# IMPLEMENTASI *DECISION TREE* PADA *SOCIAL NETWORK*

**Yunda Heningtyas**
Jurusan Teknik Informatika Institut Informatika dan Bisnis Darmajaya
Jl. Z.A. Pagar Alam No. 93 Labuhan Ratu Bandar Lampung
e-mail: yunda.heningtyas89@gmail.com

***Abstract***

*The strategy of business today is no longer done traditionally. By the presence of internet, marketing strategies can be done with the wider reach. Viral marketing is a new program in an electronic format that was created to assist in the marketing of products by making the consumer as an object to spread the advertising of a product. Abundant data from the social network has the potential information that can be utilized for virtual marketing. Social network analysis using a decision tree will produce a potential users and a pattern that could affect the user is to distribute advertising the product. Result is rules that contains the attributes which can influence the pattern indicate users to execute of the strategy business of viral marketing.*

***Key words****: social network, decision tree, virtual marketing, snowball sampling.*

## PENDAHULUAN

Strategi bisnis saat ini tidak lagi dilakukan secara tradisional. Berkembangnya internet memberikan cangkupan pemasaran menjadi sangat luas serta dapat dilakukan kapan saja dan dimana saja, tidak hanya tergantung pada tempat dan waktu. Strategi pemasaran tidak lagi dibatasi oleh lautan dan batas-batas negara. Internet menyediakan sebuah media *online* dimana penggunanya dapat berpartisipasi, berbagi, dan menciptakan isi berupa blog, *social network*, wiki, forum, dan dunia *virtual*. *Website Community* atau sering disebut *Social Network* merupakan tempat untuk para netter menjalin hubungan dengan orang lain melalui *social media sites* di internet. Dua pertiga populasi dari dunia maya selalu mengunjungi situs *social network*. Jumlah *messages* dan *content* yang dibagi oleh pengguna *social network* meningkat secara signifikan dari sebelumnya.

Artikel Yin Gui-sheng et al [1] membahas tentang *social network* yang menjadi fokus dalam *viral marketing*. Tantangan yang dibahas dalam artikel ini adalah bagaimana memilih sejumlah pengguna awal dalam seluruh populasi dengan tujuan memaksimalkan keuntungan dengan pendekatan diskrit. Gui-Sheng mencoba menerapkan algoritma *intellegent* seperti GA, DE, PSO untuk mencari "bibit" yang dapat menyelesaikan *task* ini dan mengatasi masalah skalabilitas data. Analisis data dilakukan pada dua situs *social network* yaitu www.squeak.org dan http://robots.net. Komponen yang diperhatikan dari kedua situs tersebut adalah distribusi konektifitas. Hasil yang diperoleh adalah algoritma *intellegent* yang digunakan untuk memecahkan masalah *virtual marketing* lebih baik daripada algoritma konvensional.

Model *social network* telah lama dideskripsikan dan hanya dibangun dengan parameter global yang tidak berguna untuk membuat prediksi prilaku jaringan di masa depan. Hal ini terjadi karena kurangnya data, jaringan yang tersedia kecil dan sedikit, serta hanya berisi sedikit informasi minimal tentang setiap *node*. Namun internet muncul dengan membawa sejumlah besar data pada *social network*. Data yang sangat besar tersebut memungkinkan dibuatnya sebuah model baru yang lebih detail dari model sebelumnya dalam analisis *social network*. Model *social network* yang dibuat menggunakan data dari situs www.epinions.com. Model ini memungkinkan untuk merancang rencana "*virtual marketing"* yang memaksimalkan kemungkinan "*word-of-mouth"* antar *customer*. Hasil percobaan adalah kemungkinan untuk mencapai keuntungan jauh lebih tinggi daripada mengabaikan interaksi antara pelanggan dan efek jaringan yang sesuai, seperti pemasaran tradisional [2].

Mislove et al [3] memaparkan studi pengukuran data dalam skala besar dan analisis struktur dari beberapa *social network online*

yaitu YouTube, Flickr, LiveJournal, dan Orkut. *Link* pengguna dijelajahi berdasarkan akses publik dari semua situs. Kumpulan data berisi lebih dari 11,3 juta pengguna dan 328 juta *link*. Hasil yang diperoleh menegaskan kekuatan hukum, *small word*, dan sikap *scalefree social network online*. *Node indegree user* cenderung sesuai dengan outdegree. Pengguna situs ini membentuk *social network*, yang menyediakan sarana yang kuat dari berbagi, pengorganisasian, serta mencari konten dan kontak. Jaringan mengandung inti kepadatan yang terhubung dengan *node* derajat yang tinggi dimana inti *link* terdiri dari kelompok-kelompok kecil. *Node* dengan derajat lebih rendah berada di tepi jaringan.

YouTube memiliki lebih dari satu miliar pengguna pengguna unik di seluruh dunia, belum termasuk video yang dipasang atau dilihat melalui perangkat *mobile* [4]. Lebih dari 100 juta orang berpartisipasi secara sosial di YouTube, seperti melakukan *likes*, membagikan video, atau berkomentar setiap pekannya. Pencapaian yang luar biasa ini membuat situs www.YouTube.com menjadi situs yang sangat potensial untuk proses pemasaran terutama untuk *viral marketing*. Ketersediaan data dalam YouTube menunjukan kumpulan data yang sangat besar dengan puluhan juta pengguna, masing-masing dijelaskan oleh puluhan atribut [5]. Data tersebut memiliki informasi yang berlebihan dan informasi yang tidak berguna untuk identifikasi pengguna yang relevan untuk analisis tertentu.

Masalah ini dapat diatasi dengan solusi parsial yaitu dengan membuang beberapa atribut yang tidak relevan sehubungan dengan analisis spesifik. Namun, muncul permasalahan lain yaitu beberapa atribut memiliki informasi yang berlebihan atau mengidentifikasi pengguna yang relevan untuk analisis tertentu.

Untuk mengatasi permasalahan tersebut, paper ini menggunakan salah satu metode klasifikasi yaitu *decision tree*. Konsep *decision tree* adalah mengubah data menjadi pohon keputusan dan aturan-aturan (*rule*) keputusan sehingga dapat diketahui pola-pola yang menarik. Tujuan akhir dari penelitian ini adalah mendapatkan *user* yang potensial dan suatu pola yang dapat mempengaruhi *user* tersebut untuk menyebarkan iklan produk.

## METODE

Konsep yang akan dibahas secara singkat tentang situs *social network* yaitu YouTube. YouTube menyediakan forum bagi pengguna untuk menjalin hubungan, berbagi informasi, dan menginspirasi orang di seluruh dunia dan berfungsi sebagai platform distribusi bagi pembuat konten asli dan pengiklan baik besar maupun kecil. Fasilitas yang disediakan oleh YouTube sangat memungkinkan untuk penyebaran informasi yang meluas ke seluruh dunia, terutama iklan produk atau jasa. Namun tidak semua iklan dapat tersebar dengan luas, hanya iklan dengan tema yang menarik bagi pengunjung situs yang akan tersebar secara meluas. Semakin menarik suatu iklan, semakin banyak pengunjung situs yang melihat sehingga tujuan dari iklan tersebut dapat tercapai. Hal ini sesuai dengan konsep *virtual marketing* yaitu "*word of mouth*". Pendekatan *Sowball Sampling* [6] digunakan untuk menentukan atribut yang digunakan dalam penerapan metode id3 untuk *database* YouTube.

### Pendekatan *Snowball*

*Snowball Sampling* adalah teknik survei iteratif yang bertujuan untuk mengungkapkan informasi tentang struktur jaringan dengan contoh bagian dari simpul dan tepi. *Social network* dimodelkan sebagai sebuah grafik tidak berarah dan tidak memiliki nilai dimana simpul mewakili individu dan ujung-ujungnya merupakan hubungan antara individu. Simpul dilambangkan oleh V, dengan ukuran N. Simpul terhubung ke dalam $v \epsilon V$ dengan tepi yang disimbolkan oleh K(v). Individu yang ditampilkan oleh simpul disebut *ego*.[7]

Pendekatan *Snowball* mendefinisikan prosedur iterasi survei dimana set awal adalah memilih simpul secara acak dari jaringan dan diminta melaporkan hubungan sosial mereka. Dalam setiap iterasi, individu-individu yang diwawancarai dipilih dari himpunan *user* yang telah dilaporkan sebagai kontak sosial untuk pertama kalinya pada iterasi sebelumnya. Struktur *social network* yang diselidiki mempengaruhi *user* yaitu himpunan keseluruhan *user* yang diwawancarai bukan sampel acak dari seluruh *user*.

### Konsep *Decision tree*

Konsep yang digunakan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (*rule*). ID3 (*Iterative Dichotomiser* 3)

merupakan salah satu algoritma yang digunakan untuk membuat *decision tree*. Algoritma ID3 dapat diringkas sebagai berikut:

1. Mengambil semua atribut dan menghitung entropi.
   Pemilihan atribut pada algoritma ini menggunakan ukuran berdasarkan *entrophy* yang dikenal dengan *Information Gain*. Menghitung *entrophy* atribut menggunakan rumus di bawah ini.
   $$E(S) = -p_+ log_2 p_+ - p_- log_2 p_-$$
   Adapun menghitung *Information Gain* dengan rumus berikut.
   $$Gain(A) = I_{(S_1, S_2, \ldots, S_m)} - \sum_{i=1}^{m} p_i * E(S_i)$$
2. Pilih atribut yang *entrophy* minimum
   Atribut dengan nilai *entrophy* minimum memiliki nilai *entrophy* minimun akan menyebabkan nilai dari informasi Gain maksimum. Nilai *entrophy* minimum akan dipilih untuk menjadi simpul dari *tree*.
3. Buat simpul yang mengandung atribut tersebut
   Atribut yang memiliki nilai informasi Gain terbesar merupakan simpul awal (*root*). Atribut terbesar kedua akan menjadi simpul kedua, dan seterusnya.

**Identifikasi Sumber dan Target *User***

Dataset YouTube diperoleh dengan menjelajahi grafik *social network* dengan memanfaatkan API yang disediakan oleh YouTube. Pendekatan ini umumnya digunakan dalam literatur dan memberikan akses ke data set yang besar. Penjelajahan diterapkan dengan mengumpulkan data tentang pengguna YouTube dan *Link*s sosial menggunakan pendekatan *Snowball*. Penjelajahan dimulai dari daftar *user* yang terpilih secara random. Penjelajah mengeksplorasi *Link* keluar pengguna yang belum dikunjungi untuk mendapatkan daftar kontaknya, ditambahk-an pada daftar pengguna yang akan dikunjungi pada langkah berikutnya. Tabel 1 merupakan atribut *user* ($X_1, X_2, \ldots, X_{10}$) yang diperoleh dari YouTube dan penjelasan masing-masing atribut.

Atribut tambahan tidak dikumpulkan oleh penjelajah karena dianggap tidak menarik dalam konteks pemilihan pengguna yang relevan untuk *marketing*. Atribut dalam tabel mengampilkan dua tipe informasi yaitu data tentang pengguna *Link* sosial (*subscribers, subscriptions* dan *friends*) dan data tentang interaksi *user* dengan konten yang di *shared* (*views, downloads, uploads, favorites*). Pada dasarnya, atribut tersebut mencerminkan informasi sebenarnya diberikan oleh situs YouTube sementara daftar *link* pada versi penjelajah dipotong oleh keterbatasan API YouTube. Kumpulan data yang diperoleh dari proses penjelajahan terdiri dari sebelas atribut untuk hampir dua juta pengguna. Jumlah atribut yang tinggi tersebut tidak memungkinkan untuk menggunakan dataset langsung untuk mengidentifikasi *user* yang relevan untuk *marketing* pada *social network*.

Tabel 1. Atribut *User* [5]

| | **Atribut** | **Deskripsi** |
|---|---|---|
| $X_1$ | *Subscribers* | Jumlah pengguna yang berlangganan ke *user channel* (diperoleh dari pengumpulan dan penghitungan *Link*) |
| $X_2$ | *Subscribers* | Jumlah pengguna yang berlangganan ke *user channel* (diperoleh dari profil *user*) |
| $X_3$ | *Subscriptions* | Jumlah langganan dari *user* (diperoleh dari pengumpulan dan penghitungan *link*) |
| $X_4$ | *Subscriptions* | Jumlah langganan dari *user* (diperoleh dari profil *user*) |
| $X_5$ | *Friends* | Jumlah *friends user* dalam *social nertwork* |
| $X_6$ | *Views* | Jumlah *view* untuk *upload* video oleh *user* |
| $X_7$ | *Downloads* | Jumlah video yang ditonton oleh *user* |
| $X_8$ | *Uploads* | Jumlah video yang di-*upload* oleh *user* |
| $X_9$ | *Favorites* | Jumlah video yang ditandai sebagai video yang disukai |

| | | |
|---|---|---|
| $X_{10}$ | *Rates* | Jumlah video yang diberi peringkat oleh *user* |

## HASIL DAN PEMBAHASAN

Proses pemilihan dan pengumpulan data dilakukan secara random menggunakan pendekatan *snowball sampling*. Data yang digunakan adalah 1000 data dengan *class target* adalah Yes dan No. Atribut yang diambil sesuai dengan Tabel 1. Data tersebut diproses menggunakan algoritma ID3 untuk mendapatkan pola yang dapat membantu proses *viral marketing*. Penerapan algoritma ID3 dalam proses ini ditampilkan dalam pseudocode berikut.

Pseudocode untuk mencari nilai *entrophy* dari masing-masing atribut

```
public double hitung(Vector
data) {
  int n = data.size();
  if (n == 0)
        return 0;

  int atrib = jmlhAtribut-1;
  int          jmlh          =
  domains[atrib].size();
  double Es = 0;
  for (int i=0; i< jmlh; i++)
  {
        int count=0;
        for (int j=0; j< n;
        j++) {
              DataPoint point
              =
              (DataPoint)data.
              elementAt(j);
              if
              (point.atribut[a
              trib] == i)
                    count++;
        }
        double p = count/n;
        if (count > 0)
        Es += -p*Math.log(p);
  }
}
```

Pseudocode untuk mencari nilai informasi Gain dari masing-masing atribut.

```
if (pil == false) {
      pil = true;
      pilih_entrophy =
      entrophy_S;
      pilih = i;
} else {
      if (entrophy_S <
      pilih_entrophy) {
            pil = true;
            pilih_entrophy =
            entrophy_S;
            pilih = i;
      }
}
```

Pseudocode untuk menentukan atribut yang akan dijadikan simpul.

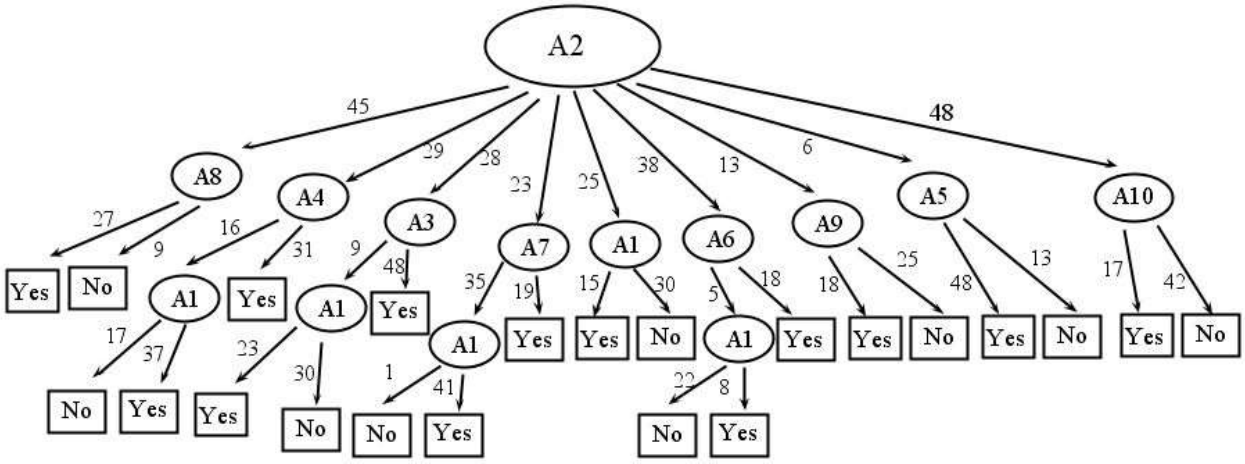
```
if (pil == false) return;

int jmlh =
domains[pilih].size();
node.pecahAtribut = pilih;
node.cabang = new pohon [jmlh];

for (int j=0; j< jmlh; j++) {
      node.cabang[j] = new
      pohon();
      node.cabang[j].akar =
      node;
      node.cabang[j].data =
      kum_data(node.data,
      pilih, j);
      node.cabang[j].pecahnilai
      = j;
}

for (int j=0; j< jmlh; j++) {
      pecahanNode(node.cabang[j]
      );
}
node.data = null;
```

Atribut yang diproses akan menghasilkan pohon dengan 3 level dimana level pertama berisi *root*, level kedua berisi *node* internal dan *leaf*, level ketiga berisi *leaf*. Hasil dari penerapan algoritma ID3 dapat dilihat pada Gambar 1. Hasil program menunjukan atribut yang digunakan sebagai *root* adalah *Subscribers* yang merupakan jumlah pengguna yang berlangganan ke *user channel* yang diperoleh dari profil *user*. Atribut lainnya berfungsi sebagai *node internal* dimana *leaf* adalah "Yes" dan "No".




Gambar 1. Atribut *Tree*

## KESIMPULAN

Penerapan metode ini pada data-data YouTube untuk mendapatkan *rules* yang baru dan dapat dimanfaatkan untuk proses *viral marketing*. Pengguna YouTube diidentifikasi untuk mengetahui *user* yang paling relevan untuk *marketing*. Identifikasi sumber dan target untuk isi diseminasi. Selanjutnya, mengklasifikasikan sasaran populasi kepada pengguna yang lebih tertarik pada *content tags, rank* dan *metadata* lainnya. Wawasan ini membantu proses pemasaran berbasis internet terutama *viral marketing* karena memungkinkan untuk menemukan pengguna yang paling sesuai untuk membantu menjalankan strategi diseminasi.

## DAFTAR PUSTAKA

[1] Yin Gui-sheng, Wei Ji-jie, Dong Hong bin, and Li Jia. 2011. Intelligent *viral marketing* algorithm over *online social network. In Proc. Second Int Networking and Distributed Computing (ICNDC) Conf, pages 319-323*.

[2] Soumya Banerjee, Hameed Al-Qaheri, and Aboul Ella Hassa-nien. 2010. Mining social networks for viral marketing using fuzzy logic. *In Proc. Fourth Asia Int Mathematical/Analytical Mo-delling and Computer Simulation (AMS) Conf, pages 24-28*.

[3] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. *Measurement and analysis of online social network*s. ACM 978-1-59593-908-1/07/0010.

[4] Statistics YouTube. 2015. YouTube statistics. https://www.youtube.com/yt/press/id/statistics.html. diakses tanggal 10 Januari 2015.

[5] C. Canali, S. Casolari, and R. Lancellotti. 2010. A quantitative me-thodology to identify relevant *user*s in *social network*s. *In Proc. IEEE Int Business Applications of Social network Analysis (BASNA) Workshop, pages 1-8*.

[6] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. 2007. Measurement and analysis of *online* social ne-tworks. *In Proc. of the 7th ACM SIGCOMM Conference on Internet measurement (IMC07)*.

[7] J. Illenberger ,Gunnar Fltterd, dan Kai Nagel. 2008. *An approach to correct biases induced by Snowball Sampling*.