E-ISSN 2460-9552
P-ISSN 2620-3227

Article history
Received 26 May 2020
Accepted 29 Dec 2020

ANALISIS SUKU DI INDONESIA MENGGUNAKAN ALGORITMA CLOSENESS CENTRALITY

Puspa Nur Indrianingtyas [1)], Fanny Azhary Formen [2)], Nur Aini Rakhmawati [3)]
[1,2,3] Departemen Sistem Informasi, Fakultas Teknik Elektro Dan Informatika Cerdas, \
Institut Teknologi Sepuluh Nopember
Email: indrityas7@gmail.com, Fannyazhari355@gmail.com, nur.aini@is.its.ac.id

Abstract

Indonesia is a country that has a lot of diversity. One of the cultural contributions made by the Indonesian state is the different languages between tribes and locations. Then with the contribution given by the Indonesian state, it is necessary to represent the mapping of Indonesian ethnic groups by using a graphical algorithm and to analyze the closeness between nodes, the proximity algorithm and node2vec are used to find the relationship of each tribe. In searching this data paper using data provided by dbpedia obtained using sparql. To display data that has dimensions that are used by t-Distributed Stochastic Neighbor Embedding (t-SNE) to change data points.

Keywords : *sparql, Graph algoritma, Node2vec, closeness algoritma, t-SNE*

Abstrak

Indonesia merupakan negara yang memiliki banyak keanekaragaman kebudayaan. Salah satu keanekaragaman budaya yang dimiliki oleh negara indonesia ini adalah bahasa yang berbeda-beda antar suku serta lokasi. Maka dengan keanekaragaman yang dimiliki negara indonesia ini perlu adanya pemetaan mengenai suku bangsa indonesia dengan menggunakan algoritma graph serta untuk menganalisis kedekatan antar node digunakan *closeness algoritma* dan *node2vec* untuk mencari hubungan dari setiap suku. Dalam pencarian data paper ini menggunakan data yang disediakan oleh dbpedia yang didapatkan menggunakan *sparql*. Untuk menampilkan data yang memiliki dimensi yang tinggi maka digunakan *t-Distributed Stochastic Neighbor Embedding* (t-SNE) untuk mengkonversi titik data.

Kata Kunci : *sparql, Graph algoritma, Node2vec, closeness algoritma, t-SNE*

1. PENDAHULUAN

Sebagai negara kepulauan negara Indonesia memiliki berbagai macam suku maupun kebudayaan. setiap suku memiliki bahasa masing-masing serta daerah sebagai tempat tinggal yang dapat menunjukkan penyebaran dari suku tersebut. maka untuk mempermudah pemetaan suku yang ada di indonesia dapat menggunakan bantuan algoritma graph yang dapat menghubungan pemetaan antar suku serta daerah yang ditempati.

Penelitian ini akan membahas tentang pemetaan suku di Indonesia dengan menggunakan algoritma graph dimana data yang digunakan adalah data-data suku bangsa di Indonesia yang bersumber dari DbPedia dengan pengambilan data menggunakan sparql.

2. METODE PENELITIAN

2.1. Dasar Teori

2.1.1. SPARQL

SPARQL adalah akronim untuk Simple Protocol And RDF Query Language. Standar SPARQL mendefinisikan protokol jaringan untuk bertukar query dan bahasa untuk mengekspresikan query. SPARQL mengadopsi sintak SQL-like untuk mengekspresikan query atau untuk mengambil data yang ditulis menggunakan RDF atau XML. Dengan adanya SPARQL, maka masing-masing sumber data bisa terhubung satu dengan lainnya.

Pada dasarnya, SPARQL merupakan query untuk mencocokan bentuk graph. Graph tersebut dicocokan dengan berbagai endpoint dari repository yang dituju untuk mencari kemungkinan dari solusi yang ada (The W3C SPARQL, 2013). SPARQL endpoint disediakan untuk melakukan query pada knowledge base DBpedia Indonesia. Untuk melakukan query sebaiknya mempelajari ontology yang digunakan DBpedia (Grover & Leskovec, 2016).

2.1.2 Node2Vec

Node2Vec merupakan framework algoritma yang digunakan untuk merepresentasikan node pada graph (Grover & Leskovec, 2016). Node2Vec belajar low-dimensional representations yaitu dengan melakukan optimasi neighborhood (Van Der Maaten & Hinton, 2008). Dalam menentukan target Node2Vec menentukannya secara acak. Dalam aplikasi machine learning Node2Vec dapat digunakan untuk memberikan prediksi (Grover & Leskovec, 2016).

2.1.3 Dbpedia

Dbpedia adalah komunitas yang bergerak untuk mengekstrak informasi terstruktur dari Wikipedia dan menyediakan informasi tersebut dalam sebuah web (Golbeck, 2013).

2.1.4 Algoritma Closeness Centrality

Algoritma yang digunakan untuk mengukur jarak rata - rata ke semua node yang ada. Node yang mempunyai skor atau nilai kedekatan paling tinggi akan memiliki jarak terpendek ke semua node yang ada (Golbeck, 2013). Berikut ini adalah rumus algoritma closeness centrality :

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}.$$

2.2 Metodologi

2.2.1 Query SPARQL

Data query yang diambil bersumber dari dbpedia.id yaitu mengenai suku-suku di Indonesia, query akan menampilkan nama suku, bahasa dan lokasi. Berikut adalah query yang digunakan sesuai dengan Query Sparql dibawah ini.

```
select distinct ?nama_suku ?bahasa ?lokasi
where
{
?suku dcterms:subject
<http://id.dbpedia.org/resource/Kategori:Suku_bangsa_di_Indonesia>.
?suku rdfs:label ?nama_suku.
?suku dbpprop-id:langs ?bahasa.
?suku dbpprop-id:popplace ?lokasi.
}
```

Berikut ini beberapa data query dalam bentuk CSV sesuai dengan yang digambarkan oleh tabel 1, data tersebut bisa dilihat di zenodo pada link berikut:https://zenodo.org/deposit/3810747

E-ISSN 2460-9552
P-ISSN 2620-3227

Tabel 1.Data query

Nama Suku	Bahasa	Lokasi
Orang Kanekes	Dialek baduy dari sunda	Banten
Suku Bali	Bahasa bali, bahasa sasak	Bali
Suku Batin	Melayu Jambi	Jambi
Suku Bauzi	Bahasa Bauzi	Papua
Suku Berau	Bahasa Banjar	Kalimantan Timur
Suku Betawi	Bahasa betawi	Jakarta
Suku Buru	Bahasa buru	Buru
Suku Dani	Bahasa Dani	Papua
Suku Melayu	Bahasa melayu	Indonesia
Suku Gayo	Bahasa gayo	Aceh Tengah

2.2.2 Pemodelan Graph

Pemodelan graph yang digunakan adalah sebagai berikut sesuai dengan gambar 1.

Model Graph

Gambar 1. Model Graph

Pada gambar 1 Model Graph dapat dilihat bahwa node yang digunakan ada tiga yaitu node suku, node lokasi dan node bahasa dimana suku memiliki relasi dengan node bahasa dan node lokasi

2.2.3 Menjalankan Node2Vec

Dalam menjalankan node2vec dipergunakan parameter sebagai berikut :
dimensions = 20
walk_length=16
num_walks=100
workers=2
dengan *code* sesuai node2vec code dibawah ini.

```
node2vec = Node2Vec(g, dimensions=20,
walk_length=16, num_walks=100,
workers=2)
model = node2vec.fit(window=10,
min_count=1)
```

2.2.4 Menjalankan Algoritma Graph

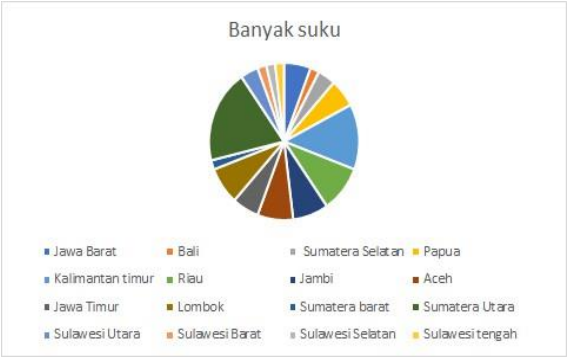
Algoritma yang kami gunakan adalah closeness centrality yang bertujuan untuk mengukur jarak rata - rata antara node yang satu dengan yang lain.

2.2.5 Visualiasi

Visualisai merupakan penggambaran data yang diperoleh setelah di proses dengan algoritma graph, dalam visualisasi ini menggunakan TSNE (Grover & Leskovec, 2016) dan Matplotlib.

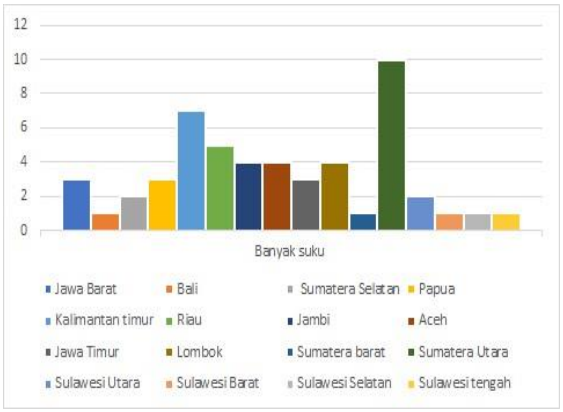
3. HASIL DAN PEMBAHASAN

Dengan menggunakan query SPARQL, didapatkan dataset suku di indonesia berdasarkan bahasa dan daerah dari tiap - tiap suku.



Gambar 2. statistika dataset

Berdasarkan gambar diagram 2 diatas, dapat dilihat bahwa suku paling banyak berasal dari daerah sumatera utara

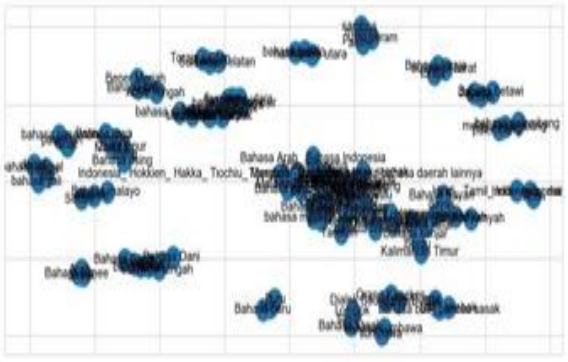


Gambar 3. statistika dataset

Berdasarkan data pada gambar 3 dapat dilihat bahwa jumlah suku yang ada didalam dataset adalah 52 suku yang berasal dari berbagai daerah di Indonesia. Dataset yang dipakai juga mempunyai hubungan atau relasi dengan bahasa berdasarkan suku.

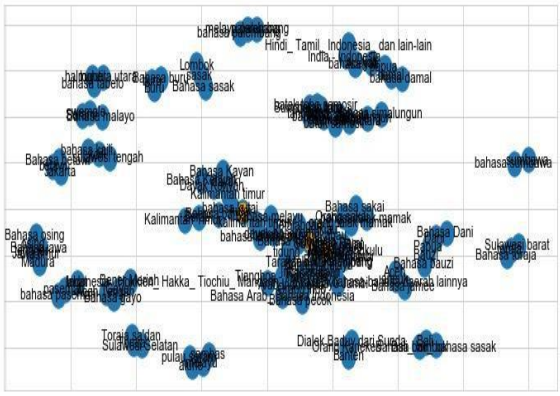
```
fig, ax = plt.subplots(figsize=(12,5))
layout = nx.spring_layout(g,iterations=50)
nx.draw_networkx_nodes(g, layout, ax =
ax, labels=True)
nx.draw_networkx_edges(g, layout, ax=ax)
_ = nx.draw_networkx_labels(g, layout,
labels, ax=ax)
```

Berdasarkan dataset diatas dengan menggunakan kode tersebut, ditemukan persebaran data menggunakan graph seperti gambar 4 dibawah ini :



Gambar 4. Graph Dataset

Setelah dilakukan analisis berdasarkan graph tersebut ditemukan 3 suku yang populer yaitu melayu, kutai dan kerinci. Adapun untuk visualisasi graph nya adalah sebagai berikut :



Gambar 5. Data Populer

Adapun kodenya adalah sebagai berikut

```
sukupopuler = [suku for suku in df.suku if
g.degree(suku) > 3]
nx.draw_networkx_nodes(g, layout,
nodelist=sukupopuler, node_color='orange',
node_size=150)
print (sukupopuler)
```

Suku populer pada gambar 5 diatas ditandai dengan data yang berwarna orange.

Berdasarkan data pada gambar 5, didapatkan node yang mempunyai kesamaan dengan melayu adalah sebagai berikut :

```
for node, _ in model.most_similar('Melayu '):
print(node)
```

- Melayu Palembang
- Melayu jambi
- bengkulu
- Melayu Riau
- Melayu bengkulu
- Sumatera Barat
- Bahasa Minang
- Palembang
- Bahasa melayu
- Jambi

Sedangkan node yang mempunyai kesamaan dengan kutai adalah sebagai berikut :

```
for node, _ in model.most_similar('kutai '):
print(node)
```

E-ISSN 2460-9552
P-ISSN 2620-3227

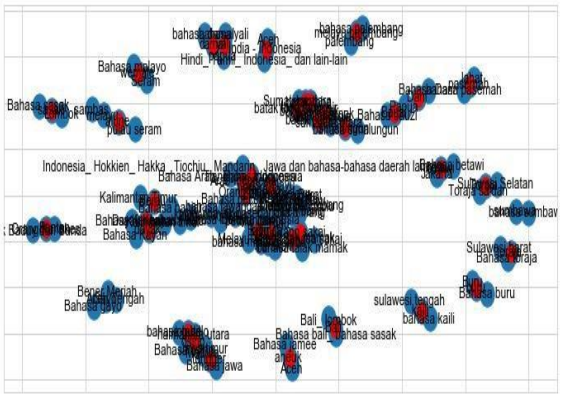
- bahasa kutai
- bahasa melayu
- Bahasa banjar
- Kalimantan timur
- kalimantan timur
- dayak
- Berau
- petalangan
- Dayak Kenyah
- dayak paser

Node yang mempunyai kesamaan dengan kerinci adalah sebagai berikut :

```
for node, _ in model.most_similar('kerinci '):
print(node)
```

- jambi
- bahasa minangkabau
- Jambi
- Bahasa indonesia
- Batin
- dayak paser
- petalangan
- Melayu Jambi
- tidung
- Melayu.

Dari dataset tersebut, juga terdapat suku - suku yang tidak populer yang divisualisasikan pada gambar 6 berikut ini.



Gambar 6. Graph Suku Tidak Popular

Berdasarkan gambar visualisasi diatas disimpulkan bahwa suku - suku yang tidak populer adalah sebagai berikut :

4. SIMPULAN

['Arab-Indonesia', 'Orang Kanekes', 'Bali ', 'Batin', 'Bauzi ', 'Berau ', 'betawi', 'buru', 'Tianghoa - Indonesia ', 'Dani ', 'India - Indonesia ', 'Orang indo ', 'dayak ', 'Dayak Kenyah ', 'Ras melayu ', 'osing ', 'Orang sakai ', 'sasak ', 'tengger ', 'tidung ', 'batak toba ', 'toraja ', 'Toraja ', 'wemale ', 'aneuk ', 'simalungun ', 'kaili ', 'orang talak mamak', 'batak samosir ', 'melayu palembang', 'togutil', 'pasemah ', 'alune ', 'batak pakpak', 'batak angkola ', 'damal', 'yali ', 'batak toba samosir ', 'tano toba ', 'karo']

Dari hasil penelitian diatas, dapat disimpulkan bahwa terdapat 3 suku yang popular yang mempunyai jarak rata - rata yang lebih dekat dibandingkan suku - suku yang lain yaitu suku melayu, suku kutai dan suk kerinci. Sehingga dapat dipastikan untuk persebaran suku tersebut di Indonesia lebih popular.

5. DAFTAR PUSTAKA

Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864).

Golbeck, J. (2013). Closeness Centrality Social network analysis : Measuring , mapping , and modeling collections of connections, Anal. Soc. Web, vol. 3.

Hansen, D. L., Shneiderman, B., Smith, M. A., & Himelboim, I. (2011). Social network analysis: measuring, mapping, and modeling collections of connections. Analyzing social media networks with NodeXL: insights from a connected world. Elsevier Inc, Burlington, 31-52.

Khosla, M., Setty, V., & Anand, A. (2019). A Comparative Study for Unsupervised Network Representation Learning. IEEE Transactions on Knowledge and Data Engineering.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579-2605.

Memahami Metode Centrality Dasar. [Online]. Available: https://budsus.wordpress.com/2013/05/01/m emahami-metode-centrality-dasar/.

node2vec: Scalable Feature Learning for Networks. Available : https://snap.stanford.edu/node2vec

World Wide Web Consortium. (2013). SPARQL 1.1 overview.