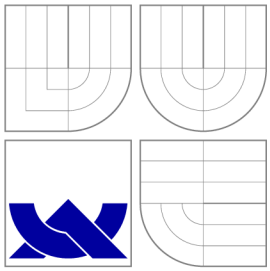
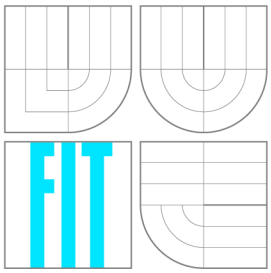


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

PŘIDÁVÁNÍ KOMENTÁŘŮ A AUTOMATICKÁ IDENTIFIKACE KLÍČOVÝCH FRÁZÍ PRO WEBOVÝ PROHLÍŽEČ AUDIO/VIDEO ZÁZNAMŮ PŘEDNÁŠEK

COMMENTS AND AUTOMATIC KEY PHRASE IDENTIFICATION FOR WEB BASED AUDIO/VIDEO LECTURE BROWSER

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

RADIM GLAJC

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. IGOR SZÖKE, Ph.D.

BRNO 2011

Abstrakt

Cílem této bakalářské práce je vytvoření rozšíření pro webový prohlížeč přednášek. Jedná se o modul umožňující uživatelům přidávat komentáře k přednáškám, dále o administrační rozhraní usnadňující práci správce přednášek a o funkci pro automatickou identifikaci klíčových frází v přepisech přednášek.

Abstract

The aim of this thesis is to create extensions for web-based lecture browser. These are modules for adding user comments, administration interface and function for automatic key phrases identification in lecture transcripts.

Klíčová slova

Přidávání komentářů, identifikace klíčových frází, administrační rozhraní, PHP, MySQL, JavaScript, AJAX.

Keywords

Adding Comments, Key Phrase Identification, Administration Interface, PHP, MySQL, JavaScript, AJAX.

Citace

Radim Glajc: Přidávání komentářů a automatická identifikace klíčových frází pro webový prohlížeč audio/video záznamů přednášek, bakalářská práce, Brno, FIT VUT v Brně, 2011

Přidávání komentářů a automatická identifikace klíčových frází pro webový prohlížeč audio/video záznamů přednášek

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Igora Szökeho. Další informace mi poskytl pan Josef Žižka. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Radim Glajc
5. května 2011

Poděkování

Rád bych poděkoval panu Igoru Szökemu a Josefu Žižkovi za užitečné rady, čas a ochotu, kterou mi věnovali při vypracování této práce.

Obsah

Kapitola 1

Úvod

Webový prohlížeč přednášek[1] je aplikace vyvinutá výzkumnou skupinou Speech@FIT[2]. Umožňuje uživatelům on-line prohlížení záznamů z přednášek, prohlížení tzv. „slajdů“ synchronizovaných s přehrávaným videem a zobrazení relevantních informací k přednášce. Především ale poskytuje vyhledávání v automaticky tvořených přepisech přednášek, což usnadňuje a zrychluje přístup k požadovaným informacím [15]. Současnou implementaci Webového prohlížeče přednášek a použité technologie popisuje kapitola 2.

Tato práce se věnuje vytvoření tří rozšíření pro Webový prohlížeč přednášek.

Prvním rozšířením je modul, umožňující uživatelům přidávat komentáře k přednáškám. Pokud student při sledování přednášky něčemu nerozumí nebo nepochopí daný problém, může na stránku s touto přednáškou vložit dotaz a přiřadit ho k příslušnému času v přednášce. Dotaz se zobrazí i všem ostatním uživatelům sledujícím tuto přednášku a kdokoliv na něj může odpovědět. Stejným způsobem lze vložit odkaz na další zdroje informací, které souvisí s přednášenými tématy, lze upozornit na otázky či příklady, které se často vyskytují u zkoušky a podobně. Hlavním přínosem tohoto modulu je shromáždění všech těchto informací souvisejících s konkrétní přednáškou na jediném místě. Modul pro přidávání komentářů je popsán v kapitole 3.

Druhým rozšířením je administrační rozhraní. Po zaznamenání nové přednášky je potřeba ji zpracovat – přidat úvodní a závěrečnou grafiku, vystřihnout přestávky, vytvořit přepis řeči přednášejícího, vložit slajdy a načasovat jejich zobrazování. Až poté je přednáška připravena k vložení do prohlížeče.

Údaje o přednáškách, kategoriích, autorech přednášek a další data jsou uložena v databázi. Tyto údaje jsou do databáze zapisovány ručně pomocí nástrojů pro práci s databází, což je časově náročné, neintuitivní a náchylné k vložení chybných informací. Administrační rozhraní tuto práci zjednoduší, urychlí a zároveň kontroluje správnost zadávaných informací. Vkládání, úprava a odstraňování přednášek, kategorií, autorů a dokumentů či správa komentářů pak bude záležitostí několika kliknutí myší. Administrační rozhraní je popsáno v kapitole 4.

Posledním rozšířením je funkce pro automatickou identifikaci klíčových frází v přepisu přednášek. Klíčové fráze jsou slovní spojení, která vystihují témata přednášky. Buď je lze získat tak, že si někdo celou přednášku poslechne a poté klíčové fráze vypíše, v dnešní době je však lze získat i automaticky. Díky práci skupiny Speech@FIT je možné automaticky vytvořit přepis řeči přednášejícího. Tento přepis není dokonalý, některá slova nejsou rozpo-

[1] http://www.prednasky.com
[2] http://speech.fit.vutbr.cz

znána správně, ale i tak postačuje k tomu, aby v něm uživatelé mohli vyhledávat, nebo aby v něm byly automaticky identifikovány klíčové fráze. Takto získané fráze je poté možné zobrazit v seznamu přednášek a/nebo kategorií, což umožní uživatelům získat přehled o tom, čeho se přednáška či kategorie týká. Na tuto práci je rovněž možné navázat rozšířením, které by v transkriptu přednášky automaticky vytvářelo odkazy na klíčové fráze do internetové encyklopedie Wikipedia. Automatická identifikace klíčových frází je popsána v kapitole 5.

V závěru práce jsou zhodnoceny dosažené výsledky a uvedeny návrhy na další vylepšení těchto tří rozšíření Webového prohlížeče přednášek.

Kapitola 2

Analýza webového prohlížeče přednášek

Webový prohlížeč přednášek je multiplatformní aplikace, kterou v roce 2009 vytvořil v rámci své diplomové práce Josef Žižka ve spolupráci s týmem Speech@FIT. Na vývoji aplikaci se dále podíleli a podílejí studenti bakalářského i magisterského studia na FIT VUT v Brně.

Prohlížeč klade důraz na poskytnutí informací k problematice probírané na dané přednášce. Tzn. aby byl záznam přednášky, slajdy, studijní opory a další informace soustředěny přehledně na jednom místě.

Aplikace funguje ve všech moderních webových prohlížečích (Internet Explorer, Mozilla Firefox, Opera, Google Chrome). Pro přehrávání videa je nutné mít nainstalovaný Adobe Flash Player, což v dnešní době (březen 2011) splňuje 99% uživatelů[1]. Dále je nutné mít povolený JavaScript (podle údajů z ledna 2008 mělo JavaScript povoleno 95% uživatelů[2]).

Předpokladem pro implementaci zvolených rozšíření bylo seznámit se podrobněji s tímto prohlížečem. Sekce 2.1 popisuje hlavní komponenty prohlížeče. V sekci 2.2 je popsáno technické řešení a technologie, zajišťující provoz prohlížeče. Sekce 2.3 popisuje strukturu databáze a jednotlivé tabulky.

2.1 Hlavní komponenty prohlížeče

Tato sekce vychází z diplomové práce Josefa Žižky [16].

Webový prohlížeč přednášek tvoří tři základní typy stránek:

- Výběr kategorie a přednášky
- Výsledky vyhledávání v přepisech přednášek
- Přehrávání přednášky

Podrobněji je rozebrána stránka pro přehrávání přednášky. Vzhled této stránky je na obrázku 2.1. Skládá z těchto komponent:

- **Videozáznam** je hlavní součástí prohlížeče, slouží k přehrávání přednášky. Je realizován pomocí JW FLV Media Playeru. Umožňuje zobrazit přímo ve videu přepis přednášky.

[1] http://www.adobe.com/products/player_census/flashplayer/
[2] http://www.w3schools.com/browsers/browsers_stats.asp

- **Vyhledávání v audiu**. Během přehrávání videa umožňuje vyhledávat v transkriptu přednášky. Vyhledávání využívá technologii AJAX.
- **Přepis přednášky (transkript)** může být zobrazován ve videu a/nebo pod videem, se kterým je synchronizován.
- **Slajdy**. Obsah plátna lze zobrazit ve vyšší kvalitě, než jaká je dostupná ve videu.
- **Odkazy na související literaturu** – např. opory k předmětu nebo slajdy ke stažení.
- **Odkazy na podobné přednášky**.

Všechny komponenty prohlížeče (kromě odkazů) jsou vzájemně synchronizovány podle časové osy přednášky. V praxi to znamená, že zatímco je přehráván videozáznam, zobrazují se k němu aktuální slajdy a transkript. Kliknutím na libovolný slajd, část transkriptu nebo výsledek vyhledávání se začne video přehrávat od daného času.

Tato synchronizace je realizována pomocí JavaScriptových událostí.

2.2 Technické řešení

Webový prohlížeč přednášek využívá zdarma dostupné technologie, které jsou popsány v této sekci. Zdrojem informací je diplomová práce Josefa Žižky [16].

2.2.1 Server

Jako webový server je používáno open-source řešení Lighttpd se zásuvným modulem mod_flv_streaming. Lighttpd je využíván mnoha populárními webovými aplikacemi, jako je např. YouTube, Wikipedie nebo Meebo. Autoři tvrdí, že Lighttpd nabízí lepší výkon, než ostatní web-servery, běžící na stejném hardware [6]. Zásuvný modul mod_flv_streaming umožňuje pseudo-streaming, založený na progresivním stahování obsahu. Proto není problémem začít přehrávat video od jakéhokoliv místa [16].

2.2.2 Přehrávač videa

Pro přehrávání videa byl zvolen JW FLV Media Player, což je nejpopulárnější open-source přehrávač videí. Umožňuje přehrát všechny multimediální formáty, které podporuje Adobe Flash Player, např. FLV, MP4, MP3 a další. Podporuje RTMP, HTTP, streaming, dále umožňuje využívat zasílání zpráv přes JavaScript API, změnu vzhledu nebo vložení zásuvných modulů [7].

2.2.3 Generování stránek a uživatelské rozhraní

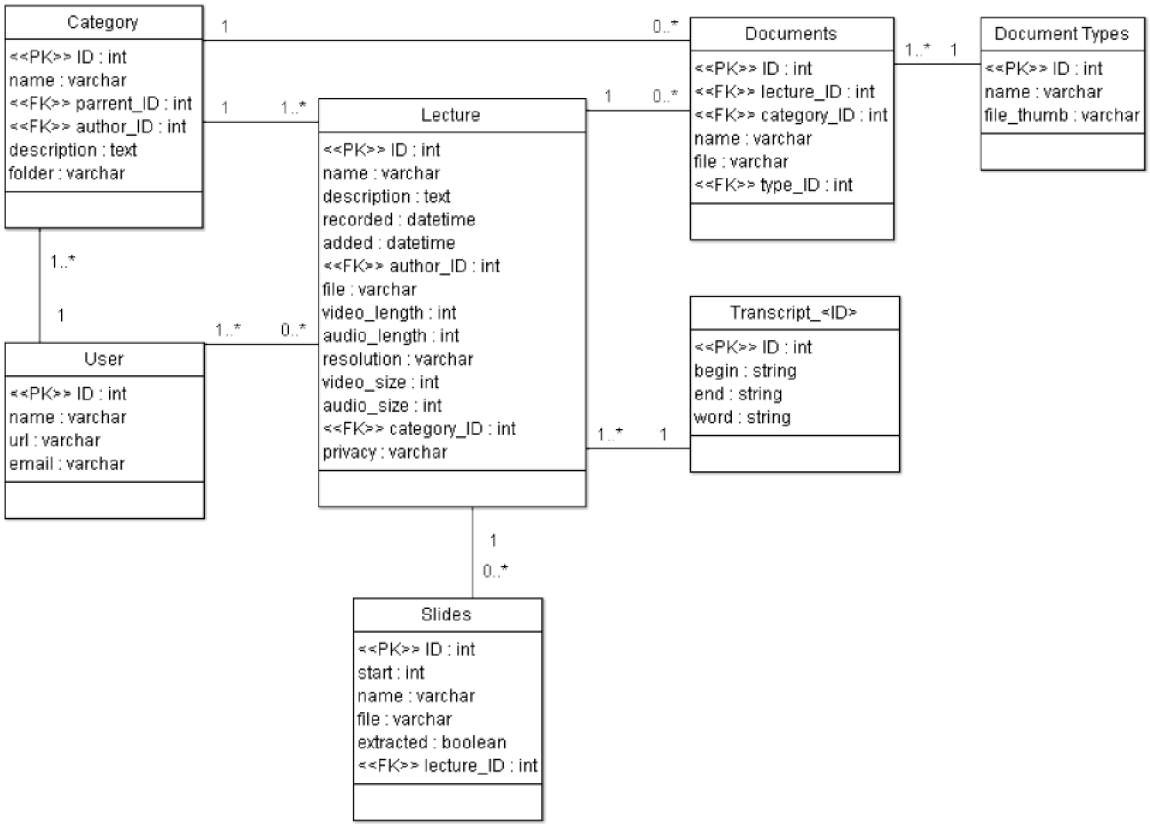
Generování HTML kódu jednotlivých stránek prohlížeče probíhá pomocí PHP[3] skriptů. Textová data (údaje o přednáškách, slajdech, transkripty aj.) byla původně uložena v XML souborech[4]. V průběhu roku 2010 byla v rámci diplomové práce Jakuba Janoviče [5] převedena do databáze MySQL[5]. Klientská část aplikace využívá JavaScript. Pro stažení dat ze serveru bez nutnosti překreslit celou stránku je použita technologie AJAX.

[3] http://www.php.net/
[4] http://www.w3.org/XML/
[5] http://www.mysql.com/

Obrázek 2.1: Prohlížeč přednášek: stránka s přehráváním

2.3 Databáze

Zde jsou stručně popsány jednotlivé tabulky (podle [5]). Struktura databáze je na obrázku 2.2.



Obrázek 2.2: Schéma databáze. Převzato z [5].

Lecture

Tabulka *Lecture* obsahuje hlavní údaje o přednášce: jméno a popis, datum a čas zaznamenání a uložení přednášky, název souboru, délku audia a videa, rozlišení a velikost audia a videa. Váže na tabulky *Category* (kategorie, do které přednáška patří) a *User* (autor/vlastník přednášky).

Category

Zde jsou uloženy údaje o kategoriích – její název, popis a jméno adresáře, ve kterém jsou uloženy jednotlivé přednášky. Váže se na tabulku *User* (autor/vlastník kategorie). Každá kategorie také může mít nadřazenou kategorii a je proto umožněno vytvořit hierarchickou strukturu kategorií.

User

Tato tabulka obsahuje jméno, e-mailovou adresu a webovou adresu uživatele systému.

Slides

V tabulce *Slides* jsou uloženy slajdy k přednáškám. Tabulka obsahuje název, počáteční čas a jméno souboru se slajdem. Příznak extracted udává, zda byl slajd extrahován z videa. Tabulka se váže na tabulku *Lectures*.

Documents

Tabulka obsahuje názvy dokumentů a jména příslušných souborů. Dokumenty mohou souviset s přednáškou (tabulka *Lectures*) nebo s kategorií (tabulka *Category*). Dále se tabulka váže na tabulku *DocumentTypes*, která obsahuje názvy typů dokumentů a cesty k souborům s ikonami.

Transcript_*id*

Jde o množinu tabulek, obsahujících přepisy přednášek, kde *id* značí identifikátor přednášky. Přepisy nejsou uloženy v jediné tabulce *Transcripts* z důvodu rychlosti. Jelikož jedna přednáška obsahuje až 15000 slov, byly by dotazy nad jedinou velkou tabulkou výpočetně náročně.

V tabulkách *Transcript_id* jsou uložena jednotlivá slova přepisu společně s časovými údaji, přičemž atribut *begin* značí počáteční čas slova a atribut *end* konečný čas celého segmentu.

Kapitola 3

Komentáře

Cílem tohoto rozšíření je umožnit uživatelům, aby upozorňovali na důležité pasáže přednášky, ptali se (a odpovídali) na nejasnosti, přidávali odkazy na další materiály a podobně. Spolu se slajdy a oporami ke stažení pak budou všechny podstatné informace na jednom místě.

V sekci 3.1 jsou uvedeny požadavky na výslednou aplikaci. Sekce 3.2 se zabývá existujícími řešeními podobných aplikací. V sekci 3.3 je navrženo vlastní řešení a v sekci 3.4 rozebrána implementace tohoto řešení. Závěrečná sekce 3.5 se věnuje zhodnocení dosažených výsledků.

3.1 Analýza požadavků

Komponenta Komentáře by měla rozšiřovat funkčnost Webového prohlížeče přednášek. Mezi hlavní požadavky na aplikaci patří:

- Intuitivní a jednoduše použitelné uživatelské rozhraní
- Eliminace vkládání spamu a komentářového vandalismu, ohodnocení kvalitních příspěvků
- Aktualizace dat bez nutnosti obnovovat celou stránku
- Možnost přiřazovat komentáře k času ve videu, odkazování na tento čas v textu komentáře

Pro splnění těchto požadavků byla nejprve prozkoumána existující řešení aplikací podobných Webovému prohlížeči přednášek a následně navrženo vlastní řešení.

3.2 Existující řešení

Tato sekce představuje aplikace podobné Webovému prohlížeči přednášek. Zaměřuje se především na možnost uživatelských reakcí formou komentářů, další vlastnosti těchto aplikací jsou popsány např. v [16].

Možnost vkládání komentářů však není u podobných aplikací příliš rozšířená, proto jsou na závěr uvedeny možnosti komentářů u služby YouTube, poskytující obsah vytvářený samotnými uživateli.

3.2.1 MIT Lecture Browser[1]

Prohlížeč vyvinutý na Massachusetts Institute of Technology umožňuje vyhledávat v přepisech přednášek a přednášky přehrávat přímo ve webovém prohlížeči. Nevýhodou je nutnost instalace RealPlayeru. MIT Lecture Browser však neumožňuje uživatelům vkládat komentáře.

3.2.2 VideoLectures.net[2]

Server VideoLectures.net nabízí ke shlédnutí přes 10000 přednášek (stav z března 2011) z mnoha vědních oborů. Vyhledávání v přepisech přednášek není možné. Uživatelé mohou vkládat komentáře k jednotlivým přednáškám, nemusí se registrovat ani zadávat CAPTCHA[3]. Přesto se zde nevyskytuje komentářový spam. Komentáře mají ale pouze základní funkcionalitu, nelze odpovídat na předchozí komentáře ani je hodnotit.

3.2.3 Videoserver Masarykovy univerzity[4]

Masarykova univerzita v Brně nabízí záznamy přednášek, akcí či studentských prací volně ke stažení. Nelze je přehrávat online. Studenti Fakulty informatiky mohou stahovat všechny přístupné záznamy přednášek nebo sledovat on-line přenos z poslucháren. Ostatní uživatelé mohou stahovat pouze vybrané záznamy. Videoserver však neumožňuje vkládat komentáře.

3.2.4 Audiovizuální centrum studentů ČVUT[5]

České vysoké učení technické v Praze nabízí nejen záznamy z vybraných přednášek, ale také z různých akcí či seminářů na těchto školách. Záznamy vznikají dobrovolnou aktivitou studentů. K dispozici je několik formátů záznamu, např. MP4, WMV, stream a je také možné přehrát video přímo ve webovém prohlížeči. Komentáře také nelze přidávat.

3.2.5 YouTube[6]

Služba YouTube je poskytovatelem videoobsahu, který na server přidávají samotní uživatelé. Videa mohou ostatní uživatelé komentovat, pokud to autor videa nezakázal. Systém komentářů je poměrně propracovaný, proto zde bude podrobněji rozebrán.

Přidávání komentářů

Komentáře mohou přidávat pouze registrovaní a přihlášení uživatelé. Lze vkládat komentáře dlouhé až 500 znaků, což je dostačující. Po překročení tohoto limitu nelze komentář odeslat, ale uživatel může psát dál, což je uživatelsky nepříjemné.

Také lze komentovat formou tzv. videoodpovědi, kdy uživatel vloží jako reakci své vlastní video.

[1] http://web.sls.csail.mit.edu/lectures/
[2] http://videolectures.net/
[3] http://www.captcha.net/
[4] http://www.video.muni.cz/
[5] http://www.avc-cvut.cz/
[6] http://www.youtube.com/

Seznam komentářů

Seznam komentářů je rozdělen do čtyř oddílů. Jako první jsou zobrazeny komentáře autora videa, případně i navazující reakce. Ve druhém oddílu následují dva komentáře s nejvyšším hodnocením (viz dále), takže kvalitní komentáře jsou více na očích. Třetí oddíl je vyhrazen pro videoodpovědi (pokud existují). Následuje textové pole pro vložení nového komentáře a pod ním jsou ostatní komentáře seřazené od nejnovějších po nejstarší. Nejsou však zobrazeny všechny, pro listování mezi stránkami komentářů jsou k dispozici tlačítka *Předchozí*, *Další* a maximálně 7 tlačítek pro přístup na konkrétní stránku. Chybí zde však tlačítka *První* (resp. *Nejnovější*) a *Poslední* (resp. *Nejstarší*), takže při velkém počtu komentářů je obtížnější prolistovat se k nejstarším komentářům.

Zobrazení komentáře

Každý komentář obsahuje text, pod ním je uveden autor a časový údaj, kdy byl komentář přidán. Časový údaj je zobrazen relativně k aktuálnímu času, např. „před X min/hod/dny“. To může působit nepřehledně a komplikuje to vyhledání komentářů přidaných například v určitý den.

Po najetí myší na komentář se zobrazí tlačítka pro hlasování o komentáři (palec nahoru nebo dolů), tlačítko pro vložení reakce a nahlášení komentáře jako spamu. Autor videa navíc vidí tlačítka pro odstranění a sdílení cizích komentářů, a autor komentáře vidí tlačítko pro odstranění svého příspěvku. Komentář označený jako spam je skrytý, ale je možné ho zobrazit a nahlásit, že nejde o spam.

Komentáře, které obdržely největší množství pozitivních hlasů, jsou zobrazeny v horní části výpisu, komentáře s velkým počtem negativních hlasů jsou trvale smazány.

Odpovědi na komentář se zobrazují odsazeně přímo pod komentářem, na který je reagováno. Pokud se ale v komentářích rozvine diskuse, čili se odpovídá na předcházející odpověď, odsazení zůstává pro všechny příspěvky stejné a nelze zjistit, na který komentář je reagováno.

Do textu komentáře je možné vložit odkaz na čas ve videu, a to zapsáním časového údaje ve formátu *mm:ss*. Po kliknutí na tento odkaz se video začne přehrávat od daného času. Tento údaj ale není nijak porovnáván s délkou videa. Časový údaj také může být omylem převeden na odkaz, i když to autor příspěvku neměl v úmyslu.

3.3 Návrh řešení

3.3.1 Uživatelské rozhraní

Komentářový systém bude rozšiřovat funkcionalitu webového prohlížeče přednášek, musí být tedy snadno použitelný. Vzhled komentářového systému by se měl podobat ostatním blokům prohlížeče a to především barvami a použitými fonty. Uživatelské rozhraní musí být pro uživatele intuitivní, příliš se nelišící od jiných komentářových systémů.

3.3.2 Vkládání komentářů

Bude možné vkládat buď samostatné komentáře nebo odpovědi na předchozí příspěvky. Pokud bude vložena odpověď, musí být jasné, na který příspěvek reaguje.

Uživatel nesmí být nucen k vyplňování mnoha zbytečných informací, aby mohl vložit komentář, proto budou požadovány pouze nejnutnější údaje – čas (viz dále) a text ko-

mentáře. Přidávání komentářů bude dostupné pouze pro přihlášené uživatele, nebude tedy nutné vyplňovat v komentářích i jméno.

Zároveň tím bude zabráněno přístupu tzv. spambotů, počítačových programů, které vkládají do komentářů reklamy na různé stránky či produkty.

3.3.3 Hlasování o kvalitě

Předpokládá se, že spam vkládaný uživateli bude výrazně redukován díky systému hlasování. U každého komentáře budou mít uživatelé možnost hlasovat o kvalitě daného příspěvku. Bude také nutné zajistit objektivitu hlasování, aby uživatel mohl hlasovat pro jeden komentář pouze jednou.

3.3.4 Karma uživatele

Toto hlasování se zároveň promítne do tzv. karmy autora příspěvku. Karmou se zde myslí suma kladných a záporných hlasů pro komentáře uživatele. Komentáře s kladnými hlasy budou zvýrazněny, naopak komentáře se zápornými hlasy budou skryty nebo označeny jako nekvalitní. Karma uživatele bude ovlivňovat nutné počty hlasů ke zvýraznění/skrytí příspěvku. Může být také využita v jiných rozšířeních webového prohlížeče přednášek, např. v manuální korektuře přepisu řeči, popsané v diplomové práci Jakuba Janoviče [5].

3.3.5 Přiřazení času

Komentáře budou rozděleny na dva typy a uživatel si při vkládání příspěvku jeden z nich zvolí:

1. komentář související s celou přednáškou (např. shrnutí, odkaz na literaturu) a
2. komentář související pouze s určitým úsekem (např. vysvětlení či odkaz ke konkrétnímu problému).

Komentáře s přiřazeným časem budou synchronizovány s videem a zvýrazňovány podle aktuálního času přehrávání. Kliknutím na časový údaj se video začne přehrávat od tohoto času, slajdy a transkript se také sesynchronizují. Bude nutné zjišťovat, zda časový údaj není větší než délka přednášky. Časový údaj bude možné vložit i přímo do textu komentáře a bude převeden na odkaz.

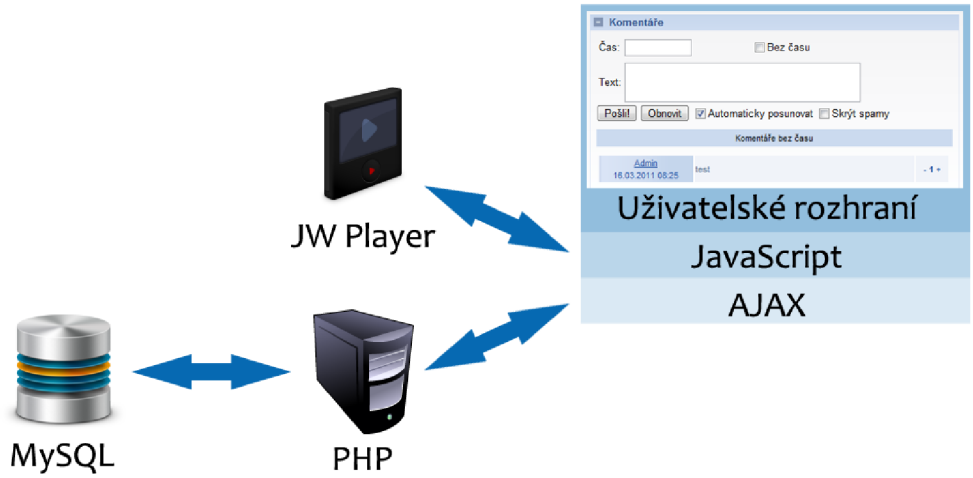
Architektura aplikace je na obrázku 3.1.

3.4 Implementace

3.4.1 Databáze

Nový návrh struktury databáze je na obrázku 3.2. Struktura databáze z [5] byla rozšířena o 2 nové tabulky – *Comments* a *Votes*. Tabulka *User* byla upravena pro použití v komentářích a v administračním rozhraní systému (viz kapitola 4). Následuje popis těchto upravených tabulek.

- **User** – z původní struktury (viz sekce 2.3) zůstávají atributy *name*, *url* a *email*. Atribut *score* určuje karmu uživatele. V současné implementaci hodnotu karmy určuje



Obrázek 3.1: Architektura komentářového systému. Byly použity ikony z [2].

suma všech kladných i záporných hlasů u komentářů, ale pro pozdější použití byla zvolena možnost ukládat hodnotu karmy právě do atributu *score* tabulky *User*. Atribut *permissions* se skládá z kombinace 8 znaků, přičemž každá pozice označuje oprávnění k různým operacím. Komentářů se týká hodnota atributu na druhé pozici, kdy znak „0“ zakazuje uživateli vkládání komentářů, jiné znaky (typicky „1“) vkládání povolují. Atribut *permissions* a ostatní využívá administrační rozhraní systému, popsané v kapitole 4.

- **Comments** – do této tabulky jsou ukládány komentáře. Komentář je přiřazen k určité přednášce (tabulka *Lecture*) a má svého autora (tabulka *User*). Dále se u komentáře ukládá datum vložení, časový údaj (viz sekce 3.3.5), skóre a text komentáře. Skóre je implicitně nastaveno na hodnotu 0 a je později ovlivněno hlasováním ostatních uživatelů. Atribut *re_ID* určuje *ID* komentáře, na který tento příspěvek reaguje. V případě že jde o samostatný komentář, má hodnotu *NULL*.

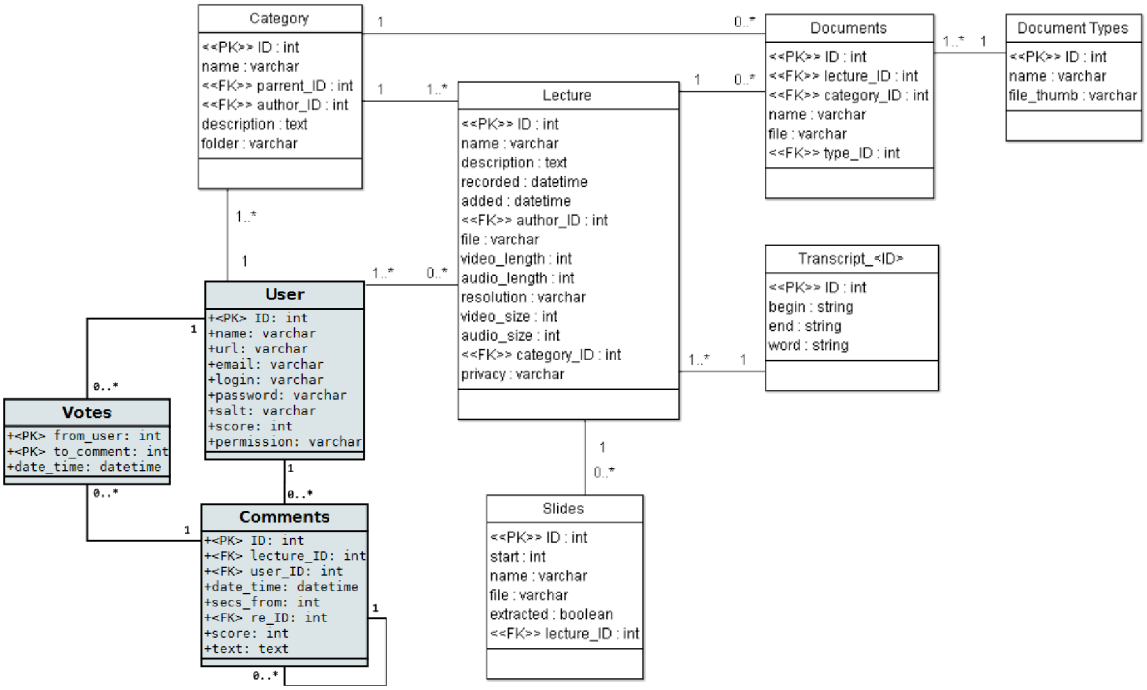
- **Votes** – tato tabulka zabezpečuje, že jeden uživatel hlasoval pro jeden komentář nejvýše jedenkrát. To zaručuje jedinečnost primárního klíče, složeného ze dvou atributů *from_user* a *to_comment*. Tabulka tedy neslouží jako záznam jak bylo hlasováno, ale zda (a kdy) bylo hlasováno.

3.4.2 Serverová část

Serverová část je implementována v jazyce PHP. Slouží k načítání dat z databáze a sestavování HTML kódu komentářů, dále k ukládání dat do databáze a ukládání informací o hlasování. Také slouží k přihlašování uživatelů, tato funkce je ale v reálném provozu webového prohlížeče přednášek implementována jiným způsobem.

Zde jsou popsány jednotlivé komponenty serverové části.

- **Začlenění do prohlížeče přednášek** je provedeno pomocí souboru *index.php*, který je vložen na patřičné místo při generování HTML kódu webového prohlížeče před-



Obrázek 3.2: Nový návrh struktury databáze. Byl použit obrázek z [5].

nášek. Skript vytváří základní HTML strukturu boxu s komentáři, kontroluje, zda je uživatel přihlášen a zda má právo vkládat komentáře nebo je pouze prohlížet.

- **Získání dat z databáze a sestavování HTML kódu komentářů** provádí skript *comments-show.php* podle identifikátoru právě zobrazené přednášky.

  Rozděluje komentáře na ty, které mají přiřazený čas v přednášce a na komentáře bez časového přiřazení. Komentáře s přiřazeným časem jsou řazeny podle tohoto času, v případě stejného času u více komentářů jsou řazeny podle data a času vložení. Komentáře bez časového přiřazení jsou řazeny pouze podle data a času vložení.

  Kvůli přehlednosti jsou všechny komentáře zobrazeny na jedné úrovni. Je však vytvořena varianta stromového zobrazování komentářů (soubor *comments-show-tree.php*), která může být volitelně použita. Algoritmus pro zobrazení stromové diskuze byl převzat z [3].

  Podle skóre komentáře a karmy jeho autora jsou komentáře zvýrazňovány nebo naopak potlačovány. Skript také převádí odkazy na čas v přednášce zapsané ve tvaru *@mm:ss* nebo *@hh:mm:ss* na javascriptový odkaz, který provede posun videa na daný čas.

  Skript *comments-bubble.php* vytváří základní HTML strukturu pro zobrazování aktuálních komentářů přímo pod videem.

- **Ukládání komentářů a hlasování** provádí skripty *comments-insert.php* a *comments-score.php*. Ověřují přihlášení a oprávnění uživatele k těmto akcím, ošetřují vstupní data, a pokud je vše v pořádku, provedou zápis dat do databáze.

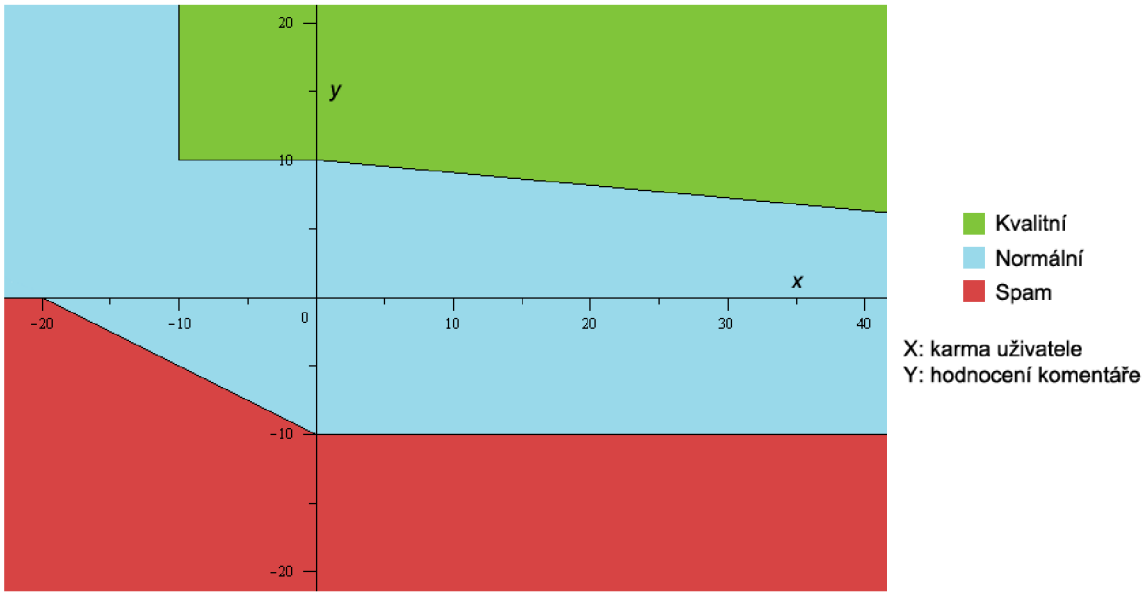
3.4.3 Karma

Zvýrazňování komentářů záleží jednak na hodnocení konkrétního komentáře (hodnota atributu *score* v tabulce *Comments*), jednak na karmě autora komentáře (hodnota atributu *score* v tabulce *User*). Cílem je zvýraznění kvalitních komentářů a potlačení nekvalitních komentářů. Pokud uživatel udělí komentáři kladný resp. záporný hlas, je inkrementováno resp. dekrementováno hodnocení daného komentáře a karma autora tohoto komentáře.

Komentáře, jejichž autoři mají vyšší karmu, by měly být oproti ostatním zvýhodněny, tzn. mělo by k jejich zvýraznění stačit menší množství kladných hlasů. Stejně tak komentáře, jejichž autoři mají nízkou karmu, musí získat větší množství kladných hlasů, aby byly označeny jako kvalitní. Takovýmto komentářům ale stačí menší množství záporných hlasů k tomu, aby byly označeny jako spam.

Způsob zvýrazňování komentářů podle karmy uživatele a hodnocení komentáře byl zvolen experimentálně a byl otestován na malém počtu uživatelů. Při reálném použití s vysokým počtem uživatelů a komentářů může být nevyhovující a může být změněn.

V následujících rovnicích je jako *usr* označena karma uživatele a jako *com* hodnocení daného komentáře. Graf zvýrazňování komentářů je znázorněn na obrázku 3.3.



Obrázek 3.3: Označení komentářů podle hodnocení a karmy uživatele

- Komentář je označen jako spam, pokud v daném intervalu hodnot *usr* platí:
  - $usr \in (-\infty, -20\rangle : com \leq 0$
  - $usr \in (-20, 0\rangle : 0.5 \cdot usr + com \leq -10$
  - $usr \in (0, \infty) : com \leq -10$

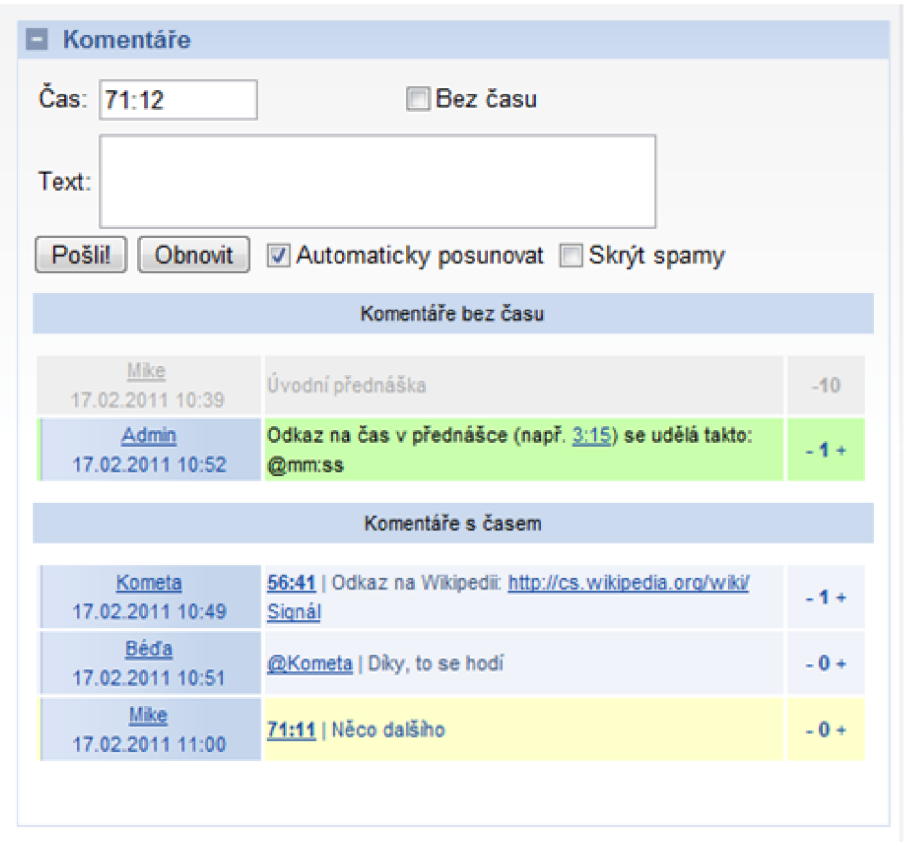
- Komentář je označen jako kvalitní, pokud v daném intervalu hodnot usr platí:
  - $usr \in (-\infty, -10\rangle$: nikdy
  - $usr \in (-10, 0\rangle$: $com \geq 10$
  - $usr \in (0, \infty)$: $0.09 \cdot usr + com \geq 10$
- V ostatních případech není komentář nijak označován.

3.4.4 Klientská část

Klientská část aplikace zprostředkovává komunikaci mezi uživatelem a serverovou částí. Je u ní důležitá především rychlost odezvy a intuitivní uživatelské rozhraní. Klientská část je implementována v jazyce JavaScript.

Rychlost odezvy je zajištěna použitím technologie AJAX. Díky ní není potřeba obnovovat celou stránku, jestliže uživatel odešle komentář, ohodnotí jiný komentář nebo aktualizuje seznam komentářů.

Uživatelské rozhraní bylo navrženo tak, aby se příliš nelišilo od podobných aplikací (např. komentářů u služby YouTube). Vzhled uživatelského rozhraní je na obrázku 3.4.



Obrázek 3.4: Uživatelské rozhraní komentářů

Textové pole v horní části komentářového boxu slouží k **vložení nového komentáře**. Uživatel musí zvolit, zda chce komentář přiřadit k určitému času v přednášce, a nebo zda bude komentář bez časového určení. V prvním případě se čas komentáře aktualizuje podle přehrávání přednášky, ale lze ho zapsat i ručně. Po kliknutí na tlačítko *Pošli!* aplikace zkontroluje, zda uživatel vyplnil text komentáře a zda zadaný čas je ve správném formátu a nepřekračuje délku přednášky. Při úspěšné kontrole je komentář odeslán.

Kliknutím na tlačítko *Obnovit* se aktualizuje seznam komentářů. Pomocí zaškrtávacích políček si uživatel nastaví, zda chce při přehrávání přednášky automaticky posunovat seznam komentářů podle příslušného času, a zda chce skrýt komentáře, které byly označeny jako spam. Tato nastavení se ukládají do cookies.

Zobrazení komentářů je rozděleno na komentáře bez času (obvykle souvisí s celou přednáškou) a na komentáře s určeným časem. Každý komentář je rozdělen do tří sloupců:

1. V prvním sloupci je jméno autora a datum vložení komentáře. Kliknutím na jméno autora je možné na tento komentář reagovat, což je uživateli oznámeno nad textovým polem pro vložení komentáře. Takovýto komentář má stejný časový údaj jako jeho předchůdce a nelze ho změnit.

2. V dalším sloupci je zobrazen vlastní text příspěvku. Před textem může být zobrazen:

   - časový údaj. Po kliknutí na odkaz pokračuje přehrávání přednášky od tohoto času.
   - jméno autora komentáře, na který tento příspěvek reaguje. Po kliknutí na odkaz je předcházející komentář zvýrazněn.

3. V posledním sloupci je uvedeno skóre komentáře a tlačítka + a –, kterými lze provést hodnocení. Tato tlačítka nejsou zobrazena u komentářů, které byly označen jako spam a u komentářů, u kterých administrátor odstranil jejich text.

Doplňkovou součástí uživatelského rozhraní je box komentářů, zobrazený pod videem, kde jsou zobrazovány komentáře podle času přehrávání přednášky. Uživatel může nastavit filtr komentářů podle jejich skóre (jak kvalitní komentáře budou zobrazovány) a na jakou dobu budou zobrazeny.

3.5 Zhodnocení řešení

Byla vytvořena aplikace rozšiřující funkčnost stávajícího Webového prohlížeče přednášek. Aplikace by měla sloužit ke komunikaci uživatelů, upozorňování na důležité pasáže přednášky nebo témata vyskytující se často u zkoušky, vysvětlení nejasností, vkládání odkazů na další literaturu apod. Díky jednoduchému uživatelskému rozhraní je použití komentářů intuitivní. Systém hodnocení komentářů a karmy umožňuje uživatelům hodnotit kvalitu příspěvků, a tím vyzdvihnout kvalitní a potlačit nekvalitní komentáře.

Aplikace byla otestována a je funkční v těchto webových prohlížečích:

- Mozilla Firefox 3.6
- Opera 11
- Google Chrome 4

V prohlížeči Internet Explorer 8 aplikace je aplikace funkční s omezeními – komentáře nejsou zvýrazňovány podle aktuálního času přehrávané přednášky.

3.5.1 Další rozšíření

Pokud bude aplikace použita v běžném provozu Webového prohlížeče přednášek, mohla by být implementována tato rozšíření:

- **Plná funkčnost v prohlížeči Internet Explorer 8**
- **Karma uživatelů** – současný systém zvýrazňování komentářů byl vytvořen experimentálně a v reálném provozu může být nevyhovující. Dle rozsáhlejších statistik by měla být upravena funkce pro zvýrazňování komentářů. Karma uživatelů by také mohla být použita v jiných rozšířeních Webového prohlížeče přednášek.
- **Formátování textu** – současná implementace umožňuje vkládat pouze prostý text příspěvků. Případné webové adresy jsou automaticky převedeny na odkazy. Bylo by vhodné umožnit uživatelům základní formátování textu (tučné písmo, kurzíva, podtržené, přeškrtnuté, odkaz) např. pomocí syntaxe Texy[7] nebo využitím některého volně dostupného WYSIWYG[8] editoru (např. TinyMCE[9]).
- **Upozorňování na nové příspěvky** – uživatel by měl mít možnost dostávat oznámení o tom, že byl vložen nový komentář. Mohlo by být vytvořeno několik variant: jednak odběr pomocí RSS, jednak informování pomocí školního e-mailu. Zajímavá varianta je taková, že by uživatel měl možnost označit určité přednášky či kurzy jako „sledované“, a po přihlášení by byl upozorněn na nové příspěvky v těchto přednáškách a kurzech.

[7] http://texy.info/
[8] What You See Is What You Get
[9] http://tinymce.moxiecode.com/

Kapitola 4

Administrační rozhraní

V současné implementaci Webového prohlížeče přednášek je nutné veškeré akce pracující s daty v databázi provádět v aplikaci phpMyAdmin[1]. Tento kvalitní nástroj umožňuje provádět veškeré potřebné databázové operace. Ale právě operace na úrovni jednotlivých tabulek nejsou pro administrátora webové aplikace příliš vhodné. Může udělat chybu při zadávání dat, pro jednu akci je nucen upravovat postupně více tabulek, práce s cizími klíči je neintuitivní apod.

Tyto problémy řeší administrační rozhraní, vytvořené přímo na míru Webového prohlížeče přednášek.

V sekci 4.1 jsou uvedeny požadavky na výslednou aplikaci. Sekce 4.2 se zabývá návrhem řešení administračního rozhraní a sekce 4.3 následnou implementací. V sekci 4.4 je provedeno zhodnocení dosaženého řešení.

4.1 Analýza požadavků

Administrační rozhraní má za úkol usnadnit práci administrátora, ne mu ji přidávat. To znamená, že má nabídnout jednodušší správu dat v prohlížeči, než jakou nabízí aplikace phpMyAdmin, a poskytovat jen funkce, které jsou využitelné při správě Webového prohlížeče přednášek. Hlavní požadavky jsou následující:

- Umožnit jednoduché provádění akcí „přidat“, „upravit“ a „odstranit“ nad všemi součástmi prohlížeče – přednáškami, kategoriemi, uživateli, dokumenty a komentáři.
- Zajistit integritu dat a definovat akce prováděné při operacích, které mohou integritu porušit. Např. při odstranění kategorie nezůstanou v databázi přednášky, které patří do této smazané kategorie.
- Kontrolovat správný formát zadávaných údajů, např. e-mailová nebo webová adresa uživatele, formát data a času apod.
- Definovat uživatelská oprávnění a před prováděním akcí tato oprávnění kontrolovat.

[1] http://www.phpmyadmin.net/home_page/

4.2 Návrh řešení

Jak bylo uvedeno v předchozí sekci, administrační rozhraní má zjednodušovat práci administrátora. Toho se docílí vytvořením jednoduchého uživatelského rozhraní a intuitivní prací s daty v databázi.

Aplikace bude víceuživatelská s právy definovanými pro každého uživatele. Bude tedy možné vytvořit různé uživatelské role, např.:

- Hlavní administrátor („Super User“), který spravuje všechna data.
- Autoři jednotlivých přednášek, kteří mohou pouze vkládat a upravovat své přednášky.
- Moderátor, který má oprávnění moderovat komentáře k přednáškám.
- Správce uživatelů, který může vytvářet nové uživatele, upravovat oprávnění, ale nemá přístup k jiným datům v databázi.

4.2.1 Případy užití

Případy užití pro administrační rozhraní jsou uvedeny na obrázku 4.1.

4.2.2 Bezpečnost hesla

Uložení hesla do databáze v takovém formátu, v jakém ho uživatel zadal, je extrémně nebezpečné a zneužitelné kýmkoliv, kdo získá přístup do databáze. Proto se běžně ukládá pouze hash hesla (např. MD5 či SHA1). Problém ovšem nastane, když si dva uživatelé zvolí náhodou stejné heslo, tedy i hashe těchto hesel budou stejné. Pokud jeden z uživatelů shodu hashů zjistí, může se přihlásit i jako druhý uživatel.

Proto se používá metoda, kdy se původní heslo smíchá s tzv. „solí“ – náhodně vygenerovaným řetězcem, a teprve poté se vypočítá hash celého tohoto řetězce. Do databáze se uloží hash a sůl. Při ověřování se zadané heslo smíchá se známou solí a z tohoto řetězce se vypočítá hash, který se porovná s uloženým hashem. Díky tomu nelze zjistit, zda dva uživatelé mají stejné heslo [12].

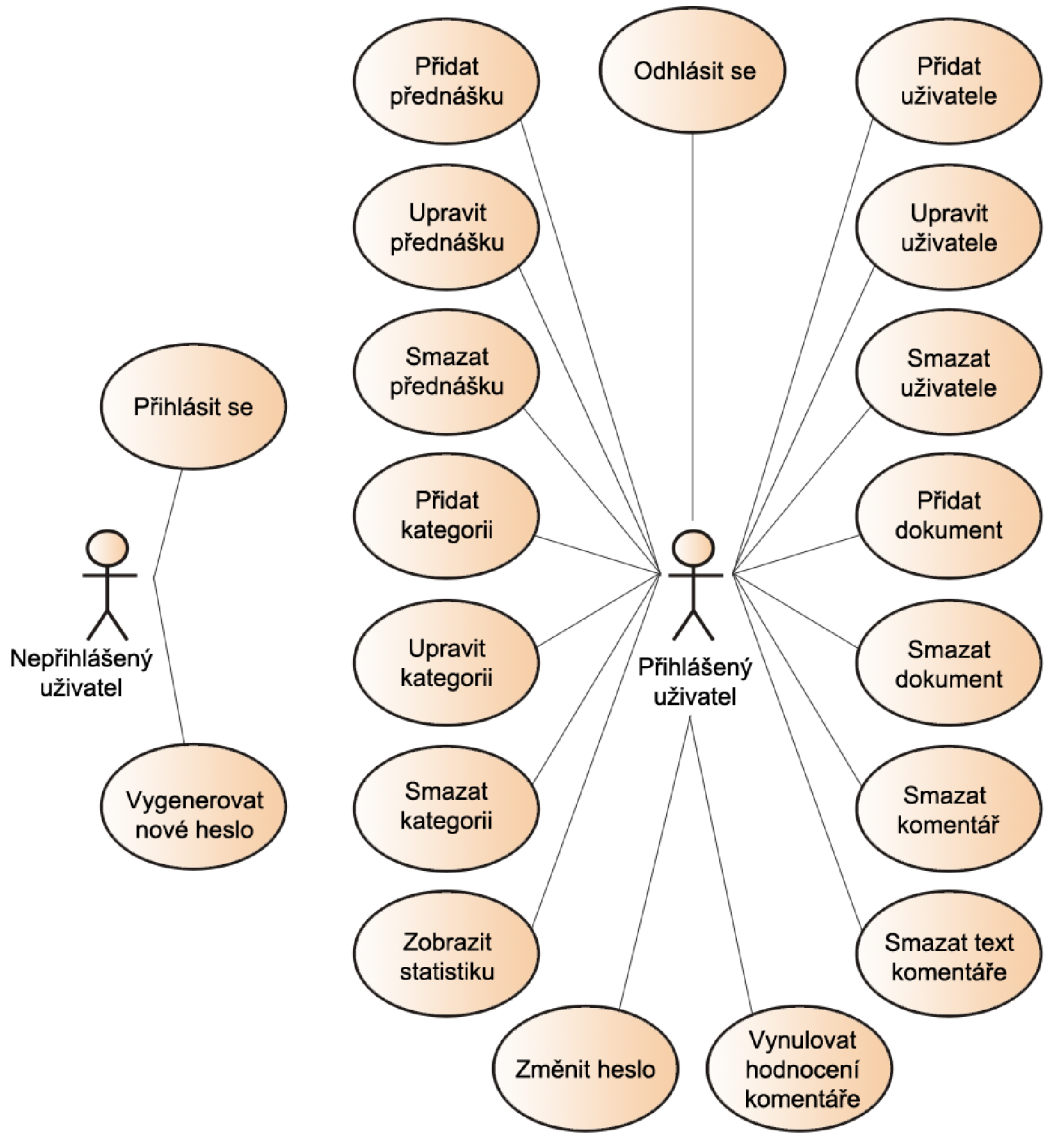
4.2.3 Oprávnění

Každý uživatel má nastavená oprávnění k provádění různých akcí. Jedná se o 8 příznaků, z nichž je jich aktuálně použito 5 a další jsou ponechány pro možné rozšíření. Podle ordinární hodnoty (počínaje od 1) mají příznaky vlastnosti, uvedené v tabulce 4.1.

Příznak	Hodnota příznaku			
	0	**1**	**2**	**3**
1: Přednášky	Prohlížet	Přidávat	Upravovat své	Upravovat všechny
2: Komentáře	Prohlížet	Přidávat	Mazat	Nulovat skóre
3: Uživatelé	Bez oprávnění	Přihlásit se	Upravovat uživatele	Upravovat oprávnění
4: Kategorie	Bez oprávnění	Upravovat své	Upravovat všechny	
5: Dokumenty	Bez oprávnění	Upravovat své	Upravovat vše	

Tabulka 4.1: Oprávnění uživatele

Vyšší oprávnění v sobě zahrnuje i všechny nižší hodnoty. Např. příznak *Uživatelé* s hodnotou 3 (Úprava oprávnění) znamená, že uživatel může nejen upravovat oprávnění ostatních



Obrázek 4.1: Případy užití

uživatelů, ale také jejich údaje (hodnota 2), a samozřejmě se může přihlásit do administrace (hodnota 1).

4.3 Implementace

Administrační rozhraní je vytvořeno jako webová aplikace. Je implementovaná v jazyce PHP a využívá databázi MySQL, stejně jako Webový prohlížeč přednášek. Aplikace je nezávislá, k jejímu spuštění a používání není nutné měnit zdrojové kódy Webového prohlížeče přednášek, pouze je potřeba upravit tabulku *User*.

4.3.1 Databáze

Struktura databáze je na obrázku 3.2. Oproti původnímu návrhu Jakuba Janoviče [5] byly přidány 3 nové tabulky. Původní tabulky jsou popsány v sekci 2.3, tabulky související s komentáři (*Comments* a *Votes*) v sekci 3.4.1. Tabulka *User* byla upravena tak, aby byla využitelná pro přihlášení do administračního rozhraní a pro kontrolu oprávnění. Tabulka obsahuje následující sloupce:

- **ID** – identifikátor uživatele a zároveň primární klíč.
- **name**, **url**, **email** – jméno uživatele, jeho webová a e-mailová adresa.
- **login** – přihlašovací jméno uživatele. Musí být unikátní.
- **password** – heslo uživatele zašifrované spolu se solí (atribut **salt** – viz sekce 4.2.2).
- **score** – karma uživatele, viz sekce 3.3.4.
- **permission** – oprávnění uživatele k operacím, viz sekce 4.2.3.

4.3.2 Případy užití

V této sekci je uveden popis jednotlivých případů užití. Před spuštěním každého případu užití jsou kontrolována oprávnění, uvedená v tabulce 4.1. Bez patřičného oprávnění není uživateli povoleno použití dané operace.

Přihlášení, odhlášení, zaslání nového hesla

Pro přihlášení uživatele slouží přihlašovací stránka. Na tuto stránku je nepřihlášený uživatel přesměrován při pokusu o přístup na kteroukoliv stránku administračního rozhraní, a také při odhlášení.

Pokud uživatel zapomene heslo, zadá své přihlašovací jméno a e-mailovou adresu, kterou má nastavenou ve svém profilu. Nově vygenerované heslo je mu zasláno na e-mail.

Přidání, úprava a smazání přednášky

Při vstupu do sekce *Přednášky* se uživateli zobrazí přednášky, které jsou uloženy v datovém adresáři Webového prohlížeče přednášek (např. adresář *data*), a které ještě nebyly uloženy do databáze. Uživatel vybere přednášky, které chce do databáze přidat, vyplní u nich popis a spustí proces importu. Tento proces využívá sadu funkcí *browser parser* Jakuba Janoviče [5] pro převod XML souborů do databáze MySQL.

Uživatel může upravovat údaje přednášky, jako je její název, popis, datum a čas, autor a zařazení do kategorie.

Při mazání přednášky si uživatel zvolí, zda chce smazat pouze data z databáze (a soubory přednášky zůstanou uloženy v datovém adresáři), nebo zda se mají smazat i tyto soubory.

Přidání, úprava a smazání kategorie

Při vytváření nové kategorie je nutné zadat její název, název adresáře, ve kterém se nachází jednotlivé přednášky, a zvolit autora této kategorie. Také je možné vybrat nadřazenou kategorii, takže lze vytvářet hierarchickou strukturu.

Všechny tyto údaje lze později upravit.

Při mazání kategorie je kontrolováno, zda obsahuje nějaké přednášky. Pokud ano, uživatel musí potvrdit, že chce tyto přednášky odstranit (podobně jako při mazání jednotlivých přednášek – pouze data v databázi, nebo i soubory). Přednášky nelze nechat nepřiřazené, je možné je postupně přesunout do jiné kategorie.

Přidání, úprava a smazání uživatele

Při přidávání nového uživatele je nutné zadat jeho jméno, přihlašovací jméno (login), heslo a oprávnění. Lze zadat své vlastní heslo a nebo zvolit náhodně vygenerované. Pokud přihlášený uživatel nemá právo na úpravu oprávnění jiných uživatelů, nový uživatel je vytvořen pouze se základními oprávněními, tzn. může nanejvýš prohlížet přednášky a přidávat komentáře. Takto vytvořený uživatel se nemůže přihlásit do administrace.

Podobně probíhá úprava uživatelských údajů. Přihlášený uživatel může změnit své heslo (je požadováno zadání starého hesla). Nikdo však z bezpečnostních důvodů nemůže měnit heslo jiného uživatele.

Smazání uživatele probíhá podobně, jako smazání kategorie. Pokud uživatel vlastní nějaké přednášky či kategorie, je vyžadováno potvrzení, zda mají být smazány i tyto přednášky a kategorie.

Přidání a smazání dokumentu

Dokument lze přidat buď do určité kategorie a nebo do konkrétní přednášky. V obou případech lze nahrát pouze dokument, jehož přípona je uvedena v tabulce *DocumentType*. Maximální velikost nahrávaného souboru je limitována nastavením serveru.

Při mazání je dokument odstraněn z databáze i fyzicky z datového adresáře.

Práce s komentáři

Pro přehlednost je zobrazení seznamu komentářů omezeno pouze na vybranou kategorii. Dále lze seznam komentářů omezit časově a lze nastavit společné kritérium, podle kterého jsou komentáře seskupovány (stejná přednáška, uživatelské jméno nebo datum).

Vymazat lze pouze takový komentář, na který neexistují žádné reakce. Pokud existují, je možné odstranit pouze jeho text. U všech komentářů lze vynulovat skóre. Po vymazání, odstranění textu a vynulování skóre je zároveň patřičně upravena karma autora komentáře (atribut *score* v tabulce *User*) a vyčištěny případné záznamy v tabulce *Votes*.

4.3.3 Uživatelské rozhraní

Uživatelské rozhraní bylo vytvořeno pouze jednoduché a proto může působit nepřívětivě. Veškeré kontroly zadávaných dat probíhají až na serveru. Proto se zde nabízí možnosti dalšího vylepšení, uvedené v sekci 4.4.1.

4.4 Zhodnocení řešení

Byla vytvořena aplikace, která usnadní práci administrátora Webového prohlížeče přednášek. Byly implementovány všechny důležité operace pro práci s přednáškami, kategoriemi, uživateli, dokumenty a komentáři. Podařilo se splnit hlavní požadavky na aplikaci, tj. jednoduchost, bezpečnost a zajišťování integrity dat. Pro uživatele jsou dostupná různá oprávnění, takže je možné vytvářet uživatelské role.

Aplikace je nezávislá na platformě. Byla testována a je plně funkční v těchto webových prohlížečích:

- Mozilla Firefox 3.6
- Opera 11
- Google Chrome 4
- Internet Explorer 8

Jelikož klientská část nepoužívá JavaScript, předpokládá se, že bude použitelná i ve starších webových prohlížečích.

4.4.1 Další rozšíření

Pokud by aplikace měla být nasazena do ostrého provozu, bylo by vhodné implementovat například tato rozšíření. Pro pohodlné použití jsou klíčové zejména první dva body:

- **Javascript** – kontroly zadávaných údajů v klientské části, vkládání data a času přes zobrazený kalendář, vytvoření „záložek“ – v každé sekci by jedna záložka sloužila pro vložení nových dat a druhá pro úpravu a mazání existujících dat, možnost filtrovat zobrazené seznamy přednášek, kategorií, autorů.
- **Vzhled aplikace** – vytvoření příjemnějšího vzhledu administračního rozhraní pomocí kaskádových stylů a upravením generovaného HTML kódu.
- **Možnost upravovat typy dokumentů** (tabulka *DocumentTypes*).
- **Propojení s aplikací pro identifikaci klíčových slov** (viz kapitola 5).

Kapitola 5

Automatická identifikace klíčových frází

Tato kapitola se zabývá rozšířením Webového prohlížeče přednášek o funkce, umožňující v přepisech přednášek automaticky identifikovat klíčové fráze.

Nabízí se 2 možnosti použití tohoto rozšíření:

- Klíčové fráze by byly zobrazeny v seznamu přednášek a umožnily by rychlý přehled o hlavních tématech přednášky.
- Transkript přednášky by byl provázán s klíčovými frázemi a byly by vytvořeny automatické odkazy na tyto fráze, např. na internetovou encyklopedii Wikipedia.

Sekce 5.1 se věnuje analýze problému a metodám identifikace klíčových slov a frází. V sekci 5.2 je popsána implementace těchto metod. V sekci 5.3 jsou prezentovány výsledky identifikace klíčových slov a frází a zhodnoceno dosažené řešení.

5.1 Analýza problému

5.1.1 Klíčové slovo, klíčová fráze

Za klíčové slovo můžeme považovat takové slovo, které popisuje obsah určitého dokumentu, obrázku, nahrávky apod. Jedna z definic klíčového slova zní: „Keyword is a word which occurs in a text more often than we would expect to occur by chance alone“ [11]. Jde tedy o slovo, které se v daném textu vyskytuje častěji než v ostatních textech. Stejně tak můžeme definovat klíčovou frázi jako dvě a více slov, která se (za sebou) v textu vyskytují častěji než v ostatních textech.

5.1.2 Metody identifikace klíčových slov

Existuje řada metod pro identifikaci klíčových slov a frází. Tyto metody můžeme rozdělit podle způsobu, jakým klíčová slova a fráze získávají [14]:

- metody založené na frekvenční analýze (např. TF-IDF, Branching Entropy)
- metody trénované na vzorových dokumentech se známými klíčovými slovy (např. KEA)

- metody založené současně na lingvistické a frekvenční analýze (např. Keyterm)

Pro tento projekt byly zvoleny metody založené na frekvenční analýze, z nichž dvě budou dále popsány.

TF-IDF

Tato sekce vychází z článku [14], kde je metoda TF-IDF použita k identifikaci klíčových slov na webových stránkách. S drobnými úpravami je však použitelná i na identifikaci klíčových slov v přepisech přednášek.

Metoda *TF-IDF* (Term Frequency–Inverse Document Frequency) je standardní metoda pro identifikaci klíčových slov. Slova, vyskytující se v jednom dokumentu častěji než v určité kolekci dokumentů, jsou označena jako klíčová. Výhodou této metody je, že porovnává jeden dokument s celou kolekcí a díky tomu jsou výsledky relevantnější. Procházení celé kolekce dokumentů (byť by byla indexovaná) je však časově náročné. Nevýhodou této metody je, že identifikuje pouze klíčová slova, nikoliv klíčové fráze (i když existují způsoby, jak toto omezení obejít).

Princip metody je popsán v těchto krocích.

1. Z dokumentů je odstraněna interpunkce a text je převeden na malá písmena.
2. Jsou odstraněna stop-slova, což jsou slova, která v daném jazyce nenesou žádnou významovou informaci [13].
3. Z každého slova je ponechán kořen a je aktualizován počet dokumentů, ve kterých se tento kořen vyskytuje.
4. Ve zkoumaném dokumentu j se vypočítá pro kořen slova i hodnota $w_{i,j}$ podle vzorce

$$w_{i,j} = \frac{n_{i,j}}{|p_j|} \cdot \log_2 \frac{N}{n_i}$$

   - kde $n_{i,j}$ je frekvence výskytu kořene slova i v dokumentu j,
   - $|p_j|$ je celkový počet slov v dokumentu j,
   - n_i je počet dokumentů obsahujících kořen slova i a
   - N je celkový počet dokumentů v kolekci.

5. Následně jsou všechny hodnoty $w_{i,j}$ normalizovány podle vzorce

$$W_{i,j} = \frac{w_{i,j}}{\sqrt{\sum_i {w_{i,j}}^2}}.$$

6. Slova s nejvyšší hodnotou $w_{i,j}$ můžeme považovat za klíčová. V [14] je uvedeno, že klíčových slov by mělo být cca 3–5, neboť to je nejčastější počet klíčových slov v technických článcích.

Branching entropy

Tato sekce vychází z článku [1].

Metoda Branching entropy slouží k identifikaci klíčových frází v přepisu mluveného slova. Je založena na poznatku, že za klíčové fráze můžeme považovat vzorky dvou a více slov, která se v textu vyskytují pohromadě častěji než jiné fráze. Pro každou frázi je vypočítán parametr zvaný *branching entropy* a podle jeho hodnoty je určeno, které fráze jsou klíčové. Výhodou této metody je její rychlost a úspěšnost nalezení klíčových frází v porovnání s jinými metodami.

Princip této metody je uveden v následujících krocích:

1. Z podřetězců v textu jsou vytvořeny tzv. *Sistrings*, což jsou částečně ohraničené řetězce. Příklad takovýchto řetězců je na obrázku 5.1.

2. Ze získaných *Sistrings* je vytvořen PAT-strom, datová struktura vycházející z trií a sloužící k ukládání řetězců. Příklad PAT-stromu je na obrázku 5.2.

3. Je vypočítána hodnota pravé entropie (*right branching entropy*) podle vzorce

$$p(x_i) = \frac{f_{x_i}}{f_X},$$

$$H_r(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i),$$

   - kde X je zkoumaná fráze (např. „hidden Markov“),
   - x_i je potomek této fráze v PAT-stromu (např. „hidden Markov model“),
   - f_{x_i} a f_X jsou četnosti výskytu X a x_i v textu a
   - n je počet různých potomků x_i fráze X.

4. Stejným způsobem je zkonstruován reverzní PAT-strom pomocí reverzních řetězců. Slova v řetězcích jsou v opačném pořadí (např. „model Markov hidden“). Následně je vypočítána hodnota levé entropie (*left branching entropy*) $H_l(\overline{X})$ pro reverzní fráze $\overline{X}$.

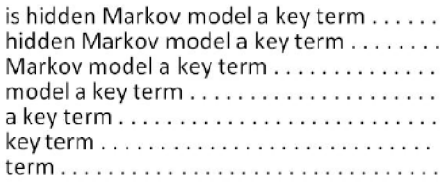
5. Po vypočítání hodnot $H_r(X)$ a $H_l(\overline{X})$ pro všechny fráze v textu je vypočítán aritmetický průměr všech hodnot $H_r(X)$ a $H_l(\overline{X})$.

6. Za klíčové fráze můžeme považovat ty fráze X, pro něž platí

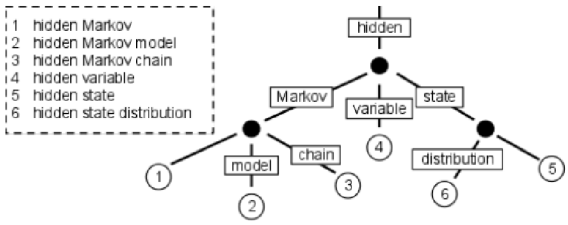
$$H_r(X) > \text{avg}(H_r) \wedge H_l(\overline{X}) > \text{avg}(H_l).$$

Pokud tedy je nějaká fráze klíčovou frází, pak je četnost jejího výskytu v textu vysoká a zároveň má vysoké hodnoty $H_r(X)$ a $H_l(\overline{X})$. To značí, že klíčová fráze je zleva i zprava obklopena mnoha různými slovy.

Příkladem může být fráze „hidden Markov model“. Tato fráze má vysoké $H_r(X)$ i $H_l(\overline{X})$, jelikož je obklopena mnoha různými slovy (např. „is“, „can“). Fráze „hidden Markov“ má ale nízké $H_r(X)$, jelikož je velmi často následována slovem „model“, ale málokdy jinými slovy.



Obrázek 5.1: Příklad tzv. *Sistrings*. Převzato z [1].



Obrázek 5.2: Příklad PAT-stromu. Převzato z [1].

5.1.3 Seznam stop-slov

V přednáškách se často opakují slova, která nenesou žádnou podstatnou informaci. Jde většinou o předložky, spojky, některá zájmena, a při identifikaci klíčových slov i slovesa. Proto se vytvářejí seznamy takovýchto slov [13].

U metody TF-IDF jsou v druhém kroku stop-slova odstraněna. V [1] není pro metodu Branching entropy uvedeno, zda jsou stop-slova odstraňována, nicméně výsledky uvedené v sekci 5.3 ukazují, že odstraňování stop-slov je i u této metody nutné.

5.2 Implementace metod

Byly implementovány dvě metody, uvedené v sekci 5.1.2 – TF-IDF pro identifikaci klíčových slov a Branching entropy pro identifikaci klíčových frází. Pro implementaci byl zvolen jazyk PHP, jelikož v tomto jazyce jsou implementována i ostatní rozšíření Webového prohlížeče přednášek. Identifikovaná klíčová slova a fráze jsou uložena v tabulkách databáze MySQL podobně, jako přepisy přednášek. Dalším důvodem pro využití kombinace PHP + MySQL je možnost začlenění funkcí pro identifikaci klíčových slov a frází přímo do administračního rozhraní (viz kapitola 4).

5.2.1 Seznam stop-slov

Pro obě implementované metody byl použit stejný seznam stop-slov.

V první fázi byly sloučeny 3 seznamy stop-slov z [8], [4] a [10]. Z výsledného seznamu byla odstraněna slova, která jsou považována za stop-slova na webových stránkách (např. „článek“, „design“, „email“), ale jako klíčová slova či ve frázích mohou být užitečná.

Následně byly spuštěny implementované metody na vybrané přednášky a do seznamu stop-slov byla přidána další slova, např. číslovky nebo výplňová slova.

V poslední fázi byl seznam stop-slov rozšířen o všechna česká slovesa a jejich tvary. Klíčová slova a fráze obsahující slovesa byly sice vyhodnoceny jako klíčové, ale nenesly žádné podstatné informace. Tato slovesa byla získána pomocí morfologického analyzátoru *ajka* [9].

5.2.2 TF-IDF

Metoda *TF-IDF* byla implementována tak, jak je popsáno v sekci 5.1.2, došlo pouze k drobným úpravám. Krok č. 1 mohl být vynechán, jelikož transkripty přednášek neobsahují interpunkci a velká písmena. V kroku č. 3 se nepracuje s kořenem slova, ale s variantou slova

s nejvyšším počtem výskytů v transkriptu. Důvodem je složitost českého jazyka a náročné určení kořene slova.

Vstupem je tabulka obsahující transkript přednášky (*transcript_<id>*) a dále seznam všech tabulek s transkripty. Aplikace si vytváří tyto pomocné tabulky:

- **processedTables** – obsahuje názvy všech tabulek s transkripty, které již byly zpracovány.
- **globalWordlist** – obsahuje seznam všech slov ve zpracovaných tabulkách a počty těchto slov.

Po spuštění aplikace zkontroluje, zda se v databázi nachází tabulky s transkripty, které ještě nebyly zpracovány. Pokud ano, jsou provedeny kroky 2 a 3, popsané v sekci 5.1.2. Pro zjednodušení nejsou do tabulky *globalWordlist* vkládány kořeny slov, ale celá slova.

Následuje zpracování požadovaného transkriptu. Ke každému slovu z transkriptu jsou vyhledána podobná slova, a podle nastaveného prahu je zkoumané slovo nahrazeno nejčastějším výskytem podobného slova. Např. pro slovo „vektorovými“ byla nalezena slova „vektorový“, „vektorovým“ a „vektorových“. Jelikož z těchto čtyř slov se slovo „vektorový“ v transkriptu vyskytuje nejčastěji, je původní slovo „vektorový“ nahrazeno tímto slovem a aktualizuje se počet jeho výskytů v dokumentu.

Poté je pro toto slovo vypočítána hodnota $w_{i,j}$ a dvojice (slovo i, hodnota $w_{i,j}$) je uložena do tabulky *transcript_<id>_keywords*. Požadovaný počet slov s nejvyšší hodnotou $w_{i,j}$ poté můžeme považovat za klíčová slova této přednášky.

5.2.3 Branching entropy

Metoda *Branching entropy* byla oproti popisu v sekci 5.1.2 výrazněji modifikována.

Bylo zjištěno (viz sekce 5.3.2), že v přepisech přednášek se vyskytují převážně dvouslovné a tříslovné klíčové fráze. Z tohoto důvodu mohla být metoda *Branching entropy* zjednodušena pouze na vyhledávání těchto frází a díky tomu odpadla nutnost vytvářet PAT-strom. Místo něj byla využita databáze MySQL a agregační funkce *GROUP BY*.

Princip identifikace klíčových frází je popsán v následujících krocích:

1. Vstupem je tabulka obsahující transkript přednášky (*transcript_<id>*). Nad touto tabulkou je volána funkce *branching_entropy()*, která vyhledá nejprve dvouslovné a poté tříslovné fráze.
   (a) Funkce si vytvoří dočasnou tabulku se sloupci *before*, *phrase*, *after*, *Hl* a *Hr*.
   (b) Z transkriptu přednášky je tato tabulka naplněna daty, přičemž *phrase* je dvou– nebo tříslovná fráze, *before* a *after* jsou slova, která této frázi předchází a následují ji. Současně jsou odstraněna stop-slova. Tento krok nahrazuje vytváření PAT-stromu.
   (c) Pro každou frázi ze sloupce *phrase* jsou vypočítány hodnoty $H_l(\overline{X})$ a $H_r(X)$ a pokud jsou větší než nastavený práh (typicky 0), jsou uloženy do příslušných sloupců *Hl* a *Hr*.
2. Z dočasné tabulky jsou vybrány ty fráze, jejichž $H_l(\overline{X})$ a $H_r(X)$ jsou větší než aritmetický průměr.

3. Pokud se ve výsledcích nacházejí podobné fráze, je fráze s menším součtem $H_l(\overline{X}) + H_r(X)$ odstraněna. Např. z frází „fourierova transformace“ a „fourierovou transformací“ měla první fráze větší $H_l(\overline{X}) + H_r(X)$, druhá fráze je proto odstraněna.

4. Pokud je dvouslovná fráze podřetězcem tříslovné fráze, je opět ta s menším $H_l(\overline{X}) + H_r(X)$ odstraněna. Příkladem mohou být fráze „objektově orientované“ a „objektově orientované programování“.

5. Takto vyfiltrované fráze můžeme považovat za klíčové.

5.3 Zhodnocení řešení

Tato sekce popisuje výsledky dvou implementovaných metod pro identifikaci klíčových frází.

Testy byly prováděny na transkriptech přednášek z bakalářských kurzů na FIT VUT z let 2007-2009 a magisterského kurzu Teoretická informatika. Celkem bylo pořízeno 380 přepisů přednášek. Průměrný počet slov v jedné přednášce je 13569, nejdelší přednáška má 22084 slov a nejkratší 2555 slov.

5.3.1 Časová náročnost

Měření časové náročnosti bylo prováděno na počítači s procesorem Intel Core 2 Duo s frekvencí 2.1 GHz a s operačním systémem Microsoft Windows 7 Professional (32bit). Byl použit webový server Apache 2.2.17 a databáze MySQL 5.5.8 v balíku *xampp*[1].

Byly vybrány přednášky, jejichž počet slov se blížil k průměrnému počtu 13569 slov.

TF-IDF

Pro metodu TF-IDF byla provedena dvě měření časové náročnosti. Prvním krokem této metody je vytvoření seznamu slov spolu s počtem transkriptů, ve kterých se tato slova vyskytují.

Časy potřebné k vytvoření tohoto seznamu jsou uvedeny v tabulce 5.1. S přibývajícím počtem záznamů v seznamu slov se čas potřebný ke zpracování jedné přednášky zvyšuje. To je způsobeno tím, že v tabulce *globalWordlist* přibývají záznamy, které je potřeba prohledat a aktualizovat při analýze nové přednášky.

Počet přednášek	Celkový čas [s]	Průměr na 1 přednášku [s]
1	0.24	0.24
5	6.24	1.25
10	20.30	2.03
100	792.10	7.92
380	6717.80	17.68

Tabulka 5.1: Čas potřebný pro vytvoření seznamu slov

V tabulce 5.2 jsou uvedeny časy potřebné k identifikaci klíčových slov. Se zvyšujícím se počtem zpracovávaných přednášek zůstává průměrný čas na jednu přednášku přibližně konstantní. Kolísání průměrného času je způsobeno rozdílným počtem slov v jednotlivých přednáškách a různým počtem odstraněných stop-slov.

[1] http://www.apachefriends.org/en/xampp.html

Počet přednášek	Celkový čas [s]	Průměr na 1 přednášku [s]
1	4.98	4.98
5	23.55	4.71
10	50.78	5.08
100	447.14	4.47

Tabulka 5.2: Čas pro identifikaci klíčových slov metodou TF-IDF

Celkový čas nutný pro identifikaci klíčových slov metodou TF-IDF je dán součtem času potřebného pro aktualizaci seznamu slov a času potřebného pro samotnou identifikaci. Se seznamem slov vytvořeným ze 380 přednášek je tento čas cca 22 s.

Branching entropy

Metoda Branching entropy nevytváří seznamy slov, pracuje pouze nad jedním transkriptem. V tabulce 5.3 jsou uvedeny časy nutné k identifikaci klíčových slov. Opět je vidět, že průměrný čas na jednu přednášku zůstává přibližně konstantní.

Počet přednášek	Celkový čas [s]	Průměr na 1 přednášku [s]
1	11.77	11.77
5	55.49	11.10
10	120.48	12.05
100	1085.87	10.86

Tabulka 5.3: Čas pro identifikaci klíčových slov metodou Branching entropy

5.3.2 Identifikovaná klíčová slova a fráze

V tabulkách 5.4 a 5.5 jsou uvedena klíčová slova identifikovaná metodou TF-IDF a klíčové fráze identifikované metodou Branching entropy ze tří přednášek – Signály a systémy (ISS), Základy počítačové grafiky (IZG) a Síťové aplikace a správa sítí (ISA).

ISS, 3.10.2008	IZG, 11.2.2009	IDS, 3.4.2009
impulsní	světlo	fronty
odezvy	barvy	paketů
impuls	grafika	služby
komplexní	vektorové	provoz
jednotkový	rastrové	kvalitu

Tabulka 5.4: Identifikovaná klíčová slova

Počet 5 slov/frází byl zvolen z důvodu, že v technickém článku se nejčastěji vyskytuje 3–5 klíčových slov [14].

Z výsledků je patrné, že jednoslovná klíčová slova jsou ve většině případů součástí víceslovné klíčové fráze (např. „impulsní“ – „impulsní odezvy“ nebo „fronty“ – „váhové fronty“). Dále jednoslovná klíčová slova nejsou sama o sobě příliš vypovídající a představu o tématech přednášky získáme až z víceslovných frází.

ISS, 3.10.2008	IZG, 11.2.2009	ISA, 27.11.2009
jednotkový impulz	vlnové délky	kvalitu služeb
diracův impuls	barevné modely	integrované služby
vstupní signál	základní barvy	maximální rychlost
impulsní odezvy	míchání barev	přenosové pásmo
komplexně sdružené	počítačová grafika	váhové fronty

Tabulka 5.5: Identifikované klíčové fráze

Z výsledků také vyplývá, že se v přednáškách nejčastěji vyskytují dvouslovné a v menší míře tříslovné klíčové fráze. Při hledání čtyř– a víceslovných klíčových frází byly nalezeny pouze dvou– a tříslovné fráze doplněné o nevýznamové slovo, které už frázi neupřesňovalo.

Hledání klíčových frází o délce dvou a tří slov tedy můžeme považovat za dostatečné a není třeba hledat víceslovné fráze, což by zpomalilo výpočet. Problémem může být situace, kdy dvouslovná fráze má vyšší váhu (součet $H_l(\overline{X}) + H_r(X)$) než tříslovná fráze, ale tříslovná fráze je subjektivně „kvalitnější“, tj. více vypovídá o tématu přednášky.

Hledání samostatných klíčových slov také nemá příliš velký význam, neboť takovéto slovo bývá součástí víceslovné fráze. Implementace metody TF-IDF byla proto využita pouze jako experiment.

Pro srovnání jsou v tabulce 5.6 uvedeny klíčové fráze identifikované metodou Branching entropy bez použití seznamu stop-slov. Jeho použití je tedy více než nutné.

ISS, 3.10.2008	IZG, 11.2.2009	ISA, 27.11.2009
na výstupu	je to	to znamená
na to	tak to	je to
to je	se to	se to
tady ten	to je	to je
je to	na to	tak to

Tabulka 5.6: Identifikované klíčové fráze bez použití seznamu stop-slov

5.3.3 Další rozšíření

Automatická identifikace klíčových slov by mohla být začleněna do Webového prohlížeče přednášek například těmito způsoby:

- **Odkazy z transkriptu** – nalezené klíčové fráze by byly automaticky vyhledávány na internetové encyklopedii Wikipedia, a pokud by pro frázi existovala příslušná stránka, byl by automaticky vytvořen odkaz v transkriptu přednášky. Po propojení s překladačem by bylo možné vyhledávat fráze i v angličtině.
- **Náhled na přednášku** – vybraný počet klíčových slov by byl zobrazen v seznamu přednášek (případně v seznamu kategorií) a uživatelé by měli přehled o hlavních tématech přednášky či kategorie.
- **Propojení s vyhledáváním v transkriptech** – při vyhledávání v transkriptech by byly uživateli prioritně nabídnuty přednášky, jejichž klíčové fráze by se shodovaly s vyhledávanou frází.

- **Hodnocení uživatelů** – uživatelé by mohli hlasovat pro nalezené klíčové fráze podle toho, jak jsou kvalitní. Uměle by se tak zvyšovalo či snižovalo ohodnocení dané fráze.
- **Označování stop-slov** – pokud by se v nalezených klíčových frázích vyskytovaly nekvalitní výsledky, uživatelé by měli možnost některá slova označit jako stop-slova. Na základě určitého počtu označení by bylo toto slovo zařazeno do seznamu stop-slov.

Kapitola 6

Závěr

Tato práce rozšiřuje současnou implementaci Webového prohlížeče přednášek o další komponenty. Prohlížeč umožňuje přehrávání záznamů z přednášek, prohlížení „slajdů“ k přednáškám, vyhledávání v přepisu řeči a jeho zobrazování přímo ve videu, zobrazení dalších odkazů k přednáškám a odkazů na podobné přednášky. Spojením těchto prvků slouží Webový prohlížeč přednášek jako komplexní výukový systém.

Výsledkem této práce jsou tři rozšíření Webového prohlížeče přednášek:

- **Možnost přidávat komentáře k přednáškám**. Tato komponenta je podrobně popsána v kapitole 3. Mezi hlavní vlastnosti patří vkládání komentářů přiřazených k určitému času přednášky i k celé přednášce. Komentáře je možné hodnotit, podle získaného hodnocení jsou kvalitní komentáře zvýrazňovány a nekvalitní potlačovány. Komentáře jsou synchronizovány s ostatními komponentami Webového prohlížeče přednášek. Při přehrávání videa jsou automaticky zvýrazňovány komentáře, které jsou přiřazené k určitému času přednášky. Kliknutím na časový údaj komentáře či na časový údaj v textu komentáře dojde k synchronizaci slajdů, transkriptu a videa. Modul pro přidávání komentářů je implementován v jazyce JavaScript, se serverem komunikuje pomocí technologie AJAX, serverová část je implementována v jazyce PHP a data jsou ukládána do databáze MySQL.

- **Administrační rozhraní** je popsáno v kapitole 4. Tato aplikace umožňuje administrátorovi přidávat, upravovat a odstraňovat přednášky, kategorie, uživatele a dokumenty a spravovat komentáře. Byla navržena tak, aby její použití bylo možné bez složitého návodu a speciálních znalostí. Správa prohlížeče je díky tomuto rozšíření jednodušší, než při použití aplikace phpMyAdmin. Administrační rozhraní je implementováno v jazyce PHP a pracuje s databází MySQL.

- **Funkce pro automatickou identifikaci klíčových frází** popisuje kapitola 5. Toto rozšíření umožňuje automaticky vyhledávat klíčové fráze v transkriptech přednášek. Dosažené výsledky ukazují, že identifikované klíčové fráze poměrně dobře vystihují hlavní témata přednášek, proto může být toto rozšíření začleněno do Webového prohlížeče přednášek. Klíčové fráze mohou sloužit jako náhled na témata přednášky či automatické vytváření odkazů na internetovou encyklopedii Wikipedia. Byly implementovány dvě metody – *TF-IDF* pro identifikaci samostatných klíčových slov a *Branching entropy* pro identifikaci klíčových frází. Po porovnání výsledků byla pro další použití zvolena metoda *Branching entropy*.

Všechna rozšíření jsou ve stavu, kdy by mohla být použita v reálném provozu Webového prohlížeče přednášek, nicméně u všech se nabízí další možnosti jejich vylepšení. Návrhy vylepšení jsou uvedeny v závěrech jednotlivých kapitol, mezi hlavní rozšíření patří zejména:

- Formátování textu při vkládání komentáře např. pomocí syntaxe Texy nebo volně dostupných WYSIWYG editorů.
- Upozorňování na nově přidané komentáře pomocí e-mailu, RSS nebo jiným způsobem.
- Vylepšení administračního rozhraní o kontrolu zadávaných údajů pomocí JavaScriptu, vytvoření lepší grafické podoby uživatelského rozhraní.
- Automatické vyhledání identifikovaných klíčových frázích na internetové encyklopedii Wikipedia a vytvoření odkazů na nalezená hesla.
- Vkládání identifikovaných klíčových frází do seznamu přednášek a/nebo kategorií jakožto rychlý přehled o tématech přednášky/kategorie.
- Uživatelské hodnocení kvality identifikovaných klíčových frází, označování chybně identifikovaných frází, označování stop-slov.

Práce vychází především z diplomové práce Ing. Josefa Žižky, který vytvořil základní verzi prohlížeče a nadále na něm pracuje, a z diplomové práce Ing. Jakuba Janoviče, zabývající se převodem prohlížeče do databáze MySQL.

Webový prohlížeč přednášek je průběžně aktualizován a je volně dostupný k vyzkoušení[1]. Doufejme, že jeho vývoj bude dále pokračovat. Na přiloženém DVD jsou k dispozici zdrojové soubory prohlížeče a vytvořených rozšíření, návod k instalaci a také elektronická podoba této práce.

[1] http://www.prednasky.com/

Literatura

[1] Chen, Y.-N.; Huang, Y.; Kong, S.-Y.; aj.: Automatic Key Term Extraction from Spoken Course Lectures Using Branching Entropy and Prosodic/Semantic Features. In *Proceedings of SLT 2010*, Berkeley, California, USA, 2010.

[2] Findicons.com: Systém vyhledávání ikon. [online]. [cit. 2011-03-31].
URL http://findicons.com/

[3] Heller, P.: Strukturovaná diskuse pod články - praxe. [online]. [cit. 2011-04-05].
URL http://interval.cz/clanky/strukturovana-diskuse-pod-clanky-praxe/

[4] Janík, M.: Česká stop slova (stop words). 2009, [online]. [cit. 2011-04-06].
URL http://www.michaljanik.cz/stop-slova

[5] Janovič, J.: *Webový prohlížeč audio/video záznamů přednášek: převod prohlížeče na MySQL databázi.* Diplomová práce, FIT VUT v Brně, 2010.

[6] Lighttpd fly light. [online]. [cit. 2011-03-29].
URL http://www.lighttpd.net/

[7] LongTail Video. [online]. [cit. 2011-03-29].
URL http://www.longtailvideo.com/

[8] Ranks.nl: Czech Stopwords. [online]. [cit. 2011-04-06].
URL http://www.ranks.nl/stopwords/czech.html

[9] Sedláček, R.; Smrž, P.: A New Czech Morphological Analyser ajka. In *Proceedings of the 4th International Conference on Text, Speech and Dialogue*, TSD '01, London, UK: Springer-Verlag, 2001, ISBN 3-540-42557-8, s. 100–107.
URL http://portal.acm.org/citation.cfm?id=647239.718437

[10] SeoServis: Stopwords. [online]. [cit. 2011-04-06].
URL http://seo-servis.cz/libs/stopwords.txt.cz

[11] Tribble, C.; Scott, M.: *Textual Patterns: Key words and corpus analysis in language education.* Philadelphia: John. Benjamins, 2006, str. 203.

[12] Vrána, J.: Ukládání hesel. [online]. [cit. 2011-04-14].
URL http://php.vrana.cz/ukladani-hesel.php

[13] Wikipedie: Stopslovo. 2010, [online]. [cit. 2011-04-05].
URL http://cs.wikipedia.org/w/index.php?title=Stopslovo&oldid=5731713

[14] Zhang, Y.; Zincir-Heywood, N.; Milios, E.: Narrative text classification for automatic key phrase extraction in web document corpora. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, WIDM '05, New York, NY, USA: ACM, 2005, ISBN 1-59593-194-5, s. 51–58.

[15] Žižka, J.; Černocký, J.; Fapšo, M.; aj.: Web-Based Lecture Browser with Speech Search. In *Znalosti 2010*, Sborník příspěvků 9. ročníku konference, Faculty of management and information, 2010, ISBN 978-80-245-1636-3, s. 287–290.
URL http://www.fit.vutbr.cz/research/view_pub.php?id=9229

[16] Žižka, J.: *Webový prohlížeč přednášek*. Diplomová práce, FIT VUT v Brně, 2009.

Příloha A

Obsah CD

- **source** – zdrojové soubory Webového prohlížeče přednášek od Ing. Josefa Žižky a Ing. Jakuba Janoviče. Zdrojové soubory implementovaných rozšíření jsou umístěny v těchto podadresářích:
  - **admin** – zdrojové soubory administračního rozhraní
  - **comments** – zdrojové soubory modulu pro přidávání komentářů
  - **keyphrases** – zdrojové soubory pro detekci klíčových frází
- **documents** – dokumentace
  - **tex** – zdrojové soubory této práce
  - **technicka_zprava.pdf** – elektronická verze této práce
- **articles** – archiv použitých on-line zdrojů
- **video** – videoprezentace modulu pro přidávání komentářů
- **README** – informace o rozšířeních a návod na instalaci