IMPLEMENTASI DETEKSI KOMUNITAS UNTUK EKSTRAKSI INFORMASI DARI DATA TEKS

Doni Pradana[1], Anggi Amilia Pratiwi[2], Saepul Lukman[3]
[1]Sistem Komputer, STMIK Jakarta STI&K
[2,3]Manajemen Informatika,, STMIK Jakarta STI&K
Doni.pradana@staff.jak-stik.ac.id, amiliaanggi11@gmail.com, fulman2012@gmail.com

Abstrak

Perkembangan teknologi dan informasi yang pesat membuat data teks dalam bentuk informasi menjadi lebih cepat dan mudah diakses. Akibatnya, diperlukan sumber daya yang besar dan pemrosesan data yang cepat. Data teks yang diterima sebagai informasi bentuknya tidak terstruktur. Ekstraksi informasi (IE) adalah tugas mengekstraksi informasi secara otomatis dari pembelajaran mesin yang tidak terstruktur. Sebagai jawaban atas tantangan ekstraksi informasi ini, maka digunakan deteksi komunitas pada grafik pengetahuan. Hubungan antar data teks disebut triplet atau berisi (subjek, predikat, objek). Evaluasi pembentukan dan pengelompokan grafik dilihat dari modularitasnya, nilainya berkisar antara -1 hingga 1. Hasil mudularitas grafik dengan data teks yang diusulkan sangat baik, yang menghasilkan 0,91. Selain itu, pengelompokan juga dilakukan dengan mengambil 5 teratas sebagian besar hubungan (edge)**.**

Kata Kunci: Grafik Pengetahuan, Deteksi Komunitas, Metode Louvain, Modularitas, Pengelompokan, Ekstraksi Informasi

Pendahuluan

Sering dengan pesatnya perkembangan informasi dan teknologi, informasi data text mengalami pertumbuhan secara pesat dan mudah diakses. Hal ini berdampak pada meningkatnya kapasitas penyimpanan menjadi besar serta pemrosesan data yang cepat dan akurat. Informasi pada data teks bersifat tidak terstruktur, sehingga data tidak dapat diproses atau dianalisis secara langsung dengan alat dan metode konvensional. Salah satu tantangan dalam pemrosesan data teks adalah kompleksitas dekonstruksi karena tidak memiliki model spesifik yang telah ditentukan.

Penelitian ini bertujuan untuk melakukan penambangan teks dan ekstraksi informasi teks [1]. Pembelajaran yang digunakan untuk melakukan supervisi, semi-supervised dan unsupervised text mining algoritma penambangan [2] (seperti penambangan hubungan semantik, klasifikasi dan pengelompokan teks, analisis orientasi teks, dan penemuan dan pelacakan topik), dan alat untuk penambangan teks [3].

Ekstraksi informasi (IE) adalah tugas mengekstraksi informasi secara otomatis dari pembelajaran mesin yang tidak terstruktur [4]. Dalam sebagian besar kasus, aktivitas ini berkaitan dengan pemrosesan bahasa atau NLP (Pemrosesan Bahasa Alami). Ekstraksi informasi tanpa pengawasan yang dapat merujuk pada IE apa pun yang mencoba memulihkan beberapa jenis informasi seperti ekstraksi informasi tabel secara terstruktur dari tabel.

Sebagai jawaban atas tantangan ekstraksi informasi ini, maka digunakanlah metode deteksi komunitas yang merupakan bagian dari pembelajaran mesin tanpa pengawasan. Definisi deteksi komunitas dan pengelompokan seringkali membingungkan dalam setiap literatur. Teknik pengelompokan lebih berfokus pada atribut node, sementara deteksi komunitas berfokus pada struktur jaringan [5]. Dengan demikian, akan tercipta cara baru untuk mendapatkan informasi dengan kemampuan yang lebih baik.

Pekerjaan terkait ekstraksi informasi, membangun grafik pengetahuan untuk pemeriksaan literatur biomedis yang

mengekstraksi data numerik (jumlah pasien, usia, dan distribusi jenis kelamin) dan data teks [6]. Selain itu, mengusulkan model jaringan saraf untuk ekstraksi bersama entitas nama dan hubungan di antara mereka, memperluas model pengenalan entitas berbasis BiLSTM - CRF, dan membuat klasifikasi menggunakan data relasional [7]. memperkenalkan grafik pengetahuan ke dalam RS dan mengusulkan model, yang secara khusus mempertimbangkan berbagai preferensi bagi pengguna untuk meningkatkan kinerja rekomendasi dan pada saat yang sama menyelesaikan fakta yang hilang dalam grafik pengetahuan [8].

Dari penelitian terkait, kita mengetahui bahwa IE penting untuk kebutuhan informasi di semua domain. Dan peneliti lain bersaing untuk mendapatkan dan mengekstrak informasi dari data mentah (tidak terstruktur), terutama pada arsitektur big data.

Secara umum tujuan penulisan makalah ini adalah untuk mengimplementasikan Community Detection pada Knowledge Graph sebagai ekstraksi informasi khususnya data teks Artikel tentang film. Kemudian, yang lain mengetahui proses dan tahapan pembentukan deteksi komunitas secara khusus dan Grafik Pengetahuan secara umum.

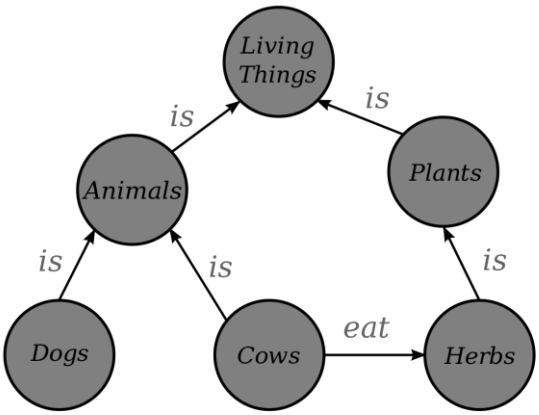
II. Grafik Pengetahuan dan Deteksi Komunitas

Grafik pengetahuan merupakan representasi dari masalah kritis dan mendasar dalam rekayasa pengetahuan dan

Kecerdasan Buatan [9] dan pada dasarnya merupakan

jaringan leksem yang terdiri dari banyak tiga (subjek, predikat, objek), yang dapat memberikan jangkauan asosiasi yang lebih dalam dan lebih luas antara pengguna dan berbagai item untuk meningkatkan properti rekomendasi [10].

Pada dasarnya, Knowledge Graph termasuk dalam kategori jaringan semantik sebagai sebuah metode. Struktur Knowledge Graph berisi konsep dan predikat. Konsep (entitas atau topik) direpresentasikan sebagai simpul dalam Knowledge Graph yang secara konseptual mengekspresikan sesuatu yang nyata di dunia. Sementara itu, predikat (hubungan) merepresentasikan hubungan antara dua konsep yang dinyatakan oleh label pada Knowledge Graph. Ilustrasi Knowledge Graph menurut Gambar 1.



Gambar 1. Ilustrasi Grafik Pengetahuan

Dalam kasus Grafik Pengetahuan, ini mengacu pada Skalabilitas Ruang-Waktu, yang berarti metode ini harus terus

berjalan dengan baik bahkan dengan data berskala besar dan harus segera menyelesaikan masalah.

Deteksi komunitas berguna untuk komputasi graf karena membantu mengungkap relasi tersembunyi. Karena terlalu banyak data yang akan diproses, terdapat klaster yang masih tersembunyi. Ada beberapa definisi untuk komunitas yang sebagian besar membutuhkan komputasi yang sangat mahal. Berikut ini adalah beberapa kebutuhan dan pentingnya mengidentifikasi komunitas:

Perhitungan deteksi komunitas diusulkan menggunakan metode Louvain, dimana metode ini berbasis pada Metode

Grafik Pengetahuan dan juga mampu Memberikan skalabilitas tinggi terlepas dari kompleksitas data yang digunakan. Secara umum, algoritma ini termasuk dalam kategori Pengelompokan Berbasis Optimasi yang mengikuti strategi untuk menemukan klaster dengan modularitas maksimum. Persamaan 1adalah metode Louvain untuk deteksi komunitas.

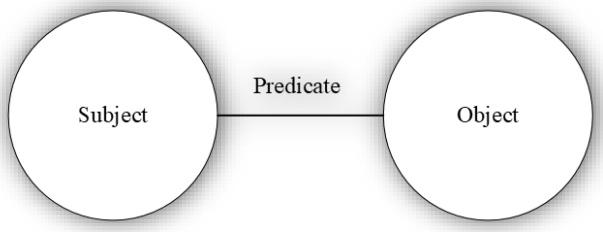
$$\Delta Q = \left[\frac{\Sigma_{in}+k_{i,in}}{2m} - \left(\frac{\Sigma_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \quad (1)$$

Σ_{in} adalah jumlah bobot pada komunitas atau kelompok di dalamnya, adalah jumlah bobot total terhubung dari

semua node di C, $k_i$adalah jumlah bobot terhubung dari hasil kejadian di node *I*, adalah jumlah bobot hubungan i ke semua node di c, dan m adalah jumlah bobot semua hubungan dalam jaringan.

Metode Penelitian

Metodologi dalam makalah ini adalah menggunakan data teks sebagai dasar untuk membangun grafik pengetahuan. Setelah membentuk grafik pengetahuan dengan semua entitas yang ada, pencarian informasi dilakukan berdasarkan relasi atau predikat. Dasar pembentukan grafik pengetahuan membutuhkan triplet (subjek-predikat-objek). Gambar 2 mengilustrasikan bentuk dasar pembentukan pengetahuan.



Gambar 2. Bentuk Dasar Informasi Pengetahuan

Data yang digunakan adalah data artikel yang berisi informasi dan artikel tentang film, bersumber dari Wikipedia dalam format CSV, dengan jumlah data sebanyak 4318 kalimat. Pengolahan data dilakukan di Google Collaboratory. Alur kerja metodologi sesuai dengan Gambar 3.

Sebagai tambahan, evaluasi grafik menggunakan perhitungan sesuai rumus 2. Rumus ini dapat diterapkan pada semua grafik pengetahuan, seperti grafik utama dan grafik relasi (tepi) 5 teratas.

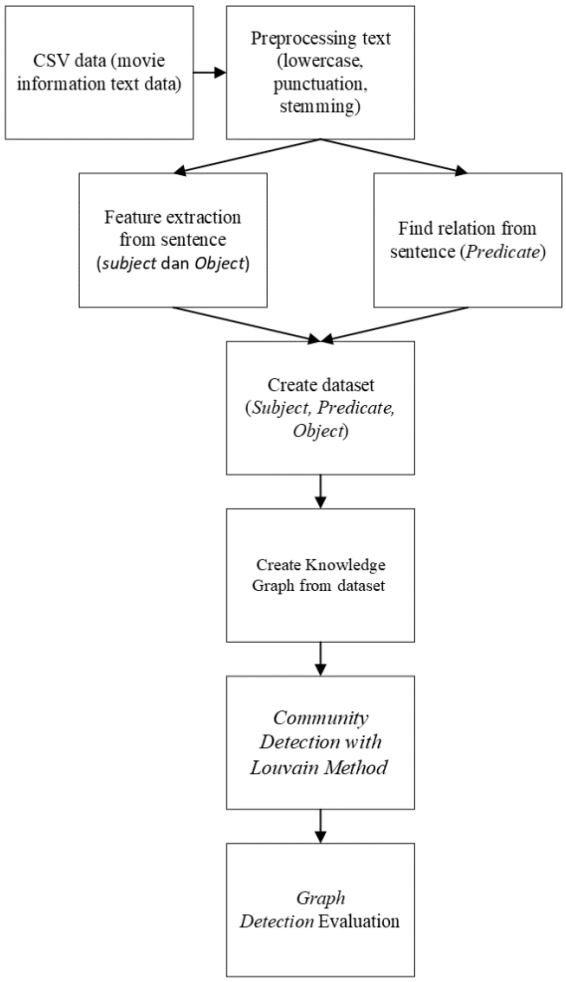
$$Q = \frac{1}{2m}\sum_{i,j}\left[A_i - \frac{k_i k_j}{2m}\right]\delta(C_i, C_j) \quad (2)$$

A_i = represent weight in *edge* between *I* and *j*

k_i, k_j = sum of weight is attached in *vertex i*

m = half of sum weight in network

$\delta(C_i, C_j)$ = 1 if i and j in a same community, otherwise 0



Gambar 3. Alur kerja metodologi

Implementasi

Implementasi penggunaan bahasa pemrograman Python dilakukan pada Google Collaboratory Notebook. Pustaka yang digunakan adalah NetworkX, pustaka khusus untuk membuat dan menganalisis jaringan, serta Pustaka Spacy untuk memproses dan mengekstrak kalimat. Sesuai dengan metodologi alur kerja program, Langkah prapemrosesan wajib dilakukan karena kalimat tersebut mengandung tanda baca. Kecuali kata henti karena mengandung relasi dalam kalimat. Pemrosesan data dilakukan denganluar angkasaPustaka. Kerangka kerja ini mampu menyediakan pemrosesan dan model bahasa.

Kemudian, proses untuk mendapatkan subjek sekaligus objek. Selain itu, ekstraksi relasi merupakan proses penting dalam alur kerja struktur. Dasar dari ekstraksi fitur adalah proses parsing, yaitu analisis sintaksis atau proses menganalisis sebuah string agar sesuai dengan aturan tata bahasa formal. Proses parsing juga memperoleh kelas kata (part of speech). Parsing kalimat secara tradisional sering dilakukan sebagai metode untuk memahami makna kata dalam suatu kalimat secara tepat. Untungnya, kerangka kerja SpaCy menyediakan alat dan model untuk melakukan parsing kalimat dengan pipeline bahasa Inggris yang telah dioptimalkan. Contoh parsing kalimat ditunjukkan pada Tabel 1 dengan kalimat *"Confused and frustrated, Connie decides to leave on her own."*

Tabel 1. Hasil Parsing

String	**Class word**
confused	advc1
and	cc
frustrated	conj
,	punct
connie	nsubj
decides	root
to	aux
leave	xcomp
on	prep
her	poss
own	pobj
.	punct

Proses ekstraksi entitas dilakukan dengan menempatkan kata berkategori subj class sebagai entitas subjek dan kata berkategori pobj class sebagai entitas objek. Selain itu, untuk ekstraksi relasi atau bagian edge, dilakukan dengan menempatkan kata berkategori root class. Setelah semua elemen berhasil diekstraksi, entitas (subjek dan objek) akan dibuat dalam bentuk dataframe bersama bagian relasi (edge), sesuai dengan Tabel 2.

Untuk membangun knowledge graph, dataframe relasi dengan data relasi (edge) yang telah dibuat sebelumnya sangat berguna. Selanjutnya, data tersebut dapat merepresentasikan proses deteksi komunitas pada tahap awal. Langkah terakhir adalah membangun komunitas dan menghitung nilai modularitas menggunakan metode Louvain serta rumus modularitas yang telah dijelaskan sebelumnya. Pada skrip Python, seluruh perhitungan dilakukan menggunakan fungsi dari pustaka NetworkX.

Hasil

Hasil yang diperoleh sesuai dengan metode dan alur kerja implementasi yang digunakan. Tabel 2 menunjukkan tabel hasil contoh ekstraksi fitur

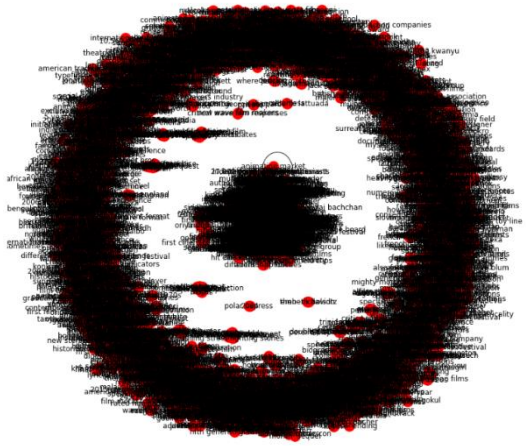
dari data teks. Kolom-kolom pada tabel tersebut merepresentasikan subjek, objek, dan predikat.

Tabel 2. Contoh Hasil Ekstraksi Fitur

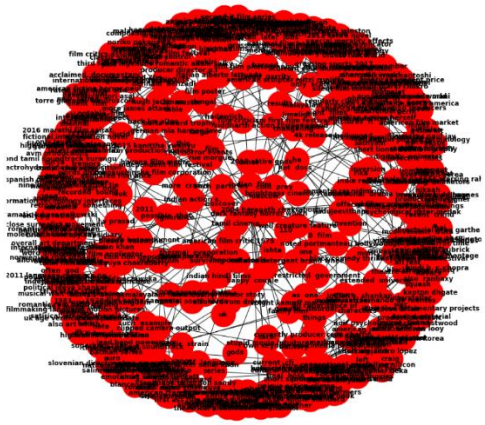
No.	Source	Target	Edge
1	connie	own	decides
2	later woman	distance	heard in
3	temple	fire	set on
4	confidencial	negatively film	responded
5	le parisien	five star rating	gave

Gambar 4 merupakan hasil knowledge graph dari seluruh dataset dalam format CSV. Dalam proses pembuatannya, digunakan tipe relasi yang bersifat undirected (tidak berarah).

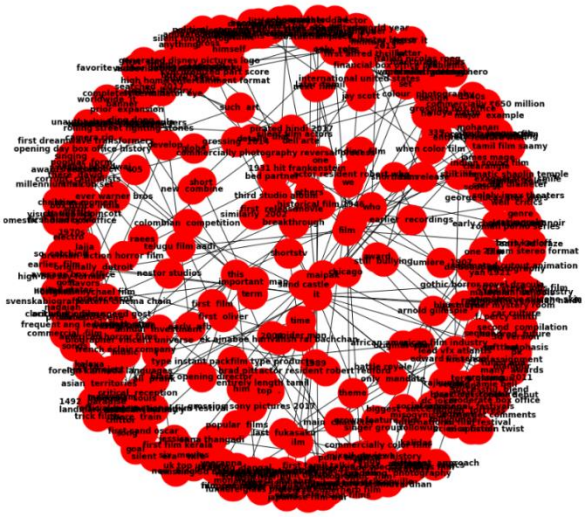
Dari gambar tersebut tampak cukup kompleks karena seluruh data saling terhubung satu sama lain. Oleh karena itu, perlu diambil salah satu sampel relasi yang ada, dengan memilih 5 jenis relasi (edge) terbanyak yang telah terbentuk. Hal ini ditunjukkan secara berurutan pada Gambar 5, Gambar 6, Gambar 7, Gambar 8, dan Gambar 9.



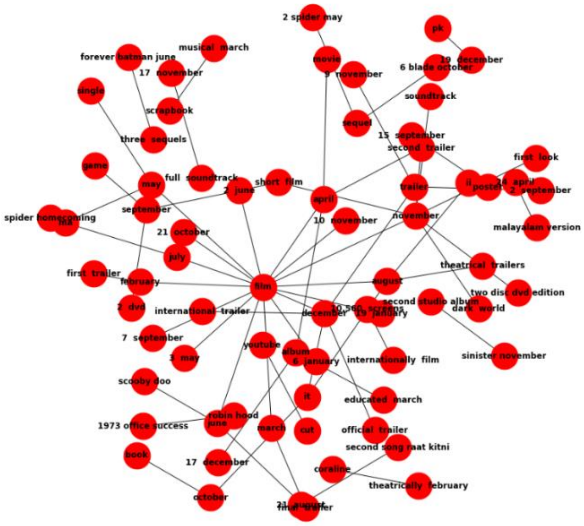
Gambar 4. Hasil dengan Semua Data di Knowledge Graph



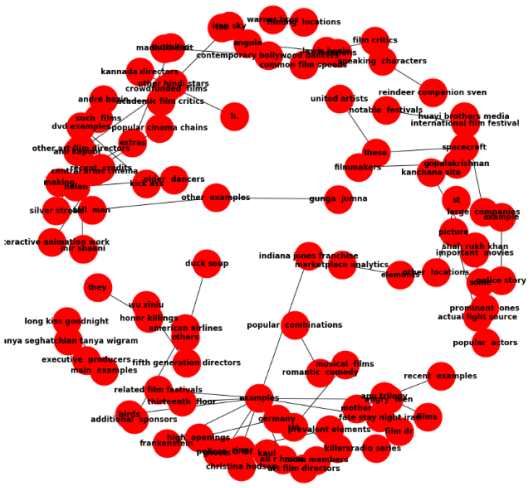
Gambar 5. Grafik dengan Relasi “adalah”



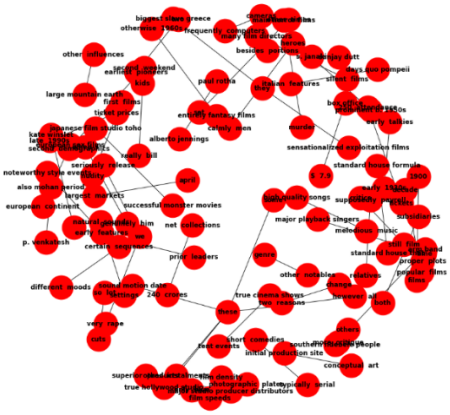
Gambar 6. Grafik dengan Relasi “was”



Gambar 7. Grafik dengan Relasi “dirilis pada”



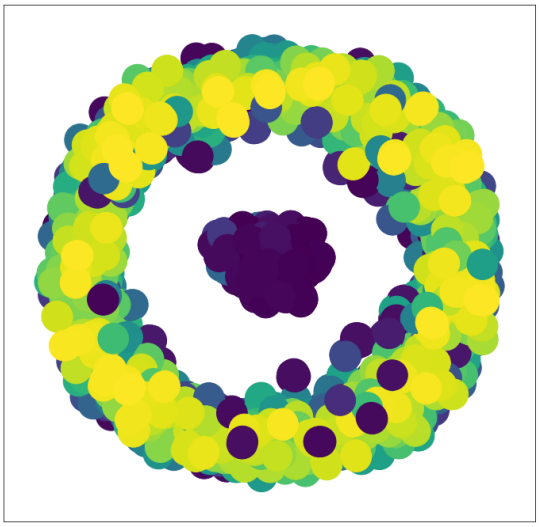
Gambar 8. Grafik dengan Relasi “include”



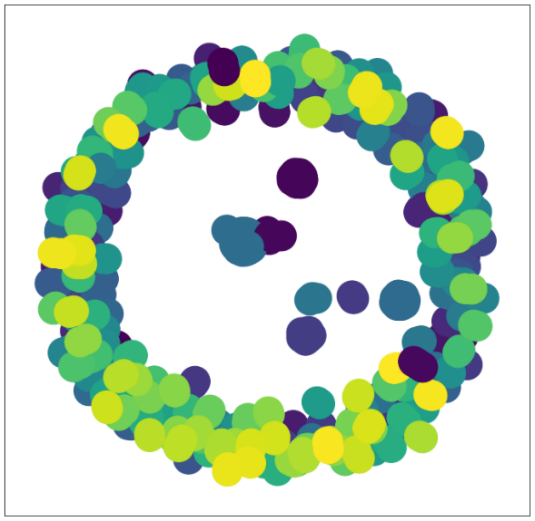
Gambar 9. Grafik dengan Relasi “were”

Dari relasi antar entitas dalam *graph*, kita dapat mengetahui bahwa melalui relasi tersebut, informasi dapat diperoleh. Sebagai contoh, jika kita ingin mengetahui suatu kasus terkait penayangan sebuah film, *knowledge graph* mampu menampilkan berbagai jadwal yang berkaitan dengan film dan tanggal rilis dari database. Dengan demikian, informasi mengenai jadwal penayangan film dapat dengan mudah dilihat.

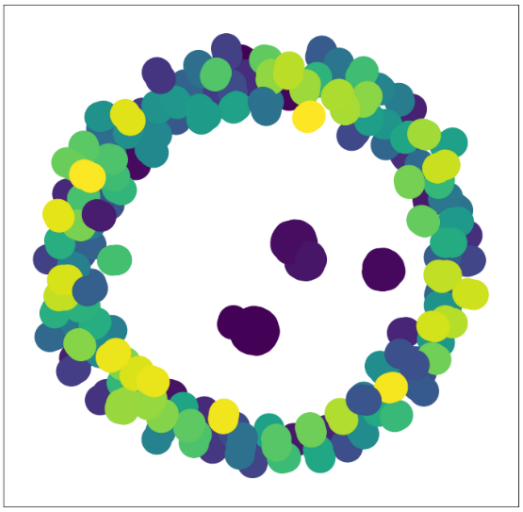
Langkah selanjutnya adalah melakukan deteksi komunitas menggunakan algoritma Louvain. Dengan algoritma Louvain, klaster dapat dibentuk. Visualisasi hasil klastering sangat penting untuk membedakan satu komunitas dengan komunitas lainnya. Dalam hal ini, penggunaan warna yang berbeda untuk setiap klaster pada *relation graph* merupakan pilihan terbaik. Gambar 10, Gambar 11, Gambar 12, Gambar 13, Gambar 14, dan Gambar 15 menunjukkan hasil deteksi komunitas menggunakan metode Louvain, di mana setiap warna merepresentasikan komunitas yang berbeda. Visualisasi ini memiliki bentuk yang hampir serupa dengan *knowledge graph* sebelumnya.



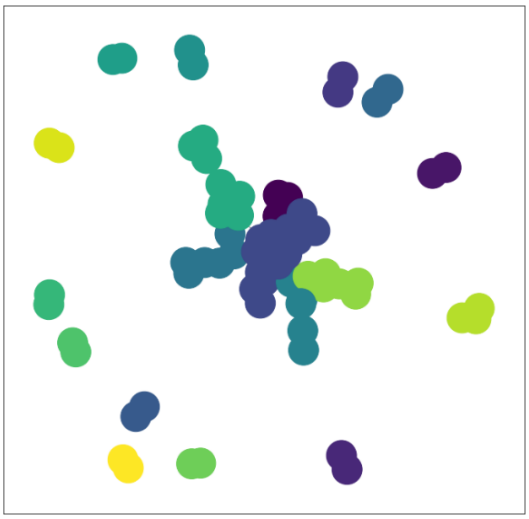
Gambar 10. Hasil Metode Lovain Semua Relasi



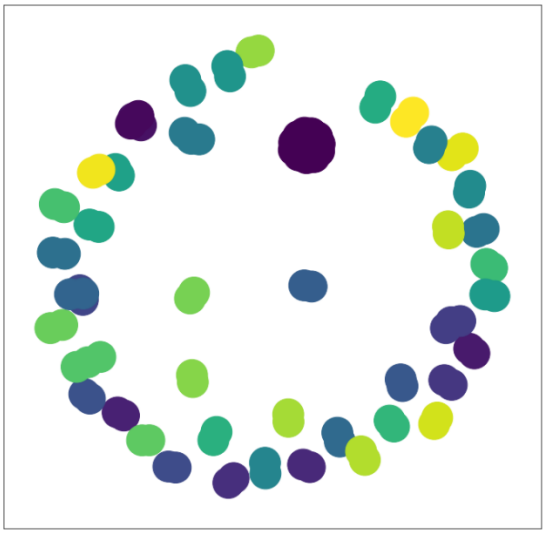
Gambar 11. Hasil Metode Lovain “adalah” Relasi



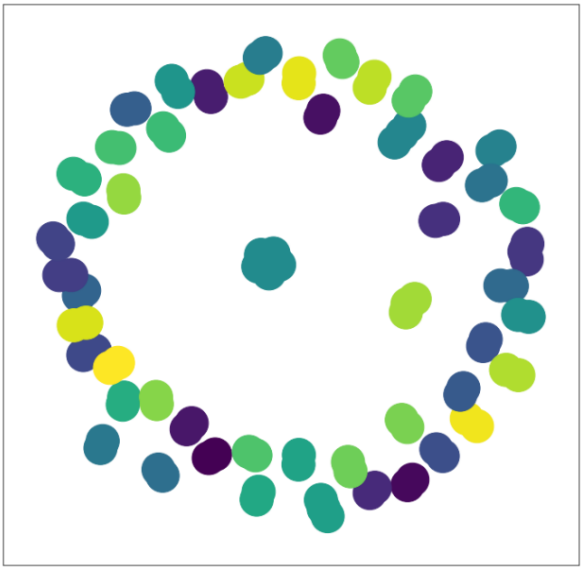
Gambar 12. Hasil Metode Lovain “was” Relasi



Gambar 13. Hasil Metode Lovain “realesd on” Relasi



Gambar 14. Hasil Relasi “include” Metode Lovain



Gambar 15.Hasil Metode Lovain “were” Relasi

Setelah proses deteksi komunitas dilakukan, sebagai bentuk evaluasi, Tabel 3 menunjukkan hasil dari 5 jenis relasi (edge) terbanyak.

Tabel 3. Relasi Top 5 (tepi)

No	Relation (edge)	Modularity value
1	Is	0.972
2	Was	0.957
3	Released on	0.681
4	Include	0.942
5	were	0.975

Dari Tabel 3 yang menunjukkan nilai modularitas, dapat diketahui bahwa pola pada knowledge graph memiliki keterkaitan dengan bentuk visual dari graph-nya. Semakin mendekati bentuk lingkaran dan terlihat solid atau rapat, maka nilai modularitasnya cenderung semakin tinggi. Selain itu, nilai modularitas tidak berbanding lurus dengan jumlah kemunculan relasi dalam tabel relasi.

Penutup

Kesimpulan

Perhitungan Knowledge Graph dan Community Detection menggunakan Metode Louvain telah dilakukan dengan memanfaatkan dataset informasi tentang film yang bersumber dari Wikipedia. Dari eksperimen ini, kita dapat memahami langkah demi langkah dalam membangun Knowledge Graph dan melakukan

Community Detection. Melalui hubungan antara subjek dan objek, informasi dapat diamati dan dianalisis lebih lanjut.

Ke depannya, Knowledge Graph dan Community Detection dapat diintegrasikan dan dikombinasikan dalam model machine learning, khususnya dalam pengembangan deep learning, mengingat besarnya volume data yang perlu diproses dan pentingnya pemahaman terhadap hubungan antar data. Selain itu, penerapan Knowledge Graph dan Community Detection juga dapat digunakan sebagai metode dalam sentiment analysis, dengan cara mengamati relasi (predicate) antara subjek dan objek yang mengarah pada sentimen positif maupun negatif.

Daftar Pustaka

[1] Jing-Tao Sun, et al. Modeling of unsupervised knowledge graph of events based on mutual information among neighbor domains and sparse representation, Defence Technology, 2021.

[2] Malik R, Franke L, Siebes A. Combination of text-mining algorithms increases the performance. Bioinformatics 2018;22(17):2151e7.

[3] Sharmin S, Zaman Z. Spam detection in social media employing machine learning tool for text mining. In: 2017 13th international conference on signal-image technology & internet-based systems (SITIS). IEEE Computer Society; 2017. p. 137e42. 1.

[4] Adnan, K., Akbar, R. An analytical study of information extraction from unstructured and multidimensional big data. J Big Data 6, 91 (2019).

[5] Isa Inuwa-Dutse, et al. A multilevel clustering technique for community detection, Neurocomputing, Volume 441, 2021, Pages 64-78.

[6] Milosevic N, Gregson C, Hernandez R, Nenadic G (February 2019). "A framework for information extraction from tables in biomedical literature". International Journal on Document Analysis and Recognition (IJDAR). 22 (1): 55–78.

[7] Dat Quoc Nguyen and Karin Verspoor (2019). "End-to-end neural relation extraction using deep biaffine attention". Proceedings of the 41st European Conference on Information Retrieval (ECIR).

[8] Cao, Y., Xiang, W., He, X., Hu, Z., & Chua, T. S. (2019). Unifying Knowledge Graph Learning and Recommendation: Towards a Better Understanding of User Preferences. In Proceedings of The World Wide Web Conference (pp. 151-161). San Francisco, CA, USA.

[9] H. Xiao, Y. Chen and X. Shi, "Knowledge Graph Embedding Based on Multi-View Clustering Framework," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 2, pp. 585-596, 1 Feb. 2021, doi: 10.1109/TKDE.2019.2931548.

[10] Xingjuan Cai, Lijie Xie, Rui Tian, Zhihua Cui, Explicable recommendation based on knowledge graph, Expert Systems with Applications, Volume 200, 2022, 117035, ISSN 0957-4174