Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-5220-47879

Bc. Róbert Móro

PERSONALIZOVANÁ SUMARIZÁCIA TEXTU

Diplomová práca

Študijný program: Softvérové inžinierstvo
Študijný odbor: 9.2.5 Softvérové inžinierstvo
Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU v Bratislave
Vedúca práce: prof. Ing. Mária Bieliková, PhD.

máj 2012

S T U
F I I T
SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
Fakulta informatiky a informačných technológií

Zadanie diplomovej práce

Meno študenta: **Bc. Móro Róbert**

Študijný program: Softvérové inžinierstvo

Študijný odbor: Softvérové inžinierstvo

Názov práce: **Personalizovaná sumarizácia textu**

Samostatnou výskumnou a vývojovou činnosťou v rámci predmetov Diplomový projekt I, II, III vypracujte diplomovú prácu na tému, vyjadrenú vyššie uvedeným názvom tak, aby ste dosiahli tieto ciele:

Všeobecný cieľ:

Vypracovaním diplomovej práce preukážte, ako ste si osvojili metódy a postupy riešenia relatívne rozsiahlych projektov, schopnosť samostatne a tvorivo riešiť zložité úlohy aj výskumného charakteru v súlade so súčasnými metódami a postupmi študovaného odboru využívanými v príslušnej oblasti a schopnosť samostatne, tvorivo a kriticky pristupovať k analýze možných riešení a k tvorbe modelov.

Špecifický cieľ:

Vytvorte riešenie, zodpovedúce návrhu textu zadania, ktorý je prílohou tohto zadania. Návrh bližšie opisuje tému vyjadrenú názvom. Tento opis je záväzný, má však rámcový charakter, aby vznikol dostatočný priestor pre Vašu tvorivosť.

Riaďte sa pokynmi Vášho vedúceho.

Pokiaľ v priebehu riešenia, opierajúc sa o hlbšie poznanie súčasného stavu v príslušnej oblasti alebo o priebežné výsledky Vášho riešenia alebo o iné závažné skutočnosti, dospejete spoločne s Vaším vedúcim k presvedčeniu, že niečo v texte zadania a/alebo v názve by sa malo zmeniť, navrhnite zmenu. Zmena je spravidla možná len pri dosiahnutí kontrolného bodu.

Miesto vypracovania: Ústav informatiky a softvérového inžinierstva FIIT STU v Bratislave

Vedúci práce: **prof. Ing. Mária Bieliková, PhD.**

Termíny odovzdania:

podľa harmonogramu štúdia platného pre semester, v ktorom máte príslušný predmet (Diplomový projekt I, II, III) absolvovať podľa Vášho študijného plánu

Predmety odovzdania:

V každom predmete dokument podľa pokynov na www.fiit.stuba.sk v časti: home > Informácie o > štúdiu > organizácia štúdia > diplomový projekt

V Bratislave dňa 14. 2. 2011

prof. Ing. Pavol Návrat, PhD.
riaditeľ Ústavu informatiky a softvérového inžinierstva

S T U
F I I T
SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
Fakulta informatiky a informačných technológií

Návrh zadania diplomovej práce

Finálna verzia do diplomovej práce[1]

Študent

Meno, priezvisko, tituly:	Róbert Móro, Bc.
Študijný program:	Softvérové inžinierstvo
Kontakt:	robo.moro@gmail.com

Výskumník:

Meno, priezvisko, tituly:	Mária Bieliková, prof. Ing. PhD.

Projekt:

Názov:	Personalizovaná sumarizácia textu
Miesto vypracovania:	Ústav informatiky a softvérového inžinierstva, FIIT STU, Bratislava
Oblasť problematiky[2]**:**	Personalizovaný web, webové systémy, spracovanie textu

Text zadania

Jeden z najvážnejších problémov súčasného webu je zahltenie informáciami. Personalizovaný web má za cieľ riešiť tento problém tým, že sa snaží prispôsobovať charakteristikám jednotlivých používateľov pri prezentácii obsahu ako aj pri navigácii a odporúčaní. Umožňuje tiež používateľom prispôsobovať existujúci a vytvárať nový obsah. Dôležitý je aj sociálny aspekt a kolaborácia používateľov pre dosiahnutie spoločných cieľov.

Analyzujte metódy prispievania a prispôsobovania obsahu na webe, zamerajte sa pritom na poznámkovanie a jeho využitie pri kolaborácii používateľov. V tejto súvislosti analyzujte aj metódy automatickej sumarizácie textu ako jedného zo spôsobov riešenia problému zahltenia informáciami.

Navrhnite metódu automatickej sumarizácie textu využívajúcej poznámkovanie textov používateľmi, ktorá sa bude prispôsobovať individuálnemu používateľovi (čitateľovi), teda bude zohľadňovať jeho ciele, záujmy a znalosti reprezentované modelom používateľa. Zohľadnite možnosť spolupráce používateľov pri poznámkovaní v podobe zdieľania poznámok navzájom, resp. v rámci skupiny používateľov s podobnými charakteristikami. Uvažujte rozšírenie metódy pre viac-dokumentovú sumarizáciu. Navrhnutú metódu experimentálne overte vo vybranej doméne (napr. výučba) a dosiahnuté výsledky porovnajte s klasickými prístupmi k sumarizácii textu, ktoré nevyužívajú personalizáciu.

150-200 slov, ktoré opisujú výskumný problém v kontexte súčasného stavu vrátane motivácie a smerov riešenia

[1] Veľkosť jednotlivých polí pre vypĺňanie nemožno meniť. Vytlačiť obojstranne na jeden list papiera.

[2] Identifikácia oblasti v rámci odboru štúdia, na ktorú sa projekt primárne viaže

Literatúra

- Zhang, H., Ma, Z.C., Cai, Q.: A study for documents summarization based on personal annotation. In: Proceedings of the HLT-NAACL Workshop on Text summarization, Association for Computational Linguistics Morristown, NJ, USA, pp. 41-48 (2003).
- Campana, M., Tombros, A.: Incremental Personalised Summarisation with Novelty Detection. In: FQAS '09 Proceedings of the 8th International Conference on Flexible Query Answering Systems, Springer-Verlag Berlin, Heidelberg, pp. 641-652 (2009).

2-3 vedecké zdroje, každý na samostatnom riadku a s údajmi zodpovedajúcimi bibliografickým odkazom podľa normy STN ISO 690, ktoré sa viažu k téme zadania a preukazujú výskumnú povahu problému a jeho aktuálnosť (uveďte všetky potrebné údaje na identifikáciu zdroja, pričom uprednostnite vedecké príspevky v časopisoch a medzinárodných konferenciách)

Vyššie je uvedený návrh diplomového projektu, ktorý vypracoval Bc. Róbert Móro, konzultovala a osvojila si ho prof. Ing. Mária Bieliková, PhD. a súhlasí, že bude takýto projekt viesť v prípade, že bude pridelený tomuto študentovi.

V BRATISLAVE dňa 25. 1. 2011

Móro
Podpis študenta

Podpis výskumníka

Vyjadrenie garanta predmetov Diplomový projekt I, II, III

Návrh zadania schválený: áno / ~~nie~~[3]

Dňa: 31. 1. 2011

Podpis garanta predmetov

[3] Nehodiace sa prečiarknite

S T U
F I I T
SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
Fakulta informatiky a informačných technológií

Návrh zadania diplomovej práce

Revízia č.[1]*: ..1.....*

Študent

Meno, priezvisko, tituly:	Róbert Móro, Bc.
Študijný program:	Softvérové inžinierstvo
Kontakt:	robo.moro@gmail.com

Výskumník:

Meno, priezvisko, tituly:	Mária Bieliková, prof. Ing. PhD.

Projekt:

Názov:	Personalizovaná sumarizácia textu
Miesto vypracovania:	Ústav informatiky a softvérového inžinierstva, FIIT STU, Bratislava
Oblasť problematiky[2]**:**	Personalizovaný web, webové systémy, spracovanie textu

Text zadania

Jedným z najvážnejších problémov súčasného webu je zahltenie informáciami. Automatická sumarizácia textu pristupuje k riešeniu tohto problému tak, že vyberá z dokumentu najdôležitejšie informácie, ktoré používateľom môžu pomôcť pri rozhodovaní, či je daný dokument pre nich relevantný, alebo nie. Iný prístup predstavuje personalizovaný web, ktorý sa snaží prispôsobovať charakteristikám jednotlivých používateľov pri prezentácii obsahu ako aj pri navigácii a odporúčaní. Umožňuje tiež používateľom prispôsobovať existujúci a vytvárať nový obsah.

Analyzujte metódy automatickej sumarizácie textu ako jeden zo spôsobov riešenia problému zahltenia informáciami. Klasické metódy sumarizácie vychádzajú len zo základného obsahu, uvažujte však aj ďalšie indikátory významnosti častí obsahu a záujmu používateľov, akými sú napríklad aktivita používateľov alebo ich interakcia s obsahom v podobe poznámkovania dokumentov. Navrhnite metódu automatickej sumarizácie textu, ktorá sa bude prispôsobovať individuálnemu používateľovi (čitateľovi), teda bude zohľadňovať jeho ciele, záujmy, prípadne znalosti reprezentované modelom používateľa. Uvažujte rozšírenie metódy pre viac-dokumentovú sumarizáciu. Navrhnutú metódu experimentálne overte vo vybranej doméne (napr. výučba) a dosiahnuté výsledky porovnajte s klasickými prístupmi k sumarizácii textu, ktoré nevyužívajú personalizáciu.

150-200 slov, ktoré opisujú výskumný problém v kontexte súčasného stavu vrátane motivácie a smerov riešenia

[1] Uvedie sa poradové číslo revízie

[2] Identifikácia oblasti v rámci odboru štúdia, na ktorú sa projekt primárne viaže

Literatúra

- Zhang, H., Ma, Z.C., Cai, Q.: A study for documents summarization based on personal annotation. In: Proceedings of the HLT-NAACL Workshop on Text summarization, Association for Computational Linguistics Morristown, NJ, USA, pp. 41-48 (2003).
- Campana, M., Tombros, A.: Incremental Personalised Summarisation with Novelty Detection. In: FQAS '09 Proceedings of the 8th International Conference on Flexible Query Answering Systems, Springer-Verlag Berlin, Heidelberg, pp. 641-652 (2009).

2-3 vedecké zdroje, každý na samostatnom riadku a s údajmi zodpovedajúcimi bibliografickým odkazom podľa normy STN ISO 690, ktoré sa viažu k téme zadania a preukazujú výskumnú povahu problému a jeho aktuálnosť (uveďte všetky potrebné údaje na identifikáciu zdroja, pričom uprednostnite vedecké príspevky v časopisoch a medzinárodných konferenciách)

Vyššie je uvedený návrh diplomového projektu, ktorý vypracoval Bc. Róbert Móro, konzultovala a osvojila si ho prof. Ing. Mária Bieliková, PhD. a súhlasí, že bude takýto projekt viesť.

V BRATISLAVE dňa 1.9.2011

Móro
Podpis študenta

Podpis výskumníka

Vyjadrenie garanta predmetov Diplomový projekt I, II, III

Návrh zadania schválený: áno / ~~nie~~[3]

Dňa: 2.9.2011

Podpis garanta predmetov

[3] Nehodiace sa prečiarknite

ANOTÁCIA

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: SOFTVÉROVÉ INŽINIERSTVO

Autor: Bc. Róbert Móro
Diplomová práca: Personalizovaná sumarizácia textu
Vedúca diplomovej práce: prof. Ing. Mária Bieliková, PhD.
máj, 2012

Automatická sumarizácia textu má za cieľ riešiť problém zahltenia informáciami pomocou extrakcie najdôležitejších informácií z dokumentu, ktoré môžu používateľom pomôcť rozhodnúť sa, či je daný dokument pre nich relevantný a mali by si ho prečítať celý, alebo nie.

Navrhli sme metódu personalizovanej sumarizácie, ktorá na rozdiel od konvenčných metód zohľadňuje rozdiely v charakteristikách používateľov. Náš prínos spočíva predovšetkým v návrhu konkrétnych hodnotičov, ale aj v spôsobe ich kombinácie, ktorý umožňuje uvažovať rôzne parametre alebo kontext sumarizácie. Zamerali sme sa na sumarizáciu pre opakovanie v doméne výučby. Pre tento účel sme tiež navrhli metódu personalizovaného výberu dokumentov na opakovanie.

Overenie sme realizovali v prostredí výučbového systému ALEF (Adaptive Learning Framework). Výsledky uskutočnených experimentov naznačujú, že zohľadnenie relevantných doménových pojmov, ako aj poznámok v procese sumarizácie prináša zlepšenie oproti generickému variantu a výsledné sumarizácie sú schopné zhrnúť dôležité koncepty obsiahnuté v dokumentoch aj pre účely opakovania.

ANNOTATION

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: SOFTWARE ENGINEERING

Author: Bc. Róbert Móro
Master's Thesis: Personalized Text Summarization
Supervisor: prof. Ing. Mária Bieliková, PhD.
2012, May

Automatic text summarization aims to address the information overload problem by extracting the most important information from a document, which can help users to decide, whether it is relevant for them and they should read the whole text or not.

In this work, we have proposed a method of personalized text summarization, which unlike the conventional methods takes into account differences in users' characteristics. Our contribution lies in the proposal of the specific raters and the method of their combination which allows considering various parameters or context of the summarization. We have focused on summarization for knowledge revision in the domain of learning. For this purpose, we have also proposed a method of personalized selection of documents for revision.

We have evaluated the proposed method of summarization in learning system ALEF (Adaptive Learning Framework). The results of our experiments suggest that considering the domain relevant terms as well as annotations in the process of summarization improves the generic variant and the resulting summaries are capable of extracting important concepts explained in the documents even for revision.

Obsah

1 Úvod

Jeden z najväčších problémov súčasného webu je zahltenie informáciami. V stave, keď na webe môžeme nájsť takmer všetko, sa stalo problémom nájsť to, čo by sme naozaj chceli, alebo potrebovali – nájsť relevantné informácie. Samotný pojem relevantné informácie je však subjektívny, pretože ako používatelia webu, líšime sa vo svojich záujmoch, cieľoch či znalostiach.

Automatická sumarizácia textu má za cieľ riešiť tento problém. Jej hlavnou myšlienkou je extrahovať najdôležitejšie informácie z dokumentu, ktoré majú používateľom napomôcť pri rozhodovaní, či je daný dokument pre nich relevantný a mali by si ho prečítať celý, alebo nie. Nevýhodou konvenčných (generických) sumarizačných metód však je, že neberú do úvahy spomínané rozdiely v záujmoch, cieľoch či znalostiach používateľov a tiež vychádzajú len zo základného obsahu, ktorý sa sumarizuje, teda neuvažujú napr. aktivitu používateľov, ktorá môže indikovať významné časti obsahu.

Iný prístup k riešeniu problému zahltenia informáciami predstavuje koncept personalizovaného (adaptívneho) webu, ktorý je alternatívou k tradičnému prístupu *jedna veľkosť pre všetkých*. Snahou je rozpoznať charakteristiky používateľa a prispôsobiť sa im, t.j. prispôsobiť obsah, ktorý je prezentovaný používateľovi, alebo prispôsobiť spôsob, akým používateľ k obsahu pristupuje.

Hlavným cieľom tejto práce je spojiť oba prístupy k riešeniu problému zahltenia informáciami – automatickú sumarizáciu textu a koncept personalizovaného webu. Navrhli sme metódu personalizovanej sumarizácie textu, ktorá extrahuje pre používateľa najdôležitejšie a najzaujímavejšie informácie. Okrem toho je nami navrhnutá metóda nezávislá od jazyka sumarizovaného textu a tiež od domény tak, aby ju bolo možné použiť v prostredí otvoreného webu.

Na overenie metódy sme si zvolili doménu výučby, v rámci ktorej sme ju realizovali v prostredí výučbového systému ALEF. Používatelia – študenti – pri učení sa vo výučbovom systéme preštudujú veľké množstvo vzdelávacích materiálov. Jedným z možných scenárov použitia sumarizácie v takomto prostredí je jej využitie pri opakovaní; práve na tento scenár sme sa v ďalšej práci zamerali.

Pri opakovaní však musíme zohľadniť aj ďalšie aspekty, predovšetkým čas opakovania a spôsob výberu dokumentov na opakovanie. Navrhli sme preto metódu personalizovaného výberu dokumentov na opakovanie, ktorá okrem iného zohľadňuje zmenu vedomostí používateľa v čase.

V nasledujúcich kapitolách sa postupne venujeme analýze problematiky. Skúmame metódy sumarizácie s dôrazom na existujúce prístupy jej personalizácie (kap. 2). V kapitole 3 uvádzame prehľad voľne dostupných riešení sumarizátorov s porovnaním ich vlastností. V kapitole 4 uvažujeme možnosti prispôsobovania sumarizácie pri zohľadnení konceptov, vedomostí či poznámok používateľov; skúmame tiež používanie rôznych typov poznámok a ich vhodnosť pre prispôsobovanie sumarizácie na reálnych používateľských dátach. Okrem toho v tejto kapitole analyzujeme špecifiká sumarizácie pre opakovanie v doméne výučby.

Vlastný príspevok opisujeme v kapitolách 5, 6 a 7. V kapitole 5 formulujeme návrh metódy personalizovanej sumarizácie založenej na kombinácii rôznych hodnotičov a podrobne

diskutujeme jednotlivé navrhnuté hodnotiče. V kapitole 6 rozoberáme uplatnenie sumarizácie v konkrétnom scenári použitia pre opakovanie v doméne výučby a navrhujeme personalizovanú metódu výberu dokumentov na opakovanie. Realizáciu navrhnutej metódy sumarizácie a jej experimentálne overenie opisujeme v kapitole 7.

Posledná kapitola (kap. 8) predstavuje zhodnotenie dosiahnutých výsledkov, prínosov našej práce a možností ďalšieho smerovania.

Prílohy tvorí dokumentácia k overeniu, technická dokumentácia, inštalačná a používateľská príručka a obsah priloženého elektronického média. Okrem toho prikladáme príspevok prijatý na medzinárodný workshop o textovom vyhľadávaní informácií TIR 2012 organizovaný v rámci konferencie DEXA 2012.

2 Automatická sumarizácia textu

Hlavnou motiváciou, prečo sa zaoberať automatickou sumarizáciou textu, je zahltenie informáciami, ktorému sme v súčasnosti (nielen) na webe vystavení. Práve sumarizácia textu je jedna z metód, ktorá pomáha používateľom rýchlo sa zorientovať v množstve zdrojov na webe a rozhodnúť sa, či je daný dokument pre nich relevantný a majú si ho prečítať celý, alebo nie.

Proces sumarizácie je definovaný ako „*vytvorenie stručnej a presnej reprezentácie obsahu dokumentu“*, resp. ako *„vyňatie najdôležitejšej informácie zo zdrojového textu, ktorá ho zostručňuje pre účely a úlohy používateľa*“ (Ježek & Steinberger, 2010). V literatúre sa však stretneme aj s inou, voľnejšou definíciou, ktorá za sumarizáciu považuje jednoducho „*text, ktorý je vytvorený z jedného alebo viacerých textov, a ktorý vyjadruje dôležité informácie z pôvodného textu (textov) a nie je dlhší ako polovica pôvodného textu (textov)*“ (Radev et al., 2002), pričom text v definícii by sme mohli nahradiť za reč, multimediálne dokumenty, hypertext a pod.

Snahou sumarizácie je teda zhrnúť na čo najmenšom priestore čo najviac najvýznamnejších tém obsiahnutých v dokumente, pričom je dôležitá aj čitateľnosť, zrozumiteľnosť a súdržnosť výsledného súhrnu.

Rozlišujeme viacero typov sumarizátorov podľa rôznych navzájom nezávislých hľadísk (Ježek & Steinberger, 2008, 2010); z pohľadu našej práce sú najdôležitejšie tieto tri:

- Forma sumarizácie
- Účel sumarizácie
- Zameranie sumarizácie

Podľa formy výslednej sumarizácie rozlišujeme, či ide o:

- *Extrakt* – výslednú sumarizáciu tvoria priamo vety vybrané (extrahované) zo sumarizovaného textu
- *Abstrakt* – výsledná sumarizácia nemusí obsahovať vety zo sumarizovaného textu, ale môže ich parafrázovať, použiť všeobecnejší, alebo naopak špecifickejší pojem, prípadne synonymum, spojiť niektoré vety dohromady, niektoré naopak rozdeliť, zmeniť ich poradie a pod; vyžaduje si preto pokročilejšie metódy spracovania prirodzeného jazyka a hlbšiu syntaktickú aj sémantickú analýzu textu

Z hľadiska účelu, na ktorý sa sumarizácia vytvára, môžeme sumarizácie rozdeliť na:

- *Indikatívna* – jej cieľom je pomôcť používateľovi rozhodnúť sa, či si má prečítať celý dokument, t.j. či je preňho relevantný; väčšinou ide o krátke zhrnutie najvýznamnejších tém
- *Informatívna* – na rozdiel od indikatívnej sumarizácie je jej cieľom poskytnúť používateľovi taký súhrn, ktorý ho oboznámi s obsahom dokumentu do tej miery, že ho už nebude musieť pre získanie základnej predstavy čítať celý (je preto spravidla dlhšia ako indikatívna sumarizácia)
- *Hodnotiaca* – typickým príkladom je recenzia, prípadne kritika, čiže okrem sumarizovania obsahu vyjadruje aj názor toho, kto ju vytvára, preto to vo všeobecnosti nie je strojovo riešiteľná úloha (je však možné sumarizovať názor používateľov, napr. na nejaký produkt)

A napokon, sumarizácie môžeme deliť tiež podľa ich zamerania na:

- *Všeobecná (generická)* – je určená pre všetkých používateľov bez ohľadu na ich záujmy, ciele alebo znalosti, preto sa snaží zhrnúť všetky dôležité témy obsiahnuté v dokumente
- *Založená na dopyte* – pri tvorbe sumarizácie sa zohľadňuje dopyt zadaný používateľom, t.j. vyberajú sa časti dokumentu, ktoré súvisia s kľúčovými slovami dopytu
- *Tematicky zameraná* – výsledná sumarizácia obsahuje informácie týkajúce sa danej témy
- *Aktualizačná* – zohľadňuje sa apriórna znalosť používateľa, tzn. predpokladáme, že ak používateľ prečítal danú množinu dokumentov, tak je oboznámený s témami, ktoré sú v nich obsiahnuté, a preto sa do sumarizácie vyberú predovšetkým nové témy
- *Zameraná na používateľa* – zohľadňujú sa záujmy, ciele, prípadne znalosti používateľa

V tejto práci ďalej uvažujeme len extraktívne sumarizácie a pod pojmom sumarizácia vždy myslíme extrakt (ak nie je explicitne uvedené inak), pričom dôraz kladieme predovšetkým na informatívne sumarizácie, ktoré sú zamerané na používateľa.

2.1 Základné sumarizačné metódy založené na obsahu

Od 50. rokov 20. storočia, kedy bola publikovaná prvá metóda automatickej sumarizácie textu, vzniklo množstvo ďalších sumarizačných metód. Ich podrobný prehľad nájdeme napr. v (Ježek & Steinberger, 2008, 2010), (Das & Martins, 2007), (Jones, 2007) alebo v staršej práci (Radev et al., 2002). Sumarizačné metódy môžeme rozdeliť do troch skupín. Medzi klasické metódy zaraďujeme:

- Heuristické metódy
- Štatistické metódy

Do skupiny metód využívajúcich spracovanie prirodzeného jazyka radíme:

- Metódy využívajúce súvislosti v texte
- Metódy modifikujúce pôvodný text

A napokon poslednú skupinu predstavujú metódy využívajúce algoritmy a prístupy z teórie grafov a algebry, t.j. ide o:

- Grafové metódy
- Algebrické metódy

Heuristické metódy predstavujú najstaršie sumarizačné metódy. Do tejto skupiny môžeme zaradiť *metódu frekvencie termov* (Luhn, 1958). Ide vôbec o prvú metódu automatickej sumarizácie textu. Jej hlavnou myšlienkou je vybrať do sumarizácie tie vety, v ktorých sa nachádza čo najviac najčastejšie sa v texte vyskytujúcich slov. Luhn najprv abecedne zoradil všetky slová dokumentu, z tejto množiny odstránil spojky, častice, predložky a i. (tzv. stop slová) a pre každé slovo vypočítal frekvenciu jeho výskytu v dokumente, pričom za rovnaké slová považoval tie, ktoré sa nelíšili o viac ako istý počet znakov (tým ošetril rôzne tvary slova s rovnakým slovným základom, t.j. koreňom). Napokon zoradil slová podľa frekvencie ich výskytu a odstránil všetky, ktorých frekvencia bola menšia ako daná prahová hodnota (a tiež tie, ktorých frekvencia bola vyššia ako prahová hodnota, pretože tieto

slová nemajú potrebnú rozlišovaciu schopnosť, keďže sa v texte vyskytujú príliš často). Zvyšné slová považoval za významné a do sumarizácie vyberal vety, ktoré obsahovali čo najviac týchto významných slov, pričom ale bral do úvahy aj ich vzájomnú vzdialenosť – medzi žiadnymi dvomi významnými slovami nesmelo byť viac ako päť nevýznamných.

Medzi heuristické metódy zaraďujeme aj *metódu pozične významných termov* (Edmundson, 1969). Pri výbere viet do sumarizácie pomocou tejto metódy nezohľadňujeme len frekvenciu slov, ale aj ich pozíciu v texte. Zvýhodňujeme slová v nadpisoch, podnadpisoch, slová v prvých a posledných odsekoch dokumentu a prvých a posledných vetách každého odseku. Okrem toho táto metóda tiež rozlišuje niektoré slová indikujúce dôležitosť vety, ako napr. „významný“, „výsledný“, „dôležitý“ a pod.

Do skupiny štatistických metód patrí metóda *tf-idf (term frequency – inverted document frequency)* a metóda *Bayesovej klasifikácie* (Ježek & Steinberger, 2008, 2010). Prvá predstavuje modifikáciu klasickej metódy frekvencie termov a je založená na predpoklade, že význam slova klesá s počtom jeho výskytov v celom dokumente (lebo najčastejšie sa v dokumente budú vyskytovať bežné slová). Preto do sumarizácie vyberáme vety tak, že pre každú vetu zostrojíme vektor frekvencií termov vo vete, ktorý vynásobíme (skalárne) vektorom inverzných frekvencií termov v celom dokumente; do sumarizácie vyberieme vety s najvyšším takto vypočítaným skóre.

Metóda Bayesovej klasifikácie (Kupiec et al., 1995) je založená na známom Bayesovom klasifikačnom vzorci. Ide o metódu strojového učenia, kedy sa klasifikátor snaží každú vetu dokumentu zaradiť do jednej z dvoch tried – patrí do sumarizácie alebo nepatrí do sumarizácie. Na natrénovanie potrebujeme dostatočne veľkú množinu dvojíc dokument – sumarizácia. Klasifikácia prebieha na základe čŕt jednotlivých viet (napr. dĺžka vety, frekvencia významných slov a pod.).

V zásade je možné na úlohu sumarizácie natrénovať ľubovoľný iný klasifikátor, ako príklad uvádzame dve nasledovné práce. Mani a Bloedern (1998) použili na tvorbu sumarizácií tri rôzne algoritmy strojového učenia – štandardizovanú kánonickú diskriminačnú funkciu (SCDF, ktorá sa snaží lineárne separovať priestor viet zaradených do sumarizácie od tých, čo do sumarizácie zaradené nie sú), metódu C4.5 a AQ15.c. Druhé dve použité metódy trénujú rozhodovacie stromy a ich výstupom sú pre človeka zrozumiteľné pravidlá, kedy nejakú vetu do sumarizácie zaradiť, a kedy nie (na základe extrahovaných čŕt). Na rozlišovanie viet použili množinu jedenástich čŕt – pozičné (poloha vety v rámci odseku, sekcii a dokumente), tematické (tf, tf-idf, termy z nadpisu) a kohézne (počty odkazov na danú vetu).

Yeh et al. (2005) navrhli trénovateľný sumarizátor založený na množine čŕt (poloha, kľúčové slová, centrálnosť, t.j. podobnosť voči ostatným vetám dokumentu, podobnosť s nadpisom) podobne ako Mani a Bloedern. Vetám priraďovali čiastkové skóre za každú črtu a tieto kombinovali do výsledného skóre vety, ktoré sa použilo pri jej výbere do sumarizácie. Na nájdenie optimálnych koeficientov kombinácie čiastkových skóre použili genetický algoritmus.

Ďalšiu skupinu sumarizačných metód predstavujú metódy využívajúce súvislosti v texte (napr. metóda založená na *teórii rétorických štruktúr*, či *metóda lexikálnych reťazcov*) a metódy modifikujúce pôvodný text (*metóda kompresie*, či *metóda „cut and paste“*) (Ježek & Steinberger, 2008, 2010). Tieto metódy vyžadujú hlbšiu znalosť štruktúry a (v prípade metód modifikujúcich pôvodný text) sémantiky textu a viet. Metódy, ktoré sú schop-

né modifikovať pôvodný text, využívajú pokročilejšie metódy spracovania prirodzeného jazyka a sú schopné produkovať abstrakty v pravom zmysle slova, tzn. že nezostávajú pri výbere viet pôvodného textu, ale dokážu meniť poradie viet, spájať ich, rozdeľovať, vynechávať časti viet (napr. vedľajšie vety), parafrázovať, nahrádzať výrazy ich synonymami, špecifickejším alebo naopak všeobecnejším pojmom a pod.

Grafové metódy používané na sumarizáciu sú založené na algoritme *PageRank* a jeho modifikáciách *TextRank* a *LexRank* (Ježek & Steinberger, 2008, 2010). Vrcholy grafu reprezentujú vety dokumentu a hrany reprezentujú väzby medzi vetami. Hrany môžu byť ohodnotené váhami, ktoré predstavujú mieru podobnosti viet (normovaný počet spoločných slov dvoch príslušných viet v prípade algoritmu TextRank a kosínusová podobnosť v prípade algoritmu LexRank). Výhodou grafových algoritmov je predovšetkým ich jazyková nezávislosť a tiež to, že nevyžadujú žiadne hlbšie lingvistické znalosti o texte.

Poslednou skupinou základných sumarizačných metód, ktorú spomenieme, je skupina algebrických metód. Z nich je na sumarizačné úlohy najčastejšie využívaná *metóda latentnej sémantickej analýzy (LSA)*, ktorá je založená na rozklade matice na singulárne hodnoty (Landauer et al., 1998). Prvýkrát bola táto metóda za účelom sumarizácie použitá v (Gong & Liu, 2001). Vstupom metódy je matica m termov a n viet dokumentu $\mathbf{A}: m \times n$, $\mathbf{A} = (\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_n})$, ktorej stĺpcové vektory $\mathbf{A_i}$ predstavujú výskyty (váhy) jednotlivých termov vo vetách dokumentu. Táto matica vstupuje do rozkladu na singulárne hodnoty SVD (angl. *singular value decomposition*):

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathbf{T}} \tag{2.1}$$

Výstupom je matica $\mathbf{U} = (u_{ij})$, ktorej stĺpce predstavujú ľavé singulárne vektory a riadky termy dokumentu, diagonálna matica $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, ktorej prvky na hlavnej diagonále sú nezáporné singulárne hodnoty usporiadané zostupne a matica $\mathbf{V} = (v_{ij})$, ktorej stĺpce predstavujú pravé singulárne vektory (a teda v transponovanej matici $\mathbf{V}^T$ sú pravé singulárne vektory riadkami matice; stĺpce predstavujú vety dokumentu).

Ako uvádzajú Gong a Liu (2001), SVD zo sémantického hľadiska odvodzuje latentnú sémantickú štruktúru dokumentu – táto metóda je schopná nájsť dôležité koncepty, ktoré sú zachytené v dokumente ako spoločný výskyt viacerých termov. Okrem toho odráža dôležitosť týchto konceptov v podobe singulárnych hodnôt, vďaka čomu vieme povedať, ktoré koncepty sú v dokumente najviac zastúpené a tiež, ktoré vety ich najlepšie zachytávajú (a tie následne vybrať do sumarizácie).

Z hľadiska sumarizačnej metódy využívajúcej LSA, ktorú navrhli Gong a Liu, sú najdôležitejšie dva kroky:

1. Konštrukcia matice **A**, resp. spôsob určenia váh jednotlivých termov.
2. Spôsob výberu viet z matice $\mathbf{V}^T$ získanej singulárnym rozkladom matice **A**.

Základom konštrukcie matice **A** termov a viet je extrakcia kľúčových slov z textu a určenie frekvencie ich výskytu pre každú vetu. Tieto hodnoty je ďalej možné váhovať podľa viacerých stratégií (schém), Gong a Liu analyzovali štyri *lokálne váhovacie stratégie* (žiadne lokálne váhovanie, binárne, logaritmické a rozšírené) a dve *globálne* (žiadne globálne váhovanie a inverznú frekvenciu výskytu termov v dokumente) a ich kombinácie (s použitím normalizácie a bez nej). Zistili, že pre generickú sumarizáciu dosahuje najlepšie výsledky binárne lokálne váhovanie v kombinácii so žiadnym globálnym váhovaním a bez použitia

normalizácie, pričom pridaním globálneho váhovania (t.j. tf-idf) sú dosahované výsledky len o niečo horšie. Normalizácia má, naopak, výrazne negatívny vplyv na výsledky sumarizácie.

Vety do sumarizácie sa vyberajú podľa konceptov – postupne prechádzame cez jednotlivé singulárne hodnoty od najvyššej po najmenšiu a pre každý koncept vyberieme vetu, ktorá ho najvýraznejšie zachytáva (t.j. nájdeme najvyššiu hodnotu v *k*-tom riadku matice $\mathbf{V}^T$ a jej stĺpcový index vyjadruje poradové číslo vety, ktorá sa vyberie do sumarizácie). Týmto je zabezpečené, že výsledná sumarizácia bude obsahovať minimum redundancie a pokryje maximum tém obsiahnutých v dokumente.

Steinberger a Ježek (2005) identifikovali dve nevýhody postupu opísaného vyššie:

1. Ak uvažujeme počet dimenzií zhodný s počtom viet vyberaných do sumarizácie, tak so zvyšujúcim počtom sa zahrnú aj čoraz menej významné vety.
2. Vety, ktoré majú vysoké skóre pri viacerých konceptoch, ale ani v jednom prípade nie je maximálne zo všetkých viet, nebudú vybrané do sumarizácie, aj keď intuitívne by do nej mali patriť.

Navrhli preto modifikáciu spôsobu výberu viet pomocou matíc získaných singulárnym rozkladom. Pre každú vetu vypočítali jej dĺžku v modifikovanom latentnom priestore:

$$s_k = \sqrt{\sum_{i=1}^{n} v_{ki}^2 \sigma_i^2} \qquad (2.2)$$

kde v_{ki} sú hodnoty z matice $\mathbf{V}$, ktoré hovoria, ako významne je daný koncept *i* zastúpený vo vete *k* a σ_i je singulárna hodnota, ktorá vyjadruje dôležitosť konceptu *i* v celom dokumente; *n* predstavuje počet dimenzií modifikovaného latentného priestoru, pričom jeho hodnota je nezávislá od počtu viet sumarizácie, ale je to parameter metódy – Ježek a Steinberger volili počet dimenzií, pre ktorý boli singulárne hodnoty väčšie nanajvýš rovné polovici z maximálnej singulárnej hodnoty. Do sumarizácie potom jednoducho vybrali *p* viet s najvyššími hodnotami *s*.

Navrhnutý prístup porovnali s pôvodnou metódou opísanou v (Gong & Liu, 2001), pričom dosiahli mierne zlepšenie okolo 3% pri metrike založenej na kosínusovej podobnosti prvých ľavých singulárnych vektorov (a okolo 6% v porovnaní s klasickou metódou tf-idf).

Latentná sémantická analýza bola úspešne použitá aj v ďalších prácach. Yeh et al. (2005) navrhli rozšírenie LSA o mapu textových vzťahov (v ktorej sú vzťahy tvorené na základe podobnosti viet). Navrhnutú metódu porovnali s tradičnými metódami sumarizácie založenými na extrakcii kľúčových slov (a prekonali ich v priemere o 20%) a tiež s trénovateľným klasifikátorom, s ktorým dosiahli porovnateľné výsledky.

Sun et al. (2005) vo svojej práci (bližšie opisujeme v ďalšej časti) porovnali výsledky dosahované LSA s klasickou metódou frekvencie termov, pričom podobne ako predchádzajúce práce zaznamenali signifikantné zlepšenie (v priemere o 10%).

2.2 Metódy sumarizácie uvažujúce ďalšie informácie

Popri klasických metódach sumarizácie existujú sumarizačné metódy, ktoré pri sumarizácii neuvažujú len samotný obsah sumarizovaného dokumentu, ale aj ďalšie informácie, napr. z logov z prezerania webu používateľmi, tagy asociované s dokumentmi, prípadne anotácie

(poznámky) pridané k dokumentom používateľmi, ale aj znalosť o tom, aké iné dokumenty už používateľ prečítal a pod.

Množstvo informácií, ktoré môžeme ďalej využiť pri sumarizácii, vieme od používateľov získať v podobe *implicitnej spätnej väzby*. Napríklad už to, či používateľ v prostredí webu klikne alebo neklikne na odkaz v dokumente (resp. na webovej stránke) alebo to, koľko času strávi čítaním daného dokumentu, nám poskytuje nepriamu spätnú väzbu o jeho záujmoch (Holub, 2010). Okrem týchto dvoch základných môžeme sledovať rôzne ďalšie implicitné indikátory, ktoré sa v zásade rozdeľujú do dvoch skupín, a to na indikátory súvisiace s dokumentom ako celkom a indikátory súvisiace s fragmentom dokumentu (Labaj, 2011a).

Do prvej skupiny patrí napr. počet stlačených klávesov, pohyb myšou, počet kliknutí, počet posúvaní dokumentu, či celkový čas posúvania dokumentu. Do druhej skupiny indikátorov zaraďujeme napr. označenia textu, presuny kurzora, pohľad, či čas zobrazenia daného fragmentu dokumentu (pri posúvaní v dokumente). Posúvanie v dokumente sa niekedy označuje ako opotrebovanie dokumentu čítaním (angl. *read wear*), editovanie textu ako opotrebovanie upravovaním (angl. *edit wear*).

Labaj (2011a, 2011b) navrhol metódu získavania implicitnej spätnej väzby sledovaním pohľadu používateľa pomocou webovej kamery v kombinácii so sledovaním interakcie používateľa s dokumentom pomocou ovládacích vstupných zariadení. Jeho cieľom bolo identifikovať práve také fragmenty textu, ktoré sú čítaním najviac opotrebované – ako jedno z možných využití uvádza sumarizáciu dokumentu.

Medzi metódy sumarizácie, ktoré zohľadňujú ďalšie informácie, môžeme zaradiť aj sumarizáciu zohľadňujúcu dopyt používateľa vyjadrený postupnosťou kľúčových slov, ktoré sa často využívajú najmä v internetových vyhľadávačoch a IR systémoch (angl. *information retrieval systems*). Výhody tohto typu sumarizácií skúmali napr. Tombros a Sanderson (1998). Porovnávali klasický prístup IR systémov, ktoré na zadaný dopyt vrátia nadpis a prvých pár viet, so sumarizáciami zohľadňujúcimi dopyt od používateľa. Zistili, že používatelia sa pomocou sumarizácií dokázali presnejšie a rýchlejšie rozhodnúť, či je daný dokument pre nich relevantný, vyššia bola tiež miera ich subjektívnej spokojnosti s takýmto prístupom oproti klasickému.

Dopyty zadané používateľmi pri vyhľadávaní zohľadňujú Sun et al. (2005), uvažujú však navyše aj informácie z logov. Používateľ zadá do vyhľadávača dopyt, vyhľadávač mu vráti výsledky. Z nich si používateľ vyberie stránku, ktorá podľa neho najviac zodpovedá tomu, čo hľadal a táto akcia sa zaznamená (nezohľadňuje sa však napríklad čas strávený na stránke, a preto sa do týchto záznamov môže dostať šum v podobe stránok, na ktoré používateľ klikol, ale hneď z nich odišiel, pretože zistil, že nezodpovedá tomu, čo chcel nájsť). Výhodou tohto prístupu je, že systém postupne získava od používateľov informácie, ktoré webové stránky skutočne obsahujú témy opísané dopytom (ak je ich dostatočný počet, môže sa čiastočne odfiltrovať spomínaný šum) – Sun et al. na tejto znalosti postavili metódu sumarizácie webových stránok, ktorá zvýhodňuje kľúčové slová dopytov, ktoré viedli na danú stránku.

Môže sa však stať, že používateľ príde na stránku, pre ktorú zatiaľ nie je v logoch žiadny záznam – pre tento prípad vytvorili tematický lexikón založený na ODP (*Open Directory Project*). ODP je projekt, ktorý obsahuje stránky manuálne zatriedené do hierarchie kategórií. Tematický lexikón obsahuje kľúčové slová dopytov pre všetky zaznamenané vyhľa-

dávania stránok danej kategórie. Ak si teraz používateľ vyžiada sumarizáciu stránky, pre ktorú nie je žiadny záznam, zoberú sa kľúčové slová zaznamenané pre kategóriu, do ktorej stránka patrí – ak nie je žiadny záznam ani pre danú kategóriu, vezme sa rodičovská kategória z hierarchie atď.

Ďalšie metódy sumarizácie (webových dokumentov) berú do úvahy hypertextové odkazy na stránke, resp. obsah dokumentov, ktoré odkazujú na sumarizovaný dokument (Delort et al., 2003). Takýto typ sumarizácie sa označuje ako *kontextová sumarizácia*. Jej výhodou oproti klasickému prístupu sumarizujúcemu priamo obsah daného dokumentu je, že si dokáže poradiť aj s dokumentmi, ktoré obsahujú málo textu, ale sú tvorené rôznym multimediálnym obsahom ako sú videá alebo obrázky.

Pri sumarizácii dokumentov môžeme zohľadniť informácie (obsah aj metadáta), ktorými začali používatelia predovšetkým s príchodom Webu 2.0 obohacovať obsah na internete. Hovoríme najmä o anotáciách (poznámkach) rôznych druhov, ako sú napr. komentáre, alebo tagy (viac o poznámkach pozri v kap. 4.5).

Delort (2006) sa zaoberal identifikovaním najviac komentovaných, resp. najviac diskutovaných častí blogov. Komentáre, ako špeciálny typ poznámky, predstavujú nepriamy odkaz na časť textu, ku ktorému sa vzťahujú. Najviac komentované, resp. diskutované časti potom môžu odrážať záujem čitateľov o dané časti dokumentu, a ak sa využijú pri sumarizácii, môže byť výsledný súhrn viac orientovaný na to, čo čitateľov skutočne zaujíma. Nie všetky komentáre sú však relevantné k textu, niektoré predstavujú názor komentujúceho, niektoré sú mierené na autora, alebo sa môžu týkať úplne inej témy.

Delort preto analyzoval rôzne typy komentárov a navrhol poloautomatickú metódu na identifikáciu relevantných komentárov a nájdenie častí textu, ku ktorým sa vzťahujú. Táto metóda spočívala v automatickej extrakcii vektorov čŕt z komentárov, na základe ktorých sa vytvorili zhluky podobných komentárov. Expert následne manuálne vyhodnotil získané zhluky, prípadne upravil parametre a nakoniec priradil zhlukom stupne relevancie voči textu. Takto získané a ohodnotené zhluky sa využili pri identifikácii najviac komentovaných, resp. diskutovaných častí textu. Porovnanie s generickou sumarizáciou odhalilo, že sa touto metódou podarilo nájsť časti textu, ktoré boli pre používateľov zaujímavé, ale do všeobecného súhrnu sa nedostali (neboli dostatočne zvýraznené autorom).

Ďalší príklad využitia dát (metadát), ktorými používatelia obohacujú dokumenty na webe, nájdeme v (Zhu et al., 2009). V tejto práci sa zamerali na využitie tagov pri tvorbe generickej sumarizácie, ktoré predstavujú „*vysoko zovšeobecnené opisy tém obsiahnutých vo webovom dokumente*“. Každé slovo v dokumente je ohodnotené ako lineárna kombinácia váhy podľa metódy td-idf a váhy tagu (ak je dané slovo zároveň aj tag asociovaný s daným dokumentom); do sumarizácie sa vyberajú vety, ktoré majú najvyššie skóre vypočítané ako súčet ohodnotení slov vety predelený celkovým počtom slov vo vete. Autori navrhli metódu ohodnotenia tagov *EigenTag* vychádzajúc z myšlienky, že dobrý tag je taký, ktorý pridalo veľa používateľov, a dobrí používatelia sú takí, ktorí pridali veľa dobrých tagov. Ide o algoritmus podobný známemu algoritmu *HITS* a jeho úlohou je minimalizovať vplyv šumu v tagoch (napr. hodnotiace tagy, osobné tagy, prípadne úmyselný spam). Taktiež navrhli spôsob rozšírenia množiny tagov o asociované tagy na základe spoločného výskytu tagov.

Využitie tagov pri tvorbe sumarizácií, resp. vo všeobecnosti poznámok, ktoré pridáva množstvo používateľov na webe, sa niekedy v literatúre označuje ako *sociálna sumarizá-*

cia. Príklady jej ďalšieho uplatnenia nájdeme v (Boydell & Smyth, 2007) a (Park et al., 2008a, 2008b).

Práca Boydella a Smytha (2007) pritom vychádza zo sumarizácie zohľadňujúcej dopyt. Je založená na predpoklade, že každú stránku môžeme asociovať s množinou dopytov, ktoré boli použité pri jej vyhľadávaní, a každý dopyt môžeme asociovať s krátkym súhrnom orientovaným na daný dopyt (ktorý vráti vyhľadávač). Tagy, ktoré zadávajú používatelia pri záložkovaní stránky (napr. pomocou sociálnej služby *Delicious*), je podľa autorov možné považovať za takýto dopyt. Výsledná sumarizácia je potom kombináciou všetkých súhrnov asociovaných s dopytmi – tagmi od všetkých používateľov, ktorí si danú stránku „ozáložkovali“, pričom vyššia váha je na fragmente, ktorý sa vyskytuje viackrát. Autori diskutujú možnosť modifikácie navrhnutej metódy s využitím virtuálnych komunít – do úvahy sa neberú tagy všetkých používateľov, ale len od členov komunity.

Iný prístup k sociálnej sumarizácii zvolili Park et al. (2008a, 2008b). Navrhli metódu, ktorá využíva komentáre a tagy vkladané používateľmi pri vytváraní záložky podobne ako Boydell a Smyth (2007) a tiež Delort (2006), ale na rozdiel od nich sumarizujú priamo tieto komentáre. Museli preto navrhnúť spôsob, ako rozpoznať, ktoré komentáre majú charakter súhrnu (a teda sú vhodné pre sumarizáciu), a ktoré nie – využívajú na to kombináciu frekvencie slov, prekryvu s nadpisom a tagmi a tiež charakteristické syntaktické vzory (postupnosti slov). Výhodou nimi navrhnutej metódy je schopnosť sumarizovať aj netextový webový obsah, naopak nevýhodou je, že dokážu poskytnúť relevantnú sumarizáciu len pre stránky s dostatočným počtom záložiek používateľov.

Doteraz spomínané metódy vždy sumarizovali jeden dokument. Ak však namiesto jedného dokumentu sumarizujeme viacero dokumentov, hovoríme o *viacdokumentovej sumarizácii* (Ježek & Steinberger, 2010). Okrem klasickej viacdokumentovej sumarizácie, pri ktorej je cieľom vytvoriť všeobecný súhrn danej množiny dokumentov, poznáme tri špeciálne typy sumarizačných úloh (Witte & Bergler, 2007):

1. *Aktualizačná sumarizácia* – predpokladá apriórnu znalosť používateľa, t.j. spolieha sa na to, že ak používateľ prečítal istú množinu dokumentov, tak pozná koncepty a témy v nich obsiahnuté, a preto má význam pri sumarizácii nového dokumentu uprednostniť tie informácie a témy, ktoré sú pre používateľa nové
2. *Cielená sumarizácia* – zameriava sa na používateľov cieľ, t.j. jeho aktuálny kontext (úlohu, ktorú rieši), ktorý môže byť vyjadrený formou výrokov opisujúcich daný kontext
3. *Kontrastná sumarizácia* – jej cieľom je zvýrazniť, čo majú dokumenty spoločné a naopak, v čom sú rozdielne, čo je dôležité napr. pri štúdiu rôznych zdrojov na jednu danú tému, pri rešerši a pod.

Všetky tri spomínané sumarizačné úlohy využívajú externé informácie buď o ďalších dokumentoch alebo o používateľovom kontexte. Witte a Bergler (2007) navrhli metódu, ktorá umožňuje dynamicky generovať všetky tri typy – základom je predspracovanie dokumentov a vytvorenie tzv. tematických zhlukov, ktoré predstavujú fuzzy množiny všetkých entít v množine dokumentov, ktoré súvisia s témou zhluku. Relatívne jednoduchým preusporiadaním týchto tematických zhlukov je potom možné vygenerovať aktualizačnú, cielenú alebo kontrastnú sumarizáciu.

Aktualizačná sumarizácia súvisí s odhaľovaním (a zohľadnením) novosti informácií (angl. *novelty*) v dokumente. Novosť informácií ako kritérium pri tvorbe sumarizácií využívajú aj

iné prístupy a metódy, opísané napr. v (Carbonell & Goldstein, 1998; Sweeney et al., 2008).

Carbonell a Goldstein (1998) navrhli metódu viacdokumentovej sumarizácie zohľadňujúcu novosť informácií na základe *maximálnej hraničnej relevancie* (angl. *maximal marginal relevance*, *MMR*). Je vhodná najmä v prípade, keď máme množstvo dokumentov s vysoko redundantnými (duplicitnými) informáciami. Snaží sa maximalizovať podobnosť dokumentu voči zadanému dopytu používateľa a zároveň minimalizovať podobnosť ľubovoľnej dvojice dokumentov vybraných do sumarizácie. Namiesto dopytu od používateľa je možné použiť aj profil používateľa (model) v podobe kľúčových slov, potom môžeme túto metódu zaradiť medzi personalizované sumarizačné metódy. Podobne môžeme túto metódu použiť aj pre sumarizáciu jedného dokumentu, kedy sa robí porovnanie na úrovni jednotlivých viet.

Sweeney et al. (2008) navrhli metódu inkrementálnej sumarizácie so zohľadnením novosti, tzn. že na začiatku používateľovi poskytli prvotnú sumarizáciu, ktorá zohľadňovala jeho dopyt a na vyžiadanie ju dopĺňala o vety, ktoré obsahovali čo najviac nových doteraz nevidených slov. Z ich zistení však vyplynulo, že zohľadnenie novosti neprinieslo výrazný rozdiel v čase ani presnosti rozhodnutia používateľa o relevancii jednotlivých dokumentov.

2.3 Metódy personalizovanej sumarizácie

Medzi sumarizačné metódy využívajúce ďalšie informácie patria aj metódy personalizovanej sumarizácie. Vyčleňujeme ich ako samostatnú skupinu, pretože ich cieľom je na rozdiel od ostatných metód využiť tieto dodatočné informácie na prispôsobenie sumarizácie používateľovi (čitateľovi), t.j. vyzdvihnúť tie témy dokumentu, o ktoré by sa mohol zaujímať, alebo ktoré by mohli doplniť či prehĺbiť jeho vedomosti.

Personalizácia (prispôsobovanie sa konkrétnemu používateľovi) predstavuje alternatívu k tradičnému prístupu, ktorý nerobí rozdiely medzi používateľmi, a ktorý nazývame aj *jedna veľkosť pre všetkých* (Brusilovsky, 2001). Prispôsobovať sa používateľovi znamená prispôsobovať sa jeho *charakteristikám* (Brusilovsky 1996, 2001):

- *Znalosti* – predstavujú jednu z najdôležitejších charakteristík predovšetkým v doméne výučbových systémov, ale aj mimo nej; snahou je modelovať znalosti používateľa o danej doméne a ich zmeny v čase. Toto môže byť vhodné pri personalizácii sumarizácie v doméne výučby, či už uvažujeme scenár opakovania, kedy z dokumentov vyberáme do súhrnu to, čo sa používateľ (študent) učil, aby si to mohol zopakovať, alebo alternatívny scenár, keď sa naopak snažíme zdôrazniť to, čo je v dokumente nové a používateľ to ešte nevie, tzn. čo nové sa môže prečítaním daného dokumentu naučiť.
- *Ciele a úlohy* – každý používateľ pri práci sleduje nejaký cieľ, prípadne pracuje na nejakej konkrétnej úlohe, preto má identifikácia tohto cieľu alebo úlohy veľký význam pre rozhodnutie o prispôsobení. Toto by sme mohli využiť aj pri personalizácii sumarizácie – napr. ak vieme, že používateľovým aktuálnym cieľom je vyhľadať konkrétne informácie, tak tieto dáme pri sumarizácii dokumentu do popredia.
- *Preferencie* – aj dvaja používatelia s rovnakými znalosťami, záujmami či cieľmi sa od seba môžu líšiť, a to najmä svojimi preferenciami, preto by adaptívne systémy mali aj tieto zohľadňovať. Pre personalizáciu sumarizácie môžu mať význam napr.

preferencie ohľadom jej dĺžky či umiestnenia (na začiatku alebo na konci dokumentu).

- *Skúsenosti* – táto skupina charakteristík predstavuje skúsenosti používateľa napr. s daným systémom, ale môže ísť aj o skúsenosti mimo systému, čo sa však ťažšie identifikuje a modeluje; nie sú vhodné (použiteľné) pre personalizáciu sumarizácie.
- *Záujmy* – táto charakteristika sa dostáva čoraz viac do popredia, pretože používatelia chodia na web nielen za prácou, ale aj (v niektorých prípadoch hlavne) za zábavou a ich cieľom je často získať nové informácie z oblastí, o ktoré sa zaujímajú; môžeme rozlíšiť dlhodobé a krátkodobé záujmy. Záujmy sú vhodné aj pre personalizáciu sumarizácií; ich zohľadnenie môže pomôcť najmä pri sumarizácii dokumentov so širším záberom tém tak, že zvýrazní tie témy, o ktoré sa používateľ aktuálne (alebo dlhodobo) zaujíma.
- *Individuálne rysy* – ide o skupinu čŕt, ktoré sú charakteristické pre každého jednotlivca a definujú jeho osobnosť, ako napr. či je extrovert alebo introvert, alebo aké štýly učenia preferuje a pod. Sú však len ťažko aplikovateľné pri personalizácii sumarizácie (predovšetkým osobnostné charakteristiky typu introvert/extrovert nemá zmysel uvažovať). Vhodná by ale mohla byť napr. informácia o štýle učenia – niekto sa možno radšej učí z definícií, iný zase môže preferovať vysvetlenie na príklade, čo by sa pri sumarizácii dalo použiť, ak by sme vedeli identifikovať typ a vlastnosti (definícia, príklad, náročnosť) jednotlivých fragmentov textu.

Súhrn charakteristík používateľa udržiavaný v systéme predstavuje *model používateľa*. Tento väčšinou vzniká spracovaním údajov, ktoré používateľ priamo alebo nepriamo poskytol systému (pomocou explicitnej alebo implicitnej spätnej väzby).

Väčšina adaptívnych systémov súčasnosti sa obmedzuje na personalizáciu v rámci jednej známej domény. Takýto prístup má svoje výhody, pretože znalosť domény umožňuje vytvoriť jej model (konceptualizáciu), ktorú následne môžeme využiť aj pri modelovaní používateľa v podobe tzv. *prekryvného modelu* (Brusilovsky, 1996).

V prípade, že doména nie je dopredu známa, musíme si síce poradiť s týmto obmedzením, no výhodou je, že takýto systém je následne nezávislý od domény a môže teda byť naraz použitý vo viacerých doménach, dokonca v prostredí celého webu. Keď sa nemôžeme spoliehať na existenciu konceptualizácie domény, môžeme použiť model používateľa založený na *kľúčových slovách* (Barla & Bieliková, 2010).

Barla a Bieliková (2010) opísali spôsob, ako môžeme v prostredí „divokého webu“ extrahovať metadáta v podobe kľúčových slov zo všetkých stránok navštívených používateľom pomocou adaptívneho proxy servera. Zberom metadát dostávame tzv. základnú (dôkazovú) vrstvu modelu používateľa. Keďže postupom času získavajú kľúčové slová reprezentujúce dlhodobé záujmy používateľa na váhe (zvyšuje sa počet ich výskytov), uvažovaním *k* najčastejšie sa vyskytujúcich kľúčových slov dostávame model používateľa, ktorý môže byť následne použitý ako základ pre personalizáciu. Využíva sa tu teda predpoklad, že navštívenie webovej stránky používateľom môžeme považovať ako nepriame vyjadrenie záujmu o témy, ktoré sú na stránke obsiahnuté, pričom sila tohto predpokladu závisí aj od času, ktorý používateľ na danej stránke strávil a ďalších akcií, ktoré so stránkou vykonal, ako napr. kopírovanie častí textu (Holub, 2010).

Pri personalizácii sumarizácie môžeme v zásade vychádzať z jedného z dvoch základných zdrojov (alebo ich kombinácie). Prvý prístup využíva ako zdroj personalizácie model pou-

žívateľa, v druhom prípade využívame metadáta väčšinou v podobe poznámok (anotácií), ktorými je obohatený sumarizovaný dokument.

Model používateľa reprezentovaný kľúčovými slovami využili Díaz et al. (2005), avšak na rozdiel od Barlu a Bielikovej (2010) nie sú tieto kľúčové slová získavané automaticky na základe aktivity používateľa, ale sú používateľom manuálne zadané. Takto získaný vektor ováhovaných kľúčových slov predstavuje model dlhodobých záujmov používateľa. Tento je rozšírený o model krátkodobých záujmov, ktorý je reprezentovaný opäť vektorom ováhovaných kľúčových slov zadaných používateľom. Pri výbere viet do sumarizácie sú následne zvýhodňované tie, ktoré sú čo najviac podobné s daným modelom. Ako sme už však naznačili, nevýhodou tohto prístupu je potreba explicitnej spätnej väzby od používateľa pri tvorbe samotného modelu.

Iný prístup zvolili Campana a Tombros (2009). Model používateľa sa konštruuje z viet dokumentov, ktoré požívateľ nedávno čítal. Model je tak reprezentovaný úplným neorientovaným grafom, kde vrcholy grafu reprezentujú vety dokumentov a jednotlivé ohodnotené hrany reprezentujú podobnosť dvoch viet (vrcholov). Model sa rozširuje o nové vety po prečítaní každého nového dokumentu, pričom do modelu sa berú najvýznamnejšie vety dokumentu. Ak počet viet prekročí stanovenú dĺžku, vynechajú sa vety z najstarších dokumentov s najnižším skóre – model tak predstavuje aktuálne (krátkodobé) záujmy používateľa. Do sumarizácie sú následne vybrané tie vety zo sumarizovaného dokumentu, ktoré sú čo najviac podobné s najreprezentatívnejšími vetami z modelu (stupeň „reprezentatívnosti“ vety sa počíta ako priemer ohodnotení všetkých hrán incidentných s vrcholom). Okrem toho je možné počiatočnú sumarizáciu rozšíriť so zohľadnením novosti – sumarizácia sa rozšíri o vety, ktoré sú síce čo najviac podobné modelu, ale zároveň čo najrozdielnejšie od viet vybraných v „prvom kole“.

V prípade, že sumarizačná metóda zohľadňuje pri personalizácii sumarizácie metadáta v podobe poznámok (anotácií), máme v zásade opäť dve možnosti. Prvou možnosťou je nechať anotovať dokument autorom už pri jeho tvorbe (Nagao & Hasida, 1998). Cieľom je použiť takú množinu značiek – anotácií – ktorá umožní strojové spracovanie (porozumenie) textu. Podobné prístupy však väčšinou zlyhávajú na nedostatočnom rozšírení kvôli zvýšeným nárokom na tvorbu takýchto metadát. Oveľa častejšie sa preto používajú metadáta, ktorými dokumenty obohacujú sami používatelia. Ide o trend, ktorý sa rozšíril predovšetkým s príchodom Webu 2.0, kedy používatelia prestali byť len pasívni príjemcovia obsahu, ale začali sa aktívne podieľať na jeho tvorbe a obohacovaní. Pre úlohu sumarizácie môžeme zohľadňovať viacero druhov anotácií, ako napr. už spomínané tagy, ale napríklad aj pri tlačených textoch populárne zvýraznenia (pozri kap. 4.5).

Používateľ zvýraznením časti textu vyjadruje svoj záujem, resp. môžeme usúdiť, že zvýraznená časť je pre používateľa dôležitá (Zhang et al., 2003). Zhang et al. vo svojej práci rozšírili klasickú tf-idf metódu o kľúčové slová extrahované zo zvýraznených častí dokumentu. Nimi publikované výsledky naznačujú, že takýto prístup je pre personalizáciu sumarizácie užitočný. Tiež z nich vyplýva, že nie všetky zvýraznené časti prispievajú ku kvalite výslednej sumarizácie, ale je potrebné vybrať vhodnú podmnožinu poznámok. Ako otvorené problémy, resp. možnosti vylepšenia identifikovali zahrnutie kolaboratívneho filtrovania, t.j. uvažovať nielen poznámky od daného používateľa, ale aj od jemu podobných používateľov a tiež možnosť rozšírenia navrhnutej metódy pre sumarizáciu viacerých dokumentov.

2.4 Zhrnutie a diskusia

Automatická sumarizácia textu predstavuje širokú oblasť, ktorá je posledných zhruba 60 rokov predmetom aktívneho výskumu. Za tento čas vzniklo množstvo metód a prístupov, ktoré sa navzájom líšia v spôsobe sumarizácie, type výstupu a ďalších vlastnostiach. Prehľad základných metód a ich vlastností uvádzame v Tab. 2.1, pričom sme sa snažili vybrať aspoň jedného zástupcu zo všetkých hlavných skupín sumarizačných metód.

Jednotlivé metódy sa okrem toho môžu líšiť v účele sumarizácie (indikatívna vs. informatívna sumarizácia), jej zameraní (generická, zohľadňujúca dopyt, tematicky zameraná, personalizovaná) a tiež v tom, či sumarizujú jeden, alebo viacero dokumentov. V prípade viacdokumentovej sumarizácie sa ďalej môžu špecializovať na aktualizačnú, cielenú a kontrastnú sumarizáciu.

Tab. 2.1 Porovnanie základných metód sumarizácie.

	Skupina	**Výstup**	**Reprezentácia textu (viet)**	**Charakteristika**
frekvencia termov	heuristické m.	extrakt	množina slov	− jednoduchosť − termy s frekvenciou ohraničenou zdola aj zhora
pozične významné termy	heuristické m.	extrakt	množina slov	− pozícia termov − rozlišujúce slová
tf-idf	štatistické m.	extrakt	množina slov	− frekvencia termov vo vete − invertovaná frekvencia termov v dokumente
Bayesov klasifikátor	štatistické m. (m. strojového učenia)	extrakt	množina slov	− potreba trénovať − efektívne po natrénovaní
TextRank	grafové m.	extrakt	množina slov	− iteratívne − vychádza z PageRank-u
LSA	algebrické m.	extrakt	množina slov	− zachytáva zložitejšiu sémantiku − väčšia pamäťová aj výpočtová zložitosť
kompresia	m. modifikujúce pôvodný text	abstrakt	zohľadnenie pravidiel syntaxe a vetnej štruktúry	− schopnosť identifikovať menej významné časti vety (súvetia) − pokročilé spracovanie prirodzeného jazyka
„cut and paste“	m. modifikujúce pôvodný text	abstrakt	zohľadnenie pravidiel syntaxe a vetnej štruktúry	− redukcia a kombinácia viet, parafrázovanie atď. − pokročilé spracovanie prirodzeného jazyka

V oblasti sumarizácie zostáva veľa otvorených problémov. Aj napriek niektorým existujúcim metódam, otvoreným problémom stále zostáva tvorba skutočných abstraktov, a nie

extraktov, ktorá zahŕňa zmenu vetnej stavby, parafrázovanie a pod. Ďalšie špecifické problémy prináša viacdokumentová sumarizácia, pri ktorej musí sumarizátor odhaliť redundantné informácie, pričom tieto môžu byť rôzne formulované.

Ak pri sumarizácii zohľadňujeme charakteristiky používateľa, hovoríme o *personalizovanej sumarizácii*. V širšom kontexte môžeme hovoriť o prispôsobovaní sumarizácie, keď výstupom môže byť aj generická sumarizácia, no zohľadňujúca nielen samotný obsah sumarizovaného dokumentu, ale aj nejaký zdroj informácií navyše. V práci sa ďalej zameriavame práve na prispôsobovanie a personalizáciu sumarizácií.

V tejto oblasti sumarizácie otvoreným problémom zostáva, čo všetko môžeme (v závislosti aj od zvolenej domény) uvažovať pri prispôsobovaní a personalizácii a ako identifikované zdroje informácií navzájom kombinovať tak, aby sme dosiahli sumarizácie lepšie prispôsobené pre danú doménu, resp. potreby a charakteristiky konkrétneho používateľa.

3 Existujúce sumarizátory

Analyzovali sme dostupné riešenia – existujúce sumarizátory, pričom sme sa zamerali na tie, ktoré sú voľne dostupné, resp. ktoré boli publikované vo vedeckých prácach a je možné ďalej skúmať ich vlastnosti. Okrem toho existujú aj komerčné riešenia sumarizátorov – z nich najznámejšie je v poslednom čase asi *Summly*[1]. Tieto sme však ďalej neuvažovali, keďže informácie o metódach použitých v týchto riešeniach nie sú dostupné, resp. sú dostupné len vo všeobecnej rovine.

3.1 SWEeT

Sumarizátor SWEeT[2] slúži na sumarizáciu tém na webe (Steinberger et al., 2008). Používateľ zadá tému opísanú množinou kľúčových slov (pozri Obr. 3.1), sumarizátor následne vyhľadá webové stránky na danú tému a zosumarizuje ich. Ide teda o tematicky zameranú viacdokumentovú sumarizáciu (aj keď tematickú nie v pravom zmysle slova, keďže téma sa zohľadňuje len pri vyhľadávaní a pri samotnej sumarizácii už nie).

Obr. 3.1 Rozhranie sumarizátora SWEeT. Používateľ zadáva opis témy pomocou kľúčových slov, môže si tiež zvoliť použitý vyhľadávač a jazyk (češtinu alebo angličtinu).

Z hľadiska implementácie ide o webovú službu postavenú na platforme Java s modulárnou architektúrou. Neimplementuje však vlastný sumarizátor, ale využíva sumarizátor Musutelsa.

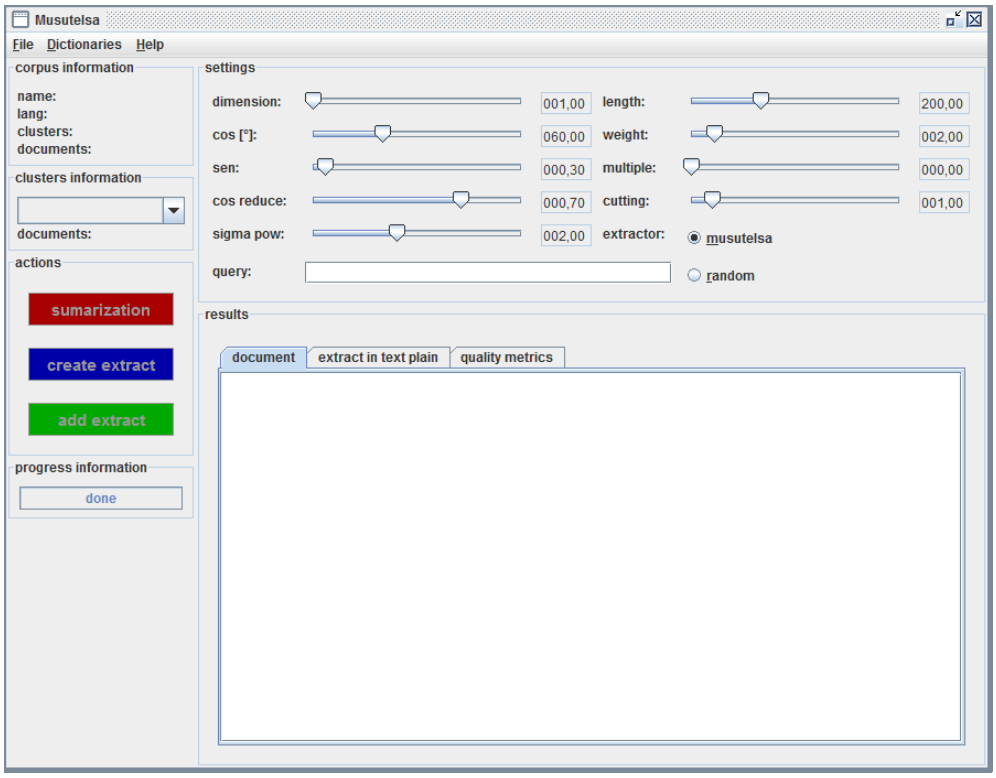
3.2 Musutelsa

Sumarizátor Musutelsa[3] vznikol v rámci diplomovej práce na Západočeskej Univerzite v Plzni. Umožňuje viacdokumentovú sumarizáciu, pričom využíva metódu latentnej sémantickej analýzy LSA. Podporuje sumarizáciu anglických a českých textov, pre podporu

[1] http://www.summly.com/en/introduction.html
[2] http://tmrg.kiv.zcu.cz:8080/sweet/
[3] http://www.musutelsa.jamstudio.eu/

ďalších jazykov je potrebné pridať príslušný slovník stop slov (zoznam bežných často sa vyskytujúcich slov ako spojky, častice a pod.) a lematizátor (slovník na prevod slova na základný slovníkový tvar – lemu).

Je nutné predspracovanie sumarizovaného textu do podporovaného XML formátu, výstup sumarizácie je tiež v XML. Ide v prvom rade o API, ale poskytuje aj konzolové a jednoduché grafické rozhranie (Obr. 3.2). Implementovaný je v jazyku Java.



Obr. 3.2 Grafické rozhranie sumarizátora Musutelsa.

3.3 Almus

Ide opäť o projekt Západočeskej Univerzity v Plzni. Dostupný sumarizátor[4] predstavuje zjednodušenú verziu sumarizátora prezentovaného autormi na konferencii TAC'08. Má podobné vlastnosti ako Musutelsa – poskytuje viacdokumentovú sumarizáciu založenú na LSA, pričom okrem základnej umožňuje aj aktualizačnú sumarizáciu. Pri základnej sumarizácii zohľadňuje aj krátky opis sumarizačnej úlohy, t.j. kľúčovým slovám z opisu je pri sumarizácii priradená vyššia váha.

Dáta sú organizované v zhlukoch, pričom každý zhluk obsahuje množinu starších dokumentov (použijú sa pri tvorbe základnej sumarizácie) a množinu súčasných dokumentov (použijú sa pri tvorbe aktualizačnej sumarizácie). Z uvedeného vyplýva, že sumarizované dokumenty je pred samotnou sumarizáciou opäť potrebné predspracovať do danej XML

[4] http://textmining.zcu.cz/downloads/almus.php

šablóny (podobnej ako v predchádzajúcom prípade). Na rozdiel od sumarizátora Musutelsa je však výstupom obyčajný text (nie v XML štruktúre).

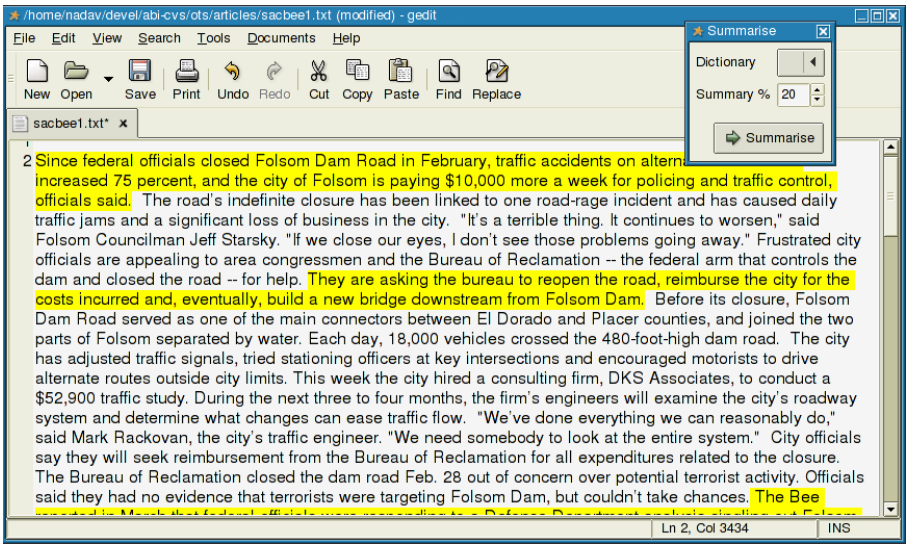
Podporovaný jazyk je len angličtina, no sumarizátor je možné rozšíriť o ďalšie jazyky rovnako ako v predchádzajúcom prípade. Implementovaný je v jazyku Java, na maticové operácie (rozklad na singulárne hodnoty) je použitá voľne dostupná knižnica JAMA[5].

3.4 OTS

Open Text Summarizer[6] (OTS) je voľne dostupný projekt textového sumarizátora. Umožňuje sumarizáciu jedného dokumentu, vstupom je text dokumentu (nie je nutné žiadne predspracovanie) a výstupom je sumarizácia v textovom formáte alebo ako HTML, kedy sú vety vybrané do sumarizácie zvýraznené podfarbením. Na rozdiel od predchádzajúcich sumarizátorov využíva tú najjednoduchšiu metódu sumarizácie, a to metódu frekvencie termov.

Má podporu pre množstvo jazykov, no slovenčina medzi nimi nie je. Nevyužíva lematizáciu, ale *stemming* (úpravu slova na koreň, t.j. odstránenie všetkých predpôn a prípon) založený na množine definovaných pravidiel (pre každý jazyk zvlášť).

Implementovaný je v jazyku C, existuje však aj nadstavba v jazyku Ruby[7]. K dispozícii je ako knižnica, konzolová aplikácia a existuje tiež v podobe rozšírení do niektorých textových procesorov a editorov (napr. Gedit, pozri Obr. 3.3).



Obr. 3.3 Rozšírenie do textového editora Gedit založené na OTS.

[5] http://math.nist.gov/javanumerics/jama/
[6] http://libots.sourceforge.net/
[7] https://github.com/ssoper/summarize

3.5 MEAD

Ide o voľne dostupnú platformu[8] pre viacdokumentovú sumarizáciu, ktorú vyvíja Univerzita v Michigane (Radev et al., 2004). Poskytuje viacero sumarizačných metód a tiež metódy na vyhodnotenie sumarizácií. Pre poslednú spomínanú vlastnosť sa často používa v literatúre spolu s OTS ako referenčný sumarizátor, napr. v (Boydell & Smyth, 2007) alebo v (Park et al., 2008a, 2008b). Jeho hlavnou výhodou je modulárna architektúra, umožňujúca rozšírenie o nové metódy.

Sumarizácia je založená na automatickej extrakcii zvolených čŕt z dokumentu a následnej klasifikácii viet podľa týchto čŕt. Implementuje viacero klasifikátorov, medzi nimi aj rozhodovací strom.

Vstup aj výstup sumarizátora je opäť v XML. Podporuje veľký počet jazykov (ale slovenčinu nie). Implementovaný je v jazyku Perl.

3.6 Diskusia

Existuje pomerne veľa (voľne dostupných) sumarizátorov, ktoré sa navzájom od seba odlišujú použitou metódou sumarizácie, ako aj ďalšími vlastnosťami (predspracovanie textu, formát výstupu, platforma a i.).

Vo väčšine analyzovaných prípadov je potrebné na vstup sumarizátora dodať už do istej miery predspracovaný dokument (t.j. rozdelený na vety) vo formáte XML. Ani jeden analyzovaný sumarizátor neposkytuje priamu podporu pre sumarizáciu slovenských dokumentov, je však možné rozšíriť ich pridaním vlastného zoznamu stop slov a lematizátora, resp. stemmera. Z uvedeného tiež vyplýva, že každý z analyzovaných sumarizátorov do istej miery závisí na jazyku sumarizovaného textu.

Keďže nejde o komerčné riešenia, rozhrania poskytované týmito sumarizátormi sú väčšinou v textovej podobe (konzola), prípadne poskytujú aj jednoduché grafické rozhranie na testovacie účely. Jedine sumarizátor SWEeT poskytuje webové rozhranie a OTS je distribuovaný aj v podobe rozšírení do textových editorov a procesorov.

Čo sa ich rozšíriteľnosti týka, najvhodnejší je MEAD, ktorý predstavuje ucelenú platformu (aplikačný rámec). Naopak OTS je skôr vhodný na použitie ako hotová knižnica.

Pokročilejšiu metódu sumarizácie (*metódu latentnej sémantickej analýzy*) využíva trojica sumarizátorov zo Západočeskej univerzity, ktoré navyše podporujú viacdokumentovú, tematickú a aj aktualizačnú sumarizáciu; a tiež MEAD (*klasifikácia na základe čŕt*). Naopak najjednoduchšiu metódu (*metódu frekvencie termov*) využíva sumarizátor OTS.

V Tab. 3.1 uvádzame porovnanie vlastností jednotlivých analyzovaných sumarizátorov. Porovnanie úspešnosti (presnosti) jednotlivých sumarizátorov nie je jednoduché, keďže v literatúre sa používajú rôzne metódy vyhodnotenia na rôznych dátových súboroch. Napriek tomu sme sa pokúsili v Tab. 3.2 uviesť publikované hodnoty úspešnosti, pričom pri každej uvádzame aj použitú metriku a zdroj, v ktorom bolo dané overenie publikované.

[8] http://www.summarization.com/mead/

Treba však mať na zreteli, že uvedené hodnoty nie sú priamo porovnateľné, ak nie sú z rovnakého zdroja.

Tab. 3.1 Porovnanie vlastností analyzovaných sumarizátorov.

	Metóda	**Vstup**	**Výstup**	**Rozhranie**	**Program. jazyk**
SWEeT	LSA, tematická sumarizácia, viacdokumentová	kľúčové slová	text	webové, konzola	Java
Musutelsa	LSA, viacdokumentová sumarizácia	XML	XML	konzola, grafické	Java
Almus	LSA, viacdokumentová, aktualizačná sumarizácia	XML	text	konzola	Java
OTS	frekvencia termov	text	HTML, text	konzola (knižnica)	C, Ruby
MEAD	klasifikácia na základe čŕt	XML	XML	konzola (platforma)	Perl

Pri všetkých vyhodnoteniach je použitá metrika ROUGE (angl. *Recall-Oriented Understudy for Gisting Evaluation*), ktorá porovnáva abstrakty vytvorené človekom s automatickými sumarizáciami na základe prekryvu n-gramov (líši sa len zvoleným *n*).

Sumarizátor SWEeT, Musutelsa a Almus boli vyhodnocované na štandardnom dátovom súbore z konferencie DUC (*Document Understanding Conference*). SWEeT bol podľa autorov vo vyhodnotení šiesty najlepší zo všetkých 27 porovnávaných sumarizátorov. Pri sumarizátore Almus nie je uvedené presné číslo, len informácia o poradí – umiestnil sa štvrtý z 24 porovnávaných sumarizátorov.

Tab. 3.2 Porovnanie úspešnosti sumarizátorov.

	Metrika	**Úspešnosť**	**Dátový súbor**	**Zdroj**
SWEeT	ROUGE-2	0,0679	DUC'05	Steinberger et al., 2008
Musutelsa[9]	–	–	–	–
Almus	ROUGE-2	x	DUC'07	Steinberger, Ježek, 2009
OTS	ROUGE-1	0.36	Delicious	Boydell & Smyth, 2007
MEAD	ROUGE-1	0.38	Delicious	Boydell & Smyth, 2007

[9] Sumarizátor Musutelsa je využívaný sumarizátorom SWEeT, preto hodnoty uvedené pre SWEeT sú platné aj preň.

Na záver treba poznamenať, že žiadny z analyzovaných voľne dostupných sumarizátorov neposkytuje personalizovanú sumarizáciu. Personalizovaná sumarizácia môže poskytnúť viaceré výhody oproti generickej vďaka prispôsobeniu sa záujmom používateľa či jeho iným charakteristikám.

Toto môže byť užitočné v rôznych scenároch použitia sumarizácie, napr. pri vyhľadávaní, keď poskytnuté sumarizácie budú cielené na to, čo používateľa zaujíma a ten sa tak bude môcť rýchlejšie a lepšie rozhodnúť, či je daný dokument preňho relevantný. Iný scenár môže byť opakovanie vo výučbovom systéme – personalizovaná sumarizácia zhrnie podstatné koncepty vysvetľované v textoch so zreteľom na to, čo sa používateľ naučil.

Domnievame sa preto, že je užitočné ďalej skúmať možnosti rozšírenia existujúcich metód za účelom ich prispôsobenia charakteristikám a potrebám používateľov.

4 Možnosti prispôsobovania sumarizácie a jej využitia

Automatickú sumarizáciu textu môžeme využiť všade tam, kde má používateľ k dispozícii veľké množstvo textových dokumentov a potrebuje sa rozhodnúť, ktoré prečítať, t.j. ktoré dokumenty sú preňho relevantné. V mnohých prípadoch pritom môže byť užitočné prispôsobovať sumarizáciu konkrétnym cieľom, záujmom či znalostiam používateľa, t.j. poskytovať personalizovanú sumarizáciu. To je možné len vtedy, ak máme k dispozícii model používateľa, alebo ďalšie informácie v podobe dát (a metadát), ktorými napr. používatelia obohacujú obsah na webe.

Typicky môžeme ako vhodné domény pre sumarizáciu identifikovať:

- novinové portály,
- systémy digitálnych knižníc,
- výučbové systémy,

ale v zásade v každej doméne, v ktorej je potrebné pracovať s veľkou bázou dokumentov, môžeme aplikovať niektorú (personalizovanú) metódu sumarizácie. Výzvou je využívať sumarizáciu priamo nad dokumentmi „divokého“ webu, kde nemáme možnosť spoliehať sa na model domény a nemôžeme preto priamo prispôsobiť danú metódu špecifikám domény.

Nie každý scenár použitia sumarizácie si vyžaduje jej prispôsobenie, resp. personalizáciu, tzn. že si vystačíme aj s generickým súhrnom daného dokumentu. Na druhej strane, existujú viaceré scenáre, v ktorých má zohľadnenie dodatočných informácií napr. o charakteristikách používateľov potenciál výrazne zlepšiť kvalitu vytváraných súhrnov, či už ide o *sumarizáciu pre opakovanie* (kedy zohľadníme napr. znalosť domény a vedomosti používateľa), alebo *sumarizácia výsledkov vyhľadávania* (kedy zohľadníme aktuálny cieľ alebo dlhodobé záujmy používateľa).

Výsledná sumarizácia závisí nielen od zvolenej dĺžky súhrnu a jeho účelu, ktoré sú väčšinou dané konkrétnym scenárom použitia, ale aj od dĺžky a charakteru sumarizovaného dokumentu. Ideálne je, ak sumarizujeme stredne dlhý voľne štruktúrovaný text (v podobe viet a odsekov).

V prípade krátkeho dokumentu môže byť problém určiť, čo z neho je relevantné – najmä v prípade klasických metód založených na frekvencii termov. Naopak, v prípade dlhého dokumentu je predpoklad, že obsahuje široké spektrum rôznych tém, konceptov a myšlienok. To môže opäť predstavovať problém pre konvenčné generické metódy sumarizácie; ak by totiž chceli obsiahnuť všetky témy, bola by taká sumarizácia príliš dlhá. Na druhej strane určiť, ktoré témy zahrnúť do súhrnu, a ktoré nie, je náročné (resp. častokrát nemožné) bez dodatočnej informácie napr. o záujmoch či cieľoch používateľa.

Podobný problém nastáva, ak máme sumarizovať text v podobe poznámok. Poznámky sú totiž väčšinou hutné a už samé o sebe predstavujú istý typ sumarizácie najdôležitejších myšlienok. Preto je náročné rozhodnúť, čo z nich je v konkrétnej situácii najrelevantnejšie. Vo všetkých týchto prípadoch má zmysel uvažovať prispôsobovanie a personalizáciu sumarizácie, ktorá majú potenciál poradiť si všade tam, kde konvenčné metódy už nestačia zahrnutím len (pre používateľa, prípadne danú doménu) najdôležitejších tém dokumentu.

4.1 Scenáre použitia

Ak uvažujeme možné scenáre použitia, musíme rozlišovať medzi indikatívnou a informatívnou sumarizáciou, ktoré sa líšia vo svojich cieľoch a teda aj požiadavkách na výstupný súhrn. Identifikovali sme štyri základné prípady použitia, ktoré môžeme rozdeliť do spomínaných dvoch skupín, t.j. či ide o indikatívnu alebo informatívnu sumarizáciu.

4.1.1 Indikatívna sumarizácia

Sumarizácia zobrazeného dokumentu: Plní funkciu abstraktu na začiatku dokumentu, typicky je preto zobrazená pod hlavným nadpisom tak, aby sa používateľ mohol veľmi rýchlo zorientovať. Čas potrebný na rozhodnutie používateľa o relevancii daného dokumentu môžeme ešte skrátiť, ak sumarizáciu zobrazíme pred zobrazením samotného dokumentu, napr. v náhľade (typický scenár použitia je, keď používateľ nadíde kurzorom nad hypertextový odkaz na dokument). V tomto scenári nemusíme sumarizáciu nutne personalizovať, ak nám vystačí všeobecný súhrn tém dokumentu. Výhodné sa použitie personalizácie stáva, ak dokument obsahuje širšie spektrum tém – vtedy je vhodné zamerať sa na tie, ktoré používateľa zaujímajú a ostatné v súhrne vynechať alebo potlačiť.

Sumarizácia výsledkov vyhľadávania: Používateľ zadá dopyt do vyhľadávača na webe alebo v informačnom systéme a výsledkom je usporiadaný zoznam odkazov na webové stránky (dokumenty) podľa ich relevancie k zadanému dopytu. Aby však mohol používateľ vyhodnotiť relevanciu vrátených odkazov bez toho, aby na ne musel kliknúť, potrebuje krátku sumarizáciu dokumentu zohľadňujúcu zadaný dopyt; zohľadniť môžeme aj ďalšie informácie, ako napr. (dlhodobé) záujmy používateľa alebo aktuálny cieľ používateľa (tento je však väčšinou už vyjadrený zadaným dopytom).

4.1.2 Informatívna sumarizácia

Opakovanie vo výučbovom systéme: Výučbové texty bývajú väčšinou dlhšie, pretože sa snažia podať vysvetľované koncepty spôsobom pochopiteľným pre svoje cieľové publikum. Študent pri opakovaní však potrebuje hutnejšie zhrnutie, ktoré zvýrazní hlavné koncepty a minimalizuje pre opakovanie zbytočné vysvetľujúce texty. Ako typický prípad použitia preto môžeme študentovi poskytnúť sumarizáciu výučbového textu (prípadne viacerých textov), ktorá mu môže poslúžiť ako hlavné oporné body pri opakovaní učiva. Vhodné pritom môže byť zohľadnenie vedomostí používateľa (t.j. mieru znalosti konceptov opisovaných v dokumente), prípadne poznámok používateľa v podobe zvýraznení, komentárov, tagov a i.

Sumarizácia zobrazeného dokumentu: Cieľom používateľa je prečítať sumarizáciu, ktorá by ho odbremenila od čítania celého textu – ide preto na rozdiel od indikatívnej sumarizácie, ktorú môžeme používateľovi poskytovať automaticky ako istý spôsob navigácie, o sumarizáciu na požiadanie. Od indikatívnej sumarizácie sa líši tiež svojou štruktúrou, do popredia vystupuje potreba čitateľnosti, súdržnosti a ucelenosti textu. Je preto okrem iného potrebné, aby mala úvod, jadro a záver podobne ako pôvodný text, čím sa dosiahne práve dojem ucelenosti. Môže vyžadovať väčšie úsilie pri konečnom spracovaní sumarizácie (nahradenie zámen za podstatné mená, t.j. vyriešenie tzv. anafor a pod.). Výhodnosť personalizácie sumarizácie pri tomto scenári použitia sa prejaví asi najmä pri dlhších, resp. tematicky širšie zameraných dokumentoch.

Mohli by sme samozrejme navrhnúť aj ďalšie prípady použitia, napr. jednoduchá rešerš v digitálnych knižniciach a i. v závislosti od zvolenej domény.

4.2 Doména výučby

Pre overenie nami navrhnutej metódy personalizovanej sumarizácie sme si zvolili doménu výučby a ako konkrétny scenár použitia *sumarizáciu pre opakovanie*, na ktorú sa zamerali pri ďalšom skúmaní možností prispôsobovania sumarizácie.

Navrhnutú metódu realizujeme v prostredí adaptívneho webového výučbového systému *ALEF (Adaptive Learning Framework)*. ALEF predstavuje aplikačný rámec, ktorý spája princípy tradičných vzdelávacích systémov (konceptualizácia domény, adaptácia navigácie a prezentácie obsahu) s princípmi webu 2.0 (obohacovanie a tvorba obsahu používateľmi, spolupráca). Stojí na troch základných princípoch (Šimko et al., 2010):

- *Modelovanie domény* s ohľadom na možnosť automatizácie niektorých krokov tvorby daného modelu a tiež jeho ďalšej modifikácie samotnými študentmi ako výsledok ich spolupráce
- *Adaptácia a personalizácia výučbového kurzu* založená na modeli používateľa
- *Aktívna účasť študentov vo vzdelávacom procese* v podobe ich spolupráce, obohacovaní existujúceho a tvorbe nového obsahu

Samotný obsah systému ALEF tvoria tzv. vzdelávacie objekty (angl. *learning objects, LO*) – pod nimi rozumieme vysvetľujúci text, otázku alebo cvičenie. Nad vrstvou obsahu sa nachádza vrstva metadát, ktorá je tvorená doménovými konceptmi (reprezentovanými v podobe tzv. relevantných doménových pojmov) a poznámkami (anotáciami) tvorenými študentmi. Keďže bol ALEF navrhnutý s ohľadom na jeho rozšíriteľnosť, bola zvolená spoločná reprezentácia vzdelávacieho obsahu ako aj k nemu prislúchajúcich metadát.

V ďalších častiach tejto kapitoly nám ALEF poslúži ako príklad konkrétnej realizácie skúmaných možností prispôsobovania.

4.3 Konceptuálny model domény

Pri automatickej sumarizácii textu je našou snahou rozpoznať najdôležitejšie termy v texte a vybrať tie vety, v ktorých sa nachádzajú, t.j. majú pre používateľa najväčšiu informačnú hodnotu. Inak povedané, princípom metód sumarizácie je rozpoznať najdôležitejšie *koncepty* daného sumarizovaného textu a následne vybrať do súhrnu termy, ktoré ich reprezentujú.

Hoci neexistuje jednotná definícia, čo je to koncept, často sú vnímané ako elementy znalostí domény, ktoré sú reprezentované termami (Šimko, 2012a, 2012b), prípadne skupinou súvisiacich termov (Cimiano, 2006). Termy, ktorými reprezentujeme koncepty, označujeme ako *relevantné doménové pojmy*, Predstavujú tzv. ľahkú sémantiku ako alternatívu voči sémantike „ťažkej“, t.j. ontológiám. Ich výhodou je, že je možné proces získavania takýchto konceptov z textu čiastočne automatizovať, čím sa odbremení doménový expert a zjednoduší sa celý proces tvorby doménového modelu.

Adaptívne vzdelávacie systémy sú postavené na konceptuálnom modeli domény. Doménoví experti na začiatku identifikujú základné koncepty (relevantné doménové pojmy), ktoré opisujú danú doménu a tiež vzťahy medzi nimi. Následne pre každý vzdelávací objekt na-

viažu množinu súvisiacich konceptov, ktoré sú v danom vzdelávacom objekte vysvetlené, resp. ktoré pomáhajú študentom pochopiť tieto koncepty či precvičiť si ich znalosť.

Konceptuálny model domény (v prípade, že ho máme k dispozícii) sa preto javí ako vhodný zdroj informácií pre prispôsobovanie sumarizácie – pri sumarizácii dokumentu zohľadníme k nemu naviazané koncepty, pretože ide o termy, ktoré už niekto (doménový expert) označil za dôležité v kontexte danej domény a zároveň aj daného dokumentu.

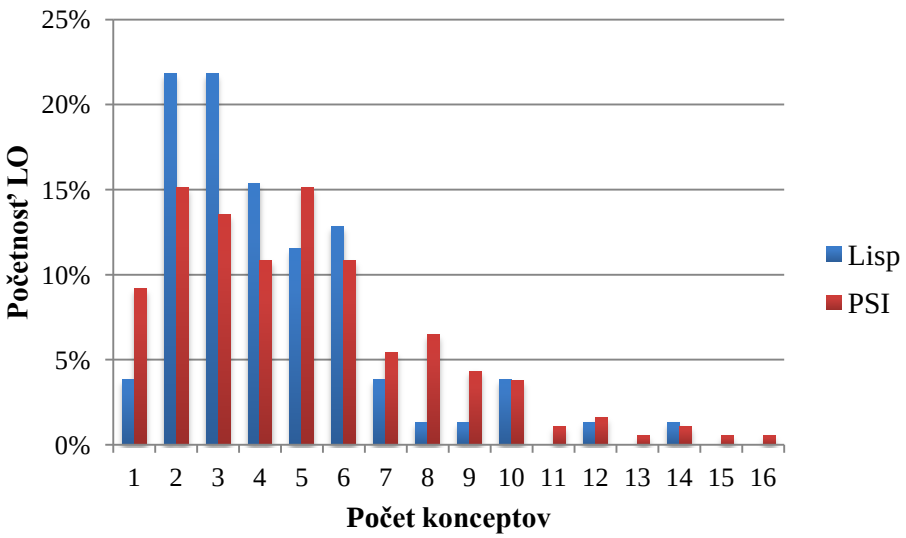
V systéme ALEF predstavujú koncepty dôležité témy obsiahnuté v daných vzdelávacích objektoch (Šimko et al., 2010). Miera súvislosti konceptu a vzdelávacieho objektu je vyjadrená číselnou váhou (Michlík, 2010). Okrem toho sa modelujú aj vzťahy medzi samotnými konceptmi:

- *Generalizácia* – vyjadruje vzťah zovšeobecnenia a špecializácie medzi dvojicou konceptov
- *Súvislosť* – vyjadruje, ako dva koncepty navzájom súvisia (miera súvislosti je daná váhou vzťahu)
- *Prerekvizita* – vyjadruje vzťah, že daný koncept *i*, ktorý je prerekvizitou konceptu *j* je nevyhnutný na pochopenie tohto konceptu; váhou je potom vyjadrená potrebná úroveň zvládnutia konceptu *i*, t.j. prerekvizity

Analyzovali sme koncepty v systéme ALEF z hľadiska možnosti ich využitia pre prispôsobenie sumarizácie. Pre každý vzdelávací kurz je v systéme ALEF doménový model vytvorený doménovým expertom, pričom pre niektoré kurzy je k dispozícii aj doménový model automaticky získaný extrakciou konceptov priamo zo vzdelávacích objektov. V oboch typoch modelov sú identifikované vzťahy súvislosti medzi konceptmi, vzťahy generalizácie a prerekvizity sa však vyskytujú len v manuálne vytvorených modeloch.

Zaujímal nás počet naviazaných konceptov na jednotlivé vzdelávacie objekty. Ako môžeme vidieť na Obr. 4.1, ktorý znázorňuje rozloženie počtu konceptov v manuálne vytvorenom doménovom modeli pre kurzy LISP a PSI, väčšina vzdelávacích objektov má naviazané dva alebo tri koncepty v prípade kurzu LISP a dva alebo päť konceptov v prípade kurzu PSI, pričom priemer aj medián sú (približne) štyri, resp. päť naviazaných konceptov. Je to podľa nás počet, ktorý môže výrazne pomôcť pri prispôsobení sumarizácie, pretože umožní zvýrazniť a sústrediť sa na skutočne dôležité termy dokumentu.

Pri sumarizácii extrahujeme termy dokumenty a prideľujeme im váhy, napr. na základe frekvencie ich výskytu v dokumente. Výhodou zohľadnenia vzťahov súvislosti konceptov (relevantných doménových pojmov) s dokumentom pri jeho sumarizácii je to, že dané váhy (miery súvislosti) určil doménový expert a nie sú nevyhnutne závislé na frekvencii výskytu daného termu v dokumente; taktiež počet naviazaných relevantných pojmov je väčšinou menší ako v prípade automatickej extrakcie termov. Automaticky extrahované termy navyše môžu zahŕňať aj také, ktoré v skutočnosti nie sú pre daný dokument, resp. doménu relevantné.



Obr. 4.1 Rozloženie počtu konceptov naviazaných na vzdelávacie objekty (LO) v doménovom modeli pre kurzy LISP a PSI.

4.4 Model používateľa

Dôležitý zdroj informácií pri prispôsobovaní (personalizácii) sumarizácií je model používateľa, ktorý zaznamenáva (modeluje) charakteristiky používateľov. V doméne výučby nás zaujímajú predovšetkým dve z nich – *znalosti* a *ciele používateľa*.

4.4.1 Znalosti

Častým prístupom pri tvorbe modelu používateľa je tzv. *prekryvný model* – pre každý koncept z doménového modelu si model používateľa pre každého používateľa zaznamenáva hodnotu odhadovanej úrovne znalosti daného konceptu (Brusilovsky & Millán, 2007).

Hoci väčšina výučbových systémov predpokladá, že znalosť (vedomosť) konceptov len rastie, používateľ v skutočnosti zabúda časť získanej vedomosti (Bieliková & Nagy, 2006). Zabúdanie môžeme modelovať pomocou známej Ebbinghausovej krivky zabúdania (1885), ktorá má tvar klesajúcej exponenciály a môžeme ju vyjadriť vzorcom

$$R = e^{-\frac{t}{S}} \tag{4.1}$$

kde R je miera uchovania (zapamätania) informácie, S je relatívna sila ľudskej pamäte a t je čas.

Dôležité z pohľadu sumarizácie a nami zvoleného scenára použitia, t.j. sumarizácie pre opakovanie je, že opakovanie predstavuje spôsob, ako je možné znovu získať zabudnutú znalosť. Opakovať môžeme (Bieliková & Nagy, 2006; Nagy, 2006):

- *Na začiatku sedenia* – tzv. prípravné opakovanie, kedy je cieľom zopakovať si vedomosti naučené v minulom sedení (resp. minulých sedeniach)
- *V priebehu sedenia* – môžeme zopakovať, čo sa študent naučil počas daného úseku sedenia, prípadne opakujeme vtedy, keď zaznamenáme istý pokles danej vedomosti (ktorú už tým pádom považujeme za zabudnutú); vtedy hovoríme o tzv. periodickom opakovaní

- *Na konci sedenia* – tzv. záverečné opakovanie, kedy je cieľom zopakovať vedomosti nadobudnuté v priebehu celého sedenia

Ak teda chceme personalizovať sumarizáciu pre opakovanie, musíme brať do úvahy úroveň znalosti jednotlivých konceptov daného používateľa. Dôležité je tiež zvýrazniť koncepty, ktoré sa používateľ nedávno naučil (či už pri prípravnom alebo záverečnom opakovaní) alebo tie, ktoré zabudol (pri periodickom opakovaní).

4.4.2 Ciele

V modeli používateľa výučbového systému sú často okrem odhadovanej znalosti používateľov zachytené aj ich ciele v rámci systému. *Cieľ* (alebo úloha) predstavuje aktuálny účel (zmysel) práce používateľa v adaptívnom systéme; odpovedá na otázku *„čo chce vlastne používateľ dosiahnuť“* (Brusilovsky & Millán, 2007). Ide o jednu z najviac premenlivých čŕt používateľa. Mení sa totiž nielen medzi jednotlivými sedeniami, ale často môže dôjsť k zmene aj niekoľkokrát počas trvania jedného sedenia.

Existuje viacero prístupov, ako modelovať ciele. Najčastejší prístup je založený na katalógu cieľov, ktorý predstavuje zoznam všetkých možných (preddefinovaných) úloh (cieľov), ktoré je schopný systém pri práci používateľa rozoznať. Možná je aj hierarchia cieľov, napr. ich rozdelenie na dlhodobé a krátkodobé, pričom väčšinou platí, že používateľ má v danom momente na jednej úrovni hierarchie práve jeden cieľ.

Rozpoznanie, aký cieľ používateľ práve sleduje, môže byť len na úrovni manuálneho zvolenia cieľov samotným používateľom (Höök et al., 1996), prípadne pedagógom, ktorý tak napr. určí ciele študentov na daný týždeň podľa aktuálne preberanej látky. Zložitejšie prístupy sú založené na pravdepodobnostnej prekryvnej vrstve nad katalógom cieľov, keď sa pre každý cieľ udržuje hodnota pravdepodobnosti, že daný cieľ je aktuálnym cieľom používateľa (Encarnação, 1997). Najpokročilejší prístup predstavuje rozpoznávanie cieľov podľa sledovania sekvencií činností vykonávaných používateľom v systéme (Hollink et al., 2005; Jin et al., 2005).

Ak ciele reprezentujeme v podobe prekryvnej vrstvy nad doménovými konceptmi, t.j. cieľ predstavuje želanú úroveň znalosti daného konceptu (alebo skupiny konceptov), vieme tieto použiť pri prispôsobení sumarizácie tak, aby odrážala aktuálne ciele používateľa.

4.5 Poznámkovanie dokumentov na webe

Keď čítame text pri učení alebo práci, veľmi často doň dopisujeme rôzne *poznámky* (*anotácie*, čo je však všeobecnejší pojem, ktorý okrem bežne chápaných poznámok môže zahŕňať ľubovoľné metadáta, prípadne značky priradené k dokumentu). Najjednoduchší a možno aj najčastejší spôsob poznámkovania je zvýrazňovanie alebo podčiarkovanie dôležitých častí dokumentu. Niektorí dokonca používajú zvýraznenia viacerých farieb, prípadne rôzne druhy podčiarknutia (rovnou čiarou, vlnovkou a pod.), aby odlíšili rôzne významy pridanej poznámky.

Časté sú aj marginálie, t.j. poznámky na okraji strany, ktoré väčšinou slúžia ako komentáre vzťahujúce sa k textu, pomocou ktorých čitateľ dopĺňa formulácie o zrozumiteľnejšie, odkazuje sa na iné časti textu, prípadne si dáva poznámku, že danej časti teraz nerozumie a musí sa k nej ešte vrátiť.

Je zrejmé, že písanie poznámok môže mať pre nás dva významy – umožňuje nám lepšie si uvedomiť dôležité časti textu, čím nám uľahčuje zapamätanie a pri ďalšom čítaní sa môžeme k našim poznámkam vrátiť, vďaka čomu sa vieme rýchlejšie zorientovať v texte napríklad pri opakovaní (Shipman et al., 2003).

Čoraz viac textu čítame nie v tlačenej, ale v elektronickej forme – na našich osobných počítačoch, mobilných zariadeniach, tabletoch či v čítačkách elektronických kníh. Osobitný dôraz pritom kladieme na webové dokumenty, keďže web sa pre nás stal asi najväčším zdrojom informácií. Je preto užitočné preniesť koncept poznámkovania aj do webových dokumentov a umožniť tak čitateľom zvýrazňovať si dôležité fragmenty textu, pridávať komentáre a pod.

Poznámky tým navyše dostávajú novú dimenziu (Marshall, 1998), pretože ak v tlačenom texte boli poznámky takmer výhradne súkromné (určené len pre ich autora s výnimkou toho, ak sa opoznámkovaný text dostal do rúk niekoho iného, napr. skriptá po staršom spolužiakovi), poznámky vo webových dokumentoch – hypertextoch, môžeme efektívne zdieľať medzi viacerými používateľmi. Hovoríme vtedy o *kolaboratívnom poznámkovaní*.

Proces poznámkovania väčšinou vychádza z cyklu *prečítaj – vyber – vlož,* t.j. pred pridaním poznámky musíme text najprv prečítať (predspracovať), následne vybrať vhodnú poznámku a nakoniec ju vložiť na vhodné miesto (Mihál & Bieliková, 2009). Ak niektoré z týchto krokov, resp. všetky vykonáva stroj, hovoríme o *(polo)automatickom poznámkovaní*. Nás však zaujíma manuálne poznámkovanie, kedy človek – čitateľ vykonáva všetky kroky uvedeného cyklu.

4.5.1 Typy poznámok

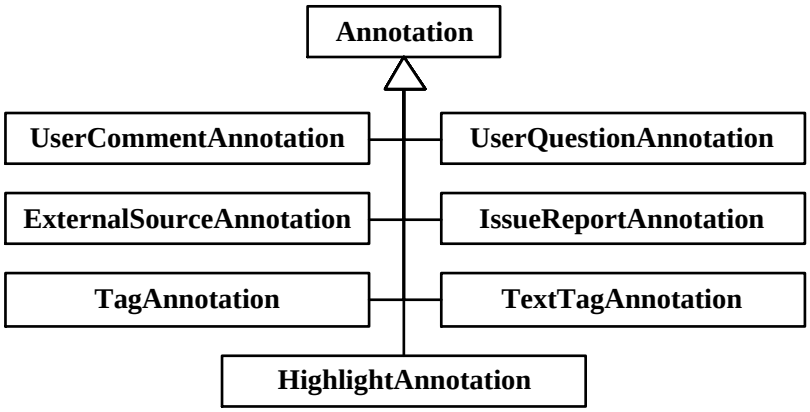
V zásade každú informáciu, ktorá rozširuje (dopĺňa) alebo pozmeňuje pôvodný text môžeme označiť za poznámku. Takáto voľná definícia nám dovoľuje zadefinovať mnoho typov poznámok. Tieto vieme rozdeliť do dvoch základných skupín: *poznámky v texte* a *poznámky k obsahu* (Šimko et al., 2011). Prvá skupina sa viaže na špecifickú časť textu, napr. slovo, frázu či odsek, ktorý bol označený počas tvorby poznámky. V prípade druhej skupiny je naproti tomu poznámka naviazaná k danému obsahu (vzdelávaciemu objektu) ako celku.

Vo výučbovom systéme ALEF existujú tieto typy poznámok:

- *Komentár* – umožňuje používateľovi pridať k označenej časti textu ľubovoľný komentár; reakciou ostatných používateľov na pridaný komentár môže vzniknúť diskusia
- *Hlásenie o chybe* – slúži na nahlasovanie preklepov alebo obsahových chýb výučbových materiálov, ale aj chýb výučbového systému
- *Používateľská otázka* – je to spôsob, ako môžu sami študenti vytvárať otázky, na ktoré následne môžu ostatní študenti odpovedať (Unčík & Bieliková, 2010)
- *Tag* – ide o kľúčové slovo alebo frázu, ktorú pridáva používateľ k výučbovému textu; väčšinou opisuje nejakú tému obsiahnutú vo výučbovom texte, môže však byť aj čisto subjektívneho charakteru (používateľ si tak napr. môže označiť preňho dôležité dokumenty a pod.)
- *Tag z textu* – podobný typ ako predchádzajúci, ale líši sa tým, že sa pridáva označením daného slova v texte a po pridaní toto slovo zostane zvýraznené

- *Externý zdroj* – môže byť pridaný k celému textu alebo jeho zvolenému fragmentu; ide o odkaz na webový dokument mimo výučbového systému, ktorý súvisí s témou dokumentu
- *Zvýraznenie* – umožňuje študentom zvýrazňovať časti textu tak, ako sú zvyknutí z tlačených textov

Na Obr. 4.2 je znázornená časť doménového modelu systému ALEF – môžeme vidieť, že všetky spomínané typy poznámok sú modelované ako špeciálne triedy všeobecnej poznámky.



Obr. 4.2 Časť doménového modelu systému ALEF.

Nás zaujíma, ako by sa dali poznámky využiť pri prispôsobení (personalizácii) sumarizácií. Ak vychádzame z predpokladu, že používateľ pridaním poznámky vyjadruje svoj záujem o danú časť textu (Zhang et al., 2003), malo by ich zohľadnenie viesť k sumarizáciám prispôsobeným používateľovi, ktorý poznámky pridal.

Otázne zostáva, aké typy poznámok sú vhodné na tento účel. Podľa publikovaných štúdií sa ako vhodné javia zvýraznenia (Zhang et al., 2003), prípadne tagy (Li et al., 2008) a komentáre (Delort, 2006). Ich vhodnosť pritom zrejme výrazne závisí aj od zvolenej domény. Rozhodli sme sa analyzovať používanie anotácií vo výučbovom systéme ALEF, aby sme získali predstavu o vhodnosti rôznych typov poznámok na základe reálnych používateľských dát; prípadovú štúdiu používania tagov sme publikovali v (Móro et al., 2011).

4.5.2 Analýza poznámok v systéme ALEF

Analyzovali sme používanie anotácií vo výučbovom systéme ALEF na predmete *Princípy softvérového inžinierstva*. Dátová vzorka obsahuje údaje zo zhruba 8 týždňov používania systému 198 študentmi predmetu (stav systému k 25.4.2011).

Tab. 4.1 zachytáva počet jednotlivých typov poznámok tak, ako ich vložili študenti pri používaní systému. Vidíme, že dominovali zvýraznenia, ktoré tvoria vyše 88% zo všetkých študentmi pridaných anotácií. Toto je spôsobené pravdepodobne tým, že na predmete boli motivovaní, aby pridávali práve tento typ poznámok, ale podľa nášho názoru aj jednoduchosťou ich pridávania. Svoju úlohu pravdepodobne zohral tiež fakt, že sú na zvýrazňovanie dôležitých častí zvyknutí z tlačených textov. Študenti pridali aspoň jednu poznámku (bez ohľadu na jej typ) k 178 vzdelávacím objektom z 317 v danom kurze, čiže sa im podarilo opoznámkovať zhruba 56% z nich (ak berieme do úvahy všetky typy vzdelávacích objektov, t.j. vrátane otázok a cvičení).

Tab. 4.1 Počet jednotlivých typov poznámok.

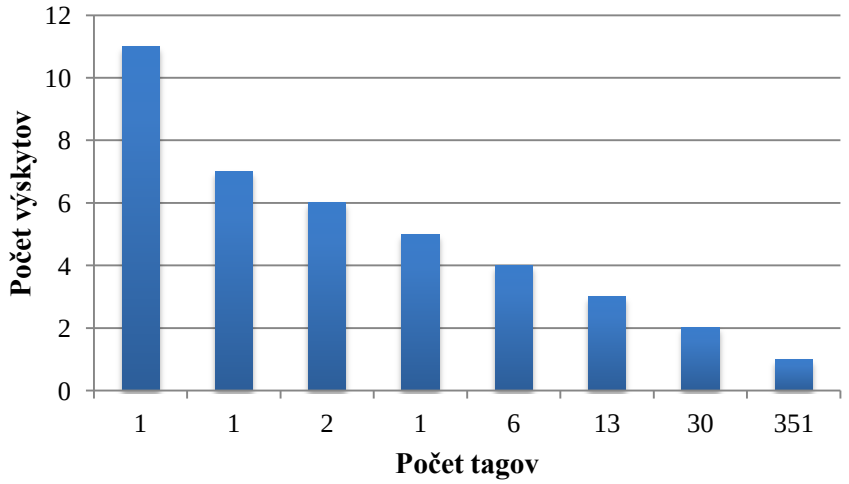
Typ	Tagy	T. z textu	Komentáre	Ext. zdroje	Hlásenia o chybe	Zvýraznenia
Počet	569	492	239	167	94	11692

Pozrime sa bližšie na jednotlivé typy poznámok s výnimkou hlásení o chybách, ktoré sme podrobnejšie neanalyzovali, pretože sú zjavne nevhodné pre účel sumarizácie.

Tagy

Študenti vložili 569 tagov, pričom unikátnych bolo 444. To znamená, že len málo tagov sa vyskytovalo viackrát, ako môžeme podrobnejšie vidieť na Obr. 4.3 znázorňujúcom distribúciu počtu výskytov tagov (išlo len o niečo vyše 13%). Preto v danom stave nemôžeme označiť tagovanie v systéme ALEF za kolaboratívne, keďže v celom kurze sú len tri populárne tagy, t.j. také, ktoré môže vidieť aj niekto iný okrem autora tagu (populárny tag je definovaný ako tag, ktorý k jednému výučbovému objektu anonymne pridali aspoň traja študenti).

Zaujímavé je tiež zistenie, že až 40% tagov bolo pridaných ako súkromné (alternatíva voči tomu sú anonymné tagy, ktoré sú verejné, ale neukáže sa ich autor), čo je relatívne vysoké číslo s ohľadom na fakt, že prednastavená hodnota je anonymný tag. Naznačuje to, že študenti nechcú svoje tagy príliš zdieľať s ostatnými, hoci nie sú súkromnej povahy (čiže nejde o tagy, ktoré by vyjadrovali subjektívny názor, alebo označenie dôležitosti a pod.).



Obr. 4.3 Distribúcia počtu výskytov tagov.

Po manuálnej kontrole tagov môžeme skonštatovať, že majú prekvapivo dobrú kvalitu a darí sa im dobre zachytávať koncepty kurzu. Na druhej strane však asi veľmi neodrážajú záujem používateľa, ktorý ich pridal, skôr jeho chápanie kľúčových slov opisujúcich daný vzdelávací objekt. Ako sme ukázali v (Móro et al., 2011), tagy a z nich tvorené folksonómie je možné a vhodné použiť pri tvorbe alebo úprave modelu domény, čím dokážeme znížiť celkovú náročnosť tohto procesu. Týmto spôsobom vieme tagy využiť aj pri prispôsobovaní sumarizácie, keď namiesto modelu domény, ktorý bol vytvorený doménovými expertmi, použijeme model vytvorený s pomocou tagov od používateľov.

Z pohľadu sumarizácie môžu byť zaujímavé aj *tagy z textu*, ktoré sa na rozdiel od klasických tagov viažu na konkrétne miesto v texte. Používatelia ich pridávajú preto, že si chcú zvýrazniť nejaký podľa nich dôležitý pojem alebo frázu (aj keď významnú úlohu zohráva zrejme aj ich jednoduchšie pridávanie zvýraznením želaného fragmentu, resp. frázy bez potreby ich ručného písania). Ide teda o akýsi kompromis medzi tagom a zvýraznením.

Zvýraznenia

Hlavnou výhodou zvýraznení je, že ich je dostatočne veľa (vyše 11000 nezmazaných zvýraznení) a zároveň pokrývajú najväčšiu časť kurzu (vyše 50% vzdelávacích objektov). Zaujímavé sú aj pre motiváciu, s ktorou ich pridávajú študenti, a to zvýrazniť si pre nich dôležité časti textu. Ide tiež, na rozdiel od tagov, nielen o kľúčové slová, ale o dlhšie úseky textu – avšak len málo zvýraznení z celkového počtu zahŕňa viacero viet, keďže priemerný počet označených slov je okolo 5,5 (medián je ešte nižšie – tri).

Jeden používateľ pridal v priemere k jednému vzdelávaciemu objektu štyri zvýraznenia. Ak by sme chceli uvažovať zvýraznenia aj od ostatných používateľov, museli by sme ich obmedziť na základe nejakého kritéria (napr. podľa ich popularity), pretože jeden vzdelávací objekt obsahuje v priemere spolu od všetkých používateľov až 67 zvýraznení.

Pri takýchto vysokých počtoch už vieme identifikovať najčastejšie zvýrazňované fragmenty dokumentov, ktoré sú zrejme objektívne dôležité a na základe nich prispôsobovať aj generické sumarizácie. Pri tvorbe personalizovaných sumarizácií sa ako vhodné javí použiť kombináciu osobných zvýraznení a najpopulárnejších zvýraznení v danom dokumente, čo je výhodné aj v prípade, že používateľ pre daný dokument nemá žiadne osobné zvýraznenia. Na identifikáciu významných fragmentov dokumentov je pritom okrem zvýraznení možné použiť aj sledovanie pohľadu (Labaj, 2011a, 2011b)

Komentáre a externé zdroje

Pre oba typy poznámok platí, že sa vyskytujú približne len pri 20% výučbových objektov. Externé zdroje je na rozdiel od komentárov možné pridávať priamo k celému výučbovému objektu, nielen k zvolenému fragmentu. Takýchto odkazov sa dokonca nachádza väčšina (vyše 65%). Zostáva preto len relatívne malý počet potenciálne využiteľných poznámok typu externý zdroj. Tieto sa už viažu ku konkrétnemu textu, ide predovšetkým o jedno- alebo dvojslovné frázy, ktoré sú pravdepodobne podrobnejšie vysvetlené v odkazovanom dokumente.

Čo sa komentárov týka, študenti ich používajú väčšinou na dovysvetlenie alebo inú formuláciu fragmentu textu, na ktorý sa komentár viaže. Niektorí študenti pomocou komentáru upozorňujú na chybu v texte. Niekoľkokrát boli komentáre použité aj na označenie dôležitej časti textu, ktorá sa vyskytla vo forme testovej otázky. Môžeme tiež pozorovať spontánny vznik diskusií; hoci nie je možné priamo reagovať na niekoho komentár, študenti jednoducho pridajú ďalší, v ktorom naň zareagujú.

Diskusia

Z analýzy poznámok v systéme ALEF jasne vyplynulo, že najvhodnejšie pre sumarizáciu sú zvýraznenia, vďaka svojmu vysokému počtu, jednoduchosti pridávania a jasnej motivácii študentov. V menšej miere by mohli byť užitočné tagy, resp. tagy z textu vďaka svojej schopnosti nájsť dôležité koncepty opisované v dokumente (a aj ich kontext v prípade tagov z textu). Ako doplnok by sme mohli uvažovať aj komentáre, prípadne externé zdroje,

no pri ich nejasnej (širšej) motivácii (najmä pri komentároch) a celkovo menšom počte je ich prínos k personalizácii sumarizácií otázny.

4.6 Odporúčanie a personalizovaná sumarizácia

Ak uvažujeme *sumarizáciu pre opakovanie* v doméne výučby ako konkrétny scenár použitia, ovplyvňuje tento výber nielen možné zdroje prispôsobovania a personalizácie sumarizácie (konceptuálny model, model používateľa, poznámky používateľov), ktoré sme analyzovali v predchádzajúcich častiach tejto kapitoly. Treba zohľadniť aj ďalšie aspekty ako čas opakovania (Bieliková & Nagy, 2006; Nagy, 2006) a predovšetkým spôsob výberu dokumentov na opakovanie.

To znamená, že potrebujeme spomedzi dokumentov vo výučbovom systéme vybrať, resp. *odporučiť* tie, ktoré sú pre konkrétneho používateľa aktuálne vhodné na opakovanie. Tým presahujeme do problematiky odporúčania, ktorá predstavuje samostatnú výskumnú oblasť.

Existujúce prístupy k odporúčaniu môžeme vo všeobecnosti rozdeliť do troch skupín (Adomavicius & Tuzhilin, 2005):

- *Odporúčanie založené na obsahu* odporúča používateľovi objekty (napr. novinové články, vzdelávacie objekty a i.), ktoré sú podobné tým, ktoré preferoval v minulosti
- *Kolaboratívne odporúčanie* odporúča používateľovi objekty na základe toho, že ich preferovali iní jemu podobní používatelia
- *Hybridné prístupy* pri odporúčaní kombinujú oba uvedené prístupy

Pri ďalšej analýze prístupov k odporúčaniu sme sa vzhľadom na vybraný scenár použitia zamerali na odporúčanie v doméne výučby.

Preferencia používateľov môže byť vyjadrená explicitne (hodnotením objektu), alebo implicitne (Labaj, 2011a, 2011b). Explicitné hodnotenie využili pri odporúčaní vzdelávacích objektov Ghauth a Abdullah (2010), ktorí navyše zohľadňovali, či hodnotenie pochádza od „dobrého“ študenta, alebo nie – priemer hodnotení dobrých študentov považovali za indikátor kvality vzdelávacieho objektu. V prípade, že daný objekt nehodnotil žiadny dobrý študent, odhadli hodnotenie pre odporúčanie na základe podobnosti objektov (t.j. ide o hybridný prístup). Dobrých študentov určovali na základe výsledkov testovania.

Pri odporúčaní preferencia často vyjadruje záujem používateľa. Doména výučby však so sebou prináša rôzne špecifiká a väčšinou v nej neuvažujeme záujem, ale skôr cieľ používateľa, jeho znalosti a pod. Vozár a Bieliková (2008) navrhli metódu adaptívneho výberu testových otázok, ktorá zohľadňuje špecifiká domény výučby kombináciou troch prístupov:

- Výber témy otázky zohľadňujúci štruktúru kurzu (graf prerekvizít)
- Výber otázok najlepšie zodpovedajúcich aktuálnej vedomosti používateľa (pomocou metódy IRT, angl. *Item Response Theory*)
- Výber otázky zohľadňujúci históriu interakcie so systémom (kedy naposledy bola položená otázka a ako na ňu používateľ odpovedal)

Podobný prístup pri odporúčaní zvolil aj Michlík v (Michlík, 2010; Michlík & Bieliková, 2010), ktorý navyše zohľadňoval, že používateľ (študent) má pri učení obmedzený čas.

Vychádzal z hypotézy, že pri obmedzenom čase na učenie je lepšie naučiť sa viac konceptov neúplne, ale na istej úrovni, ako do hĺbky zvládnuť malú podmnožinu konceptov a ostatné nevedieť vôbec.

Opísané prístupy priamo neuvažujú zabúdanie a opakovanie zabudnutých konceptov. Navyše sa zameriavajú na odporúčanie testových otázok a úloh, pričom pre sumarizáciu pre opakovanie má zmysel uvažovať skôr výber vysvetľujúcich textov (dokumentov). Explicitne zabúdanie a opakovanie uvažuje Nagy (2006), ktorý sa zameriava na opakovanie slovnej zásoby pri výučbe cudzieho jazyka – daný koncept (slovo) zaradí na opakovanie vtedy, ak jeho modelovaná vedomosť vplyvom zabúdania klesne pod istú hodnotu (t.j. používateľ ho zabudne).

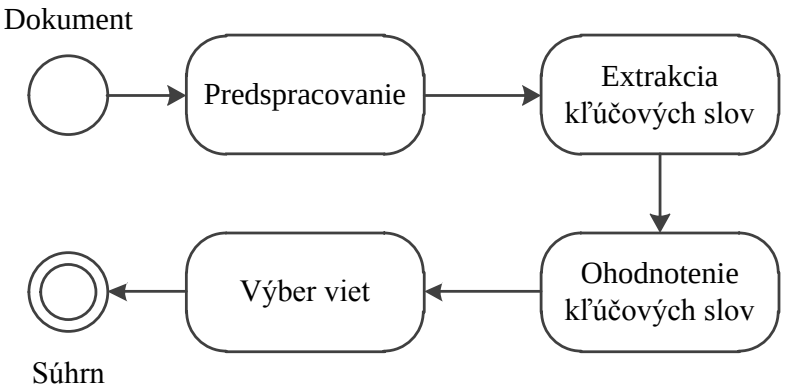
Avšak domnievame sa, že opakovať si nejaký koncept môže byť vhodné nielen pri jeho zabudnutí, ale aj vtedy, ak je pre daného používateľa nový - opakovaním sa môže utvrdiť v novonadobudnutej vedomosti, čo mu môže pomôcť lepšie si daný koncept do budúcnosti zapamätať.

5 Metóda personalizovanej sumarizácie

5.1 Základný koncept

Nenavrhujeme novú metódu či princíp sumarizácie, ale spôsob (metódu) personalizácie niektorej z existujúcich metód, prípadne ich kombinácie, ktorý by sa mohol uplatniť aj nad inou zvolenou metódou sumarizácie. Pri návrhu metódy ako personalizovať sumarizáciu textu sme preto vychádzali zo schémy na Obr. 5.1, ktorá znázorňuje proces sumarizácie spoločný pre väčšinu klasických sumarizačných metód.

Do procesu sumarizácie vstupuje dokument (v našom prípade ide o webový dokument), tento je predspracovaný podľa potreby (napr. prevod z HTML do čistého textu) a následne sa z neho niektorou metódou extrahujú kľúčové slová. Dôležitým krokom je ohodnotenie extrahovaných slov – práve tu je najväčší priestor pre personalizáciu, t.j. treba zvýhodniť tie kľúčové slová, ktoré najlepšie odrážajú záujmy, znalosti alebo ciele používateľa. Posledným krokom procesu sumarizácie je výber (extrakcia) viet z pôvodného textu do výsledného súhrnu, napríklad tak, že sa vypočíta skóre každej vety ako súčet skóre kľúčových slov vo vete a vyberú sa vety s najvyšším ohodnotením (existuje samozrejme veľa iných prístupov).



Obr. 5.1 Schéma procesu sumarizácie.

Navrhujeme metódu personalizovanej sumarizácie textu založenú na *metóde latentnej sémantickej analýzy LSA* (Gong & Liu, 2001; Steinberger & Ježek, 2005), ktorú sme si zvolili vzhľadom na jej schopnosť dosahovať lepšie výsledky sumarizácie v porovnaní s inými metódami. Nami navrhnutá metóda personalizovanej sumarizácie pozostáva z týchto krokov:

1. Predspracovanie dokumentu
2. Konštrukcia personalizovanej matice termov a viet
3. Dekompozícia matice na singulárne hodnoty (angl. *singular value decomposition*, SVD)
4. Výber viet

5.2 Predspracovanie dokumentu

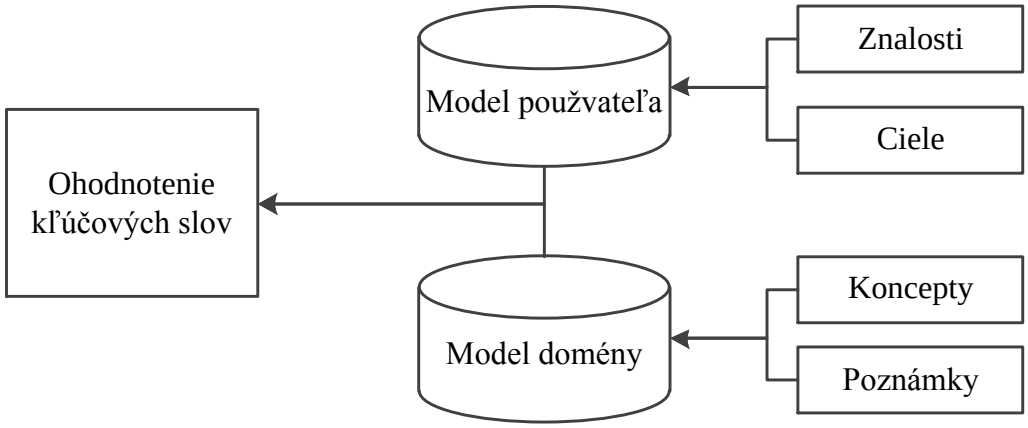
Predspracovanie dokumentu zahŕňa tieto kroky (ako vstup metódy predpokladáme čistý text; v prípade HTML dokumentu by bola navyše potrebná extrakcia textu, t.j. očistenie dokumentu od HTML značiek):

- Strojový preklad do referenčného jazyka
- Tokenizácia a extrakcia termov
- Segmentácia textu na vety

Našou snahou je dosiahnuť čo najvyššiu nezávislosť navrhnutej metódy od jazyka sumarizovaného textu. Na to využívame *strojový preklad* do jedného (referenčného) jazyka, ktorým dokážeme pokryť pomerne veľkú škálu jazykov (vrátane slovenčiny). Ako referenčný jazyk sme zvolili angličtinu. Následné kroky predspracovania, t.j. tokenizácia a extrakcia termov z textu a rozdelenie (segmentácia) textu na vety, sa vykonávajú už nad preloženým (anglickým) textom. Výsledkom sumarizácie je zoznam viet extrahovaných z textu – hoci sa prekladom môže zmeniť poradie slov vo vete, samotné poradie viet sa nezmení. To nám umožňuje nakoniec jednoducho vybrať vety z pôvodného textu, takže konečná sumarizácia je v rovnakom jazyku ako sumarizovaný dokument.

5.3 Spôsob ohodnotenia kľúčových slov

Najdôležitejším krokom metódy LSA je konštrukcia matice termov a viet a s tým súvisiace ohodnotenie kľúčových slov, t.j. priradenie istej váhy každému termu. Práve tento krok sme identifikovali ako vhodný pre personalizáciu sumarizácie. Na Obr. 5.2 je schematicky znázornené, čo všetko môžeme uvažovať pri ohodnotení kľúčových slov.



Obr. 5.2 Vstupy ohodnotenia kľúčových slov.

Klasická váhovacia schéma je založená na metóde frekvencie termov, ktorá je niekedy rozšírená o inverznú frekvenciu výskytov v dokumente, t.j. tf alebo tf-idf:

$$w(t_{ij}) = \frac{|t_{ij}|}{\sum_k |t_{kj}|} \cdot \log\left(\frac{N}{n(t_i)}\right) \qquad (5.1)$$

kde $w(t_{ij})$ je váha termu t_{ij} matice, $|t_{ij}|$ predstavuje počet výskytov termu t_i vo vete j, N je počet všetkých viet dokumentu a $n(t_i)$ je počet viet, ktoré obsahujú term t_i.

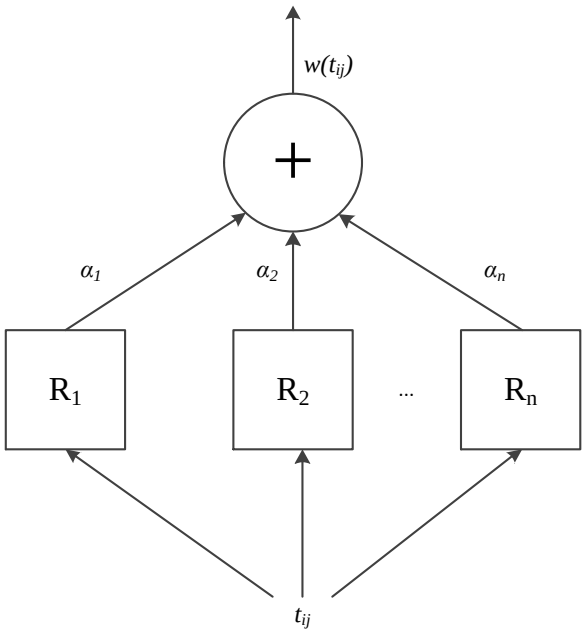
Nami navrhnutá váhovacia schéma spočíva v lineárnej kombinácii viacerých „hodnotičov“, ktorých výstupy kladne (alebo záporne) vplývajú na konečnú váhu slova (Obr. 5.3). Matematicky môžeme túto schému vyjadriť takto:

$$w(t_{ij}) = \sum_k \alpha_k R_k(t_{ij}) \tag{5.2}$$

kde $w(t_{ij})$ je váha termu t_{ij} matice a α_k je lineárny koeficient hodnotiča R_k. Hodnotič R_k je funkcia, ktorá pre zadaný term z množiny kľúčových slov T vráti jeho ohodnotenie (reálne číslo):

$$R_k : T \to \Re \tag{5.3}$$

Takýto návrh je modulárny, pretože umožňuje pridávať a meniť jednotlivé hodnotiče.



Obr. 5.3 Ohodnotenie kľúčových slov kombináciou hodnotičov.

Ide o analógiu s prístupom používaným v multiexpertných systémoch, kde každý expert ohodnocuje (klasifikuje) daný vstup a výsledné ohodnotenie (klasifikácia) je určená kombináciou (hlasovaním) zvolených expertov (klasifikátorov).

5.3.1 Návrh hodnotičov

Navrhli sme sadu hodnotičov, ktoré môžeme rozdeliť do dvoch skupín:

- *Generické hodnotiče* – zohľadňujú obsah dokumentu prípadne ďalšie informácie, ktoré prispôsobujú sumarizáciu bez ohľadu na konkrétneho používateľa
- *Personalizované hodnotiče* – zohľadňujú informácie o konkrétnom používateľovi, ako napr. jeho znalosti, ciele a pod.

Do prvej skupiny môžeme zaradiť *hodnotič frekvencie výskytu termov*, *hodnotič polohy termov v dokumente* a *hodnotič relevantných termov*. Do druhej skupiny môžeme zaradiť *hodnotič vedomostí*, *hodnotič cieľov*, *hodnotič poznámok* (zvýraznení a tagov) a *hodnotič záujmu*.

Rozoberme teraz si jednotlivé navrhnuté hodnotiče podrobnejšie.

Hodnotič frekvencie výskytu termov

Predstavuje klasickú metódu frekvencie výskytu tf, prípadne tf-idf, ktorú sme opísali vyššie pomocou vzťahu (5.1). Pracuje len s obsahom dokumentu, ktorý sumarizujeme a nezohľadňuje žiadne ďalšie informácie. Ide teda o základný prístup, ktorý je dobré kombinovať s ďalšími navrhnutými hodnotičmi.

Hodnotič polohy termov

Pri návrhu tohto hodnotiča sme vychádzali z práce Edmundsona (1969). Hlavnou myšlienkou je zvýrazniť termy, ktoré nám svojou polohou v texte napovedajú, že sú významné. Ide predovšetkým o termy v nadpise, prvej a poslednej vete a prvom a poslednom odseku.

Vychádzame z predpokladu, že nadpis väčšinou opisuje a predznamenáva obsah dokumentu, rovnako tak v mnohých dokumentoch je prvý odsek (veta) dôležitý, pretože uvádza témy ďalej opisované v dokumente. V poslednom odseku (vete) je zase zvykom zhrnúť obsah dokumentu.

Formálne môžeme hodnotič polohy vyjadriť takto:

$$\begin{aligned} w(t_{ij}) &= 1 \quad \text{ak } t_i \in S_j \cap L \\ w(t_{ij}) &= 0 \quad \text{inak} \end{aligned} \tag{5.4}$$

kde S_j je množina termov vety j, L je množina polohou významných termov (termy v nadpise, termy v prvej vete atď.).

Hodnotič relevantných doménových termov

V prípade, že máme k dispozícii doménový model adaptívneho systému, môže nám významne pomôcť pri sumarizácii dokumentu. Doménový model je totiž väčšinou vytváraný doménovými expertmi, ktorí v ňom zachytávajú svoju znalosť domény v podobe dôležitých konceptov (relevantných pojmov) a vzťahov medzi nimi.

Majme množinu relevantných pojmov (konceptov) C_d naviazaných na dokument d, pričom každý koncept v množine je reprezentovaný usporiadanou dvojicou (t_i, w_i), kde term t_i predstavuje daný relevantný pojem (koncept) i a váha w_i vyjadruje súvis dokumentu d s daným konceptom i (tak ako ju určil doménový expert, resp. niektorá z poloautomatických metód tvorby doménového modelu). Formálne potom môžeme hodnotič konceptov vyjadriť analogicky s hodnotičom polohy:

$$\begin{aligned} w(t_{ij}) &= w_i \quad \text{ak } t_i \in S_j \cap C_d \\ w(t_{ij}) &= 0 \quad \text{inak} \end{aligned} \tag{5.5}$$

Hodnotič vedomostí

Ide o hodnotič vhodný predovšetkým v doméne výučby a vzdelávania, kde máme k dispozícii úroveň znalostí (vedomostí) konceptov daného používateľa. Personalizácia je založená na rovnakom princípe ako hodnotič relevantných doménových pojmov s tým rozdielom, že namiesto váhy súvislosti konceptu a dokumentu použijeme úroveň vedomosti konceptu daného používateľa.

Toto platí v prípade, že chceme zvýrazniť koncepty, ktoré už používateľ vie, čo je užitočné napríklad pri opakovaní. Iný scenár však môže uvažovať potlačenie takýchto konceptov a zvýraznenie naopak toho, čo používateľ nevie. V takom prípade by sme ako váhy jednotlivých konceptov zobrali ich doplnok do 1, t.j.

$$k'_{ij} = 1 - k_{ij} \tag{5.6}$$

kde k_{ij} je úroveň znalosti konceptu *i* používateľom *j* (predpokladáme, že k_{ij} je reálne číslo z intervalu $\langle 0,1 \rangle$).

Hodnotič cieľov

Ak je cieľom používateľa zopakovať si daný koncept, potom by aj sumarizácia dokumentu, ktorý s týmto konceptom súvisí, mala tento cieľ podporiť. Použitie hodnotiča cieľov však nie je obmedzené len na doménu výučby, ale je ho možné uplatniť v každej doméne, kde je definovaný nejaký cieľ používateľa pomocou relevantných doménových pojmov.

Uvažujme množinu cieľov G_i daného používateľa *i*, pričom každý cieľ je reprezentovaný množinou usporiadaných dvojíc (t_j, g_{ij}), kde t_j predstavuje daný relevantný pojem (koncept) *j* a váha g_{ij} predstavuje požadovanú (cieľovú) znalosť daného konceptu *j* používateľom *i*. Potom môžeme opäť použiť prístup opísaný vzťahom (5.5) na zohľadnenie týchto cieľov pri sumarizácii, pričom váhu naviazaného konceptu w_j nahradíme rozdielom požadovanej (cieľovej) znalosti daného konceptu a aktuálnej znalosti používateľa k_{ij}:

$$w'_j = g_{ij} - k_{ij} \tag{5.7}$$

Hodnotič poznámok

Na základe analýzy poznámok v kap. 4.5.2 sme dospeli k záveru, že pri prispôsobovaní sumarizácie má zmysel uvažovať dva typy poznámok, a to *zvýraznenia* a *tagy*.

Zvýraznenia nám indikujú, že takto vyznačené časti dokumentu sú pre daného používateľa významné. Okrem zvýraznení daného používateľa však môžeme do úvahy brať aj populárne zvýraznenia, t.j. všetkými používateľmi najčastejšie zvýrazňované fragmenty dokumentu. To má význam z dvoch dôvodov: jednak využívame „múdrosť davu“, keďže skutočnosť, že nejaký fragment textu zvýraznilo veľa používateľov, indikuje, že daný fragment je asi naozaj objektívne dôležitý. Po druhé, zohľadnenie populárnych zvýraznení pomáha používateľom, ktorí nemajú žiadne vlastné zvýraznenia alebo ich majú len veľmi málo.

Zohľadnenie zvýraznení sme navrhli takto: Skonštruujeme množinu všetkých viet takých, ktorých časti boli zvýraznené používateľom *u*, formálne:

$$S_H^u = \{S_j \mid S_j \cap H_u \neq \emptyset\} \tag{5.8}$$

kde S_H^u je množina používateľom zvýraznených viet, S_j je *j*-ta veta dokumentu a H_u je množina všetkých zvýraznení v dokumente od daného používateľa *u*. Analogicky skonštruujeme množinu S_H^P všetkých viet takých, ktorých časti boli zvýraznené ľubovoľným používateľom, pričom však do tejto množiny zaradíme len tie vety, pre ktoré platí podmienka:

$$|h_j| \geq \frac{\max_l |h_l|}{2}, \tag{5.9}$$

kde h_j je počet zvýraznení vety S_j. Nakoniec vytvoríme zjednotenie týchto dvoch množín:

$$S_H = S_H^u \cap S_H^P \tag{5.10}$$

a ohodnotíme termy dokumentu:

$$\begin{aligned} w(t_{ij}) &= 1 \quad \text{ak } S_j \in S_H \\ w(t_{ij}) &= 0 \quad \text{inak} \end{aligned} \tag{5.11}$$

Tagy sú kľúčové slová či frázy, ktoré predstavujú používateľove chápanie dokumentu. Pri ich zohľadnení pri sumarizácii môžeme postupovať rovnako ako v prípade relevantných doménových pojmov (pretože v zásade ide o to isté, len relevantné doménové pojmy identifikoval doménový expert a tagy sám používateľ).

Rozdiel je len v tom, že v prípade tagov nemáme ich váhu, t.j. mieru ich súvislosti s dokumentom. To môžeme vyriešiť tak, že všetkým tagom dáme rovnakú váhu a v takom prípade (ak tá váha bude 1), pôjde o binárne váhovanie (nachádza sa v množine tagov, nenachádza sa v množine tagov). Iný možný prístup je určiť váhy na základe popularity tagov:

$$w_d(tag_i) = \frac{|tag_i|}{\sum_j |tag_j|} \tag{5.12}$$

kde $w_d(tag_i)$ je váha (súvislosť) tagu i v dokumente d a $|tag_i|$ je počet výskytov tagu i pri danom dokumente d (t.j. koľkokrát bol daný tag pridaný používateľmi k danému dokumentu).

Takýmto spôsobom vieme pomocou tagov do istej nahradiť konceptuálny model v prípade jeho absencie, resp. využiť tagy (folksonómiu) pri konštrukcii doménového modelu (Móro et al., 2011).

Hodnotič záujmu

Zohľadňuje záujmy používateľa; tieto môžu byť reprezentované v podobe kľúčových slov a k nim prislúchajúcich váh, potom je použitie tohto hodnotiča analogické so zohľadnením konceptov (relevantných doménových pojmov).

Záujem používateľa však môže byť vyjadrený aj implicitne, napr. časom zobrazenia daného fragmentu dokumentu, počtom akcií nad dokumentom (kopírovanie, označovanie textu a pod.). Identifikovať významné časti dokumentu vieme napr. aj sledovaním kurzora myši alebo pohľadu používateľa ako bolo ukázané v (Labaj, 2011). Takto získané významné časti dokumentu potom vieme pri sumarizácii pomocou nami navrhnutej metódy zohľadniť rovnako, ako zvýraznenia v dokumente.

5.3.2 Kombinovanie hodnotičov

Použitie jednotlivých hodnotičov závisí od zvolenej domény. Ich výber a kombinácia, ako aj nastavenie príslušných váh je podmienené konkrétnym scenárom použitia sumarizácie.

Väčšinou ako základ použijeme hodnotič frekvencie termov v kombinácii s hodnotičom polohy termov, pretože tieto vychádzajú len z obsahu dokumentu a vieme ich teda použiť za každých okolností. Výber ďalších hodnotičov je podmienený tým, aké zdroje informácií máme k dispozícii.

Predpokladáme teda, že váhy nastaví doménový expert, čo však nemusí byť vždy triviálna úloha. Je preto vhodné uvažovať rozšírenie navrhnutej metódy tak, aby boli váhy nastavované (polo)automaticky a mohli sa dynamicky meniť v závislosti od konkrétnej situácie. Ak kombináciu hodnotičov interpretujeme ako hlasovanie expertov, tak váhu priradenú každému hodnotiču môžeme interpretovať ako dôveru v názor daného experta, resp. istotu, s ktorou expert (hodnotič) predkladá svoj názor do hlasovania (kombinácie).

Dôveru (istotu) získame vyhodnotením vlastností daných hodnotičov, resp. vlastností sumarizovaného dokumentu, ale aj druhu, množstva a kvality informácií, ktoré majú hodnotiče k dispozícii. Napr. budeme mať inú dôveru v hodnotič poznámok, ak bude v systéme veľa poznámok od rôznych používateľov, ako keď ich je málo – populárne hodnotenia, na ktorých sa zhodol väčší počet používateľov, majú zrejme vyššiu výpovednú hodnotu. V prípade hodnotiča vedomostí môžeme mať v modeli používateľa k dispozícii okrem hodnoty vedomosti aj pravdepodobnosť, že vedomosť je skutočne taká, ako je modelovaná; od tejto pravdepodobnosti sa potom bude odvíjať dôvera v daný hodnotič. Pri hodnotiči relevantných doménových pojmov zase môžeme zohľadniť dôveru v zdroj týchto pojmov (expert, tagy používateľov a i.).

5.4 Výber viet do sumarizácie

Po ohodnotení kľúčových slov a skonštruovaní personalizovanej matice termov a viet, je táto dekomponovaná pomocou rozkladu na singulárne hodnoty (pre bližší opis pozri kap. 2.1, str. 6). Posledným krokom je výber viet do sumarizácie; vyberáme ich na základe skóre vypočítaného podľa vzťahu (2.2), ktorý navrhli Steinberger a Ježek (2005).

Hoci sa vyberú vety s najvyšším vypočítaným skóre, vo výslednej sumarizácii ich zobrazujeme v poradí, v akom sa vyskytujú v sumarizovanom dokumente. Dĺžka sumarizácie je parametrom metódy.

5.5 Diskusia

Nami navrhnutá metóda personalizovanej sumarizácie je nezávislá od zvolenej domény aj jazyka sumarizovaného dokumentu. Jazykovú nezávislosť sa snažíme maximalizovať použitím strojového prekladu dokumentu do referenčného jazyka. Nemusíme tak poskytovať rôzne algoritmy predspracovania textu (tokenizácia, extrakcia termov, segmentácia textu na vety) pre rôzne jazyky a aj napriek tomu sme schopní zosumarizovať texty zo širokej škály jazykov. Na druhej strane sme závislí od kvality poskytovaného strojového prekladu a nutne strácame istú časť informácie v porovnaní s použitím algoritmov špecializovaných pre daný jazyk.

Navrhnutý spôsob ohodnotenia kľúčových slov pri konštrukcii personalizovanej matice termov a viet nám umožňuje uvažovať rôzne parametre alebo kontext sumarizácie. Výhodou je flexibilnosť, s ktorou je možné prispôsobiť sa konkrétnemu scenáru použitia – napr. ak uvažujeme scenár opakovania vedomostí vo výučbovom systéme, môžeme s výhodou použiť hodnotič relevantných doménových pojmov, ktorý zohľadňuje znalosť experta danej domény. Môžeme ho ďalej kombinovať s hodnotičom vedomostí tak, aby boli zvýhod-

ňované tie koncepty, ktoré sa používateľ už naučil a tiež s hodnotičom poznámok, ktorý zohľadní používateľove zvýraznenia.

V inom scenári, naopak, môžeme zvýrazniť tie koncepty, ktorým používateľ ešte úplne nerozumie, a tak mu pomôcť rýchlo sa rozhodnúť, či daný dokument (ktorý predtým ešte nemusel vidieť) je preňho v danej chvíli relevantný (t.j. naučí sa niečo nové).

Za istú nevýhodu navrhnutého prístupu môžeme označiť nutnosť experimentálneho nastavenia vhodných koeficientov kombinácie tak, aby metóda dosahovala požadované výsledky. Ako vhodné sa preto javí v budúcnosti rozšíriť navrhnutú metódu tak, aby určovanie koeficientov bolo do istej miery automatizované a dynamické v závislosti od konkrétnej situácie spôsobom, ktorý sme diskutovali v kap. 5.3.2.

Ohraničenia metódy spočívajú v štruktúrovanosti a dĺžke textu sumarizovaného dokumentu. Metóda je vhodná predovšetkým pre jednoduchšie štruktúrované dokumenty (do viet a odsekov), v prípade použitia veľkého počtu odrážok, tabuliek, vzorcov či ukážok kódu môže byť problém s čitateľnosťou a zrozumiteľnosťou výsledného súhrnu.

Keďže navrhnutá metóda je založená na metóde LSA, konštruuje sa matica termov a viet, ktorá reprezentuje sumarizovaný dokument. V tom spočíva aj ohraničenie metódy, keď v prípade príliš dlhého dokumentu (veľkého počtu viet), je potrebné skonštruovať veľkú vstupnú maticu (keďže je však riedka, je možné použiť špecializované algoritmy, ktoré znížia celkovú pamäťovú náročnosť).

6 Personalizovaná sumarizácia pre opakovanie vo výučbovom systéme

Doteraz opísaný návrh personalizácie, či vo všeobecnosti prispôsobovania metódy sumarizácie je všeobecný a použiteľný v akejkoľvek doméne pri zvolení vhodných (v danej doméne použiteľných) hodnotičov a ich kombinácie.

Pre overenie nami navrhnutej metódy sumarizácie sme si zvolili konkrétny scenár použitia v doméne výučby, a to *sumarizáciu pre opakovanie*. To ovplyvňuje nielen výber navrhnutých hodnotičov, ale musíme zvážiť aj ďalšie aspekty opakovania, a to kedy opakovať a aké dokumenty vybrať na opakovanie.

6.1 Výber hodnotičov a dĺžka sumarizácie

V doméne výučby sme identifikovali tri hlavné zdroje prispôsobovania a personalizácie sumarizácie:

- *Model domény*, v ktorom sú zachytené dôležité koncepty v podobe relevantných doménových pojmov a vzťahy medzi nimi
- *Znalosť študentov* konceptov danej domény modelovanú pomocou prekryvného modelu používateľa
- *Poznámky študentov*, predovšetkým zvýraznenia dôležitých fragmentov výučbových textov, prípadne tagy, ktorými študenti značkujú tieto texty

Navrhli sme príslušné hodnotiče (bližšie opísané v kap. 5.3.1), ktoré možnosťami svojho použitia presahujú doménu výučby, a teda sa mohli uplatniť aj v iných doménach.

Pri konkrétnom scenári opakovania vo výučbovom systéme potrebujeme v sumarizácii dokumentu zachytiť vysvetľované koncepty a sústrediť sa na to, čo už študent preberal (a teda by to mal vedieť). Môže sa stať, že dokument opisuje viacero konceptov, z ktorých niektoré študent ešte nevie, pretože sú napr. náplňou výučby až neskôr (v neskorších týždňoch semestra). V takom prípade je vhodné do sumarizácie zahrnúť len tie koncepty, ktoré študent aktuálne vie (má vedieť) a ostatné potlačiť, pretože by študenta len miatli.

Pri sumarizácii pre opakovanie využívame kombináciu:

- *hodnotiča relevantných doménových pojmov*, ktorý pomáha sumarizovať koncepty vysvetľované v dokumente tak, ako ich zachytil doménový expert
- *hodnotiča vedomostí*, ktorý pomáha sústrediť sa pri sumarizácii na už naučené (osvojené a zvládnuté) koncepty a potláča tie, ktorým študent ešte dostatočne nerozumie
- *hodnotiča poznámok*, ktorý personalizuje sumarizáciu zohľadnením toho, čo je v dokumente dôležité z pohľadu konkrétneho študenta (resp. z pohľadu väčšiny, ak zahrnieme aj populárne zvýraznenia)

V prípade, ak by sme mali k dispozícii model cieľov používateľa, mohli by sme do kombinácie zaradiť aj *hodnotič cieľov*, ktorý by zohľadnil, čo je aktuálnym cieľom používateľa (napr. čo konkrétne si potrebuje zopakovať pred blížiacou sa písomkou či skúškou).

Čo sa dĺžky sumarizácie týka, pre opakovanie poskytujeme dlhšie sumarizácie (približne tretina dĺžky sumarizovaného textu). Keďže však príliš dlhá sumarizácia môže študenta

skôr odradiť a od istej dĺžky tak prestáva plniť požadovanú úlohu, obmedzujeme zhora dĺžku sumarizácie pre rozsiahlejšie dokumenty.

6.2 Čas opakovania

V kap. 4.4.1 sme analyzovali rôzne časy opakovania, ktoré pripadajú do úvahy. Z nich (opakovanie na začiatku, v priebehu a na konci sedenia) sa zameriavame na opakovanie na začiatku sedenia. Používateľ (študent) sa prihlási do systému s úmyslom učiť sa, a preto môže byť preňho prospešné, aby si na úvod zopakoval, čo sa naučil počas predchádzajúceho sedenia predtým, než sa začne učiť nové učivo. Navyše nám to dáva možnosť predspracovať odporúčania dokumentov na opakovanie ako aj ich personalizované sumarizácie, takže používateľ na ne nemusí čakať.

6.3 Výber dokumentov na opakovanie

Ďalším významným aspektom opakovania je spôsob výberu dokumentov na opakovanie. Vychádzame pritom z metód odporúčania (pozri kap. 4.6), keďže výber dokumentu v zásade predstavuje odporúčanie študentovi, aby si zopakoval koncepty vysvetlené v danom vybranom (odporúčanom) dokumente. Podobne ako pri odporúčaní môžeme brať do úvahy rôzne poznatky a informácie zachytené v modeli používateľa a domény:

- Prečítané dokumenty (spolu s časom, kedy ich používateľ čítal)
- Vzťahy medzi konceptmi (súvislosť, špecializácia/generalizácia, prerekvizita)
- Ciele používateľa (reprezentované vybranými konceptmi z domény)
- Zmena znalostí konceptov

Na ich základe sme navrhli *metódu personalizovaného výberu dokumentov na opakovanie*:

1. Z množiny všetkých vzdelávacích objektov (vysvetľujúcich textov) v doméne vyberieme podmnožinu dokumentov, ktoré používateľ čítal (môžeme obmedziť zhora, napr. najviac päť sedení dozadu).
2. Následne pre každý dokument z podmnožiny skontrolujeme, či sú splnené všetky prerekvizity. Inými slovami, prechádzame cez všetky naviazané koncepty daných dokumentov a vylúčime ten dokument, pre ktorý existuje naviazaný koncept c_i taký, že jeho prerekvizitou je koncept c_j, ktorého znalosť je pod stanovenou hodnotou (daná váhou hrany spájajúcej c_i s c_j), t.j. $k(c_j) < w(c_i,c_j)$.
3. Pre každý zostávajúci dokument d potom vypočítame skóre $S(d)$ a na opakovanie vyberieme dokumenty s najvyšším vypočítaným ohodnotením (skóre).

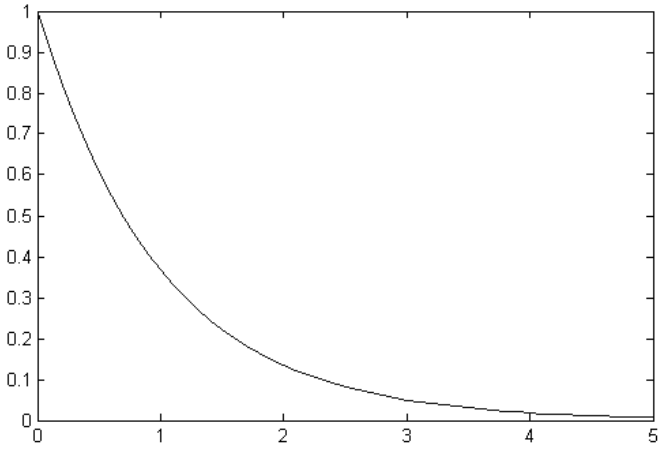
Skóre $S(d)$ vypočítame lineárnou kombináciou piatich zložiek (t.j. využívame podobný prístup ako pri konštrukcii matice termov a viet, pri ktorej sme uplatnili kombináciu hodnotičov):

$$S(d) = \alpha S_T(d) + \beta S_P(d) + \gamma S_G(d) + \delta S_K(d) + \varepsilon \qquad (6.1)$$

Parameter ε predstavuje náhodné malé číslo z normálneho rozdelenia, ktorým sa snažíme modelovať istú neurčitosť pri výbere dokumentov. Výber tým pádom nie je deterministický, ale existuje možnosť, že sa na opakovanie vyberie aj dokument s celkovo nižším ohodnotením – ovplyvniť môže najmä situácie, ak je skóre dokumentov relatívne vyrovnané a len sa mierne líši v niektorej zložke, napr. v čase prečítania dokumentu.

6.3.1 Čas prečítania dokumentu

$S_T(d)$ ohodnocuje dokumenty v závislosti od času prečítania dokumentu, pričom vychádzame z *krivky zabúdania* (Obr. 6.1). Dokumenty prečítané v poslednom sedení (v čase t_0) ohodnotí najvyšším skóre; toto ohodnotenie je nižšie s každým ďalším sedením v minulosti (ďalej na číselnej osi).



Obr. 6.1 Závislosť $S_T(d)$ ohodnotenia dokumentu d od času čítania (sedenia).

6.3.2 Popularita dokumentov

$S_P(d)$ vyjadruje mieru popularity dokumentov, t.j. ohodnocuje jednotlivé dokumenty na základe počtu videní (čítaní) daných dokumentov všetkými používateľmi za posledný čas. V tomto prípade sa spoliehame na „múdrosť davu". Vychádzame z predpokladu, že to, čo je aktuálne populárne (najviac čítané), môže používateľovi pomôcť pri opakovaní podstatných a užitočných dokumentov, najmä v situácii pred písomkou, resp. skúškou či v prípade, že on sám v poslednom čase nebol veľmi aktívny a veľa dokumentov neprečítal.

Formálne môžeme toto skóre vyjadriť takto:

$$S_P(d) = \frac{r(d)}{\sum_i r(d_i)} \tag{6.2}$$

kde $r(d)$ vyjadruje počet čítaní dokumentu všetkými používateľmi za posledný čas. Pretože môže byť problematická definícia toho, čo to znamená „v poslednom čase", môžeme výpočet skóre $S_P(d)$ upraviť tak, že použijeme skóre $S_T(d)$ zadefinované vyššie:

$$S_P(d) = \frac{1}{n}\sum_i S_T^i(d) \tag{6.3}$$

kde n je počet všetkých používateľov a $S_T^i(d)$ predstavuje skóre $S_T(d)$ vypočítané pre daného používateľa i; ide teda o priemerné skóre na základe času čítania dokumentu všetkými používateľmi (ak veľa používateľov daný dokument v poslednom čase čítalo, bude priemer vyšší ako pri dokumente, ktorý nečítal takmer nikto).

Iný variant môže pri počítaní skóre $S_P(d)$ brať do úvahy nie všetkých, ale napr. len „dobrých" používateľov (buď explicitne vyjadrené v modeli používateľa napr. ako výsledok písomky, prípadne implicitne na základe miery znalosti konceptov).

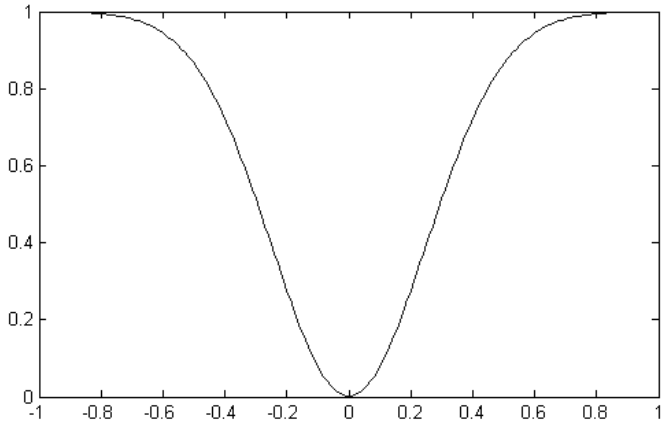
6.3.3 Podobnosť s cieľmi používateľa

$S_G(d)$ ohodnocuje dokumenty na základe ich podobnosti s cieľmi používateľa (ak sú tieto známe). Skóre dokumentu d je určené ako hodnota kosínusovej podobnosti medzi vektorom cieľov $\vec{g}$ používateľa a vektorom naviazaných konceptov $\vec{c}$ dokumentu:

$$S_G(d) = \mathrm{sim}(\vec{c}, \vec{g}) = \frac{\vec{c} \cdot \vec{g}}{\|\vec{c}\| \, \|\vec{g}\|} \tag{6.4}$$

6.3.4 Zmena znalostí konceptov

$S_K(d)$ ohodnocuje dokumenty na základe zmeny znalostí konceptov počas posledného sedenia. Pre každý koncept v doméne platí, že používateľova znalosť daného konceptu buď rástla, klesala, alebo sa nezmenila.



Obr. 6.2 Funkčné ohodnotenie konceptu c_i z domény v závislosti od používateľovej zmeny znalosti tohto konceptu (os x).

Pri určení skóre vychádzame z toho, že ak znalosť niektorého konceptu rástla, znamená to, že sa v poslednom sedení daný koncept používateľ naučil, alebo ho lepšie pochopil a chceme teda, aby si ho zopakoval a utvrdil si túto znalosť. Na druhej strane, ak znalosť konceptu klesala, znamená to, že niečo, čo predtým používateľ vedel, zabudol, alebo to predtým zle pochopil (resp. sme mali v modeli nepresnú hodnotu). Aj takýto koncept chceme podporiť, aby si používateľ osviežil zabudnuté, alebo lepšie pochopil to, čo mu zjavne robí problémy.

Každému konceptu priradíme hodnotu na základe funkcie na Obr. 6.2. Z týchto hodnôt vytvoríme vektor zmien znalostí konceptov a skóre dokumentu d vypočítame opäť pomocou kosínusovej podobnosti medzi daným vektorom a vektorom konceptov naviazaných k dokumentu d.

6.4 Diskusia

Zamerali sme sa na scenár opakovania vo výučbovom systéme, čo ovplyvnilo výber a nastavenie hodnotičov v nami navrhnutej metóde sumarizácie; okrem toho sme museli uvažovať aj ďalšie aspekty opakovania, a to predovšetkým čas opakovania a výber dokumentov na opakovanie.

Navrhnutá metóda personalizovaného výberu dokumentov zohľadňuje rôzne informácie z modelu používateľa a domény. Vychádzame z existujúcich prístupov odporúčania, pričom za náš príspevok v tejto oblasti považujeme návrh ohodnotenia dokumentov so zohľadnením zmeny znalostí používateľa.

Keďže naším cieľom je podporiť opakovanie, navrhnutá metóda vyberá (odporúča) len spomedzi dokumentov, ktoré používateľ už niekedy čítal, navyše musia byť pre tieto dokumenty splnené všetky prerekvizity. V tom spočíva aj obmedzenie našej metódy, keď nikdy nevyberie (neodporučí) nový (používateľom ešte nevidený) dokument. Aj toto obmedzenie sme však schopní odstrániť relaxovaním danej podmienky a navrhnutým zohľadnením popularity dokumentov (t.j. vieme vybrať dokumenty, ktoré sú populárne, hoci ich daný používateľ ešte sám nečítal).

7 Overenie metódy personalizovanej sumarizácie

Cieľom overenia navrhnutej metódy personalizovanej sumarizácie bolo ukázať, že zohľadnením dodatočných informácií dosiahneme lepšie výsledky sumarizácie v porovnaní s generickým variantom, t.j. vety vybrané do sumarizácie budú reprezentatívnejšie a výsledná sumarizácia bude tým pádom presnejšie reprezentovať daný dokument s ohľadom na to, čo je dôležité pre používateľa.

Metódu sme realizovali pomocou vlastného sumarizátora, ktorý je nezávislý od zvolenej domény; pre účely overenia sme si už však zvolili konkrétnu doménu – *doménu výučby*. Overenie sme realizovali v prostredí výučbového systému ALEF, ktorý sme rozšírili o sumarizačný komponent využívajúci nami implementovaný sumarizátor.

Uskutočnili sme dva experimenty na predmetoch Funkcionálne a logické programovanie a Princípy softvérového inžinierstva. V prvom experimente sme sa zamerali na porovnanie sumarizácie zohľadňujúcej relevantné doménové pojmy s generickým variantom, v druhom sme s generickým variantom porovnávali personalizované sumarizácie zohľadňujúce poznámky (zvýraznenia) používateľov.

Uvažujúc zvolený scenár použitia sumarizácie pre opakovanie, realizovali sme aj navrhnutú metódu personalizovaného výberu dokumentov na opakovanie, túto sme však experimentálne neoverovali.

7.1 Realizácia navrhnutej metódy sumarizácie

Implementovali sme vlastný sumarizátor realizujúci navrhnutú metódu ako REST (angl. *Representational State Transfer*) webovú službu. To znamená, že ide o klient – server architektúru, kde server predstavuje nami implementovaná služba sumarizátora a klienti sú používatelia alebo iné systémy, ktoré komunikujú so sumarizátorom pomocou HTTP protokolu.

Vďaka tomuto prístupu je sumarizátor nezávislý od zvolenej domény a konkrétneho systému. Umožnilo nám to tiež oddeliť časť logiky na stranu klienta, ktorý je tak zodpovedný za prvotné predspracovanie dokumentu (jeho prevod do obyčajného textu) a prípadnú komunikáciu s modelom používateľa a domény, z ktorých získava potrebné dáta a tieto posiela spolu s dokumentom v požadovanom formáte sumarizátoru. Zobrazenie výslednej sumarizácie je opäť na strane klienta.

Pre komunikáciu so sumarizátorom a jeho konfiguráciu zo strany klienta sme navrhli a implementovali vlastný protokol vo formáte JSON (podrobnejšie opisujeme v prílohe).

Predspracovanie dokumentu na strane sumarizátora pozostáva z troch krokov, pričom každý z týchto krokov sme realizovali s využitím existujúcich služieb:

1. Strojový preklad dokumentu do angličtiny
2. Tokenizácia a extrakcia termov
3. Rozdelenie, t.j. segmentácia dokumentu na vety

Strojovým prekladom dokumentu do angličtiny maximalizujeme nezávislosť sumarizátora od jazyka sumarizovaného dokumentu, čo je dôležitá požiadavka, ak chceme sumarizátor používať v prostredí otvoreného webu. Navyše aj v systéme ALEF, v ktorom sme overovali nami navrhnutú metódu sumarizácie, sú už teraz dokumenty nielen v slovenčine, ale aj

v češtine a je možné, že do budúcnosti pribudnú aj dokumenty v anglickom jazyku. Taktiež podpora existujúcich služieb pre extrakciu kľúčových slov z dokumentov v inom ako anglickom jazyku je minimálna.

Na preklad dokumentu sme využili službu *Bing Translator*[10], ktorý okrem prekladu poskytuje aj službu na detekciu jazyka dokumentu a tiež službu na segementáciu textu na vety, ktorú využívame pri treťom kroku predspracovania. Pri sumarizácii delíme na vety okrem preloženého dokumentu aj pôvodný, aby sme nakoniec mohli výslednú sumarizáciu zložiť z vybraných viet pôvodného dokumentu.

Na tokenizáciu a extrakciu termov dokumentu využívame *ElasticSearch*[11], čo je v prvom rade vyhľadávač dokumentov, no poskytuje aj službu na analýzu dokumentu a extrakciu termov. Jeho výhodou je, že sa vykonáva lokálne, čím sa odstraňuje závislosť od externej služby, ktorá nemusí vždy fungovať.

Spomedzi navrhnutých hodnotičov sme implementovali tieto:

- Hodnotič frekvencie výskytu termov (tf-idf)
- Hodnotič polohy (podľa nadpisu, prvej a poslednej vety)
- Hodnotič relevantných doménových pojmov
- Hodnotič vedomostí
- Hodnotič poznámok (zvýraznení)

Implementované hodnotiče nám umožnili porovnať základný *generický variant* sumarizácie (použitie kombinácie hodnotiča frekvencie výskytu termov s hodnotičom polohy termov) s *variantom zohľadňujúcim relevantné doménové pojmy* (kombinácia predchádzajúcich s hodnotičom relevantných doménových pojmov) a *variantom zohľadňujúcim poznámky* (zvýraznenia) používateľov (opäť prvé dva hodnotiče, tentokrát však v kombinácii s hodnotičom anotácií).

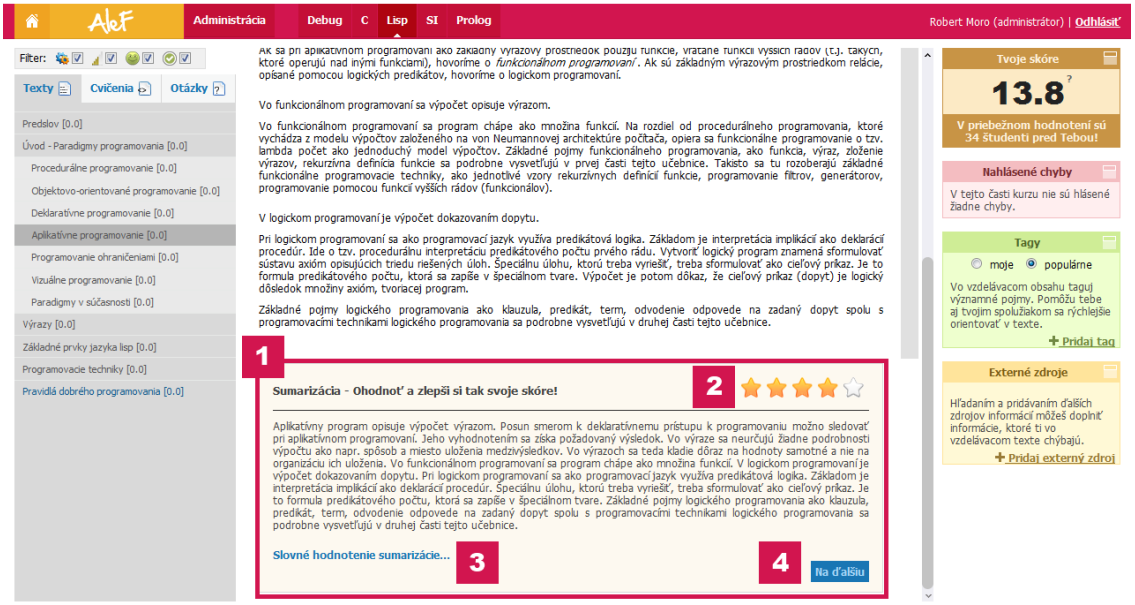
Prepojenie sumarizátora s výučbovým systémom ALEF sme realizovali pomocou sumarizačného komponentu (Obr. 7.1), ktorý zabezpečuje komunikáciu so sumarizátorom prostredníctvom poskytnutého REST rozhrania a prezentuje používateľom výsledné sumarizácie vzdelávacích textov.

Sumarizačný komponent sme využili pri experimentálnom overovaní navrhnutej metódy, preto umožňuje používateľom (študentom) hodnotiť kvalitu poskytnutých sumarizácií a tiež navigovať sa na ďalšiu sumarizáciu v poradí. Pri reálnom používaní asi nemá veľký zmysel navigovať sa takýmto spôsobom, túto funkcionalitu sme implementovali za účelom efektívnejšieho (rýchlejšieho) hodnotenia sumarizácií používateľmi.

Komponent je zároveň možné konfigurovať – zvoliť umiestnenie nad alebo pod vzdelávacím textom. Pri overovaní sme volili väčšinou umiestnenie pod vzdelávacím textom, aby mohli používatelia vyhodnotiť, ako dobre sumarizácia reprezentuje daný text; v reálnom použití má však zrejme väčší zmysel umiestnenie nad vzdelávací text (hoci si vieme pred-

[10] http://msdn.microsoft.com/en-us/library/ff512419.aspx
[11] http://www.elasticsearch.org

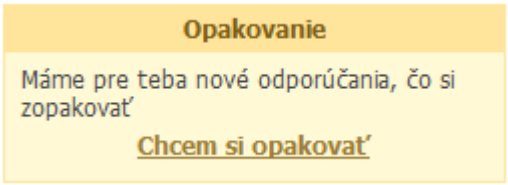
staviť aj scenár, kedy by bolo vhodné umiestnenie pod textom, napr. poskytnutie zhrnutia najdôležitejších bodov, čo si používateľ prečítal, a čo si má z textu zapamätať).



Obr. 7.1 Sumarizačný komponent v prostredí systému ALEF (1): používatelia môžu hodnotiť sumarizácie na päťstupňovej škále pomocou hviezdičiek (2), pridávať slovné hodnotenie (3) a navigovať sa na ďalšiu sumarizáciu v poradí (4).

7.2 Realizácia metódy výberu dokumentov na opakovanie

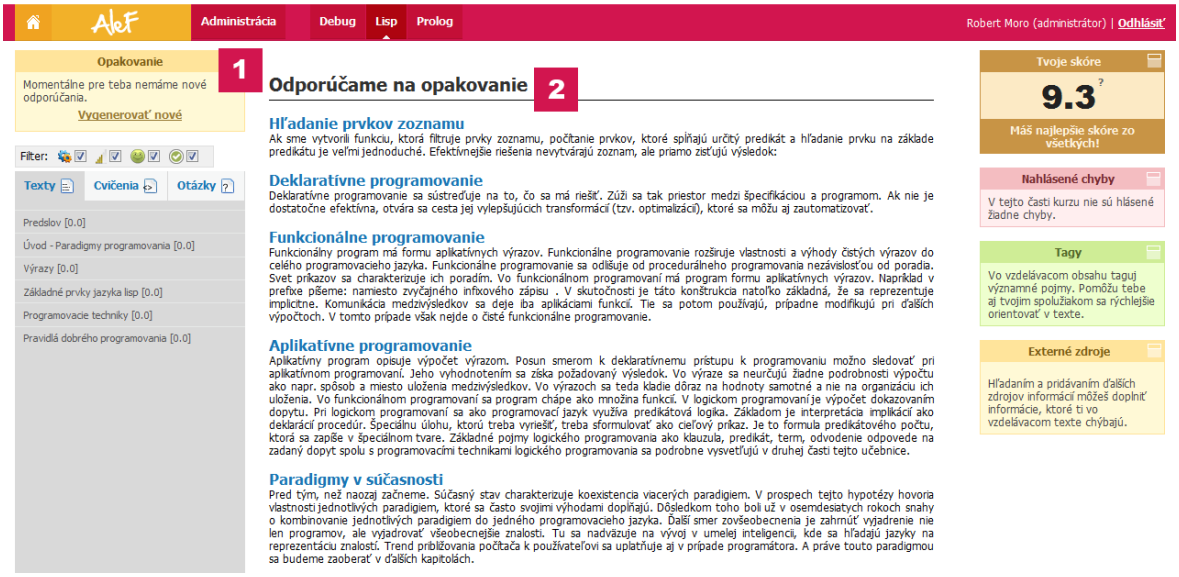
Metódu personalizovaného výberu dokumentov na opakovanie sme realizovali v prostredí výučbového systému ALEF, ktorý sme rozšírili o ďalší komponent (Obr. 7.2). Ten slúži len na to, aby si používateľ (študent) mohol zvoliť, že si chce opakovať. Následne sa mu zobrazí v hlavnej (obsahovej) časti systému zoznam dokumentov vybraných (odporúčaných) na opakovanie spolu s ich sumarizáciami (Obr. 7.3). Kliknutím na názov niektorého z odporúčaných dokumentov sa používateľ dostane k jeho celému textu.



Obr. 7.2 Komponent pre opakovanie.

V prípade, že nie sú k dispozícii ďalšie nové odporúčania dokumentov na opakovanie (napr. ak si ich používateľ práve zobrazil ako je naznačené na Obr. 7.3), môže si používateľ vyžiadať ďalšie.

Z navrhnutej metódy sme realizovali ohodnotenie dokumentov podľa zmeny znalostí používateľa v čase, tzn. že sa odporúčajú dokumenty súvisiace s konceptmi, ktorých modelovaná znalosť používateľovi v poslednom čase najviac klesla alebo naopak najviac stúpla. Keďže sme sa pri overovaní zamerali predovšetkým na vyhodnotenie navrhnutej metódy personalizovanej sumarizácie, metódu výberu dokumentov na opakovanie sme (aj pre nedostatok času) experimentálne neoverovali.



Obr. 7.3 Sumarizácia pre opakovanie: používateľ si môže vyžiadať nové odporúčania, ak nie sú k dispozícii (1), dokumenty vybrané na opakovanie (2) sú zosumarizované, pričom v prípade potreby si môže používateľ zobraziť aj ich celé texty kliknutím na názov niektorého z nich.

7.3 Spôsob overenia

S navrhnutou metódou personalizovanej sumarizácie sme experimentovali v doméne výučby uvažujúc konkrétny scenár opakovania vo výučbovom systéme. Využili sme výučbový systém ALEF, pričom náš dátový súbor pozostával z výučbových textov z kurzov *Funkcionálne a logické programovanie (FLP)* a *Princípy softvérového inžinierstva (PSI)*.

Kvalitu sumarizácií sme vyhodnocovali pomocou hodnotení používateľov (študentov), ktorým sme prezentovali náhodne niektorý z overovaných variantov sumarizácie a ich úlohou bolo hodnotiť tieto sumarizácie na päťstupňovej škále (pomocou hviezdičiek). Po každom ohodnotení sme študentom položili doplňujúcu otázku z nami navrhnutého dotazníka:

1. Sú vybrané vety reprezentatívne, t.j. dobre vystihujú (zhrnujú) obsah dokumentu (najdôležitejšie informácie z neho)?
    a. áno, vybrané vety sú reprezentatívne
    b. áno, prevažujú reprezentatívne vety
    c. čiastočne - sú zastúpené rovnako reprezentatívne aj nereprezentatívne vety
    d. nie, väčšinou sú nevhodné
    e. nie, vôbec nie sú reprezentatívne
    f. neviem rozhodnúť
2. Predstavte si situáciu, že by ste mali túto sumarizáciu k dispozícii na zopakovanie si obsahu prečítaného textu. Pomohla by vám?
    a. áno, určite
    b. áno, sumarizácia obsahovala väčšinu podstatných informácií
    c. čiastočne, niektoré dôležité časti v sumarizácii boli, ale niektoré nie
    d. nie
    e. neviem rozhodnúť
3. Predstavte si situáciu, že by ste sa mali na základe poskytnutej sumarizácie rozhodnúť, či je tento text pre vás v danom momente relevantný a treba ho prečítať celý. Pomohla by vám?

a. áno, sumarizácia dokonca zhrnula text tak, že by som ho už ani nemusel/a celý čítať a väčšinu podstatného by som sa dozvedel/a
b. áno
c. nie, nevedel/a by som na základe poskytnutej sumarizácie povedať, či je pre mňa text relevantný
d. neviem rozhodnúť

4. Je daná sumarizácia čitateľná (zrozumiteľná)?
   a. áno, nie je problém s pochopením a čitateľnosťou sumarizácie
   b. áno, sumarizácia je aj napriek niekoľkým problémom zrozumiteľná
   c. čiastočne
   d. nie, obsahuje nekompletné vety
   e. nie, obsahuje vety vytrhnuté z kontextu, ktoré sa odkazujú na niečo, čo sa v sumarizácii nespomína
   f. nie, obsahuje nekompletné vety aj vety vytrhnuté z kontextu
   g. neviem rozhodnúť
5. Je zvolená dĺžka sumarizácie vhodná vzhľadom na dĺžku (a obsah) sumarizovaného dokumentu?
   a. áno, dĺžka je vyhovujúca
   b. nie, sumarizácia je príliš krátka
   c. nie, sumarizácia je príliš dlhá
   d. uvítal/a by som, keby som si mohol/mohla meniť dĺžku podľa potreby
   e. neviem rozhodnúť

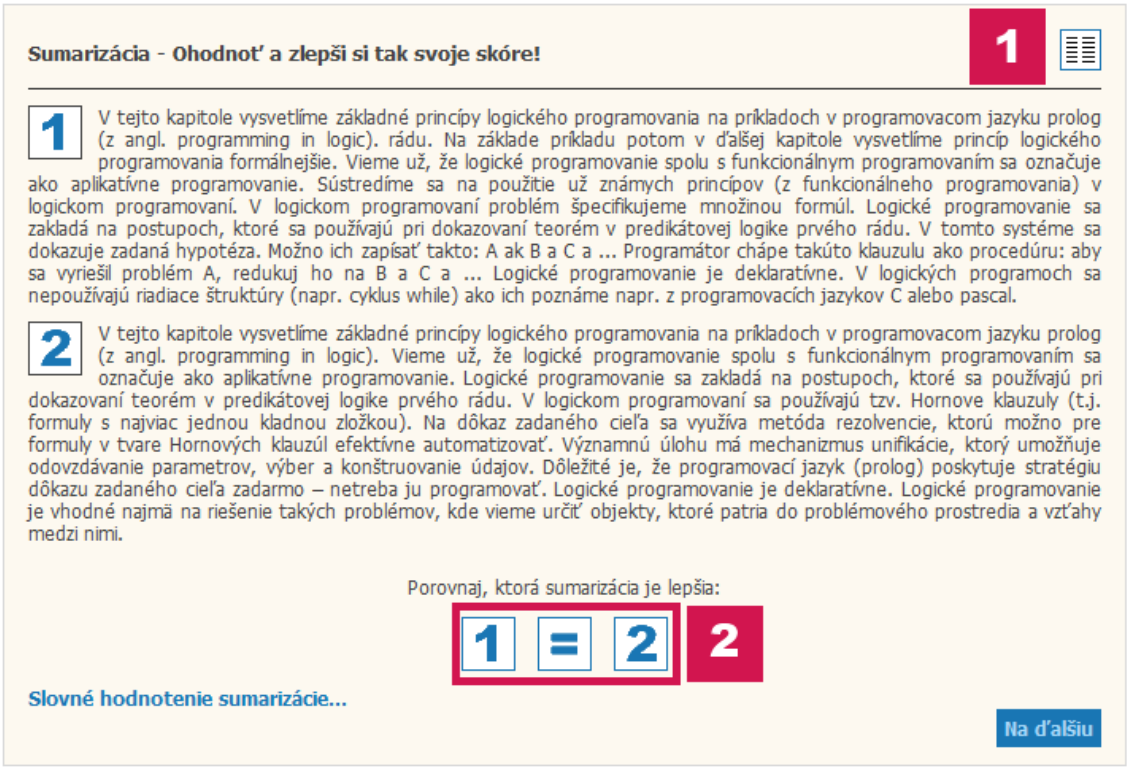
Najviac nás zaujímalo, či sa sumarizáciám darí napĺňať ich hlavný cieľ – zosumarizovať dokument tak, aby poslúžil používateľovi na zopakovanie konceptov preberaných v dokumente, na čo sme sa priamo aj nepriamo pýtali v otázkach 1-3.

Okrem hodnotenia hviezdičkami a odpovedí na doplňujúce otázky mohli študenti poskytnúť spätnú väzbu aj v podobe slovného hodnotenia.

Navyše sme spomedzi študentov vybrali „expertnú“ skupinu najlepších (podľa ich úspešnosti na predmete, resp. v doterajšom štúdiu), u ktorých bol predpoklad, že budú schopní objektívne vyhodnotiť kvalitu poskytnutých sumarizácií na základe ich znalostí z danej domény. Títo vyhodnocovali sumarizácie priamym porovnaním overovaných variantov (bez znalosti o tom, ktorý variant bol vygenerovaný ako metódou), t.j. mali označiť, ktorý z dvoch prezentovaných variantov je lepší, resp. či sú kvalitatívne rovnaké (Obr. 7.4).

Pri overovaní metód sumarizácie sa v literatúre (Ježek & Steinberger, 2010; Díaz et al., 2005) môžeme často stretnúť s použitím *kvantitatívnych metód overenia*, ktoré porovnávajú výslednú sumarizáciu s „ideálnou“, väčšinou manuálne vytvorenou (napr. autorom článku) pomocou rôznych metrík ako napr. presnosť, úplnosť, F-skóre, ROUGE a pod.

Problém však je, že v skutočnosti neexistuje niečo ako ideálna sumarizácia, a ak necháme jeden text sumarizovať rôznymi používateľmi, s veľkou pravdepodobnosťou dostaneme rôzne výsledné sumarizácie. To je ešte viac zvýraznené, ak chceme overiť personalizovanú sumarizáciu, kde už z definície je ideálna sumarizácia pre rôznych používateľov rôzna. Z týchto dôvodov nebolo možné použiť pri overení niektorý zo štandardných dátových súborov (setov). Navyše pri dostupných štandardných dátových súboroch nie sú k dispozícii informácie o používateľoch či doméne, ktoré využíva nami navrhnutá metóda (a teda by sme takýmto spôsobom mohli overiť jedine generický variant sumarizácie). Preto sme zvolili opísaný spôsob overenia v prostredí výučbového systému ALEF.



Obr. 7.4 Porovnanie variantov sumarizácie: je možné prepnúť zobrazenie do dvojstĺpcového formátu (1); používateľ (expert) sa kliknutím na jednu z možností vyjadrí, či je lepší prvý, druhý variant alebo sú kvalitatívne rovnaké (2).

7.4 Sumarizácia zohľadňujúca relevantné doménové pojmy

Uskutočnili sme experiment na predmete Funkcionálne a logické programovanie (FLP) za účelom vyhodnotenia *sumarizácie zohľadňujúcej relevantné doménové pojmy* (pomocou príslušného hodnotiča) v porovnaní s generickou sumarizáciou.

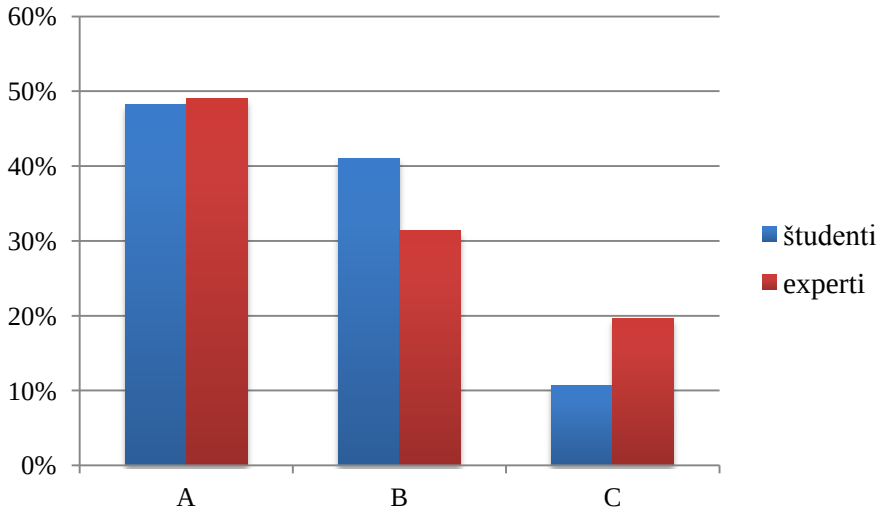
Podarilo sa nám od študentov získať hodnotenia k sumarizáciám všetkých 79 výučbových textov v danom kurze. Celkovo sme zaznamenali 278 hodnotení sumarizácií, 154 porovnaní variantov sumarizácií expertmi a 275 odpovedí na doplňujúce otázky. Variant zohľadňujúci relevantné doménové pojmy (RDT) získal v priemere vyššie skóre na päťstupňovej škále ako generický variant (pozri Tab. 7.1).

Tab. 7.1 Hodnotenia variantov sumarizácií.

	Generická	Zohľadňujúca RDT
Počet hodnotení	143	135
Priemerné hodnotenie	3,54	3,79
Rozptyl (n-1)	1,52	1,42

Podobne sme vypočítali priemerné hodnotenie pre každý variant pre jednotlivé sumarizované dokumenty. Variant zohľadňujúci RDT mal vyššie priemerné skóre v 48% prípadov, nižšie v 41% a rovnaké v 11%. Jednoznačnejšie výsledky sme dostali na základe porovnaní variantov expertmi. Variant zohľadňujúci RDT bol expertmi vyhodnotený ako lepší v 49%

prípadov, ako horší v 31% a ako kvalitatívne rovnaký v 20% (pozri graf na Obr. 7.5). Výsledky experimentu teda naznačujú, že zohľadnenie RDT v procese sumarizácie vedie k lepším výsledkom v porovnaní s generickým variantom.



Obr. 7.5 Porovnanie variantov sumarizácie: A znamená, že sumarizácia zohľadňujúca relevantné doménové pojmy bola vyhodnotená ako lepšia; B, že lepšie bola hodnotená generická sumarizácia a C, že boli vyhodnotené narovnako (ako rovnaké).

Vyhodnotili sme tiež odpovede študentov na doplňujúce otázky (presné čísla a grafy uvádzame v prílohe). Na základe nich môžeme konštatovať, že nami navrhnutá metóda vo všeobecnosti vyberá reprezentatívne vety a produkované sumarizácie sú vhodné aj pre opakovanie, resp. pomáhajú rozhodnúť sa, či je sumarizovaný dokument relevantný pre konkrétneho používateľa (t.j. mal by po prečítať celý).

Ukázalo sa tiež, že napriek niektorým problémom, sumarizácie sú vo väčšine prípadov dobre čitateľné a zrozumiteľné. Medzi pretrvávajúce problémy môžeme zaradiť problém nekompletných viet a straty kontextu. Prvý problém spôsobuje externá služba (*Bing Translate*), ktorá nedokáže vždy správne segmentovať text na vety. Druhý problém spôsobujú *koreferencie (anafory)*, t.j. odkazy v texte na niečo, čo bolo spomenuté skôr. V prípade, že sa do sumarizácie vyberie veta, ktorá sa odkazuje na niečo spomenuté v texte a to, na čo sa odkazuje, sa do sumarizácie už nedostane, zákonite nastáva problém straty kontextu. Riešením by bolo nahrádzanie koreferencií entitami, na ktoré sa odkazujú, avšak existujúce metódy sú závislé na jazyku, a preto sme to neriešili (vzhľadom na použitie strojového prekladu a snahu minimalizovať závislosť na jazyku sumarizovaného dokumentu).

Čo sa dĺžky sumarizácií týka, na základe odpovedí môžeme usúdiť, že sme kritérium dĺžky nastavili pre dané texty dobre – sumarizácie predstavovali zhruba 30% pôvodného textu, avšak obmedzili sme maximálnu dĺžku v prípade príliš dlhých dokumentov. Približne 80% študentov odpovedalo, že dĺžka je vyhovujúca, niečo cez 10% študentov by však uvítalo možnosť meniť dĺžku podľa svojich preferencií.

7.5 Sumarizácia zohľadňujúca poznámky

Na vyhodnotenie *sumarizácie zohľadňujúcej poznámky* sme uskutočnili dva nezávislé experimenty porovnávajúce tento variant s generickou sumarizáciou. Prvý experiment prebiehal opäť na predmete Funkcionálne a logické programovanie (FLP), druhý na predmete

Princípy softvérového inžinierstva (PSI). Konkrétne nastavenie parametrov metódy sumarizácie (použité hodnotiče a ich koeficienty) pre všetky opísané experimenty uvádzame v prílohe.

Výučbové texty na predmete PSI sa líšili od tých na FLP tým, že boli oveľa stručnejšie a mali viac povahu poznámok než učebnice. Tomu sme prispôsobili aj sumarizácie, ktorým sme nastavili oveľa kratšiu dĺžku – cieľom bolo vybrať najreprezentatívnejšie vety, resp. ich časti. Študentov sme nechali hodnotiť sumarizácie pomocou hviezdičiek, nepoužili sme v tomto prípade ani doplňujúce otázky, ani expertné vyhodnotenie.

V priebehu jedného týždňa, počas ktorého prebiehal experiment, sme získali hodnotenia sumarizácií pre všetkých 185 výučbový textov v kurze PSI v ALEFe. V Tab. 7.2 je uvedený počet hodnotení študentov ako aj priemerné hodnoty. Vidíme, že rozdiel v prospech zohľadnenia poznámok je len veľmi malý (0,08). Je to spôsobené zrejme povahou textov, keď takmer všetky informácie v nich sú relevantné a podstatné, a preto nech sumarizácia vyberie takmer čokoľvek, nebudú sa hodnotenia od seba asi nijak výrazne líšiť.

Tab. 7.2 Hodnotenia variantov sumarizácií na predmete PSI.

	Generická	**Zohľadňujúca poznámky**
Počet hodnotení	837	870
Priemerné hodnotenie	3,32	3,40
Rozptyl (n-1)	1,24	1,26

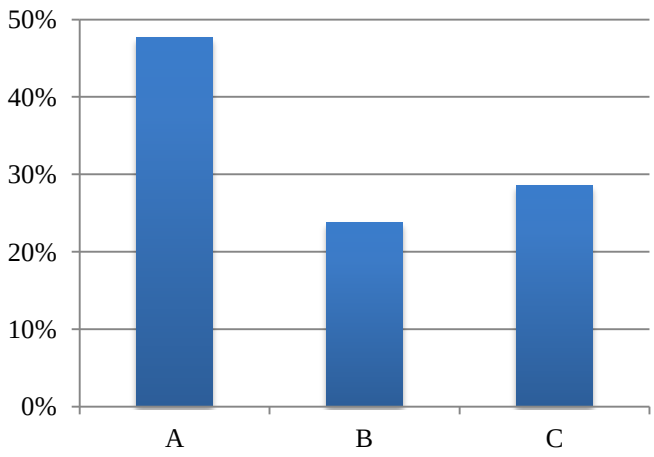
Paralelne s experimentom na PSI sme uskutočnili druhý experiment na predmete FLP. Zvolili sme rovnaký spôsob vyhodnotenia ako v prípade prvého experimentu, t.j. okrem hodnotení študentov sme zaznamenávali aj ich odpovede na doplňujúce otázky a zvolenú expertnú skupinu študentov sme nechali hodnotiť varianty sumarizácie ich priamym porovnaním.

Keďže sme tento experiment uskutočnili až v 10. týždni semestra, zapojilo sa doňho oveľa menej študentov ako pri prvom experimente. Získali sme tak len malý počet hodnotení. Zamerali sme sa preto na vyhodnotenie porovnania variantov expertmi (Obr. 7.6). Môžeme vidieť, že sumarizácia zohľadňujúca poznámky bola hodnotená ako lepšia v 48% prípadov, čo je porovnateľné s výsledkami experimentu so sumarizáciami zohľadňujúcimi RDT. Avšak oveľa nižší je podiel prípadov, kedy bola sumarizácia zohľadňujúca poznámky hodnotená ako horšia (24%). Na základe výsledkov tak môžeme konštatovať, že zohľadnenie poznámok (zvýraznení) študentov v procese sumarizácie vedie k výrazne lepším výsledkom v porovnaní s generickým variantom.

7.6 Diskusia výsledkov experimentov

Pri overovaní sme sa zamerali na porovnanie týchto variantov sumarizácie voči generickej sumarizácii:

- Sumarizácia zohľadňujúca relevantné doménové pojmy
- Sumarizácia zohľadňujúca poznámky v podobe zvýraznení



Obr. 7.6 Porovnanie variantov sumarizácie expertmi: A znamená, že sumarizácia zohľadňujúca poznámky bola vyhodnotená ako lepšia;, B, že lepšie bola hodnotená generická sumarizácia a C, že boli vyhodnotené narovnako (ako rovnaké).

Zistili sme, že zohľadnenie RDT, resp. poznámok v procese sumarizácie prináša zlepšenie oproti generickému variantu. Zároveň môžeme konštatovať, že sumarizácie získané navrhnutou metódou je možné použiť pri rozhodovaní o relevancii dokumentu, ako aj pri opakovaní vo výučbovom systéme.

Overili sme tiež ohraničenia metódy, keď v prípade textov na PSI, ktoré mali charakter poznámok, pričom takmer všetko v nich bolo podstatné, nemalo zohľadnenie poznámok v zásade nijaký efekt na ich kvalitu podľa hodnotení študentov.

V niektorých prípadoch sme zistili isté nedostatky pri segmentácii textu na vety externou službou a tiež stratu kontextu vplyvom koreferencií, čo mohlo ovplyvniť celkové hodnotenie sumarizácie (tzn. ak aj boli do sumarizácie vybrané reprezentatívne vety, vplyvom zníženej čitateľnosti a zrozumiteľnosti dostala daná sumarizácia od študentov nižšie hodnotenie).

Ďalší priestor na zlepšenie výsledkov dosahovaných navrhnutou metódou vidíme v optimalizácii zvolených koeficientov kombinácie hodnotičov a v implementácii metód spracovania prirodzeného jazyka, ktoré by napr. vyriešili spomínaný problém so stratou kontextu.

8 Zhodnotenie

Na základe analýzy existujúcich metód sumarizácie a voľne dostupných riešení sumarizátorov sme si zvolili metódu latentnej sémantickej analýzy ako východisko pre návrh spôsobu prispôsobovania (personalizácie) sumarizácie.

Podrobne sme preskúmali možnosti prispôsobovania sumarizácie zohľadnením informácií z modelu používateľa (znalosti, ciele) a z modelu domény (koncepty, resp. relevantné doménové pojmy). Analyzovali sme tiež možnosť použitia poznámok od používateľov, pričom pri hodnotení vhodnosti jednotlivých typov poznámok sme vychádzali z údajov získaných používaním systému ALEF študentmi na predmete Princípy softvérového inžinierstva; ako najvhodnejšie pre ďalšie úvahy sa ukázali byť zvýraznenia a tagy.

Navrhli sme metódu personalizácie sumarizácie založenú na kombinácii generických a personalizovaných hodnotičov. Za hlavný prínos našej práce považujeme:

- Návrh spôsobu kombinácie hodnotičov, ktoré nám umožňuje uvažovať rôzne parametre a kontext sumarizácie
- Návrh konkrétnych hodnotičov, ktoré zohľadňujú termy relevantné v danej doméne, znalosti a ciele používateľa, ako aj jeho poznámky v podobe zvýraznení a tagov

Hoci je nami navrhnutá metóda nezávislá od domény, navrhnutý spôsob kombinácie nám umožňuje vo zvolenej doméne identifikovať zdroje personalizácie a prispôsobiť metódu pre konkrétny scenár použitia.

My sme sa v práci zamerali na scenár opakovania v doméne výučby. Pre tento účel sme tiež navrhli metódu personalizovaného výberu dokumentov pre opakovanie, ktorá zohľadňuje viaceré aspekty ako čas ich čítania (zabúdanie), popularitu dokumentov a pod. Tým sme našou prácou presiahli do oblasti odporúčania, kde náš prínos vidíme v návrhu ohodnotenia dokumentov na opakovanie na základe zmeny znalostí používateľov v čase.

Navrhnutú metódu sumarizácie sme realizovali prostredníctvom sumarizátora v podobe webovej služby, pričom sme implementovali väčšinu z navrhnutých hodnotičov (hodnotič frekvencie termov, hodnotič polohy termov, hodnotič relevantných doménových pojmov, hodnotič vedomostí a hodnotič poznámok). Implementovaný sumarizátor sme za účelom overenia metódy integrovali s výučbovým systémom ALEF.

Pri overovaní sme sa zamerali na vyhodnotenie a porovnanie rôznych variantov sumarizácie. Výsledky nami vykonaných experimentov na predmete Funkcionálne a logické programovanie a predmete Princípy softvérového inžinierstva naznačujú, že zohľadnenie relevantných doménových pojmov, ako aj poznámok v procese sumarizácie prináša zlepšenie oproti generickému variantu a výsledné sumarizácie sú schopné zhrnúť dôležité koncepty obsiahnuté v dokumentoch aj pre účely opakovania.

Prácu by sme mohli ďalej rozvíjať viacerými smermi. Keďže sme sa sústredili predovšetkým na návrh hodnotičov, predpokladali sme, že koeficienty (váhy) ich kombinácie určí doménový expert v závislosti od konkrétneho scenára použitia. Vhodným rozšírením práce by mohlo byť zautomatizovanie určovania týchto váh a ich dynamické nastavovanie v závislosti od situácie napr. spôsobom, ktorý sme diskutovali v kap. 5.3.2.

Pri overovaní sme experimentovali so sumarizáciami zohľadňujúcimi relevantné doménové pojmy, ktoré zachytil expert v doménovom modeli. Doménové modelovanie je ale vo

všeobecnosti ťažká a časovo náročná úloha. Ako sme však ukázali v (Móro et al., 2011), je možné a užitočné využiť na tento účel folksonómie získané z tagov používateľov. Bolo by preto zaujímavé vyskúšať takto získaný model domény pri sumarizácii.

Ako sme zistili pri overovaní, problémom extraktívnej sumarizácie býva strata kontextu viet vybraných do súhrnu. Taktiež veta môže byť príliš veľká jednotka, najmä ak ide o dlhé súvetie zložené z viacerých vedľajších (doplňujúcich) viet. Výzvou je poskytnúť sumarizácie, ktoré sa vzdialia od jednoduchého extraktu a priblížia sa viac smerom k pravým abstraktom. To by umožnilo nahrádzať koreferencie entitami, na ktoré sa odkazujú, vynechávať menej podstatné časti viet (súvetí), parafrázovať a pod. Táto oblasť si vyžaduje použitie a návrh zložitejších metód spracovania prirodzeného jazyka a môže byť námetom pre výskum na ďalšom stupni štúdia.

V neposlednom rade sme v práci navrhli metódu personalizovaného výberu dokumentov na opakovanie. Keďže táto metóda presiahla pôvodný rámec práce, sústredili sme sa predovšetkým na overenie navrhnutej metódy sumarizácie a metódu výberu dokumentov na opakovanie sme ďalej (aj z časových dôvodov) neoverovali. Bolo by preto zaujímavé ju ďalej rozvinúť a experimentálne overiť.

Zoznam použitej literatúry

1. ADOMAVICIUS, G. – TUZHILIN, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. In: *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no.6, IEEE Educational Activities Department, pp.734-749.
2. BARLA, M. – BIELIKOVÁ, M. (2010). Ordinary web pages as a source for metadata acquisition for open corpus user modeling. In: *Proc. of IADIS WWW/Internet 2010*, IADIS Press, pp. 227-233.
3. BIELIKOVÁ, M. – NAGY, P. (2006). Considering Human Memory Aspects for Adaptation and Its Realization in AHA! In: *EC-TEL '06: Technology Enhanced Learning*, W. Nejdl, K. Tochtermann (Eds.), LNCS, vol. 4227, Springer Verlag, pp. 8–20.
4. BOYDELL, O. – SMYTH, B. (2007). From social bookmarking to social summarization: An experiment in community-based summary generation. In: *IUI '07: Proc. of the 12th Int. Conf. on Intelligent User Interfaces*, ACM Press, pp. 42-51.
5. BRUSILOVSKY, P. (1996). Methods and techniques of adaptive hypermedia. In: *User Modeling and User-Adapted Interaction*, vol. 6, no. 2-3, pp. 87-129.
6. BRUSILOVSKY, P. (2001). Adaptive hypermedia. In: *User Modeling and User-Adapted Interaction*, vol. 11, no. 1-2, pp. 87-110.
7. BRUSILOVSKY – P., MILLÁN, E. (2007). User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: *The Adaptive Web: Methods and Strategies of Web Personalization*, P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.), LNCS, vol. 4321, Springer Verlag, pp. 3-53.
8. CAMPANA, M. – TOMBROS, A. (2009). Incremental personalised summarisation with novelty detection. In: *FQAS '09: Proc. of the 8th Int. Conf. on Flexible Query Answering Systems*, 2009, Springer, Berlin, pp. 641-652.
9. CARBONELL, J., GOLDSTEIN, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: *SIGIR '98: Proc. of the 21st Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval*, ACM Press, pp. 335-336.
10. CIMIANO, P. (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag, 347 p. ISBN: 978-0-387-30632-2.
11. DAS, D. – MARTINS, A.F.T (2007). A survey on automatic text summarization. http://www.dipanjandas.com/files/summarization.pdf (2012-05-03)
12. DELORT, J.Y. – BOUCHON-MEUNIER, B. – RIFQI, M. (2003). Enhanced web document summarization using hyperlinks. In: *HYPERTEXT '03: Proc. of the 14th ACM Conf. on Hypertext and Hypermedia*, ACM Press, pp. 208-215.
13. DELORT, J.Y. (2006). Identifying commented passages of documents using implicit hyperlinks. In: *HYPERTEXT '06: Proc. of the 17th ACM Conf. on Hypertext and Hypermedia*, ACM Press, pp. 89–98.
14. DÌAZ, A. – GERVÀS, P. – GARZÌA, A. (2005). Evaluation of a system for personalized summarization of web contents. In: *User Modeling 2005*, LNCS, vol. 3538, Springer, Berlin, pp. 453–462.

15. EDMUNDSON, H.P. (1969). New methods in automatic extracting. In: *J. of the Association for Computing Machinery*, vol. 16, no. 2, pp. 264–285.
16. ENCARNAÇAO, L.M. (1997). Multi-level user support through adaptive hypermedia: A highly application-independent help component. In: *Proc. of Int. Conf. on Intelligent User Interfaces*, ACM Press, pp. 187-194.
17. GHAUTH, K.I.B. – ABDULLAH, N.A. (2010). Learning materials recommendation using good learners' ratings and content-based filtering. In: *Educational Technology Research and Development*, vol. 58, no. 6, Springer, pp. 711-727.
18. GONG, X. – LIU X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In: *SIGIR '01: Proc. of the 24th Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval*, ACM Press, pp. 19-25.
19. HOLLINK, V., SOMEREN, M.V., HAGE, S.T. (2005). Discovering stages in web navigation. In: *Proc. of 10th Int. User Modeling Conf.*, LNAI, vol. 3538. Springer Verlag, pp. 473-482.
20. HOLUB, M. – BIELIKOVÁ, M. (2010). Estimation of user interest in visited web page. In: *WWW '10: Proc. of Int. Conf. on World Wide Web*, ACM Press, pp. 1111-1112.
21. HÖÖK, K. et al. (1996). A glass box approach to adaptive hypermedia. In: *User Modeling and User-Adapted Interaction*, vol. 6, no. 2-3, Kluwer Academic Publishers, pp. 157-184.
22. JEŽEK, K. – STEINBERGER, J. (2008). Automatic text summarization: The state of the art 2007 and new challenges. In: *Proc. of Znalosti 2008*, Vydavateľstvo STU, pp. 1–12.
23. JEŽEK, K. – STEINBERGER, J. (2010). Sumarizace textů. In: *Proc. of Annual Database Conf. DATAKON*, pp. 3-23.
24. JIN, X., ZHOU, Y., MOBASHER, B. (2005). Task-Oriented Web User Modeling for Recommendation. In: *Proc. of 10th Int. User Modeling Conf.*, LNAI, vol. 3538. Springer Verlag, pp. 109-118.
25. JONES, K.S. (2007). Automatic summarizing: The state of the art. In: *J. of Information Processing and Management: an Int. Journal*, vol. 43, no. 6, Pergamon Press, pp. 1449-1481.
26. KUPIEC, J. – PEDERSEN, J. – CHEN, F. (1995). A trainable document summarizer. In: *SIGIR '95: Proc. of the 18th Annual Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval*, ACM Press, pp. 68-73.
27. LABAJ, M. (2011a). Odporúčanie a kolaborácia na základe implicitnej spätnej väzby. Diplomová práca, FIIT STU, Bratislava, 57 p.
28. LABAJ, M. (2011b). Web-based learning support based on implicit feedback. In: *Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on Student Research in Informatics and Information Technologies*, vol. 3, no. 2, Vydavateľstvo STU, pp. 76-78.
29. LANDAUER, T.K. – FOLZ, P.W. – LAHAM, D. (1998). An Introduction to Latent Semantic Analysis. In: *Discourse Processes*, vol. 25, no. 2-3, pp. 259-284.
30. LI, X. – GUO, L. – ZHAO, Y.E. (2008). Tag-based social interest discovery. In: *WWW '08: Proc. of the 17th Int. Conf. on World Wide Web*, ACM Press, pp. 675-684.
31. LUHN, H.P. (1958). The automatic creation of literature abstracts. In: *IBM J. of Research Development*, vol. 2, no. 2, pp. 159–165.

32. MANI, I. – BLOEDERN, E. (1998). Machine learning of generic and user-focused summarization. In: *AAAI '98: Proc. of the 15th National Conf. on Artificial Intelligence*, American Association for Artificial Intelligence Menlo Park, pp. 820–826.
33. MARSHALL, C.C. (1998). Toward an ecology of hypertext annotation. In: *HYPERTEXT '98: Proc. of the 9th ACM Conf. on Hypertext and Hypermedia*, ACM Press, pp. 40-49.
34. MIHÁL, V. – BIELIKOVÁ, M. (2009). An Approach to Annotation of Learning Texts on Programming within a Web-Based Educational System. In: *SMAP '09: Proc. of the 4th Int. Workshop on Semantic Media Adaptation and Personalization*, IEEE CS Press, pp. 99-104.
35. MICHLÍK, P. (2010). Personalizované odporúčanie príkladov s uvažovaním obmedzeného času učenia. Diplomová práca, FIIT STU, Bratislava, 87 p.
36. MICHLÍK, P. – BIELIKOVÁ, M. (2010). Exercises recommending for limited time learning. In: *Procedia Computer Science*, vol. 1, no. 2, Elsevier, pp. 2821–2828.
37. MÓRO, R. – SRBA, I. – UNČÍK, M. – BIELIKOVÁ, M. – ŠIMKO, M. (2011). Towards collaborative metadata enrichment for adaptive web-based learning. In: *WI-IAT '11: Proc. of the 2011 IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology*, vol. 3, IEEE CS Press, pp. 106-109.
38. NAGAO, K. – HASIDA, K. (1998). Automatic text summarization based on the Global Document Annotation. In: *COLING'98: Proc. of the 17th Int. Conf. on Computational Linguistics*, vol. 2, Association for Computational Linguistics, pp. 917-921.
39. NAGY, P. (2006). Prispôsobovanie prezentácie informácií v elektronickom kurze s využitím vlastností ľudskej pamäte. Diplomová práca, Bratislava.
40. PARK, J. et al. (2008a). Web content summarization using social bookmarking service. *NII Technical Report*. National Institute of Informatics of Japan.
41. PARK, J. et al. (2008b). Web Content summarization using social bookmarks: A new approach for social summarization. In: *WIDM '08: Proc of the 10th ACM Workshop on Web Information and Data Management*, ACM Press, pp. 103-110.
42. RADEV, D. – HOVY, E. – MCKEOWN, K. (2002). Introduction to the special issue on summarization. In: *J. of Computational Linguistics – Summarization*, vol. 28, no. 4, MIT Press Cambridge, pp. 399-408.
43. RADEV, D. et al. (2004). MEAD - a platform for multidocument multilingual text summarization. In: *LREC '04: Proc. of the 4th Int. Conf. on Language Resources and Evaluation*, pp. 699-702.
44. SHIPMAN, F. et al. (2003). Identifying useful passages in documents based on annotation patterns. In: *Proc. of the European Conf. on Digital Libraries*, pp. 101-112.
45. STEINBERGER, J. – JEŽEK, K. (2005). Text summarization and singular value decomposition. In: *ADVIS '04: Proc. of Advances in Information Systems*, LNCS, vol. 3261, Springer, Berlin, pp. 245-254.
46. STEINBRGER, J. – JEŽEK, K. – SLOUP, M. (2008). Web topic summarization. In: *ELPUB '08: Proc. of the 12th Int. Conf. on Electronic Publishing*, pp. 322-334.
47. STEINBERGER, J. – JEŽEK, K. (2009). SUTLER: Update summarizer based on latent topics. In: *TAC '08: Proc. of the 1st Text Analysis Conf.*, National Institute of Standards and Technology.

48. SUN, J. et al. (2005). Web-page summarization using clickthrough data. In: *SIGIR '05: Proc. of 28th Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval*, ACM Press, pp. 194–201.
49. SWEENEY, S. – CRESTANI, F. – LOSADA, D.E. (2008). Show me more: Incremental length summarisation using novelty detection. In: *Information Processing and Management*, vol. 44, no. 2, Elsevier, pp. 663–686.
50. ŠIMKO, M. (2012a). Automated Acquisition of Domain Model for Adaptive Collaborative Web-based Learning. Kandidátska dizertačná práca, FIIT STU Bratislava, 137 p.
51. ŠIMKO, M. (2012b). Automated Acquisition of Domain Model for Adaptive Collaborative Web-based Learning. In: *Information Sciences and Technologies Bulletin of the ACM Slovakia* (to appear).
52. ŠIMKO, M. – BARLA, M. – BIELIKOVÁ M. (2010). ALEF: A framework for adaptive web-based learning 2.0. In: *KCKS 2010: IFIP Advances in Information and Communication Technology*, vol. 324, Springer, Berlin, pp. 367-378.
53. ŠIMKO, M. – BARLA, M. – MIHÁL, V. UNČÍK, M. – BIELIKOVÁ, M. (2011). Supporting Collaborative Web-based Education via Annotations. In: ED-MEDIA '11: *Proc. of World Conf. on Educational Multimedia, Hypermedia and Telecommunications*, T. Bastiaens, M. Ebner (Eds.), Association for the Advancement of Computing in Education (AACE), pp. 2576-2585.
54. TOMBROS, A. – SANDERSON, M. (1998). Advantages of query biased summaries in information retrieval. In: *SIGIR '98: Proc. of 21th Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval*, ACM Press, pp. 2–10.
55. VOZÁR, O. – BIELIKOVÁ, M. (2008). Adaptive Test Question Selection for Web-based Educational System. In: *SMAP '08: Proc. of the 3rd Int. Workshop on Semantic Media Adaptation and Personalization*, IEEE CS Press, pp. 164-169.
56. WITTE, R. – BERGLER, S. (2007). Next-generation summarization: Contrastive, focused and update summaries. In: *RANLP '07: Proc. of Int. Conf. on Recent Advantages in Natural Language Processing*.
57. UNČÍK, M. – BIELIKOVÁ, M. (2010). Annotating Educational Content by Questions Created by Learners. In: *SMAP '10: Proc. of the 5th Int. Workshop on Semantic Media Adaptation and Personalization*, IEEE Signal Processing Society, pp. 13-18.
58. YEH, J.Y. et al. (2005). Text summarization using a trainable summarizer and latent semantic analysis. In: *Information Processing and Management*, vol. 41, no. 1, Elsevier, pp. 75–95.
59. ZHANG, H. – MA, Z.C. – CAI, Q. (2003). A study for documents summarization based on personal annotation. *In: Proc. of the HLT-NAACL Workshop on Text summarization*, Association for Computational Linguistics, pp. 41–48.
60. ZHU, J. et al. (2009). Tag-oriented document summarization. In: *WWW '09: Proc. of the 18th Int. Conf. on World Wide Web*, ACM Press, pp. 1195-1196.

PRÍLOHY

Príloha A: Dokumentácia k overeniu

A.1 Dátový súbor

Pri overení sme použili vzdelávacie texty z kurzov Funkcionálne a logické programovanie a Princípy softvérového inžinierstva, ktoré sú k dispozícii vo výučbovom systéme ALEF. Spolu ide o 303 dokumentov vo formáte XML s presne danou štruktúrou (schémou), z ktorých sa generujú HTML dokumenty zobrazované používateľom.

Príklad dokumentu *Funkcionálne programovanie* vo formáte XML:

```
<?xml version="1.0" encoding="UTF-8"?>
<article xmlns="http://docbook.org/ns/docbook" role="explanation"
xml:id="flpbook-1.2.2" version="5">
  <title>Funkcionálne programovanie</title>
  <simplesect role="definition">
    <para>Funkcionálne programovanie rozširuje vlastnosti a výhody čis-
tých výrazov do celého programovacieho jazyka. Funkcionálne programovanie
sa odlišuje od procedurálneho programovania nezávislosťou od poradia.
Svet príkazov sa charakterizuje ich poradím. Na druhej strane svet výra-
zov zahŕňa opis hodnôt a vyhodnocovanie nezávisí od poradia.</para>
    <para>Vo funkcionálnom programovaní má program formu <empha-
sis>aplikatívnych výrazov</emphasis>. Takýto výraz je buď konštanta alebo
výraz vytvorený iba aplikáciami čistých funkcií na ich argumenty, ktoré
predstavujú tiež aplikatívne výrazy.</para>
...
    <alef:embed xmlns:alef="http://fiit.stuba.sk/ns/alef">
      <alef:embeditem type="question" count="3"/>
      <alef:embeditem type="exercise" count="2"/>
    </alef:embed>
  </simplesect>
</article>
```

Pred odoslaním dokumentov na sumarizáciu sme vykonali jednoduché predspracovanie:

- Odstránili sme nadpis (značka `title`), pretože názvy dokumentov sme už získavali z dátového modelu systému ALEF
- Odstránili sme popisy obrázkov (značka `figure`)
- Odstránili sme ukážky zdrojových kódov (značka `programlisting`)
- Zo zvyšných značiek sme extrahovali čistý text, pričom sme všetky biele znaky nahradili medzerou
- Ak chýbalo interpunkčné znamienko v zoznamoch (po odrážkach), doplnili sme ho, aby sa jednoduchšie určil začiatok a koniec vety
- Keďže použitá služba Bing Translator mala problémy pri niektorých skratkách, nahradili sme ich, t.j. zo slov ako „napr.“, „tzv.“, „a pod.“ sme odstránili končiacu bodku, aby ich nepovažovala za koniec vety (vo výsledných sumarizáciách sme ich potom spätne dopĺňali)

A.2 Konfigurácie sumarizátora

Uskutočnili sme dva experimenty, pričom sme pri každom porovnávali zvolenú dvojicu variantov:

- Sumarizáciu zohľadňujúcu relevantné doménové pojmy voči generickej
- Sumarizáciu zohľadňujúcu poznámky (zvýraznenia) voči generickej

Celkovo sme však pri overovaní použili päť konfigurácii (Tab. A-1) – pri experimente na predmete Princípy softvérové inžinierstva sme kvôli povahe textov zvolili kratšie sumarizácie (uvedená max. dĺžka je len orientačná, keďže sumarizácie sa netvorili po znakoch, ale po vetách).

Tab. A-1 Konfigurácie sumarizátora použité pri experimentoch.

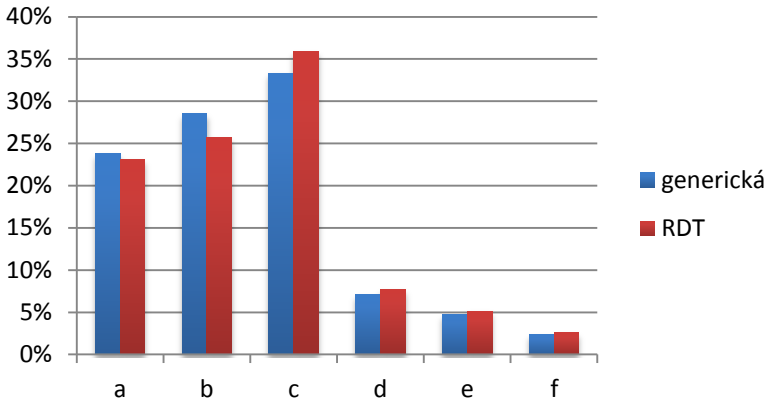
	Experiment (predmet)	**Hodnotič frekvencie**	**Hodnotič polohy**	**Hodnotič RDT**	**Hodnotič poznámok**	**Max. dĺžka (znaky)**
generická	1 (FLP)	0.9	0.1	0.0	0.0	1000
generická krátka	2 (PSI)	0.9	0.1	0.0	0.0	300
RDT	1 (FLP)	0.1	0.1	0.8	0.0	1000
poznámky	2 (FLP)	0.3	0.1	0.0	0.6	1000
poznámky krátka	2 (PSI)	0.3	0.1	0.0	0.6	300

A.3 Vyhodnotenie odpovedí na doplňujúce otázky

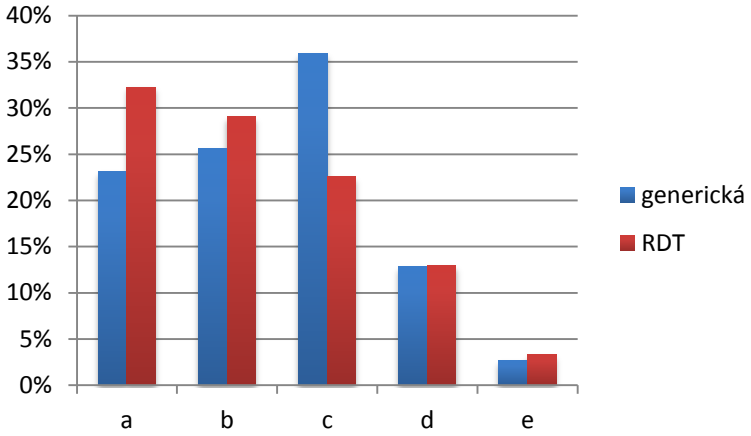
Pri prvom aj druhom experimente na predmete FLP sme použili doplňujúce otázky podľa dotazníka uvedeného v kap. 7.3. Pri prvom experimente sme získali 275 odpovedí, ktoré sme sa pokúsili vyhodnotiť. Tento počet je príliš malý vzhľadom na počet 79 vzdelávacích textov pre kurz LISP (ktorý je súčasťou FLP) v systéme ALEF na to, aby sme mohli vyhodnotiť získané odpovede po jednotlivých textoch. Pozreli sme sa aspoň na všeobecný pohľad na úrovni variantov (ktorý je však skresľujúci, lebo získané odpovede sa môžu týkať rôznej podmnožiny textov). Interpretácia uvádzaných grafov je preto problematická (a miestami aj protirečivá) a môžeme z nich odvodiť len všeobecné konštatovania k sumarizácii bez uvažovania variantov.

Na Obr. A-1 vidíme, že oba varianty sumarizácie podľa odpovedí študentov mali porovnateľnú úspešnosť pri výbere reprezentatívnych viet do sumarizácie – vety boli reprezentatívne čiastočne, t.j. niektoré boli a niektoré nie.

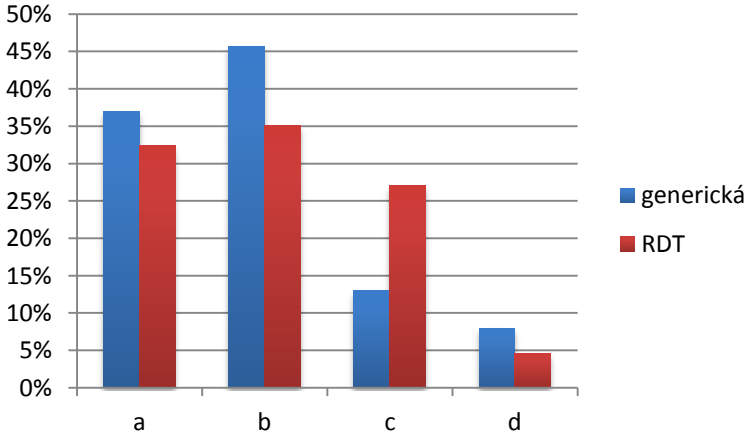
Pri druhej otázke je zaujímavé, že študenti ako odpoveď na otázku o vhodnosti sumarizácie na opakovanie zvolili možnosť A, t.j. „áno, určite“ vo vyše 30% prípadov pri variante zohľadňujúcom relevantné doménové pojmy (RDT) oproti niečo vyše 20% v prípade generického variantu (Obr. A-2). Avšak pri ďalšej otázke (Obr. A-3) je situácia opačná, keď na otázku o tom, či by im sumarizácia pomohla rozhodnúť o relevancii dokumentu odpovedali „áno, sumarizácia dokonca zhrnula text tak, že by som ho už ani nemusel/a celý čítať a väčšinu podstatného by som sa dozvedel/a“ tesne v prospech generického variantu. Keďže si tieto odpovede navzájom protirečia, naznačuje to, že ide o odpovede na rôzne podmnožiny otázok.



Obr. A-1 Reprezentatívnosť vybraných viet.

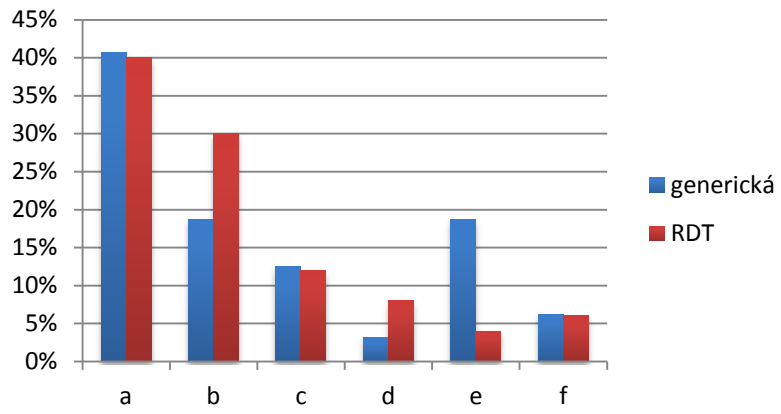


Obr. A-2 Vhodnosť sumarizácie pre opakovanie.

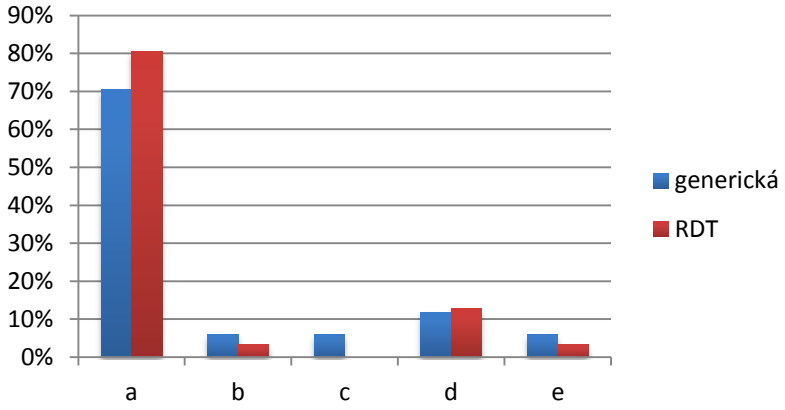


Obr. A-3 Určenie relevantnosti dokumentu na základe sumarizácie.

Z odpovedí na štvrtú otázku vidíme (Obr. A-4), že aj napriek niektorým problémom s rozdelením textu na vety stratou kontextu, väčšinou sú sumarizácie čitateľné a pochopiteľné. A napokon z odpovedí na otázku číslo päť (Obr. A-5) jednoznačne vyplýva, že dĺžka bola nastavená správne, t.j. zhruba tretina dĺžky textu je pre používateľov akceptovateľná.



Obr. A-4 Čitateľnosť a zrozumiteľnosť sumarizácie.



Obr. A-5 Vhodnosť dĺžky sumarizácie.

Keďže pri druhom experimente sme získali ešte menší počet odpovedí, ani sme ich ďalej nevyhodnocovali, pretože by mali ešte nižšiu výpovednú hodnotu.

A.4 Ukážka výslednej sumarizácie

Na dokumente s názvom *Aplikatívne programovanie* ilustrujeme schopnosti jednotlivých variantov sumarizácie zhrnúť dôležité informácie z dokumentu. Uvádzame sumarizácie vytvorené generickým variantom, variantom zohľadňujúcim relevantné doménové pojmy a napokon variantom zohľadňujúcim poznámky (zvýraznenia) používateľov.

Vidíme, že aj generická sumarizácia dokáže celkom dobre vystihnúť podstatu dokumentu. Pri zohľadnení RDT sa sumarizácia viac zameria na koncepty danej domény, preto sa do súhrnu dostala aj veta, ktorá spomína, aké všetky pojmy a techniky sú preberané v učebnici. Napokon pri sumarizácii zohľadňujúcej zvýraznenia vidíme, že vybrala väčšinu takých viet, ktoré si používateľ zvýraznil, čím by mala byť z jeho pohľadu lepšia, lebo zachytáva to, čo považuje za dôležité.

Pôvodný dokument spolu so zvýrazneniami (podčiarknuté vety):

Aplikatívne programovanie

Aplikatívny program opisuje výpočet výrazom.

Posun smerom k deklaratívnemu prístupu k programovaniu možno sledovať pri aplikatívnom programovaní. Pri aplikatívnom programovaní sa želaný výpočet opíše výrazom. Jeho vyhodnotením sa získa požadovaný výsledok. Aplikatívne programovanie svojou vysokou úrovňou abstrahovania od toho, ako sa výpočet vykonáva chápe programovanie ako deklarovanie toho, čo sa počíta, čiže o aký výpočet ide.

Na ilustráciu uvažujme tento výraz

```
(x + y) * (x - y)
```

Výraz opisuje aplikáciu niekoľkých funkcií (sčítanie, násobenie a odčítanie). Vo výraze sa neurčujú žiadne podrobnosti výpočtu ako napr. spôsob a miesto uloženia medzivýsledkov. Ďalej možno pozorovať viaceré alternatívy čo sa týka poradia vyhodnocovania jednotlivých podvýrazov a možnosť ich paralelného vyhodnotenia.

Ďalším dôležitým faktom je, že x a y vystupujú vo výraze ako číselné hodnoty a nie ako pamäťové bunky.

Na druhej strane v priraďovacom príkaze, napr. x = x + 1 prvý výskyt x označuje pamäťovú bunku a druhý výskyt x označuje číslo (obsah tejto pamäťovej bunky). Vo výrazoch sa teda kladie dôraz na hodnoty samotné a nie na organizáciu ich uloženia.

Ak sa pri aplikatívnom programovaní ako základný výrazový prostriedok použijú funkcie, vrátane funkcií vyšších rádov (t.j. takých, ktoré operujú nad inými funkciami), hovoríme o *funkcionálnom programovaní*. Ak sú základným výrazovým prostriedkom relácie, opísané pomocou logických predikátov, hovoríme o logickom programovaní.

Vo funkcionálnom programovaní sa výpočet opisuje výrazom.

Vo funkcionálnom programovaní sa program chápe ako množina funkcií. Na rozdiel od procedurálneho programovania, ktoré vychádza z modelu výpočtov založeného na von Neumannovej architektúre počítača, opiera sa funkcionálne programovanie o tzv. lambda počet ako jednoduchý model výpočtov. Základné pojmy funkcionálneho programovania, ako funkcia, výraz, zloženie výrazov, rekurzívna definícia funkcie sa podrobne vysvetľujú v prvej časti tejto učebnice. Takisto sa tu rozoberajú základné funkcionálne programovacie techniky, ako jednotlivé vzory rekurzívnych definícií funkcie, programovanie filtrov, generátorov, programovanie pomocou funkcií vyšších rádov (funkcionálov).

V logickom programovaní je výpočet dokazovaním dopytu.

Pri logickom programovaní sa ako programovací jazyk využíva predikátová logika. Základom je interpretácia implikácií ako deklarácií procedúr. Ide o tzv. procedurálnu interpretáciu predikátového počtu prvého rádu. Vytvoriť logický program znamená sformulovať sústavu axióm opisujúcich triedu riešených úloh. Špeciálnu úlohu, ktorú treba vyriešiť, treba sformulovať ako cieľový príkaz. Je to formula predikátového počtu, ktorá sa zapíše v špeciálnom tvare. Výpočet je potom dôkaz, že cieľový príkaz (dopyt) je logický dôsledok množiny axióm, tvoriacej program.

Základné pojmy logického programovania ako klauzula, predikát, term, odvodenie odpovede na zadaný dopyt spolu s programovacími technikami logického programovania sa podrobne vysvetľujú v druhej časti tejto učebnice.

Generická sumarizácia:

Aplikatívny program opisuje výpočet výrazom. Posun smerom k deklaratívnemu prístupu k programovaniu možno sledovať pri aplikatívnom programovaní. Jeho vyhodnotením sa získa požadovaný výsledok. Vo výraze sa neurčujú žiadne podrobnosti výpočtu ako napr. spôsob a miesto uloženia medzivýsledkov. Vo výrazoch sa teda kladie dôraz na hodnoty samotné a nie na organizáciu ich uloženia. Vo funkcionálnom programovaní sa program chápe ako množina funkcií. V logickom programovaní je výpočet dokazovaním dopytu. Pri logickom programovaní sa ako programovací jazyk využíva predikátová logika. Základom je interpretácia implikácií ako deklarácií procedúr. Špeciálnu úlohu, ktorú treba vyriešiť, treba sformulovať ako cieľový príkaz. Je to formula predikátového počtu, ktorá sa zapíše v špeciálnom tvare. Základné pojmy logického programovania ako klauzula, predikát, term, odvodenie odpovede na zadaný dopyt spolu s programovacími technikami logického programovania sa podrobne vysvetľujú v druhej časti tejto učebnice.

Sumarizácia zohľadňujúca relevantné doménové pojmy:

Pri aplikatívnom programovaní sa želaný výpočet opíše výrazom. Ak sú základným výrazovým prostriedkom relácie, opísané pomocou logických predikátov, hovoríme o logickom programovaní. Vo funkcionálnom programovaní sa výpočet opisuje výrazom. Vo funkcionálnom programovaní sa program chápe ako množina funkcií. Základné pojmy funkcionálneho programovania, ako funkcia, výraz, zloženie výrazov, rekurzívna definícia funkcie sa podrobne vysvetľujú v prvej časti tejto učebnice. Takisto sa tu rozoberajú základné funkcionálne programovacie techniky, ako jednotlivé vzory rekurzívnych definícií funkcie, programovanie filtrov, generátorov, programovanie pomocou funkcií vyšších rádov (funkcionálov). V logickom programovaní je výpočet dokazovaním dopytu. Pri logickom programovaní sa ako programovací jazyk využíva predikátová logika. Základné pojmy logického programovania ako klauzula, predikát, term, odvodenie odpovede na zadaný dopyt spolu s programovacími technikami logického programovania sa podrobne vysvetľujú v druhej časti tejto učebnice.

Sumarizácia zohľadňujúca poznámky (zvýraznenia):

Aplikatívny program opisuje výpočet výrazom. Ak sa pri aplikatívnom programovaní ako základný výrazový prostriedok použijú funkcie, vrátane funkcií vyšších rádov (t.j. takých, ktoré operujú nad inými funkciami), hovoríme o funkcionálnom programovaní. Ak sú základným výrazovým prostriedkom relácie, opísané pomocou logických predikátov, hovoríme o logickom programovaní. Vo funkcionálnom programovaní sa výpočet opisuje výrazom. Vo funkcionálnom programovaní sa program chápe ako množina funkcií. V logickom programovaní je výpočet dokazovaním dopytu. Pri logickom programovaní sa ako programovací jazyk využíva predikátová logika. Vytvoriť logický program znamená sformulovať sústavu axióm opisujúcich triedu riešených úloh. Výpočet je potom dôkaz, že cieľový príkaz (dopyt) je logický dôsledok množiny axióm, tvoriacej program. Základné pojmy logického programovania ako klauzula, predikát, term, odvodenie odpovede na zadaný dopyt spolu s programovacími technikami logického programovania sa podrobne vysvetľujú v druhej časti tejto učebnice.

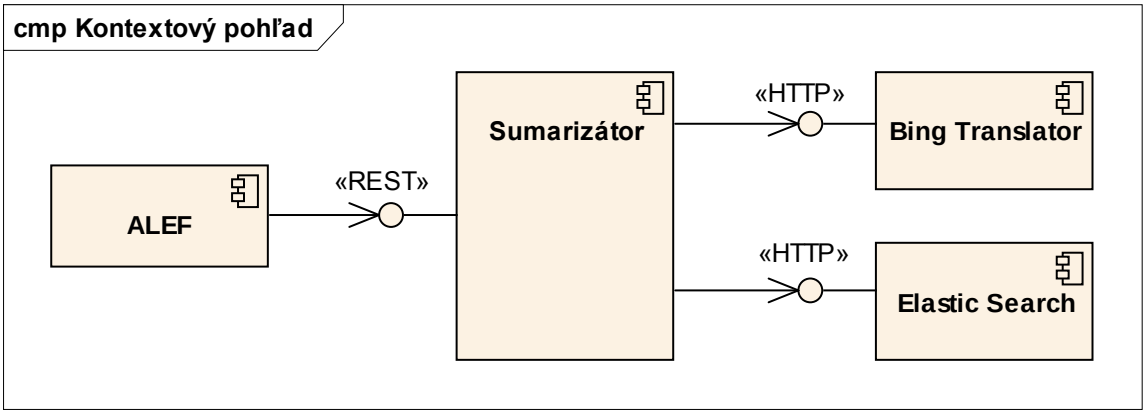
Príloha B: Technická dokumentácia

B.1 Analýza a návrh

Pri analýze a návrhu sme vychádzali zo základnej požiadavky poskytnúť možnosť sumarizácie textových dokumentov, pričom by táto sumarizácia nebola závislá od domény ani jazyka sumarizovaného dokumentu. Preto sme zvolili realizáciu sumarizátora v podobe webovej REST služby s definovaným protokolom v jazyku JSON.

Overenie metódy sumarizácie sme realizovali vo výučbovom systéme ALEF, ktorý poskytuje jednoduchú možnosť rozšírenia funkcionality pomocou pridania komponentov v rámci existujúceho komponentového rámca. Navrhli sme preto rozšírenie systému ALEF o sumarizačný komponent, ktorý využíva služby sumarizátora na tvorbu súhrnov vzdelávacích textov. Sumarizáciu sme následne využili pri opakovaní – navrhli sme rozšírenie systému ALEF o ďalší komponent, ktorý umožňuje používateľom vyžiadať si odporúčania dokumentov na opakovanie.

Základný kontextový pohľad na systém je zachytený na Obr. B-1. Vidíme, že sumarizátor využíva ďalšie služby ako *Bing Translator* a *Elastic Search*.



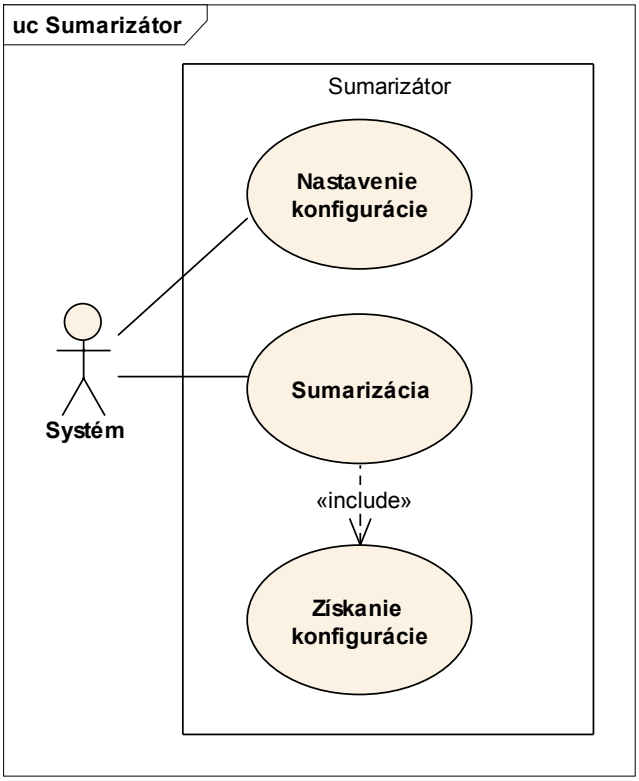
Obr. B-1 Kontextový pohľad na systém - sumarizátor a ALEF.

B.1.1 Prípady použitia

Identifikovali sme tieto služby sumarizátora (Obr. B-2):

- *Nastavenie konfigurácie* – umožňuje používateľovi nastaviť zvolenú konfiguráciu sumarizátora pomocou navrhnutého komunikačného protokolu opísaného nižšie; to zahŕňa nastavenie stratégie extrakcie kľúčových slov, viet a prekladu, voľbu hodnotičov a ich koeficientov a želaný výstup sumarizátora
- *Získanie konfigurácie* – umožňuje získať aktuálnu konfiguráciu sumarizátora na základe zaslaného identifikátora konfigurácie
- *Sumarizácia* – predstavuje najdôležitejší prípad použitia; umožňuje používateľovi získať sumarizáciu zaslaného dokumentu podľa danej konfigurácie

Používateľom systému rozumieme klientský systém, ktorý chce využívať služby sumarizátora (napr. výučbový systém ALEF).



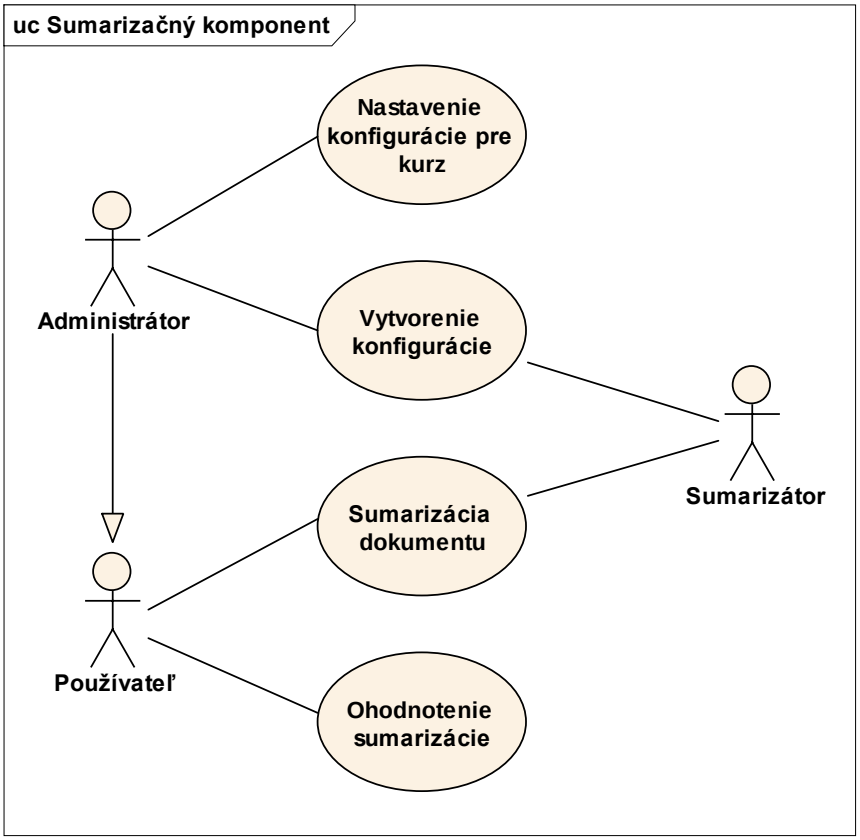
Obr. B-2 Sumarizátor: diagram prípadov použitia.

Sumarizačný komponent v systéme ALEF poskytuje tieto služby (Obr. B-3):

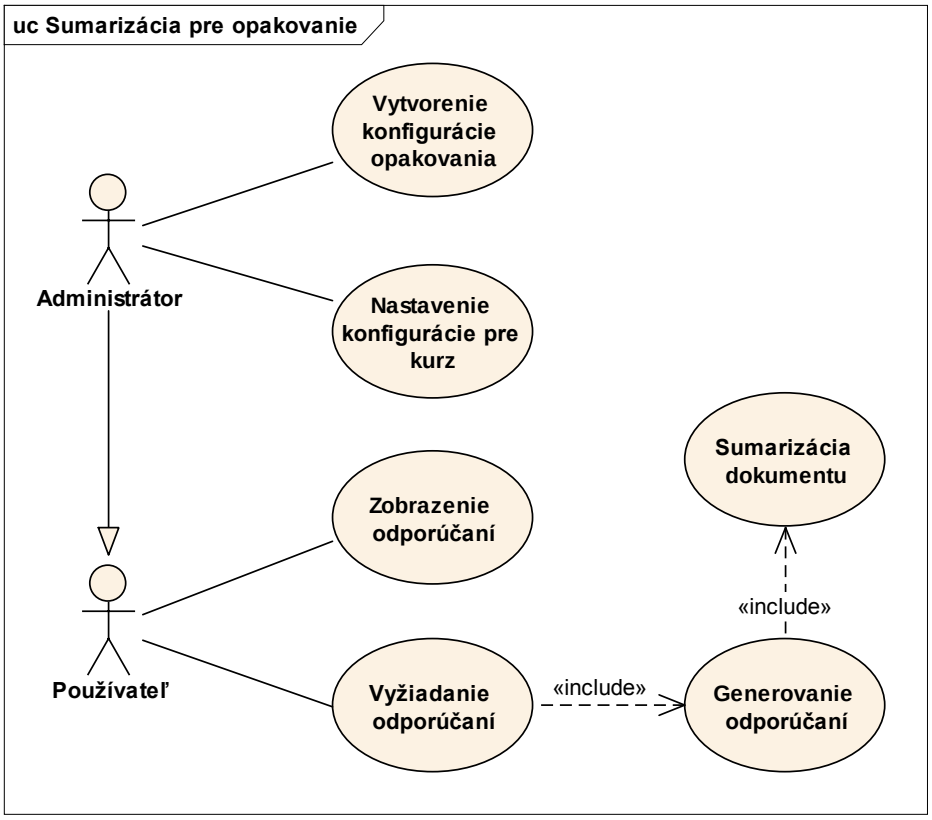
- *Nastavenie konfigurácie pre kurz* – administrátor môže nastaviť, aká konfigurácia sumarizátora sa použije pri sumarizácii dokumentov v konkrétnom kurze v systéme ALEF
- *Vytvorenie konfigurácie* – administrátor pomocou rozhrania nastaví váhy jednotlivých hodnotičov sumarizátora a ďalšie parametre, ako napr. dĺžku sumarizácie; zvolená konfigurácia sa zašle ako požiadavka na sumarizátor
- *Sumarizácia dokumentu* – zosumarizuje sa výučbový dokument podľa zvolenej konfigurácie pre daný kurz
- *Ohodnotenie sumarizácie* – používateľ môže ohodnotiť kvalitu sumarizácie

Pri sumarizácii pre opakovanie sme identifikovali tieto prípady použitia (Obr. B-4):

- *Vytvorenie konfigurácie opakovania* – administrátor môže nastaviť parametre metódy výberu dokumentov na opakovanie (váhy jednotlivých hodnotičov a počet odporúčaní, ktoré sa budú zobrazovať používateľovi)
- *Nastavenie konfigurácie pre kurz* – administrátor môže nastaviť, aká konfigurácia opakovania sa použije v konkrétnom kurze v systéme ALEF
- *Zobrazenie odporúčaní* – používateľovi sa po vyžiadaní zobrazí zoznam vygenerovaných odporúčaní – dokumentov vybraných na opakovanie
- *Vyžiadanie odporúčaní* – v prípade, že odporúčania pre používateľa nie sú k dispozícii, môže si vyžiadať ich vygenerovanie
- *Generovanie odporúčaní* – na základe zvolenej konfigurácie sa ohodnotia dokumenty a vyberú sa tie, ktoré sú vhodné pre daného používateľa na opakovanie; zahŕňa sumarizáciu vybraných dokumentov



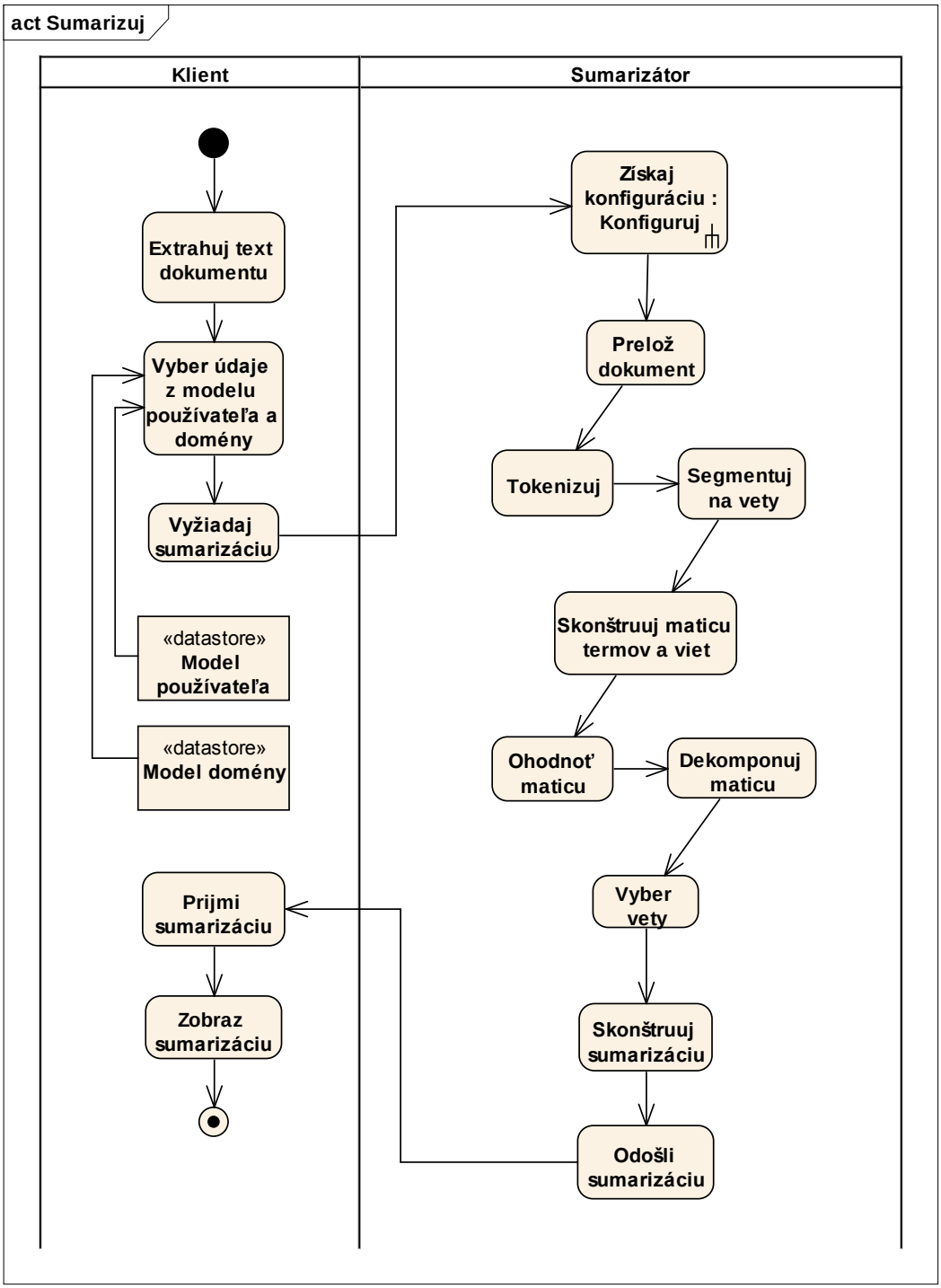
Obr. B-3 Sumarizačný komponent: diagram prípadov použitia.



Obr. B-4 Sumarizácia pre opakovanie: diagram prípadov použitia.

B.1.2 Proces sumarizácie

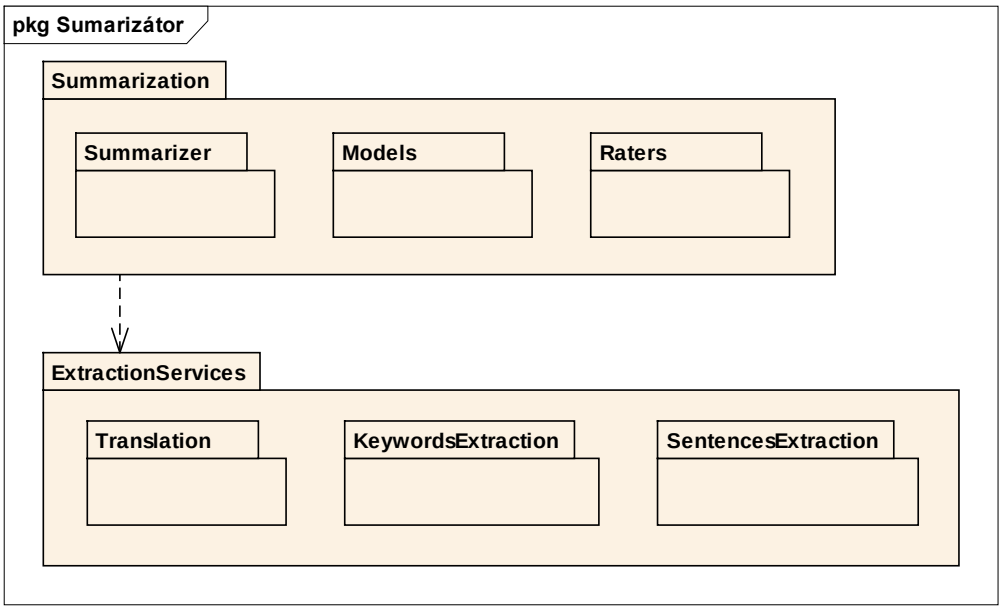
Na Obr. B-5 je znázornený proces sumarizácie. Vidíme, že časť logiky predspracovania dokumentu ako aj následné zobrazenie získaných sumarizácií používateľovi je na strane klienta; sumarizátor zabezpečuje len predspracovanie nevyhnutné pre získanie sumarizácie. Zobrazené sú tiež jednotlivé kroky metódy latentnej sémantickej analýzy (konštrukcia matice termov a viet, jej dekompozícia, výber viet), ktorá je rozšírená o nami navrhnutú metódu prispôsobovania v kroku ohodnotenia matice.



Obr. B-5 Proces sumarizácie.

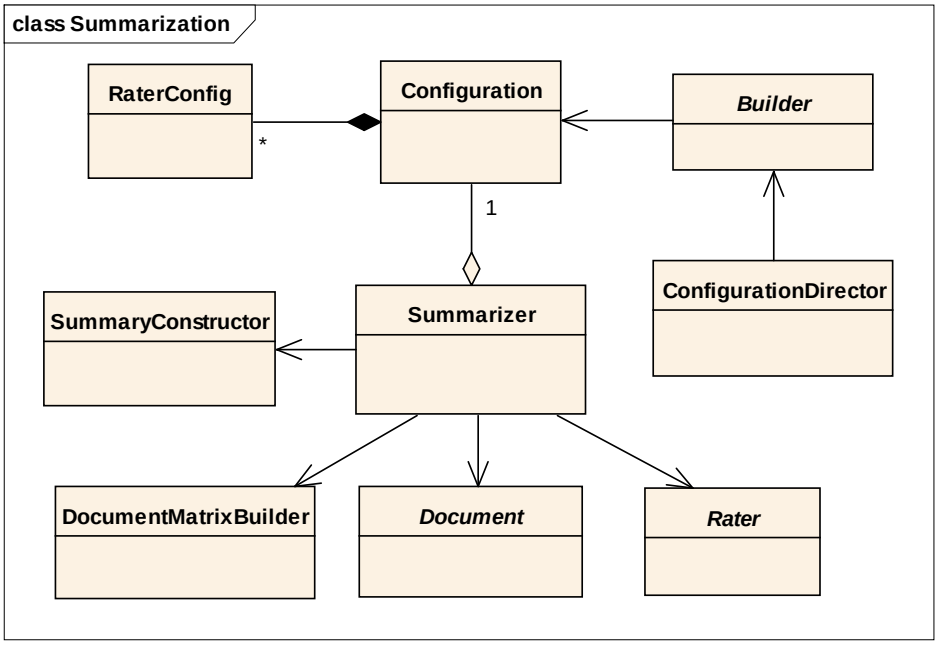
B.1.3 Moduly sumarizátora

Diagram na Obr. B-6 zachytáva moduly (balíky) sumarizátora. Modul *Summarization* združuje moduly zodpovedné za sumarizáciu dokumentov; zahŕňa modul *Summarizer*, ktorý predstavuje vlastný sumarizátor spolu s konfiguráciou, modul *Models*, ktorý obsahuje modely implementujúce hlavnú logiku a modul *Raters*, ktorý obsahuje hodnotiče využívané pri konštrukcii personalizovanej matice termov a viet. Modul *ExtractionServices* združuje triedy na extrakciu kľúčových slov z dokumentu, jeho segmentáciu na vety a strojový preklad.



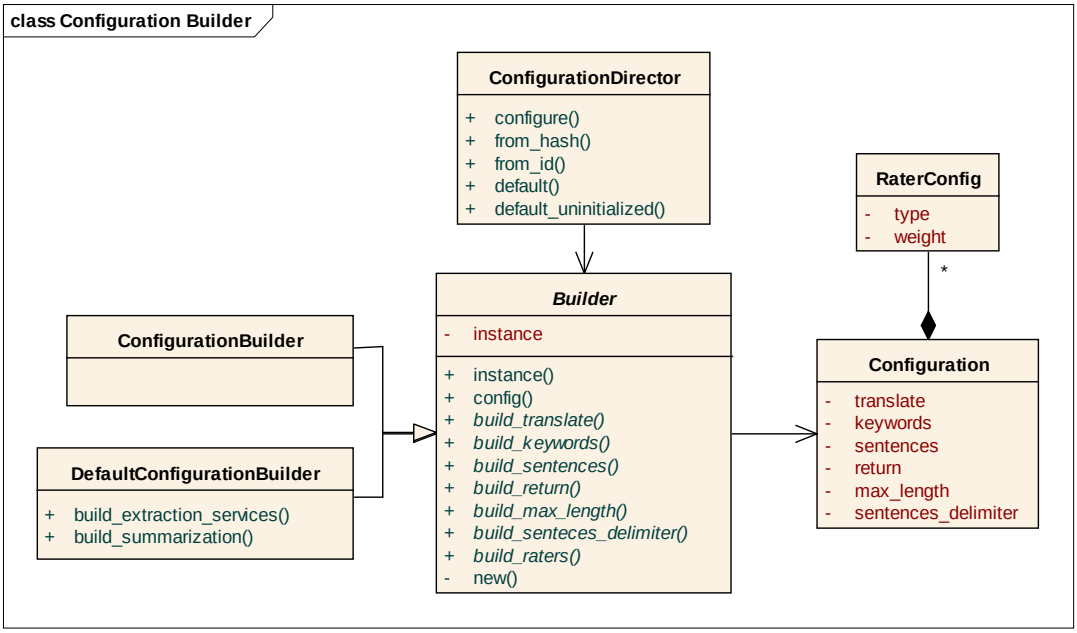
Obr. B-6 Balíky sumarizátora.

Diagram tried na Obr. B-7 predstavuje konceptuálny pohľad – najdôležitejšie triedy balíka *Summarization*.

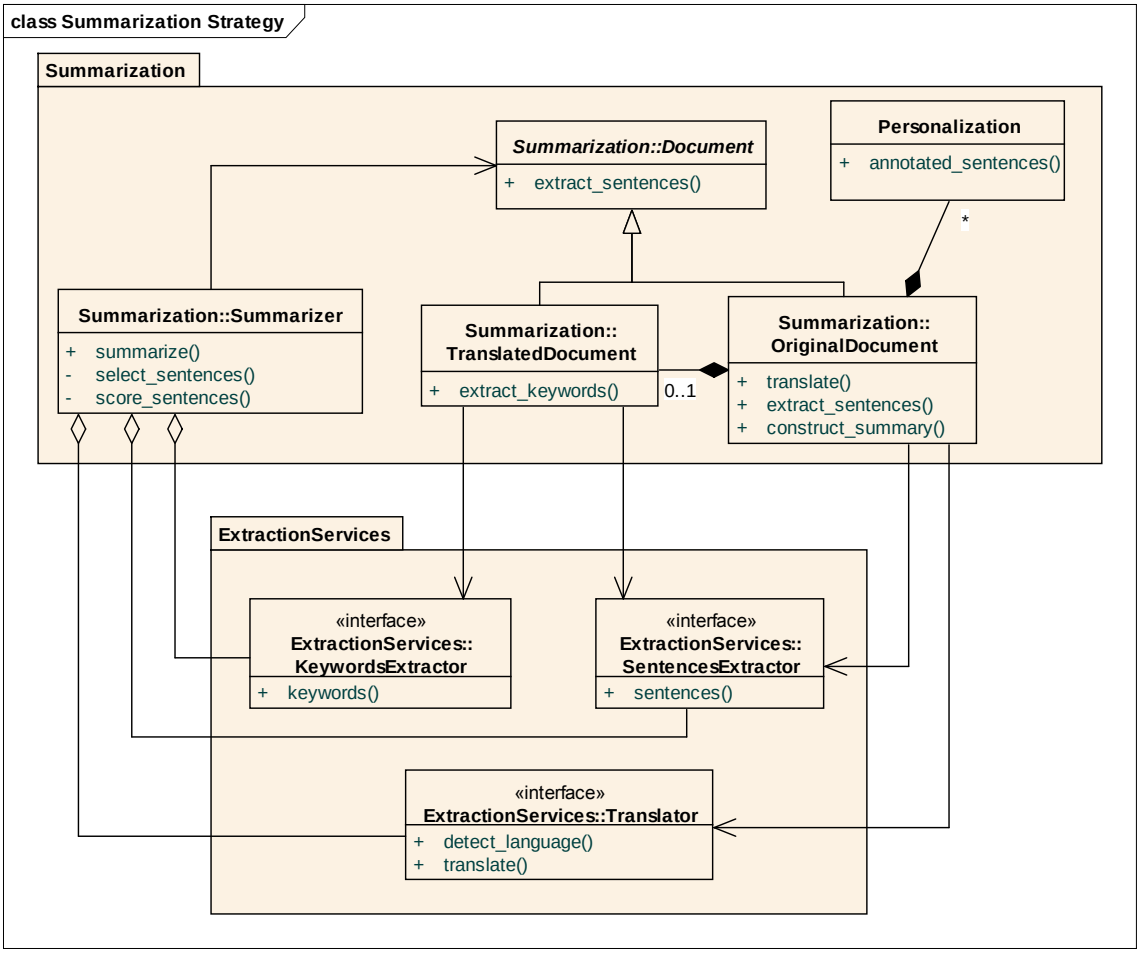


Obr. B-7 Diagram najdôležitejších tried balíka Summarization.

Na Obr. B-8 sú zachytené triedy na konfiguráciu sumarizátora (návrhový vzor *Builder*), Obr. B-9 znázorňuje výber stratégie extrakcie kľúčových slov a viet dokumentu a jeho strojového prekladu (návrhový vzor *Strategy*).

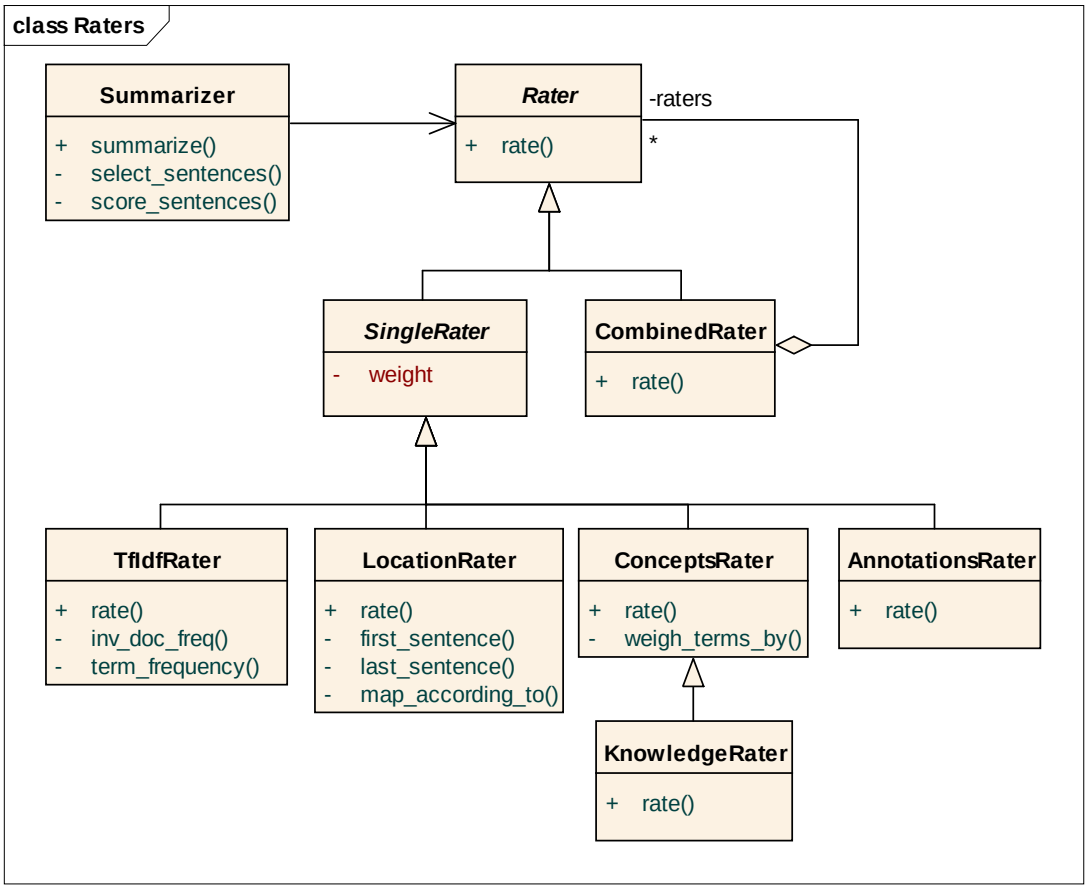


Obr. B-8 Triedy pre nastavenie konfigurácie.



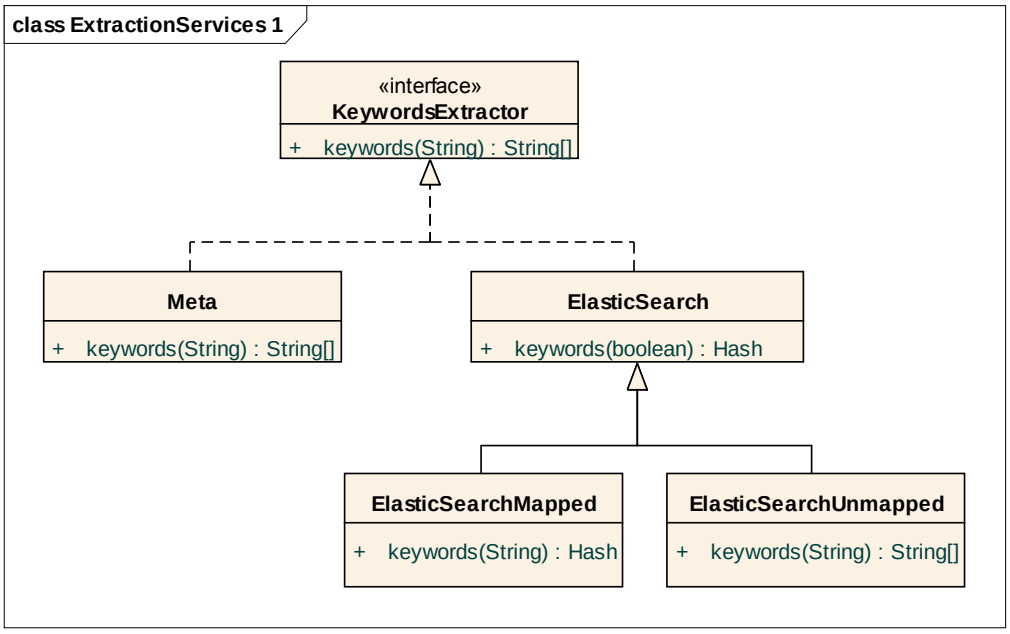
Obr. B-9 Výber stratégie extrakcie.

Pri kombinácii hodnotičov je použitý vzor *Composite* (Obr. B-10).

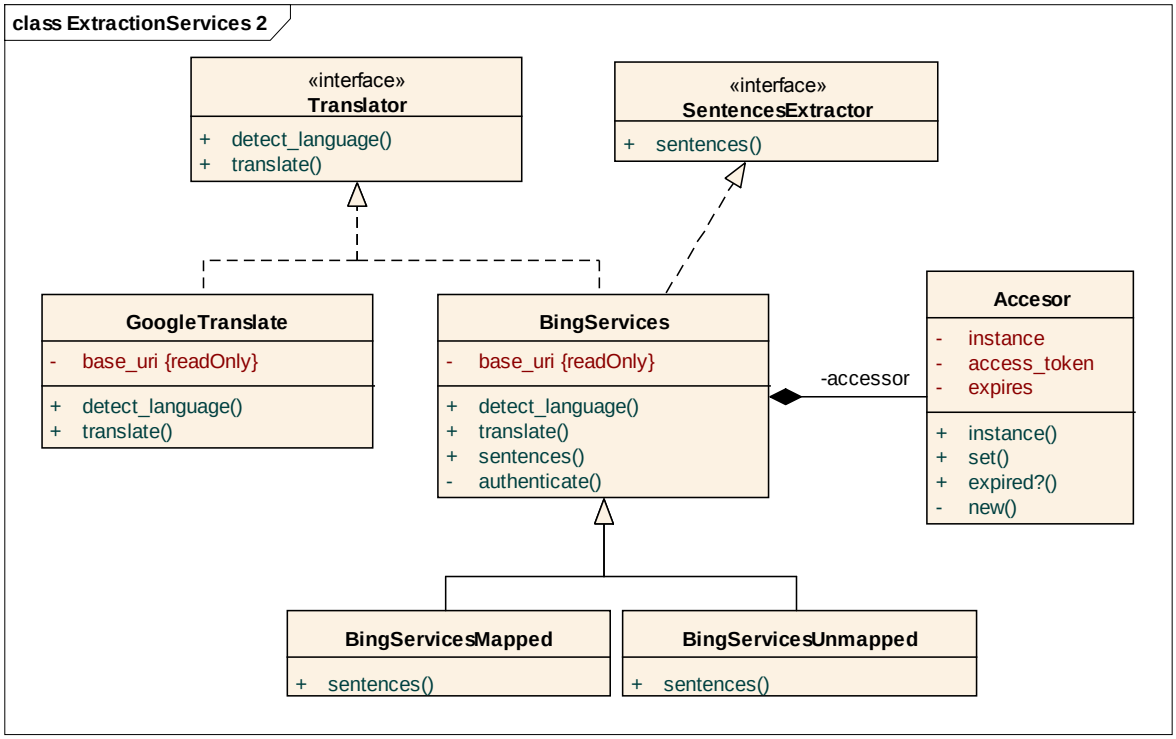


Obr. B-10 Hodnotiče a ich kombinácia.

Diagramy tried na Obr. B-11 a Obr. B-12 zobrazujú triedy balíka *ExtractionServices*, ktoré zabezpečujú extrakciu kľúčových slov z dokumentu, jeho preklad a segmentáciu na vety.



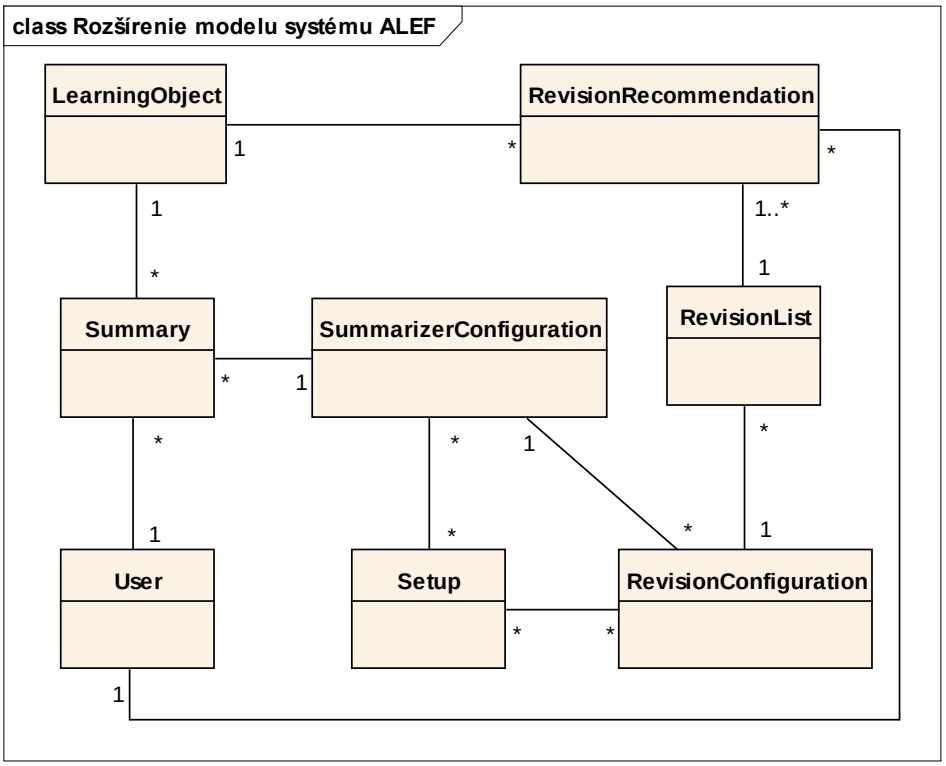
Obr. B-11 Triedy balíka ExtractionServices.



Obr. B-12 Triedy balíka ExtractionServices, pokračovanie.

B.1.4 Rozšírenie systému ALEF

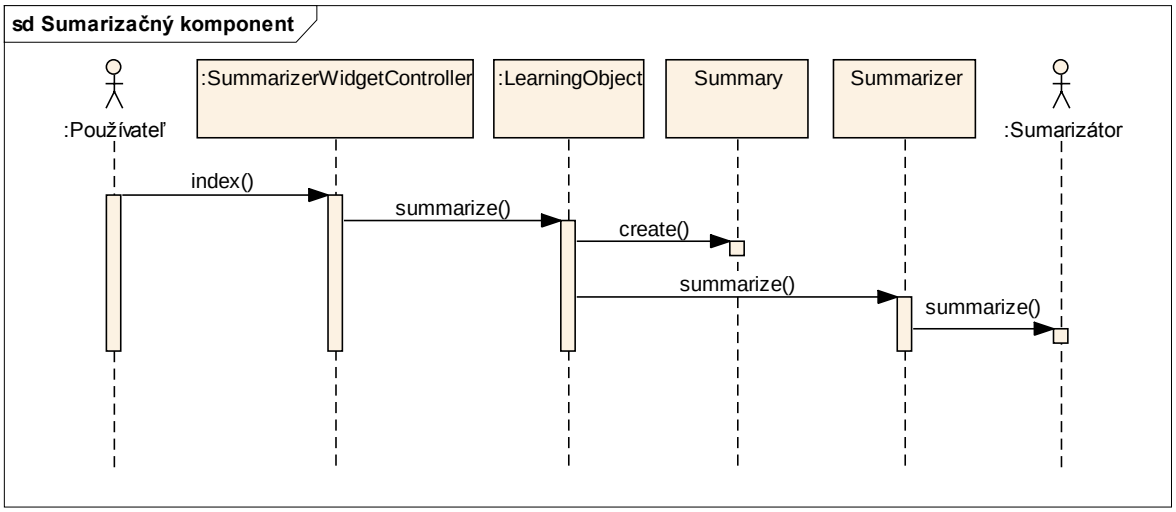
Rozšírili sme existujúci model systému ALEF (Obr. B-13) pridaním sumarizácií (entita *Summary*) a odporúčaní na opakovanie (entita *RevisionRecommendation*).



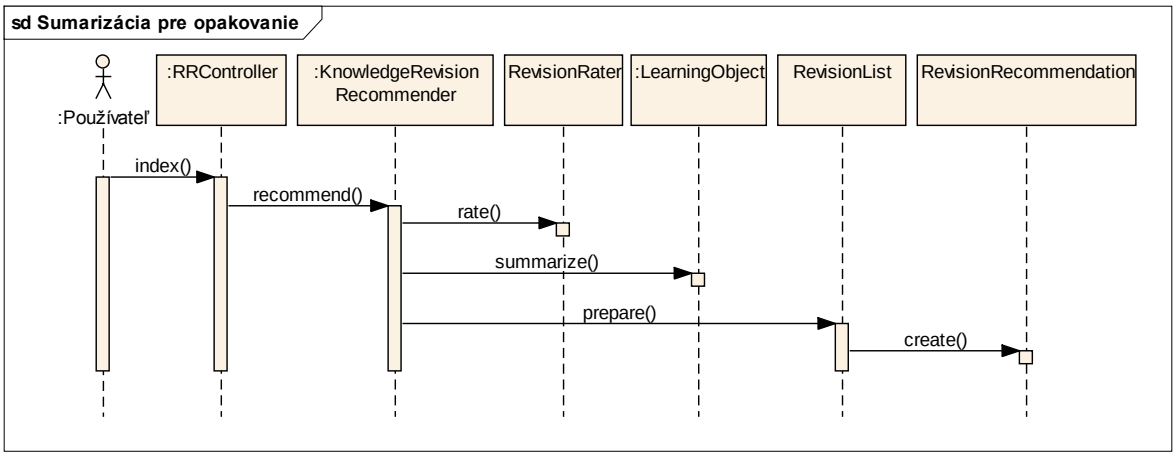
Obr. B-13 Rozšírenie modelu systému ALEF.

Okrem toho je možné vytvárať rôzne konfigurácie – nastavenia parametrov príslušných metód (entity *SummarizerConfiguration* a *RevisionRecommendation*) a pridávať ich ku kurzom (entita *Setup*) v systéme. Sumarizácie sú identifikované kombináciou používateľa, vzdelávacieho objektu a konfiguráciou sumarizátora. Odporúčania na opakovanie sú združované do zoznamov odporúčaní (*RevisionList*), pričom sa eviduje, akou konfiguráciou bol daný zoznam vygenerovaný.

Sekvenčný diagram na Obr. B-14 znázorňuje interakciu objektov pri sumarizácii pomocou sumarizačného komponentu, diagram na Obr. B-15 zas interakciu pri sumarizácii pre opakovanie.



Obr. B-14 Sumarizačný komponent: interakcia.



Obr. B-15 Sumarizácia pre opakovanie: interakcia.

B.2 Implementácia

Sumarizátor ako aj rozšírenia systému ALEF sme implementovali v jazyku Ruby. Uvádzame tu ukážky zdrojových kódov vybraných metód.

Najdôležitejšou metódou sumarizátora je metóda *summarize* triedy *Summarizer*, ktorá zabezpečuje tvorbu súhrnov zaslaných dokumentov na základe zvolenej konfigurácie:

```
def summarize(doc, personalization)
  doc.detect_language(@translator)
  doc.extract_sentences(@sentences_extractor)
  translated_doc = doc.translate(@translator)

  sentences = translated_doc.extract_sentences(@sentences_extractor)
  keywords = translated_doc.extract_keywords(@keywords_extractor)

  rater = CombinedRater.new(translated_doc, @config.rater_configs,
                            personalization)

  terms_sentences = rater.rate(DocumentMatrixBuilder.
                                  terms_by_sentences(keywords, sentences))
  u, s, vt = terms_sentences.
              singular_value_decomposition(:diagonalize => false)
  s = s.to_a.flatten

  doc.set_summary(select_sentences(doc.sentences, vt, s,
                                   doc.content.length / 3))
  doc.construct_summary(@output_strategy)
end
```

Prepojenie systému ALEF so sumarizátorom zabezpečuje proxy trieda *Summarizer* prostredníctvom metódy *summarize*:

```
def summarize
  request = {
      document: document,
      configuration: @config.nil? ? 'default' : @config.configuration_id
  }

  response = self.class.post(
      '/summarize',
      body: request.to_json,
      headers: {
          'Content-Type' => 'application/json; charset=utf-8',
          'Accept-Charset' => 'utf-8'
      }
  )

  raise SummarizationError.new(response.body) unless response.code == 200

  summary = response['summary']
  personalization = response['personalization_id']

  [summary, personalization]
end
```

Odporúčanie dokumentov na opakovanie je realizované prostredníctvo metódy *prepare_recommendations* triedy *KnowledgeRevisionRecommender*:

```
def self.prepare_recommendations(user, setup, config)
  los = filter_los(setup.domain_model_source_ids, user)
  rater = CombinedRevisionRater.new(user, config, setup)

  recommended los, scores = score los(los, rater, config.list length)
  summaries = summarize_los(recommended_los, user, setup,
                            config.summarizer_configuration)
```

```
  RevisionList.transaction do
    revision_list = RevisionList.create(user_id: user.id,
                                        setup_id: setup.id,
                                        revision_configuration_id:
                                                              config.id)
    revision_list.prepare(recommended_los, scores, summaries)
  end
end
```

B.2.1 Komunikačný protokol

Pre komunikáciu so sumarizátorom sme navrhli a implementovali komunikačný protokol vo formáte JSON. Protokol využívame pri:

- Konfigurácii sumarizátora klientom
- Zaslaní dokumentu na sumarizáciu

Používateľ môže sumarizátor konfigurovať zaslaním POST requestu, v ktorého tele je nasledovný JSON konfiguračný reťazec (znak „|“ oddeľuje jednotlivé možnosti nastavenia; používateľ zvolí práve jednu z nich):

```
{
  'configuration' : {
    'extraction_services' : {
      'translate' : 'bing' | 'google' | 'default',
      'keywords'  : 'elastic' | 'metal' | 'default',
      'sentences' : 'bing' | 'default'
    } | 'default',
    'summarization' : {
      'raters'              : [
        { 'type': 'tf_idf', 'weight': float },
        { 'type': 'location', 'weight': float },
        { 'type': 'concepts', 'weight': float },
        { 'type': 'knowledge', 'weight': float },
        { 'type': 'annotations', 'weight': float }
      ] | 'default',
      'return'              : 'text' | 'array' | 'default',
      'max_length'          : int,
      'sentences_delimiter' : string
  }
}
```

Odpoveďou na takúto požiadavku je opäť reťazec v JSON formáte, ktorý obsahuje identifikátor konfigurácie („ID“ je nahradené skutočným identifikátorom):

```
{ 'configuration_id' : 'ID' }
```

Získaný identifikátor môže používateľ následne využiť pri zaslaní požiadavky na sumarizáciu, kde ho uvedie ako želanú konfiguráciu (ak neuvedie žiadny identifikátor konfigurácie, automaticky sa zvolí predvolená konfigurácia):

```
{
  'document' : {
    'doc_id'   : 'id',
    'title'    : 'Názov',
    'content'  : 'Obsah dokumentu…',
    'concepts' : [
      ['concpet1', w1], ['concept2', w2], … ['conceptn', wn]
    ],
    'annotations'     : [
      [start1, len1], [start2, len2], … [startn, lenn]
    ],
    'personalization' : {
      'id'          : 'ID',
      'knowledge'   : [
        ['concpet1', k1], ['concept2', k2], … ['conceptn', kn]
      ],
      'annotations' : [
        [start1, len1], [start2, len2], … [startn, lenn]
      ]
    },
  },
  'configuration' : 'ID' | 'default'
}
```

Odpoveďou na takúto požiadavku je opäť reťazec v JSON formáte, ktorý obsahuje výslednú sumarizáciu, identifikátor sumarizovaného dokumentu a identifikátor personalizácie:

```
{
  'doc_id'             : 'id',
  'personalization_id' : 'p_id'
  'summary'            : 'Sumarizácia dokumentu...'
}
```

B.2.2 Použité technológie

Pri implementácii sme použili tieto technológie:

- *Ruby* – dynamický objektovo-orientovaný programovací jazyk s funkcionálnymi a aspektovými prvkami
- *Sinatra* – minimalistický webový aplikačný rámec, ktorý sme využili pri implementácii sumarizátora ako REST webovej služby

- *Ruby on Rails* – webový aplikačný rámec, ktorý je využívaný v rámci systému ALEF
- *ElasticSearch* – vyhľadávač postavený nad Lucene s REST rozhraním, jednoduchou inštaláciou a konfiguráciou; využívame ho pri extrakcii kľúčových slov
- *MongoDB* – dokumentová NoSQL databáza, ktorú využívame na ukladanie medzivýsledkov sumarizácie dokumentov a tiež konfigurácií sumarizátora
- *Linalg* – knižnica pre lineárnu algebru a prácu s maticami, ktorá je postavená nad knižnicou LAPACK (implementácia v jazyku FORTRAN); využívame predovšetkým na singulárnu dekompozíciu matíc

Príloha C: Inštalačná príručka

Opísaný je postup inštalácie sumarizátora a výučbového systému ALEF. Ide o dva nezávislé systémy, t.j. sumarizátor možno používať samostatne bez nutnosti inštalácie systému ALEF pomocou REST rozhrania opísaného v časti B.2.1 technickej dokumentácie. Systém ALEF využíva sumarizátor na tvorbu súhrnov vzdelávacích textov; integráciu zabezpečuje sumarizačný komponent – v prípade vypnutia tohto komponentu ALEF nevyžaduje, aby bol sumarizátor nainštalovaný.

C.1 Inštalácia sumarizátora

Nasledovný postup opisuje kroky potrebné pre inštaláciu (nasadenie) sumarizátora na webový server; nainštalovať a spúšťať je ho však možné aj lokálne.

Pred inštaláciou sumarizátora (na webový server) musia byť splnené tieto požiadavky:

- na webovom serveri, na ktorý plánujeme nainštalovať sumarizátor, je nainštalovaný operačný systém *Fedora 15* (aj keď opísaný postup je, možno s miernymi rozdielmi, aplikovateľný aj pre iné verzie Fedory, resp. iné linuxové distribúcie)
- je nainštalovaný webový server *Apache*
- je nainštalované prostredie *Java 6 JRE*
- používateľ má prístup k administrátorským oprávneniam

Postup inštalácie pozostáva zo piatich krokov:

1. Inštalácia ElasticSearch
2. Inštalácia MongoDB
3. Inštalácia Ruby
4. Inštalácie knižnice Linalg
5. Inštalácia samotného sumarizátora

C.1.1 Inštalácia ElasticSearch

1. Stiahnite si najnovšiu verziu ElasticSearch (celý príkaz je jeden riadok):
   ```
   $ wget
   https://github.com/downloads/elasticsearch/elasticsearch/elastic
   search-0.18.7.tar.gz
   ```
2. Extrahujte do želaného adresára:
   ```
   $ tar -zxvf elasticsearch-0.18.7.tar.gz
   ```
3. Stiahnite Java Service Wrapper pre ElasticSearch:
   ```
   $ git clone git://github.com/elasticsearch/elasticsearch-
   servicewrapper.git
   $ cp -r elasticsearch-servicewrapper/service elasticsearch-
   0.18.7/bin
   $ rm -r elasticsearch-servicewrapper
   ```
4. Nainštalujte ElasticSearch ako službu a spustite:
   ```
   $ bin/service/elasticsearch install
   $ bin/service/elasticsearch start
   ```
5. Vytvorte index `summarizer`:
   ```
   $ curl -XPUT 'http://localhost:9200/summarizer/'
   ```

C.1.2 Inštalácia MongoDB

1. Vytvorte súbor `/etc/yum.repos.d/10gen.repo` s týmto obsahom:
```
[10gen]
name=10gen Repository
baseurl=http://downloads-
distro.mongodb.org/repo/redhat/os/x86_64
gpgcheck=0
```
2. Nainštalujte MongoDB a spustite:
```
$ yum install mongo-10gen mongo-10gen-server
$ service mongod start
```

C.1.3 Inštalácia Ruby

1. Nainštalujte git a curl: `$ yum install git curl`
2. Nainštalujte rvm:
```
$ bash <<(curl -s https://rvm.beginrescueend.com/install/rvm)
$ source ~/.bashrc
```
3. Nainštalujte ďalšie balíčky odporúčané rvm: `$ rvm requirements`
4. Nainštalujte ruby 1.9.3 pomocou rvm: `$ rvm install 1.9.3`
5. Vytvorte gemset `summarizer` a nastavte ho ako predvolený:
```
$ rvm use 1.9.3
$ rvm gemset create summarizer
$ rvm --default use 1.9.3@summarizer
```

C.1.4 Inštalácia knižnice Linalg

1. Nainštalujte knižnicu LAPACK a ostatné závislosti:
```
$ yum install lapack lapack-devel blas blas-devel gcc-gfortran
compat-gcc-34 compat-gcc-34-g77
```
2. Nájdite súbor `g2c.h` a nakopírujte ho:
```
$ find / -name g2c.h
$ cp /usr/lib/gcc/x86_64-redhat-linux/3.4.6/include/g2c.h li-
nalg/ext/
$ cp linalg/ext/g2c.h linalg/ext/lapack/
$ cp linalg/ext/g2c.h linalg/ext/linalg/
```
3. Vytvorte symbolickú linku na súbor `libg2c.so.0` a prekopírujte ju:
```
$ ln -s /usr/lib64/libg2c.so.0 linalg/ext/libg2c.so
$ cp linalg/ext/libg2c.so linalg/ext/lapack/
$ cp linalg/ext/libg2c.so linalg/ext/linalg/
```
4. Spustite inštalačný skript z adresára `linalg`:
```
$ cd linalg
$ ruby linalg/install.rb
```

C.1.5 Inštalácia samotného sumarizátora

Pri inštalácii samotného sumarizátora je možné postupovať dvomi spôsobmi:

- *manuálne* – treba na serveri vytvoriť želanú adresárovú štruktúru a prekopírovať do nej zdrojové kódy sumarizátora
- *pomocou nástroja Capistrano* – sumarizátor je distribuovaný s nakonfigurovaným skriptom pre jednoduché vzdialené nasadenie na server

Opíšeme druhý spôsob, pri ktorom predpokladáme, že má používateľ prístup do git repozitára, a že má vytvorenú lokálnu kópiu sumarizátora. Okrem toho je potrebné mať správne nakonfigurovaný prístup cez SSH.

Ešte pred nasadením sumarizátora cez Capistrano je potrebné nakonfigurovať web server (tento krok je potrebný aj pri manuálnej inštalácii, pri ktorej však môže byť vykonaný až následne po nej, resp. nezáleží na poradí vykonania týchto krokov):

1. Nainštalujte Passenger (postupujte podľa pokynov inštalátora):
   ```
   $ gem install passenger
   $ passenger-install-apache2-module
   ```
2. Pridajte do konfiguračného súboru webového servera Apache tieto riadky (predpokladáme, že sumarizátor bude nasadený do adresára `/home/summarizer`):
   ```
   LoadModule passenger_module /home/summarizer/.rvm/gems/ruby-
   1.9.3-p0@summarizer/gems/passenger-
   3.0.11/ext/apache2/mod_passenger.so
   PassengerRoot /home/summarizer/.rvm/gems/ruby-1.9.3-
   p0@summarizer/gems/passenger-3.0.11
   PassengerRuby /home/summarizer/.rvm/wrappers/ruby-1.9.3-
   p0@summarizer/ruby

   <VirtualHost *:80>
     ServerName vm08.ucebne.fiit.stuba.sk # Nahraďte vlastnou
     DocumentRoot /home/summarizer/deploy/current/public
     <Directory /home/summarizer/deploy/current/public>
       Allow from all
       Options -MultiViews
     </Directory>
   </VirtualHost>
   ```

Následné nasadenie sumarizátora pomocou nástroja Capistrano je už triviálne:

```
$ cap deploy:setup   # nastaví adresárovú štruktúru
$ cap deplot:check   # skontroluje oprávnenia
$ cap deploy:setup   # spraví nasadenie, ale nespustí aplikáciu
$ cap deploy:start   # spustí aplikáciu
$ cap deploy:restart # reštartuje aplikáciu
$ cap deploy         # spraví kompletné nasadenie
```

C.2 Inštalácia výučbového systému ALEF

Pri vypracúvaní tejto príručky sme vychádzali z návodu spísaného na internej wiki projektu ALEF, ktorého autormi sú (okrem autora tejto práce) aj ďalší členovia riešiteľského kolektívu na projekte ALEF, menovite Martin Labaj, Michal Barla a Ivan Srba.

Opísaný je postup lokálnej inštalácie výučbového systému ALEF (serverová inštalácia je podobná inštalácii sumarizátora – na nasadenie je opäť možné použiť nástroj Capistrano) v prostredí operačného systému Ubuntu 11.10 32b.

Pred samotnou inštaláciou systému ALEF je potrebné mať nainštalované Ruby (návod pozri vyššie) a databázu MySQL (spolu s vývojovou knižnicou `libmysqlclient-dev`) a spravenú lokálnu kópiu zdrojových kódov systému ALEF (dostupné na priloženom elektronickom médiu).

Pri inštalácii postupujte nasledovne:

1. Nainštalujte závislosti knižnice Nokogiri (závisí od operačného systému):
   ```
   $ apt-get install libxslt-dev libxml2-dev
   ```
2. Nainštalujte balíky (tzv. gemy), ktoré potrebuje ALEF:
   ```
   $ bundle install
   ```

3. Vytvorte MySQL databázu (`alef-development` a `alef-test`) a používateľa, ktorý k nej bude mať plný prístup; prístupové údaje k nej nastavte v súbore `config/database.yml`.
4. Nainštalujte NoSQL databázu Redis:
```
$ wget http://redis.googlecode.com/files/redis-2.2.2.tar.gz
$ tar xzvf redis-2.2.2.tar.gz
$ make
$ make install
```
5. Vytvorte adresár `shared` a rozbaľte doň súbory:
```
complete_courses.zip
collected_for_manual_init.zip.
```
6. Naplňte ALEF dátami:
```
$ rake db:create RAILS_ENV='development'
$ rake db:schema:load
$ rake db:migrate
$ rake alef:init_data
```
7. Spustite inštanciu Redisu a Resque procesy (ak chcete, aby sa tieto vykonávali na pozadí, použite navyše prepínač `--daemon`):
```
$ redis-server
$ resque-pool --environment development --pidfile
tmp/pids/resque-pool.pid
```
8. Spustite server:
```
$ rails s
```
9. Overte funkčnosť na stránke `http://localhost:3000` prihlásením sa pomocou mena `staff` a hesla `staff` (ako administrátor), prípadne mena `student` a hesla `student` (ako študent).

Príloha D: Používateľská príručka

Sumarizátor sme implementovali ako webovú REST službu, tzn. že nemá grafické používateľské rozhranie – komunikácia je možná prostredníctvom poskytnutého aplikačného rozhrania, ktoré sme opísali v časti B.2.1 technickej dokumentácie. Aktuálnu verziu sumarizátora možno nájsť na adrese:

http://vm08.ucebne.fiit.stuba.sk/summarizer

Pri overovaní sumarizátora sme využili výučbový systém ALEF, ktorý sme integrovali so sumarizátorom pomocou sumarizačného komponentu. Okrem toho sme v prostredí systému ALEF implementovali metódu personalizovaného výberu dokumentov na opakovanie.

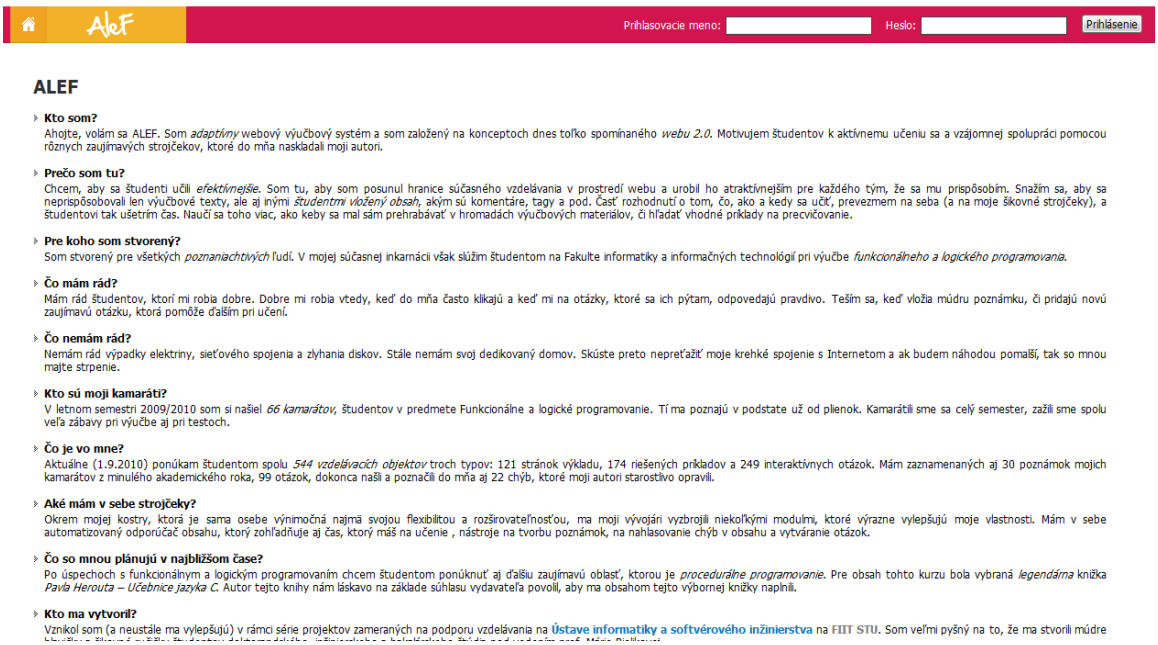
Uvádzame tu preto používateľskú príručku výučbového systému ALEF, na ktorej vypracovaní sa podieľali členovia riešiteľského kolektívu na projekte ALEF: Pavel Michlík, Martin Labaj, Vladimír Mihál, Maroš Unčík, Jakub Ševcech, Máté Fejes, Róbert Móro a Andrea Šteňová.

D.1 Prístup do systému ALEF

Systém ALEF (Adaptive Learning Framework) je adaptívny výučbový systém na podporu výučby na predmetoch Funkcionálne a logické programovanie a Procedurálne programovanie na Fakulte informatiky a informačných technológií na Slovenskej Technickej Univerzite v Bratislave. Aktuálnu ukážkovú verzia tohto systému možno nájsť na webovej adrese:

http://alef.fiit.stuba.sk/

Po návšteve webovej adresy výučbového systému ALEF sa zobrazí úvodná obrazovka, kde sa vo vrchnej časti nachádza prihlasovací formulár (Obr. D-1). V prípade, že máte vytvorené konto v systéme ALEF, zadajte svoje prihlasovacie meno a heslo, do formulára v pravej hornej časti okna.

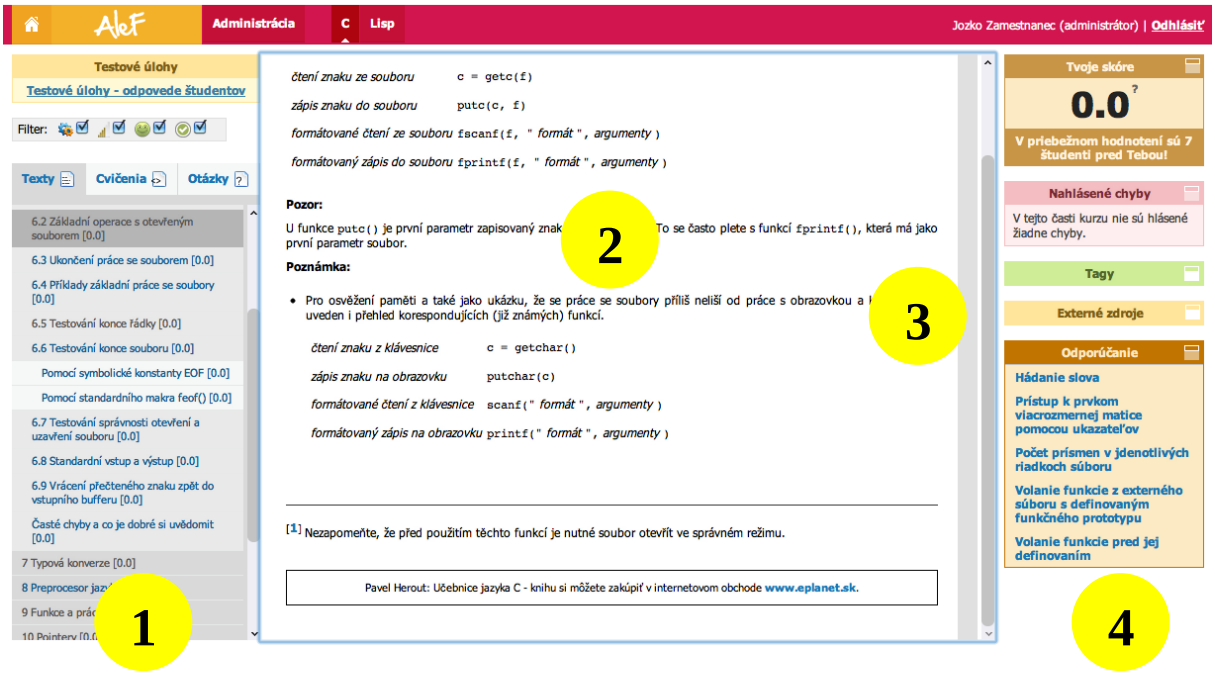


Obr. D-1 Úvodná obrazovka a prihlasovací formulár.

D.2 Práca so systémom

Po prihlásení do systému sa nezávisle od role používateľa zobrazí úvodná obrazovka systému (Obr. D-2). Ak nie je pri komponente uvedené inak, zobrazuje sa používateľom oboch rolí (t.j. študentom aj učiteľom). Úvodná obrazovka je rozdelená na štyri časti:

1. *navigačná časť*
    a. zobrazuje odporúčanie pre študentov
    b. zobrazuje menu pre navigáciu
2. *obsahová časť*
    a. zobrazuje aktuálne zvolený výučbový materiál
    b. anotačný pásik znázorňujúci poznámky a nahlásené chyby
3. *časť so zásuvnými modulmi*
    a. zobrazuje zásuvné moduly v systéme



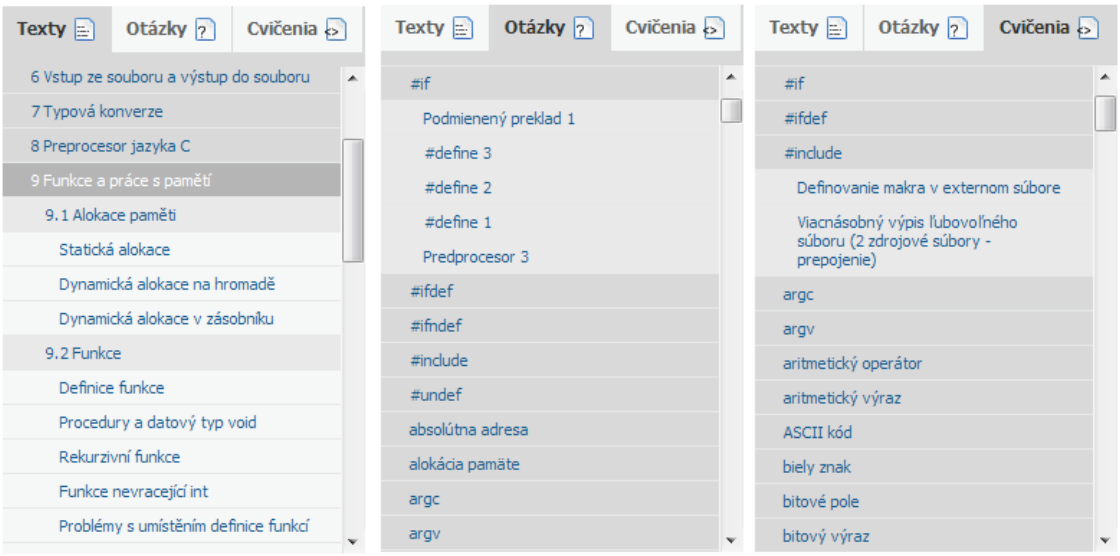
Obr. D-2 Úvodná obrazovka.

D.2.1 Navigačná časť

Hlavný komponent navigačnej časti tvorí menu. Jeho obsahom je zoznam všetkých vzdelávacích objektov. Zoznam je usporiadaný podľa kapitol, t.j. v rámci daných vzdelávacích objektov sa môžu nachádzať aj ďalšie. Kliknutím na kapitolu sa vybraná položka rozbalí a zobrazia sa k nej patriace podkapitoly.

Menu (Obr. D-3) sa skladá z troch záložiek:

- *Texty* – záložka obsahuje zoznam výučbových textov,
- *Otázky* – záložka obsahuje zoznam otázok, na ktoré je možné odpovedať,
- *Cvičenia* – záložka obsahuje zoznam cvičení, ktoré je možné vyriešiť.



Obr. D-3 Ukážky menu. Zľava záložka výučbových textov, v strede záložka otázok a vpravo záložka príkladov.

Pre rýchlejšiu navigáciu medzi cvičeniami je pridané filtrovanie otázok podľa pridaných tagov používateľov (Obr. D-4).

Obr. D-4 Ukážky menu cvičení s filtrovaním. Zľava záložka cvičení s filtrovaním podľa tagov, vpravo záložka všetkých cvičení.

V menu pri cvičeniach a otázkach sú okrem ich názvu pridané informácie o náročnosti daného objektu, ako aj ďalšie informácie o objekte. Vyznačenie informácii:

1. *Podfarbenie* – slúži na vizualizáciu náročnosti pri cvičeniach. Zelená farba značí najľahšie cvičenia, žltá stredne ľahké a červená najťažšie cvičenia.

2. *Predchádzajúce riešenie* – touto ikonou systém informuje študenta o tom, či otázku alebo cvičenie už riešil. Farba ikonu udáva správnosť alebo nesprávnosť predchádzajúceho riešenia.
3. *Hodnotenie užitočnosti* – zobrazuje predchádzajúce hodnotenie užitočnosti študenta konkrétneho objektu.
4. *Náročnosť objektu* – na rozdiel od podfarbenia objektu, táto ikona slúži na zobrazenie predchádajúceho hodnotenia náročnosti študenta konkrétneho objektu.
5. *Externý testovač* – označuje, že príslušné cvičenie je možné automaticky vyhodnotiť externým testovačom.

To, aké informácie sa pri cvičeniach a otázkach zobrazia, si môže používateľ zvoliť pomocou filtra umiestneného nad menu (Obr. D-5).

Obr. D-5 Filter informácií zobrazených pri otázkach a cvičeniach v menu.

D.2.2 Obsahová časť

V závislosti od vybraného výučbového materiálu sa môžu v obsahovej časti nachádzať štyri typy zobrazených materiálov:

1. otázka,
2. cvičenie,
3. výučbový text,
4. testová otázka s odpoveďou používateľa,
5. sumarizácie výučbových textov odporúčaných na opakovanie.

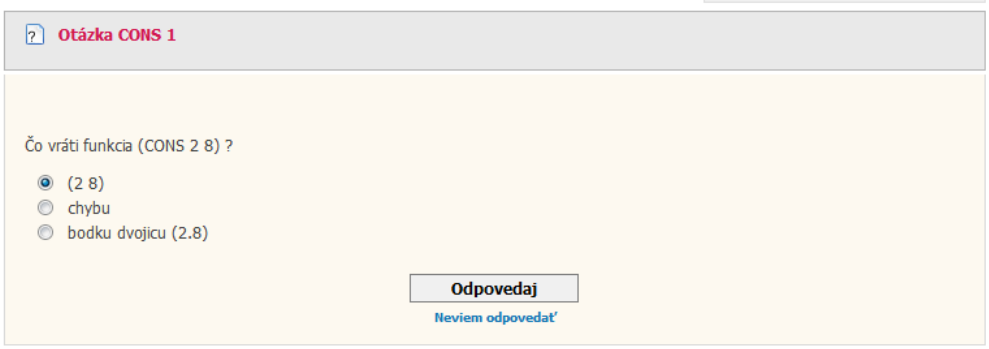
D.2.2.1 Odpovedanie na otázku

Otázka je typ výučbového materiálu, na ktorý sa dá odpovedať a systém vyhodnotí odpoveď študenta. V závislosti od typu otázok sa zobrazí príslušný formulár na jej zodpovedanie (Obr. D-6). V každom formulári študent vyplní správnu odpoveď a klikne na tlačidlo *Odpovedaj*. Systém následne vyhodnotí odpoveď študenta a zobrazí výsledok (Obr. D-7, Obr. D-8).

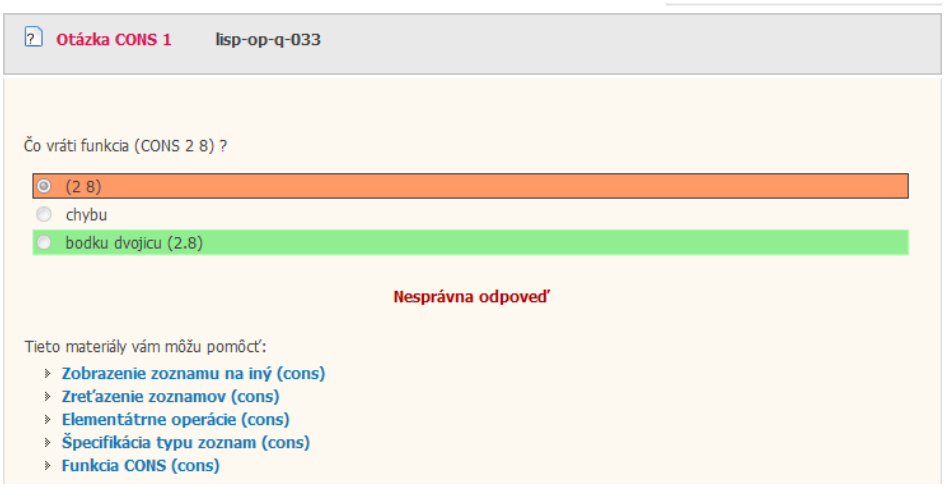
Otázky môžu byť viacerých typov:

- otázka s jednou správnou odpoveďou,
- otázka s viacerými správnymi odpoveďami,
- odpoveď voľným textom,
- odpoveď doplnením textu,
- zoradenie možností.

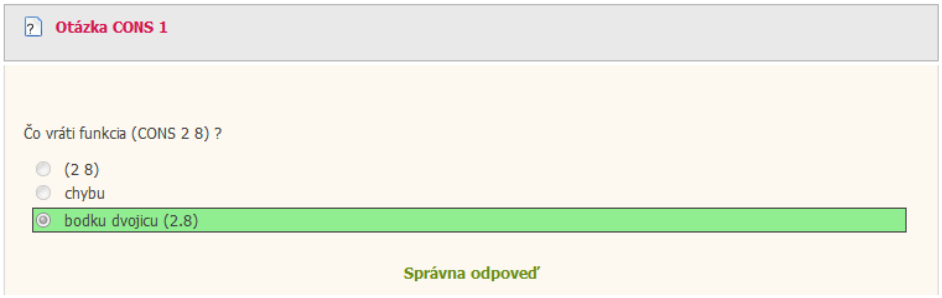
Ak odpovedá študent nesprávne, zelenou farbou sú vyznačené správne odpovede, červenou sú označené nesprávne odpovede. Čiernym rámčekom sú vyznačené odpovede, ktoré označil študent (Obr. D-7). Pod otázkou sa študentovi zobrazia texty, ktoré by mu mohli pomôcť porozumieť riešeniu.



Obr. D-6 Formulár pre zodpovedanie otázky s jednou správnou odpoveďou.



Obr. D-7 Zodpovedaná otázka, zlá odpoveď.



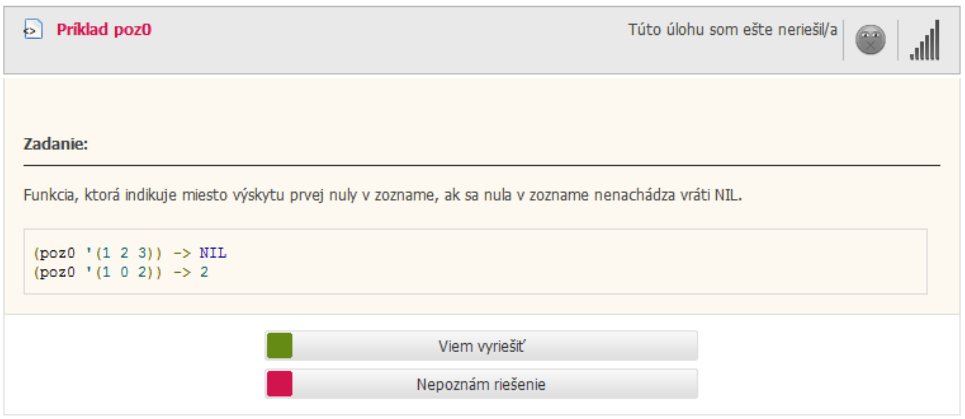
Obr. D-8 Zodpovedaná otázka, správna odpoveď.

D.2.2.2 Odpovedanie na cvičenie

Cvičenie je typ výučbového materiálu, na ktorý sa dá vyriešiť a následne oznámiť systému postup v riešení (Obr. D-9). Študent postupne kliká na jednotlivé možnosti, ktoré mu systém ponúka pri riešení. Možnosti sú nasledovné:

1. *Viem vyriešiť* – systém zobrazí správne riešenie; následne si študent vyberie z možností:
    a. *Moje riešenie je rovnaké ako vzorové riešenie*
    b. *Moje riešenie je iné, ale myslím, že správne*
    c. *Moje riešenie je nesprávne, ale už tomu rozumiem*
    d. *Moje riešenie je nesprávne a stále tomu nerozumiem*

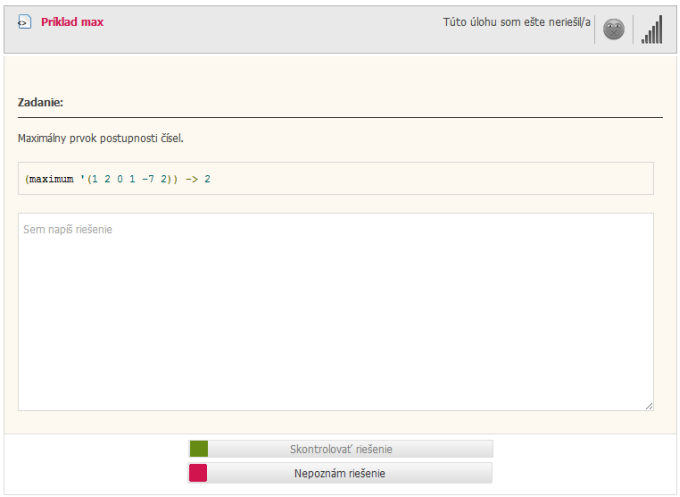
2. *Nepoznám riešenie* – systém zobrazí pomôcku; následne si študent vyberie z možností:
    a. *Už viem vyriešiť* – systém zobrazí správne riešenie; následne si študent vyberie z možností:
        i. *Moje riešenie je rovnaké ako vzorové riešenie*
        ii. *Moje riešenie je iné, ale myslím, že správne*
        iii. *Moje riešenie je nesprávne, ale už tomu rozumiem*
        iv. *Moje riešenie je nesprávne a stále tomu nerozumiem*
    b. *Stále neviem vyriešiť* – systém zobrazí správne riešenie; následne si študent vyberie z možností:
        i. *Riešeniu rozumiem*
        ii. *Riešeniu nerozumiem*



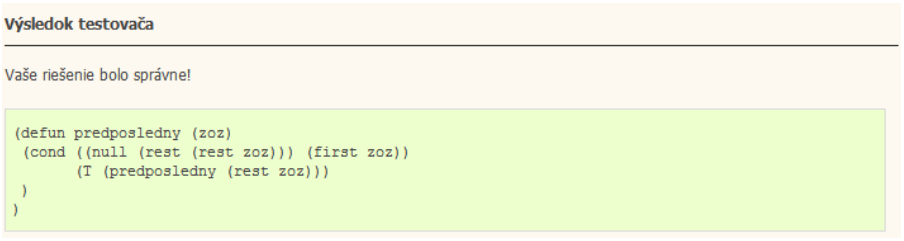
Obr. D-9 Formulár pre zodpovedanie cvičenia.

Pri niektorých cvičeniach je možnosť otestovania správneho riešenia externým testovačom (Obr. D-10). Pri tomto type cvičenia má možnosť používateľ zobraziť si pomôcku, ak nepozná riešenia, alebo vložiť a otestovať si riešenie. Následne testovač vráti výsledok, či bolo dané riešenie správne (Obr. D-11), alebo nesprávne (Obr. D-12).

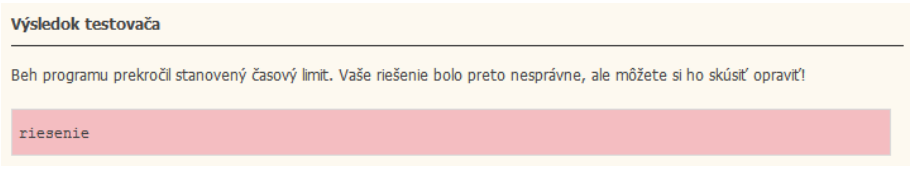
Študent má pri cvičeniach a otázkach možnosť ohodnotiť náročnosť a užitočnosť konkrétneho objektu (Obr. D-13).



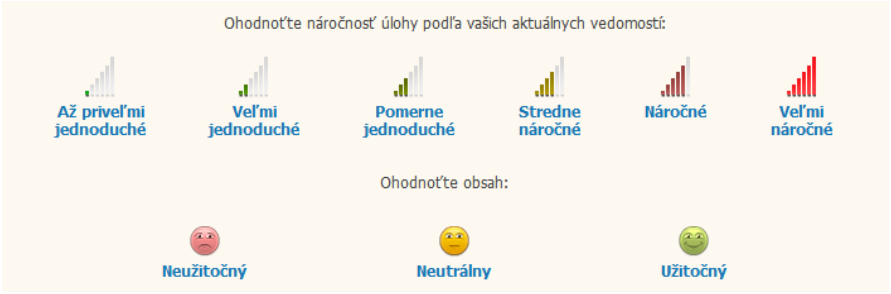
Obr. D-10 Formulár pre zodpovedanie cvičenia s testovačom.



Obr. D-11 Výsledok testovača pre správne riešenie.



Obr. D-12 Výsledok testovača pre nesprávne riešenie.



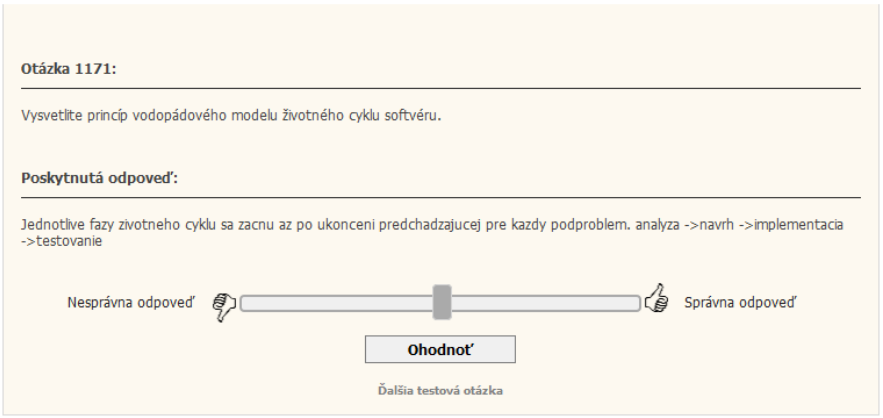
Obr. D-13 Hodnotenie objektu.

D.2.2.3 Výučbový text

Výučbový text je statický text, ktorý môže obsahovať tabuľky, obrázky, grafy, hypertextové odkazy a ďalšie prvky.

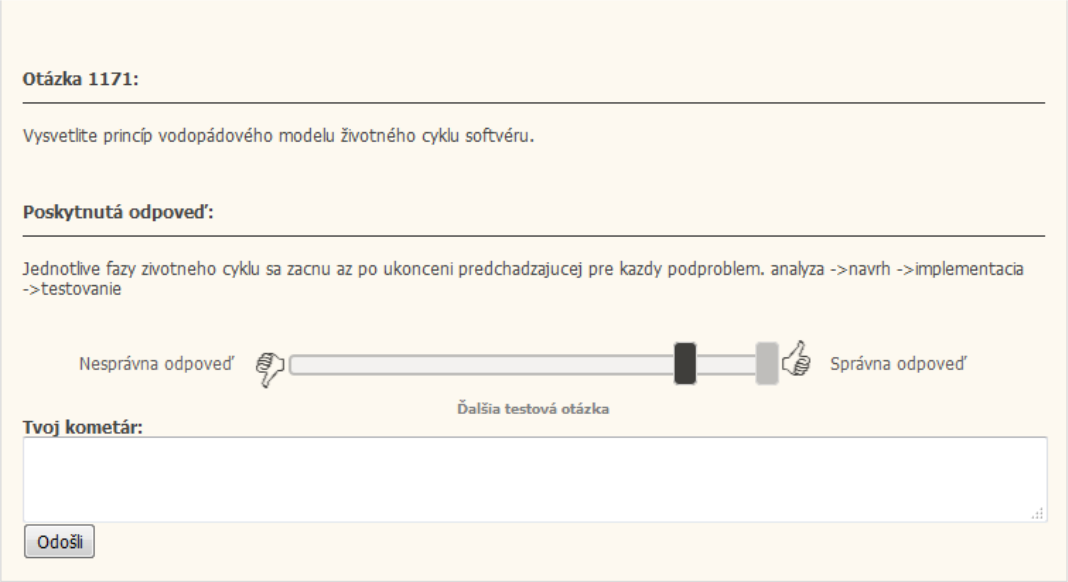
D.2.2.4 Testová otázka s odpoveďou študenta

Testová otázka predstavuje otázku zo vstupných testov, na ktorú boli získané odpovede od študentov. Používateľ sa k tomuto druhu vzdelávacieho objektu dostane kliknutím na odkaz v komponente pre zobrazenie testových úloh (pozri časť D.2.4.6). Pri tomto objekte je úlohou používateľa určiť správnosť odpovede na otázku (Obr. D-14).



Obr. D-14 Testová otázka s odpoveďou.

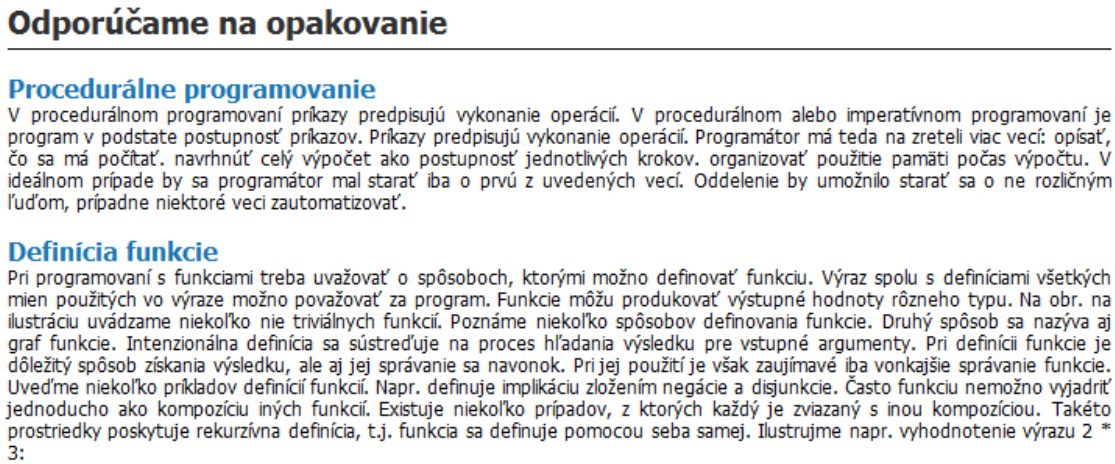
Po ohodnotení správnosti odpovede sa používateľovi zobrazí, aké je doterajšie priemerné hodnotenie správnosti získané na základe odpovedí ostatných používateľov Obr. D-15). Používateľ tiež môže po ohodnotení vložiť komentár a zároveň sa mu zobrazia komentáre, ktoré pridali ostatní používatelia.



Obr. D-15 Testová otázka s odpoveďou po ohodnotení používateľom.

D.2.2.5 Sumarizácie textov odporúčaných na opakovanie

Ak si chce používateľ opakovať, môže kliknúť na odkaz v komponente pre opakovanie (pozri časť D.2.4.7). Následne sa mu v obsahovej časti zobrazí zoznam personalizovane odporúčaných výučbových textov na opakovanie spolu s ich sumarizáciami (Obr. D-16). Po kliknutí na názov niektorého z odporúčaných textov sa mu daný výučbový text zobrazí celý, nielen jeho sumarizácia.



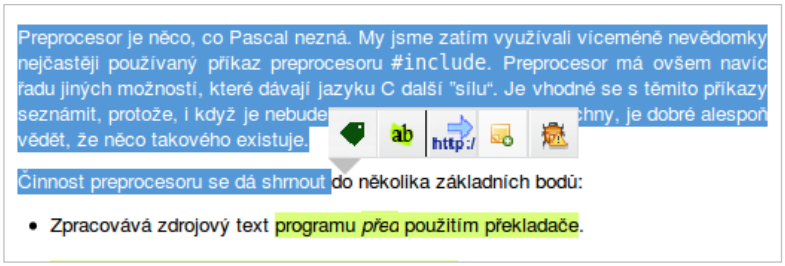
Obr. D-16 Zoznam výučbových textov odporúčaných na opakovanie spolu s ich sumarizáciami.

D.2.3 Poznámkovanie výučbových materiálov

D.2.3.1 Pridávanie poznámok

Pridávať rôzne druhy poznámok sa v systéme dá k ľubovoľnému výučbovému materiálu (textom, otázkam a aj príkladom). Používateľ označí text, ku ktorému chce pridať poznámku. Následne si z kontextového menu môže vybrať jednu z možností (Obr. D-17):

1.  *Pridanie tagu* – slúži na pridanie vyznačeného textu ako tag. Tagy sa zobrazujú zvýraznením pozadia otagovaného textu ako aj v komponente na pridávanie a zobrazovanie tagov.
2.  *Označenie textu* – vyznačenému textu sa zmení pozadie. Táto funkcia slúži na zvýraznenie dôležitých alebo inak zaujímavých častí textu.
3.  *Pridanie externého zdroju* – pomocou tejto voľby je možné k textu priradiť externý odkaz, ktorý sa bude zobrazovať v komponente na pridávanie a zobrazovanie externých odkazov.
4.  *Nový komentár* (Obr. D-18) – slúži na pridanie poznámky k výučbovému materiálu. Poznámka sa zobrazuje na mieste, ku ktorému bola pridaná. Poznámka je viditeľná pre všetkých používateľov.
5.  *Nahlásenie chyby* – umožňuje pridať poznámku, ktorá upozorňuje ostatných študentov a učiteľa, že v danom výučbovom materiáli sa nachádza chyba.

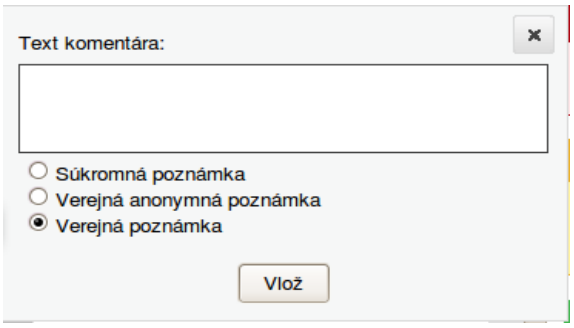


Obr. D-17 Označenie textu.

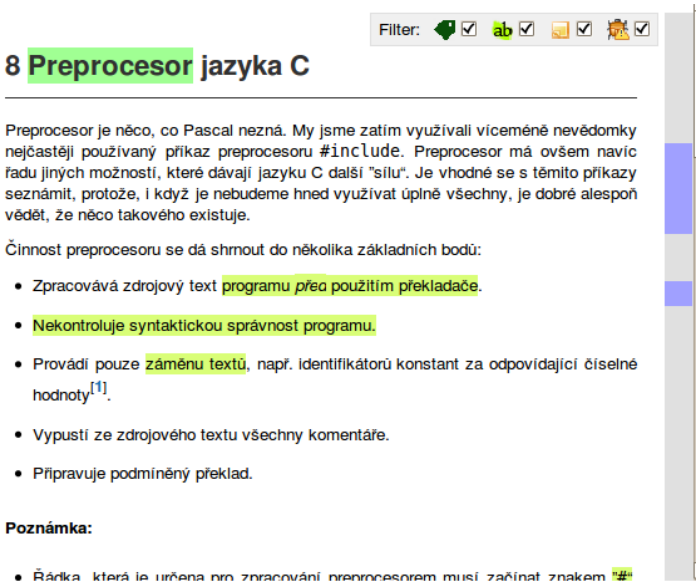
D.2.3.2 Anotačný pásik

Anotačný pásik (Obr. D-19) zobrazuje komentáre vytvorené používateľmi a nahlásené chyby vo výučbových objektoch. Ich prítomnosť reprezentuje modrý obdĺžnik. Anotácia je zobrazená ukázaním kurzorom na modrú oblasť, ktorá sa zvýrazní zelenou farbou (Obr. D-20).

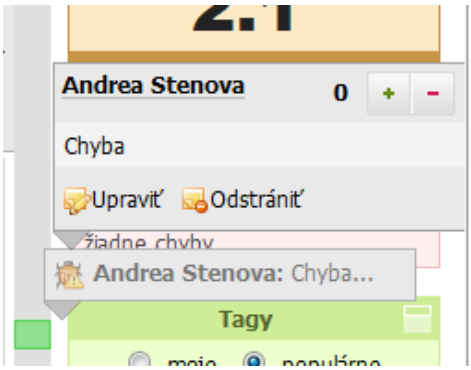
Pre filtrovanie medzi poznámkami a nahlásenými chybami slúži filter (Obr. D-19 vpravo hore). Prvá ikona predstavuje zobrazenie tagov vložených z textu, druhá ikona predstavuje zvýraznené výseky textu, tretia predstavuje zobrazenie poznámok a posledná zobrazenie nahlásených chýb. Používateľ si tak môže zvoliť, ktoré anotácie chce v danom výučbovom materiáli vidieť.



Obr. D-18 Rozhranie pre pridanie komentáru.



Obr. D-19 Anotačný pásik. Nachádza sa napravo od výučbového materiálu. Modrou sú znázornené anotácie vložené používateľmi.



Obr. D-20 Zobrazenie anotácie.

D.2.4 Časť so zásuvnými modulmi

V pravej časti obrazovky sa nachádzajú zásuvné moduly systému. V aktuálnej verzii systému sa nachádzajú štyri moduly:

1. *Komponent pre zobrazenie skóre* – zobrazuje bodový zisk prihláseného študenta. Body sú študentovi pripisované v závislosti od aktivít, ktoré vykoná v súvislosti

s rôznymi časťami systému (pridá tag do textu, nahlási chybu, pridá poznámku a pod.).

2. *Komponent pre zobrazovanie nahlásených chýb* – zobrazuje nahlásené chyby vo výučbových materiáloch od používateľov; zobrazuje sa len používateľom s rolou *Administrátor* a *Učiteľ*.
3. *Komponent pre zobrazenie a pridávanie externých odkazov* – zobrazuje externé zdroje, ktoré pridali študenti alebo vyučujúci k danému výučbovému materiálu. Umožňuje tiež pridať ďalšie externé zdroje.
4. *Komponent pre zobrazenie a pridávanie tagov* – zobrazuje tagy priradené k danému výučbovému materiálu. Umožňuje tiež pridať ďalšie tagy či už priamo jeho napísaním alebo označením časti textu, ktorý sa má použiť ako tag.

D.2.4.1 Komponent pre zobrazovanie skóre

Komponent pre zobrazenie skóre (Obr. D-21) slúži ako informatívny prvok pre študentov, ktorý sa využíva pri hre za účelom motivácie. Hra spočíva v získavaní bodov a snahe študentov dosiahnuť čo najvyššie skóre. Komponent zobrazuje aktuálnu výšku bodov, ktoré študent získal za rôzne akcie v systéme. Komponent zároveň zisťuje poradie študenta v celkovom bodovom poradí a tento údaj zobrazuje, za účelom motivácie študentov.

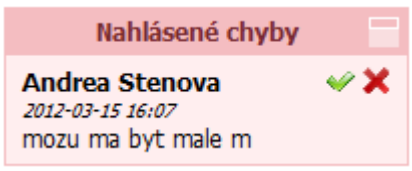
Obr. D-21 Komponent pre zobrazenie skóre.

D.2.4.2 Komponent pre zobrazenie nahlásených chýb

Komponent pre zobrazenie nahlásených chýb (Obr. D-22) slúži ako informatívny prvok. Komponent zobrazuje nahlásené chyby od používateľov v tabuľke v poradí: *používateľské meno autora nahlásenej chyby, čas nahlásenia chyby* a *text hlásenia resp. popis chyby*. Po prechode myšou nad textom hlásenia, vo výučbovom texte sa zvýrazní časť textu, ku ktorému bola chyba nahlásená. V pravej hornej časti sú ku každej nahlásenej chybe zobrazené dve ikony:

-  Slúži na označenie chyby za vyriešenú, takto označená chyba sa ďalej nebude zobrazovať.
-  Slúži na zmazanie chybového hlásenia

Tieto ikony sú dostupné len pre používateľov s rolou *Administrátor* a *Učiteľ*.

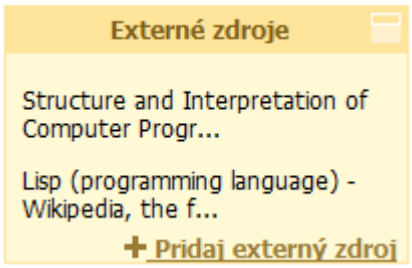


Obr. D-22 Komponent pre zobrazenie nahlásených chýb.

D.2.4.3 Komponent pre prácu s externými zdrojmi

Komponent pre prácu s externými zdrojmi (Obr. D-23) umožňuje používateľom vkladať odkazy na stránky s obsahom súvisiacim s obsahom aktuálne zobrazeného vzdelávacieho objektu. Používatelia môžu taktiež tieto odkazy hodnotiť. Rovnako ako u iných typov poznámok, pri vkladaní externých zdrojov môže používateľ špecifikovať ich viditeľnosť – teda či sú súkromné, anonymné alebo verejné. Formulár na vkladanie odkazu na externý zdroj sa zobrazí po kliknutí na odkaz „✚ Pridaj externý zdroj“.

V komponente sú zobrazené všetky externé zdroje vložené k momentálne zobrazenému vzdelávaciemu objektu. Po zobrazení objektu sú v komponente zobrazené iba tri externé zdroje s najvyšším hodnotením od používateľov. Všetky externé zdroje používateľ zobrazí kliknutím na odkaz „Ukáž ďalšie“.

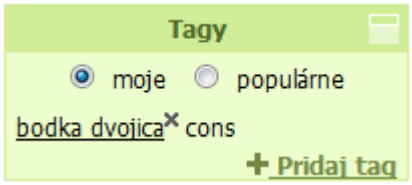


Obr. D-23 Komponent pre prácu s externými zdrojmi.

D.2.4.4 Komponent pre zobrazenie a pridávanie tagov

Komponent pre zobrazenie a pridávanie tagov (Obr. D-24) umožňuje používateľom pridávať tagy ku vzdelávacím objektom, prehliadať ich a odstraňovať. Používatelia môžu pri vkladaní tagu špecifikovať, či bude vložený ako verejný, anonymný alebo súkromný. Tag môže používateľ odstrániť nadídením nad tag a kliknutím na tlačidlo „✖“.

Používateľ môže prepínať medzi zobrazením jeho vložených tagov a populárnych tagov. Populárnym sa tag vtedy, pokiaľ je verejný a je k vzdelávaciemu objektu vložený minimálne tromi používateľmi.



Obr. D-24 Komponent pre zobrazenie a pridávanie tagov.

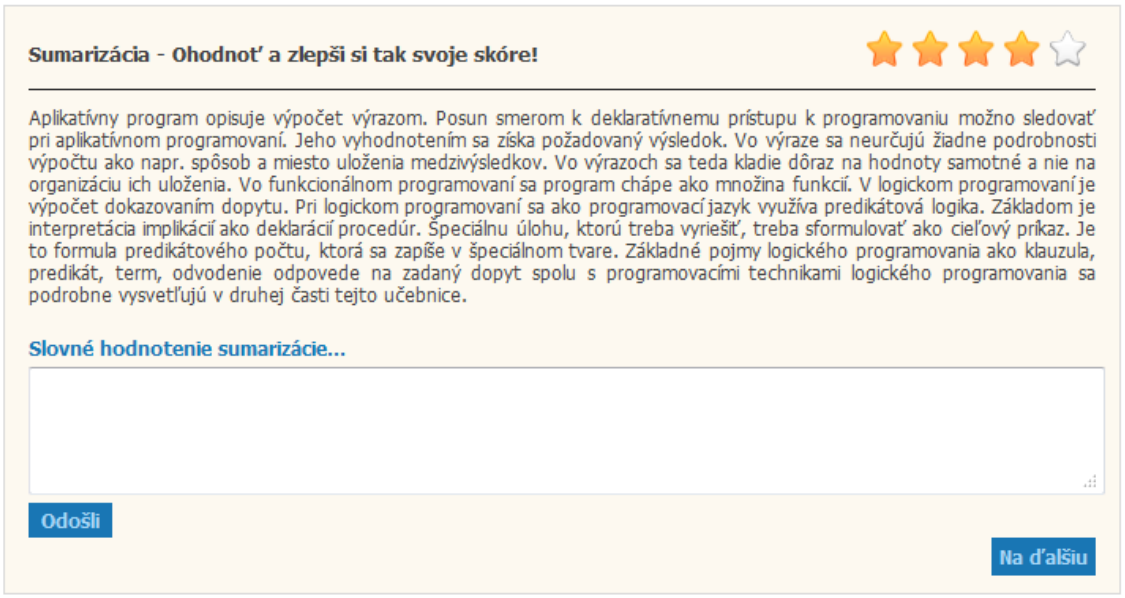
D.2.4.5 Komponent pre zobrazenie sumarizácie výučbového textu

Komponent pre zobrazenie sumarizácie výučbové textu (Obr. D-25) poskytuje používateľom súhrn daného textu. Dĺžka súhrnu závisí od nastavenia, vždy však ide o vety vybrané (extrahované) z daného textu.

Umiestnený môže byť nad alebo pod textom – ak je umiestnený nad, cieľom je pomôcť používateľovi rozhodnúť sa, či je daný text preňho relevantný a teda si ho má prečítať celý – vhodný je preto skôr kratší súhrn. Umiestnenie pod textom môže mať význam v tom, aby

si používateľ po prečítaní textu mohol opätovne prejsť najdôležitejšie časti (ale predpokladáme, že skôr sa použije umiestnenie nad textom).

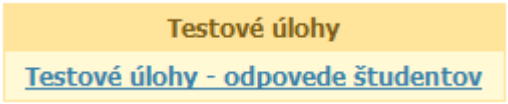
Používatelia sa môžu vyjadriť aj ku kvalite sumarizácie – buď ohodnotením pomocou hviezdičiek alebo zaslaním slovného hodnotenia; pomocou tlačidla „Na ďalšiu“ sa dostanú na ďalší výučbový text so sumarizáciou v poradí podľa menu.



Obr. D-25 Komponent pre zobrazenie sumarizácie výučbového textu.

D.2.4.6 Komponent pre zobrazenie testových úloh

Komponent s testovými úlohami (Obr. D-26) je umiestnený v navigačnej časti nad menu. Študenti sú po kliknutí na odkaz presmerovaní na testové úlohy s odpoveďami študentov, ktorých správnosť môžu hodnotiť (pozri časť D.2.2.4).

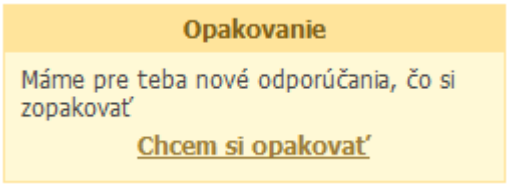


Obr. D-26 Komponent na zobrazenie testových úloh s možnosťou hodnotenia odpovedí študentov.

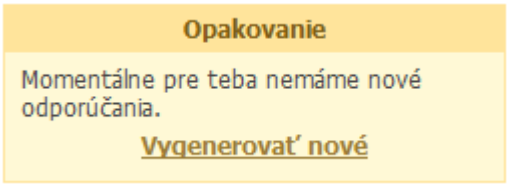
D.2.4.7 Komponent pre opakovanie

Komponent pre opakovanie (Obr. D-27) zobrazuje používateľovi odkaz, po kliknutí na ktorý sa mu v obsahovej časti zobrazí zoznam výučbových textov personalizovane odporúčaných na opakovanie (pozri časť D.2.2.5).

Môže sa stať, že nové odporúčania nie sú k dispozícii, pretože sa generujú len v istých intervaloch, spravidla raz za deň. V takom prípade sa používateľovi zobrazí príslušný oznam a tiež možnosť vygenerovať okamžite nové odporúčania (Obr. D-28). Po zvolení tejto možnosti sa zobrazí používateľovi správa, že sa mu pripravujú nové odporúčania, ktorá sa zmení na do stavu na Obr. D-27 hneď, ako budú nové odporúčania pripravené.



Obr. D-27 Komponent pre opakovanie s pripravenými odporúčaniami.



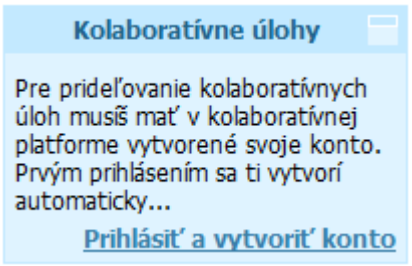
Obr. D-28 Komponent pre opakovanie - odporúčania nie sú k dispozícii; používateľ má možnosť nechať si vygenerovať nové.

D.2.4.8 Komponent pre integráciu s kolaboratívnou platformou

Komponent (Obr. D-29) zabezpečuje integráciu s kolaboratívnou platformou *Popcorm*, v rámci ktorej môžu študenti spolupracovať na riešení úloh v dynamicky vytváraných skupinách.

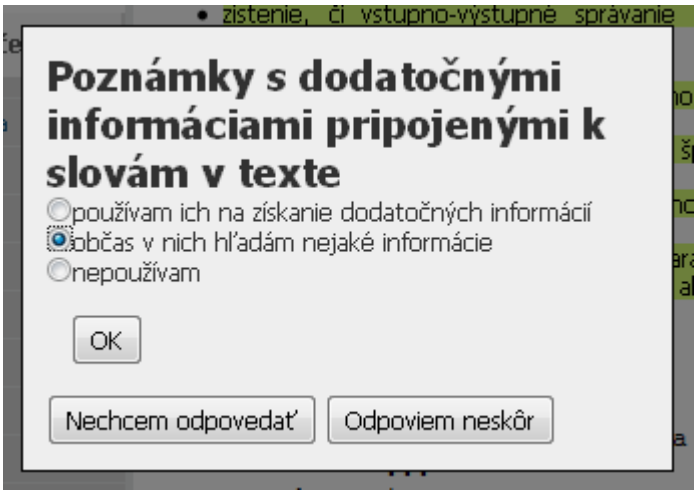
D.2.5 Adaptívne systémové otázky

Systém môže v prípade vhodnej situácie vygenerovať a zobraziť otázku (Obr. D-30), slúžiacu väčšinou na získanie odozvy od používateľa na funkcie systému. Používateľ odpovie vyplnením odpovede podľa typu otázky (výber jednej možnosti, vpísanie reakcie a pod.) a stlačením tlačidla *OK* odpoveď odošle. Ak používateľ nechce na otázku odpovedať, môže stlačením tlačidla *Nechcem odpovedať* alebo *Odpoviem neskôr* odpoveď zrušiť, respektíve odložiť.



Obr. D-29 Komponent pre integráciu s kolaboratívnou platformou.

V prípade odmietnutia odpovede sa práve zobrazená otázka deaktivuje a už sa nezobrazí (čo však nevylučuje, že systém neskôr nevygeneruje rovnakú otázku nanovo, ak to bude vhodné). V prípade odloženia odpovede sa otázka prestane zobrazovať, ale o krátku chvíľu sa zobrazí znova. Niektoré otázky sú špeciálne zamerané na aktuálny okamih a preto sa už znova nezobrazia, ani ak ich používateľ iba odloží tlačidlom *Odpoviem neskôr*.



Obr. D-30 Príklad systémovej otázky. Zobrazí sa v popredí pred výučbovým objektom.

D.3 Administrácia systému

V administrácii systému prebiehajú všetky nastavenia jednotlivých komponentov, ako aj vyhodnotenia aktivity študentov a vyhodnotenia objektov. Prihlásený administrátor má k dispozícii v hornom menu prechod z výučbovej časti do administrátorskej (Obr. D-31).

Obr. D-31 Položka menu na prechod do administrátorskej sekcie.

Po prechode do sekcie pre administrátorov sa zobrazí menu so všetkými možnosťami činností, ktoré môže administrátor vykonať (Obr. D-32) . Kliknutím na položku menu „Späť do výučby“ sa je možné prepnúť z administrátorskej časti systému znovu do výučby.

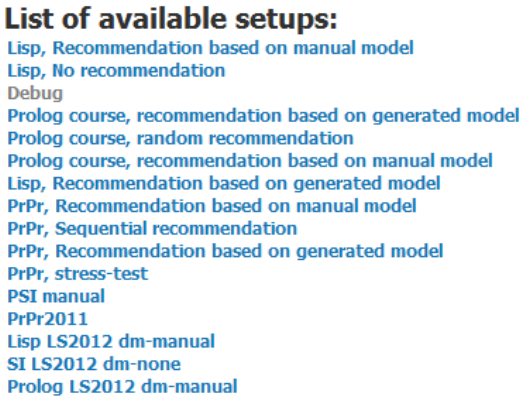
Obr. D-32 Menu pre administrátora.

Jednotlivé činnosti administrátora postupne opíšeme.

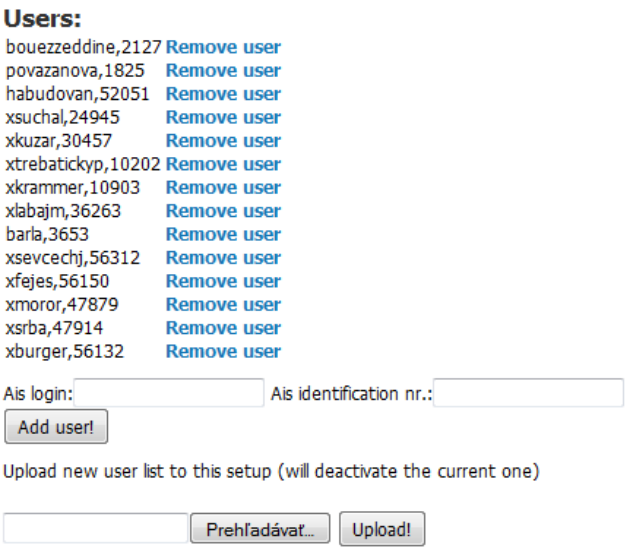
D.3.1 Konfigurácie

V tejto časti administrátor nastavuje parametre pre fungovanie jednotlivých komponentov systému. Po kliknutí na odkaz v menu, sa zobrazia všetky dostupné konfigurácie (Obr. D-33).

Do konfigurácii môže administrátor pridávať používateľov systému (Obr. D-34) zadaním ich prihlasovacieho mena do Akademického informačného systému (AIS) a ich identifikačného čísla. Alternatívne môže pridať naraz skupinu používateľov nahraním CSV súboru s potrebnými údajmi. Pridaných používateľov je možné z konfigurácie odoberať.

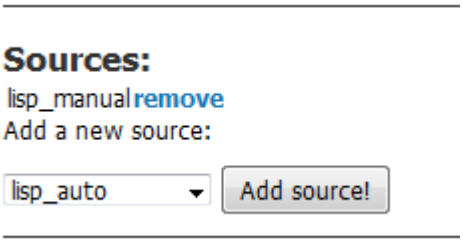


Obr. D-33 Zoznam konfigurácií v systéme.



Obr. D-34 Administrácia používateľov kurzu.

Do konfigurácie je potrebné pridať kurzy (zdroje), ktoré sa v nej budú zobrazovať (Obr. D-35). Kurzy zahŕňajú okrem samotných dát (výučbových materiálov, t.j. textov, otázok a cvičení) aj k nim naviazané metadáta v podobe konceptuálneho doménového modelu, ktorý môže byť manuálne vytvorený alebo generovaný automaticky. K dispozícii na pridanie sú všetky kurzy, ktoré boli pridané do systému. Tiež je možné zmazať niektorý z kurzov z aktuálnej konfigurácie.

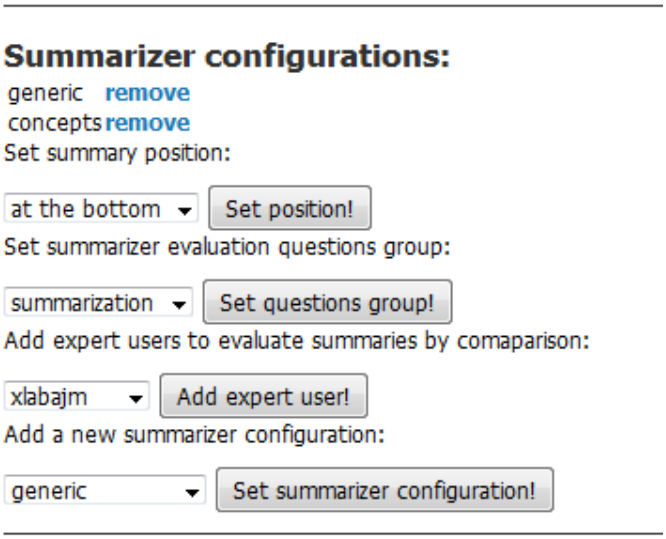


Obr. D-35 Nastavenie konfigurácie.

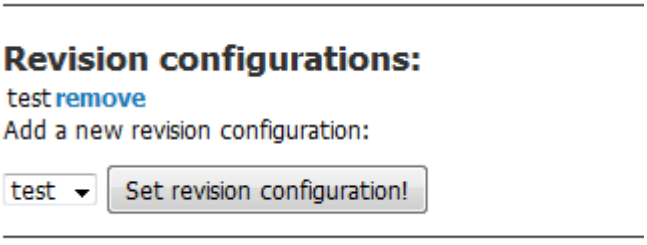
Ďalej je možné v konfiguráciách nastavovať parametre jednotlivých komponentov systému:

- Zapnutie/vypnutie integrácie s kolaboratívnou platformou Popcorm (Obr. D-36).
- Výber konfigurácie sumarizátora a nastavenie ďalších parametrov: poloha komponentu, určenie expertov za účelom vyhodnotenia a skupiny spätno-väzobných otázok (Obr. D-37).
- Výber konfigurácie opakovania (Obr. D-38).
- Nastavenie komponentu pre testové úlohy (Obr. D-39).

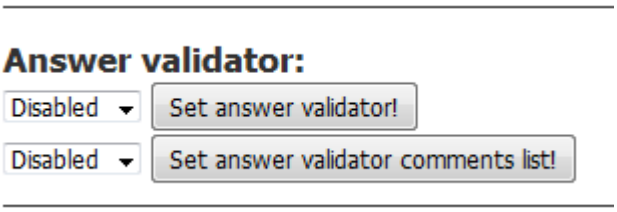
Obr. D-36 Zapnutie/vypnutie integrácie s kolaboratívnou platformou.



Obr. D-37 Nastavenie sumarizátora.



Obr. D-38 Nastavenie opakovania.



Obr. D-39 Nastavenie komponentu pre testové úlohy.

D.3.2 Texty kurzov

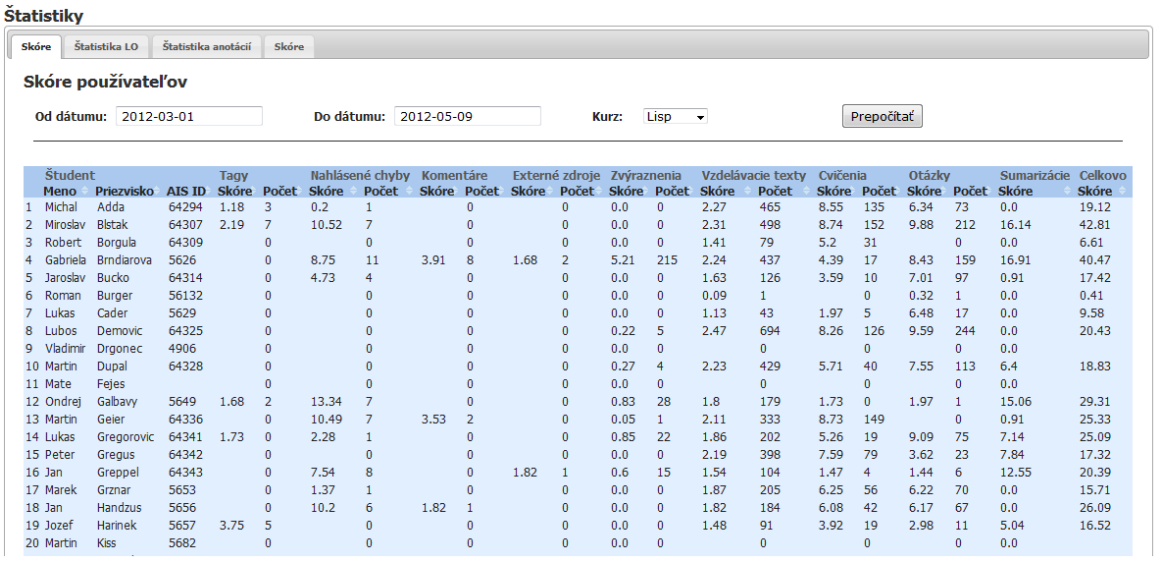
Na vytvorenie kurzu je potrebné v tejto časti administrácie nahrať ZIP súbor s informáciami o danom kurze (Obr. D-40). Tento súbor musí obsahovať adresár „uploaded_courses“ a v ňom XML súbory v adresároch pre jednotlivé kurzy. Adresáre musia byť pomenované názvom kurzu. Pri znovunahraní toho istého kurzu, sa nesmú meniť ich názvy.

Obr. D-40 Nahranie textov kurzov.

D.3.3 Štatistiky

Administrátor má k dispozícii aj štatistiky o jednotlivých používateľoch systému. Tie sa dajú filtrovať podľa kurzov a dátumov (Obr. D-41). Tiež je možné obmedziť počet zobrazených používateľov. Po načítaní zoznamu sa dá tento zoznam zoraďovať podľa jednotlivých atribútov.

V štatistikách sú zobrazené všetky aktivity používateľov, od vytvorenia anotácií, riešenia cvičení a otázok či hodnotenia sumarizácií. Pre každú položku je vypočítané skóre používateľa, takisto je zobrazené jeho celkové skóre.

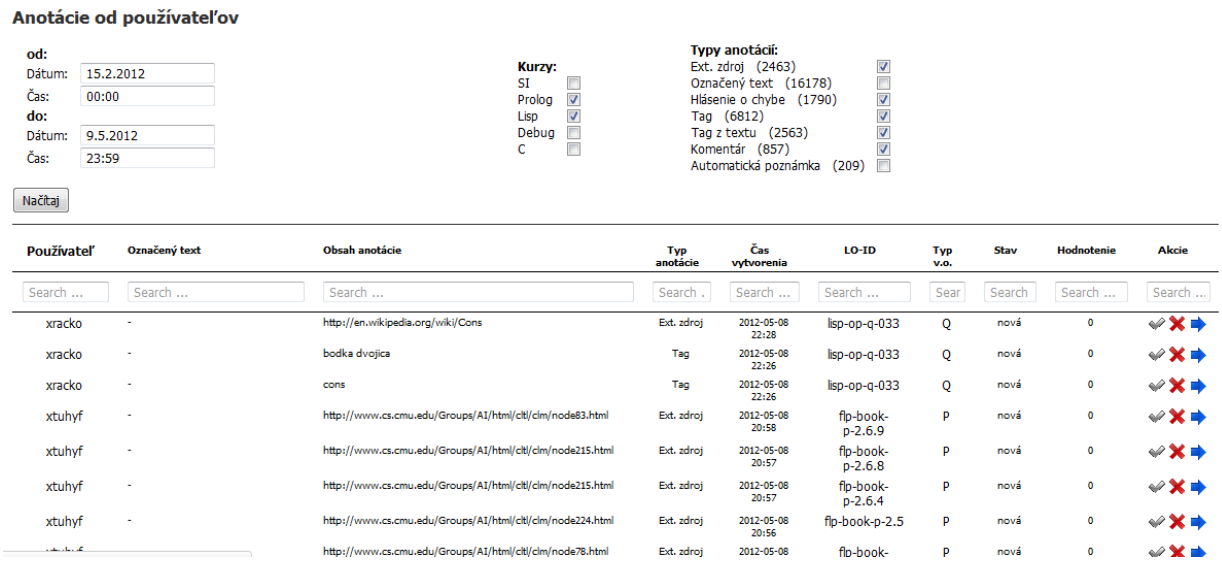


	Študent			Tagy		Nahlásené chyby		Komentáre		Externé zdroje		Zvýraznenia		Vzdelávacie texty		Cvičenia		Otázky		Sumarizácie	Celkovo
	Meno	Priezvisko	AIS ID	Skóre	Počet	Skóre	Počet	Skóre	Počet	Skóre	Počet	Skóre	Počet	Skóre	Počet	Skóre	Počet	Skóre	Počet	Skóre	Skóre
1	Michal	Adda	64294	1.18	3	0.2	1		0		0	0.0	0	2.27	465	8.55	135	6.34	73	0.0	19.12
2	Miroslav	Blstak	64307	2.19	7	10.52	7		0		0	0.0	0	2.31	498	8.74	152	9.88	212	16.14	42.81
3	Robert	Borgula	64309		0		0		0		0	0.0	0	1.41	79	5.2	31		0	0.0	6.61
4	Gabriela	Brndiarova	5626		0	8.75	11	3.91	8	1.68	2	5.21	215	2.24	437	4.39	17	8.43	159	16.91	40.47
5	Jaroslav	Bucko	64314		0	4.73	4		0		0	0.0	0	1.63	126	3.59	10	7.01	97	0.91	17.42
6	Roman	Burger	56132		0		0		0		0	0.0	0	0.09	1		0	0.32	1	0.0	0.41
7	Lukas	Cader	5629		0		0		0		0	0.0	0	1.13	43	1.97	5	6.48	17	0.0	9.58
8	Lubos	Demovic	64325		0		0		0		0	0.22	5	2.47	694	8.26	126	9.59	244	0.0	20.43
9	Vladimir	Drgonec	4906		0		0		0		0	0.0	0		0		0		0	0.0	
10	Martin	Dupal	64328		0		0		0		0	0.27	4	2.23	429	5.71	40	7.55	113	6.4	18.83
11	Mate	Fejes			0		0		0		0	0.0	0		0		0		0	0.0	
12	Ondrej	Galbavy	5649	1.68	2	13.34	7		0		0	0.83	28	1.8	179	1.73	0	1.97	1	15.06	29.31
13	Martin	Geier	64336		0	10.49	7	3.53	2		0	0.05	1	2.11	333	8.73	149		0	0.91	25.33
14	Lukas	Gregorovic	64341	1.73	0	2.28	1		0		0	0.85	22	1.86	202	5.26	19	9.09	75	7.14	25.09
15	Peter	Gregus	64342		0		0		0		0	0.0	0	2.19	398	7.59	79	3.62	23	7.84	17.32
16	Jan	Greppel	64343		0	7.54	8		0	1.82	1	0.6	15	1.54	104	1.47	4	1.44	6	12.55	20.39
17	Marek	Grznar	5653		0	1.37	1		0		0	0.0	0	1.87	205	6.25	56	6.22	70	0.0	15.71
18	Jan	Handzus	5656		0	10.2	6	1.82	1		0	0.0	0	1.82	184	6.08	42	6.17	67	0.0	26.09
19	Jozef	Harinek	5657	3.75	5		0		0		0	0.0	0	1.48	91	3.92	19	2.98	11	5.04	16.52
20	Martin	Kiss	5682		0		0		0		0	0.0	0		0		0		0	0.0	

Obr. D-41 Štatistiky používateľov.

D.3.4 Poznámky

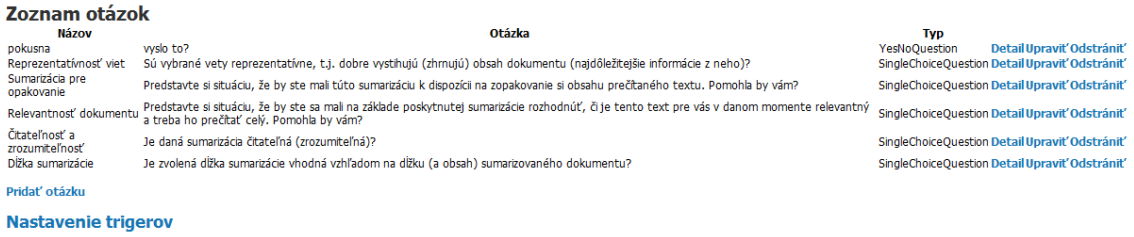
V časti poznámky si administrátor má možnosť prezerať všetky anotácie vytvorené používateľmi (Obr. D-42). Tieto anotácie je možné filtrovať podľa dátumu a kurzu, ako aj podľa typu anotácie.



Obr. D-42 Poznámky používateľov.

D.3.5 Otázky

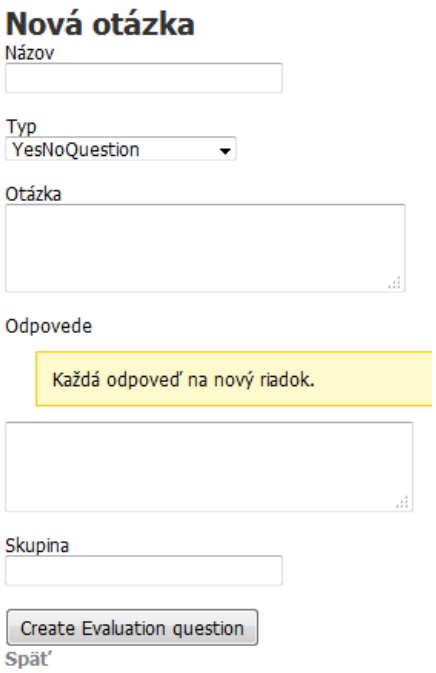
Otázky slúžia na získavanie priamej spätnej väzby od používateľov (pozri časť D.2.5). V administrátorskej časti môže administrátor vidieť, aké otázky sú v systéme, meniť ich, mazať a pridávať nové (Obr. D-43).



Obr. D-43 Zoznam adaptívnych otázok v systéme.

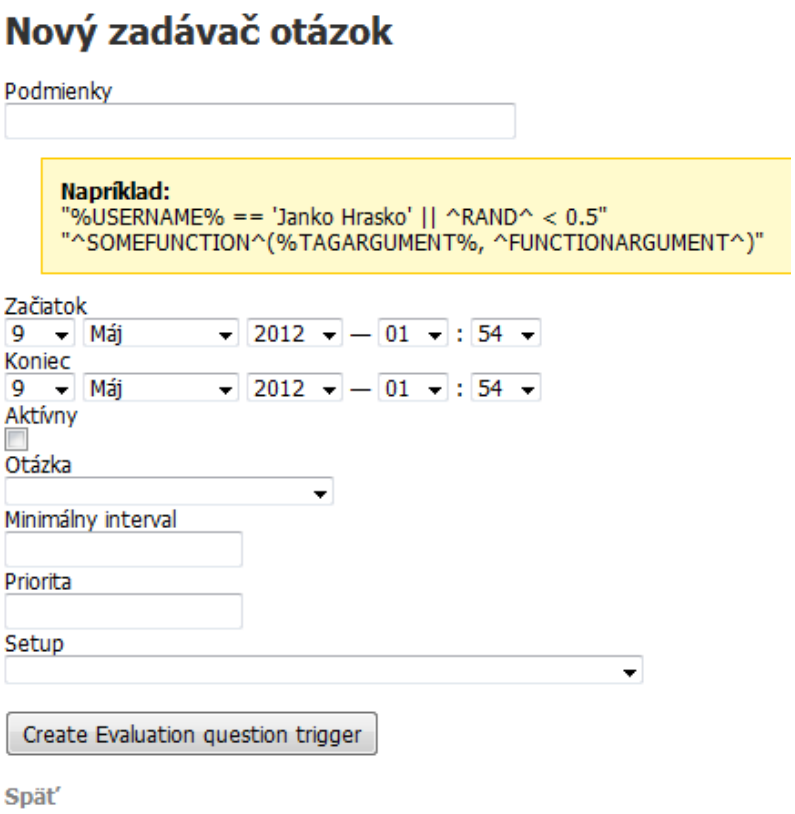
Pri tvorbe novej otázky (Obr. D-44) administrátor zadáva:

- *Názov otázky*
- *Typ* (otázka s odpoveďou Áno/Nie, otázka s jednou odpoveďou, otázka s výberom viacerých odpovedí, jednoduchá otázka s možnosťou krátkej slovnej odpovede, otázka s voľnou slovnou odpoveďou)
- *Samotný text otázky*
- *Možné odpovede na otázku* (v prípade, že ide o otázku s jednou odpoveďou, prípadne výberom viacerých odpovedí)
- *Skupinu*, pomocou ktorej vie zhlukovať súvisiace otázky



Obr. D-44 Rozhranie na pridanie otázky.

Logika výberu otázky je väčšinou implementovaná v kóde niektorého komponentu, avšak je možné pomocou grafického rozhrania vytvoriť tzv. „zadávače“, t.j. spúšťače otázok, v ktorých administrátor naskriptuje rôzne podmienky spustenia – zadania otázky (Obr. D-45).



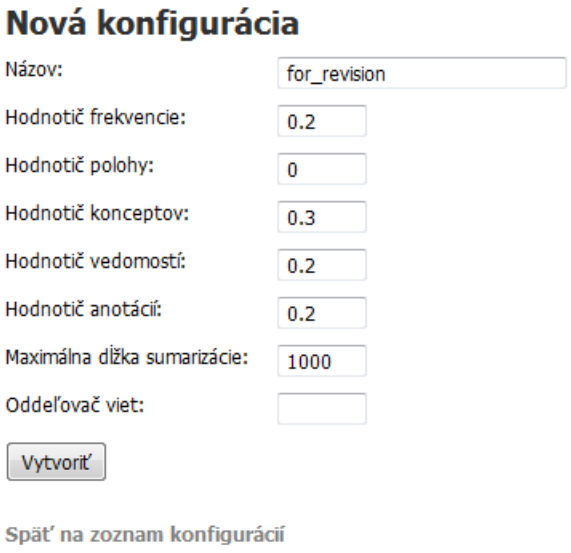
Obr. D-45 Rozhranie na tvorbu „zadávačov“ – spúšťačov otázok.

D.3.6 Sumarizátor

Administrátor môže spravovať konfigurácie sumarizátora – vytvárať nové, meniť a mazať existujúce (Obr. D-46).

Obr. D-46 Zoznam konfigurácií sumarizátora.

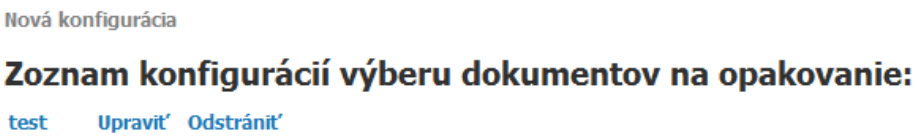
Pri vytváraní novej konfigurácie (Obr. D-47) volí jej názov a váhy jednotlivých hodnotičov. Okrem toho môže nastaviť aj ďalšie parametre ako maximálnu dĺžku či oddeľovač viet (čiarka, tri bodky a pod.).



Obr. D-47 Rozhranie na tvorbu konfigurácií sumarizátora.

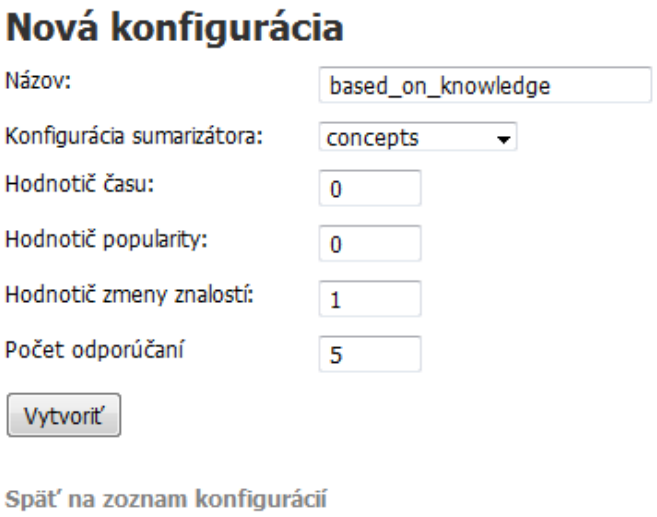
D.3.7 Opakovanie

Podobne ako pri konfiguráciách sumarizátora môže administrátor spravovať aj konfigurácie opakovania (Obr. D-48).



Obr. D-48 Zoznam konfigurácií opakovania.

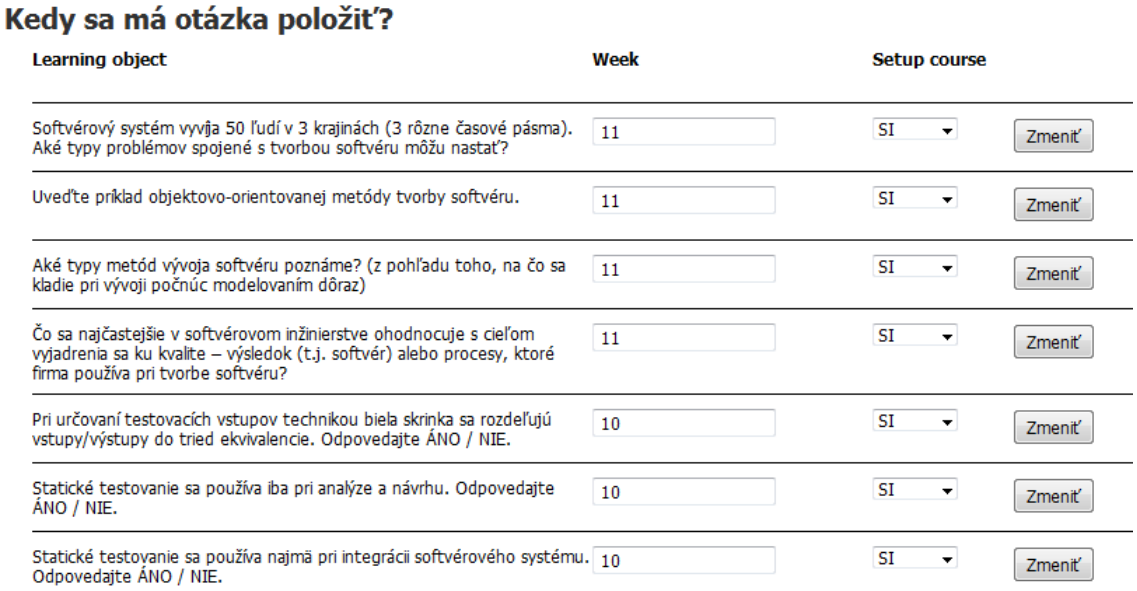
Pri tvorbe novej konfigurácie zadáva administrátor jej názov a opäť váhy hodnotičov, ktoré sa použijú pri ohodnocovaní výučbových textov pri ich výbere (odporúčaní) na opakovanie (Obr. D-49). Okrem toho je možné nastaviť počet odporúčaní, ktoré sa zobrazia používateľovi v zozname.



Obr. D-49 Rozhranie na tvorbu novej konfigurácie opakovania.

D.3.8 Validácia otázok

V tejto časti sa nastavujú parametre zobrazovania testových otázok s odpoveďami používateľov (Obr. D-50). Administrátor môže nastaviť, po ktorom týždni sa môžu jednotlivé otázky zobrazovať, a do ktorého kurzu patria. Týmto zabráni kladeniu otázok, na ktoré študent pravdepodobne nebude poznať odpovede, lebo sa to ešte v rámci kurzu nepreberalo.



Obr. D-50 Nastavenie validácie otázok.

Príloha E: Príspevok prijatý na TIR-DEXA 2012

Predkladaný príspevok bol prijatý na medzinárodný workshop o textovom vyhľadávaní informácií:

TIR 2012

9th International Workshop on Text-based Information Retrieval

ktorý sa organizuje v rámci konferencie o aplikáciách databázových a expertných systémov DEXA 2012. Vypracovali sme ho na základe príspevku, ktorý sme prezentovali na študentskej vedeckej konferencii IIT.SRC 2012, a ktorý bol ocenený Cenou dekana.

Personalized Text Summarization Based on Important Terms Identification

Róbert Móro, Mária Bieliková
Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies,
Slovak University of Technology, Ilkovičova 3, 842 47 Bratislava, Slovakia
{xmoror, maria.bielikova}@stuba.sk

***Abstract*—Automatic text summarization aims to address the information overload problem by extracting the most important information from a document, which can help a reader to decide, whether it is relevant or not. In this paper we propose a method of personalized text summarization, which improves conventional automatic text summarization methods by taking into account differences in readers' characteristics. We use annotations added by readers as one of the sources of personalization. We have experimentally evaluated the proposed method in the domain of learning, obtaining better summaries capable of extracting important concepts explained in the document, when considering the domain-relevant terms in the process of summarization.**

Keywords-automatic text summarization; personalization; annotations; domain-relevant terms

I. INTRODUCTION

Information overload is one of the most serious problems of the present-day web. There are various approaches addressing this problem; we are interested mainly in two: automatic text summarization and personalization.

Automatic text summarization aims to extract the most important information from the document, which can help readers (users) to decide whether it is relevant for them and they should read the whole text or not. However, the classical (generic) summarization methods summarize the content of the document without considering the differences in users, their needs or characteristics, i.e. their interests, goals or knowledge. On the other hand, the personalization aims to adapt the content presented to the individual user or the way she accesses the content based on her characteristics.

In this paper we propose a method of personalized summarization which extracts information from the document that is supposed to be the most important or interesting for a particular user. Because annotations (e.g. highlights) can indicate user's interest in the specific parts of the document [12], we use them as one of the sources of personalization. Our proposed method is sufficiently general to be used independently of the chosen domain; however we focus on summarization for revision in the domain of learning.

II. RELATED WORK

The first method of automatic text summarization was proposed by Luhn [6]; it was based on *term frequency*, which he used to compute the significance of terms. The idea was to extract from a document the most significant sentences, which contained the highest number of occurrences of significant terms. Edmundson [4] considered not only the frequency of terms, but also their location. Terms in the title, the first and the last paragraph and the first and the last sentence in each paragraph were assigned positive weights.

Gong and Liu [5] were first to use *latent semantic analysis (LSA)* for the text summarization. This method is capable of finding salient topics or concepts in the document and also their relative importance within the document. Each concept (topic) is represented in the summary by a sentence which captures it the best. However, Steinberger and Ježek [9] showed, that this approach fails to include into the summary sentences, which capture many concepts well, but have the highest score for none of them. They proposed a modification of the selection of sentences; sentences are selected based on their overall score computed as a combination of scores for each concept (topic).

All the methods mentioned so far summarize only the content of the document. However, there are many types of information which can indicate relevance of the sentences to extract, especially on the Web, e.g. user activity or user-added annotations (comments, highlights, tags etc.). Sun et al. [10] utilized clickthrough data which records how users find information through queries; if a user clicks on a link to a web page which is one of the results of her query, it indicates, that terms from the query describe the page and can be given more weight when summarizing the page. Park et al. [8] decided to summarize directly comments (descriptions) and tags which users add when they create a bookmark using social bookmarking service such as *Delicious*. The advantage of this approach is that it can summarize also documents with no or minimum text, but with other multimedia content. On the other hand, it depends on the number of added bookmarks and is unable to summarize documents with no bookmarks.

Methods of personalized summarization also use additional information. However, their result is not a generic summary which is the same for all the users, but a summary that is adapted (personalized) to the characteristics of a particular user. Díaz et al. [3] personalized summarization to mirror users' interests represented by a user model in the form of a vector of weighted keywords; disadvantage of this approach is that users have to manually insert keywords and weights into the model. Campana and Tombros [2] automatically built a user model from the sentences of the documents recently read by a user. Summary is constructed from the sentences which are the most similar to the most representative sentences from the user model. Zhang et al. [12] utilized user annotations in the form of highlights by extending the classical tf-

idf method. Results of their study suggest that this approach is useful for summarization personalization. They identify determining a subset of annotations suitable for summarization and including of collaborative annotating as open problems.

III. Personalized Text Summarization

We propose a method of personalized text summarization based on a method of latent semantic analysis [5][9] which consists of the following steps:

- Pre-processing
- Construction of personalized terms by sentences matrix
- Singular value decomposition (SVD)
- Sentences selection

A. Pre-processing

Pre-processing consists of three steps:

- machine translation
- tokenization and terms extraction
- breaking the text into sentences

We use machine translation of the document to a reference language (in our case English) in order to maximize our method's independency of the summarized document language. Because terms extraction and breaking the text to sentences which are also performed during the pre-processing are language-dependent, the use of machine translation enables us to provide respective algorithms only for one (reference) language and effectively cover a wide range of languages. Certainly, this can influence the quality of the resulting summaries; however, our experiments (using *Bing Translator*) show that the state-of-the-art machine translation gives reasonable results.

B. Construction of personalized terms by sentences matrix

We have identified the construction of a terms-sentences matrix representing the document as a step suitable for personalization of the summarization. In this step terms extracted from the document are assigned their respective weights. Our proposed weighting scheme extends the conventional weighting scheme based on tf-idf method by linear combination of multiple raters, which positively or negatively affect the weight of each term (see Fig. 1).

We formulate the weighting scheme as follows:

$$w(t_{ij}) = \sum_k \alpha_k R_k(t_{ij}), \tag{1}$$

where $w(t_{ij})$ is a weight of a term t_{ij} in the matrix and α_k is a linear coefficient of a rater R_k. The rater R_k is a function, which assigns each term from the extracted keywords set T its weight:

$$R_k : T \to \Re. \tag{2}$$

We have designed a set of raters which can be divided into two groups:

- *Generic raters:* terms frequency rater, terms location rater and domain-relevant terms rater
- *Personalized raters:* knowledge rater and annotations rater

Generic raters take into account content of the document and some additional information to adapt summarization regardless of a particular user. On the other hand, personalized raters consider information about the specific user and her characteristics. Selection and combination of the raters (the values of their linear coefficients) depends on a specific summarization use case and the types of additional information which we have at our disposal.

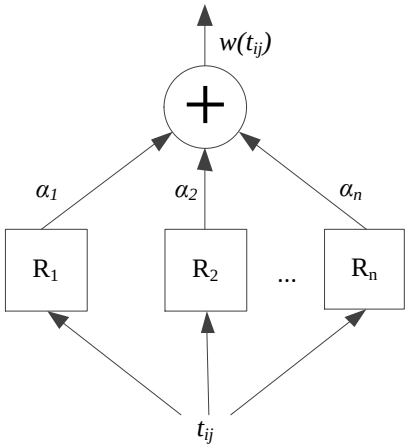


Figure 1. Term weighting by combination of raters.

Terms frequency rater and *terms location rater* are the two basic generic raters, the design of which was inspired by Luhn [6] and Edmundson [4]. The former assigns the weights based on tf-idf method, the latter based on the location of terms; terms in the title and the first and the last sentence of a document are given positive weights.

Because we focus on the domain of learning, we have identified three main sources of the summarization personalization and adaptation suitable for the chosen domain:

- Domain conceptualization in the form of the domain-relevant terms
- Knowledge of the users
- Annotations added by users, i.e. highlights or tags

We have designed a *domain-relevant terms rater* to utilize information contained in a domain model of an adaptive system. Domain models are usually constructed manually by domain experts by capturing their knowledge of the domain in the form of important concepts (domain-relevant terms) and relationships among them which makes them valuable sources of information for summarization adaptation (if they are available).

Let C_d be a set of concepts associated with a document d; each concept in the set is represented by ordered pair (t_i, w_i), where t_i is a domain-relevant term i and weight w_i represents measure of association between the document d and the concept i. We formulate the domain-relevant terms rater as follows:

$$\begin{aligned} w(t_{ij}) &= w_i \quad \text{if } t_i \in S_j \cap C_d \\ w(t_{ij}) &= 0 \quad \text{else,} \end{aligned} \tag{3}$$

where S_j is a set of terms contained in the sentence j.

Educational systems usually use overlay user models to record level of knowledge of each concept from the domain for each particular user. We have designed *knowledge rater* as a personalized version of the previous one. It uses information captured in the user model; so instead of w_i representing the measure of association between the document d and the concept i, we now use k_{iu} reflecting a level of knowledge of the concept i by a user u. This way, concepts which are better understood by the user are given more weight, which is especially useful in the knowledge revision scenario.

Although learning systems usually assume that modeled knowledge only grows in time, users in fact do forget a part of their acquired knowledge [1]. The knowledge revision represents a means of re-acquiring the forgotten knowledge. We believe that summarization is suitable for this scenario, because it can help users to remind them of the important concepts explained in the documents.

Annotations rater takes into account fragments of the text highlighted by a particular user. First we construct a set S_H of all the sentences, fragments of which were highlighted:

$$S_H = \{S_i \mid S_i \cap H \neq \emptyset\}, \tag{4}$$

where S_i is i-th sentence of the document and H is a set of all the highlights made by the particular user in the document extended by the most popular highlights made by all the users (the most highlighted fragments of the document). We compute for each text fragment h_i the number of times it was highlighted $|h_i|$ and consider as popular those, for which the following condition stands true:

$$|h_i| \geq \frac{\max_j |h_j|}{2}, \tag{5}$$

Lastly, we assign positive weights to those terms of the document, which are located in the sentences in the set S_H:

$$\begin{aligned} w(t_{ij}) &= 1 \quad \text{if } S_j \in S_H \\ w(t_{ij}) &= 0 \quad \text{else,} \end{aligned} \tag{6}$$

where S_j is j-th sentence of the document.

C. SVD and Sentences Selection

After the terms-sentences matrix **A** is constructed, it enters singular value decomposition [5]:

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathbf{T}}, \tag{7}$$

where $\mathbf{U} = (u_{ij})$ is a matrix, the columns of which represent left singular vectors and rows terms of the document, $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ is a diagonal matrix with diagonal elements representing non-negative singular values in descending order and $\mathbf{V} = (v_{ij})$ is a matrix, the columns of which represent right singular vectors and its rows represent sentences of the document.

The final step is a selection of sentences; we select sentences with the highest score computed by a method proposed in [9] (sentences are selected from the original, not translated document):

$$s_k = \sqrt{\sum_{i=1}^{n} v_{ki}^2 \sigma_i^2}, \tag{8}$$

where s_k is a score of k-th sentence, v_{ki} is a value from matrix **V** that measures how well is concept i represented by sentence k and σ_i is a singular value that represents relative relevance of the concept i in the document; n is a number of dimensions and it is a parameter of the method. The length of the generated summary is a parameter of our method as well.

IV. Evaluation

We have realized our proposed method independently of the domain using web services. We have experimented in the domain of learning choosing the knowledge revision as our specific use case. Our dataset has consisted of the educational materials from the *Functional and Logic Programming* course in the educational system ALEF.

ALEF (Adaptive Learning Framework) [11] is an application framework which merges concepts of traditional web-based educational systems with the principles of the Web 2.0, i.e. it brings several interactive features such as tagging, commenting, collaborative task solving etc. We have integrated our implementation of the summarizer with ALEF (see Fig. 2).

In our experiment, we have focused on evaluation and comparison of the two variants of summaries:

- generic summarization and
- summarization considering the domain-relevant terms identified by a domain expert.

We have generated both variants for each document in our dataset; because we focus on the summarization for revision, we have chosen the length of the generated summaries to be approximately one third of the document length.

We have asked the Functional and logic programming course students to evaluate quality of the generated summaries. They have been supposed to read educational texts in ALEF and rate presented summaries on the five-point scale without knowing what variant of summary is presented to them. The summaries were placed underneath the text, so that students would read the document first and its summary second. This has been setup specifically for our experiment; in the real usage the summary is positioned above the text.

After each summary rating, the students have been asked a follow-up question to further evaluate the summary quality. We have been interested whether the sentences selected for the summary are representative, whether the summary is suitable for the revision or whether it could help them to decide the document relevance. We have also inquired whether the length of

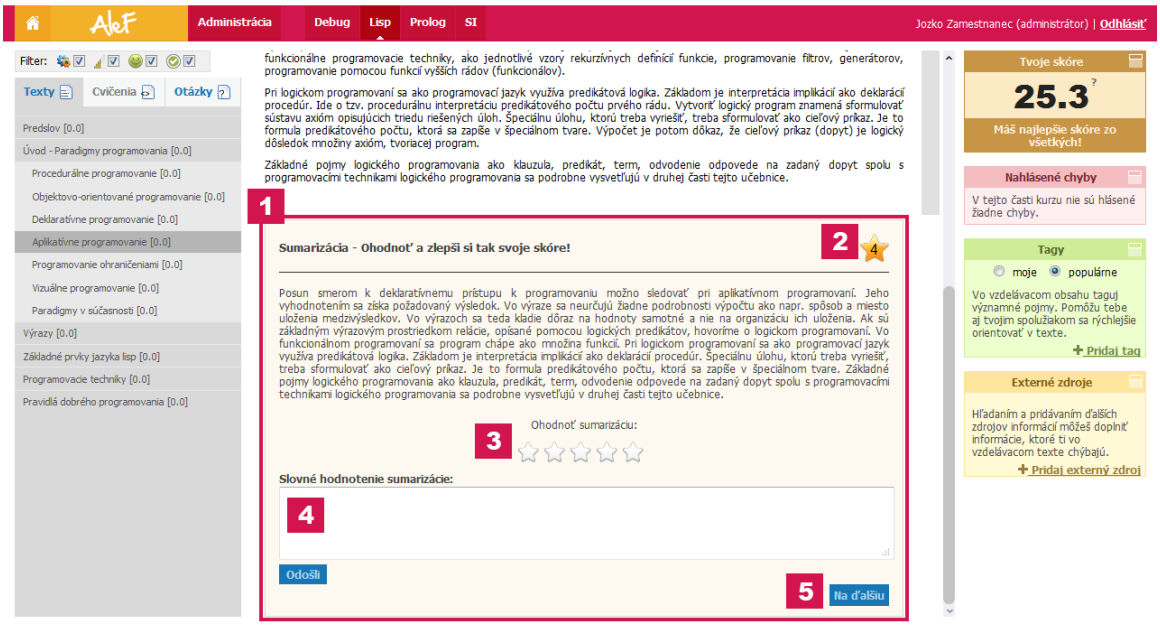


Figure 2. Example screenshot of ALEF (in Slovak) with the integrated summarizer (1 – highlighted by the border). Current user rating is shown in the right top corner (2). Users rated summaries on the five-point scale using stars (3); after each rating they were automatically asked a follow-up question. They could have also added feedback in the form of free text (4) or navigate themselves to the next summary by clicking the *Next* button in the right bottom corner (5).

the summary is suitable given the length and content of the document and if it is readable and comprehensible.

Furthermore, we have chosen a control expert group to compare their summary evaluation to that of the other students. The group has consisted of seven domain experts. In contrast to the other participants, they have been presented both summary variants (in random order) for each educational text in order to decide which variant is better or whether they are equal.

We have gathered summaries for 79 educational texts (explanation learning objects), 278 summary ratings and 154 summary variants comparisons from experts. Moreover, students have answered 275 follow-up questions.

The second variant (summarization considering the domain-relevant terms) has on average gained higher score (3.79) on the five-point scale compared to the first variant (generic summarization), which has scored 3.54 on average (see Tab. 1).

TABLE I. SUMMARY VARIANTS RATINGS

Statistic	Generic	Domain-relevant terms
No. of ratings	143	135
Mean	3.538	3.793
Variance (n-1)	1.518	1.419

We have also computed average score for each summary variant for each document. The second variant has scored more in comparison to the first one in 48% of the cases, the same in 11% and less in 41%. The comparison of summaries by the experts has given us similar results. The second variant has been evaluated as better in 49% of the cases, as equal in 20% and worse in 31% (see Fig. 3). Thus, our results suggest that considering the domain-relevant terms during the summarization process leads to better summaries in comparison to the baseline generic variant.

Lastly, we have evaluated the students' answers to the follow-up questions. They suggest that our method in general has managed to choose representative sentences and the summaries could be in many cases used for revision or to help the students to decide whether the summarized document is relevant. Furthermore, the answers to the follow-up questions show that the notion of summary quality is subjective. This confirms us in our belief that summarization personalization is useful and we can get significantly better results when we take students' annotations into consideration.

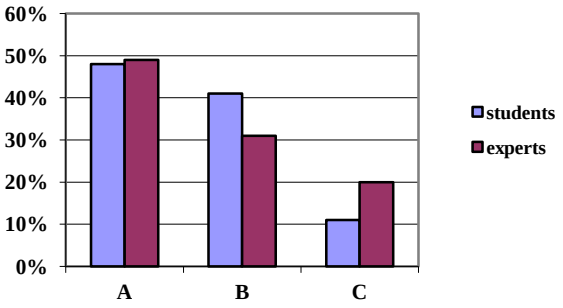


Figure 3. Comparison of summary variants, where *A* means that summary considering the domain-relevant terms was evaluated as better, *B* that generic summary was better and *C* that they were equal.

V. CONCLUSIONS

We have proposed a method of personalized summarization. Our contribution lies in the proposal of

- the specific raters that take into account terms relevant for the domain or the level of knowledge of an individual user
- the method of the raters' combination which allows considering various parameters or context of the summarization.

Even though our approach is domain independent, it allows us to identify the sources of summarization personalization specific for the chosen domain and to adapt the method for the particular scenario (in our case students' knowledge revision).

We work towards providing the users with better summaries reflecting their particular needs by taking into account also their annotations that indicate users' interest in the fragments of a document. By considering not only a user's personal annotations, but the most popular ones as well, we can potentially deal with a situation when the user has not yet added any annotations to the document and also utilize the wisdom of the crowd.

We have evaluated our approach in the domain of learning in the knowledge revision scenario. In the first phase of the evaluation, we have focused on the comparison of the two summary variants: the generic summarization and the summarization considering the domain-relevant terms. Our experimental results suggest that using the domain-relevant terms in the process of summarization can help selecting representative sentences capable of summarizing the document for revision.

VI. Future Work

As our future work, we plan more experiments with the two raters: the knowledge rater and the annotations rater. We believe that considering the knowledge as well as users' annotations in the summarization process will lead to summaries better adapted for a particular user's needs.

Considering the user's knowledge of the concepts described in the document could be especially useful for knowledge revision. Also, text highlighting is common approach of learning when students highlight the most important fragments of the document and return to them later during revision.

Selecting the appropriate documents for revision is another aspect that we have to take into account; for this purpose, we have proposed a method for personalized selection of documents for revision which considers various characteristics, e.g. recent changes of a user's knowledge supporting concepts, the knowledge of which the user has recently gained or, on the contrary, lost.

In our evaluation, we have experimented with the summaries that have considered the domain-relevant terms; these terms have been identified by domain experts and represented in the domain model of an adaptive educational system.

However, domain modeling is in general hard and time consuming task. As we have shown in [7], it is possible and useful to utilize folksonomies (users-added tags) for this purpose. Thus, we could replace the domain-relevant terms used in our evaluation by the tags associated with the summarized document and compute their weights based on their popularity. Therefore, we plan to experiment with utilizing the tags added by the users for the summarization personalization in the future as well.

Acknowledgment

This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11 and it is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

References

[1] M. Bieliková, P. Nagy, "Considering Human Memory Aspects for Adaptation and Its Realization in AHA!," in Technology Enhanced Learning (EC-TEL 06), W. Nejdl, K. Tochtermann, Eds. Springer Verlag, LNCS, vol. 4227, pp. 8–20.

[2] M. Campana, A. Tombros, "Incremental personalised summarisation with novelty detection," Proc. 8th Int. Conf. Flexible Query Answering Systems (FQAS 09), LNCS, vol. 5822, Springer, Berlin, 2009, pp. 641-652.

[3] A. Diaz, P. Gervàs, A. Garzìa, "Evaluation of a system for personalized summarization of web contents," User Modeling 2005, LNCS, vol. 3538, Springer, Berlin, 2005, pp. 453–462.

[4] H.P. Edmundson, "New methods in automatic extracting," J. of the ACM, vol. 16, no. 2, 1969, pp. 264–285.

[5] X. Gong, X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," Proc. 24th Int. ACM SIGIR Conf. Research and Development in Inf. Retrieval (SIGIR 01), ACM Press, 2001, pp. 19-25.

[6] H.P. Luhn, "The automatic creation of literature abstracts," IBM J. of Research Development, vol. 2, no. 2, 1958, pp. 159–165.

[7] R. Móro, I. Srba, M. Unčík, M. Šimko, M. Bieliková, "Towards Collaborative Metadata Enrichment for Adaptive Web-based Learning," Proc. Int. Conf. Web Intelligence and Intelligent Agent Technology (WI-IAT 11), IEEE Computer Society, 2011, pp. 106-109.

[8] J. Park, T. Fukuhara, I. Ohmukai, H. Takeda, S. Lee, "Web Content summarization using social bookmarks: A new approach for social summarization," Proc 10th ACM workshop Web information and data management (WIDM 08), ACM Press, 2008, pp. 103-110.

[9] J. Steinberger, K. Ježek, "Text summarization and singular value decomposition," Proc. Advances in Information Systems (ADVIS 04), LNCS, vol. 3261, Springer, Berlin, 2005, pp. 245-254.

[10] J. Sun et al., "Web-page summarization using clickthrough data," Proc. 28th Int. ACM SIGIR Conf. Research and Development in Inf. Retrieval (SIGIR 05), ACM Press, 2005, pp. 194–201.

[11] M. Šimko, M. Barla, M. Bieliková, "ALEF: A framework for adaptive web-based learning 2.0," IFIP Advances in Information and Communication Technology (KCKS 10), Springer, Berlin, vol. 324, 2010, pp. 367-378.

[12] H. Zhang, Z.C. Ma, Q. Cai, "A study for documents summarization based on personal annotation," Proc. HLT-NAACL Workshop Text summarization, Association for Computational Linguistics, 2003, pp. 41–48.

Príloha F: Obsah elektronického média

Prílohou tejto práce je aj CD nosič s nasledovným obsahom:

`alef/`	zdrojové kódy výučbového systému ALEF
`alef-summarization/`	zdrojové kódy rozširujúce systém ALEF o sumarizáciu
`doc/`	elektronická verzia diplomového projektu vo formáte PDF
`linalg/`	knižnica metód lineárnej algebry pre prácu s maticami potrebná pre beh sumarizátora
`papers/`	príspevok prijatý na medzinárodný workshop TIR 2012
	príspevok prijatý na konferenciu IIT.SRC 2012
	poster prezentovaný na konferencii IIT.SRC 2012
`summarizer/`	zdrojové kódy sumarizátora