SEARCH ENGINE FOR ACCESS TO INFORMATION FROM SPEECH RECOGNITION

Michal FAPŠO, Bachelor Degree Programme (3)
Dept. of Computer Graphics and Multimedia, FIT, BUT
E-mail: xfapso00@stud.fit.vutbr.cz

Supervised by: Ing. Petr Schwarz

ABSTRACT

This paper describes a system for speech data indexing and searching. Nowadays, search engines are searching in text data, such as web pages, but how can we search in speech data? We have to search in an output from speech recognizer, which is an acyclic graph of hypotheses. So it is not such a simple structre as text documents are. To search effectively in such data and to browse these graphs as fast as possible, we have to create an intelligent indexing system which is the skeleton of each fast search engine.

1 ÚVOD

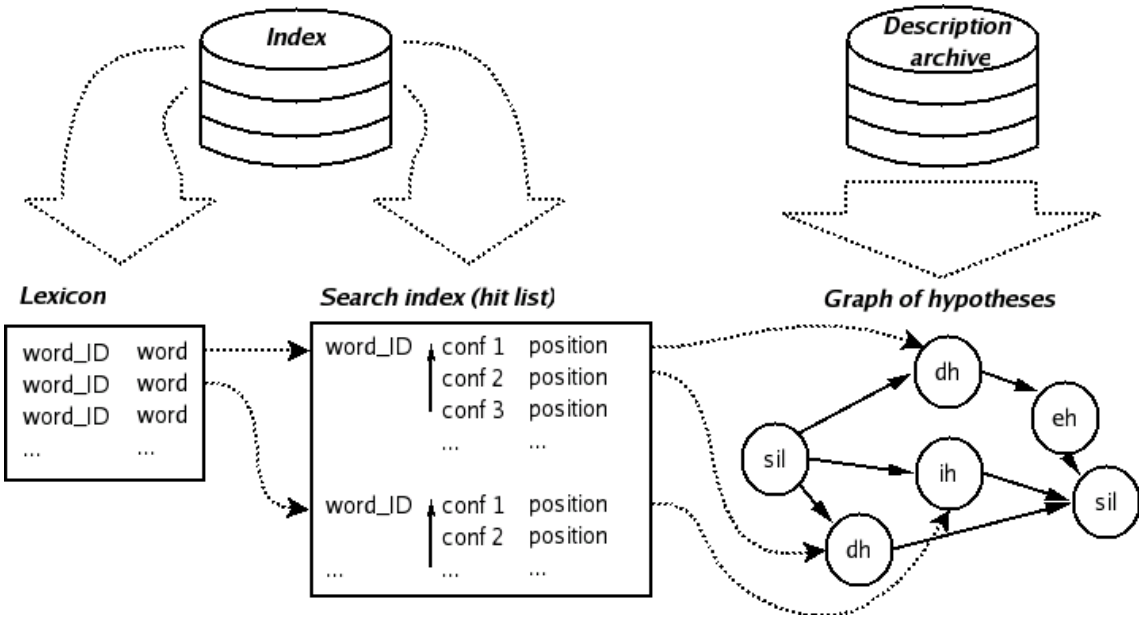
Keď hľadáme na internete nejakú informáciu, použijeme systém typu Google, ktorý prehľadáva textové dáta. Čo ale môžeme urobiť v prípade, že hľadáme informáciu, ktorú sme počuli niekde v rádiu, alebo na pracovnom mítingu pred niekoľkými týždňami? Mnoho záznamov z rádii či televízie je voľne dostupných na internete, ale pokiaľ nie je k dispozícii ich textový prepis, len veľmi ťažko v nich nájdeme konkrétnu informáciu.

Výstupom rozpoznávača reči je rozsiahla sieť hypotéz, v ktorej nie je možné vyhľadávať tak ako v textových dátach.

Hlavné parametre sledované pri vývoji systému pre vyhľadávanie v rečových dátach sú rýchlosť, pamäťové nároky a potrebný diskový priestor. V nasledujúcich odstavcoch nájdete popis a riešenie tohoto problému i s výsledkami experimentov.

2 ARCHITEKTÚRA SYSTÉMU

Systém vyvýjaný v Skupine spracovania reči na FIT v Brne bude slúžiť na vyhľadávanie informácii v rečových záznamoch. Na vstupe do systému prichádzajú akustické dáta, ktoré systém spracuje a umožní koncovému používateľovi ich prehľadávanie.



Obrázek 1: Dátové štruktúry previazané pomocou indexov

3 DÁTOVÉ ŠTRUKTÚRY

Rýchlosť vyhľadávacieho systému závisí na efektívnom indexovaní, pretože dáta v ktorých sa vyhľadáva, môžu zaberať rádovo gigabajty pamäťového priestoru.

Všetky indexované grafy hypotéz sú uložené v *indexe dokumentov* [1].

Jednotlivé *grafy hypotéz* sa prevedú z textového formátu do binárneho, ktorý je optimalizovaný pre rýchle vyhľadávanie. Vzhľadom na nutnosť prechádzania grafu pri zisťovaní najpravdepodobnejšieho kontextu nájdených slov, obsahuje každý uzol grafu odkazy na svojich predchodcov a nasledovníkov zoradené podľa pravdepodobnosti prechodu. Keďže každý uzol môže mať iný počet prechodov, táto štruktúra nemá pevnú veľkosť. Z toho dôvodu je nutné vytvoriť tabuľku indexov, v ktorej je ku každému uzlu priradená jeho pozícia v binárnom súbore. Všetky ostatné informácie o uzloch (počiatočný a koncový čas, reťazec, pravdepodobnosť, ...) sa nachádzajú v inom binárnom súbore, ktorý je možné načítať celý do pamäte, a keďže táto štruktúra má pevnú veľkosť, nie je preň potrebné vytvárať tabuľku indexov.

Pre všetky slová nachádzajúce sa v grafe hypotéz, sú vytvorené číselné identifikátory zapísané v *slovníku* pre jednoduchšie porovnávanie a šetrenie pamäťou [1]. Ku každému slovu je tiež zapísaná jeho pozícia vo *vyhľadávacom indexe*, v ktorom sú záznamy o hypotézach všetkých indexovaných grafov.

Záznamy v tomto indexe sú zoradené podľa viacerých položiek: podľa slova, pravdepodobnosti a identifikátoru príslušného grafu (spracovaného zvukového záznamu). To znamená, že ak hľadáme nejaké slovo, zistíme zo slovníka, kde sa toto slovo nachádza vo vyhľadávacom indexe a na tomto mieste sú záznamy o hypotézach daného slova zoradené podľa pravdepodobnosti, takže na začiatku zoznamu sú najpravdepodobnejšie výskyty [1]. Okrem toho obsahuje záznam o každej hypotéze aj jej pozíciu v binárnom grafe hypotéz.

Pre každý graf hypotéz je vytvorený *časový index*, ktorý obsahuje pozície prvých záznamov vo vybraných časových úsekoch. Vďaka tomu môžeme do pamäte načítať iba relevantnú oblasť grafu hypotéz okolo nájdeného slova.

4 PRIEBEH VYHĽADÁVANIA

Zo slovníka sa zistí identifikátor hľadaného slova, pomocou ktorého sa z vyhľadávacieho indexového súboru načítajú informácie o všetkých výskytoch daného slova v indexovaných dátach. Vyberú sa slová s najlepšou *confidence*, pre ktoré je potrebné zistiť kontext. Načítanie celého grafu hypotéz by ale bolo neefektívne, preto sa ”vysekne” iba malá časť grafu okolo kľúčového slova, ktorá je načítaná do pamäte. Vyhľadanie najlepšej cesty prechádzajúcej nájdeným slovom (kontextu) prebieha tak, že sa oboma smermi od nájdeného slova prechádza grafová štruktúra po najlepšie ohodnotenej ceste.

5 ZÁVER

Systém nájde hľadané slovo veľmi rýchlo (rádovo v jednotkách sekúnd) i napriek značnému objemu dát (rádovo v stovkách megabajtoch až jednotkách gigabajtov). Pre míting obsahujúci 500,232 hypotéz a 5,501,524 prepojení medzi hypotézami, trvá vyhľadávanie aj s vypísaním kontextu 15 nájdených slov cca. 2 sekundy. Využiteľnosť takéhoto systému je bez pochýb značne široká. Od vyhľadávania v rádiovom či televíznom vysielaní až po konferencie a mítingy.

Testovacia prevádzka tohoto systému bude slúžiť na vyhľadávanie v audio záznamoch z prednášok na FIT a v európskom projekte AMI, ktorý sa zameriava na spracovanie mítingov [3].

POĎAKOVANIE

Tento príspevok vznikol za podpory EC projektu Posilnenej skupinovej interakcie (AMI) č. 506811 a grantu GAČR 102/05/0278.

REFERENCE

[1] Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Science Department, Stanford University, Stanford, 2000

[2] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book, Cambridge University Engineering Department, 2002

[3] Augmented Multi-party Interaction, www.amiproject.org