# Implementasi dan Analisis *Semantic Relatedness* Dengan Menggunakan Metode *Wikipedia Link Based* (WLM)

Falanta Angya Permana
S1 Teknik Informatika
Fakultas Teknik
Universitas Telkom
Bandung, Indonesia
Falantaangya@gmail.com

***Abstract*—** Pengukuran *Semantic Relatedness* pada saat ini digunakan untuk memperkirakan nilai dari kedekatan antar pasangan kata dan teks. Pasangan kata atau teks memiliki konsep atau makna yang dapat diperoleh dari ensiklopedia. Wikipedia menyediakan puluhan ribu konsep dan makna pada seluruh artikel didalamnya. Dengan menggunakan Wikipedia sebagai penyedia konsep maka kedekatan antar pasangan kata dapat diperikirakan. Pada tugas akhir ini menggunakan link pada wikipedia yang mengindikasikan relasi antar artikel sebagai dataset dalam perhitungan *semantic relatedness* dengan inputan berupa dua kata. Pengukuran *semantic relatedness* dilakukan dengan menggunakan metode *Wikipedia Link Based* dan dua pengukuran yaitu *TFxIDF Inspired* dan *Normalized Google Distance.* Pengujian dilakukan dengan mengkorelasikan nilai yang didapat terhadap dataset WordSim 353 sebagai standar dataset. Pengukuran *semantic relatedness* pada wikipedia dilakukan dengan cara membentuk relasi antar *link* artikel yang saling berhubungan sehingga memunculkan suatu konsep yang terkait pada dua inputan kata. Hasil korelasi yang diperoleh sistem yaitu sebesar 50,2% dengan menggunakan 353 koleksi pasangan kata, sedangkan nilai acuan yaitu nilai korelasi antara sistem Wikipedia Miner dengan dataset WS 353 yaitu sebesar 69%. Untuk dapat memaksimalkan nilai korelasi setiap pasang kata harus memiliki konsep dan relasinya masing-masing.

***Keywords*— Word Relatedness, Semantic Relatedness, Wikipedia Link Based, TFxIDF, Normalized Google Distance.**

## I. PENDAHULUAN

Untuk mengetahui keterkaitan antar pasangan kata maka perlu dilakukan pengukuran nilai *Semantic Relatedness*. Contoh nya pasangan kata *football* dan *soccer* memiliki nilai *semantic relatedness* berdasarkan *human rater* sebesar 9.03[2], lalu pasangan kata *stock* dan *jaguar* memiliki nilai *semantic relatedness* berdasarkan *human rater* sebesar 0.92[2], yang artinya pada pasangan kata *football* dan *soccer* memiliki keterkaitan yang tinggi, sedangkan pada pasangan kata *stock* dan *jaguar* memiliki nilai keterkaitan yang rendah. Pengukuran *Semantic Relatedness* pada saat ini digunakan untuk mencari keterkaitan antar pasangan kata, lalu digunakan pada pasangan teks yang bertujuan untuk mendefinisikan seberapa terkait antar pasangan teks. Teks yang dimaksud dapat berupa web artikel, koran, dan dokumen lainnya. Dengan mengetahui relasi antar pasangan teks dapat memudahkan pembaca untuk memperoleh informasi yang terkait tanpa harus mencari di seluruh dokumen yang ada.

*Semantic Relatedness* merupakan suatu nilai kedekatan dari suatu pasang kata yang di dapatkan dari proses perhitungan untuk mencari keterkaitan antara dua kata yang memperhitungkan berbagai faktor yang berkaitan dengan makna atau konsep yang disampaikan oleh kata tersebut dan mencari hubungan dari antara pasangan kata-kata. Ukuran keterkaitan antar kata antar kata diukur dari leksikal dan semantik nya. *Wikipedia Link Based Measure* merupakan teknik yang digunakan untuk mendapatkan nilai dari pengukuran *semantic relatedness*. WLM menggunakan Wikipedia sebagai kamus atau *world knowledge* pada suatu kata. Metode ini menggunakan *hyperlink* yang terdapat pada Wikipedia untuk perhitungan *semantic relatedness. Semantic relatedness* dari suatu pasang kata dapat diperoleh dari relasi antar link pada artikel terkait [1].

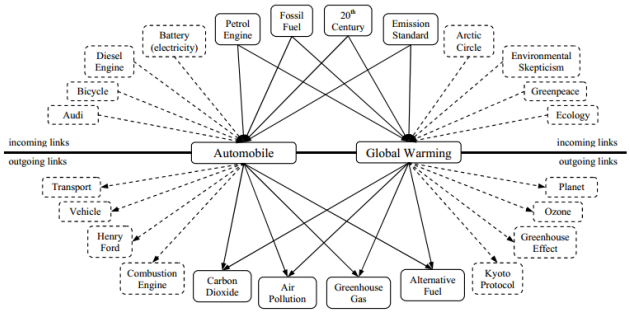
Dalam tugas akhir ini akan mengimplementasikan dan menganalisis pengukuran *Semantic Relatedness* dengan menggunakan *Wikipedia Link Based Measure*. Pengukuran *Semantic Relatedness* antar pasangan kata yang berupa judul artikel pada wikipedia dengan menggunakan *WLM* memberikan kemudahan dalam proses nya karena dataset tersedia dalam ukuran yang tidak terlalu besar. Pengukuran *Relatedness* dihitung dengan menggunakan dua pengukuran yaitu *TFxIDF Inspired* dan *Normalized Google Distance.* Pada pengukuran menggunakan *TFxIDF Inspired* data yang dibutuhkan yaitu berupa daftar relasi link pada *article of interest* terhadap artikel lain (*link out*). Sedangkan pada pengukuran *Normalized Google Distance* data yang dibutuhkan yaitu berupa daftar relasi link yang merujuk kepada *article of interest* (*link in).*

## II. DASAR TEORI DAN PERANCANGAN SISTEM

### 2.1 *Wikipedia Link Based* Measure

Pada penelitian ini menggunakan metode *Wikipedia Link Based*, Wikipedia *Link Based Measure* merupakan suatu teknik yang digunakan untuk mendapatkan nilai dan pengukuran dari *semantic relatedness*. WLM menggunakan Wikipedia sebagai kamus atau *world knowledge* tentang suatu

kata. Metode ini menggunakan *hyperlink* yang terdapat pada Wikipedia untuk perhitungan *semantic relatedness.*

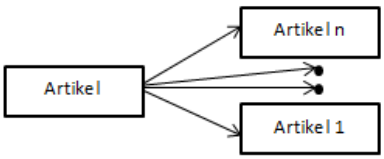



Gambar 2.1 Keterkaitan antara artikel

*Semantic relatedness* dari suatu pasang kata dapat diperoleh dari relasi antar link pada artikel terkait. Pada implementasinya, pada metode WLM terdapat dua pengukuran *semantic relatedness* yaitu *TF.IDF Inspired Measure* dan *Normalized Google Distance Measure* [1]. Penjelasan setiap pengukuran akan dijelaskan sebagai berikut:

**A. *TF.IDF***

*TF.IDF Inspired Measure* merupakan pengukuran kedekatan antara pasangan artikel yang mendefinisikan sudut antara vektor-vektor yang didapatkan dari pasangan artikel tersebut. Pengukuran ini pada dasarnya sama seperti TF.IDF yang biasanya digunakan pada *information retrieval* namun perbedaannya adalah vektor yang digunakan adalah nilai probabilitas *link count* dari setiap link yang muncul [1]. Probabilitas didefinisikan oleh jumlah link pada inputan artikel menuju seluruh artikel yang ada di wikipedia (*Link Out).*




Gambar 2.2 Gambaran *Link Out*

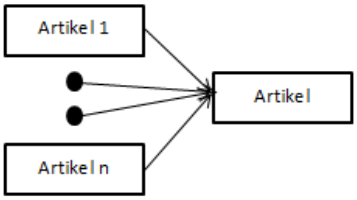
Rumus yang digunakan untuk menghitung bobot pada setiap link yang muncul yaitu :

$$w(s \to t) = log\left(\frac{|W|}{|T|}\right) \text{jika s} \in \text{T}, 0 \qquad (2.1)$$

Dimana $w(s \to t)$ adalah nilai bobot link dari artikel s terhadap semua artikel yang ada di Wikipedia, W merupakan jumlah keseluruhan artikel yang ada di Wikipedia, dan T merupakan kumpulan link s yang terhubung kepada t.
Pembobotan link digunakan untuk menghasilkan vektor-vektor yang akan menjadi inputan untuk *vector space model.*

**B. *Normalized Google Distance***

*Normalized Google Distance Measure* merupakan suatu pengukuran nilai *Semantic Relatedness* dengan menghitung jumlah link dari seluruh artikel yang ada di wikipedia yang berelasi dengan inputan artikel (*Link in).*




Gambar 2.3 Gambaran L*ink in*

Pasangan artikel yang terdeteksi mengandung link yang sama dapat diindikasikan bahwa pasangan artikel tersebut saling berelasi. Rumus dari *Normalized Google Distance Measure* [6] yaitu:

$$sr(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))} \qquad (2.2)$$

Dimana a dan b merupakan pasangan artikel yang berasal dari inputan, A merupakan jumlah link artikel lain yang terdapat pada artikel a , B merupakan jumlah link artikel lain yang terdapat pada artikel b dan W merupakan jumlah total artikel yang ada di Wikipedia. Terdapat beberapa aturan dalam perhitungan NGD yaitu nilai NGD tidak ada yang negatif dan jika pasangan kata adalah kata yang sama maka nilai NGD bernilai 0.

**2.2 Evaluasi Sistem**

Evaluasi sistem dilakukan dengan mengkorelasikan nilai yang diperoleh dari sistem dengan dataset WS 353 dan Wikipedia Miner. Korelasi yang digunakan yaitu *Pearson Product-Moment Correlation Coefficient* dengan formula sebagai berikut [8]:

$$r = \frac{\sum(x - x')(y - y')}{\sqrt{\sum(x - x')^2 \sum(y - y')^2}} \qquad (2.3)$$

Dimana x dan y merupakan sampel dari nilai dua array. Semakin dekat nilai yang didapatkan oleh sistem terhadap dataset, maka semakin baik nilai yang dihasilkan oleh sistem ini. Evaluasi sistem dilakukan pada dua kasus yaitu:

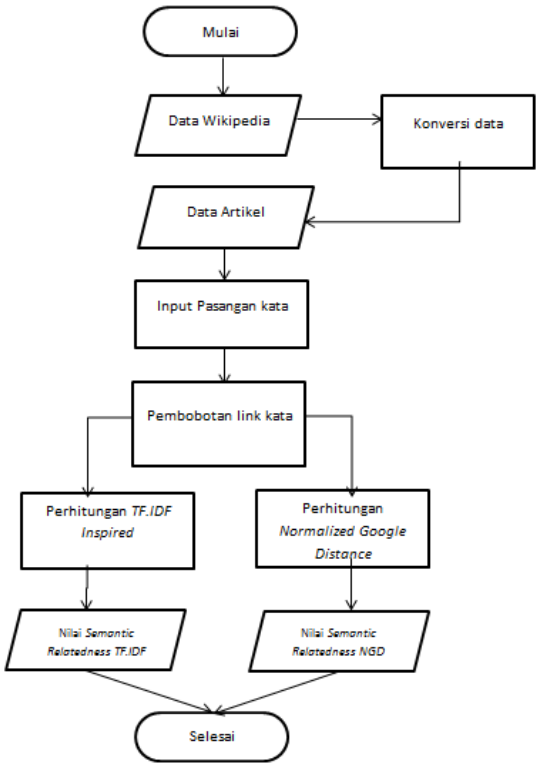
1. Hasil korelasi sistem dengan keseluruhan pasangan kata.
2. Hasil korelasi sistem dengan pemilihan nilai.

Untuk kasus yang pertama, evaluasi didapatkan dari keseluruhan hasil korelasi sistem dengan dataset sebanyak 353 pasangan kata. Sedangkan untuk kasus yang kedua evaluasi dilakukan dari hasil korelasi sistem dengan dataset dengan proses pemilihan. Proses pemilihan dilakukan dengan tidak mengkorelasikan nilai kedekatan yang bernilai 0 (tidak memiliki konsep pada dataset), sehingga pasangan kata akan kurang dari 353 kata. Kasus yang kedua dilakukan untuk meningkatkan nilai korelasi sistem terhadap dataset WS 353 dan Wikipedia Miner, karena pada dataset WS 353 seluruh

pasangan kata memiliki nilai kedekatan, namun pada dataset wikipedia dan Wikipedia Miner terdapat pasangan kata yang tidak memiliki konsep yang mengakibatkan nilai kedekatan dari pasangan tersebut bernilai 0.

### 2.3 Rancangan Sistem

Data input yang digunakan yaitu seluruh id beserta nama artikel yang ada pada wikipedia dan relasi dari setiap artikel terhadap artikel yang lain (*page to page link)*. Berikut ini gambaran umum perancangan sistem :




Gambar 3.1 Rancangan Sistem

Sistem akan membaca data link artikel dan judul artikel pada wikipedia sebagai *knowledge base*[8] dari setiap pasangan kata untuk pembobotan jumlah *link in* dan *link out* pada masing-masing kata, setelah pembobotan selesai nilai *semantic relatedness* akan dihitung dengan menggunakan pengukuran TF.IDF dan *Normalized Google Distance.*

#### 2.3.1. Data Wikipedia

Untuk memperoleh data yang berisi relasi tiap artikel yang ada pada Wikipedia dibutuhkan data *page to page link* pada Wikipedia yang sudah tersedia dalam format sql pada Wikipedia. Terdapat dua data yang dibutuhkan yaitu *page.sql.gz* dan *pagelinks.sql.gz* yang dapat di unduh dari Wikipedia yang berisi informasi link pada setiap artikel.

#### 2.3.2. Konversi Data

Data yang didapatkan dari Wikipedia akan di *convert* kedalam bentuk yang lebih mudah yaitu dalam format "*.txt".* Proses ini akan menghasilkan dua buah file *".txt"* yang berisi *page to page link* dan judul artikel yang ada pada Wikipedia.

#### 2.3.3 Pembobotan Link Artikel

Pembobotan link pada masing-masing inputan Artikel yang akan digunakan pada tahap perhitungan *Semantic Relatedness.* Tahap ini bertujuan untuk mencari *Link Out, Link In* , *Link Out* Irisan pada pasangan artikel, dan *Link In* Irisan pada pasangan artikel.

| Word 1 | Word 2 | Link out 1 | Link out 2 | Link out Irisan | Link in 1 | Link in 2 | Link in Irisan |
|---|---|---|---|---|---|---|---|
| love | sex | 261 | 115 | 3 | 1263 | 1218 | 58 |
| tiger | cat | 379 | 527 | 87 | 1120 | 1842 | 123 |
| tiger | tiger | 379 | 379 | 379 | 1120 | 1120 | 1120 |
| book | paper | 326 | 219 | 20 | 2353 | 1329 | 50 |
| computer | keyboard | 502 | 17 | 1 | 5002 | 54 | 1 |
| computer | internet | 502 | 377 | 26 | 5002 | 9297 | 351 |
| plane | car | 23 | 1 | 0 | 99 | 765 | 4 |
| train | car | 149 | 1 | 0 | 1541 | 765 | 33 |
| telephone | communication | 145 | 99 | 2 | 2153 | 1887 | 83 |

Tabel 3.1 Sample bobot link pada pasangan artikel

Suatu pasangan artikel dapat dikatakan berelasi dan memiliki keterkaitan jika banyak memiliki kesamaan pada *Link In* dan *Link Out.*

## III. PEMBAHASAN

Pengujian pada data link artikel pada Wikipedia sebagai pengukuran nilai *Semantic Relatedness* antar pasangan kata yang berupa judul artikel pada Wikipedia dengan menggunakan pengukuran TF.IDF *Inspired* dan *Normalized Google Distance* bertujuan untuk:

1. Menganalisa hasil pengujian sistem dari nilai *Semantic Relatedness* yang didapatkan dengan mengkorelasikannya terhadap dataset WS-353 dan sistem Wikipedia Miner.
2. Menganalisa pengaruh Jumlah *Link Out* dan *Link In* yang dihasilkan dari sistem terhadap nilai *Semantic Relatedness* dengan dua pengukuran *TF.IDF Inspired* dan *Normalized Google Distance.*

### 3.1 Evaluasi

Pada proses ini hasil dari perhitungan *Semantic Relatedness* oleh sistem menghasilkan 353 nilai *Semantic Relatedness* pada masing-masing pengukuran (TF.IDF & NGD). Nilai tersebut akan dikorelasikan terhadap dataset WS-353 dan sistem Wikipedia Miner dan akan dicari nilai akurasi nya terhadap

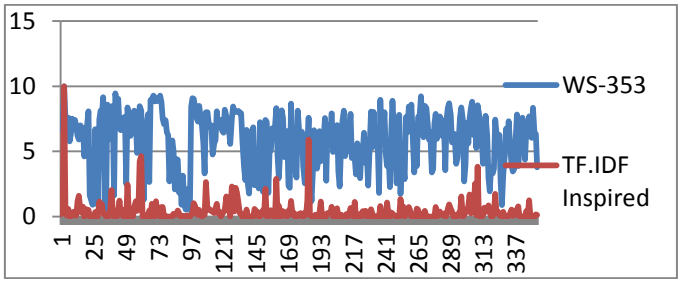
dataset tersebut. Range dari nilai akurasi yaitu 0 sampai 1, dimana 0 merupakan nilai terendah dan 1 merupakan nilai tertinggi.

### 3.2 Analisis Nilai Korelasi Sistem Pada Perhitungan Semantic Relatedness Dengan Menggunakan TF.IDF

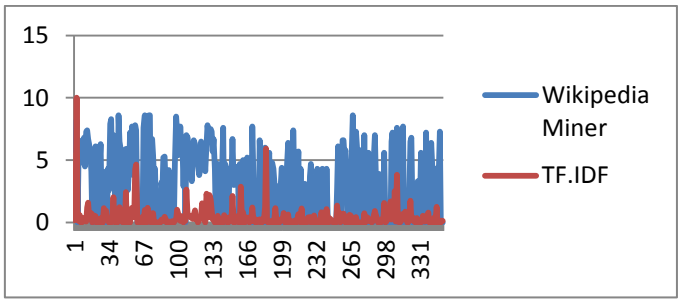
Dari hasil perhitungan Semantic Relatedness dari 353 pasang kata dengan menggunakan pengukuran TF.IDF didapatkan nilai akurasi sebagai berikut:

Tabel 3.1 Akurasi perhitungan TF.IDF Inspired *Full*

| Dataset | TF.IDF |
|---|---|
| WS-353 | 0.302235675 |
| Wikipedia Miner | 0.259531307 |




Grafik 3.1 Korelasi TF.IDF sistem terhadap dataset WS 353 *Full*




Grafik 3.2 Korelasi TF.IDF sistem terhadap dataset Wikipedia Miner *Ful*

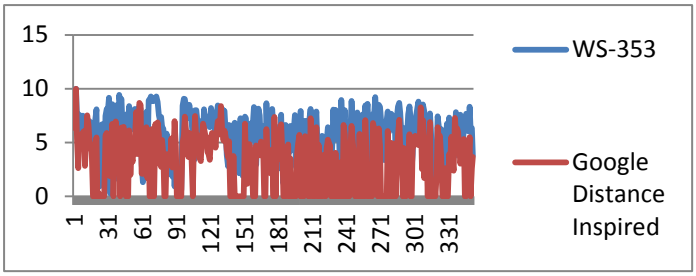
Tabel 3.1 menampilkan akurasi dari keseluruhan nilai *Semantic Relatedness* yaitu sebanyak 353 pasang kata tanpa menghilangkan nilai pasangan kata yang tidak terdapat pada Wikipedia dengan menggunakan TF.IDF.

### 3.3 Analisis Nilai Korelasi Sistem Pada Perhitungan Semantic Relatedness Dengan Menggunakan *Normalized Google Distance*

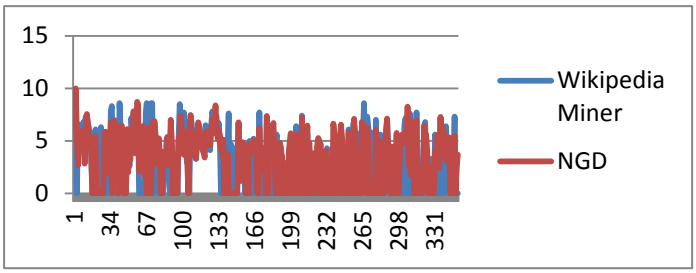
Dari hasil perhitungan Semantic Relatedness dari 353 pasang kata dengan menggunakan pengukuran Normalized Google Distance didapatkan nilai akurasi sebagai berikut:

Tabel 3.4 Akurasi perhitungan NGD full

| Dataset | NGD |
|---|---|
| WS-353 | 0.491440961 |
| Wikipedia Miner | 0.462938597 |




Grafik 3.5 Korelasi NGD sistem terhadap dataset WS 353 Full




Grafik 3.6 Korelasi NGD sistem terhadap dataset Wikipedia Miner Full

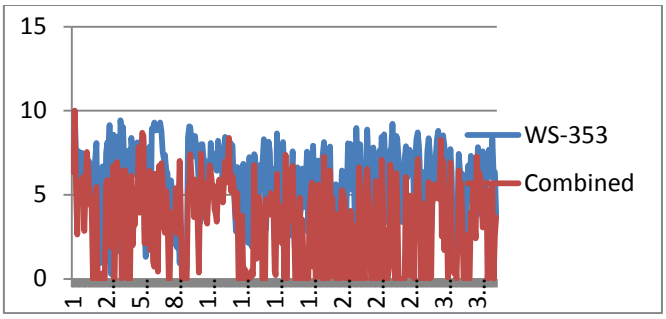
Tabel 3.4 menampilkan akurasi dari keseluruhan nilai Semantic Relatedness yaitu sebanyak 353 pasang kata tanpa menghilangkan nilai pasangan kata yang tidak terdapat pada Wikipedia dengan menggunakan NGD.

### 3.4 Analisis Nilai Korelasi Sistem Pada Perhitungan Semantic Relatedness Dengan Menggunakan TF.IDF dan *Normalized Google Distance*

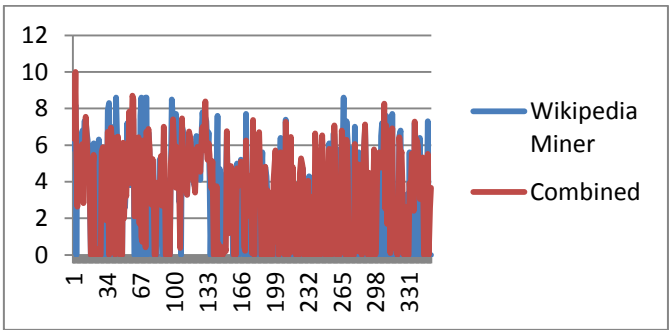
Setelah mendapatkan hasil pengukuran dari kedua pengukuran TF.IDF dan *Normalized Google Distance* dilakukan proses pengambilan nilai *Semantic Relatedness* yang maksimal dari kedua pengukuran tersebut, sehingga proses pemilihan pasangan kata yang bernilai 0 dapat diminimalisir dan dapat memperkaya jenis pasangan kata yang dapat diukur. Berikut hasil akurasi yang didapatkan dari nilai maksimal *Semantic Relatedness* yang diperoleh dari kedua pengukuran tersebut:

Tabel 3.5 Akurasi perhitungan NGD & TF.IDF *Full*

| Dataset | NGD & TF.IDF |
|---|---|
| WS-353 | 0.502243781 |
| Wikipedia Miner | 0.463470055 |




Grafik 4.9 Korelasi NGD & TF.IDF sistem terhadap dataset WS 353 *Full*




Grafik 4.10 Korelasi NGD & TF.IDF sistem terhadap dataset Wikipedia Miner *Full*

Tabel diatas merupakan hasil korelasi perhitungan NGD & TF.IDF terhadap dataset WS-353 dan Wikipedia Miner. Meski telah mengambil nilai maksimal dari kedua pengukuran tersebut tetap terdapat nilai *Semantic Relatedness* antar pasangan yang bernilai 0.

### 3.5 Analisis Pengaruh Jumlah Link Out dan Link In Pada Nilai Semantic Relatedness Dengan Dua Pengukuran TF.IDF & *Normalized Google Distance*

Setelah didapatkan hasil korelasi dari kedua pengukuran, NGD memberikan nilai akurasi yang lebih besar dibandingkan dengan pengukuran TF.IDF. Ini dapat disebabkan oleh jumlah *Link Out* yang tidak sebanding dengan jumlah *Link Out irisan*, dan jumlah *Link Out* dan *Link Out Irisan* memilki jumlah sedikit, artinya relasi yang terdapat antara artikel terkait dengan artikel yang dituju memiliki relasi yang kecil. Sedangkan pada *Link In* terdapat banyak terdapat relasi antara artikel-artikel lain yang berelasi dengan artikel terkait. Berikut jumlah data link yang diperoleh sistem:

Tabel 3.6 Jumlah link data

| Link | Jumlah Link | Link | Jumlah Link |
|---|---|---|---|
| Link Out Kata 1 | 60886 | Link In Kata 1 | 735494 |
| Link Out Kata 2 | 48822 | Link In Kata 2 | 740850 |
| Link Out Irisan | 3078 | Link In Irisan | 34168 |
| Jumlah | 112786 | Jumlah | 1510512 |

Tabel 3.6 menampilkan jumlah link dari 353 pasang kata. Jumlah *Link Out* lebih kecil dari pada jumlah *Link In* yang artinya *Link Out* memberikan relasi yang sedikit pada setiap artikel nya. Artinya hasil yang lebih baik didapatkan dengan pengukuran *Normalized Google Distance.*

## IV. SIMPULAN DAN SARAN

### 4.1 Kesimpulan

Hyperlink pada wikipedia dapat digunakan untuk menghasilkan fitur pada setiap inputan kata. Dari hasil pengujian pada 353 jenis pasangan kata yang berbedadidapatkan nilai korelasi terhadap dataset WordSim 353 sebesar 50,8% dengan pengukuran TF.IDF dan NGD (*combined*). Penurunan korelasi disebabkan oleh pasangan kata yang tidak memiliki konsep dan relasi pada hyperlink, untuk pemilihan nilai kedekatan yang bernilai 0 akan meningkatkan nilai korelasi, namun dapat menurunkan variasi jenis kata yang dapat dihitung kedekatan nya. Jadi semakin banyak jumlah link dan link yang saling beririsan maka semakin banyak konsep yang dimiliki oleh suatu kata yang berarti semakin besar relasi antara pasangan kata tersebut jika suatu konsep pada pasangan artikel saling terhubung dan nilai keterkaitan pun akan meningkat.

### 4.2 Saran

Untuk menambah *knowledge* dari dataset yang digunakan sebagai data untuk menghasilkan fitur fitur pada kata maka disarankan untuk menggunakan dataset Wikipedia yang bervariasi atau yang terbaru. Karena proses penghitungan *semantic relatedness* perpasangan nya membutuhkan waktu yang cukup lama maka disarankan untuk Membobotkan seluruh link dari setiap artikel terlebih dahulu, agar proses perhitungan tidak memakan waktu yang lama.

## REFERENSI

[1] D. Milne dan I. H. Witten, “An effective, low-cost measure of semantic relatedness”.

[2] E. Agirre, E. Alfonseca, K. Hall, M. Pasca dan A. Soroa, “ A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches,,” In Proceedings of NAACL-HLT 2009.

[3] H. Haselgrove, “Using the Wikipedia page-to-page link database, http://haselgrove.id.au/wikipedia.htm,” [Online].

[4] N. Kelly, “Information Retrieval Using Vector Spaces”.

[5] R. L. Cilibrasi dan P. M. Vitanyi, “The Google Similarity Distance,” *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 19, p. 370–383, MARCH 2007.

[6] R. J. Mooney, “Machine Learning Text Categorization,” 2006.

[7] W. B.Frakes dan R. Baeza-Yates, Information retrieval: data structures and algorithms, USA, 1992.

[8] Z. Zhang, A. L. Gentile dan F. Ciravegna, “Recent advances in methods of lexical semantic relatedness – a survey. Natural Language Engineering, 19, pp 411-479 doi:10.1017/S1351324912000125”.