Universitatea "Babeş-Bolyai", Cluj-Napoca, România
Facultatea de Matematică şi Informatică

Optimizarea Regăsirii Informaţiei în Spaţiul World Wide Web

Teză de doctorat

Student doctorand:
Ioan BĂDĂRÎNZĂ
Coordonator ştiinţific:
Prof. Dr. Florian Mircea BOIAN

2018

Rezumat

În această teză, cu titlul "Optimizarea Regăsirii Informaţiei în Spaţiul World Wide Web", am abordat un subiect de actualitate, care este intens studiat de către cercetători şi anume regăsirea informaţiei într-o colecţie de date.

În zilele noastre, numărul de documente web din spaţiul World Wide Web creşte foarte repede. Această viteză de creştere a numărului de pagini web disponibile în Internet este datorată faptului că utilizatorii pot adăuga conţinut rapid şi uşor. Ei pot încărca, într-un mod foarte facil, multiple fişiere de date, de la fişiere text la cele audio şi video şi, de asemenea, pot edita şi schimba conţinutul paginilor în fiecare zi. Problema care apare în urma acestei creşteri rapide este legată de faptul că a devenit foarte greu pentru o persoană să găsească răspunsuri la anumite nevoi de informare în toate aceste date disponibile în Internet. Pentru a face ca aceste date să fie disponibile utilzatorilor, s-au facut multe cercetări în domeniul de regăsire a informaţiei. În 1958, Luhn menţionează pentru prima dată problema de regăsire a informaţiei, când a şi propus prima modalitate de a determina utilitatea unui cuvânt pentru un anumit document din care acesta face parte. Acesată propunere a evoluat foarte mult în ultimele decenii. Regăsirea informaţiei este un proces prin care un sistem software care primeşte ca date de intrare necesitatea de informare reală, exprimată printr-un query, returnează ca răspuns o listă de documente relevante, documente care au fost extrase dintr-o colecţie foarte mare de date. Pentru a putea efectua o astfel de căutare într-un sistem de regăsire a informaţiei, fiecărui cuvânt din document îi este atribuită o anumită pondere care determină cât de reprezentativ este acel cuvânt pentru acel document. Mai apoi, aceste ponderi sunt folosite în diverşi algoritmi de clasare care returnează o lista de documente, ordonate după gradul de similaritate cu query-ul exprimat de utilizator.

Această teză se concentrează pe două subdomenii ale sistemelor de regăsire a informaţiei: căutarea textuală bazată pe conceptul de "bag od words" şi personalizarea sugestiilor de interogare. Conceptul de "bag of words" este cel mai utilizat concept în sistemele de regăsire a informaţiei, dar are un dezavantaj major: sensul cuvintelor şi implicit al propoziţiilor şi frazelor dintr-un document se pierde o dată cu indexarea lor. În acest sens, am facut unele studii în care am implementat diferite instrumente, pe care le-am folosit cu scopul de a demonstra eficienţa

propunerii noastre pentru a îmbunătaţii acest dezavantaj al sistemelor care implementează conceptul de "bag of words". O altă parte a sistemelor de regăsire a informaţiei care ne-a captat atenţia este mecanismul de sugestie a interogărilor. În ultimul timp, s-au facut multe cercetări în acest domeniu, în special în personalizarea sugestiilor de interogare. Aceste sugestii sunt prezentate utilizatorului în care momentul în încearcă să scrie query-ul prin care işi exprimă necesitatea de informare. Am identificat faptul că niciunul din sistemele propuse de alţi cercetători, nu foloseşte datele istorice personale de navigare a paginilor web, atunci când oferă lista de sugestii de interogare. În acest sens, am construit un instrument prin care am început să colectăm astfel de date despre utilizatori şi am demonstrat că acest istoric personal de navigare al paginilor web poate fi folosit pentru a îmbunătăţii calitatea sugestiilor oferite utilizatorilor.

Cuvintele cheie relevante pentru subiectul abordat de această teză sunt: regăsirea informaţiilor, căutare textuală, modele de clasificare, modelul boolean, modelul probabilistic, modelul spaţiului vectorial, index, index inversat, bm25, index grupat, algoritm de clasare, interogare, sugestie interogare, sub-query, sugestie personalizată, analiză sintactică, index sintactic.

Teza are următoarea structură:

În prima parte a tezei am descris datele generale despre regăsirea informaţiei textuale. La început am prezentat o scurtă istorie a regăsirii de informaţii într-o colecţie de date, când a fost pusă problema pentru prima dată şi cum a ajuns să fie atât de bine cunoscută. Am continuat cu prezentarea atât a conceptelor fundamentale cât şi a modului de lucru al unui sistem de regăsire a informaţiei. Mai întaâi am discutat despre componentele unui sistem de regăsire a informaţiei şi de ce este important pasul de preprocesare, după care am analizat cum este construit indexul unui sistem de regăsire a informaţiei, ca mai apoi să descriem detaliile despre diferite modele de clasare, cum ar fi modelul spaţiului vectorial, modelul probabilistic şi modelul boolean. După acestea, am discutat despre o caracteristică mai nouă a sistemelor de regăsire a informaţiei şi anume sugerarea de query-uri în timp ce utilizatorul începe să tasteze query-ul dorit. Această caracteristică ajută utilizatorul să işi definească mai bine query-ul prin care doreşte să îşi exprime necesitatea de găsire a unor informaţii. În încheiere, am descris modalităţile de evaluare a sistemelor de regăsire a informaţiei şi a rezultatelor returnate de acesta şi, nu în ultimul rând, cum putem evalua lista de sugestii returnată de un system de sugestii pentru un anumit query.

În continuare am prezentat contribuţiile majore din această teză:

Capitolul 3.1 - Indecşi Grupaţi pe Nivele: În acest capitol este prezentat un sistem de regăsire a informaţiei bazat pe indecşi grupaţi pe nivele, care are la bază modelul spaţiului vectorial. Acest sistem poate fi utilizat atât pentru căutarea exactă de fraze, cât şi pentru o căutare mai bună a cuvintelor cheie prin utilizarea ponderii de proximitate în funcţia de măsurare a similarităţii. Procesul de regăsire a informaţiei este bazat pe grupuri de documente care sunt similare între ele, iar în momentul căutării sunt verificaţi doar liderii acestor grupuri, nu şi toate documentele din baza de date.

Capitolul 3.2 - Indecşi Sintactici: Acest capitol descrie trei metode prin care se pot încorpora metadate într-un sistem de regăsire a informaţiei textuale. Prima metodă implică doar o analiză sintactică a query-ului, iar pe baza categoriei gramaticale generează o pondere diferită pentru fiecare termen din query, ponderi care mai apoi vor fi folosite în algoritmul de căutare. A doua metodă implică optimizarea spaţiului de stocare a indexului inversat prin stocarea doar a acelor termeni care apar ca subiect şi predicat într-un document. A treia metodă propune construirea unui index complet sintactic, ceea ce înseamnă că pentru fiecare termen din colecţie, indexul inversat va stoca şi categoria gramaticală a termenului, alături de frecvenţa termenului într-un document şi frecvenţa inversă a termenului.

Capitolul 3.3 - Ponderi Personalizate ale Termenilor: După efectuarea unei analize a valorilor IDF pentru termenii din colecţia Reuters, am observat că pentru interogările care conţin perechi de termeni, unul cu valori foarte mici ale IDF-ului şi altul cu valori IDF foarte mari, sistemele de regăsire a informaţiei nu returnează documente relevante. În acest capitol analizăm o metodă bazată pe modelul probailistic care permite utilizatorului să modifice ponderile termenilor din query în aşa fel încât sistemele de regăsire a informaţiei să returneze rezultate mai relevante. Această metodă nu are rolul de a spori relevanţa rezultatelor pentru căutările generale, ci are scopul de a creşte relevanţa rezultatelor returnate pentru anumite căutări specifice în care termenii query-ului folosit au valori IDF disproporţionate.

Capitolul 4.1 - Analiză a Istoriei de Navigare Web a Utilizatorului: În acest capitol subliniem rezultatele cercetării în sistemele de sugestii de interogare cu un accent deosebit acordat personalizării acestora. Sugestiile de interogare sunt un mecanism prin care utilizatorul primeşte o listă de interogări posibile de unde poate alege şi poate efectua o căutare într-un mod mult mai rapid şi mai uşor. Am evaluat utilitatea istoricului de navigare a paginilor web a unui utilizator în generarea de noi sugestii. Timp de o luna de zile am realizat un experiment în care, cu ajutorul unui plugin de Chrome, am colectat informaţii despre paginile web vizitate de utilizatori, alături de interogările pe care aceştia le făceau pe motorul de căutare Google şi am constatat că aproximativ 32% din aceste interogări ar fi putut fi prezise din acea istorie de navigare a fiecărui utilizator în parte.

Capitolul 4.2 - Personalizarea Sugestiilor de Interogare: După ce am observat

faptul că istoricul personal de navigare al paginilor web poate fi folosit pentru a prezice următorul query pe care utilizatorul va încerca să îl folosească, în acest capitol am propus o nouă metodă de personalizare a ordinii în care sunt prezentate sugestiile de interogare. Astfel, sugestiile mai relevante pentru utilizator vor fi plasate pe un loc din listă cât mai sus (să aibă un rang mai mare). Modalitatea prin care am realizat acest lucru este definirea unui scor *Personal Temporal Query Suggestion* pentru fiecare sugestie din listă. Acest scor ţine cont de istoricul personal, foarte recent şi scurt, de navigare a paginilor web, care este folosit în continuare pentru a calcula un $HybridPageScore$. Capitolul mai conţine evaluarea şi experimentele pe care le-am efectuat pentru metoda propusă.

În capitolul de concluzii, am prezentat principalele contribuţii ale acestei teze la optimizarea sistemelor de regăsire a informaţiilor întro colecţie de date şi direcţiile viitoare de cercetare în acest domeniu.