BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2016
http://excel.fit.vutbr.cz

Automatická detekce témat, segmentace a vizualizace on-line kurzů

Josef Řídký*

SuperLectures.net

Abstrakt
Tento příspěvek popisuje moji diplomovou práci, která se zabývá tématem automatické detekce témat, segmentace a vizualizace on-line kurzů. Cílem této práce je vytvořit webovou aplikaci, která dokáže, za pomoci podpůrných skriptů, automaticky detekovat a segmentovat témata z videozáznamů on-line kurzů a následně je vizualizovat. Aplikace umožňuje takovéto zpracování pouze u videozáznamů z kurzů přednášených v anglickém jazyce. Úloha se dá rozdělit na dvě části: skripty pro zpracování textu a webovou prezentaci. Tím nejdůležitějším je právě zpracování textových přepisů pomocí skriptů, které zajišťují rozdělení jednoho záznamu na jednotlivá témata, která se v záznamu objevovala, jejich následné srovnání se všemi ostatními detekovanými tématy z ostatních záznamů za využití Subspace multinomial model a výpočtu Euklidovské vzdálenosti a následné uložení výsledku do databáze. Webová prezentace je zde pouze pro zajištění uživatelsky přívětivé vizualizace výsledků a správu jednotlivých záznamů z on-line kurzů v systému. Momentálně probíhá testování celé aplikace, a proto nejsou zatím dostupné žádné relevantní výsledky. Cílem této práce je tedy nabídnout uživatelům webové aplikace při přehrávání záznamu o on-line kurzu další, tématicky podobné záznamy z jiných kurzů a to ne na základě manuálně dodaných parametrů (jako kategorie, klíčová slova apod.), ale automaticky za využití přepisu audiozáznamu do textové podoby.

Klíčová slova: Detekce témat — Automatická detekce témat — on-line kurzy — Speech@FIT — YouTube

Přiložené materiály: Internetová prezentace — http://superlectures.net

*xridky00@stud.fit.vutbr.cz, *Fakulta informačních technologií, Vysoké učení technické v Brně*

1. Úvod

Dopátrat se textových informací na požadované téma z různých internetových zdrojů je v dnešní době vcelku bezproblémové. Problém však nastává v případě, kdy chceme na základě obsahového sdělení z videozáznamu nalézt další videozáznamy, ve kterých se hovoří na stejné či podobné téma. V takovém případě nám nezbývá než se probírat názvy a anotacemi jednotlivých videí a vyhledávat podobné záznamy. Právě tuto nepraktickou činnost chci za pomoci svojí diplomové práce co nejvíce eliminovat.

Tato práce se zabývá automatickým vyhledáváním tématicky podobných částí ve videozáznamech z on-line kurzů, které jsou prezentovány v anglickém jazyce. Z videozáznamů jsou získány přepisy, které jsou následně zpracovávány. Tyto přepisy jsou rozděleny podle témat a následně porovnány s ostatními stejně rozdělenými přepisy. Výsledek porovnání je uložen

do databáze. Výsledná aplikace nabízí uživatelům při přehrávání jednotlivých videozáznamů další videozáznamy, které jsou obsahově podobné vzhledem k jednotlivým částem přehrávaného záznamu.

Přímo takto definovanému problému se žádná práce nevěnuje, ovšem práce, které se zabývají automatickou detekcí témat se v určitých fázích s touto prací prolínají [1] [2]. Má diplomová práce je postavena na výsledcích bakalářské práce pana Martina Sychry, která se věnovala tématu automatického hledání vazeb mezi částmi audiovizuálních dokumentů [3]. V roce 2015 se se svou prací účastnil konference Excel@FIT [4].

Z větší části využívá výsledků a postupů popsaných právě v bakalářské práci pana Sychry. Druhým zdrojem je práce pana Santoshe Kesirajua, který se ve své vědecké činnosti mimo jiné zabývá problémem redukce dimenzionality vektorů za pomocí Subspace multinomial model[1]. Při tvorbě své práce vycházel z [5].

První část v mé práci slouží k tomu, abych byl schopný z textového přepisu audiovizuálního záznamu získat jednotlivé segmenty textu, které se v daném záznamu věnují jednomu tématu. Druhá část je používána při hledání podobnosti mezi nalezenými segmenty. Mým úkolem je tedy zabalit tato dvě řešení do jednoho plně automatického procesu.

Snahou této práce je co nejvíce zjednodušit vyhledávání tématicky podobných videozáznamů a to na základě jejich skutečného obsahu.

2. Detekce témat

Problémem detekce témat z textu se zabývá ve své bakalářské práci pan Martin Sychra [3]. V mé práci je tato část převzata právě od něj. Z toho důvodu zde uvádím pouze přehledové shrnutí problému detekce a segmentace témat. Zjednodušený výklad lze nalézt i v jeho příspěvku do konference Excel@FIT v roce 2015 [4].

Výzkum hledání tématu[2] začal velmi expandovat kolem roku 1997. V tomto období vznikl program Transligual Information Detection, Extraction and Summarization [3], který se podílel velkou měrou na úspěchu v tomto odvětví. Iniciativa TIDES[4] rozdělila TDT problém do pěti kategorií:

- **Story segmentation** - Jedná se o první část TDT úkolu. Cílem je rozdělit delší text na menší celky, které obsahují právě jedno hlavní téma.
- **First story detection** - Druhá část postupně prochází výsledky segmentace a zjišťuje, které textové segmenty (články) pojednávají o tématu, které se v předchozích segmenteh nevyskytovalo. V této fázi se tedy detekují nová témata.
- **Topic tracking** - Ve třetí části se na základě dostatečného počtu článků, které byly přiřazeny k danému tématu, vyhledávají další články, které spadají do stejného tématu. Toto přiřazování by se mělo dít nezávisle na tématu, tedy bez využití informace o přiřazení daného článku k jiným tématům.
- **Topic detection** - V této části dochází ke shlukování jednotlivých článků dle odhadnutých témat. Jednotlivé shluky budou reprezentovat různá témata.
- **Link detection** - Tato část má za cíle určit, zda jsou dva články tématicky stejné.

Z výše uvedeného se dá odvodit, že pro moji práci jsou nejpodstatnější části *Story segmentation*, *First story detection* a *Link detection*. Pro účely této práce není závadou, že se nebude zjišťovat přesné téma dané části přepsaného textu. Stačí totiž k dané části textu nalézt další části z jiných přepisů, které jsou jí nejvíce podobné. Na základě této podobnosti pak bude uživatelům navrhováno shlédnutí dalších videozáznamů.

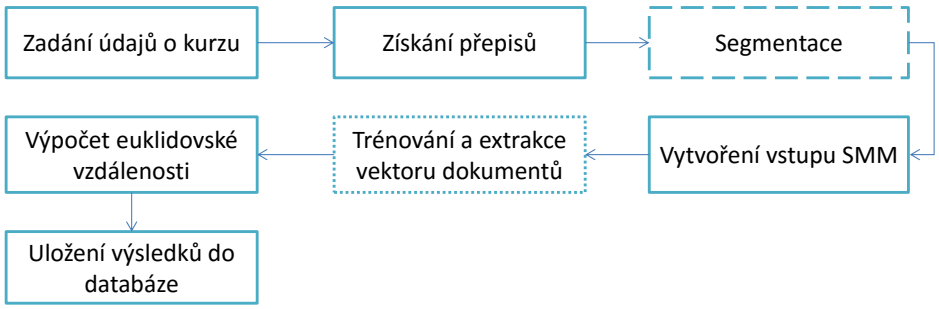
2.1 Proces segmentace a detekce témat

Na obrázku 1 je uvedeno grafické znázornění procesu zpracování jednotlivých videozáznamů. Blok, který je vyznačen čárkovaným okrajem znázorňuje zpracování dané části procesu skriptem, který byl vytvořen panem Marinem Sychrou. Blok s tečkovaným ohraničením znázorňuje skript, jehož autorem je Santosh Kesiraju. Ostatní části jsem implementoval tak, aby bylo výše uvedené skripty možné využít.

2.2 Získávání dat

Při vzniku zadání této práce, bylo nutné určit, jaké typy videozáznamů mají být ve výsledné aplikaci k dispozici. Prvotním požadavkem bylo vytvořit aplikaci, která by sdružovala relevantní, důvěryhodné a volně dostupné videozáznamy na témata z oblasti informačních technologií. Asi každého by hned napadlo posbírat prakticky libovolné videozáznamy z internetu, pojednávající na téma informačních technologií. Mezi nejčastější typy záznamů by se tak řadily tzv. videoblogy, tutoriály a další obdobná, relativně krátká, videa. Problémů je zde však hned několik. Prvním je relevantnost informací v těchto záznamech, druhým problémem je délka samotného záznamu a v neposlední

[1] dále jen SMM
[2] Topic Detection and Tracking - dále jen TDT
[3] dále jen TIDES
[4] Translingual Information Detection, Extraction, and Summarization



Obrázek 1. Diagram znázorňující proces detekce a segmentace témat

řadě je důležitý zdroj, odkud jsou jednotlivé videozáznamy přejímány. Existuje několik hlavních služeb, ze kterých je možné videozáznamy převzít, avšak jediným zdrojem, který plně vyhovuje všem zadaným kritériím je server YouTube.com. Jako jediný obsahuje volně dostupná videa, která jesou vícen než 40 minut dlouhá a ve svém obsahu pojednávají o více než jednom tématu.

Pro získávání nových záznamů z přednášek a jejich následné zpracování slouží stránkna webové prezentace dostupná pomocí odkazu v úvodu tohoto příspěvku. V této chvíli má spíše jen demonstrační charakter, neboť zápis do databáze není z bezpečnostních důvodů umožněn.

Při získávání záznamů je požadováno zadání následujících informací:

- název přednášky
- abstrakt
- URL adresa k videozáznamu na YouTube
- jméno přednášejícího
- místo působení přednášejícího
- kategorie, do které přednáška tématicky spadá

Videozáznamy lze manuálně zařadit do jedné z následujících kategorií: Computational Neuroscience, Computer Graphics, Computer Networks, Computer Science, Computer System Security, Computer Systems, Databases, Hardware, Intelligent Systems, Math, Mobile and Web, Program Languages, Programming, Signal Processing.

Pro uchování získaných a následně i vypočítaných dat je použita MySQL databáze. Po schválení daného příspěvku administrátorem dojde ke stažení videozáznamu.

2.3 Přepis textu

Základem pro přepis textu jsou stažené videozáznamy. V první verzi aplikace byly textové přepisy jednotlivých audiovizuálních souborů dodávány výzkumnou skupinou Speech@FIT ve formě .mlf souborů. V následující verzi aplikace je pro získávání přepisů využíváno služby SpokenData.com. Po nasbírání administrátorem zvoleného počtu záznamů a po získání přepisů z daných záznamů je spuštěn proces automatické detekce a segmentace témat.

2.4 Segmentace

Prvním krokem v automatickém zpracování je segmentace přepsaného textu. K tomu slouží skript, jehož autorem je pan Martin Sychra. Text je nejprve rozdělen na spoustu drobných částí dlouhých např. 50 až 90 slov. Následně jsou tyto části mezi sebou porovnávány za použití metody *cosine similarity* jejímž základem je vzorec:

$$\cos\theta = \frac{A \cdot B}{|A| \cdot |B|} \tag{1}$$

Tato metoda vychází z výpočtu cosinu úhlu mezi dvěma vektory. Čím více si jsou dva vektory podobné, menší úhel svírají a tím větší je výsledná hodnota cosinu. Skript využívá při hledání hranic témat posuvné okno, které zachovává větší délku základního bloku[5] a zároveň je tím zajištěna detekce případné hranice témat uvnitř bloku. Výsledkem této části procesu jsou textové soubory, které obsahují vždy jeden segment přepisu odpovídající jednomu tématu.

3. Vstupní data pro SMM

Dalším krokem v procesu zpracování přepisů videozáznamů je vytvoření vstupu pro SMM za pomoci dat z vytvořených segmentů. Vstupem je řídká matice, jejíž řádky odpovídají indexu ze slovníku slov, které se ve vytvořených segmentech objevují a sloupce odpovídají vytvořeným segmentům. Jednotlivé sloupce potom odpovídají vektorům s histogramy četností jednotlivých slov v daném segmentu.

[5] blok ve smyslu velikosti posuvného okna, které může být klidně i 5x90 slov dlouhé v závislosti na zdroji dat

4. Trénování a extrakce vektoru dokumentů

Po vytvoření potřebné vstupní matice je spuštěno trénování SMM a následná extrakce vektoru jednotlivých dokumentů.

Cílem SMM je redukce dimensionality. V praxi to znamená nalézt souvislou nízkodimenzionální reprezentaci vstupních dat. Například z vekoru s dimenzí 33 000 vytvořit vektor s dimenzí 300.

Tato funkcionalita je zajištěna skriptem, jehož autorem je Santosh Kesiraju. Jeho výstupem je matice, kde každý sloupec je vektor dokumentu, který má oproti své vstupní reprezentaci menší dimenzionalitu.

4.1 Hledání podobných dokumentů

Poslední výpočetní fází procesu je hledání podobných dokumentů. Vstupem této části je matice 300-rozměrných vektorů, kde každý sloupec reprezentuje jeden segment přepisu. Nejdříve je vytvořena čtvercová matice o rozměru dle počtu nalezených segmentů. Následně jsou ze vstupní matice postupně vybírány jednotlivé sloupce a je vypočítávána euklidovská vzdálenost daného vektoru vůči všem ostatním vektorům vstupní matice podle vzorce:

$$d_E(u,v) = ||u - v||_2 \tag{2}$$

Výsledek každého porovnání je uložen na odpovídající koordináty vytvořené čtvercové matice. Každý výsledek je uložen i na diagonálně souměrnou pozci, aby se zamezilo zbytečnému opakování výpočtu euklidovské vzdálenosti. Po dokončení tohoto výpočtu je následně pro každý řádek čtvercové matice vyhledáváno pět výsledků s nejmenší hodnotou euklidovské vzdálenosti. Tato hodnota společně s identifikátorem dokumentu je následně uložena do pomocného seznamu.

4.2 Uložení výsledků do databáze

V tomto posledním bloku jsou hodnoty uložené v pomocném poli postupně ukládány do MySQL databáze. Od této chvíle jsou k jednotlivým tématickým segmentům uloženy záznamy o tématicky nejpodobnějších segmentech.

5. Shrnutí

Výsledný výstup je možné shlédnout na webových stránkách uvedených v abstraktu tohoto příspěvku. Uživateli jsou zde nejdříve nabídnuty všechny dostupné videozáznamy rozdělené podle uživatelem definovaných kategorií. Z hlavní stránky má uživatel možnost si přehrát libovolmý záznam či přejít na stránku pro přidání nového záznamu.

Při přehrávání se uživateli napravo od videozáznamu zobrazuje seznam tématicky nejpodobnějších segmentů, které korespondují s nalezenými segmenty v daném videozáznamu.

Celá webová prezentace v této chvíli stále prochází vývojem a proto se rozložení, grafické provedení a funkcionalita mohou postupem času měnit.

Zkoumání a testování, zda si dané segmenty opravdu tématicky odpovídají, je silně zatíženo subjektivním hodnocením testujících. Hlavní testování však teprve proběhne.

Poděkování

Chtěl bych velice poděkovat vedoucímu mé práce panu Ing. Igoru Szökemu, Ph.D. a panu Santoshi Kesirajuovi za jejich ochotu a pomoc při řešení této práce.

Literatura

[1] Florian Kleedorfer, Peter Knees, and Tim Pohle. Oh oh oh whoah! towards automatic topic detection in song lyrics, 2008. http://ismir2008.ismir.net/papers/ISMIR2008_211.pdf.

[2] F. H. Ismail. Road map approach to automatic topic detection of diaries, Duben 2013. http://www.ijmlc.org/papers/310-K1004.pdf.

[3] Martin Sychra. *Automatické hledání vazeb mezi částmi audiovizuálních dokumentů*. Brno, FIT VUT v Brně, 2015. bakalářská práce.

[4] Martin Sychra. Automatic link detection in parts of audiovisual documents, 2015. http://excel.fit.vutbr.cz/submissions/2015/038/38.pdf.

[5] Mehdi Soufifar, Marcel Kockmann, Lukáš Burget, Oldřich Plchot, Ondřej Glembek, and Torbjorn Svendsen. *iVector Approach to Phonotactic Language Recognition*, volume 2011. International Speech Communication Association, 2011. http://www.fit.vutbr.cz/research/view_pub.php?id=9758.