Prediksi Google Search Engine Result Page (SERP) Menggunakan Classification and Regression Tree (CART)

Yanuar Ishaq
S1 Ilmu Komputasi, Fakultas Informatika, Telkom University, Bandung
ishaqyanuar@gmail.com

Abstrak

Pertumbuhan pesat internet beberapa tahun terakhir memunculkan berbagai macam media online seperti website, blog, dan social media. Dari waktu ke waktu jumlah website yang ada di dunia semakin banyak. Website menjadi salah satu media informasi, hiburan, promosi dan lain-lain. Salah satu indikator dari suksesnya sebuah website adalah trafik. Trafik dapat berasal dari berbagai macam sumber, yang paling dominan adalah trafik yang berasal dari search engine. Penelitian dalam tugas akhir ini bertujuan untuk mencari parameter penting sebuah halaman web dalam Google search engine result page (SERP). Metode yang digunakan dalam penelitian ini adalah Classification and Regression Trees (CART) untuk mendapatkan parameter-parameter yang berpengaruh terhadap peringkat hasil pencarian suatu halaman web pada Google SERP. Data yang digunakan adalah hasil pencarian 25 kata kunci atau keyword yang masing-masing hasil pencarian halaman web tersebut memiliki parameter-parameter. Parameter dari data tersebut lalu dimodelkan dengan Classification and Regression Trees dengan bantuan software Matlab. Dari hasil matlab diperoleh 2 parameter yaitu Page Authority dan Domain Authority.

Kata kunci : SERP, CART, PA, DA

Abstract

The rapid growth of the internet the past few years led to a wide variety of online media such as websites, blogs, and social media. From time to time the number of websites in the world more and more. The website became one medium of information, entertainment, promotions and others. One indicator of the success of a website is traffic. Traffic can come from a variety of sources, the most dominant is the traffic that comes from search engines. Research in this thesis aims to look for the important parameters of a web page in the Google search engine result page (SERP). The method used in this study is the Classification and Regression Trees (CART) to obtain parameters that influence the ranking of search results a web page on a Google SERP. The data used is the result of 25 search keywords or keyword that each web page of the search results have parameters. The parameters of the data was modeled by Classification and Regression Trees with the help of Matlab software. From the results obtained matlab 2 parameters: Page Authority and Domain Authority.

Key word : **SERP, CART, PA, DA**

1. Pendahuluan

Mencari informasi menggunakan search engine telah menjadi bagian dari kehidupan manusia sehari-hari. Ingin mencari informasi tentang gadget terbaru, restoran populer atau berita yang sedang populer, kebanyakan mereka akan mencari menggunakan search engine. Keberadaan search engine menjadi sumber tunggal atau utama dalam mengarahkan orang untuk mendapatkan informasi penting. Untuk alas an tersebut, search engine menempati "posisi terkemuka di dunia online"[1]. Karena banyaknya jumlah situs web yang ada di dunia, search engine memiliki tugas untuk menyortir halaman-halaman situs web dan menampilkan halaman website yang paling relevan di Search Engine Result Page (SERP) sesuai dengan permintaan pencarian yang diajukan. Dengan pertumbuhan internet dan jumlah situs web yang tersedia, akan menjadi sulit untuk mencari pengunjung situs web. Menurut sebuah penelitian ada sekitar 3 juta situs web baru yang muncul setiap bulan[2]. Studi lain menemukan bahwa lebih dari 80% dari kunjungan pertama ke sebuah situs web berasal dari search engine, dan lebih dari 76% menggunakan search engine Google[3]. Selain itu 84% dari pengguna search engine Google tidak pernah eralih ke halaman kedua dari hasil pencarian, dan 65% hampir tidak pernah memilih pada hasil pencarian dari sponsor[3]. Studi tersebut menunjukkan bahwa untuk mencapai halaman terdepan dalam mesin pencari adalah kunci untuk mendapatkan trafik yang akan meramaikan sebuah situs web. Oleh karena itu, mendapatkan posisi teratas dalam SERP mesin pencari sangat penting dan merupakan inti dan tujuan dari SEO.

CART (Classification and Regression Trees) adalah salah satu metode atau algoritma dari salah satu teknik eksplorasi data decision tree. Oleh karena itu judul tugas akhir ini adalah "Prediksi Google Search Engine Result Page (SERP) Menggunakan Classification and Regression Tree (CART)."

2. Dasar Teori

2.1 SERP (Search Engine Result Page)

Search engine seperti Google digunakan oleh sebagian besar pengguna internet dalam mencari suatu informasi. Melalui situs ini pengguna hanya menuliskan kata kunci untuk menemukan situs yang berhubungan atau relevan dengan informasi yang diharapkannya.Hasil pencarian berupa Search Engine Result Page (SERP) yang berisi daftar alamatwebsite yang terbagi dalam halaman-halaman. Namun kebiasaan dari pengguna internet hanya membaca daftar web yang ada di halaman pertama

2.2 CART

Classification tree adalah suatu metode klasifikasi atau pengelompokan yang berbentuk tree dari sekumpulan data ke dalam beberapa kelompok dengan menggunakan pemisahan tree.

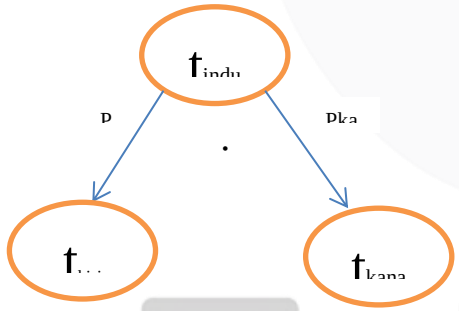
Classification tree digunakan untuk memprediksi objek atau kasus dalam kelas-kelas dengan satu variable dependent (terikat) yang kategorik dari satu atau lebih

variabel bebas.Classification tree dibangun berdasarkan splitting rule yang dipilih. Splitting rule adalah aturan pemisahan pada tree yakni aturan yang melaksanakan pemisahan data percobaan (learning sample) menjadi bagian yang lebih kecil dengan mencari kemungkinan didapatkan kehomogenitasan yang maksimum. Homogenitas maksimum adalah sebuah kondisi di mana pemisahan node berdasarkan kehomogenan kelas data sehingga pada terminal node akan didapatkan data yang lebih murni (pure). Kemurnian data yang didapat itulah yang dapat menentukan keakuratan sebuah prediksi. Semakin murni sebuah hasil yang didapat maka akan semakin akurat prediksinya, dan sebaliknya semakin tidak murni (impure) hasil yang didapat maka akan semakin tidak tepat prediksinya.

CART (Classification and Regression Trees) adalah salah satu metode atau algoritma dari salah satu teknik eksplorasi data decision tree. Metode ini dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen dan Charles J. Stone sekitar tahun 1980-an.

CART merupakan alat dari metode decision tree yang dapat dikatakan paling baik untuk memecahkan masalah data mining, pemodelan prediksi, danpengolahan data. Pada proses pengerjaannya, CART secara otomatis mencaripola-pola dan hubungan yang penting yaitu membuka struktur yangtersembunyi meskipun data yang digunakan memiliki tingkat kompleksitas yang tinggi. Dalam CART ada dua buah karakteristik penting yang harus diperhatikan untuk mendapatkan hasil tree dengan tampilan yang optimal.

Aturan pemisah tree pada metode CART digambarkan dalam diagram di bawah ini.



2.3 Algoritma pembentukan pohon klasifikasi

CART terdiri dari empat tahapan, yaitu:

1). Pemilihan pemilah (Classifier)
2). Penentuan simpul terminal
3). Penandaan label kelas
4). Penentuan pohon klasifikasi optimal

2.3.1.Pemilihan pemilah (Classifier)

Pada tahap ini dicari pemilah dari setiap simpul yang menghasilkan penurunan tingkat keheterogenan paling tinggi. Untuk mengukur tingkat keheterogenan suatu kelas dari suatu simpul tertentu dalam pohon klasifikasi dikenal dengan istilah impurity measure. Fungsi impuritas yang dapat digunakan didalam pembentukan pohon klasifikasi CART adalah Indeks Gini. Derajat impurity yang tinggi menunjukkan simpul tersebut belum homogen, sedangkan sebuah simpul dengan derajat impurity yang rendah menunjukkan simpul tersebut sudah homogen. Jika kelas obyek dinyatakan dengan k, k = 1,2,..,m, dimana m adalah jumlah kelas untuk variabel/output respon y, maka nilai impuritas dari simpul menggunakan Indeks Gini dapat dituliskan persamaannya sebagai berikut:

$$\text{Gini(t)} = 1 - \sum_{i=1}^{m} [P(k|t)]^2 \qquad (1)$$

dengan

[P(k|t)] = jumlah relatif dari kelas j pada simpul t

m = jumlah kelas

Jika nilai Indeks Gini, Gini(t)=0 maka semua data dari simpul tersebut sudah berada pada kelas yang sama (homogen). Misalkan dilakukan pemisahan (spliting) sebuah simpul menggunakan Indeks Gini. Jika simpul t di split kedalam k partisi (anak), maka kualitas split dihitung sebagai berikut:

$$Gini_{split} = \sum_{i=1}^{k} \frac{n_i}{n} Gini(t) \qquad (2)$$

dengan

ni = Jumlah record pada anak ke-i

n = Jumlah record pada simpul

2.3.1.Penentuan Simpul terminal

Suatu simpul t akan menjadi simpul terminal atau tidak akan dipilah kembali, apabila pada simpul t tidak terdapat penurunan keheterogenan secara berarti (sudah homogen) atau adanya batasan minimum n seperti hanya terdapat satu pengamatan pada tiap simpul anak. Menurut Breiman (Otok, 2009: XVI-3), pada umumnya jumlah kasus minimum dalam suatu terminal akhir adalah 5, dan apabila hal itu terpenuhi maka pengembangan pohon dihentikan. Sementara itu, menurut Steinberg dan Colla (Otok, 2009: XVI-3), jumlah kasus yang terdapat dalam simpul terminal yang homogen adalah kurang dari 10 kasus.

2.3.2. Penandaan label kelas

Penandaan label kelas pada simpul terminal dilakukan berdasarkan aturan jumlah terbanyak. Misalkan pada kasus klasifikasi keputusan pembelian computer (ya, tidak), dalam salah satu simpul terminal yang dihasilkan terdapat jumlah keputusan ya dan keputusan tidak. Jumlah terbanyak dari keputusan tersebut dijadikan label kelas simpul terminal.

2.3.3. Penentuan Pohon Klasifikasi Optimal

Pohon klasifikasi yang berukuran besar akan memberikan nilai penaksir pengganti paling kecil, sehingga pohon ini cenderung dipilih untuk menaksir nilai dari variabel respon. Tetapi ukuran pohon yang besar akan menyebabkan nilai kompleksitas yang tinggi, karena struktur data yang digambarkan cenderung kompleks, sehingga perlu dipilih pohon optimal yang berukuran sederhana tetapi memberikan nilai penaksir pengganti cukup kecil. Ada dua jenis penaksir pengganti, yaitu penaksir sampel uji (test sample estimate) dan penaksir validasi silang lipat (cross validation K-fold estimate).

Validasi silang merupakan salah satu teknik untuk menduga error rate. Beberapa teknik yang lain diantaranya adalah: holdout, leave one dan bootstrapping. K-fold cross validation membagi data menjadi k bagian terpisah, satu data menjadi data testing dan k-1 bagian menjadi data training sehingga terdapat k pasang data training-testing. K-fold cross validation dapat digunakan untuk data berukuran kecil ataupun besar. Aspek terpenting dalam validasi silang adalah kestabilan dari penaksiran yang diperoleh. Kestabilan pohon dapat bernilai rendah, jika mengandung terlalu banyak variabel prediktor.

Salah satu cara untuk mendapatkan pohon optimum yaitu dengan

pemangkasan (pruning). Pemangkasan dilakukan dengan jalan memangkas bagian pohon yang kurang penting sehingga didapatkan pohon optimal. Ukuran pemangkasan yang digunakan untuk memperoleh ukuran pohon yang layak adalah cost complexity minimum.

Sebagai ilustrasi, untuk sembarang pohon T yang merupakan sub pohon dari pohon terbesar Tmax(T<Tmax) ukuran cost complexity yaitu:

$$R_a(T_k) = R(T_k) + \alpha|\tilde{T}_k|$$

3. Perancangan Program

3.1 Parameter SEO

Parameter Search Engine Optimization yang digunakan dalam penyusunan tugas akhir ini;

1. Page Authority (PA)
2. Domain Authority (DA)
3. Title Tag
4. Description Tag
5. Link to Page

3.2 Pengumpulan Data

Pengumpulan data menggunakan search dari Google.co.id dengan kata kunci tertentu. Hasil search kemudian dicari parameternya menggunakan moz tools.

3.3 Membangun Program

Proses processing data dari Data Kata Kunci TA.xls

kunci=[{'mainan'},{'sampah'},{'makanan'},{'camilan'},{'dekorasi'},{'batik'},{'jerawat'},{'komputer'},{'fotografi'},{'drone'},{'kanker'},{'saham'},{'danareksa'},{'valas'},{'obligasi'},{'syariah'},{'wisata'},{'grafis'},{'andorid'},{'parfum'},{'asuransi'},{'pajak'},{'belajar'},{'banjir'},{'keluarga'},{'budaya'}];

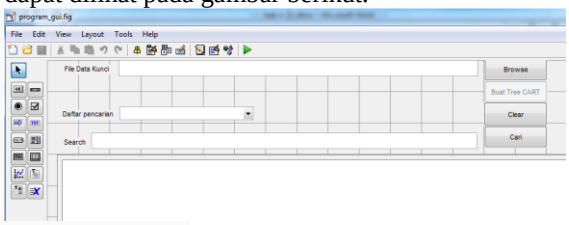
Untuk maksud tersebut, maka data yang bisa diubah menjadi data CART dipilih data pada kolom “DA”, “PA” dan “Link To Page”.

4. Pengujian Sistem Program

4.1 Program CART Matlab

Setelah perancangan diselesaikan sesuai prosedur algioritma CART maka dilakukan pengujian program sebagai berikut.

4.1.1 Langkah awal adalah masuk dalam GUI softtware Matlab R2008A yakni dengan melakukan perintah program_gui. Hasilnya dapat dilihat pada gambar berikut.



Gambar 4.1 GUI Program CART

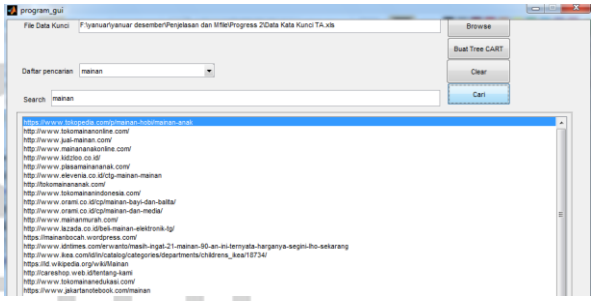
4.1.2 Pada gambar diatas terdapat tools browse yakni untuk mengambil data dari datasheet excel yang berasal dari Google SERP.

4.1.3 Pada Tools berikutnya terdapat kolom daftar pencarian untuk mengambil kata yang akan dieksekusi oleh search engine

4.1.4 Tools untuk eksekusi adalah tools cari.

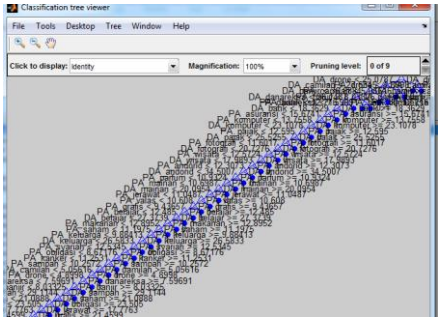
4.1.5 Dan Tools clear untuk mereset ulang

4.1.6 Hasil pengujian adalah sebagai berikut



Gambar 4.2 Hasil Keyword Makanan

4.1.7 Kemudian untuk mengetahui CART dilakukan



dengan klik tools CART yang hasilnya dapat dilihat pada gambar berikut.

4.1.7.1 Hasil Pengujian

Dari proses CART dataset beberapa kunci diatas dapat disimpulkan bahwa parameter yang berpengaruh terhadap hasil Google SERP pada sebuah website adalah *Page Authority* dan *Domain Authority*. Pada pengujian kali ini penulis akan mengetes beberapa kata kunci dan melihat apakah pada hasil Google SERP website tersebut memiliki kedua parameter tersebut.

4.1.8 Kata Kunci 1 (Sepeda)

Kata kunci yang pertama pada pengujian adalah ‘sepeda’ dan berikut ini hasil Google SERP

Posisi	URL	PA	DA
1	http://www.lazada.co.id/olahraga-sepeda/	31.5775166118469	56.2355607588808
2	http://www.lazada.co.id/sepeda/	20.9008398456575	56.2355607588808
3	https://www.tokopedia.com/p/olahraga/sepeda/sepeda	32.3393731374052	63.857656007717
4	https://www.bukalapak.com/c/sepeda	17.4070343831851	63.1555188214751
5	http://www.polygonbikes.com/id/bikes	16.9483220792864	33.2852108521829

Dari hasil Google SERP diatas dapt kita lihat bahwa Website yang tampil pada hasil pencarian Google masing-masing memiliki PA dan DA.

Posisi	URL	PA	DA
111	http://www.har.com/100-sepeda-dilley-tx-78017/homevalue_10804818	1	58.4098088182097
112	https://www.linkedin.com/pulse/blog-competition-wimcycle-sepeda-terbaikku-urip-aminoto	1	100
113	https://www.boardmanbikes.com/id_id/bikes/	1	45.23248544
			70712
114	https://www.evi.com/q/what_does_%22sepeda%22_mean_in_indonesian	1	42.8597141817698

Sedangkan hasil pencarian pada halaman 11 Google didapatkan hasil diatas dengan Website yang tidak memiliki Page authority (PA 1 memiliki arti bahwa halaman web tersebut tidak memiliki Page Authority)

4.1.9 Kata Kunci 2 (Desa)

Kata kunci yang pertama pada pengujian adalah ‘desa’ dan berikut ini hasil Google SERP

Posisi	URL	PA	DA
1	https://map-bms.wikipedia.org/wiki/Desa	29.5595607392996	100
2	https://desa.web.id/	35.3958361061185	27.3148365050799
3	https://domain.go.id/PendaftaranDomainDesa.pdf	19.3550490060533	14.2528301132137
4	http://prodeskel.binapemdes.kemendagri.go.id/	34.8347494882007	53.3991399129378
5	http://desa.kemendesa.go.id/	39.8262791723655	39.9525075820604

Dari hasil Google SERP diatas dapt kita lihat bahwa Website yang tampil pada hasil pencarian Google masing-masing memiliki PA dan DA.

Posisi	URL	PA	DA
111	http://www.casadelrio-melaka.com/media-news/sajian-desa-media-preview	17.4990505455716	38.2217922518651
112	https://www.brilio.net/wow/11-foto-menakjubkan-desa-di-dalam-gua-ada-sekolahannya-pula-1608134.html	1	41.4679700326722
113	http://www.pustaka.ut.ac.id/lib/2016/08/08/ipem4208-sistem-pemerintahan-desa/	1	44.3435580457308

114	https://buanaindonesia.com/news/aceh/2017/03/25/4537/	1	10.36661657054
115	http://manadopostonline.com/read/2017/03/24/20-Desa-Dipanggil-Inspektorat/21627	1	28.9238810612015

Sedangkan hasil pencarian pada halaman 11 Google didapatkan hasil diatas dengan halaman web yang tidak memiliki Page authority (PA 1 memiliki arti bahwa halaman web tersebut tidak memiliki Page Authority)

PA dan DA keduanya penting untuk sebuah website untuk dapat tampil di Google SERP, DA atau domain authority lebih berpengaruh terhadap Google SERP karena mencerminkan kekuatan sebuah website sehingga lebih diprioritaskan dalam hasil pencarian. Namun, PA atau page authority juga penting pada sebuah halaman web jika webmaster ingin menekankan konten atau keyword tertentu dalam hasil pencarian. Kuncinya adalah menggunakan DA dan PA secara strategis, focus pada DA atau domain authority jika ingin membangun brand atau merek, dan fokus pada PA jika ingin mendapatkan halaman web dengan hasil yang baik dalam pencarian pada keyword yang spesifik.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari seluruh proses prediksi Google Search engine result page menggunakan Classification and Regression Trees adalah sebagai berikut.

1. Parameter yang mempengaruhi posisi sebuah halaman web pada hasil pencarian Google Search Engine Result Page (SERP) adalah Page Authority dan Domain Authority.
2. Semakin tinggi Domain Authority dan Page Authority yang dimiliki sebuah halaman web akan mempertinggi posisi halaman web dalam pencarian Google SERP.

5.2. Saran

Untuk mendapatkan model dan prediksi hasil Google Search Engine Result Page (SERP) yang lebih akurat terdapat beberapa hal yang bisa dijadikan saran dan sebagai bahan pertimbangan antara lain:

1. Menambah parameter yang digunakan seperti deskripsi halaman web, judul halaman web, dan link keyword sehingga lebih akurat.
2. Hasil CART pada tugas akhir ini dapat digabung dengan teknik lain sehingga lebih mendekati hasil Google SERP.

DAFTAR PUSTAKA

[1] Wang, et al. "An Empirical Study on Search Engine Optimization Techniques and Its Outcomes." IEEE, 2011.
[2] Grzywaczewski, et al. "E-Marketing Strategy for Businesses." IEEE, 2010.
[3] Zhu, et al. "Research and Anaysis of Search Engine Optimization Factors Based on Reverse." IEEE, 2011.
[4] Ellsworth, Jill H. The Internet business book. Wiley, 1994.
[5] Cheffey, D., Ellis-Chadwick, F., Mayer, R., Johnston, K. "Internet marketing Strategy, Implementation and Practice." Prentice Hall, 2006: 349.
[6]Dover, Danny: Search Engine Optimization Secrets. Wiley Publishing, Inc., Indianapolis,
[7] Angarini, Dini. Klasifikasi Kondisi Kesehatan Jantung Menggunakan Metode Multiple Discriminant Analysis (MDA) Dan Classification and Regression Tree (CART). Jakarta : Universitas Islam Negeri. 2007.
[8] Timfee, Roman. Classification And Regression Tree (CART) Theory and Application. Master Thesis. CASE – Center of Applied Statistics and Aconomics. Humboldt University. 2003.
[9] Rakesh Kumar.” A Study on SEO Monitoring System Based on Corporate Website Development.” International Journal of Computer Science, Engineering and Information Technology.2011
Breiman, L., Friedman, J., Olsen, R.A., dan Stone, C. (1984), Classification and regression trees, Wadsworth, Belmont, California.