AI-ს კვლევა

ნახევრად სტრუქტურირებული დოკუმენტების ანალიზი

ინტერნეტი მოიცავს უზარმაზარი რაოდენობის ნახევრად სტრუქტურირებულ დოკუმენტებს, რომლებიდანაც შესაძლებელია მნიშვნელოვანი ინფორმაციის ამოღება - მოპოვება.

როგორც წესი, ასეთ ნახევრად სტრუქტურირებული დოკუმენტების პოვნა შესაძლებელია ინტერნეტის მსოფლიო ქსელში (WWW) ჰიპერტექსტური მარქაფის ენის (HTML) ტიპის დოკუმენტის სახით.

ნახევრად სტრუქტურირებული დოკუმენტების მენეჯმენტი კვლევის შედარებით ახალი სფეროა. ჯერ კიდევ 1998 წლის ზაფხულში, 16 წამყვანმა გლობალურმა ექსპერტმა მონაცემთა მენეჯმენტის სფეროში გამოაქვეყნა მანიფესტი, რომელშიც მოცემულია კვლევების სამი ძირითადი მიმართულება მომდევნო ათი წლის განმავლობაში, მათ შორისაა ნახევრად სტრუქტურირებული დოკუმენტების მენეჯმენტი.

დღესდღეობით ნახევრად სტრუქტურირებული დოკუმენტების მართვის არსებული მეთოდები სრულად არ წყვეტს ყველა გადაუდებელ პრობლემას და საჭიროებს არსებით გაუმჯობესებას. გარდა ამისა, უმეტეს ტრადიციულ DBMS- ში განხორციელებული მრავალი მექანიზმი, როგორიცაა ტრიგერები და ტრანზაქციები, მთლიანად არ არის შესწავლილი ნახევრად სტრუქტურირებული დოკუმენტების კონტექსტში. ამჟამად არსებული ნახევრად სტრუქტურირებული მონაცემთა მართვის სისტემის სიჩქარე არ არის დამაკმაყოფილებელი, ამიტომ ერთ-ერთი მთავარი ამოცანაა მათი მუშაობის ეფექტურობის გაზრდა.

Wrapper-ი არის პროგრამა, რომელიც რელაციურ ფორმაში წარდმოადგენს ინფორმაციას. იდეალურ შემთხვევაში Wrapper-ებმა მაღალი

სიზუსტით, მინიმალური ადამიანური მეთვალყურეობის პირობებში, დიდი რაოდენობით უნდა ამოიღონ ინფორმაცია ნახევრად სტრუქტურირებული დოკუმენტებიდან.

ამ დროისთვის არსებობს ამ პროცესის განხორციელების მხოლოდ რამდენიმე შემთხვევა, - Gogar et al. იყენებს ე.წ. ღრმა სწავლების (Deep learning) არქიტექტურას, ხოლო Wiedmann et al. იყენებენ მანქანური სწავლების კლასიფიკატორების მოვლენების ექსტრაქციისთვის.

ჩვენ წარმოვადგენთ რელატიური ინფორმაციის ნახევრად სტრუქტურირებული დოკუმენტებიდან ამოღების ახლებურ გზას, რომელიც იყენებს ახალი ღრმა სწავლების არქიტექტურას, ყურადღების მექანიზმზე დაფუძნებულ კონვოლუციური ქსელებისა და რეზიდუალური კავშირების, ასევე რეგულირებული ყურადღებაზე დაფუძნებული გრძელი და მოკლე მეხსიერებების ქსელისა და გრაფული ყურადღების ქსელების კომბინაციას.

კვლევის გეგმა

კვლევის ხანგრძლივობა: 10-11 თვე

კვლევის მეთოდი: თვისებრივი

კვლევა გაწერილი არის 7 ეტაპად.

I ეტაპი - *კვლევის პრობლემის ჩამოყალიბება.* (2-3 კვირა)

ნახევრად სტრუქტურირებული დოკუმენტების ანალიზთან დაკავშირებული პრობლემების და სირთულეების შესწავლა.

II ეტაპი - *არსებული ლიტერატურის შესწავლა - ანალიზი.* (3 თვე)

ნახევრად სტრუქტურირებული დოკუმენტების ანალიზის პრობლემების გადაჭრასთან, გადაჭრის ზების ხარვეზულობასთან, ზოგადი კონცეფციის

არქიტექტურასთან დაკავშირებული აკადემიური- სამეცნიერო ნაშრომების მოძიება და გაანალიზება.

III ეტაპი - *პრობლემის გადაჭრის გზების ჩამოყალიბება* (2-3 კვირა)

მოძიებული ინფორმაციის გაანალიზების შედეგად, მოცემული პრობლემის პოტენციურად გადაჭრის გზების ჩამოყალიბება.

IV ეტაპი - *ექსპერიმენტის ჩატარება.* (3 თვე)

წინა ეტაპში შერჩეული პრობლემის გადაჭრის გზების მორგება მოცემული ამოცანებისთვის. ექსპერიმენტის მსვლელობისას უნდა მოხდეს შედარება დროის და სიზუსტის განსაზღვრით.

V ეტაპი - *აპლიკაციის შექმნა.* (1-1,5 თვე)

შედეგებზე დაყრდნობით ამოცანაზე მორგებული აპლიკაციის შექმნა (ტექსტობრივი აღწერა და კოდი).

VI ეტაპი *-აპლიკაციის ტესტირება* (1-2 კვირა)

გადაჭრის გზების გამოცდა რეალურ სივრცეში.

VII ეტაპი - *სამეცნიერო-აკადემიური პუბლიკაციის გამოქვეყნება*