БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ • BULGARIAN ACADEMY OF SCIENCES

ГЕОХИМИЯ, МИНЕРАЛОГИЯ И ПЕТРОЛОГИЯ • 38 • СОФИЯ • 2001
GEOCHEMISTRY, MINERALOGY AND PETROLOGY • 38 • SOFIA • 2001

Програмата за идентификация на минерали MinMatch - решение на някои принципни проблеми при ползуване на минераложки бази данни

Томас Керестеджиян, Емил Алкалай

Kerestedjian, T., E. Alkalai. 2001. The mineral identification program MinMatch: A solution of some principal problems in the use of mineralogical databases. – *Geochem. Mineral. Petrol*., **38,** 121-124.

Abstract. The presented piece of software is a powerful mineral identification tool. It allows profound and flexible searches of mineral data, matching a set of user provided observations and analytical data. It is a search-match routine, working on a self-contained database covering over 3800 mineral species. Taking grounds of authors experience in using other existing mineralogical databases, the friendly user interface overrides some typical problems, caused by the disadvantages of the formal logic. The extremely compact bitwise internal representation of the searchable data, resulted in a very handy code of 320 K. The program is a JAVA applet, usable on any operating system and accessible through the Internet as well. The beta version of the program is available for testing and evaluation purposes and authors would be grateful for any bug reports and improvement proposals fed back.

Key words: mineral identification, database, search-match, program
Address: Geological Institute, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria; E-mail: thomas@geology.bas.bg

Ключови думи: идентификация на минерали, база данни, търсене, програма
Адрес: Геологически институт, Българска академия на науките, 1113 София

Процесът на идентификация на непознат минерал, по принцип представлява сравняване на установените свойства с литературните данни за свойствата на известните минерали. С други думи, това е задача от типа ‘search-match”, типична за приложението на компютърните бази данни. Ето защо, с навлизането на персо-налните компютри в широка употреба, една от основните задачи в минералогията – идентификацията на минерали, се прехвърли от традиционното използуване на печатни справочни източници на информация, към употребата на съответни бази данни. Ползуването на бази данни има определени преимущества, главните от които са: бързина, удобство, изчерпателност. Особено важно за качеството на извършваната идентификация е последното: при правилното използуване на бази данни не може да бъде пропусната възможност, което често се случва при работа с печатни източници. Например, ако търсите минерал с химичен състав As, при работа с печатно пособие ще намерите самороден арсен и вероятно ще спрете до там, но при употреба на база данни ще бъдете предупредени за възможността арсеноламприт, която е много близка и по редица от останалите си свойства, но се различава по пространствена група на симетрия. Така употребата на база данни бързо ще Ви насочи към допълни-

телно изследване, което еднозначно ще идентифицира минерала.

Употребата на бази данни, обаче се сблъсква с определени трудности, произтичащи от някои специфики на минераложката информация, както и от специфики на съществуващите за целта бази данни [1, 2]:

1. Значителна част от свободно (безплатно) достъпните бази данни представляват слабо структурирани текстови описания, разделени по записи (за всеки минерал), но не и по полета (за отделните свойства), а в случаите в които има полета за данни, тези полета съдържат съвкупност от данни за даденото свойство, отразяваща вариабилността на естествения природен материал. Тази ситуация има сериозно икономическо основание, тъй като функционалното структуриране на обилната и разнообразна минераложка информация изисква продължителните усилия на голям колектив от компетентни учени. Това е работа изискваща значително финансиране със слаба възвращаемост. За разлика от този подход, създаването на неструктурирана (текстова) база данни е бързо и лесно ако се използува сканиране и разпознаване (OCR) на налична печатна информация. Търсенето в такава база данни се извършва по съвпадение на текстови последователности (string), но този подход има значителни недостатъци. Най-очевидният от тях е фактът, че всяка печатна грешка, било в базата данни, било във въведената за търсене последователност, води до пропускане на смислени идентификационни възможности. Много по-неизбежен, обаче, е проблемът с различните възможности за описание на дадено свойство. Особено остро този въпрос стои с такива свойства като цвят на минерала, където традиционно се използуват интуитивни, описателни термини. Очевидно е, че вероятността такова интуитивно описание, да срещне точно текстово съвпадение в базата данни е малка.

2. Особено неуспешно е търсенето на съвпадение по химичен състав при горния подход. Така например ако търсите минерал съдържащ C, ще намерите също минералите съдържащи Ca, Ce, Cl и т.н. поради формалното съвпадение на главната буква C. Съществуват обаче и много по-тежки за решаване принципни проблеми. Ако например търсите минерал съдържащ по микросондови данни само Fe и Si и добавите O (строго погледнато би трябвало да прибавите всичките 10 елемента с атомен номер до 10, защото не се определят директно от метода, но биха могли да участвуват) ще получите 11 възможни минерала отговарящи на това условие, но верният избор най-вероятно няма да бъде сред тях. Причината за това е във факта, че значителен брой от този тип силикати съдържа във формулата си и други метали, споделящи позициите на желязото и представени в скоби заедно с него. Въпреки че вашият конкретен образец може да бъде краен член на изоморфна редица и да не съдържа други елементи, общата формула на минерала изисква формалното им добавяне. Формалната липса на тези елементи в зададения въпрос много често води до изключване на верния избор. Простото дописване на елементи в текста на заявката за търсене обаче не решава въпроса. От една страна, това е дописване "по нюх", което не е много коректно, а от друга – води до значително нарастване на броя на възможностите и обезсмисля процеса на идентификация. Причината за това е, че дописаните елементи могат да се срещат и самостоятелно, в други структурни позиции (извън скобите), прибавяйки по този начин други типове силикати към списъка на възможностите. Решението на въпроса е в разпознаване на положението на елемента във формулата, което съществуващите текстови бази данни не позволяват да бъде направено.

3. Някои физични свойства като твърдост, относително тегло и др. се описват със значителен толеранс в текстовите бази данни, което отразява реалната вариабилност на природния материал. Идентификацията по текстово съвпадение обаче не позволява употребата на сравнителни релации (повече, по-малко, над, под, и т.н.).

Така ако търсеният минерал има относително тегло 5,5–6,5, а ние сме измерили 5,8, формалната логика ще изключи верният избор, поради липсата на точно съвпадение. Освен това трябва да се има предвид, че в този тип бази данни цифровата информация от формална гледна точка е представена от буквени знаци и например 5,7 не представлява числото пет цяло и седем, а последователност от знаците пет, десетична запетая и седем. По тази причина, дори ако въведем при търсене стойност отговаряща точно на стойността в базата данни, но използуваме десетична точка вместо запетая, съвпадението няма да бъде разпознато.

4. Липсата на информация за дадено свойство в базата данни може да доведе до неоснователно изключване на правилния избор. Например, ако определяния минерал е зелен и вие по други свойства сте достигнали до малка извадка от възможности, съдържаща верния избор, но в базата данни липсва информация за цвета на този минерал, при проверка по този критерий ще се достигне до формално несъвпадение и правилната идентификация ще бъде пропусната.

5. Поради текстовата форма на информацията, съществуващите бази данни са много обемисти. Това често свежда ползуването им до невъзможност при работа с бавен мрежов обмен.

Решение на всички описани проблеми предлага създадената от нас специализирана програма за идентификация на минерали MinMatch. При създаването на тази програма информацията от няколко свободно достъпни текстови бази данни беше прочетена, анализирана и структурирана от специален програмен модул. След това структурираната информация беше представена във възможно най-икономичен бит код, който сведе обема на годните за съпоставяне данни до няколко стотин килобайта.

Потребителският интерфейс за достъп до тези битово компресирани данни беше изграден на базата на JAVA код [3, 4], позволяващ ползуването му на всякаква компютърна конфигурация и операционна система, както и чрез мрежов достъп (Internet). Самата база данни е вградена в аплета като JAVA class файл.

Като типичен JAVA аплет, програмата има висока степен на сигурност, защото няма права за модифициране съдържанието на твърдия диск и няма достъп за четене извън собствената си работна директория. Така потребителят може да бъде сигурен, че ползуването на програмата не може да навреди по никакъв начин на инсталирания софтуер и не може за разкрие лични данни за потребителя.

Цялостният обем на програмата е 720K, което я прави година за пренасяне върху една единствена дискета или може да бъде прехвърлена по мрежа с нормална скорост на обмен за няколко минути. При работа от Internet, малкият обем на програмата осигурява висока скорост на зареждане на аплета. Тъй като цялата необходима информация е вградена в аплета, след зареждане програмата работи off-line, което е практично, особено при платен достъп до Internet.

Дизайнът на интерфейса следва общоприетите принципи за изграждане на потребителски интерфейс, като предлага генериране на заявка за търсене само чрез избор от предварително зададени опции. По този начин възможността за въвеждане на некоректна заявка се елиминира напълно.

Основният интерфейс съдържа отделни табла за работа с различните свойства на минералите. Типичната употреба на програмата ползува отделните табла като филтри. Всяко табло изключва от общия брой минерали (3850) тези, които не отговарят на въведените критерии. Ако потребителят е доволен от резултата, прехвърля извадката в главния интерфейс (бутон OK), където може да види и пълния списък със свойства за всеки минерал. Ако желае да коригира критериите от това табло, може да откаже прехвърлянето на извадката (бутон CANCEL) и да опита отново. Всяко следващо табло (следващо свойство на минерала) извършва филтрация върху вече филтри-

раната от предходното табло извадка. Отделните табла могат да бъдат използвани в произволна последователност, определяна от потребителя по целесъобразност. Таблата в тази първа версия на програмата са 9: химичен състав, твърдост, относително тегло, хабитус, цвят, цепителност, сингония, блясък, класификационна принадлежност. Програмата е отворена за лесно добавяне на още табла (още свойства).

Най-значителните достойнства на програмата са в таблото химичен състав, тъй като именно там бяха решени най-тежките проблеми на съществуващите бази данни. Таблото съдържа бутони с името на всеки химичен елемент, подредени във формата на менделеева таблица. С последователно натискане на всеки отделен бутон (елемент) неговият цвят може да бъде променен, както следва: зелен – елементът задължително присъства в минерала; жълт – елементът е разрешен в произволно положение във формулата, но не е задължителен; сив – елементът е разрешен само като изоморфен (в скоби) с някой от задължителните (зелени) елементи и не е задължителен; червен – елементът е забранен. Таблото има и допълнителни бутони за опции: само задължителните (зелени) елементи са разрешени – равносилно е на забрана на всички останали елементи; всички избрани (зелени) елементи трябва да присъстват в минерала – ако тази опция не е включена филтърът работи с логическо "или", т.е. ако присъства който и да е от избраните елементи, минералът остава в извадката; използуване на групи – бутонът включва допълнителен набор от бутони за химични групи: SiO_4, CO_3, PO_4 и др. Има и бутони за улеснен избор: забрани всички елементи - всички елементи стават червени; разреши леките елементи – разрешава елементите с атомен номер до 10, което е практично при работа с данни от микросонда; разреши всички елементи – всички елементи стават жълти; разреши всички елементи като изоморфни – всички елементи стават сиви. Има и допълнителна опция, както и всички други табла, за включване в извадката на минералите за които отсъстват съответни данни. Големият набор от възможности в това табло прави изграждането на заявката изключително гъвкаво. Трябва да се отбележи, че при този набор от възможности съществува повече от един правилен начин да се извърши филтрация на данните, което позволява индивидуален подход за конкретния потребител.

Таблата твърдост и относително тегло извършват сравнение на диапазони от стойности. Таблата хабитус, сингония и блясък извършват сравнение по едно или повече определения избрани от фиксиран списък възможности. Таблото за цвят работи по подобен начин, но позволява дефинирането на до 4 цвята, като за всеки от тях може да бъде определен и нюанс. Таблото за цепителност дефинира степен и (или) посока на цепителност. Таблото класификация предлага филтрация по класификационни групи на Дейна и (или) Щрунц.

Програмата ще бъде предложена чрез Internet за тестване от заинтересуваните потребители.

Литература

1. Mineralogical database, David Barthelmy – http://www.webmineral.com/
2. Athena mineralogy, Pierre Perroud - http://un2sg4.unige.ch/athena/mineral/mineral.html
3. "JAVA tutorial for JAVA II platform" – www.java.sun.com
4. "JAVA documentation for JAVA II platform (JDK 1.3)" – www.java.sun.com

Приета на 16. 11. 2001. г.
Accepted November 16, 2001