ՏԵՔՍՏԱՅԻՆ ՓԱՍՏԱԹՂԹԵՐՈՒՄ ԻՆՏԵԼԵԿՏՈՒԱԼ ՈՐՈՆՄԱՆ ԵՎ ՀԱՄԱՊԱՏԱՍԽԱՆՈՒԹՅԱՆ ԱՍՏԻՃԱՆԻ ԳՆԱՀԱՏՄԱՆ ԱԼԳՈՐԻԹՄԻ ՄՇԱԿՈՒՄ

ՊԵՏՐՈՍՅԱՆ ԳԵՎՈՐԳ
Վանաձորի պետական համալսարանի
բնական գիտությունների ֆակուլտետի «Ինֆորմատիկա և կիրառական մաթեմատիկա» բաժնի 4-րդ կուրսի ուսանող
էլփոստ՝ g-a-petrosyan@mail.ru
ՍԱՀԱԿՅԱՆ ՌՈՒՍՏԱՄ
Տեխնիկական գիտությունների դոկտոր, պրոֆեսոր,
Վանաձորի պետական համալսարանի ռեկտոր
էլփոստ՝ rsahakyan@yahoo.com

Ժամանակակից հասարակությունում թվային տեղեկատվության ծավալների ավելացմանը զուգընթաց տեքստային փաստաթղթերում ցանկալի նյութը որոնելիս հաճախ են բարդություններ առաջանում, քանի որ տեքստային խմբագրիչներում ներկառուցված որոնիչները կատարում են միայն հստակ արժեքների որոնում: Սակայն գործնականում առավել կիրառական նշանակություն ունի ինֆորմացիայի ոչ հստակ որոնումը: Սույն հոդվածը նվիրված է տեքստային փաստաթղթերում ինտելեկտուալ որոնման գործընթացը ավելի արդյունավետ դարձնելու հիմնախնդրին: Աշխատանքի նպատակն է մշակել ալգորիթմ, որը տեքստային փաստաթղթում կիրականացնի ինտելեկտուալ որոնում և կգնահատի հարցման համապատասխանության աստիճանը: Այդ նպատակին հասնելու համար լուծվել են հետևյալ խնդիրները՝ տեղեկատվության սպացման համակարգերի հարցումների ուսումնասիրություն, տեքստային ինֆորմացիայի ոչ հստակ որոնման՝ շինգլների մեթոդի ուսումնասիրություն, տեքստային տվյալների համապատասխանության գնահատման ալգորիթմի մշակում: Կիրառվել են հետևյալ գիտական մեթոդները՝ անալիզ, համեմատում, համադրում: Իրականացված ուսումնասիրության արդյունքում տեքստային փաստաթղթերում ոչ հստակ որոնման հիմնախնդիրը վերլուծված է, իսկ ինտելեկտուալ որոնման և համապատասխանության աստիճանի գնահատման ալգորիթմը՝ առաջարկված:

Բանալի բառեր՝ *տեքստային փաստաթուղթ, ինդեքսավորում, շինգլների մեթոդ, տվյալների բազա, անորոշ բազմություն, Լևենշտեյնի հեռավորություն, Մամդանի մեթոդ:*

Ներածություն

Ինֆորմացիայի երկարաժամկետ պահպանման և մշակման խնդիրը առաջացել է առաջին համակարգիչների ստեղծումից անմիջապես հետո: 20-րդ դարի 60-ական թվականների վերջին այս խնդիրը լուծելու համար մշակվեցին մասնագիտացված ծրագրեր, որոնք կոչվում էին «տվյալների բազաների կառավարման համակարգեր» (ՏԲԿՀ) [1, 46]: Այդ ժամանակից ի վեր նման ՏԲԿՀ-ները դարձել են փաստացի ստանդարտ, և նրանց հետ աշխատանքը կառավարելու համար մշակվել է կառուցվածքավորված հարցումների լեզու՝ SQL [1, 133]:

Միակ խնդիրն այն է, որ տվյալների բազաներում պահվող բոլոր տվյալները ունեն հստակ արժեքներ, սակայն գործնականում հաճախ է անհրաժեշտ լինում կատարել գործողություններ, որոնց հիմքում ընկած են անորոշություններ: Նման դեպքերում տվյալների բազաների և առհասարակ թվային տեղեկատվության հետ աշխատելիս ավելի նպատակահարմար է օգտագործել ոչ հստակ տրամաբանություն, որը լուծում է այդ խնդիրը:

Տվյալների բազաների առավելությունը նրանց ճարտարապետությունն է և տվյալները կառուցվածքավորված պահելու հնարավորությունը, որը թույլ է տալիս տեղեկատվության հուսալի և հարմար պահպանում: Բայց տվյալների բազաներում ինֆորմացիայի մեծ քանակության դեպքում օգտագործողի մոտ անհրաժեշտ տեղեկատվություն գտնելիս կարող են առաջանալ բարդություններ: Ներկայումս ոչ հստակ արժեքներ գտնելու խնդիրները շատ արդիական են: Դրանց լուծումները կարող են կիրառվել տարբեր ոլորտներում՝ արտադրության, մարկետինգի, բժշկության և այլն:

Հիմնախնդրի նկարագրությունը

Շատ հաճախ տարբեր տեքստային խմբագրիչներում անհրաժեշտ է լինում գտնել որոշակի բառ կամ արտահայտություն: Այդ նպատակով բոլոր տեքստային խմբագրիչները ունեն ներկառուցված որոնիչ: Բայց նրանց բոլորին միավորում է մեկ հատկություն, որը կարելի է դիտարկել ինչպես դրական, այնպես էլ բացասական տեսանկյուններից. նրանք փնտրում են օգտագործողի կողմից մուտքագրված հստակ արժեք: Օրինակ՝ եթե օգտագոր-

ծողը մուտքագրի «քաղաք Վանաձոր» բառակապակցությունը, ապա որոնիչը կգտնի միայն «քաղաք Վանաձոր» արտահայտությունը: Սակայն կարող են լինել դեպքեր, երբ տեքստը պարունակի նույն արտահայտությունը միայն կրկնակի բացատով կամ «քաղաք մարզկենտրոն Վանաձոր» արտահայտությունը: Այս երկուսն էլ պարզապես անտեսվելու են՝ առանց օգտագործողին մոտավոր համապատասխանության մասին նախազգուշացնելու: Սխալ գրված որոնման արդյունքները նույնպես անտեսվելու են: Այս իրավիճակը միշտ չէ, որ ընդունելի է, բայց այն կարելի է լուծել ինտելեկտուալ որոնման միջոցով:

Խնդրի դրվածքը

Աշխատանքի նպատակն է.

- տեքստային փաստաթղթերում ինտելեկտուալ որոնման ալգորիթմի մշակում,
- հայտնաբերված տեղեկատվության համապատասխանության աստիճանի գնահատման համար մաթեմատիկական ապարատի մշակում:

Շինգլների մեթոդ

Շինգլների մեթոդը նախատեսված է որևէ տեքստի որոշակի հատվածներ կամ կրկնօրինակներ ոչ հստակ որոնելու համար:

Շինգլը տեքստի որևէ հստակ երկարություն ուեցող սիմվոլների կամ բառերի հաջորդականություն է: Շինգլների մեթոդը սկսվում է նրանով, որ տեքստերը բաժանվում են շինգլների: Այդ շինգլները պետք է հաջորդեն միմյանց, որպեսզի տեքստի որևէ բառ կամ սիմվոլ չկորչի: Ընտրված շինգլների երկարությունը կախված է պահանջվող ճշտությունից: Այնուհետև անհրաժեշտ է որոշել, թե արդյոք ընտրված շինգլների խմբերը պետք է հատվեն (յուրաքանչյուր հաջորդ խումբը պարունակի սիմվոլ նախորդից), որը ազդում է արդյունքի ճշտության վրա, որից հետո ընտրված հաջորդականությունները խեշավորվում են, և կազմվում են դրանց գումարները: Եթե ստուգման արդյունքում կոդավորված տեքստերի գումարների միջև կա համընկնում, ապա տեքստերի նմանությունն ակնհայտ է: Եվ որքան համընկնումը շատ է, այնքան մեծ է տեքստերի նմանությունը:

Տեղեկատվության ստացման համակարգերի հարցումների տեսակները

Տեղեկատվության որոնման համակարգերում որոնվող արտահայտության տեսակները կարող են դասակարգվել ըստ մի քանի պարամետրերի: Նախ կարելի է առանձնացնել կոնտեքստից կախված և կոնտեքստից անկախ հարցումներ:

Կոնտեքստից կախված հարցումները կարող են լինել բազմազան՝ փաստաթղթում ենթատողի որոնումից մինչև նախադասության կամ պարբերության մեջ որևէ բանալի բառի առկայության ստուգում:

Որոնման հարցումների լեզուն կարող է լինել ֆորմալ (բուլյան) կամ բնական: Այդ պատճառով հարցումները կարելի է բաժանել երկու խմբի՝ բուլյան (տրամաբանական) և վարկանիշավորված [6, 455]:

Բուլյան որոնման դեպքում օգտագործողը նշում է փաստաթղթում բանալի բառերի առկայության պայմանը բուլյան ֆունկցիայի միջոցով:

Վարկանիշավորված հարցումների դեպքում հարցման տեքստը սովորաբար գրվում է բնական լեզվով: Տեղեկատվության որոնման համակարգը վերլուծում է հարցումը՝ առանձնացնելով դրա մեջ բառեր կամ հաստատուն արտահայտություններ, ընդլայնում է հարցումը որոնողական հարցման բառերի հոմանիշներով և որոնում առնվազն մեկ տերմին պարունակող բոլոր փաստաթղթերը: Գտնված փաստաթղթերը դասակարգվում են ըստ հարցման համապատասխանության աստիճանի:

Հնարավոր են հարցման բանալի բառերի և փաստաթղթի համընկնելու մի շարք տարբերակներ.

- ճշգրիտ համընկնում,
- հաշվի առնելով բառերի փոփոխությունը,
- նմանությամբ:

Ճշգրիտ համընկնման դեպքում որոնելիս որոնվում են միայն այն փաստաթղթերը, որոնցում բանալի բառը ներառված է հենց օգտագործողի կողմից հարցման մեջ նշված ձևով: Ճշգրիտ համընկնումով որոնումը հնարավորություն չի տալիս գտնելու այն բանալի բառերը, որոնք փաստաթղթում առկա են այլ քերականական ձևով, ուստի տեղեկատվության որոնման համակարգերի մեծ մասը որոնում է՝ հաշվի առնելով բառի փոփոխությունները:

Էլեկտրոնային փաստաթղթերում լինում են ուղղագրական սխալներ, և օգտագործողն էլ ոչ միշտ է ճիշտ մուտքագրում հարցման տերմինները, ուս-

տի տեղեկատվության որոնման համակարգը պետք է կարողանա գտնել հնարավորինս «նման» բառեր:

Նմանության որոնման առանցքային կետը «նմանության» աստիճանի չափման ընտրությունն է: Առավել հայտնի մեթոդներից է Լևենշտեյնի հեռավորությունը: Երկու բառերի միջև Լևենշտեյնի հեռավորությունը հավասար է մեկը մյուսի վերափոխման համար անհրաժեշտ գործողությունների նվազագույն քանակին [4, 192]:

Հիմնական ալգորիթմի նկարագրություն

Տեքստային ինֆորմացիայի ոչ հստակ որոնման իրականացման համար հիմք է ընդունվել շինգլների մեթոդը:

1) Նպատակահարմար է շինգլների երկարությունը դինամիկ կերպով փոխել՝ կախված պահանջվող ճշտությունից՝ վերցնելով առնվազն 3 նիշ: Այս 3 նիշերը կլինեն նախնական շինգլը կամ համեմատության արտահայտության երկարությունը:

2) Այնուհետև, որպես վերլուծվող արտահայտություն՝ վերցվում է կորտեժ: Այս կորտեժում կատարվում է ընտրված շինգլի որոնում: Պահվում են համընկնումների քանակը և դրանց գտնվելու վայրը:

3) Հաջորդիվ պետք է վերադասավորել սիմվոլները շինգլի մեջ և կրկնել կորտեժի վերլուծությունը: Շինգլի մեջ սիմվոլների դիրքերի փոխխությունները պետք է իրականացվեն այնքան ժամանակ, քանի դեռ չեն դիտարկվել բոլոր հնարավոր տարբերակները (3 նիշից բաղկացած շինգլի համար հնարավոր տարբերակների քանակը կլինի 9):

4) Արդյունքում ստացված բոլոր տարբերակներից ընտրվում է ամենաբարձր համապատասխանության աստիճանով տարբերակը: Այն պահվում է, և շինգլին վերագրվում է այլ արժեք (սկզբնական հարցման մեջ նշված հաջորդ սիմվոլները):

5) Առաջին կորտեժի ամբողջական վերլուծությունից հետո անհրաժեշտ է որոշել դրա համապատասխանության աստիճանը որոնելի հարցման հետ:

Տեքստային տվյալների համապատասխանության գնահատման ալգորիթմ

Բոլոր կորտեժների վերլուծությունից հետո հարց է առաջանում, թե ինչպես որոշել համապատասխանության աստիճանը: Լավագույն միջոցը համապատասխանության աստիճանը տոկոսով ներկայացնելն է: Սակայն դա հաշվարկելը բարդ է, քանի որ անհրաժեշտ է հաշվի առնել այնպիսի ցու-

ցանիշներ, ինչպիսիք են համընկնող տառերի քանակը, հնարավոր դիրքային փոփոխությունները, տողերի երկարությունների տարբերությունը, և հետևաբար սիմվոլների քանակը: Բացի դրանից՝ պետք է որոշել, թե որքանով է այս ցուցանիշներից յուրաքանչյուրը ազդում նախնական հարցման համապատասխանության աստիճանի վրա:

Ելնելով վերը նշվածներից՝ առաջարկվում է օգտագործել Զադեի ոչ հստակ տրամաբանության մեթոդները:

Ոչ հստակ բազմությունների տեսության հեղինակը Կալիֆոռնիայի Բերկլիի համալսարանի պրոֆեսոր Լոտֆի Զադեն է: Զադեի հիմնական գաղափարն այն էր, որ մարդկային դատողությունների ձևը, որը հիմնված է բնական լեզվի վրա, չի կարող նկարագրվել ավանդական մաթեմատիկական ձևակերպումների միջոցով [3, 51]:

Անորոշ բազմության գաղափարը մաթեմատիկական մոդելների կառուցման համար ոչ հստակ տեղեկատվության մաթեմատիկական ձևակերպումների փորձ է: Այս հասկացությունը հիմնված է այն գաղափարի վրա, որ տվյալ բազմությունը կազմող տարրերը, ունենալով ընդհանուր հատկություն, կարող են տարբեր աստիճանի տիրապետել այդ հատկությանը և հետեւաբար պատկանել տվյալ բազմությանը տարբեր աստիճանի:

Դիտարկենք ոչ հստակ տրամաբանության կիրառումը վերևում առաջարկված ալգորիթմի իրացման շրջանակներում:

Հաճախ որևէ կոնկրետ տարրի՝ որևէ բազմությանը պատկանելության աստիճանը որոշվում է անալիտիկ բանաձևերով ներկայացվող պատկանելության ֆունկցիայի արժեքով: Առաջարկվում է ներմուծել լեզվական փոփոխականի հասկացությունը: Լեզվական փոփոխականը այս դեպքում այն փոփոխականն է, որը կարող է ընդունել արժեքներ՝ «վատ», «նորմալ» և «գերազանց», անորոշ բազմության պատկանելության աստիճանը գնահատելիս: Սա նշանակում է, որ եթե տեքստում հայտնաբերվում է այնպիսի արժեք, որի սիմվոլները ամբողջությամբ համընկնում են նախապես տրվածի հետ (նույնիսկ եթե դրանց դասավորության կարգը չի պահպանվում), ապա լեզվական փոփոխականը ստանում է «գերազանց» արժեքը: Նմանապես կարելի է գնահատել դիրքային փոփոխությունների քանակը, որի արդյունքում վերլուծվող բառից ստանում ենք որոնելին [2, 10]:

Այսպիսով, ստանում ենք երկու մեծություն (համընկնումների տոկոս և դիրքային փոփոխությունների տոկոս): Այժմ անհրաժեշտ է պարզել այս երկու մեծությունների ազդեցության աստիճանը ընդհանուր արդյունքի վրա:

Այս նպատակով ընտրված է Մամդանի ոչ հստակ եզրակացության մեթոդը [5, 1183-1191]: Ոչ հստակ տրամաբանական եզրակացությունն ըստ Մամդանի մեթոդի կատարվում է ոչ հստակ գիտելիքների բազայի հիման վրա:

Եզրակացություն

Աշխատանքի շրջանակներում ուսումնասիրվել և վերլուծվել է տեքստային փաստաթղթերում ոչ հստակ որոնման հիմնախնդիրը: Այդ նպատակի իրականացման համար ուսումնասիրվել և հիմք է ընդունվել շինգլների մեթոդը: Ուսումնասիրվել են տեղեկատվության ստացման համակարգերի հարցումների տեսակները: Աշխատանքի շրջանակներում առաջարկված է տեքստային փաստաթղթերում ինտելեկտուալ որոնման և համապատասխանության աստիճանի գնահատման ալգորիթմ, որի հիման վրա պլանավորվում է կատարել ծրագրային իրականացում:

Օգտագործված գրականության ցանկ

1. Дейт К., Введение в системы баз данных // 8-е издание, Москва, Вильямс, 2005, 1328 с.
2. Заде Л., Понятие лингвистической переменной и его применение к принятию приближенных решений//М., Мир, 1980.
3. Макаров И., Лохин В., Манько С., Романов М., Искусственный интеллект и интеллектуальные системы управления// отв. ред. И.М. Макарова, 2013, 332 с.
4. Ehrenfeucht D., Haussler - A New Distance Metric on Strings Computable in Linear Time// Discrete Applied Mathematics, 20, 1988.
5. Mamdani E. - Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Synthesis// IEEE Transactions on Computers, C-26, 1977
6. Zobel J., Moffat A., Ramamohanarao K. - Inverted files versus signature files for text indexing//Collaborative Information Technology Research Institute, Departments of Computer Science, RMIT and The University of Melbourne, 1995

DEVELOPMENT OF ALGORITHM FOR INTELLECTUAL SEARCH AND DETERMINING THE DEGREE OF CORRESPONDENCE IN TEXT DOCUMENTS

PETROSYAN GEVORG
4th-year student of the "Computer Science and Applied Mathematics"
Department of the Faculty of Natural Sciences of Vanadzor State University
e-mail:g-a-petrosyan@mail.ru
SAHAKYAN RUSTAM
Rector of Vanadzor State University,
Doctor of Technical Sciences, Professor
e-mail: rsahakyan@yahoo.com

In the modern information society, with the increase of amount of digital information, there are difficulties in finding the desired material in text documents, because the built-in search engines in text editors only search for specific values. In practice, however, the unclear search of information is more practical. This article addresses the issue of making the intellectual search process in text documents more efficient. The aim of the work is to develop an algorithm that will perform an intellectual search in a text document and will assess the degree of relevance of the query. To achieve this goal, the following tasks were solved: study of queries of information retrieval systems, study of methods of unclear search of text information (Shingles method), development of algorithm for evaluation of text data correspondence. The following scientific methods were used: analysis, comparison, combination. As a result of the research, the problem of unclear search in text documents is analyzed, and the algorithm for intellectual search and for evaluating the degree of correspondence is proposed.

Key words: *text document, intellectual search, Shingles method, database, fuzzy set, Levenshtein distance, Mamdani method.*

РАЗРАБОТКА АЛГОРИТМА ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА И ОЦЕНКИ СТЕПЕНИ СООТВЕТСТВИЯ В ТЕКСТОВЫХ ДОКУМЕНТАХ

ПЕТРОСЯН ГЕВОРГ
Студент 4-го курса отделения
«Информатика и прикладная математика»
факультета естественных наук
Ванадзорского государственного университета
электронная почта: g-a-petrosyan@mail.ru
СААКЯН РУСТАМ
Ректор Ванадзорского государственного университета,
Кандидат технических наук, профессор
электронная почта: rsahakyan@yahoo.com

В современном информационном обществе с увеличением количества цифровой информации часто возникают трудности с поиском нужного материала в текстовых документах, потому что встроенные поисковые системы в текстовых редакторах ищут только определенные значения. Однако на практике неточный поиск информации более практичен. Эта статья посвящена проблеме повышения эффективности поиска в текстовых документах. Цель работы - разработать алгоритм, который выполнит интеллектуальный поиск в текстовом документе, оценит степень релевантности запроса. Для достижения этой цели были решены следующие задачи: изучение запросов информационно-поисковых систем, изучение метода неточного поиска текстовой информации - метода Шинглера, разработка алгоритма оценки соответствия текстовых данных. Использовались следующие научные методы: анализ, сравнение, комбинирование. В результате исследования проанализирована проблема нечеткого поиска в текстовых документах и предложен алгоритм интеллектуального поиска и оценки степени соответствия.

Ключевые слова: *текстовый документ, интеллектуальный поиск, метод Шинглера, база данных, нечеткое множество, расстояние Левенштейна, метод Мамдани.*

Հոդվածը ներկայացվել է խմբագրական խորհուրդ 13.08.2021թ.:
Հոդվածը գրախոսվել է 03.09.2021թ.: