PENINGKATAN PERFORMANSI SISTEM TEMU BALIK INFORMASI DENGAN METODE *PHRASAL TRANSLATION* DAN *QUERY EXPANSION*

Ari Wibowo
Teknik Multimedia dan Jaringan, Polteknik Negeri Batam
wibowo@polibatam.ac.id

Abstract

Development of the Internet as a medium of information very rapidly today. Almost all of 24 hours a day people spend time at the computer. All this can not be separated from one branch of science called the informatics behind information retrieval systems (information retrieval). Even though most people use the internet, there are also some people who still cloud this issue. How would someone want to find an article in another language, but he did not know or forget the word (in other languages before). In this case there is a part of information retrieval called CLIRS (Cross Lingual Information Retrieval System) or through the information retrieval systems across languages. The system works like a user enters a word he was looking for an article later mentranslasikannya system and remove these articles are certainly in a different language. For this study, CLIRS viewed from several methods, namely, dictionary, phrasal translation and query expansion.

Key Word : CLIRS, query, phrasal

1. PENDAHULUAN

Saat ini jumlah informasi yang tersedia di internet semakin banyak dan terus meningkat dengan tajam. Informasi-informasi tersebut tersedia dalam berbagai format, seperti teks, *audio*, dan *visual*. Dengan semakin banyak dan beragamnya informasi yang tersedia, kebutuhan pengguna internet telah bergeser dari arah kuantitatif ke arah kualitatif. Kebutuhan yang semula berupa informasi sebanyak-banyaknya telah bergeser menjadi informasi secukupnya asalkan relevan dengan keperluan. Walaupun tersedia secara gratis dan dalam jumlah banyak, keberadaan mesin pencari (*Search Engine*) di internet dirasakan masih kurang dari aspek kualitas. Hal ini wajar, mengingat hasil pencarian yang diberikan oleh mesin-mesin pencari tersebut seringkali membludak dan kurang relevan. Oleh karena itu, kebutuhan akan suatu mekanisme pencarian dokumen yang lebih efektif dirasakan semakin mendesak.

Indikator yang lazim dipakai untuk menilai keakuratan dan kerelevansian hasil pencarian suatu dokumen adalah kesesuaian (presisi) antara *query* yang diberikan dan dokumen yang diperoleh. Di dalam bidang ilmu Sistem Temu Balik Informasi (STBI), dikenal berbagai model untuk menilai secara obyektif presisi dari suatu pencarian, antara lain model ruang-vektor (*Vector-Space Model*) dan model probabilistic (*Probabilistic Model*).

Penggunaan salah satu model di atas dapat dilihat pada *Cross Lingual Information Retrieval System* (CLIRS) atau sistem temu balik informasi lintas bahasa. Jenis *Information Retrieval* (IR) satu ini menggunakan dua atau lebih bahasa sebagai query dan hasil yang ingin didapat sehingga dapat melihat satu artikel yang artinya bisa berbeda jika diterjemahkan ke dalam bahasa yang berbeda pula.

2. METODE PENELITIAN

Metodologi yang diterapkan dalam penelitian ini adalah sebagai berikut:

1. Studi Literatur

   Eksplorasi dan studi literatur dilakukan dengan mempelajari cara kerja *phrasal translation, query expansion,* dan CLIRS melalui literatur – literatur seperti buku (*textbook*), paper dan sumber ilmiah lain seperti situs internet, artikel dokumen teks yang berhubungan.
2. Analisis dan Perancangan Perangkat Lunak

   Analisis dan perancangan perangkat lunak dilakukan untuk menentukan permasalahan

mengenai bahasa pemrograman apa yang digunakan, struktur data, input/output dari program, dan permasalahan teknik bagaimana algoritma akan diimplementasikan.

3. Implementasi Program dan Pengujian Performansi

Detail mengenai implementasi program dilakukan sesuai hasil analisis pada tahap sebelumnya. Pengujian performansi *phrasal translation* dan *query expansion* dilakukan dengan membandingkan *Non-Interpolated Average Precision* (NIAP) dari kedua metode CLIRS tersebut.

4. Analisis Hasil dan Penarikan Kesimpulan

Analisis hasil dilakukan untuk mengetahui performansi metode *phrasal translation* dan *query expansion* pada CLIRS tersebut. Jika ternyata performansi yang ditampilkan lebih baik, akan dilakukan analisis mengapa bisa demikian. Setelah analisis hasil selesai, dilakukanlah penarikan kesimpulan terhadap performansi metode *phrasal translation* dan *query expansion*.

2.1 Sistem Temu Balik Informasi

Sistem Temu Balik Informasi (Information Retrieval) adalah ilmu mencari informasi dalam suatu dokumen, mencari dokumen itu sendiri dan mencari metadata yang menggambarkan suatu dokumen. Sistem Temu Balik Informasi merupakan cabang dari ilmu komputer terapan (applied computer science) yang berkonsentrasi pada representasi, penyimpanan, pengorganisasian, akses dan distribusi informasi [KAN05]. Dalam sudut pandang pengguna, Sistem Temu Balik Informasi membantu pencarian informasi dengan memberikan koleksi informasi yang sesuai dengan kebutuhan pengguna.

2.2 Sistem Temu Balik Informasi Lintas Bahasa

Sistem temu balik informasi lintas bahasa atau dalam bahasa Inggris dinamakan *Cross-Lingual Information Retrieval System (CLIRS)* merupakan cabang dari IR yang menangani pemenuhan informasi yang dituliskan dalam bahasa yang berbeda dengan apa yang dimasukkan oleh *query user*. Misalnya *user* memasukkan *query* dalam bahasa Indonesia kemudian sistem mencari dokumen-dokumen yang relevan dalam bahasa Inggris. Penggunaan CLIRS itu sendiri sebenarnya ditekankan untuk seseorang yang misalnya dia bisa berbahasa Inggris namun pasif kemudian dia hendak mencari suatu dokumen yang berhubungan dengan kerajaan Inggris dia memasukkan *query* "kerajaan Inggris" lalu sistem me-*retrieve* semua dokumen (dalam bahasal lain) yang memuat *query* tersebut.

Workshop pertama mengenai CLIRS diadakan di Zurich ketika konferensi SIGIR-96. Hasil dari workshop ini bisa ditemukan pada buku *Cross-Language Information Retrieval* (Grefenstette, ed; Kluwer, 1998) ISBN 0-7923-8122-X. Kemudian *workshop* dilakukan secara rutin sejak tahun 2000 pada pertemuan *Cross Language Evaluation Forum (CLEF)*. *Term* "cross-language information retrieval" mempunyai banyak sinonim, biasanya yang sering digunakan adalah : *cross-lingual information retrieval, translingual information retrieval, multilingual information retrieval*. *Term* "multilingual information retrieval" bisa diartikan CLIR pada umumnya, namun juga memiliki makna yang spesifik dalam sistem temu balik informasi lintas bahasa dimana dokumen koleksinya *multilingual*.

2.3 Phrasal Translation

Gagal dalam mentranslasikan konsep multi-term sebagai frase sangat mengurangi keefektifan dari dictionary translation. Pada eksperimen di mana frase query ditranslasi secara manual [BC96], performansi meningkat sebanyak 25% melebihi automatic word-by-word (WBW) translasi query. Ada hipotesis yang mengatakan bahwa cara ini secara otomatis mengindentifikasi frase dan mendefinisikannya seperti WBW dapat meningkatkan keefektifan. Phrasal translation berbasiskan basis data frase dan kata yang telah didefinisikan terlebih dahulu. Ketika frase ditranslasikanm basis data mencari frase dalam bahasa Inggris. Jika ketemu maka mengeluarkan arti kata dalam bahasa Indonesia yang berbentuk frase juga. Jika lebih dari satu yang ditemukan maka ditambahkan ke query.

2.4 Model Probabilistik

Menurut [DIK02], model adalah pola (contoh, acuan, ragam) dari sesuatu yang akan dibuat atau dihasilkan. Selain itu, model – secara ilmiah – dapat diartikan sebagai idealisasi atau abstraksi

dari proses yang sebenarnya. Kesimpulan yang diambil berdasarkan suatu model akan sangat tergantung dari kesesuaian model tersebut dengan keadaan sebenarnya. Model dalam Sistem Temu Balik Informasi dipakai untuk menentukan detail dari sistem, yaitu bagaimana merepresentasikan dokumen dan *query*, melakukan pencarian, dan notasi kesesuaian antara dokumen dan *query*[KAN05].

Dalam model probabilistik, notasi yang lazim dipakai untuk merepresentasikan kerelevansian suatu dokumen adalah $P(X)$ dan $P(X|Y)$. $P(X)$ adalah notasi untuk kemungkinan X, sementara $P(X|Y)$ adalah notasi untuk kemungkinan X, jika diberikan Y.

Salah satu implementasi model probabilistik yang sangat populer dan lazim dipakai adalah *Binary Independence Retrieval Model* (BIR). Dalam BIR, sama seperti model probabilistik lainnya, sistem akan mencari probabilitas suatu dokumen d_m relevan terhadap *query* q_k. Notasi yang dipakai bagi nilai probabilitasnya adalah [RIJ79] $P(R|q_k, d_m)$.

Karena model probabilistik mengasumsikan bahwa setiap dokumen dideskripsikan lewat "ada" atau "tidak ada"-nya *term* indeks, maka dokumen dapat direpresentasikan menjadi vektor biner. Secara matematis [RIJ79]:

$$x = (x_1, x_2, ...x_n) \quad \text{..... (2.1)}$$

Dimana $x_1 = 0$ jika *term* indeks tidak terdapat di dalam dokumen tersebut dan $x_1 = 1$ jika *term* indeks ada di dalam dokumen tersebut. Dengan demikian, dokumen juga dapat direpresentasikan dengan notasi d_1 dan d_2.

d_1 = Dokumen adalah relevan

d_2 = Dokumen yang tidak relevan

Karena sifatnya yang biner, maka persamaan [RIJ79]:

$$P(d_1|x) + P(d_2|x) = 1 \quad \text{..... (2.2)}$$

harus terpenuhi.

Untuk memperoleh rumus yang tepat bagi penghitungan probabilitas, model probabilistik mengaplikasikan dua jenis transformasi [FUH92]:

1. Teorema Bayes, dalam bentuk

$$P(a|b) = P(b|a)\frac{P(a)}{P(b)} \quad \text{....…..... (2.3)}$$

2. Penggunaan faktor O, yaitu

$$O(y) = P(b|a)\frac{P(y)}{P(\overline{y})} = \frac{P(y)}{(1-P(y))} \quad \text{..(2.4)}$$

Dalam model probabilistik, *similarity* dihitung berdasarkan faktor O antara *query* yang menghasilkan dokumen relevan dengan *query* yang menghasilkan dokumen yang tidak relevan [FUH92].

1. *Query* yang menghasilkan dokumen relevan [FUH92]:

$$O(p) = P(x_i = 1|R) = \frac{P_i}{1-P_i} \quad \text{....(2.5)}$$

2. *Query* yang menghasilkan dokumen tidak relevan [FUH92]:

$$O(r) = P(x_i = 1|\overline{R}) = \frac{r_i}{1-r_i} \quad \text{....(2.6)}$$

3. Sehingga *similarity* (dalam bentuk logaritma) adalah [FUH92]:

$$S_i = \log\frac{p_i}{r_i}\frac{(1-r_i)}{(1-p_i)} \quad \text{……..……(2.7)}$$

Semakin besar nilai S_i, semakin besar pula probabilitas bahwa d_m relevan terhadap *query* q_k. Prinsip inilah yang dipakai dalam *Probability Ranking Principle* (PRP) dalam pengurutan dokumen.

3. ANALISIS PERANCANGAN

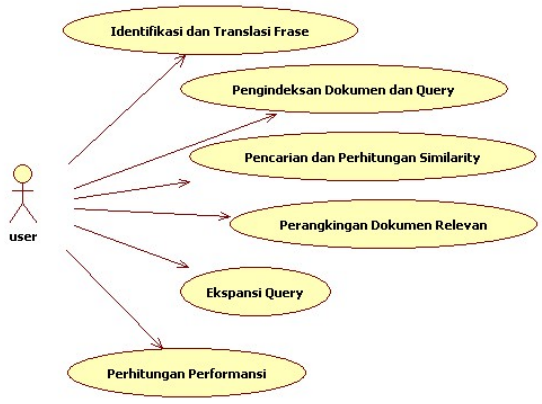
3.1 Analisis Kebutuhan Perangkat Lunak

Dalam penelitian ini, akan dibangun sebuah perangkat lunak Sistem Temu Balik Informasi Lintas Bahasa (Indonesia - Inggris) yang mengimplementasikan metode *phrasal translation* dan *query expansion*. Analisis kebutuhan perangkat lunak terdiri dari spesifikasi kebutuhan perangkat lunak, tujuan pengembangan perangkat lunak dan analisis *use case*. Perancangan perangkat lunak terdiri dari batasan perancangan perangkat lunak, perancangan arsitektur perangkat lunak, *class diagram*, *sequence diagram* dan perancangan antarmuka perangkat lunak.

Perangkat lunak yang dibangun nantinya diharapkan mampu mengimplementasikan fungsi-fungsi berikut:

1. Melakukan identifikasi frase dari dokumen dan *query*

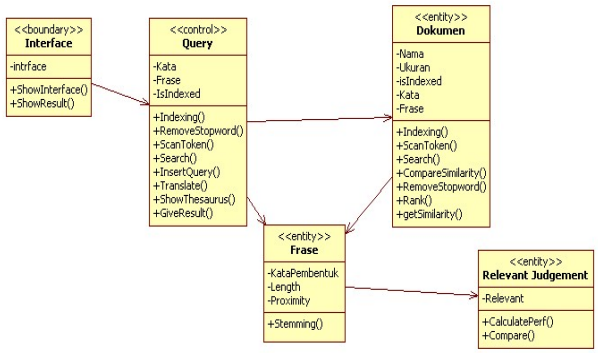
2. Melakukan translasi ke bahasa Inggris dari *query* yang dimasukkan.
3. Melakukan pengindeksan terhadap dokumen dan *query*.
4. Melakukan pencarian dokumen yang relevan dengan query.
5. Melakukan pengurutan peringkat dokumen hasil pencarian.
6. Melakukan pengindeksan ulang jika adanya kata tambahan yang dimasukkan sesuai peringkat dokumen.
7. Menghitung nilai Non-*Interpolated Average Precision* untuk menilai performansi sistem.



Gambar 1 Use Case Diagram

3.2 Diagram Kelas

Perancangan kelas perangkat lunak mengacu pada hasil analisis kelas potensial pada Tabel III-2. Hasil perancangan kelas tersebut dituangkan dalam Gambar III-3 berikut:



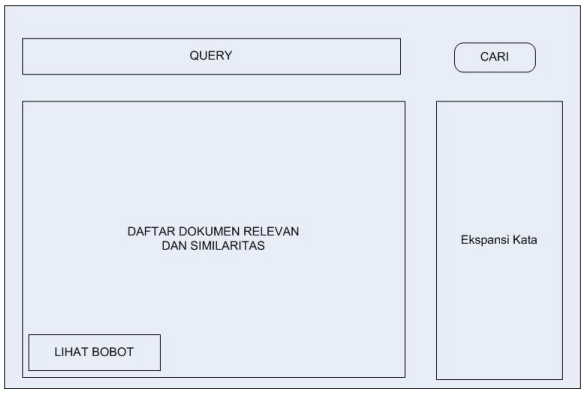
Gambar 2 Kelas Diagram

Keterangan mengenai kelas-kelas di atas adalah sebagai berikut:

1. Interface
   Kelas antarmuka, memiliki satu atribut saja, yaitu *intrface*. Kelas ini menangani operasi yang berkaitan dengan tampilan.
2. Dokumen
   Kelas ini mempunyai atribut Nama, Ukuran, *isIndexed*, Kata dan Frase. Kelas ini menangani operasi-operasi berkaitan dengan dokumen, seperti pengindeksan dan penghilangan *stopwords*.
3. Query
   Kelas ini memiliki atribut *isIndexed*, Kata dan Frase. Kelas ini menangani operasi-operasi berkaitan dengan *query*, seperti pengindeksan, translasi dan penghilangan *stopwords*.
4. Frase
   Kelas ini memiliki atribut Kata Pembentuk, dan *Length.*Kelas ini menangani operasi-operasi berkaitan dengan frase, seperti kata pembentuk, dan *length*. Kelas ini dihasilkan dari kelas dokumen dan *query*.
5. Relevant Judgement
   Kelas bawaan sistem. Kelas ini merupakan kelas yang berfungsi untuk membandingkan perhitungan sistem dengan basis data yang telah didefinisikan terlebih dahulu. Mempunyai beberapa operasi yaitu kalkulasi performansi dan perbandingan perhitungan *similarity*.

3.3 Perancangan Antarmuka Perangkat Lunak

Rancangan layar utama pada perangkat lunak diperlihatkan pada gambar III-10. Tampak ada sebuah kotak (*textbox*) untuk memasukkan *query* yang kemudian akan ditranslasikan dengan mengklik *button* "CARI". Di kiri bawah adalah kotak yang nantinya berisikan kata setelah translasi, bobotnya, dokumen-dokumen yang dihasilkan dan nilai similaritas yang didapat setelah perhitungan. Di kanan bawah ada kotak yang nantinya berisikan ekspansi kata dari *query* dari dokumen-dokumen yang memuat kata-kata sesuai *query*.



Gambar 3 Rancangan Antarmuka

4. HASIL DAN PEMBAHASAN

4.1 Batasan Pengujian

Batasan pengujian perangkat lunak adalah sebagai berikut:

1. Minimal kata pada frase adalah 2 kata dan maksimal 3 kata. Pembatasan ini dilakukan karena ada beberapa *query* yang hanya memiliki panjang 2 kata.
2. Pengujian *domain* frase tidak diperhitungkan, karena sama sekali tidak mempengaruhi hasil NIAP.
3. Pengujian untuk *query expansion* hanya dapat dilakukan jika *query* awal sudah dicoba terlebih dahulu.
4. Maksimal panjang *query* adalah 60 kata dan sebaiknya hindari penggunaan *stopwords* untuk hasil yang lebih maksimal.
5. Maksimal waktu eksekusi *query* adalah 60 detik dan jika lebih dari itu sistem akan *hang.*

Pelaksanaan Pengujian

1. Pengujian dilakukan dengan melakukan pencarian dokumen yang sesuai dengan query yang ada dalam koleksi
2. Jumlah dokumen yang hasil pencarian yang ditampilkan adalah 10 peringkat teratas
3. Perhitungan rata-rata bobot dilakukan terhadap 50 *query* yang telah terdefinisi terlebih dahulu

Hasil Pengujian

Tabel 1 – Hasil Pengujian Query

Query	Worb By Word	Frase	Query Expansion
1	0.036	0.024	0.040
2	0.096	0	0.200
3	0.048	0	0.061
4	0.040	0	0.080
5	0.125	0	0.232
6	0.072	0	0.158
7	0.058	0	0.078
8	0.204	0.335	0.282
9	0.032	0	0,282
10	0.067	0.005	0.075
11	0.020	0.041	0.043
12	0.155	0	0.195
13	0.468	0	0.468
14	0.079	0.019	0.081
15	0.034	0.151	0.042
16	0.090	0.029	0.095
17	0.050	0	0.050
18	0.030	0.089	0.194
19	0.194	0	0.330
20	0.755	0	0.755
21	0.041	0.095	0.061
22	0.109	0	0.138
23	0.125	0	0.128
24	0.058	0.005	0.061
25	0.031	0	0.032
26	0.072	0	0.139
27	0.021	0	0.021
28	0.053	0	0.064
29	0.085	0	0.096
30	0.091	0.200	0.125
31	0.229	0	0.256
32	0.127	0	0.133
33	0.014	0.077	0.018
34	0.132	0.005	0.135
35	0.144	0	0.189
36	0.051	0.031	0.061
37	0.100	0.007	0.101
38	0.130	0.119	0.145
39	0.117	0	0.153
40	0.037	0.040	0.050
41	0.030	0.033	0.035
42	0.063	0.040	0.086
43	0.026	0	0.026
44	0.159	0.007	0.303
45	0.069	0.064	0.116
46	0.089	0.021	0.094
47	0.050	0	0.050
48	0.088	0.048	0.088
49	0.068	0	0.064
50	0.068	0	0.096
Rata-rata	**0.09471**	**0.0345**	**0.118953**

4.2 Pengujian dengan Membandingkan Nilai Performansi antara Frase Dua dan Tiga Kata

Pengujian ini dilakukan untuk membandingkan nilai NIAP dari frase yang terdiri dari dua dan tiga kata. Nilai ini diperoleh dengan menggunakan aplikasi lain diluar CLIRS yang dikembangkan secara bersamaan. Adapun nilai yang dibandingkan adalah frase dua dan tiga kata dalam bahasa Inggris dan frase dua dan tiga kata setelah dilakukannya translasi.

Pelaksanaan Pengujian

1. Pengujian dilakukan dengan menghitung nilai NIAP sesuai dokumen relevan yang dihasilkan
2. Frase dibagi menjadi 2 dan 3 kata kemudian dilakukan translasi untuk *query* bahasa Indonesia
3. Jumlah dan letak dokumen relevan telah terdefinisi terlebih dahulu pada *relevant judgement* yang diberikan

Hasil Pengujian

Tabel 2 - Perbandingan Nilai NIAP antara Frase Dua dan Tiga Kata

Frase 2 Kata (translasi)	Frase 2 Kata (Inggris)	Frase 3 Kata (translasi)	Frase 3 Kata (Inggris)
0.10224980	0.1144563	0.112087	0. 126129

4.3 Analisis Hasil Pengujian

Berdasarkan Tabel 1 dapat dilihat dari 50 *query* yang digunakan untuk pengujian bahwa hampir semua nilai melalui metode *query expansion* mendapatkan hasil yang lebih baik daripada kata per kata. Sebaliknya, tidak semua *query* melalui metode *phrasal translation* mendapatkan nilai yang lebih baik dari kata per kata. Sepertinya kamu yang digunakan belum terlalu lengkap sehingga banyak kata tidak diartikan secara frase pada sistem tersebut. Selain itu, karena koleksi dokumen yang terlalu banyak, sistem hanya menggunakan 10 dokumen teratas saja untuk perhitungan (namun ini bukan menjadi penyebab utama mengapa nilai frase lebih kecil). Dari nilai rata-rata bobot masing-masing metode dapat diambil kesimpulan bahwa metode *query expansion* merupakan metode terbaik dalam hal peningkatan bobot *query* disusul kata per kata dan frase.

Pada Tabel 2 yaitu perbandingan nilai NIAP antara frase dua dan tiga kata terlihat bahwasanya untuk frase tiga kata memiliki nilai yang lebih baik. Melalui nilai ini, didapat bahwa untuk sebuah *query* akankah lebih baik jika kata yang ingin dilakukan pencarian adalah lebih dari satu kata. Hal ini dimaksudkan agar dokumen yang dihasilkan akan lebih akurat dan relevan. Seperti paragraf di atas, yaitu untuk pencarian memang bahasa asli lebih akurat daripada setelah dilakukan translasi.

5. KESIMPULAN

1. Identifikasi frase akan memberikan hasil yang lebih baik jika kamus kata yang dimiliki lebih lengkap.
2. Ekspansi *query* sangat efektif untuk mendapatkan dokumen yang sesuai karena memiliki nilai keakuratan yang tertinggi.
3. Nilai performansi dari sistem dengan translasi frase lebih tinggi dari sistem dengan translasi kata per kata.
4. Nilai performansi dari frase tiga kata lebih baik dari frase dua kata baik setelah translasi maupun sebelum.
5. Performansi asal tanpa translasi selalu lebih baik daripada setelah dilakukan translasi baik kata per kata maupun frase.

6. SARAN

1. *Term* indeks sebaiknya langsung dibuat diluar sistem namun yang dapat merangkum koleksi dokumen yang lebih banyak.
2. Untuk penelitian berikutnya, sebaiknya koleksi dokumen yang sudah ada ditambahkan dengan dokumen-dokumen baru dengan tema yang lebih ambigu. Hal ini dimaksudkan untuk menguji lebih lanjut performansi yang diberikan oleh model probabilistas.

7. DAFTAR PUSTAKA

[1] Ballesteros, L. & Croft, B. (1996). "Dictionary methods for cross-lingual information retrieval". In: *Database and Expert Systems Applications. 7th International Conference, DEXA '96 Proceedings*. Springer-Verlag Berlin, Germany.

[2] Ballesteros, L. & Croft, W. B. (1997). "Phrasal translation and query expansion techniques for cross-language information retrieval". In: *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 84 - 91. Association for Computing Machinery.

[3] Ballesteros, L. & Croft, W. B. (1998). "Resolving ambiguity for cross-language retrieval". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery.
[4] *Document Retrieval: An Examination of Syntactic and Non-Syntactic Methods*. Cornel University
[5] Fuhr, Norbert. 1992. *Probabilistic Models in Information Retrieval*. Computer Journal