a umožňovat procházení automatických přepisů nahrávek i poslech vyhledané části nahrávky.

2.6.2 SCANMail

SCANMail je systém umožnující přepis hlasové pošty do textu, jeho prohlížení a vyhledávání v něm [171]. Uživatel si tedy může hlasové zprávy uschovávat a zpětně prohlížet stejně jako klasický email.

2.6.3 Zpracování přednášek MIT

Další zajímavou aplikací vyhledávání informací v řeči je zpracování vysokoškolských přednášek. Například systém pro vyhledávání v přednáškách z MIT[15] [40] umožňuje uživateli vyhledávat požadované části přednášek, prohlížet jejich automatické textové přepisy a vybrané části přednášek si poslechnout.

2.6.4 Speechbot

Speechbot byl online systém pro vyhledávání ve zvukových datech dostupných na internetu [165]: rádiové pořady, zpravodajství, některé záznamy z videokonferencí. Systém vytvářel a indexoval automatické přepisy dostupných nahrávek a formou veřejného internetového vyhledávače umožňoval uživatelům procházet obsah online dostupných zvukových nahrávek. Vývoj a běh tohoto systému byl již společností Hewlett-Packard zastaven.

2.6.5 MALACH

Projekt MALACH[16] se zabýval zpřístupněním kulturního dědictví uchovaného ve formě audiovizuálních záznamů získaných společností Shoah Visual History Foundation. Archiv obsahuje 116 000 hodin výpovědí v 32 jazycích získaných od 52 000 přeživších a svědků holocaustu, v současnosti se jedná o největší ucelený archiv historických audiovizuálních výpovědí. Vzhledem k rozsáhlosti archivu je téměř nemožné spoléhat se na manuální přepis všech nahrávek do textové podoby, jedním z cílů projektu MALACH proto bylo vylepšit stávající metody pro automatické rozpoznávání řeči a efektivní vyhledávání v řečových archivech. Části archivu byly použity jako testovací data v CLEF CL-SR úloze [106], viz podkapitola 5.2.1 a 7.

[13] NGSW - http://www.h-net.org/~ngsw/gallery.html
[14] http://speechfind.utdallas.edu/
[15] http://web.sls.csail.mit.edu/lectures/
[16] Multilingual Access to Large Spoken Archives - http://malach.umiacs.umd.edu/

2.6.6 Televizní a rádiové zpravodajství

Televizní a rádiové stanice vlastní rozsáhlé archivy vysílaných zpravodajských pořadů a reportáží, je tedy přirozené, že velké množství systémů pro vyhledávání informací v řeči pracuje s těmito daty. V TREC/SDR úloze [38] bylo například pracováno s 557 hodinovým korpusem zpravodajských nahrávek [30].

Dalším příkladem je systém vyhledávající v kolekci tureckých nahrávek zpravodajství [109], nebo japonský systém [105].

Kapitola 3

Vyhodnocení výsledků vyhledávání informací

„Odpověď“ vyhledaná systémem na zadaný dotaz nemusí a ve většině případů nebude z mnoha důvodů přesně odpovídat potřebě uživatele. Jedním z těchto důvodů může být nepřesně zadaný dotaz, dalším důvodem může být fakt, že databáze ve které je vyhledáváno neobsahuje přesně tu informaci, kterou uživatel potřebuje. Tyto důvody se promítnou do nepřesnosti výsledků vyhledávání, není však možné je přímo ovlivnit. Další důležitou věcí, která ovlivňuje kvalitu vyhledávání je volba vyhledávacího algoritmu, použití různých metod předzpracování textu nebo řečových dat, různá nastavení metod, atd. Abychom mohli porovnávat různé metody vyhledávání informací, je nutné vhodným způsobem ohodnotit jejich výsledky. Potřebujeme tedy ohodnotit míru relevance nalezené odpovědi vzhledem k zadanému dotazu. Tento druh hodnocení bývá nazýván ohodnocení výkonu vyhledávání.

K ohodnocení použijeme testovací kolekci dokumentů a množinu uživatelských požadavků na získání informace. Tyto informační požadavky jsou pak pro potřeby systému vyhledávání informací vyjádřeny jako dotazy (například množina obsahující všechna slova z uživatelského požadavku, nebo obsahující jen podstatná jména atd.). Dále potřebujeme množinu příslušných relevantních dokumentů, určených specialisty, příslušející ke každému informačnímu požadavku. Ke zvolené strategii vyhledávání tak můžeme určit jak odpovídá množina odpovědí, získaných aplikací této strategie, relevantním dokumentům určeným specialisty.

Nejčastěji používané míry jsou *míra přesnosti* a *míra úplnosti*.

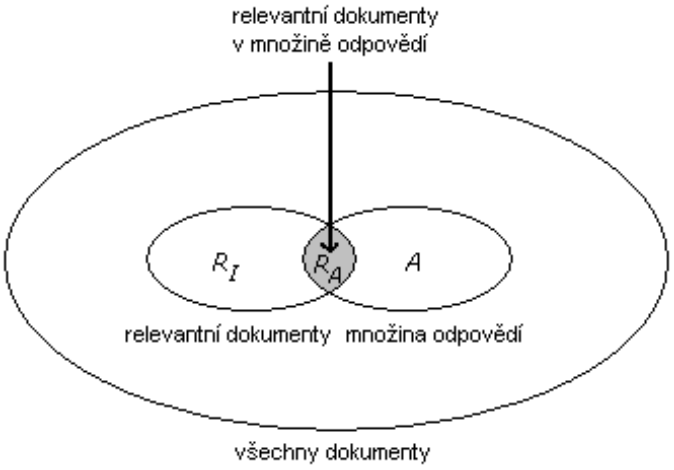
3.1 Míra přesnosti a úplnosti

Je dán dotaz q a k němu příslušná množina relevantních dokumentů R_q. Použitím určité strategie vyhledávání získáme množinu odpovědí A. Z takto získaných odpovědí vybereme ty relevantní a množinu označíme R_A. Množiny jsou znázorněny na obrázku 3.1.

3.1.1 Míra přesnosti

Míra přesnosti[1] P - odpovídá relativní četnosti relevantních dokumentů mezi nalezenými odpověďmi:

[1] Precision



Obrázek 3.1: Množiny dokumentů

$$P = \frac{|R_A|}{|A|}. \tag{3.1}$$

3.1.2 Míra úplnosti

Míra úplnosti[2] R - odpovídá relativní četnosti nalezených relevantních dokumentů mezi všemi relevantními dokumenty:

$$R = \frac{|R_A|}{|R_q|}. \tag{3.2}$$

Míry můžeme také chápat jako pravděpodobnosti. P vyjadřuje pravděpodobnost, že nalezený dokument je relevantní. R je pravděpodobnost, že byl nalezen relevantní dokument. Obrázek 3.2 ukazuje reálný funkční vztah mezi přesností a úplností pro jeden dotaz, tyto dvě míry jdou z principu proti sobě, podle požadavků úlohy vhodně vyvažujeme nastavení metody preferující vyšší přesnost nebo úplnost.

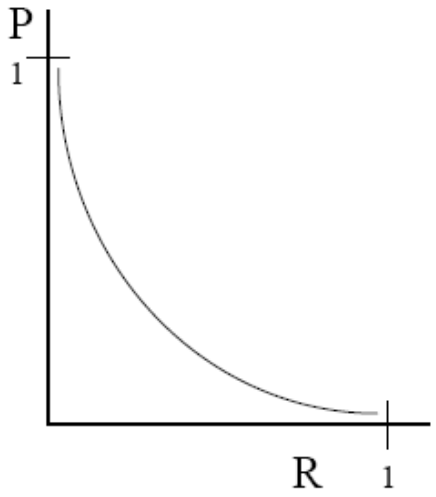
3.1.3 Postupné vyhodnocení přesnosti a úplnosti

Abychom mohli míry přesně vyjádřit, zakreslíme křivku závislosti míry přesnosti na úplnosti. Tato křivka většinou obsahuje vypočtené hodnoty míry přesnosti pro jedenáct standardních úrovní míry úplnosti: 0%, 10%, 20%, 30%, ..., 100%. Pokud nemáme hodnoty míry přesnosti určeny přesně v těchto úrovních, získáme je interpolací. Hodnoty ve standardních úrovních míry úplnosti dopočteme takto:

$$P(r_i) = \max_{r_i \leq r \leq r_{i+1}} P(r), \tag{3.3}$$

kde r_i je standardní úroveň míry úplnosti, $i \in 0, 1, 2, ..., 10$ a $P(r_i)$ je hodnota P v úrovni r_i.

[2] Recall



Obrázek 3.2: Vztah mezi přesností a úplností

Tedy dopočtená míra přesnosti v každé standardní úrovni míry úplnosti odpovídá maximální známé hodnotě míry přesnosti v jakékoli úrovni míry úplnosti mezi danou a následující standardní úrovní míry úplnosti.

Průměr z více dotazů

Takto bychom určili křivku pro jediný dotaz. Abychom však mohli dobře ohodnotit použitý algoritmus, musíme otestovat více dotazů. Pro každý dotaz získáme vlastní křivku. Abychom mohli jednotlivé algoritmy porovnávat, musíme ze všech křivek pro jednotlivé dotazy získat průměrnou křivku. Ve všech standardních úrovních míry úplnosti vypočteme průměrné hodnoty míry přesnosti $\bar{P}(r)$ takto:

$$\bar{P}(r) = \sum_{i=1}^{n} \frac{P_i(r)}{n}, \tag{3.4}$$

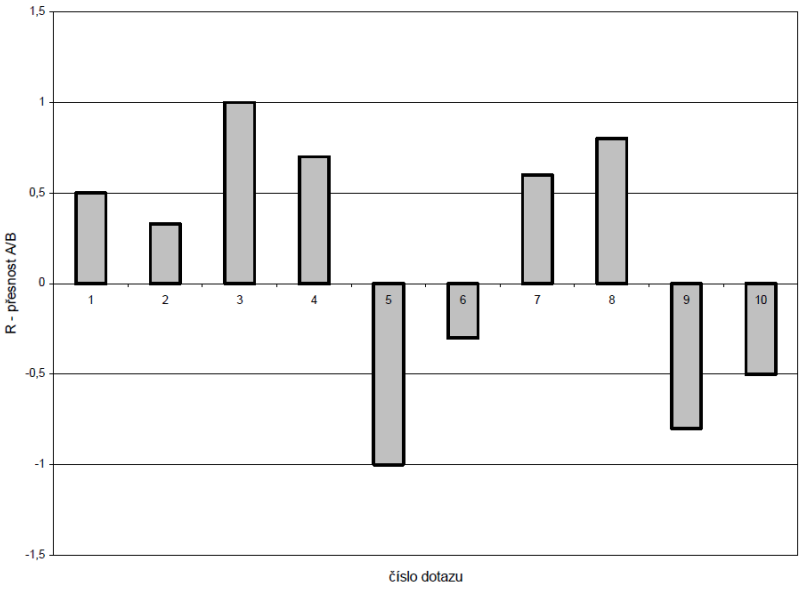
kde n je počet dotazů a $P_i(r)$ je hodnota míry přesnosti v úrovni r pro dotaz i.

3.2 R-přesnost

Hodnota míry přesnosti se počítá po získání R dokumentů, kde R je celkový počet relevantních dokumentů k danému dotazu, tedy $|R_q|$. Pokud máme například deset relevantních dokumentů k danému dotazu, spočteme míru přesnosti pro prvních deset získaných dokumentů (podle hodnocení). Mezi těmito dokumenty bude například pět relevantních dokumentů, hodnota míry R-přesnosti tedy bude 0,5.

3.2.1 Histogramy míry přesnosti

Ke srovnání schopností vyhledávání dvou algoritmů při uvažování několika dotazů můžeme použít míru R–přesnosti. Definujme rozdíl



Obrázek 3.3: Histogram R-přesnosti pro 10 dotazů

$$RP_{A/B}(i) = RP_A(i) - RP_B(i), \tag{3.5}$$

kde $RP_A(i)$ je hodnota míry R–přesnosti pro dotaz i algoritmu A a $RP_B(i)$ je hodnota míry R–přesnosti pro dotaz i algoritmu B. Pokud je $RP_{A/B}(i)$ rovno nule, pak oba algoritmy jsou stejně výkonné pro dotaz i. Kladná hodnota ukazuje lepší výkon algoritmu A, záporná algoritmu B. Na obrázku 3.3 je ukázka histogramu pro deset dotazů. Je vidět, že algoritmus A byl lepší pro šest dotazů (dotaz č. 1, 2, 3, 4, 7, 8) a algoritmus B pro zbývající.

3.3 Míra F

F-míra[3], někdy také F_1-míra vyjadřuje harmonický průměr kombinující míry přesnosti a úplnosti. Vypočteme ji takto:

$$F = 2\frac{P \cdot R}{P + R}. \tag{3.6}$$

F-míra dosahuje hodnot v intervalu $\langle 0, 1 \rangle$, hodnota 0 znamená, že nebyly získány žádné relevantní dokumenty. Hodnota 1 vyjadřuje, že všechny získané dokumenty jsou relevantní. Vysokých hodnot dosahuje tato míra pouze pokud míra přesnosti i úplnosti je vysoká, určení maximální hodnoty F-míry může sloužit k nalezení nejlepšího kompromisu mezi mírou úplnosti a přesnosti.

[3] F-measure

3.4 Míra E

Tato míra také kombinuje míru úplnosti a přesnosti. Byla představena v knize [120]. Uživatel si může stanovit, jestli je pro něj důležitější přesnost nebo úplnost. Míra je definována takto:

$$E = 1 - \frac{1 + b^2}{\frac{b^2}{R} + \frac{1}{P}}, \tag{3.7}$$

kde b je parametr určující důležitost přesnosti nebo úplnosti oproti druhé míře. Pro $b = 1$ je E doplněk k F, $b > 1$ značí větší důležitost přesnosti, $b < 1$ větší důležitost úplnosti.

3.5 Průměrná přesnost

Průměrná hodnota míry přesnosti se počítá postupně z hodnot míry přesnosti při nalezení každého relevantního dokumentu. Tedy při nalezení prvního relevantního dokumentu spočteme míru přesnosti, stejně tak při nalezení druhého, třetího, atd. Z těchto hodnot vypočteme průměr. Tato míra lépe ohodnocuje systémy, které nalézají relevantní dokumenty na počátku hodnocení. Toto dobré ohodnocení však neznamená dobrou celkovou míru úplnosti. Průměrná přesnost se spočte:

$$AveP = \frac{\sum_{i=1}^{N}(P(i) \times \mathrm{rel}(i))}{|R_q|}, \tag{3.8}$$

kde i je pořadí nalezeného dokumentu, $\mathrm{rel}(i)$ je binární funkce určující, zda je dokument relevantní a $P(i)$ je přesnost na daném pořadí i.

3.6 MAP

V současné době je MAP[4], střední průměrná přesnost, asi nejčastěji používaná míra pro ohodnocení výsledků vyhledávání:

$$MAP = \frac{\sum_{q=1}^{|Q|} AveP(q)}{|Q|}. \tag{3.9}$$

MAP se počítá pro celou množinu dotazů Q a vyjadřuje průměr z průměrné přesnosti každého dotazu.

3.7 Další míry

V některých textech se mohou objevit také další méně používané způsoby vyhodnocení vyhledávání informací. Jde zejména o míry používané v oblasti binární klasifikace textu, které jsou přizpůsobeny pro použití ve vyhledávání informací. Můžeme si definovat následující tabulku 3.1 vztahů mezi nalezenými dokumenty a relevantními / nerelevantními dokumenty:

[4] Mean Average Precision

Tabulka 3.1: Označení možných kombinací vztahů mezi relevancí dokumentu a tím, zda byl vyhledán

	Relevantní	Nerelevantní
Nalezený	pravdivě pozitivní (pp)	falešně pozitivní (fp)
Nenalezený	falešně negativní (fn)	pravdivě negativní (pn)

Míru přesnosti a úplnosti si pak můžeme alternativně vyjádřit jako:

$$P = \frac{pp}{(pp + fp)} \tag{3.10}$$

a

$$R = \frac{pp}{(pp + fn)}. \tag{3.11}$$

3.7.1 Míra správnosti

Alternativou může být míra správnosti[5] *Acc*, která určuje podíl klasifikací (relevantní / nerelevantní) vyhledávacího systému, které jsou správné:

$$Acc = \frac{(pp + pn)}{(pp + fp + pn + fn)}. \tag{3.12}$$

Systém pro vyhledávání informací je tedy chápán jako binární klasifikátor určující pro každý dokument, zda patří do třídy relevantních nebo nerelevantních dokumentů. Použití této míry pro vyhodnocení úspěšnosti vyhledávání je ale ve většině případů nevhodné. Ve vyhledávání informací bývá velmi nerovnoměrné rozložení četnosti dokumentů v jednotlivých třídách, nerelevantních dokumentů je velké množství, oproti malému množství relevantních dokumentů. Pokud tedy budeme hledat nejlepší metodu z hlediska míry správnosti, může nám jako taková vyjít metoda označující všechny dokumenty jako nerelevantní.

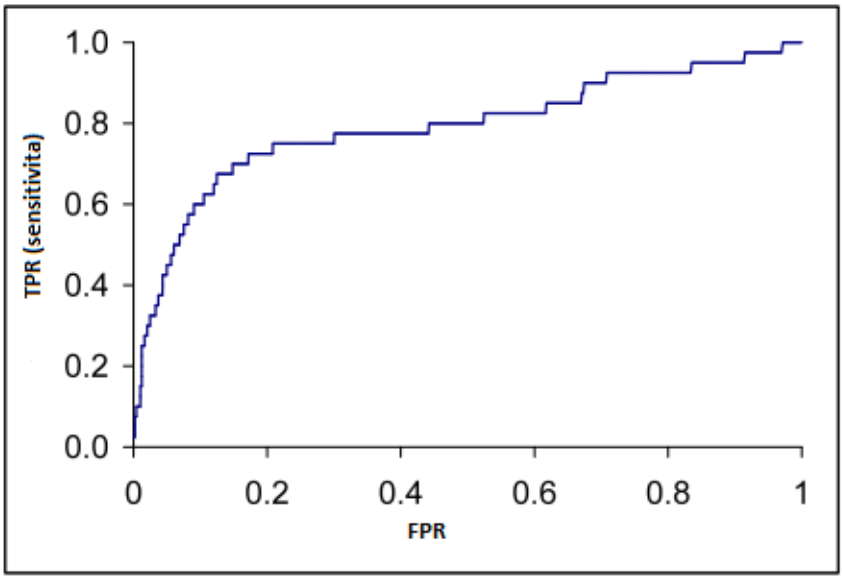
3.7.2 ROC křivka

ROC[6] křivka vyjadřuje vztah mezi podílem pravdivě pozitivních výsledků TPR[7] (tedy mírou úplnosti), někdy také nazývaným jako *sensitivita* a podílem falešně pozitivních výsledků FPR[8], který je definován jako:

$$FPR = \frac{fp}{(fp + pn)}. \tag{3.13}$$

Ukázka průběhu ROC křivky je vidět na obrázku 3.4. Použití ROC křivky pro vyhodnocení úspěšnosti vyhledávání není příliš vhodné ze stejných důvodů jako u míry správnosti. Vzhledem k velkému množství pravdivě negativních výsledků bude ROC křivka velmi podobná pro dva systémy se stejnou mírou úplnosti a rozdílnou mírou přesnosti [33]. Pro

[5] Accuracy
[6] Receiver Operating Characteristic
[7] True Positive Rate
[8] False Positive Rate



Obrázek 3.4: Ukázka ROC křivky vyjadřující vztah mezi TPR - podílem pravdivě pozitivních výsledků a FPR - podílem falešně pozitivních výsledků (převzato z [97])

porovnání výsledků vyhledávání při použití jednotlivých metod je tedy lepší použít křivku závislosti míry přesnosti na míře úplnosti (viz podkapitola 3.1.3).

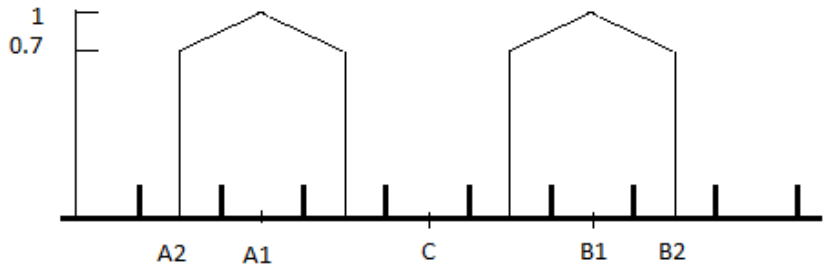
3.8 Vyhodnocení výsledků vyhledávání v řečových datech

Míry vyhodnocení výsledků vyhledávání uvedené v předchozím textu byly definovány pro vyhledávání informací v textu. Pokud se budeme zabývat vyhledáváním v řečových datech, můžeme tyto míry použít ve stejné podobě za předpokladu, že určitým způsobem definujeme podobu dokumentu. Pokud například budeme vyhledávat v archivu nahrávek a každou tuto nahrávku označíme jako jeden dokument, můžeme použít k vyhodnocení výsledků například míry přesnosti a úplnosti stejně jako u textových dat.

Pokud budou data obsahovat velmi dlouhé nahrávky, je někdy vhodné je rozdělit na kratší úseky, například podle tematického obsahu, a ty teprve označit jako dokumenty. Vyhledávací algoritmus pak bude vyhledávat přímo tyto dokumenty, a nebo v případě, že hranice tematických úseků budou známy pouze pro vyhodnocení výsledků, ale ne při běhu algoritmu, může algoritmus jako výsledek poskytovat čas v nahrávce, kde předpokládá dané téma. Jako správný výsledek pak bude hodnocen ten čas, který patří do intervalu relevantního úseku [38].

Tyto přístupy ale požadují manuální rozdělení nahrávky na úseky, které je časově i finančně velmi náročné. Další možností je hodnotit výsledky vyhledávání na základě přesnosti s jakou dokáže algoritmus nalézt počátek úseku, který tematicky odpovídá potřebě uživatele - dotazu. Takové míry jsou označovány jako jednostranné.

Pokud data, ve kterých má systém vyhledávat, sestávají z dlouhých řečových nahrávek, není pro uživatele systému možné při vyhledání dokumentu (nahrávky) jednoduše a rychle tuto nahrávku prohlédnout, zda odpovídá jeho požadavku. Ani při poskytnutí jejího přepisu do textu to není většinou pro uživatele příjemné, protože automatické textové přepisy



Obrázek 3.5: Ukázka penalizační funkce při výpočtu GAP

často obsahují velké množství chyb. Řešením je poskytnout uživateli pouze počáteční čas v nahrávce, kde začíná hledané téma.

3.8.1 GAP

Míra GAP[9] [87], tedy zobecněná průměrná přesnost, je založená na započtení přesnosti nalezení správného počátku relevantního úseku. Dobrý vyhledávací systém by měl poskytovat výsledky co nejblíže stanovenému času počátku relevantního úseku a zároveň řadit tyto úseky mezi prvními vyhledanými výsledky.

Míra je tedy nastavena tak, aby hodnotila co nejlepší pořadí nalezených relevantních úseků, ale zároveň penalizovala časovou odchylku od stanoveného začátku úseku. Zároveň je také možné pro každý relevantní úsek započítat pouze první výskyt v nalezených dokumentech, ostatní, i pokud by přesněji odpovídali definovanému počátečnímu času relevantního úseku, nejsou započítány. Jednotkou hrubosti byl zvolen čas 15 vteřin. Míra GAP je definována takto:

$$GAP = \frac{\sum_{R_k \neq 0} P(k)}{|R_q|}, \tag{3.14}$$

kde $|R_q|$ je počet relevantních úseků, $P(k)$ je přesnost na pořadí k a R_k je skóre vypočtené na základě penalizační funkce na pořadí k:

$$P(k) = \frac{\sum_{i=1}^{k} R_i}{k}. \tag{3.15}$$

Penalizační funkce může být zvolena různě, příklad je vidět na obrázku 3.5, kde $A1$, $A2$, $B1$, $B2$, C jsou systémem vyhledané počátky úseků. $A1$ a $B1$ jsou schodné s počátkem relevantního úseku, dostanou tedy nejvyšší ohodnocení, $A2$ a $B2$ nižší a C žádné.

[9] Generalized Average Precision

3.8.2 mGAP

Míra mGAP[10] vyjadřuje podobně jako míra MAP (viz podkapitola 3.6) průměr ze zobecněné průměrné přesnosti každého dotazu, počítá se pro celou množinu dotazů Q:

$$mGAP = \frac{\sum_{q=1}^{|Q|} GAP(q)}{|Q|}. \tag{3.16}$$

[10] Mean GAP

Kapitola 4

Metody vyhledávání informací

V této kapitole budou představeny nejpoužívanější metody pro vyhledávání informací v textu. Tyto metody se také používají pro vyhledávání informací v řeči, ve stejné podobě jako pro vyhledávání v textu, pokud budeme vyhledávat v nejlepších přepisech ze systému pro automatické rozpoznávání řeči, případně v upravené podobě, pokud budeme vyhledávat ve slovních mřížkách nebo jejich reprezentacích.

4.1 Model pro vyhledávání informací

Obecné schéma systému pro vyhledávání informací je vidět na obrázku 4.1. Abychom mohli vytvořit model pro vyhledávání informací, musíme nejprve zvolit určitou reprezentaci dokumentů a uživatelských požadavků - dotazů. Poté zvolíme vhodnou strukturu modelu, ta také určuje způsob vytvoření hodnotící funkce. Přestože každý z modelů má své specifické vlastnosti, je vhodné uvést definici obecného modelu pro vyhledávání informací.

4.1.1 Definice modelu pro vyhledávání informací

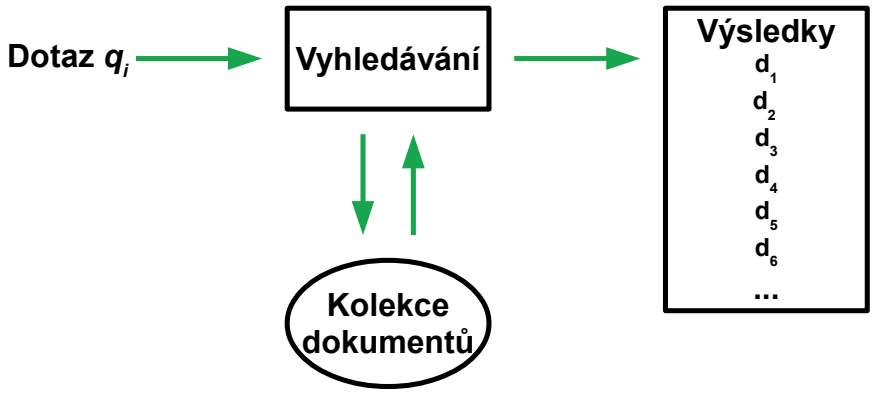
Model pro vyhledávání informací je čtveřice

$$[\mathbf{D}, \mathbf{Q}, F, R_{q,d_j}], \tag{4.1}$$

kde $\mathbf{D}$ je množina reprezentací všech obsažených dokumentů, $\mathbf{Q}$ je množina dotazů, F je struktura pro modelování reprezentací dokumentů a formulací dotazů a R_{q,d_j} je hodnotící funkce, která přiřazuje reálné číslo dotazu $q \in \mathbf{Q}$ a dokumentu $d_j \in \mathbf{D}$. Podle tohoto ohodnocení jsou seřazeny dokumenty příslušející k dotazu q.

4.1.2 Term

Abychom mohli popsat jednotlivé modely, musíme vymezit pojem term. Pojmem term označujeme klíčové slovo dokumentu, obvykle to bývá podstatné jméno. Obecně může být term jakékoli slovo z dokumentu. Dokument je vymezen svou množinou termů. Dotaz je poté také formulován pomocí termů. Každý z termů dokumentu může být pro daný dokument jinak důležitý. Některé slovo může obsahovat každý dokument, takové slovo nebude použitelné pro vyhledání konkrétního dokumentu. Naopak některá slova bude obsahovat pouze malá skupina dokumentů, tato slova jsou pro dokumenty charakteristická a jsou vhodná jako termy.



Obrázek 4.1: Schéma systému pro vyhledávání informací

4.1.3 Ohodnocení termů

Abychom mohli odlišit důležitost jednotlivých termů dokumentu pro daný dokument, přiřadíme jim číselné váhy. Označme t_i jednotlivé termy ze všech použitých termů v systému, d_j dokument a $w_{i,j} > 0$ jako váhu, přidělenou dvojici (t_i, d_j). Potom

$$w_{i,j} \begin{cases} = 0 & \text{pokud } d_j \text{ neobsahuje } t_i \\ > 0 & \text{pokud } d_j \text{ obsahuje } t_i \end{cases} \tag{4.2}$$

Takto vytvořené ohodnocení termů nerespektuje jejich možnou vzájemnou závislost. Tedy ohodnocení jednoho termu nám neříká nic o ohodnocení termu jiného. Ve skutečnosti jsou však termy v dokumentech často vzájemně závislé, například u slovních spojení.

4.2 Booleovský model

Booleovský model je nejstarší model pro vyhledávání informací. Jeho základy vznikly v padesátých letech dvacátého století. Přesto je to stále používaný model, například v knihovnických systémech. Model je založen na teorii množin a booleovské algebře. Dotazy jsou formulovány pomocí booleovských výrazů, tím je dána jasná forma dotazu. V dotazech je možno použít několik operátorů mezi termy, jak ukazuje tabulka 4.1.

Tabulka 4.1: Tabulka hodnot booleovských operátorů mezi dvěma termy

A	B	A AND B	A OR B	NOT A
1	1	1	1	0
1	0	0	1	0
0	1	0	1	1
0	0	1	0	1

Význam operátorů při vyhledávání lze popsat takto: **A AND B** – najdi dokument obsahující jak term A tak term B, **A OR B** – najdi dokument obsahující term A nebo term B, **NOT A** – najdi dokument neobsahující term A. V booleovském modelu se pouze

hodnotí, zda termy jsou nebo nejsou přítomné v dokumentu. Váhy přidělené jednotlivým termům jsou tedy pouze binární, $w_{i,j} \in \{0, 1\}$. Dotaz q je booleovský výraz a jako takový může být převeden do normální disjunktní formy. Normální disjunktní forma se skládá z disjunkcí elementárních konjunkcí, takzvaných mintermů. Každý z těchto mintermů je vektor binárních vah, vyjadřující pro všechny termy, zda musí (1) nebo nesmí (0) být obsaženy v hledaném dokumentu. Zda dokument odpovídá dotazu určíme takto:

$$sim_{d_j,q} = \begin{cases} 1 & \text{pokud existuje minterm } m \text{ takový, že } \forall t_i \; g_i(d_j) = g_i(m) \\ 0 & \text{v ostatních případech} \end{cases} \tag{4.3}$$

Kde $sim_{d_j,q}$ je podobnost dokumentu d_j a dotazu q a g_i je funkce, která vrací váhu asociovanou k termu t_i ve vektoru d_j. Pokud $sim_{d_j,q} = 1$, pak dokument odpovídá podle booleovského modelu dotazu q, pokud je rovna 0, pak dokument neodpovídá. Booleovský model pouze určuje, zda dokument je a nebo není relevantní, není tudíž možné seřadit získané dokumenty podle relevance. Model také neurčuje zda dokument odpovídá dotazu pouze částečně. Například pokud by dokument obsahoval všechny kromě jednoho hledaného termu, nebude vyhodnocen jako relevantní.

4.2.1 P-Norm model

P-Norm model je jedním z modelů označovaných jako rozšířené booleovské modely pro vyhledávání informací, představený v práci [134]. Spojuje vlastnosti booleovského modelu, tedy strukturované dotazy, a vektorového modelu, tedy použití vah termů. Model pracuje s operátory AND a OR stejně jako klasický booleovský model, ale jejich vyhodnocení není tak striktní.

Model P-Norm používá L_p vektorovou normu na měření vzdáleností. Pokud budeme uvažovat množinu termů t_1, t_2, t_3,..., t_n, můžeme označit $w_{i,j}$ váhu termu t_i v dokumentu d_j a $w_{i,q}$ váhu termu t_i v dotazu q. Podobnost dotazu a dokumentu je definována:

$$sim\left(q_{t_1 \mathrm{OR} t_2}, d_j\right) = \left(\frac{w_{1,q}^p w_{1,j}^p + w_{2,q}^p w_{2,j}^p}{w_{1,q}^p + w_{2,q}^p} \right)^{1/p}, \tag{4.4}$$

$$sim\left(q_{t_1 \mathrm{AND} t_2}, d_j\right) = 1 - \left(\frac{w_{1,q}^p \left(1 - w_{1,j}\right)^p + w_{2,q}^p \left(1 - w_{2,j}\right)^p}{w_{1,q}^p + w_{2,q}^p} \right)^{1/p}. \tag{4.5}$$

Změnou hodnoty parametru p od 1 do ∞ můžeme měnit chování modelu od čistého vektorového modelu ($p = 1$), popsaného níže, až ke konvenčnímu booleovskému modelu ($p = \infty$, binární váhy termů). Váhy termů je vhodné stanovit stejně jako ve vektorovém TF-IDF modelu (viz podkapitola 4.3.1).

Problémem tohoto modelu je požadavek strukturovaných dotazů, tedy použití operátorů AND a OR. V běžném případě jsou dotazy ve formě klíčových slov, věty, nebo krátkého textu, je tedy nutné dotazy nějakým způsobem převést na potřebnou formu, nejlépe automaticky. Metoda pro automatické vytváření booleovských dotazů byla vytvořena v rámci této práce (publikována v práci [152]) a je představena v podkapitole 7.5.1.

4.3 Vektorový model

Vektorový model [136, 135] je jedním z nejznámějších a stále ještě nejrozšířenějších modelů pro vyhledávání informací. Jako vektorové modely se dá označit celá skupina metod, které mají společný základ: skládají se z indexační funkce, která přiřadí dokumentu určitý vektor vah, a vyhledávací funkce, která pomocí porovnání vektorů vah dokumentu a dotazu přiřadí každému dokumentu určité ohodnocení.

Dokument d_j a dotaz q jsou reprezentovány vektorem vah jednotlivých svých termů t_i:

$$d_j = (w_{1,j}, w_{2,j}, ..., w_{n,j})q \quad = (w_{1,q}, w_{2,q}, ..., w_{n,q}), \tag{4.6}$$

kde $w_{i,j}$ je váha termu t_i v dokumentu d_j, $w_{i,j} \geq 0$.

Dokumenty relevantní k dotazu získáme pomocí přiřazení určitého koeficientu podobnosti mezi dokumentem a dotazem. Pro výpočet podobnosti se používá například cosinus úhlu mezi vektory:

$$sim_{d_j,q} = \frac{d_j \cdot q}{\|d_j\| \, \|q\|} = \frac{\sum_{i=1}^{t} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}, \tag{4.7}$$

$sim_{d_j,q} \in \langle 0, 1 \rangle$, při $sim_{d_j,q} = 1$ jsou dokument a dotaz shodné. Relevantní dokumenty jsou tedy ty dokumenty, které mají nejvyšší koeficient podobnosti $sim_{d_j,q}$.

Hlavní nevýhodou tohoto modelu je předpoklad, že termy v dokumentech jsou vzájemně nezávislé. Vektorový model tedy nezahrnuje do výpočtu podobnosti žádným způsobem blízkost termů, tedy například slovních spojení nebo frází.

4.3.1 TF-IDF model

Nejrozšířenějším způsobem, jak stanovit váhu $w_{i,j}$, je takzvaný TF-IDF[1] model, který vypočítává váhu termu z jeho frekvence tf_{t_i,d_j} a inverzní frekvence dokumentu idf_{t_i}. Základní myšlenkou tohoto modelu je, že term je tím důležitější (má větší váhu $w_{i,j}$), čím častěji se vyskytuje v dokumentu d_j a zároveň čím méně často se vyskytuje v ostatních dokumentech v kolekci:

$$w_{i,j} = tf_{t_i,d_j} \cdot idf_{t_i}. \tag{4.8}$$

Frekvence termu v dokumentu tf_{t_i,d_j} odpovídá nejčastěji počtu výskytů termu t_i v dokumentu d_j, bývá označeno jako f_{t_i,d_j}. Další možnosti jak tf_{t_i,d_j} stanovit [97]:

- Binární váha $tf_{t_i,d_j} \in \{0, 1\}$, podobně jako v booleovském modelu v podkapitole 4.2
- Logaritmická frekvence termu $tf_{t_i,d_j} = 1 + \log(f_{t_i,d_j})$
- Normalizovaná frekvence termu $tf_{t_i,d_j} = a + a(1-a)\frac{f_{t_i,d_j}}{f_{max}(d_j)}$, kde a je nastaveno na hodnotu mezi 0 a 1, většinou bývá $a = 0,4$ a $f_{max}(d_j)$ je maximální frekvence nějakého termu v dokumentu d_j.

[1] Term Frequency - Inverse Document Frequency

Inverzní frekvence dokumentu idf_{t_i} pro term t_i odpovídá inverzi frekvence, s jakou se term vyskytuje v jednotlivých dokumentech kolekce [68]:

$$idf_{t_i} = \log \frac{N}{n_i}. \tag{4.9}$$

N je celkový počet dokumentů v kolekci a n_i je počet dokumentů, ve kterých se vyskytuje term t_i. Někdy se používají také jiné varianty idf_{t_i}, například:

- Vyhlazená inverzní frekvence dokumentu $idf_{t_i} = \log(1 + \frac{N}{n_i})$
- Pravděpodobnostní inverzní frekvence dokumentu $idf_{t_i} = \log \frac{N - n_i}{n_i}$

Používají se také různé varianty TF-IDF modelu, například se jako váha termu použije pouze frekvence termu tf_{t_i,d_j} [90], nebo jen idf_{t_i} (často se používá pro ohodnocení termu v dotazu). Více například v práci [132].

4.4 Pravděpodobnostní model

Pravděpodobnostní modely pro vyhledávání informací se začaly vyvíjet jako alternativa k vektorovým modelům na konci sedmdesátých let dvacátého století [120, 122]. Jejich princip je založen na odhadu pravděpodobnosti s jakou se term t vyskytuje v relevantních dokumentech a podle toho poté určují, zda dokument obsahující tento term je relevantní nebo ne. Definujme si funkci R_{q,d_j}, která nabývá hodnoty 1 pro relevantní dokumenty, 0 v opačném případě. Dokumenty by tedy měly být řazeny podle odhadnuté pravděpodobnosti $P(R_{q,d_j} = 1|d_j, q)$ [120].

Jedním z tradičních způsobů jak odhadnout pravděpodobnost $P(R_{q,d_j} = 1|d_j, q)$ je *Binary Independence Model (BIM)*, který představili Stephen E. Robertson a Karen Spärck Jones [122], detaily tohoto modelu jsou dále rozvinuty v práci [120]. Dokumenty a dotaz jsou stejně jako ve vektorovém modelu zastoupeny vektory vah svých termů (viz rovnice (4.6)), váhy termů jsou pouze binární. Pravděpodobnost $P(R_{q,d_j} = 1|d_j, q)$ je modelována pomocí pravděpodobnosti vektoru výskytu termů $P(R_{q,d_j} = 1|x, q)$. Pomocí Bayesova pravidla získáme:

$$\begin{aligned} P(R_{q,d_j} = 1|x, q) &= \frac{P(x|R_{q,d_j} = 1)P(R_{q,d_j} = 1|q)}{P(x|q)} \\ P(R_{q,d_j} = 0|x, q) &= \frac{P(x|R_{q,d_j} = 0)P(R_{q,d_j} = 0|q)}{P(x|q)}, \end{aligned} \tag{4.10}$$

kde $P(x|R_{q,d_j} = 1)$ a $P(x|R_{q,d_j} = 0)$ jsou pravděpodobnosti, že pokud dokument (relevantní, respektive nerelevantní) je nalezen, pak jeho reprezentace je x. Uvedené pravděpodobnosti je nutné odhadnout pomocí statistik výskytu termů v kolekci. Apriorní pravděpodobnosti $P(R_{q,d_j} = 1|q)$ a $P(R_{q,d_j} = 0|q)$ nalezení relevantního, respektive nerelevantního dokumentu je také nutné odhadnout (pokud například známe podíl relevantních vůči nerelevantním dokumentům v kolekci). Podobnost dokumentu a dotazu lze tedy vyjádřit takto [133]:

$$\begin{aligned} sim_{d_j,q} &= \sum_{t_i \in d} t_i \log \frac{p_i(1-u_i)}{u_i(1-p_i)} + c \\ p_i &= P(x_i = 1 | R_{q,d_j} = 1) \\ u_i &= P(x_i = 1 | R_{q,d_j} = 0), \end{aligned} \tag{4.11}$$

kde c reprezentuje doplňující konstanty a hodnoty p_i a u_i pravděpodobnosti, že term t_i se vyskytuje v relevantním, respektive nerelevantním dokumentu. Tyto hodnoty je nutné odhadnout. Často se při neznalosti informace o relevanci dokumentů (například ze zpětné vazby, viz podkapitola 6.1.2) používá $p_i = 0.5$ a u_i se nastavuje jako podíl dokumentů, které obsahují term t_i, v celé kolekci:

$$u_i = \frac{n_i}{N}, \tag{4.12}$$

kde N je celkový počet dokumentů v kolekci a n_i je počet dokumentů, ve kterých se vyskytuje term t_i. Více k tomuto modelu a další možnosti jak odhadnout jednotlivé pravděpodobnosti je možné nalézt v pracích [120, 122, 125, 133, 97]

4.4.1 Okapi BM25

BM25[2] je často používaná vyhledávací funkce založená na pravděpodobnostním modelu *BIM*[3]. Tato funkce byla velice často používána v různých úpravách v TREC úlohách, poprvé byla představena na TREC-3 [123]. Existuje mnoho variant této funkce, jedna z nejpoužívanějších je například tato:

$$sim_{d_j,q} = \sum_{t_i \in q} idf_{t_i} \frac{tf_{t_i,d_j}(k_1+1)}{tf_{t_i,d_j} + k_1(1-b+b\frac{L_{d_j}}{L_{avg_d}})}, \tag{4.13}$$

kde L_{d_j} je délka dokumentu d_j a L_{avg_d} je průměrná délka dokumentů v kolekci. Parametry k_1 a b nastavují chování funkce BM25, k_1 ovlivňuje váhu tf_{t_i,d_j}, $k_1 \geq 0$, a parametr b ovlivňuje váhu délky dokumentu, $b \in \langle 0, 1 \rangle$.

Inverzní frekvenci dokumentu idf_{t_i} lze spočítat podle rovnice (4.9), často se také počítá podle vzorce

$$idf_{t_i} = \log \frac{N - n_i + 0,5}{n_i + 0,5}, \tag{4.14}$$

kde N je počet dokumentů v kolekci a n_i je počet dokumentů obsahujících term t_i. Takto počítaná idf_{t_i} je záporná, pokud se term vyskytuje ve více než půlce dokumentů, čemuž lze předejít použitím stop listu [97].

Pokud bude dotaz q dostatečně dlouhý, je možné též započítat frekvence termů v dotazu $tf_{t,q}$. Rovnici (4.13) upravíme takto:

$$sim_{d_j,q} = \sum_{t_i \in q} idf_{t_i} \frac{tf_{t_i,d_j}(k_1+1)}{tf_{t_i,d_j} + k_1(1-b+b\frac{L_{d_j}}{L_{avg_d}})} \frac{tf_{t_i,q}(k_3+1)}{tf_{t_i,q} + k_3}, \tag{4.15}$$

[2] BM - Best Match
[3] Binary Independence Model

kde k_3 je parametr udávající váhu započtení frekvence termu v dotazu $tf_{t,q}$. Parametry k_1, k_3 a b by v nejlepším případě měly být nastaveny pomocí optimalizace provedené na testovací kolekci. Pokud to není možné, bylo experimentálně zjištěno, že je vhodné volit hodnoty parametrů $k_1 \in \langle 1, 2; 2\rangle$, $k_3 = 7$ nebo $k_3 = 1000$ pro dlouhé dokumenty a parametr $b = 0,75$ [126]. Detailní popis vývoje tohoto modelu a rozsáhlé experimenty lze nalézt v práci [69].

4.5 Jazykové modelování

Základní myšlenkou přístupu k vyhledávání informací pomocí jazykového modelování je princip, že slova v dotazu by měla být často se vyskytujícími slovy z dokumentu, který je k dotazu relevantní. Pomocí představy jazykového modelu to můžeme vyjádřit i tak, že budeme hledat dokument, který by s velkou pravděpodobností mohl vygenerovat zadaný dotaz.

Přístup k vyhledávání informací přes jazykové modelování byl představen v práci [113] a [112]. Ponte a Croft zde navrhli nový přístup k hodnocení podobnosti dotazu a dokumentu, tento přístup je v současné době nazýván *Query likelihood model* (model věrohodnosti dotazu) a je základní a nejběžnější formou využití jazykového modelování při vyhledávání dokumentů [97]. Ponte a Croft ve své práci [112] ukazují, že jazykové modelování může dosáhnout lepších výsledků než TF-IDF model.

4.5.1 Query likelihood model

Princip tohoto přístupu spočívá ve vytvoření jazykového modelu θ_{d_j} z každého dokumentu d_j v kolekci a poté hledáme dokument s největší pravděpodobností v závislosti na dotazu $P(d_j|q)$. Pomocí Bayesova pravidla získáme:

$$P(d_j|q) = \frac{P(q|d_j)P(d_j)}{P(q)}. \tag{4.16}$$

Apriorní pravděpodobnost dotazu $P(q)$ je stejná pro všechny dokumenty. Apriorní pravděpodobnost dokumentu $P(d_j)$ může být nastavena v závislosti na autoru dokumentu, jeho tématu, datu vydání atd., ale v běžném případě je stejná pro všechny dokumenty. $P(q)$ i $P(d_j)$ můžeme tedy vynechat. Zbývá nám $P(q|d_j)$ respektive $P(q|\theta_{d_j})$, tedy pravděpodobnost s jakou mohl být dotaz q vygenerován z jazykového modelu dokumentu θ_{d_j}. Nejčastěji se k výpočtu této pravděpodobnosti používá multinomický unigramový jazykový model:

$$P(q|\theta_{d_j}) = K_q \prod_{t \in V} P(t|\theta_{d_j})^{tf_{t,d_j}}, \tag{4.17}$$

kde

$$K_q = \frac{L_{d_j}!}{tf_{t_1,d_j}! tf_{t_2,d_j}! \cdots tf_{t_N,d_j}!} \tag{4.18}$$

je multinomický koeficient pro dotaz q. Tento koeficient je konstantní pro určitý dotaz (množinu termů), nebude mít tedy vliv na poměr pravděpodobností s jakou mohly tento dotaz vygenerovat dva různé modely dokumentu a můžeme jej ignorovat.

Pravděpodobnost $P(q|\theta_{d_j})$ získáme pomocí odhadu maximální věrohodnosti (MLE)[4]:

$$\hat{P}(q|\theta_{d_j}) = \prod_{t \in q} \hat{P}_{MLE}(t|\theta_{d_j}) = \prod_{t \in q} \frac{tf_{t,d_j}}{L_{d_j}}, \tag{4.19}$$

kde tf_{t,d_j} je frekvence termu t v dokumentu d_j a L_{d_j} je délka dokumentu d_j. Jako relevantní dokumenty d_j k dotazu q poté označíme dokumenty s největší pravděpodobností $\hat{P}(q|\theta_{d_j})$.

4.5.2 Vyhlazování jazykového modelu

Problémem odhadu $\hat{P}(q|\theta_{d_j})$ je, že pokud dokument d_j nebude obsahovat některý term t z dotazu q, tedy $\hat{P}_{MLE}(t|\theta_{d_j}) = 0$, pak celý $\hat{P}(q|\theta_{d_j}) = 0$. Základním principem vyhlazování jazykového modelu dokumentu d_j je přiřadit nějakou pravděpodobnost i termům, které se v dokumentu nevyskytly. Tato pravděpodobnost by však neměla být větší než je pravděpodobnost výskytu termu v celé kolekci dokumentů C:

$$\hat{P}(t|\theta_{d_j}) \leq \frac{tf_{t,C}}{T}$$

kde $tf_{t,C}$ je počet výskytů termu t v kolekci C a T je celkový počet výskytů všech termů v C. Cílem vyhlazování je nejen doplnit odhad pravděpodobností chybějících termů, ale také upravit odhad termů, které se v dokumentu vyskytují, v závislosti na jejich výskytu v celé kolekci. Vyhlazování tedy plní stejnou roli jako IDF část ve vektorovém TF-IDF modelu [186].

Model kolekce θ_C získáme odhadem pravděpodobnosti:

$$\hat{P}_{MLE}(t|\theta_C) = \frac{tf_{t,C}}{T}. \tag{4.20}$$

Jelinek-Mercer vyhlazování

Nejčastěji používaným způsobem vyhlazování pro vyhledávání informací je lineární interpolace jazykového modelu - *Jelinek-Mercer vyhlazování* [66]. Pravděpodobnost $\hat{P}(t|d_j)$ zde získáme ze smíšeného modelu skládajícího se z multinomického unigramového modelu dokumentu θ_{d_j} a multinomického unigramového modelu kolekce θ_C:

$$\hat{P}(t|d_j) = \lambda P(t|\theta_{d_j}) + (1 - \lambda)P(t|\theta_C), \tag{4.21}$$

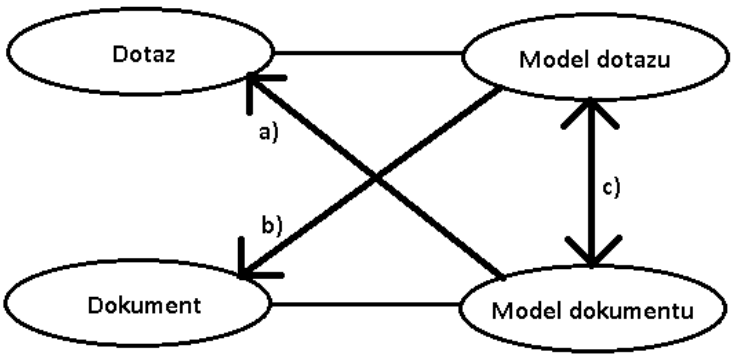
kde $\lambda \in \langle 0, 1 \rangle$ je interpolační parametr určující poměr smísení modelů.

Dirichlet vyhlazování

Další možností je *Dirichlet vyhlazování* (někdy také označované jako Bayesovské vyhlazování) [93]:

$$\hat{P}(t|d_j) = \frac{tf_{t,d_j} + \alpha \hat{P}(t|\theta_C)}{L_{d_j} + \alpha}, \tag{4.22}$$

[4]Maximum Likelihood Estimate



Obrázek 4.2: Přístupy k vyhledávání informací pomocí jazykového modelování: a) Query likelihood model, b) Document likelihood model, c) porovnání jazykových modelů dotazu a dokumentu

kde α je parametr vyhlazování. V porovnání s Jelinek-Mercer vyhlazováním získáme $\lambda = \alpha/(\alpha + L_{d_j})$, delší dokumenty budou tedy méně vyhlazovány.

Velikost vlivu vyhlazování u těchto dvou metod je kontrolována parametry λ a α. Menší hodnota λ a větší hodnota α znamená větší vliv vyhlazování. Možností také je nastavit parametry v závislosti na délce dotazu, pro delší dotaz je vhodné více vyhlazování než pro kratší [97].

Dvoufázové vyhlazování

Jeden z novějších přístupů k vyhlazování je založen na experimentech provedených v článku [186]. Zhai a Lafferty ve své studii tvrdí, že vyhlazování by mělo plnit dvě role. Mělo by vylepšovat přesnost odhadu jazykového modelu dokumentu - role odhadování a jako druhou úlohu by mělo vysvětlovat běžná a neinformativní slova v dotazu - role modelování dotazu. Na úlohu odhadování se ukázalo lepší Dirichlet vyhlazování, kdežto na úlohu modelování dotazu Jelinek-Mercer vyhlazování. Jejich závěrem tedy je, že nejlepšího efektu bude dosaženo při kombinaci obou způsobů vyhlazování. Definovali tedy *Dvoufázové vyhlazování*[5]:

$$\hat{P}(t|d_j) = \lambda \frac{tf_{t,d_j} + \alpha \hat{P}(t|\theta_C)}{L_{d_j} + \alpha} + (1 - \lambda)\hat{P}(t|\theta_U), \tag{4.23}$$

kde θ_U je jazykový model prostředí uživatelského dotazu. Obecně je tento model odlišný od modelu kolekce θ_C, ale v případě nedostatku dat lze použít model θ_C jako aproximaci modelu θ_U.

4.5.3 Jiné přístupy k jazykovému modelování

Dalšími možnými přístupy k vyhledávání informací pomocí jazykového modelování jsou například *Document likelihood model* (model věrohodnosti dokumentu), nebo přímé porovnání jazykových modelů dotazu a dokumentu například pomocí *Kullback-Leibler divergence*. Vazba mezi jednotlivými přístupy je znázorněna na obrázku 4.2.

[5] Two-stage smoothing

Document likelihood model

Document likelihood model hledá pravděpodobnost s jakou model dotazu θ_q vygeneruje dokument d_j. Hlavní nevýhodou tohoto přístupu je nedostatek dat k vygenerování modelu dotazu, protože většinou máme k dispozici pouze krátké dotazy. θ_q musí tedy být více vyhlazován nějakým jiným jazykovým modelem. Velikou výhodou tohoto modelu je snadné začlenění zpětné vazby od uživatele, prostým přidáním dalších termů z relevantních dokumentů do jazykového modelu dotazu θ_q [185, 82].

Kullback-Leibler divergence

Přímé porovnání jazykových modelů θ_q a θ_{d_j} pomocí *Kullback-Leibler divergence (KL)* [78] je založeno na změření vzdálenosti mezi dvěma pravděpodobnostními rozděleními, tedy modelem dotazu θ_q a modelem dokumentu θ_{d_j} [72]:

$$KL(\theta_{d_j}||\theta_q) = \sum_{t \in V} P(t|\theta_q) \log \frac{P(t|\theta_q)}{P(t|\theta_{d_j})}. \tag{4.24}$$

KL divergence je nesymetrická míra vzdálenosti, která vyjadřuje jak „špatně“ θ_q modeluje θ_{d_j}. Lafferty a Zhai v práci [78] ukazují, že při použití tohoto přístupu lze dosáhnout lepších výsledků, než s metodami *Query likelihood model* a *Document likelihood model*.

Do tohoto přístupu lze také velice snadno zakomponovat zpětnou vazbu jako rozšíření modelu dotazu [185] pomocí lineární interpolace. Porovnání jazykových modelů pomocí KL divergence je zobecněním *Query likelihood modelu*, ten získáme pokud jako odhad modelu θ_q použijeme

$$\hat{P}_{MLE}(t|\theta_q) = \frac{tf_{t,q}}{L_q}, \tag{4.25}$$

kde L_q je délka dotazu [184].

Modely vyššího řádu

Většina prací, zabývajících se použitím jazykového modelování v oblasti vyhledávání informací, využívá pouze unigramový jazykový model. Ve vyhledávání informací je většinou jazykový model vytvářen pouze z jednoho dokumentu, není tedy k dispozici dostatek dat pro dobrý odhad modelu vyššího řádu, a je tedy otázkou, zda více informací získaných například z bigramového modelu dokumentu vede k lepším výsledkům vyhledávání i přes jeho horší odhad.

Použití modelů vyššího řádu je představeno v práci [156], kde modely vyššího řádu jsou vyhlazovány modelem nižšího řádu:

$$\hat{P}(t_{i-1}, t_i|d_j) = \lambda_2 P(t|\theta_{d_j}) + (1 - \lambda_2) P(t_{i-1}, t_i|\theta_{bi\_d_j}), \tag{4.26}$$

kde $\theta_{bi\_d_j}$ je bigramový model dokumentu d_j. V práci [158] byl představen *Bitermový jazykový model*, na rozdíl od bigramového modelu není u toho modelu důležité pořadí slov, jen jejich společný výskyt. Slovní spojení „model jazykový“ bude mít tedy v tomto přístupu stejnou pravděpodobnost jako „jazykový model“.

V pracích [37, 14] byl představen obecný závislostní jazykový model[6], který bere v úvahu, že termy v dotazu mohou mít mezi sebou obecně nějaké závislosti a umožňuje je modelovat. Speciálním případem tohoto modelu je také bigramový model.

[6] Dependence language model

Kapitola 5

Metody vyhledávání v řečových reprezentacích

V této kapitole bude stručně popsán systém automatického rozpoznávání řeči a způsoby vyhledávání informací v různých variantách reprezentací řeči získaných z tohoto systému. Nejprve budou popsány metody vyhledávání na slovní úrovni, tedy v slovních nejlepších přepisech, slovních mřížkách a jejich reprezentacích. Poté budou zmíněny způsoby vyhledávání na úrovni subslovních jednotek a kombinované přístupy.

5.1 Automatické rozpoznávání řeči

Cílem automatického rozpoznávání řeči (ASR)[1] je získat z akustického signálu původní řečenou posloupnost slov. Obtížnost této úlohy je různá, závisí na tom, zda rozpoznáváme jednotlivá slova z malého slovníku nebo souvislou řeč s bohatým slovníkem. Kvalitu rozpoznávání dále ovlivňuje kvalita nahraného řečového signálu, tedy přítomnost hluku z okolního prostředí, kvalita mikrofonu, ale také například způsob řeči řečníka (například silný přízvuk, rýma nebo dialekt).

V reálné úloze, kdy budeme systém ASR potřebovat pro rozpoznání mluvené řeči určené dále pro vyhledávání informací, půjde téměř vždy o souvislou řeč, s různě obsáhlým slovníkem podle konkrétní aplikace (více ke konkrétním aplikacím v podkapitole 2.6). Kvalita nahrávky se bude pohybovat od dobré (např. studiové zpravodajství) po velmi špatnou (telefonní zákaznický servis).

5.1.1 Architektura ASR systému

V současné době je většina ASR systémů založena na statistickém přístupu [7] a používá velký slovník - systémy *LVCSR - Large-Vocabulary Continuous Speech Recognition*. Tyto systémy zpracovávají souvislou řeč a jsou nezávislé na řečníkovi. Většina ASR systémů pracuje na základě principu *skrytých Markovských modelů - HMM*[2] [116]. Základem tohoto přístupu je pohled na řeč jako na zarušený kanál - slova projdou tímto kanálem a za účasti rušení vznikne akustický signál. Cílem systému je najít model zarušeného kanálu a opačným procesem získat zpět posloupnost slov.

[1] Automatic Speech Recognition
[2] Hidden Markov Model

Ze všech možných slovních posloupností W v daném jazyce L hledáme tu nejpravděpodobnější za předpokladu daného akustického vstupu O. Je dána sekvence akustických znaků $O = o_1, o_2, ..., o_t$ a posloupnost slov $W = w_1, w_2, ..., w_n$. Cíl ASR, tedy odhad slovní posloupnosti W, můžeme vyjádřit takto:

$$\hat{W} = \text{argmax}_{W \in L} P(W|O). \tag{5.1}$$

Použijeme Bayesovo pravidlo:

$$\hat{W} = \text{argmax}_{W \in L} \frac{P(O|W)P(W)}{P(O)}. \tag{5.2}$$

Pravděpodobnost $P(W)$ je apriorní pravděpodobnost posloupnosti slov W, bývá odhadována *n-gramovým jazykovým modelem*. Pravděpodobnost $P(O)$, tedy apriorní pravděpodobnost akustické sekvence O může být ignorována, protože je stejná pro všechny slovní posloupnosti W. Vztah (5.2) můžeme tedy upravit:

$$\hat{W} = \text{argmax}_{W \in L} P(O|W)P(W). \tag{5.3}$$

Hledáme tedy posloupnost slov W, která bude mít největší pravděpodobnost složenou součinem pravděpodobností $P(O|W)$ - získanou z *akustického modelu* a $P(W)$ - získanou z *jazykového modelu*. Tuto posloupnost získáme pomocí dekódování HMM Viterbi algoritmem. Více o této problematice například v kapitole 9 a 10 knihy [72].

5.1.2 Výstupy ASR systému

Popsaným způsobem získáme pouze nejlepší hypotézu - *nejlepší slovní přepis* (1-best) z ASR systému. Další možností je získat *n-nejlepších přepisů* (n-best), tedy n nejpravděpodobnějších slovních posloupností $\hat{W}$ upraveným Viterbi algoritmem. Ohodnocení hypotéz v n-nejlepších přepisech můžeme dále přepočítat například použitím složitějšího jazykového modelu a teprve poté získáme druhým průchodem nejlepší hypotézu.

Pro velké n může být reprezentace hypotéz jako seznam n-nejlepších přepisů nevhodná, vhodnější bude použít *slovní mřížku* (lattice). Slovní mřížka je orientovaný graf, jehož uzly jsou body v čase a hrany reprezentují slovní hypotézu a k ní patřící informace (akustickou pravděpodobnost, pravděpodobnost z jazykového modelu, atd.). Slovní mřížku získáme při prvním průchodu dekodéru HMM zahrnutím více slovních hypotéz v každém čase. Ze slovní mřížky můžeme dalším průchodem dekodéru získat n-nejlepších hypotéz, případně nejlepší hypotézu.

5.1.3 Chybovost ASR systému

Standardní mírou pro vyhodnocení chybovosti ASR systému je míra *WER - Word Error Rate*, tedy míra slovní chybovosti. Porovnává jak moc se liší slovní posloupnost $\hat{W}$ odhadnutá ASR systémem, od originální slovní posloupnosti W. Pro výpočet WER musíme nejprve nalézt minimální počet slovních substitucí S, vložení I a smazání D, který je potřeba k získání $\hat{W}$ z W. Poté lze spočítat WER jako:

$$WER = 100 \times \frac{S + I + D}{|W|}, \tag{5.4}$$

kde $|W|$ je počet slov ve W.

Pro slovní mřížky lze definovat míru chybovosti slovní mřížky [173], která odpovídá nejmenší WER chybovosti, které dosáhne některá z cest v slovní mřížce. Tato cesta s nejmenší chybovostí se často nazývá věštecká (oracle) cesta.

5.2 Nejlepší přepisy

Nejjednodušším způsobem vyhledávání informací v řeči je použití dvou oddělených systémů. První bude použit pro přepis řeči do textu - systém pro automatické rozpoznávání řeči a jako jeho výstup vezmeme pouze nejlepší cestu v slovní nebo subslovní mřížce - nejlepší přepis. Tím převedeme úlohu vyhledávání v řeči na vyhledávání v textu. Druhým systémem bude klasický systém pro vyhledávání informací v textu pracující například na základě jednoho z modelů uvedených v kapitole 4, jehož vstupem bude nejlepší přepis z ASR systému.

5.2.1 Aplikace nejlepších přepisů ve vyhledávání

Na základě vyhledávání v nejlepších přepisech pracují systémy, jejichž cílem je poskytnout uživateli přepis řeči k nahlédnutí a vyhledávání v řečových záznamech, například hlasové poště [171] nebo historických archivech [48].

TREC/SDR

Tento přístup byl také použit v TREC/SDR úloze, cílem této úlohy bylo vyhledávání ve zpravodajských pořadech. Přibližně 550 hodin řeči bylo ručně rozděleno na 21 574 jednotlivých zpráv. Pro vyhledávání byly použity pouze nejlepší přepisy z ASR systému s chybovostí přibližně 15-20% WER (viz podkapitola 5.1.3), tedy s téměř stejnou chybovostí jako měly přibližné manuální přepisy. Bylo zjištěno, že při chybovosti ASR systému v rozmezí 15-30% nenastane žádné vážné zhoršení přesnosti vyhledávání oproti vyhledávání v přibližných manuálních přepisech [38]. Jako vyhledávací funkce byla ve většině případů použita nějaká varianta funkce BM25 (viz podkapitola 4.4.1).

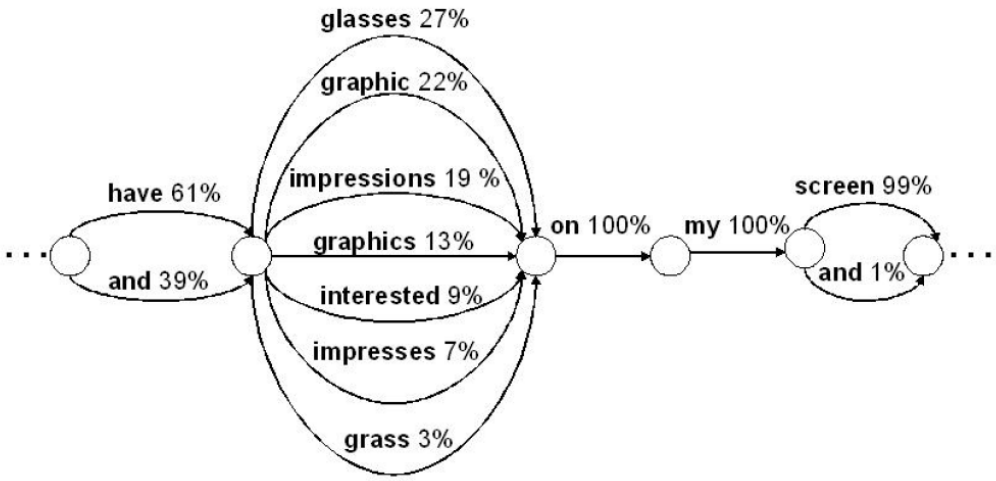
CLEF

Další z důležitých úloh, kde byl použit pouze nejlepší přepis z ASR systému, je úloha CLEF CL-SDR a CL-SR. Vyhledávání probíhalo v kolekci spontánní spojité řeči [106], získané z projektu MALACH. V letech 2006 [107] a 2007 [111] probíhalo vyhledávání též v české kolekci spontánní řeči [62]. Výpovědi byly uměle rozděleny na stejně dlouhé dokumenty získané jako nejlepší přepis z ASR systému. Výsledky experimentů ukázaly, že přesnost vyhledávání na přepisech s větší chybovostí než u TREC/SDR úlohy není dostačující.

5.2.2 Vylepšení nejlepších přepisů

V práci [95] bylo ukázáno, že pokud pro vytvoření nejlepšího přepisu nepoužijeme přímo slovní mřížku, ale mřížku první upravíme do formy WCN[3] (více v podkapitole 5.4.2)

[3] Word Confusion Networks



Obrázek 5.1: Příklad vylepšení nejlepšího přepisu při použití WCN stemmingem - dojde k sloučení pravděpodobností tvarů *graphic*, *graphics* (převzato z [95])

a teprve z WCN vytvoříme nejlepší přepis, můžeme dosáhnout snížení WER chybovosti nejlepšího přepisu a také větší přesnosti vyhledávání.

Další navržené vylepšení bylo nastíněno ve stejné práci, kde byl na slova nejprve použit *stemming*[4] a tím pádem byly sloučeny pravděpodobnosti různých tvarů stejného slova do jednoho výskytu, teprve poté byl vytvořen nejlepší přepis. Na obrázku 5.1 je vidět případ, kdy přestože nejvyšší pravděpodobnosti dosahuje slovo *glasses*, na dalších pozicích se nachází tvary stejného slova: *graphic*, *graphics*, které se součtem svých pravděpodobností dostanou na první pozici a budou tedy obsaženy v nejlepším přepisu. Podobného efektu by bylo dosaženo za použití *lemmatizace*[5].

5.2.3 Chybovost nejlepších přepisů

Navzdory dobrým výsledkům dosaženým v TREC/SDR úloze nelze říci, že by tento přístup byl obecně dostačující pro vyhledávání informací v řeči. Použité ASR systémy byly natrénovány na danou úlohu, řeč v jednotlivých zprávách je dobře srozumitelná, předem připravovaná a většinou neobsahuje šum. Díky těmto vlastnostem mohly nejlepší přepisy dosahovat relativně malé chybovosti.

Ukazuje se však, že při reálné aplikaci vyhledávání informací v řeči může chybovost nejlepších přepisů být mnohem vyšší. Například v kolekci pro vyhledávání ve spontánní řeči v projektu MALACH [106] dosahovala chybovost 40-50%, v projektu SpeechFind - vyhledávání v National Gallery of the Spoken Word [48] dosahovala chybovost až 60% WER, v oblasti telefonních konferencí [138] se chybovost pohybuje okolo 50% WER.

Mamou a kolektiv v práci [95] testovali vliv chybovosti ASR systému na přesnost vyhledávání na záznamech hovorů z telefonního centra, chybovost těchto automatických přepisů byla 30% a větší. Ukázali, že s větší chybovostí ASR přepisu klesá přesnost vyhledávání a je tedy vhodné použít nejen nejlepší přepis, ale celou slovní mřížku (v tomto případě upravenou do formy WCN).

[4] Zkrácení slova na jeho kořen

[5] Převedení slova na základní tvar

K opačným výsledkům došli v současné době autoři několika jiných prací [23, 110, 163], zabývajících se většinou nějakým vylepšením základních metod pro vyhledávání informací (například použití lemmatizace či stemmingu, rozšíření dotazu, zpětné vazby). Jejich výsledky ukazují, že není žádný nebo minimální rozdíl ve výsledcích vyhledávání při použití nejlepších hypotéz a slovní či subslovní mřížky, nebo jejich reprezentací, pokud je použita „lepší“ metoda pro vyhledávání informací, než jen základní. Závěrem experimentů tedy je, že chybovost ASR systému přímo neovlivňuje výsledky vyhledávání.

V práci [110] byl porovnáván rozdíl ve vlivu chybovosti ASR systému na výsledky vyhledávání ve dvou úlohách - vyhledávání klíčových slov a vyhledávání informací v řeči. Závěrem experimentů bylo, že přestože v úloze vyhledávání klíčových slov má chybovost ASR systému velký vliv na kvalitu výsledků, v úloze vyhledávání informací se toto nepotvrdilo. Předpokladem je, že pro vyhledávání informací je důležitější význam dokumentu než přesný obsah slov. V práci byly místo celých slov indexovány jen kořeny slov[6] ve variantě nejlepších přepisů, slovních CN, subslovních CN a kombinovaného indexu, a závěrem experimentů bylo, že při použití stemmingu dochází k shluknutí významově podobných slov do jednoho kořene a tato úprava je pro výsledky vyhledávání přínosnější než použití mřížky z ASR systému.

5.3 Slovní mřížky

Jak je vidět z předchozí podkapitoly (5.2), přestože je vyhledávání informací v řeči s použitím nejlepších přepisů z ASR systému velmi jednoduché a nevyžaduje žádnou úpravu stávajícího systému pro vyhledávání v textu, vysoká chybovost nejlepších přepisů, například v úlohách vyhledávání ve spontánní řeči, motivuje výzkumné týmy orientovat se na využití všech dostupných informací, které nám může ASR systém poskytnout.

Jedním ze způsobů, jak dosáhnout snížení vlivu chybovosti ASR systému, je použití slovních mřížek. Slovní mřížka, jak bylo řečeno v podkapitole 5.1.2, je orientovaný acyklický graf, jehož uzly jsou body v čase a hrany reprezentují slovní hypotézu a k ní patřící informace (akustickou pravděpodobnost, pravděpodobnost z jazykového modelu). Každá cesta ve slovní mřížce reprezentuje možnou posloupnost slov v řeči. Ukázka slovní mřížky a všech v ní obsažených cest je vidět na obrázku 5.2.

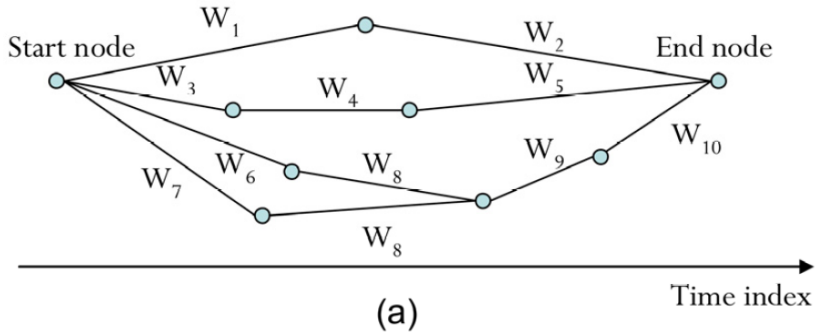
5.3.1 Vyhledávání ve slovní mřížce

Řečové dokumenty rozdělíme na segmenty s. Z ASR systému získáme slovní mřížku pro každý segment s, s cestami π ohodnocenými aposteriorními pravděpodobnostmi $P(\pi, s)$, získanými jako součin pravděpodobností jednotlivých slovních hypotéz v cestě π. Slovní mřížka je poté prořezána, abychom odstranili hypotézy s pravděpodobností menší než stanovený práh θ.

Očekávaný počet

Pro vyhledávací algoritmy obecně je jedna z důležitých informací o dokumentu informace o počtu výskytů daného slova v dokumentu. Pro dokumenty v podobě slovních mřížek získáme informaci o počtu výskytů slova ve formě *očekávaného počtu*[7]. Uvedeme například postup výpočtu použitý v pracích [27, 29].

[6] Stemming
[7] Expected Count



W_1W_2, $W_3W_4W_5$, $W_6W_8W_9W_{10}$, $W_7W_8W_9W_{10}$
(b)

Obrázek 5.2: Ukázka slovní mřížky(a) a všech cest v ní(b) (převzato z [86])

Očekávaný počet daného slova w v segmentu dokumentu s spočteme:

$$E[c(w,s)] = \sum_{\pi} c(w,\pi)P(\pi,s), \tag{5.5}$$

kde $c(w,\pi)$ je počet výskytů slova w v cestě π. Dále můžeme definovat délku úseku s:

$$E[|s|] = \sum_{\pi} |\pi|P(\pi,s). \tag{5.6}$$

Pro celý dokument d, skládající se z úseků s spočteme očekávaný počet slova w:

$$E[c(w,d)] = \sum_{s}\sum_{\pi} c(w,\pi)P(\pi,s) \tag{5.7}$$

a délku dokumnetu d:

$$E[|d|] = \sum_{s}\sum_{\pi} |\pi|P(\pi,s). \tag{5.8}$$

Když známe očekávaný počet slov v mřížce, můžeme přistoupit k samotnému vyhledávání.

Jazykové modelování

V pracích [27, 29] byl použit jazykový model pro vyhledávání informací s použitím dvoufázového vyhlazování (viz podkapitola 4.5). Pravděpodobnost $\hat{P}(q|d)$ tedy spočteme pomocí upravené rovnice (4.23):

$$\hat{P}(q|d) = \prod_{w\in q}(\lambda\frac{E[c(w,d)] + \alpha\hat{P}(w|\theta_C)}{E[|d|]+\alpha} + (1-\lambda)\hat{P}(w|\theta_U))^{tf_{w,q}}, \tag{5.9}$$

kde $tf_{w,q}$ je počet výskytů slova w v dotazu q.

BM25

Ve stejné práci [29] byl také použit model BM25 pro vyhledávání informací (viz podkapitola 4.4.1). Analogicky jako u jazykového modelování použijeme očekávaný počet slov místo frekvencí termu. Podobnost dotazu a dokumentu tedy spočteme podle upraveného vzorce (4.15).

$$sim_{d,q} = \sum_{w \in q} idf_w \frac{E[c(w,d)](k_1+1)}{E[c(w,d)] + k_1(1 - b + b\frac{E[|d|]}{Lavg_d})} \frac{tf_{w,q}(k_3+1)}{tf_{w,q}+k_3}. \tag{5.10}$$

5.3.2 Aplikace slovních mřížek pro vyhledávání

Mřížky jako reprezentace řeči pro vyhledávání klíčových slov byly poprvé použity v práci [64], v práci byly použity fónové mřížky (více v podkapitole 5.5). Přístup byl dále rozveden do vyhledávání informací v řeči [65].

Siegler v dizertační práci [144] použil upravený TF-IDF model pro vyhledávání ve slovních mřížkách a n-nejlepších přepisech. TF část modelu byla vypočtena z pravděpodobností všech hran mřížky se stejnou slovní hypotézou a IDF část jako míra vzájemné informace mezi slovem a kolekcí dokumentů.

V práci [29] Chia a kolektiv testovali vyhledávání ve slovních mřížkách pomocí metod představených v podkapitole 5.3.1 na kolekci telefonních rozhovorů v angličtině a porovnávali je s metodou vyhledávání ve WCN představenou v článku [95] (viz podkapitola 5.4.2). Výsledky jejich experimentů ukazují, že všechny tři metody založené na vyhledávání ve slovních mřížkách dosahují lepších výsledků než vyhledávání v nejlepších přepisech. Zároveň také ukazují, že obě metody založené na přímém vyhledávání ve slovních mřížkách, tedy vyhledávání založené na jazykovém modelování a BM25, dosahují lepších výsledků než metoda založená na použití TF-IDF modelu a vyhledávání v WCN reprezentaci slovní mřížky.

Chia a kolektiv také porovnávali vyhledávání ve slovních mřížkách pomocí jazykových modelů s TF-IDF vyhledáváním ve WCN na kolekci telefonních rozhovorů v čínštině [27]. Experimenty ukazují u obou metod založených na slovních mřížkách vylepšení vyhledávání oproti nejlepším přepisům. Také ukazují lepší výsledky metody založené na jazykovém modelování a slovních mřížkách oproti TF-IDF vyhledávání ve WCN.

5.4 Kompaktní reprezentace mřížek

Pro usnadnění práce s mřížkami pro vyhledávání informací byly vyvinuty různé varianty zjednodušených kompaktních reprezentací těchto mřížek. Přestože jsou tyto reprezentace ztrátové, stále obsahují informaci o aposteriorní pravděpodobnosti termu a také zjednodušují přístup k informaci o blízkosti termů. Díky své kompaktnosti mají tyto reprezentace menší paměťové nároky a také algoritmy pro jejich indexaci mohou být méně výpočetně náročné.

5.4.1 PSPL

PSPL - Position Specific Posterior Lattice [20] (Aposteriorní mřížky se specifikovanou pozicí) jsou způsobem reprezentace slovních mřížek získaných z ASR systému. Tento přístup je založen na předpokladu, že pro vyhledávání informací je velice důležitá informace o pozici slov v dokumentu a z těchto pozic získatelné vzdálenosti dvou slov. V slovních mřížkách není jasně viditelná informace o vzdálenosti dvou slov, kvůli možnosti výskytu každého slova v různých cestách v mřížce. Lze ale snadno určit pozici slova v jedné z cest mřížky, každé slovo má také přiřazenou aposteriorní pravděpodobnost. Každé slovo můžeme reprezentovat čtveřicí

$$(w, d, poz, pst)$$

kde w je slovo v mřížce, d je číslo dokumentu, poz pozice slova v mřížce a pst apriorní pravděpodobnost slova.

Konstrukce PSPL

Pokud se nějaké slovo bude vyskytovat ve více cestách v mřížce na dané pozici, sečteme příslušné aposteriorní pravděpodobnosti. Na tento výpočet může být použita varianta standardního forward-backward algoritmu. Výpočet zpětného průchodu (backward) zůstává stejný, u dopředného průchodu (forward) je nutné rozdělit dopřednou pravděpodobnost α_n v uzlu n podle délky l - počet slov v cestě mřížkou od počátečního uzlu do uzlu n:

$$\alpha_n[l] = \sum_{\pi} P(\pi), \tag{5.11}$$

π končí v n a délka π je l. Zpětná pravděpodobnost β_n zůstává standardně definována [169]. Inicializace a základní krok dopředného průchodu algoritmu může být formalizován takto:

$$\alpha_n[l+1] = \sum_{i=1}^{q} \alpha_{s_i}[l + \delta(l_i, \epsilon)] P(l_i), \tag{5.12}$$

$$\alpha_{start} = 1, l = 0$$
$$\alpha_{start} = 0, l \neq 0$$

pro uzly s_i předcházející n a $P(l_i)$ je pravděpodobnost hrany l_i z ASR systému. Aposteriorní pravděpodobnost slova w nacházejícího se na pozici l v mřížce L lze spočítat:

$$P(w, l|L) = \sum_{n} \frac{\alpha_n[l]\beta_n}{\beta_{start}} \delta(w, word(n)), \tag{5.13}$$

kde β_{start} je součet ohodnocení všech cest v mřížce. PSPL je tedy reprezentací rozdělení pravděpodobnosti $P(w, l|L)$, pro každou pozici l ukládá slovo w s pravděpodobností $P(w, l|L)$.

Vyhledávání v PSPL

Princip vyhledávání informací v PSPL mřížkách je založen na informaci o vzdálenosti termů v dokumentech, tento přístup byl inspirován přístupem, který použili Brin a Page pro Google [10].

Dokumenty v kolekci pro vyhledávání mohou být dlouhé, případně může jít o nedělenou řečovou nahrávku, proto budou rozděleny na segmenty. Každý segment převedeme do formy PSPL, tím vytvoříme takzvaný *soft index*, ukládající pro každé slovo w aposteriorní pravděpodobnost spolu s pozicí pro každý výskyt daného slova.

Pro termy t_i dotazu q a dokument d, reprezentovaný jako PSPL svých segmentů s, spočteme unigramovou podobnost součtem všech aposteriorních pravděpodobností termu t_i ve všech segmentech s a pozicích k:

$$sim_{unigram}(d,g) = \sum_{t_i \in q} \log \left[1 + \sum_s \sum_k P(w_k(s) = t_i|d) \right]. \tag{5.14}$$

Je možné také použít n-gramovou podobnost pro n-gramy nacházející se v dotazu $q : t_i ... ti + n - 1$. Pro každý řád n spočteme:

$$sim_{n-gram}(d,g) = \sum_{i=1}^{T-n+1} \log \left[1 + \sum_s \sum_k \prod_{l=0}^{n-1} P(w_{k+l}(s) = t_{i+l}|d) \right]. \tag{5.15}$$

Podobnosti získané pro každý řád n, který dovoluje dotaz $q : t_1 ... t_T$, jsou poté kombinovány s vektorem vah w_n:

$$sim(d,g) = \sum_{n=1}^{T} w_n \cdot \mathrm{sim}_{n-gram}(d,g). \tag{5.16}$$

Váha w_n se lineárně zvětšuje s rostoucím řádem n-gramů.

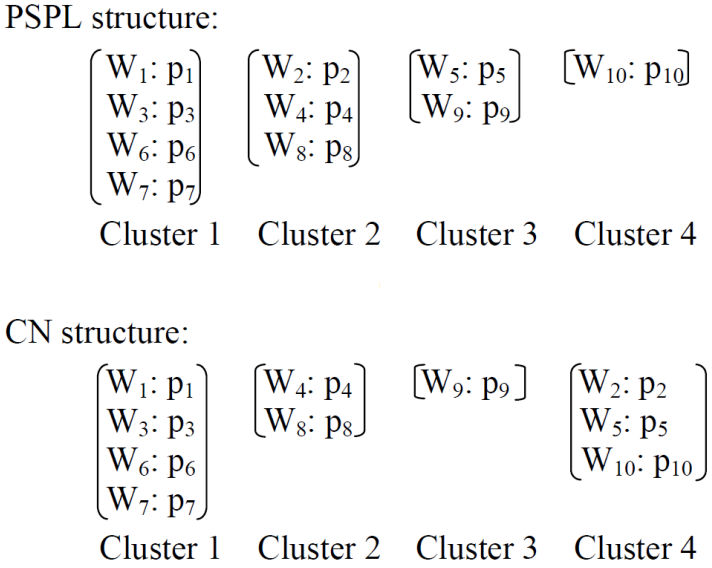
Aplikace PSPL

Chelba a Acero ve svých experimentech [21] na kolekci univerzitních přednášek iCampus [41] ukázali, že při vyhledávání v PSPL místo v nejlepším přepisu dochází k 20% relativnímu vylepšení střední průměrné přesnosti (MAP). Chybovost nejlepšího přepisu byla 45% WER a PSPL reprezentace mřížky jen 22% WER. Dochází také k podstatnému zmenšení nároků na prostor pro uložení indexu, velikost invertovaného indexu vytvořeného z PSPL je pouze 20% velikosti 3-gramové ASR mřížky [22].

5.4.2 WCN

Další možnou kompaktní reprezentací slovní mřížky jsou CN - Confusion Networks, v případě slovních mřížek WCN - Word Confusion Networks (Sítě slovních záměn). WCN reprezentace slovní mřížky byla původně navržena pro minimalizaci WER chybovosti systému pro rozpoznávání řeči [96], ale díky svým vlastnostem je také velice vhodná pro vyhledávání informací. Porovnání přístupů ke shlukování PSPL a WCN je vidět na obrázku 5.3.

Principem WCN je shlukování slovních hypotéz v mřížce do lineárních shluků slovních alternativ podle jejich stejného výskytu v čase a podobné výslovnosti.



Obrázek 5.3: Porovnání PSPL a CN shluků vytvořených ze slovní mřížky na obr. 5.2 (převzato z [86])

Konstrukce WCN

Před začátkem shlukování je nutné spočítat aposteriorní pravděpodobnosti všech slovních hypotéz v mřížce forward-backward algoritmem. Shlukování probíhá ve dvou fázích: vnitřní shlukování slov[8] a mezislovní shlukování[9]:

1. **Vnitřní shlukování slov**

   V tomto kroku shlukujeme všechny hrany e v mřížce odpovídající stejnému slovu a vyskytující se ve stejném čase. Podobnostní funkce definovaná pro množiny hran E_1, E_2 vypadá takto:

   $$sim(E_1, E_2) = \underbrace{\max}_{e_1 \in E_1, e_2 \in E_2} \text{overlap}(e_1, e_2) p(e_1) p(e_2), \tag{5.17}$$

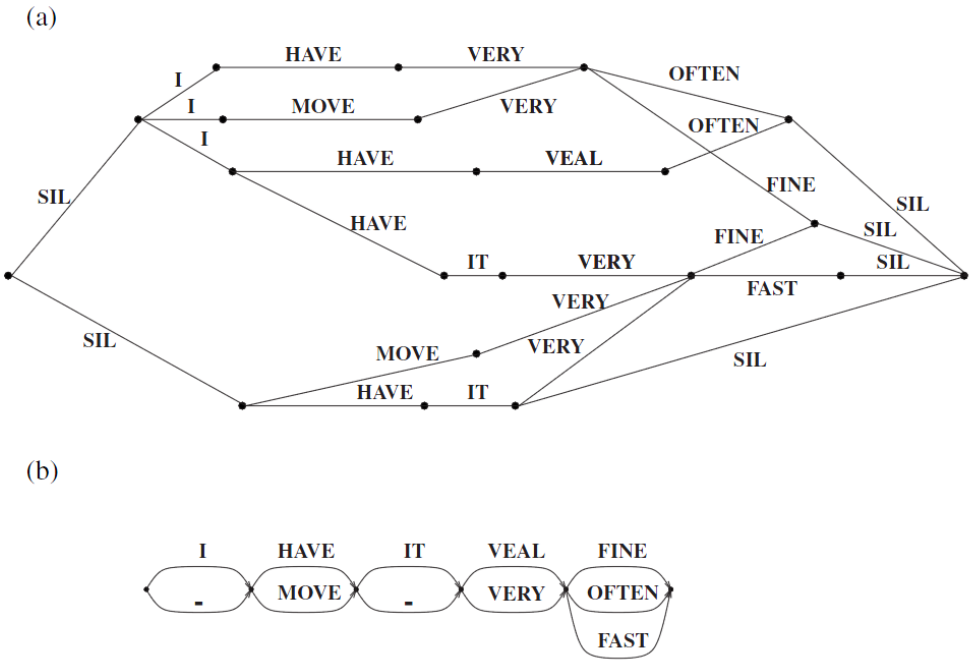
   kde funkce $\text{overlap}(e_1, e_2)$ je definována jako časový překryv dvou hran normalizovaný součtem jejich délek. Časový překryv je dále vážen aposteriorními pravděpodobnostmi jednotlivých hran, aby se znevýhodnily nepravděpodobné hypotézy.

   V každém kroku je spočtena podobnost mezi všemi možnými dvojicemi shluků a nejpodobnější shluky jsou spojeny. Na konci této fáze jsou hypotézy ve slovní mřížce rozděleny do shluků časově se překrývajících výskytů stejného slova.

2. **Mezislovní shlukování**

   V druhé fázi jsou spojovány shluky stejných slov z první fáze. Spojeny mohou být takové shluky, které nemají mezi sebou žádný vztah, tedy nenásledují za sebou.

[8] Intra-word clustering
[9] Inter-word clustering



Obrázek 5.4: Ukázka slovní mřížky(a) a odpovídající reprezentace WCN(b) (převzato z [96])

Funkce pro výpočet podobnosti shluků F_1, F_2 v této fázi je založená na fonetické podobnosti slov:

$$sim(F_1, F_2) = \underbrace{\text{avg}}_{w_1 \in Words_{F_1}, w_2 \in Words_{F_2}} \text{sim}_\text{F}(w_1, w_2) p_{F_1}(w_1) p_{F_2}(w_2), \tag{5.18}$$

kde $p_F(w) = p(\{e \in F : Word(e) = w\})$ a $\text{sim}_\text{F}(w_1, w_2)$ je fonetická podobnost slov w_1 a w_2.

Po dokončení shlukování jsou sečteny aposteriorní pravděpodobnosti všech hran příslušejících stejnému slovu v daném shluku. Ukázka WCN je na obr. 5.4.

Obecný algoritmus pro vytváření WCN ze slovní mřížky byl také představen v práci [47]. Tento algoritmus je založen na výběru jedné cesty v mřížce, ke které jsou poté přiřazovány hypotézy z dalších cest. Hypotéza je přiřazena k příslušnému překrývajícímu se shluku ve vybrané cestě a následně jsou sečteny aposteriorní pravděpodobnosti pro stejná slova.

Vyhledávání ve WCN

Mamou a kolektiv [95] použili pro vyhledávání informací ve WCN klasický vektorový TF-IDF model, upravený pro použití ve WCN. Frekvenci termu v dokumentu, reprezentovaném WCN, spočetli součtem aposteriorních pravděpodobností všech výskytů o_i termu t v dokumentu d vynásobených vahou pořadí termu mezi ostatními hypotézami $B_{rank(t|o_i)}$:

$$tf_{t,d} = \sum_{i=1}^{|occ(t,d)|} B_{rank(t|o_i,d)} \times P(t|o_i,d), \tag{5.19}$$

kde $occ(t,d)$ jsou všechny výskyty termu t v d. Inverzní frekvenci dokumentu lze spočítat takto:

$$idf_t = \log \frac{O}{O_t}, \tag{5.20}$$

kde O_t je počet všech výskytů termu t v kolekci C:

$$O_t = \sum_{d \in C} \sum_{i=1}^{|occ(t,d)|} P(t|o_i,d) \tag{5.21}$$

a O je součet výskytů všech termů O_t.

Aplikace WCN

WCN byly použity pro generování nejlepšího přepisu v práci [95], výsledky experimentů ukázaly vylepšení přesnosti vyhledávání oproti nejlepším přepisům ze slovní mřížky.

Tur a kolektiv [164] použili WCN ke klasifikaci promluv zákaznického dialogového systému. Ve svých experimentech dosáhli 5-10% zmenšení chybovosti klasifikace v porovnání s nejlepším přepisem z ASR systému [46].

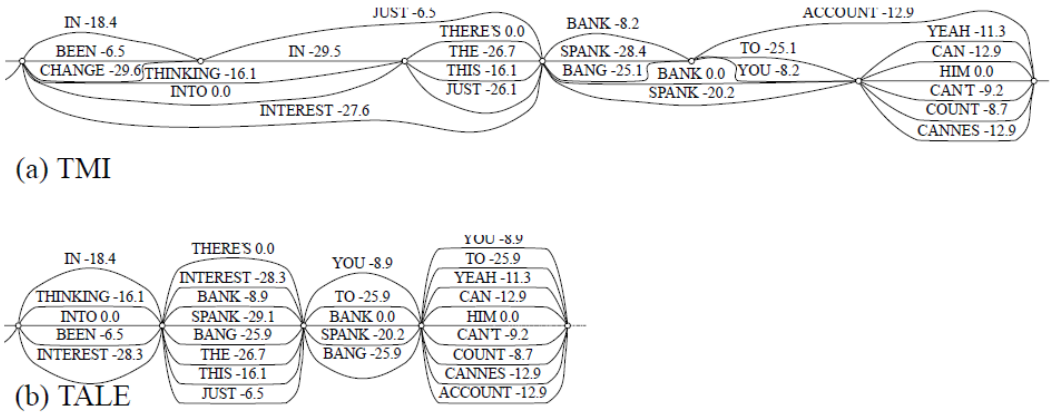
5.4.3 Další reprezentace

Další možnou reprezentací slovních mřížek je metoda zvaná Time-based Merging for Indexing - TMI[10] [187] (Časově založené slučování pro indexaci). TMI reprezentace je podobně jako PSPL založená na aposteriorních pravděpodobnostech vypočítaných forward-backward algoritmem. Hrany obsahující stejná slova jsou poté na základě shodného časového intervalu sloučeny. Zhou a kolektiv vyzkoušeli použití TMI pro úlohu vyhledávání v datech z internetu (videoklipy, online přednášky, atd.) a dosáhli 14% vylepšení přesnosti oproti nejlepším přepisům, přičemž indexy založené na TMI jsou pouze pětkrát větší než indexy vytvořené pouze z nejlepších přepisů.

TALE[11] [140] (Časově zakotvené rozšíření mřížek) je další možnou reprezentací slovních mřížek. Oproti TMI, která je zaměřená na co největší zmenšení velikosti mřížky, ale vyžaduje specifický indexovací algoritmus, umožňuje TALE použití klasického indexeru (viz obrázek 5.5). Metoda TALE je založená na převedení slovní mřížky do lineární posloupnosti shluků slov určených pro daný časový interval (podobně jako WCN). Tato metoda uchovává návaznosti posloupností až tří slov za sebou.

V článku [140] bylo ukázáno, že metody TMI i TALE dosahují 30-60% vylepšení oproti textovým přepisům.

[10] Time-based Merging for Indexing
[11] Time-Anchored Lattice Expansion



Obrázek 5.5: TMI a TALE reprezentace slovní mřížky (převzato z [140])

5.5 Použití subslovních jednotek

Předchozí podkapitoly se věnovaly vyhledávání ve slovních reprezentacích řeči, kde se výzkumné týmy použitím slovních mřížek, nebo jejich reprezentací, snažily dosáhnout větší přesnosti vyhledávání. Při použití slovních mřížek vyhledávací index obsáhne větší množství možných slovních hypotéz, lze tedy vyhledat i slova, která by díky chybě ASR systému nebyla součástí nejlepšího přepisu. Tento přístup ovšem neřeší případ slov neobsažených ve slovníku ASR systému, takzvaných *OOV*[12] slov. Pokud slovo z dotazu je takovým OOV slovem, při použití pouze slovních reprezentací řeči je nedokážeme vyhledat.

Popsané způsoby a metody vyhledávání pomocí subslovních jednotek se nicméně týkají zejména oblasti vyhledávání klíčových slov v řeči, kdy úkolem systému je nalézt přesný výskyt dotazovaného slova (nebo fráze). V oblasti vyhledávání informací v řeči se příliš nepoužívají. Například v práci [110] bylo ukázáno, že při použití v úloze vyhledávání informací nepřináší použití subslovních jednotek žádné vylepšení vyhledávání.

5.5.1 OOV slova

Woodland a kolektiv popsali v práci [174] efekt OOV slov na vyhledávání informací v řeči. Ukázali, že se stoupajícím množstvím OOV slov klesá průměrná přesnost vyhledávání a je tedy nutné se nějakým způsobem zaměřit na eliminaci OOV, nebo jejich vlivu.

Jedním ze způsobů, jak omezit vliv OOV slov je použití subslovních reprezentací řeči místo slovních. Subslovní jednotky budeme indexovat stejně jako slovní jednotky, tedy ze subslovní mřížky nebo odpovídající kompaktní subslovní reprezentace (PSPL, WCN). Slova v dotazu převedeme do odpovídajících subslovních jednotek a budeme hledat jejich sekvenci v dokumentech.

5.5.2 Výběr subslovních jednotek

Glavitsch a Schäuble v práci [42] definovali požadavky na subslovní jednotky vhodné pro vyhledávání informací v řeči:

[12] Out of Vocabulary

1. Indexované jednotky musí být jednotky snadno rozpoznatelné ASR systémem.

2. Počet různých jednotek musí být malý úměrně k množství dat potřebných pro natrénování modelu jednotky.

3. Jednotky musí být dobře rozlišitelné od sebe navzájem.

4. Frekvence jednotek v kolekci nesmí být příliš malá.

Na základě těchto požadavků navrhli trigramovou jednotku, skládající se z maximální sekvence souhlásek, obklopené z obou stran sekvencí maximálně dvou samohlásek.

Kenney Ng zkoumal v dizertační práci [104] vliv různých subslovních jednotek na přesnost vyhledávání. Ukázal, že při použití překrývajících se jednotek dosáhneme lepších výsledků než u nepřekrývajících se. Například při použití trifónů je přesnost vyhledávání téměř stejná jako při použití slov.

Výběr subslovní jednotky je také ve velké míře závislý na jazyce, ve kterém budeme vyhledávat. Pro abecedně založené jazyky, jako je například český nebo anglický jazyk jsou často voleny fonémy, grafémy [104, 167] nebo sekvence fonémů [89]. Pro jazyky založené na morfémech, jako je turecký jazyk nebo finština jsou voleny morfémy jako subslovní jednotky [109]. Pro čínštinu se například často volí slabiky [108, 26].

5.5.3 Vytvoření subslovní reprezentace řeči

Existují dvě varianty jak získat subslovní reprezentaci řeči:

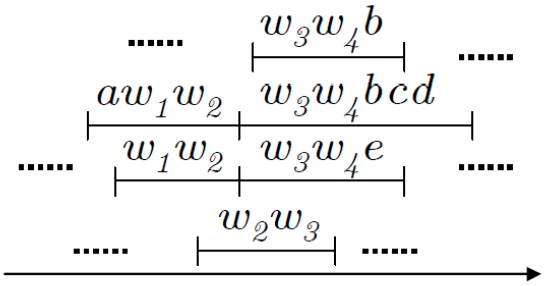
1. První možností je rozpoznávat rovnou na subslovní úrovni. Toho dosáhneme, pokud v ASR systému použijeme místo slovního jazykového modelu model subslovní.

2. Druhou možností je postup, kdy jako první krok je vytvořena slovní reprezentace řeči a ta je pak převedena na subslovní na základě výslovnosti daných slov.

Druhý přístup je často volen z důvodu vyšší přesnosti rozpoznávání na úrovni slov, než na subslovní úrovni [138] v oblasti vyhledávání klíčových slov. Jeho nevýhodou ale je, že budeme indexovat pouze slova, která jsou ve slovníku i na subslovní úrovni. Můžeme však předpokládat, že slovo, které není obsaženo ve slovníku ASR systému, bude při rozpoznávání nahrazeno hypotézami slov se stejnými, nebo podobnými subslovními jednotkami. Proto pokud rozpoznaná slova převedeme na subslovní jednotky, budeme schopni nalézt i OOV slova obsažená v dotazech tím, že nalezneme subslovní součásti slova z dotazu ve slovech v dokumentech [108]. Například OOV slovo skládající se ze subslovních jednotek $\{w_1, w_2, w_3, w_4\}$ můžeme nalézt v mřížce na obrázku 5.6 v částech jiných slov.

5.5.4 Vyhledávání v subslovních reprezentacích

Jako subslovní reprezentaci řeči můžeme použít stejné reprezentace jako ty používané na úrovni slov, tedy například subslovní mřížky, PSPL, WCN. Při vyhodnocení podobnosti dotazu a dokumentu použijeme stejné míry podobnosti jako na slovní úrovni vypočtené se subslovními jednotkami. Dotazy převedeme na základě výslovnosti jejich slov, nejlépe všech variant, na subslovní jednotky.



Obrázek 5.6: Část mřížky s označením subslovních jednotek na slovních hranách (převzato z [108])

Subslovní PSPL, CN

V práci [108] byly použity subslovní reprezentace mřížek PSPL. Aposteriorní pravděpodobnosti subslovních jednotek jsou vypočteny podobným způsobem jako v podkapitole 5.4.1. Pravděpodobnosti subslovních jednotek určitého slova jsou vypočteny z pravděpodobnosti daného slova (více v článku [108]). V experimentu byly porovnány dosažené přesnosti pro OOV dotazy i dotazy se slovy ze slovníku pro slovní PSPL, PSPL založené na písmenech, slabikách a lineární kombinace ohodnocení slovní a subslovní PSPL (písmena). Obě subslovní varianty dosáhly lepší průměrné přesnosti než slovní PSPL, nejlepších výsledků dosáhl kombinovaný přístup slovních i subslovních jednotek.

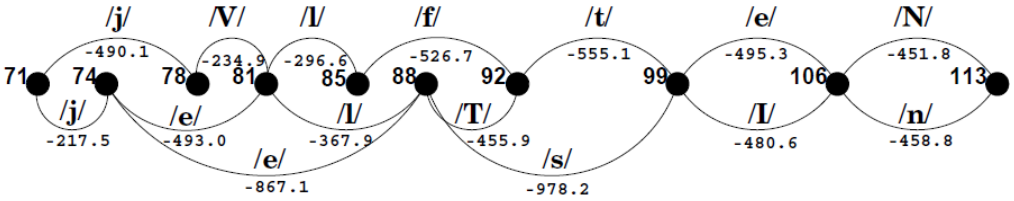
Porovnání pro subslovní varianty CN a PSPL bylo provedeno v práci [26]. Bylo ukázáno, že obě subslovní varianty dosahovaly větší přesnosti při vyhledávání než varianty slovní. Zároveň bylo ukázáno, že obě varianty PSPL dosahovaly lepších výsledků než varianty CN. Nevýhodou PSPL přístupu je větší velikost indexu než u CN přístupů.

V práci [109] byly použity morfémové CN pro vyhledávání v kolekci rozhlasových zpráv. Přestože ASR systém založený na subslovních jednotkách dosahoval menší chybovosti, nebylo dosaženo vylepšení vyhledávání. Tyto výsledky byly potvrzeny v novější práci [110], kde se ukázalo, že použití subslovních jednotek nezlepšuje vyhledávání, dosažené výsledky byly stejné jako u slovních CN.

Vyhledávání nezávislé na slovníku

Jednou z možností použití subslovních reprezentací řeči je vytvoření systému nezávislého na slovníku. ASR systém rozpozná řeč pouze do úrovně dané subslovní jednotky a převede ji do formy subslovní mřížky (například fónové mřížky viz obr. 5.7). Není tedy potřeba použít obsáhlý slovní jazykový model. Tento přístup byl použit v první práci zabývající se vyhledáváním informací v řeči [42], dále byl rozpracován v dizertační práci [65]. Jako subslovní jednotky byly použity trigramové jednotky, skládající se z maximální sekvence souhlásek, obklopené z obou stran sekvencí maximálně dvou samohlásek. Pro vyhledávání byl použit upravený TF-IDF model.

Tento přístup byl použit také v práci [11], kde byla řečová data reprezentována monofónovými nebo bifónovými mřížkami. Po zadání dotazu byla slova z dotazu nalezena v mřížkách metodou vyhledávání klíčových slov (keyword spotting). Takto nalezená slova



Obrázek 5.7: Fónová mřížka obsahující dvě hypotézy pro slovo *yeltsin* (převzato z [65])

z dotazu byla s příslušnými dokumenty indexována a poté byl pro vyhledávání použit upravený TF-IDF model.

Podobný přístup byl zvolen také v práci [31], ve fonetické reprezentaci řeči byla vyhledávána klíčová slova, fráze a booleovské dotazy.

5.6 Kombinovaný přístup

Další možností jak vylepšit úspěšnost vyhledávání informací je použít kombinaci výhod slovního a subslovního přístupu. Pro propojení obou přístupů existuje několik variant [138]:

1. **Kombinace obou přístupů** - vyhledávání je uskutečněno v slovním i subslovním indexu, výsledky jsou poté kombinovány.
2. **Kaskáda slovníků** - slova ze slovníku vyhledáváme ve slovním indexu, OOV slova v subslovním indexu.
3. **Kaskáda vyhledávání** - vyhledávání je provedeno ve slovním indexu, pokud nebude nic nalezeno, pokračuje vyhledávání v subslovním indexu.

V práci [108] byla použita kombinace slovních a subslovních PSPL, podobnost dotazu k dokumentu byla vytvořena jako lineární kombinace podobností ze slovní a subslovní PSPL. Bylo ukázáno, že tento přístup dosáhl lepší průměrné přesnosti než obě varianty samostatně.

Saraclar a Sproat v práci [138] porovnali slovní, subslovní a kombinovaný přístup na třech různých kolekcích s různou chybovostí (20%,40%,50% WER). Ukázali, že se zvyšující se chybovostí ASR přepisů klesá úspěšnost vyhledávání klíčových slov. Z jejich experimentů je vidět, že nejlepších výsledků je dosaženo při použití kombinovaného přístupu formou kaskády vyhledávání.

V práci [89, 88] byl na základě experimentů na korpusu rádiových zpráv vyhodnocen jako nejlepší kombinovaný systém, vyhledávající v slovním indexu pro slova ve slovníku a v subslovním indexu pro OOV slova. Podobný přístup byl použit také v práci [105], tedy vyhledávání v slovním indexu pro slova ze slovníku a v slabikovém indexu pro OOV slova.

Kombinovaná mřížka obsahující slovní i subslovní (fonémy) reprezentace zároveň byla představena v práci [182]. Bylo ukázáno, že tento kombinovaný způsob dosahuje lepších výsledků než slovní nebo subslovní přístupy samostatně.

Hori a kolektiv [56] použili pro vyhledávání v kolekci MIT přednášek kombinovanou CN, složenou ze slovní a fónové CN. Tento přístup dosáhl lepších výsledků než slovní nebo subslovní přístupy samostatně.

V práci [110] byl testován kombinovaný index CN v experimentech s vyhledáváním klíčových slov v řeči a vyhledáváním informací v řeči. Přestože se kombinované indexy ukázaly jako dobré pro vyhledávání klíčových slov v řeči, při vyhledávání informací nepřinesly žádné zlepšení výsledků vyhledávání.

Kapitola 6

Zpětná vazba ve vyhledávání informací

Téměř žádný systém pro vyhledávání informací nemůže uživateli poskytnout všechny relevantní dokumenty k zadanému dotazu. Záleží na typu a uplatnění systému, zda stačí uživateli předložit nějaké relevantní dokumenty, nebo je potřeba nalézt jich co nejvíce - dosáhnout co největší úplnosti[1] (viz podkapitola 3.1). Tato kapitola se bude zabývat možnostmi jak zvýšit výkonnost vyhledávacího systému pomocí úpravy vyhledávaného dotazu, zejména těmi metodami, kdy systém sám upravuje dotaz, ať již s pomocí informací získaných od uživatele, nebo zcela automaticky.

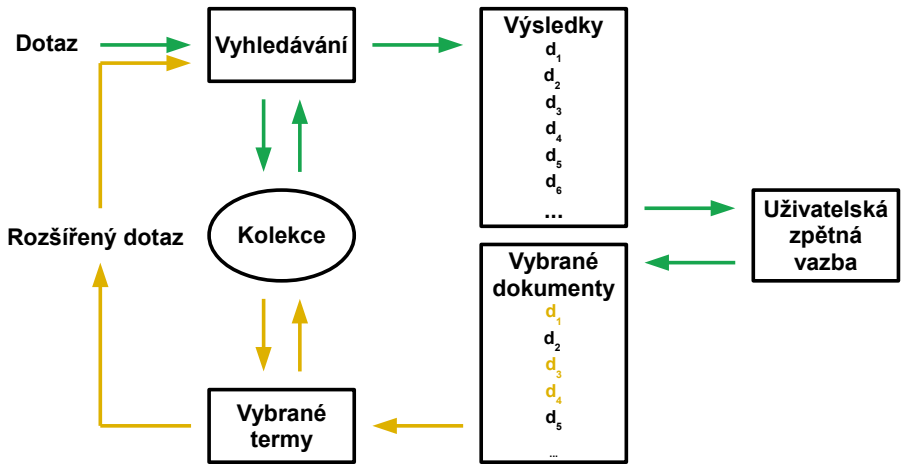
Zpětná vazba může být realizována mnoha formami, můžeme rozlišit tři hlavní způsoby přístupu: *explicitní, implicitní a slepá*[2] nebo-li „pseudo“ zpětná vazba.

- *Explicitní* zpětná vazba je realizována pomocí přímé spolupráce s uživatelem systému, kdy uživatel označí některé dokumenty jako relevantní.
- Jako *implicitní* zpětná vazba bývá označován přístup, kdy systém sbírá informace o relevanci dokumentů na pozadí, pouze z reakcí uživatele [75] (např. které dokumenty uživatel prohlíží, jak dlouho s nimi stráví, atd.).
- *Slepá* zpětná vazba probíhá bez zapojení uživatele, je čistě automatická. Systém vybere dokumenty, které budou dále považovány za relevantní a použije je k rozšíření dotazu.

Metody zpětné vazby začaly být vyvíjeny velice brzy po vývoji samotných metod pro vyhledávání informací, například v pravděpodobnostních modelech z podkapitoly 4.4 je předpoklad získání informace o relevanci dokumentů zapojen již do samotného vyhledávacího mechanismu. Efektivnost zpětné vazby pro zlepšení výsledků vyhledávání byla prokázána v mnoha studiích, například [51, 133, 12]. Možnost důkladného otestování její funkčnosti ale přinesly až TREC[3] úlohy [52] obsahující větší kolekce pro vyhledávání informací zejména z oblasti třídění dokumentů[4].

V následujícím textu budou popsány metody, které lze označit jako klasická, explicitní zpětná vazba (v podkapitole 6.1) a metody slepé zpětné vazby (v podkapitole 6.2). Metody

[1] Recall
[2] Blind relevance feedback
[3] Text REtrieval Conference
[4] Routing



Obrázek 6.1: Znázornění funkčnosti systému se zpětnou vazbou od uživatele

implicitní zpětné vazby se zabývají zejména způsoby, jak z chování uživatele odhadnout relevantní dokumenty. Tyto dokumenty se pak dají dále použít v metodách explicitní, nebo slepé zpětné vazby. Popsány jsou jak základní modely, tak i přístupy vztahující se k experimentům provedeným v této práci. Obecný detailní přehled metod a aspektů zpětné vazby a možností rozšíření dotazu může být nalezen například v rozsáhlých přehledových pracích [129] a [18].

6.1 Klasická zpětná vazba

Základní myšlenkou zpětné vazby[5] ve vyhledávání informací je upravit vyhledávání na víceprůchodový systém, kde cílem druhého průchodu (vyhledávání) je zlepšit výsledky získané v prvním průchodu. Klasická zpětná vazba počítá se zapojením uživatele do tohoto procesu, uživatel po předložení prvních výsledků vyhledávání poskytne systému informaci o relevanci, případně nerelevanci vyhledaných dokumentů. Tato informace je použita k vylepšení (rozšíření) dotazu a poté proběhne znovu vyhledání relevantních dokumentů s tímto rozšířeným dotazem.

Postup je znázorněn na obrázku 6.1 a dá se shrnout takto:

1. Uživatel předloží systému dotaz.
2. Systém vyhledá předpokládané relevantní dokumenty k zadanému dotazu a předloží je uživateli.
3. Uživatel z předložených dokumentů vybere ty, které jsou pro něj relevantní.
4. Na základě informace získané z těchto dokumentů vytvoří systém nový dotaz.
5. Pomocí nově vytvořeného dotazu jsou vyhledány dokumenty a předloženy uživateli.

Získání informace o relevantních dokumentech od uživatele a následnou úpravu dotazu je možné zopakovat i vícekrát, jde tedy o iterativní proces. Vylepšení výsledků vyhledávání při provedení více než jedné iterace zpětné vazby bylo ukázáno například v pracích [51, 1].

[5] Relevance feedback

Předpokladem úspěšnosti tohoto postupu je, že uživatel není schopen vytvořit optimální dotaz, ale je schopen vybrat pro něj relevantní dokumenty. Důvodem pro nevhodně vytvořený dotaz může být například neznalost kolekce dokumentů, v kterých je vyhledáváno, neznalost stylu jakým jsou dokumenty psány (jiná volba slov, synonym) nebo neuvědomění si vlastního požadavku - uživatel neví co přesně chce, dokud nevidí vyhledané výsledky, jeho potřeba se tedy vyvíjí s časem.

Jednu z prvních prací zmiňujících zpětnou vazbu ve vyhledávání informací publikovali v roce 1960 Maron a Kuhns [98], kde navrhovali úpravu dotazu pomocí přidání termů, které úzce souvisí s termy dotazu a v důsledku toho vylepšení výsledků vyhledávání. V dalších letech probíhal vývoj metod zpětné vazby zejména v systému SMART [130] pro vektorový model pro vyhledávání informací (viz podkapitola 4.3). V roce 1965 (znovu publikováno v roce 1971) představil Rocchio experimenty s úpravou dotazu založené na rozšíření dotazu a úpravy vah termů stávajícího dotazu [127].

6.1.1 Rocchio algoritmus pro zpětnou vazbu

Rocchio algoritmus modeluje zapojení zpětné vazby do vektorového modelu pro vyhledávání informací. Přes jeho stáří se jeho varianty stále hojně používají, zejména v úpravě pro použití jako slepá zpětná vazba [17, 179, 180, 57, 172]. Principem tohoto algoritmu je nalézt vektor dotazu q_{opt}, který by měl maximální podobnost s množinou relevantních dokumentů R_q a zároveň minimální podobnost s množinou nerelevantních dokumentů NR_q:

$$q_{opt} = \text{argmax}_q[sim_{q,R_q} - sim_{q,NR_q}], \tag{6.1}$$

kde sim_{q,R_q} je cosinová podobnost vektorů definovaná v rovnici (4.7). Optimální dotaz je tedy vzdálenost mezi centroidy množin relevantních a nerelevantních dokumentů [97]:

$$q_{opt} = \frac{1}{|R_q|} \sum_{d_j \in R_q} d_j - \frac{1}{|NR_q|} \sum_{d_j \in NR_q} d_j. \tag{6.2}$$

Pro výpočet takového optimálního dotazu by tedy bylo nutné znát všechny relevantní i nerelevantní dokumenty. Protože je taková představa nereálná, definoval Rocchio svůj algoritmus takto [127]:

$$q_1 = \alpha q_0 + \beta \frac{1}{|R_U|} \sum_{d_j \in R_U} d_j - \gamma \frac{1}{|NR_U|} \sum_{d_j \in NR_U} d_j, \tag{6.3}$$

kde q_0 je originální vektor dotazu a q_1 je vektor dotazu po první iteraci algoritmu zpětné vazby. Obecně můžeme úpravu dotazu definovat jako iterativní proces [133]:

$$q_{i+1} = \alpha q_i + \beta \frac{1}{|R_U|} \sum_{d_j \in R_U} d_j - \gamma \frac{1}{|NR_U|} \sum_{d_j \in NR_U} d_j, \tag{6.4}$$

kde R_U je množina známých relevantních dokumentů (například označených uživatelem), NR_U je množina známých nerelevantních dokumentů a α, β a γ jsou váhy, určující jak moc chceme ponechat nebo změnit původní dotaz směrem k relevantním nebo nerelevantním dokumentům. Váhy se většinou nastavují tak, aby vliv relevantních dokumentů byl větší než nerelevantních, tedy $\beta > \gamma$, s nastavením $\gamma = 0$ bude systém brát v úvahu

pouze relevantní dokumenty. V práci [133] provedli Salton a Buckley experimenty s různým poměrem β a γ, jako nejlepší nastavení se ukázalo $\alpha = 1$, $\beta = 0,75$ a $\gamma = 0,25$.

Ide v pracích [59, 58] pokračovala v experimentech s Rocchio zpětnou vazbou a představila nové varianty této metody, zejména metoda nazývaná *Ide dec hi* vychází v některých experimentech jako nejlepší varianta Rocchio metody [133]. Tato varianta je založená na tom, že místo množiny nerelevantních dokumentů je použit pouze jeden, a to ten s nejvyšším skóre (největší podobností podle vzorce (4.7)).

Opačným přístupem může být považování všech dokumentů v kolekci, kromě těch označených jako relevantní, za nerelevantní a jejich následné použití jako NR_U v rovnici (6.4). Tento přístup byl testován například v práci [145] v oblasti třídění dokumentů, kde bylo ukázáno, že lepší než všechny nerelevantní dokumenty je použít pouze nerelevantní dokumenty ze stejné oblasti jako je dotaz. Výběr nerelevantních dokumentů ze stejné oblasti jako dotaz byl realizován výběrem prvních k dokumentů seřazených podle podobnosti s dotazem, které nebyly označeny jako relevantní. Další možností bylo vybrat nerelevantní dokumenty podle nastaveného prahu podobnosti, kdy byly vybrány ty dokumenty, které měly podobnost větší než stanovený práh. Práh byl nastaven podle celkové podobnosti dotazu a dokumentů. V případě, kdy byla obecně podobnost malá, byl nastaven práh tak, aby byla menší oblast dotazu, než v případě velké podobnosti dotazu a dokumentů.

6.1.2 Pravděpodobnostní zpětná vazba

Další možností přístupu k realizaci zpětné vazby může být použití klasifikátoru. Pomocí *Naive Bayes* modelu můžeme pro každý term t_i získat pravděpodobnost jeho výskytu $x_i = 1$ nebo $x_i = 0$ v dokumentu v závislosti na relevanci dokumentu $\hat{P}(x_i = 1|R_{q,d_j})$:

$$\begin{aligned} p_i &= \hat{P}(x_i = 1|R_{q,d_j} = 1) = \frac{|R_U(t)|}{|R_U|} \\ u_i &= \hat{P}(x_i = 1|R_{q,d_j} = 0) = \frac{n_t - |R_U(t)|}{N - |R_U|}, \end{aligned} \tag{6.5}$$

kde R_{q,d_j} je hodnotící funkce, přiřazující každému dokumentu hodnotu 1 pokud je relevantní, 0 v opačném případě. $R_U(t)$ je podmnožina známých relevantních dokumentů obsahujících term t, N je celkový počet dokumentů, n_t je počet dokumentů obsahujících term t.

Pomocí této metody je možné upravit váhy termů v dotazu na základě jejich obsahu v relevantních dokumentech. Pokud hodnoty $p_i = \hat{P}(x_i = 1|R_{q,d_j} = 1)$ a $u_i = \hat{P}(x_i = 1|R_{q,d_j} = 0)$ ze vzorce (6.5) dosadíme do vzorce (4.11) v podkapitole 4.4, získáme podobnost dokumentu a dotazu pro *BIM* (viz kapitola 4.4) se zapojením zpětné vazby [122, 133]:

$$sim_{d_j,q} = \sum_{t_i \in d} t_i \log\left(\frac{|R_U(t)|}{|R_U| - |R_U(t)|} \div \frac{n_t - |R_U(t)|}{N - |R_U| - n_t + |R_U(t)|}\right). \tag{6.6}$$

Výpočet p_i a u_i v rovnici (6.5) se často upravuje na formu:

$$\begin{aligned} p_i &= \hat{P}(x_i = 1|R_{q,d_j} = 1) = \frac{|R_U(t)| + 0,5}{|R_U| + 1} \\ u_i &= \hat{P}(x_i = 1|R_{q,d_j} = 0) = \frac{n_t - |R_U(t)| + 0,5}{N - |R_U| + 1}, \end{aligned} \tag{6.7}$$

aby se předešlo nulovosti výrazu při malých hodnotách $|R_U(t)|$ a $|R_U|$. Další alternativy výpočtu p_i a u_i jsou představeny například v pracích [121, 175, 133].

6.1.3 Zpětná vazba v booleovském systému

Pro zpětnou vazbu v booleovském systému existují dvě možnosti realizace [50]. První možností je vybrat z uživatelem vyhodnocených relevantních dokumentů důležité termy (podle nějaké váhy, nebo frekvence termů) a ty předložit zpět uživateli, aby sám upravil dotaz. Druhou možností je použít algoritmus, který bude sám vytvářet, nebo upravovat uživatelem zadané booleovské dotazy (experimenty s takovým algoritmem jsou představeny v podkapitole 7.5.1). Ukázkou takového systému je například expertní systém pro vyhledávání v online katalozích představený v práci [76]. Expertní systém zde navržený upravuje booleovské dotazy například vynecháním nedůležitého termu z dotazu, záměnou operátoru AND za OR, přidáním termu, který má velkou frekvenci v relevantních výsledcích nebo vynecháním synonyma (termy spojené operátorem OR), pokud jeho vyhledáním získáme příliš málo výsledků.

6.1.4 Porovnání metod pro zpětnou vazbu

Porovnání jednotlivých metod jak realizovat zpětnou vazbu je provedeno v práci [133]. Salton a Buckley zde porovnávají dvanáct metod pro realizaci zpětné vazby (různé varianty Rocchio algoritmu a pravděpodobnostních modelů) na šesti různých kolekcích. Zjistili, že jednotlivé metody podávají konzistentní výsledky, tedy jejich pořadí podle výkonnosti je stejné na různých kolekcích. Nejlépe se osvědčila metoda *Ide dec hi* [59, 58] a obecně varianty vektorového modelu podávaly lepší výsledky než pravděpodobnostní modely.

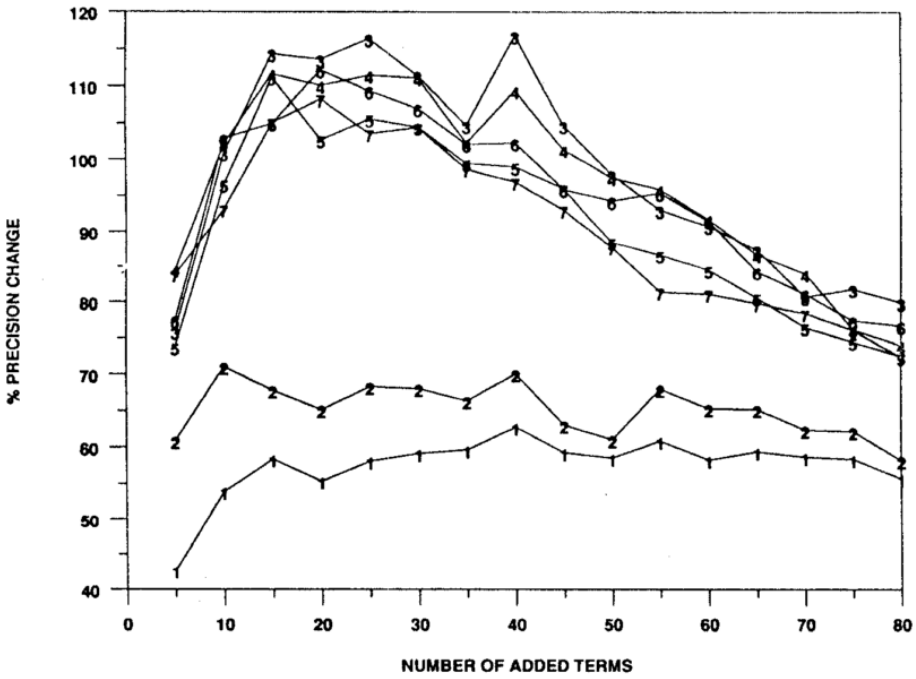
6.1.5 Výber termů pro rozšíření dotazu

Kromě výběru metody pro realizaci zpětné vazby ovlivní výsledné zlepšení výsledků vyhledávání také způsob výběru termů pro následné rozšíření dotazu z relevantních dokumentů.

Ohodnocení termů

Pro výběr termů k rozšíření dotazu pomocí zpětné vazby je nutné termy v relevantních dokumentech nejprve nějakým způsobem ohodnotit. Binární váha termu, ohodnocující pouze jeho přítomnost nebo nepřítomnost v dokumentu, není pro tento účel vhodná. V práci [133] bylo ukázáno, že vážení termů pomocí frekvence, s jakou se objevují v relevantních dokumentech, přineslo vylepšení výsledků vyhledávání. V práci [49] byly porovnány různé způsoby vážení termů použité pro jejich následné seřazení podle předpokládané důležitosti. Váhy byly zvoleny jako kombinace idf_t (viz rovnice (4.9)), frekvence termu v relevantních dokumentech $tf_{t,R_U(t)}$ a počtu relevantních dokumentů ve kterých se vyskytuje daný term. Jako nejlepší se ukázaly váhy termu obsahující frekvenci termu oproti vahám obsahujícím pouze počet dokumentů, ve kterých se term vyskytuje. Úplně nejlepších výsledků dosáhla upravená TF-IDF míra (viz podkapitola 4.3.1):

$$w_t = tf_{t,R_U(t)} \cdot idf_t, \tag{6.8}$$



Obrázek 6.2: Procentuální zlepšení průměrné přesnosti vyhledávání v závislosti na přidání různého množství termů jako rozšíření dotazu zpětnou vazbou (převzato z [51])

kde $tf_{t,R_U(t)}$ je frekvence termu v známých relevantních dokumentech. V navazující práci [51] byla TF-IDF míra dále porovnána s pravděpodobnostními váhami zahrnujícími také frekvence termu v nerelevantních dokumentech. Nově porovnávané váhy nedosáhly lepších výsledků než původně navržená TF-IDF míra ze vzorce (6.8). Váhy, které dosáhly podobných výsledků, byly například váha termu v pravděpodobnostním modelu (viz podkapitola 6.1.2):

$$w_t = \log \frac{p_i(1-u_i)}{u_i(1-p_i)}, \tag{6.9}$$

nebo váha kombinující idf_t a pravděpodobnostní váhu, jako varianta RSV[6] váhy definované Robertsonem v práci [124]:

$$w_t = idf_t \cdot (p_i - u_i), \tag{6.10}$$

kde p_i a u_i jsou definovány vzorcem (6.7). Porovnání jednotlivých metod vážení termů je možné vidět na obrázku 6.2, kde váhy z rovnic (6.8), (6.9) a (6.10) odpovídají číslům metod 6, 3 a 4.

Počet termů

V prvních experimentech na malých kolekcích pro vyhledávání informací vycházely nejlepší výsledky při přidání velkého množství termů pro rozšíření dotazu [44, 133]. Závěrem

[6] Robertson selection value

těchto prací bylo, že čím více termů je do dotazu přidáno, tím většího zlepšení je možné dosáhnout, nebo minimálně nedojde k žádnému zhoršení výsledků. V práci [12] bylo experimentováno s přidáním až 4000 termů pomocí Rocchio algoritmu, výsledky testů ukázaly, že po přidání 500 a více termů se již výsledky vyhledávání nezlepšovaly, ale zůstávaly konstantní.

V pracích [49, 51] bylo experimentováno s přidáním 0 až 80 termů v pravděpodobnostním modelu, kde bylo naopak ukázáno, že přidání 20 až 40 vybraných termů je lepší než přidání všech termů. Výsledky experimentů jsou vidět na obrázku 6.2. Tyto termy byly vybrány podle některé míry jejich důležitosti. Argumentem pro přidání pouze menšího množství vybraných termů je míra jejich důležitosti - ta bude s přidáním všech termů klesat a může tak dojít k menší specializaci dotazu a tím pádem vyhledání i nerelevantních dokumentů. Harman v práci [51] doporučuje vybírat 20 termů pro rozšíření dotazu s ohledem na vyvážení přínosu pro zlepšení výsledků a zároveň zachování rychlosti vyhledávání. Zobecnění doporučení pro výběr dvaceti termů je problematické, Harman však předpokládá, že u kolekcí s podobně dlouhými dokumenty (průměrně 50 termů na dokument po vyloučení běžných slov) by mělo platit, že takovéto množství termů je nejlepší pro rozšíření dotazu zpětnou vazbou.

6.1.6 Negativní zpětná vazba

Metody představené v této kapitole zahrnují také použití informace z nerelevantních dokumentů. Na nerelevantní dokumenty se dá pohlížet dvojím způsobem:

1. Dokumenty explicitně uživatelem označené jako nerelevantní.
2. Dokumenty, které nebyly uživatelem označené jako relevantní.

Metody zpětné vazby pro vektorový a pravděpodobnostní model obecně používají druhou skupinu dokumentů, tedy nijak neoznačené dokumenty. Metody používající dokumenty explicitně označené jako nerelevantní se dají nazvat jako *negativní zpětná vazba*[7]. V práci [129] jsou detailně shrnuty metody a postupy této negativní zpětné vazby a vysvětleny hlavní důvody, proč je její použití problematické:

- Není zřejmé, jakým způsobem zapojit negativní zpětnou vazbu do systému. Například termy, které se vyskytují v nerelevantních dokumentech, se mohou vyskytovat i v relevantních dokumentech a může záležet na jejich kontextu, jak se projeví ve výsledcích vyhledávání.
- Označení dokumentu za nerelevantní je pro uživatele mnohem složitější, než označení za relevantní. Je otázkou, kdy už se dokument stává nerelevantním a kdy pouze není relevantní.
- Časová náročnost označení nerelevantních dokumentů je mnohem větší, než označení pouze relevantních dokumentů. Uživatel tedy nemusí být ochotný tyto dokumenty označovat.

Obecně tedy metody negativní zpětné vazby nejsou doporučeny k použití, vzhledem k jejich nejistému efektu na výsledky vyhledávání.

[7] Negative relevance feedback

6.1.7 Aktivní zpětná vazba

Na rozdíl od předchozích popsaných metod, zabývajících se možnostmi jak vylepšit algoritmy zpětné vazby, aktivní zpětná vazba[8] se zabývá problémem jaké dokumenty předložit uživateli k hodnocení. Základní myšlenkou tohoto přístupu je, že označením dvou relevantních dokumentů s téměř stejným obsahem získá systém méně informací, než pokud uživatel označí dva dokumenty s různým relevantním obsahem. Cílem tedy je roztřídit vyhledané dokumenty do shluků podle jejich podobnosti a poté předložit uživateli reprezentativní vzorky jednotlivých shluků. Problém byl představen v práci [143], kde byly porovnány metody předložení uživateli prvních k dokumentů, předložení každého *k-tého* dokumentu a předložení centroidů vytvořených shluků dokumentů. Nejlepších výsledků dosáhla metoda s použitím centroidů shluků dokumentů. V práci [177] byla navržena nová metoda, vybírající dokumenty na základě jejich relevance, hustoty a rozdílnosti. V experimentech dosáhla tato metoda lepších výsledků než metody navržené v práci [143].

6.2 Slepá zpětná vazba

Na rozdíl od předchozích metod, je *slepá zpětná vazba* nebo-li *pseudo zpětná vazba* zcela automatickou metodou umožňující úpravu dotazu bez zásahu uživatele. Úloha uživatele je zde simulována úvahou, že mezi prvními k vyhledanými dokumenty jsou všechny, nebo alespoň většina z nich, relevantní a tyto dokumenty (budou dále označovány jako *pseudo relevantní*) jsou pak použity pro zpětnou vazbu.

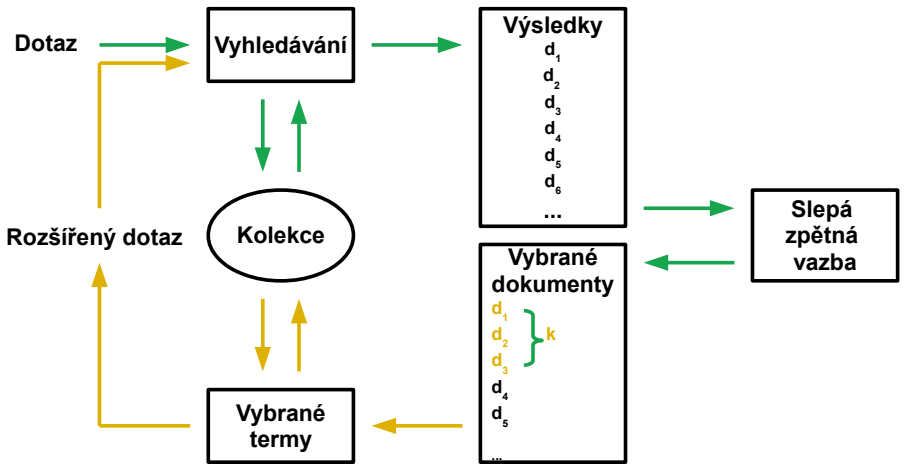
Postup je znázorněn na obrázku 6.3 a dá se popsat takto:

1. Uživatel předloží systému dotaz.
2. Systém vyhledá předpokládané relevantní dokumenty k zadanému dotazu, seřadí je podle vypočteného skóre relevance (podle zvolené vyhledávací metody).
3. Systém vezme prvních k dokumentů, ty budou nadále považovány za relevantní.
4. Na základě informace získané z těchto pseudo relevantních dokumentů vytvoří systém nový dotaz. Většinou je vybráno několik nejlepších termů ve smyslu jeho nejvyšší váhy a ty jsou použity pro rozšíření dotazu.
5. Pomocí nově vytvořeného dotazu jsou vyhledány dokumenty a předloženy uživateli.

Předpokládá se, že mezi těmito dokumenty získanými vyhledáním rozšířeného dotazu bude více relevantních dokumentů a tyto dokumenty budou lépe hodnoceny (výše v pořadí vyhledaných dokumentů).

Myšlenka slepé zpětné vazby byla poprvé představena v práci [32], kde bylo použito prvních k vyhledaných dokumentů pro výpočet p_i a u_i v pravděpodobnostním modelu (viz podkapitola 6.1.2) v případě, že nejsou označeny žádné relevantní dokumenty. Výsledky experimentů ukázaly, že tato metoda dosahuje lepších výsledků než samostatné vyhledávání bez použití zpětné vazby. V této práci byl také označen jeden ze základních problému slepé zpětné vazby, tím je *posun dotazu*[9]. Posunem dotazu je myšlena situace, kdy se rozšířením dotazu změní jeho smysl a následně se zhorší výsledky vyhledávání. Tento problém bude popsán v podkapitole 6.2.2.

[8] Active feedback
[9] Query drift



Obrázek 6.3: Znázornění funkčnosti systému se slepou zpětnou vazbou

6.2.1 Funkčnost slepé zpětné vazby

Většina studií experimentujících se slepou zpětnou vazbou potvrzuje zlepšení výsledků vyhledávání (z hlediska vyhodnocení nějakou mírou kombinující přesnost a úplnost viz podkapitola 3.1), s vylepšením dosahujícím deset a více procent [18] v různých úlohách vyhledávání informací. Práce potvrzující tyto výsledky jsou například [161, 77, 3, 101, 85, 172].

Slepá zpětná vazba obecně jako technika funguje dobře (tedy zlepšuje výsledky vyhledávání) v případě „dobrých“ počátečních dotazů, tedy těch, jejichž vyhledáním získáme relevantní dokumenty, a špatně v případě dotazů nevedoucích k relevantním dokumentům [129]. Za dobré dotazy se obecně považují dlouhé, přesně formulované dotazy [55]. Problém „špatných“ dotazů je možné řešit lepšími vyhledávacími metodami nebo lepšími metodami zpětné vazby.

6.2.2 Problém posunu dotazu

Problém posunu dotazu byl poprvé popsán v práci [32]. Je to jeden z hlavních problémů, který může nastat při použití slepé zpětné vazby. Posunem dotazu je míněna změna jeho smyslu z hlediska požadavku uživatele, a tím způsobené následné vyhledání nerelevantních dokumentů. K tomuto efektu dojde v případě, že je pro zpětnou vazbu použita množina dokumentů obsahující málo relevantních dokumentů nebo vůbec žádné. Termy vybrané z těchto dokumentů pro expanzi dotazu poté mohou být nerelevantní k požadavku uživatele.

K posunu dotazu může dojít i v případě, kdy jsou pro jeho rozšíření vybrány termy, které souvisí pouze s jedním slovem původního dotazu, ale ne se smyslem celého dotazu. Posun dotazu a následné zmenšení přesnosti vyhledávání může také nastat přidáním termů, které dotaz zobecní. Pokud například k dotazu na konkrétní herečku přidáme při rozšíření dotazu termy „herečka“ a „film“, může nastat situace, kdy při vyhledání rozšířeného dotazu bude lépe hodnocen dokument o jiné herečce, protože v něm tato slova měla vyšší váhy [16].

Tento problém byl adresován v práci [101], kde bylo cílem zvýšit přesnost vyhledávání pro první vyhledané dokumenty a tím zvýšit šanci na výskyt opravdu relevantních

dokumentů mezi pseudo relevantními dokumenty. Základní myšlenkou prezentovaného přístupu bylo přehodnocení vybraných prvních l dokumentů pomocí dodatečné informace a jejich nové seřazení. Z tohoto nově seřazeného seznamu bylo teprve vybráno prvních k dokumentů (20 v tomto případě) použitých pro slepou zpětnou vazbu. Výsledky těchto experimentů ukázaly, že efekt posunu dotazu je možné velmi omezit a zlepšit tím přesnost vyhledávání.

V další práci zabývající se tímto problémem [13] bylo testováno použití velkého množství termů (500 termů) pro rozšíření dotazu. Těchto 500 termů bylo vybráno z prvních 30 vyhledaných dokumentů. Výsledky práce ukazují, že je možné tímto způsobem dosáhnout zvýšení přesnosti vyhledávání oproti systému bez slepé zpětné vazby.

6.2.3 Metody slepé zpětné vazby

Většina prací zabývajících se slepou zpětnou vazbou používá upravené metody klasické zpětné vazby, viz podkapitola 6.1. Jak již bylo zmíněno, dojde pouze k nahrazení uživatelem vybraných relevantních dokumentů za prvních k vyhledaných dokumentů. Metody slepé zpětné vazby jsou poté použity pro úpravu dotazu jeho rozšířením, případně i změnou vah původních termů dotazu. Změna vah termů původního dotazu je součástí například Rocchio algoritmu (podkapitola 6.1.1) a pravděpodobnostní zpětné vazby (podkapitola 6.1.2). V práci [51] bylo ukázáno, že nejlepší výsledky lze dosáhnout kombinací obou přístupů, ale největší vliv na vylepšení vyhledávání má část rozšíření dotazu.

Výběr váhové funkce

Pro ohodnocení termů v pseudo relevantních dokumentech je většinou použita stejná váhová funkce jako je použita ve vyhledávacím algoritmu. Ve vektorovém modelu pro vyhledávání informací (viz podkapitola 4.3) je tedy nejčastěji použita některá z variant TF-IDF funkce (podkapitola 4.3.1):

$$w_t = \sum_{d_j \in R_P} tf_{t,d_j} \cdot idf_t, \tag{6.11}$$

kde R_P je množina pseudo relevantních dokumentů. V pravděpodobnostním modelu z podkapitoly 4.4 bývá použito ohodnocení termu definované vzorcem (6.9), kde p_i nahradíme pravděpodobností výskytu termu v pseudo relevantních dokumentech $P(t|R_P)$ a u_i pravděpodobností termu v celé kolekci $P(t|C)$:

$$w_t = \log \frac{P(t|R_P)(1 - P(t|C))}{(1 - P(t|R_P))P(t|C)}. \tag{6.12}$$

Dalšími možnými váhami jsou například RSV váha definovaná v práci [124], kde je použito stejné nahrazení p_i a u_i jako v předchozím případu:

$$w_t = \sum_{d_j \in R_P} w_{t,d_j} \cdot (P(t|R_P) - P(t|C)), \tag{6.13}$$

kde w_{t,d_j} je nějaká váha termu t v pseudo relevantních dokumentech. Další možností je KLD[10] váha definovaná v práci [17]:

[10] Kullback-Leibler distance

$$w_t = P(t|R_P) \cdot \log \frac{P(t|R_P)}{P(t|C)}, \tag{6.14}$$

nebo Chi-kvadrát[11] váha [17]:

$$w_t = \frac{(P(t|R_P) - P(t|C))^2}{P(t|C)}. \tag{6.15}$$

Pravděpodobnosti $P(t|R_P)$ a $P(t|C)$ lze spočítat pomocí odhadu maximální věrohodnosti (MLE) uvedeného v rovnici (4.19). Varianta tohoto přístupu byla s úspěchem použita v práci [172], kde byl pro odhad těchto pravděpodobností použit poměr pseudo relevantních dokumentů obsahujících term t:

$$\hat{p}(t|R_P) = \frac{|R_P(t)|}{|R_P|}. \tag{6.16}$$

V jiných pracích [13, 17] dosahovala ale tato varianta horších výsledků než odhad pomocí MLE, proto se většinou používá pouze pro RSV váhu, kde je to standardní způsob výpočtu p_i a u_i [124].

Další varianty váhových funkcí lze nalézt například v pracích [17, 172]. Experimenty v práci [17] potvrzují předchozí výsledky experimentů s klasickou zpětnou vazbou [133, 51] ukazující, že jednotlivé váhové funkce použité pro ohodnocení termů pro rozšíření dotazu přinášejí v konečném důsledku přibližně stejné vylepšení vyhledávání.

V práci [172] bylo testováno 149 variant různých váhových funkcí odvozených z 23 základních funkcí. Funkce byly rozděleny na funkce hodnotící z hlediska vnitřku dokumentu[12] (příkladem je TF), funkce hodnotící z hlediska vztahu mezi dokumenty[13] (například IDF) a jejich kombinace (TF-IDF). Jako nejlepší varianty hodnotících funkcí se ukázaly ty funkce, které kombinovaly normalizovanou frekvenci termu s nějakou vahou hodnotící z hlediska vztahu mezi dokumenty, nejlepší vyšla z experimentů funkce Chi-kvadrát. Mezi výsledky jednotlivých metod ale nebyl statisticky významný rozdíl.

Ye a Huang v práci [180] navrhli tři nové varianty výpočtu frekvence termu TF. Cílem bylo navrhnout nové míry tak, aby zahrnovaly kromě samotné frekvence termu také váhu pseudo relevantního dokumentu, relativní důležitost termu v rámci daného dokumentu a jeho blízkost k původnímu dotazu. Experimenty s novými TF mírami byly provedeny v rámci BM25 modelu pro vyhledávání informací a Rocchio algoritmu pro zpětnou vazbu, kde jsou míry použity pro ohodnocení termů vybíraných pro rozšíření dokumentu.

$$w_t = \sum_{j=0}^{n} (\lambda_j \cdot tf_j(t)) \cdot idf_t, \tag{6.17}$$

kde $tf_j(t)$ je transformační technika j pro výpočet TF termu t a λ_j vyjadřuje její důležitost. Takto navržený systém byl porovnáván s metodou jazykového modelování s použitím zpětné vazby formou modelu relevance RM3 (viz vzorce (6.21) a (6.22) v následující podkapitole). Metody byly zkoumány při výběru 5, 10, 15, 20, 30, 50 pseudo relevantních dokumentů a z nich 10, 15, 20, 25, 30, 35, 50 termů. Výsledky experimentů ukazují, že TF míry navržené v této práci dosahují lepších výsledků než klasická TF míra i než RM3 model.

[11] Chi-square
[12] Intra-document
[13] Inter-document

Jazykové modelování

Slepá zpětná vazba v jazykovém modelování pro vyhledávání informací může být provedena stejným způsobem jako u ostatních modelů, tedy výběrem termů z pseudo relevantních dokumentů a následným rozšířením dotazu a jeho opětovným vyhledáním. Tento přístup byl použit v pracích [100, 103, 112]. V práci [112] byla představena funkce pro ohodnocení termů v prostředí jazykového modelování, cílem je vybrat termy s velkou pravděpodobností v pseudo relevantních dokumentech a malou pravděpodobností v celé kolekci:

$$w_t = \sum_{d_j \in R_P} \log \frac{P(t|\theta_{d_j})}{P(t|\theta_C)}, \tag{6.18}$$

kde $P(t|\theta_C)$ je pravděpodobnost termu v celé kolekci a $P(t|\theta_{d_j})$ je pravděpodobnost v pseudo relevantním dokumentu.

Alternativou přímého přidávání termů a tím rozšíření dotazu, může být vytvoření nového jazykového modelu pro dotaz. Jazykový model specifikuje rozložení pravděpodobnosti termů a nejlepší termy jsou tedy ty s největší pravděpodobností. Jeden ze způsobů, jak využít jazykové modelování pro slepou zpětnou vazbu, byl představen v práci [185]. Zhai a Lafferty zde argumentují tím, že rozšíření dotazu klasickým způsobem je proti základní myšlence využití jazykového modelování ve vyhledávání informací a je tedy vhodné použít nějaký přístup založený na jazykovém modelu[14]. Navrhli přístup jak odhadnout nový model dotazu pro použití v KL-divergence metodě pro vyhledávání informací popsané v podkapitole 4.5.3. Je použit klasický smíšený generativní jazykový model dotazu vytvořený z pseudo relevantních dokumentů v kombinaci s modelem celé kolekce dokumentů:

$$\log P(R_P|\theta_q) = \sum_{d_j \in R_P} \sum_t c(t,d) \log((1-\lambda)P(t|\theta_q) + P(t|\theta_C)). \tag{6.19}$$

EM[15] algoritmus je poté použit na odhad modelu dotazu, tak aby maximalizoval věrohodnost pseudo relevantních dokumentů. Model bývá nazýván SMM[16].

Další možný přístup založený na modelu relevance byl navržen v práci [82]. Myšlenkou přístupu je, že dotaz i pseudo relevantní dokumenty byly vygenerovány nějakým modelem relevance θ_R. Odhad pravděpodobnosti termu v tomto modelu je pak definován:

$$\hat{p}(t|\theta_R) = \sum_{d_j \in R_P} P(d_j) P(t|\theta_{d_j}) \prod_{i=1}^{k} P(q_i|\theta_{d_j}). \tag{6.20}$$

Tento model bývá často označován jako RM1[17]. Variantou tohoto modelu je model nazývaný RM2:

$$\hat{p}(t|\theta_R) = \prod_{i=1}^{k} \sum_{d_j \in R_P} P(q_i|\theta_{d_j}) \frac{P(d_j)P(t|\theta_{d_j})}{P(t)}. \tag{6.21}$$

[14] Model-based
[15] expectation-maximization
[16] Simple Mixture Model
[17] Relevance model 1

Pravděpodobnost termu v modelu rozšířeného dotazu q' je pak často odhadnuta pomocí lineární kombinace (Jelinek-Mercer vyhlazování) modelu původního dotazu a modelu relevance, tento způsob bývá označován jako RM3, při použití RM1 odhadu, a RM4, při použití RM2 odhadu:

$$\hat{p}(t|\theta_{q'}) = \lambda P(t|\theta_R) + (1-\lambda)P(t|\theta_q). \tag{6.22}$$

V experimentech v práci [82] byl model RM1 porovnán s rozšířením dotazu pomocí váhové funkce (6.18) navrženém v práci [112] a nebylo prokázáno zlepšení výsledků vyhledávání oproti tomuto přístupu.

V práci [57] byl porovnáván model relevance RM3 s dvěma variantami KLD váhové funkce, funkcí ze vzorce (6.14) a normalizovanou variantou:

$$w_t = \sum_{d_j \in R_P} \frac{P(t|R_P) \cdot \log \frac{P(t|R_P)}{P(t|C)}}{|R_P|}. \tag{6.23}$$

Model relevance byl použit s KL divergence metodou vyhledávání informací, váhové funkce KLD s pravděpodobnostními modely (BM25), úprava vektoru nového dotazu byla provedena pomocí upraveného Rocchio algoritmu. Nejlepších výsledků dosáhla kombinace metod BM25 a normalizované KLD, ale mezi výsledky jednotlivých metod nebyly velké rozdíly. Výsledky jednotlivých metod pro vyhledávání informací byly srovnatelné.

V práci [92] bylo porovnáno pět různých metod pro odhad jazykového modelu dotazu. Metody byly porovnávány v prostředí KL divergence modelu pro vyhledávání informací, použity byly varianty modelu relevance RM3 a RM4, smíšený model SMM ze vzorce (6.19) a jeho varianta s Dirichlet vyhlazováním DMM [185], a model minimalizace vzdálenosti[18] představený v práci [160]. Nejlepších výsledků dosáhl model RM3 a SMM.

Analýza kontextu

Kromě předchozích popsaných metod lze termy pro rozšíření dotazu vybrat i jinými způsoby. Například v práci [176] byla představena metoda *lokální kontextové analýzy* založená na rozšíření dotazu *koncepty* místo jednotlivými termy. Koncept je v této souvislosti definován jako skupina sousedících podstatných jmen vybraná z prvních k vyhledaných dokumentů (v práci byly místo celých dokumentů použity jejich pasáže kvůli délce dokumentů). Pomocí vzorce (6.24) je vypočtena korelační váha $c_{u,v}$ mezi termem dotazu u a konceptem v:

$$c_{u,v} = \sum_{d_j} w_{u,j} \cdot w_{v,j}, \tag{6.24}$$

$w_{u,j}$ je frekvence termu dotazu u v dokumentu j a $w_{v,j}$ je frekvence konceptu v v dokumentu j. Korelační váhy všech termů dotazu pro konkrétní koncept jsou poté zkombinovány a výsledná korelace dotazu a konceptu je použita pro seřazení konceptů. Nejlepších m konceptů je poté použito pro rozšíření dotazu.

[18] Divergence Minimization Model

Pseudo nerelevantní dokumenty

V práci [117] byl představen koncept pseudo nerelevantních dokumentů, tedy dokumentů u nichž je velice nepravděpodobné, že by byly relevantní. Myšlenkou této metody je najít mezi vysoce hodnocenými dokumenty ty, které jsou nerelevantní a poté je použít jako zdroj informace pro výběr vhodných termů z pseudo relevantních dokumentů. Pseudo nerelevantní dokumenty byly definovány jako dokumenty s vysokým skóre mimo množinu prvních k dokumentů, které jsou zároveň velmi nepodobné prvním k dokumentům. Dokumenty byly vybrány tak, že z množiny dokumentů s vysokým skóre byly odstraněny ty, které jsou podobné prvním k dokumentům a zbytek byl označen za pseudo nerelevantní dokumenty. Z pseudo relevantních dokumentů pak byly vybrány pro rozšíření dotazu ty termy, které rozlišovaly tyto dokumenty od pseudo nerelevantních dokumentů. Použitím tohoto přístupu bylo dosaženo zlepšení výsledků vyhledávání oproti použití pouze slepé zpětné vazby s pseudo relevantními dokumenty.

6.2.4 Kombinace metod zpětné vazby

Experimenty s metodami pro výběr termů pro rozšíření dotazu formou slepé zpětné vazby ukazují, že jednotlivé metody, přestože dosahují podobných zlepšení výsledků vyhledávání, vybírají jiné termy [133, 51, 17, 172].

Lee v práci [84] využil tohoto předpokladu k vytvoření různých rozšířených dotazů. Vyhledáním těchto dotazů dojde k získání různých dokumentů. Experimenty zveřejněné v citované práci ukazují, že kombinací těchto výsledků vyhledání různých dotazů lze dosáhnout výrazného vylepšení vyhledávání. V experimentech byl použit vektorový model pro vyhledávání informací a byla kombinována zpětná vazba pomocí variant Rocchio algoritmu, Ide dec hi a pravděpodobnostní zpětná vazba. Pro zpětnou vazbu bylo použito prvních 30 vyhledaných dokumentů.

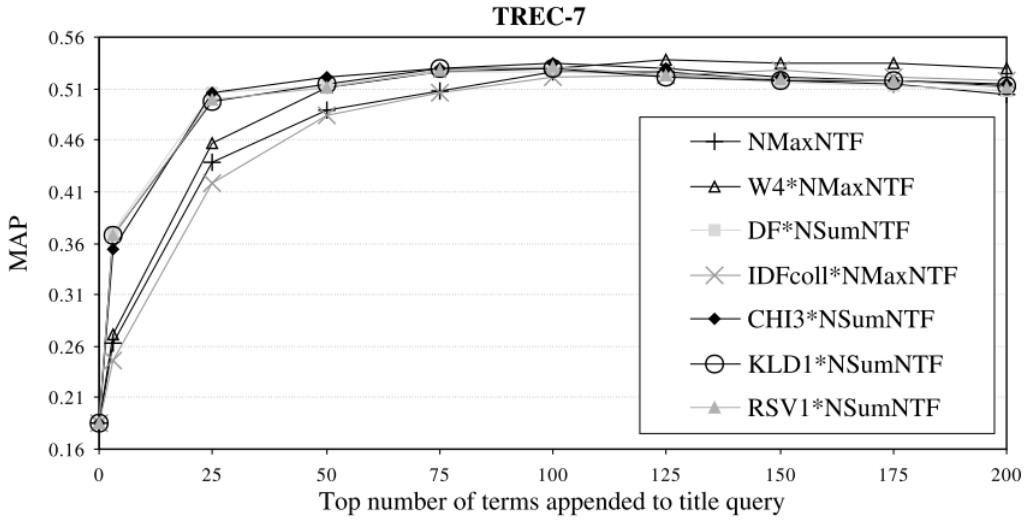
Kombinací různých metod ohodnocení termů se zabývá práce [19]. V této práci bylo popsáno a testováno několik způsobů jak metody kombinovat. První možností je použít lineární kombinaci ohodnocení vzniklých z různých metod (viz popis experimentů v práci [84]). Problémem zde je, že hodnotící funkce pracují na jiném principu a ohodnocení tak není vzájemně porovnatelné. Další možností je pomocí většinového hlasování nad množinami termů získanými různými metodami vybrat ty termy, které se vyskytly nejčastěji, a ty pak použít na rozšíření dotazu [19].

6.2.5 Výběr termů pro rozšíření dotazu

Množství termů, které je vhodné vybrat pro rozšíření dotazu je jedním z problémů slepé zpětné vazby. Pro výběr termů je možné použít doporučení v podkapitole 6.1.5 týkající se obecně zpětné vazby. Na experimenty popsané v dané podkapitole [49, 51] navazuje práce [91], kde bylo experimentováno s výběrem až 30 termů. Na základě výsledků experimentů bylo doporučeno používat pro rozšíření dotazu 1 - 20 termů.

V práci [17] bylo otestováno přidání 10 - 100 termů s krokem po 10 termech. Výsledky experimentů nebyly jednoznačné, největšího zlepšení vyhledávání bylo dosaženo při různém množství přidaných termů, většinou v rozmezí 40 - 70 termů. Větší množství termů bylo přidáno také v práci [176], bylo použito 70 konceptů, kde konceptem je myšlena skupina sousedících podstatných jmen. Ve většině případů ale koncept odpovídal pouze jednomu slovu.

Práce [15] se zabývá výběrem „dobrých“ termů pro rozšíření dotazu. Motivací byla



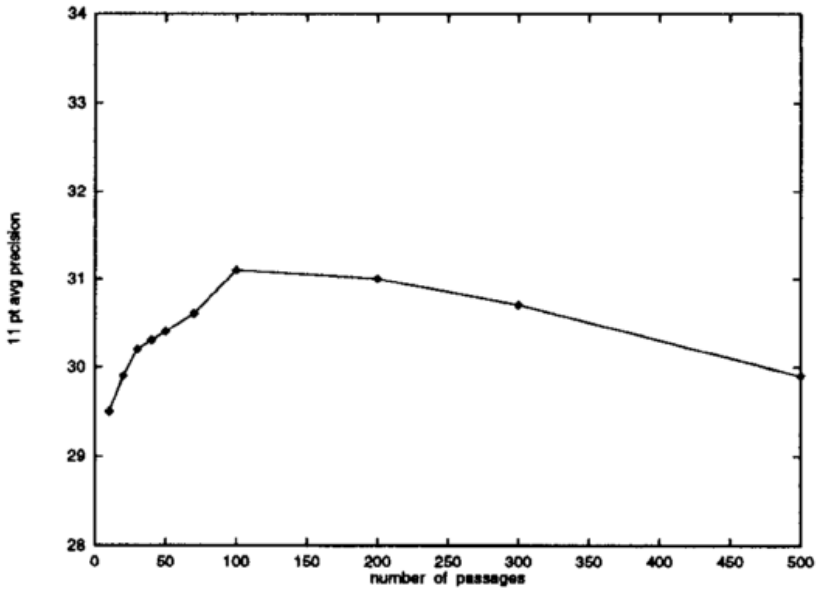
Obrázek 6.4: Porovnání vlivu množství termů použitých pro rozšíření dotazu na výsledek vyhledávání (měřený pomocí MAP) pro 7 různých funkcí ohodnocení termu na TREC-7 kolekci (převzato z [172])

úvaha, že se nikdo nezabývá efektem jaký má jeden konkrétní term na vylepšení vyhledávání, všechny studie zajímá pouze celkový vliv množiny vybraných termů na vylepšení výsledků vyhledávání. V práci bylo ukázáno, že většina termů přidávaných do dotazu nezlepšuje vyhledávání (nemají vliv nebo zhoršují), přestože celkově dojde k vylepšení výsledků. Navrhované řešení je použít SVM[19] klasifikátor k rozlišení přínosných a nepřínosných termů. Problémem tohoto přístupu je nutnost natrénovat klasifikátor na nějakých anotovaných datech.

V práci [172] bylo testováno přidání různého množství termů pro sedm nejlepších metod váhových funkcí (experiment je popsán v podkapitole 6.2.3 - Výběr váhové funkce) na kolekcích TREC-6, TREC-7, TREC-8, kde dotazy byly vytvořeny pouze z názvu tématu. Množství přidaných termů se pohybovalo mezi 0 až 200 termy (0, 3, 25, 50, 75, 100, 125, 150, 175, 200). Experimenty pro všechny možnosti váhové funkce ukázaly, že výsledky při přidání 100 termů jsou statisticky významně lepší než při přidání menšího množství termů (0 - 75 termů), ale nejsou lepší než při přidání většího množství termů (125 - 200 termů). Hodnota 100 termů byla tedy doporučena jako nejlepší možná. Porovnání výsledků je vidět na obrázku 6.4. V dalším experimentu bylo porovnáváno, zda dojde k lepším výsledkům pokud pro každý dotaz použijeme jiné (optimální) množství termů oproti použití 100 termů. Experimenty ukázaly, že při použití optimálního množství termů dojde vždy k lepším výsledkům, zlepšení ale nebylo statisticky významné.

Pokud je ohodnocení termu pravděpodobnostní, je možné vybrat termy stanovením nějakého prahu (například $p = 0,001$), kdy termy s větší pravděpodobností budou použity k rozšíření dotazu. Tento přístup byl použit například v práci [185]. Problém výběru množství termů je zde ale převeden na problém nastavení prahu.

[19] Support vector machine



Obrázek 6.5: Vliv množství použitých dokumentů (pasáží) na výsledky vyhledávání (zde měřené 11 bodovou průměrnou přesností) (převzato z [176])

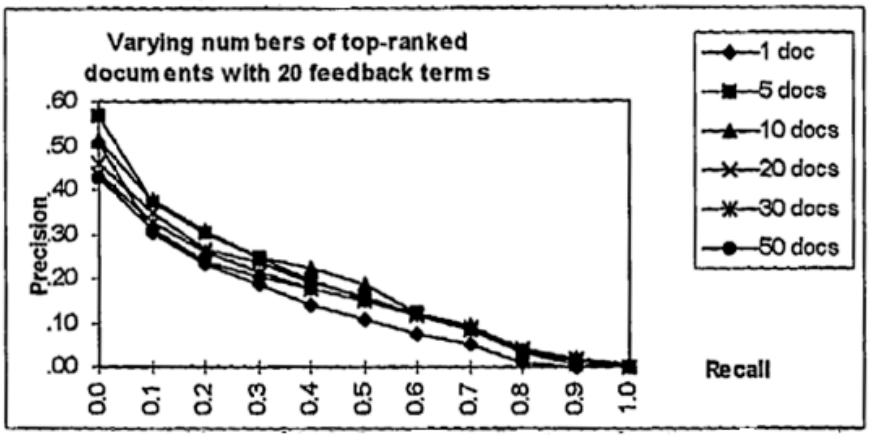
6.2.6 Počet pseudo relevantních dokumentů

Dalším z nevyřešených problému metod slepé zpětné vazby je množství pseudo relevantních dokumentů, které je vhodné použít. V teorii je většinou metod stanoveno použít „prvních k“ a tyto dokumenty považovat za relevantní. Použitý počet pseudo relevantních dokumentů u popisovaných metod se většinou pohybuje v rozmezí 20 - 30 (20 v práci [101], 30 v pracích [13, 84].

V práci [176] bylo testováno 0 - 500 dokumentů pro výběr konceptů (viz podkapitola 6.2.3), problém neexistence metody pro stanovení počtu dokumentů zde byl také zdůrazněn. Na základě experimentů bylo doporučeno, že nastavení k mezi 30 - 300 dokumenty produkuje dobré výsledky pro testované TREC kolekce. Vliv počtu použitých dokumentů je znázorněn na obrázku 6.5.

V práci [91] bylo experimentováno s použitím prvních 1, 5, 10, 20, 30 a 50 dokumentů ve vektorovém modelu. Závěrem experimentů bylo doporučení použít 5 - 20 prvních vyhledaných dokumentů jako pseudo relevantní dokumenty. Při použití menšího nebo většího množství dokumentů bylo dosaženo horších výsledků. Výsledky jsou zobrazeny na obrázku 6.6.

V práci [17] byla stanovena hypotéza, že při použití malého množství dokumentů by mělo být dosaženo většího zlepšení vyhledávání pomocí slepé zpětné vazby, díky předpokládané větší hustotě relevantních dokumentů mezi prvními vyhledanými dokumenty. Experimenty však ukázaly, že přínos zpětné vazby se zvětšuje s množstvím použitých pseudo relevantních dokumentů až do určitého množství, kdy začne pomalu klesat. Tento jev je vysvětlen tím, že v malém množství pseudo relevantních dokumentů nemusí být žádné opravdu relevantní dokumenty a i s malým množstvím relevantních dokumentů systém špatně odhadne termy pro rozšíření dotazu. Experimenty ukázaly, že nejlepších výsledků bylo dosaženo při použití 4 - 10 pseudo relevantních dokumentů (bylo testováno až do 100



Obrázek 6.6: Porovnání výsledků vyhledávání v závislosti na přidání 20 termů jako rozšíření dotazu slepou zpětnou vazbou z různého počtu dokumentů (převzato z [91])

dokumentů).

Podobné výsledky analýzy optimálního počtu pseudo relevantních dokumentů byly dosaženy v práci [180]. Experimenty byly provedeny s 5, 10, 15, 20, 30, 50 pseudo relevantních dokumentů, při zvyšování počtu dokumentů se výsledky zlepšovaly až do dosažení nějaké hranice, pak se začaly zhoršovat. Bylo zjištěno, že optimální počet dokumentů je pro každou metodu jiný, ale pro metody ohodnocení termu zohledňující pořadí pseudo relevantního dokumentu je následné zhoršení výsledků při použití více dokumentů pomalejší.

6.2.7 Použití metod slepé zpětné vazby ve vyhledávání v řeči

Velké množství výzkumných týmů se začalo zabývat vyhledáváním informací v řeči se vznikem TREC-6 a TREC-7 úloh [39, 157], kde se ukázalo, že metody vyhledávání informací jsou použitelné na automatických přepisech získaných z ASR systému. V pracích [70, 71] bylo testováno použití zpětné vazby z paralelního korpusu a slepé zpětné vazby na sedmi různých automatických přepisech stejných dat TREC-7 kolekce s WER od 25% do 61%, kde bylo zjištěno, že přestože výsledky vyhledávání jsou lepší pro přepisy s menším WER (pro nejlepší přepisy byly téměř stejné jako pro manuální transkripce), zlepšení přinášené oběma metodami zpětné vazby je konstantní. Jejich závěrem tedy bylo, že výzkum by se měl soustředit spíše na vylepšování ASR a vyhledávání informací samostatně, než na vývin nových metod pro vyhledávání informací v bohatších reprezentacích řečových dat.

Chen a kol. v práci [24] testovali různá akustická a lingvistická vylepšení ve vektorovém modelu pro vyhledávání informací na dvou různých kolekcích. Bylo testováno například použití CN (viz podkapitola 5.4.2), změna váhy termu podle slovního druhu, rozšíření dotazu podobnými slovy z kolekce a slepá zpětná vazba pomocí Rocchio algoritmu. Při použití každé metody samostatně bylo největšího vylepšení dosaženo pomocí slepé zpětné vazby (zlepšení o 10,3% na jedné kolekci, 8,2% na druhé). Celkově při použití akustických vylepšení dosáhl systém zlepšení o 3,5% a 2,8%, při použití lingvistických o 21,3% respektive 12,3%, je tedy vidět, že vylepšením metod vyhledávání informací je dosaženo většího zlepšení výsledků.

V práci [79] byla použita zpětná vazba v modelu BM25. Z prvních 5 vyhledaných dokumentů byl nejprve vytvořen souhrn o délce 6 vět a z něj bylo teprve vybráno nejlepších 5 nebo 20 termů. Ve výsledcích je porovnáno zlepšení dosažené pomocí slepé zpětné vazby na automatických přepisech a odpovídajících manuálních přepisech, tedy vyhledávání v textu. Bylo ukázáno, že pro 20 přidaných termů bylo dosaženo většího procentuálního zlepšení pro automatické přepisy než u vyhledávání v textu.

Práce [163] testuje využití slepé zpětné vazby v jazykovém modelování pro vyhledávání informací, experimenty jsou provedeny jak na nejlepších přepisech z ASR systému, tak na mřížkách. Rozdíl ve výsledcích vyhledávání informací bez použití slepé zpětné vazby mezi nejlepšími přepisy a mřížkou je méně než jedno procento absolutně, v případě přidání slepé zpětné vazby se tento rozdíl zmenšuje na 0,3%. Práce dále testuje možnost rozšíření dotazu pomocí latentních témat získaných pomocí PLSA[20] metody, kde došlo u všech přístupů ke zlepšení se stejným trendem, tedy u nejlepších přepisů bylo zlepšení větší než u mřížek. V práci je testováno množství použitých pseudo relevantních dokumentů od 0 do 100, nejlepších výsledků bylo dosaženo při použití 50 - 70 dokumentů.

V práci [25] byla testována slepá zpětná vazba pomocí modelu relevance a SSM modelu v KL divergence jazykovém modelování pro vyhledávání informací. Experimenty byly zaměřeny na možnosti získání co nejvíce informací z množiny pseudo relevantních dokumentů pro vytvoření lepšího modelu dotazu. Dokumenty pro množinu pseudo relevantních dokumentů jsou vybrány pomocí jejich ohodnocení mírou zahrnující jak jejich podobnost s dotazem získanou z prvního průběhu vyhledávání, tak i jejich vzdálenost od modelu nerelevantních dokumentů (vytvořeného z dokumentů s malým skóre z prvního běhu vyhledávání). Další součástí ohodnocení dokumentu jsou také míry hustoty a rozdílnosti (viz práce [177] popsaná v podkapitole 6.1.7). V práci bylo použito nastavení vybrat prvních 25 dokumentů a z nich poté vybrat 5 dokumentů pomocí jejich nového ohodnocení.

Lee a Lee v práci [83] testují použití slepé zpětné vazby v jazykovém modelování pomocí KL divergence, místo použití celého dokumentu pro výběr termů, navrhují použití jen jeho části - promluvy, vzhledem k možné změně tématu v průběhu řečového dokumentu.

6.3 Shrnutí poznatků

Z popsaných experimentů v této kapitole lze shrnout závěry:

- **Vylepšení výsledků** - Odborné práce zabývající se výzkumem zpětné vazby se shodují v závěru, že použití zpětné vazby zlepšuje výsledky vyhledávání. Pokud se zaměříme na slepou zpětnou vazbu, tedy bez informace o relevanci dokumentů od uživatele systému, shrnutím citovaných prací dojdeme ke stejnému závěru, tedy že použití slepé zpětné vazby zlepšuje výsledky vyhledávání.

- **Metoda vyhledávání informací** - Z popsaných experimentů je zřejmé, že není možné stanovit jednu metodu pro vyhledávání informací v kombinaci se zpětnou vazbou, o které by se dalo říci, že funguje nejlépe. Publikované experimenty ukazují velmi podobné výsledky pro metody jazykového modelování a vektorového nebo BM25 modelu.

- **Váhová funkce** - Výběr váhové funkce zásadně neovlivní zlepšení výsledků vyhledávání, pokud bude vybrána funkce, která kombinuje informaci o počtu výskytů termů

[20] Probabilistic latent semantic analysis

v (pseudo) relevantních dokumentech a informaci o jeho výskytu v celé kolekci, případně o rozložení jeho výskytu v relevantních a nerelevantních dokumentech. Na základě popsaných experimentů provedených v citovaných pracích v této kapitole se dá říci, že tento předpoklad není závislý na použité kolekci dokumentů.

- **Počet termů** - Neexistuje žádná jednotně doporučená metoda pro určení množství termů vhodných pro rozšíření dotazu. Většina prací experimentuje se zlepšením procesu výběru termů, ale jejich výsledný použitý počet je určen experimentálně, ať již na základě přímo stanovení počtu, nebo stanovením nějakého prahu. Na rozdíl od předchozího bodu, vybraný počet termů se zdá být závislý na použité kolekci. Ve většině starších prací bylo použito menší množství termů pro rozšíření dotazu (20 - 30 termů), novější práce operují s větším množstvím termů (70 - 100). Rozdíl je zejména ve velikosti použitých kolekcí, zatímco u starších prací byly použity menší kolekce, s menším počtem relevantních dokumentů, novější práce používají pro experimenty větší TREC kolekce.

- **Počet dokumentů** - Stejně jako u předchozího bodu se na základě výzkumů publikovaných v odborné literatuře dá říci, že neexistuje žádná jednotná metoda pro určení počtu dokumentů, které je vhodné použít jako pseudo relevantní dokumenty (u klasické zpětné vazby tyto dokumenty vybírá přímo uživatel). Většina prací se drží zásady „použij prvních k dokumentů“, kde k bývá určeno experimentálně u prací, které se zabývají vlivem tohoto nastavení na velikost zlepšení výsledků vyhledávání. U ostatních prací, zabývajících se pouze použitím slepé zpětné vazby pro vylepšení výsledků vyhledávání, často k diskuzi tohoto nastavení ani nedochází a je použito nějaké nastavení získané na základě jiné práce, či předchozích zkušeností. Navržení metody pro výběr počtu dokumentů a její experimentální ověření, je jedním z cílů této práce (viz podkapitola 7.7).

- **Vazba mezi počtem termů a počtem dokumentů** - V citovaných pracích nebyla zkoumána vazba mezi doporučeným počtem termů a dokumentů. Experimenty byly prováděny zvlášť, buď byl pevně nastaven počet dokumentů, nebo počet termů, maximálně byla ověřena stálost dosažených výsledků při malých změnách jednoho z parametrů. Experimenty se závislostí těchto dvou nastavení jsou provedeny v této práci a popsány v podkapitole 7.6.

Kapitola 7

Experimenty a navržené metody

Všechny experimenty, uvedené v této kapitole, byly provedeny na české kolekci pro vyhledávání informací v řeči, vytvořené v rámci úlohy CL-SR kampaně CLEF. V podkapitole 7.1 bude popsána tato kolekce a způsob vyhodnocení výsledků na ní získaných (podkapitola 7.2). Dále následují experimenty s různým předzpracováním vstupních dat (podkapitola 7.4). Poté budou uvedeny experimenty s jednotlivými metodami pro vyhledávání informací (podkapitola 7.5). V další části práce budou popsány experimenty se slepou zpětnou vazbou (podkapitola 7.6) a nově navržené metody pro stanovení počtu pseudo relevantních dokumentů (podkapitola 7.7). Na závěr této kapitoly budou uvedeny experimenty s nově navrženými metodami v podobné úloze detekce témat textu (podkapitola 7.8).

7.1 CLEF CL-SR

Úloha CL-SR v rámci kampaně CLEF navázala v roce 2005 na výsledky úlohy CL-SDR probíhající v letech 2003 [35] a 2004 [34]. Zatímco CL-SDR úloha používala kolekci dokumentů z úlohy TREC-9, pro úlohu CL-SR v roce 2005 [170] byla vytvořena nová kolekce pro vyhledávání informací ve spontánní konverzační řeči [106], získaná z výpovědí svědků holocaustu z projektu MALACH. Tato kolekce sestávala z 625 hodin rozpoznané řeči a 28 témat pro vyhledávání v anglickém jazyce, témata byla také přeložena do jiných jazyků pro mezijazykové[1] vyhledávání. Výpovědi byly manuálně rozděleny na 8104 úseků zabývajících se stejným tématem a ke každému byl vytvořen přibližně tři věty dlouhý obsah a přiřazena klíčová slova. Kolekce byla v roce 2006 rozšířena o dalších 42 témat (6 témat bylo vyřazeno) na celkem 63 témat, 38 trénovacích a 25 testovacích.

7.1.1 Česká kolekce spontánní řeči

K anglické části úlohy CL-SR přibyla v roce 2006 [107] a 2007 [111] také česká část. Oproti anglické části zde nebyly výpovědi rozděleny na úseky se stejným tématem, kolekce tedy sestávala z neděleného přepisu z ASR systému [62]. Cílem vyhledávání v této české kolekci tedy spíše než nalezení odpovídajícího dokumentu je nalezení odpovídajícího času počátku výpovědi o hledaném tématu.

[1] Cross-Language

Dokumenty

Pro vytvoření automatických přepisů byly použity dva různé ASR systémy vytvořené na Západočeské univerzitě v Plzni a Johns Hopkins University, z roku 2004 a 2006 [141, 142]. Systém z roku 2004 vytvářel přepisy v hovorové řeči, zatímco pro rok 2006 byl systém upraven pro vytváření přepisů ve spisovné češtině, což by mělo vést k lepším výsledkům, protože témata jsou také napsána spisovnou češtinou. Výsledkem získaným z ASR systému byl přepis 357 výpovědí v českém jazyce. Chybovost přepisů v této kolekci se pohybuje kolem 35% WER [61].

Pro snadnější navržení základního systému pro vyhledávání informací v této kolekci byly výpovědi rozděleny na překrývající se segmenty přibližně 3,75 minuty dlouhé, s 33% překryvem s předcházejícím a následujícím segmentem (původním cílem autorů kolekce bylo vytvořit segmenty dlouhé tři minuty s překryvem jedné minuty, kvůli chybě ve skriptu byly ale nakonec segmenty delší). Tím vzniklo 11 377 segmentů, které budeme považovat za jednotlivé dokumenty. Takto vytvořené dokumenty obsahují průměrně 400 slov.

K těmto dokumentům byla v roce 2006 také připojena anglická, a z nich přeložená česká, klíčová slova, vytvořená buď automatickým nebo manuálním přiřazením termů z thesauru. Manuálně přiřazená klíčová slova byla bohužel špatně načasována, neodpovídala tedy správným dokumentům. Automaticky přiřazená klíčová slova nepřinesla žádné vylepšení výsledků vyhledávání, proto byly oba druhy klíčových slov v roce 2007 z kolekce vynechány.

Témata

Kolekce obsahuje 118 témat k vyhledávání, 105 z nich jsou přeložená původní témata z anglické kolekce a 13 dalších témat jsou upravená - zobecněná původní témata (byla vynechána například specifická geografická omezení). Všechna témata byla vytvořena nejprve v angličtině a poté přeložena rodilými mluvčími do češtiny. V roce 2006 byl k 29 tématům vytvořen seznam relevantních úseků, obsahující celkem 1322 začátků relevantních pasáží, tedy průměrně 46 na jedno téma (s minimálním počtem 8 počátků). Tato témata byla v roce 2007 označena jako trénovací sada. V roce 2007 byl vytvořen seznam relevantních úseků k dalším 42 tématům, obsahující celkem 2389 počátků relevantních pasáží, tedy s průměrným počtem 56 pasáží na jedno téma (s minimálním počtem 6 pasáží). Tato témata byla nadále označena jako testovací sada.

Ukázka tématu je vidět na obrázku 7.1. Každé téma obsahuje pole `<title>`, `<desc>` a `<narr>`, tedy název, popis a širší popis tématu.

Ke každému tématu bylo určeno minimálně 6 relevantních úseků v kolekci. Relevantní úseky jsou definovány svým počátečním a konečným časem, tedy nejsou nijak závislé na rozdělení kolekce na segmenty.

7.2 Vyhodnocení výsledků

Pro ohodnocení výsledků vyhledávání byla použita míra mGAP [87] definovaná v podkapitole 3.8.2, založená na započtení přesnosti nalezení správného počátku relevantního úseku. Jako penalizační funkce byla v této úloze [107] zvolena symetrická lineární funkce zmenšující pro každých 9 vteřin rozdílu od stanoveného počátku relevantního úseku (na obě strany) ohodnocení o 0,1. Rozdíl větší než 90 vteřin je tedy brán jako žádná shoda. Pro každý relevantní segment je také do ohodnocení započítáván pouze první nalezený

```
<top>
<num>1166
<title>Chasidismus
<desc>Chasidové a jejich nezlomná víra
<narr>Relevantní materiál by měl vypovídat o Chasidismu
v období před holokaustem, v průběhu holokaustu a po
něm. Informace o chasidských dynastiích a založených a
zničených geografických lokalitách.
</top>
```

Obrázek 7.1: Ukázka tématu z české kolekce spontánní řeči vytvořené v rámci úlohy CLEF CL-SR

úsek, ostatní, i pokud by přesněji odpovídaly definovanému počátečnímu času relevantního úseku, nejsou započítány (obdrží nulové ohodnocení).

Princip ohodnocení tímto způsobem je založen na předpokladu, že pro uživatele systému pro vyhledávání v řeči je důležité zjistit počáteční čas relevantního úseku, který si pak bude moci poslechnout.

7.2.1 Testování hypotéz

V případě, že chceme porovnat výsledky dvou různých metod, není dostačující pouze informace o rozdílu skóre jednotlivých metod pro posouzení jejich efektu na zlepšení výsledků. Pro porovnání dvou systémů se často používá posouzení statistické významnosti výsledků experimentů. Při testování statistických hypotéz posuzujeme, zda výsledky experimentů odpovídají předpokladu, který jsme stanovili. Statistická hypotéza vyjadřuje určitý předpoklad o rozdělení náhodných veličin, při testování porovnáváme dvě hypotézy. První hypotézou je *nulová hypotéza* H_0, u této hypotézy předpokládáme její platnost, kterou budeme ověřovat. Druhou hypotézou je alternativní hypotéza H_1.

Stanovíme si hladinu významnosti α, obvykle se stanovuje hodnota $\alpha \leq 0,05$, a na základě výsledku testu stanovíme pravděpodobnost, že by testovací kritérium dosáhlo této výsledné hodnoty testu, pokud by H_0 byla platná. Tato pravděpodobnost se označuje jako *p-hodnota* testu. Pokud je tedy p-hodnota menší než stanovené α, pak můžeme hypotézu H_0 na této hladině významnosti zamítnout a přijmout alternativní hypotézu H_1.

K testování hypotéz v oblasti vyhledávání informací se často používá Wilcoxonův test[2]. Wilcoxonův test je neparametrický párový test, který nepředpokládá normalitu dat a je tedy použitelný i pro malé množství testovacích dat, u kterých není možné normalitu ověřit a použít párový t-test. V případě testování dvou metod stanovíme nulovou hypotézu H_0, že rozdíly mezi párovanými výsledky obou metod mají symetrické rozdělení okolo 0 a alternativní hypotézu H_1, že rozdíl mezi párovanými výsledky toto nesplňuje. Postup testu je následující:

1. Spočteme rozdíly mezi párovanými hodnotami jednotlivých výsledků metod a vyřadíme ty výsledky, kde je rozdíl nulový.

2. Nenulové rozdíly uspořádáme vzestupně bez ohledu na znaménko a přiřadíme jim

[2]Wilcoxon Signed-Rank Test

pořadí. Stejným rozdílům je přiřazeno průměrné pořadí.

3. Spočteme statistiku W, která je rovna menšímu ze součtů pořadí pro každé znaménko zvlášť (tedy součet pozitivních pořadí a součet negativních pořadí).

Pro malý počet nenulových rozdílů porovnáme hodnotu kriteria W s tabulkami pro příslušný počet rozdílů a zvolenou hladinu významnosti[3], pro větší počet rozdílů lze rozdělení statistiky aproximovat normálním rozdělením. Alternativně může být spočtena p-hodnota tohoto testu a test vyhodnocen podle ní.

Ukázka výpočtu Wilcoxonova testu pro ověření vlivu odstranění stop slov na výsledky vyhledávání (podkapitola 7.4.1) je uvedena v příloze A, u dalších testů budou uvedeny pouze výsledky testu. V tabulce A.1 je vidět statisticky nevýznamný rozdíl výsledků metod a v tabulce A.2 statisticky významný.

7.3 Nastavení experimentů

Pro experimenty uvedené v následující části práce bude použita kolekce popsaná v podkapitole 7.1.1. Oproti původní nastavené délce pasáží na 3 minuty, bude použito dělení nahrávek navržené v roce 2007 v práci [63]. V práci bylo navrženo zkrácení segmentů na 2 minuty s 1 minutovým překryvem (kvůli chybě skriptu byly, stejně jako u původního 3 minutového dělení, pasáže nakonec přibližně 2,5 minuty dlouhé, s 50% překryvem). Toto zkrácení pasáží vedlo ke zvýšení přesnosti výsledků vyhledávání [63]. Přestože se nakonec ukázalo toto vylepšení jako statisticky nevýznamné, takto rozdělené kratší úseky se svou délkou více přibližují délce relevantních pasáží, která je průměrně 2,83 minut [111]. Toto nové dělení automatických přepisů výpovědí vedlo ke zvýšení počtu dokumentů v kolekci z původních 11 377 na 22 581.

Pro vytvoření dotazů budou použity obě sady témat obsažené v kolekci, budou dále označovány jako trénovací a testovací sada. Nicméně je důležité si uvědomit, že metody pro vyhledávání informací se nijak „netrénují“ (například oproti klasifikátorům), nepoužívají žádná trénovací data v klasicky chápaném smyslu. Účel trénovací sady témat by se tedy dal chápat spíše jako development data pro nastavení parametrů jednotlivých metod. V případě, že nejsou testována různá nastavení metod a není vybíráno nejlepší nastavení, je možné použít pro experimenty pouze jednu sadu témat, jelikož se jedná o data vyhledávacím systémem nikdy neviděná a systém na ně není nijak nastaven.

Dotazy budou tvořeny z jednotlivých slov - termů témat, k experimentům budou použity různé kombinace polí `<title>`, `<desc>` a `<narr>`. Tyto kombinace budou označovány jako **t** - pole `<title>`, **td** - pole `<title>` a `<desc>`, **tdn** - pole `<title>`, `<desc>` a `<narr>`.

Všechny popsané a testované algoritmy byly implementovány v jazyce JAVA.

7.4 Vliv různého předzpracování vstupních dat

První část experimentů se bude věnovat metodám pro předzpracování vstupních dat - dokumentů a dotazů. Pro vylepšení výsledků vyhledávání je možné využít některé obecně používané metody, například vynechání stop slov či lemmatizaci. Na základě těchto experimentů budou pak takto upravená vstupní data použita v dalších testovaných metodách.

[3] například zde: http://www.stat.ufl.edu/winner/tables/wilcox_signrank.pdf

7.4.1 Vliv odstranění stop slov

Experimenty byly zaměřeny na testování vlivu odstranění stop slov, tedy slov, která nenesou sama o sobě žádný význam. Byly testovány tři varianty zpracování dotazů a dokumentů:

1. Všechna slova
2. Odstranění všech předložek, spojek, částic a citoslovcí (stop slova)
3. Odstranění navíc k předchozímu bodu ještě všech zájmen (stop slova + zájmena)

Pro uvedené experimenty byl použit český morfologický analyzátor dostupný v rámci Pražského závislostního korpusu 2.0 (PDT 2.0)[4] [45].

Vliv odstranění stop slov byl otestován na booleovském modelu P-Norm (p = 5) z podkapitoly 4.2.1, na vektorovém modelu TF-IDF z podkapitoly 4.3 a Query likelihood modelu s Jelínek-Mercer vyhlazováním QL ($\lambda = 0.1$) z podkapitoly 4.5.1 na lemmatizovaných datech, výsledky jsou ukázány v tabulce 7.1.

Tabulka 7.1: Vliv odstranění stop slov na mGAP v modelu P-Norm, vektorovém modelu TF-IDF a Query likelihood modelu QL

témata	**trénovací**	
P-Norm	**td**	**tdn**
Všechna slova	0,0316	0,0251
Odstranění stop slov	0,0299	0,0252
Odstranění stop slov + zájmen	0,0258	0,0257
TF-IDF	**td**	**tdn**
Všechna slova	0,0362	0,0450
Odstranění stop slov	0,0355	0,0456
Odstranění stop slov + zájmen	0,0329	0,0426
QL	**td**	**tdn**
Všechna slova	0,0312	0,0383
Odstranění stop slov	0,0306	0,0392
Odstranění stop slov + zájmen	0,0289	0,0388

Jak je vidět z tabulky 7.1, odstranění stop slov a zájmen v obou variantách nepomohlo vylepšit výsledky vyhledávání. Všechny testované metody ukazují stejný trend horších výsledků vyhledávání při odstranění více slov (slovních druhů) při použití kratších **td** dotazů.

Z těchto výsledků můžeme usoudit, že přestože tato slova nenesou sama o sobě žádný význam, v kombinaci s ostatními slovy přinášejí do vyhledávacího systému další informaci. Je však vidět, že při použití delších **tdn** dotazů je již informace získaná z ponechání všech slov nepotřebná, rozšíření dotazu o popisné pole `<narr>` přináší samo o sobě dostatek informací. Výsledky jsou téměř stejné, dokonce mírně lepší při odstranění předložek, spojek, částic a citoslovcí (stop slova). Odstranění těchto stop slov tedy můžeme použít u delších

[4]The Prague Dependency Treebank 2.0 - http://ufal.mff.cuni.cz/pdt2.0/index-cz.html

```
<top>
<num>1166
<title>Chasidismus
<desc>Chasid a jeho nezlomný víra
<narr>Relevantní materiál být mít vypovídat o chasidizmus
v období před holokaust, v průběh holokaust a po on.
Informace o chasidský dynastie a založený a zničený geografický
lokalitam.
</top>
```

Obrázek 7.2: Ukázka lemmatizovaného tématu z obrázku 7.1 z české kolekce spontánní řeči vytvořené v rámci úlohy CLEF CL-SR

tdn dotazů pro zredukování velikosti slovníku, tedy zmenšení paměťových a výpočetních nároků vyhledávání.

Výsledky metod byly testovány na statistickou významnost jejich rozdílu, u každé metody vyhledávání informací byl testován rozdíl mezi použitím všech slov v tématu a odstraněním stop slov a zájmen při použití polí tématu **td** (u všech metod je mezi těmito variantami největší rozdíl). Byly stanoveny alternativní hypotézy, že výsledek při odstranění stop slov a zájmen je horší než při použití všech slov. Rozdíly se nepotvrdily jako statisticky významné u metod P-norm (viz příloha A tabulka A.1) a QL, u metody TF-IDF test ukázal statisticky významné zhoršení výsledků vyhledávání (viz příloha A tabulka A.2). Pro vektorový model TF-IDF byl tedy ještě proveden test rozdílu mezi odstraněním stop slov a ponecháním všech slov, zde se již rozdíl neukázal statisticky významný.

V dalších experimentech budou použity dokumenty a dotazy s odstraněním stop slov, tedy všech předložek, spojek, částic a citoslovcí, jako kompromis mezi odstraněním i zájmen a použitím všech slov. Dojde tak k časové i paměťové úspoře při běhu algoritmů. Zároveň se ukázalo, že zhoršení výsledků u dotazů **td** není statisticky významné a u dotazů **tdn** k žádnému zhoršení nedochází.

7.4.2 Lemmatizace

Lemmatizace je proces, při kterém nahradíme různé tvary jednoho slova jeho lemmatem, tedy základním tvarem. Tento tvar bývá také nazýván slovníkový tvar, například pro slovesa je to infinitiv, pro podstatná jména první pád jednotného čísla. Nastavení toho co budeme považovat za lemma určitého slova se může lišit. Ukázka lemmatizovaného tématu z obrázku 7.1 je vidět na obrázku 7.2.

Vliv lemmatizace v úloze CLEF CL-SR 2006 byl ukázán v článku [60], v dalším roce se v úloze CL-SR potvrdila důležitost lingvistického předzpracování textů dokumentů a dotazů u všech zúčastněných systémů [111].

Pro všechny experimenty uvedené v této práci (pokud u nich není uvedeno jinak) byl pro lemmatizaci použit český morfologický analyzátor dostupný v rámci Pražského závislostního korpusu 2.0 (PDT 2.0)[5] [45]. Porovnání výsledků pro lemmatizovaná a nelemmatizovaná data na booleovském modelu P-Norm ($p = 5$), vektorovém modelu TF-IDF a Query likelihood modelu QL ($\lambda = 0.1$) je vidět v tabulce 7.2. Byly otestovány obě sady témat a varianty dotazů **td** a **tdn**. Je vidět, že při použití lemmatizace se výsledky vyhledávání u většiny variant zlepší téměř dvakrát. Zlepšení při použití lemmatizovaných dat

[5]The Prague Dependency Treebank 2.0 - http://ufal.mff.cuni.cz/pdt2.0/index-cz.html

oproti původním tvarům slov se ukázalo jako statisticky významné na hladině významnosti $\alpha \leq 0,05$ u všech metod a všech variant dotazů, kromě testovací sady **tdn** dotazů v modelu P-norm, kde je i vidět, že výsledky jsou obecně špatné.

Tabulka 7.2: Vliv lemmatizace v modelu P-Norm, vektorovém modelu TF-IDF a Query likelihood modelu QL (mGAP)

témata	**trénovací**		**testovací**	
P-Norm	**td**	**tdn**	**td**	**tdn**
lemmatizovaná	0,0299	0,0252	0,0135	0,0093
bez lemmatizace	0,0114	0,0125	0,0069	0,0072
TF-IDF	**td**	**tdn**	**td**	**tdn**
lemmatizovaná	0,0355	0,0456	0,0195	0,0224
bez lemmatizace	0,0198	0,021	0,0131	0,0148
QL	**td**	**tdn**	**td**	**tdn**
lemmatizovaná	0,0306	0,0392	0,0200	0,0255
bez lemmatizace	0,0176	0,0186	0,0114	0,0125

Na základě výsledků experimentů provedených na takto rozdílných přístupech k vyhledávání informací se dá usuzovat, že lemmatizace obecně zlepšuje výsledky vyhledávání na této kolekci, nehledě na použitý přístup k vyhledávání informací a měla by být součástí základního předzpracování dat. Všechny další experimenty budou provedeny na lemmatizovaných datech.

7.4.3 Automaticky vytvářený lemmatizátor

V práci [74] jsme otestovali použití automaticky vytvořeného lemmatizátoru pro úlohu vyhledávání informací. Naším cílem bylo zjistit, jak se budou lišit výsledky vyhledávání při použití automaticky a manuálně vytvářeného lemmatizátoru. Automatický lemmatizátor je trénován na datech z programu na kontrolu pravopisu Ispell, která jsou volně stažitelná z internetu[6] pro velké množství jazyků (detaily k vytváření lemmatizátorů lze nalézt v práci [74]). V tabulce 7.3 jsou vidět dosažené výsledky při použití takto lemmatizovaných dat na booleovském modelu P-Norm (p = 5), vektorovém modelu TF-IDF a Query likelihood modelu s Jelinek-Mercer vyhlazováním QL ($\lambda = 0.1$).

Rozdíly mezi oběma metodami lemmatizace se ukázaly jako statisticky nevýznamné, použití automaticky vytvářeného místo manuálně vytvářeného lemmatizátoru výsledky vyhledávání nezhorší. Díky možnosti univerzálního vytvoření lemmatizátoru pro různé jazyky je jeho použití vhodné například při nedostupnosti jiného lemmatizátoru v daném jazyce a výsledky budou srovnatelné s manuálně lingvisticky konstruovaným lemmatizátorem.

7.5 Metody pro vyhledávání informací

Všechny experimenty uvedené v této podkapitole byly provedeny s tématy a dokumenty, z jejichž textu byla odstraněna stop slova (viz experiment v podkapitole 7.4.1).

[6] https://lasr.cs.ucla.edu/geoff/ispell-dictionaries.html

Tabulka 7.3: Porovnání vlivu způsobu vytváření lemmatizátoru na vyhledávání v modelu P-Norm, vektorovém modelu TF-IDF a Query likelihood modelu QL (mGAP)

témata	**trénovací**		**testovací**	
P-Norm	**td**	**tdn**	**td**	**tdn**
manuálně vytvořený lemm.	0,0299	0,0252	0,0135	0,0093
automaticky vytvořený lemm.	0,0286	0,0284	0,0126	0,0072
TF-IDF	**td**	**tdn**	**td**	**tdn**
manuálně vytvořený lemm.	0,0355	0,0456	0,0195	0,0224
automaticky vytvořený lemm.	0,0381	0,0468	0,0185	0,0226
QL	**td**	**tdn**	**td**	**tdn**
manuálně vytvořený lemm.	0,0306	0,0392	0,0200	0,0255
automaticky vytvořený lemm.	0,0328	0,0415	0,0193	0,0227

Experimenty jsou provedeny na lemmatizovaných datech (viz experimenty v předchozí podkapitole 7.4.2).

V této podkapitole jsou představeny experimenty s vyhledáváním informací pomocí vybraných metod definovaných v kapitole 4, jmenovitě byly vyzkoušeny metody: rozšířený booleovský model P-Norm (viz podkapitola 4.2.1), vektorový model TF-IDF (podkapitola 4.3.1) a použití jazykových modelů pro vyhledávání informací s různými metodami vyhlazování (podkapitola 4.5). Důvodem pro volbu těchto metod bylo vybrat reprezentativní metody s rozdílným přístupem k vyhledávání informací (booleovský přístup, vektorový a pravděpodobnostní - jazykové modelování).

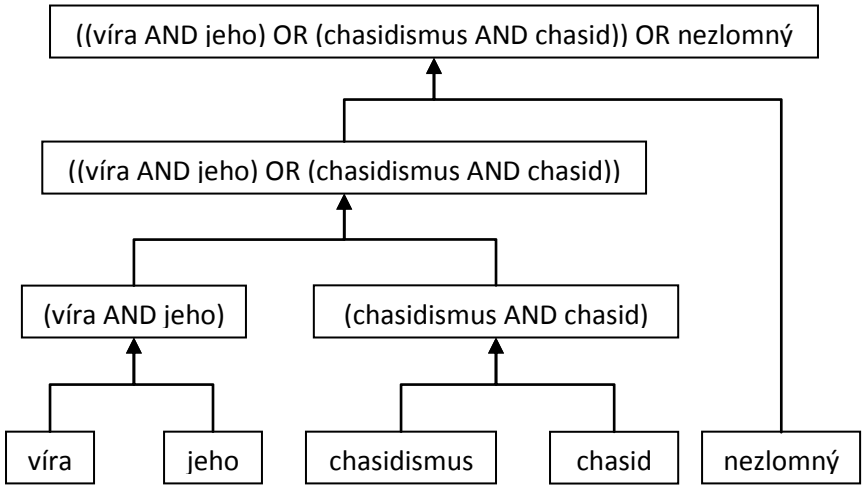
Vektorový model a jazykové modelování jsou v současnosti nejpoužívanější metody vyhledávání informací. Konkrétně vektorový model s BM25 váhovou funkcí [180] a jazykové modelování s Dirichlet vyhlazováním [92] se označují za nejlepší současné metody. Vektorový model s BM25 vyhlazováním testoval na této kolekci Ircing a kol. v práci [63], kde tento model dosáhl mnohem horších výsledků než TF-IDF. V práci bylo poukázáno na to, že model BM25 těží hlavně z normalizační části ohodnocení, která nemá žádný efekt v případě stejně dlouhých dokumentů jako jsou v této kolekci. Proto bude místo něj testován vektorový model TF-IDF. Booleovské modely se v současné době již moc nepoužívají (kromě knihovnických systémů), ale soudíme, že je to z důvodu nutnosti složitého vytváření dotazu. Proto byl do porovnání přidán i model P-norm, který překonává obecné nedostatky booleovských modelů.

7.5.1 Model P-Norm

Protože booleovský model předpokládá strukturované dotazy ve formě termů propojených operátory AND a OR, a v použité kolekci jsou dotazy ve formě témat, bylo nutné vyvinout metodu pro automatické vytváření booleovských dotazů z textu tématu. Byla navržena metoda nazvaná *Vzrůstání stromu*, založená na vytváření stromové struktury z textu tématu pomocí spojování dvojic termů dotazu pomocí AND a poté spojení těchto dvojic pomocí OR (viz obrázek 7.3). Tato metoda a provedené experimenty byly publikovány v článku [152].

Booleovský dotaz je tvořen tímto postupem:

1. Vezmeme všechny termy dotazu a seřadíme je sestupně podle jejich *idf*.



Obrázek 7.3: Vytváření strukturovaného dotazu z textu tématu

2. Vytvoříme dvojice termů spojením dvou za sebou následujících termů pomocí operátoru AND. Výsledkem je seznam dvojic termů (t_i *AND* t_j). Spočteme *idf* váhu těchto párů jako průměr z *idf* vah spojených termů: $(t_i + t_j)/2$.

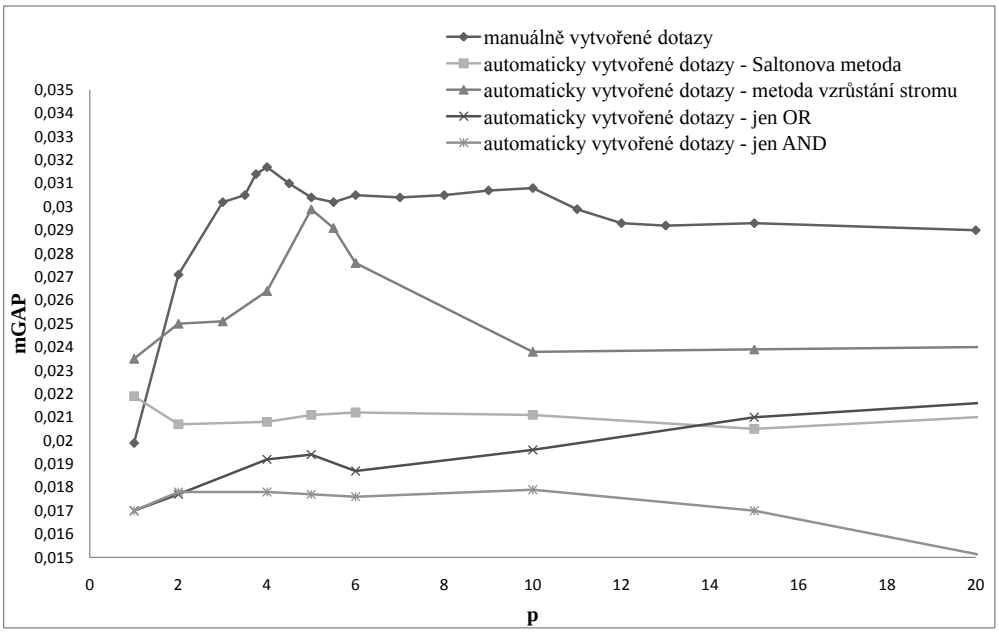
3. Znovu seřadíme seznam a vytvoříme páry z párů termů pomocí operátoru OR. Vypočteme jejich *idf* váhu, stejně jako v předchozím kroku.

4. Pokračujeme v tomto procesu iterativně, dokud nezbude poslední pár.

Pro porovnání výsledků této automatické metody byly také vytvořeny booleovské dotazy ručně a byla implementována metoda pro automatické vytváření booleovských dotazů popsaná v práci [131] (Saltonova metoda). Další metody zabývající se tímto problémem se nepodařilo nalézt. Také byly implementovány jednoduché metody spočívající pouze ve spojení termů operátorem AND nebo OR. Graf na obrázku 7.4 ukazuje porovnání těchto metod pro různé hodnoty parametru p z modelu P-Norm pro sadu trénovacích témat. Hodnoty výsledků jednotlivých metod jsou uvedeny v příloze B v tabulce B.1.

Výsledky, popsané také v práci [152], ukazují, že metoda *Vzrůstání stromu* se svým mGAP skóre velmi blíží hodnotám dosaženým při použití manuálně vytvořených dotazů. Pro množinu testovacích témat byly výsledky navržené metody lepší než výsledky při použití manuálních dotazů, výsledky ukazuje tabulka B.2 v příloze B. Pro $p = 4$, zvolené na základě nejlepších výsledků manuálně tvořených dotazů na trénovací sadě témat, byl proveden test statistické významnosti výsledků na testovací sadě témat. Lepší výsledek automaticky vytvářených dotazů se neukázal jako statisticky významný, lze tedy říci, že při použití metody *Vzrůstání stromu* pro automatickou tvorbu booleovských dotazů, lze dosáhnout stejných výsledků jako při použití manuálně vytvářených dotazů. Na druhou stranu, výsledek navržené metody *Vzrůstání stromu* se ukázal jako statisticky významně lepší než Saltonova metoda i než použití pouze operátoru OR na hladině významnosti $\alpha \leq 0,01$.

V tabulce 7.4 jsou vidět hodnoty mGAP pro vyhledávání s použitím modelu P-Norm s automaticky vytvořenými dotazy z lemmatizovaných dat pro trénovací a testovací sady dotazů pro různé hodnoty p. Je vidět, že volba p je závislá na délce dotazů, kdy pro delší



Obrázek 7.4: Porovnání metod pro automatické vytváření booleovského dotazu

dotazy **tdn** je vhodné volit spíše menší hodnotu (nastavení blíže vektorovému modelu) a pro kratší dotazy **td** je vhodná hodnota $p \in \{4, 5, 6\}$. Ukazuje se tedy, že použití strukturovaných dotazů má smysl pouze pro kratší dotazy, kdežto u delších dotazů je lepší použít spíše vektorový model.

Tabulka 7.4: Výsledky použití modelu P-Norm pro různé hodnoty p (mGAP)

témata	**trénovací**		**testovací**	
p	**td**	**tdn**	**td**	**tdn**
1	0,0235	0,0286	0,0133	0,0124
2	0,0250	0,0275	0,0131	0,0092
3	0,0253	0,0256	0,0127	0,0089
4	0,0277	0,0259	0,0133	0,0093
5	0,0304	0,0254	0,0132	0,0094
6	0,0296	0,0255	0,0134	0,0092
10	0,0255	0,0257	0,0131	0,0095
15	0,0250	0,0241	0,0124	0,0095

7.5.2 Vektorový model

Pro váhu termů byly vyzkoušeny různé varianty klasické TF-IDF váhy z rovnice (4.8), použití pouze TF nebo IDF složky jak pro váhu termu v dokumentu, tak i v dotazu a normalizovaná váha termu. V tabulce 7.5 jsou vidět výsledky těchto experimentů, nejlepších výsledků dosáhla kombinace TF-IDF vah pro termy v dotazu i dokumentech. Normalizo-

vaná váha termu nedosáhla lepších výsledků, stejně jako u BM25 modelu v práci [63] to přičítáme přibližně stejné délce dokumentů v této kolekci.

Tabulka 7.5: Otestování jednotlivých variant TF-IDF váhy ve vektorovém modelu na trénovací množině témat (mGAP)

témata		trénovací	
w_q	w_d	**td**	**tdn**
idf	$tf \cdot idf$	0,0355	0,0300
tf	$tf \cdot idf$	0,0280	0,0366
$tf \cdot idf$	idf	0,0262	0,0330
$tf \cdot idf$	tf	0,0280	0,0366
$tf \cdot idf$	$tf \cdot idf$	0,0355	0,0456
$tf \cdot idf$	$tf \cdot idf$ norm.	0,0333	0,0417

V tabulce 7.6 jsou vidět hodnoty mGAP při použití vektorového modelu TF-IDF, pro lemmatizovaná i nelemmatizovaná slova dotazů.

Tabulka 7.6: Výsledky použití vektorového modelu TF-IDF (mGAP)

témata	trénovací		testovací	
	td	**tdn**	**td**	**tdn**
lemmatizovaná	0,0355	0,0456	0,0195	0,0224
bez lemmatizace	0,0198	0,0210	0,0131	0,0148

7.5.3 Jazykové modelování

Použití jazykových modelů pro vyhledávání informací pro tuto kolekci bylo otestováno na metodách: Query likelihood model s metodami vyhlazování pomocí Jelinek-Mercer vyhlazování, Dirichlet vyhlazování a dvoufázového vyhlazování, dále přímé porovnání jazykových modelů pomocí Kullback-Leibler divergence s Jelinek-Mercer vyhlazováním. Také byla otestována možnost použití bigramových modelů v metodě Query likelihood model s Jelinek-Mercer vyhlazováním.

V následujících tabulkách jsou uvedeny dosažené výsledky pro lemmatizované dotazy, první tři tabulky (7.7, 7.8, 7.9) porovnávají různé způsoby vyhlazování jazykového modelu pro Query likelihood model. Samotný Query likelihood model bez vyhlazování nemá smysl testovat, vzhledem k tomu, že při nepřítomnosti i jen jediného termu z dotazu v hodnoceném dokumentu bude jeho skóre nulové. V tabulce 7.7 jsou uvedeny výsledky pro Jelinek-Mercer vyhlazování, v tabulce 7.8 pro Dirichlet vyhlazování a v tabulce 7.9 pro dvoufázové vyhlazování.

Z tabulek 7.7, 7.8, 7.9 je vidět, že nejlepších výsledků lze dosáhnout při použití dvoufázového vyhlazování jazykového modelu. Hodnoty parametrů λ a α byly určovány experimentálně.

Další experimenty s použitím jazykového modelování při vyhledávání informací se týkaly přímého porovnání jazykových modelů pomocí Kullback-Leibler divergence s Jelinek-

Tabulka 7.7: Výsledky použití jazykového modelování metodou Query likelihood model s Jelinek-Mercer vyhlazováním (mGAP)

témata	**trénovací**		**testovací**	
λ	**td**	**tdn**	**td**	**tdn**
0,1	0,0306	0,0392	0,0200	0,0255
0,25	0,0290	0,0376	0,0212	0,0257
0,5	0,0270	0,0343	0,0220	0,0274
0,75	0,0261	0,0317	0,0222	0,0271

Tabulka 7.8: Výsledky použití jazykového modelování metodou Query likelihood model s Dirichlet vyhlazováním (mGAP)

témata	**trénovací**		**testovací**	
α	**td**	**tdn**	**td**	**tdn**
1	0,0156	0,0163	0,015	0,0119
100	0,0238	0,0256	0,0199	0,0171
500	0,0283	0,0354	0,0212	0,0256
1000	0,0303	0,0383	0,0225	0,0260
2000	0,0331	0,0392	0,021	0,0267
5000	0,0333	0,0408	0,0186	0,0246
10000	0,0337	0,0413	0,0171	0,0214
20000	0,0337	0,0401	0,0146	0,0189

Tabulka 7.9: Výsledky použití jazykového modelování metodou Query likelihood model s dvoufázovým vyhlazováním (mGAP)

témata	**trénovací td**			**trénovací tdn**		
α/λ	**0,75**	**0,9**	**0,99**	**0,75**	**0,9**	**0,99**
1000	0,0267	0,0297	0,0314	0,0392	0,0398	0,0406
5000	0,0254	0,0309	0,0351	0,0388	0,0440	0,0445
10000	0,0252	0,0306	0,0358	0,0373	0,0431	0,0443
témata	**testovací td**			**testovací tdn**		
α/λ	**0,75**	**0,9**	**0,99**	**0,75**	**0,9**	**0,99**
1000	0,0208	0,0224	0,0232	0,0307	0,0304	0,0277
5000	0,0184	0,0202	0,0210	0,0278	0,0295	0,0295
10000	0,0185	0,0192	0,0192	0,0266	0,0273	0,0258

Mercer vyhlazováním. Jak je vidět z tabulky 7.10, při použití modelu dotazu jak byl uveden v podkapitole 4.5.3, jsou výsledky stejné jako u Query likelihood modelu (viz tabulka 7.7). Rozdílných výsledků by bylo dosaženo při použití jiného odhadu modelu dotazu než pomocí $\hat{P}_{MLE}(t|\theta_q)$, jak je uvedeno v kapitole 4.5.3.

Tabulka 7.10: Výsledky použití jazykového modelování metodou Kullback-Leibler divergence s Jelinek-Mercer vyhlazováním (mGAP)

témata	**trénovací**		**testovací**	
λ	**td**	**tdn**	**td**	**tdn**
0,1	0,0305	0,0392	0,020	0,0255
0,25	0,0289	0,0376	0,0212	0,0257
0,5	0,0269	0,0343	0,0220	0,0274
0,75	0,0260	0,0317	0,0222	0,0271

Poslední z provedených experimentů s různými variantami použití jazykového modelování ve vyhledávání informací se týkal zapojení bigramového jazykového modelu dokumentu do výpočtu $\hat{P}(t|d_j)$ v Query likelihood modelu.

Tabulka 7.11: Výsledky použití jazykového modelování metodou Query likelihood model s Jelinek-Mercer vyhlazováním s lineární interpolací přes bigramový jazykový model dokumentu (mGAP)

témata	**trénovací td**				
λ/λ_{bigram}	**0,01**	**0,05**	**0,1**	**0,2**	**0,25**
0,1	0,0283	0,0258	0,0244	0,0233	0,0228
0,25	0,0286	0,0267	0,0256	0,0246	0,0241
0,5	0,0265	0,0251	0,0244	0,0235	0,0232
0,75	0,0254	0,0242	0,0235	0,0222	0,0002
témata	**testovací td**				
λ/λ_{bigram}	**0,01**	**0,05**	**0,1**	**0,2**	**0,25**
0,1	0,0205	0,0192	0,0189	0,0177	0,0176
0,25	0,0212	0,0210	0,0206	0,0204	0,0204
0,5	0,0213	0,0212	0,0210	0,0206	0,0203
0,75	0,0213	0,0207	0,0205	0,0195	0,0003

Odhad pravděpodobnosti termu $\hat{P}(t|d_j)$ spočteme jako lineární interpolaci bigramového a unigramového modelu dokumentu s modelem celé kolekce:

$$\hat{P}(t_i|d_j) = \lambda_{bigram} P((t_i|t_{i-1})|\theta_{bigram\_d_j}) + \lambda P(t_i|\theta_{d_j}) + (1 - \lambda - \lambda_{bigram}) P(t|\theta_C). \quad (7.1)$$

V tabulce 7.11 jsou vidět výsledky experimentů pro trénovací a testovací lemmatizovaná témata. Při porovnání s použitím pouze unigramového modelu (viz tabulka 7.7) je vidět, že bigramový model výsledky vyhledávání nezlepšuje, naopak výsledky jsou horší. Je také vidět, že nejlepší výsledky bigramového modelu jsou dosaženy při co nejmenším započtení jeho skóre (nejmenší λ_{bigram}). Pro dobré natrénování bigramového modelu jed-

notlivých dokumentů jsou dokumenty v této kolekci příliš krátké, nedochází pak k dobrému odhadu pravděpodobnosti jednotlivých bigramů $P((t_i|t_{i-1})|\theta_{bigram\_d_j})$.

7.5.4 Porovnání jednotlivých metod

V tabulce 7.12 jsou porovnány nejlepší výsledky pro jednotlivé testované metody, tedy booleovský model P-Norm s navrženým automatickým vytvářením dotazu a $p = 5$, vektorový model TF-IDF s $tf \cdot idf$ vánou term a Query likelihood (QL) model s dvoufázovým vyhlazováním s $\lambda = 0,99$ a $\alpha = 5000$. Jak je vidět, vektorový model a Query likelihood model dosahují přibližně stejných výsledků, lepších než booleovský model P-Norm.

Tabulka 7.12: Porovnání nejlepších výsledků jednotlivých metod (mGAP)

témata	**trénovací**		**testovací**	
metoda	**td**	**tdn**	**td**	**tdn**
booleovský model P-Norm	0,0304	0,0254	0,0132	0,0094
vektorový model TF-IDF	0,0355	0,0456	0,0195	0,0224
QL model s dvoufázovým vyhl.	0,0351	0,0445	0,0210	0,0295

7.5.5 Shrnutí dosažených výsledků jednotlivých metod

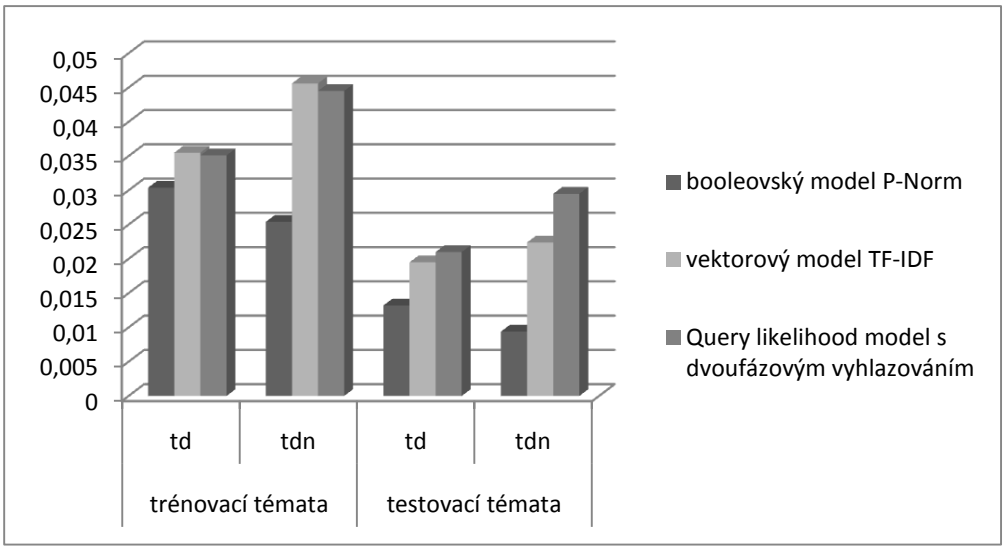
Z uvedených výsledků je vidět (viz tabulka 7.12 a obrázek 7.5), že pro další experimenty bude vhodné používat vektorový model TF-IDF nebo jazykové modelování metodou Query likelihood model s dvoufázovým vyhlazováním, booleovský model P-norm nedosahuje tak dobrých výsledků. Zvláště při použití delších dotazů **tdn**, pro které dosahují ostatní metody lepších výsledků, P-norm model selhává.

Použití vektorového modelu oproti všem metodám jazykového modelování má výhodu v tom, že není nutné nastavovat žádný parametr. Jak je vidět z výsledkových tabulek, metody jazykového modelování jsou na nastavení svých parametrů hodně závislé. Na druhou stranu metody jazykového modelování dosahují lepších výsledků na testovací kolekci (s nastavením parametrů z trénovací kolekce) než vektorový model a z tabulek je vidět, že pro jiné nastavení by mohly dosahovat výsledků ještě lepších.

Texty témat a dokumentů je vhodné před vyhledáváním lemmatizovat a používat témata v plném rozsahu, tedy všechna pole `<title>`, `<desc>` a `<narr>`. Odstranění stop slov před vyhledáváním je podle provedených experimentů sporné: nedojde k vylepšení vyhledávání, ale zároveň se zmenší velikost slovníku a díky tomu i paměťová náročnost, a tím dojde ke zvýšení rychlosti vyhledávání.

7.6 Experimenty se slepou zpětnou vazbou

V této podkapitole budou představeny experimenty týkající se využití slepé zpětné vazby pro vyhledávání informací. Budou prezentovány výsledky použití slepé zpětné vazby v jednotlivých metodách a vliv nastavení jejích parametrů. První experimenty byly provedeny i na modelu P-norm, další experimenty byly provedeny jen na vektorovém modelu a metodách jazykového modelování. Vyhledávání pomocí modelu P-norm je oproti ostatním modelům pomalé a paměťově náročné, kvůli nutnosti vytváření a poté vyhodnocení



Obrázek 7.5: Graf porovnání nejlepších výsledků testovaných metod vyhledávání

strukturovaného dotazu. Provedení rozsáhlejšího šetření by tak bylo velmi zdlouhavé a dosažené výsledky modelu k tomu nevybízejí.

Slepá zpětná vazba bude u všech metod realizována výběrem prvních k dokumentů a z nich výběrem m nejlepších termů ve smyslu zvolené váhy w_t. Pro ohodnocení termů ve zpětné vazbě se nejčastěji používá ohodnocení termů použité už ve vyhledávacím systému [18], stejně tak bude postupováno v následujících experimentech, pokud nebude uvedeno jinak.

7.6.1 Model P-norm

Na booleovském modelu P-Norm byl otestován vliv slepé zpětné vazby s různými variantami výpočtu váhy termu (tf, $tf \cdot idf$, idf) pro jejich výběr k rozšíření dotazu. Výsledky jsou vidět v tabulce 7.13, použití slepé zpětné vazby mírně zvýší dosažené hodnoty mGAP. Uvedené výsledky jsou pro trénovací sadu témat **td**.

Pokud vyhodnotíme výsledky z hlediska volby váhy termů, je vidět, že při použití pouze váhy idf dochází ke zhoršení výsledků ve všech variantách. Váha idf preferuje termy, které jsou co nejvíce jedinečné, tedy vyskytují se v co nejméně dokumentech. Tyto termy obecně nejsou dobré jako termy dotazu, dotaz příliš specializují a vedou k menší úplnosti vyhledávání. Vážení termu pouze pomocí tf vede k výběru termů, které se vyskytují nejčastěji v dokumentu, což jsou termy, které nemají žádný specifický význam. Přidání těchto termů nijak zásadně nezmění obsah a vyznění dotazu, což je vidět z výsledků experimentů, kdy je dosaženo téměř stejného mGAP ve všech variantách počtu dokumentů a přidaných termů. Nezáleží na tom, z kolika dokumentů jsou termy vybírány, přidaných 5 termů má stejný výsledek ať jsou vybrány z 20, 50, nebo 100 pseudo relevantních dokumentů, stejně tak i pro jiné počty přidaných termů. To může být způsobeno buď tím, že jsou vybrány stejné termy (tedy distribuce vybraného termu je stejná - největší ve všech dokumentech), nebo tím, že přestože jsou vybrány jiné termy, tak tyto termy stejným způsobem ovlivní výpočet podobnosti dotazu a dokumentu. Obě možnosti svědčí o tom, že vybrané termy nemají

Tabulka 7.13: Vliv použití zpětné vazby na mGAP v modelu P-Norm na trénovací množině témat

bez zpětné vazby	**0,0304**		
zpětná vazba	**použitá váha**		
dokumentů:termů	tf	$tf \cdot idf$	idf
20:5	0,0309	0,0265	0,0246
20:10	0,0307	0,0226	0,0217
50:5	0,0309	0,0298	0,0238
50:10	0,0309	0,0275	0,0220
100:5	0,0309	0,0305	0,0235
100:10	0,0308	0,0312	0,0220
100:20	0,0315	0,0313	0,0210
100:30	0,0313	0,0329	0,0185
500:20	0,0317	0,0246	0,0194

žádný význam, jde tedy o obecné termy. Důvod mírného zlepšení vyhledávání oproti variantě bez zpětné vazby není na první pohled jasně zřejmý, dochází ke stejnému efektu jako u odstranění stop slov, kdy přestože víme, že daná slova nepřináší žádný význam, ovlivní pozitivně výsledky vyhledávání.

Vhodnou variantou je použití váhy $tf \cdot idf$, tedy kombinace častého výskytu slov v dokumentech a zároveň jejich co největší jedinečnosti, kdy při vhodném nastavení dochází k výraznějšímu zlepšení výsledků vyhledávání (které se ale ani pro nejlepší variantu 100 dokumentů a 30 termů neukázalo jako statisticky významné, což se dá přičíst malému množství testovaných variant, ale také obecně problému modelu P-norm s příliš dlouhými dotazy).

Na modelu P-norm byly provedeny první experimenty se slepou zpětnou vazbou, nastavení parametrů metody bylo zvoleno na základě doporučení o „běžně“ používaném nastavení [63], v tomto případě použít prvních 20 dokumentů a z nich vzít nejlepších 5 termů. Tyto první experimenty vzbudily úvahy o důležitosti nastavení parametrů metody lépe, než podle běžně voleného doporučení. Jak je vidět z tabulky 7.13, výsledek vyhledávání je hodně závislý na nastavení počtu pseudo relevantních dokumentů a počtu vybíraných termů. Ve většině experimentů v citované literatuře také nedochází k testování vazby mezi počtem dokumentů a počtem termů, jeden parametr se vždy volí napevno a experimenty jsou provedeny pouze pro druhý z nich (viz podkapitola 6.3). Z výsledků je však zřejmá vazba mezi tímto nastavením, při výběru 10 termů z 20 dokumentů je výsledek horší než při výběru 5 termů z 20 dokumentů, na druhou stranu ale při výběru 10 termů ze 100 dokumentů je výsledek lepší než při výběru 5 termů ze 100 dokumentů.

7.6.2 Vektorový model

Vliv zpětné vazby ve vektorovém modelu na této kolekci je ukázán v práci [63], kde ale schází diskuze nad nastavením parametrů metody a je voleno typické nastavení programu Lemur[7], který byl pro experimenty použit.

Pro vektorový model byly provedeny rozsáhlé experimenty s možnostmi nastavení počtu pseudo relevantních dokumentů k a vybíraných termů m. Počet dokumentů byl volen

[7] http://www.lemurproject.org/

$k \in \{5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 150, 200, 300, 500\}$, počet termů byl volen od 5 do 40 termů, s rozestupem 5 termů. Jako váhová funkce termu w_t byla zvolena $tf \cdot idf$ váha definovaná rovnicí (6.11).

Výsledky experimentů jsou ukázány v tabulce B.3 v příloze B. Z tabulky je vidět, že výsledky všech kombinací (kromě dvou s použitím 500 dokumentů) jsou lepší než výsledky bez použití zpětné vazby. Nejlepší výsledky byly dosaženy při volbě 10-15 termů pokud bylo vybíráno z většího množství pseudo relevantních dokumentů (70-100), obecně při výběru termů z 90-100 dokumentů jsou výsledky nejlepší pro široký rozsah množství vybraných termů (5-25). Pro větší množství pseudo relevantních dokumentů (nad 100) se výsledky vyhledávání postupně zhoršují při jakémkoliv množství přidaných termů.

Na druhou stranu dobrých výsledků lze také dosáhnout při volbě více termů (25-30), pokud bylo vybíráno z menšího počtu dokumentů - kolem 30. Pokud je vybíráno z menšího množství dokumentů, pak jsou termy více vztažené k podobnému obsahu a jsou tedy obecně více relevantní a lze jich přidat větší množství. Na druhou stranu pokud vybíráme termy z většího množství dokumentů, dokumenty na nižších pozicích mohou být zaměřené trochu jiným směrem než původní dotaz, takže počet opravdu relevantních termů je nutné vybírat menší, aby nedošlo k posunu dotazu k jinému významu. Nejlepší výsledky slepé zpětné vazby jsou ukázány v tabulce 7.14.

Tabulka 7.14: Porovnání výsledků slepé zpětné vazby ve vektorovém modelu (mGAP)

témata	**trénovací**	**testovací**
metoda	**tdn**	**tdn**
vektorový model TF-IDF	0,0456	0,0224
TF-IDF, k=100, m=10	0,0563	0,0250
TF-IDF, k=70, m=10	0,0564	0,0267
TF-IDF, k=30, m=30	0,0550	0,0245

Všechny varianty na trénovací množině dat se potvrdily jako statisticky významně lepší než bez použití slepé zpětné vazby na hladině významnosti $\alpha \leq 0,01$. Pro nejlepší variantu nastavení nalezenou na trénovací množině dat, $k = 70$, $m = 10$, byl výsledek ověřen i na testovací množině, kde se potvrdila statistická významnost vylepšení výsledků vyhledávání na hladině významnosti $\alpha \leq 0,05$.

7.6.3 Jazykové modely

Méně rozsáhlé experimenty byly provedeny v Query likelihood modelu s Jelinek-Mercer vyhlazováním, byl testován výběr $k \in \{5, 10, 20, 30, 40, 50, 100\}$ pseudo relevantních dokumentů, počet termů byl volen stejně jako u vektorového modelu od 5 do 40 termů, s rozestupem 5 termů. Na základě experimentů s vektorovým modelem již nebyl testován výběr velmi velkého počtu dokumentů. Jako váhová funkce termu w_t byla zvolena váha definovaná rovnicí (6.18) [114]. Výsledky jsou ukázány v tabulce B.4 v příloze B. Nejlepších výsledků bylo dosaženo při volbě 20 dokumentů s výběrem 20 nebo 30 termů, téměř stejných výsledků s 30 dokumenty. Dalšími experimenty bylo zjištěno, že při výběru 40 termů ze 100 dokumentů je výsledek téměř stejný jako ten pro 20 dokumentů. Při výběru více dokumentů se už mGAP skóre nezlepšovalo, stejně tak při výběru 45 termů a více. Nejlepší výsledky jsou ukázány v tabulce 7.15.

Je vidět, že dochází částečně ke stejnému trendu jako u vektorového modelu, tedy nejlepších výsledků je dosaženo při výběru relativně malého množství dokumentů a většího množství termů. Oproti vektorovému modelu při výběru většího množství dokumentů dochází k lepším výsledkům při výběru také většího množství termů. Ohodnocení termů v jazykovém modelování vybírá „lepší“ termy ve smyslu jejich relevance k původnímu dotazu, i při větším přidaném množství termů tedy nedochází k posunu dotazu od původního významu a tím ke zhoršení výsledků vyhledávání.

Tabulka 7.15: Porovnání výsledků slepé zpětné vazby (BRF) v Query likelihood modelu (QL) s Jelinek-Mercer (JM) vyhlazováním (mGAP)

témata	**trénovací**	**testovací**
metoda	**tdn**	**tdn**
QL model s Jelinek-Mercer v.	0,0392	0,0255
QL model s JM, k=20,m=30	0,0513	0,0245
QL model s JM, k=20,m=20	0,0513	0,0229
QL model s JM, k=100,m=40	0,0503	0,0237

Při ověřování výsledků na testovací sadě témat došlo k zajímavému poznatku zhoršení výsledků vyhledávání při použití slepé zpětné vazby s nastavením trénovací množiny témat. Při dalším průzkumu nastavení parametrů k a m na testovací sadě témat bylo zjištěno, že alespoň stejných výsledků, jako bez použití slepé zpětné vazby, je dosaženo při malém množství přidaným termů (5-10).

Další experimenty byly provedeny i pro Query likelihood model s dvoufázovým vyhlazováním (viz tabulka B.5) a Dirichlet vyhlazováním (tabulkaB.6) pro zjištění trendu závislosti výsledků vyhledávání na volbě parametrů a nalezení nejlepšího nastavení slepé zpětné vazby. Tyto metody vykazují stejný trend jako u Jelinek-Mercer vyhlazování, s lepšími výsledky s větším počtem pseudo relevantních dokumentů, nejlepších výsledků bylo dosaženo v obou případech pro 100 dokumentů a 30 termů (nad rámec prezentované tabulky byly ještě provedeny jednotlivé experimenty se zvýšením počtu dokumentů a termů, ale dosažené výsledky vyhledávání se již nezlepšovaly). Stejně tak mají metody dobré výsledky i pro malý počet dokumentů a větší počet termů jako u Jelinek-Mercer vyhlazování. Nejlepší výsledky jsou prezentovány v tabulce 7.16.

Tabulka 7.16: Porovnání výsledků slepé zpětné vazby (BRF) v Query likelihood modelu (QL) s metodami vyhlazování: Dirichlet (D) a dvoufázovým vyhlazováním (TS) (mGAP)

témata	**trénovací tdn**
QL model s Dirichlet v.	0,0413
QL model s D, k=100,m=30	0,0527
QL model s D, k=30,m=20	0,0544
QL model s dvoufázovým v.	0,0445
QL model s TS, k=100,m=30	0,0574

Všechny prezentované nejlepší výsledky metod použití slepé zpětné vazby na trénovací množině témat se ukázaly jako statisticky významně lepší než bez použití slepé zpětné

vazby na hladině významnosti $\alpha \leq 0,05$.

7.6.4 Shrnutí experimentů se slepou zpětnou vazbou

Experimenty se slepou zpětnou vazbou ukazují důležitost jejího použití ve vyhledávání informací. Zapojením slepé zpětné vazby u vektorového modelu a u všech variant jazykového modelování dochází k statisticky významnému vylepšení výsledků vyhledávání. Tyto experimenty slouží zejména pro nalezení nejlepších možných výsledků dosažitelných při použití klasické slepé zpětné vazby s předem stanoveným počtem pseudo relevantních dokumentů. Tyto výsledky poté budeme porovnávat s výsledky metod představených v následující kapitole 7.7, abychom mohli zjistit, zda nové metody dosahují lepších výsledků, než těch kterých jsou schopny dosáhnout metody slepé zpětné vazby s pevně nastaveným počtem dokumentů.

7.7 Metody pro normalizaci skóre využité pro slepou zpětnou vazbu

V klasické slepé zpětné vazbě, tak jak byla představena v předchozích experimentech v podkapitole 7.6, je počet pseudo relevantních dokumentů nastaven pro všechny dotazy stejně. Nastavení bývá většinou provedeno na základě zvyku, případně omezených experimentů, pro výběr prvních k dokumentů. V této podkapitole budou představeny nové metody, snažící se nalézt počet pseudo relevantních dokumentů dynamicky, pro každý dotaz samostatně, na základě charakteristik dotazu a zároveň dokumentů v kolekci, které se promítají do skóre vyhledaných dokumentů.

Navržené metody jsou inspirovány metodami pro normalizaci skóre používanými v oblasti identifikace / verifikace řečníka [146, 183]. První experimenty s těmito metodami odvozenými pro použití v oblasti vyhledávání informací, byly publikovány v článku [148]. V této práci byl odvozen obecný postup normalizace skóre pro využití ve vyhledávání informací pro Query likelihood model a první normalizační metoda založená na normalizaci univerzálním modelem pozadí (viz 7.8.3). V práci [149] byly přidány další dvě normalizační metody a bylo provedeno důkladné otestování metod pro různé parametry. Následně bylo v práci [151] provedeno odvození pro vektorový model a v práci [150] otestování pro další varianty metod jazykového modelování.

7.7.1 Odvození metod pro Query likelihood model

Metody normalizace skóre jsou používány v oblasti textově nezávislé identifikace řečníka z otevřené množiny (OSTI-SI[8]) [146]. Obor identifikace řečníka se zabývá nalezením řečníka, který řekl testovanou promluvu. Úloha OSTI-SI se dá popsat jako dvoufázový problém. V první fázi musí být nalezen model řečníka, který nejlépe odpovídá řečené promluvě. V druhé fázi musí být rozhodnuto, zda byla promluva opravdu řečena tímto modelem, nebo nějakým jiným řečníkem mimo testovanou množinu. Obtížnost úkolu také spočívá v tom, že testovaná promluva nemusí být stejná jako ta, na které byl systém natrénován. Hledá se tedy takový model, který mohl vygenerovat danou promluvu. Metody normalizace skóre se používají pro kompenzaci zkreslení promluvy v druhé fázi identifikačního problému[146].

[8] Open-Set Text-Independent Speaker Identification

Stejným způsobem by se dal popsat problém slepé zpětné vazby ve vyhledávání informací: první musíme vyhledat dokumenty, které mají nejlepší skóre vzhledem k testovanému dotazu a v druhé fázi musíme určit, zda jsou dokumenty relevantní. V případě systému pro vyhledávání informací bez zpětné vazby druhý krok není potřebný, předpokládá se předložení seřazených výsledků uživateli, který si relevantní dokumenty vybere sám. V případě použití slepé zpětné vazby je na druhou stranu tento krok velmi důležitý, vyhledávací systém musí rozhodnout o výběru pseudo relevantních dokumentů pro další použití v metodách slepé zpětné vazby.

V následující části práce budou představeny odvozené metody normalizace skóre pro vyhledávání informací inspirované OSTI-SI přístupem. Metody se dají použít podobným způsobem jako v identifikaci řečníka, jen je budeme aplikovat nejen na skóre nejlepšího modelu, ale i na skóre dalších modelů v pořadí, protože na rozdíl od oblasti identifikace řečníka předpokládáme více relevantních dokumentů k jednomu dotazu.

Stejně jako v identifikaci řečníka, hledáme pomocí Query likelihood modelu ty modely dokumentu, které mohly vygenerovat daný dotaz. Po prvním běhu vyhledávacího systému získáme seřazený seznam dokumentů spolu s jejich věrohodnostními skóre $p(d|q)$ získanými pomocí rovnice (4.16). Cílem je ve vyhledaném seznamu dokumentů stanovit práh pro výběr relevantních dokumentů jiným způsobem, než standardně používaným výběrem prvních k dokumentů. Definujeme si tedy:

$$p(d_R|q) > p(d_I|q) \rightarrow q \in d_R \quad \text{jinak} \quad q \in d_I, \tag{7.2}$$

kde $p(d_R|q)$ je skóre relevantního documentu d_R a $p(d_I|q)$ je skóre nerelevantního dokumentu d_I. Pomocí Bayesova pravidla můžeme rovnici přepsat:

$$\frac{p(q|d_R)}{p(q|d_I)} > \frac{P(d_I)}{P(d_R)} \rightarrow q \in d_R \quad \text{jinak} \quad q \in d_I, \tag{7.3}$$

kde

$$l(q) = \frac{p(q|d_R)}{p(q|d_I)} \tag{7.4}$$

je normalizované věrohodnostní skóre a

$$\theta = \frac{P(d_I)}{P(d_R)} \tag{7.5}$$

je práh, který chceme nalézt. Nastavení prahu θ dopředu je obtížné, protože nevíme apriorní pravděpodobnosti relevantních a nerelevantních dokumentů $P(d_I)$ a $P(d_R)$. Vzhledem k tomu, že věrohodnostní skóre v Query likelihood modelu počítáme v logaritmické oblasti, můžeme normalizační rovnici (7.4) stejně jako v oblasti OSTI-SI [146] převést:

$$L(q) = \log p(q|d_R) - \log p(q|d_I), \tag{7.6}$$

kde $p(q|d_R)$ je skóre relevantního dokumentu a $p(q|d_I)$ je skóre nerelevantního dokumentu. Protože normalizační skóre $\log p(q|d_I)$ nerelevantního dokumentu není známé, je nutné ho nějakým způsobem odhadnout. Skóre lze odhadnout několika způsoby:

Normalizace univerzálním modelem pozadí (UBMN)

Neznámý model d_I lze odhadnout pomocí modelu kolekce M_c, který byl vytvořen jako jazykový model všech dokumentů v kolekci. Tato technika je inspirována metodou normalizace pomocí univerzálního modelu pozadí (UBMN[9]) [119]. Metoda byla použita pro první experimenty s normalizací skóre v oblasti vyhledávání informací a publikována v práci [148]. Normalizační skóre modelu nerelevantního dokumentu definujeme takto:

$$\log p(q|d_I) = \log p(q|M_c). \tag{7.7}$$

Normalizace neomezenou kohortou (UCN)

V metodě UCN[10] pro každý model dokumentu definujeme množinu (kohortu) N nejpodobnějších modelů $C = \{d_1, ..., d_N\}$ [6]. Tyto modely jsou vybrány jako ty nejbližší k testovanému modelu dokumentu, tedy modely s nejbližšími nižšími věrohodnostními skóre k danému dokumentu. Normalizační skóre je definováno:

$$\log p(q|d_I) = \log p(q|d_{UCN}) = \frac{1}{N}\sum_{n=1}^{N} \log p(q|d_n). \tag{7.8}$$

Standardizace rozložení věrohodnostního skóre (T-norm)

Další variantou normalizace skóre namísto odhadu pravděpodobnosti nerelevantního modelu je přímá transformace rozložení věrohodnostního skóre. Tato metoda byla v oblasti verifikace řečníka představena v práci [6] jako metoda T-norm[11]. Rovnice (7.6) bude převedena do tvaru:

$$L(q) = \frac{\log p(q|d_R) - \mu(q)}{\sigma(q)}, \tag{7.9}$$

kde $\mu(q)$ a $\sigma(q)$ je střední hodnota a směrodatná odchylka rozložení věrohodností dokumentů k dotazu. Přístup je podobný metodě UBMN, rozdíl je zde v použití směrodatné odchylky rozložení věrohodnostních skóre.

Nastavení prahu

Přestože jsou po aplikaci uvedených metod věrohodností skóre normalizovaná, musíme nastavit práh θ pro verifikaci relevance každého z dokumentů v seznamu výsledků vyhledávání. Nastavením tohoto prahu určíme hranici mezi relevantními a nerelevantními dokumenty. Nastavení tohoto prahu je robustnější v seznamu normalizovaných věrohodnostních skóre, protože normalizace odstraňuje vliv obecných charakteristik dotazu a zvětšuje tím rozdíl ve skóre relevantních a nerelevantních dokumentů. Práh θ definujeme jako poměr r nejlepšího normalizovaného skóre, tedy nejvýše hodnoceného modelu dokumentu.

[9] Universal Background Model Normalization
[10] Unconstrained Cohort Normalization
[11] Test normalization

7.7.2 Úprava metod pro použití ve vektorovém modelu

V předchozím textu podkapitoly 7.7.1 byly metody normalizace skóre odvozeny pro použití v Query likelihood modelu. Pokud je budeme chtít použít také ve vektorovém modelu, musíme je upravit. První je potřeba upravit normalizační rovnici (7.6). Věrohodnostní skóre $p(d|q)$ nahradíme podobností dokumentu a dotazu $sim_{d_j,q}$, protože jsou ale věrohodnostní skóre vyjádřeny logaritmy pravděpodobnosti, musíme rovnici upravit do tvaru:

$$l(q) = \frac{sim_{d_R,q}}{sim_{d_I,q}}. \tag{7.10}$$

Jednotlivé normalizační metody pro odhad podobnosti nerelevantního dokumentu musí být také upraveny:

Normalizace neomezenou kohortou (UCN)

Normalizační vztah pro metodu UCN můžeme přepsat:

$$sim_{d_I,q} = \frac{1}{N}\sum_{n=1}^{N} sim_{d_n,q}, \tag{7.11}$$

Standardizace rozložení skóre (T-norm)

Standardizace rozložení T-norm bude mít ve vektorovém modelu tvar:

$$l(q) = \frac{sim_{d_R,q} - \mu(q)}{\sigma(q)}. \tag{7.12}$$

7.7.3 Výsledky experimentů metod pro normalizaci skóre

Pro všechny metody normalizace skóre byly provedeny detailní experimenty ve všech prezentovaných metodách pro vyhledávání informací. Vzhledem k obrovskému množství možných kombinací parametrů bylo rozhodnuto kvůli lepší prezentaci výsledků pro společnou volbu počtu termů, použitých pro rozšíření dotazu pro všechny experimenty. Pro metody normalizace skóre byly provedeny experimenty s volbou počtu termů stejně jako v podkapitole 7.6, tedy volba m od 5 do 40 termů, s rozestupem 5 termů. Stejně jako při klasické slepé zpětné vazbě, metody dosahovaly dobrých výsledků při volbě většího počtu termů, kolem 30. Pro všechny metody budou tedy prezentovány výsledky při použití 30 termů pro rozšíření dotazu, i přesto, že při jiné kombinaci parametrů bylo dosaženo lepších výsledků.

Při použití metod pro normalizaci skóre je počet pseudo relevantních dokumentů stanoven pomocí volby prahu θ, definovaného jako poměr r nejlepšího normalizovaného skóre. Takto zvolený počet pseudo relevantních dokumentů je pak různý pro každý dotaz. Byly provedeny experimenty s volbou poměru r z intervalu $\langle 0,1; 0,95\rangle$ s odstupem 0,05.

Pro metodu UBMN je r jediným parametrem, který musíme zvolit. Pro metodu UCN musíme kromě poměru r nastavit také velikost kohorty C. Byly provedeny experimenty s nastavením C od 5 do 800, s rozdílem 10. Poměr r a velikost kohorty na sobě nepřímo závisí, normalizační skóre v rovnici 7.8 a 7.11 je větší pro menší kohortu (skóre metod

Tabulka 7.17: Porovnání výsledků metod pro normalizaci skóre s výsledky klasické slepé zpětné vazby (ZV) a bez použití zpětné vazby pro vektorový model TF-IDF, Query likelihood model (QL) s Jelinek-Mercer (JM), Dirichlet (D) a dvoufázovým (TS) vyhlazováním na trénovací množině témat **tdn** (mGAP)

metoda	**bez ZV**	**ZV**	**UBMN**	**UCN**	**T-norm**
parametry	**-**	**k=30**	-	**r=0,95, C=295**	**r=0,35**
TF-IDF	0,0456	0,0550	-	0,0601	0,0588
parametry	**-**	**k=20**	**r=0,5**	**r=0,25, C=85**	**r=0,55**
QL-JM	0,0392	0,0513	0,0568	0,0570	0,0564
parametry	**-**	**k=100**	**r=0,5**	**r=0,3, C=145**	**r=0,55**
QL-D	0,0413	0,0527	0,0562	0,0576	0,0546
parametry	**-**	**k=100**	**r=0,9**	**r=0,3, C=245**	**r=0,55**
QL-TS	0,0445	0,0574	0,0557	0,0614	0,0614

na vyšších vyhledaných pozicích je větší, průměr tedy bude větší). Stejně jako u metody UBMN, pro metodu T-norm je potřeba nastavit pouze poměr r. Použité parametry jsou uvedeny u jednotlivých metod. Nastavení poměru r není příliš citlivé, u všech testovaných metod se ukázalo, že změna tohoto parametru příliš neovlivní výsledky vyhledávání.

Ve vektorovém modelu nebyl navržen ekvivalent metody normalizace modelem univerzálního pozadí (UBMN), na rozdíl od metod jazykového modelování, kde ekvivalent tohoto modelu je zřejmý, ve vektorovém modelu žádný podobný model není používán.

Finální porovnání výsledků pro vektorový model TF-IDF, Query likelihood model (QL) s Jelinek-Mercer (JM), Dirichlet (D) a dvoufázovým (TS) vyhlazováním je ukázáno v tabulce 7.17. Jak je vidět, metody normalizace skóre ve všech případech dosáhly lepších výsledků než klasická slepá zpětná vazba (kromě UBMN u dvoufázového vyhlazování).

Pro každou metodu vyhledávání informací byl proveden test statistické významnosti zlepšení výsledků pomocí metod pro normalizaci skóre. Byla zvolena metoda pro normalizaci skóre s nejlepším výsledkem pro danou metodu vyhledávání informací a její výsledek byl porovnán s klasickou slepou zpětnou vazbou nastavenou pouze pomocí výběru předem stanoveného počtu dokumentů. Pro všechny testované metody dosahovala nejlepších výsledků normalizační metoda UCN. Pro Query likelihood model s Jelinek-Mercer vyhlazováním a vektorový model TF-IDF, se potvrdila statistická významnost zlepšení výsledků vyhledávání na hladině významnosti $\alpha \leq 0,05$. U druhých dvou metod se zlepšení výsledků ukázalo jako statisticky nevýznamné.

Experimenty na testovací množině témat byly provedeny pro vektorový model TF-IDF a Query likelihood model s Jelinek-Mercer vyhlazováním. Výsledky experimentů jsou ukázány v tabulce 7.18.

Na výsledcích testovací množiny témat je vidět, že pro vektorový model se potvrzují výsledky dosažené na trénovací množině témat, metody pro normalizaci skóre dosahují lepších výsledků než klasické nastavení slepé zpětné vazby. Wilcoxonův test pro ověření hypotézy vylepšení výsledků vyhledávání ukázal, že vylepšení dosažené klasickou slepou zpětnou vazbou oproti TF-IDF modelu bez zpětné vazby není statisticky významné, stejně tak u metody UCN. Vylepšení dosažené metodou T-norm oproti výsledkům bez slepé zpětné vazby se ukázalo statisticky významné na hladině $\alpha \leq 0,05$, stejně tak i vylepšení získané metodou T-norm oproti klasické slepé zpětné vazbě. Na výsledcích Query likelihood modelu s Jelinek-Mercer vyhlazováním je vidět, že při použití normalizační me-

Tabulka 7.18: Porovnání nejlepších výsledků metod pro normalizaci skóre s výsledky klasické slepé zpětné vazby (ZV) a bez použití zpětné vazby pro vektorový model TF-IDF a Query likelihood model (QL) s Jelinek-Mercer (JM)) vyhlazováním na testovací množině témat **tdn** (mGAP)

metoda	**bez ZV**	**ZV**	**UBMN**	**UCN**	**T-norm**
parametry	**-**	**k=30**	-	**r=0,95, C=295**	**r=0,35**
TF-IDF	0,0224	0,0245	-	0,0257	0,0265
parametry	**-**	**k=20**	**r=0,5**	**r=0,25, C=85**	**r=0,55**
QL-JM	0,0255	0,0245	0,0257	0,0230	

tody UCN bylo dosaženo lepšího skóre než s klasickou slepou zpětnou vazbou, nicméně bylo dosaženo pouze toho, že na rozdíl od klasické slepé zpětné vazby nedošlo ke zhoršení výsledků vyhledávání.

7.7.4 Shrnutí dosažených výsledků

Výsledky představených metod pro normalizaci skóre ukazují, že těmito metodami lze dosáhnout významného zlepšení vyhledávání oproti použití klasicky nastavené slepé zpětné vazby pomocí pevně stanoveného počtu dokumentů. U téměř všech metod se ukázalo, že s jejich pomocí lze dosáhnout lepších výsledků, než jsou nejlepší dosažitelné výsledky pomocí klasické slepé zpětné vazby. Výhodou normalizačních metod je, že stanoví pro každý vyhledávaný dotaz jiný počet pseudo relevantních dokumentů. Tento postup lépe odpovídá reálné situaci, kdy ke každému dotazu přísluší jiný počet relevantních dokumentů v kolekci.

Na výsledcích testovaných metod na testovací množině témat je vidět, že zvolené rozdělení dotazů do trénovací a testovací množiny není vhodné z hlediska požadavku na jejich reprezentaci obecného prostoru dotazů. Trénovací množina plně nepopisuje tento prostor a není tedy možné dobře nastavit vyhledávací algoritmy. K tomuto předpokladu vedou nejen obecně horší výsledky všech metod na testovací množině, ale také vliv nastavení parametrů u jazykových modelů, kdy pro jednu množinu témat je lepší co největší vyhlazování a pro druhou co nejmenší, tedy poměr mezi vlivem ohodnocení termu v testovaném dokumentu a v celé kolekci je opačný. Pro lepší vyhodnocení navrhovaných metod by bylo vhodné navrhnout jiné dělení všech témat v této kolekci tak, aby byly obě množiny více reprezentativní vzhledem k předpokládanému prostoru dotazů a mohli jsme pak předpokládat, že parametry metod navržené na trénovací množině a ověřené na testovací množině témat, budou obecně platné, tedy nejlepší i pro všechny hypotetické dotazy, které by mohl uživatel systému položit.

Pro ověření obecné platnosti navrhovaných metod normalizace skóre byly metody použity a otestovány také v oblasti multi-label detekce tématu textu, která je blízká oblasti vyhledávání informací. Metody tak budou ověřeny nejen na jiné kolekci, ale i v jiné úloze.

7.8 Multi-label detekce tématu textu

Metody představené v podkapitole 7.7 byly otestovány také v prostředí oboru multi-label detekce tématu textu, kde je každému dokumentu přiřazeno více než jedno téma. Oproti multi-class klasifikaci dokumentů, která se zabývá disjunktním roztříděním doku-

mentů mezi několik tříd, je cílem této úlohy rozhodnout pro každé téma z množiny témat, zda patří nebo nepatří k danému dokumentu. Detailní přehled metod používaných pro tuto úlohu lze nalézt v práci [162].

Experimenty byly provedeny na kolekci novinových článků, vytvořené v rámci aplikace pro získávání a ukládání velkého množství dat pro budoucí použití jako trénovací data pro odhad jazykových modelů pro zpracování řeči [166]. Protože se ukázalo, že pro dobrý odhad těchto modelů nestačí pouze velké množství dat, ale také roztřídění podle jejich zaměření [115], bylo nutné novinové články také třídit podle jejich tématu [153].

Pro řešení této úlohy se často používá množina binárních klasifikátorů, jeden pro každé téma [94], kde pro každý z nich musí být stanoven práh určující pozitivní klasifikaci. Tento přístup se dá použít v případě malého množství témat, například deseti, kde je velké množství trénovacích dat pro každé téma, ale v reálné aplikaci v případě většího množství témat (450 v tomto případě), kde některá mají velmi malé množství trénovacích dat, se nedá použít. Alternativou je použít jeden generativní klasifikátor jako například Naive Bayes klasifikátor [5, 99], jehož výstupem jsou pravděpodobnostní, nebo věrohodnostní skóre pro každé téma, že náleží k danému článku. V tomto přístupu je nutné nastavit pouze jeden práh určující rozdíl mezi „správnými“ a „nesprávnými“ tématy článku. Jde tedy o podobnou úlohu jako je stanovení prahu pro relevantní dokumenty v případě vyhledávání informací. Stejně tak jako ve vyhledávání informací, stanovení jednotného prahu pro všechny tříděné novinové články není optimální, je tedy dobré stanovit práh dynamicky pro každý článek zvlášť.

7.8.1 Stanovení prahu pro generativní klasifikátor

Problémem při použití generativního klasifikátoru je nutnost výběru „správných“ témat z výstupního rozložení věrohodností témat. Nejjednodušším způsobem jak toho dosáhnout, je vybrat témata, která mají větší věrohodnost než předem stanovený pevný práh, nebo výběr předem stanoveného počtu témat [5]. V našich prvních experimentech byla vybírána tři témata pro každý článek [153]. V práci [99] je tento problém obejit vytvořením modelů pro všechny možné kombinace témat a poté přiřazením toho smíšeného modelu, který dosáhl nejvyšší pravděpodobnosti. Tento přístup ale není vhodný pro větší množství témat.

Nalezením metody pro dynamické stanovení prahu se zabývá pouze práce [9], kde se práh θ stanoví jako střední hodnota plus směrodatná odchylka z rozložení věrohodností témat. Témata s větší věrohodností než tento práh jsou přiřazena dokumentu:

$$\theta = \frac{\sum_1^{|T|} p(A|T_i)}{|T|} + \sqrt{\frac{\sum (p(A|T_i) - \frac{\sum_1^{|T|} p(A|T_i)}{|T|})^2}{|T|}}, \tag{7.13}$$

kde T je množina všech témat, $|T|$ je počet těchto témat a $p(A|T_i)$ je věrohodnost tématu k danému dokumentu.

7.8.2 Naive Bayes klasifikátor

Pro experimenty byl použit Naive Bayes klasifikátor, který formálně odpovídá přístupu jazykového modelování ve vyhledávání informací [97], tedy Query likelihood modelu. Každé téma je definováno pomocí unigramového jazykového modelu vytvořeného z trénovacích dokumentů daného tématu a pravděpodobnost, že dokument A byl vygenerován

tématem T_i je vyjádřena podmíněnou pravděpodobností $P(T_i|A)$. Stejně jako ve vyhledávání informací v podkapitole 4.5.1 zanedbáme apriorní pravděpodobnost dokumentu $P(A)$:

$$P(T_i|A) \propto \frac{P(T_i)p(A|T_i)}{P(A)} \propto p(A|T_i) = \prod_{t \in A} p(t|T_i), \tag{7.14}$$

kde $P(T_i)$ je apriorní pravděpodobnost tématu T_i, která může být odhadnuta jako relativní frekvence výskytu tématu v trénovacích datech, nebo ji můžeme považovat za uniformní a vynechat ji [147]. Rozdělení věrohodností $p(A|T_i)$ je potom použito k určení témat náležejících k dokumentu. Pravděpodobnost $p(t|T_i)$ je odhadnuta jako relativní frekvence termu t v trénovacích datech tématu T_i.

7.8.3 Metody normalizace skóre

Výsledkem Naive Bayes klasifikace je rozdělení věrohodností témat $p(A|T)$ a je nutné stanovit práh pro rozlišení témat příslušejících k dokumentu. Stejně jako u zpětné vazby ve vyhledávání informací k tomu můžeme použít metody popsané v podkapitole 7.7. Můžeme tedy definovat následující vztah:

$$P(T_C|A) > P(T_I|A) \rightarrow A \in T_C \quad \text{jinak} \quad A \in T_I, \tag{7.15}$$

kde $P(T_C|A)$ je skóre modelu „správného“ tématu T_C a $P(T_I|A)$ je skóre modelu „nesprávného“ tématu T_I. Pomocí Bayesova pravidla můžeme rovnici (7.15) přepsat:

$$\frac{p(A|T_C)}{p(A|T_I)} > \frac{P(T_I)}{P(T_C)} \rightarrow A \in T_C \quad \text{jinak} \quad A \in T_I, \tag{7.16}$$

kde

$$l(A) = \frac{p(A|T_C)}{p(A|T_I)} \tag{7.17}$$

je normalizované skóre věrohodnosti a

$$\theta = \frac{P(T_I)}{P(T_C)} \tag{7.18}$$

je požadovaný práh. Nastavení prahu θ je obtížné, protože neznáme pravděpodobnosti $P(T_I)$ a $P(T_C)$. Často používaná forma normalizační rovnice (7.17) v oblasti identifikace řečníka [146] může být převedena pro tuto úlohu na:

$$L(A) = \log p(A|T_C) - \log p(A|T_I). \tag{7.19}$$

Protože skóre „nesprávného“ tématu $\log p(A|T_I)$ neznáme, musíme ho stejně jako v podkapitole 7.7 odhadnout pomocí některé z normalizačních metod upravených pro úlohu detekce tématu.

Normalizace univerzálním modelem pozadí (UBMN)

Stejně jako v podkapitole 7.7, model pozadí (UBMN[12]) [119] je zde chápán jako obecný model prostředí dokumentů. Tato metoda byla publikována jako General topic model normalization (GTMN) v článku [147]. Model T_I je aproximován obecným modelem tématu G, který byl vytvořen jako jazykový model všech dokumentů v trénovací kolekci. Normalizační skóre T_I je definováno jako:

$$\log p(A|T_I) = \log p(A|G). \tag{7.20}$$

Normalizace neomezenou kohortou (UCN)

Metoda normalizace neomezenou kohortou (UCN[13]) [6] pro detekci tématu, byla publikována v práci [154]. Pro každý model tématu je vybrána množina N modelů nejpodobnějších témat $C = \{T_1, ..., T_N\}$, jsou to modely s nejvyšší věrohodností menší než daný model. Normalizační skóre je pak dáno vzorcem:

$$\log p(A|T_I) = \log p(A|T_{UCN}) = \frac{1}{N}\sum_{n=1}^{N} \log p(A|T_n). \tag{7.21}$$

Normalizace kohortou (CoN)

Metoda normalizace kohortou (CoN[14]) [188] používá množinu podobných modelů C, definovanou dopředu, před začátkem klasifikace. V oblasti identifikace řečníka jsou nejpodobnější řečníci vybráni jako ti nejblíže v prostoru řečníka [119, 128]. V oblasti detekce tématu jsme v práci [155] navrhli podobnost dvou modelů tématu na základě jejich blízkosti v stromu témat. Pro daný model tématu se kohorta C skládá z témat, která jsou na stejné úrovni a mají stejný nadřazený uzel. Takto definovaná množina má jinou velikost N pro každé téma. Normalizační skóre je definováno stejnou rovnicí jako u neomezené kohorty (7.21), jen výběr množiny C je rozdílný.

Standardizace rozložení skóre (T-norm)

Další možností je použít metodu T-norm[15] definovanou v práci [146] upravující přímo rozložení věrohodnosti témat. Upravená rovnice (7.19) získá tvar:

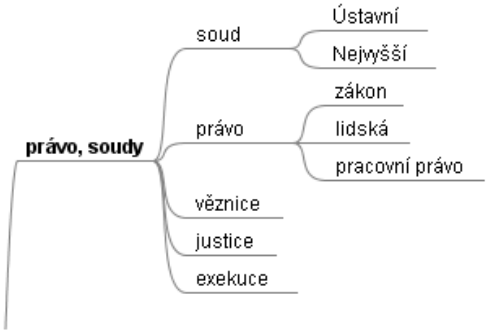
$$L(A) = \frac{\log p(A|T_C) - \mu(A)}{\sigma(A)}, \tag{7.22}$$

kde $\mu(A)$ a $\sigma(A)$ je střední hodnota a směrodatná odchylka celého rozdělení věrohodnosti témat.

Nastavení prahu

Stejně jako v úloze výběru relevantního dokumentu ve vyhledávání informací i zde musíme po normalizaci ještě nastavit práh θ v rovnici (7.16). V experimentech publikovaných

[12] Universal Background Model
[13] Unconstrained Cohort Normalization
[14] Cohort Normalization
[15] Test normalization



Obrázek 7.6: Ukázka jedné větve ze stromu témat

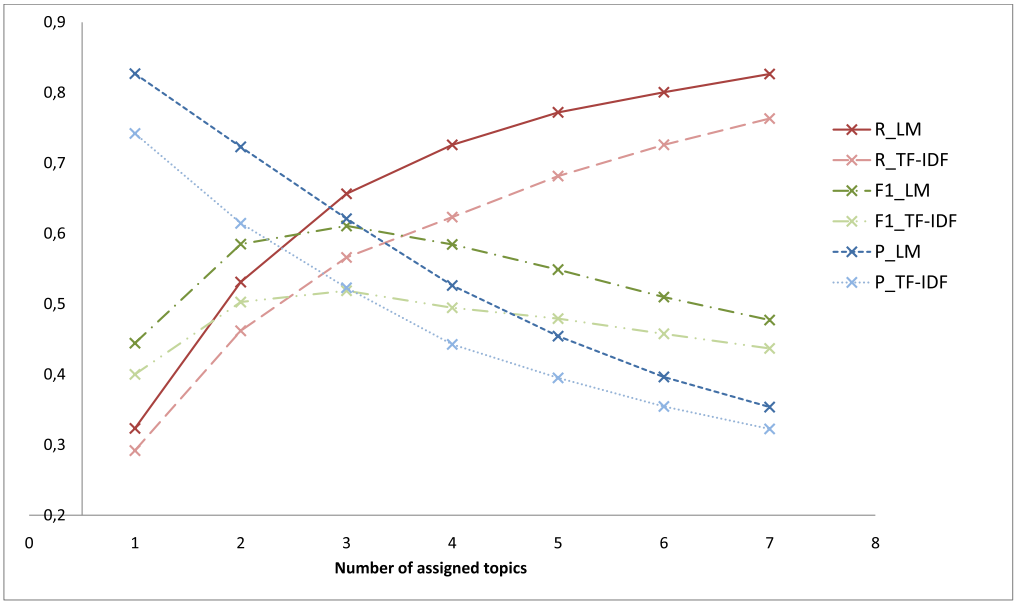
v pracích [147, 154] jsme nastavili práh na 80% normalizovaného skóre nejlepšího tématu, v práci [155] byly publikovány další experimenty s tímto nastavením popsané v podkapitole 7.8.5. Práh θ je definován jako poměr r normalizovaného skóre nejlepšího tématu.

7.8.4 Nastavení experimentů

Pro experimenty byl použit systém popsaný v článku [166], podrobný popis předzpracování textů použitých novinových článků je uveden v pracích [147, 154, 155], všechny texty byly lemmatizovány pomocí algoritmu testovaného v podkapitole 7.4.2 [73]. Z celého korpusu článků byla oddělena kolekce použitá k experimentům publikovaným v pracích [147, 154, 155]. Kolekce obsahuje 31 tisíc novinových článků publikovaných v roce 2011 (od ledna do října) a je rozdělena na 27 tisíc trénovacích a 4 tisíce testovacích článků. Testovací články byly v práci [155] a v zde popsaných experimentech použity jako development data a byla vytvořena nová sada 5 tisíc testovacích článků z roku 2012. Články jsou řazeny podle data jejich vydání, to znamená že všechny development články byly vydány až po trénovacích článcích a testovací články až po development datech. Tento způsob vytvoření kolekce lépe odpovídá reálné situaci, kdy systém klasifikuje nově vydané novinové články na základě natrénovaných modelů z dřívějších článků. V případě použití náhodného rozdělení článků do jednotlivých množin by nejspíše docházelo k lepším výsledkům klasifikace, protože trénovací a testovací množina by mohla obsahovat podobné nebo téměř totožné články (z různých novin vydané na stejné téma, případně z jednoho období opakující se téma).

Témata jsou vybírána z hierarchického stromu témat vytvořeného na základě výskytu témat v českých online zpravodajských denících, strom obsahuje přibližně 450 témat [159]. Stromová struktura je využita pouze pro pozdější zpracování dokumentů příslušných k tématu a pro výběr témat pro metodu CoN, při klasifikaci jsou všechna témata brána jako na stejné úrovni. Ukázka větve stromu témat je vidět na obrázku 7.6.

Výběr 3 témat pro každý článek je založen na prvních experimentech s multi-label detekcí témat na této úloze. V experimentech publikovaných v práci [153] bylo testováno jaké je nejlepší množství témat pro přiřazení ke každému článku na větší kolekci skládající se z 140 tisíc trénovacích článků a 15 tisíc testovacích. Zároveň bylo testováno použití klasifikace pomocí jazykových modelů a vektorového TF-IDF modelu pro vyhledávání informací (viz podkapitola 4.3). Výsledky experimentů jsou vidět na obrázku 7.7, na základě těchto výsledků bylo rozhodnuto pro další používání jazykového modelování jako metody pro detekci tématu a nejlepších výsledků bylo dosaženo při volbě 3 témat pro každý článek.



Obrázek 7.7: Porovnání výsledků pro jazykové modelování (LM) a vektorový model (TF-IDF) v závislosti na volbě počtu přiřazených témat ke každému článku (převzato z [153])

Pro vyhodnocení výsledků klasifikace se v oblasti multi-label detekce tématu používají podobné míry jako pro vyhodnocení výsledků vyhledávání informací. Testovaný dokument (novinový článek) je považován za dotaz a pro vyhodnocení získaných odpovědí (témat) můžeme použít míry přesnosti a úplnosti definované v podkapitole 3.1 [94, 8, 43]. Pro množinu klasifikovaných dokumentů D a klasifikátor H můžeme definovat průměrnou přesnost $P(H, D)$:

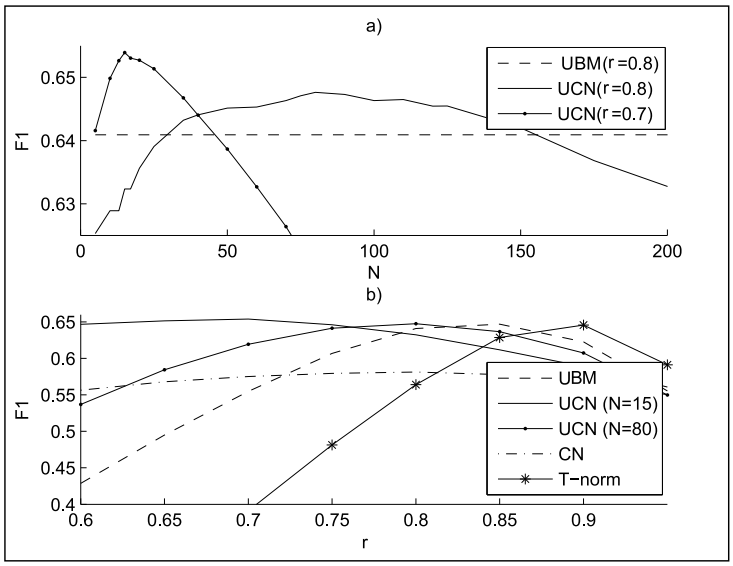
$$P(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{T_C}{T_A} \tag{7.23}$$

a průměrnou úplnost $R(H, D)$:

$$R(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{T_C}{T_R}, \tag{7.24}$$

kde T_A je počet témat přiřazených k článku , T_C je počet správně přiřazených témat a T_R je počet relevantních témat. $F_1(H, D)$ míra (viz podkapitola 3.3), používaná pro jednoduché porovnání dvou metod klasifikace se spočte:

$$F_1(H, D) = 2\frac{P(H, D) \cdot R(H, D)}{P(H, D) + R(H, D)} \tag{7.25}$$



Obrázek 7.8: Porovnání metod pro normalizaci skóre na množině development dat a) závislost UCN na velikosti kohorty N pro pevný podíl $r = 0,7$ a $0,8$ (porovnání s UBMN) b) závislost na podílu r pro nastavení prahu (pro velikost kohorty v UCN $N = 15$ a $N = 80$)

7.8.5 Výsledky experimentů

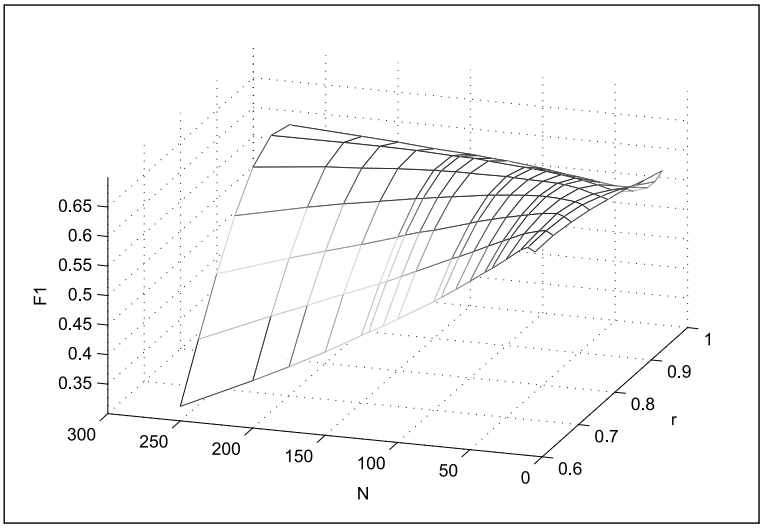
Experimenty s nastavením parametrů jednotlivých metod byly provedeny na development kolekci z roku 2011. Pro metody UBMN, T-norm a CoN je potřeba zjistit nastavení prahu θ. Na obrázku 7.8b) je vidět závislost těchto metod na různém nastavení prahu θ, definovaného jako podíl r z nejlepšího dosaženého věrohodnostního skóre.

Pro metodu UCN je nutné nalézt nejlepší kombinaci nastavení prahu a velikosti kohorty. Na obrázku 7.9 je vidět závislost F míry výsledků na nastavení podílu r a velikosti kohorty. Je vidět, že velikost podílu r je přímo úměrná nastavení velikosti kohorty, protože normalizační skóre v rovnici (7.21) je větší (průměr z větších věrohodností témat) pro malou kohortu. Obrázek 7.8a ukazuje porovnání závislosti F míry výsledků UCN na velikosti kohorty pro dvě různá nastavení podílu r. Pro nastavení prahu jako 80% ($r = 0.8$) nejlepšího dosaženého skóre jsou výsledky téměř stabilní pro různá nastavení velikosti kohorty. Na druhou stranu lepší výsledky se dají získat pro nastavení prahu na 70% ($r = 0.7$) při použití menší kohorty ($N = 15$). Pokud se na to podíváme z jiného úhlu pohledu, při použití malé kohorty je metoda UCN stabilní pro rozdílné nastavení prahu než UCN s větší kohortou a také než ostatní metody (viz obrázek 7.8b).

Tabulka 7.19 ukazuje porovnání nejlepších dosažených výsledků na množině development dat. Pro metodu UCN jsou ukázány dvě možnosti nastavení, menší kohorta $N = 15$ pro větší stabilitu při nastavení prahu a nastavení podílu $r = 0.8$ pro lepší stabilitu při nastavení velikosti kohorty. Výsledky jsou porovnány s pevným nastavením výběru pouze 1 nebo 3 témat, použitým v článcích [5] respektive [153] a nastavením prahu jako střední hodnoty plus směrodatné odchylky rozdělení věrohodnosti témat (MpSD[16]), jak bylo použito v práci [9].

V tabulce 7.20 je vidět porovnání výsledků experimentů na testovací kolekci z roku

[16]Mean plus Standard Deviation



Obrázek 7.9: Závislost F míry výsledků metody UCN na velikosti kohorty N a nastavení podílu r na množině development dat

Tabulka 7.19: Výsledky použití metod normalizace skóre v porovnání s ostatními metodami pro nalezení prahu pro výběr „správných“ témat dokumentu. Nejlepší výsledky experimentů na development kolekci 2011.

metoda	**1 t.**	**3 t.**	**MpSD**	**UBMN**	**CoN**	**UCN**		**T-norm**
parametry	-	-	-	r=0,8	r=0,8	r=0,7 N=15	r=0,8 N=80	r=0,9
$P(H,D)$	0,812	0,586	0,055	0,592	0,605	**0,689**	0,665	0,678
$R(H,D)$	0,319	0,616	0,961	0,699	0,560	**0,622**	0,631	0,616
$F_1(H,D)$	0,458	0,600	0,105	0,641	0,581	**0,654**	0,648	0,646

2012. Pro metody normalizace skóre bylo zvoleno nastavení dosahující nejlepších výsledků F míry na development kolekci.

7.8.6 Shrnutí výsledků

- Z porovnání výsledků v tabulkách 7.19 a 7.20 je vidět, že metody podávají stabilní výkony i na testovací kolekci. Mírně horší výsledky jsou způsobeny větším časovým rozestupem trénovacích a testovacích dat oproti development datům. To ukazuje na jeden z problémů specifický pro detekci témat novinového článku, kdy se obsah tématu mění v závislosti na čase (například změna politické orientace státu) a je tedy nutné neustále aktualizovat modely témat pomocí nově vydaných článků.

- Při výběru pouze jednoho tématu pro každý článek lze dosáhnout vysoké přesnosti, protože první nalezené téma je většinou správné. Naopak úplnost je nízká, což vychází z podstaty dat, která mají většinou přiřazeno více než jedno téma.

- Výběr 3 témat je optimální pro tuto úlohu z hlediska přiřazení pevně předem stanoveného počtu témat. Tato volba je ale závislá na charakteru dat, pokud by články měly větší nebo menší počet referenčních témat, bylo by dosaženo nejlepších výsledků pro

Tabulka 7.20: Výsledky použití metod normalizace skóre v porovnání s ostatními metodami pro nalezení prahu pro výběr „správných“ témat dokumentu. Výsledky experimentů na testovací kolekci 2012 pro nejlepší nastavení metod získané z development kolekce 2011.

metoda	**1 t.**	**3 t.**	**MpSD**	**UBMN**	**CoN**	**UCN**		**T-norm**
parametry	-	-	-	r=0,8	r=0,8	r=0,7 N=15	r=0,8 N=80	r=0,9
$P(H,D)$	0,797	0,570	0,059	0,575	0,598	**0,677**	0,662	0,667
$R(H,D)$	0,308	0,596	0,952	0,669	0,526	**0,583**	0,591	0,584
$F_1(H,D)$	0,444	0,583	0,110	0,618	0,559	**0,626**	0,625	0,623

jinou pevně stanovenou hodnotu. Každému článku je také přiřazeno stejné množství témat, nehledě na jejich věrohodnost.

- Při použití metody MpSD dosáhneme vysoké míry úplnosti, bohužel přesnost je velmi nízká. Metoda přiřazuje každému článku kolem 50 témat, díky tomu je dosaženo nalezení většiny referenčních témat a tudíž vysoké úplnosti. Velmi špatné výsledky této metody oproti prezentovaným dobrým výsledkům v práci [9] můžeme přisoudit tomu, že v původní práci bylo klasifikováno pouze do 10 témat, kdežto v našem případě 450 témat metoda selhává.

- Metody normalizace skóre použité v úloze multi-label detekce tématu textu dosahují lepších výsledků než ostatní metody pro výběr prahu generativního klasifikátoru (kromě metody CoN). Lepších výsledků je dosaženo díky dynamickému výběru počtu přiřazených témat pro každý článek zvlášť oproti předem pevně stanovenému počtu. Principem normalizačních metod je zdůraznit rozdíl ve skóre „správných“ a „nesprávných“ témat, takže je poté jednodušší tyto skupiny pomocí nastaveného prahu od sebe oddělit. Přestože je tedy stále nutné nastavit práh pro výběr „správných“ témat, toto nastavení je více robustní.

- Metoda normalizace skóre pomocí předem stanovené kohorty (CoN) nedosahuje tak dobrých výsledků jako ostatní normalizační metody. Prvním důvodem je špatné počáteční nastavení nejbližších témat pomocí stromu témat, kdy v některých uzlech spolu listová témata téměř nesouvisí, takže článek může mít velkou věrohodnost vůči jednomu z nich, ale malou vůči ostatním. Druhým důvodem je multi-label povaha dat, tedy že každý článek se týká více témat. K článku budou tedy nejbližší nejen formálně podobná témata, ale i témata naprosto rozdílná (například článek o nehodách na silnici v důsledku špatného počasí), metoda ale pro každé téma počítá pouze s okruhem podobných témat. To je rozdíl oproti původní oblasti identifikace řečníka, kde metoda dosahuje dobrých výsledků. V oblasti identifikace řečníka má testovaná promluva podobné charakteristiky jako uložené modely řečníků, při výběru kohorty k jednotlivým modelům dopředu tedy dojde k výběru podobného složení kohorty jako u výběru metodou UCN a tyto metody tak dosahují podobných výsledků. Při použití v oblasti detekce tématu ale metoda UCN vybírá kohortu podle přímé podobnosti článku s tématy, takže ve výsledné kohortě jsou zastoupena vzájemně formálně nepodobná témata, jejichž smíšením ale vznikl testovaný článek.

- Metoda UCN dosahuje lepších výsledků než ostatní metody normalizace skóre, navíc je tato metoda stabilnější při výběru prahu pomocí podílu r při nastavení menší kohorty ($N = 5 - 15$). Téměř stejných výsledků dosahuje i metoda T-norm.

- Testy statistické významnosti výsledků byly provedeny pomocí párového t-testu na výsledcích z testovacích dat kolekce z roku 2012. Byly porovnány metody normalizace skóre oproti nejlepšímu dosažitelnému výsledku při pevné volbě počtu témat, tedy volbě 3 témat. Pro metody UCN, UBMN a T-norm potvrdil párový t-test statistickou významnost zlepšení výsledků oproti výběru 3 témat na hladině významnosti $\alpha = 0,0005$ ($p \leq 0{,}0001$). Zlepšení výsledků metody UCN oproti UBMN se neukázalo statisticky významné, u metody T-norm oproti tomu ano, na hladině významnosti $\alpha = 0,005$ ($p = 0{,}001$). Lze z toho usoudit, že zlepšení u metody T-norm, ač je víceméně stejné jako u UCN, je více stabilní, nedochází tedy k výkyvům u jednotlivých článků (což se i potvrdilo při posouzení výsledků pro jednotlivé články).

Kapitola 8

Závěr

Vyhledávání informací v řeči prochází v současné době velmi rychlým vývojem. Stále více informací je ukládáno v multimediální podobě a jsou veřejně dostupné například prostřednictvím internetu. Je tedy nutné věnovat se vývoji metod pro vyhledávání v takovýchto datech. Výzkum v této oblasti je směřován zejména do dvou hlavních větví. První z nich je zaměřena na lepší propojení ASR systému a systému pro vyhledávání informací využitím informace získané z bohatších reprezentací řeči z ASR systému a cílí na situace, kdy dochází k velké chybovosti automatických přepisů řeči. Druhá větev výzkumu je zaměřena na zlepšení metod vyhledávání informací a zapojení metod rozšíření dotazu a slepé zpětné vazby do vyhledávání informací v řeči. Tato práce se zabývá druhou větví výzkumu, zejména zapojením a vylepšením metod slepé zpětné vazby.

Práce prezentuje způsoby jak lze vyhledávat informace v řeči, od nejjednoduššího postupu - použití metod pro vyhledávání informací představených v kapitole 4 na nejlepší automatické přepisy z ASR systému, po využití kompletních slovních nebo subslovních mřížek a jejich reprezentací pro vyhledávání (viz kapitola 5). Práce se dále zaměřuje na využití slepé zpětné vazby při vyhledávání informací. V kapitole 6 je představen přehled existujících metod zpětné a slepé zpětné vazby, nejen přehled publikovaných vědeckých prací, ale i analýza používaných postupů z hlediska společných vlastností se zaměřením na identifikaci problémů metod slepé zpětné vazby. Jedním ze společných problémů všech metod slepé zpětné vazby je volba počtu pseudo relevantních dokumentů, z kterých jsou vybírány termy pro rozšíření dotazu. Analýza a návrh metod pro řešení tohoto problému je součástí této práce.

V kapitole 7 byly představeny experimenty s vyhledáváním informací v řeči provedené na české kolekci spontánní řeči, vytvořené v rámci CLEF CL-SR úlohy v roce 2006 a 2007. Práce se postupně věnuje jednotlivým aspektům vyhledávání informací, od testování vlivu různého předzpracování vstupních dat, přes porovnání nejpoužívanějších metod pro vyhledávání informací a zapojení slepé zpětné vazby do těchto metod. Pro metody slepé zpětné vazby byla provedena detailní analýza vlivu nastavení jejích parametrů a jejich vzájemného ovlivnění, zvláště vlivu nastavení počtu vybíraných pseudo relevantních dokumentů. Výsledky experimentů jsou shrnuty v podkapitolách 7.4.1, 7.5.5 a 7.6.4.

Všechny dosud používané metody pracovaly za předpokladu, že prvních k dokumentů nalezených systémem pro vyhledávání informací je relevantních. Hodnota k musela být vždy stanovena apriori, většinou byla nastavena na základě zvyku, či několika málo experimentů. Tato hodnota byla pak pro všechny dotazy stejná. V této práci byly navrženy nové metody pro normalizaci skóre dokumentu, představené v podkapitole 7.7. Navržené metody umožňují dynamicky určit počet pseudo relevantních dokumentů pro následné po-

užití v slepé zpětné vazbě. Počet pseudo relevantních dokumentů k je tedy pro každý dotaz jiný, závislý na obsahu daného dotazu. Experimenty představené v práci ukázaly, že s využitím předložených metod pro výběr relevantních dokumentů pro zpětnou vazbu je možné dosáhnout výrazného zlepšení přínosu slepé zpětné vazby a tím i statisticky významného zlepšení výsledků vyhledávání. Výsledky experimentů jsou shrnuty v podkapitole 7.7.3 a 7.7.4.

Představené metody normalizace skóre byly v závěru práce, v podkapitole 7.8, otestovány také v úloze multi-label detekce tématu textu pro výběr „správných“ témat textu z výstupu generativního klasifikátoru. Metody pro normalizaci skóre dosáhly statisticky významného zlepšení výsledků detekce tématu textu oproti ostatním používaným metodám pro výběr témat.

8.1 Shrnutí přínosů práce

- Popsány, programově realizovány a otestovány
  - klasické metody vyhledávání informací, včetně jejich mnoha variant
  - efekty předzpracování dat: lemmatizace, vynechání stop slov
  - metody pro využití slepé zpětné vazby v různých modelech pro vyhledávání informací
- Navrženy a experimentálně ověřeny
  - metoda pro automatickou tvorbu booleovského dotazu
  - možnosti nastavení metod slepé zpětné vazby, jejich vliv na výsledky vyhledávání
  - nové metody pro výběr pseudo relevantních dokumentů pro slepou zpětnou vazbu
  - varianty představených metod pro jednotlivé metody vyhledávání informací
  - nové metody také v příbuzné oblasti detekce tématu textu

Stanovené cíle disertační práce byly splněny, tato práce navíc poskytuje teoretický základ k rozšíření provedených experimentů a k dalšímu testování představených metod pro výběr pseudo relevantních dokumentů ve slepé zpětné vazbě. Budoucí pokračování prezentovaného výzkumu je tedy možné v těchto směrech:

- Rozšíření experimentů do šířky, tedy otestování nově představených metod pro výběr pseudo relevantních dokumentů v prostředí dalších metod pro vyhledávání informací a slepé zpětné vazby.
- Přestože byly metody navrženy pro úlohu vyhledávání informací v řeči, nejsou na tento typ úlohy nijak vázané a lze tedy pokračovat v jejich výzkumu a dalším testování na řádově větších kolekcích pro vyhledávání informací v textu.
- V letošním roce byla znovu vydána kolekce pro vyhledávání informací v české spontánní řeči, na jejíž původní verzi byly provedeny experimenty představené v této práci. Aktuální kolekce obsahuje novou verzi automatického přepisu nahrávek, nejen ve formě nejlepších přepisů, ale i slovních mřížek z ASR systému. Bylo by tedy možné

nejen porovnat vliv nového automatického rozpoznávače řeči na úspěšnost vyhledávání, ale také upravit představené metody pro použití s vyhledáváním ve slovních mřížkách.

Literatura

[1] Aalbersberg, I. J.: Incremental Relevance Feedback. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, New York, NY, USA: ACM, 1992, ISBN 0-89791-523-2, s. 11–22.

[2] Akiba, T.; Nishizaki, H.; Aikawa, K.; aj.: Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop. In *NTCIR*, 2011.

[3] Allan, J.: Relevance Feedback with Too Much Data. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, New York, NY, USA: ACM, 1995, ISBN 0-89791-714-6, s. 337–343.

[4] Allan, J.: Perspectives on Information Retrieval and Speech. In *Information Retrieval Techniques for Speech Applications*, editace A. R. Coden; E. W. Brown; S. Srinivasan, Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, ISBN 978-3-540-45637-7, s. 1–10.

[5] Asy'arie, A. D.; Pribadi, A. W.: Automatic news articles classification in Indonesian language by using Naive Bayes Classifier method. In *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*, New York, USA: ACM, 2009, ISBN 978-1-60558-660-1, s. 658–662.

[6] Auckenthaler, R.; Carey, M.; Lloyd-Thomas, H.: Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, ročník 10, č. 1-3, 2000: s. 42 – 54, ISSN 1051-2004.

[7] Bahl, L. R.; Jelinek, F.; Mercer, R. L.: A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, ročník PAMI-5, č. 2, 1983: s. 179 –190, ISSN 0162-8828.

[8] Boutell, M. R.; Luo, J.; Shen, X.; aj.: Learning multi-label scene classification. *Pattern Recognition*, ročník 37, č. 9, 2004: s. 1757 – 1771, ISSN 0031-3203.

[9] Bracewell, D. B.; Yan, J.; Ren, F.; aj.: Category Classification and Topic Discovery of Japanese and English News Articles. *Electron. Notes Theor. Comput. Sci.*, ročník 225, 2009: s. 51–65, ISSN 1571-0661.

[10] Brin, S.; Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Computer networks and ISDN systems*, Elsevier Science Publishers B. V., 1998, s. 107–117.

[11] Brown, M. G.; Foote, J. T.; Jones, G. J. F.; aj.: Open-vocabulary speech indexing for voice and video mail retrieval. In *Proceedings of the fourth ACM international conference on Multimedia*, MULTIMEDIA '96, New York, NY, USA: ACM, 1996, ISBN 0-89791-871-1, s. 307–316.

[12] Buckley, C.; Salton, G.; Allan, J.: The Effect of Adding Relevance Information in a Relevance Feedback Environment. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, New York, NY, USA: Springer-Verlag New York, Inc., 1994, ISBN 0-387-19889-X, s. 292–300.

[13] Buckley, C.; Salton, G.; Allan, J.; aj.: Automatic Query expansion using SMART : TREC 3. *NIST special publication*, 1995: s. 69–80, ISSN 1048-776X.

[14] Cao, G.; Nie, J.-Y.; Bai, J.: Integrating Word Relationships into Language Models. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, New York, NY, USA: ACM, 2005, ISBN 1-59593-034-5, s. 298–305.

[15] Cao, G.; Nie, J.-Y.; Gao, J.; aj.: Selecting good expansion terms for pseudo-relevance feedback. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, 2008: str. 243.

[16] Carmel, D.; Farchi, E.; Petruschka, Y.; aj.: Automatic Query Wefinement Using Lexical Affinities with Maximal Information Gain. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, New York, NY, USA: ACM, 2002, ISBN 1-58113-561-0, s. 283–290.

[17] Carpineto, C.; de Mori, R.; Romano, G.; aj.: An Information-theoretic Approach to Automatic Query Expansion. *ACM Trans. Inf. Syst.*, ročník 19, č. 1, Leden 2001: s. 1–27, ISSN 1046-8188.

[18] Carpineto, C.; Romano, G.: A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.*, ročník 44, č. 1, Leden 2012: s. 1:1–1:50, ISSN 0360-0300.

[19] Carpineto, C.; Romano, G.; Giannini, V.: Improving Retrieval Feedback with Multiple Term-ranking Function Combination. *ACM Trans. Inf. Syst.*, ročník 20, č. 3, Červenec 2002: s. 259–290, ISSN 1046-8188.

[20] Chelba, C.; Acero, A.: Position specific posterior lattices for indexing speech. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, s. 443–450.

[21] Chelba, C.; Acero, A.: SPEECH OGLE: Indexing Uncertainty for Spoken Document Search. In *Proc. of ACL. Ann Arbor*, 2005.

[22] Chelba, C.; Silva, J.; Acero, A.: Soft indexing of speech content for search in spoken documents. *Comput. Speech Lang.*, ročník 21, 2007: s. 458–478, ISSN 0885-2308.

[23] Chen, B.: Word Topic Models for Spoken Document Retrieval and Transcription. ročník 8, č. 1, Březen 2009: s. 2:1–2:27, ISSN 1530-0226.

[24] Chen, B.; min Wang, H.; shan Lee, L.: Improved spoken document retrieval by exploring extra acoustic and linguistic cues. In *in Proc. European Conf. Speech Communication and Technology (INTERSPEECH)*, 2001, s. 299–302.

[25] Chen, Y. W.; Chen, K. Y.; Wang, H. M.; aj.: Effective pseudo-relevance feedback for spoken document retrieval. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, ISSN 1520-6149, s. 8535–8539.

[26] Cheng Pan, Y.; lin Chang, H.; shan Lee, L.: Analytical comparison between position specific posterior lattices and confusion networks based on words and subword units for spoken document indexing. In *IEEE Workshop on Automatic Speech Recognition Understanding*, Prosinec 2007.

[27] Chia, T. K.; Li, H.; Ng, H. T.: A Statistical Language Modeling Approach to Lattice-Based Spoken Document Retrieval. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague: Association for Computational Linguistics, 2007, s. 810–818.

[28] Chia, T. K.; Sim, K. C.; Li, H.; aj.: A lattice-based approach to query-by-example spoken document retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, New York, NY, USA: ACM, 2008, ISBN 978-1-60558-164-4, s. 363–370.

[29] Chia, T. K.; Sim, K. C.; Li, H.; aj.: Statistical lattice-based spoken document retrieval. *ACM Trans. Inf. Syst.*, ročník 28, 2010: s. 2:1–2:30, ISSN 1046-8188.

[30] Cieri, C.; Graff, D.; Liberman, M.; aj.: The TDT-2 Text And Speech Corpus. In *Proceedings of DARPA broadcast news workshop*, Morgan Kaufmann, 1999, s. 57–60.

[31] Clements, M.; Robertson, S.; Miller, M.: Phonetic searching applied to on-line distance learning modules. In *Proceedings of 2002 IEEE 10th Digital Signal Processing Workshop and the 2nd Signal Processing Education Workshop*, 2002, s. 186 – 191.

[32] CROFT, W.; HARPER, D.: Using Probabilistic Models of Document Retrieval Without Relevance Information. *Journal of Documentation*, ročník 35, č. 4, 04 1979: s. 285–295, ISSN 0022-0418.

[33] Davis, J.; Goadrich, M.: The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, New York, NY, USA: ACM, 2006, ISBN 1-59593-383-2, s. 233–240.

[34] Federico, M.; Bertoldi, N.; Levow, G.-A.; aj.: CLEF 2004 Cross-Language Spoken Document Retrieval Track. In *Multilingual Information Access for Text, Speech and Images*, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2005, s. 816–820.

[35] Federico, M.; Jones, G. J. F.: The CLEF 2003 Cross-Language Spoken Document Retrieval Track. In *Comparative Evaluation of Multilingual Information Access Systems*, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2004, s. 467–475.

[36] Galuščáková, P.: Segmentation Strategies for Passage Retrieval in Audio-visual Documents. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, New York, NY, USA: ACM, 2013, ISBN 978-1-4503-2034-4, s. 1143–1143.

[37] Gao, J.; Nie, J.-Y.; Wu, G.; aj.: Dependence Language Model for Information Retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, New York, NY, USA: ACM, 2004, ISBN 1-58113-881-4, s. 170–177.

[38] Garofolo, J. S.; Auzanne, C. G. P.; Voorhees, E. M.: The TREC Spoken Document Retrieval Track: A Success Story. In *Text Retrieval Conference (TREC) 8*, 2000, s. 16–19.

[39] Garofolo, J. S.; Voorhees, E. M.; Auzanne, C. G. P.; aj.: Spoken Document Retrieval: 1998 Evaluation and Investigation of New Metrics. *Proceedings of the Workshop on Accessing Information in Spoken Audio*, 1999: s. 1–7.

[40] Glass, J.; Hazen, T. J.; Cyphers, S.; aj.: Recent Progress in the MIT Spoken Lecture Processing Project. In *Proc. Interspeech*, 2007, s. 2553–2556.

[41] Glass, J.; Hazen, T. J.; Hetherington, L.; aj.: Analysis and processing of lecture audio data: preliminary investigations. In *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*, SpeechIR '04, Stroudsburg, PA, USA: Association for Computational Linguistics, 2004, s. 9–12.

[42] Glavitsch, U.; Schäuble, P.: A system for retrieving speech documents. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, New York, NY, USA: ACM, 1992, ISBN 0-89791-523-2, s. 168–176.

[43] Godbole, S.; Sarawagi, S.: Discriminative Methods for Multi-labeled Classification. In *Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004. Proceedings*, editace H. Dai; R. Srikant; C. Zhang, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, ISBN 978-3-540-24775-3, s. 22–30.

[44] Haines, D.; Croft, W. B.: Relevance Feedback and Inference Networks. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, New York, NY, USA: ACM, 1993, ISBN 0-89791-605-0, s. 2–11.

[45] Hajič, J.; Hladká, B.: Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of ACL/COLING'98*, 1998, s. 483–490.

[46] Hakkani-Tür, D.; Béchet, F.; Riccardi, G.; aj.: Beyond ASR 1-best: Using word confusion networks in spoken language understanding. *Computer Speech & Language*, ročník 20, č. 4, 2006: s. 495 – 514, ISSN 0885-2308.

[47] Hakkani-Tur, D.; Riccardi, G.: A general algorithm for word graph matrix decomposition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, ročník 1, 2003, ISSN 1520-6149, s. I–596 – I–599 vol.1.

[48] Hansen, J.; Huang, R.; Zhou, B.; aj.: SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word. *IEEE Transactions on Speech and Audio Processing*, ročník 13, č. 5, 2005: s. 712 – 730, ISSN 1063-6676.

[49] Harman, D.: Towards Interactive Query Expansion. In *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '88, New York, NY, USA: ACM, 1988, ISBN 2-7061-0309-4, s. 321–331.

[50] Harman, D.: Information Retrieval. kapitola Relevance Feedback and Other Query Modification Techniques, Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1992, ISBN 0-13-463837-9, s. 241–263.

[51] Harman, D.: Relevance Feedback Revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, New York, NY, USA: ACM, 1992, ISBN 0-89791-523-2, s. 1–10.

[52] Harman, D.: Overview of the First TREC Conference. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, New York, NY, USA: ACM, 1993, ISBN 0-89791-605-0, s. 36–47.

[53] Hauptmann, A. G.; Witbrock, M. J.: Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. In *Intelligent multimedia information retrieval*, AAAI Press, 1997, s. 213–239.

[54] Hauptmann, A. G.; Witbrock, M. J.; Rudnicky, A. I.: Speech for multimedia information retrieval. In *Proceedings of the 8th annual ACM symposium on User interface and software technology*, UIST '95, New York, NY, USA: ACM, 1995, ISBN 0-89791-709-X, s. 79–80.

[55] Hearst, M.: User interfaces and visualization. In *Modern information retrieval*, editace R. Baeza-Yates; B. Ribeiro-Neto, 1999, ISBN 0-201-39829-X, s. 1–75.

[56] Hori, T.; Hetherington, I. L.; Hazen, T. J.; aj.: Open-vocabulary spoken utterance retrieval using confusion neworks. In *Proceedings of ICASSP*, 2007, s. 73–76.

[57] Hui, K.; He, B.; Luo, T.; aj.: A Comparative Study of Pseudo Relevance Feedback for Ad-hoc Retrieval. In *Advances in Information Retrieval Theory: Third International Conference, ICTIR 2011, Bertinoro, Italy, September 12-14, 2011. Proceedings*, editace G. Amati; F. Crestani, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, ISBN 978-3-642-23318-0, s. 318–322.

[58] Ide, E.: New Experiments in Relevance Feedback. In *The SMART Retrieval System*, editace G. Salton, Englewood Cliffs, N.J.: Prentice Hall, s. 337–54.

[59] Ide, E. R. C.: *Relevance Feedback In An Automatic Document Retrieval System.* Master of science, Cornell University, 1969.

[60] Ircing, P.; Müller, L.: Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006. *Lecture Notes in Computer Science*, 2007: s. 759–765.

[61] Ircing, P.; Oard, D. W.; Hoidekr, J.: First experiments searching spontaneous Czech speech. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, New York, NY, USA: ACM, 2007, ISBN 978-1-59593-597-7, s. 835–836.

[62] Ircing, P.; Pecina, P.; Oard, D. W.; aj.: Information retrieval test collection for searching spontaneous Czech speech. In *Proceedings of the 10th international conference on Text, speech and dialogue*, TSD'07, Berlin, Heidelberg: Springer-Verlag, 2007, ISBN 3-540-74627-7, 978-3-540-74627-0, s. 439–446.

[63] Ircing, P.; Psutka, J.; Vavruška, J.: What Can and Cannot Be Found in Czech Spontaneous Speech Using Document-Oriented IR Methods – UWB at CLEF 2007 CL-SR Track. Berlin, Heidelberg: Springer-Verlag, 2008, ISBN 978-3-540-85759-4, s. 712–718.

[64] James, D.; Young, S.: A fast lattice-based approach to vocabulary independent wordspotting. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, ročník i, 1994, s. I/377 –I/380 vol.1.

[65] James, D. A.: *The application of classical information retrieval techniques to spoken documents*. Dizertační práce, University of Cambridge, 1995.

[66] Jelinek, F.; Mercer, R. L.: Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North-Holland, 1980, s. 381–397.

[67] Jones, G.; Foote, J.; Spärk Jones, K.; aj.: Robust talker-independent audio document retrieval. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, ročník 1, 1996, s. 311 –314 vol. 1.

[68] Jones, K. S.: A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, ročník 28, č. 1, 01 1972: s. 11–21, ISSN 0022-0418.

[69] Jones, K. S.; Walker, S.; Robertson, S. E.: A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. *Inf. Process. Manage.*, ročník 36, č. 6, Listopad 2000: s. 779–808, ISSN 0306-4573.

[70] Jourlin, P.; Johnson, S. E.; Jones, K. S.; aj.: Improving Retrieval on Imperfect Speech Transcriptions (Poster Abstract). In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, New York, NY, USA: ACM, 1999, ISBN 1-58113-096-1, s. 283–284.

[71] Jourlin, P.; Johnson, S. E.; Jones, K. S.; aj.: Spoken document representations for probabilistic retrieval. *Speech Commun.*, ročník 32, 2000: s. 21–36, ISSN 0167-6393.

[72] Jurafsky, D.; Martin, J. H.: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall PTR, první vydání, 2000, ISBN 0130950696.

[73] Kanis, J.; Müller, L.: Automatic Lemmatizer Construction with Focus on OOV Words Lemmatization. In *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, ročník 3658, editace V. Matoušek; P. Mautner; T. Pavelka, Springer Berlin / Heidelberg, 2005, ISBN 978-3-540-28789-6, s. 742–742.

[74] Kanis, J.; Skorkovská, L.: Comparison of different lemmatization approaches through the means of information retrieval performance. *Lecture Notes in Artificial Intelligence*, ročník 2010, 2010: s. 93–100, ISSN 0302-9743.

[75] Kelly, D.; Teevan, J.: Implicit Feedback for Inferring User Preference: A Bibliography. *SIGIR Forum*, ročník 37, č. 2, Září 2003: s. 18–28, ISSN 0163-5840.

[76] Khoo, C. S.; Poo, D. C.: An expert system approach to online catalog subject searching. *Information Processing and Management*, ročník 30, č. 2, 1994: s. 223 – 238, ISSN 0306-4573.

[77] Kwok, K. L.; Grunfeld, L.; Lewis, D. D.: TREC-3 Ad-Hoc, Routing Retrieval and Thresholding Experiments using PIRCS. In *In Proceedings of TREC'3*, Publication, 1995, s. 247–255.

[78] Lafferty, J.; Zhai, C.: Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, New York, NY, USA: ACM, 2001, ISBN 1-58113-331-6, s. 111–119.

[79] Lam-Adesina, A. M.; Jones, G. J. F.: Using String Comparison in Context for Improved Relevance Feedback in Different Text Media. In *String Processing and Information Retrieval: 13th International Conference, SPIRE 2006, Glasgow, UK, October 11-13, 2006. Proceedings*, editace F. Crestani; P. Ferragina; M. Sanderson, Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, ISBN 978-3-540-45775-6, s. 229–241.

[80] Larson, M.; Eskevich, M.; Ordelman, R.; aj.: Overview of MediaEval 2011 rich speech retrieval task and genre tagging task. CEUR Workshop Proceedings, 2011.

[81] Larson, M.; Jones, G. J.: Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval*, ročník 5, č. 4-5, 2012: s. 235–422.

[82] Lavrenko, V.; Croft, W. B.: Relevance Based Language Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, New York, NY, USA: ACM, 2001, ISBN 1-58113-331-6, s. 120–127.

[83] Lee, H. Y.; Lee, L. S.: Improved Semantic Retrieval of Spoken Content by Document/Query Expansion with Random Walk Over Acoustic Similarity Graphs. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, ročník 22, č. 1, Jan 2014: s. 80–94, ISSN 2329-9290.

[84] Lee, J. H.: Combining the evidence of different relevance feedback methods for information retrieval. *Information Processing and Management*, ročník 34, č. 6, 1998: s. 681–691, ISSN 0306-4573.

[85] Lee, K. S.; Croft, W. B.; Allan, J.: A Cluster-based Resampling Method for Pseudo-relevance Feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, New York, NY, USA: ACM, 2008, ISBN 978-1-60558-164-4, s. 235–242.

[86] Lee, L.-S.; Pan, Y.-C.: Voice-based information retrieval - how far are we from the text-based information retrieval? In *IEEE Workshop on Automatic Speech Recognition Understanding, 2009.*, 2009, s. 26 –43.

[87] Liu, B.; Oard, D. W.: One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, New York, NY, USA: ACM, 2006, ISBN 1-59593-369-7, s. 673–674.

[88] Logan, B.; Moreno, P.; Deshmukh, O.: Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, s. 31–35.

[89] Logan, B.; Van Thong, J.-M.; Moreno, P.: Approaches to reduce the effects of OOV queries on indexed spoken audio. *IEEE Transactions on Multimedia*, ročník 7, č. 5, 2005: s. 899 – 906, ISSN 1520-9210.

[90] Luhn, H. P.: A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, ročník 1, č. 4, Oct 1957: s. 309–317, ISSN 0018-8646.

[91] Lundquist, C.; Grossman, D. A.; Frieder, O.: Improving Relevance Feedback in the Vector Space Model. In *Proceedings of the Sixth International Conference on Information and Knowledge Management*, CIKM '97, New York, NY, USA: ACM, 1997, ISBN 0-89791-970-X, s. 16–23.

[92] Lv, Y.; Zhai, C.: A Comparative Study of Methods for Estimating Query Language Models with Pseudo Feedback. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, New York, NY, USA: ACM, 2009, ISBN 978-1-60558-512-3, s. 1895–1898.

[93] MacKay, D. J.; Peto, L. C. B.: A Hierarchical Dirichlet Language Model. *Natural Language Engineering*, ročník 1, 1995: s. 1–19.

[94] Madjarov, G.; Kocev, D.; Gjorgjevikj, D.; aj.: An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, ročník 45, č. 9, Září 2012: s. 3084–3104, ISSN 00313203.

[95] Mamou, J.; Carmel, D.; Hoory, R.: Spoken document retrieval from call-center conversations. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, New York, NY, USA: ACM, 2006, ISBN 1-59593-369-7, s. 51–58.

[96] Mangu, L.; Brill, E.; Stolcke, A.: Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, ročník 14, č. 4, 2000: s. 373 – 400, ISSN 0885-2308.

[97] Manning, C. D.; Raghavan, P.; Schütze, H.: *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008, ISBN 0521865719, 9780521865715.

[98] Maron, M. E.; Kuhns, J. L.: On Relevance, Probabilistic Indexing and Information Retrieval. *J. ACM*, ročník 7, č. 3, Červenec 1960: s. 216–244, ISSN 0004-5411.

[99] McCallum, A. K.: Multi-label text classification with a mixture model trained by EM. In *AAAI 99 Workshop on Text Learning*, 1999.

[100] Miller, D. R. H.; Leek, T.; Schwartz, R. M.: A Hidden Markov Model Information Retrieval System. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, New York, NY, USA: ACM, 1999, ISBN 1-58113-096-1, s. 214–221.

[101] Mitra, M.; Singhal, A.; Buckley, C.: Improving Automatic Query Expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, New York, NY, USA: ACM, 1998, ISBN 1-58113-015-5, s. 206–214.

[102] Mooers, C. N.: *Application of random codes to the gathering of statistical information.* Diplomová práce, Massachusetts Institute of Technology, 1948.

[103] Ng, K.: A Maximum Likelihood Ratio Information Retrieval Model. 1999.

[104] Ng, K.: *Subword-based approaches for spoken document retrieval.* Dizertační práce, Massachusetts Institute of Technology, 2000.

[105] Nishizaki, H.; Nakagawa, S.: Japanese spoken document retrieval considering OOV keywords using LVCSR system with OOV detection processing. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, s. 157–164.

[106] Oard, D. W.; Soergel, D.; Doermann, D.; aj.: Building an information retrieval test collection for spontaneous conversational speech. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, New York, NY, USA: ACM, 2004, ISBN 1-58113-881-4, s. 41–48.

[107] Oard, D. W.; Wang, J.; Jones, G. J. F.; aj.: Overview of the CLEF-2006 cross-Language speech retrieval track. In *Proceedings of the CLEF 2006: Workshop on Cross-Language Information Retrieval and Evaluation*, Springer, 2007.

[108] Pan, Y.-c.; Chang, H.-l.; Chen, B.; aj.: Subword-based position specific posterior lattices (s-PSPL) for indexing speech information. In *Proc. Interspeech*, 2007, s. 318–321.

[109] Parlak, S.; Saraclar, M.: Spoken information retrieval for turkish broadcast news. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, New York, NY, USA: ACM, 2009, ISBN 978-1-60558-483-6, s. 782–783.

[110] Parlak, S.; Saraclar, M.: Performance Analysis and Improvement of Turkish Broadcast News Retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, ročník 20, č. 3, March 2012: s. 731–741, ISSN 1558-7916.

[111] Pecina, P.; Hoffmannová, P.; Jones, G. J.; aj.: Overview of the CLEF-2007 Cross-Language Speech Retrieval Track. Berlin, Heidelberg: Springer-Verlag, 2008, ISBN 978-3-540-85759-4, s. 674–686.

[112] Ponte, J. M.: *A language modeling approach to information retrieval.* Dizertační práce, University of Massachusetts, Amherst, MA, USA, 1998.

[113] Ponte, J. M.; Croft, W. B.: A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, New York, NY, USA: ACM, 1998, ISBN 1-58113-015-5, s. 275–281.

[114] Ponte, J. M.; Croft, W. B.: A language modeling approach to information retrieval. In *Proceedings of SIGIR '98*, New York, NY, USA: ACM, 1998, ISBN 1-58113-015-5, s. 275–281.

[115] Psutka, J.; Ircing, P.; Psutka, J. V.; aj.: Large Vocabulary ASR for Spontaneous Czech in the MALACH Project. In *Proceedings of Eurospeech 2003*, Geneva, 2003, s. 1821–1824.

[116] Rabiner, L. R.: A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, 1989, s. 257–286.

[117] Raman, K.; Udupa, R.; Bhattacharya, P.; aj.: On Improving Pseudo-relevance Feedback Using Pseudo-irrelevant Documents. In *Proceedings of the 32Nd European Conference on Advances in Information Retrieval*, ECIR'2010, Berlin, Heidelberg: Springer-Verlag, 2010, ISBN 3-642-12274-4, 978-3-642-12274-3, s. 573–576.

[118] Renals, S.; Abberley, D.; Kirby, D.; aj.: Indexing and Retrieval of Broadcast News. *Speech Commun.*, ročník 32, č. 1-2, Září 2000: s. 5–20, ISSN 0167-6393.

[119] Reynolds, D. A.; Quatieri, T. F.; Dunn, R. B.: Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, ročník 10, č. 1, 2000: s. 19 – 41, ISSN 1051-2004.

[120] Rijsbergen, C. J. V.: *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann, druhé vydání, 1979, ISBN 0408709294.

[121] Robertson, S.: On Relevance Weight Estimation and Query Expansion. *Journal of Documentation*, ročník 42, č. 3, 03 1986: s. 182–188, ISSN 0022-0418.

[122] Robertson, S.; Jones, K.: Relevance weighting of search terms. *Journal of American Society of Information Science*, ročník 27, č. 3, may 1976: s. 129–146, ISSN 1097-4571.

[123] Robertson, S.; Walker, S.; Jones, S.; aj.: Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC–3)*, 1996, s. 109–126.

[124] Robertson, S. E.: On Term Selection for Query Expansion. *Journal of Documentation*, ročník 46, č. 4, Leden 1990: s. 359–364, ISSN 0022-0418.

[125] Robertson, S. E.; Sparck Jones, K.: *Relevance weighting of search terms*. London, UK, UK: Taylor Graham Publishing, 1988, ISBN 0-947568-21-2, s. 143–160.

[126] Robertson, S. E.; Walker, S.: Okapi/Keenbow at TREC-8. In *The Eighth Text REtrieval Conference (TREC 8)*, NIST, Gaithersburg, MD: NIST, 2000, s. 151–162.

[127] Rocchio, J. J.: Relevance Feedback in Information Retrieval. In The SMART Retrieval System—Experiments in Automatic Document Processing [130].

[128] Rosenberg, A. E.; DeLong, J.; Lee, C.-H.; aj.: The Use of Cohort Normalized Scores for Speaker Verification. In *Second International Conference on Spoken Language Processing*, 1992.

[129] Ruthven, I.; Lalmas, M.: A Survey on the Use of Relevance Feedback for Information Access Systems. *Knowl. Eng. Rev.*, ročník 18, č. 2, Červen 2003: s. 95–145, ISSN 0269-8889.

[130] Salton, G.: *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1971.

[131] Salton, G.: A blueprint for automatic Boolean query processing. *SIGIR Forum*, ročník 17, 1982: s. 6–24, ISSN 0163-5840.

[132] Salton, G.; Buckley, C.: Term-weighting approaches in automatic text retrieval. In *Information processing and management*, 1988, s. 513–523.

[133] Salton, G.; Buckley, C.: Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, ročník 41, 1990: s. 288–297.

[134] Salton, G.; Fox, E. A.; Wu, H.: Extended Boolean information retrieval. *Commun. ACM*, ročník 26, 1983: s. 1022–1036, ISSN 0001-0782.

[135] Salton, G.; Lesk, M. E.: Computer Evaluation of Indexing and Text Processing. *J. ACM*, ročník 15, 1968: s. 8–36, ISSN 0004-5411.

[136] Salton, G.; Wong, A.; Yang, C. S.: A vector space model for automatic indexing. *Commun. ACM*, ročník 18, 1975: s. 613–620, ISSN 0001-0782.

[137] Sanderson, M.; Shou, X. M.: Search of Spoken Documents Retrieves Well Recognized Transcripts. In *Advances in Information Retrieval: 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007. Proceedings*, editace G. Amati; C. Carpineto; G. Romano, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, ISBN 978-3-540-71496-5, s. 505–516.

[138] Saraclar, M.; Sproat, R.: Lattice-based search for spoken utterance retrieval. In *Proceedings of HLT-NAACL 2004*, 2004, s. 129–136.

[139] Schäuble, P.: *Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases*. Norwell, MA, USA: Kluwer Academic Publishers, 1997, ISBN 0792398998.

[140] Seide, F.; Yu, P.; Shi, Y.: Towards spoken-document retrieval for the enterprise: Approximate word-lattice indexing with text indexers. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2007.

[141] Shafran, I.; Byrne, W.: Task-Specific Minimum Bayes-Risk Decoding using Learned Edit Distance. In *Proceedings of ICSLP 2004, Jeju Island, South Korea*, 2004, s. 1945–1948.

[142] Shafran, I.; Hall, K.: Corrective models for speech recognition of inflected languages. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, ISBN 1-932432-73-6, s. 390–398.

[143] Shen, X.; Zhai, C.: Active Feedback in Ad Hoc Information Retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, New York, NY, USA: ACM, 2005, ISBN 1-59593-034-5, s. 59–66.

[144] Siegler, M. A.: *Integration of continuous speech recognition and information retrieval for mutually optimal performance*. Dizertační práce, Carnegie Mellon University, Pittsburgh, PA, USA, 1999.

[145] Singhal, A.; Mitra, M.; Buckley, C.: Learning Routing Queries in a Query Zone. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, New York, NY, USA: ACM, 1997, ISBN 0-89791-836-3, s. 25–32.

[146] Sivakumaran, P.; Fortuna, J.; Ariyaeeinia; aj.: Score normalisation applied to open-set, text-independent speaker identification. In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva*, Geneva, 2003, s. 2669–2672.

[147] Skorkovská, L.: Dynamic Threshold Selection Method for Multi-label Newspaper Topic Identification. In *Text, Speech, and Dialogue*, *Lecture Notes in Computer Science*, ročník 8082, editace I. Habernal; V. Matoušek, Springer Berlin Heidelberg, 2013, ISBN 978-3-642-40584-6, s. 209–216.

[148] Skorkovská, L.: First Experiments with Relevant Documents Selection for Blind Relevance Feedback in Spoken Document Retrieval. In *Speech and Computer*, *LNCS*, ročník 8773, Springer International Publishing, 2014, ISBN 978-3-319-11580-1, s. 235–242.

[149] Skorkovská, L.: Score Normalization Methods for Relevant Documents Selection for Blind Relevance Feedback in Speech Information Retrieval. In *Text, Speech, and Dialogue, TSD 2015, Proceedings*, LNCS, Cham: Springer International Publishing, 2015, ISBN 978-3-319-24033-6, s. 316–324.

[150] Skorkovská, L.: Comparison of Retrieval Approaches and Blind Relevance Feedback Methods Within the Czech Speech Information Retrieval. In *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings*, editace A. Ronzhin; R. Potapova; G. Németh, Cham: Springer International Publishing, 2016, ISBN 978-3-319-43958-7, s. 182–190.

[151] Skorkovská, L.: Relevant Documents Selection for Blind Relevance Feedback in Speech Information Retrieval. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno , Czech Republic, September 12-16, 2016, Proceedings*, editace P. Sojka; A. Horák; I. Kopeček; K. Pala, Cham: Springer International Publishing, 2016, ISBN 978-3-319-45510-5, s. 418–425.

[152] Skorkovská, L.; Ircing, P.: Experiments with Automatic Query Formulation in the Extended Boolean Model. In *Proceedings of the 12th International Conference on Text, Speech and Dialogue*, TSD '09, Berlin, Heidelberg: Springer-Verlag, 2009, ISBN 978-3-642-04207-2, s. 371–378.

[153] Skorkovská, L.; Ircing, P.; Pražák, A.; aj.: Automatic Topic Identification for Large Scale Language Modeling Data Filtering. In *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, ročník 6836, editace I. Habernal; V. Matoušek, Springer Berlin / Heidelberg, 2011, ISBN 978-3-642-23537-5, s. 64–71.

[154] Skorkovská, L.; Zajíc, Z.: Score Normalization Methods Applied to Topic Identification. In *TSD 2014*, *LNCS*, ročník 8655, Springer International Publishing, 2014, ISBN 978-3-319-10815-5, s. 133–140.

[155] Skorkovská, L.; Zajíc, Z.; Müller, L.: Comparison of score normalization methods applied to multi-label classification. In *2014 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Dec 2014, ISSN 2162-7843, s. 000433–000437.

[156] Song, F.; Croft, W. B.: A General Language Model for Information Retrieval. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, CIKM '99, New York, NY, USA: ACM, 1999, ISBN 1-58113-146-1, s. 316–321.

[157] Sparck Jones, K.: Further Reflections on TREC. *Inf. Process. Manage.*, ročník 36, č. 1, Leden 2000: s. 37–85, ISSN 0306-4573.

[158] Srikanth, M.; Srihari, R.: Biterm Language Models for Document Retrieval. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, New York, NY, USA: ACM, 2002, ISBN 1-58113-561-0, s. 425–426.

[159] Švec, J.; Lehečka, J.; Ircing, P.; aj.: General framework for mining, processing and storing large amounts of electronic texts for language modeling purposes. *Language Resources and Evaluation*, ročník 48, č. 2, 2014: s. 227–248, ISSN 1574-0218.

[160] Tao, T.; Zhai, C.: Regularized Estimation of Mixture Models for Robust Pseudo-relevance Feedback. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, New York, NY, USA: ACM, 2006, ISBN 1-59593-369-7, s. 162–169.

[161] Thompson, P.; Yang, B.; Flood, J.: TREC-3 Ad Hoc Retrieval and Routing Experiments using the WIN System. 1995.

[162] Tsoumakas, G.; Katakis, I.: Multi-label classification: An overview. *Int J Data Warehousing and Mining*, ročník 2007, 2007: s. 1–13.

[163] w. Tu, T.; y. Lee, H.; y. Chou, Y.; aj.: Semantic query expansion and context-based discriminative term modeling for spoken document retrieval. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, ISSN 1520-6149, s. 5085–5088.

[164] Tur, G.; Hakkani-Tur, D.; Riccardi, G.: Extending boosting for call classification using word confusion networks. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. (ICASSP '04).*, ročník 1, 2004, ISSN 1520-6149, s. I – 437–40 vol.1.

[165] Van Thong, J.-M.; Moreno, P.; Logan, B.; aj.: Speechbot: an experimental speech-based search engine for multimedia content on the web. *IEEE Transactions on Multimedia*, ročník 4, č. 1, 2002: s. 88 –96, ISSN 1520-9210.

[166] Švec, J.; Hoidekr, J.; Soutner, D.; aj.: Web Text Data Mining for Building Large Scale Language Modelling Corpus. In *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, ročník 6836, editace I. Habernal; V. Matoušek, Springer Berlin / Heidelberg, 2011, ISBN 978-3-642-23537-5, s. 356–363.

[167] Wang, D.; Frankel, J.; Tejedor, J.; aj.: A comparison of phone and grapheme-based spoken term detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, ISSN 1520-6149, s. 4969 –4972.

[168] Wang, Y.-Y.; Yu, D.; Ju, Y.-C.; aj.: An introduction to voice search. *Signal Processing Magazine, IEEE*, ročník 25, č. 3, 2008: s. 28 –38, ISSN 1053-5888.

[169] Wessel, F.; Schluter, R.; Macherey, K.; aj.: Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, ročník 9, č. 3, 2001: s. 288 –298, ISSN 1063-6676.

[170] White, R. W.; Oard, D. W.; Jones, G. J. F.; aj.: Overview of the CLEF-2005 cross-language speech retrieval track. In *Proc. CLEF 2005*, Springer, 2006, s. 744–759.

[171] Whittaker, S.; Hirschberg, J.; Amento, B.; aj.: SCANMail: a voicemail interface that makes speech browsable, readable and searchable. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, CHI '02, New York, NY, USA: ACM, 2002, ISBN 1-58113-453-3, s. 275–282.

[172] Wong, W. S.; Luk, R. W. P.; Leong, H. V.; aj.: Re-examining the Effects of Adding Relevance Information in a Relevance Feedback Environment. *Information Processing and Management*, ročník 44, č. 3, Květen 2008: s. 1086–1116, ISSN 0306-4573.

[173] Woodland, P.; Leggetter, C.; Odell, J.; aj.: The 1994 HTK large vocabulary speech recognition system. In *International Conference on Acoustics, Speech, and Signal Processing*, ročník 1, 1995, ISSN 1520-6149, s. 73 –76 vol.1.

[174] Woodland, P. C.; Johnson, S. E.; Jourlin, P.; aj.: Effects of out of vocabulary words in spoken document retrieval (poster session). In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, New York, NY, USA: ACM, 2000, ISBN 1-58113-226-3, s. 372–374.

[175] Wu, H.; Salton, G.: The Estimation of Term Relevance Weights Using Relevance Feedback. *Journal of Documentation*, ročník 37, č. 4, 04 1981: s. 194–214, ISSN 0022-0418.

[176] Xu, J.; Croft, W. B.: Query Expansion Using Local and Global Document Analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, New York, NY, USA: ACM, 1996, ISBN 0-89791-792-8, s. 4–11.

[177] Xu, Z.; Akella, R.; Zhang, Y.: Incorporating Diversity and Density in Active Learning for Relevance Feedback. In *Proceedings of the 29th European Conference on IR Research*, ECIR'07, Berlin, Heidelberg: Springer-Verlag, 2007, ISBN 978-3-540-71494-1, s. 246–257.

[178] Yaguchi, Y.; Watanabe, Y.; Naruse, K.; aj.: Speech and Song Search on the Web: System Design and Implementation. *Computer and Information Technology, International Conference on*, ročník 0, 2007: s. 270–275.

[179] Ye, Z.; He, B.; Huang, X.; aj.: Revisiting Rocchio's Relevance Feedback Algorithm for Probabilistic Models. In *Information Retrieval Technology: 6th Asia Information Retrieval Societies Conference, AIRS 2010, Taipei, Taiwan, December 1-3, 2010. Proceedings*, editace P.-J. Cheng; M.-Y. Kan; W. Lam; P. Nakov, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, ISBN 978-3-642-17187-1, s. 151–161.

[180] Ye, Z.; Huang, J. X.: A Simple Term Frequency Transformation Model for Effective Pseudo Relevance Feedback. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, New York, NY, USA: ACM, 2014, ISBN 978-1-4503-2257-7, s. 323–332.

[181] Yu, D.; cheng Ju, Y.; yi Wang, Y.; aj.: Automated Directory Assistance System - from Theory to Practice. In *Proceedings of INTERSPEECH*, 2007, str. 2709.

[182] Yu, P.; Seide, F.: A hybrid-word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech. In *Proc. INTERSPEECH 2004, Korea*, 2004, str. 293296.

[183] Zajíc, Z.; Machlica, L.; Padrta, A.; aj.: An Expert System in Speaker Verification Task. In *Proceedings of Interspeech*, ročník 9, Brisbane, AU: International Speech Communication Association, 2008, s. 355–358.

[184] Zhai, C.: *Statistical Language Models for Information Retrieval (Synthesis Lectures on Human Language Technologies)*. Morgan and Claypool Publishers, 2008, ISBN 159829590X.

[185] Zhai, C.; Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, New York, NY, USA: ACM, 2001, ISBN 1-58113-436-3, s. 403–410.

[186] Zhai, C.; Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, ročník 22, 2004: s. 179–214, ISSN 1046-8188.

[187] Zhou, Z.-Y.; Yu, P.; Chelba, C.; aj.: Towards spoken-document retrieval for the internet: lattice indexing for large-scale web-search architectures. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, s. 415–422.

[188] Zigel, Y.; Cohen, A.: On Cohort Selection for Speaker Verification. In *Proceedings of EUROSPEECH*, Geneva, 2003, s. 2977–2980.

Příloha A

Ukázky výpočtu testu statistické významnosti

Tabulka A.1: Ukázka výpočtu Wilcoxonova testu pro test vlivu odstranění stop slov a zájmen (stop sl. + záj.) proti použití všech slov pro model P-norm, trénovací sadu témat, dotaz vytvořený z polí **td**. Stanovíme si jednostrannou alternativní hypotézu, že výsledek při odstranění stop slov je horší než při použití všech slov. Součet pozitivních pořadí je roven 143,5 a součet negativních pořadí 87,5. Hodnota kriteria se tedy rovná $W = 87,5$. Podle tabulek je pro 21 vzorků (dalších 8 má stejné výsledky) kritická hodnota na hladině významnosti $\alpha = 0,05$ 67. Protože $W = 87,5$ není menší než 67, výsledek není statisticky významný na hladině významnosti $\alpha = 0,05$.

téma	všechna sl.	stop sl. + záj.	znam.	rozdíl	pořadí	znam. pořadí
1166	0,0001	0	1	0,0001	3	3
1181	0,1	0	1	0,1	21	21
1185	0,0019	0,0019	0	0	n/a	n/a
1187	0,002	0,002	0	0	n/a	n/a
1225	0,0658	0,0608	1	0,005	14	14
1286	0,0234	0,0369	-1	0,0135	19	-19
1288	0,0005	0,0005	0	0	n/a	n/a
1310	0,0062	0,0062	0	0	n/a	n/a
1311	0,0166	0,006	1	0,0106	18	18
1321	0,0017	0,0024	-1	0,0007	9	-9
1508	0,0358	0,0358	0	0	n/a	n/a
1620	0,0023	0,0023	0	0	n/a	n/a
1630	0,1728	0,1812	-1	0,0084	15	-15
1663	0,0165	0,0166	-1	0,0001	1,5	-1,5
1843	0,0276	0,0185	1	0,0091	16	16
2198	0,1366	0,0812	1	0,0554	20	20
2253	0,0045	0,0049	-1	0,0004	7,5	-7,5
3004	0,0073	0,0077	-1	0,0004	7,5	-7,5
3005	0,0326	0,0327	-1	0,0001	4	-4
3009	0,0059	0,0037	1	0,0022	11	11
3014	0,089	0,0795	1	0,0095	17	17
3015	0,0081	0,0055	1	0,0026	12	12
3017	0,0001	0,0001	0	0	n/a	n/a
3018	0,0003	0,0005	-1	0,0002	6	-6
3020	0,0212	0,0198	1	0,0014	10	10
3025	0,0135	0,0134	1	0,0001	1,5	1,5
3033	0,1092	0,1119	-1	0,0027	13	-13
4000	0,0004	0,0004	0	0	n/a	n/a
14312	0,0145	0,0147	-1	0,0002	5	-5

Tabulka A.2: Ukázka výpočtu Wilcoxonova testu pro test vlivu odstranění stop slov a zájmen (stop sl. + záj.) proti použití všech slov pro vektorový model TF-IDF, trénovací sadu témat, dotaz vytvořený z polí **td**. Stanovíme si jednostrannou alternativní hypotézu, že výsledek při odstranění stop slov je horší než při použití všech slov. Součet pozitivních pořadí je roven 196,5 a součet negativních pořadí 79,5. Hodnota kriteria se tedy rovná $W = 79,5$. Podle tabulek je pro 23 vzorků (dalších 6 má stejné výsledky) kritická hodnota na hladině významnosti $\alpha = 0,05$ 83. Protože $W = 79,5$ je menší než 83, výsledek je statisticky významný na hladině významnosti $\alpha = 0,05$.

téma	všechna sl.	stop sl. + záj.	znam.	rozdíl	pořadí	znam. pořadí
1166	0,0001	0	1	0,0001	3	3
1181	0,0377	0	1	0,0377	22	22
1185	0,0054	0,0056	-1	0,0002	5	-5
1187	0,0027	0,0027	0	0	n/a	n/a
1225	0,0419	0,0412	1	0,0007	11,5	11,5
1286	0,0403	0,0387	1	0,0016	15	15
1288	0,0033	0,0036	-1	0,0003	8	-8
1310	0,0105	0,0105	0	0	n/a	n/a
1311	0,0172	0,0125	1	0,0047	19	19
1321	0,0084	0,0079	1	0,0005	10	10
1508	0,0398	0,0398	0	0	n/a	n/a
1620	0,0009	0,0009	0	0	n/a	n/a
1630	0,2076	0,2025	1	0,0051	20	20
1663	0,0165	0,0166	-1	0,0001	2	-2
1843	0,024	0,0183	1	0,0057	21	21
2198	0,1098	0,0712	1	0,0386	23	23
2253	0,0172	0,017	1	0,0002	4	4
3004	0,0108	0,0112	-1	0,0004	9	-9
3005	0,0475	0,0486	-1	0,0011	14	-14
3009	0,0321	0,032	1	0,0001	1	1
3014	0,1041	0,105	-1	0,0009	13	-13
3015	0,0138	0,0108	1	0,003	18	18
3017	0,0001	0,0001	0	0	n/a	n/a
3018	0,0012	0,001	1	0,0002	6	6
3020	0,0279	0,0308	-1	0,0029	17	-17
3025	0,0662	0,0659	1	0,0003	7	7
3033	0,1106	0,1078	1	0,0028	16	16
4000	0,0004	0,0004	0	0	n/a	n/a
14312	0,0511	0,0518	-1	0,0007	11,5	-11,5

Příloha B

Tabulky výsledků

Tabulka B.1: Porovnání metod pro vytváření booleovského dotazu na trénovací sadě témat (mGAP)

p	manuální d.	vzrůstání stromu	jen OR	jen AND	Saltonova m.
1	0,0199	0,0235	0,017	0,017	0,0219
2	0,0271	0,025	0,0177	0,0178	0,0207
4	0,0317	0,0264	0,0192	0,0178	0,0208
5	0,0304	0,0299	0,0194	0,0177	0,0211
6	0,0305	0,0276	0,0187	0,0176	0,0212
10	0,0308	0,0238	0,0196	0,0179	0,0211
15	0,0293	0,0239	0,021	0,017	0,0205
25	0,0292	0,0241	0,0222	0,0133	0,0215

Tabulka B.2: Porovnání metod pro vytváření booleovského dotazu na testovací sadě témat (mGAP)

p	manuální d.	vzrůstání stromu	jen OR	Saltonova m.
1	0,0106	0,0133	0,0073	0,0049
2	0,0102	0,0135	0,0068	0,0052
4	0,0103	0,0136	0,0072	0,0058
5	0,0097	0,0135	0,0073	0,0058
6	0,0095	0,0135	0,0078	0,0058
10	0,0094	0,013	0,0075	0,0058

Tabulka B.3: Porovnání výsledků nastavení počtu dokumentů k a počtu termů m ve slepé zpětné vazbě ve vektorovém modelu pro trénovací sadu **tdn**. Bez zpětné vazby dosahuje vektorový model mGAP 0,0456. Tučně jsou vyznačeny výsledky s mGAP větším než 0,0530. (mGAP)

tdn	**0,0456**							
k / m	5	10	15	20	25	30	35	40
5	0,0505	0,0518	0,0503	0,0502	0,0489	0,0494	0,0483	0,0472
10	0,0506	0,0494	0,0490	0,0507	0,0485	0,0488	0,0507	0,0507
20	0,0515	0,0521	**0,0532**	**0,0534**	**0,0544**	**0,0531**	0,0525	0,0529
30	0,0490	0,0506	0,0491	0,0517	**0,0531**	**0,0550**	**0,0541**	**0,0537**
40	0,0524	0,0525	**0,0532**	**0,0538**	0,0518	0,0513	0,0515	0,0521
50	**0,0536**	**0,0541**	**0,0546**	0,0527	0,0510	0,0514	0,0510	0,0522
60	**0,0551**	**0,0530**	0,0528	0,0515	0,0507	0,0513	0,0500	0,0496
70	**0,0553**	**0,0564**	0,0518	0,0525	0,0528	0,0511	0,0511	0,0511
80	**0,0545**	**0,0562**	**0,0562**	0,0512	0,0513	0,0520	0,0512	0,0517
90	**0,0545**	**0,0555**	**0,0547**	**0,0543**	**0,0555**	0,0520	0,0519	0,0521
100	**0,0530**	**0,0563**	**0,0562**	**0,0548**	**0,0560**	**0,0538**	**0,0532**	**0,0546**
110	0.0520	**0.0542**	**0.0532**	**0.0553**	**0.0549**	**0.0551**	**0.0549**	**0.0541**
120	0.0519	0.0518	**0.0544**	**0.0544**	**0.0544**	**0.0527**	**0.0532**	**0.0530**
150	0,0496	0,0484	0,0484	0,0490	0,0499	0,0498	0,0513	0,0513
200	0,0490	0,0498	0,0480	0,0489	0,0500	0,0495	0,0485	0,0477
300	0,0499	0,0482	0,0477	0,0467	0,0453	0,0471	0,0473	0,0464
500	0,0482	0,0470	0,0460	0,0461	0,0448	0,0449	0,0456	0,0468

Tabulka B.4: Porovnání výsledků nastavení počtu dokumentů k a počtu termů m ve slepé zpětné vazbě ve Query likelihood modelu s Jelinek-Mercer vyhlazováním pro trénovací sadu **tdn**. Bez zpětné vazby dosahuje Query likelihood model mGAP 0,0392. Tučně jsou vyznačeny výsledky s mGAP větším než 0,0480. (mGAP)

tdn	**0,0392**							
k / m	5	10	15	20	25	30	35	40
5	0,0426	0,0448	0,0449	0,0451	0,0448	0,0440	0,0444	0,0430
10	0,0436	0,0437	0,0449	0,0452	0,0450	0,0453	0,0459	0,0466
20	0,0432	**0,0491**	**0,0502**	**0,0513**	**0,0496**	**0,0513**	**0,0494**	0,0473
30	0,0438	0,0463	0,0476	0,0474	**0,0504**	**0,0493**	**0,0506**	**0,0498**
40	0,0436	0,0461	0,0462	0,0477	0,0477	0,0475	**0,0489**	**0,0497**
50	0,0449	0,0475	0,0476	**0,0488**	**0,0489**	**0,0482**	**0,0494**	0,0479
100	0,0401	0,0460	0,0462	0,0477	**0,0484**	**0,0481**	**0,0486**	**0,0503**

Tabulka B.5: Porovnání výsledků nastavení počtu dokumentů k a počtu termů m ve slepé zpětné vazbě ve Query likelihood modelu s dvoufázovým vyhlazováním pro trénovací sadu **tdn**. Bez zpětné vazby dosahuje Query likelihood model s dvoufázovým vyhlazováním mGAP 0,0445. Tučně jsou vyznačeny výsledky s mGAP větším než 0,0540. (mGAP)

tdn	**0,0445**							
k / m	5	10	15	20	25	30	35	40
5	0,0469	0,0473	0,0486	0,0476	0,0477	0,0465	0,0451	0,0450
10	0,0516	0,0508	0,0517	0,0513	0,0503	0,0497	0,0499	0,0500
20	0,0501	0,0536	0,0498	0,0524	0,0512	0,0534	0,0528	0,0511
30	0,0477	0,0513	0,0505	0,0509	0,0511	0,0517	0,0524	0,0536
40	0,0488	0,0513	**0,0541**	**0,0544**	0,0530	**0,0547**	**0,0567**	**0,0561**
50	0,0497	0,0508	0,0515	0,0539	**0,0556**	**0,0559**	**0,0558**	**0,0557**
100	0,0461	0,0497	0,0522	**0,0553**	**0,0557**	**0,0574**	**0,0566**	**0,0565**

Tabulka B.6: Porovnání výsledků nastavení počtu dokumentů k a počtu termů m ve slepé zpětné vazbě ve Query likelihood modelu s Dirichlet vyhlazováním pro trénovací sadu **tdn**. Bez zpětné vazby dosahuje Query likelihood model mGAP 0,0414. Tučně jsou vyznačeny výsledky s mGAP větším než 0,0510. (mGAP)

tdn	**0,0414**							
k / m	5	10	15	20	25	30	35	40
5	0,0456	0,0444	0,0461	0,0456	0,0467	0,0446	0,0427	0,0422
10	0,0509	0,0504	0,0494	0,0490	0,0488	0,0506	0,0481	0,0473
20	0,0459	**0,0517**	0,0485	0,0500	**0,0520**	**0,0511**	0,0501	0,0481
30	0,0449	0,0497	**0,0513**	**0,0544**	**0,0521**	**0,0515**	**0,0517**	**0,0519**
40	0,0465	0,0484	0,0505	0,0500	0,0503	**0,0518**	**0,0520**	0,0506
50	0,0461	0,0496	0,0489	0,0489	0,0504	0,0507	**0,0515**	**0,0522**
100	0,0438	0,0469	0,0467	0,0497	**0,0516**	**0,0527**	**0,0523**	0,0508

Seznam publikovaných prací

1. Skorkovská, L.; Ircing, P.: Experiments with Automatic Query Formulation in the Extended Boolean Model. In *Proceedings of the 12th International Conference on Text, Speech and Dialogue*, TSD '09, Berlin, Heidelberg: Springer-Verlag, 2009, ISBN 978-3-642-04207-2, s. 371–378.

2. Kanis, J.; Skorkovská, L.: Comparison of different lemmatization approaches through the means of information retrieval performance. *Lecture Notes in Artificial Intelligence*, ročník 2010, 2010: s. 93–100, ISSN 0302-9743.

3. Skorkovská, L., Ircing, P., Pražák, A., Lehečka, J.: Automatic Topic Identification for Large Scale Language Modeling Data Filtering. *Lecture Notes in Computer Science*, 2011, roč. 2011, č. 6836, s. 64-71. ISSN 0302-9743.

4. Skorkovská, L.: Application of Lemmatization and Summarization Methods in Topic Identification Module for Large Scale Language Modeling Data Filtering. *Lecture Notes in Computer Science*, 2012, roč. 7499, s. 191-198. ISSN 0302-9743.

5. Skorkovská, L.: Dynamic Threshold Selection Method for Multi-label Newspaper Topic Identification. In *Text, Speech and Dialogue*. Lecture Notes in Computer Science. Heidelberg: Springer, 2013. s. 209-216. ISBN: 978-3-642-40584-6 , ISSN 0302-9743.

6. Švec, J., Lehečka, J., Ircing, P., Skorkovská, L., Pražák, A., Vavruška, J., Stanislav, P., Hoidekr, J.: General framework for mining, processing and storing large amounts of electronic texts for language modeling purposes. *Language Resources and Evaluation*, 2014, roč. 48, č. 2, s. 227-248. ISSN 1574-020X.

7. Skorkovská, L., Zajíc, Z., Müller, L.: Comparison of Score Normalization Methods Applied to Multi-label Classification. In *IEEE International Symposium on Signal Processing and Information Technology*. Institute of Electrical and Electronics Engineers (IEEE), Noida, India. 2014.

8. Skorkovská, L.: First Experiments with Relevant Documents Selection for Blind Relevance Feedback in Spoken Document Retrieval. In *Speech and Computer, 16th International Conference, SPECOM 2014, Novi Sad, Serbia, October 5-9, 2014, Proceedings*. Heidelberg: Springer, 2014. s. 235-242. ISBN 978-3-319-11580-1, ISSN 0302-9743.

9. Skorkovská, L., Zajíc, Z.: Score Normalization Methods Applied to Topic Identification. In *Text, Speech, and Dialogue, 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings*. Heidelberg: Springer, 2014. s. 133-140. ISBN 978-3-319-10815-5 , ISSN 0302-9743.

10. Skorkovská, L.: Score Normalization Methods for Relevant Documents Selection for Blind Relevance Feedback in Speech Information Retrieval. In *Text, Speech, and Dialogue, 18th International Conference, TSD 2015, Plzeň, Czech Republic. Proceedings*. Heidelberg: Springer, 2015. s. 353-361. ISSN 0302-9743.

11. Skorkovská, L.: Relevant Documents Selection for Blind Relevance Feedback in Speech Information Retrieval. In *Text, Speech, and Dialogue, 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*. Springer International Publishing, 2016. s. 418-425. ISSN 0302-9743.

12. Skorkovská, L.: Comparison of Retrieval Approaches and Blind Relevance Feedback Methods Within the Czech Speech Information Retrieval. In *Speech and Computer, 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings*. Springer International Publishing, 2016. s. 182-190. ISSN 0302-9743.

Další práce

1. Skorkovská, L.: Booleovský model pro vyhledávání informací v textu. Bakalářská práce, Západočeská Univerzita v Plzni, 2006.

2. Skorkovská, L.: Srovnání Booleovského a rozšířeného Booleovského modelu. Diplomová práce, Západočeská Univerzita v Plzni, 2008.

3. Skorkovská, L.: First Experiments with Automatic Topic Identification of Czech Newspaper Articles. In *SVK 2010 - magisterské a doktorské studijní programy, sborník rozšířených abstraktů*. Plzeň: Západočeská univerzita v Plzni, 2010. s. 49-50. ISBN 978-80-7043-903-6.

4. Skorkovská, L.: JMZW: Topic Identification in Czech Newspaper Articles. In *SVK 2011 - magisterské a doktorské studijní programy, sborník rozšířených abstraktů*. Plzeň: Západočeská univerzita v Plzni, 2011. s. 95-96. ISBN 978-80-261-0000-3.

5. Skorkovská, L.: Vyhledávání informací v řeči. Odborná práce ke státní doktorské zkoušce. Plzeň : 2011.

6. Skorkovská, L.: JMZW: Application of Summarization Methods in Topic Identification Module for Large Scale Language Modeling Data Filtering. In *SVK 2012 - magisterské a doktorské studijní programy, sborník rozšířených abstraktů*. Plzeň: Západočeská univerzita v Plzni, 2012. s. 91-93. ISBN 978-80-261-0127-7.

7. Skorkovská, L.: Multi-label Classification of Newspaper Articles. In *SVK 2013 - magisterské a doktorské studijní programy, sborník rozšířených abstraktů*. Plzeň: Západočeská univerzita v Plzni, 2013. s. 83-84. ISBN 978-80-261-0238-0.

8. Skorkovská, L.: The Use of the Unconstrained Cohort Normalization Technique for Multi-label Classification Score Normalization. In *SVK 2014 - magisterské a doktorské studijní programy, sborník rozšířených abstraktů*. Plzeň: Západočeská univerzita v Plzni, 2014. s. 99-100. ISBN 978-80-261-0365-3.