PENERAPAN METODE STEMMING UNTUK SISTEM TEMU KEMBALI INFORMASI ARTIKEL KESEHATAN

Iin Ernawati
Program Studi Sistem Informasi, Fakultas Ilmu Komputer UPN "Veteran" Jakarta
Jl. RS. Fatmawati Pondok Labu Jakarta Selatan - 12450
Telp. 021 7656971 E-mail: iin_ernawati@yahoo.com

Abstract

Information retrieval is a process which aims at finding some collected documents in the form of text in order to meet the desires of the users for retrieving information. The desired information inquired by the users may contain one or more terms to be used when searching, but it can also add other information such as the weight of each word. Data mining is a set of processes to explore the added value of a data set of knowledge that has been known to manually. Data Mining is the process of finding patterns and relationships hidden in large amounts of data in order to perform classification, estimation, prediction, association rules, clustering, description and visualization. This process can be applied to several types of documents, one of them is the documents which contain article about health. The process will run more smoothly by applying a method called Stemming Method which can be combined with the steps of documents processing using the formula of tf and idf.

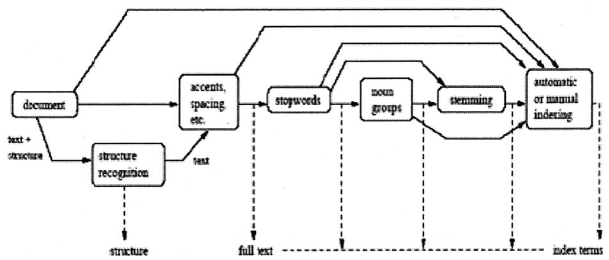
Key Words: *information, retrieval, system, data, mining*

PENDAHULUAN

Information Retrieval pada dasarnya merupakan proses untuk menentukan dokumen dalam koleksi yang harus ditemu kembalikan sesuai keingginan pengguna informasi. Informasi yang diinginkan pengguna direpresentasikan dalam bentuk *query* dan mengandung satu atau lebih term yang digunakan dalam pencarian, dapat juga ditambahkan informasi lain seperti bobot tiap term tersebut. Oleh karena itu, keputusan menemu kembalikan dokumen dibuat dengan membandingkan *term–term* dari *query* dengan term indeks yang terdapat pada dokumen tersebut. Keputusan yang diambil merupakan keputusan biner *(retieve/reject),* atau melibatkan perkiraan *relevance degree* dokumen terhadap *query.* Suatu sistem temu kembali informasi dikatakan ideal jika sistem tersebut dapat menemukan seluruh dokumen yang relevan dan sistem hanya menemukan dokumen tersebut, tetapi, *term–term* yang terdapat pada dokumen dan *query* sering memiliki banyak varian *morfologik,* sehingga pasangan *term* seperti *"computing* dan *computation"* tidak dianggap ekuivalen oleh sistem tanpa suatu bentuk *Natural Language Processing (NLP).* Dalam sistem temu kembali informasi *(Information Retrieval/IR)* terdapat kumpulan dokumen, yang setiap dokumennya dideskripsikan dengan kata–kata (istilah). Istilah yang memiliki akar kata *(stem)* yang sama umumnya memiliki arti yang sama, misalnya hubung, hubungan, hubungkan, hubungi. Jika keempat istilah ini dikelompokkan kedalam satu kelompok dengan menghilangkan akhirannya, kinerja sistem IR menjadi meningkat. Proses penghilangan akhiran kata mengurangi jumlah total istilah dalam sistem IR sehingga mengurangi ukuran dan kompleksitas data dalam sistem. Sebagai contoh sederhana, jika dicari suatu dokumen dengan judul *"How to Write"* dengan menggunakan *query "writting",* dokumen yang dimaksud tidak pernah terdapat dalam hasil pencarian. Tetapi, jika *query* dipotong mengakibatkan writting diubah menjadi *write,* maka pencarian akan berhasil. Hal ini tidak hanya berarti varian yang berbeda dari suatu term dapat direduksi sebagai satu bentuk *representative,* tetapi juga berdampak terhadap berkurangnya ukuran *inverted file* (jumlah term yang diperlukan untuk merepresentasikan suatu koleksi dokumen). Ukuran *inverted file* yang kecil dapat menghemat *storage space* dan mempercepat waktu pemrosesan.

Adapun rumusan masalah yang ditemukan yaitu untuk menyusun koleksi dokumen teks berbahasa Indonesia sebagai bagian perangkat pengujian sistem temu kembali informasi. Tujuan penelitian untuk menerapkan perangkat uji sistem temu kembali informasi, dan menyusun koleksi dokumen teks berbahasa Indonesia sebagai bagian perangkat pengujian sistem temu kembali informasi. Batasan masalah dalam penelitian ini adalah (1) data yang digunakan adalah data artikel Bahasa Indonesia, (2) koleksi dokumen yng digunakan untuk penelitian ini merupakan berkas teks dengan *query* yang telah ditentukan sebelumnya, (3) menerapkan serta mengembangkan proses pemrograman PHP serta database MySQL untuk mengolah dan menganalisis data atau dokumen teks atau artikel bahasa Indonesia, dan (4) mengetahui hubungan antara sistem temu kembali informasi dengan *data mining*.

Dokumen dalam koleksi seringkali direpresentsikan melalui suatu *set term indeks* atau *keywords*. *Term indeks* dapat diekstrasi secara langsung dari teks dokumen atau didefinisikan secara manual (dibuat oleh spesialis seperti banyak dilakukan pada bidang *information science)*. Komputer modern memungkinkan representasi suatu dokumen dengan menggunakan seluruh *set term* yang terdapat pada dokumen tersebut. Sistem temu kembali informasi disebut sebagai mengadopsi *full text logical view* dari dokumen. Jika koleksi dokumen yang digunakan sangat besar, bahkan komputer modern sekalipun mungkin harus mengurangi jumlah *set term* indeks. Hal ini dapat dilakukan dengan membuang *stopwords*, menggunakan *stemming*, ataupun melakukan identifikasi noun groups (yang akan mengeliminasi *adjective, adverbs, dan verbs).* Operasi–operasi tersebut disebut *text operations* (transformasi). *Text operations* mengurangi kompleksitas dari representasi dokumen dan memungkinkan perubahan *logical view* dari *full text* menjadi set *term indeks.*



Gambar 1. *Logical view* suatu dokumen dari f*ull text* menjadi *set term indeks*

PEMBAHASAN

Penelitian ini adalah merupakan suatu kajian yang dilaksanakan tehadap sebuah bahasa guna meneliti struktur bahasa secara mendalam. Data yang digunakan dalam penelitian ini adalah dataset yang bertipe artikel, berupa data hasil dari penghitungan *tf, idf,* serta *similarity* data artikel berbahasa Indonesia yang diambil dari sumber Kompas Gramedia Jakarta Barat dengan jumlah data yang diperoleh sebanyak 35 artikel. Salah satu artikel bahasa Indonesia yang diambil dari Pusat Informasi Kompas yaitu artikel yang ditulis oleh Mohamad Harli dengan judul Alpukat Sehat Untuk Penderita HIV/AIDS *(Kompas Minggu, Tanggal 3 Juni 2001, hal. 22).*

Perkembangan tingkat keparahan penyakit sejak seseorang dinyatakan positif HIV hingga ke tahap AIDS sangat dipengaruhi oleh perawatan medis dan penanganan gizi. Mereka yang mendapat perawatan medis dan gizi secara baik berusia lebih panjang. Salah satu makanan yang sehat untuk mereka adalah buah alpukat (Persea Americana). Ada beberapa faktor yang membuat buah alpukat sehat untuk orang-orang dengan HIV/AIDS ini. Pertama, sumber energi, terutama lemak, yang aman, dilaporkan bahwa orang dengan HIV/AIDS memerlukan energi/kalori yang lebih banyak dibandingkan orang yang sehat. Salah satu sumber utama energi tersebut berasal dari lemak. Alpukat merupakan salah satu buah yang tinggi kadar lemaknya, sekitar 16 persen. Hasil penelitian diketahui bahwa lemak yang terkandung dalam buah alpukat aman dan malah menyehatkan. Hal itu karena sekitar 63 persen unsur penyusunnya adalah asam lemak tidak jenuh, terutama asam lemak tidak jenuh tunggal (*MUFA, monounsaturated fatty acids)*. Diet rendah lemak sering menurunkan kolesterol HDL *(high density lipoprotein)* yang menyehatkan. Diet alpukat yang kaya MUFA dapat menurunkan kadar kolesterol *LDL (low density lipoprotein)* yang merugikan kesehatan darah, tanpa menurunkan kadar HDL. Lemak MUFA juga mempunyai aktivitas antioksidan yang menjaga tubuh dari kerusakan arteri akibat keganasan kolesterol LDL. Alpukat dapat melindungi arteri dari kerusakan oksidatif dan mengamankan kolesterol sehingga tidak menjadi ganas dan berbahaya. Kedua, orang-orang dengan HIV/AIDS memerlukan masukan vitamin dan mineral yang kuantitasnya cukup dan kualitasnya baik. Alpukat mengandung vitamin A dan karoten yang baik. Dalam 100 gram buah

alpukat terkandung sekitar 300-400 IU vitamin A, dan sekitar 165 mikrogram karoten. Terkandung pula tiamin, riboflavin, dan niasin, yang tergolong vitamin B-kompleks. Kadar vitamin C alpukat cukup baik, sekitar 14 mg per 100 gram buah alpukat. Buah alpukat kaya akan mineral kalium (604 mg/100 g) dan rendah mineral natrium (4 mg/100g). Dilaporkan makanan yang kadar kaliumnya tinggi dan natriumnya rendah dengan rasio K:Na lebih besar dari 5:1 adalah makanan yang sehat untuk menjaga kesehatan jantung dan pembuluh darah. Serat alpukat juga tinggi sekitar 1,6 gram/100 g. Hal ini bermanfaat untuk membantu sistem pencernaan dan membuang sisa-sisa pencernaan yang beracun. Ketiga, buah alpukat mengandung kadar glutation tertinggi diantara buah-buahan, yaitu 21 mg per 100 gram buah segar. Senyawa *glutation* tersebut berfungsi sebagai unsur pertahanan tubuh dari berbagai serbuan perusak kesehatan tubuh. Ia berfungsi sebagai antioksidan, bersama vitamin C, E, dan karoten, dalam meningkatkan sistem kekebalan tubuh. *Glutation* dilaporkan berfungsi sebagai zat antikanker yang dapat menonaktifkan sedikitnya 30 zat penyebab kanker. *Glutation* juga membantu menghambat kerusakan tubuh akibat senyawa beracun, misalnya bahan pencemar lingkungan seperti pestisida, logam-logam berat (timah), yaitu dengan cara menawarkan racun tersebut lalu membuangnya lewat sistem pembuangan (feses, urin, atau keringat).

Kaitan *glutation* dengan AIDS tampak dalam hasil penelitian laboratorium Dr Alton Meister dari Fakultas Kedokteran Universitas Cornell, AS. Meister menemukan bahwa *glutation* menghentikan penyebaran/penjalaran atau reflikasi virus HIV secara in vitro. Makin banyak senyawa glutation yang ditambahkan makin besar efeknya. Kadar glutation pada orang dengan HIV/AIDS terus menurun, hal ini tentu saja akan mempercepat tingkat keparahan HIV ke tahap AIDS. Artikel diatas dapat dicari nilai *tf, idf, dan similarity*-nya.

Berikut contoh data laporan berupa hasil penghitungan dari data artikel bahasa Indonesia tentang kesehatan. Tabel di bawah ini merupakan tabel yang berisi contoh dataset laporan hasil perhitungan data artikel bahasa Indonesia.

Tahap pertama yang dilakukan adalah penghilangan *stopwords* yang berasal dari masing–masing artikel. *Stopwords* tersebut berupa kata penghubung, seperti dan, dengan, ke, di, pada, serta, yang dan sebagainya.

Tahap kedua yang dilakukan adalah penghitungan nilai tf yaitu tahap pemilihan atau pentuan query yang ditentukan. Kemudian dicari pada setiap artikel yang telah di kumpulkan serta dihitung, seberapa banyak query yang dihasilkan pada setiap artikel.

Tabel 1. Nilai tf

Kata	tf										
	Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Penyakit	21	1	0	2	0	18	10	2	0	0	0
Kanker	9	1	0	0	0	2	45	0	0	0	2
Sehat	5	4	0	3	0	0	0	0	0	0	0
Gizi	9	3	0	0	0	0	0	1	20	18	0
HIV	5	7	0	0	0	0	0	0	0	0	0
Jantung	8	0	0	2	0	2	0	0	0	0	0
Diabetes	8	0	0	4	2	0	0	16	0	0	0
Kolesterol	4	4	0	29	0	0	0	0	0	0	0
Darah	13	2	3	2	12	1	0	13	0	0	0
Depresi	1	0	0	0	0	28	0	0	0	0	0
Ginjal	3	0	0	0	44	0	0	0	0	0	0
Infeksi	9	0	0	0	0	0	1	0	0	0	4
Kesehatan	15	3	0	2	0	0	1	1	7	4	0
Osteoporosis	1	0	0	0	0	0	0	0	0	0	0
Lupus	2	0	0	0	1	0	0	0	0	0	0

Tahap ketiga adalah pencarian atau penghitungan nilai idf adalah tahap yang melakukan penghitungan atau pncarian nilai bobot pada setiap artikel yang telah dikumpulkan.

Tabel 2. Hasil Pencarian Nilai idf

N/n	idf
1.7	0.22
3.9	0.59
7	0.85
3.9	0.59
7	0.85
4.4	0.64
4.4	0.64
8.8	0.94
2.7	0.43
35	1.54
12	1.07
3.9	0.59
2.3	0.37
35	1.54
18	1.24
35	1.54
8.8	0.94
12	1.07
35	1.54
5	0.7

Nilai idf diatas, dapat dicari nilai bobot (w) dari masing–masing artikel bahasa Indonesia. Sehingga didapatkan hasil dari penghitungan bobot (w) tersebut berupa tabel 3 terhadap dokumen artikel bahasa Indonesia tentang kesehatan.

Tabel 3. Hasil Perhitungan Nilai Bobot (w)

Idf	Bobot (w)										
	Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
0.22	4.659	0.222	0	0.44	0	3.99	2.22	0.44	0	0	0
0.59	5.308	0.59	0	0	0	1.18	26.5	0	0	0	1.18
0.85	4.225	3.38	0	2.54	0	0	0	0	0	0	0
0.59	5.308	1.769	0	0	0	0	0	0.59	11.8	10.6	0
0.85	4.225	5.916	0	0	0	0	0	0	0	0	0
0.64	5.128	0	0	1.28	0	1.28	0	0	0	0	0
0.64	5.128	0	0	2.56	1.28	0	0	10.3	0	0	0
0.94	3.768	3.768	0	27.3	0	0	0	0	0	0	0
0.43	5.592	0.86	1.29	0.86	5.16	0.43	0	5.59	0	0	0
1.54	1.544	0	0	0	0	43.2	0	0	0	0	0
1.07	3.201	0	0	0	46.9	0	0	0	0	0	0
0.59	5.308	0	0	0	0	0	0.59	0	0	0	2.36
0.37	5.52	1.104	0	0.74	0	0	0.37	0.37	2.576	1.47	0
1.54	1.544	0	0	0	0	0	0	0	0	0	0
1.24	2.486	0	0	0	1.24	0	0	0	0	0	0
1.54	1.544	0	0	0	0	0	0	0	0	0	34
0.94	3.768	0	0	0	0	0.94	0	0	0	0	0
1.07	3.201	0	0	0	0	0	0	0	0	0	0
1.54	1.544	0	0	0	0	0	0	0	0	0	0
0.7	4.893	0	0	0	0	0	0	0	2.097	0	0.7

Setelah melakukan tahap praproses maka dilanjutkan dengan tahap data mining. Pada proses ini kita akan menentukan nilai data-data record tersebut, kemudian kita dapat mencari nilai *similarity* atau kemiripan data dari masing–masing artikel tersebut. Langkah pertama yang harus dilakukan adalah dengan mencari nilai similarity pada setaip dokumen artikel dengan menggunakan rumus. Sehingga didapatkan hasil dari penghitung nilai similarity tersebut berupa tabel terhadap dokumen artikel bahasa Indonesia tentang kesehatan.

Tabel 4. Hasil Perhitungan Nilai Similariti

Kata Similariti											
Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	
Penyakit	21.7	1.03	0	2.067	0	18.6	10.3	2.067	0	0	0
Kanker	28.18	3.13	0	0	0	6.26	141	0	0	0	6.262
Sehat	17.85	14.3	0	10.71	0	0	0	0	0	0	0
Gizi	28.18	9.39	0	0	0	0	0	3.131	62.6	56.4	0
HIV	17.85	25	0	0	0	0	0	0	0	0	0
Jantung	26.29	0	0	6.574	0	6.57	0	0	0	0	0
Diabetes	26.29	0	0	13.15	6.57	0	0	52.59	0	0	0
Kolesterol	14.2	14.2	0	102.9	0	0	0	0	0	0	0
Darah	31.27	4.81	7.22	4.81	28.9	2.41	0	31.27	0	0	0
Depresi	2.384	0	0	0	0	66.8	0	0	0	0	0
Ginjal	10.25	0	0	0	150	0	0	0	0	0	0
Infeksi	28.18	0	0	0	0	0	3.13	0	0	0	12.52
Kesehatan	30.47	6.09	0	4.062	0	0	2.03	2.031	14.2	8.12	0
Osteoporosis	2.384	0	0	0	0	0	0	0	0	0	0
Lupus	6.181	0	0	0	3.09	0	0	0	0	0	0
Tumor	2.384	0	0	0	0	0	0	0	0	0	52.45
Asma	14.2	0	0	0	0	3.55	0	0	0	0	0
Narkotika	10.25	0	0	0	0	0	0	0	0	0	0
Hepatitis C	2.384	0	0	0	0	0	0	0	0	0	0
Rokok	23.94	0	0	0	0	0	0	0	10.3	0	3.42

Untuk lebih jelasnya mengenai hasil perhitungan similarity di atas, menggunakan tabel 5 berikut.

Tabel 5.
Hasil Perhitungan Nilai Bobot (w) dari Beberapa Contoh Artikel dan Query

Bobot (w)										
Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
4.659	0.222	0	0.44	0	3.99	2.22	0.44	0	0	0
5.308	0.59	0	0	0	1.18	26.5	0	0	0	1.18
4.225	3.38	0	2.54	0	0	0	0	0	0	0
5.308	1.769	0	0	0	0	0	0.59	11.8	10.6	0
4.225	5.916	0	0	0	0	0	0	0	0	0

Kemudian pada tahap selanjutnya yaitu tahap data mining, dimana pada tahap ini dicari nilai similaritynya.

Tabel 6.
Hasil Perhitungan Nilai Similariti dari Beberapa Contoh Artikel dan Query

Kata Similariti											
Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	
Penyakit	21.7	1.03	0	2.067	0	18.6	10.3	2.067	0	0	0
Kanker	28.18	3.13	0	0	0	6.26	141	0	0	0	6.262
Sehat	17.85	14.3	0	10.71	0	0	0	0	0	0	0
Gizi	28.18	9.39	0	0	0	0	0	3.131	62.6	56.4	0
HIV	17.85	25	0	0	0	0	0	0	0	0	0

SIMPULAN

Temu kembali informasi meliputi representasi, penyimpanan, pengorganisasian, dan pengaksesan informasi. Deskripsi informasi yang diberikan pengguna pada sistem direpresentasikan sebagai suatu *query* yang dapat diproses oleh *search engine (Information Retrieval System)*. Data *mining* merupakan proses pencarian pola dan relasi-relasi yang tersembunyi dalam sejumlah data yang besar.

Metode *stemming* dapat memudahkan pembobotan untuk menetukan nilai *query* yang telah ditentukan. Setiap bobot dokumen diketahui, maka dilakukan proses pemeringkatan atau perankingan dokumen berdasarkan besarnya tingkat kerelevanan (kesesuaian) dokumen terhadapa *query*, di mana semakin besar nilai bobot dokumen terhadap *query* maka semakin besar tingkat similaritas dokumen tersebut terhadap *query* yang dicari.

Data *mining* juga mempunyai hubungan dengan sistem temu kembali informasi, di mana hubungan itu salah satunya yaitu memiliki cakupan yang lebih luas bila dibandingkan dengan sistem temu kembali informasi. Sistem temu kembali informasi sebagai salah satu bagian dari data *mining*.

Penerapan sistem temu kembali informasi tersebut berupa search engine atau dengan nama lain yaitu temu kembali informasi yang telah diproses oleh data mining.

DAFTAR PUSTAKA

Hung Wei Chia, Tsun Li Chang.*Design of Content based Multimedia Retrieval.* Department of Computer Science. University of Warwick. United Kingdom.

Multimedia Data Mining. 2008. Departemen Ilmu Komputer.Bogor: Institut Pertanian Bogor.

Multimedia Retrieval/Multimedia Mining. 2008. Fujitsu.

Tesic, Jelena S. *Multimedia Data Mining.*Vision Research Laboratory. Department of Electrical and Computer Engineering. University of California.Santa Barbara

ww.dcs.warwick.ac.uk/~ctli/papers/Chapter3.pdf