Journal of Information System Research (JOSH)
Volume 4, No. 3, April 2023, pp 748−754
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i3.2867

Pemrosesan Query dan Pemeringkatan Hasil dalam Information Retrieval: Sebuah Kajian Literatur

Roberto Kaban*, Poltak Sihombing, Mahdianta Pandia, Purwanto Simamora

Fakultas Ilmu Komputer dan Teknologi Informasi, Teknik Komputer, Universitas Sumatera Utara, Medan
Gedung C Fasilkom-TI, Universitas Sumatera Utara, Jl. Alumni No.3, Padang Bulan, Kec. Medan Baru, Kota Medan, Sumatera Utara, Indonesia
Email: [1,*]roberto.kaban@yahoo.com, [2]poltak@usu.ac.id, [3]mahdianta03@gmail.com, [4]purwantosimamora@gmail.com
Email Penulis Korespondensi: roberto.kaban@yahoo.com
Submitted: **07/01/2023**; Accepted: **12/04/2023**; Published: **30/04/2023**

Abstrak−Dalam penelitian ini, kami mengkaji literatur tentang temu balik informasi (Information Retrieval) mulai dari dasar-dasar Information Retrieval (IR), komponen dan tantangan kedepannya. Tujuan dari penelitian ini adalah mengamati teknik yang pernah dilakukan oleh peneliti sebelumnya dalam IR, khususnya pemrosesan kueri dan pemeringkatan hasil pencarian. Kami menggunakan metode tinjauan literatur dengan mengidentifikasi, meninjau, dan mengamati teknik dalam IR berdasarkan hasil beberapa penelitian sebelumnya. Kami menggunakan lebih banyak sumber literatur dari ACM Digital Library, Researchgate dan MDPI. Dari penelitian ini, kami menemukan beberapa model IR untuk pencarian dokumen (informasi) yang relevan dan pemeringkatan hasil yang akurat antara lain EXplanaTion RAnking (EXTRA), Deep-QPP, ColBERT-PRF dan UQSCM-RFD.

Kata Kunci: Information Retrieval (IR); Temu Balik Informasi; Model IR; Pemrosesan Kueri; Pemeringkatan Pencarian; Pencarian Cerdas

Abstract−In this study, we reviewed the literature on information retrieval starting from the basics of Information Retrieval (IR), components and future challenges. The purpose of this study is to observe techniques that have been used by previous researchers in IR, especially query processing and ranking of search results. We used a literature review method by identifying, reviewing, and observing techniques in IR based on the results of several previous studies. We collected more literature sources from the ACM Digital Library, Researchgate and MDPI. From this research, we found several IR models for searching relevant documents (information) and ranking accurate results, including EXplanaTion RAnking (EXTRA), Deep-QPP, ColBERT-PRF and UQSCM-RFD.

Keywords: Information Retrieval (IR); IR Models; Query Processing; Retrieval Ranking; Intelligent Search

1. PENDAHULUAN

Perluasan pengolahan data terlihat telah berkembang secara dramatis dalam beberapa tahun terakhir, hal ini tentu saja menimbulkan tantangan tambahan bagi para peneliti untuk menemukan pendekatan inovatif untuk mengekstrak informasi terkait dalam waktu yang lebih singkat. Berbagai teknik telah digunakan untuk pengolahan data penting dari sebuah repositori sebagai pangkalan data maupun dalam bentuk database. Data yang di kelola biasanya berisi berbagai jenis, seperti data yang sudah terstruktur, semi-terstruktur, tidak terstruktur, atau bahkan yang bersifat heterogen.

Berbagai solusi, mulai dari penyimpanan hingga pencarian informasi, sedang dipelajari untuk mengakomodasi kecepatan temu balik informasi (Information Retrieval). Information Retrieval System (IRS) adalah semua aspek yang mencakup struktur, analisis, penyimpanan, pencarian, dan temu kembali informasi. IRS digunakan untuk menemukan kembali dokumen-dokumen yang sesuai dengan kueri (kata kunci) yang diberikan pengguna pada suatu koleksi dokumen. Dalam hal ini, akan muncul pertanyaan dengan sendirinya, yaitu apakah sistem temu balik informasi (IRS) benar-benar mengambil dokumen (hasil) yang akurat dan relevan sesuai kebutuhan pengguna?.

Sebagian besar IRS masih memberikan hasil yang tidak konsisten dan tidak relevan dengan informasi yang dibutuhkan pengguna. Kueri dengan kata kunci yang pendek yang terdiri dari satu atau dua kata akan menyulitkan IRS untuk menyajikan hasil yang akurat. Kemudian, bagaimana merepresentasikan konten sedemikian rupa sehingga dapat dibaca oleh IRS juga menjadi masalah tersendiri. Selain itu, terdapat juga tantangan dalam model pemeringkatan atas kesesuaian hasil pencarian dengan kueri kata kunci yang di inputkan oleh pengguna dan pemberian skor untuk perangkingan pada hasil pencarian. Penyajian informasi yang tidak akurat akibat keterbatasan kemampuan pengguna untuk menjelaskan informasi apa yang dibutuhkan bersumber dari pengguna yang memberikan kata kunci yang ambigu dengan kata yang memiliki makna ganda.

Menyusun ulang kueri (query reformulation) yang diinputkan oleh pengguna adalah hal yang sering dilakukan dalam Information Retrieval (IR). Hal ini dilakukan untuk mengatasi ketidaksesuaian antara kueri yang didinputkan dengan informasi yang didinginkan pengguna. Pada IRS, terkadang dokumen yang di-retrieve dari kueri yang dimasukkan pengguna tidak relevan. Hal ini bisa saja disebabkan human error (salah ketik) ataupun karena pengguna tidak tahu kueri apa yang harus dimasukkan ke sistem. Oleh karena itu, diimplementasikan relevance feedback melalui refulmasi kueri agar didapatkan return hits yang relevan dan sesuai dengan kebutuhan user. Relevance feedback merupakan proses formulasi ulang kueri awal berdasarkan informasi umpan balik relevansi dari pengguna terhadap dokumen-dokumen hasil pencarian awal. Dengan menggunakan relevance

Journal of Information System Research (JOSH)
Volume 4, No. 3, April 2023, pp 748−754
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i3.2867

feedback, pengguna dapat memilih informasi yang relevan terhadap kebutuhannya, dan dari feedback-feedback pengguna tersebut yang selanjutnya digunakan untuk menentukan kueri baru.

Kajian literatur merupakan uraian tentang teori, temuan, dan bahan penelitian lainnya yang berhubungan dengan objek yang diteliti. Penelitian ini mengamati teknik yang pernah dilakukan oleh peneliti sebelumnya dalam IR, khusunya dalam perosesan kueri dan bagaimana model pemeringkatan hasil pencarian dalam IRS. Penelitian sebelumnya [1], dibahas berbagai fase teknik prapemrosesan dokumen juga membandingkan beberapa pendekatan yang ada seperti KNN, SVM, Naïve Bayes, dan Algoritma Rocchio. Sebagian besar penelitian berfokus pada metode statistik dengan menggunakan pendekatannya ruang vektor[2]. Kesamaan antara dokumen atau kata kunci (kueri) dikuantifikasi dalam bidang pencarian informasi dengan menunjuknya ke dalam vektor frekuensi kata dalam ruang vektor, selanjutnya ditentukan sudut antara kedua vektor, sehingga jarak antara dua vektor dalam ruang menentukan kesamaan kata. Jarak kosinus dapat digunakan sebagai ukuran kesamaan antara dua kata dalam rentang 0 dan 1[3].

Perluasan kueri juga merupakan salah satu metode yang digunakan untuk meningkatkan sistem pencarian untuk hasil yang cepat dan relevan. Banyak frasa penting mungkin hilang dari kueri yang di inputkan pengguna, menyebabkan IRS merespons dengan tidak efektif, sehingga hasil yang ditemukan kurang relevan dengan kueri. Masalah ini awalnya dipecahkan oleh Rocchio [4], yang kemudian menghadirkan IRS yang relevan yang secara otomatis memperluas pencarian asli berdasarkan umpan balik pengguna atau restrukturisasi kueri dengan menyediakan lebih banyak istilah seperti sinonim, jamak, pengubah, dan sebagainya[5]. Dalam hal kata kunci kueri yang pendek berupa kata, IRS umumnya sudah bekerja dengan baik, tetapi akan bermasalah dan tidak efektif untuk kueri yang panjang berupa kalimat (verbose). Teknik pemrosesan kueri dapat diterapkan pada kueri yang sifatnya verbose sebelum dikirim ke IRS untuk meningkatkan hasil pencarian. Penggunaan struktur stop dengan frase stop yang dimulai pada kata pertama dalam kueri, dapat meningkatkan kinerja kueri[6].

Penelitian ini dilakukan untuk mengidentifikasi, meninjau, dan mengamati teknik dalam IR berdasarkan hasil beberapa penelitian sebelumnya khususnya pemrosesan kueri dan pemeringkatan hasil pencarian. Penelitian ini dibagi menjadi tiga bagian yang berbeda. Bagian pertama memberikan gambaran singkat tentang Information Retrieval (IR), IRS beserta komponennya dan tantangannya. Bagian selanjutnya menjelaskan proses pencarian informasi beserta tahapan pemrosesan teks dalam IR. Bagian terakhir menyajikan penelitian-penelitian terkait teknik untuk memproses kueri, dan algoritma pemeringkatan yang sudah pernah dilakukan dalam bidang IR.

2. METODOLOGI PENELITIAN

2.1 Metode Pengumpulan Data

Penelitian ini menggunakan metode tinjauan literatur untuk mengidentifikasi, meninjau, dan mengamati teknik dalam IR berdasarkan hasil beberapa penelitian sebelumnya. Tujuan dari penelitian ini adalah mengamati teknik yang pernah dilakukan oleh peneliti sebelumnya dalam IR khususnya pemrosesan kueri dan pemeringkatan hasil pencarian. Sumber literatur yang di akses berasal dari ACM Digital Library, ResearchGate, IEEE, IGI Global, MDPI, Semanticscholar, Springer dan Science Direct. ACM Digital Library mejadi rujukan utama yang kami lakukan dalam hal pengumpulan literatur tentang teknik IR khsusnya pemrosesan kueri dan pemeringkatan hasil pencarian.

2.2 Information Retrieval (IR)

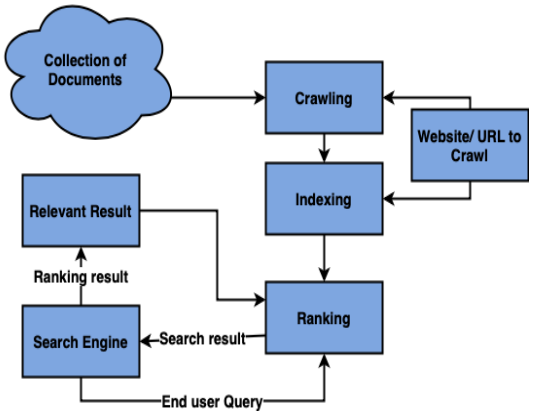
Information Retrieval (IR) adalah bidang ilmu yang cukup mendasar dan diminati oleh para peneliti di bidang ilmu komputer dan sistem informasi. Sekarang ini, pengguna yang membutuhkan informasi akan memanfaatkan mesin pencari untuk melakukan pencarian dan menemukan informasi yang dibutuhkan. Mesin pencari merupakan perangkat penelusur informasi yang diinginkan, dapat berupa pencarian alamat website, gambar, video, file ataupun topik-topik tertentu yang dinginkan pengguna.

Kata kunci pencarian yang seringkali pendek dan ambigu membuat mesin pencari sulit untuk mengindentifikasi maksud pengguna yang sebenarnya [8]. Mesin pencari yang cepat dan akurat dalam menemukan informasi yang relevan dan sesuai dengan keinginan pengguna akan lebih bermanfaat untuk mendukung kebutuhan informasi. Google adalah contoh paling terkenal dari sistem pencarian informasi yang telah digunakan semua orang. IR dalam hal ini berperan untuk menemukan kembali informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. Dengan kata lain, dapat disimpulkan bahwa IR adalah prosedur merepresentasikan, menyimpan, dan mencari kumpulan data besar untuk tujuan penggalian pengetahuan dan akses untuk menemukan hasil yang relevan yang memenuhi kebutuhan pengguna sebagai reaksi terhadap suatu kueri pengguna.

2.3 Komponen Information Retrieval

Tujuan utama IR adalah untuk menemukan informasi yang relevan atau dokumen yang memenuhi kebutuhan pengguna. Untuk mencapai tujuan tersebut [9], IRS menggunakan empat komponen utama, yaitu pengindeksan dokumen, pemrosesan kueri, proses pencarian dan pemeringkatan.

Journal of Information System Research (JOSH)
Volume 4, No. 3, April 2023, pp 748−754
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i3.2867



Gambar 1. Komponen Information Retrieval

Seperti gambar 1 diatas [1], komponen pertama adalah Crawling (perayapan). Web Crawler (perayap web) mencari dan mengambil dokumen untuk mesin pencari. Web Crawler memungkinkan untuk dibatasi pada situs tertentu sebagai titik awal pencarian website [10]. Teknik merepresentasikan dokumen merupakan komponen kedua yang biasa disebut dengan indexing yang berfungsi untuk membuat indeks keseluruhan dokumen (informasi). Selanjutnya proses representasi kueri. Pada fase ini, pengguna menginputkan kueri ke sistem untuk mencari informasi yang dibutuhkan. Kemudian, sistem mencari indeks untuk dokumen yang relevan dan terkait dengan kueri yang diberikan pengguna, dan menyajikan hasil pencarian dengan pemeringkatan berdasarkan relevansi kueri. Pengguna juga dapat memberikan umpan balik atas hasil pencarian kepada mesin pencari [9].

2.4 Information Retrieval System

IRS pertama kali dibuat untuk membantu pengelolaan data dalam jumlah besar. IRS menemukan informasi yang umumnya dalam bentuk dokumen dengan susunan data yang tidak terstruktur dan ditempatkan dalam media penyimpanan (database). IRS merupakan suatu sistem yang menemukan informasi yang sesuai dengan kebutuhan pengguna dari kumpulan informasi secara otomatis [11].

Saat ini, IRS digunakan dalam banyak bidang seperti perpusatakaan, pemerintahan, institusi pendidikan, perusahaan lain sebagainya. Aplikasi IRS meliputi pencarian media (file, gambar, video, kata, kalimat), mesin pencari (seperti google, bing), perpusatakaan digital dan aplikasi bidang tertentu seperti pencarian domain, letak geografis [9] hingga arsip surat administrasi pemerintahan [12]. Fungsi utama IRS adalah sebagai berikut [13]:

a. Mengidentifikasi sumber informasi yang relevan dengan kebutuhan pengguna
b. Menganalisis isi sumber informasi (dokumen)
c. Merepresentasikan isi sumber informasi dengan cara tertentu yang memungkinkan untuk dipertemukan dengan pertanyaan pengguna
d. Merepresentasikan kueri pengguna dengan cara tertentu yang memungkinkan untuk dipertemukan dengan informasi yang terdapat dalam basis data.
e. Mempertemukan pernyataan pencarian dengan data yang tersimpan dalam basis data
f. Menemukembalikan informasi yang relevan
g. Menyempurnakan unjuk kerja sistem berdasarkan umpan balik yang diberikan oleh pengguna

2.5 Tantangan dalam Information Retrieval

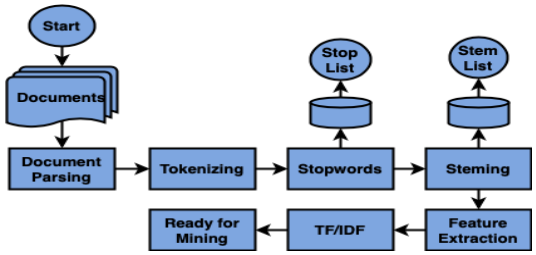
Ketidaksesuaian antara bagaimana pengguna menginputkan informasi yang dicari kedalam mesin pencari dan hasil yang disajikan oleh mesin pencari adalah tantangan utama dalam IRS. Sebagian besar IRS masih memberikan hasil yang tidak konsisten, tidak lengkap, tidak akurat, dan tidak relevan dengan informasi yang dibutuhkan pengguna. Kueri dengan kata kunci yang pendek yang terdiri dari satu atau dua kata akan menyulitkan mesin pencari untuk menyajikan hasil yang akurat. Berikut ini beberapa tantangan dalam IR [9]:

a. Representasi Konten
   Representasi dokumen dan kueri yang tidak memadai untuk diproses oleh IRS. Tantangan dalam hal ini adalah, bagaimana merepresentasikan konten sedemikian rupa sehingga dapat dibaca oleh IRS.
b. Pencocokan dan perangkingan
   Sejauh mana kesesuaian hasil pencarian dengan kueri kata kunci yang di inputkan oleh pengguna dan pemberian skor untuk perangkingan pada hasil pencarian.
c. Ambiguitas kata kunci
   Ini tantangan paling umum pada IR dan biasanya bersumber dari pengguna yang memberikan kata kunci yang ambigu dengan kata yang memiliki mana ganda.
d. Konteks Permintaan
   Penyajian informasi yang tidak akurat akibat keterbatasan kemampuan pengguna untuk menjelaskan informasi apa yang dibutuhkan.

Journal of Information System Research (JOSH)
Volume 4, No. 3, April 2023, pp 748−754
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i3.2867

2.6 Text Processing

Proses dalam IR meliputi preprocessing (pra-pemrosesan) dan pencocokan data. Pada tahapan preprocessing mencakup proses tokenization, stopwords removal, pendeteksian kalimat, stemming, lemmatisasi dan pembobotan [14].



Gambar 2. Tahapan pemrosesan teks dalam IR [1]

a. Document Parsing
Penguraian dokumen adalah proses mengidentifikasi isi dan bentuk dokumen teks yang ditulis dan disajikan dalam berbagai rangkaian karakter, bentuk dan bahasa. Penguraian dokumen memerlukan identifikasi dan pemisahan struktur dokumen menjadi komponen yang berbeda untuk membentuknya menjadi sebuah kesatuan [15].

b. Tokenizing
Tokenisasi atau analisis leksikal adalah operasi pembuatan kata dari rangkaian huruf (karakter) dalam suatu dokumen [1]. Setelah proses parsing, dilakukan analisis leksikal untuk memecah atau menandai dokumen yang dianggap sebagai aliran input menjadi kata, frase, atau simbol [15]. Salah satu masalah tokenisasi adalah konversi akronim dan singkatan ke dalam format standar. Kesulitan tokenisasi bervariasi tergantung pada bahasanya.

c. Stopwords/ filtering
Tujuan pada fase ini untuk menghapus semua istilah umum yang tidak penting dari proses tokenisasi. Pada tahap ini dilakukan penghapusan pengambilan kata-kata yang tidak berkontibusi banyak pada konten dokumen. Contoh stopword dalam bahasa Indonesia adalah “dan”, “yang”,”di” dan seterusnya. Pada proses ini, jika teks berisi kata sambung, spasi, kata depan, nama hari, nama bulan, nama tempat, serta tanda titik, koma, kurung, buka, kurung tutup, tanda tanya, kata tanya, tanda seru maka akan dihilangkan, sehingga akan menghasilkan kata - kata yang penting saja. Stopwords dapat dilakukan pada dokumen sebelum dilakukan indexing atau selama kueri melakukan proses eksekusi [16].

d. Stemming
Stemming adalah proses pengurangan sebuah kata menjadi kata dasar atau semantik. Stemming menjadi tahapan processing teks dasar yang sering digunakan untuk efisiensi dan efektifitas IR [16]. Identifikasi kata dari bentuk kata dasar umum meningkatkan sensitifitas dari hasil pencarian dengan meningkatkan kemampuannya menemukan dokumen yang relevan.

e. Ekstraksi Fitur
Teknik menghilangkan karakter yang tidak ingin digunakan dari dataset dikenal sebagai ekstraksi fitur. Saat menetapkan teks ke satu bagian atau lebih, akurasi akan meningkat dengan menggunakan teknik ekstraksi fitur. Ekstraksi fitur disarankan untuk klasifikasi dokumen. Ini membantu akurasi menjadi lebih baik, mengurangi dimensi, dan mengurangi waktu pemrosesan [17]. Algoritma ekstraksi fitur tergantung pada model ruang vektor, di mana kalimat direpresentasikan sebagai titik dalam ruang N-dimensi. Dimensi setiap titik menunjukkan aspek yang berbeda dari teks dalam bentuk digital. Salah satu metode untuk mengekstraksi fitur adalah dengan menggunakan skema pembobotan Term Frequency Inverse Document Frequency (TF- IDF). Premis dalam TF-IDF adalah istilah dalam dokumen yang dapat dibedakan menjadi dua kategori yaitu kata unik dan non-unik, terlepas dari apakah istilah tersebut penting untuk isi dokumen atau tidak [18]. Berikut adalah rumus perhitungannya [18]:

$$W_{i,j} = tf_{i,j} * \log\left(\frac{|N|}{d_{fj}}\right) \quad (1)$$

berdasarkan rumus (1) tersebut, **Wij** adalah bobot term (**tj**) terhadap dokumen (**di**). Sedangkan **tfij** adalah jumlah kemunculan term (**tj**) dalam dokumen (**di**). **d** adalah adalah jumlah semua dokumen yang ada dalam database dan **dfj** adalah jumlah dokumen yang mengandung term (**tj**) (minimal ada satu kata yaitu term (**tj**)).

f. Evaluasi
Pada dasarnya, ada dua kriteria berbeda untuk menilai kualitas IRS. Menurut [19], yang pertama adalah Precission. Kemampuan IRS untuk mengembalikan dokumen (informasi) terkait, serta keakuratan dan ketepatan dari dokumen yang diambil, seperti yang ditunjukkan oleh persamaan (2). Presisi merupakan proporsi dokumen yang dikembalikan yang benar-benar relevan dengan kueri yang diberikan.

Journal of Information System Research (JOSH)
Volume 4, No. 3, April 2023, pp 748−754
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i3.2867

$$\text{Precission} = \frac{|\text{relevant documents} \cap \text{retrived documents}|}{|\text{retrived documents}|} \quad (2)$$

Kriteria kedua adalah Recall, seperti yang ditunjukkan pada persamaan (3). Recall merupakan proporsi dokumen yang telah ditemukan dan terkait dengan kueri.

$$\text{Recall} = \frac{|\text{relevant documents} \cap \text{retrived documents}|}{|\text{relevant documents}|} \quad (3)$$

Kedua kriteria ini membantu menghitung metrik IR lainnya yaitu F dengan persamaan berikut [19]

$$\text{F} - \text{measure} = \frac{2 * \text{precission} * \text{recall}}{\text{precission} + \text{recall}} \quad (4)$$

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini melakukan studi pustaka dari penelitian terdahulu yang terkait dengan IR khususnya dalam bidang pemrosesan kueri dan perangkingan. Hasil penelitian yang dijabarkan disini (tabel 1) lebih fokus kedalam model yang digunakan pada IR dan terdiri dari 14 judul dalam rentang waktu tahun 2020 sampai dengan tahun 2022.

Tabel 1. Penelitian terdahulu beserta teknik yang ditawarkan

Metode	Judul Artikel	Tahun	Penulis
Dataset LOVBench	LOVBench: Ontology Ranking Benchmark	2020	Kolbe, et.all
DELTR	Reducing Disparate Exposure in Ranking:A Learning To Rank Approach	2020	Zehlike, et.all
RLPer	RLPer: A Reinforcement Learning Model for Personalized Search	2020	Yao, et.all
SetRank	SetRank: Learning a Permutation-Invariant Ranking Model for Information Retrieval	2020	Pang, et.all
Generating Clarifying Questions	Generating Clarifying Questions for Information Retrieval	2020	Zamani, et.all
Coeus	Coeus: A System for Oblivious Document Ranking and Retrieval	2021	Ahmad, et.all
Boosting algorithm Pseude-relevance feedback	Pseudo-relevance feedback based query expansion using boosting algorithm	2021	Rasheed, et.all
Tuned neural ranking models following cross- encoder	Improving Neural Models for the Retrieval of Relevant Passages to Geographical Queries	2021	Coelho, et.all
EXplanaTion RAnking (EXTRA)	EXTRA: Explanation Ranking Datasets for Explainable Recommendation	2021	
UQSCM-RFD	UQSCM-RFD: A query–knowledge interfacing approach for diversified query recommendation in semantic search based on river flow dynamics and dynamic user interaction	2022	Deepak, et.all
Framework Query performance prediction (QPP) model-agnostik	A Relative Information Gain-based Query Performance Prediction Framework with Generated Query Variants	2022	Datta, et.all
Efficient DAAT and SAAT	Eficient Document-at-a-Time and Score-at-a-Time uery Evaluation for Learned Sparse Representations	2022	Mackenzie, et.all
Deep-QPP	Deep-QPP: A Pairwise Interaction-based Deep Learning Model for Supervised Query Performance Prediction	2022	Datta, et.all
ColBERT-PRF	ColBERT-PRF: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval	2022	Wang, et.all

Peneltian tentang sistem pemberi rekomendasi atas umpan balik pencarian telah menjadi perhatian oleh banyak akademisi dan industri sehingga menghasilkan banyak model pada bidang tersebut. Mekanisme umpan balik Pseudo-relevance (relevansi semu) menunjukkan manfaat yang signifikan untuk memperluas kueri awal pengguna. ColBERT-PRF merupakan pengembangan dari Pseudo-relevance yang mengekstrak hasil embedding umpan balik yang representatif dari sekumpulan hasil pencarian yang kemudian ditambahkan ke representasi kueri asli [20].

Journal of Information System Research (JOSH)
Volume 4, No. 3, April 2023, pp 748−754
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i3.2867

EXplanaTion RAnking (EXTRA), salah satu model untuk memberikan penjelasan atas hasil rekomendasi dengan metrik berorientasi peringkat. EXTRA mengidentifikasi kalimat yang hampir identik dari kueri pengguna dengan kompleksitas waktu proses untuk memperkirakan kesamaan antara dua kalimat, selanjutnya mengkategorikan kalimat dalam kumpulan data secara efisien ke dalam kelompok yang berbeda [21].

Kata kunci pencarian seringkali pendek, hal ini menyulitkan mesin pencari untuk memprediksi maksud pengguna. Untuk mengatasi masalah ini, mesin pencari sering mendiversifikasi daftar hasil dan menyajikan dokumen yang relevan. Untuk mengurangi masalah ini [18], pertama-tama mengidentifikasi taksonomi klarifikasi untuk kueri pencarian dengan menganalisis data reformulasi kueri berskala besar yang diambil sampelnya dari log pencarian Bing. Taksonomi ini akan membawa ke sekumpulan template pertanyaan dan algoritma pengisian slot yang sederhana namun efektif. Kemudian, model ini secara otomatis menghasilkan pertanyaan klarifikasi untuk pelatihan. Didapatkan sebuah metode untuk menghasilkan kandidat jawaban pada setiap pertanyaan klarifikasi, sehingga pengguna dapat memilih dari serangkaian jawaban yang telah ditentukan sebelumnya.

SetRank salah satu model pemeringkatan dengan permutasi-invarian yang ditentukan pada kumpulan dokumen dengan berbagai ukuran. SetRank menggunakan tumpukan (induksi) blok sebagai komponen kuncinya untuk mempelajari dokumen yang diambil secara bersamaan. Hasil eksperimen pada tiga tolak ukur menunjukkan bahwa SetRank secara signifikan mengungguli baseline termasuk model learning-to-rank tradisional dan model Neural IR [22]. Deep-QPP model pemeringkat dengan neural query performance prediction (QPP) yang menggunakan pendekatan saraf tiruan memprediksi kinerja query. Deep-QPP tidak bergantung pada pemanfaatan informasi dari interaksi semantik atas hasil kueri yang ada pada peringkat teratas. Deep-QPP menggunakan arsitektur yang terdiri dari beberapa lapisan konvolusi 2D diikuti lapisan parameter feed-forward. [23]

Penelitian ini [24] menghasilkan peningkatan substansial dalam evaluasi kueri untuk mengatasi kekurangan pada pendekatan Document-at-a-time (DAAT) dan Score-at-a-time (SAAT). Dengan analisis empiris terperinci menunjukkan bahwa DAAT dan SAAT memiliki permasalahan pada pembatasan nilai pareto efektivitas dan efisiensi. Framework Query performance prediction (QPP) model-agnostik [25], mengumpulkan bukti tambahan dengan memanfaatkan informasi dari pola karakteristik distribusi Retrieval Status Values (RSVs) yang dihitung melalui serangkaian varian kueri yang dihasilkan secara otomatis. Framework ini dapat menghasilkan varian dengan cara yang terkontrol dengan mengganti istilah dari kueri asli dengan yang baru dan diambil sampelnya dari distribusi berbobot, baik melalui model relevansi atau dengan bantuan representasi kueri.

Selanjutnya, UQSCM-RFD [26], sebuah strategi semantik berbasis grafik untuk rekomendasi kueri. UQSCM-RFD bertujuan untuk membangun grafik dengan konsep Query Sense dengan pemilihan jalur konsep terbaik untuk mengintegrasikan pengetahuan dunia nyata dari wiki semantik. Pemeringkatan neural network dengan strategi cross-encoder [27] mengeksplorasi model neural network dalam konteks kueri geo-spasial dengan menganalisis strategi pemeringkatan ulang berdasarkan jarak geografis, selanjutnya melakukan penyempurnaan dengan strategi bi-encoder atau cross-encoder.

Boosting algorithm Pseude-relevance feedback [28] menggunakan pengujian teknik hibrid yang berbeda dengan metode perluasan kueri tradisional. Analisis kueri menunjukkan hasil yang lebih efektif dalam mengidentifikasi kata kunci yang paling relevan, bahkan untuk kueri pendek. Coeus [29], menilai relevansi dokumen menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF) yang banyak digunakan dalam Private Information Retrieval (PIR). Coeus meningkatkan latensi pada pengguna dengan mengurangi overhead PIR dengan memisahkan pengambilan metadata pribadi dari pengambilan dokumen, selanjutnya menskalakan produk matriks-vektor ke matriks TF-IDF.

RLPer [30], model personalisasi untuk melacak interaksi secara berurutan antara pengguna dan mesin pencari dengan Markov Decision Process (MDP). Mesin pencari dalam RLPer berinteraksi dengan pengguna untuk memperbarui model peringkat pencarian yang mendasarinya secara terus-menerus secara real-time. DELTR [31], framework berbasis learning-to-rank (LTR) yang mengeksplorasi hasil pemeringkatan dengan mengukur masalah dalam hal perbedaan rata-rata relevansi hasil pencarian. Kemudian yang terakhir, Dataset LOVBench [32] sebagai tolak ukur untuk pemeringkatan yang memungkinkan pembaruan tolak ukur secara terus-menerus. LOVBench mempertimbangkan dekomposisi fitur untuk meningkatkan performa peringkat secara signifikan.

4. KESIMPULAN

Dalam penelitian ini, dijabarkan secara mendasar tentang Information Retrieval (IR) dan Information Retrieval System (IRS) mulai dari komponen, fungsi utama IR dan text processing dalam IR. Kami juga membahas beberapa teknik yang digunakan dalam IR. Dalam penelitian selanjutnya, kami menyarankan sebaiknya dibandingkan lebih banyak lagi teknik dalam IR khususnya dalam optimasi dalam query processing agar kueri dengan kata kunci yang pendek (satu atau dua kata) dapat menyajikan hasil pencarian yang akurat dan relevan serta sesuai dengan kebutuhan pengguna.

REFERENCES

[1] S. Ibrihich, A. Oussous, O. Ibrihich and M. Esghir, “A Review on recent research in information retrieval,” in Procedia Computer Science, 2022.

Journal of Information System Research (JOSH)
Volume 4, No. 3, April 2023, pp 748−754
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i3.2867

[2] E. L. Terra and C. L. Clarke, "Frequency estimates for statistical word similarity measures.," in Proceedings of the human language technology conference of the North American Chapter of the Association for Computational Linguistics, 2003.
[3] G. Qian, S. Sural and S. Pramanik, "Similarity between Euclidean and cosine angle distance for nearest neighbor queries," in In Proceedings of the ACM symposium on Applied computing, 2004.
[4] S. E. Robertson, "The probability ranking principle in IR," Journal of documentation, vol. 33, no. 4, pp. 294-304., 1977.
[5] F. Diaz, "Pseudo-query reformulation," in European Conference on Information Retrieval, 2016.
[6] S. Huston and W. B. Croft, "Evaluating Verbose Query Processing Techniques," in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 2010.
[7] H. Zamani, S. Dumais and N. Craswell,, "Generating Clarifying Questions for Information Retrieval," in IW3C2 (International World Wide Web Conference Committee), Taipei, Taiwan, 2020.
[8] R. Kumar and S. Sharma, "Information Retrieval System: An Overview, Issues, and Challenges," International Journal of Technology Diffusion, vol. 9, no. 1, pp. 1-10, 2018.
[9] W. B. Croft, D. Metzler and T. Strohman, Search Engine Information Retrieval in Practice, New York: Pearson Education, Inc., 2015.
[10] F. Amin and P. Purwatiningtyas, "Rancang Bangun Information Retrieval System (IRS) Bahasa Jawa Ngoko pada Palintangan Penjebar Semangad dengan Metode Vector Space Model (VSM)," Jurnal Teknologi Informasi DINAMIK, vol. 20, no. 1, pp. 25-35, 2015.
[11] jdih.anri.go.id. [Online]. Available: https://jdih.anri.go.id. [Accessed 27 12 2022].
[12] G. Salton, Automatic Text Processing, The Transformation, Analysis, and Retrieval of information by computer., USA: Addison – Wesly Publishing Company, Inc, 1989.
[13] S. Silva, . A. S. Vieira, P. Celard, E. L. Iglesias and L. Borrajo, "A Query Expansion Method Using Multinomial Naive Bayes," MDPI: Applied Sciences, vol. 11, no. 21, pp. 1-14, 2021.
[14] S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali and S. Quarteroni, Web Information Retrieval (Data-Centric Systems and Applications), New York: Springer, 2013.
[15] S. Vijayarani, J. M. Ilamathi and B. Nithya, "Preprocessing techniques for textmining-anoverview," International Journal of Computer Science and Communication Networks, vol. 5, no. 1, pp. 7-16, 2014.
[16] S. Vidhya, A. Singh and E. Leavline, "Feature extraction for document classification," International Journal of Innovative Research in Science, Engineering and Technology, vol. 4, no. 6, pp. 50-56, 2015.
[17] R. Dzisevic and D. Šešok, "Text classification using different feature extraction approaches," Open Conference of Electrical, Electronic and Information Sciences (eStream), no. 10.1109/eStream.2019.8732167, pp. 1-4, 2019.
[18] B. Saini, V. Singh and S. Kumar, "Information retrieval models and searching methodologies: Survey," International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE), vol. 1, no. 2, p. 20, 2014.
[19] X. WANG, C. MACDONALD, N. TONELLOTTO and I. OUNIS, "ColBERT-PRF: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval," Association for Computing Machinery (ACM), 2022.
[20] L. Li, Y. Zhang and L. Chen, "EXTRA: Explanation Ranking Datasets for Explainable Recommendation," in Proceedings of the 44th Inter- national ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), Canada, 2021.
[21] L. Pang, J. Xu, Q. Ai, Y. Lan, X. Cheng and J. Wen, "SetRank: Learning a Permutation-Invariant Ranking Model for Information Retrieval," in Conference on Research and Development in Information Retrieval (SIGIR '20), New York, 2020.
[22] S. Datta, D. Ganguly, D. Greene and M. Mitra, "Deep-QPP: A Pairwise Interaction-based Deep Learning Model for Supervised Query Performance Prediction," in Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), Tempe, Arizona, USA, 2022.
[23] J. MACKENZIE, A. TROTMAN and J. LIN, "Eficient Document-at-a-Time and Score-at-a-Time uery Evaluation for Learned Sparse Representations," Association for Computing Machinery, 2022.
[24] S. DATTA, D. GANGULY, M. MITRA and D. GREENE, "A Relative Information Gain-based Query Performance Prediction Framework with Generated Query Variants," Association for Computing Machinery, 2022.
[25] G. Deepak and A. Santhanavijayan, "UQSCM-RFD: A query–knowledge interfacing approach for diversified query recommendation in semantic search based on river flow dynamics and dynamic user interaction," Neural Comput. Appl (Springer-Verlag), vol. 34, no. 1, p. 651–675, 2022.
[26] J. Coelho, J. Magalhães and B. Martins, "Improving Neural Models for the Retrieval of Relevant Passages to Geographical Queries," in Proceedings of the 29th International Conference on Advances in Geographic Information Systems, Association for Computing Machinery, Beijing, China, 2021.
[27] I. Rasheed, H. Banka and H. M. Khan, "Pseudo-relevance feedback based query expansion using boosting algorithm," Artificial Intelligence Review, Springer, vol. 54, no. 1, p. 6101–6124, 2021.
[28] I. Ahmad, S. Laboni, A. Divyakant, A. E. Amr and T. Gupta, "Coeus: A System for Oblivious Document Ranking and Retrieval," in Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles, German, 2021.
[29] J. Yao, Z. Dou, J. Xu and J.-R. Wen, "RLPer: A Reinforcement Learning Model for Personalized Search," in In Proceedings of The Web Conference 2020 (WWW '20), Taiwan, 2020.
[30] M. Zehlike and C. Castillo, "Reducing Disparate Exposure in Ranking: A Learning To Rank Approach," in Proceedings of The Web Conference 2020, Taipei, Taiwan, 2020.
[31] N. Kolbe, P.-Y. Vandenbussche, S. Kubler and Y. L. Traon, "LOVBench: Ontology Ranking Benchmark," in Proceedings of The Web Conference 2020, Taipei, Taiwan, 2020.
[32] R. C. Shiveric and R. B. Polloct, "Information Retrieval and Methods". United States of America Patent US9922078B2, 20 3 2018.