Sistem Rekomendasi Obat berdasarkan Gejala dengan Information Retrieval

Adam Yogisyah Putra
Dep. Ilmu Komputer dan Elektronika
Universitas Gadjah Mada

Hizkya Firstadipa Hartoko
Dep. Ilmu Komputer dan Elektronika
Universitas Gadjah Mada

Muhammad Arsya Putra
Dep. Ilmu Komputer dan Elektronika
Universitas Gadjah Mada

I. Introduction

A. Latar Belakang

Pada era sekarang, penyakit semakin bervariasi, bersamaan dengan itu, banyak orang yang masih awam dalam memilih obat yang tepat untuk dapat mengobati penyakitnya sesuai gejala yang dirasakan. Padahal, di sisi lain obat-obatan sekarang mudah diperoleh dan telah banyak dikembangkan untuk menangani penyakit-penyakit tertentu. Hal tersebut menjadi masalah yang perlu diselesaikan agar masyarakat tidak kebingungan lagi dalam memilih obat yang tepat untuknya.

Dengan semakin berkembangnya teknologi, kita dapat memanfaatkan keberadaanya untuk memecahkan masalah di sekitar. Teknologi yang sangat canggih di zaman sekarang memungkinkan kita untuk mendapatkan informasi yang luas dalam waktu yang singkat. Dengan kemampuan tersebut, kita dapat membuat sebuah sistem rekomendasi yang dapat membantu masyarakat dalam memilih obat-obatan sesuai dengan gejala yang sedang dirasakan.

B. Related Work

Telah muncul beberapa aplikasi yang menyediakan sistem rekomendasi/pencarian obat seperti halodoc, alodokter, dan guesehat. Halodoc menyediakan layanan seperti konsultasi dengan dokter, layanan bidan, dan juga toko kesehatan yang di dalamnya, terdapat banyak jenis obat yang telah dikategorikan sesuai gejala yang disembuhkan, sedangkan alodokter dan guesehat menyediakan list obat berdasarkan nama obat tersebut.

C. Proposed approach

Perlu adanya sistem temu kembali informasi yang dapat memberikan rekomendasi obat yang paling sesuai berdasarkan input gejala yang dirasakan oleh pasien. Sistem ini perlu mengembalikan informasi obat-obatan yang akurat dengan input dari user. User dapat memasukan gejala yang sedang dirasakan, lalu sistem akan memberi daftar obat mulai dari obat paling relevan sesuai dengan gejala tersebut.

Data akan dikumpulkan dengan cara web scraping pada website-website yang menyediakan data obat-obatan, kemudian data-data tersebut akan diproses terlebih dahulu sebelum diolah agar menghasilkan keseragaman pada data. Hal ini agar pengolahan data lebih mudah dan cepat. Selanjutnya, akan dilakukan pembobotan/weighting terhadap konten di dalam data. Pembobotan dilakukan untuk mengurutkan data obat-obat menurut kecocokannya dengan gejala yang dimasukkan oleh pengguna.

Metode pengumpulan data dilakukan dengan cara web scraping. Metode tersebut dipilih karena informasi data-data obat dapat diambil dengan mudah di internet. dinilai mudah dan sederhana dalam pelaksanaanya. Scraping dilakukan terhadap website halodoc.com yang menyediakan data obat yang cukup detail mulai dari dosis hingga penggunaannya.

Metode pembuatan indeks dilakukan dengan metode Inverted Index. Pada inverted index term sudah diurutkan secara leksikografis. Kemudian dihitung kemunculannya pada setiap dokumen. Selain itu pada inverted index juga disimpan referensi atau pointer yang menunjukkan dimana term tersebut berada. Metode inverted index dipilih karena efektif, efisien, dan sederhana diimplementasikan.

Metode kompresi indeks dilakukan dengan model dictionary as a string, metode ini bertujuan untuk memperkecil ukuran data yang disimpan pada main memory. Term disambung menjadi satu string term yang panjang kemudian disimpan letak term tersebut pada dokumen, dan indeks term tersebut pada string term. Metode ini dipilih karena dinilai dapat memperkecil ukuran file dengan signifikan. Selain itu, metode ini sudah cukup populer dan mudah untuk diimplementasikan.

II. Methods

A. Akuisisi Data

Akuisisi data dilakukan dengan cara web scraping pada situs halodoc.com dengan menggunakan tools yaitu Python dengan bantuan library Beautiful Soup 4. Scraping dilakukan pada tiap-tiap obat berdasarkan kategorinya, kemudian hasil scraping disimpan dalam bentuk dictionary dan diexport ke dalam format CSV.

B. Transformasi Data

Metode transformasi data dilakukan dengan bantuan library Pandas yang bertujuan untuk merubah data mentah yang didapatkan dari akuisisi data ke format atau struktur yang lebih cocok digunakan untuk pembuatan indeks.

Pertama, semua dokumen dataset yang diperoleh dari web scraping digabungkan menjadi satu dokumen. Dokumen-dokumen ini berisi kumpulan data obat yang dikelompokkan sesuai dari gejala yang disembuhkan. Ketika dataset digabungkan, terdapat banyak data yang tidak memiliki nilai

pada beberapa kolom seperti komposisi, segmentasi, dan kontra indikasi, sehingga perlu dilakukan cleaning dataset. Kolom-kolom yang memiliki banyak nilai kosong dihapus karena tidak mewakilkan nilai sebenarnya dari dataset. Data-data yang memiliki nilai kosong pada minimal satu kolom juga dihapus sehingga kita memiliki dataset dengan nilai yang lengkap.

Selanjutnya, dilakukan pre-processing pada dataset yang terdiri dari case folding dan stopwords removal. Pre-processing ini dilakukan untuk menyeragamkan dataset sehingga mempermudah proses pengolahan data. Case folding berfungsi untuk merubah nilai karakter pada dataset menjadi huruf kecil, sedangkan stopwords removal berfungsi untuk menghilangkan nilai karakter seperti titik dan koma dari dataset. [2]

C. Pembuatan Indeks

Tahapan pembuatan indeks adalah tahapan yang bertujuan untuk mengubah data menjadi indeks-indeks dan kemudian akan disimpan dalam bentuk indeks tersebut [3]. Indeks dibuat dalam bentuk Inverted Index. Inverted Indeks adalah sebuah struktur data index yang dibangun untuk memudahkan query pencarian yang memotong tiap kata (term) yang berbeda dari suatu daftar term dokumen. [1] Inverted index bertujuan untuk meningkatkan kecepatan dan efisiensi dalam melakukan pencarian pada sekumpulan dokumen dan menemukan dokumen-dokumen yang mengandung query user.

1) Tokenisas: Tokenisasi merupakan tahapan memecah dokumen/kalimat menjadi lebih kecil yang disebut dengan term. Kemudian data dalam bentuk kolom-kolom tersebut digabung/di-merge menajadi satu kesatuan entitas data sehingga mempermudah proses-proses berikutnya

2) Formation of Vector: Token yang sudah digabungkan kemudian diambil termnya dan letak asal dokumennya, sehingga membuat suatu formasi term dan id dokumen. Selain itu, frekuensi kemunculan term-term pada setiap dokumen juga dihitung. Kemudian term diurutkan secara leksikografis.

3) Inverted Index: Inverted index adalah hasil akhir dari proses pembuatan index. Frekuensi kemunculan term dalam dokumen dapat diketahui melalui matriks inverted index, sekaligus letak term tersebut pada indeks

D. Kompresi Indeks

Kompresi indeks adalah tahapan untuk mengurangi ukuran dari indeks yang telah dibuat dengan tujuan untuk menghemat ruang penyimpanan dan juga mempercepat performa sistem agar lebih efisien.

Kompresi indeks dilakukan dengan menggunakan metode dictionary as a string yaitu menggabungkan seluruh term yang telah diperoleh menjadi satu string, kemudian dibuat pointer untuk tiap-tiap term sebagai penunjuk lokasi term dalam string tersebut. Pointer tersebut disimpan dalam suatu dataset bersamaan dengan frekuensi dan indeks dokumennya.

E. Arsitektur

1) Pengguna: Pengguna/User adalah seseorang yang menggunakan sistem untuk mencari rekomendasi obat.

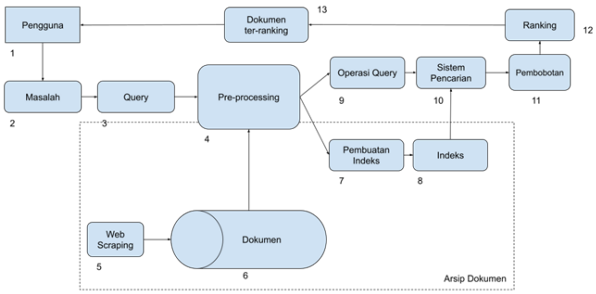

Fig. 1. Arsitektur Sistem Rekomendasi Obat

2) Masalah: Masalah adalah sesuatu yang ingin diselesaikan/ditemukan solusinya melalui sistem, dalam hal ini masalah adalah cikal bakal dari query. Masalah dari user adalah gejala yang sedang diderita.

3) Query: Gejala dari user dimasukkan ke sistem sebagai query.

4) Pre-Processing: Data dokumen dan query diproses terlebih dahulu agar memiliki bentuk yang seragam. Akan dilakukan transformasi data teks seperti seperti stopword removal, stemming, dan lain-lain.

5) Web Scraping: Proses ekstraksi data/informasi dari website-website kesehatan yang menyediakan informasi mengenai obat-obatan.

6) Dokumen: Hasil ekstraksi dari web scraping dikumpulkan dan disimpan pada database.

7) Pembuatan Indeks: Proses pemberian indeks pada dokumen-dokumen yang tersedia

8) Indeks: hasil pembuatan indeks pada dokumen-dokumen

9) Operasi Query: Pemrosesan query agar dapat digunakan pada sistem TKI

10) Sistem Pencarian: Proses pencarian data yang bersesuaian dengan query yang diinputkan user

11) Pembobotan: Pembobotan adalah proses kalkulasi ketersesuaian dokumen dengan query yang user inputkan

12) Ranking: Pengurutan dokumen berdasarkan hasil kalkulasi pembobotan dari yang paling sesuai

13) Dokumen Terurut: Dokumen yang telah di-ranking akan digunakan sebagai acuan dalam menentukan hasil pencarian berupa nama-nama obat dari yang paling sesuai

III. Results

1) Akuisisi Data: Metode akuisisi data dengan web scraping menghasilkan beberapa dokumen data obat yang dikategorikan berdasarkan gejala yang diobati. Web scraping menggunakan python dengan library Beautiful Soup 4 pada situs halodoc.com mengalami beberapa masalah seperti data-data pada kategori tertentu tidak berhasil terambil oleh algoritma web scraping dan algoritma web scraping tidak dapat digunakan pada beberapa halaman karena struktur html yang berbeda sehingga perlu penyesuaian.

	product_name	category	deskripsi	indikasi_umum	dosis	aturan_pakai
0	ABDI Baterai Alat Bantu Dengar Ukuran 10 (Kuning)	Alat Bantu Dengar	Baterai Alat Bantu Dengar ABDI tersedia dalam...	Baterai untuk alat bantu dengar.\r\nSpesifika...	-	-
17	Vikacare Hearing Aid Golden Vk-115	Alat Bantu Dengar	-	Alat bantu dengar	-	-
16	Vikacare Hearing Aid Diamond Vk-125	Alat Bantu Dengar	-	Alat bantu dengar	-	-
15	Tubing Earmold Alat Bantu Dengar (Hearing Aids...	Alat Bantu Dengar	Tubing Earmold Hearing Aids merupakan selang ...	Earmold tubing untuk penghubung antara alat b...	-	-
19	Widex Evoke 50 Fp	Alat Bantu Dengar	WIDEX EVOKE 50 FP merupakan alat bantu dengar...	Alat bantu dengar digital dan programmable de...	Sesuai kebutuhan	-
...	...	...	...	...	...	...
3972	NaN	NaN	NaN	NaN	NaN	NaN
3973	NaN	NaN	NaN	NaN	NaN	NaN

Fig. 2. Dataset Hasil Scraping

2) Transformasi Data: Metode transformasi data dengan melakukan cleaning, stopword removal, merging data, dan case folding menghasilkan data yang serupa dan seragam. Namun beberapa data perlu dihilangkan karena beberapa tidak memiliki atribut yang lengkap dan tidak sesuai dengan kriteria data yang dipakai. Data yang sudah ditransformasikan memiliki struktur yang siap untuk dilakukan pembuatan indeks.

	product_name	category	deskripsi	indikasi_umum	dosis	aturan_pakai
6	cetirizine sirup 60 ml	alergi	cetirizine merupakan antihistamin yang secara...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	dikonsumsi sesudah makan
11	ocuson 10 tablet	alergi	ocuson tablet adalah obat kombinasi kortikost...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	diminum setelah makan
10	interhistin 50 mg 10 tablet	alergi	interhistin 50 mg tablet merupakan obat denga...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	diminum setelah makan
8	ctm 4 mg 12 tablet	alergi	ctm 4 mg tablet merupakan obat anti alergi ya...	obat ini digunakan untuk mengatasi gejala ale...	dewasa \r\n1 tablet diminum 3 4 kali per h...	diberikan bersama atau tanpa makanan
7	desteem plus 10 tablet	alergi	desteem plus tablet adalah obat kombinasi kor...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	obat diminum sesudah makan

Fig. 3. Dataset Clean hasil transformasi data

3) Pembuatan Indeks: Metode pembuatan indeks menghasilkan data inverted index yang berisi term, frekuensi dari term tersebut, dan posting list (Posisi term pada dokumen asli). Struktur data seperti ini dapat mempermudah proses pencarian dan mudah untuk dikompresi.

	term	docFrequency	postingList
2000	kuning	1	[94]
2001	kunyah	1	[382]
2002	kurang	19	[5, 183, 2, 111, 111, 165, 111, 159, 6, 165, 1...
2003	kurangnya	1	[271]
2004	kurma	3	[258, 258, 258]
...	...	...	...
2095	letakkan	4	[48, 268, 376, 304]
2096	leukosit	1	[221]
2097	levofloxacin	4	[267, 19, 19, 267]
2098	levonogestrel	2	[204, 215]
2099	lewat	1	[67]

100 rows × 3 columns

Fig. 4. Inverted Index hasil index construction

4) Kompresi Indeks: Kompresi indeks merupakan tahapan yang bertujuan untuk memperkecil ukuran file yang disimpan pada main memory. Tahapan ini menghasilkan indeks terkompresi dengan metode dictionary as a string, tetapi metode ini kurang signifikan, terlihat dampaknya disebabkan oleh dataset tidak terlalu besar ukurannya. Dataset berukuran 279 Kilobyte berhasil dikompresi menjadi 250 Kilobyte. Metode ini jika dilakukan pada dataset dengan ukuran yang lebih besar akan lebih memperlihatkan keefektifannya

	Freq	Posting_ptr	Term_ptr
0	8	[273, 344, 344, 273, 45, 18, 276, 29]	0
1	2	[421, 421]	1
2	1	[381]	3
3	6	[252, 250, 254, 250, 252, 254]	10
4	4	[210, 210, 210, 210]	18
...	...	...	...
3859	2	[385, 385]	29481
3860	2	[160, 146]	29487
3861	4	[388, 388, 390, 390]	29489
3862	4	[342, 384, 363, 387]	29498
3863	2	[233, 233]	29502

3864 rows × 3 columns

Fig. 5. Dataset hasil

'aaaabdomenabnormalabortsiabsesabsolutabsopsiabsorpsinyaaceacetateacetonideacetylcysteineacetylsalicylicacidacnacareacneacneactacquiredacsbpactifed actingactionacycloviradadaadalahadanyaadekuatadenosinadenosinaadenosinaadhesiadhesiveadidasadipositadpadrenergicadrenergikadrenoceptoradrenokortikalad+enoreceptoradrenoreseptoradukadvancaaereusaerobaerogenesaeruginosaafricanumaftosaagalactiasaagaragenagentagonisagonistagregasiahilainsairakanakarakhirakhirnyaakibatakilenaknaeksiaktifaktifitasaktivasiaktivatoraktivitasaktivitasaktivitasanyaakulakuminatasakurasiakuratakutalamalamiamiahalasalatalbicansalbothylalcoholaldosteronaldosteronealergenalergialfaalialinanalkalosisalkoholallergyallorisalociairaloealparaalphaalternatifalternatifnyaaluminiumalumunaluminumamanambilambroxolamenoreameneroreasaminefronaminoaminoacylaminoglikosidasaminopenisilinamlodipinamlodipineammoniumamoebiasiamokisilinamoxicillinamoxiclavamoxsanampulanaerobanafilaksisanakanalgesikanalgetikanaloganalognyaanastesiandaandalanamnesa'

Fig. 6. Dictionary as String

IV. Discussion

Berdasarkan metode yang telah dijelaskan sebelumnya, dihasilkan sebuah dataset obat dalam bentuk inverted index dengan ukuran yang kecil dan format dictionary as a string yang mudah untuk diakses. Arsitektur dari dataset efisien untuk digunakan karena semua term dan index dari dokumen saling terhubung melalui posting list yang berisi list dokumen yang memiliki term tersebut didalamnya.

Namun indeks dan dataset sistem rekomendasi obat ini masih memiliki beberapa kekurangan dan limitasi dalam penggunaannya. Jumlah dataset obat yang dimiliki masih sangat sedikit karena banyak data obat yang tidak berhasil diambil saat dilakukan scraping dan beberapa data obat terpaksa dihapus akibat memiliki nilai informasi yang kosong. Selain dari jumlah dataset, format inverted index menggunakan posting list juga membuat data term dalam dokumen saling terpisah satu sama lainnya sehingga dapat mempengaruhi akurasi dari sistem rekomendasi obat. Dikarenakan ukuran dari dataset yang kecil, performa dari metode kompresi indeks tidak bisa diukur dengan pasti. Metode pembuatan indeks yang digunakan juga dibuat dengan anggapan koleksi

dataset bersifat statis, artinya jarang terjadi penambahan dan modifikasi data. Karena itu update pada dataset tidak efisien.

Dataset dan indeks dalam sistem ini dapat disempurnakan dengan menambah jumlah data obat yang disimpan dan menggunakan metode pembuatan dan kompresi indeks yang lebih baik seperti dynamic indexing dengan logarithmic merge dan kompresi menggunakan metode blocking ditambah front coding.

REFERENCES

[1] S. Singh, "What is an inverted index?," Educative, 29-Sep-2022. [Online]. Available: https://www.educative.io/answers/what-is-an-inverted-index. [Accessed: 29-Sep-2022].

[2] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to Information Retrieval," Capitalization/case-folding., 2008. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/capitalizationcase-folding-1.html. [Accessed: 29-Sep-2022].

[3] P. Pecina, "Index construction, Distributed and dynamic indexing, Index compression," NPFL103: Information retrieval. [Online]. Available: https://ufal.mff.cuni.cz/ pecina/ir/. [Accessed: 29-Sep-2022].