ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი

მიხეილ წერეთელი

მონაცემთა ბაზებში ინფორმაციის საძიებო სისტემის მოდელი

სამაგისტრო ნაშრომი შესრულებულია საინფორმაციო ტექნოლოგიების მაგისტრის აკადემიური ხარისხის მოსაპოვებლად

ხელმძღვანელი: მანანა ხაჩიძე

მაია არჩუაძე

თბილისი 2013

ანოტაცია

ნაშრომში განხილულია ინფორმაციის ძებნა, როგორც პროცესი და ის თანამედროვე ალგორითმები, რომლებიც გამოიყენება ძებნის სისტემებში. გაანალიზებულია ამ ალგორითმების სუსტი და ძლიერი მხარეები. წარმოდგენილია მონაცემთა ძებნის პროცედურები მონაცემთა ბაზებში და სემანტიკური ძებნის თავისებურებები. შემუშავებულია ძებნის „ჰიბრიდული“ მოდელი, რომელიც წარმოადგენს არსებული სამი მოდელის ეტაპობრივ გამოყენებას. მოდელში ძებნის პროცესი განიხილება, როგორც სამ დონიანი პროცესი, რომელსაც წინ უძღვის მოთხოვნის ტიპის იდენტიფიცირების ეტაპი და შედეგების რანჟირების ეტაპი.

Annotation

The paper deals with the search of information as to the process and the algorithms that are used in search engines. Analyzed the strengths and weaknesses of these algorithms. The database search procedures for databases and semantic search features. Designed for "hybrid" model, which represents the gradual application of the model. Model of the search process is treated as a three-level process is preceded by the step of identifying the type of request and the results of the ranking stage.

შინაარსი

შესავალი

თანამედროვე ტექნოლოგიების განვითარებასთან ერთად, სულ უფრო მეტად იზრდება ინფორმაციის მოცულობა, რაც ბუნებრივია ართულებს ჩვენთვის საჭირო ინფორმაციის ადვილად და დროულად მოძიებას. უკანასკნელი ათწლეულის განმავლობაში ინფორმაციის ძიების ოპტიმიზაციის პროცესი გასაკუთრებით მნიშვნელოვანი გახდა. ინფორმაციის ძებნისათვის უმეტეს წილად გამოიყენება ვებ სივრცე, რომელიც მალე გახდა ყველაზე ხშირად გამოყენებადი სივრცე. მაგალითისთვის ჯერ კიდევ 2004 წელს ინტერნეტ მომხმარებლებს შორის ჩატარებულმა გამოკითხვამ აჩვენა, რომ გამოკითხულთა 92% ფიქრობდა: ინტერნეტ სივრცე არის ყველაზე კარგი საშუალება ყოველდღიური ინფორმაციის მოსაძიებლად.

მიუხედავად იმისა, რომ ინფორმაციული ტექნოლოგიების სამომხმარებლო სივრცეში მუდმივად ჩნდებიან სხვადასხვა საძიებო სისტემები ინფორმაციის ძებნის ეფექტური ალგორითმების შექმნა კვლავ აქტუალურია. განსაკუთრებით აქტუალურია ინფორმაციის ნაკადის მზარდ მოცულობაში მომხმარებლის მოთხოვნის შესაბამისი (რელევანტური) ინფორმაციის მიღების შესაძლებლობა. აქ იკვეთება ორი ძირითადი მიმართულება: პასუხი მოთხოვნაზე იყოს რაც შეიძლება ზუსტი (სემანტიკური თვალსაზრისით) და 2. მოთხოვნაზე პასუხი იყოს რაც შეიძლება სწრაფი. სემანტიკური თვალსაზრისით ზუსტი პასუხის მისაღებად აუცილებელია თვით მომხმარებლის მოთხოვნა წარმოსდგეს ისეთი ფორმალიზებული სახით, რომ არ დაირღვეს მისი შინაარსობრივი მნიშვნელობა.

ინფორმაციულ ძებნასთან დაკავშირებული საკითხები შეიძლება განვიხილოთ ორი მიმართულებით: ძებნა ინტერნეტ ნაკადში - ვებ სივრცეში და ძებნა მონაცემთა ბაზებში. თუ გავითვალისწინებთ მონაცემთა ბაზების მიმართ მომხმარებელთა მზარდ ინტერესს, ნებისმიერი ახალი მიგნება მონაცემთა ბაზაში ინფორმაციის ძებნის ეფექტურობის თვალსაზრისით მეტად აქტუალურია.

ნაშრომში შემოთავაზებულია მონაცემთა ბაზებში ინფორმაციის ძებნის „ჰიბრიდული“ ალგორითმი, რომლის რეალიზაციაც შესაძლებელია დანართის სახით მონაცემთა ბაზის მართვის სისტემებში.

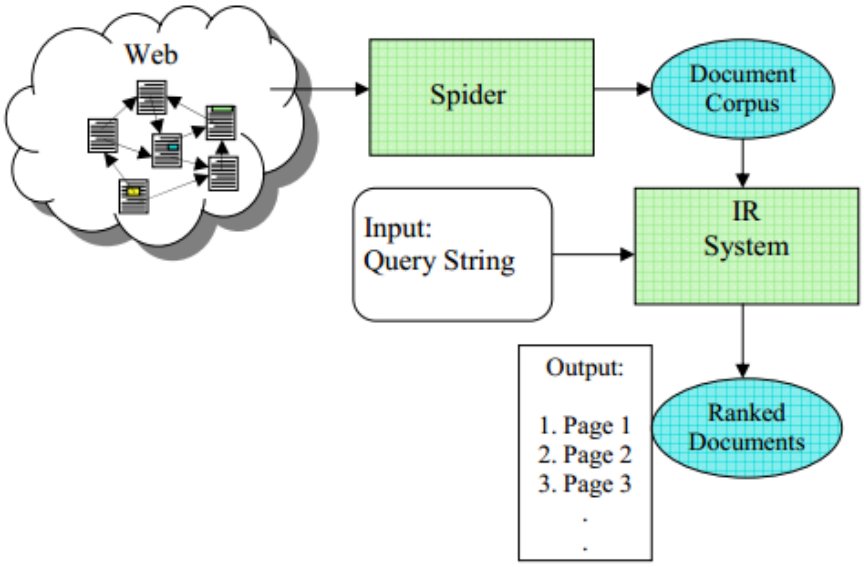


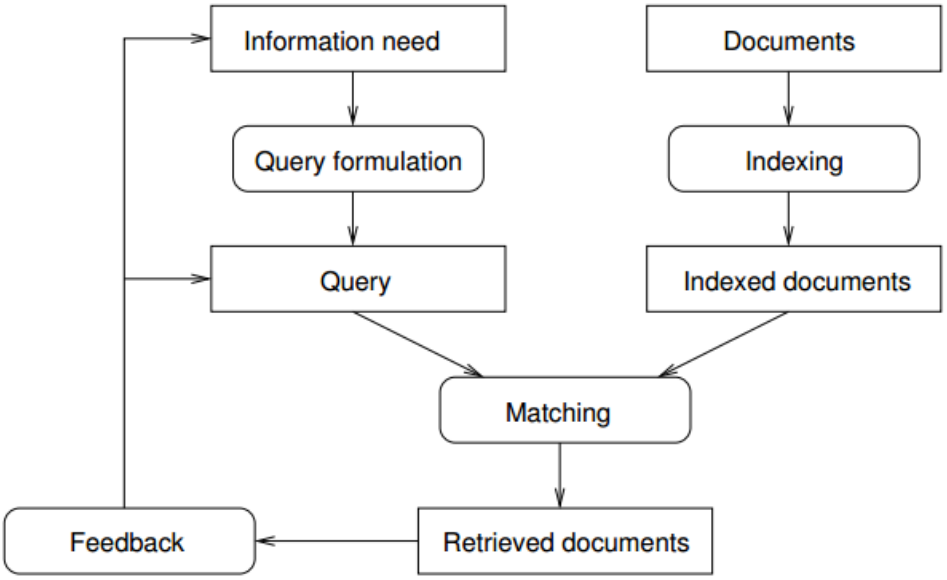
Figure 1. The simplified architecture of a search engine.

ინფორმაციის ძებნა, როგორც პროცესი და ინსტრუმენტი

ინფორმაციის ძებნის, როგორც პროცესის, აქტუალურობა ზრდის მოთხოვნას შეიქმნას ისეთი ტიპის პროგრამული სისტემები და ინსტრუმენტები (მათ შორის დანართებიც) რომლებიც მომხმარებლის მოთხოვნაზე დააბრუნებენ ადექვატურ პასუხს რაც შეიძლება მცირე დროში. მეტ წილად ყურადღება მახვილდება ინფორმაციის საძიებო სისტემის/პროგრამის შემუშავებაზე. ინფორმაციის ძებნის სისტემა არის პროგრამა, რომელიც ინახავს, მართავს ინფორმაციას უმეტესად ტექსტურ დოკუმენტებში, თუმცა შესაძლოა შენახული იქნას ასევე მულტიმედიურ ფაილებში. სისტემა ეხმარება მომხმარებლებს მისთვის საჭირო ინფორმაციის მოძიებაში. მას არ შეუძლია ცალსახად დააბრუნოს ინფორმაცია ან უპასუხოს კითხვებს, ნაცვლად ამისა სისტემა ატყობინებს მომხმარებელს იმ დოკუმენტის არსებობის და ადგილმდებარეობის შესახებ, რომელიც შეიცავს მისთვის სასურველ ინფორმაციას. ასეთ დოკუმენტს ეწოდება *რელევანტური* დოკუმენტი. კარგმა ძებნის სისტემამ უნდა დააბრუნოს მხოლოდ რელევანტური დოკუმენტები, თუმცა ასეთი ძებნის სისტემა არ არსებობს და არც იარსებებს, ვინაიდან საძიებო ტექსტის ფორმულირება უმეტეს წილად არასრული იქნება, შესაბამისად რელევანტური დოკუმენტის ძებნა დამოკიდებულია მომხმარებლის სუბიექტურ აზრზე. პრაქტიკაში შესაძლებელია ორმა მომხმარებელმა საძიებელ სისტემას მისცეს ერთი და იგივე მოთხოვნა, სისტემამ დააბრუნოს მაქსიმალურად რელევანტური დოკუმენტი, თუმცა მოძიებული ინფორმაცია ერთი მომხმარებლისთვის შესაძლოა იყოს სასურველი, ხოლო მეორე მომხმარებლისთვის - არა.

არსებობს სამი საბაზისო პროცესი, რომელსაც მხარს უნდა უჭერდეს ინფორმაციის საძიებო სისტემა: დოკუმენტის შიგთავსის წარმოდგენა, მომხმარებლის მიერ მოთხოვნილი ინფორმაციის წარმოდგენა, აღნიშნული ორი მოთხოვნის შედარება.

სქემატურად აღნიშნული პროცესი შეიძლება წარმოვადგინოთ ნახაზის სახით (ნახ 1.)



ნახ 1. ინფორმაციის ძებნის პროცესი

ინფორმაციის ძებნის სისტემაში დოკუმენტის წარმოდგენის პროცესს ეწოდება ინდექსაცია, რომელიც მიმდინარეობს ავტომატურად, მომხმარებელი არ იღებს უშუალო მონაწილეობას პროცესში. ხშირად საძიებო სისტემა იყენებს ძალიან ტრივიალურ, მარტივ ალგორითმებს ინდექსირებული ტექსტის წარმოსადგენად, მაგალითად ალგორითმი რომელიც ეძებს სიტყვებს ინგლისურ ტექსტში უბრალოდ ასოებს აყენებს ერთ რეგისტრში, რის შედეგადაც მომხმარებელს შეუძლია უგულებელყოს რეგისტრები. ინდექსაციის პროცესი მოიცავს მის შენახვას სისტემაში, ხშირად ინახება მხოლოდ დოკუმენტის ნაწილი მაგალითად სათაური, ძირითადი აბსტრაქტული ინფორმაცია პლიუს ინფორმაცია დოკუმენტის ადგილმდებარეობის შესახებ.

მომხმარებლისათვის საჭირო ინფორმაციის მოძიება ხშირად დამოკიდებულია, მის მიერ მოთხოვნის ფორმულირების პროცესზე. მომხმარებლის მოთხოვნისა და სისტემის მიერ წარმოდგენილი დოკუმენტის თანაფარდობას ეწოდება შედარებას პროცესი (*matching*). ამ პროცესის შედეგი არის დოკუმენტების სია, რომელთაგანაც მომხმარებელმა უნდა მოძებნოს ის დოკუმენტი, რომელიც

ყველაზე მეტად შეესაბამება მის მოთხოვნებს. სისტემის მიერ დაბრუნებული დოკუმენტების სიაში რელევანტური დოკუმენტები უნდა იყოს სიის სათავეში, ამ შემთხვევაში მცირდება მომხმარებლის მიერ დახარჯული დრო, რომელიც ჭირდება მას აღნიშნული დოკუმენტის წასაკითხად. მარტივი მაგრამ ეფექტური რეიტინგული ალგორითმი იყენებს დოკუმენტში საძიებო სიტყვების სიხშირული განაწილების მეთოდს და ასევე სტატისტიკურ ინფორმაციას, დოკუმენტში ბმულების რაოდენობის შესახებ. სტატისტიკაზე დაფუძნებული რეიტინგული ალგორითმები ანახევრებს მომხმარებლის დროს, რომელიც ჭირდება მას დოკუმენტის წასაკითხად.

ინფორმაციის ძებნის თანამედროვე მეთოდები

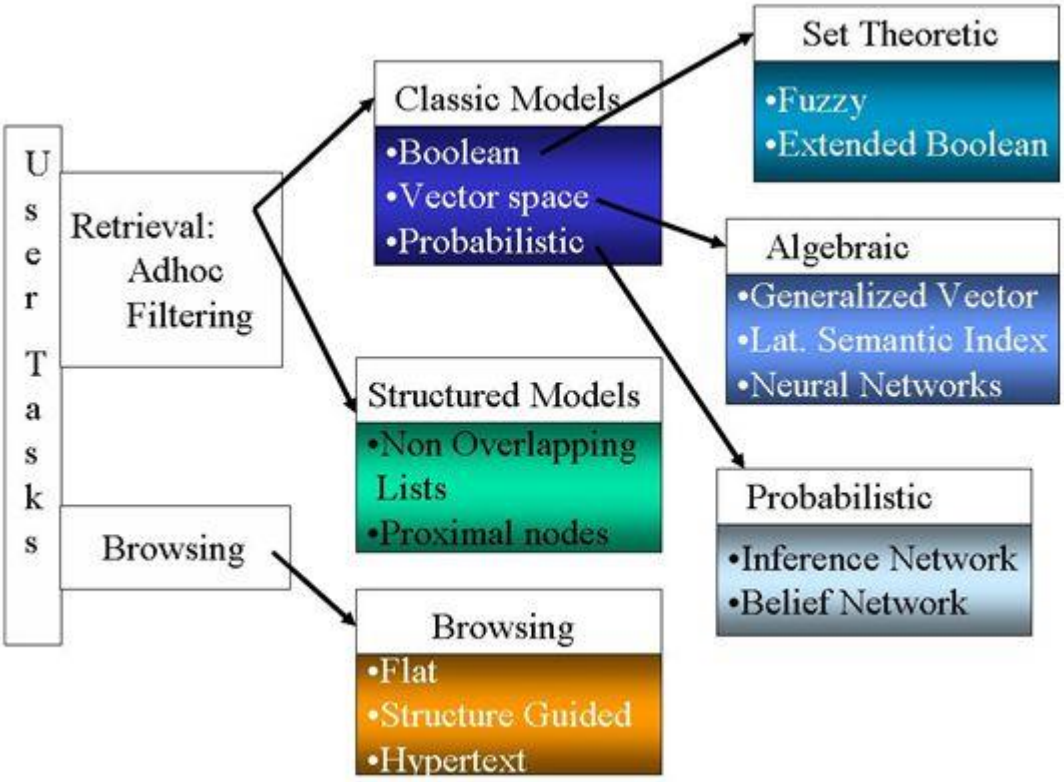
ინფორმაციის ძებნის მეთოდების შეფასების 2 ძირითადი შეფასების კრიტერიუმი არსებობს: I. ძებნის მეთოდები ეძებენ არა რელევანტურ დოკუმენტებს, II. ისინი ვერ ეძებენ ყველა რელევანტურ დოკუმენტს. პირველ შეფასებას ეწოდება ზუსტი შეფასება (*precision rate*) ანუ პროპორციული შეფარდება მოძებნილი დოკუმენტი არის თუ არა რელევანტური, ხოლო მეორე შეფასებას ეწოდება მსგავსი/გახსენებითი შეფასება (*recall rate*) - პროპორციული თანაფარდობა, ყველა მოძებნილი დოკუმენტი არის თუ არა რელევანტური. თუ ინფორმაციის მაძიებელს სურს მეტი სიზუსტე, მან უნდა შეავიწროვოს /შეამციროს მოთხოვნა. ნახ.2-ზე წარმოდგენილია ინფორმაციის ძებნის ცნობილი თანამედროვე მოდელები და მათი კლასიფიცირება გამოყენების თვალსაზრისით.

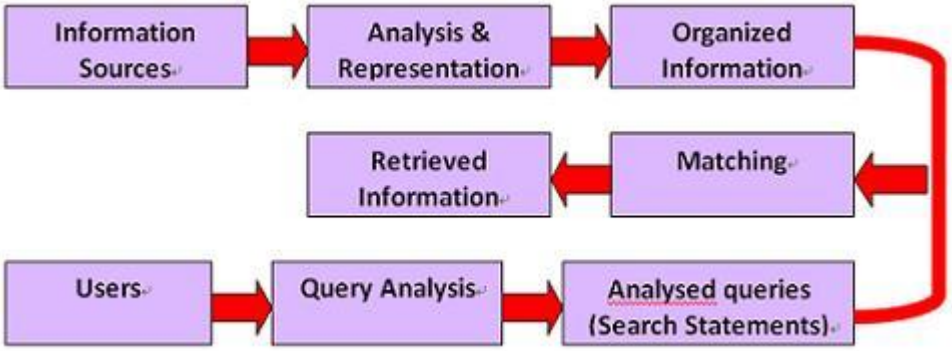
ინფორმაციის ძებნის გავრცელებული მეთოდი იყოფა ორ მოდელად ესენია:

- ლოგიკური (boolean) მოდელი
- სტატისტიკური (Statistical) მოდელი:

პირველ მოდელს მიეკუთვნება Boolean Information Retrieval (BIR) ინფორმაციის ძებნის ლოგიკური მოდელი, ხოლო სტატისტიკურ მოდელს:

- Vector Space Model ინფორმაციის ძებნის ვექტორული სივრცის მოდელი
- Probabilistic Model ინფორმაციის ძებნის ალბათური მოდელი


User Tasks
Retrieval: Adhoc Filtering
Browsing
Classic Models
•Boolean
•Vector space
•Probabilistic
Structured Models
•Non Overlapping Lists
•Proximal nodes
Browsing
•Flat
•Structure Guided
•Hypertext
Set Theoretic
•Fuzzy
•Extended Boolean
Algebraic
•Generalized Vector
•Lat. Semantic Index
•Neural Networks
Probabilistic
•Inference Network
•Belief Network


Information Sources
Analysis & Representation
Organized Information
Retrieved Information
Matching
Users
Query Analysis
Analysed queries (Search Statements)

ნახ.2

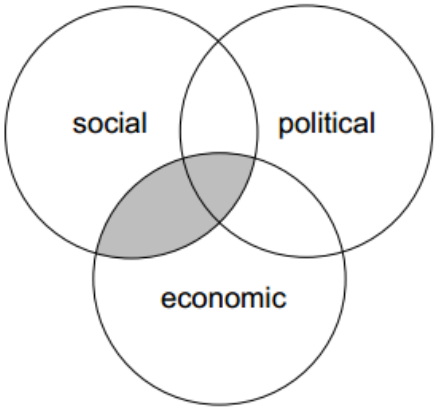
ინფორმაციის ძებნის ლოგიკური მოდელი

IR Boolean model

პირველი ინფორმაციის ძებნის მოდელი არის ინფორმაციის ლოგიკური ძებნის მოდელი. ეს მოდელი ახდენს მოთხოვნის შინაარსობრივ ანალიზს და მას უსადაგებს შესაბამის რელევანტურ დოკუმენტს. მაგალითად მოთხოვნა სახელწოდებით economic მარტივად განსაზღვრავს, იმ დოკუმენტებს რომლების ინდექსირებული არიან განმარტებით economic.

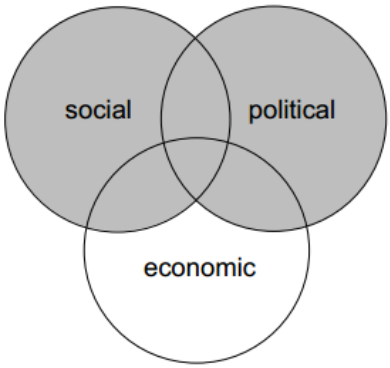
ლოგიკური მოდელი იყენებს George Bool-ის მათემატიკურ ლოგიკას. იგი მოთხოვნის ტერმინებს და დოკუმენტის ინდექსირებულ აღწერებს უსადაგებს ერთმანეთს, რათა შექმნას ახალი ტიპის დოკუმენტები. ამ მოდელში განსაზღვრულია სამი საბაზისო ოპერატორი. ესენია: ოპერატორები “AND“ , “OR” და “NOT”.

“AND” ლოგიკური ოპერატორის საშუალებით დაკავშირებულ მოთხოვნების კომბინირება ხდება შემდეგნაირად: უნდა მოხდეს მოთხოვნაში ოპერატორით დაკავშირებულ თითოეულ საძიებო სიტყვის დამთხვევა დოკუმენტში ინდექსირებისას განსაზღვრულ განმარტებებთან. მაგალითისათვის მოთხოვნამ “Social AND Economic” უნდა დააბრუნოს დოკუმენტი, რომელიც დაინდექსებული იქნება ორივე მათგანით როგორც განმარტებით „Social“ ასევე განმარტებით Economic“.

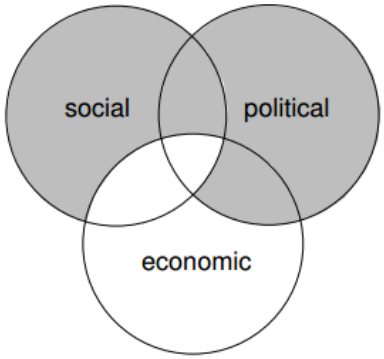


social AND economic

“OR” ლოგიკური ოპერატორის საშუალებით დაკავშირებულ მოთხოვნების კომბინირება ხდება შემდეგნაირად: მოთხოვნაში ოპერატორით დაკავშირებული საძიებო სიტყვების დამთხვევა უნდა მოხდეს რომელიმე განმარტებით ინდექსირებულ დოკუმენტთან. მაგალითისათვის მოთხოვნამ “Social OR Political” უნდა დააბრუნოს „Social“ განმარტებით დაინდექსებული დოკუმენტი ან „Political“ განმარტებით დაინდექსებული დოკუმენტი ან ორივე მათგანით ერთად.



social OR political



(social OR political) AND NOT (social AND economic)

განვიხილოთ სტანდარტული ლოგიკური ძებნის მოდელი და განვსაზღვროთ მისი დადებითი და უარყოფითი მხარეები:

<u>ძლიერი/ დადებითი მხარეები:</u>

- არის ადვილად დასანერგი და კომპიუტერულად ეფექტური/ქმედითი (სტანდარტული მოდელი ფართომაშტაბიანი, ოპერატიული ძებნის სისტემა, რომელიც დიდ ონლაინ სერვისებს იყენებს).
- საშუალებას აძლევს მომხმარებელს გადმოსცეს სტრუქტურული და კონცეპტუალური შეზღუდვები, რათა შეძლოს მნიშვნელოვანი ლინგვისტური თვისებების აღწერა. (მოთხოვნაში შესაძლებელია გამოყენებულ იქნას სინონიმური სპეციფიკაციები და ფრაზები).
- ლოგიკური ძებნის მეთოდს გააჩნია მკაფიოდ და ნათლად გამოხატვის შესაძლებლობა. (მოთხოვნა უნდა იყოს ამომწურავი და არაორაზროვანი).
- ეს მეთოდი გვთავაზობს ბევრ მეთოდს მოთხოვნის გასაფართოებლად ან დასაკონკრეტებლად.
- აქვს მეტი ეფექტი ძებნის შემდეგ ეტაპზე, მეტი სიზუსტე ცნებებსა და დამოკიდებულებებს შორის.

<u>უარყოფითი მხარეები:</u>

- მომხმარებელს უჭირს ეფექტური ლოგიკური მოთხოვნის შემუშავება (მომხმარებლები იყენებენ ენის კავშირებს “AND“ „OR“,“NOT“, რომლებსაც მოთხოვნის შემთხვევაში გააჩნია განსხვავებული მნიშვნელობა, ვინაიდან აღიქვამენ, როგორც ინგლისური ენის კავშირებს)

აღნიშნული ძებნის მოდელი მუშაობს „ზუსტი დამთხვევის“ (exact match) პრინციპით

	Standard Boolean
Goal	• Capture conceptual structure and contextual information
Methods	• Coordination: AND, OR, NOT • Proximity • Fields • Stemming / Truncation
(+)	• Easy to implement • Computationally efficient => all the major on-line databases use it • Expressiveness and Clarity Synonym specifications (OR-clauses) and phrases (AND-clauses).
(-)	• Difficult to construct Boolean queries. • All or nothing AND too severe, and OR does not differentiate enough. • Difficult to control output: Null output <-> Overload. • No ranking • No weighting of index or query terms • No uncertainty measure

ნახ.3 ინფორმაციის ლოგიკური ძებნის მეთოდი.

„Vector Space Model“ ინფორმაციის ძებნის ვექტორული სივრცის მოდელი და „Probabilistic Model“ ინფორმაციის ძებნის ალბათური მოდელი არის ორი ძირითადი მიდგომა ინფორმაციის სტატისტიკური ძებნის მეთოდისათვის. ორივე მოდელი იყენებს სტატისტიკურ ინფორმაციას სიხშირის განსაზღვრაში, რათა შეუსაბამოს რელევანტური დოკუმენტი მოთხოვნას. ინფორმაციის ძებნის სტატისტიკურ მეთოდებში გადაჭრილია ლოგიკური ძებნის მეთოდის ზოგიერთ სუსტი მხარე თუმცა მას აქვს თავისი ნაკლოვანებებიც.

ინფორმაციის ძებნის ვექტორული სივრცის მოდელი

„Vector Space Model“

Peter Luhn იყო პირველი, რომელმაც ინფორმაცის ძებნისას გამოიყენა სტატისტიკური მიდგომა. მისი ვარაუდის მიხედვით, მომხმარებელმა საჭირო დოკუმენტების მოსაძებნად უნდა მოამზადოს ისეთი დოკუმენტი რომელიც მაქსიმალურად მსგავსი იქნება საჭირო დოკუმენტებისა. მსგავსების ხარისხი, მომზადებულ დოკუმენტსა და დოკუმენტების კოლექციას შორის გამოიყენება ძიების რეზულტატის რეიტინგის განსასაზღვრად.

რაც უფრო მეტი იქნება თანხვედრა/შეთანხმება წარმოდგენებსა და მოცემულ ელემენტებს და მის განაწილებას შორის, მით მეტი იქნება ალბათობა იმისა, რომ მოძიებული იქნება საძიებო ინფორმაციის ანალოგიური/მსგავსი.

აღნიშნულ მოდელში დოკუმენტები და მოთხოვნები წარმოდგება მრავალგანზომილებიან სივრცეში, რომლის განზომილებები გამოიყენება ინდექსის შესაქმნელად, დოკუმენტის აღწერისათვის. ინდექსის შექმნის მიზანია მნიშვნელოვანი ასპექტების ლექსიკური სკანირება, მორფოლოგიური ანალიზი საძიებო სიტყვის „აზრობრივად“ მსგავსი ალტერნატივების შერჩევისათვის. დოკუმენტის ან მისი სუროგატის/შემცვლელის მოთხოვნის დროს შედარება ხდება ვექტორების გამოყენებით, მაგალითად გამოიყენება კოსინუსის მსგავსი გაზომვა.

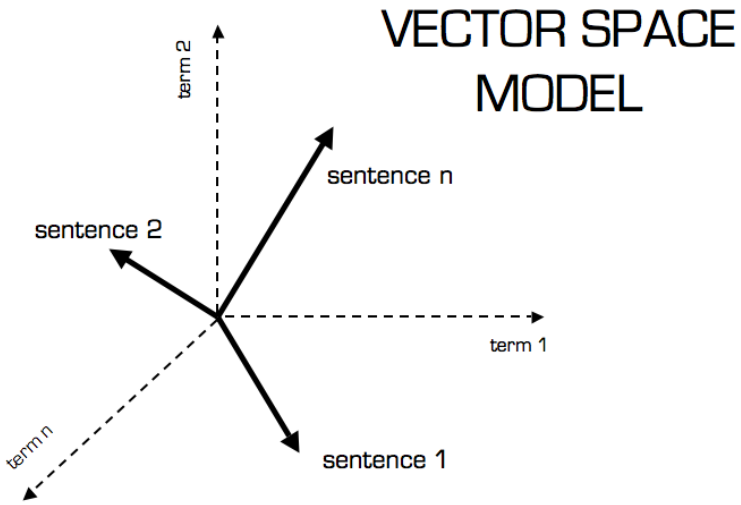
თუ დოკუმენტის ინდექსური წარმოდგენა არის ვექტორი $\vec{d} = (d_1, d_2, \dots d_m)$, რომლის თითოეული კომპონენტი $d_k (1 \leq k \geq m)$ დაკავშირებულია ინდექსის აღწერასთან და მოთხოვნაც არის ვექტორი $\vec{q} = (q_1, q_2, \dots q_m)$ რომლის კომპონენტები დაკავშირებულია იმავე აღწერასთან, მაშინ პირდაპირი მსგავსების ზომა არის ვექტორების გაერთიანება.

$$Score\left(\vec{d}, \vec{q}\right) = \sum_{k=1}^{m} d_k * q_k$$

თუ ვექტორის კომპონენტები ორობითი რიცხვებია, მაშინ კომპონენტის მნიშვნელობა იქნება 1 თუ აღწერა არსებობს დოკუმენტში ან მოთხოვნაში და აღწერის მნიშვნელობა იქნება 0 თუ არ არსებობს. შესაბამისად პირდაპირი მსგავსების ზომა არის საერთო აღწერების რაოდენობა. უფრო ზოგადი წარმოდგენისათვის ვექტორის კომპონენტებად გამოყენებულ უნდა იქნას ნატურალური ან მთელი რიცხვები.

აღნიშნულ მეთოდში კარგად მუშაობს დოკუმენტებთან, რომლებიც შესაძლოა შეიცავდეს მოთხოვნის რამდენიმე ზედმეტ ტერმინს, თუ ეს ტერმინები ხვდებიან იშვიათად კოლექციაში, მაგრამ ხშირად დოკუმენტში. ვექტორულ სივრცის მოდელი შედგება შემდეგი მოსაზრებებისაგან:

- უფრო მეტად დოკუმენტთან მიახლოებული მოთხოვნის ვექტორი ზრდის იმის ალბათობას, რომ დოკუმენტი ეკუთვნის/შეესაბამება მომხმარებლის მოთხოვნას.
- საძიებო სიტყვა გამოიყენება დოკუმენტის სივრცის ზომის განსასაზღვრად (როდესაც მოხდება საძიებო სიტყვის მიახლოებულ მნიშვნელობის მოძიება, ითვლება რომ ეს სიტყვა არ არის რეალური, იდენტური)



Issues	• How to express NOT ? Proximity searches ? • Limited expressive power • Computationally intensive • Assumes that terms are independent.	• Estimation of needed probabilities • Prior knowledge needed. • Independence assumption • Boolean relations lost. • Which model is best ?
	• Lack of structure to represent important linguistic features • How to better visualize the retrieved set ?	

ინფორმაციის ძებნის ალბათური მოდელი

(Probabilistic Model)

ინფორმაცის ძებნის ალბათური მოდელი დაფუძნებულია მონაცემების ალბათური/რეიტინგული შეფასების პრინციპზე, რომელიც გულისხმობს, რომ ინფორმაციის საძიებო სისტემას უნდა შეეძლოს მოთხოვნის და დოკუმენტის საუკეთესო შესაბამისობის მოძიება. მეთოდში შესაძლოა იყოს განსხვავებული მტკიცებულების წყაროები, რომლებსაც იყენებს მოცემული მეთოდი. უმეტეს შემთხვევაში დოკუმენტში არსებობს როგორც რელევანტური ასევე არა რელევანტური ინფორმაციის სტატისტიკური განაწილება.

განვიხილოთ ინფორმაციის ძებნის სტატისტიკური მოდელის ძლიერი და სუსტი მხარეები:

უპირატესობები:

- მეთოდები აწვდიან მომხმარებლებს რელევანტური რეიტინგით მოძიებულ დოკუმენტს.
- მოთხოვნა შესაძლებელია იყოს მარტივად ფორმულირებადი, რადგან მომხმარებელს არ მოუწიოს მოთხოვნის ენის შესწავლა.
- მოთხოვნის კონცეფციის უჩვეულოდ დამახასიათებელ არჩევა შეიძლება იყოს წარმოდგენილი.

უარყოფითი მხარეები:

- არ ყოფნის სტრუქტურა, რათა მოძებნოს საჭირო აზრობრივად მიახლოებული ფრაზები.
- რელევანტური გამოთვლა შესაძლოა შეფასებული იყოს როგორც ძვირადღირებული(მოითხოვს მეტ რესურსს) გამოთვლა.

Statistical	**Vector Space**	**Probabilistic**
Motivation	Simplify query formulation Ability to control output	Address uncertainty in query representations
Goal	Rank the output based on Similarity	Probability of Relevance
Methods	Cosine measure	Use of different models
Source	**Query Term Statistics** Vector-Space: • similarity(Q,D) = Σ (w_{iq} x w_{ij}) / "normalizer" where w_{iq} = (0.5 + 0.5 $freq_{iq}$ / $maxfreq_q$) x **idf**(i) w_{ij} = $freq_{ij}$ x **idf**(i) • inverse term freq. in collection **idf**(i) = $\log_2$ (N–n(i)) / n(i). Probabilistic: • term weight = log [(r_t / R–r_t) / ((n_t –r_t) / ((N–n_t) – (R–r_t)))] ="(hits / misses) / (false alarms / correct misses)" • similarity $_{jk}$ = Σ (C + **idf**(i)) x **tf**(i,j) where **tf**(i,j) = K + (1–K) (freq(i,j) / maxfreq(j)).	
Issues	• How to express NOT ? Proximity searches ? • Limited expressive power • Computationally intensive • Assumes that terms are independent.	• Estimation of needed probabilities • Prior knowledge needed. • Independence assumption • Boolean relations lost. • Which model is best ?
	• Lack of structure to represent important linguistic features • How to better visualize the retrieved set ?	

ძებნა მონაცემთა ბაზებში - საძიებო სისტემის მოდელი

ვებ სივრცეში ინფორმაციის მოძიებასთან ერთად არანაკლებ მნიშვნელოვანია ძებნის მოდელების შემუშავება მონაცემთა ბაზისათვის, რომელიც საშუალებას მოგვცემს დიდი მონაცემთა ბაზებისათვის საჭირო (რელევანტური) ინფორმაციის მოძიებას, ნაკლები რესურსებით და ოპტიმალურ დროში.

დღეს გავრცელებულ ინფორმაციის ძებნის ტექნოლოგიები (როგორც ვებ სივრცეში ასევე მონაცემთა ბაზაში) ეფუძნება ზევით აღწერილ მეთოდებიდან ერთ-ერთ მეთოდს. ნაკლებად მოიძებნება ტექნოლოგიები, რომელიც კომბინირებულად მოახდენდა ამ ძებნის მოდელების გამოყენებას. მონაცემთა ბაზის მართვის სისტემაში ძებნის უზრუნველყოფის სერვისი ერთ-ერთ მნიშვნელოვანი პრინციპია, რომლებზეც დამოკიდებულია ბაზის ძირითადი ფუნქციონალი: მომხმარებლისათვის მოთხოვნისამებრ, საჭირო ინფორმაციის მიწოდება.

მონაცემთა ბაზისათვის საძიებო სისტემის შემუშავება ცალკე პროგრამის სახით თითქმის არ განიხილება. უმთავრესად ძებნის პროცედურა წარმოდგება, როგორც მონაცემთა ბაზის მართვის სისტემის დანართი. ამ ფუნქციონალური დანართის მუშაობის ეფექტურობა დამოკიდებულია ბაზის მოდელზე, ბაზის სტრუქტურაზე, სისტემურ რეალიზაციაზე და ასევე მის ფიზიკურ განლაგებაზე (განლაგებულია ერ სერვერზე, თუ დაცილებული წვდომის საცავებში).

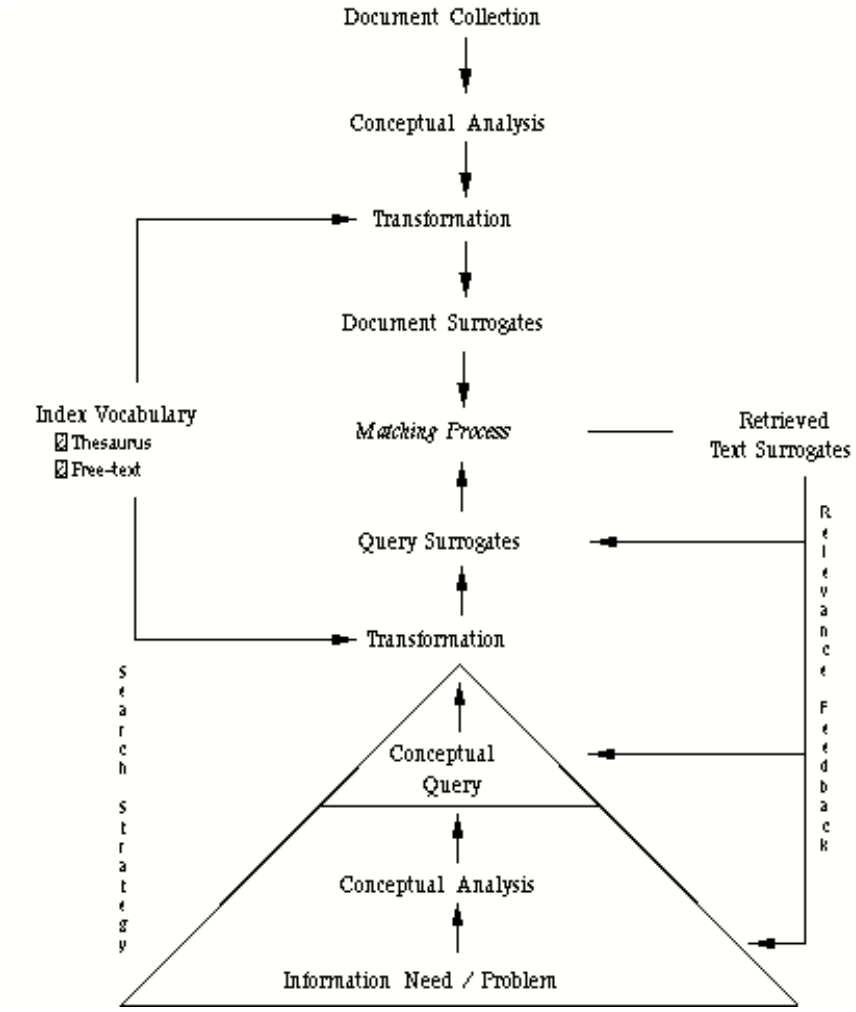
მონაცემთა ბაზების თეორია იცნობს იერარქიულ, ქსელურ, რელაციურ და ჰიბრიდულ მონაცემთა სტრუქტურებს, რომლებიც ხანგრძლივი დროის განმავლობაში ვითარდებოდა და კვლავაც ვითარდება. მრავალი მონაცემთა ბაზების მართვის სისტემა ისტორიას ჩაბარდა, ბევრი ახალიც გამოჩნდა, რომელთა შორის ლიდერებიცაა, კერძოდ Oracle, MsSQL Server, MySQL, Intebase, №სცცესს, dBase, Visual_FoxPro, Paradox და სხვ. ყველა ეს «გადარჩენილი» თუ «ახალდაბადებული» ბაზების მართვის სისტემები რელაციურ მოდელზეა აგებული (ან სუბრელაციურია).

მონაცემთა ბაზა დისკრეტული დინამიკური სისტემაა, რომლის მდგომარეობები (რელაციები) მიეკუთვნება ფიქსირებულ მონაცემთა მოდელს. მონაცემთა მოდელი

შეიძლება განვმარტოთ, როგორც ალგებრული სისტემა, რომელიც შედგება ლექსიკონური სიმრავლეების (საგნობრივი სფეროს ამსახველი სიტყვების ერთობლიობა) და მათ შორის დამოკიდებულებებისგან, რომლებზეც შესაძლებელია განხორციელდეს სხვადასხვა სახის ოპერაცია. მონაცემთა რელაციური მოდელი ფორმალური მათემატიკური ობიექტია და მისი საშუალებით საგნობრივი სფეროს არა ფორმალიზებული თვისებების ასახვას მივყავართ რთულ სემანტიკურ პრობლემამდე. იგულისხმება სემანტიკის (შინაარსის) მათემატიკური მოდელირების სპეციფიკური პრობლემები. მათი სირთულე ძირითადად განისაზღვრება მონაცემთა ბაზის რეორგანიზაციის (განახლების) პროცედურების სირთულით, სტატიკური და დინამიკური შეზღუდვების (პრედიკატებს) სისწორის შემოწმებით.

შეზღუდვები გამოხატავს მონაცემთა ბაზის მდგომარეობის (რელაციები) ზოგად, აბსტრაქტულ თვისებებს ანუ მონაცემთა ბაზის სემანტიკას. შეზღუდვების გამოხატვის ყველაზე ბუნებრივი ხერხია გამონათქვამები I რიგის პრედიკატებს ენაზე, რომელიც განსაზღვრავს მონაცემთა ბაზის დასაშვებ მდგომარეობათა სიმრავლეს. ამ გამონათქვამების ერთობლიობა მოდელიდან ამოსარჩევი ელემენტის სახელებთან ერთად (ატრიბუტების დასახელება) ქმნის მონაცემთა ბაზის ე.წ. სტატიკურ სქემას.

მონაცემთა ბაზა, გადადის რა ერთი მდგომარეობიდან მეორეში (რელაციური ცვალებადობა), აღწერს მონაცემთა მოდელში გარკვეულ ტრაექტორიას. ყოველი მომდევნო მდგომარეობა შეიძლება დამოკიდებული იყოს მის წინა მდგომარეობაზე. ამ მდგომარეობებს შორის კავშირები აღიწერება მთლიანობის დინამიკური შეზღუდვებით (პრედიკატებით), რომლებიც ქმნის მონაცემთა ბაზის ე.წ. დინამიკურ.



ნახ. 1 საძიებო სისტემის მუშაობის მოდელი

თუ გავაანალიზებთ ინფორმაციის ძიების საყოველთაოდ მიღებულ სქემას, ადვილი იქნება მისი შეთანადება მონაცემთა ბაზაში და მის სამართავ სისტემაში მიმდინარე პროცესებთან და დამუშავების ეტაპებთან.

ამდენად თუ შევიმუშავებთ ინფორმაციის ძიების სინთეზირებულ მეთოდს, რომელიც გააერთიანებს ზემოთ აღწერილ სამივე მეთოდის ძლიერ მხარეებს, მას გადავიტანთ მონაცემთა ბაზის ტერმინებში და შევუსაბამებთ პროცედურებს, შესაძლებელი იქნება შევქმნათ მონაცემთა ბაზებში ინფორმაციის საძიებო სისტემის მოდელი.

მონაცემთა ბაზაში ინფორმაციის ძებნის „ჰიბრიდული“ მოდელი

ინფორმაციის ძებნის თანამედროვე ალგორითმები, რომლებიც ძირითადად გამოიყენება სხვადასხვა საძიებო სისტემებში ერთი მხრივ აპრობირებულია და ფართოდ გამოყენებადია, მაგრამ მეორეს მხრივ მისი გამოყენება მონაცემთა ბაზებისათვის სემანტიკურად რელევანტური ძებნის თვალსაზრისით თითქმის არ განიხილება. ძებნა სტრუქტურირებულ მონაცემთა ბაზაში, არააქტუალურად ითვლება, რადგან არსებობს ძებნის კლასიკური ალგორითმები და მათი სხვადასხვა მოდიფიკაციები, რომლებიც უფრო სწრაფია.

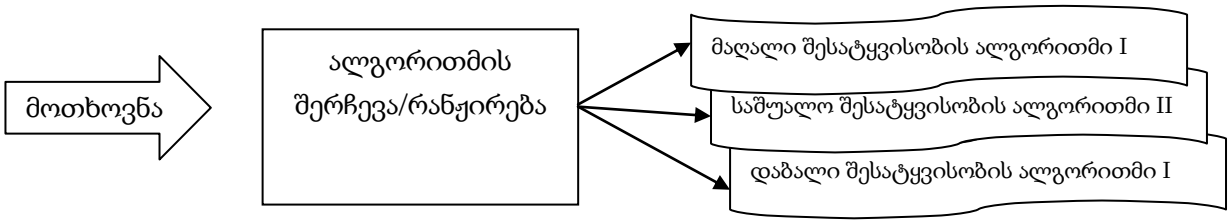
მონაცემთა ბაზებში ინფორმაცის ძებნის მოდულის შესამუშავებლად მოხდა არსებული თანამედროვე ძებნის ალგორითმების კომბინირება და მის საფუძველზე მივიღეთ ეგრეთწოდებული ინფორმაციის ძებნის „ჰიბრიდული“ მოდელი.

ძებნის ალგორითმების ანალიზის საფუძველზე ირკვევა, რომ სხვადასხვა ტიპის მონაცემების ძებნისათვის (სკალარული, ვექტორული, სიმბოლური, ტექსტური, ლოგიკური და ა.შ) სხვადასხვა ალგორითმს სხვადასხვა შედეგები აქვს. ჩვენ აქცენტს ვაკეთებთ სემანტიკურ ძებნაზე ამიტომ ალგორითმების პრიორიტეტულობა ლაგდება სემანტიკურად ზუსტი შედეგის მიხედვით (დროის ფაქტორი ამ ეტაპზე არ განიხილება).

ჩვენს მიერ შექმნილი მოდელი წარმოადგენს ზემოთ აღწერილი სამივე მოდელის ეტაპობრივ გამოყენებას. მოდელში ძებნის პროცესი განიხილება, როგორც სამ დონიანი პროცესი, რომელსაც წინ უძღვის მოთხოვნის ტიპის იდენტიფიცირების ეტაპი და შედეგების რანჟირების ეტაპი.

საწყისი ეტაპი - მოთხოვნის ტიპის იდენტიფიცირება

ამ ეტაპზე ხდება მოთხოვნის ტიპის იდენტიფიცირება. ეს გულისხმობს - საწყისს ეტაპზე იმისდა მიხედვით თუ მომხმარებელი რა ტიპის (სკალარული, ვექტორული, სიმბოლური, ტექსტური, ლოგიკური და ა.შ) მონაცემის მოძიებას სურს, სისტემა ირჩევს მონაცემს ტიპის მიხედვით სამი წარმოდგენილი ალგორითმიდან ყველაზე მისადაგებულ ალგორითმს.



ნახ.3. საწყისი ეტაპი

ძებნის პირველი ეტაპი

საწყის ეტაპზე შერჩეული ალგორითმით ხორციელდება ძებნა მონაცემთა ბაზაში. გამოიყოფა ცალკე მოძიებული ინფორმაცია.

ძებნის მეორე ეტაპი

საწყის ეტაპზე რანჟირების საფუძველზე მეორე ალგორითმით ხორციელდება ძებნა მონაცემთა ბაზაში. გამოიყოფა ცალკე მოძიებული ინფორმაცია.

ძებნის მესამე ეტაპი

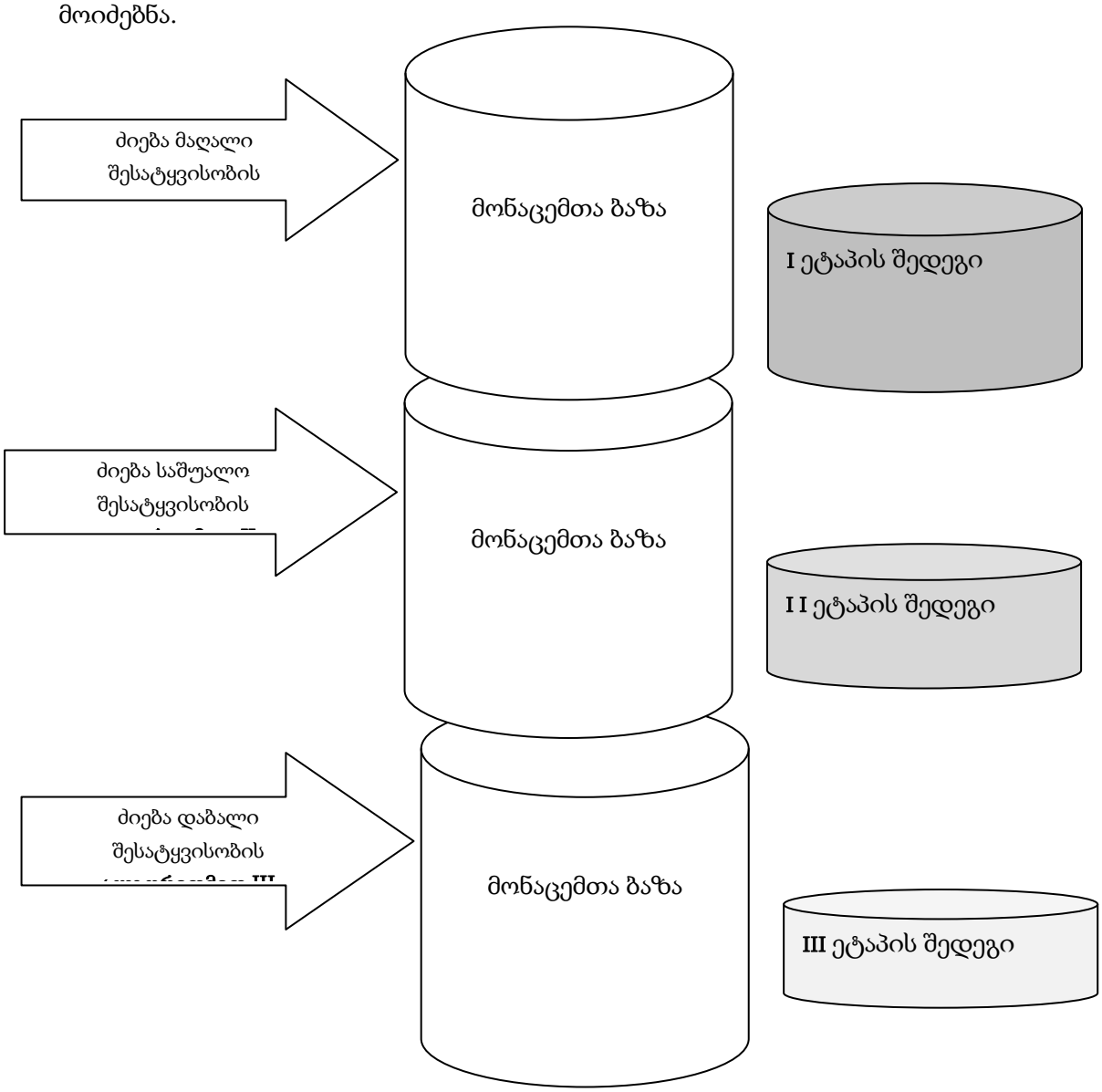
დარჩენილი ალგორითმით ხორციელდება ძებნა მონაცემთა ბაზაში. გამოიყოფა ცალკე მოძიებული ინფორმაცია.

შედეგების რანჟირების ეტაპი

ძებნის სამი დონის გავლის შედეგად მიიღება მონაცემთა ბაზიდან ამოკრეფილი სამი ქვესიმრავლე, რომლებსაც ლოგიკურად შესაძლებელია ჰქონდეთ თანა კვეთა. ამ შედეგთა გაერთიანებით მიღებული მონაცემები რანჟირება ხდება შემდეგი პრინციპით:

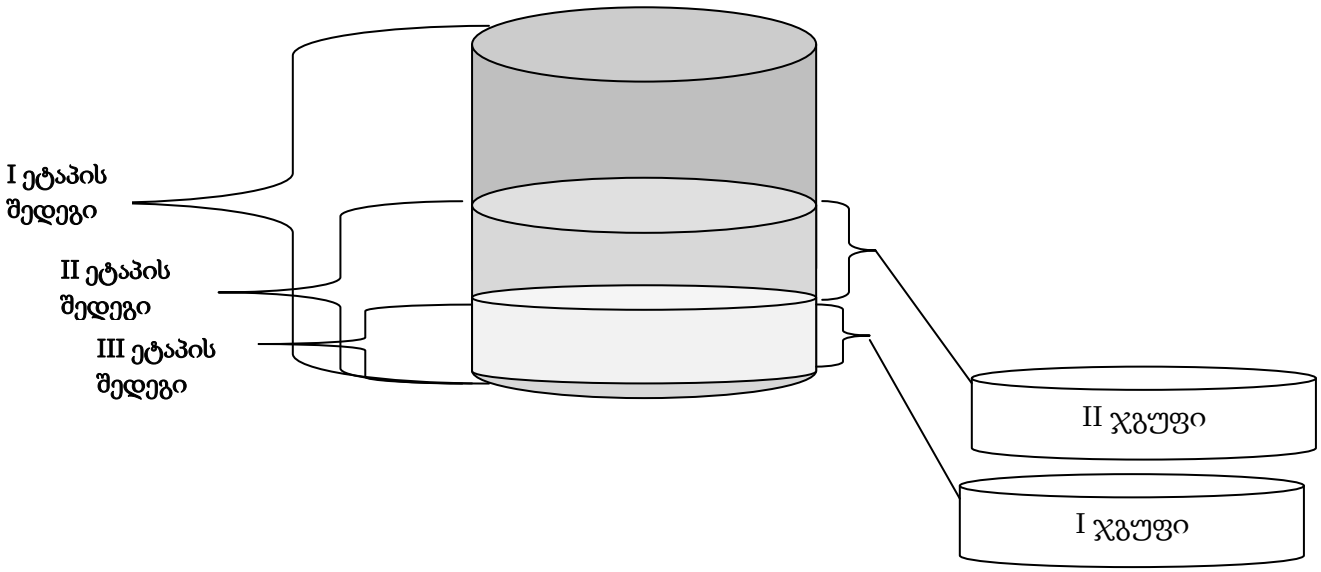
- პირველი ჯგუფი - მონაცემები რომელიც სამივე ალგორითმით მოიძებნა.
- მეორე ჯგუფი - მონაცემები რომელიც პირველი და მეორე ალგორითმით მოიძებნა.

- მესამე ჯგუფი - მონაცემები რომელიც პირველი და მესამე ალგორითმით მოიძებნა.



- მეოთხე ჯგუფი - მონაცემები რომლებიც დარჩა წინა ამორჩევების შემდეგ პირველ ჯგუფში
- მეხუთე ჯგუფი - მონაცემები რომელიც მეორე და მესამე ალგორითმით მოიძებნა.
- მეექვსე ჯგუფი - მონაცემები რომლებიც დარჩა წინა ამორჩევების შემდეგ მეორე ჯგუფში

- მეშვიდე ჯგუფი - მონაცემები რომლებიც დარჩა წინა ამორჩევების შემდეგ მესამე ჯგუფში.



ჩვენს მიერ შემუშავებული მეთოდი იმის გამო, რომ მოითხოვს ბაზაზე სამჯერ ერთი და იმავე ინფორმაციის მოძიებას, ძებნის სისწრაფის თვალსაზრისით არაეფექტურია, მაგრამ მოთხოვნაზე იძლევა პასუხს მაღალი სიზუსტით, ასევე მომხმარებელს საშუალება ეძლევა მიიღოს პრიორიტეტულობის მიხედვით იერარქიულად დალაგებული სხვა ალტერნატიული პასუხებიც.

შემუშავებული მოდელის რეალიზაციისათვის თუ ჩვენ გამოვიყენებთ ისეთ პროგრამულ ინსტრუმენტს, რომელიც ნებისმიერ მონაცემთა ბაზის მართვის სისტემასთან იქნება თავსებადი, მიღებული პროდუქტი იქნება უნივერსალური და მოსახერხებელი ყველა მომხმარებლისათვის.

შექმნილი მოდელის საფუძველზე შექმნილია მცირე სატესტო პროგრამა, რომელიც მოითხოვს დახვეწას.

დასკვნა

სამაგისტრო ნაშრომის ფარგლებში ჩატარებული სამუშაოების შედეგად გამოიკვეთა, რომ არ არსებობს ინფორმაციის ძებნის უნივერსალური ალგორითმი, რომელიც ერთნაირად ეფექტური იქნება ვებ- სივრცეში, სხვადასხვა მონაცემთა საცავებში და ასევე მონაცემთა ბაზებში. გარდა ამისა მონაცემთა ბაზებში ინფორმაციის ძებნისათვის ძირითადად გამოიყენება სტანდარტული გადარჩევის ალგორითმები, რომლებიც უზრუნველყოფენ ძებნას ველების მიხედვით.

შემუშავებული მოდელის დახვეწის და მისი პროგრამული რეალიზაციის შედეგად შესაძლებელი იქნება შეიქმნას ისეთი პროგრამულ ინსტრუმენტი, რომელიც ნებისმიერ მონაცემთა ბაზის მართვის სისტემასთან იქნება თავსებადი, მიღებული პროდუქტი იქნება უნივერსალური და მოსახერხებელი ყველა მომხმარებლისათვის.

გამოყენებული ლიტერატურა

1. Manning D., Raghavan P., Schütze H. (2008), " Boolean Retrieval", in Introduction to Information Retrieval,Cambridge University Press, pp. 1-18.
2. Manning D., Raghavan P., Schütze H. (2008), "Scoring, term weighting and thevector space model", in Introduction to Information Retrieval,Cambridge University Press, pp. 109-133.
3. Jun Zhai, Yiduo Liang, Yi Yu and Jiatao Jiang (2008), *"Semantic Information Retrieval Based on Fuzzy Ontology for Electronic Commerce"*, School of Management, Dalian Maritime University, Dalian 116026, P. R. China, Journal of Software, Vol. 3., N9, pp. 20-29.
4. Margulis, E. (1993). Modelling documents with multiple poisson distributions. Information Processing and Management 29, 215{227
5. Fuhr, N. (1989). Models for retrieval with probabilistic indexing. Information processing and management 25(1), 55{72.
6. Sparck - Jones, K., S. Walker, and S. Robertson (2000). A probabilistic model of information retrieval: Development and comparative experiments (part 1 and 2). Information Processing & Management 36(6), 779{840.