Information Retrieval Pada *Frequently Asked Questions* (FAQ) dengan metode *String Similarity*

Information Retrieval on Frequently Asked Questions (FAQ) using String Similarity method

Gede Herdian Setiawan[1], I Made Budi Adnyana[2]
[1,2]Fakulas Informatika dan Komputer, Program Studi Sistem Komputer,
Institut Teknologi dan Bisnis STIKOM Bali, Denpasar, Indonesia
E-mail: [1]herdian@stikom-bali.ac.id, [2]budi.adnyana@stikom-bali.ac.id

Abstrak

Information retrieval merupakan sebuah sarana untuk menemukan informasi berdasarkan kumpulan informasi pada data terstruktur maupun tidak terstruktur secara otomatis. implementasi information retrival seperti mesin pencari menggunakan query dari pengguna dengan bahasa alami manusia kemudian sistem dapat menemukan dokumen atau informasi yang berkaitan dengan query dari pengguna. Pada penelitian ini di usulkan sistem information retrieval pada Frequently Asked Questions atau FAQ dengan mencari pertanyaan yang mirip (similar) pada daftar pertanyaan di basis data terhadap pertanyaan yang diberikan oleh pengguna menggunakan algoritma Cosine similarity untuk mencari kesamaan kosinus tertinggi. Selanjutnya memberikan respon jawaban yang sebelum nya sudah di berikan label terhadap pertanyaan yang relevan dan memiliki similaritas paling tinggi. Telah dihasilkan dataset FAQ dan dilakukan preprocessing, penerapan algoritma Cosine Similarity terhadap input pertanyaan (query) dengan dataset dan menghasilkan bobot pada setiap pertanyaan (label) pada dataset. Melalui evaluasi akurasi pemberian bobot similaritas yang dilakukan dengan memberikan dua belas input pertanyaan dibagai pada empat kategori berdasarkan tingkat kemiripan memiliki akurasi mencapai 75% terkait kemampuan memberikan bobot sesuai dengan tingkat similaritas pertanyaan (query) dengan dataset pertanyaan pada FAQ.

Kata kunci: Information Retrieval, Frequently Asked Questions, Cosine Similarity

Abstract

Information retrieval is a means to find information based on a collection of information on structured and unstructured data automatically. implementation of information retrieval such as search engines using queries from users with natural human language then the system can find documents or information related to queries from users. In this study, an information retrieval system is proposed on Frequently Asked Questions or FAQs by looking for similar questions (similar) in the list of questions in the database to questions given by users using the Cosine similarity algorithm to find the highest cosine similarity. Next, provide answers that have previously been labeled with relevant questions and have the highest similarity. The FAQ dataset has been generated and preprocessing has been carried out, applying the Cosine Similarity algorithm to the input questions (queries) with the dataset and generating weights for each question (label) in the dataset. Through the assessment, presenting the weight of similarity which is carried out by providing twelve input questions in various categories based on the assessment having an assessment of having an accuracy of up to 75% related to ability according to the level of the question (query) with the question dataset in the FAQ.

Keywords: Information Retrieval, Frequently Asked Questions, Cosine Similarity

1. PENDAHULUAN

Frequently Asked Questions atau FAQ adalah sebuah tulisan atau kumpulan informasi yang dirancang untuk memberikan jawaban terhadap pertanyaan yang umumnya berkiatan dengan penggunakan suatu produk, layanan dan palform/aplikasi [1]. Kumpulan pertanyaan dan jawaban pada FAQ umumnya disediakan pada sebuah halaman yang dapat di akses oleh pengguna. Pengguna secara aktif mencari pertanyaan yang sesuai dengan kondisi dan menemukan jawaban nya, hal ini akan menjadi kendala apabila daftar pertanyaan cukup banyak sehingga membutuhkan waktu untuk mencari pertanyaan yang relevan dengan kebutuhan [2]. Untuk mengatasi hal ini diperlukan sebuah sistem yang dapat menyediakan *information retrieval* berdasarkan input dari pengguna [3].

Information retrieval merupakan sebuah sarana untuk menemukan informasi berdasarkan kumpulan informasi pada data terstruktur maupun tidak terstruktur secara otomatis [4]. implementasi *information retrival* seperti mesin mencari menggunakan *query* dari pengguna dengan bahasa alami manusia kemudian sistem dapat menemukan dokumen atau informasi yang berkaitan dengan *query* dari pengguna [5]. Pada sistem *information retrieval* selain menerima *input* dari pengguna yang disebut dengan *query*, dilakuan proses menentukan dokumen yang paling mirip (similar) dan selanjutnya memberikan respon ke pengguna[6][7]

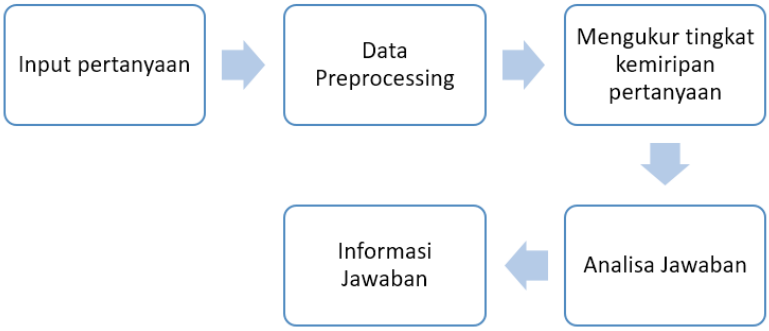
Kunci dari *information ratrival* terletak pada bagaimana menemukan kemiripan / similaritas *query* dari pengguna dengan dokumen atau informasi pada basis data. Penerapan metode *string similarity* untuk menentukan kemiripan teks maupun dokumen telah banyak dilakukan untuk berbagai kebutuhan pengembangan sistem seperti yang dilakukan [8] menerapkan *string similarity* Dengan algoritma *cosine similarity* untuk sistem *question answering*, metode yang digunakan adalah mencari dokumen paling relevan berdasarkan pertanyaan yang diberikan dan memperoleh jawaban berdasarkan *rules* untuk setiap pertanyaan, hasil nya metode yang diusulkan memperoleh akurasi sebesar 40%. Penelitian lain dilakukan oleh [9] klasifikasi buku berdasarkan sinopsis dengan melalukan pembobotan TF/IDF dan *cosine similarity* pada model VSM untuk mengukur tingkat kemiripan dokumen, hasil nya sistem mampu memberikan 10 kategori berdasarkan 120 dokumen dengan nilai *precision* sebesar 90,91%. Penerapan *string similarity* pada *information retrival* pernah dilakukan oleh [3] menggunakan metode VSM untuk pembobotan tiap dokumen berbahasa jawa ngoko yang ada pada basis data sehingga antar dokumen memiliki bobot yang berbeda untuk menentukan dokumen mana yang paling mirip (similar) dengan *query*, hasil nya sistem mampu melakukan pencarian dokumen disertai bobot setiap dokumen dan letak dokumen. *Information retrieval* untuk menentukan kemiripan dokumen juga dilakukan oleh [10], proses *information retrieval* pada dokumen dengan memperhatikan urutan kata (*sequential of word*) dalam kalimat, hasil nya *information retrival* dapat menjaga kata – kata yang berurutan dalam kalimat secara semantic. Sampai saat ini belum pernah dilakukan penelitian mengenai pengembangan sistem *Frequently Asked Questions* (FAQ) berbasis *information retrieval* dengan metode *string similarity*.

Pada penelitian ini di usulkan sistem *information retrieval* pada *Frequently Asked Questions* atau FAQ dengan mencari pertanyaan yang mirip (similar) pada daftar pertanyaan di basis data terhadap pertanyaan yang diberikan oleh pengguna dengan menggunakan algoritma *Cosine similarity* untuk mencari kesamaan kosinus tertinggi. Selanjutnya memberikan respon jawaban yang sebelum nya sudah di berikan label terhadap pertanyaan yang relevan dan memiliki similaritas paling tinggi [11][12].

Penelitian ini menggunakan FAQ pada Sistem Informasi Online Mahasiswa (SION) sebagai *dataset* pertanyaan. Tahapan yang dilakukan pada penelitian yaitu 1) Menerima input pertanyaan, 2). Melakukan data *preprocessing* dengan menggunakan Bag of Word [13], 3) Mengukur tingkat kemiripan dengan algoritma Cosine Similarity, 4) Melakukan Analisa Jawaban berdasarkan bobot similaritas. Evaluasi hasil pada penelitian ini dilakukan dengan menguji akurasi pemberian bobot pada *information retrieval* melalui memberikan *input* pertanyaan (*query*) dengan tiga kriteria berdasarkan tingkat kemiripan pada masing-masing pertanyaan. Hasil yang diperoleh Sistem *Information Retreival* mampu memberikan bobot

sesuai dengan tingkat similaritas pertanyaan (*query*) dengan *dataset* pertanyaan pada FAQ.

2. METODE PENELITIAN



2.1 Input Pertanyaan

Pertanyaan dimasukkan oleh user bersifat deskriptif dalam Bahasa Indonesial.

2.2 Data Preprocessing

Tahapan data *pre-processing* yang dilakukan adalah tokenisasi, pembentukan vector dengan Bag of Word (BoW) [14] [15].

2.3 Mengukur Tingkat Kemiripan Pertanyaan

Mengukur tingkat kemiripan pertanyaan dengan daftar pertanyaan sesuai dengan FAQ pada *dataset*, tahapan ini menerapkan Cosine similarity sesuai dengan persamaan 1 [16].

$$similarity(p_j, q) = \frac{p_j.q}{|p_j|.|q|} = \frac{\sum_{i=1}^{n}(W_{i,j}.W_{i,q})}{\sqrt{\sum_{i=1}^{n} W_{i,j}^2 . \sum_{i=1}^{n} W_{i,q}^2}} \quad (1)$$

Keterangan :

$W_{i,j}$ merupakan bobot kata *i* pada pertanyaan (*label*) *j*

$W_{i,q}$ merupakan bobot kata *i* pada pertanyaan (*query*) *q*

2.4 Analisa Jawaban dan Pengujian bobot

Tahapan analisa jawaban dilakukan untuk memperoleh jawaban berdasarkan tingkat kemiripan (bobot) pertanyaan yang telah di ukur pada tahap sebelumnya.

3. HASIL DAN PEMBAHASAN

3.1 Data Pertanyaan

Information retrieval bersumber dari *dataset* pertanyaan dan jawaban FAQ. Pada penelitian ini menggunakan FAQ akun microsoft dari SION (Sistem Informasi Online), tabel 1 menunjukkan data FAQ.

Tabel 1. *Dataset* FAQ SION

No	Pertanyaan dan jawaban
1	**Pertanyaan** : Siapa saja yang mendapatkan akun email Microsoft dengan suffix "@stikom-bali.ac.id"? **Jawaban** : Seluruh mahasiswa aktif mendapatkan akun email Microsoft.
2	**Pertanyaan** : Apa alamat akun email Microsoft nya? **Jawaban :** Format alamat akun email Microsoft adalah @stikom-bali.ac.id. Misal NIM nya

	adalah 200010001, maka akun email nya menjadi 200010001@stikom-bali.ac.id
3	**Pertanyaan** : Untuk passwordnya bagaimana? **Jawaban :** Password dari masing-masing akun adalah unik. Terdiri dari 8 alphanumeric yang di-generate random oleh sistem Microsoft. Apabila baru pertama kali menggunakan akun email Microsoft, silakan reset saja di web sion yang ada di menu Personal > Password Microsoft. Silakan lengkapi form yang ada, dan perhatikan error yang muncul apabila salah mengisi form
4	**Pertanyaan** : Saya sudah reset password di SION, lalu apa yang harus saya lakukan **Jawaban :** Apabila sudah benar mengisi form di Reset Password Microsoft, sistem akan mengirimkan email petunjuk reset password ke email yang terdaftar di biodata SION. Proses pengiriman memerlukan waktu paling lama 5 menit.
5	**Pertanyaan** : Saya sudah menunggu 5 menit, tapi tidak muncul di inbox email pribadi saya? **Jawaban :** Pastikan hal berikut ini: Email yang terdaftar di biodata SION bisa diakses. Jangan lupa refresh dulu inbox/spam/junk dari email nya. Silakan cek di bagian Junk/Spam mail dengan alamat pengirim : support_microsoft365@stikom-bali.ac.id. Jika Email yang terdaftar di biodata SION tidak dapat diakses, silakan ubah dulu di Update Biodata, lalu ulangi lagi reset password Microsoft nya.
6	**Pertanyaan** : Layanan apa saja yang saya dapatkan saat memiliki akun email Microsoft? **Jawaban :** Layanan yang diberikan secara gratis selama Anda masih menjadi mahasiswa aktif dari ITB STIKOM Bali adalah sebagai berikut: Email dengan suffix "@stikom-bali.ac.id", Storage cloud sebesar 1 TB di onedrive.
7	**Pertanyaan** : Mengapa harus login terlebih dahulu ke outlook.office.com ? **Jawaban :** Semua layanan Microsoft yang ada di point 6, memerlukan aktivasi di email Microsoft. Satusatunya penanda bahwa user sudah mengaktifkan layanan dari Microsoft adalah dengan cara login ke inbox dari outlook.office.com
8	**Pertanyaan** : Apakah saya bisa masuk ke Microsoft Teams menggunakan email pribadi untuk join perkuliahan? **Jawaban :** Bisa, tapi kami tidak menyarankan hal tersebut. Kami lebih menyarankan Anda menggunakan email stikom-bali.ac.id untuk dapat join perkuliahan karena email stikom-bali.ac.id sudah include untuk lisensi layanan Microsoft (lihat point 6).
9	**Pertanyaan** : Kenapa harus menggunakan email stikom-bali.ac.id ? **Jawaban :** Seluruh layanan yang disebutkan di point 6 merupakan layanan yang hanya akan didapatkan jika menggunakan email stikom-bali.ac.id (lisensi dari layanan tersebut terkoneksi dengan akun email stikom-bali.ac.id). Jika menggunakan email pribadi, maka Anda harus membeli sendiri layanan tersebut di toko online resmi Microsoft.

10	**Pertanyaan** : Saya tidak bisa masuk Ms Teams **Jawaban :** Pastikan beberapa hal ini sudah dilakukan: Pastikan username dan password benar, Sudah pernah login ke outlook.office.com dan sudah mengisi semua form yang ada, Sudah pernah masuk ke inbox outlook.office.com, Selisih durasi aktivasi akun email dan login pertama ke Ms Teams sudah lebih dari 24 jam.

3.2 Preprocessing

Tahap *preprocessing* bertujuan untuk menghasilkan tipe data vektor, tahap ini dilakukan transformasi data teks dari pertanyaan dan label jawaban pada *dataset* dengan vektor. Setiap kalimat di ekstrak dan direpresentasikan dalam bentuk vektor, proses ini dilakukan dengan Teknik Bag of Word dengan langkah 1). Menyusun setiap kata unik dalam angka 2) Menghitung jumlah setiap kata yang unik. Tabel 2 menunjukkan proses pembentukan vektor dengan Bag of Word untuk sebuah input pertanyaan dari *user* (*query*) dan satu pertanyaan pada *dataset* (label).

Pertanyaan dari user (*query*) : Alamat akun Microsoft

Pertanyaan pada *dataset* (label) : Apa alamat akun email Microsoft nya

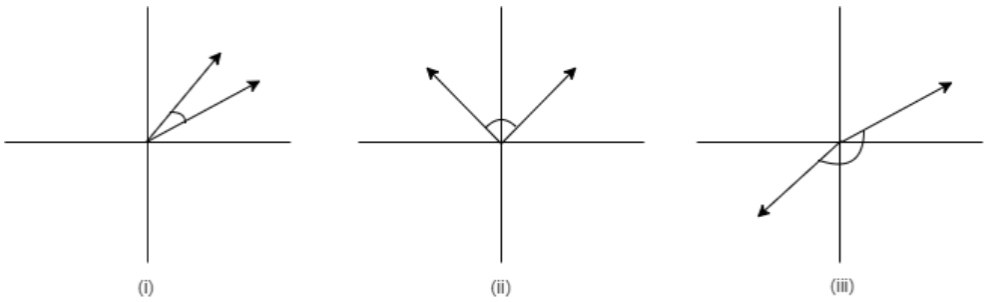
Tabel 2. Pembentukan Bag of Word

	Apa	**Alamat**	**Akun**	**Email**	**Microsoft**	**nya**
	1	**2**	**3**	**4**	**5**	**6**
query	0	1	1	0	1	0
label	1	1	1	1	1	1

Vektor pertanyaan (*query*) adalah [1,1,1,0,1,0]
Vektor pertanyaan (*label*) adalah [1,1,1,1,1,1]

3.3 Mengukur Kemiripan Pertanyaan

Setelah tahap *preprocessing* selesai dan telah terbetuk kedua vektor, selanjutnya menghitung nilai cosinus sudut antar dua vektor. Gambar 1 menunjukkan tiga kemungkinan kategori kemiripan



Gambar 1. Kategori Kemiripan

Rentan nilai Cosine similarity adalah dari -1 sampai dengan 1. Berdasarkan Gambar 1 maka terdapat tiga kemungkinan kategori kemiripan sebagai berikut :

(i) Kedua teks mirip ketika sudut antar vektor mendekati nol dengan nilai cosinus mendekati 1
(ii) Kedua teks tidak berkaitan atau tidak mirip ketika sudut antar vektor mendekati 90 derajat dan nilai cosinus mendekati 0
(iii) Kedua teks berlawanan ketika sudut antar vektor mendekati 180 derajat dan nilai cosinus negative dan medekati minus 1.

Vektor yang telah terbentuk pada tahap *preprocessing* kemudian di hitung dengan persamaan (1) maka di dapat tingkat kemiripan antar teks pada pembetukan vektor Tabel 2 adalah 0,471

3.4 Analisa Jawaban

Setelah tahap perhitungan kemiripan selesai, pada tahap Analisa jawaban dilakukan pembentukan urutan pertanyaan dan jawaban pada *dataset* berdasarkan bobot yang diperoleh. Tabel 3 menunjukan pertanyaan (*label*) yang memiliki bobot paling mendekati 1

Tabel 3. Bobot Similaritas Pertanyaan

No	Pertanyaan Similar *(label database)*	Bobot
1	Apa alamat akun email Microsoft nya	0.47140452079103173
2	Siapa saja yang mendapatkan akun email Microsoft dengan suffix @stikom-bali.ac.id	0.18257418583505536
3	Layanan apa saja yang saya dapatkan saat memiliki akun email Microsoft	0.17407765595569785

3.5 Evaluasi Hasil

Evaluasi hasil dilakukan dengan menguji akurasi pemberian bobot pada *information retrieval* dengan Cosine Similarity melalui memberikan input pertanyaan (*query*) dengan empat kriteria : 1) *query* dengan pertanyaan mengandung kata yang sama dengan salah satu pertanyaan pada *dataset* pertanyaan FAQ, 2) *query* dengan pertanyaan mengandung beberapa kata yang sama pada salah satu pertanyaan *dataset* FAQ, 3) *query* dengan pertanyaan tidak mengandung satupun kata pada *dataset* pertanyaan FAQ dan 4) *query* dengan kalimat menyerupai namun memiliki makna yang berlawanan dengan daftar pertanyaan pada *dataset*. Daftar pertanyaan ditunjukkan pada Tabel 4.

Tabel 4. Daftar Pertanyaan

No	Kode Pertanyaan	Kriteria	Pertanyaan
1	P1	1	Apa alamat akun email Microsoft nya?
2	P2	1	Layanan apa saja yang saya dapatkan saat memiliki akun email Microsoft?
3	P3	1	Saya tidak bisa masuk Ms Teams
4	P4	2	Akun email Microsoft?
5	P5	2	Layanan akun Microsoft?
6	P6	2	Cara masuk ms teams?
7	P7	3	Jadwal kuliah?
8	P8	3	Nilai matakuliah?
9	P9	3	Jadwal perwalian?
10	P10	4	Layanan yang tidak saya dapatkan saaat memiliki akun email Microsoft?
11	P11	4	Saya bisa masuk ms teams
12	P12	4	Saya belum reset password di SION

Tabel 5. Bobot Similartitas Pertanyaan *(query)*

No	Kode Pertanyaan	Bobot Similaritas	Ket. Pengamatan
1	P1	**2 (1.0)**	Sesuai
2	P2	**6 (1.0)**	Sesuai
3	P3	**10 (1.0)**	Sesuai
4	P4	**2 (0.7071)** 1 (0.5477)	Sesuai
5	P5	**6 (0.5222)**	Sesuai

6	P6	**10 (0.6123)**	Sesuai
7	P7	0	Sesuai
8	P8	0	Sesuai
9	P9	0	Sesuai
10	P10	**6 (0.6674)**	Tidak Sesuai
11	P11	**10 (0.9128)**	Tidak Sesuai
12	P12	**4 (0.5455)**	Tidak Sesuai

Tabel 5 menunjukkan bobot similaritas pada pertanyaan (*query*) tabel 5.4. Hasil pembobotan menunjukkan pada pertanyaan kriteria 1 secara konsisten memberikan bobot similaritas 1. Untuk pertanyaan dengan kode P1 memiliki kesamaan dengan *dataset* pertanyaan No. 2, P2 memiliki kesamaan dengan *dataset* pertanyaan No. 6 dan P3 memiliki kesamaan dengan *dataset* pertanyaan No 10. Pada pertanyaan dengan kriteria 2 sistem telah mampu memberikan bobot sesuai dengan tingkat kemiripan atau similaritas pertanyaan. Untuk pertanyaan dengan kode P4 memiliki kemiripan dengan *dataset* pertanyaan No. 2 dengan bobot 0.7071, P5 memiliki kemiripan dengan *dataset* pertanyaan No. 6 dengan bobot 0.5222 dan P6 memiliki kemiripan dengan *dataset* pertanyaan No. 10 dengan bobot 0.6123. Pada pertanyaan dengan kriteria 3 dengan pertanyaan uji tidak mengandung satupun kata pada *dataset* pertanyaan FAQ menghasilkan bobot similaritas 0. Sedangkan pada pertanyaan dengan kriteria 3 sistem tetap memberikan pembobotan terhadap pertanyaan pada *dataset* meskipun pertanyaan memiliki makna yang berlawanan. Jika dihitung persentase dengan persamaan (2), tingkat akurasi pada seluruh pertanyaan uji mencapai 75%.

$$s = \frac{t}{n}\ x\ 100 \quad (2)$$

Keterangan :
s : hasil perhitungan akurasi
t : pemberian bobot yang sesuai pada setiap pertanyaan
n : jumlah pertanyaan

Dengan demikian berdasarkan empat kriteria pertanyaan uji sistem *information retrieval* dengan Cosine Similarity telah mampu memberikan bobot sesuai dengan tingkat similaritas pertanyaan (*query*) dengan *dataset* pertanyaan pada FAQ namun sistem belum mampu membedakan makna pada setiap kalimat pertanyaan.

4. KESIMPULAN DAN SARAN

Penelitian ini menerapkan Algoritma Cosine similarity untuk mengukur tingkat kemiripan input pertanyaan (*query*) dengan daftar pertanyaan pada *dataset* FAQ pada Sistem Informasi Online (SION) yang digunakan untuk menyediakan *information retrival*. Telah dihasilkan *dataset* FAQ dan dilakukan *preprocessing*, menerapkan algoritma Cosine Similarity terhadap input pertanyaan (*query*) dengan *dataset* dan menghasilkan bobot pada setiap pertanyaan (label) pada *dataset*. Melalui evaluasi akurasi pemberian bobot similaritas yang dilakukan dengan memberikan dua belas input pertanyaan dibagi pada empat kategori berdasarkan tingkat kemiripan memiliki akurasi mencapai 75%, sistem *information retrieval* dengan Cosine similarity telah mampu memberikan bobot sesuai dengan tingkat similaritas pertanyaan (*query*) dengan *dataset* pertanyaan pada FAQ. Namun pada pemberian bobot similaritas masih terbatas medeteksi kemiripan pertanyaan dari segi leksikal belum mampu membedakan makna dari setiap kalimat pertanyaan. Pada penelitian selanjutnya akan ditingkatkan agar mampu memberikan bobot similaritas berdasarkan makna atau dapat mendeteksi kalimat pada pertanyaan yang memiliki informasi berlawanan.

UCAPAN TERIMA KASIH

Ucapan terima kasih penulis sampaikan kepada Institut Teknologi dan Bisnis STIKOM Bali yang telah mendukung penulis dari segi moril maupun finansial dalam menyelesaikan penelitian ini.

DAFTAR PUSTAKA

[1] F. Razzaghi, H. Minaee, and A. A. Ghorbani, "Context Free Frequently Asked Questions Detection Using Machine Learning Techniques," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2016, pp. 558–561, doi: 10.1109/WI.2016.0095.

[2] K. Miyamoto, A. Koseki, and M. Ohno, "Effective data curation for frequently asked questions," in *2017 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, 2017, pp. 7–12, doi: 10.1109/SOLI.2017.8120960.

[3] F. Amin and Purwatiningtyas, "Rancang Bangun Information Retrieval System (IRS) Bahasa Jawa Ngoko pada Palintangan Penjebar Semangad dengan Metode Vector Space Model (VSM)," *J. Teknol. Inf. Din.*, vol. 20, no. 1, pp. 25–35, 2015.

[4] F. Ramli, S. A. Noah, and T. B. Kurniawan, "Ontology-based information retrieval for historical documents," in *2016 Third International Conference on Information Retrieval and Knowledge Management (CAMP)*, 2016, pp. 55–59, doi: 10.1109/INFRKM.2016.7806335.

[5] X. Li and X. Xie, "Research of intelligent word segmentation and information retrieval," in *2010 2nd International Conference on Education Technology and Computer*, 2010, vol. 5, pp. V5-411-V5-414, doi: 10.1109/ICETC.2010.5529961.

[6] A. Latreche and L. Guezouli, "Similarity measure for semi-structured information retrieval based on the path and neighborhood," in *2012 International Conference on Information Technology and e-Services*, 2012, pp. 1–5, doi: 10.1109/ICITeS.2012.6216597.

[7] D. Soyusiawaty and Y. Zakaria, "Book Data Content Similarity Detector With Cosine Similarity (Case study on digilib.uad.ac.id)," in *2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 2018, pp. 1–6, doi: 10.1109/TSSA.2018.8708758.

[8] M. A. P. Subali and P. Wijaya, "Sistem Question Answering untuk Bahasa Bali menggunakan Metode Rule-Based dan String Similarity," *Techno.Com*, vol. 20, no. 2, pp. 300–308, 2021, doi: 10.33633/tc.v20i2.4390.

[9] M. M. umilasari Syabani reni, "Penerapan Metode Cosine Similarity dan Pembobotan TF/IDF pada Sistem Klasifikasi Sinopsis Buku di Perpustakaan Kejaksaan Negeri Jember," *JUSTINDO (Jurnal Sist. dan Teknol. Inf. Indones.*, no. Vol 3, No 1 (2018): JUSTINDO, pp. 31–42, 2018, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO/article/view/2345.

[10] G. N. M. Nata, "Information Retrival untuk Pencarian Dokumen Tugas Akhir Menggunakan Sequential Pattern Mining," *Semin. Multimed. \& Artif. …*, no. 84, pp. 81–86, 2018, [Online]. Available: http://papersmai.mercubuana-yogya.ac.id/index.php/smai/article/view/13.

[11] M. Alodadi and V. P. Janeja, "Similarity in Patient Support Forums Using TF-IDF and Cosine Similarity Metrics," in *2015 International Conference on Healthcare*

Informatics, 2015, pp. 521–522, doi: 10.1109/ICHI.2015.99.

[12] S. Pattnaik and A. K. Nayak, "Summarization of Odia Text Document Using Cosine Similarity and Clustering," in *2019 International Conference on Applied Machine Learning (ICAML)*, 2019, pp. 143–146, doi: 10.1109/ICAML48257.2019.00035.

[13] R. Shekhar and C. V Jawahar, "Word Image Retrieval Using Bag of Visual Words," in *2012 10th IAPR International Workshop on Document Analysis Systems*, 2012, pp. 297–301, doi: 10.1109/DAS.2012.96.

[14] T. S. Kartikasari, H. Setiawan, and P. Lucky Tirma Irawan, "Implementasi Text Mining Untuk Analisis Opini Publik Terhadap Calon Presiden," *J. Simantec*, vol. 7, no. 1, pp. 39–47, 2020, doi: 10.21107/simantec.v7i1.6528.

[15] P. Yu, X. Ruan, and X. Zhu, "The loop closure Detection Algorithm Based on Bag of Semantic Word For Robot Navigation," in *2020 IEEE International Conference on Information Technology,Big Data and Artificial Intelligence (ICIBA)*, 2020, vol. 1, pp. 54–58, doi: 10.1109/ICIBA50161.2020.9277317.

[16] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *J. Tek. Elektro Univ. Negeri Semarang*, vol. 9, no. 1, pp. 18–23, 2017, [Online]. Available: https://journal.unnes.ac.id/nju/index.php/jte/article/download/10955/6659.