მაია არჩუაძის სადოქტორო დისერტაციის

„ქართულენოვანი ტექსტების კლასიფიკაციის ამოცანა ინფორმაციულ ძებნაში“

(სამეცნიერო ხელმძღვანელი: თსუ პროფესორი მანანა ხაჩიძე)

## ანოტაცია

ნაშრომში წარმოდგენილია არასტრუქტურირებული დოკუმენტების დაჭდევების ახალი მეთოდი, რომელიც გამოყენებულია ქართულენოვანი ტექსტების კლასიფიცირების პროცესის განსახორციელებლად. მეთოდი ეფუძნება ანალიტიკური ევრისტიკების მეთოდით, ცნების „პატერნების“ ცოდნის ბაზის ფორმირებას ტექსტების კლასიფიკაციისათვის და პირველად იქნა გამოყენებული ასეთი ტიპის ამოცანაში.

ინფორმაციის ძებნის პროცესი არ წარმოადგენს ერთგვაროვან ოპერაციას. მისი წარმატებულობა და რელევანტურობა დამოკიდებულია ძებნის ციკლის ადეკვატურობაზე და სისრულეზე. ამ ციკლში ერთ-ერთი მნიშვნელოვანი ადგილი უკავია კლასიფიკაციის ეტაპს, რომლითაც, როგორც წესი, იწყება ძებნის პროცესი.

ძებნის სტრატეგიები, იდენტიფიცირებულია, როგორც ძებნის მოდელები. ინფორმაციული ძებნისთვის განხილულია ბულის მოდელი, ვექტორული სივრცის მოდელი და ალბათური ძებნის მოდელი, მათი მუშაობის თავისებურებანი და ის დადებითი და უარყოფითი თვისებები, რომლებიც ახასიათებს თითოეულ მათგანს. ძებნის მოდელები ეფუძნება ტერმინების წონის დათვლის პრინციპს. ტერმინს წონა არის სტატისტიკური სიდიდე, რომელიც განისაზღვრება დოკუმენტში ტერმინის შეხვედრის სიხშირით და განსაზღვრავს ტერმინის მნიშვნელოვნებას. შესაბამისად, ძებნის მოდელების ფუნქციონირების ჩარჩოებში, განხილულია ტერმინის წონის განსაზღვრის მეთოდები, რომელთა ნაწილი მუშავდება ალბათური ძებნის მოდელების საზღვრებში, ხოლო ნაწილი რეალიზდება ვექტორული სივრცის მოდელის ფარგლებში.

ამასთანავე, აღწერილია ინფორმაციული ძებნის ამოცანებში ბუნებრივი ენის ანალიზის მეთოდები, მათი ძირითადი ეტაპები და ის მნიშვნელოვანი თვისებები, რომელიც ახასიათებს თითოეულ მათგანს. ბუნებრივი ენის ანალიზი მოიცავს სინტაქსსა და სემანტიკაზე დაყრდნობით ცოდნის ამოღებას ბუნებრივ ენაზე დაწერილი დოკუმენტებიდან. ასეთი მიდგომა შეიძლება განვიხილოთ, როგორც „სემანტიკური“ მიდგომა იმ ლოგიკით, რომ დოკუმენტის შინაარსი და სტრუქტურა განისაზღვრება არასტატისტიკური მეთოდებით. დღეისათვის სტატისტიკური/ალბათური მეთოდებისა და სინტაქსური/სემანტიკური მეთოდების ინტეგრაცია იდეალური გამოსავალია ძებნის პროცესის ეფექტურობის გაზრდისათვის.

ნაშრომში განხილულია ტექსტის საწყისი დამუშავების პროცესები, რომელთა განხორციელება აუცილებელია კლასიფიკაციის საწყის ეტაპზე. განხვიხილეთ სტემინგისა და ლემატიზაციის პროცესი, რომელიც წარმოადგენს დოკუმენტების დამუშავების უმნიშვნელოვანეს ეტაპს. საუბარია სტემინგის პოპულარულ ალგორითმებზე - ლოვინსის(Lowins), პორტერის (Porter) და პაის/ჰასკის (Pice/Hask) ალგორითმებზე.

განხილულია სტემინგის ალგორითმების გამოყენების თავისებურებანი ანალიზურ და სინთეზურ ენებში და, ამ თავისებურებათა გათვალისწინებით, მათი მოდიფიკაციის აუცილებლობა სხვადასხვა ენის კორპუსისათვის, თუმცა არსებობს ისეთი ენებიც, რომელთა დამუშავება მოითხოვს საერთოდ ახალი სტემერის შექმნას.

არსებული სტემინგის ალგორითმების გამოყენება ქართული ენის თავისებურებებიდან გამომდინარე შეუძლებელი გახდა, ამიტომ ქართულენოვანი ტექსტების კლასიფიკაციის ამოცანაში (მსგავსად სხვა ენებისა), ტექსტის დამუშავებისათვის შემუშავებულ იქნა სტემინგის ახალი ალგორითმი. იგი ეფუძნება სიტყვების და სუფიქსების ბაზას და ეფექტურად მუშაობს სიტყვის კვეცის პრობლემებზე.

განხილულიაბუნებრივიენისდამუშავებისმეთოდებიკლასიფიკაციისამოცანებში, კერძოდ, კონცეპტებზედაფუძნებულიინფორმაციულიძებნა.

აღწერილია ინფორმაციული ძებნის ორი მნიშვნელოვანი პრობლემა: სინონიმია და პოლისემია, გადაჭრის სხვადასხვა მეთოდი და მათი ქმედითი ასპექტები. აღნიშნული მეთოდებიდან გამოვყავით ლატენტური სემანტიკური და ზუსტი სემანტიკური ანალიზის მეთოდები, როგორც საუკეთესო აღნიშნული პრობლემების გადასაჭრელად.

დღეისათვის კლასიფიკაციის ამოცანა შეიძლება განიხილოს, როგორც მანქანური სწავლებისა და ინფორმაციული ძებნის მეთოდების ერთობლიობა. ნაშრომში დავახასიათეთ კლასიფიკაციის ამოცანებში ყველაზე ხშირად გამოყენებადი სამი ალგორითმი, რომლებიც გამოყენებული იქნა ჩვენს მიერ განხორციელებულ კვლევებში. ესენია: უახლოესი მეზობლის ალგორითმი, მხარდამჭერი ვექტორების ალგორითმი და ბაიესის ალგორითმი. განვიხილეთ ყველა თავისებურება, რომელიც ახასიათებთ თითოეულს მუშაობის სხვადასხვა საფეხურზე.

სემანტიკური მიდგომის თავისებურებას უმთავრესად წარმოადგენს ის ფაქტი, რომ გამოიყენება დოკუმენტების კონცეპტუალური წარმოდგენა, რომელიც იქმნება საგნობრივი არის ცოდნის სემანტიკურ მოდელებზე დაყრდნობით, ხოლო ცოდნის წარმოდგენის არსებულ ინსტრუმენტებს შორის, ონტოლოგია წარმოადგენს ყველაზე გამოსახვით ხერხს. ჩვეულებრივ ონტოლოგიებში საგნობრივი არეების ცოდნა აღიწერება ცნებებისა და თვისებების იერარქიით, ასევე შეერთებული ცნებების ეგზემპლარების სემანტიკური ქსელებით. არსებულ მიდგომებთან შედარებით, ონტოლოგიის გამოყენებამ შესაძლოა მოგვცეს საშუალება გავაუმჯობესოთ ძებნის ხარისხი. ამიტომ მნიშვნელოვანია საგნობრივი არის აღმწერი ცოდნის ბაზის შემუშავების მოქნილი ალგორითმის შექმნა. ჩვენს მიდგომაში ეს ალგორითმი ეფუძნება ანალიტიკური ევრისტიკების მეთოდს.

ეს მეთოდი ეფუძნება აკად. ვლადიმერ ჭავჭანიძის მეთოდს, რომელიც ცნობილია კონცეპტების ფორმირების ანალიტიკური ევრისტიკების მეთოდის სახელით.

მეთოდი მოიცავს სხვადასხვა ეტაპებს: ნიშანთვისებათა ბინარიზაცია; ნიშანთვისებათა გადაკოდირება; ორთონორმირებული ბინარული მდგომარეობის ვექტორების აგება; ფილტრაციის ოპერაცია; დიზიუნქციური სუპერპოზიციის ოპერაცია; ბულის ცვლადებზე პირობითი გადასვლის ოპერაცია.

ანალიტიკური ევრისტიკების მეთოდის გამოყენებით, დოკუმენტების კოლექციიდან მოხდა კონკრეტული ცნების აღმწერი „კონცეპტ-პატერნების“ შემუშავება. პორტერის ალგორითმის მოდიფიცირებული ვარიანტით განხორციელდა კლასის აღმწერი ნიშან-თვისებების სიმრავლის გამოყოფა და შესაბამისად წონების დათვლა tf-idf სტანდარტული სქემით.

ამ ამოცანის განხორციელების შემდეგ, მოხდა მისი პრაქტიკული რეალიზაცია ქართულენოვანი სამედიცინო ჩანაწერების კლასიფიკაციისათვის. წარმოდგენილი იქნა სამდიცინო ჩანაწერების კლასიფიცირების მეთოდი ქართულენოვანი ტექსტებისათვის.

კვლევებისათვის გამოყენებული იქნა 24.855 ჩანაწერი. დოკუმენტების კლასიფიკაცია განხორციელდა სამ ძირითად ჯგუფად (ულტრასონოგრაფია, ენდოსკოპია, რენტგენი) და 13 ქვეჯგუფად. ამოცანის გადაწყვეტისათვის გამოყენებული იქნა ორი კარგად ცნობილი მანქანური სწავლების ალგორითმი: მხარდამჭერი ვექტორების (SVM) დაუახლოესი მეზობლის (KNN). შედეგებმა აჩვენა რომ მანქანური სწავლების ორივე მეთოდი საკმაოდ შედეგინია, მაგრამ უკეთესი შედეგი გამოვლინდა SVM-ის გამოყენებისას. კლასიფიკაციის პროცესში ჩვენს მიერ შემუშავებული იქნა თვისებათა ამოკრების მეთოდის ჩვენი ვარიანტი, ე. წ. „ „შეკუმშვის“ მეთოდი, რომელმაც კლასიფიკაციის პირველ დონეზე საკმაოდ კარგი შედეგი მოგვცა. თუმცა მეორე დონეზე, რომელიც ქვეკლასებად კატეგორიას მოიცავდა შედეგები ცოტა გაუარესდა. დოკუმენტების 23%-ის მიკუთვნება ინდივიდუალური კლასებისათვის არ განხორციელდა. შედეგის გაუარესება გამოიწვია დაავადებათა აღწერაში გამოყენებული ტერმინების ერთგვაროვნებამ. ქვეკლასების მიხედვით კვლევის შედეგების დაფიქსირებისას მოხდა ერთიდაიგივე ტერმინების გამოყენება, რაც შედეგის შეცვლის მიზეზი გახდა.

აღსანიშნავია, რომ ეს არის ასეთი ტიპის ტექსტების კლასიფიკაციის პირველი მცდელობა. ზოგადად, ქართულენოვანი ტექსტებისათვის მსგავსი ამოცანა აქამდე არ განხორციელებულა.

## Annotation

In the presented work the document "labeling" method for classification process is provided. The method is based on knowledge base formation using concept "patterns" for text classification.

Information retrieval process does not represent the outcome of only one type of operation. Its success and relevance is based on retrieval cycle recall and adequacy. One of the important parts of this cycle is the stage of classification- which represents the initial stage of text retrieval.

The strategies of retrieval are identified as the models of retrieval. The following basic retrieval models along with their peculiarities, pros and cons are considered: the Boolean model, Space Vector Model (SVM) and the probabilistic model. The models of retrieval are based on the principle of term weight calculation. The Weight of term represents the statistical value defined according to the frequency of its appearance in text and is defining the term value. Thus, in frames of retrieval model functionality, the methods of term frequency determination, partially in statistical retrieval models and partially in Vector Space models, is considered.

The methods on natural language processing, along their main stages and significant properties, for information retrieval task is described as well. The Natural Language Processing (NLP) contains the knowledge acquisition based on syntax and semantics of the provided natural language text. Such an approach can be considered as "semantic" based on logic that the content and structure of document is defined using non-statistical methods. Nowadays the integration of statistic/probabilistic models with syntactic/semantical models are considered to be the best solution for increasing the effectiveness of the retrieval process.

Work address the text initial Processing as a necessary part of text classification initial stage, particularly the process of Stemming and Lemmatization. The well-known and popular algorithms such as Lovins, Porter and Paice/Husk stemming algorithms are considered.

The peculiarities of stemming algorithm applications in analytical and Synthetic languages are provided and the necessity of their modification for different language Corps are underlined. However, there are languages requiring the development of a new stemmers for them.

In the work provided we considered natural language processing methods for classification task – namely the task of concept pattern based information retrieval.

We have described the two main problems of Information Retrieval: Polysemy and Synonyms along with their solution methods. From the methods considered we have selected the methods of Latent Sematic Analysis and Exact sematic analysis as the best suitable methods for the posed problem solution.

The problem of classification can be considered as a union of machine learning and IR methods. The following most common three algorithms in problems of classification: the K Nearest Neighbor (KNN), Support Vector Machine (SVM) and Bayes algorithms have been described, used later on in our research.

The semantic approach peculiarity is that the conceptual representation of document is applied, based on semantic models of subject area knowledge. From tools of knowledge representation the ontology represents the most suitable one. Generally in ontologies the subject area knowledge is described using hierarchy of notions and characteristics as well as joints notions semantic network entities. Application of Ontologies might lead to improvement of retrieval quality. This is why it is so important to develop an algorithm of subject area describing knowledgebase construction. In our case such algorithm is based on method of Heuristic Analytics.

The method considered is based on the method called "Pattern Formation Heuristic Analysis" developed by Academician Vladimer Chavchanidze.

The method consists of several stages: binarization of characteristics, re-coding of characteristics, orthonormal binary vector construction, operation of filtration, disjunctive superposition operation and Boolean variable transfer conditional operation.

Using the method of Heuristic Analysis, the certain notion defining concept-patterns were developed. Using the modified Porter algorithm, the class defining characteristics have been selected and appropriate weights have been calculated using the standard tf-idf scheme.

The practical realization of the task considered was fulfilled for medical data based document collection. The Georgian language based medical data classification method was presented.

The presented work introduces the instrument for Georgian-language-based medical records classification. It is the first attempt of classification of the Georgian-language-based medical records. Totally, 24.855 examination records were studied. The documents were classified into three main groups (Ultrasonography, Endoscopy, X-Ray) and 13 subgroups using two well-known methods: Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). The results obtained demonstrated that both machine learning methods performed successfully, with a little supremacy of SVM. In the process of classification a "shrink" method - based on features selection - was introduced and applied. At the first stage of classification the results of the "shrink" case were better, however on the second stage of classification into subclasses 23 % of all documents could not be linked to only one definite individual subclass (liver or binary system) due to common features characterizing these subclasses. The overall results of the study were successful.

We have to note that it was the first attempt of classification for such type of texts.