მულტიმედიურ მონაცემთა ბაზაში რელატიური დოკუმენტის იდენტიფიკაცია

ლილი პეტრიაშვილი, მაია ოხანაშვილი, ნინო აბრამიშვილი, გიორგი ბასილაძე
საქართველოს ტექნიკური უნივერსიტეტი

რეზიუმე

განიხილება მულტიმედიალური მონაცემთა ბაზების გამოყენების საკითხები კორპორაციული მართვის სისტემებში. აქ ინტერნეტ სივრცეში გავრცელებული ჰეტეროგენული მონაცემები: ტექსტური, გრაფიკული, აუდიო-ვიდეო და ანალოგური ტიპის ფაილების სახით, სხვადასხვა პროგრამული პაკეტების დახმარებით გარდაიქმნება ციფრული ფორმატის ინფორმაციად. წარმოდგენილია კორპორაციული მართვის სისტემებში მულტიმედიალური მონაცემთა ბაზების გამოყენების ამოცანა, სადაც რელევანტური დოკუმენტების წვდომის და იდენტიფიკაციის საკითხი ერთ-ერთი მნიშვნელოვანი პროცესია. შემოთავაზებულია საინფორმაციო-რელატიური სისტემა, რომელიც უზრუნველყოფს მულტიმედიურ მონაცემთა ბაზაში განთავსებულ ბიბლიოთეკის ლექსიკონში ტექსტის ინდექსირების პროცესს, გასაღებური სიტყვის დახმარებით.

საკვანძო სიტყვები: მულტიმედიალური სისტემა. მულტიმედიალური მონაცემთა ბაზა. რელატიური დოკუმენტი. იდენტიფიკაცია. ჰეტეროგენული ინფორმაცია. მედია ობიექტები.

1. შესავალი

თანამედროვე ადამიანი ვირტუალური სამყაროს განუყოფელი ნაწილია. პრაქტიკულად საქმიანობის ყველა სფეროში მეცნიერება, კულტურა, ბიზნესი, განათლება, მულტიმედიური ტექნოლოგია მნიშვნელოვან როლს თამაშობს. მულტიმედიური ტექნოლოგიის გამოყენებით შესაძლებელია სხვადასხვა სახის და ფორმატის ინფორმაცია მომხმარებელს მიეწოდოს, როგორც ერთი სახის პროდუქტი, სადაც სინთეზურად ურთიერთდაკავშირებულია ტექსტური, გრაფიკული, ვიდეო, აუდიო და სხვადასხვა ვიზუალური ეფექტებით სრულყოფილი ინფორმაცია.

მულტიმედიური ტექნოლოგია – მომხმარებელს აძლევს საშუალებას მიიღოს სხვადასხვა ფორმატის (ტექსტური, გრაფიკული, ანიმაციური, ვიდეო-აუდიო) ინფორმაცია ინტერაქტიულ რეჟიმში.

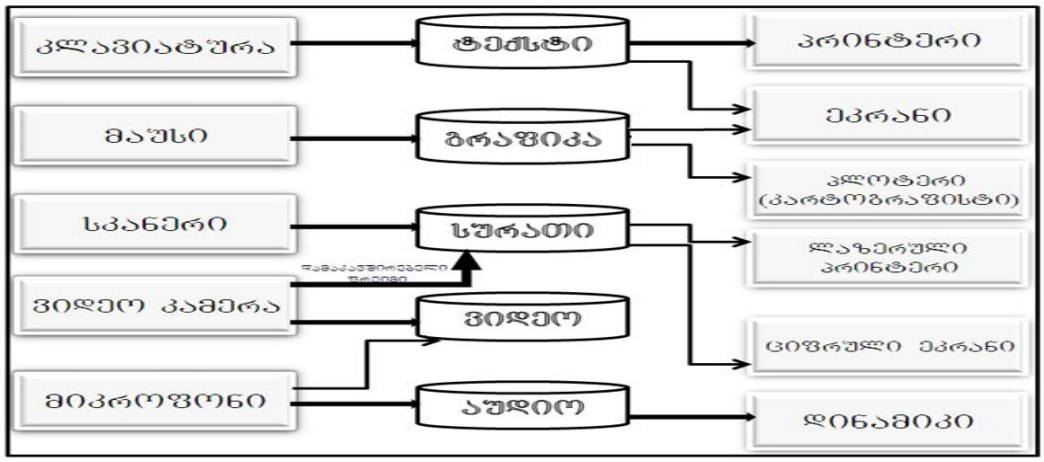
მონაცემთა შეკრება პირველადი წყაროებიდან, მათი გაფილტვრა (მეთოდურად, სემანტიკურად და ტექნიკურად), ტრანსფორმაცია და კონვერტაცია (მონაცემთა წინასწარ განსაზღვრული სტრუქტურების მისაღებად), მეტაინფორმაციის იერარქიულიად ორგანიზება (მონაცემთა კატალოგებისა და არქივების სამართავად), უმნიშვნელოვანეს ამოცანათა კლასს შეადგენს თანამედროვე მართვის ინტეგრირებული სისტემებისათვის [2]. 1-ელ ნახაზზე მოცემულია ინტეგრირებული მულტიმედიური მონაცემთა ბაზა.

ნახ.1

თანამედროვე საინფორმაციო ობიექტების ეფექტურად მართვისათვის აუცილებელია მძლავრი მულტიმედიალური მონაცემთა ბაზის შექმნა, რომელიც განთავსდება ინტერნეტში, კორპორატიულ ინტრანეტში და კლიენტ/სერვერ სისტემაში. ინტერნეტ სივრცეში გავრცელებული ჰეტეროგენული მონაცემები (ტექსტური, ვიდეო, გრაფიკული, აუდიო და ანალოგური) სხვადასხვა პროგრამული პაკეტების დახმარებით გადაიქცევა ციფრული ფორმატის ინფორმაციად და განთავსდება მულტიმედიურ მონაცემთა ბაზებში [1,2].

ინტეგრირებულ მულტიმედიურ მონაცემთა ბაზების ეფექტური გამოყენებით და ოპტიმალური მართვის შედეგად ბიზნესისათვის იხსნება ახალი ბაზარი, სადაც დამატებითი ხარჯები მინიმუმამდეა დაყვანილი და მკაცრი კონკურენციის პირობებში შესაძლებელია კონკურენტუნარიანი მდგომარეობის შენარჩუნება.

როგორც ზემოთ აღვნიშნეთ, ჰეტეროგენული ინფორმაცია მუშავდება და ინფორმაციის მიმღებ მოწყობილობებში განთავსდება მედია ობიექტების სახით. მედია ობიექტებისათვის არის ეგრეთწოდებული მმართველი ანუ არქივირების სისტემები, რომელიც იყენებს მონაცემთა ბაზებს. ამ პროცესში მნიშვნელოვანია მონაცემთა წვდომის ოპერაცია, რადგან ერთი და იგივე მონაცემი ხშირ შემთხვევაში წარმოდგენილია სხვადასხვა სინტაქსით. მედია ობიექტებს ზოგადად აქვს მე-2 ნახაზზე მოცემული სახე:



ნახ.2. მედია ობიექტები, ინფორმაციის მიმღები და გამცემი მოწყობილობები

მედია ობიეტებს შორის ერთ-ერთი ყველაზე მეტად გავრცელებული სახეა ტექსტური დოკუმენტი, რომელიც ინტერნეტ სივრცეში შესაძლოა მოხვდეს, როგორც კლავიატურიდან შეტანილი სიმბოლოებით, ასევე სკანერის საშუალებით და ციფრული გამოსახულებით.

ყოველი დოკუმენტი შეიცავს ტერმინთა ჩამონათვალს (index terms), ხოლო ყოველი ტერმინის ინვერსიული ინდექსი ასოცირდება დოკუმენტების ჩამონათვალთან, სადაც ხდება მოცემული ტერმინის იდენტიფიცირება. ანუ ვლებულობთ ინვერსიულ შესაბამისობას, ყოველ ტერმინს შეესაბამება დოკუმენტების ჩამონათვალი.

არის შემთხვევები, როდესაც ერთი და იგივე სხვადასხვა მნიშვნელობის ტერმინი (ომონიმები) და ერთი და იგივე მნიშვნელობის სხვა სახით ჩაწერილი ტერმინი (სინონიმები) ერთსა და იმავე დოკუმენტში რამდენჯერმე გვხვდება, რაც ძებნის პროცესში იწვევს სირთულეებს. ამოცანა მდგომარეობს ტექსტში გამოყენებული ერთი და იგივე სიტყვის ან ტერმინის სხვადასხვა მნიშვნელობით წარმოდგენის შემთხვევაში, როგორ განვსაზღვროთ სწორი პასუხი.

შემოვიტანოთ აღნიშვნა tf_{ij}, სადაც t - ტექსტში მოცემული ტერმინია, f - *Frequency* (ფიქსირებული ასახვა), ხოლო ij აღნიშნავს, თუ რა სიხშირით მოხვდება i დოკუმენტში

j დესკრიპტორი. ახეთივე სახით წარმოვადგინოთ *Frequency* დოკუმენტი df_j(დოკუმენტების რაოდენობა, სადაც დესკრიპტორი j მეორდება).

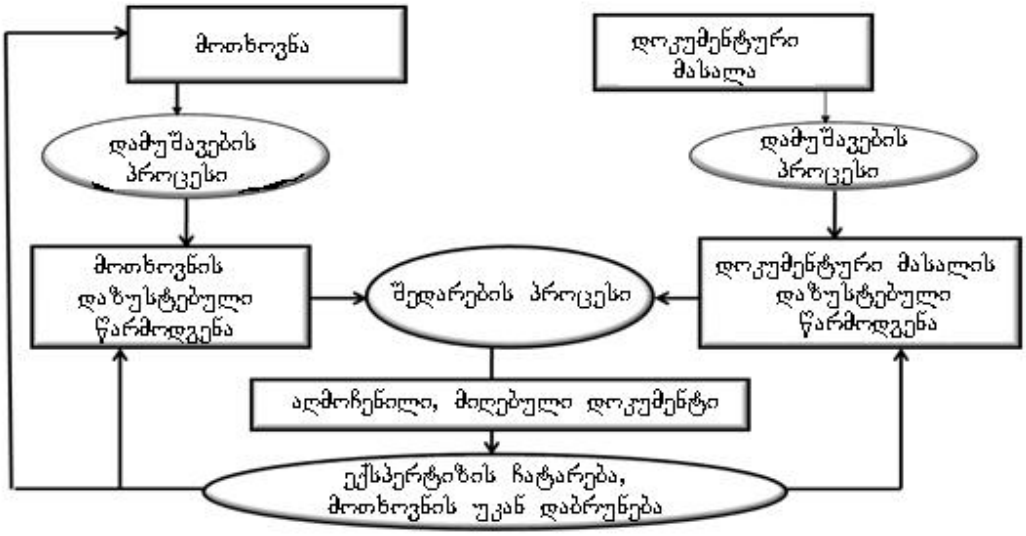
i დოკუმენტისათვის j დესკრიპტორის დამოკიდებულება გამოისახება ფორმულით [1]:

$$W_{ij} = tf_{ij} \times \log(N/df_i) \quad (N - \text{დოკუმენტების რაოდენობა})$$

აქედან გამომდინარეობს, რომ გამოყენებული ტერმინის სიხშირე პირდაპირ პროპორციულია ინვერსირებული დოკუმენტის სიხშირის.

თუ ტექსტში გამოყენებულ ერთსა და იმავე სიტყვას დავალაგებთ გამოყენების სიხშირის მიხედვით, მაშინ ამ სიტყვის გამოყენების სიხშირე პირდაპირ პროპორციულია მისი რიგითი ნომრისა.

განვიხილოთ შემთხვევა, როდესაც $df_j = N$ ანუ ყოველ დოკუმენტში ერთი დესკრიპტორი ხვდება, მაშინ მისი განმეორების მნიშვნელობა ნულს უტოლდება. მე-3 ნახაზზე წარმოდგენილია რეტრივიალური პროცესი, ანუ როდესაც არ ხდება სასურველი შედეგის მიღება.



ნახ.3. ინფორმაციის რეტრივიალური პროცესი

წარმოდგენილ ლოგიკურ რეტრივიალურ მოდელს აქვს დიდი უპირატესობა, მას ფართოდ იყენებს კომერციული სისტემები. დაყენებული მოთხოვნის ფორმულირება არის საკმაოდ რთული და აქვს შედეგისათვის გადამწყვეტი მნიშვნელობა[3].

ცნობილი რეტრივიალური ანუ განმეორებითი მოდელი არის ვექტორულ-რეტრივიალური-მოდელი [1]. დოკუმენტში დესკრიპტორთა რაოდენობა განისაზღვრება ფორმულით

$$D_i = (T_{i1}, T_{i2}, \ldots, T_{ik}, \ldots, T_{iN}),$$

სადაც T_{ik} არის k დესკრიპტორის მოცულობა i დოკუმენტში. N არის ყველა დესკრიპტორთა რიგითი ნომერი. წარმოდგენილი მოთხოვნა ანალოგიურია ჩანაწერისა:

$$Q_j = (Q_{j1}, Q_{j2}, \ldots, Q_{jk}, \ldots, Q_{jN}),$$

სადაც Q_{jk} არის k დესკრიპტორის მოცულობა j მოთხოვნაში. ამ შემთხვევაში მოცულობა არის ბინარული სიდიდე (0 ან 1), ისევე როგორც ზემოთ იყო გამოთვლილი W_{ij} სიდიდისათვის, ამიტომ D_i და Q_j არის მსგავსი სიდიდეები, და გამოისახება შემდეგი ფორმულით:

$$S(D_i, Q_j) = \sum_{k=1}^{N} (T_{ik} \times Q_{jk})$$

მოთხოვნათა მოდიფიცირების შემთხვევაში დესკრიპტორები ასოცირდება დოკუმენტებთან, და ანალოგიურად დოკუმენტები მოდიან დესკრიპტორებთან შესაბამისობაში. ეს პროცესი გამოისახება შემდეგი ფორმულით:

$$Q^{(i+1)} = Q^{(i)} + \alpha * \sum_{D_i \in Rel} D_i - \beta * \sum_{D_i \in NonRel} D_i \ ,$$

სადაც $Q -$ არის საწყისი მოთხოვნა, ხოლო $Q^{(i+1)}$ არის ახალი მოთხოვნა, რომელიც დადგება წინა მოთხოვნაზე მიღებული შედეგის განსაზღვრის შემდეგ. α და β ასახავს დაყენებული მოთხოვნის შესრულების ხარისხს, თუ რამდენად ზუსტადაა განსაზღვრული მიღებული შედეგი.

ზოგადად, დოკუმენტის მოდიფიკაციის განსაზღვრისათვის დაცული უნდა იყოს შემდეგი წესები [1]:

- მოთხოვნათა რელატიური დესკრიპტორი, რომელიც წარმოდგენილ დოკუმენტში არ კლასიფიცირდებოდა, ახდენს დოკუმენტის ინიციალიზაციას;
- მოთხოვნათა რელატიური დესკრიპტორი, რომელიც წარმოდგენილ დოკუმენტში კლასიფიცირდება, ამ დოკუმენტს ანიჭებს განსაზღვრულ მდგომარეობას და შემდგომში ხდება დოკუმენტის წარდგენა მოთხოვნის შესაბამისად;
- დესკრიპტორები, რომლებიც მოთხოვნაში არ ხვდება, ამცირებს დოკუმენტის მნიშვნელობას და ფაქტიურად არ ხდება დოკუმენტის მოძებნა.

ასეთი სახის მოდიფიკაცია გამართლებულია იმ შემთხვევაში, როდესაც მოთხოვნა შეესაბამება წარმოდგენილ მოდელს. ბოლოს განხილულ წესს უწოდებენ რეტრივიალურ მოდელს და იგი უნდა აკმაყოფილებდეს ოთხ პარამეტრს:

1. $P(rel)$ - ალბათობა იმისა, რომ დოკუმენტი რელატიურია;
2. $P(nonrel)$ - ალბათობა იმისა, რომ დოკუმენტი რელატიური არ არის;
3. a1 დანახარჯი, რომელიც შეესაბამება უკან დაბრუნებულ არარელატიურ დოკუმენტს;
4. a2 დანახარჯი, რომელიც შეესაბამება მოძიებულ რელატიურ დოკუმენტს.

ამ პარამეტრთა გათვალისწინებით შეიძლება ჩავწეროთ: $a2 \times P(rel) \geq a1 \times P(nonrel)$, საიდანაც ჩანს, რომ როდესაც დოკუმენტი რალატიურია მისი ღირებულება უფრო დაბალია, ვიდრე არარელატიური დოკუმენტის შემთხვევაში.

3. დასკვნა

კორპორაციული მართვის სისტემებში მულტიმედიური მონაცემთა ბაზების, გამოყენებით მომხმარებელს საშუალება ეძლევა ინტერაქტიულ რეჟიმში მიიღოს ინტერნეტ სივრცეში გავრცელებული ჰეტეროგენული მონაცემები. შედგენილია საინფორმაციო-რელატიური-სისტემა, სადაც ხდება მონაცემთა ინდექსირება. შეფასებულია დოკუმენტის რელატიურობის ხარისხი. წარმოდგენილია მათემატიკური მოდელი, რომლის გამოყენებით შესაძლებელია რელევანტური დოკუმენტების წვდომის სიზუსტის განსაზღვრა, დოკუმენტში გამოყენებული ტერმინის სიხშირე პირდაპირ პროპორციულია ინვერსირებული დოკუმენტის სიხშირის.

ლიტერატურა:

1. Meyer-Wegener K. Multimediale Datenbanken. Germany - 2003
2. სურგულაძე გ., პეტრიაშვილი ლ. მონაცემთა საცავების აგების ტექნოლოგია ინტერნეტული ბიზნესის სისტემებისათვის. სტუ. თბ., 2005
3. Гогичаишвили Г.Г., Сургуладзе Г.Г. Разработка прикладного программного обеспечения интегрированных информационных систем управления на основе UML. Georgian Electronic Scientific Journal, 2002, N1.

IDENTIFYING RETRIEVAL OF DOCUMENTS IN A MULTIMEDIA DATABASE

Petriashvili Lili, Okhanashvili Maia, Abramishvili Nino, Basiladze Giorgi
Georgian Technical University

Summary

The application of multimedia database in corporate governance systems is analyzed. The heterogenic data like text, graphics; audio - video and analogue type of files spread in Internet will be converted into the digital format through various software packages.The multimedia documents are presented where the entries are identified pursuant to the quairies. A mathematical model that supports and simplifies this process is presented.

ИДЕНТИФИКАЦИЯ ПОИСКОВЫХ ДОКУМЕНТОВ В МУЛЬТИМЕДИЙНОЙ БАЗЕ ДАННЫХ

Петриашвили Л., Оханашвили М., Абрамишвили Н., Басиладзе Г.
Грузинский Технический Университет

Резюме

Рассмотрены вопросы использования мультимедийных баз данных в системах корпоративного управления. Распространяемые ими в интернете гетерогенные данные: тексты, графики, аудио-видео и аналоговые типы файлов, с использованием различных программных пакетов будут преобразованы в цифровой формат. В статье представлены мультимедийные документы, где в соответствии с запросами происходит идентификация записи. Предлагается математическая модель, которая поддерживает эти процессы и легко решает указанную проблему.