Analisis dan Implementasi Graph Clustering pada Berita Online Menggunakan Algoritma Chinese Whispers

Moch. Fitrah Eka P[1] , Kemas Rahmat Saleh W, S.T., M.Eng [2], Anisa Herdiani, S.T., M.T [3]
[1.2.3] *Teknik Informatika, Fakultas Teknik Informatika, Telkom University Jalan Telekomunikasi No.1, Dayeuh Kolot, Bandung 40257* fitrah.eka@gmail.com[1], bagindok3m45@gmail.com[2], anisaherdiani@gmail.com[3]

Abstrak

Berita *online* saat ini merupakan sesuatu yang sangat umum dikalangan masyarakat Indonesia. Data berita *online* yang telah tersimpan pada suatu penyimpanan data mencapai ratusan miliar berita. Oleh sebab itu diperlukan suatu permodelan, agar memudahkan proses pencarian, manipulasi atau pengolahan data tersebut. Salah satu model yang sangat cocok untuk data berita tersebut adalah model *graph*. Untuk memudahkan pembaca maka berita *online* tersebut perlu dikelompokkan berdasarkan keterkaitan isi beritanya. Salah satu metode yang bisa dimanfaatkan untuk mengelompokkan berita adalah dengan *graph clustering*.

Sebelum melakukan *graph clustering*, data berita *online* harus diubah menjadi model *graph*. Langkah pertama untuk mengubah data berita kebentuk *graph* adalah melakukan *preprocessing*, lalu dihitung bobot keterkaitan isi beritanya dengan memanfaatkan *cosine similarity*, setelah itu bobot hasil *cosine similarity* dinormalisasi untuk dijadikan *edge* yang menghubungkan *node* dokumen berita. Setelah berbentuk *graph*, barulah dilakukan proses *graph clustering*. Dalam penelitian ini algoritma *graph clustering* yang digunakan adalah *Chinese Whispers*, karena *Chinese Whispers* mampu membentuk *cluster* dari data *graph* yang besar dengan waktu yang relatif cepat, sehingga sangat cocok digunakan untuk kasus *clustering* berita *online*.

Pada penelitian ini telah diuji performasi algoritma *Chinese Whispers* dari segi kualitas serta tingkat akurasi *cluster* yang dihasilkan. Setelah dilakukan pengujian diperoleh bahwa kuliatas hasil *cluster Chinese Whisper* cukup bagus karena hampir 95% *node* hasil *cluster* sudah memiliki nilai *intra-cluster* yang lebih tinggi dari pada *inter-cluster*-nya, sedangkan rata-rata akurasi dari proses *clustering* menggunakan algoritma *Chinese Whispers* adalah 80.0 %.

Kata kunci : *Graph, Graph Database, Clustering, Graph Clustering, Chinese Whispers*

1. Pendahuluan

Berita *online* saat ini merupakan sesuatu yang sangat umum di kalangan masyarakat Indonesia bahkan dunia. Hal ini disebabkan karena berita *online* merupakan sarana penyebaran informasi yang mudah dan cepat serta fleksibel, dapat diakses kapan saja, dan dimana saja [1].

Data berita *online* yang telah tersimpan pada suatu penyimpanan data mencapai puluhan hingga ratusan miliar berita. Dari miliaran data berita tersebut diperlukan suatu permodelan, agar memudahkan proses pengolahan data tersebut. Salah satu model yang cocok untuk data berita adalah model *graph*, karena dengan model *graph* akan lebih mudah dimengerti oleh pikiran manusia, selain itu model *graph* juga cocok digunakan untuk data yang berskala besar dan juga yang memiliki sifat semi terstruktur [2] [3] seperti data berita *online*.

Pengelompokkan berita berdasarkan kemiripan/keterkaitan isi beritanya adalah salah satu kebutuhan dalam berita *online*. Untuk itu diperlukan sebuah metode yang disebut *clustering*, pada dasarnya *clustering* sendiri bertujuan untuk membuat kelompok-kelompok berdasarkan kemiripan elemen di dalamnya [4] [5]. Salah satu teknik *clustering* yang ada adalah *graph clustering*. *Graph clustering* adalah proses pengelompokan *node* dalam sebuah *graph*, sehingga *node* tersebut berada dalam satu *cluster*/kelompok dengan *node* lainnya dengan sifat yang sama [6].

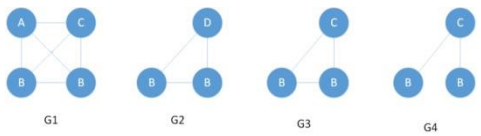
Algoritma *graph clustering* sangatlah banyak, diantaranya MST *Clustering*,

Chameleon, *Makarov Clustering*, dan *Star Clustering* [7]. Namun, dalam penelitian ini algoritma yang digunakan adalah *Chinese Whisper*. Hal ini dikarenakan algoritma *Chinese Whisper* mampu membentuk *cluster* dari data *graph* yang besar dengan waktu yang relatif cepat [8], sehingga sangat cocok digunakan untuk kasus *clustering* berita *online*.

2. Landasan Teori

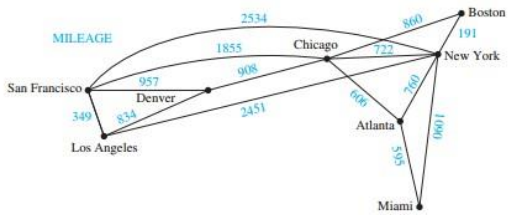
A. Teori Graph

Secara definisi, sebuah *graph* terdiri dari kumpulan dari *vertex* (V) dan *edge* (E) dan dapat disimbolkan G = (V, E). Berdasarkan keberadaan arah pada *edge* pada sebuah *graph* terdiri dari *direct graph* dan *indirect graph*. *Direct graph* memiliki sebuah *vertex* yang menjadi *start* dan sebuah *vertex* yang menjadi *end*, sedangkan *undirect graph* tidak memiliki *vertex* yang menjadi *start* dan *end* [9]. Contoh *simple undirect graph* dapat dilihat pada gambar 2.1.



Gambar 2.1 Contoh *simple undirect graph*

Selain itu *graph* juga dibedakan menjadi *graph* berbobot dan *graph* tidak berbobot. Disebut *graph* berbobot jika suatu *graph* memiliki angka/nomor yang dituliskan di setiap *edge* pada *graph* tersebut [9]. Namun, jika semua bobot pada *graph* tersebut bernilai 1 maka *graph* tersebut bisa dikatakan *graph* tidak berbobot [8]. Contoh *graph* berbobot bisa dilihat pada gambar 2.2



Gambar 2.2 Contoh *graph* berbobot

B. Graph Clustering

Cluster analysis adalah salah satu studi matematika yang bertujuan untuk mengetahui *natural group* dengan berdasarkan beberapa kesamaannya. Sedangkan proses untuk mengindetifikasi kesamaan suatu elemen data disebut dengan *Clustering* [10]. *Graph Clustering* adalah pengelompokan *node* dalam sebuah *graph*, sehingga *node* tersebut berada dalam satu *cluster*/kelompok yang memiliki kesamaan sifat [6]. Beberapa ahli juga berpendapat bahwa *graph clustering* adalah proses pengelompokan *node* pada sebuah *graph*, sehingga *node* tersebut memiliki nilai *inta-connectivity* yang tinggi terhadap *node* lainnya yang berada dalam *cluster* yang sama serta memiliki nilai *inter-connectivity* yang rendah [6]. Biasanya *cluster analysis* atau *graph clustering* digunakan untuk bidang biologi khususnya dalam mempelajari gen, VLSI *chip design*, *social network*, dan lainnya [10].

C. Chinese Whispers

Chinese Whispers (CW) adalah algoritma yang sangat mudah dan efektif dalam mengelompokkan/ *clustering node* dari *weighted, undirected graphs*. Ide dari CW bermula dari permainan anak-anak, dimana sang anak harus menyebarkan sebuah kata/kalimat dengan cara berbisik. Langkah-langkah dalam algoritma CW adalah sebagai berikut [8] :

1. Setiap *node* dalam *graph* tersebut di pisahkan dalam *cluster* yang berbeda.
2. Tunjuk satu *node* secara acak, lalu masukkan *node* tersebut kedalam *cluster* tetangga dengan nilai tertinggi. Nilai tertinggi dihitung dengan melihat total bobot keterhubungan dengan *cluster* lainnya.
3. Lakukan pengulangan proses 2 hingga semua *node* dalam *graph* tersebut selesai di tunjuk.
4. Lakukan proses 2-3 sesuai dengan jumlah iterasi yang diinginkan.
5. Jika pada proses 2 terdapat 2 atau lebih *cluster* yang meiliki nilai tertinggi, lakukan penunjukan secara acak untuk menentukan *cluster* mana yang akan dituju dari *node* yang ditunjuk tersebut.

D. Cosine Similarity

Cosine similarity adalah metode pengukuran kemiripan yang sering digunakan dalam *text mining* khususnya dalam *information retrieval* dan juga *clustering* [11]. *Cosine similarity* merupakan ukuran sudut antara vektor dokumen (titik (*ax*,*bx*)) dan (titik (*ay*,*by*)). Tiap vektor tersebut merepresentasikan setiap kata dalam setiap dokumen yang dibandingkan dan membentuk sebuah segitiga. Ketika dua dokumen identik, maka akan menbentuk sudut nol derajat (0°) dan bobot kesamaannya adalah satu (1); dan ketika dua dokumen tidak identik sama sekali, maka sudut yang akan dibentuk adalah 90 derajat (90°) dan bobot kesamaannya adalah nol (0) [12]. Berikut adalah rumus untuk mencari Cosine Similarity [13] [14]:

$$\textit{Cosine Similarity}\ (d1,d2) = \frac{}{\| \ \| \ \| \ \|}$$

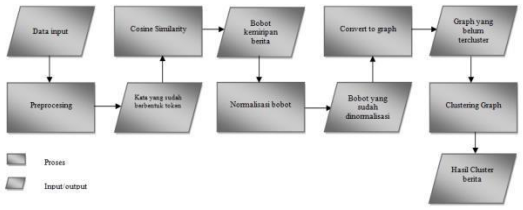
Sebelum melakukan penghitungan terhadap nilai Cosine Similarity, terlebih dahulu harus menghitung nilai TF, IDF, dan TFxIDF. Gambar 2.4 adalah proses yang harus dilewati terlebih dahulu sebelum menghitung nilai *Cosine Similarity*.

Gambar 2.3 Proses sebelum penghitungan *Cosine Similarity*

3. Gambaran Umum Sitem dan Skenario Pengujian

A. Gambaran Umum Sistem

Pada tugas akhir ini akan dibangun sistem *clustering graph* dengan memanfaatkan algoritma *Chinese Wishpers*, dimana setiap *node* dalam *graph* tersebut merupakan permodelan dari sebuah berita *online* berbahasa Indonesia. Gambar 3.1 adalah gambaran umum dari sistem yang dibentuk :



Gambar 3.1 Gambaran umum sistem

a. Pada tahap *preprocessing* datainput yang ada akan di rubah menjadi kata-kata berupa *token*. Pada tahap ini akan dilakukan *tokenizing* + *case folding*, *stopword removal*, serta *stemming*.
b. Pada tahap penghitungan *Cosine Similarity*, *token* yang didapatkan dari tahap *preprocessing* akan dihitung menggunakan rumus *cosine similarity*, namun sebelumnya dilakukan penghitungan terhadap TF, IDF, dan juga TFxIDF pada setiap dokumen.
c. Hasil penghitungan *Cosine Similarity* berkisar antara 0 hingga 1, untuk itu perlu dilakukan normalisasi agar memudahkan proses pengamatan penelitihan. Normalisasi dilakukan dengan melakukan perkalian dengan angka 10 terhadap semua bobot yang ada.
d. Setelah dilakukan normalisasi, selanjutnya adalah memodelkan berita kebentuk *graph*. Dimana setiap *node* dalam *graph* tersebut berikan informasi mengenai suatu berita, seperti nama pengarang, judul, tanggal terbit, isi berita, dan lainnya. Sedangkan untuk relasi antar *node* diperoleh dari bobot kedua dokumen yang dihasilkan dari proses normalisasi. Suatu dokumen dikatakan saling terkait jika mereka mempunya bobot antara 1-9.
e. Berita-berita yang sudah berbentuk *graph* tadi selanjutnya akan di *clustering*, sehingga berita-berita tersebut akan terkelompok berdasarkan keterkaitan isi dari berita tersebut.

B. Skenario Pengujian

a. Akan dilakukan pengamatan hasil akhir *cluster* dengan menggunakan iterasi 1-6 dengan masing-masing iterasi akan diambil 5 hasil *cluster* dengan menggunakan data uji yang merupakan data sampel. Selanjutanya setiap *node* dari hasil akhir

*cluste*r akan dilihat nilai *inter-cluster* dan nilai *intra-cluster*-nya, nantinya akan dihitung presentase sebuah *node* sudah berada pada *cluster* yang memiliki nilai *intra-cluster* yang lebih tinggi dari *inter-cluster*-nya.

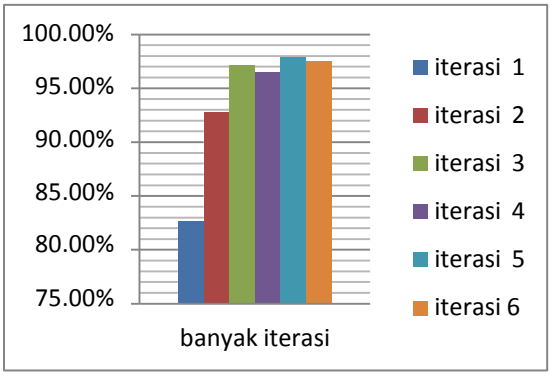
b. Akan dibandingkan hasil *clustering* dari sistem dengan hasil *clustering* yang dilakukan oleh *expert*. *Expert* yang dipilih dalam skenario uji ini adalah guru Bahasa Indonesia yang juga lulusan program studi pendidikan bahasa dan sastra Indonesia. Hasil *cluster* yang digunakan adalah hasil *clustering* dengan 6 kali iterasi,

4. Analisis Pengujian

A. Skenario 1

Pada gambar 4.1 terlihat bahwa pada hasil *cluster* dengan 1 kali iterasi memiliki rata-rata presentase hasil *cluster* memiliki *nilai intra-cluster* yang lebih besar daripada nilai *inter-cluster* sebesar 82.69% hal ini disebabkan karena pada hasil *clustering*-nya masih belum stabil, sehingga masih banyak node yang belum memiliki nilai *intra-cluster* yang lebih tinggi dari pada nilai *inter-cluster*-nya. Untuk hasil *clustering* dengan menggunakan 2 kali iterasi presentasenya meningkat menjadi 92.79%, hal ini berbanding lurus dengan semakin stabilnya *cluster* yang dihasilkan.

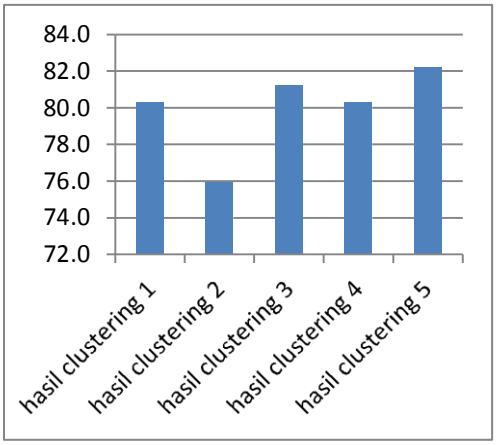
Untuk hasil akhir dengan menggunakan 3 dan 4 kali iterasi presentase yang dihasilkan sudah lebih dari 95%, namun masih didapatkan ada beberapa *node* yang masih belum memiliki nilai *intra-cluster* yang lebih tinggi dari nilai *inter-cluster*-nya yang disebabkan *cluster* tersebut masih belum cukup stabilnya *cluster* yang dihasilkan. Sedangkan untuk hasil akhir dengan menggunakan 5 dan 6 iterasi presentase yang dihasilkan juga sudah lebih dari 95%, namun *cluster* yang dihasilkan sudah stabil karena semua *node* tidak ada yang memiliki nilai *inter-cluster* yang lebih tinggi dari pada nilai *intra-cluster*-nya. Penyebab nilai presentase pada iterasi 5 dan 6 tidak mencapai 100% adalah beberapa *node* memiliki nilai *intra-cluster* dan *inter-clsuter* yang sama, hal ini disebabkan oleh proses penunjukan *cluster* secara *random* jika terdapat *cluster* yang memiliki bobot yang sama



Gambar 4.1 Hasil pengujian skenario 1

B. Skenario 2

Berdasarkan skenario 2 pengujian bertujuan untuk mengetahui nilai akurasi hasil clustering, pada gambar 4.2 adalah hasil penghitungan nilai akurasi dari hasil *clustering*.



Gambar 4.2 Hasil pengujian skenario 2

Tidak konsistennya nilai akurasi yang dihasilkan disebabkan karena di setiap hasil *clustering* memiliki hasil *cluster* yang berbeda walaupun jumlah *cluster* yang dihasilkan sama. Faktor utama penyebab hasil *cluster* menjadi berbeda-beda terletak pada proses pemilihan *node* secara *random* dan proses pemilihan *cluster* secara *random* jika terdapat bobot *cluster* yang sama pada saat *clustering* berlangsung, penunjukan secara *random* pada awal iterasi sangat berpenagruh terhadap hasil akhir *clustering*. Contohnya adalah antara hasil *clustering* 1 dengan hasil *clustering* 3 meskipun keduanya mempunyai jumlah *cluster* yang sama 19 akan tetapi isi dari tiap *cluster* berbeda, itulah mengapa nilai akurasi yang dihasilkan juga berbeda.

Faktor lainnya yang mempengaruhi nilai akurasi terletak pada proses *text mining* yang tidak sempurna. Seperti yang telah di jelaskan pada bab 4 bahwa pada proses *preprocessing text mining* menggunakan *library* dan luarannya tidak akurat 100%. Akibatnya ada beberapa dokumen yang harusnya memiliki bobot yang cukup besar menjadi lebih kecil dari seharusnya dan dampaknya juga berpengaruh terhadap hasil *clustering*.

5. Kesimpulan

1. Dokumen berita *online* dapat dimodelkan kedalam bentuk *graph*, dimana setiap dokumen berita akan dirubah menjadi sebuah *node*. Penghubung antar tiap *node* dapat dicari dengan memanfaatkan keterkaitan isi berita yang dihitung menggunakan *cosine similarity*.
2. Algoritma *Chinese Whispers* mampu digunakan untuk mengelompokkan berita *online* sesuai dengan keterkaitan isi beritanya.
3. Hasil *clustering* yang dihasilkan oleh algoritma *Chinese Whispers* sangat bergantung pada banyaknya iterasi yang digunakan. Iterasi bertujuan untuk membuat hasil *clustering* semakin stabil. Jika *clustering* yang dihasilkan telah berada pada kondisi yang stabil maka nilai presentase hasil *clustering* tersebut memiliki nilai *intra-cluster* yang lebih besar dari pada nilai *inter-cluster*-nya mencapai lebih dari 95%.
4. Nilai akurasi hasil *clustering* terhadap berita *online* sangat tergantung dengan luaran proses *clustering* itu sendiri. Hal ini disebabkan oleh *clustering* menggunakan algoritma *Chinese Whispers* sangat bergantung terhadap penunjukkan secara *random* pada awal iterasi. Adapun rata-rata nilai akurasi hasil *clustering* terhadap berita *online* adalah 80.0%.

6. Daftar Pusaka

[1] Husnul Khatimah, *ANALISIS FAKTOR-FAKTOR WEBSITE QUALITY DALAM MENGGUNAKAN SITUS BERITA ONLINE KOMPAS.COM 2013*. Bandung: FEB Universitas Telkom, 2013.

[2] isjhar kautsar, *Analisis Performansi Metode Graph Decomposition Index pada Graph Database*. Bandung, Indonesia: Fakultas Informatika Telkom University, 2014.

[3] Fahri Firdausillah, Erwin Yudi Hidayat, and Ika Novita Dewi, *NoSQL: Latar Belakang, Konsep, dan Kritik*. Semarang, Indonesia: Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang, 2012.

[4] Yang Zhou, Hong Cheng, and Jeffery Xu Yu, *Graph Clustering Based on Structural/Attribute Similarities*. Hong Kong: Engineering Management University of Hong Kong.

[5] Satu Elisa Schaeffer, *Graph CLustering*. Helsinki, Finland: Laboratory for Theoretical Computer Science, Helsinki University of Technology, 2007.

[6] Reena Mishra, Shashwat Shukla, Deepak Arora, and Mohit Kumar, *An Effective Comparison of Graph Clustering Algorithms via Random Graphs*. India: Department of Computer Science and Engineering Amity University, 2011.

[7] Derry Tanti Wijaya, *Graph, Clustering and Applications*. Singapore: Depatment of Computer Science National University of Singapore, 2008.

[8] Chris Biemann, *Chinese Whispers - an Efficient Graph Clustering Algorithmand its Application to Natural Language Processing Problems*. Leipzig, Germany: University of Leipzig, NLP Department.

[9] Kenneth H. Rosen, *Discrete Mathematics and Its Applications*, Seventh Edition ed. New York: McGraw-Hill, 2007.

[10] Mousumi Dhara and K. K. Shukla, *Characteristic of Restricted Neighbourhood Search Algorithm and Markov Clustering on Modified Power-Law Distributor*. India: Departement of Computer Engineering, 2012.

[11] Bhanu Prasad A. and Venkata Gopala Rao S., *Space and Cosine Similarity measures for Text Document Clustering*. Hyderabad, India: Vardhaman College of Engineering, 2013.

[12] Radiant Victor Imbar, Adelia , Mewati Ayub, and Alexander Rehatta, *Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks*. Bandung, Indonesia: Fakultas Teknologi Informasi Universitas Kristen Maranatha, 2014.

[13] Jana Vembunarayanan. (2013, Oktober) Seeking Wisdom. [Online]. https://janav.wordpress.com/2013/10/27/tf-idf-

and-cosine-similarity/

[14] Jon Borglund, *Event-Centric Clustering of News Articles*. Uppsala, Sweden: Department of Information Technology, Uppsala University, 2013.