Jurnal SUTASOMA
(Science Teknologi Sosial Humaniora)

JURNAL SUTASOMA| ISSN 2964-8114 | E-ISSN 2964-8114
Vol.03 No.01 – Juni 2025 | https://s.id/Sutasoma
DOI : https://doi.org/10.58878/sutasoma.v3i2.392
Publishing : Universitas Tabanan

STUDY LITERATUR *INFORMATION RETRIEVAL* MODEL: TEKNIK DAN APLIKASI

I Gede Nyoman Agung Jayarana [1], I Gede Wira Darma[2], I Wayan Ady Juliantara[3], I Made Agus Widiana Putra[4]

[1]Institut Teknologi dan Bisnis STIKOM BALI
Denpasar, Indonesia

[2]Program Studi Teknik Komputer, Fakultas Teknik dan Perencanaan, Universitas Warmadewa
Denpasar, Indonesia

[34] Program Studi Sistem Informasi, Fakultas Sain dan Teknologi, Universitas Tabanan
Tabanan, Indonesia

agung_jayarana@stikom-bali.ac.id[1], igedewiradarma@warmadewa.ac.id[2],
adyjuliantara1@gmail.com[3], imadeagusclass@gmail.com[4]

Received: Juni, 2025	Accepted: Juni, 2025	Published: Juni, 2025

Abstrack

Information retrieval (IR) is a field in computer science that focuses on searching and retrieving relevant information from large-scale data sets. With the development of technology and the explosion of digital information, the need for efficient and accurate IR systems is increasing. This study aims to examine various IR models, both classical and modern, and the supporting techniques used in the information retrieval process. Classical models such as the Boolean Model, Vector Space Model (VSM), and BM25 are the initial foundations of IR systems, while modern neural network-based approaches such as DSSM, DPR, and Retrieval-Augmented Generation (RAG) offer more semantic and contextual search performance. In addition, basic techniques such as indexing and tokenization, as well as advanced techniques such as query expansion and relevance feedback, are also discussed, which also increase the effectiveness of the system. The performance evaluation of the IR system is carried out using various metrics such as precision, recall, F1-score, MAP, and NDCG. The results of this study indicate that the combination of the right IR models and techniques can produce an information retrieval system that is more relevant, efficient, and adaptive to the needs of modern users.

Keywords: *Information retrieval (IR), Indexing, IR Mode, Searching, Vector Space Model (VSM)*

Abstrak

Information retrieval (IR) merupakan bidang dalam ilmu komputer yang berfokus pada pencarian dan pengambilan informasi relevan dari kumpulan data dalam skala besar. Seiring perkembangan teknologi dan ledakan informasi digital, kebutuhan akan sistem IR yang efisien dan akurat semakin meningkat. Studi ini bertujuan untuk mengkaji berbagai model IR, baik klasik maupun modern, serta teknik-teknik pendukung yang digunakan dalam proses temu kembali informasi. Model klasik seperti *Boolean* Model, *Vector Space* Model (VSM), dan BM25 menjadi fondasi awal sistem IR, sedangkan pendekatan modern berbasis neural network seperti DSSM, DPR, dan *Retrieval-Augmented Generation* (RAG) menawarkan performa pencarian yang lebih semantik dan kontekstual. Selain itu, dibahas pula teknik dasar seperti *indexing* dan *tokenisasi*, serta teknik lanjutan seperti *query expansion* dan *relevance feedback*, yang turut meningkatkan efektivitas sistem. Evaluasi kinerja sistem IR dilakukan dengan berbagai metrik seperti *precision*, *recall*, F1-score, MAP, dan NDCG. Hasil

studi ini menunjukkan bahwa penggabungan model dan teknik IR yang tepat dapat menghasilkan sistem pencarian informasi yang lebih relevan, efisien, dan adaptif terhadap kebutuhan pengguna modern.

***Kata Kunci**: Information retrieval (IR), Indexing, IR Mode, Searching, Vector Space Model (VSM)*

1. PENDAHULUAN

1.1 Sejarah *Information Retrieval*

Pencarian informasi tidak dimulai dengan Internet. Baru pada dekade terakhir ini sistem *Information retrieval* (IR) ditemukan dalam aplikasi komersial. Sejarah *information retrieval* (IR) dimulai dengan metode pengatalogan manual yang digunakan di perpustakaan, seperti katalog kartu dan sistem klasifikasi hierarkis seperti Dewey Decimal (1876), dan pada tahun 1930-1940-an beberapa ilmuwan mulai berpikir tentang cara mekanis untuk mencari informasi di koleksi besar seperti Emanuel Goldberg yang menciptakan mesin "*Stastitical Machine*" untuk pencarian dokumen di film mikro. Dengan kemajuan teknologi, perangkat elektromekanis dan sistem kartu *punch* dikembangkan untuk mengotomatisasi pencarian. Pada tahun 1950-an, komputer mulai digunakan untuk IR, memungkinkan pengindeksan dan pengambilan dokumen secara otomatis (Azizah, 2022).

Dampak komputer dalam IR disorot ketika Hollywood menarik perhatian publik terhadap inovasi dengan film komedi *Desk Set*, yang keluar pada tahun 1957. Film ini berpusat pada sekelompok pustakawan referensi yang akan digantikan oleh komputer. IR sebagai disiplin penelitian mulai muncul pada masa ini dengan dua perkembangan penting, yaitu *Indexing* dan bagaimana cara mengembalikannya kembali. Sistem IR awal merepresentasikan dokumen dan query sebagai vektor, yang memungkinkan pengukuran kemiripan melalui teknik seperti *cosine similarity*, yang meningkatkan akurasi pencarian. Seiring berjalannya waktu, metode seperti stemming dan pemrosesan bahasa alami meningkatkan efektivitas sistem (Sanderson & Croft, 2016).

Tahun 1960-an salah satu tokoh utama yang muncul pada periode ini adalah Gerard Salton, yang membentuk dan memimpin kelompok IR yang besar, pertama di Harvard *University* (Cambridge, MA), dan kemudian di Cornell *University* (Ithaca, NY). Sanderson & Croft, (2016) menjelaskan kelompok ini mengembangkan *system* SMART (*System for the Mechanical Analysis and Retrieval of Text*) dan memunculkan model-model penting seperti *Boolean Moden* dan *Vector Space* Model (*Salton*).

Selanjutnya tahun 1970 salah satu perkembangan utama dari periode ini adalah Luhn's bobot term *frequency* (tf) *weight* (berdasarkan kemunculan kata dalam sebuah dokumen) dilengkapi dengan Karya Spa¨rck Jones tentang kemunculan kata di seluruh dokumen dalam sebuah koleksi. Pada tahun ini IR sudah mulai digunakan disistem *database* teks seperti bidang hukum, medis dan ilmiah. Berdasarkan perkembangan pada tahun 1970-an variasi dari tf idf skema pembobotan diproduksi (Salton dan Buckley). Pada pertengahan tahun 1980 sampai pertengahan 1990 dikembangkanlah proyek penting untuk evaluasi sistem IR yaitu TREC (*Text Retrieval Conference*) dengan di perkenalkannya model *probabilistic* seperti BM25 (Sanderson & Croft, 2016).

Munculnya internet menyebabkan munculnya mesin pencari web, ledakan World Wide Web yang diciptakan oleh Berners-Lee tahun 1990 dan diawali oleh Web Crawler pada tahun 1994 yang menggunakan pola klik pengguna dan *log query* untuk menyaring hasil. Munculnya mesin pencari seperti Archie, AltaVista dan Google yang memperkenalkan *Page Rank* yang mengintegrasikan *hyperlink* sebagai sinyal relevansi. Saat ini, IR terus berkembang dengan AI, pemahaman semantik, dan algoritme yang lebih canggih, yang bertujuan untuk pencarian informasi yang lebih cepat dan lebih relevan. Abad ke-20 dan awal abad ke-21 merupakan abad yang penuh perubahan dalam cara orang mengakses informasi (Sanderson & Croft, 2016).

Arti istilah *Information retrieval* bisa sangat luas. Hanya mengambil kartu kredit dari dompet Anda sehingga Anda dapat mengetikkan nomor kartu adalah salah satu bentuk pencarian informasi. Namun, sebagai sebuah bidang studi akademis, IR dapat didefinisikan menemukan materi (biasanya dokumen) yang tidak terstruktur (biasanya teks) yang memenuhi kebutuhan informasi dari dalam koleksi yang besar (biasanya disimpan di komputer) (HOWARD, 1946). *Information retrieval* secara umum dianggap sebagai sub bidang ilmu komputer yang berhubungan dengan representasi, penyimpanan, dan akses informasi. *Information retrieval* berkaitan dengan pengorganisasian dan pengambilan informasi dari koleksi basis data yang besar. *Information retrieval* (IR) adalah proses di mana kumpulan data direpresentasikan, disimpan, dan dicari untuk tujuan penemuan pengetahuan

sebagai respons terhadap permintaan pengguna (*query*). Proses ini melibatkan berbagai tahapan yang dimulai dengan merepresentasikan data dan diakhiri dengan mengembalikan informasi yang relevan kepada pengguna. Tahapan meliputi operasi penyaringan, pencarian, pencocokan, dan pemeringkatan. Tujuan utama dari sistem temu kembali informasi (IRS) adalah untuk "menemukan informasi yang relevan atau dokumen yang memenuhi kebutuhan informasi pengguna"(James & Kannan, 2017).

Sistem *Information retrieval* (IR) didasarkan, baik secara langsung maupun tidak langsung pada model-model proses temu kembali. Model-model pencarian ini menentukan bagaimana representasi dokumen teks dan kebutuhan informasi harus dibandingkan dalam untuk memperkirakan kemungkinan bahwa sebuah dokumen akan dinilai relevan. Model-model tersebut perkiraan relevansi dokumen dengan *query* yang diberikan adalah dasar untuk peringkat dokumen yang sekarang menjadi bagian yang tidak asing lagi dalam sistem IR (Belkin, 2017).

1.2 Pentingnya *Information Retrieval* Saat ini

Information retrieval (IR) memiliki peranan yang sangat krusial di era digital karena kemampuannya dalam menemukan, mengorganisasi, dan menyediakan informasi yang tepat, relevan, dan cepat dari lautan data yang sangat besar. Berikut beberapa poin penting yang menggambarkan signifikansi IR dalam berbagai aspek kehidupan dan teknologi modern:

1. Menemukan Informasi Secara Cepat dan Efisien
   IR memungkinkan pencarian informasi relevan dari jutaan hingga miliaran dokumen dalam waktu singkat, sehingga menghemat waktu dan tenaga. Contohnya adalah mesin pencari seperti Google *Search* yang menggunakan teknik IR canggih untuk mencocokkan permintaan pengguna dengan halaman web relevan secara *real-time* (Xie et al., 2021).
2. Mendukung Pengambilan Keputusan
   Dalam sektor kritikal seperti kedokteran, hukum, dan bisnis, IR menyediakan akses ke informasi berkualitas tinggi yang mendukung keputusan strategis. Contohnya, dokter dapat dengan cepat menemukan penelitian medis terbaru, dan pengacara dapat mengakses dokumen hukum penting sehingga proses pengambilan keputusan menjadi lebih tepat dan berinformasi (Liu et al., 2023).
3. Akses ke Pengetahuan dan Pendidikan
   IR memperkuat akses terbuka ke sumber belajar dan literatur akademik melalui platform seperti *Semantic Scholar* dan Europe PMC, memudahkan mahasiswa dan peneliti dalam menemukan materi pembelajaran dan penelitian terbaru (Wang et al., 2022).
4. E-Commerce dan Rekomendasi Produk
   Industri *e-commerce* sangat bergantung pada IR untuk memungkinkan pengguna menemukan produk yang mereka inginkan dengan kata kunci, kategori, atau riwayat pencarian sebelumnya. Contohnya di platform besar seperti Amazon dan Tokopedia, yang juga memanfaatkan IR untuk sistem rekomendasi produk yang personal dan relevan (Zhou et al., 2020).
5. Asisten Virtual dan Pencarian Suara
   IR menjadi pondasi dari teknologi asisten virtual seperti Siri, Alexa, dan Google *Assistant*. Sistem ini membutuhkan IR untuk memproses dan memahami pertanyaan suara serta memberikan jawaban yang akurat dan cepat, menjadikan interaksi pengguna lebih alami dan efisien (Zhou et al., 2020).
6. Penegakan Hukum dan Keamanan
   Dalam konteks hukum dan keamanan, IR digunakan dalam e-discovery dan investigasi digital untuk mencari dan menganalisis bukti dari kumpulan data besar secara efektif, membantu penegak hukum mengungkap fakta dengan cepat (Xie et al., 2021).
7. Pencarian Multimedia (Gambar, Video, Audio)
   IR telah berkembang tidak hanya untuk teks melainkan juga konten multimedia. Fitur pencarian gambar seperti Google *Image Search* dan video di YouTube adalah contoh aplikasi IR yang memungkinkan pencarian berbasis konten multimedia (Xie et al., 2021).
8. Riset Ilmiah dan Big Data
   IR mendukung analisis dan pencarian informasi dalam *dataset* besar yang ada di riset ilmiah. Peneliti dapat dengan cepat menyaring publikasi, *dataset*, dan literatur relevan untuk mendukung penelitian mereka, mempercepat kemajuan ilmiah (Xie et al., 2021).
9. Optimalisasi Sistem AI dan NLP
   IR diintegrasikan dalam kecerdasan buatan dan pemrosesan bahasa alami (NLP), seperti pada model *Retrieval-Augmented Generation* (RAG) yang digunakan dalam *chatbot* dan asisten digital. Ini meningkatkan kemampuan AI dalam menyediakan jawaban yang relevan dan kontekstual (Lewis et al., 2020).
10. Personalisasi Konten
    IR memungkinkan penyajian konten yang dipersonalisasi berdasarkan preferensi dan riwayat interaksi pengguna. Platform seperti Netflix atau Spotify menggunakan IR untuk merekomendasikan film, musik, atau konten lain yang sesuai dengan minat individu.

2. MODEL *INFORMATION RETRIEVAL*

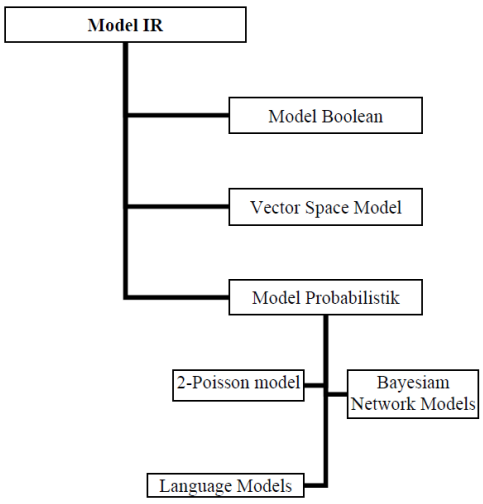
Evolusi model IR menunjukkan perkembangan dari pendekatan tradisional yang menggunakan metode statistik dan *lexical* hingga model canggih berbasis neural *network* dan integrasi dengan teknologi *generative* seperti RAG.

2.1 Model Tradisional (Statistik/Lexical)

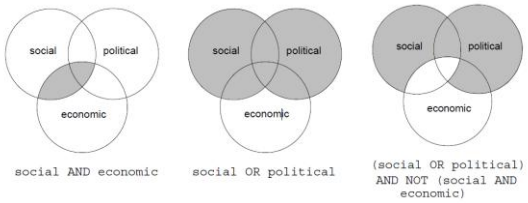
Model-model tradisional merupakan dasar dari system IR yang sudah ada sejak awal perkembangan IR.

1. Boolean Model

   Model ini menggunakan logika Boolean dengan operator AND, OR, dan NOT untuk mencocokkan *query* dengan dokumen. Sistem ini mudah dipahami dan diimplementasikan, tetapi kurang fleksibel karena hasil pencarian bersifat biner (relevan atau tidak relevan). Untuk penggunaan operator logika tersebut, akan dijelaskan pada diagram Venn berikut:



Gambar 1. Usulan Taksonomi untuk Permodel *Information Retrieval*



Gambar 2. Kombinasi Operator Logika menggunakan Diagram Venn

Banyak web semantik yang memakai model Boolean di mana dokumen dijadikan kumpulan variabel berhubungan dengan klause WHERE pada query SPARQL, dan model boolean mengembalikan semua query yang memuaskan. Beberapa langkah yang dapat dilakukan dalam proses *Booelan Retrieval* ini antara lain:

a. Lakukan *indexing*, dalam hal ini *inverted index*.
b. Temukan kata/term query di dalam kamus dan daftar *posting*.
c. Lakukan operasi dari operator logika yang diinginkan dengan mencari irisan dari *posting list*.

Model Boolean tidak melakukan pengambilan peringkat, namun, model ini masih dijadikan pilihan utama untuk *search engine* dengan memasukkan tambahan seperti operator kedekatan panjang. Sebuah operator kedekatan adalah cara menentukan bahwa dua istilah dalam *query* harus terjadi dekat satu sama lain dalam dokumen, di mana kedekatan dapat diukur dengan membatasi jumlah kata intervensi yang diizinkan atau dengan mengacu pada unit struktural seperti kalimat atau paragraf.

2. *Vector Space Model* (VSM)

   Model ini merepresentasikan dokumen dan *query* sebagai vektor dalam ruang multi-dimensi, biasanya menggunakan bobot TF-IDF (*Term Frequency-Inverse Document Frequency*). Kemiripan dihitung menggunakan metrik *cosine similarity* sehingga memberikan hasil pencarian yang lebih gradien berdasarkan skor relevansi. Pada VSM setiap term i, di dalam dokumen maupun *query*, j, diberikan suatu bobot (*weight*) yang bernilai real. Dokumen dan query diekspresikan sebagai vektor t-dimensi dan diasumsikan terdapat n dokumen di dalam *database*. Selain itu pada VSM, *database* dari semua dokumen direpresentasikan oleh *matrik term-document* (atau *term frequency*), di mana setiap sel pada matriks berkorespondensi dengan bobot yang diberikan dari suatu term dalam dokumen yang ditentukan. Nilai nol berarti bahwa *term idak* terdapat dalam dokumen. Pada Gambar 3 diperlihatkan *matrik term document* dengan n dokumen dan t term.

	T_1	T_2	T_3	$T_{...}$	T_t
D_1	W_{11}	W_{21}	W_{31}	$\square_{...}$	T_{t1}
D_2	W_{12}	W_{22}	W_{32}	$\square_{...}$	T_{t2}
D_3	W_{13}	W_{23}	W_{33}	$\square_{...}$	T_{t3}
$D_{...}$	$\square_{...}$	$\square_{...}$	$\square_{...}$	$\square_{...}$	$\square_{\square}$
D_n	W_{1n}	W_{2n}	W_{3n}	$\square_{...}$	T_{tn}

Gambar 3. *Matrik term-document*

Vector Space Model meranking dokumen berdasarkan kemiripan *vector-space* antara

vektor *query* dan vektor dokumen. Ada banyak cara untuk mengkomputasi kesamaan dari dua vektor tersebut, salah satunya dengan *inner product* atau kesamaan *cosine.*

3. *Probabilistic* Model (BM25, *Language* Model)
   Model ini mengurutkan dokumen dalam urutan menurun terhadap peluang relevansi sebuah dokumen pada informasi yang dibutuhkan Dalam model probabilistik (peluang), IR tergantung pada dua komponen utama yaitu sekumpulan dokumen yang diidentifikasi sebagai *record* yang relevan dan yang tidak relevan. Adapun Karakteristik model probabilistik adalah sebagai berikut:
   a. Melakukan pendugaan *page* relevansi dengan menggunakan probabilistik
   b. Mempunyai *teoritical framework* yang jelas, yaitu berdasarkan prinsip statistik, relevansi dokumen dapat diupdate, adanya feed back/timbal balik dari *user*.

   Ide dasarnya yaitu *query* dapat menghasilkan jawaban yang benar, menggunakan indeks term, menggunakan pendugaan awal, menggunakan *initial* hasil, dan *feed back* dari *user* dapat memperbaiki probabilitas dari relevansi.

2.2 Neural IR Models

Information retrieval (IR) berbasis *neural network* telah menjadi pendekatan dominan dalam sistem pencarian modern karena kemampuannya dalam memahami semantik dan menangkap hubungan kompleks antar kata. Pendekatan ini melibatkan transformasi dokumen dan *query* ke dalam bentuk vektor berdimensi tinggi yang dihasilkan oleh model pembelajaran mendalam. Berikut dijelaskaan tiga model neural IR yang populer digunakan:

1. *Deep Structured Semantic Model* (DSSM)
   DSSM diperkenalkan oleh Huang et al. (2015) sebagai pendekatan awal untuk merepresentasikan query dan dokumen dalam ruang vektor semantik menggunakan jaringan neural. Tujuan utama DSSM adalah untuk memetakan pasangan query-dokumen ke dalam representasi berdimensi rendah di mana kemiripan semantik diukur dengan fungsi kesamaan (*similarity function*), seperti *cosine similarity*:

   $$\text{sim}(q, d) = \frac{v_q \cdot v_d}{||v_q|| ||v_d||}$$

   Di mana v_q dan v_d masing-masing adalah vektor representasi *query* dan dokumen. Model ini menggunakan beberapa lapisan *fully connected* untuk mengekstrak fitur semantik dari representasi *input*, yang umumnya berasal dari *character-level n-grams*. DSSM mampu mengatasi keterbatasan pencocokan kata kunci dengan memanfaatkan representasi semantik. Namun, karena arsitekturnya yang sederhana, model ini kurang efektif dalam menangani konteks panjang dan ketergantungan antar kata yang kompleks.
2. *Dense Passage Retrieval* (DPR)
   Dense Passage Retrieval (DPR), yang dikembangkan oleh Karpukhin et al. (2020), merupakan sistem IR berbasis *dense representation*. DPR menggunakan dua *encoder* independen yang masing-masing mengubah query q dan dokumen d menjadi vektor embedding, umumnya dengan model transformer seperti BERT:

   $$v_q = \text{Encoder}_q(q), \quad v_d = \text{Encoder}_d(d)$$

   Kemiripan antara keduanya dihitung menggunakan *dot product*. Dalam implementasinya, DPR memungkinkan *pre-computing* dokumen karena dokumen *di-encode* satu kali dan disimpan dalam indeks vektor. Proses pencarian menjadi efisien karena hanya melibatkan komputasi vektor *query* dan pencocokan vektor. Keunggulan DPR adalah skalabilitas dan efisiensi tinggi, namun karena arsitekturnya berupa *bi-encoder*, interaksi langsung antara kata dalam *query* dan dokumen tidak diperhitungkan.
3. *Bi-Encoder* dan *Cross-Encoder*
   Arsitektur *bi-encoder* dan *cross-encoder* merupakan dua strategi utama dalam desain sistem IR berbasis transformer.
   a. *Bi-Encoder*: Query dan dokumen diproses oleh *encoder* secara terpisah untuk menghasilkan *embedding*:

   $$v_q = \text{Encoder}(q), \quad v_d = \text{Encoder}(d)$$

   Kemudian dihitung skor kesamaan antar vektor dengan *dot product* atau *cosine similarity*. Model ini efisien untuk *retrieval* skala besar karena dokumen bisa diproses terlebih dahulu. Namun, *bi-encoder* memiliki keterbatasan dalam menangkap hubungan semantik mendalam antar kata
   b. *Cross-Encoder*: *Query* dan dokumen digabung dalam satu *input* dan diproses bersamaan oleh model:

   $$\text{Input} = [\text{CLS}]\, q\, [\text{SEP}]\, d\, [\text{SEP}]$$

   Model memproses seluruh *token* sekaligus dan menghasilkan skor relevansi sebagai *output* akhir. Karena mempertimbangkan interaksi penuh antar *token*, *cross-encoder* mampu memberikan prediksi relevansi yang lebih akurat, tetapi dengan biaya komputasi yang jauh lebih tinggi.

2.3 *Retrieval-Augmented Generation* (RAG)

Retrieval-Augmented Generation (RAG) merupakan pendekatan generatif terbaru dalam *Information retrieval* (IR) yang menggabungkan dua komponen utama: *retrieval* dan *text generation*. Pendekatan ini bertujuan untuk meningkatkan kualitas jawaban atau teks yang dihasilkan dengan mengakses informasi eksternal yang relevan dari basis pengetahuan atau korpus dokumen. Model RAG sangat relevan dalam pengembangan sistem *question answering*, asisten cerdas, dan sistem dialog berbasis pengetahuan seperti ChatGPT yang terhubung dengan modul *retrieval*. *Retrieval-Augmented Generation* (RAG) merupakan pendekatan dalam NLP yang menggabungkan model *retrieval* dan generatif untuk meningkatkan akurasi dan keluasan pengetahuan dalam menghasilkan teks. Dengan memanfaatkan informasi dari dokumen eksternal, RAG mampu mengurangi kesalahan faktual (*hallucination*) dan memperluas jangkauan informasi yang dapat digunakan model.

1. Sistem *Question Answering* (QA)
   RAG sangat efektif dalam *sistem open-domain* QA karena mampu mengambil informasi dari koleksi dokumen besar sebelum menghasilkan jawaban. Hal ini menjadikan sistem lebih faktual, responsif, dan fleksibel terhadap pertanyaan baru.
2. Asisten Virtual dan *Chatbot*
   Dalam asisten cerdas, RAG memungkinkan jawaban yang berbasis dokumen eksternal seperti PDF, web, atau basis data organisasi. Ini berguna untuk tugas seperti penjelasan kebijakan, membaca regulasi, atau memberikan jawaban berbasis data terkini.
3. Mesin Pencari Generatif
   RAG menjadi fondasi dalam mesin pencari modern, yang tidak hanya mencari, tetapi juga merangkum informasi menjadi satu jawaban naratif berbasis referensi nyata.
4. Ringkasan Dokumen dan Laporan
   Model ini juga digunakan untuk membuat ringkasan dari berbagai dokumen panjang, seperti laporan penelitian, putusan hukum, atau kebijakan publik, dengan merujuk langsung ke sumber data yang diambil secara dinamis.

3. TEKNIK DALAM *INFORMATION RETRIEVAL*

Information retrieval (IR) merupakan bidang yang terus berkembang, mulai dari teknik dasar berbasis *keyword* hingga teknik lanjutan berbasis pembelajaran mesin dan representasi semantik. Teknik-teknik dalam IR bertujuan untuk meningkatkan kualitas pencarian dan relevansi dokumen terhadap kebutuhan informasi pengguna.

3.1 Teknik Dasar Klasik

1. Indexing
   Indexing adalah proses representasi dokumen dalam struktur data yang efisien, seperti *inverted index*, yang memetakan term ke daftar dokumen yang mengandung term tersebut (Azizah, 2022).
2. Tokenisasi
   Tokenisasi adalah tahap awal preprocessing teks, di mana teks dipecah menjadi unit kata (*token*). Proses ini penting untuk memastikan bahwa setiap kata dikenali dan diproses dalam pencocokan *query* dan *indexing* (Manning, Raghavan, & Schutze, 2018).
3. Term *Weighting*
   Term *weighting* mengukur pentingnya sebuah term dalam dokumen dan seluruh koleksi dokumen. Skema yang paling umum adalah TF-IDF (*Term Frequency-Inverse Document Frequency*):

$$\text{TF-IDF}(t,d) = tf(t,d) \cdot \log\left(\frac{N}{df(t)}\right)$$

4. Boolean Retrieval
   Model Boolean menggunakan operator logika seperti AND, OR, dan NOT untuk pencocokan dokumen. Pendekatan ini bersifat biner: dokumen dianggap relevan atau tidak, tanpa urutan prioritas (Mitra, B., & Craswell, N. 2017).
5. Probabilistic Model – BM25
   BM25 (*Best Matching* 25) merupakan model probabilistik yang memperhitungkan frekuensi term dan panjang dokumen. Skor BM25 untuk dokumen d dan *query* q didefinisikan sebagai:

$$\text{Score}(d,q) = \sum_{t \in q} IDF(t) \cdot \frac{f(t,d) \cdot (k_1 + 1)}{f(t,d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{\text{avgdl}})}$$

   Dengan $f(t,d)$ adalah frekuensi term t, dan parameter k_1 serta b digunakan untuk mengatur sensitivitas terhadap panjang dokumen3. BM25 adalah salah satu metode retrieval klasik yang paling efektif dan masih digunakan secara luas.

3.2 Teknik Lanjutan

1. *Query Expansion*
   Query Expansion memperluas *query* awal dengan sinonim atau istilah terkait untuk meningkatkan recall. Pendekatan ini dapat menggunakan sumber leksikal seperti *WordNet* atau *embedding* semantik (Voorhees, 2015).
2. *Relevance Feedback*
   Relevance feedback memungkinkan sistem memperbaiki hasil pencarian berdasarkan dokumen yang dinilai relevan atau tidak oleh

pengguna. Salah satu metode klasik adalah *Rocchio Algorithm* yang memperbarui vektor *query* berdasarkan dokumen relevan dan non-relevan (Abass, O. A., & Arowolo, O. A. 2017).

3. *Reranking*
   Reranking adalah proses menyusun ulang hasil pencarian awal berdasarkan fitur tambahan atau model lanjutan seperti *Learning to Rank* atau *Cross-Encoder*. Teknik ini digunakan untuk meningkatkan presisi pada tahap akhir pencarian.
4. *Embedding-Based Retrieval*
   Metode ini menggunakan representasi vektor berdimensi tinggi yang dihasilkan dari model neural seperti BERT atau DPR, sehingga dapat menangkap makna semantik lebih baik daripada pencocokan kata kunci (Karpukhin et al., 2020). Pencocokan dilakukan dengan menghitung kesamaan vektor menggunakan *cosine similarity* atau *dot product*.

4. EVALUASI *INFORMATION RETRIEVAL*

Evaluasi dalam *Information retrieval* (IR) digunakan untuk mengukur seberapa baik sistem pencarian menemukan dan menyajikan informasi yang relevan bagi pengguna. Evaluasi ini menjadi landasan penting dalam pengembangan dan perbandingan berbagai model IR (Manning, Raghavan, & Schütze, 2018). Berikut ini adalah metrik evaluasi yang umum digunakan:

1. *Precision* dan *Recall*
   a. *Precision* mengukur akurasi hasil pencarian, yaitu proporsi dokumen yang diambil dan benar-benar relevan.
   $$\text{Precision} = \frac{\text{Relevant Retrieved}}{\text{Total Retrieved}}$$
   b. *Recall* mengukur kelengkapan, yaitu proporsi dokumen relevan yang berhasil ditemukan dari keseluruhan dokumen relevan.
   $$\text{Recall} = \frac{\text{Relevant Retrieved}}{\text{Total Relevant}}$$
   Precision dan *recall* merupakan dasar evaluasi sistem IR tradisional maupun modern (Baeza-Yates & Ribeiro-Neto, 1999).
2. F1-score
   F1-score adalah rata-rata harmonik dari *precision* dan *recall*, digunakan untuk menyeimbangkan keduanya ketika keduanya dianggap penting. F1 memberikan skor tunggal dalam evaluasi:
   $$\text{F1} = 2 . \frac{\text{Precision} . \text{Recall}}{\text{Precision+Recall}}$$
   F1-score sering digunakan dalam konteks klasifikasi atau *retrieval* biner (Manning et al., 2018)
3. MAP (*Mean Average Precision*)
   MAP mengukur kualitas ranking dari seluruh hasil pencarian. Metrik ini menghitung rata-rata *precision* pada posisi setiap dokumen relevan yang ditemukan, kemudian dirata-ratakan untuk semua *query*. MAP sangat efektif dalam mengevaluasi *retrieval multi-query*.
4. NDCG (*Normalized Discounted Cumulative Gain*)
   NDCG mempertimbangkan peringkat posisi dokumen dalam hasil pencarian. Dokumen relevan yang muncul lebih awal dalam daftar hasil dianggap lebih bernilai. NDCG banyak digunakan dalam sistem berbasis ranking seperti pencarian web atau rekomendasi (Jarvelin & Kekalainen, 2022).
5. Recall@K
   Recall@K mengukur berapa banyak dokumen relevan yang berada di antara K hasil teratas. Metrik ini populer dalam *deep retrieval* dan *recommendation systems* di mana pengguna hanya melihat sebagian kecil hasil:
   $$\text{Recall@K} = \frac{\text{Dokumen Relevan dalam Top}-\text{K}}{\text{Total Dokumen Relevan}}$$
   Recall@K sangat relevan dalam model *retrieval* berbasis *embedding* (Karpukhin et al., 2020).

5. KESIMPULAN

Berdasarkan studi literatur yang telah dilakukan, dapat disimpulkan bahwa *information retrieval* (IR) merupakan bidang yang terus berkembang dan sangat penting dalam era digital saat ini. Sistem IR memungkinkan pencarian dan pengambilan informasi yang relevan dari kumpulan data besar, baik dalam bentuk teks maupun multimedia, dengan tujuan utama memenuhi kebutuhan informasi pengguna secara efisien dan akurat. Model-model IR telah berevolusi dari pendekatan klasik seperti Boolean Model, *Vector Space* Model (VSM), dan *Probabilistic* Model (BM25) yang berbasis pada statistik dan logika, menuju pendekatan modern berbasis pembelajaran mendalam seperti *Deep Structured Semantic Model* (DSSM), *Dense Passage Retrieval* (DPR), dan *Retrieval-Augmented Generation* (RAG). Model neural ini memungkinkan sistem IR menangkap makna semantik yang lebih dalam dan memberikan hasil yang lebih kontekstual serta adaptif.

Selain model, keberhasilan sistem IR juga dipengaruhi oleh teknik pendukung seperti *indexing*, *tokenisasi*, *term weighting*, *relevance feedback*, *query expansion*, *reranking*, serta *embedding-based retrieval*. Teknik evaluasi seperti *Precision*, *Recall*, F1-score, MAP, NDCG, dan *Recall@K* sangat penting untuk mengukur efektivitas performa sistem IR dalam menjawab

query pengguna. Secara keseluruhan, IR tidak hanya mendukung pengambilan informasi dalam mesin pencari umum, tetapi juga berperan dalam pengambilan keputusan di bidang medis, hukum, e-*commerce*, riset ilmiah, keamanan digital, dan pengembangan kecerdasan buatan. Kemajuan dalam teknologi IR akan terus mendorong kemampuan sistem untuk menghadirkan informasi yang lebih cepat, akurat, dan sesuai konteks di berbagai sektor kehidupan.

PERNYATAAN PENGHARGAAN

Puji dan syukur Penulis panjatkan kepada Tuhan Yang Maha Esa, karena atas berkat dan rahmatnya Penulis dapat menyelesaikan penelitian ini. Penulis menyadari bahwa tanpa bantuan dan informasi dari berbagai pihak, sangat sulit bagi Penulis untuk menyelesaikan penelitian ini, oleh sebab itu pada kesempatan ini Penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada semua pihak yang tidak dapat Penulis sebutkan satu-per satu atas bantuan secara langsung maupun tidak langsung dalam penulisan penelitian ini.

Akhir kata Penulis mengucapkan terima kasih kepada semua pihak yang telah membantu dalam penulisan penelitian ini dan Penulis berharap semoga penelitian ini dapat bermanfaat bagi kita semua dan menjadi bahan masukan dalam dunia pendidikan ke depannya.

DAFTAR PUSTAKA

Abass, O. A., & Arowolo, O. A. (2017). *Information retrieval models, techniques and applications. International Research Journal of Advanced Engineering and Science*, 2(2), 197–202.

Allan, J. (Ed.). (2015). *Topic Detection and Tracking: Event-based Information Organization*. Springer. https://doi.org/10.1007/978-1-4615-1335-6

Azizah, E. N., & Handayani, A. N. (2022). *Permodelan pada Information Retrieval: Literature review*. Jurnal Inovasi Teknik dan Edukasi Teknologi, 2(11), 527–535. https://doi.org/10.17977/um068v2i112022p527-535

Belkin, N. J., & Croft, W. B. (2017). *Information filtering and information retrieval: Two sides of the same coin?* Communications of the ACM, 35(12), 29–38.

Guo, J., Fan, Y., Ai, Q., & Croft, W. B. (2016). A *deep relevance matching model for ad-hoc retrieval. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management* (pp. 55–64). https://doi.org/10.1145/2983323.2983769

Huang, P. S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2015). *Learning deep structured semantic models for web search using clickthrough data. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM)* (pp. 2333–2338). https://doi.org/10.1145/2505515.2505665

James, N. T., & Kannan, R. (2017). *A survey on information retrieval models, techniques and applications. International Journal of Advanced Research in Computer Science and Software Engineering*, 7(7), 16–22. https://doi.org/10.23956/ijarcsse.v7i7.90

Jarvelin, K., & Kekalainen, J. (2022). *Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS)*, 20(4), 422–446. https://doi.org/10.1145/582415.582418

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). *Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://arxiv.org/abs/2004.04906

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuksa, P., Yih, W. T., *Rocktäschel, T., & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2005.11401

Liu, Y., Li, X., Zhang, H., & Wu, J. (2023). *Smart health information retrieval: A review. Journal of Biomedical Informatics*, 139, 104296. https://doi.org/10.1016/j.jbi.2023.104296

Manning, C. D. (2022). *The deep learning tsunami and the future of information retrieval*. Information Retrieval Journal, 25(1–2), 1–8. https://doi.org/10.1007/s10791-021-09406-y

Manning, C. D., Raghavan, P., & Schütze, H. (2018). Introduction to *Information retrieval*. Cambridge University Press.

Mitra, B., & Craswell, N. (2017). *Neural models for information retrieval (arXiv:1705.01509)*. arXiv. https://arxiv.org/abs/1705.01509

Mitra, B., & Craswell, N. (2018). *An introduction to neural information retrieval. Foundations and Trends in Information Retrieva*l, 13(1), 1–126.

Ndengabaganizi, T. J., & Kannan, R. (2017). *A survey on information retrieval models, techniques and applications*. International Journal of Advanced Research in Computer Science and Software Engineering, 7(7), [Artikel 0112]. https://doi.org/10.23956/ijarcsse/V7I7/0112

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: *Sentence Embeddings using Siamese BERT-Networks*. EMNLP. https://arxiv.org/abs/1908.10084

Roshdi, A., & Roohparvar, A. (2015). Review: *Information retrieval techniques and applications*. International Journal of Computer Networks and Communications Security, 3(9), 373–377. http://www.ijcncs.org

Sanderson, M., & Croft, W. B. (2015). The history of *information retrieval* research. *Proceedings of the IEEE*, *100*(SPL CONTENT), 1444–1451. https://doi.org/10.1109/JPROC.2012.2189916

Voorhees, E. M. (2015). *Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing & Management*, 36(5), 697–716. https://doi.org/10.1016/S0306-4573(00)00010-8

Wang, L. L., Lo, K., & Kohlmeier, S. (2022). COVID-19 literature search: Semantic Scholar's answer to the pandemic. *Patterns*, *3*(2).

Xie, Q., Wang, Y., & Zhang, J. (2021). Deep learning for multimedia *information retrieval*: A survey. *IEEE Transactions on Multimedia*, *23*, 1483–1499.

Zhou, T., Liu, W., & Song, J. (2020). *A deep learning framework for e-commerce product retrieval and recommendation*. *IEEE Access*, *8*, 193218–193227.