Jurnal Inovasi Teknik dan Edukasi Teknologi, 2(11), 2022, 527-535
ISSN: 2797-7196 (online)
DOI: 10.17977/um068v2i112022p527-535

Permodelan pada Information Retrieval: Literature Review

Erwina Nurul Azizah, Anik Nur Handayani*
Universitas Negeri Malang, Jl. Semarang No. 5 Malang, Jawa Timur, Indonesia
*Penulis korespondensi, Surel: aniknur.ft@um.ac.id

Paper received: 06-11-2022; revised: 15-11-2022; accepted: 29-11-2022

Abstract
Information Retrieval (IR) is a technique for finding information stored in relevant sources according to user needs. There are various ways to use IR, but this paper focuses on modeling which is used as a framework for information retrieval. There are three types of IR models developed in this paper. IR modeling techniques will be explained as proposed in the literature with a detailed description of the modes, such as Boolean and Region models. Also included are the advantages and disadvantages of each model.

Keywords: information retrieval; boolean models; regional models; vector space approach

Abstrak
Information Retrieval (IR) adalah teknik untuk menemukan sebuah informasi yang tersimpan pada sumber yang relevan sesuai dengan kebutuhan pengguna. Ada berbagai cara memanfaatkan IR, namun pada paper ini difokuskan pada permodelan yang dipakai sebagai kerangka dalam pengambilan informasi. Ada tiga jenis model IR yang dikembangkan pada paper ini. Teknik permodelan IR akan dijelaskan sesuai yang diusulkan di literatur dengan penjabaran terperinci tentang mode-model tersebut, seperti model Boolean dan Region. Disertakan juga keunggulan dan kelemahan masing-masing model.

Kata kunci: information retrieval; model boolean; model region; pendekatan ruang vektor

1. Pendahuluan

Jika dilihat dalam kehidupan sehari-hari, manusia tak pernah lebah dari berinteraksi dengan sebuah tulisan dalam berbagai macam dan berbagai kebutuhan pula. Membaca majalah, menonton televisi atau pergi ke toko buku; dengan tujuan untuk mengetahui tentang dunia, mendapatkan hiburan, atau untuk belajar. Semua kegiatan itu tentu saja memerlukan sebuah interaksi dengan tulisan. Yang dimaksud dengan interaksi adalah manusia bukan menjadi pihak pasif yang hanya menerima informasi, namun juga aktif mencari sekaligus memberikan informasi[1].

Teknologi masa kini memberikan banyak data baru yang tersimpan pada berbagai tempat. Macam-macam informasi digitalpun mulai beragam, mulai dari informasi kependudukan sampai informasi seputar kegemaran dan keunggulan seseorang. Informasi-informasi tersebut tentulah berguna untuk suatu hal, terutama pada sebuah halaman web. Dengan berkembangnya halaman web seperti Blog, wiki, dan aplikasi seperti RSS, Tag, dan seterusnya semenjak tahun 2004, pengguna telah menjadi pusat dari penghasil dan pemakai informasi, pengguna juga memiliki lebih banyak media untuk menyalurkan informasi. Perkembangan ini tak hanya mengubah perkembangan industri jaringan, tetapi juga sangat berdampak pada metode pengumpulan data karena memberikan tuntutan baru ke depannya. [2]

Banyak tipe informasi yang bisa didapat dari sebuah halaman web. Walau begitu menulis informasi mengenai Web adalah tantangan yang cukup sulit, terutama untuk Internet-based IR yang menjadi subjek yang paling luas untuk dipelajari dan selalu menjadi topik hangat

perdebatan. Karena kenyataannya informasi pada internet tak bisa dikatakan tepat dan penuh kejujuran [3]. Di sisi lain pandangan standar pada IR adalah membandingkan informasi yang didapat pada user [1] sehingga bisa menghasilkan IR yang efektif [4].

Banyak sekali jumlah text, audio, video, dan dokumen lain yang tersedia di internet, dan hampir mencangkup semua topik.Pengguna membutuhkan kemampuan untuk menemukan informasi yang tepat dan memuaskan kebutuhan mereka. Terdapat dua cara mencari informasi di internet: menggunakan mesin pencari atau menelusuri direktori sesuai dengan kategori. Walau begitu masih banyak bagian interned yang tak bisa diakses dengan bebas, contohnya database [5].

Dari semua itu, ada berbagai teknik yang bisa dipakai untuk mengumpulkan informasi dari halaman web.Hal ini juga yang dilakukan mesin pencari seperti Google, sebagai contoh Google Schoolar memberikan informasi yang diinginkan user melalui tema spesifik yaitu literatur pendidikan [6]. Banyak teknik yang dikembangkan, semua tergantung pada media apa yang kita inginkan informasinya. Untuk media halaman web sendiri bisa meggunakan beberapa metode pendekatan, seperti basis CBR (Case –based reasoning) yang memgkombinasikan penyelesaian masalah dengan yang lainnya [7][8]. Ada juga Ontology Mapping yang mencari elemen yang sama pada ontology lain, yang sangat penting untuk mencapai probabilitas semantik dari World Wide Web (WWW) [9] . Namun ini adalah contoh penggunaan lebih lanjut, pada dasarnya semua metode di atas didasari oleh model IR.

Pada paper ini, tujuan yang diinginkan adalah memberikan review keadaan permodelan dalam IR dan memberikan penjelasan kelebihan serta kelemahan dari permodelan tersebut. Selebihnya pada paper ini dikategorikan menjadi beberapa bab. Bab kedua menjelaskan lebih lanjut tentang permodelan yang ada. Pada bab ketiga akan dibahas perbandingan masing-masing permodelan. Bab kelima adalah konklusi.

1.2. Model Pada Information Retrieval

Ada dua alasan kenapa lebih baik memiliki sebuah model dalam IR. Pertama, model bekerja sebagai penunjuk arah dan memberikan nilai tengah pada diskusi akademik. Dan yang kedua adalah model bisa dianggap sebagai kerangka awal untuk pengaplikasian sistem retrieval yang sebenarnya [10]. Hampir seluruh sistem IR yang ada bergantung pada keputusan data yang hanya berada pada query dan dokumen; informasi seputar user dan konteks pencarian sebagian besar di acuhkan [11].

Seperti yang sudah dikatakan sebelumnya, ada beberapa model IR yang akan dibahas di literatur ini. Untuk awalnya dijelaskan pada bagan Taksonomi di dalam Fig. 1, dimana permodelan IR terdapat 6 jenis [10].

1.3. Model Boolean

Model Boolean adalah model pertama dan sekaligus paling banyak dikritik [10][12]. Model Boolean merupakan IR sederhana yang berdasarkan atas teori aljabar Boolean. Untuk singkatnya, kita dapat menganggap query adalah kumpulan dokumen yang jelas [13]. Boolean memiliki 3 operator logika AND, OR, dan NOT. Di dalam struktur data, Boolean merupakan sebuah tipe data yang bernilai “True” atau “False” (benar atau salah) [14]. Sehingga didalam IR, logika boolean diartikan bahwa data yang di telusuri sesuai atau tidak antara variable –

variablenya. Untuk penggunaan operator logika tersebut, akan dijelaskan pada diagram Venn yang ada pada Fig. 2.

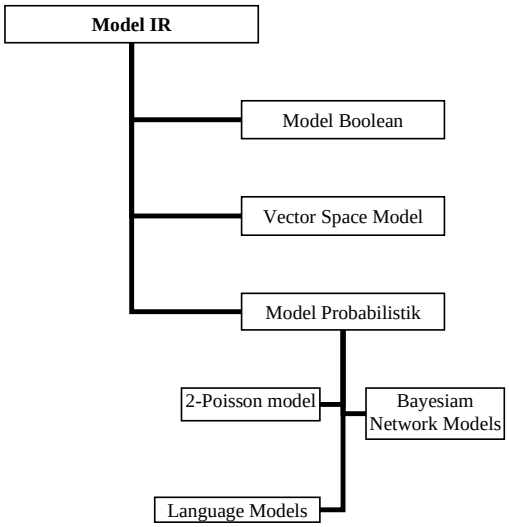

Fig. 1. Usulan Taksonomi untuk permodel Information Retrieval

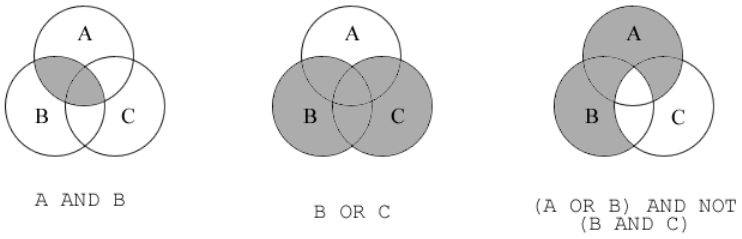

Fig. 2. Kombinasi Operator Logika menggunakan diagram Venn.

Banyak web semantik yang memakai model Boolean [15] dimana dokumen dijadikan kumpulan variabel berhubungan dengan klause WHERE pada query SPARQL, dan model boolean mengembalikan semua query yang memuaskan [16]. Beberapa langkah yang dapat dilakukan dalam proses Booelan Retrieval ini antara lain:

1. Lakukan indexing, dalam hal ini inverted index.
2. Temukan kata/term query didalam kamus dan daftar posting.
3. 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, dimana kedekatan dapat diukur dengan membatasi jumlah kata intervensi yang diizinkan atau dengan mengacu pada unit struktural seperti kalimat atau paragraf .

1.4. Vector Space Model

Model Vector Space digunakan untuk menyatukan entitas pada seluruh dokumen adalah standar yang banyak digunakan pada IR [17]. Vector Space Model (VSM) mempresentasikan setiap dokumen yang terdapat dalam database dan query ke dalam vektor multidimensi. Dimensi dari vektor berkorespondensi dengan jumlah setiap term dalam database dan kumpulan term tersebut membentuk suatu ruang vektor [18].

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). Dimana 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 [19] Fig 3 menunjukkan 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}

Fig. 4. 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 [20].

1.5. Model Probabilistik

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 [21][10][22].

Adapun Karakteristik model probabilistik adalah sebagai berikut:

1. Melakukan pendugaan page relevansi dengan menggunakan probabilistik
2. Mempunyai teoritical framework yang jelas, yaitu berdasarkan prinsip statistik, relevansi dokumen dapat diupdate, adanya feed back/timbal balik dari user.
3. 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.

Adapun beberapa jenis dari permodelan probabilistik antara lain.

1.6. 2-Poisson Model

Bookstein dan Swanson (1974) mempelajari masalah membangun aturan statistic untuk tujuan mengindentifikasi index term pada dokumen. Mereka menyarankan urutan kejadian tf pada dokumen dapat di modelkan dengan campuran 2 distribusi Poisson seperti di bawah ini, dimana X adalah variabel acak untuk urutan kejadian.

Permodelan ini mengasumsikan bahwa dokumen dibuat oleh arus acak dari kejadian. Setiap masa, kumpulan dokumen tersebut dapat dibagi menjadi dua subsets [23].

Apabila dokumen diambil acak pada subset pertama, maka hubungan probabilitas dianggap sama, atau lebih tinggi dari hubungan di subses kedua; karena hubungan probabilitas dianggap saling berkolekasi dengan sejauh mana subjek mengacu dari waktu dipulihkan [24].

1.7. Model Bayesian Network

Bayesian Network berasal dari teorema Bayes, sebuah pendekatan untuk sebuah ketidakpastian yang diukur dengan probabilitas [25]. Teorema ini dikemukakan oleh Thomas Bayes, namun untuk istilah "jaringan Bayesian" diciptakan oleh Yudea Pearl pada tahun 1985 untuk menekankan pada tiga aspek yaitu:

1. Sifat sering subjektif dari informasi masukan (input).
2. Ketergantungan pada pengkondisian Bayes sebagai dasar untuk memperbarui informasi.
3. Perbedaan antara kausal dan mode bukti penalaran.

Bayesian network adalah graph asiklik (tak memiliki direktori path A -> -> Z seperti A = Z) yang meng-enkode hubungan ketergantungan probabilitas antar variabel acak. Pada percobaannya, model ini dibagi menjadi 4 layer; dokumen, representasi, query dan informasi yang dibutuhkan. Figure 5 menunjukkan versi sederhana dari model tersebut, semua node mewakili variabel binary acak dengan isian {0,1} [26][27].

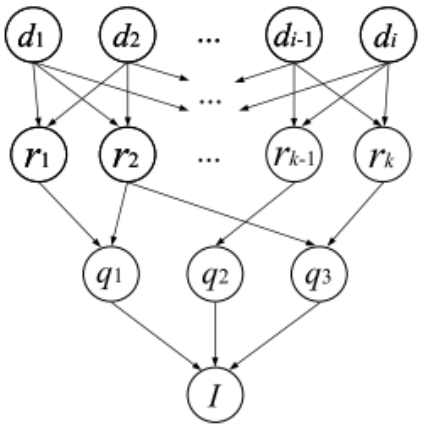

Fig. 5. Versi sederhana tampilan network

Sesuai dengan bagan di atas, maka bisa disimpulkan bahwa pribabilitas node r2, q1, q3 dan I adalah :

P(r2,q1,q3,I) = P(r2)P(q1|r2)P(q3|r2,q1)P(I|r2,q1,q3)

1.8. Language Models

Language modeling (LM) memberikan pendekatan dalam bentuk seperti novel dalam mencari problem di pencarian teks, dimana sembali menghubungkan dengan banyak penelitian baru pada proses percakapan dan bahasa. Sebagai aplikasi statistik, LM memang merupakan penerapan dari teori-teori Markov yang antara lain juga sudah dipakai oleh Zipf dalam bibliometrika dan Shannon yang berupaya menerapkan teori informasi dalam penggunaan bahasa manusia. Kini LM dipakai untuk pengenal bahasa lisan (automatic speech recognition). Sejak 1980, LM menjadi komponen penting dalam penerjemahan otomatis (machine traslation) dan pelacakan kesalahan eja (error spelling). Bahkan kemudian juga dipakai untuk mengembangkan perangkat lunak pengolah bahasa alamiah (natural language processing task), dan pembuatan ringkasan teks otomatis (summarization). Di penghujung era1990an teori dan aplikasi LM diperkenalkan ke bidang information retrieval (IR) dan kini menjadi salah satu cabang penting penelitian di bidang ini. [28].

Dalam bentuk rumus matematika, LM mengasumsikan S sebagai kata-kata (words) yang beruntaian:

$$S = W_1, W_2, \ldots, W_3$$

Untuk sejumlah k kata-kata, maka S mencerminkan Markov process dengan hitungan probabilitas:

$$P_n(S) = \prod_{i=1}^{k} P\left(w_1 \middle| w_{i-1}, w_{i-2}, w_{i-3}, \ldots, w_{i-n+1}\right)$$

Ketika n = 2, kita mengatakannya sebagai biagram language model, yang kemudian dapat diestimasi menggunakan informasi tentang keberadaan-bersama (co-occurance) pasangan kata-kata. Jika n = 1 maka kita menamakannya unigram language model, yang menggunakan hanya probabilitas dari kata-kata secara sendiri-sendiri (individual). Dalam bidang penelitian IR, orang lebih banyak menggunakan unigram model karena urut-urutan kata tidak terlalu dipermasalahkan, tidak seperti dalam pengenal suara otomatis (speech recognition) yang sangat bergantung kepada kemampuan mesin memahami urutan kata-kata [29].

Salah satu model IR yang menggunakan LM adalah Query-Likelihood Model yang pertama diusulkan oleh Ponte dan Croft (1998). Dalam model ini diasumsikan bahwa para pemakai sistem sudah memiliki gambaran yang cukup tentang istilah-istilah yang akan ada di dokumen "ideal" yang akan memenuhi kebutuhan informasi mereka. Lalu, diasumsikan pula bahwa istilah yang digunakan untuk mencari dokumen itu (atau biasa disebut query) dapat memisahkan yang "ideal" dari yang tidak.

Dalam artikelnya, Ponte dan Croft [29] menggunakan Bernoulli multivariat untuk menghitung P(Q|D). Mereka menganggap sebuah query sebagai sebuah vektor dari atribut biner, masing-masing atribut untuk sebuah istilah yang unik di dalam kosakata indeks, dan menandakan ada-tidaknya istilah tersebut di dalam query. Jumlah kemunculan istilah tersebut

di dalam query sendiri tidaklah diperhitungkan. Ada dua asumsi yang mendasari model ini, yaitu:

1. Semua atribut bernilai biner. Jika sebuah istilah ada di query, maka atribut yang mewakili istilah tersebut bernilai 1. Jika tidak, bernilai 0.
2. Istilah dianggap tidak berkaitan (independen) di dalam sebuah dokumen. Asumsi ini mirip dengan asumsi yang digunakan dalam teori-teori IR probabilistrik.

Berdasarkan dua asumsi di atas, maka query likelihood P(Q|D) dapat dirumuskan sebagai hasil dari dua probabilitas, yaitu probabilitas kemunculan istilah pada query dan probabilitas ketidak-munculan istilah itu. Atau dalam rumus formal:

P(w|D) dihitung dengan metode non-parametrik yang memanfaatkan probabilias rata-rata dari w (words, kata-kata) di dalam dokumen yang mengandungnya. Untuk istilah-istilah yang tidak muncul, maka probabilitas umum di dalam koleksi lah yang digunakan. Juga perlu diketahui bahwa statistik tentang koleksi, seperti frekuensi kemunculan istilah (term frequency) dan frekuensi dokumen merupakan bagian integral dari LM, walaupun tidak digunakan secara menyeluruh/heuristik seperi halnya di dalam teori-teori probabilitas [30].

Model	Kelebihan	Kekurangan
Model Boolean	Model sederhana yang basisnya hanyalah 3 operasi pencarian. Jadi mudah diimplementasikan.[10] Karena masih mendasar, masih bisa diperluas atau digabungkan dengan model lain.[12] Memberikan User kontrol penuh pada sistem. [12]	Tak ada ranking pada dokumen. Sulit dalam mengambil keputusan dari dokumen yang didapat [12]. *Query* bisa jadi sangat kompleks tergantung syarat pencariannya [13].
Vector Space Model	Model yang simpel karena basisnya dari aljabar linear [17]. Berat term bukan *binary*. Memungkinkan komputing untuk tingkat berkelanjutan antara *query* dan dokumen [20].	Kesulitan dalam hal mencari sinonim dan *polysery* [18]. Secara teoritis mengasumsikan bahwa *term* adalah independent secara statistik[20].
2-Poisson Model	Tidak memerlukan penambahan waktu untuk menanggung beban algoritma yang dimasukkan [23]. Lebih memahani konsep IR dan menginspirasi para peneliti untuk membuat model yang lebih baru [25].	Masalah paling besar berada pada estimasi dari parameter [24]. Dari setiap waktu pengerjaan, terdapat 3 parameter misterius yang tidak bisa diestimasi secara langsung dari data [26]. Model yang terlalu kompleks.
Bayesian Network	Bayesian sudah sangat umum ditemukan di bidang ilmu komputer. Jadi akan lebih banyak menemukan "bantuan" untuk solusi IR [27]. Sangat cocok untuk data kecil dan belum sempurna. Karena Bayesian Network tak memerlukan jumlah minimum dokumen.[27]	Membutuhkan sampel dalam jumlah banyak karena node yang begitu spesifik [26][10]. .
Lamguage Model	Language modeling memberikan pendekatan baru pada scoring dokumen dengan *query* dan lebih meringankan beban yang dipakai [28].	Resiko ketidaksesuaian bahasa antara dokumen dan *query* masih tinggi [30]. Berlaku independen.

Model	Kelebihan	Kekurangan
	Hasil dari IR berkonsep simpel, bersifat matematis, komputasi yang dilakukan terurut dan intuitif menarik [28].	

2. Simpulan

Information Retrieval mulai menjadi peranan penting seiring berkembangnya jaman, manusia terus menciptakan dan meminta informasi karena itu Information Retrieval sangat dibutuhkan, walau begitu metode information retrieval sebaiknya menggunakan permodelan dasar. Ada dua alasan kenapa lebih baik memiliki sebuah model dalam IR. Pertama, model bekerja sebagai penunjuk arah dan memberikan nilai tengah pada diskusi akademik. Dan yang kedua adalah model bisa dianggap sebagai kerangka awal untuk pengaplikasian sistem retrieval yang sebenarnya.Ada berbagai macam model Information Retrieval mulai dari yang paling mudah dengan pendekatan Boolean, dan yang paling rumit dan baru Language Model.

Daftar Rujukan

Belkin, N. J. (1993). Interaction with texts: Information retrieval as information seeking behavior. *Information retrieval*, *93*(55-66).

Zhang, Z., & Tang, J. (2007). Information Retrieval in Web2. 0. In *Integration and Innovation Orient to E-Society Volume 1: Seventh IFIP International Conference on e-Business, e-Services, and e-Society (I3E2007), October 10–12, Wuhan, China* (pp. 663-670). Springer US.

Kobayashi, M., & Takeda, K. (2000). Information retrieval on the web. *ACM computing surveys (CSUR)*, *32*(2), 144-173.

Kobayashi, M., & Takeda, K. (2000). Information retrieval on the web. *ACM computing surveys (CSUR)*, *32*(2), 144-173.

Mazur, Z., & Wiklak, K. (2013). MSALSA–A Method of Positioning Search Results in Music Information Retrieval Systems. In *Multimedia and Internet Systems: Theory and Practice* (pp. 221-228). Springer Berlin Heidelberg.

Giustini, D., & Boulos, M. N. K. (2013). Google Scholar is not enough to be used alone for systematic reviews. *Online journal of public health informatics*, *5*(2)..

Duan, Y., Burrell, P., Mullins, R., & Jin, H. (2005). A Case-Based Reasoning Approach to Enhance Web-Based Training on Internet Marketing. In *Artificial Intelligence Applications and Innovations: IFIP TC12 WG12. 5-Second IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI2005), September 7–9, 2005, Beijing, China 2* (pp. 557-566). Springer US.

Ceausu, V., & Despres, S. (2007). A semantic case-based reasoning framework for text categorization. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings* (pp. 736-749). Springer Berlin Heidelberg.

Mao, M. (2007). Ontology mapping: An information retrieval and interactive activation network-based approach. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings* (pp. 931-935). Springer Berlin Heidelberg.

Azizah, E. N., & Handayani, A. N. (2022). Permodelan pada Information Retrieval: Literature Review. *Jurnal Inovasi Teknologi dan Edukasi Teknik (JITET)*, *2*(11)..

Aknouche, R., Asfari, O., Bentayeb, F., & Boussaid, O. (2012). Integrating query context and user context in an information retrieval model based on expanded language modeling. In *Multidisciplinary Research and Practice for Information Systems: IFIP WG 8.4, 8.9/TC 5 International Cross-Domain Conference and Workshop on Availability, Reliability, and Security, CD-ARES 2012, Prague, Czech Republic, August 20-24, 2012. Proceedings 7* (pp. 244-258). Springer Berlin Heidelberg.

HO LEE, J. O. O. N., HO KIM, M. Y. O. U. N. G., & JOON LEE, Y. O. O. N. (1993). Information retrieval based on conceptual distance in IS-A hierarchies. *Journal of documentation*, *49*(2), 188-207.

Van Opijnen, M., & Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, *25*, 65-87.

Cadoli, M., Giovanardi, A., & Schaerf, M. (1998). An algorithm to evaluate quantified Boolean formulae. *AAAI/IAAI*, *98*, 262-267.

Michalowski, M., Ambite, J. L., Thakkar, S., Tuchinda, R., Knoblock, C. A., & Minton, S. (2004). Retrieving and semantically integrating heterogeneous data from the web. *IEEE Intelligent Systems*, *19*(3), 72-79.

Gronski, J. (2009). Semantic web for search. In *The Semantic Web-ISWC 2009: 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings 8* (pp. 957-964). Springer Berlin Heidelberg.

Bagga, A., & Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Abdillah, A. A., Muktyas, I. B., Matematika, P., & Surya, S. (2013). Implementasi vector space model untuk pencarian dokumen. In *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika* (Vol. 2013, pp. 1-7).

Erk, K., & Padó, S. (2008, October). A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 897-906).

Mole, "Vector Space Model," 1999.

Croft, W. B., & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevant information. *Journal of documentation*.

Lafferty, J., & Zhai, C. (2003). Probabilistic relevance models based on document and query generation. *Language modeling for information retrieval*, 1-10.

V. Lavrenko, "Introduction to Probabilistic Models for Information Retrieval," Proc. 33rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. Ed. by Hsin-His Chen, Efthismis N. Efthimiadis, Jacques Savoy, Fabio Crestani Lugano Stephane Marehand-Maillet, p. p. 905, 2010.

Gehler, P. V., Holub, A. D., & Welling, M. (2006, June). The rate adapting poisson model for information retrieval and object recognition. In *Proceedings of the 23rd international conference on Machine learning* (pp. 337-344).