Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

Пребарување на медицински документи со мултимодални податоци

Докторска дисертација

Иван Китановски

Ментор:
Проф. д-р Сузана Лошковска

Скопје, Јули 2017

Комисија

Вон. проф. д-р Ивица Димитровски, претседател
Факултет за информатички науки и компјутерско инженерство
Универзитет Св. „Кирил и Методиј“, Скопје, Македонија

Проф. д-р Сузана Лошковска, ментор
Факултет за информатички науки и компјутерско инженерство
Универзитет Св. „Кирил и Методиј“, Скопје, Македонија

Проф. д-р Дејан Ѓорѓевиќ, член
Факултет за информатички науки и компјутерско инженерство
Универзитет Св. „Кирил и Методиј“, Скопје, Македонија

Вон. проф. д-р Ѓорѓи Мацаров, член
Факултет за информатички науки и компјутерско инженерство
Универзитет Св. „Кирил и Методиј“, Скопје, Македонија

Проф. д-р Цвета Мартиновска-Банде, надворешен член
Факултетот за информатика
Универзитетот „Гоце Делчев“, Штип, Македонија

Посветено на моето семејство

Пребарување на медицински документи со мултимодални податоци

Докторска дисертација

Иван Китановски

Апстракт

Во оваа докторска дисертација е истражувана областа на пребарување на медицински документи. Истражувањата резултираа со комплексен систем за пребарување на медицински документи во кој се имплементирани истражуваните методи. Имплементираниот систем е поделен на повеќе подсистеми и тоа: подсистем за класификација на медицински слики според модалитет, подсистем за пребарување на медицински слики и повеќе подсистеми за пребарување на медицински трудови.

Подсистемот за класификација на медицински слики според модалитет на автоматски начин ги означува сликите со соодветниот модалитет на основа на нивните визуелни и/или текстуални карактеристики. Во овој подсистем се имплементирани повеќе дескриптори за извлекување на визуелните карактеристики на сликите, како LBP, FCTH, CEDD, SIFT и OSIFT. Текстуалните карактеристики се оформени на основа на трудовите во кои се појавуваат сликите. Класификацијата е изведена со помош на машини со носечки вектори. Евалуацијата на овој подсистем прикажува дека OSIFT дескрипторот има најдобри перформанси во однос прецизноста на класифицирањето, споредено со останатите дескриптори, но комбинацијата од сите визуелни дескриптори прикажува уште подобри перформанси. Конечно, комбинирање на сите визуелни дескриптори и текстуалните карактеристики придонесува за најдобри резултати кои воедно се врвни за множества кои се користат за евалуација.

Подсистемот за пребарување на медицински слики функционира на основа на визуелните и/или текстуалните карактеристики на сликите. Делот од подсистемот за пребарување на медицинските слики кој е текстуално базиран ги користи текстуалните податоци на трудовите во кои се појавуваат сликите и креира нивна репрезентација со помош на векторскиот модел. Пребарувањето во овој дел на подсистемот е помогнато од метод на проширување на прашањата со псевдо-релевантна повратна врска. Делот од подсистемот за пребарување кој е содржински (визуелно) базиран користи RGB хистограми, енкодирани во Фишер вектори за опишување на сликите. Овде е применет метод на квантизација на карактеристиките, со цел да се овозможи скалабилност и брзи пребарувања на големи колекции на слики. Подсистемот за пребарување функционира во комбинација со подсистемот за класификација на медицински слики според модалитет на тој начин што инцијално вратените слики се класифицираат и повторно се рангираат на основа на нивниот модалитет во однос на поставеното прашање. Овој

подсистем е евалуиран на повеќе стандардизирани бази на податоци и прикажува врвни перформанси.

Подсистемите поврзани со пребарување на медицински трудови функционираат на основа на текстуалните податоци во нив. Заедничко за нив е што креираат текстуална репрезентација на медицинските трудови преку збогатување на содржината со концепти извлечени од медицински и/или генерички бази на знаење. Прашањата кои ги поставува корисникот, во фазата на пребарување, се модифицираат со помош на надворешни алатки за извлекување на медицинско знаење и/или со псевдо-релевантна повратна врска со цел допрецизирање на истите. Сите подсистеми за пребарување на медицински трудови се евалуирани и прикажуваат резултати кои се објавени во повеќе публикации.

Medical document retrieval using multimodal data

PhD thesis

Ivan Kitanovski

Abstract

In this thesis we have researched the field of medical documents retrieval. Our work in the field resulted in a complex medical document retrieval system which contains the proposed methods. The implemented system is divided into multiple subsystems: subsystem for modality classification of medical images, subsystem for medical image retrieval and multiple subsystems for medical articles retrieval.

The subsystem for modality classification of medical images annotates the images with the appropriate modality based on their visual and/or textual characteristics. This subsystem contains multiple descriptors for feature extraction from the images such as LBP, FCTH, CEDD, SIFT and OSIFT. The textual characteristics are formed based on the articles where the images appear. The classification is done using support vector machines. The evaluation of the subsystem shows that the OSIFT descriptor provides the best performance compared to the other descriptors, but combining all visual descriptors provides even better results. Finally, combining the visual descriptors and the textual characteristics provide the best overall performance, which are state-of-the-art for the datasets we used for evaluation.

The subsystem for medical image retrieval works based on the visual (content) and/or textual characteristics of the images. The text-based part of the medical image retrieval subsystem uses the textual data of the articles where the images appear and creates a representation using the vector model. The retrieval of this part is boosted by query expansion using pseudo-relevance feedback. The content-based part of the subsystem uses RGB histograms which are encoded into Fisher vectors to describe the images. The method of product quantization is applied here, so that this part of the subsystem is scalable and allows fast retrieval over large image collections. The medical image retrieval subsystem works in combination with the modality classification subsystem, in such a way that the initially retrieved images are classified and are re-ranked based on their relation with the submitted query. This subsystem is evaluated over several standardized datasets and provides state-of-the-art results.

The subsystems related to medical articles retrieval perform the retrieval based on the textual data in them. The common thing for the subsystems is that they create a textual representation of the medical articles, by first enriching them using external medical or generic knowledge databases. The queries provided by the users are modified, in the retrieval phase, using external medical knowledge tools and/or pseudo-relevance feedback. All subsystems for medical articles retrieval are evaluated and provide good results which have been published in multiple papers.

Благодарност

Огромна благодарност за мојот ментор, проф. д-р Сузана Лошковска, за нејзината помош и целокупна поддршка, како во активностите поврзани со изработката на докторската дисертација, така и во секојдневните активности. Особено што ме поттикнуваше да ги поставувам вистинските прашања во текот на научно-истражувачката работа. Соработката со неа за мене претставува голема чест.

На вонр. проф. д-р. Ивица Димитровски му должам посебна благодарност за несебичното пренесување на искуството и стручната помош што постојано ми ја даваше во текот на целиот процес. Особено за насоките кои ми ги даваше во текот на практичниот дел од докторската дисертација. Тој никогаш не ги штедеше своето време и енергија за да понуди искрена помош.

Благодарност би сакал да искажам и на останатите членови од комисијата за корисните совети и насоките кои ми ги дадоа во изработката на докторската дисертација. Дополнително, сакам да искажам благодарност и до сите колеги за постојаната поддршка и поттик, како и за заедничката соработка.

На сестра ми Соња и нејзината фамилија кои секогаш се тука за мене им должам посебна благодарност. Вечно ќе им бидам благодарен на моите родители, Лидија и Кире, за нивната безрезервна поддршка и љубов. Тие се најзаслужни за тоа што сум денес.

На крај, бескрајно благодарен на мојата сопруга, Зорица, за нејзината љубов и разбирање. Само со нејзината поддршка имам мотивација и енергија да ги постигнам сите мои цели.

Содржина

Листа на слики

Листа на табели

Глава 1

Вовед

1.1 Тема на докторската дисертација

Во оваа докторска дисертација истражувана е областа на пребарување на медицински документи со помош на мултимодални податоци. Конкретно истражувани се полињата на пребарување на генерички медицински слики со мултимодални податоци и пребарување на медицински случаи (или медицински трудови) со помош на методи за проширување на прашањата и генерички бази на знаење. Во рамките на докторската дисертација е избработен систем за пребарување на генерички медицински слики и трудови во кој се имплементирани техники добиени како резултат на истражувањата.

Во делот на системот за пребарување на медицински слики имплементирани се методи за текстуално базирано, содржински базирано пребарување, како и класификација на слики. Текстуално базираното пребарување формира текстуални репрезентации за медицинските слики на основа на контекстот во кои се појавуваат, а може да користи различни модели за утежнување во фазата на пребарување. Содржински базираното пребарување ја користи визуелната содржина на сликите со цел да изгради ефикасна и компактна репрезентација и да овозможи пребарување во реално време над големи колекции на слики. Класификацијата на медицинските слики автоматски ги класифицира сликите на основа на нивниот модалитет со помош на алгоритми за машинско учење, користејќи текстуални и визуелни податоци. Оваа информација се користи во фазата за пребарување.

Делот од системот за пребарување на медицински слики е евалуиран на стандардизирано множество за евалуација со одредено множество на слики, предефинирани тест прашања и одговори. Целта на евалуација е да се утврди релевантноста на резултатите што ги враќа системот за дадено прашање. Релеваноста е изразена преку прецизноста на вратените слики. Експериментите извршени во овој дел од истражувањата се објавени во неколку публикации и покажуваат *state-of-the-art* резултати во ова поле.

Во рамки на делот за пребарување на медицински трудови имплементирани се техники за текстуално базирано пребарување. Техниките креираат соодветна репрезентација на трудовите преку збогатување на содржината со концепти извлечени од бази

на знаење. Во фазата на пребарување се користат различни модели за утежнување. Дополнително, имплементирани се техники за модифицирање на прашањата кои ги поставува корисникот со цел на автоматски начин да се допрецизира барањето. Во тој поглед имплементирани се повеќе техники за проширување на прашањата со помош на генерички бази на знаење и методи на релевентна повратна врска.

Делот од системот посветен на пребарување на медицински трудови е евалуаиран на специјализирано множество за евалуација кое содржи поголем број медицински трудови (случаи) и предефинирани прашања и одговори. Евалуацијата се однесува на прецизноста на резултатите што ги враќа системот. Експериментите извршени на овој дел од системот покажуваат значајни резултати и истите се објавени во бројни публикации.

1.2 Придобивки од докторската дисертација

Целта на докторската дисертација е имплементација на систем за пребарување на медицински слики и трудови. За таа цел ги имплементиравме компонентите за системот кои треба да ги овозможат различните операции вклучени во процесот на пребарување; како текстуално и содржински базирано пребарување на слики, класификација на слики според модалитет, текстуално базирано пребарување на трудови и проширување на прашања за текстуално пребарување на трудови. Клучните придобивки од докторската дисертација може да се сумираат на следниот начин:

- **Текстуално базирано пребарување на слики:** Имплементиран е подсистем за текстуално базирано пребарување на слики. Подсистемот го обработува текстот асоциран со сликите и трудовите во кои се појавуваат со цел формирање на компактни текстуални репрезентации на сликите. Текстуалните репрезентации се индексираат со едно изминување (анг. *single pass indexing*), а пребарувањето се прави со помош на модели за утежнување. Овој дел од системот е евалуиран на повеќе бази на податоци и добиените резултати покажуваат дека истиот има state-of-the-art перформанси.
- **Содржински базирано пребарување на слики:** Имплементиран е подсистем за содржински базирано пребарување на слики. Подсистемот овозможува анализирање на содржината на сликите со помош на дескриптори базирани на Fisher вектори добиени од RGB хистограми на сликата. Во фазата на пребарување се употребува метод на кодирање на векторите со цел побрзо пребарување. Главниот предизвик во овој дел од системот беше да се постигне пребарување во реално време без загуба на перформанси.
- **Класификација на слики според модалитет:** Развиен е подсистем за автоматска класификација на слики според нивниот модалитет базиран на машини со носечки вектори (анг. *Support Vector Machines - SVM*). Сликите се опишани со opponentSIFT дескрипторот, кој овозможува ефикасна репрезентација на слики во контекст на разграничување на модалитетот на сликите. Подсистемот е евалу-

иран на различни бази на податоци за класификација на слики според модалитет и истиот покажува state-of-the-art перформанси во тоа поле.

- **Текстуално базирано пребарување на трудови:** Подсистемите за пребарување на медицински трудови функционираат на основа на текстуалната содржина на трудовите. Главниот дел од овие подсистеми е механизмот за проширување на прашањата со псевдо-релевантна повратна врска, проширување со генерички бази на знаење, проширување со медицински бази на знаење. Подсистемите прикажуваат добри резултати над множествата кои ги употребивме за евалуација на истиот.

Важно е да се напомене дека првите три подсистеми функционираат независно и како целина со цел овозможување на мултимодално пребарување на медицински слики и на тој начин се постигнуваат најдобрите пријавени резултати на базите за евалуација искористени за тестирање на системот.

1.3 Структура на докторската дисертација

Докторската дисертација е организирана на следниот начин:

- **Втора глава:** Оваа глава ја дефинира проблематиката на докторската дисертација. Презентиран е општиот проблем на пребарување на информации, причините за појавувањето на системите за пребарување и општите пристапи кон пребарувањето. Во вториот дел од оваа глава е презентиран проблемот на пребарување информации во контекст на медицината, опишана е улогата на системите за пребарување во медицината, како и специфичните предизвици во тоа поле. Во последниот дел од главата се презентирани конкретните проблеми кои се цел на истражување на докторската дисертација, односно дефинирани се проблемите и предзивиците при пребарувањето на медицински слики и трудови.
- **Трета глава:** Во оваа глава е направен преглед на литературата. Претставени се методите за пребарување на медицински трудови со осврт на нивните недостатоци и посебно се анализирани начините за проширување на прашањата при пребарување на медицински трудови како најефикасни методи за пребарување во тој контекст. Дефиниран е проблемот на пребарување на генерички медицински слики. Даден е опширен преглед на постоечките техники и предизвици кај текстуално и содржински базираното пребарување на генерички медицински слики со посебен осврт на нивните недостатоци, отворените истражувачки прашања и предизвици. Анализирани се методи за класификација на медицински слики според модалитет. Прикажани се постоечките процеси и нивните предности и недостатоци. На крајот од главата се презентирани постоечки техники за пребарување на слики од медицински трудови, при што посебен акцент е ставен на мултимодалните пристапи за пребарување.

- **Четврта глава:** Методологијата на работа е претставена во оваа глава. Опишан е општиот процес на пребарување на информации. Презентирани се моделите за пребарување на информации и клучните техники при пребарување на информации. Презентирани се методите за извлекување на карактеристики од податоците, поделени според видот на карактеристиките т.е. текстуални и визуелни карактеристики и посебен акцент е ставен на модалитетот на сликите во поддршка на пребарувањето. Дадени се платформите за пребарување кои се користат во контекст на пребарување на медицински слики. Опишан е процесот на мултимодално пребарување на медицински слики и објаснати се чекорите од аспект на текстуално и содржински базирано пребарување, како и нивното спојување. Прикажани се базите на медицинско знаење кои се користат како поддршка во процесот на пребарување. На крај се опишани базите со слики и трудови за евалуација и метриките кои ги употребуваме за евалуација на перформансите на имплементираниот систем.
- **Петта глава:** Во оваа глава е презентирана архитектурата на имплементираниот систем и неговите подсистемите. Прво е презентирана генералната рамка на систем за пребарување на медицински документи со мултимодални податоци. Во останиот дел од главата се презенирани архитектурите на подсистемите и тоа на: подсистемот за класификација на слики според модалитет, подсистемот за пребарување на медицински слики, подсистемот за пребарување на медицински трудови со спојување на зборови и медицински концепти, подсистемот за пребарување на медицински трудови со проширување на прашања и подсистемот за пребарување на медицински трудови со генерички бази за знаење.
- **Шеста глава:** Експериментите се презентирани во оваа глава. Експериментите се групирани според подсистемот кој се евалира. За секој од експериментите се дефинирани поставувањата за извршување и прашањата кои треба се одговорат. Резултатите се прикажани според метриките од интерес и за истите е опширно дисктурирано.
- **Седма глава:** Во последната глава од докторската дисертација се презентирани заклучоците од анализата, имплементацијата и евалуацијата на системот за пребарување. Наведени се насоки за идни истражувања и идеи за понатамошни подобрувања на системите за пребарување на медицински слики и трудови.

1.4 Листа на објавени трудови поврзани со докторската дисертација

Резултатите од работата на докторската дисертација се потврдени во следните објавени публикации:

1. Ivan Kitanovski, Gjorgji Strezoski, Ivica Dimitrovski, Gjorgji Madjarov, and Suzana Loskovska. Multimodal medical image retrieval system. *Multimedia Tools and Appli-*

cations, pages 1–24, 2016

2. Ivan Kitanovski, Ivica Dimitrovski, Gjorgji Madjarov, and Suzana Loskovska. Medical image retrieval using multimodal data. In *International Conference on Discovery Science*, pages 144–155. Springer International Publishing, 2014
3. Ivica Dimitrovski, Dragi Kocev, Ivan Kitanovski, Suzana Loskovska, and Sašo Džeroski. Improved medical image modality classification using a combination of visual and textual features. *Computerized Medical Imaging and Graphics*, 39:14–26, 2015
4. Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loshkovska. Merging words and concepts for medical articles retrieval. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 25–28. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013
5. Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loskovska. Improving medical cases retrieval using an online fact database. In *ICT Innovations 2016: Cognitive Functions and Next Generation ICT Systems*, volume 9. Springer International Publishing, 2016
6. Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loskovska. Multimodal medical image retrieval. In *ICT Innovations 2012*, pages 81–89. Springer, 2013
7. Ivan Kitanovski, Ivica Dimitrovski, and Suzana Loskovska. Fcse at medical tasks of imageclef 2013. In *CLEF (Working Notes)*, 2013
8. Ivan Kitanovski, Ivica Dimitrovski, and Suzana Loskovska. Fcse at imageclef 2012: Evaluating techniques for medical image retrieval. In *CLEF (Online Working Notes / Labs / Workshop)*, 2012
9. Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loskovska. Query expansion methods for text-based retrieval of medical articles. In *ICT Innovations 2015: Emerging technologies for better living*, 2015
10. Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loshkovska. Web-based system for textual retrieval of medical images. In *Proceedings of the 12th International Conference for Informatics and Information Technology*, 2015
11. Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loskovska. Text-based medical image retrieval using query modification methods. In *ICT Innovations 2014: World of Data*, 2014
12. Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loshkovska. Word-space approach to case-based retrieval. In *Proceedings of the 11th International Conference for Informatics and Information Technology*, 2014

Глава 2

Опис на проблемот

Во оваа глава е опишан генералниот проблем на пребарување на информации, причината за потребата од системи за пребарување и општите пристапи за пребарувањето. Во вториот дел од главата ги објаснуваме специфичностите и отворените проблеми на пребарувањето на информации во областа на медицината.

2.1 Пребарување на информации

Архивирањето на пишаните податоци може да се забележи како практика уште од 3000 година п.н.е. кога Сумерите одредувале специјални места каде ги чувале глинените табли на кои запишувале податоци. Дури и тогаш, Сумерите сфатиле дека соодветна организација и пристап до архивите е клучно за ефикасно искористување на информациите што тие ги содржат. Тие развиле специјални класификации да ја идентификуваат секоја табла и нејзината содржина [13].

Потребата за чување и преземање на пишаните податоци стана особено важна со текот на годините и развојот на цивилизацијата, особено со појавувањето на изуми како хартијата и печатниците. Штом се појавија компјутерите, луѓето сфатија дека тие може да се користат за складирање и автоматско преземање на големи количества информации. Во 1950-тите почнуваат да се појавуваат идеи за автоматско пребарување на компјутерските архиви на податоци [14]. Притоа, во една идеја, која произлегува како доминантна, се предлага зборовите да се користат како елементи за индексирање на документите, а како критериум за пребарување да се користи бројот на заеднички зборови меѓу даденото барање и документите во архивата [15].

Во 1960-тите се направени неколку клучни напредоци во ова поле. Најзначајни се развојот на SMART системот [16] и Cranfield евалуациите [17]. Cranfield тестовите придонесоа за развој на методологија за еваулација на системите за пребарување која сé уште се користи и денес. Од друга страна, SMART системот овозможи развој на експерименти за подобрување на квалитетот на пребарувањето. Всушност, методологијата за евалуација во комбинација со флексибилен систем за експериментирање овозможија брз развој на полето.

Во текот на 1970-тите и 1980-тите развиени се голем број модели за пребарување на документи кои придонесоа за подобрување на сите аспекти од пребарувањето. Овие модели беа експриментално евалуирани на мали множества на податоци (неколку илјади документи) кои биле достапни во тоа време и се покажале како ефикасни. Но, поради недостигот на големи множества на податоци, многу прашања останаа отворени во поглед на скалабилноста на методите. Сето тоа се смени во 1992 со почетокот на конференцијата за текстуално пребарување (анг. *Text Retrieval Conference - TREC*) [18]. TREC е серија на конфереции финансирана од владата на САД со цел поттикнување на истражувањата на областа на пребарување на информации од големи множества на текстуални податоци. Достапноста до големите множества на текстуални податоци во рамките TREC придонесе кон модифицирање на многу од старите техники и развој на комплетно нови техники за пребарување (и сé уште се развиваат).

2.2 Специфичности на пребарувањето на медицински информации

Медицинските лица често се преоптоварени со информации, а од друга страна истите информации може да имаат важна улога при носење на одлуки поврзани со медицинската дијагностика и третирање на пациентот, или пак за образовни и истражувачки цели. Во тој контекст, компјутерските системи се појавуваат како поддршка во нивната работа, преку складирање, но и овозможување на лесен и брз пристап до валидни и релевантни информации за конкретниот проблем за кој медицинските лица се интересираат. Квалитетот на доставени информации во голема мера влијае на одлуката што тие ја носат. Во таа насока, компјутерските системи за поддршка во медицината се нераздвоиви од ефикасното пребарување на информации во областа на медицината [19].

Пребарување на информации во медицината е особено важно, бидејќи добивање на точна информација за дадено прашање е многу тешко, односно процесот на добивање на релевантен одговор на дадено прашање се смета за комплексна операција од повеќе причини. Една причина е фактот што бројот на медицински податоци експоненцијално расте. Имено, само базата на медицински трудови PubMed содржела 27 милиони записи заклучно со април 2017 година, а просечно секоја година во неа се додаваат околу 500 000 нови записи [20]. Друг предизвик е тоа што форматот во кои се чуваат податоците е различен за различни видови податоци. Форматот во кои се запишани медицинските појави зависи од опремата со која се направени испитувањата. На пример, скршена рака може да се репрезентира преку рентгенски слики, додека слабокрвност се репрезентира со бројни вредности организирани во извештаи. Во некои случаи потребно е да се користат повеќе различни видови на репрезентации за прецизно да се опише некоја појава. Ако се земе предвид дека има 12400 различни категории на медицински болести кои може да се репрезентираат со визуелни или текстуални податоци, тогаш од особена важност е да се овозможи ефикасно пребарување на медицинските податоци [21]

Терминот „пребарување на медицински информации“ (анг. *Medical Information Retrieval*) се однесува на методологии и технологии кои се обидуваат да го подобрат пристапот до архивите со медицински податоци преку процес на пребарување на информации (анг. *Information Retrieval - IR*). Таков вид на информации може да бидат достапни преку различни веб-страници, социјални мрежи, бази на медицински трудови и болнички архиви. Човековото здравје е меѓу најпопуларните теми за пребарување на интернет и како таква е од особен интерес за полето на пребарување на информации. Медицинските информации се од интерес на различни категории на луѓе, почнувајќи од пациентите и нивните блиски, истражувачите, општите лекари и специјалистите, радиолозите и тн. Веќе постојат многу сервиси кои се обидуваат да ги направат тие информации подостапни. Еден таков пример е системот за пристап до медицински податоци „Health on the Net“ [1].

Но, сепак, постојат бројни предизвици кои треба да се надминат за креирање на ефикасни сервиси за пребарување на медицински информации. Предизивиците може да се групираат според:

- **Видот на информации:** Пациенти кои скоро биле дијагностицирани со некоја болест повеќе би имале корист од едноставни или воведни информации за болеста или терапијата. Од друга страна, пациент кој има одредена болест подолг период може да има потреба од понапредни информации. На сличен начин, општ лекар можеби би имал потреба од некои едноставни или генерални податоци за болеста додека советува некој пациент, но, доколку станува збор за одредување на дијагноза или терапија, тогаш би имал потреба од подетални информаци. Специјалист би имал потреба од разгледување на слични медицински случаи или истражувачки трудови поврзани со состојбата на пациентот за кој се обидува да одреди дијагноза. Според тоа, еден од тековните предизвици на пребарување на медицинските информации, и воопшто во полето на пребарување на информации, е да се препознаат различните видови на корисници и информациите што им се потребни.
- **Нивото на медицинско знаење:** Различни групи на корисници имаат различно ниво на медицинско знаење, а и во рамките на истата група несомнено е дека нивото на знаење нема да биде исто меѓу сите луѓе. Тоа влијае на начинот на кој корисниците поставуваат прашања на системите за пребарување, како и на видот и комплексноста на информациите кои треба да бидат вратени.
- **Видот на податоците:** Форматот, квалитетот на биомедицинските и медицинските информации многу се разликуваат. Еден медицински запис може да содржи клинички записи, технички патолошки податоци, слики, историја од пациенти со слични состојби.

[1] http://www.hon.ch

2.3 Пребарување на медицински информации во рамки на истражувањата

Имајќи ги предвид различните форми на пребарување на медицински информации, истражувањата во докторската дисертација се однесуваат на пребарувања поврзани со медицински трудови како сеопфатни елементи на разнородни податоци. Медицинските трудови содржат големо количество на биомедицински податоци, што често се употребуваат во секојдневната медицинска практика. Вообичаено трудовите содржат слики, односно визуелни податоци кои служат за подобро објаснување на темата која се обработува во трудот. Стандардно на сликите им е придружен краток текстуален опис (анг. *caption*) што ја опишува нивната содржина. Сликите може да бидат фотографии, графици, дијаграми итн. Тие значајно придонесуваат за разбирање на содржината на трудот, а може да бидат употребени за образовни цели, истражувања или како поддршка при носење одлуки во фаза на дијагностицирање, одредување терапија или практика. Затоа, целокупниот проблем е поделен во два апликативни контексти: *пребарување на медицински слики* и *пребарување на медицински трудови (случаи)*.

Проблемот на пребарување на медицински слики може да се дефинира како пребарување каде за дадено текстуално и/или визуелно прашање треба да се врати подредена листа од слики на основа на нивниот текстуален опис (репрезентација) и/или визуелна содржина. Ова поле е особено атрактивно и веќе постојат многу методи кои се обидуваат да го решат проблемот. Во следната глава се опишани дел од најзначајните методи, но главниот проблем со кои се соочуваат истите се однесува на прецизноста на вратените резултати. Прецизноста е најважниот фактор кој треба да се анализира, бидејќи претставува директен индикатор за релевантноста на резултатите што ги добива крајниот корисник од системот. Системот може да помогне во секојдневната медицинска практика единствено ако дава резултати кои се релевантни за корисникот. Токму затоа, истражувањата направени во рамките на докторската дисертација го адресираат проблемот на *прецизноста* на пребарувањето. Во таа насока, предизвикот е да се пронајдат оптимални методи на комбинирање на различните видови на податоци, текстуални и визуелни, како и методи за класификација на податоците во фазата на индексирање и пребарување на големи колекции на медицински слики.

Пребарувањето на медицински трудови (случаи) се дефинира како пребарување каде за дадено опширно текстуално и/или визуелно прашање треба да се врати подредена листа на медицински трудови на основа на нивната содржина и/или сликите што тие ги содржат. Веќе постојат пристапи кон решавање на овој проблем, кои се презентирани во повеќе детали во следната глава. Но, главниот предизик со кои се соочуваат овие пристапи, повторно се однесува на прецизноста на пребарувањето. Според тоа, истражувањата направени во тој дел од докторската дисертација се однесуваат на подобрување на *прецизноста* на пребарувањето. Пребарување на основа на визуелни податоци не се покажало како особено ефикасно, поради што повеќе истражувања се насочуваат кон методи за збогатување/проширување на текстуалните податоци. Главните предизвици

во овој дел се однесуваат во наоѓање на начини на репрезентација на медицинските трудови и прашањата, како и имплементирање на нови и поефикасни методи за проширување на прашањата.

Глава 3

Преглед на литература

Во оваа глава е даден опширен преглед на техниките за пребарување на медицински документи. Даден е преглед на техниките за пребарување на медицински трудови и потенцирани се нивните недостатоци, како и генералните отворени предизвици. Дефиниран е проблемот на пребарување на генерички медицински слики. Образложени се методите и предизвиците за текстуалното и содржински базираното пребарување на медицински слики. Анализирани се методи за класификација на медицински слики според модалитет. Прикажани се постоечките методи и потенцирани се нивните недостатоци и предности. На крај, направен е преглед на методите за пребарување на медицински слики од медицински трудови.

3.1 Пребарување на медицински трудови

Пребарувањето на медицинските трудови се смета за задача која е поблиска до медицинската практика [22]. Кај овој вид на пребарување како влез се поставува детален текстуален опис за медицинската состојба која го интересира корисникот и/или прашални слики, а како резултат треба да се добие листа на медицински трудови. Во овој контекст, како главен предизвик претставува репрезентацијата на трудовите, како и проширувањето на прашања (анг. *Query Expansion*) кои треба да го збогатат текстуалниот влез на системот со цел прецизирање на прашањето и насочување кон соодветниот вид на трудови.

Покрај тоа што постојат онлајн системи како Pubmed [20], eTBLAST [23], Pubget [24] кои се употребуваат во практика, сé уште има многу обиди да се реши проблемот на пребарување на медицински трудови.

Групата Medgift [25] стандардно ги индексира медицинските трудови со помош на Lucene и прави пребарување над креираниот индекс. Интересно е да се напомене дека не применуваат никакви оптимизации на параметри или некакво збогатување на содржината. Покрај тоа, пријавуваат многу добри резултати.

Abdulahhad et al. [26] употребуваат метод на основа на концепти, односно тие ги мапираат сите документи и трудови во UMLS концепти со помош на Metamap [27].

Потоа, се имплементира стратегија за броење што треба да ја пресмета релацијата на даден документ во однос на дадено прашање. Во трудот се дефинира т.н. *frequency shift*, што всушност значи бројот на заеднички термини кои даден документ и прашање ги имаат е различен во просторот на зборови (анг. *word-space*) и во просторот на концепти (анг. *concept-space*) и според тоа треба да се третираат различно. Основната идеја во овој пристап е тоа што концептите кои се сочинуваат од повеќе зборови се позначајни, додека двозначните концепти се помалку значајни.

Wu et al. [28] употребуваат комбиниран пристап од текстуално и семантичко пребарување. Во овој пристап, документите и прашањата се мапирааат во MESH дескриптори со помош на MESHUP [29]. Во фазата на пребарување се употребува специјална мерка за одредување на семантичка сличност меѓу документите и прашањата. Конкретно, се употребува асиметрична мерка за сличност што ја пресметува сличноста меѓу концептите на основа на нивната хиерархиска релација на соодветните MESH десктиптори. Дополнително, се врши и стандардно текстуално пребарување со помош на Lemur алатката за пребарување и TF-IDF моделот на утежнување. Конечно, со линеарна комбинација се спојуваат резултатите од поединечните пребарувања и се добива крајниот резултат.

Интересен пристап со помош на Google Search API е претставен во [30]. Пристапот употребува две алатки за пребарување: Lucene и Essie. Текстуалното прашање прво се поставува кон Google Search API, а потоа се извлекуваат првите пет документи кои ќе бидат вратени од Google. Во следниот чекор, се анализираат сите медицински концепти кои се појавуваат во тие документи со помош на Metamap и се употребуваат трите најчести концепти за проширување на прашањето.

Vanegas et al. [31] со помош на Python NLP алатката имплементирале нивна верзија на Okapi BM25 моделот на утежнување за текстуално пребарување на слики. Иако, овој модел не е директно насочен кон пребарување на документи, сепак го прикажува потенцијалот кој може да се искористи и во овој домен.

Simpson et al [32] го користат системот Essie како платформа за пребарување. Пристапот прави проширување на прашањата со UMLS концепти со помош на Essie системот, кој овозможува таков вид на проширување. Овој пристап се покажа како добар за текстуално базирано пребарување на слики, но, покажал прилично лоши перформанси за пребарување на медицински трудови.

Во [33] е прикажан интересен пристап за пребарување на основа на текст. Пристапот го користи Indri системот [34] како платформа за пребарување. Клучен дел од овој пристап е делот со проширување на прашањата. Конкретно, се употребува Medline базата [35] како надворешен библиографски ресурс. Секое прашање, прво се изршува над Medline базата и се извлекуваат првите n документи. За секој вратен документ се извлекуваат соодветните MESH дескриптори и се додаваат во оригиналното прашање. Потоа, збогатеното прашање се извршува над тековната база на системот. Овој пристап се покажал како доста ефикасен, иако не на различни множества на податоци.

3.2 Пребарување на медицински слики

Генеричкото пребарување на медицинските слики може да биде на основа на визуелната содржина на сликата и/или придружените текстуални описи (анотации). Пристапите за пребарување на слики на основа на текстуални описи генерално користат стандардни методи за пребарување на текстуални документи. Главните проблеми кај овие пристапи се формирањето на текстуалната репрезентација на сликата на автоматски и надежен начин [22]. Во случај на генерички медицински слики кои се употребуваат како дополнителен материјал во медицинските трудови, текстуалните описи може да се генерираат на основа на трудовите во кои се појавуваат.

Пребарувањето на сликите на основа на нивните визуелни карактеристики се нарекува содржински базирано пребарување на слики (анг. *Content-based Image Retrieval - CBIR*) [36]. Системите за содржински базирано пребарување на слики на влез добиваат прашална слика, а како резултат прикажуваат слики со слични визуелни карактеристики на прашалната слика. Визуелните карактеристики се извлекуваат на автоматски или полуавтоматски начин со помош на програмски алгоритми (дескриптори). Овој пристап е скалабилен, бидејќи не зависи од рачното обележување на слики кое е скапо во однос на човечки ресурси и е подложно на субјективност која произлегува од човечкиот фактор [37].

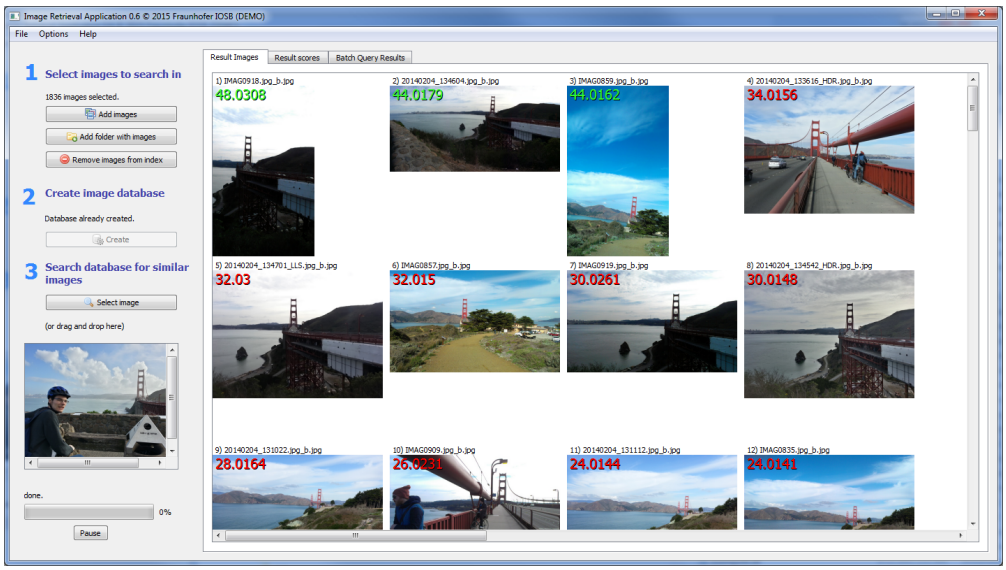
Дескрипторите се еден од најважните делови во системите за содржински базирано пребарување, бидејќи тие служат да се опише визуелната содржина на сликата. Кога дескрипторите ја опишуваат целата содржина на сликата, тие се нарекуваат глобални дескриптори. Дескрипторите кои опишуваат различни делови од сликата се нарекуваат локални дескриптори.

Глобални дескриптори се: хистограми на боја [38], Габор филтри [39], вектор на кохерентни бои [40], Тамура карактеристики [41] итн. Вообичаено, глобалните дескриптори не ја земаат предвид просторната распределба на бојата/текстурата и само даваат информација за нејзината процентуална застапеност.

Локалните десктриптори се отпорни на препокривање и сечење на објекти и различни геометриски трансформации [42]. Имено, кога сликата ќе претрпи геометриски промени или кога истата сцена се набљудува од друг агол, тогаш мора повторно да се пронајдат истите карактеристики. Овој вид на дескриптори се пресметуваат на одредени делови од сликата како некои региони, рабови или точки од интерес. Точките од интерес се меѓу најпопуларните каратеристики заради нивната робусност. Тие наоѓаат одредени точки од интерес и пресметуваат визуелни карактеристики на само мал дел околу најдените точки [43], [44]. Еден од најпопуларните локални дескриптори за оваа проблематика е SIFT дескрипторот (трансформација на дескриптори инваријантни на размер, анг. *Scale Invarian Feature Transform*) [45], [46], кој детектира клучни точки од интерес и генерира дескриптори инваријантни на трансформации на сликата.

Системите за содржински базирано пребарување на сликите потребно е да ги најдат сликите со слични визуелни карактеристики на прашалната слика. Во базата можно е

да постојат повеќе слики што одговараат на барањето на корисникот, па затоа системот најчесто враќа листа од сите слики подредени според нивото на сличност со прашалната слика. На страната на корисничкиот интерфејс на корисникот обично му се прикажуваат првите n слики, а му се овозможува и опција да ги погледне и следните. Постојат многу комерцијални и експериментални системи за содржински базирано пребарување на слики, а поширок и поопсежен преглед може да се најде во [47] и [48]. На Слика 3.1 е прикажан на интерфејсот на системот за содржински базирано пребарување развиен од Fraunhofer-Institute IOSB и претставува типичен интефејс за ваков вид на системи [49].



Слика 3.1: Интерфејс на систем за содржински базирано пребарување.

Во [50] е предложена архитектура на систем за содржински базирано пребарување на слики. Имено, станува збор за веб базиран систем кој овозможува складирање, пребарување, манипулација и анотација на медицински слики со помош на глобални и локални дескриптори. Овој систем го извршува пребарувањето преку споредување на генерираниот дескриптор на прашалната слика и сите слики во базата, пресметување на мерка на сличност и подредување на сите слики од базата во однос на нивната сличност со прашалната слика. Во [51] е претставено скалабилно содржински базирано пребарување со визуелен речник, со што се пресликува однесувањето на текстуално базираното пребарување. Овој пристап е скалабилен и може да поддржи големо множество на податоци.

3.3 Класификација на медицински слики

Медицинските лица вообичаено сакаат да ги филтрираат сликите добиени од пребарувањата според модалитетот на сликите [22]. Во трудовите, сликите многу ретко се обележани со нивниот модалитет, а да се направи тоа рачно е тешко и подложно на грешки. Затоа од особена важност е развој на механизам за автоматско означување на

слики со нивниот модалитет, односно автоматска класификација на сликите според модалитет на основа на нивната визуелна содржина. Вклучувајќи го овој механизам автоматски во процесот на пребарување може да доведе до подобри перформанси.

Во таа насока разгледани се различни пристапи за автоматско означување на медицински слики [52], [7], [53], [54], [55], [56]. Во [57], [58], [59] се презентирани повеќе пристапи за класификација/означување на медицински слики. Важно е да се напомене, дека кога станува збор за означување во рамките на докторската дисертација тоа се однесува на означување на модалитетот на слика. Автоматското означување на слики е овозможено со помош на техники на машинско учење, кои прво користат множество на претходно означени слики за обучување на класификатор за потоа да можат автоматски да означат нови слики [58]. Во [60] е потенцирано дека модалитетот на сликите може да се извлече и од визуелните карактеристики на сликите.

Најчесто користени техники за машинско учење кои се употребуваат во контекст на класификација на сликите според модалитет се машини со носечки вектори (анг. *Support Vector Machines - SVM*) [7], [32], [56], k-најблиски соседи (анг. *k-nearest neighbour - KNN*) [52], [54], логистички модел на регресија (анг. *logistic regression model*) [56] итн.

Машините со носечки вектори се популарен метод на машинско учење кој веќе успешно се употребува за препознавање на објекти, класификација на текст итн. Во однос на класификација на модалитети на медицински слики во [7] се употребуваат машини со носечки вектори со пристап на *еден-против-сите*(анг. *one-vs-all*), каде во фазата на обучување за секоја класа се креира по еден класификатор. Во фазата на тестирање сите неозначени слики поминуваат низ бинарните класификатори и најчесто се зема модалитетот од класификаторот кој дал најголема веројатност.

Методот на k-најблиски соседи е еден од концепциски наједноставните алгоритми за класификација. Методот е базиран на пресметување на растојанија меѓу примероците. Работи на принцип на наоѓање на подмножество од k примероци, кои се најблиски до примерокот што треба да се класифицира. Класната припадност на примерокот се определува на основа на доминантната класа од подмножеството. Во контекст на класификација на медицински слики според модалитет во [61] се потенцираат главните предизвици како одредување на k и мерката на сличност меѓу сликите.

3.4 Пребарување на слики од медицински трудови

Вообичаените методи за пребарување на генерички медицински слики се на основа на текст, односно сликите се пребаруваат на основа на нивните текстуални анотации/описи. Еден дел од проблемот е изнаоѓање на соодветна текстуална репрезентација на сликите. Сликите често се користат подобро да се опише некој дел од медицинските трудови и се важен податок во дадениот контекст. Според тоа, текстот во медицинските трудови може да се користи да се добие текстуална репрезентација на сликите.

Во [62] е претставен системот за пребарување на медицински трудови BioText кој го употребува Lucene како платформа за индексирање и пребарување. Во рамките на овој

систем постои дел за пребарување на медицински слики. Сликите се пребаруваат на основа на делови од текстот каде сликите се референцирани, како и самиот текстуален опис кој го содржи сликата.

Интересен преглед на текстуалното наспроти содржински базираното пребарување на медицински слики е претставено во [63] користејќи множество на слики на кои не им е придружена текстуална репрезентација. Текстуалните репрезентации за сликите се извлечени од описите кои сликите ги содржат како и од параграфите од текстот каде сликите се референцирани. Во рамките на тоа истражување беа направени неколку видови евалуцаии: евалуација на креирање на индекси, каде произведената IRMA [64] анотација се проверува и евалуација на пребарувањето, каде за дадено прашање се проверува листата на слики кои ќе ги врати. Резултатите кои ги пријавуваат се во насока дека индексирањето и пребарувањето со помош на визуелните карактеристики даваат подобри перформанси со тоа што во одредени случаи индексирањето со помош на текстуални описи давало подобри IRMA анотации. Дополнително, резултатите покажале дека подобри текстуални репрезентации може да се генерираат на основа на текстуални описи на сликите, отколку на параграфите во кои сликите се референцирани. Но, авторите извршиле и мултимодален експеримент, каде се комбинираат текстуалните и визуелните податоци, при што се добиле подобри резултати во фазата на индексирање и мало подобрување во фазата на пребарување. Недостакот на овој пристап е што не се обидува да ги вклучи и другите делови од трудот при формирање на текстуална репрезентација на сликите.

Во [65] е поставена тезата дека во „реални“ медицински услови и податоци, техниките на основа на содржински базираното пребарување не даваат резултати доволни за активна примена во индустријата. Според тоа, тие се неприменливи во активна медицинска работа. Во рамките на истото истражување е пријавено дека пристапите каде се комбинираат текстуални и визуелни податоци прикажуваат подобри перформанси отколку пребарување само на основа на текстуални или само на основа на визуелни податоци.

Rahman et al. во [66] предлагаат пребарување на генерички медицински слики на основа на мултимодални податоци. Системот што го имплементирале користи текстуални и визуелни податоци и т.н. метод за филтрирање. Текстуалниот дел од системот ги индексира и пребарува низ текстуалните репрезентации на сликите. Како платформа за пребарување на текстуалниот дел се користи Essie. Во однос на содржински базираното пребарување, пристапот се состои од исфрлање на сликите кои не се од ист модалитет како прашалната слика, а потоа се извршува пребарување низ преостанатите. Детекцијата на модалитетот се извршува само на основа на визуелната содржина на сликата. Системот прави текстуално и содржински базирано пребарување, а потоа со линеарна комбинација се спојуваат поединечните резултати и се добиваат конечните резултати. Главниот недостаток на овој пристап е во поглед на ефикасноста, бидејќи се употребуваат карактеристики од ниско ниво за опишување на сликите, а соодветно со тоа и за класификција на сликите според модалитетот.

Пребарувањето во Yale Image Finder (*YIF*) [67] системот се состои од генерирање на текстуална репрезентација на сликите според трудовите каде се појавуваат и извлекување на потенцијална текстуална содржина во рамките на самата слика. Текстуалната репрезентација се генерира од насловот, апстрактот и текстуалните описи на сликите во медицинските трудови каде се појавуваат. Извлекувањето на текстот од сликата се прави со помош на техники за оптичко детектирање на знаци (анг. *Optical Character Recognition - OCR*).

Системот Goldminer презентиран во [68] е специјализиран за пребарување на радиолошки слики на основа на нивните текстуални описи во трудовите кои се појавуваат. Трудовите се извлечени од Radiological Society of North America (*RSNA*). Системот ги мапира текстуалните описи со UMLS концепти и извршува пребарување на таа основа.

Stathopoulos el al. [56] креираат структурирана текстуална репрезентација од сликите и извршуваат индексирање со помош на Lucene и соодветните механизми справување со документи организирани во полиња (анг. *field-based documents*). Фазата на пребарување се одвива преку додавање на различни тежини на полињата што ја содржат текстуалната репрезентација во зависност од делот од трудот од каде информацијата е извлечена. Идејата позади овој пристап е дека различни делови од трудот носат различно ниво на информација при опишување на сликата.

Ozturkmenoglu et al [53] предлагаат систем за пребарување на основа на Terrier IR системот за пребарување. Пребарувањето во системот се одвива во две фази. Прво, системот го детекира модалитетот на прашалната слика и ги филтрира сликите кои не припаѓаат на таа класа. Втората фаза се состои од пребарување низ филтрираното подмножество на слики. Најголемиот проблем е во неговата ефикасност. Имено, системот прикажува многу слаби резулати во однос на слични системи. Пребарувањето примарно зависи од филтрирањето и квалитетот на детекција на модалитет, што не е секогаш одлучувачки фактор во пребарувањето.

Повеќето од претходните методи функционираат на основа на текстуалните податоци асоцирани со сликите, но постојат и содржински базирани методи кои даваат релативно послаби резултати при пребарување на генерички медицински слики. Дескрипторите кои се употребуваат во овој случај се фокусирани на глобални бои (на пример, хистограми), рабови (на пример, Габор карактеристики), текстури (на пример, Тамура карактеристики) или комбинирани информации (на пример, CEDD). Но, најчесто се употребува комбинација од сите тие карактеристики [69]. Поновите системи, се повеќе користат локални карактеристики на основа на клучни точки [37].

Мултимодално пребарување на медицински слики подразбира комбинирање на текстуалните и визуелните податоци во фазата на индексирање и/или пребарување. Клучен проблем при мултимодалното пребарување е спојувањето (фузијата) на податоците од различни модалитети. Постојат три главни начини на спојување [70]:

- рана фузија (анг. *early fusion*): спојување на ниво на карактеристики
- доцна фузија (анг. *late fusion*): спојување на ниво на одлука
- хибридна фузија (анг. *hybrid fusion*): комбинација од претходните две

Методите со рана фузија комбинираат различни карактеристики во единствен дескриптор [25]. Потоа, овој комбиниран дескриптор се употребува во фазата на индексирање, пребарување и/или класификација. Наједноставен пример за рана фузија е конкатенација на различни вектори на карактеристики во единствен вектор. Раната фузија овозможува полесна анализа на корелацијата меѓу различните видови на податоци. Најголем недостаток и предизвик кај овие методи е фактот што треба да се трансформираат различните видови на податоци во заеднички формат, како и утежнувањето и скалирањето на вредностите.

Методите со доцна фузија ги користат резултатите од повеќе моно-модалитетни системи да се формира конечниот резултат [71]. Повеќе од методите со доцна фузија во основа се ненадгледувани техники кои го користат рангот или *score*-от на документите за да се пресмета конечниот резултат. Главниот недостаток на овие методи е што покрај ранговите треба да биде придружена нумеричка вредност за секој документ/слика. Добивањето на тие вредности е различно од еден систем до друг и нормализирањето на истите може многу да влијае на перформансите [72]. Но, од друга страна овозможува повеќе контрола врз тоа кој модалитет колку влијае на крајниот резултат со што се добива модуларност и скалабилност.

Глава 4

Методологија

Еден систем за пребарување на информации се оценува на основа на тоа колку *добри* резултати враќа за прашањата кои ги поставил корисникот. Според тоа, целта на нашите истражувања е да се имплементира систем кој ќе биде евалуиран за *квалитетот* на вратените резултати. Квалитетот го разгледуваме на основа на бројот на релевантни документи што ги враќа системот.

За тестирање и евалуирање на перформансите на системите за пребарување на информации се употребуваат специјално дизајнирани колекции/множества на документи (анг. *document test collection*). Тие се состојат од три дела: документи, прашања (анг. *queries/topics*) и одговори (анг. *relevance judgements*). Дадените прашања се користат да се направи пребарување на системот кој се тестира, а потоа добиените резултати се споредуваат со предодредените одговори и на таа основа се прави еваулацијата на системот. Во таа насока методите (моделите) на пребарување кои се имплементирани во системот влијаат на квалитетот на крајните резултати.

Имајќи го тоа предвид во оваа глава ја опишуваме методологијата на работа во текот на истражувањето, целите кои сме ги поставиле, како и начинот на кој ги евалуираме имплементираните решенија. Тоа го опфаќа општиот процес на пребарување, моделите на пребарување, клучните техники при пребарување на информации, методи за извлекување на карактеристики, платформите за пребарување, процесот на пребарување на медицински слики, базите на медицинско знаење, како и базите за евалуација и метриките за оценување на работата на имплементираните методи.

4.1 Општ процес на пребарување на информации

Општиот процес на пребарување на информации е прикажан на Слика 4.1. Процесот се состои од три чекори [73]:

- Репрезентирање на множеството документи низ кои пребарува системот
- Репрезентирање на барањето на корисникот
- Споредување на двете репрезентации

Слика 4.1: Општ процес на пребарување на информации.

Чекорот на репрезентација на содржината на документите се нарекува *индексирање* (анг. *indexing*). Во овој чекор се обработува множеството на документи низ кои се пребарува и се формира репрезентација т.е. се извлекуваат карактеристики за секој документ. Чекорот на репрезентација на барањето на корисникот се вика *формулација на прашање* (анг. *query formulation*). Во оваа фаза корисникот задава влезни податоци на чија основа се врши пребарувањето. Од системот се очекува да врати подредена листа на документи, при што најрелевантниот документ треба да е позициониран најгоре во листата. Ова се нарекува *ad-hoc* пребарување. Со цел да се поддржи пребарување низ множеството на документи, прашањата треба да се репрезентираат на ист начин како и документите, односно и за прашањата треба да се извлечат карактеристики кои ќе бидат споредливи со карактеристиките на документите. Во реално сценарио, индексирањето на документите се извршува предвремено (офлајн фаза), а процесирањето на прашањето и споредувањето на неговата репрезентација со документите се прави откако корисникот ќе го зададе во системот (онлајн фаза). Почетното прашање може дополнително да се модифицира на автоматски или полуавтоматски начин.

Во чекорот на споредување, системот мора да има начин како да ја споредува содржината на прашањето и документите од множеството, со цел да генерира подредена листа на документи. Тоа значи дека системот мора да има имплементирано метод (вообичаено математички модел) кој за дадена репрезентација на прашање и документ враќа нумеричка вредност за нивната сличност. Со помош на оваа квантитативна метрика системот може да ги подреди документите.

4.2 Модели на пребарување на информации

Пребарување на информации овозможува наоѓање на релевантни податоци за поставено прашање од дадено множество, а системите кои ја реализираат постапката се нарекуваат системи за пребарување на информации. Почетните системи за пребарување на информации функционираа на основа на булова логика (анг. *boolean systems*) каде на корисниците им беше овозможено да ги специфицираат прашања со помош на комплексни комбинации од булови оператори, односно со И (анг. ANG), ИЛИ (анг. OR) и НЕГАЦИЈА (анг. NOT). Тие системи имаа низа на недостатоци. На пример, кај овие системи немаше концепт на рангирање на резултантните документи, а воедно беше многу тешко за корисникот добро да го формира прашањето. Документите кои системите ги враќаа беа подредени според некој принцип, како датум на ажурирање на документот или некоја друга особина, но, во основа немаше концепт за релевантноста на документот во однос на даденото прашање. Корисниците денес очекуваат дека системите за пребарување ќе враќаат резултати подредени според нивната релевантност. Системите за пребарување ги рангираат документите според нивната проценка за сличноста на документот со даденото прашање. Постојат повеќе модели за дефинирање на таа проценка, но најчесто користени методи се векторскиот модел (анг. *vector space model*), моделот на основа на веројатност (анг. *probabilistic model*) и inference network моделот.

4.2.1 Векторски модел

Во векторските модели текстот е опишан со вектор од термини [74]. Дефиницијата за *термин* не е дадена во моделот, но генерално за термини се сметаат зборовите или фразите во текстот. Доколку зборовите се употребуваат како термини, тогаш секој збор во речникот станува индекс во повеќедимензионален вектор. Во овој случај секој текст може да се претстави како вектор. Доколку даден термин припаѓа на одреден текст, тогаш тој добива вредност (што не е нула) во текстуалниот вектор во индексот што соодветствува на тој термин. Секој текст има конечно множество на термини (речникот може да има милиони термини), па според тоа повеќето текстуални вектори ќе имаат многу елементи со вредност 0, бидејќи доколку терминот не е присутен во текстот ќе има вредност 0 за соодветниот индекс.

Со цел да се додели нумеричка вредност на даден документ во однос на некое прашање, моделот пресметува *сличност* меѓу векторот на прашањето (бидејќи и прашањето е краток текст кој може да се претвори во вектор) и векторот на документот. Сличноста меѓу два вектори не е директно дефинирана во векторскиот модел. Вообичаено се пресметува косинусно растојание меѓу векторите и добиената вредност се употребува како мерка за сличност. Косинусното растојание како резултат дава 1.0 за исти вектори, а 0.0 за ортогонални вектори. Како алтернатива може да се пресмета скаларен производ меѓу два вектори. Доколку $\vec{D}$ е векторот на документот и $\vec{Q}$ е векторот на прашањето, тогаш сличноста на документот D во однос на прашањето Q може да се претстави на

следниот начин:

$$Sim(\vec{D}, \vec{Q}) = \sum_{t_i \in Q,D} w_{t_iQ} \cdot w_{t_iD} \tag{4.1}$$

каде w_{t_iQ} е вредноста на i-тата компонента во векторот на прашањето $\vec{Q}$, а w_{t_iD} е i-тата компонента во векторот на документот $\vec{D}$. Бидејќи секој термин што не е присутен во прашањето или документот ќе има за w_{t_iQ}, односно w_{t_iD} вредност 0, соодветно, тогаш доволно е само да се направи сума само на оние термини кои се заеднички за прашањето и документот. Векторскиот модел не дефинира како да се пресметаат w_{t_iQ} и w_{t_iD}, но тие се од особена важност за ефикасноста на еден систем за пребарување. w_{t_iD} се именува како *тежина* (анг. *weight*) на i-тиот термин во документот D.

4.2.2 Модел на основа на веројатност

Оваа група на модели на пребарување на информации функционира на основа на тоа дека документите во дадено податочно множество (колекција) треба да бидат рангирани според веројатноста за нивната релевантност во однос на дадено прашање во опаѓачки редослед (анг. *probabilistic ranking principle - PRP*) [75]. Целта на овие модели е со пресметка да се процени веројатноста за релевантноста на документите во однос на некое прашање. Разликите меѓу поединечните модели во рамките на оваа група е во начинот на кој се пресметуваат веројатностите [76].

4.2.3 Inference модел

Овој на модел на пребарување користи техники за пресметување на дистрибуција на постериорна веројатност во контекст на мрежи (анг. *inference network*) [77]. Многу од постоечките техники за пребарување на информации може да се претстават со овој модел. Во наједноставната имплементација на овој модел, за даден документ се инстанцира термин со одредена вредност (тежина) и се собираат вредностите од сите термини што ги содржи во однос на дадено прашање и како резултат се добива нумеричка вредност. Рангирањето на документите е многу слично со векторскиот модел и моделите на основа на веројатност. Начинот на кој се пресметува вредноста на терминот не е дефинирана во рамките на овој модел и може да се употребува која било формулација.

4.3 Клучни техники за поддршка на пребарување на информации

Најважниот податок кој се употребува во сите модели на рангирање на документи е тежината на даден термин во одреден документ. Многу истражувања се направени во насока на пресметување на тежината на термините, а според тоа се појавиле и методи за

изведување на постапката т.н. модели на утежнување (анг. *weighting model*). Друга техника која се покажала како ефикасна во рангирањето на документите е модификација на прашањето преку релевантна повратна врска (анг. *relevance feedback*). Еден современ систем за рангирање треба ефикасно да користи модел за утежнување во комбинација со добра техника за проширување на прашањето.

4.3.1 Модели на утежнување

Истражувањата во полето на пребарување на информации резултираат со појава на различни методи за пресметување на тежината на термините. Моделите за утежнување во рамките на методите за рангирање на основа на веројатност се фокусираат на различни начини на пресметување на одредени веројатности кои ја репрезентираат тежината [78]. Моделите за утежнување во рамките на векторските модели на рангирање се развиени на основа на опсежни експерименти и искуства со системите за пребарување. Во секој случај постојат три главни фактори кои се земаат предвид при пресметувањето на тежината на даден термин:

Фреквенција на терминот (анг. *term frequency - tf*): Оваа мерка се однесува на тоа колку пати терминот се појавува во даден документ. Термини што се повторуваат повеќе пати во документот се сметаат за важни. За даден документ, множеството тежини одредени со помош на tf за термините што ги содржи може да се гледа како репрезентација на документот. Кај овој начин на репрезентирање на документите (т.н. *bag-of-words* модел) се игнорира точниот редослед на термините во документот, а се анализира бројот на појавувања на термините.

Фреквенција на документот (анг. *document frequency - df*): Оваа мерка всушност означува во колку документи се појавува дадениот термин. Проблемот со tf е во тоа што користејќи ја само таа мерка, сите термини се третираат како еднакви. Терминте кои се појавуваат во многу документи се вообичаени и не се особено важни за содржината на документот. Според тоа, се воведува инверзна фреквенција на документот (анг. *inverse document frequecy - idf*), која треба да ја нормализира тежината произведена со помош на tf. Нека N е бројот на документи во податочното множество, тогаш idf за терминот t се пресметува на следниот начин:

$$idf_t = log\frac{N}{df_t} \tag{4.2}$$

Според 4.2 тежината за ретки термини е висока, а ниска за термини кои често се наоѓаат во документите.

Должина на документот (анг. *document length*): Множествата на податоци имаат документи со различни должини. Поголеми документи може да се појават погоре во рангирањето, бидејќи содржат повеќе зборови. Овој ефект се намалува на тој начин што се нормализира должината на документите во утежнувањето на терминте.

Во продолжение се претставени најчесто користените модели за утежнување.

TF-IDF

TF-IDF (*term frequency-inverse document frequency*) е модел за утежнување кој задава квантитаивна мерка за важноста на даден термин t во даден документ d. Според овој модел, тежината се пресметува на следниот начин:

$$tf - idf_{t,d} = tf_{t,d} \times idf_t \tag{4.3}$$

Формулата 4.3 доделува тежина $tf - idf_{t,d}$ на терминот t која е:

1. највисока кога терминот t се појавува многу пати во мал број на документи
2. ниска кога терминот t се појавува малку пати во даден документ
3. најниска кога терминот t се појавува во сите документи

На основа на тежината може да се пресметаат различни мерки на сличност меѓу дадено прашање и документ. Една основна мерка е т.н. мерка на преклопување (анг. *overlap score measure*) и се пресметува на следниот начин:

$$Score(q, d) = \sum_{t \in q} tf - idf_{t,d} \tag{4.4}$$

Оваа формула ни кажува колку колку заеднички термини имаат дадено прашање и документ.

BM25

BM25 моделот на утежнување или уште познат како *Okapi* модел, според системот каде прв пат се употребил, претставува математички модел, кој ги зема предвид фрекфенциите на термините и должината на документот при пресметување на тежината и додава неколку параметри за оптимизирање [79]. Наједноставен начин за пресметување на релевантноста (RSV) на даден документ d во однос на прашање q е преку инверзната фреквенција на документот:

$$RSV_d = \sum_{t \in q} log \frac{N}{df_t} \tag{4.5}$$

Некогаш се употребува алтернативна верзија на idf.

$$RSV_d = \sum_{t \in q} log \frac{N - df_f + \frac{1}{2}}{df_t + \frac{1}{2}} \tag{4.6}$$

Верзијата на формулата 4.6 може да се однесува невообичаено во одредени случаи. На пример, ако даден термин се појави во повеќе од половина од документите од множеството тогаш овој модел ќе произведе негативна тежина, што е проблематично, бидејќи сите тежини се очекуваат да бидат позитивни. Но, доколку се елиминираат термините кои се појавуваат често, овој проблем нема да се појави.

Формулата 4.5 може да се подобри на тој начин што ќе се земат предвид и фрекфенциите на термините и должината на документот:

$$RSV_d = \sum_{t \in q} [log \frac{N}{df_t}] \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1 - b) + b \times (L_d / L_{ave}) + tf_{td})} \tag{4.7}$$

Според формулата 4.7, tf_{td} е фревенцијата на терминот t во документот d, L_d и L_{ave} се должината на документот и просечната должина на документите во податочното множество, соодветно. Променливата k_1 е параметер за оптимизација кој треба да биде позитивен и служи за контролирање на влијанието на фреквенцијата на терминот во крајниот резултат. Друг параметер за оптимизација е b $(0 \leq b \leq 1)$ кој го одредува влијанието на должината на документот на крајниот резултат.

Доколку прашањето е подолго, тогаш може да се додаде нормализација и во тој дел од утежнувањето:

$$RSV_d = \sum_{t \in q} [log \frac{N}{df_t}] \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1 - b) + b \times (L_d / L_{ave}) + tf_{td})} \cdot \frac{(k_3 + 1) tf_{tq}}{k_3 + tf_{tq}} \tag{4.8}$$

Каде tf_{tq} е фреквенцијата на терминот t во прашањето q, а k_3 е параметар за оптимизација која го контролира влијанието на фреквенцијата на терминот во прашањето. Препорачано е овие параметри, прво да се оптимизраат над некое тест множество (рачно или со некои методи како *grid search*), а потоа да употребуваат над реално множество. Експериментите покажале дека препорачливи вредности за k_1 и k_3 се вредности меѓу 1.2 и 2, а додека за b препорачана вредност е 0.75.

Горенаведените формули ја претставуваат суштината на BM25 моделот за утежнување и нашироко се употребуваат и се покажале како доста успешни во различни множества на податоци.

4.3.2 Модификација на прашањата

Уште во првите системи за пребарување беше забележано дека е прилично тешко корисниците да формулираат адвектватни прашања според кои би се направило ефикасно пребарување. Се сметало дека додавање на синоними на прашалните зборови би го подобрило пребарувањето. Првичните системи за пребарување на информации користеле речник на зборови за наоѓање на синоними и нивно додавање во прашањето. Но, бидејќи е прилично тешко да се обезбеди квалитетен генерички речник, истражувањата повеќе се насочуваа кон автоматско генерирање на зборови што би се користеле за проширување/променување на прашањата. Методите за справување со овој проблем се делат на две групи: глобални и локални методи [80]. Глобалните методи се техники за проширување на прашањата или реформулирање на некои термини во прашањата независно од прашањата и вратените резултати. Глобалните методи вклучуваат:

- Проширување/преформулирање на прашањата со помош на речник
- Проширување на прашањата со помош на автоматски генерирани речници

- Техники за поправање на печатни грешки

Локалните методи го приспособуваат прашањето според документите кои иницијално се вратени од системот. Основите методи од оваа група се:

- Релевантна повратна врска
- Псевдо-релевантна повратна врска
- Индиректна релевантна повратна врска

Методот на релевантна повратна врска (анг. *relevance feedback*) е предложен во 1965 од Rocchio со цел изменување на прашањата [81] и подобрување на резултатите од пребарувањето. Во кратки црти постапката се одвива на следниот начин:

- Корисникот поставува прашање кон системот
- Системот враќа почетни резултати
- Корисникот ги означува вратените документи како релевантни или нерелевантни
- Системот пресметува подобра репрезентација за информациските потреби на корисникот на основа на тоа како ги означил документите
- Системот прави уште едно пребарување и повторно враќа резултати

Овој процес може да се повторува во повеќе итерации. Идејата користена во методот е дека е тешко да се формулира прашање кога не се знае добро податочното множество што се пребарува. Но, сепак, корисникот знае да процени дали одредени документи се релевантни или нерелевантни за него, и на тој начин итеративно да го подобрува пребарувањето.

Во текот на 1990-тите се развија многу техники за автомаско проширување на прашањата без никаква интеракција од страна на корисникот. Меѓу нив од особен интерес е методот на псевдо-релевантна врска (анг. *pseudo-relevance feedback*). Псевдо-релеванта врска е варијанта на релевантната врска која е уште позната како „слепа“ релевантна врска (анг. *blind relevance feedback*) [80]. Овој метод овозможува автоматска локална анализа. Го автоматизира рачниот дел од релевантната повратна врска, со цел корисникот да добие подобри резултати без дополнителна интеракција со системот. Постапката кај овој метод е следна:

- Корисникот поставува прашање кон системот
- Системот враќа почетни резултати
- Системот ги означува првите n документи за релевантни
- Системот ги извлекува m-те *најинформативни* термини од релевантните документи и ги додава на прашањето
- Системот прави уште едно пребарување со проширеното прашање и повторно враќа резултати

Параметрите n и m се конфигурираат и може да се прават експерименти за нивна оптимизација зависно од податочното множество.

4.4 Извлекување на карактеристики

Податоците што се користат во системот, односно документите низ кои се пребарува, како и прашањата кои се упатуваат до системот треба соодветно да се репрезентирани со цел да може да се споредуваат. Репрезентацијата се прави со методи на извлекување на карактеристики од податоците. Методите за извлекување можеме да ги поделиме според видот на податоците од кои извлекуваат карактеристиките: текстуални и визуелни карактеристики. Имајќи предвид дека во нашите истражувања се фокусираме на пребарување на медицински слики и трудови, тогаш од интерес ни се двата видови на карактеристики.

4.4.1 Текстуални карактеристики

Медицинските трудови и слики (може) да имаат своја текстуална репрезентација. Во рамките на истражувањето се користи векторскиот модел за опишување на текстуалните репрезентации. На почеток, текстуалните податоци за дадена слика/документ се претпроцесираат. Прво, се применува токенизација со што текстот се разложува на поединечни зборови. Во следниот чекор се бришат сите стоп зборови. Бришењето на стоп зборовите е неопходно, бидејќи тие зборови се употребуваат често и само го зголемуваат шумот, а не носат дополнителна инфромативна вредност (на пример, *a, from, of, to* итн.). Веднаш потоа се применува стемирање со помош на Porter stemmer [82]. Стемирањето ги редуцира зборовите во нивната основна форма со цел нормализација на различните форми во кои зборовите може да се најдат.

4.4.2 Визуелни карактеристики

Визуелните карактеристики се релевантни за медицинските слики. Визуелни карактеристики (дескриптори) се еден од најважните делови во системите за (содржински базирано) пребарување и класификација, бидејќи тие служат да се опише визуелната содржина на сликата. Кога дескрипторите ја опишуваат целата содржина на сликата, тие се нарекуваат глобални дескриптори. Дескрипторите кои опишуваат различни делови од сликата се нарекуваат локални дескриптори. Во продолжение ќе бидат објаснети некои од најкористените во контекст на пребарување и класификација на медицински информации.

Локални бинарни шаблони

Локалните бинарни шаблони (анг. *Local Binary Patters - LBP*) се едни од најдобрите дескриптори за опишување на текстури во сликите [83]. Тие се инваријантни на монотоните промени на нијансите на сиво и се брзи за пресметување. Уште повеќе, тие се способни да детектираат микро шаблони, како рабови и унформни точки и региони.

LBP се темелат на извлекување на информации за текстурата од локално соседство на точки. Прво, се дефинира радиус R на локалното соседство над кое се врши

извлекувањето. Следно, LBP операторот генерира бинарен код што ја опишува локалната текстура на локалните соседства на P пиксели. Бинарниот код се пресметува на тој начин што вредноста на централниот пиксел од локалното соседство се употребува како прагова вредност. Бинарниот код се конвертира во декаден број што го репрезентира LBP кодот. Според дефиниција за даден пиксел со позиција (x_c, y_c) LBP кодот се пресметува на следниот начин:

$$LBP_{(P,R)}(x_c, y_c) = \sum_{n=0}^{P-1} S(i_n - i_c)2^n \tag{4.9}$$

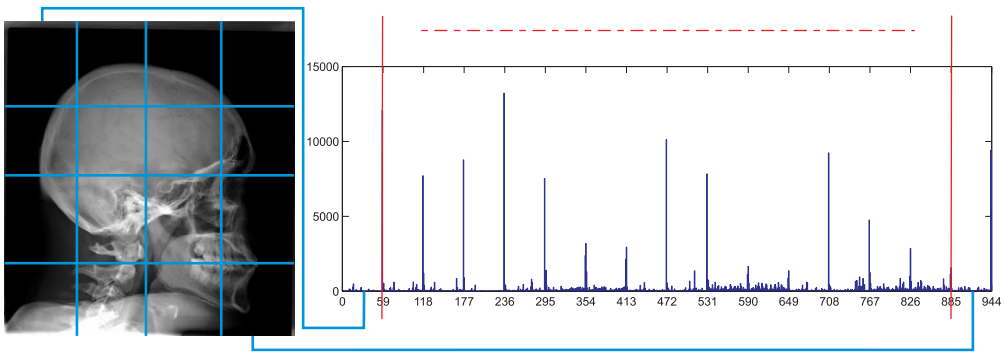
каде n ги изминува сите P соседи на централниот пиксел (x_c, y_c), а i_c и i_n се вредностите за нијансите на сиво на централниот пиксел и соседниот пиксел, а $S(x)$ е дефинирана на следниот начин:

$$S(x) = 1, x \geq 0; \;\; 0, x < 0 \tag{4.10}$$

LBP операторот ја изминува целата слика пиксел по пиксел и добиените резултати се акумулираат во хистограм.

Но, не сите LBP кодови се информативни. Кодови кој опишуваат фундаментални карактеристики и својства на текстурата се нарекуваат униформни шаблони. Униформните шаблони се околу 90% од вкупниот број на шаблони присутни во текстурата која се опишува [83]. Овие шаблони се региони од сликата со многу мал број на просторни промени и служат како темплејти за репрезентирање на микроструктурите во сликата.

Со цел подобрување на перформансите на дескрипторот се предлага да се анализираат предефинирани подрегиони/подслики од дадена слика (на пример, 1x1, 2x2, 4x4 итн.) [61]. Дескрипторите генерирани од различните подслики се агрегираат/спојуваат во еден поопишен дескриптор. Според оваа сугестија во рамките на овој дел од истражувањето сликите се поделени на неколку региони и за секој од регионите се пресметува LBP дескриптор. Сите пресметани дескрипотори за секој од регионите се спојуваат во еден агрегатен дескриптор за секоја слика. На Слика 4.2 е прикажан начинот како се гради еден LBP хистограм со 16 x 243 = 3888 бинови, за секоја слика. Сликата е поделена на 4x4 региони.



Слика 4.2: Сликата е поделена на 4x4 региони за кои се пресметуваат LBP хистограми и сите хистограми се спојуваат во еден агрегатен хистограм за целата слика.

Дескриптор за насока на бои и рабови

Дескрипторот за насока на бои и рабови (анг. *Color Edge Directivity Descriptor - CEDD*) вклучува информации за бојата и тесктурата во еден хистограм [84]. CEDD хистограмот се состои од 6 региони/бинови кои се одредени од текстурата во сликата. Секој регион се содржи од 24 подрегиони кои произлегуваат од бојата во тој регион. Според тоа, крајниот хистограм се состои од 6 x 24 = 144 региони/бинови.

Хистограмот се пресметува на тој начин што сликата се дели на региони [84] и за секој регион се пресметуваат податоци за текстурите и боите. Информацијата за текстурите се пресметува со помош на 5 дигитални филтри и дескрипторот за рабови (анг. *edge histogram descriptor*) од MPEG-7 [85], [86]. Петте филтри се всушност насоките на рабовите: вертикални рабови, хоризонтални рабови, рабови под 45°, рабови под 135° и рабови без насока. Од овој дел се добива хистограм со 6 бинови од кои 5 се однесуваат на сите видови рабови и преостанатиот се однесува на ситуации кога нема рабови.

Информацијата за бојата се извлекува со процесирање на регионите на сликата во HSV просторот на бои, при што како резултат се добива хистограм со 24 бинови за следните бои: црна, сива, бела, темно црвена, црвена, светло црвена, темно порткалова, портокалова, светло портокалова, темно жолта, жолта, светло жолта, темно зелена, зелена, светло зелена, темно тиркизна, тиркизна, светло тиркизна, темно сина, сина, светло сина, темно магента, магента и светло магента.

Фази хистограми за боја и текстура

Фази хистограми за боја и текстура (анг. *Fuzzy Color and Texture Histograms - FCTH*) во себе вклучуваат информации за текстурата од хистограм со 8 бинови добиени од фази систем кои користи Haar вејвлети [87]. Информациите за бојата се репрезентирани во хистограм со 24 бинови произведен од фази-поврзан систем [87]. Конечниот хистограм се состои од 8x24 = 192 региони/бинови.

Постапката на генерирање на хистограмот се состои од следните чекори. Прво, сликата се дели на предефиниран број на региони. Секој региони се претвора во YIQ просторот на бои и се трансформира со Haar вејвлети. Вредностите за f_{LH}, f_{HL}, f_{HH} се пресметуваат со помош на фази систем кои ги класифицира f коефициентите. Секој регион се класифицира во еден од осумте излезни бинови.

Следно, истиот регион се трансформира во HSV просторо на бои и се пресметуваат средните вредности за H, S и V вредностите во рамките на регионот. Овие вредности се користат како влез во фази систем кој генерира хистограм на бои со 10 бинови. Следниот фази систем ги употребува средните вредности на S и V, како и позицијата на бинот и го пресметува H каналот за бојата и генерира хистограм со 24 бинови. Истиот процес се повторува за сите региони во сликата. Овој дескриптор е сличен на CEDD со таа разлика што информацијата за текстурите овде се опфаќа преку Haar вејвлети.

Спротивна трансформација на дескриптори инваријантни на размер

Спротивна трансформација на дескриптори инваријантни на размер (анг. *Opponent Scale Invarian Feature Transform - OSIFT*) претставува варијанта на SIFT дескрипторот, кој пак, во комбинација со речник од визуелни зборови спаѓа меѓу најчесто користените пристапи за пребарување и класификација на медицински слики [61], [88]. Главните предизвици со кои се соочуваме при примена на овој пристап е начинот на избирањето на деловите/регионите од сликата од кои ќе се генерира дескрипторот и конструирањето на визуелниот речник.

Во рамките на докторската дисертација се употребува густо избирање на делови/региони, при што сликата униформно се дели на региони (анг. grid fashion). Растојанието меѓу регионите е фиксно и изнесува 6 пиксели, а избирањето е направено на различни размери на сликата ($\sigma = 1.2$ и $\sigma = 2.0$) [89]. Густото избирање на деловите/регионите се употребува, бидејќи некои од медицинските слики имаат низок контраст (на пример, радиографските слики) и тешко може да се применат детектори на клучни точки. На секој регион се пресметуваат SIFT и OSIFT дескриптори [46], [89], [90]. OSIFT дескрипторот ги опишува сите канали во спротивниот простор на бои (анг. *opponent color space*) со помош на SIFT дескриптори. Информацијата во O3 каналот дава опис на интензитетот, а другите канали ги опишуваат боите во сликата. Другите канали содржат информација за интензитетот, но, заради нормализацијата на SIFT дескрипторот тие се инваријантни на промени во интензитетот на светлината [89].

Клучниот аспект на визуелниот речник е неговото конструирање. Во [91] е даден исцрпен преглед на различни начини на креирање и репрезентација на визуелни речници. Во рамките на ова истражување се употребува *k-means* кластерирање/групирање на 250000 случајно избрани дескриптори од множеството на слики за обучување. *k-means* алгоритамот за кластерирање го дели просторот на дескриптори со минимизирање на варијансата меѓу предифинирано множество од k кластери/групи. Во нашето истражување k е поставено на 1000 и на тој начин се креира речник со 1000 визуелни зборови [92].

4.4.3 Модалитетот на медицински слики како поддршка при пребарување

Медицинските бази на податоци кои се употребуваат за пребарување или за образовни цели често содржат слики од различни модалитети, како рентгенски слики (анг. *X-ray*), слики со компјутерска томографија (анг. *CT scan*), ултразвук (анг. *ultrasound*) итн. Дополнителна компликација е што вообичаено сликите се направени под различни услови и според тоа прецизноста на нивните анотации е променлива и неконзистентна [93]. Ова е особено точно за слики кои се наоѓаат во онлајн ресурси, вклучувајќи ги онлајн колекциите на медицински трудови.

Модалитетот на сликата е основна визуелна карактеристика во медицината и може да се искористи во насока на подобрување на процесот на пребарување. Но, анотациите

или описите кои им се придружени на сликите, често не ги содржат информациите за нивниот модалитетот. Во заглавјето на DICOM постојат тагови кои објаснуваат кој дел од телото е прикажан на сликата, позицијата на пациентот и модалитетот [94]. Некои од таговите се поставуваат автоматски од системот кои ги прави сликите според протоколот кој се користи да се опфатат визуелните детали. Останатите тагови се поставуваат рачно од докторот или радиологот во процесот на документирање на сликата. Оваа процедура не може секогаш да се смета за надежна, бидејќи често се случува некои записи да не се внесат или да се комплетно неточни [95].

Базите на медицински слики содржат слики направени со различни медицински техники. Со цел сликите да се соодветно репрезентирани потребно е да се употребуваат различни техники за извлекување на визуелните карактеристики да се опфатат различните аспекти на сликите (на пример, текстура, форма, дистрибуција на бои итн.) [93]. Текстурата е особено важна, бидејќи е многу тешко да се класифицираат медицински слики само на основа на облик или со едноставна пресметка на дистрибуцијата на нијансите на сиво (повеќето од медицинските слики се претставени само со нијанси на сиво). Потребно е да се најде ефективна репрезентација на текстурата да се разграничат сликите од ист модалитет. Во тој контекст локалните дескриптори се клучни за репрезентација на сликите, бидејќи тие ги опфаќаат и опишуваат деталите на сликата, за разлика од глобалните, кои ја опшуваат сликата во целина. Локалните дескриптори овозможуваат решавање на проблемот на меѓу- и внатре-класната варијабилност на медицинските слики [92]. Токму тоа е причината поради која ги избравме горенаведените дескриптори.

4.5 Платформи за пребарување на информации

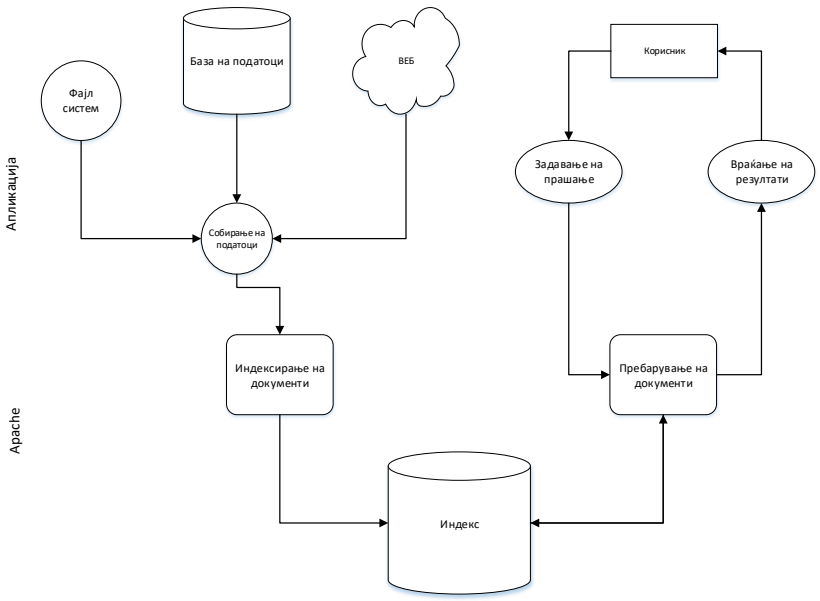
Сите концепти кои претходно беа образложени околу пребарувањето веќе се имплементирани во многу постоечки платформи за пребарување. Во продолжение ќе бидат објаснети некои од најкористените во контекст на пребарување на медицински информации.

4.5.1 Lucene

Apache Lucene [96] е бесплатна библиотека за пребарување на информации со отворен код. Библиотеката е оргинално развиена во програмскиот јазик Java од страна на Даг Катинг (анг. *Doug Cutting*). Моментално е поддржана од фондацијата за софтвер Apache и се издава под Apache лиценцата. Библиотеката има верзии кои поддржуваат други јазици како Object Pascal, Perl, C#, C++, Python, Ruby и PHP.

Оригиналната намена на библиотека е овозможување на пребарување во апликации кои имаа потреба од индексирање и пребарување низ големи текстуални документи, но, во практиката се покажала како корисна и за имплементација системи за пребарување низ интернетот, како и во рамките на поединечни веб-страници.

Во центарот на логичката структура на Lucene е парадигмата дека секој документ содржи полиња со текст. На тој начин се овозможува Lucene да се употребува независно од форматот на документите. Текст од PDF, HTML, Word, OpenDocument документи и многу други формати, може да се индексира, се додека постои начин како да се извлечи текстуалната содржината.



Слика 4.3: Стандардна интеграција со Lucene.

На Слика 4.3 е прикажана стандардна интеграција на Lucene библиотеката во некоја апликација.

4.5.2 Essie

Системот за пребарување Essie [97] е развиен во 2000 во Националната Библиотека за Медицина во САД (анг. *Nationanl Library of Medicine - NLM*) за поддршка на ClinicalTrials.gov, онлајн регистарот на истражувачки медицински случаи. Уште од почеток, Essie, беше дизајниран да користи синоними извлечени од системот за унифициран медицински јазик (анг. *Unified Medical Language System ®- UMLS*), со цел да се овозможи поквалитетно пребарување за крајните корисници на веб-страницата [98]. Речникот на UMLS содржи концепти од повеќе од 100 медицински речници [99]. Секој UMLS концепт може да има повеќе имиња, односно термини. Многу од корисниците на ClicalTrials.gov не биле многу запознати со медицинската терминологија која се употребува во медицинските документи и пребарувале со поошт речник на зборови. На пример, повеќе од документите на системот за срцеви удари не ја содржат фразата „срцев удар“ (анг. *heart attack*), туку го користат медицинскиот термин „myocardial infarction“. Според тоа, Essie, системот се обидува да воведе пребарување со автоматско детектирање на концепти и пребарување на нивна основа (анг. *concept-based searching*).

Essie системот се состои од две посебни фази: индексирање и пребарување. Индексирањето ги идентификува и зачувува позициите на секој збор што се појавува во податочното множество. Пребарување се извршува со помош на техника на проширување на прашањата, при што се генерираат зборови или фрази, кои треба да се додадат во прашањата. Во системот постојат различни начини на генерирање на зборовите и тоа:

- Проширување со зборови: Овој метод вклучува различни варијанти на истите зборови во прашањето (еднина, множина и сл.).
- Проширување со концепти: Вклучува синоними на зборовите во прашањето.
- Релаксирано проширување (анг. *Relaxation Expansion*): Овој метод го разделува прашањето на повеќе фрагменти и додава синоними за сите фрагменти.
- Проширување со загуба (анг. *Lossy Expansion*): Овој метод го разделува прашањето на повеќе фрагменти и додава синоними за сите фрагменти, но исто така, може да испушти некои фрагменти.

Користењето на овие техники за проширување на прашањата станува проблематично за подолги прашања, бидејќи системот треба да преработи поголем број на комбинации за проширување. Во практиката се покажало дека интерактивните системи употребуваат релаксирано проширување, додека системите без повратна врска од корисникот се користи lossy проширување. Релаксираното проширување работи подобро кај интерактивните системи, бидејќи корисниците даваат повратна информација и го реформулираат прашањето соодветно, доколку не се задоволни од резултатите.

4.5.3 Terrier IR

Terrier (*Terabyte Retriever*) IR [100] е проект, започнат во рамките на Универзитетот на Глазгов во 2000 година, со цел да се развие флексибилна платформа за брз развој на апликации за пребарување на информации со голем капацитет, како и платформа за истражување и експериментирање со најновите методи во областа на пребарување на информации. Terrier IR е со отворен код изработен во програмскиот јазик Java и содржи различни модели за пребарување информации, многу базирани на DFR рамката (*Divergence FROM Randomness*) [1]. DFR рамката подразбира: колку е поголема дивергенцијата на фреквенцијата на терминот t во рамките на документот во однос на дистрибуцијата во остатокот од податочното множество, толку повеќе информација носи терминот t во документот. Terrier содржи преку 50 модели за утежнување, меѓу кои и популарните BM25, TF-IDF и сл.

Архитектурата на Terrier е дизајнирана да овозможи ефикасно скалирање на големината на податочните множества и при тоа, овозможува оперирање во централизирана или дистрибуирана околина. Главните податочни структури на Terrier се директниот индекс (анг. *direct index*), индексот на документи (анг. *document index*), инвертираниот индекс (анг. *inverted index*) и лексиконот (анг. *lexicon*). Директниот индекс ги чува

[1] Повеќе информации за DFR во Terrier може да се најдат на следниот линк: http://ir.dcs.gla.ac.uk/terrier/description.html.

термините кои се појавуваат во секој документ и соодветните фреквенции. Овој индекс се употребува во фазата на проширување на прашањата. Индексот на документи чува информација за должината на документите и покажувачи до директниот индекс. Инвертираниот индекс чува податоци кој термин во кој документ се појавува, а лексиконот ги чува сите зборови кои се појавуваат во податочното множество.

Во процесот на индексирање секој документ се парсира и дели на токени, односно се извлекуваат поединечните зборови во документот. Зависно од поставувањата на апликацијата, се филтрираат стоп зборовите, односно зборовите кои многу често се појавуваат и немаат информативна вредност, а потоа се применува стемирање (анг. *stemming*) со што сите зборови се доведуваат во нормализирана форма.

Процесот на пребарување се одвива на сличен начин. Даденото прашање се процесира на тој начин што прво се бришат стоп-зборовите и се применува стемирање, зависно од поставувањата на апликацијата. Доколку не се поставени детали околу моделите на утежнување и начинот на пребарување (на пример, проширување на прашањето или анализа на линкови, ако станува збор за веб сценарио), тогаш Terrier автоматски ги поставува овие детали. Во случај да се употребува проширување на прашањето, тогаш автоматски се бира модел за утежнување и бројот на документи што треба да се анализираат да се добијат термините за проширување.

Terrier постојано се надополнува со нови техники кои овозможуваат покомплексно пребарување, како различни модели за дефинирање на структурата на документот, методи за нормализација [101], техники за обработување на природни јазици [102], извлекување на информации [103] итн.

4.6 Процес на мултимодално пребарување на медицински слики

Мултимодалното пребарување на медицински слики може да го поделеме на два вида на пребарување: текстуално базирано и содржински базирано пребарување. Мултимодалниот дел всушност се третира како проблем во фаза на постпроцесирање или фузија на податоците. Текстуално базираното пребарување ги користи карактеристиките генерирани од текстуална репрезентација на сликите. Според тоа, ова пребарување треба да функционира на основа на контекстот во кои се прикажани сликите. Додека, содржински базираното пребарување врши пребарување на основа на визуелните карактеристики кои се генерирани од визуелната содржина на сликите во базата на податоци, што значи ова пребарување ќе се извршува според тоа што е визуелно прикажано во тие слики.

Текстуално базирано пребарување

Во овој вид на пребарување текстуалните репрезентации на сликите и прашањето се процесираат на начинот прикажан во поглавје 4.4.1 и се применува модел за утеж-

нување со цел да се најдат релевантни слики. Моделот за утежнување се употребува да се пресмета нумеричка вредност, која ја претставува сличноста на дадена слика во однос на прашањето поставено од корисникот. Колку е поголема пресметаната вредност - толку е порелевантна сликата. Откако вредноста ќе се пресмета за сите слики од множеството, тие се подредуваат во опаѓачки редослед во однос на пресметаната вредност и се враќаат. Имајќи предвид дека постојат различни модели на утежнување, ние прво експериментираме со текстуално базирано пребарување со шест различни модели на утежнување: PL2 [104], BM25 [104], BB2 [104], DFR-BM25 [104], TF-IDF [105], DirichletLM [106]. Откако го определуваме најсоодветниот модел за овој проблем, тогаш тој модел го вклучуваме во мултимодалното пребарување на медицински слики.

Текстуално базираното пребарување може да се подобри со помош на метод за проширување на прашањата. Поентата на методот за проширување на прашањата има за цел да ги модифицира прашањата да се обезбедат подобри резултати. Во овој контекст употребувавме псевдо-релевантна повратна врска која ги зема предвид првите n инцијално вратени документи за дадено прашање и ги наоѓа најинформативните m термини во нив. Најдените термини се додаваат во оригиналното прашање и се прави уште едно пребарување, но сега со модифицираното прашање. Резултатите добиени со тоа пребарување се конечните резултати од текстуално базираното пребарување.

Содржински базирано пребарување

Главната цел овој вид на пребарување е скалабилноста и можноста за пребарување во реално време. Според тоа се осврнуваме на пристапот презентиран во [107] т.е. енкодирање на векторите на карактеристиките со метод на квантизација за побрзо пребарување. Основната идеја позади овој пристап е да се земат генерираните карактеристики за сите слики во множеството и да се кодираат во компактен вектор, кој ќе овозможи побрзо пребарување. Конкретно, за даден d димензионален влезен вектор, сакаме да генерираме репрезентација на слика преку b битови, така што најблиските соседи во некодираниот вектор може да се пронајдат во множество од n кодирани вектори.

4.7 Бази на медицинско знаење

Во рамките на истражувањето употребуваме различни медицински бази на знаење за поддршка на процесот на пребарување:

- **MeSH:** Medical Subject Headings (MeSH) претставува строго контролиран речник кој се користи за индексирање на трудови и книги од природните науки и често се користи при пребарување [108]. Речникот е креиран и одржуван од Националната Библиотека за Медицина во САД (анг. *National Library of Medicine - NLM*). Речникот се употребува во базата MEDLINE/PubMed.
- **UMLS:** Unified Medical Language System (UMLS) претставува комбинација од различни медицински речници [98]. Множеството има структура на мапирање меѓу

различните речници и во таа смисла овозможува нормализирање на различните терминологии. UMLS се смета за релевантна и обемна онтологија на биомедицински концепти. Всушност, UMLS се состои од повеќе бази на податоци и софтверски алатки.

Овие бази се меѓу најголемите медицински бази на знаење кои воопшто постојат и затоа се користат во истражувањето.

4.8 Бази за евалуација

Во рамките на истражувањето ги употребувавме базите од ImageCLEF развиени како дел од Cross-Language Evaluation Forum (CLEF), што претставува организација чија цел е промоција на истражување, иновација и развој на системи за пристап до информации со посебен акцент на информации во различни јазици, модалитети и структура. ImageCLEF како подмножество на оваа организација е фокусирано на истажувања поврзани со анотација и пребарување на слики. Мотивирани од потребата за поддршка на корисници со различни јазични побарувања над експоненцијално растечко множество на мултимедијални податоци, целта на ImageCLEF е поддршка на развој, анализа, индексирање, класификација и пребарување на визуелни информации, преку имплементирање на соодветната инфраструктура за евалуација на системите за пребарување на визуелни информации. Поконктрено, целта на ImageCLEF е овозможување на реискористливи ресурси кои може да се употребуваат за еваулација на ваквите системи [109]. Во тој контекст се развиени стандардизираните бази за евалуација кои беа употребувани во рамките на истражувањата на докторската дисертација.

4.8.1 База за евалуација на методи за пребарување на медицински трудови

Експериментите поврзани со пребарување на медицински трудови се извршуваат над базите со медицински трудови од ImageCLEF (case-based retrieval subtask) 2012 и 2013 година. Базите содржат по 74 654 медицински трудови (случаи), главно трудови извлечени од PubMed. Секој труд е зачуван во XML документ со определена структура која ги содржи следните делови: наслов, апстракт, содржина на трудот и текстуални описи на сликите кои се користат во трудот. На Слика 4.4 е прикажана XML структурата во кои се складирани медицинските трудови.

Главната разликата меѓу ImageCLEF 2012 и 2013 е бројот на дадени прашања, а тие се 26 и 35, соодветно. Прашањата се состојат од кратки описи на некој случај. Во продолжение интегрално се прикажани неколку прашањата дадени за евалуација на методи за пребарување на медицински трудови:

- **Прашање 1**. *A 50-year-old man with severe right flank pain and hematuria. Renal ultrasound shows a markedly echogenic lesion with a posterior acoustic shadow measuring about 8x10mm in the right kidney.*

Слика 4.4: XML структура на медицински труд од ImageCLEF базите за евалуација.

- **Прашање 2**. *A 49-year-old woman with a prolapsed mass in the opening of her urethra. Pelvic CT shows a heterogeneously enhanced mass on the female urethra. Pathology shows ramifying papillae, high nuclear/cytoplasmic ratio, and brisk mitotic activity.*
- **Прашање 3**. *A 56-year-old woman with Hepatitis C, now with abdominal pain and jaundice. Abdominal MRI shows T1 and T2 hyperintense mass in the left lobe of the liver which is enhanced in the arterial phase.*

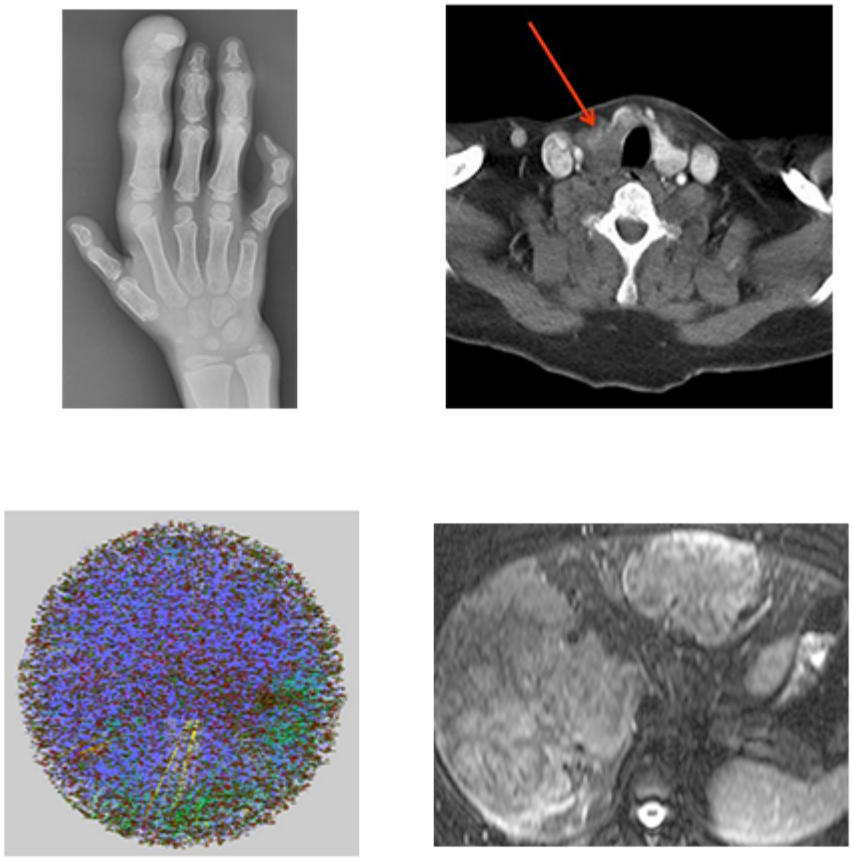
Секое прашање е придружено од 2-3 слики со цел евалуација на мултимодално или содржински-базирано пребарување на трудови.

4.8.2 База за евалуација на методи за пребарување на медицински слики

Експериментите поврзани со пребарување на медицински слики се извршуваат над базите со генерички медицински слики од ImageCLEF (ad-hoc retrieval subtask) 2011, 2012 и 2013 година, прикажани во Табела 4.1.

Множествата се содржат од одреден број на генерички медицински слики кои се извлечени од PubMed медицински трудови во кои се употребуваат. На Слика 4.5 се прикажани неколку примероци од базите на слики. Покрај сликите, дадени се одреден број на прашања за евалуација. Прашањата се состојат од неколку клучни зборови и/или неколку медицински слики. Во продолжение интегрално се прикажани неколку од прашањата дадени за евалуација на методи за пребарување на медицински слики:

- **Прашање 1**. *osteoporosis x-ray*
- **Прашање 2**. *nephrocalcinosis ultrasound images*
- **Прашање 3**. *lymphoma MRI images*



Слика 4.5: Пример слики од множествата за евалуација на алгоритми за пребарување на медицински слики на ImageCLEF.

4.8.3 База за евалуација на методи за класификација на медицински слики според модалитет

Во рамките на ImageCLEF множествата постојат и колекции за евалуација на методи за класификација на медицински слики според модалитетот. Имајќи предвид дека една од целите на докторската дисертација е вклучување на методи за класификација на слики според модалитет во процесот на пребарување, потребно е да се евалуираат и тие делови од истражувањето над некое множество. Во таа насока се користат ImageCLEF 2011, 2012 и 2013 (modality classification subtask) множествата за класификација на слики според модалитет. Имено, секое од множествата се состои од одреден број на

Табела 4.1: Бројот на слики и прашања во базите на ImageCLEF од 2011, 2012 и 2013 година

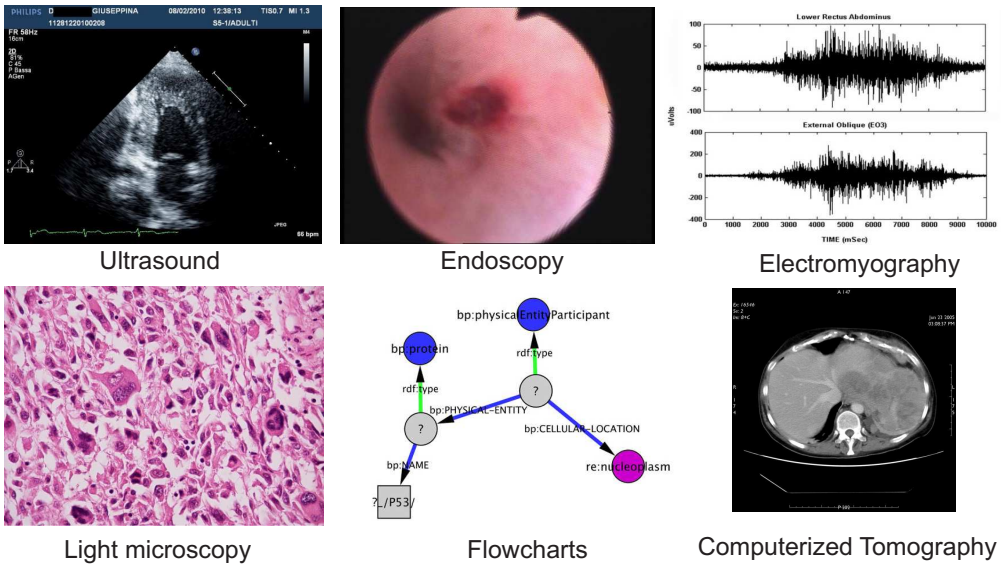
База	#Слики	#Прашања
2011	230088	30
2012	306539	26
2013	306539	35

Табела 4.2: Детали за базите за евалуација на алгоритмите за класификација на слики според модалитет

База	#Слики за обучување	#Слики за тестирање	#Класи
2011	988	1024	18
2012	1001	1000	31
2013	2901	2582	31

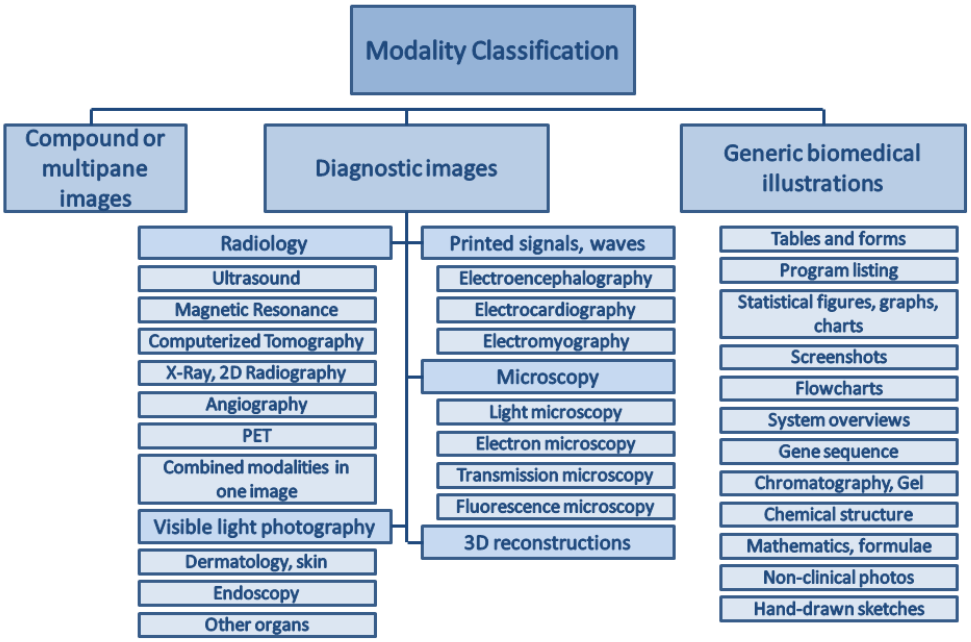
слики за обучување, валидација и тестирање. Бројот на модалитети (класи) е различен зависно од множеството. Во Табела 4.2 се прикажани деталите за множествата.

Сликите за обучување се употребуваат за тренирање на класификаторите и за нив е однапред познат модалитетот. Сликите за валидација се слики за кои модалитет е однапред познат и кои се користат за оптимизирање на параметрите на класификаторите. Сликите за тестирање се сликите за кои не се познати модалитетите и перформансите на класификаторите се одредуваат врз основа на тоа како ќе ги класифицираат сликите. На Слика 4.6 се прикажани неколку примероци од различни класи со цел да се воочат визуелните разлики.



Слика 4.6: Пример слики од множествата за евалуација на алгоритми за класификација на слики според модалитет на ImageCLEF.

Класите за множествата 2012 и 2013 се прикажани на Слика 4.7.



Слика 4.7: Хиерархиска организација на класите од множествата ImageCLEF 2012 и 2013.

4.9 Метрики за еваулација

Две главни метрики за одредување на ефикасноста на алгоритми за пребарување се *прецизност* (анг. *precision*) и *одзив* (анг. *recall*). Прецизноста е мерка која кажува колку се точни вратените резултати на системот, додека одзивот се однесува на опфатноста на релевантните документи од системот. Прецизноста се дефинира како процент на релевантни документи добиени при пребарувањето од вкупниот број на добиени документи:

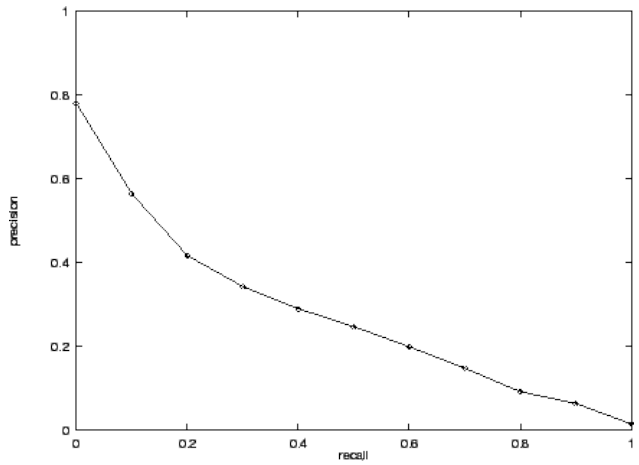
$$precision = \frac{|R \bigcap p|}{|p|} \tag{4.11}$$

каде, R е множеството од сите документи кои се релевантни, p е множество на сите документи кои вратени при пребарувањето. Со $|X|$ е означена големина (бројот на елементи) на множеството X. Одзивот е определен како однос меѓу бројот на вратени релевантни документи и вкупниот број на релевантни документи во множеството:

$$precision = \frac{|R \bigcap p|}{|R|} \tag{4.12}$$

Според претходно дадените дефиниции може да се заклучи дека двете метрики се подеднакво важни при евалуирање на системи за пребарување. Добар систем за пребарување треба да обезбеди соодветен баланс меѓу прецизноста и одзивот, бидејќи тие две метрики се меѓусебно зависни. Генералната зависност меѓу прецизноста и одзивот

може да се видат на Слика 4.8 каде е прикажана пример крива за прецизност-одзив за пребарување.



Слика 4.8: Пример прецизност-одзив крива за пребарување.

Кога системот треба да врати повеќе резултати, очигледно се зголемува веројатноста повеќе релевантни документи да се појават во добиените резулати, што значи се зголемува одзивот. Но, во исто време се намалува прецизноста на системот, бидејќи се зголемува веројатноста во добиените резултати да се појават и поголем број на нерелевантни слики. Ако системот врати помал број на документи, но поточни, тогаш се зголемува прецизноста на системот, но се намалува одзивот, бидејќи ќе останат уште многу релевантни документи кои нема да се појават во резултатите.

4.9.1 Средна прецизност

Вообичаено мерките за прецизност и одзив се комбинираат, бидејќи корисниците би биле заинтересирани за висока прецизност и висок одзив. Една таква мерка е средната прецизност (анг. *average precision*) и најчесто се користи за евалуација на системите за пребарување [110]. За висока средна прецизност неопходно е системот да има висока прецизност, но истовремено потребно е и висок одзив. Конкретно, потребно е системот да рангира повеќе релевантни документи што е можно повисоко во резултатите.

Средната прецизност се пресметува на следниот начин:

$$AP = \frac{1}{R} \sum_{r=1}^{N} P(d_r) \cdot Rel(d_r) \tag{4.13}$$

$$R = \sum_{r=1}^{N} Rel(d_r) \tag{4.14}$$

каде, N е бројот на документи во множеството, d_r е документ на ранг r, $P(d_r)$ е прецизноста за ранг r, R е бројот на релевантни документи во множеството за даденото прашање, а $Rel(d_r)$ е 1, d_r е релевантен документ за даденото прашање, а 0 во спротивен случај.

Со цел робусно тестирање на ефективноста на даден систем за пребарување на информации, секој систем се тестира на определен број прашања (анг. *topics*). Средната прецизност се пресметува за секое прашање и за сите пресметани вредности се агрегира. Оваа метрика се вика усреднета средна прецизност (анг. *mean average precision*) и претставува стандардна метрика за евалуација на системи за пребарување на информации.

Глава 5

Архитектура

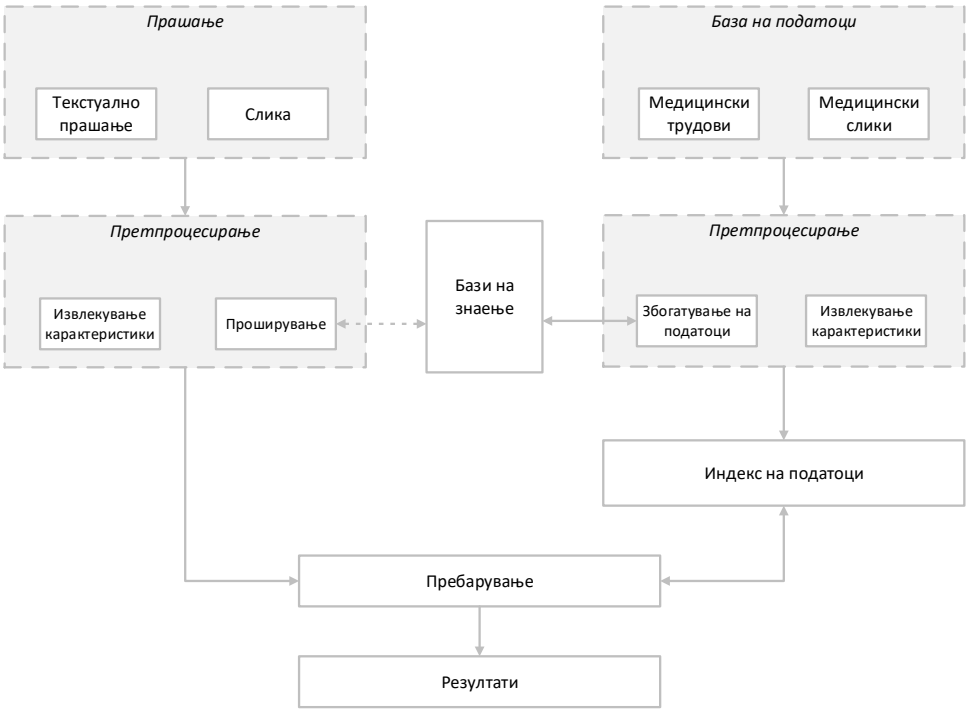
Во оваа глава е презентирана архитектурата на системот за мултимодално пребарување на медицински документи. Системот се состои од повеќе подсистеми, при што секој дел е фокусиран на решавање на одреден проблем. Прво е презентирана генералната рамка на системот, а потоа подсистемите. Првиот подсистем се фокусира на решавање на проблемот на класификација на медицински слики според модалитет. Вториот подсистем нуди решение за пребарување на медицински слики со мултимодални податоци. Следните подсистеми нудат решенија за различни начини на пребарување на медицински трудови.

5.1 Генерална рамка

Главна цел на истражувањата на докторската дисертација се методите за пребарување на медицински документи со помош на мултимодални (разнородни) податоци. За истражувањата кои ги направивме во текот на работата имплементиравме комплексен систем за пребарување на медицински документи. Генералната рамка на системот е прикажана на Слика 5.1.

Системот има две главни линии на работа и тоа: справување со податоците во базата на податоци и справување со прашањата од корисникот. На десната страна на дијаграмот е прикажан делот од системот што ја обработува базата на податоци. Базата на податоци на системот се состои од разнородни медицински податоци, односно од медицински трудови и слики. Овие податоци се процесираат од системот во фазата на претпроцесирање. Процесирањето се состои од извлекување на карактеристики од документите, зависно од видот на документот кој се обработува т.е. извлекувањето на карактеристики за медицинските трудови е различно од извлекување на карактеристиките за медицинските слики. Покрај извлекување на карактеристики, во оваа фаза системот врши и збогатување на податоците со помош на бази на знаење или други компоненти. Откако претпроцесирањето ќе заврши, податоците се индексираат.

На левата страна на дијаграмот е прикажан делот од системот кој се справува со прашањата од корисникот. Корисникот може да постави (како прашање) текстуален



Слика 5.1: Генерална рамка на систем за пребарување на медицински документи со мултимодални податоци.

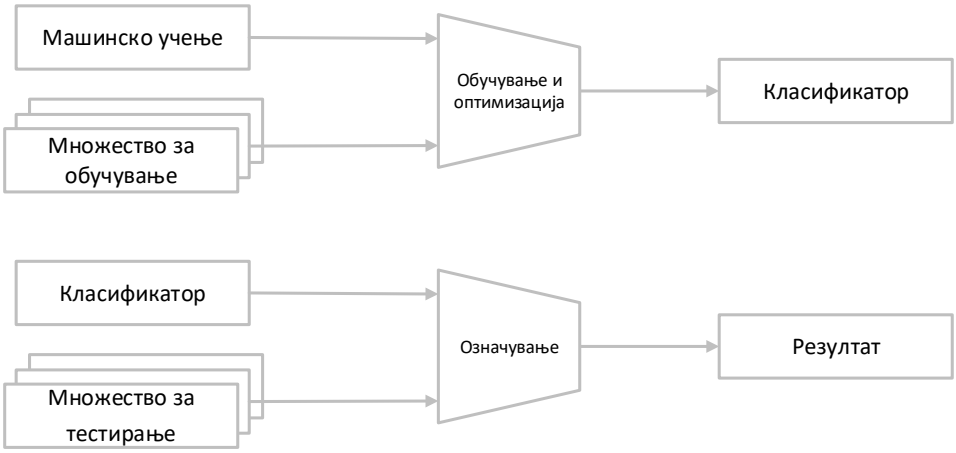
опис и/или слика. Овој влез, системот го претпроцесира на тој начин што извлекува карактеристики за дадените податоци. Дополнително, системот го проширува прашањето на различни начини, меѓу кои и со помош на бази на знаење. На крај, процесираното прашање се испраќа до делот за пребарување кој го користи индексот на податоци да ги пресмета и прикаже резултатите од пребарувањето.

Овој дијаграм претставува генерална рамка на нашиот систем за пребарување. Во практика системот е поделен на повеќе подсистеми кои заедно функционираат и ја извршуваат целокупната работа. Во продолжение се опишани неговите подсистеми.

5.2 Подсистем за класификација на медицински слики според модалитет

Подсистемот за класификација на медицински слики според модалитет се обидува на автоматски начин да го одреди модалитетот на дадена слика. Самиот процес на класификација се состои од два главни чекори кои се прикажани на Слика 5.2.

Во првата фаза се обучува и оптимизира класификатор со помош на множество за обучување. Во практика, од податоците во множеството се извлекуваат карактеристики и истите се задаваат како влез во процесот на обучување. Во втората фаза, добиениот класификатор се применува на множество за тестирање и се евалуираат неговите перформанси.



Слика 5.2: Архитектура на подсистемот за класификација на медицински слики според модалитет.

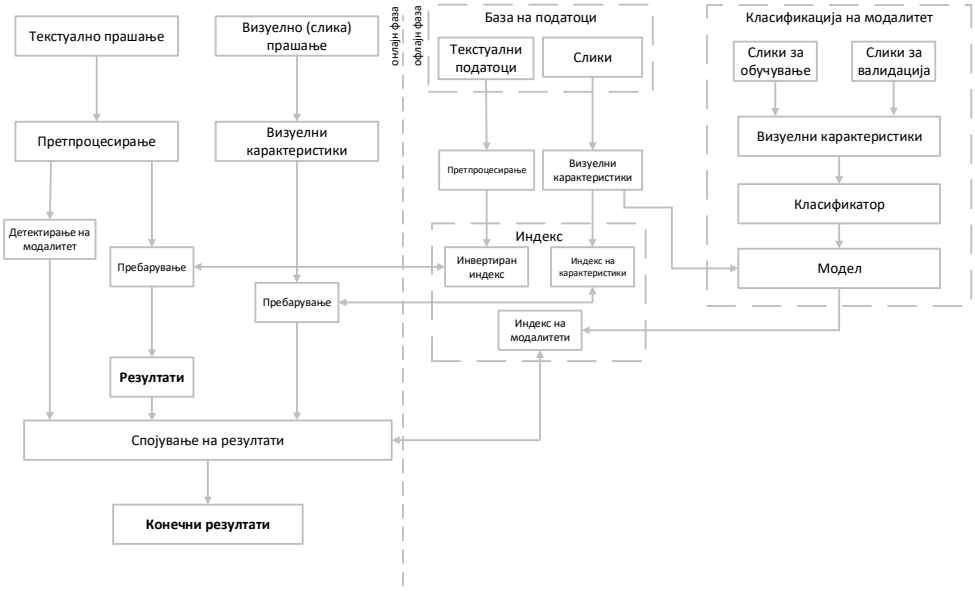
Вршени се истражувања кои укажуваат на тоа дека доколку се користат визуелни карактеристики извлечени со помош на различни дескриптори, а со тоа и опфаќаат различен вид на визуелни информации, се добиваат подобри резултати во контекст на класификација отколку со карактеристики извлечени од поединечни дескриптори [60], [92], [59]. Дополнително, текстуалните карактеристики носат информации кои можат да се искористат во процесот на класификација. Согласно на тоа во нашиот подсистем за класификација на слики овозможивме комбинација на горенаведените визуелни и текстуални карактеристики извлечени од трудовите во кои се поставени сликите. Во системот имплементирани се следните визуелни карактеристики: локални бинарни шаблони, дескриптор за насока на бои и рабови, фази хистограми за боја и текстура и спротивна трансформација на дескриптори инваријантни на размер (анг. *Opponent Scale Invariant Feature Transform - OSIFT*). За текстуалните информации се употребува bag-of-words репрезентација.

5.3 Подсистем за пребарување на медицински слики

Подсистемот за пребарување на медицински слики кој го развивме во рамките на докторската дисертација е комплексен и се состои од повеќе компоненти. Дијаграмот од архитектурата на подсистемот е прикажан на Слика 5.3. Функционалноста на подсистемот е поделена на онлајн и офлајн фаза.

Текстуалната репрезентација на сликата е извлечена од трудовите каде се содржат сликите. Генерираните текстуални репрезентации се праќаат на делот за индексирање, кој извршува некои стандардни претпроцесирачки техники како: токенизација, бришење на стоп зборови и стемирање. Откако овие процеси ќе завршат се креира инвертиран индекс.

Паралелно на текстуалното процесирање функционира и содржински базиран прис-



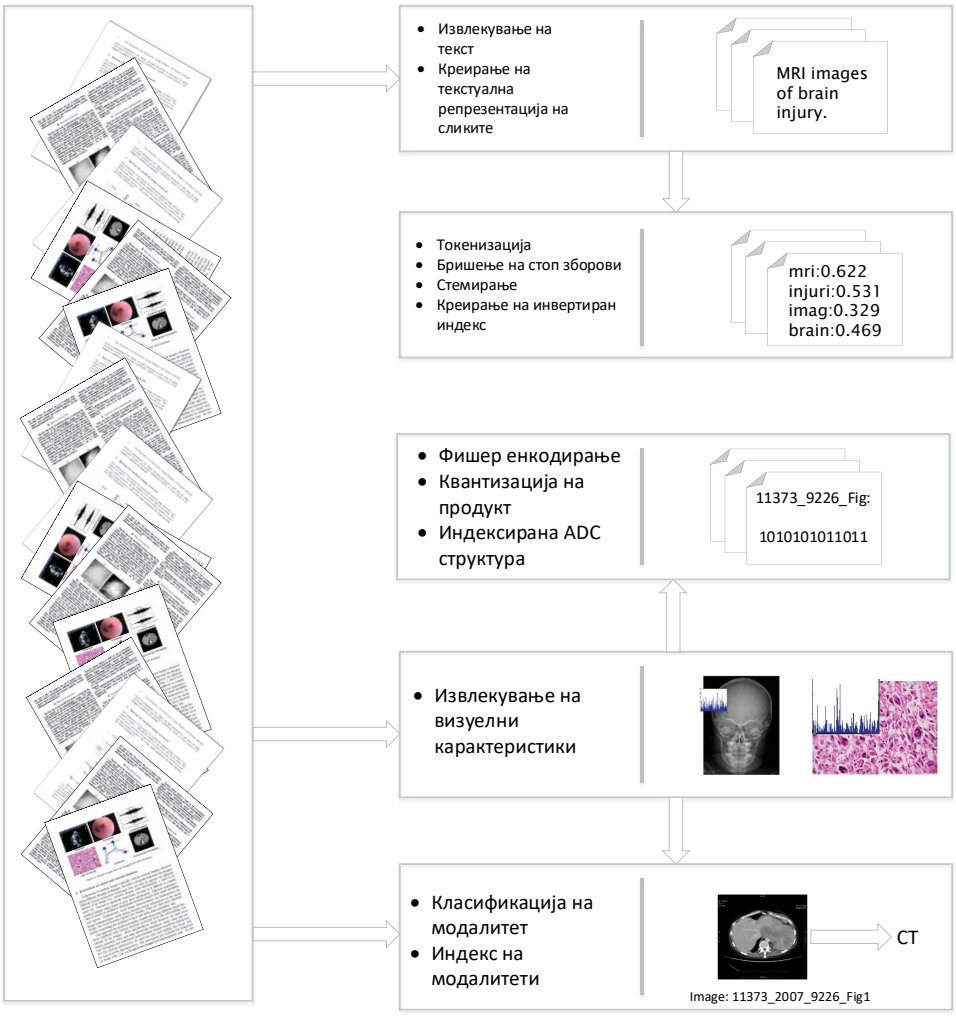
Слика 5.3: Архитектура на подсистемот за пребарување на медицински слики со текстуални и/или визуелни податоци.

тап. На сите слики кои се наоѓаат во медицинските трудови им се извлекуваат визуелните карактеристики, со цел подсистемот да генерира карактеристики со кои ќе се опише нивната визуелна содржина. Генерираните карактеристики се чуваат во индексот на карактеристики и се користат за содржински базираното пребарување.

Од друга страна, визуелни карактеристики се извлекуваат од сликите со цел обучување на класификатори за класификација на модалитетот, односно за препознавање на модалитетот на дадена слика. За таа цел, се користи посебно множество на слики за обучување. Обучениот класификатор се оптимизира на валидациско множество на анотирани слики т.е. слики со претходно познати модалитети. Штом се добие оптималниот модел, подсистемот ги класифицира сите слики во базата на податоци на основа на нивните визуелни карактеристики. На крајот за секоја слика во базата на податоци, подсистемот генерира соодветен модалитет, кој се чува во индексот на модалитети. Индексот се чува во *csv* документ и се вчитува во меморија при стартување на подсистемот за побрз пристап.

Онлајн фазата го репрезентира процесот кога се прави пребарување со помош на подсистемот. Подсистемот како влез добива текстуални и/или визуелни (слика) прашања. Текстуалните прашања се претпроцесираат и се извршува пребарувањето. Текстуалните прашања често содржат клучни зборови што го опшуваат модалитетот на сликата која ја бараат. Во оваа фаза, подсистемот го извлекува посакуваниот модалитет преку едноставно детектирање на соодветните клучни зборови, кои се однесуваат на модалитет. Кога корисникот ќе даде слики како прашање, тогаш подсистемот ги извлекува визуелните карактеристики и ги употребува за пребарување. Во случај на мултимодално пребарување, подсистемот врши повторно рангирање на вратените слики со помош на индексот на модалитет и ги пресметува конечните резултати.

На Слика 5.4 е претставен секвенцен дијаграм на подсистемот за пребарување на слики. Дијаграмот претставува поапстрактна репрезентација на податоците и концептите кои се употребуваат во подсистемот.



Слика 5.4: Секвенцен дијаграм на подсистемот за пребарување на медицински слики.

5.4 Подсистем за пребарување на медицински трудови со спојување на зборови и медицински концепти

Архитектурата на делот од подсистемот за пребарување на медицински трудови со спојување на зборови и медицински концепти е прикажана на Слика 5.5. Архитектурата се состои од два независни дела, кои функционираат паралелно, а на крај нивните поединечни резултати се спојуваат.

Првиот метод на пребарување (на основа на зборови) на медицински трудови функционира на стандардниот начин на индексирање и пребарување на текстуални документи. Медицинските трудови се претпроцесираат. Претпроцесирање се состои од три

Слика 5.5: Архитектура за пребарување на медицински трудови со спојување на зборови и концепти.

фази: токенизација, бришење на стоп зборови и стемизација. Откако ќе се претпроцесираат текстуалните репрезентации се креира индекс (анг. *inverted index*) со чија помош потоа се изведува пребарувањето. Во фазата на пребарување, текстуалното прашање се претпроцесира на истиот начин како и медицинските трудови. Се применуваат модели на утежнување да се пресмета релевантноста на секој медицински труд во однос на даденото текстуално прашање. Откако ќе се пресмета релевантноста, тогаш сите документи се подредуваат и враќаат.

Во вториот метод (пребарување на основа на концепти) се врши анализа на текстот во медицинските трудови и поставените прашања со цел да се извлечат медицинските концепти кои се појавуваат во нив. Извлекувањето на медицинските концепти, односно мапирањето може да се направи со некои постоечки алатки, библиотеки или сервиси како Metamap [27], MeshUp [29] итн. Проблемот кој се јавува во овој начин на репрезентација е во однос на пресметувањето на релевантноста на трудовите и прашањата. Бројот на заеднички зборови кои медицинските трудови и прашањата ги имаат, не дава директна индикација за тоа колку имаат заеднички медицински концепти. Конкретно, доколку со помош на алатките се обидеме да го мапираме зборот *x-ray*, тогаш за него ќе добиеме шест различни медицински концепти во кои тој може да припаѓа, доколку, пак, ако се обидеме да го мапираме терминот *lung x-ray*, алатките ќе извлечат само еден концепт [26]. Ова значи, дека бројот на заеднички зборови и заеднички медицински концепти за два документи, не е линеарно пропорционален. Затоа, во рамките на овој дел од истражувањето како влез на алатките за мапирање се предава целосната текстуална содржина на трудовите и прањата, а се земаат предвид сите концепти кои се добиваат како резултат. Со добиените концепти се креираат нови репрезентации за трудовите и прашањата. Врз овие репрезентации се применуваат стандардни методи на

индексирање и пребарување. Овој пристап може да се смета како еден вид претпроцесирање на стандардното пребарување. Според [26] ваквата репрезентација на документи кои содржат медицински текст треба да даде добри резултати при индексирање и пребарување низ документите. Од тука следи идејата за креирање на ваква репрезентација за медицинските трудови и прашања.

Во последната фаза се комбинираат (спојуваат) резултатите од двата претходни методи. Спојувањето се одвива во два дела. Во првиот дел се прави нормализација на поединечните податоци со цел доведување на резултатите во ист опсег на вредности. Во вториот дел може да се употребуваат кои било методи на доцна фузија [28]. Овде се користи линеарна комбинација на нормализираните резултати, бидејќи овозможува модуларност, скалабилност и лесна контрола над тоа кој дел колкаво влијание има врз пресметаните резултати. Резултатите добиени во оваа фаза се конечните резултати.

5.5 Подсистем за пребарување на медицински трудови со проширување на прашања

Во овој дел од истражувањето фокусот е ставен на различни методи за проширување на прашањата при пребарување на медицински трудови. Самото процесирање и пребарување низ трудовите се извршува на стандарден начин користејќи техники за индексирање и пребарување имплементирани во многу платформи за пребарување (како Terrier IR).

Опфатени се следните методи на проширување:

5.5.1 Проширување со MeSH термини

Целта на овој дел од подсистемот е да се користи надворешна алатка за мапирање, која може да го анализира прашањето и да ги извлече MeSH термините поврзани со истото. Алатките кои можат да направат такво мапирање се MTI (анг. *Medical Text Indexer*) [111] или MeshUp [29] и двете развиени од NLM. Извлечените термини се додаваат на оригиналното прашање и пребарувањето се извршува на основа на истото.

5.5.2 Проширување со UMLS термини

Целта на овој дел од подсистемот е слична на претходниот со тоа што овде се користат UMLS концепти при проширувањето. Постојат алатки кои можат да мапираат UMLS концепти од даден текст. Најпозната алатка е развиена од NLM и се вика MetaMap [27]. Откако концептите се извлекуваат, тие се додаваат на оригиналното прашање и пребарувањето се извршува на основа на истото.

5.5.3 Псевдо-релевантна повртана врска

Проширување на прашања со пседво-релевантна врска (анг. *pseudo-relevance feedback*) веќе претставува индустриски стандард при пребарување на информации [112]. Процесот се одвива на тој начин што прво се пребарува со оригиналното прашање. Откако ќе се добијат првичните резултати, се анализираат првите n документи за да се најдат најинформативните m термини. Параметрите n и m може да се нагодуваат зависно од условите. Најдените термини се додаваат на оригиналното прашање и повторно се извршува пребарување со истото. Добиените резултати од второто пребарување се конечните резултати.

5.6 Подсистем за пребарување на медицински трудови со генерички бази за знаење

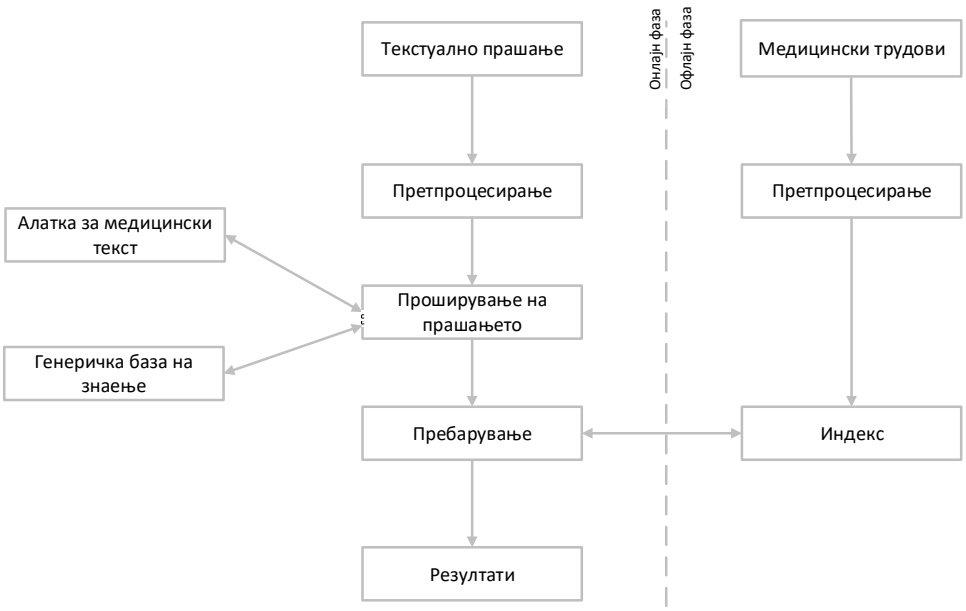
Овој подсистем се фокусира на пребарување на медицински трудови со помош на генерички бази за знаење. Во овој пристап употребуваме алатки за детектирање на медицински термини во поставените прашања до системот, а потоа со генерички бази за знаење наоѓаме синоними за тие термини. Идејата на овој пристап е дека генеричките бази за знаење (на пример, Freebase [113]) се одржуваат од огромни групи на корисници со различни профили.

Дијаграмот на подсистемот е прикажан на Слика 5.6. Подсистемот е поделен на онлајн и офлајн фаза. Офлајн фазата се состои од претпроцесирање и индексирање на медицински трудови. Претпроцесирањето и индексирањето е изведено на ист начин како во Глава 4.6 во делот со пребарување на основа на зборови.

Во онлајн фазата корисникот го дава својот влез во системот во форма на опширно прашање. Прво, прашањето се процесира на ист начин како и медицинските трудови. Потоа, подсистемот се обидува да го прошири прашањето со дополнителни термини на следниот начин:

1. Од прашањето се извлекуваат медицински термин со помош на специјализирана алатка за обработка на медицински текстови
2. Извлечените медицински термини се пребаруваат во генеричката бази за знаење и се извлекуваат нивните синоними
3. Извлечените синоними се додаваат во оригиналното прашање

Откако ќе се добие променетото прашање се применуваат модели на утежнување на ист начин како во Глава 4.6 во делот со пребарување на основа на зборови и се добиваат конечните резултати.



Слика 5.6: Архитектура за пребарување на медицински трудови со генерички бази за знаење.

Глава 6

Експерименти и дискусија

Во оваа глава се презентирани експериментите изведени во рамките на истражувањето. Образложено е како сите делови вклучени во експериментите се поставени со нивните технички детали. За секој од експериментите се дефинирани прашањата кои се обидуваме да ги одговориме. Резултатите од експериментите се прикажани со метриките од интерес и истите опиширно се дискутирани со цел да се даде одговор на поставените експериментални прашања.

6.1 Подсистем за класификација на медицински слики според модалитет

6.1.1 Експериментални поставувања

Во овој дел се објаснуваат техничките детали и подесувањата на различните алгоритми и алатки што се користат во делот за класификација на медицински слики според модалитет, како и начините на кои тие се оптимизирани и валидирани. Прво, се презентирани деталите за текстуалните карактеристики. Потоа се образложени деталите околу алгоритмите за машинско учење кои се употребени за класификација на сликите според модалитет. Следно, опишани се деталите за оптимизација на визуелните дескриптори. Прикажан е методот за спојување на податоците со цел да се добие подобра предиктивна моќ на класификаторот. На крај се дефинирани експерименталните прашања кои сакаме да ги одговориме во овој дел од истражувањето.

Поставување на текстуалните карактеристики

Текстуалната репрезентација на сликите е формирана на основа на трудовите во кои се појавуваат. Во нашиот случај репрезентацијата е формирана со спојување на насловот на трудот и текстуалниот опис на сликата во трудот. Добиената репрезентација се процесира на начинот објаснет во поглавјето 4.4.1. За секој термин од секоја процесирана репрезентација пресметуваме тежини со стандардниот TF-IDF моделот.

Пресметаните тежини ги нормализираме со L2 нормализација и добиените резултати ги користиме како текстуални карактеристики за класификацијата.

Поставување на класификаторот

Како алгоритам за класификација се користат машините со носечки вектори од библиотеката LibSVM [114], [115]. За решавање на класификацискиот проблем со повеќе класи е искористен пристапот *еден-против-сите*. Имено, се обучува бинарен класификатор за секоја класа/модалитет т.е. сликите кои се асоцирани со тој модалитет/класа се означени како позитивни примероци, а сите останати слики се третираат како негативни примероци. Со овој начин на обучување има голема нерамнотежа меѓу бројот на позитивни и негативни примероци, што може да претставува проблем за класификаторот. Овој проблем е решен со додавање на тежини на позитивната и негативната класа [89]. Конкретно, тежината на позитивната класа се поставува како $((\#pos + \#neg)/\#pos)$, а тежината на негативната класа на следниот начин $((\#pos + \#neg)/\#neg)$, каде $\#pos$ е бројот на позитивни примероци, а $\#neg$ е бројот на негативни примероци во множеството за обучување.

Текстуалните и визуелните дескриптори се многу различни по природа. Текстуалните дескриптори имаат многу нулеви вредности во векторите, што значи дека класификаторот треба да го зема тоа предвид. За визуелните дескриптори се користи χ^2 кернел. За текстуалните дескриптори се користи однапред рачно пресметан кернел со помош на косинусно растојание и L2 нормализирани TF-IDF тежини. Параметарот C е оптимизиран на автоматски начин со пребарување во одреден опсег на вредности, при тоа, 20% од множеството за обучување се одвојува како множество за валидација. Откако ќе се пронајде оптималниот параметар, SVM класификаторот се обучува на целото множество од слики за обучување и се евалуира на сликите за тестирање.

Евалуацијата на перформансите се прави со пресметување прецизност на класификацијата, односно процентот на тест слики кои точно беа означени во фазата на класификација.

Оптимизирање на дескрипторите

Со цел да се постигнат најдобри резултати во крајните експерименти прво мора да се изврши оптимизација на дескрипторите преку прелиминарни експерименти за подесување на нивните параметри. Параметарот кој го оптимизираме е бројот на подслики/просторни пирамиди за дескрипторите. Експериментите за оптимизација се изведуваа на следниот начин.

Генерираме визуелни дескриптори за сликите за обучување и тестирање користејќи различни вредности за бројот на подслики (кој ја делат сликата на униформни делови). Исто како кај сите други експерименти, прво го оптимизираме C параметарот на SVM класификаторот со одделување на 20% од множеството за обучување како множество за валидација. Откако ја определиме оптималната вредност за C, SVM класификаторот го обучиме со целото множество за обучување и го евалуираме на множеството за

тестирање. Оптималните вредности за бројот на подслики за секој дескриптор го избравме по ревидирање на резултатите од евалуацијата. Во агрегатниот дескриптор, кој го користиме во останатите експерименти, ги спојуваме дескрипторите со оптималните параметри добиени од оваа фаза.

LBP. Со цел да се збогати просторната информација и да се подобрат перформансите на десктрипторот предложено е да се извлечат дескриптори од подслики/подрегиони на сликите. Според воспоставената практика, сликите ги поделивме со решетка (анг. *grid*) на униформни делови, односно, ги поделивме на 1x1, 2x2, 3x3, 4x4, 5x5, 6x6, 7x7 и 8x8 делови. Според нашите евалуации најдобри резултати се добиваат со 16 региони (4x4) за LBP дескрипторот, а со помал или поголем број на региони се намалуваат перформансите.

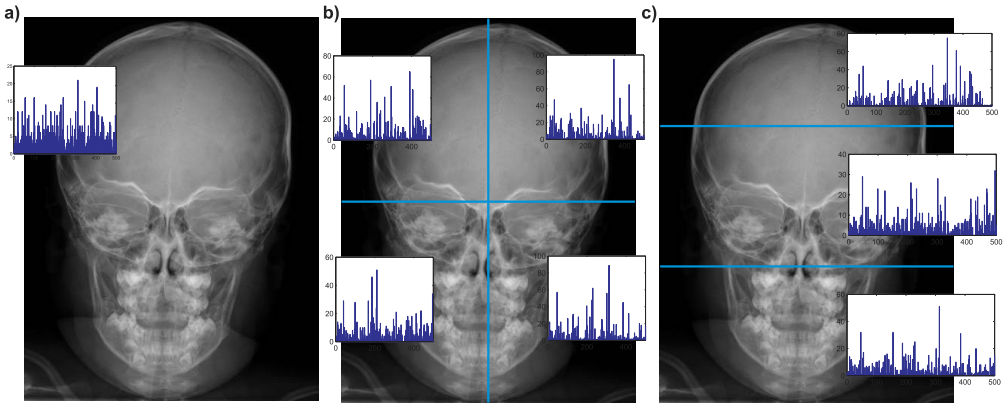
CEDD. Истата процедура ја повторивме за CEDD дескрипторот со истиот број на региони и според резулатите заклучивме дека оптимален број на региони е 6x6. Имајќи предвид дека се генерираат 192 бинови за секој региони, крајниот дескриптор ќе има 6 x 6 x 144 = 5184 бинови.

FCTH. Процедурата за FCTH е иста како за LBP и CEDD дескрипторот. Евалуацијата на овој дескриптор ни покажа дека 6x6 е оптималниот број на региони. Овој дескриптор генерира 192 бинови за секој региони, па според тоа конечниот дескриптор ќе има 6 x 6 x 192 = 6912 бинови.

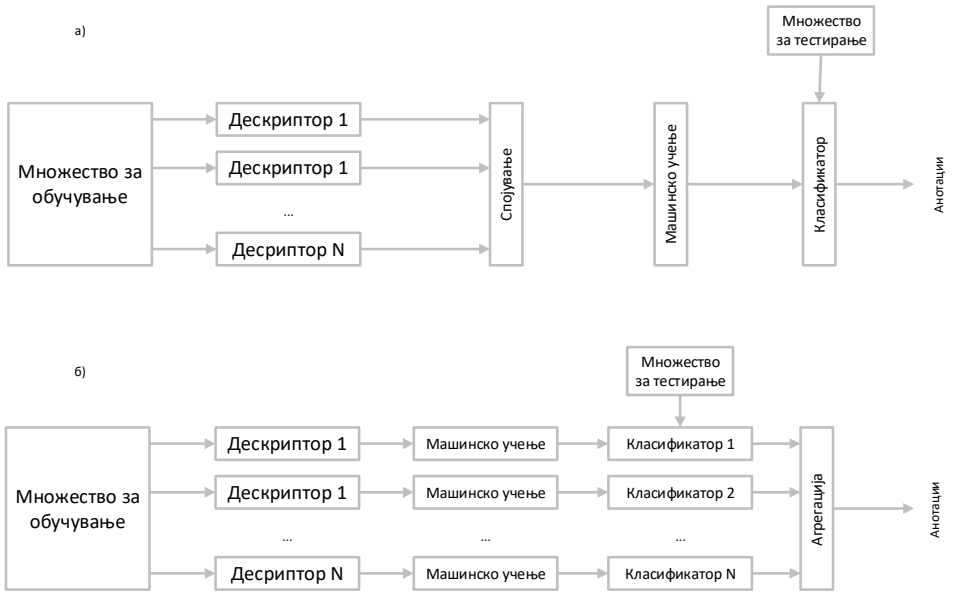
SIFT. Густото избирање на клучни точки им дава иста тежина на сите точки, независно од нивната просторна распределба во сликата. За овој проблем да се надмине, го применуваме методот на просторни пирамиди [116]. Со овој метод, сликата се дели на фиксни региони, како 1x1, 2x2, 4x4 итн. и различните резолуции се агрегираат во т.н. просторна пирамида. Просторната пирамида може да се користи во комбинација со густото бирање, бидејќи секој регион претставува посебна (помала) слика. Според [116] се тврди дека 2x2 е доволно грануларно за просторна пирамида, додека во [117] се тврди дека потребно е да се вклучи и 1x3 поделба. Евалуациите кои ги направивме покажуваат дека SIFT дескриптор во контекст на класификација на медицински слики според модалитет, најдобро се конфигурира со просторни пирамиди 1x1 + 2x2 + 1x3, што значи дека резултантиот вектор има 8000 бинови кој е добиен со спојување на пресметаните хистограми од секој регион (1000 бинови по регион). На Слика 6.1 е прикажано извлекување на карактеристики со просторни пирамиди за поделба со 1x1, 2x2 и 1x3 региони.

Методи за спојување на податоците

Во подсистемот за класификација на сликите според модалитет правиме два начини на спојување на карактеристиките прикажани на Слика 6.2: спојување на ниско и високо ниво. При спојувањето на ниско ниво, дескрипторите се спојуваат во еден вектор. Овој вектор се користи како влез во класификаторот. Кај спојувањето на високо ниво се обучува класификатор за секој дескриптор поединечно, а класификацијата се прави со усреднување на предикциите од сите класификатори.



Слика 6.1: Во рамките на нашите експерименти креиравме три различни просторни пирамиди: а) 1 x 1, б) 2 x 2 и в) 1 x 3. Пристапот ги извлекува карактеристиките во форма на хистограм од секој регион.



Слика 6.2: Спојување на дескрипторите на ниско (а) и високо (б) ниво.

Стратегијата за спојување на ниско ниво се користи за спојување на визуелните карактеристики, бидејќи дава подобри резултати од спојување на високо ниво во контекст на анотација на медицински слики [59]. Но, поради различната природа на текстуалните и визуелните дескриптори, обучуваме два посебни класификатори за текстуалните дескриптори и споените визуелни дескриптори, а конечната класификација се прави со усреднување на предикциите од двата класификатори. Тежината за предикциите од класификаторот на основа на текстуални карактеристики е 0.5, а тежината на класификаторот на основа на визуелните карактеристики е истотака 0.5. Тежините ги одредивме со помош на горенаведените техники за оптимизирање.

6.1.2 Експериментални прашања

Целта на експерименталната еваулација на подсистемот за класификација на медицински според модалитет е да се одговори на следните прашања:

1. Кој визуелен дескриптор е најдобар за класификација на медицински слики според модалитет?
2. Дали комбинирање на различни визуелни дескриптори придонесува за подобри перформанси на класификацијата?
3. Кои карактеристики се подобри за класификација на медицински слики според модалитет, текстуалните или визуелните?
4. Дали спојувањето на текстуалните и визуелните карактеристики придонесува за подобри перформанси на класификацијата?

Со цел да се определи кој дескриптор е најсоодветен за класификација на слики според модалитет, ќе ги споредиме резултатите од класификаторите изградени поединечно за секој дескриптор (прашање 1). Потоа, ќе ги споредиме резултатите од класификаторите изградени за поединечните дескриптори со класифиакаторот изграден на споениот дескриптор (прашање 2). Со споредување на перформансите од најдобриот визуелен дескриптор и класификаторот изграден над текстуалните карактеристики ќе дадеме одговор на третото прашање. На крај, со споредување на перформансите на класфикатори со и без текстуални карактеристики ќе дадеме одговор на четвртото прашање.

6.1.3 Резултати и дискусија

Во Табела 6.1 се презентирани резултатите добиени во однос на прецизност на класификација со помош на горенаведните експериментални поставувања. Дискусијата за резулатите ќе ја започнеме со перформансите на визуелните дескриптори, претставени во првите пет редови од Табела 6.1. Може да забележиме дека најдобри предиктивни перформанси има OSIFT дескрипторот над сите множества. Тоа значи дека додавање на информации за бојата при пресметувањето на SIFT дескрипторот (во opponent простор на бои) придонесува за подобри перформанси за 1%-2% во однос на перформансите при страндардното пресметување на SIFT дескрипторот над сликите трансформирани во нијанси на сиво. Подобрувањето изгледа логично, ако се земе предвид дека само бојата како информација помага во разликување меѓу одредени модалитети на слики.

Разликите во перформансите меѓу OSIFT и SIFT дескрипторите, од една страна, и CEED, FCTH и LBP дескрипторите, од друга страна е значителна. Експерименталните резултати покажуваат дека визуелните карактеристики кои ги опишуваат сликите на локално ниво (на пример, SIFT дескрипторите) даваат подобри резултати од оние кои даваат глобални описи на сликите. Локалните карактеристики ги опфаќаат деталите, а глобалните карактеристики креираат генерален опис на целата слика. Уште повеќе,

Табела 6.1: Предиктивни перформанси на класификаторот обучен од дескриптори пресметани со различните методи за извлекување на карактеристики и нивните комбинации. Резултатите се прикажани за множествата: ImageCLEF 2011, 2012 и 2013. Ознаката CONCAT се однесува на спојување на карактеристиките на ниско ниво, односно конкатанација на сите визуелни дескриптори. Ознаката CONCAT+TEXT се однесува на спојување на карактеристиките на високо ниво за конкатанираните визуелни карактеристики и текстуални карактеристики.

	2011	2012	2013
LBP	65.52	48.00	67.37
FCTH	72.16	47.70	61.81
CEDD	72.36	50.30	68.00
SIFT	80.95	66.50	76.36
OSIFT	82.03	67.30	78.10
TEXT	72.65	63.80	63.88
CONCAT	84.66	70.40	80.31
CONCAT+TEXT	87.10	77.10	82.25

SIFT дескрипторите се отпорни на шум, промени во осветлувањето, размерот, ротацијата и друг вид дисторзија. Според тоа, можеме да заклучиме дека локалните дескриптори се посоодветни отколку глобалните дескриптори во контекст на класификација на медицински слики според модалитет.

Вклучувањето на повеќе видови визуелни карактеристики во процесот на класификација придонесува за подобра репрезентација на визуелната содржина на сликите и помага за дополнително подобрување на предиктивните перформанси. Класификаторот обучен со споените визуелни дескриптори (CONCAT во Табела 6.1) користи информација за различните аспекти на сликите опфатени преку сите вклучени дескриптори. Оваа дополнителна информација е комплементарна и придонесува класификаторот да произведе подобри перформанси. Резултатите на класификаторот обучен со споените визуелни дескриптори прикажува подобри резултати од класификаторите обучени на поединечните дескриптори над сите три множества.

Според прикажаните резултати за текстуалните дескриптори (TEXT во Табела 6.1) можеме да забележиме дека единствено OSIFT и SIFT дескрипторите имаат подобри перформанси. Резултатите на класификаторот обучен на основа на текстуалните карактеристики има споредливи или (во некои случаи) подобри перформанси од класификаторите на основа на CEDD, FCTH и LBP дескрипторите. Тоа е добар сигнал дека текстуалните карактеристики имаат важни информации што можат дополнително да ги подобрат перформансите на класификаторите.

Текстуалните карактеристики во комбинација (на високо ниво) со споените визуелни карактеристики генерираат најдобри предиктивни перформанси за трите множества (последниот ред во Табела 6.1). Подобрување е особено видливо за множеството ImageCLEF 2012, каде комбинираниот пристап дава подобрување за 6.7% во однос на вториот најдобар пристап (CONCAT). Подобрувањето кај другите две множества е

Табела 6.2: Детален приказ на резултати за експериментите на 2011 множеството.

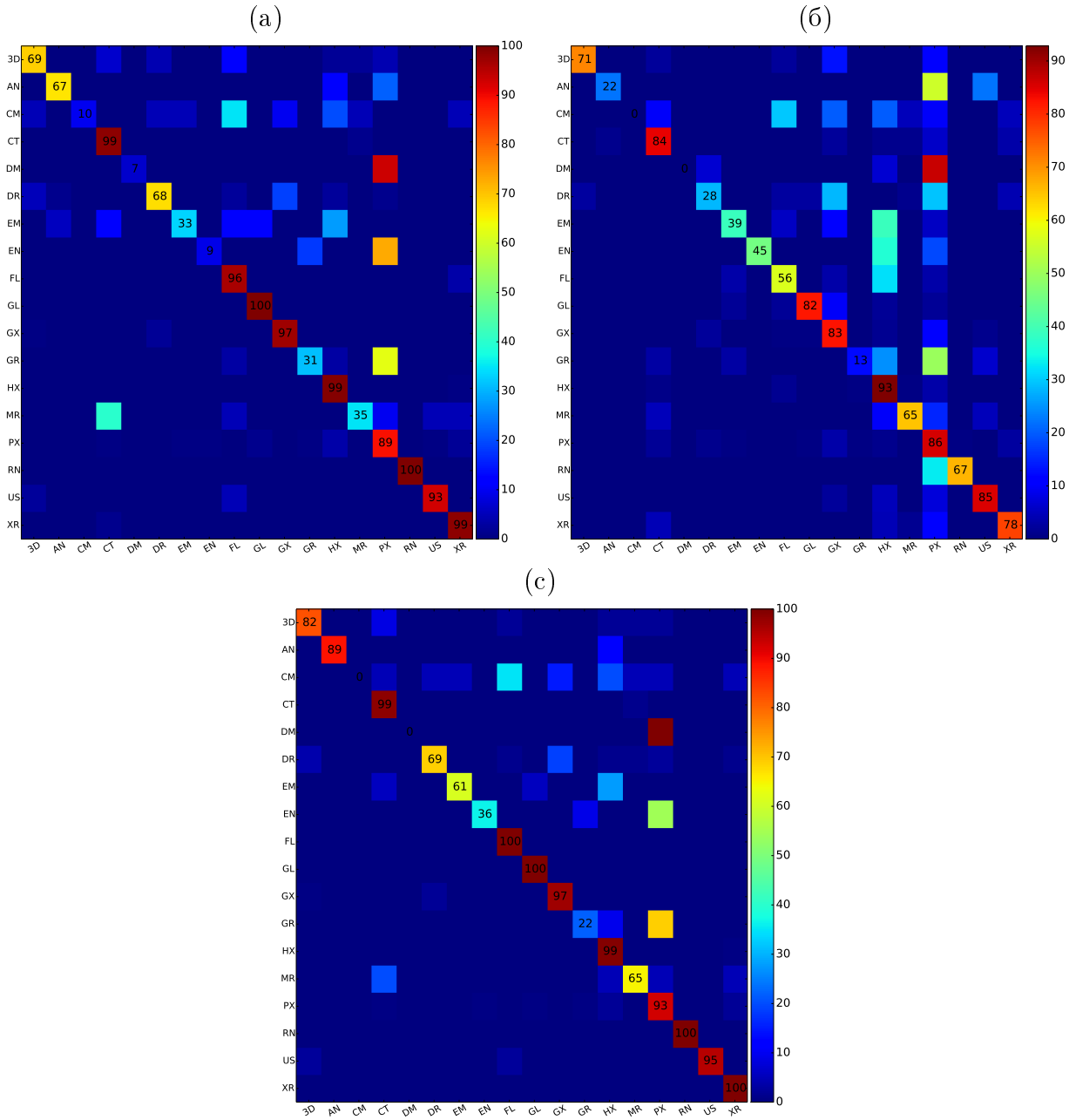
Опис	Код на класа	Слики		Прецизност		
		#обучување	#тестирање	Визуелно	Текстуано	Комбинирано
Electron microscopy	EM	16	18	33.33	38.89	61.11
Histopathology	HX	208	195	99.49	92.82	99.49
Dermatology	DM	7	15	6.67	0.00	0.00
Gross pathology	GR	43	32	31.25	12.50	21.88
Compound figure	CM	17	20	10.00	0.00	0.00
Fluoresence	FL	43	28	96.43	57.14	100.00
Graphs	GX	161	172	97.09	82.56	97.09
Ultrasound	US	30	41	92.68	85.37	95.12
Angiography	AN	11	9	66.67	22.22	88.89
Gel	GL	50	50	100.00	82.00	100.00
Endoscopic imaging	EN	10	11	9.09	45.45	36.36
Magnetic resonance imaging	MR	16	20	35.00	65.00	65.00
X-ray	XR	59	67	98.51	77.61	100.00
Retinography	RN	5	3	100.00	66.67	100.00
3D reconstruction	3D	32	45	68.89	71.11	82.22
Drawing	DR	43	74	67.57	28.38	68.92
General photo	PX	166	141	89.36	85.82	92.91
Computed tomography	CT	71	83	98.80	84.34	98.80

2.44% и 1.94% за множеството ImageCLEF 2011 и 2013, соодветно.

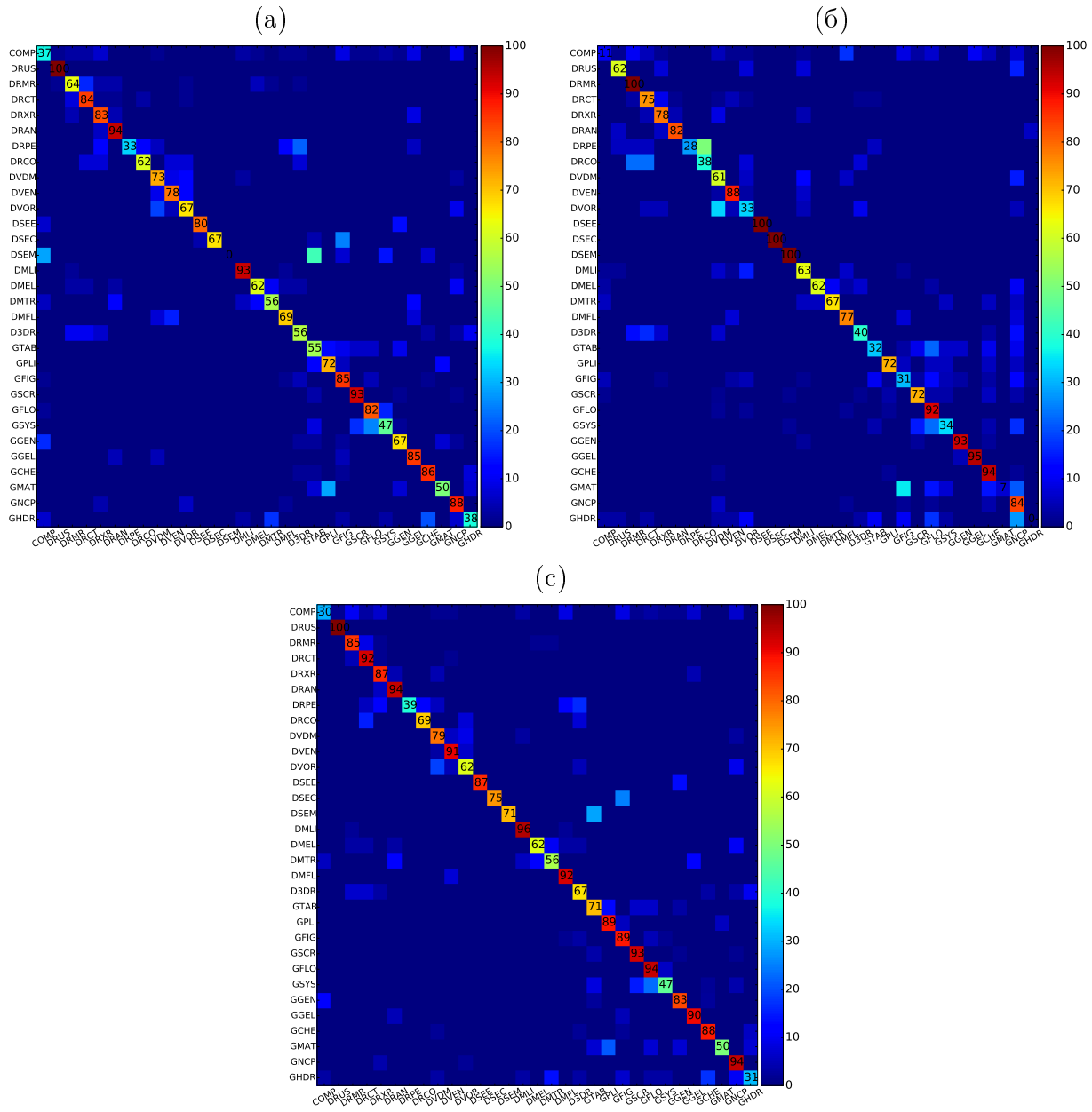
Подобрувањето на перформансите за 2012 множеството најдобро може да се објасни ако ги разгледаме матриците на грешки (анг. *confusion matrix*) за трите видови на карактеристики презентирани на Слика 6.3, 6.4 и 6.5 и деталните перформанси според модалитет презентирани на Табела 6.2, 6.3 и 6.4. Ако ги споредиме споените визуелни карактеристики (Слика 6.4 (а)) и текстуалните карактеристики (Слика 6.4 (б)) може да забележиме дека постојат комплементарни информации особено за класите DSEM, DSEC и DSEE (во средниот дел на дијагоналата). Според тоа логично е да се очекува подобрување на перформансите од комбинираните дескриптори (текстуални и визуелни) за истото множество.

Доколку ги разгледаме матриците на грешки за множествата ImageCLEF 2011 (Слика 6.3) и ImageCLEF 2013 (Слика 6.5), можеме да забележиме дека генерално текстуалните карактеристики произведуваат полоши перформанси од споените визуелни карактеристики за поголемиот број од класите. Уште повеќе нема комплементарност на информации меѓу текстуалните и споените визуелни карактеристики, што е особено видливо во множетсвото ImageCLEF 2013. Затоа комбинирањето на текстуалните и споените визуелни карактеристики резултира со помал раст на перформансите споредено со множеството ImageCLEF 2012.

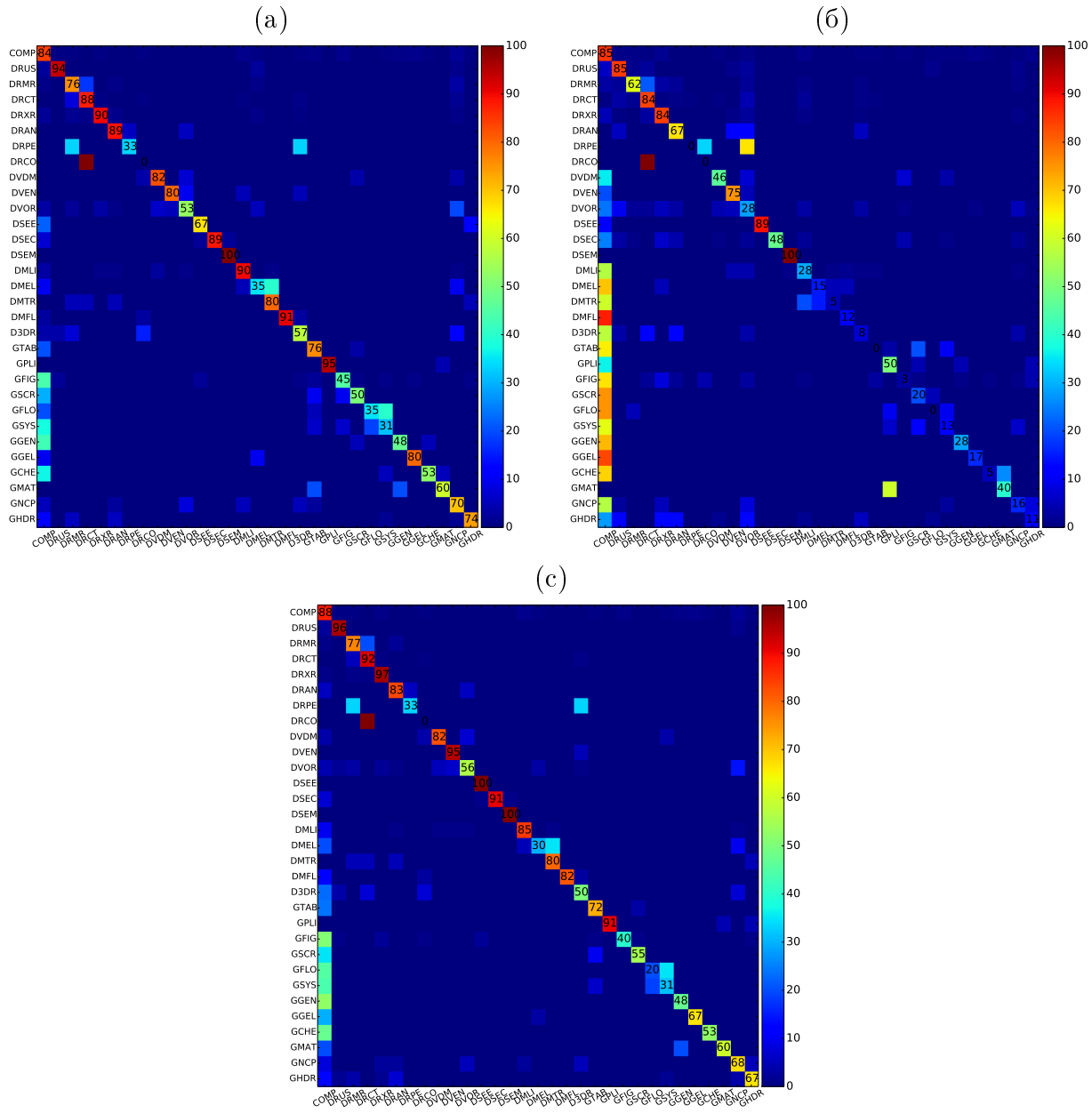
Во Табела 6.5 се прикажани официјалните резултати од најдобрите пристапи над истите множества. Резултатите се групирани според видот на податоците на основа на кои се прави класификацијата, што може да биде на основа на текстуални, визуелни или комбинирани податоци.



Слика 6.3: Матрица на грешка за множеството 2011: (а) споени визуелни карактеристики (б) текстуални карактеристки (в) доцна фузија на споените визуелни карактеристики и текстуалните карактеристики.



Слика 6.4: Матрица на грешка за множеството 2012: (а) споени визуелни карактеристики (б) текстуални карактеристки (в) доцна фузија на споените визуелни карактеристики и текстуалните карактеристики.



Слика 6.5: Матрица на грешка за множеството 2013: (а) споени визуелни карактеристики (б) текстуални карактеристки (в) доцна фузија на споените визуелни карактеристики и текстуалните карактеристики.

Табела 6.3: Детален приказ на резултати за експериментите на 2012 множеството.

Опис	Код на класа	Слики		Прецизност		
		#обучување	#тестирање	Визуелно	Текстуано	Комбинирано
Tables and forms	GTAB	38	31	54.84	32.26	70.97
Fluorescence microscopy	DMFL	21	13	69.23	76.92	92.31
Chromatography, Gel	GGEL	49	20	85.00	95.00	90.00
Statistical figures, graphs, charts	GFIG	48	61	85.25	31.15	88.52
Other organs	DVOR	48	21	66.67	33.33	61.90
Chemical structure	GCHE	21	50	86.00	94.00	88.00
Light microscopy	DMLI	46	46	93.48	63.04	95.65
Angiography	DRAN	38	17	94.12	82.35	94.12
Screenshots	GSCR	40	54	92.59	72.22	92.59
Endoscopy	DVEN	32	32	78.13	87.50	90.63
Hand-drawn sketches	GHDR	17	29	37.93	0.00	31.03
Gene sequence	GGEN	47	42	66.67	92.86	83.33
System overviews	GSYS	48	47	46.81	34.04	46.81
Compound or multipane images	COMP	49	57	36.84	10.53	29.82
Ultrasound	DRUS	48	13	100.00	61.54	100.00
Combined modalities in one image	DRCO	12	13	61.54	38.46	69.23
Electromyography	DSEM	5	14	0.00	100.00	71.43
Program listing	GPLI	10	18	72.22	72.22	88.89
Electroencephalography	DSEE	6	15	80.00	100.00	86.67
Mathematics, formulae	GMAT	6	14	50.00	7.14	50.00
Electrocardiography	DSEC	5	24	66.67	100.00	75.00
X-Ray, 2D Radiography	DRXR	48	23	82.61	78.26	86.96
Transmission microscopy	DMTR	29	18	55.56	66.67	55.56
Flowcharts	GFLO	48	50	82.00	92.00	94.00
Dermatology, skin	DVDM	47	33	72.73	60.61	78.79
Electron microscopy	DMEL	22	29	62.07	62.07	62.07
Computerized Tomography	DRCT	49	64	84.38	75.00	92.19
PET	DRPE	9	18	33.33	27.78	38.89
Non-clinical photos	GNCP	47	49	87.76	83.67	93.88
Magnetic Resonance	DRMR	43	55	63.64	100.00	85.45
3D reconstructions	D3DR	25	30	56.67	40.00	66.67

Табела 6.4: Детален приказ на резултати за експериментите на 2013 множеството.

Опис	Код на класа	Слики		Прецизност		
		#обучување	#тестирање	Визуелно	Текстуано	Комбинирано
Tables and forms	GTAB	65	29	75.86	0.00	72.41
Fluorescence microscopy	DMFL	33	33	90.91	12.12	81.82
Chromatography, Gel	GGEL	55	30	80.00	16.67	66.67
Statistical figures, graphs, charts	GFIG	102	102	45.10	2.94	40.20
Other organs	DVOR	70	92	53.26	28.26	56.52
Chemical structure	GCHE	62	19	52.63	5.26	52.63
Light microscopy	DMLI	91	121	90.08	28.93	85.12
Angiography	DRAN	54	18	88.89	66.67	83.33
Screenshots	GSCR	91	20	50.00	20.00	55.00
Endoscopy	DVEN	64	20	80.00	75.00	95.00
Hand-drawn sketches	GHDR	46	54	74.07	11.11	66.67
Gene sequence	GGEN	68	21	47.62	28.57	47.62
System overviews	GSYS	89	16	31.25	12.50	31.25
Compound or multipane images	COMP	1105	1014	84.22	85.21	87.67
Ultrasound	DRUS	60	85	94.12	84.71	96.47
Combined modalities in one image	DRCO	22	1	0.00	0.00	0.00
Electromyography	DSEM	18	1	100.00	100.00	100.00
Program listing	GPLI	28	22	95.45	50.00	90.91
Electroencephalography	DSEE	21	9	66.67	88.89	100.00
Mathematics, formulae	GMAT	20	5	60.00	40.00	60.00
Electrocardiography	DSEC	29	96	88.54	47.92	90.63
X-Ray, 2D Radiography	DRXR	70	344	90.12	84.01	96.80
Transmission microscopy	DMTR	46	20	80.00	5.00	80.00
Flowcharts	GFLO	94	20	35.00	0.00	20.00
Dermatology, skin	DVDM	79	28	82.14	46.43	82.14
Electron microscopy	DMEL	51	20	35.00	15.00	30.00
Computerized Tomography	DRCT	113	186	87.63	83.87	92.47
PET	DRPE	16	3	33.33	0.00	33.33
Non-clinical photos	GNCP	96	37	70.27	16.22	67.57
Magnetic Resonance	DRMR	97	90	75.56	62.22	76.67
3D reconstructions	D3DR	46	26	57.69	7.69	50.00

Табела 6.5: Официјални резултати за експерименти на множеството за класификација на медицински слики според модалитет за ImageCLEF 2011, 2012 и 2013. Експериментите се поделени според видот на податоците на кои се прави класификација (визуелни, текстуални и комбинирани).

Група	Вид на под.	Прецизност
2011		
XRCE [22]	Комбинирани	86.91
XRCE [22]	Визуелни	83.59
IPL [22]	Текстуални	70.41
2012		
medGIFT [118]	Комбинирани	66.20
IBM Multimedia Analytics [118]	Визуелни	69.60
ITI [118]	Текстуални	41.30
2013		
IBM Multimedia Analytics [61]	Комбинирани	81.68
IBM Multimedia Analytics [61]	Визуелни	80.79
IBM Multimedia Analytics [61]	Текстуални	64.17

6.2 Подсистем за пребарување на медицински слики

6.2.1 Експериментални поставувања

Карактеристиките кои ги имплементиравме во подсистемот за опишување на сликите може да се поделат во две групи: текстуални и визуелни карактеристики. Поставувањата за секоја од групите се дадени во продолжение.

Поставување на текстуални карактеристики

Текстуалните карактеристики се извлечени од текстуалната репрезентација на сликите. Текстуалната репрезентација на дадена слика од базата се формира со спојување на следните делови од трудот во кој се појавува: насловот, апстрактот, MeSH термините, текстуалниот опис на сликата (анг. *caption*) и деловите (реченициte) во кои сликата е референцирана. Овие делови се избрани на основа на претходни емпириски проверки [61].

На основа на текстуалните карактеристики се креира инвертиран индекс за ефективно и ефикасно пребарување. За овој дел е интегриран системот Terrier IR како платформа за пребарување, заради неговата флексибилност и можноста за работа со големи бази на податоци.

Поставување на визуелни карактеристики

Во однос на визуелните карактеристики се користи комплексна репрезентација да се опише визуелната содржина на сликите. Конкретно, во различни фази од подсистемот се користат различни дескриптори т.е. за содржински базираното пребарување се користи еден вид на карактеристики, а за класификација на модалитети се користи друго множество на карактерстики.

Во делот од подсистемот кој се занимава со класификација на модалитет е имплементирана техника на основа на карактеристики за текстурата. Класифицирање на медицинските слики на основа на облици или нијанси на сиво е тешко и токму затоа текстурите се клучни во овој процес [52]. Текстурите мора ефикасно да се репрезентираат на тој начин што слики од истиот модалитет и визуелен распоред може меѓусебно да се разликуваат. Според тоа, во овој дел од подсистемот се употребува SIFT дескрипторот или попрецизно opponentSIFT (OSIFT) дескрипторот [46]. Карактеристиките на основа на SIFT се локални карактеристики. Нашата претходна работа во склоп на ова поле покажала дека локалните карактеристики се подобри за ваков вид на проблем [3]. Подсистемот имплементира локални карактеристики, бидејќи тие овозможуваат детален опис на сликите, но, истовремено се толерантни за разлики меѓу сликите во рамките на истата класа, а строги во однос на разлики со слики од други класи. Локалните карактеристики ги опфаќаат деталите во сликите, за разлика од глобалните карактеристики, кои даваат опис на целата слика. Клучната идеја на овој пристап е да се најдат делови од сликата кои може да се употребуваат како примероци (со детекција

на клучни точки, по случаен избор или на точно определен опсег - densely) врз кои може да се пресмета визуелен дескриптор. На крај се генерира речник (анг. *codebook*) и хистограми на основа на тој речник.

Во однос на делот за содржински базирано пребарување на слики имплементиран е RGB хисторам [89]. Зошто се одлучивме да ги опишуваме сликите со RGB во фазата на пребарување? Одговорот е: бидејќи нудат рамнотежа меѓу перформанси во поглед на времето на пребарување и квалитетот во опишување на сликите. Во рамките на истражувањата беа направени пребарувања и со opponentSIFT, но, времето на пребарување беше премногу бавно. RGB хистограмите ни овозможија слични перформанси во однос на прецизноста на пребарувањето, но, со драматично намалување на времето на пребарување. Подсистемот извршува класификација на модалитет во офлајн фазата. Во таа фаза, може да се жртвува времето на процесирање со цел да се добијат поквалитетни карактеристики. Во фазата на пребарување потребни се компактни каратеристики со високи перформанси со цел пребарување да може да се изврши во разумно време. Според тоа тука е потребно да се добие подобро време на процесирање, а не само високо квалитетни карактеристики.

RGB хистограмот е комбинација од три 1-D хистограми на основа на R, G и B каналите во RGB просторот на бои. Овој хистограм нема карактеристики кои се однесуваа на инваријантност. Откако ќе се изгенерираат RGB дескрипторите се применува енкодирање со Фишер вектори [119] со помош на дистрибуции на карактеристики (анг. *feature distributions*). Во контекст на пребарување на слики Фишер кернелите го прошируваат познатиот концепт на визуелни зборови (анг. *bag of visual words - BOV*) и се покажало дека произведуваат добри резултати. Конечната репрезентација со Фишер кернел е соодветна за пребарување на големо множество на податоци. Току затоа го избравме овој метод на репрезентација на сликите во делот за содржински базирано пребарување. Друга причина за употребување на Фишер кернели е тоа што тие природно овозможуваат начин на одредување на сличноста меѓу сликите преку веројатносниот распоред. Овој начин на репрезентација во споредба со BOV репрезентацијата има потреба од помалку визуелни зборови.

Поставување на содржински базираното пребарување

Основната идеја на содржински базираното пребарување е генерираните карактеристики за сите слики во множеството и нивно енкодирање во компактен вектор, кој ќе овозможи побрзо пребарување. Целиот процес се состои од два чекори. Прво, се генерира проекција, која ја намалува димензионалноста на векторот, а потоа, резултантните вектори се индексираат преку метод на квантизација. Наједноставен начин за решавање на овој проблем е преку методи за апроксимативно (анг. *approximate*) пребарување на најблиски соседи. Повеќето од овие техники имаат потреба постојано да чуваат хеш табели во меморија, што ги прави неприменливи на големи множества на податоци (како што е случајот со нашите множества). Затоа се осврнуваме кон методот на Spectral Hashing [120], бидејќи го трансформира векторот во бинарен формат и има помали ме-

мориски пребарувања. Со овој метод, секоја слика од множеството е репрезентирана преку компактен бинарен код. Слични слики во множеството имаат слични бинарни репрезентации и таа сличност може да се измери со Хамингово растојание.

За определување на растојание, во подсистемот е имплементирано пресметување на асиметрично растојание (анг. *Asymmetric Distance Computation - ADC*). Во [121] е претставена компактна шема за кодирање со систем за инвертиран индекс, а во овој контекст ADC само ги подобрува резултатите во смисла на рамнотежа меѓу квалитетот на пребарување и мемориските барања. Специфичноста на овој метод е во тоа што ги кодира сите вектори во множеството, но, не и векторот на прашањето. На пример, нека x е векторот на прашањето за кој треба да ги најдеме најблиските соседи $NN(x)$ во множеството $Y = y_1, y_2, ..., y_n$. Пристапот со ADC се состои од кодирање на секој вектор y_i со метод на квантизација $c_i = q(y_i)$. За квантизација употребуваме квантизатор $q(.)$ од k центроиди. Тоа значи дека вектрот ќе се кодира со $log_2(k)$ битови, каде k е 2 на некој степен. Откако ќе заврши квантизацијата, за наоѓање на најблиските a соседи на x, потребно е само да се реши $a - argmin||x - q(y_i)||$. Можеме да заклучиме дека нема грешка во апроксимацијата, бидејќи влезниот вектор не се квантифицира.

6.2.2 Експериментални прашања

Целта на експерименталната еваулација на подсистемот за пребарување на медицински слики е да се одговори на следните прашања:

1. Кој модел за утежнување е најдобар за пребарување на генерички медицински слики?
2. Како нашиот подсистем за пребарување се споредува во однос на метриките со останатите пристапи за пребарување над истите множества?

Одговорот на првото прашање го истражувавме со пребарување на сликите од множеството ImageCLEF 2013 со различните моделите за утежнување кои ги разгледавме во Глава 4.6 и со споредување на добиените резултати. Да одговориме на второто прашање направивме три видови експерименти со цел да се опфатат сите комбинации на пребарување кои ги нуди подсистемот (текстуално базирано, содржински базирано и комбинирано пребарување) и добиените резултати ги споредивме со пријавените резултати од други истражувачи над истите множества.

6.2.3 Резултати и дискусија

Во Табела 6.6 се прикажани резултатите од пребарувањето над множеството ImageCLEF 2013 со различни модели на утежнување. Во овие експерименти го земаме текстуалниот дел од прашањата и пребаруваме низ текстуалните репрезентации на индексираните слики. Резултатот е подредена листа од слики, подредени според релевантноста, односно

слики со текстуална репрезентација кои имаат повеќе заеднички термини со текстуалното прашање треба да се наоѓаат во погорните резултати. Подредеността на сликите во голема мера зависи од моделот на утежнување.

Во табелата се опфатени мерките за утежнета средна прецизност (MAP), прецизноста при првите 10 вратени слики (P10) и вкупниот број на вратени релевантни слики. Можеме да забележиме дека најдобра утежнета средна прецизност се појавува кај моделот за утежнување BM25 и дава ∼2% подобри резултати од останатите модели. Исто така, бројот на вратени релевантни слики е најголем кај овој модел. Интересно е што TF-IDF моделот има најдобри перформанси за P10 со 2,2% подобра прецизност од BM25. Но, имајќи предвид дека MAP е многу поопсежна мерка за прецизност и ако се земе предвид бројот на вратени релевантни слики, понатамошните експерименти во контекст на текстуално базирано пребарување на слики ги правевме со BM25 моделот.

Табела 6.6: Резултати од пребарување со различни модели за утежнување над множеството ImageCLEF 2013.

Модел	***MAP***	***P10***	***# број на рел. док. вратени***
BB2	0.2056	0.3429	473
BM25	0.2266	0.3381	494
DFR-BM25	0.2091	0.3476	474
PL2	0.2055	0.3429	472
TF-IDF	0.2085	0.3524	471
DirichletLM	0.1601	0.2619	434

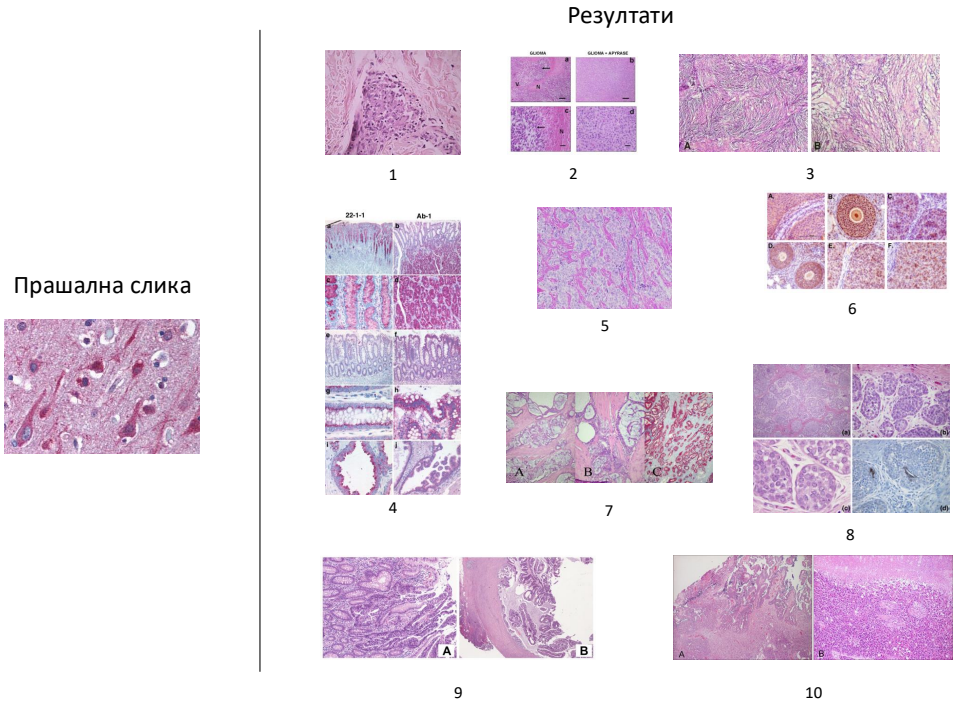
Преостанатите експерименти се прикажани во Табела 6.7. Прво, направивме експерименти со текстуално базирано пребарување (*текстуално*) со помош на BM25 моделот и метод за проширување на прашањата. Редовите означени со *содржински* се однесуваат на експериментите со содржински базирано пребарување. Пријавените резултати за содржински базираното пребарување за сите множества покажуваат послаби перформанси од текстуално базираното пребарување. Заради тоа, содржински базираното пребарување не го вклучивме во дополнителните експерименти. Слабите резултати на содржински базираното пребарување се должат на начинот на кој се формирани одговорите во множествата. Ова преставува проблем со сите пристапи со содржински базирано пребарување [122]. Според дизајнот на множествата, целта на пребарувањето било да се пронајдат семантични слични слики, а содржински базираното пребарување враќа визуелно слични слики. Пример на прашална слика и вратените резултати за таа слика се прикажани на Слика 6.6.

Редовите во табелата означени со *текстуално + модалитет* се однесуваат на експериментите кои го земаат предвид модалитетот на сликите. Кај овие експерименти се зема предвид само текстуалниот дел од секое прашање (заради лошите перформанси на содржински базираното пребарување). По извршување на текстуалното прашање, првичните резултати се анализираат. Текстуално базираното пребарување враќа листа од подредени слики. Сликите се подредени според пресметаната вредност за нивните текстуални карактеристики наспроти прашањето. За секоја слика во првично вратените

Табела 6.7: Резултати од пребарувањето на медицински слики над множествата од ImageCLEF на основа на текстуални и мултимодални податоци.

	Текстуално			Текстуално + модалитет			Содржински		
Година	MAP	P10	P30	MAP	P10	P30	MAP	P10	P30
2011	0.20	0.35	0.28	0.23	0.37	0.30	0.0014	0.0033	0.0022
2012	0.22	0.25	0.19	0.23	0.24	0.20	0.0004	0.0000	0.0015
2013	0.27	0.36	0.25	0.32	0.39	0.25	0.0004	0.0000	0.0000

резултати подсистемот го проверува индексот на модалитети да се одреди дали сликата го има истиот модалитет како некој од модалитетите извлечени од прашањето. Доколку модалитетот на сликата и прашањето се исти, тогаш пресметаната вредност на сликата за зголемува за одреден фактор [2]. Во експериментите факторот се определува за секое множество посебно. Оптимизацијата на факторот за секое множество се прави на основа на претходната верзија на множеството (со исклучок на верзијата 2011, каде за оптимизација се користеше верзијата 2013). Подсистемот го проверува модалитетот на секоја слика во иницијалните резулати и ја модифицира пресметаната вредност по потреба. Потоа, сите слики се рангираат повторно според ново-пресместаните вредности. Дополнително, овој процес ни покажува дали модалитетот го подобрува пребарувањето. Можеме да забележиме дека овој пристап дава подобрување на усреднетата средна



Слика 6.6: Пример на прашална слика и вратени резултати при содржински базирано пребарување.

прецизност за приближно 2% над стандардното текстуално базирано пребарување. Интегрирање на информацијата за модалитетот во процесот на пребарување ги подобрува перформансите кај сите три множества.

Подетални резултати за секое множество се презентирани во Табела 6.8, Табела 6.9, Табела 6.10 за множествата ImageCLEF 2011, 2012 и 2013, соодветно.

За множеството од 2011, можеме да забележиме дека резултатите за повеќето од прашањата се исти, но, за некои има драстични подобрувања во перформансите. Пример, прашањата 4, 10, 13, 19 имаат подобрување од околу 20% на MAP. Во продолжение се прикажани споменатите прашања:

- *4 - chest CT images with emphysema*
- *10 - medial meniscus MRI*
- *13 - all x-ray images containing one or more fractures*
- *19 - mediastinum PET*

Доколку ги провериме одговорите за дадените прашања, може да видиме дека тие се наменети да најдат слики со модалитетот кој е споменат во истите (пр. CT, MRI и слично). Од друга страна, има прашања за кои резултатите требаше да бидат најдени според некоја семантика. Тоа значи дека резултатите за тие прашања не се слики кои имаат слична текстуална репрезентација како прашањата ниту се визуелни слични со прашалните слики.

Деталните резултати за множеството 2012 прикажуваат дека има мали подобрувања на перформансите во некои случаи и никаков ефект или намалување во други случаи. Сепак, ова најмногу зависи од начинот на кој се дадени одговорите на прашањата. Пример, прашањето 11, каде имаме подобрување на MAP и P10 е поставено на следниот начин: *CT images of small bowel obstruction*. Првите слики во дадениот одговор за ова прашање припаѓаат на модалитетот на компјутерска томографија. Според тоа очигледно е тоа што модулот за класификација на слики според модалитет, овозможува подобрување на резулатите кај тоа прашање. Од друга страна, имаме деградација на перформансите за прашањето *4 - pneumothorax CT images*. Првите слики во одговорот на ова прашање сугерираат дека имало потреба од семантичко пребарување, бидејќи во сликите кои се очекуваа како одговор нема многу или воопшто слики од дадената болест или модалитет. Нашиот подсистем пребарува на основа на дадениот текст и визуелните податоци, па според тоа не може да направи ваков вид на корелации. Интересни случаи се појавуваат во прашањата 14 и 15. Прашањето *14 - handdrawn figure* е премногу општо и не може да се интерпертира на некој конкретен начин и затоа во тој случај се прикажуваат лоши резултати. Прашањето 15 е дефинирано попрецизно (*ovarian torsion MRI image*), но дадените одговори (горните 10 слики) за прашањето се состојат од графици, што го прави ова прашање многу тешко за анализа со помош на нашите методи.

Анализата на множеството 2013 е слична како кај претходното множество. Постојат прашања кај кои нашиот пристап овозможува подобрување на резултатите и прашања

Табела 6.8: Резултати за множеството ImageCLEF 2011 на ниво на прашање.

Прашање	Текстуално			Текстуално + модалитет		
	MAP	P10	P30	MAP	P10	P30
1	0.0449	0.0000	0.0000	0.0525	0.2000	0.1000
2	0.0437	0.0000	0.1333	0.0313	0.0000	0.1000
3	0.5411	1.0000	0.8333	0.5800	0.8000	0.8333
4	0.2989	0.2000	0.3333	0.4913	0.7000	0.3000
5	0.1001	0.2000	0.1333	0.0821	0.1000	0.1000
6	0.1125	0.5000	0.3333	0.1125	0.5000	0.3333
7	0.3076	0.7000	0.3667	0.2908	0.6000	0.2333
8	0.0380	0.2000	0.1000	0.1063	0.3000	0.3000
9	0.4793	1.0000	0.9333	0.4811	0.8000	0.8667
10	0.2242	0.2000	0.1333	0.5167	0.4000	0.1333
11	0.2684	0.8000	0.4667	0.2716	0.8000	0.4667
12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
13	0.1731	0.0000	0.1333	0.4309	0.7000	0.7667
14	0.0870	0.2000	0.2333	0.0671	0.0000	0.1333
15	0.2497	0.6000	0.4333	0.1180	0.4000	0.2333
16	0.8257	1.0000	1.0000	0.8257	1.0000	1.0000
17	0.0196	0.1000	0.0667	0.0248	0.1000	0.1000
18	0.0951	0.1000	0.0333	0.0132	0.0000	0.0333
19	0.0696	0.1000	0.0333	0.2823	0.2000	0.1000
20	0.6261	0.4000	0.2000	0.6977	0.4000	0.2000
21	0.0519	0.0000	0.1000	0.0519	0.0000	0.1000
22	0.1701	0.1000	0.2333	0.1701	0.1000	0.2333
23	0.1680	0.7000	0.5333	0.1680	0.7000	0.5333
24	0.0259	0.3000	0.1667	0.0235	0.2000	0.0667
25	0.1734	0.5000	0.2667	0.1869	0.7000	0.4667
26	0.0074	0.0000	0.0000	0.0082	0.0000	0.0333
27	0.0274	0.1000	0.1333	0.0274	0.1000	0.1333
28	0.2489	0.3000	0.1000	0.2489	0.3000	0.1000
29	0.4127	0.2000	0.0667	0.4127	0.2000	0.0667
30	0.3504	1.0000	0.9333	0.3504	1.0000	0.9333

кај кои не придонесува за подобрување на резултатите или пак предизвикува спротивен ефект. Но, повторно, тоа е најмногу заради природата на прашањата т.е. некои прашања се наменети да бидат најдени со помош на семантички техники и тоа е најголемиот недостаток на нашиот пристап. На пример, резултатите за прашањето *17 - avascular necrosis MRI* се особено интересни. Со помош на текстуално базираното пребарување речиси и да нема релевантни слики во горните резултати. Но, по повторното рангирање на сликите со помош на информацијата за модалитетот на сликите, порелевантните слики се поместуваат кон погорните ранкови. Во одговорот на ова прашање има слики од магнетна резонанција во горните резулати и затоа се добива подобрување со нашиот пристап. Од друга страна, постојат прашања кои се премногу општи или небулозни (или не предвидуваат релевантни одговори), како 8, 18, 19 и нашиот пристап

Табела 6.9: Резултати за множеството ImageCLEF 2012 на ниво на прашање.

Прашање	Текстуално			Текстуално + модалитет		
	MAP	P10	P30	MAP	P10	P30
1	0.0000	0.0000	0.0000	0.0204	0.0000	0.0000
2	0.5150	0.8000	0.6333	0.5080	0.7000	0.6333
3	0.0475	0.2000	0.1000	0.0634	0.1000	0.2000
4	0.5040	0.7000	0.4000	0.3457	0.5000	0.4000
5	0.0387	0.1000	0.1667	0.0196	0.1000	0.1000
6	0.0769	0.1000	0.0333	0.0769	0.1000	0.0333
7	0.3194	0.2000	0.2333	0.2517	0.2000	0.2000
8	0.1744	0.1000	0.0333	0.1748	0.1000	0.0333
9	0.4167	0.2000	0.0667	0.7000	0.2000	0.0667
10	0.1382	0.1000	0.1667	0.1651	0.1000	0.1667
11	0.5915	0.8000	0.8000	0.6893	0.9000	0.8667
12	0.1870	0.3000	0.2000	0.1631	0.1000	0.3000
13	0.4738	0.5000	0.6333	0.5000	0.8000	0.6333
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
16	0.0703	0.1000	0.0333	0.1350	0.2000	0.0333
17	0.0290	0.0000	0.0667	0.0252	0.0000	0.0333
18	0.8167	0.4000	0.1333	0.8167	0.4000	0.1333
19	0.1134	0.2000	0.0667	0.1001	0.2000	0.0667
20	0.0551	0.0000	0.0333	0.0550	0.0000	0.0333
21	0.2413	0.6000	0.4000	0.2437	0.6000	0.4000
22	0.1617	0.2000	0.1333	0.1682	0.2000	0.1667

не даде релевантни резултати во ниту една од тие ситуации. Но, генерално подсистемот со помош на информацијата за модалитетот на сликите овозможува подобрување на пребарувањето.

Нашите анализи прикажуваат дека добиените резултати од подсистемот за пребарување се најдобрите резултати воопшто добиени за сите три множества според досега објавените трудови [61], [118], [22]. Уште повеќе, нашиот пристап покажува дека вметнување на информацијата за модалитетот на сликите во ваков вид на пребарување може да ги подобри целокупните перформанси. Секако, мора да споменеме дека големо влијание на ова има начинот на кој е поставено прашањето и тоа што се очекува од тоа како резултат. Во рамките на истражувањето на докторската дисертација, ние развивме онлајн систем, каде е имплементиран подсистемот за пребарување, кој е достапен за сите кои сакаат да видат како функционира [1].

[1] http://194.149.136.27/images/home

Табела 6.10: Резултати за множеството ImageCLEF 2013 на ниво на прашање.

Прашање	Текстуално			Текстуално + модалитет		
	MAP	P10	P30	MAP	P10	P30
1	0.3719	0.3000	0.1000	0.0373	0.0000	0.0667
2	0.4381	0.7000	0.6000	0.4501	0.7000	0.4667
3	0.0557	0.3000	0.2667	0.1286	0.3000	0.3000
4	0.4178	0.5000	0.6000	0.4804	0.6000	0.5333
5	0.3425	0.2000	0.1000	0.2173	0.1000	0.1000
6	0.1494	0.1000	0.1000	0.3333	0.1000	0.0333
7	0.6186	0.3000	0.1333	0.5714	0.3000	0.1333
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.5574	0.5000	0.2000	0.5097	0.5000	0.1667
10	0.3403	0.6000	0.3333	0.2673	0.1000	0.3000
11	0.3608	0.6000	0.3000	0.4530	0.8000	0.3333
12	0.3025	0.6000	0.3667	0.2465	0.2000	0.3000
13	0.4859	0.6000	0.6333	0.3792	0.6000	0.4667
14	0.0014	0.1000	0.0333	0.0000	0.0000	0.0000
15	0.3617	0.3000	0.1000	0.7725	0.7000	0.2333
16	0.6336	0.7000	0.3667	0.5591	0.7000	0.3333
17	0.0014	0.0000	0.0000	0.2091	0.5000	0.2000
18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0315	0.0000	0.0000
21	0.2066	0.4000	0.1667	0.3243	0.6000	0.3000
22	0.2500	0.1000	0.0333	0.5091	0.1000	0.0333
23	0.3340	0.1000	0.0333	0.3342	0.1000	0.0333
24	0.3707	0.4000	0.1667	0.4600	0.4000	0.2333
25	0.6004	0.9000	0.8667	0.6638	1.0000	0.9333
26	0.2506	0.4000	0.1667	0.2991	0.4000	0.2000
27	0.4167	0.8000	0.6000	0.3919	0.6000	0.7000
28	0.4411	0.8000	0.3333	0.4514	0.8000	0.4000
29	0.2350	0.5000	0.6667	0.2023	0.5000	0.5333
30	0.0700	0.1000	0.0667	0.1931	0.3000	0.2000
31	0.0000	0.0000	0.0000	0.0556	0.1000	0.0333
32	0.2650	0.3000	0.1667	0.4959	0.5000	0.2000
33	0.2869	0.7000	0.7000	0.2777	0.7000	0.5333
34	0.2425	0.5000	0.2333	0.2762	0.5000	0.2000
35	0.2827	0.4000	0.3000	0.2082	0.4000	0.1333

6.3 Подсистем за пребарување на медицински трудови со спојување на зборови и медицински концепти

6.3.1 Експериментални поставувања

Во овој дел од докторската дисертација се презентирани подесувањата на двата делови на подсистемот за пребарување на медицински трудови со спојување на зборови

и медицински концепти, а тоа се делот за пребарување на основа на зборови и делот за пребарување на основа на медицински концепти.

За пребарувањето според зборови се употребува системот за пребарување Terrier IR. Претпроцесирањето се изведува со токенизатор за англиски јазик, стемизацијата со Портер стемер [82], а бришењето на стоп зборовите според предефинираната листа во Terrier IR. Во фазата на пребарување се евалуираа повеќе модели на утежнување, како PL2 [104], BM25 [104], BB2 [104], DFR-BM25 [104], TF-IDF [105], DirichletLM [106]

Пребарувањето според медицинските концепти зависи од систем за мапирање (извлекување) на концепти од даден текст. За таа цел се користи Metamap, а извлечените концепти се UMLS [27] концепти. Мапирањето се изведува над текстот од медицинските трудови и прашањата. На основа на добиените концепти се генерираат нови репрезентации на медицинските трудови и прашања. Новогенерираните податоци се предаваат на Terrier IR за индексирање и пребарување. Претходно наведените модели за споредување се употребуваат и во овој дел.

Спојувањето на двата методи се прави со доцна фузија. Меѓутоа, претходен чекор е нормализација на поединечните резултати, бидејќи разнородни податоци може да имаат различни опсези на вредности. Со помош на мин-макс методот [123] резултатите од двата методи се доведуваат во опсег од 0 до 1. На овој начин се доведуваат податоците до ист опсег на вредности, без да се наруши дистрибуцијата. Впрочем, овој вид на нормализација се употребува и кај најдобрите методи за овој вид на пребарување над оваа база на податоци [61].

6.3.2 Експериментални прашања

Целта на овој дел од истражувањето е да се одговори на следните прашања:

1. Кој модел дава најдобри резултати при пребарување на медицински трудови на основа на зборови?
2. Какви резултати ќе даде пребарувањето на медицински трудови според концепти?
3. Дали спојувањето на двата начини на пребарување ќе даде подобри резултати?

Првото прашање се одговара со споредување на сите модели на пребарување. Со споредување на резултатите од пребарувањето според концепти и стандардното пребарување се дава одговор на второто прашање. Третото прашање се одговара со споредување на резултатите од третиот метод со резултатите од поединечните методи.

За евалуација се употребуваат следните метрики: усреднета средна прецизност (MAP), прецизност кога сите релевантни слики се најдени (Rprec), прецизност на првите 10 вратени слики (P10) и прецизност на првите 20 вратени слики (P20). Според стандардните практики за евалуација на овој проблем MAP се пресметува на првите 1000 слики за дадено прашање. Експериментите беа изведени над множеството ImageCLEF 2012.

6.3.3 Резултати и дискусија

Резултатите од пребарувањето на медицински трудови според зборови со различни модели на утежнување се прикажани на Табела 6.11. Според резултатите може да се забележи дека најдобри перформанси се добиваат со BM25 моделот со 0.1818 MAP.

Табела 6.11: Резултати од пребарување на медицински трудови на основа на зборови со различни модели за утежнување.

Модел	***MAP***	***Rprec***	***P10***	***P20***
BB2	0.1598	0.1604	0.1435	0.1326
BM25	0.1818	0.1757	0.1522	0.1391
DFR-BM25	0.1816	0.1767	0.1522	0.1413
TF-IDF	0.1805	0.1662	0.1522	0.1326
PL2	0.1780	0.1861	0.1478	0.1370
DirichletLM	0.1811	0.1744	0.1652	0.1283

Резултатите од пребарувањето на медицински трудови според медицински концепти со различни модели на утежнување се прикажани на Табела 6.12. Според резултатите може да се забележи дека најдобри перформанси се добиваат со DirichletLM моделот со 0.1073 MAP. Тоа е претежно заради начинот на кој се направени репрезентациите на трудовите и прашањата (без дуплирање на концептите). Имено, при креирањето на репрезентации на основа на концепти, секој концепт се додава само еднаш во репрезентацијата, независно дали го има повеќе пати во текстот. DirichletLM не го зема предвид бројот на појавувања на концептот во дадена репрезентација, туку само проверува дали го има концептот или не. Другите модели на утежнување зависат од овој параметар и затоа имаат послаби резултати за пребарување според медицински концепти.

Табела 6.12: Резултати од пребарување на медицински трудови на основа на медицински концепти со различни модели за утежнување.

Модел	***MAP***	***Rprec***	***P10***	***P20***
BM25	0.0705	0.0815	0.1000	0.0652
DFR-BM25	0.0706	0.0888	0.1000	0.0652
TF-IDF	0.0690	0.0815	0.0957	0.0609
BB2	0.0691	0.0874	0.1000	0.0630
PL2	0.0686	0.0835	0.0826	0.0674
DirichletLM	0.0841	0.0988	0.0957	0.0565

Табела 6.13: Резултати од пребарување на медицински трудови со доцна фузија.

Модел	***MAP***	***Rprec***	***P10***	***P20***
доцна фузија	0.1840	0.1917	0.1652	0.1391
MedGift	0.1690	/	0.1885	/

Резултатите од спојувањето на двата методи се прикажани на Табела 6.13. Овој експеримент е изведен со комбинирање на најдобрите експерименти од поединечните

методи. Резултатите покажуваат дека има подобрување на перформансите во однос на поединечните методи од приближно 0.2%. Подобрувањето настанува генерално заради тоа што постои одреден број на медицински трудови кои имале подобро рангирање при пребарување според медицински концепти, отколку при стандардното пребарување. Во овој случај, со едноставно конфигурирање на доцната фузија тие разлики во рангирањето на двата методи може до одредена мера да се намалат, а со тоа и да се зголеми прецизноста на крајните резултати.

6.4 Подсистем за пребарување на медицински трудови со проширување на прашања

6.4.1 Експериментални прашања

Целта на овој дел од истражувањето е да се одговори на следните прашања:

1. Дали може резултатите од основното пребарување да се подобрат со техники за проширување на прашањата?
2. Која техника за проширување на прашањата ќе даде најдобри резултати?

На првото прашање ќе одговориме преку споредување на резултатите од основното пребарување со резултатите од останатите пребарувања. Второто прашање ќе го одговориме преку споредување на резултатите од пребарувањата со проширување на прашањата.

Според горенаведените резултати за пребарување на медицински трудови на основа на зборови, како модел за утежнување се употребува BM25 во рамките на овие експерименти. Експериментите се изведени на множеството ImageCLEF 2013, а евалуацијата на резултатите се извршува со MAP, P10 и P30.

6.4.2 Резултати и дискусија

Со цел да се одговорат експерименталните прашања, направени се четири експерименти: 1. Основен експеримент - резултатите од основното пребарување, без проширување на прашањата; 2. Пребарување со MeSH концепти - резултатите од пребарувањето кога прашањата се проширени со MeSH термини; 3. Пребарување со UMLS концепти - резултатите од пребарувањето кога прашањата се проширени со UMLS термини; 4. Пребарување со псевдо-релевантна повратна врска - резултатите од пребарувањето кога прашањата се проширени со метод на псевдо-релевантна повратна врска.

Резултатите од експериментите се прикажани во Табела 6.14. Резултатите прикажуваат дека проширувањето на прашањата може да ги подобри перформансите на основното пребарување. Најголемото подобрување се појавува кога се употребува псевдо-релевантна повратна врска, каде е видливо подобрувањето во P10 и P30 метриките. Овие метрики се многу важни, во смисла на тоа дека ја репрезентираат прецизноста

што би ја искусиле крајните корисници, бидејќи корисниците обично ги гледаат најгорните вратени резултати.

Табела 6.14: Резултатите од евалуација на методите за проширување на прашањата.

	MAP	**P10**	**P30**
основното	0.2004	0.2029	0.1381
mesh концепти	0.1556	0.1800	0.1238
umls концепти	0.1625	0.1943	0.1343
псевдо-рел.	0.2005	0.2286	0.1667

Интересно е да се забележи дека основното пребарување се појавува како второ според перформансите во сите метрики. Тоа значи дека додавањето на MeSH и UMLS термини не придонесуваат во процесот на пребарување т.е. тие додаваат повеќе документи кои не се релевантни во горните резултати. Ова може да биде поради различни причини. Еден проблем може да биде тоа што прашања не се доволно дескриптивни за алатките за мапирање да извлечат соодветни концепти. Тоа значи дека извлечените термини не ги носат клучните концепти кои треба да се додадат на прашањето, што придонесува за други, нерелевантни трудови да се појават во горните резултати.

6.5 Подсистем за пребарување на медицински трудови со генерички бази за знаење

Експериментите кај овој дел од истражувањето треба да дадат одговор на следното прашање: *Дали проширувањето на прашањата со генеричка база на знаење може да придонесе за подобри перформанси на пребарувањето?*

За таа цел направивме два експерименти. Првиот експеримент е стандардното пребарување, каде подсистемот пребарува со оригиналното прашање без дополнителни модификации. Вториот експеримент се состои од пребарување со помош на методот за проширување на прашањата. Резултатите од експериментите се прикажани на Табела 6.15.

Табела 6.15: Резултати од подсистемот за пребарување на медицински трудови со генеричко знаење.

	MAP	**P10**	**P30**
основно	0.2004	0.2029	0.1476
freebase	0.2179	0.2086	0.1695

Според презентираните резултати може да се забележи дека вклучувањето на генеричките бази на знаење може да придонесе за подобрување на перформансите од пребарувањето. Резултатите дури и прикажуваат минимално зголемување на P10 и P30 метриките. Овие метрики се важни, бидејќи тие ги мерат резултатите за горните вратени трудови, што најчесто ги гледаат корисниците. Генерално, подобрувањето

на перформансите се должи на дескриптивноста на термините кои се извлечени како синоними од генеричките бази на знаење.

Глава 7

Заклучок

Во оваа докторска дисертација предложивме систем за пребарување на медицински документи со мултимодални податоци, што подразбира пребарување на медицински слики и трудови (случаи). Со цел да ги утврдиме карактеристиките на еден ваков систем, прво го опишавме проблемот кој се обидуваме да го решиме, а потоа направивме опширен преглед на постоечките техники за пребарување на медицински слики и трудови со посебен осврт на нивните недостатоци и отворените предизвици. На основа на тоа разививме нови алгоритми за мултимодално пребарување на медицински слики и методи за проширување на прашањата при пребарување на медицински трудови. Развиените алгоритми и методи ги имплементиравме во рамките на предложениот систем составен од повеќе подсистеми, при што секој решава различен дел од генеричкото пребарување.

Подсистемот за класификација на медицински слики според модалитет употребува текстуални и визуелни податоци за класификација на сликите. Во оваа фаза евалуиравме повеќе видови на визуелни дескриптори (LBP, FCTH, CEDD, SIFT и opponentSIFT) со кои ја репрезентиравме визуелната содржина на сликите. Текстуалниот дел го репрезентиравме на стандарден начин преку *bag-of-words* со модел за утежнување. Како класификатор користевме машини со носечки вектори со стратегија *еден-против-сите* за повеќе класна класификација. Експериментите ни покажаа дека рана фузија на визуелните дескриптори (во еден споен, комплескен дескриптор) дава најдобри перформанси за класификација на слики на основа на нивната визуелна содржина. Доцна фузија на класификаторите на основа на споениот визуелен дескриптор и текстуалните податоци прикажаа најдобри резултати воопшто пријавени за класификација над множествата на кои ние тестиравме.

Во подсистемот за пребарување на слики со мултимодални податоци демонстрираме три видови на пребарување кои ги поддржува истиот: текстуално базирано, содржински базирано и комбинирано со класификација на модалитет. Текстуално базираното пребарување го имплементиравме со Terrier IR и проширување на прашањата со псевворелевантна врска. Во однос на содржински базираното пребарување се фокусираме на RGB дескриптор со ADC структура за брзо пребарување. Пребарувањето со помош

на текстуални и визуелни карактеристики покажа подобри резултати од поединчените пребарувања. Вклучување на подсистемот за класификација на медицински слики според модалитет придонесе за подобрување на резултатите за 2-3%. Подобрувањето на резултатите најмногу се должи на фактот што прашањата кои беа праќани до подсистемот соджеа клучни зборови кои се однесуваат на модалитетот на барана слика, што ни овозможи подобро да ги рангираме сликите со тој модалитет. Добиено подобрување во перформансите ни овозможи да прикажеме најдобри пријавени резултати за множествата за пребарување на медицински слики од ImageCLEF 2011,2012 и 2013.

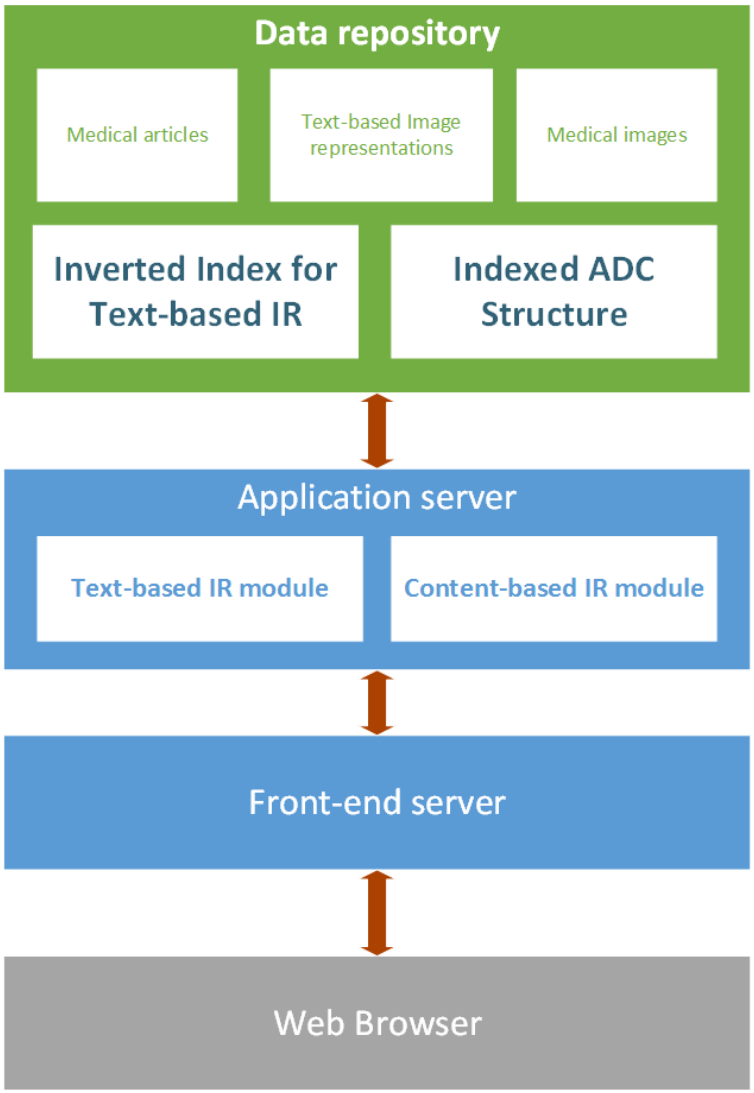
Во однос на пребарување на медицински трудови презентиравме неколку различни пристапи за решавање на овој проблем. Во подсистемот за пребарување на медицински трудови со спојување на зборови и медицински концепти претставивме метод кој употребува алатка за излекување на медицински концепти и истите ги користи за пребарување. Подсистемот креира две репрезентации за трудовите и прашањата. Една репрезентација е стандардна текстуална репрезентација, другата репрезентација е на основа на медицински концепти во трудот/сликата. Подсистемот креира посебни индекси за двата видови на репрезентации, а во фазата на пребарување врши две посебни пребарувања на двата индекси. Добиените резултати од двата индекси ги спојува и враќа на корисникот. Експриментите во овој дел ни прикажаа дека комбинирањето на двата методи дава подобри резлутати отколку само стандардно текстуално базирано пребарување.

Во делот за пребарување на медицински трудови со проширување на прашањата експериментиравме со неколку методи на проширување и тоа: проширување со MESH термини, проширување со UMLS термини и проширување со псевдо-релевантна врска. Проширувањето со псевдо-релевантна даде најдобри резултати во тој дел.

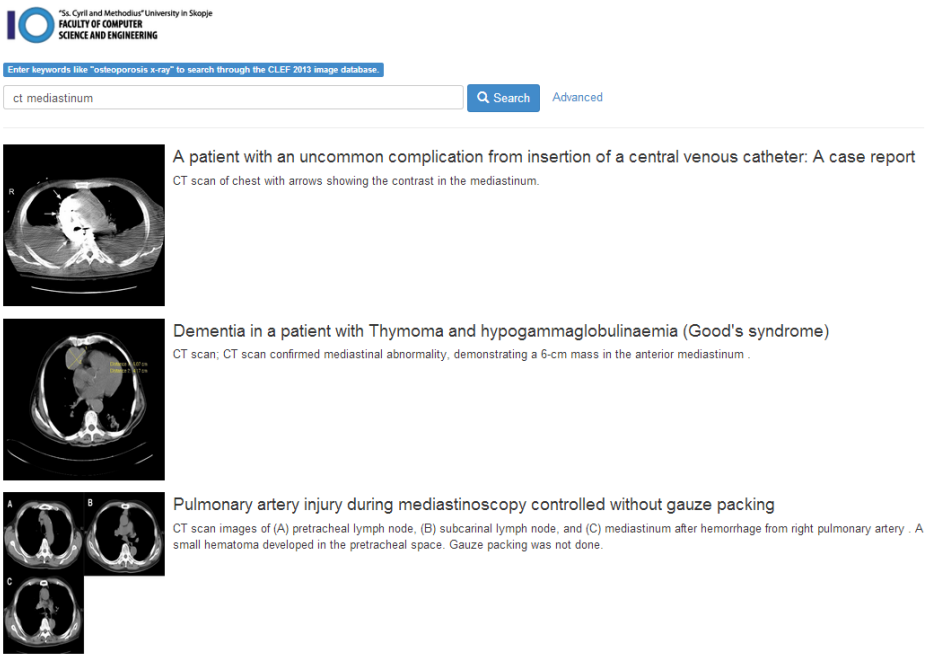
Презентиравме подсистем за пребарување на медицински трудови со генерички бази на знаење. Подсистемот фунционира на тој начин што за дадено прашање ги детектира медицинските концепти, а потоа ги наоѓа синонимите за тие концепти во генеричка база на знаење и ги додава во оригиналното прашање. Експериментите во овој подсистем ни демонстрираа дека вклучување на дополнителни бази на знаење може да доведе до подобрување на перформансите на пребарувањето.

Глава А

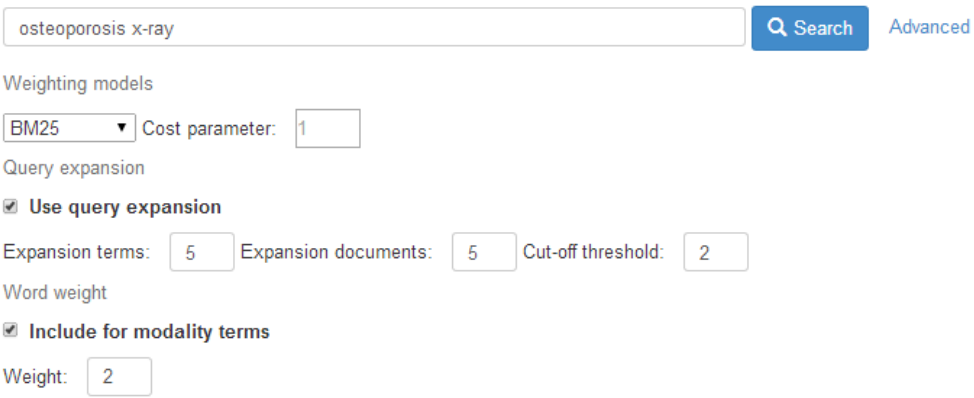
Додаток



Слика А.1: Архитектура на веб систем за пребарување на генерички медицински слики.



Слика А.2: Кориснички интерфејс на веб систем за пребарување на генерички медицински слики.



Слика А.3: Кориснички интерфејс за конфигурација на веб систем за пребарување на генерички медицински слики.

Табела А.1: Резултати од пребарување на медицински трудови со спојување на зборови и медицински концепти со проширување на прашањата со псевдо-релевантна повратна врска за најдобрите моделите за утежнување кај секој вид на пребарување. (1) Пребарување на медицински трудови на основа на зборови со BM25 со проширување на прашања. (2) Пребарување на медицински трудови на основа на медицински концепти со DirichletLM со проширување на прашања. (3) Доцна фузија на горните два експерименти.

Експеримент	***MAP***	***Rprec***	***P10***	***P20***
1	0.1928	0.1844	0.1826	0.1457
2	0.1073	0.1069	0.1261	0.0870
3	0.2127	0.2018	0.2000	0.1630

Табела А.2: Резултати од експерименти за оптимизација на пребарување на медицински слики на основа на зборови на ImageCLEF 2012. BM25-ww е пребарувањето со BM25 каде на одредени медицински зборови им е дадена поголема тежина.

Модел	***MAP***	***P10***	***P20***	***Rprec***	***# на рел. документи***
BB2	0.2056	0.3429	0.2714	0.2411	473
BM25	0.2266	0.3381	0.3000	0.2559	494
DFR-BM25	0.2091	0.3476	0.2738	0.2236	474
PL2	0.2055	0.3429	0.2643	0.2353	472
TF-IDF	0.2085	0.3524	0.2714	0.2194	471
DirichletLM	0.1601	0.2619	0.2024	0.1614	434
BM25-ww	0.2407	0.3619	0.2929	0.2620	490

Табела А.3: Резултати од експерименти за оптимизација на пребарување на медицински слики на основа на медицински концепти на ImageCLEF 2012.

Модел	***MAP***	***P10***	***P20***	***Rprec***	***# на рел. документи***
BB2	0.1257	0.1700	0.1025	0.1433	173
BM25	0.1230	0.1550	0.1025	0.1441	172
DFR-BM25	0.1227	0.1550	0.1000	0.1441	173
PL2	0.1065	0.1500	0.0950	0.1137	168
TF-IDF	0.1226	0.1550	0.1025	0.1402	172
DirichletLM	0.1568	0.2450	0.1475	0.1888	232

Табела А.4: Резултати од спојување на експериментите за оптимизација на пребарување на медицински слики на основа на зборови и медицински концепти на ImageCLEF 2012.

Тип	***MAP***	***P10***	***P20***	***Rprec***	***# на рел. документи***
Комбинирано	0.2385	0.3762	0.2738	0.2496	492
Комбинирано-ww	0.2528	0.3857	0.2690	0.2600	488

Библиографија

[1] Ivan Kitanovski, Gjorgji Strezoski, Ivica Dimitrovski, Gjorgji Madjarov, and Suzana Loskovska. Multimodal medical image retrieval system. *Multimedia Tools and Applications*, pages 1–24, 2016.

[2] Ivan Kitanovski, Ivica Dimitrovski, Gjorgji Madjarov, and Suzana Loskovska. Medical image retrieval using multimodal data. In *International Conference on Discovery Science*, pages 144–155. Springer International Publishing, 2014.

[3] Ivica Dimitrovski, Dragi Kocev, Ivan Kitanovski, Suzana Loskovska, and Sašo Džeroski. Improved medical image modality classification using a combination of visual and textual features. *Computerized Medical Imaging and Graphics*, 39:14–26, 2015.

[4] Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loshkovska. Merging words and concepts for medical articles retrieval. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 25–28. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013.

[5] Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loskovska. Improving medical cases retrieval using an online fact database. In *ICT Innovations 2016: Cognitive Functions and Next Generation ICT Systems*, volume 9. Springer International Publishing, 2016.

[6] Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loskovska. Multimodal medical image retrieval. In *ICT Innovations 2012*, pages 81–89. Springer, 2013.

[7] Ivan Kitanovski, Ivica Dimitrovski, and Suzana Loskovska. Fcse at medical tasks of imageclef 2013. In *CLEF (Working Notes)*, 2013.

[8] Ivan Kitanovski, Ivica Dimitrovski, and Suzana Loskovska. Fcse at imageclef 2012: Evaluating techniques for medical image retrieval. In *CLEF (Online Working Notes / Labs / Workshop)*, 2012.

[9] Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loskovska. Query expansion methods for text-based retrieval of medical articles. In *ICT Innovations 2015: Emerging technologies for better living*, 2015.

[10] Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loshkovska. Web-based system for textual retrieval of medical images. In *Proceedings of the 12th International Conference for Informatics and Information Technology*, 2015.

[11] Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loskovska. Text-based medical image retrieval using query modification methods. In *ICT Innovations 2014: World of Data*, 2014.

[12] Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, and Suzana Loshkovska. Word-space approach to case-based retrieval. In *Proceedings of the 11th International Conference for Informatics and Information Technology*, 2014.

[13] Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[14] Vannevar Bush. As we may think. In *Computer-supported cooperative work*, pages 17–34. Morgan Kaufmann Publishers Inc., 1988.

[15] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957.

[16] Gerard Salton. The smart retrieval system—experiments in automatic document processing. 1971.

[17] Cyril Cleverdon. The cranfield tests on index language devices. In *Aslib proceedings*, volume 19, pages 173–194. MCB UP Ltd, 1967.

[18] Donna K Harman. The first text retrieval conference (trec-1) rockville, md, usa, 4–6 november, 1992. *Information Processing & Management*, 29(4):411–414, 1993.

[19] P Pluye and RM Grad. How information retrieval technology may impact on physician practice: an organizational case study in family medicine. *Journal of evaluation in clinical practice*, 10(3):413–430, 2004.

[20] Pubmed. `http://www.ncbi.nlm.nih.gov/pubmed`. Accessed: 2017-04-22.

[21] Christopher Dye, John C Reeder, and Robert F Terry. *Research for universal health coverage*. World Health Organization, 2013.

[22] Jayashree Kalpathy-Cramer, Henning Muller, Steven Bedrick, Ivan Eggel, Alba Garcia Seco de Herrera, and Theodora Tsikrika. Overview of the clef 2011 medical image classification and retrieval tasks. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[23] Mounir Errami, Jonathan D Wren, Justin M Hicks, and Harold R Garner. etblast: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic acids research*, 35(suppl 2):W12–W15, 2007.

[24] Pubget. `http://www.nlm.nih.gov/pubs/factsheets/medline.html`. Accessed: 2016-06-09.

[25] Alba Garcia Seco de Herrera, Dimitrios Markonis, Ivan Eggel, and Henning Müller. The medgift group in imageclefmed 2012. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[26] Karam Abdulahhad, Jean-Pierre Chevallet, and Catherine Berrut. Mrim at imageclef2012. from words to concepts: A new counting approach. In *CLEF 2012-Conference on Multilingual and Multimodal Information Access Evaluation*, page 13p, 2012.

[27] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[28] Hong Wu, Kuangkai Sun, Xianzhi Deng, Yi Zhang, and Bili Che. Uestc at imageclef 2012 medical tasks. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[29] Dolf Trieschnigg, Piotr Pezik, Vivian Lee, Franciska De Jong, Wessel Kraaij, and Dietrich Rebholz-Schuhmann. Mesh up: effective mesh text classification for improved document retrieval. *Bioinformatics*, 25(11):1412–1418, 2009.

[30] Matthew S Simpson, Daekeun You, Md Mahmudur Rahman, Dina Demner-Fushman, Sameer Antani, and George R Thoma. Iti's participation in the imageclef 2012 medical retrieval and classification tasks. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[31] Jorge A Vanegas, Juan C Caicedo, Jorge E Camargo, Raul Ramos-Pollán, and Fabio A González. Bioingenium at imageclef 2012: Textual and visual indexing for medical images. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[32] Md Mahmudur Rahman Dina Demner-Fushman Sameer Antani Matthew S. Simpson, Daekeun You and George Thoma. Iti's participation in the 2013 medical track of imageclef. In *CLEF (Online Working Notes/Labs/Workshop)*, 2013.

[33] Sungbin Choi, Jeongeun Lee, and Jinwook Choi. Snumedinfo at imageclef 2013: Medical retrieval task. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, 2013.

[34] Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6. Citeseer, 2005.

[35] Pubget. `http://pubget.com`. Accessed: 2015-05-09.

[36] Thomas M Lehmann, Berthold B Wein, Joerg Dahmen, Joerg Bredno, Frank Vogelsang, and Michael Kohnen. Content-based image retrieval in medical applications: a novel multistep approach. In *Electronic Imaging*, pages 312–320. International Society for Optics and Photonics, 1999.

[37] Ashnil Kumar, Jinman Kim, Weidong Cai, Michael Fulham, and Dagan Feng. Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data. *Journal of digital imaging*, 26(6):1025–1039, 2013.

[38] Greg Pass and Ramin Zabih. Comparing images using joint histograms. *Multimedia systems*, 7(3):234–240, 1999.

[39] Dengsheng Zhang, Aylwin Wong, Maria Indrawan, and Guojun Lu. Content-based image retrieval using gabor texture features. In *IEEE Pacific-Rim Conference on Multimedia, University of Sydney, Australia*, pages 392–395, 2000.

[40] Greg Pass, Ramin Zabih, and Justin Miller. Comparing images using color coherence vectors. In *Proceedings of the fourth ACM international conference on Multimedia*, pages 65–73. ACM, 1997.

[41] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(6):460–473, 1978.

[42] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[43] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Citeseer, 1988.

[44] Krystian Mikolajczyk. *Interest point detection invariant to affine transformations*. PhD thesis, PhD thesis, Institut National Polytechnique de Grenoble, 2002.

[45] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[46] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[47] Yong Rui, Thomas S Huang, and Shih-Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of visual communication and image representation*, 10(1):39–62, 1999.

[48] Fuhui Long, Hongjiang Zhang, and David Dagan Feng. Fundamentals of content-based image retrieval. In *Multimedia Information Retrieval and Management*, pages 1–26. Springer, 2003.

[49] Stephen Crabbe, Peter Ambs2 Sue Black, Caroline Wilkinson, Jan Bikker, Norbert Herz, Daniel Manger, René Pape, and Helmut Seibert. Results of the fastid project.

[50] Ivica Dimitrovski, Dejan Gorgevik, and Suzana Loskovska. Web-based medical image retrieval system. In *Proceedings of the 10th International Multiconference INFORMATION SOCIETY*, pages 10–11, 2007.

[51] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sašo Džeroski. Fast and scalable image retrieval using predictive clustering trees. In *International Conference on Discovery Science*, pages 33–48. Springer Berlin Heidelberg, 2013.

[52] Seco de Herrera, Alba Garcia, Dimitrios Markonis, Roger Schaer, Ivan Eggel, and Henning Müller. The medgift group in imageclefmed 2013. In *CLEF working Notes 2013*, pages 1–12, 2013.

[53] Nefise Meltem Ceylan Okan Ozturkmenoglu and Adil Alpkocak. Demir at imageclefmed 2013: The effects of modality classification to information retrieval. In *CLEF (Online Working Notes/Labs/Workshop)*, 2013.

[54] Xin Zhou, Miaofei Han, Yanli Song, and Qiang Li. Fast filtering techniques in medical image classification and retrieval. In *CLEF (Working Notes)*, 2013.

[55] Flavio Martins Andre Mouruao and Jo ao Magalh aes. Novasearch on medical imageclef 2013. In *CLEF (Online Working Notes/Labs/Workshop)*, 2013.

[56] Antonia Kyriakopoulou Spyridon Stathopoulos, Ismini Lourentzou and Theodore Kalamboukis. Ipl at clef 2013 medical retrieval task. In *CLEF (Online Working Notes/Labs/Workshop)*, 2013.

[57] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sašo Džeroski. Fast and efficient visual codebook construction for multi-label annotation using predictive clustering trees. *Pattern Recognition Letters*, 38:38–45, 2014.

[58] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sašo Džeroski. Hierarchical classification of diatom images using ensembles of predictive clustering trees. *Ecological Informatics*, 7(1):19–29, 2012.

[59] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Saso Dzeroski. Hierarchical annotation of medical images. *Pattern Recognition*, 44(10-11):2436–2449, 2011.

[60] Henning Müller, Jayashree Kalpathy-Cramer, Ivan Eggel, Steven Bedrick, Saïd Radhouani, Brian Bakke, Charles E Kahn Jr, and William Hersh. Overview of the clef

2009 medical image retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 72–84. Springer, 2009.

[61] Alba G Seco de Herrera, Jayashree Kalpathy-Cramer, D Demner-Fushman, Sameer Antani, and Henning Müller. Overview of the imageclef 2013 medical tasks. *Working notes of CLEF*, 2013.

[62] Marti A Hearst, Anna Divoli, Harendra Guturu, Alex Ksikes, Preslav Nakov, Michael A Wooldridge, and Jerry Ye. Biotext search engine: beyond abstract search. *Bioinformatics*, 23(16):2196–2197, 2007.

[63] Aurélie Névéol, Thomas M Deserno, Stéfan J Darmoni, Mark Oliver Güld, and Alan R Aronson. Natural language processing versus content-based image analysis for medical document retrieval. *Journal of the American Society for Information Science and Technology*, 60(1):123–134, 2009.

[64] Thomas M Lehmann, Henning Schubert, Daniel Keysers, Michael Kohnen, and Berthold B Wein. The irma code for unique classification of medical images. In *Medical Imaging 2003*, pages 440–451. International Society for Optics and Photonics, 2003.

[65] Henning Müller, Jayashree Kalpathy-Cramer, Charles E Kahn Jr, and William Hersh. Comparing the quality of accessing medical literature using content-based visual and textual information retrieval. In *SPIE Medical Imaging*, pages 726405–726405. International Society for Optics and Photonics, 2009.

[66] Md Mahmudur Rahman, Daekeun You, Matthew S Simpson, Sameer K Antani, Dina Demner-Fushman, and George R Thoma. Multimodal biomedical image retrieval using hierarchical classification and modality fusion. *International Journal of Multimedia Information Retrieval*, 2(3):159–173, 2013.

[67] Songhua Xu, James McCusker, and Michael Krauthammer. Yale image finder (yif): a new search engine for retrieving biomedical images. *Bioinformatics*, 24(17):1968–1970, 2008.

[68] Charles E Kahn Jr and Cheng Thao. Goldminer: a radiology image search engine. *American Journal of Roentgenology*, 188(6):1475–1478, 2007.

[69] Jayashree Kalpathy-Cramer and William Hersh. Effectiveness of global features for automatic medical image classification and retrieval–the experiences of ohsu at imageclefmed. *Pattern recognition letters*, 29(15):2032–2038, 2008.

[70] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[71] Hugo Jair Escalante, Carlos A Hérnadez, Luis Enrique Sucar, and Manuel Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 172–179. ACM, 2008.

[72] Mark Montague and Javed A Aslam. Relevance score normalization for metasearch. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 427–433. ACM, 2001.

[73] Djoerd Hiemstra. *Using language models for information retrieval*. Taaluitgeverij Neslia Paniculata, 2001.

[74] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[75] Stephen E Robertson. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977.

[76] Melvin Earl Maron and John L Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244, 1960.

[77] Howard Turtle and W Bruce Croft. Inference networks for document retrieval. In *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1–24. ACM, 1989.

[78] Stephen E Robertson and K Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146, 1976.

[79] K Sparck Jones, Steve Walker, and Stephen E. Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information Processing & Management*, 36(6):809–840, 2000.

[80] Hinrich Schütze. Introduction to information retrieval. In *Proceedings of the international communication of association for computing machinery conference*, 2008.

[81] Joseph John Rocchio. Relevance feedback in information retrieval. 1971.

[82] Craig Macdonald, Vassilis Plachouras, Ben He, Christina Lioma, and Iadh Ounis. University of glasgow at webclef 2005: Experiments in per-field normalisation and language specific stemming. In *Accessing Multilingual Information Repositories*, pages 898–907. Springer, 2006.

[83] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[84] Savvas A Chatzichristofis and Yiannis S Boutalis. Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *International Conference on Computer Vision Systems*, pages 312–322. Springer, 2008.

[85] Djemel Ziou and Salvatore Tabbone. Edge detection techniques an overview. *Pattern Recognition and Image Analysis*, 8(24):537–559, 1998.

[86] Dong Kwon Park, Yoon Seok Jeon, and Chee Sun Won. Efficient use of local edge histogram descriptor. In *ACM workshops on Multimedia*, pages 51–54, 2000.

[87] Savvas A Chatzichristofis and Yiannis S Boutalis. Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval. In *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, pages 191–196. IEEE, 2008.

[88] Jianguo Zhang, Marcin Marszalek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

[89] Koen van de Sande, Theo Gevers, and Cees Snoek. Evaluating color fescriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[90] Koen EA van de Sande and Theo Gevers. University of amsterdam at the visual concept detection and annotation tasks. In *ImageCLEF*, pages 343–358. Springer, 2010.

[91] Jan C Van Gemert, Cor J Veenman, Arnold WM Smeulders, and Jan-Mark Geusebroek. Visual word ambiguity. *IEEE transactions on pattern analysis and machine intelligence*, 32(7):1271–1283, 2010.

[92] Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. Discriminative cue integration for medical image annotation. *Pattern Recognition Letters*, 29(15):1996–2002, 2008.

[93] Jayashree Kalpathy-Cramer and William Hersh. Automatic image modality based classification and annotation to improve medical image retrieval. In *MedInfo*, pages 1334–1338, 2007.

[94] National Electrical Manufacturers Association. Digital imaging and communications in medicine - DICOM. http://dicom.nema.org/, 2009.

[95] Mark O. Guld, Michael Kohnen, Daniel Keysers, Henning Schubert, Berthold B. Wein, Joerg Bredno, and Thomas M. Lehmann. Quality of DICOM header information for image categorization. In *SPIE vol. 4685 - Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation*, pages 280–287, 2002.

[96] Otis Gospodnetic and Erik Hatcher. *Lucene*. Manning, 2005.

[97] Nicholas C Ide, Russell F Loane, and Dina Demner-Fushman. Essie: a concept-based search engine for structured biomedical text. *Journal of the American Medical Informatics Association*, 14(3):253–263, 2007.

[98] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[99] William Hersh, Susan Price, and Larry Donohoe. Assessing thesaurus-based query expansion using the umls metathesaurus. In *Proceedings of the AMIA Symposium*, page 344. American Medical Informatics Association, 2000.

[100] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier information retrieval platform. In *Advances in Information Retrieval*, pages 517–519. Springer, 2005.

[101] Ben He and Iadh Ounis. A study of the dirichlet priors for term frequency normalisation. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 465–471. ACM, 2005.

[102] Christina Lioma and Iadh Ounis. Examining the content load of part of speech blocks for information retrieval. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 531–538. Association for Computational Linguistics, 2006.

[103] Giambattista Amati. Information theoretic approach to information extraction. In *International Conference on Flexible Query Answering Systems*, pages 519–529. Springer, 2006.

[104] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[105] Djoerd Hiemstra. A probabilistic justification for using tf× idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2):131–139, 2000.

[106] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM, 2001.

[107] Matthijs Douze, Arnau Ramisa, and Cordelia Schmid. Combining attributes and fisher vectors for efficient image retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 745–752. IEEE, 2011.

[108] Henry J Lowe and G Octo Barnett. Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *Jama*, 271(14):1103–1108, 1994.

[109] Imageclef. `http://www.imageclef.org/`. Accessed: 2016-10-30.

[110] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[111] Alan R Aronson, James G Mork, Clifford W Gay, Susanne M Humphrey, and Willie J Rogers. The nlm indexing initiative's medical text indexer. *Medinfo*, 11(Pt 1):268–72, 2004.

[112] Rong Yan, Alexander Hauptmann, and Rong Jin. Multimedia search with pseudo-relevance feedback. In *Image and Video Retrieval*, pages 238–247. Springer, 2003.

[113] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008.

[114] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[115] Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68:267–276, 2007.

[116] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

[117] Marcin Marszałek, Cordelia Schmid, Hedi Harzallah, and Joost van de Weijer. Learning object representations for visual object class recognition, 2007. In *Visual Recognition Challenge workshop, in conjunction with ICCV, Rio de Janeiro, Brazil*.

[118] Henning Müller, Alba Garcia Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, Sameer Antani, and Ivan Eggel. Overview of the imageclef 2012 medical image retrieval and classification tasks. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[119] Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. 2011.

[120] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *Advances in neural information processing systems*, pages 1753–1760, 2009.

[121] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):117–128, 2011.

[122] Lei Zheng, Arthur W Wetzel, John Gilbertson, and Michael J Becich. Design and analysis of a content-based pathology image retrieval system. *Information Technology in Biomedicine, IEEE Transactions on*, 7(4):249–255, 2003.

[123] Anil Jain, Karthik Nandakumar, and Arun Ross. Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285, 2005.